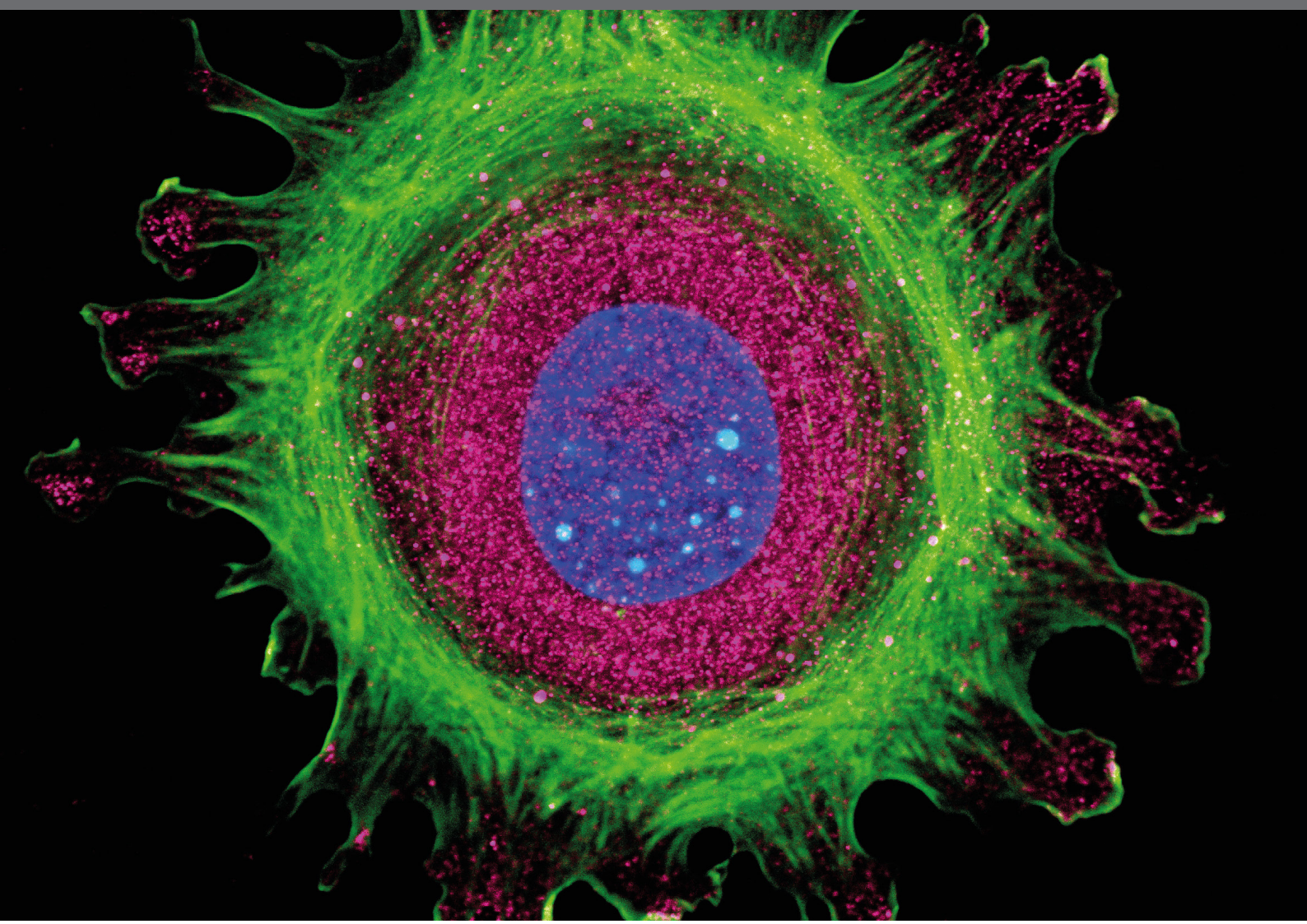
INTRODUCTION TO SINGLE CELL OMICS

EDITED BY: Xinghua Pan, Shixiu Wu and Sherman M. Weissman
PUBLISHED IN: Frontiers in Cell and Developmental Biology, Frontiers in Genetics and Frontiers in Bioengineering and Biotechnology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-920-9
DOI 10.3389/978-2-88945-920-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

INTRODUCTION TO SINGLE CELL OMICS

Topic Editors:
Xinghua Pan, Southern Medical University, Guangdong Provincial Key Lab of Single Cell Technology and Application, China; Yale University School of Medicine, United States
Shixiu Wu, Hangzhou Cancer Hospital, China
Sherman M. Weissman, Yale University School of Medicine, United States

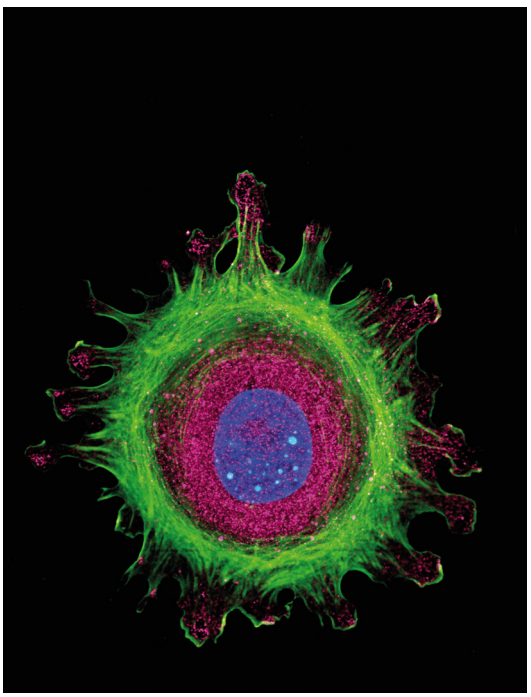


Image: DrimaFilm/Shutterstock.com

Single-cell omics is a progressing frontier that stems from the sequencing of the human genome and the development of omics technologies, particularly genomics, transcriptomics, epigenomics and proteomics, but the sensitivity is now improved to single-cell level. The new generation of methodologies, especially the next generation sequencing (NGS) technology, plays a leading role in genomics related fields; however, the conventional techniques of omics require number of cells to be large, usually on the order of millions of cells, which is hardly accessible in some cases. More importantly, harnessing the power of omics technologies and applying those at the single-cell level are crucial since every cell is specific and unique, and almost every cell population in every systems, derived in either vivo or in vitro, is heterogeneous. Deciphering the heterogeneity of the cell population hence

becomes critical for recognizing the mechanism and significance of the system. However, without an extensive examination of individual cells, a massive analysis of cell population would only give an average output of the cells, but neglect the differences among cells.

Single-cell omics seeks to study a number of individual cells in parallel for their different dimensions of molecular profile on genome-wide scale, providing unprecedented resolution for the interpretation of both the structure and function of an organ, tissue or other system, as well as the interaction (and communication) and dynamics of single cells or subpopulations of cells and their lineages. Importantly single-cell omics enables the identification of a minor subpopulation of cells that may play a critical role in biological process over a dominant subpolulation such as a cancer and a developing organ. It provides an ultra-sensitive tool for us to clarify specific molecular mechanisms and pathways and reveal the nature of cell heterogeneity. Besides, it also empowers the clinical investigation of patients when facing a very low quantity of cell available for analysis, such as noninvasive cancer screening with circulating tumor cells (CTC), noninvasive prenatal diagnostics (NIPD) and preimplantation genetic test (PGT) for in vitro fertilization. Single-cell omics greatly promotes the understanding of life at a more fundamental level, bring vast applications in medicine. Accordingly, single-cell omics is also called as single-cell analysis or single-cell biology.

Within only a couple of years, single-cell omics, especially transcriptomic sequencing (scRNA-seq), whole genome and exome sequencing (scWGS, scWES), has become robust and broadly accessible. Besides the existing technologies, recently, multiplexing barcode design and combinatorial indexing technology, in combination with microfluidic platform exampled by Drop-seq, or even being independent of microfluidic platform but using a regular PCR-plate, enable us a greater capacity of single cell analysis, switching from one single cell to thousands of single cells in a single test. The unique molecular identifiers (UMIs) allow the amplification bias among the original molecules to be corrected faithfully, resulting in a reliable quantitative measurement of omics in single cells. Of late, a variety of single-cell epigenomics analyses are becoming sophisticated, particularly single cell chromatin accessibility (scATAC-seq) and CpG methylation profiling (scBS-seq, scRRBS-seq). High resolution single molecular Fluorescence *in situ* hybridization (smFISH) and its revolutionary versions (ex. seqFISH, MERFISH, and so on), in addition to the spatial transcriptome sequencing, make the native relationship of the individual cells of a tissue to be in 3D or 4D format visually and quantitatively clarified. On the other hand, CRISPR/cas9 editing-based *in vivo* lineage tracing methods enable dynamic profile of a whole developmental process to be accurately displayed. Multi-omics analysis facilitates the study of multi-dimensional regulation and relationship of different elements of the central dogma in a single cell, as well as permitting a clear dissection of the complicated omics heterogeneity of a system. Last but not the least, the technology and biological noise, sequence dropout, and batch effect bring a huge challenge to the bioinformatics of single cell omics. While significant progress in the data analysis has been made since then, revolutionary theory and algorithm logics for single cell omics are expected. Indeed, single-cell analysis exert considerable impacts on the fields of biological studies, particularly cancers, neuron and neural system, stem cells, embryo development and immune system; other than that, it also tremendously

motivates pharmaceutic RD, clinical diagnosis and monitoring, as well as precision medicine.

This book hereby summarizes the recent developments and general considerations of single-cell analysis, with a detailed presentation on selected technologies and applications. Starting with the experimental design on single-cell omics, the book then emphasizes the consideration on heterogeneity of cancer and other systems. It also gives an introduction of the basic methods and key facts for bioinformatics analysis. Secondary, this book provides a summary of two types of popular technologies, the fundamental tools on single-cell isolation, and the developments of single cell multi-omics, followed by descriptions of FISH technologies, though other popular technologies are not covered here due to the fact that they are intensively described here and there recently. Finally, the book illustrates an elastomer-based integrated fluidic circuit that allows a connection between single cell functional studies combining stimulation, response, imaging and measurement, and corresponding single cell sequencing. This is a model system for single cell functional genomics. In addition, it reports a pipeline for single-cell proteomics with an analysis of the early development of Xenopus embryo, a single-cell qRT-PCR application that defined the subpopulations related to cell cycling, and a new method for synergistic assembly of single cell genome with sequencing of amplification product by phi29 DNA polymerase. Due to the tremendous progresses of single-cell omics in recent years, the topics covered here are incomplete, but each individual topic is excellently addressed, significantly interesting and beneficial to scientists working in or affiliated with this field.

Citation: Pan, X., Wu, S., Weissman, S. M., eds. (2019). Introduction to Single Cell Omics. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-920-9

Table of Contents

frontiers in Cell and Developmental Biology

MINI REVIEW
published: 04 September 2018
doi: 10.3389/fcell.2018.00108

Experimental Considerations for Single-Cell RNA Sequencing Approaches

*Quy H. Nguyen[1], Nicholas Pervolarakis[2], Kevin Nee[1] and Kai Kessenbrock[1]**

[1] Department of Biological Chemistry, University of California, Irvine, Irvine, CA, United States, [2] Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, United States

Single-cell transcriptomic technologies have emerged as powerful tools to explore cellular heterogeneity at the resolution of individual cells. Previous scientific knowledge in cell biology is largely limited to data generated by bulk profiling methods, which only provide averaged read-outs that generally mask cellular heterogeneity. This averaged approach is particularly problematic when the biological effect of interest is limited to only a subpopulation of cells such as stem/progenitor cells within a given tissue, or immune cell subsets infiltrating a tumor. Great advances in single-cell RNA sequencing (scRNAseq) enabled scientists to overcome this limitation and allow for in depth interrogation of previously unexplored rare cell types. Due to the high sensitivity of scRNAseq, adequate attention must be put into experimental setup and execution. Careful handling and processing of cells for scRNAseq is critical to preserve the native expression profile that will ensure meaningful analysis and conclusions. Here, we delineate the individual steps of a typical single-cell analysis workflow from tissue procurement, cell preparation, to platform selection and data analysis, and we discuss critical challenges in each of these steps, which will serve as a helpful guide to navigate the complex field of single-cell sequencing.

Keywords: single-cell genomics, single-cell analysis, cell isolation, computational biology, cellular heterogeneity

OPEN ACCESS

Edited by:
Xinghua Victor Pan,
Yale University, United States

Reviewed by:
Lasse Dahl Ejby Jensen,
Linköping University, Sweden
Alexander D. Borowsky,
University of California, Davis, United States

****Correspondence:***
Kai Kessenbrock
kai.kessenbrock@uci.edu

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *29 April 2018*
Accepted: *20 August 2018*
Published: *04 September 2018*

Citation:
Nguyen QH, Pervolarakis N, Nee K and Kessenbrock K (2018) Experimental Considerations for Single-Cell RNA Sequencing Approaches. Front. Cell Dev. Biol. 6:108. doi: 10.3389/fcell.2018.00108

INTRODUCTION

Elucidating cellular heterogeneity represents a major scientific challenge in many areas of biology and biomedical research including developmental and stem cell biology, immunology, neurobiology, and cancer research (Wagner et al., 2016). Recent convergence of next generation sequencing (NGS) and bioengineering approaches to manipulate individual cells has led to unbiased single-cell DNA (Navin et al., 2011), RNA (Pollen et al., 2014; Treutlein et al., 2014; Tanay and Regev, 2017), and ATAC (Buenrostro et al., 2015) sequencing. These technological advances are redefining our understanding of how biological systems function and have formed the basis for large-scale, international collaborations such as the Human Cell Atlas project (Rozenblatt-Rosen et al., 2017). In this spirit, a recent endeavor using microwell-based single-cell RNAseq (scRNAseq) created the first cell atlas to map out most tissues of the mouse (Han et al., 2018). Moreover, scRNAseq has provided critical new insights into key developmental processes such as the earliest steps of cardiovascular lineage segregation in mice (Lescroart et al., 2018), and our recent work utilized scRNAseq to reveal the spectrum of cellular heterogeneity within the human

breast epithelium identifying three major cell types each harboring multiple distinct cell states (Nguyen et al., 2018).

Due to the high sensitivity of these methods, in particular scRNAseq, it can be difficult to choose an adequate approach to minimize batch effects and unwanted technical variation that may overshadow true biological insights. Here, we provide helpful insights and delineate a step-wise approach for designing single-cell analysis workflows (**Figure 1**).

CELL DISSOCIATION AND SINGLE-CELL PREPARATION

The process of single-cell preparation is arguably the greatest source of unwanted technical variation and batch effects in any single-cell study (Tung et al., 2017). Different tissues can vary significantly in extracellular matrix (ECM) composition, cellularity, and stiffness, and therefore dissociation protocols must be optimized for the specific tissue type of interest. Conventional protocols for single-cell preparation typically involve the following steps: (1) tissue dissection, (2) mechanical mincing, (3) enzymatic/proteolytic ECM breakdown (e.g., dispase, collagenase, trypsin) often accompanied by mechanical agitation, and (4) optional enrichment for cell types of interest by flow cytometry, bead-based immune-selection, differential centrifugation, or sedimentation. Each step can affect the cells' expression signatures, and should therefore be carefully optimized to introduce the least artifact. An optimal tissue dissociation protocol will yield as many viable cells as possible in the shortest possible duration without preferentially depleting or significantly altering the frequencies of certain cell types.

Recent advances in bioengineering of innovative microfluidic cell dissociation devices (Qiu et al., 2014) have the potential to radically change the way tissue samples are dissociated into single cells, while avoiding inter-assay variation due to human handling of the tissue. Several microfluidic devices have been optimized for streamlined tissue digestion, cell dissociation, filtering, and polishing. In brief, these devices were designed to work with tissue sequentially through progressively smaller size scales, starting from tissue specimen, through cellular aggregates and clusters, and finally eluting a solution containing close to 100% single cells, which will be ideal for scRNAseq applications. In addition, new semi-automated commercially available systems can help streamline tissue dissociation (e.g., Miltenyi gentleMACS). These devices offer tissue-type specific kits that may allow more reproducible, time-saving and efficient tissue dissociation and single-cell preparation (Meeson et al., 2013; Baldan et al., 2015). Ultimately, determining a "best practices" dissociation strategy through heuristic optimization will be critical for downstream single-cell library quality.

Cell Type Enrichment

There are various methods for isolating specific cell populations or removal of unwanted populations that should be optimized for any specific tissues type. Manual isolation utilizing magnetic beads or gradient purification are potential methods for removal of unwanted cells such as dead cells. Flow cytometry is a widely used, high-throughput method to enrich for rare cells such as hematopoietic stem cells (Radbruch and Recktenwald, 1995; Will and Steidl, 2010). However, these methods are not without drawbacks, since they can introduce artificial stress on cells and change their expression profile (Van Den Brink et al., 2017). Methods that involve antibody binding for purification can also affect the cell expression profile if binding of the antibodies to cell surface molecules induce intracellular signaling (Kornbluth and Hoover, 1989; Christaki et al., 2011). Flow cytometry-isolated cells are exposed to high pressure during sorting and these osmotic and pressure changes introduced to cells during cell sorting and handling can induce change to the cell expression profile of multiple cell types (Xiong et al., 2002; Romero-santacreu et al., 2009; Van Den Brink et al., 2017).

Quality Control

Due to the high cost of single-cell sequencing experiments, careful quality control measurements should be executed. The performance of alternative protocols can be assessed using a number of readouts. A useful first metric can be acquired using imaging of viability such as using the Countess platform (Thermo Fisher Scientific). Flow cytometry is particularly valuable to measure several critical metrics simultaneously, such as cell viability, and contamination with doublets and small cell clusters which can confound single-cell sequencing results. Flow cytometry can also be used to evaluate whether cell populations of interest, such as immune cells, stromal fibroblasts, or stem cell populations, are maintained in the cell preparation and in the appropriate frequency. Finally, an additional metric on RNA quality can be acquired using the RNA integrity number (RIN) method (Schroeder et al., 2006).

SINGLE-CELL TRANSCRIPTOMIC PLATFORM

Protocols for transcriptome analysis have advanced rapidly, resulting in several robust methods which range in cell and mRNA capture strategy, barcoding, throughput, and level of automation (Fan et al., 2015; Macosko et al., 2015). Selection of the optimal approach depends largely on the research question. Recent high-throughput protocols for scRNAseq have dramatically increased scalability through automation, increasing the number of cells that can be processed simultaneously, and decreasing reagent cost through reaction miniaturization. Using microwell-based (Cytoseq, Wayfergen), microfluidics-based (Fluidigm C1 HT), or droplet-based (inDrop, Drop-seq, and 10× Chromium) approaches, hundreds to thousands of cells can be captured in a single experiment (Islam et al., 2014; Picelli et al., 2014; Klein et al., 2015; Heath et al., 2016; Zheng et al., 2017). The newest of these protocols utilize beads functionalized with oligonucleotide primers, which each contain a universal PCR priming site, a cell-specific barcode, an mRNA capture sequence, and Unique Molecular Identifiers (UMI). Individual cells are captured in wells or droplets with a single bead. Cell-specific barcode are similar within a droplet but unique UMI sequence on the primer allows for individual transcripts within

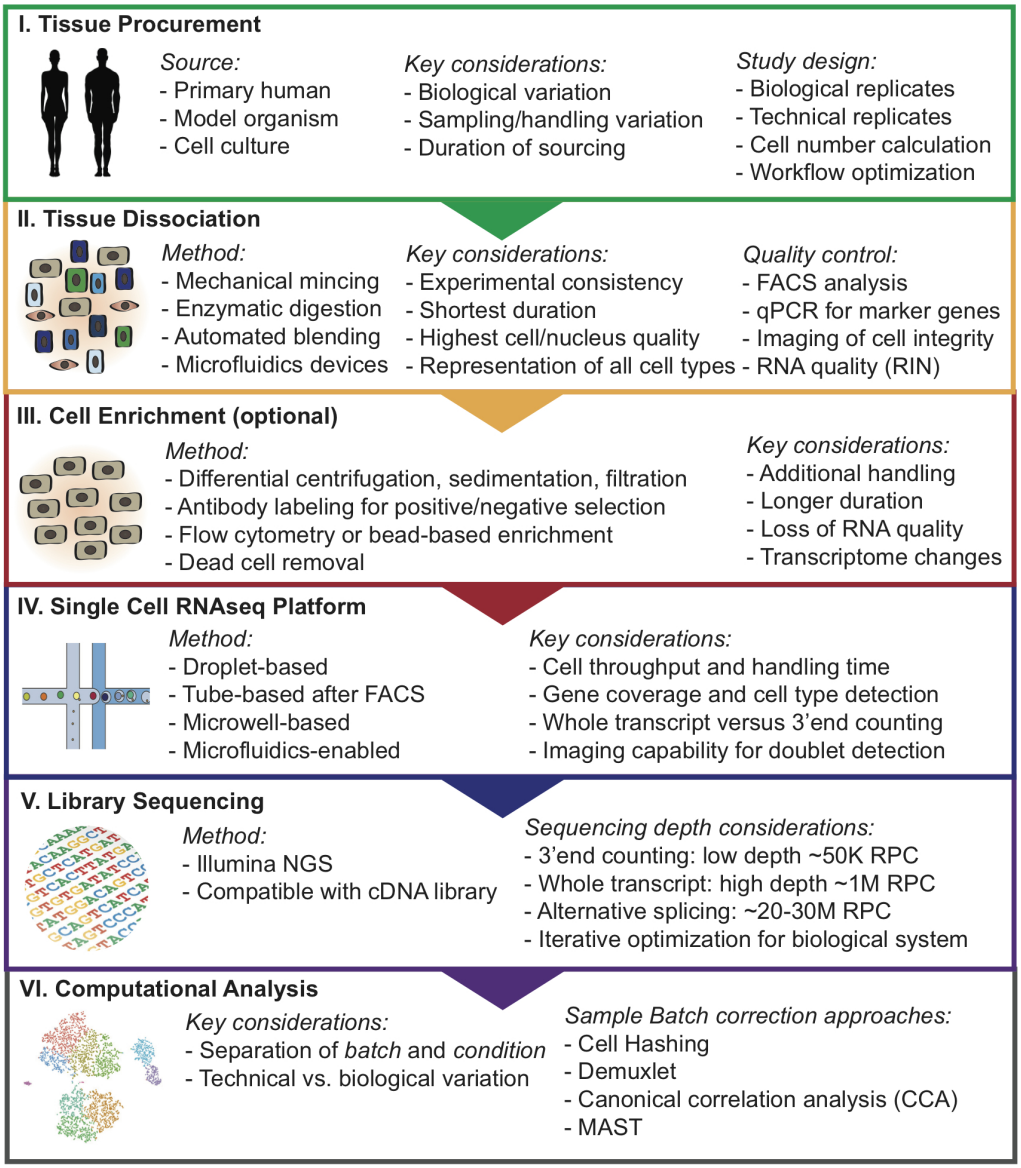


FIGURE 1 | Overview of step-wise approach to designing single-cell analysis workflows. RNA integrity number (RIN); Reads per cell (RPC).

a cell to be counted. This provides a quantitative readout of the number of transcripts of each gene detected in a cell, thereby reducing the effects of amplification duplicates that occur with earlier technologies (Ramsköld et al., 2012; Patel et al., 2014). High-throughput 3′-end counting approaches have several important limitations. Since only the 3′-end of each mRNA are sequenced, differential splicing analyses are not feasible (Macosko et al., 2015; Heath et al., 2016). High-throughput approaches typically only achieve ∼10% transcriptome coverage, relative to ∼40% for full-length scRNAseq protocols that use Switching Mechanism at 5′End of RNA Template (SMART) chemistry (Tirosh et al., 2016; Yuan et al., 2017). This is partly due to lower mRNA capture efficiency, but also due to lower sequencing depth. Single-cell qPCR platforms (e.g., Fluidigm C1 and Biomark) remain superior in sensitivity for detecting low-expressed genes (Lawson et al., 2015).

Protocols for processing rare cells usually involve an upstream capture step by flow cytometry or micromanipulation, followed by dispensing single cells into microtubes or microwell plates. Studies investigating rare cell populations that require selection via specific markers (e.g., adult tissue stem cell populations), are best performed using these protocols. Single-cell libraries are prepared using SMART-based chemistry, which utilizes a template-switching oligonucleotide (TSO) (Tirosh et al., 2016). This TSO can be used to prime off of the untemplated nucleotides added by the reverse transcriptase, enabling subsequent PCR using a single primer and capture of full length transcripts (Tirosh et al., 2016; Yuan et al., 2017). cDNAs are then

amplified by PCR and libraries are prepared for sequencing using standard protocols. Although there have been several large scale projects utilizing these protocols, because they are manual in nature and utilize larger microliter reaction volumes, they limit the number of cells that can be processed at reasonable cost.

Another area of ongoing debate is how to determine how many cells one should be analyzed to reach sufficient statistical power. Several methods have been developed using power analysis statistics, such as Scotty[1] or web-based tools[2], but one must estimate the number and expected frequencies of cell populations present in the sample, and such information is often not available. Therefore, these decisions are usually made based on logistical restraints (i.e., the number of cells available), financial considerations, or re-iterative experiments where an initial sample of cells is sequenced to get a sense for overall population structure, and then increasing numbers of cells are sequenced until one is satisfied that all the main populations have been identified.

SINGLE NUCLEI ISOLATION AND SEQUENCING

Single-cell RNA sequencing methods are optimal when cells can be harvested intact and viable (Grindberg et al., 2013). However, certain cell types (e.g., neurons, adipocytes), are not amenable to standard organ dissociation protocols, since enzymatic and mechanical forces easily disrupt the cytoplasmic contents (Habib et al., 2017). In these cases, an option could be to isolate intact nuclei for single-nucleus RNAseq (snRNAseq) (Grindberg et al., 2013; Habib et al., 2016, 2017; Krishnaswami et al., 2016; Lacar et al., 2016; Lake et al., 2016). To prepare single nuclei, cells are lysed with detergent and dounce homogenized to expel cytoplasmic contents and nuclei from the cellular membrane, (Habib et al., 2016), which may avoid transcriptomic changes (Van Den Brink et al., 2017). Nuclei can then be purified by flow cytometry or gradient centrifugation (Grindberg et al., 2013; Ambati et al., 2016; Habib et al., 2016). When cell-type specific nuclear proteins exist, they can be used for nuclei isolation from specific cell types using antibody labeling (Lacar et al., 2016; Habib et al., 2017).

Single-nucleus RNAseq is not only amenable for difficult to isolate cell types, but can also be used for archived tissues such as flash-frozen clinical samples. Individual nuclei isolated from frozen adult mouse and human brain tissues have been successfully sequenced, demonstrating that snRNAseq has sufficient resolution to identify many different cell types from frozen and post-mortem tissue (Grindberg et al., 2013). With the rapid development of many applications for snRNAseq, nuclei are amenable to other studies not easily done by scRNAseq.

An important question remains: To what degree is the nuclear transcriptome representative of the whole cell? Recent studies have demonstrated that many transcripts of cell and nucleus are equally represented and that nuclear RNA represents an important and significant population of transcripts that contribute greatly to the overall diversity of transcripts (Barthelson et al., 2007; Trask et al., 2009). Comparative studies of scRNAseq and snRNAseq in neural progenitor cells have also demonstrated that genes are expressed in equal proportion between whole cell and nuclei (Grindberg et al., 2013). Nanogrid single-cell and nuclei RNA sequencing studies in the same breast cancer lines found that overall copy number, expression level, and abundance had a high (r_s = 0.95) Spearman's correlation (Gao et al., 2017). Similarly, the transcriptomes of single cells and nuclei of 3T3 cells have also demonstrated high correlation (Pearson, r = 0.87) (Habib et al., 2017). Together these results suggest that nuclei and cells have highly correlated relative gene expression.

Despite the similarities between single-cell and nuclei transcriptomic profiles there remain notable differences. Not surprisingly, nuclear transcriptomes are enriched for several types of nuclear RNAs (Grindberg et al., 2013; Habib et al., 2016, 2017; Krishnaswami et al., 2016; Gao et al., 2017). Since ncRNAs are only polyadenylated in the nucleus, snRNAseq provides a feasible strategy to capture the heterogeneity of ncRNA transcription in single-cell resolution (Krishnaswami et al., 2016). In addition, nuclear transcriptomes are enriched for lncRNAs and nuclear-function genes (Gao et al., 2017). Another difference between cell and nuclear RNAseq is the higher abundance of intronic sequences in snRNAseq, which ranged between 10–40% of mapped reads (Grindberg et al., 2013; Gao et al., 2017; Habib et al., 2017). These features need to be accounted for when comparing datasets from cellular versus nuclear transcriptome analyses.

In conclusion, snRNAseq has emerged as a promising avenue for profiling archived samples or cell types that are hard to viably isolate from tissues.

SINGLE-CELL LIBRARY SEQUENCING

The next critical part of designing single-cell workflows is to align the analysis pipeline with the respective NGS platform and sequencing depth. It is important to confirm that the chemistry used for library construction is compatible with the sequencing technology. Currently, there are two major outputs for libraries from scRNAseq: full-length transcript or 3′-end counted libraries, which each require different read depths (Haque et al., 2017). Full-length transcript libraries are typically sequenced at a depth of 10^6 reads per cell, but may still yield important biological information at as low as 5×10^4 reads per cell (Pollen et al., 2014). For specific applications such as alternative splicing analysis on the single-cell level, much higher sequencing depth up to 15– 25×10^6 reads per cell is necessary. On the other hand, 3′-end counting libraries are sequenced at much lower depth of around 10^4 or 10^5 reads per cells (Haque et al., 2017). Reaching the optimal sequencing depth can be an iterative process and may require multiple rounds of optimization. Sequencing saturation can be estimated by plotting down-sampled sequencing depth in mean reads per cell (e.g., 10× Genomics Cell Ranger).

[1] http://scotty.genetics.utah.edu/

[2] http://satijalab.org/howmanycells

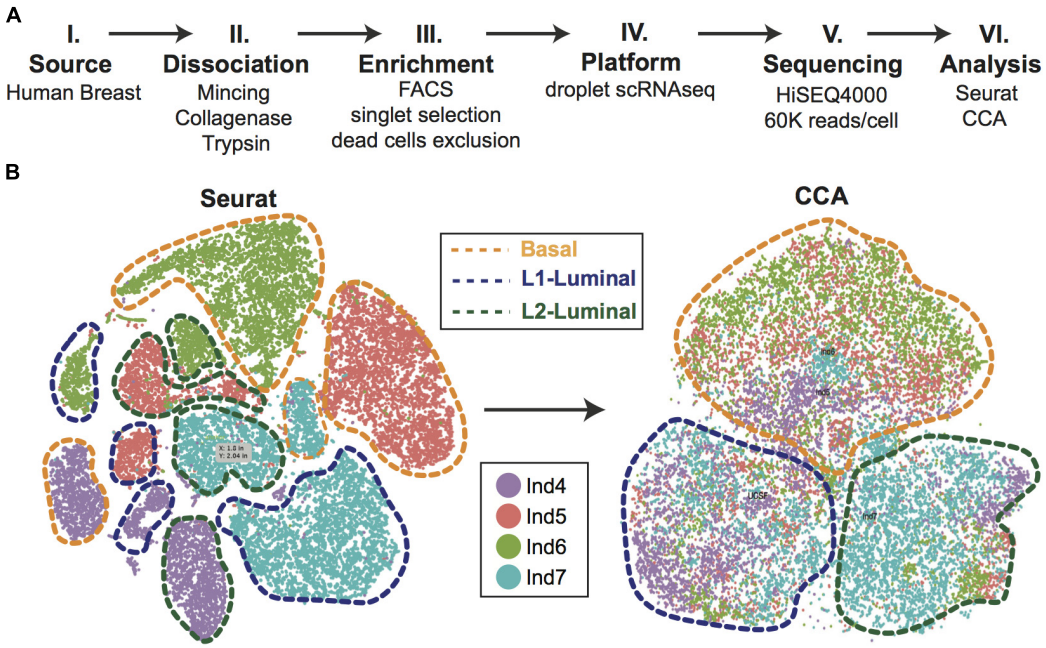


FIGURE 2 | Single-cell analysis of primary human breast epithelial cells. **(A)** Approach overview summarizing individual steps of single-cell analysis approach from primary human breast epithelial cells using scRNAseq. **(B)** Combined computational analysis of 24,465 single-cell transcriptomes from primary breast epithelial cells harvested from four human individuals. Standard Seurat analysis shows clear separation by individual as shown in tSNE plot. Applying canonical correlation analysis (CCA) successfully removes individual-specific clustering giving rise to three major clusters corresponding to the three main breast epithelial cell types, namely Basal, L1-Luminal, and L2-Luminal (outlined by dotted lines).

STUDY DESIGN AND DATA ANALYSIS

In the following section, we highlight several key considerations from a data analysis perspective for adequately designing a successful scRNAseq study. As mentioned, many single-cell technologies can be greatly affected by technical variation, and without proper study design the results can be difficult to interpret. One critical aspect of this is the separation of *batch* and *condition*. *Batch* refers to a library that was singularly generated in a contained workflow (i.e., harvesting tissue specimen, disassociating into single-cell suspension, and generating scRNAseq library). *Condition* refers to a biological state or experimental treatment that is being analyzed in the study. Technical variation can be difficult to separate from relevant biological variation when conditions are interrogated individually. To help correct for this, the generation of replicates (biological or technical) whenever possible is strongly recommended.

In addition to replicates, an option is to mix samples and conditions within a batch, such that they can be treated without confounding each other (Hicks et al., 2015). One example is the Demuxlet workflow, where samples from genetically distinct individuals can be processed within the same library generation protocol and sequenced together (Kang et al., 2018). Prior to library generation, genotyping of distinct samples is performed and subsequently used in conjunction with the scRNAseq library to demultiplex the mixed cell sample into the samples of origin. In situations where genetically identical samples are used, or genotypic data is not readily available, cellular hashing can be employed (Stoeckius et al., 2017). This involves oligo-tagged antibodies specific to each sample in the study and then pooling and generating the scRNAseq library from the sample mixture. The antibodies labeled with unique barcodes can be traced back to its sample of origin (Stoeckius et al., 2017).

Efforts can be made computationally to mitigate batch-to-batch variation. Batch effects are not unique to scRNAseq data, but the assumptions made by correction algorithms are not always appropriate for the bimodality of gene expression in zero-inflated scRNAseq data. Here, we highlight recent analytical frameworks that may be used to correct for this phenomenon. A recently developed approach by Haghverdi et al. (2018) builds a mixed nearest neighbor model for cells between datasets or samples that does not require known or equal proportions of cell types between data sets. In addition, the widely used Seurat pipeline for scRNAseq analysis recently employed canonical correlation analysis (CCA) that allows for discovery of co-correlated gene modules between datasets that can then be used to cluster upon (Butler et al., 2018). This approach identifies the cell types common between datasets and samples, as well as those that are unique to an individual set by finding common sources of variation in gene expression. As an illustration of this method, we applied CCA to our recently published droplet-enabled scRNAseq dataset from four individual primary human breast tissue samples (**Figure 2**). Finally, the single-cell batch correction framework MAST (Finak et al., 2015) models the positive expression mean and the over-the-background

expression of transcripts, and calculates a fraction of detected genes per cell and uses this as a covariate that is independent of a previously specified control set of genes. Together, these methods serve as recent examples to handle batch-to-batch variation computationally, resulting in improved dimensionality reduction and clustering for meaningful scRNAseq data analysis.

Beyond accounting for technical variation, a common question that researchers address is the relatedness of described cell populations through the lens of a differentiation processes. The key assumption of pipelines that seek to address this is that the tissue sample analyzed using scRNAseq contains cell types/states that represent not only the ends of a differentiation process, but also stem/progenitor cells and transitional cell states along the path of differentiation. Common analysis suites that seek to reconstruct these differentiation trajectories are Monocle (Qiu et al., 2017), TSCAN (Ji and Ji, 2016), and CellTree (duVerle et al., 2016). Each use different methods, but their goal is to visualize differentiation trajectories and identify expression signatures that change through pseudotime.

CONCLUSION

To fully harness the potential of single-cell analysis tools to decipher complex biological systems on the level of individual cells, careful study design and rigorous optimization of every step along the experimental procedure are mandatory. Here, we delineate a step-wise experimental approach for optimizing tissue handling, cell dissociation and enrichment, single-cell platform selection, library sequencing, and data analysis for designing single-cell workflows. A move toward standardized and automated processing of tissues will minimize changes introduced by tissue handling that may obscure biologically relevant transcriptomic profiles. For tissues that are problematic to dissociate into high-quality and viable single-cell suspensions, snRNAseq offers a solution to this problem, and can be used to achieve uniform extraction and sequencing of multiple cell types for cross comparison. Numerous computational frameworks are currently emerging that help mitigate batch effects to separate biological variation from unwanted technical variation.

AUTHOR CONTRIBUTIONS

KK outlined concept and overview of review. QN, NP, and KN wrote the manuscript. KK and QN designed and prepared the figures.

FUNDING

This study was supported by funds from the National Cancer Institute (R00 CA181490), Chan/Zuckerberg Initiative (HCA-A-1704-01668), and the University of California Cancer Research Coordinating Committee (CTN-18-515073).

REFERENCES

Alles, J., Karaiskos, N., Praktiknjo, S. D., Grosswendt, S., Wahle, P., Ruffault, P. L., et al. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* 15:44. doi: 10.1186/s12915-017-0383-5

Ambati, S., Yu, P., McKinney, E. C., Kandasamy, M. K., Hartzell, D., Baile, C. A., et al. (2016). Adipocyte nuclei captured from VAT and SAT. *BMC Obes.* 3:35. doi: 10.1186/S40608-016-0112-6

Baldan, V., Griffiths, R., Hawkins, R. E., and Gilham, D. E. (2015). Efficient and reproducible generation of tumour-infiltrating lymphocytes for renal cell carcinoma. *Br. J. Cancer* 112, 1510–1518. doi: 10.1038/bjc.2015.96

Barthelson, R. A., Lambert, G. M., Vanier, C., Lynch, R. M., and Galbraith, D. W. (2007). Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. *BMC Genomics* 8:340. doi: 10.1186/1471-2164-8-340

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490. doi: 10.1038/nature14590

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Christaki, E., Opal, S. M., Keith, J. C., Kessimian, N., Palardy, J. E., Parejo, N. A., et al. (2011). A monoclonal antibody against RAGE alters gene expression and is protective in experimental models of sepsis and pneumococcal pneumonia. *Shock* 35, 492–498. doi: 10.1097/SHK.0b013e31820b2e1c

duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H., and Tsuda, K. (2016). CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics* 17:363. doi: 10.1186/s12859-016-1175-6

Fan, H. C., Fu, G. K., and Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science* doi: 10.1126/science.1258367 [Epub ahead of print].

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5

Gao, R., Kim, C., Sei, E., Foukakis, T., Crosetto, N., Chan, L. K., et al. (2017). Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* 8:228. doi: 10.1038/s41467-017-00244-w

Grindberg, R. V., Yee-Greenbaum, J. L., McConnell, M. J., Novotny, M., and O'Shaughnessy, A. L. (2013). RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19802–19807. doi: 10.1073/pnas.1319700110

Habib, N., Avraham-davidi, I., Basu, A., and Burks, T. (2017). Massively-parallel single nucleus RNA-seq with DroNc-seq. *Nat. Methods* 14, 955–958. doi: 10.1038/nmeth.4407

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., et al. (2016). Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* 353, 925–928. doi: 10.1126/science.aad7038

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the mouse cell atlas by microwell-Seq. *Cell* 172, 1091.e17–1097.e17. doi: 10.1016/j.cell.2018.02.001

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9, 1–12. doi: 10.1186/s13073-017-0467-4

Heath, J. R., Ribas, A., and Mischel, P. S. (2016). Single-cell analysis tools for drug discovery and development. *Nat. Rev. Drug Discov.* 15, 204–216. doi: 10.1038/nrd.2015.16

Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* [Preprint]. doi: 10.1101/025528

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi: 10.1038/nmeth.2772

Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44:e117. doi: 10.1093/nar/gkw430

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94. doi: 10.1038/nbt.4042

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044

Kornbluth, J., and Hoover, R. G. (1989). "Anti-HLA class i antibodies alter gene expression in human natural killer cells," in *Immunobiology of HLA*, ed. B. Dupont (New York, NY: Springer), 150–152.

Krishnaswami, S. R., Grindberg, R. V., Novotny, M., Venepally, P., Lacar, B., Bhutani, K., et al. (2016). Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* 11, 499–524. doi: 10.1038/nprot.2016.015

Lacar, B., Linker, S. B., Jaeger, B. N., Krishnaswami, S. R., Barron, J. J., Kelder, M. J. E., et al. (2016). Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* 7:11022. doi: 10.1038/ncomms11022

Lake, B., Shen, R., Ronaghi, M., Fan, J., Wang, W., and Zhang, K. (2016). Neuronal subtypes and diverstiy revealed by single-nucleus RNA sequencing of human brain. *Science* 352, 1586–1590. doi: 10.1126/science.aaf1204

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* 526, 131–135. doi: 10.1038/nature15260

Lescroart, F., Wang, X., Lin, X., Swedlund, B., Gargouri, S., Sànchez-Dànes, A., et al. (2018). Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* 359, 1177–1181. doi: 10.1126/science.aao4174

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Meeson, A., Fuller, A., Breault, D. T., Owens, W. A., and Richardson, G. D. (2013). Optimised protocols for the identification of the murine cardiac side population. *Stem Cell Rev. Rep.* 9, 731–739. doi: 10.1007/s12015-013-9440-9

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–95. doi: 10.1038/nature09807

Nguyen, Q. H., Pervolarakis, N., Blake, K., Ma, D., Davis, R., James, N., et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* 9:2028. doi: 10.1038/s41467-018-04334-1

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 8–13. doi: 10.1126/science.1254257

Picelli, S., Faridani, O. R., Björklund, ÅK., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.* 9, 171–181. doi: 10.1038/nprot.2014.006

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Qiu, X., De Jesus, J., Pennell, M., Troiani, M., and Haun, J. B. (2014). Microfluidic device for mechanical dissociation of cancer cell aggregates into single cells. *Lab Chip* 15, 339–350. doi: 10.1039/c4lc01126k

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979–982. doi: 10.1038/nmeth.4402

Radbruch, A., and Recktenwald, D. (1995). Detection and isolation of rare cells. *Curr. Opin. Immunol.* 7, 270–273. doi: 10.1016/0952-7915(95)80014-X

Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282

Romero-santacreu, L., Moreno, J., Perez-Ortin, J. E., and Alepuz, P. (2009). Specific and global regulation of mRNA stability during osmotic stress in *Saccharomyces cerevisiae*. *RNA* 15, 1110–1120. doi: 10.1261/rna.1435709

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: from vision to reality. *Nature* 550, 451–453. doi: 10.1038/550451a

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7:3. doi: 10.1186/1471-2199-7-3

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B., Smibert, P., et al. (2017). Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *bioRxiv* [Preprint]. doi: 10.1101/237693

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338. doi: 10.1038/nature21350

Tirosh, I., Izar, B., Prakadan, S. M., Ii, M. H. W., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular exosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501. Dissecting

Trask, H. W., Cowper-Sal-lari, R., Sartor, M. A., Gui, J., Heath, C. V., Renuka, J., et al. (2009). Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs. *RNA* 15, 1917–1928. doi: 10.1261/rna.1677409

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375. doi: 10.1038/nature13173

Tung, P. Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., et al. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7:39921. doi: 10.1038/srep39921

Van Den Brink, S. C., Sage, F., Vértesy, Á, Spanjaard, B., Peterson-Maduro, J., Baron, C. S., et al. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* 14, 935–936. doi: 10.1038/nmeth.4437

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711

Will, B., and Steidl, U. (2010). Multi-parameter fluorescence-activated cell sorting and analysis of stem and progenitor cells in myeloid malignancies. *Best Pract. Res. Clin. Haematol.* 23, 391–401. doi: 10.1016/j.beha.2010.06.006

Xiong, L., Lee, H., Ishitani, M., and Zhu, J. K. (2002). Regulation of osmotic stress-responsive gene expression by the LOS6/ABA1 locus in *Arabidopsis*. *J. Biol. Chem.* 277, 8588–8596. doi: 10.1074/jbc.M109275200

Yuan, G. C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., et al. (2017). Challenges and emerging directions in single-cell analysis. *Genome Biol.* 18:84. doi: 10.1186/s13059-017-1218-y

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 01 March 2019
doi: 10.3389/fgene.2019.00008

The Impact of Heterogeneity on Single-Cell Sequencing

*Samantha L. Goldman[1,2], Matthew MacKay[1,2], Ebrahim Afshinnekoo[1,2,3], Ari M. Melnick[4], Shuxiu Wu[5,6] and Christopher E. Mason[1,2,3,7]**

[1] Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, United States, [2] The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, United States, [3] WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, United States, [4] Department of Medicine, Weill Cornell Medicine, New York, NY, United States, [5] Hangzhou Cancer Institute, Hangzhou Cancer Hospital, Hangzhou, China, [6] Department of Radiation Oncology, Hangzhou Cancer Hospital, Hangzhou, China, [7] The Feil Family Brain and Mind Research Institute, New York, NY, United States

The importance of diversity and cellular specialization is clear for many reasons, from population-level diversification, to improved resiliency to unforeseen stresses, to unique functions within metazoan organisms during development and differentiation. However, the level of cellular heterogeneity is just now becoming clear through the integration of genome-wide analyses and more cost effective Next Generation Sequencing (NGS). With easy access to single-cell NGS (scNGS), new opportunities exist to examine different levels of gene expression and somatic mutational heterogeneity, but these assays can generate yottabyte scale data. Here, we model the importance of heterogeneity for large-scale analysis of scNGS data, with a focus on the utilization in oncology and other diseases, providing a guide to aid in sample size and experimental design.

Keywords: single-cell sequencing, heterogeneity, scRNA-seq, NGS, RNA, single cells

OPEN ACCESS

Edited by:
Xinghua Victor Pan, Southern Medical University, China

Reviewed by:
Saheli Sarkar, Northeastern University, United States
Guangshuai Jia, Max-Planck-Institut für Herz- und Lungenforschung, Germany

***Correspondence:**
Christopher E. Mason
chm2042@med.cornell.edu

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *20 July 2018*
Accepted: *09 January 2019*
Published: *01 March 2019*

Citation:
Goldman SL, MacKay M, Afshinnekoo E, Melnick AM, Wu S and Mason CE (2019) The Impact of Heterogeneity on Single-Cell Sequencing. Front. Genet. 10:8. doi: 10.3389/fgene.2019.00008

INTRODUCTION

It has been well-documented, both theoretically (Elsasser, 1984) and experimentally, that nearly all cellular systems are heterogenous (Altschuler and Wu, 2010). Heterogeneity may arise for a number of different reasons, and at many different levels, in order to improve survival and functionality. Both single-celled and multicellular organisms employ population-level survival strategies such as bet-hedging in order to achieve a better chance of survival when faced with new stresses though having a diverse population (Grimbergen et al., 2015). At a single-organism level, diversity further enables the existence of specialization and, within metazoan organisms, differentiation (Hadjantonakis and Arias, 2016).

Cellular heterogeneity can be measured in several different ways, most commonly via genomic, epigenomic, transcriptomic, and proteomic studies. However, the level of heterogeneity at one level of expression or regulation may not be the same at another level. Cells within a given person have nearly identical genomes, yet through specific modifications throughout development and disease, may generate many distinct cell types with unique expression profiles. Even the genome itself may be specifically rewired to generate increased genetic diversity within specific cell types, most notably B- and T-cells through V(D)J recombination. Uncovering the true diversity of cells is crucial to better understand cellular communication and responsibility within both healthy and disease states. It is now well understood that differentiation throughout development allows for

the necessary cellular specialization required for complex multicellular system function. Further, specific epigenomic modifications allow for this precise differentiation which inevitably results in the cascade of cellular diversity present in humans, and also is important in cancer (Li et al., 2014, 2016).

Next generation sequencing (NGS) is continuously being used more and more due to its rapidly decreased costs and ability to generate a large amount of data (Mason et al., 2014), with new data sets even being generated in zero gravity (McIntyre et al., 2016; Castro-Wallace et al., 2017). Within bulk-NGS analyses, many, typically hundreds of thousands to millions, of cells are analyzed at once. This generates an averaged picture of a given population of cells, and thus majority of our understanding of different cell and tissue types comes from the analysis of bulk experimentation which may underestimate the true heterogeneity of cells. Bulk-NGS is simply ill-equipped to address some important questions revolving around cellular heterogeneity. Single-cell NGS (scNGS) attempts to resolve issues facing bulk-NGS through the ability to relate sequences to a given cell, across the genetic, transcriptomic, epigenomic, and proteomic levels. This approach reduces the issue of data generalization which is prevalent in some bulk-NGS studies. However, scNGS is not without its faults. One of the main issues with scNGS is its cost and, though it has considerably decreased in recent years, it is still a large factor when designing experimentations, as well as technical issues and challenges in sensitivity. Here, we will outline the importance of cellular heterogeneity, assess factors of scNGS heterogeneity, and provide a practical sample size guide to aid in experimental design.

THE IMPORTANCE OF CELLULAR HETEROGENEITY

Having a heterogeneous (i.e., diverse) population is beneficial for cellular systems for the same reason why it is beneficial for there to be variation among many organisms in a single species – bet-hedging (Beaumont et al., 2009). Bet-hedging is a population-level survival strategy in which less-fit individuals are maintained in a population as a precaution; if the environment were to drastically change, the originally less-fit organisms may be adapted to the new environment, thereby assuring the survival of the population (Grimbergen et al., 2015). In an ever-changing environment, a population has a greater overall fitness if there is greater diversity. In this way, the evolution adaptation of all cellular systems can be modeled in terms of Darwinian evolution.

There are many causes of cellular heterogeneity. Firstly, populations of cells will naturally contain individuals that develop random mutations. These unique subclones can become significant portions of the population if that mutation confers a selective advantage and proliferates. However, not all cellular heterogeneity is genetic. Rather, much heterogeneity is phenotypic, and is frequently expressed in transcriptomes that vary from cell to cell. This heterogeneity can arise via external or internal factors. Extrinsic heterogeneity can lead to phenotypic plasticity in response to an environmental change, and only affects the part of the population that is exposed to the causative environment (Huang, 2009). It can also include variables such as cell-cycle stage and cell size (Singh and Soltani, 2013). Intrinsic heterogeneity is a more nuanced phenomenon, and is a result of stochastic events, such as gene expression noise (Huang, 2009), rather than a changing intracellular environment (Elowitz et al., 2002).

Because of stochastic gene fluctuation, there are varying levels of protein abundance in different cells in a population at any given time. This is most easily visualized via flow cytometry, which yields a bell-shaped curve (Brock et al., 2009). Stochastic gene expression may have its evolutionary advantages, as well. In the same way that populations of cells maintain random mutations in bet-hedging, populations of clonal, unicellular organisms may maintain variation via stochastic gene expression to ensure overall survival (Raj and van Oudenaarden, 2008). Although stochastic gene expression is a significant contributor to heterogeneity, it is not the only cause. The sub-state of any given genome/cell depends on a number of factors, including epigenetics, alternative splicing sites, post-translational modifications, and sometimes even microbial interactions (Shabaan et al., 2018). These processes are not always stochastic, and can therefore lead to "directed" heterogeneity, instead of the more random "non-directed" heterogeneity of stochastic gene expression (Chang and Marshall, 2017).

Interestingly, non-genetic, cellular heterogeneity also plays an important role in development. Early in the developmental process, before the small population of cells is beginning to differentiate, these cells are theoretically identical. However, as the cells begin to differentiate, they display non-genetic heterogeneity. The body of research on the role of heterogeneity in development is largely focused on transcriptional heterogeneity (Griffiths et al., 2018), which is a driver of differentiation of pluripotent stem cells. More recent work has also shown that RNA modifications, called the epitranscriptome (Saletore et al., 2012), can also lead to differential response of human cells to both disease and infection (Gokhale et al., 2016; Vu et al., 2017). Also, some transcriptional sub-states are heritable through several generations of cell divisions. Signaling factors, developmental regulators, and chromatin regulators contribute to transcriptional heterogeneity in stem cells (Kumar et al., 2014). "Directed" heterogeneity has been shown to lead the process behind the development of a body plan in *Drosophila melanogaster* (Chang and Marshall, 2017).

Even after development, all human tissue systems experience some level of differentiation. This allows cells to specialize, leading to a more flexible biological system. This principle has been most notably studied in the nervous and immune systems. In the central nervous system, for instance, there are dozens of different types of neurons. Subsets of these neurons form the myriad different regions within the brain (Emery and Barres, 2008). One phenotypic hallmark of heterogeneity in the nervous system, for example, is the distribution of mitochondria within the neuron. This heterogeneity is exhibited both regionally within the brain (e.g., brain regions that require more energy are composed of neurons with more mitochondria) (Dubinsky, 2009) and within individual neurons. This distribution differs greatly depending on the immediate

and current needs of the neuron, and is regulated by a complex system of proteins (Course and Wang, 2016). In the immune system, monocytes, macrophages (Gordon and Taylor, 2005), B-cells, and T-cells show heterogeneity. As an example, T-cell heterogeneity is essential for an effective immune response, since subtle differences in T-cell receptors (TCRs) enable the identification and elimination of foreign invaders (Durlanik and Thiel, 2015). However, in autoimmune disease, faulty TCR diversification can result in the improper identification of "self" as an invader resulting in normal tissue destruction.

Different diseases leverage heterogeneity to their advantage. A "survival of the fittest" model for cellular heterogeneity can be applied not only to populations of single-celled organisms, but also to tumors. Cancer cells continuously acquire and pass down genetic and epigenetic modifications to subsequent generations of cancer cells resulting in heterogeneity. These genetic mutations and epigenetic shifts may further lead to changes in fitness (Li et al., 2016). Cancer cells are often exposed to hostile environments, such as chemotherapy and radiation, during treatment (Afshinnekoo and Mason, 2016). Through bet-hedging, and therefore maintenance of a heterogeneous population, the chance of resistance or relapse from treatment is dramatically increased. As these cancer cells are all in the same small environment and are all competing for the same limited resources, there are complex interactions between different subclones that further reinforce these Darwinian relationships (Tabassum and Polyak, 2015). Cancer cells can be further driven into a "survival of the fittest" scenario via treatment with a chemotherapeutic drug, as this may lead to the selection for cancer-variants that are resistant to the drug. Over time, this could lead to chemotherapeutic resistance within the whole tumor (Dagogo-Jack and Shaw, 2017), as well as tumor subtypes (Shih et al., 2017). Indeed, it has been shown that a single tumor biopsy dramatically underrepresents the genetic diversity present within an entire tumor (Gerlinger et al., 2012). However, heterogeneity is not only clinically relevant in regards to chemotherapy. Immunotherapies can also be profoundly impacted by heterogeneity. Liver cancer-targeted immunotherapy is designed around tumor-infiltrating T-cells. Through the use of single-cell RNA sequencing, 11 tumor-infiltrating T-cell sub-states have been identified. Each of these sub-states has a unique profile of up- and downregulated genes, which may impact the efficacy of any immunotherapies (Zheng et al., 2017).

Intratumoral heterogeneity has been extensively studied through single-cell sequencing methods. For example, single-cell RNA sequencing has revealed significant heterogeneity in primary glioblastomas (Patel et al., 2014). Additionally, increased levels of heterogeneity in these tumors was inversely correlated with survival, indicating that intratumor heterogeneity should be an essential clinical factor, including events from DNA transposition (Henssen et al., 2017). Metastatic melanoma is also highly transcriptionally heterogeneous, and this heterogeneity is multifaceted; it is associated with a number of factors, including cell cycle stage, location, and chemotherapeutic resistance (Tirosh et al., 2016). The use of RNA sequencing here is key, as transcriptomics captures fine details of non-genetic heterogeneity that other sequencing methods may have missed. Shifting of cellular heterogeneity is not just a hallmark of cancer, but of many other diseases, but here we will focus on the relevance for cancer.

ASSESSING HETEROGENEITY

Heterogeneity itself is a gradient which may be based on variable changes in the transcriptome or more permanent changes within the genome. Differences seen between cells may be temporal due to cell-cycle states, or spatial due to external stimuli (Dagogo-Jack and Shaw, 2017). Also, differences between cells may exist at any processing level of the cell, from the genome to transcriptome to proteome, or due to any additional modifications which may exist. With this in mind, it could be possible to define all cells as heterogeneous. However, two disparate cells might not behave functionally different, and their heterogeneity would therefore not be considered impactful (Altschuler and Wu, 2010). The overall assessment of cellular heterogeneity is therefore context-specific and the technologies used to assess cellular differences need to be considered carefully.

Proteomic and cell-marker classification has been historically used to discern cell types. Immunohistochemistry (IHC) can be used to distinguish immune cell types within healthy systems (Reuben et al., 2017b) or even the cancer subtyping such as HER2 expression within breast cancer (Potts et al., 2012). Surface markers help to distinguish cell types into broad classification, but this type of analysis required prior gene expression knowledge and specific antibody usage. Other approaches, such as whole genome sequencing (WGS), bisulfite sequencing, and RNA sequencing, allow for genome-wide analysis (Mason et al., 2017). Historically these techniques are done on heterogeneous tissue samples, generating an averaged picture of the tissue of interest (bulk-NGS). Although bulk-NGS has a tendency to generalize heterogeneity, certain biological understanding and computational modeling can mitigate this effect within genomic and epigenomic analyses.

Bulk-WGS can be directly used to assess the existence of subclonal mutations through the use of variant allele frequencies (VAFs). Through the modeling of VAFs and copy number changes, an understanding of the clonal architecture may be inferred from such bulk-NGS data. One such method, *Canopy*, uses a Bayesian analysis to identify subpopulations and build a phylogenetic tree detailing their likely evolutionary history (Jiang et al., 2016). Long read bulk TCR sequencing can also be used directly to assess clonal structures under the assumption that there is a unique V(D)J recombination per subclone. As such, the quantity of a given TCR gene can be directly related to the abundance of that subclone and the number of different TCR genes relates to the overall heterogeneity and diversity of the T-cell population. TCR sequencing has also been used, and has shown intratumoral heterogeneity in localized lung carcinomas, which may confer post-surgical recurrence (Reuben et al., 2017a). As epigenetics also plays a significant role in heterogeneity, bisulfite sequencing can be used to study patterns of DNA methylation and estimate clonality, such as with the algorithm methclone (Li et al., 2014). Bisulfite sequencing has also been

used to reveal heterogeneity in DNA methylation of the *MLH1* (a mismatch repair gene) promoter across several endometrial tumors (Varley et al., 2009).

While many bulk-NGS methods rely on mixture models of the VAFs to analyze small indels and point mutations, these methods often rely on the copy number of the gene in question, which can be altered in cancers, and are unable to relate multiple mutations which exist at low frequencies (Jiang et al., 2016). Additionally, bulk sequencing has a tendency to report what an "average" cell in a population would look like and for that reason would not be usable in the analysis of an all-or-nothing response (Altschuler and Wu, 2010). For example, *Xenopus* oocytes, have a binary response when signaled by progesterone to begin a process of maturation; they either mature or they do not (Ferrell and Machleder, 1998). In this case, looking at an average of two distinct oocyte subpopulations – one that has been signaled to mature and one that has not – would artificially yield a biologically impossible "mean oocyte" that has committed to maturation half-way (Altschuler and Wu, 2010).

There has been a significant effort within the field to quantitatively measure heterogeneity and relate it to a functional change. One approach to this is to quantify stochastic gene expression. This has been done through dividing stochastic gene expression into its intrinsic and extrinsic components via a two-color reporter experiment and deriving analytical formulas to measure each component of noise (Singh and Soltani, 2013). Systems have also been developed to quantify the individual contribution of unique processes to stochastic gene expression, and therefore to heterogeneity. For example, experimentally generated models have been used to quantify the individual contribution to chromatin dynamics in isogenic chicken-cell populations (Viñuelas et al., 2013). Also, shifted gene expression dynamics have been shown to drive cell fate choice for hematopoietic progenitors (Kleppe et al., 2017), induced pluripotent stem cells (iPSCs), and the mouse inner-cell mass during embryogenesis (Mojtahedi et al., 2016; Bargaje et al., 2017; Mohammed et al., 2017).

UTILIZATION OF scNGS

To best understand cellular heterogeneity, single cells must be studied individually through the use of scNGS. Since assessing cellular co-occurrence is the main drawback of bulk-NGS, many studies have also been conducted to further elucidate clonal structures using single-cell DNAseq [including whole exome sequencing (WES) or WGS], bisulfite sequencing, and ATACseq (assay for transposable accessible chromatin, ATAC). Given the variability and importance of gene expression, sc-RNAseq is one of the most used single-cell sequencing techniques (**Supplementary Table S1**). Single-cell multi-omic analyses are also possible to uncover the true level of heterogeneity across expression levels within cells (Macaulay et al., 2017), which enable examination of the genome, transcriptome, and epigenome at once. scNGS has the ability to resolve noise in bulk-NGS through the additional ability to trace generated reads back to their cell of origin. Though, this added benefit comes at a steep monetary cost, as single-cell sequencing is still much more expensive than more traditional bulk NGS given the need to sequence more (**Supplementary Table S2**). Also, subpopulations of cancer cells can be found by scATAC-seq, which has the power to identify specific chromatin motifs. Indeed, when combined with RNA-seq, it has been used to identify epigenetic plasticity between two cell subpopulations (Litzenburger et al., 2017).

There are currently dozens of variations of techniques to study the genome, epigenome, transcriptome, and epitranscriptome of cells, and here, we focus on those most commonly in use (**Supplementary Table S1**). Each of these technologies has had a significant impact on numerous fields, including immunology, oncology, and microbiology. Because the scope of the benefits of single-cell analysis is so wide, there is tremendous pressure to advance the technologies in the field. This is evident in the dramatic increase in recent years in publications referencing single-cell technologies (Wang and Navin, 2015). These techniques are highly varied, from manual manipulation (Pan et al., 2013) to droplet microfluidics used for sc-WGS (Hosokawa et al., 2017) to the creation of an RNA-library (Hedlund and Deng, 2018), such as bisulfite sequencing, can also be used on the single-cell level (Clark et al., 2017). A novel approach that combines Raman spectroscopy with an algorithmic biomolecular component analysis (microRaman-BCA) allows for the profiling of single organelles from a cell. Because this technique does not destroy the cell during analysis, the study can be performed multiple times on the same cell, providing a better picture of heterogeneity over time (Kuzmin et al., 2017).

While much of the current knowledge of cellular heterogeneity is transcriptional, newer techniques such as single-cell epigenomics have tremendous potential to study heterogeneity (Hassan et al., 2017) and may be able to provide further insights into the characterization and mechanisms of heterogeneity (Clark et al., 2016). Several topics in epigenomics are best suited to study with single-cell methods, including the relationship between transcriptional heterogeneity and epigenetic heterogeneity, which may vary greatly from cell to cell. Another application of single-cell sequencing is to study tumor resistance and therapeutic response to decrease the chance of resistance or relapse. scNGS can be used to not only detect heterogenous subclones within a tumor, but also to characterize these cells. Additionally, it can be used to characterize metastases and to create an effective treatment plan that minimizes the chance of chemotherapeutic resistance of specific subclones (Liang and Fu, 2017). In one study, analysis via deep whole-exome sequencing revealed that 75% of relapsed tumors in pediatric B-acute lymphoblastic leukemia were descendants of originally rare subclones (Ma et al., 2015). Given technical and sampling limitations, it is possible that resistant subclones existed within more patients. Although scNGS is currently expensive, treatment for cancer is often much more expensive. For this reason, any possible technique that could lead to a more effective therapy (even an expensive one like scRNA-seq) has clinical potential (Shalek and Benson, 2017).

Additionally, subclones can communicate and interact with each other, leading to complex relationships that may only be fully elucidated via scNGS. Although some of these

interactions are neutral, they can also be positive (leading to a commensalistic/mutualistic relationships in which one or both of the subclones benefit), or negative (leading to competition between subclones, e.g.), and can contribute to the chemotherapeutic resistance of one or more subclones within a tumor (Tabassum and Polyak, 2015). For instance, one study demonstrated that various clonal lineages in a case of colorectal cancer responded differently to treatment with chemotherapy (Kreso et al., 2013). Additionally, there is evidence that parallel evolution of various subclones within a tumor can lead to polyclonal resistance (Gerlinger et al., 2014). Additionally, intra-tumor heterogeneity makes it more difficult to precisely identify either histologically or genetically a tumor via a traditional biopsy (Tellez-Gabriel et al., 2016).

The implications of tumor heterogeneity in cancer evolution, clinical treatment, and tumoral spatial organization are not yet fully understood (Alizadeh et al., 2015), but scNGS provides a mechanism for beginning to unravel these relationships. Although heterogeneity makes the histological and genetic identity of a tumor more ambiguous, if the mechanisms

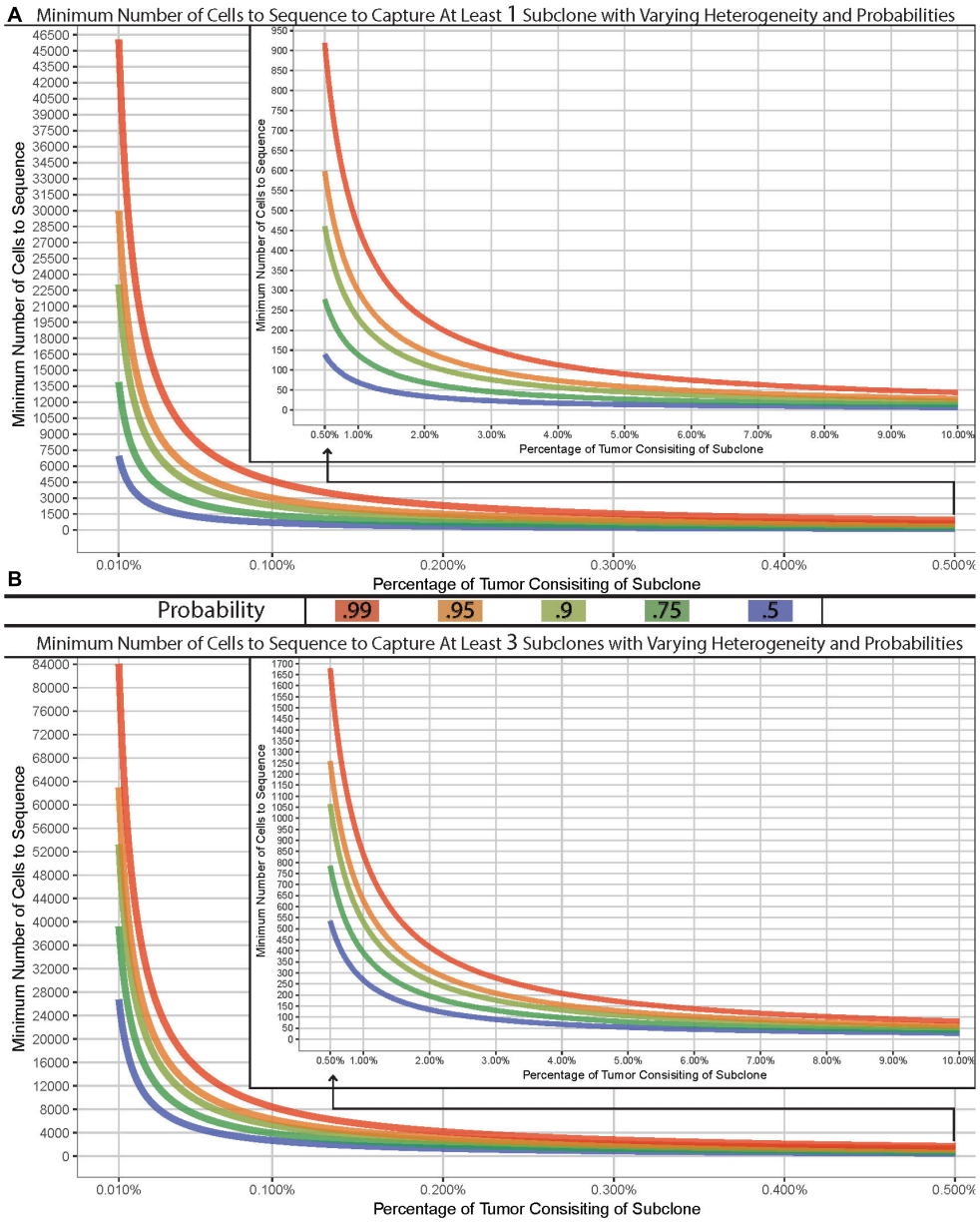


FIGURE 1 | Model of cells required for detection of variants. Minimum number of cells to sample to capture at least one **(A)** or three **(B)** subclone with varying probabilities (lines) across varying concentrations in a tissue with 1 billion cells. Hypergeometric calculations were done using R's phyper() function with lower.tail = F and $q = 0$ **(A)** or two **(B)** across varying sample sizes and clonal frequencies such that m+n = 1,000,000,000.

driving heterogeneity are further elucidated, they may lead to a better understanding of carcinogenesis (Gay et al., 2016). Moreover, data gathered from single-cell sequencing may help to clarify the methods of cancer progression and subclone resistance to chemotherapeutic treatment by sequencing both smaller transcripts and whole genomes in single cellular representatives of heterogeneous populations (Baslan and Hicks, 2017).

Interestingly, scNGS also has implications in lineage tracking in the development of differentiated tissues, as it may help to further clarify the developmental pathways involved in tissue differentiation (Kester and van Oudenaarden, 2018). As discussed, the nervous and immune systems are both well-studied examples of cellular systems that display cellular heterogeneity. For example, this technique can be used to study the central nervous system, and has the potential to not only molecularly classify various neurons or groups of neurons, but also to further study the molecular mechanisms behind, and possible therapies for, neurological diseases (Ofengeim et al., 2017). Indeed, this application can also be utilized to type sperm and oocytes, allowing for the confirmation and subsequent study of recombination events and polymorphisms in these haploids (Zhang et al., 1992).

DESIGN OF scNGS EXPERIMENTS

One of the key questions in planning the methodology of a single-cell study is how many cells to sequence. Sequencing more cells enables a greater representation of the cells in a population, giving a more accurate model of the diversity of subclones. The number of single-cells sequenced in a study has scaled exponentially with the development of new technologies. In 2009, for example, only one cell could be sequenced at a time. By 2017, however, the technology has advanced enough to permit the analysis of hundreds of thousands of cells at once (Svensson et al., 2018) and the possibility to generate exobytes and even yottabytes of data in the future.

Many complexities exist with scNGS analyses and need to be carefully considered. Other work have covered the specific differences, benefits, and drawbacks between the various scNGS protocols (Kanter and Kalisky, 2015; Clark et al., 2016; Haque et al., 2017; Liang and Fu, 2017). Previous data have shown that the best scNGS technology should be used for a given hypothesis, in tandem with a proper experimental design for the number of cells. Due to this, the required number of cells necessary to address a given question or tissue model will largely vary depending on the overall hypothesis. However, the question of "how many cells should I sequence" can be simplified to how many cells do you need to sample in order to capture at least one subclonal cell. The chance of sampling a subclone from a tissue of interest depending on the subclonal prevalence, the size of the tissue, and the size of the sample. Therefore, this question can be modeled using the hypergeometric distribution with varying degrees of probability (**Figure 1A**). It is common within sc-NGS analysis to require multiple cells to contain a given phenotype, and therefore may be more appropriate to ask the question of "how many cells should I sample to capture at least three subclonal cells" (**Figure 1B**).

We have built a model to demonstrate the number of cells required for a sampling design can widely vary. As an example, if the goal was to sample a tissue which has 1 billion cells for a previously undefined stem-cell which exist at a population of 0.01%, you would have a 99% chance of sampling at least one stem-cell if you analyzed approximately 46,000 cells. However, to truly characterize and identify this subclonal population or to detect a lower threshold, the number of cells required could easily reach, or even surpass, 100,000 depending on tissue size (**Figure 1B**). Given the recent advances in scNGS and decreases in costs, this is now possible to do. Such a design – while completely impossible 5 years ago – should be strongly considered when designing experimentations today.

THE FUTURE OF SINGLE-CELL ANALYSES

While single-cell sequencing has many advantages, it certainly is not a perfect technique. There are many different techniques for obtaining single-cell sequencing data and single-cell whole genome sequencing (sc-WGS), and each of these methods presents its own unique strengths and weaknesses. Multiple displacement amplification (MDA) and other PCR-based sequencing techniques often experience significant amplification bias (de Bourcy et al., 2014; Ahsanuddin et al., 2017). This could lead to incorrect interpretation of the prevalence and diversity of certain genes. Nonetheless, thanks to the breakthroughs in scNGS, the long-sought goal of sequencing of single cells is possible. This has created significant opportunities for advancement in the study of heterogeneity, especially as it applies to cancer. While it may be necessary to sample thousands or even millions of cells to encounter a unique subclone at low prevalence within a large tissue, sequencing continues to get cheaper, and thus scNGS will continue to open up many new research directions into the mechanisms of heterogeneity study variation on cell-by-cell resolution.

AUTHOR CONTRIBUTIONS

CM and SG conceived and designed the study. CM, SG, and MM analyzed the data. SG, MM, EA, AM, and CM wrote the paper. All authors, reviewed, edited, and approved the manuscript.

FUNDING

This work was supported by funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G and NNX17AB26G), the National Institutes of Health (R25EB020393, R01NS076465, R01AI125416, R01ES021006, 1R21AI129851, and 1R01MH117406), the Bill and Melinda Gates Foundation (OPP1151054).

ACKNOWLEDGMENTS

We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, as well as the Starr Cancer Consortium (I9-A9-071).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00008/full#supplementary-material

REFERENCES

Afshinnekoo, A., and Mason, C. E. (2016). Epigenetic therapy in a new era of medicine: creating and integrating molecular profiles of patients. *Ann. Transl. Med.* 4:436. doi: 10.21037/atm.2016.11.19

Ahsanuddin, S., Afshinnekoo, E., Gandara, J., Hakyemezoğlu, M., Bezdan, D., Minot, S., et al. (2017). Assessment of Repli-G Multiple Displacement Whole Genome Amplification (WGA) for Metagenomic Sequencing. *J. Biomol. Tech.* 28, 46–55. doi: 10.7171/jbt.17-2801-008

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., et al. (2015). Toward understanding and exploiting tumor heterogeneity. *Nat. Med.* 21, 846–853. doi: 10.1038/nm.3915

Altschuler, S. J., and Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell* 141, 559–563. doi: 10.1016/j.cell.2010.04.033

Bargaje, R., Trachana, K., Shelton, M. N., McGinnis, C. S., Zhou, J. X., Chadick, C.,et al. (2017). Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc. Natl. Acad. Sci. U.S.A.* 114, 2271–2276. doi: 10.1073/pnas.1621412114

Baslan, T., and Hicks, J. (2017). Unravelling biology and shifting paradigms in cancer with single- cell sequencing. *Nat. Rev. Can.* 17, 557–569. doi: 10.1038/nrc.2017.58

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C., and Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature* 462, 90–93. doi: 10.1038/nature08504

Brock, A., Chang, H., and Huang, S. (2009). Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet*. 10, 336–342. doi: 10.1038/nrg2556

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Nat. Sci. Data* 7:18022. doi: 10.1038/s41598-017-18364-0

Chang, A. Y., and Marshall, W. F. (2017). Organelles – understanding noise and heterogeneity in cell biology at an intermediate scale. *J. Cell Sci.* 130, 819–826. doi: 10.1242/jcs.181024

Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* 17:72. doi: 10.1186/s13059-016-0944-x

Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome- wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* 12, 534–547. doi: 10.1038/nprot.2016.187

Course, M. M., and Wang, X. (2016). Transporting mitochondria in neurons. *F1000Res* 5:F1000 Faculty Rev-1735. doi: 10.12688/f1000research.7864.1

Dagogo-Jack, I., and Shaw, A. T. (2017). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi: 10.1038/nrclinonc.2017.166

de Bourcy, C. F. A., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., and Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* 9:e105585. doi: 10.1371/journal.pone.0105585

Dubinsky, J. M. (2009). Heterogeneity of nervous system mitochondria: location, location, location! *Exp. Neurol.* 218, 293–307. doi: 10.1016/j.expneurol.2009.05.020

Durlanik, S., and Thiel, A. (2015). Requirement of immune system heterogeneity for protective immunity. *Vaccine* 33, 5308–5312. doi: 10.1016/j.vaccine.2015.05.096

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* 297, 1183–1186. doi: 10.1126/science.1070919

Elsasser, W. M. (1984). Outline of a theory of cellular heterogeneity. *Proc. Natl. Acad. Sci. U.S.A.* 81, 5126–5129. doi: 10.1073/pnas.81.16.5126

Emery, B., and Barres, B. A. (2008). Unlocking CNS cell type heterogeneity. *Cell* 135, 596–598. doi: 10.1016/j.cell.2008.10.031

Ferrell, J. E. Jr., and Machleder, E. M. (1998). The biochemical basis of an all-or-none cell fate switch in *Xenopus oocytes*. 280, 895–898. doi: 10.1126/science.280.5365.895

Gay, L., Baker, A.-M., and Graham, T. A. (2016). Tumour Cell Heterogeneity. *F1000Res.* 5:F1000 Faculty Rev-238. doi: 10.12688/f1000research.7210.1

Gerlinger, M., McGranahan, N., Dewhurst, S. M., Burrell, R. A., Tomlinson, I., and Swanton, C. (2014). Cancer: evolution within a lifetime. *Annu. Rev. Genet.* 48, 215–236. doi: 10.1146/annurev-genet-120213-092314

Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Gokhale, N. S., McIntyre, A. B. R., McFadden, M. J., Roder, A. E., Kennedy, E. M., Gandara, J. A., et al. (2016). N6-methyladenosine in Flaviviridae viral RNA genomes regulates infection. *Cell Host Microbe* 20, 654–665. doi: 10.1016/j.chom.2016.09.015

Gordon, S., and Taylor, P. R. (2005). Monocyte and macrophage heterogeneity. *Nat. Rev. Immunol.* 5:953. doi: 10.1038/nri1733

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14:e8046. doi: 10.15252/msb.20178046

Grimbergen, A. J., Siebring, J., Solopova, A., and Kuipers, O. P. (2015). Microbial bet-hedging: the power of being different. *Curr. Opin. Microbiol.* 25, 67–72. doi: 10.1016/j.mib.2015.04.008

Hadjantonakis, A., and Arias, A. (2016). Single-cell approaches: pandora's box of developmental mechanisms. *Dev. Cell* 38, 574–578. doi: 10.1016/j.devcel.2016.09.012

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9:75.

