Computational Methods for the Analysis of Single-Cell Transcriptomic Data and Their Applications to Cancer

Computational Methods for the Analysis of Single-Cell Transcriptomic Data and their Applications to Cancer

A dissertation presented

by

Daphne Tsoucas

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

June 2018

Dissertation Advisor: Professor Guo-Cheng Yuan Daphne Tsoucas

Computational Methods for the Analysis of Single-Cell Transcriptomic Data and their Applications to Cancer

Abstract

Single-cell sequencing methods have allowed for a closer view into the heterogeneity of cell populations, down to the level of the individual cell. In particular, single-cell transcriptomic data provides a detailed map of the diverse gene expression profiles present throughout a sample. These new methods have shown promise in many fields, especially the field of cancer genomics. However, despite the technological advances, many computational challenges remain.

In Chapter 1, we provide an overview of single-cell technological and computational methods through the perspective of cancer genomics. Cancer is a notoriously heterogeneous disease whose characteristics can differ greatly from person to person and cell to cell, making treatment a very challenging proposition. Single-cell methods allow us to dissect within-tumor heterogeneity, opening up the possibility of developing individualized therapies that target specific cancer cell subpopulations within a patient.

We next address two current challenges associated with the analysis of single-cell transcriptomic data. The first challenge is the difficulty of detecting rare cell populations. Current clustering methods either only detect prevalent cell types,

or specifically target only the detection of rare cell types. In Chapter 2, we develop a clustering method, GiniClust2, that can accurately identify both rare and common cell types using a novel, cluster-aware ensemble method that combines clustering results from rare and common cell-type-specific clustering methods.

The second challenge we address is the inference of cell-type composition from bulk gene expression data when single-cell data is unavailable. In Chapter 3, we propose a gene expression deconvolution method that estimates cell-type composition using a gene signature derived from single-cell data. We also introduce a novel dampened weighted least squares algorithm (DWLS) for estimation that adjusts for biases present in existing estimation methods, to create a method that can more accurately detect diverse cell types.

Finally, in Chapter 4, we conclude with two applications of single-cell RNA-sequencing data analysis to the discovery of immune response mechanisms in cancer. This highlights the impact such methods can have on helping cancer immunologists identify drug targets and assess their effects.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to first and foremost thank my advisor, Dr. Guo-Cheng Yuan, for providing me with support and guidance throughout my graduate school career. From introducing me to the bioinformatics field, to helping me publish my first, first-author paper, to advising me on my next career move, he has been instrumental to my development as a scientist. He has encouraged me to think independently and creatively when faced with challenges, all while being available to give advice anytime I needed.

I want to thank the rest of my committee, Drs. John Quackenbush and Martin Aryee, for providing a fresh perspective on my research. Their insights have helped strengthen my manuscripts and allowed me to make consistent progress with my research.

I want to thank everyone at the Guo-Cheng Yuan lab for fostering such a welcoming and supportive environment. They were willing to help with anything and everything, and taught me nuances of the field that would have taken me much longer to figure out myself. I will especially miss the group lunches where I gained new perspectives on so many things outside the realm of bioinformatics.

I want to thank my family for instilling in me a strong sense of scientific curiosity and encouraging my passions since before I can remember. Last but not least, I want to thank David Yang for being there every step of the way. I am so lucky to have someone who shares my goals and passions for not only research, but also so many other things in life.

Recent Progress in Single-Cell Cancer Genomics

Daphne Tsoucas and Guo-Cheng Yuan

A modified version of this text is published in *Current Opinion in Genetics and Development*, Volume 42, February 2017, Pages 22-32.

1.1 Abstract

The advent of single-cell sequencing has been revolutionary to the field of cancer genomics. Perfectly suited to capture cancer's heterogeneous nature, single-cell analyses provide information bulk sequencing could never hope to uncover. Many mechanisms of cancer have yet to be fully understood, and single-cell approaches are showing promise in their abilities to uncover these mysteries. Here we focus on the most recent single-cell methods for cancer genomics, and how they are not only providing insights into the inner workings of cancer, but are also transforming individualized therapy and non-invasive monitoring and diagnosis.

1.2 Introduction

Genomic analysis has been widely applied in cancer studies. The identification of genomic, epigenomic, and transcriptomic changes in cancer has led to precise classification, biomarker discovery, and mechanical understanding of cancer, and has played an essential part in cancer diagnosis, monitoring, and treatment [1]. However, until recently, bulk sequencing has been the only viable option for cancer genomic analysis. One major limitation is that bulk sequencing cannot detect the heterogeneity within a tumor. This limitation has important

clinical consequences. For example, cancer is often composed of multiple clones, and the most aggressive clone is difficult to identify and target since it may not be the one that metastasizes.

Throughout every stage of cancer, cells accumulate distinct mutations, which define the further evolution and progression of the disease. It is commonly viewed that cancer originates from an accumulation of mutations in oncogenes and tumor suppressors such that cell growth becomes unregulated and invasive [2]. The progeny of these cells in turn accumulate further mutations and selective pressures drive clonal evolution. The cancer will eventually metastasize, spreading to other parts of the body through the circulatory or lymphatic systems to form further distinct subpopulations. In addition, targeted cancer therapy may drive further evolution and eventually lead to drug resistance.

The recent advent of single-cell sequencing has revolutionized the field of cancer genomics, opening the door to a vast number of possibilities. From the ability to resolve intra-tumoral heterogeneity [3-6], map clonal evolution [7,8], and track the development of therapy resistance [3,9], to the capacity to analyze rare tumor cell populations such as tumor stem cells and circulating tumor cells [10-12], single-cell techniques have opened new avenues for cancer research. A better understanding of the mechanisms of cancer can in turn inform more effective and personalized treatments.

In this chapter, we review recent progress in single-cell analysis techniques and their applications in cancer genomics (Fig. 1.1), focusing on topics that have not been covered by previous reviews [13-16].

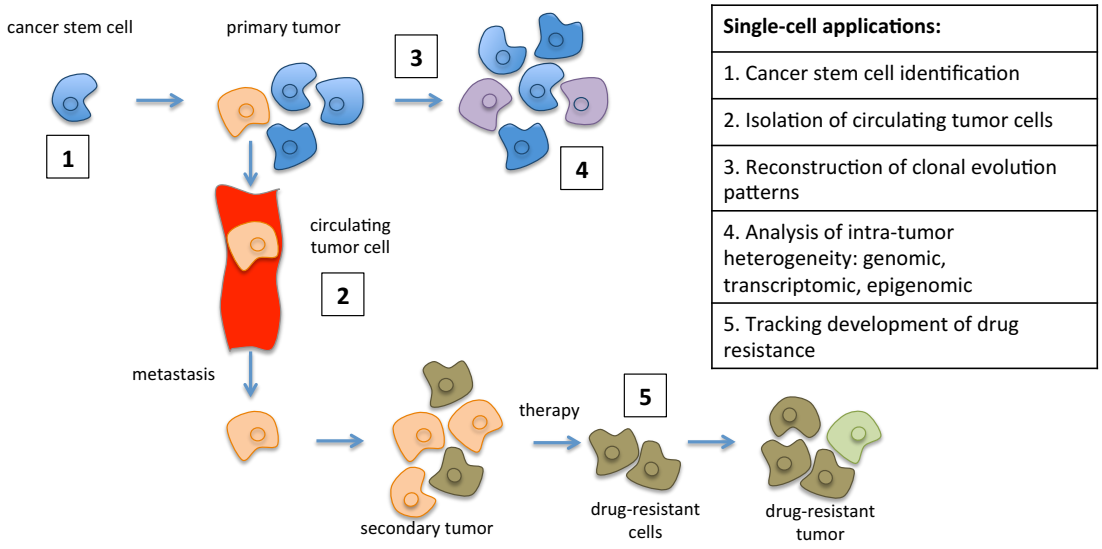


Figure 1.1. Single-cell methods provide novel insights into every stage of cancer progression, from primary tumor development to metastasis, to the development of drug resistance.

1.3 Intra-tumor genome sequence heterogeneity

Understanding the genomic heterogeneity of cancer cells first and foremost necessitates methods for single-cell DNA sequencing. The earliest developments for single-cell genomics involve whole genome amplification, providing ample amounts of DNA for subsequent sequencing. Degenerate oligonucleotide primed PCR (DOP-PCR) is appropriate for CNV detection, with low coverage but uniform amplification [17]. Multiple displacement amplification (MDA) is a linear amplification method capable of higher coverage through the use of Phi-29 polymerase, making it suitable

for SNP detection [18]. MALBAC (multiple annealing and looping-based amplification cycles) combines MDA and PCR for a high coverage, uniform amplification method suitable for either CNV or SNP detection [19]. These methods have been extensively applied to the characterization of intra-tumor CNVs and SNPs in various cancer types.

However, one major limitation of the aforementioned methods is that spatial information is lost as soon as single cells are isolated. Such information is integral to understanding the interaction of the cell with its micro-environment and may prove valuable for evaluating drug responsiveness. Recently, a new technology, STAR-FISH (specific-to-allele PCR-FISH) [3], has been developed which can detect the spatial distribution of both SNVs and CNVs using a combination of *in situ* PCR and FISH. PCR primers are built to target mutant and wild type mRNAs, one gene at a time. Amplification is followed by hybridization of fluorophores to a 5' overhang built into each probe. Janiszewska et al. use their method to study the commonly reported His1047Arg mutation in *PIK3CA* and *ERBB2* (also commonly known as *HER2*) amplification in *HER2+* breast cancer, before and after chemotherapy. They were able to identify changes in mutational frequency of mutated cells, which help gain an understanding of the development of drug resistance in *HER2+* breast cancer [3]. When combined with longitudinal analysis, this method was used to pinpoint migratory cells [3]. Currently, the technology can only be used to detect the location of known mutations.

The introduction of spatial methods to single-cell cancer genomics allows genomic heterogeneity to be mapped in space. This presents new opportunities in

studying cell-to-cell interactions, and in identifying migratory cancer cells and their roles in metastasis.

1.4 Intra-tumor transcriptomic heterogeneity

Like single-cell genome analysis, the first efforts in single-cell transcriptomics were in the amplification of the transcriptome to allow for quantification and sequencing of the transcriptome. Whole transcriptome amplification methods include poly-A tailing methods [20] and template-switching methods like Smart-seq [21]. Targeted gene expression profiles can also be quantified by multiplexing qPCR with high sensitivity [22].

In conjunction with single-cell RNA sequencing and qPCR, these methods have been used in various cancer studies. Cancer-specific gene expression signatures and alternative-splicing events have been identified for melanoma [21]. Gene expression signatures have led to the identification of cancer cell types, such as cancer stem cells [23]. The relative contributions of clonal evolution and multi-lineage differentiation in transcriptomic heterogeneity have been studied in the context of colon cancer [24].

Recent technologies have been developed to quantify gene expression levels *in situ*, thereby preserving spatial information. Here we review recent single-cell spatial transcriptomic methods and their potential for future use in cancer studies. These methods share the same fundamental principle as single-molecule fluorescence *in situ* hybridization (smFISH), whereby fluorescently-labeled DNA

oligonucleotide probes are hybridized to their complementary target mRNA, and are then identified via fluorescence microscopy [25,4]. The newer techniques described below have greatly enhanced detection efficiency and throughput.

SeqFISH (sequential FISH) is an adaptation of smFISH that uses sequential hybridization to allow for multiplexing [26]. Each mRNA is assigned a unique sequence of fluorophores that create a barcode through which each mRNA can be decoded. In the first round of this process, probes that target the same mRNA are labeled with the same fluorophore. These probes are hybridized, imaged, and then purged. In the next round, the same probes are labeled with a different fluorophore, and the same sequence of steps is followed. Several rounds of this create a unique barcode of colors for the particular mRNA. Each probe set targeting a particular mRNA is labeled with a unique barcode in this way. For F fluorophores and N hybridization rounds, this means F^N mRNAs can be visualized. As this number scales up rapidly with an increasing number of fluorophores and hybridization rounds, this technique can potentially be used to sequence all known genes with limited numbers of fluorophores and hybridization cycles. The authors initially applied this method to immobilized yeast cells and mouse embryonic stem cells [26], but have since extended the method so that it is now applicable to deep tissues such as the brain [27].

MERFISH (multiplexed error-robust FISH) is a similar approach that also allows for error correction by using a smart choice of barcodes [28]. Specifically, barcode sequences are chosen to include only those that are separated by a certain Hamming distance (Hamming distance=number of changes in a barcode sequence

required to transform one sequence into another). Since not all possible barcodes encode a particular mRNA, this encoding scheme provides a means to error detection and correction. The authors use this approach to simultaneously measure 1001 genes in human fibroblast cells. Two fluorophores and 14 hybridization rounds allow all encoding sequences to be separated by a Hamming distance of 2 [28]. Of note, these authors show that their barcode design helps reduce the error rate significantly.

FISSEQ is another *in situ* technique which is based on sequencing. RNA is first reverse-transcribed and amplified [29]. The amplicons are crosslinked to the cellular matrix and sequenced by using the SOLiD SBL (sequencing-by-ligation) technique. The method has been applied to a simulation of the wound healing response in primary fibroblasts where the authors found differentially expressed genes between migrating cells and contact-inhibited cells [29]. Such a method could similarly be applied to find differentially expressed genes in migratory vs. non-migratory tumor cells.

In addition, transcriptomic profiles can also be measured *in vivo* by using a technology called TIVA (transcriptome *in vivo* analysis). In this approach, a photoactivatable biotin-labeled TIVA-tag is inserted into live cells, attached to mRNA upon selective photoactivation, and recaptured via streptadavin beads. The captured mRNA is subsequently sequenced [30]. TIVA was used on live mouse and human brain tissue, as well as mouse brain cells in culture. A comparison of live and culture mouse brain cells shows significant differences in gene expression levels, emphasizing that cells removed from their natural environment may not be

representative of the same cells *in vivo* [30].

The aforementioned methods give increasingly multiplexed ways of spatially resolving gene expression patterns. While most of the applications to date have been limited to cell culture, we expect that soon they will be applicable to tissue samples. If they can be adapted to tumor cross-sections, these methods will have great impact on investigating the cancer progression path. For example, the location of tumor-like stem cells could be mapped within the tumor. If longitudinal measurements are taken, cell migratory paths may be traced.

1.5 Intra-tumor epigenetic heterogeneity

Epigenetics plays an important role in regulating gene expression in cancer, and exploring the heterogeneity of epigenetic patterns may aid in understanding underlying transcriptomic heterogeneity. As a dynamic process, epigenetics may contribute to the phenotypic plasticity of cancer cells, for example aiding in the differentiation of cancer stem cells [31]. Studies have shown abnormally low levels of global DNA methylation along with hyper-methylation in specific regions, such as tumor suppressor gene promoter regions, giving strong evidence for the role of epigenetic aberrations in cancer proliferation [32].

The characterization of intra-tumor epigenetic heterogeneity has been less extensively studied due to its technical difficulty. Nonetheless, multiple epigenetic methods have recently been adapted for single-cell purposes. Determining DNA methylation patterns has traditionally been performed by bisulfite sequencing

methods, but bulk techniques have performed poorly in the single-cell setting due to DNA degradation during bisulfite conversion. Methods have adapted bisulfite sequencing for single-cell, including scRRBS (reduced representation bisulfite sequencing) [33] and PBAT (post-bisulfite adapter-tagging) [34]. In each, a modified version of bisulfite sequencing is applied to each cell individually. ScRRBS mitigates the issue of high DNA loss by replacing the multiple purification steps prior to bisulfite sequencing with a single-tube reaction. A restriction enzyme that recognizes CpG islands is used to cut the genome, selecting CpG island regions for subsequent conversion and sequencing. By sequencing only these regions, this method provides low-cost but low-coverage sequencing [33]. ScRRBS has been applied to human hepatocellular carcinoma tissue in conjunction with simultaneous transcriptome sequencing (discussed in greater detail in the next section) [5]. Methylation levels at all CpG sites were measured and subsequently used to cluster the tissue into two subpopulations via unsupervised hierarchical clustering. A large amount of heterogeneity was found between and within these subpopulations. Interestingly, when the same clustering method was applied using CNV patterns, an identical clustering was found [5].

PBAT is a more unbiased whole-genome approach that addresses the issue of bisulfite-conversion-induced DNA degradation by performing suitable library preparation after bisulfite sequencing. Traditionally, adapter-tagging is performed before bisulfite conversion and sequencing templates become degraded, but switching the order of these events alleviates this problem [34,35]. In an application of PBAT, differential methylation of distal regulatory elements was

discovered in mouse embryonic stem cells [35]. These elements cannot commonly be captured by scRRBS, making it promising for higher-coverage cancer methylation studies.

Chromatin structure also plays an important role in gene regulation. Most transcription factors can only bind to open chromatin regions, whereas a small number of pioneer factors may bind to closed chromatin, opening it up so that other factors can bind. The genome-wide landscape of chromatin accessibility can be measured by using either ATAC-seq (assay for transposase-accessible chromatin) [36,37] or DNase-seq [38]. The difference between these two methods is the DNA-cutting enzymes, corresponding to Tn5 and DNase I, respectively. Both methods have been adapted to single-cell analysis. Two single-cell methods have modified ATAC-seq. A combinatorial indexing approach [36] tags nuclei with unique barcodes so they can then be grouped and processed together. Groups of nuclei are placed in wells, barcoded, and then passed through a second set of wells and barcoded again. Given that each nuclei is highly likely to pass through a unique combination of wells, the barcoding is overwhelmingly cell-specific [36]. In a microfluidic approach [37], cells are captured and assayed separately. The microfluidic technique has been used to find a high variability of transcription factor motif accessibility in cancer cell lines [37]. For DNase-seq, a single-cell method called Pico-Seq [38] sorts cells using FACS before DNaseI treatment. To prevent a large loss of digested DNA during subsequent library preparation, circular carrier DNA is added after digestion. This DNA will not be amplified in the PCR that follows due to its incompatibility with the adaptor ligation process. Of note, the authors

applied their method to formalin-fixed paraffin-embedded follicular thyroid cancer patient tissue and, in one patient, found a SNV that prevents the binding of tumor suppressor protein p53 [38].

The aforementioned methods have started to provide new mechanistic insights into cancer heterogeneity. In addition, two additional single-cell methods, Hi-C and ChIP-seq, have been recently developed and show potential for use in future cancer epigenetic studies. A type of chromosome conformation capture that quantifies interactions between genomic loci, Hi-C can be used to find *trans*-regulatory elements and their targets [39]. ChIP-seq, which characterizes interactions between DNA and DNA-binding proteins, can determine transcription factor-regulatory element interactions [40].

1.6 Simultaneous multiple omic analysis

Ideally, the different omic approaches should be applied to study a particular tumor so that the information can be integrated. However, this multiple-omic approach is much more technologically challenging. We review some recent studies in this direction.

Simultaneous transcriptomic and genomic sequencing for single cells has recently been achieved by the G&T-seq method [6]. Cells are first isolated and lysed to release mRNA and genomic DNA. Poly-A mRNA is then separated from genomic DNA through the use of biotinylated oligo-dT primers coupled with streptavidin-coated magnetic beads. The primers are hybridized directly to the poly-A tail, and

subsequently recruited by streptavidin-coated magnetic beads through a strong biotin-streptavidin interaction. Standard single-cell techniques can then be used to separately sequence the isolated mRNA and genomic DNA [6].

The ability to measure transcriptomic and genomic landscapes in the same cells opens a window into understanding the direct effect of genomic variation on transcriptomic variation. Macaulay et al. use their method on HCC38 breast cancer cells to discover the chromosomal rearrangement responsible for the fusion transcript *MTAP-PCDH7*, found in a majority of HCC38 cells [6]. They also conclude that a trisomy found in a subset of HCC38_BL (B lymphoblastoid) cells results in proportionally increased mRNA expression in these cells [6]. To date, the application of G&T has been limited to cell lines; however, it provides hope to analyze the direct effect of copy number variants on transcript levels in tumor samples in the near future.

An extension of this idea of concurrent sequencing has been implemented via the scTrio-seq method [5]. This technique simultaneously sequences not only the genome and transcriptome, but the DNA methylome as well. In this method, separation of genomic DNA and mRNA is performed through centrifugation of lysed single cells, where a special centrifugation technique allows for the separation of cytoplasm from intact nuclei. The mRNA found in the cytoplasm is sequenced separately from the genomic DNA, which is subjected to scRRBS, providing methylomic and genomic data. The ability to simultaneously quantify genomic, transcriptomic, and epigenomic changes in the same cells has provided new insights into the gene expression regulatory mechanisms. The authors use their method in

the analysis of the heterogeneity of human hepatocellular carcinoma. Their results corroborate those of Macaulay et al. in that CNV gene dosage is found to have a proportional effect on transcript levels. DNA methylome results, however, show that CNVs have no similar effect on methylation levels [5].

The transcriptome is often used as a proxy for protein levels, as single-cell proteomic analyses have not reached the degree of multiplexing that single-cell transcriptomic analyses have. However, mRNA molecules have shorter half-lives than proteins, and previous studies have shown that the mRNA and protein levels may not correspond well [41]. However, their relationship remains unclear at the single-cell level. Recently, Darmanis et al. have developed a new technique to simultaneously measure the transcriptomes and proteomes of single cells [42]. This is achieved by the splitting of cell lysate and independent processing of each fraction, much like the methods above. The mRNA fraction is subjected to qPCR, and the protein fraction to proximity extension assay (PEA). During PEA, pairs of oligo-labeled antibodies bind to target proteins, where each pair's oligos are complementary to one another and bind upon being brought in proximity, creating a PCR amplicon, which is then quantified with PCR. The authors apply this technique to quantify cancer pathway proteins that were determined a priori to be of relevance in BMP4-treated glioblastoma cells, and find poor correlation between mRNA and protein levels in these cells. They conclude that protein levels are better predictors of treatment response, leading to the conclusion that perhaps single-cell transcriptomic methods are not sufficient in determining treatment response [42].

1.7 Computational methods for analyzing single-cell genomic and transcriptomic data

With the advent of single-cell technologies comes the necessity for new computational methods to process the data collected. These methods fall into two categories. First are methods that modify bulk sequencing methods to adjust for nuances unique to single-cell data: sparse, noisy data that lacks technical replicates. The second set of methods implement new applications possible only with single cell data. Here we mention methods of the second variety that are of special relevance to cancer genomics. Other methods are extensively covered in previous reviews [5,6].

1.7.1 Inference of spatial patterns

As described above, exciting technologies have been developed to profile single-cell gene expression patterns *in situ*. Computational methods are still lacking to systematically detect the spatial patterns and classify samples using such patterns.

In some cases, spatial patterns can be inferred by integrating single-cell RNA-seq data collected from isolated cells with *in situ* expression patterns of a small number of landmark genes [43,44]. Location of the cells is inferred through correlation between their expression levels and those of the *in situ* data landmark genes. This approach has been used in developmental biology for the analysis of

embryos, where cells are predictably distributed across the dorsal-ventral and animal-vegetal axes [43]. An analogous method has been used to map cells back to annelid brain regions [44]. However, there is a possibility for difficulties in measuring spatial heterogeneity in tumors due to their typical lack of spatial patterning [43].

1.7.2 Pseudo-time ordering with bifurcation

Single-cell RNA-seq data is only capable of producing a static view of gene expression levels within cells. Pseudo-time ordering computational methods now allow for a window into continuous changes in gene expression levels, which have thus far given insights into the transcriptional kinetics of cell differentiation. Making the assumption that cells at various stages of differentiation can be found in one scRNA-seq dataset, a time series of transcriptional changes is produced, onto which each cell is mapped. Applying these methods to cancer data can be used to track genes activated at various stages of differentiation from cancer stem cell to matured cancer cell.

Monocle was the first of a series of pseudo-time-ordering algorithms, and uses a combination of dimensionality reduction and a minimal spanning tree (MST) algorithm to build a differentiation trajectory [45]. Monocle2 has since been released, which uses reverse graph embedding and is capable of handling data from much larger scRNA-seq experiments than before [46]. TSCAN (pseudo-Time reconstruction in Single-Cell RNA-seq Analysis) was built as an improvement upon

the original Monocle method, reporting more robust results. Instead of creating an MST on all cells, cells are first clustered via hierarchical clustering, and these clusters are used as the MST inputs [47]. A reduced space from which to build a trajectory allows for more stable inference, hence more robust final results. Waterfall is a similar method that also conducts clustering before MST creation [48]. An alternative approach to reconstruct pseudo-time is by fitting the data by a principal curve [49]. This method has been applied to analyzing CyTOF data.

Cell differentiation often involves bifurcation, where two or more distinct cell-types may emerge from a common stem/progenitor cell population. If the temporal information is known, SCUBA can be used to detect bifurcation events [49]. However, in most cases, the temporal information is unavailable. Some pseudo-time methods also build bifurcation events into their models. Instead of assuming one trajectory for all cells, these methods allow for a branching trajectory to account for differentiation into multiple cell types. One method, Wishbone [50], is an updated version of Wanderlust [51] with the added ability to account for bifurcations. The initial Wanderlust algorithm represents cells as nodes in a graph, where the shortest path between two nodes represents their phenotypic distance. An early cell is chosen and distances are calculated between each cell and the early cell. To adjust for the fact that longer paths are noisier than shorter paths, random waypoint cells are introduced, and each cell's position is iteratively refined with respect to these waypoint cells. Repeating the graph-building process several times and averaging cell positions from all these graphs mitigates "short circuits," or edges that occur erroneously between developmentally distant cells [51]. Wishbone

updates this algorithm by introducing a step to identify branch points through discrepancies in waypoint distances. Additionally, "short circuits" are avoided via a different approach, where the initial graph is rebuilt in a reduced space to remove noise [50].

The ability to order cells of complex lineage relationship may have important applications in development. Already, these methods have been used to study the development of cells such as human B lymphocytes [51] and human neural cells [48]. In the future, pseudo-time ordering may be used in mapping the altered mechanisms of cell development in cancer.

1.7.3 Rare cell-type detection

The detection of rare cell types is pertinent to cancer, where the ability to identify circulating tumor cells (CTCs), cancer stem cells, or drug resistant cells will have important clinical implications. Most clustering methods to date are only able to identify major cell groups.

RaceID [52] is a method aimed at detecting rare cell types from scRNA-seq data. Cells are first clustered into major groups by k-means. Outliers of each cluster, which are determined not to be a cause of technical or biological noise, are then grouped into rare cell clusters based on transcriptome correlation [52]. RaceID was recently updated for more robust clustering, where the newer RaceID2 [10] replaces k-means with k-medoid clustering. Grün et al. have integrated RaceID2 into a stem-cell detection algorithm named StemID, which uses the

identified cell clusters to guide inference of a lineage tree. Stem cells are then defined by this differentiation trajectory. In this manner, the authors were able to classify stem cells from mouse bone marrow cells, and predict novel pancreatic pluripotent cells [10].

GiniClust [11] is an alternative approach for detecting rare cell types, by using an innovative approach to choose genes that are likely to be associated with rare cells types, using a statistic called the Gini index. The high Gini genes are identified and subsequently used as input into DBSCAN (density-based spatial clustering of applications with noise) [53]. The authors used this approach on both scRNA-seq and qPCR data. Among other findings, they were able to discover a novel stem cell type characterized by a high expression of *ZSCAN4* in mouse embryonic stem cells, and were able to identify rare normal cells in glioblastoma primary tumor samples [11]. While GiniClust is not able to simultaneously detect common cell types, our method GiniClust2 [12], discussed in Chapter 2, uses a cluster-aware weighted consensus clustering algorithm to combine results from GiniClust and a clustering algorithm designed to detect common cell types into a final clustering result that includes both common and rare cell types.

1.7.4 Clonal evolution inference

Cancer undergoes a process of clonal expansion and selection that can be inferred through single-cell sequencing data using computational tools. Two such methods are OncoNEM (oncogenetic nested effects model) [7] and SCITE (single cell

inference of tumor evolution) [8], which create tumor lineage trees from the single-cell sequencing data. Building lineage trees can guide understanding of the development of therapy resistance; if a sample is taken post-treatment, we can infer a timeline of mutational events that take place before, during and after treatment. Furthermore, these methods can identify mutations that occur early on in tumor development and are propagated throughout each subsequent clone, and guide treatment targeted towards these mutations. These two methods differ in their algorithms—SCITE uses Markov chain Monte Carlo and OncoNEM uses a heuristic search—but importantly, both implement a probabilistic model instead of the traditional maximum parsimony model. Single-cell sequencing data suffers from a large amount of technical error as compared to bulk data that can easily be propagated through subsequent tree-building methods. Using maximum likelihood principles, SCITE and OncoNEM build sequencing error estimation into their models to account for this [7,8].

1.8 Biological insights obtained through single-cell analyses

1.8.1 Cancer stem cells

The cancer stem cell hypothesis postulates that there exists a sub-population of self-renewing cells with differentiation potential that serves to initiate and maintain the larger tumor cell population. These cells are estimated to make up less than 1% of the total tumor cell population [54]. Single-cell techniques have

provided a powerful tool for identifying and molecularly characterizing cancer stem cells.

As a starting point, Patel et al. [23] use scRNA-seq to analyze the transcriptomes of cells from 5 human glioblastomas in search of glioblastoma stem-like cells (GSC). The authors derive a transcriptome signature that corresponds with "stemness" by comparing the transcriptomes of GSCs and DGCs (differentiated glioblastoma cells) modeled in culture. They then use this signature to identify GSCs *in vivo*, and find a continuous gradient of stemness-indicating gene expression [23]. Lawson et al. similarly identifies stem-like cells in metastatic breast cancer tumors by a stem-cell-like gene expression signature [55]. Early stage metastases contain these stem-like cells, while later stage metastases contain cells closer to primary tumor cells in gene expression, supporting the theory that as cancer progresses, tumor cells with stem-like properties initiate and propagate metastatic tumors [55].

1.8.2 Circulating tumor cells

Single-cell analysis has also provided a powerful tool for the detection and characterization of circulating tumor cells, which are cells that are shed from the tumor into the vasculature or lymphatics and circulate through the bloodstream. Monitoring the presence of CTCs may be used to track the evolution of tumors over time with a simple series of blood tests. However, at an estimated frequency of as little as 1 in 10^9 of all blood cells [56], it is extremely challenging to capture and analyze these cells. Because of the large amount of heterogeneity in these cells,

which may derive from the original tumor or any metastases, single cell methods are necessary. The rarity of these cells requires tools for isolation from hematological cells. A common method involves identifying circulating tumor cells (CTCs) through the presence of EpCAM (epithelial cell adhesion molecule)—found in epithelial cells but not blood cells—on the surface of the cell. Separation of these cells from the blood is then performed using magnetic beads coated with anti-EpCAM antibody. Other recent methods have been developed to overcome a major limitation of this method: the expression of EpCAM is variable from tumor cell to tumor cell, especially those in the epithelial-mesenchymal transition. These alternative methods include isolation of CTCs by microscopic imaging, cell size, and passive capture through removal of all other blood cells.

Genomic and transcriptomic profiling of CTCs have been applied to studying cancer progression. Ni et al. elucidate the pathway of metastasis in lung cancer through the whole-genome sequencing of CTCs from lung cancer patients [57]. As these circulating tumor cells reproducibly share similar CNV patterns to the same patient's metastatic tumors, the CNV patterns of these CTCs can be used as proxies for the metastatic tumors. These CNV patterns are different from those of the primary tumors, suggesting that metastasis may occur through a set of copy number changes [57]. Several papers point to the sequencing of CTCs as a tool for noninvasively tracking the development therapy resistance. Miyamoto et al. and Dago et al. study the progression of prostate cancer over the course of androgen receptor inhibitor treatment, discussed in the next section [58,59].

1.8.3 Development of therapy resistance

The ability to detect mutations at a single-cell level has lead to yet another possibility: tracking the development of cancer therapy resistance. The main approach towards this goal is longitudinal single-cell measurements before and after various therapies. A common method for treating cancer is chemotherapy before a round of targeted therapy; longitudinal data therefore may consist of measurements before and after each of these events. Noting differences in mutational frequencies over time gives insight into how tumor cells respond to therapy and the mechanisms by which they develop resistance. These studies may in addition be used to validate two prevalent theories of therapy resistance: adaptive resistance, in which a mutation present at low frequency in the original population is selected for during therapy and rises in frequency, or acquired resistance, whereby resistance-conferring mutations arise as a consequence of therapy.

One study evaluates the response of $BRAF^{V600E}$ melanoma to treatment with RAF or combined RAF/MEK inhibitors in both cell culture and tissue [9]. A comparison of scRNA-seq data from biopsies taken from patients before and after treatment with either RAF or RAF/MEK inhibitors finds that post-treatment tissues contain a higher proportion of cells overexpressing *AXL*, a known marker of resistance. A follow-up experiment in melanoma cell lines, in which cells are treated to increasing doses of RAF/MEK inhibitors, also reveals an increase in AXL-positive cells. These AXL-positive cells preexisted in the treatment-naïve sample and were

selected for by treatment, a demonstration of the adaptive resistance mechanism [9].

The Dago et al. and Miyamoto et al. studies mentioned above use CTC tracking to analyze the development of resistance to androgen receptor (AR)-targeted therapy in prostate cancer patients [58,59]. Through whole-genome sequencing of CTCs before and after treatment, the former find the emergence of two distinct resistant subpopulations with *AR* amplification. One of these subpopulations is found to be a descendant of a clone found in the therapy-naïve population, indicating support for the adaptive resistance hypothesis [58]. Miyamoto et al. use scRNA-seq of CTCs to show the acquisition of heterogeneous resistance-conferring changes in the AR-independent Wnt signaling pathway [59]. Both studies demonstrate the relevance of CTCs in the non-invasive monitoring of therapy resistance.

Authors of the aforementioned Janiszewska et al. paper [3] use their STAR-FISH technique to study the implications of chemotherapy in the development of resistance to subsequent ERBB2 (HER2)-targeted trastuzumab therapy in *HER2*$^+$ breast cancer patients. *HER2* amplification and frequency of the His1047Arg mutation in *PIK3CA* were observed before and after chemotherapy in *HER2*$^+$ breast tumor samples. Chemotherapy is found to result in an increased frequency of *PIK3CA* mutants (known to be a determinant of resistance to trastuzumab) and a decreased frequency of *HER2* amplification (giving trastuzumab less target sites). These results suggest that trastuzumab may be ineffective for patients who have already received chemotherapy. The spatial information provided by the STAR-

FISH method may also be informative in studying resistance, as the authors found that chemotherapy increases the dispersion of cancer cells with the *PIK3CA* mutation. This increased dispersion may be an indicator of poor prognosis [3].

A new study extends this type of study to single-cell proteomic data [60]. Wei et al. collect proteomic data on 12 proteins and phosphoproteins in cells of a patient-derived *in vivo* brain cancer glioblastoma model before and after treatment with an mTOR kinase inhibitor. Correlations between protein expression levels are then used to build signaling networks, and these networks are compared pre- and post- targeted therapy. The drug decreases mTORC1/C2 signaling (the intended target), but upon reaching a state of resistance, the signaling is reactivated, once again an example of adaptive resistance. In addition, upon reaching a state of resistance, new correlations can be seen in the ERK/Src pathways. This is an indication that increased signaling in these pathways may promote downstream mTOR signaling, and consequently that an effective targeted therapy must simultaneously target both mTOR and ERK/Src [60].

1.9 Conclusion

Single-cell biology is a fast-evolving field. As discussed in the paper, a lot of the technical and computational development has been made in just a few years. These methods have greatly empowered researchers to systematically interrogate the cellular heterogeneity within a tumor especially in terms of spatial heterogeneity and multi-omics integration. All the methods reviewed here share a

common goal: improving our understanding of tumor cell heterogeneity for the purpose of informing personalized cancer treatment.

Studying intra-tumoral heterogeneity and the spatial orientation of subclones in the primary tumor via new spatial transcriptome methods and simultaneous multiple omic sequencing will allow for the proper drug targeting of the subclones. Examining the nature of stem-like tumor cells and the transcriptomic mechanisms required to give rise to new tumor populations will give clarity to the origination of metastases. Targeting these stem-like cells could hamper the spread of cancer throughout the body. Being able to isolate and longitudinally sample CTCs will permit non-invasive diagnosis and monitoring over the course of treatment. Treatment approaches can be constantly updated upon tracking the response and evolution of CTCs throughout treatment. Finally, treatment resistance can be prevented with a more accurate modeling of the development of resistance to current drugs.

Much work remains to make these possibilities realities. But as single-cell sequencing methods continue to become cheaper, capable of higher coverage, and able to process a greater number of cells faster, no doubt these goals will become more and more attainable.

Method Type	Specific Methods	Application to Cancer Genomics	Refs
Experimental Methods			
Single-cell whole genome amplification	DOP-PCR, MDA, MALBAC	Used in conjunction with next-generation sequencing to detect intra-tumor CNVs and SNPs.	[7,8,9]
Single-cell spatial genomics	STAR-FISH	Detects the spatial distribution of intra-tumor CNVs and SNPs. Can be combined with longitudinal analysis to reveal migratory cells.	[10·]
Single-cell transcriptome amplification	Smart-seq, Tang et al. method, single-cell qPCR	Identifies cancer-specific gene expression signatures, cancer cell types, alternative-splicing events.	[11,12,13]
Single-cell spatial transcriptomics	smFISH, SeqFISH, MERFISH, FISSEQ, TIVA	Can provide spatially-resolved gene expression signatures in tumors. Has potential applications in tracing cell migratory paths and locating tumor-like stem cells.	[16,17··,18·,19,20·,21,22]
Single-cell DNA methylomics	scRRBS, PBAT	Enables the discovery of differential methylation in cancer cells. Potential for broadening understanding of phenotypic plasticity of cancer cells.	[25,26,27·]
Single-cell chromatin accessibility	ATAC-seq, Pico-Seq	Can give insight into the differential binding of transcription factors in cancer cells.	[29·,30·,31]
Chromosome conformation capture	Hi-C, ChIP-seq	Potential for understanding the mechanisms of cancer heterogeneity through mapping transcription factor-regulatory element interactions.	[32,33]
Simultaneous multiple single-cell omics	G&T-seq, scTrio-seq, Darmanis et al. method	Provides an integrated view of intra-tumoral heterogeneity through measuring direct interactions between genomic, transcriptomic, epigenetic, and proteomic variation.	[34··,27·,36·]
Computational Methods			
Single-cell spatial transcriptomic inference	Seurat, Achim et al. method	Infers cell location through scRNA-seq data and an *in situ* RNA reference map of several landmark genes, enabling mapping of intra-tumor spatial heterogeneity.	[37·,38]
Pseudo-time ordering	Monocle, TSCAN, Waterfall, SCUBA, Wanderlust, Wishbone	Projects gene expression values from a single time-point to a continuous trajectory over cell differentation. Potential use in understanding differentiation from stem-like cancer cell to matured cancer cell.	[39,40,41,42,43,44·,45]
Rare cell-type detection	RaceID, StemID, GiniClust	Potential use in the detection of circulating tumors cells and stem-like cancer cells.	[46,47,48]
Clonal evolution inference	SCITE, OncoNEM	Builds lineage trees for understanding evolutionary events such as the development of therapy resistance.	[50,51]

Table 1.1. A summary of relevant single-cell methods and their applications to cancer genomics.

GiniClust2: A Cluster-Aware, Weighted Ensemble Clustering Method for Cell-Type Detection

Daphne Tsoucas and Guo-Cheng Yuan

A modified version of this text is published in *Genome Biology*, Volume 19, May 2018, Page 58.

2.1 Abstract

Single-cell analysis is a powerful tool for dissecting the cellular composition within a tissue or organ. However, it remains difficult to detect rare and common cell types at the same time. Here, we present a new computational method, GiniClust2, to overcome this challenge. GiniClust2 combines the strengths of two complementary approaches, using the Gini index and Fano factor, respectively, through a cluster-aware, weighted ensemble clustering technique. GiniClust2 successfully identifies both common and rare cell types in diverse datasets, outperforming existing methods. GiniClust2 is scalable to large datasets.

2.2 Introduction

Genome-wide transcriptomic profiling has served as a paradigm for the systematic characterization of molecular signatures associated with biological functions and disease-related alterations, but traditionally this could only be done using bulk samples that often contain significant cellular heterogeneity. The recent development of single-cell technologies has enabled biologists to dissect cellular heterogeneity within a cell population. Such efforts have led to an increased

understanding of cell-type composition, lineage relationships, and mechanisms underlying cell-fate transitions. As the throughput of single-cell technology increases dramatically, it has become feasible not only to characterize major cell types, but also to detect cells that are present at low frequencies, including those that are known to play an important role in development and disease, such as stem and progenitor cells, cancer-initiating cells, and drug-resistant cells [61, 62].

On the other hand, it remains a computational challenge to fully dissect the cellular heterogeneity within a large cell population. Despite the intensive effort in method development [63-68], significant limitations remain. Most methods are effective only for detecting common cell populations, but are not sensitive enough to detect rare cells. A number of methods have been developed to specifically detect rare cells [69-72], but the features used in these methods are distinct from those distinguishing major populations. Existing methods cannot satisfactorily detect both large and rare cell populations. A naïve approach combining features that are either associated with common or rare cell populations fails to characterize either type correctly, as a mixed feature space will dilute both common and rare cell-type-specific biological signals, an unsatisfactory compromise.

To overcome this challenge, we have developed a new method, GiniClust2, to integrate information from complementary clustering methods using a novel ensemble approach. Instead of averaging results from individual clustering methods, as is traditionally done, GiniClust2 selectively weighs the outcomes of each model to maximize the methods' respective strengths. We show that this cluster-

aware weighted ensemble approach can accurately identify both common and rare cell types and is scalable to large datasets.

2.3 Results

2.3.1 Overview of the GiniClust2 method

An overview of the GiniClust2 pipeline is shown in Fig. 2.1. We begin by independently running both a rare cell-type detection method and a common cell-type detection method on the same data set (Fig. 2.1a). In a previous study [71], we showed that different strategies are optimal for identifying genes associated with rare cell types than for common ones. Whereas the Fano factor is a valuable metric for capturing differentially expressed genes specific to common cell types, the Gini index is much more effective for identifying genes that are associated with rare cells [71]. Therefore, we were motivated to develop a new method that combines the strengths of these two approaches. To facilitate a concrete discussion, here we choose GiniClust as the Gini-index based method and k-means as the Fano-factor based method. However, the same approach can be used to combine any other clustering methods with similar properties. We call this new method GiniClust2.

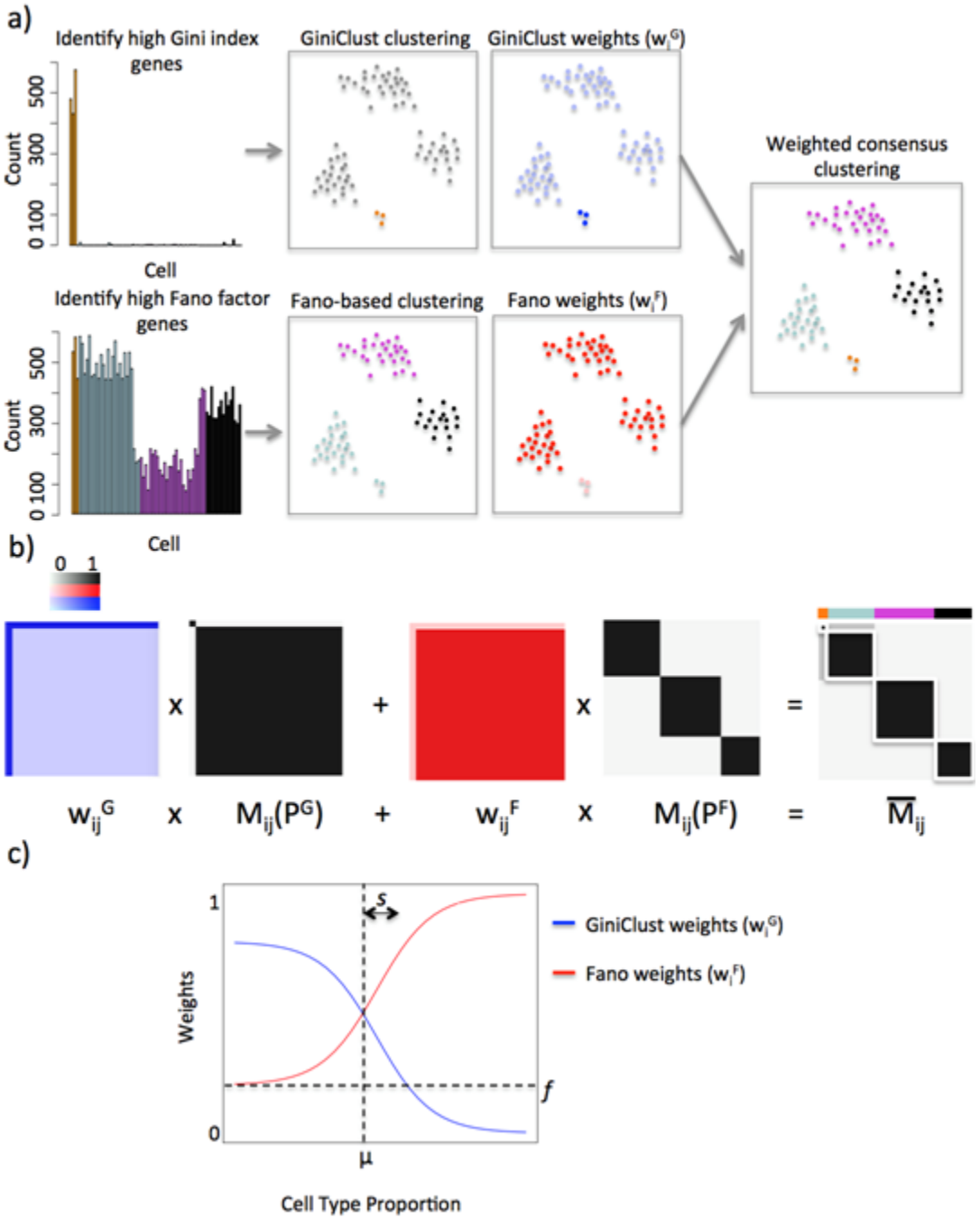


Figure 2.1. An overview of the GiniClust2 pipeline. (a) The Gini index and Fano factor are used (left), respectively, to select genes for GiniClust and Fano-based clustering (middle left). A cluster-aware, weighted ensemble method is applied to each of these, where cell-specific cluster-aware weights w_i^F and w_i^G are represented by the shading of the cells (middle right), to reach a consensus clustering (right). (b) A schematic of the weighted consensus association calculation, with association matrices in black and white, weighting schemes in red and blue, and final GiniClust2 clusters highlighted in white. (c) Cell-specific GiniClust and Fano-based weights are defined as a function of cell type proportion, where parameters μ, s, and f define the shapes of the weighting curves.

Our goal is to consolidate these two differing clustering results into one consensus grouping. The output from each initial clustering method can be represented as a binary-valued connectivity matrix, M_{ij}, where a value of one indicates cells i and j belong to the same cluster (Fig. 2.1b). Given each method's distinct feature space, we find that GiniClust and Fano-factor-based k-means tend to emphasize the accurate clustering of rare and common cell types, respectively, at the expense of their complements. To optimally combine these methods, a consensus matrix is calculated as a cluster-aware, weighted sum of the connectivity matrices, using a variant of the weighted consensus clustering algorithm developed by Li and Ding [73] (Fig. 2.1b). Since GiniClust is more accurate for detecting rare clusters, its outcome is more highly weighted for rare cluster assignments, while Fano-factor-based k-means is more accurate for detecting common clusters and therefore its outcome is more highly weighted for common cluster assignments. Accordingly, weights are assigned to each cell as a function of the size of the cluster to which the cell belongs (Fig. 2.1c). For simplicity, the weighting functions are modeled as logistic functions which can be specified by three tunable parameters: μ is the cluster size at which GiniClust and Fano-factor-based clustering methods have the same detection precision, s represents how quickly GiniClust loses its ability to detect rare cell types, and f represents the importance of the Fano cluster membership in determining the larger context of the membership of each cell. The values of parameters μ and s are specified as a function of the smallest cluster size detectable by GiniClust and the parameter f is set to a constant (Materials and Methods, Supplemental Information). The resulting cell-specific weights are

transformed into cell-pair-specific weights w_{ij}^{G} and w_{ij}^{F} (Materials and Methods), and multiplied by their respective connectivity matrices to form the resulting consensus matrix (Fig. 2.1b). An additional round of clustering is then applied to the consensus matrix to identify both common and rare cell clusters. The mathematical details are described in the Materials and Methods section.

2.3.2 Accurate detection of both common and rare cell types in a simulated dataset

We started by evaluating the performance of GiniClust2 using a simulated scRNA-seq dataset, which contains two common clusters (of 2000 and 1000 cells, respectively) and four rare clusters (of 10, 6, 4, and 3 cells, respectively) (Materials and Methods, Fig. 2.2a). We first applied GiniClust and Fano-factor-based k-means independently to cluster the cells. As expected, GiniClust correctly identifies all 4 rare cell clusters, but merges the two common clusters into a single large cluster (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). In contrast, Fano-factor-based k-means (with k=2) accurately separates the two common clusters, while lumping together all four rare cell clusters into the largest group (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). Increasing k past k=3 results in dividing each common cluster into smaller clusters, without resolving all rare clusters, indicating an intrinsic limitation of selecting gene features using the Fano factor (Supplemental Fig. S2.2a). We find this limitation to be independent of the clustering method used, as applying alternative clustering methods to the Fano-

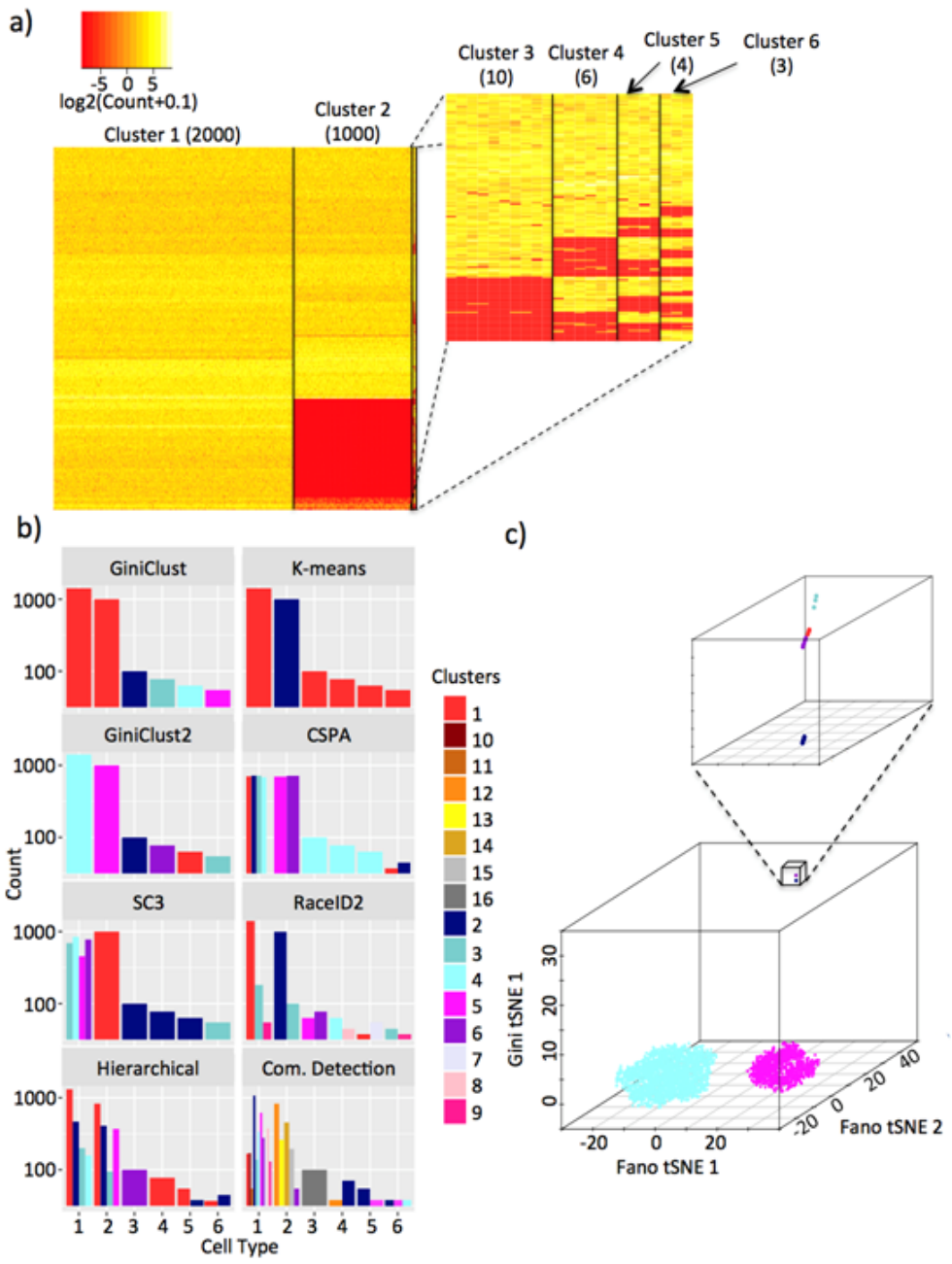


Figure 2.2. The application of GiniClust2 and comparable methods to simulated data. (a) A heatmap representation of the simulated data with 6 distinct clusters, showing the genes permuted to define each cluster. A zoomed-in view of the rare clusters is shown in the smaller heatmap. (b) A comparison between the true clusters (x-axis) and clustering results from GiniClust2 and comparable methods (y-axis). Each cluster is represented by a distinct color bar. Multiple bars are shown if a true cluster is split into multiple clusters by a clustering method. (c) A three-dimensional visualization of the GiniClust2 clustering results using a composite tSNE plot, combining two Fano-based tSNE dimensions and one Gini-based tSNE dimension. The inset shows a zoomed-in view of the corresponding region.

factor based feature space, such as hierarchical clustering and community-detection on a kNN graph, also results in the inability to resolve rare clusters (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). Furthermore, simply combining the Gini and Fano feature space fails to provide a more satisfactory solution (Supplemental Information, Supplemental Fig. S2.3). These analyses signify the importance of feature selection in a context-specific manner.

We next used the GiniClust2 weighted ensemble step to combine the results from GiniClust and Fano-factor-based k-means. Of note, all six cell clusters are perfectly recapitulated by GiniClust2 (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1), suggesting that GiniClust2 is indeed effective for detecting both common and rare cell clusters. To aid visualization, we created a composite tSNE plot, projecting the cells into a three-dimensional space based on a combination of a two-dimensional Fano-based tSNE map and a one-dimensional Gini-based tSNE map (Fig. 2.2c). A three-dimensional space is required because, although the Fano-based dimensions are able to clearly separate the two common clusters, the rare clusters are overlapping and cannot be fully discerned. The third (Gini) dimension results in complete separation of the rare clusters. Unlike a traditional tSNE plot, this composite view does not correspond to a single projection of a high-dimensional dataset into a three-dimensional space but integrates two orthogonal views obtained from different high-dimensional features. Although the distance does not have a simple interpretation, it provides a convenient way to visualize data from complementary views.

Since the number of common clusters is unknown in advance, we also tested the robustness of GiniClust2 with respect to other choices of k. We found that setting k=3 provides the same final clustering, while further increase results in poorer performance by splitting of the larger clusters (Supplemental Fig. S2.2b). By default, the value of k was chosen using the gap statistic, which coincided with the number of common clusters (k=2) [74]. However, this metric may not be optimal in various cases when the underlying distribution is more complex [75], therefore additional exploration is often needed to select the optimal value for k. Since the clustering outcome is sensitive to the choice of k (Supplemental Information), we recommend using the gap statistic as a starting point for choosing k, and then evaluating this choice of k by checking the resulting clusters for adequate separation in the Fano-factor-based tSNE plot and expression of distinct and biologically relevant genes.

For comparison, we evaluated the performance of two unweighted ensemble clustering methods. First, we used the cluster-based similarity partitioning algorithm (CSPA) [76] to combine the GiniClust and Fano-factor-based k-means (k=2) clustering results. The consensus clustering splits the common clusters into six subgroups, whereas cells in the four rare clusters are assigned to one of two clusters shared with the largest common cell group (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). Without guidance, the consensus clustering treats all clustering results equally and attempts to resolve any inconsistency via suboptimal compromise. The second method we considered, known as SC3 [64], is specifically designed for single-cell analysis. This method performs an unweighted

ensemble of k-means clusterings for various parameter choices without specifically targeting rare cell detection. Regardless of the specific parameter choices, k-means cannot resolve the rarest clusters, and the final ensemble clustering splits the largest group into three and differentiates only one of the four rare clusters (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). These analyses suggest that our cluster-aware, weighted ensemble approach is important for optimally combining the strengths of different methods.

We also compared the performance of GiniClust2 with other rare cell type detection methods. In particular, we compared with RaceID2 [70], which is an improved version of RaceID [69] developed by the same group. For fair comparison, we considered k=2, the exact number of common cell clusters, and k=12, the parameter value recommended by authors Grün et al. as determined by a within-cluster dispersion saturation metric [70]. In both cases, RaceID2 over-estimated the number of clusters, and split both common and rare cell clusters into smaller subclusters (Fig. 2.2b, Supplemental Information, Supplemental Fig. S2.1). This tendency of over-clustering is consistent with our previous observations [71].

2.3.3 Robust identification of rare cell types over a wide range of proportions

In order to evaluate the performance of GiniClust2 on analyzing real scRNA-seq datasets, we focused on one of the largest public scRNA-seq datasets generated by 10X Genomics [77]. The dataset consists of transcriptomic profiles of about 68,000 peripheral blood mononuclear cells (PBMCs) [77], which were classified into

11 subpopulations based on transcriptomic similarity with purified cell-types (Fig. 2.3a). It was noted that the transcriptomic profiles of several subpopulations are nearly indistinguishable [77].

To reduce the effects of stochastic variation and technical artifacts, we started by considering only a subset of cell types whose transcriptomic profiles are distinct from one another. In particular, we focused on three large subpopulations: CD56+ natural killer (NK) cells, CD14+ Monocytes, and CD19+ B cells. To ensure our analysis is not affected by within-cell-type heterogeneity, additional known gene markers were used to further remove heterogeneity within each subpopulation (see Materials and Methods for cell type definition details). In the end, three populations were selected, corresponding to NK, macrophage, and B cells, respectively (Fig. 2.3a). To systematically compare the ability of different methods in detecting both common and rare cell types, we created a total of 140 random subsamples that mix different cell types at various proportions (Supplemental Table S2.1), with the rare cell type (macrophage) proportions ranging from 0.2% to 11.6% (see Materials and Methods for details).

We applied GiniClust2 and comparable methods to the down-sampled datasets generated above. Each method was evaluated based on its ability to detect each cell type using three Matthews correlation coefficients (MCC) [78] (Fig. 2.3b). The MCC is a metric that quantifies the overall agreement between two binary classifications, taking into account both true and false positives and negatives. The

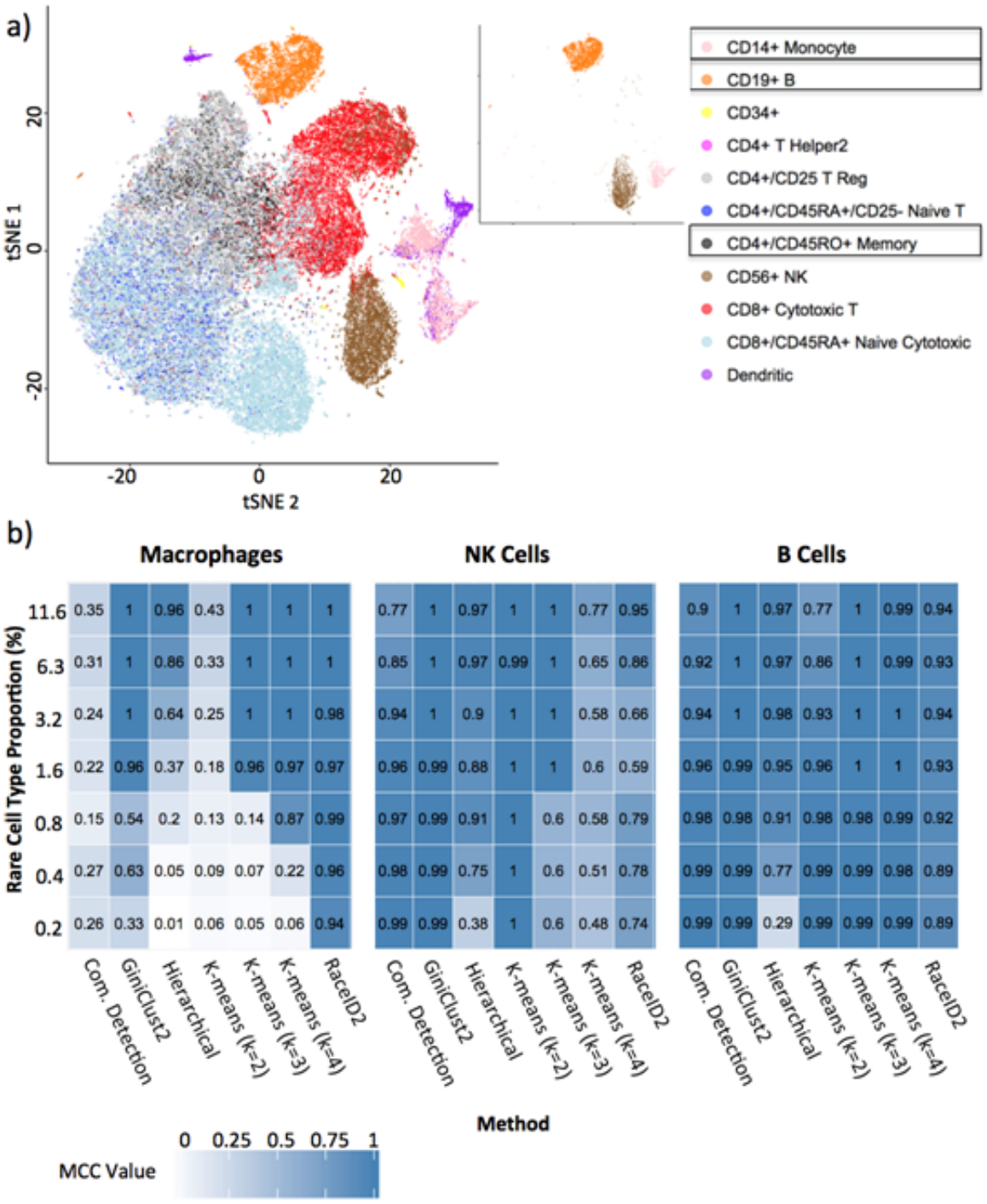


Figure 2.3. Analysis of the 68k PBMC dataset [77]. (a) A visualization of reference labels for the full data set (left), along with the 3 cell subtypes selected for detailed analysis (right). (b) Comparison of the performance of different clustering methods, quantified by a Matthews correlation coefficient (MCC) [78] for each of the three cell subtypes.

MCC value ranges from -1 to 1, where 1 means a perfect agreement between a clustering and the reference, 0 means the clustering is as good as a random guess, and -1 means a total disagreement between a clustering and the reference. In addition, we also evaluated the performance of each method using several additional metrics (Supplemental Information). While each metric typically generates a different value, the relative performance across different clustering methods is highly conserved (Supplemental Fig. S2.4).

RaceID2 is the best method for detecting the rare macrophage cell type at a frequency of 1.6% or lower, and GiniClust2 is the next best method. As expected, the performance of GiniClust degrades as the "rare" cell type becomes more abundant, whereas Fano-factor-based k-means becomes more powerful in such cases. Combining these two methods enables GiniClust2 to perform among the top over a wide range of rare-cell proportions. The remaining methods cannot detect rare cell clusters well. For the common groups, Fano-factor-based k-means tends to perform better, but only if the parameter is chosen correctly. For example, Fano-factor-based k-means with k=4 systematically splits the largest NK cell group and leads to a relatively low MCC value. Other clustering methods that use Fano-factor-based feature selection (such as hierarchical clustering and community detection) also adequately pick up common clusters. This strong performance is preserved by the GiniClust2 method. In comparison, RaceID2 does not perform as well here, since some of the cells in the common groups are falsely identified as rare cells. Taken together, these comparative results suggest that GiniClust2 reaches a good balance

for detecting both common and rare clusters. The same conclusion can be arrived using alternative evaluation metrics (Supplemental Fig. S2.4).

2.3.4 Detection of rare cell types in differentiating mouse embryonic stem cells

To test if GiniClust2 is useful for detecting previously unknown, biologically relevant cell types, we analyzed a published dataset associated with leukemia inhibitory factor (LIF) withdrawal induced mouse embryonic stem cell (mESC) differentiation [79]. Previously, we applied GiniClust to analyze a subset containing undifferentiated mESCs, and identified a rare group of Zscan4-enriched cells [71]. As expected, these rare cells were rediscovered using GiniClust2.

In this study, we focused on the cells assayed on Day 4 post LIF withdrawal, and tested if GiniClust2 might uncover greater cellular heterogeneity than previously recognized. GiniClust2 identified two rare clusters consisting of 5 and 4 cells respectively, corresponding to 1.80% and 1.44% of the entire cell population. The first group contains 25 differentially expressed genes when compared to the rest of the cell population (MAST likelihood ratio test p-value<1e-5, fold change>2), including known primitive endoderm (PrEn) markers such as *Col4a1*, *Col4a2*, *Lama1*, *Lama2*, and *Ctsl*. These genes are also associated with high Gini index values. Overall there is a highly significant overlap between differentially expressed and high Gini genes (Fisher exact test p-value<1e-18). The second group contains 10 differentially expressed genes (MAST likelihood ratio test p-value<1e-5, fold

change>2), including maternally imprinted genes *Rhox6*, *Rhox9*, and *Sct*, all of which are also high Gini genes. Once again there is a significant overlap between differentially expressed and high Gini genes (Fisher exact test p-value<1e-12). Although these clusters were detected in the original publication [79], this was achieved based on *a priori* knowledge of relevant markers. Here, the strength of GiniClust2 is that it can identify these clusters without previous knowledge.

In addition, GiniClust2 identified 2 common clusters. The first group specifically expresses a number of genes related to cell growth and embryonic development, including *Pim2*, *Tdgf1*, and *Tcf15* (MAST likelihood ratio test p-value<1e-5, fold change>2), indicating it corresponds to undifferentiated stem cells. The second group is strongly associated with a number of genes related to epiblast cells, including *Krt8*, *Krt18*, *S100a6*, *Tagln*, *Actg1*, *Anxa2*, and *Flnc* (MAST likelihood ratio test p-value<1e-5, fold change>2), suggesting this group corresponds to an epiblast-like state. Of note, 114 of the 128 genes (Fisher exact test p-value<1e-88) specifically expressed in this group were selected as high Fano-factor genes, confirming the utility of Fano factor in detecting common cell-types. Both populations were discovered in the original publication [79]. The dissimilarity between these cell types is evident in the heatmap (Fig. 2.4a) and composite tSNE plot (Supplemental Fig. S2.5).

For comparison, we applied RaceID2 to analyze the same dataset. Unlike GiniClust2, RaceID2 broke each cluster into multiple subclusters, and failed to identify the rare cell clusters (Fig. 2.4b). With k=2, RaceID2 found a total of 11 clusters. Clusters 1, 2, 4, and 9 display an epiblast-like signature, clusters 7 and 10

overexpress genes relating to maternal imprinting, and clusters 8 and 11 correspond to PrEn cells. From these results it appears that RaceID2 has difficulty in differentiating rare, biologically meaningful cell types from outliers.

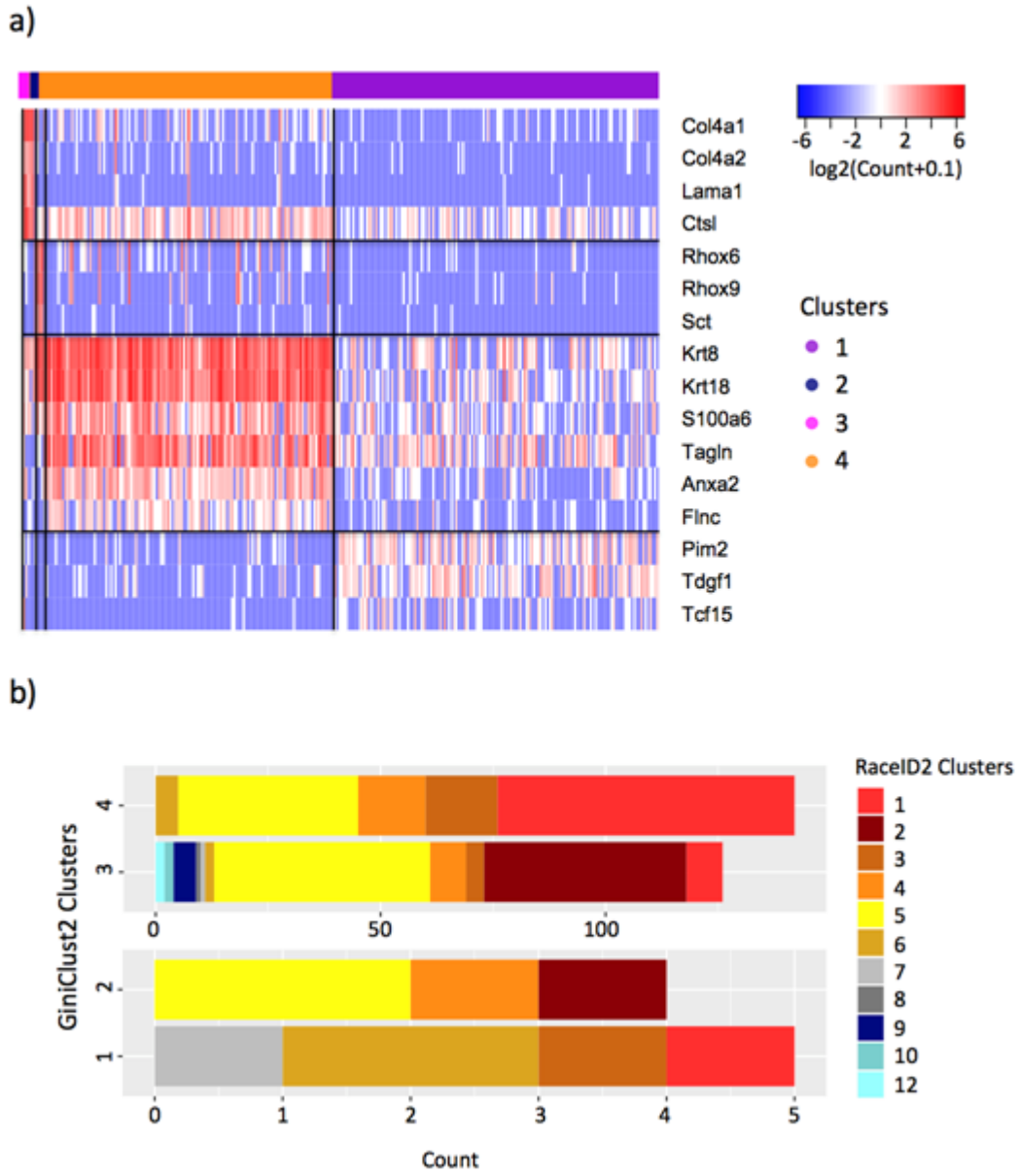


Figure 2.4. Analysis of the inDrop dataset for day 4 post-LIF mESC differentiation [19]. (a) A heatmap of top differentially expressed genes for each GiniClust2 cluster. The colorbar above the heatmap indicates the cluster assignments. (b) A comparison of GiniClust2 and RaceID2 clustering results, for common (above) and rare (below) cell types. The same color-coding scheme is used in all panels.

2.3.5 Scalability to large data sets

With the rapid development of single-cell technologies, it has become feasible to profile thousands or even millions of transcriptomes at single-cell resolution. Thus, it is desirable to develop scalable computational methods for single-cell data analysis. As a benchmark, we applied GiniClust2 to analyze the entire 68k PBMC data set [77] described above to uncover hidden cell types. The complete analysis took 2.3 hours on one core of a 2 GHz Intel Xeon CPU and utilized 237 GB of memory (not optimized for speed or memory usage). For comparison, RaceID2 analysis could not be completed for this large dataset. One possible explanation is this method may be limited to handling data sets with less than 65,536 data points due to an intrinsic vector size restriction in R. Our implementation of GiniClust2 circumvents this restriction by splitting up larger vectors into several smaller ones, with no changes to the functionality of the code. In principle, the same strategy can be implemented in RaceID2 to overcome this limitation. Comparisons of computational run-times between RaceID2 and GiniClust2 on smaller data sets show that the runtime of GiniClust2 scales better with the number of cells in the data set (Supplemental Information, Supplemental Fig. S2.6). For example, for a data set of 80 cells GiniClust2 and RaceID2 take the same amount of time, whereas for the simulated data set of 3023 cells GiniClust2 takes just under 10 minutes while RaceID2 takes 1 hour and 13 mins. Despite the advantages of GiniClust2, it should be noted that GiniClust2 still requires a

considerable amount of memory to run on very large data sets, presenting a limitation to the application of this method to even larger data.

Our analysis identified 9 common clusters and two rare clusters (Fig. 2.5a). In general, the results of GiniClust2 and Fano-factor-based k-means are similar; both agree well with the reference cell types (Fig. 2.5b). To quantify this agreement, we use normalized mutual information (NMI), which is an entropy-based method normalized by cluster size that can be applied to multi-class classification problems [80]. A value of 1 indicates perfect agreement, whereas a value of 0 means the performance is as good as random guess. Here, values are 0.540 for GiniClust2 and 0.553 for Fano-factor-based k-means. Most of the discrepancy between the clustering results and reference labels are associated with T-cell subtypes. As noted by the original authors [77], these subtypes are difficult to separate because they share similar gene expression patterns and biological functions. The common clusters detected by GiniClust2 and Fano-factor-based k-means express marker genes known to be specific to the cell types represented in the reference [81] (Fig. 2.5c).

With respect to rare cell types, our first group contains a homogeneous and visually distinct subset of 171 of 262 total CD34+ cells (Cluster 2, Fig. 2.5a). This cluster was partially detectable using Fano-factor-based k-means, although it was partially mixed with major clusters. The second rare cell cluster is previously unrecognized (Cluster 3, Fig. 2.5a). This cluster contains 118 cells (0.17%) within a large set of 5433 immune cells with similar gene expression patterns. Among these 118 cells, 101 cells are classified as monocytes, whereas 16 are classified as

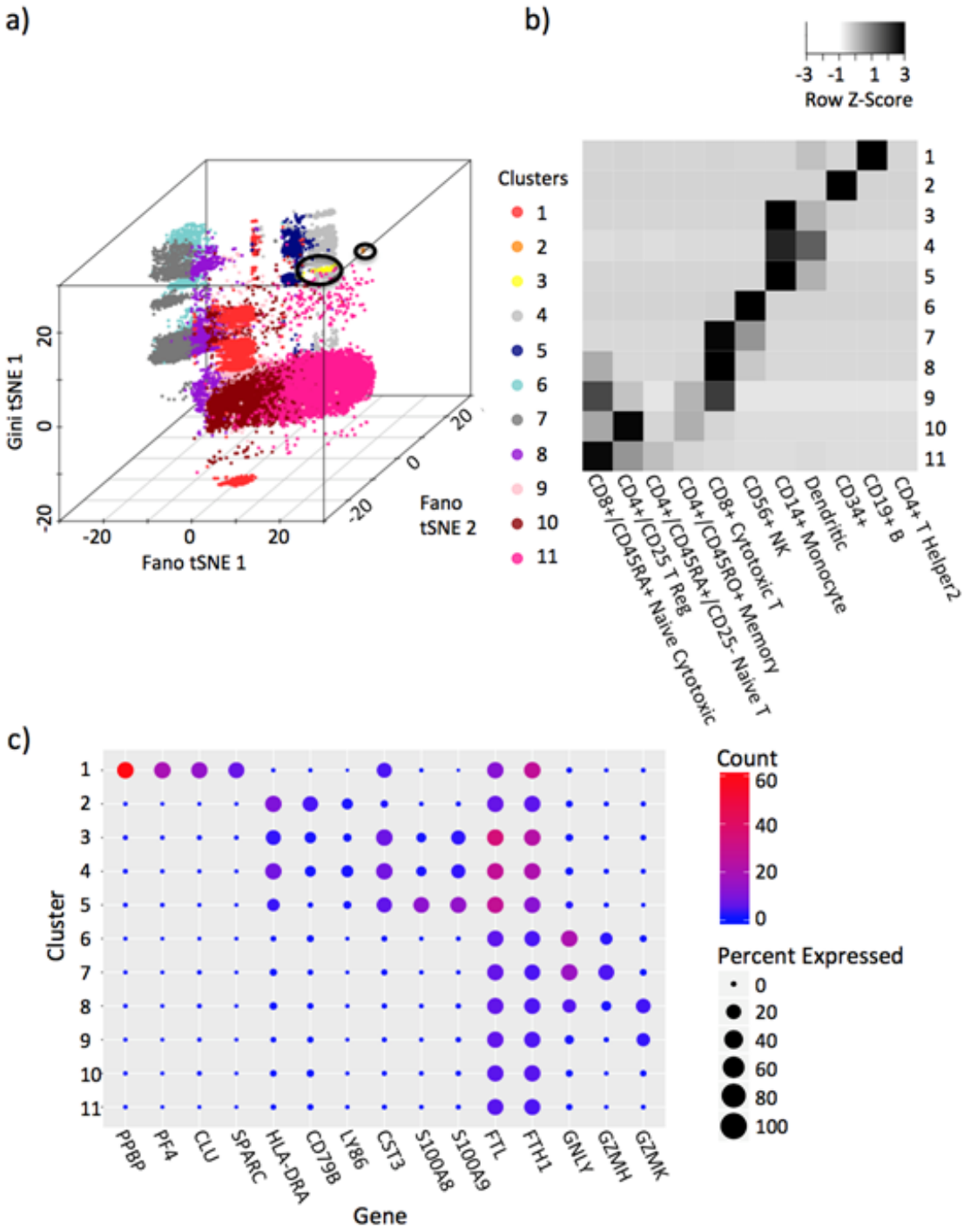


Figure 2.5. Results from the full 68k PBMC data analysis. (a) A composite tSNE plot of the GiniClust2 results; rare cell types are circled. (b) A confusion map showing similarities between GiniClust2 clusters and reference labels. Values represent the proportion of cells per reference label that are in each cluster. (c) A bubble plot showing expression of cluster-specific genes. Size represents the percentage of cells within each cluster with non-zero expression of each gene, while color represents the average normalized UMI counts for each cluster and gene.

dendritic cells, and 1 is classified as a CD34+ cell. Differential expression analysis (MAST likelihood ratio test p-value<1e-5, fold change>2) identified 187 genes that are specifically expressed in this cell cluster, including a number of genes associated with tolerogenic properties, such as *Ftl*, *Fth1*, and *Cst3* [82], suggesting these cells may be associated with elevated immune response and metabolism. Additional validation would be necessary to determine whether this cluster is functionally distinct. Taken together, these results strongly indicate the utility of GiniClust2 in analyzing large single-cell datasets.

2.4 Discussion and Conclusions

According to the "no free lunch" theorems [83], an algorithm that performs well on a certain class of optimization problems is typically associated with degraded performance for other problems. Therefore it is expected that clustering algorithms optimized for detecting common cell clusters are unable to detect rare cell clusters, and vice versa. While ensemble clustering is a promising strategy to combine the strengths of multiple methods [76, 64, 65], our analysis shows that the traditional, unweighted approach does not perform well.

To optimally combine the strengths of different clustering methods, we have developed GiniClust2, which is a cluster-aware, weighted ensemble clustering method. GiniClust2 effectively combines the strengths of Gini-index- and Fano-factor-based clustering methods for detecting rare and common cell clusters, respectively, by assigning higher weights to the more reliable clusters for each

method. By analyzing a number of simulated and real scRNA-seq datasets, we find that GiniClust2 consistently performs better than other methods in maintaining the overall balance of detecting both rare and common cell types. This weighted approach is generally applicable to a wide range of problems.

GiniClust2 is currently the only rare-cell-specific detection method equipped to handle such large data sets, as demonstrated by our analysis of the 68k PBMC dataset from 10X Genomics. This property is important for detecting hidden cell types in large datasets, and may be particularly useful for annotating the Human Cell Atlas [84].

2.5 Materials and Methods

2.5.1 Data preprocessing

The processed mouse ESC scRNA-seq data is represented as UMI filtered-mapped counts. Removing genes expressed in fewer than 3 cells, and cells expressing fewer than 2000 genes, we were left with a total of 8055 genes and 278 cells.

The processed 68k PBMC dataset, represented as UMI counts, was filtered and normalized using the code provided by 10X Genomics (https://github.com/10XGenomics/single-cell-3prime-paper). The resulting data consists of a total of 20387 genes and 68579 cells. Cell-type labels were assigned

based on the maximum correlation between the gene expression profile of each single cell to 11 purified cell populations, using the code provided by 10X Genomics.

2.5.2 GiniClust2 method details

The GiniClust2 pipeline contains the following steps.

Step 1: Clustering cells using Gini-index based features

The Gini index for each gene is calculated and normalized as described before [71]. Briefly, the raw Gini index is calculated as twice the area between the diagonal and the Lorenz curve, taking a range of values between 0 and 1. Raw Gini index values are normalized by removing the trend with maximum expression levels using a two-step LOESS regression procedure as described in [71]. Genes whose normalized Gini index is significantly above zero (p-value < 0.0001 under the normal distribution assumption) are labeled high Gini genes and selected for further analysis.

A high-Gini-gene-based distance is calculated between each pair of cells using the Jaccard distance metric. This is used as input into DBSCAN [85], which is implemented using the dbscan function in the fpc R package, with method= "dist". Parameter choices for eps and MinPts are discussed in the Supplemental Information.

Step 2: Clustering cells using Fano-factor based features

The Fano factor is defined as the variance over mean expression value for each gene. The top 1000 genes are chosen for further analysis. Principal component analysis (PCA) is applied to the gene expression matrix for dimensionality reduction, using the svd function in R. The first 50 principal components are reserved for clustering analysis. Cell clusters are identified by k-means clustering, using the kmeans function in R with default parameters. Optimal choice of k is discussed in the Supplemental Information. To improve robustness, 20 independent runs of k-means clustering with different random initializations are applied to each dataset, and the optimal clustering result is selected.

Step 3. Combining the results from Steps 1 and 2 via a cluster-aware, weighted ensemble approach.

We adapted the weighted consensus clustering algorithm developed by Li and Ding [73] by further considering cluster-specific weighting. For GiniClust, higher weights are assigned to the rare cell clusters and lower weights to common clusters, whereas the opposite scheme is used to weight the outcome from Fano-factor-based k-means clustering. This allows us to combine the strengths of each clustering method. The mathematical details are described as follows, and visualized in Fig. 2.1b.

Let P^G be the partitioning provided by GiniClust, and P^F the partitioning provided by Fano-factor-based clustering. Each partition consists of a set of clusters: $C^G = C_1^G, C_2^G, \dots, C_{k_G}^G$, and $C^F = C_1^F, C_2^F, \dots, C_{k_F}^F$. Define the connectivity matrices as:

$$M_{ij}(P^G) = \begin{cases} 1, (i,j) \in C_k(P^G) \\ 0, otherwise \end{cases}, \text{ and } M_{ij}(P^F) = \begin{cases} 1, (i,j) \in C_k(P^F) \\ 0, otherwise \end{cases}$$

If two cells are clustered together in the same group, their connectivity is 1, while if they are clustered separately, their connectivity is 0. Define the weighted consensus association as:

$$\overline{M_{ij}} = w_{ij}^G M_{ij}(P^G) + w_{ij}^F M_{ij}(P^F)$$

where $w_{ij}^G + w_{ij}^F = 1$, $w_{ij}^G, w_{ij}^F \geq 0 \; \forall \; i, j \in [1, n]$, n represents the number of cells. Weights w_{ij}^G and w_{ij}^F are specific to each pair of cells, and are determined based on $\widetilde{w}_i^G$ and $\widetilde{w}_i^F$, weights that are specific to each cell.

For simplicity, we set the cell-specific weights for the Fano-factor-based clusters as a constant: $\widetilde{w}_i^F = f$ '. The cell-specific GiniClust ($\widetilde{w}_i^G$) weights are determined as a function of the size of the cluster containing the particular cell. Our choices for these weights derive from the observation that as the proportion of the rare cell type increases, the utility of GiniClust begins to decline. For simplicity, we

model the cell-specific GiniClust weights using a logistic curve, specified by the following function:

$$\widetilde{w}_i^G(x_i) = 1 - \frac{1}{1 + e^{-(x_i - \mu\prime)/s\prime}}$$

where x_i is the proportion of the GiniClust cluster to which cell i belongs, μ' is the rare cell type proportion at which GiniClust and Fano-factor-based clustering methods have approximately the same ability to detect rare cell types, and s' represents how quickly GiniClust loses its ability to detect rare cell types above μ'. The parameters s', μ', and f' can be viewed as intermediate variables that are closely associated with the parameters s, μ, and f, schematically shown in Fig. 2.1c. Specifically, $f = \frac{f\prime}{1+f\prime}$, $s = s'$, and μ is obtained relative to the other parameters through the following relationship: $f' = 1 - \frac{1}{1+e^{(\mu-\mu\prime)/s\prime}}$. The selection of the parameter values for s', μ', and f', as well as a sensitivity analysis, are described in the Supplemental Information.

To set the cell-pair-specific weights, we first define

$$\widetilde{w}_{ij}^G = \max(\widetilde{w}_i^G, \widetilde{w}_j^G) \text{ and } \widetilde{w}_{ij}^F = \widetilde{w}_i^F$$

Then, weights are normalized to 1:

$$w_{ij}^{G} = \frac{\widetilde{w}_{ij}^{G}}{\widetilde{w}_{ij}^{G} + \widetilde{w}_{ij}^{F}} \text{ and } w_{ij}^{F} = \frac{\widetilde{w}_{ij}^{F}}{\widetilde{w}_{ij}^{G} + \widetilde{w}_{ij}^{F}}$$

Each cell-cell pair will thus be assigned a weighted consensus association between 0 and 1, which is a weighted average of both GiniClust and Fano-factor-based clustering associations, where the weights are functions of the size of the cell clusters.

At this point, the weighted consensus association matrix provides a probabilistic clustering for each cell, where each entry represents the probability that cell i and cell j reside in the same cluster. To transform this into a final deterministic clustering assignment, we optimize the following:

$$min_{U} \left|\left|\overline{M} - U\right|\right|^{2},$$

where U is any possible connectivity matrix. In Li and Ding [73], this optimization problem is solved via symmetric non-negative matrix factorization (NMF) to yield a soft clustering. To obtain a hard clustering we add an orthogonality constraint, leading to k-means clustering [86], implemented once again using the kmeans R function.

2.5.3 tSNE visualization

Dimension reduction by tSNE [87] is performed using the Rtsne R package. The tSNE algorithm is first run using the Gini-based distance to obtain a one-dimensional projection of each cell. For large data sets, tSNE is run on the first 50 principal components of the Gini-based distance to prevent tSNE from becoming prohibitively slow. Then, the tSNE algorithm is run using the first 50 principal components of our Fano-based Euclidean distance to obtain a separate two-dimensional projection. The three resulting dimensions (one for Gini-based distance and two for Fano-based distance) are plotted to visualize cluster separation.

2.5.4 Differential expression analysis on resulting clusters

Differentially expressed genes for each cluster are determined by comparing their gene expression levels to all other clusters. This is performed using the zlm.SingleCellAssay function in the R MAST package [88], with method= "glm". P-values for differentially expressed genes are calculated using the lrTest function, with a hurdle model.

2.5.5 SC3 analysis

SC3 [64] was accessed through the SC3 Bioconductor R package. SC3 was applied to the simulated data set post-filtering using default parameters, with k=6 to

match the true number of clusters. The author-recommended choice of k using the Tracy-Widom test yielded a k of 55, and was deemed inappropriate for this analysis.

2.5.6 CSPA analysis

Matlab code for the CSPA [76] was accessed through http://strehl.com/soft.html, under "ClusterPack_V2.0." CSPA was applied to the Gini and Fano-based clustering results for the simulated data set, using the clusterensemble function, specifying the CSPA option. Results are shown for k=5, the default parameter, and k=6, the true number of clusters.

2.5.7 RaceID2 analysis

RaceID2 [70] R scripts were accessed through https://github.com/dgrun/StemID. RaceID2 was applied to already-filtered data sets as above to make results directly comparable to GiniClust2, with default parameters. Results are shown for k set to the default parameter as determined by a within-cluster dispersion saturation metric [70], and k to match the corresponding GiniClust2 k parameter specification.

2.5.8 Hierarchical clustering analysis

Hierarchical clustering was performed on a Fano-based Euclidean distance using the hclust function in R. For the simulated data analysis, results are shown for choices k=6, to match the true number of clusters, and k=2, the parameter value as determined by the gap statistic through the clusGap function in R. For the subsampled PBMC analysis, results are shown for k=3, to match the true number of clusters.

2.5.9 Community detection analysis

Community detection was performed on a k-nearest neighbor (kNN) graph, using a high Fano feature space, for simulated and subsampled data sets. Function nn2 in the RANN R package was used to compute a kNN distance with default parameters. The igraph R package was used to perform community detection, using the cluster_edge_betweenness function with default parameters.

2.5.10 Simulation details

We created synthetic data following the same approach as Jiang et al. [71], specifying one large 2000 cell cluster, one large 1000 cell cluster, and four rare clusters of 10, 6, 4 and 3 cells, respectively. Gene expression levels are modeled using a negative binomial distribution, and distribution parameters are estimated using an intestinal scRNA-seq data set using a background noise model as in Grün et al. [69]. To create clusters with distinct gene expression patterns, we permute 100

lowly (mean<10 counts) and 100 highly (mean>10 counts) expressed gene labels for each cluster (see Jiang et al. [71] for more details). This results in a 23,538 gene by 3023 cell data set. After filtering (as above) we are left with 3708 genes and 3023 cells.

2.5.11 10X Genomics data subsampling

The full 68k 10X Genomics PBMC dataset is down-sampled for model evaluation. We consider only 3 cell types here. CD19+ B cells are defined by their correlation to reference transcriptomes as in Zheng et al. [77]. CD14+ monocytes and CD56+ NK cells are defined in the same way, but here we recognize that these broadly defined cell types actually consist of two subtypes each. We therefore use additional known markers to refine each cell type definition. With regard to CD14+ monocytes, we use macrophage markers *Cd68* and *Cd37* [81] to separate macrophages and monocytes, and we define macrophage cells as those with positive expression of both markers. These cells are selected for subsampling. The CD56+ NK cells are composed of NK and NKT cells, so we use T-cell markers *Cd3d*, *Cd3e*, and *Cd3g* [81] to separate the groups, and define the NK cells as those with zero expression of these three markers. There is some additional heterogeneity in this NK group, so we choose to include only those NK cells that were most highly correlated (top 50%) to the reference transcriptomes. Given these cell type definitions, we created 7 sets of 20 subsampled data sets each for a total of 140 data sets in the following manner: five cells were randomly sampled from the

macrophage cell population to form a "rare" cell group for all 120 datasets. Then, for each set of 20 data sets, cells were randomly sampled from the NK and B cells in specified numbers to form "common" cell clusters, the details of which are listed in Supplemental Table S1.

2.5.12 Availability of data and materials

GiniClust2 is implemented in R and the source code has been deposited at https://github.com/dtsoucas/GiniClust2. This open-source software is released under the MIT license, and accessible under the DOI: https://doi.org/10.5281/zenodo.1211359 [89].

The intestinal scRNA-seq data used in the creation of the simulated data set is available through the Gene Expression Omnibus (GEO) under the accession number GSE62270 [90]. The mouse ESC scRNA-seq data is available through GEO under the accession number GSE65525 [91]. The 10X PBMC data is available through NCBI Sequence Read Archive (SRA) under the accession number SRP073767 [92].

Accurate Estimation of Cell-Type Composition from Gene Expression Data

Daphne Tsoucas, Hyde Chen, Guoji Guo and Guo-Cheng Yuan

3.1 Abstract

The rapid development of single-cell transcriptomic technologies has helped uncover the cellular heterogeneity within cell populations and tissue samples. However, bulk RNA-seq continues to be the main workhorse for quantifying gene expression levels due to technical simplicity and low cost. In order to most effectively extract information from bulk data given the new knowledge gained from single-cell methods, we have developed a novel algorithm to estimate the cell-type composition of bulk data from a single-cell RNA-seq-derived cell-type signature. By thorough comparison with a number of existing methods using various real RNA-seq datasets, we find that our new approach is more accurate and comprehensive than previous methods, especially for the estimation of rare cell types. More importantly, our method is able to detect cell-type composition changes in response to external perturbations, thereby providing a valuable, cost-effective method for dissecting the cell-type-specific effects of drug treatments or condition changes. As such, our method is applicable to a wide range of biological and clinical investigations.

3.2 Introduction

Gene expression profiling is widely used in biology and medicine for the systematic characterization of cellular or disease states. Identifying gene expression changes across conditions can help generate hypotheses as to underlying biological

mechanisms. However, one common problem is that each sample has considerable cellular heterogeneity that bulk RNA-seq methods are not able to capture. As the overall signature generated from these methods only measures the average behavior, it is often the case that changes in gene expression only reflect changes in cell-type composition, rather than fundamental changes in cell states [93]. To alleviate such problems, a series of computational methods have been developed with the common goal of estimating the cell-type composition within a tissue sample from bulk RNA-seq data [94,95]. These methods, often referred to as deconvolution methods, provide an important means to distinguishing between changes in cell-type composition and changes in cell-state. Various estimation approaches have been used, including least squares regression [96], constrained least squares regression [97], quadratic programming [98-100], and ν-support vector regression [101].

However, existing methods have a number of important limitations. Most importantly, the underlying cell-type signatures must be known in advance. Most studies assume that such signatures can be identified from the bulk transcriptomic profiling of purified cell types. The success of cell-type purification relies heavily on the knowledge of specific markers as well as the ability to isolate cells from surrounding tissues. Moreover, it is now known that even the 'purified' cells may still contain significant cellular heterogeneity [102].

Recent single-cell transcriptomic methods [103,104] have provided a very powerful approach to systematically characterizing cellular heterogeneity, thereby enabling the identification of new cell types/states and the reconstruction of

developmental trajectories. Applications of single-cell methods in medicine have led to novel insights into disease progression and drug response [105-107]. Single-cell data provides an alternative approach to deriving cell-type signatures. In fact, a few recent studies [108,109] have extended deconvolution methods by estimating cell-type signatures from single-cell data, where cell types are inferred by clustering. While these methods are useful, a number of significant challenges remain. In particular, their estimates tend to be biased against cells types that either: 1) make up a small proportion of the total bulk cell population, or 2) are characterized by lowly expressed genes. To remove these biases, we develop a cell-type-sensitive method for the estimation of the underlying cell fractions, using a novel weighted least squares approach.

3.3 Results

3.3.1 A weighted least squares approach to deconvolution

We aimed to build a method that can accurately and comprehensively estimate the relative abundance of both common and rare cell types within a bulk sample. Much like recent studies [108,109], we use single-cell RNA-seq data to extract cell-type-specific gene expression signatures. Simply, the cell types are identified by clustering analysis. For each cell type, marker genes are identified by differential expression analysis, after which gene expression levels for each of these

genes are averaged across all cells associated with the cell type. This results in a gene by cell type signature matrix, which is denoted by S (see Methods for details).

In order to accurately and comprehensively estimate the cell-type composition, we made a number of significant modifications to the standard ordinary least squares (OLS) approach, which underlies most existing methods [96-100]. In this approach, the deconvolution problem is represented as a system of linear equations: $Sx = t$, where S is an nxk gene signature matrix (n=number of genes, k=number of cell types), t is an nx1 vector representing the bulk RNA-seq data, and x is a kx1 vector containing the cell type numbers. Since typically n >> k, this is an over-determined equation with no exact solution. In the OLS approach, the solution x minimizes the total squared absolute error. This leads to two undesirable consequences. Firstly, the estimation error for rare cell types is typically large since such a term has little impact on the total estimation error. Secondly, not all informative genes are effectively taken into account. The contribution of a gene can be minimal if its mean expression level is low, even if it is highly differentially expressed between different cell types.

To illustrate these effects, we carried out a highly idealized simulation. We generated a single-cell data set consisting of three cell types, each characterized by two differentially expressed marker genes. A portion of the data was used to create the signature matrix, while a non-overlapping portion was used to create the bulk data by averaging gene expression values across the cells. First, to see how the OLS formulation affects rare cell type estimation, we varied the abundance of one cell type from 0.02% to 33.3% (see Methods for details). When the abundance is very

low, the relative percent error (RPE) of estimation, defined as $RPE = \frac{\left|\frac{x_l}{\sum_{j=1}^{k} x_j} - \frac{\hat{x}_l}{\sum_{j=1}^{k} \hat{x}_j}\right|}{\frac{x_l}{\sum_{j=1}^{k} x_j}} * 100$, is very high (Fig. 3.1a), supporting our intuition that the OLS framework is not appropriate for estimating the prevalence of rare cell types. In addition, we varied the mean gene expression level of the two highly differentially expressed genes (fold change = 10) pertaining to one cell type such that the ratio of mean expression level between genes in this cell type vs. the other two cell types ranges from 0.001 to 0.2. As expected, the deconvolution accuracy is significantly affected by the mean expression level of these genes (Fig. 3.1b).

To mitigate these issues, we designed a weighted least squares approach to properly adjust the contribution of each gene. Accordingly, the weighted error term becomes: $Err = \sum_{i=1}^{n} w_i (t_i - (Sx)_i)^2$. Our mathematical derivation indicates that setting $w_i = \frac{1}{(Sx)_i^2}$ optimally reduces the biases (see Methods for details). To test this idea empirically, we applied this weighted approach to analyze the aforementioned simulated data. It is clear that both biases are significantly reduced (Fig. 3.1).

When applying our weighted least squares method in all real applications, we make a few adjustments required to make the weighting formulation tractable in all situations. Given that the weights are a function of the solution, we use an iterative method in which weights are initialized according to the solution from the unweighted method, and then subsequently updated by the weighted least squares solution until convergence (see Methods for details). Next, given that cell type proportions must be non-negative, the weighted least squares solution is

constrained such that x >=0. Finally, a dampening constant is introduced to prevent infinite weights resulting from low cell type proportions and/or low marker gene expression, which will lead to unstable solutions driven by only one or a few genes (see Methods for details). Because of this last step, we subsequently refer to our method as Dampened Weighted Least Squares (DWLS).

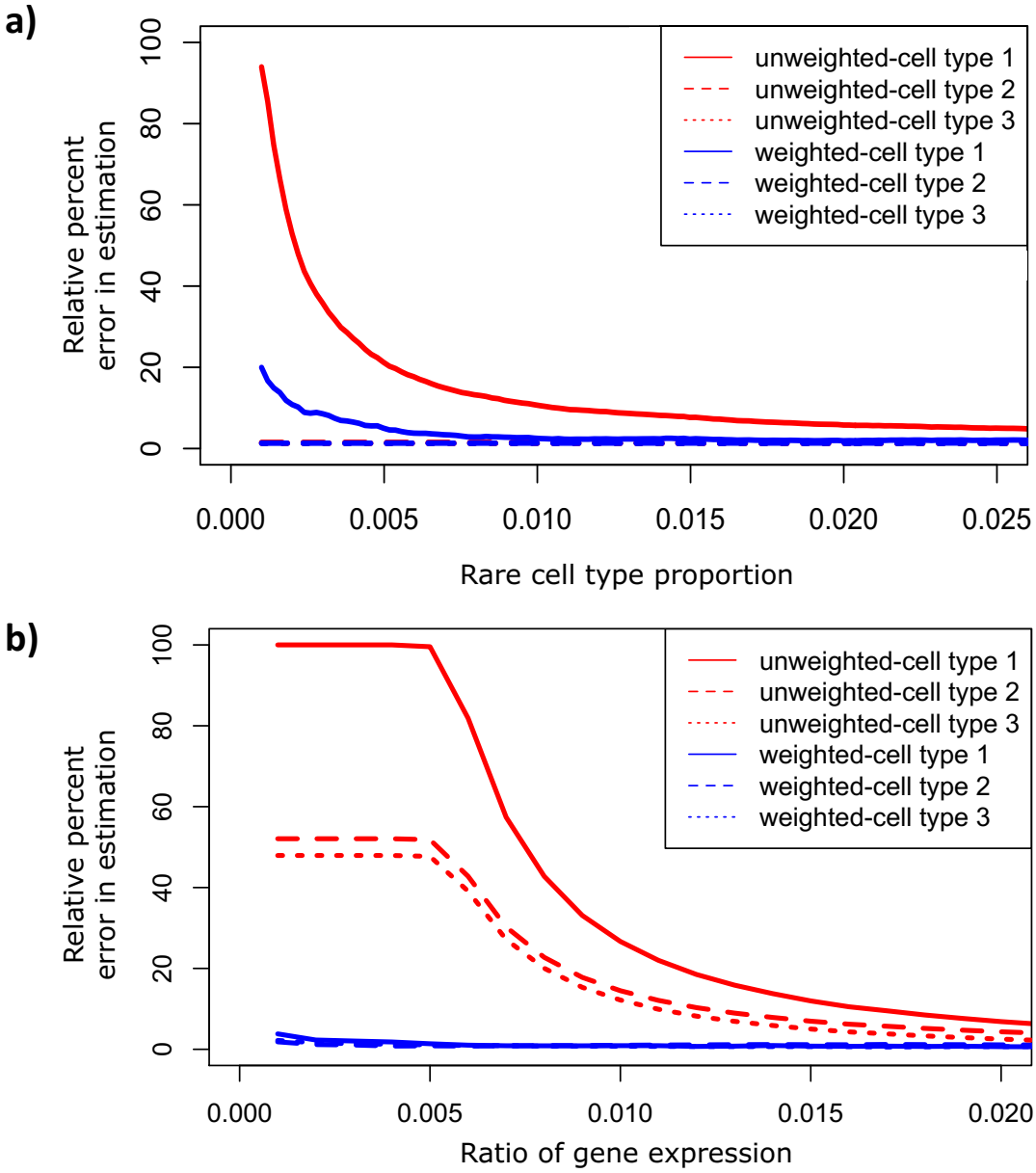


Figure 3.1. A simple simulation shows the advantages of a weighted least squares method. (a) A plot of relative percent error in estimation using both unweighted and weighted least squares approaches, for each of three cell types across various proportions of cell type 1, the rare cell type. Because of the increased influence of rare-cell-type-specific marker genes in the weighted sum of squares error, the weighted least squares method performs better in the estimation of rare cell types than the unweighted method. (b) A plot of relative percent error in estimation using both unweighted and weighted least squares approaches, for each of three cell types across various ratios of mean gene expression level between marker genes of cell type 1 and marker genes of cell types 2 and 3. Because of the increased influence of lowly expressed marker genes in the weighted sum of squares error, the weighted least squares method performs better in the estimation of all cell types than the unweighted method.

3.3.2 Benchmarking of weighted least squares on simulated PBMC data

To evaluate the performance of our DWLS method, we first consider a benchmark data set introduced by Schelker et al. [109], who were among the first to consider the application of a single-cell derived gene expression signature to the problem of deconvolution. This data set is a compilation of 27 single-cell data sets from immune and cancer cell populations, derived from human donor peripheral blood mononuclear cells (PBMCs), tumor-derived melanoma patient samples, and ovarian cancer ascites samples. Since no bulk data was provided, we created 27 simulated bulk data sets by averaging expression values for each gene across all cells obtained from each donor, assuming that the bulk data is equivalent to the pooled data from individual cells. A similar assumption was made previously [109]. In addition, the cell-type-specific gene expression matrix was estimated by clustering the combined 27 single-cell data sets. Marker genes were then chosen to match the genes used in the immune-cell-specific signature from CIBERSORT [101], and expression values for each marker gene were averaged within each cell type.

We applied ν-support vector regression (ν-SVR), quadratic programming (QP) and DWLS to the deconvolution of these 27 simulated bulk data sets. To quantify the overall performance of each method, we use two metrics. The first is a modified relative percent error metric, which quantifies the difference in true and estimated cell type proportions, normalized by the mean of true and estimated cell type proportions (see Methods for details). Averaged across all cell types, the modified relative percent error is lowest for DWLS, at 53.3%, second lowest for ν-

SVR, at 57.0%, and highest for QP, at 62.9%. The second is a more standard metric of absolute error between estimated and true cell type proportions, in which we can see that absolute errors across cell types are again on average lowest for DWLS (Table S3.1).

We further compared the accuracy of different methods on a per-cell-type basis (Fig. 3.2a). While ν-SVR performs well for the largest cell subpopulation, DWLS performs better over a wide range of cell types, especially the rarest cell groups. In particular, DWLS preserves a good balance between rare and common cell-type estimation. A similar trend can be seen from the standpoint of absolute error (Table S3.1).

We took a closer look at the two rarest cell types across the 27 samples: dendritic and endothelial cells. Dendritic cells contribute to a maximum of 4.89% of the total cells in any given sample, with an average 0.999% prevalence across samples. Endothelial cells contribute to a maximum of 6.99% of the total cells in any given sample, with an average 0.831% prevalence across samples. For both cell types, DWLS is able to maintain high estimation accuracy ($\rho_{dendritic,DWLS} = 0.93, \rho_{endothelial,DWLS} = 0.81$), outperforming ν-SVR $\left(\rho_{dendritic,SVR} = 0.91, \rho_{endothelial,SVR} = 0.54\right)$ and QP ($\rho_{dendritic,QP} = 0.66, \rho_{endothelial,QP} = 0.44$). Overall, these analyses indicate that DWLS exhibits greater accuracy in estimating rare cell types than existing methods.

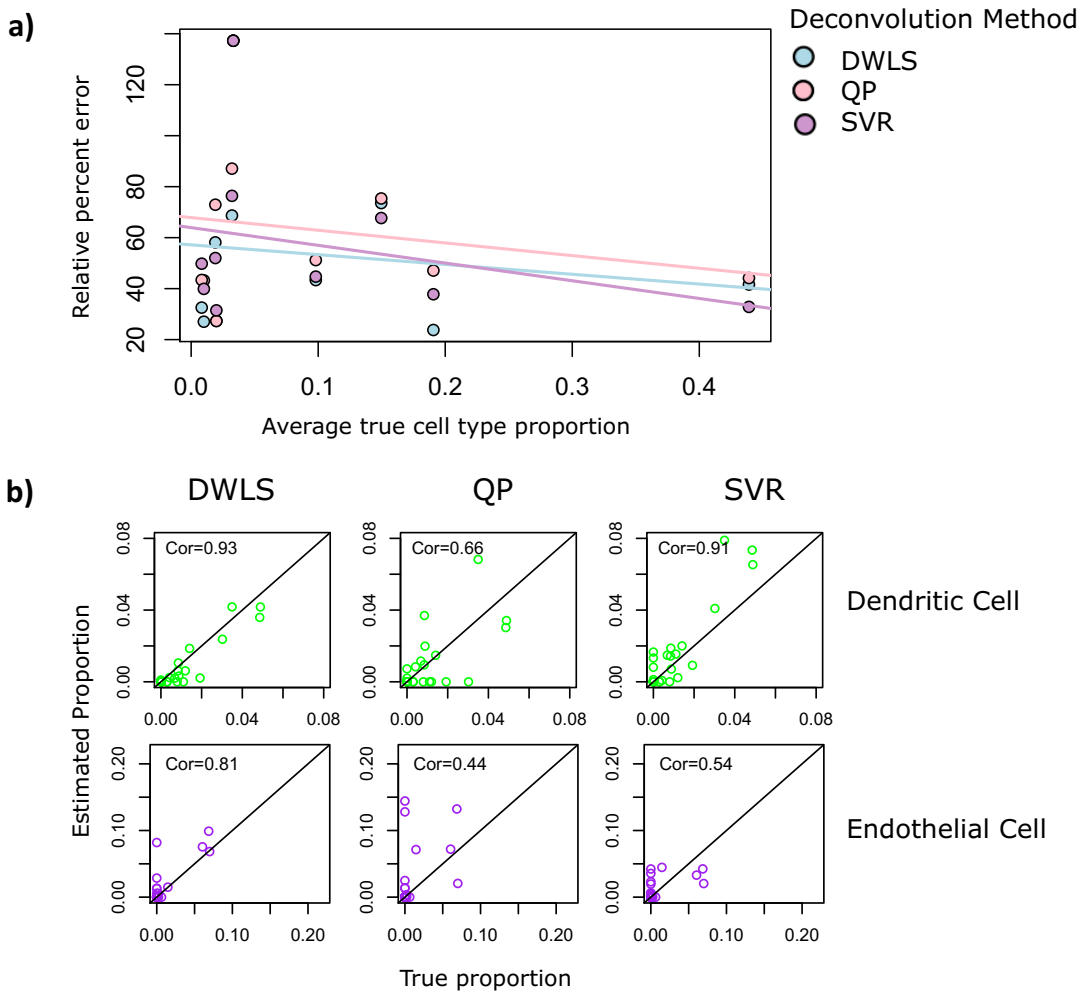


Figure 3.2. Results from the deconvolution of 27 simulated bulk data sets from donor, melanoma, and ovarian cancer patient immune and tumor cells, using dampened weighted least squares (DWLS), quadratic programming (QP), and ν-support vector regression (ν-SVR) estimation methods. (a) The mean relative percent error in estimation for each cell type across the 27 data sets, plotted against the average true proportion of the cell type, for each method. The fitted lines represent the trend in estimation accuracy as a function of cell type proportion. (b) A subset of the deconvolution cell type proportion estimates, plotted against the true cell type proportions. Here, only the rarest cell types, dendritic and endothelial cells, are shown. Correlation values between true and estimated proportions are used to quantify estimation accuracy. The 45-degree line in each plot represents the optimal estimate.

3.3.3 DWLS extends to real bulk data characterized by the Mouse Cell Atlas

Recently, Han et al. have characterized forty-three healthy mouse tissues at single-cell resolution to create the Mouse Cell Atlas [110]. Based on a combined single-cell data set of 61k cells, they have identified 52 distinct cell types spread across all tissues. Here we selected four represented tissues—kidney, lung, liver and small intestine—and generated two bulk RNA-seq data sets per tissue. Obtaining both bulk and single-cell data from the same tissue provides an opportunity to rigorously evaluate the accuracy of our deconvolution method, where we assume cell-type composition in bulk and single-cell data sets to be approximately equal. We use the entire single-cell data set to provide a comprehensive gene expression signature.

We calculate estimates using various deconvolution methods: DWLS, ν-SVR and QP. Overall, we find a high replicability of our results within each pair of tissues, each of which come from separate mice. DWLS estimates for each pair have correlations between 0.84 and 0.99, showing that cell type composition differences between mice are small.

Here, DWLS again performs favorably over other methods, which we demonstrate in two ways. We first look at a representative example, the deconvolution of bulk kidney data (Fig. 3.3a,b). We plot deconvolution estimates against the predicted true cell type composition, and find that DWLS estimates are most highly correlated to the predicted true proportion ($\rho_{kidney,DWLS} = 0.89$), with ν-SVR and QP performing less favorably ($\rho_{kidney,SVR} = 0.87, \rho_{kidney,QP} = 0.092$) (Fig.

3.3a). DWLS is the only method able to correctly predict the presence of all four kidney cell types. QP misses three out of these four groups entirely, while ν-SVR misses one (Fig. 3.3b). ν-SVR also significantly overestimates the presence of other rarer cell types (Fig. 3.3b), which should make up around 6% of the total kidney cell population, but are estimated by ν-SVR to make up 43% instead.

Second, we look more generally at the estimates of all eight tissue samples analyzed. DWLS remains the most accurate method, with an average correlation of 0.78 for DWLS, compared to average correlations of 0.21 and 0.59 for QP and ν-SVR, respectively (Fig. 3.3c). QP once again fails to detect biologically relevant cell types across the eight bulk samples. This can be quantified by a sensitivity metric, defined as the fraction of all true cell types that are detected by the deconvolution method. Across the eight bulk samples, QP deconvolution results are characterized by a low sensitivity (Fig. 3.3c). ν-SVR once again erroneously predicts the presence of cell types that are known to be biologically irrelevant to the given tissue. This is measured using a specificity metric, defined as the fraction of all false cell types that are correctly undetected by the deconvolution method. Across the eight bulk samples, ν-SVR deconvolution results are characterized by a low specificity (Fig. 3.3c). Overall, DWLS strikes the best balance between these two metrics by being able to both detect correct cell types and ignore false cell types (Fig. 3.3c).

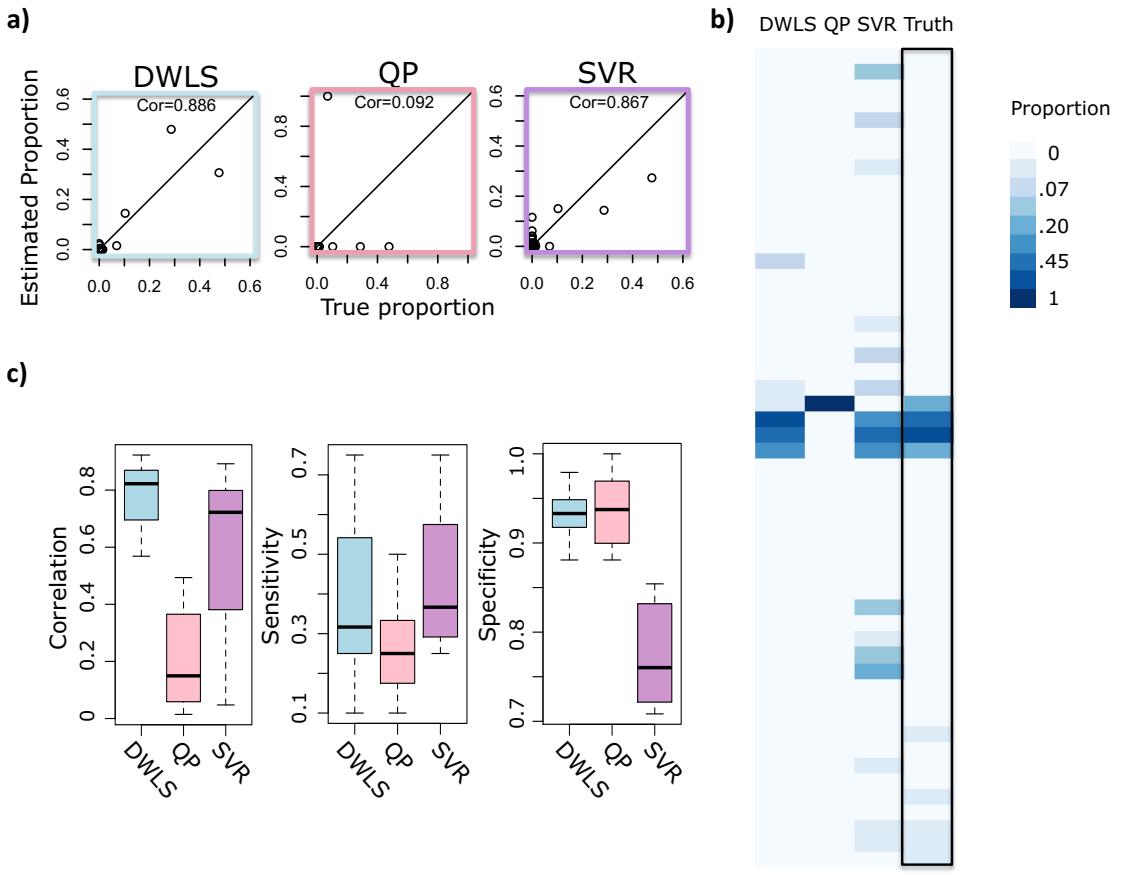


Figure 3.3. The deconvolution of eight normal mouse bulk data sets using a signature constructed from the mouse cell atlas (MCA), using three deconvolution methods: dampened weighted least squares (DWLS), quadratic programming (QP), and ν-support vector regression (ν-SVR). (a) Estimates for all cell types characterized by the MCA for a bulk mouse kidney data set, plotted against an approximate true cell type proportion as defined by the MCA data. Correlation values between true and estimated proportions are used to quantify estimation accuracy for each method. The 45-degree line in each plot represents the optimal estimate. (b) Another view of the kidney deconvolution estimates under each deconvolution method via a heatmap, where each box corresponds to a cell type proportion estimate, and a darker color corresponds a higher estimated proportion. (c) A summary of deconvolution results across all eight bulk samples, quantified by: 1. correlation between true and estimated cell type proportions for each tissue (left panel), 2. sensitivity of each deconvolution method (middle panel), and 3. specificity of each deconvolution method (right panel).

3.3.4 The deconvolution of bulk intestinal stem cell data by DWLS across various conditions accurately captures associated changes in cell type composition

One of the most important applications of deconvolution methods is in the identification of cell-type composition variations across conditions. To test the utility of our deconvolution method, we turned to a public dataset where mouse intestinal stem cell (ISC) compartments are perturbed by drug treatments. In particular, Yan et al. [111] explore the effects of R-spondin ligand (RSPO1-4) inhibition and gain-of-function on intestinal stem cell regeneration and differentiation through bulk gene expression profiling. Since bulk RNA-seq analysis alone does not provide information regarding cell-type composition, they followed up with single-cell RNA-seq analysis, and observed dramatic changes of cell-type composition in four distinct cell-type compartments: non-cycling ISC, cycling ISC, transit amplifying (TA), and differentiated cells. Here we use this dataset to test whether our deconvolution method can reveal such changes based on bulk RNA-seq data alone.

We applied DWLS to estimate the cell-type composition changes due to these drug treatments, using the single-cell data only to estimate the cell-type-specific gene expression signature matrix (Fig. 3.4). We found that treatment with Ad-LGR5-ECD almost entirely removed the intestinal stem cell population (on average, from 53.3% to 1.76%), while increasing the proportion of transit amplifying cells by 2.07-fold (25.5% to 52.8%) on average and differentiated cell types by 2.15-fold

(21.1% to 45.4%) on average. On the other hand, treatment with Ad-RSPO1 completely removed the transit amplifying cell population, while increasing the size of the intestinal stem cell population by an average 1.50-fold (53.3% to 79.8%). These observations are highly consistent with the single-cell RNA-seq data, which were used to deduce the biological functions of these treatments. That is, Ad-LGR5-ECD treatment drives differentiation, while Ad-RSPO1-treatment promotes stem-cell renewal. Here, we were able to draw the same conclusions without the need to generate single-cell RNA-seq data from every condition.

In comparison, inconsistencies arose when estimation was performed using QP and ν-SVR approaches. Specifically, neither method was consistently able to detect any cycling intestinal stem cells, whose proportion was estimated to be 29% in the control condition and 44% in the Ad-RSPO1 condition based on the single-cell RNA-seq data, and on average 31.8% and 31.4% according to the DWLS estimates. ν-SVR also predicted an increase in differentiated cell types due to Ad-RSPO1 treatment (7.64% to 45.2%), which is inconsistent with the results of the other estimation methods, the single-cell RNA-seq data, and the underlying biological mechanisms [111].

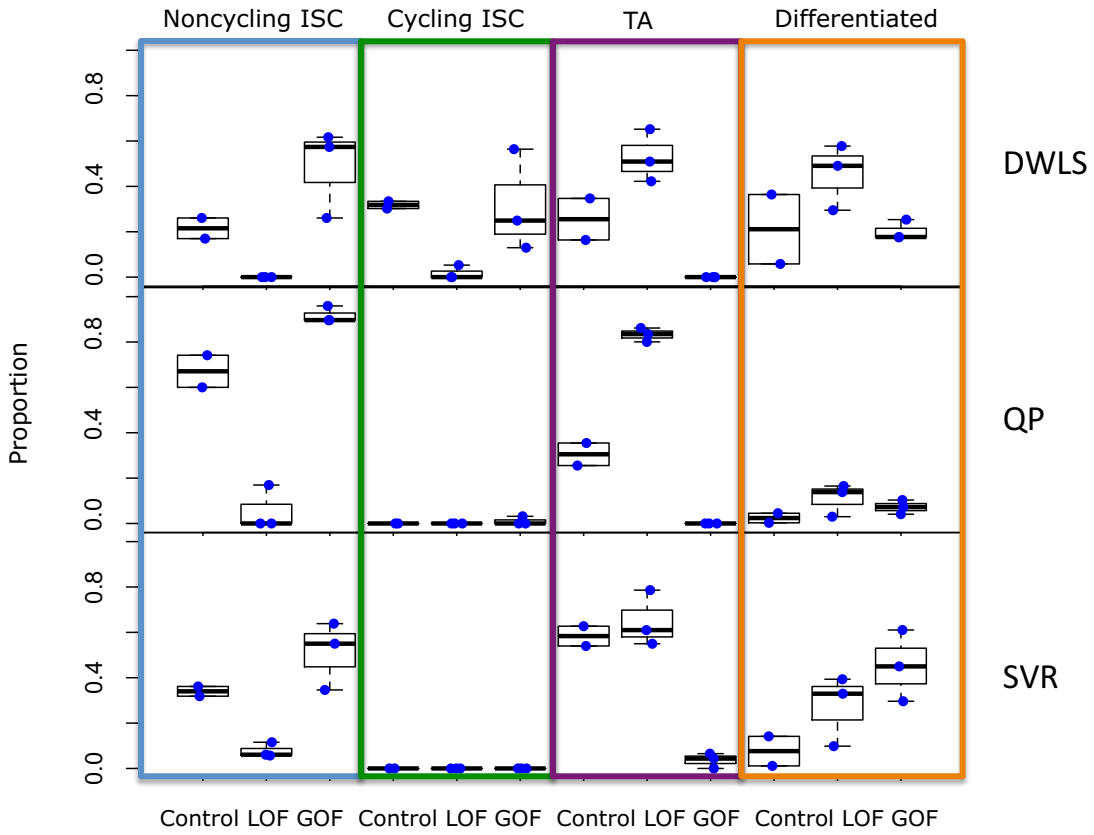


Figure 3.4. Deconvolution estimates of bulk mouse intestinal stem cell data for dampened weighted least squares (DWLS), quadratic programming (QP), and ν-support vector regression (ν-SVR) deconvolution methods, across various conditions. The Control condition corresponds to Lgr5- eGFP+ intestine cells 1.5 days post treatment with Ad-Fc, the loss of function (LOF) condition corresponds to Lgr5- eGFP+ intestine cells 1.5 days post treatment with Ad-LGR5-ECD, and the gain of function (GOF) condition corresponds to Lgr5- eGFP+ intestine cells 1.5 days post treatment with Ad-RSPO1. Each point corresponds to the deconvolution estimate of a cell type for a single bulk data set. Cell types include cycling and non-cycling intestinal stem cells (ISCs), transit amplifying (TA) cells, and various differentiated cell types.

3.4 Conclusion

Cellular heterogeneity must be taken into account when comparing gene expression data from bulk samples. As large efforts are under way to thoroughly characterize cell types of different organisms through single-cell analyses [112], we are facing a new opportunity to systematically quantify cell-type composition using the detected cell-type signatures. We envision that such deconvolution methods will be routinely used to precisely determine gene expression pattern changes in development and disease. Towards this goal, we have developed a new and more accurate computational method for deconvolution.

Using the mouse cell atlas dataset as an example, we have demonstrated that the tissue of origin of a bulk sample can be accurately predicted from deconvolution given a comprehensive signature of all cell types in an organism. In the meantime, we also recognize the danger of detecting irrelevant cell types, which is especially acute when many irrelevant cell types are included in the signature. Cell types from different tissues may share similar functions and therefore may be difficult to differentiate due to high collinearity. To minimize this risk, we advise that after a general deconvolution with a broad signature, irrelevant cell types be removed from the signature matrix to build a more specific signature matrix from only the most appropriate single-cell data sets. Such a multi-step approach may result in both more specific cell-type designations and more accurate estimates, although further investigation is needed to validate this approach.

At the other end of the spectrum, deconvolution accuracy is always dependent on the completeness of the cell-type signature, and incomplete cell-type information will compromise estimates of all cell types in the signature. Care must

always be taken to create the most appropriate signature matrix given the extent of information known about the sample. Overall, the flexibility of signature matrix definitions made possible by large quantities of single-cell data has promising implications.

Another challenge in deconvolution is the accurate estimation of rare cell types. In part, this is because detecting rare cell types from a large population in single-cell data is a challenging task, and precise signatures are difficult to build [113-116]. Additionally, the estimation of rare cell proportions by deconvolution is notoriously difficult due to the increased stochasticity of small sample sizes [94]. While our method presents an improvement over previous methods in rare cell-type detection, we hope to further improve rare cell-type detection accuracy in future work.

3.5 Methods

The DWLS method is implemented in R and is available at: https://github.com/dtsoucas/DWLS.

3.5.1 Creation of the signature matrix

The cell-type signature matrix is constructed using a representative single-cell data set, such that all cell types expected in the bulk data are also represented in the single-cell data (the converse need not be true). The single-cell data is first

clustered to reveal its constituent cell types. The optimal clustering method is dependent on the data set, but generally, a rare-cell-type-sensitive clustering method is preferred [113-116]. Further inspection of differentially expressed genes between each of these clusters is important, as this will confirm whether the detected clusters consist of biologically relevant cell types. Upon characterization of the cell types, differential expression analysis is performed to identify marker genes for each cell type. We define marker genes as genes with an FDR adjusted p-value of less than 0.01 (defined using the hurdle model in the MAST R package), and a log2 mean fold change greater than 0.5. For very large single-cell data sets like the Mouse Cell Atlas, p-values are instead determined using the Seurat R package under the "bimod" likelihood ratio test for single-cell gene expression [117], due to the faster runtime. To create the final signature matrix S, we create many candidate matrices (151 in total), which include between 50 and 200 marker genes from each cell type. The expression values of these chosen genes are averaged across each cell type, so that each resulting candidate matrix is an n by k matrix, where n is the number of genes and k is the number of cell types. The final signature matrix S is chosen as the candidate matrix with the lowest condition number, in a manner similar to CIBERSORT [101].

3.5.2 Derivation of weighted least squares

To be more precise, we rewrite the deconvolution problem as $\hat{S}\hat{x} = t$, where $\hat{S}$ is the signature matrix derived above, $\hat{x}$ is the estimated cell type number, and t is

the bulk data. Most notably, $\hat{S}$ is used to denote that the single-cell-derived signature is only an estimate of the true cell type signature, S, which is unknown. Similarly, $\hat{x}$ is the solution to $\hat{S}\hat{x} = t$, which will almost always differ from the true cell type number, x, which is only known in the case of simulated bulk data. Suppose we have k cell types and n signature genes. Let $t = (t_1 t_2 \dots t_n)'$, $\hat{x} = (\hat{x}_1 \hat{x}_2 \dots \hat{x}_k)'$, and $\hat{S} = \begin{bmatrix} \hat{S}_{11} & \dots & \hat{S}_{1k} \\ \vdots & \ddots & \vdots \\ \hat{S}_{n1} & \dots & \hat{S}_{nk} \end{bmatrix}$. This system of equations can be solved in various ways.

In the traditional setting, we obtain an estimate, $\hat{x}$, of the true cell type x by minimizing the squared error:

$$\hat{\mathrm{x}} = \underset{\tilde{x}}{\operatorname{argmin}}\, Err\left(t, \hat{S}, \tilde{x}\right) = \underset{\tilde{x}}{\operatorname{argmin}} \sum_{i=1}^{n} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2$$

Assume $\tilde{x}_j = \frac{x_j}{x_1}\tilde{x}_1$, for $j = 2, \dots, k$. Then,

$$Err = \sum_{i=1}^{n} (t_i - \hat{S}_{i1}\tilde{x}_1 - \sum_{j=2}^{k} \hat{S}_{ij}\frac{x_j}{x_1}\tilde{x}_1)^2$$

$$= \sum_{i=1}^{n} (t_i - \hat{S}_{i1}\tilde{x}_1\hat{k}_i)^2, where\ \hat{k}_i = (1 + \sum_{j=2}^{k} \frac{x_j}{x_1}\frac{\hat{S}_{ij}}{\hat{S}_{i1}})$$

$$= \sum_{i=1}^{n} (S_{i1}x_1k_i - \hat{S}_{i1}\tilde{x}_1\hat{k}_i)^2\ , where\ k_i = (1 + \sum_{j=2}^{k} \frac{x_j}{x_1}\frac{S_{ij}}{S_{i1}})$$

$$= \sum_{i=1}^{n} (S_{i1}x_1k_i - \hat{S}_{i1}x_1\hat{k}_i + \hat{S}_{i1}x_1\hat{k}_i - \hat{S}_{i1}\tilde{x}_1\hat{k}_i)^2$$

$$= \sum_{i=1}^{n} (S_{i1}x_1k_i - \hat{S}_{i1}x_1\hat{k}_i)^2 + (\hat{S}_{i1}x_1\hat{k}_i - \hat{S}_{i1}\tilde{x}_1\hat{k}_i)^2$$

(assuming orthogonality of the cross terms)

$$= \sum_{i=1}^{n} (x_1(S_{i1}k_i - \hat{S}_{i1}\hat{k}_i))^2 + (\hat{S}_{i1}\hat{k}_i(x_1 - \tilde{x}_1))^2$$

The first term is driven by a difference between the true and estimated signatures, and this error cannot be controlled for. We concern ourselves with the second term:

$$Err \approx \sum_{i=1}^{n} \left(\hat{S}_{i1}\hat{k}_i(x_1 - \tilde{x}_1)\right)^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \hat{S}_{ij}x_j\right)^2 \left(\frac{(x_1 - \tilde{x}_1)}{x_1}\right)^2$$

We can see that this error term corresponds to the relative error of estimation for cell type 1, multiplied by a function of *S* and *x* such that genes with high expression and genes pertaining to prevalent cell types will have a larger impact on the error term. Because we would like all cell types to be estimated with equal accuracy, we would like the error term to be a function of the relative error of estimation only. To mitigate this problem, we use a weighted least squares approach to solve the equation, which is represented as the following optimization:

$$\min_{\tilde{x}} \sum_{i=1}^{n} w_i \left(t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j\right)^2$$

The weights are chosen as to remove the extra term in the error function above. If we let:

$$w_i = \frac{1}{(\sum_{j=1}^{k} \hat{S}_{ij}x_j)^2} = \frac{1}{(\hat{S}_{i.}x)^2}$$

we are now minimizing:

$$Err = \sum_{i=1}^{n} w_i \, (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2$$

$$\approx \sum_{i=1}^{n} \frac{1}{(\hat{S}_{i.}x)^2} (\hat{S}_{i.}x)^2 \left(\frac{(x_1 - \tilde{x}_1)}{x_1}\right)^2 = \sum_{i=1}^{n} \left(\frac{(x_1 - \tilde{x}_1)}{x_1}\right)^2 = n\left(\frac{(x_1 - \tilde{x}_1)}{x_1}\right)^2,$$

such that the total error is now a function of the relative error in cell type number for cell type 1. Without loss of generality, we can similarly show this relationship for any cell type j ∈ $\{1, \dots, k\}$, such that

$$Err \approx n\left(\frac{(x_j - \tilde{x}_j)}{x_j}\right)^2 \forall \text{ j } \in \{1, \dots, k\}$$

Compared to the ordinary least squares approach, the relative error is reduced.

3.5.3 Additional adjustments to improve performance

Using the framework derived above, we would like to formulate the estimation of cell type proportion as a weighted least squares problem with weights $w_i = \frac{1}{(\hat{S}_{i.}x)^2}$. Several modifications are required to make this a viable approach:

1. The weights are a function of x, the true cell type number, which is unknown. We can approximate this with our estimated cell type number, $\hat{x}$, but since this also the variable being solved for, iteration is required to reach a solution. Let:

$$\hat{x}^{(0)} = \underset{\tilde{x}}{\text{argmin}} \sum_{i=1}^{n} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2$$

$$\hat{x}^{(1)} = \underset{\tilde{x}}{\text{argmin}} \sum_{i=1}^{n} w_i^{(1)} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2 \text{ , where } w_i^{(1)} = \frac{1}{(\hat{S}_{i.}\hat{x}^{(0)})^2}$$

…

$$\hat{x}^{(l)} = \underset{\tilde{x}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i^{(l)} (t_i - \sum_{j=1}^{k} \hat{S}_{ij} \tilde{x}_j)^2 \, , where \; w_i^{(l)} = \frac{1}{(\hat{S}_{i.} \hat{x}^{(l-1)})^2}$$

Convergence is reached when $\left|\left|\hat{x}^{(l)} - \hat{x}^{(l-1)}\right|\right| \leq .01$.

2. The weights are unbounded from above and may approach infinity in the case of very rare cell types ($\hat{x} \approx 0$) and/or lowly expressing genes ($\hat{S}_{ij} \approx 0$ for all cell types). This will lead to a solution driven by only a few genes. To rectify this, a dampening constant d is introduced, which defines the maximum value any weight can take on. For ease of use, we first linearly scale the weights such that the minimum weight takes on a value of 1: $w_i^S = \frac{w_i}{\min(w_j)}, j \in \{1, \ldots n\}$. The resulting optimization is equivalent. The dampened weights $\widetilde{w}_i$ are then defined as:

$$\widetilde{w}_i = \begin{Bmatrix} w_i^S, if w_i^S < d \\ d, otherwise \end{Bmatrix}$$

Cross-validation is used to select d, as follows. The possible values for d are defined as $d = 2^q, where \; q \in \{0,1,2, \ldots \max(\text{noninfinite } \log2(\mathrm{w_i^S}))\}$. Then, 100 subsets of signature genes of half the size of the full signature gene set are randomly selected. For each subset, the cell type proportion is estimated using weighted least squares on the dampened weights, for each possible value of d. The variance of the estimates over the 100 subsets for each choice of d is calculated, and the d that leads to the lowest variance is selected.

3. As specified above, $\hat{x}$ need not be positive. However, cell type numbers are inherently nonnegative. To set a constraint on $\hat{x}$, such that $\hat{x}_j \geq 0 \; \forall j$, we

solve the constrained dampened weighted least squares problem via quadratic programming, using the function "solve.QP" in the R package "quadprog". The new minimization problem is then:

$$\min_{\tilde{x},\tilde{x}\geq 0} \sum_{i=1}^{n} \tilde{w}_i (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2$$

Jointly implementing all of these alterations, we reach the final deconvolution process:

$$\hat{x}^{(0)} = \underset{\tilde{x}}{\operatorname{argmin}} \sum_{i=1}^{n} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2$$

$$\hat{x}^{(1)} = \underset{\tilde{x},\tilde{x}\geq 0}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \tilde{w}_i^{(1)} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2 \text{, where } \tilde{w}_i^{(1)} =$$

$$damp\left(\frac{1}{\left(\hat{S}_{i\cdot}\hat{x}^{(0)}\right)^2}\right), \text{and } damp(w_i) = \left\{\begin{matrix} \frac{w_i}{\min(w_j)}, if \frac{w_i}{\min(w_j)} < d \\ d, otherwise \end{matrix}\right\}$$

...

$$\hat{x}^{(l)} = \underset{\tilde{x},\tilde{x}\geq 0}{\operatorname{argmin}} \sum_{i=1}^{n} w_i^{(l)} (t_i - \sum_{j=1}^{k} \hat{S}_{ij}\tilde{x}_j)^2 \text{, where } \tilde{w}_i^{(l)} = damp\left(\frac{1}{\left(\hat{S}_{i\cdot}\hat{x}^{(l-1)}\right)^2}\right)$$

Convergence is reached when $\left|\left|\hat{x}^{(l)} - \hat{x}^{(l-1)}\right|\right| \leq .01$.

3.5.4 Simulation details

Counts for the simulated single-cell data set are generated using a Poisson distribution, for a total of six genes and three cell types. In the first simulation, two

genes are upregulated in each cell type, where $\lambda = 50$ for an upregulated gene and $\lambda = 5$ otherwise. Fifty cells from each cell type are used to create a signature matrix, where the six genes are averaged over each cell type to create a reference gene expression profile. Between 10001 and 15000 cells are used to simulate bulk data, by summing up gene expression values across cell types. Specifically, 5000 bulk data sets are created by combining 5000 cells from cell types two and three, and between 1 and 5000 cells from cell type one. Overall, this creates bulk data sets with a rare cell type proportion spanning between 0.1% and 33.3%. Bulk data simulation is repeated 10 times for each rare cell type proportion, and all metrics reported are based on an average of these 10 samples.

In the second simulation, two genes are again upregulated in each cell type, but the mean expression level of the marker genes corresponding to the first cell type is lower, such that λ ranges from 0.05 to 10 for an upregulated gene and from .005 to 1 otherwise. Fifty cells from each cell type are again used to create a signature matrix, where gene expression levels are scaled for each choice of λ, for a total of 200 signature matrices. To simulate the bulk data, 5000 cells from each cell type are aggregated so that each cell type is present in equal proportion. Bulk data simulation is repeated 10 times for each choice of λ, and all metrics reported are based on an average of these 10 samples.

3.5.5 Estimation using other deconvolution methods

Nu-SVR

Nu-support vector regression was performed using the "svm" function in the "e1071" package in R. Parameters were set to nu=0.5, type= "nu-regression", kernel= "linear", cost=1, and all others to the default values. Bulk data and signature matrices were scaled to [-1, 1]. These parameter and scaling choices match those specified in Schelker et al. [109] in their matlab code, accessed through: https://figshare.com/s/865e694ad06d5857db4b. As in Newman et al. [101], model coefficients are extracted from the svm model using t(model$coefs) %*% model$SV, and any negative coefficients are set to zero. The coefficients are then scaled by the sum of the coefficients, such that the scaled coefficients will sum to one.

Quadratic programming

Quadratic programming is implemented using the "solve.QP" function in the "quadprog" package in R. Default parameters are used, and the constraints are specified such that all coefficients must be greater than or equal to zero.

3.5.6 Data sources and processing

MCA Bulk RNA-seq library construction, sequencing, and processing

6-10 week-old male C57BL/6J mice were purchased from the Shanghai Laboratory Animal Center (SLAC). From each mouse, 4 non-sexual tissues (liver, small intestine, lung, kidney) were excised. The excised tissues were immediately washed in PBS. After washing, each tissue was ground into powder with liquid N2.

RNA extraction was performed using Trizol. We used mRNA Capture Beads (VAHTS mRNA-seq v2 Library Prep Kit for Illumina, Vazyme) to extract mRNA from total RNA. A PrimeScript Double Strand cDNA Synthesis Kit (TaKaRa) was used to synthesize double-stranded cDNA from purified polyadenylated mRNA templates according to the manufacturer's protocol. We used TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme) to prepare cDNA libraries for Illumina sequencing (VeritasGenetics).

Bulk sequencing reads containing multiplexed data were filtered using the bbduk function of the bbmap tool to select reads containing the appropriate sample index. STAR 2.5.3a [118] with default parameters was used to map filtered reads to the Ensembl release 75 mouse reference genome. Aligned reads were normalized by library size to fragments per kilobase of transcript per million mapped reads (FPKM) using the "fpkm" function in the "DESeq2" package in R.

Mouse Cell Atlas single-cell data

The mouse cell atlas (MCA) single-cell data [110] and annotations were accessed through: https://figshare.com/s/865e694ad06d5857db4b. The single-cell data is quantified as UMI counts. The signature matrix was built using the 61k cell subset consisting of randomly sampled cells from 43 tissues. Cell types were defined by collapsing the 98 clusters identified by Han et al. into 52 unique cell types.

Intestinal stem cell bulk and single-cell data

Intestinal stem cell (ISC) single-cell and bulk RNA-seq data sets from Yan et al. [111] were accessed through the Gene Expression Omnibus (GEO) repository under accession numbers GSE92865 and GSE92377, respectively. The single-cell data is quantified as UMI counts. All Lgr5-eGFP+ and Lgr5-eGFP- cells were used in the construction of the signature matrix. The single-cell data cell type labels shown in Yan et al. Fig. 5a [111] were obtained from the authors upon request, and these were used to generate the signature matrix. The bulk data is quantified in terms of FPKM values.

3.5.7 Schelker et al. simulation details

Source code and data from the Schelker et al. {Schelker 2017} simulation analysis was accessed through: https://figshare.com/s/711d3fb2bd3288c8483a. The single-cell RNA-seq data used in Schelker et al. [109] includes tumor cells from 19 melanoma patients, PBMCs from four healthy subjects, and ascite samples from four ovarian cancer patients. A signature matrix was built using all cells, using the clusters found by DBSCAN in Schelker et al. [109], and using the genes from the CIBERSORT immune cell signature [101]. 27 patient-specific simulated bulk data sets were built by summing up gene expression values of signature genes across all cell types, for each patient.

3.5.8 Modified relative percent error calculation

Modified relative percent error measures the absolute difference between estimated and true cell type proportions, normalized by the mean of the estimated and true cell type proportions. A pseudo count of 0.005 is added so that for very small cell type proportions, relative error does not become unreasonably high. It is defined as:

$$MRPE = \left\{ \begin{array}{c} 0, if\ x_l = 0\ and\ \hat{x}_l = 0 \\ \dfrac{\left| \dfrac{x_l}{\sum_{j=1}^{k} x_j} - \dfrac{\hat{x}_l}{\sum_{j=1}^{k} \hat{x}_j} \right|}{\dfrac{\dfrac{x_l}{\sum_{j=1}^{k} x_j} + \dfrac{\hat{x}_l}{\sum_{j=1}^{k} \hat{x}_j}}{2} + 0.005} * 100, otherwise \end{array} \right\},$$

where x_l and $\hat{x}_l$ are the true and estimated cell type numbers, respectively, for cell type *l*.

Hands-on applications of single-cell RNA-sequencing technologies to discoveries in cancer immunology

4.1 Antibody-mediated inhibition of MICA and MICB shedding promotes NK cell-driven tumor immunity

By Lucas Ferrari de Andrade, Rong En Tay, Deng Pan, Adrienne M. Luoma, Yoshinaga Ito, Soumya Badrinath, Daphne Tsoucas, Bettina Franz, Kenneth F. May Jr., Christopher J. Harvey, Sebastian Kobold, Jason W. Pyrdol, Charles Yoon, Guo-Cheng Yuan, F. Stephen Hodi, Glenn Dranoff, Kai W. Wucherpfennig

The full text is published in *Science*, Vol. 359, Issue 6383, March 2018, pp. 1537-1542.

When cells are damaged, they normally express MICA and MICB proteins that flag the cells for removal. These cell stress markers bind to the natural killer group 2D (NKG2D) receptor of cytotoxic lymphocytes, recruiting these cytotoxic lymphocytes for the destruction of damaged cells. However, cancer cells are able to cleave MICA and MICB in order to escape this immune response. In order to prevent

MICA/B shedding, our collaborators investigated treatment with MICA/B $\alpha 3$ domain-specific antibody 7C6-hIgG1. They showed that MICA/B antibody treatment decreases tumor growth and metastasis in human and mouse models (Fig. 4.1).

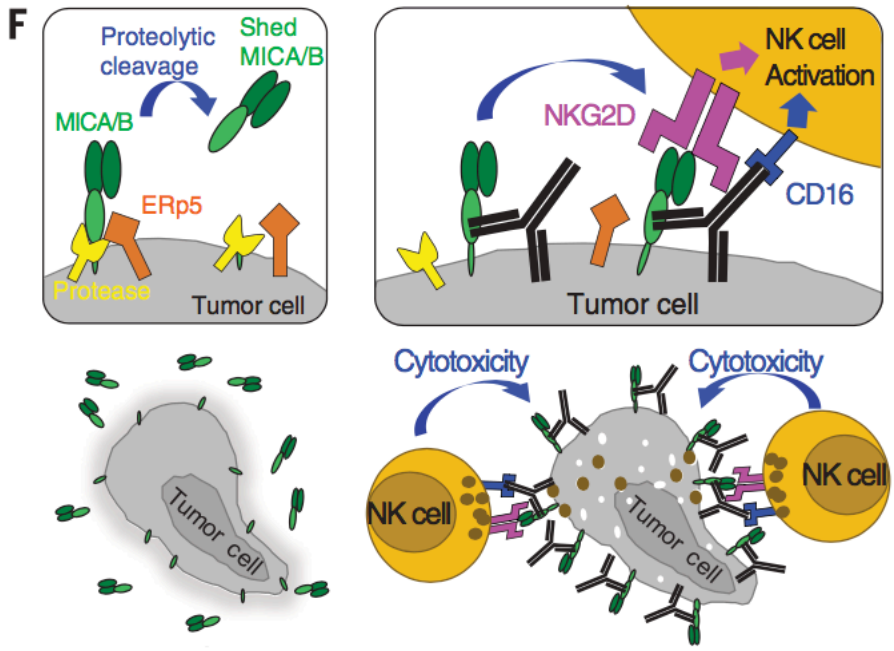


Figure 4.1. A cartoon depicting the mechanisms behind NK cell recognition of damaged cells. On the left, tumor cells shed MICA/B through proteolytic cleavage, and escape detection by NK cells. On the right, antibody binding at the MICA/B $\alpha 3$ domain prevents cleavage of MICA/B, resulting in recruitment of NK cells through the NKG2D receptor.

To further explore the effects of such treatment on the tumor microenvironment, we analyzed single-cell RNA-sequencing data from sorted group 1 innate lymphoid cells (ILCs) from mouse metastatic lung tissue, both untreated and treated with 7C6-mIgG2a. Through clustering, tSNE visualization, and differential expression analyses, we saw a striking difference in cell type composition between the two states (Fig. 4.2). The treated sample was composed of

63.2% activated, cytotoxic NK cells, characterized by high expression of EOMES (eomesodermin), GZMA (granzyme A), GZMB (granzyme B), and PRF1 (perforin 1), up from 6.41% in the control condition (Isotype). On the other hand, the control sample was composed of 49.4% ILC1 cells characterized by high expression of CXCR3, CXCR6, and LTB (lymphotoxin β), associated with cytokine and chemokine signaling and inflammation, up from 7.69% in the treated sample. These differences support the finding that MICA/B antibody treatment prevents MICA/B cleavage, and results in increased activity of tumor-infiltrating NK cells.

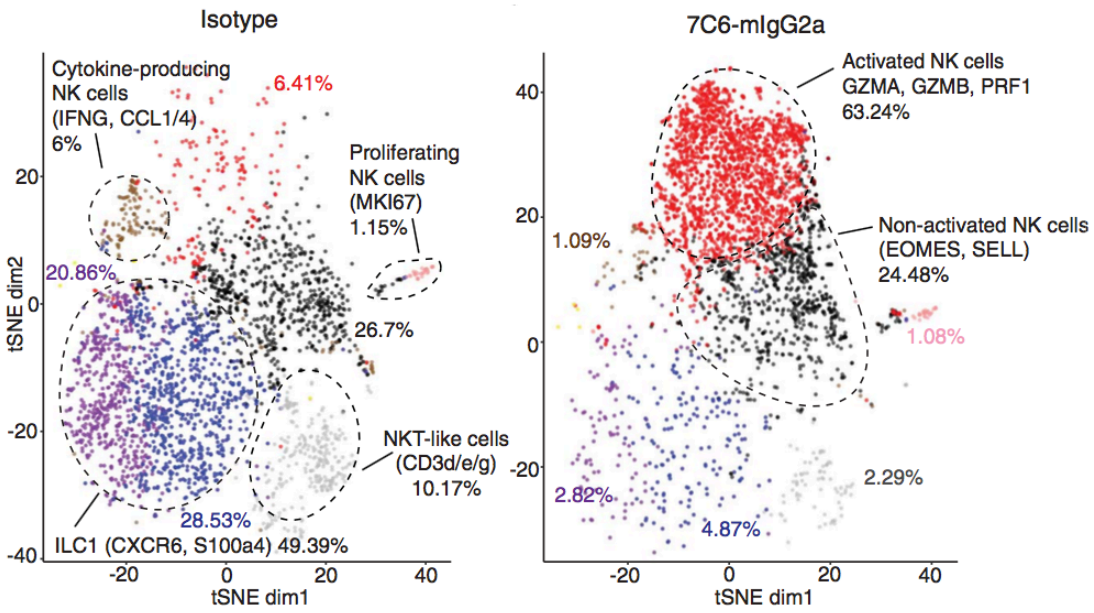


Figure 4.2. A t-SNE visualization of group 1 innate lymphoid cells (ILCs) for control (Isotype, left), and antibody-treated (7C6-mIgG2a, right) samples. There is a striking increase in the proportion of activated NK cells upon antibody treatment.

4.2 A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing

By Deng Pan, Aya Kobayashi, Peng Jiang, Lucas Ferrari de Andrade, Rong En Tay, Adrienne Luoma, Daphne Tsoucas, Xintao Qiu, Klothilda Lim, Prakash Rao, Henry W. Long, Guo-Cheng Yuan, John Doench, Myles Brown, Shirley Liu, Kai W. Wucherpfennig

The full text is published in *Science*, Vol. 359, Issue 6377, February 2018, pp. 770-775.

Tumor cells are often resistant to killing by cytotoxic T cells, and the mechanisms of resistance are unclear. In order to find which genes and processes contribute to this resistance, our collaborators performed a genome-wide CRISPR/Cas9 screen on mouse melanoma cells. Inactivation of certain genes, including *Pbrm1*, *Arid2*, and *Brd7* of the SWI/SNF chromatin-remodeling complex, results in greater recruitment of effector T cells. Specifically, inactivation of these genes results in loss of function of the PBAF complex, a part of the SWI/SNF complex, which then increases chromatin accessibility to promoters and enhancers of interferon-γ induced genes. This in turn results in greater secretion of chemokines that recruit effector T cells.

To gain a greater understanding of this process, we analyze single-cell RNA-seq data from sorted CD45+ immune cells from both *Pbrm1*-deficient and control mouse tumors (Fig. 4.3). A comparison of the two conditions shows an increased percentage of dendritic cells and tumor-inhibitory M1-like macrophages in the *Pbrm1*-deficient mice, and an increased percentage of tumor-promoting M2-like macrophages in the control mice. Additionally, a gene set enrichment analysis finds that the *Pbrm1*-deficient mice express a high level of genes related to anti-tumor immunity relative to the control. Our analysis demonstrates that *Pbrm1* deficiency alters the tumor microenvironment, making it more receptive to T cell therapies.

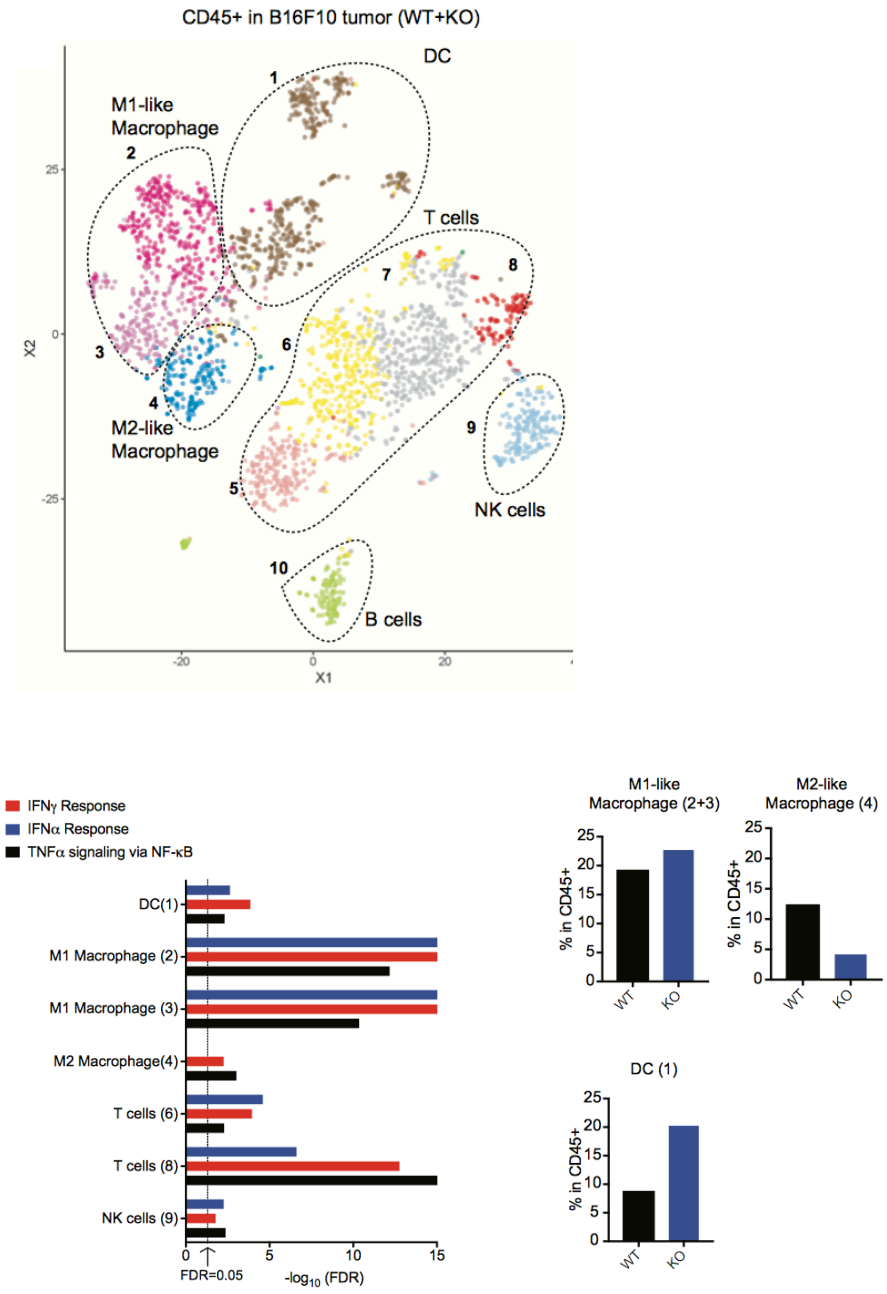


Figure 4.3. Single-cell RNA-sequencing data from CD45+ immune cells from *Pbrm1*-deficient and control mouse tumors. Top, a t-SNE visualization of cells from pooled wild type and knock-out conditions. Bottom left, a gene set enrichment analysis of the wild type condition shows enrichment for signatures associated with anti-tumor immunity. Bottom right, proportions of macrophage and dendritic cell populations in wild type and knock-out conditions.

References

References from Chapter 1

1. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011; 17:297–303.

2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674.

3. Janiszewska M, Liu L, Almendro V, Kuang Y, Paweletz C, Sakr RA, Weigelt B, Hanker AB, Chandarlapaty S, King TA, et al. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. Nat Genet. 2015; 47:1212– 1219.

4. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008; 5:877–879.

5. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016; 26:304–319.

6. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015; 12:519–522.

7. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biol. 2016; 17:86.

8. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. 2016; 17:69.

9. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016; 352:189–196.

10. Grun D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell. 2016; 19:266–277.

11. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016; 17:144.

12. Tsoucas D, Yuan GC: GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol* 2018, 19:58.

13. Van Loo P, Voet T. Single cell analysis of cancer genomes. Curr Opin Genet Dev. 2014; 24:82–91.

14. Navin NE. Cancer genomics: one cell at a time. Genome Biol. 2014; 15:452.

15. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015; 16:133–145.

16. Saadatpour A, Lai S, Guo G, Yuan GC. Single-Cell Analysis in Cancer Genomics. Trends Genet. 2015; 31:576–586.

17. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, et al. Genome-wide copy number analysis of single cells. Nat Protoc. 2012; 7:1024– 1041.

18. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012; 148:886–895.

19. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy- number variations of a single human cell. Science. 2012; 338:1622–1626.

20. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6:377–382. [PubMed: 19349980]

21. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012; 30:777–782.

22. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell. 2010; 18:675–685.

23. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344:1396–1401.

24. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol. 2011; 29:1120–1127.

25. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts in situ. Science. 1998; 280: 5–590.

26. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods. 2014; 11:360–361.

27. Shah S, Lubeck E, Zhou W, Cai L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron. 2016; 92:342–357.

28. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015; 348:aaa6090.

29. Lovatt D, Ruble BK, Lee J, Dueck H, Kim TK, Fisher S, Francis C, Spaethling JM, Wolf JA, Grady MS, et al. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat Methods. 2014; 11:190–196.

30. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014; 343:1360–1363.

31. Easwaran H, Tsai HC, Baylin SB. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Mol Cell. 2014; 54:716–727.

32. Notaro S, Reimer D, Fiegl H, Schmid G, Wiedemair A, Rossler J, Marth C, Zeimet AG. Evaluation of folate receptor 1 (FOLR1) mRNA expression, its specific promoter methylation and global DNA hypomethylation in type I and type II ovarian cancers. BMC Cancer. 2016; 16:589.

33. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013; 23:2126–2135.

34. Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 2012; 40:e136.

35. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014; 11:817–820.

36. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348:910–914.

37. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490.

38. Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R, et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature. 2015; 528:142–146.

39. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502:59–64.

40. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. Single-cell ChIP- seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015; 33:1165–1172.

41. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N. Noise in protein expression scales with natural protein abundance. Nat Genet. 2006; 38:636–643.

42. Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, Fredriksson S, Assarsson E, Lundberg M, Nelander S, et al. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. Cell Rep. 2016; 14:380–389.

43. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33:495–502.

44. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC. High- throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol. 2015; 33:503–509.

45. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32:381–386.

46. Trapnell, C. Monocle: Differential expression and time-series analysis for single-cell RNA-Seq. 2016.

47. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016; 44:e117.

48. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming GL, et al. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell. 2015; 17:360–372.

49. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. Bifurcation analysis of single- cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014; 111:E5643–5650.

50. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016; 34:637–645.

51. Bendall SC, Davis KL, Amir e-A, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014; 157:714–725.

52. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015; 525:251–255.

53. Ester, M., Kriegel, H-P., Sander, J., Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 2nd International Conference on Knowledge Discovery and Data Mining; Portland, OR: AAAI; 1996. p. 226-231.

54. Shipitsin M, Polyak K. The cancer stem cell hypothesis: in search of definitions, markers, and relevance. Lab Invest. 2008; 88:4–463.

55. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, Zhou A, Eyob H, Balakrishnan S, Wang CY, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015; 526:131–135.

56. Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Ulkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, et al. Isolation of rare circulating tumour cells in cancer patients by microchip technology. Nature. 2007; 450:1235–1239.

57. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, Zong C, Bai H, Chapman AR, Zhao J, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci U S A. 2013; 110:21083–21088.

58. Dago AE, Stepansky A, Carlsson A, Luttgen M, Kendall J, Baslan T, Kolatkar A, Wigler M, Bethel K, Gross ME, et al. Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. PLoS One. 2014; 9:e101777.

59. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, Desai R, Fox DB, Brannigan BW, Trautwein J, et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science. 2015; 349:1351–1356.

60. Wei W, Shin YS, Xue M, Matsutani T, Masui K, Yang H, Ikegami S, Gu Y, Herrmann K, Johnson D, et al. Single-Cell Phosphoproteomics Resolves Adaptive Signaling Dynamics and Informs Targeted Combination Therapy in Glioblastoma. Cancer Cell. 2016; 29:563–573.

References from Chapter 2

61. Tsoucas D, Yuan GC. Recent progress in single-cell cancer genomics. Curr Opin Genet Dev. 2017;42:22–32.

62. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16:133–45.

63. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.

64. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.

65. Giecold G, Marco E, Garcia SP, Trippa L, Yuan GC. Robust lineage reconstruction from high-dimensional single-cell data. Nucleic Acids Res. 2016;44:e122.

66. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166:1308–1323.e1330.

67. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347:1138–42.

68. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19:335–46.

69. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5.

70. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al. De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell. 2016; 19:266–77.

71. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17:144.

72. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546:431–5.

73. Li T, Ding C. Weighted consensus clustering. In: SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics; 2008.

74. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol. 2001;63:411–23.

75. Kodinariya T, Makwana P. Review on determining number of cluster in k- means clustering. Int J. 2013;1(6):90–5.

76. Strehl A, Ghosh J. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3:583–617.

77. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.

78. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405:442–51.

79. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.

80. Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech Theory Exp:P09008.

81. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.

82. Schinnerling K, García-González P, Aguillón JC. Gene expression profiling of human monocyte-derived dendritic cells - searching for molecular regulators of tolerogenicity. Front Immunol. 2015;6:528.

83. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput. 1997;1:67–82.

84. The Human Cell Atlas. https://www.humancellatlas.org. Accessed 12 Dec 2017.

85. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: 2nd International Conference on Knowledge Discovery and Data Mining; Portland, OR. Menlo Park: AAAI; 1996. p. 226–31.

86. Ding C, He X, Simon H. On the equivalence of nonnegative matrix factorization and spectral clustering. In: SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics; 2005. p. 606–10.

87. Maaten LVD, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008; 9:2579–605.

88. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.

89. Tsoucas D, Yuan G. GiniClust2. Zenodo. 2018. https://doi.org/10.5281/zenodo.1211359.

90. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell mRNA sequencing reveals rare intestinal cell types. NCBI GEO database. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62270. Accessed 2 Apr 2018.

91. Klein A, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz D, Kirschner M. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. NCBI GEO database. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525. Accessed 2 Apr 2018.

92. Zheng G, Terry J, Belgrader P, Ryvkin P, Bent Z, Wilson R, Ziraldo S, Wheeler T, McDermott G, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. NCBI Sequence Read Archive. 2017. https://www.ncbi.nlm.nih.gov/sra/?term=SRP073767. Accessed 2 Apr 2018.

References from Chapter 3

93. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M: Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* 2010, 11:27.

94. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K: Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* 2018, 34:1969-1979.

95. Shen-Orr SS, Gaujoux R: Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr Opin Immunol* 2013, 25:571-578.

96. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF: Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* 2009, 4:e6098.

97. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al.: Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol* 2016, 17:174.

98. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD: Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* 2011, 6:e27156.

99. Gong T, Szustakowski JD: DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 2013, 29:1083-1085.

100. Zhong Y, Wan YW, Pang K, Chow LM, Liu Z: Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* 2013, 14:89.

101. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015, 12:453-457.

102. Guo G, Luc S, Marco E, Lin TW, Peng C, Kerenyi MA, Beyaz S, Kim W, Xu J, Das PP, et al: Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* 2013, 13:492-505.

103. Sandberg R: Entering the era of single-cell transcriptomics in biology and medicine. *Nat Methods* 2014, 11:22-24.

104. Kalisky T, Oriel S, Bar-Lev TH, Ben-Haim N, Trink A, Wineberg Y, Kanter I, Gilad S, Pyne S: A brief review of single-cell transcriptomic technologies. *Brief Funct Genomics* 2018, 17:64-76.

105. Saadatpour A, Lai S, Guo G, Yuan GC: Single-Cell Analysis in Cancer Genomics. *Trends Genet* 2015, 31:576-586.

106. Tsoucas D, Yuan GC: Recent progress in single-cell cancer genomics. *Curr Opin Genet Dev* 2017, 42:22-32.

107. Wang Y, Navin NE: Advances and applications of single-cell sequencing technologies. *Mol Cell* 2015, 58:598-609.

108. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al.: A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* 2016, 3:346-360.e344.

109. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, Schoeberl B, Raue A: Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat Commun* 2017, 8:2032.

110. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, et al.: Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 2018, 172:1091-1107.e1017.

111. Yan KS, Janda CY, Chang J, Zheng GXY, Larkin KA, Luca VC, Chia LA, Mah AT, Han A, Terry JM, et al.: Non-equivalence of Wnt and R-spondin ligands during Lgr5. *Nature* 2017, 545:238-242.

112. The Human Cell Atlas. https://www.humancellatlas.org. Accessed 12 Dec 2017.

113. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A: Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015, 525:251-255.

114. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al.: De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* 2016, 19:266-277.

115. Jiang L, Chen H, Pinello L, Yuan GC: GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* 2016, 17:144.

116. Tsoucas D, Yuan GC: GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol* 2018, 19:58.

117. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, Roederer M, Gottardo R: Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 2013, 29:461-467.

118. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29:15-21.

Supplemental References

119. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval: Cambridge University Press; 2008.

120. Van Rijsbergen CJ. Information retrieval: Butterworths; 1979.

121. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.

122. Tolman RC. Principles of statistical mechanics: Courier Corporation; 1938.

A. Supplemental Materials for Chapter 2

A.1 Supplemental Information

Evaluation of clustering performance using additional metrics

Several additional metrics are used to compare clustering accuracy across various clustering methods, for both simulated and subsampled data. Results for simulated data are shown in Supplemental Fig. S2.1, and results for a subset of the subsampled data sets (corresponding to a rare cell type of 1.6%) are shown in Supplemental Fig. S2.4. Full results for the subsampled data are not shown for reasons of brevity, but results for this particular rare cell type proportion are representative of the overall results. The additional metrics used to evaluate clustering accuracy are: purity, normalized mutual information (NMI), micro-averaged F-measure, adjusted rand index (ARI), and entropy. These various metrics are introduced to give a more complete view of the clustering results, as each method measures accuracy in a slightly different way. Purity is a measure of how often a cluster contains a single cell type, where 1 indicates that all clusters contain a single cell type, and a value of 0 indicates poor clustering [119]. Unfortunately, this metric does not penalize for overclustering, and a perfect clustering can be achieved by clustering each cell separately. Often, more complex metrics are required. NMI is an entropy-based method normalized by cluster size [80], where a value of 1 indicates perfect agreement, whereas a value of 0 means the performance is as good as random guess. The micro-averaged F-measure is the harmonic mean

of micro-averaged precision and recall rates, which are computed by summing true positive, true negative, false positive and false negative values over all cell types. Values also range from 0 to 1, with values close to 1 implying better clustering [120]. The ARI is a version of the rand index that is corrected for chance, and also takes into account both false positives and negatives. A value of 1 indicates a perfect clustering, a value of 0 is the expected value for a random clustering, and the metric takes on negative values if the clustering is worse than expected [121]. Entropy is a measure of disorder within the clustering results that ranges from 0 to 1, where values close to 0 imply less disorder and better clustering [122]. As this is the only metric where a lower value means a better clustering, we show 1-entropy for a more intuitive visualization.

Results for a naïve combination of high Fano and Gini genes in simulated data

The most obvious approach to combining the superior performances of Gini and Fano-based feature spaces for detecting rare and common cell types, respectively, may be to combine the two feature spaces, and perform clustering on this combined space. Supplemental Fig. S2.3 shows a two-dimensional tSNE of the Jaccard distance of this combined feature space, colored with the true cell clusters, followed by clustering results on this same space using DBSCAN and k-means clustering methods. Neither of these clustering methods is able to recapitulate all six clusters, and the visualization gives an indication as to why. The two larger clusters are visually separable, but the smaller clusters are in close proximity to the largest cluster, and are indistinguishable from each other. This combination of two

distinct feature spaces is undesirable because it dilutes the signal from each feature space, and further demonstrates the need for a consensus clustering approach.

Parameter choice details

Parameter choice for DBSCAN

DBSCAN has two parameters: MinPts and eps. MinPts is specified as 3 for all data sets except for the PBMC data, where MinPts is set to 100, in accordance with the larger size of this data set. This corresponds to the minimum cluster size for which we would expect to see a biologically relevant cluster. In general, we find an appropriate MinPts specification to be about 0.1% of the total number of cells.

The eps parameter is determined by a k-nearest-neighbors (kNN) plot as recommended by the authors of DBSCAN [85]. According to their approach, distance from each point to its kth nearest neighbor is plotted in ascending order, where k=MinPts. This will form a line featuring an inflection point, at which lies the recommended choice for eps [85]. If multiple inflection points exist, this may suggest that multiple values of eps are worth exploring; however, in our case as we are concerned with rare clusters, we only consider the smallest choice of eps, corresponding to the first inflection point. Here, our Gini-based distance metric is particularly low-dimensional due to the use of Jaccard distance and a small number of high Gini genes. This causes cells with similar expression profiles to have pairwise distances of zero, which distorts the traditional kNN-distance curve shape and makes the inflection point harder to visualize (Supplemental Fig. S2.7). We provide an alternate numerical approach to approximating the inflection point: after

removing all zero distances, the inflection point roughly corresponds to the kNN-distance of the (0.00125*total number of cells*MinPts)th cell. For the 68k data set, the computation of all kNN distances was prohibitive, so we subsampled 2057 cells and computed 3-NN distances to maintain the ratio of MinPts to the total sample size.

Choice of k for k-means clustering

We give the option of automatically determining k using the gap statistic. However, observing differentially expressed genes and visualizing k-means clusters gives the best intuition as to the optimal k. We also do not suggest using the gap statistic for large data sets due to its computational demands. For the simulated data, we chose k=2, in accordance with both the number of large clusters and the gap statistic, but show that k=3 will also yield the same result (Supplemental Fig. S2.2). For subsampled PBMC data sets, k was chosen as 2 or 3 depending on the ability of k-means to pick up the rare NK cell group. k was chosen as 2 for the day 4 post-LIF mouse embryonic stem cell data as we found this number to best group biologically meaningful cell types.

For the 68k PBMC data, we chose k=10 to allow for direct comparison with clustering results from Zheng et al. [77]. For comparison to the k=10 parameter choice, we additionally show results for both choices k=8 and k=12 for the Fano-based clustering step (Supplemental Fig. S2.8). All parameter choices perform comparably, with NMI values of 0.542, 0.541 and 0.498, respectively, when compared to the reference labels. The k=8 clustering results in two fewer clusters

within the CD56+ NK, CD8+ Cytotoxic T, CD8+/CD45RA+ Naïve Cytotoxic continuum. The k=12 clustering further splits clusters containing overlapping CD8+/CD45RA+ Naïve Cytotoxic, CD4+/CD45RA+/CD25- Naïve T, CD4+/CD25 T Reg, and CD4+/CD45RO+ Memory cells, as well as adding another cluster to the aforementioned CD56+ NK, CD8+ Cytotoxic T, CD8+/CD45RA+ Naïve Cytotoxic continuum. These changes are minor as they occur predominantly in regions of unclear identity.

Parameter choices for weighted consensus clustering

As discussed in the Materials and Methods section, the parameter values for μ, s, and f are derived through intermediate variables μ', s', and f'. The values for these intermediate variables are determined empirically using the following procedure. First, we set μ' = 4*(MinPts/total number of cells), where MinPts represents the minimum cluster size allowed by DBSCAN. We find that this is the approximate cell fraction where GiniClust and Fano-factor-based clustering perform equally. Next, using the same logic, we set the value for s' such that the 99th percentile of the GiniClust weighting distribution is reached at 6*(MinPts/total number of cells). We find that this is the approximate cell fraction in which GiniClust can no longer detect the rare cell type. Finally, we set f' = 0.1. While these parameter settings cannot guarantee optimal performance, results from our sensitivity analysis (see next section) strongly suggest that the clustering results are robust over a wide range of parameter values.

Sensitivity analysis on simulated data

GiniClust2 parameters were varied one at a time on simulated data to test the robustness of the method to specific parameter choices. The following parameters were varied: DBSCAN parameters MinPts and eps, k-means parameter k, Gini and Fano gene thresholds, and weighting scheme parameters μ, s, and f. In addition, the behavior of GiniClust2 across various signal:noise ratios was evaluated by running the method on several variations of the original data set. The noise level was simulated by varying the scale parameter of the generative negative binomial distribution (see Methods). Clustering accuracy was evaluated using several metrics: NMI, ARI, entropy, purity, and the micro-averaged F-measure. Results of these analyses are shown in Supplemental Fig. S2.9.

Our analysis suggests that the clustering results are strongly affected by the choice of k. A small k results in combining the large clusters, while a large k results in splitting the large clusters into smaller subgroups. This resolution uncertainty is intrinsic to all clustering methods. For all other parameter changes, metrics do not dip below 0.96, indicating the robustness of the clustering results to these parameter choices. Perhaps more importantly, clustering results are perfect over a wide range of many of these parameter values.

Analysis of the 10X Genomics data supports a logistic function model

We test whether the consensus clustering weighting function $w_i^G(x_i)$ accurately represents the power of GiniClust to detect rare cell types over a range of cell type proportions. In the Results section we discuss a subsampling analysis

performed by selecting macrophage, NK and B cells at varying proportions from a 10X Genomics dataset consisting of about 68,000 peripheral blood mononuclear cells (PBMCs) [77]. Cells are classified based on transcriptomic similarity with purified cell-types and additional known gene markers (see Methods for full details). Cell types are sampled 140 times according to Supplemental Table S2.1 such that the rare macrophage group ranges in cell type proportion from 0.2% to 11.6%.

To capture the power of GiniClust and Fano-factor-based k-means to detect cell types of varying rarity, we define "detection" of the rare cell type as the clustering together of at least 3 out of the 5 rare cells, while including at most 2 other cells in this rare group. For the subsampled PBMC data we calculate the rare cell type detection rates of both GiniClust and Fano-factor-based k-means for each of the rare cell type proportions (Supplemental Fig. S2.10a). We next calculate the ratio between the GiniClust detection ability and the sum of both GiniClust and Fano-factor-based k-means detection abilities (Supplemental Fig. S2.10b). This is a measure of the ability of GiniClust over Fano-factor-based k-means in detecting the rare cell type, which we tried to capture in our GiniClust weighting function. We can see that the shape of the curve in Supplemental Fig. S2.10b closely mimics that of the logistic GiniClust weighting function, also pictured in Supplemental Fig. S2.10b, and suggests that such a logistic function shape is appropriate for defining GiniClust weights.

Comparison of the computational performance of GiniClust2 and RaceID2

GiniClust2 and RaceID2 were run on all smaller data sets (<=3023 cells) using a 2.5 GHz Intel Core i7 CPU with 16 GB memory. Runtimes for these methods are shown in Supplemental Fig. S2.6. Both methods were run using default—and where applicable, automatic—parameter choices. For datasets above 155 cells, GiniClust2 is faster than RaceID2, and scales better than RaceID2 for an increasing number of cells. Only GiniClust2 can be run for very large data sets (68k cells), and therefore, a comparison cannot be shown. It should be noted that for these large data sets, a few code alterations (see Methods) make running GiniClust2 a faster process without sacrificing accuracy, and the runtime does not scale with the runtimes for these smaller data sets.

A.2 Supplemental Figures

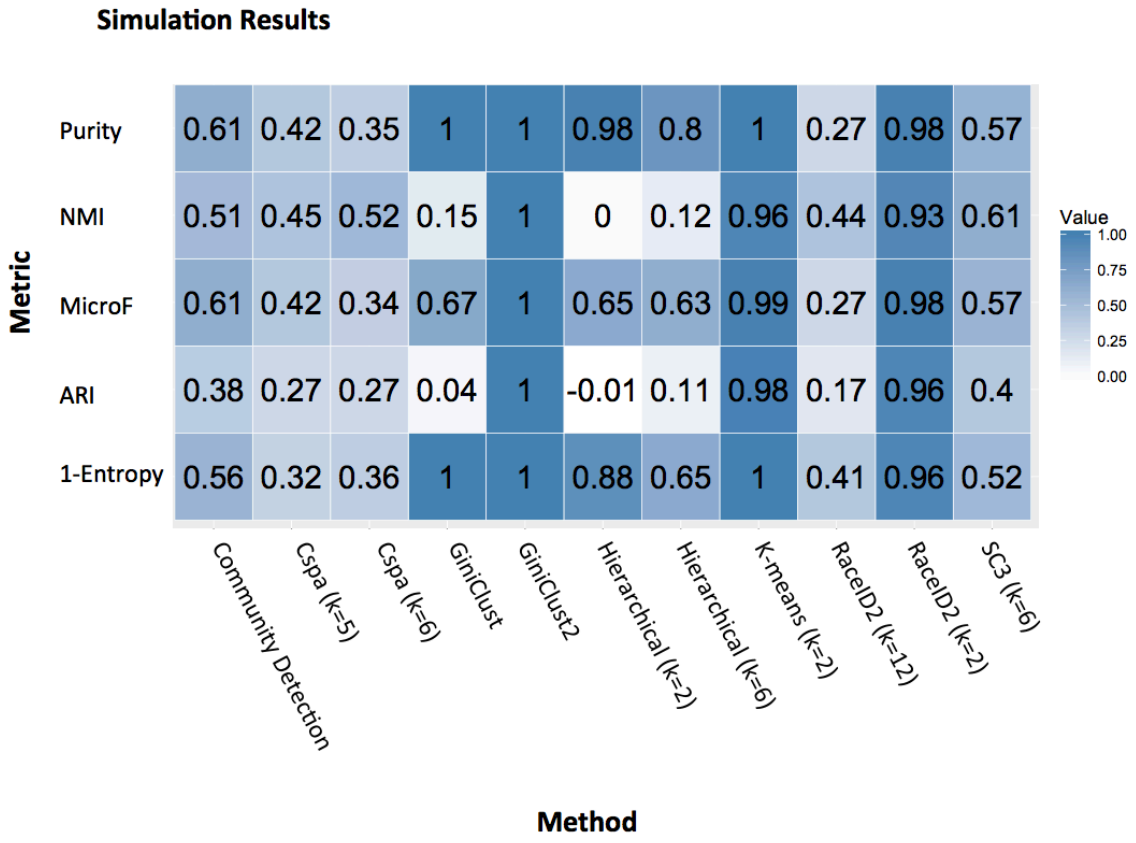
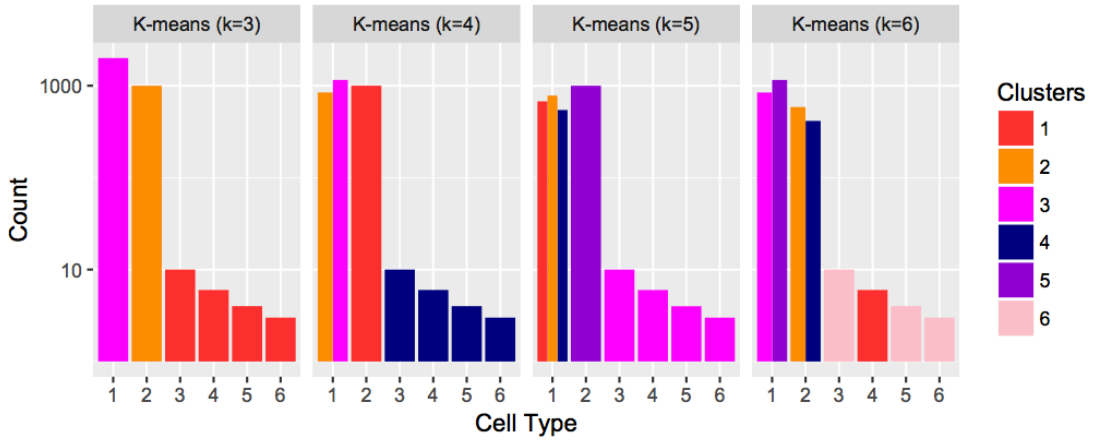


Figure S2.1. A summary of the clustering results of GiniClust2, RaceID2, and other comparable methods on simulated data. Clustering accuracy is measured using several metrics: purity, normalized mutual information (NMI), micro- averaged F-measure, adjusted rand index (ARI), and entropy.

a)



b)

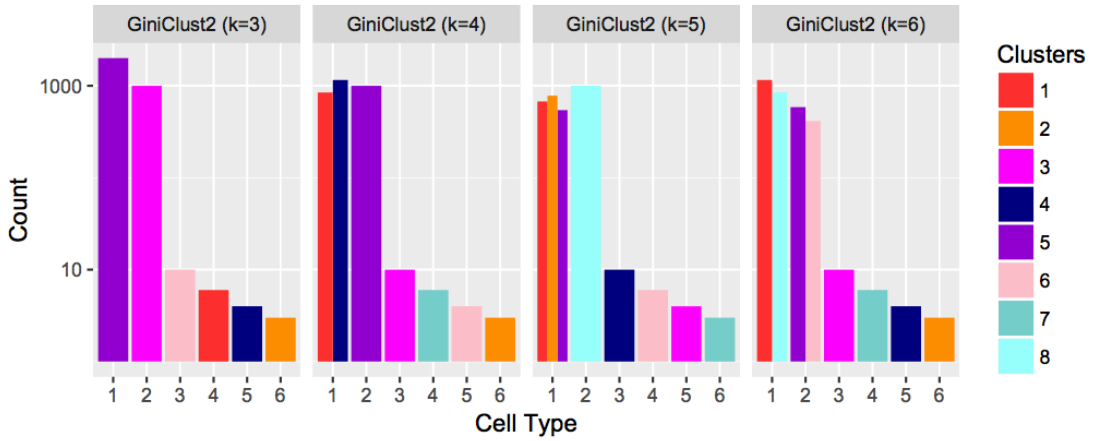


Figure S2.2. The effect of various choices of k on (a) the k-means step and (b) the overall clustering of our GiniClust2 method for the simulated data. Each bar represents the contribution of a cluster to the total number of cells in each reference type.

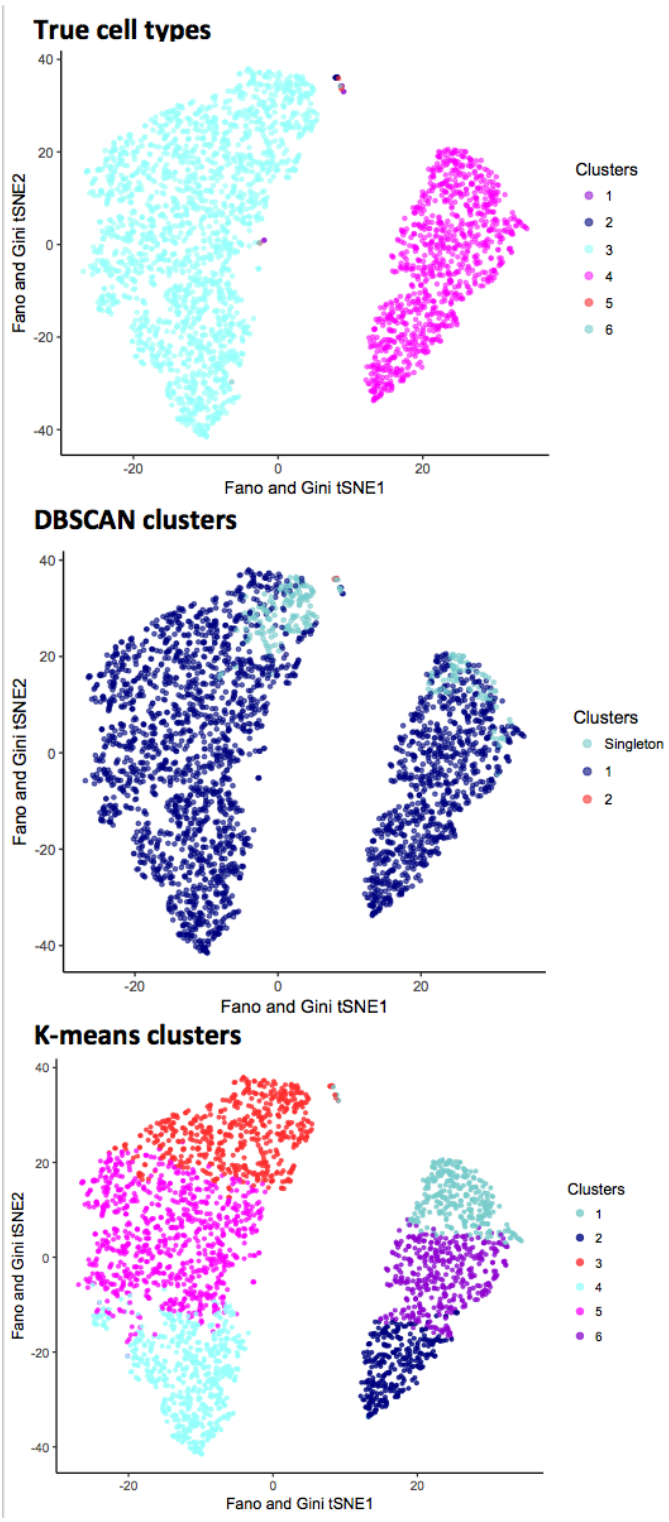


Figure S2.3. Two-dimensional tSNE representations of the simulated data, using a feature space based on a naïve combination of high Gini and Fano genes. Colors represent the true cell types, followed by clustering results using DBSCAN and k-means clustering methods on this naïve feature space, respectively.

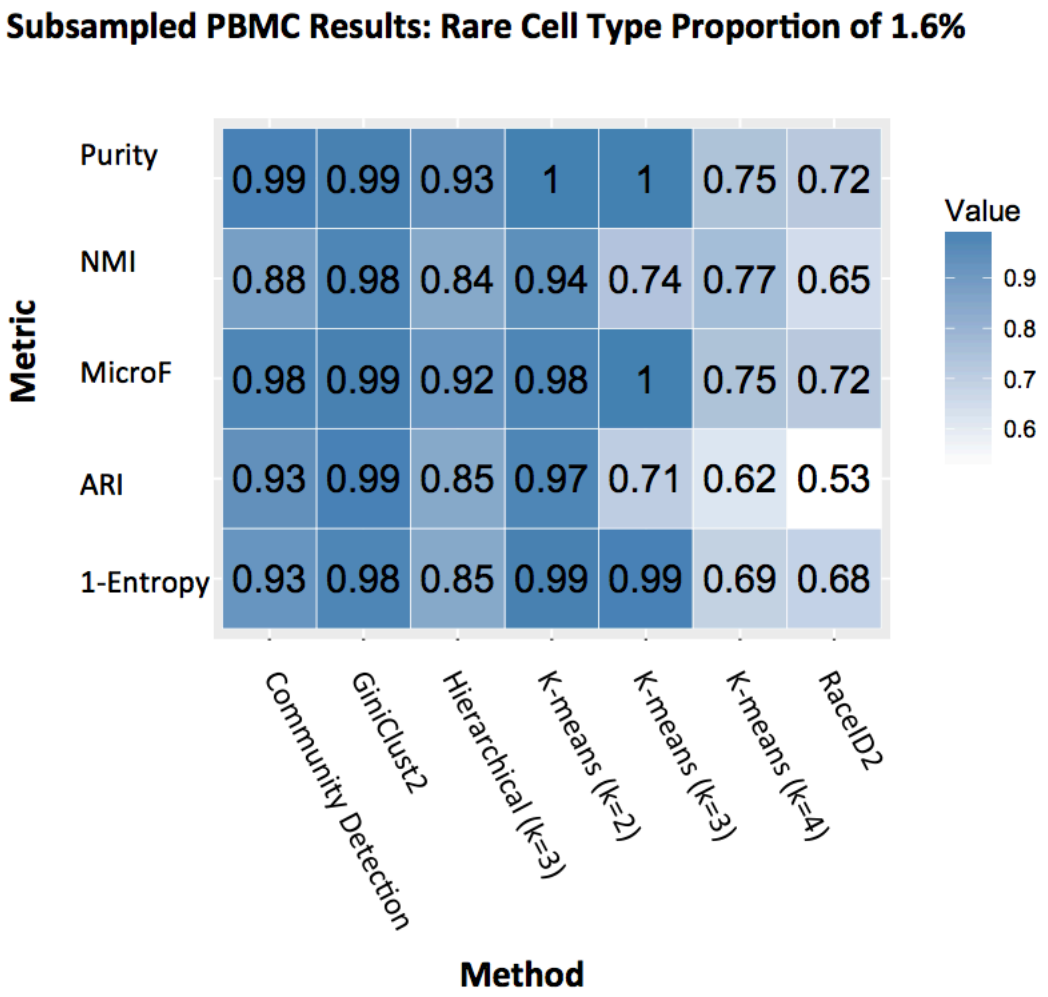


Figure S2.4. Clustering results for a subset of the subsampled PBMC data sets containing a representative rare cell type proportion of 1.6%, for GiniClust2, RaceID2, and other comparable methods. These were measured using purity, normalized mutual information (NMI), micro-averaged F-measure, adjusted rand index (ARI), and entropy.

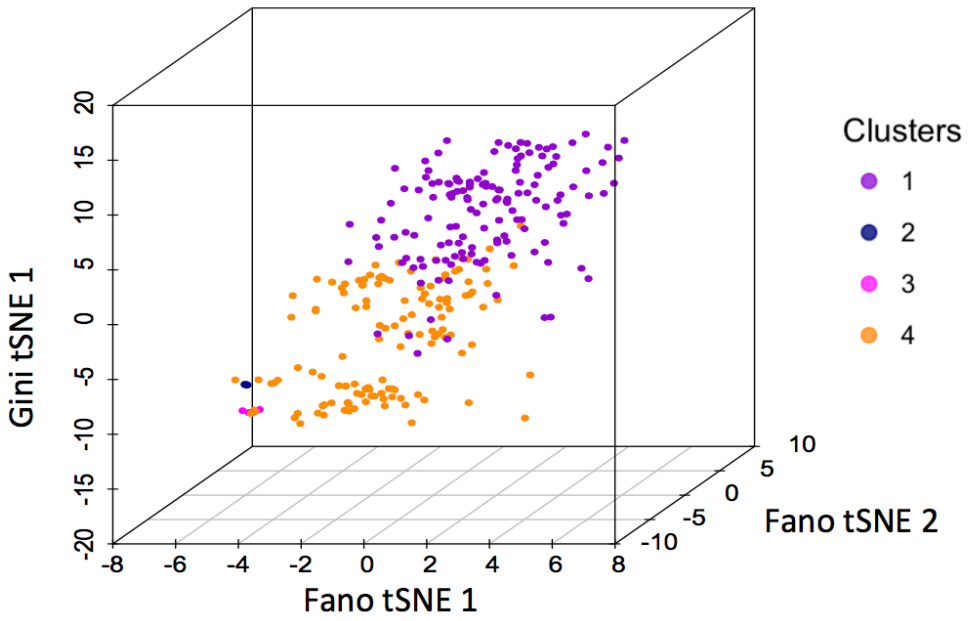


Figure S2.5. A composite tSNE plot representing the GiniClust2 clustering results for the inDrop dataset for day 4 post-LIF mESC differentiation [79].

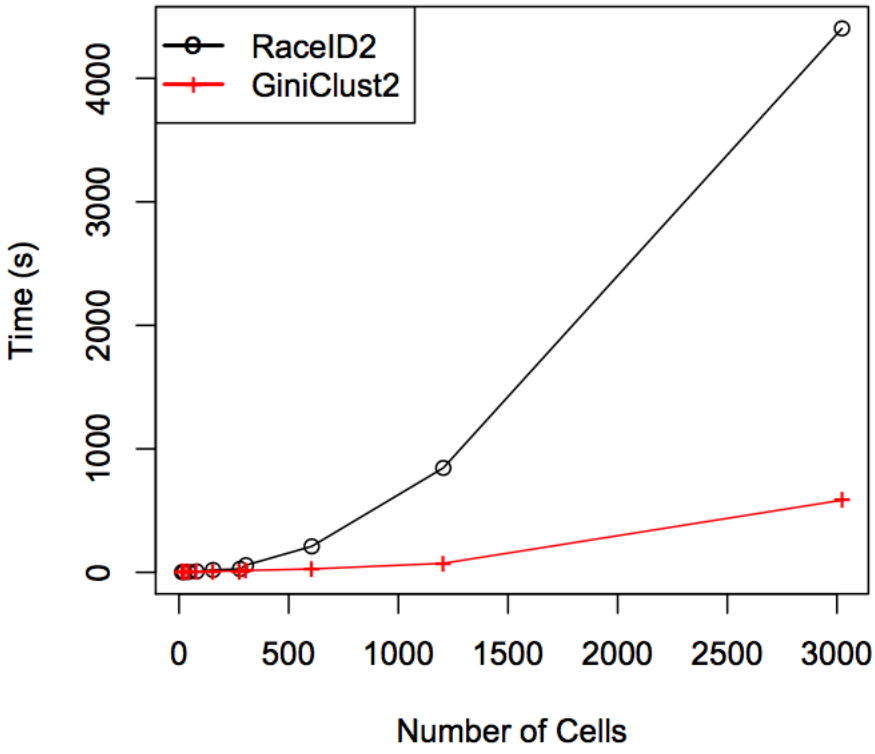


Figure S2.6. A comparison of computational runtimes for GiniClust2 and RaceID2, for data sets ranging from 43 to 3023 cells. The methods were run on a 2.5 GHz Intel Core i7 CPU with 16 GB memory.

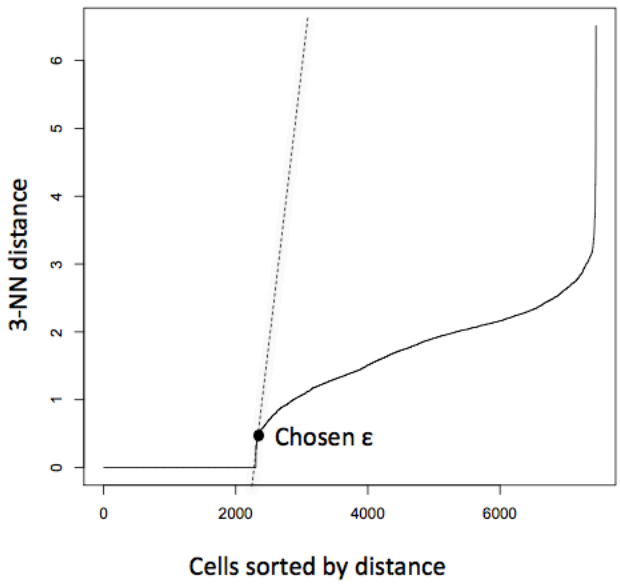


Figure S2.7. An illustration of the eps parameter selection process for DBSCAN, the clustering method used in GiniClust. Eps is chosen as the distance at the inflection point in the k-nearest-neighbors distance plot.

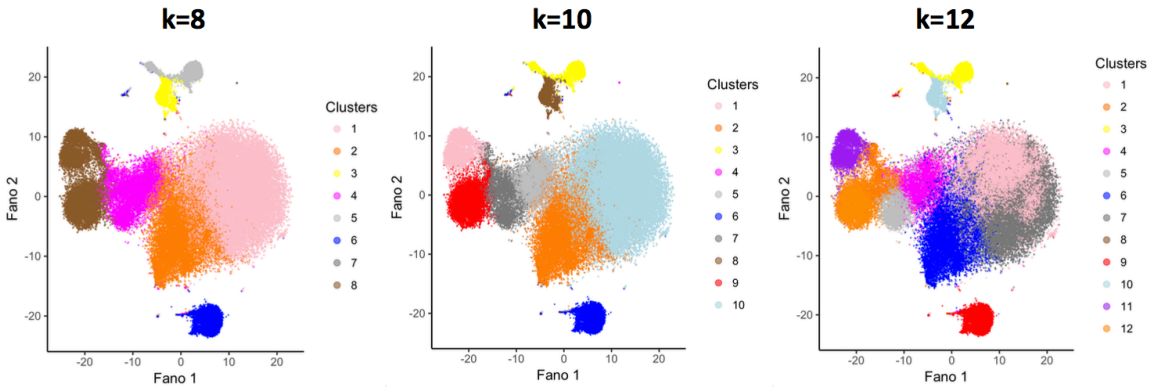


Figure S2.8. The effect of various choices of k on the Fano-factor-based k-means step of GiniClust2 for the full 68k PBMC data. Clustering results for Fano-factor- based kmeans using k=8, k=10, and k=12, respectively, are shown using three two-dimensional tSNE plots colored with each set of corresponding cluster labels.

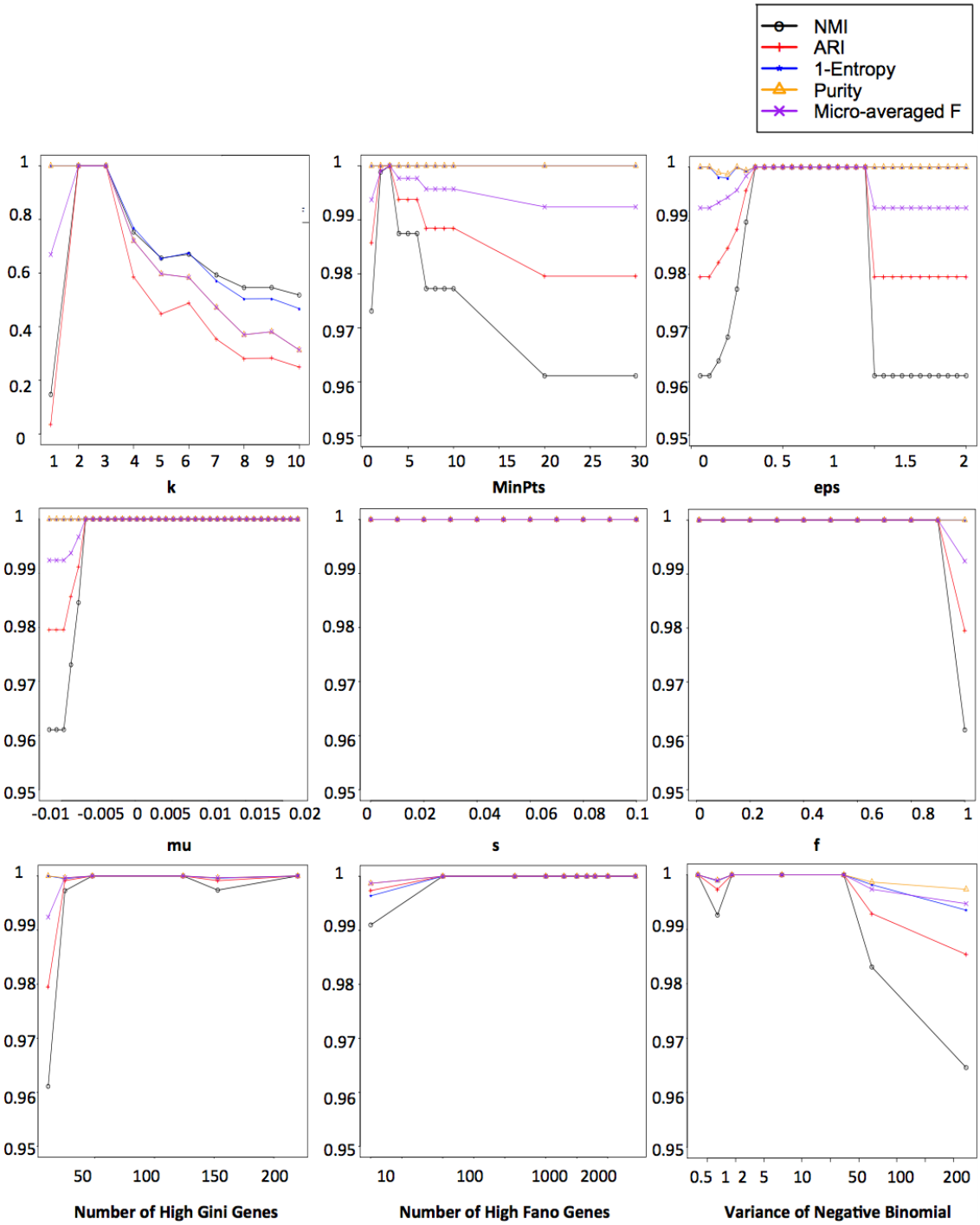


Figure S2.9. A sensitivity analysis for GiniClust2 on simulated data. Eight parameters were independently varied: k-means parameter k, DBSCAN parameters MinPts and eps, weighting scheme parameters μ, s, and f, and Gini and Fano gene thresholds. Additionally, the variance of the simulated data was varied. The effects of each of these changes on the final clustering accuracy of GiniClust2 were measured using several metrics: normalized mutual information (NMI), adjusted rand index (ARI), entropy, purity, and micro-averaged F-measure.

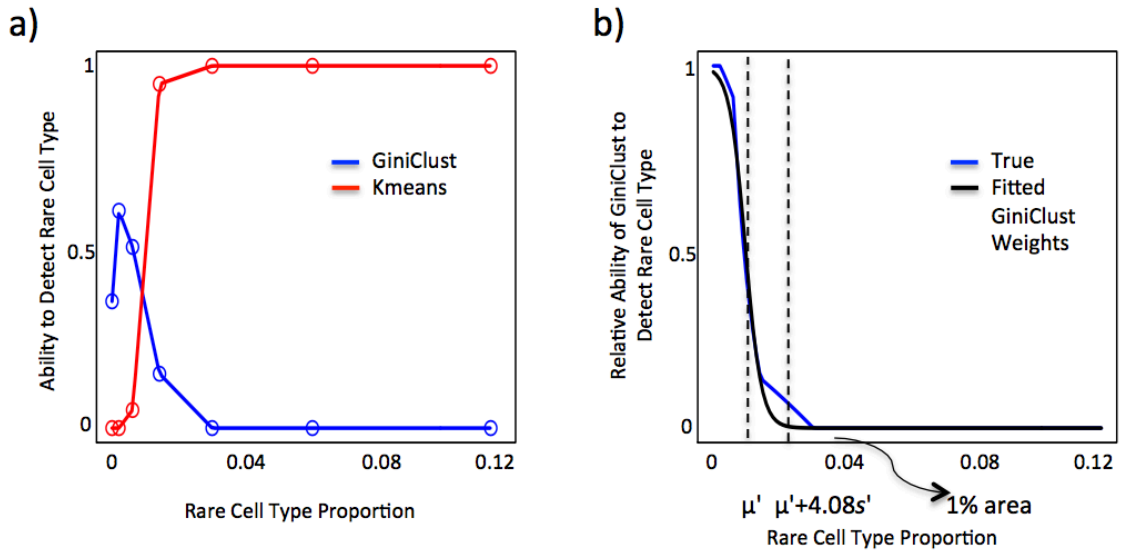


Figure S2.10. An evaluation of the abilities of GiniClust and Fano-factor-based k-means to detect rare cells, performed on the subsampled PBMC data sets. (a) Rare cell type detection abilities of GiniClust and Fano-factor-based k-means over a range of rare cell type proportions. (b) A representation of the ability of GiniClust to detect rare cell types over Fano-factor-based k-means, and its logistic fit. Parameters μ' and s' determine the shape of the curve.

A.3. Supplemental Table

Simulations	Macrophage	NK	B	Rare Cell Type Proportion
1-20.	5	1600	800	.002
21-40.	5	800	400	.004
41-60.	5	400	200	.008
61-80.	5	200	100	.016
81-100.	5	100	50	.032
101-120.	5	50	25	.063
121-140.	5	25	13	.116

Table S2.1. Cell numbers in three different cell types for each of 140 subsampled datasets from 68k PBMCs.

B. Supplemental Materials for Chapter 3

B.1 Supplemental Table

	T cell (0.439)	B cell (0.098)	Macrophage/Monocyte (0.150)	Dendritic cell (.010)	NK cell (.032)	Endothelial cell (.008)	Cancer Associated Fibroblast (0.019)	Ovarian cancer cell (.020)	Melanoma cell (0.191)	Overall
DWLS	.086	.039	.044	.004	.027	.008	.024	.023	.029	.032
QP	.075	.030	.045	.008	.040	.019	.057	.019	.058	.038
ν-SVR	.103	.030	.043	.008	.037	.011	.014	.010	.065	.035

Table S3.1. The accuracy of deconvolution results for the simulated bulk data created from 27 different donor and patient immune and tumor cell single-cell data sets. Estimation accuracy is measured using absolute error, and is calculated for three different deconvolution methods: DWLS, ν-SVR, and QP. Average true proportions for each cell type are listed alongside each cell type name.