Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Statistical single cell multi-omics integration

M. Colomé-Tatché[1,2,3] and F. J. Theis[1,4]

Abstract

Single cell high throughput genomic measurements are revolutionizing the fields of biology and medicine, providing a means to tackle biological problems that have thus far been inaccessible, such as the systematic discovery of new cell types, the identification of cellular heterogeneity in health and disease, or the cell-fate decisions taking place during differentiation and reprogramming. Recently implemented multi–omics measurements of genomes, transcriptomes, epigenomes, proteomes and chromatin organization are opening up new avenues to begin to disentangle the causal relationship between -omics layers and how these co-determine higher-order cellular phenotypes. This technological revolution is not restricted to basic science but promises major breakthroughs in medical diagnostics and treatments. In this paper we review existing computational methods for the analysis and integration of different -omics layers and discuss what new approaches are needed to leverage the full potential of single cell multi-omics data.

Addresses

[1] Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany

[2] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands

[3] TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

[4] Department of Mathematics, Technical University of Munich, Garching, Germany

Corresponding authors: Theis, F.J. (fabian.theis@helmholtz-muenchen.de); Colomé-Tatché, M. (maria.colome@helmholtz-muenchen.de)

Current Opinion in Systems Biology 2018, **7**:54–59

This review comes from a themed issue on **Future of systems biology (2018)**

Edited by **Marija Cvijovic** and **Stefan Hohmann**

For a complete overview see the Issue and the Editorial

Available online 2 February 2018

https://doi.org/10.1016/j.coisb.2018.01.003

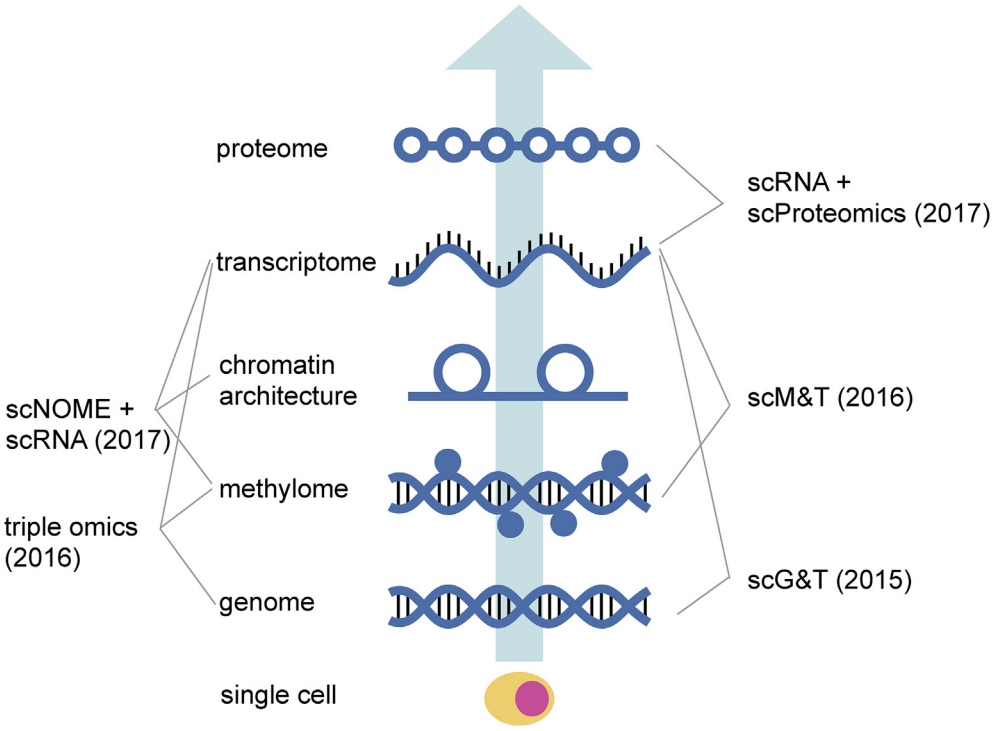
Introduction

Single-cell biology is revolutionizing biomedical research with the promise to profile the basic entities of life at an unprecedented level of resolution. Single cell genomics has had a major impact in fields ranging from developmental biology to immunology and tumor biology [1]. It is increasingly realized that when dealing with inherently heterogeneous cellular populations, such as in stem cell research or in many diseases such as cancer, the corresponding molecular markers have to be scrutinized on a single cell level, instead of at the average level in the cellular population. Promising international collaborations such as the Human Cell Atlas [2] are shaping up to address these questions on a large-scale level.

