Single-cell genomics

Tomer Kalisky & Stephen R Quake

Methods for genomic analysis at single-cell resolution enable new understanding of complex biological phenomena. Single-cell techniques, ranging from flow cytometry and microfluidics to PCR and sequencing, are used to understand the cellular composition of complex tissues, find new microbial species and perform genome-wide haplotyping.

In 1839, Jakob Schleiden and Theodor Schwann formulated the 'cell theory' according to which cells are the basic structural and functional units of living organisms, from bacteria to animals and plants. Shortly thereafter, Rudolph Virchow extended this theory to state that new cells are formed from existing cells by cell division, and that tissues are formed by cell multiplication. Consequently, much effort has been devoted to developing measurement technology that allows one to interrogate biology in single cells. Here we will discuss approaches that enable single-cell analysis, both at the level of gene expression and by genome sequencing.

Brief history of single-cell analysis

Throughout its history, cell biology has been dependent on technological advances. The first of these, the invention of the microscope and the development of staining methods, permitted both tissues and individual cells to be viewed. Modern techniques provide more quantitative measures by combining molecular labeling with computational analysis of optical micrographs in a variety of powerful combinations: antibody staining of proteins allows a direct measure of gene expression products, RNA fluorescence *in situ* hybridization can be used for the sensitive detection and counting of RNA molecules in fixed cells or tissue sections[1], and fluorescent fusion proteins have revolutionized the ability to measure protein dynamics in living cells[2]. These methods generally allow one to measure a handful of gene products in each sample for a moderate number of cells. Laser capture microdissection enables extraction of single cells from specific locations in sectioned tissue samples[3]. Patch-clamp techniques can be used to record electrophysiological signals from single live neurons, and subsequent aspiration of intracellular contents for gene expression profiling provides a connection between physiological and molecular characteristics of single cells[4].

The throughput of single-cell analysis was radically improved by the invention of flow cytometry fluorescence-activated cell sorting (FACS) in the 1970s, and this approach is now widely used for single-cell measurements in medicine and biology[5]. Briefly, cells are suspended in a narrow liquid stream such that they pass single-file through the path of multiple laser beams, each of a different wavelength. Optical detectors convert fluorescent light emitted from each cell into an electrical signal. Often the cells are labeled with fluorescent antibodies to specific membrane proteins. Based on the intensity of signal emitted at different wavelengths, the cells can be analyzed one by one for various properties such as size, granularity and expression of membrane-bound proteins. Flow cytometers and cell sorters can process thousands of single cells per hour and can analyze up to 18 protein markers at a time. They can be used to purify subsets of cell populations based on combinations of membrane protein expression levels and to isolate single cells for gene expression and transplantation assays.

In the last 30 years FACS has enabled the identification and purification of a variety of cell types, including stem cells in tissues and tumors. Recent advances include phospho-specific antibodies that enable measurement of phosphorylation states in multiple proteins, thus making it possible to monitor signaling networks in thousands of single cells[6]. Another variation combines flow cytometry with mass spectrometry by attaching specially designed multiatom elemental tags to antibodies in place of fluorescent labels. This can increase the number of measurable markers by overcoming the limitations that arise from the spectral overlap between signals from different fluorescent labels and represents an exciting new avenue of research, although as generally practiced flow cytometry is still limited by the requirement of having antibodies to the target of interest[7].

The consequences of noise

In bacterial colonies, genetically identical cells have considerable phenotypic variability. This variability, also referred to as 'noise', arises from the intrinsic stochastic nature of biochemical processes regulated by a small number of molecules as well as from extrinsic sources such as the cell cycle[8]. Noise in gene expression has been a popular area of research whose full treatment is beyond the scope of this paper, which we will summarize by saying that even 'boring' housekeeping genes in

Tomer Kalisky is in the Department of Bioengineering, Stanford University, Stanford, California, USA. Stephen R. Quake is in the Departments of Applied Physics and Bioengineering, Stanford University, Stanford, California, USA and Howard Hughes Medical Institute, Chevy Chase, Maryland, USA.
e-mail: quake@stanford.edu

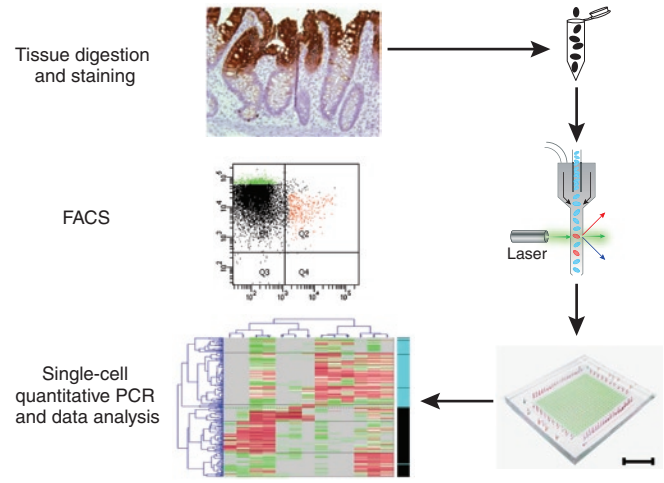

Figure 1 | High-throughput single-cell gene expression using microfluidic chips for studying the cellular hierarchy of solid tissues and tumors. A typical workflow is shown: a tissue sample is disaggregated into a single-cell suspension and stained for the desired surface markers. Single cells are sorted into individual wells using flow cytometry. Predetermined gene targets are reverse-transcribed and amplified using multiplex PCR. Subsequently, the amplified cDNA is multiplexed on a microfluidic chip (scale bar, 1 cm) with up to 96 gene-specific primers and probes, and quantified by PCR. Statistical analysis of the single-cell gene expression data can be used to identify cellular subpopulations comprising the tissue.

genetically identical cells have an extremely broad range of noisy gene expression values. Loosely speaking, the distribution is well described phenomenologically by a log-normal function, although the precise mathematical description is probably more complicated.

One may reasonably ask, 'if gene expression is so noisy in each cell, what is the point of it all? I should just analyze bulk samples so that the noise gets averaged out'. One answer to this question is that it is often of interest to know the exact distribution or variability of gene expression values from cell to cell. For example, cells may respond to an external stimulus uniformly or in a 'digital' manner such that only a fraction of cells respond and others do not[9]. This information is often lost in bulk measurements.

Another answer is that it is often difficult to obtain pure populations of a given cell type. In multicellular organisms, tissues are made up of multiple cell types, often including stem cells, progenitors in various stages of differentiation and mature cells of various types. Any bulk measurement will average over all of these cell types, but using single-cell analysis, in principle, one can deconvolve these cell types and discover the identities and gene expression profiles of various subpopulations, including rare ones such as stem cells.

Single-cell gene expression

Although FACS permits the analysis of more than just a few targets per single cell, methods for gene expression analysis at a still larger scale are needed to comprehensively understand cellular heterogeneity. PCR was developed in the 1980s and has been used to detect and amplify DNA targets from single cells for various applications from genetics[10] to immunology[11]. Quantitative reverse transcription–PCR (RT-PCR) has been used to quantify gene expression in single cells[12]. Using a targeted multiplexed preamplification technique, RT-PCR can be used to detect tens to hundreds of mRNA[13] or microRNA[14] targets from a single cell. Furthermore, microfluidic arrays can be used to combinatorially mix samples and assays, and perform thousands of PCRs in a single device[15], thus paving the way for simultaneous, high-throughput single-cell gene expression measurements from hundreds of individual cells and genes.

Using these techniques, it is possible to dissect heterogeneous tissues into cell subpopulations according to their unique gene expression profiles[16] (**Fig. 1**). Indeed, identifying different cell populations in the presence of noisy gene expression often requires that many genes be examined in many cells. For instance, with our collaborator Mike Clarke and his group, we have used microfluidic multiplexed PCR to map the cellular subpopulations in normal solid tissues and tumors. This approach is powerful because we can use the genes both to classify cells into subpopulations and simultaneously to interrogate the fundamental biological properties of those subpopulations[17].

Single-cell measurements often require precise counting of small numbers of molecules. Digital PCR is a method to count DNA or RNA molecules by limiting dilution; the sample is partitioned into many small isolated chambers such that each chamber is expected to contain on average one molecule or less[18]. A PCR occurs in each chamber, and the presence or absence of product is detected by a fluorescence signal. The total number of molecules is then estimated by counting chambers with products. Because of their capacity for miniaturization and parallelization, microfluidic chips are ideal for realization of digital PCR. For example, Warren *et al.*[19] counted the number of *Sfpi1* (also known as *PU.1*) transcripts encoding the low-copy transcription factor PU.1 and the high-abundance metabolic gene *Gapdh* in single cells sorted by FACS from hematopoietic cell lineages. These measurements showed that cells in the common myeloid progenitor compartment had large variation in *Sfpi1* expression and suggested that this heterogeneity is related to their committed differentiation to distinct cell fates[19].

Microarrays, developed in the 1990s, can be used to measure the expression of thousands of genes but usually require 1–2 µg of mRNA, which corresponds to ~10^6–10^7 cells. To analyze single cells, the RNA from the single-cell sample has to be reverse-transcribed and amplified using PCR-based methods[20] or T7 amplification (*in vitro* transcription)[21]. Transcriptome sequencing (RNA-seq) provides higher sensitivity and additional information not easily attainable by microarrays, and has also been adapted to single-cell samples by whole-transcriptome amplification[22].

Currently, both single-cell microarrays and single-cell RNA-seq are costly, limited to analyzing a small number of cells and have uncontrolled amplification bias. We expect that these technical limitations will be mitigated to some extent in the near future, and that sequencing and real-time PCR will then be used in conjunction: a small number of cells analyzed by sequencing to identify candidate genes, with follow-up studies on larger numbers of cells with real-time PCR.

Single-cell genome sequencing

As single-cell technology matures, direct analysis of genomes from single cells is becoming possible. This approach has been most broadly applied to the microbial universe, where understanding the genomic diversity and evolution of bacterial ecosystems is essential for applications ranging from understanding climate to the treatment of infectious disease. Only a fraction of all microbes has been identified, and an even smaller fraction has been grown in culture. Therefore, there is great interest in developing single-cell genome sequencing approaches that enable culture-independent characterization of microbes.

Otteson *et al.*[23] achieved a partial solution; they used microfluidic digital PCR to amplify arbitrary pairs of genes from single bacteria. This showed that gene function could be mapped to organism identity in complex ecosystems with few members that could be grown in culture, using the termite gut as a model. This allowed the identification of about a dozen new bacterial species that have a crucial role in acetogenesis—the process by which termites can eat wood and survive by metabolizing the acetate produced by commensal bacteria in their guts[23]. Tadmor *et al.*[24] recently extended this work; they used a similar approach to map the ecology of viruses infecting the bacteria that colonize the termite gut.

It is also of interest to try to obtain an entire genome sequence from single cells, and most work has used the biochemical amplification scheme of multiple displacement amplification (MDA)[25] for this purpose. MDA is particularly powerful in that it amplifies all DNA in a sample without the need for a priori sequence knowledge. However, background amplification of contaminating DNA from the sample and the reagents themselves presents a challenge. Many approaches have been used to control this problem, including UV-light treatment of reagents[26], the development of digital MDA[27], the preparation of ultrapure reagents[27] and the use of nanoliter volumes to perform amplification in microfluidic devices[28]. In two early reports, single bacteria from laboratory-cultured samples had been sorted by FACS and partially sequenced after whole-genome amplification[26], and uncultured bacteria from the human mouth were isolated in a microfluidic device, amplified and partially sequenced[28]. The microfluidic approach allows for extensive characterization of the cells to be processed and can be used for small clinical samples from individuals as well as for environmental samples containing diverse or unknown bacterial species. Since then, a few other examples of single-cell genomes have been published[29–32], and in the next few years we expect many new interesting bacterial genomes to be sequenced using single-cell methods (**Fig. 2**).

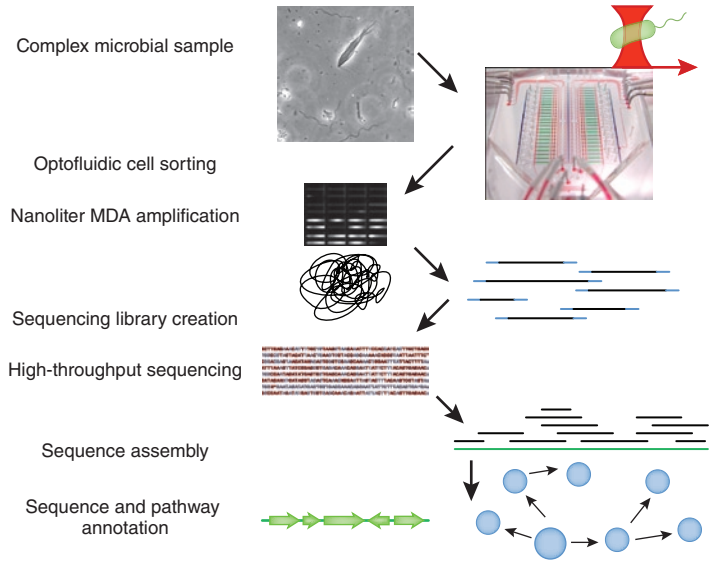

Figure 2 | Single-cell genome sequencing using microfluidics. A mixture of cells sampled from a complex microbial ecosystem is introduced into the chip. Single cells are selected using an optical trap, and are sorted into chambers for cell lysis and genome amplification. Genomes are amplified in nanoliter MDA reactions to produce larger quantities of DNA (shown are SYBR Green–stained products in microfluidic reaction chambers). Sequencing libraries are created from the amplified genomic DNA for sequencing on a high-throughput DNA sequencer. The sequence reads are assembled to recover the genome sequence, which is annotated to identify genes and pathways present in the original cell. The microfluidics image was reprinted from ref. 35.

Are single-cell genome analyses useful for human cells? Up until recently, the two most notable examples were karyotyping, intrinsically a single-cell measurement, and preimplantation genetic diagnosis, a PCR test performed during *in vitro* fertilization to check the preimplanted embryo for genetic abnormalities that may cause implantation failure, miscarriage or disease.

Just as in the case of microbes, however, we are beginning to see a new generation of approaches to single-cell human genome analysis. One such example is the lineage mapping approach developed by Ehud Shapiro and colleagues, wherein hypervariable parts of the genome are measured in single cells to map the lineages of a given tissue, with applications ranging from development to cancer[33]. Another example is our recent work using a microfluidic approach to amplify genomes from single human cells. In our experiments, the chips were designed with 48 amplification chambers, so that the individual chromosomes from the cell could be dispersed randomly to different chambers. Independent amplification and recovery of the material in these chambers has allowed us to use this approach to solve a problem that has bedeviled human genome sequencing: correct analysis of haplotype phase[34]. This approach will also be useful for numerous medical applications involving rare cells, such as circulating tumor cells in individuals with cancer and circulating fetal cells in pregnant women. We expect to see many innovative applications in this fertile area in the future.

ACKNOWLEDGMENTS

We thank P. Dalerba, M. Rothenberg and M. Clarke for providing the tissue section and FACS plot images for **Figure 1**, and P. Blainey for preparing **Figure 2** and for reading the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

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

2. Taniguchi, Y. *et al. Science* **329**, 533–538 (2010).
3. Tietjen, I. *et al. Neuron* **38**, 161–175 (2003).
4. Eberwine, J. *et al. Proc. Natl. Acad. Sci. USA* **89**, 3010–3014 (1992).
5. Shapiro, H.M. *Practical flow cytometry* 4th edn. (Wiley-Liss, New York, 2003).
6. Irish, J.M. *et al. Cell* **118**, 217–228 (2004).
7. Bandura, D.R. *et al. Anal. Chem.* **81**, 6813–6822 (2009).
8. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. *Science* **297**, 1183–1186 (2002).
9. Tay, S. *et al. Nature* **466**, 267–271 (2010).
10. Li, H.H. *et al. Nature* **335**, 414–417 (1988).
11. Maryanski, J.L., Jongeneel, C.V., Bucher, P., Casanova, J.L. & Walker, P.R. *Immunity* **4**, 47–55 (1996).
12. Lambolez, B., Audinat, E., Bochet, P., Crepel, F. & Rossier, J. *Neuron* **9**, 247–258 (1992).
13. Bengtsson, M., Stahlberg, A., Rorsman, P. & Kubista, M. *Genome Res.* **15**, 1388–1392 (2005).
14. Tang, F. *et al. Nat. Protoc.* **1**, 1154–1159 (2006).
15. Liu, J., Hansen, C. & Quake, S.R. *Anal. Chem.* **75**, 4718–4723 (2003).
16. Guo, G. *et al. Dev. Cell* **18**, 675–685 (2010).
17. Diehn, M. *et al. Nature* **458**, 780–783 (2009).
18. Sykes, P.J. *et al. Biotechniques* **13**, 444–449 (1992).
19. Warren, L., Bryder, D., Weissman, I.L. & Quake, S.R. *Proc. Natl. Acad. Sci. USA* **103**, 17807–17812 (2006).
20. Chiang, M.K. & Melton, D.A. *Dev. Cell* **4**, 383–393 (2003).
21. Luo, L. *et al. Nat. Med.* **5**, 117–122 (1999).
22. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
23. Ottesen, E.A., Hong, J.W., Quake, S.R. & Leadbetter, J.R. *Science* **314**, 1464–1467 (2006).
24. Tadmor, A.D., Ottesen, E.A., Leadbetter, J.R. & Phillips, R. *Science* (in the press).
25. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. *Genome Res.* **11**, 1095–1099 (2001).
26. Zhang, K. *et al. Nat. Biotechnol.* **24**, 680–686 (2006).
27. Blainey, P.C. & Quake, S.R. *Nucleic Acids Res.* **39**, e19 (2011).
28. Marcy, Y. *et al. Proc. Natl. Acad. Sci. USA* **104**, 11889–11894 (2007).
29. Woyke, T. *et al. PLoS ONE* **4**, e5299 (2009).
30. Woyke, T. *et al. PLoS ONE* **5**, e10314 (2010).
31. Blainey, P.C., Mosier, A.C., Potanina, A., Francis, C.A. & Quake, S.R. *PLoS ONE* **6**, e16626 (2011).
32. Rodrigue, S. *et al. PLoS ONE* **4**, e6864 (2009).
33. Frumkin, D. *et al. Cancer Res.* **68**, 5924–5931 (2008).
34. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. *Nat. Biotechnol.* **29**, 51–57 (2011).
35. Leslie, M. *Science* **331**, 24–26 (2011).