Advances in Experimental Medicine and Biology 1129

Yutaka Suzuki *Editor*

Single Molecule and Single Cell Sequencing

Springer

Advances in Experimental Medicine and Biology

Volume 1129

More information about this series at http://www.springer.com/series/5584

Yutaka Suzuki
Editor

Single Molecule and Single Cell Sequencing

Springer

Editor
Yutaka Suzuki
Department of Computational Biology and Medical Sciences
University of Tokyo
Chiba, Chiba, Japan

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-981-13-6036-7 ISBN 978-981-13-6037-4 (eBook)
https://doi.org/10.1007/978-981-13-6037-4

Library of Congress Control Number: 2019934708

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Strategies for Converting RNA to Amplifiable cDNA for Single-Cell RNA Sequencing Methods

Yohei Sasagawa, Tetsutaro Hayashi, and Itoshi Nikaido

Abstract This review describes the features of molecular biology techniques for single-cell RNA sequencing (scRNA-seq), including methods developed in our laboratory. Existing scRNA-seq methods require the conversion of first-strand cDNA to amplifiable cDNA followed by whole-transcript amplification. There are three primary strategies for this conversion: poly-A tagging, template switching, and RNase H-DNA polymerase I-mediated second-strand cDNA synthesis for in vitro transcription. We discuss the merits and limitations of these strategies and describe our Reverse Transcription with Random Displacement Amplification technology that allows for direct first-strand cDNA amplification from RNA without the need for conversion to an amplifiable cDNA. We believe that this review provides all users of single-cell transcriptome technologies with an understanding of the relationship between the quantitative performance of various methods and their molecular features.

Background

Single-cell transcriptome analysis is a useful tool to identify novel cell types and states in a population. In the early stages of single-cell transcriptome analysis, several methods were developed on the microarray platform (Eberwine et al. 2001; Iscove et al. 2002; Tietjen et al. 2003; Kamme et al. 2003; Kurimoto 2006; Subkhankulova and Livesey 2006). Nearly all molecular biology techniques used in these methods—several of which have been reported since 2009 (Tang et al. 2009; Islam et al. 2011, 2013; Hashimshony et al. 2012, 2016; Ramsköld et al. 2012; Sasagawa et al. 2013, 2018; Wu et al. 2013; Picelli et al. 2013; Streets et al. 2014; Jaitin et al. 2014; Soumillon et al. 2014; Fan et al. 2015; Nakamura et al. 2015;

Y. Sasagawa · T. Hayashi · I. Nikaido (✉)
Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research, Wako, Saitama, Japan
e-mail: youhei.sasagawa@riken.jp; tetsutaro.hayashi@riken.jp; itoshi.nikaido@riken.jp

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_1

Matsunaga et al. 2015; Klein et al. 2015; Bose et al. 2015; Macosko et al. 2015; Muraro et al. 2016; Sheng et al. 2017; Yang et al. 2017; Avital et al. 2017; Hochgerner et al. 2017; Zheng et al. 2017; Gierahn et al. 2017; Hashimoto et al. 2017; Herman et al. 2018; Rosenberg et al. 2018; Han et al. 2018; Hayashi et al. 2018)—are applied for single-cell RNA-sequencing (scRNA-seq) (Fig. 1). Cell barcoding is a key technology for improving the throughput of scRNA-seq (Islam et al. 2011), which allows for obtainment of transcriptome data from thousands of single cells in a single experiment (Klein et al. 2015; Macosko et al. 2015; Zheng et al. 2017). Full-length read coverage across all transcripts from single cells has also improved (Ramsköld et al. 2012; Picelli et al. 2013; Sheng et al. 2017), enabling the detection of fusion genes, copy number variations, single nucleotide variations, allele-specific expression, isoforms, and splice variants. These features make scRNA-seq methods invaluable in medical and basic research (Regev et al. 2017). However, these methods have a limited capability to assess the expression of many genes and elucidate their cellular function by these methods.

We recently improved the reaction efficiency of molecular biology techniques for scRNA-seq methods (Sasagawa et al. 2018) and devised a novel reaction (Hayashi et al. 2016) (manuscript in preparation), which increase the sensitivity and quantitative capability of our Quartz-Seq2 and RamDA-seq methods (Sasagawa et al. 2018; Hayashi et al. 2018). Other studies have focused on improving the reaction efficiency of molecular biology techniques for quantitative scRNA-seq (Picelli et al. 2013; Hashimshony et al. 2016; Zajac et al. 2013). Herein, we describe the features of molecular biology techniques for scRNA-seq, including our methods. This review will provide end-users and developers of single-cell transcriptome technologies with an understanding of the relationship between the quantitative performance of various methods and their molecular features.

scRNA-Seq Methods Require cDNA Amplification

Single mammalian cells have a limited amount of total RNA (between 1 and 30 pg). scRNA-seq methods typically target poly-adenylated (polyA) RNA, which accounts for 1–5% of total RNA. However, sequencing platforms require approximately 5 ng of DNA from a library. This requires the conversion of polyA RNA to amplifiable, cDNA followed by whole-transcript amplification. Reverse transcription is performed to convert mRNA to first-strand cDNA. A universal adaptor is then added to the 5′ end. However, the resulting first-strand cDNA cannot be amplified. An additional step is required to convert first-strand cDNA to amplifiable cDNA. Unconverted cDNA cannot be amplified and detected in the sequencer. Low conversion efficiency can prevent the detection of the target gene; improving this process is therefore critical for quantitative scRNA-seq. There are three common strategies for this conversion: poly-A tagging, template-switching, and RNase H-DNA polymerase I-mediated second-strand cDNA synthesis for in vitro transcription (IVT). We describe each of these in strategies in the following paragraphs.

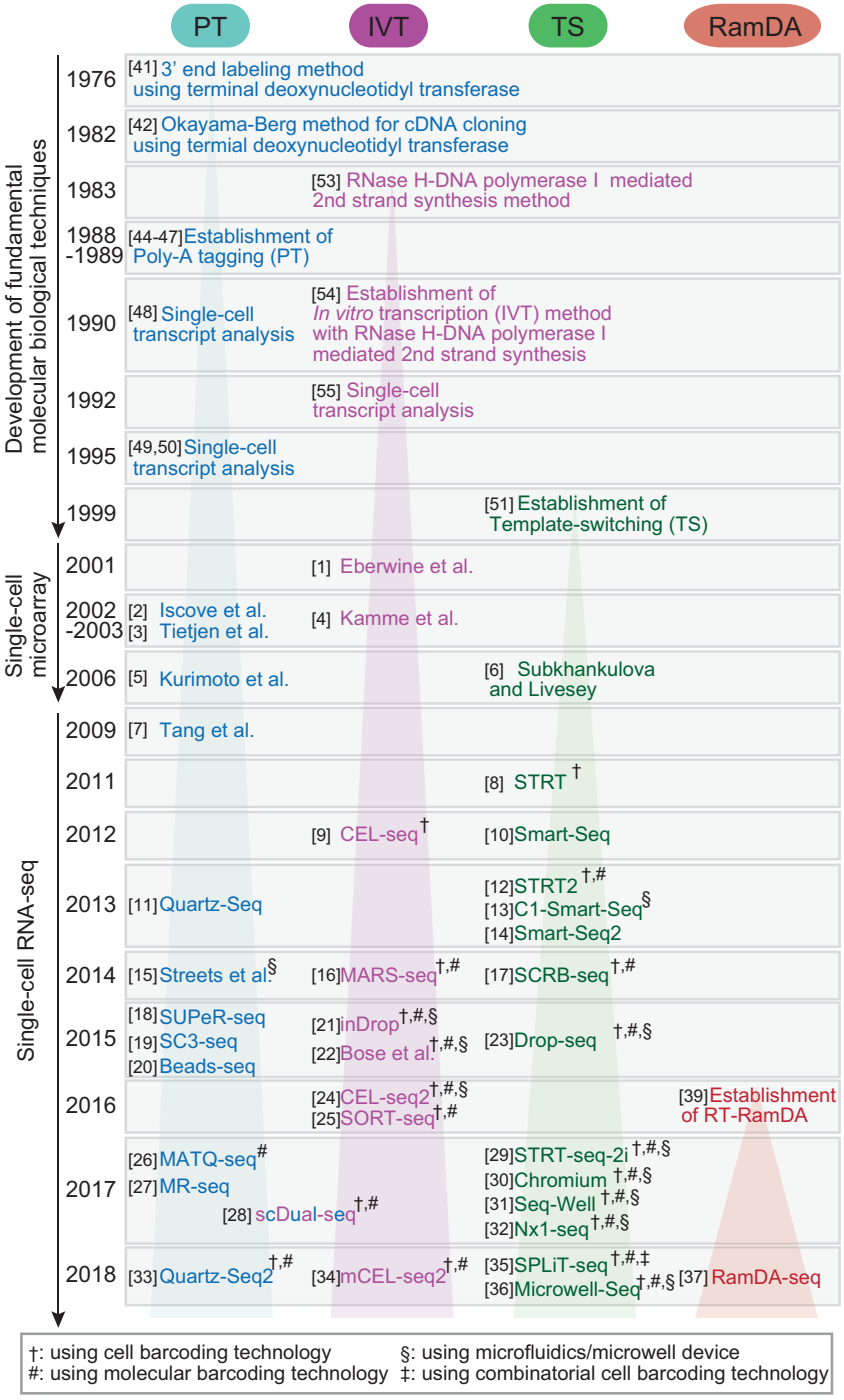

Fig. 1 Chronology of single-cell transcriptome analysis from the standpoint of molecular biology methods. The methods were classified in accordance with four general strategies (IVT, IVT with RNase H-DNA polymerase I-mediated second-strand synthesis; PT, poly-A tagging; RT-RamDA, reverse transcription with random displacement amplification; TS, template-switching)

Poly-A Tagging for scRNA-Seq

Terminal deoxynucleotidyl transferase (TdT) is used to label the 3′ end of DNA with a homopolymeric sequence (Roychoudhury et al. 1976). TdT was first used in cDNA cloning, such as the Okayama-Berg method (Okayama and Berg 1982; Land et al. 1981). Thereafter, the poly-A tagging strategy had been developed to clone full-length cDNA from minute amounts of RNA via a combination of cDNA labeling by TdT and PCR amplification (Frohman et al. 1988; Ohara et al. 1989; Loh et al. 1989; Belyavsky et al. 1989).

The poly-A tagging strategy is categorized as 5′ RACE (Rapid Amplification of cDNA Ends) technique. The strategy came into use for single-cell transcript analysis and single-cell microarray analysis (Iscove et al. 2002; Tietjen et al. 2003; Kurimoto 2006; Brady et al. 1990, 1995; Dulac and Axel 1995). In 2009, Tang et al. developed the first scRNA-seq method, using a poly-A tagging strategy (Tang et al. 2009). The poly-A tail was added to the 3′ end of first-strand cDNA by TdT (Fig. 2). A poly-T primer was then annealed to the poly-A tail, yielding amplifiable cDNA containing binding sites for PCR primers at both ends. In reverse transcription, truncated first-strand cDNA is usually formed as a long transcript. Poly-A tagging targets both full-length and truncated first-strand cDNAs (Sasagawa et al. 2013), and the conversion efficiency is independent of transcript length. Thus, poly-A tagging has advantages for gene detection (Subkhankulova and Livesey 2006). However, no studies have attempted to improve the efficiency of poly-A tagging itself for scRNA-seq.

We recently established the optimal buffer and temperature conditions for poly-A tagging, which improved the efficiency of this process by 3.6-folds (Fig. 3). We also enhanced the efficiency of reverse transcription at low enzyme concentrations. We incorporated these improvements into a scRNA-seq method that we named Quartz-Seq2, which increased the number of detected genes and unique molecular identifiers (Fig. 3). Moreover, the number of detected genes was higher with Quartz-Seq2 than with other methods (Fig. 3). Quartz-Seq2 analysis also yielded more information on functional terms and biological pathways. Thus, improving the efficiency of conversion from first-strand cDNA to amplifiable cDNA is essential for quantitation in scRNA-seq.

Template Switching

Template switching (also known as Switching Mechanism at 5′ End of RNA Template [SMART]) had been developed similar to 5′ RACE (Petalidis 2003; Matz 1999). In reverse transcription, first-strand cDNA is extended from the polyadenylated site. When the 3′ end of first-strand cDNA reaches the 5′ end of total RNA, non-template nucleotides are added to the former by the terminal transferase activity of reverse transcriptase. Deoxycytidine trimer is predicted to account for

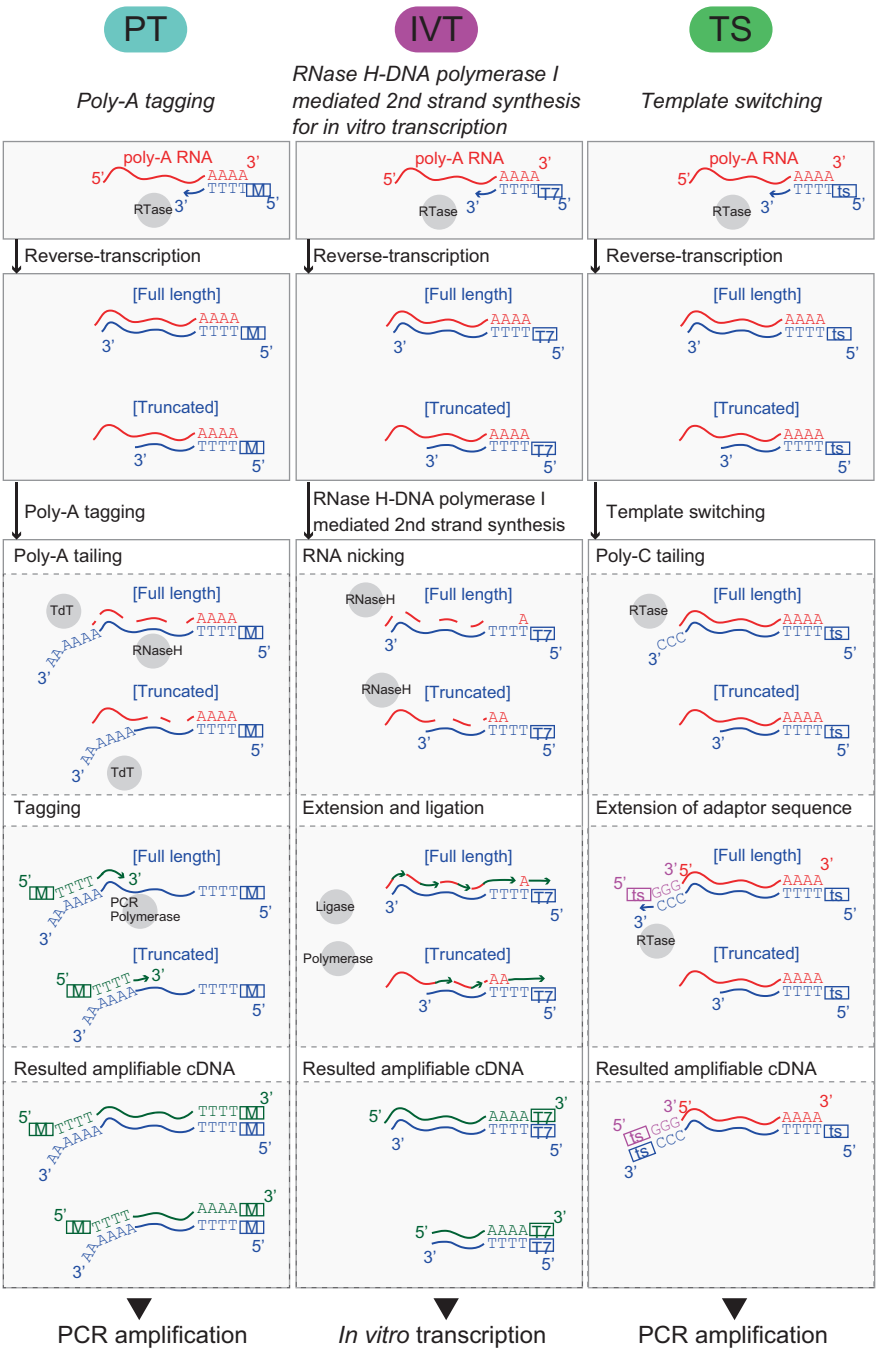

Fig. 2 Schematic representation of strategies for converting first-strand cDNA to amplifiable cDNA. Red, blue, and green lines represent RNA, first-strand cDNA, and second-strand cDNA, respectively. Purple primers represent template-switching oligos. "T7" refers to the promoter sequence of T7 RNA polymerase. Ligase, *Escherichia coli* DNA ligase; PCR polymerase, thermostable DNA polymerase for PCR reaction; polymerase, DNA polymerase I; RNaseH, RNase H enzyme; RTase, reverse transcriptase; TdT, terminal deoxynucleotidyl transferase

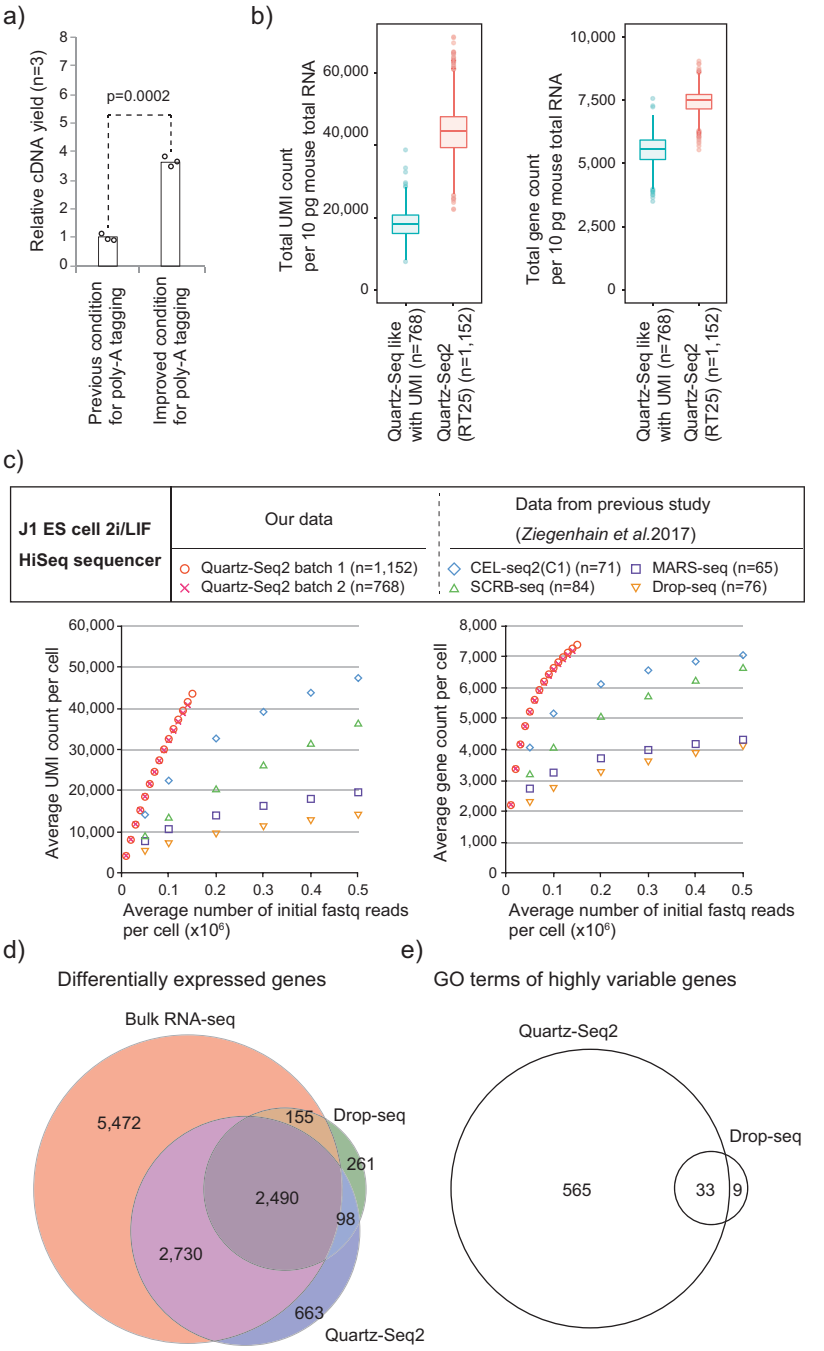

Fig. 3 Improved efficiency of poly-A tagging in Quartz-Seq2. (**a**) Improving the conditions of poly-A tagging increased the amount of amplified cDNA 3.6-fold. (**b** and **c**) The quantitative performance of Quartz-Seq2 is superior to that of previous approaches, including our earlier Quartz-Seq method. (**d** and **e**) Quartz-Seq2 detected more differentially expressed genes and Gene Ontology terms

46% of the nucleotide sequence (Zajac et al. 2013); accordingly, template-switching primers contain three guanines at the 3′ end. By including template-switching primers in the reverse transcription reaction, the enzyme further extends the complementary sequence of the primer at the 3′ end of first-strand cDNA (Fig. 2), which has binding sites for the PCR primer at both ends. The cDNA can then be amplified.

Template-switching, which is globally marketed as the SMART/SMARTer kit, is exclusively applicable to reverse transcription and is widely used in various scRNA-seq methods owing to its relative ease of use (Islam et al. 2011, 2013; Ramsköld et al. 2012; Wu et al. 2013; Picelli et al. 2013; Soumillon et al. 2014; Macosko et al. 2015; Hochgerner et al. 2017; Zheng et al. 2017; Gierahn et al. 2017; Hashimoto et al. 2017; Rosenberg et al. 2018; Han et al. 2018). Template-switching is specific to full-length cDNA and is therefore suitable for detecting full-length or the 5′ end of a transcript (Islam et al. 2011, 2013; Ramsköld et al. 2012; Picelli et al. 2013); its efficiency has been improved with locked nucleic acids (Picelli et al. 2013). Notably, truncated first-strand cDNA cannot be converted into amplifiable cDNA via template switching for detection by a sequencer. Therefore, the completeness of first-strand cDNA for various transcript lengths is very important for this strategy. Moreover, the preference of TdT activity for non-templated nucleotides is non-uniform (Zajac et al. 2013). A uniform non-templated nucleotide preference is necessary for further improvement of the strategy.

RNase H-DNA Polymerase I-Mediated Second-Strand cDNA Synthesis for IVT

RNase H-DNA polymerase I-mediated second-strand cDNA synthesis was developed to achieve cloning of full-length cDNA (Gubler and Hoffman 1983). An IVT strategy for cDNA amplification has been established on the basis of this method, using T7 RNA polymerase (Van Gelder et al. 1990; Eberwine et al. 1992). In this method, RNA-DNA hybrids are synthesized following reverse transcription. RNase H specifically targets and nicks the RNA strand of the RNA/cDNA hybrid. Subsequently, DNA polymerase I synthesizes and extends second-strand cDNA from the 3′ end of the nicked RNA; DNA ligase then joins the second-strand cDNA fragments and the resultant cDNA can be amplified. The IVT strategy was first applied for single-cell transcriptome analysis on a microarray in neuronal cells (Eberwine et al. 2001). CEL-seq was developed on the basis of IVT and cell barcoding (Hashimshony et al. 2012), with CEL-seq2 showing improved quantitation capabilities (Hashimshony et al. 2016). In their study, the authors selected optimal enzymes for reverse transcription and second-strand cDNA synthesis. However, the ideal reaction conditions (e.g., buffer or enzyme) have not been optimized for RNase H-DNA polymerase I-mediated second-strand cDNA synthesis. Optimization of the current method or development of novel approaches can potentially improve IVT-based scRNA-seq.

Limitations of the Three Strategies

As described above, we described three basic strategies (poly-A tagging, IVT, template-switching) used in almost all scRNA-seq methods. In all of these approaches, cell barcoding and unique molecular identifier sequences can be used to quantify the number of transcripts expressed in a single cell (Islam et al. 2011, 2013; Hashimshony et al. 2012; Jaitin et al. 2014; Soumillon et al. 2014). Although they are useful for high-throughput analysis of single-cell transcriptomes, these strategies have some technical limitations. For instance, non-polyadenylated RNA cannot be detected (Tang et al. 2009; Islam et al. 2011, 2013; Hashimshony et al. 2012, 2016; Ramsköld et al. 2012; Sasagawa et al. 2013, 2018; Wu et al. 2013; Picelli et al. 2013; Streets et al. 2014; Jaitin et al. 2014; Soumillon et al. 2014; Fan et al. 2015; Nakamura et al. 2015; Matsunaga et al. 2015; Klein et al. 2015; Bose et al. 2015; Macosko et al. 2015; Muraro et al. 2016; Yang et al. 2017; Hochgerner et al. 2017; Zheng et al. 2017; Gierahn et al. 2017; Hashimoto et al. 2017; Herman et al. 2018; Rosenberg et al. 2018; Han et al. 2018), since the methods target poly-A RNA, using oligo-dT primers through a reverse transcription reaction (Fig. 4a). Moreover, reverse transcription with oligo-dT primer shows a 3′ bias in RNA-seq read coverage (Wang et al. 2009). cDNA synthesis by reverse transcriptase from the

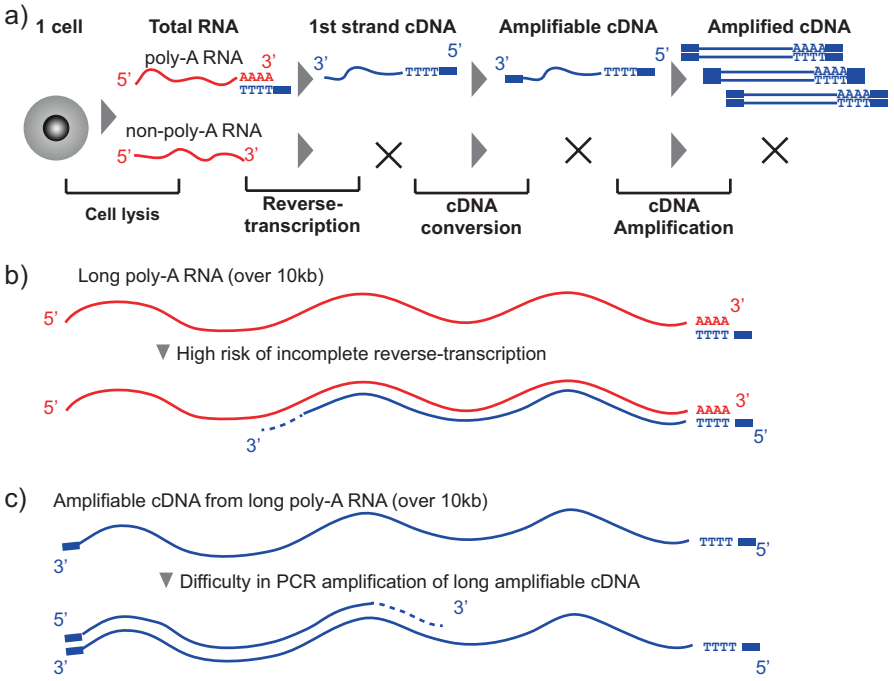

Fig. 4 Limitations of oligo-dT primer-based single-cell RNA sequencing method. Red and blue lines represent RNA strand and first-strand cDNA, respectively

3′ end of RNA is occasionally halted by RNA secondary structures or enzyme inhibitors. Thereafter, regions further downstream cannot be detected. Reverse transcriptase has an extension limit of first-strand cDNA for a long transcript (i.e., >10 kb) (Hayashi et al. 2018; Tang et al. 2012; Archer et al. 2016). Even if a long transcript is converted to first-strand cDNA and then to amplifiable cDNA, it is difficult to amplify effectively. Thus, in theory, there is a bias in the PCR amplification of short vs. long RNA in these strategies. Additionally, read coverage across a transcript is low when using an oligo-dT primer (Fig. 4). Ideally, more uniform read coverage across all transcript types (including poly-A and non-poly-A RNA) is desired for scRNA-seq.

The three strategies involve many biological reactions to convert target RNA to an analyzable DNA sequence library (Fig. 5a), which collectively comprises over 10 components (e.g., enzymes, primers, purification column/beads, etc.). Theoretically, the conversion efficiency from RNA to DNA library can be expressed as a multiple of the reaction efficiencies of individual components; that is, if 10 components have reaction efficiencies of 80%, then the cumulative efficiency is only 10% (Fig. 5b). A reduced efficiency of individual reactions has a major impact on the ability to detect an RNA molecule. It is therefore important to tune all components to maximize quantitative performance in scRNA-seq. Notably, up to one molecule of first-strand cDNA can be generated from one poly-A RNA via reverse transcription. Thus, low conversion efficiency from first-strand cDNA to an analyzable sequence library results in many low-copy genes that cannot be detected by the sequencer (Fig. 5c). This issue can be resolved by generating many first-strand cDNAs a single RNA molecule (Fig. 5c). Hence, we developed the Reverse Transcription with Random Displacement Amplification (RT-RamDA) technology, which allows for direct first-strand cDNAs from RNA without conversion to an amplifiable cDNA form (Hayashi et al. 2018) (Fig. 6). We established a single-cell full-length total RNA sequencing method based on RT-RamDA, which is described in detail in the following section.

cDNA Amplification During Reverse Transcription Without Conversion to an Amplifiable Form

In the previous sections, we described the importance of conversion efficiency to an amplifiable cDNA form to reduce the probability of dropout genes and achieve uniform read coverage across all transcripts. We also attempted to reduce dropout genes via RT-RamDA. This has two major advantages: it eliminates the requirement for conversion to amplifiable cDNA and allows for random priming during reverse transcription, thereby increasing the types of RNA that can be detected.

Random primers can anneal non-specifically to an RNA molecule, resulting in cDNA synthesis from multiple sites. This inevitably reduces the number of dropout genes during reverse transcription. Since they can anneal even on the 5′ side of RNA, random primers can in theory cover the entire length of the molecule regardless of

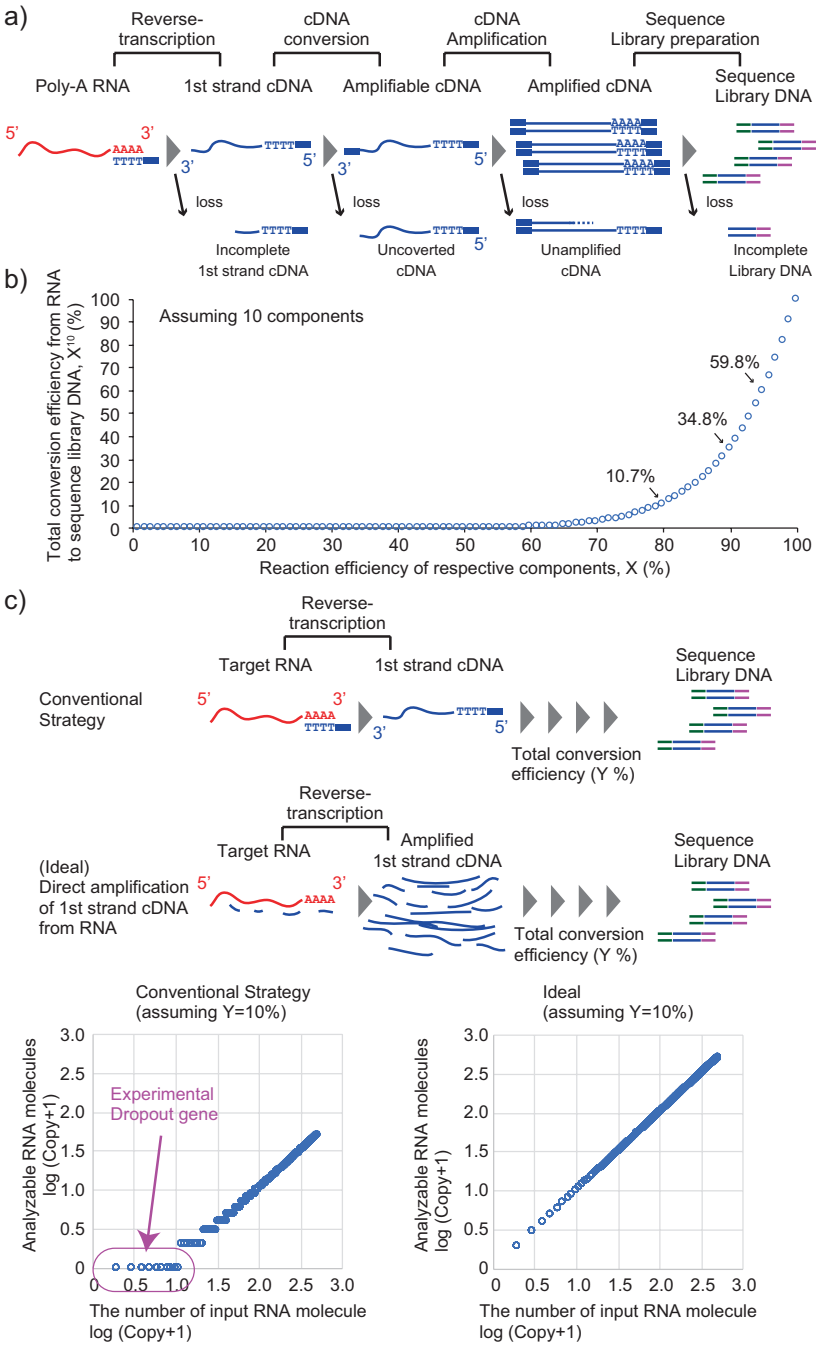

Fig. 5 Simulation of conversion efficiency from first-strand cDNA to analyzable DNA sequence library. (**a**) Schematic representation of molecule loss during conventional RNA-seq. (**b**) X axis represents the reaction efficiency of each component; Y axis represents the tenth power of the reaction efficiency of each component. (**c**) Simulation of experimental dropout genes. In theory, 10 molecules of first-strand cDNA are generated from one poly-A RNA

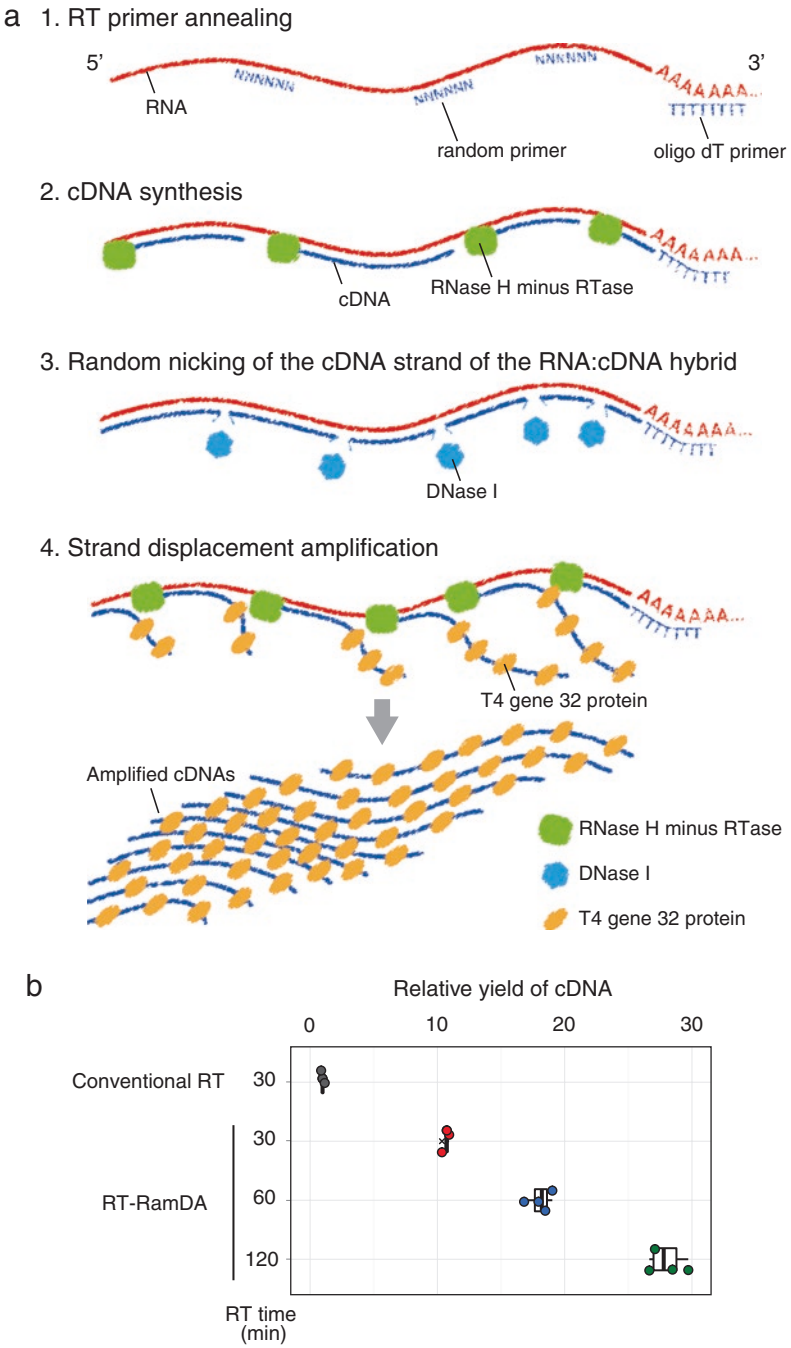

Fig. 6 Reverse transcription with random displacement amplification (RT-RamDA) can amplify cDNA directly from RNA through reverse transcription. (**a**) Schema of RT-RamDA. (**b**) Relative yield of cDNA molecules in RT-RamDA from 10 pg total RNA, calculated using conventional reverse transcription as a standard

gene length. They can also detect non-poly-A RNAs through their ability to anneal to RNA independent of poly-A sequences. In contrast, oligo-dT primers can only recognize poly-A sequences at the 3′ end of RNA; hence, one primer molecule can only anneal to a single RNA molecule, which increases the risk of gene dropout. Furthermore, as mentioned earlier, oligo-dT priming cannot detect non-poly-A RNA and does not offer full coverage of long transcripts.

