

---

# Normalisr: inferring single-cell differential and co-expression with linear association testing

---

Lingfei Wang<sup>1 2</sup> Jacques Deguine<sup>1</sup> Ramnik J. Xavier<sup>1 2</sup>

## Abstract

ScRNA-seq may provide unprecedented technical and statistical power to study gene expression and regulation, but differential expression (DE) and co-expression remain challenging due to its sparsity and technical variations. Here we present Normalisr, a parameter-free normalization-association two-step inferential framework for scRNA-seq that unifies case-control DE, co-expression, and pooled CRISPRi scRNA-seq screen under linear association testing. Normalisr addresses those challenges with posterior mRNA abundances, nonlinear cellular summary covariates, and mean and variance normalization. Consequently, Normalisr achieves optimal sensitivity, specificity, and speed in all above scenarios. Normalisr recovers high-quality transcriptome-wide co-expression networks from conventional scRNA-seq and robust gene regulations from pooled CRISPRi scRNA-seq screens. Normalisr provides a unified framework for optimal, scalable hypothesis testings in scRNA-seq.

## 1. Introduction

Understanding gene regulatory networks forms a major part of most biological studies. ScRNA-seq provides a unique glance into cellular transcriptomic variations beyond the capabilities of bulk technologies, enabling novel biological questions such as single-cell DE, co-expression, and causal network inference on cell subsets at will. However, neither existing bulk computational methods nor newly proposed single-cell methods could account for its low read counts and cell-to-cell technical variations (Soneson & Robinson, 2018). It also remains unclear how to unify DE, co-

expression, and causal network inference in one framework (Lähnemann et al., 2020).

ScRNA-seq normalization (Hafemeister & Satija, 2019) and imputation (Dijk et al., 2018; Eraslan et al., 2019) aim to address those challenges and recover the true, biological, but hidden expression levels, which any analyses may then operate upon. However, most such methods were clustering focused. Sensitivity and specificity in DE and particularly co-expression are much more direct, loyal reflections of true expression recovery, and are also major goals of normalization, but were left mostly uncharted. Consequently, high-quality DE and gene-level co-expression remain challenging in scRNA-seq (Lähnemann et al., 2020).

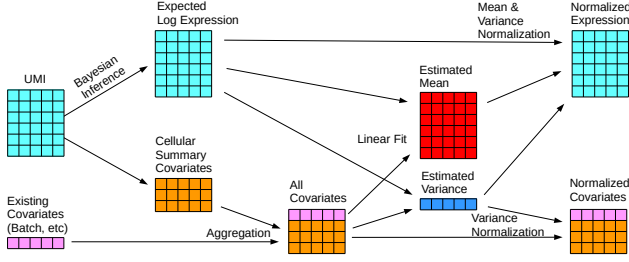
Here, we present Normalisr, a non-parametric normalization-association two-step framework for statistical inference of gene regulations and co-regulations in scRNA-seq. The normalization step estimates the pre-measurement biological mRNA levels from the scRNA-seq read counts. The association step, utilizing linear association testing, unifies DE, co-expression, and potentially beyond to interrogate the biological system with numerous advantages: (i) exact P-values, (ii) native removal of covariates, (iii) non-parametric robustness, (iv) unbeatable time and memory complexities, and (v) extensibility in genetic relatedness. We present Normalisr’s superior sensitivity, specificity, and speed in differential and co-expression on evaluation datasets. We demonstrate its applications in two scenarios — the reconstruction of a transcriptome-wide co-expression network of dysfunctional T cells in melanoma, and the search for gene regulations in pooled CROP-seq screens.

## 2. Method

Normalisr first estimates the expectation of logCPM for the binomial (approximate of multinomial) mRNA sampling process (Svensson, 2020), using a posterior Beta distribution for every gene in every cell (Fig 1). This avoids the artificial choice of constant in the conventional transformation  $\log(\text{CPM} + \text{constant})$ . To account for technical variations, we introduced two covariates that are known to confound gene expression: the number of 0-UMI genes (Finak et al.,

---

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA <sup>2</sup>Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. Correspondence to: Lingfei Wang <lingfei@broadinstitute.org>, Ramnik J. Xavier <xavier@molbio.mgh.harvard.edu>.



