Published in final edited form as:

Trends Mol Med. 2017 July ; 23(7): 583–593. doi:10.1016/j.molmed.2017.05.004.

De Novo Gene Expression Reconstruction in Space

Je H. Lee

Abstract

The biological function of a gene often depends on spatial context, and an atlas of transcriptional regulation could be instrumental in defining functional elements across the genome. Despite recent advances in single-cell RNA sequencing and *in situ* RNA imaging, fundamental barriers limit the speed, genome-wide coverage, and resolution of *de novo* transcriptome assembly in space. Here, we discuss potential next-generation approaches for *de novo* assembly of the transcriptome in space, and propose more efficient methods of detecting long-range spatial variations in gene expression. Finally, we discuss future *in situ* sequencing chemistries for visualizing biological pathways and processes in tissues so that genomics technologies might be more easily applied to conditions of human health and disease.

Keywords

gene expression atlas; single cell RNA sequencing; multiplex single-molecule FISH; *in situ* sequencing; FISSEQ

Gene expression in multi-cellular organisms is cell-cell interaction- and context-dependent, and the observed transcriptional signature often depends on the degree of cellular heterogeneity and on sampling size. Therefore, gene expression profiling with single-cell resolution accompanied by spatial information is much needed to investigate the regulatory relationship between cells in development and cancer[1–3]. Recent advances in single-cell RNA sequencing now permit gene expression profiling in as many as 10^4 single cells[4–8]. In addition, newer single-molecule *in situ* RNA hybridization methods can visualize up to 10^3 genes simultaneously[9–13]. Together, it may be possible to identify the major cell types and pin their location in tissues in an unbiased manner. Unsurprisingly, the current trend is to scale these technologies using miniaturization, multiplexing, automation, and computational analysis[14]. Certainly, developing effective strategies for producing a comprehensive map of molecular signatures, especially in clinical specimens, could represent a major breakthrough.

In this Opinion article, we discuss why improvements in the existing approaches might be insufficient for studying cell-cell interactions driving cellular decisions *in vivo*, and we postulate on technologies needed to efficiently assemble a gene expression atlas *de novo*.

Correspondence: jlee@cshl.edu (J. Lee).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Here, we draw comparisons to issues encountered in reference genome assembly projects. Given the degree of overlap in cell types and states, the noisy nature of single-cell measurements[15, 16], and the similarities among many tissue niches, we describe various challenges in assembling or mapping gene expression data in space. We hypothesize on ways to incorporate different categories of spatial information into sequencing or sequential hybridization reactions in order to efficiently assemble a gene expression atlas *de novo*. Our aim is to highlight conceptual and technological parameters in different phases and applications of *de novo* gene expression assembly and to outline a path toward technology development capable of addressing key questions in biology.

Profiling Gene Expression in Space

Comparative Spatial Gene Expression Analysis

Over the years, researchers have developed multiple strategies for isolating RNA from tissues using morphologic features, molecular biomarkers, or genetically engineered reporters. Typically, tissues-of-interest (or cells in a tissue) are surgically or enzymatically separated for RNA extraction, destroying the tissue and averaging the cellular heterogeneity within the sampled region (Fig. 1A and 1B)[17]. Then, the isolated tissues are sorted into chambers from which nucleic acids are extracted for sequencing analysis. Alternatively, the RNA can be labeled *in situ* or *in vivo*, followed by the isolation and sequencing of the purified RNA *ex vivo*[18–20]; however, these approaches are labor-intensive and low-throughput, as well as being dependent on the availability of specific features or biomarkers that encapsulate the essence of the desired tissue region. In addition, they do not examine the surrounding cells or regions, introducing a potential confounding factor during comparative gene expression analysis.

Conceptual Parameters in Gene Expression Atlas Assembly

In order to interpret a particular molecular signature in the global tissue context, a comprehensive whole tissue atlas is needed. For this reason, the Allen Brain Atlas utilized industrialized implementation of *in situ* hybridization (ISH), interrogating 2000 genes across thousands of mouse brain sections from a specific developmental stage[21, 22] (Fig. 1C and 1D). Despite the semi-qualitative nature, such atlases had a significant impact on neurobiology and inspired multiple computational and experimental methods. Because this approach requires a significant investment in capital, researchers are utilizing single-cell RNA sequencing and single-molecule RNA ISH technologies (e.g. RNA-fluorescent *in situ* hybridization or RNA-FISH) as a potentially more efficient and scalable alternative in assembling gene expression atlases[23].