Reverse transcription using random primers can prevent gene dropout. However, in scRNA-seq analysis, it has the disadvantage that even rRNA is converted to cDNA; the efficiency of reverse transcription is greatly reduced when an additional adaptor sequence is added to random primers for the amplification reaction. We avoided rRNA contamination by using not-so-random primers (Armour et al. 2009; Ozsolak et al. 2010), in which are sequences complementary to rRNA are removed beforehand. This approach has been used in bulk-RNA-seq; however, it was first applied in scRNA-seq, using RamDA-seq. However, there was no cDNA amplification method targeting whole transcripts without adding universal adaptors to random primers in reverse transcription. Therefore, we needed to develop RT-RamDA method.

In RT-RamDA, DNase I randomly introduces multiple nicks in only the first strand of cDNA of the RNA:cDNA hybrid during reverse transcription. From these nicked sites, the downstream cDNA strand is isolated via strand displacement activity of reverse transcriptase, which then synthesizes a new cDNA strand to re-form an RNA:cDNA hybrid. Importantly, T4 gene 32 protein, a single-stranded DNA-binding protein, promotes the strand displacement activity of reverse transcriptase and protects isolated cDNAs from the nuclease activity of DNase I (Fig. 6a). By repeating these reactions, first-strand cDNA can be amplified 30 times or more directly from an RNA template (Fig. 6b) (Hayashi et al. 2018). Moreover, RT-RamDA is very convenient because these reactions occur in one tube in a single step. Since only the RNA is used as a template for RT-RamDA, nonspecific amplification of contaminating DNA does not occur, as is observed in multiple displacement amplification (Lizardi et al. 1998; Dean et al. 2002; Takahashi et al. 2016). Thus, utilizing the characteristic of RT-RamDA that is a cDNA amplification method with few dropout genes, RamDA-seq can detect a larger number of genes than existing methods, including from RNAs >10 kb and non-poly(A) RNAs (Fig. 7). This is particularly useful for detecting long introns exceeding 150 kb, recursive splicing, and enhancer RNAs (Hayashi et al. 2018).

Fig. 7 (continued) (Fan et al. 2015) and the following oligo-dT primer-based methods: Switching Mechanism at 5′ End of RNA Template (SMART-Seq v4), a commercially available kit based on Smart-seq2 (Ramsköld et al. 2012; Picelli et al. 2013); and Quartz-Seq (Sasagawa et al. 2013). (**a**) Number of detected transcripts with twofold or lower changes in expression relative to rRNA depleted RNA-seq (rdRNA-seq). paRNA-seq, poly(A) RNA-selected RNA-seq. (**b**) Heat map showing the sensitivity for detecting histone transcripts, which serve as an indicator of non-poly(A) transcript detection capability. Each row represents a histone transcript, and each column represents a sample analyzed with the indicated scRNA-seq method. Expression levels are indicated as log10 (transcript per million [TPM] + 1) in accordance with the color key. (**c**) Percentage of sequence read coverage throughout the length of the transcript. The X axis shows the absolute distance (bp) from the 3′ end of the transcript. (**d**) Summary of scRNA-seq by RamDA-seq

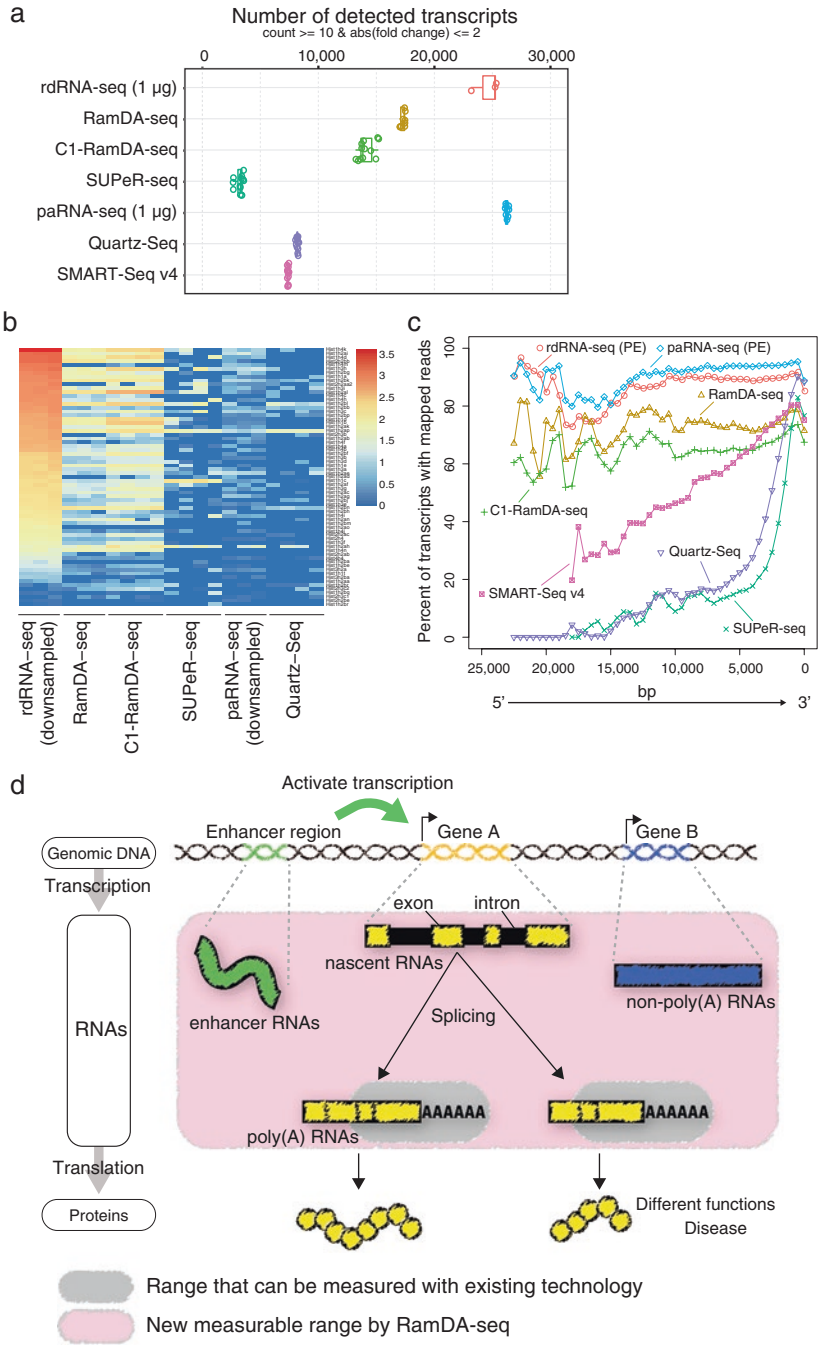

Fig. 7 Full-length total RNA sequencing from single cells via reverse transcription with random displacement amplification sequencing (RamDA-seq). (**a**–**c**) Comparison of the performances of RamDA-seq and single-cell universal poly(A)-independent RNA sequencing (SUPeR-seq)

Concluding Remarks

The three molecular biology methods used in scRNA-seq discussed in this review have a history of approximately 40 years (Fig. 1). However, various improvements have been made in these methods for high-throughput scRNA-seq. Moreover, our developed RT-RamDA technology generate new value to detect full length total RNA. RamDA-seq could help investigate the dynamics of gene expression, RNA-processing events and transcriptional regulation in single cells. However, RT-RamDA and RamDA-seq are technologies that have been developed shortly. One disadvantage of RamDA-seq is that it does not enable cell barcoding for multiplex library preparation. Additionally, to achieve uniform read coverage across full-length transcripts, the number of reads required per cell is larger than for other high-throughput scRNA-seq methods. Since the amplification rate is lower than that of other cDNA amplification methods, for RT-RamDA to become a standard technique, it will be necessary to enhance the amplification capacity. The same is true for existing full-length sequencing methods for single cells; however, RamDA-seq has yet to achieve directional sequencing. This issue is expected to be resolved in the near future.

For targeting an entire molecule of long pre-mRNA and lincRNA via RamDA-seq, we required numerous sequence reads. Currently, RamDA-seq needs 1–4 M reads per cell to over a detectable number of a transcript in typical single-cell polyA RNA-seq. To generate numerous sequencing libraries of RamDA-seq in the future, the throughput of DNA sequencing needs to be increased. We hope to increase the throughput of DNA sequencing by 10–100 times of the current sequencing platforms. We hope to develop an altogether new DNA sequencing technology.

References

Archer N, Walsh MD, Shahrezaei V, Hebenstreit D. Modeling enzyme processivity reveals that RNA-seq libraries are biased in characteristic and correctable ways. Cell Syst. 2016;3:467–479.e12.

Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods. 2009;6:647–9.

Avital G, Avraham R, Fan A, Hashimshony T, Hung DT, Yanai I. scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing. Genome Biol. 2017;18:19.

Belyavsky A, Vinogradova T, Rajewsky K. PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. Nucleic Acids Res. 1989;17:2919–32.

Bose S, Wan Z, Carr A, Rizvi A, Vieira G, Pe'er D, et al. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol. 2015;16:120.

Brady G, Barbara M, Iscove NN. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol Cell Biol. 1990;2:17–25.. [referencex]

Brady G, Billia F, Knox J, Hoang T, Kirsch IR, Voura EB, et al. Analysis of gene expression in a complex differentiation hierarchy by global amplification of cDNA from single cells. Curr Biol. 1995;5:909–22.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002;99:5261–6.

Dulac C, Axel R. A novel family of genes encoding putative pheromone receptors in mammals. Cell. 1995;83:195–206.

Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, et al. Analysis of gene expression in single live neurons. Proc Natl Acad Sci U S A. 1992;89:3010–4.

Eberwine J, Kacharmina JE, Andrews C, Miyashiro K, McIntosh T, Becker K, et al. mRNA expression analysis of tissue sections and single cells. J Neurosci. 2001;21:8310–4.

Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16:236.

Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A. 1988;85:8998–9002.

Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14:395–8.

Gubler U, Hoffman BJ. A simple and very efficient method for generating cDNA libraries. Gene. 1983;25:263–9.

Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018;173:1307.

Hashimoto S, Tabuchi Y, Yurino H, Hirohashi Y, Deshimaru S, Asano T, et al. Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues. Sci Rep. 2017;7:439.

Hashimshony T, Wagner F, Sher N, Yanai I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73.

Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. Genome Biol. 2016;17:892.

Hayashi T, Sasagawa Y, Nikaido I. RIKEN. Method for nucleic acid amplification. In: Patent WO2016052619A1; 2016.

Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9:1435.

Herman JS, Sagar, Grün D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nat Methods. 2018;15:379–86.

Hochgerner H, Lönnerberg P, Hodge R, Mikes J, Heskol A, Hubschle H, et al. STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. Sci Rep. 2017;7:566.

Iscove NN, Barbara M, Gu M, Gibson M, Modi C, Winegarden N. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. Nat Biotechnol. 2002;20:940–3.

Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21:1160–7.

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2013;11:163–6.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776–9.

Kamme F, Salunga R, Yu J, Tran D-T, Zhu J, Luo L, et al. Single-cell microarray analysis in Hippocampus CA1: demonstration and validation of cellular heterogeneity. J Neurosci. 2003;23:3607–15.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.

Kurimoto K. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res. 2006;34:e42.
Land H, Grez M, Hauser H, Lindenmaier W, Schütz G. 5′-Terminal sequences of eucaryotic mRNA can be cloned with high efficiency. Nucleic Acids Res. 1981;9:2251–66.
Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. 1998;19:225–32.
Loh E, Elliott J, Cwirla S, Lanier L, Davis M. Polymerase chain reaction with single-sided specificity: analysis of T cell receptor delta chain. Science. 1989;243:217–20.
Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
Matsunaga H, Goto M, Arikawa K, Shirai M, Tsunoda H, Huang H, et al. A highly sensitive and accurate gene expression analysis by sequencing ("bead-seq") for a single cell. Anal Biochem. 2015;471:9–16.
Matz M. Amplification of cDNA ends based on template-switching effect and step- out PCR. Nucleic Acids Res. 1999;27:1558–60.
Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst. 2016;3:385–394.e3.
Nakamura T, Yabuta Y, Okamoto I, Aramaki S, Yokobayashi S, Kurimoto K, et al. SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. Nucleic Acids Res. 2015;43:e60.
Ohara O, Dorit RL, Gilbert W. One-sided polymerase chain reaction: the amplification of cDNA. Proc Natl Acad Sci U S A. 1989;86:5673–7.
Okayama H, Berg P. High-efficiency cloning of full-length cDNA. Mol Cell Biol. 1982;2:161–70.
Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, et al. Digital transcriptome profiling from attomole-level RNA samples. Genome Res. 2010;20:519–25.
Petalidis L. Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. Nucleic Acids Res. 2003;31:142e–142.
Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82.
Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. eLife. 2017;6:503.
Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360:176–82.
Roychoudhury R, Jay E, Wu R. Terminal labeling and addition of homopolymer tracts to duplex DNA fragments by terminal deoxynucleotidyl transferase. Nucleic Acids Res. 1976;3:863–78.
Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31.
Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol. 2018;19:14049.
Sheng K, Cao W, Niu Y, Deng Q, Zong C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods. 2017;14:267–70.
Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-seq; 2014. https://doi.org/10.1101/003236.
Streets AM, Zhang X, Cao C, Pang Y, Wu X, Xiong L, et al. Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci U S A. 2014;111:7048–53.
Subkhankulova T, Livesey F. Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. Genome Biol. 2006;7:R18.

Takahashi H, Satoh T, Kanahara H, Kubota Y, Hirose T, Yamazaki H, et al. Development of a bench-top extra-cleanroom for DNA amplification. BioTechniques. 2016;61:42–6.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82.

Tang DTP, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, et al. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. Nucleic Acids Res. 2012;41:e44.

Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. Single-cell transcriptional analysis of neuronal progenitors. Neuron. 2003;38:161–75.

Van Gelder RN, Zastrow von ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A. 1990;87:1663–7.

Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2013;11:41–6.

Yang L, Ma Z, Cao C, Zhang Y, Wu X, Lee R, et al. MR-seq: measuring a single cell's transcriptome repeatedly by RNA-seq. Sci Bull. 2017;62:391–8.

Zajac P, Islam S, Hochgerner H, Lönnerberg P, Linnarsson S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. Menéndez-Arias L, editor. PLoS One. 2013;8:e85270.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.

Integrated Fluidic Circuits for Single-Cell Omics and Multi-omics Applications

Mark Lynch and Naveen Ramalingam

Abstract Single-cell genomics plays a crucial role in several aspects of biology, from developmental biology to mapping every cell in the human body through the Cell Atlas initiative. To meet these various applications, single-cell methods are rapidly evolving to increase throughput; improve sensitivity, quantification accuracy, and usability; and reduce nucleic-acid amplification bias and cost. In addition to improvement in single-cell methods, there is a huge interest in analyzing multiple analytes such as genome, epigenome, transcriptome, and protein from the same single cell. This approach is generalized as single-cell multi-omics. Automation of multi-step single-cell methods is highly desired to achieve a reproducible workflow; reduce human error and avoid contamination; and introduce technical variability to an existing stochastic process. Typically single-cell reactions start with a low level of nucleic acid, in the range of picograms. Miniaturization in microfluidic devices leads to a gain in reaction efficiency in Nanoliter or picoliter reaction volumes and active mixing help ensure that solid-state microfluidic devices provide the broadest flexibility and best sensitivity in single-cell reactions, compared to other methods. In this chapter, we will present integrated fluidic circuit (IFC) microfluidics for various single-cell multi-omics applications, and show how this technology fits into the current single-cell technology portfolio available from various vendors. We will then discuss possible uses for IFCs in multi-omics applications that are on the horizon.

Introduction

In the last decade single-cell methods have enabled us to understand the role of cell heterogeneity in many basic biological processes, particularly developmental biology (Briggs et al. 2018; Farrell et al. 2018; Wagner et al. 2018), oncology

M. Lynch (✉) · N. Ramalingam
Fluidigm Corporation, South San Francisco, CA, USA
e-mail: mark.lynch@fluidigm.com

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_2

(Brady et al. 2017), neurology (Lake et al. 2016), and immunology (See et al. 2017). The first single-cell RNA seq method were published in 2009 using two blastomeres (Tang et al. 2009). In this work, the authors targeted the poly-A^+ RNA transcripts with an oligo-dT primer with anchor sequence (UP1). Poly(A) tail was then added to first strand cDNA at 3′ end. The second strand cDNA was the synthesized using poly(A) with another anchor sequence (UP2). The cDNA was finally amplified using UP1 and UP2, fragmented, and sequenced on a SOLiD™ system. Following this there is a huge surge in number of single-cell methods to improve throughputs through the use of 96-well plates (Islam et al. 2011), 384-well plates (Jaitin et al. 2014), microfluidics (Bose et al. 2015; Fan et al. 2015; Goldstein et al. 2017; Pollen et al. 2014), droplets (Klein et al. 2015; Macosko 2015), and more recently in fixed cells or nuclei (Rosenberg et al. 2018); improve sensitivity through various modifications to the single-cell chemistry through methods such as single-cell tagged reverse transcription (STRT) (Islam et al. 2011), Smart-Seq and Smart-seq2 (Picelli et al. 2013; Ramsköld et al. 2012), SCRB-seq (Soumillon et al. 2014), Cell Expression by Linear amplification and Sequencing (CEL-Seq and CEL-Seq2) (Hashimshony et al. 2016; Hashimshony et al. 2012) and Massively parallel RNA single-cell sequencing (MARS-seq) (Jaitin et al. 2014). The ease of use in workflow is improved by barcoding template and pooling analytes at initial stages of chemistry steps (Jaitin et al. 2014; Soumillon et al. 2014).

Single-Cell Applications

Over the last 9 years, various single-cell applications for genome (Szulwach et al. 2015), transcriptome (Shalek et al. 2014), protein (Gong et al. 2016), chromatin accessibility (Buenrostro et al. 2015), methylation profiling (Guo et al. 2013) through reduced representation bisulfite sequencing (RRBS), three-dimensional architecture of whole genomes through scHi-C (Nagano et al. 2013), and chromatin state through chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Rotem et al. 2015) have been published. To understand the functional aspects of single cells, recently there have been numerous reports analyzing multiple analytes per cell and correlating these measurements. There are reports analyzing transcriptomic and genomic information (Dey et al. 2015; Macaulay et al. 2015), transcriptomic and epigenomic information (Angermueller et al. 2016), or transcriptomic and proteomic information (Darmanis et al. 2016; Frei et al. 2016; Peterson et al. 2017; Stoeckius et al. 2017; Gong et al. 2017) from the same single cell. Going one step further, there are a couple of reports analyzing three analytes (genome, epigenome and transcriptome) simultaneously such as simultaneous single-cell methylome and transcriptome sequencing (scMT-seq) (Hu et al. 2016) and single-cell triple omics sequencing (scTrio-seq) (Hou et al. 2016).

Fluidigm Single-Cell Applications

The C1™ system has over 30 individual applications. The system provides the most comprehensive application breadth of any commercial system housed on Script Hub™, as shown in the table below.

Application	Description	Compatible Fluidigm IFCs
T-ATAC Seq	Multi-omic application combined T cell receptor (TCR) and ATAC to allow investigation of the epigenetic landscape and the TCR simultaneously. This method allows the discovery of antigens that drive T cell fate or cis and trans regulators that drive the expansion of a T cell clone.	Open App
CEL-Seq 2	CEL-Seq2 is a 3′-end counting mRNAseq method that uses in vitro transcription for initial amplification. CELseq2 incorporates numerous improvements over the original method including increased sensitivity and incorporation of unique molecular identifiers (UMIs).	Open App
STRT-Seq	STRT mRNA Seq protocol is an optimized SMARTer® protocol using optimized components and UMIs to identify individual molecules.	Open App
SMART-seq2	Modification of Picelli et al Nature Methods 2014. Protocol modified from plate based method to ensure performance on C1.	Open App
C1 CAGE	C1 CAGE is a method for single-cell transcriptome analysis for molecular counting of RNA 5′-ends. Paired-end sequencing, random priming and unique molecular identifiers are used for single-molecule fragment assembly of mRNAs and long non-coding RNAs, including non-polyadenylated transcripts.	Open App
TCR-Seq	This protocol is intended to produce data from T lymphocytes compatible with T cell receptor (TCR) sequence determination using the TraCeR computational method, as well as standard gene expression analysis.	mRNA sequencing and Open App
ATAC-Seq	To reveal the landscape and principles of cellular DNA regulatory variation by developing a robust method for mapping the accessible genome of individual cells via assay for transposase-accessible chromatin using sequencing (ATAC-seq).	Open App
CORTAD-Seq	A new multi-omic approach enabling concurrent measurement of full-length mRNA and targeted genomic DNA from the same single cell. CORTAD Seq offers an unbiased and flexible approach for single-cell multi-omic analysis. As described recently in Clinical Chemistry, C1 Open App IFCs are used to generate high-quality RNA sequencing data with good coverage of the targeted genomic loci, which is essential for accurate detection of single-nucleotide variations, deletion mutations, copy number variation and haplotype construction	Open App

(continued)

Application	Description	Compatible Fluidigm IFCs
HT REAP-Seq	The RNA expression and protein sequencing (REAP-seq) assay uses DNA-barcoded antibodies to measure protein expression levels in conjunction with gene expression on the same single cells. This method leverages the DNA polymerase activity of reverse transcriptase to extend antibody barcodes (Ab BC) containing a poly(dT) tail, and synthesize cDNA from mRNA in the same reaction.	HT

Enhanced Efficiency Due to Microfluidics

Numerous reports highlight advantages of performing multi-step chemistry in microfluidics devices in nanoliter or picoliter reaction volume (Wu et al. 2014; Svensson et al. 2017; Ziegenhain et al. 2017; Hashimshony et al. 2016). One highly desired feature of the single-cell methods is the sensitivity to detect single-cell transcripts at single-molecule resolution. Wu et al. reported reduced bias when SMARTer chemistry was performed using Fluidigm C1 IFCs. Using ERCC spike-in, the authors reported quantum efficiency of amplification and detection of a single molecule in C1 to be approximately 0.63. This detection has been the best to date. It is postulated that the increased efficiency in microfluidic chambers is due to an increase in effective concentration of reactants leading to increased interaction of templates with enzymes compared to enzyme interaction with non-specific templates. Based on this idea, enhanced molecular crowding has been reported in well-plates using PEG800 (Bagnoli et al. 2017).

The original CEL-seq protocol had an efficiency of approximately 6%. Further improvement to this chemistry (CEL-seq2) by shortening RT primers, optimizing reverse transcriptase vendor, and switching to bead cleanup instead of column cleanup improved the efficiency to 19.7%. Further improvement to efficiency (22%) was noted when the chemistry was imported to Fluidigm C1 (Hashimshony et al. 2016). A recent comparison study of 15 distant experimental protocols concluded that SMARTer and CEL-seq2 on C1 have the potential to detect single-digit spike-in molecule (Svensson et al. 2018). This is in addition to two other protocols, STRT-seq and inDrop. The authors also compared use of the 96-well plate to the Fluidigm C1 system for the CEL-seq2 protocol and reported poor sensitivity for the 96-well plate compared the Fluidigm C1 system. In addition to sensitivity that translates to number of genes detected, Dueck et al. (2016) report high reproducibility of replicates with the C1 IFC.

Future Perspective

There is a huge interest in understanding and correlating the gene expression level to the protein expression in a high-throughput manner. Recent publications have demonstrated this application in droplet-based microfluidics (Peterson et al. 2017; Stoeckius et al. 2017). Gong et al. analyzed 31 proteins and mRNA in three cell lines at single-cell resolutions and noted that the distribution profile between mRNA and protein was dependent on the gene. Researchers at Merck quantified 82 barcoded antibodies with over 20,000 genes. Recently Fluidigm adjusted the workflow of the C1 Single-Cell mRNA Seq HT 10–17 μm IFC to enable C1 REAP-seq. Our new application is a powerful multi-omic single-cell application that enables deep characterization of unique cellular subtypes and functional states by measuring the expression of both cellular proteins and RNAs. Capable of pairing with functional imaging assays that measure differences in cell size, morphology, or phenotype within the clear C1 microfluidic cell chambers, C1 REAP-seq represents a significant step forward in multi-omic analysis for basic and translational research.

Conclusion

The Fluidigm C1 system and IFCs provide the single-cell community with the most comprehensive tools to study single-cell omics and single-cell multi-omics. The applications provided through Script Hub enable the C1 to provide a full localite and cellular characterization system that profiles thousands of single cells in parallel.

C1 Script Hub is the most up-to-date repository for new applications, and the C1 Script Builder™ provides the single-cell community the tools to develop custom applications and scripts to build the applications of importance to their research area.

While ultrahigh-throughput methods using droplets have emerged to provide high-throughput, cell identification and classification, our solid-state microfluidic devices with multi-step chemistry and active mixing provide the increased sensitivity to perform cell characterization upstream or downstream of droplet technologies.

References

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13(3):229–32. https://doi.org/10.1038/nmeth.3728.

Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard W. mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. bioRxiv. 2017.

Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, Pe'er D, Sims PA. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol. 2015;16(1):1–16. https://doi.org/10.1186/s13059-015-0684-3.

Brady SW, McQuerry JA, Qiao Y, Piccolo SR, Shrestha G, Jenkins DF, Layer RM, Pedersen BS, Miller RH, Esch A, Selitsky SR, Parker JS, Anderson LA, Dalley BK, Factor RE, Reddy CB, Boltax JP, Li DY, Moos PJ, Gray JW, Heiser LM, Buys SS, Cohen AL, Johnson WE, Quinlan AR, Marth G, Werner TL, Bild AH. Combating subclonal evolution of resistant cancer phenotypes. Nat Commun. 2017;8:1231. https://doi.org/10.1038/s41467-017-01174-3.

Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science. 2018;360:eaar5780.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90. https://doi.org/10.1038/nature14590.

Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, Fredriksson S, Assarsson E, Lundberg M, Nelander S, Westermark B, Landegren U. Simultaneous multiplexed measurement of RNA and proteins in single cells. Cell Rep. 2016;14(2):380–9. https://doi.org/10.1016/j.celrep.2015.12.021.

Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing from the same cell. Nat Biotechnol. 2015;33(3):285–9. https://doi.org/10.1038/nbt.3129.

Dueck HR, Ai R, Camarena A, Ding B, Dominguez R, Evgrafov OV, Fan J-B, Fisher SA, Herstein JS, Kim TK, Kim JM, Lin M-Y, Liu R, Mack WJ, McGroty S, Nguyen JD, Salathia N, Shallcross J, Souaiaia T, Spaethling JM, Walker CP, Wang J, Wang K, Wang W, Wildberg A, Zheng L, Chow RH, Eberwine J, Knowles JA, Zhang K, Kim J. Assessing characteristics of RNA amplification methods for single cell RNA sequencing. BMC Genomics. 2016;17:966. https://doi.org/10.1186/s12864-016-3300-3.

Fan HC, Fu GK, Fodor SPA. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015;347:1258367. https://doi.org/10.1126/science.1258367.

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science. 2018;360:eaar3131.

Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, Gherardini PF. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat Methods. 2016;13:269. https://doi.org/10.1038/nmeth.3742.. https://www.nature.com/articles/nmeth.3742#supplementary-information

Goldstein LD, Chen Y-JJ, Dunne J, Mir A, Hubschle H, Guillory J, Yuan W, Zhang J, Stinson J, Jaiswal B, Pahuja KB, Mann I, Schaal T, Chan L, Anandakrishnan S, Lin C-w, Espinoza P, Husain S, Shapiro H, Swaminathan K, Wei S, Srinivasan M, Seshagiri S, Modrusan Z. Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics. 2017;18(1):519. https://doi.org/10.1186/s12864-017-3893-1.

Gong H, Holcomb I, Ooi A, Wang X, Majonis D, Unger MA, Ramakrishnan R. Simple method to prepare oligonucleotide-conjugated antibodies and its application in multiplex protein detection in single cells. Bioconjug Chem. 2016;27(1):217–25. https://doi.org/10.1021/acs.bioconjchem.5b00613.

Gong H, Wang X, Liu B, Boutet S, Holcomb I, Dakshinamoorthy G, Ooi A, Sanada C, Sun G, Ramakrishnan R. Single-cell protein-mRNA correlation analysis enabled by multiplexed dual-analyte co-detection. Sci Rep. 2017;7(1):2776. https://doi.org/10.1038/s41598-017-03057-5.

Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23(12):2126–35. https://doi.org/10.1101/gr.161679.113.

Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73.

Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, Yanai I. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17(1):77. https://doi.org/10.1186/s13059-016-0938-8.

Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, Peng J. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016;26:304. https://doi.org/10.1038/cr.2016.23.. https://www.nature.com/articles/cr201623#supplementary-information

Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang C-Y, Xue Z, Fan G. Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. 2016;17(1):88. https://doi.org/10.1186/s13059-016-0950-z.

Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21:1160–7. https://doi.org/10.1101/gr.110882.110.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–9. https://doi.org/10.1126/science.1247651.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung H-L, Chen S, Vijayaraghavan R, Wong J, Chen A, Sheng X, Kaper F, Shen R, Ronaghi M, Fan J-B, Wang W, Chun J, Zhang K. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016;352(6293):1586–90.

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, Smith M, Van der Aa N, Banerjee R, Ellis PD, Quail MA, Swerdlow HP, Zernicka-Goetz M, Livesey FJ, Ponting CP, Voet T. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12:519. https://doi.org/10.1038/nmeth.3370.. https://www.nature.com/articles/nmeth.3370#supplementary-information

Macosko EZ. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502:59. https://doi.org/10.1038/nature12593.. https://www.nature.com/articles/nature12593#supplementary-information

Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35:936. https://doi.org/10.1038/nbt.3973.. https://www.nature.com/articles/nbt.3973#supplementary-information

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8. https://doi.org/10.1038/nmeth.2639.

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, Ramalingam N, Sun G, Thu M, Norris M, Lebofsky R, Toppani D, Kemp Ii DW, Wong M, Clerkson B, Jones BN, Wu S, Knutsson L, Alvarado B, Wang J, Weaver LS, May AP, Jones RC, Unger MA, Kriegstein AR, West JAA. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053–8. https://doi.org/10.1038/nbt.2967.. http://www.nature.com/nbt/journal/v32/n10/abs/nbt.2967.html#supplementary-information

Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR. Full-length mrNA-seq from single-cell levels of rNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82. https://doi.org/10.1038/nbt.2282.

Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, Gray L, Peeler DJ, Mukherjee S, Chen W, Pun SH, Sellers DL, Tasic B, Seelig G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360:176–82.

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015;33(11):1165–72. https://www.nature.com/articles/nbt.3383.

See P, Dutertre C-A, Chen J, Günther P, McGovern N, Irac SE, Gunawan M, Beyer M, Händler K, Duan K, Sumatoh HRB, Ruffin N, Jouve M, Gea-Mallorquí E, Hennekam RCM, Lim T, Yip CC, Wen M, Malleret B, Low I, Shadan NB, Fen CFS, Tay A, Lum J, Zolezzi F, Larbi A, Poidinger M, Chan JKY, Chen Q, Rénia L, Haniffa M, Benaroch P, Schlitzer A, Schultze JL, Newell EW, Ginhoux F. Mapping the human DC lineage through the integration of high-dimensional techniques. Science. 2017;356(6342):eaag3009.

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014;510:363–9.

Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv. 2014.

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8. https://doi.org/10.1038/nmeth.4380.. http://www.nature.com/nmeth/journal/v14/n9/abs/nmeth.4380.html#supplementary-information

Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–7. https://doi.org/10.1038/nmeth.4220. http://www.nature.com/nmeth/journal/v14/n4/abs/nmeth.4220.html#supplementary-information

Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599. https://doi.org/10.1038/nprot.2017.149.. https://www.nature.com/articles/nprot.2017.149#supplementary-information

Szulwach KE, Chen P, Wang X, Wang J, Weaver LS, Gonzales ML, Sun G, Unger MA, Ramakrishnan R. Single-cell genetic analysis using automated microfluidics to resolve somatic mosaicism. PLoS One. 2015;10(8):e0135007. https://doi.org/10.1371/journal.pone.0135007.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82. https://doi.org/10.1038/nmeth.1315.

Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science. 2018;360:981–7.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, Quake SR. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11(1):41–6. https://doi.org/10.1038/nmeth.2694.

Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–643.e634. https://doi.org/10.1016/j.molcel.2017.01.023.

Single-Cell DNA-Seq and RNA-Seq in Cancer Using the C1 System

Masahide Seki, Ayako Suzuki, Sarun Sereewattanawoot, and Yutaka Suzuki

Abstract Heterogeneous phenotypes of cancer cells enable them to adapt to various environments. The heterogeneity results from diversity of genome, transcriptome, and epigenome at a single-cell level. The C1 system can automatically perform single-cell capture and whole genome amplification (WGA) or whole transcription amplification (WTA) by MDA or Smart-Seq, respectively. Here, we describe the protocols for WGA and WTA from a single cell by using the C1 system and the protocols for sequence library preparation from amplified gDNA and cDNA. We also described about the computational analysis for single-cell data of cancer.

