STAT 6358 Project Proposal - Adopting the CAMP Framework of Censoring Zeros for RNA-Seq.

Callum Doyle

2025-04-16

Abstract

RNA-seq data is sparse and an ongoing challenge in this field is the choice of zero-handling method when the goal is to test for differential expression. CAMP (Chan and Li (2024)) is a recently developed method that treats zero counts as censored values, facilitating the use of survival analysis methods for differential abundance analysis (DAA). The goal of this project is to assess the feasibility of treating zero counts as censored values in RNA-seq data, in contrast to traditional methods which remove lowly-expressed genes or add a pseudocount to zeros. In particular, we aim to investigate the impact on differential analysis of RNA-seq data when zeros are treated as censored by exploring the following interesting questions. 1. Does treating zeros as censored change the differentially expressed genes (DEGs) results meaningfully? 2. Do the unique DEGs detected by the CAMP framework make sense biologically? 3. Under what conditions (for example, sample size or sparsity) is there a large overlap between CAMP and more-traditional RNA-seq methods, and equivalently, when do they disagree? By answering these questions, we aim to determine if treating zeros as censored is a valid approach that could, in-turn, open up new ideas to existing methodology by including the idea that a zero-count is partial information. Although CAMP was developed for microbiome data, the authors suggest that it could extend to RNA-seq data due to the similarity in sparsity.

This project will be uploaded to https://github.com/cjd04/stat6358project. Ideally, it will include the simulation and real data analysis code, the simulated datasets, references, and whatever else that might be relevant (if space/memory limits allow, particularly for the datasets).

Introduction

RNA Seq background, pipeline, and some common methods

RNA-Seq is arguably the most popular method for measuring differentially expressed genes (DEGs) between specific conditions and is key to learning about variation in the genome. The general protocol of RNA-seq is covered by Koch et al. (2018) and the pipeline is essentially as follows. RNA is fragmented into cDNA and sequenced with a highthroughput platform. The raw reads are aligned and mapped to genes to produce a count matrix. The data are then processed and downstream analysis is performed with the goal to identify the DEGs. The latter stage of data processing has no consensus on how to "best" prepare the data for analysis, and the choice of method at each stage can have an effect on the end results. The downstream analysis often consists of exploratory analysis such as clustering, heat maps, and volcano plots, along with testing for differential gene expression with correction for multiple comparisons. Some common tools for differential expression analysis make different assumptions about how best to model the count data. For example, edgeR - developed by Robinson, McCarthy, and Smyth (2010), and DESeq2 - developed by Love, Huber, and Anders (2014), both assume a negative binomial model. However, their normalisation steps differ. Indeed, edgeR uses the trimmed mean of M values (TMM) as the reference value, whereas DESeq2 uses the median of the ratios of each gene for that sample to the geometric mean of each gene for all samples Quinn, Crowley, and Richardson (2018). Alternatively, the ALDEx2 approach (Fernandes et al. (2013)) assumes the data are compositional and uses a Monte Carlo approach for the count matrix with sampling from the Dirichlet distribution.

Zero-inflation problem

The nature of high-throughput sequencing data leads to high sparsity, with many zeros appearing in the count matrix across samples. Many zeros in a dataset are problematic and multiple methods have been proposed to tackle the challenges that sparse datasets pose. Furthermore, the nature of how exactly the zeros arise are unknown. There is no consensus for the best choice of zero-handling method, and the choice appears to be based on preference and little else. Silverman et al. (2020) explored different zero-handling methods across a variety of scenarios (simulated and real datasets) and demonstrated that the choice of method drastically influences the outcome. Additionally, the behaviour of the different methods was examined under different zero-generating processes (ZGPs). ZGPs were categorised into the following four categories. The first category is sampling zeros, this is when a zero count for a given sample is not due to absence, but due to a sampling effect. The second category is biological zeros, this is when a zero count is truly zero. We consider that a sequence with a zero count is absent from the biological system. The third and fourth categories are technical zeros, partial and complete; where the technical bias either partially or completely inhibits measurement. Although the purpose of this project is not to investigate how well methods handle different ZGPs, it is useful to keep the types of zeros in mind when examining results for meaningful

interpretation. For instance, treating biological and sampling zeros as equivalent could cause bias in the statistical analysis, it is important to distinguish between them to the best of our ability. Overall, there is still a gap in the literature for the best way of handling zeros. It is reasonable to assume that not all zeros are the same. This gives way to the idea that some zeros actually contain partial information.

RNA-Seq data as Compositional

RNA-seq data is commonly treated as count data (Anders et al. (2013)), in which the observations are counts and analysis is performed on the count matrix to examine any changes in gene expression among the groups of interest. Read depth in RNA-seq counts is arbitrary, if the read depth across genes is not accounted for, then it is possible for one sample with a high library size to dominate the analysis. For this reason, it is important to take into account the relative abundance across samples and group conditions. McGee et al. (2019) argue that traditional count-based methods lead to distorted results and that RNA-seq data must be treated compositionally. Additionally, Quinn et al. (2018) suggest that most methods are left invalid when count data is not subject to normalisation or transformation. It can be useful to think of RNA-seq counts as portions of a whole, instead of counts. Quinn et al. (2018) provide a more detailed discussion on treating sequencing data as compositional.

Handling Zeros

As mentioned previously, there is no consensus on the best method for handling zeros in RNA-seq data and the choice appears to mostly be based on preference. Two common approaches used to deal with zeros are zero-inflated models and pseudocounts. Zero-inflated models typically model the count matrix using a mixture distribution to inflate the abundance of zeros. For example, one may model the count matrix as negative binomial, but add an additional layer of probability that any value is zero, this particular model is named zero-inflated negative binomial (ZINB). Pseudocounts will avoid any issues caused by zeros by adding a small value to all counts in the matrix. Silverman et al. (2020) found that both zero-inflated and pseudocount approaches performed sub-optimally compared to simple count models (e.g., negative binomial without the zero-inflated component) across multiple settings. The results were sensitive to model misspecification and therefore leaves room for an alternative perspective with a robust approach for handling zeros.

The framework introduced by Chan and Li (2024), Censoring-based analysis of Microbiome Proportions (CAMP), introduces the idea of treating zero counts as censored observations, then transforming the count matrix into data that can be analysed by survival analysis methods. The idea is to treat the zero counts as partially observed data and perform some transformations to convert the data to time-to-event-like data that is likely to be without ties, which addresses the low-power issue that nonparametric tests face in the presence of ties. CAMP assumes compositionality of the data and results show that type 1 error is well-controlled and

powerful. The authors suggest the CAMP framework can be applied to RNA-seq data, if treated as compositional. We propose to adapt the CAMP framework to RNA-seq data and compare the results to three methods developed for RNA-seq, edgeR, DESeq2, and ALDEx2. edgeR and DESeq2 are more traditional methods that do not assume relative abundance, whereas ALDEx2 assumes the data is compositional.

CAMP approach

The general idea of CAMP is as follows. Once the raw count matrix $X = \{x_{ij}\}$ has been obtained (OTU table in the case of CAMP), where $i = 1, ..., n$ indexes the samples and $j = 1, ..., p$ indexes the taxa (translation for RNA-seq data: genes), define a surrogate read count matrix and an indicator matrix $X^\star = \{x^\star_{ij}\}$ and $\Delta = \{\delta_{ij}\}$ such that $x^\star_{ij} = x_{ij}$ if $x_{ij} > d$ or $x^\star_{ij} = d$ if $x_{ij} = 0$, and $\delta_{ij} = 1$ if $x_{ij} > d$ and 0 otherwise, where $d > 0$ is a predefined detection limit in a sequencing study. Typically, d is just the minimum non-zero count across the whole count matrix. The next step is to get the relative abundance matrix which is formed from the surrogate matrix, so it is left-censored and there are no zero counts. Next, the relative abundance matrix is transformed into right-censored time-to-event data by applying a negative log-transformation. This new "time" matrix facilitates the use of survival analysis techniques and the analogous interpretation (of whether the event has occured or not) for differential gene expression is based on the presence or absence of a taxon (gene) at a specific relative abundance level. In survival analysis, we are interested in the "at-risk" population at a given time point which consists of the subjects (taxa) who have not yet experienced the event. For us, this translates to: at a given cut-off point of relative abundance, the taxa (gene) is either present (differentially expressed) or absent (not DE). This creates a 2x2 table of presence/absence for condition 1 vs 2, and a log-rank test can be conducted to test for significance. Note that no distributional assumptions are required for this test, and the transformed data is likely to be without ties, which Chan and Li (2024) acknowledge tackles the issue of ties causing reduced power in nonparametric tests.

Simulation Results

We conducted simulations under various scenarios comparing the four methods. Data generation was conducted using `compcodeR` (Soneson (2014)), where options for generation are straightforward, which makes it ideal for conducting multiple simulations. We assumed the default setting for each method; the reasoning for this is that the most optimal setting for each method under a particular scenario is not of great interest. Indeed, the interest lies in the capability of the CAMP framework to outperform other methods that were originally designed for RNA-seq data. That is, if CAMP performs similarly or better than the other traditional methods, it is reasonable to believe that treating zero counts as censored is a valid approach that would be worth exploring in greater detail.

We conducted a simulation study to compare the performance of CAMP with the three traditional methods. All settings assumed two conditions (group 1 and group 2), where the following settings were modified to produce a total of 18 simulation scenarios.

- The initial number of genes in dataset: $10,000; 15,000; 20,000$.
- The number of samples in each condition, $\frac{n}{2}$: $10; 25; 50$.
- The proportion of truly differentially expressed genes: $0.05; 0.1$.

The simulation results suggest that CAMP performs similarly to the other RNA-seq methods. Indeed, Figure 1 demonstrates that the power of CAMP follows the same pattern as the other methods across simulation settings, and although CAMP was never the most powerful method, it was consistently the second or third-best performing, and with a large enough sample size was often the second-best. Additionally, CAMP had a well-controlled type 1 error rate (see table 2 in the appendix). Performance appears to primarily depend on sample size and the number of true DEGs, rather than the initial number of genes in the dataset.

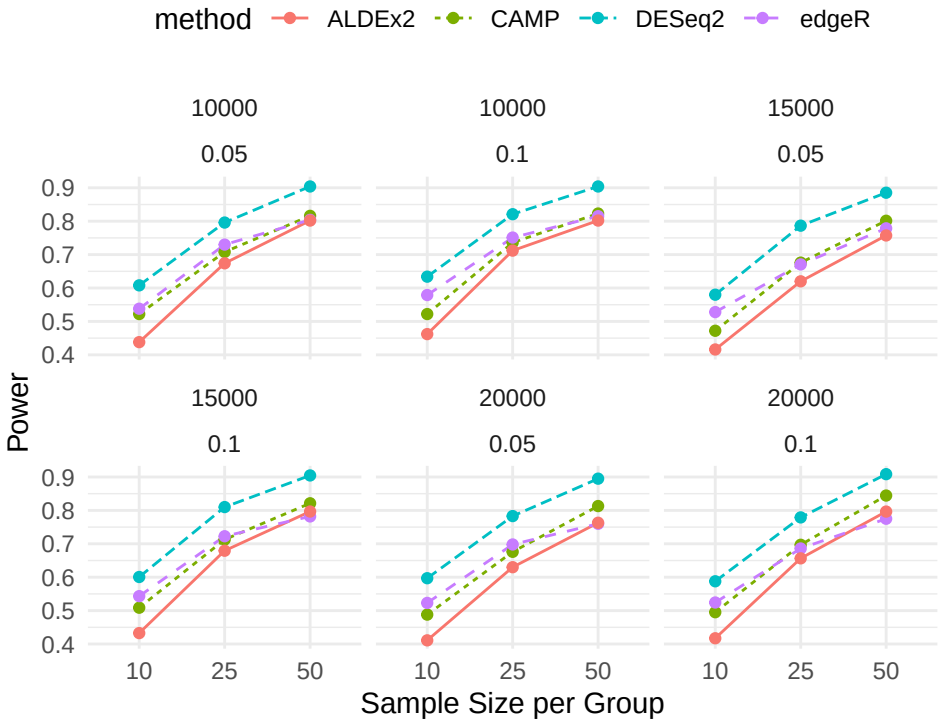

Figure 1: Power table displaying results across different simulation scenarios. There are two numbers above each chart: the top number is the total number of initial genes in the dataset; the bottom number is the proportion of true DEGs. For example, there are 1000 DEGs in the second chart in the top row.