However, using dissociated single-cell data for spatial reconstruction[23–25] faces a number of fundamental barriers that require careful planning. Although the representative gene expression signature can be identified using single-cell RNA sequencing methods, computationally assembling or mapping such signatures in space requires a pre-existing reference atlas[23]. Even with a comprehensive reference atlas, cells of different state, function, and location can be assigned to a single gene expression cluster due to the noisy

nature of single-cell RNA sequencing. Consequently, the spatial mapping remains especially difficult in large tissues containing many degenerate or similar cellular elements (Fig. 2A).

Over the past fifty years, different types of DNA sequencing technologies were developed to assemble reference genomes. From the earliest days, the impact of the genome size, the complexity, and the physical genetic distance was well appreciated[26, 27]. In fact, the development of statistical concepts, computational tools, and technologies to address these concerns were central to the completion of many reference genome projects[28]. For instance, many short sequencing-reads (i.e. 50–300 bases) from modern DNA sequencers cannot be mapped uniquely due to the sheer number of highly similar sequences in the genome (Fig. 2B). However, if one introduces a spacer (i.e. ~400–10,000 bases) between a pair of sequences to skip over nearby repetitive elements, even short sequences can be assembled into large contiguous regions using massively parallel sequencing of random DNA fragments (i.e. ‘paired-end Shotgun sequencing’)[29, 30]. Without such strategies, our modern DNA sequencers with relatively short sequencing-read length cannot assemble the genome efficiently. In order to construct the genome from end-to-end, the latest sequencing technologies can generate extremely long sequencing-read lengths (i.e. >10,000 bases) and are capable of detecting large genomic structural variants *de novo* (Fig. 2C)[31–34]. By combining long sequencing-read technologies (fast *de novo* genome assembly) with short sequencing-read technologies (low sequencing error-rate), it might be possible to assemble various genomes accurately, efficiently, and rapidly across model organisms and patients.

In assembling a gene expression atlas using single-cell RNA sequencing, it might be helpful to consider dissociated single-cells as ‘short DNA fragments in solution’, and the gene expression signature, as a ‘short DNA sequencing read’. Similarly to the genome assembly in 1-D, there might be multiple plausible strategies for mapping single-cells in 2-D or 3-D in organisms of different complexity and sizes; however, the key question is whether such strategies can be generated in a fast, efficient, and scalable manner across many specimens or individuals. The answer to this question could determine how such technologies are used for comparing genetic and functional differences across different individuals, organisms, and diseases.

Current Approaches

Recently, advances in single-cell RNA sequencing have enabled the identification of cell types and cell states from complex tissues in an unbiased manner[14]. Here, single-cells are enzymatically dissociated and delivered into reaction chambers using microfluidics[4–6, 24]. Once in a microfluidics device, the cells’ mRNAs are reverse-transcribed and barcoded[35, 36], and the sample is pooled prior to massively parallel sequencing. Although single-cell RNA sequencing-reads are sparse and noisy for each cell, the population analysis of many single cells ($\sim 10^2$ – 10^4) can identify infrequent cell types or states missed in bulk-tissue RNA sequencing. Consequently, co-expression analysis can infer the transcriptional phenotype of a representative cell type or state; however, it cannot measure cell-cell interactions because each cell has lost the positional information during single-cell dissociation.

Another major advance in recent years has been in the field of single-molecule RNA imaging in situ. One critical feature to consider in this methodology is the use of fluorescently labeled oligonucleotides that co-localize on the same RNA transcript; in this manner, it is possible to discriminate specific from non-specific oligonucleotide binding to RNA[37–39]. Modern advances now allow for sequential hybridization of the same transcript, so that each transcript molecule can be represented by different color sequences over time[9–13]. Here, it is necessary to find and align each ‘spot’ or ‘signal’ to decode the identity of individual transcripts; however, signal crowding can become a major issue when using low-resolution optical microscopy for whole tissue imaging[40]. While single-molecule RNA-FISH methods have unparalleled sensitivity and specificity (over >90% in many cases)[41], this property works against massively multiplexed profiling in single cells ($\sim 10^4$ – 10^5 mRNAs/cell) when using optical imaging, as it may result in signal crowding[40].

