Chapter 10

Investigating Inter- and Intrasample Diversity of Single-Cell RNA Sequencing Datasets

Meghan C. Ferrall-Fairbanks and Philipp M. Altrock

Abstract

Tumor heterogeneity can arise from a variety of extrinsic and intrinsic sources and drives unfavorable outcomes. With recent technological advances, single-cell RNA sequencing has become a way for researchers to easily assay tumor heterogeneity at the transcriptomic level with high resolution. However, ongoing research focuses on different ways to analyze this big data and how to compare across multiple different samples. In this chapter, we provide a practical guide to calculate inter- and intrasample diversity metrics from single-cell RNA sequencing datasets. These measures of diversity are adapted from commonly used metrics in statistics and ecology to quantify and compare sample heterogeneity at single-cell resolution.

Key words Tumor heterogeneity, Single-cell RNA sequencing, Diversity index, Tumor progression, Cancer evolution

1 Introduction

Intratumor heterogeneity (ITH) is a major determinant of tumor progression, the evolution of resistance to therapy, and can fuel tumor evolution and the development of metastasis. ITH is present on multiple different levels, ranging from genetic [1] to epigenetic/cell phenotypic [2, 3] and metabolic [4] to microenvironmental heterogeneity [5]. Single-cell DNA and RNA sequencing have made it possible to identify ITH in a way that cannot be captured by bulk sample profiling [6, 7], because they can, in principle, characterize important differences or common features on the level of the individual cell. Estimating cellular heterogeneity by way of diversity and uncertainty about the identity of an individual in the context of others in a sample is thus an important task. One important quantitative method to assess heterogeneity it by calculating the degree of variation between individual entities, which can be achieved using the concept of a diversity index [8]. Here, we present a method to use single-cell RNA sequencing data and clustering algorithms to calculate a general diversity index

Joseph Markowitz (ed.), *Translational Bioinformatics for Therapeutic Development*, Methods in Molecular Biology, vol. 2194, https://doi.org/10.1007/978-1-0716-0849-4_10, © Springer Science+Business Media, LLC, part of Springer Nature 2021

in order to estimate intratumor heterogeneity, and use it as a starting point for clinical correlations [9], or mathematical modeling [10, 11].

ITH is of clinical interest because it serves as a reservoir for therapeutic resistance and is likely a driver of clinical progression with single and combination therapies, when targeted therapies. The clinical implications of ITH have not been explored in all types of cancer, on all scales of heterogeneity. Further, it is unknown whether certain therapeutics could directly decrease ITH and thus serve to mitigate this critical resistance mechanism. The primary objective of this manuscript is to introduce a multiscale approach to measure ITH using single-cell RNA sequencing. This method can be applied downstream of a number of computational and statistical approaches that integrates scRNA-seq data, and will become an important step in the quest to generate foundational evidence that ITH as a relevant clinical factor in those cancer types that have been lacking behind in terms of describing and clinically assessing tumor heterogeneity. Eventually, it would be the goal to describe ITH such that it can be modified by, for example, epigenetic therapeutics that either increase or reduce it to avert rapid resistance evolution.

Single-cell RNA sequencing (scRNA-seq) can be used to estimate cellular diversity, especially in the context of intratumor heterogeneity. Novel scRNA-seq technologies have become a cost-effective method to identify transcriptomic changes at high resolution. Intratumor heterogeneity can be identified for many disease at various stages [12], and have the potential to bring about novel ways to understand tumor evolution [13]. We build our methodology on the fact that single cell transcriptome profiling of leukemias can directly measure intraleukemic heterogeneity (ILH). A scRNA-seq study in chronic myeloid leukemia (CML) has demonstrated that scRNA-seq was capable of segregating patients with discordant responses to targeted tyrosine kinase inhibitor therapy [14], and it was recently shown that scRNA-seq data-based cellular diversity quantification can segregate various other healthy and cancer states [15].

As a summary statistic for ILH we show how to calculate and use a diversity index often applied in ecology [16, 17], using the general nonspatial diversity index [18] called ${}^{q}D$. This approach considers diversity on all possible orders q, and allows to compare states according to specific diversity indices, which emerge as special cases (e.g., ${}^{0}D$, or ${}^{2}D$). The species (clonal) richness of a sample is given by ${}^{0}D$. The Simpson index, that is, the probability that any two cells are identical, emerges from ${}^{2}D$. Most notably, the Shannon index [19]—a measure of uncertainty about the state of the heterogeneous cell population estimated from a subsample of it—can be derived from the limit of q approaching the value of 1. These indices have been used previously to quantify cancer heterogeneity

[10, 11, 20]. This general representation of diversity allows flexibility in the choice of the optimal *q*, potentially tailored for its biological or clinical application.

1.1 Chapter Outline

This chapter describes one established framework for quantifying inter- and intrasample heterogeneity of single-cell RNA sequencing experiments, followed by an example previously described comparing these diversity metrics for acute myeloid leukemia (AML) patients to healthy donors and discussion of interpretation of these diversity metrics. The chapter concludes with notes about how robustness and error of these types of metrics. The outline for this chapter is as follows. Subheading 2 contains the computational materials, including R libraries, and example single-cell RNA sequencing datasets that can be used to gain an intuition of the method. In Subheading 3, we present a single-cell RNA sequencing quality control analysis, methods of data clustering, and the diversity score calculation pipeline. Subheading 4 contains a worked example of calculating a universal diversity metric, applied to an AML dataset of four patient samples. Finally, in Subheading 5 we present notes/discussion from the example R code.

2 Materials

The methods presented in this chapter are one way to calculate diversity metrics applied to FASTQ files generated from single-cell RNA sequencing experiments. The materials required include:

1. FASTQ dataset.
2. Cell Ranger Pipelines.
3. Statistical programming package R.
4. R libraries.
5. UMAP.

2.1 Count Matrix Generation

Single-cell RNA sequencing (scRNA-seq) data is often exported or postprocessed into a FASTQ file. FASTQ files are the most common way scRNA-seq data is stored in publicly available datasets.

2.1.1 Downloading and Installing Tools

1. Download and install the Cell Ranger tar file on a Linux distribution from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation).
2. Install *Seurat* (*version 3.0.0*) package available on CRAN in R on version 3.4 or greater [21, 22]. Additional packages that are useful to speed up computation, especially if running these analyses on a cluster include *future* to access multiple process for parallelizing the Seurat commands and *bigmemory*.

3. Install UMAP (and required dependencies) with python as described on GitHub (https://github.com/lmcinnes/umap/blob/master/README.rst) [23]. UMAP visualization for clustered scRNA-seq data can be performed in R using the Seurat package but requires that UMAP first be installed via python.

2.2 Preparing Data for Diversity Analysis

1. Run *cellranger count* on each sample fastq file. This creates a number of output files, including a folder with the filtered feature-barcode matrix containing a MEX formatted counts matrix. *cellranger count* should be run on each sample's FASTQ file individually.
2. Once FASTQ files have been run through *cellranger count*, open R Studio and initiate libraries using *library(packagename)* for the following libraries: Matrix, dplyr, readr, rdetools, data.table, ggplots2, iterators, Seurat.
3. Import data into R using the *Read10X* and then the *CreateSeuratObject* commands. Use these two commands for each additional sample. To create a multisample dataset, use the *merge* command to merge the individual Seurat Objects.

3 Methods

This method was adapted and expanded from the tutorials for analysis of single-cell RNA sequencing data from the developers of the Seurat package [21, 22] in R available at: https://satijalab.org/seurat/. These tools are not the only way to analyze single-cell RNA sequencing data, but are best practices that we have found useful in quantifying differences in diversity measures across different polyclonal and malignant populations.

3.1 Quality Control and Normalization of Count Matrix

Raw datasets need to be corrected to remove batch effects. This can be done by assessing the distribution of genes captured per cell in the dataset and the individual dataset's distribution of mitochondrial gene.

Typical gene distribution cutoff to determine which cells to include in further analysis involved a lower limit of cells with at least 200 genes detected and an upper limit of genes detected as:

Mean number of genes detected ± 2 × (standard deviation of genes detected)

These gene distribution cutoffs are an attempt to avoid counting doublet cells as distinct single-cells [24].

Typical mitochondrial DNA content upper cutoff is very problem specific. A general rule of thumb is to exclude cells with over 5% mitochondrial content. This cutoff is an attempt to exclude cells that are dying and may not be capturing the biology researchers are interested in exploring [24]. This cutoff needs to be increased for cell- and disease-specific cases, for example, cardiac cells are known to have increased mitochondrial content and so the maximum mitochondrial content cutoff would need to be increased.

These cutoffs to correct for doublet and high mitochondrial content concerns are implemented with by using the *subset* function on the full Seurat object dataset.

Then researchers can normalize and scale the expression data using *NormalizeData* and *ScaleData* functions. One common normalization technique is to normalize the feature expression for each cell by the total expression. Scaling the data allows researchers to removed unwanted sources of variation, including RNA count information and mitochondrial content by shifting the expression level to have a mean around 0 and variance across cells of 1.

3.2 Cluster Detection

3.2.1 Dimension Reduction

Principal component analysis (PCA) is often performed on single-cell RNA sequencing datasets to identify the largest sources of variation in the dataset (i.e., the principal components) as well as that some clustering algorithms require dimension reduced datasets to be used with those algorithms. PCA is implemented on a Seurat object using the *RunPCA* function. Dimension reduction can be performed on the entire scaled dataset or on the subset of variable feature gene set.

3.2.2 Community-Based Detection Methods

Clustering the single cells based on similar expression profiles offers an axis on which diversity can be quantified by across individual samples. One approach used here is a graph-based clustering approach, where the cells are embedded in a graph structures with edges drawn between similar cells [21, 22]. The graph was then partitioned into highly connected communities and the Louvain algorithm is applied to optimize based on modularity. The modularity scores the quality of the optimized clusters. High modularity reflects the presence of community structure in the graph [25]. For our analysis, modularity less than 0.6 were further refined for lack of community structure present in the network analyzed. High confidences in community network structure are present in networks with modularity greater than 0.8. Using Seurat, cluster determination is implemented by *FindNeighbors* and *FindClusters* functions, respectively. In the *FindClusters* function, there is a resolution parameter (defaulted at 0.6) that allows researchers to adjust the granularity of downstream clustering. Increasing the resolution parameter in *FindClusters*, increases the number of distinct clusters identified and should be optimized for large datasets. Visualization of these clusters can be implemented using the *RunUMAP* function that utilizes the principal component dimension reduction to create a 2D visualization of the clustered data, which can be displayed using the *DimPlot* function to plot the clustered dataset either by cluster identity or sample identity (or any other meta data groupings added to the dataset).

3.3 Diversity Score Calculation

3.3.1 Generalized Diversity Index

The generalized diversity index takes the frequency of each sample identity in each of the clusters identified in the dataset and quantifies the diversity score that can be calculated over a range of resolution scales. The mathematical formulation is:

$$^{q}D = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}}$$

where n is the number of clusters identified, p_i is the frequency of each cluster, and q is the resolution or "order of diversity." The most common diversity metrics, Shannon Entropy and Simpson Index are permutations of this generalized index [26]. Shannon entropy [27] is calculated by $q = 1$ with $\log(^{1}D)$ and the Simpson index [28] is calculated by $q = 2$ corresponding to $1/^{2}D$.

This generalized diversity index, ^{q}D, can be calculated from the clustered data set by first counting the number of unique barcodes per cluster, then grouping cells by cluster and then by sample type. From the raw per-cluster-per-type grouping, the cell counts can be converted to frequencies, which can be directly input the equation for ^{q}D, which can be solved over a range of q (one range capturing most dynamics if from 10^{-2} to 10^{2}).

3.3.2 Kolmogorov Smirnov Distance

Another metric that can be used to quantify the differences between samples is using the Kolmogorov Smirnov (KS) distance between two discrete distributions. The KS distances is a nonparametric test, where similar distributions have smaller KS distances. The KS distance can be calculated by taking the probability mass function for a given sample across all the clusters identified by the aggregate dataset, converting that to a cumulative probability distribution. Then the KS distance is calculated using the supremum, or least upper bound (practically, the maximum value of a finite set of numbers)

$$d_{\mathrm{KS}} = \sup\left(\mathrm{abs}\left(c_{1,i} - c_{2,i}\right)\right)$$

where c_1 and c_2 are the cumulative probability function of two different samples. The maximum value of the absolute differences of the cumulative distributions is what is known as the KS distance. Previously, we have shown the KS distances is smaller for like samples (two healthy or two AML) and larger between different samples (for example, healthy versus AML) [15].

4 Example

4.1 Background

In this section, we demonstrate how inter- and intrasample diversity may be used to quantify the bone marrow mononuclear cell (BMMC) heterogeneity of two acute myeloid leukemia (AML) patients compared to BMMCs of two healthy donors previously described in Ferrall-Fairbanks et al. [15] and code publicly available. This draws upon existing, publicly available data from Zheng et al. [6].

4.2 Method

Download the AML dataset (https://github.com/MathOnco/scRNA-seqITH/blob/master/Pipelines/data/AML-Data.zip) and R code (https://github.com/MathOnco/scRNAseqITH/blob/master/Pipelines/MiMB-Diversity-Pipeline.R) from GitHub to follow along with the worked-example of quantifying heterogeneity between two healthy donor and two AML BMMCs (*see* **Notes 1** and **2**).

5 Notes

1. Metric interpretation

   Quantifying a generalized diversity metric enabled us to distinguish between leukemic states based on the high-dimensional single-cell patient samples. From an ecological perspective, diversity can be measured across a number of different spatial scales and by solving for a continuum of diversity indices we can examine a sample's diversity across these scales. For low q, generalized diversity index (GDI) represents the clonal richness, assuming that clusters of similar gene expression represent a 'clone' and as q approaches 0, ^{0}D becomes the number of clusters identified. On the other side of the spatial scale, for high q, the contribution of the major clone(s) is weighted more, attempting to quantify species evenness. In a clinical setting, this would likely represent the dominant one or two phenotypes of tumor, that are easily detected by clinicians and may drive therapy selection. Intermediate values of q correspond to classical measures of sample diversity, such as Shannon index (H, $q = 1$) [27] that has been used in oncology to analyze tumor evolution and single-cell tumor imaging data [29]. We have seen that diversity scores can be very similar around $q = 1$ and as a result the Shannon index can therefore be a problematic diversity indicator. However, GDI at a range of q, point to differences in the number of major drivers of tumor evolution, possibly prior to detection/sampling.

2. Robustness

   One limitation of scRNA-seq is that missing data does not necessarily reflect that those transcripts are not expressed in the sample. In scRNAseq, any given cell only captures at most about 10% of the whole transcriptome, but in aggregation with other single cells, the entire genome is covered. As a result, in order to test the robustness this diversity metric, one can down-sampled their dataset and cluster to determine how the diversity index may change with the subset dataset. In Ferrall-Fairbanks et al., we down-sampled the dataset by randomly removing as much as 50% of the cells from each of the healthy and AML samples (this was repeated 1000 times) and found that the clustering did not change more than 1–2 clusters in either direction. With these new clustering, if the AML diversity curve was shifted down two units and healthy diversity curve was shifted up two units, you still would have separation between the conditions, suggesting that this metric is pretty robust. Furthermore, during this down-sampling exercise, we found that if we down-sampled to roughly 1500 cells, we would often capture the same number of clusters as the full dataset, suggesting that at least for this AML versus healthy BMMC comparisons, 1500 cells is a lower limit of cells needed to capture these diversity dynamics.

6 Summary

Translational bioinformatics is an emerging field at the intersection of molecular bioinformatics, statistics, and clinical applications. In the age of ever refined molecular insights in large data sets, it is important to develop pipelines that allow effective integration and biological interpretation with a focus on cancer evolution. Importantly, these pipelines have to scale when applied to large data sets, as well as deliver biological interpretation. With the pipeline described here, we have provided the theoretical and computational basis for tools that allow quantification and comparing diversity sample diversity at single cell resolution. The generalized diversity score applied to single cell sequencing samples allows comparing heterogeneity across different normal and malignant samples, and is based on clustering of single cell data. Hereby, particular choices of imputation and clustering procedures [30, 31] and batch correction [32, 33]—likely to undergo further development in the near future [34]—can be integrated into this concept—our method does not rely on a particular clustering algorithm.

Further therapeutic target development can be gleaned from the diversity analysis by exploring differentially expressed gene signatures between normal and malignant patients, or between different patients of different stages. How the measure of diversity, and the associated gene signatures, change over time in a given patient may also offer important insights into therapeutically relevant targets, which needs to be explored further.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW (2013) Cancer genome landscapes. Science 339 (6127):1546–1558
2. Marusyk A, Almendro V, Polyak K (2012) Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer 12(5):323–334
3. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. Nature 501(7467):328–337
4. Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson AR (2015) Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. Cancer Res 75 (8):1567–1579
5. Tabassum DP, Polyak K (2015) Tumorigenesis: it takes a village. Nat Rev Cancer 15 (8):473–483
6. Zheng GX, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049
7. Paguirigan AL, Smith J, Meshinchi S, Carroll M, Maley C, Radich JP (2015) Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. Sci Transl Med 7(281):281re282
8. Lou J (2006) Entropy and diversity. Oikos 113 (2):363–375
9. Dagogo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15(2):81–94
10. Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514(7520):54–58
11. Altrock PM, Liu LL, Michor F (2015) The mathematics of cancer: integrating quantitative models. Nat Rev Cancer 15(12):730–745
12. Park Y, Lim S, Nam JW, Kim S (2016) Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci Rep 6:37767
13. Hu Z, Sun R, Curtis C (2017) A population genetics perspective on the determinants of intra-tumor heterogeneity. Biochim Biophys Acta 1867(2):109–126
14. Giustacchini A, Thongjuea S, Barkas N et al (2017) Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat Med 23 (6):692–702
15. Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM (2019) Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity. JCO Clin Cancer Informatics 3:1–10
16. MacArthur RH (1965) Patterns of species diversity. Biol Rev 40:510–533
17. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432
18. Tuomisto H (2010) A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia 164(4):853–860
19. Shannon CE (1997) The mathematical theory of communication. 1963. MD Comput 14 (4):306–317
20. Almendro V, Cheng YK, Randles A et al (2014) Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. Cell Rep 6(3):514–527
21. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36 (5):411–420
22. Stuart T, Butler A, Hoffman P et al (2019) Comprehensive integration of single-cell data. Cell 177(7):1888–1902.e1821
23. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. J Open Source Softw 3 (29)
24. AlJanahi AA, Danielsen M, Dunbar CE (2018) An introduction to the analysis of single-cell RNA-sequencing data. Mol Ther Methods Clin Dev 10:189–196
25. Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103(23):8577–8582
26. Morris EK, Caruso T, Buscot F et al (2014) Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecol Evol 4 (18):3514–3524

27. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
28. Simpson EH (1949) Measurement of diversity. Nature 163:688
29. Almendro V, Kim HJ, Cheng YK et al (2014) Genetic and phenotypic diversity in breast tumor metastases. Cancer Res 74 (5):1338–1348
30. Qi R, Ma A, Ma Q, Zou Q (2019) Clustering and classification methods for single-cell RNA-sequencing data. Brief Bioinform. https://doi.org/10.1093/bib/bbz062
31. Petegrosso R, Li Z, Kuang R (2019) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief Bioinform. https://doi.org/10.1093/bib/bbz063
32. Hicks SC, Townes FW, Teng M, Irizarry RA (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19(4):562–578
33. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ (2019) A test metric for assessing single-cell RNA-seq batch correction. Nat Methods 16(1):43–49
34. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20 (5):273–282