Stratified Test Alleviates Batch Effects in Single-Cell Data

Shaoheng Liang[1,2], Qingnan Liang[3], Rui Chen[3], and Ken Chen[2](✉)

[1] Rice University, Houston, TX 77005, USA
[2] The University of Texas MD Anderson Cancer Center,
Houston, TX 77030, USA
{sliang3,kchen3}@mdanderson.org
[3] Baylor College of Medicine, Houston, TX 77030, USA

Abstract. Analyzing single-cell sequencing data across batches is challenging. We find that the Van Elteren test, a stratified version of Wilcoxon rank-sum test, elegantly mitigates the problem. We also modified the common language effect size to supplement this test, further improving its utility. On both simulated and real patient data we show the ability of Van Elteren test to control for false positives and false negatives. The effect size also estimates the differences between cell types more accurately.

Keywords: scRNA-seq analysis · Differential expression analysis · Batch effect · Wilcoxon rank-sum test · Van Elteren test

1 Introduction

Large-scale studies such as the Human Cell Atlas [15] involve hundreds of laboratories, thousands of patients, and millions of cells, bringing about opportunities and challenges in analyses. When comparing cell types or groups, discrepancies across experiments and differences among participants lead to omissions and false discoveries in differentially expressed genes. Even the trend (upregulated or downregulated) can be reversed in a phenomenon called Simpson's paradox [1]. Although multiple methods have been proposed to tackle such batch effects, no such option for the widely used Wilcoxon rank-sum test [11,18] has been applied to single-cell studies, to the best of our knowledge. Here, we show that the stratified rank-sum test (known as Van Elteren test [17]) and our modified common language effect size may benefit single-cell studies.

We briefly review and conceptually compare related works on correcting batch effect in Sect. 1.1. Then, in Sect. 2, we revisit Wilcoxon rank-sum test,

Supported by grant number 2018-182735 to KC, Human Breast Cell Atlas Seed Network Grant (HCA3-0000000147) to NN and KC from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation, grant RP180248 to KC from Cancer Prevention & Research Institute of Texas, and the Cancer Center Support Grant P30 CA016672 to PP from the National Cancer Institute.

C. Martín-Vide et al. (Eds.): AlCoB 2020, LNBI 12099, pp. 167–177, 2020.
https://doi.org/10.1007/978-3-030-42266-0_13

and introduce the Van Elteren test supplemented by our direct extension of the common language effect size [5,12]. Simulation studies and applications to real data in Sect. 3 show that the test controls for the batch effects and leads to more accurate biological discovers, compared with Wilcoxon rank-sum test. More discussions and explanations, are shown in Sect. 4.

1.1 Related Works

Mainstream methods to mitigate batch effect fall into two categories, batch correction methods and batch-aware statistical tests. The former includes methods reducing batch effect in the data to facilitate downstream analysis, while the latter includes analyses that control for the batch effect.

Batch Correction Methods. Batch correction methods eliminate the discrepancy among batches to create an integrated dataset. The most conspicuous manifestation of batch effect is splitting one cell type into multiple clusters. To solve this problem, many methods match and combine clusters across samples based on similarities. A commonly adopted one, Mutual nearest neighbor (MNN) [3], uses similar cells across datasets as anchors, and based on them correct the gene expression of other cells. Scanorama [4] and Seurat [16] are both based on the MNN methodology. Another method, Harmony [7], iteratively corrects the data by clustering the cells and moving neighboring clusters toward each other. These methods typically produce a unified data matrix, which can be conveniently used in visualization and downstream analysis. However, these empirical corrections usually lack negative control and raise uncertainty in the discovery [14].

Batch-Aware Statistical Tests. Instead of manipulating the data directly, statistical methods may handle batch effect by considering it as a covariate in the model. This is possible in Student's t-test, Poisson test, negative binomial test, etc. Notably, all these tests are parametric, meaning that a distribution must be given in advance. However, the debate of the true distribution of single-cell gene expression has never ceased, which is a reason why the nonparametric Wilcoxon rank-sum test is widely used. To allow modeling covariates, one may use a generalized version of rank-sum test, the proportional odds model [2]. However, modeling batches by using a covariate also makes unnecessary assumptions upon them. Stratification, which only combines statistics from batches, is the "as simple as possible, but no simpler" way to handle batches. The Van Elteren test we use, is the stratified version of Wilcoxon rank-sum test.

It is worth noting that methods like scVI [9] have combined statistical modeling with batch effect correction. However, the effect of batches is modeled by a black-box neural network, making the interpretation elusive.

2 Methods

2.1 Wilcoxon Rank-Sum Test

We briefly revisit the Wilcoxon rank-sum test (also known as Mann–Whitney U test) [11,18]. The test statistics U is defined as

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\mathbb{1}_{a_i > b_j} + \frac{1}{2} \mathbb{1}_{a_i = b_j} \right), \tag{1}$$

where $A = \{a_i\}_{i=1}^{n_1}$ and $B = \{b_j\}_{j=1}^{n_2}$ are the two samples to be compared (e.g., two cell types in one experiment), with sample sizes n_1 and n_2, respectively. Function "$\mathbb{1}$" takes value 1 when its condition holds true, and 0 otherwise. When n_1 and n_2 are both at least 10, which is common in single-cell studies, the distribution of U approximately follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where

$$\mu = \frac{n_1 n_2}{2} \tag{2}$$

and

$$\sigma^2 = \frac{n_1 n_2}{12} \left((n+1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{n(n-1)} \right), \tag{3}$$

in which t_i stands for ties (corresponding to the second term in Eq. 1).

2.2 Van Elteren Test

The Van Elteren test [17] is the stratified version of Wilcoxon rank-sum test. For example, if there are m patients, they maybe treated as strata. In that case, a U statistic may be obtained from each patient $g \in \{1, \cdots, m\}$, denoted as $U_g \sim \mathcal{N}(\mu_g, \sigma_g^2)$. A new statistics V is constructed by

$$V = \frac{\left[\sum_{g=1}^{m} w_g (U_g - \mu_g) \right]^2}{\sum_{g=1}^{m} w_g^2 \sigma_g^2} \sim \chi_1^2, \tag{4}$$

where w_g is a weight for each sample to be discussed later. When $m = 1$, the formula degenerates to $V = (U_g - \mu_g)^2 / \sigma_g^2 \sim \chi_1^2$, which is consistent with the rank-sum test.

Weights. As discussed by Van Elteren [17], the weights w_g can be assigned in different ways. It should be noted that the U_g for a batch g ranges from 0 to $n_{g1} n_{g2}$, the product of two sample sizes in the batch. Should the weights all be equal, a patient with more cells available will dominate the test results. It is proven in [17] that weight

$$w_g = \frac{1}{n_{g1} n_{g2}} \tag{5}$$

eliminates such effect, and a test utilizing such weight is thus named as "design-free test". However, given that a batch with more instances available (e.g., a patient with more cells sequenced) may be more convincing, another weight

$$w_g = \frac{1}{n_{g1} + n_{g2} + 1} \tag{6}$$

is introduced, which gives more power to larger batches. It also effectively assigns larger weights to batches whose samples are more balanced, when the batch sizes are the same. It is shown in [17] that this choice yields largest statistical power against randomized alternatives, and is thus named as "locally-best test". The comparison of two weights are shown in Fig. 1.

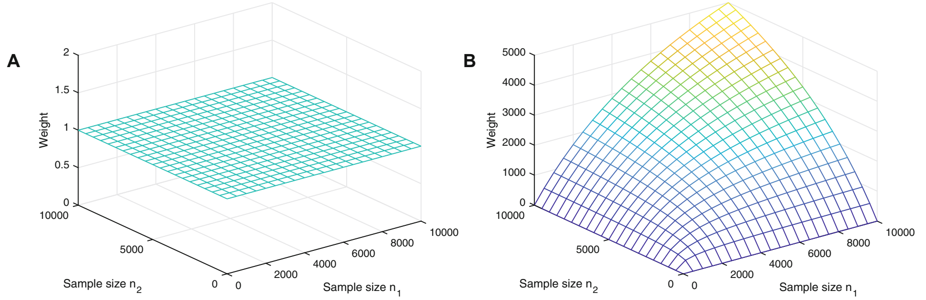

Fig. 1. The weight for each batch when using the (A) design-free test and (B) locally-best test. For design-free test all the batches have equal weights, while for locally-best test higher weights are given to batches with higher and balanced sample sizes.

Effect Sizes. For Wilcoxon rank-sum test, a simple definition of effect size is

$$f = \frac{U}{n_1 n_2}, \tag{7}$$

which is centered at 50%, meaning the probability $P(a > b)$ when a and b are randomly drawn from sample A and sample B, respectively. An effect size greater than 50% generally means that A is higher, and vice versa. It may be easily extended for Van Elteren test by taking average using desired weights. For the design-free test, the effect size is

$$f = \frac{1}{m} \sum_{g=1}^{m} \frac{U_g}{n_{g1} n_{g2}}, \tag{8}$$

as all batches are treated equally regardless of the sample sizes. It may be interpreted as the probability of $P(a_g > b_g)$ for a_g and b_g randomly drawn from

A_g and B_g, after randomly choosing a batch g. For the locally-best test, the effect size becomes

$$f = \frac{1}{\sum_{g=1}^{m} \frac{n_{g1}n_{g2}}{n_{g1}+n_{g2}+1}} \sum_{g=1}^{m} \frac{n_{g1}n_{g2}}{n_{g1}+n_{g2}+1} \frac{U_g}{n_{g1}n_{g2}} \tag{9}$$

$$= \frac{\sum_{g=1}^{m} \frac{U_g}{n_{g1}+n_{g2}+1}}{\sum_{g=1}^{m} \frac{n_{g1}n_{g2}}{n_{g1}+n_{g2}+1}}, \tag{10}$$

which changes the probability of choosing a group g to be in proportion to $\frac{n_{g1}n_{g2}}{n_{g1}+n_{g2}+1}$, giving higher weights to batches with higher and balanced sample sizes (Fig. 1). Generally, any w_g may be used to define f, as in

$$f = \frac{\sum_{g=1}^{m} U_g w_g}{\sum_{g=1}^{m} n_{g1}n_{g2}w_g}, \tag{11}$$

the two previous options being its special cases.

3 Results

We implemented the Van Elteren test with the effect size in R, available at our GitHub repository (https://github.com/KChen-lab/stratified-tests-for-seurat), based on Seurat 3.0 by utilizing its differential expression analysis part (but irrelevant to the data integration) [16]. When there are two groups of cells, `A` and `B`, denoted as `type`, and patient identity, denoted as `batch`, the Van Elteren Test can be called as follows.

```
FindMarkers(obj, ident.1 = ’A’, ident.2 = ’B’, group.by = ’type’,
            test.use = ’VE’, logfc.threshold = 0,
            latent.vars = ’batch’, genre = ’locally-best’)
```

The `genre` may be set to either `locally-best` or `design-free`, as introduced in Sect. 2, based on which p-values and effect sizes are calculated. Typical results are shown in Table 2. An effect size of larger than 0.5 indicates a higher expression in cell type `A`, and vice versa. The `avg_logFC`, average logarithmic fold changes, are calculated automatically by Seurat, where a positive value indicates a higher expression. It may show different trends compared with the effect sizes. Generally, the effect sizes are more indicative after controlling for the batch effect.

3.1 Simulation Study

We simulated datasets to illustrate the key utilities of Van Elteren test. The parameters are specified in Table 1. Poisson distribution is used to model sequencing depth. Visualization is available in Fig. 2 for illustration. We assume that the library size of each sample is equalized by other genes beyond the simulated ones.

The testing result of Van Elteren test and Wilcoxon rank-sum test is shown in Table 2. Trend (A over B) are indicated by arrows and insignificant p-values are grayed out. For Van Elteren test, the locally-best version and the design-free version return very similar results.

Table 1. Simulated datasets

Patient	Number 1		Number 2	
Cell type	A	B	A	B
Cell amount	101	30	31	100
Gene1	Pois($\lambda = 9$)	Pois($\lambda = 10$)	Pois($\lambda = 8$)	Pois($\lambda = 9$)
Gene2	Pois($\lambda = 8$)	Pois($\lambda = 6$)	Pois($\lambda = 15$)	Pois($\lambda = 14$)
Gene3	Pois($\lambda = 3$)	Pois($\lambda = 3$)	Pois($\lambda = 5$)	Pois($\lambda = 5$)
Gene4	Pois($\lambda = 5$)	Pois($\lambda = 6$)	Pois($\lambda = 5$)	Pois($\lambda = 6$)

Suppressing False Negatives. Batch effect may introduce false negatives, where a significantly differentially expressed gene is overshadowed. For gene 1, on which cell type B always have higher expression on both patients, the Wilcoxon rank-sum test did not pass the threshold of 0.05 to reject the null hypothesis, while Van Elteren test yields a significant p-value. The effect size, smaller than 0.5, also correctly suggests that the expression of gene 1 in cell type B is higher than that in cell type A, compared with the average logarithmic fold change, which wrongfully indicates otherwise.

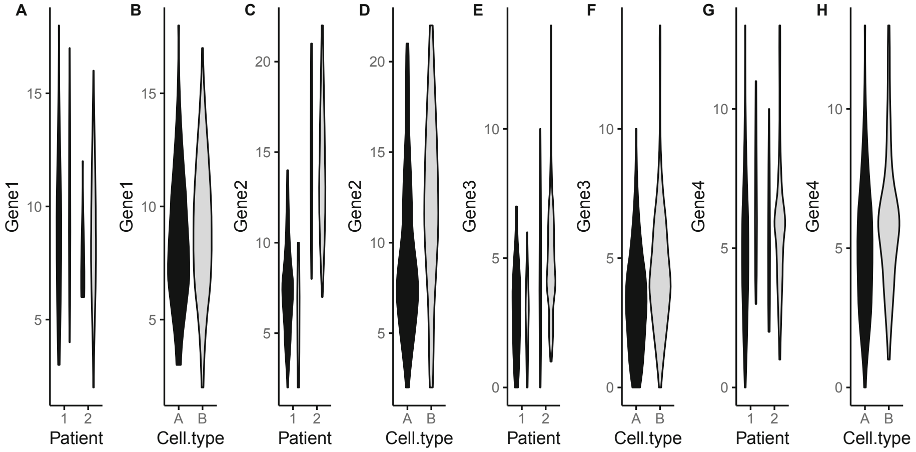

Fig. 2. Illustration of four simulated genes. Two shades correspond to two cell types (dark: cell type A; light: cell type B). For each gene, the left panel is stratified by patients and the right panel shows aggregated distribution.

Table 2. Results of the two tests in simulation studies

	P-value			Effect size		
	Wilcoxon	Van Elteren		Log fold change	Van Elteren f	
		locally-best	design-free		locally-best	design-free
Gene1	6.336E−02	3.119E−03	3.109E−03	0.673 ↑	0.375 ↓	0.375 ↓
Gene2	3.700E−08	8.469E−03	8.193E−03	−1.182 ↓	0.611 ↑	0.611 ↑
Gene3	5.770E−05	7.831E−01	7.905E−01	−3.664 ↓	0.488 ↓	0.489 ↓
Gene4	2.465E−03	2.245E−03	2.416E−03	−0.817 ↓	0.373 ↓	0.372 ↓

Suppressing Reversed Conclusions. Batch effect may also lead to reversed conclusion (i.e., which cell type has higher expression). For gene 2, on which cell type A always have higher expression value, both tests reject the null hypothesis. However, the effect size of Van Elteren test, larger than 0.5, correctly identifies that the expression of gene 2 in cell type A is higher than that in cell type B, while the average logarithmic fold change wrongfully indicates otherwise.

Suppressing False Positives. False discoveries are also possible outcome of batch effect. As is shown in gene 3, the distribution of both cell types are exactly the same in each patient. Nevertheless, Wilcoxon test yields a very significant p-value. The average logarithmic fold change also has a large magnitude. Van Elteren test returns a p-value of 0.7831, together with a effect size close to 0.5, suggesting that the difference is neither significant nor large.

Consistency. As a negative control, when the three issues above are not present, p-value from Van Elteren test is consistent with Wilcoxon rank-sum test, as is shown by gene 4. The effect size and the log fold change also both show that the cell type B has higher expression in gene 4.

3.2 Retina Data

We have tested the Van Elteren test on real retina single-cell gene expression data gathered from three patients [8]. Two regions, macula (i.e., the center area) and peripheral, are labeled in the data. We question which genes differentially express for the same cell type between two regions. We run Wilcoxon rank-sum test and Van Elteren test on 2,295 rod cells and 203 cone cells. We compare the results in Fig. 3, where genes with large differences in p-values between two tests are labeled with gene names.

Rod Cells. The results of two tests are largely comparable, showing a diagonal pattern. Meanwhile, some exceptions are present (see Table 3 for the p-values

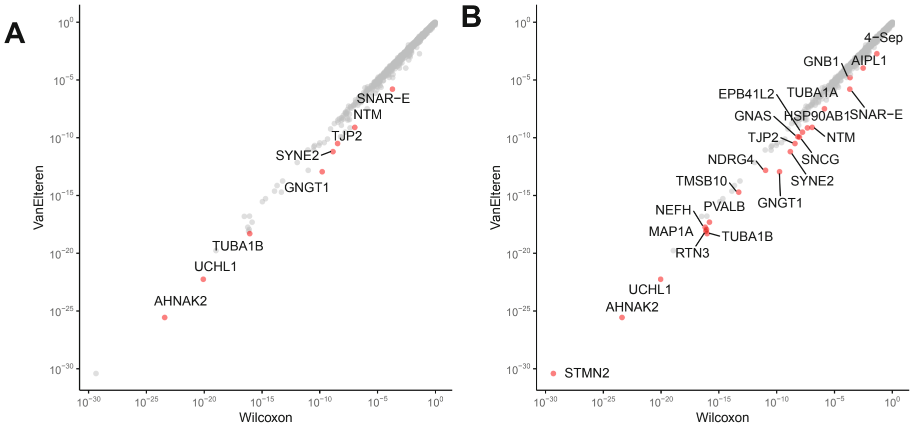

Fig. 3. Comparison of p-values returned from Wilcoxon rank-sum test and Van Elteren test on (A) rod cells and (B) cone cells. Each dot is a gene, whose p-value from Wilcoxon rank-sum test and Van Elteren test are shown by its x-coordinate and y-coordinate, respectively. Genes with largely changed p-values (10^2 for rod cells and 10^1 for cone cells) are labeled with gene names.

and effect sizes), among which we observed that p-values for gene *GNGT1* and *SYNE2* change the most.

For *GNGT1*, the reversed conclusion effect is also observed, as the Van Elteren f effect size suggests that the peripheral region has a higher expression, while the log fold change indicates otherwise. We further inspected the distributions to validate and interpret the differences. In Fig. 4A, the left panel does show generally higher expression of *GNGT1* in each individual patient, while the aggregated distribution on the right shows a reversed effect, which is an instance of the aforementioned Simpson's paradox. For *SYNE2* (Fig. 4B), conspicuous discrepancy among batches is also shown, which leads to a less precise rank-sum test result. Indeed, these two genes were found playing roles in macular degeneration diseases [6,10].

Cone Cells. For cone cells, we observed similar results. Some genes show changes while most genes are consistent across the tests. The p-values and effect sizes are shown in Table 4.

For gene *PCBP4*, Van Elteren test shows more significant p-value, and an effect size indicating smaller expression in macula, which is different from the log fold change. Decrease in *PCBP4* has also been linked with age-related macular degeneration [13]. Figure 5 shows that batch effect in distribution of *PCBP4* may have misled the rank-sum test and the logarithmic fold change.

Table 3. Gene with large p-value change in rod cells

	P-value			Effect size	
	Wilcoxon	Van Elteren	Log10 ratio	Log fold change	Van Elteren f
GNGT1	1.694E−10	1.229E−13	3.14	89.679 ↑	0.390 ↓
SYNE2	1.451E−09	6.654E−12	2.34	−48.321 ↓	0.400 ↓
TUBA1B	9.317E−17	5.521E−19	2.23	233.992 ↑	0.631 ↑
UCHL1	9.044E−21	6.300E−23	2.16	243.679 ↑	0.639 ↑
NTM	1.116E−07	8.324E−10	2.13	3.679 ↑	0.410 ↓
AHNAK2	4.179E−24	3.140E−26	2.12	178.679 ↑	0.644 ↑
SNAR-E	1.991E−04	1.664E−06	2.08	Inf ↑	0.431 ↓
TJP2	3.650E−09	3.358E−11	2.04	75.678 ↑	0.411 ↓

Table 4. Gene with large p-value change in cone cells

	P-value			Effect size	
	Wilcoxon	Van Elteren	Log10 ratio	Log fold change	Van Elteren f
PCBP4	1.734E−02	1.285E−03	1.13	5.397 ↑	0.335 ↓

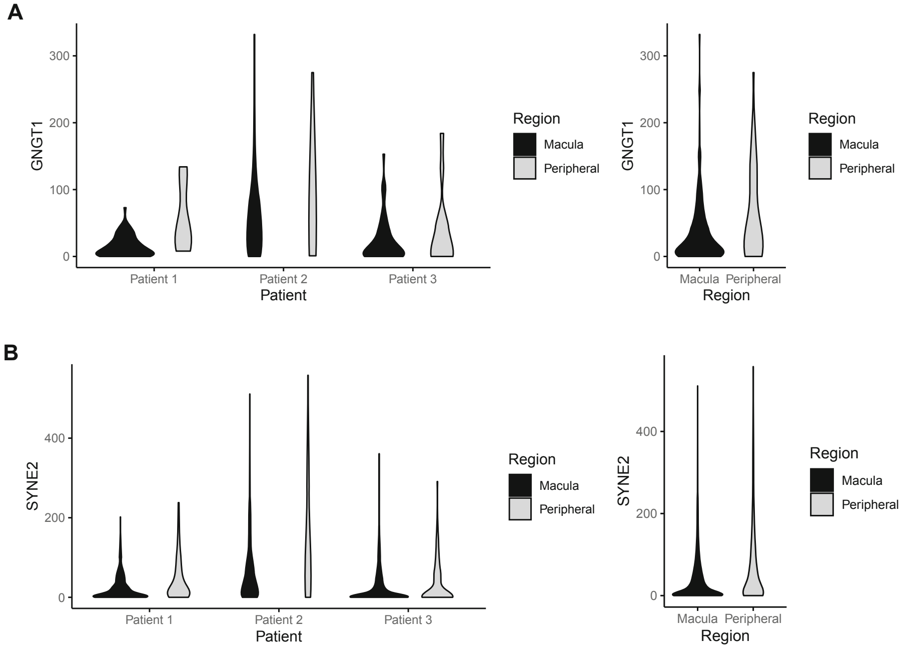

Fig. 4. Distribution of counts of (A) *GNGT1* and (B) *SYNE2* in rod cells. Left panels are stratified by patients and right panels show aggregated distributions.

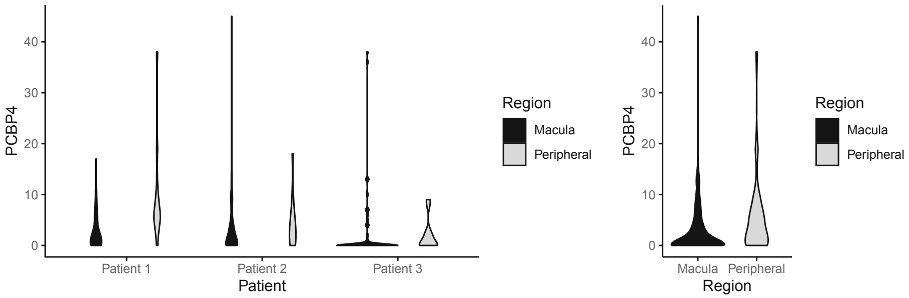

Fig. 5. Distribution of counts of *PCBP4* in rod cells. The left panel is stratified by patients and the right panel shows aggregated distributions.

4 Discussion

The results have clearly shown that Van Elteren test benefits biological studies in precisely identifying differentially expressed genes. Although the results we show do not include multiple comparison correction, Seurat 3.0 will automatically give corrected p-value based on the raw p-value using Bonferroni correction. Generally, any correction based on p-values will also apply.

The result also indicates that stratified test is a neat way to handle batch effect. Although covariate has the ability to control for explanatory variables, it is generally more suitable for continuous variables. It also casts more assumptions when modeling covariate. Stratified test, on the other hand, does not infer the influence of the discrete batches. Rather, it directly aggregates the statistical power of multiple samples.

Admittedly, for the rod cells, although changes in p-values are observed, the significance threshold was well passed by both. However, it should be noted that the retina data are collected from relatively healthy tissues and are considered clean, while Van Elteren test is expected to make a more meaningful difference on noisy pathological and tumor data. In addition, rod cell is the most populous cell in retina. For rare cell types that take smaller proportions, like the cone cells, the difference Van Elteren test makes can be crucial.

The caveat of stratified test is that for it to work the strata shall not overlap with the variable of interest. For instance, it may not find the difference, meanwhile also control for the batch effect, between two patients. Nonetheless, neither is covariate applicable to such cases. As the batch effect and biological effect are convolved, more prior knowledge is generally needed to distinguish them.

5 Summary

We have adopted Van Elteren test, an underappreciated statistical test, and our weighted common language effect size to single-cell sequencing data. When batch effect is severe, the test control for false positives. Otherwise, it is consistent with Wilcoxon rank-sum test. The modified common language effect size also faithfully depicts the trends. This work may increase the precision of differential expression analysis to help identify genes of interests.

References

1. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)
2. Everitt, B., Skrondal, A.: The Cambridge Dictionary of Statistics. BusinessPro collection, 4th edn. Cambridge University Press, Cambridge (2010)
3. Haghverdi, L., Lun, A.T., Morgan, M.D., Marioni, J.C.: Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. **36**(5), 421 (2018)
4. Hie, B., Bryson, B., Berger, B.: Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. **37**(6), 685 (2019)
5. Kerby, D.S.: The simple difference formula: an approach to teaching nonparametric correlation. Compr. Psychol. **3**, 1–10 (2014). 11-T
6. Kolesnikov, A.V., et al.: G-protein $\beta\gamma$-complex is crucial for efficient signal amplification in vision. J. Neurosci. **31**(22), 8067–8077 (2011)
7. Korsunsky, I., et al.: Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods **16**, 1–8 (2019)
8. Liang, Q., et al.: Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling. Nat. Commun. **10**(1), 1–12 (2019)
9. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nat. Methods **15**(12), 1053 (2018)
10. Maddox, D.M., et al.: A mutation in Syne2 causes early retinal defects in photoreceptors, secondary neurons, and Müller glia. Invest. Ophthalmol. Vis. Sci. **56**(6), 3776–3787 (2015)
11. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**, 50–60 (1947)
12. McGraw, K.O., Wong, S.: A common language effect size statistic. Psychol. Bull. **111**(2), 361 (1992)
13. Meyer, J.G., Garcia, T.Y., Schilling, B., Gibson, B.W., Lamba, D.A.: Proteome and secretome dynamics of human retinal pigment epithelium in response to reactive oxygen species. Sci. Rep. **9**(1), 1–12 (2019)
14. Nygaard, V., Rødland, E.A., Hovig, E.: Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics **17**(1), 29–39 (2016)
15. Regev, A., et al.: Science forum: the human cell atlas. Elife **6**, e27041 (2017)
16. Stuart, T., et al.: Comprehensive integration of single-cell data. Cell **177**, 1888–1902 (2019)
17. Van Elteren, P.: On the combination of independent two-sample tests of Wilcoxon. Bull. Inst. Int. Staist. **37**, 351–361 (1960)
18. Wilcoxon, F.: Individual comparisons by ranking methods. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics), pp. 196–202. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-4380-9_16