Potential Solutions and Caveats

It is well known that two main weaknesses of single-cell RNA sequencing are the lack of spatial information and the sparse and noisy data structure in single-cells. Although the latter can be addressed using population-based co-expression analysis, the lack of spatial information is harder to resolve. One could speculate about impregnating single-cells with oligonucleotide barcodes in situ. If such barcodes could be synthesized on the opposite ends of micron-sized particles to preserve distance and angular information between single-cells, it might be possible to perform the 2D or 3D spatial equivalent of paired-end shotgun sequencing for tissue assembly using multiple specimens (in paired-end shotgun sequencing, RNA is randomly broken up into numerous small fragments, and both ends are sequencing to generate high-quality, alignable sequence data). Alternatively, each tissue could be randomly fragmented into uniformly-sized, jigsaw-shaped cell clusters, and single-cells at the opposing ends could be barcoded and sequenced. While theoretically plausible, purely computational approaches are fraught with technical challenges and experimental variabilities. In addition, there would be limited scalability of single-cell RNA sequencing over the whole tissue (i.e. a tissue specimen of 1-cm in diameter contains $\sim 10^9$ cells).

Ideally, single-cells and their molecular signatures should be profiled in the intact 3D environment. If one could scale up ISH techniques so that tens of thousands of genes might be interrogated in parallel, challenges associated with single-cell RNA sequencing might be solved[10]. The bottleneck in highly multiplexed single-molecule RNA ISH is the optical resolution required to resolve tens of thousands of spots in each cell. Recently, the Boyden laboratory at MIT showed that one could embed tissue specimens in an expandable acrylamide gel and isometrically increase the cell dimensions while retaining RNA localization patterns [42, 43]. In fact, the expanded specimen could be embedded iteratively for imaging single-molecules at ultra-high resolution using a traditional microscope[44]. With further advances in imaging techniques, comprehensive profiling using multiplexed single-molecule RNA-FISH might be possible. Combined with single-cell RNA sequencing, a large number of molecular signatures could be defined and mapped in tissues with single-molecule resolution; however, approaches that depend on single-molecule imaging may not scale favorably for comparisons of comprehensive sets of genes across large numbers of

specimens, individuals, or species. Might there be a way to generate a sparse, or lower-complexity representation in order to obtain a more efficient spatial reconstruction of gene expression *de novo* in cells and tissues?

Assembling Tissue Atlases Using Stochastic Gene Expression Data in Space

Single-cell RNA reconstructs the transcriptome of a representative cell type using sparse data from thousands of single-cells. However, to reconstruct the transcriptome associated with specific tissue phenotypes, it might be necessary to profile gene expression in situ, even if it may yield low sensitivity, and then reconstruct the full transcriptome across multiple images and specimens. In this case, signal de-crowding is achieved by distributing the signal detection across more specimens, rather than by relying on ultra high-resolution microscopy. In fact, the stochastic signal detection over time (Fig. 3A) or space (Fig. 3B) followed by computational reconstruction has changed the face of high-resolution microscopy in cell biology[45], and cryo-electron microscopy in protein structural biology[46], respectively.

One such solution might be found in technologies such as Fluorescent In Situ Sequencing (FISSEQ), a random transcriptome sampling technique with low sensitivity (~0.01%) that can detect thousands of genes in parallel, testing the spatial enrichment of pathways and gene expression clusters (Fig. 3C)[47–49]. Here, fixed cells or tissues are infused with reagents capable of randomly converting RNA molecules into cDNA fragments that are amplified in situ. Then, standard sequencing reagents reliant on fluorescent probes and optical microscopy (i.e. SOLiD and Illumina platforms) are added for massively parallel RNA sequencing on a 3-D imaging microscope (Fig. 4A). Our laboratory is now using high-speed 3D imaging to scan 4×4-mm regions in just over two hours per cycle. (The entire run requires 32 cycles, or ~2-TB images per run) [47–49]. One of our current goals is to establish the number of arrayable *Drosophila* embryos required to detect key developmental regulators in early embryo patterning (Fig. 3C and 4B); for instance, ~100 embryos can be sequenced in parallel using a 2-hour FISSEQ 3D image acquisition set-up.