Hassan, C., Afshinnekoo, E., Wu, S., and Mason, C. E. (2017). Genetic and epigenetic heterogeneity and the impact on cancer relapse. *Exp. Hematol.* 54, 26–30. doi: 10.1016/j.exphem.2017.07.002

Hedlund, E., and Deng, Q. (2018). Single-cell RNA sequencing: technical advancements and biological applications. *Mol. Aspects Med.* 59, 36–46. doi: 10.1016/j.mam.2017.07.003

Henssen, A. G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., et al. (2017). Human PGBD5 DNA transposase promotes site-specific oncogenic mutations in rhabdoid tumors. *Nat. Genet.* 49, 1005–1014. doi: 10.1038/ng.3866

Hosokawa, M., Nishikawa, Y., Kogawa, M., and Takeyama, H. (2017). Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci. Rep.* 7:5199. doi: 10.1038/s41598-017-05436-4

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development* 136, 3853–3862. doi: 10.1242/dev.035139

Jiang, Y., Qiu, Y., Minn, A. J., and Zhang, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5528–E5537. doi: 10.1073/pnas.1522203113

Kanter, I., and Kalisky, T. (2015). Single cell transcriptomics: methods and applications. *Front. Oncol.* 5:53. doi: 10.3389/fonc.2015.00053

Kester, L., and van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell.* 23, 166–179. doi: 10.1016/j.stem.2018.04.014

Kleppe, M., Spitzer, M. H., Li, S., Hill, C. E., Dong, L., Papalexi, E., et al. (2017). Jak1 Integrates Cytokine Sensing to Regulate Hematopoietic Stem Cell Function and

Stress Hematopoiesis. *Cell Stem Cell* 21, 48.e7–501.e7. doi: 10.1016/j.stem.2017.08.011

Kreso, A., O'Brien, C. A., van Galen, P., Gan, O. I., Notta, F., Brown, A. M. K., et al. (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* 339, 543–548. doi: 10.1126/science.1227670

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., Daleykeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56–61. doi: 10.1038/nature13920

Kuzmin, A. N., Levchenko, S. M., Pliss, A., Qu, J., and Prasad, P. N. (2017). Molecular profiling of single organelles for quantitative analysis of cellular heterogeneity. *Sci. Rep.* 7:6512. doi: 10.1038/s41598-017-06936-z

Li, S., Garrett-Bakelman, F., Perl, A. E., Luger, S. M., Zhang, C., To, B. L., et al. (2014). Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol.* 15:472. doi: 10.1186/s13059-014-0472-5

Li, S., Garrett-Bakelman, F. E., Chung, S. S., Sanders, M. A., Hricik, T., Rapaport, F., et al. (2016). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* 22, 792–799. doi: 10.1038/nm.4125

Liang, S.-B., and Fu, L.-W. (2017). Application of single-cell technology in cancer research. *Biotechnol. Adv.* 35, 443–449. doi: 10.1016/j.biotechadv.2017.04.001

Litzenburger, U. M., Buenrostro, J. D., Wu, B., Shen, Y., Sheffield, N. C., Kathiria, A., et al. (2017). Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol.* 18:15. doi: 10.1186/s13059-016-1133-7

Ma, X., Edmonson, M., Yergeau, D., Muzny, D. M., Hampton, O. A., Rusch, M., et al. (2015). Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.* 6:6604. doi: 10.1038/ncomms7604

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-Cell Multiomics: multiple measurements from single Cells. *Trends Genet.* 33, 155–168. doi: 10.1016/j.tig.2016.12.003

Mason, C. E., Afshinnekoo, E., Tighe, S., Wu, S., and Levy, S. (2017). International standards for genomes, transcriptomes, and metagenomes. *J. Biomol. Tech.* 28, 8–18. doi: 10.7171/jbt.17-2801-006

Mason, C. E., Porter, S., and Smith, T. (2014). Characterizing Multi-omic data in Systems Biology. *Adv. Exp. Med. Biol.* 799, 15–38. doi: 10.1007/978-1-4614-8778-4_2

McIntyre, A. B. R., Rizzardi, L., Yu, A. M., Alexander, N., Rosen, G. L., Botkin, D. J., et al. (2016). Nanopore sequencing in microgravity. *Nat. Partner. J. Microgravity* 2:16035. doi: 10.1038/npjmgrav.2016.35

Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., et al. (2017). Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* 20, 1215–1228. doi: 10.1016/j.celrep.2017.07.009

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I. G., Leong-Quong, R. Y. Y., Chang, H., et al. (2016). Cell fate decision as high-dimensional critical state transition. *PLoS Biol.* 14:e2000640. doi: 10.1371/journal.pbio.2000640

Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C., and Yuan, J. (2017). Single-cell RNA sequencing: unraveling the brain one cell at a time. *Trends Mol. Med.* 23, 563–576. doi: 10.1016/j.molmed.2017.04.006

Pan, X., Durrett, R. E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., et al. (2013). Two methods for full-length RNA- seq for low quantities of cells and single cells. *Proc. Natl. Acad. Sci. U.S.A.* 110, 594–599. doi: 10.1073/pnas.1217322109

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257

Potts, S. J., Krueger, J. S., Landis, N. D., Eberhard, D. A., Young, G. D., Schmechel, S. C., et al. (2012). Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Lab. Invest. A J. Tech. Methods Pathol.* 92, 1342–1357. doi: 10.1038/labinvest.2012.91

Raj, A., and van Oudenaarden, A. (2008). Stochastic gene expression and its consequences. *Cell* 135, 216–226. doi: 10.1016/j.cell.2008.09.050

Reuben, A., Gittelman, R., Gao, J., Zhang, J., Yusko, E. C., Wu, C.-J., et al. (2017a). TCR repertoire intratumor heterogeneity in localized lung adenocarcinomas: an association with predicted neoantigen heterogeneity and postsurgical recurrence. *Cancer Discov.* 7, 1088–1097. doi: 10.1158/2159-8290.CD-17-0256

Reuben, A., Spencer, C. N., Prieto, P. A., Gopalakrishnan, V., Reddy, S. M., Miller, J. P., et al. (2017b). Genomic and immune heterogeneity are associated with differential responses to therapy in melanoma. *NPJ Genom. Med.* 2:10. doi: 10.1038/s41525-017-0013-8

Saletore, Y., Meyer, K., Korlach, J., Vilfan, I., Jaffrey, S., and Mason, C. E. (2012). The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* 13:175. doi: 10.1186/gb-2012-13-10-175

Shabaan, H., Westfall, D. A., Mohammad, R., Danko, D., Bezdan, D., Afshinnekoo, E., et al. (2018). The microbe directory: an annotated, searchable inventory of microbes' characteristics. *Gates Open Res.* 2:3. doi: 10.12688/gatesopenres.12772.1

Shalek, A. K., and Benson, M. (2017). Single-cell analyses to tailor treatments. *Sci. Transl. Med.* 9:eaan4730. doi: 10.1126/scitranslmed.aan4730

Shih, A., Meydan, C., Shank, K., Garrett-Bakelman, F., Ward, F., and Levine, R. (2017). Combination targeted therapy to disrupt aberrant oncogenic signaling and reverse epigenetic dysfunction in IDH2- and TET2-mutant acute myeloid leukemia. *Cancer Discov.* 7, 494–505. doi: 10.1158/2159-8290.CD-16-1049

Singh, A., Soltani, M. (2013). Quantifying intrinsic and extrinsic variability in stochastic gene expression models. *PLoS One* 8:e84301. doi: 10.1371/journal.pone.0084301

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604. doi: 10.1038/nprot.2017.149

Tabassum, D. P., and Polyak, K. (2015). Tumorigenesis: it takes a village. *Nat. Rev. Can.* 15, 473–483. doi: 10.1038/nrc3971

Tellez-Gabriel, M., Ory, B., Lamoureux, F., Heymann, M.-F., and Heymann, D. (2016). Tumour heterogeneity: the key advantages of single-cell analysis. *Int. J. Mol. Sci.* 17 2142 . doi: 10.3390/ijms17122142

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J., and Mitra, R. D. (2009). Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res.* 37, 4603–4612. doi: 10.1093/nar/gkp457

Viñuelas, J., Kaneko, G., Coulon, A., Vallin, E., Morin, V., Mejia-Pous, C., et al. (2013). Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC Biol.* 11:15. doi: 10.1186/1741-7007-11-15

Vu, L., Pickering, B. F., Cheng, Y., Zaccara, S., Nguyen, D., Minuesa, G., et al. (2017). The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal and leukemia cells. *Nat. Med.* 23, 1369–1376. doi: 10.1038/nm.4416

Wang, Y., and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol. Cell* 58, 598–609. doi: 10.1016/j.molcel.2015.05.005

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci. U.S.A.* 89, 5847–5851. doi: 10.1073/pnas.89.13.5847

Zheng, C., Zheng, L., Yoo, J. K., Guo, H., Zhang, Y., Guo, X., et al. (2017). Landscape of Infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 169, 1342.e16–56.e16. doi: 10.1016/j.cell.2017.05.035

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

MINI REVIEW
published: 21 September 2016
doi: 10.3389/fgene.2016.00163

Single-Cell Transcriptomics Bioinformatics and Computational Challenges

Olivier B. Poirion[1†], *Xun Zhu*[1,2†], *Travers Ching*[1,2] *and Lana Garmire*[1*]

[1] *Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA,* [2] *Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI, USA*

The emerging single-cell RNA-Seq (scRNA-Seq) technology holds the promise to revolutionize our understanding of diseases and associated biological processes at an unprecedented resolution. It opens the door to reveal intercellular heterogeneity and has been employed to a variety of applications, ranging from characterizing cancer cells subpopulations to elucidating tumor resistance mechanisms. Parallel to improving experimental protocols to deal with technological issues, deriving new analytical methods to interpret the complexity in scRNA-Seq data is just as challenging. Here, we review current state-of-the-art bioinformatics tools and methods for scRNA-Seq analysis, as well as addressing some critical analytical challenges that the field faces.

Keywords: single-cell genomics, single-cell analysis, bioinformatics, heterogeneity, microevolution

OPEN ACCESS

Edited by:
H. Steven Wiley, Pacific Northwest National Laboratory, USA

Reviewed by:
Seth G. N. Grant, University of Edinburgh, UK
Milind Ratnaparkhe, Indian Institute of Soybean Research (ICAR), India

***Correspondence:**
Lana Garmire
lgarmire@cc.hawaii.edu

[†] These authors have contributed equally to this work.

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *01 May 2016*
Accepted: *02 September 2016*
Published: *21 September 2016*

Citation:
Poirion OB, Zhu X, Ching T and Garmire L (2016) Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front. Genet. 7:163. doi: 10.3389/fgene.2016.00163

INTRODUCTION

Characterization of genomic signatures in individual patients is a key step toward the realization of precision medicine. Recently, next-generation sequencing (NGS) based RNA expression profiling (RNA-seq) has made broad impacts on biomedical fields. However, population-averaged RNA-seq has limited discovery power, and it can also mask the presence of rare subpopulations of cells (such as cancer stem cells) and thus may overlook important biological insights. The emerging single-cell RNA-Seq (scRNA-Seq) technology is designed to overcome these limitations by investigating expression profiles at the cell level. In just a few years, the number scRNA-Seq experiments has grown beyond exponentially. This new approach offers the potential to revolutionize our understanding of diseases and associated biological processes, with the capacity to reveal the intercellular heterogeneity within a specific tissue at an unprecedented resolution (Yan et al., 2013; Trapnell et al., 2014). Using single-cell level features, we can infer cell lineages (Treutlein et al., 2014), identify subpopulations (Trapnell et al., 2014) and highlight cell-specific biological characteristics (Tang et al., 2010). Moreover, single-cell analyses have already demonstrated their utilities in the clinical applications, ranging from characterizing cancer cells subpopulations (Navin et al., 2011; Patel et al., 2014; Ting et al., 2014), highlighting specific resistance mechanisms (Kim, K. T. et al., 2015; Miyamoto et al., 2015) to being used as diagnostic tools (Ramsköld et al., 2012; Kvastad et al., 2015).

Despite the expansion of scRNA-Seq studies and rapid maturing of experimental methods, major analytical challenges remain as the consequences of experimentation. One major challenge is that scRNA-Seq datasets present a very high level of noise (Brennecke et al., 2013; Kharchenko et al., 2014). Much of the noise is due to the nature of single-cell technologies. Because of the extremely low amount of starting biological material in the single cell, amplification processes are

required. These procedures are prone to distortion and contamination (Leng et al., 2015). To tackle these issues, rigorous efforts have been made to develop analytical methods for scRNA-Seq data. Here, we summarize current state-of-the-art bioinformatics analysis tools and methods for scRNA-Seq (**Figure 1** and **Table 1**), and address some critical analytical challenges that we are facing. The first section describes specific pre-processing steps for noise removal of scRNA-Seq datasets. The second section reviews specific scRNA-Seq bioinformatics analysis procedures with emphasis on subpopulation detection. The third section focuses on microevolution analysis for scRNA-Seq data. In the last section, we highlight the challenges to be addressed and work to be accomplished in scRNA-Seq bioinformatics field.

DATA PREPROCESSING AND NOISE REMOVAL

Quality Control

scRNA-Seq experiments generate FASTQ files from the sequencing machine, which contain millions of reads composed of RNA sequences and add-on sequences (UMI tag and the cell tag etc). These reads need to be pre-processed before being aligned back to the reference genome. For scRNA-seq, pre-processing and quality control (QC) analyses similar to bulk RNA-seq are used. Cutadapt (Martin, 2011) is a tool that removes adapter sequences, and Trimmomatic (Bolger et al., 2014) performs quality-based trimming in addition to removing adapter sequence. These tools are commonly used in scRNA-seq experiments (Treutlein et al., 2014; Handel et al., 2016; Hou et al., 2016). Other generic quality control tools such as FASTQC or HTQC (Yang et al., 2013) might also be useful to produce quality metrics. Finally, it is worth noting that platform-specific QC tools such as SolexaQA (Cox et al., 2010) provide QC pipelines specific for Illumina sequencing, with trimming and quality-based filtering.

Other QC procedures for scRNA-seq involve the analysis of the expression of housekeeping genes (Ting et al., 2014; Treutlein et al., 2014), overall gene expression patterns (Zeisel et al., 2015) and the number of genes or reads detected per cell (Kumar et al., 2014). However, one issue of these approaches is that the thresholds chosen for filtering are arbitrary and should differ according to the dataset (Jiang, P. et al., 2016). SinQC (Jiang, P. et al., 2016) and SCell (Diaz et al., 2016) are two QC tools specifically designed for scRNA-seq data. SinQC uses sequencing library quality to confirm gene expression outliers. It computes different quality metrics (e.g., total number of mapped reads, mapping rate and library complexity) to identify a user-specified fraction of the dataset as noise. SCell is a versatile tool that allows for outlier detection. It estimates genes that are expressed at the background level using Gini index, which measures statistical dispersion, and removes samples whose background fraction is significantly higher than the average. Recently, a new mapping and quality assessment pipeline Celloline detects low quality cells from expression profiles, using curated biological and technical features (Ilicic et al., 2016).

Alignment

To our knowledge, there are currently no specific aligners dedicated to scRNA-seq, and scRNA-seq studies use existing aligners made for bulk RNA-Seq. Tophat is one of the most popular aligners capable of detecting novel splice (Trapnell et al., 2009; Kim et al., 2013), and it is widely used in scRNA-seq studies (Treutlein et al., 2014; Fan et al., 2016; Freeman et al., 2016; Handel et al., 2016; Hou et al., 2016). RNA-Seq by Expectation Maximization, or RSEM, is a popular framework that includes an aligner (Li and Dewey, 2011). It is also used in some scRNA-seq studies (Gao et al., 2016; Kimmerling et al., 2016; Meyer et al., 2016). Other aligners used in scRNA-Seq studies include MapSplice (Wang et al., 2010), GSNAP (Brennecke et al., 2013; Buettner et al., 2015; Wu et al., 2016), and STAR (Dobin and Gingeras, 2015; Moignard et al., 2015; Petropoulos et al., 2016). Among these aligners, TopHat and STAR were found to be about one to two magnitudes faster than GSNAP and MapSplice (Engström et al., 2013). More recently developed aligners include Kallisto (Bray et al., 2016) and HISAT (Kim, D. et al., 2015). Kallisto uses pseudo-alignment with hashing de Bruijn graphs and avoids alignment altogether, which drastically improves the speed of expression quantification. HISAT (hierarchical indexing for spliced alignment of transcripts) seems also promising in term of the speed and accuracy. It is worth mentioning that some major scRNA-Seq methods do not get enough coverage across the gene to measure alternative splicing, therefore algorithms for isoform measurements are not as critical in scRNA-Seq, at least at this stage.

Feature Quantification

Feature quantification is the process of converting alignment results into a gene expression profile. An expression profile is conventionally represented as a numeric matrix where rows are genes and columns are cells. Each entry in the matrix is the abundance of a particular gene or transcript in a particular sample. Just as is the case for aligners, most scRNA-Seq studies use canonical feature quantification methods applied to bulk RNA-Seq.

Quantification methods for gene expression differ dramatically. The simplest approach, employed by programs such as HTSeq (Anders et al., 2014) and FeatureCounts (Liao et al., 2013), is to count the number of reads located within the boundaries of a gene (Liao et al., 2013; Anders et al., 2014). These programs have simple but flexible parameters for determining read counts in the case of overlapping genes, and were used in some scRNA-Seq studies (Brennecke et al., 2013; Moignard et al., 2015; Fan et al., 2016; Handel et al., 2016). More sophisticated approaches calculate probabilistic estimates of gene expression. For example, RSEM and Cufflinks both employ a maximum likelihood approach (Trapnell et al., 2010; Li and Dewey, 2011). These programs are based on statistical models where reads in a RNA-Seq sample are observed random variables predicted from the latent variables, such as the transcript sequence, strand and length. The new Kallisto pipeline (Bray et al., 2016) as described before, is shown to have up to two orders of magnitude speed improvement over previous aligner-quantifier combinations (Ntranos et al., 2016). Interestingly, while

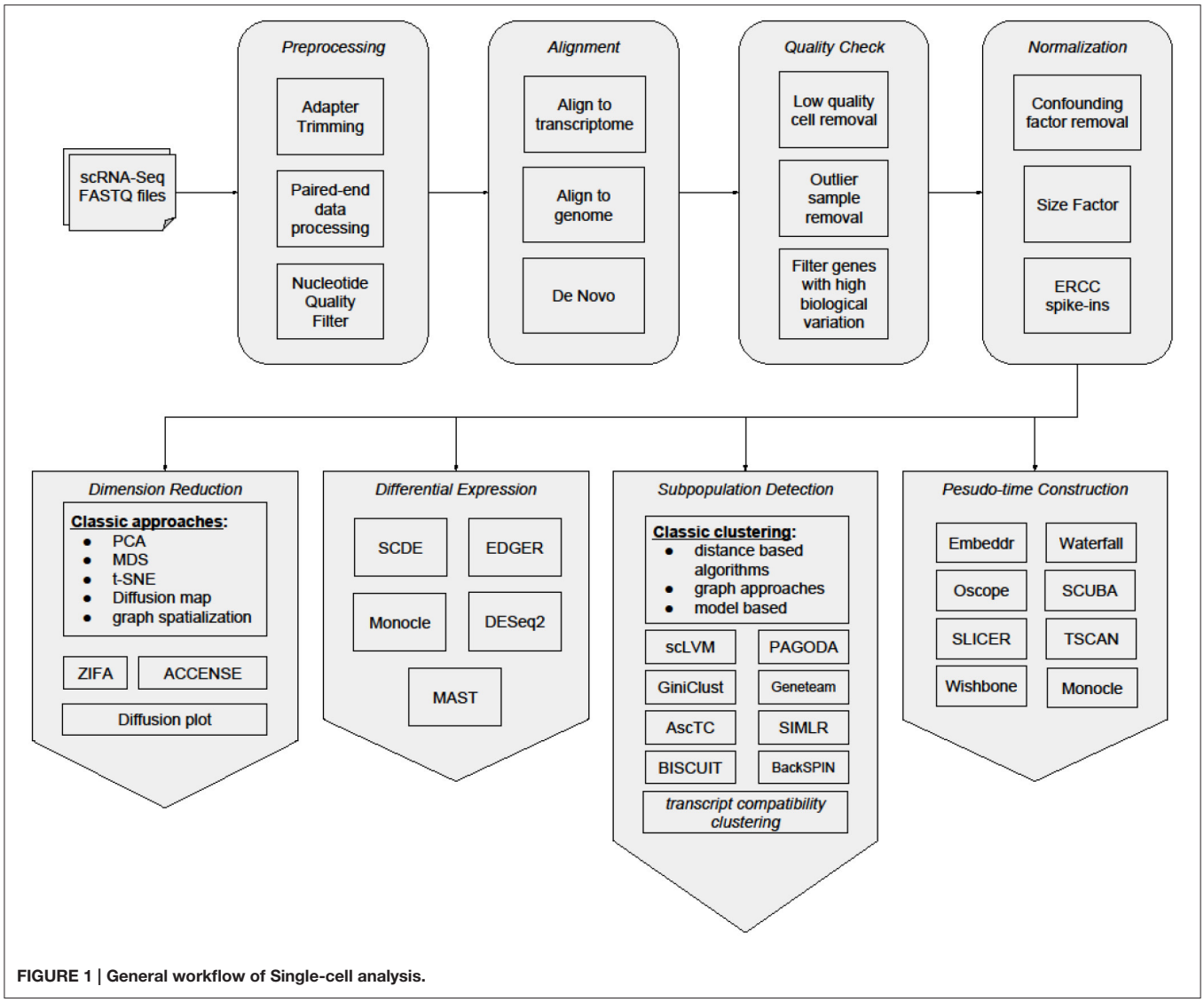


FIGURE 1 | General workflow of Single-cell analysis.

probabilistic approaches are conceptually more refined, simple counting programs such as HTSeq and FeatureCounts showed comparable or even stronger performance (Chandramohan et al., 2013; Fonseca et al., 2014), suggesting that these probabilistic models are yet to be improved.

Given the uncertainties of quantifying fragments post-amplification, a new technique was shown to reduce amplification noise by introducing random sequences called unique molecular identifiers, or UMIs (Islam et al., 2014). UMIs are tagged on individual RNA molecules before amplification and used for tracking transcripts directly rather than using sophisticated statistical modeling. This approach may lead to a different workflow than conventional fragment-based quantification methods (e.g., gene filtering and normalization).

Gene Filtering

Due to the high level of noise in scRNA-Seq datasets, it is necessary to filter out low quality genes and samples. Various practices have been made to filter out genes that are expressed in too few samples (Brennecke et al., 2013; Treutlein et al., 2014; Petropoulos et al., 2016). Usually, a gene is defined as "expressed" by a minimal expression level threshold. For experiments that quantify gene expression with fragment counting, an FPKM (Fragment per Kilobase per Million Reads) threshold is appropriate. Common FPKM thresholds are 1 (Freeman et al., 2016) and 10 (Petropoulos et al., 2016). Other studies also set the threshold by Transcript Per Million (TPM) instead of FPKM (Meyer et al., 2016). Yet better filtering reference could come from External RNA Controls Consortium (ERCC) spike-ins added to the experiment, which provides calibration of the relative amount of starting material (Brennecke et al., 2013; Treutlein et al., 2014).

Recently, specific methods have been developed to filter genes from scRNA-seq dataset. OEFinder is designed to identify artifact genes from scRNA-seq experiments using the Fluidigm C1 platform for cell capture (Leng et al., 2016). For experiments that

TABLE 1 | List of single-cell analytical tools mentioned in this chapter.

Category	Tool name	References	Availability
Preprocessing	cutadapt	Martin, 2011	https://cutadapt.readthedocs.org/en/stable/index.html
Preprocessing	Trimmomatic	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic
Preprocessing	FASTQC	Andrews, 2010	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Preprocessing	SolexaQA	Cox et al., 2010	http://solexaqa.sourceforge.net/
Preprocessing	BIGpre	Zhang et al., 2011	https://sourceforge.net/projects/bigpre/
Preprocessing	HTQC	Yang et al., 2013	https://sourceforge.net/projects/htqc/
Preprocessing	SinQC	Jiang, P. et al., 2016	http://www.morgridge.net/SinQC.html
Preprocessing	SCell	Diaz et al., 2016	https://github.com/diazlab/scell
Preprocessing	celloline	Ilicic et al., 2016	https://github.com/Teichlab/celloline
Alignment	Tophat	Trapnell et al., 2009; Kim et al., 2013	https://ccb.jhu.edu/software/tophat/index.shtml
Alignment	RSEM	Li and Dewey, 2011	http://deweylab.github.io/RSEM/
Alignment	GSNAP	Wu et al., 2016	http://research-pub.gene.com/gmap/
Alignment	STAR	Dobin and Gingeras, 2015	https://github.com/alexdobin/STAR
Alignment	Mapsplice	Wang et al., 2010	http://www.netlab.uky.edu/p/bioinfo/MapSplice2
Quantification	Cufflinks	Trapnell et al., 2010	http://cole-trapnell-lab.github.io/cufflinks/
Quantification	HISAT	Kim, D. et al., 2015	https://ccb.jhu.edu/software/hisat2/index.shtml
Quantification	HTSeq	Anders et al., 2014	http://www-huber.embl.de/HTSeq/doc/overview.html
Quantification	FeatureCounts	Liao et al., 2013	http://bioinf.wehi.edu.au/featureCounts/
Quantification	Kallisto	Bray et al., 2016	https://pachterlab.github.io/kallisto/about.html
Gene filtering	OEFinder	Leng et al., 2016	https://github.com/lengning/OEFinder
Cofounding factor removal	scLVM	Buettner et al., 2015	https://github.com/PMBio/scLVM
Cofounding factor removal	COMBAT	Johnson et al., 2007	https://github.com/brentp/combat.py
Normalization	GRM	Ding et al., 2015	http://wanglab.ucsd.edu/star/GRM/
Normalization	BASICS	Vallejos et al., 2015	http://journals.plos.org/ploscompbiol/article/asset?unique&id=info:doi/10.1371/journal.pcbi.1004333.s009
Normalization	SAMstrt	Katayama et al., 2013	https://github.com/shka/R-SAMstrt
Normalization	Deconvolution	Aaron et al., 2016	https://github.com/MarioniLab/Deconvolution2016
Dimension Reduction	pcaReduce	Zurauskiene and Yau, 2015	https://github.com/JustinaZ/pcaReduce
Dimension Reduction	*t*-SNE	der Maaten and Hinton, 2008	https://lvdmaaten.github.io/tsne/
Dimension Reduction	ACCENSE	Shekhar et al., 2014	http://www.cellaccense.com/
Dimension Reduction	ZIFA	Pierson and Yau, 2015	https://github.com/epierson9/ZIFA
Differential Expression	SCDE	Kharchenko et al., 2014	http://hms-dbmi.github.io/scde/
Differential Expression	PAGODA	Fan et al., 2016	http://hms-dbmi.github.io/scde/
Differential Expression	EdgeR	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Differential Expression	DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Differential Expression	MAST	Finak et al., 2015	https://github.com/RGLab/MAST
Subpopulation Detection	GiniClust	Jiang, L. et al., 2016	https://github.com/lanjiangboston/GiniClust
Subpopulation Detection	Geneteam	Harris et al., 2015	
Subpopulation Detection	AscTC	Ntranos et al., 2016	https://github.com/govinda-kamath/clustering_on_transcript_compatibility_counts
Subpopulation Detection	SIMLR	Wang et al., 2016	https://github.com/BatzoglouLabSU/SIMLR
Subpopulation Detection	BISCUIT	Prabhakaran et al., 2016	http://www.c2b2.columbia.edu/danapeerlab/html/pub/prabhakaran16-supp.pdf
Subpopulation Detection	BackSPIN	Zeisel et al., 2015	https://github.com/linnarsson-lab/BackSPIN
Microevolution	Moncole	Trapnell et al., 2014	http://cole-trapnell-lab.github.io/monocle-release/
Microevolution	embeddr	Campbell et al., 2015	https://github.com/kieranrcampbell/embeddr
Microevolution	SCUBA	Marco et al., 2014	https://github.com/gcyuan/SCUBA
Microevolution	Oscope	Leng et al., 2015	https://www.biostat.wisc.edu/~kendzior/OSCOPE/
Microevolution	SLICER	Welch et al., 2016	https://github.com/jw156605/SLICER
Microevolution	TSCAN	Ji and Ji, 2016	http://bioconductor.org/packages/release/bioc/html/TSCAN.html
Workflow	SINCERA	Guo et al., 2015	https://research.cchmc.org/pbge/sincera.html

Links for their availability are attached.

quantify gene expression with UMI counting, one can directly set up a molecule number threshold, e.g., 25 (Zeisel et al., 2015). It is also recommended to remove UMIs that have reads <1/100 of average non-zero UMI reads, in order to avoid erroneous UMIs generated during amplification.

Removal of Confounding Factors

When the entire data set consists of several runs of experiments with potentially varied conditions, systematic variations called batch effects might be introduced. These artifacts may pose substantial problems to downstream statistical analysis, or even mask biological signals. For studies concerning over-dispersion of gene expression, it is necessary to factor out the extra variance caused by the systematic differences between batches (Fan et al., 2016). The appropriate way to compensate for batch effect depends on the quantification method as well as the downstream analysis. For most studies batch effects can be eliminated by using down-sampling methods, however the complexity is reduced (Wang et al., 2012; Dey et al., 2015; Grün and van Oudenaarden, 2015). For studies that use traditional fragment counting, COMBAT (Johnson et al., 2007) is a batch effect eliminating method based on empirical Bayes frameworks and purports to be robust to outliers for small sample sizes. It was originally designed for microarray data but was used in scRNA-Seq experiments (Kim, K. T. et al., 2015). Although unsupervised batch effect detection or removal methods exist (Leek, 2014), the batches called by such methods often correlate highly with subpopulations detected by other scRNA-Seq methods (Finak et al., 2015). Since it is usually desirable to consider subpopulations for valuable biological insights, unsupervised batch effect removal methods should be used with discretion in single-cell experiments.

Besides batch-effect removal, it is also important to remove technical variability within the noise. The technical noise level of a genes correlates with its average expression level. Thus, a probabilistic model can be built to fit this correlation using technical spike-ins and further infer the biological variability of each gene (Brennecke et al., 2013). For most studies, it is also desirable to avoid the ubiquitous cell-cycle induced variation to mask other interesting biological variations. scLVM is a package that tries to introduce a cell-cycle factor removal step before subpopulations detection (Buettner et al., 2015). Recently, a new package called ccRemover was developed to remove the principal components that are identified as cell-cycle affected, which claimed to perform better than scLVM in several simulated and real datasets (Barron and Li, 2016).

Normalization

In scRNA-seq experiments, technical factors such as read depth, cell capture efficiency, 3′ bias or full sequence coverage due to particular library prep methods, might differ among different scRNA-Seq data sets. Thus, raw read counts should be normalized before downstream analyses. This procedure maximally ensures that the difference between the values in the matrix correctly reflects the abundance difference of transcripts or genes between the cells. When experiments are designed with ERCC spike-ins, ERCC can be used as internal controls and serve as anchors for normalization. GRM is a scRNA-seq normalization tool fitting a Gamma Regression Model between the reads (FPKM, RPKM, TPM) and spike-ins (Ding et al., 2015). The trained model is then used to estimate gene expression from the reads. BASICS, another recent workflow, provides a Bayesian model allowing to infer cell-specific normalization factor (Vallejos et al., 2015). This workflow estimates the technical variability using spike-ins. Finally, SAMstrt (Katayama et al., 2013) is an earlier algorithm that applies the resampling normalization procedure of the SAMseq algorithm to spike-ins, which was originally developed for bulk RNA-seq (Li and Tibshirani, 2013).

For experiments without spike-ins, if the quantification is count-based, one can normalize the expression profile by the scaling methods used in DESeq and edgeR etc. (Love et al., 2014). A new specific scRNA-seq procedure proposes a de-convolution approach on the pooled counts of gene expression for multiple cells, thus allows to infer the size factor for individual cells without using spike-ins (Aaron et al., 2016). The authors claimed that their approach improved the accuracy of the normalization compared with existing methods. However, experiments designed with UMIs as mentioned earlier quantify gene expression on an absolute basis and thus they do not need computational normalization.

Differential Expression

Differential expression (DE) analysis is the process of calling gene expression that show statistically significant difference between pre-specified groups of samples. Although DE is typically not the main objective of a single-cell experiment design, as it requires pre-defined grouping information among cells of interest, it is nevertheless common in scRNA-Seq experiments. Simple statistical methods such as *t*-test and Wilcoxon rank sum test are used in scRNA-Seq workflows such as SINCERA (Guo et al., 2015). Interestingly, EdgeR and DESeq2, two DE methods developed for bulk RNA-Seq, gave the best results for some scRNA-Seq data (Schurch et al., 2016).

The dropout event is a unique type of noise of scRNA-Seq that rarely occurs in bulk RNA-Seq experiments. It refers to the phenomenon that a gene is shown expressed abundantly in one cell but not detectable in another cell, as a consequence of the transcript loss in the reverse-transcription step. To account for frequent dropout events and biological variability within cell population, more sophisticated algorithms have been developed for scRNA-Seq data. Single-Cell Differential Expression (SCDE) is a package developed specifically for single-cell differential expression (Kharchenko et al., 2014). The model assumes that observed expression levels in scRNA-Seq data follow a mixture of negative binomial distribution for amplified genes, as proposed before (Anders and Huber, 2010); and a low-mean poisson distribution for dropout genes, as is observed in transcriptionally silenced genes. This model is then fit using Expectation Maximization (EM) algorithm (Kharchenko et al., 2014). It claimed higher sensitivity of differentially expressed genes compared to DESeq and CuffDiff. More recently, PAGODA improved upon SCDE's method in several aspects, including optimization of the computational process and a refined model

for better fitting (Fan et al., 2016). MAST is another scRNA-Seq differential expression detection method that uses a two-part generalized linear model and adjusts for the fraction of cells that express a certain gene (Finak et al., 2015).

Another challenge unique to scRNA-Seq is that some genes may exhibit bimodality, meaning that the expression levels across a group of cells concentrate around two modes instead of one. A beta-Poisson distribution was proposed in order to provide a more accurate differential expression analysis that captures bimodality (Vu et al., 2016). Another tool Monocle (Trapnell et al., 2014) also has a module for differential expression, which fits the data with a non-parametric generalized additive model. Finally, the workflow of BASICS as described earlier, provides an criterion to detect high- or low-variable genes within the single cells dataset (Vallejos et al., 2015). However, it is not clear which methods have generally superior performance.

SUBPOPULATION AND MODULE DETECTION

General Machine-Learning Approaches

Different classical unsupervised approaches have been used to highlight single cell subgroups among a population. Principal Component Analysis (PCA) and its variants (e.g., Robust PCA and Kernel PCA) have been used in different single cell studies (Amir et al., 2013; Yan et al., 2013; Pollen et al., 2014; Trapnell et al., 2014; Treutlein et al., 2014; Satija et al., 2015; Fan et al., 2016; Ilicic et al., 2016). *K*-means and other distance based clustering algorithms such as hierarchical clustering or WARD are also widely used (Yan et al., 2013; Jaitin et al., 2014; Kharchenko et al., 2014; Lohr et al., 2014; Marco et al., 2014; Pollen et al., 2014; Shin et al., 2015). For example, Jaitin et al. combined hierarchical clustering and probabilistic mixture models to classify single cells from different tissues (Jaitin et al., 2014). A refined clustering method called pcaReduce (Zurauskiene and Yau, 2015) was designed for scRNA-Seq. It iteratively uses PCA combined with *K*-means to produce the hierarchical tree of the cells. For distance metrics employed by these methods, Euclidean distance, Pearson and Spearman correlation coefficients have been popular (though may not be optimal) choices (Pollen et al., 2014; Rotem et al., 2015).

Machine-Learning Approaches Tailored for scRNA-Seq Analysis

More sophisticated machine-learning algorithms have great potentials to overcome some issues of scRNA-Seq functional analysis. A main issue of scRNA-Seq analysis is that gene expression data cannot be expressed as a linear combination of the relationships between two cells in general (Buettner and Theis, 2012; Bendall et al., 2014; Levine et al., 2015). Also classical similarities (such as cosine or Euclidean distances) are less meaningful as the dimensionality increases (Beyer et al., 1999), and may not be appropriate for scRNA-Seq (Xu and Su, 2015). Possible irrelevant associations may arise with inappropriate metrics, while searching for the nearest neighbors on noisy data (Balasubramanian and Schwartz, 2002). Adequate analytical methods for scRNA-Seq data should also be able to highlight "rare events," such as the small fraction of metastatic cancer cells amongst a large cell population (Bose et al., 2015; Shin et al., 2015). We describe the scRNA-Seq specific algorithms below in the order of dimension reduction, clustering, and other clustering variant methods. The datasets that were used to test these algorithms are listed in **Table 2**.

Among the dimension reduction methods, Zero-inflated factor analysis (ZIFA) algorithm is a new method that includes dropout events by representing the probability of gene dropout as an exponential function of its mean expression (Pierson and Yau, 2015). Using a latent variable model based on factor analysis, ZIFA reduces the dimension of scRNA-Seq dataset and allows the probability of each gene expression to be zero. Experiments in the original study suggest that ZIFA is a more robust alternative to PCA. As mentioned earlier, scLVM is another method for identifying cell subpopulations, which features removal of confounding factor like cell-cycle effects (Buettner et al., 2015). It first computes cell-to-cell covariance using a set of marker genes related to biological hidden factors of interest (such as the cell cycle). Another approach, PAGODA as mentioned before, uses a weighted PCA to characterize multiple aspects of heterogeneity in mouse neuronal progenitors (Fan et al., 2016). PAGODA evaluates over-dispersion of individual genes using error models.

SIMLR is a new clustering method designed to learn a distance metric that best fits the structure of the data. It infers a distance function as a linear combination of several distance metrics (Wang et al., 2016). It is designed to tackle the heterogeneity observed amongst single-cell datasets related to both technological difference across platforms as well as biological difference across studies. In another single-cell clustering approach named analysis of scRNA-seq based on transcript-compatibility counts (AscTC), read counts from scRNA-Seq dataset are transformed into probabilities using transcript-compatibility counts, rather than the conventional transcript abundance (Ntranos et al., 2016). Individual cells are clustered using an affinity propagation algorithm, a derivative of spectral clustering.

A few other hierarchical clustering approaches are worth mentioning. Geneteam is a multi-level recursive clustering method that searches for bipartitions of cells sharing exclusive expression profiles for a subset of genes (Harris et al., 2015). Similarly, Backspin is another hierarchical dividing clustering algorithm, allowing to cluster both genes and cells (Zeisel et al., 2015). It uses the SPIN algorithm (Tsafrir et al., 2005) at each iteration to sort the expression matrix and then separates genes (rows) and cells (columns) into two groups by a specific splitting criterion. Alternatively, BISCUIT is a new iterative normalization and clustering procedure based on Dirichlet Process, which was designed to correct technical variation in scRNA-seq together with cell clustering (Prabhakaran et al., 2016).

Graph Approaches beyond Clustering

Traditional clustering methods lack the function of inferring the inherent lineage between cells. Common approaches for cell lineage inferences require the creation of a graph or a tree, where single cells are represented as nodes and edges

TABLE 2 | Description of the main datasets for subpopulation and module detection analysis.

Dataset description	Accession	References	Species	Number of cells	Original analysis	Applied algorithms
Cortex and hippocampus cells	GSE60361	Zeisel et al., 2015	Mouse	3005	BackSPIN	Geneteam, PAGODA, AscTC, BISCUIT, GiniClust
11 different cell types	SRP041736	Pollen et al., 2014	Human	301	PCA and hierarchical clustering	ZIFA, SILMR, pcaReduce
Myoblast differentiation	GSE52529	Trapnell et al., 2014	Human	372	MONOCLE	ZIFA, AscTC, TSCAN, Embeddr
Embryomic T-cells under different cell cycle stages	E-MTAB-2512	Buettner et al., 2015	Mouse	182	scLVM	ZIFA, SLIMR
Preimplementation embryos and embryonic stem cells at different stages	GSE36552	Yan et al., 2013	Human	124	PCA and hierarchical clustering	scLVM, SNN-Cliq
Cells from developing bronchioalveolar at four different stages of development	GSE52583	Treutlein et al., 2014	Mouse	202	PCA and hierarchical clustering	SLICER, EMBEDDR

between the cells indicate their similarities. The lengths of the edges are computed from a similarity matrix based on a given metric. Before constructing the graph, a de-noising procedure is necessary. A useful de-noising procedure is to compute the *k*-Nearest-Neighbor graph (kNNG; Bendall et al., 2014; Levine et al., 2015; Xu and Su, 2015). Samples from the kNNG could then be compared using the geodesic distance, defined as the shortest path between two nodes (Bendall et al., 2014). Such an approach can remove "shortcuts" between irrelevant pairs of samples due to the curse of high dimensionality (Tenenbaum et al., 2000). Clustering analysis can then be performed on the graph using community detection algorithms (Fortunato, 2010). Xu and Su first used Euclidean distance to compute Shared Nearest-Neighbor (SNN) graph, then searched for quasi-cliques to obtain clusters of cells (Xu and Su, 2015). Quasi-cliques are communities of nodes, densely but not necessarily fully connected. Highly Connected Sub-graph (HPC) is another community detection algorithm that showed very similar performances as SNN (Hartuv and Shamir, 2000).

MICROEVOLUTION OF SINGLE CELLS

Inference without Spatial and Temporal Information

scRNA-Seq data are also informative to reveal single-cell microevolution. Different algorithms have been specifically designed for scRNA-Seq to infer a pseudo temporal ordering of single cells. Moncole is the first scRNA-Seq bioinformatics tool to infer the temporal ordering of single cells (Trapnell et al., 2014). It first uses Independent Component Analysis (ICA) to reduce the dimension, then computes a Minimum Spanning Tree (MST) on the graph constructed by Euclidean distance between cell pairs. MST connects all nodes of a graph using edges with a minimal total weighting, based on the hypothesis that the longest path through the MST corresponds to the longest series of transcriptionally similar cells. Another similar method, Waterfall, uses PCA coupled with *k*-means to produce clusters, then connects the cluster centroids with MST (Shin et al., 2015). Similar to Waterfall, TSCAN is a new approach based on MST. Cells are first clustered using a model-based approach before constructing an MST, allowing the reduction of the tree space complexity (Ji and Ji, 2016).

Embeddr is a method that uses the correlation metric between cells to construct kNNG, then projects the samples into a low-dimensional embedding using Laplacian eigen maps. The pseudo time order is then fitted using the principal curves (Campbell et al., 2015). Embeddr aims to tackle the drawbacks of Monocle, where gene expression is modeled as a linear combination and the result is highly sensitive to outliers. This scheme is also used in the workflow of SLICER, a recent algorithm using Locally Linear Embedding (LLE) to project the dataset and to construct a kNNG among cells (Welch et al., 2016).

Since visualization is key in understanding reconstructed single-cell trajectories, better visualization algorithms are as important as methods to reconstruct the single-cell microevolution. *t*-SNE is a popular method to visualize single cells, as part of a more complex workflow (Jiang, L. et al., 2016; Petropoulos et al., 2016). Another approach derived from diffusion map was developed, allowing one to visualize a clear bifurcation event among the cells which may be missed by independent component analysis (ICA) or *t*-SNE (Haghverdi et al., 2015; Moignard et al., 2015).

Modeling Microevolution with Spatial and Temporal Information

Cell subpopulations can also be characterized by different temporal and/or spatial gene expressions. Several approaches have been designed to exploit datasets with explicit temporal information. SCUBA is a method to detect bifurcation events using time course data (Marco et al., 2014). It assumes that the switch between cell states is a stochastic punctual process. To infer cellular hierarchy, it iteratively divides cells using *k*-means algorithm and uses a gap statistic to determine if a bifurcation event should occur. This process creates a binary tree, which can then be used to model gene expression dynamics (Marco et al., 2014). However, one drawback of SCUBA is that it requires

data with temporal features. Free from such a requirement, Oscope is another method to infer oscillatory genes among single cells collected from a single tissue (Leng et al., 2015). It hypothesizes that these cells represent distinct states according to an oscillatory process. Oscope fits a two-dimensional sinusoidal function for each pair of genes, clusters gene pairs by frequency and reconstructs the order of the cells in a cyclic fashion. However, Oscope is unable to infer bifurcation events.

Other models also consider the spatial organization of cells in a tissue. Seurat is an approach that infers the spatial localization of single cells by integrating RNA-Seq with *in situ* RNA patterns (Satija et al., 2015). Seurat divides a cellular tissue into distinct spatial bins, linked by the expression of landmark genes per RNA *in-situ* hybridization. Within each bin, it builds a mixture model using expression values among correlated genes. The posterior probability is generated for each cell and assigned to a given bin. Another approach models the tissue as a 3D map and assumes that cells spatially close share common scRNA-Seq profiles (Pettit et al., 2014). This method uses a hidden markov random field to assign each bin of the map to a given cluster. Similar to Seurat, it takes the input of spatial gene expression measurement using whole mount *in situ* Hybridizations (WiSH) technology, a confocal microscopic approach that detects the presence of mRNA linked to a fluorescent probe.

CHALLENGES AND FUTURE WORK

Compared to bulk-cell analysis, single-cell genomics has the advantage of exploring cellular processes with a more accurate resolution, but it is more vulnerable to disturbances. Besides perfecting the experimental protocols to deal with issues such as dropouts in gene expression and biases in amplification, deriving new analytical methods to reveal the complexity in scRNA-Seq data is just as challenging. In this review, we have listed the different bioinformatics algorithms dedicated to single-cell analysis. Although the initial few steps of workflow for scRNA-Seq analysis are similar to bulk-cell analysis (data pre-processing, batch removal, alignment, quality check, and normalization), the subsequent analyses are largely unique for single cells, such as subpopulations detection, and microevolution characterization (**Figure 1**). With the increasing popularity of single-cell assays and ever increasing number of computational methods developed, these methods need to be more accessible to research groups without bioinformatics expertise. Moreover, datasets where cell classes have already been previously charaterized should be identified as benchmark data, in order to accurately assess the performance of new bioinformatics methods.

Although this review focuses on scRNA-Seq analyses, with the rapid development of technologies, coupled DNA-based genomics data can be obtained from the same cell, in parallel with scRNA-Seq data (Han et al., 2014; Dey et al., 2015; Kim, K. T. et al., 2015; Macaulay et al., 2015). This will further increase the analytical challenges. Previous multi-omics bioinformatics tools applied to bulk samples could be leveraged. The use of graphs and tensor approaches that integrate heterogeneous features in bulk samples may be good starting points for multi-dimensional single cell data (Li et al., 2009; Levine et al., 2015; Katrib et al., 2016; Zhu et al., 2016). Efforts should also be made toward developing computational methods to make use of spatial information (possibly guided by imaging) in combination of scRNA-Seq (Pettit et al., 2014; Satija et al., 2015). Also most emphasis in scRNA-Seq by far has been made on protein coding genes, and the dynamics and roles of non-coding RNAs such as lncRNAs (Travers et al., 2015; Ching et al., 2016) and micro-RNAs are poorly explored. Finally, a large number of single-cells ($n = 4645$) in a single data set was reported recently (Tirosh et al., 2016), and the scRNA-Seq data volume is expected to continue growing exponentially. Foreseeably, this poses a large spectrum of challenges from developing more efficient aligners to better data storage and data sharing solutions.

AUTHOR CONTRIBUTIONS

LG envisioned this project, OP, XZ, TC, and LG wrote the manuscript, all authors have read and agreed on the manuscript.

ACKNOWLEDGMENTS

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, 1R01LM012373 awarded by NLM, and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to LG.

REFERENCES

Aaron, T. L. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Amir, E. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., and Bendall, S. C. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* 31, 545–552. doi: 10.1038/nbt.2594

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Balasubramanian, M., and Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science* 295:7. doi: 10.1126/science.295.5552.7a

Barron, M., and Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell rna-sequencing data. arXiv:1605.04492.

Bendall, S. C., Davis, K. L., Amir el-D., Tadmor, M. D., Simonds, E. F., Chen, T. J., et al. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* 157, 714–725. doi: 10.1016/j.cell.2014.04.005

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). "When Is 'Nearest Neighbor' Meaningful?," in *DATABASE Theory–ICDT'99* (Jerusalem: Springer), 217–235.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bose, S., Wan, Z., Carr, A., Rizvi, A. H., Vieira, G., Pe'er, D., et al. (2015). Scalable microfluidics for single cell rna printing and sequencing. *Genome Biol.* 16:120. doi: 10.1186/s13059-015-0684-3

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095. doi: 10.1038/nmeth.2645

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 55–160. doi: 10.1038/nbt.3102

Buettner, F., and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* 28, i626–i632. doi: 10.1093/bioinformatics/bts385

Campbell, K., Ponting, C. P., and Webber, C. (2015). Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell rna-seq profiles. *bioRxiv* 27219. doi: 10.1101/027219

Chandramohan, R., Wu, P.-Y., Phan, J. H., and Wang, M. D. (2013). "Benchmarking RNA-Seq quantification tools," in *Engineering In Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka), 647–650.

Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., et al. (2016). Pan-Cancer analyses reveal long intergenic non-coding rnas relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine* 7, 62–72. doi: 10.1016/j.ebiom.2016.03.023

Cox, M. P., Peterson, D. A., and Biggs, P. J. (2010). SolexaQA: at-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinformatics* 11:485. doi: 10.1186/1471-2105-11-485

der Maaten, L., and Hinton, G. (2008). Visualizing data using T-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33, 285–289. doi: 10.1038/nbt.3129

Diaz, A., Liu, S. J., Sandoval, C., Pollen, A., Nowakowski, T. J., Lim, D. A., et al. (2016). SCell: integrated analysis of single-cell RNA-Seq data. *Bioinformatics* 32, 2219–2220. doi: 10.1093/bioinformatics/btw201

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., et al. (2015). Normalization and noise reduction for single cell RNA-Seq experiments. *Bioinformatics* 31, 2225–2227. doi: 10.1093/bioinformatics/btv122

Dobin, A., and Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinform.* 51, 11.14.1–11.14.19. doi: 10.1002/0471250953.bi1114s51

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi: 10.1038/nmeth.2722

Fan, J.-B., Jean, J. J.-B., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi: 10.1038/nmeth.3734

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5

Fonseca, N. A., Marioni, J., and Brazma, A. (2014). RNA-Seq gene profiling-a systematic empirical comparison. *PloS ONE* 9:e107026. doi: 10.1371/journal.pone.0107026

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Freeman, B. T., Jung, J. P., and Ogle, B. M. (2016). Single-Cell RNA-seq reveals activation of unique gene groups as a consequence of stem cell-parenchymal cell fusion. *Sci. Rep.* 6:23270. doi: 10.1038/srep23270

Gao, Y., Wang, F., Eisinger, B. E., Kelnhofer, L. E., Jobe, E. M., and Zhao, X. (2016). Integrative single-cell transcriptomics reveals molecular networks defining neuronal maturation during postnatal neurogenesis. *Cereb. Cortex*. doi: 10.1093/cercor/bhw040. [Epub ahead of print].

Grün, D., and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell* 163, 799–810. doi: 10.1016/j.cell.2015.10.039

Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. (2015). SINCERA: a Pipeline for Single-Cell RNA-Seq profiling analysis. *PLoS Comput. Biol.* 11:e1004575. doi: 10.1371/journal.pcbi.1004575

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31, 2989–2998. doi: 10.1093/bioinformatics/btv325

Han, L., Zi, X., Garmire, L. X., Wu, Y., Weissman, S. M., Pan, X., et al. (2014). Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. *Sci. Rep.* 4:6485. doi: 10.1038/srep06485

Handel, A. E., Chintawar, S., Lalic, T., Whiteley, E., Vowles, J., Giustacchini, A., et al. (2016). Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. *Hum. Mol. Genet*. 25, 989–1000. doi: 10.1093/hmg/ddv637

Harris, K., Magno, L., Katona, L., Lönnerberg, P., Muñoz Manchado, A. B., Somogyi, P., et al. (2015). Molecular organization of CA1 interneuron classes. *bioRxiv* 34595. doi: 10.1101/034595

Hartuv, E., and Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Inf. Process. Lett.* 76, 175–181. doi: 10.1016/S0020-0190(00)00142-3

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., et al. (2016). Single-Cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* 26, 304–319. doi: 10.1038/cr.2016.23

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17:29. doi: 10.1186/s13059-016-0888-1

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-Cell RNA-Seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi: 10.1038/nmeth.2772

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel Single-Cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779. doi: 10.1126/science.1247651

Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in Single-Cell RNA-Seq analysis. *Nucl. Acids Res*. 44:e117. doi: 10.1093/nar/gkw430

Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biol.* 17:144. doi: 10.1186/s13059-016-1010-4

Jiang, P., Thomson, J., A., and Stewart, R. (2016). Quality control of Single-Cell RNA-seq by SinQC. *Bioinformatics*. doi: 10.1093/bioinformatics/btw176. [Epub ahead of print].

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Katayama, S., Töhönen, V., Linnarsson, S., and Kere, J. (2013). SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29, 2943–2945. doi: 10.1093/bioinformatics/btt511

Katrib, A., Hsu, W., Bui, A., and Xing, Y. (2016). Radiotranscriptomics: a synergy of imaging and transcriptomics in clinical assessment. *Quant. Biol.* 4, 1–12. doi: 10.1007/s40484-016-0061-6

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S. L., et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kim, K. T., Lee, H. W., Lee, H. O., Kim, S. C., Seo, Y. J., Chung, W., et al. (2015). Single-Cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* 16:127. doi: 10.1186/s13059-015-0692-3

Kimmerling, R. J., Szeto, G. L., Li, J. W., Genshaft, A. S., Kazer, S. W., Payer, K. R., et al. (2016). A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat. Commun.* 7:10220. doi: 10.1038/ncomms 10220

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56–61. doi: 10.1038/nature13920

Kvastad, L., Solnestam, B. W., Johansson, E., Nygren, A. O., Laddach, N., Sahlén, P., et al. (2015). Single cell analysis of cancer cells using an improved RT-MLPA method has potential for cancer diagnosis and monitoring. *Sci. Rep.* 5:16519. doi: 10.1038/srep16519

Leek, J. T. (2014). Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. 42. doi: 10.1093/nar/gku864

Leng, N., Choi, J., Chu, L. F., Thomson, J. A., Kendziorski, C., and Stewart, R. (2016). OEFinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data. *Bioinformatics* 32, 1408–1410. doi: 10.1093/ bioinformatics/btw004

Leng, N., Chu, L. F., Barry, C., Li, Y., Choi, J., Li, X., et al. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* 12, 947–950. doi: 10.1038/nmeth.3549

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir el, A. D., Tadmor, M. D., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162, 184–197. doi: 10.1016/j.cell.2015.05.047

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, J., and Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. *Stat. Methods Med. Res.* 22, 519–536. doi: 10.1177/0962280211428386

Liao, Y., Smyth, G. K., and Shi, W. (2013). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. doi: 10.1093/bioinformatics/btt656

Lohr, J. G., Adalsteinsson, V. A., Cibulskis, K., Choudhury, A. D., Rosenberg, M., Cruz-Gordillo, P., et al. (2014). Whole exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.* 32:479. doi: 10.1038/nbt.2892

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. doi: 10.1101/002832

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., et al. (2015). G&T-Seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522. doi: 10.1038/nmeth.3370

Marco, M., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., et al. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci.* 111, E5643–E5650. doi: 10.1073/pnas.1408993111

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17:10. doi: 10.14806/ej.17.1.200

Meyer, S. E., Qin, T., Muench, D. E., Masuda, K., Venkatasubramanian, M., Orr, E., et al. (2016). Dnmt3a haploinsufficiency transforms Flt3-ITD myeloproliferative disease into a rapid, spontaneous, and fully-penetrant acute myeloid leukemia. *Cancer Discov*. 6, 501–515. doi: 10.1158/2159-8290.CD-16-0008

Miyamoto, D. T., Zheng, Y., Wittner, B. S., Lee, R. J., Zhu, H., Broderick, K. T., et al. (2015). RNA-seq of single prostate CTCs implicates noncanonical wnt signaling in antiandrogen resistance. *Science* 349, 1351–1356. doi: 10.1126/science.aab0917

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* 33, 269–276. doi: 10.1038/nbt.3154

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94. doi: 10.1038/nature09807

Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L., and Tse, D. N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *bioRxiv*. 17:112. doi: 10.1186/s13059-016-0970-8

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257

Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., et al. (2016). Single-cell RNA-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell* 165, 1012–1026. doi: 10.1016/j.cell. 2016.03.023

Pettit, J.-B., Tomer,R., Achim,K., Richardson,S., Azizi, L., and Marioni, J. (2014). Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput Biol* 10:e1003824. doi: 10.1371/journal.pcbi.1003824

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*16, 1–10. doi: 10.1186/s13059-015-0805-z

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat,A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Prabhakaran, S., Azizi, E., and Pe'er, D. (2016). "Dirichlet process mixture model for correcting technical variation in single-cell gene expression data." in *Proceedings of The 33rd International Conference on Machine Learning* (New York, NY), 1070–1079.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282

Robinson, M. D., McCarthy,D. J., and Smyth,G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., et al. (2015). Single-Cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33, 1165–1172. doi: 10.1038/nbt.3383

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22, 839–851. doi: 10.1261/rna.053959.115

Shekhar, K., Brodin, P., Davis, M. M., and Chakraborty, A. K. (2014). Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc. Natl. Acad. Sci.U.S.A.* 111, 202–207. doi: 10.1073/pnas.1321405111

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., et al. (2015). Single-Cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17, 360–372. doi: 10.1016/j.stem.2015.07.013

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., et al. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. *Cell Stem Cell* 6, 468–478. doi: 10.1016/j.stem.2010.03.015

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Ting, D. T., Wittner, B. S., Ligorio, M., Jordan,N. V., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* 32, 381. doi: 10.1038/nbt.2859

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Travers, C., Masaki, J., Weirather, J., Garmire, L. X., Ching, T., Masaki, J., et al. (2015). *Non*-coding yet non-trivial: a review on the computational genomics *of lincRNAs. BioData Min.* 8:44. doi: 10.1186/s13040-015-0075-z

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375. doi: 10.1038/nature13173

Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D. A., and Domany, E. (2005). Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* 21, 2301–2308. doi: 10.1093/bioinformatics/bti329

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32, 2128–2135. doi: 10.1093/bioinformatics/btw202

Wang, B., Zhu, J., Pierson, E., and Batzoglou, S. (2016). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *bioRxiv*. 52225. doi: 10.1101/052225

Wang, J.-Y., Bensmail, H., and Gao, X. (2012). Multiple graph regularized protein domain ranking. *BMC Bioinformatics* 13:307. doi: 10.1186/1471-2105-13-307

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 38:e178. doi: 10.1093/nar/gkq622

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). *SLICER:* inferring branched, nonlinear cellular trajectories from single cell *RNA*-seq data. *Genome Biol.* 17:106. doi: 10.1186/s13059-016-0975-3

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016). GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Stat. Genomics Methods Protoc*. 1418, 283–334. doi: 10.1007/978-1-4939-3578-9_15

Xu, C., and Su,. Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al. (2013). *HTQC:* a fast quality control toolkit for illumina sequencing data. *BMC Bioinformatics* 14:33. doi: 10.1186/1471-2105-14-33

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., Manno, G. L., Juréus, A., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934

Zhang, T., Luo, Y., Liu, K., Pan, L., Zhang, B., Yu, J., et al. (2011). BIGpre: a quality assessment package for next-generation sequencing data. *Genomics, Proteomics Bioinformatics* 9, 238–244. doi: 10.1016/S1672-0229(11)60027-2

Zhu, Z., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., et al. (2016). Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* 7:10812. doi: 10.1038/ncomms10812

Zurauskiene, J., and Yau, C. (2015). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *bioRxiv*. 26385. doi: 10.1186/s12859-016-0984-y

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Cell and Developmental Biology

REVIEW
published: 25 October 2016
doi: 10.3389/fcell.2016.00116

Single Cell Isolation and Analysis

Ping Hu[1†], *Wenhua Zhang*[2†], *Hongbo Xin*[1] *and Glenn Deng*[1,3,4*]

[1] The Center for Biotechnology and Biopharmaceutics, Institute of Translational Medicine, Nanchang University, Nanchang, China, [2] Laboratory of Fear and Anxiety Disorders, Institute of Life Science, Nanchang University, Nanchang, China, [3] Yichang Research Center for Biomedical Industry and Central Laboratory of Yichang Central Hospital, Medical School, China Three Gorges University, Yichang, China, [4] Division of Surgical Oncology, Stanford University School of Medicine, Stanford, CA, USA

Individual cell heterogeneity within a population can be critical to its peculiar function and fate. Subpopulations studies with mixed mutants and wild types may not be as informative regarding which cell responds to which drugs or clinical treatments. Cell to cell differences in RNA transcripts and protein expression can be key to answering questions in cancer, neurobiology, stem cell biology, immunology, and developmental biology. Conventional cell-based assays mainly analyze the average responses from a population of cells, without regarding individual cell phenotypes. To better understand the variations from cell to cell, scientists need to use single cell analyses to provide more detailed information for therapeutic decision making in precision medicine. In this review, we focus on the recent developments in single cell isolation and analysis, which include technologies, analyses and main applications. Here, we summarize the historical background, limitations, applications, and potential of single cell isolation technologies.

Keywords: heterogeneity, single cell, isolation, analysis, sequencing

OPEN ACCESS

Edited by:
Ashok Kumar,
University of Louisville, USA

Reviewed by:
Wen-Shu Wu,
University of Illinois at Chicago, USA
Sandra Orsulic,
Cedars-Sinai Medical Center, USA
Adriana Simionescu Bankston,
University of Louisville, USA

***Correspondence:**
Glenn Deng
yaguangdeng@126.com

[†]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *10 June 2016*
Accepted: *07 October 2016*
Published: *25 October 2016*

Citation:
Hu P, Zhang W, Xin H and Deng G (2016) Single Cell Isolation and Analysis. Front. Cell Dev. Biol. 4:116. doi: 10.3389/fcell.2016.00116

INTRODUCTION

The cell is the fundamental unit of biological organisms. Despite the apparent synchrony in cellular systems, analyzed single cell results show that even the same cell line or tissue, can present different genomes, transcriptomes, and epigenomes during cell division and differentiation (Schatz and Swanson, 2011). For example, a developing embryo, brain, or tumor have intricate structures consisting of numerous types of cells that may be spatially separated. Thus, the isolation of distinct cell types is essential for further analysis and will be valuable for diagnostics, biotechnological and biomedical applications.

Conventional cell-based assays mainly measure the average response from a population of cells, assuming the average response is representative of each cell. However, in doing this important information about a small but potentially relevant subpopulation maybe lost, particularly in cases where that subpopulation determines the behavior of the whole population. For example, the tumor microenvironment is a complex heterogeneous system that consists of multiple intricate interactions between tumor cells and its neighboring non-cancerous stromal cells. The stromal cells are composed of endothelial cells, fibroblasts, macrophages, immune cells, and stem cells. Due to the variation in genetic and environmental factors, different kinds of cells have unique behaviors and present different implications in pathogenic conditions (Schor and Schor, 2001). These challenges make conventional analysis insufficient. Therefore, new technologies to isolate individual single cells from a complex sample and study the genomes and proteomes of single cells could provide great insights on genome variation and gene expression processes. It is believed that single cell analyses have influences on various fields including life sciences and biomedical research (Blainey and Quake, 2014).

In early times, researchers have applied low-throughput single cell analysis techniques, such as immunofluorescence, fluorescence *in situ* hybridization (FISH) and single cell PCR, to detect certain molecular markers of single cells (Taniguchi et al., 2009; Citri et al., 2012). These techniques allow quantification of a limited number of parameters in single cells. On the other hand, high-throughput genomic analysis, such as DNA and RNA sequencing are now widely used. However, genomic studies rely on studying collective averages obtained from pooling thousands to millions of cells, precluding genome-wide analysis of cell to cell variability. Therefore, single cell sequencing developed alongside its necessity in research awarding it "method of the year" by Nature Methods in 2013 (2014). By using single cell analysis, researchers have profiled many biological processes and diseases at the single cell level including tumor evolution, circulating tumor cells (CTCs), neuron heterogeneity, early embryo development, and uncultivatable bacteria.

In this review, we discuss the technologies recently developed for single cell isolation, genome acquisition, transcriptome, and proteome analyses, and their applications. We also briefly discuss the future potentials of single cell isolation technologies and analyses.

TECHNOLOGIES FOR SINGLE CELL ISOLATION

Before initiating a single cell analysis, scientists need to isolate or identify single cells. The performance of cell isolation technology is typically characterized by three parameters: efficiency or throughput (how many cells can be isolated in a certain time), purity (the fraction of the target cells collected after the separation), and recovery (the fraction of the target cells obtained after the separation as compared to initially available target cells in the sample). The current techniques show different advantages for each of the three parameters.

Based on the variety of principles used, current existing cell isolation techniques can be classified into two groups. The first group is based on physical properties like size, density, electric changes, and deformability, with methods including density gradient centrifugation, membrane filtration and microchip-based capture platforms. The most advantageous physical properties is single cell isolation without labeling. The second group is based on cellular biological characteristics, comprising of affinity methods, such as affinity solid matrix (beads, plates, fibers), fluorescence-activated cell sorting, and magnetic-activated cell sorting, which are based upon biological protein expression properties (Dainiak et al., 2007). Thus, in what follows we briefly summarize the principle of each method, as well as the advantage and limitation of their applications (**Table 1**). We will not discuss limiting dilution since it is well known in the field of monoclonal cell cultures production.

Fluorescence Activated Cell Sorting (FACS)

Fluorescence Activated Cell Sorting (FACS), a specialized type of flow cytometry with sorting capacity, is the most sophisticated and user-friendly technique for characterizing and defining different cell types in a heterogeneous cell population based on size, granularity, and fluorescence. FACS allows simultaneous quantitative and qualitative multi-parametric analyses of single cells (Gross et al., 2015). Before separation, a cell suspension is made and the target cells are labeled with fluorescent probes. Fluorophore-conjugated monoclonal antibodies are the most widely used fluorescent probes (mAb) that recognize specific surface markers on target cells. As the cell suspension runs through the cytometry, each cell is exposed to a laser, which allows the fluorescence detectors to identify cells based on the selected characteristics. The instrument applies a charge (positive or negative) to the droplet containing a cell of interest and an electrostatic deflection system facilitates the collection of the charged droplets into appropriate collection tubes for later analysis (**Figure 1A**). Although FACS has been widely used for isolation of highly purified cell populations, it has been reported that FACS can also be used to sort single cells (Schulz et al., 2012). For example, BD cell sorting systems (such as the BD FACSAria III Cell Sorter) are able to isolate single cells of interest from thousands of cells in a population using up to 18 surface markers.

Since the late 1960s, remarkable advances have been made on the FACS technology including the instrumentation and the availability of a large number of highly specific antibodies. The capability of FACS technology has improved significantly from a technique limited to measuring 1–2 fluorescent species per cell to 10–15 species. The maximum number of proteins that can be simultaneously measured has progressively increased (Wu and Singh, 2012). Due to this progress, our understanding of immunology and stem cell biology has improved tremendously alongside the discovery of scores of functionally diverse cell populations (Bendall et al., 2012). It has also been reported that using the next generation cytometry, "post-fluorescence" single cell technology termed mass cytometry is theoretically capable of measuring 70–100 parameters.

Although FACS has been widely used in both basic and clinical research, there are several limiting disadvantages. First, FACS requires a huge starting number of cells (more than 10,000) in suspension. Therefore, it fails to isolate single cells from a low quantity cell population. Second, the rapid flow in the machine and non-specific fluorescent molecules can damage the viability of the sorted cells rendering the isolation a failure. Moreover, cells or cell cultures must be subjected to stimulation experiments and treated in a separate environment before FACS analysis.

Magnetic-Activated Cell Sorting (MACS)

Magnetic-Activated Cell Sorting (MACS) is another commonly used passive separation technique to isolate different types of cells depending on their cluster of differentiation. It has been reported that MACS is capable of isolating specific cell populations with a purity >90% purification (Miltenyi et al., 1990). MACS is based on antibodies, enzymes, lectins, or strepavidins conjugated to magnetic beads to bind specific proteins on the target cells. When a mixed population of cells is placed in an external magnetic field, the magnetic beads will activate and the labeled cells will polarize while other cells are washed out. The remaining cells can be acquired by elution after the magnetic field is turned off (**Figure 1B**). With this technique, the cells can be separated by

TABLE 1 | Overview of single cell isolation techniques.

Techniques	Throughput	Advantage	Disadvantage	References
Fluorescence-activated cell sorting (FACS)	High	High specificity multiple parameters	Large amount of material, dissociated cells, high skill needed	Gross et al., 2015
Magnetic-activated cell sorting (MACS)	High	High specificity, cost effective	Dissociated cells, non-specific cell capture	Welzel et al., 2015
Laser capture microdissection (LCM)	Low	Intact fixed and live tissue	Contaminated by neighboring cells, high skill needed	Espina et al., 2007; Datta et al., 2015
Manual cell picking	Low	Intact live tissue	High skill needed, low throughput	Citri et al., 2012
Microfluidic	High	Low sample consumption, integrated with amplification	Dissociated cells, high skill needed	Bhagat et al., 2010; Lecault et al., 2012

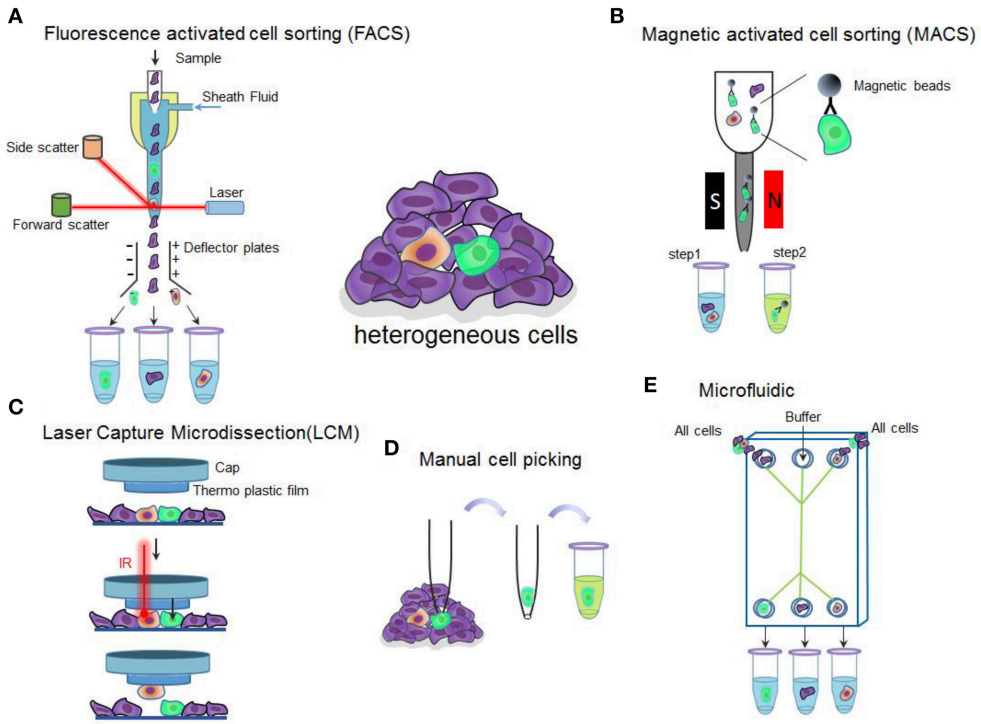


FIGURE 1 | Overview of single-cell isolation technologies discussed in the section. (A) Schematic of fluorescence-activated cell sorting. The suspended labeled cells are passed as a stream in droplets with each containing a single cell in front of a laser. The fluorescence detection system detects the fluorescent and light scatter characteristics. Based on their characteristics, the instrument applies a charge to the droplet containing a cell of interest and an electrostatic deflection system facilitates collection of the charged droplets into different collecting tubes. Cells labeled with green, purple, and yellow indicate different cell types. **(B)** Schematic of magnetic-activated cell sorting. Cells of interest are labeled with specific antibody conjugated magnetic beads. An external magnetic field is used to separate the labeled cells from the cell suspension. S and N indicate magnetic field. **(C)** Schematic of laser capture microdissection. The technique utilizes a laser which fired through the cap over the cells of interest to melt the membrane to let the cells adhere to the melted membrane. When the cap is removed, captured cells are removed, leaving the unwanted cells behind. **(D)** Schematic of manual cell picking. The cells of interest are monitored under a microscope. By using a glass pipette connected to a micromanipulator, single cells can be collected and transferred to a new tube for following analysis. **(E)** Schematic of microfluidic used for single cell isolation. Before starting the experiments, cells need to be dissociated then flow into a chip. Thus, the cells may be separated into different tubes containing only one cell.

charge with respect to the particular antigens. Positive separation techniques use coated magnetic beads and attract cells. The cells of interest are labeled while the unlabeled cells are discarded. In contrast, if species-specific substances are unavailable, a good choice is to use negative separation techniques which employ a cocktail of antibodies to coat untreated cells. In this case, labeled cells are discarded while unlabeled are retained (Grützkau and Radbruch, 2010).

Of the two most common affinity-based techniques for specific cell isolation, MACS technology is comparatively simple and cost-effective. However, the MACS system's obvious shortcoming lies in its initial costs in the separation magnet, and running costs including not only the price of the conjugated magnetic beads, but also replacement columns. In addition, the final purity of isolated cells in MACS devices depends on the specificity and the affinity of the antibodies used to select the

target cells. It also depends on the amount of non-specific cell capture. Non-specific contamination can be from adsorption of background cells to the capturing device or their entrapment within the large excess of magnetic particles needed for labeling rare cells in large volumes. Using new materials can eliminate contamination from non-specific adsorption or entrapment of other blood cells. Another disadvantage of MACS is that it can only utilize cell surface molecules as markers for separation of live cells. Furthermore, it should be noted that MACS is far more limited than FACS because of immunomagnetic techniques that can only separate cells into positive and negative populations. High and low expression of a molecule cannot be separated while it is possible by using FACS sorting.

Laser Capture Microdissection (LCM)

Laser Capture Microdissection (LCM) is an advanced technology for isolating pure cell populations or a single cell from mostly solid tissue samples on a microscope slide (Emmert-Buck et al., 1996). It can accurately and efficiently target and capture the cells of interest to fully exploit emerging molecular analytical technologies, including PCR, microarrays and proteomics (Espina et al., 2007). Today, there are two general classes of laser capture microdissection systems: infrared (IR LCM) and ultraviolet (UV LCM). The LCM system consists of an inverted microscope, a solid state near infrared laser diode, a laser control unit, a joy stick controlled microscope stage with a vacuum chuck for slide immobilization, a CCD camera, and a color monitor (Datta et al., 2015). The basic principle of LCM starts with visualizing the cells of interest through an inverted microscope, then a fixed-position, short duration and focused laser pulse is delivered to melt the thin transparent thermoplastic film on a cap above the targeted cells. The film melts and fuses with the underlying cells of choice. When the film is removed, the target cells remain bound to the film while the rest of the tissue is left behind. Finally, transfer the cells to a microcentrifuge tube containing buffer solutions required for a wide range of downstream analysis (Kummari et al., 2015; **Figure 1C**).

The most important advantage of LCM is its speed while maintaining precision and versatility (Fend and Raffeld, 2000). LCM provides a rapid, reliable method to procure pure populations of target cells from a wide range of cell and tissue preparations via microscopic visualization (Bonner et al., 1997). Conventional techniques for molecular analysis require dissociation of tissue. This may introduce inherent contamination problems and reduce the specificity and sensitivity to subsequent molecular analysis. On the other hand, LCM is a "no touch" technique that does not destroy adjacent tissues after initial microdissection. Morphology of both the captured cells as well as the residual tissue is well preserved and reduces the danger of tissue loss (Esposito, 2007). In addition, after removing the chosen cells, the remaining tissue on the slide is fully accessible for further capture, allowing comparative molecular analysis of adjacent cells.

The major requirement for effective LCM is correct identification of cell subpopulations or single cells in a complex tissue. Thus, the major limitation is the need to identify cells of interest through visual microscopic inspection of morphological characteristics, which in turn, requires a pathologist, cytologist, or technologist trained in cell identification (Espina et al., 2007). Another significant limitation is that the microdissected tissue section does not have a cover slip. Cover slipping would prevent physical access to the tissue surface, which is crucial to any current microdissection method. Without a cover slip, and the index matching between the mounting media and the tissue, the dry tissue section has a refractile quality, which might obscure cellular detail at high magnifications. Moreover, LCM introduces a number of technical artifacts, including slicing the cells during the preparation of tissue sections and UV damage to DNA or RNA from the laser cutting energy (Allard et al., 2004).

Manual Cell Picking/Micromanipulation

Manual cell picking is a simple, convenient, and efficient method for isolating single cells. Similar to LCM, manual cell picking micromanipulators also consists of an inverted microscope combined with micro-pipettes that are movable through motorized mechanical stages. Each isolated single cell can be observed and photographed under the microscope, thus enabling unbiased isolation (**Figure 1D**). Unlike LCM that mainly isolates single cells from sections of fixed tissue, micromanipulation plays an important role in isolating live culture cells or embryo cells.

Micromanipulation can be easily performed in an electrophysiology lab equipped with a patch clamp system. For example, after investigating neuronal function in brain slices preparations after standard whole-cell patch-clamp electrophysiological recordings, scientists would apply negative pressure through the patch pipette so that the cytosolic material containing cellular mRNA can be aspirated for further analysis (Eberwine et al., 1992; Citri et al., 2012). However, the throughput is limited and it requires highly skilled professionals to perform, it has the utility limitation when detecting complex changes.

Microfluidics

Microfluidics is recognized as a powerful enabling technology for investigating the inherent complexity of cellular systems as it provides precise fluid control, low sample consumption, device miniaturization, low analysis cost, and easy handling of nanoliters-volumes (Whitesides, 2006; **Figure 1E**). Cell Sorting by a microfluidic chip can be divided into four categories: cell-affinity chromatography based microfluidic (Nagrath et al., 2007), physical characteristics of cell based microfluidic separation, immunomagnetic beads based microfluidic separation, and separation methods based on differences between dielectric properties of various cell types.

Cell-affinity chromatography based microfluidic is the most commonly used method for microfluidic chip analysis. It is based upon highly specific interactions between antigen and antibody, ligand and receptor. At the beginning of the process, the micro-channel in the chip is modified with specific antibodies capable of binding to cell surface antigen or aptamer, such as an epithelial cell adhesion molecule. Once the sample flows through the micro-channels, its cell surface antigen can bind to the specific antibodies or aptamer immobilizing the cells on the chip, while the remaining cells flow off the chip with the buffer. Finally,

using a different buffer, we can elute the immobilized cells for downstream analysis. Compared to other separation methods, affinity based systems have higher specificity and sensitivity because of the recognition-binding event.

Today, microfluidics can be combined with different separation methods, such as filtration and sedimentation or affinity-based technologies like FACS and MACS. In the recent years, numerous investigations and applications in microfluidic devices have been reported, including cancer research, microbiology, single-cell analysis, stem cell research, drug discovery, and screening (Arora et al., 2010; Li et al., 2012a). Recently, microfluidic chips have been fabricated from silicon or glass, elastomer, thermosets, hydrogel, thermoplastics, and paper (Ren et al., 2013, 2014). The advantages and disadvantages of the materials used in microfluidic chips have been well-summarized previously (Ren et al., 2014). Microfluidics are used to manipulate liquids (dimensions from 1 to 1000 μm) in networks of micro-channels in a single device. At such ultralow volumes, fluids exhibit different physico-chemical properties compared to their behavior at the macro-scale (Squires and Quake, 2005). Other common fluids can be used in microfluidic devices include bacterial cell suspensions, whole blood samples, protein or antibody solutions, and various buffers.

Taking advantages of integrating cell handling and processing concurrently, microfluidic chips show potential applications in DNA sequencing (Hashimoto et al., 2007; Liu et al., 2007), protein analysis (Emrich et al., 2007), cell manipulation, and cell composition analysis (VanDijken et al., 2007; Bhagat et al., 2010). For example, Fluidigm developed a commercially available valve-based microfluidic qPCR system called the Dynamic Array™. This system advanced on providing low-volume (nanoliter) and high-throughput (thousands of PCR reactions per device) methods to the researchers and has become increasingly popular for large-scale single cell studies. Moreover, microfluidic technology has shown increasing applications in studying diversity and variations in single cell genomes, spanning from cancer biology to environmental microbiology and neurobiology. Beyond genomics applications, the scalability and small volume advantages of microfluidic methods have found applications in the measurement of intracellular and secreted proteins from single cells.

SINGLE CELL ANALYSIS

Single cell analysis tools can be divided into three groups: genomics, transcriptomics, and proteomics. Due to next generation sequencing (NGS) technologies as well as whole genome/transcriptome amplification (WGA/WTA) approaches, a new scientific field of single cell genome studies have been established. A combination of high-throughput and multiparameter approaches is used in single cell analysis which can reflect cell to cell variability and heterogeneous differences in the individual cells. Therefore, the development of efficient single cell analysis methods requires attention. In this section, we discuss novel technologies designed for single cell analysis of genomics, transcriptomics, and proteomics (**Table 2**).