**Figure 1. Normaliser overview.** Left to right: Normaliser starts by computing the expectation of posterior distribution of mRNA proportion in each cell. Meanwhile, Normaliser appends existing covariates with cellular summary covariates. Normaliser then estimates the mean and variance of gene expression by linearly regressing out covariates. The estimated mean and variance are then applied to normalize gene expression and covariates. Their normalized values are ready for downstream linear association testing, such as differential and co-expression.

2015) and log total UMI count (Hafemeister & Satija, 2019). Using only such unbiased cellular summary statistics minimizes spurious gene inter-dependencies tampered by normalization or imputation, and their interference with true co-expressions.

We introduced Taylor expansion to account for nonlinear effects of cellular summary covariates. In an iterative decision process, we found that the square of log total UMI count alone would suffice as an extra covariate, to recover the uniform null distribution of co-expression P-values (not shown). Those technical covariates are linearly regressed out at mean and log variance levels to maintain possible biologies of variance fluctuations. DE and co-expression tests then take the form:

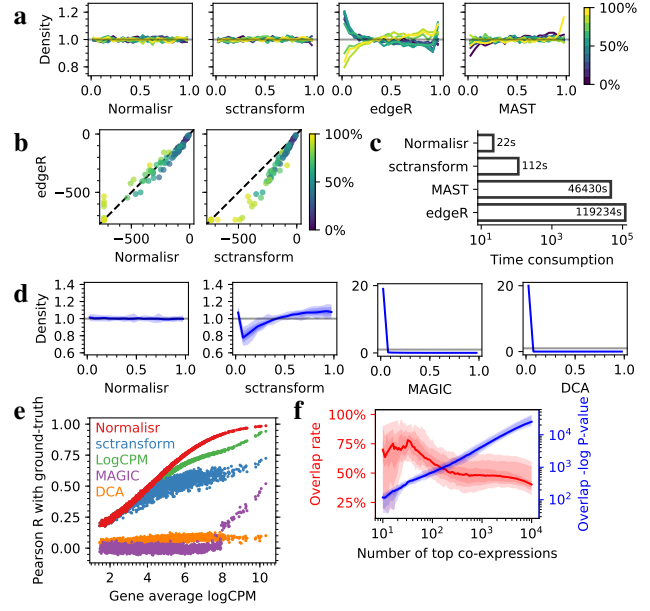
$$\text{lcpm}_i = \alpha x + \beta C + \varepsilon, \quad \varepsilon \sim i.i.d N(0, \sigma^2), \quad (1)$$

where  $\text{lcpm}_i$  is the normalized expression of gene  $i$ ,  $C$  is normalized covariates, and  $x$  is binary grouping of cells for DE and  $\text{lcpm}_j$  for co-expression. Exact P-values are computed from Beta distribution without permutation.

### 3. Results

#### 3.1. Normaliser obtained optimal sensitivity and specificity in single-cell DE and co-expression

P-values and effect sizes are essential (and irreplaceable by rankings) in comparisons across experiments or even genes for differential or co-expression. By definition, P-values need to follow the (standard) uniform distribution under the null hypothesis, *i.e.* without differential or co-expression. Biased null P-values towards 0 or 1 penalize specificity or sensitivity. Expression-dependent biases introduce extra noises and also reduce sensitivity. Method evaluation in a controlled setting is critical.



**Figure 2. Normaliser achieved optimal sensitivity, specificity, speed, and robustness in single-cell DE & co-expression.** **a** DE specificity in terms of null P-value histograms, for genes in 10 equally sized and separately colored bins stratified by expression (proportion of cells expressing gene). **b** DE sensitivity comparison in terms of log P-values on CRISPRi gRNA-target positive control pairs. Genes are colored by the proportion of expressed cells. Dashed line indicates  $X=Y$ . **c** Time consumption comparison for null DE. **d** Co-expression specificity in terms of null P-value histograms. Genes were split into 10 equal bins by expression. P-value histogram curves for each bin-pair were aggregated. **e** Pearson R between (log) normalized/imputed expressions and ground-truth of each gene across cells. **f** Overlap rate (left, red) and P-values (right, blue) of Normaliser inferred co-expression networks between each batch. Curves for each batch-pair were aggregated. **Histograms:** Gray lines indicate the expected uniform distribution for null P-value. **Central curves and shades** for a set of curves show median, 50%, 80%, and 100% of all curves.

We used the Perturb-seq dataset (Adamson et al., 2016) for method evaluation. We tested different normalization and state-of-the-art DE methods between randomly grouped cells for 100 times. Normaliser and sctransform (Hafemeister & Satija, 2019) could recover uniform distributions of null DE P-values at all expression levels (Fig 2a). EdgeR and MAST, the best single-cell DE methods benchmarked in Sonesson & Robinson (2018), had expression-dependent, 0- or 1-biased null P-values. Meanwhile, in a sensitivity evaluation of DE of CRISPRi targeted genes, sctransform yielded weaker P-values for those true positives than Normaliser and edgeR (Fig 2b). Normaliser was much faster than others, especially being over 2,000 times faster than edgeR and MAST (Fig 2c). Normaliser had the optimal sensitivity, specificity, and speed in differential gene expression studies.

We then evaluated normalization and imputation methods

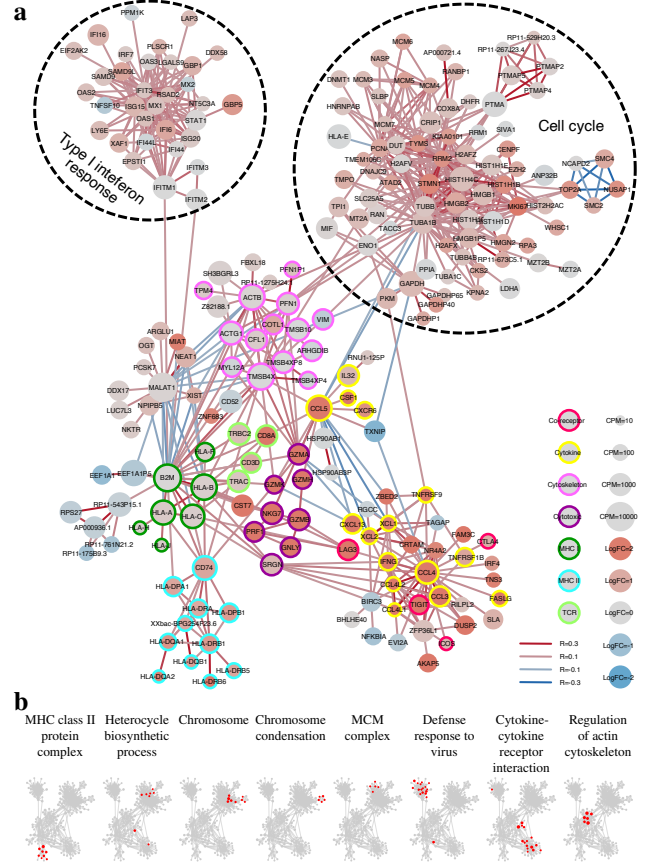
on their accounting for technical confounding, in terms of null P-value distribution of single-cell co-expression. For this, we generated a synthetic null scRNA-seq dataset to mimic the gRNA-free Perturb-seq cells but without any co-expression, using log-normal and multinomial distributions for biological and technical variations respectively (Vieth et al., 2017). Normalisr recovered uniformly distributed P-values irrespective of expression (Fig 2d). Sctransform incurred distortions in null P-value distribution whilst imputation methods MAGIC (Dijk et al., 2018) and DCA (Eraslan et al., 2019) inflated gene-gene associations. Only Normalisr could account for technical confounding and recover the absence of co-expression, a necessity for recovering true expressions and co-expressions.

Normalisr also performed highly in other evaluations. The synthetic dataset recorded pre-measurement biological ground-truths of mRNA proportions before technical variations were introduced. Normalisr obtained optimal Pearson correlations with those true expression levels (Fig 2e). Since a high-quality co-expression ground-truth is absent, we evaluated Normalisr's co-expression robustness through the overlap of transcriptome-wide co-expression networks from each of 10 sequencing batches of Perturb-seq. Normalisr was highly robust in single-cell co-expression inference, with ~50% overlap among the top 1,000 co-expressions (Fig 2f). Normalisr provides a unique normalization framework with optimal sensitivity, specificity, efficiency, and reproducibility for linear association testings of single-cell differential and co-expression.

### 3.2. Normalisr discovered functional gene modules with single-cell DE and co-expression network of dysfunctional T cell in human melanoma

Organisms are evolved to efficiently modulate various functional pathways, partly through the regulation and co-regulation of gene expression. Manifested at the mRNA level, gene co-expression may provide unique insights into cellular and gene functions. However, no existing co-expression detection or network inference method could control for false discovery at single-cell, single-time level.

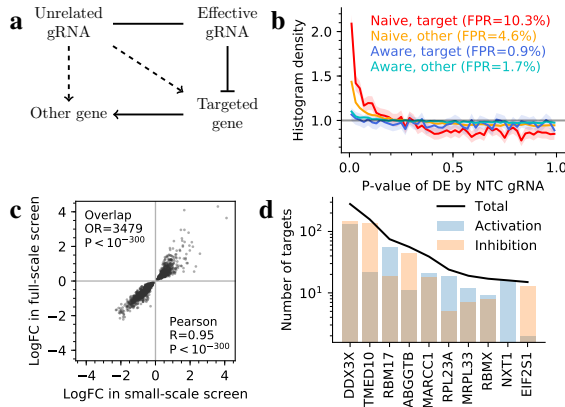
We used Normalisr to infer single-cell transcriptome-wide co-expression networks from dysfunctional T cells in human melanoma MARS-seq (Li et al., 2019). Normalisr recovered biologically meaningful gene co-expression and clustering patterns from cell-to-cell variations (Fig 3a). We observed two distinct gene clusters of cell cycle and type I interferon response. We annotated the remaining, interconnected genes according to their known roles in T cell function. Genes in the same functional category formed obvious regional co-expression clusters. The cytokine program was divided into two negatively associated gene clusters. This suggests a potential functional diversification in the dysfunctional T



**Figure 3. Normalisr revealed gene-level cellular pathways and functional modules in the single-cell co-expression network of dysfunctional T cells in human melanoma.** **a** Single-cell transcriptome-wide co-expression network (major connected component) highlighted functional gene sets for dysfunctional T cells. Edge color indicates positive (red) or negative (blue) co-expression in Pearson R. Node color indicates DE logFC between dysfunctional and naive T cells. Node size indicates average expression level in dysfunctional T cells. Node boundary color indicates functional annotation. **b** Single-cell co-expression recovered significant gene functional similarities, according to GO and KEGG pathway memberships highlighted in red.

cell population, in agreement with recent discoveries in this field.

We then systematically evaluated the functional associations recoverable from single-cell co-expression, based on the over-abundance of co-expression edges between genes in each GO or KEGG annotation. Genes in 33 GO and 8 KEGG pathways had significantly more co-expressions than randomly assigned annotations (Bonferroni  $P \leq 0.05$ ), which encompassed a wide range of cell-type specific and generic functions (Fig 3b). Normalisr recovered gene-level cellular pathways and functional modules in the high-quality single-cell transcriptome-wide co-expression network.



**Figure 4. Normalisr detected robust and specific gene regulations from high-MOI CRISPRi systems.** **a** Example scenario of false positives of gRNA-gene associations (dashed) arising from negative gRNA cross-associations and true regulations (solid), for genes directly or indirectly targeted by any gRNA. **b** Competition-aware method can account for gRNA competitions and reduce the false positives of competition-naïve method, for genes **targeted** by positive control gRNAs at TSS and for **other** genes. Gray line indicates the expected uniform distribution. Shades indicate errors estimated as  $2\sqrt{N} + 1$ . **c** Significant gRNA-gene associations were highly reproducible between small- and full-scale CRISPRi screens, in terms of hypergeometric overlap odds ratio (OR) and P-value, and the logFC Pearson correlation among 1,857 overlaps (each dot). **d** Numbers of inferred targets of top regulators showed distinct activation and inhibition preferences.

### 3.3. Normalisr learned robust and specific gene regulations from high-MOI CRISPRi systems

High multiplicity of infection (MOI) CRISPRi systems are highly efficient DE screens for gene regulations, gene functions (Adamson et al., 2016), and regulatory elements (Gasperini et al., 2019). However, gRNAs may compete for dCas9 and the limited read count, leading to false positive associations between an unrelated gRNA and genes targeted by other gRNAs (Fig 4a). To test such effects, we used a recent and largest-to-date enhancer-screening dataset (Gasperini et al., 2019) beyond the capability of existing DE methods. Its gRNAs were categorized into non-targeting-controls (NTC, as negative controls), transcription start site (TSS) targeting (as positive controls), and candidate enhancers. Using Normalisr, we found that false positive rates (FPR) estimated from NTCs with competition-naïve DE analysis that disregards other, untested gRNAs were much higher than the competition-aware method that accounted for all other gRNAs as additional covariates. This effect was much stronger among genes targeted by any gRNA at their TSS, supporting our hypothesis that false positives were mediated through other, especially TSS-targeting gRNAs (Fig 4b).

To validate the reproducibility of the gRNA-gene associ-

ations discovered by competition-aware method, we performed the same inference on a small-scale screen of the same study. We found major overlap of significant associations between the two screens, with highly correlated logFCs and all effect directions matched (Fig 4c). Normalisr's regulation inference was highly reproducible across CROP-seq screens.

We then inferred gene regulations by searching for TSS-targeting gRNA→targeted gene→trans-gene relationships via gRNA-gene trans-association. To account for off-target effects and mediation through nearby genes, we excluded gRNAs that inhibited another gene within 1Mb from the TSS, and gene regulations irreproducible across gRNAs targeting the same TSS or across screens. In total, we found 833 high-confidence putative gene regulations.

We cross-validated some of the top identified regulators (Fig 4d) with published datasets or literature. An ENCODE bulk RNA-seq of K562 cells with DDX3X knock-down and control shRNAs exhibited a strong agreement in logFC (Pearson  $R=0.38$ ,  $P < 10^{-300}$ , not shown). TMED10, RBM17, RABGGTB exhibited strong activating or inhibitory preferences, and their inferred targets were all enriched with GO categories of their known functions (not shown). Normalisr accurately detected specific, robust, and validated gene regulations in high-MOI CRISPRi screens.

## 4. Discussion

The current explosion of scRNA-seq data generation represents a tremendous opportunity to understand gene regulations at the single-cell level. Here, we described Normalisr, a unique inferential framework and unified solution for single-cell DE, co-expression, and pooled screens, with unparalleled sensitivity, specificity, and speed.

Linear models possess immense capacities and flexibilities to be unleashed on scRNA-seq, such as 'soft' groupings for DE along differentiation trajectories, expression quantitative loci (eQTL) studies, and integrative, distributed analyses of hundreds of millions of cells. At the same sequencing depth with bulk RNA-seq, scRNA-seq additionally partitions the reads between cells and cell types. This additional information provides substantial statistical gain and cell-type stratification that Normalisr as a unified framework can liberate for scRNA-seq, independently or as an emergent superior substitute for bulk RNA-seq.

## Software and Data

Normalisr will be available on github.



## References

- Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., and Weissman, J. S. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.048. URL <http://www.sciencedirect.com/science/article/pii/S0092867416316609>.
- Dijk, D. v., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bieri, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe’er, D. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.05.061. URL [https://www.cell.com/cell/abstract/S0092-8674\(18\)30724-4](https://www.cell.com/cell/abstract/S0092-8674(18)30724-4).
- Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, January 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-07931-2. URL <https://www.nature.com/articles/s41467-018-07931-2>.
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, December 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0844-5. URL <https://doi.org/10.1186/s13059-015-0844-5>.
- Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390.e19, January 2019. ISSN 00928674. doi: 10.1016/j.cell.2018.11.029. URL <https://linkinghub.elsevier.com/retrieve/pii/S009286741831554X>.
- Hafemeister, C. and Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1874-1. URL <https://doi.org/10.1186/s13059-019-1874-1>.
- Li, H., van der Leun, A. M., Yofe, I., Lubling, Y., Gelbard-Solodkin, D., van Akkooi, A. C., van den Braber, M., Rozeman, E. A., Haanen, J. B., Blank, C. U., Horlings, H. M., David, E., Baran, Y., Bercovich, A., Lifshitz, A., Schumacher, T. N., Tanay, A., and Amit, I. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell*, 176(4):775–789.e18, February 2019. ISSN 00928674. doi: 10.1016/j.cell.2018.11.043. URL <https://linkinghub.elsevier.com/retrieve/pii/S009286741831568X>.
- Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korb, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schönhuth, A. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1926-6. URL <https://doi.org/10.1186/s13059-020-1926-6>.
- Soneson, C. and Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, April 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4612. URL <https://www.nature.com/articles/nmeth.4612>.
- Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, February 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-019-0379-5. URL <http://www.nature.com/articles/s41587-019-0379-5>.
- Vieth, B., Ziegenhain, C., Parekh, S., Enard, W., and Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, November 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx435. URL <https://academic.oup.com/bioinformatics/article/33/21/3486/3952669>. Publisher: Oxford Academic.