Method of the Year 2013

Methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine.

Once considered a technical challenge reserved for a few specialized labs, single-cell transcriptome and genome sequencing is becoming robust and broadly accessible. Exciting insights from recent studies are revealing the potential to understand biology at the unitary resolution of life, and last year marked a turning point in the widespread adoption of these methods to address a variety of research questions. For these reasons, single-cell sequencing is our choice of Method of the Year for 2013.

Every cell is unique—it occupies an exclusive position in space, carries distinct errors in its copied genome and is subject to programmed and induced changes in gene expression. Yet most DNA and RNA sequencing is performed on tissue samples or cell populations, in which biological differences between cells can be obscured by averaging or mistaken for technical noise.

Single-cell methods offer a way to dissect this heterogeneity. Single-cell DNA sequencing can reveal mutations and structural changes in the genomes of cancer cells, which tend to have high mutation rates. This information can be used to describe the clonal structure and to trace the evolution and spread of the disease. These approaches are also revealing a surprising level of mosaicism in somatic tissues such as the brain, the functional consequences of which will need to be elucidated in the coming years.

Differences between cells can be greater yet at the RNA level, even within seemingly uniform populations such as immune cells that have been purified on the basis of cell-surface markers. Single-cell transcriptome profiling can identify biologically relevant differences in cells, even when cells may not be distinguishable by marker genes or cell morphology, and can be used to group cells in an unbiased way.

Another advantage of single-cell sequencing is that it makes rare cells more accessible to analysis, provided that methods are available to isolate or enrich these cells from their heterogeneous environments. Cells taken from very specific spatiotemporal contexts, including microbes sampled from the environment, can be evaluated at the genome scale. In the clinic, single-cell sequencing can help with preimplantation screening of *in vitro*–fertilized embryos; and cancer diagnostics based on rare circulating tumor cells that can seed cancer at distant body sites becomes possible.

The central challenge of scaling down to the cellular level is capturing such a tiny amount of template and amplifying it to generate enough material for high-throughput sequencing. Maintaining fidelity and avoiding biases during heavy amplification is not trivial, but doing so is critical to ensuring adequate sequence coverage, accurate quantification and detection of sequence variation.

Recent protocol improvements and commercial offerings are helping to ease the adoption of single-cell sequencing approaches. Microfluidics and microwell technologies are also improving reproducibility and scale. We outline some basic workflows and considerations in a Primer (p. 18). In a News Feature (p. 13), Kelly Rae Chi highlights how single-cell sequencing approaches are already being effectively applied in the areas of biological development, cancer and neurobiology.

Single-cell genome sequencing reduces the sequence complexity of cell mixtures. In a Commentary (p. 19), Paul Blainey and Stephen Quake discuss how this can be leveraged to determine recombination frequencies in cells undergoing meiosis, to tease apart the maternal and paternal genomic contributions, or haplotypes, and to enable the assembly of microbial genomes sampled directly from complex mixtures in the environment.

In another Commentary, Rickard Sandberg argues that we are entering an era of single-cell transcriptome sequencing that will deepen our understanding of gene regulation and cellular transcriptional states, improve our ability to identify differences between healthy and diseased tissues, and profile rare cancerous cells (p. 22).

By focusing on genome and transcriptome sequencing, we do not mean to discount the importance of alternative single-cell approaches. Other methods such as *in situ* hybridization can effectively interrogate sequences in single cells in addition to providing the physical address of transcripts or DNA in intact tissue. Epigenomic profiling of single cells will add important information on gene regulation. Beyond sequence, approaches such as mass cytometry and mass spectrometry will help to characterize protein expression in single cells on a large scale. A final Commentary by James Eberwine and colleagues (p. 25) discusses the directions that such complementary technologies will need to take to understand single cells at the level of function.

We also present our Methods to Watch (p. 28), a selection of methods or areas of methodological development that we believe have particularly interesting potential in the coming years.

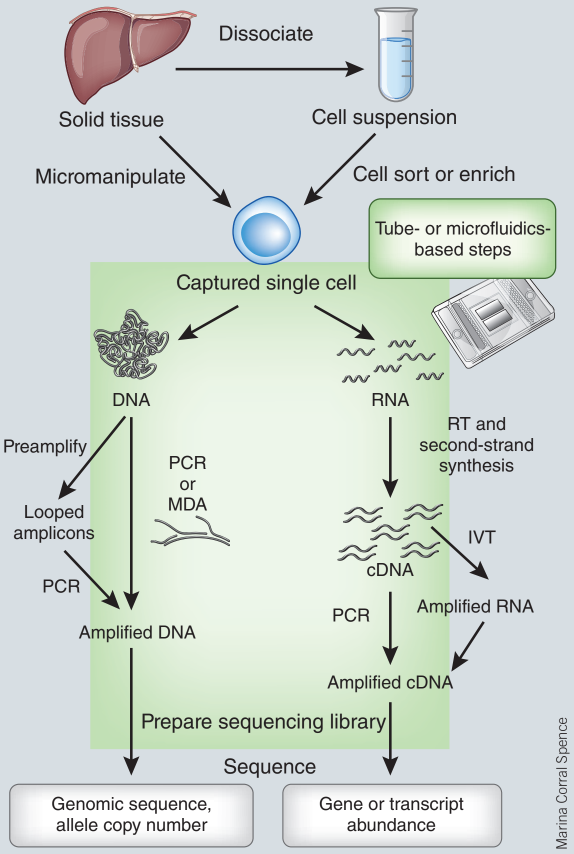
We hope that you enjoy our special feature. A happy 2014 to all our readers!

Single-cell sequencing

A brief overview of how to derive a genome or transcriptome from a single cell.

Genome and transcriptome sequencing require orders of magnitude more starting material than what is found in an individual cell, pushing the limits of amplification technology. Handling such small quantities means that degradation, sample loss and contamination can have a pronounced effect on sequence quality and robustness. Heavy amplification also propagates errors and biases, which can lead to uneven coverage, noise and inaccurate quantification.

Recent technical advances have helped mitigate these challenges, making single-cell sequencing an appealing way to address an expanding set of problems. Rare cell types, heterogeneous samples, phenotypes associated with mosaicism or variability, and microbes that cannot be cultured are good candidates for single-cell approaches. Single-cell sequencing can enable the discovery of clonal mutations, cryptic cell types or transcriptional features that would be diluted or averaged out in bulk tissue.



Marina Corral Spence

Workflows for amplifying and sequencing the RNA or genomic DNA of a cell. MDA, multiple displacement amplification; RT, reverse transcription; IVT, *in vitro* transcription.

Picking the right cell

Micromanipulation is a precise but laborious way to target a single cell, and microcapillaries can be used to extract a cell's contents directly. Many tissues can be dissociated to produce cell suspensions, which are easier to handle and allow cells expressing specific markers to be enriched with a cell sorter. Instruments that trap very rare cells on the basis of their surface markers are also being used to isolate tumor cells that circulate in blood.

Single transcriptomes

Many single-cell RNA sequencing protocols are now available, though every variation begins with the conversion of RNA to the first strand of complementary DNA (cDNA) by reverse transcriptase. Some methods sequence full transcripts and others sequence tags only at the 5′ or 3′ end.

The common goal of these methods is to capture the original RNA population and amplify it evenly and accurately. Capture efficiency is influenced by how comprehensively reverse transcriptase samples RNA from the cell—a stochastic process that can be improved using small reaction volumes and, potentially, better enzymes. In addition, a technique known as template switching may ensure that a greater proportion of captured transcripts are full length.

Amplification can also be improved by minimizing the number of cycles, inhibiting primer byproduct amplification by 'suppression PCR', and pooling barcoded cDNA from different samples to provide enough starting material for linear amplification by *in vitro* transcription. Unique molecular identifier sequences can also be used to label individual RNA molecules, allowing the absolute number of original molecules to be counted directly even after subsequent (uneven) amplification.

Single genomes

Whole-genome amplification starts with a tiny amount of material: just a single molecule of DNA. This makes uneven coverage and preferential amplification of one allele—or loss, known as allelic dropout—a widespread problem. The most common approach is multiple displacement amplification, which uses random primers that bind throughout the genome and a polymerase that can displace other fragments on the same template that it copies, forming iterative branching structures that massively amplify DNA. Early cycles have a strong effect on the uniformity of amplification. One variation uses special primers that cause amplicons to form closed loops and prevent further copying, allowing a few cycles of linear amplification prior to PCR. Scaling up and monitoring reactions in real time can help to overcome low success rates in genome amplification, and lower-input sequencing library preparations that rely on less amplification are another promising direction.

One cell for all

Scaling up is important to ensure that biological variability is well sampled in single-cell studies. Microfluidics or microwell technologies provide higher throughput and standardized handling, and are often efficient because reactions are concentrated in small volumes; however, microfluidics can be restricted to certain cell size ranges. Barcoding and pooling will also help to increase throughput.

Technologies for single-cell amplification and sequencing are maturing. As the cost and ease of examining individual cells improves, the approach will enter the hands of more researchers as a standard tool for understanding biology at high resolution.

Tal Nawy

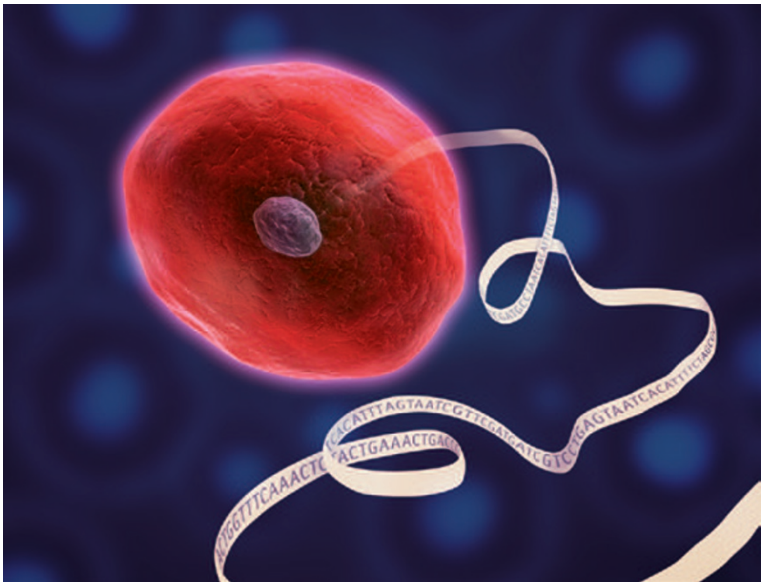
ACKNOWLEDGMENTS
We thank J. Levin and K. Zhang for help preparing this Primer.