It is too early to tell whether the research community will adopt FISSEQ or similar technologies; regardless, the key lesson here is that the sequencing-read can be generated and fixed to a position in space. By randomly sampling the transcriptome in an unbiased manner, FISSEQ enables statistical testing of position- and window size-specific gene expression. Similarly to genome sequencing, more than one cell, embryo, or tissue phenotype can be arrayed and sequenced in parallel to increase the coverage and spatial resolution. Although conceptually enticing, it is unknown how this approach compares to other previously discussed approaches in terms of actually yielding novel molecular targets or biomarkers involved in multi-cellular development and pathophysiology. Consequently, a direct side-by-side comparison will be essential in the near future to address this issue.

Visualizing Genomics Information in Clinical Specimens without Molecular Imaging

Single-cell RNA sequencing, multiplexed single-molecule RNA-FISH, and fluorescence in situ RNA sequencing (FISSEQ) generate a large amount of data or images, and they require sophisticated tools for data analysis and interpretation; however, such features make it less likely that global genomics- or transcriptomics-scale information is used for routine bench

experiments or patient tissue diagnostics. In comparison, general tissue stains such as hematoxylin & eosin (H&E) are widely used in research and medicine precisely because they are universally applicable, easy to implement, and highly informative in distinguishing cellular features with high spatial resolution. Might there be easier ways to detect gene expression changes in tissues at a spatial scale in a medically or biologically meaningful, unbiased manner (Box 1)?

Box 1

Clinician's Corner

- General tissue stains in histopathology are fast, easy-to-implement, and highly informative for tissue diagnosis, but they do not utilize a large amount of genetic data embedded within each tissue section.
- Methods and assays are needed to extract the genetic information from each tissue specimen along with morphological, histological, or architectural tissue/cell information.
- In the future, there might be simpler and easier-to-use technologies to paint or stain the genomic information present in tissue specimens in an unbiased manner using modified sequencing chemistries.
- Tumor growth, wound healing, microbial infection, immunotherapy, and most other medical conditions are highly dependent on their local tissue context. Finding diagnostic or prognostic markers and therapeutic candidates warrants new technologies capable of quantifying genetic variants in an actual tissue-specific context.

Here, we turn to yet another example in the reference genome assembly projects. Even with advanced DNA sequencers generating long sequencing-reads, using *de novo* genome assembly is an impractical option in medicine. This is why technologies such as DNA nanopore sequencing might change the way in which we utilize genomics-scale information[31–33]. Specifically, in DNA nanopore sequencing, one long DNA molecule is threaded through a protein pore channel embedded within a membrane, and changes in the electrical current are measured as the DNA molecule slips through. Because it does not rely on optical imaging, DNA nanopore sequencers fit on a device the size of a USB flash-drive with minimal input DNA preparation (Fig. 5A). In the future, an array of DNA nanopore sequencers might make *de novo* assembly of the human genome a practical and affordable option in detecting large genomic variants, such as DNA translocations or inversions in cancers.

Similarly, our laboratory is focusing on in situ sequencing methods that do not rely on single-molecule or high-resolution imaging. Instead, our approach is to emphasize the utility of simple tissue stain-like approaches in assessing tissue architecture (Fig. 5B). Here, the critical elements are: i) synthesizing a large number of short oligonucleotides that bind RNA targets with no off-target binding in situ, eliminating the need to see probe co-localization using single-molecule imaging; ii) sequential stripping and hybridization with a universal set

of oligonucleotides that define orthogonal cell types, pathways, or processes defined by single-cell RNA sequencing; and iii) automated image analysis for the rapid visualization and interpretation of gene expression pathways and biological processes. Our laboratory might now be in a position to address the first issue with a new conceptualized sequencing technology enabling labeled oligonucleotides to bind directly to ~14–16-bp RNA target sequences with single-nucleotide resolution: the premise of this method is that incorrectly paired oligonucleotides on the RNA target might not be joined together and thus would be degraded by exonucleases, resulting in little or no background. Evidently, more extensive work is needed to precisely design universal oligonucleotide pools based on single-cell RNA sequencing, and be able to infer the biological pathways from sequential staining characteristics in tissues. Nevertheless, we believe that it is vital that concepts and technologies in high-end genomics be able to yield tools and assays to facilitate genomics-based medicine in a routine clinical setting.

