Abstract of “Analyzing RNA-seq data using prior knowledge of gene and cell relationships” by Rebecca Elyanow, Ph.D., Brown University, May 24, 2020.

Ten years ago, the first RNA-seq study was published. Since then over 200 thousand RNA-seq studies have been published, spanning many different organisms, tissue types, and experimental conditions. However, until recently RNA-seq could only be used to investigate differences in gene expression between samples. This is because the expression of a sample is measured from pooled mRNA from hundreds of thousands to millions of cells. Recently, new RNA-seq technologies have begun to emerge, such as single cell RNA-seq (scRNA-seq), which allows for profiling of individual cells from a sample. This allows for the study of cellular heterogeneity within a tissue. Another new RNA-seq technology called Spatial Transcriptomics RNA-seq (STRNA-seq) profiles the mRNA transcripts from a tissue slice while retaining the spatial location of the transcripts in the tissue. Both methods produce high-dimensional transcript count matrices but are limited by extremely low coverage, with roughly 80% zero entries. In this dissertation, we introduce two methods that use known gene and cell dependencies to recover signal from scRNA-seq and STRNA-seq data. The first method, netNMF-sc, is a matrix factorization method which utilizes gene co-expression networks obtained from prior RNA-seq studies to perform dimensionality reduction and imputation of sparse scRNA-seq data, improving clustering performance and recovery of coexpressed genes over existing methods. The second method, STCNA, uses hidden Markov models to infer genomic copy number aberrations (CNAs) from STRNA-seq data of tumor tissues. Copy number aberrations, a subset of genomic rearrangements, are acquired as a tumor evolves and are a driving force of cancer development. STCNA uses spatial information to uncover subclonal CNAs, which are present in only a subset of cells in the tissue. Finally, we present a third method, NAIBR, which identifies genomic rearrangements, including those which do not result in copy number changes, such as inversions and translocations, from barcoded DNA sequencing data.

Analyzing RNA-seq data using prior knowledge of gene and cell relationships

by

Rebecca Elyanow

BS Computer Science and Computational Biology, Carnegie Mellon University, 2014

MS Computer Science, Brown University, 2017

A dissertation submitted in partial fulfillment of the

requirements for the Degree of Doctor of Philosophy

in the Center for Computational Molecular Biology at Brown University

Providence, Rhode Island

May 24, 2020

This dissertation by Rebecca Elyanow is accepted in its present form by
the Center for Computational Molecular Biology as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____
Benjamin Raphael, Director

Recommended to the Graduate Council

Date _____ _____
William Fairbrother, Reader

Date _____ _____
Ritambhara Singh, Reader

Date _____ _____
Ashley Webb, Reader

Acknowledgements

My PhD has been marked by changes, from moving to a new university to adapting to the new normal of working from home. Through all of these changes, the support from my family has remained constant. My parents have been there as a sounding board and have shared in the struggles and successes during my PhD and I am forever grateful for their love and support. My boyfriend Logan (who I would not have met if Ben had not decided to move to Princeton, thank you Ben!) has also helped me immensely in finishing this thesis. We have pushed each other during our PhDs, providing both emotional (and sometimes technical) support. My dog Naga also became an integral part of my experience as a grad student, coming with me to the office every day and brining joy into many people's lives. While I'm sure I could have completed my PhD without her company, it would have been a whole lot less fun. Of course, I am grateful for the support and guidance of my advisor Ben and everyone in the Raphael group. All of my work is not mine alone but the product of many meetings, revisions, and thoughtful discussions with Ben and others in the group. I am excited to move to the next chapter of my life which I'm sure will be filled with similar unexpected twists and turns but which I know I can handle with a little help.

Contents

⋆ Parts of this dissertation appeared in proceedings of conferences or in journals. In particular, Chapter 2 is an extended version of [2], and Chapter 4 is an extended version of [1].

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 The past and present of DNA sequencing

After the advent of Sanger sequencing in 1977 [17], new sequencing technologies are constantly being developed. The most commonly used sequencing technology today is called *next-generation sequencing* or *second generation sequencing*. This massively parallel sequencing technology starts by randomly segmented DNA into small fragments. An adapter is ligated to each fragment which then binds to a flow cell. Fragments are then amplified via PCR using fluorescently labelled nucleotides. As each nucleotide is incorporated, the flow cell is imaged and the emission from each cluster of amplified DNA fragments is recorded, where each nucleotide has a specific emission wavelength. Each sequenced fragment is called a *read*. This method of "sequencing by synthesis", implemented by Illumina and several other companies, produces highly accurate reads, with an error rate around 0.1%, but is limited in the length of fragment that can be synthesized. This means that reads from second generation sequencing technologies are limited to about 300bp. One technological development that helps overcome this limitation is *paired-end* reads. Paired-end sequencing involves synthesizing DNA from both ends of the fragment, resulting in two paired reads from each fragment, one from the beginning of the fragment and one from the end. Another technological advancement is the advent of *third generation sequencing* technologies, which allow sequencing of molecules up to 1Mb in length. There are two major players in third-generation, long-read, sequencing: PacBio and Oxford Nanopore. In PacBio sequencing, long molecules are sequenced by immobilizing each molecule in an individual well

along with a single DNA polymerase and recording fluorescence as nucleotides are synthesized. In Oxford Nanopore sequencing, changes in electrical signal are measured as a molecule passes through a pore, where each nucleotide has a distinct electrical signature. Both methods currently have higher cost than short-read sequencing and a significantly higher error rate ($10-15\%$).

More recently, another third-generation sequencing technology called *linked-read sequencing* was developed by the company 10X Genomics. This technology combines the low error rate of short-read sequencing with the long-range information of long-read sequencing. It does so by encapsulating long molecules in droplets of oil, each containing a unique string of nucleotides called a barcode. The molecules are then fragmented within the oil droplets and barcodes are ligated to each short fragment. Then, standard Illumina sequencing is performed to generate short paired-end reads which can be mapped back to their long molecule of origin. The limitation of this technology, however, is that each long molecule is only covered by a handful of short reads, resulting in coverage of only $0.1X$ per molecule. This is the first of a new type of sequencing technologies which we will call *high-dimensional* sequencing technologies, because we obtain a set of sequenced reads for thousands of individual molecules. These high-dimensional sequencing technologies are characterized (1) high resolution, in this case resolution at the molecule level, and (2) low-coverage, in this case coverage of about $0.1X$ per molecule.

1.1.2 The past and present of RNA sequencing

RNA-sequencing (RNA-seq), which sequences cDNA reverse-transcribed from mRNA transcripts, was developed significantly after DNA-sequencing in 2009 [184]. While RNA-seq is vital for understanding how DNA is spliced to form alternate isoforms, researchers performing RNA-seq are most often not interested in the sequences themselves, but rather in the relative abundances of mRNA transcripts in a sample. mRNA transcripts, the precursors to proteins, are transcribed from DNA at different rates depending on a cell's function. This results in a wide variety of distinct cell types, from neurons to skin cells to blood cells, which each have diverse and specific functions. Measuring the relative abundance of mRNA transcripts under different experimental conditions and disease states has led to countless discoveries including breakthroughs in drug discovery and prognostic gene signatures [16]. However, until recently RNA-seq could only be used to investigate differences in gene expression between samples. This is because the expression profile (quantity of mRNA molecules per gene) of a sample is measured from pooled mRNA from hundreds of thousands to millions of cells.

Like DNA-seq, new technologies for RNA-seq are constantly being developed and improved. The recent advent of single-cell RNA-sequencing (scRNA-seq) provides the ability to measure gene expression at the resolution of a single cell. scRNA-seq combines high-throughput single-cell isolation techniques with second-generation sequencing, enabling the measurement of gene expression in hundreds to thousands of cells in a single experiment. This capability overcomes the limitations of microarray and RNA-seq technologies, which measure the combined expression in a bulk sample, and thus is able to quantify heterogeneity of gene expression in individual cells and subpopulations of cells [184]. The advantages of scRNA-seq compared to bulk RNA-seq are tempered by undersampling of transcript counts in single cells due to inefficient RNA capture and low numbers of reads per cell. The output of a scRNA-seq experiment is a high-dimensional gene $\times$ cell matrix of transcript counts, where each column of the matrix represents the expression profile of a single cell. The coverage per cell from scRNA-seq experiments is extremely low. This results in a transcript count matrix which contains many *dropout events* which occur when no reads from a gene are sequenced in a particular cell, even though the gene is expressed in that cell. The frequency of dropout events depends on the sequencing protocol and depth of sequencing. Cell-capture technologies, such as Fluidigm C1, sequence hundreds of cells with high coverage (1-2 million reads) per cell, resulting in dropout rates $\approx 20-40\%$ [193]. Microfluidic scRNA-seq technologies, such as 10X Genomics' Chromium platform, Drop-Seq, and inDrops sequence thousands of cells with low coverage (1K-200K reads) per cell, resulting in higher dropout rates, up to 90% [194].

In addition to having low-coverage per cell, scRNA-seq is also limited by the fact that cells are dissociated from their tissue of origin prior to sequencing, so important information regarding a cell's location in the tissue as well as spatial relationships between cells is lost. To address this limitation, spatial transcriptomics, also called STRNA-seq was developed by [136]. With this technology, a tissue section is placed on an array comprising of a grid of spots. Each spot contains surface probes each with unique molecular barcodes. The mRNA within each circular spot, covering 10μm in diameter, is then quantified using standard RNA-seq protocols. STRNA-seq provides a gene$\times$spot transcript count matrix, where each spot contains the mRNA from about $10-100$ cells. Each spot is associated with a unique coordinate representing its spatial location in the tissue. Like scRNA-seq, STRNA-seq data also has very low coverage per spot, about .15X corresponding to about 80% zero entries, so analysis of STRNA-seq data poses similar challenges to scRNA-seq.

Another spatial mRNA quantification approach developed at roughly the same time is called seqFISH (sequential fluorescence in situ hybridization). In this technology, fluorescently labelled probes are hybridized

and then removed from complementary transcripts and the fluorescence of each cell is recorded. This technology allows for single-cell resolution of spatial transcriptional patterns, but is limited by in the number of genes that can be profiled per cell. Cells contain so many mRNA transcripts that only small proportion of genes (about $100-1000$) can be accurately measured due to optical crowding [20]. Recent improvements to the seqFISH protocol have demonstrated the ability to resolve up to $10,000$ genes by increasing the number of color channels from five to 60 "pseudocolors" [19]. This technology is called seqFish+ and it potentially offers a breakthrough in spatial mRNA quantification technologies due to its ability to capture mRNA expression at single-cell resolution with relatively high accuracy. However, there have currently been no published studies using seqFISH+ besides the original publication [19] while there have been many studies using the STRNA-seq technology [21, 22, 24, 125], so time will tell whether cost, ease of use, and momentum drive STRNA-seq or seqFISH+ to be the dominant technology for spatial mRNA profiling.

1.1.3 Using sequencing for personalized cancer diagnosis and treatment

The applications of DNA and RNA sequencing are numerous and varied. In this dissertation we will mainly focus on applications of these technologies to cancer and how they can be used to improve our understanding of tumor heterogeneity and give insights into cancer prognosis and treatment. Cancer is a disease characterized by the accumulation of deleterious mutations which cause cancer cells to divide uncontrollably and spread throughout the body. Some of these mutations are point mutations called *single nucleotide variants* (SNVs) which occur when a single nucleotide is changed. Most SNVs have no effect but some can alter the function of a protein which can lead to an increase in cell division or a disruption of cell-cycle checkpoints. Other mutations are called *genomic rearrangements* or *structural variants*. These are mutations that involve structural changes to the genome and include deletions, the removal of one or more nucleotides, insertions, the addition of one or more nucleotides (often duplicated from another location in the genome), and inversions, which change the orientation of a genomic segment. Collectively, structural variants affect a larger portion of the human genome than single nucleotide variants [84]. Inherited germline structural variants have been implicated in several diseases including Crohn's disease, rheumatoid arthritis, Type I diabetes, and autism [83, 89, 100]. In addition, somatic structural variants are common in cancer genomes [72, 86]. These include deletions of tumor suppressor genes and amplifications of oncogenes which can promote aggressive cell growth and drive the development of cancer. Cancer genomes can also undergo dramatic rearrangement events such as chromothripsis, the shattering and random repair of chromosomes in a single catastrophic event [95], or chromoplexy [68], both of which result in a large number of complex structural variants in a

cancer genome.

The identification of structural variants from high-throughput DNA sequencing data is generally more challenging than the identification of single nucleotide variants. This difficulty is primarily a result of the fact that many structural variants are significantly longer than the DNA sequence reads produced by second generation DNA sequencing technologies, whose fragment sizes are ~300-500 nucleotides. In addition, such reads are too short for *de novo* genome assembly. Thus, structural variants are inferred from atypical, or aberrant, alignments of reads to a reference genome.

Numerous methods have been developed over the past several years to identify different types of structural variants from paired-end whole-genome DNA sequencing. Most of these methods rely on first aligning the paired-end reads to the human reference genome and then looking for two signals: erroneously mapped reads and changes in read depth. Erroneously mapped reads include *discordant read-pairs*, where a pair of reads align too close, too far, or with the opposite orientation of what would be expected. These discordant read-pairs indicate that the reference genome and cancer genome to not match at genomic region between these read pairs. Split reads are another example of erroneously mapped reads, where a read has no continuous alignment to the reference genome but rather has at least two partial alignments. Both discordant read pairs and split reads are signatures of a *novel adjacency* in the cancer genome; that is, two intervals that are non-adjacent in the reference genome are adjacent in a cancer genome. Change in read depth is another signal of a subset of structural variants called *copy number aberrations* (CNAs). These include deletions and amplifications where a DNA segment is either removed or added, resulting in fewer or more read-pairs mapping to that region respectively.

Extensive work has gone in to identifying novel adjacencies and copy number aberrations from whole-genome DNA sequencing data of cancer data. Methods that utilize discorded paired reads and split reads include: BreakDancer [71], GASV [90], VariationHunter [75], Pindel [102], DELLY [85], and LUMPY [80]; many others are reviewed in [97]. Methods that use read depth to identify CNAs include BIC-Seq [101], CNVnator [66], and TITAN [111]. Other methods, such as GASVPro [92] and SV-Bay [78] combine signals from discordant read-pairs and read depth signals to identify structural variants. These methods are limited by the fact that many structural variants are significantly longer than the DNA sequence reads produced by second generation DNA sequencing technologies, causing some structural variants to not be reported by these methods. As described earlier, third-generation sequencing technologies can overcome this limitation, however current long-read technologies Oxford Nanopore and PacBio are often not used clinically due to their high cost and high error rate. However, linked-read sequencing is roughly the same cost and error rate

as paired-end sequencing, so it offers an attractive option when it comes to discovery of structural variants due to the addition of long-range information which can span many large-scale structural variants.

In addition to uncovering mutations underlying cancer development, much can be gleaned from studying the transcriptional profiles of tumors. Microarray and bulk RNA-seq have long been used to identify *gene signatures*, sets of genes with significant differences in expression specific to certain cancer types, subtypes, or responses to treatment. These gene signatures have been used to diagnose patients [11] as well as to determine which treatments are most appropriate for an individual patient [12, 13], and to predict prognosis or recurrence of cancer after treatment [14, 15]. These gene signatures are one of the initial applications of *personalized medicine*, where a patient's unique gene expression signature is used to tailor treatment.

However, gene signatures from bulk microarray or RNA sequencing are limited in their diagnostic and prognostic power due to heterogeneity within a tumor. Tumor heterogeneity refers to the existence of subpopulations of cells with distinct genotypes and phenotypes. These subpopulations, called clones or subclones, may have drastically different phenotypes and even a small subpopulation of highly malignant or drug-resistant cells can result in poor prognosis. If tumors containing drug-resistant subclones are treated with that drug, the resistant subclone, which is often highly malignant, will become dominant, resulting in worse prognosis than before treatment [10].

Single cell RNA-seq (scRNA-seq) and spatial transcriptomics (STRNA-seq) technologies can help uncover this tumor heterogeneity and lead to better personalized diagnostic and treatment options. Heterogeneity is typically first classified by clustering the cells (or spots in the case of STRNA-seq) and classifying the clusters into cell types and cell states based on genes which are differentially expressed between the clusters. Several methods have been developed to cluster scRNA-seq data, including BISCUIT [146], CIDR [170], and SC3 [9]. Standard linear regression methods such as limma [7] and DESeq [8] are typically used to determine differentially expressed genes between clusters. STRNA-seq provides the ability to not only profile tumor heterogeneity in gene expression but to understand the spatial landscape of a tumor tissue. This spatial landscape is important. For example, cells on the outside of the tissue are more susceptible to treatment, while cells at the center are more protected. Several methods have been developed to use this spatial information to identify spatially distributed differentially expressed genes. These include SpatialDE [123] and Spatial Variance Component Analysis (SVCA) [139].

1.2 Contributions

We have introduced several high-dimensional sequencing methods: linked-read DNA sequencing, single cell RNA sequencing (scRNA-seq), and spatial transcriptomics (STRNA-seq). Each of these sequencing technologies offer attractive opportunities to understand cell biology, and specifically cancer biology, with unprecedented precision. However, each of these technologies suffer from low-coverage, which limits the applicability of existing methods. Linked-read sequencing data has low-coverage per molecule, with about 0.1X coverage on average. scRNA-seq data has low-coverage of each cell, with about 90% of genes having zero counts. Similarly, STRNA-seq data has low-coverage per spot, with about 80% of genes having zero counts.

This motivates the need for analysis methods designed specifically for these technologies, which can make use of the benefits of high-dimensional data and overcome the negative effect of low-coverage. To address this need, we have developed three methods for these three sequencing technologies which utilize known dependencies to improve the analysis and interpretation of sparse high-dimensional sequencing data.

We first introduce a matrix factorization method netNMF-sc, which makes use of known correlations in expression between gene pairs obtained from prior RNA-seq and microarray experiments. By incorporating gene-gene correlations from prior experiments in the form of a gene coexpression network, netNMF-sc is able to accurately recover cell clusters from scRNA-seq data.

We next introduce our method STCNA which uses prior knowledge of gene and spot dependencies to infer copy number aberrations (CNAs) from spatial transcriptomics RNA-seq (STRNA-seq) data. Unlike netNMF-sc, the prior knowledge of these dependencies comes directly from the dataset of interest. This is ideal because this information will always be available for any STRNA-seq experiment of any organism, whereas there may be limited prior knowledge available from rarely studied organisms/tissues for use with netNMF-sc. To our knowledge, STCNA is the first method which incorporates spatial information to infer CNAs from STRNA-seq data.

Finally, we introduce our method NAIBR which infers novel adjacencies created by structural variants in a tumor genome from linked-read sequencing data. Linked-read sequencing data consists of barcoded paired-end reads which originate from long molecules $\sim$ 50Kb in length. The probability of a paired-end read originating from a molecule that spans a novel adjacency is dependent on paired-end reads of other molecules sharing the same barcode. NAIBR incorporates these dependencies into a probabilistic model to infer the most likely set of novel adjacencies in the data. We show that by incorporating these dependencies,

NAIBR outperforms other methods at recovering validated novel adjacencies from tumor genomes.

Chapter 2

netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis

Abstract

Single-cell RNA-sequencing (scRNA-seq) enables high throughput measurement of RNA expression in single cells. However, due to technical limitations, scRNA-seq data often contain zero counts for many transcripts in individual cells. These zero counts, or *dropout events*, complicate the analysis of scRNA-seq data using standard methods developed for bulk RNA-seq data. Current scRNA-seq analysis methods typically overcome dropout by combining information across cells in a lower dimensional space, leveraging the observation that cells generally occupy a small number of RNA expression states. We introduce netNMF-sc, an algorithm for scRNA-seq analysis that leverages information across *both* cells and genes. netNMF-sc learns a low-dimensional representation of scRNA-seq transcript counts using network-regularized non-negative matrix factorization. The network regularization takes advantage of prior knowledge of gene-gene interactions,

encouraging pairs of genes with known interactions to be nearby each other in the low-dimensional representation. The resulting matrix factorization imputes gene abundance for both zero and non-zero counts and can be used to cluster cells into meaningful subpopulations. We show that netNMF-sc outperforms existing methods at clustering cells and estimating gene-gene covariance using both simulated and real scRNA-seq data, with increasing advantages at higher dropout rates (e.g., above 60%). We also show that the results from netNMF-sc are robust to variation in the input network, with more representative networks leading to greater performance gains.

2.1 Introduction

Single-cell RNA-sequencing (scRNA-seq) technologies provide the ability to measure gene expression within and among organisms, tissues, and disease states at the resolution of a single cell. These technologies combine high-throughput single-cell isolation techniques with second-generation sequencing, enabling the measurement of gene expression in hundreds to thousands of cells in a single experiment. This capability overcomes the limitations of microarray and RNA-seq technologies, which measure the average expression in a bulk sample, and thus have limited ability to quantify gene expression in individual cells or subpopulations of cells present in low proportion in the sample [184].

The advantages of scRNA-seq are tempered by undersampling of transcript counts in single cells due to inefficient RNA capture and low numbers of reads per cell. The result of scRNA-seq is a gene $\times$ cell matrix of transcript counts containing many *dropout events* that occur when no reads from a gene are measured in a cell, even though the gene is expressed in the cell. The frequency of dropout events depends on the sequencing protocol and depth of sequencing. Cell-capture technologies, such as Fluidigm C1, sequence hundreds of cells with high coverage (1-2 million reads) per cell, resulting in dropout rates $\approx 20-40\%$ [193]. Microfluidic scRNA-seq technologies, such as 10x Genomics Chromium platform, Drop-Seq, and inDrops sequence thousands of cells with low coverage (1K-200K reads) per cell, resulting in higher dropout rates, up to 90% [194]. Furthermore, transcripts are not dropped out uniformly at random, but in proportion to their true expression levels in that cell.

In recent years, multiple methods have been introduced to analyze scRNA-seq data in the presence of dropout events. The first three steps that constitute most scRNA-seq pipelines are: (1) imputation of dropout events; (2) dimensionality reduction to identify lower-dimensional representations that explain most of the variance in the data; (3) clustering to group cells with similar expression. Imputation methods include

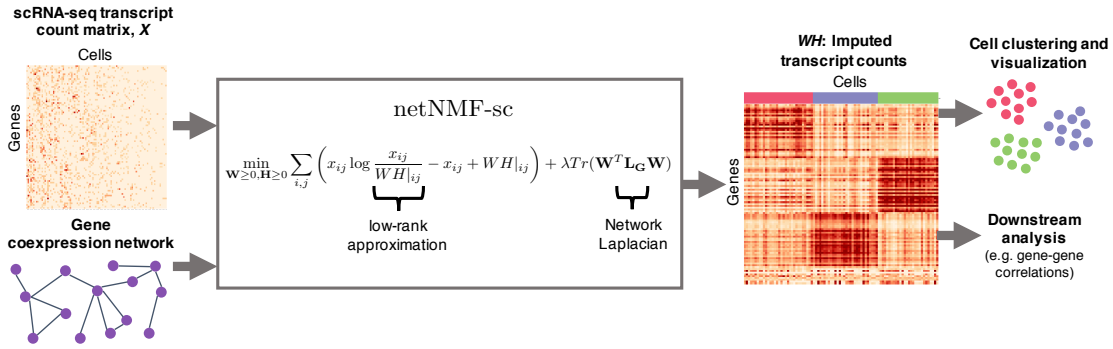


Figure 2.1: Overview of netNMF-sc. Inputs to netNMF-sc are: a transcript count matrix $\mathbf{X}$ from scRNA-seq and a gene coexpression network. netNMF-sc factors $\mathbf{X}$ into two lower-dimensional matrices, a gene matrix $\mathbf{W}$ and a cell matrix $\mathbf{H}$, using the network to constrain the factorization. The product matrix $\hat{\mathbf{X}} = \mathbf{WH}$ imputes dropped out values in the transcript count matrix $\mathbf{X}$. $\mathbf{H}$ is useful for clustering and visualizing cells in lower-dimensional space while $\mathbf{WH}$ is useful for downstream analysis such as quantifying gene-gene correlations.

MAGIC [182], a Markov affinity-based graph method, scImpute [168], a method that distinguishes dropout events from true zeros using dropout probabilities estimated by a mixture model, and SAVER [160], a method that uses gene-gene relationships to infer the expression values for each gene across cells. Dimensionality reduction methods include ZIFA [175], a method that uses a zero-inflated factor analysis model, SIMLR [183], a method that uses kernel based similarity learning, and two matrix factorization methods, pCMF [151] and scNBMF [179], which use a gamma-Poisson and negative binomial model factor model respectively. Clustering methods include BISCUIT, which uses a Dirichlet process mixture model to perform both imputation and clustering [146], and CIDR, which uses principal coordinate analysis to cluster and impute cells [170]. Other methods, such as Scanorama, attempt to overcome limitations of scRNA-seq by merging data across multiple experiments [158].

We introduce a new method, netNMF-sc, which leverages prior information in the form of a gene co-expression or physical interaction network during imputation and dimensionality reduction of scRNA-seq data. netNMF-sc uses network-regularized non-negative matrix factorization (NMF) to factor the transcript count matrix into two low-dimensional matrices: a gene matrix and a cell matrix. The network regularization encourages two genes connected in the network to have a similar representation in the low-dimensional gene matrix, recovering structure that was obscured by dropout in the transcript count matrix. The resulting matrix factors can be used to cluster cells and impute values for dropout events. While netNMF-sc may use any type of network as prior information, a particularly promising approach is to leverage tissue-specific gene coexpression networks derived from earlier RNA-seq and microarray studies of bulk tissue, and recorded in

large databases such as COXPRESdb [174], COEXPEDIA [188], GeneSigDB [149], and others [166, 186]. netNMF-sc provides a flexible and robust approach for incorporating prior information about genes in imputation and dimensionality reduction of scRNA-seq data.

2.2 Results

2.2.1 netNMF-sc algorithm

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a matrix of transcript counts from an scRNA-seq experiment for m transcripts and n single cells. It has been observed that the majority of variation in transcript counts is explained by a small number of gene expression signatures that represent cell types or cell states. Since $\mathbf{X}$ is a non-negative matrix, non-negative matrix factorization (NMF) [165] can be used to find a lower dimensional representation. NMF factors $\mathbf{X}$ into an $m \times d$ gene matrix $\mathbf{W}$ and a $d \times n$ cell matrix $\mathbf{H}$, where $d \ll m, n$, and the elements of both $\mathbf{W}$ and $\mathbf{H}$ are non-negative. We formulate this factorization as a minimization problem,

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{\mathbf{WH}|_{ij}} - x_{ij} + \mathbf{WH}|_{ij} \right), \tag{2.1}$$

where $\geq$ indicates non-negative matrices whose entries are ≥ 0.

The original NMF publication [165] proposed two cost functions to measure the difference between $\mathbf{X}$ and $\mathbf{WH}$: the Kullback-Leibler (KL) divergence given above and the Euclidean distance, $||\mathbf{X} - \mathbf{WH}||^2$. We use KL divergence because it is equivalent to maximizing the likelihood of the Poisson model $x_{ij} \sim Pois(\hat{x_{ij}})$, where $\hat{\mathbf{X}} = \mathbf{WH}$ [152]. The Poisson distribution [181] and the negative binomial distribution [157, 180] without zero inflation have been shown to provide a good fit for droplet-based transcript (UMI) count data. The Poisson model can be applied directly to transcript count matrices, eliminating the need to log-transform the transcript counts to better fit a Gaussian distribution [168, 176].

Log-transformation has been shown to introduce bias transcript in count data [157, 181]. Due to high dropout rates and other sources of variability in scRNA-seq data, the direct application of NMF to the transcript count matrix $\mathbf{X}$ may lead to components of $\mathbf{W}$ and $\mathbf{H}$ that primarily reflect technical artifacts rather than biological variation in the data. For example, [153] observe that the number of dropped out transcripts in a cell is the primary source of variation in several scRNA-seq experiments.

To reduce the effect of technical artifacts on the factorization, we propose to combine information across transcripts using prior knowledge in the form of a gene-gene interaction network. We incorporate network

information using graph regularized NMF [148] which includes a regularization term to constrain $\mathbf{W}$ based on prior knowledge of gene coexpression. The resulting method, netNMF-sc, performs matrix factorization by solving the following optimization problem:

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{\mathbf{WH}|_{ij}} - x_{ij} + \mathbf{WH}|_{ij} \right) + \lambda Tr(\mathbf{W}^T \mathbf{L} \mathbf{W}), \tag{2.2}$$

where λ is a positive real constant, $\mathbf{L}$ is the Laplacian matrix of the gene-gene interaction network, and $Tr(\cdot)$ indicates the trace of the matrix.

netNMF-sc uses the resulting matrix $\mathbf{H}$ to cluster cells, and the product matrix $\hat{\mathbf{X}} = \mathbf{WH}$ to impute values in the transcript count matrix $\mathbf{X}$, including dropout events (Fig 1).

We also derive a formulation of netNMF-sc with the Euclidean distance cost function $||\mathbf{X} - \mathbf{WH}||^2$ (Section 2.3.2), which is useful for (log-transformed) data with zero-inflation; e.g., read count data lacking UMIs. We show that netNMF-sc with the Euclidean distance cost function has similar clustering performance (ARI) to netNMF-sc with the KL divergence cost function on read count data from [147] (Fig 2.2A-D).

We select the regularization parameter λ as well as the dimension d of the factor via holdout validation (see Section 2.3.4 and Fig 2.3).

2.2.2 Evaluation on simulated data

We compared netNMF-sc and several other methods for scRNA-seq analysis on a simulated dataset containing 5000 genes and 1000 cells and consisting of 6 clusters with 300, 250, 200, 100, 100, and 50 cells per cluster respectively. We generated this data using a modified version of the SPLATTER simulator [189], modeling gene-gene correlations using a gene coexpression network from [188]. We simulated dropout events using one of two models: a *multinomial dropout model* [171, 192] and a *double exponential dropout model* [146, 168]. Further details are in Methods.

We compared the performance of netNMF-sc to PCA, scNBMF, MAGIC, scImpute, and NMF at dropout rates ranging from 0 (no dropout) to 0.80 (80% of the values in the data are zero), using 20 simulated datasets for each dropout rate. We clustered the output from each method using k-means clustering with $k = 6$ to match the number of simulated clusters (See Section 2.3.6 for more details on clustering). We selected $d = 10$ for NMF, scNBMF, and netNMF-sc and $\lambda = 10$ for netNMF-sc based on holdout validation (see Section 2.3.4).

For netNMF-sc, we used a randomly selected subnetwork $S = (V_S, E_S)$ of the same gene coexpression network $G = (V, E)$ [188] used to create the simulated data (see Section 4.4). This is intended as a positive

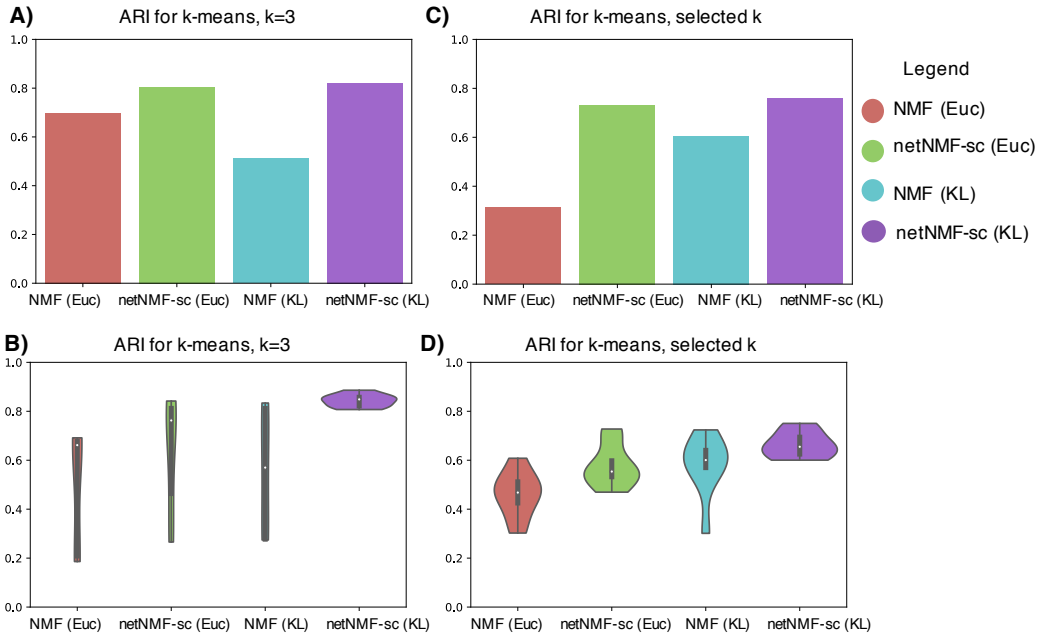


Figure 2.2: A) Clustering performance of NMF and netNMF-sc on scRNA-seq of 182 cells from [147] with Euclidean (Euc) and KL divergence cost functions, and k-means clustering with $k = 3$. The factor matrices $\mathbf{W}$ and $\mathbf{H}$ are randomly initialized by sampling i.i.d from the standard normal distribution, taking the absolute value of each entry to ensure non-negativity. The result that minimizes the netNMF-sc objective value across 10 random initializations is displayed. B) Variance in clustering performance across 10 initializations of NMF or netNMF-sc. C) Clustering performance of NMF and netNMF-sc with Euclidean and KL divergence distance functions clustered with k-means. For each initialization, the k which produces the highest silhouette score within the range $2 \leq k \leq 20$ is selected. D) Variance in clustering performance across 10 initializations of NMF or netNMF-sc with k selected using silhouette score.

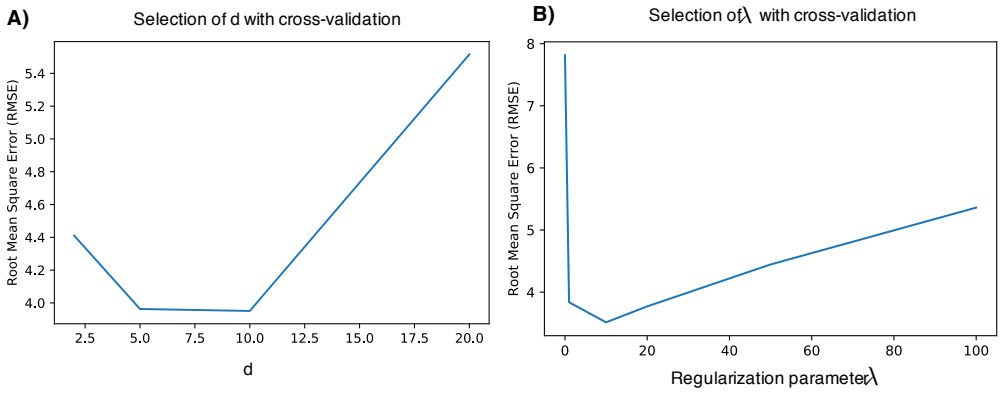


Figure 2.3: A) RMSE between held-out entries of $\mathbf{X}$ and corresponding imputed entries of $\mathbf{WH}$ on simulated data. Here $d = 10$ has the lowest root mean squared error. B) RMSE between held-out entries of $\mathbf{X}$ and corresponding imputed entries of $\mathbf{WH}$ with $d = 10$. Here $\lambda = 10$ has the lowest RMSE.