Singled out for sequencing

Single-cell genome and transcriptome sequencing methods are generating a fresh wave of biological insights into development, cancer and neuroscience. Kelly Rae Chi reports.

Conceiving a child is an emotionally painful and exhausting process for those who struggle with infertility, and the worries don't stop with achieving pregnancy: all expectant parents hope for healthy babies. For individuals with known risks who are undergoing *in vitro* fertilization (IVF), preimplantation genetic diagnosis—in which clinicians remove a cell from an early embryo and screen it for genetic disorders—is a way to select an unaffected embryo, though current techniques analyze only one or a few sites in the genome. The cells of an early embryo are few and precious, so clinicians are keen to learn as much as possible from the limited numbers of cells.

That's one big problem that single-cell whole-genome sequencing methods are promising to resolve in early embryonic development and other fields. Thanks to improved approaches for isolating individual cells and for amplifying and sequencing their tiny complement of DNA or RNA, scientists can scan entire genomes or transcriptomes rather than a few targeted sites, and at higher resolution than was previously possible.



Erin Dewalt

In 2013, single-cell sequencing methods made their way to the mainstream.

One of several groups applying single-cell genome sequencing to IVF, Sunney Xie at Harvard University and his collaborators have tested their new whole-genome amplification methods on the first and second polar bodies, small cellular castoffs of the fertilized donor egg that reflect its chromosomal health. In a recent paper, Xie's team showed that in eight female donors, polar-body biopsy and single-cell sequencing could correctly infer both embryo aneuploidy—too many chromosomes, as in the case of Down's syndrome, or too few—and single-nucleotide variations inherited from either parent (*Cell*, doi:10.1016/j.cell.2013.11.040 19 December 2013). Detecting aneuploidy may require sequencing as little as one out of every hundred genomic regions on average, making the strategy cheaper and more accurate than traditional methods, Xie says.

Xie and his collaborators on the paper, Fuchou Tang of Peking University and Jie Qiao of Peking University Third Hospital, have launched a clinical study of women undergoing IVF. The team will amplify and sequence whole genomes of the polar bodies of participants' embryos to see whether they are fit for transfer. Such a step toward the clinic seemed impossible only 2 years ago, says Xie, adding that people desperate to have a baby free of a devastating genetic disorder have been e-mailing him. The study's first baby could be born within the year. "I didn't anticipate that [our technique] would be used so quickly for patients," he says.

Sequencing in 2013

Single-cell sequencing is no small feat. The amount of DNA or RNA in a single cell starts at a few picograms—not even close to the quantity that today's sequencing machines demand. So scientists must amplify these molecules and do so in ways that minimize technical errors while surveying sequences as broadly and evenly as possible. Until recently, many researchers doubted that sequencing of single cells could be reliably conducted by any but a few experts.

Although a handful of groups sowed the seeds for single-cell genome and transcriptome sequencing approaches years ago, the methods have more recently started to make their way to the masses, and a community has formed around their application in areas including neuroscience, cancer and microbial ecology. "Almost since the first day that PCR was invented, people began trying to use it to do single-cell gene expression and genome analysis," says Stephen Quake at Stanford University, cofounder of Fluidigm. "But [single-cell sequencing] really is just taking off for a bunch of reasons."

Updated protocols for DNA and RNA amplification, especially those disseminated in the last two years, have given new

Z. He

Sunney Xie may soon see his group's genome amplification method used for preimplantation genetic diagnosis.

users greater choice for their experiments. Industry has also contributed countless kits for amplifying genetic material from single cells, and readout technologies have lowered in price. Fluidigm introduced the first single-cell automated prep system for RNA-seq in 2013. All these advancements are lowering barriers for beginners. "People have been wanting to do this for decades," says Rickard Sandberg of Karolinska Institutet in Sweden, referring to single-cell RNA sequencing. "It's just that technology is now allowing us to do it in a much cheaper and much better way than before. It's becoming really accessible for lots of labs."

At the heart of the approach is the question: why go to the single-cell level? The rationale is that the alternative—pooling cells by the thousands or millions—blurs potential insights into the heterogeneity of complex systems such as the brain, blood and immune system, or even their component cell types. "When you go to the level of the single cell, you lose the information in the total system," says James Eberwine of the University of Pennsylvania. "But if you can do multiple cells within that system, then you can build up that system, I think, in a more informative way."

Numerous fields in which bulk tissue approaches may be insufficient are beginning to benefit from the new tools. And not only are single-cell sequencing methods helping define heterogeneity among cells, they also are allowing a level of comparison that is expected by many to redefine what a cell type is.

Tempering some of the enthusiasm are myriad challenges inherent to the process, from the isolation of cells, to amplification of their genomes or transcriptomes, to making sense of the data. Cost is also a consideration—single cells typically need to be sampled at higher numbers than tissues do—leaving good reason to carefully select situations that justify going to the single-cell level. "Do we need to analyze single cells to meet the objective? If the answer is no, you shouldn't do single cells. It's hard, expensive, and you start encountering a lot of variability," says Paul Blainey at the Broad Institute and MIT.

From a few molecules of RNA

Sequencing a cell's transcriptome hinges on the ability to amplify large amounts of the complementary DNA (cDNA) that is synthesized from RNA. Capturing small amounts of RNA as cDNA and amplifying the cDNA extensively are difficult to do evenly and efficiently.

In 1990, transcriptome analysis at the resolution of single cells was made possible by Norman Iscove's group, who amplified cDNAs exponentially using PCR. In the early 1990s, Eberwine and his colleagues came up with a technique that generated cDNA from single live neurons and performed linear amplification by transcribing RNA from the cDNA. With the advent of microarrays, scientists used both linear and exponential amplification strategies to identify differences in gene expression among single cells.

High-throughput RNA sequencing (RNA-seq) came onto the scene in 2008, and shortly after, researchers coupled it to such amplification techniques to get a more detailed look at single-cell transcriptomes. For a 2009 study, Tang, then working in M. Azim Surani's laboratory at the Gurdon Institute at the University of Cambridge, showed that it was possible to detect—from a single mouse blastomere—the expression of thousands more genes than had been revealed using microarrays (*Nat. Methods* **6**, 377–382, 2009).

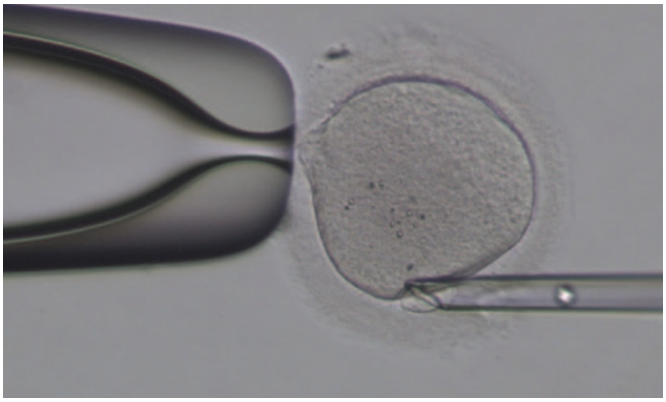
That same year, Cold Spring Harbor Laboratory hosted its first single-cell meeting, and fewer than 50 scientists—developers and early adopters—attended. "I remember everybody was trying to do RNA-seq and trying to figure out what they had, how to believe what was real and figure out reproducibility," says Mike McConnell, now at the University of Virginia.

Methods development has since come a long way, researchers say. Now that there are protocols and product offerings for single-cell sequencing, says Sten Linnarsson at Karolinska Institutet, "the phase of pure method development has culminated this year, and it's now become possible to actually use these methods at a pretty large scale to address biological questions." Rather than hundreds of cells, some groups are aiming to analyze tens of thousands.

For example, as part of the Single Cell Analysis Program supported by the US National Institutes of Health Common Fund, Kun Zhang's team will generate full transcriptomes from 10,000 cells in three areas of the human cortex. They will group the transcripts into cell types—perhaps redefining those cell types in the process—and map the transcripts back to cortical slices of the brain. Single-cell RNA-seq itself is no longer a barrier. "If you have a good cell, and you want to get a measure of the transcriptome, there is more than one option that can lead you to that goal," Zhang says. In general, however, extracting the neurons posthumously, minimizing RNA degradation and preserving some of the neuronal spatial information is challenging, and the group is evaluating several approaches, Zhang says.

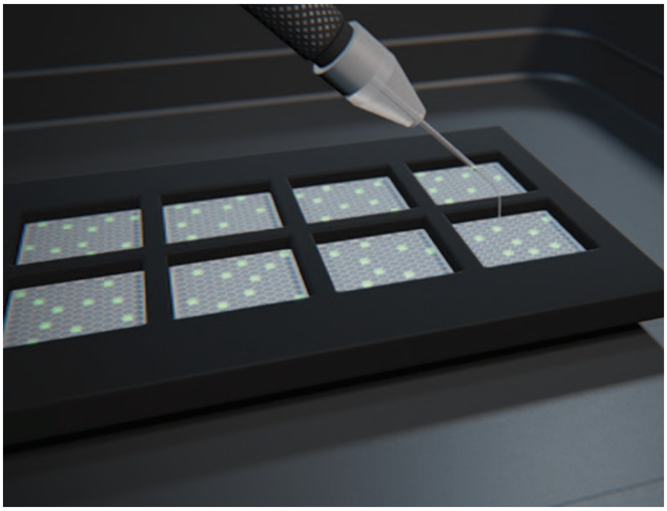
Amplifying the genome

Developing a way to amplify whole genomes of single cells took a bit longer because only one or two unique copies of DNA exist in the cell. The method lagged behind RNA amplification until 2005, when Roger Lasken's group became the first to amplify and sequence DNA from a single cell, that of an *Escherichia coli* bacterium, using the multiple displacement amplification (MDA) method that they had developed. That sparked a vigorous effort by microbiologists to generate reference genomes for diverse, uncultivable bacterial species.



F. Tang, L. Yan

Sequencing the genomes of polar bodies associated with an *in vitro*–fertilized egg can help with preimplantation diagnostic screens.



K. Zhang

MIDAS is a scaled-up strategy for isolating, amplifying and sequencing genomes from single cells.