Simple Sequencing Tools with High Spatial Resolution Can Transform Biology

The advent of CRISPR/Cas9 technology has enabled unbiased transcriptional activation, silencing, and perturbation genome-wide[50–53], as well as cell lineage tracing using induced somatic mutations[54, 55]; however, pooled genetic screens using CRISPR/Cas9 often use cell lines in vitro[51, 53, 56]. Because the functional phenotype often depends on the tissue context, it might be important to perform pooled CRISPR/Cas9 screens in vivo[57–59], and subsequently assess cellular phenotype side-by-side in situ (i.e. behavior, morphology, function). Several laboratories are already developing in situ DNA or RNA sequencing methods for Cas9-induced mutations in transcriptional reporters[55, 60]. To use the simplified in situ sequencing chemistry described above without extra signal amplification, our laboratory has developed a method of directing hundreds of RNA barcode molecules to a specific subcellular compartment (Fig. 6A). Here, the concentrated RNA barcodes (6-bases) are sequenced using serial sequencing reactions followed by imaging, generating millions of fluorescence signatures that can be used for cell segmentation and tracing [54, 55]. Furthermore, such reporters can be mutated over time using Cas9 in vivo, and then sequenced in situ to reconstruct cellular phylogenetic relationships in space[54, 55] (Fig. 6B). This might allow one to track in vivo, stem or cancer cell divisions, lineages, and co-migration properties across various types of microenvironments, and with single-cell resolution.

Concluding Remarks

In this Opinion article, we build on concepts and technologies established by many pioneers in the field of *de novo* genome assembly. We suggest that there might be additional parallels that could be helpful in developing strategies for efficient and scalable reconstruction of gene expression atlases in tissues (see Outstanding Questions). Again, we emphasize the need to develop novel methods, utilizing a reductionist representation of the transcriptome signature for fast, unbiased assessment of biological pathways and processes in situ. The physical length of sequencing reads have provided a rallying cry in sequencing technology development, leading to efficient *de novo* genome assembly techniques. Analogous concepts are likely to play a vital role in developing technologies that will assemble tissue atlases of

gene expression in space. We believe that these advances are fundamental to achieving accurate comparisons of many different phenotypic variants within populations, and to revealing potential new insights in biology and medicine.

Outstanding Questions Box

- How many cells and tissue types will we need to sequence-catalog all major cell types in the human body in an unbiased manner?
- How do we standardize such efforts for systemic and comprehensive sequence-cataloguing?
- Are advances in high-resolution optical microscopy, data storage, and computational analysis for multiplexed single-molecule RNA imaging sufficiently scalable for comparative genomics-phenomics?
- Beyond generating reference atlases in the future, what are the conceptual and technological issues in functional genomics that need to be addressed by these technologies?

Acknowledgments

We would like to acknowledge the support of NIGMS R35, The V Foundation, STARR Cancer Consortium, Human Scientific Frontier Program, and Pershing Square Foundation, NCI, and Northwell/CSHL for enabling the development of various concepts, tools, and technologies discussed in this manuscript.