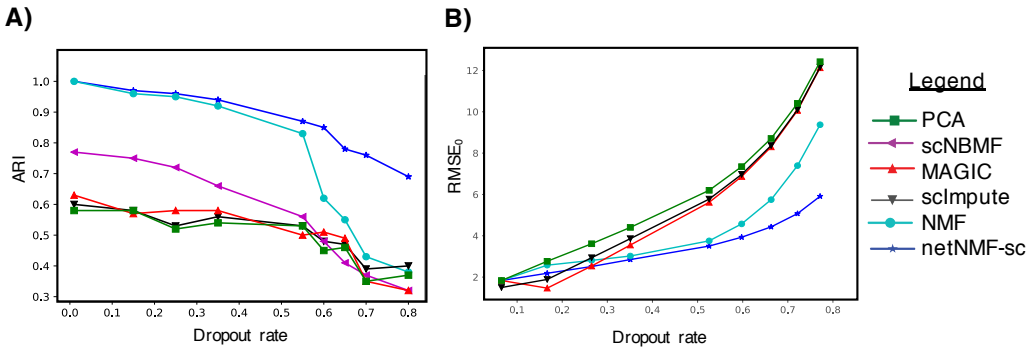


Figure 2.4: Comparison of netNMF-sc and other methods on a simulated scRNA-seq dataset containing 1000 cells and 5000 genes, with dropout simulated using a multinomial dropout model. (A) Adjusted Rand Index (ARI) between the true and inferred cell clusters obtained as a function of dropout rate. (B) Root Mean Square Error at dropped-out entries ($RMSE_0$) between true and imputed transcript counts.

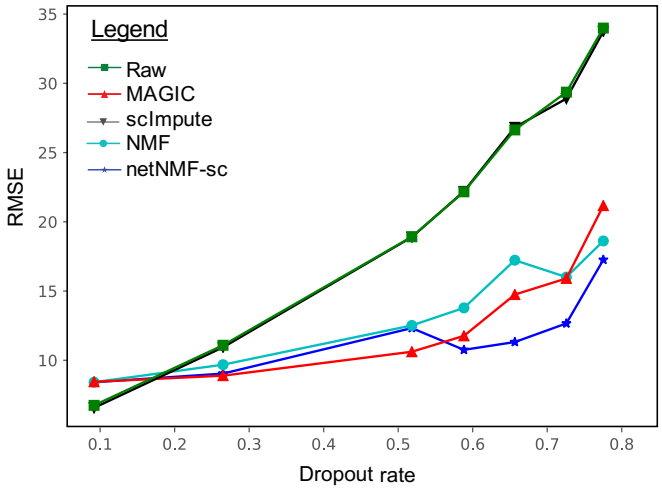


Figure 2.5: Root mean square error (RMSE) on simulated data using the *multinomial dropout model*.

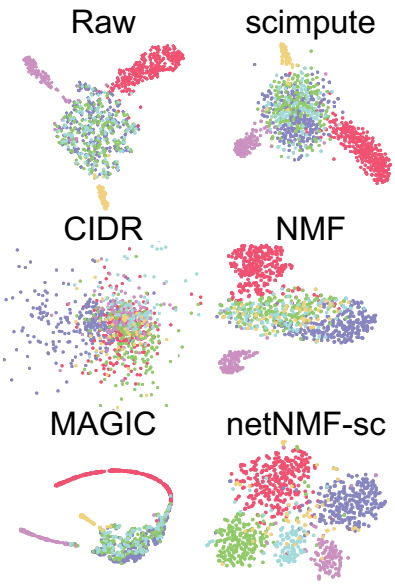


Figure 2.6: t-SNE projections of imputed simulated data with 5 simulated cell clusters.

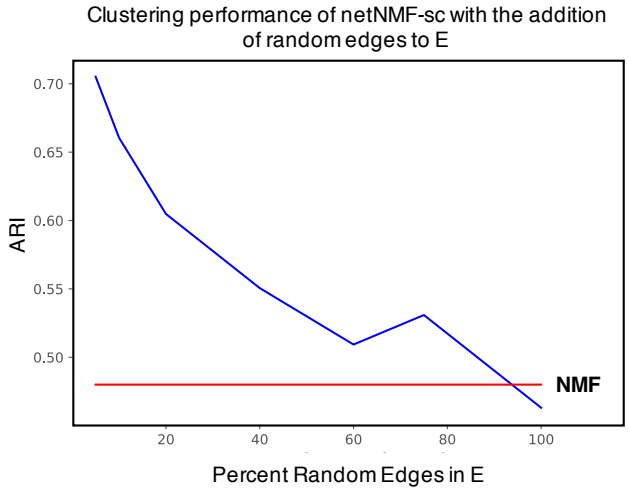


Figure 2.7: Clustering performance of netNMF-sc run on simulated data with 5000 genes, 1000 cells, and 6 clusters. Dropout was simulated using the multinomial dropout model with a dropout rate of 0.7. The x-axis measures the number of random edges added to the original graph $G = (V, E)$, where the number of random edges is $x|E|$. The red line shows the performance of NMF on the same data.

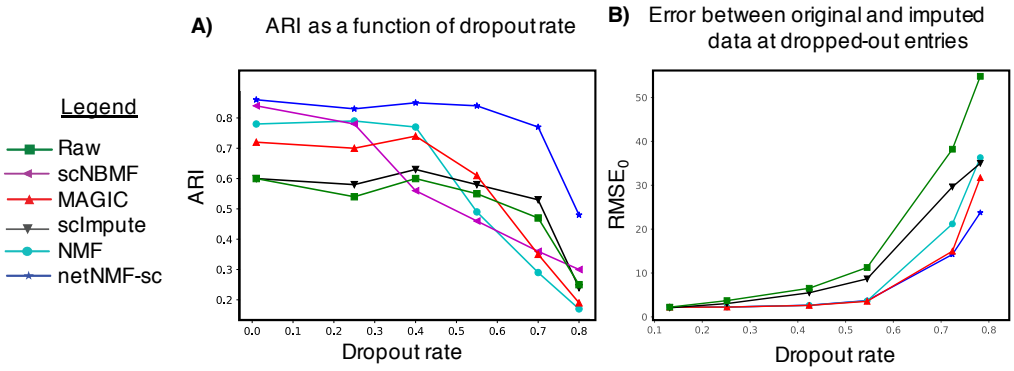


Figure 2.8: Comparison of netNMF-sc and other methods on clustering and imputation for a simulated scRNA-seq dataset containing 1000 cells and 5000 genes, with dropout simulated using a double exponential model. (A) Clustering results for several scRNA-seq methods on simulated data with different dropout rates. (B) Imputation results with different dropout rates.

control, to demonstrate the benefit of netNMF-sc when a highly informative network is available. We note that while S may correlate more strongly with the underlying coexpression structure of the data than we would expect from biological datasets, the edges in S do not perfectly correspond to coexpressed genes in the simulated data. This is due to the fact that only a subset of genes from S are differentially expressed in the simulated data and some pairs of differentially expressed genes in the simulated data are not represented by an edge in S. When we compare the correlation matrix of the simulated data to S, we observe 317 gene pairs with $R^2 \geq 0.5$ are captured by edges in S while 828 gene pairs with $R^2 \geq 0.5$ are not.

We found that the clusters identified using netNMF-sc across all dropout rates had higher overlap with true clusters compared to the clusters identified using other methods (Fig 2A). The improvement for netNMF-sc was especially pronounced at higher dropout rates; for example, at a dropout rate of 0.7, netNMF-sc had adjusted Rand index (ARI) $= 0.78$, compared to 0.47 for the next best performing method, NMF. We observe a similar improvement in clustering performance using the *double exponential dropout model* (Fig 2.8A-B). At a dropout rate of 0.7, netNMF-sc had ARI $= 0.79$, compared to 0.41 for the next best performing method, scImpute (Fig 2.8A).

We compared the performance of netNMF-sc and other methods on the task of imputation by computing the root mean squared error (RMSE) between $\mathbf{X}'$, the simulated transcript count matrix before dropout, and the imputed matrix $\hat{\mathbf{X}} = \mathbf{WH}$. We first compute RMSE_0, the RMSE between $\mathbf{X}'$ and the imputed matrix $\hat{\mathbf{X}}$ restricted to entries where dropout events were simulated. At low dropout rates (< 0.25), netNMF-sc had similar RMSE_0 as other methods, but at higher dropout rates netNMF-sc had lower values of RMSE_0 (Fig

2B). For example, at a dropout rate of 0.7, netNMF-sc had $\text{RMSE}_0 = 4.8$ compared to 7.4 for the next best performing method, NMF (Fig 2B). Similar results were observed on data simulated using the *double exponential dropout model*. At a dropout rate of 0.7, netNMF-sc had $\text{RMSE}_0 = 8.3$, slightly above MAGIC ($\text{RMSE}_0 = 7.9$) but substantially better than NMF ($\text{RMSE}_0 = 15.9$) and scImpute ($\text{RMSE}_0 = 18.3$). When we compute the RMSE between all entries of the transcript count matrix, scImpute outperforms other methods at low dropout rates (< 0.25) because scImpute does not attempt to impute non-zero counts. However, at dropout rates above 0.6, netNMF-sc has the lowest RMSE (Fig 2.5). Additionally, we investigated the contribution of the input network to the performance of netNMF-sc. We found that the addition of up to 70% random edges did not have a large effect on the performance (Fig 2.7).

2.2.3 Evaluation on cell clustering

We compared netNMF-sc and other scRNA-seq methods in their ability to cluster cells into meaningful cell types using three scRNA-seq datasets. For all datasets, we normalized the transcript count matrices following [191] to reduce the effect of differences in the library size, or total number of transcripts sequenced in each cell (Section 2.3.5). We used the normalized count data for all methods except PCA and scImpute. For these two methods we applied a log-transformation ($\log_2(\mathbf{X}+1)$) to the transcript count matrix as these methods assume the data were generated from a Gaussian distribution.

The first dataset contains 182 mouse embryonic stem cells (mESCs) that were flow sorted into one of three cell cycle phases: G, S, and G2/M and sequenced using the Fluidigm C1 platform combined with Illumina sequencing [147]. The data contain 9571 genes and a zero-proportion of 0.41. We computed cell clusters for each method as described in Section 2.3.6. We ran NMF, scNBMF, and netNMF-sc with $d = 5$ dimensions, a value selected via holdout validation. For netNMF-sc, we used a network from the ESCAPE database [187] which contains $153,920$ protein-mRNA regulatory interactions from mESCs, with edge weights of 1 for positive correlations and -1 for negative correlations. We selected $\lambda = 5$ via holdout validation. For PCA we used the top 136 principal components, which explained 90% of the variance. We compared the cell clusters obtained by running each method followed by k-means clustering on the low-dimensional representation, using both the true cluster number $k = 3$ as well as the value k that produced the highest silhouette score in the range $2 \leq k \leq 20$. We also ran Phenograph [167], a graph clustering method, but found that performance was similar or worse than k-means for all methods (Fig 2.9A-D). We found that netNMF-sc outperformed other methods at clustering cells into the three cell cycle stages with an adjusted Rand index (ARI) $= 0.84$ compared to 0.24 for MAGIC and 0.37 for scImpute (Fig 3A-B). Note that while MAGIC did not perform as

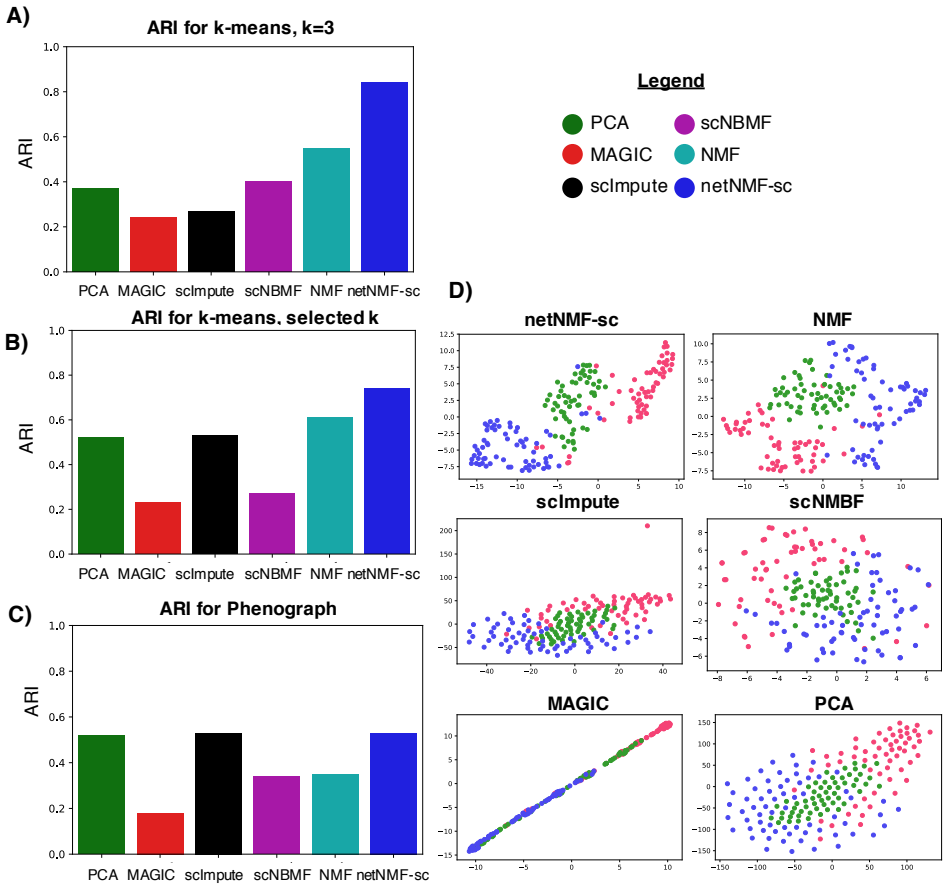


Figure 2.9: Clustering results on the mouse embryonic stem cell (mESC) dataset from [147], which has 3 clusters of cell determined by flow-sorting according to 3 cell cycle stages. (A) k-means clustering results for $k = 3$. (B) k-means clustering results for the value k that produced the highest silhouette score in the range $2 \leq k \leq 20$ for each method. (C) Phenograph clustering results. (D) t-SNE projections of k-means clustering results for $k = 3$.

well as netNMF-sc in clustering the cells into distinct cell cycle phases, it did identify a trajectory between the phases of the cell cycle, which may be biologically meaningful. However, MAGIC also identified a trajectory between clusters in the simulated data above although no trajectory was present (Fig 2.6).

To quantify the contribution of the network to the performance of netNMF-sc, we ran netNMF-sc with three additional networks: a generic gene coexpression network from COEXPEDIA [188], a k-nearest neighbors network (KNN), and a random network with the same degree distribution as the ESCAPE network. The k-nearest neighbors network was constructed by placing an edge between the ten nearest neighbors of each gene in the input data matrix $\mathbf{X}$, based on Euclidean distance (see Section S6 for more details). We found that the ESCAPE coexpression network gave the best performance, with an ARI of 0.84 compared to 0.76 for COEXPEDIA, 0.68 for KNN, 0.63 for the random network, and 0.60 for NMF (Fig 2.17A-B). This result is consistent with the fact that the ESCAPE network was constructed using the same cell type as the scRNA-seq data, mESCs, while the COEXPEDIA network was constructed using cells from many different cell types. This demonstrates the benefit of prior knowledge that is matched to the cell types in the scRNA-seq data. We note that netNMF-sc with any of the networks outperformed NMF, although the difference for the random network was negligible, suggesting that some of the advantage of netNMF-sc may be due to enforcing sparsity on $\mathbf{W}$.

The second dataset, from [190], contains 3005 mouse brain cells from 9 cell types sequenced with the STRT-seq (single-cell tagged reverse transcription) protocol. The data contain $8,345$ genes and a zero-proportion of 0.60. For netNMF-sc we used a gene coexpression network from [173] containing $157,306$ gene-gene correlations across brain cell types (astrocytes, neurons, endothelial cells, microglia, and oligodendrodytes), and selected $\lambda = 1$ via holdout validation. NMF, scNBMF, and netNMF-sc were run with $d = 30$ dimensions, selected via holdout validation. For PCA we used the top 82 principal components, which explained 90% of the variance. For each method we ran k-means with $k = 9$. We found that netNMF-sc outperformed other methods with an ARI $= 0.82$ compared to the next best performing methods NMF and MAGIC with ARIs $= 0.72$ and 0.71 respectively (Fig 3C-D). netNMF-sc also outperformed other methods with k selected using the silhouette score as well as using the clustering method Phenograph, with scNBMF performing second-best (Fig 2.10A-D).

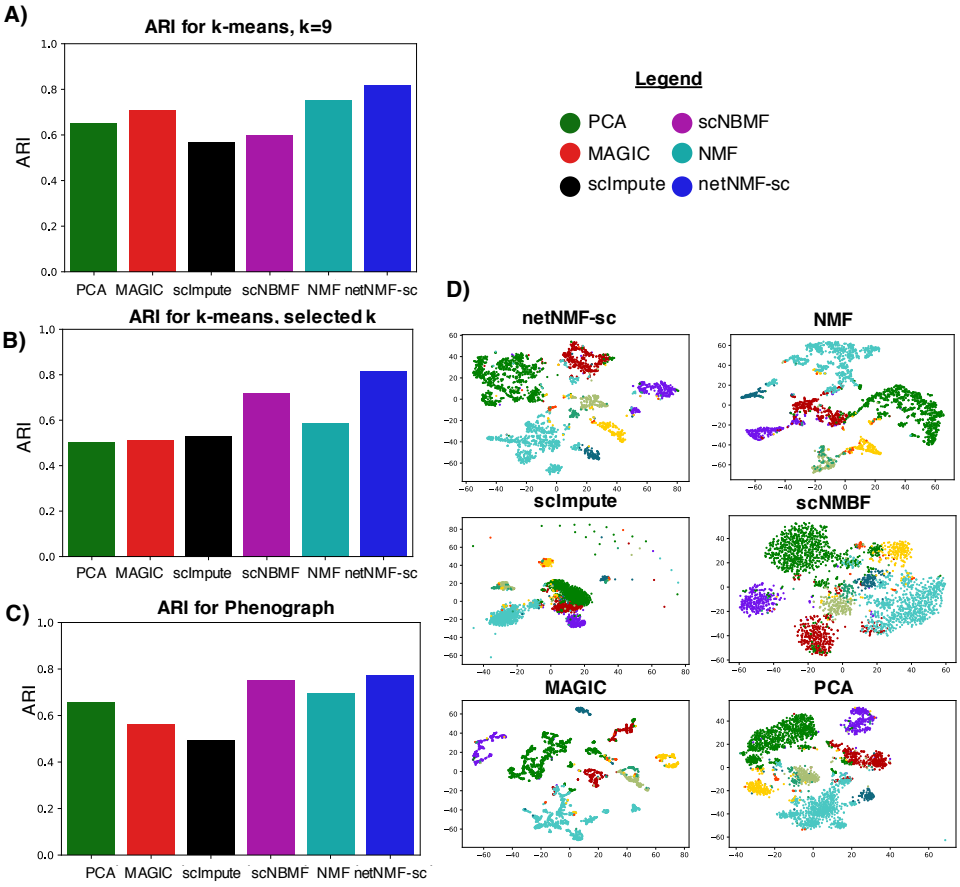


Figure 2.10: Clustering results on brain cell dataset from [190] who identified 9 cell types. (A) k-means clustering results for $k = 9$. (B) k-means clustering results for the value k that produced the highest silhouette score in the range $2 \leq k \leq 20$ for each method. (C) Phenograph clustering results. (D) t-SNE projections of k-means clustering results for $k = 9$.

The third dataset contains 2022 brain cells from an E18 mouse sequenced using 10x Genomics scRNA-seq platform (`https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons_2000`). The data contain $13,509$ genes with transcript counts ≥ 10 and a zero proportion of 0.84. Since this dataset does not have known cell clusters, we compare the cell clusters computed by each method with the 16 brain cell types reported in a separate 10x Genomics scRNA-seq dataset of 1.3 million cells from the forebrains of two different E18 mice that was analyzed using bigSCale [161], a framework for analyzing large-scale transcript count data. For netNMF-sc, we used a gene coexpression network from [173] containing $157,306$ gene-gene correlations across brain cell types (astrocytes, neurons, endothelial cells, microglia, and oligodendrodytes) and selected $\lambda = 50$ via holdout validation. NMF and netNMF-sc were run with $d = 20$ dimensions, selected via holdout validation. For PCA we used the top 372 principal components, which explained 90% of the variance. We used $k = 16$ in k-means clustering to match the number of brain cell types reported in bigSCale. We matched the cell clusters output by each method to the 16 cell types reported in bigSCale as follows. We computed the overlap between the top 200 over-expressed genes in each cluster (calculated with a one-sided t-test between cells in and out of the cluster) and the published marker genes for each of the 16 cell types, and selected the cell type with the lowest p-value of overlap (Fisher's exact test). If the cluster was not enriched for any cell type with Bonferroni-corrected $p < 0.05$ then we marked the cluster as unclassified.

While the true class assignment for each cell is unknown, both scRNA-seq datasets were generated from the forebrains of E18 mice, and thus we expect that the proportions of each cell type should be similar across both datasets. We found that the proportions of each cell type identified by netNMF-sc (Fig 4E) were the closest (many within 2%) to the proportions reported by [161] (Fig 4F). In both cases, the cell type with the largest proportion is glutamatergic neurons, followed by interneurons and then radial glia and post-mitotic neuroblasts. Other cell types, such as dividing GABAergic progenitors and Cajal–Retzius neurons, were found in smaller proportions. In contrast, MAGIC (Fig 4B) finds a large population (13%) of Cajal–Retzius neurons, while scImpute (Fig 4C) finds a large population (18%) of dividing GABAergic progenitor cells – both proportions more than 3-fold greater than in bigSCale or netNMF-sc. Clusters computed from PCA (Fig 4A) and from NMF (Fig 4D) also differed substantially from the proportions reported in bigSCale (Fig 4F); for example, the proportion of post-mitotic neuroblasts was 0% in PCA, 20% in NMF, but 10% in bigSCale. We found that the number of unclassified cells varied substantially across the methods. Clusters computed from scImpute and netNMF-sc had no unclassified cells while PCA, MAGIC, and NMF had 10%, 25%, and 1% of cells unclassified, respectively.

We further examined the smallest cell cluster identified by netNMF-sc, containing only 14 cells. This cluster was enriched ($p \leq 2.2 \times 10^{-16}$) for microglia marker genes reported by bigSCale, including well-studied marker genes such as *Csf1r*, *Olfml3*, and *P2ry12* [185]. These 14 cells represented 0.7% of the 2022 sequenced cells, closely matching the proportion of microglia reported by bigScale (1%). NMF and MAGIC also identified clusters of microglia cells, but the differentially expressed genes in these clusters were less enriched for microglia marker genes ($p \leq 4.1 \times 10^{-13}$ and $p \leq 5.5 \times 10^{-3}$, respectively). The NMF cluster contained 65 cells but did not include any of the 14 cells classified as microglia by netNMF-sc. In addition, these 65 cells were equally enriched for erythrocyte marker genes ($p \leq 3.2 \times 10^{-11}$). The MAGIC cluster contained 174 cells, a much larger proportion (9%) of the cell population than the 1% reported by bigSCale. This cluster included the 14 microglia identified by netNMF-sc but also 160 other cells. The additional 160 cells present in the cluster were not enriched for microglia marker genes ($p \leq 1.2 \times 10^{-1}$) but were enriched for glutamatergic marker genes ($p \leq 1.5 \times 10^{-2}$). This suggests that MAGIC erroneously grouped together different types of cells.

We found 436 genes were differentially expressed between the 14 microglia identified by netNMF-sc and the other 2008 cells (FDR ≤ 0.01). All 50 microglia marker genes from bigScale were included in this set, including the two most highly differentially expressed genes *Cc14* (fold change 12.5) and *C1qc* (fold change 8.7). Of the top 20 differentially expressed genes identified in the netNMF-sc microglial cells, several were reported in other studies as microglia genes [178] but not bigScale; these include *Hexb* (fold change 7.8) and *Lgmn* (fold change 5.8). Several potential novel marker genes were in the 20 differentially expressed genes, including *Cstdc5* (fold change 4.5) and *Stfa1* (fold change 4.2). These results suggest that netNMF-sc improves clustering of cells into biologically meaningful cell types from scRNA-seq data with high dropout – even when the cell type is represented by only a small number (< 20) of cells – and facilitates the identification of potentially novel marker genes.

2.2.4 Recovering marker genes and gene-gene correlations from cell cycle data

Finally, we investigated how well each method recovers differentially expressed marker genes and gene-gene correlations from scRNA-seq data. First, we examined cell cycle marker genes. We obtained a set of 67 *periodic marker genes* whose expression has been shown to vary over the cell cycle across multiple cell types [150]. This set contains 16 genes with peak expression in G1/S phase and 51 genes with peak expression during G2/M phase. We expect to observe a significant number of these periodic genes amongst the top differentially expressed genes between G1/S phase and G2/M phase cells in the cell cycle dataset

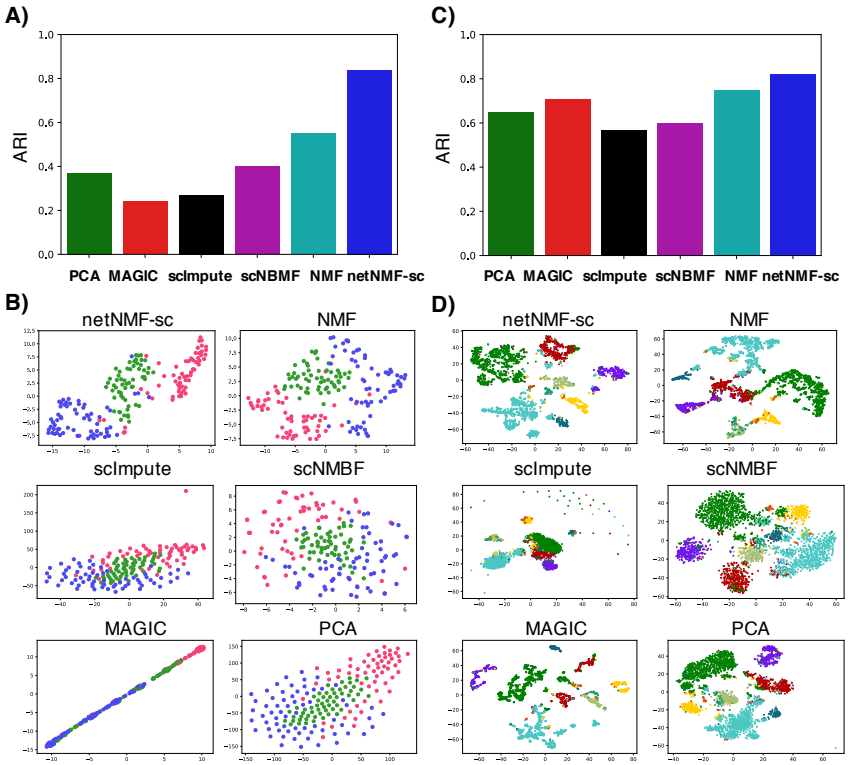
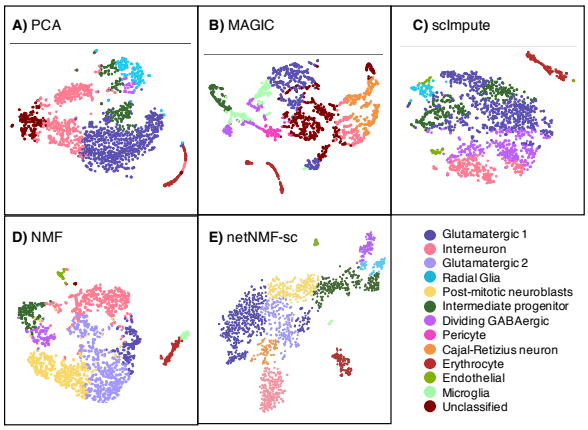


Figure 2.11: (A) Adjusted Rand index (ARI) for cell clusters obtained by methods on mouse embryonic stem cell (mESC) scRNA-seq data from [147], with cell cycle labels obtained by flow sorting. (B) 2D t-SNE projections of cells in reduced dimensional space. (C) Clustering results on brain cell dataset from [190] into 9 cell types. (D) 2D t-SNE projections of cells in reduced dimensional space.

from [147]. We compared the ranked list of differentially expressed genes from data imputed by netNMF-sc to the ranked lists of differentially expressed genes from the untransformed data and data imputed NMF, MAGIC, scImpute. We found that periodic genes ranked very highly in netNMF-sc results ($p \leq 3.2 \times 10^{-11}$, Wilcoxon rank sum), an improvement compared to their ranking in the untransformed data ($p \leq 4.5 \times 10^{-3}$, Wilcoxon rank sum, Fig 5A). In contrast, the data imputed with NMF, MAGIC, and scImpute resulted in a lower ranking of the periodic genes ($p \geq 0.05$, Wilcoxon rank sum). Additionally, we found that in data imputed by MAGIC, some periodic genes had expression patterns that were out of phase with the cell cycle. For example, *Exo1*, which peaks in G1/S phase, had lower expression in G1/S phase cells compared to G2/M phase cells ($p \leq 2.2 \times 10^{-16}$, Wilcoxon rank sum) in MAGIC imputed data (Fig 5B). In contrast, the peak in *Exo1* expression during G1/S phase is observed in the results from netNMF-sc ($p \leq 6.7 \times 10^{-12}$, Wilcoxon rank sum), while *Exo1* is not differentially expressed in the untransformed data ($p \leq 0.17$, Wilcoxon rank sum) (Fig 5B).



F) Cell type proportions

	PCA	MAGIC	scImpute	NMF	netNMF-sc	bigSCale
Glutamatergic 1	39%	21%	39%	9%	23%	**25%**
Interneuron	26%	8%	15%	22%	17%	**19%**
Glutamatergic 2	0%	0%	0%	27%	15%	**15%**
Radial Glia	8%	0%	4%	0%	6%	**11%**
Post-mitotic neuroblasts	0%	0%	0%	20%	11%	**10%**
Intermediate progenitor	10%	6%	15%	7%	10%	**9%**
Dividing GABAergic	2%	7%	18%	6%	5%	**4%**
Pericyte	0%	4%	0%	0%	0%	**2%**
Cajal-Retizius neuron	0%	13%	0%	0%	4%	**2%**
Erythrocyte	5%	7%	7%	5%	6%	**1%**
Endothelial	0%	0%	2%	1%	2%	**1%**
Microglia	0%	9%	0%	2%	1%	**1%**
Unclassified	10%	25%	0%	1%	0%	**0%**

Figure 2.12: (A-E) t-SNE projections of scRNA-seq data from 2022 brain cells from an E18 mouse. Colors indicate cell types as derived in bigSCale analysis of 1.3 million E18 mouse brain cells [161]. (F) Proportions of each cell type predicted by each method. Entries highlighted in *blue* are within 2% of the proportions from bigSCale. Entries highlighted in *orange* differ by more than 10% from the proportions from bigSCale.

We also investigated whether each method could recover gene-gene correlations between periodic marker genes in the cell cycle data. We expect pairs of periodic genes whose expression peaks during the same phase of the cell cycle to be positively correlated and pairs of genes that peak during different phases to be negatively correlated. Across all 2211 pairs of periodic marker genes, we found that the mean R^2 was 0.54 for netNMF-sc, compared to 0.73 for MAGIC, 0.29 for NMF, 0.02 for scImpute and 0.03 for untransformed data (Fig 5C). Setting a stringent cutoff for significant correlation ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$), we found

that 15% of the pairs of periodic genes were correlated in data imputed by netNMF-sc compared to 68% in data imputed by MAGIC, 0.8% in data imputed by NMF, and nearly 0% in data imputed by scImpute. While the higher percentage of correlated gene pairs in MAGIC seems to be an advantage, the MAGIC-imputed data also contained a number of cell cycle marker genes, such as *Exo1*, whose expression signature was the *opposite* of expected. Such cases can result in incorrect correlations between pairs of marker genes. For example, marker genes *Exo1* and *Dtl* both peak during G1/S phase and are expected to be positively correlated. However, MAGIC found negative correlation ($R = -0.58, p \leq 3.6 \times 10^{-16}$) between these two genes. In contrast, netNMF-sc recovers the positive correlation ($R = 0.56, p \leq 2.2 \times 10^{-16}$), while scImpute ($R = 0.03, p \leq 0.66$) and NMF ($R = 0.06, p \leq 0.46$) do not (Fig 5D).

Overall, we found that in the data imputed by MAGIC 19% of correlated periodic genes were correlated in the *opposite* direction than expected; i.e., genes that peaked during the same phase were negatively correlated or genes which peaked during different phases were positively correlated. In contrast, in the data imputed by netNMF-sc only 1% of the correlated periodic genes were correlated in the opposite direction than expected (Table 1). These results from MAGIC may be explained by the fact that MAGIC introduces a large number gene-gene correlations during imputation, many of which may be spurious, as was previously reported by [160]. In fact, the majority (78%) of the gene pairs in the correlation matrix generated from data imputed by MAGIC were correlated ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$), compared to only 0.2% in the correlation matrix generated from data imputed by netNMF-sc and 0.005% in the correlation matrix generated from the untransformed data (Table 1).

To examine whether these correlations identified by MAGIC and netNMF-sc represented real biological signal, we ran both methods on permuted data where the transcript counts were permuted independently in each cell. We found that 85% of the gene pairs were correlated ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$) in MAGIC imputed data compared to only 0.2% of gene pairs in netNMF-sc imputed data (Table 1). This observation suggests that many of the gene-gene correlations found in the MAGIC imputed cell cycle data may be spurious. Further investigation on simulated data suggests that such spurious correlations may be a consequence of the small number of cells: we found that MAGIC imputed data had many correlations in transcript count matrices with ~ 200 cells but fewer correlations in imputed data with many (~ 1000) cells (Fig 2.15). We also observed the number of gene-gene correlations found by MAGIC on permuted data increased rapidly with the diffusion parameter t before reaching a plateau (Fig 2.14B). In contrast, the number of gene-gene correlations found by netNMF-sc on permuted data decreased as the number of latent dimensions d increased (Fig 2.14A).

Method	Fraction of gene pairs with correlation ($R^2 \geq .8$)	Fraction of periodic gene pairs with correlation ($R^2 \geq .8$) in correct/incorrect orientation
Untransformed	$1x10^{-5}$	0.00 / 0.00
MAGIC	0.78	0.49 / 0.19
scImpute	$1x10^{-5}$	0.00 / 0.00
NMF	$2x10^{-3}$	$8x10^{-3}$ / $1x10^{-3}$
netNMF-sc	$2x10^{-3}$	0.14 / 0.01
Bulk (COXPRESdb)	$9x10^{-5}$	0.14 / 0.00
Permuted data	$1x10^{-4}$	$2x10^{-2}$ / $1x10^{-2}$
MAGIC on permuted data	0.85	0.40 / 0.39
netNMF-sc on permuted data	$2x10^{-3}$	$2x10^{-3}$ / $3x10^{-4}$

Table 2.1: Fraction of all pairs of genes and pairs of periodic genes (defined by [150]) with correlations ($R^2 \geq 0.8$, $p \leq 2.2 \times 10^{-16}$, Student's t-test) in the cell cycle dataset [147]. *Correct* orientation means that a pair of genes with peak expression in the same stage of the cell cycle has positive correlation, and a pair of genes with peak expression in different stages of the cell cycle has negative correlation. Grey rows denote correlations on permuted data.