One of the most common strategies still used today, MDA is carried out by polymerases such as Phi29, which elongates random primers that have annealed throughout the genome. Each polymerase can displace neighboring elongating strands to produce large quantities of long (7- to 10-kilobase) overlapping copied fragments for sequencing.

M. Nemchuk/Broad Institute

Single-cell RNA-seq worked on the first try, says Aviv Regev.

In 2011, researchers coupled single-cell genome amplification with high-throughput sequencing. Working in Michael Wigler's group at Cold Spring Harbor Laboratory, Nicholas Navin profiled—at 50-kilobase resolution—large deletions or duplications of DNA called copy-number variants (CNVs) across the genomes of breast tumor cells from two individuals (*Nature* **472**, 90–94, 2011).

One of the biggest challenges in single-cell sequencing of genomes is that some portions of a string of DNA get amplified more than others. In 2012, Xie's group described a new strategy called MALBAC, or multiple annealing and looping-based amplification cycles, that involves five cycles of MDA 'preamplification' during which newly amplified fragments form closed loops (*Science* **338**, 1622–1626, 2012). The loops prevent the fragments from being copied again, and the amplification thus stays linear. Normal PCR follows the preamplification but is less prone to bias because of the more evenly amplified starting template. Using MALBAC, Xie's group obtained enough coverage to sequence 93% of the human genome and detect CNVs in a single cancer cell.

Scientists will soon be able to probe genomes more deeply in each cell, which will allow them to see smaller deletions and duplications or even single-nucleotide variations. Amplifying the genome evenly still poses a challenge, but experts believe that scaling down reaction volumes will help reduce error.

For example, Zhang, at the University of California, San Diego, and his colleagues recently described MIDAS (micro-well displacement amplification system), a strategy for conducting MDA reactions in thousands of nanoliter-sized compartments etched onto a glass slide (*Nat. Biotechnol.*, **31**, 1126–1132, 2013). Researchers extract amplified fragments manually or with a robot and then sequence them. MIDAS allowed the group to detect single-copy-number changes in human neurons, with very little sequencing, at 1- to 2-megabase resolution.

Cells expressing their differences

At the Broad Institute, Aviv Regev, Joshua Levin and their colleagues were comparing RNA-seq methods for low-quantity and degraded bulk samples, when it occurred to them to try RNA-seq on a single cell. They decided to use a protocol called Smart-Seq on bone marrow–derived dendritic cells, postmitotic immune cells known to generate strong transcriptional responses to antigens.

That pilot study used 18 single cells and Regev allotted a week for the experiment. "You try out many things and they fail," but this worked on the first try, she says. Each cell uniformly expressed a set of 'housekeeping' genes, but the individual cells also revealed a surprise: genes important for immune regulation were expressed either at high levels or not at all. Such bimodality had never before been seen in dendritic cells because differences among cells are averaged out when populations are sequenced. The results, published last June, suggested the presence of a cryptic cell type—a rare 'first responder' among what was thought of as a highly pure population (*Nature* **498**, 236-240, 2013). More broadly, the findings help reshape our understanding of these cells' identity, signaling and behavior.

Single-cell transcriptome sequencing is also helping researchers study gene expression and regulation in early development, and in far greater detail than what was previously possible for such rare samples. For a study published last August, Guoping Fan from the University of California, Los Angeles, and his collaborators in China sequenced transcriptomes from 33 single cells in multiple stages of development, identifying the order in which clusters of genes are expressed through the initial stages of development and how the timing of gene expression differs between early human and mouse embryonic development (*Nature* **500**, 593–597, 2013).

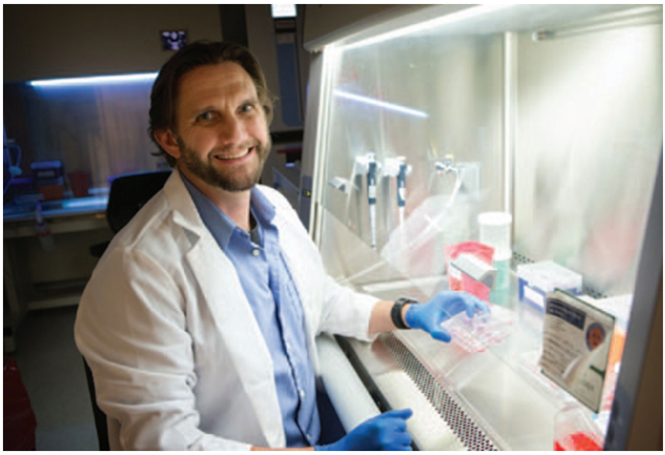
Meanwhile, Tang's group carefully dissociated each cell of several early human embryos and sequenced their transcriptomes individually. "It's quite stressful. It's so important and a rare sample," he says. But the pressure paid off: they discovered more than 2,700 new long noncoding RNAs in the embryos that may play roles in early gene regulation (*Nat. Struct. Mol. Biol.* **20**, 1131–1139, 2013). Before this, all single-cell RNA-seq work had analyzed known genes or, at most, novel alternative splicing isoforms of known genes, Tang says.

Univ. Texas MD Anderson Cancer Center

Single-cell sequencing is a powerful tool for understanding genomic variation in cancer cells, says Nicholas Navin.

The cellular patchwork of cancer

From prognostics to disease monitoring, cancer research stands to benefit enormously from single-cell sequencing approaches. Cancer cells often undergo high mutation rates, and tumors tend to be heterogeneous. Identifying which subsets of cells, called clones, are present and evolve into metastases or respond in a certain way to chemotherapy is critical to understanding and fighting the disease. In particular, circulating tumor



D. Addison

Mike McConnell found single neurons in the human brain with large DNA deletions or duplications.

cells (CTCs)—which break off from a tumor and seed a cancer's metastasis—are those rare cells whose genomes or transcriptomes might offer clues for diagnosis, monitoring or treatment.

In Navin's 2011 *Nature* study, for example, profiling the genomes of single cells for CNVs revealed a punctuated model of tumor evolution: bursts of genomic instability following a stable expansion of tumor mass. "That was surprising, because people believed … mutations gradually accumulate over time," says Navin, now at the University of Texas MD Anderson Cancer Center. "It showed how powerful these single-cell tools could be for understanding at least copy-number alterations in human cancers." He and his collaborators have continued to study copy-number evolution in triple-negative breast cancers—a heterogeneous and aggressive group of cancers—and also hope to better understand metastasis.

Besides Navin's, several other groups are applying single-cell sequencing approaches to cancer. For example, Xie, working with Fan Bai at Peking University and Jie Wang at Peking University Cancer Hospital, found a shared pattern of CNVs among CTCs of people with one subtype of lung cancer but not another (*Proc. Natl. Acad. Sci. USA*, doi:10.1073/pnas.1320659110, 9 December 2013). This recent finding offers potential for early diagnosis, Xie says.

Transcriptional differences may also hold the key to understanding cancer progression. Sandberg's group used their Smart-Seq method to sequence RNA from a single CTC as a proof of concept for their methods. Using their new version, Smart-Seq2, they can look at many more cells at a fraction of the cost. The technical noise plaguing CTC studies will greatly improve with more cells. "We are really hoping to make a more systematic effort, to better understand the heterogeneity in CTCs and to better understand their gene expression programs when they go into circulation," he says.

More elusive still than the genome or transcriptome is the epigenome, the chemical marks on the genome that guide gene expression. Although current techniques have proven insufficient for the single cell—the traditional methods for detecting epigenetic marks on DNA also tend to degrade it—researchers are eager to see what the epigenome might reveal about cancer.

Tang's group has developed a tube-based method to reveal genome-wide DNA methylation in a single cell (*Genome Res.* **23**, 2126–2135, 2013). "[Epigenomes] really need the single-cell approach" for researchers to see how tumor cells are different from their neighbors, whether that's through methylation or other mechanisms, he says. Wolf Reik's team at the Wellcome Trust Sanger Institute has taken methylome analysis to 50–100 cells, and, he says, "We are very interested in pushing the boundaries further."

Uncharted territory in the brain

Neurons are among the newest cell types to receive the single-cell treatment, and scientists aren't exactly sure what to expect from them. Experimental support for the idea that neurons harbor diverse genomes came about only recently. Even with these results, the diversity is baffling. In 2001, Jerold Chun's group, then at the University of California, San Diego, found aneuploidy in the mouse brain (and in 2005, in human neurons). "No one knew what to do with it," says McConnell, a graduate student in Chun's lab at the time. "The tip of the iceberg was how I saw it. If there's aneuploidy, there's got to be a lot more changes in these genes or in these genomes."

In the meantime, scientists demonstrated that there are 80–100 potentially active L1s—bits of DNA that copy and paste themselves throughout genomes—in every human genome and that L1s are active in neurons. Those studies and others delivered potential suspects in genomic diversity, but how extensive the variation is remains unclear today.

"We are just beginning to understand the molecular diversity of cells in the brain," says Thomas Insel, director of the US National Institute of Mental Health. "Single-cell methods will be critical, not only for defining the taxonomy of neurons and glia but for revealing the effects of experience or development on profiles of expression within a brain region."

Researchers are unlocking single-cell genomic variation in a few different ways. Christopher Walsh's team at Harvard Medical School scanned the genomes of 300 single human neurons isolated posthumously for L1 insertions (*Cell* **151**, 483–496, 2012). They found few, a result suggesting that L1 is not a major player in genomic diversity, at least in the cortex and caudate nucleus.

In 2013, other groups scanned entire genomes of single human neurons. For a study published in November, researchers took 110 frontal cortex cells from three healthy human brains, sequenced their genomes and found a surprisingly high proportion of neurons with large CNVs (*Science* **342**, 632–637, 2013). Neurons derived from the skin cells of healthy people also had more CNVs than did the skin cells themselves; these findings suggest that human neurons derived from induced pluripotent stem cells might work well for studying the functional implications of cell heterogeneity.

Babraham Institute/Wellcome Trust Sanger Institute

Wolf Reik hopes to see epigenetic methods reach single-cell resolution.

Indeed, neuroscientists are still wrapping their heads around what somatic variation could mean. It might make the brain more robust to perturbation, says geneticist Ira Hall at the University of Virginia, a corresponding author on the *Science* study. On the other hand, genomic mosaicism could affect risk for cancer or other diseases, he adds. To know whether some brain regions are more affected than others, or how much variability there is from one individual's brain to another's, researchers will have to look at many more cells. "I don't know if we know yet how many," adds McConnell, who led the study.