References

1. Venteicher AS, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (New York, NY)*. 2017;355.
2. Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Current opinion in biotechnology*. 2017; 46:126–133. [PubMed: 28346891]
3. Crosetto N, et al. Spatially resolved transcriptomics and beyond. *Nature reviews. Genetics*. 2015; 16:57–66.
4. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015; 161:1202–1214. [PubMed: 26000488]
5. Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015; 161:1187–1201. [PubMed: 26000487]
6. Zheng GX, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*. 2017; 8:14049.
7. Cao J, et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv*. 2017:104844.
8. Shekhar K, et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. 2016; 166:1308. [PubMed: 27565351]
9. Lubeck E, Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature methods*. 2012; 9:743–748. [PubMed: 22660740]
10. Lubeck E, et al. Single-cell in situ RNA profiling by sequential hybridization. *Nature methods*. 2014; 11:360–361. [PubMed: 24681720]
11. Shah S, et al. Single-molecule RNA detection at depth via hybridization chain reaction and tissue hydrogel embedding and clearing. *Development*. 2016
12. Chen KH, et al. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. 2015; 348:aaa6090. [PubMed: 25858977]

13. Moffitt JR, et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 113:11046–11051. [PubMed: 27625426]
14. Ziegenhain C, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell*. 2017; 65:631–6430000. [PubMed: 28212749]
15. Kim JK, et al. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature communications*. 2015; 6:8687.
16. Svensson V, et al. Power analysis of single-cell RNA-sequencing experiments. *Nature methods*. 2017; 14:381–387. [PubMed: 28263961]
17. Molyneaux BJ, et al. DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. *Neuron*. 2015; 85:275–288. [PubMed: 25556833]
18. Lovatt D, et al. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nature methods*. 2014; 11:190–196. [PubMed: 24412976]
19. Miller MR, et al. TU-tagging: cell type specific RNA isolation from intact complex tissues. *Nature methods*. 2009
20. Stahl PL, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016; 353:78–82. [PubMed: 27365449]
21. Lein ES, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007; 445:168–176. [PubMed: 17151600]
22. Jones AR, et al. The Allen Brain Atlas: 5 years and beyond. *Nature Reviews Neuroscience*. 2009; 10:821–828. [PubMed: 19826436]
23. Achim K, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature biotechnology*. 2015; 33:503–509.
24. Treutlein B, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*. 2014; 509:371–375. [PubMed: 24739965]
25. Satija R, et al. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*. 2015
26. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*. 1988; 2:231–239. [PubMed: 3294162]
27. Fleischmann RD, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995
28. Schatz MC, et al. Assembly of large genomes using second-generation sequencing. *Genome research*. 2010; 20:1165–1173. [PubMed: 20508146]
29. Roach JC, et al. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*. 1995
30. Port E, et al. Genomic mapping by end-characterized random clones: A mathematical analysis. *Genomics*. 1995
31. Goodwin S, et al. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*. 2015; 25:1750–1756. [PubMed: 26447147]
32. Lee H, et al. Third-generation sequencing and the future of genomics. *bioRxiv*. 2016:48603.
33. Ashton PM, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*. 2014
34. Peters BA, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*. 2012; 487:190–195. [PubMed: 22785314]
35. Hashimshony T, et al. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports*. 2012; 2:666–673. [PubMed: 22939981]
36. Ramsköld D, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*. 2012; 30:777–782.
37. Femino AM, et al. Visualization of single RNA transcripts in situ. *Science*. 1998; 280:585–590. [PubMed: 9554849]
38. Raj A, et al. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods*. 2008; 5:877–879. [PubMed: 18806792]