We performed a second analysis of differentially expressed marker genes and gene-gene correlations in scRNA-seq data from the MAGIC publication [182] containing 7415 human transformed mammary epithelial cells (HMLEs) which were induced to undergo epithelial to mesenchymal transition (EMT) and then sequenced using the inDrops platform [163]. We assessed how well each method recovered differential expression of 16 canonical EMT marker genes from [154] (3 genes with high expression in epithelial (E) cells and 13 genes with high expression in mesenchymal (M) cells). We found that the EMT marker genes ranked highly in netNMF-sc results ($p \leq 1.4 \times 10^{-5}$, Wilcoxon rank sum), an improvement compared to their ranking in the untransformed data ($p \leq 3.1 \times 10^{-3}$, Wilcoxon rank sum, Fig 2.18A). MAGIC was the second best method, ranking EMT genes highly ($p \leq 1.1 \times 10^{-4}$, Wilcoxon rank sum) but below the performance of netNMF-sc. We observed that in data imputed by MAGIC, the E marker gene *TJP1* had higher average expression in M cells than E cells ($p \leq 2.2 \times 10^{-16}$) (Fig 2.18B). This resulted in *TJP1* being negatively correlated ($R = -0.57, p \leq 2.2 \times 10^{-16}$) with another epithelial marker gene, *CDH1* in the MAGIC imputed data. In contrast, these E marker genes showed positive correlation ($R = 0.66, p \leq 6.4 \times 10^{-16}$) in the netNMF-sc

imputed data and this correlation that was not apparent in the untransformed data (Fig 2.18C). We also investigated whether netNMF-sc could recover gene-gene correlations between EMT marker genes in E and M cells. We expect that pairs of E or M genes would exhibit positive correlation, while pairs containing one E and one M gene would exhibit negative correlations. In data imputed by netNMF-sc, 12% of the EMT gene pairs were correlated ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$), with all gene pairs correlated in the expected orientation (Fig 2.19). In data imputed by MAGIC, 23% of EMT gene pairs were correlated, but 5% were correlated in the *opposite* direction than expected (Fig 2.18D).

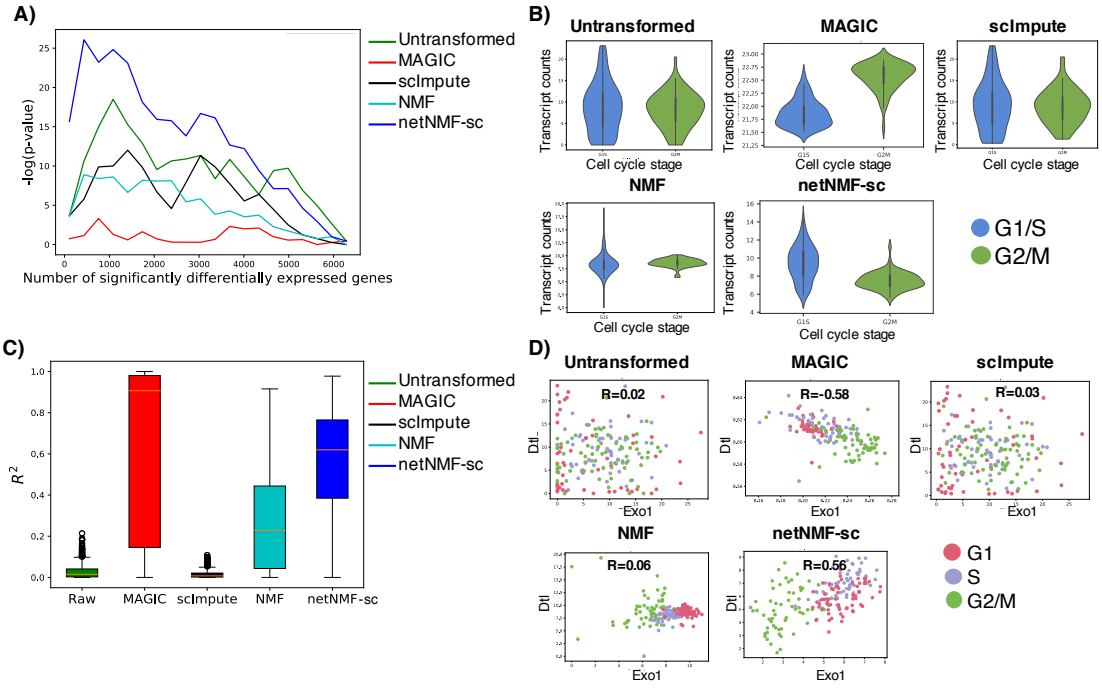


Figure 2.13: Comparison of differential expression of marker genes and gene-gene correlations in untransformed data from [147] and data imputed using netNMF-sc, NMF, scImpute, and MAGIC. (A) Overlap between differentially expressed genes and periodic genes (log p-values from Fisher's exact test). (B) Expression of the G1/S phase marker gene *Exo1* in cells labeled as G1/S (blue) and cells labeled as G2/M (green) in data imputed by each method. In netNMF-sc imputed data, *Exo1* is overexpressed in G1/S cells compared to G2/M cells ($p \leq 6.7 \times 10^{-12}$), as expected. In contrast, in data imputed by MAGIC, *Exo1* is *underexpressed* in G1/S cells compared to G2/M cells ($p \leq 2.2 \times 10^{-16}$). Exo1 shows no difference in expression in untransformed and scImpute data. (C) Distribution of R^2 correlation coefficients between pairs of periodic genes in the cell cycle data. (D) Scatter plot of expression of two G1/S phase genes, *Dtl* and *Exo1*, across cells. These genes are positively correlated in data imputed by netNMF-sc ($p \leq 2.2 \times 10^{-16}$), negatively correlated in data imputed by MAGIC ($p \leq 2.2 \times 10^{-16}$), and uncorrelated in other methods.

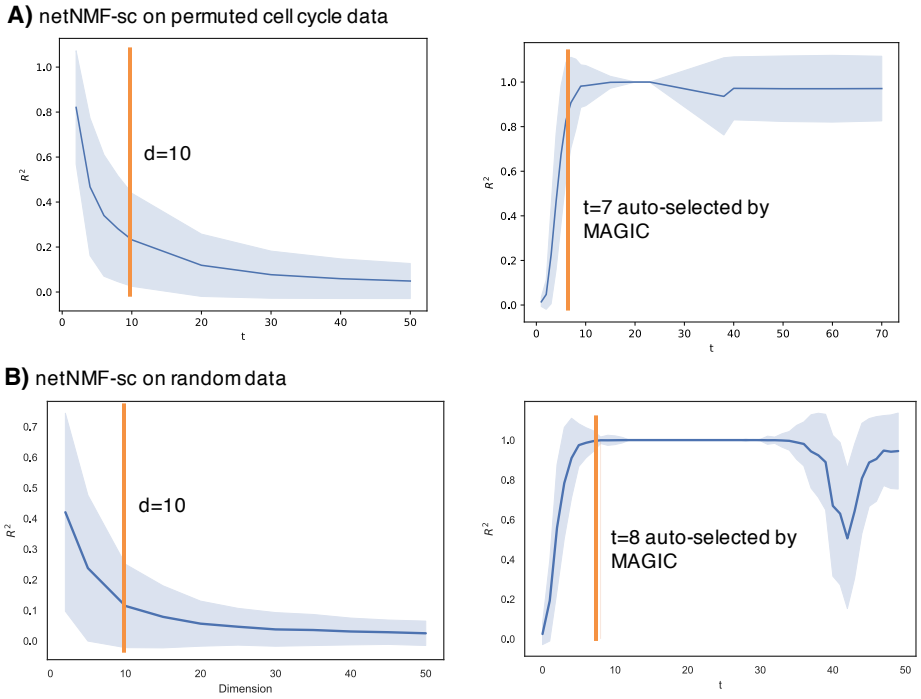


Figure 2.14: (A) Average R^2 correlation over gene pairs on permuted cell cycle data as a function of the number d of dimensions in the matrix factorization from netNMF-sc. (B) Average R^2 correlation over gene pairs on permuted cell cycle data as a function of the diffusion operator, t, used by MAGIC (light blue indicates standard deviation). $t = 5$ is auto-selected by MAGIC according to the Procrustes disparity of the diffused data. (C) netNMF-sc run on random data drawn from $N(2,2)$. (D) MAGIC run on random data drawn from $N(2,2)$.

Random expression matrices drawn from N(2,2) and imputed using MAGIC

Size (genes,cells)	Mean R^2	Percent significant correlations ($R^2 > 0.8$)	Auto-selected t
(10000,100)	0.997	0.997	5
(10000,200)	0.997	0.96	5
(10000,300)	0.73	0.60	21
(10000,400)	0.13	$5\text{x}10^{-3}$	20
(10000,500)	0.16	$7\text{x}10^{-3}$	21
(10000,1000)	0.08	$1\text{x}10^{-3}$	19
(10000,2000)	0.07	$1\text{x}10^{-3}$	20

Figure 2.15: Gene-gene correlations introduced by MAGIC on expression matrices simulated from a $N(2,2)$ distribution.

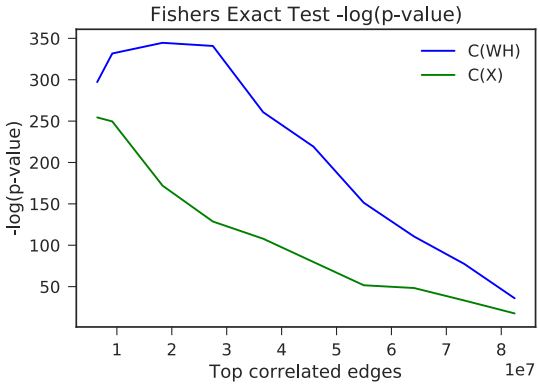


Figure 2.16: Enrichment of the imputed count matrix **WH** (blue) and the raw count matrix **X** for edges in the input network (ESCAPE).

2.2.5 Clustering on cell cycle data

To quantify the effect our choice of network has on the performance of netNMF-sc, we ran netNMF-sc with two different external networks as well as a network containing randomized edges. The first network is the previously described network obtained from the ESCAPE database [187]. The second network is a generic gene-gene co-expression network which is the result of combining expression data from $2,486$ mouse microarray experiments [188]. Next, we constructed a k-nearest neighbors network, constructed by representing the 10 nearest neighbors of each gene in the input data matrix as edges with weight 1 in the network. Finally, we constructed a randomized network that maintains the same node degree as the ESCAPE network by performing the double_edge_swap procedure from the python library networkx.

We found that all networks besides the random network significantly improved clustering results compared to NMF (Fig 2.17A-B), with the mESC-specific network obtained from the ESCAPE database performing the best.

2.2.6 Recovering marker genes and gene-gene correlations from EMT data

Using a set of 16 canonical EMT marker genes (3 genes overexpressed in epithelial cells and 13 genes overexpressed in mesenchymal cells) [154], we defined the set of all 120 gene pairs as our gold standard. We note that this set includes several gene pairs not investigated in the MAGIC paper [**?**]. To validate our approach, we looked for positive correlations between pairs of mesenchymal or epithelial genes and negative correlations between pairs containing one epithelial and one mesenchymal gene.

We clustered cells by *CDH1* and *VIM* expression, two canonical marker genes for epithelial (*CDH1*)

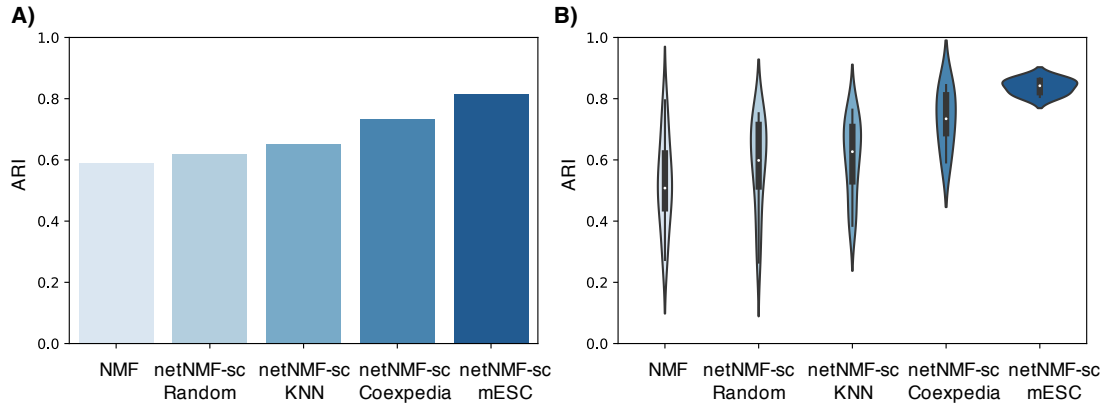


Figure 2.17: (A) Clustering results for cell cycle data from [147]. The result that minimizes the netNMF-sc objective value across 10 random initializations is displayed. NMF is compared with netNMF-sc run with different networks used as input. Coexpedia is a generic gene-gene co-expression network, ESCAPE is a gene-gene co-expression network specific to mESCs, and KNN is a k-nearest neighbors network constructed from the 10 nearest neighbors of each gene in the input data matrix. Random is a random network constructed to have the same number of edges and degree as the ESCAPE network. (B) Variance in clustering performance across 10 random initializations.

and mesenchymal cells (*VIM*), respectively. We labeled the 200 cells with the highest *CDH1* expression epithelial and the 200 cells with the highest *VIM* expression mesenchymal. We compared the ranked list of differentially expressed genes from data imputed by netNMF-sc to the ranked lists of differentially expressed genes from the raw data and data imputed NMF, MAGIC, scImpute. We found that the EMT marker genes ranked very highly in netNMF-sc results ($p \leq 1.4 \times 10^{-5}$, Wilcoxon rank sum), a significant improvement compared to their ranking in the raw data ($p \leq 3.1 \times 10^{-3}$, Wilcoxon rank sum) (Fig 2.18(a)). In contrast, the next best performing method MAGIC had a smaller improvement in the ranking of EMT marker genes compared to the raw data ($p \leq 1.1 \times 10^{-4}$, Wilcoxon rank sum).

We observed that in data imputed by MAGIC, the E marker gene *TJP1* had higher average expression in M cells than E cells ($p = 1.5 \times 10^{-33}$) (Fig 2.18(b)). This resulted in *TJP1* being negatively correlated ($R = -0.57, p = 3.4 \times 10^{-50}$) with another E marker gene, *CDH1* in the MAGIC imputed data; in contrast, these E marker genes showed positive correlation ($R = 0.66, p = 6.4 \times 10^{-78}$) in the netNMF-sc imputed data, correlation that was not apparent in the raw data (Fig 2.18(c)). We also investigated whether netNMF-sc could recover gene-gene correlations between EMT marker genes in E and M cells. We expect that pairs of E or M genes would exhibit positive correlation, while pairs containing one E and one M gene would exhibit negative correlations. In data imputed by netNMF-sc, 12% of the EMT gene pairs were significantly correlated ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$), with all gene pairs correlated in the expected orientation (Fig 2.19).

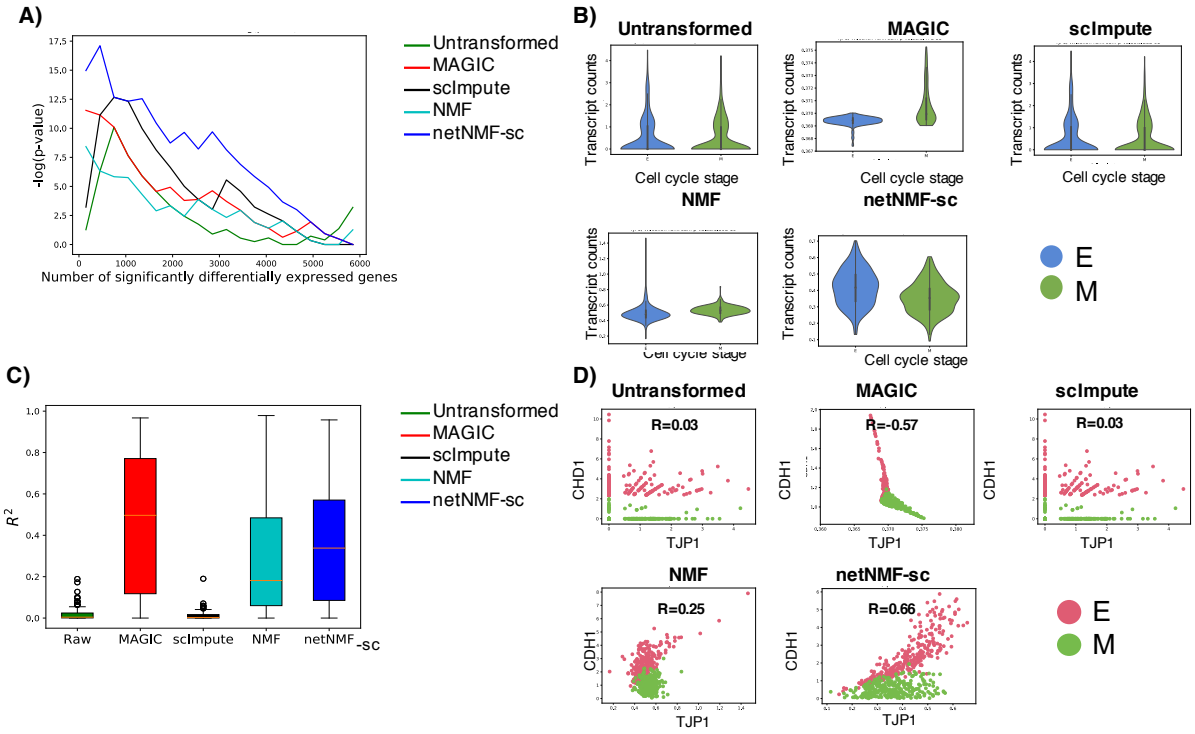


Figure 2.18: Comparison of gene-gene correlations and differential gene expression in raw data from [182] and data imputed using netNMF-sc, NMF, scImpute, and MAGIC. (A) Overlap between differentially expressed genes and EMT marker genes (log p-values from Fisher's exact test). (B) Expression of the E marker gene *TJP1* in cells labeled as E (blue) and cells labeled as M (green) in data imputed by each method. In netNMF-sc imputed data, *TJP1* is overexpressed in E cells compared to M cells ($p = 1.4 \times 10^{-12}$), as expected. In contrast, in data imputed by MAGIC, *TJP1* is *underexpressed* in E cells compared to M cells ($p = 1.5 \times 10^{-33}$), and shows no significant difference in expression in raw and scImpute data. (C) Correlation between pairs of periodic genes in cell cycle data. (D) Scatter plot of two E phase genes: *CDH1* and *TJP1*. The genes are positively correlated in data imputed by netNMF-sc ($p = 6.3 \times 10^{-78}$) but negatively correlated in data imputed by MAGIC ($p = 3.4 \times 10^{-50}$).

In data imputed by MAGIC, 23% of EMT gene pairs were significantly correlated, but 5% were correlated in the *opposite* direction than expected (Fig 2.18(d)).

Method	Gene pairs with significant ($R^2 \geq .8$) correlation	Periodic gene pairs with significant ($R^2 \geq .8$) correlation in correct/incorrect orientation
Raw	6e-8	0.00 / 0.00
MAGIC	0.05	0.18 / 0.05
scImpute	6e-8	0.00 / 0.00
NMF	0.02	0.06 / 0.00
netNMF-sc	6e-3	0.12 / 0.00

Figure 2.19: Fraction of all gene pairs and EMT gene pairs (defined by [154]) with significant correlations ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$) in the EMT dataset. *Correct* orientation means that a pair of E-E or M-M genes have positive correlation while E-M genes have negative correlation.

2.3 Methods

2.3.1 netNMF-sc algorithm

netNMF-sc uses graph-regularized NMF [148] with KL divergence, which solves the following optimization problem:

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{\mathbf{WH}|_{ij}} - x_{ij} + \mathbf{WH}|_{ij} \right) + \lambda Tr(\mathbf{W}^T \mathbf{L} \mathbf{W}), \tag{2.3}$$

for a positive real constant λ, where $\mathbf{L}$ is the Laplacian matrix of the gene-gene interaction network, and $Tr(\cdot)$ indicates the trace of the matrix. The regularization term $Tr(\mathbf{W}^T \mathbf{L} \mathbf{W})$ encourages pairs of genes to have similar representations in the matrix $\mathbf{W}$ when they are connected in the network. Graph-regularized NMF has previously been used in bioinformatics to analyze somatic mutations in cancer [159].

We derive the graph Laplacian $\mathbf{L}$ for the gene-gene interaction network as follows. Let $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{m \times m}$ denote a gene-gene similarity matrix whose entry s_{ij} is the weight of an interaction between genes i and j. A positive weight s_{ij} indicates a positive correlation between gene i and gene j, while a negative weight indicates negative correlation. We use the signed graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D} = \text{Diag}(|\mathbf{S}|1)$ is the degree matrix and $|\mathbf{S}|$ is the entry-wise absolute value of $\mathbf{S}$. The signed Laplacian, like the Laplacian, is symmetric and positive semidefinite, [156, 164]. Performing Laplacian embedding using the signed version of the graph Laplacian produces an embedding where positive edges between a pair of genes correspond to high similarity and negative edges correspond to low similarity [164].

We implemented netNMF-sc using the TensorFlow Python library [155] and tested the performance of netNMF-sc with four different optimizers: Adam, momentum, gradient descent, and Adagrad. We found

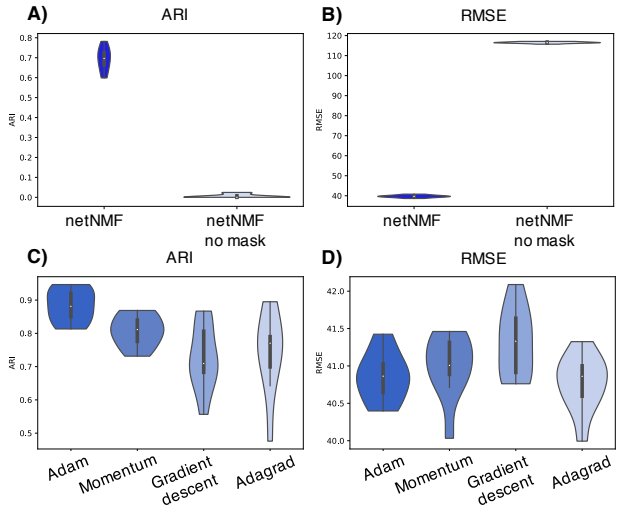


Figure 2.20: (A-B) Adjusted rand index (ARI) and Root mean square error (RMSE) of netNMF-sc with Euclidean distance on simulated data with and without masking of zero entries. (C-D) Clustering performance (ARI) and imputation error (RMSE) of netNMF-sc with Euclidean distance using different optimizers (Adam, Momentum, Gradient descent, and Adagrad).

Adam to perform the best at recovering clusters embedded in the data as well as reducing error between the imputed data and the original data prior to dropout (Fig 2.20A-D). Adam (Adaptive Moment Estimation) uses the first and second moments of gradient of the cost function to adapt the learning rate for each parameter [162]. This allows Adam to perform well on noisy data as well as sparse matrices [162].

netNMF-sc has a shorter runtime on large-scale scRNA-seq datasets than other methods. On a simulated dataset with 5000 genes and 2000 cells, netNMF-sc ran in 1.2 minutes on one 2.60GHz Intel Xeon CPU and in 34 seconds on one NVidia Tesla P100 GPU. In comparison, MAGIC was the fastest method, taking only 13 seconds, while scNBMF and scImpute were both significantly slower than netNMF-sc, taking 2.1 and 6.9 minutes respectively (Fig 2.21). On a real dataset from [172] with 9291 genes and $44,808$ mouse retinal cells, netNMF-sc ran in 34 minutes on one NVidia Tesla P100 GPU. In comparison, MAGIC was the fastest, running in 1.3 minutes, while scNBMF and scImpute were significantly slower than netNMF-sc, failing to complete in 5 hours.

2.3.2 netNMF-sc with Euclidean distance

We also formulated netNMF-sc with a Euclidean distance cost function. This cost function is equivalent to maximizing the Gaussian likelihood the data $\mathbf{X}$ given its factors $\mathbf{W}$ and $\mathbf{H}$ [152]. Graph-regularized NMF

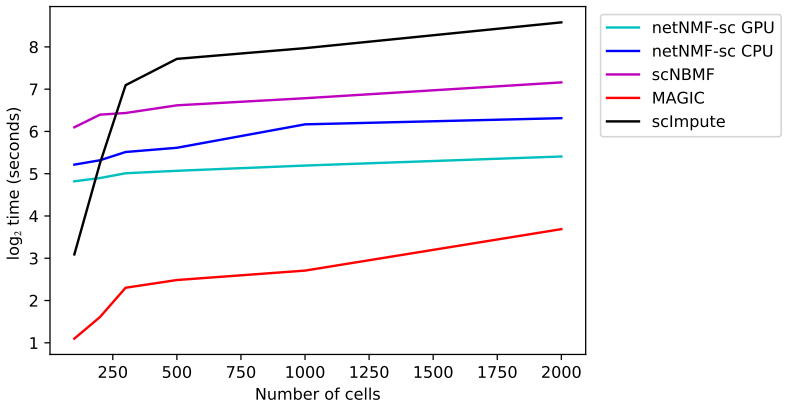


Figure 2.21: Runtime ($\log^2$)of imputation methods as a function of the number of cells (with 5000 genes).

[148] is the following:

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} ||\mathbf{X}-\mathbf{WH}||_F^2 + \lambda Tr(\mathbf{W}^T\mathbf{LW}), \tag{2.4}$$

where λ is a positive real constant, $\mathbf{L}$ is the Laplacian matrix of the gene-gene interaction network, and $Tr(\cdot)$ indicates the trace of the matrix. We allow for zero inflation using a binary matrix $\mathbf{M}$ that masks zero entries in $\mathbf{X}$, such that a non-zero entry in a_{ij} in $\mathbf{WH}$ is not penalized when the corresponding entry x_{ij} of $\mathbf{X}$ is equal to 0. $\mathbf{M}$ has the same dimensions as $\mathbf{X}$ with entries

$$m_{ij} = \begin{cases} 1 & \text{if } x_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Incorporating the mask, the final formulation of netNMF-sc with a Euclidean distance cost function is

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} ||\mathbf{M}\circ\mathbf{X}-\mathbf{M}\circ\mathbf{WH}||_F^2 + \lambda Tr(\mathbf{W}^T\mathbf{LW}), \tag{2.6}$$

where $\circ$ indicates element-wise multiplication (or Schur product of matrices).

To meet the Gaussian assumptions of this model, we set $\mathbf{X}$ to be the log-transform of the transcript counts with a pseudocount of 1, as in many scRNA-seq models which assume an underlying Gaussian distribution [168, 176]. The zero entry mask is not implemented in many commonly used NMF methods [**? ?**], but has a profound effect on improving clustering performance and imputation accuracy at high dropout rates (Fig 2.20(a-b))

2.3.3 Generation of simulated scRNA-seq data

We used the simulator SPLATTER [189] to generate transcript count data, estimating the parameters of the model from mouse embryonic stem cell scRNA-seq data [147] using the SplatEstimate command. We modified SPLATTER to introduce correlations between genes that are differentially expressed in each cluster using a gene coexpression network from [188]. See Section 4.4 for further details.

After simulating transcript counts to obtain a count matrix $\mathbf{X}'$, we generated dropout events using one of two models. The first is a *multinomial dropout model*, used previously to model dropout in scRNA-seq data [171, 192]. In this model, the observed transcript counts in a cell are multinomial distributed, where the probability of observing a transcript from gene i in cell j is $\frac{x'_{ij}}{\sum_{r,s} x'_{rs}}$ and the number of trials is the sum of all transcripts in the count matrix, $\sum_{r,s} x'_{rs}$, multiplied by the capture efficiency, ranging from 0 to 1. The resulting count matrix $\mathbf{X}$ contains dropout proportional to the capture efficiency. The second model is the *double exponential dropout model*, used previously in the scImpute [168] and BISCUIT [146] publications. In this model, an entry x_{ij} of the count matrix is set to zero with probability $p = \exp(-\delta x'^2_{ij})$, where δ is the dropout rate.

2.3.4 Parameter selection via holdout validation

We use the following holdout validation procedure to select the number of latent dimensions d and the regularization parameter λ.

1. Select 20% of the entries of $\mathbf{X}$ to be held-out at random. Let $\mathscr{V}$ denote the indices of these data in $\mathbf{X}$.

2. Run netNMF-sc for a range of latent dimensions d with $\lambda = 0$, masking out held-out entries using the matrix $\mathbf{M}$

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{(\mathbf{M} \circ \mathbf{WH}|_{ij})} - x_{ij} + (\mathbf{M} \circ \mathbf{WH}|_{ij})|_{ij} \right) + \lambda Tr(\mathbf{W}^T \mathbf{LW}), \tag{2.7}$$

   where $\mathbf{M}$ contains zeros for $m_{ij} \in \mathscr{V}$ and ones for $m_{ij} \notin \mathscr{V}$. We hold out random entries rather than rows or columns to prevent overfitting as proposed by [**?**].

3. Calculate root mean squared error (RMSE) $= \sqrt{\frac{\sum_{(i,j)\in\mathscr{V}} (\mathbf{WH}_{ij} - \mathbf{X}_{ij})^2}{|\mathscr{V}|}}$ between the held-out data from $\mathbf{X}$ and the reconstructed data $\mathbf{WH}$, where $|\mathscr{V}|$ denotes the number of held-out entries

4. Select the value of d which results in the lowest RMSE

We perform the analogous procedure to select the regularization parameter λ using the value of d selected in the previous step.

2.3.5 Library size normalization

For a transcript count matrix $\mathbf{X}$, the library size l_j of each cell j is the sum of all transcript counts across every gene,

$$l_j = \sum_{i \in n} x_{ij}.$$

To normalize $\mathbf{X}$, we divide each entry x_{ij} in a cell's expression profile by the cell's library size and then multiply x_{ij} by the median library size q across all cells,

$$\bar{x}_{ij} = q \frac{x_{ij}}{l_j},$$

where $\bar{x}_{ij}$ is an entry in the normalized transcript count matrix $\bar{\mathbf{X}}$.

2.3.6 Clustering low-dimensional cell matrices

To compare the results of the dimensionality reduction and imputation methods PCA, scNBMF, NMF, netNMF-sc, MAGIC, and scImpute, we cluster cells by running k-means on the output from each method. For dimensionality reduction methods (scNBMF, NMF, netNMF-sc) we cluster by running k-means on the low-dimensional cell matrix, $\mathbf{H}$, where the number of dimensions d is selected using holdout validation (Section 2.3.4). For PCA we cluster by running k-means on the top principal components which explain 90% of the variance in the data. For imputation methods (MAGIC and scImpute) we run PCA on the imputed matrices to reduce the dimensionality of the data and cluster by running k-means on the top principal components which explain 90% of the variance in the data. For each method, k-means is run with 100 random initializations and the clusters corresponding to the optimal objective value are reported.

2.3.7 Data simulation

We use a real gene-gene co-expression network obtained from Coexpedia [188] and randomly select 5000 genes to be retained using the *random.sample* command. To define differentially expressed genes, for each of the k clusters, we randomly sample 5 genes and their neighbors to be differentially expressed. If this results in more than 10% of genes being differentially expressed in each cluster, we downsample, at random,

these selected genes such that at most 10% of the genes in each cluster are differentially expressed. Each differentially expressed gene is scaled by a *differential expression factor* as described by Splatter [189], however we ensure that if a gene is overexpressed in a cluster (differential expression factor > 1), then its selected neighbors are also overexpressed. The same is true for underexpressed genes (differential expression factor < 1). Dropout of transcripts is performed following either the double exponential or the multinomial dropout model.

2.4 Discussion

We present netNMF-sc, a method for performing dimensionality reduction and imputation of scRNA-seq data in the presence of high ($> 60\%$) dropout rates. These high dropout rates are common in droplet-based sequencing technologies, such as 10x Genomics Chromium, which are becoming the dominant technology for scRNA-seq. netNMF-sc leverages prior knowledge in the form of a gene coexpression network. Such networks are readily available for many tissue types, having been constructed from bulk RNA-sequencing data, or from other experimental approaches. To our knowledge, the only other method that uses network information to perform dimensionality reduction on scRNA-seq data is [169]. However, this method assumes that there is no dropout in the data, and its performance with high dropout rates is unknown. Moreover, this method uses a neural network that is trained on a specific protein-protein interaction (PPI) network, while netNMF-sc can use any gene-gene interaction network. Another method, netSmooth [177] – published during the preparation of this manuscript – uses network information to smooth noisy scRNA-seq matrices but does not perform dimensionality reduction.

We demonstrate that netNMF-sc outperforms state-of-the-art methods in clustering cells in both simulated and real scRNA-seq data. In addition, netNMF-sc is better able to distinguish cells in different stages of the cell cycle and to classify mouse embryonic brain cells into distinct cell types whose proportions mirror the cellular diversity reported in another study with a substantially greater number of sequenced cells. netNMF-sc imputes values for every entry in the input matrix, similar to MAGIC and in contrast to scImpute which imputes values only for zero counts. Since transcript counts in scRNA-seq data are reduced for all genes, imputation of all values can improve clustering performance and better recover biologically meaningful gene-gene correlations. On multiple datasets, we show that netNMF-sc yields more biologically meaningful gene-gene correlations than other methods. However, one potential downside of imputation is “oversmoothing" of the data resulting in the introduction of artificial gene-gene correlations.

There are multiple directions for future improvement of netNMF-sc. First, netNMF-sc relies on existing gene-gene interaction networks. While we have demonstrated that generic gene coexpression networks [188] can improve clustering performance on human and mouse scRNA-seq data, netNMF-sc may not offer substantial improvements over existing methods on tissues or organisms where high-quality gene-gene interaction networks are not available. In the future, other prior knowledge could be incorporated into netNMF-sc, such as cell-cell correlations, which might be obtained from underlying knowledge of cell types or from spatial or temporal information. Second, there are several additional sources of variation in scRNA-seq data in addition to dropout, such as cell cycle and batch effects. netNMF-sc may be able to assist in removing these confounding effects by encouraging correlations between genes that are connected in the network, thus down-weighting correlations induced by these or other confounding effects. Evaluating the effectiveness of netNMF-sc in the presence of these additional sources of variation is left as future work. Finally, there remains the issue of whether one should identify discrete cell clusters or continuous trajectories in scRNA-seq data. Here we focused on clustering cells in the low-dimensional space obtained from netNMF-sc. An potential future direction is to investigate how to leverage prior knowledge in trajectory inference from scRNA-seq data.

