Agarwal, Divyansh; **Wang, Jingshu**; **Zhang, Nancy R.**

Data denoising and post-denoising corrections in single cell RNA sequencing. (English)

Zbl 07292499
Stat. Sci. 35, No. 1, 112-128 (2020).

Summary: Single cell sequencing technologies are transforming biomedical research. However, due to the inherent nature of the data, single cell RNA sequencing analysis poses new computational and statistical challenges. We begin with a survey of a selection of topics in this field, with a gentle introduction to the biology and a more detailed exploration of the technical noise. We consider in detail the problem of single cell data denoising, sometimes referred to as "imputation" in the relevant literature. We discuss why this is not a typical statistical imputation problem, and review current approaches to this problem. We then explore why the use of denoised values in downstream analyses invites novel statistical insights, and how denoising uncertainty should be accounted for to yield valid statistical inference. The utilization of denoised or imputed matrices in statistical inference is not unique to single cell genomics, and arises in many other fields. We describe the challenges in this type of analysis, discuss some preliminary solutions, and highlight unresolved issues.

MSC:

62 Statistics

Keywords:

single cell biology; RNA sequencing; imputation; post-denoising inference; empirical Bayes; deep learning

References:

[1] Andrews, T. S. and Hemberg, M. (2018). False signals induced by single-cell imputation. F1000Res 7.

[2] Agarwal, D., Wang, J. and Zhang, N. R (2020). Supplement to "Data Denoising and Post-Denoising orrections in Single Cell RNA Sequencing." https://doi.org/10.1214/19-STS7560SUPP.

[3] Arkin, A., Ross, J. and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage \(\lambda \)-infected Escherichia coli cells. Genetics 149 1633-1648.

[4] Badsha, M. B., Li, R., Liu, B., Li, Y. I., Xian, M., Banovich, N. E. and Fu, A. Q. (2018). Imputation of single-cell gene expression with an autoencoder neural network. BioRxiv 504977.

[5] Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S. et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Systems 3 346-360.

[6] Barroso, G. V., Puzovic, N. and Dutheil, J. Y. (2018). The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. Genetics 208 173-189.

[7] Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10 1093-1095.

[8] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36 411-420.

[9] Chen, M. and Zhou, X. (2018). VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol. 19 196.

[10] Chen, R., Wu, X., Jiang, L. and Zhang, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. Cell Reports 18 3227-3241.

[11] Chiron, L., van Agthoven, M. A., Kieffer, B., Rolando, C. and Delsuc, M.-A. (2014). Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry. Proc. Natl. Acad. Sci. USA 111 1385-1390.

[12] Clevers, H., Rafelski, S. and Elowitz, M. et al. (2017). What is your conceptual definition of 'cell type' in the context of a mature organism? Cell Systems 4 255-259.

[13] Degrelle, S. A., Hennequet-Antier, C., Chiapello, H., Piot-Kaminski, K., Piumi, F., Robin, S., Renard, J.-P. and Hue, I. (2008). Amplification biases: Possible differences among deviating gene expressions. BMC Genomics 9 46.

[14] Di Gregorio, A., Bowling, S. and Rodriguez, T. A. (2016). Cell competition and its role in the regulation of cell fitness from development to cancer. Developmental Cell 38 621-634.

[15] Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M. and Coleman, P. (1992). Analysis of gene expression in single live neurons. Proc. Natl. Acad. Sci. USA 89 3010-3014.

[16] Eldar, A. and Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. Nature 467 167-173.

[17] Elowitz, M. B., Levine, A. J., Siggia, E. D. and Swain, P. S. (2002). Stochastic gene expression in a single cell. Science 297 1183-1186.

[18] Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K. and Quake, S. R. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. Cell 171 321-330.

[19] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. 10 390.

[20] Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. and Garry, D. J. (2018). DrImpute: Imputing dropout events in single cell RNA sequencing data. BMC Bioinform. 19 220.

[21] Gossett, D. R., Henry, T., Lee, S. A., Ying, Y., Lindgren, A. G., Yang, O. O., Rao, J., Clark, A. T. and Di Carlo, D. (2012). Hydrodynamic stretching of single cells for large population mechanical phenotyping. Proc. Natl. Acad. Sci. USA 109 7630-7635.

