Commentary | Published: 30 December 2013

The promise of single-cell sequencing

James Eberwine, Jai-Yoon Sul, Tamas Bartfai & Junhyong Kim

Nature Methods **volume 11**, pages 25–27 (2014) | Download Citation

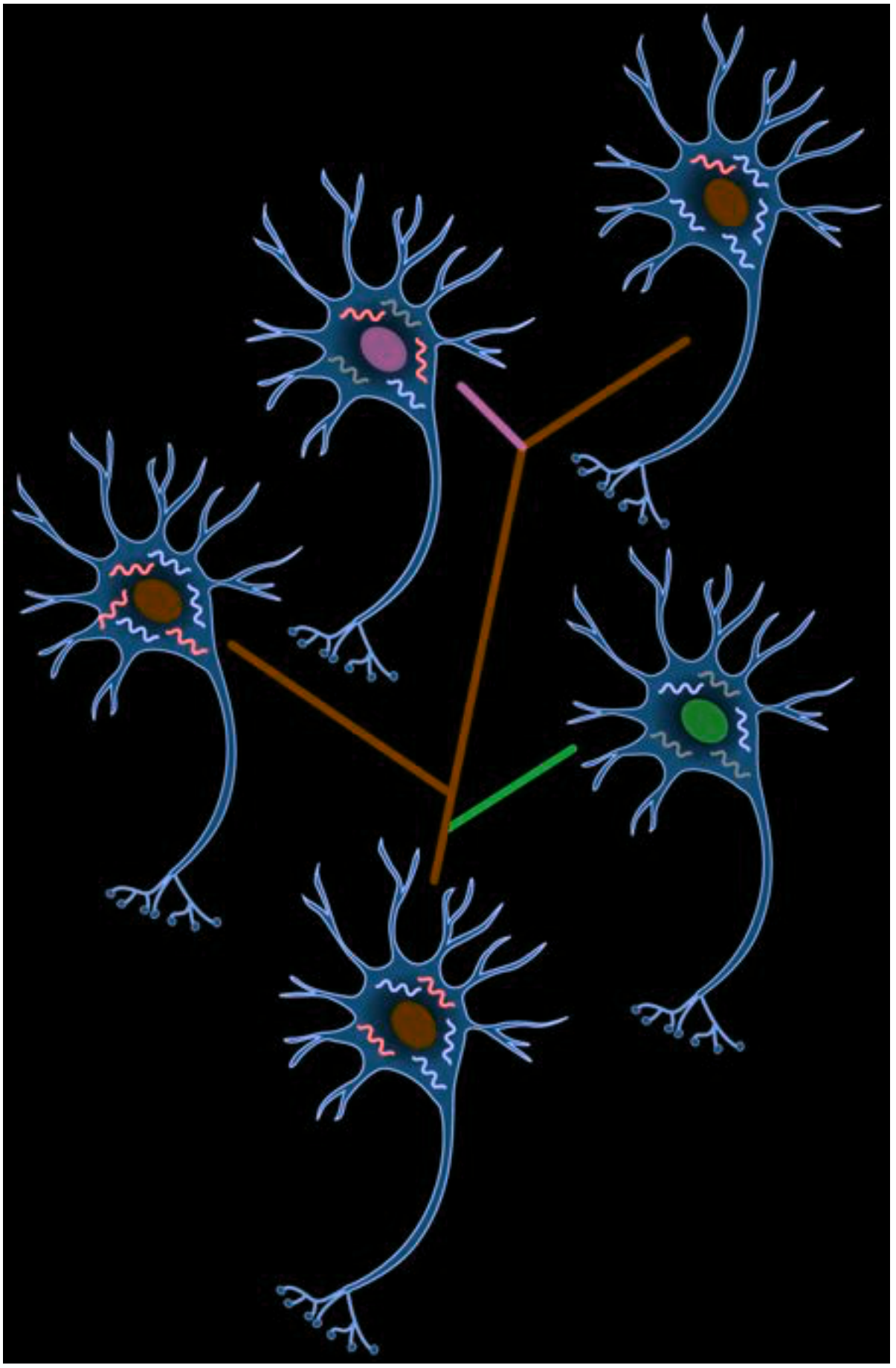
Individual cells of the same phenotype are commonly viewed as identical functional units of a tissue or organ. However, the deep sequencing of DNA and RNA from single cells suggests a more complex ecology of heterogeneous cell states that together produce emergent system-level function. Continuing development 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)?

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. *Image: Katie Vicari*

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 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[11,12,13]. 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.

ReferencesReferences

1.

Navin, N. *et al. Nature* **472**, 90–94 (2011).

2.

X... Y... *et al. Cell* **143**, 826–835 (2010).

Author informationAuthor

information

Rights and permissionsRights and permissions

About this articleAbout this article

Further reading Further reading

Nature Methods

ISSN 1548-7105 (online)