39. Player AN, et al. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. *J Histochem Cytochem.* 2001; 49:603–612. [PubMed: 11304798]
40. Coskun AF, Cai L. Dense transcript profiling in single cells by image correlation decoding. *Nature methods.* 2016; 13:657–660. [PubMed: 27271198]
41. Itzkovitz S, van Oudenaarden A. Validating transcripts with probes and imaging technology. *Nature methods.* 2011; 8:S12–19. [PubMed: 21451512]
42. Chen F, et al. Nanoscale imaging of RNA with expansion microscopy. *Nature methods.* 2016; 13:679–684. [PubMed: 27376770]
43. Chen F, et al. Expansion microscopy. *Science.* 2015; 347:543–548. [PubMed: 25592419]
44. Chang J-BB, et al. Iterative expansion microscopy. *Nature methods.* 2017
45. Huang B, et al. Breaking the diffraction barrier: super-resolution imaging of cells. *Cell.* 2010; 143:1047–1058. [PubMed: 21168201]
46. Frank, J. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press; 2006.
47. Lee JH. Quantitative approaches for investigating the spatial context of gene expression. Wiley interdisciplinary reviews. *Systems biology and medicine.* 2016
48. Lee JH, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature protocols.* 2015; 10:442–458. [PubMed: 25675209]
49. Lee JH, et al. Highly multiplexed subcellular RNA sequencing in situ. *Science.* 2014
50. Gilbert LA, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell.* 2013; 154:442–451. [PubMed: 23849981]
51. Shalem O, et al. High-throughput functional genomics using CRISPR-Cas9. *Nature reviews. Genetics.* 2015; 16:299–311.
52. Konermann S, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature.* 2015; 517:583–588. [PubMed: 25494202]
53. Shalem O, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science.* 2014; 343:84–87. [PubMed: 24336571]
54. McKenna A, et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science.* 2016; 353:aaf7907. [PubMed: 27229144]
55. Frieda KL, et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature.* 2017; 541:107–111. [PubMed: 27869821]
56. Dixit A, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell.* 2016; 167:1853. [PubMed: 27984732]
57. Platt RJ, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell.* 2014; 159:440–455. [PubMed: 25263330]
58. Chen S, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell.* 2015; 160:1246–1260. [PubMed: 25748654]
59. Xue W, et al. CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature.* 2014; 514:380–384. [PubMed: 25119044]
60. Kalhor R, et al. Rapidly evolving homing CRISPR barcodes. *Nature methods.* 2017; 14:195–200. [PubMed: 27918539]
61. Betzig E, et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science.* 2006; 313:1642–1645. [PubMed: 16902090]
62. Rust MJ, et al. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods.* 2006; 3:793–795. [PubMed: 16896339]
63. Bates M, et al. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science.* 2007; 317:1749–1753. [PubMed: 17702910]
64. Huang B, et al. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science.* 2008; 319:810–813. [PubMed: 18174397]
65. Shaikh TR, et al. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature protocols.* 2008; 3:1941–1974. [PubMed: 19180078]

Trends Box

- Single-cell RNA sequencing will soon be able to sequence up to 10^5 cells, and it can already catalog common and rare cell types in an unbiased manner; however, it lacks the spatial information for investigating cell-cell interactions.
- Single-molecule RNA in situ hybridization can visualize hundreds of genes in parallel, and new microscopy techniques might allow comprehensive spatial RNA profiling.
- Fast and efficient *de novo* assembly of the transcriptome in 3D might require a new conceptual framework as well as novel technologies, as in the case of reference genome DNA assembly projects.
- The stochastic optical reconstruction of in situ sequencing or spatial transcriptomics data might assemble the transcriptome of spatial phenotypes (i.e. morphology, behavior, patterns) in a scalable manner.
- Simplifying in situ sequencing or spatial transcriptomics technologies could have a major impact on basic, translational, and clinical problems that tightly depend on an increased understanding of tissue context.

Figure 1. Challenges in Spatial Gene Expression Analysis

(A) In most cases, tissue regions are dissected (e.g. laser-capture microdissection) using anatomic or genetic markers for RNA sequencing. Alternatively, the RNA can be tagged using light or genetically encoded enzymes in vivo (photorelease oligo-dT in vivo), followed by tag-assisted RNA purification. (B) In traditional genetics, a genetic locus is defined experimentally and subsequently cloned for sequencing. In both cases, finding positional markers can be time-consuming and laborious. (C) Gene expression atlases using RNA imaging do not use region-specific markers, but their creation depends on large industrialized implementations. (D) Similarly, reference genome projects had once relied on industrialized implementations of genomic cloning, sequencing, and assembly (e.g. shotgun sequencing); however, such approaches do not scale to thousands of individuals.