Introduction

Cell Heterogeneity

Individual cancer cells occasionally have significant heterogeneity in their phenotypic behaviors (Meacham and Morrison 2013). Even within a single population of the same cancer origin, the cells show distinct features regarding their growth rate, nutrient and oxygen requirements, and above all, eventual malignancy. Acquired mutations, which constitute so-called cancer evolution, are believed to underlie this diversification. A number of examples have been reported in which a single cell acquires advantageous mutations and expands within a cancer population, despite otherwise adverse environmental conditions (Navin et al. 2011; Hou et al. 2012; Xu et al. 2012; Wang et al. 2014). Moreover, even though not fully proven, it is thought that changes in the cell's epigenome, without accompanying genomic mutations,

M. Seki (✉) · A. Suzuki · S. Sereewattanawoot · Y. Suzuki
Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan
e-mail: mseki@edu.k.u-tokyo.ac.jp; asuzuki@edu.k.u-tokyo.ac.jp; sereewattanawoot_sarun_15@stu-cbms.k.u-tokyo.ac.jp; ysuzuki@edu.k.u-tokyo.ac.jp

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_3

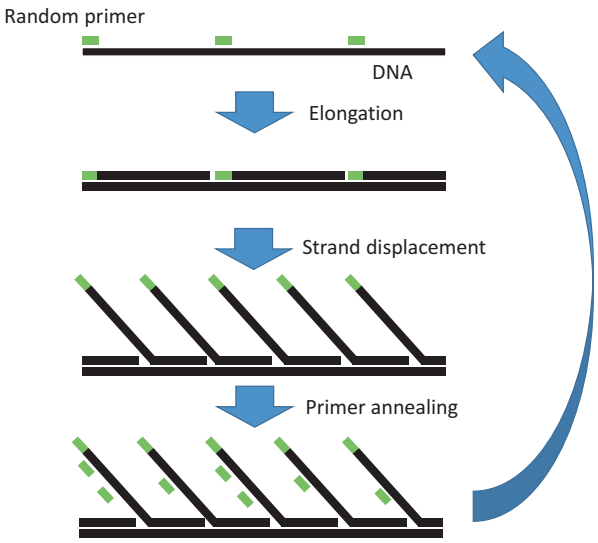

Fig. 1 Overview of whole genome amplification by multiple displacement amplification

can invoke changes in gene expression and similar phenotypic changes (Hitchins 2015). Indeed, except for several exceptional genes, comprehensive data explaining how each cancer cell accommodates to its varying microenvironment remain elusive. This knowledge is particularly important because only a limited number of cancer cells within a population that shows distinct behaviors will eventually develop into metastatic cells or cells resistant to anti-cancer drugs. Although current next-generation sequencing technology is powerful, its application in single-cell analysis has been hampered mainly by technical difficulties in handling the extremely small amounts of DNA/RNA contained within a single cell. However, very recently, these technical difficulties have begun to be overcome by refined experimental protocols as well as the automation of the procedure, which excludes human error in manipulating samples on the picogram scale.

Single-Cell DNA-Seq Method

Various whole genome amplification (WGA) methods for single-cell DNA-Seq have been developed, such as MDA (multiple displacement amplification), MALBAC (Multiple Annealing and Looping Based Amplification Cycles), and PicoPLEX (Spits et al. 2006; Zong et al. 2012; Langmore 2002). MDA is the most used method for single-cell DNA-Seq (Fig. 1). MDA employs Phi29 polymerase, which has strand displacement activity, and random primers (Spits et al. 2006). First, random primers hybridize with gDNA, and are elongated by Phi29 polymerase (Fig. 1). When double-stranded DNA exists in the direction of elongation, Phi29 polymerase dissociates it by using its strand displacement activity. Moreover,

random primers hybridize with the generated single-stranded DNA by strand displacement and are elongated. By repeating these reactions, the entire genome is amplified under isothermal conditions. By library preparation from amplified gDNA and sequencing following it, single-cell whole genome sequencing (scWGS) is conducted. Because of the high DNA replication fidelity of Phi29 polymerase, MDA has fewer amplification errors than MALBAC and PicoPlex (de Bourcy et al. 2014). Thus, MDA is suited to single nucleotide polymorphism (SNP) or single nucleotide variant (SNV) analysis of single cells. However, because MDA shows high amplification bias, MALBAC and PicoPlex are better suited for structural variation detection from single cells, such as copy number variation.

Single-Cell RNA-Seq Method

Various methods for single-cell RNA-Seq (scRNA-Seq) have also been developed and include Smart-Seq, CEL-Seq, and Quatz-Seq (Ramsköld et al. 2012; Hashimshony et al. 2012; Sasagawa et al. 2013). Although it is difficult to synthesize a sequence library directly from the small amount of mRNA from a single cell, these methods enable scRNA-Seq via whole transcriptome amplification (WTA) by PCR or *in vitro* transcription of cDNA. Smart-Seq is one of the most used methods for scRNA-Seq (Fig. 2). In Smart-Seq, reverse transcriptase, derived from Moloney murine leukemia virus (MMLV-RT), is employed (Ramsköld et al. 2012). In addition to its reverse transcriptase activity, MMLV-RT has two other important activities: terminal transferase activity to the 3′ end of cDNA and template switch activity. MMLV-RT reverse transcribes mRNA by using a primer that contains the PCR primer and a poly-dT sequence, and MMLV-RT adds dG-rich sequence to the 3′ end of the cDNA by using its terminal transferase activity. The template switch oligo (TSO), containing the PCR primer sequence and the sequence complementary to the 3′ end of the cDNA, is hybridized to cDNA. By the template switch reaction from mRNA to TSO, MMLV-RT performs further elongation and addition of the PCR primer sequence. By utilizing both ends of the PCR primer sequence of the cDNA, cDNAs are amplified. Subsequently, the sequence library is synthesized from amplified cDNA by fragmentation and ligation of adapters.

About the C1 System

The C1 system released by Fluidigm is a microfluidics-based machinery for single-cell analysis. The C1 system can automatically perform single-cell separation of up to 96 cells and WTA or WGA by Smart-Seq or MDA, respectively (Wu et al. 2014; Szulwach et al. 2015). The C1 system is also applicable to other scRNA-Seq methods, such as CEL-Seq and STRT-Seq, and the single-cell open chromatin analysis method, scATAC-Seq (Hashimshony et al. 2012; Islam et al.

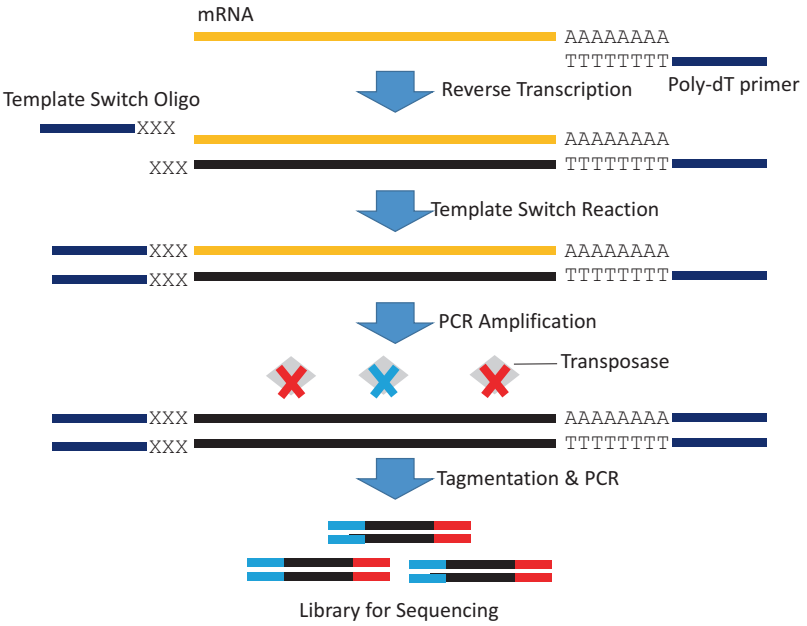

Fig. 2 Overview of single-cell RNA-Seq by Smart-Seq

2014; Buenrostro et al. 2015). The C1 mRNA Seq HT kit, which can capture up to 800 cells and prepare scRNA-Seq libraries, is also available.

Here, we describe the standard protocols of WGA and WTA using the C1 system and library preparation protocols for Illumina sequencing from amplified gDNA and cDNA. There are several methods for library preparation. We describe library preparation protocols for the KAPA HyperPlus Kit, which is an enzymatic fragmentation and ligation-based method, and the Nextera XT DNA Library Prep Kit, which is a method utilizing TD5 transposase, for scDNA-Seq and scRNA-Seq, respectively.

Materials

Reagents for scDNA-Seq

Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, cat. 25-6600-30)
C1 Single-Cell Auto Prep Reagent Kit for DNA Seq (Fluidigm, cat. 100-7357).
C1 Single-Cell Auto Prep IFC for DNA Seq 5–10 μm (Fluidigm, cat. 100-5762).
C1 Single-Cell Auto Prep IFC for DNA Seq 10–17 μm (Fluidigm, cat. 100-5763).
C1 Single-Cell Auto Prep IFC for DNA Seq 17–25 μm (Fluidigm, cat. 100-5764).
KAPA HyperPlus Kit (KAPA Biosystems, cat. KK8514).

Agencourt AMPure XP (Beckman Coulter, cat. A63880).
Nuclease-Free Water (Thermo Fisher Scientific, cat. AM9932).
Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. Q32851).
SureSelect XT2 Reagent Kit HSQ (Agilent Technologies, cat. G9661B).
10 mM Tris-HCl (pH 8.0).
High Sensitivity DNA kit (Agilent Technologies, cat. 5067-4626).

(Optional)
LIVE/DEAD Viability/Cytotoxicity Kit for mammalian cells (Thermo Fisher Scientific, cat. L-3224).
SureSelect XT2 capture library Human All Exome (Agilent Technologies, cat. 5190-9312).
Dynabeads MyOne Streptavidin T1 (Thermo Fisher Scientific, cat. 65601).
High Sensitivity D1000 kit (Agilent Technologies, cat. 5067-5584 and 5067-5585).

Reagents for scRNA-Seq

SMARTer Ultra Low RNA Kit for the Fluidigm C1 System (Clontech, cat. 634833).
C1 Reagent Kit for mRNA Seq (Fluidigm, cat. 100-6201).
C1 Single-Cell Auto Prep IFC for mRNA Seq 5–10 μm (Fluidigm, cat. 100-5759).
C1 Single-Cell Auto Prep IFC for mRNA Seq 10–17 μm (Fluidigm, cat. 100-5760).
C1 Single-Cell Auto Prep IFC for mRNA Seq 17–25 μm (Fluidigm, cat. 100-5761).
Nextera XT DNA Sample Prep Kit (Illumina, cat. FC-131-1096).
Nextera XT Index Kit (Illumina, cat. FC-131-1002).
Agencourt AMPure XP (Beckman Coulter, cat. A63880).
High Sensitivity DNA kit (Agilent Technologies, cat. 5067-4626).

(Optional)
LIVE/DEAD Viability/Cytotoxicity Kit for mammalian cells (Thermo Fisher Scientific, cat. L-3224).
ArrayControl RNA Spikes (Thermo Fisher Scientific, cat. AM1780).
THE RNA Storage Solution (Thermo Fisher Scientific, cat. AM7000).

Equipment

C1 Single-Cell Auto Prep system (Fluidigm, cat. 100-7000).
Agilent 2100 Bioanalyzer (Agilent Technologies).
Qubit 4 Fluorometer (Thermo Fisher Scientific, cat. Q33226).
Centrifuge.
Optical microscope.

Cell counter.
Thermal cycler.
Magnetic stand.

(Optional)
Fluorescence microscope.
Agilent TapeStation (Agilent Technologies).
Agilent Bravo NGS Workstation (Agilent Technologies).

Methods

In the C1 system, the efficiency of single-cell capture is largely affected by the size of the cells. In particular, when analyzing a mixture of cells of various sizes, the single cells captured by the C1 system do not necessarily reflect the original population. To collect single-cell data without cell size bias, cell size-independent methods, such as manual manipulation, microwell-based systems, and droplet-based systems, are appropriate (Gierahn et al. 2017; Macosko et al. 2015).

In WGA and WTA amplification steps, use RNase- and DNase-free pipette tips and water, and wear latex gloves to avoid RNA degradation and contamination of extraneous DNA and RNA. Unless otherwise noted, perform all experiments on ice. It is possible that leaving the dissociated cells for a long time may influence their expression patterns. Prior to conducting scRNA-Seq, prepare the reagents and perform Integrated Fluidic Circuit (IFC) priming first to minimize the time after single-cell preparation.

Preparation of Single Cells

1. For adherent cells and tissues, dissociate the cells by treating with a dissociation reagent, such as trypsin or collagenase (see Note 1). This step is unnecessary for non-adherent cells.
2. Centrifuge at 300 g for 3 min and discard the supernatant. Resuspend the cells in PBS or cell culture medium.
3. Repeat step 2.
4. Filter the cell suspension by using a 40-μm cell strainer to remove cell clumps and large debris (see Note 2).
5. Count the number of cells, and measure the cell size using a cell counter.
6. Dilute the cell suspension to 66–800 cells/μl.
7. (Optional) When distinguishing whether the captured single cells are alive or dead, prepare the C1 LIVE/DEAD solution. Mix 1.25 ml of Cell Wash Buffer, 2.5 μl of Ethidium homodimer-1, and 0.625 μl of Calcein AM. Vortex and spin down.

Whole Genome Amplification for scDNA-Seq

An outline of the C1 IFC for scDNA-Seq is indicated in Fig. 3a.

Priming of C1 DNA Seq IFC

When injecting reagents into the IFC, do not place a bubble on the bottom of the wells.

1. Choose proper C1 DNA Seq IFC that is appropriate for the cell size. Peel off the seal at the bottom of the IFC (see Note 3).
2. Push the lid of the well with a pipette and inject 200 μl of C1 Harvest Reagent into 2 wells shown by the P1 arrows.
3. Inject 20 μl of C1 Harvest reagent into 40 wells shown by the P2 arrows.
4. Inject 20 μl of C1 Preloading reagent into inlets 2 and 5 shown by arrow P3.
5. Inject 20 μl of C1 DNA Seq Cell Wash Buffer into a well shown by arrow P4
6. Inject 15 μl of C1 Blocking Reagent into two wells shown as by arrows P5 and P6.
7. Set the IFC to the C1 system and touch the icon "DNA Seq: Prime". When priming is finished, eject the IFC from the C1 system (see Note 4).

Loading Cells

1. Inject 20 μl of C1 DNA Seq Cell Wash Buffer into inlet 1 shown by arrow L1.
   (Optional) If distinguishing whether the captured single cells are alive or dead, inject C1 LIVE/DEAD solution instead of C1 Cell Wash Buffer.
2. Remove reagents from two wells shown by arrows P5 and P6.
3. Mix 12 μl of cell suspension and 8 μl of C1 Cell Suspension Reagent. Mix well by pipette and proceed immediately to the next step.
   (Optional) When the rate of cell capture is low, it is possibly improved by changing the volume of cell suspension and C1 Cell Suspension Reagent to 14 μl and 6 μl, respectively.
4. Inject the mixture into the well shown by arrow P5.
5. Set the IFC to the C1 system and touch the icon "DNA Seq: Cell Load".
   (Optional) If distinguishing whether the captured single cells are alive or dead, touch the icon "DNA Seq: Cell Load & Stain" instead of "DNA Seq: Cell Load".
6. When cell loading is finished, eject the IFC from the C1 system. Check the number of captured cells in the chamber by using an optical microscope (see Note 5).

(Optional) When the cells are stained with C1 LIVE/DEAD solution, the status of the single cells can be distinguished by fluorescence microscope. Live cells show green fluorescence, whereas dead cells show red fluorescence (Fig. 4).

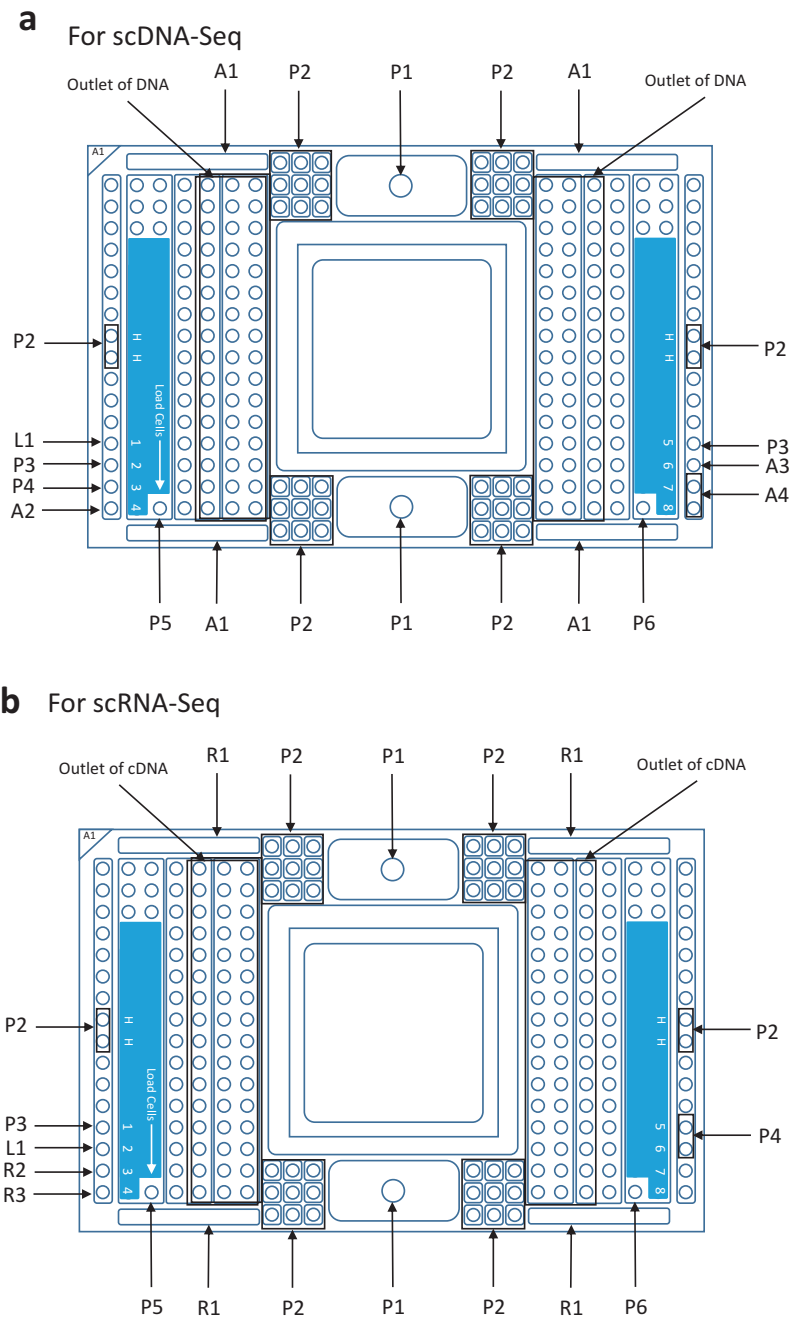

Fig. 3 Outline of IFCs of the C1 system. An outline of the C1 IFCs for DNA-Seq and RNA-Seq are shown in (**a**) and (**b**), respectively. The wells where the reagents are injected and the outlets of amplified DNA are indicated by arrows

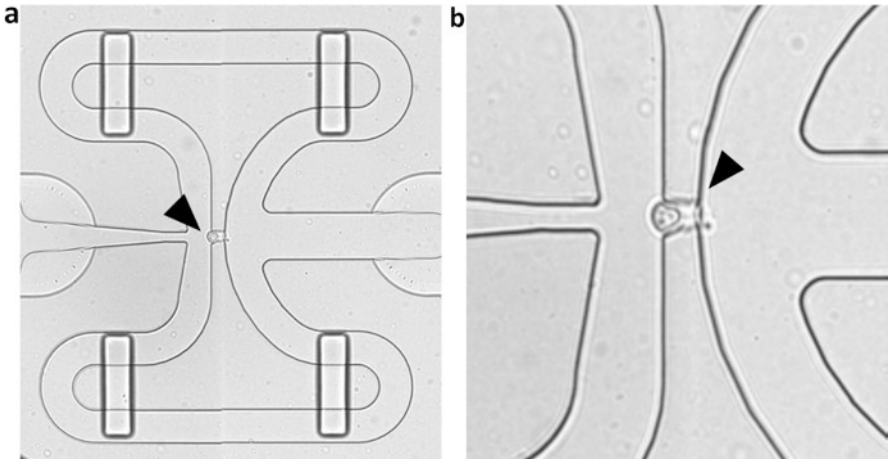

Fig. 4 Single-cell capture by the C1 system. (**a**) Picture of the chamber containing a single cell (arrowhead). (**b**) Enlarged view of a single cell is shown below. A single cell is shown by the arrowhead

Preparation of C1 Reagents

1. Mix 193.1 µl of Nuclease-Free Water, 2.3 µl of Sample Buffer, 2.3 µl of Reaction Buffer, and 2.3 µl of C1 DTT in a tube labeled "DTT Mix". Vortex and spin down.
2. Mix 13.5 µl of C1 DNA Seq Lysis Buffer and 1.5 µl of C1 DTT in a tube labeled "Lysis Mix". Vortex and spin down.
   (Optional) If preparing the control for gDNA amplification, mix 19.8 µl of C1 DNA Seq Lysis Buffer and 2.2 µl of C1 DTT in a tube labeled "Lysis Mix". Vortex and spin down.
3. Mix 30 µl of C1 DNA Seq Reaction Buffer, 21 µl of DTT Mix, and 3 µl of Enzyme Mix in a tube labeled "Reaction-Enzyme Mix". Vortex and spin down.

(Optional) If preparing the control for gDNA amplification, mix 45 µl of C1 DNA Seq Reaction Buffer, 31.5 µl of DTT Mix, and 4.5 µl Enzyme Mix in a tube labeled "Reaction-Enzyme Mix". Vortex and spin down.

Cell Lysis and gDNA Amplification

1. Add 180 µl of C1 Harvest Reagent to four rectangular wells in the four corners of the IFC shown by the A1 arrows.
2. Inject 10 µl of Lysis Mix into inlet 4 shown by arrow A2.
3. Inject 10 µl of C1 DNA Seq Stop Buffer into inlet 6 shown by arrow A3.
4. Inject 24 µl of Reaction-Enzyme Mix into inlets 7 and 8 shown by arrow A4.
5. Place the IFC into the C1 system and touch the icon "DNA Seq: Amplify". Adjust the ending time of this step (see Note 6).

Table 1 Correspondence table between the number of chambers and positions of the outlet

Left side			Right side		
C03	C02	C01	C49	C50	C51
C06	C05	C04	C52	C53	C54
C09	C08	C07	C55	C56	C57
C12	C11	C10	C58	C59	C60
C15	C14	C13	C61	C62	C63
C18	C17	C16	C64	C65	C66
C21	C20	C19	C67	C68	C69
C24	C23	C22	C70	C71	C72
C25	C26	C27	C75	C74	C73
C28	C29	C30	C78	C77	C76
C31	C32	C33	C81	C80	C79
C34	C35	C36	C84	C83	C82
C37	C38	C39	C87	C86	C85
C40	C41	C42	C90	C89	C88
C43	C44	C45	C93	C92	C91
C46	C47	C48	C96	C95	C94

Collection of Amplified gDNA

1. Dispense 10 µl of C1 DNA Dilution Reagent to a new 96-well plate. Peel off the tape sealing the outlets of the IFC.
2. Pipette the amplified gDNA from the outlets of the IFC and transfer to the plate. The chamber numbers correspond to positions of outlets indicated in Table 1. Seal the plate, vortex, and spin down.
3. Perform quantification of the amplified gDNA by the Qubit dsDNA HS assay kit (see Note 7).
4. Add Nuclease Free Water to 50 ng of the amplified cDNA to bring the volume to 35 µl.

(Optional) Preparation of Control of gDNA Amplification

1. Prepare 1 ml of a 200 cells/µl cell suspension from the same pool used for cell loading.
2. Centrifuge the suspension at 300 g for 3 min. Remove the supernatant.
3. Suspend the cell pellet in 1 ml of C1 Cell Wash Buffer. Centrifuge the suspension at 300 g for 5 min. Remove the supernatant. Repeat this step once more.
4. Resuspend the cell pellet in 0.9 ml of C1 Cell Wash Buffer.
5. Mix 1 µl of washed cell suspension and 2 µl of Lysis Mix in a new PCR tube labeled “Cell Mix”.
6. Mix 1 µl of 10 ng/µl gDNA from the Illustra kit and 2 µl of Lysis Mix in a new PCR tube labeled “gDNA Mix”.
7. Mix 1 µl of Nuclease Free Water and 2 µl of Lysis Mix in a new PCR tube labeled “NTC”.

8. Vortex and spin down. Incubate on ice for 10 min.
9. Add 4 μl of C1 DNA Seq Stop Buffer to each tube. Vortex and spin down. Incubate at room temperature for 3 min.
10. Mix 1.05 μl of the sample and 8.95 μl of Reaction-Enzyme Mix in three new tubes labeled with each sample name. Incubate the tubes in a thermal cycler using the following cycle for amplification by MDA.

    38 °C for 2 hours.
    70 °C for 15 min.
    Hold at 4 °C.

11. Vortex and spin down. Dilute 1 μl of each control sample after amplification with 9 μl of C1 DNA Dilution Reagent.
12. Perform quantification by using the Qubit dsDNA HS assay kit.

scDNA-Seq Library Preparation from Amplified gDNA

We describe the library preparation method using the KAPA HyperPlus Kit. Procedures for library preparation and purification are automatable by using the Agilent Bravo Workstation. Other methods are also applicable, such as the Nextera Rapid Capture Kit and Truseq Nano DNA Kit combined with sonication.

Fragmentation and Adapter Ligation

1. Mix 35 μl of diluted DNA, 5 μl of KAPA Frag Buffer, and 10 μl of KAPA Frag Enzyme in a 96-well plate. Vortex gently and spin down.
2. Incubate the plate in a thermal cycler using the following cycle for fragmentation (see Note 8).

    Pre-cooling at 4 °C.
    37 °C for 20 min.
    Hold at 4 °C.

3. Once the sample reaches 4 °C, immediately proceed to the next step.
4. Add 7 μl of End Repair & A-Tailing Buffer and 3 μl of End Repair & A-Tailing Enzyme Mix to the plate. Vortex and spin down.
5. Incubate the plate in a thermal cycler using the following cycle for end repair and dA-tailing.

    65 °C for 30 min.
    Hold at 4 °C.

6. Add 5 μl of Nuclease Free Water, 30 μl of Ligation Buffer, 5 μl of SureSelect XT2 Pre-capture Index Adapter, and 10 μl of DNA Ligase to the plate. Incubate the plate in a thermal cycler at 20 °C for 15 min (see Note 9).

Purification After Ligation

Perform this step at room temperature.

1. Add 88 μl of AMPure XP to the plate (see Note 10). Mix well by pipette and incubate for 5 min.
2. Place the plate on a magnetic stand and incubate until the liquid is completely clear.
3. Remove supernatant. Add 200 μl of 80% ethanol and incubate for 30 s. Repeat this step once more.
4. To remove the ethanol completely, air-dry the sample for 5 min (see Note 11). Remove the tube from the magnetic stand.
5. Suspend the bead in 21 μl of 10 mM Tris-HCl (pH 8.0). Disperse dried beads by pipette and incubate for 2 min.
6. Place the tube on the magnetic stand until the liquid is completely clear. Transfer 20 μl of the supernatant to a new 96-well plate.

Library Enrichment and Purification

1. Add 25 μl of 2× KAPA HiFi Hotstart ReadyMix and 5 μl of 10× Library Amplification Primer Mix to the plate. Incubate the tube in a thermal cycler using the following cycle for PCR amplification (see Note 9 and 12).

   98 °C for 45 s.
   8 cycles of

      98 °C for 15 s.
      60 °C for 30 s.
      72 °C for 30 s.

   72 °C for 60 s.
   Hold at 4 °C.

   The following procedure is performed at room temperature.
2. Add 50 μl of AMPure XP to the plate (see Note 10). Mix well by pipette and incubate for 5 min.
3. Place the plate on a magnetic stand and incubate until the liquid is completely clear.
4. Remove the supernatant. Add 200 μl of 80% ethanol and incubate for 30 s. Repeat this step once more.
5. To remove the ethanol completely, open the lid of the tube and air-dry the sample for 5 min (see Note 11). Remove the tube from the magnetic stand.
6. Resuspend the beads in 21 μl of Nuclease Free Water. Disperse dried beads by pipette and incubate for 2 min.
7. Place the tube on a magnetic stand until the liquid is completely clear. Transfer approximately 20 μl of the supernatant to a new 96-well plate.

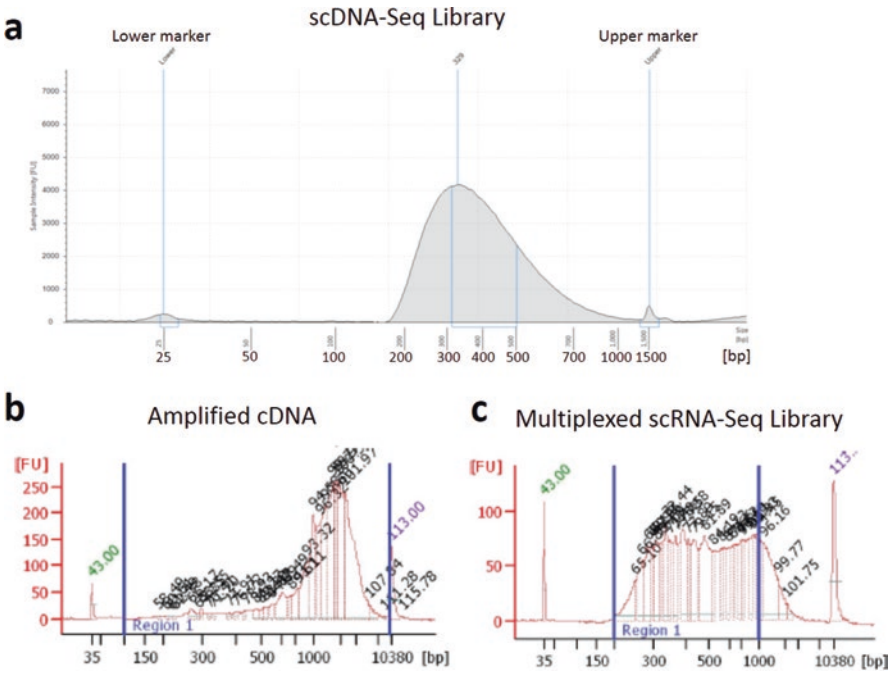

Fig. 5 Quantification and quality analysis of amplified DNA and libraries. (**a**) scDNA-Seq library quantified by the HighSensitivity D1000 kit from Agilent TapeStation. (**b** and **c**) Amplified cDNA (**b**) and multiplexed scRNA-Seq library (**c**) quantified by the High Sensitivity DNA kit from the Agilent 2100 Bioanalyzer

8. Perform quantification and quality analysis of the purified sequence library by using the High Sensitivity Kit of the Agilent Bioanalyzer or TapeStation (Fig. 5a).
9. (Optional) For multiplexing libraries, mix the libraries with different indexes at equal molarity.

Sequencing and Computational Analysis of the scDNA-Seq Library

By sequencing the enriched library on the Illumina platform, you can conduct scWGS of a single cell. If scExome-Seq is desired, perform exon capture and PCR amplification by using the SureSelect XT2 Target Enrichment System before sequencing.

The inherent advantage of using scDNA-Seq for identifying SNVs in a single or limited number cells is that the SNV data can be treated in a digital manner. Only three statuses are possible for each SNV: namely, null (wild type homo), hetero and homo. Therefore, the problem is how to distinguish these three statuses from given sequence data. This outcome is conceptually different from the SNV calling of bulk

Fig. 6 SNV detection of single cells. SNV detection from 5 cells SNV of leukocyte detected by scExome-Seq is shown

cells, where the SNV should be represented as a frequency of alleles within the population and this frequency can sometimes be very low. Indeed, assuming that an error rate of the sequencer is 0.1%, mutations occurring in 1/1000 cells cannot be detected from bulk cells in theory. Moreover, in the analysis of scDNA-Seq data, all the sequence data should indicate the presence or absence of the SNV in the corresponding cells. After identifying whether a cell has a SNV, the number of SNV-containing cells is counted within a population (Fig. 6 for examples).

Whole Transcriptome Amplification for scRNA-Seq

An outline of the C1 IFC for scRNA-Seq is indicated in Fig. 3b.

Preparation of Control RNA Spikes (Optional)

1. Dilute 1.5 μl of ArrayControl RNA Spikes 7, 4, and 1 with 13.5 μl, 12.0 μl, and 148.5 μl of the RNA Storage Solution in tubes labeled “1”, “2”, and “3”, respectively. Vortex and spin down all tubes.
2. Add 1.5 μl of the mixture of tube “1” to tube “2”. Vortex and spin down tube B.
3. Add 1.5 μl of the mixture of tube “2” to tube “3”. Vortex and spin down tube C.
4. Aliquot the RNA Spikes mixture of tube “3” into 1.25 μl aliquots. Store at −80 °C, if it will not be immediately used.

Preparation of C1 Reagents (see Note 13)

1. (Optional) Mix 1 μl of the RNA Spikes mixture and C1 Loading Reagent. Vortex and spin down.
2. Mix 11.5 μl of Dilution Buffer, 7 μl of 3' SMART CDS Primer IIA, 0.5 μl of RNA Inhibitor, and 1 μl of C1 Loading Buffer in a tube labeled "Mix A". Mix by pipette.

   (Optional) When using RNA Spikes, add a 1 μl dilution of the RNA Spikes mixture instead of C1 Loading Buffer.
3. Mix 11.2 μl of 5X First-Strand Buffer, 5.6 μl of 10 mM dNTP, 1.4 μl of DTT, 1.4 μl of RNase Inhibitor, 5.6 μl of SMARTer IIA Oligonucleotide, 5.6 μl of SMARTScribe Reverse Transcriptase, and 1.2 μl of C1 Loading Reagent in a tube labeled "Mix B". Vortex gently and spin down.
4. Mix 63.5 μl of PCR Water, 10 μl of 10X Advantage 2 PCR Buffer, 4 μl of 50X dNTP Mix, 4 μl of IS PCR Primer, 4 μl of 50X Advantage 2 Polymerase Mix, and 4.5 μl of C1 Loading Reagent in a tube labeled "Mix C". Vortex and spin down.
5. (Optional) When distinguishing whether captured single cells are alive or dead, prepare the C1 LIVE/DEAD solution. Mix 1.25 ml of Cell Wash Buffer, 2.5 μl of Ethidium homodimer-1, and 0.625 μl of Calcein AM. Vortex and spin down.

Priming of C1 IFC

When injecting reagents, do not place a bubble on the bottom of the wells.

1. Choose proper C1 IFC matching that is appropriate for the cell size. Peel the seal from the bottom of the IFC.
2. Push the lid of the well by pipette and inject 200 μl of C1 Harvest Reagent into 2 wells shown by the P1 arrows.
3. Inject 20 μl of C1 Harvest Reagent into 40 wells shown by the P2 arrows.
4. Inject 20 μl of C1 Preloading Reagent into inlet 1 shown by arrow P3.
5. Inject 20 μl of C1 Cell Wash Buffer into two wells shown by arrow P4
6. Inject 15 μl of C1 Blocking Reagent into two wells shown by arrows P5 and P6.
7. Set the C1 IFC to C1 system and push the button "RNA-Seq: Prime". When the priming is finished, eject the C1 IFC from the C1 system.

Loading Single Cells

1. Inject C1 Cell Wash Buffer into inlet 1 shown by arrow L1.

   (Optional) If distinguishing whether the captured single cells are alive or dead, inject C1 LIVE/DEAD solution instead of C1 Cell Wash Buffer.
2. Remove the reagents from two wells shown by arrows P5 and P6.

3. Mix 12 μl of cell suspension and 8 μl of C1 Cell Suspension Reagent. Mix well by pipette and proceed to the next step.
   (Optional) If the single cell capture rate is low, it is possibly improved by changing the volume of cell suspension and C1 Cell Suspension Reagent to 14 μl and 6 μl, respectively.
4. Inject the cell mixture into the well shown by arrow P5.
5. Set the C1 IFC on the C1 system and push the button "RNA-Seq: Cell Load".
   (Optional) If distinguishing whether the captured single cells are alive or dead, push the button "RNA Seq: Cell Load & Stain" instead of "RNA Seq: Cell Load".
6. When cell loading is finished, eject the C1 IFC from the C1 system. Check the number of captured cells in each capture site by using an optical microscope (see Note 5).

(Optional) Whether single cells stained with the C1 LIVE/DEAD solution are alive or dead can be distinguished using a fluorescence microscope. Live cells show green fluorescence, whereas dead cells show red fluorescence.