Beyond proof of concept

Although the single-cell field is abuzz with biological questions, methods development is by no means over. Researchers must sort out biological variation from technical noise in their own systems. Single-cell RNA and DNA sequencing technologies are not yet robust, notes Joakim Lundeberg of the KTH Royal Institute of Technology in Sweden, whose group has developed a preliminary approach to sequencing RNA in the tissue context. "We will need many more single cells to be analyzed within a single experiment to address biological and stoichiometric noise or at least to achieve a better understanding of cell-to-cell variation within tissue," he adds.

Because of the challenges faced on many fronts, including cell isolation and computation, and in different fields of application, Blainey expects more waves of technological advancements in the next few years.

For newcomers, the choice of which transcriptome-sequencing method to use might be daunting. The answer likely depends on one's research goals—whether it's analyzing many cells, looking for transcript isoforms or detecting low-abundance RNA. But "there's more than one way to skin a cat, which is good," Quake says. In October, his group showed that single-cell qPCR and RNA-seq perform similarly when sample prep is done in nanoliter-scale reaction volumes—in this case, with Fluidigm's C1 system (*Nat. Methods* **11**, 41–46, 2014). "That's a really good thing in terms of trust in the measurement," he says.

The selection of genome amplification strategies will improve, too, as new platforms become commercially available and single-cell researchers introduce their own versions of protocols. On the other hand, because everyone does amplification differently, it's hard to directly compare strategies across different studies. "In our hands MALBAC is much better than MDA, but it depends on how you do MDA," Xie says. With further development both methods will be obsolete, he says, and he plans to be part of that effort. "By no means is MALBAC the end game. We're trying to do better."

Meanwhile, research in cancer, neuroscience, microbiology, drug development and a host of other fields will continue to reap the rewards as new and improved methods impart the ultimate, cellular resolution of transcripts and genomes. The approaches will attract newcomers such as Reik, who is well established in the epigenetics field. Reik attended his first single-cell meeting in 2013. "I haven't been part of the single-cell community before," he says. "I thought it was hugely exciting to see the field galvanized by these technologies. It is primarily still technology driven, and that's great, but very quickly now we'll get to some super exciting biology questions."

Corrected after print 5 February 2014.

Kelly Rae Chi is a freelance writer based in Cary, North Carolina, USA.

Erratum: Singled out for sequencing

Kelly Rae Chi

Nature Methods 11, 13–17 (2014); published online 30 December 2013; corrected after print 5 February 2014

In the version of this article initially published, an incorrect institutional affiliation was given for Jie Wang: this affiliation was listed as Harvard when it should have been Peking University Cancer Hospital. The error has been corrected in the HTML and PDF versions of the article.

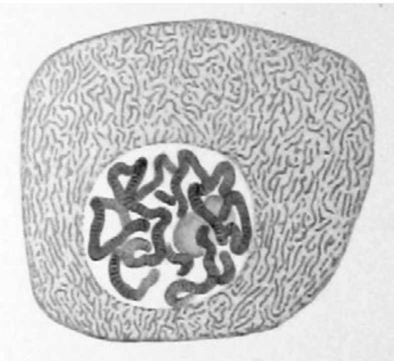
Dissecting genomic diversity, one cell at a time

Paul C Blainey & Stephen R Quake

Emerging technologies are bringing single-cell genome sequencing into the mainstream; this field has already yielded insights into the genetic architecture and variability between cells that highlight the dynamic nature of the genome.

The idea of analyzing genomes at the single-cell level is quite old: the banded structure of polytene chromosomes was first reported in images of single cells from the salivary glands of insects in 1882 (ref. 1). In 1935, Calvin Bridges published a map of the *Drosophila* genome based on such images, which enabled the identification of large-scale genomic rearrangements that distinguished different individuals, lines and species[2]. More recently, there has been sustained effort to apply the polymerase chain reaction and other biochemical amplification technologies to single cells. Notable results include the analysis of recombination hot-spot usage in single sperm cells two decades ago[3] as well as the routine analysis of single cells from embryos for preimplantation genetic diagnoses[4]. Given the century-long history of this field, it is quite reasonable to ask, why has there been a sudden recent flood of attention on single cells?

We argue that the answer has to do with phenomenal recent advances in the ability to analyze detailed sequence information from single cells. These are defined by a confluence of three factors: advances in technology that enable effective whole genome and transcriptome amplification, the relentless improvement in DNA-sequencing instruments with ever higher throughput and lower costs, and the invention of technologies for single-cell manipulation such as microfluidics and fluorescence-activated cell sorting. The last 5 years have seen a burst of papers from labs around the world that have developed expertise in single-cell gene expression and genome analysis, and commercial vendors have played a crucial role in helping expand access to these technologies. Single-cell genome analysis is now influencing areas as diverse as microbial ecology, cancer, prenatal genetic diagnosis and the study of human genome structure and variation (reviewed in refs. 5 and 6). In this Commentary, we will focus on recent highlights and our best guesses of where the field might be going next.



A single-cell genome image of polytene chromosomes from the 1882 monograph by Flemming (ref. 1).

Single-cell sequencing of single-celled life

Microbial ecology is an ideal arena for single-cell genomics because the vast majority of microbes—99% of species, by most estimates—cannot currently be cultured. Uncultured species exist as biological 'dark matter', as they can only be observed indirectly by methods such as marker-gene sequence surveys. Although metagenomic approaches can help establish gene inventories from such complex environments, the fundamental link between organism and gene is lost. It is only through single-cell genomic approaches that one can understand the connections between a unicellular organism's identity and the functional capabilities provided by its genome. A consequence of this is that an enormous fraction of genetic and evolutionary diversity on earth is not fully accounted for in current genomic databases.

The first uncultured microorganism to be analyzed with single-cell sequencing was a bacterium that lives on human tooth plaque[7]. In recent years there have been more than a dozen publications on single-cell genomes from yet-uncultured microbes, and we believe that this trend is going to increase exponentially as technologies continue to improve. As these data accumulate, we may see the discovery of previously unknown microbial functions and metabolites, the identification of many new species that relate to human health in both positive ways (such as through the microbiome) and negative ways (as newly recognized pathogens), and even major revisions to the structure of the 'tree of life' and insights into the evolutionary relationship between eukaryotes, bacteria and archaea.

The diversity of morphology, physiology and genotypes to explore among microbes also creates technical challenges for single-cell analysis. Numerous sample-specific considerations come into play when choosing reaction formats and chemistries[5]. For instance, microbes often require stringent lysis conditions, and the need to match

Paul C. Blainey is in the Department of Biological Engineering, Massachusetts Institute of Technology, and Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA, and Stephen R. Quake is in the Departments of Applied Physics and Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford, California, USA.
e-mail: pblainey@broadinstitute.org or quake@stanford.edu

conditions to different microbe types may complicate protocols. Because DNA purification is not commonly applied before amplification, the amplification reagents must also be compatible with the lysis reagents and contents of the lysed cells. Complex lysis and amplification protocols are well-suited to microwell plates and integrated microfluidic devices where protocol steps can be automated; interestingly, the performance of biochemical amplifiers improves as reaction volumes are shrunk to nanoliters in microfluidic formats[8]. Simple protocols may be well suited to reverse-emulsion liquid-droplet systems, where thousands of individual microreactors can be quickly produced and processed. Nearly all of the microbial single-cell sequencing results to date have used one particular whole-genome amplification chemistry: multiple displacement amplification (MDA), which is an isothermal amplification scheme that uses random primers and that is based on the strand-displacement ability of Φ29 DNA polymerase[9].

Human haplotypes

Human-genome analysis has rapidly progressed from determining the reference sequence for the 'average' human genome to prolific sequencing of personal genomes, and it may seem surprising that single-cell approaches have anything more to contribute. However, some aspects of the human genome have been very challenging to determine using conventional techniques. For example, all of us have two genomes within each of our cells—one from our mother and another from our father—and the location of sequence variation in each haploid genome can have a significant effect on gene expression, protein function and disease.

The best-known example of this is variation in human leukocyte antigen (HLA) genes, whose haplotype is important to understand for bone marrow transplants, but it applies just as well to compound heterozygous mutations—two mutations at a single locus that may be harmless when they reside in the same haplotype, but deleterious when distributed between the maternal and paternal allele. Current techniques have not been able to resolve these differences, known as haplotype determination, at a genome-wide level with any degree of precision. The best conventional methods for haplotype determination require additional sequencing within a family pedigree, mainly from parents. Clearly, this is not a practical approach in most clinical situations.

Single-cell chromosome isolation enabled the first genome-wide haplotype measurement, with haplotype phase determined across the lengths of entire chromosomes[10]. This work was quickly followed by related single-cell sequencing approaches using small numbers of cells[11], or in the case of males, large numbers of individual sperm cells[12]. We expect that further application of these techniques and long-read sequencing technologies that determine haplotypes of genomic segments will accelerate analysis of refractory parts of the human genome. The HLA region, which is the most polymorphic part of the human genome and intimately involved in the immune system and many aspects of human health, is a particularly interesting target, but up to now it has been sequenced in only a limited number of individuals owing to its haplotype complexity.

Another area touched by single-cell genomics is the analysis of recombination patterns across human diversity. Recombination is the cutting and pasting of large blocks of the maternally and paternally inherited chromosomes to create entirely new genomes in sperm and egg cells, and it is a major contributor to genetic diversity in the human species. It is known that recombination does not happen with uniform probability across the genome; rather, certain 'hot spots' experience frequent recombination. One of the earliest contributions of single-cell genomics was to show that there is differential hot-spot usage between individuals: some spots may be hot for one person but not another[13]. More recently, single-cell approaches have enabled the measurement of genome-wide recombination patterns and mutation rates in individual sperm cells, enabling the first studies of genome-wide hot-spot behavior within individuals[12,14]. We expect that further genomic analyses of single sperm cells will enable the study of recombination mutants (for example, in individuals carrying rare alleles of *PRDM9*) as well as the potential diagnosis of those with meiotic dysfunction related to sterility and infertility.

Somatic variation

The value of sequencing individual human genomes is increasingly being recognized, yet a personal genome is actually an average of cellular genomes in the body, which also vary. Genomic differences in some cell types are well characterized and have been appreciated for decades; these include B cells of the immune system, which make such a strong commitment to expressing a particular unique antibody that they irreversibly reprogram that gene in their genome. Germ cells also differ through the well-characterized process of meiosis and recombination, as discussed above. Less well understood but still important are the gradual accumulation of mutations through errors in cell division and the movement of mobile genetic elements[15].