Single Cell Genomics

Single cell genome sequencing allows us to identify chromosomal variations, such as copy number and single-nucleotide variations. It also allows us to study tumor evolution, gamete genesis, and somatic mosaicism, which is reflected in the genomic heterogeneity among a population of cells. However, in humans, it often faces the low amount of genome materials, for example, the weight of one genomic DNA is only 6 pg and each gene in the genome only has two copies in a single normal cell which is not quite enough for the current NGS use. However, amplification using traditional PCR suffers from severe biases and allelic dropout across the genome when it is applied to single cells. Therefore, a precise, unbiased amplification of the DNA is critical to single cell genome sequencing. Lots of attempts were made, mostly by modifying the traditional PCR methodology to linker-adapter PCR (LA-PCR) (Klein et al., 1999), interspersed repetitive sequence PCR (IRS-PCR), primer extension pre-amplification PCR (PEP-PCR) (Hubert et al., 1992), degenerate oligonucleotide-primed PCR (DOP-PCR) (Telenius et al., 1992), and its variant displacement DOP-PCR (D-DOP-PCR) (Langmore, 2002). For example, by using DOP-PCR, Navin and colleagues demonstrated accurate and robust determination of genome wide copy number in rearranged cancer genomes (Navin et al., 2011). This is the first report of single cell genome sequencing applied to a cancer genomic heterogeneity study. However, these methods also have some limitations in low coverage, amplification bias, and allele dropout.

The multiple displacement amplification (MDA) is the most popular method applied in genome analysis due to its high fidelity and simplicity. It can amplify DNA in a 30°C isothermal reaction with random hexamer primers and phi29 DNA polymerase. The kernel of MDA is that phi29 DNA polymerase can extend the primers with high fidelity and strong processivity, which exhibits powerful strand displacement ability during the new strand synthesis (Dean et al., 2002). The displacement process generates single stranded DNA templates, which are reprimed and extended, thereby amplifying the DNA in an isothermal reaction. Based on MDA, Xu and colleagues provided the first intratumoral genetic landscape at a single-cell level and demonstrated that clear cell renal cell carcinoma (ccRCC, the most common kidney cancer) may be more genetically complex than previously thought (Xu et al., 2012). However, MDA also suffers from strong biases and high allelic dropout rate across the genome, making the reaction vulnerable to generating "chimeras," resulting in unwanted noise and false results.

Another new method, multiple annealing and looping-based amplification cycles (MALBAC) showed faithful copy number variation detection (Zong et al., 2012), which can amplify the genome of a single cell with high uniformity. MALBAC is based upon strand displacement pre-amplification that generates amplicons with complementary ends. Thus, the full amplicons generated in the reaction seal themselves to form loops to prevent them from being amplified again. This also ensures that each new amplicon is replicated from the original templates. Therefore, the obvious advantage of MALBAC is that it can reduce the

TABLE 2 | Techniques for single cell analyses.

	Methods	Classification	Throughput	Advantage	Disadvantage	References
Genome	PCR*	LA-PCR*, IRS-PCR*, PEP-PCR*, DOP-PCR*	High	High coverage	Uneven coverage, amplification bias, allele dropout	Klein et al., 1999
	MDA*	None	High	Homogeneous coverage	Amplification bias, allele dropout, "chimera" structure	Spits et al., 2006
	MALBAC*	None	High	Homogeneous coverage	Amplification bias, allele dropout	Lu et al., 2012: Van Loo and Voet, 2014
Transcriptome	PCR-based amplification	RNA-seq, TPEA*, SMART*	High	Amplify quickly	Distort the difference	Pan, 2014
	IVT*	CEL-seq Quartz-seq	High	Specificity, ratio fidelity	Low efficiency	Hebenstreit, 2012; Liu et al., 2014
	Phi29 DNA polymerase	TTA* PMA*	High	High efficient, low bias	RNA need to be selected from the gDNA	Pan et al., 2013; Liu et al., 2014
Protein	Flow cytometry	None	High	More species	Spectral overlap	Haselgrübler et al., 2014
	Microfluidic flow cytometry	None	High	Small number of cells	Dissociated cells, high skill needed	Wu and Singh, 2012
	Mass spectrometry	LDI-MS*, SIMS* (MALDI)-MS*	High	Low sensitivity	No molecular labels, Femtomolar sensitivity	Haselgrübler et al., 2014; Liu et al., 2014

**PCR, Polymerase chain reaction; *LA-PCR, linker-adapter PCR; *IRS-PCR, Interspersed repetitive sequence PCR; *PEP-PCR, Primer extension pre-amplification PCR; *DOP-PCR, degenerate oligonucleotide-primed PCR; *MDA, Multiple displacement amplification; MALBAC, Multiple annealing and looping-based amplification cycles; *TPEA, 3′-end amplification; *SMART, strand-switch-mediated reverse transcription amplification; *IVT, in vitro transcription; TTA, Total transcript amplification; *PMA, Phi29 mRNA amplification; LDI-MS, Laser desorption and ionization mass spectrometry; *SIMS, Secondary ion mass spectrometry; *MALDI-MS, Matrix-assisted laser desorption/ionization mass spectrometry.*

amplification errors and biases as the starting materials of the exponential amplification are amplicon separately copied from the original template. However, it is still needed to improve the fidelity and lower the bias (Marcy et al., 2007; Wu et al., 2014).

Single Cell Transcriptomics

Single cell transcriptome sequencing has recently emerged as a powerful technology for revealing differential gene expression and diverse RNA splicing patterns during early embryonic development, differentiation and reprogramming. The main application of single-cell transcriptomics is to connect a cell's genotype to phenotype. It is able to detect thousands of transcripts in various kinds of tissues and cells (Cloonan et al., 2008; Mortazavi et al., 2008). Although mRNA is not as rare as DNA in a single cell, there are still thousands of copies. This is ideal since NGS transcriptome sequencing also requires a large amount of RNA as the starting material. The mRNA from single cells needs to be reverse-transcribed to cDNA followed by cycles of PCR amplification (Sandberg, 2014). The key process in completing single cell mRNA amplification successfully is based on performing reverse transcription to double-strand DNA with high efficiency and low biases.

PCR-based amplification was first reported in single-cell transcriptome analysis of the preparation of single-cell cDNAs using cDNA microarray and RNA-seq analysis (Brady and Iscove, 1993). The disadvantage of a microarray is the low detection sensitivity that would likely miss many low-level but key transcripts. Compared to microarray analysis, RNA-seq analysis expanded the spectrum of detected genes with high accuracy and effectively increased the proportion of full-length cDNA. One advantage of PCR-based mRNA transcriptome amplification bias is that it makes the expression difference more visible between samples and any RNA starting amount can be employed. But on the other hand, it may distort the original difference when it is marginal. Several modified PCR-based methods of cDNA amplification have been developed, such as global PCR amplification (GA), 3′-end amplification (TPEA), and strand-switch-mediated reverse transcription amplification (SMART) (Pan, 2014).

In vitro transcription (IVT)-based amplification linear RNA amplification is the first strategy that has been used to successfully amplify RNA for molecular profiling studies, which promoted the birth of the era of single cell analysis (Liu et al., 2014). It is based on T7 RNA polymerase-mediated IVT and requires three rounds of amplification. The main advantages of the IVT strategy include its specificity, ratio fidelity, and reducing accumulation non-specific products, but has the drawback of low efficiency and a time consuming procedure.

Recently, single cell RNA amplification methods have been raised based on the Phi29 DNA polymerase (Blanco and Salas, 1984; Dean et al., 2002). This polymerase is a highly processive enzyme with strong strand displacement activity that allows for highly efficient isothermal DNA. The phi29 DNA polymerase-based transcriptome amplification method is a simple, fast and isothermal reaction (Liu et al., 2014). The primary advantage of this method is the highly efficient, low bias, and uniform nature of amplification.

Furthermore, in order to retain the spatial and temporal information of RNAs in cells, several new RNA sequencing methods have been developed, including transcriptome *in vivo*

analysis (TIVA), single molecule fluorescent *in situ* hybridization (smFISH), fluorescent *in situ* RNA sequencing (FISSEQ), and so on (Lee et al., 2014; Lovatt et al., 2014). These technologies become powerful tools for unraveling longstanding biomedical questions.

Single Cell Proteomics

Single cell analysis of DNA and RNA can provide qualitative information about protein expression. However, they cannot give information on protein concentration, location, post-translational modifications, or interactions with other proteins. Thus, single-cell proteomics help us obtain much more information that is crucial in cell signaling and cell to cell heterogeneity. Traditional protein analysis techniques, such as gel electrophoresis, immunoassays, chromatography, and mass spectrometry require numerous cells for analysis. Therefore, the major challenges of analyzing proteins at the single-cell level are the exceedingly small copy number of individual proteins and the lack of amplification methods. However, recent advances in multiparameter flow cytometry, microfluidics, mass spectrometry, mass cytometry, and other techniques have led to new single cell proteomics studies that could be performed with greater sensitivity and specificity.

Not only widely used in cell sorting, flow cytometry is also the most established and user-friendly method for both qualitative and quantitative multiparameter analysis of single cells. As mentioned before, by using multiparameter flow cytometry, scientists can simultaneously measure 10–15 key proteins in signaling pathways in individual cells (De Rosa et al., 2001; Perez and Nolan, 2002). In addition, in an immunological proof-of-concept study, as many as 19 separate parameters including 17 fluorescent colors and 2 physical parameters were analyzed (Perfetto et al., 2004). This strong ability has turned flow cytometry into a powerful tool to semi-quantitatively analyze pathways underlying many diseases (Irish et al., 2004; Sachs et al., 2005). The main limitation is the spectral overlap due to the broad spectral emission bands of organic fluorescent dyes. Quantum dots mitigate but do not eliminate the problem. Hence, complex correction algorithms are required for spectral deconvolution. Moreover, commercial flow cytometers use cell suspensions, which in turn allow individual interrogation of cells. The sample preparation is still done manually and therefore, requires a large numbers of cells (More than 10,000). This makes it hard to analyze small samples, such as cells recovered from a biopsy, tissue specimens or small volumes of blood.

To overcome these limitations, efforts have been made to develop microfluidic-based miniaturized flow cytometers which permit analysis of small numbers of cells (100–1000) (Lindström and Andersson-Svahn, 2010). For example, Su and colleagues developed a microscope-based label-free microfluidic cytometer. It is capable of acquiring two dimensional light scatter patterns from the smallest mature blood cells (platelets), cord blood hematopoietic stem/progenitor cells (CD34 + cells), and myeloid precursor cells (Su et al., 2011). Srivastava et al. (2009) developed an integrated microfluidic device which retro-fitted to commercial. The major advantage of this microfluidic device is its ability to perform cell culture, stimulation and sample preparation in combination with conventional fluorescence imaging and microfluidic flow cytometry to monitor immune response in macrophages. These microfluidic devices not only drastically reduced the amount of sample and reagent required, but also provided a means to perform two orthogonal modes of measurements-imaging and cytometry, in one experiment.

Mass spectrometry (MS) is the most powerful tool for protein analysis. However, MS's use for analyzing proteins in single cells is limited due to the lack of sensitivity to detect low amounts of proteins. Fractionation of the cell lysate by capillary electrophoresis (CE) prior to MS offers a good way to improve sensitivity. Recently, a format for flow cytometry has been developed that leverages the precision of mass spectrometry which is termed mass cytometry. It can uniquely enable the measurement of over 40 simultaneous cellular parameters on single cells with the throughput capacity to survey millions of cells from an individual sample (Mellors et al., 2010).

APPLICATION OF SINGLE CELL ANALYSIS

The exponential growth in studies applying single cell analysis is explicitly tied to the acceptance of the technique by biologists. Single cell analysis has influenced and impacted different domains of science including cancer biology, neuroscience, and immunology and so on. It is impossible to document each of these developments. Therefore, a short overview of the fields of applications that are typically addressed by single cell analysis is presented in the research and application for cancer, brain and stem cell, etc.

Application of Single Cell Analysis in Cancer, Neuron Research

Intra-tumor heterogeneity has been widely reported in numerous human cancer types. Tumors are frequently composed of individual, molecularly distinct clones that differ in their proliferation rates and metastatic potential, most critically, in their sensitivities and responses to drug treatment. Those cells that can cause distant metastases should possess unique characteristics when compared to the remaining subpopulation. Exome sequencing of single cells isolated from primary renal carcinomas showed that only 31–37% of the genetic lesions within a tumor are identical to the rest of the tumor cells (Gerlinger et al., 2012; Xu et al., 2012). Therefore, analyzing the occurrence, development and metastasis of these tumors at a single cell level provides much more detailed information on how a drug will respond to the tumor cells. It has been reported that the PIK3CA mutations were detected in primary and metastatic tumor tissues, but it is different periodically in single cells of CTCs and DTCs indicated the drug efficacy (Deng et al., 2014).

Several important types of cancer cells have been discovered, including primary tumor cells, metastatic tumor cells, cancer stem cells (CSC), circulating tumor cells (CTC), and disseminated tumor cells (DTC) (Zhang et al., 2016). CTC and DTC play a vital role in cancer dissemination, self-renewal, and distant metastases. They are being increasingly recognized for their potential utility in disease monitoring and therapeutic

targeting. Many cancer patients are diagnosed with early-stage cancer with no clinical symptoms of metastasis but subsequently succumb to metastatic relapse. One important reason is that CTCs in the blood and DTCs have already reached a secondary organ but have not yet grown to become clinical metastasis. However, the CTCs are so rare among massive numbers of blood cells, as few as one cell per 10 million white blood cells and 5 billion red blood cells, that the accurate identification of CTCs turns out to be the most difficult step in the isolation process (Deng et al., 2008). In recent years, a variety of enrichment and detection techniques have been developed, making significant progress in CTC detection. For example, the CellSearch® system (Janssen Diagnostics, NJ, USA) is the first and the only technique that has been approved by the US FDA for the detection, enrichment and quantification of CTCs in peripheral whole blood samples (Riethdorf et al., 2007). This system utilizes magnets with ferrofluid nanoparticles conjugated to antibodies that target epithelial cell adhesion molecules, such as EpCAM and CD45. EpCAM is the most commonly used epithelial marker that is present on epithelial tumor cells while CD45 is an immunocyte marker that is present on many blood cells but absent in epithelial cells. Thus, the findings of EpCAM-positive and CD45-negative cells indicate the presence of CTCs. Another new immunomagnetic separation technology, called MagSweeper (Illumina), involves dipping a rotating magnetic rod with bound EpCAM antibodies in order to isolate CTCs. Then moving the magnetic rod into a new buffer to release the CTCs (Talasaz et al., 2009; Powell et al., 2012). The MagSweeper can be used reliably to extract functional human CTCs from the blood of mice inoculated with human tumor xenografts, while retaining both their tumor-initiating and metastasizing capacities (Ameri et al., 2010). This highlights the most advantageous aspect of MagSweeper is that CTCs can be completely isolated while preserving the integrity and viability of these fragile cells.

In recent years, a large number of studies have been reported using single cell analysis to analyze individual tumor cells isolated from breast cancer (Navin et al., 2011; Deng et al., 2014; Wang et al., 2014; Eirew et al., 2015), colon cancer (Zong et al., 2012; Yu et al., 2014), pancreatic adenocarcinomas (Ruiz et al., 2011), muscle-invasive bladder cancer (Li et al., 2012b), intestinal cancer (Grün et al., 2015), lung adenocarcinoma cancer (Kim et al., 2015), renal cell carcinoma (Gerlinger et al., 2012; Li et al., 2012b), and acute myeloid leukemia (Ding et al., 2012; Hughes et al., 2014; Paguirigan et al., 2015). For example, Navin and colleagues investigated copy number variation in single tumor cells using DOP WGA followed by DNA sequencing to determine cell population structure and tumor evolution patterns in a single breast tumor (Navin et al., 2011). This study provided an important breakthrough for research on tumor evolution and offered a way to assess the genetic details of tumor structure. Hou and colleagues applied MDA based single cell sequencing technology for the first time to analyze primary thrombocytosis disease (essential immature, ET) in patients at single bone marrow cell level (Hou et al., 2012). Thus, understanding tumor heterogeneity via single cell analysis is considered the biggest challenge in cancer research and if elucidated would enhance our ability to determine the best treatment options.

It is no exaggeration to say that the brain is the most complex structure in the human body. There are more than 100 billion neurons in the human brain. Each of them can make approximately 10,000 direct connections with others, totaling some 100 trillion nerve connections. This makes the brain a complicated network (Herculano-Houzel, 2009). The brain is divided into several regions. Each region consist of various morphologically and/or neurochemically distinct neurons surrounded by various types of glial cells (oligodendrocytes, microglia, and astrocytes). Additionally, distinct regions in the brain, such as areas of the cerebral cortex, hippocampus have specific functions. The cerebral cortex is responsible for many "higher-order" functions like language and information processing while the hippocampus is involved in spatial learning and memory. Increasing evidence shows that each brain region contains different types of neurons according to their location, neurotransmitter identity, connectivity, electrophysiological properties, and molecular markers. Changes of genomic content and epigenetic profiling of specific neuronal or glia subtypes are involved in the pathogenesis of neuropsychiatric diseases, such as Parkinson's and Alzheimer's diseases and autism spectrum disorders(Citri et al., 2012).

Hence there is no doubt that single cell isolation and analysis have made increasingly significant contributions to our understanding of the role that somatic genome variations play in neuronal diversity and behaviors. For example, MACS based technique has been successfully applied to isolating immature neuronal cells from a large number of embryonic zebrafish; the antibody of PSA-NCAM conjugated microbeads were used within a semi-automated dissociation process. (Welzel et al., 2015). Moreover, the MACS was also used for the isolation of embryonic spinal oligodendroglial progenitor cell populations from the rat embryonic spinal cord. By using superparamagnetic MicroBeads combined with A2B5 antibodies (a specific oligodendroglial development marker) and the Mini-MACS separator column, the oligodendroglial cells were isolated with a cell purity of 58–61% in comparison to 6–12% in an unseparated population (Cizkova et al., 2009).

Moreover, basolateral amygdala (BLA) neurons are used to activate distinct populations of the lateral central nucleus of the amygdala (CeL) neurons to either promote fear or reduce anxiety. Namburi and colleagues identified two populations of neurons in the basolateral amygdala neurons that undergo opposing synaptic changes following fear (negative emotion) or reward (positive emotion) conditioning. By using RNA-seq they identified few differentially expressed candidate genes between these two population neurons that may mediate the effects (Namburi et al., 2015). Usoskin and colleagues used comprehensive transcriptome analysis of 622 single mouse neurons from sensory system and discovered 11 fundamentally distinct types of sensory neurons. Interestingly, each neuron is associated with a different type of sensation (Usoskin et al., 2015). Even cells that appear to be morphologically similar may show marked differences in expression patterns. In neuroscience research, electrophysiological analysis combined with molecular

biology within the same cell will provide convincing results for us to better understand of how changes at the molecular level are manifested in functional properties (Eberwine et al., 1992).

Applications of Single Cell Analysis in Stem Cell Research

Stem cells are undifferentiated cells that are characterized as both being capable of self-renewal and having the potential to differentiate into specialized types of cells. How stem cells balance their self-renewal capacity and their ability to differentiate are central questions in stem cell research. Stem cells can be generally classified into pluripotent stem cells, which can give rise to cells of all three germ layers (the ectoderm, mesoderm, and endoderm) or tissue-specific stem cells (also referred to as somatic or adult stem cells), which play essential roles in the development of embryonic tissues and the homeostasis of adult tissues. Both of these two types of stem cells are intermingled with a variety of differentiated and intermediate cell types in the embryonic or adult tissues, forming heterogeneous populations. Therefore, isolation, analysis, and development of specific therapies that target stem cells give cancer patients hope for improvement in terms of survival and quality of life, (Li et al., 2008; Sharma et al., 2010).

Cancer stem cells (CSCs) are hypothesized to persist in tumors as a distinct population and cause relapses and metastases by forming new tumors. CSC are intrinsically more refractory to the effects of a variety of anticancer drugs possibly via enhanced drug efflux (Trumpp and Wiestler, 2008). These cells are especially resistant to therapeutic drugs. Due to the limited number of CSCs in cancer tissues, isolation and analysis CSCs are still a hard work. Single cell sequencing provides powerful tools for identifying these cells providing new insight into complex intra-tumoral heterogeneity. For example, Patel et al. (2014) used single-cell RNA sequencing to profile 672 single cells from five primary. Each tumor showed high intra-tumoral cell heterogeneity in many aspects, including copy number variations as well as cell cycle, immune response and hypoxia. By examining a set of "stemness" genes, they identified continuous, rather than discrete, stemness-related expression states among the individual cells of all five tumors, reflecting the complex stem cell states within a primary tumor. It has been suggested that CSCs are more resistant to chemo—and radiotherapy than other cells in a tumor. This could be one explanation to why most tumors relapse after therapy. Thus, understanding how cancer stem cells resist medical therapy could lead to the development of new, more efficient cancer treatments. Although the existence of these CSCs is still controversial in many cancer types, there is no doubt that CSCs have the potential to provide a foundation for new innovative treatment targeting the roots of cancer.

The neural stem cells (NSCs) in the subventricular zone (SVZ) and the subgranular zone (SGZ) of the dentate gyrus continually divide and differentiate into mature neurons and glia in the adult rodent brain (Aimone et al., 2014). Although it has been documented that endogenous NSCs can be activated to produce multiple types of progeny to contribute to brain repair after brain injury, people do not know how distinct pools of NSCs may react to brain injury and which molecules trigger injury-induced activation of NSCs. Single-cell sequencing reveals a population of dormant neural stem cells in the SVZ that become activated upon brain injury by down regulation of glycolytic metabolism and a concomitant up regulation of lineage-specific transcription factors and protein synthesis (Llorens-Bobadilla et al., 2015).

Increasing evidence shows that multiple molecularly distinct groups of stem cells that respond differently to physiological stimuli coexist in the tissues. Understanding and implementing this molecular diversity will be critical in harnessing the potential of disease treatment.

CONCLUSION AND OUTLOOK

The biological relevance of cell to cell variations and the high potential of single cell analysis in both basic research and clinical diagnostics have drawn the attention of the scientific community. Single cell gene expression analysis can be used for tumor cell identification; single cell DNA mutation analysis can be used for tumor cell monitoring and clinical decision making (Powell et al., 2012; Deng et al., 2014). Understanding cellular heterogeneity has been a major thrust of technological development over the past decade, resulting in an increasingly powerful suite of instrumentation, protocols, and methods for analyzing single cells at the DNA sequence, RNA expression and protein abundance levels (Kalisky et al., 2011; Wu and Singh, 2012). As remarkable examples, technical developments, and appropriate clinical solutions based on single cell analyses of CTCs and CSCs showed the promise to uncover personalized medicine to fight against cancer.

Although much progress has been made during the recent years in single cell gene analysis, live single cell isolation and molecular analyses are more favorable for global profiling of RNA expression and DNA mutation (Powell et al., 2012). We are still only beginning to face the measurement challenges of cellular heterogeneity. There is still more room for improvement in enabling new modes of analysis and improving the sensitivity, precision, speed and throughput (Lecault et al., 2012).

For single cell genomic and gene expression analyses, the greatest obstacle for direct detection of diverse genomic, transcriptomic, and epigenetic events is whether there is a sufficient amount of DNA or RNA. On the one hand, purification of high-quality nucleotides from a single sample plays a pivotal role for the following studies. A problem that is commonly faced is tube absorption which causes loss of sample materials. Low absorption material containers instead of ordinary tubes and single tube reaction analysis are recommended to reduce the loss of DNA and RNA, single cell direct PCR/RT-PCR without nucleotide isolation are also often used. Another problem is the low replication efficiency of secondary structure DNA sequences. Methods for current single cell sequencing still have relatively high technical noise. It is acceptable when studying highly expressed genes, but the biological variations of genes that are expressed at low levels may be masked. Thus, the efficiency of reverse transcription and PCR amplification should be urgently improved. On the other hand, this problem could be overcome

by the third-generation sequencing platforms, which are based on sequencing single molecules and real-time signal monitoring (Schadt et al., 2010; Liu et al., 2012). Within third-generation sequencing technology, no amplification is required and it also overcomes the issue of PCR amplification bias. However, the detection sensitivity, accuracy of sequencing reads, sample handling, recovery, and sequence assembly still need to be further improved.

Protein analysis is far more challenging than nucleic acid analysis. Undoubtedly, the complexity of the proteome, lack of amplification methods and highly specific high-affinity probes make protein analysis technically demanding. Because the cell contents are highly diluted after lysis, high affinity probes (not only monoclonal antibodies), and highly sensitive detection methods are needed to detect low abundance proteins and post-translational modifications.

To summarize, single cell analysis now stands poised to illuminate this new layer of biological complexity under normal development and disease conditions. Considering the rapid progress in either the development of single cell isolation or analysis technology, many of the problems mentioned above will be solved in the near future. Nevertheless, further developments and interdisciplinary co-operative work between technologists, scientists, and clinicians will be necessary. In the distant future, we expect that the single cell techniques will become a powerful tool to unravel longstanding questions in both biological research and clinical diagnostics.

AUTHOR CONTRIBUTIONS

GD and HX conceived the structure of the manuscript; PH and WZ wrote the manuscript; GD and HX read, edited, and approved the manuscript; Mr. Brian Deng (Stanford University) helped the discussion and correction of English writing.

ACKNOWLEDGMENTS

This work was supported by grants of National Basic Research Program of China (2013CB531103 to HX), National Natural Science Foundation of China (91339113, 81270202 to HX, 81601179 to WZ), Natural Science Foundation of Jiangxi Province of China (20161BAB204166 to WZ, 20161BAB205212 to PH), Shenzhen Basic Research Program (20140825105648 to GD), and Wuhan Science and Technology Bureau grant (2014060202010125 to GD), Hubei 100 Talents and Wuhan 3551 Talents Program (to GD).

REFERENCES

(2014). Method of the year 2013. *Nat. Methods* 11:1. doi:10.1038/nmeth.2801

Aimone, J. B., Li, Y., Lee, S. W., Clemenson, G. D., Deng, W., and Gage, F. H. (2014). Regulation and function of adult neurogenesis: from genes to cognition. *Physiol. Rev.* 94, 991–1026. doi: 10.1152/physrev.00004.2014

Allard, W. J., Matera, J., Miller, M. C., Repollet, M., Connelly, M. C., Rao, C., et al. (2004). Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin. Cancer Res.* 10, 6897–6904. doi: 10.1158/1078-0432.CCR-04-0378

Ameri, K., Luong, R., Zhang, H., Powell, A. A., Montgomery, K. D., Espinosa, I., et al. (2010). Circulating tumour cells demonstrate an altered response to hypoxia and an aggressive phenotype. *Br. J. Cancer* 102, 561–569. doi: 10.1038/sj.bjc.6605491

Arora, A., Simone, G., Salieb-Beugelaar, G. B., Kim, J. T., and Manz, A. (2010). Latest developments in micro total analysis systems. *Anal. Chem.* 82, 4830–4847. doi: 10.1021/ac100969k

Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012). A deep profiler's guide to cytometry. *Trends Immunol.* 33, 323–332. doi: 10.1016/j.it.2012.02.010

Bhagat, A. A., Bow, H., Hou, H. W., Tan, S. J., Han, J., and Lim, C. T. (2010). Microfluidics for cell separation. *Med. Biol. Eng. Comput.* 48, 999–1014. doi: 10.1007/s11517-010-0611-4

Blainey, P. C., and Quake, S. R. (2014). Dissecting genomic diversity, one cell at a time. *Nat. Methods* 11, 19–21. doi: 10.1038/nmeth.2783

Blanco, L., and Salas, M. (1984). Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proc. Natl. Acad. Sci. U.S.A.* 81, 5325–5329. doi: 10.1073/pnas.81.17.5325

Bonner, R. F., Emmert-Buck, M., Cole, K., Pohida, T., Chuaqui, R., Goldstein, S., et al. (1997). Laser capture microdissection: molecular analysis of tissue. *Science* 278, 1481, 1483.

Brady, G., and Iscove, N. N. (1993). Construction of cDNA libraries from single cells. *Meth. Enzymol.* 225, 611–623. doi: 10.1016/0076-6879(93)25039-5

Citri, A., Pang, Z. P., Sudhof, T. C., Wernig, M., and Malenka, R. C. (2012). Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.* 7, 118–127. doi: 10.1038/nprot.2011.430

Cizkova, D., Cizek, M., Nagyova, M., Slovinska, L., Novotna, I., Jergova, S., et al. (2009). Enrichment of rat oligodendrocyte progenitor cells by magnetic cell sorting. *J. Neurosci. Methods* 184, 88–94. doi: 10.1016/j.jneumeth.2009.07.030

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619. doi: 10.1038/nmeth.1223

Dainiak, M. B., Kumar, A., Galaev, I. Y., and Mattiasson, B. (2007). Methods in cell separations. *Adv. Biochem. Eng. Biotechnol.* 106, 1–18. doi: 10.1007/10_2007_069

Datta, S., Malhotra, L., Dickerson, R., Chaffee, S., Sen, C. K., and Roy, S. (2015). Laser capture microdissection: big data from small samples. *Histol. Histopathol.* 30, 1255–1269. doi: 10.14670/HH-11-622

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261–5266. doi: 10.1073/pnas.082089499

Deng, G., Herrler, M., Burgess, D., Manna, E., Krag, D., and Burke, J. F. (2008). Enrichment with anti-cytokeratin alone or combined with anti-EpCAM antibodies significantly increases the sensitivity for circulating tumor cell detection in metastatic breast cancer patients. *Breast Cancer Res.* 10:R69. doi: 10.1186/bcr2131

Deng, G., Krishnakumar, S., Powell, A. A., Zhang, H., Mindrinos, M. N., Telli, M. L., et al. (2014). Single cell mutational analysis of PIK3CA in circulating tumor cells and metastases in breast cancer reveals heterogeneity, discordance, and mutation persistence in cultured disseminated tumor cells from bone marrow. *BMC Cancer* 14:456. doi: 10.1186/1471-2407-14-456

De Rosa, S. C., Herzenberg, L. A., and Roederer, M. (2001). 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nat. Med.* 7, 245–248. doi: 10.1038/84701

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481, 506–510. doi: 10.1038/nature10738

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., et al. (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U.S.A.* 89, 3010–3014. doi: 10.1073/pnas.89.7.3010

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., et al. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 518, 422–426. doi: 10.1038/nature13952

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., et al. (1996). Laser capture microdissection. *Science* 274, 998–1001. doi: 10.1126/science.274.5289.998

Emrich, C. A., Medintz, I. L., Chu, W. K., and Mathies, R. A. (2007). Microfabricated two-dimensional electrophoresis device for differential protein expression profiling. *Anal. Chem.* 79, 7360–7366. doi: 10.1021/ac0711485

Espina, V., Heiby, M., Pierobon, M., and Liotta, L. A. (2007). Laser capture microdissection technology. *Expert Rev. Mol. Diagn.* 7, 647–657. doi: 10.1586/14737159.7.5.647

Esposito, G. (2007). Complementary techniques: laser capture microdissection–increasing specificity of gene expression profiling of cancer specimens. *Adv. Exp. Med. Biol.* 593, 54–65. doi: 10.1007/978-0-387-39978-2_6

Fend, F., and Raffeld, M. (2000). Laser capture microdissection in pathology. *J. Clin. Pathol.* 53, 666–672. doi: 10.1136/jcp.53.9.666

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., and Koltay, P. (2015). Technologies for single-cell isolation. *Int. J. Mol. Sci.* 16, 16897–16919. doi: 10.3390/ijms160816897

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966

Grützkau, A., and Radbruch, A. (2010). Small but mighty: how the MACS-technology based on nanosized superparamagnetic particles has helped to analyze the immune system within the last 20 years. *Cytometry A* 77, 643–647. doi: 10.1002/cyto.a.20918

Haselgrübler, T., Haider, M., Ji, B., Juhasz, K., Sonnleitner, A., Balogi, Z., et al. (2014). High-throughput, multiparameter analysis of single cells. *Anal. Bioanal. Chem.* 406, 3279–3296. doi: 10.1007/s00216-013-7485-x

Hashimoto, M., Barany, F., Xu, F., and Soper, S. A. (2007). Serial processing of biological reactions using flow-through microfluidic devices: coupled PCR/LDR for the detection of low-abundant DNA point mutations. *Analyst* 132, 913–921. doi: 10.1039/b700071e

Hebenstreit, D. (2012). Methods, challenges and potentials of single cell RNA-seq. *Biology (Basel).* 1, 658–667. doi: 10.3390/biology1030658

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* 3:31. doi: 10.3389/neuro.09.031.2009

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., et al. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 148, 873–885. doi: 10.1016/j.cell.2012.02.028

Hubert, R., Weber, J. L., Schmitt, K., Zhang, L., and Arnheim, N. (1992). A new source of polymorphic DNA markers for sperm typing: analysis of microsatellite repeats in single cells. *Am. J. Hum. Genet.* 51, 985–991.

Hughes, A. E., Magrini, V., Demeter, R., Miller, C. A., Fulton, R., Fulton, L. L., et al. (2014). Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. *PLoS Genet.* 10:e1004462. doi: 10.1371/journal.pgen.1004462

Irish, J. M., Hovland, R., Krutzik, P. O., Perez, O. D., Bruserud, O., Gjertsen, B. T., et al. (2004). Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* 118, 217–228. doi: 10.1016/j.cell.2004.06.028

Kalisky, T., Blainey, P., and Quake, S. R. (2011). Genomic analysis at the single-cell level. *Annu. Rev. Genet.* 45, 431–445. doi: 10.1146/annurev-genet-102209-163607

Kim, K. T., Lee, H. W., Lee, H. O., Kim, S. C., Seo, Y. J., Chung, W., et al. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* 16:127. doi: 10.1186/s13059-015-0692-3

Klein, C. A., Schmidt-Kittler, O., Schardt, J. A., Pantel, K., Speicher, M. R., and Riethmuller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4494–4499. doi: 10.1073/pnas.96.8.4494

Kummari, E., Guo-Ross, S. X., and Eells, J. B. (2015). Laser capture microdissection–a demonstration of the isolation of individual dopamine neurons and the entire ventral tegmental area. *J. Vis. Exp.* 96:e52336. doi: 10.3791/52336

Langmore, J. P. (2002). Rubicon genomics, Inc. *Pharmacogenomics* 3, 557–560. doi: 10.1517/14622416.3.4.557

Lecault, V., White, A. K., Singhal, A., and Hansen, C. L. (2012). Microfluidic single cell analysis: from promise to practice. *Curr. Opin. Chem. Biol.* 16, 381–390. doi: 10.1016/j.cbpa.2012.03.022

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343, 1360–1363. doi: 10.1126/science.1250212

Li, P., Gao, Y., and Pappas, D. (2012a). Multiparameter cell affinity chromatography: separation and analysis in a single microfluidic channel. *Anal. Chem.* 84, 8140–8148. doi: 10.1021/ac302002a

Li, X., Lewis, M. T., Huang, J., Gutierrez, C., Osborne, C. K., Wu, M. F., et al. (2008). Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J. Natl. Cancer Inst.* 100, 672–679. doi: 10.1093/jnci/djn123

Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., et al. (2012b). Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience* 1:12. doi: 10.1186/2047-217X-1-12

Lindström, S., and Andersson-Svahn, H. (2010). Overview of single-cell analyses: microdevices and applications. *Lab Chip* 10, 3363–3372. doi: 10.1039/c0lc00150c

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364. doi: 10.1155/2012/251364

Liu, N., Liu, L., and Pan, X. (2014). Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell. Mol. Life Sci.* 71, 2707–2715. doi: 10.1007/s00018-014-1601-8

Liu, P., Seo, T. S., Beyor, N., Shin, K. J., Scherer, J. R., and Mathies, R. A. (2007). Integrated portable polymerase chain reaction-capillary electrophoresis microsystem for rapid forensic short tandem repeat typing. *Anal. Chem.* 79, 1881–1889. doi: 10.1021/ac061961k

Llorens-Bobadilla, E., Zhao, S., Baser, A., Saiz-Castro, G., Zwadlo, K., and Martin-Villalba, A. (2015). Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell* 17, 329–340. doi: 10.1016/j.stem.2015.07.002

Lovatt, D., Ruble, B. K., Lee, J., Dueck, H., Kim, T. K., Fisher, S., et al. (2014). Transcriptome *in vivo* analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* 11, 190–196. doi: 10.1038/nmeth.2804

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A. R., et al. (2012). Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338, 1627–1630. doi: 10.1126/science.1229112

Marcy, Y., Ishoey, T., Lasken, R. S., Stockwell, T. B., Walenz, B. P., Halpern, A. L., et al. (2007). Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.* 3:e155. doi: 10.1371/journal.pgen.0030155

Mellors, J. S., Jorabchi, K., Smith, L. M., and Ramsey, J. M. (2010). Integrated microfluidic device for automated single cell analysis using electrophoretic separation and electrospray ionization mass spectrometry. *Anal. Chem.* 82, 967–973. doi: 10.1021/ac902218y

Miltenyi, S., Müller, W., Weichel, W., and Radbruch, A. (1990). High gradient magnetic cell separation with MACS. *Cytometry* 11, 231–238. doi: 10.1002/cyto.990110203

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Nagrath, S., Sequist, L. V., Maheswaran, S., Bell, D. W., Irimia, D., Ulkus, L., et al. (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* 450, 1235–1239. doi: 10.1038/nature06385

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., et al. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature* 520, 675–678. doi: 10.1038/nature14366

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94. doi: 10.1038/nature09807

Paguirigan, A. L., Smith, J., Meshinchi, S., Carroll, M., Maley, C., and Radich, J. P. (2015). Single-cell genotyping demonstrates complex clonal

diversity in acute myeloid leukemia. *Sci. Transl. Med.* 7:281re282. doi: 10.1126/scitranslmed.aaa0763

Pan, X. (2014). Single cell analysis: from technology to biology and medicine. *Single Cell Biol.* 3:106. doi: 10.4172/2168-9431.1000106

Pan, X., Durrett, R. E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., et al. (2013). Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc. Natl. Acad. Sci. U.S.A.* 110, 594–599. doi: 10.1073/pnas.1217322109

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S.M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257

Perez, O. D., and Nolan, G. P. (2002). Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat. Biotechnol.* 20, 155–162. doi: 10.1038/nbt0202-155

Perfetto, S. P., Chattopadhyay, P. K., and Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nat. Rev. Immunol.* 4, 648–655. doi: 10.1038/nri1416

Powell, A. A., Talasaz, A. H., Zhang, H., Coram, M. A., Reddy, A., Deng, G., et al. (2012). Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. *PLoS ONE* 7:e33788. doi: 10.1371/journal.pone.0033788

Ren, K., Chen, Y., and Wu, H. (2014). New materials for microfluidics in biology. *Curr. Opin. Biotechnol.* 25, 78–85. doi: 10.1016/j.copbio.2013.09.004

Ren, K., Zhou, J., and Wu, H. (2013). Materials for microfluidic chip fabrication. *Acc. Chem. Res.* 46, 2396–2406. doi: 10.1021/ar300314s

Riethdorf, S., Fritsche, H., Muller, V., Rau, T., Schindlbeck, C., Rack, B., et al. (2007). Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clin. Cancer Res.* 13, 920–928. doi: 10.1158/1078-0432.CCR-06-1695

Ruiz, C., Lenkiewicz, E., Evers, L., Holley, T., Robeson, A., Kiefer, J., et al. (2011). Advancing a clinically relevant perspective of the clonal nature of cancer. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12054–12059. doi: 10.1073/pnas.1104009108

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529. doi: 10.1126/science.1105809

Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11, 22–24. doi: 10.1038/nmeth.2764

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240. doi: 10.1093/hmg/ddq416

Schatz, D. G., and Swanson, P. C. (2011). V(D)J recombination: mechanisms of initiation. *Annu. Rev. Genet.* 45, 167–202. doi: 10.1146/annurev-genet-110410-132552

Schor, S. L., and Schor, A. M. (2001). Phenotypic and genetic alterations in mammary stroma: implications for tumour progression. *Breast Cancer Res.* 3, 373–379. doi: 10.1186/bcr325

Schulz, K. R., Danna, E. A., Krutzik, P. O., and Nolan, G. P. (2012). Single-cell phospho-protein analysis by flow cytometry. *Curr. Protoc. Immunol.* Chapter 8: Unit 8. 17, 11–20. doi: 10.1002/0471142735.im0817s96

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., et al. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 141, 69–80. doi: 10.1016/j.cell.2010.02.027

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., et al. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* 1, 1965–1970. doi: 10.1038/nprot.2006.326

Squires, T. M., and Quake, S. R. (2005). Microfluidics: fluid physics at the nanoliter scale. *Rev. Mod. Phys.* 77, 977–1026. doi: 10.1103/RevModPhys.77.977

Srivastava, N., Brennan, J. S., Renzi, R. F., Wu, M., Branda, S. S., Singh, A. K., et al. (2009). Fully integrated microfluidic platform enabling automated phosphoprofiling of macrophage response. *Anal. Chem.* 81, 3261–3269. doi: 10.1021/ac8024224

Su, X., Kirkwood, S. E., Gupta, M., Marquez-Curtis, L., Qiu, Y., Janowska-Wieczorek, A., et al. (2011). Microscope-based label-free microfluidic cytometry. *Opt. Express* 19, 387–398. doi: 10.1364/OE.19.000387

Talasaz, A. H., Powell, A. A., Huber, D. E., Berbee, J. G., Roh, K. H., Yu, W., et al. (2009). Isolating highly enriched populations of circulating epithelial cells and other rare cells from blood using a magnetic sweeper device. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3970–3975. doi: 10.1073/pnas.0813188106

Taniguchi, K., Kajiyama, T., and Kambara, H. (2009). Quantitative analysis of gene expression in a single cell by qPCR. *Nat. Methods* 6, 503–506. doi: 10.1038/nmeth.1338

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 13, 718–725. doi: 10.1016/0888-7543(92)90147-K

Trumpp, A., and Wiestler, O. D. (2008). Mechanisms of disease: cancer stem cells–targeting the evil twin. *Nat. Clin. Pract. Oncol.* 5, 337–347. doi: 10.1038/ncponc1110

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153. doi: 10.1038/nn.3881

VanDijken, J., Kaigala, G. V., Lauzon, J., Atrazhev, A., Adamia, S., Taylor, B. J., et al. (2007). Microfluidic chips for detecting the t(4;14) translocation and monitoring disease during treatment using reverse transcriptase-polymerase chain reaction analysis of IgH-MMSET hybrid transcripts. *J. Mol. Diagn.* 9, 358–367. doi: 10.2353/jmoldx.2007.060149

Van Loo, P., and Voet, T. (2014). Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.* 24, 82–91. doi: 10.1016/j.gde.2013.12.004

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512, 155–160. doi: 10.1038/nature13600

Welzel, G., Seitz, D., and Schuster, S. (2015). Magnetic-activated cell sorting (MACS) can be used as a large-scale method for establishing zebrafish neuronal cell cultures. *Sci. Rep.* 5:7959. doi: 10.1038/srep07959

Whitesides, G. M. (2006). The origins and the future of microfluidics. *Nature* 442, 368–373. doi: 10.1038/nature05058

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46. doi: 10.1038/nmeth.2694

Wu, M., and Singh, A. K. (2012). Single-cell protein analysis. *Curr. Opin. Biotechnol.* 23, 83–88. doi: 10.1016/j.copbio.2011.11.023

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148, 886–895. doi: 10.1016/j.cell.2012.02.025

Yu, C., Yu, J., Yao, X., Wu, W. K., Lu, Y., Tang, S., et al. (2014). Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* 24, 701–712. doi: 10.1038/cr.2014.43

Zhang, X., Marjani, S. L., Hu, Z., Weissman, S. M., Pan, X., and Wu, S. (2016). Single-Cell sequencing for precise cancer research: progress and prospects. *Cancer Res.* 76, 1305–1312. doi: 10.1158/0008-5472.CAN-15-1907

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622–1626. doi: 10.1126/science.1229164

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer ASB and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Cell and Developmental Biology

REVIEW
published: 20 April 2018
doi: 10.3389/fcell.2018.00028

Single Cell Multi-Omics Technology: Methodology and Application

Youjin Hu[1], Qin An[2], Katherine Sheu[2], Brandon Trejo[2], Shuxin Fan[1] and Ying Guo[3]**

[1] Zhongshan Ophthalmic Center, State Key Laboratory of Ophthalmology, Sun-Ye-Sat University, Guangzhou, China, [2] Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA, United States, [3] The Second Affiliated Hospital, Xiangya School of Medicine, Central South University, Changsha, China

In the era of precision medicine, multi-omics approaches enable the integration of data from diverse omics platforms, providing multi-faceted insight into the interrelation of these omics layers on disease processes. Single cell sequencing technology can dissect the genotypic and phenotypic heterogeneity of bulk tissue and promises to deepen our understanding of the underlying mechanisms governing both health and disease. Through modification and combination of single cell assays available for transcriptome, genome, epigenome, and proteome profiling, single cell multi-omics approaches have been developed to simultaneously and comprehensively study not only the unique genotypic and phenotypic characteristics of single cells, but also the combined regulatory mechanisms evident only at single cell resolution. In this review, we summarize the state-of-the-art single cell multi-omics methods and discuss their applications, challenges, and future directions.

Keywords: single cell transcriptome, single cell multi-omics profiling, single cell epigenome, single cell proteome, gene regulation, epigenetics

OPEN ACCESS

Edited by:
Xinghua Victor Pan,
Yale University, United States

Reviewed by:
Zhibin Wang,
Johns Hopkins University, United States
Leonard C. Edelstein,
Thomas Jefferson University, United States
Stephen Clark,
Babraham Institute (BBSRC), United Kingdom

****Correspondence:***
Youjin Hu
huyoujin@gzzoc.com
Ying Guo
mytyl.g@hotmail.com

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *15 December 2017*
Accepted: *08 March 2018*
Published: *20 April 2018*

Citation:
Hu Y, An Q, Sheu K, Trejo B, Fan S and Guo Y (2018) Single Cell Multi-Omics Technology: Methodology and Application. Front. Cell Dev. Biol. 6:28. doi: 10.3389/fcell.2018.00028

INTRODUCTION

According to the central dogma, also known as the DNA-RNA-protein axis, DNA provides the code for RNA, which is translated to produce proteins that fulfill biological functions (Crick, 1970). To discover the regulatory mechanisms behind RNA transcription and protein translation, the most straightforward approach is to analyze both DNA and RNA, or both RNA and protein, from the same sample. Despite the complexity of tissues comprised of heterogeneous cell populations, such as cancer, most experimental results to date have been based on analysis of bulk samples, which theoretically read an averaged signal from the population and prevent resolution of cellular variation (Navin et al., 2011; Huang et al., 2015; Gawad et al., 2016). To decipher the mechanism of heterogeneous gene transcriptional regulation, integrated measurement and co-analysis of multiple types of molecules, such as DNA, RNA, and protein, at single cell level is required.

The invention of PCR methods in 1983 made it possible to analyze the picogram amounts of DNA in single cells, although these initial methods could only amplify small, targeted regions of the genome. However, the development of whole genome amplification (WGA) and whole transcriptome amplification (WTA) methods (Tang et al., 2009; Zong et al., 2012; Huang et al., 2015; Wang and Navin, 2015; Gawad et al., 2016) soon allowed quantitative measurement of DNA and RNA for multiple genes in single cells. At the same time, the development of next generation sequencing technology has enabled genome-wide analysis of DNA and RNA in single cells. Inspired by the very first report of single cell DNA sequencing

and single cell RNA sequencing, scientists have developed numerous methods to measure other omics at single cell level, including single cell DNA methylation, single cell chromatin sequencing and single cell proteome analysis [**Figure 1,** A detailed introduction of single cell sequencing methods has been reviewed elsewhere (Wang and Navin, 2015; Gawad et al., 2016)].

Single cell genome-wide approaches provide a valuable opportunity to measure different molecules, such as DNA, RNA, protein, and chromatin with ultimate resolution. By isolating multiple types of molecules (DNA, RNA, or protein) from a single cell simultaneously, it is feasible to profile different types of molecules in parallel. For example, genomic DNA can be used to assay the single cell genome, methylome or chromatin accessibility, while RNA from the same cell can be used to profile the transcriptome, and protein the proteome. Utilizing these different single cell omics profiling strategies as building blocks, we can construct a multi-omics profile for the same cell. Here, we summarize current single cell multi-omics approaches, such as scG&T-seq (single cell Genome & Transcriptome sequencing), scMT-seq (single cell Methylome and Transcriptome sequencing), scM&T-seq (single cell Methylome & Transcriptome sequencing), scTrio-seq (single-cell triple omics sequencing), and scCOOL-seq (single cell Chromatin Overall Omic-scale Landscape Sequencing) (MacAulay et al., 2015; Angermueller et al., 2016; Hou et al., 2016; Hu et al., 2016), with each of them measuring a different combination of omics data (**Figure 2**). We also review the bioinformatics advances that have been necessary to understand the large amounts of multi-dimensional data arising from single cell multi-omics profiling, and we examine the potential for this technology to elucidate numerous biological enigmas.

METHODS FOR ISOLATING MULTIPLE TYPES OF MOLECULES FROM A SINGLE CELL

Isolating multiple types of molecules from a single cell is the starting point for single cell multi-omics measurement, and generally can be divided into two steps.

The first step is to collect a single cell randomly from a population with heterogeneity. The standard protocol is to get viable, intact cells by mechanical or enzymatic dissociation and then capture single cells from the dissociated cell suspension. Several approaches can be used, including mouth pipetting, serial dilution, robotic micromanipulation, flow-assisted cell sorting (FACS), and microfluidic platforms (Wang and Navin, 2015). Although these collection approaches are borrowed from methods developed for single cell mono-omics sequencing, additional considerations must be taken for multi-omics to ensure that multiple types of molecules can be viably measured in the same cell. The success of this first collection step is critical for preserving an accurate representation of the DNA, RNA, and protein within the cell for downstream measurements. The method used for the initial dissociation of tissues into single cells—mechanical or enzymatic—needs to be selected with consideration for both the nature of the starting material and the types of sequencing to be performed. Clinical samples such as solid tumors are often obtained flash frozen or embedded in paraffin (FFPE), making multi-omics measurements that include cytoplasmic RNA or protein more challenging. However, because this type of freezing process perturbs the cytoplasmic membrane while keeping the nuclear membrane intact, multi-omics measurements that involve the genome, epigenome, and chromatin-associated RNA are still possible after creation of nuclear suspensions (Navin, 2015). For fresh tissues, choice of mechanical or enzymatic dissociation reflects the need for both cell integrity and dissociation quality. Prolonged exposure to common dissociation enzymes such as papain, collagenase, dispase, and neutral protease can result in degradation of RNA and proteins, or generation of cell debris that aberrantly activate cell signaling pathways and cell surface proteins (Autengruber et al., 2012; Volovitz et al., 2016). Mechanical mincing of the starting material through trituration or nanofiltration may also disrupt accurate representation of the proteome or transcriptome in cells that contain long projections such as neurons. These pitfalls in turn can complicate the subsequent computational analyses performed on the data, which often involve identification of correlative relationships among the different layers of multi-omics data obtained. Thus, both tissue-specific and measurement-specific aspects of obtaining multi-omics measurements need to be considered in order to achieve optimized single cell suspensions.

Next, the technique used to select single cells after separation of bulk tissues also has an impact on the feasibility of combinatorial multi-omics measurements. The advantages of techniques such as mouth pipetting and serial dilution include the simplicity and rapidity of moving single cells from the cell suspension to individual reaction chambers. This helps limit the degradation of more volatile molecules such as RNA or protein and may reduce the possibility of non-physiologic changes in chromatin accessibility and chromatin conformation (Wang and Navin, 2015; Svensson et al., 2017). Robotic manipulation, FACS, and microfluidic capture platforms have the advantage of the ability to sort through subpopulations by cell labeling, but require more extensive manipulation of single cells using expensive equipment (Ortega et al., 2017). Of the numerous options, selection of a protocol for isolating single cells for multi-omics data collection will ultimately depend on the molecules that need to be preserved, the type of tissue obtained, and the cost.

The second step is to isolate multiple types of molecules from the same cell, for which there are four main strategies: To isolate DNA and RNA of a single cell, the first strategy is physical separation, including separation of nucleus from cytosol, as genomic DNA is contained in the nucleus and the majority of mRNAs are located in the cytosol. Single cells are treated with a membrane-selective lysis buffer, through which the cell membrane is broken down while the nucleus is kept intact. Then, single nuclei are separated from cytoplasm by micropipetting, centrifugation, or antibody-conjugated magnetic microbeads (Hou et al., 2016; Hu et al., 2016; Han et al., 2018; **Table 1**). This method has been demonstrated to be highly efficient by several research groups, including our lab. Our data indicates that profiling of cytosolic RNA can resemble the transcriptome of the

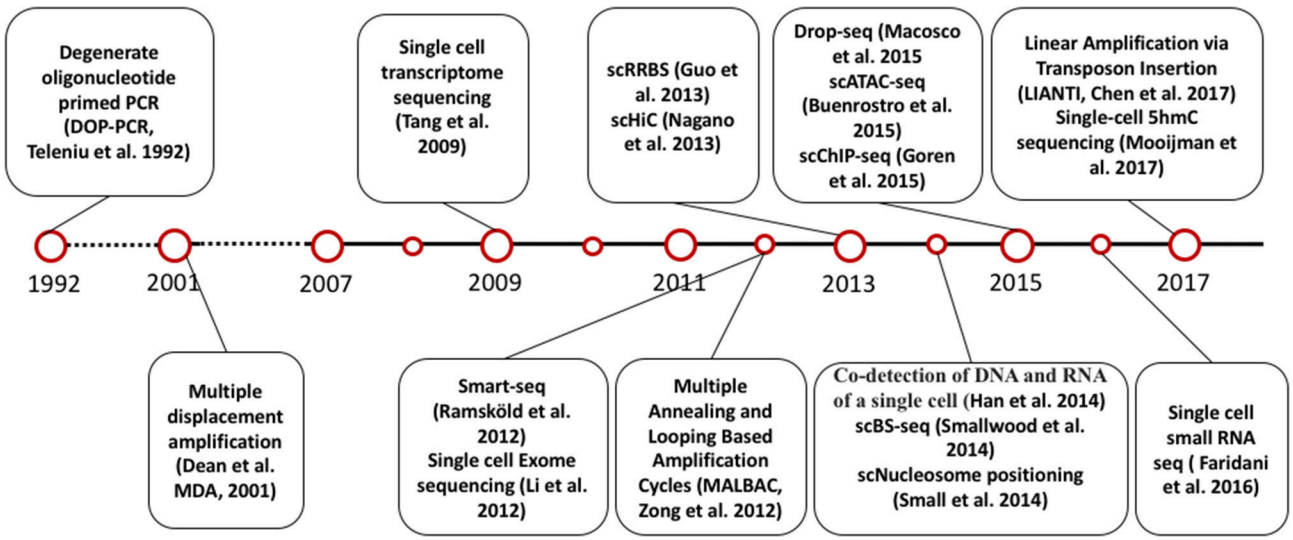


FIGURE 1 | Timeline of single cell sequencing methods milestones.

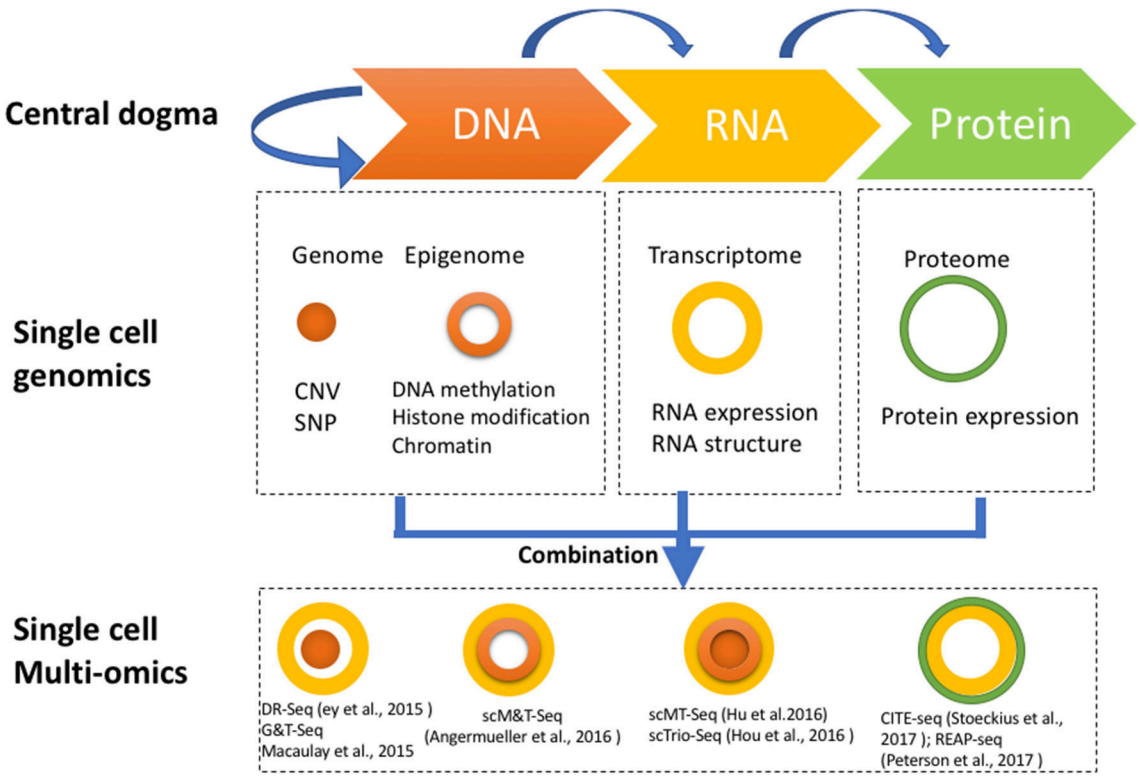


FIGURE 2 | Strategies for multi-omics profiling of single cells. Three major types of molecules relating to biological central dogma **(Top)**. Single cell genomics methods profiling the genome, epigenome, transcriptome, and proteome are shown by different shapes with variable colors **(Middle)**. Single cell multi-omics methods are built by combining different single cell sequencing methods to simultaneously profile multiple types of molecules of a single cell genome wide **(Bottom)**. For example, G&T-seq was built by combining genome (orange) and transcriptome (yellow) to simultaneously detect DNA and RNA of the same cell genome wide.

whole cell. However, this method is low throughput (Hu et al., 2016), as the nucleus-picking procedure is manual and cannot be automated easily. Methods based on centrifugation (Hou et al., 2016) or antibody conjugated magnetic microbeads (Han et al., 2018) can achieve relatively higher throughput in isolating DNA and RNA from single cells.

TABLE 1 | Current multi-omics methods.

Methods (time)	Cell type	Measurement	Approach	Single cell isolation	Major discovery	References
DR-seq (single cell gDNA and mRNA sequencing) (2015)	Mouse embryonic stem cell line (E14) and breast cancer cell line (SK-BR-3).	Genome, mRNA transcriptome	Amplify gDNA and synthesize cDNA without physically separating the nucleic acids. Product is then split for scWGS (single cell Whole Genome Sequencing) and scRNA-seq.	Mouth pipet	Genome copy number variation could drive transcriptome variability.	Dey et al., 2015
G&T-seq (single cell genome & Transcriptome sequencing) (2016)	HCC38, HCC38-BL and iPSCs carrying trisomy 21.	Genome, mRNA transcriptome	Cell is lysed, genomic DNA and poly(A)+ mRNA is separated by magnetic beads for scWGS and scRNA-seq.	Flow cytometry	Identified transcriptional consequences of chromosomal aneuploidies and inter-chromosomal fusions.	MacAulay et al., 2015
scMT-seq (single cell Methylome and Transcriptome sequencing) (2016)	Mouse dorsal root ganglion neurons.	DNA methylome, mRNA transcriptome	Cell is lysed, cell nucleus containing genomic DNA and cell lysis poly(A)+ mRNA is separated for scRRBS (single cell Reduced Representation Bisulfite Sequencing) and scRNA-seq.	Mouth pipet	Methylation of non-CGI promoters is better anti-correlated with gene transcription, gene body methylation of CGI promoter genes has higher correlation with transcription, potentially reveal allelic specific methylation and allelic expression in single cells.	Hu et al., 2016
scM&T-seq (single cell Methylome& Transcriptome sequencing) (2016)	Mouse embryonic stem cell line (E14), in serum and 2i conditions.	DNA methylome, mRNA transcriptome	Cell is lysed, genomic DNA and poly(A)+ mRNA is separated by magnetic beads for scWGBS (single cell Whole Genome Bisulfite Sequencing) and scRNA-seq (single cell RNA sequencing).	Flow cytometry	Non-CGI promoter methylation and transcription in single cell is negatively correlated; methylation and transcription can be both positively and negatively correlated in distal regulatory regions.	Angermueller et al., 2016
sc-GEM (genotype single cells genotype, gene expression, DNA methylation) (2016)	Human fibroblast, hIPSC, hESC and NSCLC sample.	Genotype single cells while simultaneously interrogating gene expression and DNA methylation at multiple loci	Cell is captured and lysed on C1 Fluidigm chip. RNA is measured using single cell RT-qPCR and methylation is measured using Single Cell Restriction Analysis of Methylation.	Microfluidic device	Tight coupling between the timing of DNA methylation changes and transcription in individual cells; cells have EGFR mutations show a distinct epigenetic signature.	Cheow et al., 2016
scTrio-seq (single-cell triple omics sequencing) (2017)	HepG2 cell line, mESCs and hepatocellular carcinoma tissue sample.	DNA methylome, CNV (copy number variation), mRNA transcriptome	Cell is lysed. DNA and RNA are separated using centrifuge. mRNA is measured using scRNA-seq, methylation is measured using scRRBS. CNV is computationally inferred form scRRBS coverage.	Mouth pipet	Detected subpopulations of cancer cells according to the large-scale CNV, and detected relationships between CNV, methylation and transcription.	Hou et al., 2016
Simultaneous multiplexed measurement of RNA and proteins in single cells (2016)	Cell culture from glioblastoma multiforme patient sample.	Protein and multiple mRNA	Cells are sorted and lysed. Protein is detected by homogeneous affinity-based proximity extension assay (PEA), and RNA is measured by microfluidic qPCR.	Cell sorting and microfluidic devices	RNA and protein data provide complementary information in defining cell states, and significant heterogeneity in cell culture derived from a glioblastoma multiform patient.	Darmanis et al., 2016

(Continued)

TABLE 1 | Continued

Methods (time)	Cell type	Measurement	Approach	Single cell isolation	Major discovery	References
scCOOL-seq (single cell Chromatin Overall Omic-scale Landscape Sequencing) (2017)	Mouse preimplantation embryos at different developmental stages.	Chromatin state, DNA methylation, and CNV	Cell is lysed. Chromatin is treated by GpC methyltransferase, and treated DNA is sequenced by scWGBS. *In vivo* methylation is detected as CpG methylation, and chromatin accessibility is computationally inferred by GpC methylation level.	Mouth pipet	DNA methylation is different between paternal and maternal alleles, but their chromatin accessibility states are similar.	Guo et al., 2017
CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing) (2017)	Cord blood mononuclear cells.	Protein and mRNA transcriptome	mRNA is sequenced using 10X genomics platform. Protein is detected by oligo-labeled antibody, which can be read out during sequencing.	Compatible with 10X genomics, adaptable to other platforms	Multimodal data enable to reveal phenotypes that could not be discovered by using scRNA-seq alone.	Stoeckius et al., 2017
REAP-seq (RNA expression and protein sequencing assay)	human lymphocytes	Protein and mRNA transcriptome	mRNA is sequenced using 10X genomics platform. Protein is detected by oligo-labeled antibody, which can be read out during sequencing.	Flow cytometry	assess the costimulatory effects of a CD27 agonist on human CD8+ lymphocytes and to identify and characterize an unknown cell type	Peterson et al., 2017
scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) (2018)	Mouse embryonic stem cells	Nucleosome status, DNA methylation and mRNA transcription	Similar with scM&T methods, DNA and mRNA were isolated. DNA was cut with GpC methyltransferase M.CviPI before bisulfite treatment.	FACS	Novel links between all three molecular layers and revealing dynamic coupling between epigenomic layers during differentiation	Clark et al., 2018
SIDR-seq simultaneous isolation of genomic DNA and total RNA (SIDR) and sequencing. (2018)	Human lung cancer and breast cancer cells, MCF7, HCC827, and SKBR3 cell lines.	Genome, mRNA transcriptome	Nucleus and cytosol of a single cell were separated by antibody-conjugated magnetic microbeads. mRNA is measured using smart-seq2, gDNA is measured using ingle-cell whole-genome amplification (Repli-g single cell kit)	Manually diluted to 48-well	copy-number variations positively correlated with the corresponding gene expression levels	Han et al., 2018

The second strategy uses oligo-dT primer coated magnetic beads to bind and separate polyadenylated mRNA from DNA (MacAulay et al., 2015; Angermueller et al., 2016). Genome wide sequencing of single cell DNA and RNA purified by this method indicated that breadth of genome coverage and number of genes were not affected by the process of separation, indicating high efficiency in the recovery of DNA and RNA. Since this strategy is adaptable to liquid-handling robots or automated work stations, higher throughput can be achieved. However, coverage of isolated DNA was less evenly distributed across the genome compared to that of the whole single cell sequencing, which may result in less accuracy for copy number analysis of certain genomic regions at a suboptimized sequencing depth.

Besides direct physical isolation of DNA and RNA at the beginning, the third strategy is to preamplify DNA and RNA simultaneously, followed by separation into two parts (Dey et al., 2015). Whole transcriptome sequencing of preamplified RNA of one part showed a similar number of genes covered compared to that of whole single cells. However, as the amplified DNA does not retain methylation states, this method is not suitable for methylome analysis.

The fourth strategy is to split the material of a single cell into two parts directly. For example, a recent report used the splitting strategy to split a single cell into two parts and simultaneously analyze the RNA and protein of the same cell (Darmanis et al., 2016). This splitting strategy is not an ideal method to isolate substrates such as DNA because some material will inevitably be lost due to the uneven split. However, for RNA and protein molecules with high copy number in the single cells, this method is feasible as long as the split is even between the two parts.

INTEGRATION OF GENOME AND TRANSCRIPTOME

The first single cell transcriptome analysis was reported in 2009 (Tang et al., 2009), and many additional single cell RNA sequencing methods have been developed since, such as Quartz-seq (Sasagawa et al., 2013), smart-seq (Switching mechanism at 5′ end of the RNA transcript) (Goetz and Trimarchi, 2012; Picelli et al., 2014), Cel-seq (Cell expression by linear amplification and sequencing) (Hashimshony et al., 2012) etc., which were developed using different strategies for different purposes. For example, Quartz-seq detects the 3′ end of transcripts, while Smart-seq detects full length transcripts. Cel-seq barcodes and pools samples before linearly amplifying mRNA to multiplex single cell samples. In parallel, due to the development of single-cell whole-genome amplification (WGA) methods, single cell genome sequencing technologies have also been established. At present, four major WGA methods have been reported: DOP (degenerate oligonucleotide-primed polymerase chain reaction) (Telenius et al., 1992), MDA (Multiple Displacement Amplification) (Dean et al., 2001), MALBAC (Multiple Annealing and Looping Based Amplification Cycles) (Zong et al., 2012) and PicoPLEX (Rubicon Genomics PicoPLEX Kit). In 2013, Han et al. first reported a co-detection of DNA and RNA from the same single cell (Han et al., 2014), which was achieved by physical isolation of cytoplasm (containing cytoplasm RNAs) from nucleus (containing the intact genome) from the same single cells, followed by separate amplification of the transcriptome and genome, and further by respective sequencing of both. Although the initial report showed only the data of the whole transcriptome but not the whole genome, instead of Sanger sequencing of a selected set of genomic sequences, it paved a way to establish multi-omic profiling methods. Later, experimental protocols that simultaneously sequenced the genome and transcriptome were developed by elegantly integrating existing single cell sequencing methods, namely DR-seq (gDNA and mRNA sequencing) (Dey et al., 2015) and G&T-seq (Genome & Transcriptome sequencing) (MacAulay et al., 2015). In DR-seq, a cell is lysed completely, releasing its DNA and RNA into the same reaction system. Genomic DNA and cDNA initially being amplified at the same time is split into two halves: one for RNA-seq using the CEL-seq protocol, and the other half for genome sequencing using MALBAC (Dey et al., 2015). Different from DR-seq, G&T-seq separated poly-A tailed mRNAs from DNA by using oligo-dT-coated magnetic beads. Separated mRNA and DNA were then sequenced using SMART-seq2 and various WGA protocols (MDA or PicoPLEX), respectively (MacAulay et al., 2015). Most recently, Han et al. reported a novel method for simultaneous isolation of genomic DNA and total RNA (SIDR) from single cells by using hypotonic lysis to preserve nuclear lamina integrity and subsequently capturing the cell lysate using antibody-conjugated magnetic microbeads. They found that copy-number variations positively correlated with the corresponding gene expression levels (Han et al., 2018). In summary, using DR-seq, G&T-seq and SIDR, researchers were able to directly determine the correlation between large-scale copy number variation and transcription levels in the CNV regions.

As discussed previously by MacAulay et al. (2017), a substantial advantage of direct measurement of multiple molecular types from the same single cell over separate measurement of each type of molecule from different cells is that genotype-phenotype correlation can be determined unambiguously. First, the genomic variation can be directly linked to the transcriptional variation without being confounded by cell heterogeneity, enabling the dissection of potential molecular mechanisms underlying variable phenotypes among single cells. Second, coupled with lineage record technology, simultaneous sequencing of the genome and transcriptome can be used for reconstruction of lineage trees. Genomic profiling of single cells can divulge the lineage relationship among single cells, based on inherited mutations. The transcriptome profiling of the same single cells can in parallel provide information about the cell's phenotype and function. One intriguing application of this method is to dissect the mechanism of heterogeneity of tumor cells to inform our knowledge of tumor formation and potential therapeutic targets (Shapiro et al., 2013). Third, simultaneous sequencing of DNA and RNA of the same cell can detect DNA mutations with higher accuracy, as the mutations found in DNA or RNA can be verified by each other. This strategy can be

very helpful in situations where highly accurate mutation calling from a single cell is required, such as genetic diagnosis screening during *in vitro* fertilization, when only 1–2 single blastomeres are available (Vermeesch et al., 2016). Of note, post-transcriptional modification such as RNA editing (Tan et al., 2017) which may affect the concordance of variations in both DNA and RNA, should be taken into consideration to precisely call the mutations.

INTEGRATION OF EPIGENOME WITH TRANSCRIPTOME

Based on the development of technologies for single cell epigenome and transcriptome profiling, the methods for the integrated analysis of the epigenome and transcriptome were developed (Angermueller et al., 2016; Hou et al., 2016; Hu et al., 2016). DNA methylation has been demonstrated to have key regulatory functions on gene expression in many biological process, so the relationship between the DNA methylome and transcriptome from the same single cell is of great interest. Two major methods for single cell methylome analysis are single cell reduced representative bisulfite sequencing (scRRBS) (Guo et al., 2013) and single cell whole genome bisulfite sequencing (scWGBS) (Smallwood et al., 2014). The first reported combined DNA methylome and transcriptome profiling method is scM&T-seq (single cell methylome and transcriptome sequencing), which is developed using the procedure of G&T-seq to isolate DNA and RNA from the same single cell. The protocols for mRNA capture, amplification and sequencing are the same as those in G&T-seq. In parallel, the genomic DNA is subjected to bisulfite treatment and sequencing, allowing the simultaneous profiling of the DNA methylome and RNA transcriptome from the same single cell (Angermueller et al., 2016). Subsequently, scMT-seq (Hu et al., 2016) and scTrio-seq (Hou et al., 2016) were reported using a different strategy to isolate DNA and RNA from a single cell, in which cell membrane but not nucleus was selectively lysed to release RNA, and then intact nucleus was physically separated from the cell lysate (Hou et al., 2016; Hu et al., 2016; Guo et al., 2017). In the scMT-seq method, the single cell nucleus is collected by micropipette and subjected to scRRBS, and mRNA in the lysate is amplified by a modified Smart-seq2 protocol. In the scTrio-seq, the nucleus and cytosol are separated by centrifugation, and genomic DNA contained in the nucleus is sequenced by scRRBS while mRNA is amplified by the scRNA-seq protocol reported by Tang et al. (2009).

The simultaneous profiling of methylome and transcriptome of a single cell provides a unique opportunity to directly measure DNA methylation and gene transcription within the same single cell, and to study the correlation of DNA methylation differences with gene transcription variance across single cells. For example, scM&T-seq investigated the relationship between the transcriptome and DNA methylome, and found that low methylated regions (LMR) showed high variance in methylation level, which is consistent with their role as distal regulatory elements that control gene expression (Angermueller et al., 2016). Our results using scMT-seq found that variable CpG sites were significantly enriched at non-CGI (non-CpG island) promoters but depleted at CGI (CpG island) promoters, suggesting that non-CGI promoters could be the major region contributing to methylome heterogeneity among dorsal root ganglion single cells. We also found that transcription level was positively correlated with genebody methylation, but negatively correlated with promoter methylation. In addition, by integrating the genomic SNP information, we found a correlation between allelic gene body methylation and allelic expression at single cell level. Thus, scMT-seq allows us to profile genome, DNA methylome and transcriptome in parallel within a single cell (Hu et al., 2016). Similarly, scTrio-seq enables profiling of DNA methylome, genome (CNV) and transcriptome at the same time, in which the copy number variation is computationally inferred from the scRRBS (Hou et al., 2016). Most recently, Guo et al. from the same group reported another single cell multi-omics sequencing method called single-cell COOL-seq that can profile DNA methylation and chromatin state/nucleosome positioning, copy number variation and ploidy simultaneously from the same cell (Guo et al., 2017). Although they did not incorporate the RNA sequencing in this protocol (which is theoretically possible), this method provided new insights into the comprehensive study of genome-wide gene regulation at single cell level. Most recently, Clark et al. reported the scNMT-seq (single-cell nucleosome, methylation, and transcription sequencing), which can simultaneously profile single cell nucleosome, DNA methylation and transcription. By profiling the mouse embryonic stem cell, they found novel links between all three molecular layers and revealed dynamic coupling between epigenomic layers during differentiation (Clark et al., 2018).

PARALLEL PROFILING OF RNA AND PROTEIN

RNA and protein have distinctive biochemical properties. Compared to genomic sequencing methods, the throughput in terms of the number of proteins that can be detected by the single cell proteome profiling is limited. Until now, a few single cell proteomic methods have been developed based on different strategies, including fluorescence-activated cell sorting (FACS), western blot, metal-tagged antibodies followed by mass cytometry, and oligonucleotide labeled antibodies. Although the multiplexing of these approaches were still limited to tens of proteins for a single cell, they still demonstrated the feasibility of detection of protein and RNA expression, paving a way to discover the dynamics of RNA and protein within the same cell. Darmanis et al. developed a method based on homogeneous affinity-based proximity extension assay that converts protein abundance into tag-oligo levels (Darmanis et al., 2016), and both transcript level and protein level were quantified by qPCR. This method has succeeded in capturing parallel profiles of protein and RNA for up to 96 genes (Darmanis et al., 2016). Another approach to simultaneously detect the RNA and protein of the same cell is PLAYR (proximity ligation assay for RNA). Briefly, the RNA transcripts are bound by and ligated to isotope labeled probes. Transcript levels are converted into isotope label levels

that can be easily measured together with elemental isotope-labeled protein using mass cytometry (Frei et al., 2016). With this method, simultaneous quantification of more than 40 different mRNAs and proteins can be achieved, although improvement is required to achieve genome-wide measurement with higher throughput. Most recently, two methods named REAP-seq and CITE-seq with higher throughput have been reported, in which oligonucleotide-labeled antibodies are used to integrate cellular protein and transcriptome measurements into an efficient, single-cell readout (Peterson et al., 2017; Stoeckius et al., 2017). Quantified proteins with 82 barcoded antibodies and more than 20,000 genes can be detected in a single workflow.

STRATEGIES FOR BIOINFORMATICS ANALYSIS OF SINGLE CELL SEQUENCING DATA

Single cell sequencing technologies for genome wide profiling of DNA and RNA, as well as the subsequent integrative computational analysis methods, are central to the interpretation of single cell multi-omics data. The prelude to this type of analysis hinges first on the development of bioinformatics approaches for single cell single-omics sequencing data for various individual types of molecular measurements. Because technical characteristics of various single cell sequencing protocols are different, the bioinformatics methods involved must also be customized to correctly analyze each data type. The need to address the specific characteristics of different single cell sequencing approaches has inspired many computational methods that allow us to better analyze sequencing datasets involving multiple layers.

Single Cell Genome Sequencing

Two major purposes of single-cell genome sequencing are identifying copy number variation and identifying point mutations/SNPs. Both these questions have been addressed in bulk WGS, and the methods developed for bulk WGS data have provided guidance for single cell WGS analysis.

Copy number variation can be robustly identified using Hidden Markov Model (HMM) or Circular Binary Segmentation (CBS), and these methods have proved effective for scWGS data (Knouse et al., 2016). Although these two methods perform similarly in many situations, user-defined parameter adjustments within the algorithms can affect the sensitivity and specificity of copy number calls. For example, comparison of these two methods on scWGS data with a range of parameters indicated that CBS was more sensitive in calling copy number losses, while HMM was more sensitive in calling gains (Knouse et al., 2016). In the context of single cell CNV analysis, one strategy to reconcile the two approaches has been to take the overlap of CNVs identified by CBS and HMM to increase confidence (Knouse et al., 2016). Considerations in choosing between the methods involve the biological properties of the samples, such as the expected sizes of the CNVs, which could range from whole-arm changes seen in aneuploid tumors to dinucleotide changes observed in inherited polymorphisms or in microsatellite instability. CBS is more flexible than HMM in that the algorithm recursively searches for segmentation points in an unsupervised approach, while HMM depends on the assumption that segmentation points follow a homogenous Poisson process, which is not always the case and may therefore compromise flexibility (Wineinger et al., 2008).

Many tools have been developed for detecting variations in bulk WGS data (Depristo et al., 2011; Koboldt et al., 2012), and these methods, in principle, should perform well in scWGS data. However, scWGS data suffers from high allele coverage bias and high PCR amplification error, which could impair the performance of variant calling methods if not corrected. Recently, with increased understanding of coverage bias in scWGS data (Zhang et al., 2015), Dong et al. reported a computational method that can correct amplification bias to reduce false positive SNPs resulting from PCR or sequencing errors (Dong et al., 2017). Although this new method still partially relies on GATK to identify new variants, it achieved better accuracy by removing false positive variants resulting from PCR error.

Single Cell Transcriptome Sequencing

Single cell RNA-seq data enables the discovery of exciting and new biological phenomena while presenting new challenges for data analysis. For example, single-cell RNA-seq can help us identify cell subtypes with unprecedented resolution, and reconstruct continuous cell lineages. Some early studies showed that identification of cell subtypes or reconstruction of cell lineage could be done manually by experts with sufficient biological prior knowledge using basic statistical methods (Xue et al., 2013; Treutlein et al., 2014). However, recently, huge datasets with extremely heterogeneous cell populations have precluded the feasibility of manual annotation, and many computational pipelines have been developed. For example, tools based on different theoretical frameworks have been developed to cluster cells based on their gene expression similarity, such as SINCERA (Guo et al., 2015), pcaReduce (Žurauskiene and Yau, 2016), SC3 (Kiselev et al., 2017), and SNN-Cliq (Xu and Su, 2015). Additional tools have been developed to reconstruct cell lineage by ordering cells according to computationally inferred pseudo-time (Trapnell et al., 2014; Cannoodt et al., 2016; Qiu et al., 2017). However, despite the availability of myriad computational software packages for clustering and lineage inference, few benchmarking studies have been done to compare their performance.

In addition to those two classical biological questions, the technical problem of imputation of missing values in single-cell RNA-seq data has recently attracted increasing attention. Single-cell RNA-seq, especially for cells captured by droplet-based methods, is often plagued by missing values due to drop-out events, leading to an exceedingly sparse depiction of the single cell transcriptome. Simply removing genes containing missing values restricts the analysis to only highly expressed genes. To overcome this problem, much effort has been made to impute missing values (Kiselev et al., 2017; Lin et al., 2017). These imputation methods can not only enable us to investigate lowly expressed genes but can also improve the performance of

existing computational methods for other purposes by reducing noise from drop-out events.

