**Benchmarking compositional data analysis methods for single-cell RNA-seq data**

**Abstract**

Compositional analysis is one of the downstream analyses for scRNA-seq, it allows us to compare the composition of cellular heterogeneity between different conditions. However, the compositionality of the negative correlation and the sum-constrained makes most of the data assumption for univariate statistical tests invalid. Recently, there have been several compositional data analysis methods developed that claim to model the compositional data more accurately. Here, we benchmark 7 common methods for compositional data analysis and compute their false positive rate, statistical power, and their performance with low sample size. We found that the traditional methods had better performance in statistical power compared to Bayesian-based methods while they had similar performance in terms of false positive rate.

**Introduction**

The increasing prevalence of single-cell RNA sequencing (scRNA-seq) has enabled biologists to study genomics at individual cell level which allows for analysis of rare cell type and decode cellular heterogeneity within a population (Quinn et al. 2019). Previously studies have elucidated rare cell types and trajectory analysis for cell dynamics. The downstream analysis of scRNA-seq includes both analysis at the gene level as well as at the cellular level (Haber et al. 2017). Compositional data analysis (CoDA) is a downstream analysis at the cellular level. It studies how the proportion of each cell type differs between conditions, for example, if the sample is compared between control and treatment group for a drug trial or wild type versus gene knockout for a scientific experiment. Additionally, it can offer insight into the developmental stage in which the progenitor cells differentiated into another cell type. In Haber et al, the researchers used Salmonella infection of mouse intestine to demonstrate that the composition of enterocytes increased compared to the control samples. The study of cell composition is enabled by the rapid advancement in scRNA-seq, yet it also presents various problems that need to be accounted for. One of the problems with compositional data is the compositionality (Greenacre, 2021). Most univariate statistical tests require assumption about the data for the result to be valid, one of the assumptions that is violated in the compositional data is the observations are independent of each other (Buttner et al, 2021). By nature, compositional data exist in a simplex space with each cell type interdependent and are negatively correlated with each other, meaning an increase in one cell type would result in the decrease of other cell type and vice versa. Due to the limitations of scRNA-seq technique, it is impossible to capture all cells present in the tissue of interest, thus the cell counts can only be viewed as relative counts instead of absolute counts (Cao et al. 2019). Additionally, because there is variability between samples within a condition from both biological and technical variabilities, CoDA method would account for this sample-to-sample variability as well. To exacerbate the problem, scRNA-seq is still considered to be an expensive assay, therefore the replicate number of samples is usually low which decreases the statistical power.

To evaluate CoDA methods comprehensively, a simulated dataset or some sort is required to evaluate the performance. Yet, since different methods require different statistical distributions, generating a dataset from one distribution will risk overlap with the method, possibly leading to overfit. At the same time, sample to sample variability is hard because the variability itself varies between species and experimental conditions. From the current best practice of scRNA-seq, it was recommended that in the absence of dedicated tools, visual comparison can be used. Yet, this is not the case most of the time due to the nature of compositionality, as one cell type decreases its proportion, other cells would increase but it is impossible to know from a simple visualization whether the former or the latter truly happened (Mangiola et al, 2023). All methods benchmarked require compositional data from scRNA-seq which always include the cell type label, which is the annotation of cell types, sometimes methods require a reference cell type; the condition label which shows control vs treatment/diseased/developmental stage; and sample label which shows sample-to-sample variability needs to be accounted for within a condition. Currently, numerous methods have been published since the best practice was published in 2019, and a benchmark was published on bioRxiv (Simmons, 2022) but there have been newer methods with Bayesian based distribution that claims to be more accurate.



**Figure 1.** A) An example of compositional data with three cell type and two conditions, each conditions contains 10 samples. Any increase in one cell type would result in the decrease of other cell types due to the constrained simplex. B) Stacked bar plot showing the composition of three cell types across all samples, a common visualization of compositional data. C) Ground truth change of the compositional data, with type C being decreased and other cell type remained unchanged. D) Composition of immune data with all control samples each consist of 6 immune cell type.