This gradual accumulation of errors is associated both with aging in general and with cancer in particular, so it is not surprising that these areas will become important for single-cell genomic applications. To date, single-cell methods have been used to directly measure the *de novo* mutation rate in human sperm[14] as well as immortalized human cell lines[16]. They have also been used to determine the order of mutations that occur in normal hematopoietic stem cells before their transformation to acute myelogenous leukemia[17], to understand the lineage structure of leukemia tumors[18] and to estimate the clonal structure of breast tumors[19].

Mosaic variation is also known to exist in adult neural tissue and has been implicated in neurodegenerative disorders such as Alzheimer's disease[20]. Recently, single-cell genomic techniques have been used to detect megabase-scale copy number variation in a large fraction of induced pluripotent stem cell–derived neurons and normal postmortem brain cells using single-cell MDA and genomic analysis[21]. Similarly, single-cell MDA and PCR-based whole-genome amplification were used to show how retrotransposition of an L1 element is a potential driver of somatic mosaicism in the brain, and further, how a variant present in a third of cells or less can drive serious brain disease (hemimegalencephaly)[22]. A complementary method—fluorescence *in situ* hybridization—was used to show how the proportion of aneuploid neurons increases in aging mice[23]. This is a fascinating area, as there are varying degrees of evidence that mosaic somatic variation is functional in development[24], is present in normal mature neural tissue[21], may explain variability among 'normal' neural phenotypes, can cause neurological disease[22], may contribute to psychiatric disease and increases with aging[23].

When to go single-cell

When does it make economic sense to invest in sequencing single-cell genomes? Tumor genomes are highly heterogeneous and accrue mutations at widely varied rates, so they may be obvious targets for single-cell sequencing. While bulk tumor sequencing does not allow the unambiguous deconvolution of component clonal populations, it does point to genomic loci with sequence heterogeneity where targeted single-cell sequencing can reveal further detail. Such a staged approach greatly reduces sequencing costs, increasing the number of single cells that can be sequenced for a given tumor.

It is not clear that it will ever become cost effective to sequence entire genomes from large numbers of single cells from a given tumor, but most of the same benefits can come from analyzing important subsets of the genome or using shallow sequencing to measure lower-resolution copy number changes in these cells, in a manner quite reminiscent of that of Bridges and his *Drosophila* genomes from 80 years ago! An intriguing potential alternative to the staged approach is to use a single-step method in which exome sequences from many single cells are directly measured, thereby allowing the bulk tumor exome to be 'calculated' while also revealing the true clonal diversity within the tumor; this approach can be comparable in cost to whole-tumor sequencing.

Sequence before life

Single-cell sequencing is sometimes the only alternative for probing rare or unique cells. Preimplantation genetic diagnosis (PGD) is a procedure used for couples trying to conceive a child through *in vitro* fertilization: a single cell is extracted from the embryo before implantation and the contents of this cell's genome are analyzed. Although meta-analyses of older clinical trials found that PGD is not an effective way to screen for genetic disease[25], more modern procedures have had better success in randomly controlled trials and have shown that live birth rates can nearly double[26]. The application of genome-wide analysis methods, such as array comparative genomic hybridization, in this area opens up the possibility of higher-resolution measurements of the embryo genome before implantation[27]. We expect that higher-resolution genomic techniques will soon be applied to PGD, enabling the routine calling of structural variation and even point mutations in individual embryos. These data in turn will be used to make more careful assessments about which embryos are most likely to lead to successful pregnancies.

Future of the technology

The cost of sequencing will hopefully continue to decline. The past decade has also been a fertile one for the development of biochemical DNA amplifiers, and there are now several choices for single-cell experiments[5]. No single amplifier has emerged as the overall winner, and we would be surprised if that becomes the case. It is very difficult to point to a 'best' amplifier, since there are several different performance parameters to consider. In particular, one would like to understand the following as they relate to the specific application, sample type and reaction format: convenience (isothermal or thermocycled; single step or multiple reagent steps), cost (commercial or homemade), fidelity (off-target and contaminant amplification, uniformity or bias in amplification, coverage, error rates, artifacts such as chimeras) and gain (amount of amplification required).

Furthermore, comparing different amplification chemistries on a statistically relevant sample of single cells is a significant undertaking, and requires care to avoid confounding effects such as reaction volume, gain, reaction format, lysis conditions, contamination, sample-specific differences and random cell-to-cell variability. Such well-controlled comparisons are needed in order to identify the best matches among sample types, applications, amplification chemistries, reaction formats and sequencing approaches.

Finally, there is a continuing need for innovative technologies that automate single-cell isolation and genome amplification. The current state of the art operates at the level of hundreds of cells, and one can use commercially available cell sorters for the isolation steps, pipetting robots for cell lysis and amplification, or microfluidic devices that combine both steps in a fully automated and integrated fashion[28]. Automation and miniaturization are important to single-cell sequencing because sufficient sampling is important to fully characterize genomic diversity in the bulk material. We expect quite a bit of creativity in the development of arrayed and serialized microfluidic and microfabricated approaches. These will increase the throughput and decrease the cost of steps required for single-cell sequencing by orders of magnitude and will enable microbial and targeted human cell studies on thousands of cells in a single experiment. We believe that it is only a matter of time until large projects are launched which will systematically characterize the genomes and transcriptomes of hundreds of thousands of single cells.

Single-cell genomic analysis represents a suite of rapidly developing technologies that touch on a wide variety of fundamental and applied problems in the life sciences. We look forward to the continuing impact of single-cell sequencing as amplification chemistries and reaction formats diversify, and as the community innovates ways of applying this technology to extract information from biological systems.

Corrected after print 27 January 2014.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nmeth.2783).

1. Flemming, W. *Zellsubstanz, Kern und Zelltheilung* (Vogel, 1882).
2. Bridges, C.B. *J. Hered.* **26**, 60–64 (1935).
3. Hubert, R., MacDonald, M., Gusella, J. & Arnheim, N. *Nat. Genet.* **7**, 420–424 (1994).
4. Handyside, A.H. *et al. N. Engl. J. Med.* **327**, 905–909 (1992).
5. Blainey, P.C. *FEMS Microbiol. Rev.* **37**, 407–427 (2013).
6. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
7. Marcy, Y. *et al. Proc. Natl. Acad. Sci. USA* **104**, 11889–11894 (2007).
8. Marcy, Y. *et al. PLoS Genet.* **3**, 1702–1708 (2007).
9. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. *Genome Res.* **11**, 1095–1099 (2001).
10. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. *Nat. Biotechnol.* **29**, 51–57 (2011).
11. Peters, B.A. *et al. Nature* **487**, 190–195 (2012).
12. Lu, S. *et al. Science* **338**, 1627–1630 (2012).
13. Arnheim, N., Calabrese, P. & Tiemann-Boege, I. *Annu. Rev. Genet.* **41**, 369–399 (2007).
14. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. *Cell* **150**, 402–412 (2012).
15. Lynch, M. *Trends Genet.* **26**, 345–352 (2010).
16. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. *Science* **338**, 1622–1626 (2012).
17. Jan, M. *et al. Sci. Transl. Med.* **4**, 149ra118 (2012).
18. Shlush, L.I. *et al. Blood* **120**, 603–612 (2012).
19. Navin, N. *et al. Nature* **472**, 90–94 (2011).
20. Arendt, T. *Mol. Neurobiol.* **46**, 125–135 (2012).
21. McConnell, M.J. *et al. Science* **342**, 632–637 (2013).
22. Evrony, G.D. *et al. Cell* **151**, 483–496 (2012).
23. Gundry, M., Li, W., Maqbool, S.B. & Vijg, J. *Nucleic Acids Res.* **40**, 2032–2040 (2012).
24. Martin, S.L. *Nature* **460**, 1087–1088 (2009).
25. Mastenbroek, S., Twisk, M., van der Veen, F. & Repping, S. *Hum. Reprod. Update* **17**, 454–466 (2011).
26. Rubio, C. *et al. Fertil. Steril.* **99**, 1400–1407 (2013).
27. Yang, Z. *et al. Mol. Cytogenet.* **5**, 24 (2012).
28. Landry Z.C., Giovanonni S.J., Quake S.R. & Blainey P.C. *Methods Enzymol.* **531**, 61–90 (2013).

Erratum: Dissecting genomic diversity, one cell at a time

Paul C Blainey & Stephen R Quake

***Nat. Methods* 11, 19–21 (2014); published online 30 December 2013; corrected after print 27 January 2014**

In the version of this article initially published, references 16–24 were incorrectly cited as references 15–23. The error has been corrected in the HTML and PDF versions of the article.

Entering the era of single-cell transcriptomics in biology and medicine

Rickard Sandberg

Recent technical advances have enabled RNA sequencing (RNA-seq) in single cells. Exploratory studies have already led to insights into the dynamics of differentiation, cellular responses to stimulation and the stochastic nature of transcription. We are entering an era of single-cell transcriptomics that holds promise to substantially impact biology and medicine.

Our notion of transcriptomes has been forged mainly by population-level observations that have been the mainstream in biology over the last two decades. We are used to thinking about differences in expression in terms of graded or subtle fold changes when comparing data across entire tissues or conditions. But the actual differences between cells may be far larger. Subsets of cells may experience dramatic changes that are averaged out or diluted by the presence of a large number of nonresponsive cells. In fact, it was shown over 60 years ago that inductive cues often result in all-or-none responses in single cells but these responses are observed as a gradual increase when quantified across the population[1].

It is clear that assessing gene expression in single cells is critical to better understand cellular behaviors and compositions in developing, adult and pathological tissues. To this end, a long-standing goal has been to enable genome-wide RNA profiling, or transcriptomics, in single cells[2,3]. Only recently has the technology matured so that biologically meaningful differences can be robustly detected with single-cell RNA-seq. Detailed protocols[4–6] for sequencing library preparations and the introduction of commercial automation (for example, Fluidigm C1) have lowered the barriers for researchers to access these methods. Widespread adoption of these techniques will have a major impact on our understanding and appreciation of cellular states, the nature of transcription and gene regulation, and our ability to characterize pathological states in disease.

