Cell

Leading Edge
Minireview

Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology

Jan Philipp Junker[1,*] and Alexander van Oudenaarden[1]
[1]Hubrecht Institute, KNAW and University Medical Center Utrecht, 3584 CT Utrecht, the Netherlands
*Correspondence: j.junker@hubrecht.eu
http://dx.doi.org/10.1016/j.cell.2014.02.010

Single-cell analyses have provided invaluable insights into studying heterogenity, signaling, and stochastic gene expression. Recent technological advances now open the door to genome-wide single-cell studies.

Introduction

From populations of unicellular organisms to complex tissues, cell-to-cell variability in genotypic and/or phenotypic traits seems to be universal. To study heterogeneity and its biological consequences, researchers have used low-throughput approaches—such as fluorescent reporters and fluorescence in situ hybridization (FISH) techniques—that allow quantification of a limited number of parameters in single cells. On the other hand, genomic technologies allow for high-throughput approaches and are now used in laboratories around the world. However, genomic studies have hitherto relied on studying ensemble averages obtained from pooling thousands to millions of cells, precluding genome-wide analysis of cell-to-cell variability. Improvements in sequencing technology and molecular biology are now leading to the emergence of genome-wide quantitative analysis of single cells and, hence, to the convergence of genomics and single-cell biology (Figure 1). We discuss the importance of studying cell-to-cell variability using large-scale and genome-wide techniques, while highlighting technological breakthroughs and new frontiers in single-cell biology. Variation between individual cells can originate from genetic differences, developmental and functional states, and environmental cues. Even in cells with an identical genome, fluctuations of regulator molecules and stochastic gene expression can cause significant deviation of individual cells from the population average. We also describe methods and selected applications for studying single-cell heterogeneity on these different levels.

Genomic Differences as a Source of Cellular Heterogeneity

Genomic differences are arguably the most fundamental source of cellular variability. Sequencing genomic DNA from single cells has been a challenge. In cases in which DNA yield from samples is limiting, such as in single cells, whole-genome amplification methods are generally important for ensuring that the starting material is sufficient for sequencing. However, amplification bias can lead to suboptimal genome coverage. Multiple displacement amplification (MDA) reduces amplification bias by annealing of random hexamers to denatured DNA, followed by isothermal strand-displacement synthesis of DNA products (Dean et al., 2002). Multiple annealing and looping-based amplification cycles (MALBAC) (Zong et al., 2012) is a novel protocol for whole-genome amplification that might potentially reduce bias even further by ensuring that amplification products cannot serve as new templates for amplification, thereby achieving quasilinear amplification. But what is the benefit of sequencing DNA from single cells? We briefly discuss several fascinating applications.

Recently, Rinke et al. (2013) obtained genomic information using MDA from 3,300 single, uncultured microbial cells from different habitats. They performed 16S rRNA analyses on a subset of them and were able to discover novel phylogenetic relationships and unexpected metabolic capabilities, possibly caused by ancient lateral gene transfer events between eukaryotes and archaea.

Tumors are composed of genetically heterogeneous cell types. Using a new approach for whole-nucleus isolation and sequencing, Navin et al. (2011) were able to determine copy number variations of individual cells extracted from breast cancer samples and identify distinct cell populations likely corresponding to separate waves of clonal expansion. Importantly, by macrodissecting the tumor into different sectors prior to cell dissociation, the authors were able to retain spatial information and measure how this clonal architecture varied in space.

Single-cell genomics was also applied in the study of homologous recombination in single germ cells. Homologous recombination creates unique hybrid haploid genomes in individual germ cells by shuffling genetic information between homologous chromosomes. Using microfluidics for parallelized whole-genome amplification of single sperm cells, Wang et al. (2012) mapped recombination sites in 91 human sperm cells. On a coarse-grained level, the authors found similar patterns of recombination site distributions as in previous population-level analyses of pedigree data. More detailed analysis, however, revealed individual-specific recombination hot spots and allowed direct measurement of the germline de novo mutation rate.

In a study on somatic mosaicism, McConnell et al. (2013) investigated copy number variation in neurons obtained from human induced pluripotent stem cells and postmortem human brains at the single-cell level. The ability to detect rare copy

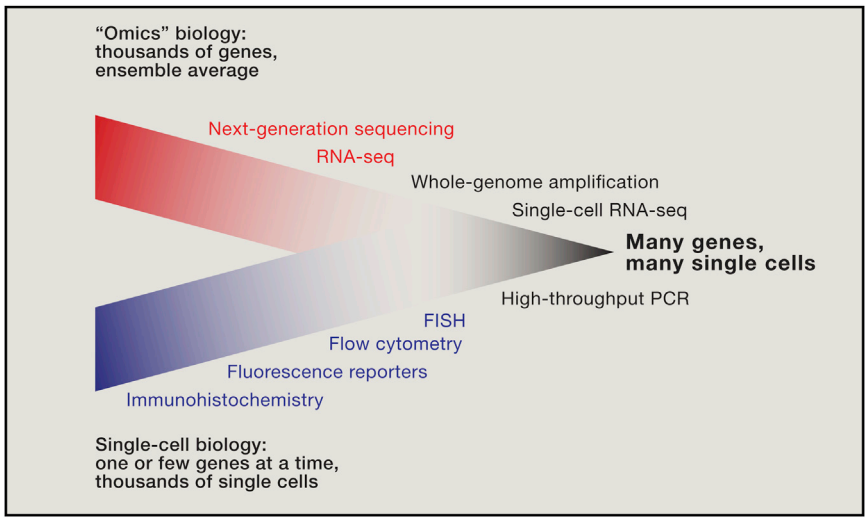

Figure 1. Convergence of "Omics" Biology and Single-Cell Biology
Technology that allows researchers to obtain genome-wide information from single cells is extending the boundaries of a field that has thus far been limited to the analyses of a select gene in eukaryotes.

number variations is an important advantage of single-cell analysis compared to previous bulk experiments. The authors found that neurons have significantly higher copy number variation than other cell types, with deletions being more frequent than duplications. It is interesting to speculate that copy number variation in neurons might play a role in functional diversification. However, the physiological consequences of mosaic copy number variation in human neurons are currently unclear.

Genome-wide single-cell techniques also hold great promise for insights into nuclear organization. Chromosome conformation capture (3C) has revealed general principles of chromosome folding by studying the ensemble average of large cell populations. Single-cell analysis of nuclear organization, however, has so far only been possible using microscopy and was limited to selected loci. By developing a single-cell variant of the 3C technique, Nagano et al. (2013) discovered extensive cell-cell variability of intra- and interchromosomal contact formation in T helper cells. Although limited by relatively sparse genome coverage, averaged results of single-cell experiments agree well with ensemble studies. In a fascinating application of this technique, the authors use their single-cell chromatin interaction data to reconstruct the full 3D topology of the X chromosome.

Phenotypic and Developmental States as a Source of Cellular Heterogeneity

Multicellular organisms consist of different cell types that share the same genomic DNA but exhibit widely different morphological and phenotypic properties. Transcriptome analysis of individual cells is a major focus of single-cell biology. Single-cell expression profiling experiments have been performed using microarrays, high-throughput real-time PCR combined with microfluidics, and, more recently, single-cell RNA sequencing (RNA-seq).

In an interesting application of parallelized real-time PCR, Guo et al. (2010) analyzed the expression of 48 genes in more than 500 cells from the 1-cell-stage to the 64-cell-stage mouse embryo. The authors identified early markers of three different cell lineages (trophectoderm, primitive endoderm, and epiblast) and found that, at early stages, cells coexpress markers for different lineages. Lineage specification seems to proceed via downregulation of transcription factors of opposing lineages rather than via upregulation of specific transcription factors. The authors used a mathematical approach called principal component analysis (PCA) (Figure 2A) to distinguish different phenotypic groups at the 64-cell stage, where 3 cell lineages could be clearly defined. This then allowed them to identify early markers for cell lineage specification and show that expression of early marker genes significantly differed between inner and outer cells in the morula, confirming their role in early embryonic pattern formation.

Using the same high-throughput approach, Buganim et al. (2012) studied the dynamics of cellular reprogramming by analyzing expression of 48 genes in thousands of cells extracted from clonal populations at different time points. The authors identified two distinct stages of reprogramming, an early "stochastic" phase and a later "hierarchical" phase, and found genes that are predictive of successful reprogramming. By applying Bayesian network analysis, the authors inferred the topology of the transcriptional network governing the later stages of reprogramming directly from real-time PCR data. Importantly, the network structure was verified using coexpression analysis based on single-molecule RNA FISH (Raj et al., 2008) and genetically modified cell lines.

Large-scale real-time PCR allows analysis of many genes in thousands of cells. However, the fact that the entire genome cannot be probed at once in such assays means that researchers have to select genes of interest based on hypotheses or assumptions. Single-cell RNA-seq has the advantage of providing genome-wide information, leading to an unbiased and complete description of transcriptional heterogeneity in cell populations. Several protocols for single-cell RNA-seq have been published (reviewed in Shapiro et al. [2013]). In a pioneering study from the Surani lab, reverse transcription was performed directly on cell lysates from individual cells using oligo-dT primers (Tang et al., 2009). The cDNA library was then PCR amplified, fragmented, and subjected to sample preparation for deep sequencing. However, this approach suffers from two important shortcomings: the method is expensive when applied to a large number of cells, and PCR amplification can lead to biases in the mRNA counts. The first problem was addressed in two recent publications that made use of single-cell barcoding to allow multiplexing (Hashimshony et al., 2012; Islam et al., 2011): cell-specific barcodes are added to mRNA molecules during reverse transcription. Multiplexing comes at the

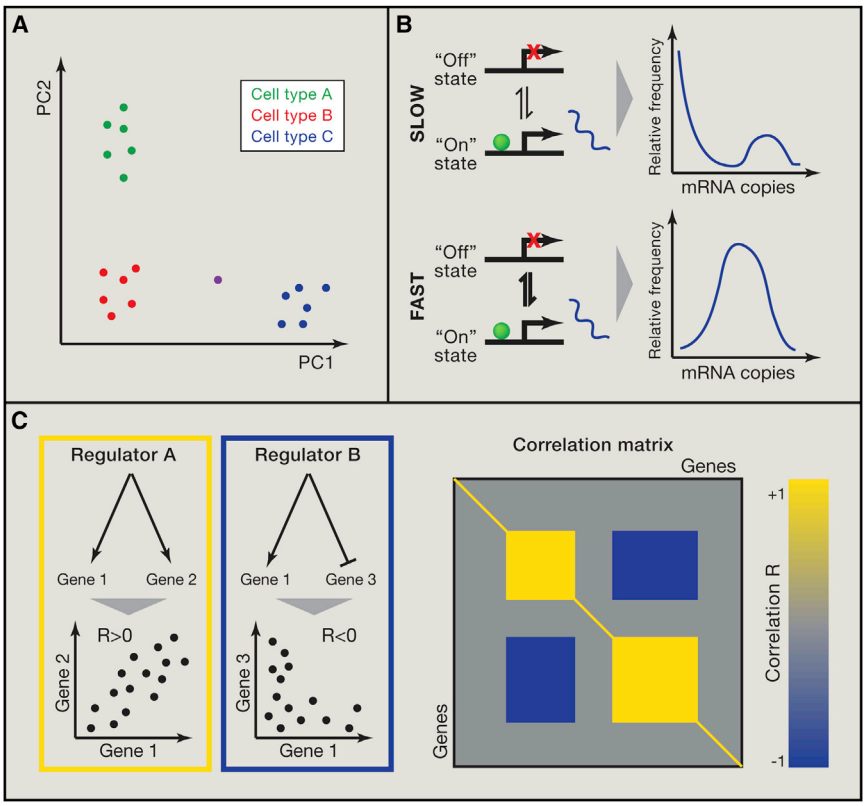

Figure 2. Mathematical Approaches for Analyzing Single-Cell RNA-Seq Data

(A) Principal Component Analysis reduces multidimensional data by identifying linear combinations of genes that are responsible for cell-to-cell variability. Here, each dot corresponds to a single cell. This analysis is often used for separating different cell populations in large-scale single-cell studies.

(B) Slow transitions between the "on" and "off" state of a promoter can give rise to bimodality (top). Fast transitions, however, lead to unimodal copy number distributions (bottom). This example illustrates that measurement of mRNA distributions allows determining the values of the kinetic parameters controlling expression of individual genes.

(C) Genes controlled by the same upstream regulator are expected to be positively or negatively correlated across single cells (left). Clusters of coregulated genes can be identified via calculation of pairwise correlations. This approach allows identification of regulatory modules in unperturbed wild-type cell populations.

expense of whole transcript coverage because these techniques are limited to sequencing either the 5′ or 3′ end of mRNAs. Addressing the second problem, linear amplification via in vitro transcription has been proposed as a way to reduce technical noise due to PCR bias (Hashimshony et al., 2012). The use of random barcodes incorporated in the primers for reverse transcription is another promising option (Kivioja et al., 2012). Applications relying on full transcript coverage, such as detection of gene isoforms, require methods for reducing the 3′ bias (e.g., by making use of template switching) (Picelli et al., 2013).

In an exciting application of single-cell RNA-seq, Shalek et al. (2013) investigated heterogeneity in the response to lipopolysaccharide of mouse bone-marrow-derived dendritic cells. The authors found extensive bimodality in the cells' response to lipopolysaccharide. Using PCA, they identified closely related yet different subpopulations, which they attributed to different maturation states of the cells. They then identified clusters of genes belonging to the same regulatory modules by calculating pairwise correlations of induced genes across all single cells (Figure 2C). By confirming selected coregulated genes identified by this analysis with real-time PCR and FISH, the authors provided an important proof-of-principle for systematic and genome-wide identification of regulatory circuits based on single-cell RNA-seq. These experiments were based on only 18 cells, and we anticipate that technological advances and reduction of sequencing costs will allow sequencing of much larger numbers of cells, leading to detailed insights into cell differentiation programs and regulatory modules.

Whereas protocols for single-cell RNA-seq work reliably and are now routinely used, single-cell proteomics is yet to follow. Single-cell mass cytometry (Bendall et al., 2011) allows parallel detection of a large number of proteins in single cells by using specific antibodies labeled with heavy metals. Antibodies coupled to distinct transition element isotopes are used to bind to their epitopes. Individual cells are then vaporized and ionized in a plasma, and elemental ions are detected by time-of-flight mass spectroscopy. Although having lower sensitivity than fluorescence-based flow cytometry, this novel approach enables simultaneous detection of a much higher number of different proteins and therefore allows more detailed analysis of cell states. Proximity ligation assays, where binding of two antibodies labeled with unique DNA tag sequences to the same protein molecule enables rolling circle amplification of the tags, allow for multiplexed protein detection through microscopy or next-generation sequencing (Leuchowius et al., 2013).

Stochastic Gene Expression as a Source of Cellular Heterogeneity

Gene expression is inherently stochastic. Random fluctuations of the mechanisms underlying mRNA and protein production cause heterogeneity among otherwise-identical cell populations (Raj and van Oudenaarden, 2008). Studying stochastic gene expression in cells of the same type requires analysis of hundreds of cells, as well as accurate quantification. So far, the necessary number of cells and the required level of accuracy have not been accessible to genome-wide techniques. Consequently, imaging-based techniques such as single-molecule FISH or GFP-based approaches have been used. We anticipate that genome-wide analysis of gene expression noise will become possible in the near future and discuss two possible applications of such data sets.

Distributions of mRNA or protein molecules can be used to understand gene regulation. As illustrated in Figure 2B, transition

rates between "on" and "off" states of a promoter have a direct influence on mRNA copy number distributions. Hence, quantitative genome-wide measurements of mRNA distributions would yield information about the parameters describing promoter function. Precise quantitative analysis of single-cell gene expression data for selected genes has demonstrated the potential of this approach for clarifying gene regulation (Neuert et al., 2013).

Although analysis of mRNA copy number distributions of individual genes can provide information about the kinetics of the on and off states of the promoter (Figure 2B), genome-wide data sets also allow identifying clusters of coregulated genes. Correlation analysis can identify clusters of cofluctuating genes, which are likely to be controlled by the same upstream regulators (Figure 2C). By performing flow cytometry analysis of pairwise correlations of GFP and mCherry-tagged proteins in unperturbed yeast cells, Stewart-Ornstein et al. (2012) were able to identify clusters of functionally related genes that are regulated by the same upstream factors. Importantly, their approach could be predictive of the magnitude of the response upon stimulation (Stewart-Ornstein et al., 2012). Single-cell RNA-seq has the potential to extend such analyses to the genomic scale and to mammalian cells.

Outlook and Challenges

Genome-wide single-cell analysis of gene expression is still in its infancy, and we anticipate that important breakthroughs will continue to revolutionize this young field. Automation and robotization using microfluidics will enable analysis of larger numbers of cells, allowing identification of rare intermediate cell states or weak regulatory interactions. Applying the concept of stochastic gene expression at the genomic scale will require innovative metrics that discriminate technical noise from biologically meaningful variation (Brennecke et al., 2013), perhaps such as the use of random barcodes as unique molecular identifiers (Kivioja et al., 2012). In the meantime, microscopy-based methods such as single-molecule FISH, which allow direct and amplification-free mRNA detection, can be used as a robust experimental framework to validate selected genes and quantify amplification bias. The detection efficiency of mRNA molecules in RNA-seq methods is typically low. Using spike-ins of synthetic RNA molecules at defined concentrations, Hashimshony et al. (2012) found that, in their single-cell RNA-seq protocol, around 90% of all mRNA molecules remained undetected, probably due to inefficiency in reverse transcription. However, by using microfluidics to reduce the reaction volume and optimizing buffers and reagents, the Linnarsson lab has recently reported detection efficiencies > 40% (Islam et al., 2014). Single-cell proteomics will require novel experimental techniques but is likely to add a new dimension to our understanding of cellular variability.

Investigating mechanisms underlying heterogeneous gene expression such as transcription factor binding, methylation, histone modifications, or nucleosome occupancy at the single-cell level remains largely uncharted territory, as does the spatial analysis of single cells in tissue. No matter what the next breakthroughs are, there is no doubt that the convergence of single-cell biology and genome-wide analysis will continue to foster amazing discoveries.

ACKNOWLEDGMENTS

We thank members of the van Oudenaarden laboratory for discussions and comments on the manuscript.

REFERENCES

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, A.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., et al. (2011). Science *332*, 687–696.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Nat. Methods *10*, 1093–1095.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Cell *150*, 1209–1222.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Proc. Natl. Acad. Sci. USA *99*, 5261–5266.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Dev. Cell *18*, 675–685.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). Cell Rep. *2*, 666–673.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Genome Res. *21*, 1160–1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Nat. Methods *11*, 163–166.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Nat. Methods *9*, 72–74.

Leuchowius, K.-J., Clausson, C.-M., Grannas, K., Erbilgin, Y., Botling, J., Zieba, A., Landegren, U., and Söderberg, O. (2013). Mol. Cell. Proteomics *12*, 1563–1571.

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., and Gage, F.H. (2013). Science *342*, 632–637.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Nature *502*, 59–64.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Nature *472*, 90–94.

Neuert, G., Munsky, B., Tan, R.Z., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Science *339*, 584–587.

Picelli, S., Björklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Nat. Methods *10*, 1096–1098.

Raj, A., and van Oudenaarden, A. (2008). Cell *135*, 216–226.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Nat. Methods *5*, 877–879.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Nature *499*, 431–437.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Nature *498*, 236–240.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Nat. Rev. Genet. *14*, 618–630.

Stewart-Ornstein, J., Weissman, J.S., and El-Samad, H. (2012). Mol. Cell *45*, 483–493.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). Nat. Methods *6*, 377–382.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Cell *150*, 402–412.

Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Science *338*, 1622–1626.