Figure 2. Challenges in Mapping the Position of Gene Expression and DNA Sequences
 (A) Single cells are dissociated from tissues, often under the microscope. Thousands of such cells are individually sequenced to identify gene expression variations between distinct cell populations. However, even with information on the cell type or state, mapping such signatures to a reference map (i.e. brain atlas) is challenging. (B) Similarly, modern sequencers generate short sequencing-reads that often cannot be mapped uniquely to the reference genome due to the sheer number of related sequences. (C) A long DNA fragment can be sequenced with lower resolution bi-directionally using newer sequencing technologies. The longer sequencing-read length generates more unique overlaps and yields faster and longer *de novo* assembly[32]. DNAPol: DNA polymerase

Figure 3. Massively Parallel RNA Sequencing in 3-D Space Can Use Modern Image Reconstruction Methods

Emerging *in situ* sequencing technologies can add positional coordinates to RNA sequencing-reads; however, the detection frequency is low and stochastic. (A) In super-resolution microscopy, thousands of stochastic single-molecule images over time are used to calculate the true position, breaking the optical microscopy resolution barrier[61–64]. (B) In cryo-electron microscopy (cryoEM), thousands of single proteins are imaged side-by-side for reconstruction tomography, radically accelerating protein structural biology[46, 65]. (C) Similarly, 3D reconstruction of gene expression might be possible from spatial RNA sequencing technologies (*in situ* sequencing) with low sensitivity, as long as each phenotype class can be grouped based on morphology, orientation, size, or gene expression patterns.

Figure 4. Applications for Optical Reconstruction Tomography in Massively Parallel RNA Sequencing In Situ

(A) Fluorescent In Situ Sequencing (FISSEQ) stochastically reverse transcribes, circularizes, and amplifies the transcriptome (amplicons) for sequencing-by-ligation in situ, which is then imaged in 3D. (B) Representative micrographs of RNA amplicons are shown in one of four colors that change over time. The gene identity is defined by aligning the color sequence to the reference transcriptome. By combining RNA sequences from spatially defined features (i.e. movement, morphology, polarity, patterning), the transcriptome for representative phenotypes can be constructed. If enough tissues can be imaged in parallel, which is a function of the tissue size and the imaging speed, it might be possible to rapidly reconstruct a comprehensive tissue atlas. Examples of images taken of astrocytes, embryonic stem cells and a *Drosophila sp.* blastoderm are shown.

Figure 5. Trade-offs in Molecular Resolution, Spatial Information, and Usability

(A) Less accurate and lower-resolution DNA sequencing technologies (i.e. DNA nanopore sequencing) can generate long stretches of DNA sequences for low-cost *de novo* genome assembly. Because it does not use optical microscopy like other DNA sequencers (i.e. Illumina), it can be miniaturized and made portable. (B) Similarly, it might be necessary to develop multiplexed in situ RNA detection methods (defining cell type/state-specific (sp)) that do not require single-molecule imaging. The main challenge is to develop in situ RNA detection chemistry that will enable pooling of many DNA or RNA probes with little or no cross-reactivity for sequential hybridization of major transcriptional signatures in situ. CT1, CT2, etc: cell type 1,2, etc.).

Figure 6. Prioritizing Simpler In Situ Sequencing Chemistries

One of the most tantalizing opportunities for *in situ* sequencing is the potential for massively parallel functional screens and cell tracing *in vivo*. (A) Synthetic transcripts bearing an RNA barcode of length N (4^N possibilities) might be localized to specific subcellular organelles using a protein fusion tag, including the plasma membrane, to digitally segment all 4^N cells using *in situ* RNA sequencing of the barcode sequence (RNA BC). (B) Such barcodes could be designed to evolve for phylogenetic reconstruction to track cell fate decisions over space and time [54, 55, 60], or indicate specific genetic perturbations (i.e. CRISPR/Cas9-sgRNA) *in vivo* to understand how the neighboring cell or the microenvironment modifies the gene function [58]. The molecular recorder is a biological molecular reporter construct that is altered by stochastic or periodic events over time to ‘record’ the occurrence of events. Typically, DNA is used since its alterations can be sequenced, and the progressive changes (i.e. mutation, deletion, expansions) can be massively assayed in parallel for temporal reconstruction. Our current focus is finding sequencing chemistries that are easy-to-use, sensitive, and independent of on single-molecule imaging. If these criteria could be met, they might lead to a new paradigm in experimental research in developmental and cancer

biology. sgRNA: single guide RNA; MS2-LS:localization signal;; freq: frequency; sp.
specific nt: nucleotide(s).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript