RNA-seq data analysis using sequential multiple permutation testing

Lasse, Tim, Aaditya

October 4, 2024

1 Implementation

We implemented the sequential permutation test based on the anytime-valid BC p-value and the BH procedure in an R/C++ package, `adaptiveperm`. In the context of the general algorithm for sequential permutation testing (Algorithm 1), checking whether to move hypothesis i from the active set A into the rejection set R — i.e., line 7 — involves performing a BH correction on all hypotheses contained within the active set A. To avoid performing a BH correction at every iteration of the algorithm (which would require a computationally costly sort of the p-values), we instead check whether the maximum p-value $p^{\max}$ among all hypotheses in the active set can be rejected by the BH procedure:

$$p^{\max} = \max_{i \in A}\{p^i\} \leq \frac{|A|m}{\alpha}.$$

If the above criterion is satisfied, we move all hypotheses from the active set into the rejection set. This procedure produces the same discovery set as the algorithm that performs a BH correction at each iteration. We implemented several other accelerations to accelerate the algorithm, including a high-performance Fisher-Yates sampler for permuting the data (Ting 2021).

2 Genomics data analysis

We applied our sequential permutation testing method to perform a differential expression analysis of RNA sequencing (RNA-seq) data. Typically, RNA-seq data are analyzed using a parametric regression method, such as DESeq2 (Love, Huber, and Anders 2014) or Limma (Ritchie et al. 2015). However, in a recent, careful analysis of 13 RNA-seq datasets, Li et al. 2022 found that popular tools for RNA-seq analysis — including DESeq2 — did not control type-I error on negative control data (i.e., data devoid of signal). This likely was because the parametric assumptions of these methods failed to hold. Li et al. 2022 instead recommended use of the Mann-Whitney (MW) test — a classical, nonparametric, two-sample test — for RNA-seq data analysis (Mann and Whitney 1947).[1]

---

[1] According to Li et al. 2022, the RNA-seq data should contain at least 16 samples for the MW test to have sufficient power.

Li et al. 2022 applied the MW test to analyze several datasets, including a dataset of human adipose (i.e., fat) tissue generated by the Genotype-Tissue Expression (GTEx) project (Lonsdale et al. 2013). Subcutaneous (i.e., directly under-the-skin) adipose tissue and visceral (i.e., deep within-the-body) adipose tissue samples were collected from 581 and 469 subjects, respectively. RNA sequencing was performed on each of these $n = 1,050$ tissue samples to measure the expression of each of $m = 54,591$ genes. This experimental procedure yielded a tissue-by-gene expression matrix $Y \in \mathbb{N}^{n \times m}$ and a binary vector $X \in \{0,1\}^n$ indicating whether a given tissue was subcutaneous ($X_i = 0$) or visceral ($X_i = 1$). Li et al. 2022 sought to determine which genes were differentially expressed across the subcutaneous and visceral tissue samples. To this end, for each gene $j \in \{1, \ldots, m\}$, Li et al. 2022 performed a MW test to test for association between the gene expressions $Y_{1,j}, \ldots, Y_{n,j}$ and the indicators $X_1, \ldots, X_n$, yielding p-values $p_1, \ldots, p_m$. Finally, Li et al. 2022 subjected the p-values to a BH correction to produce a discovery set. Li et al. 2022 used the asymptotic version of the MW test, as it is much more computationally efficient than the finite sample version of the MW test in high-multiplicity settings. The finite sample MW test generally is calibrated via permutations.

We sought to explore whether our sequential permutation testing procedure would enable computationally efficient application of the finite-sample MW test to the RNA-seq data. To this end we applied three methods to analyze the RNA-seq data: (i) the asymptotic MW test (implemented via the `wilcox.test()` function in R); (ii) a permutation test based on the MW statistic using a fixed number $B = 5m/\alpha$ of permutations across hypotheses; and (iii) the sequential permutation testing procedure based on the MW test statistic, the anytime-valid BC p-value, and the BH procedure. We set the tuning parameter h to 15 in the latter method. We applied all three methods to analyze the data, setting the nominal FDR to $\alpha = 0.1$. We compared the methods with respect to their running time and number of discoveries.

The classical permutation test was slowest, taking over three days and 14 hours to complete. The asymptotic test and the anytime-valid permutation test, on the other hand, were much faster, running in about two minutes and three and a half minutes, respectively. All three methods rejected a similar proportion of the null hypotheses ($56.4 - 57.9\%$, depending on the method). We concluded on the basis of this experiment that the anytime-valid permutation test inherited the strenghts of both the asymptotic test and the classical permutation test: like the asymptotic test, the anytime-valid permutation test was fast, and like the classical permutation test, the anytime-valid permutation test avoided asymptotic assumptions.

It is important to note that the MW test is limited in that it does not allow for the straightforward inclusion of covariates, such as the sex of the subject from which a given tissue sample was derived. Our goal in this section was to reproduce the analysis of Li et al. 2022, swapping their asymptotic MW test for a computationally efficient, finite-sample alternative. How to test for differential expression under minimal assumptions — while adjusting for covariates — is an ongoing topic of research in statistical genomics (Niu, Choudhury, and Katsevich 2024).

Table 1: Comparison of the asymptotic MW test, the classical permutation test based on the MW statistic, and the (proposed) anytime-valid permutation test based on the MW statistic. All three methods were applied to analyze the adipose RNA-seq data. The anytime-valid permutation test was nearly as fast as the asymptotic test while avoiding asymptotic assumptions.

Method	Proportion rejected	Running time
Asymptotic test	56.6%	2m 4s
Classical permutation test	56.4%	3d 14h 24m
Anytime-valid permutation test	56.9%	3m 27s

3 Additional notes about RNA-seq data analysis

The subcutaneous tissue data, visceral adipose tissue data, and code to reproduce the RNA-seq analysis are available at the following links:

- `gtexportal.org/home/tissue/Adipose_Subcutaneous`

- `gtexportal.org/home/tissue/Adipose_Visceral_Omentum`

- `github.com/timothy-barry/sequential_perm_testing`

A standard step in RNA-seq analysis is to adjust for differences in library size across samples, where the library size l_i of the ith sample is defined as the sum of the expressions across the genes in that sample:

$$l_i = \sum_{j=1}^{m} Y_{ij}.$$

(This step sometimes is called "normalization.") We normalized the data by dividing a given count Y_{ij} by the library size of its sample, i.e. Y_{ij}/l_i. Li et al. 2022 instead employed the normalization strategy used by the R package EdgeR (Robinson, McCarthy, and Smyth 2010), which is slightly more sophisticated than the division strategy described above but similar in spirit. We filtered out genes with an expression level of zero across samples, as is standard in RNA-seq analysis (e.g., Love, Huber, and Anders 2014).

References

Li, Yumei et al. (2022). "Exaggerated false positives by popular differential expression methods when analyzing human population samples". In: *Genome biology* 23.1, p. 79.

Lonsdale, John et al. (2013). "The genotype-tissue expression (GTEx) project". In: *Nature genetics* 45.6, pp. 580–585.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Mann, Henry B and Donald R Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics*, pp. 50–60.

Niu, Zhaonan, Jayanta R Choudhury, and Eugene Katsevich (July 2024). "Computationally efficient and statistically accurate conditional independence testing with spaCRT". In: *arXiv preprint arXiv:2407.08911*.

Ritchie, Matthew E et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7, e47–e47.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1, pp. 139–140.

Ting, Daniel (2021). "Simple, optimal algorithms for random sampling without replacement". In: *arXiv preprint arXiv:2104.05091*.