Single Cell Methylome Analysis

Compared to bulk WGBS (whole genome bisulfite sequencing) data, the analysis of single cell WGBS requires distinct bioinformatics techniques due to the sparse and uneven coverage of scWGBS (single cell WGBS) libraries across the genome. Although many tools have been developed for bulk WGBS data analysis, these methods will fail if applied to scWGBS data directly. To make scGWBS data analysis possible, the first strategy is to merge data from single cells and analyze the merged data as a sample (Farlik et al., 2016). By combining data from many single cells (usually hundreds), the data coverage becomes high, and the bias from allele dropout is averaged out. However, this strategy cannot be used to address the heterogeneity of methylation among different single cells, because methylation data are merged and averaged among the cell population.

Aside from adapting scWGBS data to existing computational pipelines by merging data, the second strategy is to develop new methods specifically for scWGBS data, and many of these methods aim to aggregate methylation levels from adjacent CpG sites or regions with similar biological properties to overcome the sparseness of scWGBS data. For example, Smallwood et al. segment the genome into 5-kbp, non-overlapping bins and use average methylation level among bins as the feature for subsequent analysis (Smallwood et al., 2014). Similarly, by aggregating methylation signal on regulatory elements, we can reveal regulatory mechanisms behind the changes in the DNA methylome (Farlik et al., 2015). In these methods, each single cell is treated as a sample separately, thus enabling the discovery of DNA methylome heterogeneity among single cells.

Interestingly, besides aggregating existing methylation information to reduce noise, a method based on the deep neuronal network was recently developed, which infers missing methylation information from sequencing motifs (Angermueller et al., 2017). Although this method achieved high prediction accuracy for whole genome, its performance on low-methylated regions, the regulatory regions where methylation level influences gene expression greatly, were not satisfying. However, we believe that the prediction accuracy on LMRs can be further improved by incorporating more features into the same deep learning framework.

Single Cell Sequencing for Chromatin Status Analysis

Success in single cell genome and transcriptome sequencing inspired the development of single cell epigenome sequencing. So far, single cell ChIP-seq (Rotem et al., 2015) (Chromatin Immunoprecipitation Sequencing), DNase-seq, and ATAC-seq (Buenrostro et al., 2015) (Assay for Transposase-Accessible Chromatin using sequencing) has been reported from different groups. Since this type of single cell epigenome data has just begun to emerge, the related computational analysis methods are still in their infancy and only a few methods have been developed specifically for single cell data. For example, scChIP-seq and scATAC-seq have been developed to investigate histone modification and chromatin accessibility landscapes at single cell level (Buenrostro et al., 2015; Rotem et al., 2015; Corces et al., 2016), and the reads from one single cell are extremely sparse due to the low amount of DNA in a cell. To identify the regions that have histone modification or regions with open chromatin, reads from several dozen to hundred single cell libraries were pooled together, and only this "pooled library" has enough reads for conventional peak calling methods. In the subsequent analysis, these putative peaks will be used as guidance to aggregate sparse signal and remove background signal. Although this method enables the meaningful analysis of scChIP-seq and scATAC-seq without requirement of any new computational methods, concerns have been raised about the sensitivity of this strategy (Zamanighomi et al., 2017). Interestingly, methods designed for scATAC-seq analysis are emerging, such as chromVAR (Schep et al., 2017) and scABC (Zamanighomi et al., 2017). We believe these pipelines will also inspire the development of effective pipelines for scChIP-seq data.

APPLICATION OF SINGLE CELL MULTI-OMICS METHODS

As described above, single cell multi-omics analysis integrates multiple data sets from the genome, epigenome, transcriptome, proteome, providing a unique chance to uncover novel biological processes. By extending and integrating methods developed for single-omics analysis, we can obtain a multi-channel molecular readout and utilize these features from multiple omics types to achieve a more comprehensive depiction of the state of a single cell. In combination with continuously advancing bioinformatic algorithms and computational resources, experimental collection of multi-omics data has allowed us to uncover increasingly important and complex insights.

The first application of single cell multi-omics methods is to identify cell subtypes from a heterogeneous cell population. Previously, for example, single cell RNA-seq approaches were shown to be effective in identifying cell subtypes such as human blood dendritic cells, monocytes, and neurons in human brain cortex (MacOsko et al., 2015; Ofengeim et al., 2017; Villani et al., 2017). Recently, single cell DNA methylation sequencing was also applied to study human brain cortex. By examining non-CpG methylation among single cells, they identified novel cell subtypes that were masked in scRNA-seq analysis (Luo et al., 2017). Epigenetic modifications such as DNA methylation are developmentally regulated and cell type-specific, yet stable over the life span, and therefore profiling the epigenome and transcriptome simultaneously can compensate for the limitation of single cell RNA-seq, which mainly yields information about highly expressed transcripts. Thus, different omics measurements can provide non-redundant information about cell identity and enable more detailed and more accurate dissection of complicated tissues.

Second, single cell multi-omics can be used to reconstruct cell lineage trajectories. Understanding cell lineage trajectories during the complete time course of multicellular animal development is the holy grail of developmental biology. DNA

mutations, as well as epigenetic modifications gained during the cell division and passed to the daughter cells, can be used for lineage tracing, while the transcriptome of the matching single cells can reveal the concomitant alteration of gene expression and transcriptional cell fate change during cell proliferation and differentiaion. For example, cancer cells have extremely unstable genomes, and understanding cancer genome evolution is crucial for revealing "driver" mutations or copy number changes that cause carcinogenesis. Single cell multi-omics can not only help us determine the occurrence order of different mutations during cancer evolution, but can also reveal their functional consequences, such as alteration in gene expression, which will eventually help us identify the causal mutations that induce the transition from normal cell to cancer cell.

Lastly but most importantly, single cell multi-omics data provides the resolution to definitively reveal the relationship between different omics readouts. Correlation analysis between different omics is a prevailing approach to generate regulatory hypotheses between two omics data types. For example, cytosine methylation is among the best-studied epigenetic modifications and has been shown to regulate many critical biological processes. With both DNA and RNA sequencing data, DR-seq and G&T-seq have allowed us the ability to reveal correlation between copy number variation and gene expression level at a single cell scale. Further, scTrio-seq showed that large-scale CNVs caused proportional changes in RNA expression of genes within the gained or lost genomic regions, whereas these CNVs generally do not affect DNA methylation in these regions. Our work using scMT-seq not only showed allele-specific expression patterns based on SNV information, but also showed correlation of DNA methylation with allele-specific expression, providing new insight into the study of imprinting and its underlying mechanism. In the near future, multi-omics methods may be helpful for understanding the correlation between DNA mutations with epigenetic modifications and their effects on gene expression to reveal the mechanisms underlying interesting biological questions such as dosage compensation and X-inactivation, among others (Livernois et al., 2012; Graves, 2016). Inevitably, even with single cell multi-omics technology, we are still limited to identifying correlation but not causality. We therefore believe that single cell multi-omics, once combined together with experimental perturbation, will be effective in allowing us to understand causal relationships among omics data types.

Essential to all these applications is the development of computational approaches that help to integrate multiple data layers and to recover information lost due to the sequencing of minute amounts of biological material. Bioinformatic and computational techniques have advanced single cell multi-omics technology in several arenas, such as (1) imputation of "dropped-out" single cell measurements, (2) indirect measurement of another omics layer from a measured one (Farlik et al., 2015; Bock et al., 2016), and (3) mathematical and statistical quantification of multi-dimensional associations (Lane et al., 2017). Imputation methods pull information from groups of similar cells to help to restore measurements for molecules originally in very low abundance, such as lowly expressed RNA transcripts, filling in sparse data matrices for better representations of the original relationships (Van Dijk et al., 2017; Li and Li, 2018). Furthermore, as our knowledge of biological regulatory relationships increases, one data type may be able to serve as proxy for inference of another omics layer. For example, transcription factor binding or copy number alterations have been indirectly inferred from single cell methylation data (Farlik et al., 2015; Hou et al., 2016). Likewise, copy number information can be inferred from the single cell transcriptome (Tirosh et al., 2016), and chromatin state from the methylome (Guo et al., 2017). In addition, as single cell multi-omics technology becomes progressively high throughput, computational resources and time needed for processing of the raw data will be an important aspect in the flexibility of data analysis. Pipelines and new algorithms that streamline and shorten the computational time needed for data processing will be important for increasingly complex, multi-dimensional experiments. Raw files for each omic type must be separately processed, aligned, filtered, and quality-controlled in a manner that accounts for complications inherent in single cell measurements, such as low signal-to-noise ratio, technical amplification artifacts, and technical variation (Bock et al., 2016). Each omics layer of processed data is then assigned back to the single cell and co-analyzed with both mathematical and statistical models to reveal patterns of regulation. These new computational methods, while still nascent, allow us the capacity to bypass experimental limitations and expose excitingly novel relationships.

CONCLUSIONS AND FUTURE DIRECTIONS

Single cell multi-omics methods have provided countless opportunities to systematically understand biological diversity, and to identify rare cell types and their characteristics with unprecedented accuracy through integration of information from multiple omics levels, including DNA, RNA, and protein. These single cell multi-omics methods will play an important role in many diverse fields, and their applications are rapidly expanding, including (1) delineating cellular diversity, (2) lineage tracing, (3) identifying new cell types, and (4) deciphering the regulatory mechanisms between omics. Although some of the applications have been reported in initial studies, there are still many avenues open for exploration, and the further development of new multi-omics methods will also facilitate their increasing utility. It is anticipated that better performance of multi-omics methods will be generated based on the optimization of current single cell sequencing methods. There are currently several main challenges and thus opportunities for further development of single cell multi-omics technology: (1) Overcoming the limitations of current single cell sequencing methods will facilitate the development of more types of omics measurements on single cells. For example, outside of single cell DNA methylome analysis, there are other single cell epigenome sequencing methods such as scAba-seq (DNA hydroxymethylation) (Mooijman et al., 2016), single cell ATAC-seq (open chromatin) (Buenrostro et al., 2015), single cell Hi-C

(chromatin conformation) (Nagano et al., 2013), and single cell ChIP-seq (histone modifications) (Rotem et al., 2015). However, due to limitations such as low genome coverage and high noise signals derived from locus dropout and PCR amplification, no reliable multi-omics approach based on these methods has been reported yet. Optimization of the existing single cell sequencing methods as well as newly developed methods will provide more opportunities to integrate diverse methods with transcriptomic analysis to reveal the relationship between epigenetic states and RNA transcription variation. (2) New approaches to isolate and label multiple types of molecules of the same single cell will help to increase the number of omics profiled in parallel, from dual-omics to triple-omics or more. Even multiple functional parameters of single cells could be included, such as with the development of patch-seq, which combined whole-cell electrophysiological patch-clamp recordings, single-cell RNA-sequencing, and morphological characterization to identify new cell types in the nervous system (Cadwell et al., 2016, 2017). (3) In contrast to the rich resources of experimental protocols, computational methods for single cell multi-omics data analysis have just started to emerge. New computational approaches tailored to the analysis of single cell multi-omics data will also substantially facilitate the application of the methods (Yan et al., 2017). In summary, with further development of multi-omics methods, the future will witness an even wider application of single cell multi-omics technology that will result in meaningful findings never before achieved.

AUTHOR CONTRIBUTIONS

YH and YG: Conceived the structure of the manuscript; YH, YG, and QA: Wrote the manuscript; QA, BT, KS, and SF: Read and edited the manuscript.

ACKNOWLEDGMENTS

The work was supported by National Key R&D Program of China (2017YFA0104100, 2017YFC1001300), National Natural Science Foundation of China (31700900).

REFERENCES

Angermueller, C., Clark, S. J., Lee, H. J., MacAulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232. doi: 10.1038/nmeth.3728

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18:67. doi: 10.1186/s13059-017-1189-z

Autengruber, A., Gereke, M., Hansen, G., Hennig, C., and Bruder, D. (2012). Impact of enzymatic tissue disintegration on the level of surface molecule expression and immune cell function. *Eur. J. Microbiol. Immunol.* 2, 112–120. doi: 10.1556/EuJMI.2.2012.2.3

Bock, C., Farlik, M., and Sheffield, N. C. (2016). Multi-omics of single cells: strategies and applications. *Trends Biotechnol.* 34, 605–608. doi: 10.1016/j.tibtech.2016.04.004

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490. doi: 10.1038/nature14590

Cadwell, C. R., Palasantza, A., Jiang, X., Berens, P., Deng, Q., Yilmaz, M., et al. (2016). Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* 34, 199–203. doi: 10.1038/nbt.3445

Cadwell, C. R., Scala, F., Li, S., Livrizzi, G., Shen, S., Sandberg, R., et al. (2017). Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using Patch-seq. *Nat. Protoc.* 12, 2531–2553. doi: 10.1038/nprot.2017.120

Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* 46, 2496–2506. doi: 10.1002/eji.201646347

Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., et al. (2016). Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* 13, 833–836. doi: 10.1038/nmeth.3961

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* 9:781. doi: 10.1038/s41467-018-03149-4

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* 48, 1193–1203. doi: 10.1038/ng.3646

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563. doi: 10.1038/227561a0

Darmanis, S., Gallant, C. J., Marinescu, V. D., Niklasson, M., Segerman, A., Flamourakis, G., et al. (2016). Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* 14, 380–389. doi: 10.1016/j.celrep.2015.12.021

Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* 11, 1095–1099. doi: 10.1101/gr.180501

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33, 285–289. doi: 10.1038/nbt.3129

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., et al. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* 14, 491–493. doi: 10.1038/nmeth.4227

Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., et al. (2016). DNA methylation dynamics of human hematopoietic stem cell differentiation. *Cell Stem Cell* 19, 808–822. doi: 10.1016/j.stem.2016.10.019

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., et al. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* 10, 1386–1397. doi: 10.1016/j.celrep.2015.02.001

Frei, A. P., Bava, F. A., Zunder, E. R., Hsieh, E. W., Chen, S. Y., Nolan, G. P., et al. (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* 13, 269–275. doi: 10.1038/nmeth.3742

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* 17, 175–188. doi: 10.1038/nrg.2015.16

Goetz, J. J., and Trimarchi, J. M. (2012). Transcriptome sequencing of single cells with Smart-Seq. *Nat. Biotechnol.* 30, 763–765. doi: 10.1038/nbt.2325

Graves, J. A. (2016). Evolution of vertebrate sex chromosomes and dosage compensation. *Nat. Rev. Genet.* 17, 33–46. doi: 10.1038/nrg.2015.2

Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., et al. (2017). Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* 27, 967–988. doi: 10.1038/cr.2017.82

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos

analyzed using reduced representation bisulfite sequencing. *Genome Res.* 23, 2126–2135. doi: 10.1101/gr.161679.113

Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. (2015). SINCERA: a pipeline for single-cell rna-seq profiling analysis. *PLoS Comput. Biol.* 11:e1004575. doi: 10.1371/journal.pcbi.1004575

Han, K. Y., Kim, K. T., Joung, J. G., Son, D. S., Kim, Y. J., Jo, A., et al. (2018). SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* 28, 75–87. doi: 10.1101/gr.223263.117

Han, L., Zi, X., Garmire, L. X., Wu, Y., Weissman, S. M., Pan, X., et al. (2014). Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. *Sci. Rep.* 4:6485. doi: 10.1038/srep06485

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2, 666–673. doi: 10.1016/j.celrep.2012.08.003

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* 26, 304–319. doi: 10.1038/cr.2016.23

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., et al. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* 17:88. doi: 10.1186/s13059-016-0950-z

Huang, L., Ma, F., Chapman, A., Lu, S., and Xie, X. S. (2015). Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genomics Hum. Genet.* 16, 79–102. doi: 10.1146/annurev-genom-090413-025352

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236

Knouse, K. A., Wu, J., and Amon, A. (2016). Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res.* 26, 376–384. doi: 10.1101/gr.198937.115

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Lane, K., Van Valen, D., Defelice, M. M., MacKlin, D. N., Kudo, T., Jaimovich, A., et al. (2017). Measuring signaling and RNA-Seq in the same cell links gene expression to dynamic patterns of NF-kappaB activation. *Cell Syst.* 4, 458.e455–469.e455. doi: 10.1016/j.cels.2017.03.010

Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9:997. doi: 10.1038/s41467-018-03405-7

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

Livernois, A. M., Graves, J. A., and Waters, P. D. (2012). The origin and evolution of vertebrate sex chromosomes and dosage compensation. *Heredity* 108, 50–58. doi: 10.1038/hdy.2011.106

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* 357, 600–604. doi: 10.1126/science.aan3351

MacAulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522. doi: 10.1038/nmeth.3370

MacAulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-Cell multiomics: multiple measurements from single cells. *Trends Genet.* 33, 155–168. doi: 10.1016/j.tig.2016.12.003

MacOsko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mooijman, D., Dey, S. S., Boisset, J. C., Crosetto, N., and van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat. Biotechnol.* 34, 852–856. doi: 10.1038/nbt.3598

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., et al. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64. doi: 10.1038/nature12593

Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome Res.* 25, 1499–1507. doi: 10.1101/gr.191098.115

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94. doi: 10.1038/nature09807

Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C., and Yuan, J. (2017). Single-Cell RNA sequencing: unraveling the brain one cell at a time. *Trends Mol. Med.* 23, 563–576. doi: 10.1016/j.molmed.2017.04.006

Ortega, M. A., Poirion, O., Zhu, X., Huang, S., Wolfgruber, T. K., Sebra, R., et al. (2017). Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clin. Transl. Med.* 6:46. doi: 10.1186/s40169-017-0177-y

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939. doi: 10.1038/nbt.3973

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181. doi: 10.1038/nprot.2014.006

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979–982. doi: 10.1038/nmeth.4402

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., et al. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33, 1165–1172. doi: 10.1038/nbt.3383

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., et al. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14:R31. doi: 10.1186/gb-2013-14-4-r31

Schep, A. N., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* 14, 975–978. doi: 10.1038/nmeth.4401

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14, 618–630. doi: 10.1038/nrg3542

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11, 817–820. doi: 10.1038/nmeth.3035

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868. doi: 10.1038/nmeth.4380

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., MacAulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi: 10.1038/nmeth.4220

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550, 249–254. doi: 10.1038/nature24041

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. A., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 13, 718–725. doi: 10.1016/0888-7543(92)90147-K

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H. II., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375. doi: 10.1038/nature13173

Van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., et al. (2017). MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*. doi: 10.1101/111591

Vermeesch, J. R., Voet, T., and Devriendt, K. (2016). Prenatal and pre-implantation genetic diagnosis. *Nat. Rev. Genet.* 17, 643–656. doi: 10.1038/nrg.2016.97

Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356:eaah4573. doi: 10.1126/science.aah4573

Volovitz, I., Shapira, N., Ezer, H., Gafni, A., Lustgarten, M., Alter, T., et al. (2016). A non-aggressive, highly efficient, enzymatic method for dissociation of human brain-tumors and brain-tissues to viable single-cells. *BMC Neurosci.* 17:30. doi: 10.1186/s12868-016-0262-y

Wang, Y., and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol. Cell* 58, 598–609. doi: 10.1016/j.molcel.2015.05.005

Wineinger, N. E., Kennedy, R. E., Erickson, S. W., Wojczynski, M. K., Bruder, C. E., and Tiwari, H. K. (2008). Statistical issues in the analysis of DNA copy number variations. *Int. J. Comput. Biol. Drug Des.* 1, 368–395. doi: 10.1504/IJCBDD.2008.022208

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593–597. doi: 10.1038/nature12364

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2017). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief. Bioinform*. 2017, 1–12. doi: 10.1093/bib/bbx066

Zamanighomi, M., Lin, Z., Daley, T., Schep, A., Greenleaf, W. J., and Wong, W. H. (2017). Unsupervised clustering and epigenetic classification of single cells. *bioRxiv*. doi: 10.1101/143701

Zhang, C. Z., Adalsteinsson, V. A., Francis, J., Cornils, H., Jung, J., Maire, C., et al. (2015). Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.* 6:6822. doi: 10.1038/ncomms7822

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622–1626. doi: 10.1126/science.1229164

Žurauskiene, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17:140. doi: 10.1186/s12859-016-0984-y

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Cell and Developmental Biology

REVIEW
published: 05 September 2016
doi: 10.3389/fcell.2016.00089

Fluorescence *In situ* Hybridization: Cell-Based Genetic Diagnostic and Research Applications

Chenghua Cui[1,2], *Wei Shu*[1,3] *and Peining Li*[1]*

[1] *Laboratory of Clinical Cytogenetics, Department of Genetics, Yale School of Medicine, New Haven, CT, USA,* [2] *Department of Pathology, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences, Tianjin, China,* [3] *Department of Cell Biology and Genetics, Guangxi Medical University, Nanning, China*

Fluorescence *in situ* hybridization (FISH) is a macromolecule recognition technology based on the complementary nature of DNA or DNA/RNA double strands. Selected DNA strands incorporated with fluorophore-coupled nucleotides can be used as probes to hybridize onto the complementary sequences in tested cells and tissues and then visualized through a fluorescence microscope or an imaging system. This technology was initially developed as a physical mapping tool to delineate genes within chromosomes. Its high analytical resolution to a single gene level and high sensitivity and specificity enabled an immediate application for genetic diagnosis of constitutional common aneuploidies, microdeletion/microduplication syndromes, and subtelomeric rearrangements. FISH tests using panels of gene-specific probes for somatic recurrent losses, gains, and translocations have been routinely applied for hematologic and solid tumors and are one of the fastest-growing areas in cancer diagnosis. FISH has also been used to detect infectious microbias and parasites like malaria in human blood cells. Recent advances in FISH technology involve various methods for improving probe labeling efficiency and the use of super resolution imaging systems for direct visualization of intra-nuclear chromosomal organization and profiling of RNA transcription in single cells. Cas9-mediated FISH (CASFISH) allowed *in situ* labeling of repetitive sequences and single-copy sequences without the disruption of nuclear genomic organization in fixed or living cells. Using oligopaint-FISH and super-resolution imaging enabled *in situ* visualization of chromosome haplotypes from differentially specified single-nucleotide polymorphism loci. Single molecule RNA FISH (smRNA-FISH) using combinatorial labeling or sequential barcoding by multiple round of hybridization were applied to measure mRNA expression of multiple genes within single cells. Research applications of these single molecule single cells DNA and RNA FISH techniques have visualized intra-nuclear genomic structure and sub-cellular transcriptional dynamics of many genes and revealed their functions in various biological processes.

Keywords: fluorescence *in situ* hybridization (FISH), genetic diagnosis, aneuploidy, pathogenic copy number variants (CNV), microdeletion/microduplication syndromes, Cas-9 mediated FISH (CASFISH), oligopaint-FISH, single molecule RNA FISH (smRNA-FISH)

OPEN ACCESS

Edited by:
Shixiu Wu,
Hangzhou Cancer Research Institute, China

Reviewed by:
Frederick Charles Campbell,
Queen's University Belfast, UK
Marco Ghezzi,
University of Padova, Italy

***Correspondence:**
Peining Li
peining.li@yale.edu

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *01 May 2016*
Accepted: *11 August 2016*
Published: *05 September 2016*

Citation:
Cui C, Shu W and Li P (2016) Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. Front. Cell Dev. Biol. 4:89. doi: 10.3389/fcell.2016.00089

INTRODUCTION

Fluorescence *in situ* hybridization (FISH) uses DNA fragments incorporated with fluorophore-coupled nucleotides as probes to examine the presence or absence of complementary sequences in fixed cells or tissues under a fluorescent microscope. This hybridization-based macromolecule recognition tool was very effective in mapping genes and polymorphic loci onto metaphase chromosomes for constructing a physical map of the human genome (Langer-Safer et al., 1982; Lichter et al., 1993). FISH technology offers three major advantages including high sensitivity and specificity in recognizing targeted DNA or RNA sequences, direct application to both metaphase chromosomes and interphase nuclei, and visualization of hybridization signals at the single-cell level. These advantages increased the analytic resolution from Giemsa bands to the gene level and enabled rapid detection of numerical and structural chromosomal abnormalities (Klinger et al., 1992; Ried et al., 1992). Clinical application of FISH technology had upgraded classical cytogenetics to molecular cytogenetics. With the improvement in probe labeling efficiency and the introduction of a super resolution imaging system, FISH has been renovated for research analysis of nuclear structures and gene functions. This review presents the recent progress in FISH technology and summarizes its diagnostic and research applications.

CELL BASED GENETIC DIAGNOSIS BY FISH

Analytical and Clinical Validities and Practice Guidelines

Most DNA fragments used as probes are extracted from bacterial artificial clones (BACs) which contain cloned human genomic DNA sequences in the size of 100–200 Kilobases (Kb). These DNA fragments could be directly labeled by nick translation to incorporate nucleotides coupled with different fluorophores such as coumarins, fluoresceins, rhodamine, and cyanines (Cy3, Cy5, and Cy7) (Morrison et al., 2003). According to the targeted regions and labeling design, FISH probes can be divided into locus-specific probes targeted to specific regions or genes and regional painting probes for specific chromosomal bands, an entire chromosome or whole genome. Commonly used locus-specific probes include alpha repetitive sequences for centromeric regions and single copy sequences for subtelomeric and gene regions. Multi-color locus-specific probes allow simultaneously detection of numerical abnormalities of two to three regions in one FISH assay. For structural rearrangements, locus-specific probes with different fluorophores for two genes or for the 5′ and 3′ regions of a gene have been used to detect "double-fusion" signals resulting from a reciprocal translocation or "break apart" signals from a gene rearrangement, respectively. Painting probes have been used mostly in a research setting to dissect chromosome domains within a nucleus or structural rearrangements in metaphase chromosomes. **Figure 1** shows representative FISH applications of locus-specific and chromosome painting probes in the detection of numerical and structural chromosomal abnormalities.

Earlier studies had evaluated signal-to-noise ratios, spatial resolution of the fluorescent signals, and hybridization/detection efficiencies of FISH tests on lymphocytes and aminocytes (Klinger et al., 1992; Ried et al., 1992). These studies led to the commercialization of FISH probes with optimized probe selection and standardized labeling, and the clinical utility of FISH testing in large case series (Ward et al., 1993). To ensure safe and effective diagnostic application, a clinical cytogenetics laboratory needs to establish the analytical and clinical validities for every FISH assay. The analytical validity of a FISH assay is evaluated by its targeted accuracy, sensitivity, specificity, and normal reference ranges following a standardized laboratory procedure (Wolff et al., 2007; Ciolino et al., 2009). FISH testing could be used as an adjunctive assay or a stand-alone diagnostic assay for constitutional and somatic abnormalities. The clinical validity for its intended use should be evaluated by calculating the sensitivity from patients with targeted abnormalities and the specificity from normal controls. Other analytical and clinical considerations include possible false positive or negative results, continuous monitoring of signal variations, periodical evaluation and batch-to-batch comparisons of probe performances (Test and Technology Transfer Committee, 2000).

FISH technology enabled the detection of an increased spectrum of genetic disorders from chromosomal abnormalities to submicroscopic copy number variants (CNVs) and extended the cell-based analysis from metaphases to interphases (Xu and Li, 2013). The analytical resolution of FISH is in the range of 100–200 Kb as determined by the probe size, which is 50-fold higher than the 5–10 megabase (Mb) Giesma banding of a high resolution karyotyping. Locus-specific probes detected submicroscopic CNV and led to the identification of a group of genomic disorders (also termed contiguous gene syndromes or microdeletion syndromes), such as DiGeorge syndrome (OMIM#188400) by a deletion at 22q11.2, Prader-Willi syndrome (OMIM#176270) and Angelman syndrome (OMIM#105830) by a deletion at 15q11.2. FISH can be performed directly on interphase nuclei, which eliminated the time consuming cell culture procedure and extended its diagnostic application toward rapid screening of chromosomal and genomic abnormalities. In the following sections, the diagnostic applications of FISH technology are focused on three main areas: prenatal screening and postnatal diagnosis of constitutional chromosomal abnormalities and submicroscopic pathogenic CNVs, identification and monitoring of acquired chromosomal abnormalities in hematopoietic and solid tumors, and the detection of infectious diseases caused by microbials and parasites.

Detection of Constitutional Chromosomal Abnormalities and Pathogenic CNVs

A Multiplex FISH panel with differentially labeled probes has been developed for prenatal screening of common aneuploidies involving gains or losses of chromosomes X, Y, 13, 18, and 21 (Ried et al., 1992; Ward et al., 1993). Pregnant women

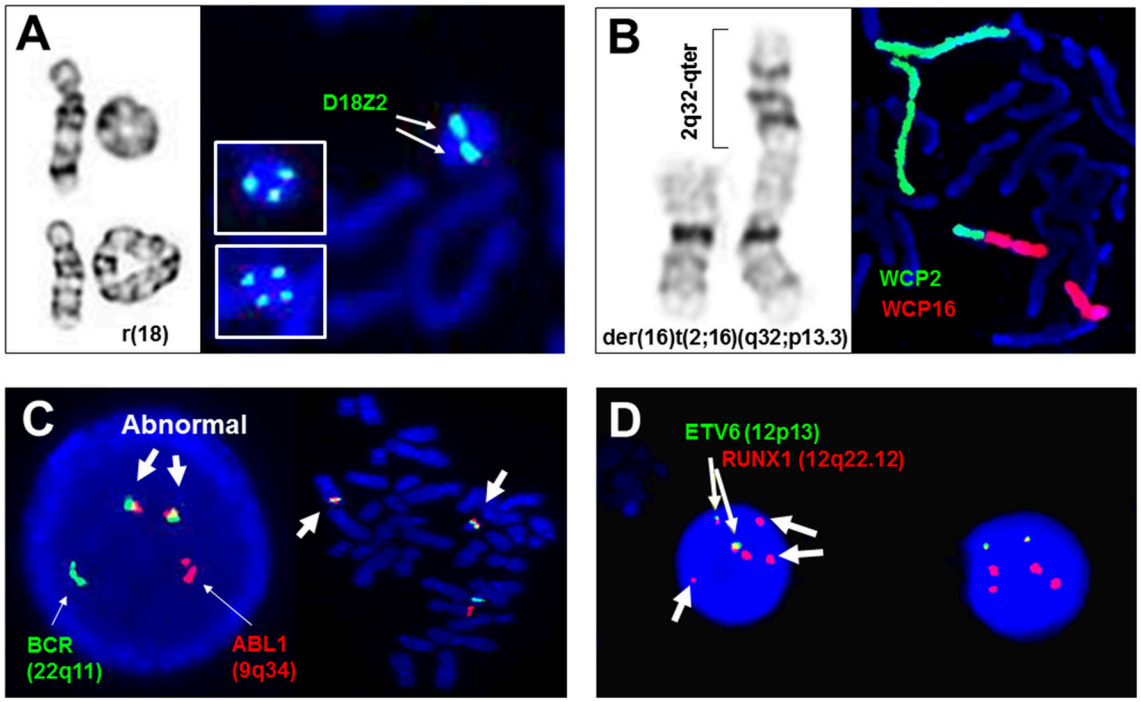


FIGURE 1 | Adjunctive and diagnostic assays of FISH in clinical cytogenetics. (A) The detection of di-centric, tri-centric, and tetra-centric ring chromosome 18 using a centromeric probe D18Z2 for chromosome 18. Left panel shows normal chromosome 18, dicentric ring 18 in top, and tetracentric ring 18 in bottom, right panel shows dicentric ring 18 and tricentric/tetracentric ring 18 in insets by FISH. **(B)** The detection of a derivative chromosome 16 from a 2q32/16p13.3 translocation by whole chromosome painting probes for chromosomes 2 (WCP2) and 16 (WCP16). **(C)** The detection of ABL1/BCR gene fusions in interphase and metaphase cells by dual color double fusion probes (thin arrows point to the normal signal and thick arrows point to the abnormal fusion signals). **(D)** Diagnostic use of ETV6 and RUNX1 probes for the detection of two fusion signals for a cryptic t(12;21)(p13;q22), loss of an ETV6 signal and gain of three extra RUNX1 signals (thin arrows point to the fusion signals and thick arrows to extra RUNX1 signals). All images are from Yale clinical cytogenetics laboratory.

with a single indication or combined indications of advanced maternal age, abnormal ultrasound findings, or abnormal maternal serum screening have an increased risk of 4–30% for carrying numerical and structural chromosomal abnormalities; among these abnormalities, 84% were numerical abnormalities mostly detectable by the multiplex FISH panel, and 16% were structural abnormalities required further microarray analysis (Li et al., 2011). For prenatal cases with cardiac anomalies detected by prenatal ultrasound examination, DiGeorge syndrome was detected by FISH. Recently, the application of non-invasive prenatal testing by massive parallel sequencing on maternal cell-free fetal DNA significantly improved the accuracy of aneuploidy screening, which resulted in a 57% decline in invasive prenatal procedures and an increase of diagnostic yield of chromosomal abnormalities (Xu Z. Y. et al., 2013; Meng et al., 2015). Despite these technology advances in prenatal diagnosis, the multiplex FISH panel is still used as an adjunctive assay for rapid detection of common aneuploidies. It should be noted that false positive or negative results as well as maternal cell contamination have been noted in prenatal FISH analysis. Therefore, an irreversible therapeutic action should not be initiated on the basis of FISH results alone. The current guideline recommended that clinical decisions should be made based on two of three pieces of available information: FISH results, conventional cytogenetic analysis and clinical information (Test and Technology Transfer Committee, 2000). Furthermore, aneuploidies and polyploidies have been detected in about 50% of first trimester spontaneous abortions by chromosome analysis and in 35% of products of conception culture failure cases by microarray analysis; it is recognized that an extended FISH panel for chromosomes X/Y/18, 13/21, and 15/16/22 will detect all polyploidies, 84% of aneuploidies, and 69% of multiple aneuploidies causing miscarriages (Zhou et al., 2016).

Developmental delay, intellectual disabilities, and multiple congenital anomalies are present in 1–5% of newborns, and chromosome microarray analysis as the first tier genetic testing has detected a spectrum of cytogenomic abnormalities in 10∼20% of these patients (Miller et al., 2010; Li et al., 2015). Analysis of abnormal findings from consecutive pediatric cases observed genomic disorders (microdeletion/microduplication syndromes), subtelomeric rearrangements, interstitial imbalances, chromosomal structural rearrangements, and aneuploidies in about 37, 26, 19, 10, and 8% of these cases, respectively (Xu et al., 2014). Cell-based FISH testing has been a cost-effective adjunctive assay to confirm microarray detected genomic disorders and then to detect carrier statues in a follow-up parental study. Microdeletions can be detected as a loss of one signal in metaphases and interphases, while

microduplications can be detected as "twin-spot" like two signals in the interphase nuclei. For subtelomeric rearrangements, a complete set of subtelomeric FISH probes for all human chromosomes was developed (Ning et al., 1996) and have been used routinely as an adjunctive assay in visualizing cryptic and complex subtelomeric rearrangements (Li et al., 2006; Rossi et al., 2009). For many newly defined loci of genomic disorders and interstitial imbalances, there are no commercially available FISH probes. Therefore, "home-brew" targeted BAC clone FISH probes were used for these unique cases (Li et al., 2006; Khattab et al., 2011).

Structural rearrangements like ring chromosomes and small supernumerary marker chromosomes (sSMC) present not only segmental gains or losses but also a mosaic pattern due to their dynamic behavior in mitosis. As shown in **Figure 1A**, centromeric FISH probes are routinely used to track the changes from dicentric, tricentric, and tetracentric ring chromosomes to loss of the ring through mitosis. Subtelomeric and interstitial FISH probes have been used to define the intactness of the ring chromosome and the level of mosaicism (Zhang et al., 2004; Xu F. et al., 2013). A cytogenomic approach combining chromosome, FISH, and microarray analyses has been recommended for characterizing the genomic structure, mitotic instability, and mechanisms of ring formation for cases with a ring chromosome (Zhang et al., 2012). sSMC are extra centric chromosome fragments usually in the forms of an inverted duplication or a small ring chromosome and present in 0.043% of newborn children. Several sSMC have syndromic phenotypes such as inv dup(22q11.2) for cat-eye syndrome (OMIM∗607576) and i(12p) for Pallister Killian syndrome (OMIM#601803), and others like inv dup(15q) and i(18p) can have variable phenotypes (Liehr et al., 2004, 2006). About 30% of sSMC are derived from chromosome 15; the D15S10 or *SNRPN* probes are routinely used to assess inv dup(15q) (Wang et al., 2015). The euchromatic material in sSMC can be detected by a microarray analysis. A set of pericentric core probes for each arm of human chromosomes has been validated for characterizing unambiguously the chromosomal origin of sSMC and the level of mosaicism (Castronovo et al., 2013).

Identification and Monitoring of Acquired Chromosomal Abnormalities

The discovery of Philadelphia chromosome in chronic myeloid leukemia (CML) followed by the characterization of t(9;22)(q34;q11) with underlying ABL1/BCR gene fusions supported the causative role of chromosomal abnormalities in carcinogenesis and set the foundation for cancer cytogenetics (Mitelman et al., 2007). Cancer is considered a genetic disease at the cellular level resulting from either a progressive process or a one-off catastrophic event (Stephens et al., 2011; Li and Cui, 2016). The two main pathogenetic pathways for hallmarks of cancer development are the inactivation of tumor suppressor genes by deletions, mutations, miRNA upregulation, or epigenetic mechanisms, and the activation or deregulation of oncogenes as a consequence of point mutations, amplification or balanced cytogenetic abnormalities (Vogelstein and Kinzler, 2004; Hanahan and Weinberg, 2011). Recurrent chromosomal abnormalities including translocations, deletions, duplications, and gene amplifications associated with distinct tumor entities have been characterized; specifically designed FISH panels have been widely used in the diagnosis and monitoring of acquired chromosomal abnormalities in hematologic and solid tumors (Hu et al., 2014; Liehr et al., 2015; Mikhail et al., 2016).

Current guidelines recommend an integrated approach for cancer cytogenetic diagnosis (Wolff et al., 2007). In general, both conventional karyotyping and FISH testing are used for initial diagnosis and follow up monitoring of clonal abnormalities. For hematopoietic and lymphoid tumors, the most commonly used FISH probes and disease-specific panels in a clinical cytogenetics laboratory are listed in **Table 1**. Results from a FISH panel offer a quick evaluation of targeted abnormal patterns and their percentage within the bone marrow cells or leukocytes. Chromosome analysis will then reveal the clonal abnormalities and clonal evolution. For leukemias requiring urgent treatment, such as acute promyelocytic leukemia (APL) caused by the t(15;17)(q24;q21) with underlying PML/RARa fusions, rapid FISH result is mandated for the administration of all-trans retinoic acid (ATRA). Targeted therapy against the ABL1/BCR fusion protein by small molecule tyrosine inhibitors like imatinib mesylate (Gleevec), dasatinib (Sprycel), and nilotinib (Tasigna) has increased the 10-year overall survival from 20 to 80–90% (Li et al., 2013). For many cryptic rearrangements undetectable by routine chromosome analysis, such as t(12;21)(p13;q22) with ETV6/RUNX1 gene fusions, t(4;14)(p16.3;q32) with FGFR3/IGH gene fusions, deletions of 12p13 (ETV6), 13q14 (RB1), and 17p13 (TP53), FISH tests are considered a stand-alone diagnostic assay. Adjunctive use of FISH probes to further define ambiguous or hidden chromosomal abnormalities is required for many cases (Kamath et al., 2008; Massaro et al., 2011). Additionally, FISH is a sensitive and timely method to monitor residual diseases with known clonal abnormality and bone marrow transplantation by sex-mismatch donor at cellular level. Considering some hematologic tumors may be morphologically similar and the abnormalities may not be detected by low-resolution karyotyping and/or in low percentage of leukemic cells, FISH could be important for differential diagnosis between these diseases. For example, cyclin D1 (*CCND1*) translocation can be detected by FISH as a characteristic abnormality in mantle cell lymphoma, which provides differential diagnosis for morphologically similar chronic lymphoid leukemia (CLL). Furthermore, FISH for nuclear DNA can be combined with immunostaining of cytoplasmic markers for simultaneous identification of chromosomal abnormalities and cell types. For example, *IGH* translocation is present in multiple myeloma and monoclonal gammopathy of undetermined significance (MM/MGUS) with high frequency, which is usually detected in plasma cells. In a two-step assay with first the hybridization of *IGH* probe and then immune-staining by fluoresceinisothiocyanate (FITC)-conjugated antibodies against κ- or λ-light chain, the FITC-stained cytoplasm and *IGH* break apart signals within the nuclei were visualized in plasma cells simultaneously. This modified immuno-FISH was expected to improve the diagnostic

TABLE 1 | List of FISH panels and probes for hematopoietic and lymphoid tumors.

Gene (G-band)	Probe Design	Myeloid leukemia			Lymphocytic leukemia			Lymphoma	MM/MGUS	MPD
		CML	MDS	AML	CLL	B-ALL	T-ALL			
CKS1B (1q21), CDKN2C (1p32)	DCE								1p/1q+	
PBX1 (1q23.3), TCF3 (19p13.3)	DCDF					t(1;19)		t(1;19)		
ALK (2p23)	DCBAP							ALK		
MECOM (3q26)	DCBAP			inv(3)						
BCL6 (3q27)	DCBAP							BCL6		
D4Z1 (4cen), D10Z1 (10cen), D17Z1 (17cen)	TCE					+4/10/17				
PDGFRA (4q12)	DCBAP									PDGFRA
FGFR3 (4p16.3), IGH (14q32)	DCDF								t(4;14)	
TAS2R1 (5p15.31), EGR1 (5q31)	DCE		5q−/−5							
PDGFRB (5q33)	DCBAP					PDGFRB				PDGFRB
MYB (6q23), D6Z1 (6cen)	DCE								6q−	
RELN (7q22), TES (7q31)	DCE		7q−/−7							
TCRB (7q34)	DCBAP						TCRB			
FGFR1 (8p11)	DCBAP									FGFR1
RUNX1T1 (8q21), RUNX1 (21q22)	DCDF			t(8;21)						
cMYC (8q24)	DCBAP							cMYC		
cMYC (8q24), D20S108 (20q12)	DCE		+8/20q−							
PAX5 (9p13.2)	DCBAP					PAX5				
CDKN2A (9p21), D9Z3 (9cen)	DCE					9p−	9p−			
ABL (9q34), BCR (22q11)	DCDF	t(9;22)				t(9;22)	t(9;22)			
CCND1 (11q13), IGH (14q32)	DCDF							t(11;14)	t(11;14)	
ATM (11q22), TP53 (17p13)	DCE				11q−/17p−					
KMT2A (11q23)	DCBAP			KMT2A		KMT2A	KMT2A			
ETV6 (12p13), RUNX1 (21q22)	DCDF					t(12;21)				
DLEU1 (13q14), D13S25 (13q34)	DCE								13q−	
DLEU1 (13q14), D13S25 (13q34), D12Z3 (12cen)	TCE				13q−/+12					
TCRA/D (14q11)	DCE						TCRA			
IGH (14q32)	DCBAP				IGH			IGH	IGH	
IGH (14q32), BCL2 (18q21)	DCDF							t(14;18)		
SNRPN (15q11.2), TP53 (17p13)	DCE								+15/17p−	
PML (15q24), RARA (17q21)	DCDF			t(15;17)						
MYH11 (16p13), CBFB (16q22)	DCDF			inv(16)						
MALT1 (18q21)	DCBAP							MALT1		
CRLF2 (Xp22.33)	DCBAP					CLFR2				

DCE,dual-color enumerate; TCE, tri-color enumerate; DCBAP, dual-color break apart; DCDF, dual-color double fusion; CML, Chronic myeloid leukemia; MDS, Myelodysplastic syndrome; AML, Acute myeloid leukemia; CLL, Chronic lymphocytic leukemia; B-ALL, B-cell acute lymphocytic leukemia; T-ALL, T-cell acute lymphocytic leukemia; MM/MGUS, Multiple myeloma/Monoclonal mopathy of undetermined significance; MPD, Myeloproliferative disorder. Shaded for recurrent abnormalities detected by a primary FISH panel, unshaded for secondary FISH probes for specific abnormalities. For references see (Hu et al., 2014; Liehr et al., 2015), and (Mikhail et al., 2016).

accuracy but the low sensitivity limited its application only in follow-up study (Boersma-Vreugdenhil et al., 2003).

FISH tests are widely used in various types of solid tumors. For example, FISH can define gene rearrangements in congenital fibrosarcoma with a novel complex translocation (Marino-Enriquez et al., 2008) and validate subclone markers in heterogeneous melanoma biopsies (Parisi et al., 2011). FISH results can be used to guide cancer treatment. For example, Herceptin-targeted therapy is effectively against *HER2* over-expressed breast cancer. For routine clinical specimen, immunohistochemistry, real-time polymerase chain reaction, and FISH were used to assess the *HER2* protein level, RNA expression, and DNA copy numbers, respectively. Among these methods, FISH offered a cell-based evaluation for the ratio of *HER2* gene copy number to the number of copies of chromosome 17 (*HER2*/CEP17 ratio). The FISH scoring criteria for *HER2*/CEP17 ratio and the interpretive guidelines were reported (Hicks et al., 2005). Many targeted therapies for recurrent translocations in various types of solid tumors have been either approved by FDA or are under clinical trials. For example, lapatinib, sorafenib, sunitinib, termsirolimus, and pazopanib have been used for papillary renal cell carcinoma with translocations involving the *TFE3* gene at Xp11.2; cixutumumab and mithramycin are in phase II clinical trial for Ewing sarcoma with translocation involving the EWSR1 gene at 22q12 (Li et al., 2013). FISH assays using probes for specific recurrent translocations from different solid tumors could guide effective targeted therapy. FISH tests were also used to evaluate sperm aneuploidy frequencies before and after chemotherapy in patients with testicular cancer and Hodgkin's lymphoma; significantly increased frequencies of aneuploidies for a duration up to 24 months were noted (De Mas et al., 2001; Tempest et al., 2008). It was recommended that genetic counseling about potentially increased reproduction risk from chemotherapy should be offered to cancer patients.

Detection of Infectious Diseases by FISH

The majority of FISH probes target to specific chromosomal and genomic abnormalities in the human genome. Rapid phylogenetic identification of single microbial cells was achieved using fluorescently labeled oligonucleotides complementary to 16S ribosomal RNA (rRNA) (DeLong et al., 1989). Some segments in the 16S rRNA are invariant in all organisms but phylogenetic group-specific 16S rRNA in different groups of organism can be used as oligonucleotide FISH probes (length 17–34 nucleotides) to identify infectious agents in clinical samples. For example, FISH probes complementary to specific sequence of 16s rRNA can detect malaria infection in blood samples. The *Plasmodium* Genus (P-Genus) FISH assay has a *Plasmodium* genus specific probes that detect all five species of *Plasmodium* known to cause the disease in humans. The sensitivity of this FISH assay is better than the Giemsa staining method. A LED light source may be an available device to read FISH result, which can extend the clinical application of FISH especially in the resource-limited areas. Since rRNA has a short life and is present in a live organism with plenty of copies, FISH should be done in the live pathogens (Shah et al., 2015).

SINGLE-CELL DNA STRUCTURAL AND RNA TRANSCRIPTIONAL ANALYSES

FISH assays using locus-specific and regional painting probes are still a powerful tool in visualizing simple and complex chromosomal and genomic rearrangements. Fiber-FISH by locus-specific BAC clone probes within a 900 Kb 17q12 inversion hybridizing onto stretched DNA fibers correlated the inversion orientations with associated haplotypes, which allowed the evaluation of inversion frequencies among human populations globally (Donnelly et al., 2010). Pericentriomeric heterochromatin probes were used in a three dimensional FISH (3D-FISH) to study intra-nuclear centromeric positions in cultured cells from patients with ICF syndrome (immunodeficiency, centromeric region instability, facial anomalies) and Robert syndrome (cohesion defect by mutations in the *ESCO2* gene) (Dupont et al., 2012, 2014). Multi-color FISH (M-FISH) by painting probes specific for a human chromosome and multi-color banding FISH (M-BAND) by painting probes specific for every band in a chromosome were used to visualize complex chromosomal rearrangements from chromothripsis in two patients with acute myeloid leukemia (Mackinnon and Campbell, 2013). Chromothripsis are seen as regional clustering of breakpoints and regularity of oscillating copy-number states by microarray analysis and as heterogeneous staining regions, marker or ring chromosomes, and other undefinable rearrangements by chromosome analysis (Stephens et al., 2011). Selected FISH probes targeting to the oscillating copy-number gains and losses could be used to monitor the abnormal clones with chromothripsis.

FISH technology has made significant progress with the innovation of novel labeling methods and the introduction of super resolution imaging systems for fine mapping of intra-nuclear genomic structures and for single cells single molecule profiling of cytoplasmic RNA transcription. Recently, a novel FISH method using nuclease-deficient clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated caspase 9 (dCas9) system was developed. The initial design used enhanced green fluorescent protein (EGFP) tagged dCas9 and small guide RNA (sgRNA) targeting to repetitive telomere sequences or sgRNAs tiling along a non-repetitive genomic sequences at the *MUC4* locus. This method enabled the visualization of intra-nuclear locations and dynamics of telomeres and *MUC4* loci during mitosis in living human cells (Chen et al., 2013). Further modification by using both fluorophore-coupled sgRNA and fluorophore-coupled dCas9 was termed Cas9-mediated FISH (CASFISH); rapid and robust labeling of repetitive DNA elements in preicentromere, centromere, G-rich telomere, and *MUC4* gene by CASFISH was demonstrated (**Figure 2A**; Deng et al., 2015). This CASFISH did not require the denature treatment for targeted DNA and therefore preserved the nature spatiotemporal organization of the nucleus. The CASFISH process is remarkably rapid (within 1 h) and can be used directly on fixed tissues or living cells. However, using tiling sgRNAs for single-copy gene regions could have low labeling efficiency and higher background. Further optimization of this CASFISH technology

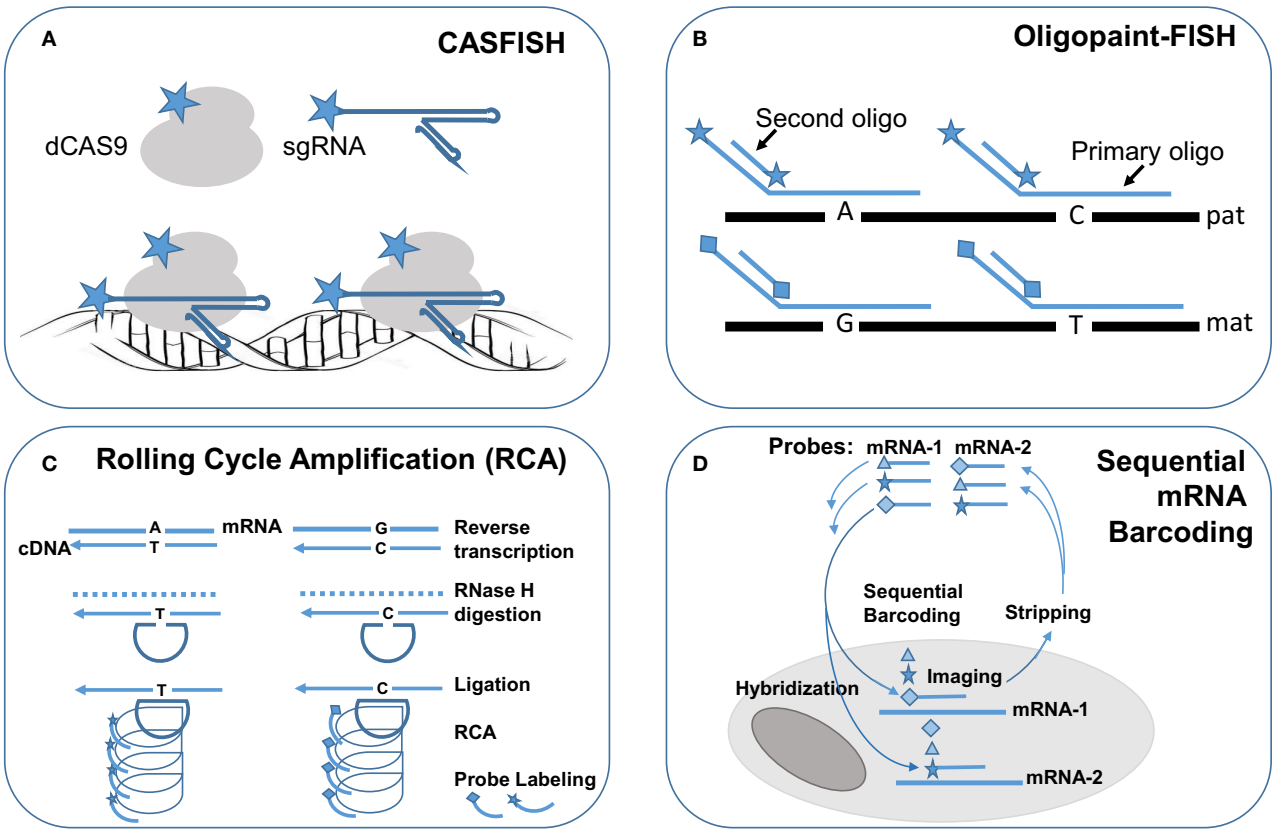


FIGURE 2 | Single molecule FISH techniques for research application in single cells. Schematic drawings of single molecule FISH methods: **(A)** Cas-9 mediated FISH (CASFISH) using fluorophore-coupled sgRNAs and dCAS9. **(B)** Oligopaint-FISH using fluorophore-coupled primary oligonucleotides for targeted SNP loci and fluorophore-coupled second oligonucleotide to enhance labeling efficiency shows differential labeling of paternal (pat) and maternal (mat) chromosomes. **(C)** single molecule RNA FISH by rolling cycling amplification (RCA) using padlock probes targeting to reverse transcripted cDNA with different alleles followed by ligation, cycling amplification and specific fluorophore-couple probe hybridization and visualization. **(D)** Sequential barcoding of multiplex different mRNAs by repeat rounds of hybridization, imaging, and stripping. Star, diamond, and triangle are symbols for different fluorophores.

is needed before its application for basic research and genetic diagnosis.

A synthesized primary single-strand oligonucleotide library targeting to a single copy region of the genome along with fluorophore-coupled second oligonucleotides complementary to a portion of the primary oligonucleotides were developed for so-called oligopaint FISH (Beliveau et al., 2015). Co-hybridization of a set of hundreds to thousands of primary fluorophore-coupled oligopaint probes (30–42 bases in length for targeted genome region and hinged 14–32 bases for second oligonucleotides) with fluorophore-coupled second oligonucleotide (14–32 bases) can visualize a 52 Kb–3 Mb regions in nuclei with a 96–100% hybridization efficiency. Oligopaint FISH probes designed with one fluorophore for specified single nucleotide polymorphisms (SNPs) in a targeted region from one chromosome and another fluorophore for these SNPs in the homology chromosome enabled differential labeling of the two homologous chromosomes. Stochastic optical reconstruction microscope (STORM) was used for single-molecule super-resolution imaging. Therefore, with prior information of the specific SNP alleles from the two homologous chromosomes, oligopaint FISH showed *in situ* haplotyping for paternal and maternal chromosomes (**Figure 2B**). The oligopaint probes are chosen bioinformatically to avoid repetitive DNA sequences and they can be selected to target any organisms whose genomes have been sequenced. With further improvement on signal pattern recognition from the SNP loci, oligopaint FISH should enable direct analysis of fine-scale chromatin structure, differential visualization of homologous chromosomes, and allele-specific studies of gene expression.

RNA FISH is a cell-based technique for detecting mRNA transcripts. With the advance of various methods for signal amplification and super-resolution imaging, single molecule RNA FISH (smRNA-FISH) techniques have been developed. Several approaches, including branched DNA probes, tyramide signal amplification, quantum dots, and padlock-rolling circle amplification (RCA), have been used for signal enhancement (Kwon, 2013). RCA is the only method capable of distinguishing single nucleotide allelic changes in transcripts. Briefly, reverse transcription was performed *in situ* on cells and tissue sections to

generate complementary DNA (cDNA), the mRNA was degraded by ribonuclease H, and then padlock probes were hybridized to targeted cDNA with 5′ and 3′ arms circularized by a T4 DNA ligase. The circularized padlock probes served as a template for RCA by Φ29 DNA polymerase, and then fluorophore-couple oligonucleotide probes specific for each padlock probe could be hybridized and visualized (**Figure 2C**; Larsson et al., 2010). To increase the capacity for multiplex detection of different mRNA molecules in single cells, combinatorial labeling, and optical super-resolution microscope were used to measure mRNA levels of 32 genes simultaneously in single *Saccharomyces cerevisiae* cells (Lubeck and Cai, 2012). Further modification introduced a sequential barcoding scheme for multiplex different mRNA quantitation (Lubeck et al., 2014). In this scheme, the mRNAs in cells were barcoded by sequential rounds of hybridization, imaging and probe stripping (**Figure 2D**). Theoretically, the multiplexing capacity scaled up quickly as the number of fluorophores and rounds of hybridization increased. In practice, the available fluorophores were limited and each round of hybridization introduced loss of the RNA integrity in the tested cells.

Various smRNA-FISH methods have been used in imaging cell-type specific RNA profiles and sub-cellular localization patterns of mRNAs in *in vitro* cellular systems (Ronander et al., 2012; Lalmansingh et al., 2013; Shaffer et al., 2013; Sinnamon and Czaplinski, 2014) and model animals such as *Drosophila* (Zimmerman et al., 2013), *Caenorhabditis elegans* (Bolková and Lanctôt, 2015), and Zebrafish (Hauptmann et al., 2016). Additionally, smRNA FISH has been used to study the subcellular localization and cell-to-cell variability of long non-coding RNAs (lncRNA); systematically quantification and categorization based on the subcellular localization patterns were achieved for a representative set of 61 lncRNAs in three different cell types (Cabili et al., 2015). Knowledge of lncRNA subcellular localization patterns is essential to understand its biological processes. An interesting application of smRNA FISH is the study on nuclear RNA foci in genetic diseases resulting from the expansion of tri-, tetra-, penta-, and hexa-nucleotide repeats; a detailed protocol was reported for detecting mRNAs containing expanded CAG and CUG repeats in fibroblasts, lymphoblasts, and induced pluripotent stem cells (Urbanek and Krzyzosiak, 2016).

Simultaneous detection of mRNA and protein quantity and their subcellular distribution in single cells by combining a RNase-free modification of the immunofluorescence (IF) technique and the smRNA FISH method observed direct interaction of RNase MCPIP1 with IL-6 mRNA (Kochan et al., 2015). Real-time live imaging using laser-scanning confocal microscope with photon-counting detectors for quantitative studies of transcription in culture cells and model animals have been achieved by smRNA-FISH and GFP-tagged reporter gene for RNA polymerase (Gregor et al., 2014). Using *Drosophila* embryo as a testing system, smRNA-FISH observed stochastic transcriptional activity of four critical patterning genes and co-packaging of transcripts as multi-copy heterogeneous granules to selected subcellular domains (Little et al., 2013, 2015). These results indicated that there are

TABLE 2 | FISH applications in genetic diagnosis and research.

Genetics diagnosis	References	Research applications	References
Constitutional chromosomal and genomic abnormalities		**Analysis complex chromosomal rearrangements**	
Rapid screening of common aneuploidies	Ried et al., 1992	Mapping breakpoints and genomic orientation	Donnelly et al., 2010
Detection of microdeletion/microduplication syndromes	Wei et al., 2013	The study of 3D chromosomal structures	Dupont et al., 2012
Characterization of subtelomeric rearrangements	Ning et al., 1996	Define complex rearrangements	Mackinnon and Campbell, 2013
Analysis of supernumerary marker and ring chromosomes	Zhang et al., 2012	**Characterizing nuclear genomic structures**	
Somatic recurrent chromosomal abnormalities		Spatiotemporal organization of centromeres/telomeres	Chen et al., 2013
Detection of translocations, deletions, duplications/amplifications	Hu et al., 2014	Chromatin interaction during cell cycle	Deng et al., 2015
Monitoring disease progression and clonal evolution	Mikhail et al., 2016	*in situ chromosome haplotyping*	Beliveau et al., 2015
Assessment of sex-mismatch bone marrow transplantation	Liehr et al., 2015	**Profiling RNA transcription and localization**	
Infectious diseases		Quantitation of multiplex mRNAs in single cells	Lubeck et al., 2014
Detection of malaria by 16s rRNA	Shah et al., 2015	Subcellular localization of mRNAs and non-coding RNAs	Cabili et al., 2015

conserved mechanisms of precision mRNA transcription and localization for spatiotemporal control of protein synthesis in regulating cellular and embryo development.

CONCLUSIONS AND FUTURE DIRECTIONS

In summary, FISH has a wide spectrum of diagnostic and research applications as shown in **Table 2**. FISH has the advantage that it can be used in metaphase chromosomes and interphase nuclei, and thus offers a cell-based genetic diagnosis in complementary to DNA-based molecular testing (Xu and Li, 2013). FISH has been used as adjunctive and diagnostic assays for both constitutional and somatic cytogenomic abnormalities. FISH analysis of uncultured interphase cells from amniotic fluid or chorionic villus samples is a standard procedure for rapid prenatal testing of common aneuploidy and genomic disorders, which alleviates much anxiety for patients and physicians. The use of interphase FISH has been particularly fruitful for cancer cytogenetics, where the detection of recurrent chromosomal abnormalities and clonal evolution is crucial for classifying different types of tumors, selecting treatment protocols, and monitoring outcomes. Even with the introduction of genomic technologies like microarray analysis and exome sequencing, FISH analysis will still be an integral part of genetic diagnosis (Parisi et al., 2012; Wei et al., 2013; Martin and Warburton, 2015). Microfluidic devices for miniaturized and automatic FISH applications are currently under development (Vedarethinam et al., 2010; Kwasny et al., 2012; Kao et al., 2015). The validation of these devices in the near future and the available of more disease-specific probes will further enhance and expand the diagnostic FISH application.

Novel FISH techniques and super-resolution imaging systems have been introduced to study the spatiotemporal changes of intra-nuclear genomic organization and cytoplasmic RNA profiling. These FISH techniques such as CASFISH, oligopaint-FISH, and smRNA-FISH have been developed mainly for genetic research applications. A current trend in FISH is toward simultaneous single-cell measurement of DNA, RNA, cell surface proteins, and intracellular proteins (Lai et al., 2016; Soh et al., 2016). The translation of these single molecule single cells FISH techniques into cell-based genetic diagnosis is expected to improve the analytical resolution and capacity for a spectrum of genetic defects from chromosomal and genomic abnormalities to epigenetic aberrations.

AUTHOR CONTRIBUTIONS

CC drafted the cell-based genetics diagnosis by FISH. WS drafted the single cell DNA structural and RNA transcriptional analysis by FISH. PL organized, modified, and edited the manuscript. We would like to thank Audrey Meusel for proofreading and editing this manuscript.

ACKNOWLEDGMENTS

Funding from China Scholarship Council to WS (Project No. 201308455012) supported part of this study.

REFERENCES

Beliveau, B. J., Boettiger, A. N., Avendaño, M. S., Jungmann, R., McCole, R. B., Joyce, E. F., et al. (2015). Single-molecule super-resolution imaging of chromosomes and *in situ* haplotypevisualization using oligopaint FISH probes. *Nat. Commun.* 6, 7147. doi: 10.1038/ncomms8147

Boersma-Vreugdenhil, G. R., Peeters, T., Bast, B. J., and Lokhorse, H. M. (2003). Translocation of the IgH locus is nearly ubiquitous in multiple myeloma as detected by immune-FISH. *Blood* 101:1653. doi: 10.1182/blood-2002-09-2968

Bolková, J., and Lanctôt, C. (2015). Quantitative gene expression analysis in *Caenorhabditis elegans* using single molecule RNA FISH. *Methods* 98, 42–49. doi: 10.1016/j.ymeth.2015.11.008

Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A., et al. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* 16:20. doi: 10.1186/s13059-015-0586-4

Castronovo, C., Valtorta, E., Crippa, M., Tedoldi, S., Romitti, L., Amione, M. C., et al. (2013). Design and validation of a pericentromeric BAC clone set aimed at improving diagnosis and phenotype prediction of supernumerary marker chromosomes. *Mol. Cytogenet.* 6:45. doi: 10.1186/1755-8166-6-45

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer. J., Zhang, W., Li, G. W., et al. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 155, 1479–1491. doi: 10.1016/j.cell.2013.12.001

Ciolino, A. L., Tang, M. E., and Bryant, R. (2009). Statistical treatment of fluorescence *in situ* hybridization validation data to generate normal reference ranges using Excel functions. *J. Mol. Diagn.* 11, 330–333. doi: 10.2353/jmoldx.2009.080101

DeLong, E. F., Wickham, G. S., and Pace, N. R. (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science* 243, 1360–1363. doi: 10.1126/science.2466341

Deng, W., Shi, X., Tjian, R., Lionnet, T., and Singer, R. H. (2015). CASFISH: CRISPR/Cas9-mediated *in situ* labeling of genomic loci in fixed cells. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11870–11875. doi: 10.1073/pnas.1515692112

De Mas, P., Daudin, M., Vincent, M. C., Bourrouillou, G., Calvas, P., Mieusset, R., et al. (2001). Increased aneuploidy in spermatozoa from testicular tumour patients after chemotherapy with cisplatin, etoposide and bleomycin. *Hum. Reprod.* 16, 1204–1208. doi: 10.1093/humrep/16.6.1204

Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S. Q., Kajuna, S. L. B., et al. (2010). The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am. J. Hum. Genet.* 86, 161–171. doi: 10.1016/j.ajhg.2010.01.007

Dupont, C., Bucourt, M., Guimiot, F., Kraoua, L., Smiljkovski, D., Le Tessier, D., et al. (2014). 3D-FISH analysis reveals chromatid cohesion defect during interphase in Roberts syndrome. *Mol. Cytogenet.* 30:7. doi: 10.1186/s13039-014-0059-6

Dupont, C., Guimiot, F., Perrin, L., Marey, I., Smiljkovski, D., Le Tessier, D., et al. (2012). 3D position of pericentromeric heterochromatin within the nucleus of a patient with ICF syndrome. *Clin. Genet.* 82, 187–192. doi: 10.1111/j.1399-0004.2011.01697.x

Gregor, T., Garcia, H. G., and Little, S. C. (2014). The embryo as a laboratory: quantifying transcription in *Drosophila*. *Trends Genet.* 30, 364–375. doi: 10.1016/j.tig.2014.06.002

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hauptmann, G., Lauter, G., and Söll, I. (2016). Detection and signal amplification in zebrafish RNA FISH. *Methods* 98, 50–59. doi: 10.1016/j.ymeth.2016.01.012

Hicks, D. G., Raymond, R., and Tybbs, R. R. (2005). Assessment of the HER2 status in breast cancer by fluorescence *in situ* hybridization: a technical review with interpretive guidelines. *Hum. Pathol.* 36, 250–261. doi: 10.1016/j.humpath.2004.11.010

Hu, L., Ru, K., Zhang, L., Huang, Y., Zhu, X., Liu, H., et al. (2014). Fluorescence *in situ* hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomark. Res.* 2:3. doi: 10.1186/2050-7771-2-3

Kamath, A., Tara, H., Xiang, B., Bajaj, R., He, W., and Li, P. (2008). Double minute MYC amplification and deletion of MTAP, CDKN2A, CDKN2B and ELAVL2 in an acute myeloid leukemia characterized by oligonucleotide-array comparative genomic hybridization. *Cancer Genet. Cytogenet.* 183, 117–120. doi: 10.1016/j.cancergencyto.2008.02.011

Kao, K.-J., Tai, C.-H., Chang, W.-H., Yeh, T.-S., Chen, T.-C., and Lee, G.-B. (2015). A fluorescence *in situ* hybridization (FISH) microfluidic platform for detection of *HER2* amplification in cancer cells. *Biosens. Bioelectorn.* 69, 272–279. doi: 10.1016/j.bios.2015.03.003

Khattab, M., Xu, F., Li, P., and Bhandari, V. (2011). A *de novo* 3.54 Mb deletion of 17q22-q23.1 associated with hydrocephalus: a case report and review of literature. *Am. J. Med. Genet. A* 155, 3082–3086. doi: 10.1002/ajmg.a.34307

Klinger, K., Landes, G., Shook, D., Harvey, R., Lopez, L., Locke, P., et al. (1992). Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence *in situ* hybridization (FISH). *Am. J. Hum. Genet.* 51, 55–65.

Kochan, J., Wawro, M., and Kasza, A. (2015). Simultaneous detection of mRNA and protein in single cells using immunofluorescence-combined single-molecule RNA FISH. *Biotechniques* 59, 209–212. doi: 10.2144/000114340

Kwasny, D., Vedarethinam, I., Shah, P., Dimaki, M., Silahtaroglu, A., Tumer, Z., et al. (2012). Advanced microtechnologies for detection of chromosome abnormalities by fluorescent *in situ* hybridization. *Biomed. Microdevices* 14, 453–460. doi: 10.1007/s10544-011-9622-7

Kwon, S. (2013). Single-molecule fluorescence *in situ* hybridization: quantitative imaging of single RNA molecules. *BMB Rep.* 46, 65–72. doi: 10.5483/BMBRep.2013.46.2.016

Lai, L. T., Meng, Z., Shao, F., and Zhang, L. F. (2016). Simultaneous RNA-DNA FISH. *Methods Mol. Biol.* 1402, 135–145. doi: 10.1007/978-1-4939-3378-5_11

Lalmansingh, A. S., Arora, K., Demarco, R. A., Hager, G. L., and Nagaich, A. K. (2013). High throughput RNA FISH analysis by imaging flow cytometry reveals that pioneer factor Foxa1 reduces transcriptional stochasticity. *PLoS ONE* 8:e76043. doi: 10.1371/journal.pone.0076043

Langer-Safer, P. R., Levine, M., and Ward, D. C. (1982). Immunological method for mapping geneson *Drosophila* polytene chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* 79, 4381–5438. doi: 10.1146/annurev-genom-090413-025346

Larsson, C., Grundberg, I., Söderberg, O., and Nilsson, M. (2010). *In situ* detection and genotypingof individual mRNA molecules. *Nat. Methods* 7, 395–397. doi: 10.1038/nmeth.1448

Li, M. M., Ewton, A. A., and Smith, J. L. (2013). Using cytogenetic rearrangements for cancer prognosis and treatment (pharmacogenetics). *Curr. Genet. Med. Rep.* 1, 99–112. doi: 10.1007/s40142-013-0011-9

Li, P., and Cui, C. H. (2016). A broader view of cancer cytogenetics: from nuclear aberrations to cytogenomic abnormalities. *J. Mol. Genet. Med.* 10:e108. doi: 10.4172/1747-0862.1000E108

Li, P., Pomianowski, P., DiMaio, S. M., Stanis, J. R., Rossi, M. R., Xiang, B., et al. (2011). Genomic characterization of prenatally detected chromosomal structural abnormalities using oligonucleotide array comparative genomic hybridization. *Am. J. Med. Genet. A* 155, 1605–1615. doi: 10.1002/ajmg.a.34043

Li, P., Xu, F., and Shu, W. (2015). The spectrum of cytogenomic abnormalities in patients with developmental delay and intellectual disabilities. *N. Am. J. Med. Sci.* 8, 166–172. doi: 10.7156/najms.2015.0804172

Li, P., Zhang, H. Z., Huff, S., Nimmakayalu, M., Qumsiyeh, M., Yu, J. W., et al. (2006). Karyotype-phenotype insights from 11q14.1-q23.2 interstitial deletions: FZD4 haploinsufficiency and exudative vitroretinopathy in a patient with a complex chromosome rearrangement. *Am. J. Med. Genet.* 140, 2721–2729. doi: 10.1002/ajmg.a.31498

Lichter, J. B., Difilippantonio, M. J., Pakstis, A. J., Goodfellow, P. J., Ward, D. C., and Kidd, K. K. (1993). Physical and genetic maps for chromosome 10. *Genomics* 16, 320–324. doi: 10.1006/geno.1993.1192

Liehr, T., Claussen, U., and Starke, H. (2004). Small supernumerary marker chromosomes (sSMC) in humans. *Cytogenet. Genome Res.* 107, 55–67. doi: 10.1159/000079572

Liehr, T., Mrasek, K., Weise, A., Dufke, A., Rodriguez, L., Guardia, N. M., et al. (2006). Small supernumerary marker chromosomes-progress towards a genotype-phenotype correlation. *Cytogenet. Genome Res.* 112, 23–34. doi: 10.1159/000087510

Liehr, T., Othman, M. A. K., Rittscher, K., and Alhourani, E. (2015). The current state of molecular cytogenetics in cancer diagnosis. *Expert Rev. Mol. Diagn.* 15, 517–526. doi: 10.1586/14737159.2015.1013032

Little, S. C., Sinsimer, K. S., Lee, J. J., Wieschaus, E. F., and Gavis, E. R. (2015). Independent and coordinate trafficking of single *Drosophila* germ plasm mRNAs. *Nat. Cell Biol.* 17, 558–568. doi: 10.1038/ncb3143

Little, S. C., Tikhonov, M., and Gregor, T. (2013). Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* 154, 789–800. doi: 10.1016/j.cell.2013.07.025

Lubeck, E., and Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* 9, 743–748. doi: 10.1038/nmeth.2069

Lubeck, E., Coskum, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods.* 11, 360–361. doi: 10.1038/nmeth.2892

Mackinnon, R. N., and Campbell, L. J. (2013). Chromothripsis under the microscope: a cytogenetic perspective of two cases of AML with catastrophic chromosome rearrangement. *Cancer Genet.* 206, 238–251. doi: 10.1016/j.cancergen.2013.05.021

Marino-Enriquez, A., Li, P., Samuelson, J., Rossi, M. R., and Reyes-Mugica, M. (2008). Congenital fibrosarcoma with a novel complex three way translocation t(12;15;19) and unusual histological features. *Hum. Pathol.* 39, 1844–1848. doi: 10.1016/j.humpath.2008.04.013

Martin, C. L., and Warburton, D. (2015). Detection of chromosomal aberrations in clinical practice: from karyotype to genome sequence. *Ann. Rev. Genomics Hum. Genet.* 16, 309–326. doi: 10.1146/annurev-genom-090413-025346

Massaro, S. A., Bajaj, R., Pashankar, F. D., Ornstein, D., Gallager, P. G., Krause, D. S., et al. (2011). Bi-allelic deletions within 13q14 and transient trisomy 21 with absence of GATA1s in pediatric acute megakaryoblastic leukemia. *Ped. Blood Cancer* 57, 516–519. doi: 10.1002/pbc.23156

Meng, J., Matarese, C., Crivello, J., Wilcox, K., Wang, D., DiAdamo, A., et al. (2015). Changes in and efficacies of indications for invasive prenatal diagnosis of cytogenomic abnormalities: 13 years of experience in a single center. *Med. Sci. Monit.* 21, 1942–1948. doi: 10.12659/MSM.893870

Mikhail, F. M., Heerema, N. A., Rao, K. W., Burnside, R. D., Cherry, A. M., and Cooley, L. D. (2016). Section E6.1-6.4 of the ACMG technical standards and guidelines: chromosome studies of neoplastic blood and bone marrow-acquired chromosomal abnormalities. *Genet. Med.* 18, 635–642. doi: 10.1038/gim.2016.50

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* 86, 749–764. doi: 10.1016/j.ajhg.2010.04.006

Mitelman, F., Johansson, B., and Metens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* 7, 233–245. doi: 10.1038/nrc2091

Morrison, L. E., Ramakrishnan, R., Ruffalo, T. M., and Wilber, K. A. (2003). "Labeling fluorescence *in situ* hybridization probes for genomic targets," in *Molecular Cytogenetics Protocols and Applications,* ed Y.-S. Fan *(Totowa, NJ: Humana Press Inc.),* 21–40.

Ning, Y., Roschke, A., Smith, A. C. M., Macha, M., Precht, K., Riethman, H., et al. (1996). A complete set of human telomeric probes and their clinical application. *Nat. Genet.* 14, 86–89. doi: 10.1038/ng0996-86

Parisi, F., Ariyan, S., Narayan, D., Bacchiocchi, A., Hoyt, K., Cheng, E., et al. (2011). Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. *BMC Genomics* 12:230. doi: 10.1186/1471-2164-12-230

Parisi, F., Micsinai, M., Strino, F., Ariyan, S., Narayan, D., Bacchiocchi, A., et al. (2012). Integrated analysis of tumor samples sheds light on tumor heterogeneity. *Yale J. Biol. Med.* 85, 347–361.

Ried, T., Landes, G., Dackowski, W., Klinger, K., and Ward, D. C. (1992). Multicolor fluorescence *in situ* hubridization for the simultaneous detection of probe sets for chromosomes 13, 18, X and Y in uncultured amniotic fluid cells. *Hum. Mol. Genet.* 1, 307–313.

Ronander, E., Bengtsson, D. C., Joergensen, L., Jensen, A. T., and Arnot, D. E. (2012). Analysis of single-cell gene transcription by RNA fluorescent *in situ* hybridization (FISH). *J. Vis. Exp.* 68:4073. doi: 10.3791/4073

Rossi, M. R., DiMaio, M., Xiang, B., Lu, K., Hande, K., Seashore, G., et al. (2009). Clinical and genomic characterization of distal duplications and deletions of chromosome 4q: study of two cases and review of the literature. *Am. J. Med. Genet. A* 149, 2788–2794. doi: 10.1002/ajmg.a.33088

Shaffer, S. M., Wu, M. T., Levesque, M. J., and Raj, A. (2013). Turbo FISH: a method for rapid single molecule RNA FISH. *PLoS ONE* 8:e75120. doi: 10.1371/journal.pone.0075120

Shah, J., Mark, O., Weltman, H., Barcelo, H., Lo, W., Wronska, D., et al. (2015). Fluorescence *in situ* hybridization (FISH) assays for diagnosing malaria in endemic areas. *PLoS ONE* 10:e0136726. doi: 10.1371/journal.pone.0136726

Sinnamon, J. R., and Czaplinski, K. (2014). RNA detection *in situ* with FISH-STICs. *RNA* 20, 260–266. doi: 10.1261/rna.041905.113

Soh, K. T., Tario, J. D. Jr., Colligan, S., Maguire, O., and Pan, D., Minderman, H., et al. (2016). Simultaneous, single-cell measurement of messenger RNA, cell surface proteins, and intracellular proteins. *Curr. Protoc. Cytom.* 75, 7.45.1–7.45.33. doi: 10.1002/0471142956.cy0745s75

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40. doi: 10.1016/j.cell.2010.11.055

Tempest, H. G., Ko, E., Chan, P., Robaire, B., Rademaker, A., and Martin, R. H. (2008). Sperm aneuploidy frequencies analysed before and after chemotherapy in testicular cancer and Hodgkin's lymphoma patients. *Hum. Reprod.* 23, 251–258. doi: 10.1093/humrep/dem389

Test and Technology Transfer Committee. (2000). Technical and clinical assessment of fluorescence *in situ* hybridization: an ACMG/ASHG position statement. I. Technical considerations. *Genet. Med.* 2, 356–361. doi: 10.1097/00125817-200011000-00011

Urbanek, M. O., and Krzyzosiak, W. J. (2016). RNA FISH for detecting expanded repeats in human diseases. *Methods* 98, 115–123. doi: 10.1016/j.ymeth.2015.11.017

Vedarethinam, I., Shah, P., Dimaki, M., Tumer, Z., Tommerup, N., and Svendsen, W. E. (2010). Metaphase FISH on a chip: miniaturized microfluidic device for fluorescence *in situ* hybridization. *Sensors* 10, 9831–9846. doi: 10.3390/s101109831

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799. doi: 10.1038/nm1087

Wang, Q., Wu, W., Xu, Z., Luo, F., Zhou, Q., Li, P., et al. (2015). Copy number changes and methylation patterns in an isodicentric and a ring chromosome of 15q11-q13: report of two cases and review of literature. *Mol. Cytogenet.* 8:97. doi: 10.1186/s13039-015-0198-4

Ward, B. E., Gersen, S. L., Carelli, M. P., McGuire, N. M., Dackowski, W. R., Weinstein, M., et al. (1993). Rapid prenatal diagnosis of chromosomal aneuploidies by fluorescence *in situ* hybridization: clinical experience with 4,500 specimens. *Am. J. Hum. Genet.* 52, 854–865.

Wei, Y., Xu, F., and Li, P. (2013). Technology-driven and evidence-based genomic analysis for integrated pediatric and prenatal genetic evaluation. *J. Genet. Genomics* 40, 1–14. doi: 10.1016/j.jgg.2012.12.004

Wolff, D. J., Bagg, A., Cooley, L. D., Dewald, G. W., Hirsch, B. A., and Jacky, P. B. (2007). Guidance for fluorescence *in situ* hybridization testing in hematologic disorders. *J. Mol. Diagn.* 9, 134–143. doi: 10.2353/jmoldx.2007.060128

Xu, F., DiAdamo, A. J., Grommisch, B., and Li, P. (2013). Interstitial duplication and distal deletion in a ring chromosome 13 with pulmonary atresia and ventricular septal defect: a case report and review of the literature. *N. Am. J. Med. Sci.* 6, 208–212. doi: 10.7156/najms.2013.0604208

Xu, F., Li, L., Schulz, V. P., Gallager, P. G., Xiang, B., Zhao, H. Y., et al. (2014). Cytogenomic mapping and bioinformatic mining reveal interacting brain expressed genes for intellectural disabilities. *Mol. Cytogenet.* 7:4. doi: 10.1186/1755-8166-7-4

Xu, F., and Li, P. (2013). "Cytogenomic abnormalities and dosage-sensitive mechanisms for intellectual and developmental disabilities," in *Developmental Disabilities—Molecules Involved, Diagnosis and Clinical Care*, ed A. Salehi (Rijeka: InTech), 1–30.

Xu, Z. Y., Xie, J. S., Meng, J. L., Li, P., Pan, X. H., and Zhou, Q. H. (2013). Non-invasive prenatal diagnosis: a comparison of cell free fetal DNA (cffDNA) based screening and fetal nucleated red blood cell (fnRBC) initiated testing. *N. Am. J. Med. Sci.* 6, 194–199. doi: 10.7156/najms.2013.0604194

Zhang, H. Z., Li, P., Wang, D., Huff, S., Nimmakayalu, M., Qumsiyeh, M., et al. (2004). *FOXC1* gene deletion is associated with eye anomalies in ring chromosome 6. *Am. J. Med. Genet.* 124, 280–287. doi: 10.1002/ajmg.a.20413

Zhang, H. Z., Xu, F., Seashore, M., and Li, P. (2012). Unique genomic structure and distinct mitotic behavior of ring chromosome 21 in two unrelated cases. *Cytogenet. Genome Res.* 136, 180–187. doi: 10.1159/000336978

Zhou, Q., Wu, S. Y., Amato, K., Diadamo, A., and Li, P. (2016). Spectrum of cytogenomic abnormalities revealed by array comparative genomic hybridization in products of conception culture failure and normal karyotype samples. *J. Genet. Genomics.* 43, 121–131. doi: 10.1016/j.jgg.2016.02.002

Zimmerman, S. G., Peters, N. C., Altaras, A. E., and Berg, C. A. (2013). Optimized RNA ISH, RNA FISH and protein-RNA double labeling (IF/FISH) in *Drosophila* ovaries. *Nat. Protoc.* 8, 2158–2179. doi: 10.1038/nprot.2013.136

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Cell and Developmental Biology

METHODS
published: 11 April 2018
doi: 10.3389/fcell.2018.00042

Single-Cell *in Situ* RNA Analysis With Switchable Fluorescent Oligonucleotides

*Lu Xiao and Jia Guo**

Biodesign Institute and School of Molecular Sciences, Arizona State University, Tempe, AZ, United States

Comprehensive RNA analyses in individual cells in their native spatial contexts promise to transform our understanding of normal physiology and disease pathogenesis. Here we report a single-cell *in situ* RNA analysis approach using switchable fluorescent oligonucleotides (SFO). In this method, transcripts are first hybridized by pre-decoding oligonucleotides. These oligonucleotides subsequently recruit SFO to stain their corresponding RNA targets. After fluorescence imaging, all the SFO in the whole specimen are simultaneously removed by DNA strand displacement reactions. Through continuous cycles of target staining, fluorescence imaging, and SFO removal, a large number of different transcripts can be identified by unique fluorophore sequences and visualized at the optical resolution. To demonstrate the feasibility of this approach, we show that the hybridized SFO can be efficiently stripped by strand displacement reactions within 30 min. We also demonstrate that this SFO removal process maintains the integrity of the RNA targets and the pre-decoding oligonucleotides, and keeps them hybridized. Applying this approach, we show that transcripts can be restained in at least eight hybridization cycles with high analysis accuracy, which theoretically would enable the whole transcriptome to be quantified at the single molecule sensitivity in individual cells. This *in situ* RNA analysis technology will have wide applications in systems biology, molecular diagnosis, and targeted therapies.

Keywords: transcriptomics, genomics, fluorescence *in situ* hybridization, strand displacement reactions, RNA expression, oligonucleotides, fluorescent probes, single-cell

OPEN ACCESS

Edited by:
Xinghua Victor Pan,
Yale University, United States

Reviewed by:
Jeffrey C. Petruska,
University of Louisville, United States
Saurabh Chattopadhyay,
University of Toledo, United States

****Correspondence:***
Jia Guo
jiaguo@asu.edu

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *08 December 2017*
Accepted: *26 March 2018*
Published: *11 April 2018*

Citation:
Xiao L and Guo J (2018) Single-Cell in Situ RNA Analysis With Switchable Fluorescent Oligonucleotides. Front. Cell Dev. Biol. 6:42. doi: 10.3389/fcell.2018.00042

INTRODUCTION

The ability to profile a large number of distinct transcripts in single cells *in situ* is crucial for our understanding of cancer, neurobiology, and stem cell biology (Crosetto et al., 2014). The differences between individual cells in complex biological systems may have significant consequences in the function and health of the entire systems. Thus, single cell analysis is required to explore such cell heterogeneity. Due to the inherent complexity of gene expression regulatory networks, comprehensive molecular profiling is required to systematically infer the functions and interactions of different RNA species. The precise location of cells in a tissue and transcripts in a cell is critical for effective cell-cell interactions and gene expression regulation, which can determine cell fates and functions. Therefore, to fully understand the organization, regulation, and function of a heterogeneous biological system, highly multiplexed single-cell *in situ* RNA analysis is critically needed.

Next-generation sequencing (Guo et al., 2010; Metzker, 2010) and microarray technologies (Hoheisel, 2006) have been widely used to study gene expression regulation in health and disease by profiling RNA on a genome-wide scale. However, as transcripts are extracted, purified and then analyzed in these approaches, the RNA location information is lost. Imaging-based methods, such as molecular beacons (Guo et al., 2012; Huang and Martí, 2012), templated fluorescence activation probes (Franzini and Kool, 2009), and fluorescence *in situ* hybridization (FISH) (Raj et al., 2008), allow transcripts to be quantified in their native spatial contexts in single cells. Nonetheless, due to the spectral overlap of commonly available fluorophores, these methods can only detect a handful of different RNA species in one sample.

To enable comprehensive single-cell *in situ* RNA analysis, several approaches have been investigated. For instance, *in situ* sequencing (Ke et al., 2013; Lee et al., 2014) has been explored to enable transcriptome profiling in individual cells. However, this method has limited detection efficiency and may miss low-expression transcripts. Combinatorial labeling (Levsky et al., 2002; Lubeck and Cai, 2012; Levesque and Raj, 2013) and reiterative hybridization (Xiao and Guo, 2015; Guo, 2016; Shaffer et al., 2017; Mondal et al., 2018) offer single-molecule detection sensitivity, but these approaches suffer from limited multiplexing capacities. Recently, sequential hybridization (Lubeck et al., 2014; Shah et al., 2016) and multiplexed error-robust fluorescence *in situ* hybridization (MER-FISH) (Chen et al., 2015; Moffitt et al., 2016a,b) have been developed for highly multiplexed single-molecule RNA detection. In these methods, to stain the same RNA molecules in different analysis cycles, several approaches have been explored to remove the fluorescence signals at the end of each cycle. Such approaches include probe degradation by DNase, photobleaching, and disulfide based chemical cleavage. Nevertheless, probe degradation by DNase is limited by its low signal removal efficiency. In addition, DNase removes all the probes, including the large oligonucleotides library hybridized to their RNA targets. Consequently, this expensive oligonucleotides library has to be re-hybridized in every analysis cycle, which will increase the assay time and cost. Photobleaching erases fluorescence signals in different imaging areas sequentially. As a result, it is less time-effective and has low sample throughput. The disulfide based probes can cross-react with the endogenous thiol groups and the thiol groups generated by fluorophore cleavage in previous cycles, which will lead to high background and false positive signals.

Here, we report a single-cell *in situ* RNA analysis approach using switchable fluorescent oligonucleotides (SFO). In this method, RNA molecules are first hybridized by pre-decoding oligonucleotides, which subsequently recruit SFO to stain their RNA targets. After imaging, SFO are removed by strand displacement reactions. Upon continuous cycles of target staining, fluorescence imaging, and SFO removal, varied RNA species are identified by unique fluorophore sequences at the optical resolution. To demonstrate the feasibility of this approach, we show that the hybridized SFO can be efficiently removed by strand displacement reactions within the cellular environment in 30 min. We also demonstrate that this probe removal process maintains the RNA integrity and keeps the pre-decoding oligonucleotides hybridized to their RNA targets. Additionally, we show that RNA can be quantified with high accuracy in at least eight continuous hybridization cycles, which theoretically would allow the whole transcriptome to be profiled in individual cells *in situ*.

MATERIALS AND METHODS

General Information

Chemicals and solvents were purchased from Sigma-Aldrich or Ambion and were used without further purification, unless otherwise noted. Biogreagents were purchased from Invitrogen, unless otherwise indicated.

Cell Culture

HeLa CCL-2 cells (ATCC) were maintained in Dulbecco's modified Eagle's Medium supplemented with 10% fetal bovine serum, 10 U mL^{-1} penicillin and 100 g mL^{-1} streptomycin in a humidified atmosphere at 37°C with 5% CO_2. Cells were plated on chambered coverglass (Thermo Scientific) and allowed to reach 60% confluency in 1–2 days.

Cell Fixation

Cultured HeLa CCL-2 cells were first washed with 1 X PBS at room temperature for 5 min, fixed with fixation solution [4% formaldehyde (Polusciences) in 1 X PBS] at room temperature for 10 min, and subsequently washed another 2 times with 1 X PBS at room temperature, each for 5 min. The fixed cells were then permeabilized with 70% (v/v) EtOH at 4°C at least overnight.

Probe Design

The pre-decoding probes with a length of 70 nt contain three 20 nt sequences: (i) a target-binding sequence for *in situ* hybridization to the target RNA, and (ii) two repeated readout sequences for decoding hybridization. The three sequences are separated from each other by a flanking 5T spacer. The target-binding sequence was designed by the Stellaris Probe Designer provided by Biosearch Technology. The sequences of pre-decoding probes are provided in **Table S1**.

The decoding probe (SFO) with a length of 40 nt contains two 20 nt sequences: (i) a binding sequence complimentary to the readout sequence of the pre-decoding probes, and (ii) a toehold sequence for strand displacement reactions. The decoding probe is conjugated to fluorophores with the 5′-amino modification. The sequence of the decoding probe is provided in **Table S1**.

The eraser oligonucleotide with a length of 40 nt is complimentary to the decoding probe. The sequence of the eraser oligonucleotide is provided in **Table S1**.

The SFO-orthogonal oligonucleotide with a length of 40 nt is conjugated to fluorophores with the 5′-amino modification. The sequence of the SFO-orthogonal oligonucleotide is provided in Tabl S1.

To further ensure the specificity, all the sequences above were screened against the human transcriptome by using Basic Local Alignment Search Tool (BLAST) (Camacho et al., 2009) to ensure there were no more than 10 nt of homology. Sequence alignment

test were also performed by BLAST within these sequences to ensure there were no more than 8 nt of homology.

Probe Preparation

Pre-decoding oligonucleotides belonging to one library (IDT) were mixed and then stored as pre-decoding probe stock solution (10 mM in 0.01X Tris EDTA, pH 8.0) at 4°C.

The 5′-amino modified decoding probe or the SFO-orthogonal oligonucleotide (IDT), at a scale of 1 nmol, was dissolved in 3 μL of nuclease-free water. To this solution was added sodium bicarbonate aqueous solution (1M, 3 μL) and Cy3 (AAT Bioquest) or Cy5 (AAT Bioquest) in DMF (20 mM, 5 μL). The mixture was incubated at room temperature for 2 h and then purified using a nucleotide removal kit (Qiagen). The fluorophore conjugated oligonucleotides were subsequently purified via an HPLC (Agilent) equipped with a C18 column (Aligent) and a dual wavelength detector set to detect DNA absorption (260 nm) and the fluorophore absorbtion (555 nm for Cy3, 650 nm for Cy5). For the gradient, triethyl ammonium acetate (Buffer A) (0.1 M, pH 6.5) and acetonitrile (Buffer B) (pH 6.5) were used, ranging from 7 to 30% Buffer B over the course of 30 min, then at 70% Buffer B for 10 min followed by 7% Buffer B for another 10 min, all at a flow rate of 1 mL min^{-1}. The collected fraction was then dried in a Savant SpeedVac Concentrator and stored as decoding probe stock solution or SFO-orthogonal oligonucleotide stock solution at 4°C in 100 μL 0.01X Tris EDTA (pH 8.0).

The eraser oligonucleotide was dissolved and stored as displacement stock solution (10 mM in 0.01X Tris EDTA, pH 8.0) at 4°C.

Pre-decoding Hybridization

To 100 μL of pre-decoding hybridization buffer (100 mg mL^{-1} dextran sulfate, 1 mg mL^{-1} Escherichia coli tRNA, 2 mM vanadyl ribonucleoside complex, 20 μg mL^{-1} bovine serum albumin, and 10% formamide in 2 X SSC) was added 1 μL of pre-decoding probe stock solution. Then the mixture was vortexed and centrifuged to obtain pre-decoding hybridization solution.

HeLa CCL-2 cells after fixation and permeabilization were first incubated with wash buffer (2 mM vanadyl ribonucleoside complex and 10% formamide in 2 X SSC) for 5 min at room temperature, then incubated with 100 μL of pre-decoding hybridization solution at 37°C overnight. Cells were then washed three times with wash buffer, each for 30 min, at 37°C.

Cells were then post-fixed with post-fixation solution [4% formaldehyde (Polusciences) in 2X SSC] at room temperature for 10 min, and subsequently washed another three times with 2X SSC at room temperature, each for 5 min.

Decoding Hybridization

To 100 μL of decoding hybridization buffer (100 mg mL^{-1} dextran sulfate, 2 mM vanadyl ribonucleoside complex, and 10% formamide in 2 X SSC) was added 5 μL of decoding probe stock solution with or without 5 μL of SFO-orthogonal oligonucleotide stock solution. Then the mixture was vortexed and centrifuged to obtain decoding hybridization solution.

Cells labeled with pre-decoding probes were directly incubated with 100 μL of decoding hybridization solution at 37°C for 30 min, and washed once with wash buffer at 37°C for 30 min. After incubation with GLOX buffer (0.4% glucose and 10 mM Tris HCl in 2 X SSC) for 1–2 min at room temperature, the stained cells were imaged in GLOX solution (0.37 mg mL^{-1} glucose oxidase and 1% catalase in GLOX buffer).

Displacement of Decoding Probes

To 100 μL of displacement buffer (100 mg mL^{-1} dextran sulfate, 2 mM vanadyl ribonucleoside complex, and 10% formamide in 2 X SSC) was added 5 μL of displacement stock solution. Then the mixture was vortexed and centrifuged to obtain displacement solution.

Cells after imaging were incubated with 100 μL of displacement solution at 37°C for 30 min, and washed 3 times with 1X PBS at 37°C, each for 15 min, then followed by the next cycle of decoding hybridization.

Imaging and Data Analysis

Cells were imaged under a Nikon Ti-E epofluorescence microscope equipped with a 100X objective, using a 5 μm range and 0.3 μm z spacing. Images were captured using a CoolSNAP HQ2 camera and NIS-Elements Imaging software. Chroma filters 49004 and 49009 were used for Quasar 579 and Cy5, respectively.

Fluorescent spots in each hybridization cycle were identified and localized by SpotDetector (Olivo-Marin, 2002). For the detected FISH spots, their intensities in the Cy3 and Cy5 channels were compared to determine the color of the spots. Raw images of the same cells in different cycles of hybridization were aligned to the same coordination system established by the images collected in the first cycle of hybridization based on one specific spot reappearing in each cycle. Spots in the first hybridization cycle with the distance less than 2 pixels (320 nm) to those in the second hybridization cycle were extracted as the barcodes, which corresponded to a potential mRNA molecule. Spots in the following hybridization cycles that shared the distance less than 2 pixels (320 nm) with the barcodes were identified as the reappearance of the barcodes. And the barcode reappearance percentage in each hybridization cycle was then calculated.

RESULTS

Platform Design

In this SFO-based RNA profiling approach (**Figure 1**), individual RNA target is first hybridized by a set of non-fluorescent pre-decoding oligonucleotides with varied target binding sequences. These oligonucleotides also have one or multiple decoding oligonucleotides binding sequences, which can recruit SFO as decoding probes. Each of the subsequent analysis cycles consists of three steps. First, SFO are hybridized to pre-decoding probes to stain the RNA targets. In the second step, fluorescence images are acquired with each RNA molecule visualized as a single spot. Finally, oligonucleotide erasers, which are perfectly complementary to SFO, are applied to remove SFO by strand displacement reactions (Zhang and Seelig, 2011). These oligonucleotide erasers hybridize to the

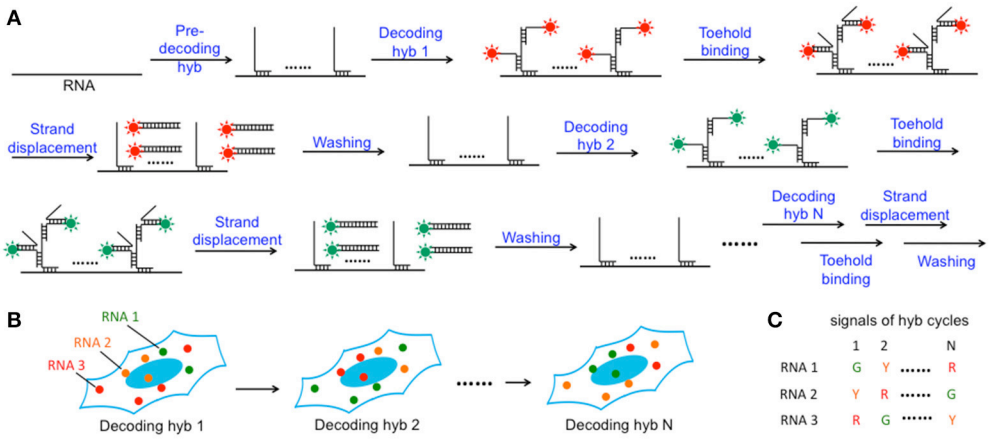


FIGURE 1 | Highly multiplexed single-cell *in situ* RNA analysis with SFO. **(A)** Each transcript is first hybridized with a set of pre-decoding probes, which have varied target-binding sequences to hybridize to the different regions on the target RNA and the shared decoding sequence to recruit SFO as decoding probes. After imaging, the hybridized SFO is removed by strand displacement reactions. Through reiterative cycles of SFO hybridization, fluorescence imaging and strand displacement, the target RNA is sequentially stained by a set of SFO labeled with varied fluorophores. **(B)** Schematic diagram of the N cycles of hybridization images. In each cycle, individual transcript is visualized as a single spot with a specific color. **(C)** As RNA molecules remain in place during different hybridization cycles, different RNA species can be identified by the unique color sequences.

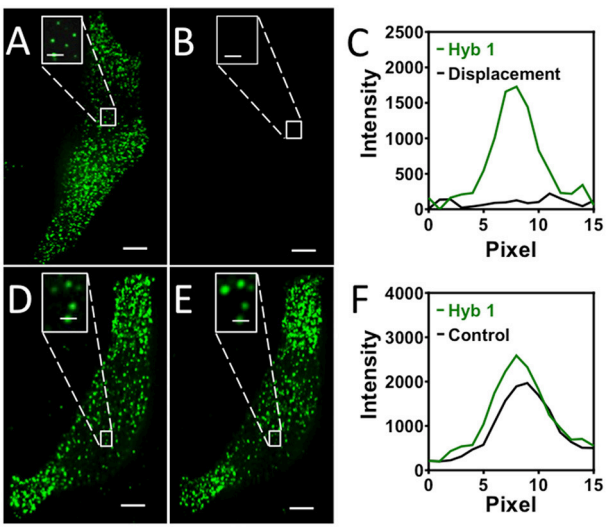


FIGURE 2 | (A) GAPDH transcripts are stained by Cy3 labeled SFO. **(B)** SFO is removed by the eraser oligonucleotide. **(C)** Signal intensity profiles corresponding to the marked FISH spot in **(A,B)**. **(D)** GAPDH transcripts are stained by Cy3 labeled SFO. **(E)** The stained cells are incubated with an orthogonal oligonucleotide. **(F)** Signal intensity profiles corresponding to the marked FISH spot in **(D,E)**. Scale bars, 5 μm.

toehold on SFO, branch migrate and release SFO from the pre-decoding probes. Through reiterative cycles of target staining, fluorescence imaging and SFO release, each transcript is identified by a fluorescence sequence barcode. With M fluorophores applied in each cycle and N sequential cycles, a total of M^N RNA species can be quantified in single cells *in situ*.

SFO Removal Efficiency

One requirement for the success of this SFO-based RNA profiling technology is that fluorescent decoding probes need to be removed very efficiently at the end of each analysis cycle. In this way, the minimized fluorescence signal leftover will not lead to false positive signals in the subsequent cycles. Additionally, the efficient removal of SFO will regenerate the single-stranded SFO-binding sequences on pre-decoding probes, so that SFO can be recruited in the following cycle to stain the target RNA again. To assess the SFO stripping efficiency, we stained mRNA GAPDH with Cy3 labeled decoding probes (**Figure 2A**). After incubating the stained cells with the oligonucleotide eraser for 30 min at 37°C, almost all the original FISH spots become undetectable (**Figures 2B,C**). We also performed control experiments by incubating the stained cells with an SFO-orthogonal oligonucleotide (**Figure 2D**). The fluorescence intensities of the Cy3 stained GAPDH remained largely the same before and after the oligonucleotide incubation (**Figures 2E,F**). These results indicate that SFO can be efficiently removed by strand displacement reactions.

Effects of the Strand Displacement Reactions

Another requirement for the success of this SFO-based approach is that the strand displacement reactions should maintain the RNA integrity, so that the same transcripts can be restained in the subsequent cycles. Additionally, it is preferred to keep the pre-decoding probes hybridized to their RNA targets throughout the assay, rather than to apply them in every analysis cycle. This is essential for the following reasons. First, due to the theoretical hybridization efficiency of ∼75% (Lubeck and Cai, 2012), a small percentage of transcripts are not hybridized with enough pre-decoding probes to make them detectable. And

these undetectable RNA can be different transcripts in different analysis cycles, if the pre-decoding probes are removed and rehybridized in each cycle. Consequently, many missing spots in the aligned fluorophore sequences will be generated, leading to the increased error rate. Furthermore, as the hybridization of the pre-decoding probes takes overnight to 36 h, it is time-consuming to apply this step in each cycle. Finally, for highly multiplexed RNA profiling, the pre-decoding probes library is usually composed of thousands of oligonucleotides. Thus, it will make the assay less cost-effective if the expensive pre-decoding library is removed and re-hybridized in every cycle.

To assess the effects of the strand displacement reactions on the RNA targets and the hybridized pre-decoding probes, we stained mRNA GAPDH in three continuous hybridization cycles (**Figure 3**). In each cycle, Cy3 or Cy5 labeled SFO were applied to stain the transcripts, and were subsequently removed very efficiently using the same oligonucleotide eraser. We counted 1032 and 1045 spots in the first and second cycle, respectively. Among these spots, 803 spots were colocalized. These results are consistent with the ones obtained by using two sets of different colored FISH probes to stain the same transcripts (Raj et al., 2008). The small fraction of spots that did not colocalize may correspond to the non-specifically bound probes. To exclude these off-target signals, we define only the spots colocalized in the first two cycles as true mRNA signals. With our approach, 99% of the true signals reappeared in the third cycle. In comparison, when both pre-decoding and decoding probes are degraded using DNase, only 78% of spots reoccur in the third cycle (Lubeck et al., 2014). These results suggest that the DNA displacement reactions do not damage the RNA integrity, and the pre-decoding probes remain hybridized to their RNA targets throughout the assay. In this way, the analysis accuracy is improved and the assay time and cost are reduced.

Eight-Cycle RNA Restaining

To demonstrate the multi-cycle potential of our approach, we stained mRNA GAPDH in eight consecutive hybridization cycles using SFO (**Figure 4**). To evaluate the target staining specificity, we incubated the cells with Cy3 conjugated SFO together with a Cy5 labeled orthogonal oligonucleotide in the odd hybridization cycles, and with Cy5 conjugated SFO and a Cy3 labeled orthogonal oligonucleotide in the even cycles. In the first cycle, the FISH spots were only observed in the Cy3 channel, suggesting that mRNA GAPDH is specifically stained by the corresponding SFO. After signal detection and strand displacement reactions, we imaged the cells again to confirm the efficient stripping of SFO. This process of staining, imaging and stripping was repeated eight times to obtain the 8-bit fluorophore sequence barcode for the target mRNA. For the spots co-localized in the first two cycles ($n = 1470$), more than 97% of these spots reappeared in each of the following cycles (**Figure 5**). And over 95% of the spots were successfully identified in all the hybridization cycles (**Figure 6**). A plot of the signal intensities of the FISH spots in both the Cy3 and

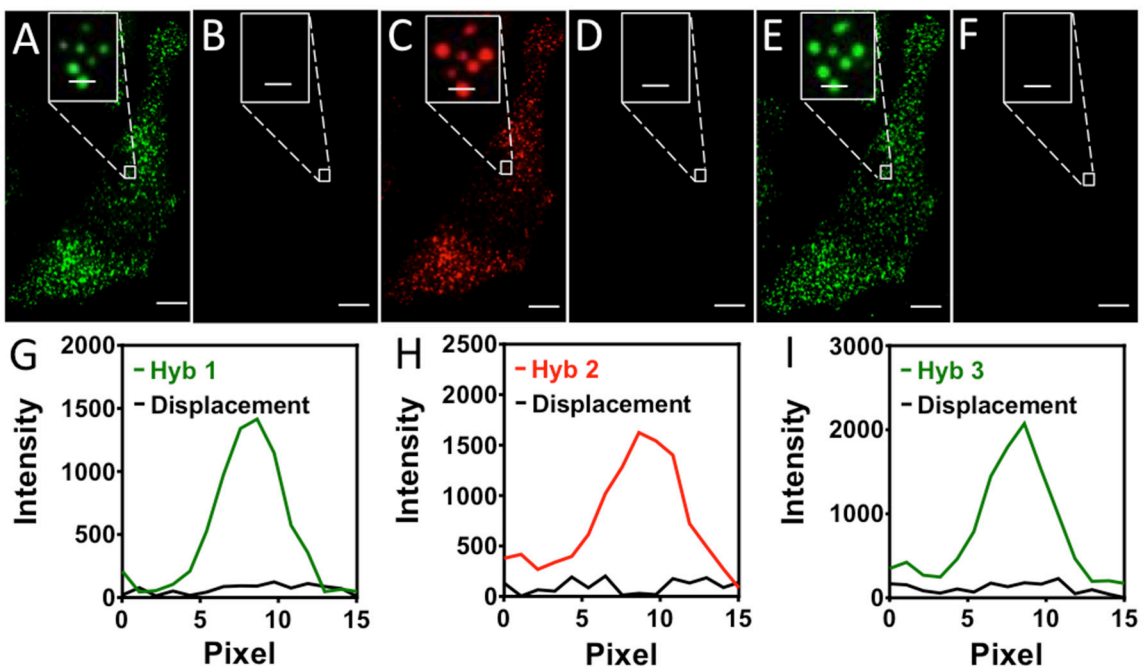


FIGURE 3 | (A) In the first hybridization cycle, GAPDH transcripts are stained by Cy3 labeled SFO. **(B)** SFO is removed by the eraser oligonucleotide. **(C)** In the second hybridization cycle, GAPDH transcripts are stained by Cy5 labeled SFO. **(D)** SFO is removed by the eraser oligonucleotide. **(E)** In the third hybridization cycle, GAPDH transcripts are stained by Cy3 labeled SFO. **(F)** SFO is removed by the eraser oligonucleotide. **(G)** Signal intensity profiles corresponding to the marked FISH spot in **(A,B)**. **(H)** Signal intensity profiles corresponding to the marked FISH spot in **(C,D)**. **(I)** Signal intensity profiles corresponding to the marked FISH spot in **(E,F)**. Scale bars, 5 μm.

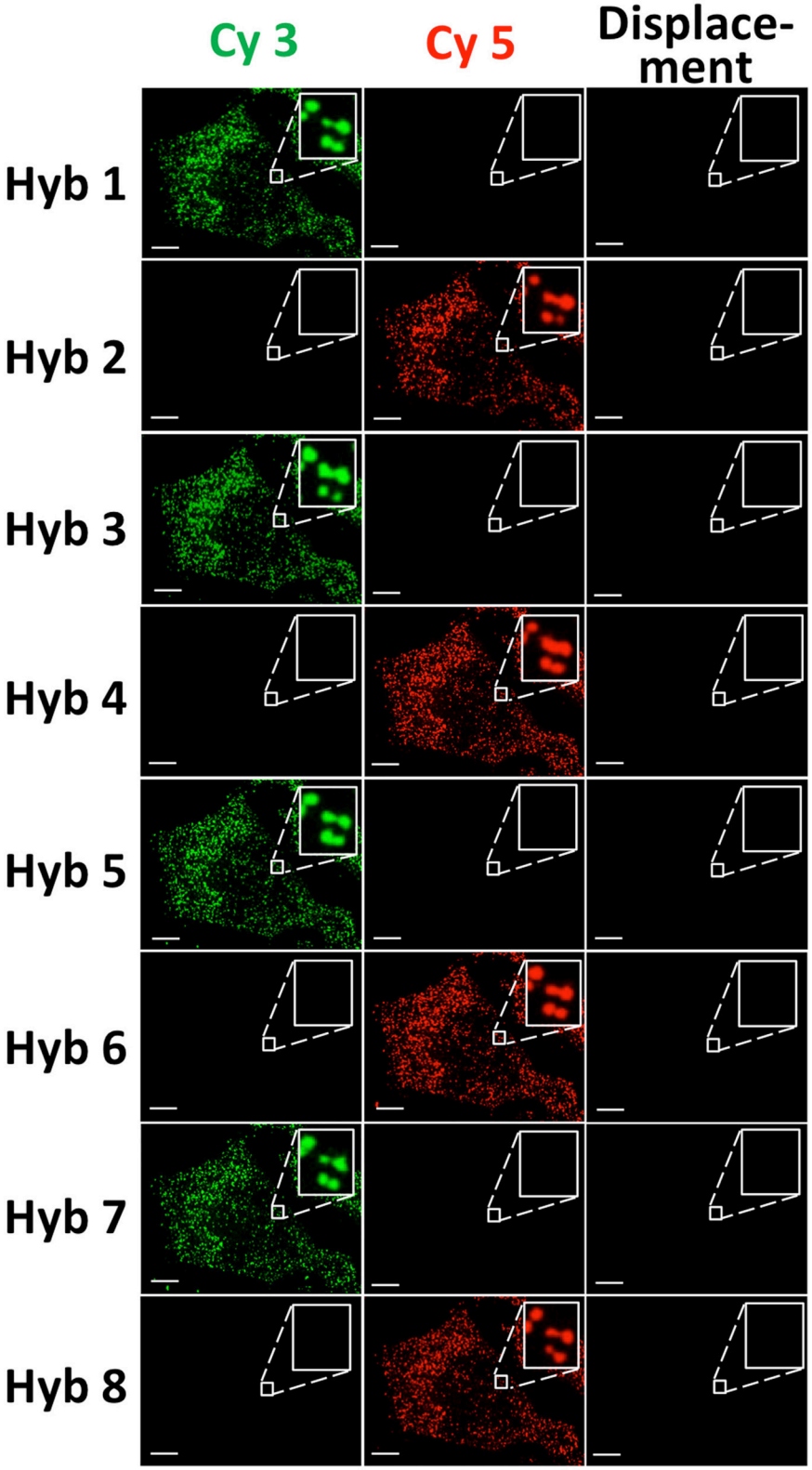


FIGURE 4 | GAPDH transcripts are stained by SFO in eight consecutive hybridization cycles. In the odd cycles, cells are incubated with Cy3 conjugated SFO and a Cy5 labeled orthogonal oligonucleotide. In the even cycles, cells are incubated with Cy5 conjugated SFO and a Cy3 labeled orthogonal oligonucleotide. After target staining, images are captured in the Cy3 and Cy5 fluorescence channels. Following strand displacement reactions, images are captured in the Cy3 channel in the odd cycles and in the Cy5 channel in the even cycles. Scale bars, 5 μm.

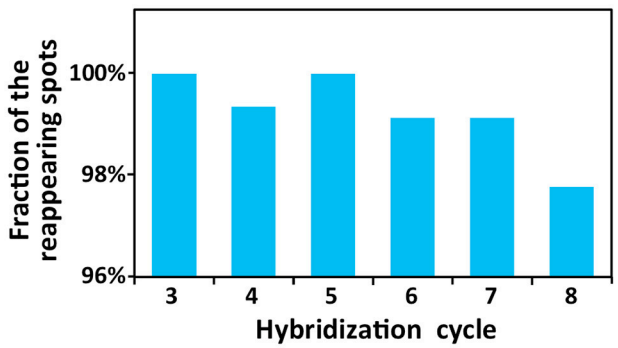


FIGURE 5 | Fractions of the spots colocalized in the first two hybridization cycles that reappear in the following analysis cycles.

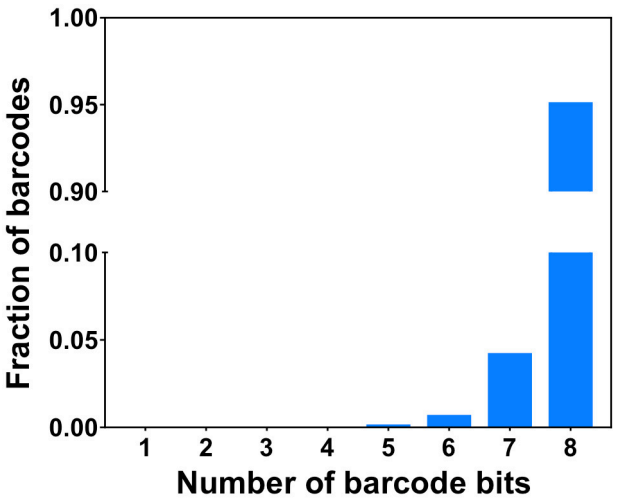


FIGURE 6 | Fractions of the identified barcodes with different numbers of bits.

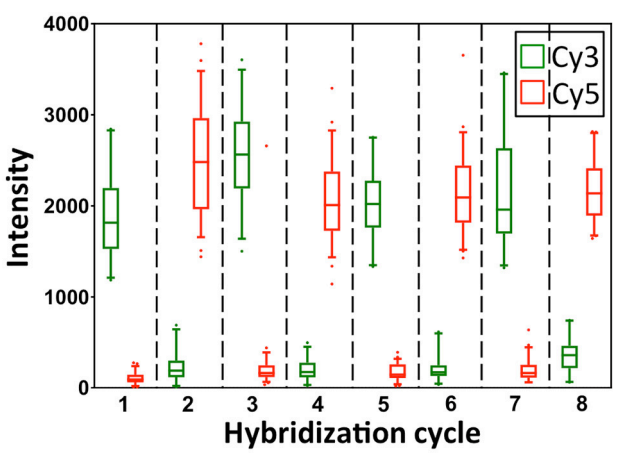


FIGURE 7 | Intensity distributions of GAPDH FISH spots ($n = 60$ spots) in Cy3 and Cy5 channels over the eight hybridization cycles.

Cy5 channels vs. the hybridization cycles is shown in **Figure 7**. Due to the high staining specificity, all the FISH spots were unambiguously detected in the correct fluorescence channels. We also performed control experiments to stain mRNA GAPDH using the conventional smFISH method. The copy numbers per cell obtained by the two methods (**Figure 8**), together with those reported previously using RNA-Seq (Uhlén et al., 2015), are consistent with each other. These results suggest that transcripts can be quantitatively profiled in single cells *in situ* by multi-cycle staining using the SFO-based approach.

In each cycle of MER-FISH, only certain transcripts are stained and other RNA targets remain unlabeled. Thus, to determine which transcripts are stained in a specific cycle, a detection threshold has to be manually selected by comparing the signal intensities of different FISH spots. However, due to the imperfect probe hybridization efficiency, RNA secondary structures, proteins bound to transcripts and other factors, even individual transcripts from the same RNA species can have significantly different staining intensities (**Figure 7**). As a result, the artificial detection threshold can lead to false negative signals, if the stained transcripts have low signal intensities. This threshold will also result in false positive signals, if the un-stained transcripts have high fluorescence intensities, which are generated as the signal leftovers from the previous cycles. In contrast, all the RNA targets are stained simultaneously in every cycle in the SFO-based approach. Rather than using a threshold to identify the stained transcripts, we compare the signal intensities of the same spot in different fluorescence channels to determine which SFO is hybridized to the specific RNA target. In this way, the correct fluorescence sequence can be unambiguously identified for both the weak spots (**Figure 9A**) and the strong spots (**Figure 9B**) in each analysis cycle. These results suggest that the SFO-based approach avoids the false positive and negative signals generated by the artificial threshold, and have enhanced detection sensitivity and analysis accuracy.

DISCUSSSION

We have developed an SFO-based technology for *in situ* RNA profiling. Compared with the existing methods, our approach has the following advantages. (i) By detecting transcripts directly without target sequence amplification, our technology enables RNA analysis at the single-molecule sensitivity. (ii) In this method, different RNA species can be distinguished by the varied color sequences, whose number increases exponentially with the number of hybridization cycles. Thus, our approach has the potential to enable highly multiplexed RNA analysis. (iii) All the distinct SFO in the whole specimen can be simultaneously removed by their corresponding eraser oligonucleotides. Therefore, our approach has high sample throughput, and allows a large number of cells to be quantified in a short time. (iv) As SFO can be very efficiently removed and have minimized cross-reactions with endogenous biomolecules and other probes, our approach has enhanced signal to noise ratio. (v) By keeping the pre-decoding oligonucleotides hybridized to their targets throughout the assay, our method has increased

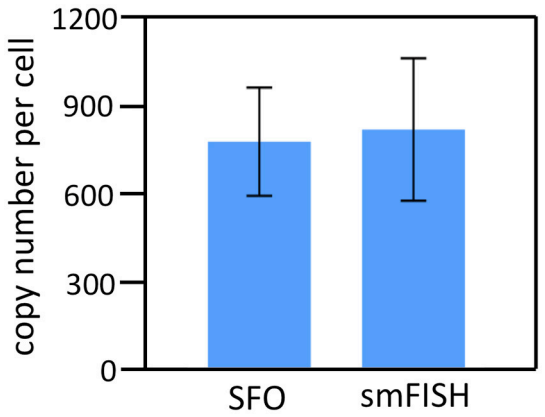


FIGURE 8 | The GAPDH mean copy numbers per cell ($n = 45$ cells) obtained using the SFO-based approach and the conventional smFISH ($p > 0.65$).

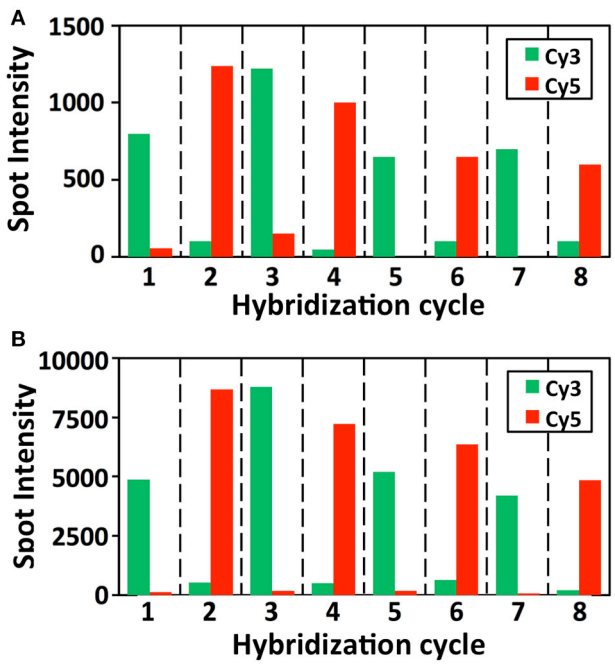


FIGURE 9 | Signal intensities of **(A)** a weak and **(B)** a strong GAPDH FISH spot in Cy3 and Cy5 channels over the eight hybridization cycles.

analysis accuracy and decreased assay time and cost. (vi) With each transcript stained in every cycle, this SFO-based approach avoids the false positive and false negative signals generated by the manually selected detection thresholds.

The number of RNA species that can be quantified using this SFO-based approach depends on two factors: the number of hybridization cycles and the number of different fluorophores used in each cycle. As we have demonstrated, at least eight hybridization cycles with high analysis accuracy can be carried out in the same set of cells. And it is well-established that hundreds of thousands of oligonucleotides can be prepared cost-effectively by massively parallel synthesis on a microarray slide (Murgha et al., 2014). Thus, further implementation of the SFO-based approach with four classical fluorophores applied in each cycle will potentially enable the whole transcriptome to be profiled using the 65, 536 (4^8) distinct fluorophore sequences. Additionally, multispectral fluorophores (Dai et al., 2011; Guo et al., 2011; Wang et al., 2012) coupled with the hyperspectral imaging (Garini et al., 2006) can be applied to allow more fluorophores to be distinguished and applied in each hybridization cycle. In this way, the cycle number together with the assay time can be further reduced. Furthermore, following the RNA profiling by this SFO-based approach, the nuclear and cellular membranes can be counterstained using nuclear staining dyes (such as DAPI) and fluorescent antibodies targeting membrane proteins (such as E cadherin), respectively. With individual cells precisely segmented by this counterstaining approach, the SFO-based approach will allow RNA analysis in single cells of intact tissues. Finally, the combination of this SFO-based approach with multiplexed *in situ* protein analysis technologies (Bodenmiller, 2016; Mondal et al., 2017, in press) will enable the comprehensive and integrated RNA and protein profiling in single cells *in situ*. This molecular imaging platform will bring new insights into systems biology, signaling network regulation, molecular diagnosis and cellular targeted therapy.

AUTHOR CONTRIBUTIONS

LX and JG designed the experiments. LX performed the experiments. LX and JG analyzed the data and wrote the manuscript.

FUNDING

This research is supported by funding from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R21AI132840), Arizona State University startup funds, Arizona State University/Mayo Clinic seed grant (ARI-219693), and Cystic Fibrosis Foundation (FIRTH17XX0).

ACKNOWLEDGMENTS

We would like to thank members of the Guo lab for their input and helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2018.00042/full#supplementary-material

Table S1 | Sequences of the pre-decoding probes, decoding probes, eraser oligonucleotide and SFO-orthogonal oligonucleotide.

REFERENCES

Bodenmiller, B. (2016). Multiplexed epitope-based tissue imaging for discovery and healthcare applications. *Cell Syst.* 2, 225–238. doi: 10.1016/j.cels.2016.03.008

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 1363, 1360–1363. doi: 10.1126/science.aaa6090

Crosetto, N., Bienko, M., and Van Oudenaarden, A. (2014). Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* 16, 57–66. doi: 10.1038/nrg3832

Dai, N., Guo, J., Teo, Y. N., and Kool, E. T. (2011). Protease probes built from DNA: multispectral fluorescent DNA-peptide conjugates as caspase chemosensors. *Angew. Chem. Int. Ed. Engl.* 50, 5105–5109. doi: 10.1002/anie.201007805

Franzini, R. M., and Kool, E. T. (2009). Efficient nucleic acid detection by templated reductive quencher release. *J. Am. Chem. Soc.* 131, 16021–16023. doi: 10.1021/ja904138v

Garini, Y., Young, I. T., and McNamara, G. (2006). Spectral imaging: principles and applications. *Cytometry A* 69, 735–747. doi: 10.1002/cyto.a.20311

Guo, J. (2016). *System and Method for Iterative Detection of Biological Molecules*. US Patent Application 20160054308A1. Arizona Board of Regents on behalf of Arizona State University.

Guo, J., Ju, J., and Turro, N. J. (2012). Fluorescent hybridization probes for nucleic acid detection. *Anal. Bioanal. Chem.* 402, 3115–3125. doi: 10.1007/s00216-011-5526-x

Guo, J., Wang, S., Dai, N., Teo, Y. N., and Kool, E. T. (2011). Multispectral labeling of antibodies with polyfluorophores on a DNA backbone and application in cellular imaging. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3493–3498. doi: 10.1073/pnas.1017349108

Guo, J., Yu, L., Turro, N. J., and Ju, J. (2010). An integrated system for DNA sequencing by synthesis using novel nucleotide analogues. *Acc. Chem. Res.* 43, 551–563. doi: 10.1021/ar900255c

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* 7, 200–210. doi: 10.1038/nrg1809

Huang, K., and Martí, A. A. (2012). Recent trends in molecular beacon design and applications. *Anal. Bioanal. Chem.* 402, 3091–3102. doi: 10.1007/s00216-011-5570-6

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., et al. (2013). *in situ* sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* 10, 857–860. doi: 10.1038/nmeth.2563

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343, 1360–1363. doi: 10.1126/science.1250212

Levesque, M. J., and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* 10, 246–248. doi: 10.1038/nmeth.2372

Levsky, J. M., Shenoy, S. M., Pezo, R. C., and Singer, R. H. (2002). Single-cell gene expression profiling. *Science* 297, 836–840. doi: 10.1126/science.1072241

Lubeck, E., and Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* 9, 743–748. doi: 10.1038/nmeth.2069

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods* 11, 360–361. doi: 10.1038/nmeth.2892

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Moffitt, J. R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., and Zhuang, X. (2016a). High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14456–14461. doi: 10.1073/pnas.1617699113

Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., and Zhuang, X. (2016b). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence *in situ* hybridization. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11046–11051. doi: 10.1073/pnas.1612826113

Mondal, M., Liao, R., and Guo, J. (in press). Highly multiplexed single-cell protein analysis. *Chemistry* doi: 10.1002/chem.201705014

Mondal, M., Liao, R., Nazaroff, C. D., Samuel, A. D., and Guo, J. (2018). Highly multiplexed single-cell *in situ* RNA and DNA analysis with bioorthogonal cleavable fluorescent oligonucleotides. *Chem. Sci.* 9, 2909–2917. doi: 10.1039/C7SC05089E

Mondal, M., Liao, R., Xiao, L., Eno, T., and Guo, J. (2017). Highly multiplexed single-cell *in situ* protein analysis with cleavable fluorescent antibodies. *Angew. Chem. Int. Ed. Engl.* 56, 2636–2639. doi: 10.1002/anie.201611641

Murgha, Y. E., Rouillard, J. M., and Gulari, E. (2014). Methods for the preparation of large quantities of complex single-stranded oligonucleotide libraries. *PLoS ONE* 9:e94752. doi: 10.1371/journal.pone.0094752

Olivo-Marin, J.-C. (2002). Extraction of spots in biological images using multiscale products. *Pattern Recognit.* 35, 1989–1996. doi: 10.1016/S0031-3203(01)00127-3

Raj, A., Van den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5, 877–879. doi: 10.1038/nmeth.1253

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* 546, 431–435. doi: 10.1038/nature22794

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). *in situ* transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92, 342–357. doi: 10.1016/j.neuron.2016.10.001

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Wang, S., Guo, J., Ono, T., and Kool, E. T. (2012). DNA polyfluorophores for real-time multicolor tracking of dynamic biological systems. *Angew. Chem. Int. Ed. Engl.* 51, 7176–7180. doi: 10.1002/anie.201201928

Xiao, L., and Guo, J. (2015). Multiplexed single-cell *in situ* RNA analysis by reiterative hybridization. *Anal. Methods* 7, 7290–7295. doi: 10.1039/C5AY00500K

Zhang, D. Y., and Seelig, G. (2011). Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem.* 3, 103–113. doi: 10.1038/nchem.957

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

TECHNOLOGY REPORT
published: 21 September 2016
doi: 10.3389/fbioe.2016.00070

Fluidic Logic Used in a Systems Approach to Enable Integrated Single-Cell Functional Analysis

*Naveen Ramalingam†, Brian Fowler†, Lukasz Szpankowski†, Anne A. Leyrat, Kyle Hukari, Myo Thu Maung, Wiganda Yorza, Michael Norris, Chris Cesar, Joe Shuga, Michael L. Gonzales, Chad D. Sanada, Xiaohui Wang, Rudy Yeung, Win Hwang, Justin Axsom, Naga Sai Gopi Krishna Devaraju, Ninez Delos Angeles, Cassandra Greene, Ming-Fang Zhou, Eng-Seng Ong, Chang-Chee Poh, Marcos Lam, Henry Choi, Zaw Htoo, Leo Lee, Chee-Sing Chin, Zhong-Wei Shen, Chong T. Lu, Ilona Holcomb, Aik Ooi, Craig Stolarczyk, Tony Shuga, Kenneth J. Livak, Cate Larsen, Marc Unger and Jay A. A. West**

New Technologies Research Department, Fluidigm Corporation, South San Francisco, CA, USA

OPEN ACCESS

Edited by:
Xinghua Pan, Yale University, USA

Reviewed by:
Senentxu Lanceros-Mendez, University of Minho, Portugal
Lin Han, Shandong University, China

***Correspondence:**
Jay A. A. West
jay.west@fluidigm.com

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *31 May 2016*
Accepted: *23 August 2016*
Published: *21 September 2016*

Citation:
Ramalingam N, Fowler B, Szpankowski L, Leyrat AA, Hukari K, Maung MT, Yorza W, Norris M, Cesar C, Shuga J, Gonzales ML, Sanada CD, Wang X, Yeung R, Hwang W, Axsom J, Devaraju NSGK, Angeles ND, Greene C, Zhou M-F, Ong E-S, Poh C-C, Lam M, Choi H, Htoo Z, Lee L, Chin C-S, Shen Z-W, Lu CT, Holcomb I, Ooi A, Stolarczyk C, Shuga T, Livak KJ, Larsen C, Unger M and West JAA (2016) Fluidic Logic Used in a Systems Approach to Enable Integrated Single-Cell Functional Analysis. Front. Bioeng. Biotechnol. 4:70. doi: 10.3389/fbioe.2016.00070

The study of single cells has evolved over the past several years to include expression and genomic analysis of an increasing number of single cells. Several studies have demonstrated wide spread variation and heterogeneity within cell populations of similar phenotype. While the characterization of these populations will likely set the foundation for our understanding of genomic- and expression-based diversity, it will not be able to link the functional differences of a single cell to its underlying genomic structure and activity. Currently, it is difficult to perturb single cells in a controlled environment, monitor and measure the response due to perturbation, and link these response measurements to downstream genomic and transcriptomic analysis. In order to address this challenge, we developed a platform to integrate and miniaturize many of the experimental steps required to study single-cell function. The heart of this platform is an elastomer-based integrated fluidic circuit that uses fluidic logic to select and sequester specific single cells based on a phenotypic trait for downstream experimentation. Experiments with sequestered cells that have been performed include on-chip culture, exposure to various stimulants, and post-exposure image-based response analysis, followed by preparation of the mRNA transcriptome for massively parallel sequencing analysis. The flexible system embodies experimental design and execution that enable routine functional studies of single cells.

Keywords: single-cell, mRNA-seq, functional studies, Fluidigm, Polaris

INTRODUCTION

Recent single-cell transcriptomic analyses have documented the importance of cellular heterogeneity in studying cancer (Ennen et al., 2014; Saadatpour et al., 2014; Kim et al., 2015), immunology (Shalek et al., 2014), developmental biology (Briggs et al., 2015), stem cell research (Wilson et al., 2015), and neurobiology (Pollen et al., 2015). It has been estimated that the human body contains 37.2 trillion cells (Bianconi et al., 2013), excluding the complex microbiome that lives in the human body. High-throughput single-cell mRNA sequencing provides an unbiased path to classifying this

vast number of cells into cell types. This endeavor has stimulated the development of methods to increase throughput (Fan et al., 2015; Klein et al., 2015; Macosko et al., 2015). The classification of cell types can be thought of as a high-resolution anatomy. At the single-cell level, moving from anatomy to physiology or from description to mechanism means moving from cell type to cell function. This will require integrating transcriptional data with other cellular measurements. In this regard, progress has been made in obtaining transcriptomic and genomic information (Dey et al., 2015; Macaulay, 2015), transcriptomic and epigenomic information (Angermueller et al., 2016), or transcriptomic and proteomic information (Darmanis et al., 2016; Frei et al., 2016) from the same single cell.

Moving from cell type to cell function will also require understanding how single-cell profiles change in response to perturbations. It is important to examine these effects at the single-cell level because cell-to-cell heterogeneity has been observed in a diverse set of circumstances, such as the response of macrophages to bacterial invasion (Avraham et al., 2015), the response of hematopoietic cells to various drugs (Bendall et al., 2011), and drug resistance in adenocarcinoma cells (Kim et al., 2015). Progress in the long-term culture of circulating tumor cells (Gao et al., 2014; Yu et al., 2014; Cayrefourcq et al., 2015; Alix-Panabières et al., 2016) enables single-cell functional studies on this important class of cells, which should lead to improved cancer diagnosis and therapy. Performing perturbation experiments on single cells requires care in maintaining the appropriate microenvironment. Examining the effects of serum on mouse embryonic stem cells (ESCs), researchers (Guo et al., 2016) concluded that "a large proportion of intracellular network variability is due to the extracellular culture environment." Microfluidic-based approaches are attractive for the precise control of the microenvironment because they enable structures at a size appropriate for single cells. Microfluidic systems for high-throughput preparation of sequencing libraries, though, have cell lysis as the initial step and thus are not suitable to maintain single cells for experimentation. What is required is a system specifically designed to capture, maintain, perturb, and observe single cells and then prepare these cells for high-dimensional analysis.

In this paper, we report development of an integrated fluidic circuit (IFC) that uses fluidic logic to actively select and sequester desired single cells based on particular biological markers of interest. This Polaris™ IFC can sequester up to 48 single cells. If required, the cells can be cultured in appropriate medium in order to control and manipulate the microenvironment around the sequestered cells. For adherent cells, appropriate extracellular matrix (ECM) can be coated inside the culture chambers. The single cells can be perturbed with a drug or other stimuli (i.e., mRNA, cytokines, bacteria, or viruses), with the response to perturbation monitored and measured by fluorescence imaging. Subsequently, the single cells are processed for cell lysis, reverse transcription (RT), and full-length transcriptome amplification using template-switching chemistry. Following harvest from the IFC, sequencing libraries are generated using a modified Nextera® protocol and sequenced on any Illumina® platform (**Figure 1**).

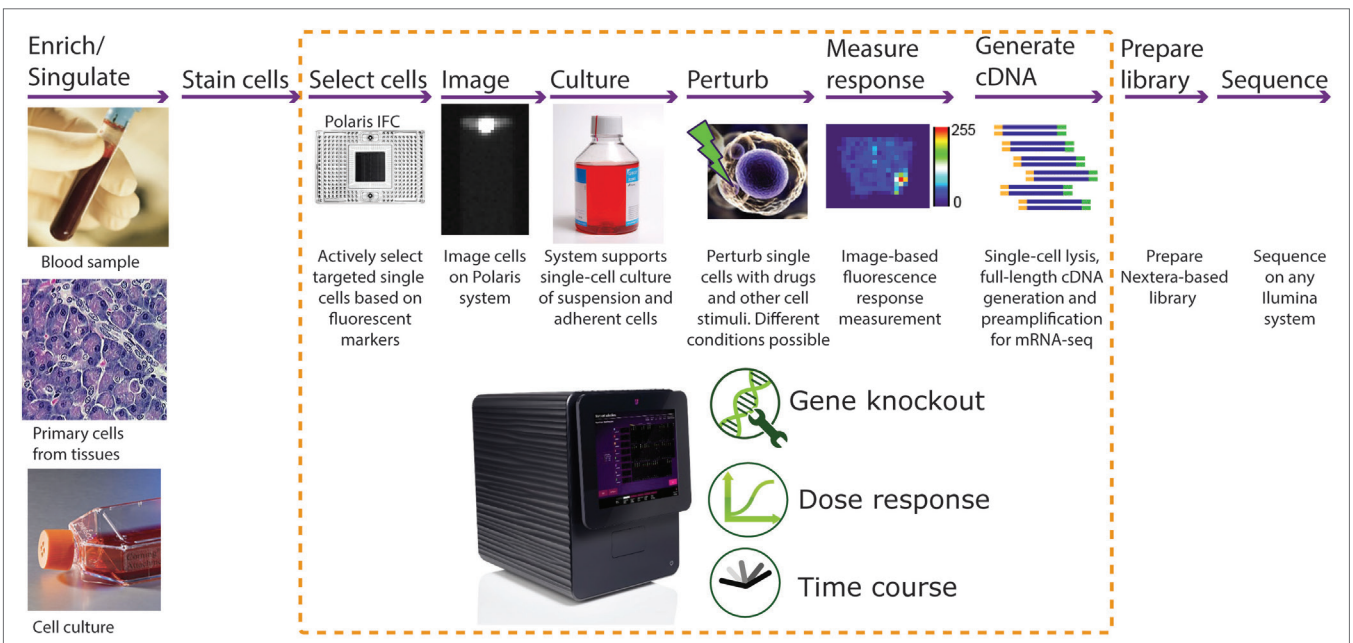


FIGURE 1 | Typical workflow for single-cell functional studies. The input single-cell suspension can be obtained from blood, primary cells, or cell culture. In the case of rare cells or a subset of cell population, there is an option to enrich them prior to use with Polaris using either fluorescence-assisted cell sorting (FACS) or other methods. Subsequently, the cells are labeled with a universal fluorescent marker for tracking the cells on the Fluidigm® Polaris system. Most of the single-cell functional study steps are automated on the Polaris system. The Polaris system generates preamplified full-length cDNA, which can be further processed for library preparation and massive parallel sequencing for mRNA sequencing.

MATERIALS AND METHODS

Design and Fabrication of Logic-Based Integrated Fluidic Circuit

The nanoscale IFC consists of a plastic carrier and a polydimethylsiloxane (PDMS) core (**Figure 2A**) or fluidic circuit. The carrier contains reservoir wells for input and output of reagents and circuit control. It provides a platform to facilitate interfacing with the fluidic circuit. The fluidic circuit with the desired microfluidic control components was fabricated using multilayer soft lithography (MSL®) process (Unger et al., 2000). Fluidic circuit components include flow and control channels, valves, multiplexors, and logic devices [such as serial-to-parallel shift register (SR)]. Fabrication and operational details of the fluidic logic circuits and devices were reported earlier (Devaraju and Unger, 2012). The IFC is designed to have the capability to actively select single cells based on fluorescent markers, isolate them to a desired holding location (cell capture site), apply individual conditions (feed medium and dose reagents to cells), and finally study the functional response. Execution of all these complex functions in a routine fashion requires flexible, programmable operational control, which in turn requires many controls in a parallel manner. Traditionally, in microfluidics, a dedicated external control line is required to independently control a set of valves. This imposes a limitation on the number of practical on-chip control operations and poses a challenge for scalability by requiring more external hardware. On-chip control architecture capable of receiving and processing data by elementary computation and decision-making can integrate programmability of controls on-chip and allows an increase in the number of on-chip control lines for the same number of external chip connections (**Figure 2B**). We developed such a microfluidic fluidic logic and implemented it on our Polaris IFC.

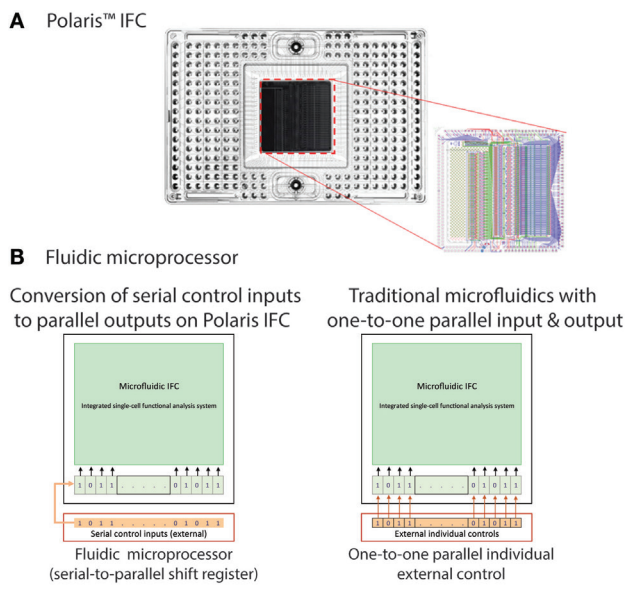


FIGURE 2 | **(A)** Polaris mRNA-seq dosing integrated fluidic circuit (IFC). The polydimethylsiloxane (PDMS) elastomer is attached to a plastic carrier. In order to image the cells with low-fluorescence background on the Polaris system, the PDMS is backed with a black material. This black backing is removed after cell perturbation and fluorescence response measurement. Removal of black backing enables high-resolution imaging of cells on a fluorescent microscope. The CAD drawing of the microfluidic components is shown on the right. The green channels are the control lines. **(B)** Polaris IFC uses fluidic processor to receive serial control inputs and converts them to parallel shift register elements. Traditional microfluidic control elements use one external control for every internal control.

The state-based microfluidic fluidic logic devices and circuits utilize static gain and normally closed valves (NCVs). NCVs are fabricated by filling specialized control channels with a flash curable prepolymer and curing while the valve is closed. The resulting closed valve exerts certain force against fluidic pressure to keep the valve closed. The valves are characterized by breakthrough pressure: the threshold pressure in the flow channel required to push open the valve and restore the continuity of the flow. Breakthrough pressure for an NCV can be tailored by controlling the pressure at which they are cured. Using these NCVs, we have developed static gain valves (SGV) that have the ability to control higher (or equal) fluidic pressure using a lower pressure. This type of valve is essential to create any logic/feedback structures (to account for signal strength losses), which can receive the output of the previous element/gate and use it as an input for decision making.

Utilizing the SGV, we next built an inverter (NOT gate), which was further used to build more complex circuits including bistable flip flops, clocked flip flops (latches), delay flip flops (D-FF, one bit of the SR), and complex microprocessors (SR). A SR that is capable of processing $n + 1$ bits of data is formed by combining n D flip flops (bits of SR). The SR presented here uses air as the medium and receives three active high-pressure inputs: source, clock, and data (**Figure 3**). The pneumatic output of the SR cannot be used to control the flow of liquids in microchannels directly, due to risk of introducing bubbles. In order to address this issue, the signal medium is converted from air to liquid using an inverter.

The Polaris IFC microprocessor receives 28 external signals serially and processes them into 28 parallel independent controls capable of controlling individual valves or a set of valves. Five dedicated high-pressure external active signals are required for a SR. The CAD drawing of the various microfluidic components on a Polaris IFC is shown in **Figure 3**. The IFC can accept up to 20 independent reagents. The fluorescently labeled cells are loaded in a serpentine partition channel. Based on a desired combination of up to three fluorescent markers (refer to Section "Polaris Instrument Design" for excitation and emission details), single cells are selected and sequentially isolated to the cell capture sites through a multiplexer. Up to 48 single cells can be sequestered on a single Polaris IFC. Subsequently, these 48 cells are processed through template-switching chemistry for full-length cDNA generation for mRNA-seq. In brief, the cells are lysed and reverse-transcribed, and full-length cDNA is preamplified by long and accurate PCR.

Polaris Instrument Design

The Fluidigm Polaris system (**Figure 4A**) consists of four major modules: (1) thermal control module; (2) imaging module; (3) pneumatic control module; and (4) environmental control

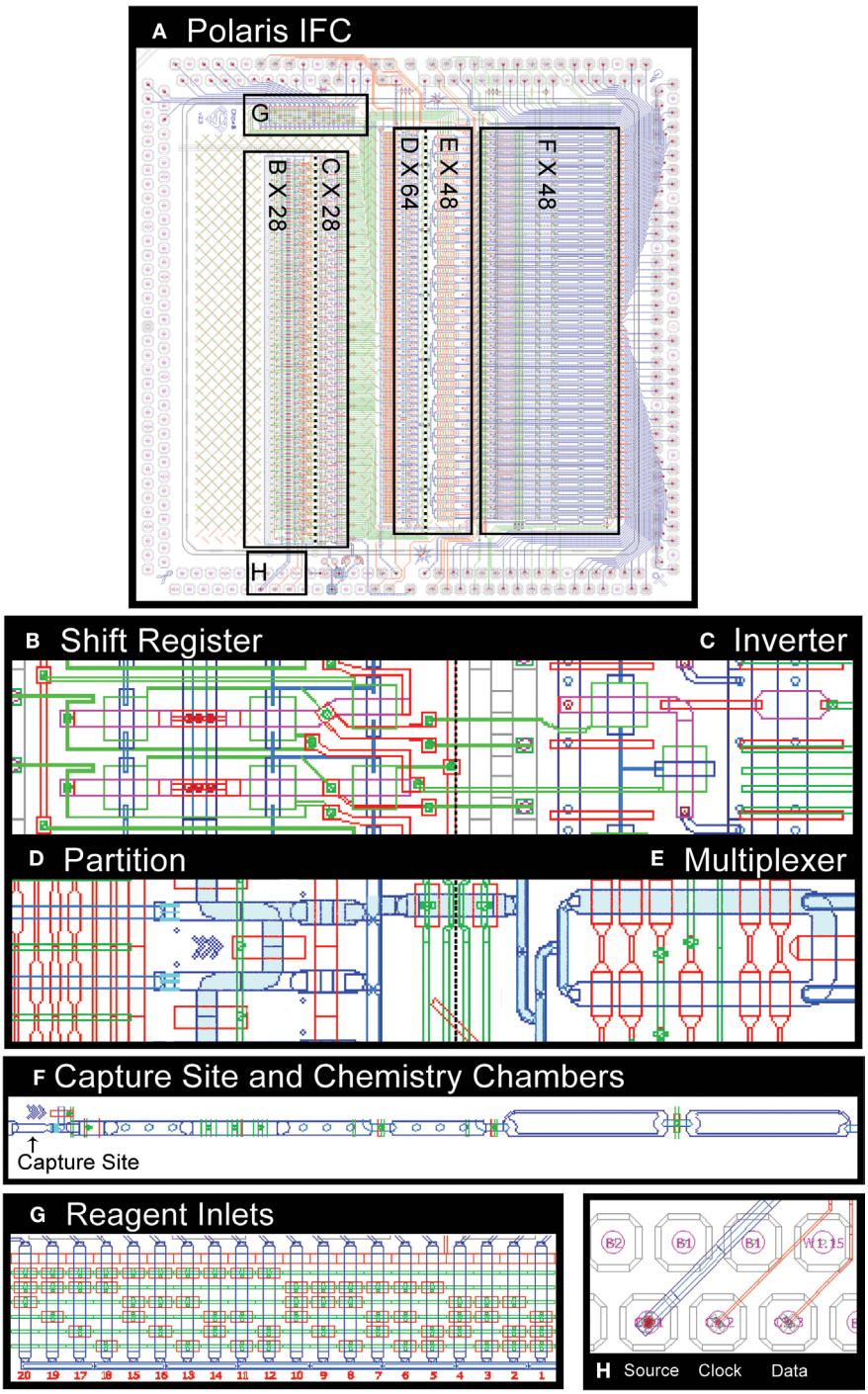


FIGURE 3 | CAD drawing of microfluidic control components on Polaris mRNA-seq dosing IFC (A). The shift register **(B)** enables active selection of single cells. The dilute single-cell suspension is loaded into a serpentine partition channel **(D)**. The cell suspension liquid flow is stopped, and the partition channel is imaged to identify single cells based on a particular set of fluorescence markers. The selected cells are then microfluidically moved downstream to a cell capture site **(F)** through a multiplexer **(E)**. The IFC is capable of accepting 20 reagents **(G)** as input. The shift register uses inverter **(C)** and a set of source, clock, and data **(H)**.

(EC) module. The thermal module consists of a Peltier-based thermoelectric couple (TEC) device for heating/cooling. The TEC module can provide temperature in the range of 4–99°C. Vacuum grooves on the thermal module are designed to enable tight contact with the glass-based integrated heat spreader (IHS) on the Polaris IFC. This ensures thermal uniformity across the fluidic circuit. The imaging module contains a five-color LED light engine for excitation (Ex wavelengths: 438, 475, 530, 575,

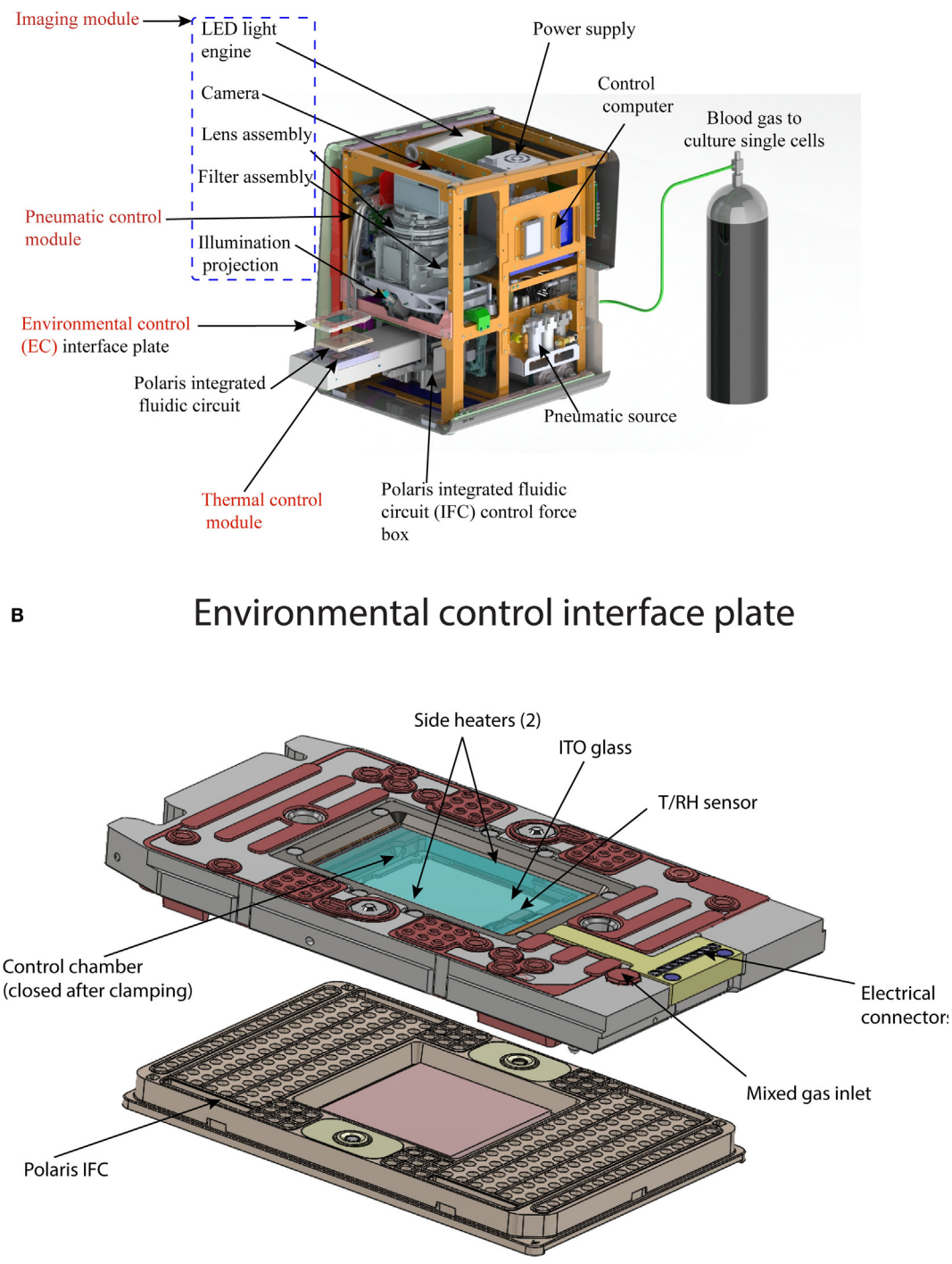


FIGURE 4 | **(A)** Components of Polaris instrument. The instrument consist of four major modules: (1) thermal module, enables preparative chemistry on sequestered single cells; (2) imaging module, consists of LED excitation and emission collection by a camera; (3) pneumatic module, controls the movement of reagents inside microfluidic channel by application of positive pressure on the IFC carrier; and (4) environmental control module, maintains the temperature, humidity, and blood gas flow rate for single-cell culture on-IFC. **(B)** Components of environmental control interface plate. The top of the interface plate contains glass coated with indium tin oxide. Internal heaters are used to maintain the temperature of the closed chamber between the interface plate and Polaris IFC. The temperature and relative humidity inside the closed chamber (after clamping with Polaris IFC) are measured by T/RH sensor. Blood or premixed gas required for cell culture is pumped through mixed gas inlet port on the interface plate. Polaris IFC is shown for reference.

and 632 nm). The light source from the engine is collected and projected onto the fluidic circuit using fiber optics. The emitted signal from the fluidic circuit passes through an emission filter (five Em wavelengths: 488, 525, 570, 630, and 700 nm) and is collected by CCD camera with 6-μm pixel resolution through a custom-designed collimator lens.

The pneumatic control module generates and stores air with volume up to 1 L. The system can achieve a maximum pressure of 100 psi. The pneumatic controller generates a vacuum on the thermal chuck, clamps the Polaris IFC against the EC interface plate (IP) to enable a closed environment around the IFC, and loads reagents from the inlets on the IFC carrier to the microchannels and reagent chambers of the fluidic circuit. The system contains segregated zones to regulate four different pressures simultaneously. The EC module provides an environment suitable for cell culture using user-desired gas composition. Environmental parameters such as temperature, relative humidity (RH), and mix gas flow rate across the fluidic circuit are monitored and controlled. The gas inside the closed chamber is heated by two heater coils. The gas inlet on the EC IP is used to regulate the flow of gas across the fluidic circuit. The EC IP (**Figure 4B**) contains an indium tin oxide (ITO) coated glass on the top to maintain thermal control in the EC while allowing imaging through the EC IP. During cell culture operation, the ITO glass is heated to prevent condensation. During cell culture, the environment around the fluidic circuit is maintained by blood gas (5% CO_2, 5% oxygen, and 90% nitrogen) or premixed gas of choice (for example, 5% CO_2, 20% oxygen, and 75% nitrogen). Before on-IFC cell culture, a rectangular sponge saturated with water is installed inside the closed chamber to provide the desired humidity through heating. The EC IP is equipped with two sensors (T/RH) to measure and maintain temperature at 37°C and RH at 90%.

K562 Cell Culture and CD59 Staining

K562 cells (ATCC® CCL-243) are cultured in T25 flasks in a volume ranging from 10 to 15 mL in an incubator (37°C, 5% CO_2). The culture medium contains IMDM + GlutaMAX™-I + 25 mM HEPES + 3.024 g/L sodium bicarbonate (Gibco, 31980-030) and is supplemented with 10% FBS. The cells were fed every 2–3 days by dilution to 200,000 cells/mL. The K562 cells were stained with CellTracker™ Orange (CTO) CMRA Dye (Thermo Fisher Scientific, C34551) as universal marker and Alexa Fluor® 647 conjugated CD59 antibody. The recommended dyes and corresponding excitation and emission filters on the Polaris system are shown in **Table 1**. Immediately before use, the cell staining solution was prepared by adding 0.6 μL of 1 mM CTO to 2 mL of HBSS without calcium or magnesium (−/−) at a final concentration of 0.3 μM. The cell staining solution was protected from light until use within 30 min. A total of ~1.5×10^6 cells was aliquoted in a 15 mL non-pyrogenic conical tube. The cell suspension was centrifuged at 300 × *g* for 3 min. Following this, the medium was aspirated without disturbing the pellet, and 2 mL of cell staining solution was added to the pellet and gently suspended by pipetting up and down three times. The cells were then incubated in the dark at 37°C for 20 min with occasional inverting and flicking. Following this, the cells were washed by adding 12 mL of HBSS to the cells in the 2 mL of staining buffer and then centrifuged at 300 × *g* for 5 min. Supernatant was aspirated and discarded without disturbing the pellet. The pellet was then resuspended in 200 μL of HBSS. The CTO-stained K562 cells were split into two tubes of 100 μL each. One tube was used as negative surface-stained cell population, and the other tube was processed further to stain CD59 epitope. In order to stain the surface CD59 epitope, 10 μL of CD59 biotinylated antibody (BD Biosciences, 555762, 100 tests, 2.0 mL) was added to 100 μL of CTO-stained cells. For negative surface-stained cell control, 10 μL of HBSS was added. Both the tubes were incubated at room temperature for 20 min with occasional inverting and flicking. Subsequently, 13 mL of HBSS was added to each tube and centrifuged at 300 × *g* for 5 min. The pellet was resuspended in 100 μL of HBSS. To this, 0.5 μL of Streptavidin Alexa Fluor® 647 (Thermo Fisher Scientific, S32357, 2 mg/mL stock) was added to positive-stain tube with CD59 biotinylated antibody in 100-μL cell suspension. This solution was mixed gently by pipetting up and down five times. Following this, the stain solution was incubated at room temperature for 15 min with occasional flicking. Again, 13 mL of HBSS was added to each tube, mixed by gently pipetting up and down, and centrifuged at 300 × *g* for 5 min. The supernatant was removed, and the pellet was resuspended in ~100–150 μL culture medium with FBS, but without phenol red, to prevent high background fluorescence during cell selection on the Polaris system. The resuspension volume of culture medium accounts for cell losses during the staining procedure and was chosen to yield a cell concentration greater than the target concentration of 550 cells/μL. Typically, 10 μL of cell mix is loaded into a C-Chip™ Disposable Hemocytometer (INCYTO, DHC-N01) and imaged on the Polaris system to estimate the staining intensity and purity. In order to achieve optimal buoyancy, cells in the range of 333–550 cells/μL are mixed with suspension reagent (Fluidigm, 101-0434). Typically, the ratio of cells to cell suspension reagent is 3:2. However, this ratio might need optimization depending on the cell type.

TABLE 1 | Recommended stains and corresponding excitation and emission filter on Polaris system.

Channel name	Excitation[a]	Emission	Recommended stains
FAM™	475/40	525/25	Alexa Fluor 488 (selection marker)
VIC®	530/20	570/30	CellTracker Orange CMRA Dye (universal marker)
Cy5®	632/28	700/30	Alexa Fluor 647 (selection marker)

[a]*Excitation values are center wavelength/band pass (≥90%).*

IFC Operation

The Polaris IFC is first primed to fill the control lines on the fluidic circuit, load cell capture beads, and the inside of PDMS channels is blocked to prevent non-specific absorption/adsorption of proteins. In order to capture and maintain the single cells in the sites, the capture sites (48 sites) are preloaded with beads that are linked on-IFC to fabricate a tightly packed bead column during the IFC prime step. In the case of adherent cells, ECM is coated inside the cell capture chambers during prime

step. After completion of the prime step, the cell mix (cells with suspension reagent) is loaded on the Polaris IFC and single CTO$^+$/CD59$^+$ cells are selected to capture sites. We extensively tested the performance of the Polaris IFC and system at three different cell purities (3, 10, and 50%). The cell purity is defined as the ratio of CTO$^+$/CD59$^+$ cells to CTO$^+$/CD59$^-$ cells. During the cell selection step, the suspended cells are loaded into the serpentine partition channel (**Figure 3**). Subsequently, the flow inside the partition channel is stopped (**Figure 5A**), and the cells are imaged in the partition channel for different fluorescent markers, as selected by the user. Based on automated image analyses by the system's software, only single cells with the desired combination of fluorescent markers are selected and isolated to the cell capture site. Any doublets or single cells with undesired fluorescent combinations are not selected by the software for further experimentation (**Figure 5B**). The selected single cells are moved to capture sites through a multiplexer (**Figure 5C**). The system takes images of the capture sites to confirm the arrival of single cells from a particular position in the partition to a particular capture site number. **Figure 5C** shows a typical image from the Polaris system showing K562 single cells captured inside sites packed with a column of beads. The system will then select and isolate all available single cells from a partition fill as per desired fluorescent marker combination. Once it completes selection of candidate cells, the system refills the serpentine partition to look for more candidate single cells. The system repeats this process to select and isolate single cells until it fills all 48 capture sites.