Cell Lysis, Reverse Transcription, and Amplification

1. Add 180 μl of C1 Harvest Reagent to four rectangular wells in the four corners of the IFC shown by the R1 arrows.
2. Inject 9 μl of Mix A into inlet 3 shown by arrow R2.
3. Inject 9 μl of Mix B into inlet 4 shown by arrow R3.
4. Inject 24 μl of Mix C into inlets 7 and 8 shown by arrow R4.
5. Set the IFC on the C1 system and push the button "mRNA Seq: RT+Amp". Adjust the ending time of this step. Leaving the IFC for a long time causes the samples to evaporate. Therefore, eject the IFC from the C1 system and collect the samples from the IFC within 2 h after this step is completed.

Collection of Amplified cDNA

1. Dispense 10 μl of C1 DNA Dilution Reagent to a new 96-well plate. Peel the tape that seals the outlets of the amplified cDNA.
2. Pipette samples from the outlets of the IFC and transfer to the plate. The numbers of capture sites correspond to the numbers indicated in Table 1. Seal, vortex, and spin down the plate.
3. Perform quantification and quality analysis of the amplified cDNA using the High Sensitivity DNA Kit of the Agilent 2100 Bioanalyzer (see Note 14). Calculate the concentration of DNA from 100 bp to 10,000 bp by region setting (Fig. 5b).
4. Dilute 2 μl of cDNA solution with an adequate amount of C1 Harvest Buffer to a concentration of 0.1–0.3 ng/μl.

(Optional) cDNA Preparation of Bulk Cell Control

1. Prepare 1 ml of 200 cells/μl cell suspension from the same pool used for cell loading.
2. Centrifuge the suspension at 300 g for 5 min. Remove the supernatant.
3. Suspend the cell pellet in 1 ml of C1 Cell Wash Buffer. Centrifuge the suspension at 300 g for 5 min. Remove the supernatant. Repeat this step once more.
4. Resuspend the cell pellet in 0.9 ml of C1 Cell Wash Buffer.
5. Mix 1 μl of washed cell suspension and Mix A in a PCR tube. Vortex gently and spin down. Incubate the tube in a thermal cycler using the following cycle for cell lysis and primer hybridization.

   72 °C for 3 min.
   4 °C for 10 min.
   25 °C for 1 min.
   Hold at 4 °C.

6. Spin down and add 4 μl of Mix B to the tube. Vortex gently and spin down. Incubate the tube in a thermal cycler using the following cycle for reverse transcription.

   42 °C for 90 min.
   70 °C for 10 min.
   Hold at 4 °C.

7. Vortex and spin down. Mix 1 μl of Mix C and 1 μl of first strand cDNA in a new PCR tube. Vortex and spin down. Incubate the tube in a thermal cycler using the following cycle for PCR amplification.

   95 °C for 1 min.
   5 cycles of

   - 95 °C for 20 s.
   - 58 °C for 4 min.
   - 68 °C for 6 min.

   9 cycles of

   - 95 °C for 20 s.
   - 64 °C for 30 s.
   - 68 °C for 6 min.

   7 cycles of

   - 95 °C for 30 s.
   - 64 °C for 30 s.
   - 68 °C for 7 min.

   72 °C for 10 min.
   Hold at 4 °C.

8. Vortex and spin down. Mix 1 μl of PCR product and 45 μl of C1 DNA Dilution Reagent in a new tube. Vortex and spin down.
9. Perform quantification and quality analysis of amplified cDNA using the High Sensitivity DNA Analysis Kit of the Agilent 2100 Bioanalyzer.
10. Dilute 2 μl of cDNA solution with an adequate amount of C1 Harvest Buffer to a concentration of 0.1–0.3 ng/μl.

scRNA-Seq Library Preparation from Amplified cDNA

Library Preparation by Tn5 Tagmentation and PCR

1. Add 2.5 μl of Tagmentation DNA Buffer, 1.25 μl of Amplicon Tagment Mix, and 1.25 μl of diluted cDNA to a new 96-well plate. Seal the plate, vortex gently, and spin down.
2. Incubate the plate in a thermal cycler using the following cycle.

   55 °C for 10 min.
   Hold at 10 °C.

3. When at the plate reaches 10 °C, immediately add 1.25 μl of Neutralize Tagment Buffer to stop the tagmentation reaction. Seal the plate, vortex, and spin down.
4. Add 3.75 μl of Nextera PCR Master Mix, 1.25 μl of Index Primer 1, and 1.25 μl of Index Primer 2 to the plate. Incubate the plate in a thermal cycler using the following cycle.

   72 °C for 3 min.
   95 °C for 30 s.
   12 cycles of

   - 95 °C for 10 s.
   - 55 °C for 30 s.
   - 72 °C for 60 s.

   72 °C for 5 min.
   Hold at 10 °C.

Multiplexing of Libraries and Purification

These steps are performed at room temperature.

1. Incubate AMPure XP at room temperature for 30 min and vortex until the precipitate disperses completely.
2. For multiplexing the libraries, mix the libraries with different indexes at equal molarity in a tube. Add AMPure XP in an amount that is 0.9 times the amount of the mixed library. Mix by pipette and incubate for 5 min.

3. Place the tube on a magnetic stand and incubate for 2 min.
4. Remove the supernatant. Add 180 μl of 70% ethanol and incubate for 30 s. Repeat this step once more.
5. To remove the ethanol, open the lid of the tube and air-dry the sample for 9 min. Remove the tube from the magnetic stand.
6. Add the same amount of Resuspension Buffer as the amount of the library mixture before purification. Disperse dried beads by pipette and incubate for 2 min.
7. Place the tube on a magnetic stand and incubate for 2 min.
8. Transfer the supernatant to a new tube. Add AMPure XP in an amount that is 0.9 times the amount of the supernatant. Mix by pipette and incubate for 5 min.
9. Place the tube on a magnetic stand and incubate for 2 min.
10. Remove the supernatant. Add 180 μl of 70% ethanol and incubate for 30 s. Repeat this step once more.
11. To remove the ethanol completely, open the lid of the tube and air-dry the sample for 9 min. Remove the tube from the magnetic stand.
12. Add Resuspension Buffer in an amount that is 1.5 times the amount of the library mixture before purification. Disperse the dried beads by pipette and incubate for 2 min.
13. Place the tube on a magnetic stand for 2 min. Transfer the supernatant to a new tube.
14. Perform quality analysis of the multiplexed library using the kit of the Agilent 2100 Bioanalyzer (Fig. 5c).

Sequencing and Computational Analysis of the scRNA-Seq Library

By sequencing the enriched library on an Illumina platform, you can conduct scRNA-Seq. Although the number of samples is generally large, 100–1000 RNA-Seq datasets, each of which represents a single-cell transcriptome, the basic data process itself is the same as that of standard RNA-Seq. For mapping and converting the mapping information to gene expression information, general RNA-Seq software, such as TopHat2 and Cufflinks, can be used (Trapnell et al. 2009, 2010). For visual inspection, IGV and the UCSC Genome Browser can be used (Robinson et al. 2011; Thorvaldsdóttir et al. 2013; Kent et al. 2002) (Fig. 7a). Notably, in some methods, such as CEL-Seq and STRT-Seq, only one end of the transcripts is represented, although the entire transcript is represented in the most popular method for this purpose, Smart-Seq (Ramsköld et al. 2012; Hashimshony et al. 2012; Islam et al. 2011). When begun from data collected using specialized methods, the method of tag counts should be modified accordingly.

A unique feature of the scRNA-Seq analysis is present in the later step. To extract relevant information, the gene expression information collected from hundreds of cells should be collectively analyzed. The easiest analysis may be to calculate

deviations of gene expression among cells. For this purpose, relative divergence (standard deviation divided by average) is often used. When the divergence within a population is unknown, it may provide clues for understanding the molecular bases of diverse cellular phenotypes. Another frequently used approach is the clustering of cells based on their expression information (Ramsköld et al. 2012; Patel et al. 2014; Suzuki et al. 2015). As exemplified in Fig. 7b, even within the same cancer cell line, significant expression divergence is sometimes observed. This clustering can be performed using either the entire genes or a group of selected genes, such as cancer-related genes. In fusion gene detection, it is difficult to detect significant fusion transcripts from single-cell data. Thus, it is better to first detect fusion genes from data merged from reads of single cells by using fusion detection software, such as TopHat-fusion and DEFUSE, and then to distinguish whether the data for each single cell contain sequence tags covers the detected junction (Suzuki et al. 2015; Kim and Salzberg 2011; McPherson et al. 2011) (Fig. 7c).

Notes

1. The incomplete dissociation of cells causes the capture of multiple cells in a chamber. The cells should be dissociated completely.

   Dead cell shows an aberrant profile. When a sample contains many dead cells, remove dead cells using a cell sorter or magnetic bead-based methods.
2. If the cell suspension contains substantial amounts of small debris, increase the number of washing steps.
3. An unsuitable choice of IFC causes a decrease in single-cell capture efficiency.
4. After priming, the IFC can be stored at 4 °C for several hours.
5. Living cells have a high degree of transparency. Thus, even without staining, it is possible to distinguish whether cells are alive or dead roughly by the degree of transparency.

   If a single cell is not captured at the capture site, judge single-cell capture as successful when the chamber contains a single cell (Fig. 4). In contrast, if a single cell is captured in the capture site, judge single-cell capture as a failure when the chamber contains two cells or more.
6. When using the IFC for small cells and others, this step takes approximately 6.25 and 7.5 h, respectively, at the shortest. Leaving the IFC for a long time after the ending time causes the samples to evaporate.
7. The Quant-iT PicoGreen dsDNA Assay Kit is also applicable for quantification. The concentration of amplified gDNA is approximately 5–40 ng/μl.
8. By changing the incubation time at 37 °C, it is possible to optimize the average size of DNA after fragmentation.

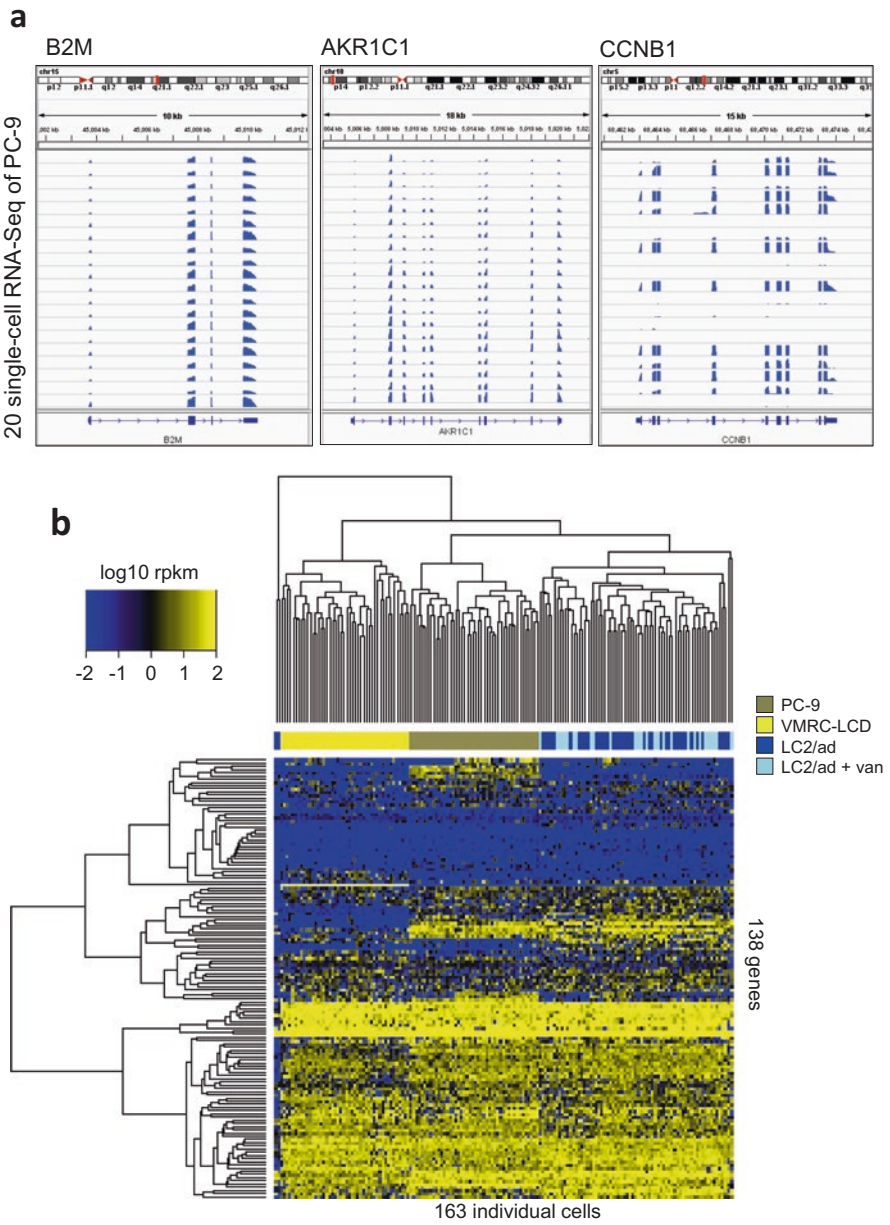

Fig. 7 Single-cell RNA-Seq analysis in lung adenocarcinoma cell line. (**a**) Selected 20 single-cell RNA-Seq data of PC-9 were visualized by IGV. (**b**) Hierarchical clustering analysis was performed by expression patterns of 138 cancer-driver genes (Vogelstein et al. 2013). Data of 163 individual cells were used including 46, 46, 43, and 28 cells from PC-9, VMRC-LCD, LC2/ad, and LC2/ad with vandetanib stimulation, respectively. (**c**) A fusion transcript CCDC6-RET in LC2/ad single-cell and bulk RNA-Seq data. Junction sequence tags of CCDC6-RET in single-cell RNA-Seq data. The junction point of CCDC6-RET was identified from bulk RNA-Seq data. Bulk RNA-Seq tags around the junction point in CCDC6 and RET were visualized by IGV

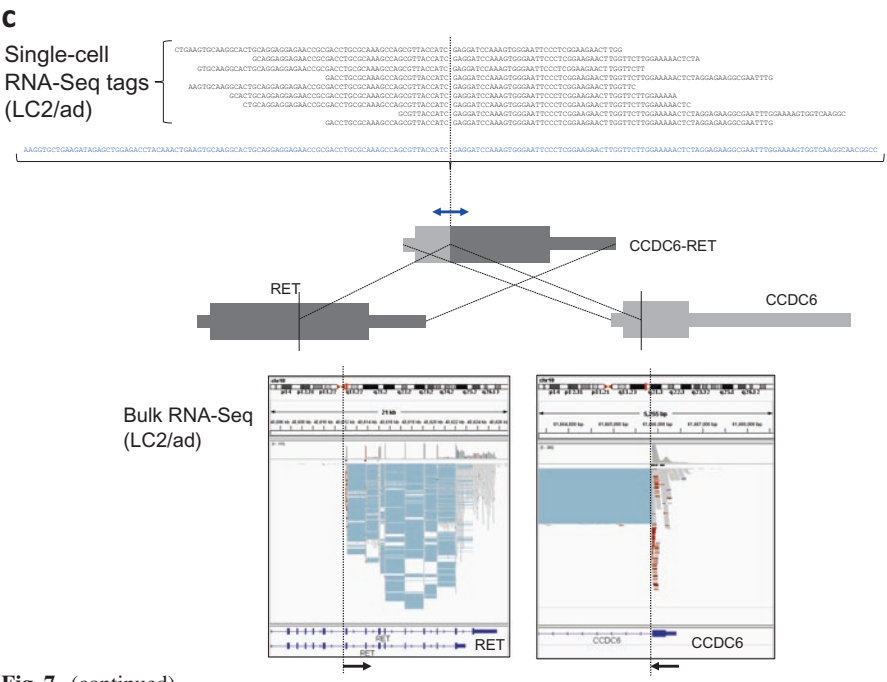

Fig. 7 (continued)

9. If necessary, dilute the adapter oligo mixture according to the instructions of the KAPA HyperPlus Kit.

   If you do not want to perform scExome-Seq (single-cell exome sequencing) using the SureSelect XT2 Target Enrichment System, you can also use the adapters released by other manufacturers.
10. Before use, incubate AMPure XP at room temperature for 30 min and vortex until the precipitate disperses completely.
11. Remaining ethanol or excess drying causes a low recovery rate of DNA.
12. The required PCR cycle differs depending on the size and concentration of the sample. Optimize the number of cycles according to the manufacturer's instructions.
13. These reagents can be stored at 4 °C for a few hours.
14. The Quant-iT PicoGreen dsDNA Assay Kit is also applicable for quantification.

Acknowledgments We would like to express our gratitude to Y Kuze, T Horiuchi, K Kunigo, Y Ishikawa, and K Imamura for helpful advice in writing this manuscript. This work was supported by MEXT KAKENHI Grant Number 221S0002.

References

Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501:328–37. https://doi.org/10.1038/nature12624.

Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4. https://doi.org/10.1038/nature09807.

Hou Y, Song L, Zhu P, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148:873–85. https://doi.org/10.1016/j.cell.2012.02.028.

Xu X, Hou Y, Yin X, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148:886–95. https://doi.org/10.1016/j.cell.2012.02.025.

Wang Y, Waters J, Leung ML, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60. https://doi.org/10.1038/nature13600.

Hitchins MP. Constitutional epimutation as a mechanism for cancer causality and heritability? Nat Rev Cancer. 2015;15:625–34. https://doi.org/10.1038/nrc4001.

Spits C, Le Caignec C, De Rycke M, et al. Whole-genome multiple displacement amplification from single cells. Nat Protoc. 2006;1:1965–70. https://doi.org/10.1038/nprot.2006.326.

Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338:1622–6. https://doi.org/10.1126/science.1229164.

Langmore JP. Rubicon Genomics, Inc. Pharmacogenomics. 2002;3:557–60. https://doi.org/10.1517/14622416.3.4.557.

de Bourcy CF, De Vlaminck I, Kanbar JN, et al. A quantitative comparison of single-cell whole genome amplification methods. PLoS One. 2014;9:e105585. https://doi.org/10.1371/journal.pone.0105585.

Ramsköld D, Luo S, Wang Y-C, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82. https://doi.org/10.1038/nbt.2282.

Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003.

Sasagawa Y, Nikaido I, Hayashi T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31. https://doi.org/10.1186/gb-2013-14-4-r31.

Wu AR, Neff NF, Kalisky T, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11:41–6. https://doi.org/10.1038/nmeth.2694.

Szulwach KE, Chen P, Wang X, et al. Single-cell genetic analysis using automated microfluidics to resolve somatic mosaicism. PLoS One. 2015;10:e0135007. https://doi.org/10.1371/journal.pone.0135007.

Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11:163–6. https://doi.org/10.1038/nmeth.2772.

Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90. https://doi.org/10.1038/nature14590.

Gierahn TM, Wadsworth MH, Hughes TK, et al. Seq-Well: portable, low-cost rna sequencing of single cells at high throughput. Nat Methods. 2017;14:395–8. https://doi.org/10.1038/nmeth.4179.

Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11. https://doi.org/10.1093/bioinformatics/btp120.

Trapnell C, BA W, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5. https://doi.org/10.1038/nbt.1621.

Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6. https://doi.org/10.1038/nbt.1754.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92. https://doi.org/10.1093/bib/bbs017.

Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006. https://doi.org/10.1101/gr.229102.

Islam S, Kjällquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21:1160–7. https://doi.org/10.1101/gr.110882.110.

Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401. https://doi.org/10.1126/science.1254257.

Suzuki A, Matsushima K, Makinoshima H, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. 2015;16:66. https://doi.org/10.1186/s13059-015-0636-y.

Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12:R72. https://doi.org/10.1186/gb-2011-12-8-r72.

McPherson A, Hormozdiari F, Zayed A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011;7:e1001138. https://doi.org/10.1371/journal.pcbi.1001138.

Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science. 2013;339:1546–58. https://doi.org/10.1126/science.1235122.

Nx1-Seq (Well Based Single-Cell Analysis System)

Shinichi Hashimoto

Abstract Research on the hierarchical nature of cell differentiation and heterogeneity in tissues has been performed by isolating and identifying cells by the use of monoclonal antibodies, cell sorting, microdissection, and functional assays. However, it is difficult to analyze continuous changes in cell differentiation and the identification of cells for which cell markers are unclear. Furthermore, cell populations considered identical were shown to be diverse. Recently, single cell gene expression analysis was performed to help understand the complexity of cell populations. Single-cell analysis can analyze the diversity of individual cell populations as well as the tissue microenvironment, and is extremely useful for research on intercellular interactions in diseases and identifying specific marker genes. Recent advances in technology have made it possible to analyze hundreds of single cells. In this paper, we introduce our newly developed well-based single-cell transcriptome method, which includes other methods.

Introduction

The clarification of a cell phenotype by gene expression analysis is widely used in the fields of biology and medicine. Previous measurements of gene expression have been performed on bulk samples. However, to characterize the complexity/diversity of a tissue it is necessary to analyze the gene expression of a single cell. Gene expression analysis of a classical single cell is commonly performed by PCR amplification of RNA or microarray, which attempts to identify the characteristics of cells by analyzing known genes. However, if there is no change in the gene examined, no result is obtained. Therefore, methods such as CEL-seq (Hashimshony et al. 2012), Quartz-Seq (Sasagawa et al. 2013) and Smart-Seq (Picelli et al. 2014) were developed to make cDNA from all mRNAs in a cell for analysis. Recently, analyses were

S. Hashimoto (✉)
Kanazawa University, Kanazawa, Ishikawa, Japan
e-mail: hashimoto@med.kanazawa-u.ac.jp

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_4

performed using Fluidigm Instrument C1, which can analyze the genes of hundreds of cells (Pollen et al. 2014; Treutlein et al. 2014). However, at this level, it is not suitable when the cell diversity is large. Therefore, methods such as Drop-seq (Macosko et al. 2015), iDrop RNA sequencing (Klein et al. 2015), Cyto-Seq (Fan et al. 2015a) and Nx1-Seq have been developed to analyze mRNA expression in hundreds of individual cells. Currently, the diversity of cells in tissues is studied using these methods. To separate specific cells from various cell populations, an efficient data analysis procedure in a single cell transcriptome is also provided (Jaitin et al. 2014; Buettner et al. 2015; Patel et al. 2014; Li and Li 2018).

Principle of Cellular Gene Analysis

Every method is based on the classification and identification of individual cells by adding barcode sequences to nucleic acids. The basic procedure is as follows: (a) lysis of cells in a limited space in a microwell plate or a droplet; and (b) collection of RNA from the cell and its capture by microbeads or hydrogel. For these methods, specific equipment is necessary. Specifically, one-cell analysis using a microwell plate consists of the following steps: (1) prepare a single cell suspension from the tissue; (2) insert barcode microparticles and a single cell into one well with pico-nanoliter scale liquid volume; (3) add lysis buffer to dissolve cells; (4) capture single cell-derived mRNA via oligo dT on the barcode microparticle that forms a bead to obtain mRNA from a single cell; (5) collect beads in one tube and subject them to reverse transcription; and (6) analysis where cells are identified using the barcodes. The microparticles currently used are about 20–30 μm in size and bind oligonucleotides containing a barcode sequence on the surface. Oligonucleotides on beads consist of the following four parts: (i) a sequence for use as a priming site for downstream PCR and sequencing; (ii) a "cell barcode" (identical to the on the surface of one bead – when we use a 12 bp barcode, the diversity of the barcode on the bead is 4^{12} = 16,777,216); (iii) a unique molecular identifier (UMI) (identifying duplication of PCR); and (iv) an oligo dT sequence for capturing polyadenylated mRNA and priming reverse transcription. The barcode portion of this oligonucleotide can be synthesized directly on beads by the "split-and-pool" DNA synthesis method or by using the emulsion PCR method described later. That is, the 12 bp barcode sequence is used to identify the cells and the 8 bp random barcode sequence is used to eliminate gene duplication bias by PCR.

Nx1-Seq (Hashimoto et al. 2017)

We developed a method to analyze one cell from thousands to several tens of thousands simultaneously and named Nx1-seq (Fig. 1). This new approach is simple and can be used to analyze hundreds to tens of thousands of cells without any special equipment. In addition, this microwell with barcode beads in a Lab-Tek chamber

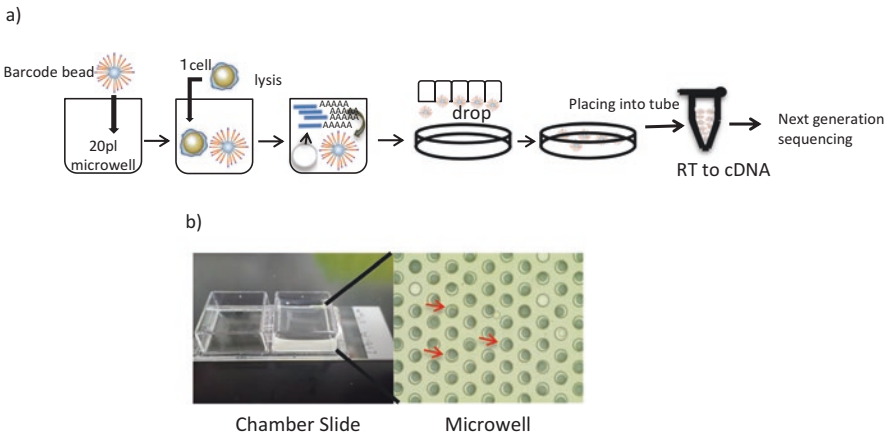

Fig. 1 Schematic of the Nx1-Seq analysis. (**a**) A cell suspension and barcoded beads are mixed on a chamber slide containing a PDMS microwell. In brief, a cell suspension is placed on each PDMS microwell slide and allowed to settle into the wells by gravity. The cell seeding process is 90% efficient by measuring the cell concentration in seeding buffers for both pre- and post-cell seeding. The distribution of cells as single or multiple cells per well was calculated using Poisson statistics. The wells are then rinsed with PBS. Next, cold cell lysis solution is applied to wells for 12 min at room temperature. The PDMS slide is inverted in a dish containing 2 mL of cold lysis solution to force the beads out of the microwells. Subsequently, the beads are washed. The RNAs bound to microbeads are converted into cDNA. (**b**) Red arrows indicate wells that contain a single cell

slide can be stored with the buffer for several months before use. Therefore, the plate can be carried anywhere and experiments can be performed immediately after cell separation without preparing the plate. In addition, the size of the well varies with the degree of morphology and film composition depending on the size of the target cell: a larger microwell for cancer cells or a smaller microwell for leukocytes and other non-cancer cells. It can also be modified: if the cell size is unknown, empty wells without beads can easily be checked with a microscope.

However, to obtain more information for a cell, we should measure the gene expression and mutations of a specific gene at the same time. In single cell gene expression analysis, the barcode is at the 3′ end; thus, in many cases, only the sequence at the 3′ end is required. However, utilizing the principle of emulsion (em) PCR, we can create barcode beads that trap specific sites of genes and genomes using specific primers.

Procedure of Nx1-Seq

The method is a scalable approach for the digital gene expression profiling of thousands of single cells without the use of robotics. The method has three steps: (1) preparation of beads conjugated with barcode nucleotides and of oligo-dTs by means of emulsion PCR; (2) placing a barcode bead into 20-pL wells molded in

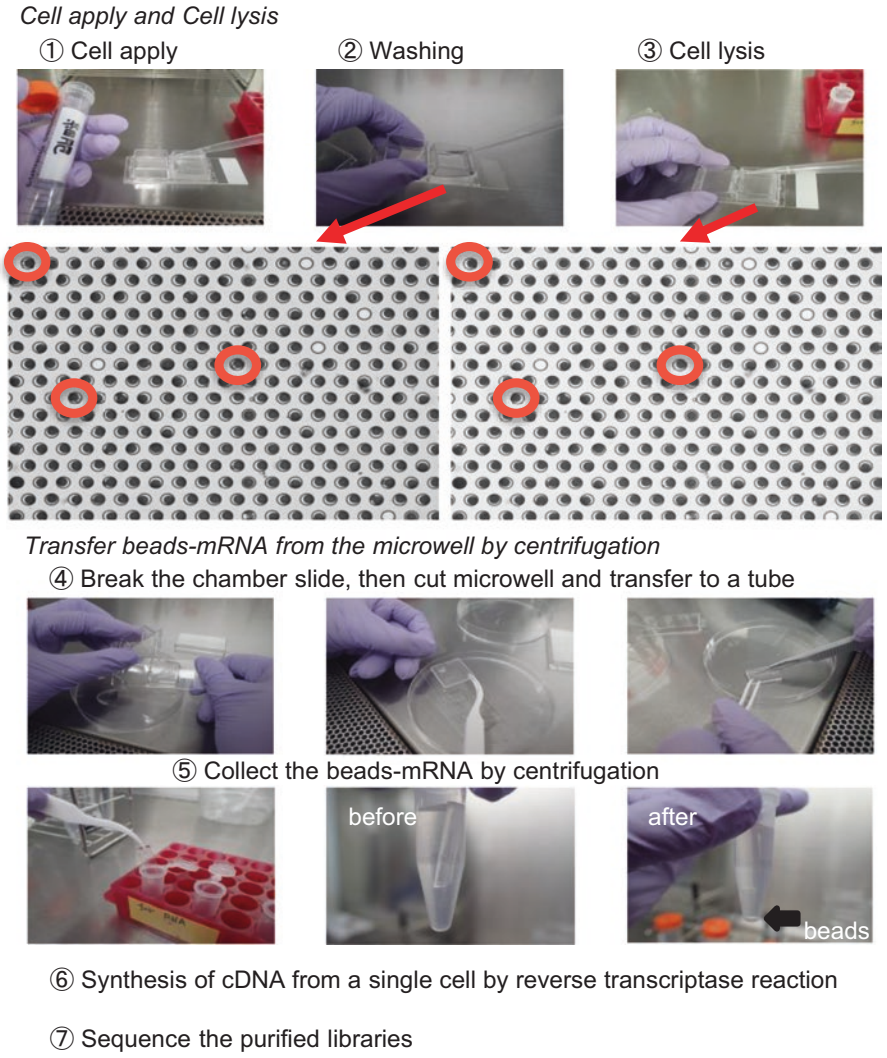

Fig. 2 Procedure of the Nx1-Seq analysis. Numbers indicate each step. A red circle around a microwell indicates a well containing a cell. The cell is lysed after adding lysis buffer to the well

polydimethylsiloxane (PDMS) slides (Fig. 2); and (3) adding a population of heterogeneous cells into wells. Each slide can contain 1.6×10^5 wells (2×2 cm). Poly(dT) barcoded beads with a diameter of 20 μm are first added to the microwell slide to achieve 1 bead/well. Approximately 10,000 cells are allowed to settle into the wells of a slide by gravity (Fig. 2①). The slides are then incubated with a cell lysis solution containing 1% lithium dodecyl sulfate (Fig. 2③). After lysis, cellular mRNA binds to the poly(dT) barcoded beads, which are collected and used for reverse transcription.

Production of Barcode Beads Using Emulsion PCR

In addition to oligo dT that traps mRNA by creating barcode beads using emulsion PCR (emPCR), it is possible to insert chains complementary to specific genes and genomic sequences on beads (Fig. 3). Barcode beads are prepared by using beads to which arbitrary DNA sequences are bound beforehand and amplifying one molecule of synthetic oligo to be a barcode on the beads. The specific primer is designed

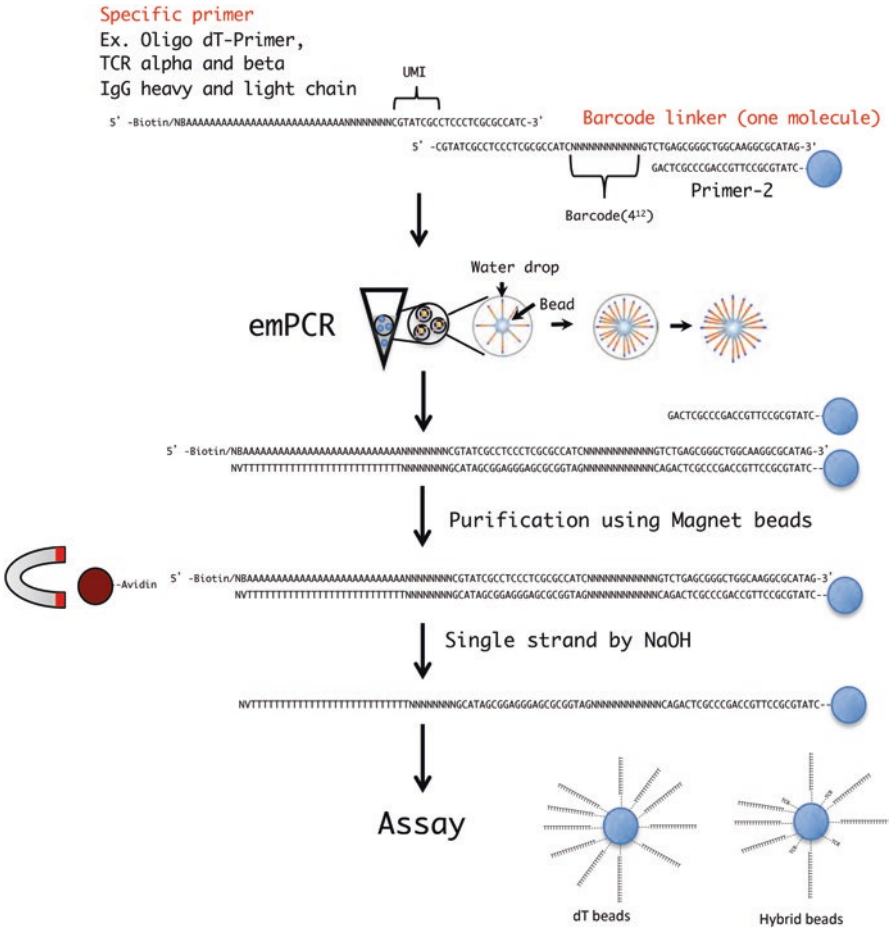

Fig. 3 Strategy to generate barcoded oligonucleotides. One barcode linker molecule enters a single water drop with a DT-primer containing an UMI and forward primer. We then place an insert sequence for any binding site such as TCR, or IgG, into the capture beads using specific primers. After emulsion PCR, barcode beads are purified as described in the manual for the emPCR kit. Then, the barcode beads are treated with *Exo*I for 30 min. After the inactivation of *Exo*I at 80 °C, the beads are rendered single stranded by the removal of the secondary strand through incubation in NaOH solution. Before assay, the barcode beads are stored at −20 °C

to contain the UMI sequence and dT. In addition, when trapping an arbitrary part of a gene such as TCR or IgG, emPCR is performed by changing the dT part of a specific primer to a TCR or IgG insertion sequence. It is also possible to insert sequences such as Oligo dT and TCR at the same time into beads by increasing the type of specific primers. After emulsion PCR, barcode beads are purified and barcode beads with single strands with barcode sequences are produced by the removal of secondary strands by incubation of the beads in NaOH solution.

Reproducibility and Sensitivity of Nx1-Seq

We used Nx1-seq to analyze single cell transcriptomes in PC9 cells, a lung cancer cell line, to ensure the reproducibility of technical replicates. The top 100 libraries from PC9 cells were compared to the bulk cell library, and Pearson correlation coefficients of approximately 0.94 were obtained (Fig. 4a). The pooling of Nx1-seq data from single cells of two homogeneous cell populations provided rich and highly reproducible transcriptional profiles. In addition, gene expression patterns among libraries with large sequencing reads were similar ($r = 0.97$), as shown in the mean of the scatter plots (Fig. 4b). Moreover, Nx1-seq data was compared between two libraries using mouse cell lines. The gene detection levels per read for each cell line were similar to those of Drop-seq (Fig. 4c, d). These data are similar to other Droplet-methods.