**Method**

The methods can be classified into three broad categories. The first category is the traditional statistical methods which include t-test, linear regression, and negative binomial. They are widely used in the scientific fields and precede most of the statistical software. The second category contains Dirichlet-multinomial regression which is the in-house method that our lab currently employs for compositional data analysis. The last category is the published methods that contain dedicated R packages that were published, most of the methods were published within the last 5 years, we chose corncob, propeller, and sccomp because each uses a different statistical model. For some of the traditional methods, the original data which contains raw count would need to be transformed, often with centered log-ratio transformation (CLR), but other log-ratio transformations would also work. The transformation ensures the data is normalized to convey relative information as compositional data are both interdependent and constrained to 1 as they are proportional. For this benchmark, CLR transformation is performed by dividing each compositional part by the geometric mean of all parts.

To evaluate the performance of each method, we require a ground truth dataset that includes both sample-to-sample variabilities as well as a difference in cell type abundance. The dataset is referred to as *Immune Data*, which is from *An immune-cell signature of bacterial sepsis* on Nature Medicine (cite). The immune data consists of 36 control samples and 29 sepsis samples with 106,545 cells. To establish ground truth, we filtered the data to be both control only and CD45 sorted, which resulted in 19 control samples left with 6 cell types. The cell type can be classified into 3 categories based on their proportions: abundant (T & monocyte), intermediate (NK & B cell), and rare (Megakaryocytes & DC). By doing so, we established that there is no biological significance between these samples since they are all controls, the only difference in proportion between samples of the same cell type is attributed to the sample-to-sample variabilities presented in the dataset, which is often a problem needed to be overcome by the various CoDA methods.

A false positive test is conducted to determine each method’s type I error, such that if there is no true change, the likelihood that the method would detect a statistically significant change. This was done by randomly assigning half (n = 9, rounded down to the nearest integer) of the control data (n = 19) label to treatment without changing any cell counts. Thus, we obtained biological data of control/treatment labels but without any true difference between them. Then, we input the count data to all of the CoDA methods to obtain an effect estimate with their 95% CI for each cell type respectively. To ensure robustness, the process above is bootstrapped 100 times so the sample-to-sample variability for each run will be different and allows for a more comprehensive result.

A power test is needed to determine if there is a true biological change in the data, would the CoDA methods be able to reliably detect it. It can be classified into either the addition or deletion of a cell type, or whether the cell type changed is considered to be abundant or rare. In addition, a cell type might undergo transient change to differentiate itself into another cell type, as commonly seen in progenitor cells during various developmental stages. First, we decide to model the deletion of a common cell type, monocyte. After half of the samples were assigned as treatment, we removed cell counts of monocytes at a threshold level of either 10%, 30%, or 50% of their original cell counts. Then the data is inputted into various CoDA methods to determine the effect estimate and 95% CI for all cell types. This process is bootstrapped 50 times for each of the removal thresholds, resulting in 150 bootstrapped samples per method. For cell type addition, the process remained the same as above except for changing the cell type from monocytes to DC which is a rare cell type. The addition threshold is set to be 50%, 75%, and 100% since DC has low cell counts. Both of these setups attempted to replicate two common scenarios with biological count data from scRNA-seq; the deletion of abundant cell type and the addition of rare cell type. The final scenario we have not considered is the transient change of one cell type to another. We have done so by converting 10%, 30%, and 50% of the NK cells to B cells in the treatment samples, although biologically impossible, their respective proportion in the samples is relatively similar as both are present in intermediate quantities. All three of these experiments have similar setups with modifications done on the “treatment” samples which are randomly bootstrapped from the control-only data, thus allowing for establishing a ground truth of proportional change in the data.