If desired, the single cells can then be cultured in the capture sites. It is possible to culture either suspension (e.g., K562) or adherent (e.g., BJ fibroblast) cells. For adherent cells, extracellular matrix can be coated inside the capture site during the IFC priming step. **Figure 5D** shows a Polaris image of a cultured BJ fibroblast (adhered). Based on the experimental design, it is possible to dose these single cells and on-IFC-cultured single cells with drugs or other cell stimuli. Finally, the single cells are processed through template-switching mRNA-seq chemistry for full-length cDNA generation and preamplification on-IFC.

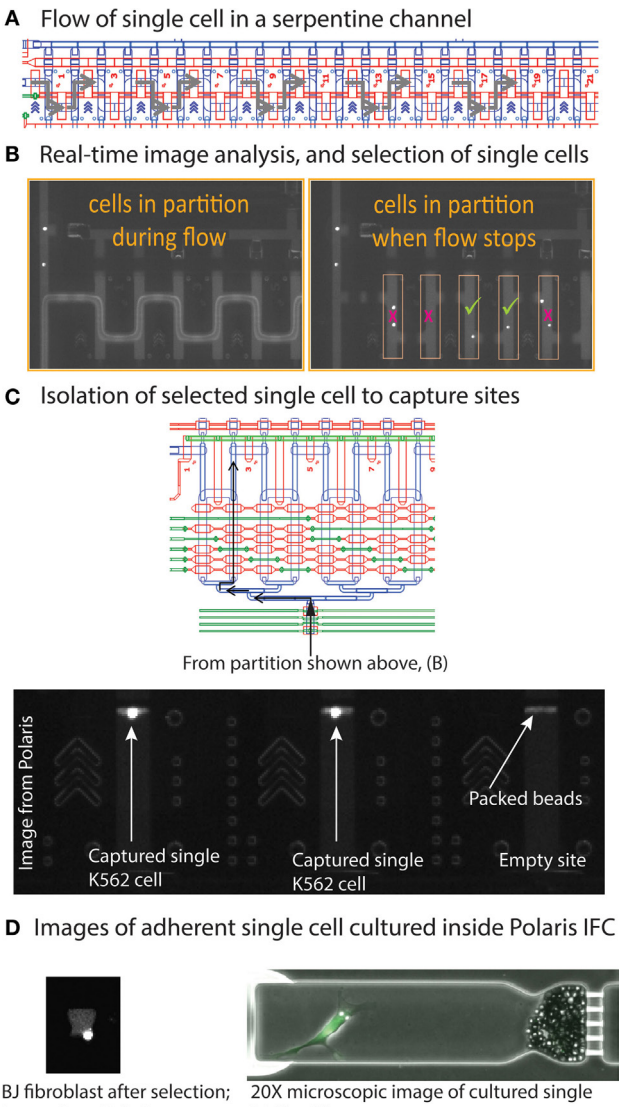


FIGURE 5 | **(A)** Cells in suspension are loaded into a serpentine partition channel (CAD design image). **(B)** Image analysis shows movement of fluorescent cells inside the serpentine channel during reagent flow. Once the reagent flow is stopped, single cells are separated from each other. The system software identifies single cells based on desired fluorescent markers. **(C)** After identification, single cells are isolated by moving them to the capture sites through a fluidic multiplexer. **(D)** Image of cultured BJ fibroblast cells. The Polaris IFC can be imaged on a microscope to obtain high-resolution micrographs.

Full-Length cDNA Generation

Preamplified full-length cDNA of selected single cells are generated on-IFC, and the amplicons are harvested through 48 different outlets. We used the SMARTer Ultra® Low RNA Kit for Illumina Sequencing (Clontech®, 634936) to generate preamplified cDNA. The selected and sequestered single cells were lysed using Polaris cell lysis mixture. The 28-µL cell lysis mix consists of 8.0 µL of Polaris Lysis Reagent (Fluidigm, 101-1637), 9.6 µL of Polaris Lysis Plus Reagent (Fluidigm, 101-1635), 9.0 µL of 3′ SMART™ CDS Primer II A (12 µM, Clontech, 634936), and 1.4 µL of Loading Reagent (20X, Fluidigm, 101-1004). Synthetic RNA spikes can be optionally used with cell lysis mix. We typically use ArrayControl™ RNA spikes 1, 4, and 7 (Thermo Fisher Scientific, AM1780) to establish the functionality of RT and PCR on-IFC. We also use ERCC spikes at 1:50,000 dilution (final in lysis mix) for efficiency and quantification estimations. In order to implement synthetic RNA spikes, we thoroughly mix 96.5 µL of loading reagent with 2.5 µL of SMARTer Kit RNase Inhibitor (40 U/µL; Clontech, 634936) and subsequently add 1 µL of synthetic RNA spike to this spike mix. If RNA spike is used, then 1.4 µL of the loading reagent is replaced with the spike mix. The thermal profile for single-cell lysis is 37°C for 5 min, 72°C for 3 min, 25°C for 1 min, and hold at 4°C.

The 48-µL preparation volume for RT contains 1X SMARTer Kit 5X First-Strand Buffer (5X; Clontech, 634936), 2.5-mM

SMARTer Kit Dithiothreitol (100 mM; Clontech, 634936), 1-mM SMARTer Kit dNTP Mix (10 mM each; Clontech, 634936), 1.2-μM SMARTer Kit SMARTer II A Oligonucleotide (12 μM; Clontech, 634936), 1-U/μL SMARTer Kit RNase Inhibitor (40 U/μL; Clontech, 634936), 10-U/μL SMARTScribe™ Reverse Transcriptase (100 U/μL; Clontech, 634936), and 3.2 μL of Polaris RT Plus Reagent (Fluidigm, 101-1366). All the concentrations correspond to those found in the RT chambers inside the Polaris IFC. The thermal protocol for RT is 42°C for 90 min (RT), 70°C for 10 min (enzyme inactivation), and a final hold at 4°C.

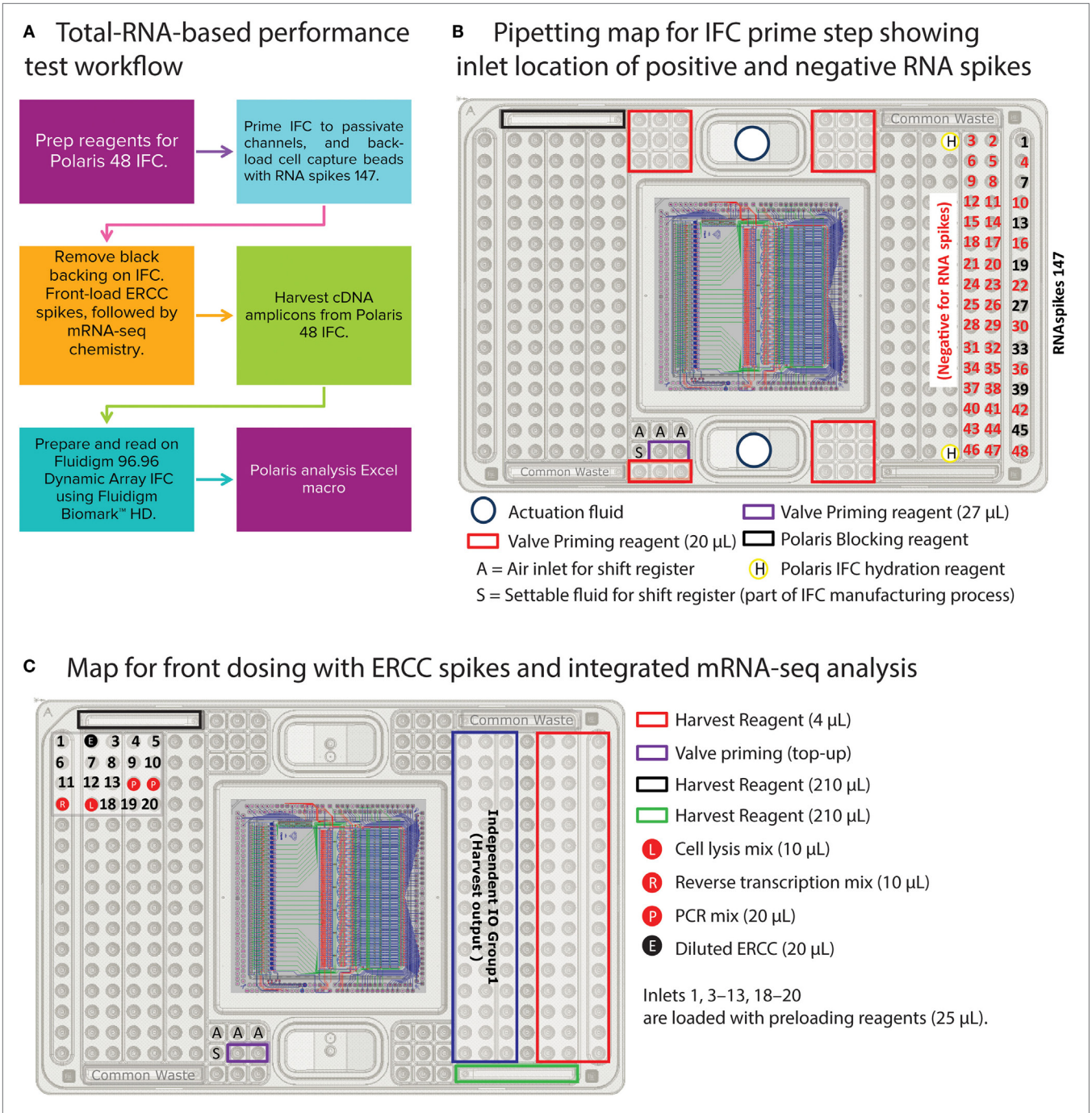
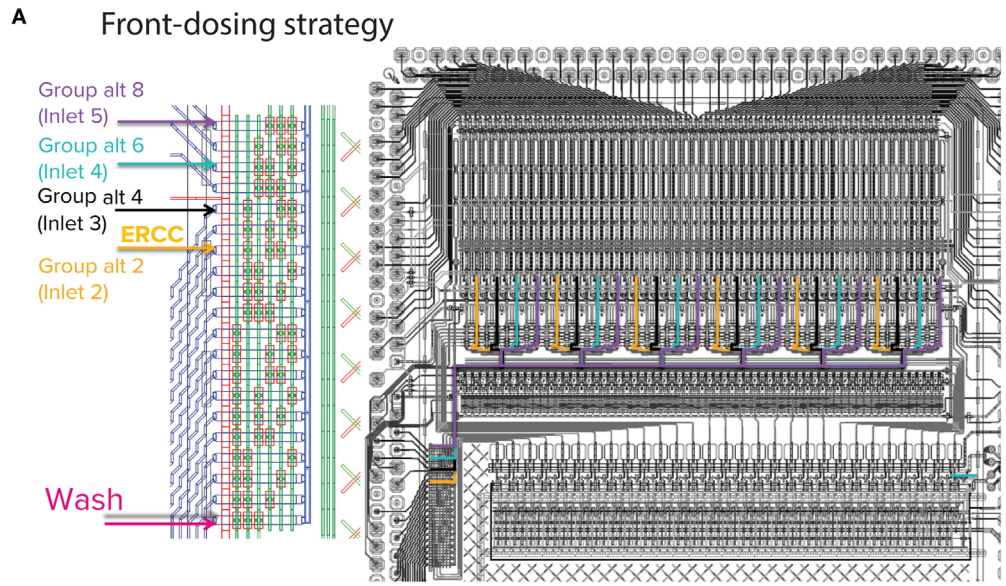


FIGURE 6 | Total-RNA-based performance test (two-step) workflow and pipetting map. **(A)** Workflow to prime Polaris IFC with beads, back-load them with RNAspikes 147, simulate front dosing with diluted ERCC spikes, and generate cDNA using mRNA-seq chemistry. The cDNA amplicons from Polaris IFC were analyzed on the Fluidigm M96.96 IFC using 85 qPCR assays for specific genes, 8 assays for ERCC RNA spikes, and 3 assays for RNAspikes 147. **(B)** Pipetting map for IFC prime step. During prime step, RNAspikes 147 are back-loaded to 8 specific inlets with cell capture beads. **(C)** Pipetting map for front-dosing simulation with ERCC RNA spikes, followed by mRNA-seq chemistry.

The 90-µL preparation volume for PCR contains 1X Advantage 2 PCR Buffer [not short amplicon (SA)] (10X, Clontech, 639206, Advantage® 2 PCR Kit), 0.4-mM dNTP Mix (50X/10 mM, Clontech, 639206), 0.48-µM IS PCR Primer (12 µM, Clontech, 639206), 2X Advantage 2 Polymerase Mix (50X, Clontech, 639206), and 1X Loading Reagent (20X, Fluidigm, 101-1004). All the concentrations correspond to those found in the PCR chambers inside the Polaris IFC. The thermal protocol for preamplification consists of 95°C for 1 min (enzyme activation), five cycles (95°C for 20 s, 58°C for 4 min, and 68°C for 6 min), nine cycles (95°C for 20 s, 64°C for 30 s, and 68°C for 6 min), seven cycles (95°C for 30 s, 64°C for 30 s, and 68°C for 7 min), and final extension at 72°C for 10 min. The preamplified cDNAs are harvested into 48 separate outlets on the Polaris IFC carrier.

qPCR Analysis on Biomark™

Harvested samples from Polaris IFCs were analyzed by qPCR using 96.96 Dynamic Array™ IFCs and the Biomark™ HD system from Fluidigm. Processing of the IFCs and operation of the instruments were performed according to the manufacturer's procedures. For detection using the RNA expression and splice variant assays, a Master Mix was prepared consisting of 360-µL SsoFast™ EvaGreen® Supermix with Low ROX (BioRad 172-5211) and 36-µL 20 × DNA Binding Dye Sample Loading Reagent (Fluidigm 100-5360), and 3.3 µL of this mix was dispensed to



C Sequence of chemistry loading with its mode and function

Chamber	Component	Loading Inlet Port	Loading Mode	Function	Sites	Cell Capture Site Numbers
0	RNA spikes (1, 4, 7)	1, 7, 13, 19, 27, 33, 39, 45	Back-loading with beads	Back dosing cross-talk	8 sites	1, 7, 13, 19, 27, 33, 39, 45
0	ERCC RNA spikes	2	Front-loading	Front dosing cross-talk	Group alt 2 (6 sites)	2, 10, 18, 26, 34, 42
0	Preloading reagent (negative ERCC spikes)	3	Front-loading	Front dosing cross-talk	Group alt 4	4, 12, 20, 28, 36, 44
0	Preloading reagent (negative ERCC spikes)	4	Front-loading	Front dosing cross-talk	Group alt 6	6, 14, 22, 30, 38, 46
0	Preloading reagent (negative ERCC spikes)	5	Front-loading	Front dosing cross-talk	Group alt 8	8, 16, 24, 32, 40, 48
1	Cell lysis mix with K562 total RNA	17	Front-loading	Chemistry	48 sites	
2,3	Reverse transcription mix	16	Front-loading	Chemistry	48 sites	
4,5	PCR mix	14, 15	Front-loading	Chemistry	48 sites	

FIGURE 7 | **(A)** Front-dosing strategy. The capture sites on Polaris IFC are segregated as six groups of eight sites. ERCC spike-in mix is loaded into the second capture site of every group (Group alt 2). For negative control, 1X Preloading Reagent was loaded into Group alt 4, 6, and 8. **(B)** Sequential chambers for multistep chemistry. The capture site with beads holds 2 nL of liquid. The lysis mix is loaded into chamber 1 (9 nL), then reverse transcription mix is loaded into chambers 2 and 3 (9 nL each), and PCR mix is loaded into chambers 4 and 5 (135 nL each). **(C)** Chemistry loading sequence and its function.

each well of a 96-well assay plate. Diluted harvest product (2.7 μL) was added to each well, and the plate was briefly vortexed and centrifuged. Following priming of the IFC in the IFC Controller HX, 5 μL of the sample + Master Mix were dispensed to each sample inlet of the 96.96 IFC. Five microliters of each 10 × assay [5 μM each primer, 1 × assay Loading Reagent (Fluidigm 100-5359)] were dispensed to each Detector Inlet of the 96.96 IFC. After loading the assays and samples into the IFC in the IFC Controller HX, the IFC was transferred to the Biomark HD, and PCR was performed using the thermal protocol GE Fast 96 × 96 PCR + Melt v2.pcl. This protocol consists of a thermal mix of 70°C, 40 min; 60°C, 30 s, hot start at 95°C, 1 min, PCR cycle of 30 cycles of 96°C, 5 s; 60°C, 20 s, and melting using a ramp from 60 to 95°C at 1°C/3 s. Data were analyzed using Fluidigm Real-Time PCR Analysis software using the Linear (Derivative) Baseline Correction Method and the Auto (Global) Ct Threshold Method. The data are exported as a.csv file into an Excel® macro to compile and compare the data against in-house specifications.

RESULTS

Performance Evaluation of Polaris IFC

In order to statistically evaluate the performance of the Polaris IFC, we designed and developed two performance tests: (1) total-RNA-based performance test (RNA PT) and (2) single-cell-based key performance test (KPT). Since single cells are heterogeneous, it would be difficult to evaluate the performance uniformity across 48 capture sites using a cell-based test method. Hence, we developed a 20-cell-equivalent total-RNA PT to evaluate and improve the performance of the Polaris IFC during the initial phase of the IFC development process.

Total RNA-Based Performance Test

The primary objective of this test is to statistically validate a workflow that is very close to the cell-based experiments on the Polaris system and yet collects critical information about uniformity of cDNA synthesis across IFC, reaction line cross-talk (on-IFC), and IFC-to-IFC correlation. To achieve this objective, we simulated steps such as loading of cell capture beads and the thermal step for cell lysis in the total-RNA PT. The workflow of the total-RNA PT is shown in **Figure 6A**. Briefly, the RNA-PT is a two-step procedure. In the first step, the control lines on the Polaris IFC are primed, channels are blocked, and cell capture beads are back-loaded with ArrayControl RNA SPIKES (1, 4, and 7 only, Thermo Fisher Scientific, AM1780; henceforth referred to as RNAspikes 147) in eight specific capture sites (**Figure 6B**).

The ArrayControl RNA Spikes are used to evaluate the back-dosing cross-talk using highly sensitive qPCR assay designed to detect RNA spikes 1, 4, and 7 (three total ArrayControl RNA Spikes). After the priming step, six specific capture sites are loaded with ERCC RNA Spike-In Mix (Thermo Fisher Scientific, 4456740) to estimate the cross-talk for front-loaded reagents and dosing agents. Although the ERCC RNA mix contains 92 spike-ins, only 8 ERCC spike-ins were probed using qPCR assays. The front-dosing strategy is illustrated in **Figure 7A** and the pipetting map is shown in **Figure 6C**.

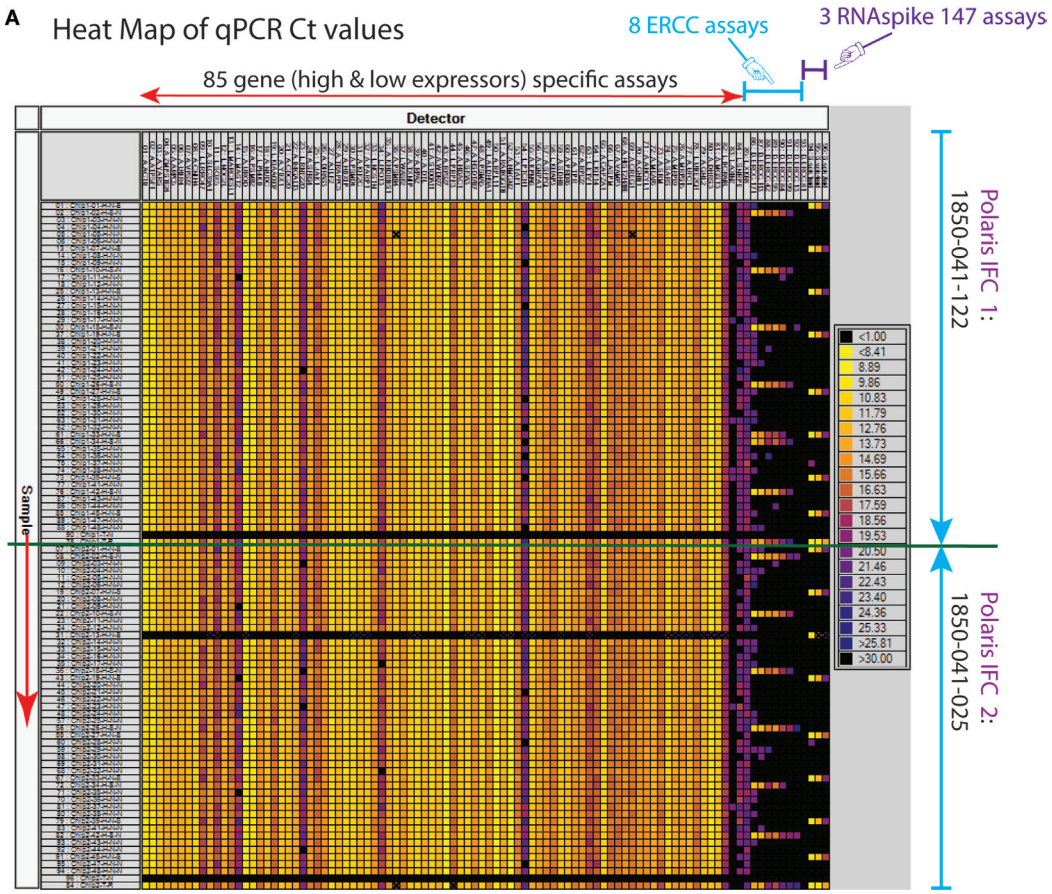
For negative control, 1X Preloading Reagent (Fluidigm, 100-9942) was loaded into specific inlets and capture sites (**Figure 7C**). After completion of front dosing, the mRNA-seq chemistry prep is integrated with the dosing step. The lysis mixture for the RNA-PT contains Leukemia (K562) Total RNA (Thermo Fisher Scientific, AM7832) at a concentration equivalent to 20 cells of total RNA in every cell capture site (48 sites). The cell capture site is serially connected to five chambers to enable multistep reaction chemistry (**Figure 7B**). Cell lysis mixture is loaded into the first 9-nL chamber. Then, RT mixture is loaded in 18-nL volume (two 9-nL chambers). Finally, PCR mixture for preamplification of full-length cDNA is loaded in 270-nL volume (two 135-nL chambers).

The preamplified cDNA is harvested in ~7 μL volume. The harvest is further diluted by addition of 10 μL of DNA Dilution RGT (Fluidigm, 100-9167). In order to evaluate IFC uniformity and other performance metrics, we designed 88 Delta Gene™ assays (Fluidigm) for K562 (85 genes covering high and low expressors) and RNAspikes 147. In addition to these 88 assays, we used 8 ERCC qPCR assays from a published work (Devonshire et al., 2011). In total, we used 96 intercalating dye-based qPCR assays for read-out of RNA PT on an M96.96 Dynamic Array™ IFC (Fluidigm). We routinely test positive and negative tube controls for every chemistry preparation by qPCR assays on the M96.96. In order to do this, 2 out of the 48 samples from a Polaris IFC are replaced by positive and negative tube controls on M96.96. The positive tube control contains total RNA from K562, RNAspikes 147, and ERCC. The tube controls are used to validate the functionality of chemistry preparation on a particular day. Harvest products from two Polaris IFC are tested on a single M96.96 Dynamic Array IFC run. A typical qPCR Ct heat map and associated Excel macro for two Polaris RNA PTs are shown in **Figures 8A,B**. In order to statistically validate the performance, we tested 44 Polaris IFCs with RNA PTs. IFC and reagents from minimum of three manufacturing lots were used. Tolerance limit or interval analyses were performed on more than 40 Polaris IFC runs. The distribution of data and tolerance limit analyses for different metrics for the RNA PTs are shown in **Figure 9**.

Single-Cell-Based Key Performance Test

Key performance test was developed and validated using one cell type each for suspension (K562) and adherent (BJ fibroblast) cells. As described in Section "Materials and Methods," cells are stained with the universal fluorescent marker, CTO. A subset of these cells were stained for surface marker using antibody conjugated with Alexa 647. In the case of K562, we used Anti-Human CD59-Biotin (BD Biosciences, 555762) with Streptavidin Alexa 647, and for BJ fibroblast, we used Anti-Mouse/Human CD44-Alexa 647 (BioLegend 103018). The double-stained cells (universal CTO and surface marker Alexa 647) were mixed with cells stained with CTO only to achieve three different purity percentages (3, 10, and 50%). The cells were selected for universal CTO and surface marker. For BJ fibroblasts, we tested two different workflows, one with cell

A Heat Map of qPCR Ct values



B Typical macro output results for two Polaris IFC (above heat map)

Worksheet, Polaris 48 mRNASeq RNA Perf Test (100-9194 A6)

Import CSV and Calculate Metrics — <-- Click button to import csv file, and calculate the pass/fail and information only metrics

Chip Run Information		Chip 1 Final Test Results	Chip 2 Final Test Results
CSV Filename (Barcode)	1362170343		
Chip Run Filename	L:\CassandraGreene\Polaris\Superman Performance Tests\Superman Perform		
Application Version	4.0.1		
Application Build	20130423.08		
Quality Threshold	0.65		
Baseline Method	Linear (Derivative)		
Ct Threshold Method	Auto (Global)		
Chip Type	96.96		

Pass or Fail Metrics	Chip 1 Value	Chip 2 Value	Threshold 1	Threshold 2	Chip 1 Pass/Fail	Chip 2 Pass/Fail
Average Assay STD	0.21	0.22			Pass	Pass
% Assay Dropouts	0.1%	0.0%			Pass	Pass
Overall % Assay Dropouts	1.5%	1.7%			Pass	Pass
Reaction Line Cross-talk	0.0%	0.0%			Pass	Pass
Dosing Cross-talk	0.6%	0.0%			Pass	Pass
Chip and Ref Correlation (Slope)	1.023	1.013	0.9	1.1	Pass	Pass
Chip and Ref Correlation (R2)	0.976	0.985	0.9		Pass	Pass
Number of No Calls	0	96		96	Pass	Pass
Metrics for Information Only	**Chip 1 Value**	**Chip 2 Value**	**Target 1**	**Target 2**	**Chip 1 Meet Target**	**Chip 2 Meet Target**
Number of Assays Pass STD	100.0%	98.4%	50%		Meet	Meet
Number of Correlated Assays	100.0%	100.0%	95%	2	Meet	Meet

Pass | Pass

FIGURE 8 | **(A)** Typical heat map of high-throughput qPCR assay for RNA-based performance test. The M96.96 IFC (96 samples) can accept amplicons from two Polaris IFCs (48 samples each). For every Polaris IFC, we replace two samples with positive and negative control samples. The columns are assays (85 high- and low-expressing assays; 8 ERCC spike assays; and 3 RNAspike 147 assays). The rows are diluted amplicons from Polaris IFCs. **(B)** Excel macro for the RNA-based performance test.

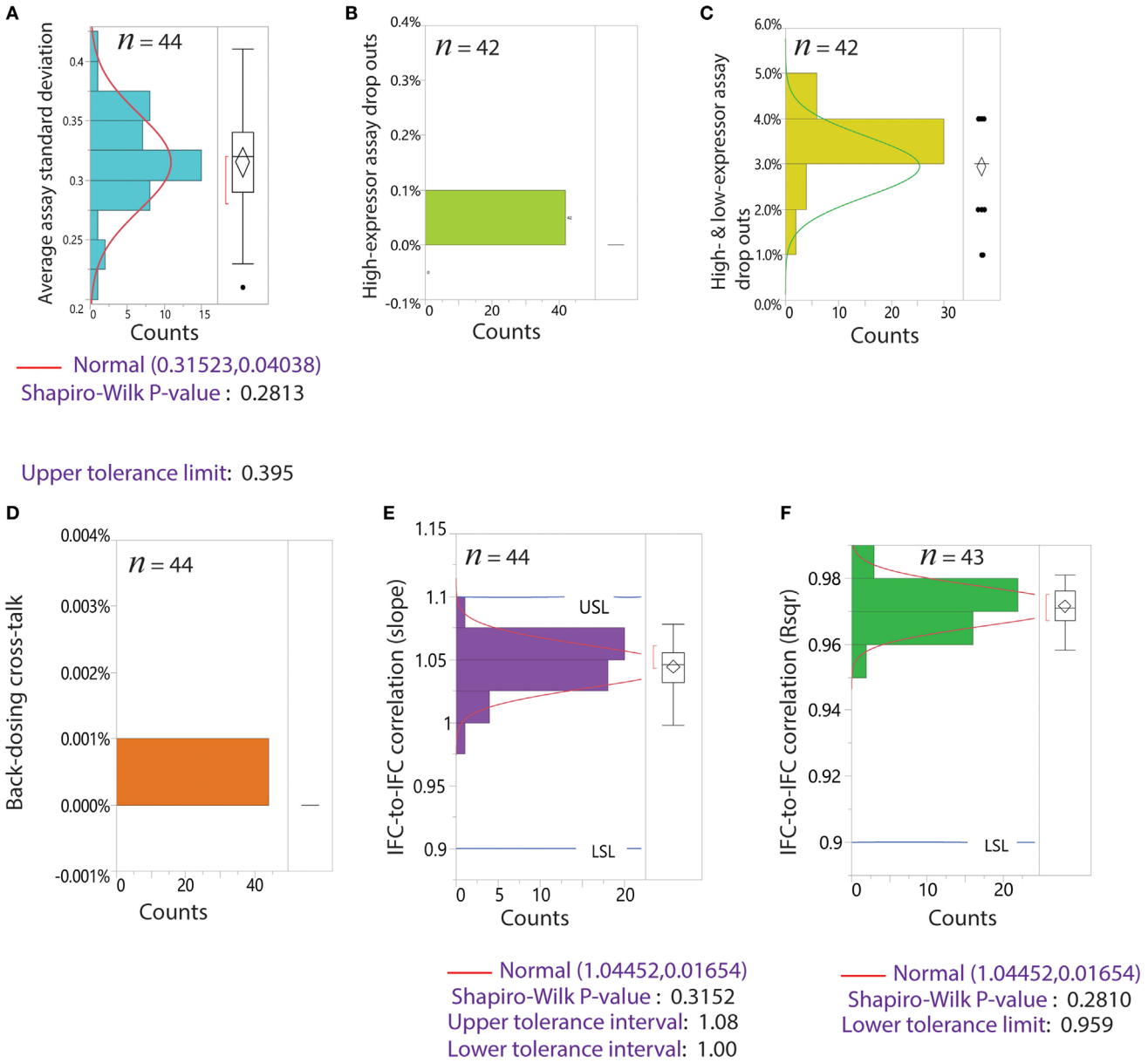


FIGURE 9 | Tolerance limit analyses from 44 Polaris IFC runs using RNA-based performance test. *n* is the sample number after removal of outliers. Outliers are estimated by Tukey's box plot. The percentage of outliers removed ranged from 0 to 5%. **(A)** Average assay SD. The data fit a normal distribution with goodness of fit *P*-value of 0.2813. The upper tolerance interval for this metric is 0.395. **(B)** Distribution of drop outs for high-expressing assays only. **(C)** Distribution of drop outs for both high and low expressors. **(D)** Back-loaded dosing cross-talk. **(E)** Distribution of slope for IFC-to-IFC correlation. The data fit a normal distribution with tolerance interval of 1–1.08. **(F)** Distribution of R^2 for IFC-to-IFC correlation. The data fit a normal distribution with lower tolerance limit of 0.959.

selection followed by chemistry (immediate) and another to dose the BJ fibroblasts with medium every 4 h for 24-h adherent culture, followed by chemistry (BJ dosing). In order to evaluate the cell viability prior to the cell lysis step, we used Zombie Yellow™ cell viability stain (BioLegend, 423103; λ_{ex} = 396 nm and λ_{em} = 572 nm), which stains dead cells. Performance metrics, such as number of sites occupied with single cells out of the total 48 sites (cell selection), number of cells retained after dosing and prior to cell lysis (cell retention), and number of viable cells prior to lysis, were evaluated. On average from 20 Polaris IFC runs, our cell selection was ~95% for K562 and BJ fibroblast with different purity percentages (**Figure 10A**). For cell retention,

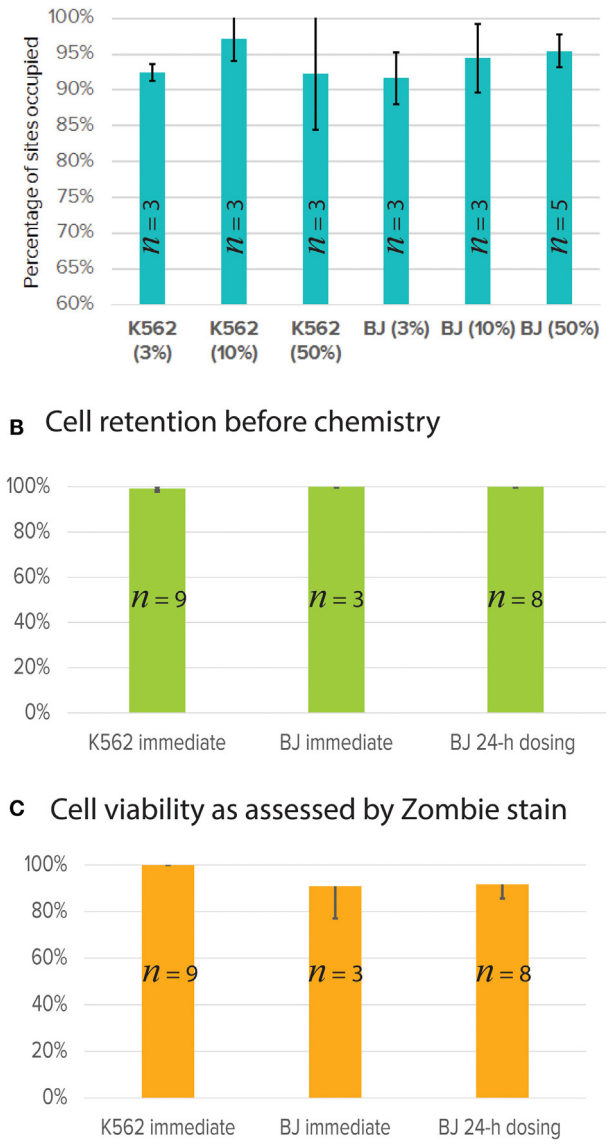


FIGURE 10 | Single-cell selection, retention, and viability performance using single-cell key performance test. (A) Percentage of single cells selected (out of 48 total capture sites) for K562 (suspension) and BJ fibroblast (adherent) at different cell purity concentrations (3, 10, and 50%). *n* is number of Polaris IFCs tested. **(B)** Single cells retained after dosing and culture. Cells are imaged and counted before the cell lysis step. K562 and BJ fibroblasts were selected and immediately processed for mRNA-seq chemistry preparation. BJ fibroblasts were also selected, cultured, and dosed with cell culture medium for 24 h (BJ dosing). Data from three IFCs are presented here. **(C)** Single-cell viability as assessed by Zombie stain on-IFC.

>47/48 sites showed presence of single cells as enumerated after the cell selection step and prior to cell lysis step (**Figure 10B**). The average cell viability was ~90% as estimated from 20 Polaris IFC runs (**Figure 10C**).

The Polaris system generates very high quality (size distribution) and quantity (yield) of preamplified cDNA from single cells. The size distribution of preamplified cDNA from single cells, as evaluated using Bioanalyzer 2100 and the DNA high-sensitivity chip (Agilent), is typically in the size range of 0.3–7 kb (**Figure 11A**). For yield, preamplified cDNA from single cells was quantified using PicoGreen-based dsDNA quantification assay (Quant-iT™ PicoGreen® dsDNA Assay Kit, Thermo Fisher Scientific, P7589). The average total cDNA yield per single K562 cell is 38.42 ± 8.08 ng (**Figure 11B**). We randomly selected ~14 single cells from three Polaris IFC runs, barcoded them using modified Nextera library prep, and pooled them to generate a single sequencing library. A representative library profile from 42 single cells is shown in **Figure 11C**. The majority of the single-cell library falls in the range of 200–2,000 bp. For three sequencing libraries from nine Polaris IFCs tested with K562 immediate chemistry, sequencing data from three MiSeq™ runs using v2 150 bp PE kit were compiled, and tolerance limits (90% confidence with 95% population coverage) were estimated for two key sequencing metrics (**Figure 12**). The average percentage of reads mapping to rRNA/total reads is 0.122%. The Box–Cox transformed data fit a normal distribution with a Shapiro–Wilk *P*-value of 0.0983. Based on the normal distribution, the upper tolerance limit for percentage of reads mapping to rRNA is 0.3% (**Figure 12A**). The mean number of genes detected is 6,967 ± 115. The data fit a normal distribution with a lower tolerance limit of 5,919 genes as estimated from 115 single-cell datapoints (**Figure 12B**).

We extensively analyzed our single-cell sequencing data for transcript coverage bias and possible positional bias of single cells selected across 48 capture sites on the Polaris IFC (**Figure 13**). We noted uniform coverage along the transcript length (**Figure 13B**) without any positional bias on the Polaris IFC. The plot of normalized coverage vs. normalized distance along the transcript with respect to capture sites (2, 3, 4 and 40, 41, 42) from a Polaris IFC is shown in **Figure 13B**. A plot of median 3′ end bias of transcript coverage with respect to capture site number indicates no positional bias across three Polaris IFC runs (**Figure 13A**). In order to understand if there is any possible effect of hypoxia on single cells due to spatial location of capture sites on the Polaris IFC, we analyzed the expression value of *HIF1A* gene across different capture sites. Up-regulation of hypoxia-induced factor 1 gene (*HIF1A*) is a known consequence of hypoxia (Choudhry and Mole, 2015). Expression analyses of *HIF1A* did not show any positional bias with respect to the capture sites (**Figure 13C**). It should be noted that we recommend strictly following the Polaris workflow as described in the Polaris protocol document (Fluidigm, 101-0082). Any deviation from the validated workflow might lead to introduction of possible bias at multiple levels.

Sensitivity Studies Using ERCC Spike-Ins

An alternative way to evaluate performance of single-cell mRNA-seq on the Polaris system is to implement use of the ERCC RNA Spike-In Mix 1 in the lysis mix. The ERCC control mix consists of 92 polyadenylated transcripts with a size range of 273–2,022 bases and six orders of magnitude range in concentration. We tested both qPCR- and sequencing-based methods for detection of ERCC spikes. Ninety-two primer pairs were designed

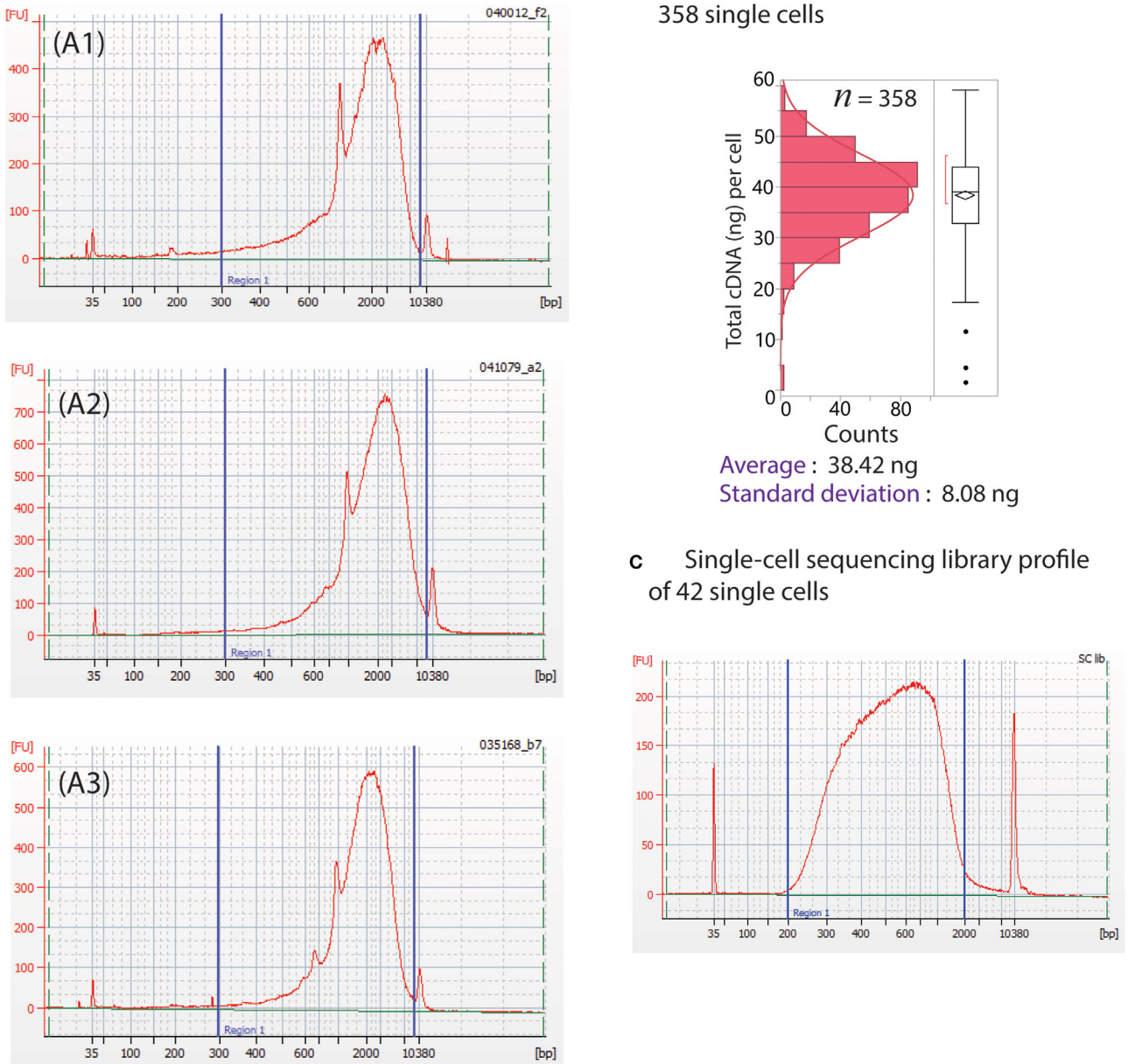


FIGURE 11 | **(A)** Typical cDNA distribution from single cells from three different Polaris IFC runs (A1–A3). The peak for RNAspikes 147 is around 1,000 bp. **(B)** Distribution of cDNA yield from 358 single K562 cells from multiple Polaris IFC runs. The average cDNA yield per cell is 38.42 ng with a SD of 8.08 ng. **(C)** Library profile of 42 single cells.

to target the corresponding transcripts for qPCR testing. qPCR was performed on a Fluidigm 96.96 Dynamic Array IFC. Stochastic distribution of transcripts was observed when the input concentration was 25 copies per reaction or less on the Polaris IFC followed by qPCR detection on the 96.96 Dynamic Array (**Figure 14A**). Single-copy RNA detection is demonstrated, although intermittently, likely due to sampling at the reaction site. Transcripts at 1.6 copies per reaction were intermittently detected by qPCR on the 96.96 Dynamic Array IFC. We also evaluated the detection rate of ERCC spikes (>1.6 copies/reaction) using an approach based on massive parallel sequencing. There were 7 ERCC spikes (ERCC-00170; ERCC-00148; ERCC-00126; ERCC-00099; ERCC-00054; ERCC-00163; ERCC-00059), which were at a concentration of 1.6 copies per Polaris reaction chamber. One of the 7 ERCC spikes (ERCC-00054) was not detected in any of the 19 single-cell samples. If we remove this datapoint as an outlier, the average detection rate of ~1.6 copies is 28%. Based on Poisson estimates, single-copy detection rate should be ~33%

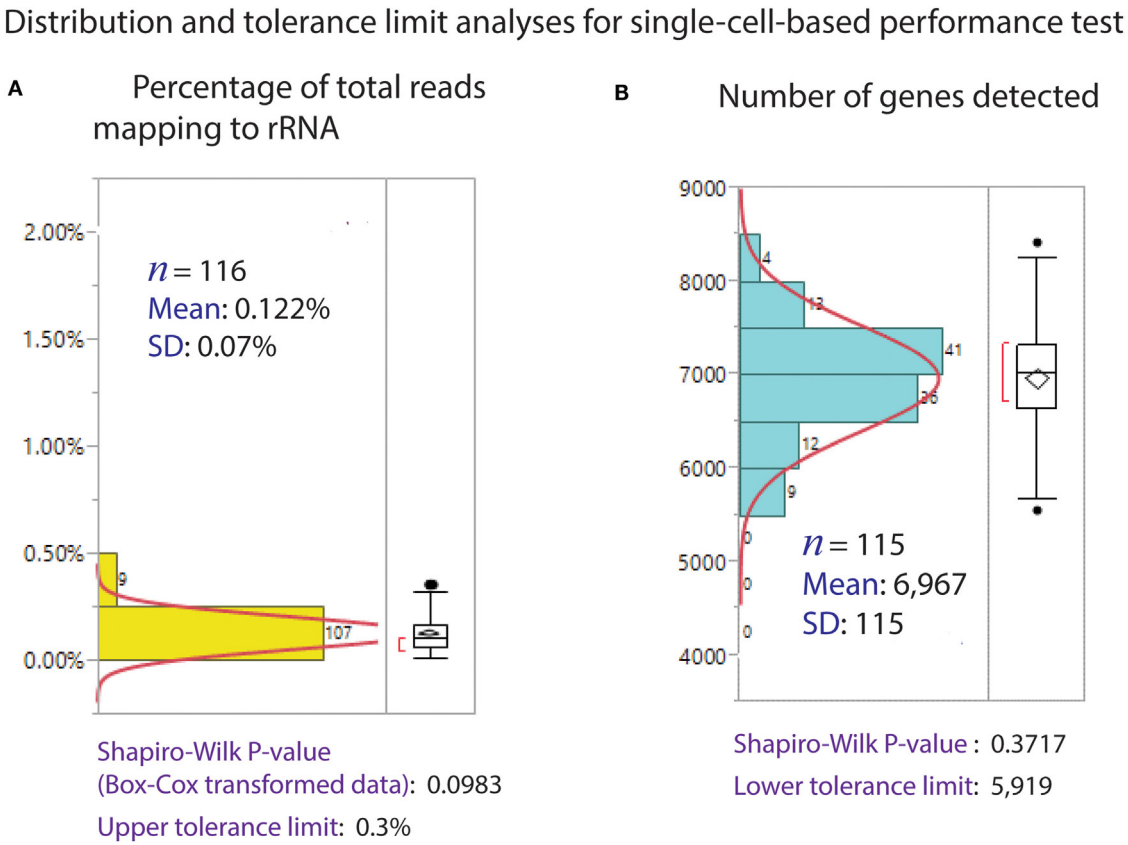


FIGURE 12 | Distribution of sequencing metrics for single cells. **(A)** Percentage of reads mapped to rRNA/total reads. The average mapping to rRNA is 0.122% from 116 single-cell sequence datapoints. The upper tolerance limit for percentage rRNA mapping is 0.3%. **(B)** Number of genes detected (TPM > 1). Distribution of genes detected from 115 single-cell sequence data. The average genes detected is 6,967, with a SD of 115. The data fit a normal distribution, with a lower tolerance limit of 5,919.

(67% should be a failure event). The single-copy detection rate (28%) from the Polaris system is very close to expected theoretical estimates based on Poisson statistics (**Figure 14B**).

Single-Cell Transfection of nGFP mRNA and GFP Expression Analyses

In order to demonstrate dosing and functional response analyses, we transfected single K562 cells with nuclear green fluorescent protein (nGFP) mRNA and cultured the transfected single K562 for 16 h. During this culture duration, the cells translated the nGFP mRNA and expressed the GFP inside the cell. The imaging capability of Polaris enables monitoring of GFP expression. Subsequently, the cells were processed for mRNA-seq chemistry on-IFC and sequenced on MiSeq to quantify the reads mapped to GFP. K562 cells stained with CTO were selected on Polaris IFC. To carry out the single-cell transfection, 10 µL of Stemfect RNA transfection reagent was mixed with 240-µL Stemfect transfection buffer (Stemgent® Stemfect™ RNA Transfection Kit, 00-0069) (Mix A). The stock nGFP mRNA (Stemgent, 05-0019) at 100 ng/µL was diluted with Stemfect transfection buffer first and then further diluted with Mix A to make mRNA transfection complex. This complex was incubated at room temperature for 15 min and further diluted with K562 cell culture medium (refer to Section "K562 Cell Culture and CD59 Staining") to achieve different final concentrations (0.5 and 1 ng/µL) of nGFP mRNA. Selected single K562 cells were cultured with the nGFP mRNA transfection complex with culture medium at 37°C with 5% CO_2 on the Polaris IFC. During this cell culture incubation time, images were taken every hour to monitor the onset of GFP expression. Image analyses (**Figure 15**) showed that single cells picked up nGFP mRNA at 0.5 and 1 ng/µL concentrations and expressed the green fluorescent proteins, thereby reinforcing the fact that single cells on Polaris IFC are healthy and are able to uptake naked mRNA and translate it to protein capable of being transfected. **Figure 15A** shows typical time-series images, which can be obtained from the Polaris system. It should be noted that for this particular experiment, the imaging interval was set to 1 h, but the Polaris system is capable of taking successive images in a rapid mode. We noted onset of GFP gene expression around the 3-h time frame at single-cell resolution. The cDNA pool showed a length range from 0.3 to 9.2 kb, with an average length ~2 kb. It is also noted that >85% of the total cDNA pool lies between 0.5 and 9.2 kb (**Figure 15B**). Sequencing data show that the cells transfected with nGFP mRNA harbored the

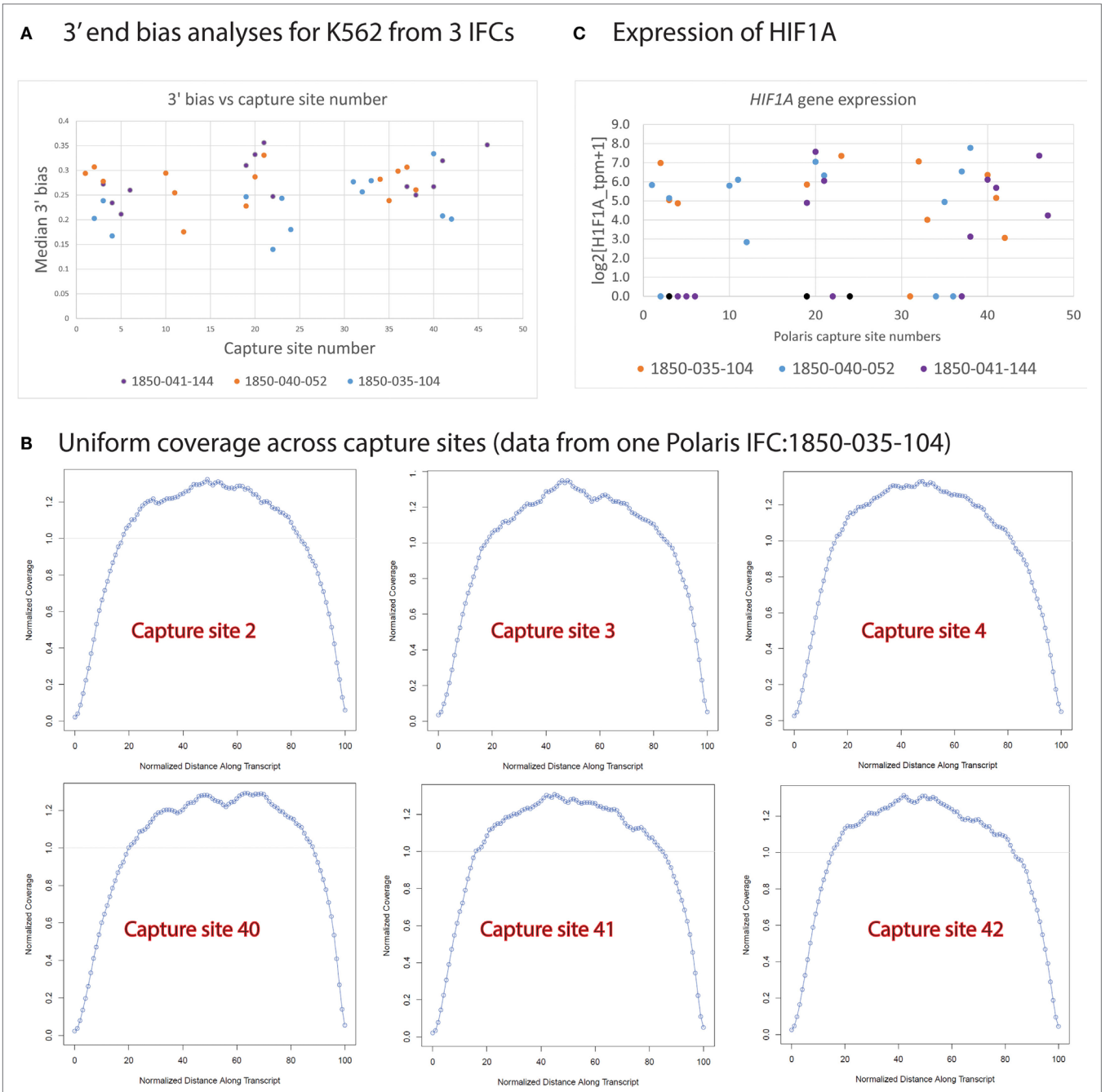


FIGURE 13 | Bias vs. capture site number. **(A)** Median 3′ end bias was analyzed for random single cells picked for sequencing. No 3′ bias was noted for cells from three Polaris IFC runs **(B)** *HIF1A* gene expression of cells across different capture sites. **(C)** Sequencing coverage as estimated by Picard for single cells. Uniform coverage noted across different capture sites from a Polaris IFC.

extracellular mRNA even after 16 h of culture. As expected, the control cells without nGFP transfection did not show any mapping to GFP sequence. The transfection of nGFP did not alter the mapping to genome and transcriptome when compared to the control cells (**Figure 15C**). The nGFP-transfected cells showed percentage average reads mapping of 0.57, 87.99, and 47.82% to GFP, genome, and transcriptome, respectively ($n = 7$), while the control K562 showed percentage average mapping of 0, 88.03, and 49.22 ($n = 7$).

DISCUSSION

In this work, we report design and development of an integrated system to perform functional studies on single cells.

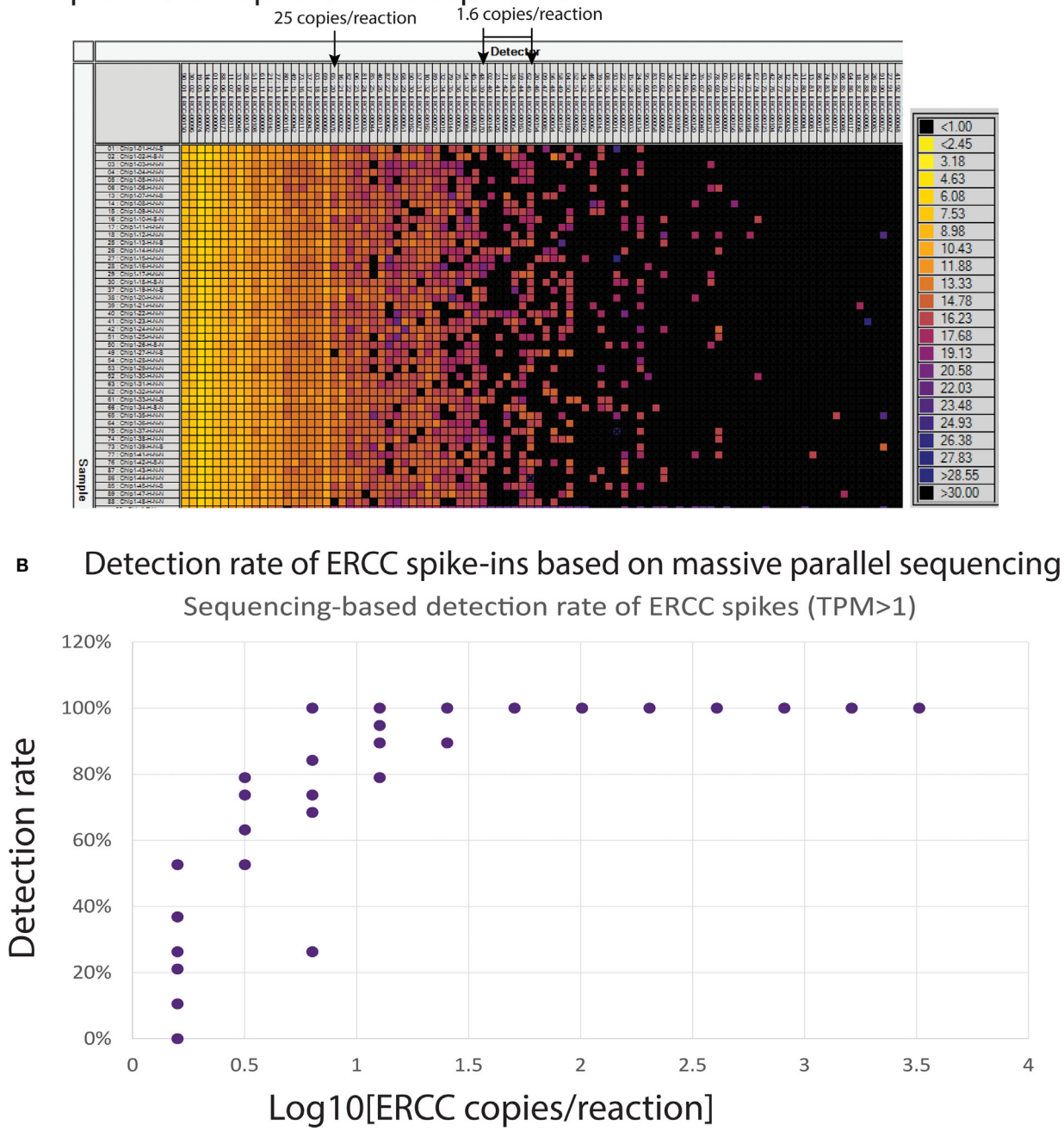


FIGURE 14 | **(A)** Heat map (qPCR assay) for ERCC spike-in with single SKBR3 cells. The columns are 92 qPCR assays designed and developed for ERCC spike-in. The rows are amplicons from single SKBR3 cells with ERCC spike-in. The ERCC assays on the columns are sorted by concentrations. **(B)** Detection rate of ERCC spikes with >1.58 copies per Polaris IFC chambers.

We developed a nanoscale IFC, which employs fluidic logic to actively select single cells, and a system capable of performing multiple functionalities. The performance of the developed IFC and system was extensively tested using RNA-based and single-cell-based performance tests. These tests were specifically designed to evaluate different functionalities of the IFC and system. The functional capability of the Polaris IFC and system has been successfully demonstrated using transfection of naked nGFP mRNA, followed by monitoring of nGFP expression and finally analysis of the whole set of mRNA transcripts by massive parallel sequencing. It is noted that it is not currently possible to perform studies reported in this work on any other single-cell platforms. The limitation of the current system includes limited number of cells for functional studies

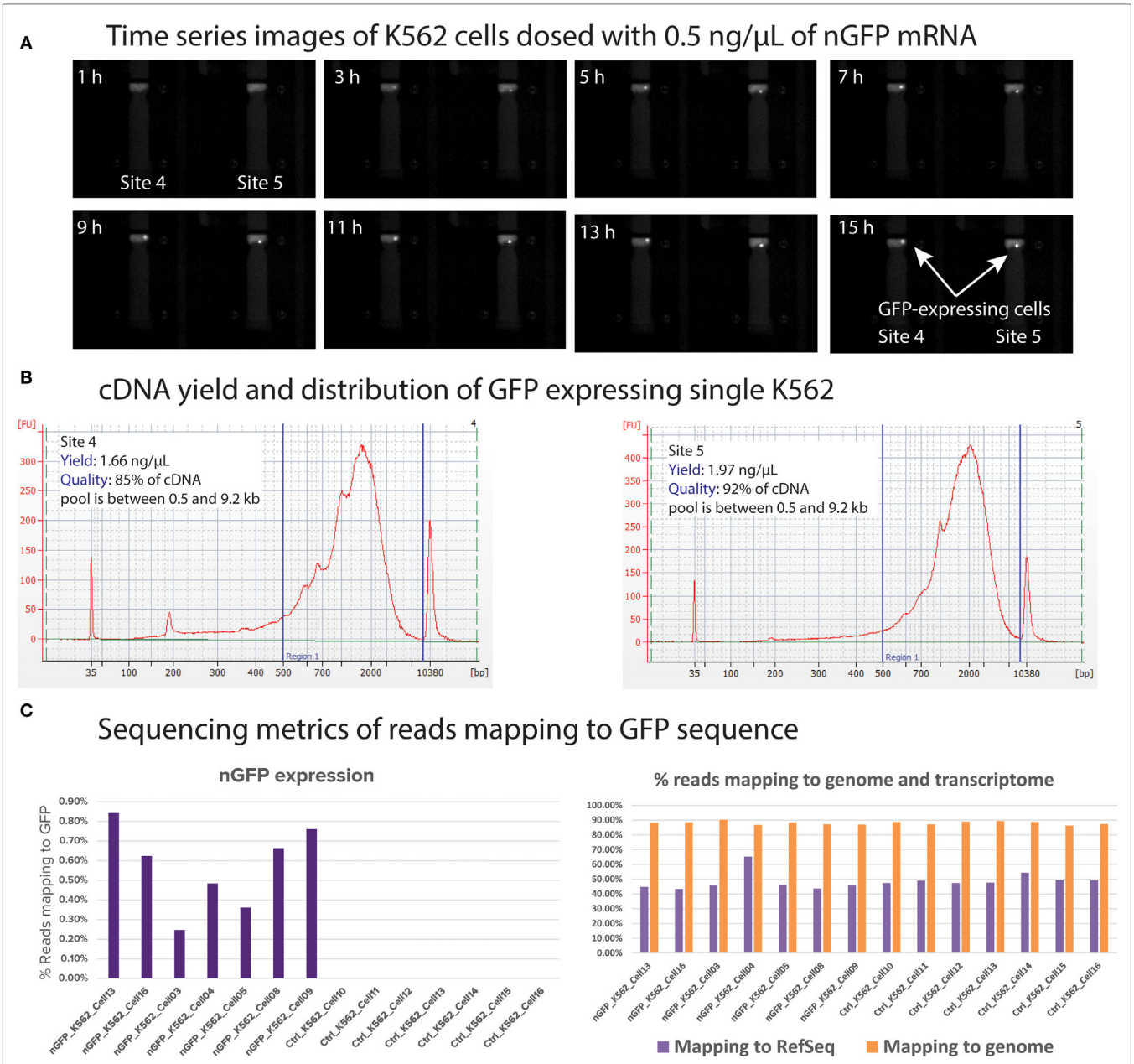


FIGURE 15 | **(A)** Time-series images from the Polaris system taken every hour to monitor onset of GFP expression. For visual clarity, only two cell capture sites are shown. The onset GFP expression is noted around 3 h from the time of transfection. **(B)** cDNA yield and distribution from single K562 processed through mRNA-seq chemistry. **(C)** Mapping metrics of GFP-transfected single K562 cells to genome, transcriptome, and GFP sequence.

(up to 48 cells). However, the requirement on the number of cells depends on the biological question, and it is possible to expand the capability of the IFC consumable to process more cells in the future.

AUTHOR CONTRIBUTIONS

NR, BF, JS, and JAAW conceived and designed the RNA-based performance test; BF, LS, AAL, JS, and JAAW conceived and designed the single-cell-based performance test; NR, LS, AAL, JS, MLG, CDS, NDA, CG, CTL, IH, AO, CS, and JAAW performed experiments; BF, NSGKD, MZ, EO, and CP were involved in Polaris IFC development; KH, MTM, WY, MN, CC, ML, HC, ZH, LL, CC, and ZS were involved in Polaris system development; RY, WH, JA, and ZH were involved in Polaris software development; NR, LS, JS, CDS, XW, and JAAW analyzed the data; TS edited the manuscript; CL drafted the Polaris user guide; MU and JAAW supervised the project, helped with design and interpretation, and provided laboratory space and financial support; and NR, LS, KJL, and JAAW wrote the manuscript with input from all authors. All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Alix-Panabières, C., Bartkowiak, K., and Pantel, K. (2016). Functional studies on circulating and disseminated tumor cells in carcinoma patients. *Mol. Oncol.* 10, 443–449. doi:10.1016/j.molonc.2016.01.004

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232. doi:10.1038/nmeth.3728

Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H. B., Trombetta, J. J., et al. (2015). Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell* 162, 1309–1321. doi:10.1016/j.cell.2015.08.027

Bendall, S. C., Simonds, E. F., Qiu, P., Ad, A., Krutzik, P. O., Finck, R., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332, 687–696. doi:10.1126/science.1198704

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., et al. (2013). An estimation of the number of cells in the human body. *Ann. Hum. Biol.* 40, 463–471. doi:10.3109/03014460.2013.807878

Briggs, S. F., Dominguez, A. A., Chavez, S. L., and Reijo Pera, R. A. (2015). Single-cell XIST expression in human preimplantation embryos and newly reprogrammed female induced pluripotent stem cells. *Stem Cells* 33, 1771–1781. doi:10.1002/stem.1992

Cayrefourcq, L., Mazard, T., Joosse, S., Solassol, J., Ramos, J., Assenat, E., et al. (2015). Establishment and characterization of a cell line from human circulating colon cancer cells. *Cancer Res.* 75, 892–901. doi:10.1158/0008-5472.CAN-14-2613

Choudhry, H., and Mole, D. R. (2015). Hypoxic regulation of the noncoding genome and NEAT1. *Brief. Funct. Genomics*. 15, 174–185. doi:10.1093/bfgp/elv050

Darmanis, S., Gallant, C. J., Marinescu, V. D., Niklasson, M., Segerman, A., and Flamourakis, G. (2016). Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* 14, 380–389. doi:10.1016/j.celrep.2015.12.021

Devaraju, N. S. G. K., and Unger, M. A. (2012). Pressure driven digital logic in PDMS based microfluidic devices fabricated by multilayer soft lithography. *Lab. Chip* 12, 4809–4815. doi:10.1039/c2lc21155f

Devonshire, A. S., Elaswarapu, R., and Foy, C. A. (2011). Applicability of RNA standards for evaluating RT-qPCR assays and platforms. *BMC Genomics* 12:118. doi:10.1186/1471-2164-12-118

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33, 285–289. doi:10.1038/nbt.3129

Ennen, M., Keime, C., Kobi, D., Mengus, G., Lipsker, D., Thibault-Carpentier, C., et al. (2014). Single-cell gene expression signatures reveal melanoma cell heterogeneity. *Oncogene* 34, 3251–3263. doi:10.1038/onc.2014.262

Fan, H. C., Fu, G. K., and Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science* 347, 1258367. doi:10.1126/science.1258367

Frei, A. P., Bava, F.-A., Zunder, E. R., Hsieh, E. W. Y., Chen, S.-Y., Nolan, G. P., et al. (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* 13, 269–275. doi:10.1038/nmeth.3742

Gao, D., Vela, I., Sboner, A., Iaquinta, P. J., Karthaus, W. R., Gopalan, A., et al. (2014). Organoid cultures derived from patients with advanced prostate cancer. *Cell* 159, 176–187. doi:10.1016/j.cell.2014.08.016

Guo, G., Pinello, L., Han, X., Lai, S., Shen, L., Lin, T.-W., et al. (2016). Serum-based culture conditions provoke gene expression variability in mouse embryonic stem cells as revealed by single-cell analysis. *Cell Rep.* 14, 956–965. doi:10.1016/j.celrep.2015.12.089

Kim, K.-T., Lee, H. W., Lee, H.-O., Kim, S. C., Seo, Y. J., Chung, W., et al. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* 16, 1–15. doi:10.1186/s13059-015-0692-3

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi:10.1016/j.cell.2015.04.044

Macaulay, I. C. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522. doi:10.1038/nmeth.3370

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002

Pollen, A. A., Nowakowski, T. J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C. R., et al. (2015). Molecular identity of human outer radial glia during cortical development. *Cell* 163, 55–67. doi:10.1016/j.cell.2015.09.004

Saadatpour, A., Guo, G., Orkin, S. H., and Yuan, G.-C. (2014). Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis. *Genome Biol.* 15, 1–13. doi:10.1186/s13059-014-0525-9

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369. doi:10.1038/nature13437

Unger, M. A., Chou, H.-P., Thorsen, T., Scherer, A., and Quake, S. R. (2000). Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* 288, 113–116. doi:10.1126/science.288.5463.113

Wilson, N. K., Kent, D. G., Buettner, F., Shehata, M., Macaulay, I. C., Calero-Nieto, F. J., et al. (2015). Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* 16, 712–724. doi:10.1016/j.stem.2015.04.004

Yu, M., Bardia, A., Aceto, N., Bersani, F., Madden, M. W., Donaldson, M. C., et al. (2014). Ex vivo culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science* 345, 216–220. doi:10.1126/science.1253533

Conflict of Interest Statement: All authors are employees of Fluidigm Corporation.

frontiers in Cell and Developmental Biology

METHODS
published: 05 October 2016
doi: 10.3389/fcell.2016.00100

High-Sensitivity Mass Spectrometry for Probing Gene Translation in Single Embryonic Cells in the Early Frog (*Xenopus*) Embryo

Camille Lombard-Banek[1], *Sally A. Moody*[2] *and Peter Nemes*[1]*

[1] *Department of Chemistry, The George Washington University, Washington, DC, USA,* [2] *Department of Anatomy and Regenerative Biology, The George Washington University, Washington, DC, USA*

Direct measurement of protein expression with single-cell resolution promises to deepen the understanding of the basic molecular processes during normal and impaired development. High-resolution mass spectrometry provides detailed coverage of the proteomic composition of large numbers of cells. Here we discuss recent mass spectrometry developments based on single-cell capillary electrophoresis that extend discovery proteomics to sufficient sensitivity to enable the measurement of proteins in single cells. The single-cell mass spectrometry system is used to detect a large number of proteins in single embryonic cells in the 16-cell embryo of the South African clawed frog (*Xenopus laevis)* that give rise to distinct tissue types. Single-cell measurements of protein expression provide complementary information on gene transcription during early development of the vertebrate embryo, raising a potential to understand how differential gene expression coordinates normal cell heterogeneity during development.

Keywords: single-cell analysis, mass spectrometry, proteomics, cell differentiation, *Xenopus laevis*

OPEN ACCESS

Edited by:
Xinghua Pan,
Yale University, USA

Reviewed by:
Raman Chandrasekar,
Kansas State University, USA
Vasudevan Seshadri,
National Centre for Cell Science, India
Qing-Yu He,
Jinan University, China

***Correspondence:**
Peter Nemes
petern@gwu.edu

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *01 June 2016*
Accepted: *29 August 2016*
Published: *05 October 2016*

Citation:
Lombard-Banek C, Moody SA and Nemes P (2016) High-Sensitivity Mass Spectrometry for Probing Gene Translation in Single Embryonic Cells in the Early Frog (Xenopus) Embryo. Front. Cell Dev. Biol. 4:100. doi: 10.3389/fcell.2016.00100

INTRODUCTION

Single-cell analysis technologies are essential to understanding cell heterogeneity during normal development and disease. Characterization of the genomes and their expression at the levels of the transcriptome, proteome, and metabolome provides a molecular window into basic cell processes. Singe-cell measurements complement traditional cell population-averaging approaches by enabling studies at the level of the building blocks of life, where many critical processes unfold (Raj and van Oudenaarden, 2008; Altschuler and Wu, 2010; Singh et al., 2010; Zenobi, 2013). For example, by studying individual cells, it is possible to ask how cells give rise to all the different types of tissues in the body (stem cells) and specialize for defense (immune cells), communication (neurons), and support (glia). This information in turn lays the foundation to developing diagnosis and treatments for addressing pressing health concerns, such as emergence of drug resistant bacteria, onset and development of neurodegeneration, and cancer, as well as infections.

Single-cell investigations take advantage of rapid developments in technology. With more than million-fold amplification of DNA and RNA and the commercialization of high throughput DNA and RNA sequencing, it is now possible to query cell-to-cell differences (Kolisko et al., 2014; Mitra et al., 2014), including but not limited to chromosomal mosaicism in tissues (Vijg, 2014; Gajecka, 2016) and embryonic somatic cells (Liang et al., 2008; Jacobs et al., 2014), establishment of cell heterogeneity in the nervous system

(McConnell et al., 2013), and mutations during disease states (Junker and van Oudenaarden, 2015; Kanter and Kalisky, 2015). How gene expression translates into the functionally important proteins and how they then feedback to modulate gene expression is essential to systems cell biology. Multiple reports found differences between transcription and translation (Vogel and Marcotte, 2012; Smits et al., 2014; Peshkin et al., 2015), and transcription is known to be controlled by translational factors during development (Radford et al., 2008); therefore, characterization of the proteome is critical to understanding cell heterogeneity. Translational cell heterogeneity has traditionally been measured by immunohistochemistry and Western blot analyses. Protein-targeted assays have recently gained substantial throughput by the development of mass cytometry (CyTOF), which uses inductively coupled plasma and mass spectrometry (MS) to simultaneously quantify ∼35 different proteins tagged with rare earth elements in thousands of cells. This level of multidimensionality has promoted applications in cell differentiation during erythropoiesis (Bendall et al., 2011), and was recently coupled to laser-ablation to spatially survey cell heterogeneity in the tumor environment (Giesen et al., 2014).

Cell heterogeneity has functional implications during embryonic development. Over four decades of innovative embryological manipulations combined with gene-by-gene identifications and functional characterizations in *Xenopus* have shown that molecular asymmetries in the distribution of maternal mRNAs occur upon fertilization and lead to the formation of the three primary germ layers and the germ line (King et al., 2005; Lindeman and Pelegri, 2010). Recent approaches have defined the spatial and temporal changes of mRNAs and abundant proteins and metabolites in the whole embryo (Flachsova et al., 2013; Wuhr et al., 2014; De Domenico et al., 2015). However, very little is known about how these molecules change over time in individual blastomere lineages as they acquire germ layer and body axis fates. In many animals, mRNAs that are synthesized during oogenesis are sequestered to different cytoplasmic domains (Davidson, 1990; Sullivan et al., 2001), which after fertilization then specify the germ cell lineage (King et al., 2005; Haston and Reijo-Pera, 2007; Cuykendall and Houston, 2010) and determine the anterior-posterior and dorsal-ventral axes of the embryo (Heasman, 2006b; Kenyon, 2007; Ratnaparkhi and Courey, 2007; White and Heasman, 2008; Abrams and Mullins, 2009). For example, in *Xenopus* several mRNAs are localized to the animal pole region, which later gives rise to the embryonic ectoderm and the nervous system (Grant et al., 2014), whereas localization of VegT mRNA to the vegetal pole specifies endoderm formation (Xanthos et al., 2001), and region-specific relocalization of the Wnt and Dsh maternal proteins govern the dorsal-ventral patterning of the embryo (Heasman, 2006a; White and Heasman, 2008). However, there is abundant evidence that in developing systems not all transcripts are translated into proteins; therefore, analyses of the mRNAs may not reveal the activity state of the cell. In fact, different animal blastomeres of the 16-cell *Xenopus* embryo that are transcriptionally silent can have very different potentials to give rise to neural tissues (Gallagher et al., 1991; Hainski and Moody, 1992; Yan and Moody, 2007), even though they appear to express common mRNAs (Grant et al., 2014; Gaur et al., 2016).

High-resolution MS is the technology of choice for the analysis of the proteome (Aebersold and Mann, 2003; Guerrera and Kleiner, 2005; Walther and Mann, 2010; Zhang et al., 2013). Using millions of cells, contemporary MS enables the discovery (untargeted) characterization of the encoded proteomes of various species in near complete coverage, as recently demonstrated for the yeast (Hebert et al., 2014), mouse (Geiger et al., 2013), and human (Wilhelm et al., 2014). Recent whole-embryo analyses by MS revealed that transcriptomic events are accompanied by gross proteomic and metabolic changes during the development of *Xenopus* (Sindelka et al., 2010; Vastag et al., 2011; Flachsova et al., 2013; Shrestha et al., 2014; Sun et al., 2014), raising the question whether these chemical changes are heterogeneous also between individual cells of the embryo at different embryonic developmental stages. However, the challenge has been to collect high-quality signal from the miniscule amounts of molecules contained within single blastomeres for analysis. Since different blastomeres in *Xenopus* are fated to give rise to different tissues (Moody, 1987a,b; Moody and Kline, 1990), elucidating the proteome in individual cells of the embryo holds a great potential to elevate our understanding of the cellular physiology that regulates embryogenesis. For a deeper understanding of the developmental processes that govern early embryonic processes, it would be transformative to assay the ultimate indicator of gene expression downstream of transcription: the proteome.

To address this cell biology question, we and others have developed platforms to extend MS to single cells (see reviews in References Mellors et al., 2010; Rubakhin et al., 2011; Passarelli and Ewing, 2013; Li et al., 2015). For example, targeted proteins have been measured in erythrocytes (Hofstadler et al., 1995; Valaskovic et al., 1996; Mellors et al., 2010). Discovery MS has been used in the study of protein partitioning in the nucleus of the *Xenopus laevis* oocyte (Wuhr et al., 2015). Recently, we have developed single-cell analysis workflows and custom-built microanalytical capillary electrophoresis (CE) platforms for MS to enable the discovery (untargeted) characterization of gene translation in single embryonic cells (blastomeres). Using single-cell CE, we have measured hundreds–thousands of proteins in blastomeres giving rise to distinct tissues in the frog (*X. laevis*), such as neural, epidermal, and gut tissues (Moody, 1987a). We have also established quantitative approaches to compare gene translation between these cell types. Quantification of ∼150 different proteins between the blastomeres has captured translational cell heterogeneity in the 16-cell vertebrate embryo (Lombard-Banek et al., 2016a). These results complement known transcriptional cell differences in the embryo, but also provide previously unknown details on how differential gene expression establishes cell heterogeneity during early embryonic development.

In this contribution, we give an overview of the major steps of the single-cell CE-MS workflow (**Figure 1**). Protocols are provided to isolate single cells, extract and process proteins,

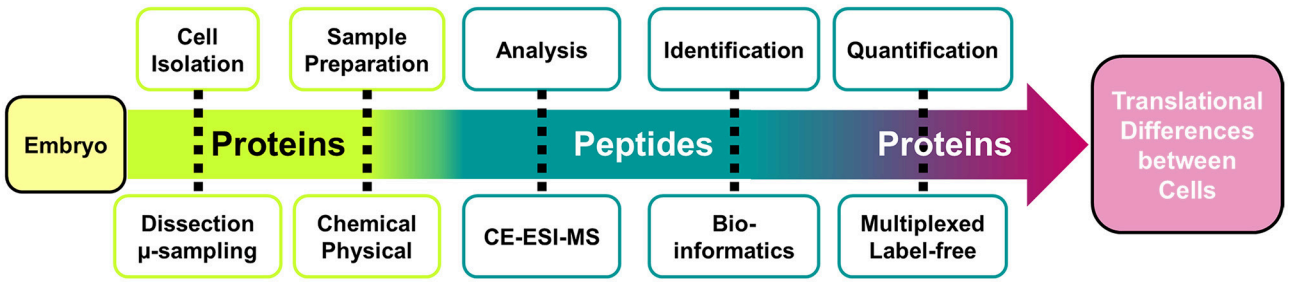


FIGURE 1 | Analytical workflow for the bottom-up measurements of protein expression in single embryonic cells. A custom-built high-sensitivity capillary electrophoresis electrospray ionization mass spectrometer (CE-ESI-MS) is used to identify and quantify proteins.

TABLE 1 | Troubleshooting advice for CE-ESI-MS for bottom-up proteomics.