Real Data Analysis of the Airway data

We applied the four differential analysis methods to the well-explored airway data from Himes et al. (2014). A brief explanation of how the differentially expressed genes (DEGs) were obtained for each method follows. For CAMP, we simply applied the `camp()` function to obtain the p-values and used the Benjamini-Hochberg FDR adjustment. For edgeR, we used the default filtering and normalisation approach with the functions `filterByExpr, calcNormFactors, estimateDisp` and performed the DE analysis, then adjusted the p-values. For DESeq2, we applied the `DESeqDataSetFromMatrix, DESeq` functions to obtain the p-values, followed by FDR adjustment. Lastly, ALDEx2 was performed with the default setting of 128 Monte Carlo samples and obtained the FDR adjusted p-values from the output. The number of DEGs varied substantially across the four methods, as seen in table 1. An interesting difference between the results is the clear split in the number of DEGs between the compositional-based methos and the absolute count-based methods. Although one may have like to have seen some DEGs under the CAMP approach, the result provides assurances by being consistent with its compositional companion, ALDEx2; both are very conservative. This is expected from the assumption of compositionality. We do not check for any overlapping DEGs from the airway data because our primary focus is on the CAMP approach (there are no DEGs to compare). For a more thorough comparison of the other three methods, an interested reader may refer to Quinn, Crowley, and Richardson (2018).

The code used to perform the airway analysis can be found in the file `airwayAnalysis.R` in the supplementary material (or the GitHub page).

Method	**Num. DEGs**
CAMP	0
edgeR	1,989
DESeq2	3,993
ALDEx2	9

Table 1: Number of differentially expressed genes (DEGs) detected by each method on the airway dataset. Genes were considered significant if the adjusted FDR p-value was $<$ 0.05.

Appendix

Table 2: Power and type 1 error table across each simulation scenario. The output is raw because the plot is much easier to understand and the type 1 error is pretty good all across the board for these settings (except for edgeR later on).

method	power	Type1 error	n.genes	n.per.grp	n.trueDEGs
CAMP	0.522	0.004	10000	10	0.05
edgeR	0.538	0.031	10000	10	0.05
DESeq2	0.608	0.005	10000	10	0.05
ALDEx2	0.438	0.001	10000	10	0.05
CAMP	0.472	0.003	15000	10	0.05
edgeR	0.528	0.030	15000	10	0.05
DESeq2	0.580	0.004	15000	10	0.05
ALDEx2	0.416	0.000	15000	10	0.05
CAMP	0.488	0.004	20000	10	0.05
edgeR	0.523	0.030	20000	10	0.05
DESeq2	0.597	0.005	20000	10	0.05
ALDEx2	0.411	0.000	20000	10	0.05
CAMP	0.708	0.004	10000	25	0.05
edgeR	0.730	0.041	10000	25	0.05
DESeq2	0.796	0.003	10000	25	0.05
ALDEx2	0.674	0.001	10000	25	0.05
CAMP	0.676	0.004	15000	25	0.05
edgeR	0.671	0.038	15000	25	0.05
DESeq2	0.787	0.003	15000	25	0.05
ALDEx2	0.620	0.001	15000	25	0.05
CAMP	0.676	0.003	20000	25	0.05
edgeR	0.698	0.039	20000	25	0.05
DESeq2	0.783	0.003	20000	25	0.05
ALDEx2	0.630	0.001	20000	25	0.05
CAMP	0.816	0.005	10000	50	0.05
edgeR	0.804	0.047	10000	50	0.05
DESeq2	0.904	0.004	10000	50	0.05
ALDEx2	0.802	0.002	10000	50	0.05
CAMP	0.801	0.013	15000	50	0.05
edgeR	0.779	0.044	15000	50	0.05
DESeq2	0.885	0.003	15000	50	0.05
ALDEx2	0.757	0.001	15000	50	0.05
CAMP	0.813	0.003	20000	50	0.05
edgeR	0.760	0.043	20000	50	0.05
DESeq2	0.895	0.003	20000	50	0.05

method	power	Type1 error	n.genes	n.per.grp	n.trueDEGs
ALDEx2	0.763	0.001	20000	50	0.05
CAMP	0.522	0.008	10000	10	0.10
edgeR	0.579	0.068	10000	10	0.10
DESeq2	0.634	0.007	10000	10	0.10
ALDEx2	0.462	0.001	10000	10	0.10
CAMP	0.509	0.006	15000	10	0.10
edgeR	0.543	0.064	15000	10	0.10
DESeq2	0.601	0.006	15000	10	0.10
ALDEx2	0.433	0.001	15000	10	0.10
CAMP	0.495	0.006	20000	10	0.10
edgeR	0.524	0.062	20000	10	0.10
DESeq2	0.588	0.007	20000	10	0.10
ALDEx2	0.417	0.001	20000	10	0.10
CAMP	0.736	0.007	10000	25	0.10
edgeR	0.751	0.089	10000	25	0.10
DESeq2	0.821	0.006	10000	25	0.10
ALDEx2	0.712	0.002	10000	25	0.10
CAMP	0.713	0.007	15000	25	0.10
edgeR	0.723	0.085	15000	25	0.10
DESeq2	0.810	0.006	15000	25	0.10
ALDEx2	0.679	0.001	15000	25	0.10
CAMP	0.697	0.007	20000	25	0.10
edgeR	0.686	0.080	20000	25	0.10
DESeq2	0.779	0.005	20000	25	0.10
ALDEx2	0.656	0.001	20000	25	0.10
CAMP	0.823	0.009	10000	50	0.10
edgeR	0.815	0.097	10000	50	0.10
DESeq2	0.904	0.007	10000	50	0.10
ALDEx2	0.802	0.002	10000	50	0.10
CAMP	0.821	0.013	15000	50	0.10
edgeR	0.782	0.092	15000	50	0.10
DESeq2	0.905	0.005	15000	50	0.10
ALDEx2	0.797	0.001	15000	50	0.10
CAMP	0.845	0.009	20000	50	0.10
edgeR	0.775	0.091	20000	50	0.10
DESeq2	0.908	0.005	20000	50	0.10
ALDEx2	0.796	0.002	20000	50	0.10

References

Anders, Simon, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. 2013. "Count-Based Differential Expression Analysis of RNA Sequencing Data Using r and Bioconductor." *Nature Protocols* 8 (9): 1765–86.

Chan, Lap Sum, and Gen Li. 2024. "Zero Is Not Absence: Censoring-Based Differential Abundance Analysis for Microbiome Data." *Bioinformatics* 40 (2): btae071.

Fernandes, Andrew D, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. 2013. "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PloS One* 8 (7): e67019.

Himes, Blanca E, Xiaofeng Jiang, Peter Wagner, Ruoxi Hu, Qiyu Wang, Barbara Klanderman, Reid M Whitaker, et al. 2014. "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene That Modulates Cytokine Function in Airway Smooth Muscle Cells." *PloS One* 9 (6): e99625.

Koch, Clarissa M, Stephen F Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M Ridge, Elizabeth T Bartom, and Deborah R Winter. 2018. "A Beginner?s Guide to Analysis of RNA Sequencing Data." *American Journal of Respiratory Cell and Molecular Biology* 59 (2): 145–57.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 1–21.

McGee, Warren A, Harold Pimentel, Lior Pachter, and Jane Y Wu. 2019. "Compositional Data Analysis Is Necessary for Simulating and Analyzing RNA-Seq Data." *BioRxiv*, 564955.

Quinn, Thomas P, Tamsyn M Crowley, and Mark F Richardson. 2018. "Benchmarking Differential Expression Analysis Tools for RNA-Seq: Normalization-Based Vs. Log-Ratio Transformation-Based Methods." *BMC Bioinformatics* 19: 1–15.

Quinn, Thomas P, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. 2018. "Understanding Sequencing Data as Compositions: An Outlook and Review." *Bioinformatics* 34 (16): 2870–78.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Silverman, Justin D, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. 2020. "Naught All Zeros in Sequence Count Data Are the Same." *Computational and Structural Biotechnology Journal* 18: 2789–98.

Soneson, Charlotte. 2014. "compcodeR?an r Package for Benchmarking Differential Expression Methods for RNA-Seq Data." *Bioinformatics* 30 (17): 2517–18.