Currently cell identity has been mostly considered on the level of RNA via a host of single cell RNA-sequencing (scRNA-seq) approaches. Common analysis approaches typically proceed from initial preprocessing and quality control to dimension reduction for visualization and clustering to identify cell types, sometimes followed by lineage inference and/or differential comparison across conditions [3]. Although many methods initially proposed for bulk cell populations can be used for such analyses, it became clear early on that due to different noise properties and data sparseness, amongst others, new computational strategies would have to be taken to fully make use of the richness of the data [4]. This has been pointed out for scRNA-seq but increasingly becomes clear also for other single cell -omics techniques that have subsequently been developed, and in particular for the integration of the different -omics datasets.

Since 2009, when the first scRNA-seq was performed [5], other single cell sequencing techniques have been developed. In 2011 Navin et al. [6] reported the first genome sequencing of a single human cell, while in 2012 Falconer et al. [7] demonstrated the first single-cell single-strand genome sequencing. More recently, several other single cell measurements have been introduced, such as single cell DNA methylation [8], single cell chromatin immunoprecipitation for assessing histone modification status [9], single cell open chromatin (scATAC-seq and scDNase-seq [10–12]), single cell chromosomal conformation [13] and single cell chromosome–lamina interactions [14], together with single cell proteomics [15] and metabolomics [16]. These methods complement single cell transcriptomic measurements by providing an extra layer of regulatory information. However, true regulatory inferences can only be made if the different -omics layers are measured on the same single cell. Such combined measurements are now becoming available (Figure 1). In particular, one can currently combine measurements of transcriptome and genome [17,18], as well as transcriptome and methylome [19–22], which in turn provides information about genomic gains and losses [21]. The combination

Figure 1



Scheme of the different experimentally available single cell multi-omics measurements.

of transcriptome, nucleosome positioning and DNA methylation has also been obtained thanks to the single cell NOME-seq techniques [23—25]. Finally, transcriptome and proteome have also been measured for the same cell [26,27].

These above-described multi-omics measurements constitute a major breakthrough in the life sciences. They enable — for the first time - systematic studies of the relationship between genome, epigenome, transcriptome and proteome, and thus allow us to investigate various questions of outstanding biological and medical relevance that have thus far been inaccessible with other technologies. Among these questions is how molecular states lead to different cell fates in development or disease, what mechanisms govern transient cell responses, or what role cellular heterogeneity plays in the onset and progression of diseases such as cancer, diabetes or Alzheimer's. These developments are resulting in the generation of big data in the life sciences, as it is projected that the data generated from single cell sequencing experiments will reach 1 zetta-bases = 10^6 peta-bases/year in 2025, the same range as data acquisition in astronomy [28]. A key challenge in the analysis of this data is to device efficient computational tools to process, integrate and characterize multiple functional measurements in a biologically meaningful manner, and overcome the extensive amount of missing data inherent in single cell sequencing experiments.

Current multi-omics integration approaches in bulk

Analysis approaches to single cell multi-omics data have the goal to infer regulatory relationships between the multiple -omics layers, and to describe the unique cellular states in more detail. From a statistical perspective, the task of integrating multiple -omics levels is also known as multi-view learning [29], and includes methods ranging from kernel learning and Bayesian modeling to matrix factorization and multimodal deep learning [30]. Integration of multiple -omics levels has been discussed in detail on bulk-averaged cell populations, resulting in analyses integrating genotypes, DNA methylation, histone modifications, RNA expression and splicing, as well as protein expression [31,32]. Methodologically, we can roughly outline three strategies: (i) *statistical integration*: statistical approaches that explore common variation across species, essentially using multi-view machine learning, resulting in regulatory networks with multiple node types of graphical models; (ii) *QTL-based analysis*: integration of 'downstream' -omics levels by anchoring them by genetic variation, thereby defining e.g. expression QTLs, methylation QTLs as well as histone modification QTLs and overlaps thereof; (iii) *mechanism-based integration*: mechanism-based integration strategies that link -omics levels on individual units using approximations to assumed regulation levels, e.g. mRNA and

protein using models for transcriptional and/or post-transcriptional regulation. Examples of these strategies outlined above are the construction of a correlation network of both transcripts and metabolites in blood to study transcriptional regulation of metabolism and its link to clinical markers [33] (for statistical integration); the analysis of the genetic drivers of both epigenetic and transcriptional variation in human immune cells, from the International Human Epigenome Consortium [34] (for QTL-based methods); and the linking of multiple -omics levels (such as mRNA and protein or mRNA and microRNA) using a model-based mechanistic Bayesian approach, with subsequent joint enrichment for ontologies [35] (for mechanism-based integration).

Current multi-omics integration approaches in single cells

For the analysis and integration of single cell -omics layers measured in single cells, existing approaches have so far considered either computing correlations between -omics layers, or obtaining separate single cell maps for every measured -omics followed by integration (as outlined in Ref. [36]).

At the genomic level, correlations between copy number variation and transcription levels have been studied to delineate the effect of genomic duplications and deletions on gene expression, and to study the effect of ploidy on the cell-to-cell transcriptional variability [17,18,37]. From a regulatory point of view, correlations between DNA methylation and transcription levels (both at the promoter and at the gene body) have been computed to investigate the direct role of DNA methylation on transcription and to study the effect of heterogeneous DNA methylation in the cell population on gene expression [19,21,22]. Since single cell (reduced representation) bisulfite sequencing for measuring DNA methylation gives access to genomic information as well, it has also been possible to correlate both DNA methylation and gene expression to genomic gains and losses at the same time [21].

Similar strategies have been implemented for other chromatin signatures, such as nucleosome positioning in combination with DNA methylation and gene expression, to determine the role of chromatin architecture on expression and to study the DNA methylation levels of loci displaying highly heterogeneous chromatin signatures [23–25]. Finally, for the combined measurement of transcriptome and proteome, correlations have been calculated between gene expression and protein abundance for a few genes [26], a first result towards the study of post-transcriptional regulation.

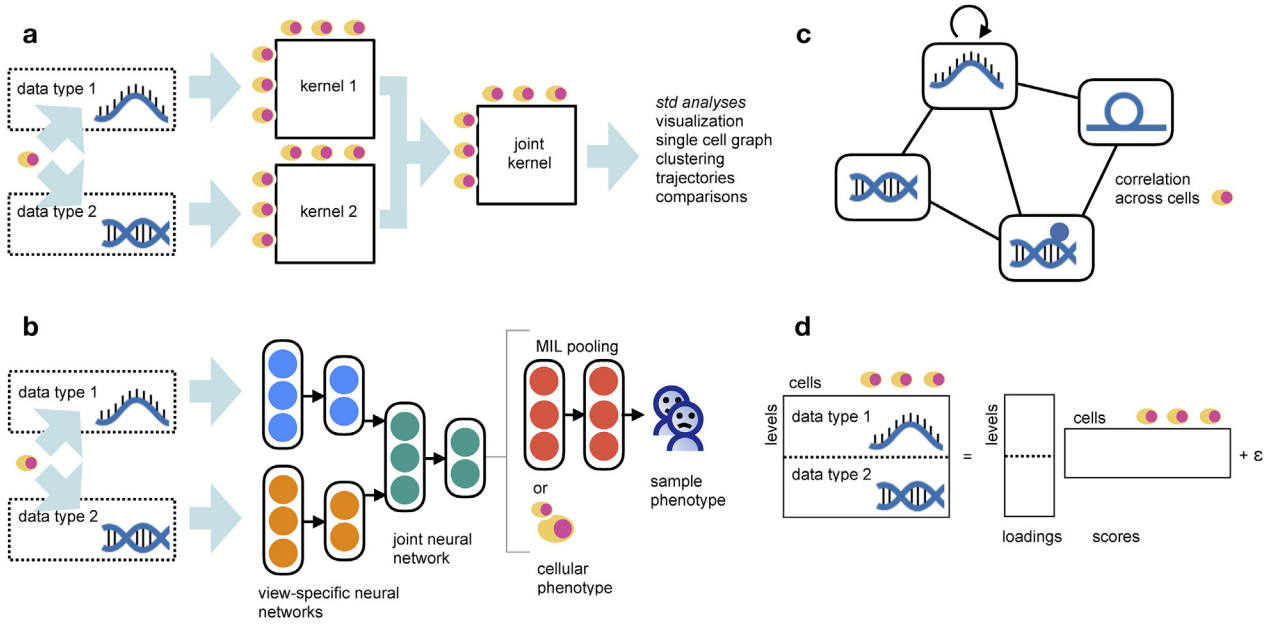
The second most common analysis approach has been to construct a separate single cell map for every -omics layer, followed by dimension reduction and clustering. The independent clustering outcomes for the different -omics layers are subsequently combined or compared [19–21,23,24,26,27]. Such an approach provides information on what -omic layer is most efficient at clustering cell types, and infers how much information is shared between layers by comparing clustering outputs. In the analysis of single cell transcriptomics and proteomics instead a common approach has been to cluster cells based on their genome-wide transcriptional profile and to then explore protein levels in the different clusters [26,27].

Finally, other analysis strategies have been devised for the integration of multiple -omics layers measured in different cells, and not the same single cell, sampled from the same cell population. A method called MATCHER [38] has been developed to perform manifold alignment of transcriptomic and epigenomic measurements from different cells. The method uses first a Gaussian process latent variable model to obtain pseudotime values for every cell in every -omic layer, and then aligns the quantiles of the pseudotime distribution to match the quantiles of a uniform distribution, making them directly comparable. Another approach is the one from Butler and Satija [39], who have devised a strategy for performing integrated or comparative analysis of scRNA-seq datasets produced across different platforms, conditions or species. The method uses diagonal canonical correlation analysis (CCA) to learn a shared gene correlation structure which is conserved between the different RNA-seq datasets, and aligns the datasets into a conserved low-dimensional space. CCA and other matrix factorization techniques have been used in bulk data analysis for the integration of multiple -omics layers [40–42], and therefore their single cell extensions could be used to incorporate other sources of variation such as DNA methylation or proteomic measurements (Figure 2d), where the resulting loadings would indicate cell and view specific contributions to the overall data variation.

Outlook

Instead of the above-described methods, which treat the -omics layers separately, we argue that an ideal approach would be to construct single cell maps based on a joint kernel that incorporates all measured -omic layers (Figure 2a), similar to the above described statistical approaches in bulk. The goal would be to construct a multi-space similarity measure, i.e. a real-valued function that takes as inputs the n -omics measurements for every pair of cells and outputs a single similarity value between them, based on a combination of the n -omics layers (Figure 2a). Learning joint kernel or similarity matrices across the -omics levels allows standard analyses on the integrated cell–cell distances, such as t-SNE based visualization [43], pseudotime estimation [44], or clustering [45,46]. The underlying

Figure 2



General strategies for statistical integration of mutli-omics single cell measurements beyond only result integration [36]. (a) Multi-view kernel learning to allow standard downstream analyses. (b) Multi-view classification via neural networks, with multi-instance learning. (c) Network estimation. (d) Multi-view matrix factorization.

multi-omics single cell maps will increase the power for detecting differences between single cells and, especially in the cases where the independent single -omic analysis generate different results, will help identify intermediate populations and will help determine the contribution of the different regulatory layers to the cellular identity.

Instead of distance learning, one could reformulate questions to multi-omics observations as classification tasks. Due to the added single cell resolution, we may have labels on cellular level but often only per sample or condition (such as control vs knockout). In machine learning this question is called multi-instance learning [47], and may be combined with multi-view learning. For simplicity, we would suggest to approach this by training view-specific classifiers such as neural networks on each view level first, and then integrating them via a joint network. This may be combined with a feature selection stage, determining which -omics level contributes most and least to the classification performance, and potentially include interaction terms as well. Similarly, it would be interesting to ask how much expression values in each -omics level help classification versus using quantifiers of heterogeneity e.g. within a cell-type or cluster. Depending on the classification task, one may follow this up with a cell specific phenotype or reformulate the problem as a multi-view learning task (Figure 2b). The latter would help to extend the above question of how much heterogeneity within a cell ('bag' in multi-instance learning) contributes to the classification, and how much so for each -omics level.

Besides prediction, correlation based ideas can be extended to multiple views by determining how specific measurements jointly change across the cells. This corresponds to the mechanism-based integration approach in the bulk situation. One would simply ask if a particular value within an -omics level is correlated to another value either within the same layer or within an adjacent/known to be influenced one; an example would be to determine if the expression of a transcription factor is correlated to the expression of its targets, similarly for a microRNA and its targets, or a methylated region in a particular gene's promotor with the expression of that gene. Beyond correlation, other dependency measures such as mutual information may be used for the same questions, or if sample size and noise permits, even causal dependency indices (e.g. within a pseudotime). These dependencies can be visualized as in bulk using networks (Figure 2c), allowing for analysis of jointly regulated hubs or dense network regions.

Finally, QTL based analysis and mechanism-based integration can also be adapted to single cell multi-omics data integration. For QTL based analysis, genetic variation can be linked to population single cell metrics such as transcription heterogeneity, as has been done for scRNAseq only in a pilot study in human PBMCs using multiplexing by genetic barcodes [48].

For the mechanism-based integration, mechanistic models can be explored, linking the -omics levels in the population of individual single cells like has been reviewed above (section Current multi-omics integration approaches in single cells), where the links between -omics layers are not obscured by the population average measurements like in bulk.

Single cell RNA-seq approaches are rapidly scaling up to truly large sample numbers [49], going into the millions [50] and promising even billions of transcriptomes in the near future [51]. With other single cell -omics catching up, the field has started to develop more sophisticated multi-omics statistical models for data integration, while addressing related computational issues using efficient implementations, distributed computing and even distributed data storage [52]. Many of the outlined integration algorithms are significantly more complex than the univariate analyses and are hence more difficult to upscale, but promise to exploit the full potential of single cell multi-omics sequencing techniques.

Acknowledgments

M.C.T. acknowledges support from the Impuls-und Vernetzungsfonds of the Helmholtz-Gemeinschaft (grant VH-NG-1219). F.J.T. acknowledges financial support by the German Science Foundation (SFB 1243 and Graduate School QBM) as well as the Federal Ministry of Education and Research (Single Cell Genomics Network Germany).

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Linnarsson S, Teichmann SA: **Single-cell genomics: coming of age**. *Genome Biol* 2016, **17**:97.

2. ** Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M., *et al.*: Human Cell Atlas Meeting Participants: **Science forum: the human cell Atlas**. *Elife* 2017, **6**, https://doi.org/10.7554/eLife.27041.
Description of the idea as well as the potential of the Human Cell Atlas Project, aiming to build an open molecular reference map of cell states in healthy human tissues.

3. Sandberg R: **Entering the era of single-cell transcriptomics in biology and medicine**. *Nat Meth* 2014, **11**:22–24.

4. * Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in single-cell transcriptomics**. *Nat Rev Genet* 2015, **16**:133–145.
Systematic review of the main challenges in the analysis of single cell transcriptomics data.

5. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Meth* 2009, **6**:377–382.

6. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, *et al.*: **Tumour evolution inferred by single-cell sequencing**. *Nature* 2011, **472**:90–94.

7. Falconer E, Hills M, Naumann U, Poon SSS, Chavez EA, Sanders AD, Zhao Y, Hirst M, Lansdorp PM: **DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution**. *Nat Meth* 2012, **9**:1107–1112.

8. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G: **Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity**. *Nat Meth* 2014, **11**:817–820.

9. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE: **Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state**. *Nat Biotechnol* 2015, **33**:1165–1172.

10. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals principles of regulatory variation**. *Nature* 2015, **523**:486–490.

11. Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R, Levens D, Zhao K: **Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples**. *Nature* 2015, **528**:142–146.

12. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J: **Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing**. *Science* 2015, **348**:910–914.

13. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure**. *Nature* 2013, **502**:59–64.

14. Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, Zhan Y, Lajoie B, de Graaf CA, Amendola M, *et al.*: **Genome-wide maps of nuclear lamina interactions in single human cells**. *Cell* 2015, **163**:134–147.

15. Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, Gherardini PF: **Highly multiplexed simultaneous detection of RNAs and proteins in single cells**. *Nat Meth* 2016, **13**:269–275.

16. Fessenden M: **Metabolomics: small molecules, single cells**. *Nature* 2016, **540**:153–155.

17. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, *et al.*: **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes**. *Nat Meth* 2015, **12**:519–522.

18. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A: **Integrated genome and transcriptome sequencing of the same cell**. *Nat Biotechnol* 2015, **33**:285–289.

19. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T: **Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity**. *Nat Meth* 2016, **13**:229–232.

20. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, Tan DSW, Robson P, Loh Y-H, Quake SR, Burkholder WF: **Single-cell multimodal profiling reveals cellular epigenetic heterogeneity**. *Nat Meth* 2016, **13**:833–836.

21. ** Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, Peng J: **Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas**. *Cell Res* 2016, **26**:304–319.
One of a few papers describing the measurement and integration of more than two -omics layers of the same single cell, in particular genomics, epigenomics and transcriptomics.

22. Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang C-Y, Xue Z, Fan G: **Simultaneous profiling of transcriptome and DNA methylome from a single cell**. *Genome Biol* 2016, **17**:88.

23. Pott S: **Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells**. *Elife* 2017, **6**, https://doi.org/10.7554/eLife.23203.

24. Guo F, Li L, Li J, Wu X, Hu B, Zhu P, Wen L, Tang F: **Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells**. *Cell Res* 2017, **27**:967–988.

25. ** Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W: **Joint profiling of chromatin accessibility, DNA methylation and transcription in single cells**. *bioRxiv* 2017: 138685, https://doi.org/10.1101/138685.

One of a few papers profiling and integrating more than two -omic layers of the same single cell, combining NOME-seq (chromatin accessibility and DNA methylation) with scRNA-seq.

26. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA: **Multiplexed quantification of proteins and transcripts in single cells**. *Nat Biotechnol* 2017, **35**:936–939.

27. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P: **Simultaneous epitope and transcriptome measurement in single cells**. *Nat Meth* 2017, **14**:865–868.

28. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE: **Big data: astronomical or genomical?** *PLoS Biol* 2015, **13**:e1002195.

29. ** Sun S: **A survey of multi-view machine learning**. *Neural Comput Appl* 2013, **23**:2031–2038.

A comprehensive review and categorization of machine learning methods dealing with the integration of multi-modal ("multi-view") data sets, i.e. samples with multiple alternative multivariate observations.

30. Li Y, Wu F-X, Ngom A: **A review on machine learning principles for multi-view biological data integration**. *Brief Bioinform* 2016, https://doi.org/10.1093/bib/bbw113.

31. Huang S, Chaudhary K, Garmire LX: **More is better: recent progress in multi-omics data integration methods**. *Front Genet* 2017, **8**:84.

32. Civelek M, Lusis AJ: **Systems genetics approaches to understand complex traits**. *Nat Rev Genet* 2014, **15**:34–48.

33. Bartel J, Krumsiek J, Schramm K, Adamski J, Gieger C, Herder C, Carstensen M, Peters A, Rathmann W, Roden M, *et al.*: **The human blood metabolome-transcriptome interface**. *PLoS Genet* 2015, **11**:e1005274.

34. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yan Y, Kundu K, Ecker S, *et al.*: **Genetic drivers of epigenetic and transcriptional variation in human immune cells**. *Cell* 2016, **167**:1398–1414. e24.

35. Sass S, Buettner F, Mueller NS, Theis FJ: **A modular framework for gene set analysis integrating multilevel omics data,**. *Nucleic Acids Res* 2013, **41**:9622–9633.

36. Bock C, Farlik M, Sheffield NC: **Multi-omics of single cells: strategies and applications**. *Trends Biotechnol* 2016, **34**:605–608.

37. Han KY, Kim K-T, Joung J-G, Son D-S, Kim YJ, Jo A, Jeon H-J, Moon H-S, Yoo CE, Chung W, *et al.*: **SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells**. *Genome Res* 2017, https://doi.org/10.1101/gr.223263.117.

38. * Welch JD, Hartemink AJ, Prins JF: **MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics**. *Genome Biol* 2017, **18**:138.

Description of a computational method to align multiple single cell omics layers, albeit assuming they have been measured in *different* single cells.

39. * Butler A, Satija R: **Integrated analysis of single cell transcriptomic data across conditions, technologies, and species**. *bioRxiv* 2017:164889, https://doi.org/10.1101/164889.

Computational integration of single cell transcriptomic measurements performed on different populations of single cells

40. Witten DM, Tibshirani R, Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009, **10**:515–534.

41. Lê Cao K-A, Martin PGP, Robert-Granié C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinf* 2009, **10**:34.

42. Waaijenborg S, Verselewel de Witt Hamer PC, Zwinderman AH: **Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis**. *Stat Appl Genet Mol Biol* 2008, **7**. Article3.

43. van der Maaten L, Hinton G: **Visualizing data using t-SNE**. *J Mach Learn Res* 2008, **9**:2579–2605.

44. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ: **Diffusion pseudotime robustly reconstructs lineage branching**. *Nat Meth* 2016, **13**:845–848.

45. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks**. *J Stat Mech* 2008, **2008**:P10008.

46. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, *et al.*: **Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis**. *Cell* 2015, **162**:184–197.

47. Amores J: **Multiple instance classification: review, taxonomy and comparative study**. *Artif Intell* 2013, **201**:81–105.

48. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, Wan E, Wong S, Byrnes L, Lanata C, Gate R, *et al.*: **Multiplexing droplet-based single cell RNA-sequencing using natural genetic barcodes**. *bioRxiv* 2017:118778, https://doi.org/10.1101/118778.

49. * Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ: **Single cells make big data: new challenges and opportunities in transcriptomics**. *Curr Opin Struct Biol* 2017, **4**:85–91.

Review of computational challenges arising from the increasing size of scRNAseq data sets, highlighting opportunities and challenges in the context of big data analytics.

50. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, *et al.*: **Massively parallel digital transcriptional profiling of single cells**. *Nat Commun* 2017, **8**:14049.

51. Rosenberg AB, Roco C, Muscat RA, Kuchina A, Mukherjee S, Chen W, Peeler DJ, Yao Z, Tasic B, Sellers DL, Pun SH, Seelig G: **Scaling single cell transcriptomics through split pool barcoding**. *bioRxiv* 2017:105163, https://doi.org/10.1101/105163.

52. Hon C-C, Shin JW, Carninci P, Stubbington MJT: **The human cell Atlas: technical approaches and challenges**. *Brief Funct Genomics* 2017, https://doi.org/10.1093/bfgp/elx029.