Chapter 3

Identifying CNAs from Spatial Transcriptomics RNA-seq data

3.1 Abstract

Tumors are highly heterogeneous, consisting of cell populations with transcriptional and genetic diversity. These diverse cell populations are spatially organized, creating spatial structure within the tumor microenvironment. A new technology called *spatial transcriptomics* can recover spatial patterns of gene expression within a tissue by sequencing the RNA from a grid of spots, each containing a small number of cells. In a tumor, these gene expression patterns represent the combined contribution of gene regulatory mechanisms, which alter the rate at which a gene is transcribed, and genetic diversity, such as copy number aberrations (CNAs), which affects the number of copies of a gene in the genome. The presence of CNAs in a tumor cell can be inferred from the cell's gene expression profile, but is complicated by transcriptional variation as well as sparsity of the data. We use the observation that cells nearby in space are likely to share similar CNAs and propose a method to incorporate known spatial relationships between spots to aid in inferring CNAs from spatial transcriptomics sequencing of tumors. We find that incorporating spatial information improves CNA inference on simulated data.

3.2 Introduction

A new technology called *spatial transcriptomics* [136] can recover spatial patterns of gene expression within a tissue by sequencing the RNA from a grid of spots, each containing a small number of cells. In a tumor, these gene expression patterns represent the combined contribution of gene regulatory mechanisms, which alter the rate at which a gene is transcribed, and genetic diversity, which affects the number of copies of a gene in the genome. In cancer cells, the variance in expression of a given mRNA transcript is a combination of regulatory mechanisms as well as Copy Number Aberrations (CNAs). These CNAs can amplify or delete copies of the gene in the genome, resulting in more or fewer mRNA transcripts of that gene. If we knew the CNAs present in a cell's genome, then we could correct the cell's expression profile for variance due to CNAs and be able to determine if a gene is differentially expressed due to regulatory mechanisms or changes in copy number.

Calling CNAs from DNA sequencing has a long history. Many methods, such as APOLLOH [110], TITAN [111], and others employ HMMs to model dependencies between adjacent segments of the genome, where the hidden states of the HMM are the copy number states. Inferring CNAs RNA-seq data has only recently become an active area of research. Several recently developed methods attempt to match cells from a single-cell RNA-seq (scRNA-seq) experiment to predefined clones (obtained from single-cell DNA-seq or other methods). These include CloneAlign [144], which uses variants and read depth to match cells to clones, Cardelino [135], which uses a probabilistic model to match cells to clones, and LIAYSON [143], which deconvolutes bulk CNV profiles into cell-specific CNAs. Several methods [112, 113] have been developed to infer chromosome-level CNAs from scRNA-seq data. Two methods, HoneyBADGER [120] and InferCNV [142] directly infer megabase or smaller CNAs from scRNA-seq data. Both methods use HMMs which emit either Deletion, Neutral, or Amplification for each gene, where neighboring genes on the chromosome are connected by an edge in the HMM. HoneyBADGER also incorporates SNPs to identify loss of heterozygosity events.

Since cells replicate by dividing, it is reasonable to assume that nearby cells are likely to have a recent common ancestor. That means that they are also likely to contain the same set of CNAs. We propose to use the spatial relationships between spots to infer CNAs from spatial transcriptomics data. Other recently developed methods for spatial transcriptomics and its counterpart single-molecule fluorescence in situ hybridization (smFISH) have incorporated spatial relationships between cells to cluster cells or identify spatially distributed differentially expressed genes. SpatialDE [123] identifies spatially distributed differentially expressed genes

from spatial transcriptomics or smFISH data using Gaussian process regression, decomposing expression variability into spatial and non-spatial components. Spatial Variance Component Analysis (SVCA) [139] also uses a Gaussian process framework to quantify the effect of cell-cell interactions on gene expression. [138] use a Poisson factorization model to perform factor analysis on spatial transcriptomics data, identifying factors which incorporate both gene activity and spatial activity. Two methods, by [134] and [145], developed for smFISH us a Hidden Markov Random Field (HMRF) to identify spatially distributed cell clusters.

We propose to model both the spatial relationships between spots and the relationships between genes on a chromosome using a combination of a HMRF and an HMM. Given a gene and spot pair, the copy number state of each gene will be influenced by the copy number states of neighboring genes, similar to HoneyBADGER and InferCNV. However, by utilizing the additional spatial information, the clone assignment of each spot will be influenced by the clone assignment of neighboring spots.

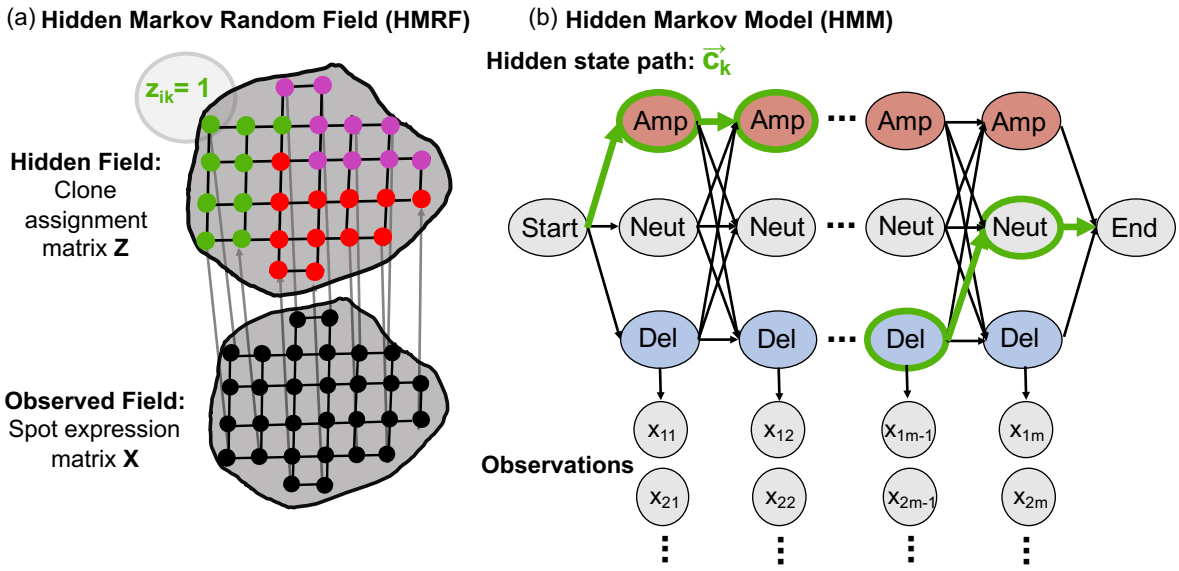


Figure 3.1: (a) Hidden Markov Random Field (HMRF) for modeling spot clone assignment matrix **Z**. The observed field is spot expression matrix **X**. The hidden field is clone assignment matrix **Z**. $z_{ik} = 1$ if spot i is assigned to clone k and 0 otherwise. (b) Hidden Markov Model (HMM) for modeling CNA profile c_k for clone k. The observations are the expression profiles $\vec{x}_i$ for spots assigned to clone k. The hidden state path $\vec{c}_k$ represents the sequence of CNA states from the alphabet $S = \{\text{Deletion,Neutral,Amplification}\}$. The hidden state path for the green clone k is highlighted.

3.3 Method

3.3.1 Problem Definition

The observed expression of each gene in a cell is determined both by the copy number state of that gene as well as regulatory mechanisms which alter its rate of transcription. We normalize out the majority of the variance in expression due to regulatory mechanisms by combining the expression of multiple nearby genes into *bins*. This relies on the assumption that nearby genes will participate in many different regulatory pathways and thus have different rates of transcription. See section 3.3.8 for full details. We further assume that the cells contained within each spot have the same copy number profile. While this is a strong assumption, it has been shown that nearby cells are close evolutionary ancestors [107, 109]. Thus, cells contained within the same 100μl spot are likely to share the same set of CNAs.

Due to length of the average CNA being much larger than the length of the average gene, we expect many CNAs to span multiple bins along a chromosome (see Section 3.3.8 for full details). Thus, we expect the copy number of bin $i+1$ to be dependent on the copy number of bin i. Similarly, neighboring spots in a tissue are likely to originate from the same clone, i.e. have the same CNA profile [107] . We can represent these spatial dependencies by connecting neighboring spots by an edge, creating a grid network $G = (V,E)$, where an edge $(v_i, v_j) \in E$ represents two neighboring spots (Figure 3.1).

Given an n spot by m bin observed expression matrix $\mathbf{X}^{n\times m}$, where row $\vec{x}_i = [x_{i1}, ..., x_{im}]$ is the expression profile of spot i, and a spot grid network $G = (V,E)$, we wish to (1) infer clone assignment matrix $\mathbf{Z}^{n\times K}$, where $z_{ik} = 1$ if spot i belongs to clone k and 0 otherwise, and (2) infer CNA profile matrix $\mathbf{C}^{K\times m}$, where row $\vec{c}_k = [c_{k1}, ..., c_{km}]$ is the copy number profile of clone k.

In brief, our method consists of iterating between updating the entries of the CNA profile matrix $\mathbf{C}$ given the clone assignment matrix $\mathbf{Z}$ and updating the entries of the clone assignment matrix $\mathbf{Z}$ given the CNA profile matrix $\mathbf{C}$. We model dependencies between adjacent bins along a chromosome arm by a hidden Markov model (HMM) and model dependencies between neighboring spots indicated by $G = (V,E)$ by a hidden Markov random field (HMRF). The details of the HMM and HMRF are described below.

3.3.2 Hidden Markov Model (HMM) for predicting CNA profiles

Because CNAs do not span across different chromosomes or chromosome arms, the observed expression across each chromosome arm can be treated as independent of the expression of any other arm. Thus, CNA

profiles can be independently inferred for each chromosome arm. Furthermore, we make the assumption that observed spots originate from one of K clones with distinct CNA profiles. Thus, we can separately infer CNA profiles for pairs (a,k) of chromosome arm a and clone k. For ease of explanation, in the following section we will introduce our model under a simplified assumption that all bins $1,...,m$ originate from a single chromosome arm and all spots $1,...,n$ originate from a single clone.

We can model the expression profile $\vec{x}_i = [x_{i1},...,x_{im}]$ of spot i as an ordered sequence of observed symbols emitted from an HMM with underlying state sequence $\vec{c} = [c_1,...,c_m]$, where each symbol $x_{ij} \in \mathbb{R}$ and each state c_j takes one of the values from the set of states $S = \{\text{Deletion, Neutral, Amplification}\}$. We wish to infer the sequence $\vec{c}$ of hidden states which emitted the observed expression profile $\vec{x}_i$.

We assume that the hidden state sequence $\vec{c}$ is a first-order Markov chain. This means that the probability of being in state a at position $p+1$ depends only on the state at position p. The probability of transitioning from state a to state b is given by transition matrix T. For the initial state c_1, we denote the initial probability $\pi(s) = P(c_1 = s), \forall s \in S$. We model the emission probability at state s by a Gaussian with parameters (μ_s, σ_s), such that

$$P(x_{ij}|c_j = s) = G(x_{ij}; \mu_s, \sigma_s).$$

The full HMM can be specified completely by (1) the transition matrix T, (2) the initial state probability $\pi(s)$, and (3) the parameters $\Theta = \{\mu_{\text{Del}}, \sigma_{\text{Del}}, \mu_{\text{Neut}}, \sigma_{\text{Neut}}, \mu_{\text{Amp}}, \sigma_{\text{Amp}}\}$ of the Gaussian emission probability. We let $\lambda = (T, \pi, \Theta)$ define the HMM. The formulation of the HMM is similar to the formulations used by CNA callers designed for DNA sequencing data, such as APOLLOH [110] and TITAN [111].

Now that we have defined the HMM, we relax the assumption that our data originates from a single clone and assume the data originates from K clones. We can estimate the values of the parameters $\lambda = (T, \pi, \Theta)$ using expectation maximization EM. The parameters λ of the HMM are assumed to be shared across each chromosome arm and clone, so we only learn a single set of parameters for the observed data $\mathbf{X}$. For the expectation step, we use the forward-backward algorithm to compute the joint posterior log-likelihood of the CNA matrix $\mathbf{C}$ given the observed data $\mathbf{X}$, the clone assignment matrix $\mathbf{Z}$, and the parameters λ. Since the expression profiles of each spot are independent of each other spot given the model parameters λ, we obtain the following function for the posterior log-likelihood:

$$P(\mathbf{C}|\mathbf{X}, \mathbf{Z}, \lambda) = \sum_{i=1}^{n} \log z_{ik}(P(c_k|\vec{x}_i, \lambda)), \tag{3.1}$$

where $\vec{x}_i$ is the expression profile of spot i and $z_{ik} = 1$ if spot i is assigned to clone k and 0 otherwise. For the

maximization step of EM, we estimate the values of the parameters $\lambda = (T, \pi, \Theta)$ using coordinate descent until convergence criteria are met. The Viterbi algorithm is then applied independently for each chromosome arm and clone to find the state path which maximizes the posterior log-likelihood given the learned values for λ.

3.3.3 Markov Random Field (MRF) model

We can model clone assignment matrix $\mathbf{Z}$ as a configuration of a random field with respect to graph $G = (V, E)$, where E is the edges (v_i, v_j) between neighboring spots. Let Z be a random field with finite state space A, where state $a \in A$ is a length K vector with all entries $a_i = 0$ except for $a_k = 1$. and let $\mathbf{Z}$, where $z_i \in A$, be a configuration of random field Z with respect to graph $G = (V, E)$. Z is a MRF with respect to neighborhood system $\{N_i | i \in V\}$ only if Z obeys the local Markov property:

$$P(z_i | z_j, i \neq j) = P(z_i | z_j j \in N_i),$$

meaning that the clone assignment z_i of vertex V_i depends only on the clone assignments z_j of its neighboring vertices $V_j \in N_i$.

3.3.4 Hidden Markov Random Field (HMRF)

We seek to infer a clone assignment matrix $\hat{\mathbf{Z}}$ which is an estimate of the true clone assignment matrix $\mathbf{Z}^*$ according to the MAP criterion

$$\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z} \in \mathscr{Z}} P(\mathbf{X}|\mathbf{Z}) P(\mathbf{Z}), \tag{3.2}$$

where $\mathscr{Z}$ is the set of all possible configurations of $\mathbf{Z}$. Given $\hat{\mathbf{Z}}$, we can model the observed data $\mathbf{X}$ by a HMRF. The HMRF model is characterized by three properties: (1) a **hidden random field**, where the configuration of the field is unobservable, (2) an **observable random field**, where the random variables of the observable random field follow a known emission probability distribution given a specific configuration of the hidden random field, (3) given a configuration of the hidden random field, the random variables of the observed random field are **conditionally independent**.

We define a HMRF with hidden random field Z and observable random field X such that $\mathbf{Z}$ is a configuration of random field Z with respect to graph $G = (V, E)$. We assume the observed random variables $\vec{x}_i$ of field X follow a multivariate Gaussian emission probability. Given a configuration of hidden field Z where $z_{ik} = 1$,

$\vec{x}_i = x$ is emitted with probability $G(x; d_k, \Sigma_k)$, where d_k is the mean vector of clone k and Σ_k is covariance matrix of clone k.

3.3.5 HMRF algorithm for predicting clone assignments

Given the CNA profile matrix $\hat{\mathbf{C}}$ and parameters Θ of the emission probability we can infer the clone assignment matrix $\hat{\mathbf{Z}}$

$$\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z} \in \mathscr{Z}} P(\mathbf{X}|\mathbf{Z}, \hat{\mathbf{C}}, \Theta) P(\mathbf{Z})$$

The conditional probability distribution $P(\mathbf{X}|\mathbf{Z}, \hat{\mathbf{C}}, \Theta)$ is the emission probability function. We model the emission probability of clone k by a multivariate Gaussian distribution G with parameters $\vec{\mu}_k$ and Σ_k. Let c_{kj} denote the state of bin j in clone k. The j^{th} entry of mean vector $\vec{\mu}_k$ is the expected value of bin j in clone k, which we have estimated to be $\mu_{c_{kj}}$ in the previous step. Thus, we define the mean vector $\vec{\mu}_k = \mu_{c_{k1}}, ..., \mu_{c_{km}}$. Similarly, we define the covariance matrix Σ_k such that the diagonal entries $\Sigma_{jj} = \sigma^2_{c_{kj}}$ and all other entries are 0. We can thus express the joint likelihood probability,

$$\begin{aligned} P(\mathbf{X}|\mathbf{Z}, \hat{\mathbf{C}}, \Theta) &= \prod_{i=1}^{n} P(\mathbf{x}_i|z_i, \hat{\mathbf{C}}, \Theta) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} z_{ik} G(x_i; \vec{\mu}_k, \Sigma_k) \end{aligned}$$

The prior probability $P(\mathbf{Z})$ is a Gibbs distribution

$$P(\mathbf{Z}) = \frac{1}{q} \exp(-U(\mathbf{Z})) \tag{3.3}$$

$$U(\mathbf{Z}) = \sum_{i,j \in E} Y(z_i, z_j) \tag{3.4}$$

$$Y(z_i, z_j) = \begin{cases} 0 & \text{if } z_i = z_j \\ \beta & \text{if } z_i \neq z_j, \end{cases} \tag{3.5}$$

where $q = 2\pi^{|S|/2}$ is a normalizing constant. Using these equations we use Iterated Conditional Modes (ICM) to infer a locally optimal $\mathbf{Z}$ by iteratively updating z_i conditioned on i's neighbors until convergence. We then iterate between optimizing for $\hat{\mathbf{Z}}$ given $\hat{\mathbf{C}}$ and optimizing $\hat{\mathbf{C}}$ given $\hat{\mathbf{Z}}$ until convergence.

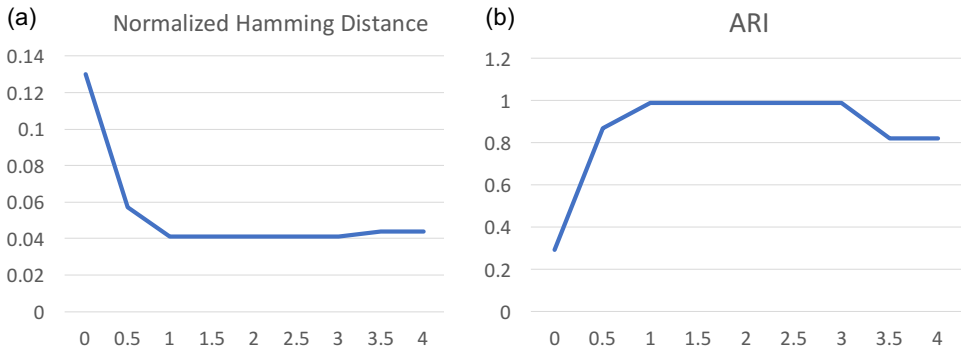


Figure 3.2: (a) Normalized Hamming distance between true and inferred CNA profiles on simulated data for STCNA run with different range of β. (b) Adjusted Rand Index (ARI) between true and inferred clone assignments on simulated data for STCNA run with range of β.

3.3.6 Selecting the value of parameter β

To select the optimal value of parameter β we measured the distance between true and inferred CNA matrices $\mathbf{C}^*$ and $\hat{\mathbf{C}}$ as well as the difference between true and inferred clone assignment matrices $\mathbf{Z}^*$ and $\hat{\mathbf{Z}}$ over a range of β values. We found that both the inferred CNAs and clone assignments were stable across a range of β from 1 to 3 (Figure 3.2). The distance between true and inferred CNA matrices, measured by normalized Hamming distance, as well as the difference between true and inferred clone assignment matrices, measured by ARI, increased for β values less than 1 or higher than 3 (Figure 3.2). We use $\beta = 2$ for all results on simulated and real data.

3.3.7 Parameter initialization

We initialize the transition matrix T,

$$T = \begin{bmatrix} 1-2t & t & t \\ t & 1-2t & t \\ t & t & 1-2t \end{bmatrix},$$

where $t = 10^{-5}$. The clone assignment matrix $\mathbf{Z}$ is initialized by performing K-means clustering on the rows of the observed expression matrix $\mathbf{X}$. The number of clones K may be selected either prior knowledge or by computing the average silhouette score for a range of K and selecting the value of K with the highest average silhouette score. After initializing the clone assignment matrix $\mathbf{Z}$ we initialize the CNA profile matrix $\mathbf{C}$ and

the values for parameters (μ_s, σ_s). For each clone k, we compare the observed expression of spots assigned to clone k to the expression of normal spots in each bin $1 \leq j \leq m$. The distance between these samples is measured by a two-sample Kolmogorov–Smirnov (KS) test. If the hypothesis that the two samples are drawn from the same distribution is rejected ($p \leq .001$), then $c_{kj} =$ Amplification or $c_{kj} =$ Deletion depending on whether the mean of clone k spots is higher or lower than the normal spots. The parameters (μ_s, σ_s) are then initialized to the means and standard deviations of bins assigned to each state s.

3.3.8 Binning genes

The median length of a human gene is approximately 24Kb [115], with average intergenic distances between genes being 4Kb or smaller [114]. Comparatively, the median length for a amplification or deletion across many cancer types has been found to be approximately 900Kb and 700Kb respectively, with focal CNAs being much longer at approximately 19600Kb and 22700Kb respectively [117]. Thus we expect the median CNA to span 28 genes or more. We bin genes in windows of w genes with a step size of s. We bin based on genes rather than genomic intervals because the observed measurements are at the gene level and this allows measurements to be directly comparable across bins. For all experiments we bin with a window size w such that there is a median of 50X coverage per gene. This typically results in window size between $30-50$. For all experiments we use a step size of 1, so we get a copy number call for all m genes in the dataset.

3.3.9 Data Preprocessing

Prior to running STCNA the raw expression matrix $\mathbf{X}$ is preprocessed via the following steps inspired by InferCNV. First, genes with non-zero counts in fewer than 20 cells are filtered out. The data is then library-size normalized, where counts for each gene in a spot are divided by the sum of counts for the spot. This normalization removes variance due to differences in size or number of cells in each spot. Next, the data is log-normalized with pseudocount 1 ($\log(\mathbf{X}+1)$). Then, the data is binned to remove variation in expression from regulatory mechanisms acting at the gene level (see Section 3.3.8), retaining only large-scale variations due to copy number changes. For each bin, the median of the normal spots is subtracted from the tumor spots to remove differences in total expression between bins and allow for direct comparison of expression across bins. Finally, the log-transformation is inverted $(\exp(\mathbf{X})-1)$.

3.3.10 Normalized Hamming Distance

The Hamming distance between two CNA profiles a and b is the total number of bins in which the two profiles differ. Let the hamming distance between a and b be denoted $H(a,b)$ and let $|a|$ denote the total number of non-neutral bins in CNA profile a and $|b|$ denote the total number of non-neutral bins in CNA profile b, where a non-neutral bin is either an amplification or deletion. We define normalized Hamming distance to be

$$H_n(a,b) = \frac{H(a,b)}{|a|+|b|}$$

. Thus, if all non-neutral bins are the same between a and b, then $H_n(a,b) = 0$ and if all non-neutral bins differ between a and b then $H_n(a,b) = 1$. Without this normalization, CNA profiles with few non-neutral bins will almost always have lower Hamming distance than CNA profiles with many non-neutral bins even if none of the bins match between profiles a and b.

3.4 Results

3.4.1 Results on Simulated Data

We simulated a tumor tissue with 150 spots from 3 spatially distributed clones, A (75 spots), B (50 spots), and C (25 spots). The CNAs present in each of these clones are the same as those identified by scDNA-seq of a triple-negative breast cancer patient-derived xenograft SA501 sequenced and analyzed by [144]. [144] also sequenced and analyzed scRNA-seq expression data from the same patient-derived xenographt SA501 and assigned each of the 1152 cells to one of the clones (A, B, C) identified by scDNA-seq. After preprocessing (see Section 3.3.9) the scRNA-seq data contained 6557 genes. We simulated a 6557×150 expression matrix $\mathbf{X}$ using a three-component Gaussian mixture model with parameters $(\mu_{\text{Del}}, \sigma_{\text{Del}})$,$(\mu_{\text{Neut}}, \sigma_{\text{Neut}})$, and $(\mu_{\text{Amp}}, \sigma_{\text{Amp}})$ respectively,

$$\mathbf{x}_i \sim \begin{cases} G(\mu_{\text{Del}}, \sigma_{\text{Del}}) & \text{if } c_i = \text{Deletion}, \\ G(\mu_{\text{Neut}}, \sigma_{\text{Neut}}) & \text{if } c_i = \text{Neutral}, \\ G(\mu_{\text{Amp}}, \sigma_{\text{Amp}}) & \text{if } c_i = \text{Amplifciation}. \end{cases} \tag{3.6}$$

The values of μ_{Del} and σ_{Del} are the mean and standard deviation of deleted segments from the scRNA-seq data of sample SA501 and similarly for neutral and amplified segments. We define a graph $G = (V,E)$ of spot

relationships. Each spot is connected by an edge to at most 4 neighboring spots. We align the 150 spots in a grid such that spots from the same clone are likely to be near each other. We select 25 spots from clone A to represent normal spots and use those to normalize the simulated expression matrix $\mathbf{X}$.

We ran STCNA both with the spatial information and without. We denote STCNA run with the spatial information ($\beta = 1$) and without the spatial information ($beta = 0$) as STCNA-HMRF and STCNA-HMM respectively. We also compare to another copy-number inference method developed for scRNA-seq called InferCNV. By incorporating spatial information, STCNA-HMRF outperformed STCNA-HMM and InferCNV at recovering the simulated CNAs, measured by normalized hamming distance (see Section 3.3.10), and assigning spots to their respective clones, measured by ARI (Figure 3.3).

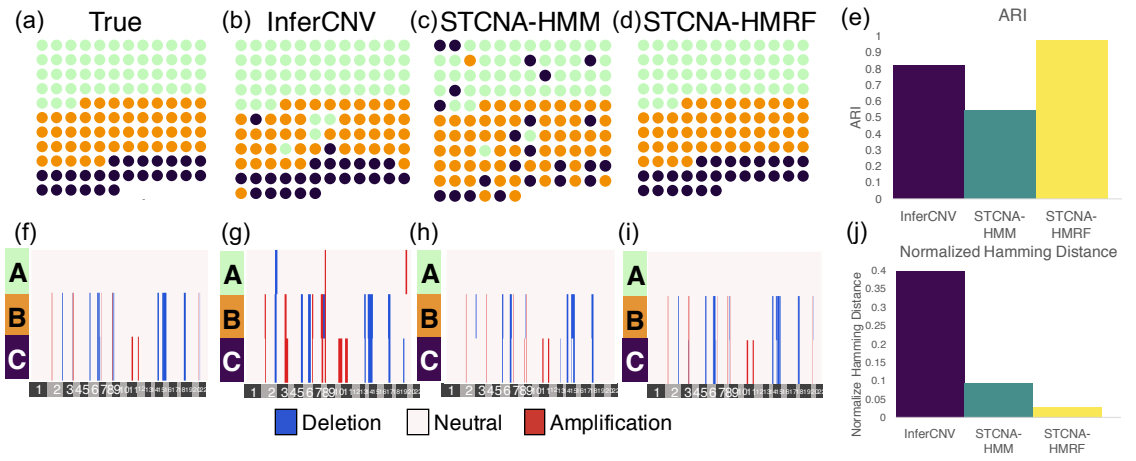


Figure 3.3: (a-d) Assignment of spots to clones A (green), B (orange) and C (purple). (f-i) CNA profiles for each clone. (e) Similarity between true and inferred clone assignments for each method measured by Adjusted Rand Index (higher is better). (j) Normalized Hamming distance between true and inferred CNA profiles (lower is better).

3.4.2 Pseudo-spatial transcriptomics from matched scDNA-seq and scRNA-seq

Patient-derived xenographt SA501

As described in the previous section, we obtained scRNA-seq expression data of 1152 cells from triple-negative breast cancer patient-derived xenograft SA501 as well as the copy number profiles of three clones, derived from scDNA-seq, present in this tissue from [144]. Each of the 1152 cells was matched to one of three clones (A, B, C) by clonealign [144]. We simulated pseudo-spatial relationships between these 1152 cells assigned to the same clone (Figure 3.4). The data also did not include matched normal (reference) cells

to which we could normalize the scRNA-seq data. Because of differences in mean expression of each gene, it is necessary to normalize the expression to a reference so that relative expression can be compared between genes. To overcome this, we randomly selected 200 of the 930 cells assigned to clone A and defined them as normal cells, normalizing each gene by the median of its expression in these normal cells.

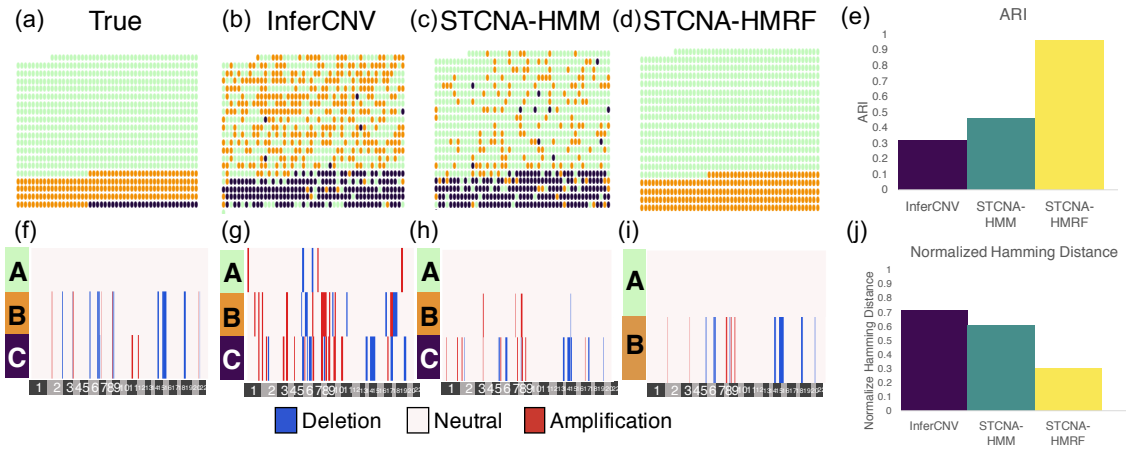


Figure 3.4: Clone assignment and CNA inference results from InferCNV, STCNA-HMM, and STCNA-HMRF on patient-derived xenograft SA501. (a-d) Assignment of spots to clones A (green), B (orange) and C (purple). (f-i) CNA profiles for each clone. (e) Similarity between true and inferred clone assignments for each method measured by Adjusted Rand Index (higher is better). (j) Normalized Hamming distance between true and inferred CNA profiles (lower is better).

Using the expression profiles obtained from scRNA-seq and the simulated spatial relationships between cells, we ran STCNA-HMRF, STCNA-HMM, and InferCNV and compared the inferred CNAs to the true CNAs. We ran each method with the number of clones $K = 3$. CNA inference from this real scRNA-seq data is more difficult than the simulated data because (1) there are additional sources of variation due to variance in gene expression and dropout that were not captured in the simulated data, (2) each spot may contain cells from multiple clones result in in mixed signal, and (3) the proportions of each cell type are very unequal, with 930 spots assigned to clone A, 192 spots assigned to clone B and only 30 spots assigned to clone C.

We found that by incorporating spatial information STCNA-HMRF successfully assigned cells to clones A and B with high accuracy (Figure 3.4). However, none of the methods were able to identify the 30 spots from clone C which differ from clone A by only two amplifications on chromosome 11. STCNA-HMRF assigned all of these 30 spots to clone B, which has a very similar expression profile to clone C. InferCNV and STCNA-HMM, however, assigned these spots to three different clones, none of which matched the true expression profile of clone C (Figure 3.4). Though none of the methods correctly inferred the CNA profiles

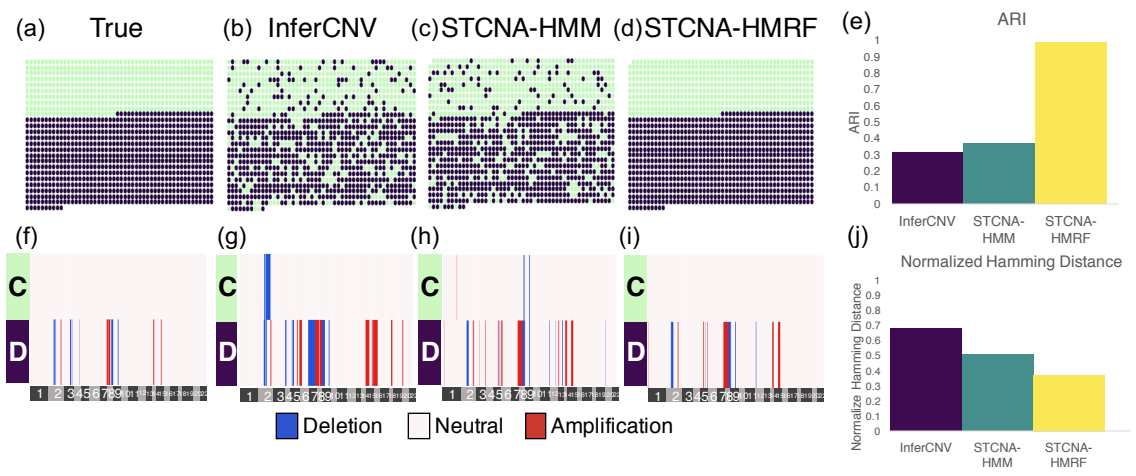


Figure 3.5: Clone assignment and CNA inference results from InferCNV, STCNA-HMM, and STCNA-HMRF on high grade serous carcinoma cell line OV2295. (a-d) Assignment of spots to clones A (green), B (orange) and C (purple). (f-i) CNA profiles for each clone. (e) Similarity between true and inferred clone assignments for each method measured by Adjusted Rand Index (higher is better). (j) Normalized Hamming distance between true and inferred CNA profiles (lower is better).

of all clones, STCNA-HMRF inferred the CNA profile of clones A and B with low error and assigned most of the spots to their respective clones (ARI .97), while InferCNV and STCNA-HMM identified many CNAs not present in the data and failed to correctly assign spots to clones (ARI of .32 and .46 respectively).

high grade serous carcinoma cell line OV2295

We also obtained matched scRNA-seq and scDNA-seq data from a high grade serous carcinoma cell line, OV2295, from [144]. The scRNA-seq data contains 1460 cells assigned to clones C (674 cells) and D (786 cells), with CNA profiles inferred from scDNA-seq data. We found that by incorporating spatial information STCNA-HMRF successfully assigned cells to clones C and D with high accuracy (ARI=.99), while InferCNV had low accuracy (ARI=.31) and STCNA-HMM failed at assigning cells to clones (ARI=0.02) (Figure 3.5). In addition, STCNA-HMRF had the lowest error in its inferred CNA profiles, with a normalized hamming distance of .44 compared to InferCNV with .68 and STCNA-HMM with .76.