Issues	Potential causes	Advice
No peptides detected	Failed enzymatic digestion	Repeat analysis; if problem persists, repeat protein digestion (use standard proteins as quality control)
CE current drops drastically	Capillary is clogged or a bubble was injected	Flush the capillary with the BGE for ~10–15 min; repeat analysis
Electrospray is unstable	Electrolysis in the CE-ESI interface; the sheath flow connection is loose	Lower the spray voltage; revise connections; repeat analysis
Low number of protein identifications	Erroneous injection; inaccurate calibration of the mass spectrometer	Repeat analysis; calibrate the mass spectrometer

and use the CE-MS platform to identify and quantify protein expression. Additional details on technology development and validation are available elsewhere (Nemes et al., 2013; Onjiko et al., 2015; Lombard-Banek et al., 2016a,b). These protocols have allowed us to study proteins (Lombard-Banek et al., 2016a,b) and metabolites (Onjiko et al., 2015, 2016) in single blastomeres in 8-, 16-, and 32-cell *X. laevis* embryos. Additionally, trouble-shooting advice (**Table 1**) is provided to help others adopt single-cell MS toward the systems biology characterization of molecular processes in cells and limited amounts of specimens.

MATERIALS AND EQUIPMENT

Single Blastomere Dissection

a. Fine sharp forceps (e.g., Dumont #5). One forceps should have a squared tip, while the other should be sharpened to a fine tip.
b. Sterile Pasteur pipets.
c. Hair loop: place a fine hair (~10 cm long) into a 6″ Pasteur pipet to form a 2–3 mm loop and secure it in place with melted paraffin. Sterilize the hair loop before usage by dipping it in 70% methanol.
d. 0.6 mL centrifuge tubes.
e. 60 and 90 mm Petri dishes.
f. Incubator set to 14°C.
g. Dejellying solution: 2% cysteine hydrochloride in water, pH 8, prepared by adding 20 g of crystalline cysteine hydrochloride into 1 L of distilled water. pH is adjusted to 8 by adding 10 N NaOH drop-wise.
h. 100% Steinberg's solution (SS): Dissolve the following salts into 1 L of distilled water: 3.5064 g NaCl, 49.9 mg KCl, 99.9 mg $MgSO_4$, 55.8 mg $Ca(NO_3)_2$, 0.6302 g Tris-HCl, and 80.0 mg Tris-base. Adjust the pH to 7.4. Autoclave and store in 14°C incubator.
i. 50% Steinberg's solution: Dilute 50 mL of 100% SS with 50 mL of distilled water.
j. Dissection dish: add 2 g of agarose in 100 mL of 100% Steinberg's solution. Dissolve the agarose by autoclaving. Once the bottle is cool enough to handle, pour the agarose mixture to ~1 mm in thickness into 60 mm in diameter Petri dishes. Alternatively, the agarose mixture can be stored at 4°C, and reheated in a microwave before use. Dishes should be stored wrapped in plastic at 4°C to prevent dehydration of the agarose.
k. *X. laevis* (adult male and female). Protocols related to the handling and manipulation of animals must adhere to Institutional and/or Federal guidelines; the work reported here was approved by the George Washington University Institutional Animal Care and Use Committee (IACUC #A311).

Protein Extraction, Enzymatic Digestion, and Quantification

a. Refrigerated centrifuge (4°C)
b. Heat blocks (2) set to 60 and 37°C.
c. A −20°C freezer.
d. Sonication bath (e.g., Brandson CPX 2800).
e. A vacuum concentrator (e.g., CentriVap, LabConco).

f. Lysis buffer: for 1 mL of lysis buffer, mix 100 µL of 10% sodium dodecyl sulfate (SDS), 100 µL of 1.5 M NaCl, 20 µL of 1 M Tris-HCl (pH 7.5), 10 µL of 0.5 M EDTA, and 770 µL of H_2O.
g. Enzymatic digestion solution, 50 mM ammonium bicarbonate: add 0.1976 g of crystalline ammonium bicarbonate to HPLC grade water.
h. Dithiothreitol (1 M): Dissolve 0.1543 g of solid dithiothreitol into 1 mL of 50 mM ammonium bicarbonate. Divide in 50–100 µL aliquots and store at −20°C for months.
i. Iodoacetamide (1 M): Dissolve 0.1850 g of crystalline iodoacetamide into 1 mL of 50 mM ammonium bicarbonate. Iodoacetamide is light sensitive and therefore should be kept away from any light sources. It is suggested to make freshly before use, but storage in 50–100 µL aliquots at −20°C is acceptable for up to 2 months. Aliquots are only for single use, do not freeze-thaw.
j. Trypsin solution 0.5 µg/µL: dissolve a 20 µg vial in 40 µL of 1 mM HCl in water.
k. Tandem mass tags kit (e.g., TMT10plex, Thermo Scientific).

CE-ESI-MS Analysis

a. HPLC grade solvents and reagents: water, acetonitrile, methanol, formic acid, and acetic acid.
b. Regulated high voltage power supplies (2) outputting up to 5 kV for maintaining the electrospray (e.g., P350, Stanford Research Systems), and up to 30 kV for CE separation (e.g., Bertan 230-30R, Spellman).
c. Separation capillary: 40/110 µm (i.d./o.d.) bare fused silica capillary from Polymicro.
d. Sample solvent: mix 500 µL methanol with 500 µL water and 0.5 µL acetic acid.
e. Sheath solution: add 50 mL of methanol to 50 mL of water and 50 µL of formic acid.
f. Background electrolyte: to prepare 50 mL, mix 12.5 mL of acetonitrile, and 1.887 mL of formic acid with 35.613 mL of water.
g. High-resolution mass spectrometer (e.g., Orbitrap Fusion, Thermo).

PROCEDURES

Sample Preparation

The goal of sample preparation is to extract proteins from single cells and process the proteins for MS analysis. The workflow (**Figure 1**) starts with the identification of blastomeres in the embryo in reference to established cell fate maps (Moody, 1987a,b; Moody and Kline, 1990; Lee et al., 2012) and differences in cell size and pigmentation. Cells are microdissected using sharp forceps and collected into individual microcentrifuge tubes. **Figure 2** shows the dissection of the V11 cell. Next, isolated blastomeres are lysed using chemical (detergent) and physical (ultrasonication) methods, and their proteins are extracted. The proteins are processed via standard bottom-up proteomics protocols (Zhang et al., 2013), whereby reduction, alkylation, and enzymatic digestion are performed to convert proteins into peptides that are more readily analyzable by MS.

Single Blastomere Dissection and Isolation

As detailed protocols are available on the identification and dissection of blastomeres (Moody, 2012; Grant et al., 2013), only a brief summary of the major steps follows.

(1) Prepare consumables:

- 2% cysteine solution
- 100% Steinberg solution (SS)
- 50% Steinberg solution (SS)
- Sterile Pasteur pipet
- Petri dish filled with 2% agarose (w/v in 100% SS)
- Sharp forceps
- Hair loop
- 0.6 mL microcentrifuge tubes

(2) Remove jelly coats that naturally surround the embryos:

a. Add 4× volume of the cysteine solution to the embryos (**Table 2**) and gently swirl the solution for ∼4 min.
b. Once the embryos are free of the jelly coat, immediately wash them with 100% SS (**Table 2**) 4 times for 2 min each.

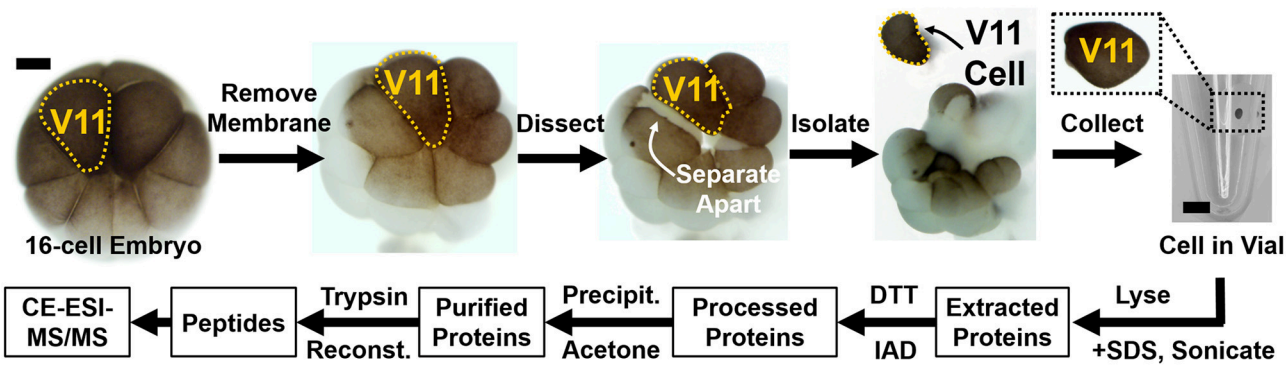


FIGURE 2 | Isolation of identified cells and processing of their protein content. Example shows how the epidermal-fated ventral-animal cell (named V11) was identified in the 16-cell *X. laevis* embryo based on pigmentation, cell size, and location in reference to established cell fate maps (Moody, 1987a). The cell was processed via bottom-up proteomic workflow, and the resulting peptides collected for proteomic analysis. Key: DTT, dithiothreitol; IAD, iodoacetamide. Scale bar = 200 µm (embryo), 1.25 mm (vial).

TABLE 2 | Solutions and their uses.

Solution/buffer	Composition	Usage	Storage conditions
Cysteine Hydrochloride	2% (w/v) cysteine hydrochloride, pH 8 adjusted with 10 N NaOH drop wise	Removes the jelly coats surrounding embryos	Make fresh
Steinberg's Solution (SS)	60 mM NaCl, 0.67 mM KCl, 0.83 mM $MgSO_4$, 0.34 mM $Ca(NO_3)_2$, 4 mM Tris–HCl, 0.66 mM Tris base, in distilled water, pH 7.4. Autoclaved. Store in incubator for months.	Provides media for culturing embryos	4–14°C
Lysis Buffer	1% sodium dodecyl sulfate (SDS), 150 mM NaCl, 20 mM Tris-HCl pH 8, 5 mM EDTA in distilled water	Lyses cells/tissues	4°C
Sample Solvent	50–60% acetonitrile in water, 0.05% acetic acid (all solvents are LC-MS grade)	Reconstitutes protein digest	4°C
Background Electrolyte (BGE)	25% acetonitrile in water, 1 M formic acid (all solvents are LC-MS grade)	Electrolyte for CE	4°C
Electrsopray Sheath Liquid	50% methanol in water, 0.1% formic acid (all solvents are LC-MS grade)	Stabilizes ESI-MS operation	4°C

c. Transfer the embryos to a clean Petri dish filled with 100% SS and store them at 14–20°C in an incubator.

(3) Dissect cells from the embryos as published elsewhere (Grant et al., 2013). A representative example is shown in **Figure 2**. Briefly:

a. Transfer the selected embryos to a 60 mm Petri dish coated with 2% agarose and filled with 50% SS.
b. Place the embryo of interest in a groove made in the agarose coating.
c. Orient the embryo for easy handling of the cell of interest using a hair loop.
d. Remove the vitelline membrane gently using sharp forceps. During this step, take care not to damage the embryo.
e. Hold the embryo using sharp forceps on the opposite side of the cell of interest, and gently pull on either side to isolate the cell.
f. Transfer isolated cells using a sterile Pasteur pipet into a micro-centrifuge tube.

Protein Extraction and Enzymatic Digestion

(1) Prepare consumables:

- Lysis buffer
- Acetone chilled to −20°C
- 50 mM ammonium bicarbonate
- 1 M dithiothreitol
- 1 M iodoacetamide
- Sonication bath (e.g., Brandson CPX 2800)

(2) Lyse the cells to release their content:

a. Remove the excess 50% SS from around the cell. Take care not to disrupt the cell.
b. Add 10 μL of lysis buffer (**Table 2**) and vortex for ∼30 s.
c. Sonicate for ∼5 min, vortex for ∼30 s. Repeat this step 3 times.
d. (Optionally) Add protease inhibitor to the lysis buffer to minimize/avoid protein degradation during this step.

(3) Reduce and alkylate protein disulfide bonds:

a. Add 0.5 μL of 1 M dithiothreitol to the sample, and incubate for 20–30 min at 60°C.
b. Add 1 μL of 1 M iodoacetamide and incubate for 15 min in the dark at room temperature.
c. Quench the reaction by adding 0.5 μL of 1 M dithiothreitol.

(4) Purify proteins by cold acetone precipitation.

a. Add to the cell extract a volume of pure acetone that is 5 times that of the cell extract (∼50 μL), and incubate at −20°C overnight.
b. Recover the precipitated proteins by centrifugation at 10,000 × g for 10 min and 4°C.
c. Remove the supernatant.
d. Dry the pellet using a vacuum concentrator.
e. (Optional) Store the protein pellet at –20 or −80°C for up to 3 months.

(5) Digest proteins for bottom-up proteomics analysis. A variety of enzymes or a combination of enzymes can be used for this task (e.g., trypsin, lysine C). We choose trypsin due to its benefits for MS analysis (Zhang et al., 2013).

a. Reconstitute the protein pellet in 50 mM ammonium bicarbonate.
b. Add 0.3 μL of 0.5 μg/μL trypsin (trypsin in 1 mM HCl), equivalent to a protease/protein ratio of ∼1/50.
c. Incubate overnight at 37°C.

(6) (Optional) Store the digest at −80°C for up to 3 months.

Quantification

The presented technology is compatible with well-established protocols in quantitative proteomics. Stable isotope labeling with amino acids in cell culture (SILAC) allows barcoding of proteins with isotopic labels for multiplexing quantification (Geiger et al., 2013). Label-free quantification (LFQ) is an alternative strategy whereby peptide signal abundance is used as a proxy for protein concentration. We have recently demonstrated LFQ for single blastomeres of neural fates in the 16-cell embryo using

the protocol presented here (Lombard-Banek et al., 2016b). Alternatively, relative quantification can be performed using designer mass tags. In this approach, proteins are digested to peptides and the peptides barcoded with isotopic labels that can be distinguished by high-resolution MS. Multiple protocols allow for quantifying protein expression at the level of peptides in high throughput via multiplexing, including tandem mass tags (TMT) (Thompson et al., 2006; McAlister et al., 2014), and isobaric tag for relative and absolute quantitation (iTRAQ; Ross et al., 2004), and di-Leu (Xiang et al., 2010; Frost and Li, 2016). We have recently downscaled TMT-based multiplexed quantification to the protein content of single blastomeres using the following strategy (adapted from the vendor), which we then used to compare protein expression between the D11, V11, and V21 cells (Lombard-Banek et al., 2016a) that are fated to give rise to different types of tissues (neural, epidermal, and hindgut, respectively):

a. Add 15 μL of TMT reagent to each digest and incubate for 1 h at room temperature.
b. Add 3.5 μL of hydroxylamine and incubate for 15 min at room temperature.
c. Mix the samples together at a 1:1 ratio (volume or total protein content)
d. Dry the sample using a vacuum concentrator.
e. Add 5 μL of 60% acetonitrile containing 0.05% formic acid.

Sample Analysis Using CE-ESI-MS

Peptides are analyzed using a custom-built CE-ESI-MS platform (Nemes et al., 2013; Onjiko et al., 2015; Lombard-Banek et al., 2016a). Instructions regarding the construction and operation of the platform are available from elsewhere (Nemes et al., 2013). Schematics of the CE-ESI-MS instrument are shown in **Figure 3**. CE is selected to electrophoretically separate peptides in a fused silica capillary by applying voltage difference across the capillary ends. As a general rule, peptides with smaller size and higher charge state migrate faster through the capillary. A high resolution mass spectrometer is used to sequence peptides via data-dependent acquisition. In this approach, eluting peptides are detected based on single-stage (full) scans (MS^1) and are sequenced by tandem-MS (MS^2 scans) using collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), or other fragmentation technologies. The tandem mass spectra reveal sequence information for the peptides, as also exemplified for LGLGLELEA in **Figure 4**. During quantification experiments, the TMT labels also dissociate from the peptide, and the relative abundance of these TMT signals serves as quantitative measure of protein abundance (**Figure 4C,** right panel).

CE-ESI-MS Measurements

(1) Build the CE-ESI-MS system as described elsewhere (Nemes et al., 2013; Onjiko et al., 2015). For bottom-up proteomics of single *Xenopus* blastomeres, operate the system as recently established (Lombard-Banek et al., 2016a,b).
(2) Prepare the CE system ∼15 min prior to start the experiments as follow:
    a. Flush the capillary with background electrolyte (25% acetonitrile with 1 M formic acid).
    b. Flush the sheath capillary with electrospray solution (50% methanol with 0.1% formic acid)
    c. Turn on the electronics (high voltage power supplies, syringe pumps, mass spectrometer, etc.) for ∼30 min to stabilize operation.
(3) Inject the sample into the capillary as follows:
    a. Transfer the capillary into the background electrolyte vial.
    b. Deposit ∼1 μL of sample onto the sample microvial (see **Figure 3**).
    c. Transfer the capillary from the BGE vial to the sample vial.
    d. Elevate the injection stage by ∼15 cm for ∼3 min to siphon ∼20 nL of the sample into the CE capillary.
    e. Lower the injection stage to level the capillary inlet to the outlet, and transfer the capillary inlet end into the BGE vial.
    f. Apply ∼10,000 V to the background electrolyte vial to start electrophoretic separation of the peptides.

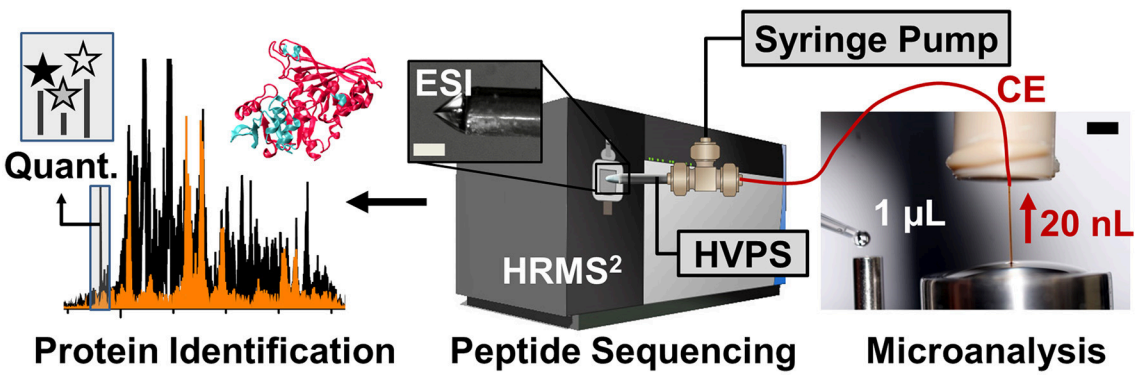


FIGURE 3 | Schematics of the high-sensitivity proteomic analyzer. The platform integrates microanalytical capillary electrophoresis (CE), electrospray ionization (ESI), and high-resolution tandem mass spectrometry ($HRMS^2$). Scale bar = 150 μm (ESI), 1.5 mm (CE panel). Key: HVPS, high-voltage power supply. Figure adapted with permission from Lombard-Banek et al. (2016a).

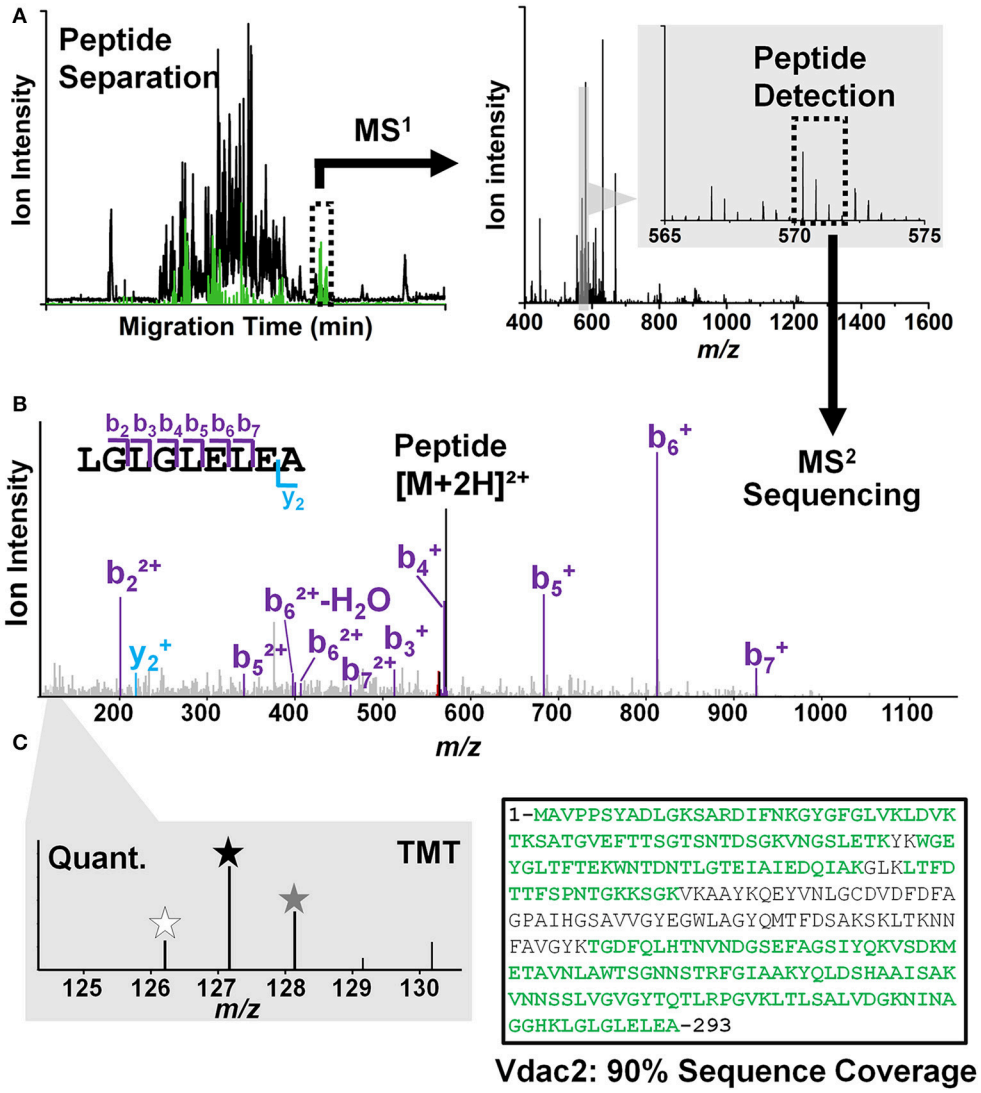


FIGURE 4 | Peptide identification/quantification in CE-ESI-HRMS^2 using a bottom-up strategy. (A) Peptides are electrophoretically separated (left panel) and their accurate mass is measured (right panel). **(B)** Peptide signals are sequenced by tandem MS (MS^2). For example, a signal was detected with *m/z* 572.33 at ~50 min separation, which was assigned to the sequence LGLGLELEA based on the MS^2 data. **(C)** Peptides are quantified and assigned to the source protein. Tandem mass tags (TMT) with different *m/z* values (indicated by asterisks of different color in left panel) are used to barcode peptides from different cells, allowing their simultaneous analysis (multiplexing) with higher throughput (left panel). For example, the sequence LGLGLELEA was unique to the voltage-dependent anion channel 2 protein in the *Xenopus* proteome. The presence of other peptides allowed identifying this protein in high sequence coverage; see detected sequence in green (right panel).

g. Increase the electrospray voltage gradually until the cone jet mode is established for efficient ionization (Nemes et al., 2007). Using a long-distance microscope, carefully inspect the electrospray emitter to avoid electrical breakdown; electrical discharge, spark, or arc risks the mass spectrometer. In our experiments, the electrospray emitter is positioned ~0.5 cm from the mass spectrometer orifice and is biased to 3000 V to generate the cone-jet spray.
h. Ramp the separation voltage to ~18,000 V. In our system, we limit the separation voltage to keep the CE current <8 μA to prevent/minimize electrolysis or solvent heating. Monitor the CE current and adjust the separation voltage as necessary. For instructions on how to measure the current, refer to Nemes et al. (2013).
i. Start MS acquisition with data-dependent acquisition as specified by the mass spectrometer vendor. For example, we use the following settings for a quadrupole-orbitrap linear ion trap mass spectrometer (Fusion, Thermo Scientific): MS^1 analyzer resolution (orbitrap), 60,000 FWHM; *m/z* scan range, 350–1600; injection time, 100 ms; precursor ion selection window, 0.8 Da

in the quadrupole cell; fragmentation, HCD with 30% normalized energy in the multipole cell using nitrogen collision gas; MS^2 analyzer rate, rapid scan; MS^2 maximum injection time, 50 ms.

Protein Identification

Last, peptide sequences are compared to the proteome of the specimen (*X. laevis* here) to identify proteins (see **Figure 4**). This step is facilitated by readily available proteomes from SwissProt, UniProt, and experimentally determined RNA expression (Wang et al., 2012; Smits et al., 2014; Wuhr et al., 2014). Well-established bioinformatics software packages are used to process raw mass spectrometric data. For example, Proteome Discoverer (Thermo Scientific), ProteinScape (Bruker Daltonics), and MaxQuant (Cox and Mann, 2008) interpret MS–tandem-MS datasets by executing well-established search engines, such as SEQUEST (Eng et al., 1994), Mascot (Perkins et al., 1999), and Andromeda (Cox et al., 2011). The general strategy of bottom-up proteomics has recently been reviewed in detail (Sadygov et al., 2004; Cox et al., 2011; Zhang et al., 2013). We typically acquire tens of thousands to a million mass spectra, which identify 2000–4000 peptides in single blastomeres in the 16-cell embryo. These data allow us to identify ~1700 protein groups and quantify hundreds of proteins between the D11, V11, and V21 cells.

Anticipated Results

The CE-ESI-MS can be used to identify gene translational differences between cells. As shown in **Figure 5**, we have used this approach to assess protein differences between blastomeres of the 16-cell *X. laevis* embryo (Lombard-Banek et al., 2016a,b). Cell types with different tissue developmental fates were analyzed: the midline dorsal-animal cell (named D11) develops mainly into the retina and brain, the midline ventral-animal cell (named V11) gives rise primarily to the head and trunk epidermis, and the midline ventral-vegetal cell (named V21) is the primary precursor of the hindgut. The approach allowed the identification of 1709 protein groups (<1% false discovery rate, FDR) from ~20 ng of protein digest, corresponding to ~0.2% of the total protein content of the blastomere (Lombard-Banek et al., 2016a). Many of the identified proteins are known to be involved in different cell fates. For example, Geminin (Gem) and Isthmin (Ism) were detected in the D11 cells in our measurements,

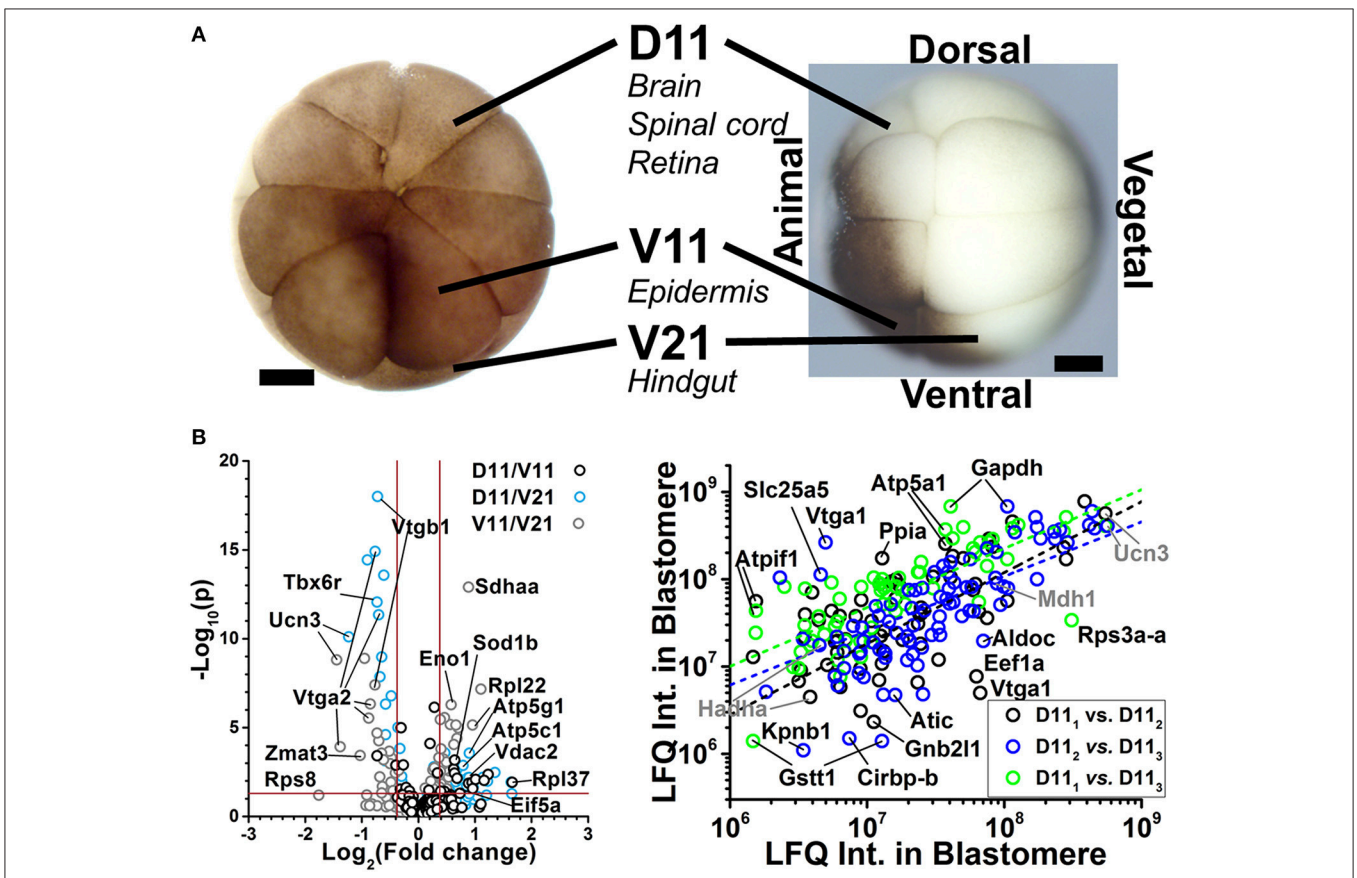


FIGURE 5 | Examples of protein identification–quantification between single embryonic cells. (A) The D11, V11, and V21 cells have different tissues fates in the frog *X. laevis*. Scale bars: 250 μm. Figure reprinted with permission from Onjiko et al. (2015). **(B)** These cells were dissected from different 16-cell *X. laevis* embryos and analyzed using multiplexed (left panel) and label-free quantification (right panel). Volcano plots reveal gene translation differences between the V11, D11, and V21 cell types (left). Pearson correlation analysis of protein expression finds similar protein expression for the majority of proteins between D11 blastomeres, and detectable differences for others (right panel). Figures adapted with permission from Lombard-Banek et al. (2016a,b).

and these proteins are involved in brain development (Pera et al., 2002; Seo et al., 2005), which is the stereotypical fate of D11 cells (Moody, 1987a). Multiplexed quantification by TMTs provided comparative evaluation for 152 non-redundant protein groups between the cell types (**Figure 5B,** left), including many that were significantly differentially expressed between the cell types ($p < 0.05$, fold change $\geq$1.3). We have also performed label free quantitation (LFQ) to compare D11 cells that were isolated at similar developmental phase of the 16-cell *X. laevis* embryos (**Figure 5A**). A Pearson correlation analysis showed similar expression levels for the majority of proteins between the D11 cells (see proteins along linear fits). The study also found 25 proteins that were differentially accumulated in the respective cells, suggesting highly variable expression (**Figure 5B,** right; Lombard-Banek et al., 2016b). These data on translational cell heterogeneity complement transcriptomic information on cell differences (Flachsova et al., 2013), but also provide new insights into how differential gene expression sets up different cell fates and the major developmental axes of the early embryo.

CONCLUSIONS

High-sensitivity MS enables the identification and quantification of a sufficiently large number of proteins to study cell and developmental processes at the level of individual cells. Advances in sampling (smaller single cells), protein processing, microanalytical MS, and bioinformatics have enabled the discovery characterization of hundreds to thousands of proteins in single cells. Unbiased measurement of protein translation by MS complements genomic and transcriptomic information, essentially laying down the foundation of the molecular characterization of cell heterogeneity. Knowledge of genomic, transcriptomic, proteomic, and metabolomic processes paves the way to understanding how differential gene expression establishes cell heterogeneity during normal development and disease states.

AUTHOR CONTRIBUTIONS

CL, SM, and PN wrote the manuscript.

FUNDING

This research was supported by National Science Foundation Grant DBI-1455474 (to PN and SM) and the George Washington University Start-Up Funds (to PN) and Columbian College Facilitating Funds (to PN and SM). The content of the presented work was solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

Abrams, E. W., and Mullins, M. C. (2009). Early zebrafish development: it's in the maternal genes. *Curr. Opin. Genet. Dev.* 19, 396–403. doi: 10.1016/j.gde.2009.06.002

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207. doi: 10.1038/nature01511

Altschuler, S. J., and Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell* 141, 559–563. doi: 10.1016/j.cell.2010.04.033

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E. A. D., Krutzik, P. O., Finck, R., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332, 687–696. doi: 10.1126/science.1198704

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372. doi: 10.1038/nbt.1511

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10, 1794–1805. doi: 10.1021/pr101065j

Cuykendall, T. N., and Houston, D. W. (2010). Identification of germ plasm-associated transcripts by microarray analysis of *Xenopus* vegetal cortex RNA. *Dev. Dyn.* 239, 1838–1848. doi: 10.1002/dvdy.22304

Davidson, E. H. (1990). How embryos work: a comparative view of diverse modes of cell fate specification. *Development* 108, 365–389.

De Domenico, E., Owens, N. D. L., Grant, I. M., Gomes-Faria, R., and Gilchrist, M. J. (2015). Molecular asymmetry in the 8-cell stage *Xenopus tropicalis* embryo described by single blastomere transcript sequencing. *Dev. Biol.* 408, 252–268. doi: 10.1016/j.ydbio.2015.06.010

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate tandem mass-spectral data of peptides with amino-acids sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989. doi: 10.1016/1044-0305(94)80016-2

Flachsova, M., Sindelka, R., and Kubista, M. (2013). Single blastomere expression profiling of *Xenopus laevis* embryos of 8 to 32-cells reveals developmental asymmetry. *Sci. Rep.* 3:2278. doi: 10.1038/srep02278

Frost, D. C., and Li, L. J. (2016). "High-throughput quantitative proteomics enabled by mass defect-based 12-plex diLeu isobaric tags," in *Quantitative Proteomics by Mass Spectrometry, 2nd Edn.,* ed S. Sechi (Totowa, NJ: Humana Press Inc.), 169–194.

Gajecka, M. (2016). Unrevealed mosaicism in the next-generation sequencing era. *Mol. Genet. Genomics* 291, 513–530. doi: 10.1007/s00438-015-1130-7

Gallagher, B. C., Hainski, A. M., and Moody, S. A. (1991). Autonomous differentiation of dorsal axial structures from an animal cap cleavage stage blastomere in *Xenopus*. *Development* 112, 1103–1114.

Gaur, S., Mandelbaum, M., Herold, M., Majumdar, H. D., Neilson, K. M., Maynard, T. M., et al. (2016). Neural transcription factors bias cleavage stage blastomeres to give rise to neural ectoderm. *Genesis* 54, 334–349. doi: 10.1002/dvg.22943

Geiger, T., Velic, A., Macek, B., Lundberg, E., Kampf, C., Nagaraj, N., et al. (2013). Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* 12, 1709–1722. doi: 10.1074/mcp.M112.024919

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* 11, 417–422. doi: 10.1038/nmeth.2869

Grant, P. A., Herold, M. B., and Moody, S. A. (2013). Blastomere explants to test for cell fate commitment during embryonic development. *J. Vis. Exp.* 71:e4458. doi: 10.3791/4458

Grant, P. A., Yan, B., Johnson, M. A., Johnson, D. L. E., and Moody, S. A. (2014). Novel animal pole-enriched maternal mRNAs are preferentially expressed in neural ectoderm. *Dev. Dyn.* 243, 478–496. doi: 10.1002/dvdy.24082

Guerrera, I. C., and Kleiner, O. (2005). Application of mass spectrometry in proteomics. *Biosci. Rep.* 25, 71–93. doi: 10.1007/s10540-005-2849-x

Hainski, A. M., and Moody, S. A. (1992). *Xenopus* maternal Rnas from a dorsal animal blastomere induce a secondary axis in host embryos. *Development* 116, 347–355.

Haston, K. M., and Reijo-Pera, R. A. (2007). "Germ line determinants and oogenesis," in *Principles of Developmental Genetics,* ed S. A. Moody (New York, NY: Academic Press), 150–172.

Heasman, J. (2006a). Maternal determinants of embryonic cell fate. *Semin. Cell. Dev. Biol.* 17, 93–98. doi: 10.1016/j.semcdb.2005.11.005

Heasman, J. (2006b). Patterning the early *Xenopus* embryo. *Development* 133, 1205–1217. doi: 10.1242/dev.02304

Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., et al. (2014). The one hour yeast proteome. *Mol. Cell. Proteomics* 13, 339–347. doi: 10.1074/mcp.M113.034769

Hofstadler, S. A., Swanek, F. D., Gale, D. C., Ewing, A. G., and Smith, R. D. (1995). Capillary electrophoresis electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for direct analysis of cellular proteins. *Anal. Chem.* 67, 1477–1480. doi: 10.1021/ac00104a028

Jacobs, K., Mertzanidou, A., Geens, M., Nguyen, H. T., Staessen, C., and Spits, C. (2014). Low-grade chromosomal mosaicism in human somatic and embryonic stem cell populations. *Nat. Commun.* 5, 4227–4237. doi: 10.1038/ncomms5227

Junker, J. P., and van Oudenaarden, A. (2015). Single-cell transcriptomics enters the age of mass production. *Mol. Cell* 58, 563–564. doi: 10.1016/j.molcel.2015.05.019

Kanter, I., and Kalisky, T. (2015). Single cell transcriptics: methods and applications. *Front. Oncol.* 5:53. doi: 10.3389/fonc.2015.00053

Kenyon, K. L. (2007). "Patterning the anterior-posterior axis during *Drosophila* embryogenesis," in *Principles of Developmental Genetics,* ed S. A. Moody (New York, NY: Academic Press), 173–200.

King, M. L., Messitt, T. J., and Mowry, K. L. (2005). Putting RNAs in the right place at the right time: RNA localization in the frog oocyte. *Biol. Cell* 97, 19–33. doi: 10.1042/BC20040067

Kolisko, M., Boscaro, V., Burki, F., Lynn, D. H., and Keeling, P. J. (2014). Single-cell transcriptomics for microbial eukaryotes. *Curr. Biol.* 24, R1081–R1082. doi: 10.1016/j.cub.2014.10.026

Lee, H. S., Sokol, S. Y., Moody, S. A., and Daar, I. O. (2012). *Using 32-Cell Stage Xenopus Embryos to Probe PCP Signaling.* New York, NY: Springer.

Liang, Q., Conte, N., Skarnes, W. C., and Bradley, A. (2008). Extensive genomic copy number variation in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 105, 17453–17456. doi: 10.1073/pnas.0805638105

Lindeman, R. E., and Pelegri, F. (2010). Vertebrate maternal-effect genes: insights into fertilization, early cleavage divisions, and germ cell determinant localization from studies in the zebrafish. *Mol. Reprod. Dev.* 77, 299–313. doi: 10.1002/mrd.21128

Li, S. Y., Plouffe, B. D., Belov, A. M., Ray, S., Wang, X. Z., Murthy, S. K., et al. (2015). An integrated platform for isolation, processing, and mass spectrometry-based proteomic profiling of rare cells in whole blood. *Mol. Cell. Proteomics* 14, 1672–1683. doi: 10.1074/mcp.M114.045724

Lombard-Banek, C., Moody, S. A., and Nemes, P. (2016a). Single-cell mass spectrometry for discovery proteomics: quantifying translational cell heterogeneity in the 16-cell frog (*Xenopus*) embryo. *Angew. Chem. Int. Ed. Engl.* 55, 2454–2458. doi: 10.1002/anie.201510411

Lombard-Banek, C., Reddy, S., Moody, S. A., and Nemes, P. (2016b). Label-free quantification of proteins in single embryonic cells with neural fate in the cleavage-stage frog *(Xenopus laevis)* embryo using capillary electrophoresis electrospray ionization high-resolution mass spectrometry CE-ESI-HRMS. *Mol. Cell. Proteomics* 15, 2756–2768. doi: 10.1074/mcp.M115.057760

McAlister, G. C., Nusinow, D. P., Jedrychowski, M. P., Wühr, M., Huttlin, E. L., Erickson, B. K., et al. (2014). MultiNotch MS^3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* 86, 7150–7158. doi: 10.1021/ac502040v

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., et al. (2013). Mosaic copy number variation in human neurons. *Science* 342, 632–637. doi: 10.1126/science.1243472

Mellors, J. S., Jorabchi, K., Smith, L. M., and Ramsey, J. M. (2010). Integrated microfluidic device for automated single cell analysis using electrophoretic separation and electrospray ionization mass spectrometry. *Anal. Chem.* 82, 967–973. doi: 10.1021/ac902218y

Mitra, A. K., Stessman, H., Linden, M. A., and Van Ness, B. (2014). Single-cell transcriptomics identifies intra-tumor heterogeneity in Human myeloma cell lines. *Blood* 124, 3385.

Moody, S. A. (1987a). Fates of the blastomeres of the 16-cell stage *Xenopus* embryo. *Dev. Biol.* 119, 560–578.

Moody, S. A. (1987b). Fates of the blastomeres of the 32-cell-stage *Xenopus* embryo. *Dev. Biol.* 122, 300–319.

Moody, S. A. (2012). *Testing Retina Fate Commitment in Xenopus by Blastomere Deletion, Transplantation, and Explant Culture.* New York, NY: *Springer.*

Moody, S. A., and Kline, M. J. (1990). Segregation of fate during cleavage of frog (*Xenopus laevis*) blastomeres. *Anat. Embryol.* 182, 347–362. doi: 10.1007/BF02433495

Nemes, P., Marginean, I., and Vertes, A. (2007). Spraying mode effect on droplet formation and ion chemistry in electrosprays. *Anal. Chem.* 79, 3105–3116. doi: 10.1021/ac062382i

Nemes, P., Rubakhin, S. S., Aerts, J. T., and Sweedler, J. V. (2013). Qualitative and quantitative metabolomic investigation of single neurons by capillary electrophoresis electrospray ionization mass spectrometry. *Nat. Protoc.* 8, 783–799. doi: 10.1038/nprot.2013.035

Onjiko, R. M., Moody, S. A., and Nemes, P. (2015). Single-cell mass spectrometry reveals small molecules that affect cell fates in the 16-cell embryo. *Proc. Natl. Acad. Sci. U.S.A.* 112, 6545–6550. doi: 10.1073/pnas.1423682112

Onjiko, R. M., Morris, S. E., Moody, S. A., and Nemes, P. (2016). Single-cell mass spectrometry with multi-solvent extraction identifies metabolic differences between left and right blastomeres in the 8-cell frog (*Xenopus*) embryo. *Analyst* 141, 3648–3656. doi: 10.1039/c6an00200e

Passarelli, M. K., and Ewing, A. G. (2013). Single-cell imaging mass spectrometry. *Curr. Opin. Chem. Biol.* 17, 854–859. doi: 10.1016/j.cbpa.2013.07.017

Pera, E. M., Kim, J. I., Martinez, S. L., Brechner, M., Li, S. Y., Wessely, O., et al. (2002). Isthmin is a novel secreted protein expressed as part of the Fgf-8 synexpression group in the *Xenopus* midbrain-hindbrain organizer. *Mech. Dev.* 116, 169–172. doi: 10.1016/S0925-4773(02)00123-5

Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R. M. Jr., Gerhart, J. C., et al. (2015). On the relationship of protein and mRNA dynamics in vertebrate embryonic development. *Dev. Cell* 35, 383–394. doi: 10.1016/j.devcel.2015.10.010

Radford, H. E., Meijer, H. A., and de Moor, C. H. (2008). Translational control by cytoplasmic polyadenylation in *Xenopus* oocytes. *BBA-Gene Regul. Mech.* 1779, 217–229. doi: 10.1016/j.bbagrm.2008.02.002

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135, 216–226. doi: 10.1016/j.cell.2008.09.050

Ratnaparkhi, G. S., and Courey, A. J. (2007). "Signaling cascades, gradients, and gene networks in dorsal/ventral patterning," in *Principles of Developmental Genetics,* ed S. A. Moody (New York, NY: Academic Press), 216–240.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., et al. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 3, 1154–1169. doi: 10.1074/mcp.M400129-MCP200

Rubakhin, S. S., Romanova, E. V., Nemes, P., and Sweedler, J. V. (2011). Profiling metabolites and peptides in single cells. *Nat. Methods* 8, S20–S29. doi: 10.1038/nmeth.1549

Sadygov, R. G., Cociorva, D., and Yates, J. R. (2004). Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* 1, 195–202. doi: 10.1038/nmeth725

Seo, S., Herr, A., Lim, J. W., Richardson, G. A., Richardson, H., and Kroll, K. L. (2005). Geminin regulates neuronal differentiation by antagonizing Brg1 activity. *Genes Dev.* 19, 1723–1734. doi: 10.1101/gad.1319105

Shrestha, B., Sripadi, P., Reschke, B. R., Henderson, H. D., Powell, M. J., Moody, S. A., et al. (2014). Subcellular metabolite and lipid analysis of *Xenopus laevis* eggs by LAESI mass spectrometry. *PLoS ONE* 9:e115173. doi: 10.1371/journal.pone.0115173

Sindelka, R., Sidova, M., Svec, D., and Kubista, M. (2010). Spatial expression profiles in the *Xenopus laevis* oocytes measured with qPCR tomography. *Methods* 51, 87–91. doi: 10.1016/j.ymeth.2009.12.011

Singh, D. K., Ku, C. J., Wichaidit, C., Steininger, R. J., Wu, L. F., and Altschuler, S. J. (2010). Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.* 6, 369. doi: 10.1038/msb.2010.22

Smits, A. H., Lindeboom, R. G. H., Perino, M., van Heeringen, S. J., Veenstra, G. J. C., and Vermeulen, M. (2014). Global absolute quantification reveals tight regulation of protein expression in single *Xenopus* eggs. *Nucleic Acids Res.* 42, 9880–9891. doi: 10.1093/nar/gku661

Sullivan, S. A., Akers, L., and Moody, S. A. (2001). foxD5a, a *Xenopus* winged helix gene, maintains an immature neural ectoderm via transcriptional repression that is dependent on the C-terminal domain. *Dev. Biol.* 232, 439–457. doi: 10.1006/dbio.2001.0191

Sun, L. L., Bertke, M. M., Champion, M. M., Zhu, G. J., Huber, P. W., and Dovichi, N. J. (2014). Quantitative proteomics of *Xenopus laevis* embryos: expression kinetics of nearly 4000 proteins during early development. *Sci. Rep.* 4:4365. doi: 10.1038/srep04365

Thompson, A., Schaefer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., et al. (2006). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 78, 4235. doi: 10.1021/ac060310l

Valaskovic, G. A., Kelleher, N. L., and McLafferty, F. W. (1996). Attomole protein characterization by capillary electrophoresis mass spectrometry. *Science* 273, 1199–1202. doi: 10.1126/science.273.5279.1199

Vastag, L., Jorgensen, P., Peshkin, L., Wei, R., Rabinowitz, J. D., and Kirschner, M. W. (2011). Remodeling of the metabolome during early frog development. *PLoS ONE* 6:e16881. doi: 10.1371/journal.pone.0016881

Vijg, J. (2014). Somatic mutations, genome mosaicism, cancer and aging. *Curr. Opin. Genet. Dev.* 26, 141–149. doi: 10.1016/j.gde.2014.04.002

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi: 10.1038/nrg3185

Walther, T. C., and Mann, M. (2010). Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* 190, 491–500. doi: 10.1083/jcb.201004052

Wang, X. J., Slebos, R. J. C., Wang, D., Halvey, P. J., Tabb, D. L., Liebler, D. C., et al. (2012). Protein identification using customized protein sequence databases derived from RNA-seq data. *J. Proteome Res.* 11, 1009–1017. doi: 10.1021/pr200766z

White, J. A., and Heasman, J. (2008). Maternal control of pattern formation in *Xenopus laevis*. *J. Exp. Zool. Part B* 310B, 73–84. doi: 10.1002/jez.b.21153

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587. doi: 10.1038/nature13319

Wühr, M., Freeman, R. M. Jr., Presler, M., Horb, M. E., Peshkin, L., Gygi, S. P., et al. (2014). Deep proteomics of the *Xenopus laevis* egg using an mRNA-derived reference database. *Curr. Biol.* 24, 1467–1475. doi: 10.1016/j.cub.2014.05.044

Wühr, M., Guttler, T., Peshkin, L., McAlister, G. C., Sonnett, M., Ishihara, K., et al. (2015). The nuclear proteome of a vertebrate. *Curr. Biol.* 25, 2663–2671. doi: 10.1016/j.cub.2015.08.047

Xanthos, J. B., Kofron, M., Wylie, C., and Heasman, J. (2001). Maternal VegT is the initiator of a molecular network specifying endoderm in *Xenopus laevis*. *Development* 128, 167–180.

Xiang, F., Ye, H., Chen, R. B., Fu, Q., and Li, L. J. (2010). N,N-dimethyl leucines as novel isobaric tandem mass tags for quantitative proteomics and peptidomics. *Anal. Chem.* 82, 2817–2825. doi: 10.1021/ac902778d

Yan, B., and Moody, S. A. (2007). The competence of *Xenopus* blastomeres to produce neural and retinal progeny is repressed by two endo-mesoderm promoting pathways. *Dev. Biol.* 305, 103–119. doi: 10.1016/j.ydbio.2007.01.040

Zenobi, R. (2013). Single-cell metabolomics: analytical and biological perspectives. *Science* 342, 1201. doi: 10.1126/science.1243259

Zhang, Y. Y., Fonslow, B. R., Shan, B., Baek, M. C., and Yates, J. R. III, (2013). Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* 113, 2343–2394. doi: 10.1021/cr3003533

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 January 2017
doi: 10.3389/fgene.2017.00001

Cell Cycle and Cell Size Dependent Gene Expression Reveals Distinct Subpopulations at Single-Cell Level

Soheila Dolatabadi[1], Julián Candia[2,3], Nina Akrap[1], Christoffer Vannas[1], Tajana Tesan Tomic[1], Wolfgang Losert[3], Göran Landberg[1], Pierre Åman[1] and Anders Ståhlberg[1]**

[1] Department of Pathology and Genetics, Sahlgrenska Cancer Center, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden, [2] Center for Human Immunology, Autoimmunity and Inflammation, National Institutes of Health, Bethesda, MD, USA, [3] Department of Physics, University of Maryland, College Park, MD, USA

Cell proliferation includes a series of events that is tightly regulated by several checkpoints and layers of control mechanisms. Most studies have been performed on large cell populations, but detailed understanding of cell dynamics and heterogeneity requires single-cell analysis. Here, we used quantitative real-time PCR, profiling the expression of 93 genes in single-cells from three different cell lines. Individual unsynchronized cells from three different cell lines were collected in different cell cycle phases (G0/G1 – S – G2/M) with variable cell sizes. We found that the total transcript level per cell and the expression of most individual genes correlated with progression through the cell cycle, but not with cell size. By applying the random forests algorithm, a supervised machine learning approach, we show how a multi-gene signature that classifies individual cells into their correct cell cycle phase and cell size can be generated. To identify the most predictive genes we used a variable selection strategy. Detailed analysis of cell cycle predictive genes allowed us to define subpopulations with distinct gene expression profiles and to calculate a cell cycle index that illustrates the transition of cells between cell cycle phases. In conclusion, we provide useful experimental approaches and bioinformatics to identify informative and predictive genes at the single-cell level, which opens up new means to describe and understand cell proliferation and subpopulation dynamics.

Keywords: cell cycle, cell size, single-cell gene expression, machine learning, variable selection, random forests, cell subpopulations, cell transitions

OPEN ACCESS

Edited by:
Xinghua Pan,
Yale University, USA

Reviewed by:
David Loose,
University of Texas Medical School, USA
Haiying Zhu,
Second Military Medical University, China

****Correspondence:***
Julián Candia
julian.candia@nih.gov
Anders Ståhlberg
anders.stahlberg@gu.se

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *07 June 2016*
Accepted: *06 January 2017*
Published: *25 January 2017*

Citation:
Dolatabadi S, Candia J, Akrap N, Vannas C, Tesan Tomic T, Losert W, Landberg G, Åman P and Ståhlberg A (2017) Cell Cycle and Cell Size Dependent Gene Expression Reveals Distinct Subpopulations at Single-Cell Level. Front. Genet. 8:1. doi: 10.3389/fgene.2017.00001

INTRODUCTION

Cell proliferation is a tightly organized process that involves cell division and cell growth, where cell division can be divided into distinct cell cycle phases: G0, G1, S, G2, and M. Transitions through the phases are regulated by several layers of checkpoints and control mechanisms (Baserga, 1981; Lubischer, 2007; Bertoli et al., 2013; Grant et al., 2013). The molecular processes behind cell cycle progression have been dissected by numerous morphological studies on live or fixed single cells using a plethora of techniques to visualize components and processes during cell division. Many more investigations have been made on cells, sorted according to size, or artificially arrested at various cell cycle checkpoints. However, most of our knowledge about cell proliferation comes from studies that average data from large and mixed cell populations. Such data are only indirectly

related to quantitative changes in cells at different states of division and growth. Analysis at the single-cell level can overcome most of these limitations. Detailed single-cell analyses have shown that transcript numbers fluctuate in individual cells, even in seemingly homogeneous populations (Raj et al., 2006), and that features of the typical or average cell in a population cannot be deduced from measurements on cell population samples (Bengtsson et al., 2005). Variations in transcript numbers allow cells to produce unique responses to internal and external cues that lead to defined paths of cell proliferation and differentiation (Levine et al., 2013). Recent development of single-cell analytical platforms opens up new possibilities to define the molecular profiles of cells at different states and to determine the importance of cell heterogeneity on cellular processes and cell fate decisions (Kalisky et al., 2011; Ståhlberg et al., 2011b; Sanchez and Golding, 2013; Shapiro et al., 2013).

Here, we employed single-cell gene expression profiling to describe the dynamic transition between cell proliferative states in three different cell lines using a panel consisting of 93 marker genes. Function of selected genes related to cell proliferation, cell cycle regulation, TP53 function, stemness, differentiation, cell signaling, and housekeeping functions (for gene details, see Table S1). We assessed cell division by collecting cells in the G0/G1, S and G2/M phases, and cell growth by selecting small and large cells in respective cell cycle phase. In contrast to cell population data, single-cell data are reported as transcripts per cell without any further normalization (Ståhlberg et al., 2013), allowing total transcript levels to be determined and compared between cell states (Sanchez and Golding, 2013). To determine if, and to what degree, the gene expression profile of individual cells were associated with cell division and growth we applied the random forests algorithm (Hastie et al., 2009; Gareth et al., 2013), which is a supervised machine learning approach. By applying variable selection, a recursive feature elimination (RFE) scheme (James et al., 2013; Candia et al., 2015), we were able to identify the genes with strongest cell proliferation association and to define distinct subpopulations. Finally, we calculated a cell cycle index based on the most predictive genes that allowed us to visualize and biologically interpret cell cycle progression.

MATERIALS AND METHODS

Cell Culture

All cell lines were cultured at 37°C and in 5% CO_2. The myxoid liposarcoma cell line MLS 402-91 was cultured in RPMI 1640 GlutaMAX medium supplemented with 10% fetal bovine serum, 100 U/mL penicillin, and 100 μg/mL streptomycin (all Life Technologies). Cells were passaged with 0.25% trypsin and 0.5 mM EDTA (both Life Technologies). The breast cancer cell line MCF7 was cultured in DMEM medium supplemented with 2 mM L-glutamine, 1% penicillin/streptomycin (all PAA Laboratories), 10% fetal bovine serum (Lonza), and 1% non-essential amino acids (Sigma-Aldrich). MCF7 cells were passaged with 0.05% trypsin-EDTA (PAA Laboratories). Mesenchymal stem cells (MSC) derived from human embryonic stem cells (hES-MP 002.5, Takara Bio), were cultured in DMEM GlutaMAX, supplemented with 10% fetal bovine serum, 100 U/mL penicillin, 100 μg/mL streptomycin, and 4 ng/mL fibroblast growth factor 2 (all Life Technologies) as described (Karlsson et al., 2009). MSCs were passaged with TrypLE Select (Life Technologies). Dissociation enzyme inactivation was performed using complete medium, containing fetal bovine serum for all cell lines. Cell cultures were confirmed as mycoplasma-free using the Mycoplasma PCR Detection Kit (Applied Biological Materials).

Fluorescent Activated Cell Sorting

Vybrant DyeCycle violet stain (Life Technologies) and CellVue Claret far red dye (Sigma-Aldrich) were used to stain genomic DNA and membrane lipids, respectively. Suspension of 10^6 cells in 1 mL Hanks' balanced salt solution (Life Technologies) was first stained with Vybrant DyeCycle violet stain (5 μM, final concentration) at 37°C for 30 min. Then, 1 mL CellVue Claret far red dye diluted in diluent C (Sigma-Aldrich, 3.3 μM, final concentration) was added followed by an incubation step at 37°C for 5 min. Staining was inactivated by complete medium and the cells were finally resuspended in Hanks' balanced salt solution.

G1/S cell cycle arrest was performed using a double thymidine block (Sigma-Aldrich). Thymidine (2 mM, final concentration) was added to 25–30% confluent cells for 18 h. Cells were then released by addition of fresh medium without thymidine. Finally, after 9 h cells were re-exposed to thymidine for additional 17 h. Complete cell cycle arrest was confirmed by Vybrant DyeCycle violet staining followed by fluorescence activated cell sorting analysis.

Cell aggregates were removed by filtering with a 40 μm cell strainer (BD Biosciences) and single cells were sorted with a BD FACSAria II (BD Biosciences) into 96-well-plates (Life Technologies), each well-containing 5 μL 1 mg/mL bovine serum albumin (Thermo Scientific; Svec et al., 2013). Collected single cells were frozen on dry ice and kept at −80°C until subsequent analysis. Gating strategies for cell size and cell cycle phase are shown in Figure S1. The cell size/cell volume was estimated from the average CellVue Claret far red signal, assuming a spherical cell shape. All single-cells from respective biological condition were collected from an individual culture, to minimize batch-to-batch differences as described (Wills et al., 2013).

Single-Cell Gene Expression Profiling

Reverse transcription was performed with SuperScript III (Life Technologies). Lysed single cells, 0.5 mM dNTPs (Sigma-Aldrich), 5.0 μM Oligo(dT_{12-18}), and 5.0 μM random hexamers (both Life Technologies) were incubated in 6.5 μL at 65°C for 5 min. Next, 50 mM Tris–HCl, 75 mM KCl, 3 mM $MgCl_2$, 5 mM dithiothreitol, 10 U RNaseOut, and 50 U SuperScript III (all Life Technologies) were added to a final volume of 10 μL. Final reaction concentrations are shown. Reverse transcription was performed at 25°C for 5 min, 50°C for 60 min, 55°C for 10 min, and terminated by heating to 70°C for 15 min. All samples were diluted to 30 μL with water.

Targeted cDNA preamplification was performed with the iQ Supermix (BioRad) in 50 μL reactions. Each reaction contained 10 or 15 μL diluted cDNA and 40 nM of each primer. Primer sequences are shown in Table S1. Optimization and validation of

good performing qPCR assays and preamplification are described elsewhere (Ståhlberg and Bengtsson, 2010; Andersson et al., 2015). The temperature profile was 95°C for 3 min followed by 20 cycles of amplification (95°C for 20 s, 60°C for 3 min, and 72°C for 20 s). All preamplified samples were chilled on ice and diluted 1:20 in TE-buffer (pH 8.0; Life Technologies). Preamplification was performed as two separate reactions for each single cell, each containing half of the assays. The products of the two reactions were pooled after preamplifciation. Reproducibility and efficiency of the preamplification were evaluated by standard curve analysis using cDNA from MLS 402-91 (Figure S2). The overall preamplification efficiency was assessed using five different cDNA concentrations ($n = 4$) generated from 0.04, 0.2, 1, 5, 25 ng total RNA, respectively. The average cycle of quantification value of all genes expressed in four or more dilutions were used to determine the overall preamplification efficiency.

The BioMark real-time PCR system with 96 × 96 dynamic arrays (Fluidigm) was used for gene expression profiling according to the manufacturer's instructions. The 5 μL sample reaction mixture contained 1X SsoFast EvaGreen Supermix (BioRad), 1X ROX (Life Technologies), 1X GE Sample Loading Reagent (Fluidigm), and 2 μL diluted preamplified cDNA. The 5 μL primer reaction contained 1X Assay Loading Reagent (Fluidigm) and 5 μM of each primer. Preamplification and qPCR were performed with the same primers (Table S1). The chip was first primed with the NanoFlex IFC Controller (Fluidigm) and then loaded with the sample and primer reaction mixtures. The cycling program was 3 min at 95°C for polymerase activation, followed by 40 cycles of amplification (96°C for 5 s and 60°C for 20 s). After qPCR, all samples were analyzed by melting curve analysis (60–95°C with 0.33°C per s increment). All assays were confirmed to generate correct PCR product length by agarose gel electrophoresis. Data pre-processing was performed with GenEx (v.6, MultiD) as described (Ståhlberg et al., 2013). Briefly, samples with aberrant melting curves were removed and cycle of quantification values larger than 25 were replaced with 25. Data were transformed to relative quantities assuming that a cycle of quantification value of 25 equals one molecule. Missing data were replaced with 0.5 molecules. All data were calculated per cell if not stated otherwise. For all data analysis we assumed 100% PCR efficiency. The impact of the chosen cut-off value and applied PCR efficiency had negligible effect on downstream analysis.

Immunofluorescence

MLS 402-91 and MCF-7 cells were seeded on Millicell EZ SLIDE 4-well-glasses (Merck Millipore). After 24 h, cells were rinsed with phosphate buffer saline (Life Technologies) and fixed in 3.7% formaldehyde for 5 min (Sigma-Aldrich), washed three times with phosphate buffer saline and permeabilized in AB buffer (phosphate buffer saline supplied with 1% bovine serum albumin and 0.5% Triton X, Sigma-Aldrich). Cells were stained with anti-MCM6 antibody (HPA004818 rabbit, diluted 1:50, Sigma-Aldrich). Detection was performed with a Cy3 conjugated secondary antibody (PA43004, diluted 1:1000, GE Healthcare Life Sciences). Slides were mounted using Prolong Gold anti-fade with 4′,6-diamidino-2-phenylindole (Life Technologies). Cellular fluorescence was imaged using a Zeiss Axioplan 2 microscope (Zeiss). Relative protein level per cell was estimated using Volocity 3D Image Analysis Software (PerkinElmer).

Single-Cell Data Analysis and Statistics

Principal component analysis, hierarchical clustering, and Kohonen self-organizing maps were performed in GenEx software using autoscaled gene expression data as described (Ståhlberg et al., 2011a). The Ward's algorithm and Euclidean distance measure were applied for hierarchical clustering. Parameters for Kohonen self-organizing maps were: 3–4 × 1 map, 2 neighbors, 0.4 learning rate, and 150 iterations. The resulting clusters were not sensitive to parameter choice.

A random forests algorithm was implemented to pairwise classify different cell cycle phases and cell sizes. Two cell states were compared at a time. Random forests are collections of decision trees. At the top-most level of each decision tree, all genes are scanned one by one, to determine the best gene, and corresponding gene expression threshold to optimally partition the original cells into two branches. The optimal partition is algorithmically determined based on the minimization of a quality function such as the cross-entropy or the Gini index (Hastie et al., 2009; Gareth et al., 2013), which aim to increase the class purity of each branch. Subsequently, each branch is considered for further separation based on the expression values of other genes. The process continues until the full decision tree is grown in such a manner that each of its leaves, i.e., the endpoint of each branch, contains cells of a single class. To generate robust solutions and avoid data overfitting, additional parameters are usually incorporated to the model in order to either limit the length of the tree (or, alternatively, the size of the nodes that can undergo further branching) or to prune the tree. In this context, a popular technique is to generate a so-called random forest that contains a large number of partially decorrelated trees built out of bootstrapped samples from the original data set. Compared to single decision trees, random forests are less intuitive, since they lack a direct visualization of the structure and relations among predictor genes, but random forests are more powerful and robust. In this study, we implemented a random forest analysis using the random Forest (v4.6-10) package in R. This implementation uses the decrease of Gini index impurity as a splitting criterion and selects the splitting predictor from a subset of predictors, randomly chosen at each split. Each random forest consisted of 1000 trees. For each random forest we scanned the size of the predictor subset in the full range from one to the total number of predictors and selected the smallest subset that minimized the out-of-bag error. The so-called out-of-bag error is calculated from predictions on out-of-bag instances, i.e., those cells that have not been used in building a particular tree. Moreover, in order to assess model variance, for each class comparison we generated ensembles consisting of 100 different random forests. Only genes with detectable expression in at least 50% of the cells in at least one cell class were included in our analysis. We report averages and standard deviations calculated over these random forest ensembles throughout.

Cell classification performance can be quantified by several measures. In addition to the out-of-bag error, another measure is

the balanced accuracy. The balanced accuracy is the classification accuracy averaged over all classes, where the classification accuracy for each class is the percentage of cells in the class that are correctly classified by the random forest. Yet another measure is Fisher's *p*-value obtained by applying Fisher's exact test on the confusion matrix, which consists of the number of correctly and incorrectly classified cells in each class. Moreover, we also computed the so-called gene importance, a quantitative measure of the impact of the gene on the node purity.

To address the question of which, and how many, genes are needed to best separate two classes we applied a recursive feature elimination (RFE) scheme, a standard approach for feature selection (Tarca et al., 2007; Candia et al., 2013). In the first RFE cycle, we generated a random forest ensemble using all (N) genes and computed classification statistics, including confusion matrices with associated Fisher's *p*-value, balanced accuracy, out-of-bag error, and gene importance. We determined the least significant gene based on gene importance and removed it. Then, in the second RFE cycle we used the remaining N–1 genes and repeated the random forest analysis to eliminate the second least significant gene. The procedure was subsequently iterated until one gene was left. By comparing the classification performance across all RFE cycles we could then determine the number of genes in the optimal gene signature. We verified that, for this optimal gene signature, the out-of-bag error and Fisher's *p*-value were minimized, while the balanced accuracy was maximized. The intended redundancy of separately considering three classification performance metrics allowed us to ensure the robustness of the optimally obtained gene signature.

The most predictive genes identified by RFE was used to calculate a cell cycle index as the sum of all G1 to S and/or G2/M upregulated genes subtracted by the sum of all G1 to S and/or G2/M downregulated genes divided by the number of genes used. The lg2 expression value of each gene was used.

RESULTS

Gene expression and cell heterogeneity of proliferating cells were studied by fluorescence activated cell sorting combined with single-cell gene expression profiling. Three different cell lines were investigated: a genetically stable myxoid liposarcoma cell line (MLS 402-91) (Aman et al., 1992); a breast cancer adenocarcinoma derived cell line (MCF7; Soule et al., 1973) and mesenchymal stem cells (MSC) differentiated from an embryonic stem cell line (Karlsson et al., 2009). Cells were stained with lipid and DNA binding dyes, visualizing cell size, and DNA content. Utilizing this double-labeling approach we collected small and large cells in the G0/G1, S, and G2/M phases (Figure S1). DNA staining cannot distinguish between G0 and G1 phase cells, or between G2 and M phase cells. We refer the G0/G1 phase as G1 phase only, since few G0 cells are expected in our continuously passaged cell cultures. The average volume ratio between large and small collected cells was 2.8 for MLS 402-91, 2.5 for MCF7, and 4.5 for MSC (Figure S1). Expression of 93 genes were analyzed in each cell using reverse transcription quantitative real-time PCR. One gene (*FUS*) was assessed by two assays. Assay information and gene function are shown in Table S1. All basic data, including number of positive cells expressing each gene and mean single-cell expression with standard deviation, are shown in Table S2. We tested the reproducibility of our data by collecting individual MLS 402-91 cells in the G1, S, and G2/M phases without any cell size selection in an independent experiment.

Total Transcript Level Correlates with Cell Cycle Phase at the Single-Cell Level

Transcript numbers were measured per single cell without any further normalization between cells (Ståhlberg et al., 2011a, 2013). Hence, the total transcript level could be calculated as the sum of all measured transcripts per cell. **Figure 1A** and **Table 1** show that the total transcript level correlated with cell cycle phase, but not with cell size. In MLS 402-91 the total transcript level reached maximum in G2/M phase cells with about two-fold higher levels compared to G1 phase cells. In MCF7 the total transcript level reached maximum in S phase cells and remained at the same level in G2/M phase cells. MSC only displayed a weak correlation between total transcript level and cell cycle phase.

The total transcript level varied highly between individual cells (**Figure 1B**). The distributions were skewed with few cells containing high total transcript levels. The total transcript level was 17, 120, and 820 times higher in the cell with highest total transcript level compared to the cell with lowest total transcript level in MLS 402-91, MCF7, and MSC, respectively (all cells included). Correlation analysis between transcript levels of individual genes at single-cell level showed positive correlations between most genes: 74% in MLS 402-91 (total number of comparisons = 4278), 85% (total number of comparisons = 3081) in MCF7 and 90% (total number of comparisons = 3486) in MSC. Consequently, cells with high total transcript level also displayed elevated transcript numbers of most individual genes.

Identification of Genes with Cell Cycle Phase and Cell Size Dependent Expression

Principal component analysis (PCA) showed that individual cells partly clustered based on their cell cycle phase in all three cell lines (MLS 402-91 in **Figure 2A**, MCF7 in **Figure 3A**, and MSC in **Figure 4A**), but only MSC displayed cell size depended clustering. However, large overlaps between cells of different cell cycle phases and cell sizes were observed for all cell lines. Double thymidine treated MLS 402-91 cells showed a completely divergent expression profile compared to non-treated G1, S, or G2/M phase cells, demonstrating that artificial cell synchronization result in severe and unintended side effects (**Figure 2A**).

To determine if individual cells can be correctly classified into cell cycle phase or cell size based on their gene expression profile we applied the random forests algorithm, a machine-learning approach based on decision trees. As a classifier, a decision tree is a hierarchically organized structure that optimally can separate cell cycle phases and cell sizes (see Section Materials and Methods for details). **Figures 2B**, **3B**, **4B** show how well-cell cycle phase and cell size could be distinguished using a multi-gene signature at the single-cell level. In MLS 402-91, we obtained best classification comparing G2/M with G1 phase cells, while the

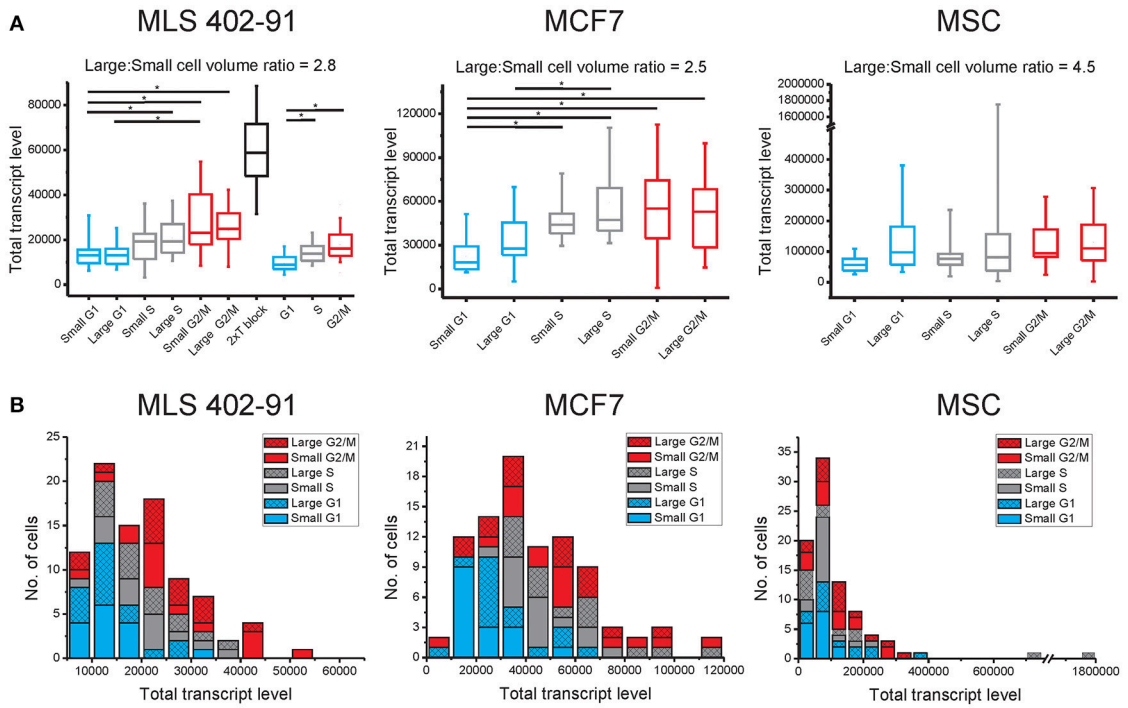


FIGURE 1 | Total transcript levels are mainly cell cycle phase dependent. (A) The total transcript level for small and large cells in the G1 (blue), S (gray), and G2/M (red) phases are shown (MLS 402-91: $n_{small-G1} = 15$, $n_{large-G1} = 16$, $n_{small-S} = 15$, $n_{large-S} = 15$, $n_{small-G2/M} = 15$, $n_{large-G2/M} = 15$; MCF7: $n_{small-G1} = 16$, $n_{large-G1} = 15$, $n_{small-S} = 15$, $n_{large-S} = 15$, $n_{small-G2/M} = 15$, $n_{large-G2/M} = 16$, and MSC: $n_{small-G1} = 16$, $n_{large-G1} = 13$, $n_{small-S} = 16$, $n_{large-S} = 12$, $n_{small-G2/M} = 14$, $n_{large-G2/M} = 15$). In addition, G1/S phase arrested MLS 402-91 cells with any cell size were analyzed, using a double thymidine block ($n = 61$). As a separate experiment, MLS 402-91 cells were collected and analyzed based on cell cycle phase only ($n_{G1} = 30$, $n_S = 29$ and $n_{G2/M} = 30$). Box-Whisker plots are shown; the box ranges between the 25 and 75% and the whiskers range between the 5 and 95% of all data. *indicate 95% significance using the Mann-Whitney *U*-test with Holm-Bonferroni correction for multiple testing. **(B)** Distribution of total transcript levels among individual cells in MLS 402-91, MCF7, and MSC. The total transcript level per cell is calculated as the sum of all measured transcript for all 93 genes.

classifications between other cell cycle phases were less efficient (**Figure 2B**). For example, 29.86 ± 0.35 out of 31 MLS 402-91 cells were correctly classified as G1 phase cells, while 1.14 ± 0.35 G1 phase cells were falsely predicted to be G2/M phase cells. The ability to classify MCF7 cells was similar (**Figure 3B**). The gene expression profile was less predictive to classify cell size than cell cycle phase in both MLS 402-91 and MCF7 cells (**Figures 2B**, **3B**). Similar gene expression profiles and classifications were also observed for the independent MLS 402-91 data set (Figure S3). The gene expression profile of individual MSC was less predictive for cell cycle phases compared to the two other cell lines, but the ability to classify cell size was more efficient in MSC (**Figure 4B**). We also compared small and large cells within respective cell cycle phase, but no distinct cell size dependency was found in any of the three cell lines (data not shown). The random forests approach also allowed us to rank the individual genes based on their importance in the classification (Figure S4). **Figures 2C**, **3C**, **4C** show the genes with strongest cell cycle phase and cell size dependent expression. Even if the median expression level of these predictive genes correlated well with their ability to classify cell cycle phase or cell size, individual cells showed highly variable, and overlapping gene expression (**Figures 2C**, **3C**, **4C**).

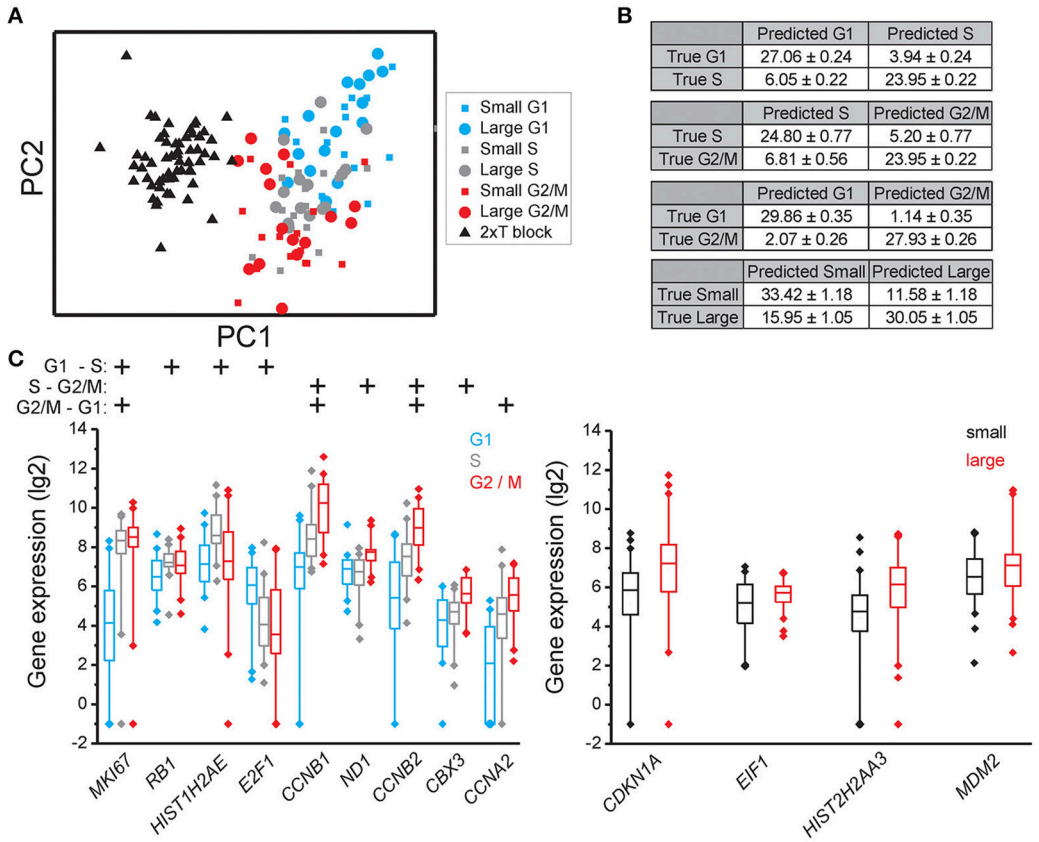
TABLE 1 | Spearman's correlation coefficient between total transcript level and cell proliferation parameters at single-cell level.

	MLS 402-91	MCF7	MSC
Cell cycle phase combined with cell size	0.27*	0.34**	0.28**
Cell cycle phase	0.51**	0.47**	0.23*
Cell size	0.03		0.19

*$p < 0.05$, **$p < 0.01$.