It is necessary to have a gold standard dataset where cell type changes under different conditions. Thus, we require another dataset to evaluate the methods’ performance. The data referred to as *COVID data* is from Impaired local intrinsic immunity to SARS-COV-2 infection in severe COVID-19 published in Cell (cite). The data consists of control-healthy, control-ICU, and COVID with severity ranked on WHO scales. For this thesis, we filtered the samples with control healthy and COVID WHO 6-8 indicating severe outcomes, which leaves us with 36 samples: with 15 control samples and 21 COVID samples. There is a total of 18 cell types but plasmacytoid DCs and mast cells contained zeros in over 80% of the samples and thus removed accordingly. Other cell types with zeros in less than 80% of the total samples were replaced using a multiplicative simple replacement approach with R package *zCompositions* which will preserve the compositionality of the data. We input the COVID data into all the CoDA methods to check whether there is a significant change between control and COVID patients. We also compare the results from each method to the compositional changes claimed in the original paper.

The downsampling test evaluates each CoDA method on whether it can detect a statistically significant change should there be one under a low sample size, often limited by the prohibitive cost of single-cell sequencing. We chose to subsample 3, 5, and 10 samples from each condition respectively in the COVID dataset. For each subsampling threshold, we repeat the subsampling 10 times and check with the original finding in the previous section. The only cell types we will compare against would be the ciliated and secretory cells, as all methods detected statistically significant changes for these cell types between different conditions.

**Result**

In the false positive test, none of the cell types have true proportional changes as their conditions are randomly assigned from the control samples only. If a method detects a statistically significant change, it would be classified as a false positive (FP) while if the change is not statistically significant, it would be classified as a true negative (TN). The statistical significance is determined by whether the CI includes zero. A significant CI would contain a solid colour while a non-significant CI would have a transparent colour (Figure 2a). Each method contains 600 CIs, aggregated by 6 cell types and 100 bootstrap iterations. Corncob contains 9 FPs, the least number of FP out of all 7 methods while negative binomial contains 57 FPs, the greatest number of FP. The t-test, propeller, Dirichlet-multinomial, sccomp, and rlm had 33, 34, 32, 52, and 55 FPs respectively, note that Dirichlet-multinomial requires a reference cell type to compare the compositional change to that cell type, in which the T cell was chosen. This resulted in 100 fewer comparisons than other methods. The false positive rate was obtained by the formula FP / (FP + TN), which is shown in Figure 2b with corncob having the lowest FPR of 0.015 and the negative binomial with the highest FPR of 0.095. The most common threshold for false positive rate when conducting compositional data analysis is 5%, thus we compared each method’s FPR with the threshold, we are aware that the bootstrap introduces some degree of uncertainty and computed the confidence interval of the FPR using prop.test in R. With the confidence interval considered, we conclude that sccomp, rlm, and negative binomial did not meet the threshold of 5% FPR. Aggregating all the cell type masked the differences between cell types, therefore we decided to look at the effect estimate for each cell type separately (Figure 2c). Although the median of the effect estimates remained around 0 for all the cell types, it is evident that NK had more outliers in their effect estimates. This was a surprising finding since we would assume the rare cell to have more fluctuations due to their low absolute counts. To investigate it further, we also looked at the number of statistically significant hits by cell type per method (Figure 2d). We found that monocyte and NK cells are more prone to false positives than other cell types. Monocyte and NK cells contained 65 and 68 FPs respectively while Megakaryocytes had only 34 FPs and the remaining cell types (B, DC, T) all had 35 FPs. Computing the false positive rate for each cell type, we obtained that while Megakaryocyte, B, DC, and T cells had between 0.049 to 0.058, NK and monocytes had a much higher FPR of around 0.093 and 0.097 respectively. It appears that both NK and monocytes had higher dispersion than the rest of the cell types, meaning that their presence in the samples is less consistent, which offers the possible explanation for the method to incorrectly detect the sample-to-sample variability as true proportional change.



**Figure 2.** A) All confidence intervals computed by bootstrapping 100 times the immune for each method. Each bootstrap iteration consists of 6 immune cell type from 19 samples. B) False positive rate computed by bootstrapping 100 times with the confidence interval of FPR computed by prop.test. C) Boxplot of effect estimate for each cell type by method. D) Heatmap showing the number of statistically significant result for each cell type per method.

For power analysis, we start with the deletion of monocytes which is considered an abundant cell type and the addition of DC which is regarded as a rare cell type. In this case, after running the analysis, any statistically significant result on the targeted cell type (monocyte for deletion and DC for addition) will be a true positive while a non-significant result would be a false negative. We know because we modified the absolute count by a certain threshold on these targeted cell types. This way we can calculate the power by subtracting the type 2 which is computed by FN / (TP + FN). In the deletion of monocyte scenario, most methods had low power when the percent deleted was at 10%. As the deletion percentage increased, the power of all methods also increased. At 50% deletion of monocyte, three methods (rlm, negative binomial, and t-test) achieved 100% power, followed by sccomp and Dirichlet-multinomial (Figure 3a) and lastly, by propeller and corncob. The overall ranking was an aggregation of the power at each threshold level, in this case, the negative binomial had the highest power of 0.55 followed by t-test and rlm with 0.52 and 0.49 respectively. It is then followed by Dirichlet-multinomial, sccomp, and propeller with 0.44, 0.42, and 0.23 respectively. The method with the lowest power in this case is corncob which had an average power of 0.14. In the addition of DC cells, all methods had a power below 0.4 at the 50% addition threshold. As the addition percentage increased, the power also increased but the overall power was lower than the deletion of monocyte, likely because DC as a rare cell type was harder to determine the true proportional change. The highest power was achieved by the negative binomial method with 0.53 and trailed by rlm, propeller, t-test, sccomp, corncob, and Dirichlet-multinomial, in that order. In this scenario, the Dirichlet-multinomial had the lowest power with only 0.007; out of the 150 cases of cell type proportional change, it is only able to detect 1 case. The effect estimates for the deletion of monocytes showed that at 10% deletion, none of the methods detected a credible shift in the effect, with their median centered around 0. At 30% deletion, we see all methods showed a negative shift in the effect estimate accurately around the threshold level. While at 50% deletion, all methods showed a negative shift, however, the effect estimate is an overestimate of the true effect, with most methods having their median estimate around 0.75 (Figure 3b, c). For the addition of DC cells, we observe an opposite effect such that the methods all tend to provide an underestimate of the true effect. At around each addition threshold of DC cells, most of the methods had a lower estimate than the actual effect size (Figure 3b, d). More importantly, both Bayesian-based methods sccomp and Dirichlet-multinomial did not detect a statistically significant change at the 50% addition threshold. When the addition threshold is increased to 75%, only the aggregated confidence interval of the Dirichlet-multinomial contains 0. All methods detected statistically significant change when the addition threshold increased to 100%.



**Figure 3.** A) Statistical power of each method at each modification threshold (left: deletion of monocyte, right: addition of DC cells). B) Effect estimate of the cell type modified by each method at different modification threshold (left: deletion of monocyte, right: addition of DC cells). C) Aggregated confidence interval of each method at different deletion threshold for monocyte, the dotted line represents the true deletion change and the solid line represent the null effect. D) Aggregated confidence interval of each method at different addition threshold for DC cells, the dotted line represents the true addition change and the solid line represent the null effect.

When the NK cells change their proportions to B cells, we observe an overall decrease in power across all methods compared to the deletion/addition experiment. We see that sccomp is the best-performing method which had highest power for B cells and the second highest for NK cells (Figure 4a). While other methods had higher power for one cell type but not the others. A consistent worst performer is the beta-binomial corncob package, with a lower than 0.1 power across both NK and B cells. However, when we analyzed the number of statistically significant results by cell type, it appears that sccomp had more significant hits across the cell type that is neither NK cells nor B cells, with monocyte as the most abundant significant results (Figure 4c). Furthermore, we see that most methods except for propeller and corncob, had more significant hits in cell types that are neither NK cells nor B cells.



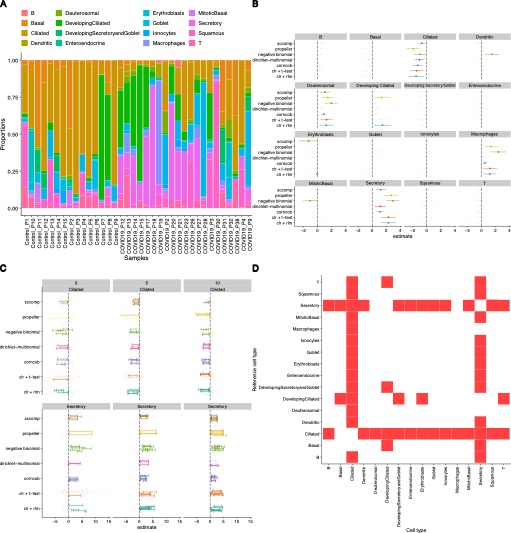
**Figure 4.** A) Statistical power of each method for the transient change from NK cell to B cell at different percentage changed. B) Heatmap showing the number of statistically significant change per method for each cell type. C) Bar plot showing the number of statistically significant hit for the cell type that was changed (NK & B cells) vs. other cell types. D) Runtime of each iteration for all methods measured in seconds.

The runtime for each method was compiled with 100 bootstrap iterations and divided accordingly to obtain the average runtime per iteration. We can see that frequentist methods all only needed a fraction of a second while the Bayesian-based methods required a median of 9 seconds for sccomp and a median of 43 seconds for the inhouse Dirichlet-multinomial method (Figure 4d). This is likely due to the initialization and resampling of these Bayesian methods, causing them to take a longer time than the frequentist methods.

We then ran each of the methods on the COVID data to determine the performance of each method on a more complicated dataset. COVID data is a heterogeneous dataset consisting of both healthy and COVID participants (Figure 5a). We can see that within the condition, there is a lot of sample-to-sample variability. For example, participants 7 & 8 in the healthy condition expressed a higher proportion of developing ciliated, goblet, and secretory cells than other healthy participants. Additionally, participant 19 in the COVID condition group had a higher proportion of macrophages than other COVID participants. The paper associated with this COVID dataset reported that there is a significant difference in proportion between deutersomal cells, developing ciliated cells, ciliated cells, and secretory cells. While we observed a statistically significant change in proportion for ciliated and secretory cells for all methods benchmarked, we did not see such a concordance for deutersomal cells and developing ciliated cells (Figure 5b). For deutersomal cells, the Dirichlet-multinomial methods did not detect a statistically significant change. For developing ciliated cells, only propeller and robust linear regression determined the proportion to be significantly different. Interestingly, the proportion of macrophages was not deemed significant in the original paper, yet we found that all methods except for sccomp and Dirichlet-multinomial detected a significant change in the proportion of macrophages. We can only ascertain the proportional change in secretory cells, as the original paper used flow cytometry to verify the increase of secretory cells.