[22] Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. BioRxiv 576827.

[23] Haghverdi, L., Lun, A. T. L., Morgan, M. D. and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36 421-427.

[24] Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H. et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell 172 1091-1107.

[25] Hedlund, E. and Deng, Q. (2018). Single-cell RNA sequencing: Technical advancements and biological applications. Mol. Aspects Med. 59 36-46.

[26] Hicks, S. C., Townes, F. W., Teng, M. and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19 562-578.

[27] Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L. and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6 211-226. · Zbl 1071.62104

[28] Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. Nat. Methods 15 539. · Zbl 1416.62625

[29] Hwang, B., Lee, J. H. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental \& Molecular Medicine 50 1-14.

[30] Islam, S., Zeisel, A., Joost, S., Manno, G. L., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11 163-166.

[31] Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Illicic, T., Teichmann, S. A. and Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. 6 8687.

[32] Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H. and Yang, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. Brief. Bioinform. 20 2316-2326.

[33] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161 1187-1201.

[34] Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. Molecular Cell 58 610-620.

[35] La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R., Toledo, E. M. et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. Cell 167 566-580.

[36] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P. et al. (2018). RNA velocity of single cells. Nature 560 494-498.

[37] Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat. Commun. 9 997.

[38] Linderman, G. C., Zhao, J. and Kluger, Y. (2018). Zero-preserving imputation of scRNA-seq data using low-rank approximation. BioRxiv 397588.

[39] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods 15 1053-1058.

[40] Losick, R. and Desplan, C. (2008). Stochasticity and cell fate. Science 320 65-68.

[41] Martinez-Jimenez, C. P., Eling, N., Chen, H.-C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J. T. et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science 355 1433-1436.

[42] McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA 94 814-819.

[43] Novick, A. and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. Proc. Natl. Acad. Sci. USA 43 553-566.

[44] Papalexi, E. and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev., Immunol. 18 35-45.

[45] Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. Sci. Rep. 6 25533.

[46] Park, J., Shrestha, R., Qiu, C., Kondo, A., Huang, S., Werth, M., Li, M., Barasch, J. and Suszták, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science 360 758-763.

[47] Pearson, K. (1982). The Grammer of Science. Cambridge Univ. Press, Cambridge.

[48] Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. Cell 135 216-226.

[49] Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P. et al. (2017). Science forum: The human cell atlas. eLife 6 e27041.

[50] Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A. and Teichmann, S. A. (2017). The human cell atlas: From vision to reality. Nature News 550 451.

[51] Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. 37 547-554.

[52] Skinnider, M. A., Squair, J. W. and Foster, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. Nat. Methods 16 381-386.

[53] Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods 15 255-261.

[54] Song, R., Sarnoski, E. A. and Acar, M. (2018). The systems biology of single-cell aging. IScience 7 154-169.

[55] Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. 20 257-272.

[56] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. et al. (2019). Comprehensive integration of single-cell data. Cell 177 1888-1902.

[57] Svensson, V., Vento-Tormo, R. and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 13 599-604.

[58] Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A. and Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods 14 381-387.

[59] The Tabula Muris Consortium (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562 367.

[60] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6 377.

[61] Teschendorff, A. E. and Enver, T. (2017). Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat. Commun. 8 15599.

[62] Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A. et al. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat. Methods 16 479-487.

[63] Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. 25 1491-1498.

[64] Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K. and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Sci. Rep. 7 39921.

[65] Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L. et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell 174 716-729.

[66] Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D. and Eberwine, J. H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc. Natl. Acad. Sci. USA 87 1663-1667.

[67] Wagner, F., Yan, Y. and Yanai, I. (2017). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. BioRxiv 217737.

[68] Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M. and Zhang, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. Proc. Natl. Acad. Sci. USA 115 E6437-E6446. · Zbl 1416.62625

[69] Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C. and Zhang, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. Nat. Methods 16 875-878.

[70] Zappia, L., Phipson, B. and Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol. 14 e1006245.

[71] Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347 1138-1142.

[72] Zhang, L. and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. IEEE/ACM Trans. Comput. Biol. Bioinform.

[73] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8 14049.

[74] Ziegenhain, C.