Identification of Predictive Genes and Cell Line Specific Subpopulations

Expression data for all genes were used in the random forests classification algorithm to predict cell cycle phase and cell size. To determine if a similar prediction model could be generated with fewer genes, we applied a recursive feature elimination (RFE) approach. In RFE, the least informative gene is eliminated from the random forests analysis. This procedure is repeated until only one gene remains. Figure S5 shows how well the random forests algorithm performed with decreasing number of genes. We found that expression data from the following gene sets were almost as accurate as the complete gene panel in classifying cell cycle phase in MLS 402-91: G1 vs. S: *MKI67*, *RB1*, *E2F1*,



	Predicted G1	Predicted S
True G1	27.06 ± 0.24	3.94 ± 0.24
True S	6.05 ± 0.22	23.95 ± 0.22

	Predicted S	Predicted G2/M
True S	24.80 ± 0.77	5.20 ± 0.77
True G2/M	6.81 ± 0.56	23.95 ± 0.22

	Predicted G1	Predicted G2/M
True G1	29.86 ± 0.35	1.14 ± 0.35
True G2/M	2.07 ± 0.26	27.93 ± 0.26

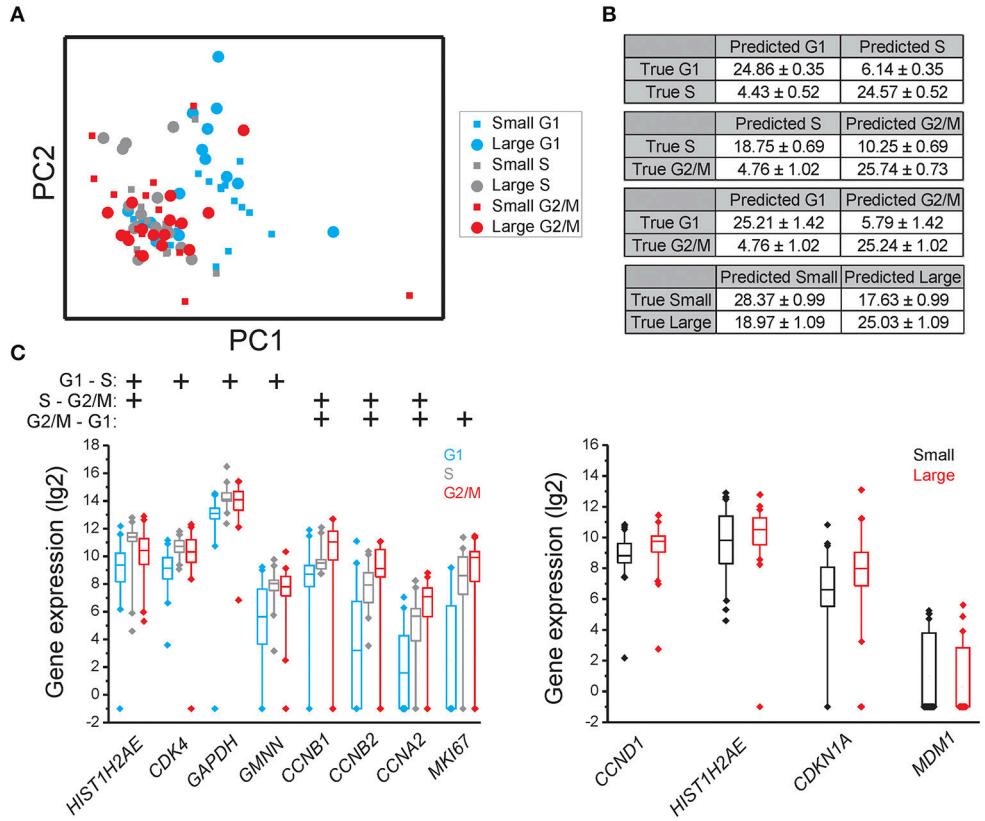
	Predicted Small	Predicted Large
True Small	33.42 ± 1.18	11.58 ± 1.18
True Large	15.95 ± 1.05	30.05 ± 1.05

FIGURE 2 | Cell cycle phase and cell size dependent gene expression in MLS 402-91. (A) PCA of small and large MLS 402-91 cells in the G1, S, and G2/M phases. Note that the double thymidine treated cells (T block) show a completely different expression profile than non-treated cells. Each dot, square, and triangle represents a single cell. **(B)** Confusion matrices of cell classifications using the random forests algorithm. Fisher's exact test was used to calculate significance ($p <$ 0.0001) for all matrices. **(C)** Box-Whisker plots for the genes with highest importance to classify cell cycle phase and cell size using the random forests algorithm. The box ranges between the 25 and 75%, the whiskers range between the 5 and 95% of all data and outliers are indicated as diamonds.

HIST1H2AE, and *CCNB1*; S vs. G2/M: *CCNB1*, *CBX3*, and *ND1* and G2/M vs. G1: *MKI67*, *GAPDH*, *CCNB1*, and *CCNB2*. The gene lists are ordered with the most predictive gene listed first. Refined PCA using only these nine predictive genes revealed a distinct subpopulation that was not clearly visible using all genes (**Figure 5A**). The same subpopulation was also identified using other algorithms, including hierarchical clustering and Kohonen self-organizing maps (Figure S6). This new subpopulation mainly consisted of G1 cell cycle phase cells and was characterized by upregulation of *MCM6* and downregulation of 21 other genes, mainly cell cycle related genes (**Figures 5B,C**). We refer to this subpopulation as the G1′ subpopulation. The total transcript level in the G1′ subpopulation was on average 32% lower compared to the other G1 phase cells ($p < 0.01$, Mann-Whitney *U*-test), suggesting a distinct G1 cell state with low transcriptional activity. We also confirmed the presence of the same G1 subpopulation with almost an identical gene expression profile in the independent MLS 402-91 data set (Figure S7).

In MCF7, the following sets of predictive genes were identified by RFE: G1 vs. S phase: *HIST1H2AE*, *CCNB1*, *CDK4*, and *GMNN*; S vs. G2/M phase: *CCNB1*, *CCNB2*, and *HIST1H2AE* and G2/M vs. G1 phase: *MKI67*, *CCNB1*, *RPS10*, *RPL7*, and *EIF1*. Refined PCA revealed a G1 subpopulation with similar characteristics as the G1′ subpopulation found in MLS 402-91 (**Figures 5D–F**). The existence of the MCF7 defined G1′ subpopulation was confirmed by hierarchical clustering and Kohonen self-organizing maps (data not shown). The total transcript level was 47% lower in the G1′ subpopulation compared to the other G1 phase cells ($p < 0.01$, Mann-Whitney *U*-test). One gene, *MCM6*, displayed opposite regulation in the G1′ subpopulation in MCF7 compared to MLS 402-91. The variable and divergent *MCM6* expression prompted us to analyze its protein expression. Immunofluorescence analysis showed variable MCM6 protein expression in both MLS 402-91 and MCF7 with somewhat higher variability in MCF7 cells (Figure S8).

In MSC, RFE generated the following sets of predictive genes: G1 vs. S phase: *HIST1H2AE*, *MKI67*, *ATF4*, and *YWHAZ*; S vs. G2/M phase: *HIST1H2AE*, *E2F4*, *TAF15*, and *RB1* and G2/M vs. G1 phase: *CCNA2*, *NOTCH1*, *CCNB1*, and *VIM*. In contrast to MLS 402-91 and MCF7, MSC displayed a distinct subpopulation



	Predicted G1	Predicted S
True G1	24.86 ± 0.35	6.14 ± 0.35
True S	4.43 ± 0.52	24.57 ± 0.52

	Predicted S	Predicted G2/M
True S	18.75 ± 0.69	10.25 ± 0.69
True G2/M	4.76 ± 1.02	25.74 ± 0.73

	Predicted G1	Predicted G2/M
True G1	25.21 ± 1.42	5.79 ± 1.42
True G2/M	4.76 ± 1.02	25.24 ± 1.02

	Predicted Small	Predicted Large
True Small	28.37 ± 0.99	17.63 ± 0.99
True Large	18.97 ± 1.09	25.03 ± 1.09

FIGURE 3 | Cell cycle phase and cell size dependent gene expression in MCF7. (A) PCA of small and large MCF7 cells in the G1, S, and G2/M phases. Each dot and square represents a single cell. **(B)** Confusion matrices of cell classifications using the random forests algorithm. Fisher's exact test was used to calculate significance ($p < 0.0001$) for all matrices. The confusion matrix of small compared to large cells was not significant. **(C)** Box-Whisker plots for the genes with highest importance to classify cell cycle phase and cell size using the random forests algorithm. The box ranges between the 25 and 75%, the whiskers range between the 5 and 95% of all data and outliers are indicated as diamonds.

of small S and G2/M phase cells that was characterized by upregulated cell proliferation genes (**Figures 5G–I**). The existence of this MSC specific subpopulation was also confirmed by other algorithms (data not shown).

Cell Cycle Progression Can Be Visualized By a Cell Cycle Index Based on Gene Expression

Multi-gene profiles are usually hard to visualize and interpret. Hence, we calculated and plotted a cell cycle index based on the expression of all cell cycle regulated genes identified by RFE for each cell line (**Figure 6**). The index correlated with the cell cycle progression for all three cell lines, where G1 phase cells showed low indexes, while G2/M phase cells displayed high indexes. The cell cycle index varied most between individual G1 phase cells in MLS 402-91 and MCF7, where a distinct index crossover point could be identified for cells in the transition from G1 to S phase. In contrast, MSC showed a different pattern with a more uniform G1 to S phase transition. The cells in the G1′ subpopulations identified in MLS 402-91 and MCF7 displayed the lowest cell cycle indexes, while the cells in the subpopulation defined in MSC showed the highest indexes.

DISCUSSION

The mechanisms governing cell growth and division of mammalian cells have long been a subject of intense research. Many of the decisive regulatory events occur by post translational modifications of pre-existing proteins (Pagliuca et al., 2011), but underlying this regulatory level is also synchronized *de novo* production of cell cycle regulated components. A large number of genes have been reported to be timely transcribed as part of cell cycle progression (Sun et al., 2007; Simmons Kovacs et al., 2008; Muller and Engeland, 2010). Here, we have taken advantage of emerging technology to study gene expression profiles in single cells of different cell cycle phases and of different cell sizes. To date, most studies aimed at cell cycle regulated gene transcription were based on large cultures and artificial cell synchronization. We and others (Cooper, 2002, 2003) have observed that standard synchronization strategies affect cell states in unintended ways as they cause cell stress and abnormal expression profiles (**Figures 1A**, **2A**). Our approach to collect unsynchronized individual cells avoids these issues and our data clearly demonstrate some of the benefits using single-cell analysis. Both the observed cell-to-cell variability and the identified

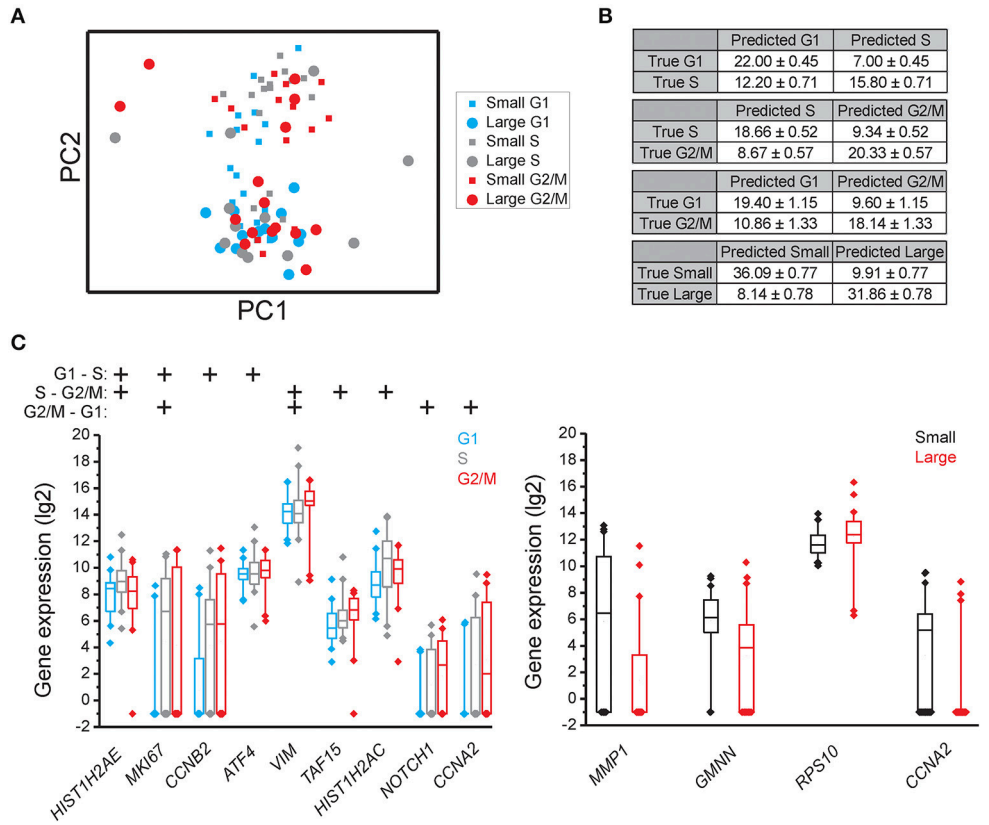


FIGURE 4 | Cell cycle phase and cell size dependent gene expression in MSC. (A) PCA of small and large MSC cells in the G1, S, and G2 phases. Each dot and square represents a single cell. **(B)** Confusion matrices of cell classifications using the random forests algorithm. Fisher's exact test was used to calculate significance ($p < 0.05$) for all matrices. **(C)** Box-Whisker plots for the genes with highest importance to classify cell cycle phase and cell size using the random forests algorithm. The box ranges between the 25 and 75%, the whiskers range between the 5 and 95% of all data and outliers are indicated as diamonds.

subpopulations would have been challenging to study at cell population level.

Traditional expression analysis usually involves normalization processes before samples can be compared. Normalization assumes that selected house-keeping genes, i.e., reference genes, or the total amount of transcripts is essentially identical across samples. However, single-cell RT-qPCR data are reported as transcripts per cell without the need of additional normalization between cells, which enable us to calculate the total transcript level of all analyzed genes (Ståhlberg et al., 2011a, 2013). This strategy is possible, since single cells are analyzed directly without any extraction steps. Our data show that the assumption of equal total transcription levels between individual cells is not valid. Instead, we observed that the total transcript level correlated with the cell cycle phase (**Table 1**). This was further tested by analyzing an additional published single-cell astrocyte data set generated directly from dissociated mice brains (Figure S9; Rusnakova et al., 2013). Taken together, our data show a considerable cell-to-cell variation in total transcript levels where most genes are positively correlated. In addition, only a minority of cells displayed elevated total transcript levels. Consequently, these few cells expressed high number of transcripts of most genes. The absolute values of the calculated total transcript levels are dependent on the applied gene panel. However, the observation of subpopulations expressing elevated levels of transcripts for most genes is not gene panel dependent. Our results are in agreement with earlier observations that transcription occurs in bursts (Raj et al., 2006; Sanchez and Golding, 2013), generating skewed distributions of transcripts among individual cells (Bengtsson et al., 2005).

In many organisms cell size is strongly correlated to cell division and growth rate (Dungrawala et al., 2010; Marguerat and Bahler, 2012), but the role of cell size in mammalian cells is less clear (Echave et al., 2007; Tzur et al., 2009). Our cell size data are in line with these reports. We observed increased numbers of small cells in the G1 phase using fluorescence activated cell sorting (Figure S1), but no clear correlation between cell size and total transcript levels were observed in any cell line. In MSC, we identified a subpopulation of small S and G2/M phase cells with distinct gene expression profile. The divergent results of MSC could be connected to the larger span in size variation of these cells compared to the other two cell lines (**Figure 1A** and Figure S1).

A large number of genes displayed correlations between their expression levels and cell cycle phase, while the number of correlations between expression level and cell size was fewer (**Table 1** and Table S2). However, even for the genes with highest correlations we observed large overlap in gene expression levels among individual cells of different cell cycle

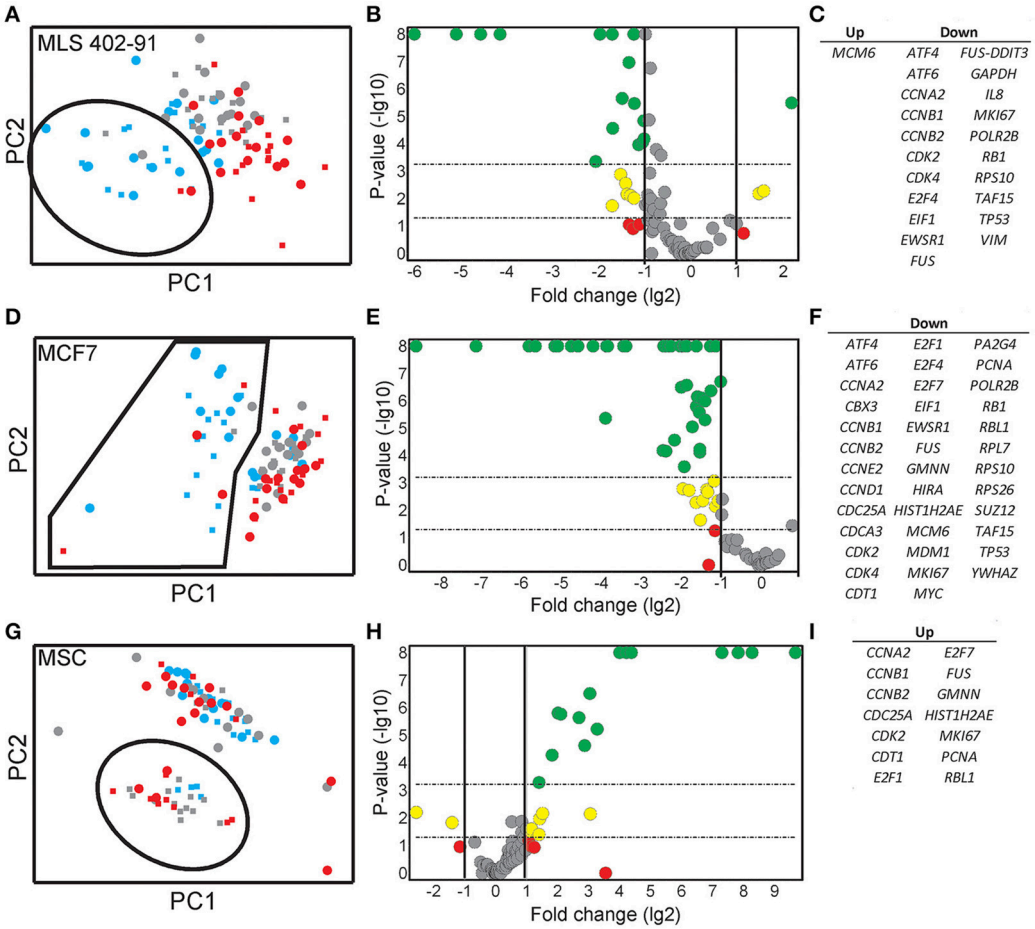


FIGURE 5 | Identification and characterization of distinct subpopulations. (A) A MLS 402-91 subpopulation (encircled) was defined using PCA and RFE identified genes (*MKI67*, *RB1*, *HIST1H2AE*, *CCNB1*, *CBX3*, *ND1*, *GAPDH*, *CCNB2*, and *E2F1)*. Individual small (squares) and large (dots) MLS 402-91 cells in G1 (blue), S (gray), and G2/M (red) phase are shown. **(B)** The volcano plot shows regulation and significance of all analyzed genes, comparing the defined G1 subpopulation and the remaining G1 phase cells. Dunn-Bonferroni correction for multiple testing ($p < 0.00054$) was applied using 95% significance. Red ($p > 0.05$), yellow ($0.05 > p > 0.00054$), and green ($p < 0.00054$) dots indicate at least two-fold regulated genes. **(C)** All significantly MLS 402-91 regulated genes identified in the volcano plot are listed. **(D)** A MCF7 subpopulation (encircled) was defined using PCA and RFE identified genes (*HIST1H2AE*, *CCNB1*, *CDK4*, *GMNN*, *CCNB2*, MKI67, RPS10, *RPL7*, and *EIF1*). Individual small (squares) and large (dot) MCF7 cells in G1 (blue), S (gray), and G2/M (red) phase are shown. **(E)** The volcano plot shows regulation and significance for all analyzed genes in MCF7, comparing the defined G1 subpopulation and the remaining G1 phase cells. Dunn-Bonferroni correction for multiple testing ($p < 0.00062$) was applied using 95% significance. Red ($p > 0.05$), yellow ($0.05 > p > 0.00062$) and green ($p < 0.00062$) dots indicate at least two-fold regulated genes. **(F)** All significantly MCF7 regulated genes identified in the volcano plot are listed. **(G)** A MSC subpopulation (encircled) was defined using PCA and RFE identified genes (*HIST1H2AE*, *MKI67*, *ATF4*, *YWHAZ*, E2F4, TAF15, *RB1*, *CCNA2*, *NOTCH1*, *CCNB1*, and *VIM*). Individual small (squares) and large (dot) MCF7 cells in G1 (blue), S (gray), and G2/M (red) phase are shown. **(H)** The volcano plot shows regulation and significance for all analyzed genes in MSC, comparing the defined subpopulation and the remaining cells. Dunn-Bonferroni correction for multiple testing ($p < 0.0006$) was applied using 95% significance. Red ($p > 0.05$), yellow ($0.05 > p > 0.0006$), and green ($p < 0.0006$) dots indicate at least two-fold regulated genes. **(I)** All significantly MSC regulated genes identified in the volcano plot are listed.

phases and cell sizes (**Figures 2C**, **3C**, **4C** and Table S2). To further analyze the relations between gene expression and cell cycle phase respective cell size we applied the supervised random forests learning algorithm. This strategy generated a multi-gene signature that optimally separated pre-defined cell populations. Further, to identify the most predictive genes we applied RFE. Most of the predictive genes were similar in MLS 402-91 and MCF7, while MSC displayed a different gene list. Some genes, including *CCNB1* and *MKI67*, were predictive in all three cell lines. The RFE results showed that none of the measured genes alone or in combination could predict all cells into correct cell cycle phase or cell size in any cell line.

By excluding non-informative genes in the PCA we identified distinct G1′ subpopulations in both MLS 402-91 and MCF7. The G1′ subpopulations were characterized by low total transcript levels and downregulation of several proliferation associated genes. We speculate that these G1 phase cells are cells that have recently divided (Martinsson et al., 2005). One gene, *MCM6*, was upregulated in MLS 402-91, while downregulated in MCF7. *MCM6* belongs to the *MCM* gene family, where the MCM complex is loaded on chromatin exclusively during the G1 phase

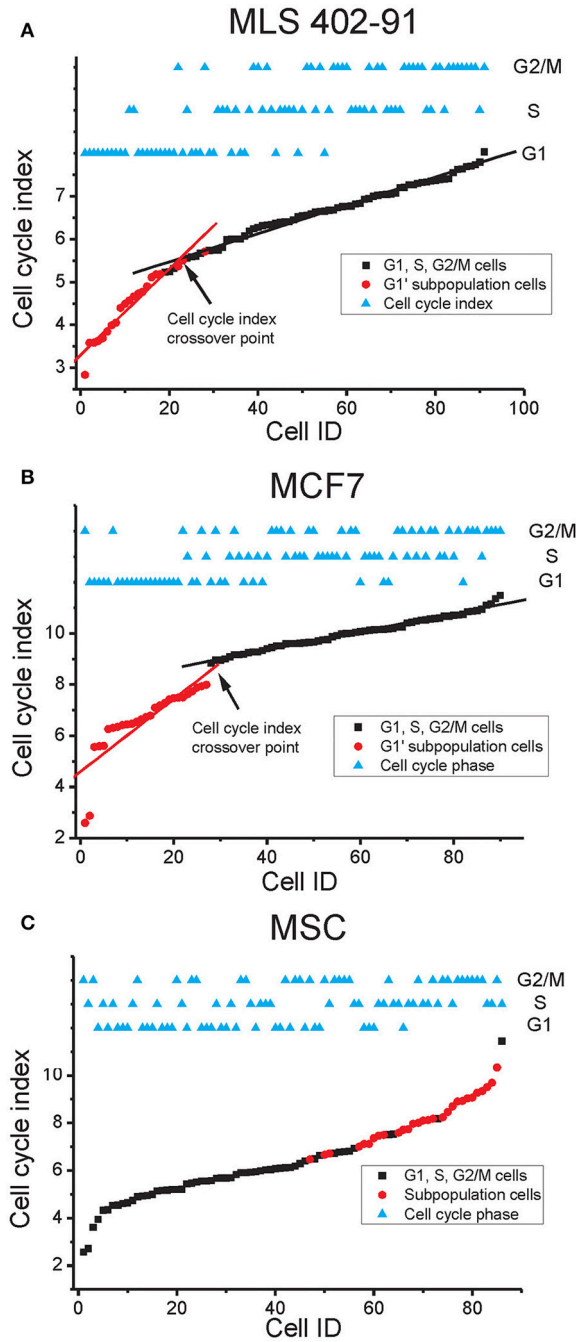


FIGURE 6 | Cell cycle index. The cell cycle index of each cell is shown in relation to its cell cycle phase. Subpopulation cells identified in **Figure 5** are also indicated. **(A)** The MLS 402-91 index was calculated as: (*MKI67* + *RB1* + *HIST1H2AE* + *CCNB1* + *CBX3* + *ND1* + *GAPDH* + *CCNB2* – *E2F1*)/9. The lg2 expression value of each gene was used. The cell cycle index crossover point where the index enters a plateau is indicated. The linear fits are shown to guide the eye. **(B)** The MCF7 index was calculated as: (*HIST1H2AE* + *CCNB1* + *CDK4* + *GMNN* + *CCNB2* + *MKI67* + *RPS10* + *RPL7* +*EIF1*)/9. The lg2 expression value of each gene was used. The cell cycle index crossover point where the index enters a plateau is indicated. The linear fits are shown to guide the eye. **(C)** The MSC index was calculated as: (*HIST1H2AE* + *MKI67* + *ATF4* + *YWHAX* + *E2F4* + *TAF15* + *RB1* + *CCNA2* + *NOTCH1* + *CCNB2* + *VIM*)/11. The lg2 expression value of each gene was used.

with help of other proteins, including CDT1 and CDC6 (Shetty et al., 2005). Interestingly, the second most upregulated gene in the MLS 402-91 G1′ subpopulation was *CDT1*, further indicating that the MCM complex may be differently regulated in MLS 402-91 compared to MCF7. The heterogeneously *MCM6* expression also translated into variable protein expression levels. Transcript data suggest that the cells with high MCM6 protein level in MLS 402-91 correspond to the G1′ subpopulation, while the opposite seems true for MCF7. Further, analyses are needed to define the cell line specific regulation of *MCM* genes.

A single parameter is easier to visualize and interpret than a multi-gene signature. Hence, we developed a cell cycle index to illustrate cell cycle progression. The index shows that cells are in continuous transition throughout the cell cycle until mitosis. In MLS 402-91 and MCF7 we observed a distinct cell cycle index crossover point for cells that were in the G1 to S phase transition (**Figures 6A–B**). We speculate that this cell cycle index breakpoint is related to the G1 restriction check point (Lubischer, 2007). The identified G1′ subpopulations in MLS 402-91 and MCF7 were characterized by low indexes, illustrating that these cells are not likely to enter the S phase in the near future. However, further analysis of more cell lines in different conditions, degree of differentiation and various genetic backgrounds is needed to determine general cell proliferation constraints. In addition, whole transcriptome analysis would most likely reveal more predictive genes allowing for a more detailed understanding of cell transitions between cell cycle phases.

AUTHOR CONTRIBUTIONS

AS conceived and designed the study. AS, SD, NA, CV, TT performed the experiments. AS, JC, WL performed data analysis. All authors were involved in data interpretation and manuscript drafting. All authors approved the final manuscript.

FUNDING

Barncancerfonden, BioCARE, Cancerfonden, Johan Jansson Stiftelsen för tumörforskning och cancerskadade, Sahlgrenska Akademin-ALF, Stiftelsen Assar Gabrielssons Fond, Stiftelserna Wilhelm och Martina Lundgrens Vetenskapsfond, VINNOVA, Åke Wiberg Stiftelse.

ACKNOWLEDGMENTS

We acknowledge the Centre for Cellular Imaging at the Sahlgrenska Academy, University of Gothenburg for imaging support and Dr. Daniel Andersson at the Sahlgrenska Cancer Center, University of Gothenburg, Gothenburg, Sweden for comments on the manuscript draft.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00001/full#supplementary-material

REFERENCES

Aman, P., Ron, D., Mandahl, N., Fioretos, T., Heim, S., Arheden, K., et al. (1992). Rearrangement of the transcription factor gene CHOP in myxoid liposarcomas with t(12;16)(q13;p11). *Genes Chromosomes Cancer* 5, 278–285. doi: 10.1002/gcc.2870050403

Andersson, D., Akrap, N., Svec, D., Godfrey, T. E., Kubista, M., Landberg, G., et al. (2015). Properties of targeted preamplification in DNA and cDNA quantification. *Expert Rev. Mol. Diagn.* 15, 1085–1100. doi: 10.1586/14737159.2015.1057124

Baserga, R. (1981). The cell cycle. *N.Engl. J. Med.* 304, 453–459. doi: 10.1056/NEJM198102193040803

Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* 15, 1388–1392. doi: 10.1101/gr.3820805

Bertoli, C., Skotheim, J. M., and de Bruin, R. A. (2013). Control of cell cycle transcription during G1 and S phases. *Nat. Rev. Mol. Cell Biol.* 14, 518–528. doi: 10.1038/nrm3629

Candia, J., Banavar, J. R., and Losert, W. (2015). "Uncovering phenotypes with supercells: applications to single-cell sequencing," in *Single Cell Sequencing and Systems Immunology*, ed X. Wang (Dordrecht: Springer), 11–30.

Candia, J., Maunu, R., Driscoll, M., Biancotto, A., Dagur, P., McCoy, J. P., et al. (2013). From cellular characteristics to disease diagnosis: uncovering phenotypes with supercells. *PLoS Comput. Biol.* 9:e1003215. doi: 10.1371/journal.pcbi.1003215

Cooper, S. (2002). Minimally disturbed, multicycle, and reproducible synchrony using a eukaryotic "baby machine." *Bioessays* 24, 499–501. doi: 10.1002/bies.10108

Cooper, S. (2003). Rethinking synchronization of mammalian cells for cell cycle analysis. *Cell. Mol. Life Sci.* 60, 1099–1106. doi: 10.1007/s00018-003-2253-2

Dungrawala, H., Manukyan, A., and Schneider, B. L. (2010). Gene regulation: global transcription rates scale with size. *Curr. Biol.* 20, R979–R981. doi: 10.1016/j.cub.2010.09.064

Echave, P., Conlon, I. J., and Lloyd, A. C. (2007). Cell size regulation in mammalian cells. *Cell Cycle* 6, 218–224. doi: 10.4161/cc.6.2.3744

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York, NY; Heidelberg; Dordrecht; London: Springer.

Grant, G. D., Brooks, L. III, Zhang, X., Mahoney, J. M., Martyanov, V., Wood, T. A., et al. (2013). Identification of cell cycle-regulated genes periodically expressed in U2OS cells and their regulation by FOXM1 and E2F transcription factors. *Mol. Biol. Cell* 24, 3634–3650. doi: 10.1091/mbc.E13-05-0264

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn*. New York, NY: Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY; Heidelberg; Dordrecht; London: Springer.

Kalisky, T., Blainey, P., and Quake, S. R. (2011). Genomic analysis at the single-cell level. *Annu. Rev. Genet.* 45, 431–445. doi: 10.1146/annurev-genet-102209-163607

Karlsson, C., Emanuelsson, K., Wessberg, F., Kajic, K., Axell, M. Z., Eriksson, P. S., et al. (2009). Human embryonic stem cell-derived mesenchymal progenitors–potential in regenerative medicine. *Stem Cell Res.* 3, 39–50. doi: 10.1016/j.scr.2009.05.002

Levine, J. H., Lin, Y., and Elowitz, M. B. (2013). Functional roles of pulsing in genetic circuits. *Science* 342, 1193–1200. doi: 10.1126/science.1239999

Lubischer, J. L. (2007). The cell cycle, principles of control. *David O. Morgan. Integr. Comp. Biol.* 47, 794–795. doi: 10.1093/icb/icm066

Marguerat, S., and Bahler, J. (2012). Coordinating genome expression with cell size. *Trends Genet.* 28, 560–565. doi: 10.1016/j.tig.2012.07.003

Martinsson, H. S., Zickert, P., Starborg, M., Larsson, O., and Zetterberg, A. (2005). Changes in cell shape and anchorage in relation to the restriction point. *J. Cell. Physiol.* 203, 27–34. doi: 10.1002/jcp.20204

Muller, G. A., and Engeland, K. (2010). The central role of CDE/CHR promoter elements in the regulation of cell cycle-dependent gene transcription. *FEBS J.* 277, 877–893. doi: 10.1111/j.1742-4658.2009.07508.x

Pagliuca, F. W., Collins, M. O., Lichawska, A., Zegerman, P., Choudhary, J. S., and Pines, J. (2011). Quantitative proteomics reveals the basis for the biochemical specificity of the cell-cycle machinery. *Mol. Cell* 43, 406–417. doi: 10.1016/j.molcel.2011.05.031

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309. doi: 10.1371/journal.pbio.0040309

Rusnakova, V., Honsa, P., Dzamba, D., Ståhlberg, A., Kubista, M., and Anderova, M. (2013). Heterogeneity of astrocytes: from development to injury - single cell gene expression. *PLoS ONE* 8:e69734. doi: 10.1371/journal.pone.0069734

Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. *Science* 342, 1188–1193. doi: 10.1126/science.1242975

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14, 618–630. doi: 10.1038/nrg3542

Shetty, A., Loddo, M., Fanshawe, T., Prevost, A. T., Sainsbury, R., Williams, G. H., et al. (2005). DNA replication licensing and cell cycle kinetics of normal and neoplastic breast. *Br. J. Cancer* 93, 1295–1300. doi: 10.1038/sj.bjc.6602829

Simmons Kovacs, L. A., Orlando, D. A., and Haase, S. B. (2008). Transcription networks and cyclin/CDKs: the yin and yang of cell cycle oscillators. *Cell Cycle* 7, 2626–2629. doi: 10.4161/cc.7.17.6515

Soule, H. D., Vazguez, J., Long, A., Albert, S., and Brennan, M. (1973). A human cell line from a pleural effusion derived from a breast carcinoma. *J. Natl. Cancer Inst.* 51, 1409–1416.

Ståhlberg, A., Andersson, D., Aurelius, J., Faiz, M., Pekna, M., Kubista, M., et al. (2011a). Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res.* 39, e24. doi: 10.1093/nar/gkq1182

Ståhlberg, A., and Bengtsson, M. (2010). Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods* 50, 282–288. doi: 10.1016/j.ymeth.2010.01.002

Ståhlberg, A., Kubista, M., and Aman, P. (2011b). Single-cell gene-expression profiling and its potential diagnostic applications. *Expert Rev. Mol. Diagn.* 11, 735–740. doi: 10.1586/erm.11.60

Ståhlberg, A., Rusnakova, V., Forootan, A., Anderova, M., and Kubista, M. (2013). RT-qPCR work-flow for single-cell data analysis. *Methods* 59, 80–88. doi: 10.1016/j.ymeth.2012.09.007

Sun, A., Bagella, L., Tutton, S., Romano, G., and Giordano, A. (2007). From G0 to S phase: a view of the roles played by the retinoblastoma (Rb) family members in the Rb-E2F pathway. *J. Cell. Biochem.* 102, 1400–1404. doi: 10.1002/jcb.21609

Svec, D., Andersson, D., Pekny, M., Sjoback, R., Kubista, M., and Ståhlberg, A. (2013). Direct cell lysis for single-cell gene expression profiling. *Front. Oncol.* 3:274. doi: 10.3389/fonc.2013.00274

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., and Draghici, S. (2007). Machine learning and its applications to biology. *PLoS Comput. Biol.* 3:e116. doi: 10.1371/journal.pcbi.0030116

Tzur, A., Kafri, R., LeBleu, V. S., Lahav, G., and Kirschner, M. W. (2009). Cell growth and size homeostasis in proliferating animal cells. *Science* 325, 167–171. doi: 10.1126/science.1174294

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., et al. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* 31, 748–752. doi: 10.1038/nbt.2642

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

METHODS
published: 23 May 2016
doi: 10.3389/fbioe.2016.00042

Efficient Synergistic Single-Cell Genome Assembly

*Narjes S. Movahedi[1], Mallory Embree[2], Harish Nagarajan[2], Karsten Zengler[2] and Hamidreza Chitsaz[3]**

[1]Department of Computer Science, Wayne State University, Detroit, MI, USA, [2]Department of Bioengineering, University of California San Diego, San Diego, CA, USA, [3]Department of Computer Science, Colorado State University, Fort Collins, CO, USA

As the vast majority of all microbes are unculturable, single-cell sequencing has become a significant method to gain insight into microbial physiology. Single-cell sequencing methods, currently powered by multiple displacement genome amplification (MDA), have passed important milestones such as finishing and closing the genome of a prokaryote. However, the quality and reliability of genome assemblies from single cells are still unsatisfactory due to uneven coverage depth and the absence of scattered chunks of the genome in the final collection of reads caused by MDA bias. In this work, our new algorithm Hybrid *De novo* Assembler (HyDA) demonstrates the power of coassembly of multiple single-cell genomic data sets through significant improvement of the assembly quality in terms of predicted functional elements and length statistics. Coassemblies contain significantly more base pairs and protein coding genes, cover more subsystems, and consist of longer contigs compared to individual assemblies by the same algorithm as well as state-of-the-art single-cell assemblers SPAdes and IDBA-UD. Hybrid *De novo* Assembler (HyDA) is also able to avoid chimeric assemblies by detecting and separating shared and exclusive pieces of sequence for input data sets. By replacing one deep single-cell sequencing experiment with a few single-cell sequencing experiments of lower depth, the coassembly method can hedge against the risk of failure and loss of the sample, without significantly increasing sequencing cost. Application of the single-cell coassembler HyDA to the study of three uncultured members of an alkane-degrading methanogenic community validated the usefulness of the coassembly concept. HyDA is open source and publicly available at http://chitsazlab.org/software.html, and the raw reads are available at http://chitsazlab.org/research.html.

Keywords: genome assembly, single-cell genomics, uncultivable bacteria, colored de Bruijn graph, genome coassembly

Edited by:
Xinghua Pan,
Yale University, USA

Reviewed by:
Malek Faham,
Sequenta Inc., USA
Xuefeng Wang,
State University of New York at Stony Brook, USA

****Correspondence:***
Hamidreza Chitsaz
chitsaz@chitsazlab.org

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *24 February 2016*
Accepted: *06 May 2016*
Published: *23 May 2016*

Citation:
Movahedi NS, Embree M, Nagarajan H, Zengler K and Chitsaz H (2016) Efficient Synergistic Single-Cell Genome Assembly. Front. Bioeng. Biotechnol. 4:42. doi: 10.3389/fbioe.2016.00042

1. INTRODUCTION

Enormous progress toward DNA sequencing has brought a realm of exciting applications within reach, including genomic analysis at single-cell resolution. Single-cell genome sequencing holds great promise for various areas of biology including environmental biology (McLean et al., 2013). In particular, myriad unculturable environmental microorganisms have been studied using single-cell genome sequencing powered by high-throughput DNA amplification methods (Dean et al., 2001, 2002;

Hosono et al., 2003; Gill et al., 2006; Rusch et al., 2007). Since the majority of microbes to date are unculturable, single-cell sequencing has enabled significant progress in elucidating the genome sequences and metabolic capabilities of these previously inaccessible microorganisms.

Single-cell sequencing, which was challenging and limited for years, is now accessible and attractive for many scientific fields according to the Nature Method of the year 2013. It helps various types of projects such as antibiotics discovery (Li and Vederas, 2009), Earth Microbiome Project (EMP) (Caporaso et al., 2012), and Human Microbiome Project (HMP) (Gill et al., 2006). The importance of single-cell sequencing is particularly due to the fact that only 1% of environmental bacteria have been cultured in the laboratory as they need their natural habitant for cultivation (Lasken, 2007). Also, single-cell sequencing can preserve the uniqueness of each cell and its individual mutations and structural variations, which are valuable information, especially in cancer studies.

Nevertheless, single-cell sequencing is still far from perfect as whole-genome amplification procedures are needed to augment femtograms of DNA material of one cell into micrograms. All known amplification reactions to date introduce some form of bias. Today, the dominant amplification method in single-cell sequencing technology is the Multiple Displacement Amplification (MDA) (Dean et al., 2001, 2002; Lasken and Egholm, 2003). Another popular amplification method is MALBAC, which causes its type of amplification artifact (Lu et al., 2012; Zong et al., 2012).

Multiple Displacement Amplification (MDA) is the preferred amplification method for single-cell sequencing, since it is an isothermal (without thermo cycling) process as opposed to PCR (Illumina, 2013, 2014). Compared to PCR-based amplification methods, it produces less amplification coverage bias and error (Tindall and Kunkel, 1988; Esteban et al., 1993; Pinard et al., 2006).

Recently, a new whole-genome amplification method has been demonstrated on individual human cells, which is called Multiple Annealing and Looping Based Amplification Cycles (MALBAC) (Lu et al., 2012; Zong et al., 2012). MALBAC coverage of the human genome has less bias than that of MDA. Nevertheless, amplification bias is still a challenge despite the improvements achieved by MALBAC (Daley and Smith, 2014). Furthermore, sensitivity of MALBAC to background noise makes it not suitable for many applications, such as *de novo* assembly (de Bourcy et al., 2014).

Although single-cell sequencing methods have passed important milestones, such as capturing ≥90% of genes in a prokaryotic cell (Chitsaz et al., 2011) or finishing and closing the genome of a prokaryote using MDA (Woyke et al., 2010), the quality and reliability of genome assemblies from single cells lag behind those of sequencing methods from multi cells due to a bias arising from MDA. The main factors that affect quality are uneven coverage depth and the absence of scattered chunks of the genome in the final collection of reads. There is no known deterministic pattern for the preferred amplified regions, and they are currently treated as the result of a random process. Also, the outcome of MDA is widely variable ranging from total loss of the sample and any information therein to nearly complete reconstruction of the genome. In this sense, an MDA-based single-cell sequencing experiment is currently a gamble that can potentially lead to the loss of the sample and sequencing expenses.

The uneven depth of coverage of a single-cell data set makes the result of *de novo* assembly with uniform sequencing depth assumption inaccurate (Rodrigue et al., 2009; Woyke et al., 2009). This makes the challenges of single-cell sequencing more computational than experimental (Rodrigue et al., 2009). A novel computational solution proposed by Chitsaz et al. (2011) overcomes some of the complications caused by uneven depth of coverage. That method is implemented into a tool called Velvet-SC and adapted by other subsequent single-cell assembly tools, such as SPAdes (Bankevich et al., 2012) and IDBA-UD (Peng et al., 2012), which introduce further advanced algorithmic features and outperform Velvet-SC.

No matter how sophisticated the algorithmic features of an assembler, there is no way to assemble those regions of the genome that are not amplified enough to be captured in sequencing. Chitsaz et al. (2011) called those absent parts of the genome *blackout regions*. We propose an elegant solution to retrieve those blackout regions using the information vested in other single-cell data sets. Coverage data of identical DNA molecules suggest that the MDA process has a strong random component to the extent that it is likely that the blackout regions in one reaction are fully covered in another one. We introduce a coassembly strategy, which can fill the blackout regions in a data set by using the information in another coassembled data set using the idea of colored de Bruijn graph (Iqbal et al., 2012).

Colored de Bruijn graph was initially introduced for structural variation detection. We modified and implemented the algorithm for single-cell coassembly. Furthermore, our algorithm modifies the iterative k assembly algorithm, which is implemented by SPAdes (Bankevich et al., 2012) and IDBA-UD (Peng et al., 2012), and adapts it to the colored graph (Shariat Razavi et al., 2014). It has been shown that the weakness of the coassembly is related to breaking contigs due to various colored branches (Movahedi et al., 2012). Iterative assembly with variable k overcomes that contiguity weakness.

We demonstrate in this work how to hedge against the risk of poor assembly results through sequencing and coassembly of few single cells. Our method replaces a single-cell deep sequencing experiment with multiple single-cell shallow sequencing experiments, allowing for the simultaneous acquisition of supposedly synergistic information about multiple single cells.

2. MATERIALS AND METHODS

2.1. Media and Cultivation of the Methanogenic Alkane-Degrading Community

The microbial community was enriched from sediment from a hydrocarbon-contaminated ditch in Bremen, Germany (Zengler et al., 1999). The consortium was propagated in the laboratory in anoxic medium containing 0.3 g NH_4Cl, 0.5 g $MgSO_4 \cdot 7H_2O$, 2.5 g $NaHCO_3$, 0.5 g K_2HPO_4, 0.05 g KBr, 0.02 g H_3BO_3, 0.02 g KI, 0.003 g $Na_2WO_2 \cdot 2H_2O$, 0.002 g $NiCl_2 \cdot 6H_2O$, trace elements, and trace minerals as previously described

(Zengler et al., 1999). The medium was sparged with a mixture of N_2/CO_2 (80:20 v/v), and the pH was adjusted to 7.0. After autoclaving, anoxic $CaCl_2$ (final concentration 0.25 g/L) and filter-sterilized vitamin solution (Zengler et al., 1999) were added. Cells were supplemented with anoxic hexadecane as previously described (Embree et al., 2013). Bottles were degassed as necessary to relieve over-pressurization.

2.2. Single-Cell Sorting, MDA, and Genomes Sequencing

Individual cells from the alkane-degrading consortium were obtained by staining (SYTO-9 DNA stain) and sorting of single cells by FACS (Embree et al., 2013). Single cells were lysed as previously described, and the genomic DNA of individual cells was amplified using whole-genome multiple displacement amplification (MDA) (Swan et al., 2011). Amplified genomic DNA was screened for Smithella-specific 16S rDNA gene sequences. Six amplified Smithella genomes were selected for Next-Generation Sequencing. The MDA amplified genomes were prepared for Illumina sequencing using the Nextera kit, version 1 (Illumina) using the Nextera protocol (ver. June 2010) and high molecular weight buffer. Libraries with an average insert size of 400 bp were created for these samples and sequenced using an Illumina Genome Analyzer IIx. The 34-bp paired-end reads were generated for K05 (20.9 million reads), C04 (23.3 million reads), F02 (26.9 million reads), and A17 (22.2 million reads). The 58-bp single-end reads were generated for MEB10 (41.3 million reads), MEK03 (54.1 million reads), and MEL13 (18.0 million reads). The 36-bp paired-end reads were generated for F16 (11.0 million reads), K04 (27.2 million reads), and K19 (22.9 million reads).

2.3. Assembly of Single-Cell Genomes

Assemblies were obtained using HyDA version 1.1.1, SPAdes version 2.4.0, and IDBA-UD version 1.0.9. SPAdes and IDBA-UD were run with the default parameters in the single-end mode. The scripts to generate all of the assemblies are provided in Supplementary Material. The length of k-mers in the de Bruijn graph was 25, and the coverage cut off to trim erroneous branches in the graph was selected to be 100. The contigs were then annotated using RAST (Aziz et al., 2008), and the resulting annotation was used to generate a draft metabolic reconstruction using Model SEED (Henry et al., 2010). The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AWGX00000000. The version described here is version AWGX01000000.

3. RESULTS

3.1. Colored de Bruijn Graph

Algorithmic paradigms for fragment assembly, such as overlap-layout-consensus and de Bruijn graph, depend on the characteristics of sequencing reads, particularly read length and error profile. Overlap-layout-consensus is a paradigm that is usually applied to assembly projects using long reads, and the de Bruijn graph is another widely adopted paradigm that is used for short-read data sets (Compeau et al., 2011). A number of consecutive *k*-mers (a sequence of length *k* nucleotides) replace each read in the de Bruijn graph paradigm. Each *k*-mer is represented by a unique vertex. An edge is present between two vertices if there is a read in which the two respective *k*-mers are consecutively overlapping. When there are at least *k* consecutive common bases, reads share a vertex (respectively, $k + 1$ common bases for an edge) along which contigs are efficiently constructed.

Colored de Bruijn graph is a method proposed for coassembly of multiple short-read data sets (Iqbal et al., 2012). It is an extension of the classical approach by superimposing different uniquely colored input data sets on a single de Bruijn graph. Each vertex, which is a representation of a *k*-mer, accompanies an array of colored multiplicities. In this way, input data sets are virtually combined while they are almost fully tracked, enabling separation after assembly. Iqbal et al. (2012) proposed the colored de Bruijn graph in Cortex for variant calling and genotyping, whereas our tool Hybrid *De novo* Assembler (HyDA) (Movahedi et al., 2012) is developed for *de novo* assembly of short-read sequences with non-uniform coverage, which is a dominant phenomenon in MDA-based single-cell sequencing (Chitsaz et al., 2011). To fill the gaps and compare colors, contigs in HyDA are constructed in a color-oblivious manner, solely based on the branching structure of the graph. First, this method rescues a poorly covered region of the genome in one data set when it is well covered in at least one of the other input data sets (**Figure 1A**; **Table 1**). Second, it allows comparison of colored assemblies by

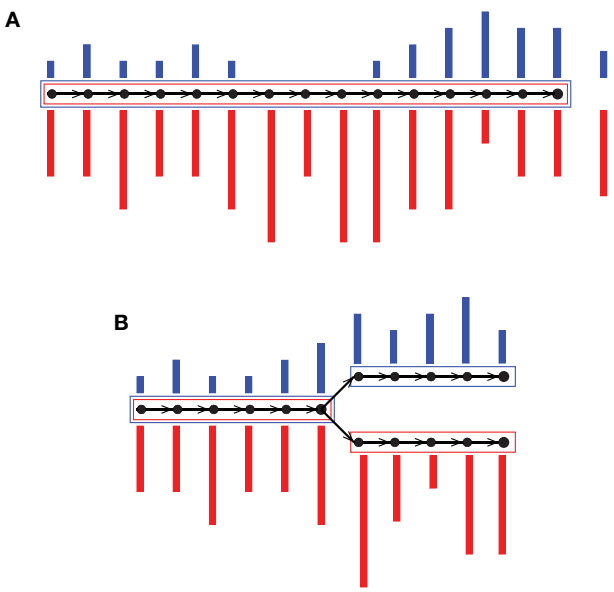


FIGURE 1 | Two sample colored de Bruijn graphs with colors red and blue. Nodes are *k*-mers and edges represent $k + 1$-mers. A colored bar shows multiplicity of the *k*-mer in the corresponding colored input data set. Each box is an output contig, and the color of a box shows non-zero colored average coverage, which is shown on the right hand side of the contig in **(A)**. Our coassembly algorithm **(A)** rescues a poorly covered region of the genome in one color when it is well covered in the other, and **(B)** allows pairwise comparison of colored assemblies through revealing all of their shared and exclusive pieces of sequence.

revealing all shared and exclusive pieces of sequence not shorter than *k* (**Figure 1B**; **Table 2**).

3.2. Coverage Characteristics of Single-Cell Read Data Sets

Genomes amplified from single cells exhibit highly non-uniform genome coverage and multiple gaps, which are called blackout regions (Chitsaz et al., 2011). For the evaluation of such coverage characteristics in this study, we used amplified DNA originating from two single *Escherichia coli* cells as well as from one single *Staphylococcus aureus* cell (Chitsaz et al., 2011). Although these amplified DNAs were quality checked for preselected genomic loci using quantitative PCR (Rodrigue et al., 2009), they still did not cover the entire genome (Table S1 in Supplementary Material; **Figure 2**). One single *E. coli* cell was sequenced in four technical replicate lanes (1–4), and the other was sequenced in three technical replicate lanes (6–8) each with a sequencing depth of 600 per lane. The single *S. aureus* cell was sequenced in two technical replicate lanes each with a sequencing depth of 1,800. All nine lanes were sequenced on Illumina GAIIx platform in paired 2–100 bps read mode.

The coverage bias in technical replicates is almost identical, which suggests that the vast majority of bias is caused by MDA. The coverage bias, particularly of the blackout regions, does not always occur at the same genomic loci for different cells of the same genome (Chitsaz et al., 2011). Blackout regions in *E. coli* lanes 1 and 6 sequenced from two independently amplified single cells make up 1.8 and 0.1% of the genome, respectively, but there are no common blackout regions between these two data sets (Table S1 in Supplementary Material). This means that combining the two data sets could fill all gaps and yield a complete genome, which is the property that HyDA exploits with colored coassembly.

3.3. Colored Coassembly of *E. coli* and *S. aureus* Mitigates the Effect of Dropout Regions due to Amplification Bias

Single-cell read data sets have highly variable coverage (Raghunathan et al., 2005; Rodrigue et al., 2009) (Table S1 in Supplementary Material; **Figure 2**), which poses serious challenges for downstream applications such as *de novo* assembly. A number of single-cell assemblers, including EULER + Velvet-SC (Chitsaz et al., 2011), SPAdes (Bankevich et al., 2012), and IDBA-UD (Peng et al., 2012), have been developed to mitigate the adverse effects of non-uniform coverage and maximize the transfer of sequencing information into the final assembly. These efforts have been successful, and the existing single-cell assemblers are able to extract nearly all of the information contained in the input data set. However, the vast majority of single-cell data sets do not encompass the entire genome. We report that combining multiple data sets from the same or closely related species significantly improves the final assembly by filling genome gaps (Table S1 in Supplementary Material). The challenge presented by this method is the subsequent deconvolution of single-cell genomes to avoid chimeric assemblies.

TABLE 1 | The GAGE (Salzberg et al., 2012) statistics of HyDA assemblies for the six scenarios in Figure S1 in Supplementary Material.

	Lane 1 Single color	Lane 6 Single color	Identical cells Mixed	Identical cells Colored	Non-identical cells Mixed	Non-identical cells Colored
Assembly size	4,532,221	4,642,640	5,262,077	5,204,061	8,273,488	5,212,674
Missing *E. coli* reference bases (%)	314,009 (6.77%)	123,687 (2.67%)	1,555 (0.03%)	2,023 (0.04%)	1,289 (0.03%)	2,136 (0.05%)
Extra bases (%)	280,998 (6.20%)	198,072 (4.27%)	653,307 (12.42%)	584,534 (11.23%)	3,661,052 (44.25%)	597,088 (11.45%)
SNPs	60	19	11	3	5	5
Indels < 5 bp	6	4	10	6	8	6
Indels ≥ 5 bp	13	14	6	5	4	4
Inversions	0	0	0	0	0	0
Relocations	12	11	2	3	2	3
NG50	42,257	54,422	41,964	34,752	54,505	37,794
Corrected NG50	39,975	44,872	39,334	32,876	39,334	36,868

GAGE (Salzberg et al., 2012) was based on MUMmer 3.23 aligner (Kurtz et al., 2004).

TABLE 2 | Pairwise relationships between three coassembled data sets, *E. coli* lanes 1 and 6 and *S. aureus* lane 7, in a coassembly of *E. coli* lanes 1–4, 6–8, and *S. aureus* lanes 7 and 8.

Pair of data sets	Pair 1		Pair 2		Pair 3	
	E. coli lane 1		*E. coli* lane 6		*S. aureus*	*E. coli* lane 1
Total (bps)	5,228,480		5,240,302		3,366,622	5,228,480
Shared (bps)	5,210,548		335,648		336,184	
Exclusive (bps)	179,32	29,754	4,904,654	3,030,974	3,030,438	4,892,296
Exclusivity ratio[a]	0.003	0.005	0.9359	0.9003	0.9001	0.9357

Total is the total size of those contigs that have non-zero coverage in the corresponding color. Shared is the size of those contigs that have non-zero coverage in both colors. Exclusive is the size of those contigs that have non-zero coverage in the corresponding color and zero coverage in the other color in the pair.

[a]*Exclusivity ratio = exclusive/total.*

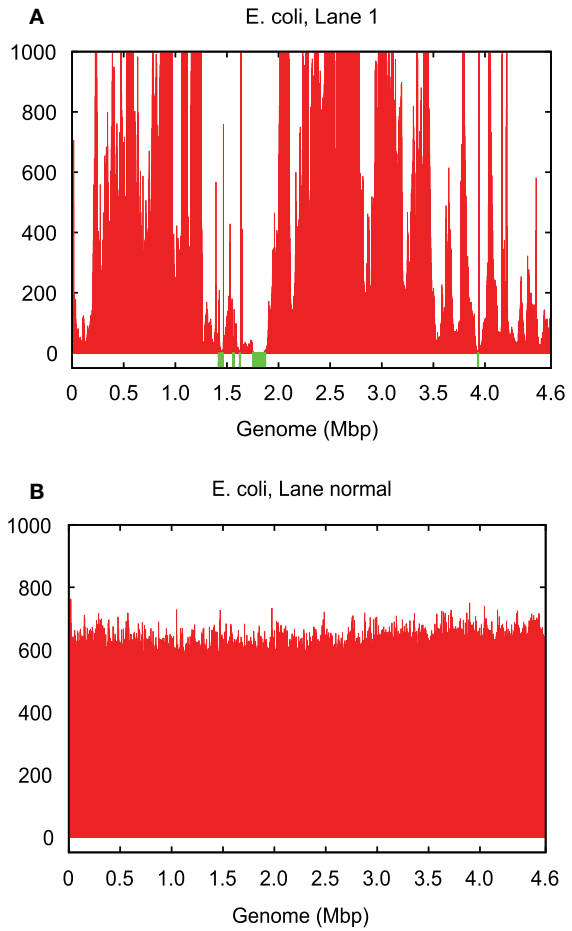


FIGURE 2 | Genome coverage in (A) single-cell *E. coli* lane 1 vs. (B) normal multicell *E. coli*. Both have an average coverage of ~600×.

The ideal solution involves the coassembly of multiple data sets without explicitly mixing sequencing reads such that individual assemblies can benefit from the synergy without suffering from chimerism. We propose and implement this solution using the colored de Bruijn graph in HyDA.

We report in **Table 1** the coassembly results for six distinct scenarios (Figure S1 in Supplementary Material), each consisting of a combination of the input read data sets: (i) single-cell assembly of *E. coli* lane 1; (ii) single-cell assembly of *E. coli* lane 6; (iii) mixed monochromatic assembly of *E. coli* lanes 1–4 and 6–8, technical replicates of two biologically replicate single cells; (iv) multichromatic coassembly of *E. coli* lanes 1–4 and 6–8; (v) mixed monochromatic assembly of non-identical cells: *E. coli* lanes 1–4 and 6–8, and *S. aureus* lanes 7 and 8; and (vi) multichromatic coassembly of non-identical cells: *E. coli* lanes 1–4 and 6–8, and *S. aureus* lanes 7 and 8, each assigned a unique color. GAGE, a standard genome evaluation tool, which reports the size statistics and number of substitution, indel, and chimeric errors of an assembly, was used to evaluate our assemblies (Salzberg et al., 2012). In all six scenarios, GAGE results (**Table 1**) comparing the assembly of color 0 with the *E. coli* reference genome are reported. Color 0 corresponds to *E. coli* lane 1 in (i), (iv), and (vi); *E. coli* lane 6 in (ii); and the mixture in (iii) and (v) (Figure S1 in Supplementary Material).

TABLE 3 | Evaluation results obtained from GAGE (Salzberg et al., 2012) for assembly of *E. coli* lanes 1 and 6 using E + V-SC (Chitsaz et al., 2011), SPAdes (Bankevich et al., 2012), and IDBA-UD (Peng et al., 2012).

Tool	Missing ref. bases (%)	
	Lane 1	Lane 6
E + V-SC	281,060 (6.06%)	109,994 (2.37%)
SPAdes	128,600 (2.77%)	15,831 (0.34%)
IDBA-UD	145,536 (3.14%)	28,583 (0.62%)

While the state-of-the-art individual single-cell *E. coli* assemblies by SPAdes (SPAdes outperforms IDBA-UD and Euler + Velvet-SC in this case) miss 128,600 (2.77%) and 15,831 (0.34%) base pairs of the reference genome in the two different single cells (**Table 3**), our coassembly misses only 2,023 (0.04%) of the genome (**Table 1**), an improvement of 126,577 (2.72%) base pairs of the *E. coli* cell 1. Our coassembly of the two single *E. coli* cells and one *S. aureus* cell misses only 2,136 (0.05%) of the genome. The coassembly algorithm in this work, without any error correction, *k*-mer incrementation, or scaffolding, increases the total assembly size for both *E. coli* lanes 1 and 6 using only the synergy in the input data sets. Our exclusivity ratio (defined below) obtained from the coassembly results completely differentiates *E. coli* and *S. aureus* data sets (**Table 2**).

3.4. Quantification of Similarities and Differences between Colors

Input data sets can be clustered based on the similarity between their assemblies. For a pair of colors *i* and *j*, contigs belonging to both colors are considered shared and contigs belonging to color *i* but not to color *j* are considered exclusive of color *i* with respect to color *j*. We define the exclusivity ratio of color *i* with respect to color *j* as the ratio of the size of exclusive color *i* contigs to the total assembly size of color *i*. The exclusivity ratio for *E. coli* lane 1-lane 6 (Pair 1 in **Table 2**) is less than 0.5%, while that ratio for *E. coli* and *S. aureus* in the two other pairs (Pair 2 and 3 in **Table 2**) is greater than 90%. This large difference in exclusivity ratio between Pair 1 and Pairs 2 and 3 is expected in this case, as *E. coli* and *S. aureus* are phylogenetically divergent species belonging to different phyla.

3.5. *De Novo* Single-Cell Coassembly of Members of an Alkane-Degrading Methanogenic Consortium

The genomes of 10 cells from three dominant but uncultured bacterial members of a methanogenic consortium (Zengler et al., 1999; Embree et al., 2013), belonging to the families *Syntrophacea* and *Anaerolineaceae* were sequenced from their amplified single-cell whole DNAs: six cells belonging to *Smithella*, two cells belonging to *Anaerolinea*, and two cells belonging to *Syntrophus*. Single cells were isolated from the consortium by fluorescence-activated

cell sorting, and the genomes of individual cells were amplified using MDA. MDA products were sequenced using an Illumina GAIIx with 34, 36, or 58 base pair reads. In total, 10 data sets, one per cell, were obtained. The 10 data sets were coassembled with HyDA in a *ten-color* setup, and to exhibit the advantage of the coassembly method, each data set was assembled individually by HyDA. Individual assemblies created by SPAdes and IDBA-UD were used as comparison. The QUAST (Gurevich et al., 2013) length statistics of the resulting assemblies (≥100 bp contigs) are compared in **Table 4** and Figures S2–S11 in Supplementary Material. The comparison between individual assembly and coassembly by HyDA demonstrates that coassembly rescues on average 101.4% more total base pairs for all 10 cells (Table S2 in Supplementary Material). Although HyDA does not use advanced assembly features such as variable *k*-mer sizes and paired read information, it can assemble 3.6–54% more total base pairs than both SPAdes and IDBA-UD do in all cells except two cases: *Anaerolinea* F02 and *Smithella* MEK03 (**Table 4**; Table S2 in Supplementary Material). When all contigs are considered, HyDA coassemblies of *Anaerolinea* F02 and *Smithella* MEK03 are 11% smaller and 41% larger than their SPAdes counterparts, respectively. *Smithella* MEK03 input reads are longer (58 bp) than the reads in some of the other data sets; therefore, the *Smithella* MEK03 assembly contains many short contigs and suffers because of the small *k*-mer size (k = 25) dictated by the shorter reads.

3.6. Exclusivity Analysis of Ten Assemblies from Single Uncultured Bacterial Cells

Exclusivity analysis revealed that the six Smithella cells clustered into a consistent group as their exclusivity ratios with respect to the two *Anaerolinea* and two *Syntrophus* cells are almost identical (**Table 5**). It is important to note that *Anaerolinea* A17 and *Syntrophus* C04 assemblies are relatively short, meaning the exclusivity ratios must be interpreted with caution. Although *Syntrophus* K05s exclusivity signature with respect to the six *Smithella* cells is indistinguishable from the six *Smithella* signatures with respect to themselves, the exclusivity ratios of *Syntrophus* K05 with respect to the two *Anaerolinea* cells and *Syntrophus* C04 differentiate *Syntrophus* K05 from the six *Smithella* cells. Slight differences between the *Syntrophus* C04 and K05 exclusivity signatures are not surprising because of the existence of potential intraspecies variations.

TABLE 4 | Quast (Gurevich et al., 2013) analysis of 10 cells from *Anaerolinea*, *Smithella*, and *Syntrophus* single-cell data sets assembled with HyDA (individual assembly), HyDA (10-color coassembly), SPAdes, and IDBA-UD.

		Anaerolinea		*Smithella*						*Syntrophus*	
		A17	**F02**	**F16**	**K04**	**K19**	**MEB10**	**MEK03**	**MEL13**	**C04**	**K05**
HyDA	Total	54,237	1,278,742	604,769	449,148	371,311	1,182,622	1,666,233	1,150681	252,402	502,469
	N50	**2,935**	8,461	**8,303**	**9,959**	5,416	5,718	6,167	7,315	**5,578**	4,963
HyDA-Cl	Total	**260,386**	1,352,341	**1,323,536**	**720,188**	**840,236**	**1,569,709**	1,945,701	**1,590,259**	**465,091**	**1,265,548**
	N50	850	8,201	6,088	5,239	**7,295**	5,887	5,952	6,977	1,928	3,782
SPAdes	Total	169,413	**1,698,195**	982,263	618,500	653,866	1,514,813	**1,960,722**	1,415,399	390,923	869,586
	N50	1,187	5,944	5,366	9,332	3,834	**8,861**	**11,372**	**10,475**	4,234	3,128
IDBA-UD	Total	144,512	1,441,353	927,009	56,6327	613,399	1,327,742	1,746,656	1,351,465	318,914	804,313
	N50	2,894	**8,756**	3,163	**3,178**	5,751	6,851	8,209	1,0253	4,706	**5,618**

All statistics are based on contigs of size ≥100 bp. Only those HyDA contigs that have a coverage of at least 1 in the corresponding color are considered. Coverage cutoff was chosen to be 24 for all HyDA assemblies (–c = 24). Total is the total assembly size and N50 is the assembly N50 (the size of the contig, the contigs larger than which cover half of the assembly size). Best result is in bold face.

TABLE 5 | The exclusivity ratio (%) of row with respect to column for the 10 cells from *Anaerolinea*, *Smithella*, and *Syntrophus* single-cell data sets coassembled using 10 colors with Squeezambler (Taghavi et al., 2013), a tool in the HyDA package.

		Anaerolinea		*Smithella*						*Syntrophus*	
		A17	**F02**	**F16**	**K04**	**K19**	**MEB10**	**MEK03**	**MEL13**	**C04**	**K05**
Anaerolinea	A17	0	24	87	95	96	80	82	86	22	19
	F02	77	0	96	98	99	71	68	72	12	5
Smithella	F16	96	96	0	73	73	37	22	38	96	55
	K04	97	97	49	0	67	42	25	45	97	73
	K19	98	98	54	68	0	35	32	32	98	55
	MEB10	96	96	74	48	69	0	24	39	95	57
	MEK03	97	97	73	54	74	38	0	37	96	58
	MEL13	97	97	76	51	68	39	22	0	97	59
Syntrophus	C04	44	39	89	96	97	85	86	90	0	64
	K05	77	75	54	76	75	45	41	49	73	0

Only the contigs of coverage at least 1 in the corresponding color are considered. Coverage cutoff was chosen to be 24 for all HyDA assemblies (–c = 24).

3.7. Annotation of the *Anaerolinea*, *Smithella*, and *Syntrophus* Assemblies

To assess the quality of coassemblies with HyDA, IDBA-UD, and SPAdes, we used the RAST server to predict the coding sequences and subsystems present in each assembly. The HyDA assemblies are superior to those of SPAdes and IDBA-UD in terms of the number of coding sequences and captured subsystems for one *Anaerolinea*, four *Smithella*, and both *Syntrophus* assemblies (**Table 6**). For *Smithella* MEB10 and MEK03, the HyDA assembly closely follows the SPAdes assembly, which provides the largest annotation (**Table 6**). For *Smithella* F16 and *Syntrophus* K05, HyDA assemblies contain significantly more coding sequences (33 and 39%, respectively) and cover more subsystems (29 and 57%, respectively) in comparison to the best of SPAdes and IDBA-UD assemblies.

To confirm the accuracy of the assemblies, the closest related species to each assembly was computed by the RAST server. For the HyDA, SPAdes, and IDBA-UD *Anaerolinea* F02 assemblies, the closest species was *Anaerolinea thermophila* UNI-1 (GenomeID 926569.3) (no closest genomes data found for *Anaerolinea* A17 by the RAST server). For the HyDA, SPAdes, and IDBA-UD *Smithella* and *Syntrophus* assemblies, the closest species is *Syntrophus aciditrophicus SB* (GenomeIDs 56780.10 and 56780.15). Note that *Syntrophus aciditrophicus SB* is the closest finished genome to the *Smithella* family. This verifies that coassembly does not create chimeric assemblies; otherwise, we would see *Syntrophus aciditrophicus SB* among close neighbors of the *Anaerolinea* assemblies and/or *Anaerolinea thermophila UNI-1* among close neighbors of the *Smithella* and *Syntrophus* assemblies by HyDA.

3.8. Metabolic Reconstruction of *Anaerolinea, Smithella*, and *Syntrophus*

Assembly and subsequent annotation of these genomes enables the elucidation of the functional roles of individual, unculturable constituents within the community. *Anaerolinea*, *Syntrophus*, and *Smithella* each represent genera with very few cultured members and only two sequenced genomes – *Anaerolinea thermophila* (no genome paper) and *Syntrophus aciditrophicus* (McInerney et al., 2007) are the only available sequenced genomes from these genera to date. The only member of *Smithella* that has been isolated, *Smithella propionica* (Liu et al., 1999), has not been sequenced yet. In addition to understanding the genetic basis for the unique metabolic capability of this microbial community, the genomes of these particular organisms present an opportunity to explore the breadth of genetic diversity in these elusive genera. Using the advanced genome assembly algorithm, we recently identified the key genes involved in anaerobic metabolism of hexadecane and long-chain fatty acids, such as palmitate, octadecanoate, and tetradecanoate, in *Smithella* (Embree et al., 2013). Based on sequence homology, *Syntrophus* is closely related to Smithella, but we cannot determine if it is also actively degrading hexadecane at this point in time. Only two species of *Anaerolinea* have been isolated and characterized thus far. These species, both isolated from anaerobic sludge reactors, form long, multicellular filaments and are strictly anaerobic (Sekiguchi et al., 2003; Yamada et al., 2006). Each species is capable of growing on a large number of carbon sources, and both isolates produce acetate, lactate, and hydrogen as the main end products of fermentation. Comparison of the Anaerolinea sp. genome derived from single-cell sequencing with the genome of *Anaerolinea thermophila* UN-1 revealed many similarities in potential metabolic capability. The *Anaerolinea* genome obtained from a single cell contains genes for the utilization of galactose and xylose, consistent with a previous physiological characterization of *A. thermophila* (Sekiguchi et al., 2003). Additionally, the single-cell *Anaerolinea sp*. genome encoded for several transporters and genes related to trehalose biosynthesis, suggesting extended metabolic capabilities of this strain. Furthermore, the genome has an extracellular deoxyribonuclease, an enzyme required for catabolism of external DNA, hinting at the strains ability to scavenge deoxyribonucleosides.

4. DISCUSSION

We demonstrated the power of genome coassembly of multiple single-cell data sets through significant improvement of the assembly quality in terms of predicted functional elements and length statistics. Coassemblies without any effort to scaffold or close gaps contain significantly more protein coding genes, subsystems, base pairs, and generally longer contigs compared

TABLE 6 | Summary of coding sequences and subsystems predicted by the RAST server (Aziz et al., 2008) for HyDA, IDBA-UD, and SPAdes assemblies of the three alkane-degrading bacterial genomes.

		HyDA-colored		Spades		IDBA-UD	
		Coding sequence	Subsystem	Coding sequence	Subsystem	Coding sequence	Subsystem
Anaerolinea	A17	**212**	8	146	**9**	132	7
	F02	1,283	122	**1,653**	**153**	1,375	121
	F16	**1,197**	**117**	899	91	866	89
	K04	**659**	**89**	559	75	508	66
Smithella	K19	**757**	**82**	581	54	572	57
	MEB10	1,491	151	**1,504**	**156**	1,297	138
	MEK03	1,856	180	**1,955**	**200**	1,178	170
	MEL13	**1,535**	**165**	1,435	154	1,384	148
Syntrophus	C04	**416**	48	375	**49**	320	36
	K05	**1,216**	**121**	873	68	854	77

Best result is in bold face.

to individual assemblies by the same algorithm as well as the state-of-the-art single-cell assemblers (SPAdes and IDBA-UD). The new algorithm is also able to avoid chimeric assemblies by detecting and separating shared and exclusive pieces of sequence for input data sets. This suggests that in lieu of single-cell assembly, which can lead to failure and loss of the sample or significantly increase sequencing expenses, the coassembly method can hedge against that risk. Our single-cell coassembler HyDA proved the usefulness of the coassembly concept and permitted the study of three bacteria. The improved assembly gave insight into the metabolic capability of these microorganisms, thereby proving a new tool for the study of uncultured microorganisms. Thus, the coassembler can readily be applied to study genomic content and the metabolic capability of microorganisms, and increase our knowledge of the function of cells related to environmental processes as well as human health and disease. The colored de Bruijn graph uses a single k-mer size for all input data sets, which has to be chosen based on the minimum read length across all data sets. For instance, Smithella MEK03 input reads are longer (58 bp) than the reads in some of the other data sets, while the Smithella MEK03 assembly contains many short contigs because of the small k-mer size ($k = 25$) dictated by the shorter reads. This minor disadvantage can be remedied by using advanced assembly features such as variable k-mer size, alignment of reads back to the graph and threading, and utilization of paired-end information.

AUTHOR CONTRIBUTIONS

NM carried out genome assembly and evaluation, helped with metabolic reconstruction analysis, participated in development of HyDA, and drafted the manuscript. ME and HN participated in acquisition of the alkane-degrading consortium genomic data and drafted the manuscript. KZ participated in the project conception, participated in acquisition of the alkane-degrading consortium genomic data, and drafted the manuscript. HC participated in the project conception, developed HyDA, carried out interpretation of results, and drafted the manuscript.

FUNDING

Funding for this work was partially provided by NSF DBI-1262565 grant to HC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fbioe.2016.00042

REFERENCES

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi:10.1186/1471-2164-9-75

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi:10.1038/ismej.2012.8

Chitsaz, H., Yee-Greenbaum, J. L., Tesler, G., Lombardo, M.-J., Dupont, C. L., Badger, J. H., et al. (2011). Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* 29, 915–921. doi:10.1038/nbt.1966

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991. doi:10.1038/nbt.2023

Daley, T., and Smith, A. D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics* 30, 3159–3165. doi:10.1093/bioinformatics/btu540

de Bourcy, C. F., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., and Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* 9:e105585. doi:10.1371/journal.pone.0105585

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261–5266. doi:10.1073/pnas.082089499

Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* 11, 1095–1099. doi:10.1101/gr.180501

Embree, M., Nagarajan, H., Movahedi, N., Chitsaz, H., and Zengler, K. (2013). Single-cell genome and metatranscriptome sequencing reveal metabolic interactions of an alkane-degrading methanogenic community. *ISME J.* 8, 757–767. doi:10.1038/ismej.2013.187

Esteban, J., Salas, M., and Blanco, L. (1993). Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.* 268, 2719–2726.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi:10.1126/science.1124234

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi:10.1038/nbt.1672

Hosono, S., Faruqi, A. F., Dean, F. B., Du, Y., Sun, Z., Wu, X., et al. (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* 13, 954–964. doi:10.1101/gr.816903

Illumina. (2013). TruSeq nano DNA sample preparation kit@ONLINE. Pub. No. 770-2013-012.

Illumina. (2014). Nextera DNA sample preparation kit.@ONLINE. Pub. No. 770-2011-021.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44, 226–232. doi:10.1038/ng.1028

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi:10.1186/gb-2004-5-6-p12

Lasken, R. S. (2007). Single-cell genomic sequencing using multiple displacement amplification. *Curr. Opin. Microbiol.* 10, 510–516. doi:10.1016/j.mib.2007.08.005

Lasken, R. S., and Egholm, M. (2003). Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol.* 21, 531–535. doi:10.1016/j.tibtech.2003.09.010

Li, J. W.-H., and Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science* 325, 161–165. doi:10.1126/science.1168243

Liu, Y., Balkwill, D. L., Aldrich, H. C., Drake, G. R., and Boone, D. R. (1999). Characterization of the anaerobic propionate-degrading syntrophs *Smithella propionica* gen. nov., sp. nov. and *Syntrophobacter wolinii*. *Int. J. Syst. Bacteriol.* 49, 545–556. doi:10.1099/00207713-49-2-545

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A. R., et al. (2012). Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338, 1627–1630. doi:10.1126/science.1229112

McInerney, M. J., Rohlin, L., Mouttaki, H., Kim, U., Krupp, R. S., Rios-Hernandez, L., et al. (2007). The genome of *Syntrophus aciditrophicus*: life at the

thermodynamic limit of microbial growth. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7600–7605. doi:10.1073/pnas.0610456104

McLean, J. S., Lombardo, M. J., Badger, J. H., Edlund, A., Novotny, M., Yee-Greenbaum, J., et al. (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl. Acad. Sci. U.S.A.* 110, E2390–E2399. doi:10.1073/pnas.1219809110

Movahedi, N. S., Forouzmand, E., and Chitsaz, H. (2012). "De novo co-assembly of bacterial genomes from multiple single cells," in *IEEE Conference on Bioinformatics and Biomedicine*, (Philadelphia, PA: IEEE), 561–565.

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi:10.1093/bioinformatics/bts174

Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., Tartaro, K. R., Plant, R. N., et al. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7:216. doi:10.1186/1471-2164-7-216

Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M., and Lasken, R. S. (2005). Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* 71, 3342–3347. doi:10.1128/AEM.71.6.3342-3347.2005

Rodrigue, S., Malmstrom, R. R., Berlin, A. M., Birren, B. W., Henn, M. R., and Chisholm, S. W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* 4:e6864. doi:10.1371/journal.pone.0006864

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., et al. (2007). The Sorcerer II Global Ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5:e77. doi:10.1371/journal.pbio.0050077

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 22, 557–567. doi:10.1101/gr.131383.111

Sekiguchi, Y., Yamada, T., Hanada, S., Ohashi, A., Harada, H., and Kamagata, Y. (2003). *Anaerolinea thermophila* gen. nov., sp. nov. and *Caldilinea aerophila* gen. nov., sp. nov., novel filamentous thermophiles that represent a previously uncultured lineage of the domain bacteria at the subphylum level. *Int. J. Syst. Evol. Microbiol.* 53, 1843–1851. doi:10.1099/ijs.0.02699-0

Shariat Razavi, S. B., Movahedi Tabrizi, N. S., Chitsaz, H., and Boucher, C. (2014). HyDA-Vista: towards optimal guided selection of *k*-mer size for sequence assembly. *BMC Genomics* 15:S9. doi:10.1186/1471-2164-15-S10-S9

Swan, B. K., Martinez-Garcia, M., Preston, C. M., Sczyrba, A., Woyke, T., Lamy, D., et al. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* 333, 1296–1300. doi:10.1126/science.1203690

Taghavi, Z., Movahedi, N. S., Draghici, S., and Chitsaz, H. (2013). Distilled single-cell genome sequencing and de novo assembly for sparse microbial communities. *Bioinformatics* 29, 2395–2401. doi:10.1093/bioinformatics/btt420

Tindall, K. R., and Kunkel, T. A. (1988). Fidelity of DNA synthesis by the thermus aquaticus DNA polymerase. *Biochemistry* 27, 6008–6013. doi:10.1021/bi00416a027

Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., et al. (2010). One bacterial cell, one complete genome. *PLoS ONE* 5:e10314. doi:10.1371/journal.pone.0010314

Woyke, T., Xie, G., Copeland, A., Gonzalez, J. M., Han, C., Kiss, H., et al. (2009). Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4:e5299. doi:10.1371/journal.pone.0005299

Yamada, T., Sekiguchi, Y., Hanada, S., Imachi, H., Ohashi, A., Harada, H., et al. (2006). *Anaerolinea thermolimosa* sp. nov., *Levilinea saccharolytica* gen. nov., sp. nov. and *Leptolinea tardivitalis* gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes anaerolineae classis nov. and caldilineae classis nov. in the bacterial phylum chloroflexi. *Int. J. Syst. Evol. Microbiol.* 56, 1331–1340. doi:10.1099/ijs.0.64169-0

Zengler, K., Richnow, H. H., Rosselló-Mora, R., Michaelis, W., and Widdel, F. (1999). Methane formation from long-chain alkanes by anaerobic microorganisms. *Nature* 401, 266–269. doi:10.1038/45777

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622–1626. doi:10.1126/science.1229164

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership