Statistical methods for pattern discovery with cell-aggregated single-cell RNA-seq data

Hector Roux de Bézieux
Group in Biostatistics, and
Center for Computational Biology
University of California, Berkeley
Berkeley, CA 94704
hector.rouxdebezieux@berkeley.edu

Nima Hejazi
Group in Biostatistics, and
Center for Computational Biology
University of California, Berkeley
Berkeley, CA 94704
nhejazi@berkeley.edu

1 Introduction

Numerous biotechnological innovations for the quantification of mRNA abundance have been developed and used in addressing the need to cope with the deluge of biological sequencing data. Next-generation sequencing (NGS) approaches, such as bulk RNA sequencing (RNA-seq), have increasingly displaced microarrays in the last decade, largely owing to the fact that NGS approaches have the significant advantage of enabling whole exome sequencing. In comparison, microarrays were limited to investigating only a select number of pre-specified probes, leading to possible selection bias in the analysis of sequencing data.

More recently, a slew of new technologies — broadly grouped under the umbrella of *single-cell* RNA sequencing (scRNA-seq) — have made possible the quantification of transcription at the level of individual cells. By comparison to the tissue-level analyses that bulk RNA-seq facilitated, the wealth of biological insights made accessible by scRNA-seq (e.g., cell lineage inference) is staggering. Such advances may be used to distinguish between nuances in cell types, to probe developmental phenomena, or even to elucidate variability in cell populations in responses to a given stimulus (e.g., drug intervention). Despite the new types of questions that may be answered by scRNA-seq, bulk RNA-seq has not been made obsolete — that is, it is still useful when the scientific question of interest concerns biological phenomena at the tissue level; for example, in studies of how the pancreas may react to prospective new anti-diabetes drugs.

While both bulk RNA-seq and scRNA-seq have their uses, the emergence of the newer biotechnologies has artificially dichotomized the available levels of scientific inquiry — in fact, many biologists feel limited to analyses at the tissue or single-cell levels. Yet, there are numerous scientific questions that exist outside of these two levels of analysis, where the dichotomy between tissue-level and single-cell investigations is simply too restrictive. When the objective of an investigation concerns only differences in how general *types* of cells (in a given tissue) respond to changes, the single-cell level analysis provides far too much information, and therefore represents a waste of resources, while the tissue-level analysis is simply insufficient for probing the underlying biology at the requisite depth.

In this paper, we propose to investigate examples where an intermediary level analysis — at the level *between* bulk RNA-seq and scRNA-seq — would sufficiently answer the biological question of interest. Such an investigation would be largely statistical (and, perhaps, computational), framed in terms of what is to be gained and lost in combining single-cell level counts to represent groups of cells. Using two example, mouse Embryogenesis development, and spatial differentiation in a zebrafish embryo, we demonstrate that an intermediate scale recover most of the biological meaningful results of the single-cell RNA-seq analysis.

Preprint. Work in progress.

2 Background

2.1 Embryogenesis

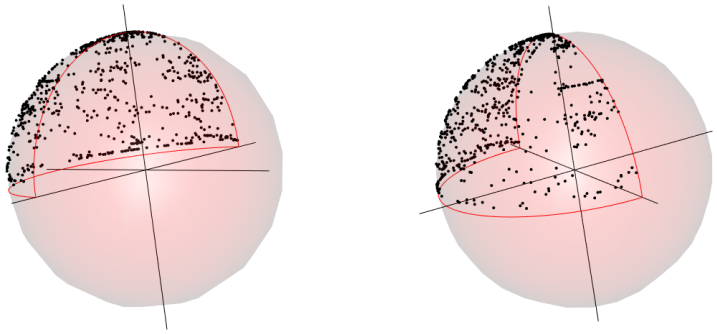
Recent scRNA-seq analysis has demonstrated that, during the first few stages of development, the first 64 divisions, within-embryo variability is much smaller than between-embryo variability [5]. Therefore, a standard scRNA-seq analysis would yield nearly identical results at the single-cell level or at the embryo level. Thus, ascertaining whether standard statistical methods provide concordant results at the single-cell level and at the embryo level would provide first-pass confirmation of this hypothesis. A few scRNA-seq data sets are publicly available for this purpose, providing information on cell origin (i.e., which cells came from the same embryo); we plan to focus on two in particular [5, 8].

2.2 Tumor heterogeneity

A tumor is far from a uniform environment. Cancer cells neighbor normal tissues, various cell types can be found. Even among cancer cells, a wide diversity exist. While genetic diversity has been observed and studied for a longer time [14], epigenetic differences [6], as well as differences of environment (oxygen influx for example) [12] have recently been observed and seem to play an important role. Characterizing this heterogeneity is a key part of cancer research, since heterogeneity usually drives tumor resistance to treatment[7].

Gene-expression diversity in tumor is therefore one crucial aspect of this diversity. Our hypothesis is that, to fully characterize this heterogeneity, the fine-scope of scRNA-se is too narrow. Indeed, spatially nearby cells are very likely to have similar gene-expression. Sequencing more neighboring cells is not likely to analysis results. On the other hand, merging those cells together before sequencing will reduce dropout of count and in fact, produce more stable results.

To test this assumption, we select a zebrafish embryo dataset from [23]. In their original work, the authors sequenced 851 cells at the surface of the spherical embryo and used known spatial gene markers to recover the coordinates of each cell. We use those coordinates as a given to merge cells and measure of much the result change when compared to single-cell level analysis. An overview of the cells' location on the zebrafish embryo can be found in fig. 5. *** HRB: Should be one or two figures? Or maybe one in appendix as well



(a) (b)

Figure 1: View of the cells (black points) collected on the zebrafish embryo (red sphere). The delimitation of the collections area is with the red lines.

3 Methodology

3.1 Merging the single cells

We want to simulate sequencing data from groups of cells. In order to do so, we numerically - and therefore artificially - collapse the read count data from single-cells into broader groups. To consolidate data to such an intermediary level, read counts from individual scRNA-seq cells could be combined, based on an arbitrary summarization, into those groups. Standard statistical analysis of RNA-Seq data (see 3.3) can then be performed on those centers, considered as bulk RNA-Seq samples.

To better simulate the sequencing mechanism, summing the reads is the most sensible. Indeed, it will give the same number of reads overall, to fewer samples. We therefore have a similar sequencing "budget" to reproduce the results. The intermediate step has fewer samples but deeper sequencing, compared to the single-cell. However, normalization is an essential first step of RNA-Seq data analysis. To mimic this normalization step as well, we instead collapse the count as the mean read count for any given gene.

3.2 Random merging of the cells into larger groups based on spatial location

As detailed in the background section 2, embryo-derived single cell have a natural level at which to be merged: the embryo from which they came from. Indeed ,those embryos are easily sequenced and collected. Breaking them to a single-cell level only increases the cost of the experiment. So, the "centers" of our new embryo data are the individual embryo themselves.

However, no such obvious level exists for the zebrafish dataset. Hence, we imagine a sampling mechanism that tries to reproduce a real-life setting: cells that are physically close are much more likely to be sampled together. The repartition of the n single cells into k groups ($k < n$) was done in the following manner:

1. Sample k centers on the sphere randomly.
2. Assign each cell to its nearest center, based on orthodromic distance, the distance measured along the surface of the sphere.

Then, each center's count data was just the mean of all the count of the cells that got assigned to it. To assess how the randomness in sampling would affect the result, we repeated the value 15 times for each value of k.

3.3 Single-cell RNA-Seq tools

We plan to evaluate a subset of several standard single-cell RNA-seq (scRNA-seq) methodologies, including clustering, differential expression, and, time permitting, lineage inference.

3.3.1 Clustering

We use the clusterExperiment package[18] and framework. We also make use of the Hierarchical Ordered Partitioning And Collapsing Hybrid (HOPACH), a clustering method that combines the strengths of both hierarchical and partitioning approaches, and its associated implementation in the hopach package [16, 27]. Both software implementations are available through the Bioconductor project for the R language and environment for statistical computing [9, 11, 19]. The clusterExperiemnt package is built for finding small stable clusters of cells based on gene expression.

The clustering algorithm of clusterExperiment works in three steps:

1. Apply a number of clustering algorithms with various (sensible) tuning parameter values. Those algorithms are applied to random subsample of the cells (or genes in our case).
2. Build a co-clustering matrix C where $C_{i,j} \in [0, 1]$ denotes the proportion of times cells i and j clustered together, among all the times those cells were both sampled for the subset.

3. Based on that co-clustering matrix, find stable clusters, i.e., groups of cells that cluster together at least $p\%$ of the time (p is user-defined).

4. Merge clusters that are too similar.

Because we impose a minimum size for a cluster in steps 1 & 3, some cells will be left un-clustered. For more information, see [18].

The clustering algorithm of HOPACH works in the following general manner:

1. The algorithm operates on a matrix of pairwise distances between the observations to be clustered (in our case, these are a set of selected genes).

2. A hierarchical tree of nested clusters is built by recursively partitioning the observations.

3. At each level of the tree, clusters are split into two or more clusters using a partitioning clustering procedure such as partitioning around medoids (PAM), i.e., splits are not restricted to be binary.

4. At each level of the tree, clusters are ordered and possibly collapsed.

5. To assess confidence in cluster assignment, the bootstrap may be used.

To examine how clustering may be perturbed between standard scRNA-seq data and combined data, we present the results of clustering a distance-matrix of the set of genes found to be differentially expressed in the standard scRNA-seq data. Distance matrices comparing the two approaches are presented in section 4.

3.3.2 Differential expression

We use `edgeR`, a procedure that makes use of the negative binomial model, and `limma`, a method that assumes a Gaussian model but transforms count data to achieve good, empirically validated differential expression results. In the `edgeR` model, the mean counts μ_{ij} for gene i sample j is a function of the total number of reads for sample i and the relative abundance of gene j for sample i as they relate to the experiment group within which sample i is situated. This relative abundance is the parameter of interest in our differential expression analysis. To estimate the degree of overdispersion, an empirical Bayes procedure is used [21].

The model of `limma` does not differ significantly from a standard linear modeling procedure, though there are two key alterations that make `limma` suitable for differential expression analysis of RNA-seq data. In particular, [13] proposed the so-called `voom` transformation, which is used in assessing the mean-variance relationship of count data via a standard (Gaussian) linear model — empirically, this procedure has been shown to work quite well [20]. The original chief innovation of the `limma` procedure was an shrinkage estimator of the standard error, derived though a hierarchical Bayesian model — in particular, the proposed variance estimator is useful in creating a *moderated* t-statistic, which has enhanced performance in small samples owing to the fact that it borrows across samples to create a pooled variance estimate:

$$\tilde{t}_b = \frac{\hat{\beta}_{1,b}}{\tilde{\sigma}_n^b} \quad \text{where} \quad \tilde{\sigma}_{b,n}^2 = \frac{d_0\sigma_0^2 + d_b(\sigma_n^b)^2}{d_0 + d_b},$$

where d_b is the degrees of freedom for the b^{th} gene/biomarker, d_0 is the degrees of freedom for the remaining $(B-1)$ genes/biomarkers, σ_n^b is the standard deviation for the b^{th} biomarker and σ_0 is the common standard deviation across all biomarkers towards which shrinkage is performed [24]. Since there are B hypothesis tests performed (of null hypotheses $H_{0,b} : \beta_{1,b} = 0$, for genes/biomarkers $b = 1, \ldots, B$), we make use of the standard Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR) [4]. In discussions of differential expression results, genes are labeled as significant if their adjusted p-values were below the cutoff of 0.05, ensuring that, in expectation, the number of false discoveries is limited to 5%.

4 Results

4.1 Zebrafish embryo

We collapsed the single-cell data into groups, from 30 to 759, compared with originally 851 single cells, following the process described in 3.2. An example of how cells are assigned, with just seven groups for better visualization, can be found in Fig. 2.

Figure 2: Random assignment of the cells. Cells are colored by their assignment, following the process in 3.2. The zebrafish embryo is in light red.

After collapsing the data, we identified stable clusters using clusterExperiment[18], described in 3.3.1, and then picked the top 1000 DE genes, according to edgeR [21]. We only look at the top 1000 since usually, biologists just focus on the top genes, regardless of the associated p-values. We then compared the concordance of this list of 1000 genes to the list found when following the same framework with the single-cell dataset and obtained a concordance score. Since the collapsing algorithm is random, we collapse 15 times for each number of groups. Results can be seen in Fig. 3.

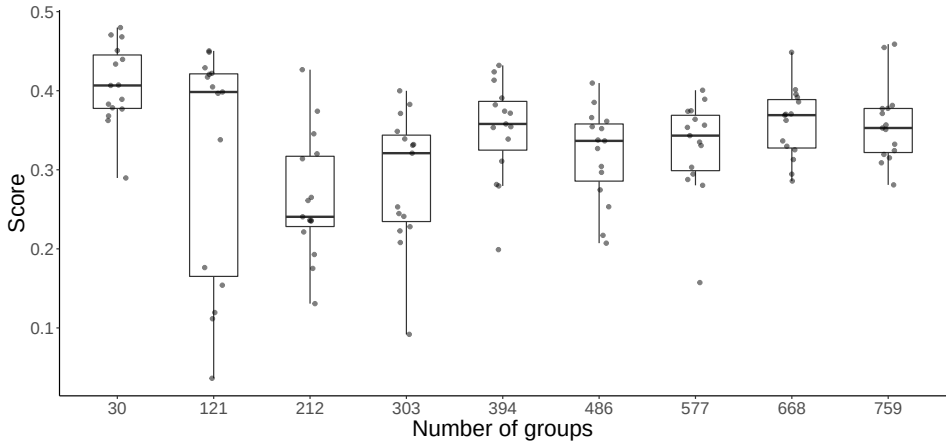
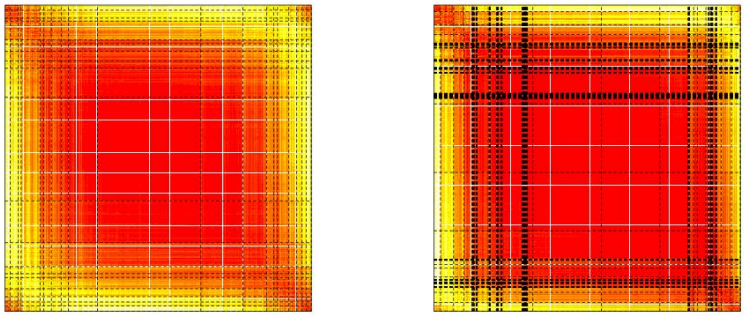


Figure 3: Score of the collapsed data compared to the single-cell, for 15 collapse per group number

As we can see, the score for the lowest value of k is of the same order of magnitude as the score for the highest value, where we are near single-cell levels. For intermediate values of k, we have a strong heterogeneity. With only 30 groups collected, we can achieve the same results as with 760. Our initial hypothesis is correct.

4.2 HOPACH clustering concordance

To assess the impact of cell-level aggregation on clustering, we employed the HOPACH clustering algorithm, briefly detailed in 3.3.1. To assess concordance between cluster assignments, HOPACH clustering was performed on the set of genes found to be differentially expressed in the non-aggregated scRNA-seq data, using both the non-aggregated counts and their aggregated counterparts. The distance matrix between gene-level counts was assessed, after imposing the cluster-based ordering of HOPACH, for both the non-aggregated and aggregated data.



(a) HOPACH with non-aggregated data (b) HOPACH with aggregated data

Figure 4: Distance matrices of the non-aggregated and aggregated scRNA-seq data with HOPACH ordering based on the relevant matrix of gene-level counts for the Biase et al. data.

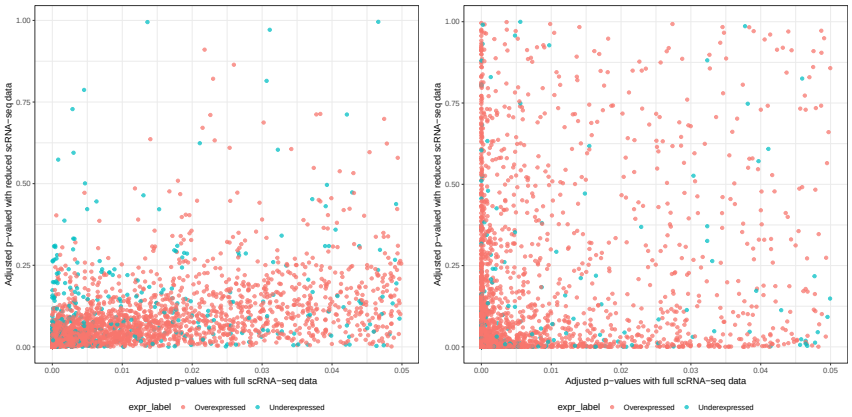
With HOPACH, clusters with genes of similar expression are represented as blocks on the diagonal of the matrix. Indeed, we notice some discrepancies in the ordering and distance matrices generated by the HOPACH algorithm — while this does not readily allow us to draw any scientific conclusions, it points to an area of further investigation. In general, we would expect the aggregated data to behave more like standard (bulk) RNA-seq data, in that it will be less prone to the effects of dropout sparsity and the like.

4.3 Differential expression concordance

To assess concordance in differential expression, we used the approach based on moderated statistics as implemented in the `limma` `R` package [24]. Separately, for both the data sets of Biase et al. and Deng et al. , both the non-aggregated and aggregated scRNA-seq data was analyzed in the following set of steps:

1. Perform quanntile normalization to remove the potential effects of unwanted variation.
2. Perform the "`voom`" transformation, which alters count data into a representation based on the mean-variance trend of the original counts, and estimate sample-level weights so as to downweight possibly uninformative (e.g., contaminated) samples.
3. Fit a linear model of the form $Y = X\beta$, estimating a regression coeffcient β_1 for the cell state (e.g., 2 cells, 4 cells, etc.)
4. Compute the moderated t-statistic (or F-statistic) using a shrinkage estimator of the standard error that pools across samples.
5. Correct for multiple testing using the Benjamini-Hochberg procedure to control the False Discovery Rate at 0.05 [4].

For the set of genes identified as differentially expressed in the non-aggregated scRNA-seq data (with adjusted p-values below 0.05), we examine the distribution of adjusted p-values in the aggregated data below



(a) Biase et al.'s data (b) Deng et al.'s data

Figure 5: Comparison of the distribution of adjusted p-values for genes differentially expressed in non-aggregated data.

We note that there generally appears to be good concordance between the two approaches, as most of the points appear in the bottom left-hand corner (low adjusted p-values under both analyses). The data of Deng et al. appears to show much more variability in the aggregated adjusted p-values for potentially highly statistically significant genes — at first glance, it appears that cell-level aggregation may be drowning out some potentially significant findings, though this may equivalently be interpreted as the aggregration procedure leading to more conservative hypothesis tests. As with the clustering procedure, this requires further investigation before any conclusions can definitively be drawn. We note here that this interpretation of cell-level aggregation as a possibly more conservative testing procedure may be advantageous in assessing the quality of scRNA-seq findings in future studies.

5 Future directions

The present article is merely an initial investigation into how pattern discovery in single-cell RNA-seq data may be performed when information from groupings of cells are collapsed — including in both spatially and developmentally related groups of cells. Much remains to be done. In particular,

we initially proposed the comparative assessment of lineage inference techniques between scRNA-seq and collapsed data as well as the use of nonparametric methods for biomarker discovery. A general manner in which lineage inference may be compared is given in the sequel while details on nonparametric methods is deferred to a later draft.

5.1 Lineage inference

A variety of statistical methods have been developed to probe developmental processes and estimate the location or pseudotime of each cell along lineages based on single-cell RNA-Seq expression measures.

Lineage inference techniques usually follow the following framework:

1. Project the cells in a lower dimensional space.
2. In that lower dimensional space, find curves in a data driven way. Depending on the methods, the number of curves or the endpoints may need to be specified.
3. Each curve now represent a lineage. Each cell can be projected on each of those lineages. The distance between the projection and the beginning of the curve is called pseudotime.

Several observations can be made in this framework. Since the space in which the curves exists is a low-dimensional space of gene expression, the pseudotime measures the amount of genomic variation since the start. As such, the pseudotime is more of a distance than a time, and another word used in the literature instead of lineages is trajectories.

For more information and an overview of these techniques, see [22, 25]. Based on a recent review[22], we selected the Slingshot algorithm [25].

Slingshot requires as input a low-dimension representation of the cells, as well as cluster labels for each cell. It then construct a minimal spanning tree between cluster centers. Once this tree is constructed, it fit principal curves simultaneously on all lineages. This, in essence, smooths the tree in the reduced space. If end clusters or starting clusters are given, it is then possible to determine the location of the starting point of the trajectories. Pseudotime of a cell along a trajectory is then computed as the distance from this starting point to the orthogonal projection of the cell onto the associated curve.

6 Discussions

For the zebrafish embryo example, we can however see that, even at the highest value of k, we only achieve a score of around 40%, which is quite low in term of reproducibility. One possible reason is that the tuning of the hyper-parameters of the clustering algorithms was done on the single-cell data and may not be the most appropriate at every step. Poor clustering would have a strong impact on downstream DE analysis and could therefore explain the gap. Using clustering algorithm that require less user-tuned hyper-parameters would help improve the pipeline, since manual tuning of the hyper-parameters is not possible here (we have 150 simulated datasets). Another hypothesis is the assignation mechanism. For higher number of groups, the number of cells assigned to each group is highly variable, which does not reflect a biologically meaningful collection process. A more realistic mechanism would limit this variation. Likewise, the merging through a simple summation of count data is likely not representative of real-world situation. Indeed, sequencing amplification bias are numerous. Using genes GC content to simulate that bias could improve the realism of the simulation. Finally, any conclusion drawn on this type of data would require extensive validation on real tumor dataset. Emergence of new tools that make use of joint FISH and RNA sequencing to assess spatial location of single-cells will help to generate more realistic datasets that will serve as more reliant benchmarks. All code for our analyses is available on GitHub at `https://github.com/HectorRDB/CS924_Final_Project`.

Acknowledgments

We thank Profs. Jennifer Listgarten, Sandrine Dudoit and Mark van der Laan for insightful discussions throughout the development of the present work.

References

[1] K. Achim, J.-B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt, and J. C. Marioni, "High-throughput spatial mapping of single-cell rna-seq data to tissue of origin," *Nature Biotechnology*, vol. 33, pp. 503 EP –, 04 2015.

[2] A. Alpert, L. S. Moore, T. Dubovik, and S. S. Shen-Orr, "Alignment of single-cell trajectories to compare cellular expression dynamics," *Nature Methods*, 2018.

[3] O. Bembom, M. L. Petersen, S.-Y. Rhee, W. J. Fessel, S. E. Sinisi, R. W. Shafer, and M. J. van der Laan, "Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection," *Statistics in medicine*, vol. 28, no. 1, pp. 152–172, 2009.

[4] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[5] F. H. Biase, X. Cao, and S. Zhong, "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing," *Genome Research*, 2014.

[6] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution," *Nature*, vol. 501, pp. 338 EP –, 09 2013. [Online]. Available: https://doi.org/10.1038/nature12625

[7] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature Reviews Clinical Oncology*, vol. 15, pp. 81 EP –, 11 2017. [Online]. Available: https://doi.org/10.1038/nrclinonc.2017.166

[8] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, 2014.

[9] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[10] S. Gruber and M. J. van der Laan, "An application of collaborative targeted maximum likelihood estimation in causal inference and genomics," *The International Journal of Biostatistics*, vol. 6, no. 1, 2010.

[11] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke *et al.*, "Orchestrating high-throughput genomic analysis with bioconductor," *Nature methods*, vol. 12, no. 2, pp. 115–121, 2015.

[12] M. R. Junttila and F. J. de Sauvage, "Influence of tumour micro-environment heterogeneity on therapeutic response," *Nature*, vol. 501, pp. 346 EP –, 09 2013. [Online]. Available: https://doi.org/10.1038/nature12626

[13] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: Precision weights unlock linear model analysis tools for rna-seq read counts," *Genome biology*, vol. 15, no. 2, p. R29, 2014.

[14] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?" *Nature Reviews Cancer*, vol. 12, pp. 323 EP –, 04 2012. [Online]. Available: https://doi.org/10.1038/nrc3261

[15] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, vol. 14, no. 4, pp. 417–419, 4 2017. [Online]. Available: http://www.nature.com/articles/nmeth.4197

[16] K. S. Pollard and M. J. Van Der Laan, "Cluster analysis of genomic data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005, pp. 209–228.

[17] K. S. Pollard and M. J. van der Laan, "Supervised distance matrices," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.

[18] E. Purdom and D. Risso, "clusterExperiment: Compare Clusterings for Single-Cell Sequencing," 2018.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[20] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 139–40, 1 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19910308http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2796818

[22] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods: towards more accurate and robust tools," *bioRxiv*, p. 276907, 3 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/03/05/276907

[23] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature Biotechnology*, vol. 33, pp. 495 EP –, 04 2015. [Online]. Available: https://doi.org/10.1038/nbt.3192

[24] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[25] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, 2018. [Online]. Available: https://doi.org/10.1186/s12864-018-4772-0

[26] C. Tuglus and M. J. van der Laan, "Targeted methods for biomarker discovery," in *Targeted Learning*. Springer, 2011, pp. 367–382.

[27] M. J. van der Laan and K. S. Pollard, "A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap," *Journal of Statistical Planning and Inference*, vol. 117, no. 2, pp. 275–303, 2003.