3.4.3 Results on STRNA-seq of breast cancer biopsy

We have demonstrated that STCNA can recover CNA profiles and accurately assign cells to their respective clones from scRNA-seq data with simulated spatial organization. Now, we demonstrate the performance on

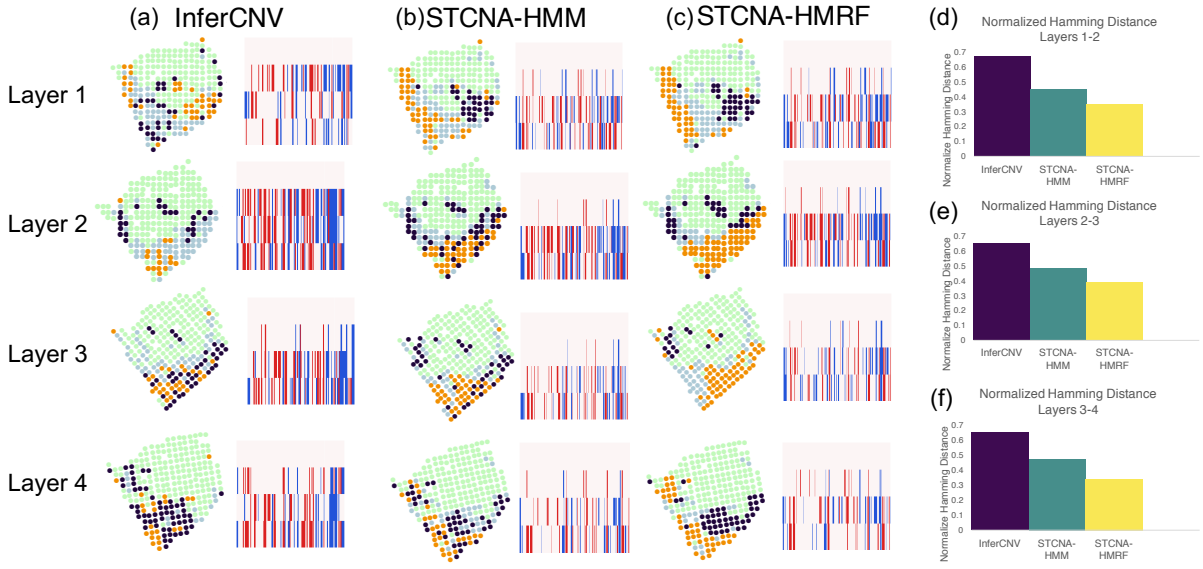


Figure 3.6: Results of applying CNA inference methods to STRNA-seq of four layers from a breast cancer biopsy. For each method, CNA inference was performed separately for each of the four layers. (a) Results from InferCNV. (b) Results from STCNA-HMM (STCNA with $\beta = 0$). STCNA-HMM does not use spatial information. (c) Results from STCNA-HMRF (STCNA with $\beta = 2$). STCNA-HMRF uses spatial information.

STCNA on STRNA-seq data. Currently, no matched STRNA-seq and scDNA or bulk DNA-seq datasets exist, so we cannot compare CNA profiles inferred by STCNA to a ground truth. However, [] sequenced four layers of a breast cancer tissue biopsy. If we assume that the same clones are likely to be present between any two tissue layers, then we can assess the performance of CNA inference methods by measuring the normalized Hamming distance between adjacent layers (1-2,2-3, and 3-4).

For each layer, we first have to differentiate tumor from normal spots so that we can normalize the expression of tumor spots relative to the normal spots as described in Section 3.3.9. It relatively easy to distinguish by eye the tumor and normal sections of the tissue from the histopathology images. From analyzing many STRNA-seq datasets, we also notice that clustering into two clusters based on just the first principal component of the expression matrix results in a segmentation that visually matches that of the histopathology images, indicating that the differences between tumor and normal spots is the largest source of variation in the data. We also note that the first principal component correlates with the library-size of each spot as well as the proportion of zero entries, with normal spots having on average smaller library size and higher proportion of zero entries. For each of the four layers of the breast cancer biopsy, we use the first principal component to segment spots into two clusters and denote spots in the cluster with the lower average library-size as the

normal spots.

We then ran STCNA-HMRF, STCNA-HMM, and InferCNV on the tumor spots from each layer independently with $K = 3$. For each method, we compared the inferred CNA profiles between layers 1-2, 2-3, and 3-4 respectively. The CNA profiles inferred by STCNA had lower normalized Hamming distance $(.35, .39, .34)$ than InferCNV $(.67, .65, .65)$ and STCNA-HMM $(.45, .48, .47)$ (Figure 3.6(d-f)). The better correspondence of CNA profiles between layers indicates that STCNA is able to recover more accurate CNA profiles by utilizing the spatial relationships between spots. In addition, the spots assigned to each of the three clones inferred by STCNA-HMRF are more contiguous within the tissue than the other two methods, where spots from each clone are more mixed across the tissue (Figure 3.6(a-c))).

3.5 Conclusion

We propose a new method, STCNA, to infer CNAs from spatial transcriptomics (STRNA-seq) data. STCNA jointly infers the assignment of cells to one of K clones and the CNA profiles of each clone by using relationships between neighboring spots and neighboring genes to improve clone assignment and CNA profile inference. We demonstrate that by incorporating spot relationships, STCNA outperforms other methods at recovering the true clone assignments and CNA profiles of simulated data. We also demonstrate that STCNA outperforms other methods on real STRNA-seq data, inferring CNA profiles that are consistent across adjacent tissue layers of a breast cancer biopsy.

There are several potential extensions to STCNA that we leave as future work. The first is incorporating information variant calling of raw reads from STRNA-seq. CNA inference methods such as TITAN [111] developed for DNA-seq and HoneyBADGER [120] developed for scRNA-seq use SNPs to identify loss-of-heterozygosity (LOH) events. While it is unclear how informative SNPs called from STRNA-seq data will be due to the data having low coverage per spot, it is an interesting future direction. Another potential extension to STCNA is allowing for more than three hidden states (Deletion, Neutral, Amplification). For example, we could allow for loss or gain of one and two copies, resulting in five hidden states.

Chapter 4

Identifying structural variants using linked-read sequencing data

4.1 Abstract

Structural variation, including large deletions, duplications, inversions, translocations, and other rearrangements, is common in human and cancer genomes. A number of methods have been developed to identify structural variants from Illumina short-read sequencing data. However, reliable identification of structural variants remains challenging because many variants have breakpoints in repetitive regions of the genome and thus are difficult to identify with short reads. The recently developed linked-read sequencing technology from 10X Genomics combines a novel barcoding strategy with Illumina sequencing. This technology labels all reads that originate from a small number (~5-10) DNA molecules ~50Kbp in length with the same molecular barcode. These barcoded reads contain long-range sequence information that is advantageous for identification of structural variants. We present Novel Adjacency Identification with Barcoded Reads (NAIBR), an algorithm to identify structural variants in linked-read sequencing data. NAIBR predicts novel adjacencies in a individual genome resulting from structural variants using a probabilistic model that combines multiple signals in barcoded reads. We show that NAIBR outperforms several existing methods for structural variant identification – including two recent methods that also analyze linked-reads – on simulated sequencing data and 10X whole-genome sequencing data from the NA12878 human genome and the HCC1954 breast cancer cell line. Several of the novel somatic structural variants identified in HCC1954 overlap known cancer genes.

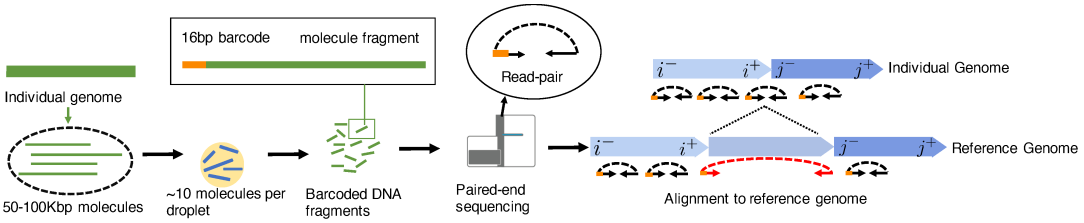


Figure 4.1: (Left) Linked-read sequencing with the 10X Genomics Chromium platform begins by fragmenting the individual genome into large DNA molecules, which are isolated into individuals beads that contain several large molecules and sequencing reagents. Within the bead, molecules are sheared into smaller fragments (500 bp) and labeled with a 16bp barcode indicating its bead of origin. Illumina paired-read sequencing of each fragment results in barcoded paired-end reads. (Right) Alignment of read-pairs to a reference genome results in concordant reads (black) and discordant reads (red). Discordant reads indicate candidate novel adjacencies that are a result of structural variants that distinguish individual genomes from the reference genome.

4.2 Introduction

Recent whole genome sequencing (WGS) analysis of human genomes has shown that *structural variation*, including insertions, deletions, duplications, and rearrangements of genomic segments greater than 50bp, are a key component of human variation [96]. Collectively, structural variants affect a larger portion of the human genome than single nucleotide variants [84]. Inherited germline structural variants have been implicated in several diseases including Crohn's disease, rheumatoid arthritis, Type I diabetes, and autism [83, 89, 100]. In addition, somatic structural variants are common in cancer genomes [72, 86]. These include deletions of tumor suppressor genes and amplifications of oncogenes which can promote aggressive cell growth and drive the development of cancer. Cancer genomes can also undergo dramatic rearrangement events such as chromothripsis, the shattering and random repair of chromosomes in a single catastrophic event [95], or chromoplexy [68], both of which result in a large number of complex structural variants in a cancer genome.

The identification of structural variants from high-throughput DNA sequencing data is generally more challenging than the identification of single nucleotide variants. This difficulty is primarily a result of the the fact that many structural variants are significantly longer than the DNA sequence reads produced by current (second generation) DNA sequencing technologies, whose read lengths are ~300-500 nucleotides. In addition, such reads are too short for *de novo* genome assembly. Thus structural variants are inferred from atypical, or aberrant, alignments of reads to a reference genome.

Numerous methods have been developed over the past several years to identify different types of structural variants from read alignments. Each of these methods use some combination of three signals that can

be extracted from read alignments: *discordant read-pairs*, *split-reads*, and *read depth* (Figure S12a). A discordant read-pair is a pair of reads from the same fragment/insert whose alignments to the reference genome have distance and/or orientation that differ from expected if the entire fragment was contiguous on the reference genome. A split read is a read with no contiguous alignment to the reference genome, but rather with at least two partial alignments to the reference. (In practice, only a single partial, or clipped, alignment may be reported by the read alignment software.) Discordant read pairs and split reads are signatures of a *novel adjacency* in the individual genome; that is, two intervals that are non-adjacent in the reference genome are adjacent in an individual genome. Methods that rely on discordant paired reads and/or split reads include BreakDancer [71], GASV [90], VariationHunter [75], Pindel [102], DELLY [85], and LUMPY [80]; many others are reviewed in [97]. Read depth is the (normalized) number of reads that map to a particular region of the genome. Read depth can be used to identify copy-number aberrations, such as deletions and duplications, by identifying regions of unexpectedly low or high coverage in the genome. Examples of read depth methods include BIC-Seq [101], and CNVnator [66]. In addition, methods such as GASVPro [92] and SV-Bay [78] combine signals from discordant read-pairs and read depth signals to identify structural variants. Local assembly approaches such as SvABA [98], and novoBreak [73] achieve nucleotide level resolution of novel adjacencies. However, local assembly approaches require the identification of candidate regions for assembly using the signals described above; thus, local assembly generally increases specificity more than sensitivity. In addition, assembly-based approaches requires high coverage and is typically more time consuming than read-pair or read depth based methods [67].

The fundamental limitations in structural variant detection and whole-genome assembly are the fact that the human genome is diploid and highly repetitive [82]. Short reads have low signal to assign variants to haplotypes and to identify structural variants whose breakpoints lie in repetitive sequences. While algorithms can help extract information from this low signal, longer reads provide stronger signal. New *3rd-generation* sequencing technologies developed by Pacific Biosciences and Oxford Nanopore produce much longer reads (exceeding 10Kbp). However, these technologies are practically limited by their high per-base error rate and high cost compared to Illumina short-read sequencing. Additionally, structural variants that are larger than the average read size or that fall in repetitive regions still remain difficult to identify using these technologies [99]. An alternative sequencing technology called *linked-read sequencing* was recently developed by 10X Genomics, and commercialized in their Chromium platform. In this technology, long DNA molecules, $50-100$Kbp in size are partitioned into one of several million droplets using microfluidics. Each droplet contains a small number (~10) of molecules [100]. The molecules in each droplet are sheared into smaller fragments

and labeled with a 16bp molecular barcode that is unique to each droplet. The fragments are then amplified and sequenced using Illumina paired-end sequencing protocol (Figure 4.1). 10X Genomics' linked-read technology thus provides both the low error rate and low cost of Illumina sequencing as well as long-range sequencing information provided by 3rd-generation sequencing technologies (Figure S12b). The technology is similar in some respects to the strobe sequencing technology that was prototyped by Pacific BioSciences but never commercially released [87, 88].

Here, we introduce Novel Adjacency Identification with Barcoded Reads (NAIBR, pronounced "neighbor"), a method that identifies novel adjacencies resulting from structural variants in an individual genome from linked-read sequencing data. NAIBR combines a novel split-read type signal from linked-reads with traditional signals of structural variants in the underlying paired-reads in the data. We demonstrate that NAIBR outperforms existing structural variant detection algorithms – both paired-read methods and two recently developed methods [93, 105] for linked-reads – using simulated and real linked-read sequencing data. NAIBR also leverages haplotype phasing information from linked-reads, improving the detection of heterozygous structural variants.

4.3 Methods

Consider two genomes, a *reference genome* and an *individual genome*, each represented by an interval, $G = [1,n]$ and $G' = [1,n']$ respectively. We let $[i^-,i^+]$ denote an interval in the genome, where i^- and i^+ indicate the start and end of the interval respectively. We consider a structural variant to be any difference between an individual genome and a reference genome due to DNA breakage and repair, that results in the joining of two non-adjacent intervals $[i^+,i^-]$ and $[j^+,j^-]$ in the reference genome. The ends of these intervals may be joined in one of four orientations, and we indicate the four possible *novel adjacencies* in the individual genome by the pairs of joined ends: (i^+,j^-),(i^+,j^+),(i^-,j^-), or (i^-,j^+). Note that novel adjacencies are formed by most of the usual structural variants (segmental deletions/insertions, inversions, and translocations), with the notable exception of the deletion of a chromosome to the telomere.

NAIBR aims to identify such novel adjacencies using linked-read sequencing data. Our algorithm differs from previously published methods in that it incorporates signals from both paired-end reads and linked-reads into a unified model. Before defining the model, we will describe the signals we observe in paired-end and linked-read data and how these signals are combined to identify novel adjacencies arising from structural variants in an individual genome.

4.3.1 Paired-end sequencing data

In Illumina paired-end sequencing, chromosomes are sheared into small fragments and size selected such that most fragment lengths are within the interval $[l_{\min}, l_{\max}]$. Each fragment is sequenced from both ends from opposite DNA strands; thus one read will originate from the forward (+) strand and one from the reverse (-) strand. Paired reads are aligned to the reference genome. Each aligned read, x, is represented by a tuple $x = (l_x, r_x, o_x, q_x)$, where l_x is the leftmost position of x in the reference genome, r_x is the rightmost position in the reference genome, $o_x \in \{+,-\}$ is the orientation of x, and q_x is the mapping probability of x. We define a read-pair to be the ordered pair $\langle x, y \rangle$, where read x has the smaller starting coordinate. A read-pair $\langle x, y \rangle$ is *concordant* provided the distance between aligned reads $f = r_y - l_x$ isbetween $l_{\min}$ and $l_{\max}$ and the orientations are $o_x = +, o_y = -$. Concordant reads are consistent with the fragment aligning contiguously to the reference genome with no rearrangement. A read pair $\langle x, y \rangle$ not satisfying this condition is *discordant*. Discordant read-pairs arise from either (1) errors in sequencing and/or alignment or (2) novel adjacencies in an individual genome with respect to the reference.

4.3.2 Linked-read sequencing data

The linked-read sequencing technology developed 10X Genomics adds a layer of structure to Illumina paired-end sequencing by tagging each DNA fragment with a barcode prior to sequencing (Figure 4.1). Barcoded read-pairs are aligned to the reference using a linked-read aware aligner called Lariat [70]. Lariat processes all read-pairs from a single barcode simultaneously, using the knowledge that reads originate from a small number of long molecules. Using this prior knowledge, it finds more unique mappings than other tools and can map to highly repetitive regions.

Read-pairs $p = \langle x, y \rangle$ originating from the same long molecule will each be tagged with the same barcode $b_p \in \mathbb{N}$. Because molecules are partitioned into droplets uniformly at random, the likelihood of assigning the same barcode to two molecules from nearby locations on the reference genome is low. Thus, we assume that read-pairs with the same barcode that map near each other on an individual genome are likely to have originated from the same long molecule. We partition such read-pairs into sets called *linked-reads*.

A linked-read is a set of concordant read-pairs that share the same barcode and have a maximum distance of δ from another read-pair in the linked-read. Each linked-read is assumed to have originated from a contiguous strand of DNA in the reference genome. The set of all concordant read-pairs in the genome is partitioned into linked-reads such that any two read-pairs $p = \langle x, y \rangle$ and $p' = \langle x', y' \rangle$, where $l_x < l'_x$, are both

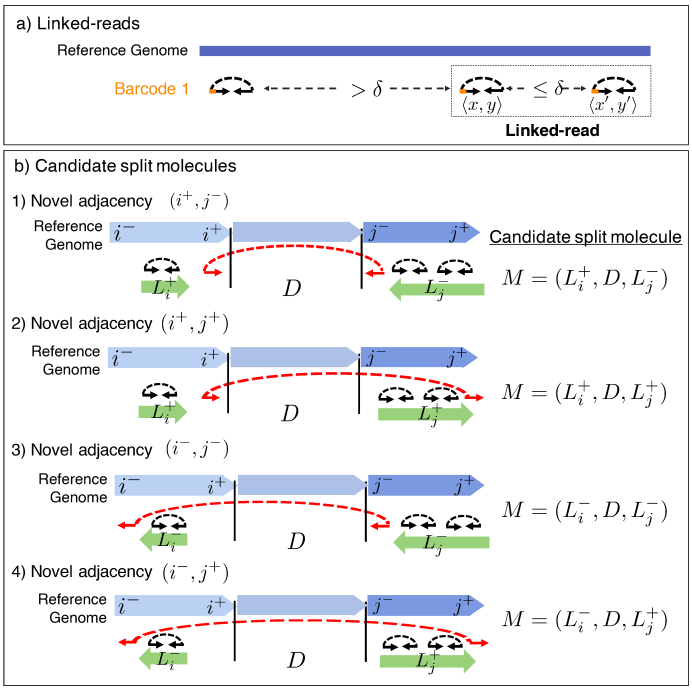


Figure 4.2: (a) A linked-read is defined by read-pairs separated by a distance $\leq \delta$ on the reference genome. (b) Linked-reads L_i and L_j may have originated from one of 4 candidate split molecules, each supporting a novel adjacency with a different orientation. $M = (L_i^+, D, L_j^-)$ supports a novel adjacency (i^+, j^-) and indicates that the end of linked-read L_i is adjacent to the start of linked-read L_j (the arrows points to the location of the novel adjacency).

partitioned into the same linked-read if $b_p = b_{p'}$, and $l_{x'} - r_y \leq \delta$ (see Figure 4.2a).

Any pair of linked-reads sharing a barcode may have originated from a molecule that is split with respect to the reference genome due to the presence of a novel adjacency in an individual genome. We define a *candidate split molecule* as a tuple $M = (L_1, D, L_2)$, where L_1 and L_2 are linked-reads and D is a set of discordant read pairs, satisfying the following conditions: (1) All read pairs in L_1, L_2, and D share the same barcode. (2) All discordant read pairs in D have the same orientations, and the distance between any two discordant read-pairs in D is at most $l_{\max}$. Formally, for read pairs $p = \langle x, y \rangle, p' = \langle x', y' \rangle \in D$, $|x - x'| < l_{\max}$ and $|y - y'| < l_{\max}$. (3) The linked reads are located within δ of the discordant read-pairs in D, in the direction consistent with the orientation of D. To indicate the last condition, we assign an orientation to each linked read in M. For example, $M = (L_1^+, D, L_2^-)$ is a candidate split molecule provided: the *rightmost* position, $r_{L_1} = \max\{r_y \mid \langle x, y \rangle \in L_1^+\}$, of L_1 is within δ of discordant read-pairs in D and the *leftmost* position, $l_{L_2} = \min\{l_x \mid \langle x, y \rangle \in L_2^-\}$, of L_2 is within δ of discordant read-pairs in D. See Figure 4.2). Note that we also define a candidate split molecule in the case where D is the empty set. In this case, a candidate split molecule can be formed for any of the four possible orientations of linked-reads L_1 and L_2: $(L_1^+, \emptyset, L_2^-), (L_1^+, \emptyset, L_2^+), (L_1^-, \emptyset, L_2^-), (L_1^-, \emptyset, L_2^+)$.

We say that a candidate split molecule M *supports* a novel adjacency provided that the distances and orientations of the linked and discordant reads in M are consistent with the novel adjacency. For example, candidate split molecule $M = (L_1^+, D, L_2^-)$ supports the novel adjacency (i^+, j^-) provided: (1) The orientation of the candidate split molecule matches the orientation of the novel adjacency. (2) Each read-pair $p = \langle x, y \rangle \in D$, is at most $l_{\max}$ from breakends i and j: $i - l_{\max} \leq r_x \leq i$ and $j \leq l_y \leq j + l_{\max}$. (3) Oriented linked-reads L_i^+ and L_j^- are at most a distance δ from positions i and j respectively: $i - \delta \leq r_{L_1} \leq i$ and $j \leq l_{L_2} \leq j + \delta$.

According to the definitions above, for a given barcode and novel adjacency (i^+, j^-), there is at most one candidate split molecule that supports this novel adjacency. We define the set $\mathscr{M}$ to be the set of all candidate split molecules supporting a novel adjacency (i^+, j^-). For ease of exposition, we will describe the model below for a novel adjacency of the form (i^+, j^-), but the model may be applied to novel adjacencies with any of the four orientations.

4.3.3 Likelihood ratio score

We use a likelihood ratio score to evaluate the evidence support a novel adjacency. Given a potential novel adjacency (i^+, j^-), let A_{i^+, j^-} be the event of a novel adjacency (i^+, j^-) in an individual genome and let event $\bar{A}_{i^+, j^-}$ to be the absence of this novel adjacency. Let $\mathscr{M}$ be the set of all candidate split molecules supporting

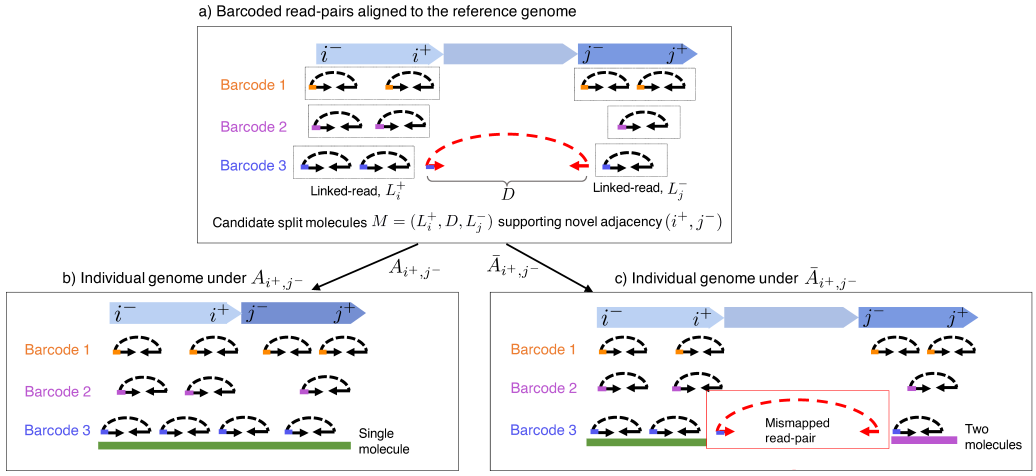


Figure 4.3: (a) A candidate split molecule for novel adjacency (i^+, j^-) consists of a linked-read L_i^+ a linked-read L_j^-, within a distance δ of position j and a set of discordant reads D. Barcoded reads aligned to the reference genome may originate from an individual genome that either contains a novel adjacency (A_{i^+,j^-}) or does not contain a novel adjacency $(\bar{A}_{i^+,j^-})$. (b) Under the alternative hypothesis A_{i^+,j^-} that i^+ and j^- are adjacent in an individual genome, reads in barcode 3 are close and are likely to have originated from a single molecule. (c) Under the null hypothesis $\bar{A}_{i^+,j^-}$ that i and j are non-adjacent in an individual genome, reads in barcode 3 are separated by a large distance and are likely to have originated from two molecules. Barcode 2 contains a read that is discordant under $\bar{A}_{i^+,j^-}$ and therefore assumed to be mismapped.

novel adjacency (i^+, j^-). We compare the likelihood $P(\mathscr{M} \mid A_{i^+,j^-})$ of A_{i^+,j^-} and the likelihood $P(\mathscr{M} \mid \bar{A}_{i^+,j^-}$ of $\bar{A}_{i^+,j^-}$ given the set of observed candidate split molecules (Figure 4.3) using the log-likelihood ratio

$$\Lambda_{i^+,j^-} = \log \frac{P(\mathscr{M} \mid A_{i^+,j^-})}{P(\mathscr{M} \mid \bar{A}_{i^+,j^-})}. \tag{4.1}$$

We report novel adjacencies with log-likelihood ratio, $\Lambda_{i^+,j^-} > c$ (selection of c is described in the Supplement) as the set of predicted novel adjacencies in the individual genome.

We now describe how we compute each of the terms in the log-likelihood ratio. First, we assume that the locations of molecules with different barcodes are independent. Combining this assumption with the requirement that candidate split molecules with different barcodes must originate from different molecules, we have that $P(\mathscr{M} \mid A_{i^+,j^-})$ and $P(\mathscr{M} \mid A_{i^+,j^-})$ are each the product of the individual probabilities of observing each candidate split molecule $M = (L_i^+, D, L_j^-) \in \mathscr{M}$. Next, a candidate split molecule M may have either originated from a single molecule that spans the interval $[i^+, j^-]$ or from two molecules, each of which does not span the interval $[i^+, j^-]$. Let E_M be the event that candidate split molecule M originates from a single molecule and let $E_{L_i^+}$ and $E_{L_j^-}$ be the events that linked-reads L_i^+ and L_j^- originate from unique molecules. Thus, the probability of observing candidate split molecule M is the probability of two disjoint events: that

M originates from one molecule (E_M), or that L_i^+ and L_j^- originate from two different molecules ($E_{L_i^+} \cap E_{L_j^-}$). Thus, the log-likelihood ratio Λ_{i^+,j^-} is calculated as follows,

$$\Lambda_{i^+,j^-} = \frac{\prod_{M \in \mathscr{M}} \left[P(E_M \mid A_{i^+,j^-}) + P(E_{L_i^+} \cap E_{L_j^-} \mid A_{i^+,j^-}) \right]}{\prod_{M \in \mathscr{M}} \left[P(E_M \mid \bar{A}_{i^+,j^-}) + P(E_{L_i^+} \cap E_{L_j^-} \mid \bar{A}_{i^+,j^-}) \right]}.$$

We now describe how we calculate the probabilities of E_M and $E_{L_i^+} \cap E_{L_j^-}$ given the the events A_{i^+,j^-} and $\bar{A}_{i^+,j^-}$. Given a candidate split molecule $M = (L_i^+, D, L_j^-)$ supporting a novel adjacency (i^+, j^-), $P(E_M \mid A_{i^+,j^-})$ is the probability that M was sequenced from a single molecule when i^+ and j^- are adjacent in the individual genome; correspondingly, $P(E_M \mid \bar{A}_{i^+,j^-})$ is the probability that M was sequenced from a single molecule when i^+ and j^- are not adjacent in the individual genome. We model the sequencing of a molecule as the generation of mapped reads through three sequential processes: (1) molecule size selection, (2) molecule sequencing, and (3) read mapping. We assume that the fragmentation of chromosomes into long molecules and the sequencing of reads from a molecule are independent processes. This assumption of independence is reasonable because sequencing occurs after molecules are sheared into short fragments.

Thus, we model the probability $P(E_M \mid A_{i^+,j^-})$ as,

$$\begin{aligned} P(E_M \mid A_{i^+,j^-}) &= P(S(M) = s) \cdot P(R(M) = \rho) \cdot P(Q_M), \\ \text{where } S(M) &= \text{the size of molecule } M, \\ R(M) &= \text{the sequencing rate of molecule } M, \text{ and} \\ Q_M &= \text{the event that all reads in } M \text{ are correctly mapped,} \end{aligned}$$

and the probability $P(E_M \mid \bar{A}_{i^+,j^-})$ as,

$$\begin{aligned} P(E_M \mid \bar{A}_{i^+,j^-}) &= P(S(M) = \bar{s}) \cdot P(R(M) = \bar{\rho}) \cdot P(Q_{M \setminus D}) \\ \text{where } Q_{M \setminus D} &= \text{the event that reads in } M \setminus D \text{ are correctly} \\ &\qquad \text{mapped and reads in } D \text{ are mismapped.} \end{aligned}$$

We model the size $S(M)$ of a molecule M by a negative binomial distribution, $P(S(M) = \cdot) = NB(\cdot)$, with parameters estimated from the collection of all aligned linked reads. We assume that the vast majority of linked-reads in the data originate from molecules that are contiguous with respect to the reference genome.

Formally, the size of a linked-read L is $r_L - l_L$, where $r_L = \max\{r_y \mid \langle x,y\rangle \in L\}$ and $l_L = \min\{l_x \mid \langle x,y\rangle \in L\}$ are the rightmost and leftmost positions, respectively, of paired reads in L. This approximation tends to slightly underestimate the true length of a molecule due to missing reads at the ends of the molecule. Under A_{i^+,j^-}, the size of candidate split molecule M is approximately $s = (i - l_{L_i}) + (r_{L_j} - j)$, the sum of the portions of the molecule aligning to the left and right of the novel adjacency. Similarly, under $\bar{A}_{i^+,j^-}$, the size of M is approximated by $\bar{s} = r_{L_j} - l_{L_i}$, the distance between the leftmost position in L_i and the rightmost position in L_j. For the case where i^+ and j^- are on different chromosomes, $P(S(M) = \bar{s}) = 0$.

We model the sequencing rate $R(M)$ of a molecule M by a Gamma distribution, $P(R(M) = \cdot) = \Gamma(\cdot)$. The negative binomial and Gamma distributions provide a good fit to the empirical distributions of molecule size and sequencing rate respectively (Figure S2), but other distributions can be used (Table S1). The sequencing rate of a candidate split molecule M under A_{i^+,j^-} is approximated by $\rho = \frac{|L_i^+ \cup D \cup L_j^-|}{s}$, the number of reads sequenced from M normalized by the size s of M. The sequencing rate of M under $\bar{A}_{i^+,j^-}$ is approximated by $\bar{\rho} = \frac{|L_i^+ \cup L_j^-|}{\bar{s}}$. Under $\bar{A}_{i^+,j^-}$, we exclude D from the set of reads sequenced from M because these reads are assumed to be mismapped.

We now define the probabilities $P(Q_M)$ and $P(Q_{M\setminus D})$ of mapping reads to the reference genome. Under A_{i^+,j^-}, all reads in M are correctly mapped, an event with probability

$$P(Q_M) \approx \prod_{\langle x,y\rangle \in L_i^+ \cup D \cup L_j^-} \left(\frac{q_x + q_y}{2}\right),$$

where q_x and q_y are the mapping probabilities of reads x and y obtained from the alignment software. We chose to approximate the mapping probability of the read-pair $p = \langle x,y\rangle$ to be the average of the mapping probabilities of each read, which we found to perform well on data. Under $\bar{A}_{i^+,j^-}$, concordant reads in L_i^+ and L_j^- are correctly mapped and discordant reads in D are mismapped, an event with probability,

$$P(Q_{M\setminus D}) \approx \prod_{\langle x,y\rangle \in D} \left(1 - \frac{q_x + q_y}{2}\right) \cdot \prod_{\langle x,y\rangle \in L_i^+ \cup D \cup L_j^-} \left(\frac{q_x + q_y}{2}\right).$$

We calculate the probability $P(E_{L_i^+} \cap E_{L_j^-}) \mid A_{i^+,j^-})$ that L_i^+ and L_j^- were sequenced from two different molecules as the product of the probabilities that L_i^+ and L_j^- were sequenced from independent molecules with corresponding sizes s_i, s_j and sequencing rates p_i, p_j, and that the reads from the two molecules were

then properly mapped to the reference genome, the latter event denoted by event Q_M. Formally,

$$P(E_{L_i^+} \cap E_{L_j^-}) \mid A_{i^+,j^-}) = \\ P(S(L_i) = s_i) \cdot P(R(L_i) = \rho_i) \cdot \\ P(S(L_j) = s_j) \cdot P(R(L_j) = \rho_j) \cdot P(Q_M),$$

where $s_k = r_{L_k} - l_{L_k}$ and $\rho_k = \frac{|L_k|}{s_k}$.

Under $\bar{A}_{i^+,j^-}$, the probabilities of molecule size selection and sequencing remain the same. However any discordant reads in D will be mismapped under $\bar{A}_{i^+,j^-}$, resulting in the last term being the probability $P(Q_{M \setminus D})$.

4.3.4 Incorporating haplotype phase

10X Genomics provides phasing as part of its alignment pipeline, using linked-reads to phase SNPs into large phase blocks. Each position i in the reference genome is assigned a phase block m_i and any SNP within phase block m_i will be assigned to either haplotype 1 or haplotype 2. [105] report that the current software can phase $> 95\%$ of SNPs to a phase blocks of size > 0.5Mb, with an average error rate of 0.03%. Let $m_x \in \mathbb{N}$ denote the phase block of a read x, aligned to the reference genome, and let $h_x \in \{1,2\}$ denote the haplotype of read x. We define $\mathscr{M}^{\alpha,\beta} \subseteq \mathscr{M}$ be the subset of candidate split molecules $M = (L_i^+, D, L_j^-)$ such that the haplotype h_{L_i} of linked-read L_i is α, the haplotype h_{L_j} of linked-read L_j is β, and for all read pairs $\langle x, y \rangle$ in D, $h_x = \alpha$ and $h_y = \beta$. On the linked-read datasets described in the Results below, we found that $> 99\%$ of linked-reads ($\delta = 10$Kbp) contain read-pairs that are assigned to a unique phase block and haplotype. We omit the small number of linked-reads containing read-pairs from multiple haplotypes or phase blocks from further analysis.

We define a haplotype-specific log-likelihood ratio,

$$\Lambda^{\alpha,\beta}_{i^+,j^-} = \log \frac{P(\mathscr{M}^{\alpha,\beta} \mid A_{i^+,j^-})}{P(\mathscr{M}^{\alpha,\beta} \mid \bar{A}_{i^+,j^-})}. \tag{4.2}$$

Each novel adjacency in a diploid genome is the result a heterozygous structural variant on haplotype 1, a heterozygous structural variant on haplotype 2, or a homozygous structural variant affecting both haplotypes. If positions i and j are on the same phase block $m_i = m_j$, then the haplotypes of i and j must match. Thus, we calculate the log-likelihood ratio $\Lambda^h_{i^+,j^-}$ to be the maximum of $\Lambda^{1,1}_{i^+,j^-}$, $\Lambda^{2,2}_{i^+,j^-}$, and Λ_{i^+,j^-}, corresponding to

a novel adjacency resulting from a heterozygous structural variant on haplotype 1, a heterozygous structural variant on haplotype 2, or a homozygous structural variant,

$$\Lambda^{h}_{i^+,j^-} = \max(\Lambda^{1,1}_{i^+,j^-}, \Lambda^{2,2}_{i^+,j^-}, \Lambda_{i^+,j^-}). \tag{4.3}$$

However, if positions i and j are not on the same phase block, $m_i \neq m_j$, then the haplotypes of i and j may not match. For example, SNPs assigned to haplotype 1 of phase block m_i may originate from the same chromosome as SNPs assigned haplotype 2 of phase block m_j. In this case, we calculate the log-likelihood ratio $\Lambda^{h}_{i^+,j^-}$ as,

$$\Lambda^{h}_{i^+,j^-} = \max(\Lambda^{1,1}_{i^+,j^-}, \Lambda^{2,2}_{i^+,j^-}, \Lambda^{1,2}_{i^+,j^-}, \Lambda^{2,1}_{i^+,j^-}\Lambda_{i^+,j^-}), \tag{4.4}$$

which accounts for the ambiguity in haplotype assignment. NAIBR reports both the phased log-likelihood ratio $\Lambda^{h}_{i^+,j^-}$ as well as the inferred haplotype for each novel adjacency.

4.4 Simulating structural variants

We run each method using its default parameters. For NAIBR we use the the phased log-likelihood ratio $\lambda^{h}_{i^+,j^-}$ described in section 2.4 and set $\delta = 10$Kbp. This value of δ is selected to be the 95^{th} percentile of distances between read pairs within each barcode, calculated using reads from three datasets, NA12878, HCC1954T, and HCC1954N, sequenced on the 10X Chromium platform. The 95^{th} percentile is chosen because it results in few read-pairs originating from the same molecule being erroneously separated into two linked-reads.

Some methods output a pair of intervals for each novel adjacency, allowing for some uncertainty in the location of the novel adjacency, while others output a precise pair of coordinates. To account for this difference, we take the midpoint of each reported window and define each called novel adjacency to be a pair of 500bp windows centered on their respective midpoints. A call is considered correct if the left coordinate of the true novel adjacency overlaps the left window of the called novel adjacency and the right coordinate of the true novel adjacency overlaps the right window of the called novel adjacency.

We use chromosomes 17 and 18 from human reference genome hg19 as our reference for the purpose of simulation. We simulate several types of structural variants in a range of sizes. To assess NAIBR's ability to detect novel adjacencies that occur on a single haplotype, we simulate two test genomes, one that contains 400 homozygous structural variants across the genome and one that contains 400 different structural

variants on each haplotype. We also add SNPs with a uniform 1% mutation rate across the genome to simulate genetic variance between an individual and the reference. The 400 structural variants consist of 200 deletions, 100 insertions (including duplications, and translocations), and 100 inversions between 50Bp and 100Kbp. Structural variants are simulated using the R package, *RSVSim* with default parameters for the size distribution of each structural variant type. RSVSim estimates these default size distributions for each structural variant type by fitting a beta distribution to structural variant sizes obtained from the Database of Genomic Variants [77]. Figure 4.4 shows the distribution of structural variants sizes for 400 simulated structural variants. Because structural variants tend to co-occur with other, smaller mutations, such as small indels or SNPs, we randomly generate additional SNPs and indels within 50bp of the novel adjacencies introduced by each structural variant, with a probability $p = 0.25$ of generating a SNP and $p = 0.5$ of generating a small indel.

The structural variants for each haplotype of the two test genomes are non-overlapping and cumulatively affect 4.6% of the haplotype. Each structural variant creates between 1 and 2 novel adjacencies in the data: deletions create 1 novel adjacency while inversions and insertions create 2 novel adjacencies.

Sequencing of chromosomes 17 and 18 is simulated using LRSIM [81], which simulates reads generated with 10X Genomics' linked-read technology. 30X coverage 100bp paired-end read sequencing data was generated with an average molecule size of 85Kb (consistent with the current state of the technology), a per base error rate of 0.1%, and a mean insert size of 340bp. Molecules were assigned barcodes such that on average each barcode is assigned to 4 molecules. Though the current technology assigns the same barcode to approximately 10 molecules, because the test genome is an order of magnitude smaller than the human genome, this number must be reduced to achieve a similar average distance between molecules with the same barcode. Reads are mapped to chromosomes 17 and 18 and assigned to haplotype blocks using Long Ranger [105]. Mapped reads are used as input to each structural variant detection method.

To assess NAIBR's ability to detect heterozygous novel adjacencies, we simulated a genome that contains 400 different structural variants – including duplications, deletions, translocations, and inversions – on each haplotype (800 in total), creating a total of 1027 novel adjacencies, 734 novel adjacencies larger than 10Kbp, and 545 novel adjacencies larger than 30Kbp. This dataset, sequenced at 30X total coverage, has 15X coverage of each haplotype, resulting in fewer discordant read-pairs crossing each novel adjacency than in the homozygous dataset. NAIBR identifies novel adjacencies from the heterozygous dataset with nearly equal precision and recall as the homozygous dataset, correctly identifying $967/1027$ novel adjacencies (Fig. 4.5b). The introduction of heterozygous variants results in a significant decrease in precision and/or recall for all other methods except for GASVPro. This could be due to GASVPro's use of *breakend read depth*,

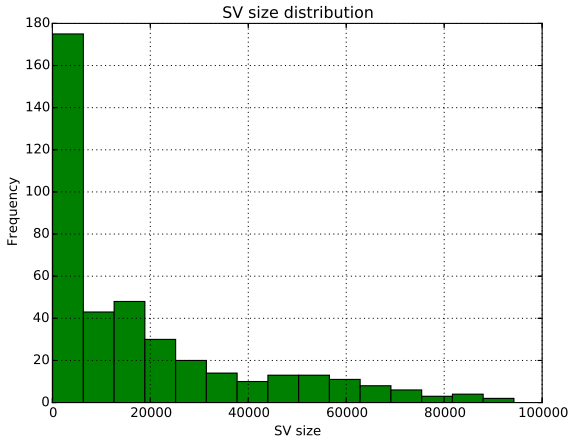


Figure 4.4: Size distribution of 400 simulated structural variants (deletions, insertions, and inversions).

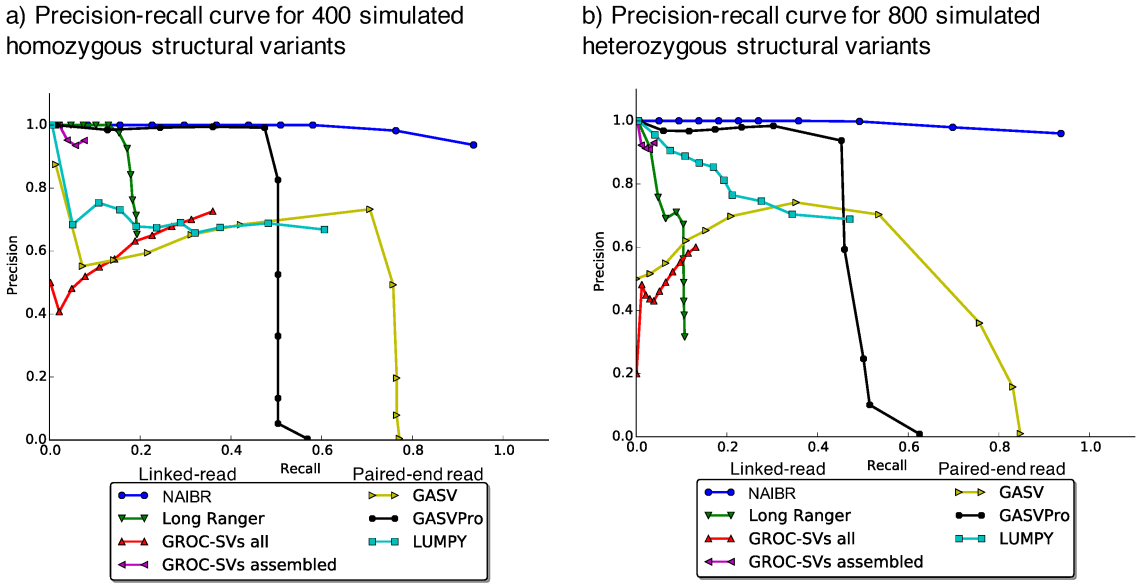


Figure 4.5: a) Precision-recall curve for NAIBR, Long Ranger, GROC-SVs, GASV, GASVPro, and LUMPY on 30X simulated data from chromosomes 17 and 18, containing 400 homozygous structural variants. b) Precision-recall curve for NAIBR, Long Ranger, GROC-SVs, GASV, GASVPro, and LUMPY on 30X simulated data from chromosomes 17 and 18, containing 800 heterozygous structural variants.

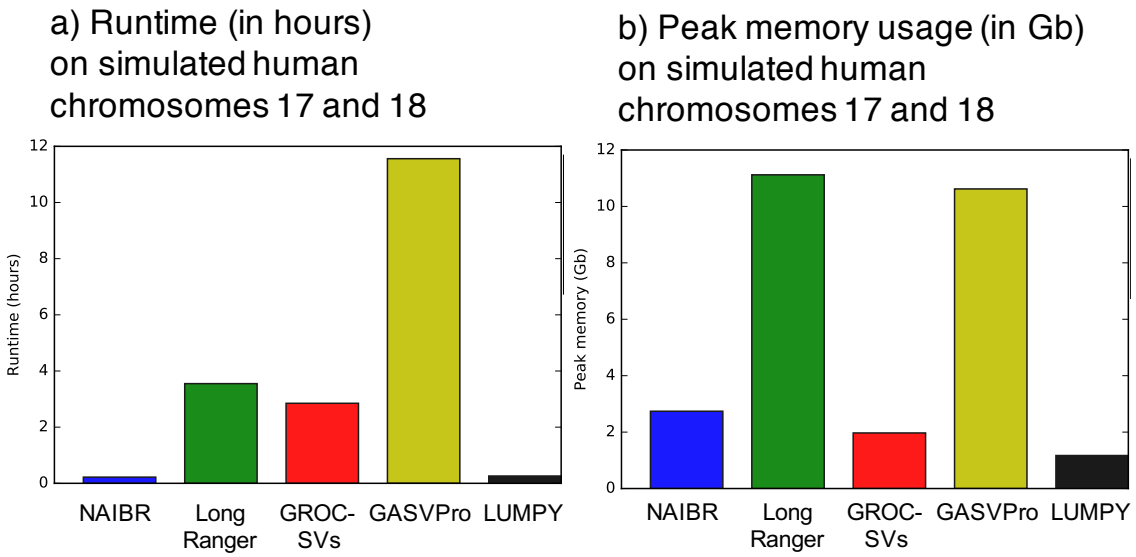


Figure 4.6: a) Runtime of NAIBR, Long Ranger, GROC-SVs, GASVPro, and LUMPY on simulated human chromosomes 17 and 18. NAIBR and LUMPY both ran in about 15 minutes, while Long Ranger and GROC-SVs took 3.55 and 2.85 hours to complete respectively. GASVPro had the longest runtime at 11.56 hours. b) Peak memory usage for NAIBR, Long Ranger, GROC-SVs, GASVPro, and LUMPY. NAIBR requires a similar amount of memory as GROC-SVs and LUMPY and significantly less memory than both Long Ranger and GASVPro.

a signal of a lower concordant read depth surrounding novel adjacency breakends. Breakend read depth is present for both balanced and unbalanced rearrangements and provides additional signal in the absence of discordant reads.

4.4.1 Runtime analysis

We compared runtimes and memory usage of NAIBR, Long Ranger, GROC-SVs, GASVPro, and LUMPY on the 400 homozygous variants simulated data described above. We ran each method on 1 core of a 2.6Ghz 512Gb machine. Note that Long Ranger's structural variant calling algorithm is incorporated as part of its phasing pipeline, so runtime and peak memory consumption include both phasing and structural variant calling. Figure 4.6 shows that NAIBR is more than 10 times faster than Long Ranger and GROC-SVs, which both utilize linked-reads, and consumes less than half the amount of memory as Long Ranger. NAIBR also performs well against paired-end methods, achieving a slightly lower running time than LUMPY and a significantly lower running time than GASVPro while requiring half the amount of memory as GASVPro.

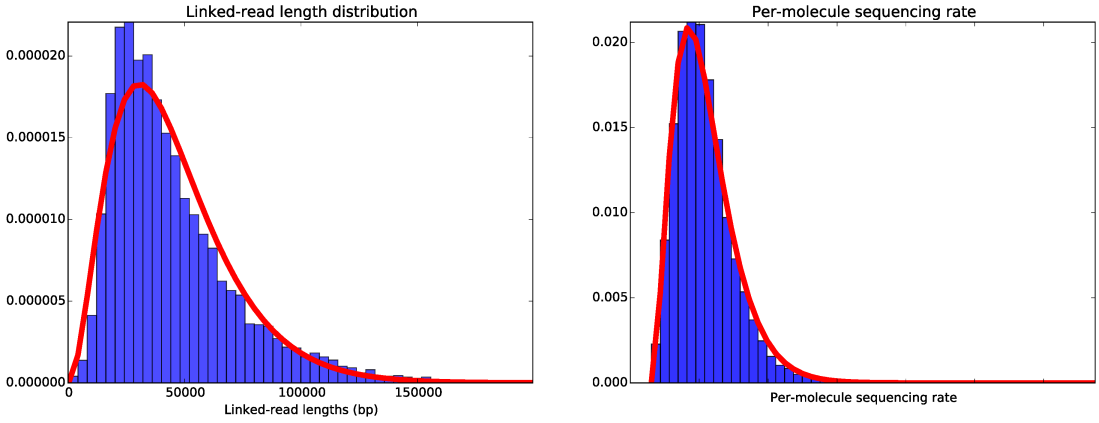


Figure 4.7: (left) The negative binomial distribution (red) fit to the empirical linked-read length distribution (blue) from 35X linked-read sequencing data from individual NA12878 of the 1000 genomes project. (right) The gamma distribution (red) fit to the empirical distribution of sequencing rate per molecule (blue).

a) Simulated chromosomes 17 and 18 with 400 homozygous structural variants

		Sequencing rate per molecule			
		Gamma		Normal	
		Precision	Recall	Precision	Recall
Molecule size	Negative binomial	**98.3%**	**94.2%**	97.5%	92.9%
	Poisson	97.5%	92.4%	98.0%	91.9%

b) Simulated chromosomes 17 and 18 with 800 heterozygous structural variants

		Sequencing rate per molecule			
		Gamma		Normal	
		Precision	Recall	Precision	Recall
Molecule size	Negative binomial	**96.9%**	**94.2%**	95.7%	91.4%
	Poisson	98.0%	91.2%	94.3%	93.4%

Table 4.1: (a) Precision and recall for 400 simulated structural variants on human chromosomes 17 and 18. (b) Precision and recall for 800 simulated structural variants on human chromosomes 17 and 18.

4.5 Modeling empirical distributions

We model molecule size by a negative binomial distribution with parameters (p,r) and model the sequencing rate per molecule by a Gamma distribution with parameters (α,β). These distributions provide a close fit to the empirical distributions (see Figure S2). We examine the importance of distribution choice by comparing the NAIBR's performance using either a negative binomial or a Poisson distribution to model molecule size and either a Gamma or Normal distribution to model the sequencing rate per molecule.

We ran NAIBR on two simulated datasets containing chromosomes 17 and 18 with 400 homozygous structural variants and 800 heterozygous structural variants respectively. We report the precision and recall of both datasets using all combinations of distributions to model molecule size and sequencing rate per

molecule. On both datasets, the combination of negative binomial to model molecule size and Gamma to model sequencing rate per molecule performs slightly better than the other distributions, with precision and recall of 98.3% and 94.2% respectively for the dataset with homozygous variants (Table 4.1a) and precision and recall of 96.9% and 94.2% respectively for the dataset with heterozygous variants (Table 4.1b). The difference in performance between the different distributions is quite small, only accounting for at most a 2.6% difference in precision and at most a 3% difference in recall.

4.6 Determining a value for Λ_{i^+,j^-}

As the number of split molecules spanning a novel adjacency increases, the value of the log-likelihood ratio Λ_{i^+,j^-} will increase proportionally. This means that the value of Λ_{i^+,j^-} is proportional to the coverage of the sequencing data. We estimate an appropriate cutoff parameter c for Λ_{i^+,j^-} by running NAIBR on simulated data. We define c such that the set of novel adjacencies with $\Lambda_{i^+,j^-} \geq c$ can recall $\geq 90\%$ of simulated novel adjacencies.

We simulated data containing 1027 heterozygous novel adjacencies (as described in Section 4.4) at seven levels of coverage: 60X, 50X, 30X, 30X, 20X, 10X, and 5X. We then ran NAIBR on each of the datasets. NAIBR was able to recall at least 90% of the novel adjacencies at all levels of coverage except for 5X coverage, where it could only recall 84% of the novel adjacencies. Figure 4.8 shows that, as expected, c increases linearly as the coverage increases. The equation of the best-fit line $y = 6.943 * x - 37.33$ is used to determine c. NAIBR automatically determines the cutoff parameter c based on the coverage of the data and labels each prediction with score $\Lambda_{i^+,j^-} \geq c$.

4.7 Benchmarking HCC1954

Three studies – [69] [94], and [74] – have sequenced the tumor cell line HCC1954 and reported sets of predicted structural variants. Bignell et al. used BAC sequencing to identify amplifications in the tumor cell line HCC1954. 21 significantly amplified BACs were shotgun-sequenced and aligned to the human genome, identifying 69 novel adjacencies. These novel adjacencies were confirmed as somatic events by PCR of both the tumor and matched normal DNA. Stephens et al. performed paired-end sequencing of 24 cancer genomes, including HCC1954. The study identified 1832 novel adjacencies based on discordantly mapped reads and confirmed 246 of these to be somatic through PCR sequencing of the tumor and matched normal. Galante

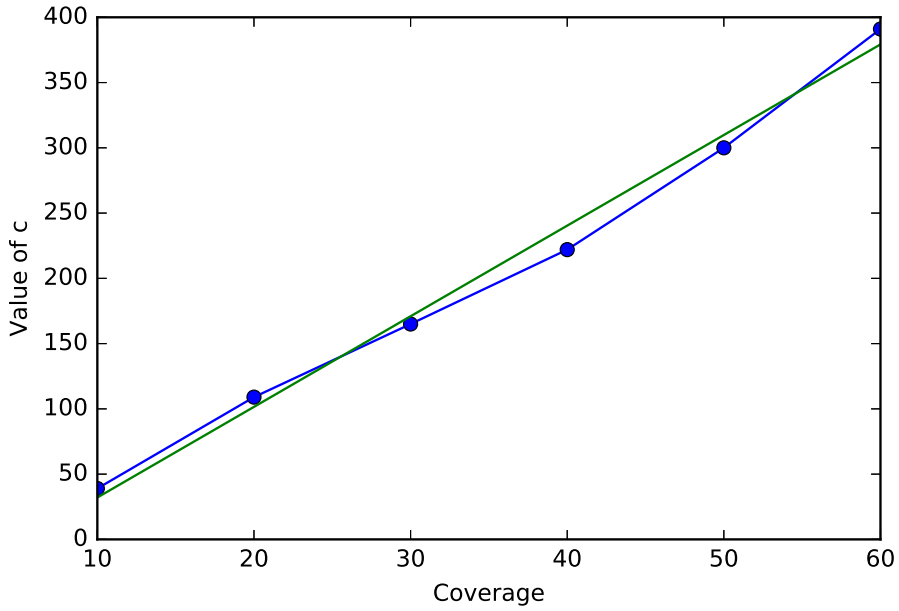


Figure 4.8: c is the maximal cutoff value such that 90% recall is achieved. c increases linearly as the coverage increases. The green line is the best-fit line to the data.

et al. performed whole paired-end sequencing of HCC1954 and identified 89 somatic novel adjacencies validated by PCR sequencing of the tumor and matched normal.

Though the genomic rearrangements reported by each study were experimentally validated, relatively few were reported by more than one study and only 1 variant was reported by all three studies (Figure 4.9). This indicates that experimental design plays a large role in the dictating which novel adjacencies are identified and that there are still potentially many more that have yet to be identified.

We combined the three sets of validated novel adjacencies into a single set, combining overlapping novel adjacencies. A novel adjacency defined by breakends (a,b) overlaps a novel adjacency defined by breakends (c,d) if $|c-a| \leq 500$bp and $|d-b| \leq 500$bp. The novel adjacencies are combined by defining each breakend as the interval $[\min(a,c), \max(a,c)]$ and $[\min(b,d), \max(b,d)]$ respectively. Each study reports novel adjacencies created by four types of structural variation events: deletions, inversions, duplications, and interchromsomal events. Deletions and classified as having concordantly oriented read-pairs (+-) and a decrease in copy number. Duplications are classified as having -+ or +- orientation and an increase in copy number. Inversion are classified as having – or ++ orientation. Interchromosomal events are classified as having novel adjacency breakends on different chromosomes.

We ran NAIBR, Long Ranger, GASV, and GASVPro on 35X coverage sequencing data for tumor cell

Figure 4.9: Venn diagram of novel adjacencies reported by Bignell et al., Stephens et al., and Galante et al..

line HCC1954T and normal cell line HCC1954N provided by 10X Genomics. Novel adjacencies reported by each method that appear in HCC1954T and not HCC1954N are considered to be *somatic*. We will consider novel adjacencies separated by a distance at least 30Kb, as the vast majority (92%) of the PCR validated novel adjacencies were at least 30Kb in size and Long Ranger only reports novel adjacencies $\geq$ 30Kb.

NAIBR reported 549 somatic novel adjacencies with scores $>$ 205. The score cutoff is determined using the equation derived in section 4.8. Long Ranger reported 555 somatic novel adjacencies passing quality thresholds and labeled as 'CALLS'. GASV reported 13920 somatic novel adjacencies with at least 4 supporting discordant reads. LUMPY reported 1342 somatic novel adjacencies with at least 2 split reads.

Figure 4.12 shows the number of PCR validated novel adjacencies created by deletions, duplications, inversions, or interchromosomal events predicted by each method. NAIBR identifies significantly more novel adjacencies created by duplication events than the other methods and significantly more novel adjacencies created by interchromosomal events than Long Ranger. Bignell et al. and Stephens et al. note that tandem duplications are the most common type of structural variant observed in breast cancer tumors, followed by interchromosomal events. Both suggest that a defect in DNA maintenance may generate this particular class duplication events. NAIBR's ability to detect more novel adjacencies introduced by duplications and interchromosomal events than other methods without introducing many potential false positive predictions

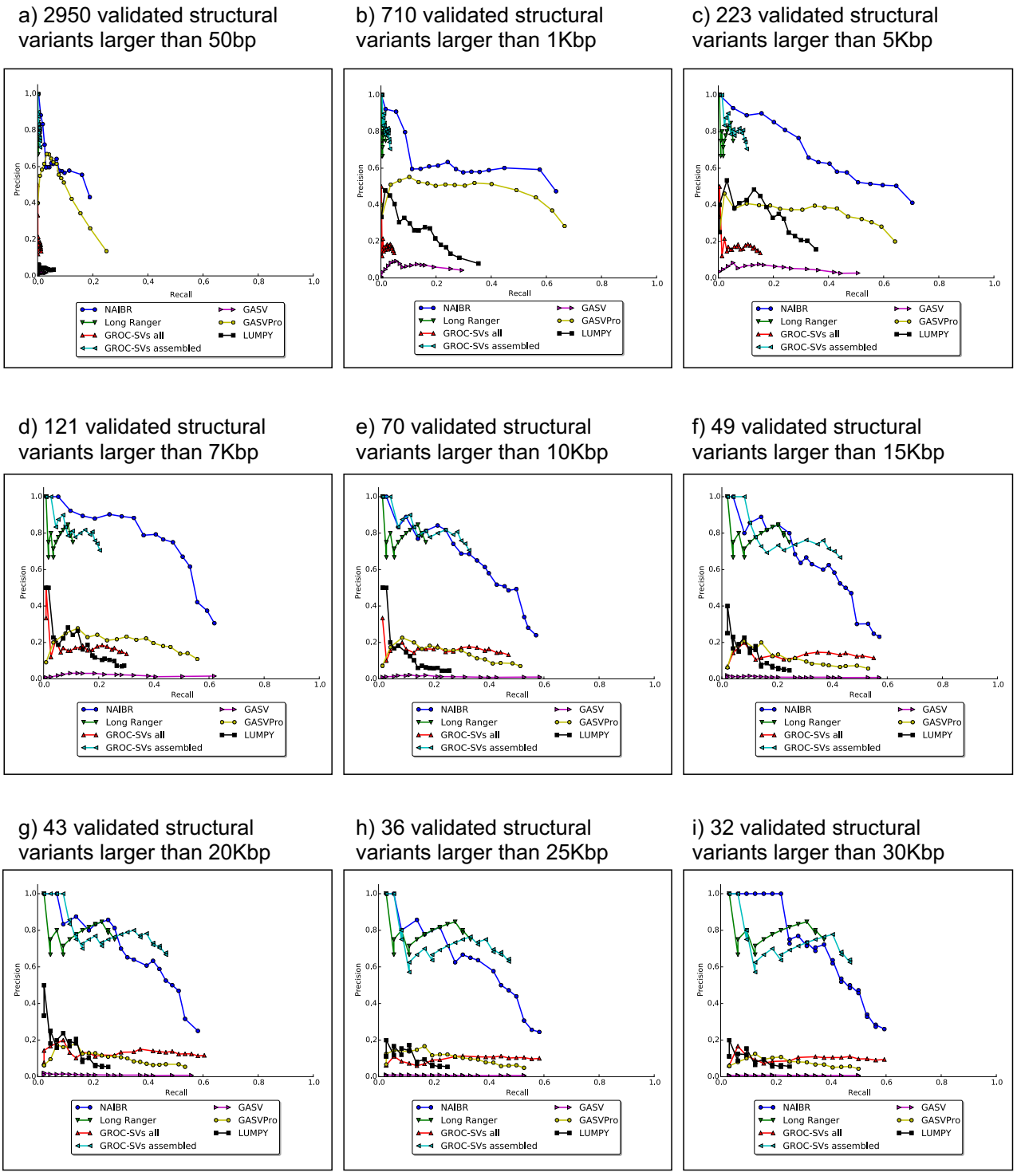


Figure 4.10: Precision-recall curves (a-i) for human cell line NA12878 with 2950 validated structural variants of different sizes. All methods were run on 35X linked-read sequencing data for cell line NA12878 provided by 10X Genomics. (f-i) For large structural variants $\geq$ 15Kb, NAIBR (dark blue) performs similarly to other linked-read structural variant detection methods, Long Ranger and GROC-SVs. GROC-SVs (light blue) performs with slightly higher precision due to its use of local assembly to verify predicted variants. (b-e) For mid-range structural variants $\geq$ 1Kbp, NAIBR demonstrates significant improvement over other methods. NAIBR predicts more true variants than linked-read methods Long Ranger and GROC-SVs and performs with higher precision than paired-end read methods GASV, GASVPro, and LUMPY. (a) NAIBR was designed to identify structural variants significantly larger than the insert size of a concordant paired-end read, which ranges between 250bp and 850bp in this dataset. On small structural variants, linked-reads

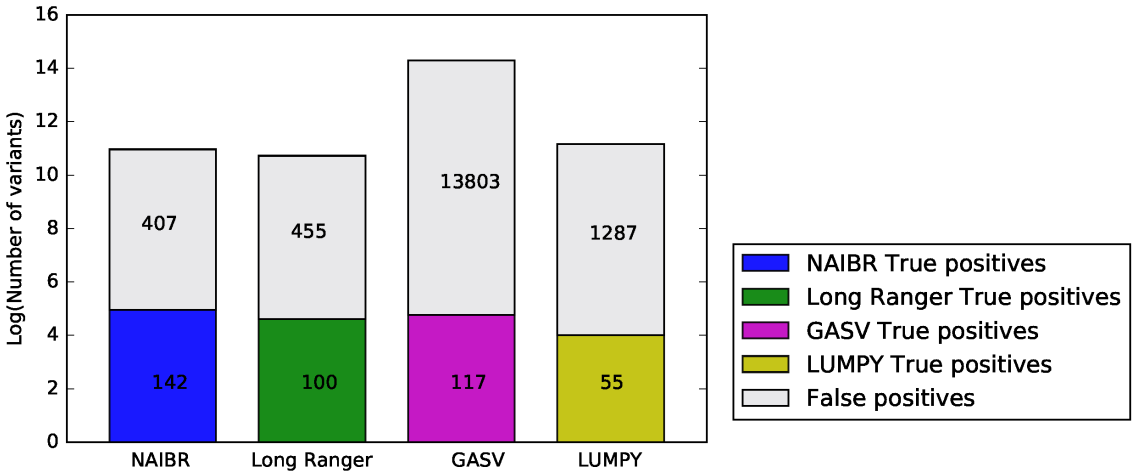


Figure 4.11: Precision of PCR validated novel adjacencies reported by Bignell et al., Stephens et al., and Galante et al. for structural variant calling methods: NAIBR, Long Ranger, GASV, and GASVPro. Colored bars represent true positive events and grey bars represent false positive events. NAIBR reports the highest number of true positives and reports fewer false positives than Long Ranger and GASV.

may allow it to recover more clinically relevant novel adjacencies.

NAIBR recalls significantly more novel adjacencies than other methods (142 recalled by NAIBR compared to 117 recalled by GASV, 100 recalled by Long Ranger, and 55 recalled by LUMPY). Figure 4.11 shows total number of true positive and false positive predictions made by each method (reported on the log scale). GASV reported significantly more predictions than the other methods. NAIBR and Long Ranger made a similar number of total predictions, but NAIBR recalled more PCR validated novel adjacencies than Long Ranger. We plot the number of PCR validated variants predicted by each method at different levels of recall in Figure 4c. To obtain different levels of recall, the results from NAIBR and Long Ranger are sorted by log-likelihood score and the predictions by GASV are sorted by number of supporting discordant reads. Figure 4c shows that at all levels of recall, NAIBR performs with higher precision than Long Ranger, GASV, and LUMPY.

We compare the predictions made by NAIBR to those made LUMPY, which uses only paired-end reads. NAIBR and LUMPY both identified 42 novel adjacencies from the PCR-validated set. NAIBR additionally identified 100 PCR-validated novel adjacencies not identified by LUMPY while LUMPY identified 13 PCR-validated novel adjacencies not identified by NAIBR (Figure 4.15). NAIBR utilizes signals from both discordant read-pairs as well as candidate split molecules, while LUMPY relies entirely on discordant read-pairs and split reads. Figure 4.16a shows the distribution of the number of discordant read-pairs and candidate split molecules in NAIBR predictions, but not predicted by LUMPY or present in the set of validated novel

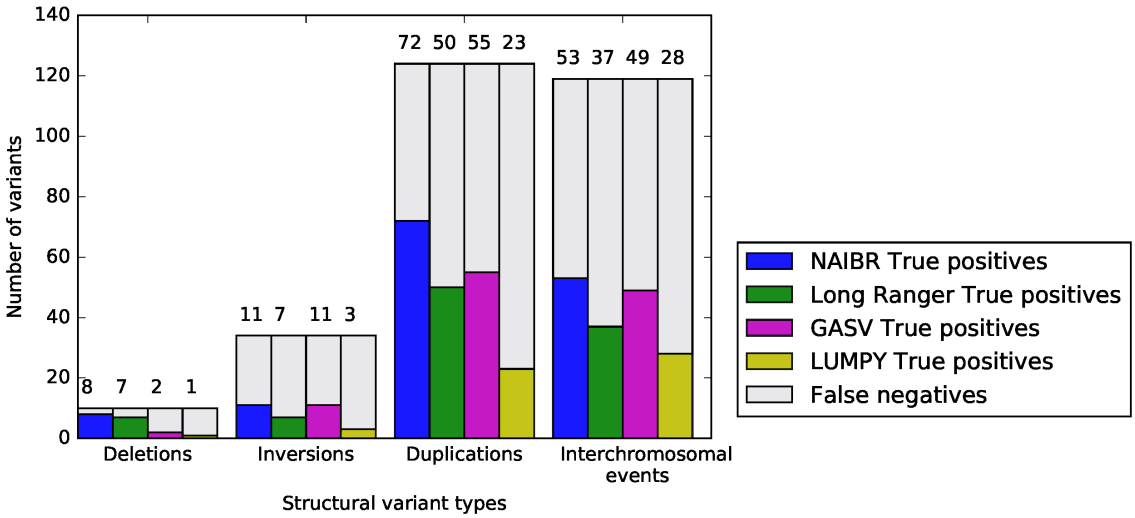


Figure 4.12: Recall of PCR validated novel adjacencies reported by Bignell et al., Stephens et al., and Galante et al. for structural variant calling methods: NAIBR, Long Ranger, GASV, and GASVPro. Colored bars represent true positive events and grey bars represent false negative events.

adjacencies. This set of predictions contains on average the lowest number (3.25) of discordant read-pairs, with 24% of predictions containing 0 discordant read-pairs. The mean numbers of candidate split-molecules (116) and discordant pairs (10.0) present in predictions shared by NAIBR and LUMPY (Figure 4.16b) are significantly larger ($P = 2.5 \cdot 10^{-4}, P = 2.38 \cdot 10^{-21}$) than the corresponding numbers (46.0 and 3.25) in Figure 4.16a. Figure 4.16d shows the distribution of the number of discordant read-pairs and candidate split molecules predicted by both NAIBR and LUMPY and also present in the set of validated novel adjacencies. Most predictions in this set contain multiple discordant read-pairs. The mean number of discordant pairs (12.5) in Figure 4.16d that overlap PCR-validated novel adjacencies is significantly larger ($P = 1.44 \cdot 10^{-7}$) than the corresponding number (10.0) in Figure 4.16b.

We also explored how NAIBR's predictions varied with sequence coverage. Figure 4.14 shows precision recall curves for NAIBR on the HCC1954 breast cancer cell line at coverage: 35X, 15X, and 10X. Not surprisingly, the total recall increases with increasing coverage. Precision remains approximately the same across different coverages, with the exception of a slight decrease in precision for recall $> 20\%$ in the 10X coverage dataset. In Figure 4.13 we see that the total number of predictions as well as those matching validated novel adjacencies decreases as coverage decreases. While the average number of discordant read-pairs decreases by nearly 50% when coverages drops from 35X to 15X, the number of candidate split molecules decreases by only 15%. This results in a decrease from 142 predictions matching validated novel adjacencies

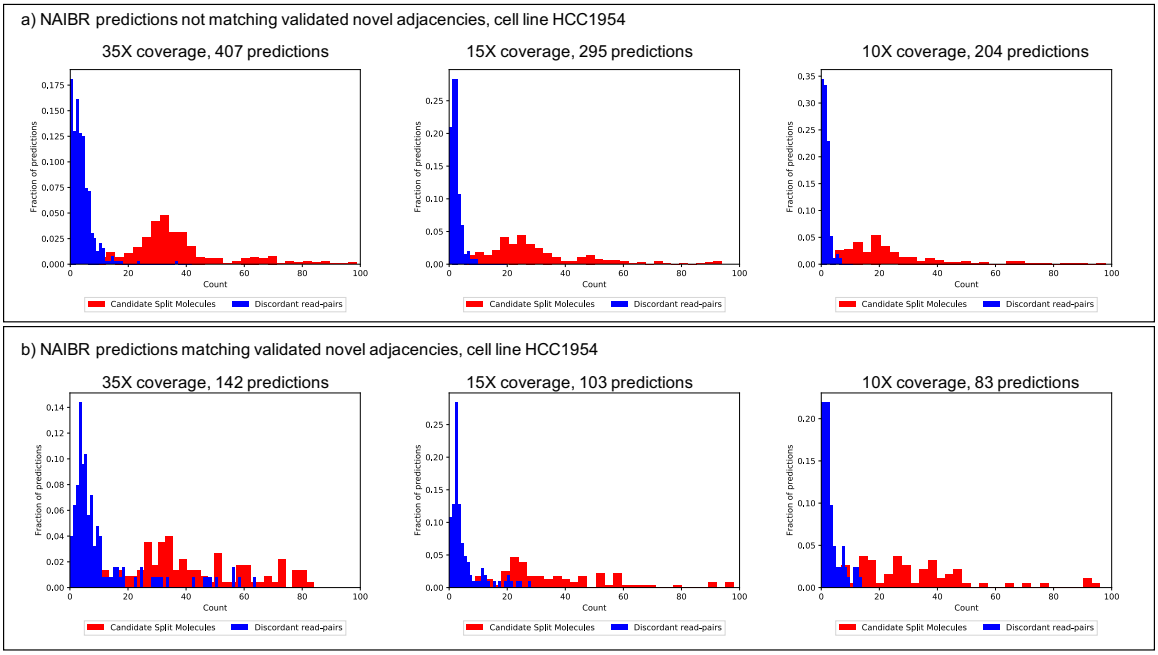


Figure 4.13: Distributions of discordant read pairs and candidate split molecules supporting NAIBR predictions, at 35X coverage, 15X coverage, and 10X coverage on the HCC1954 breast cancer cell line. (a) Distribution for NAIBR predictions not matching validated novel adjacencies evaluated on 35X, 15X, and 10X coverage datasets. NAIBR predicted 407 novel adjacencies at 35X coverage, 295 at 15X coverage, and 204 at 10X coverage. At 35X coverage the mean number of discordant pairs (3.58) is significantly larger ($P = 6.79 \cdot 10^{-12}$) than at 15X coverage (1.87) while the number of candidate split molecules (50.6) is not significantly larger ($P = 0.056$) than at 15X coverage (40.6). (b) Distribution for NAIBR predictions matching validated novel adjacencies evaluated on 35X, 15X, and 10X coverage datasets. NAIBR predicted 142 novel adjacencies at 35X coverage, 103 at 15X coverage, and 83 at 10X coverage. At 35X coverage the mean number of candidate split-molecules (118.9) and discordant pairs (10.8) is significantly larger ($P = 1.64 \cdot 10^{-5}, P = 2.95 \cdot 10^{-22}$) than the corresponding numbers (50.6 and 3.58) in (a). These differences are also significant at 15X ($P = 5.70 \cdot 10^{-5}, P = 1.51 \cdot 10^{-18}$) and 10X coverage ($P = 1.17 \cdot 10^{-4}, P = 3.60 \cdot 10^{-17}$).

in the 35X dataset to 103 predictions matching novel adjacencies in the 15X dataset, a 28% reduction in the number of validated predictions.

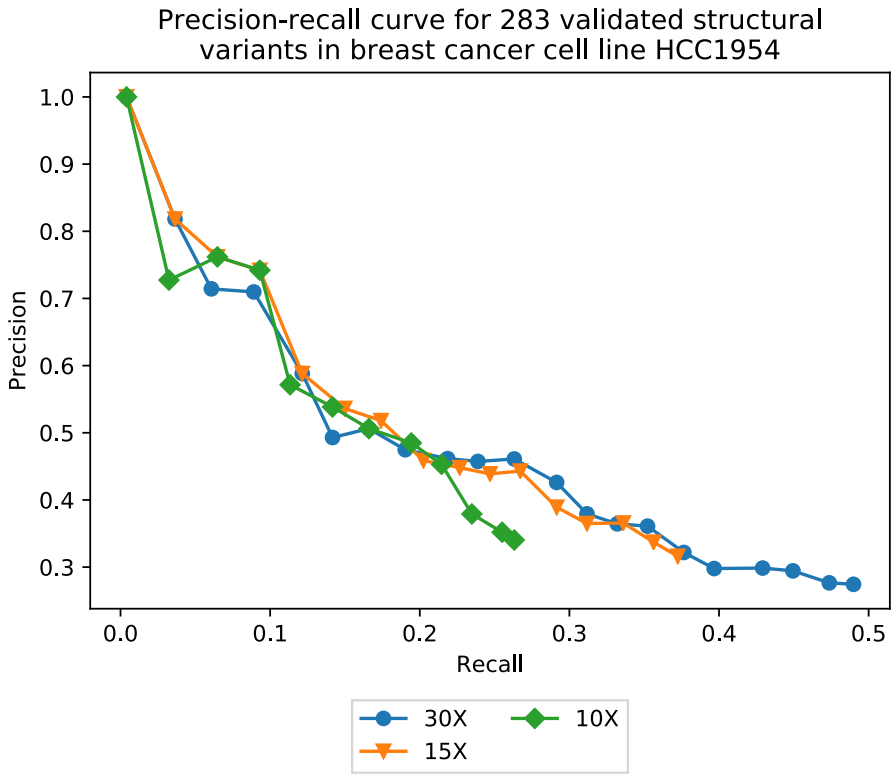


Figure 4.14: Precision-recall curve for 283 validated structural variants in breast cancer cell line HCC1954 predicted by NAIBR at three levels of coverage: 35X, 15X, and 10X. A total of 142 variants were predicted at 35X coverage, 103 at 15X coverage, and 69 at 10X coverage. The total recall increases with increasing coverage, with precision remaining approximately the same across different coverage, with the exception of a slight decrease in precision for recall $>$ 20% in the 10X coverage dataset.

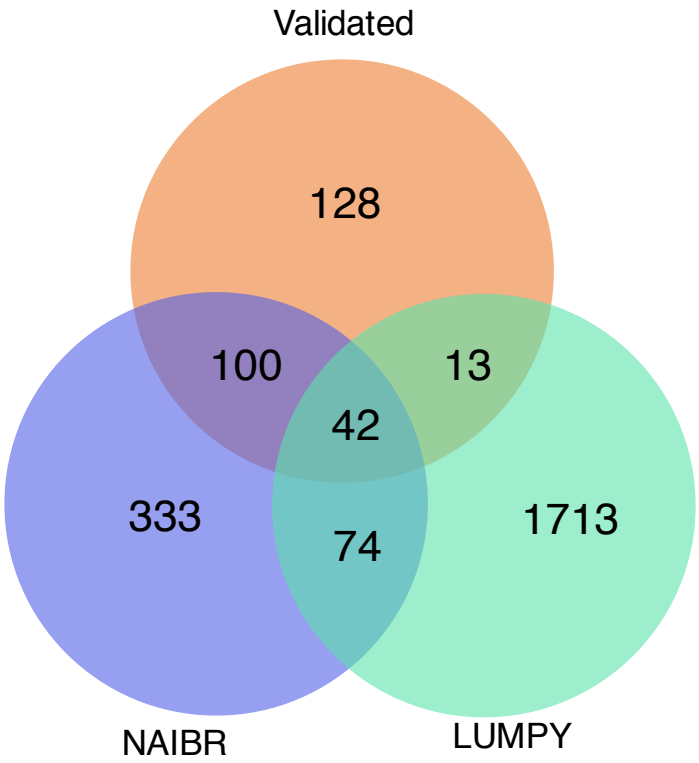


Figure 4.15: Venn diagram comparing NAIBR predictions, LUMPY predictions, and PCR-validated novel adjacencies on the HCC1954 breast cancer cell line.

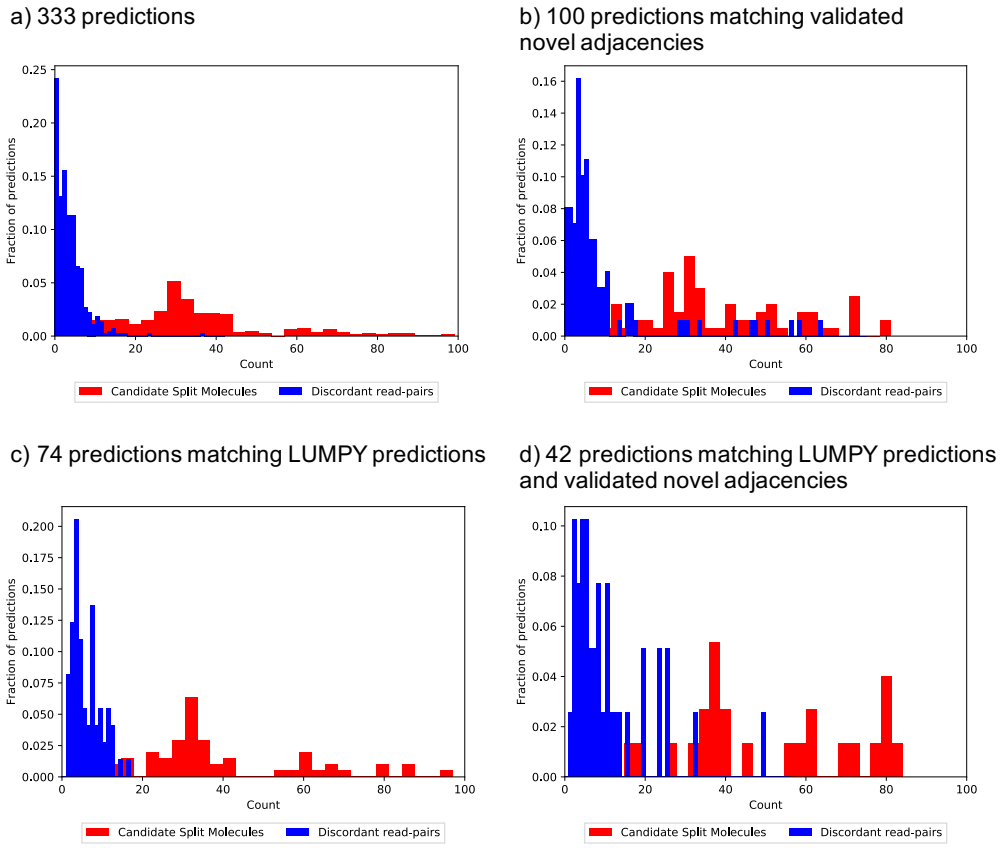


Figure 4.16: Distributions of discordant read pairs and candidate split molecules supporting NAIBR predictions, LUMPY predictions, and PCR-validated novel adjacencies on the HCC1954 breast cancer cell line. (a) Distribution for 333 NAIBR-unique predictions not matching validated novel adjacencies. (b) Distribution for 100 NAIBR-unique predictions matching validated novel adjacencies. The mean numbers of candidate split-molecules (116) and discordant pairs (10.0) are significantly larger ($P = 2.5 \cdot 10^{-4}, P = 2.38 \cdot 10^{-21}$) than the corresponding numbers (46.0 and 3.25) in (a). (c) Distribution for 74 predictions shared by NAIBR and LUMPY, but not matching validated novel adjacencies. (d) Distribution for 42 predictions shared by NAIBR and LUMPY that match validated novel adjacencies. The mean number of discordant pairs (12.5) is significantly larger ($P = 1.44 \cdot 10^{-7}$) than the corresponding number (10.0) in (b).

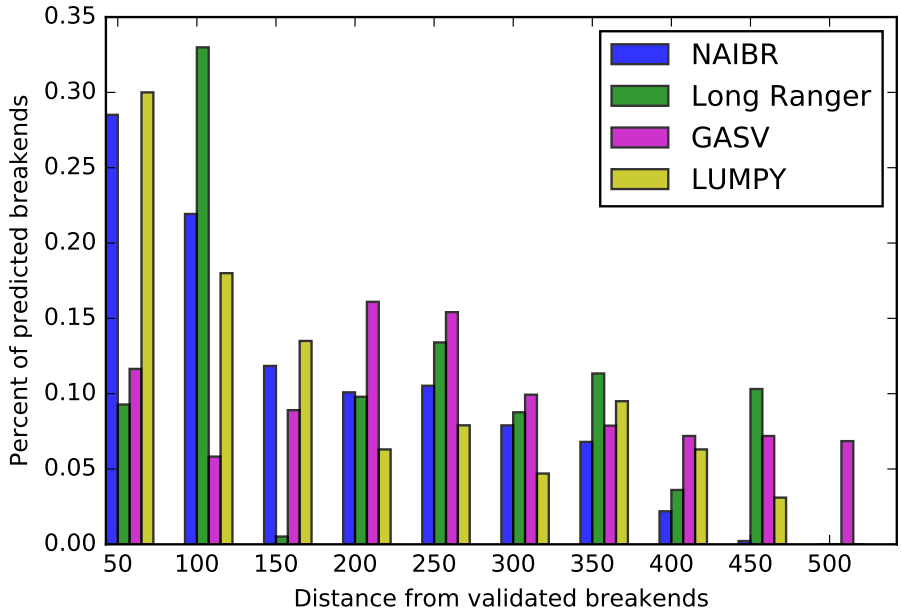


Figure 4.17: Percentage of validated novel adjacencies within a distance, *x*, from the true breakends. Distance is measured as the absolute value of the distance between the true breakends and the breakends predicted by each structural variant caller.

For the HCC1954 cancer cell line we compared the specificity of breakends predicted by NAIBR to those predicted by Long Ranger, GASV, and LUMPY. For the set of true positive predictions made by NAIBR, Long Ranger, GASV, and LUMPY we plotted the percentage of breakends within distances ranging from 0-500bp from the breakends of the PCR validated novel adjacencies (Figure 4.17). 30% of the breakends predicted by NAIBR fall within 50bp of the PCR-validated breakends, while only 9% and 12% of the breakends predicted by Long Ranger and GASV respectively fall within 50bp of the PCR-validated breakends.

Finally, we assess the distance between breakends predicted by NAIBR to breakends from validated novel adjacencies from human cell lines NA12878 and HCC1954 as well as simulated novel adjacencies. A predicted novel adjacency is determined to match a validated novel adjacency if both breakends of a predicted novel adjacency lie within 1000bp of breakends from a validated novel adjacency. Figure 4.18 shows the percentage of predicted breakends that lie within distances ranging from 0-700bp from validated breakends. We find that for each dataset, the vast majority of breakends lie within 500bp of validated breakends and that over half of the breakends lie within 100bp of validated breakends.

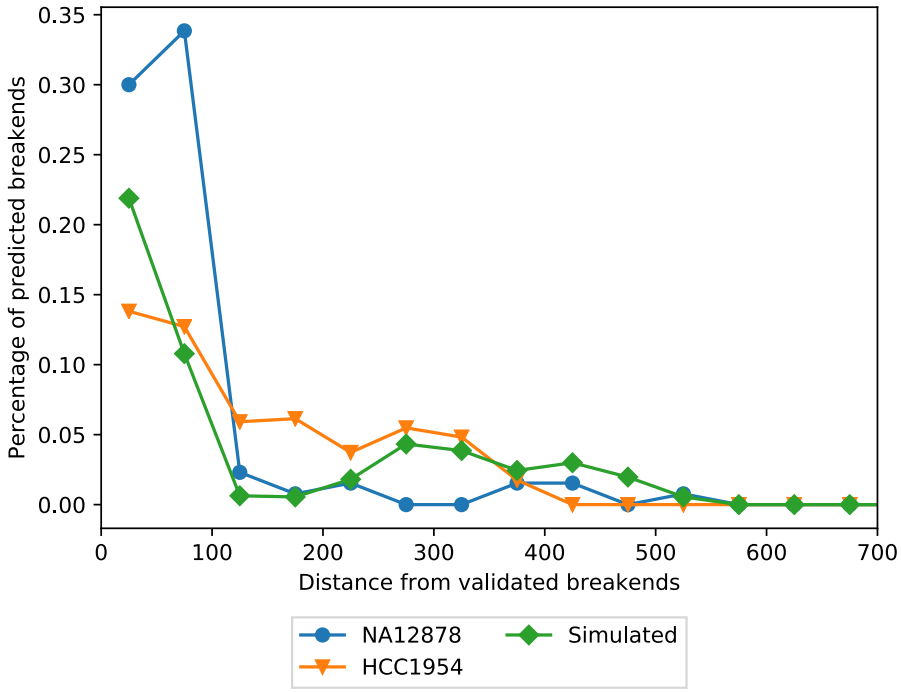


Figure 4.18: Percentage of breakends predicted by NAIBR that lie within a given distance from experimentally validated novel adjacencies in human cell lines NA12878 (blue), HCC1954 cancer cell line (orange), and simulated data (green). For all datasets, the majority of novel adjacencies predicted by NAIBR lie within 100bp of validated novel adjacencies. Distance is measured as the absolute value of the distance between the true and predicted breakends.

4.8 Identifying candidate novel adjacencies

Each candidate split molecule is defined with respect to a pair of oriented breakends i^+ and j^- representing a novel adjacency between an interval ending at i^+ and an interval starting at j^-. NAIBR provides the option to identify candidate novel adjacencies using discordant read-pairs, linked-reads, or user-defined novel adjacencies. Split reads could also be utilized to either define candidate novel adjacencies or to refine the breakends of adjacencies defined by discordant read-pairs or linked-reads. This is currently not employed by NAIBR and is left as future work. We define candidate novel adjacencies using discordant read-pairs as follows.

For a discordant read-pair $\langle x,y\rangle$, the location of the novel adjacency is determined by the orientations o_x and o_y of each read, x and y. For each of the four possible pairs of orientations, the candidate novel adjacencies are defined,

$$\begin{aligned}
(i^+, j^+) &= (r_x, r_y) \quad \text{if } o_x = + \text{ and } o_y = +,\\
(i^+, j^-) &= (r_x, l_y) \quad \text{if } o_x = + \text{ and } o_y = -,\\
(i^-, j^+) &= (l_x, r_y) \quad \text{if } o_x = - \text{ and } o_y = +,\\
(i^-, j^-) &= (l_x, l_y) \quad \text{if } o_x = - \text{ and } o_y = -.
\end{aligned}$$

Candidate novel adjacencies can also be defined using linked-reads. For each pair of linked-reads L_1, L_2 in a barcode β, L_1 and L_2 may have originated from a split-molecule with a novel adjacency in one of four orientations:

$$\begin{aligned}
(i^+, j^+) &= (e_1, e_2),\\
(i^+, j^-) &= (e_1, s_2),\\
(i^-, j^+) &= (s_1, e_2),\\
(i^-, j^-) &= (s_1, s_2),\\
\text{where } e_k &= \max\{r_y \mid \langle x,y\rangle \in L_k\},\\
s_k &= \min\{l_x \mid \langle x,y\rangle \in L_k\}.
\end{aligned}$$

L_1 and L_2 may have alternatively originated from distinct molecules assigned the same barcode by chance. To reduce the number of candidate novel adjacencies, we only apply NAIBR to candidate novel adjacencies with at least k overlaps. Two candidate novel adjacencies (i^+, j^-) and (a^+, b^-) overlap if $|i^+ - a^+| < \delta$ and $|j^- - b^-| < \delta$. k may be defined by the user with a default value of $k = 3$.

Each candidate novel adjacency is assigned a log-likelihood ratio score by NAIBR. If the breakends of two predicted novel adjacencies each fall within a distance $l_{\max}$ of each other than the novel adjacency with the higher log-likelihood ratio is reported.

By default, NAIBR utilizes only discordant read-pairs to define candidate novel adjacencies because they typically fall within several hundred bases of the true adjacency whereas linked-reads may be as much as a distance δ from the true novel adjacency. Using simulated data we found that we could obtain over 90% recall using candidate novel adjacencies defined by discordant read-pairs on data ranging from 10X-60X coverage.

4.9 Results

We assess NAIBR's ability to detect novel adjacencies in simulated and real 10X long-read sequencing data and benchmark against 5 other methods: Long Ranger [105], GROC-SVs [93], GASV [91], GASVPro [92], and LUMPY [80]. We chose these methods for comparison because they utilize different combinations of signals to identify and rank novel adjacencies. Long Ranger is 10X Genomics' structural variant detection program. Long Ranger identifies novel adjacencies by computing overlapping pairs of linked-reads and computes a likelihood score based on the number of overlaps observed in the data. GROC-SVs is also designed for linked-read sequencing data. GROC-SVs identifies structural variants by performing local assembly on barcoded reads. GROC-SVs operates in two steps. First it assigns p-value to novel adjacencies, and then it performs local assembly, labelling novel adjacencies as assembled or unassembled. In some cases, unassembled variants have smaller p-values than assembled variants. GASV, GASVPro, and LUMPY analyze paired-end sequencing data. When running these algorithms on 10X Genomics data we ignore barcodes and treat the data as Illumina paired-end sequencing data. GASV uses only discordant read-pairs and ranks novel adjacencies by the number of supporting discordant read-pairs. GASVPro uses a combination of discordant read-pairs and breakend read depth to assign a log-likelihood score to each predicted novel adjacency. LUMPY uses discordant and split reads to call novel adjacencies and reports a p-value for each adjacency.

We benchmark NAIBR against each method on both simulated and real data. Reported novel adjacencies are ranked from highest to lowest confidence according to the metrics used by each method. We run each

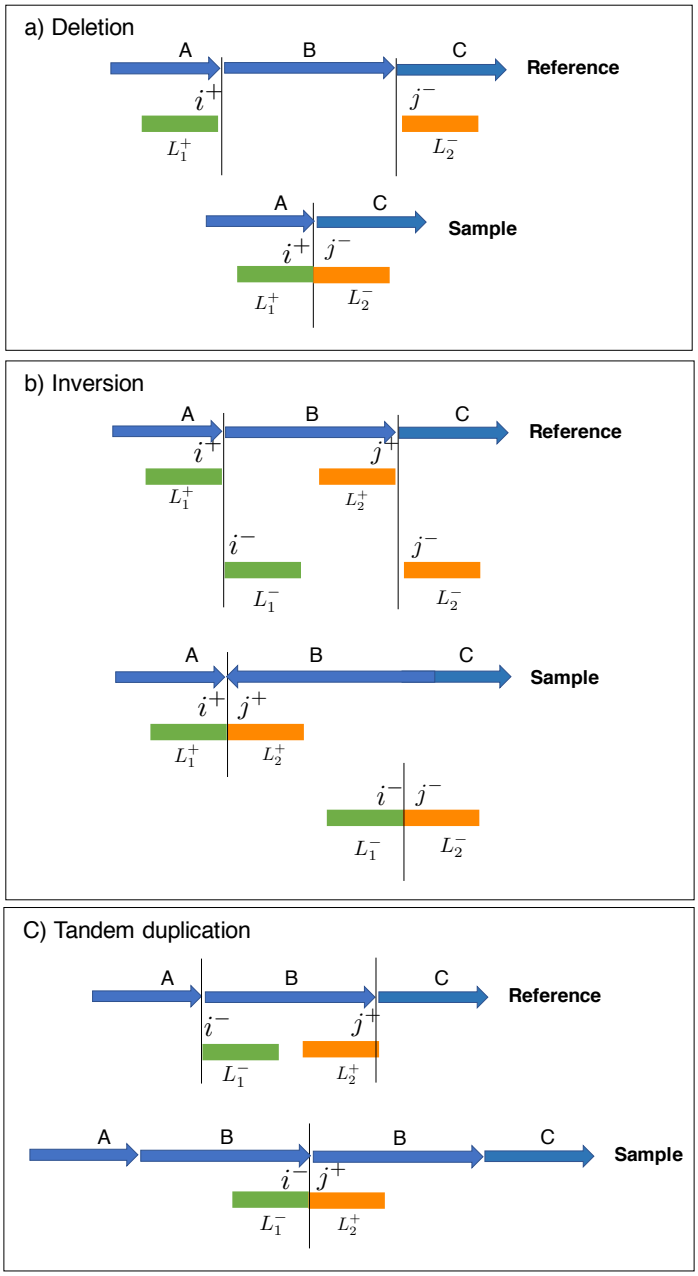


Figure 4.19: Signals observed by molecules spanning novel adjacencies produced by different structural variation events. a) A molecule spanning a novel adjacency produced by a deletion of B will be split if the size $|B|$ of interval B is $> \delta$. b) An inversion of the interval B will result in two novel adjacency. A molecule spanning a novel adjacency between the end of A and the end of B will be split if $|B| - |L_j^+| > \delta$. A molecule spanning a novel adjacency between the start of B and the start of C will be split if $|B| - |L_i^-| > \delta$. c) A tandem duplication of the interval B will result in a single novel adjacency between the end of B and the start of B. A molecule spanning this novel adjacency will be split if $|B| - |L_i^-| - |L_j^+| > \delta$.

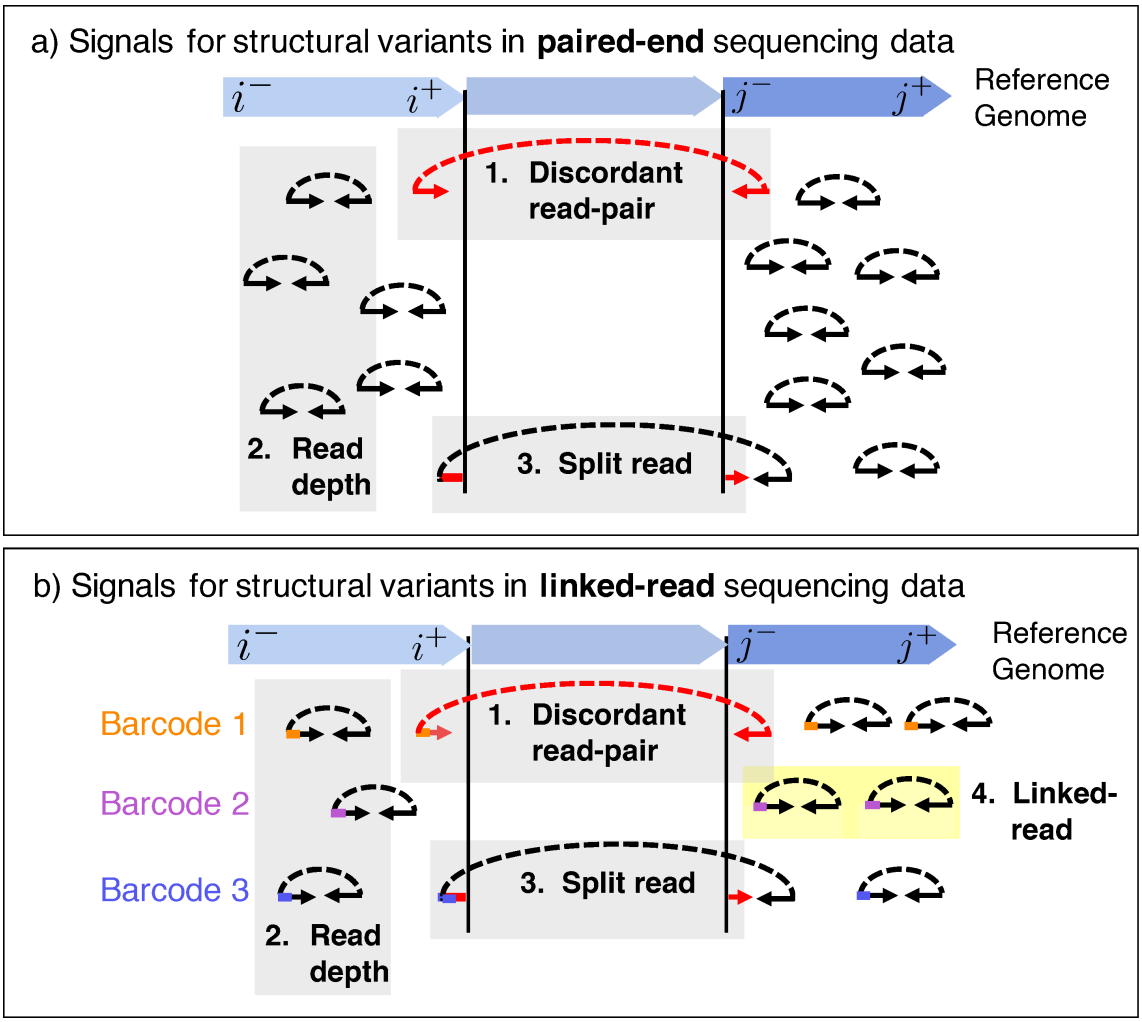


Figure 4.20: (a) Three signals of structural variants in paired-end sequencing data. (1) Discordant read-pairs occur when a read-pair aligns to the reference genome with non-concordant $(+,-)$ orientation or an insert size smaller than l_{min} or larger than l_{max}. (2) Read depth measures the number of reads mapping to a genomic region. Read depth will be lower in regions spanning a deletion and higher in regions spanning a duplication. (3) A split read occurs when a novel adjacency lies within one of the reads of the pair, causing it to be unmapped. (b) Linked-read sequencing contains all the signals of paired-end sequencing (discordant read-pairs, read depth, and split reads) and also adds linked-reads, which are formed from nearby read-pairs sharing the same barcode.

method using its default parameters. More specifics can be found in the Supplement.

We simulate several types of structural variants – including duplications, deletions, translocations, and inversions – on chromosomes 17 and 18 of the human reference genome hg19. To assess NAIBR's ability to detect novel adjacencies that occur on a single haplotype, we simulate two test genomes, one that contains 400 homozygous structural variants and one that contains 400 different structural variants on each haplotype. Translocations and inversions create more than 1 novel adjacency in the simulated genome, resulting in 508 homozygous novel adjacencies in the first simulated genome and 1027 heterozygous novel adjacencies in the second simulated genome. We simulate linked-read sequencing to 30X coverage. Details on simulation can be found in the Supplement. Figure 4.21a shows the precision-recall curve for all 5 methods run on the 30X test dataset containing 508 homozygous novel adjacencies. NAIBR has the highest recall of all methods, correctly identifying $479/508$ homozygous novel adjacencies. GASV correctly identified $309/508$ variants, however GASV reported several thousand variants, resulting in very low precision at high values of recall. LUMPY performed similarly to GASV, correctly identifying $308/508$ true variants, however the algorithm only reports variants with high probability scores according to their scoring metric, resulting in lower recall. GASVPro identified as many true variants as NAIBR at 50% recall, but only reported $289/508$ true variants in total, compared to 479 reported by NAIBR.

Long Ranger and GROC-SVs are each designed to utilize linked-reads, however both methods are limited to the identification of certain types of variants. Long Ranger reports variants larger than 30Kbp and GROC-SVs only reports variants larger than 10Kbp. The simulated dataset contains 369 novel adjacencies larger than 10Kbp and 269 structural variants larger than 30Kbp. Both Long Ranger and GROC-SVs perform with lower precision than NAIBR. For GROC-SVs, $38/39$ of assembled novel adjacencies were present in the truth set, showing that the local assembly approach has high precision. However an additional 59 true novel adjacencies were predicted by GROC-SVs but failed to assemble, indicating that local assembly removes many true positives. We perform the same comparison on a simulated dataset containing 800 heterozygous novel adjacencies with similar results. We also compare the runtime and memory usage of NAIBR to other methods and find that NAIBR outperforms other linked-read methods (see Supplement). These results demonstrate that NAIBR's incorporation of both linked-read and paired-end read data improves performance over other methods without significant additional time or memory requirements.

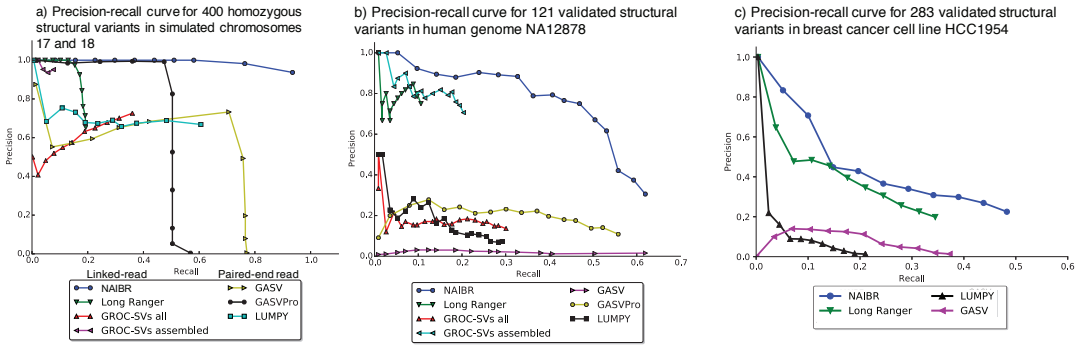


Figure 4.21: a) Precision-recall curve for NAIBR, Long Ranger, GROC-SVs, GASV, GASVPro, and LUMPY on 30X simulated data from chromosomes 17 and 18, containing 400 homozygous structural variants. b) Precision-recall curve for NAIBR, GASV, GASVPro, LUMPY, GROC-SVs, and Long Ranger evaluated against the set of 123 validated structural variants $>$ 7Kbp from NA12878 [80]. c) Precision-recall curve for NAIBR, GASV, GASVPro, LUMPY, GROC-SVs, and Long Ranger evaluated against the set of validated structural variants $\geq$ 30Kbp from breast cancer cell line HCC1954 [69, 74, 94].

4.9.1 Benchmarking on NA12878

To assess NAIBR's ability to detect known variants from a real dataset, we obtained whole-genome sequencing data of individual NA12878 of the 1000 Genomes Project from 10X genomics (`https://support.10xgenomics.com/genome-exome/datasets/NA12878_WGS_210`). The data was sequenced on the Chromium platform to 35X sequencing coverage with the Illumina Hiseq2500 to produce 845 million 98bp reads with mean insert size of 340bp and a mean molecule size of 68Kbp. We used the 2,950 validated novel adjacencies in NA12878 reported in [80] as the truth set. Novel adjacencies in this dataset were validated by split-read mapping analysis of independent long-read sequencing data from PacBio or Illumina Moleculo platforms.

Figure 4.21b shows the precision-recall curves each method against 121 validated structural variants larger than 7Kbp. NAIBR correctly predicted $73/121$ novel adjacencies $>$ 7Kbp. GASV also correctly predicted 73 novel adjacencies but displayed poor precision. GROC-SVs reported variants that it was able to assemble using its local assembly pipeline as well as variants that were not assembled. The local assembly step of GROC-SVs drastically reduces the number of reported false positives but also fails to assemble several predictions matching the truth set. Long Ranger reported 17 novel adjacencies, correctly identifying $12/32$ variants larger than 30Kbp from the truth set, compared to $17/32$ variants larger than 30Kbp detected by NAIBR. Figure S9 shows the precision-recall curves for each of the 5 methods on different structural variant sizes, ranging from 50bp-30Kbp. NAIBR outperforms other linked-read methods at detecting variants between 50bp and 10Kbp and outperforms paired-end methods at detecting variants larger than 5Kbp. NAIBR outperforms both linked-read and paired-end methods at detecting variants between 1Kbp and 10Kbp. In

summary, NAIBR detects large structural variants ($>$ 1Kbp) with better precision than paired-end callers GASV, GASVPro, and LUMPY and has higher recall and precision for variants smaller than 10Kbp compared to linked-read callers Long Ranger and GROC-SVs.

4.9.2 Tumor cell line HCC1954

We test NAIBR's ability to detect somatic structural variants in tumor cell line HCC1954T. The cell line was derived from a grade 3 invasive ductal carcinoma and sequenced by 10X Genomics to 35X coverage with a mean molecule size of 85Kbp. The matched normal HCC1954N was sequenced by 10X Genomics to 35X coverage with a mean molecule size of 88Kbp. We identify novel adjacencies in both HCC1954T and HCC1954N using 4 different methods: NAIBR, Long Ranger, GASV, and LUMPY. We formed a set of 369 true novel adjacencies by combining PCR-validated novel adjacencies from three previous studies: [69], [94], and [74].

NAIBR identifies 142 PCR-validated novel adjacencies, significantly more than Long Ranger (100), GASV (117), and LUMPY (55) (Figure S6). NAIBR also demonstrates better precision at all levels of recall than other methods (Figure 4.21c). Notably, GASV has significantly lower precision than Long Ranger and NAIBR, predicting over ten times as many novel adjacencies with lower recall than the other methods (Figure S6). NAIBR significantly outperforms Long Ranger at identifying duplications and interchromosomal events (Figure S5), identifying 72 duplications compared to 50 identified by Long Ranger and 53 interchromosomal events compared to 37 predicted by Long Ranger. Over 30% of the breakends predicted by NAIBR lie within 50bp of the breakends of the PCR-validated novel adjacencies, compared to approximately 10% of breakends predicted by Long Ranger and GASV (Figure S7).

Several of the novel adjacencies in HCC1954T that were not identified by Long Ranger and GASV affect known oncogenes and tumor suppressors. For example, a novel adjacency between Chr11:93153935 and Chr11:93160223 occurs within the gene CCDC67, potentially leading to loss of function of the gene. CCDC67 has recently been identified as a tumor suppressor gene [103, 106]. A novel adjacency between Chr14:89829140 and Chr7:155683934 affects the forkhead transcription factor checkpoint suppressor, CHES1. CHES1 expression has been shown to be reduced across many cancer types [76]. A novel adjacency between Chr11:69059340 and Chr11:69089741, surrounding the MYEOV gene (Chr:69061613-69064754), is potentially a result of a duplication of MYEOV. MYEOV has been shown to be amplified in breast cancer patients and has been identified as a candidate oncogene [79].

4.10 Discussion

We present NAIBR, a probabilistic algorithm for the identification of novel adjacencies using linked-read sequencing data. Linked-read sequencing combines low per-base error rate of short-read sequencing technologies with long-range linking information of long-read technologies. Linked-read sequencing offers drastically improved mapping and phasing results compared to paired-end sequencing [70] with similar cost, making it an attractive option for researchers. NAIBR is one of the first algorithms that identifies structural variation by using signals unique to linked-read sequencing data. NAIBR uses discordant read-pairs obtained from paired-end reads combined with candidate split molecule obtained from linked-reads to identify and rank novel adjacencies. NAIBR detects novel adjacencies with higher accuracy and precision than existing methods on both simulated and real linked-read sequencing data.

We also demonstrate NAIBR's ability to predict somatic novel adjacencies from cancer data by applying it to cell line HCC1954. Several novel adjacencies detected by NAIBR were not identified by other methods, including novel adjacencies affecting tumor suppressor genes CCDC67 and CHES1 and candidate oncogene MYEOV. While some of the novel adjacencies predicted by NAIBR might be verified PCR, it is possible that long-read sequencing data (e.g.from PacBio or Oxford Nanopore) would be required, since these adjacencies were not readily apparent from short-read sequencing data.

As future work, we plan to incorporate additional signals, such as read depth and split-reads, into our probabilistic model for identifying novel adjacencies. Our algorithm can also be extended by performing local assembly on linked-reads supporting novel adjacencies to reconstruct structural variants, as is done by GROC-SVs [93].

The utility of linked-reads in identifying novel adjacencies between nearby positions on the reference genome (such as a small deletion) is limited by the fact that each molecule is sequenced to low coverage, introducing large gaps between read-pairs sequenced from the same molecule. Thus, with the current Chromium technology from 10X Genomics, linked-reads provide substantial additional power to detect large structural variants but provide little additional power above that of paired-end sequencing in the detection of small structural variants. However, linked-read sequencing is a very new technology and is likely to improve in the coming years. As molecules are sequenced to higher coverage, NAIBR's ability to detect novel adjacencies using candidate split molecules will improve significantly, enabling the identification of novel adjacencies arising from smaller structural variants.

Chapter 5

Conclusions and Future Directions

Recently developed high-dimensional sequencing technologies such as scRNA-seq, STRNA-seq, linked-read sequencing, and others provide increased resolution over standard DNA-seq and RNA-seq technologes, boosting our ability to study biological mechanisms underlying disease with unprecedented precision. However, it is currently only feasible to obtain this high resolution at the cost of low-coverage. This means that analysis methods, such as clustering, differential expression analysis, and copy number calling, designed for high-coverage "bulk" sequencing data will not be effective on the sparse high-dimensional data from these new technologies. This motivates the need for analysis methods designed specifically for these technologies, which can make use of their benefits and overcome the negative effect of low-coverage. To address this need, we have developed three methods for three different sequencing technologies, scRNA-seq, STRNA-seq, and linked-read sequencing, which utilize known dependencies to improve the analysis and interpretation of sparse high-dimensional sequencing data.

These known dependencies can come in many forms. First, we introduce a method, netNMF-sc, which makes use of known correlations in expression between gene pairs obtained from prior RNA-seq and microarray experiments. By incorporating gene-gene correlations from prior experiments in the form of a gene coexpression network, netNMF-sc is able to accurately recover cell clusters from scRNA-seq data. Incorporating these known gene-gene correlations gives netNMF-sc an advantage over other methods because of the sparsity of scRNA-seq data, often containing as many as 90% zero entries. We have shown that netNMF-sc performs well on several real scRNA-seq datasets with a variety of gene coexpression networks, however the method relies on the strong assumption that gene-gene correlations are the same for every cell in the scRNA-seq experiment. While it is generally true that many gene regulatory pathways are conserved across

organisms of the same species, a one-size-fits all approach may not be best. An interesting future direction would be to investigate cell-type-specific gene coexpression networks and model the coexpression of the scRNA-seq data as a combination of contributions from each of these cell-type-specific networks.

We next introduce a method STCNA which uses prior knowledge of gene and spot dependencies to infer copy number aberrations (CNAs) from spatial transcriptomics RNA-seq (STRNA-seq) data. Unlike netNMF-sc, the prior knowledge of these dependencies comes directly from the dataset of interest. This is ideal because this information will always be available for any STRNA-seq experiment of any organism, whereas there may be limited prior knowledge available from rarely studied organisms/tissues for use with netNMF-sc. To our knowledge, STCNA is the first method which incorporates spatial information to infer CNAs from STRNA-seq data. We demonstrate that by incorporating both dependencies between adjacent genes on the genome and adjacent spots in a tissue, we can significantly improve the inference of clone CNA profiles and the assignment of spots to clones. Exploring violations of these dependencies in STRNA-seq data is an interesting avenue for future work. For example, a clone which does not show significant spatial clustering and has spots distributed throughout a tumor tissue could indicate a population of cells which has increased mobility. These mobile cells may suggest poor prognosis with a higher chance of metastasis. Another potential future direction is investigating the relationship between copy number aberrations and gene expression in tumor tissues. With STCNA we can currently only identify genes that are directly affected by copy number changes. However, the deletion of DNA coding for a transcription factor, for example, may set of a cascade of regulatory changes, resulting in higher or lower expression of several genes in a shared regulatory pathway. Further research of paired gene expression and CNA profiles across many individuals is needed to determine these complex causal relationships between copy number aberrations and downstream changes in gene expression.

The third method we introduce is NAIBR which infers novel adjacencies created by structural variants in a tumor genome from linked-read sequencing data. Linked-read sequencing data consists of barcoded paired-end reads which originate from long molecules $\sim$ 50Kb in length. The probability of a paired-end read originating from a molecule that spans a novel adjacency is dependent on paired-end reads of other molecules sharing the same barcode. By incorporating these dependencies into a probabilistic model, we demonstrate that NAIBR improves the identification of novel adjacencies over other existing methods on multiple datasets. Directions for future work include incorporating additional signals, such as read depth and split-reads into the probabilistic model. The probabilistic model could also be extended by performing local assembly of linked-reads supporting novel adjacencies to reconstruct structural variants. Combining read

depth, split-reads, and local assembly into the existing model would reduce the number of false positive calls and potentially facilitate the inference of smaller events.

Bibliography

[1] Elyanow, R., Wu, H. and Raphael, B. Identifying structural variants using linked-read sequencing data. *Bioinformatics*. **34**, 353–360 (2018)

[2] Elyanow, R., Dumitrascu, B., Engelhardt, B. and Raphael, B. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Research*. **30**, 195–204 (2020)

[3] Ståhl, P., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. and Huss, M. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016)

[4] Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L., Melnyk, N., Mcpherson, A., Bashashati, A. and Laks, E. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*. **24**, 1881–1893 (2014)

[5] Svensson, V., Teichmann, S. and Stegle, O. SpatialDE: identification of spatially variable genes. *Nature Methods*. **15**, 343 (2018)

[6] Arnol, D., Schapiro, D., Bodenmiller, B., Saez-rodriguez, J. and Stegle, O. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*. **29**, 202–211 (2019)

[7] Smyth, G. Limma: linear models for microarray data. (Springe,2005)

[8] Anders, S. and Huber, W. Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (embl)*. **10** pp. f1000research (201)

[9] Kiselev, V., Kirschner, K., Schaub, M., Andrews, T., Yiu, A., Ra, T., Natarajan, K., Reik, W., Barahona, M. and Green, A. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*. **14**, 483–486 (2017)

[10] Choi, S., Henderson, M., Kwan, E., Beesley, A., Sutton, R., Bahar, A., Giles, J., Venn, N., Pozza, L. and Baker, D. Relapse in children with acute lymphoblastic leukemia involving selection of a preexisting drug-resistant subclone. *Blood, The Journal Of The American Society Of Hematology*. **110**, 632–639 (2007)

[11] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E. and Mesirov, J. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings Of The National Academy Of Sciences*. **98**, 15149–15154 (2001)

[12] Farmer, P., Bonnefoi, H., Anderle, P., Cameron, D., Wirapati, P., Becette, V., André, S., Piccart, M., Campone, M. and Brain, E. A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nature Medicine*. **15**, 68 (2009)

[13] Wan, Y., Sabbagh, E., Raese, R., Qian, Y., Luo, D., Denvir, J., Vallyathan, V., Castranova, V. and Guo, N. Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction. *Plos One*. **5** (2010)

[14] Larsen, J., Pavey, S., Passmore, L., Bowman, R., Hayward, N. and Fong, K. Gene expression signature predicts recurrence in lung adenocarcinoma. *Clinical Cancer Research*. **13**, 2946–2954 (2007)

[15] Liu, R., Wang, X., Chen, G., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K. and Clarke, M. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New England Journal Of Medicine*. **356**, 217–226 (2007)

[16] Lachmann, A., Torre, D., Keenan, A., Jagodnik, K., Lee, H., Wang, L., Silverstein, M. and Ma'ayan, A. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*. **9**, 1–10 (2018)

[17] Sanger, F., Nicklen, S. and Coulson, A. DNA sequencing with chain-terminating inhibitors. *Proceedings Of The National Academy Of Sciences*. **74**, 5463–5467 (1977)

[18] Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. **10**, 57–63 (2009)

[19] Eng, C., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C. and Yuan, G. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. **568**, 235–239 (2019)

[20] Lubeck, E., Coskun, A., Zhiyentayev, T., Ahmad, M. and Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*. **11**, 360 (2014)

[21] Andersson, A., Bergenstråhle, J., Asp, M., Navarro, J., Bergenstråhle, L., Jurek, A. and Lundeberg, J. Spatial mapping of cell types by integration of transcriptomics data. *Biorxiv*. (2019)

[22] Carlberg, K., Korotkova, M., Larsson, L., Catrina, A., Ståhl, P. and Malmström, V. Exploring inflammatory signatures in arthritic joint biopsies with Spatial Transcriptomics. *Scientific Reports*. **9**, 1–10 (2019)

[23] Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. and Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Research*. **78**, 5970–5979 (2018)

[24] Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J. and Salmén, F. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*. **179**, 1647–1660 (2019)

[66] Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, **21**(6), 974–984.

[67] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, **12**(5), 363–376.

[68] Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M., *et al.* (2013). Punctuated evolution of prostate cancer genomes. *Cell*, **153**(3), 666–677.

[69] Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., *et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome research*, **17**(9), 1296–1303.

[70] Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D. E., West, R., Sidow, A., and Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. *Genome research*, **25**(10), 1570–1580.

[71] Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, **6**(9), 677–81.

[72] Chittenden, T., Chittenden, T., Howe, E., Howe, E., a.C. Culhane, a.C. Culhane, Sultana, R., Sultana, R., Taylor, J., Taylor, J., Holmes, C., Holmes, C., Quackenbush, J., and Quackenbush, J. (2008). Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics*, **455**(7216), 1061–1068.

[73] Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A. Y., Boutros, P., Chen, J., *et al.* (2017). novobreak: local assembly for breakpoint detection in cancer genomes. *Nature methods*, **14**(1), 65–67.

[74] Galante, P. A., Parmigiani, R. B., Zhao, Q., Caballero, O. L., De Souza, J. E., Navarro, F. C., Gerber, A. L., Nicolás, M. F., Salim, A. C. M., Silva, A. P. M., *et al.* (2011). Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic acids research*, **39**(14), 6056–6068.

[75] Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**(12), i350–i357.

[76] Huot, G., Vernier, M., Bourdeau, V., Doucet, L., Saint-Germain, E., Gaumont-Leclerc, M.-F., Moro, A., and Ferbeyre, G. (2014). Ches1/foxn3 regulates cell proliferation by repressing pim2 and protein biosynthesis. *Molecular biology of the cell*, **25**(5), 554–565.

[77] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, **36**(9), 949–951.

[78] Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., and Boeva, V. (2016). Sv-bay: structural

variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability. *Bioinformatics*, **32**(7), 984–992.

[79] Janssen, J. W., Imoto, I., Inoue, J., Shimada, Y., Ueda, M., Imamura, M., Bartram, C. R., and Inazawa, J. (2002). Myeov, a gene at 11q13, is coamplified with ccnd1, but epigenetically inactivated in a subset of esophageal squamous cell carcinomas. *Journal of human genetics*, **47**(9), 460–464.

[80] Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, **15**(6), R84.

[81] Luo, R. (2017). https://github.com/aquaskyline/lrsim.

[82] Mak, A. C., Lai, Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K., Poon, A., Mostovoy, Y., Hastie, A. R., Stedman, W., Anantharaman, T., *et al.* (2016). Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, **202**(1), 351–362.

[83] Mccarroll, S. a., Huett, A., Kuballa, P., Chilewski, S. D., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Judy, H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Daly, M. J., and Xavier, R. J. (2009). NIH Public Access. **40**(9), 1107–1112.

[84] Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. a., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., and Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, **11**.

[85] Rausch, T., Zichner, T., Schlattl, A., St??tz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), 333–339.

[86] Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Paul, A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Pleier, S., Cin, H., Hawkins, C., Beck, C., Von, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Benes, V., Kulozik, A. E., Witt, O., and Martin, D. (2013). NIH Public Access. *NIH Public Access*, **148**, 59–71.

[87] Ritz, A., Bashir, A., and Raphael, B. J. (2010). Structural variation analysis with strobe reads. *Bioinformatics*, **26**(10), 1291–1298.

[88] Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, **30**(24), 3458–3466.

[89] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science*, **316**(5823), 445–449.

[90] Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009a). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), 222–230.

[91] Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009b). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), i222–i230.

[92] Sindi, S. S., Onal, S., Peng, L. C., Wu, H.-T., and Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, **13**(3), R22.

[93] Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., Salit, M., West, R. B., Batzoglou, S., and Sidow, A. (2016). Genome-wide reconstruction of complex structural variants using read clouds. *bioRxiv*, page 074518.

[94] Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**(7276), 1005.

[95] Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., *et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, **144**(1), 27–40.

[96] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

[97] Tattini, L., D?Aurizio, R., and Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, **3**.

[98] Wala, J., Bandopadhayay, P., Greenwald, N., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S. E., Li, Y., Weischenfeldt, J., Yao, X., *et al.* (2017). Genome-wide detection of structural variants and indels by local assembly. *bioRxiv*, page 105080.

[99] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, **14**(2), 125–138.

[100] Weisenfeld, N. I., Kumar, V., Shah, P., Church, D., and Jaffe, D. B. (2016). Direct determination of diploid genome sequences. *bioRxiv*, page 070425.

[101] Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M., and Park, P. J. (2010). Bic-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome biology*, **11**(S1), O10.

[102] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**(21), 2865–2871.

[103] Yin, D.-T., Xu, J., Lei, M., Li, H., Wang, Y., Liu, Z., Zhou, Y., and Xing, M. (2016). Characterization of the novel tumor-suppressor gene ccdc67 in papillary thyroid carcinoma. *Oncotarget*, **7**(5), 5830.

[105] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

[105] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

[106] Zhu, X.-J., Liu, J., Xu, X.-Y., Zhang, C.-D., and Dai, D.-Q. (2014). Novel tumor-suppressor gene epidermal growth factor-containing fibulin-like extracellular matrix protein 1 is epigenetically silenced and associated with invasion and metastasis in human gastric cancer. *Molecular medicine reports*, **9**(6), 2283–2292.

[66] Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, **21**(6), 974–984.

[67] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, **12**(5), 363–376.

[68] Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M., *et al.* (2013). Punctuated evolution of prostate cancer genomes. *Cell*, **153**(3), 666–677.

[69] Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., *et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome research*, **17**(9), 1296–1303.

[70] Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D. E., West, R., Sidow, A., and Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. *Genome research*, **25**(10), 1570–1580.

[71] Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, **6**(9), 677–81.

[72] Chittenden, T., Chittenden, T., Howe, E., Howe, E., a.C. Culhane, a.C. Culhane, Sultana, R., Sultana, R., Taylor, J., Taylor, J., Holmes, C., Holmes, C., Quackenbush, J., and Quackenbush, J. (2008). Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics*, **455**(7216), 1061–1068.

[73] Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A. Y., Boutros, P., Chen, J., *et al.* (2017). novobreak: local assembly for breakpoint detection in cancer genomes. *Nature methods*, **14**(1), 65–67.

[74] Galante, P. A., Parmigiani, R. B., Zhao, Q., Caballero, O. L., De Souza, J. E., Navarro, F. C., Gerber, A. L., Nicolás, M. F., Salim, A. C. M., Silva, A. P. M., *et al.* (2011). Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic acids research*, **39**(14), 6056–6068.

[75] Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**(12), i350–i357.

[76] Huot, G., Vernier, M., Bourdeau, V., Doucet, L., Saint-Germain, E., Gaumont-Leclerc, M.-F., Moro, A., and Ferbeyre, G. (2014). Ches1/foxn3 regulates cell proliferation by repressing pim2 and protein biosynthesis. *Molecular biology of the cell*, **25**(5), 554–565.

[77] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, **36**(9), 949–951.

[78] Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., and Boeva, V. (2016). Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability. *Bioinformatics*, **32**(7), 984–992.

[79] Janssen, J. W., Imoto, I., Inoue, J., Shimada, Y., Ueda, M., Imamura, M., Bartram, C. R., and Inazawa, J. (2002). Myeov, a gene at 11q13, is coamplified with ccnd1, but epigenetically inactivated in a subset of esophageal squamous cell carcinomas. *Journal of human genetics*, **47**(9), 460–464.

[80] Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, **15**(6), R84.

[81] Luo, R. (2017). https://github.com/aquaskyline/lrsim.

[82] Mak, A. C., Lai, Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K., Poon, A., Mostovoy, Y., Hastie, A. R., Stedman, W., Anantharaman, T., *et al.* (2016). Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, **202**(1), 351–362.

[83] Mccarroll, S. a., Huett, A., Kuballa, P., Chilewski, S. D., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Judy, H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Daly, M. J., and Xavier, R. J. (2009). NIH Public Access. **40**(9), 1107–1112.

[84] Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. a., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., and Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, **11**.

[85] Rausch, T., Zichner, T., Schlattl, A., St??tz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), 333–339.

[86] Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Paul, A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Pleier, S., Cin, H., Hawkins, C., Beck, C., Von, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Benes, V., Kulozik, A. E., Witt, O., and Martin, D. (2013). NIH Public Access. *NIH Public Access*, **148**, 59–71.

[87] Ritz, A., Bashir, A., and Raphael, B. J. (2010). Structural variation analysis with strobe reads. *Bioinformatics*, **26**(10), 1291–1298.

[88] Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, **30**(24), 3458–3466.

[89] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science*, **316**(5823), 445–449.

[90] Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009a). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), 222–230.

[91] Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009b). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), i222–i230.

[92] Sindi, S. S., Onal, S., Peng, L. C., Wu, H.-T., and Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, **13**(3), R22.

[93] Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., Salit, M., West, R. B., Batzoglou, S., and Sidow, A. (2016). Genome-wide reconstruction of complex structural variants using read clouds. *bioRxiv*, page 074518.

[94] Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**(7276), 1005.

[95] Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., *et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, **144**(1), 27–40.

[96] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

[97] Tattini, L., D?Aurizio, R., and Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, **3**.

[98] Wala, J., Bandopadhayay, P., Greenwald, N., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S. E., Li, Y., Weischenfeldt, J., Yao, X., *et al.* (2017). Genome-wide detection of structural variants and indels by local assembly. *bioRxiv*, page 105080.

[99] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, **14**(2), 125–138.

[100] Weisenfeld, N. I., Kumar, V., Shah, P., Church, D., and Jaffe, D. B. (2016). Direct determination of diploid genome sequences. *bioRxiv*, page 070425.

[101] Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M., and Park, P. J. (2010). Bic-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome biology*, **11**(S1), O10.

[102] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**(21), 2865–2871.

[103] Yin, D.-T., Xu, J., Lei, M., Li, H., Wang, Y., Liu, Z., Zhou, Y., and Xing, M. (2016). Characterization of the novel tumor-suppressor gene ccdc67 in papillary thyroid carcinoma. *Oncotarget*, **7**(5), 5830.

[105] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

[105] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

[106] Zhu, X.-J., Liu, J., Xu, X.-Y., Zhang, C.-D., and Dai, D.-Q. (2014). Novel tumor-suppressor gene epidermal growth factor-containing fibulin-like extracellular matrix protein 1 is epigenetically silenced and associated with invasion and metastasis in human gastric cancer. *Molecular medicine reports*, **9**(6), 2283–2292.

[107] Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J. and Grubor, V. Inferring tumor progression from genomic heterogeneity. *Genome Research*. **20**, 68–80 (2010)

[108] Burkhardt, L., Grob, T., Hermann, I., T, E., Choschzick, M., Jänicke, F., Müller, V., Bokemeyer, C., Simon, R. and Sauter, G. Gene amplification in ductal carcinoma in situ of the breast. *Breast Cancer Research And Treatment*. **123**, 757–765 (2010)

[109] Merlo, L., Pepper, J., Reid, B. and Maley, C. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*. **6**, 924–935 (2006)

[110] Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A. and Shumansky, K. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*. **22**, 1995–2007 (2012)

[111] Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L., Melnyk, N., Mcpherson, A., Bashashati, A. and Laks, E. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*. **24**, 1881–1893 (2014)

[112] Tirosh, I., Venteicher, A., Hebert, C., Escalante, L., Patel, A., Yizhak, K., Fisher, J., Rodman, C., Mount, C. and Filbin, M. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*. **539**, 309–313 (2016)

[113] Patel, A., Tirosh, I., Trombetta, J., Shalek, A., Gillespie, S., Wakimoto, H., Cahill, D., Nahed, B., Curry, W. and Martuza, R. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. **344**, 1396–1401 (2014)

[114] Djebali, S., Davis, C., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W. and Schlesinger, F. Landscape of transcription in human cells. *Nature*. **489**, 101–108 (2012)

[115] Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I. and Oren, M. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology*. **15**, R69 (2014)

[116] Jareborg, N., Birney, E. and Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research*. **9**, 815–824 (1999)

[117] Zack, T., Schumacher, S., Carter, S., Cherniack, A., Saksena, G., Tabak, B., Lawrence, M., Zhang, C., Wala, J. and Mermel, C. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*. **45**, 1134–1140 (2013)

[118] Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. and Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Research*. **78**, 5970–5979 (2018)

[119] Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal Of The Royal Statistical Society: Series B (methodological)*. **36**, 192–225 (1974)

[120] Fan, J., Lee, H., Lee, S., Ryu, D., Lee, S., Xue, C., Kim, S., Kim, K., Barkas, N. and Park, P. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Research*. **28**, 1217–1227 (2018)

[121] Wang, H., Wellmann, J., Li, Z., Wang, X. and Liang, R. A segmentation approach for stochastic geological modeling using hidden Markov random fields. *Mathematical Geosciences*. **49**, 145–177 (2017)

[122] Kenny, P. InferCNV, a python web app for copy number inference from discrete gene-level amplification signals noted in clinical tumor profiling reports. *F1000research*. **8** (201)

[123] Svensson, V., Teichmann, S. and Stegle, O. SpatialDE: identification of spatially variable genes. *Nature Methods*. **15**, 343 (2018)

[124] Ståhl, P., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. and Huss, M. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016)

[125] Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. and Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Research*. **78**, 5970–5979 (2018)

[126] He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition.

[127] Nowell, P. The clonal evolution of tumor cell populations. *Science*. **194**, 23–28 (1976)

[128] Zhang, Y., Brady, M. and Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Ieee Transactions On Medical Imaging*. **20**, 45–57 (2001)

[129] Shang, F., Jiao, L. and Wang, F. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition*. **45**, 2237–2250 (2012)

[130] Zhang, T., Xia, Y. and Feng, D. An Evolutionary HMRF approach to brain MR image segmentation using clonal selection algorithm. *Ifac Proceedings Volumes*. **45**, 6–11 (2012)

[131] Huang, K., Zhao, Z., Gong, Q., Zha, J., Chen, L. and Yang, R. Nasopharyngeal carcinoma segmentation via HMRF-EM with maximum entropy.

[132] Ibrahim, M., John, N., Kabuka, M. and Younis, A. Hidden Markov models-based 3D MRI brain segmentation. *Image And Vision Computing*. **24**, 1065–1079 (2006)

[133] Pradhan, S. and Patra, D. Unsupervised brain magnetic resonance image segmentation using HMRF-FCM framework.

[134] Zhu, Q., Shah, S., Dries, R., Cai, L. and Yuan, G. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature Biotechnology*. **36**, 1183 (2018)

[135] Mccarthy, D., Rostom, R., Huang, Y., Kunz, D., Danecek, P., Bonder, M., Hagai, T., Wang, W., Gaffney, D. and Simons, B. Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. *Biorxiv*. pp. 413047 (2018)

[136] Ståhl, P., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. and Huss, M. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016)

[137] Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J., Baron, M., Hajdu, C., Simeone, D. and Yanai, I. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*. pp. 1–10 (2020)

[138] Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S. and Larsson, L. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications*. **9**, 1–13 (2018)

[139] Arnol, D., Schapiro, D., Bodenmiller, B., Saez-rodriguez, J. and Stegle, O. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*. **29**, 202–211 (2019)

[140] äijö, T., Maniatis, S., Vickovic, S., Kang, K., Cuevas, M., Braine, C., Phatnani, H., Lundeberg, J. and Bonneau, R. Splotch: Robust estimation of aligned spatial temporal gene expression data. *Biorxiv*. pp. 757096 (2019)

[141] Nitzan, M., Karaiskos, N., Friedman, N. and Rajewsky, N. Gene expression cartography. *Nature*. **576**, 132–137 (2019)

[142] Paraic A., Kenny, InferCNV, a python web app for copy number inference from discrete gene-level amplification signals noted in clinical tumor profiling reports. *F1000Research*. **8** (2019)

[143] Andor, N., Lau, B., Catalanotti, C., Kumar, V., Sathe, A., Belhocine, K., Wheeler, T., Price, A., Song, M. and Stafford, D. Joint single cell DNA-Seq and RNA-Seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *Biorxiv*. pp. 445932 (2018)

[144] Campbell, K., Steif, A., Laks, E., Zahn, H., Lai, D., Mcpherson, A., Farahani, H., Kabeer, F., O'flanagan, C. and Biele, J. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology*. **20**, 54 (2019)

[145] Pettit, J., Rajutomer, K., Richardson, S., Azizi, L. and Marioni, J. Identifying cell types from spatially referenced single-cell expression datasets. *Plos Computational Biology*. **10** (2014)

[146] Elham Azizi, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er. Bayesian inference for single-cell clustering and imputing. *Genomics and Computational Biology*, 3(1):e46–e46, 2017.

[147] 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. *Nature biotechnology*, 33(2):155, 2015.

[148] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 63–72. IEEE, 2008.

[149] Aedín C Culhane, Thomas Schwarzl, Razvan Sultana, Kermshlise C Picard, Shaita C Picard, Tim H Lu, Katherine R Franklin, Simon J French, Gerald Papenhausen, Mick Correll, et al. Genesigdb—a curated database of gene expression signatures. *Nucleic acids research*, 38(suppl_1):D716–D725, 2009.

[150] Daniel Dominguez, Yi-Hsuan Tsai, Nicholas Gomez, Deepak Kumar Jha, Ian Davis, and Zefeng Wang. A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell research*, 26(8):946, 2016.

[151] Ghislain Durif, Laurent Modolo, Jeff E Mold, Sophie Lambert-Lacroix, and Franck Picard. Probabilistic count matrix factorization for single cell expression data analysis. In *RECOMB*, pages 254–255. Springer, 2018.

[152] Cédric Févotte and A Taylan Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. In *2009 17th European Signal Processing Conference*, pages 1913–1917. IEEE, 2009.

[153] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.

[154] Don L Gibbons and Chad J Creighton. Pan-cancer survey of epithelial–mesenchymal transition markers across the cancer genome atlas. *Developmental Dynamics*, 247(3):555–564, 2018.

[155] Sanjay Surendranath Girija. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.

[156] Chen Gong, Dacheng Tao, Jie Yang, and Keren Fu. Signed laplacian embedding for supervised dimension reduction. In *AAAI*, pages 1847–1853, 2014.

[157] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *bioRxiv*, page 576827, 2019.

[158] Brian L Hie, Bryan Bryson, and Bonnie Berger. Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv*, page 371179, 2018.

[159] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108, 2013.

[160] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, page 1, 2018.

[161] Giovanni Iacono, Elisabetta Mereu, Amy Guillaumet-Adkins, Roser Corominas, Ivon Cuscó, Gustavo Rodríguez-Esteban, Marta Gut, Luis Alberto Pérez-Jurado, Ivo Gut, and Holger Heyn. bigscale: an analytical framework for big-scale single-cell data. *Genome research*, 2018.

[162] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[163] 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, 2015.

[164] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 559–570. SIAM, 2010.

[165] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[166] Homin K Lee, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14(6):1085–1094, 2004.

[167] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[168] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

[169] Chieh Lin, Siddhartha Jain, Hannah Kim, and Ziv Bar-Joseph. Using neural networks for reducing the dimensions of single-cell rna-seq data. *Nucleic acids research*, 45(17):e156–e156, 2017.

[170] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):59, 2017.

[171] George C Linderman, Jun Zhao, and Yuval Kluger. Zero-preserving imputation of scrna-seq data using low-rank approximation. *bioRxiv*, page 397588, 2018.

[172] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[173] Andrew T McKenzie, Minghui Wang, Mads E Hauberg, John F Fullard, Alexey Kozlenkov, Alexandra Keenan, Yasmin L Hurd, Stella Dracheva, Patrizia Casaccia, Panos Roussos, et al. Brain cell type specific gene expression and co-expression network architectures. *Scientific reports*, 8, 2018.

[174] Yasunobu Okamura, Yuichi Aoki, Takeshi Obayashi, Shu Tadaka, Satoshi Ito, Takafumi Narise, and Kengo Kinoshita. Coxpresdb in 2015: coexpression database for animal species by dna-microarray and rnaseq-based expression data with multiple quality assessment systems. *Nucleic acids research*, 43 (D1):D82–D86, 2014.

[175] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[176] Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

[177] Jonathan Ronen and Altuna Akalin. netsmooth: Network-smoothing based imputation for single cell rna-seq. *F1000Research*, 7, 2018.

[178] Carole Sousa, Knut Biber, and Alessandro Michelucci. Cellular and molecular characterization of microglia: a unique immune cell population. *Frontiers in immunology*, 8:198, 2017.

[179] Shiquan Sun, Yabo Chen, Yang Liu, and Xuequn Shang. A fast and efficient count-based matrix factorization method for detecting cell types from single-cell rnaseq data. *BMC systems biology*, 13(2): 28, 2019.

[180] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *bioRxiv*, 2019. doi: 10.1101/582064. URL `https://www.biorxiv.org/content/early/2019/03/19/582064`.

[181] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single cell rna-seq based on a multinomial model. *bioRxiv*, page 574574, 2019.

[182] David Van Dijk, Roshan Sharma, Juoas Nainys, Kristina Yim, Pooja Kathail, Ambrose Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. 2018.

[183] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414, 2017.

[184] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57, 2009.

[185] Paul D Wes, Inge R Holtman, Erik WGM Boddeke, Thomas Möller, and Bart JL Eggen. Next generation transcriptomics and genomics elucidate biological complexity of microglia in health and disease. *Glia*, 64(2):197–213, 2016.

[186] Guanming Wu, Xin Feng, and Lincoln Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5):R53, 2010.

[187] Huilei Xu, Yen-Sin Ang, Ana Sevilla, Ihor R Lemischka, and Avi Ma'ayan. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS computational biology*, 10(8):e1003777, 2014.

[188] Sunmo Yang, Chan Yeong Kim, Sohyun Hwang, Eiru Kim, Hyojin Kim, Hongseok Shim, and Insuk Lee. Coexpedia: exploring biomedical hypotheses via co-expressions associated with medical subject headings (mesh). *Nucleic acids research*, 45(D1):D389–D396, 2016.

[189] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[190] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[191] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

[192] Lingxue Zhu, Jing Lei, Bernie Devlin, Kathryn Roeder, et al. A unified statistical framework for single cell and bulk rna sequencing data. *The Annals of Applied Statistics*, 12(1):609–632, 2018.

[193] 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. *Molecular cell*, 65(4):631–643, 2017.

[194] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *nature protocols*, 12 (1):44, 2017.