For downsampling analysis, we found that when the sample size is downsampled to only 3 samples per condition, negative binomial has the most consistent statistically significant results for ciliated and secretory cells (Figure 5c). This was followed by corncob that uses a beta-binomial model with half of the results remaining significant. This was followed by linear regression, sccomp, Dirichlet-multinomial, t-test, and propeller in that order. With the propeller method only able to detect 2 statistically significant results out of 20 expected. When the downsampling threshold is increased to 5 samples per condition, we observe a similar trend with linear regression and corncob tied for being the best performer, classifying 16 statistically significant results out of the ground truth of 20 (10 for ciliated and secretory, respectively). It is then followed by negative binomial and sccomp, each classifying 14 and 13 statistically significant results respectively. Trailed by Dirichlet-multinomial, t-test, and propeller. With the latter only containing 4 statistically significant results. When the sample size is increased to 10 samples per condition, we see the similar trend with corncob, linear regression, and negative binomial being the top three methods for remaining the statistical significance after downsampling. Each of the methods contained 17, 14, and 14 hits respectively. The number of significances after downsampling are then followed by t-test, propeller, sccomp, and Dirichlet-multinomial, in that order, with Dirichlet-multinomial achieving the lowest with 7 statistically hits out of 20 supposed hits. Overall, as we increased the number of downsampling threshold, all of the methods appear to increase its power, yet we see that corncob, negative binomial, and linear regression consistently perform well on these tasks.

We also investigated the effect of selecting reference cell type on the result. The only method from this benchmark study that requires a reference cell type is the in-house Dirichlet-multinomial method implemented with brms in R. We see that when we select the reference cell type to be secretory or ciliated cells, the results would be bogus as most of the cell type for the COVID data would be statistically significant (Figure 5d). This is because the composition of both secretory and ciliated cells are likely to be shifting with confirmation of composition change in secretory cells. When the reference cell type is developing ciliated cells, we observe that there are more cell types registered as statistically significant by the Dirichlet-multinomial method. Interestingly, developing ciliated cells is reported in the original paper to be changing in their composition between healthy and COVID participants. With most cell types as the reference, we only observe ciliated and secretory cells to be statistically significant for the most part.



**Figure 5.** A) Bar plot of the composition of the COVID data with 15 controls and 21 COVID participants. B) Confidence interval and effect estimate for each cell type of the COVID data between conditions. C) Downsampling of ciliated and secretory cells at different downsampled threshold hold of 3, 5, and 10 per condition. D) Binary heatmap showing the statistical significance of cell type based on the reference cell type chosen for Dirichlet-multinomial method.

**Discussion**

The false positive test evaluates when there is no true change in cell type composition, how likely is a method to detect a statistically significant change. In this case, the filtered immune data provide a way to evaluate the FPR as it only contains false positives and true negatives. With 100 bootstrap iterations, we are able to reliably formulate false positive rates for each method. Ideally, we want the FPR to be low, but practically it is impossible and would reduce the power of a method if the false positive rate is too low. Since most research studies choose the FPR to be at 0.05; corncob, t-test, propeller, and Dirichlet-multinomial met this threshold with their FPR confidence interval computed. Due to the uncertainty of compositional data, some researchers might use a more liberal FPR threshold of 0.2. In that case, all of the methods evaluated in this benchmark study would satisfy that requirement. Interestingly, the abundance of a cell type does not appear to have a sizable impact on the FPR. A cell type is more likely to be considered as a false positive when its dispersion is high, that is, if the sample-to-sample variability is large within a condition. Another important metric to evaluate is the power of a method, such that, if there is a true change in proportion, how likely is a method to detect it. For this, we simulated the deletion of cell type and the addition of cell type, ranging from both abundant cell type to rarer cell type that might not be detected by each method. For the deletion of an abundant cell type, we chose monocyte in the immune data and discovered that the traditional methods like robust linear regression, negative binomial, and t-test achieved higher statistical power than the methods that were published recently like sccomp which uses Bayesian framework of a sum-constrained beta-binomial distribution. We also concluded that all methods had higher power if the composition of an abundant cell type is changing when compared to when the composition of a rare cell type is changing. The effect estimate for the deletion of abundant cell type proves to be more accurate than the addition of rare cell type. We did find that the in-house method, Dirichlet-multinomial, performs particularly poorly on the addition of rare cell types. This might be attributed to the selection of reference cell type which would mask the true effect size change, this is also investigated later in the benchmark. Additionally, sccomp had the highest number of “off-target” hits, meaning that they detected more change in the cell type that was not added manually. Although the proportion of all cell types will change in response to addition of counts to a particular cell type, it is still imperative that the method detect the change mostly from the cell type that is being modified. With the “off-target” result and the poor performance on the addition of rare cell type, two of the Bayesian-based methods, sccomp and Dirichlet-multinomial did not perform well on these tasks. For transient change in which one cell type shifts into another cell type, often useful in trajectory analysis or developmental stages, all the CoDA methods achieved low statistical power. This is likely because of the compositionality where an increase in one cell type would result in the decrease of others. The method accounts for this and does not register the change for other cell types as statistically significant. Overall, the transient change between cell types proved to be difficult for all methods benchmarked. For the COVID dataset, all the methods showed high concordance with the original paper in which the dataset originated. This includes the change of composition in secretory cells which was detected by all methods and experimentally verified using flow cytometry in the paper. The paper claimed the composition change in ciliated, deutersomal, developing ciliated, and secretory cells. Out of these cell types, all methods detected a composition change in ciliated and secretory while only propeller and robust linear regression detected change in developing ciliated cells. All methods except for Dirichlet-multinomial detected the change in deutersomal cells. Furthermore, most of the methods except for the Bayesian-based methods, sccomp and Dirichlet-multinomial called the change of composition in macrophage to be significant. Overall, the result showed not much difference between the traditional and Bayesian methods.

**Conclusion**

Overall, this benchmark compared the performance of 7 commonly used methods for compositional data analysis with scRNA-seq data. We evaluated key metrics like false positive rates in which corncob and t-test achieved the best performance while linear regression and negative binomial had the worst performance. Although an overall ranking would be difficult, it is clear to see that the traditional methods still perform relatively well over some of the newer methods that use Bayesian statistics to model the compositional data. This benchmark effort enables researchers with scRNA-seq data to better understand the approach to compositional analysis and which methods to pick depending on whether they value false positive rate or statistical power more.

**Reference**

Büttner, M. et al. (2021) ‘scCODA is a Bayesian model for compositional single-cell data analysis’, Nature Communications, 12(1), p. 6876. Available at: <https://doi.org/10.1038/s41467-021-27150-6>.

Cao, Y. et al. (2019) ‘scDC: single cell differential composition analysis’, BMC Bioinformatics, 20(19), p. 721. Available at: <https://doi.org/10.1186/s12859-019-3211-9>.

Greenacre, M. (2021) ‘Compositional Data Analysis’, Annual Review of Statistics and Its Application, 8(Volume 8, 2021), pp. 271–299. Available at: <https://doi.org/10.1146/annurev-statistics-042720-124436>.

Mangiola, S. et al. (2023) ‘sccomp: Robust differential composition and variability analysis for single-cell data’, Proceedings of the National Academy of Sciences, 120(33), p. e2203828120. Available at: <https://doi.org/10.1073/pnas.2203828120>.

Phipson, B. et al. (2022) ‘propeller: testing for differences in cell type proportions in single cell data’, Bioinformatics, 38(20), pp. 4720–4726. Available at: <https://doi.org/10.1093/bioinformatics/btac582>.

Quinn, T.P. et al. (2019) ‘A field guide for the compositional analysis of any-omics data’, GigaScience, 8(9), p. giz107. Available at: <https://doi.org/10.1093/gigascience/giz107>.

Simmons, S. (2022) ‘Cell Type Composition Analysis: Comparison of statistical methods’. Available at: <https://doi.org/10.1101/2022.02.04.479123>.

Smillie, C.S. et al. (2019) ‘Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis’, Cell, 178(3), pp. 714-730.e22. Available at: <https://doi.org/10.1016/j.cell.2019.06.029>.

Haber, A., Biton, M., Rogel, N. et al. A single-cell survey of the small intestinal epithelium. Nature 551, 333–339 (2017). https://doi.org/10.1038/nature24489