Rickard Sandberg is at the Ludwig Institute for Cancer Research, Stockholm, Sweden, and Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.
e-mail: rickard.sandberg@ki.se

Above the noise

Single-cell transcriptomics relies on the reverse transcription of RNA to complementary DNA and subsequent amplification by PCR or *in vitro* transcription before deep sequencing—procedures prone to losses or biases. The biases are exaggerated by the need for very high amplification from the small amounts of RNA found in an individual cell. Although technical noise confounds precise measurements of low-abundance transcripts, modern protocols have progressed to the point that single-cell measurements are rich in biological information. For example, a recurrent theme in single-cell transcriptome studies is that cells reliably group by their cell type or state when subjected to unsupervised clustering[7–10]. Gene expression associated with cell identity or developmental stages thus has a stronger signal than technical noise or biological variability related to dynamic processes such as phase of the cell cycle. Moreover, the power to detect meaningful biological differences from single-cell data is demonstrated by the identification of hundreds to thousands of genes with differences in abundances between cell types[7,9]. Recent refinements will improve the signal-to-noise ratio even further by enhancing the efficiencies of reverse transcription and PCR[11] or applying molecular barcoding strategies that control for amplification bias[12].

Challenges in single-cell transcriptomics

Currently available single-cell RNA-seq methods were developed with several different objectives. Full-length transcripts can be profiled, such that sequence reads cover the entire gene to quantify both gene and transcript isoforms and also monitor single-nucleotide polymorphisms and mutations[9,11]. In contrast, tag-based sequencing of 5′ or 3′ ends[10,13] provides only an estimate of transcript abundance at the cost of coverage across gene structures but allows the assay to be scaled up and combined with molecule counting[12].

The unified goal in the field is to develop cost-effective, high-throughput methods that detect all RNA present inside the cell at full-length RNA coverage. Lowering RNA losses and enhancing the conversion of RNA to cDNA before amplification are areas where further development would boost RNA detection. Another important goal is to augment procedures for the dissociation, sorting and picking of individual cells[14] so that complex tissues can be dissociated into single-cell suspension without inducing changes in gene expression related to cell handling or picking. Finally, simultaneous detection of poly(A)$^+$ and poly(A)$^-$ RNA, irrespective of transcript length, and RNA modifications (for example, m^{6}A in ref. 15) are also desirable features for future development.

One of the mind-boggling features of transcription that only becomes apparent in single-cell analysis is that expression of a gene that is reliably detected in a population may be anywhere from absent, to low, to high in a given cell because of random fluctuations. Such variability may be explained by models that describe transcription as occurring in discrete bursts[16] driven by stochastic molecular processes. The stochastic nature of transcription has been studied in greatest detail in prokaryotes and unicellular eukaryotes[16], but more and more lines of evidence point to similar phenomena in mammalian cells[17,18]. We must therefore take into account such transcriptional behavior in our strategies for analyzing single-cell transcriptome data and in our biological interpretation of the results. For example, standard differential expression tests might not be suitable for single-cell data that contain a fair number of cells with no detectable expression. Indeed, new tests have been proposed[19] that combine differences in transcript abundance with differences in the fraction of cells with expression.

Single-cell transcriptome studies to date require cells in suspension (for example, dissociated tissues or cultures) so that the spatial organization of the population is often lost, unless cells had been picked from defined areas. Spatial information can be recovered to some extent through RNA *in situ* hybridization analyses of marker genes for identified cell types, allowing cell type–specific expression profiles to be projected onto complex tissue structures. However, methods that simultaneously capture spatial structures and transcriptome-wide profiles at single-cell resolution are being developed but have yet to be described (for example, building on *in situ* sequencing or array-based multiplexing strategies). The ability to perform spatial single-cell transcriptomics on developing, adult or pathological tissues promises to dramatically elevate our understanding of life and disease, revealing the transcriptomes related to specific states of intercellular communication, polarity formation and local gradients.

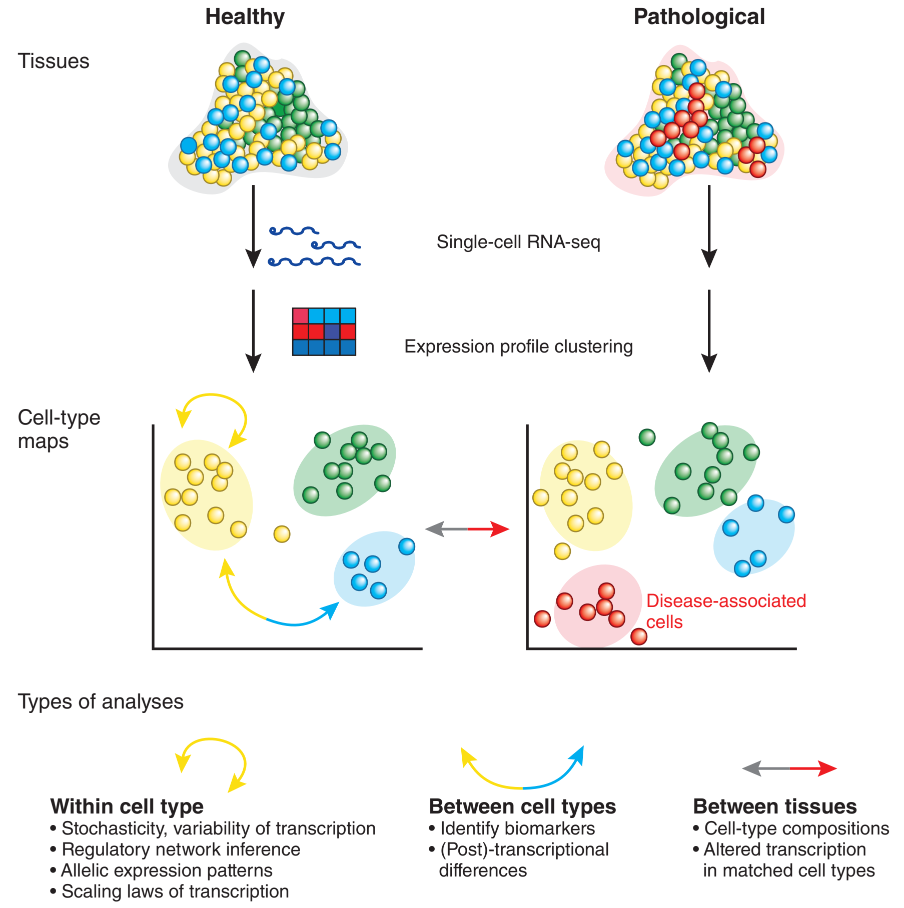

Healthy
Pathological
Tissues
Single-cell RNA-seq
Expression profile clustering
Cell-type maps
Disease-associated cells
Types of analyses
Within cell type
• Stochasticity, variability of transcription
• Regulatory network inference
• Allelic expression patterns
• Scaling laws of transcription
Between cell types
• Identify biomarkers
• (Post)-transcriptional differences
Between tissues
• Cell-type compositions
• Altered transcription in matched cell types

Figure 1 | Single-cell transcriptome analyses of tissues and cell types. Cells from a healthy or pathological tissue are dissociated, analyzed independently with single-cell RNA-seq and clustered based on their gene expression profiles. Clustering of cells reveals a cell-type map that can be used to assess the composition of the tissue including the identification of new cell types or subtypes. These rich data can be used to address many questions of gene expression and regulation within or between cell types and between tissues.

Implications for biology

The measurement of gene expression in single cells will revolutionize our understanding of gene regulation and resolve many longstanding debates in biology. Cells cluster by cell type or developmental state when grouped according to their expression profiles[7–10]. Thus, expression-based clustering allows for the unbiased reconstruction or 'reverse engineering' of cell types in any population or tissue after sequencing enough individual cells (**Fig. 1**). If the sampling of cells is extensive and sufficiently free from biases, such clustering can reveal all cell types present, including new ones. All cells in a cluster can also be used to derive robust cell-type expression profiles, again in a data-driven manner and without previous knowledge of which marker genes define a tissue or cell type. Single-cell profiling of RNAs is therefore the first method that could lay a foundation for a quantitative, data-driven classification of cell types.

Single-cell transcriptomics will also enable high-resolution transcriptional maps of both stable and transient cellular states during differentiation or reprogramming. Important for these aims is to sample sufficient individual cells that span the entire process, so that analyses can later zoom in on the subset of cells at critical bifurcation points of differentiation. The sample size should reflect how often cell types or events are expected to occur. Also, it is debated to what extent the human genome is transcribed, as several studies have identified very rare transcripts (for example, those present in one copy per 10,000 cells)[20]. These transcripts could either be expressed at high levels in rare cells (for example, ten copies in one of 100,000 cells) or have low (leaky) expression in a larger subset of cells. Analyses across hundreds or thousands of individual cells will likely resolve these questions and improve our understanding of cellular transcriptional landscapes and regulatory networks.

RNA-seq analyses across human tissues and cell populations have demonstrated the pervasive use of RNA processing to diversify the transcriptome and the proteome[21]. A large fraction of differences are subtle when comparing tissues, but it is possible that patterns of alternative splicing, polyadenylation and transcription start-site usage will have a more bimodal (on or off) distribution at single-cell resolution, as suggested by a pioneering study on single cells[22]. Studies of the regulation of alternative polyadenylation have revealed a general shortening of 3′ untranslated regions in more highly proliferating cells[23] and in transformed cells *in vitro*[24]. Analyses of *in vivo* tumors would benefit greatly from single-cell RNA-seq to separately extract transcript abundance and isoform information from the mixture of transformed cells, stroma and other infiltrating cells. Single-cell transcriptomics of dissociated tumor and healthy tissues will enable the precise identification of mRNA isoforms that are important for the transformed state.

Implications for medicine

Transcriptomic approaches in medicine are often based on comparing pathological with matched healthy tissue[25] or analyzing a large number of pathological tissues to find subclassifications[26]. Cancer tissues are often characterized by changes in both cellular compositions (for example, infiltrating immune cells) and alterations in gene-expression programs in both the transformed cells and the surrounding stroma. Thus, observations at the tissue level contain several differential expression profiles superimposed on top of each other. High-throughput single-cell analysis of pathological tissues would simultaneously monitor changes in cellular composition (based on clustering) and associated gene expression profiles[27]. Comparisons could then be made between specific cell types observed in both the healthy and pathological tissues to reveal more precise gene expression programs of disease (**Fig. 1**). However, regional variations in cellular composition may necessitate sampling in multiple regions from the same tumor[28].

Areas of research that stand to benefit in particular from single-cell transcriptomics are those in which the clinically relevant cells are too rare to be studied using population-level techniques. For example, only a few circulating tumor cells (CTCs) are typically present in a milliliter of blood, which has precluded their genome-wide profiling. Two pioneering studies demonstrated the utility of single-cell RNA-seq analyses of CTCs of melanoma[9] or pancreatic[29] origin, as the transcriptome profiles both validated the cellular isolation procedure and were used to identify alterations in the gene expression programs. Single-cell RNA-seq with full-length transcript coverage[11] should enable simultaneous measurement of gene expression programs and detection of mutations that arise in the tumor through analyses of the CTCs. Transcriptome analyses of single CTCs is a noninvasive strategy to select treatment based on the inferred mutations[30] and also to monitor the development of drug resistance. It is time to determine to what extent CTC transcriptome profiling can be a future method for cancer diagnostics and treatment selection, and provide biomarkers for future therapies targeting CTCs.

Outlook

As we are just entering an era of single-cell transcriptomics, the near future will likely unravel many surprising and new characteristics of transcriptomes. It will be interesting to investigate whether certain scaling laws exist between RNA abundance profiles and cellular phenotypes such as cell or nucleus size. For example, to maintain protein concentrations inside membranes or subcellular compartments in cells of varying size, different abundances would be needed as volume and area scale differently with cell size. Sets of genes are likely to scale with characteristics such as plasma or nuclear membrane area, cytoplasmic volume and nuclear volume. Only with such knowledge at hand can we begin to resolve how cellular heterogeneity and cell type composition confound population-level transcriptome analyses. For example, comparisons of two tissues composed of cells of differing size might reveal differences in expression related to size, rather than the differences of interest. A better understanding of single-cell expression profiles will also provide a more rational basis for the design of future studies at the most appropriate level of resolution (for example, tissue, cell type, single cell or combinations of the three).

With the maturation of single-cell transcriptomics, I expect that studies of gene expression and regulation in single cells will boom in the coming years and the research community will soon obtain precise transcript-isoform quantifications across hundreds of thousands to even millions of individual cells. This information will answer many outstanding questions (**Fig. 1**) and lay the foundation for a quantitative definition of cell types and their variation in homogeneous as well as heterogeneous cell populations. Based on this knowledge it will become feasible to determine the transcriptome profiles of nearly all cell types in complex multicellular organisms. Single-cell profiling will also dramatically improve gene-regulatory network inferences[31], as the vast amounts of single-cell profiles are bona fide biological perturbations that should improve the power of inference.

ACKNOWLEDGMENTS
I am grateful to B. Reinius, J. Muhr, J. Holmberg and members of the Sandberg laboratory for comments on the manuscript. R.S. is supported by the European Research Council Starting grant 243066, the Swedish Foundation for Strategic Research FFL4 and the Swedish Research Council grant 2008-4562.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Novick, A. & Weiner, M. *Proc. Natl. Acad. Sci. USA* **43**, 553–566 (1957).
2. Brady, G., Barbara, M. & Iscove, N.N. *Methods Mol. Biol.* **2**, 17–25 (1990).
3. Eberwine, J. *et al. Proc. Natl. Acad. Sci. USA* **89**, 3010–3014 (1992).
4. Islam, S. *et al. Nat. Protoc.* **7**, 813–828 (2012).
5. Tang, F. *et al. Nat. Protoc.* **5**, 516–535 (2010).
6. Picelli, S. *et al. Nat. Protocols*, doi:10.1038/nprot.2014.006 (2 January 2014).
7. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
8. Tang, F. *et al. Cell Stem Cell* **6**, 468–478 (2010).
9. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
10. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
11. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
12. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep* **2**, 666–673 (2012).
14. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
15. Dominissini, D. *et al. Nature* **485**, 201–206 (2012).
16. Raj, A. & van Oudenaarden, A. *Cell* **135**, 216–226 (2008).
17. Chubb, J.R., Trcek, T., Shenoy, S.M. & Singer, R.H. *Curr. Biol.* **16**, 1018–1025 (2006).
18. Wills, Q.F. *et al. Nat. Biotechnol.* **31**, 748–752 (2013).
19. McDavid, A. *et al. Bioinformatics* **29**, 461–467 (2013).
20. Mercer, T.R. *et al. Nat. Biotechnol.* **30**, 99–104 (2012).
21. Wang, E.T. *et al. Nature* **456**, 470–476 (2008).
22. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
23. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. *Science* **320**, 1643–1647 (2008).
24. Mayr, C. & Bartel, D.P. *Cell* **138**, 673–684 (2009).
25. Rhodes, D.R. *et al. Proc. Natl. Acad. Sci. USA* **101**, 9309–9314 (2004).
26. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
27. Dalerba, P. *et al. Nat. Biotechnol.* **29**, 1120–1127 (2011).
28. Gerlinger, M. *et al. N. Engl. J. Med.* **366**, 883–892 (2012).
29. Yu, M. *et al. Nature* **487**, 510–513 (2012).
30. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
31. Kim, H.D., Shay, T., O'Shea, E.K. & Regev, A. *Science* **325**, 429–432 (2009).

The promise of single-cell sequencing

James Eberwine[1,2], Jai-Yoon Sul[1], Tamas Bartfai[3] & Junhyong Kim[2,4]

Individual cells of the same phenotype are commonly viewed as identical functional units of a tissue or organ. However, the deep sequencing of DNA and RNA from single cells suggests a more complex ecology of heterogeneous cell states that together produce emergent system-level function. Continuing development of high-content, real-time, multimodal single-cell measurement technologies will lead to the ultimate goal of understanding the function of an individual cell in the context of its microenvironment.

Since 1665, when Robert Hooke used the term "cell" to describe a structure in cork that he observed under his microscope, cells have been objects of intense study. Although the existence of distinct cell types was obvious from the earliest morphological studies, recent advances are revealing a surprising diversity of individual cell states. A typical human cell contains an estimated 6 billion base pairs of DNA and 600 million bases of mRNA—an immense capacity for coding function. Deep sequencing of DNA and RNA from single cells can read these blueprints for cellular function more comprehensively and at higher resolution than previously possible[1,2]. Such specificity in our recognition of cell states provides hope for a better understanding of cell function and dysfunction.

The higher resolution of cellular differences detected by single-cell sequencing also raises a host of new questions. Perhaps most fundamental is that measuring differences does not mean that these differences have consequences: which are the 'functional' cell states? Given that a typical human cell contains just tens of each mRNA on average, do these small numbers of molecules participate in cellular regulation, such as the precise dynamics seen in early development? How single cells interact to produce emergent function at the tissue level, and the nature of this cellular ecology, is fertile territory for exploration. Furthermore, if we posit that cellular phenotype is a function of the local ecology of individual cells[3], how many such distinct ecologies exist in a multicellular individual, and are they interchangeable (**Fig. 1**)?

Although there is tremendous excitement about single-cell sequencing, it is still not a routine experimental procedure. Improvements in the basic technology as well as in data analysis and interpretation will be important for obtaining the precision of measurement and large sample sizes needed to understand the role of individual cells in their system-level function. We present some of these issues in this Commentary, and highlight future directions and emerging technologies that complement sequencing and promise to further expand our knowledge of how single cells function within their ecological context.

Critical considerations for single cells

Several important considerations influence the quality of data generated from a single cell. Of particular note is the inevitable problem that the transcriptome will change in response to manipulation, which is likely to be more acute in individual cells. With this consideration in mind, single-cell transcriptome data should be interpreted partly as a perturbation experiment until less disruptive RNA isolation methods can be developed.

Cell isolation. The isolation of single cells is the single-cell technique that is arguably in greatest need of development and standardization. Using a patch pipette or nanotube to harvest the cytoplasmic contents of single cells is a common method for the isolation of cellular RNA, but it may leave cellular subcompartments behind. Microfluidic devices or similar tools can capture single cells in isolated reaction chambers but require detaching cells from substrates, which will perturb transcriptional states. Altered transcriptional states are also a concern for cells that are dissociated and enriched by cell sorting. Dispersed cultured cells are the easiest to isolate, but when using these one must carefully craft experimental questions so that the lack of microenvironmental influences does not pose an interpretation problem. Ideally, we need procedures to isolate the contents of a cell that is still within its tissue matrix or natural microenvironment. In this way, mRNA measurements will reflect the natural state of a cell in a population context and also show minimal transcriptional changes due to manipulation.

Amplification. The classic constraint of single-cell approaches—in the absence of mature and robust single-molecule sequencing technologies—is the need for substantial amplification, which may misrepresent the original DNA sequence or RNA population. The problem is especially acute when working with DNA, where only a single molecule is available. For DNA, the main problem is coverage. Extensive PCR-based amplification may yield higher coverage, but this is typically at the expense of uneven representation and error amplification. Both error correction and the detection of single-nucleotide variants

[1]Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]The Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California, USA. [4]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. e-mail: eberwine@mail.med.upenn.edu

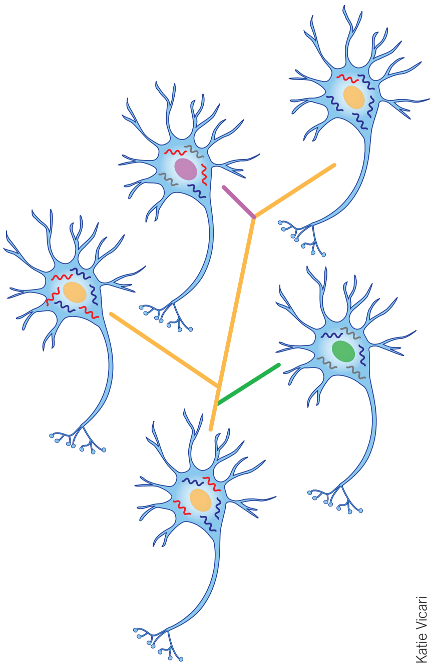

Katie Vicari

Figure 1 | A diversity of individual cell states revealed by single-cell sequencing. The schematic shows how sequence differences (colors) in DNA (nuclei) and RNA (curved lines) often distinguish cells that are related or seem identical.

will require additional statistical methods. Error correction will be especially difficult for single cells, as there is no good control for a single cell's genomic DNA variability, and a priori it is impossible to know how much DNA variation to expect between single cells.

For RNA, the challenge is to maintain the initial relative abundances of cellular RNAs through the amplification process. The first step in amplifying RNA is its conversion into complementary DNA (cDNA) by reverse transcriptase (RT). This is the most critical factor in single-cell transcriptomics, as the efficiency of the RT reaction will dictate the percentage of a cell's RNA population that is ultimately analyzed by the sequencer. RT was originally isolated from a picornavirus-infected mammalian cell; it is highly efficient as only one copy of the virus RNA is present in the host cell, which must be copied to its full length. Although the processivity of this reaction *in vitro* has been reported to be as low as 10% of full length, it can reach as high as 90% after optimization. Mutagenesis of recombinant RT may enable longer cDNA production, which is especially important with suboptimal concentrations of RNA.

Single-cell PCR permits the exponential amplification of the cDNA copy of RNA. Although many studies have used PCR to make libraries[4], one must be conscious of the fact that biases in PCR efficiency for particular sequences (for example, GC content and snapback structures) will also be amplified exponentially. Most investigators try to limit the number of PCR cycles in an effort to reduce this error. However, because the bias is likely to be sequence specific and gene expression is variable to begin with, the degree of bias may be difficult to predict. Linear amplification based on *in vitro* transcription of cDNA into amplified RNA (aRNA) alleviates some of these problems[5,6], although it is conceivable that specific sequences are transcribed inefficiently, resulting in shorter amplified sequences or some sequence drop-out. Shorter sequences may not be an issue if the goal is quantification of RNAs rather than splice-variant analysis and relative amounts of the amplified products are maintained. Sequencing *in vitro*–diluted control RNA and comparing read counts to Poisson distribution suggests that an expected resolution of two to four molecules can be quantitatively achieved with aRNA, though performance depends on recovery as well as amplification.

One idea to overcome amplification bias is to incorporate unique sequence tags into the first cDNA product. Given a large enough diversity of tags, each cDNA copy of an individual RNA could be labeled by a unique tag sequence; after PCR, distortions from amplification would not affect the counting of tags (unless there was sequence dropout), which reflect the original number of template RNA molecules in the cell[7]. However, the protocol for such digital tagging is difficult and is currently still being optimized.

Dynamic range and number of cells. Current estimates suggest that 5,000-15,000 different genes are transcribed in a typical mammalian cell. If we think of each as a variable, ideally we would want 10- to 30-fold more measurements than the degrees of freedom to characterize the covariance of the transcriptome, or more if the variation is nonlinear and complex. The degrees of freedom for the transcriptome of a single cell is an open problem, but it is likely to be at least in the thousands, suggesting the need to measure tens of thousands of cells. Projects of this magnitude are currently ongoing but limited to a small number of target molecules at low sequencing coverage. Therefore, an important question is how to determine the number of cells that need to be measured in order to obtain adequate coverage of the transcriptome.

Various estimates suggest that the most highly expressed genes have steady-state values of 3,000–5,000 molecules per cell. But current single-cell transcriptome data from the literature and from our own labs suggest that 90% of the transcriptome is expressed at less than 50 molecules per cell. The key question is whether such low-level expression is critical to a cell's function and phenotype. What is clear is that many genes show binary 'on' and 'off' states that vary across individual cells in a population, and many of the weakly expressed genes are never seen in tissue-level measurements. The complement of genes with fewer than 50 transcripts per cell include many critical regulators such as transcription factors and signal-transduction proteins. Thus, the sensitivity issue cannot be ignored, and fully covering the dynamic range of individual transcriptomes is just as important as obtaining data from a large number of cells.

Gaining spatial context

One method to assess RNA within cells in their natural microenvironment is fluorescence *in situ* hybridization (FISH). Current implementations of FISH typically use multiple short fluorescence-labeled probes that can diffuse into tissues and anneal to target RNA[8]. While there have been great advances in sensitivity, it is difficult to be confident of the selectivity of hybridization (as is the case for microarrays), and it is unclear how much RNA is available for hybridization after cellular cross-linking. Most importantly, the limited number of fluorescent molecules with distinct emission spectra are incapable of simultaneously measuring 'transcriptome-scale' numbers of RNAs. Current probe multiplexing is reported to identify up to ~30 different mRNAs in a cell, which is already a vast improvement over previous FISH approaches.

Several groups are developing *in situ* sequencing and combinatorial tagging methods, but even if the RNA were evenly spaced, only a maximum of ~13,000 total spots or pixels can be distinguished (given two tissue sections through a typical mammalian cell of 20 × 20 μm, at 250-nm optical resolution), a fraction of the estimated 100,000 to 300,000 mRNA molecules in a cell. Regardless, the ability to assess spatial resolution for multiple RNAs provides additional biological insight into cellular function and phenotype.

Beyond genomes and transcriptomes

The transcriptome is usually used as a surrogate to infer the functional proteome of cells. The relationship between mRNA and protein abundances is not clear, and methods that permit a direct correlation of the transcriptome with the functional proteome are needed. The chemical complexity of proteins has made them significantly more difficult to quantify than RNA; however, as mass spectrometry becomes more sensitive and better ways of volatizing proteins are developed, there is hope that this technology will permit analyses of protein mixtures at the level of single cells. Alternatively, as tighter-binding antibodies, antibody derivatives (nanobodies, single-chain variable fragments) or aptamers are developed, there is hope that the greater sensitivity offered by such enhanced affinity technologies will permit single-cell proteomics to eventually become a reality.

We also need to expand single-cell measurements to other genomic regulatory features beyond sequence, including DNA structural states of the epigenome. Chromosomal conformation, DNA methylation, open chromatin and small-molecule metabolome assays are all moving toward single cell-level detection. Ideally, what is needed is a real-time, live-cell multiplex measurement of sufficient variables within each cell's tissue context regardless of molecule type, such that we will have a true picture of the multidimensional spatial dynamics of the cellular ecology. For RNA, this might be accomplished by single-molecule detection of transcription as it is occurring in live cells. Such measurements will show the molecular building blocks upon which biology occurs and will lead to a fuller understanding of biological processes.

Taking a step beyond measurement, we need to perform perturbations at the single-cell level to dynamically probe cell function. Using a population of RNA as a modulator of cellular function may lead to functional insights and perhaps have therapeutic potential. The ability to transfect quantitatively titrated pools of RNA was first described as the transcriptome-induced phenotype remodeling (TIPeR) methodology[9]. Whole transcriptomes or groups of RNAs introduced into cells using the approach have induced a change in cellular phenotypes toward that of target cells. The idea behind TIPeR is that the transfer of RNA memory can be used to create cells of specific function; the methodology has been used as a functional genomics approach to modulate cellular function[10] and phenotype[11–13]. The ability to measure and quantitatively manipulate the transcriptome will allow us to modulate cell phenotypes for both basic research and therapeutic purposes.

Prospects for single-cell biology

At the level of individual cells, all diseases show heterogeneity in their pathology. Single-cell studies may lead to a better understanding of why some cells degenerate while adjacent cells are normal, or why some cells are drug responsive but others are not. In many cases, the cells or tissues that are most affected by a disease or control its onset and severity have been identified. Pinpointing the molecular states specific to disease will help to identify and exploit drug targets, but achieving this depends on how well we can characterize 'pathology-important cells'.

For example, we know that dopaminergic neurons lose their ability to synthesize and secrete dopamine and subsequently die during the progression of Parkinson's disease. Every receptor, ion channel or transporter that is identified specifically in these neurons may be targeted by drugs to assist in slowing disease progression and treating symptoms. Currently, pharmacological approaches exploit only four proteins from these cells (the Dopa transporter, muscarinic receptor M1, monoamine oxidase (MAO) and the adenosine A2A receptor). Previous tissue-level studies highlighted druggable targets, but many were not present in the cells of interest. The sensitivity and specificity provided by single-cell studies have shown that as many as 300–400 druggable genes are expressed in many cell types. Presuming this is true for Parkinson's-affected neurons, one might expect 30–40 genes to be selectively expressed in these neurons and at different stages of this decades-long disease.

Beyond its translational applications, single-cell analysis has the potential to fundamentally change our view of how multicellular organisms work and to generate new research questions. How many distinct cell types are there among the 100 trillion cells of the human body? What is the role of somatic DNA alteration in cell identity and diversity? If somatic changes are prevalent, are they random or part of a genomic program? Is the phenotype of a cell programmed by its genome or the result of community-coupled cell-state dynamics? That is, to use a metaphor, is DNA the program or just informational storage[14]?

Microbiome sequencing data increasingly suggest that single-celled microbes may be integral to the multicellular host body[15,16]. At the other end, sequencing the DNA and RNA of individual cells from tissues of multicellular organisms is suggesting much greater heterogeneity of these cells. This raises the possibility that the cells of a multicellular body are not so much the uniform units of tissues; rather, tissues and organs might be functionally coherent assemblies arising from ecologies of cells, whose interactions characterize the system-level phenotype, in a fashion akin to that for microbiome data. If this is an organizing principle across species, then the characterization of single cells, their diversity and their ecology will be an undeniable imperative for understanding the individual organism.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nmeth.2769).

1. Navin, N. *et al. Nature* **472**, 90–94 (2011).
2. Xu, X. *et al. Cell* **148**, 886–895 (2012).
3. Buss, L.W. *The Evolution of Individuality* (Princeton Univ. Press, 1987).
4. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
5. Buckley, P.T. *et al. Neuron* **69**, 877–884 (2011).
6. Hashimshony, T. *et al. Cell Rep.* **2**, 666–673 (2012).
7. Shiroguchi, K. *et al. Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
8. Singer, R.H. & Ward, D.C. *Proc. Natl. Acad. Sci. USA* **79**, 7331–7335 (1982).
9. Sul, J.Y. *et al. Proc. Natl. Acad. Sci. USA* **106**, 7624–7629 (2009).
10. Kim, T.K. *et al. Cell Rep.* **5**, 114–125 (2013).
11. Yakubov, E. *et al. Biochem. Biophys. Res. Commun.* **394**, 189–193 (2010).
12. Sul, J.Y. *et al. Trends Biotechnol.* **30**, 243–249 (2012).
13. Zangi, L. *et al. Nat. Biotechnol.* **31**, 898–907 (2013).
14. Kim, J. & Eberwine, J. *Trends Cell Biol.* **20**, 311–318 (2010).
15. Giannoukos, G. *et al. Genome Biol.* **13**, R23 (2012).
16. Turnbaugh, P.J. *et al. Proc. Natl. Acad. Sci. USA* **107**, 7503–7508 (2010).