Measurement Example

As an example of the single-cell analysis of cancer tissues using Nx1-seq, tissue samples were removed from the myometrial infiltration side (M-side) and endometrial side (E-side) of a human endometrioid adenocarcinoma tissue (Hashimoto et al. 2017). Human endometrioid adenocarcinoma tissues are comprised of varying ratios of different cell populations of infiltrating immune cells and cancer cells (Fig. 5a). Myometrial invasion is an independent prognostic parameter of endometrioid carcinomas and is correlated with the risk of metastasis to the lymph nodes. In addition, it is believed that cells expressing some malignancy-related genes increase at the myoinvasive front. However, in this study, cancer cells in the E-side were highly malignant compared with those in the M-side. Many cells on the E-side were positive for spheroid-specific markers such as *SOX2*, *CTSV* and *GNG2*, which are related to tumorigenesis. In addition to the increased frequency of cells with a cancer stem cell marker in the E-side, the population of EMT in the E-side was higher than in the M-side. Moreover, the population of infiltrating T cells in the tumor, which is a marker for prognosis, was smaller in the E-side compared with the M-side. These data, produced using Nx1-seq, demonstrated that cells with high malignant potential (HMP) were present in the site of the same cancer tissue

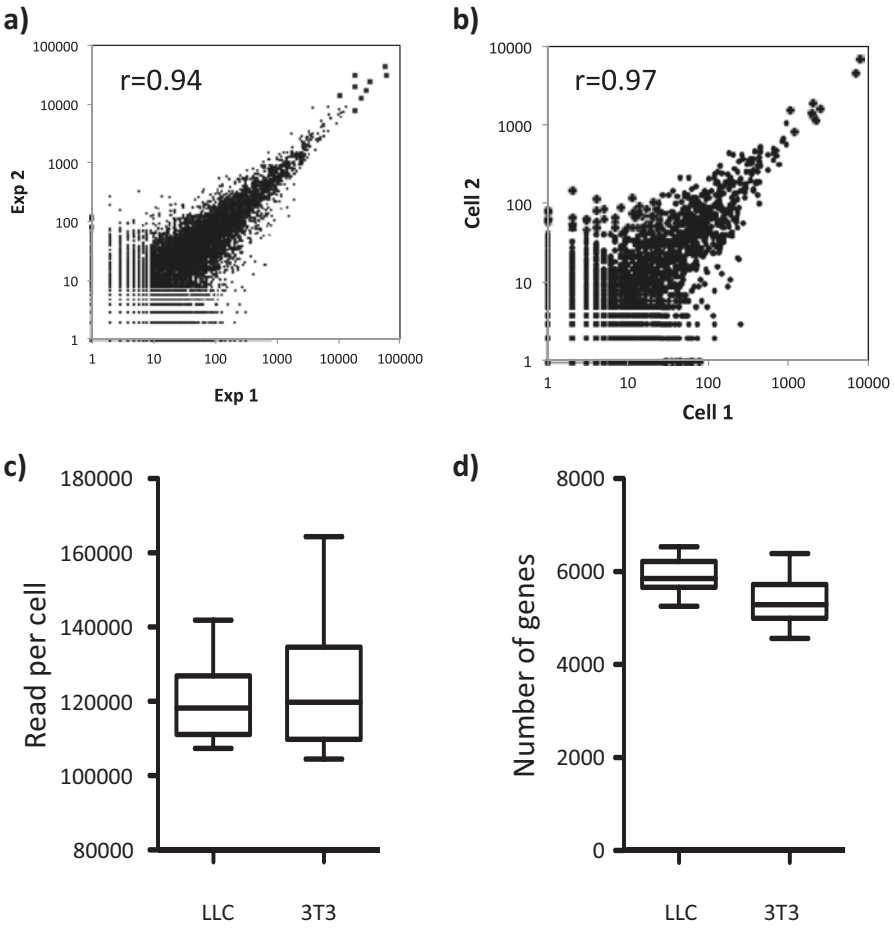

Fig. 4 Reproducibility and sensitivity of Nx1-seq. (**a**) Comparison of Nx1-seq gene expression measurements (mean of 100 single cell data) between two independent experiments. (**b**) Comparison of Nx1-seq gene expression measurements between two independent cells. (**c** and **d**) Comparison of Nx1-seq gene expression measurements (mean of 50 single cell data) between two independent cells

(Fig. 5b). Furthermore, when the gene expressions of cancer cells from the E-side and M-side were compared, several genes were identified at sites where cancer cells had HMP. Interestingly, the differential gene expression profile of macrophages and T cells in the E-side and the M-side were similar to those of cancer cells. To confirm this phenomenon, immunostaining of the cancer tissue was carried out. Immunostaining of UCHL1, a highly expressed gene in cancer in the E-side, showed a similar staining pattern in macrophages and T cells as well as cancer cells. In addition, other differential genes were also observed in macrophages and T cells as well as cancer cells. These data showed that many immune cells have the properties (genes) of cancer cells in cancer tissues with HMP. To elucidate this phenomenon,

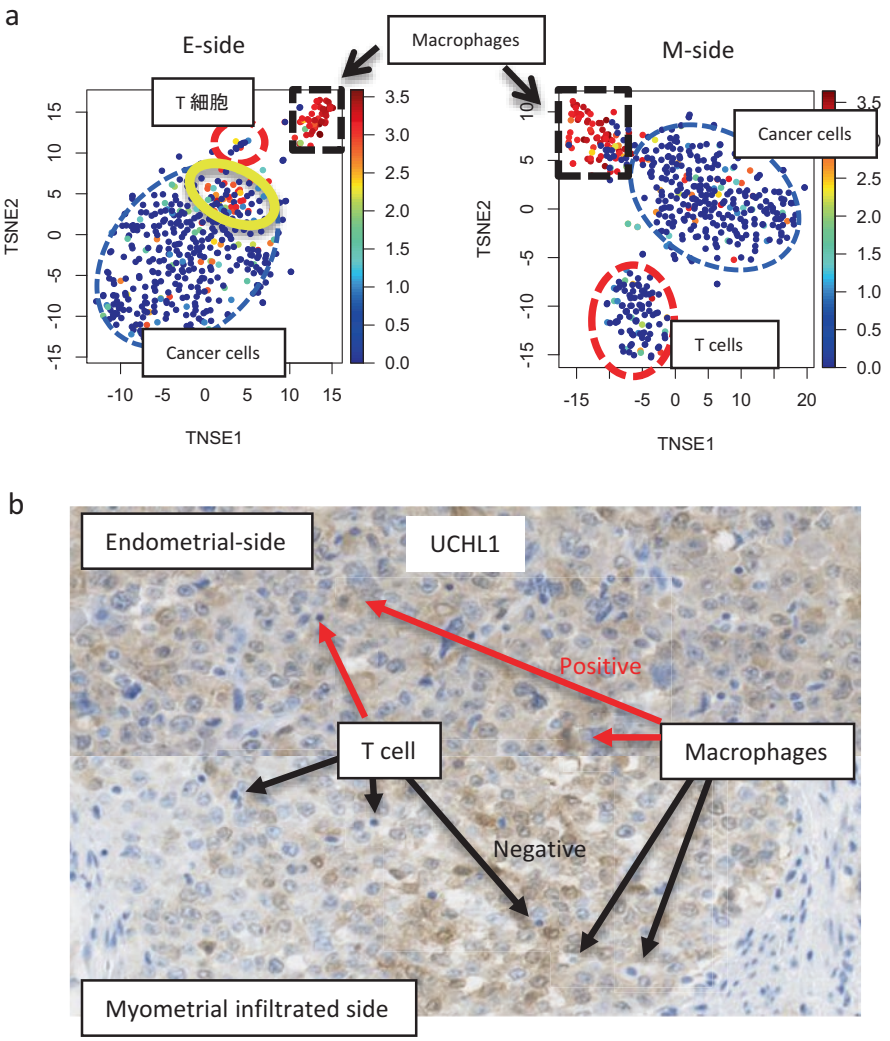

Fig. 5 Nx1-seq analysis was performed after separating 1 cell from the endometrium side or the muscle layer of the infiltrated advanced part of a fresh human endometrioid adenocarcinoma tissue. (**a**) The tSNE analysis from approximately 1000 single cells is shown. Clusters are different on both sides of cancer cells (blue enclosure) and infiltrating T cells (red enclosure). Black box: macrophages. Red dots indicate the expression levels of CD14. The yellow enclosure indicates genes expressed commonly in cancer cells and macrophages. (**b**) Immunostaining of UCHL1 in the E-side and M-side

it is very important to clarify the mechanism of new cancer development. Therefore, to achieve more effective cancer therapies, the diversity of cancer cells at each distinct site of tumor tissues must be considered.

Other Well-Based Single Cell Analysis Methods

CytoSeq: Fan et al. developed an approach to enable the cytogenetic expression of thousands of single cells without using robots and automation, termed CytoSeq (Fan et al. 2015b), which is similar to other well-based methods described previously. Cells settle into wells by gravity. Next, the bead library is loaded onto the microwell to saturation so that most wells become filled. After bead placing, the microwell is washed to delete the remaining beads. Then, mRNA captured with oligo(dT) are sequenced. As an example, a complex complication of the human hematopoietic system was characterized using CytoSeq. They examined cytokines, transcription factors, and genes that encode intracellular proteins of various cellular functions that may not be easily analyzed by flow cytometry. They demonstrated the ability to identify major subsets within human peripheral blood mononuclear cells and showed cellular heterogeneity in resting CD3+ T cells compared with those stimulated with antibodies to CD3 and CD28, as well as resting CD8+ T cells compared with those stimulated with CMV peptides. In addition, they found that the upregulation of a number of genes in the stimulated samples originated from only a few cells.

Seq-Well: Gierahn et al. also devised a well-based method, named Seq-Well (Gierahn et al. 2017). They improved some of the problems of CytoSeq. To demonstrate the efficiency of cell lysis and mRNA trapping from a single cell, capture beads and single cells were incubated in semipermeable membranes on the microwell. An important feature of Seq-Well is the use of selective chemical functionalities to promote the reversible deposition of semipermeable polycarbonate membranes (10 nm pore size) in physiological buffers. This enables a rapid exchange for efficient cell lysis and capture of transcripts and it reduces cross-contamination. As an example, they used cell gene expression profiling for human macrophages treated with *Mycobacterium tuberculosis*.

Microwell-Seq: Hanler et al. reported how to capture mRNA on magnetic beads using an agarose microarray to establish a broadly accessible, cost-effective single technology, and called Microwell-Seq (Han et al. 2018). They analyzed over 400,000 single cells covering all major mouse organs and built a basic screening system of the mouse cell atlas.

Conclusion

It has long been claimed that tumor heterogeneity contributes to the progression of the disease and has a major influence on the therapeutic effect. A few studies have reported the heterogeneity of both cancer cells and stroma cells. Our newly developed single cell transcriptome analysis, which we termed Nx1-seq, can overcome this problem and provide novel insights into tumor microenvironments. This new approach is simple and can be used to analyze several hundreds to tens of thousands

of cells without special equipment. The method makes use of beads with barcodes that are distributed singly into microwells. In addition, the size of the well can be modified depending on the size of the target cells: larger microwells for cancer cells or smaller microwells for leukocytes and other non-cancerous cells. Furthermore, microwells equipped with barcode-beads in a Lab-Tek chamber slide can be stored with the buffer for several months before use. In addition, microwells can be carried anywhere and used immediately after cell separation without preparing the plates. Alternatively, by using emulsion (em)PCR, we are able to place an insert sequence for any binding site such as TCR, or IgG, into the capture beads using specific primers (Fig. 3). These beads can be used for circulating tumor cell analysis, immune disorders and infections, immunotherapy and vaccination. They might also be useful for new clinical applications such as monitoring diagnosis. In conclusion, Nx1-seq analysis is a powerful approach for characterizing and understanding cellular diversity under physiological or pathological conditions. We hope it will be useful for driving new clinical applications such as monitoring diagnosis.

Acknowledgements We are most grateful to T Torigoe, Y Hirohashi and Y Takamura for technical assistance. This research is (partially) supported by JST CREST Grant Number JPMJCR15G3, Japan, and Japan Agency for Medical Research and Development (AMED).

References

Buettner F, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33:155–60. https://doi.org/10.1038/nbt.3102.

Fan X, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015a;16:148. https://doi.org/10.1186/s13059-015-0706-1.

Fan HC, Fu GK, Fodor SP. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015b;347(6222):1258367.

Gierahn TM, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14:395–8. https://doi.org/10.1038/nmeth.4179.

Han X, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018;172:1091–1107.e17. https://doi.org/10.1016/j.cell.

Hashimoto S, et al. Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues. Sci Rep. 2017;7:14225. https://doi.org/10.1038/s41598-017-14676-3.

Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003.

Jaitin DA, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776–9. https://doi.org/10.1126/science.1247651.

Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9:997. https://doi.org/10.1038/s41467-018-03405-7.

Macosko EZ, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Patel AP, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401. https://doi.org/10.1126/science.1254257.

Picelli S, et al. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9:171–81. https://doi.org/10.1038/nprot.2014.006.

Pollen AA, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32:1053–8. https://doi.org/10.1038/nbt.2967.

Sasagawa Y, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31. https://doi.org/10.1186/gb-2013-14-4-r31.

Treutlein B, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371–5. https://doi.org/10.1038/nature13173.

Quantitation of mRNA Transcripts and Proteins Using the BD Rhapsody™ Single-Cell Analysis System

Eleen Y. Shum, Elisabeth M. Walczak, Christina Chang, and H. Christina Fan

Abstract In this review, we describe the BD Rhapsody™ Single-Cell Analysis System, a platform that allows high-throughput capture of nucleic acids from single cells using a simple cartridge workflow and a multitier barcoding system. The resulting captured information can be used to generate various types of next-generation sequencing (NGS) libraries, including whole transcriptome analysis for discovery biology and targeted RNA analysis for high sensitivity transcript detection. The BD Rhapsody system can be used with emerging applications, such as BD™ AbSeq assays, to profile gene expression in both mRNA and protein level to provide ultra-high resolution analysis of single cells.

Principles of BD Rhapsody Single-Cell Analysis System

Mechanism of Cell Capture

The BD Rhapsody Single-Cell Analysis System enables gene expression measurements from 100 to >10,000 single cells simultaneously. The technology, originally described in 2015 (Fan et al. 2015), makes use of two main elements: (1) an array consisting of hundreds of thousands of microwells, and (2) a diverse library of barcoded beads. First, cells are loaded sparsely onto the microwell array, such that after falling into wells by gravity, each cell occupies a single well according to Poisson statistics. Barcoded magnetic BD Rhapsody Cell Capture Beads are then loaded close to saturation, such that each cell is paired with a bead. The geometry and dimension of the microwell and the bead are designed for single-bead occupancy, with sufficient space for a cell of ≤25 μm in diameter. Upon cell lysis, the mRNA content of each cell can be captured by probes via polyA/polyT

E. Y. Shum (✉) · E. M. Walczak · C. Chang · H. Christina Fan
BD Biosciences, San Jose, CA, USA
e-mail: Eleen.Shum@bd.com; Elisabeth.Walczak@bd.com; Christina.Chang@bd.com; Christina.Fan@bd.com

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_5

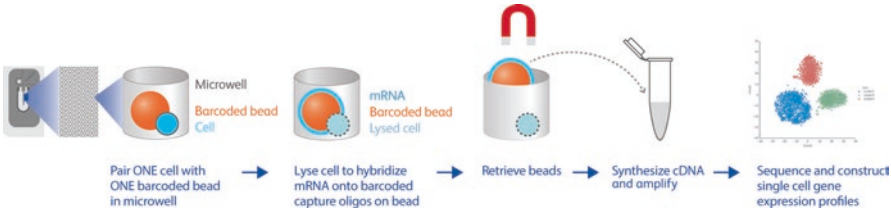

Fig. 1 Overview of the BD Rhapsody Single-Cell Analysis System

hybridization. BD Rhapsody beads are subsequently retrieved from the microwells by magnets and pooled into a single tube to allow simplified downstream NGS library preparation (Fig. 1).

Multitier Molecular Barcoding Scheme

BD Rhapsody beads contain a two-level barcoding system to differentiate each captured polyadenylated RNA by both its cell origin and transcript origin. First, each of the millions of oligo-dT (dT) primers carried on a single bead have the same cell-label (CL) sequence to allow differentiation of gene detection between different cells. Second, adjacent to the CL sequence, a unique molecular identifier (UMI) sequence is used to distinguish different mRNA transcripts captured by the bead. The use of UMI sequences mitigates biases from downstream PCR amplification and enables the counting of transcript molecules (Fu et al. 2011). The number of available CL sequences is in excess as compared to the number of cells analyzed per cartridge, such that each cell is paired with a bead that has a unique CL. Similarly, the diversity of UMIs that coat each bead allows almost all transcripts of the same gene from a cell to receive a different UMI. During cDNA synthesis, the reverse transcriptase enzyme extends a cDNA complementary to the original mRNA transcript from dT, allowing the CL and UMI to be paired with each cDNA molecule (Fig. 2a). The resulting reaction generates a barcoded transcriptome of thousands of single cells that can be archived.

Sequencing Library Generation Using BD Rhapsody

To study the single-cell gene expression profiles, cDNA is amplified and converted into sequencing libraries. Multiple amplification schemes can be employed depending on user application. For exploratory and discovery work, the user can append a universal adaptor at the 3′ end of the cDNA molecule via ligation, followed by

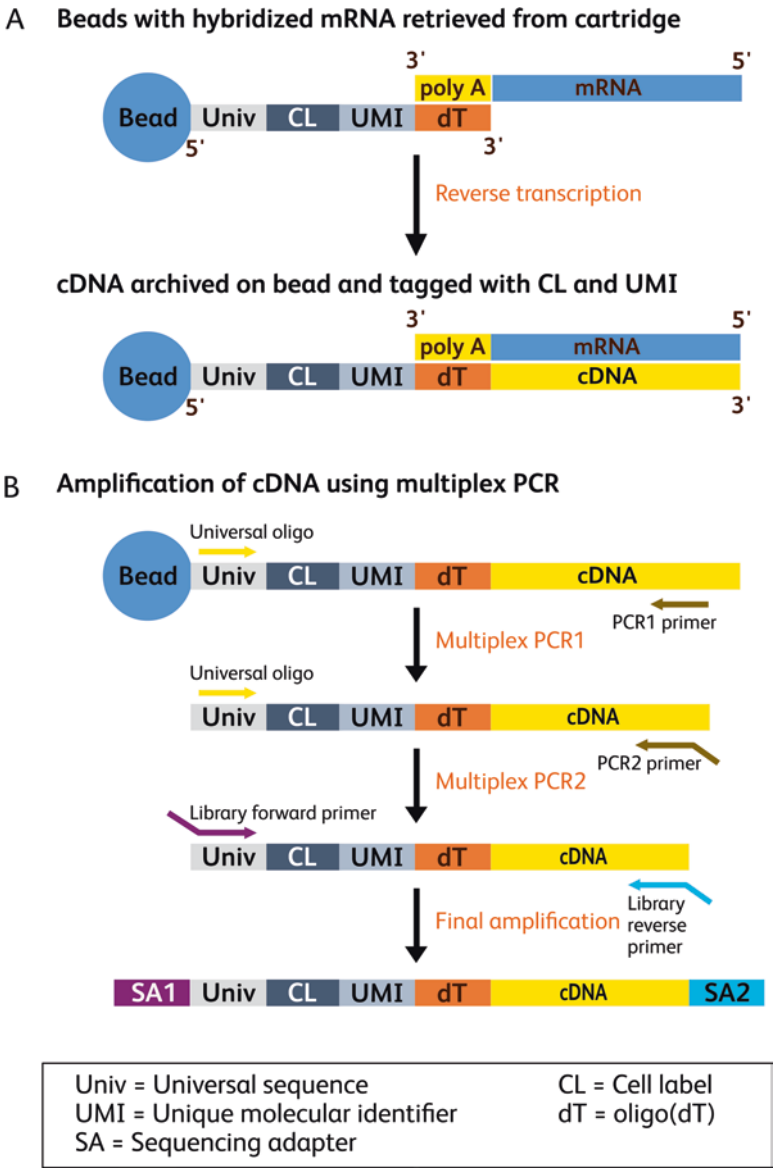

Fig. 2 Molecular biology steps of the BD Rhapsody system to convert barcoded transcripts to sequencing libraries. (**a**) Attachment of cell label (CL) and unique molecular identifier (UMI) sequences to cDNA molecules via reverse transcription with oligo capture probes on BD Rhapsody beads. (**b**) Targeted amplification and sequencing of cDNA captured on BD Rhapsody beads

universal amplification to amplify the whole transcriptome. The tradeoff of such an approach is a bias for detecting housekeeping genes that are several-fold higher in expression as compared to key cell-type identifiers such as transcription factors. Alternatively, a targeted approach can be used for routine analysis of larger sample sizes, in which a selected set of genes of interest is amplified using multiplex PCR (Fig. 2b).

Generally, Illumina® paired-end sequencing is used as a read out for the gene identity, CL, and UMI sequences. Sequencing data is then processed using bioinformatics pipelines to decode the CL and UMI and adjust for PCR or sequencing errors, before assigning each read to the cell and transcript molecule of origin. The end result is a data matrix, containing counts of transcripts per gene per cell. The data matrix can then be used for various kinds of clustering analysis and high-dimensional visualization, such as t-stochastic embedded neighboring (t-SNE) (van der Maaten and Hinton 2008).

To demonstrate successful single-cell capture by the system, a 1:1 mixture of human 293T cells and mouse NIH/3T3 cells was loaded at a range of cell densities and assayed using whole transcriptome profiling. At the current recommended operating range of ≤20,000 cells, high single-cell purity can be achieved since the proportion of wells filled with cells is low (Fig. 3). The lower the cell number input, the higher the single-cell purity that is achieved; for example, at a cell load density of ~1000 cells, no multiplets can be identified in the experiment (Fig. 3b). Multiplets are events where two or more cells are captured by a single BD Rhapsody Cell Capture Bead. The occurrence of multiplets increases as more cells are loaded into the BD Rhapsody Cartridge, following Poisson distribution calculations.

The system also yields low inter-cell noise, exemplified by extremely low mouse signal detected by human cells, and vice versa (Fig. 3a). The overall cell capture rate from cell loading to sequencing is ~65%, and can vary due to the use of different cell types and user handling. The system is shown to produce reproducible gene expression and population profiles ranging from 100 to 10,000 cells (Fig. 4).

Characteristics of the BD Rhapsody System

Several characteristics of the BD Rhapsody system distinguish it from other single-cell analysis systems (Wu et al. 2013; Klein et al. 2015; Macosko et al. 2015; Zheng et al. 2017):

1. *Does not use microfluidics.* The microwell array is housed in a fluidic cartridge that has a volume of ~600 μL, which is the minimum volume of the cell suspension to be loaded. This implies that the cell suspension loaded can be rather dilute and that concentrating the cell suspension (which is often required in microfluidic- or microdroplet-based systems) is not necessary. Additionally, because no microfluidic channels are involved, clogging of channels with cells is less of a concern. The interior of the cartridge is specially treated to prevent cells from adhering to the surfaces.

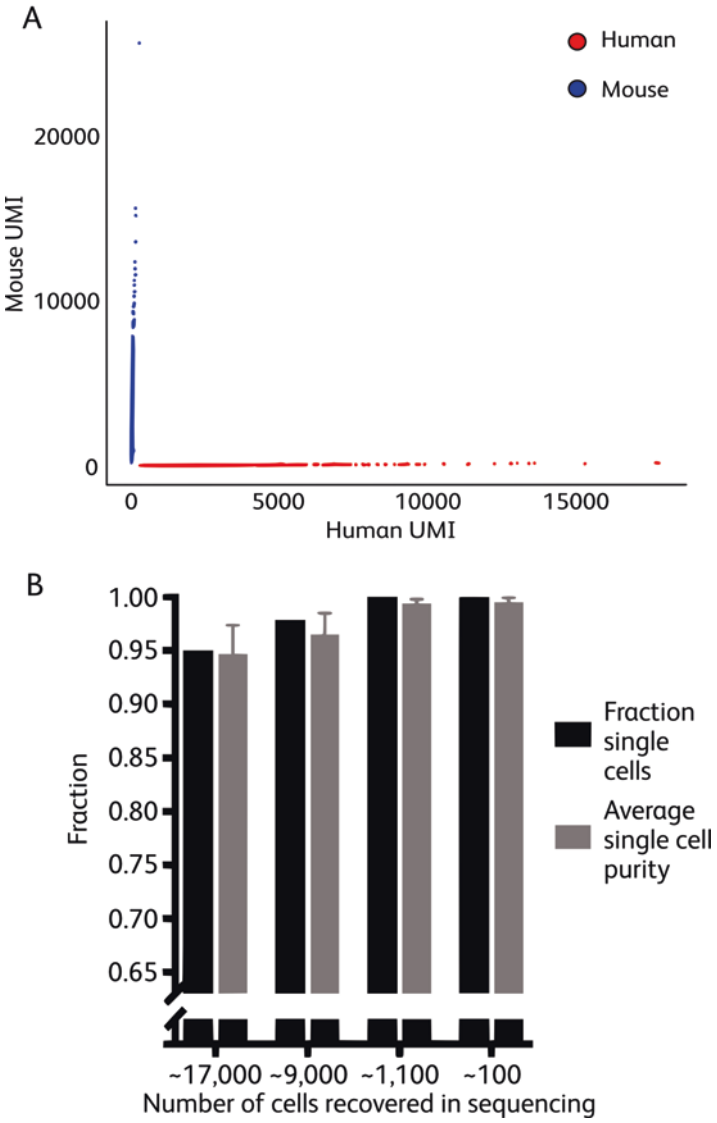

Fig. 3 Multiplet rate and single-cell purity demonstrated by the BD Rhapsody system. Mixtures of 1:1 human 293T and mouse NIH/3T3 cells were loaded onto a cartridge at different concentrations. cDNA molecules were universally amplified and sequenced shallowly (~8700 reads per cell). (**a**) At the loading condition in which 1109 cells were detected in sequenced data, an average of 99.4% of molecules from each cell were identified as either only human, or only mouse (i.e. purity), indicating minimal molecular crosstalk between cells. (**b**) Observed cell multiplet rate and purity at various loaded cell densities of 1:1 human 293T and mouse NIH/3 T3 cells

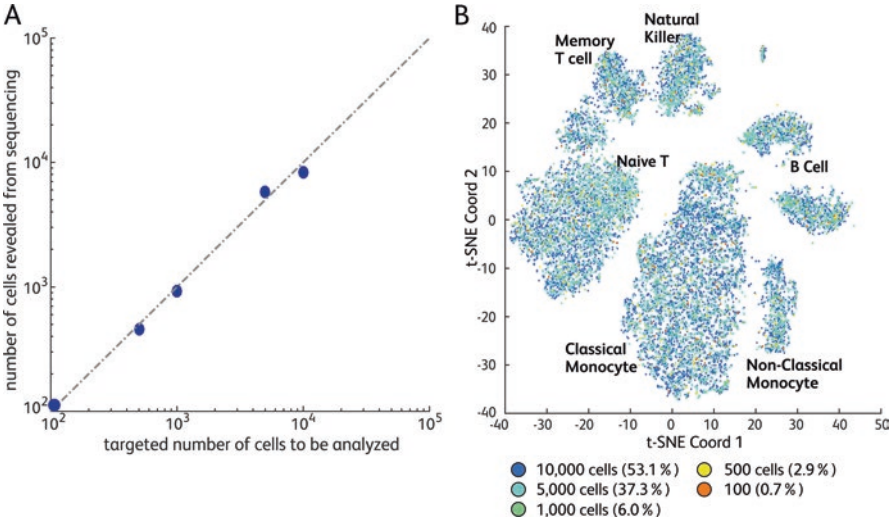

Fig. 4 Demonstrating dynamic range of input cell number on the BD Rhapsody system. (**a**) Human PBMCs, ranging from 100 to 10,000 cells, were profiled using a panel of ~400 genes designed to analyze human immune cells. The number of cells recovered from sequencing correlated closely to that of the expected number based on number of cells loaded. (**b**) Known PBMC cell types with similar expression profiles were consistently identified across a wide range of sample-size inputs

2. *Requires a minimal set of benchtop equipment.* Fluidic exchange in the BD Rhapsody Cartridge is performed using a set of automated pipettes on a mechanical station for magnet operations and fluid collection. Unlike microfluidic- or microdroplet-based systems, no fluidic pumps are required.
3. *Ability to visualize single-cell capture.* The BD Rhapsody Cartridge has an optically clear window, enabling visual inspection of the contents of the cartridge and of each microwell. The BD Rhapsody Scanner can be used to provide quality control measures at different stages of the workflow, such as cell-capture rate and cell-multiplet rate, by direct imaging. These measures can be useful for troubleshooting and can provide users some expectation of the number of cells recovered by sequencing. Additionally, errors that occur during the cartridge workflow are reflected by abnormal quality control metrics, enabling users to decide whether to proceed to sequencing the libraries, as sequencing can be expensive.
4. *BD Rhapsody beads can be retained for later experimentation.* The BD Rhapsody beads remain intact throughout the workflow and cDNA molecules can be stored for several months on the beads. One can therefore subsample beads for creation of multiple sequencing libraries. For instance, a user can first capture mRNA from ~10,000 cells from a precious sample, create a library from ~1000 cells using ~10% of the beads to study one set of gene targets, and later revisit the archived beads to prepare a new library from another subset of cells to analyze a different set of gene targets.

5. *The platform is flexible and compatible with various downstream molecular biology workflows.* The design of the cartridge enables various reagents and buffers to be loaded. Bead-bound cDNA is amenable to standard molecular biology manipulations. Purifications and washes after each molecular biology step are relatively straightforward using benchtop tube magnets. These properties can potentially lead to new applications from creative users.

Sample Multiplexing on BD Rhapsody

The BD Rhapsody cartridge is designed to process a single sample of cell suspension. However, there are situations in which it is desirable to analyze multiple samples at a time; for example, a user might want to compare various test conditions versus a control. Instead of performing an experiment over multiple BD Rhapsody cartridges, an alternative is to use the BD™ Single-Cell Multiplexing Kit to differentiate multiple samples within a single cartridge (Fig. 5a).

The technology uses a prescreened antibody that targets a universally-expressed cell-surface antigen across multiple tissues and cell types. The same antibody is conjugated to one of 12 sample tags (STs). Each ST is a unique 45-nucleotide barcode sequence flanked by a universal PCR handle and poly(A) tail that allows each ST to be captured by oligo-dT beads, such as the BD Rhapsody Cell Capture Bead, and then amplified by PCR using the ST universal PCR handle region (Fig. 5b).

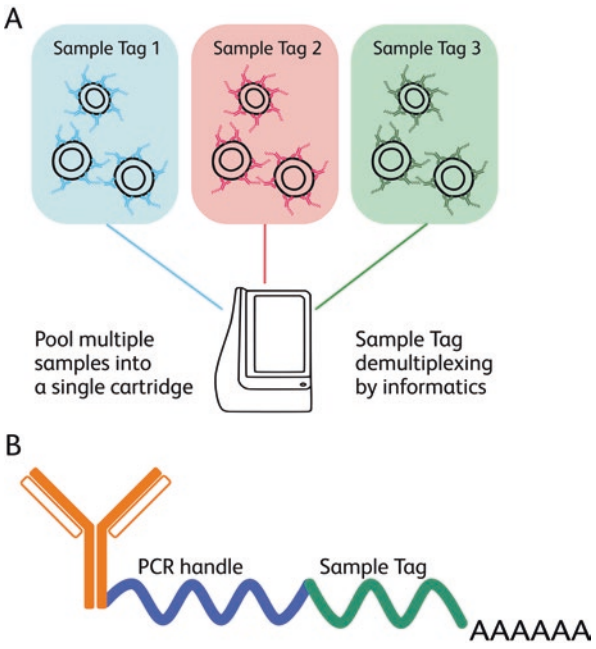

Fig. 5 Sample multiplexing on BD Rhapsody using Sample Tags (STs). (**a**) Workflow of upstream ST labeling prior to BD Rhapsody cartridge loading. (**b**) Structure of a ST and its compatibility for capture by a BD Rhapsody Cell Capture Bead via the poly(A) tail to mimic polyadenylated RNA transcripts

In addition to sample calling, STs allow identification of a subset of multiplets in a single-cell sample that have >1 ST associated to the same CL. Even though BD Rhapsody microwell technology provides low multiplet occurrence compared to conventional droplet-based technologies, the user can still make a decision to limit the number of cells loaded into a cartridge to maintain a low multiplet rate. In some cases, multiplets are not easily identifiable and can be misinterpreted as biologically meaningful.

The labeling of samples using ST is analogous to standard antibody staining prior to loading into the BD Rhapsody or comparable single-cell 3′ RNA-seq systems. After library preparation of the ST library, it can be combined with the RNA-based library into a single sequencing run. Sample origin is identified after sequencing using the BD Rhapsody pipeline with the sample multiplex module. This technology greatly enhances sample throughput and flexibility in experimental design, while reducing technical errors between samples.

BD AbSeq Technology: Utilizing Oligonucleotide-Conjugated Antibodies for High Parameter Protein Profiling

The emerging use of antibody-oligonucleotides in single-cell sequencing enables dual measurement of mRNA and protein expression in each cell. This *multi-omic* analysis (also called BD AbSeq) builds on top of existing 3′ single cell RNA-seq capture systems by conjugating a polyadenylated antibody-specific barcode (ABC) onto an antibody (Shahi et al. 2017; Stoeckius et al. 2017; Peterson et al. 2017). After antibody-labeling of cells, these ABCs act as RNA mimics and are captured in the same manner as cellular polyadenylated RNAs (Fig. 6). The ease-of-use and simple companion workflow to the BD Rhapsody system allows simultaneous profiling of RNA and protein together.

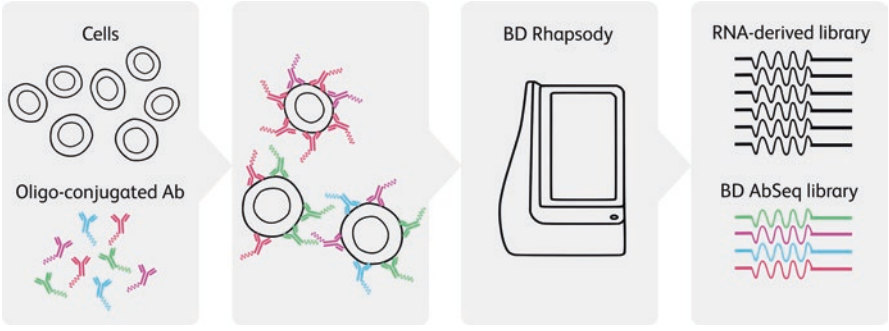

Fig. 6 BD AbSeq workflow. Current workflow utilizes a conventional antibody labeling step prior to BD Rhapsody cartridge loading, followed by a split library preparation workflow to prepare for sequencing

BD AbSeq technology enables researchers to understand single cells in a more sophisticated and in-depth manner than standard antibody-based assays such as flow cytometry and mass cytometry. In flow cytometry, where antibodies are conjugated to fluorophores for single-cell analysis, the number of antibody parameters are limited by cytometer hardware configurations, as well as the intrinsic limitation of the color spectrum (also called spectral overlap). In recent years, the continued improvement of dyes and cytometers, such as the BD FACSymphony™ flow cytometer, has pushed the upper limit of parameters to 30–50 per sample. Moreover, mass cytometry instruments are also available to assay a similar number of parameters using antibodies conjugated to isotypically pure elements. While mass cytometry overcomes the spectral overlap of flow cytometry, the catalog of elements available for antibody conjugation and discernable by mass cytometry still appears to be limited. Most importantly, the ability to profile gene expression at the protein-only level is limited by the availability of highly specific antibodies to the antigen of interest. BD AbSeq bridges this gap by profiling RNA expression as well, through targeted panels of hundreds of genes, or the whole transcriptome.

The ability to profile both mRNA and proteins allows unprecedented resolution in the understanding of cellular mechanisms. While genetic information encoded by DNA is transcribed to RNA, followed by translation to protein (Crick 1958), this flow of information can be affected by complex molecular mechanisms governed by cell-type and cell-state regulation pathways. For example, post-transcriptional mechanisms governed by microRNAs can affect target RNA stability and/or translation efficiency (Bartel 2009). Moreover, protein turnover can also affect how genes are dynamically regulated (Hochstrasser 1995). As a result, the relationship between the expression levels of mRNA and its corresponding protein do not always correlate (Gygi et al. 1999), requiring the need to pursue technologies that can profile both arenas of gene expression. The BD AbSeq assay allows researchers to scrutinize intricate gene-regulation patterns in single cells, paving the way to understand intricate biological systems.

Demonstrated Use of the BD Rhapsody System and Companion Technologies

Whole Transcriptome Profiling

The BD Rhapsody system has been used to analyze mammalian cells from various sources, such as blood, lung lavage fluids, and dissociated tissues and tumors. Generally, any cell suspension that is prepared using processes suitable for flow cytometry (except for fixation methods that degrade mRNA, such as formaldehyde) can be studied with this system. For example, Birey prepared single-cell whole transcriptome libraries from dissociated cortical spheroids and subpallium spheroids derived from differentiation of human induced pluripotent stem cells (Birey et al. 2017). The authors used the BD Rhapsody system to profile single cells to verify the correct neuronal subtypes in these spheroids.

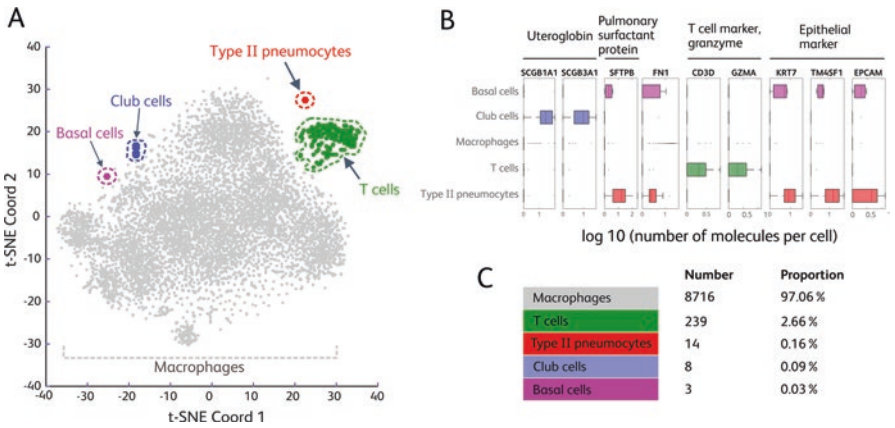

Fig. 7 Analysis of a human lung lavage fluid sample on BD Rhapsody using whole-transcriptome analysis. (**a**) t-SNE projection showing various cell types detected. (**b**) Boxplot showing expression levels of various cell type-specific genes across the detected cell populations (numbered the same as in **a**). (**c**) Proportions of the identified cell types

In another example, cells from human lung lavage fluids were analyzed to understand the functional state of the epithelial cells and their potential association with various lung diseases. These cells were rare, fragile, and lack robust cell-surface markers, making enrichment of these cells by flow sorting challenging. The use of massively parallel single-cell gene expression analysis was especially useful, as the analysis of large numbers of cells enabled sampling of rare cells without extensive sample manipulation or enrichment beforehand. By studying ~9000 cells on the BD Rhapsody system, various epithelial cell types were detected, including club cells, basal cells, and pneumocytes (Fig. 7a, b). Each of these rare cell types constituted as few as 0.03% of the entire population (Fig. 7c). Unlike flow cytometry, the detection of a rare but distinct population requires only a small number of cells, especially if the rare cells co-express very specific combinations of highly abundant marker genes.

Targeted Sequencing in Single Cells

In the two cases above, the whole transcriptome approach was used because relatively little was known about the gene-expression profiles of the cell populations in the sample types. These exploratory studies provide a shallow survey of population architecture, but tend to have lower sensitivity because the preparation of whole-transcriptome sequencing libraries requires multiple molecular biology steps, with each step having non-ideal efficiency. Moreover, sequencing whole transcriptomes to saturation (that is, to observe all transcript molecules in sequencing data at least once) requires as many as 100,000 sequencing reads per cell, making these analyses

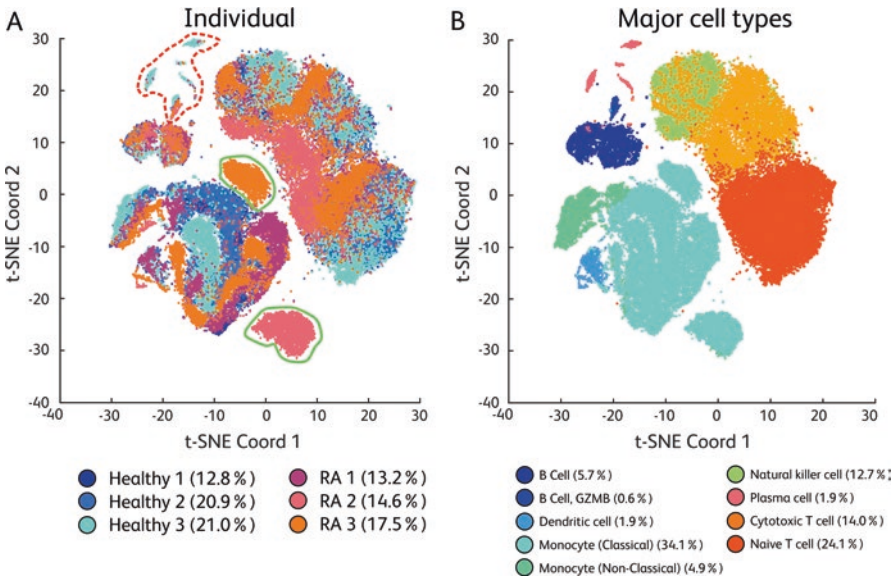

Fig. 8 Analysis of PBMCs from 3 healthy donors and 3 patients with rheumatoid arthritis (RA) using a panel of 454 targets. Gene expression profiles of ~10,000 cells per sample were analyzed. (**a**) t-SNE projection of a total of ~60,000 cells from all 6 samples. Red dotted circle indicates presence of large number of plasma cells specific to one of the donors. Green solid circles indicate distinct population of classical monocytes of two of the patients. (**b**) t-SNE projection colored by the major immune cell populations

not very scalable nor cost-effective. In many biological systems, less than 1000 genes of the entire transcriptome (which has >50,000 annotated genes in humans) are variable and are responsible for driving clustering of cellular expression profiles (Macosko et al. 2015; Birey et al. 2017). Most abundant housekeeping genes are not variable but they occupy much of the sequencing space (Wassarman 1995).

Many users may find a targeted approach (focusing on a particular set of genes) to be more sensitive and cost-effective for analyzing a large number of samples. For instance, the whole-transcriptome data from the above two cases can be used to derive lists of population-specific mRNA markers for a targeted panel. Additional genes of interest that might not be detected well in whole-transcriptome analysis but warrant deeper analysis due to their biological significance can be added to the panel.

As an alternative to whole-transcriptome profiling, a multiplex PCR panel of 454 targets was designed for BD Rhapsody beads to test in human peripheral blood mononuclear cells (PBMCs), a very well-characterized, heterogeneous biological system. We used this assay to study the gene expression profiles of ~10,000 single PBMCs each from a set of 6 individuals comprising 3 healthy donors and 3 patients with rheumatoid arthritis (RA) (Fig. 8). In addition to measuring proportions of major immune populations, in which we found unusual elevation of plasma-cell populations in one of the healthy donors (Fig. 9a), single-cell gene expression profiling

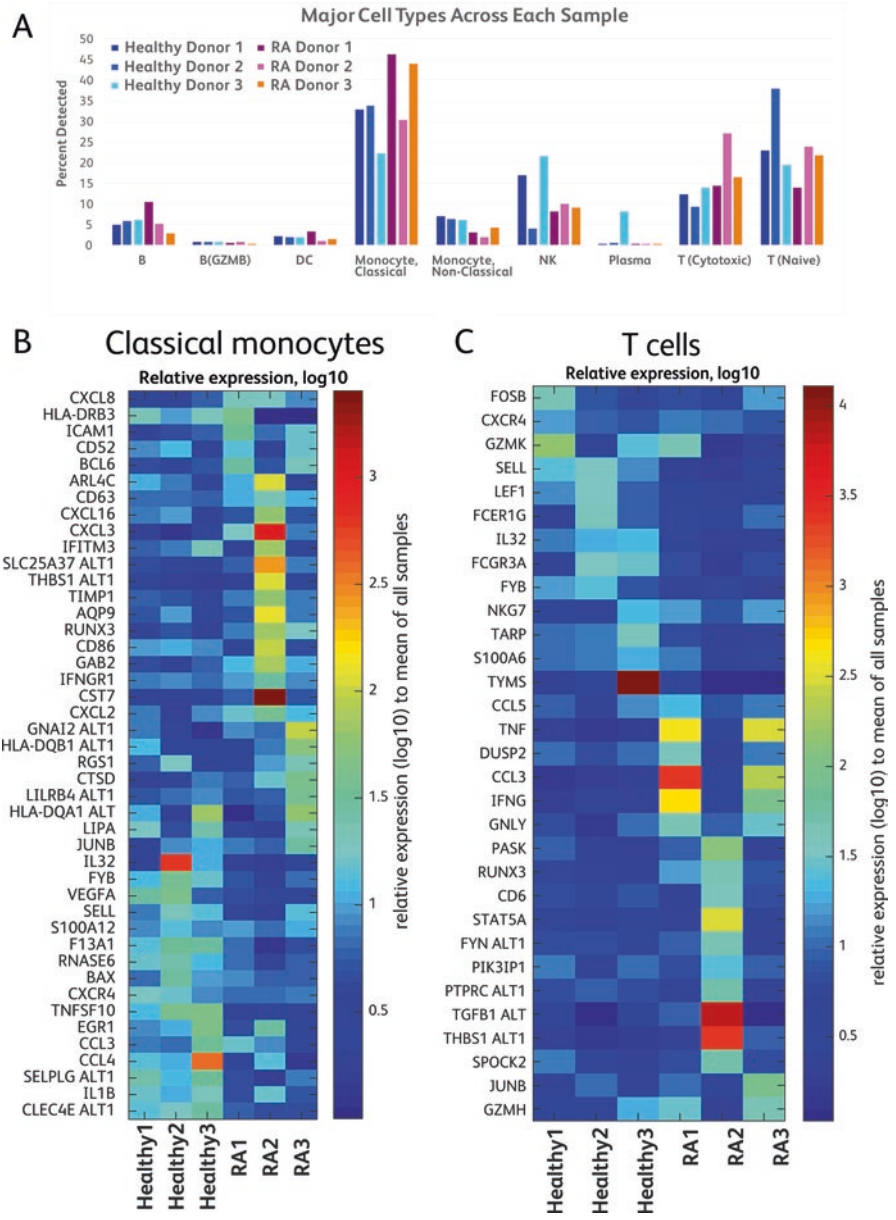

Fig. 9 (**a**) Proportion of each major immune cell type in each sample. Heatmaps showing relative levels of selected differentially expressed genes across the six individuals (three healthy donors and three RA patients) in classical monocytes (**b**) and T cells (**c**)

allowed for differential gene-expression analysis within each immune population across individuals (Fig. 9b, c). In particular, we found expression patterns in classical monocytes and T cells specific to each of the RA patients. This method of high-throughput single-cell analysis of blood cells can potentially reveal many more biological insights than microarray or RNA-seq analysis of bulk RNA from blood, especially when further samples are analyzed to provide statistical power.

Sample Multiplexing and Multiplet Identification Using STs

In addition to increasing sample throughput, another advantage of using STs is that one can load a higher number of cells while maintaining a low rate of unidentified multiplets. To illustrate this capability, a dataset of four sample types –PBMCs, Ramos B cells, Jurkat T cells, and T47D breast cancer cells – split across 12 STs (Fig. 10a) was used to demonstrate the assay's ability to identify multiplets, which are cell labels associated with more than one ST (Fig. 10b, c). In this ~20,000 cell experiment, the theoretical multiplet occurrence by Poisson statistics was ~4.7%; using STs, 4.3% of the putative cells were identified as multiplets (Fig. 10b, c). When mRNA expression profiles were projected using t-SNE and overlaid with the ST annotation, many of the multiplets identified by the ST determination algorithm resided in small clusters between the major cell populations (Fig. 10b). These small clusters expressed gene markers from more than one cell type (not shown in this publication), thereby validating the use of STs for multiplet identification. In addition, the assay using these demonstration cell types achieved 98.6% sensitivity and >99% specificity, as defined by percentage of cells positive for ST detection and percentage of cells assigned to the correct ST, respectively.

Simultaneous Analysis of Protein and mRNA Expression in Single Cells

Using BD AbSeq assays on the BD Rhapsody system, protein and mRNA content of PBMCs from different healthy donors were measured. To accomplish this, a high-parameter oligo-conjugated antibody panel against immune-relevant cell-surface markers was paired with a targeted gene expression panel consisting of 399 mRNA targets. PBMCs from both donors were prepared in the same workflow using the BD Single-Cell Multiplexing Kit to minimize technical errors and reduce library preparation cost. The combination of protein and mRNA profiling allows flexible data analysis options, exemplified by the ability to perform high dimension t-SNE visualization by mRNA only, protein only, or both (Fig. 11). While mRNA and protein-only t-SNE provided good distinction of cell types within these PBMCs, t-SNE driven by both protein and mRNA (Fig. 11c) provided the most robust

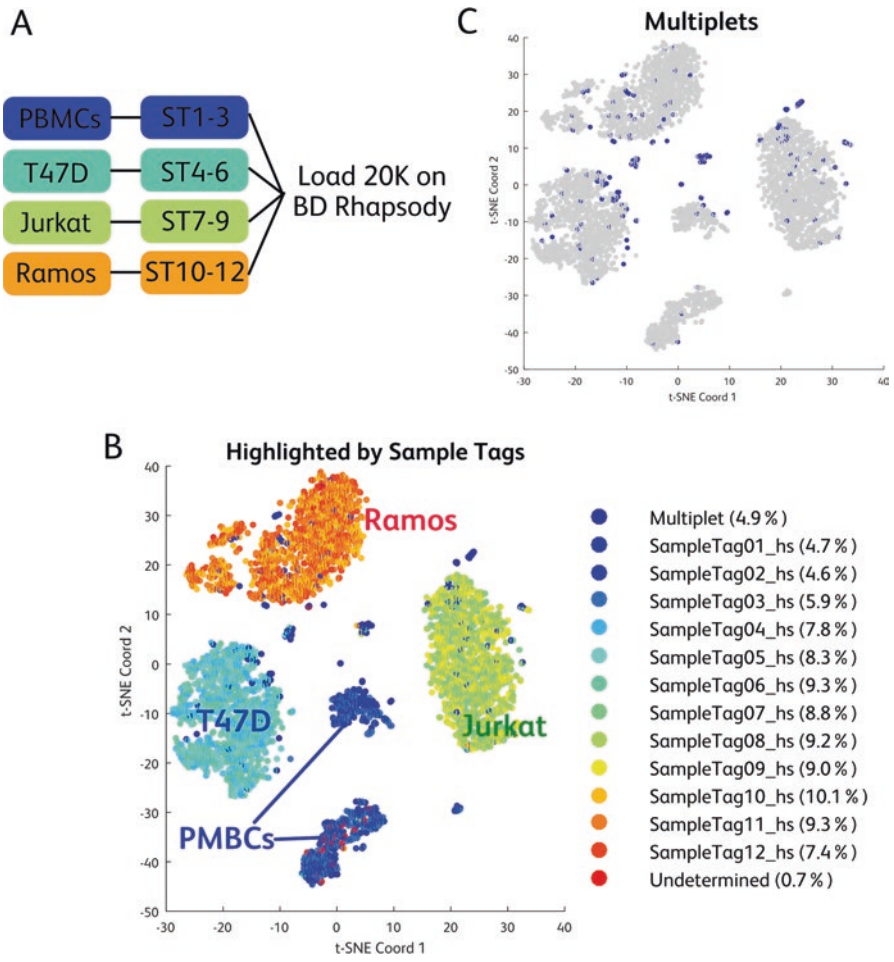

Fig. 10 The use of STs in single-cell analysis to increase sample throughput and detect cell multiplets. (**a**) Schematic of the 12-ST experiment with 20,000 cells of mixed sample types loaded on the BD Rhapsody. ~10,000 cell-equivalent BD Rhapsody beads were used to generate libraries for sequencing analysis. (**b**) t-SNE visualization of cell clusters using 12 STs and the BD Rhapsody Immune Response Panel. Annotation of major cell types identified by their ST call with >99% specificity, defined by the percentage of STs calling to the right cell type based on cell-specific mRNA markers (not shown here). (**c**) Highlight of multiplets identified using ST-algorithm, which are single cells that are assigned to two or more STs. In addition to mixed-cell multiplets, those formed by the same cell types can also be identified using Sample Tags even though they are embedded within main clusters

distinction between critical cell types such as naïve and effector T cells. When comparing multiple PBMC donors, we were able to distinguish expression profile differences between some populations, such as classical monocytes (Fig. 12). Together, BD Rhapsody and its companion technologies allow high resolution analysis of complex biological samples.

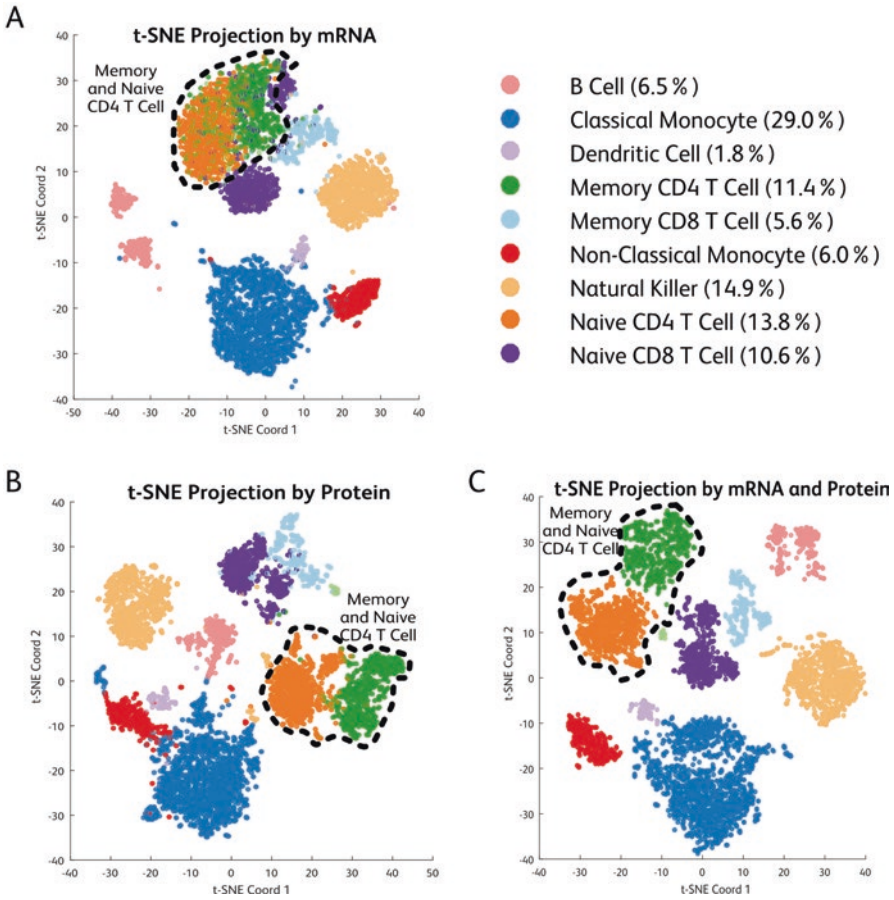

Fig. 11 Application of BD AbSeq assays on the BD Rhapsody system using PBMCs from two healthy donors with major cell types highlighted by different colors. (**a**) t-SNE projection using BD Rhapsody Immune Response Panel (RNA) only; while major immune cell types can be identified, similar cell types – such as memory and naïve CD4 T cells – are not distinctly separated. (**b**) t-SNE projection using BD AbSeq (protein) profiles only, showing increased separation between memory and naïve CD4 T cells. (**c**) t-SNE projection using both, BD Rhapsody Immune Response Panel and BD AbSeq profiles, providing clear distinction between similar cell types, including memory and naïve CD4 T cells and more

Summary

BD Rhapsody is a single-cell analysis system that allows high-throughput capture and library preparation of single cells. The system utilizes a microwell technology that results in high purity of single cells and a low multiplet rate, providing cleaner single-cell analysis. Moreover, the multilevel barcoding scheme of CLs and UMIs allows simple downstream workflow and the ability to digitally quantify expression

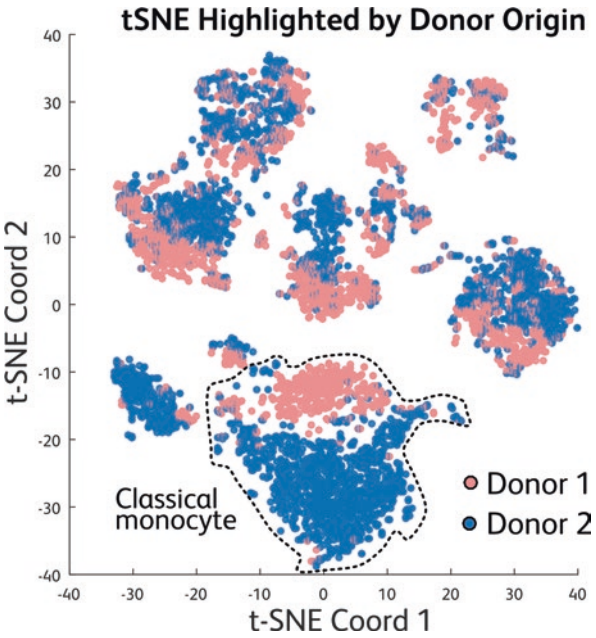

Fig. 12 t-SNE projection using both BD Rhapsody Immune Response Panel (RNA) and BD AbSeq (protein) profiles. Using ST technology, multiple donor samples can be pooled together to perform BD Rhapsody and BD AbSeq assays in a single cartridge, providing low technical variability. Pooled library preparation and sequencing reveal differences in classical monocyte profiles between healthy donor 1 and healthy donor 2, providing insights into single-cell heterogeneity between different biological samples

of gene transcripts. With emerging technologies such as the BD AbSeq assays, BD Rhapsody can be used for single-cell multi-omics profiling and pave the way for a new generation of understanding of gene regulation and biological heterogeneity, and potentially lead to new clinical applications.

BD FACSymphony™ is a Class 1 Laser Product. BD products described here are For Research Use Only. Not for use in diagnostic or therapeutic procedures.

References

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–33. https://doi.org/10.1016/j.cell.2009.01.002.

Birey F, Andersen J, Makinson CD, et al. Assembly of functionally integrated human forebrain spheroids. Nature. 2017;545:54–9. https://doi.org/10.1038/nature22330.

Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.

Fan HC, Fu GK, Fodor SPA. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015;347:1258367. https://doi.org/10.1126/science.1258367.

Fu GK, Hu J, Wang P-H, Fodor SPA. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci. 2011;108:9026–31. https://doi.org/10.1073/pnas.1017621108.

Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999;19:1720–30.

Hochstrasser M. Ubiquitin, proteasomes, and the regulation of intracellular protein degradation. Curr Opin Cell Biol. 1995;7:215–23. https://doi.org/10.1016/0955-0674(95)80031-X.

Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Peterson VM, Zhang KX, Kumar N, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35:936–9. https://doi.org/10.1038/nbt.3973.

Shahi P, Kim SC, Haliburton JR, et al. Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci Rep. 2017;7:44447. https://doi.org/10.1038/srep44447.

Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14:865–8. https://doi.org/10.1038/nmeth.4380.

van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

Wassarman PM. Advances in developmental biochemistry, vol. 3B. Amsterdam: Elsevier; 1995.

Wu AR, Neff NF, Kalisky T, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2013;11:41–6. https://doi.org/10.1038/nmeth.2694.

Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. https://doi.org/10.1038/ncomms14049.

An Informative Approach to Single-Cell Sequencing Analysis

Yukie Kashima, Ayako Suzuki, and Yutaka Suzuki

Abstract Recent advances in sequencing technologies enable us to obtain genome, epigenome and transcriptome data in individual cells. In this review, we describe various platforms for single-cell sequencing analysis across multiple layers. We mainly introduce an automated single-cell RNA-seq platform, the Chromium Single Cell 3′ RNA-seq system, and its technical features and compare it with other single-cell RNA-seq systems. We also describe computational methods for analyzing large, complex single-cell datasets. Due to the insufficient depth of single-cell RNA-seq data, resulting in a critical lack of transcriptome information for low-expressed genes, it is occasionally difficult to interpret the data as is. To overcome the analytical problems for such sparse datasets, there are many bioinformatics reports that provide informative approaches, including imputation, correction of batch effects, dimensional reduction and clustering.

Summary of Single-Cell Sequencing Methods

If we analyze cells in bulk, we tend to overlook subtle molecular profiles and changes in each individual cell because they are buried in major population-level changes. Single-cell analysis can be a strong tool for detecting distinct molecular profiles in individual cells and understanding complex systems in organisms (Fig. 1). The number of papers using single-cell analysis has been rapidly increasing, especially in cancer and developmental biology studies (Wang and Navin 2015). In this section, we introduce a brief summary of single-cell sequencing methods and their applications for various fields.

Y. Kashima · A. Suzuki (✉) · Y. Suzuki
Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan
e-mail: ykashima@edu.k.u-tokyo.ac.jp; asuzuki@edu.k.u-tokyo.ac.jp; ysuzuki@edu.k.u-tokyo.ac.jp

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_6

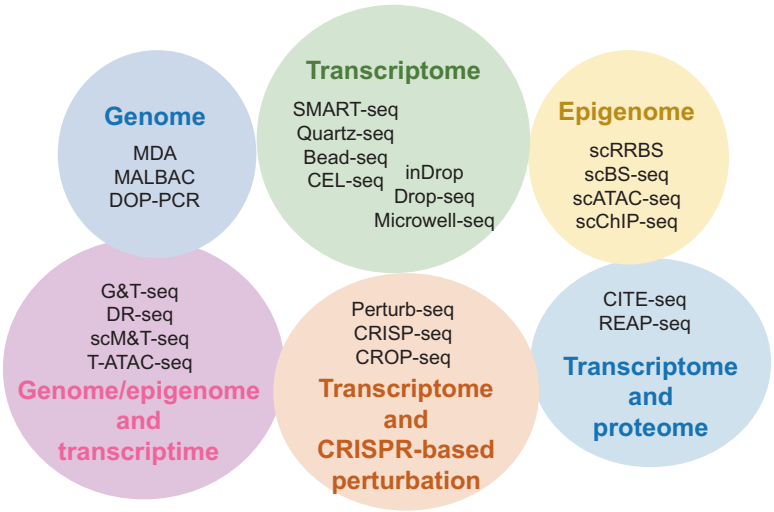

Fig. 1 Summary of methods for single-cell sequencing
Single-cell sequencing methods have been developed to analyze various layers of omics profiles at the single-cell level. There are diverse methods for scRNA-seq. SMART-seq (Ramsköld et al. 2012; Picelli et al. 2013), Quartz-seq (Sasagawa et al. 2013), bead-seq (Matsunaga et al. 2015) and CEL-seq (Hashimshony et al. 2012) are methods for whole-transcriptome amplification, molecular barcoding and library construction. Researchers have also developed inDrop (Klein et al. 2015) as well as Drop-seq (Macosko et al. 2015) and Microwell-seq (Han et al. 2018) for droplet-based and well-based platforms of scRNA-seq, respectively. For single-cell genome sequencing, MDA (Lasken 2007), MALBAC (Zong et al. 2012) and DOP-PCR (Telenius et al. 1992) are common methods of whole-genome amplification. There are also single-cell epigenome sequencing methods, namely, scRRBS (Guo et al. 2013) and scBS-seq (Smallwood et al. 2014) for DNA methylation and scATAC-seq (Buenrostro et al. 2015) and scChIP-seq (Rotem et al. 2015) for chromatin statuses. In an individual cell, we can analyze multi-layered molecular profiles using G&T-seq (Macaulay et al. 2015), DR-seq (Dey et al. 2015) (genome and transcriptome), scM&T-seq (Angermueller et al. 2016) (DNA methylation and transcriptome) and T-ATAC-seq (Satpathy et al. 2018) (open chromatin statuses and TCR sequence). CITE-seq (Stoeckius et al. 2017) and REAP-seq (Peterson et al. 2017) have been developed for analysis of transcriptome and protein statuses in a cell. Perturb-seq (Dixit et al. 2016), CRISP-seq (Jaitin et al. 2016) and CROP-seq (Datlinger et al. 2017) are provided for scRNA-seq with CRISPR-based genetic perturbation screening

Single-Cell RNA-Seq

Currently, we can use various platforms of single-cell RNA-seq (scRNA-seq) to analyze transcriptome profiles in each individual cell. There are several methods in which researchers should manually prepare reagents and instruments. These methods, including Drop-seq (Macosko et al. 2015) and Microwell-seq (Han et al. 2018), make costs reasonable and do not require expensive machines. In contrast, we can also utilize commercialized platforms such as the C1, ICELL8 and Chromium systems. With these platforms, processes including single cell separation, cell lysis, reverse transcription and amplification are almost automated. Even though the initial investment and running costs are high, these platforms could broaden the use of single-cell sequencing analysis for many laboratories and researchers.

The Drop-seq method was reported by the McCarroll lab in 2015. It enables us to analyze the transcriptome profiles of thousands of cells in parallel (Macosko et al. 2015). In Drop-seq, cells are individually separated into droplets with barcoded beads, resulting in scRNA-seq libraries with different barcodes associated with each of the cells. The researchers reported scRNA-seq data of 44,808 mouse retinal cells and identified cell populations that showed distinct transcriptome profiles. On the other hand, Microwell-seq was developed based on microwell arrays (Han et al. 2018). Cells are separated into microwells containing different barcoded beads. Using Microwell-seq, a Chinese research group reported a "mouse cell atlas" to obtain transcriptome profiles of more than 400 k cells in mice and their cellular networks. They could detect 98 major cell clusters and identify cell types/subtypes that were unrecognized in previous studies.

We can also exploit scRNA-seq platforms provided by several commercial companies. The C1 Single-Cell Auto Prep system (C1) is commercialized by Fluidigm. This instrument is one of the automated scRNA-seq platforms based on microfluidics technology. Cells are separated into an individual chamber of the C1 Integrated Fluidic Circuit (IFC). The C1 system also automatically conducts cell lysis, reverse transcription and amplification and finally generates amplified cDNAs of up to 96 cells. ICELL8 (Takara) and Rhapsody (BD) are two other scRNA-seq platforms based on microwells. These microwell-based platforms enable us to monitor the viability of the cells in each well before the construction of scRNA-seq libraries. In addition, there are several scRNA-seq instruments using microdroplet technologies, such as ddSEQ (Illumina/BioRad) and Chromium (10x Genomics). Chromium is one of the most popular scRNA-seq platforms used worldwide. Detailed information about the Chromium system is described below.

Single-Cell Genome Sequencing

Not all cells harbor identical genome sequences, as is the case with the occurrence of mosaicism in somatic/germline cells and accumulation of somatic mutations in cancer cells. To grasp such genomic heterogeneity in a population of cells, we can conduct single-cell analysis of whole-genome/exome sequencing. Focusing on cancers, Navin's lab reported single-cell genome sequencing analysis of breast cancers, mainly to analyze copy number variants (CNVs) in each tumor cell and to characterize cancer evolution (Navin et al. 2011; Wang et al. 2014; Gao et al. 2016).

For single-cell genomics analysis, we need to uniformly amplify genomic DNAs from each cell, termed whole genome amplification (WGA); these methods include multiple displacement amplification (MDA) (Lasken 2007), multiple annealing and looping-based amplification cycles (MALBAC) (Zong et al. 2012) and degenerate oligonucleotide-primed PCR (DOP-PCR) (Telenius et al. 1992). However, there are some technical problems. Amount of DNA samples is limited for single-cell genome sequencing. In addition, allelic dropout is another big problem. This event generally occurs during WGA and results in non-uniform coverage and false positive/negative detection of alterations. In addition to the scRNA-seq application, the C1 system

provides single-cell genome sequencing with automatic cell separation, cell lysis and WGA. Futher, 10x Genomics released an application for sinlge-cell CNV (scCNV) analysis within the Chromium system. Although the number of cells depends on the number of sequencing reads, Chromium can process 5000 cells per cell at most (10x Chromium Single Cell DNA Users Guide). Using Chromium's scCNV and scRNA-seq method, Ador et al reported a study of gastic cancer. They estimated copy number information of individual cancer cells and also used data from peripheral blood mononuclear cells as genome stable diploid cells. They integrated CNV and RNA datasets by their novel method (Ador et al, bioRxiv, 2018). Similar to scRNA-seq, rapid development of protocols, instruments and computational methods is required for successful genome analysis at the single-cell level.

Single-Cell Epigenome Sequencing

Using single-cell sequencing techniques, we can also decipher epigenome profiles in individual cells. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a useful method for identifying regions of open chromatin status (Buenrostro et al. 2013). ATAC-seq identifies nucleosome-bound and nucleosome-free positions in regulatory regions and infers binding patterns of transcription factors (TFs). In 2015, Cusanovich et al. reported ATAC-seq analysis in a large number of single cells (Cusanovich et al. 2015). Buenrosto et al. also reported single-cell ATAC-seq (scATAC-seq) with the C1 system (Buenrostro et al. 2015). Recently, Chromium scATAC-seq is also released. These scATAC-seq analysis give us an understanding of the heterogeneity of transcriptional regulations.

For analyzing DNA methylation in single cells, there are several bisulfite sequencing methods at the single-cell level. Single-cell reduced-representation bisulfite sequencing (scRRBS) is one method for single-cell DNA methylation analysis (Guo et al. 2013). In this method, genomic DNAs are fragmented by restriction enzymes and enriched with CpG-rich regions. Guo et al. analyzed methylation statuses of mouse embryonic stem cells (mESCs) using scRRBS. Smallwood et al. reported single-cell bisulfite sequencing (scBS-seq) (Smallwood et al. 2014). The authors also focused on DNA methylation statuses in each of the individual mESCs to analyze epigenetic heterogeneity in the DNA methylome.

Multi-omics Analysis in Single Cells

There are many papers related to single-cell sequencing protocols in which researchers independently analyze genome, epigenome and transcriptome patterns of each individual cell. In addition, several studies reported using a combination of these single-cell sequencing assays across multiple layers. Here, we focus on single-cell multi-layered analysis.

Genome Sequencing Combined with RNA-Seq at the Single-Cell Level

Cancer research is one of the most important applications of single-cell sequencing due to the complicated microenvironment and intratumor heterogeneity in cancer tissues (Wang and Navin 2015). Kim et al. revealed the evolution of breast cancers by combining datasets from 900 cells' single-nucleus DNA sequencing, 6,862 cells' single-nucleus RNA-seq and bulk exome deep sequencing (Kim et al. 2018). They used frozen tissues from 20 triple-negative breast cancer patients who underwent neoadjuvant chemotherapy (NAC). The authors showed that genomic aberrations associated with resistance to chemotherapy are pre-existing in tumors. Thus, integration of multi-layerted single-cell sequencing would give us new insights that we overlooked in independent studies.

In addition, there are several methods for obtaining genomic and transcriptomic statuses from the same single cell. In G&T-seq reported by Macaulay et al. (2015), RNA and DNA are separated using oligo-dT primers and magnetic beads and sequenced separately. Another research group reported DR-Seq in which DNA and RNA are amplified without separation (Dey et al. 2015).

scATAC-Seq Combined with RNA Sequencing Analysis

Within 3 years of publishing the scATAC-seq paper, the authors reported an integrative analysis of scATAC-seq and scRNA-seq data (Buenrostro et al. 2018). They used the C1 system for scATAC-seq and the Chromium system for scRNA-seq to individually obtain transcriptional profiles of diverse cell types. To understand transcriptional heterogeneity in hematopoiesis, they analyzed TF dynamics and enhancer-gene correlation in the hematopoietic system.

Immune system researchers combined T cell receptor (TCR) sequencing and scATAC-seq (Satpathy et al. 2018). They developed transcript-indexed ATAC-seq (T-ATAC-seq) to simultaneously obtain information about TCR specificity and epigenomic statuses from a T cell. T-ATAC-seq reveals the clone-specific epigenome patterns of T cells.

Epitope Analysis Combined with scRNA-Seq

Proteins are considered primary targets for various applications, for example, drug development. Hence, unbiased detection of proteins is needed. Recently, two independent groups published papers describing tools to detect both mRNA and proteins simultaneously at the single-cell level.

Stoeckius et al. developed a new method, CITE-seq, to describe the transcriptome profiling together with cell-surface protein levels (Stoeckius et al. 2017). To recognize epitopes, they employed antibody-derived tags (ADTs). According to the authors, CITE-seq is compatible with most of the known scRNA-seq platforms, including 10x Genomics Chromium. Another group, Peterson et al., published a

paper introducing the RNA expression and protein sequencing assay (REAP-seq) (Peterson et al. 2017). REAP-seq makes it possible to quantify proteins and mRNAs simultaneously at single-cell level. In the paper, they measured proteins with 82 antibodies and RNA expressions of more than 20,000 genes. These two methods differ in how the antibody conjugates the DNA barcode. Compared to CITE-seq, REAP-seq uses small and stable covalent bonds between them (Peterson et al. 2017).

Chromium; One of the Most Commonly Utilized Platforms for Single-Cell Analysis

In 2016, 10x Genomics released a platform for single-cell RNA-seq analysis, the Chromium Single Cell 3′ RNA-seq. This platform can analyze 1200–6000 cells per sample and eight samples simultaneously. The updated version, v2, makes it possible to process 500–10,000 cells per sample at a time. Using this platform, researchers can analyze samples in single-cell level. 10x Genomics also announced a version upgrade of scRNA-seq, Chromium Single-cell 3' Solution version 3. Chromium scRNA-seq becomes to be able to process 500–80,000 cells per sample (10x Genomics webpage).

A Brief Summary of the Chromium scRNA-Seq Platform

Chromium has three components for making droplets: gel beads, cell suspension with RT reagent, and partitioning oil (10x Genomics Single Cell 3′ Reagent Kits v2 User Guide). These three components are poured into a microfluidic plate with eight channels. It enables us to process eight samples at most. In the microfluidic plate, cells are divided into each micro-droplet containing a gel bead and reagents in the instrument (Fig. 2). This process takes only about 7 min. Gel bead includes primers with sequencing adaptors, cell barcode (CB) and unique molecular identifier (UMI). CB is unique to each cell and a UMI is a 10 bp randomer associated with each mRNA molecule. After sequencing, each cell is identified by the CB and UMI.

After this procedure, we obtain thousands of GEMs (Gel bead in Emulsion). In each GEM, an initial cell lysis process is followed by binding the 3′ end of the mRNA to a barcoded bead using 30 bp oligo dTs. Then, reverse transcription starts. After the first strand is synthesized, template switching begins with a template switching oligo (TSO). GEMs are broken, and then, barcoded cDNAs are amplified by PCR in bulk. Following fragmentation, end-repir, A-tailing, adaptor ligation and sample index PCR, a constructed library is loaded into a next generation sequencing platform.

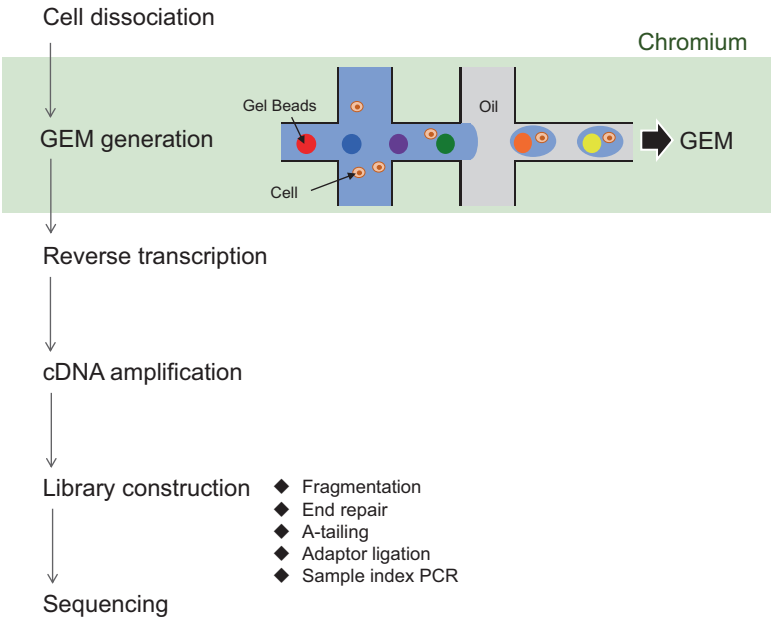

Fig. 2 Chromium single cell 3′ RNA-seq
A brief workflow of single-cell RNA-seq using the Chromium Single Cell 3′ RNA-seq protocol. Cells are dissociated by a user-provided protocol depending on tissue/cell types. Gel Bead-In-Emulsions (GEMs) are created by the Chromium system, which includes dT primers with a cell barcode and unique molecular identifiers (UMIs), reagent mix for reactions, and a cell. After GEM generation, reverse transcription and cDNA amplification are performed, and the amplified cDNAs are quantified by BioAnalyzer (Agilent). For sequencing analysis, sequence libraries are constructed with fragmentation, end repair, A-tailing, adaptor ligation and PCR amplification. The constructed sequence libraries are sequenced by Illumina sequencers

scRNA-Seq Papers Using Chromium

Confirmation of the Chromium Single Cell 3′RNA-Seq Platform

Almost 3 years have already passed since 10x Genomics released Chromium. Hence, there are some papers published using the Chromium scRNA-seq system. The first paper to report using Chromium was Zheng et al. (2017b). In this paper, they conducted a technical demonstration and validation of the sensitivity and ability of Chromium. According to the paper, Chromium can capture 50% of loaded cells. For the microdroplet-based protocols, including Chromium, doublet rates truly matter. Using a mixture of mouse and human cell lines, they observed that the multiplet rate was 1.6% when they obtained 1012 GEMs. Next, they tested its ability to distinguish distinct cellular populations and revealed heterogeneity of healthy donors' peripheral blood mononuclear cells by clustering analysis. In this study, they used the pipeline called Cell Ranger, which was developed for single-cell

RNA-seq datasets produced by Chromium. In conclusion, they demonstrated that Chromium enables high-throughput scRNA-seq.

Comparison of Chromium with Other scRNA-Seq Platforms

There are several papers comparing Chromium to other single-cell platforms. Svensson et al. highlighted the technological developments of scRNA-seq (Svensson et al. 2017). Based on the External RNA Controls Consortium (ERCC) spike-in standards, they computationally compared 15 protocols and experimentally compared four protocols, including CEL-seq (Hashimshony et al. 2012), Chromium (Zheng et al. 2017b), inDrop (Klein et al. 2015) and Drop-seq (Macosko et al. 2015). To define the sensitivity, they used the minimal number of input RNA molecules required for detection. They also defined accuracy based on the input molecules and showed that the accuracy of scRNA-seq protocols is generally high and that their sensitivity depends on sequencing depth, even though the comparison using ERCC spike-in is not perfect.

Our group also attempted to reveal differences between Chromium Single-cell 3' version 1 and another scRNA-seq platform (Suzuki et al. 2015; Kashima et al. 2018). It is sometimes difficult to detect gene expression of low-expressed genes in Chromium due to insufficient depths. In contrast, bead-seq (Matsunaga et al. 2015), which generates datasets similar to those from C1, generates enough sequencing reads but is limited by the number of cells that can be processed at one time. As a conclusion, we suggested combining datasets from the two scRNA-seq protocols to complement the datasets with each other.

Screening Analysis Using Chromium

Recently, some papers related to clustered regularly interspaced short palindromic repeats (CRISPR) screening combined with Chromium have been reported. Combining scRNA-seq and pooled CRISPR-based perturbation, Dixit et al. developed a novel genetic screening method called Perturb-seq (Dixit et al. 2016). To produce perturbations, cells are infected with a pool of lentiviral constructs encoding single-guide RNAs (sgRNAs). The authors used a CRISPR lentiviral vector that delivers an sgRNA to a cell and reports the identity of the sgRNA by expressing the guide barcode. They also developed a computational framework, Multi-Input-Multi-Output-Single-Cell-Analysis (MIMOSCA). Using Perturb-seq and MIMOSCA, the authors demonstrated the effect of TFs on genes, programs and states in the immune cell lipopolysaccharide (LPS) response. Another group published CRISP-seq, which is an integrated method for scRNA-seq and CRISPR screening (Jaitin et al. 2016).

Summary of Chromium

1. Chromium is an automated platform for single-cell sequencing. For capsuling, it takes less than 10 min in total. In addition, the results from this platform are rarely affected by technicians' skill.
2. Chromium enables us to process thousands of cells per sample and eight samples in parallel at one time.
3. Chromium is easy to handle, but the cost is high.

Computational Approach for Single-Cell Analysis

With the rapid spread of single-cell analysis in various research areas, issues with data processing have emerged due to the large and sparse datasets. The equipment makers provide their original bioinformatics tools, such as Cell Ranger for Chromium. However, we need to extract distinct additional information for its application in each new study. To overcome these problems, various bioinformatics reports have introduced new ways of interpreting single-cell datasets. Here, we focus on computational methods, especially for scRNA-seq datasets, including tools for basic analysis of scRNA-seq, imputation of missing values and trajectory analysis (Table 1).

General Analysis of scRNA-Seq Data

After sequencing scRNA-seq libraries and computational primary analysis (mapping and read count), we obtain a gene-cell matrix with UMI (or read) counts. Recently, computational biologists have developed many useful methods and tools to analyze these data. Dimensionality reduction is one of the essential methods for high-dimensional data such as single-cell expression matrices. It is conducted by several methods such as principal component analysis (PCA). This step is followed by clustering analysis for inferring cell types and groups and differential expression analysis to identify differentially expressed genes (DEGs) or marker genes. The scRNA-seq data are often visualized into two dimensions by t-distributed stochastic neighbor embedding (t-SNE) or viSNE based methods (Amir et al. 2013). In this section, we introduce four tools for computational analysis.

Seurat

Satija et al. introduced Seurat in 2015. In this paper, Seurat inferred cellular localization by combining scRNA-seq with in situ RNA-patterns. In 2015, Macosko et al. used Seurat for interpreting scRNA-seq datasets produced by

Table 1 A list of computational approaches for scRNA-seq data

Category	Method/tool	Description and references
General analysis	Seurat	A commonly used R toolkit (Butler et al. *Nat Biotechnol.* 2018)
	SC3	Unsupervised clustering (Kiselev et al. *Nat Methods.* 2017)
	RCA	Semi-supervised clustering (Li et al. *Nat Genet.* 2017)
	SCDE	Differential expression analysis (Kharchenko et al. *Nat methods.* 2014)
Dimensionality reduction and imputation for zero-inflated datasets	ZIFA	Dimensionality reduction considering zero counts (Pierson et al. *Genome Biol*. 2015)
	ZINB-WaVE	Dimensionality reduction of zero-inflated scRNA-seq datasets (Risso et al. Nat Commun. 2018)
	MAGIC	Imputation (van Dijk et al. *Cell.* 2018)
	scImpute	Imputation (Li and Li. *Nat Commun.* 2018)
	SAVER	Imputation by Bayesian approach (Huang et al. *Nat Methods*. 2018)
Trajectory analysis	SCUBA	Trajectory anlaysis (Marco et al. *Proc Natl Acad Sci U S A.* 2014)
	P-Creode	Unsupervised trajectory analysis (Herring et al. *Cell Syst.* 2018)
	CellAlign	Trajectory analysis (Alpert et al. *Nat Methods.* 2018)
	Monocle2	Trajectory analysis; A commonly used R toolkit (Qiu et al. *Nat Methods.* 2017)
	CellTree	Trajectory analysis based on a LDA model (duVerle et al. *BMC Bioinformatics*. 2016)
	CellRouter	Trajectory analysis (Lummertz da Rocha et al. *Nat Commun.* 2018)

microdroplet-based scRNA-seq. Currently, *Butler et al* reported the renewed version of Seurat in 2018 (Butler et al. 2018). Seurat got new skill to analyze datasets with each renewal (Seurat). Nowadays, it is thought to be the most commonly used tools for scRNA-seq analysis. It includes various functions, such as data filtering, dimensionality reduction, clustering and drawing figures. Seurat can also be used to compare two scRNA-seq datasets (for example, stimulated vs. unstimulated) using the canonical correlation analysis (CCA) function.

SC3 for Clustering

Single-cell consensus clustering (SC3) is a tool for unsupervised clustering, combining multiple clustering methods (Kiselev et al. 2017). This method is used by Zheng et al. to analyze single-cell transcriptomes of tumor-infiltrating T cells in liver cancers (Zheng et al. 2017a). They identified 11 subsets of T cells with unique signatures.

Reference Component Analysis (RCA) for Clustering

Reference component analysis (RCA) is a semi-supervised clustering approach for scRNA-seq data, which is based on bulk transcriptome data obtained from various types of tissue and cell types (Li et al. 2017). In the paper, RCA is used for accurate clustering of scRNA-seq data of colon cancers and matched normal tissues. The authors identified subpopulations of cancer-associated fibroblasts and single-cell transcriptome signatures associated with the prognosis.

Single-Cell Differential Expression (SCDE) for Differential Expression Analysis

Single-Cell Differential Expression (SCDE) is a tool for analyzing scRNA-seq datasets using a Bayesian approach (Kharchenko et al. 2014). It contains two frameworks for single-cell analysis, scde and pagoda. This method assesses differential expression between groups of cells, considering dropout and amplification bias in scRNA-seq data. It also enables us to analyze pathway and gene set overdispersion analysis (Fan et al. 2016).

Imputation for Missing Values

Depending on sequencing depth, each scRNA-seq dataset has a limited numbers of reads. Insufficient sequencing reads result in "dropout" of information with zero counts. The dropouts lead to incorrect conclusions in data interpretation. There are several computational methods to investigate whether zero counts indicate real low expression or technical dropout.

Zero Inflated Factor Analysis (ZIFA)

In 2015, Pierson et al. introduced a new method to consider dropout in single-cell datasets. Zero Inflated Factor Analysis (ZIFA) is a method of dimensionality reduction that incorporates zero counts (Pierson and Yau 2015). They tested ZIFA in both simulation-based and experimental-based datasets. It treats dropouts not as outliers but as real observations. The results suggested that ZIFA out-performed standard dimensionality-reduction algorithms. Even though ZIFA improves the accuracy of single-cell profiles, there are some limitations. First, it strictly models zero measurements. Second, its framework depends on the linear transformation, even though nonlinear dimensionality-reduction methods are said to be effective in single-cell analysis. Third, the performance of ZIFA rests heavily on the intrinsic separability of the cell types and dropout rates.

Zero-Inflated Negative Binomial-Based Wanted Variation Extraction (ZINB-WaVE)

Risso et al. reported Zero-Inflated Negative Binomial-based Wanted Variation Extraction (ZINB-WaVE), which achieves dimensional reduction of zero-inflated datasets (Risso et al. 2018). This method implements more accurate dimensional reduction compared with previously employed models, such as PCA and ZIFA.

Other Imputation Methods

Some imputation methods directly infer expression levels in genes with zero counts. Markov Affinity-based Graph Imputation of Cells (MAGIC) is a method for imputing dropouts in scRNA-seq datasets, which infers expression levels by referring to similar cell profiles (van Dijk et al. 2018). ScImpute is another method for imputing dropouts in scRNA-seq data (Li and Li 2018). Single-cell Analysis Via Expression Recovery (SAVER) is a Bayesian-based imputation method (Huang et al. 2018), which was also used by Savas et al. to estimate missing values of marker genes specific to distinct T cell populations in breast cancers (Savas et al. 2018).

Trajectory Analysis

Single-cell trajectory analysis is important for understanding cell-fate transition and identifying initiating cells, such as stem cells. There are various methods for tracking cell transition trajectories, for example, methods based on minimum spanning tree and nonlinear embedding. Here, we focus on novel pipelines recently reported for trajectory analysis.

Single-Cell Clustering Using Bifurcation Analysis (SCUBA)

In 2014, Marco et al. introduced a novel approach for single-cell clustering (Marco et al. 2014). Single-cell Clustering Using Bifurcation Analysis (SCUBA) reveals lineage relationships and their dynamic changes using single-cell transcriptome data. This method is based on a two-step approach. First, SCUBA estimates the locations of stage-specific attractors and their relationships (the cellular hierarchy) using a binary tree model. Second, SCUBA quantitatively models the dynamics in each direction. In this paper, the authors showed analysis of human B-cell differentiation using SCUBA and other methods, such as Wanderlust and Monocle. In SCUBA, we do not need to choose an initialization cell. The inferred pseudotime of SCUBA and Wanderlust showed high correlation ($R^2 = 0.70$). SCUBA is acceptable for RT-PCR, RNA-seq, mass cytometry, and various other experimental data. SCUBA is said to be well suited for investing in developmental processes.

p-Creode

Herring et al. developed an unsupervised algorithm called p-Creode (Herring et al. 2018). This method enables us to derive multibranching trajectories from single-cell transcriptome data. In this paper, they combined inDrop, multiplex immunofluorescence (MxIF) and mass cytometry datasets. P-Creode consists of six steps. Among these steps, there are three novel points: automatic identification of the end point, a hierarchical placement strategy for placing data points on branches and N resampled topologies to depict the relative robustness of data. Using p-Creode, the authors predicted alternative tuft cell origins and *Atoh1*-driven developmental programs in the gut. They also applied p-Creode to two published scRNA-seq datasets from Fluidigm C1 and MARS-seq, to assess their performance. In conclusion, they showed that p-Creode is a reliable method to depict branching cell transition trajectories.

CellAlign

CellAlign is a quantitative method to compare expression dynamics among trajectories and is based on dynamic time warping (Alpert et al. 2018). It enables to reveal the phenomenon that would be masked in conventional methods. In the paper, the authors evaluated the ability of cellAlign using publicly available scRNA-seq data. As a result, cellAlign appears to be an accurate and robust method. They also showed that cellAlign could be used for analysis of single-cell trajectories using mass cytometry data.

Other Methods for Trajectory Analysis

To conduct pseudotime analysis using scRNA-seq data, we can also use other methods such as Monocle2 (Trapnell et al. 2014; Qiu et al. 2017), CellTree (duVerle et al. 2016) and CellRouter (Lummertz Da Rocha et al. 2018). Monocle2 is one of the most common methods for revealing single-cell trajectories by "pseudotime". This method applies reversed graph embedding, which is a machine learning strategy for learning the graph structures of single-cell trajectories.

Recently, numerous methods and tools have been developed for single-cell trajectory analysis. It is difficult to choose suitable methods. Saelens et al. provided an assessment of trajectory methods (Saelens et al. 2018). They compared 29 trajectory methods and suggested useful guidelines for method selection.

Summary

Single-cell sequencing analysis is a powerful tool for understanding heterogeneity in cellular populations. In particular, scRNA-seq provides comprehensive transcriptome profiles of individual cells for inferring cell types, their states and trajectories.

The development of protocols and tools has promoted the accumulation of huge scRNA-seq datasets. However, analysis of scRNA-seq data is still difficult because the data are sparse and zero-inflated, with insufficient sequencing depths for each cell. This problem sometimes results in misleading conclusions. Recent studies of single-cell analysis have used diverse computational methods designed for scRNA-seq data. We need to follow the very rapid advance of experimental and computational methods for single-cell sequencing analysis.

References

10x Genomics Single Cell 3′ Reagent Kits v2 User Guide. https://support.10xgenomics.com/single-cell-gene-expression/index/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v2-chemistry.

Alpert A, Moore LS, Dubovik T, Shen-Orr SS. Alignment of single-cell trajectories to compare cellular expression dynamics. Nat Methods. 2018;15:267–70. https://doi.org/10.1038/nmeth.4628.

Angermueller C, Clark SJ, Lee HJ, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13:229–32. https://doi.org/10.1038/nmeth.3728.

Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8. https://doi.org/10.1038/nmeth.2688.

Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90. https://doi.org/10.1038/nature14590.

Buenrostro JD, Corces MR, Lareau CA, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell. 2018;173:1535–1548.e16. https://doi.org/10.1016/j.cell.2018.03.074.

Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20. https://doi.org/10.1038/nbt.4096.

Cusanovich DA, Daza R, Adey A, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348:910–4. https://doi.org/10.1126/science.aab1601.

Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14:297–301. https://doi.org/10.1038/nmeth.4177.

Dey SS, Kester L, Spanjaard B, et al. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33:285–9. https://doi.org/10.1038/nbt.3129.

Dixit A, Parnas O, Li B, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167:1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

duVerle DA, Yotsukura S, Nomura S, et al. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. BMC Bioinformatics. 2016;17:363. https://doi.org/10.1186/s12859-016-1175-6.

Fan J, Salathia N, Liu R, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016;13:241–4. https://doi.org/10.1038/nmeth.3734.

Gao R, Davis A, McDonald TO, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48:1119–30. https://doi.org/10.1038/ng.3641.

Guo H, Zhu P, Wu X, et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23:2126–35. https://doi.org/10.1101/gr.161679.113.

Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by Microwell-Seq. Cell. 2018;172:1091–1097.e17. https://doi.org/10.1016/j.cell.2018.02.001.
Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003.
Herring CA, Banerjee A, McKinley ET, et al. Unsupervised trajectory analysis of single-cell RNA-Seq and imaging data reveals alternative tuft cell origins in the g. Cell Syst. 2018;6:37–51.e9. https://doi.org/10.1016/j.cels.2017.10.012.
Huang M, Wang J, Torre E, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15:539–42. https://doi.org/10.1038/s41592-018-0033-z.
Jaitin DA, Weiner A, Yofe I, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. Cell. 2016;167:1883–1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.
Kashima Y, Suzuki A, Liu Y, et al. Combinatory use of distinct single-cell RNA-seq analytical platforms reveals the heterogeneous transcriptome response. Sci Rep. 2018;8:3482. https://doi.org/10.1038/s41598-018-21161-y.
Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2. https://doi.org/10.1038/nmeth.2967.
Kim C, Gao R, Sei E, et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. Cell. 2018;173:879–893.e13. https://doi.org/10.1016/j.cell.2018.03.041.
Kiselev VY, Kirschner K, Schaub MT, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6. https://doi.org/10.1038/nmeth.4236.
Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.
Lasken RS. Single-cell genomic sequencing using multiple displacement amplification. Curr Opin Microbiol. 2007;10:510–6. https://doi.org/10.1016/j.mib.2007.08.005.
Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9:997. https://doi.org/10.1038/s41467-018-03405-7.
Li H, Courtois ET, Sengupta D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49:708–18. https://doi.org/10.1038/ng.3818.
Lummertz Da Rocha E, Rowe RG, Lundin V, et al. Reconstruction of complex single-cell trajectories using CellRouter. Nat Commun. 2018;9:892. https://doi.org/10.1038/s41467-018-03214-y.
Macaulay IC, Haerty W, Kumar P, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12:519–22. https://doi.org/10.1038/nmeth.3370.
Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.
Marco E, Karp RL, Guo G, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014;111:E5643–50. https://doi.org/10.1073/pnas.1408993111.
Matsunaga H, Goto M, Arikawa K, et al. A highly sensitive and accurate gene expression analysis by sequencing ("bead-seq") for a single cell. Anal Biochem. 2015;471:9–16. https://doi.org/10.1016/j.ab.2014.10.011.
Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–5. https://doi.org/10.1038/nature09807.
Peterson VM, Zhang KX, Kumar N, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35:936–9. https://doi.org/10.1038/nbt.3973.
Picelli S, Björklund ÅK, Faridani OR, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–100. https://doi.org/10.1038/nmeth.2639.
Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241. https://doi.org/10.1186/s13059-015-0805-z.

Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14:979–82. https://doi.org/10.1038/nmeth.4402.

Ramsköld D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82. https://doi.org/10.1038/nbt.2282.

Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9:284. https://doi.org/10.1038/s41467-017-02554-5.

Rotem A, Ram O, Shoresh N, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015;33:1165–72. https://doi.org/10.1038/nbt.3383.

Saelens W, Cannoodt R, et al. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. bioRxiv. 2018:276907. https://doi.org/10.1101/276907.

Sasagawa Y, Nikaido I, Hayashi T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. Genome Biol. 2013;14:R31. https://doi.org/10.1186/gb-2013-14-4-r31.

Satpathy AT, Saligrama N, Buenrostro JD, et al. Transcript-indexed ATAC-seq for precision immune profiling. Nat Med. 2018;24:580–90. https://doi.org/10.1038/s41591-018-0008-8.

Savas P, Virassamy B, Ye C, et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat Med. 2018;24:986–93. https://doi.org/10.1038/s41591-018-0078-7.

Seurat. https://satijalab.org/seurat/.

Smallwood SA, Lee HJ, Angermueller C, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11:817–20. https://doi.org/10.1038/nmeth.3035.

Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14:865–8. https://doi.org/10.1038/nmeth.4380.

Suzuki A, Matsushima K, Makinoshima H, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. 2015;16:66. https://doi.org/10.1186/s13059-015-0636-y.

Svensson V, Natarajan KN, Ly LH, et al. Power analysis of single-cell rnA-sequencing experiments. Nat Methods. 2017;14:381–7. https://doi.org/10.1038/nmeth.4220.

Telenius H, Carter NP, Bebb CE, et al. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics. 1992;13:718–25. https://doi.org/10.1016/0888-7543(92)90147-K.

Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6. https://doi.org/10.1038/nbt.2859.

van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018. https://doi.org/10.1016/j.cell.2018.05.061.

Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell. 2015;58:598–609.

Wang Y, Waters J, Leung ML, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60. https://doi.org/10.1038/nature13600.

Zheng C, Zheng L, Yoo JK, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell. 2017a;169:1342–1356.e16. https://doi.org/10.1016/j.cell.2017.05.035.

Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017b;8:14049. https://doi.org/10.1038/ncomms14049.

Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338:1622–6. https://doi.org/10.1126/science.1229164.

Bionano Genome Mapping: High-Throughput, Ultra-Long Molecule Genome Analysis System for Precision Genome Assembly and Haploid-Resolved Structural Variation Discovery

Sven Bocklandt, Alex Hastie, and Han Cao

Abstract Next Generation Sequencing (NGS) has rapidly advanced genomic research with tremendously increased throughput and reduced cost, through reading the fragmented genome content in massively parallel fashion. We have been able to sequence and map genomes to reference sequences with relative ease compared to the past. However, this mapping can only be accurately accomplished in the single copy regions of the genome, leaving out most duplicated genes and structural variation. Additionally, assembly of long genomic segments remains elusive since multi copy regions of the genome produce ambiguity when short read sequence is used.

Most of the large genomes are complex in that they contain not only millions of single or multiple base level variants called SNPs (Single Nucleotide Polymorphism) and indels (small insertions and deletions), they also contain many thousands of much larger structural variants, repetitive regions composed of identical or similar stretches of sequences, mobile elements such as transposons, large insertions, deletions, translations and inversions up to millions of bases, even partial or entire chromosomes altered. Often more than half of the genome is composed of these non-unique and highly variable regions such as in human and up to 90% in certain plants (Jiao et al. 2017). And now through studying thousands upon thousands of genomes, we have come to realize that each genome from each individual bears the mark of its own evolutionary journey and environment. This is seen in the different code in each of the two haplotypes from each family or different ethnicity specific signatures in populations (Sudmant et al. 2015), and even the genomes in different cells derived from the same gamete carry non-static sequence variation accumulated throughout its lifetime, sometimes leading to tumorigenesis and contributing to the natural aging process.

S. Bocklandt (✉) · A. Hastie · H. Cao
Bionano Genomics, San Diego, CA, USA
e-mail: SBocklandt@bionanogenomics.com

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_7

It is not enough just to re-sequence each genome by aligning short reads from NGS to an existing relatively contiguous reference genome, calling only the SNPs and small indels. To realize the full potential of the so-called precision medicine, we need to get to the true, accurate, and complete genome information *de novo* to understand how these large structural variations might affect biological functions. The first step is creating and identifying technologies that are able to preserve and access native long range genomic content, including SNPS, small indels and all classes of SVs, without gaps.

Large structural variations (SVs) are less common than SNPs and indels in the population in numbers of events but collectively account for a significantly larger number of base pair variations, although the impact on genetic variation and diseases is yet unknown. While single nucleotide mutations might impact the 2% of protein coding regions and key small regulatory elements such as the transcription factor binding sites, larger structural variations could have additional large effects, including eliminating, truncating or altering the coding regions or regulatory elements directly, and also changing the copy number, position or orientation of these genes or promoters, placing them into different genomic context. Moreover, large SVs can alter the complex three-dimensional folding of the chromatin within the cell and how genomic, epigenomic and protein elements interact with each other dynamically in a much more profound time and spatial order. However, the existing prevailing methods cannot comprehensively and cost effectively detect all of the large structural variations due to the limited read lengths of the existing technologies (Huddleston and Eichler 2016).

To address these challenges, Bionano Genomics applies a high-throughput, native, single molecule level genome mapping technology to comprehensively determine genome wide structure using *de novo* assembly of sequence motif-specific labeled long molecules (>150 kb), linearized in massive parallel nanofluidic channels fabricated on a solid-state material (Lam et al. 2012). Exploiting this technology, structurally accurate whole genome *de novo* assemblies can be generated. Typically, comparing to the human reference, thousands of SVs (>500 bp) are obtained in a single human genome. Because ultra-long read technology is so new, currently only a fraction of SVs is verified in SVs databases while a large portion are novel. Due to the nature of ultra-long molecules, haplotypes preserving native structural information are phased in differently clustered molecules by the association of the same labeling patterns in a straightforward fashion. In addition, these data are validated instantly by supporting raw images of the long molecules, not inferred by an algorithm (Bickhart et al. 2017). Without excessive processing performed in a typical sample prep such as PCR, adaptor addition, cloning, or library construction, the longest possible genomic DNA molecules (150 kb to megabases) are isolated directly from cells to be labeled, and imaged at single molecule level, ensuring that the integrity of the most native information is preserved at the genomic and epigenomic level (Grunwald et al. 2017).

The genome mapping technology enabled by NanoChannel arrays, Bionano optical mapping, provides valuable information for endogenous highly variable regions such as areas related to immunity (MHC, KIR, TCR, etc.) as well as exogenous elements such as free or integrated viral sequence without a priori knowledge on a whole genome scale (Cao et al. 2014). Furthermore, dynamic intra-

cellular genomic events such as DNA replication can be imaged with this platform process. DNA replication is often implicated as a major cause of genomic error generation eventually causing genome instability and cancers (Klein et al. 2017).

Bionano maps often reach chromosome arm lengths and are therefore highly informative for de novo genome assembly projects where they scaffold fragmented NGS assemblies and correct assembly errors. Many of the resulting assemblies are among the most contiguous and accurate assembled to date.

By employing Bionano mapping's long-range genome analysis, large SVs can be identified in each individual and across multiple ethnic populations where population-specific structural variation sets are seen. These results highlight the need for a comprehensive set of alternate haplotypes derived from different populations to resolve structural variation patterns in complex regions of the genome, providing evidence for population genomic based diagnosis and drug development.

Bionano mapping technology is a high throughput, high fidelity and versatile platform with high potential to transform clinical cytogenetic and genetic analysis in a fully automated and standardized fashion in a cost-effective way. It has been recently demonstrated that comprehensive large SVs can be profiled in prostate cancer samples, where novel potential causal events were discovered efficiently *de novo* (Jaratlerdsiri et al. 2017). In a separate study involving rare and undiagnosed diseases, a very large 5.1 Mbp inversion in the genome of a patient with Duchenne Muscular Dystrophy was discovered with Bionano technology in a single, 1 week experiment, leading to the definitive molecular mechanism caused by a truncation of Dystrophin gene (Barseghyan et al. 2017). This inversion had previously evaded a wide range of standard clinical and molecular tests.

These clinical studies have paved the way for demonstration of the potential of routine comprehensive genomic analysis for complex diseases in precision medicine era.

Background

Existing technologies including chromosomal microarrays and whole genome sequencing diagnose less than 50% of patients with genetic disorders (Lee et al. 2014; Miller et al. 2010). This leaves a majority of patients without ever receiving a molecular diagnosis. Undiagnosed disorders are individually rare but their combined incidence and the associated diagnostic odyssey, with resultant delays in treatment, are a drain on families and the healthcare system. Many of these diseases remain medical mysteries with no root cause or clear basis for treatment.

To close this diagnostic sensitivity gap and get a better understanding of the genetic causes of disease, we need better tools to access the entire genome, and large translational research studies to apply these tools to the discovery of novel biomarkers. Genetic disorders for which no molecular basis is currently known are either caused by genomic events that are poorly detected with current technology, events occurring in inaccessible parts of the genome, or a combination of events that

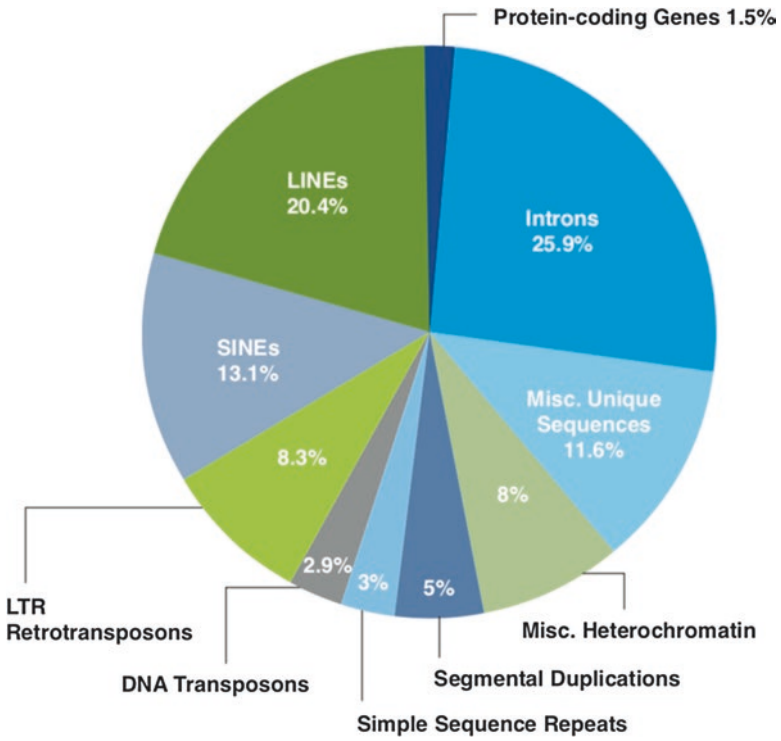

Fig. 1 Repetitive structures in the human genome

is too complex to analyze using existing tools. Better molecular tools are needed to analyze the entire range of genomic variations. Armed with such tools, large translational research studies are needed to identify disease correlated biomarkers spanning all genomic variants in patients with genetic disorders.

Two thirds of the human genome consist of repetitive sequences (Fig. 1). Exome sequencing accesses just 1.5% of the genome (de Koning et al. 2011), and Whole Genome Sequencing (WGS) does not align correctly with the repetitive parts of the genome. The most common repetitive sequences in the genome are LINEs, SINEs, retrotransposons and segmental duplications. The short-read sequences Next-Generation Sequencing (NGS) provides, map with poor accuracy to these repeats. Alignment algorithms typically fail to identify the exact genomic location to align these short-reads to. When they do align, the limited 100–150 bp read length and spacing of paired-end reads does not allow for a correct sizing of larger repeats.

Structural variants make up the majority of human genomic variation, but Next-Generation Sequencing technology can't correctly identify them. Clinical exome sequencing solves about 30% of rare diseases (Lee et al. 2014). NGS, consisting of Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) reliably identify single nucleotide variants and small insertions and deletions. However, NGS relies on short-read sequences that are mapped to a reference human genome

and fails to identify most large insertions, deletions, or copy-number variations in repetitive regions of the genome. It is incapable of easily detecting other structural variations (SVs) such as inversions and translocations. Non-allelic homologous recombination of repetitive sequences is thought to be a predominant mechanism for the origin of many large SVs. The non-unique sequences flanking these SVs often make them invisible to sequencing-based detection methods. Together, structural variable regions cover 13% of the genome and individuals show structural variation covering as much as 30 Mbp between each other (Sudmant et al. 2015).

Methods

Mechanism of Bionano Technology and Workflow

Ultra-Long Range Linear DNA Analysis Technology Enabled by NanoChannel Array Technology

An overview of the molecular and bioinformatics method is shown in Fig. 2.

Since structurally accurate genome interrogation and assembly requires long molecules, traditional purification methods are not suitable for DNA isolation for optical mapping. Bionano Genomics adapted the plug lysis strategy commonly used to construct BAC libraries for optical mapping. Briefly, cells/nuclei are embedded into an agarose matrix to protect DNA from mechanical shearing during the purification process. Agarose is then melted and solubilized, and the resulting megabase DNA is further cleaned by drop dialysis prior to labeling at sequence-specific sites.

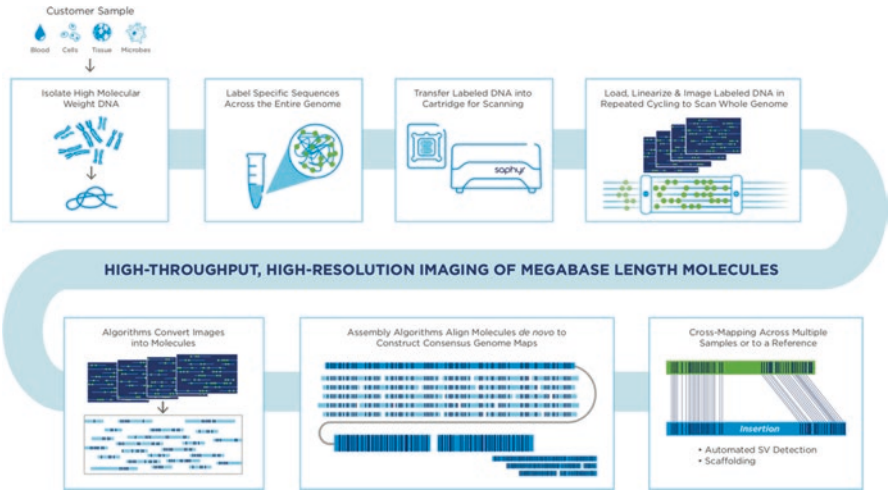

Fig. 2 Bionano workflow for DNA isolation, labeling, imaging, analysis

Megabase size molecules of genomic DNA are labeled at a specific 6 or 7 basepair sequence motif, occuring approximately 8–28 times per 100 kbp, depending on its frequency in a particular genome. The label patterns allow each long molecule to be uniquely identified and aligned.

Labeled DNA is loaded onto the Saphyr chip and placed into the Saphyr instrument where Saphyr initiates electrophoresis to move megabase length molecules from bulk solution into the silicon chip micro environment before unwinding and linearizing the DNA in the NanoChannel arrays. The instrument uses machine learning initially and throughout the run to provide adaptive loading of DNA, optimize run conditions and maximize throughput.

When molecules are fully loaded into the NanoChannels of one flow cell, electrophoresis is halted and the entire surface of the NanoChannel array of that flow cell is rapidly imaged. During the imaging phase of the run, electrophoresis in the second flow cell is initiated. Cycles of loading of the NanoChannels followed by imaging are performed until sufficient data is collected.

Bioinformatics of Bionano Mapping Using Sequence Motif Pattern Specific Labeling

Bionano image detection software creates extracts molecules from raw image data. The backbone stain signal of the DNA molecules is used to identify molecules and to determine their position and size. The distance between the labels on each molecule is recorded to generate an extracted molecule file called a BNX file. The BNX file is the only input needed for the Bionano *de novo* assembly process.

Images generated by Saphyr are sent to the analysis server for real time data extraction during the run. Image detection is typically completed shortly after the run is finished and *de novo* assembly can be automatically initiated (for human genomes).

Using pairwise alignment of the single molecules, an assembly graph is constructed and a consensus genome map is produced, refined, extended and merged. Molecules are then clustered into two alleles, where there is heterozygous structural variation, and a diploid assembly is created to allow for heterozygous SV detection. Genome maps can be created using different enzymes labeling different sequence motifs to generate broader coverage and higher label density.

A standard automated pipeline for *de novo* assembly and SV calling was developed by Bionano Genomics to enable comprehensive SV analysis. The Python-based pipeline manages job submission, drives execution of alignment and assembly tools, and provides data summary information. It features a haplotype-aware assembler designed to detect and differentiate parental alleles. The *de novo* assembly algorithm is a custom implementation of the overlap-layout-consensus strategy. The assembler assembles extracted molecules from raw image data, and the final consensus maps are used as input for SV calling.

Examples of Applications with Bionano Mapping

De novo Assembly of Complex Genomes – Long Contiguity with Accurate Complex Structural Context

Hybrid Assembly Combining Mapping and Sequencing Data Derived from All Platforms – Single and Multiple Sequence Motif-Based Assembly

The *de novo* Bionano genome maps are a whole genome *de novo* assembly and can be used to learn about various characteristics of the genome such as size, repetitive content, and extent of heterozygosity. They can also be integrated with a sequence assembly to order and orient sequence fragments, identify and correct potential chimeric joins in the sequence assembly, and estimate the gap size between adjacent sequences. In order to do so, the Bionano Solve software imports the sequence assembly and identifies the recognition sites for the specifying nick sites in the sequence based on the nicking endonuclease-specific recognition site. These *in silico* maps for the sequence contigs are then aligned to the *de novo* Bionano genome maps. Conflicts between the two are identified and resolved, and hybrid scaffolds are generated in which sequence maps are used to bridge Bionano maps and vice versa. Finally, the sequence assembly corresponding to this hybrid scaffold is generated and exported as FASTA and AGP files.

The pipeline is fully integrated with Bionano Access which provides a convenient interface for running Hybrid Scaffold and viewing scaffolding results.

The hybrid scaffolding process considerably reduces the number of contigs found in the initial NGS assembly, improving assembly accuracy and quality while reducing the need for deep sequencing coverage.

The hybrid scaffolding approach can yield significant improvements in contiguity, as expressed by the assembly N50 values. Assembly contiguity can be further increased by performing hybrid scaffolding with maps using two separate nicking enzymes. Two sets of Bionano maps, each generated with a different nicking enzyme, can be integrated with NGS sequences together. This enables the NGS sequences to function as a bridge to merge single-enzyme Bionano maps into two-enzyme maps that contain the sequence motif patterns from both nicking enzymes. Since the Bionano maps are generated independently they serve as orthogonal sources of evidence to detect and correct assembly errors in input data. The complementarity of different data also greatly improves the contiguity of the merged Bionano map while doubling the information density, which substantially increases the ability to anchor short NGS sequences in the final scaffolds.

The two-enzyme approach was validated on the human NA12878 genome, a model data set for which sequence data is publicly available. Three different assemblies were tested: Illumina-D, 51x of 250 bp pair-end sequence; Illumina-S, 40x of 101 bp pair-end and 25x of 2.5–2.5 kbp mate-pair sequence; and PacBio, 46x with mean read length of 3.6 kbp. Compared to the input NGS, the two-enzyme approach improves the scaffold contiguity up to 100-fold, Fig. 3), anchors 30% more sequence

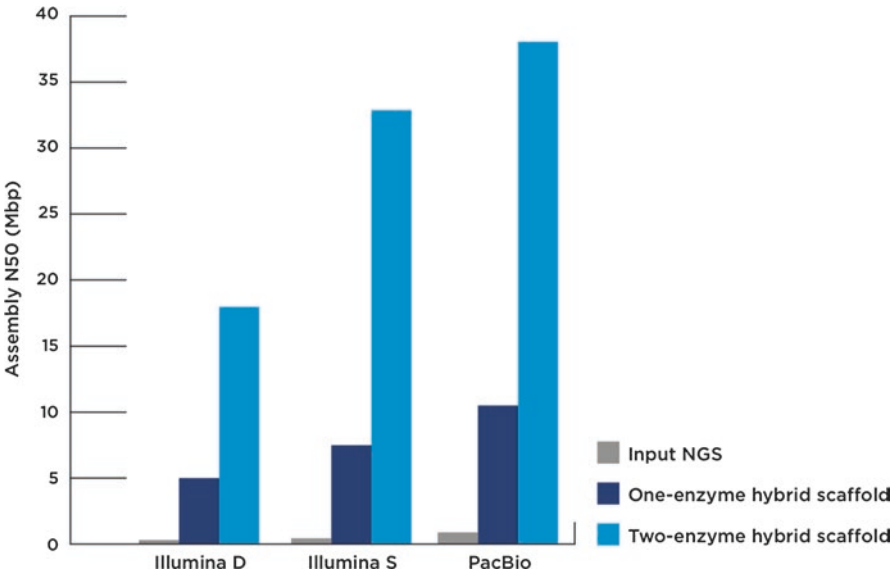

Fig. 3 Improvements in NA12878 assembly contiguity after hybrid scaffold with one-enzyme and two-enzyme genome maps. Illumina-D: 51x of 250 bp pair-end sequence; Illumina-S: 40x of 101 bp pair-end and 25x of 2.5–2.5 kbp mate-pair sequence; PacBio: 46x with mean read length of 3.6 kbp

contigs in the final scaffolds and corrects 50% more assembly errors in NGS sequences. The pipeline performs robustly in both animal and plant genomes as well (Fig. 4). This approach greatly expands the type of NGS data that can be integrated with Bionano maps to produce highly accurate and contiguous assemblies for complex genomes.

At the time of writing, all published data using Bionano mapping has been generated by labeling DNA with nicking endonucleases. These highly sequence-specific enzymes create a single stranded nick at the presence of a 6- or 7 bp motif. At the site of the nicked DNA, fluorescently labeled nucleotides are inserted by polymerization and the molecules are repaired. This method (Nick Label Repair Stain, or NLRS) performs with extremely high specificity but can create double stranded breaks when nick site appear within about 200 bp on opposite strands. Recently Bionano Genomics has developed a novel labeling technology that avoids nicking, and instead uses a direct labeling method where the fluorophore is attached directly to the DNA at the location of a specific sequence motif. Since this Direct Labeling and Staining (DLS) method does not create systematic double stranded breaks, Bionano maps created from molecules labeled with DLS typically show a 50x improvement in contiguity compared to NLRS maps. Bionano maps now typically reach chromosome arm length, and the contiguity of sequence assemblies built using DLS reaches chromosome arm or full chromosome length in a variety of species.

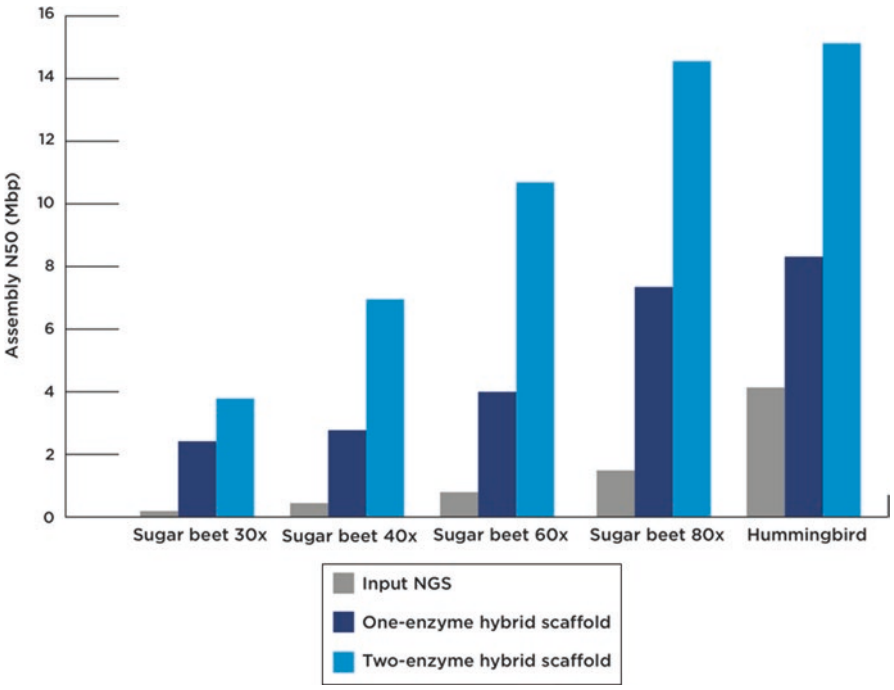

Fig. 4 Improvements in sugar beet and hummingbird assembly contiguity after hybrid scaffolding with Bionano genome maps using one-enzyme and two-enzymes. For sugar beet, the fold coverage of the PacBio de novo assemblies is shown

Error Correction and Validation of Sequencing Data

The Bionano hybrid scaffold pipeline detects and resolves chimeric joins. Chimeric joins are typically formed when short reads, molecules, or paired-end inserts are unable to span across long DNA repeats. The errors appear as conflicting junctions in the alignment between the Bionano map and NGS assemblies.

When the hybrid scaffold pipeline detects a conflict, it analyzes the single-molecule data that underlies a Bionano map and assesses which assembly was incorrectly formed. If the Bionano map has long molecule support at the conflict junction, the sequence contig is automatically cut, removing the putative chimeric join (Fig. 5). If it does not have strong molecule support, then the Bionano map is automatically cut. Both assemblies must have coverage spanning both sides of a chimeric join to detect and resolve these conflicts.

Automated cuts using Bionano Solve help to resolve conflicts with a high level of accuracy. The majority of cuts made using Bionano Solve can be confirmed by comparison to the species' reference assembly. There are several reasons why some cuts cannot be confirmed: the reference assembly is incomplete, the two separate input assemblies may represent different alleles, or the chimeric joins may have been caused by segmental duplications that are too long for Bionano molecules to resolve.

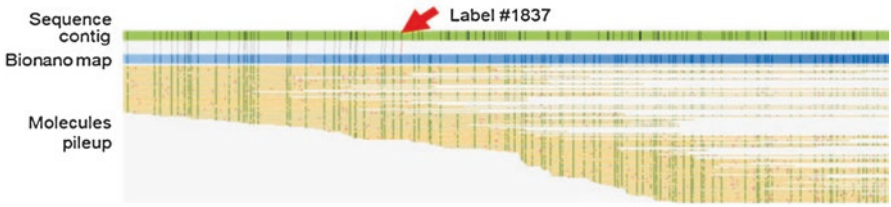

Fig. 5 Example of a conflict between a sequence contig and a Bionano map. The conflict junction as shown by the red arrow in the alignment between the sequence contig and the Bionano map. There is strong molecule support spanning the junction region on the genome map, so the sequence is cut at the label indicated

The two-enzyme scaffolding method improves the error correction even further. Since the Bionano maps were generated independently they serve as orthogonal sources of evidences to detect and correct assembly errors in input data. Compared to the published one-enzyme hybrid-scaffolds, the two-enzyme approach corrects up to 50% more assembly errors in NGS sequences.

Users can manually inspect all conflict resolution results. Bionano Solve notes the IDs and coordinates of the sequences and maps where conflicts have been detected and the corresponding resolution approaches taken. This file can be edited and modified, and then run again in the hybrid scaffold pipeline to produce a new set of scaffolds based on the manual conflict resolution. This manual enhancement process can be performed multiple times, giving users fine control in generating high-quality, complete hybrid scaffolds.

Comprehensive Genomic Structural Variation Discovery and Identification

Detecting All Classes of Structural Variants, Mobile Elements and Repeats, at Haploid Resolved Level

Bionano genome mapping is the only technology that detects all SV types, homozygous and heterozygous, starting at 500 bp up to millions of bp. Bionano maps are built completely *de novo*, without any reference guidance or bias. This differentiates Bionano from NGS, where short-read sequences are typically aligned to a reference. This alignment often fails to detect true structural variants by forcing the short-reads to map to an incorrect or too divergent reference, or by excluding mismatched reads from the alignment. Only *de novo* constructed genomes, like Bionano maps, allow for a completely unbiased, accurate assembly.

Bionano's SVs are observed, and not inferred as with NGS. When short-read NGS sequences are aligned to the reference genome, algorithms piece together sequence fragments in an attempt to rebuild the actual structure of the genome. SVs are inferred from the fragmented data, with mixed success. With Bionano mapping, megabase-size native DNA molecules are imaged, and most large SVs or their

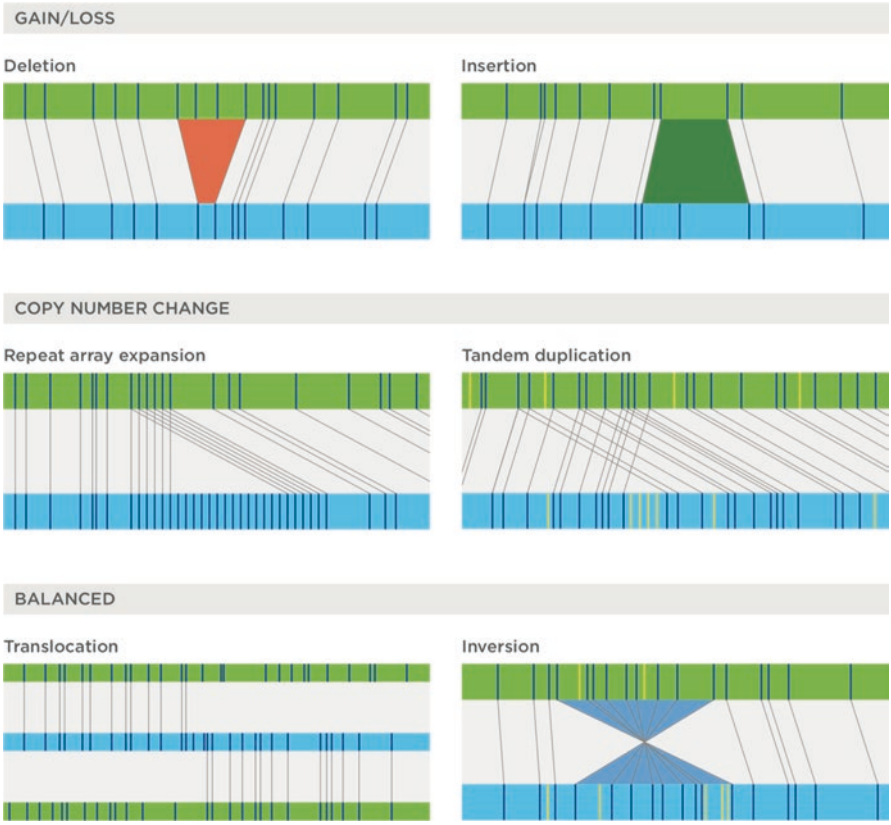

Fig. 6 Structural variant types detected by Bionano mapping. SVs are identified by comparing label patterns in the sample of interest (blue) with those in the reference genome, or in a reference sample (green). Major types detected are

breakpoints (in the case of inter-chromosomal translocations) can be observed directly in the label pattern on the molecules. If a native-state DNA molecule with a specific SV exists, then that SV call cannot be wrong.

SV calls are made based on analyses of a multiple local alignment between consensus maps and the reference (Fig. 6). The pipeline supports calling of major SV types: insertions, deletions, inversions, and translocation breakpoints. Bionano Access also supports visualization and confidence-based filtering of these SV types. Poorly aligned or unaligned regions flanked by well-aligned regions are called as deletions or insertions, depending on whether there is gain or loss of sequence relative to the reference. Junctions of neighboring alignments with opposite orientations are identified as inversion breakpoints. Fusion points between distant regions of the genome are identified as translocation breakpoints. Intrachromosomal translocation breakpoints involve regions on the same chromosome but at least 5 Mbp away from each other. Interchromosomal translocation breakpoints involve regions on different chromosomes.

Gain/Loss of material: Labels moving closer together, with or without loss of labels, are evidence of deletions. Label spacing that increases with or without additional labels detected are called as insertions.

Copy number change: Expansions or contractions of tandem arrays or segmental duplications. Duplications are called automatically in direct or inverted orientation.

Balanced events: Genome maps aligning partially with two or more different chromosomes or genomic locations indicate translocations. When label patterns are inverted relative to the reference, an inversion is called.

Zygosity and confidence are assigned to each SV call to facilitate downstream analysis. An SV call can be labeled as homozygous, heterozygous, or unknown. Confidence scores are scaled such that they range from 0 to 1.

SV calls can be exported in a dbVar compliant VCF file. This file format contains all genomic variants identified in sample including SNVs, small indels, and SVs of various sizes. The VCF file generated by Bionano Access can be used in downstream analysis using a variety of existing tools.

Bionano algorithms call SVs by comparing genome structures. To identify a structural variation, a *de novo* genome map assembly can be aligned to a reference genome, or two samples can be aligned to each other directly. When aligning a genome map to a reference assembly, Bionano software identifies the location of the same recognition sequence used to label the DNA molecules in the reference genome and aligns matching label patterns in the sample and reference. This alignment provides all the annotation of the reference to the *de novo* assembled genome.

By observing changes in label spacing and comparisons of order, position, and orientation of label patterns, Bionano's automated structural variation calling algorithms detect all major structural variation types.

Bionano detects seven times more SVs larger than 5 kbp compared to NGS. Professor Pui-Yan Kwok at the University of California, San Francisco, demonstrated the robustness of Bionano mapping for genome-wide discovery of SVs in a trio from the 1000 Genomes Project. Since high quality NGS data on these samples is publicly available, structural variation analysis using short-read data has been performed with over a dozen different algorithms. Using Bionano maps, hundreds of insertions, deletions, and inversions greater than 5 kbp were uncovered, 7 times more than the large SV events previously detected by NGS (Mak et al. 2016). Several are located in regions likely leading to disruption of gene function or regulation.

Bionano has exceptional sensitivity and specificity to detect insertions and deletions over a wide size range as demonstrated using simulated data. Insertions and deletions were randomly introduced into an in-silico map of the human reference genome hg19. The simulated events were at least 500 kbp from each other or N-base gaps. They ranged from 200 bp to 1 Mbp, with smaller SVs more frequent than larger ones.

Based on the edited and the unedited hg19, molecules were simulated to resemble actual molecules collected on a Bionano system and mixed such that all events would be heterozygous. Two sets of molecules were simulated, each labeled with a

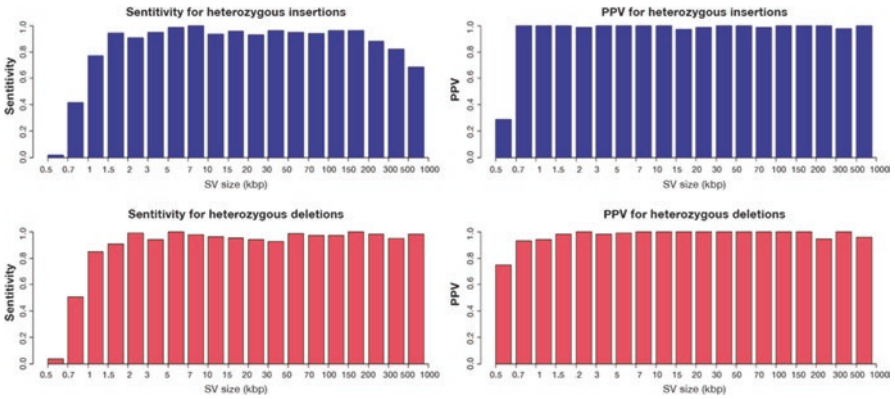

Fig. 7 Heterozygous SV calling performance from a simulated dataset. Molecules were simulated from unedited and edited versions of hg19 (with insertions and deletions of different sizes) and used for assembly and SV calling

different nicking endonuclease. Datasets with 70x effective coverage were generated. The simulated molecules were used as input to the Bionano Solve pipeline and SV calls were made by combining the single-enzyme SV calls from both nicking endonucleases using the SV Merge algorithm. SV calls were compared to the ground truth.

Figure 7 shows sensitivity and positive predicted value (PPV) for heterozygous insertions and deletions within a large size range. SV size estimates were typically within 500 bp of the actual SV sizes, while reported breakpoints were typically within 10 kbp of the actual breakpoint coordinates. Additional large insertions (>200 kbp) were found but classified as end-calls.

Bionano mapping has exceptional sensitivity and specificity to detect heterozygous insertions and deletions over a wide size range as demonstrated using experimental data. Since there is no perfectly characterized human genome that can be considered the ground truth, a diploid human genome was simulated by combining data from two hydatidiform mole derived cell lines. These moles occur when an oocyte without nuclear DNA gets fertilized by a sperm. The haploid genome in the sperm gets duplicated, and the cell lines resulting from this tissue (CHM1 and CHM13) are therefore entirely homozygous.

Structural variants detected in the homozygous cell lines were considered the (conditional) ground truth. An equal mixture of single molecule data from two such cell lines was assembled to simulate a diploid genome, and SV calls made from this mixture were used to calculate the sensitivity to detect heterozygous SVs.

Table 1 shows the number of insertions and deletions larger than 1.5 kbp detected in the CHM1 and CHM13 homozygous cell lines relative to the reference, and the in silico CHM1/13 mixture. SVs detected in CHM1 only or CHM13 only are heterozygous and those detected in both are homozygous. Bionano has a sensitivity of 92% for heterozygous deletions and 84% for heterozygous insertions larger than 1.5 kbp. The largest detected deletion was 4.28 Mbp in size and the largest insertion 412 kbp

Table 1 Two homozygous cell lines, CHM1 and CHM13 were independently de novo assembled and insertions and deletions >1.5 kbp called

	PacBio				Bionano			
	CHM1 and CHM13 assemblies	Mixture assembly	Sensitivity (%)	PPV (%)	CHM1 andCHM13 assemblies	Mixture assembly	Sensitivity (%)	PPV (%)
Homozygous insertions	467	353	75.6	96.1	707	700	99.0	97.9
Heterozygous insertions	586	252	43.0		663	554	83.6	
Homozygous deletions	221	183	82.8	94.9	269	268	99.6	97.1
Heterozygous deletions	501	337	67.3		517	477	92.3	

Raw data was mixed together, assembled and SVs called (Mixture assemblies column). The sensitivity and positive predictive value (PPV) to detect heterozygous relative to homozygous SVs is shown

A similar experiment on PacBio long-read sequencing was described recently (Huddleston et al. 2017). Structural variants were called with the SMRT-SV algorithm in CHM1 and CHM13, and compared to those called in an equal mixture of both. The sensitivity to detect homozygous SVs using PacBio was 87%, compared to 99.2% using Bionano. The sensitivity to detect heterozygous SVs using PacBio was only 41%, which is less than half the 86% sensitivity for heterozygous SV detection using Bionano. Even when the PacBio SV calls were limited to insertions and deletions larger than 1.5 kbp, the sensitivity for homozygous SVs was only 78%, and for heterozygous SVs 54% (Table 1).

Bionano genome mapping detects 98% of large inversions. Inversions are the invisible variants and have traditionally been the hardest to detect structural events. They are balanced, without gain or loss of sequence, and unlike translocations they don't create easily visible changes in genomic context. Inversions often escape detection by traditional cytogenetic techniques. Chromosomal Microarray can not identify balanced events, and metaphase chromosome spreads can only visualize some megabase size inversions. Next Generation Sequencing approaches tend to miss inversions because reads from inside the inversion map back to the reference without any indication that the orientation has changed. Detection of the breakpoints often fails, especially if the inversion is flanked by segmental duplications, repeat arrays or other non-unique sequences.

Bionano's imaging of extremely long molecules overcomes these obstacles to identifying inversions. Simulations of thousands of heterozygous inversions of various sizes demonstrated that our SV detection algorithms have high sensitivity to detect inversions larger than 30 kbp, reaching 98% sensitivity to pick up inversions larger than 70 kbp throughout the genome.

Bionano far outperforms other technologies in the detection of translocations. Thousands of translocations were simulated similarly to insertions and deletions in an in-silico map of the human reference genome hg19. The sensitivity for heterozygous translocations was shown to be 98% for breakpoint detection in both balanced and unbalanced translocations. Genome mapping can define the true positions of breakpoints within a median distance of 2.9 kbp, which is approximately 1000 times more precise than karyotyping and FISH. This accuracy is often sufficient for PCR and sequencing if single nucleotide resolution of the fusion point is desired for subsequent gene function studies.

In addition, translocation detection sensitivity was verified in two reference samples, NA16736 and NA21891, which are lymphoblast cell lines produced from blood cells from patients. One patient had a developmental disorders resulting in deafness with DNA repair deficiency caused by a t(9;22) translocation, and a second patient had Prader-Willi syndrome associated with a t(4;15) translocation. Both cell lines had been characterized by traditional cytogenetic methods. Bionano was able to detect both expected translocations as well as the reciprocal translocation breakpoints. Additionally, NA16736 contained a t(12:12) rearrangement which flanked an inverted segmental duplication. In NA21891, one translocation breakpoint could be localized within a gene, resulting in a predicted truncation (Fig. 8).

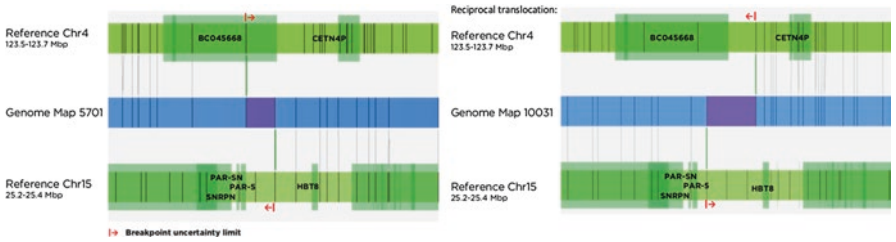

Fig. 8 Example of a translocation detected by Bionano mapping, associated with Prader-Willi syndrome. Blue bars are Bionano maps, and vertical lines represent Nt.BspQI label sites. For each of the reciprocal translocation breakpoints, maps are shown with alignments of the maps to chromosome 4 (top) and chromosome 15 (bottom) of the human reference hg19. Breakpoint resolution can be determined by the distance between matched and unmatched labels

Bionano Genomics developed a variant annotation pipeline (VAP) to help prioritize variants and to determine if a variant is relevant to the disease or phenotype of interest. In particular, it is useful for family-based and case-control studies. The two main components of the VAP are: (1) variant annotation, and (2) variant validation. The pipeline provides gene annotation and compares a given variant to variants detected in phenotypically normal control samples, including tumor versus control from the same patient. For a trio analysis, the pipeline annotates whether variants in the proband are found in the parents to help identify inherited and *de novo* variants. To validate variants, the pipeline examines assembly quality scores and aligns molecules against the assembly of interest to determine if the detected variants are well supported.

By using a control database of common variants, VAP filters the thousands of identified variants down to hundreds that are rare, or to a handful of e novo variants. It also identifies the genes they overlap with or are closest to in the genome. The VAP is part of Bionano Access, which provides an interface for setting up experiments on Saphyr, starting and monitoring instrument runs, launching *de novo* assemblies and SV calling, visualizing SVs, and annotating variants with the VAP. The results can be exported as a dbVar compliant VCF file, for easy integration with variants identified with NGS or other methods.

SV Detection in Cancer and Genetic Disease

Bionano mapping correctly diagnoses genetic disorders: In a publication in Genome Medicine, professor Eric Vilain of Children's National Medical Center, Washington, DC, presents molecular diagnoses using Bionano mapping of patients with Duchenne Muscular Dystrophy (DMD) (Barseghyan et al. 2017).

His team successfully mapped deletions, a duplication, and an inversion affecting the X-linked dystrophin gene, identifying deletions 45–250 kbp in size and an insertion of 13 kbp. The Bionano maps refined the location of deletion break points within introns compared to current PCR-based clinical techniques. They detected heterozygous SVs in carrier mothers of DMD patients as well, demonstrating the

ability of Bionano mapping to ascertain carrier status for large SVs. Vilain's team identified a 5.1 Mbp inversion involving the DMD gene, previously only identified by RNA sequencing of a muscle biopsy sample but missed by standard clinical methods (Fig. 9).

Bionano mapping also identifies genomic rearrangement in prostate cancer: Professor Vanessa Hayes at the Garvan Institute of Medical Research published a complete tumor-normal comparison from a primary prostate cancer (Jaratlerdsiri et al. 2017). Her team identified 85 large somatic deletions and insertions, of which half directly impact potentially oncogenic genes or regions. One such insertion, disrupting a gene known to be involved in cancer, is shown in Fig. 10.

Only one-tenth of these large SVs were detected using high-coverage short-read NGS and bioinformatics analyses using a combination of the best SV calling algorithms for NGS data. A manual inspection of NGS reads corresponding with the Bionano derived target regions verified 94% of the total SVs called with Bionano mapping. Many SVs detected with Bionano were flanked by repetitive sequences, making them all but invisible to short-read sequencing.

Targeted Known SV Detection as Biomarkers in Diagnostics and Companion Tests – Cytogenetics, Immuno-Repertoire Variation Mapping

Custom Labeling of Specific Sequences

A team from Drexel University has published several papers on a novel method to label any sequence of choice before imaging on a Bionano system (McCaffrey et al. 2016, 2017). An in vitro CRISPR/Cas9 RNA-directed nickase directs the specific labeling of a specific sequence motif that guide RNAs are designed against. In one application, they label human (TTAGGG)n DNA tracts in genomes that have also been barcoded using Bionano's standard labeling kits. High-throughput imaging and analysis of large DNA single molecules from genomes labeled in this fashion using Bionano's Irys or Saphyr permits mapping through subtelomere repeat element (SRE) regions to unique chromosomal DNA while simultaneously measuring the (TTAGGG)n tract length at the end of each large telomere-terminal DNA segment. This method enables global subtelomere and haplotype-resolved analysis of telomere lengths at the single-molecule level. Similarly, this team labeled HIV insertion sites and a variety of other repeat sequences.

With this custom labeling method, virtually any part of the genome can be studied in detail with Bionano mapping, even those parts which don't have identifiable patterns using Bionano's standard motif labeling.

Targeted Enriched Genomic Regions

Bionano mapping is typically performed on a whole genome scale. To enable collection of higher depth coverage of genomic regions of interest, or map a region much faster, a team from Tel Aviv University published a method for isolation and enrichment of a large genomic region of interest for targeted analysis based on Cas9

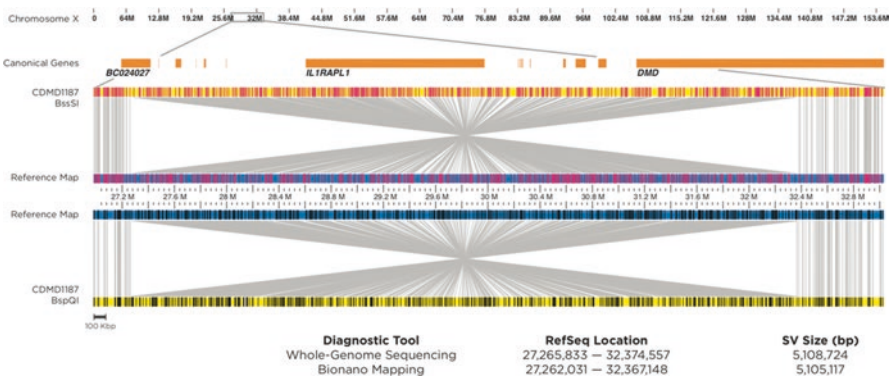

Fig. 9 A 5.1 Mbp inversion affecting the dystrophin gene detected in a patient with Duchenne Muscular Dystrophy. The inversion was detected twice, independently, in maps generated from patient DNA labeled after nicking with Nb.BssSI (top) and Nt.BspQI (bottom) nicking endonucleases. In both cases the inverted alignment of patient maps (top and bottom) relative to the reference (middle) is shown. Label sites are represented by red (Nb.BssSI) or black (Nt.BspQI) vertical lines in patient maps and reference, with grey match lines showing the aligned sites. RefSeq genes (orange) and the location of the inversion on the X-chromosome are shown at the top. (Barseghyan et al. 2017)

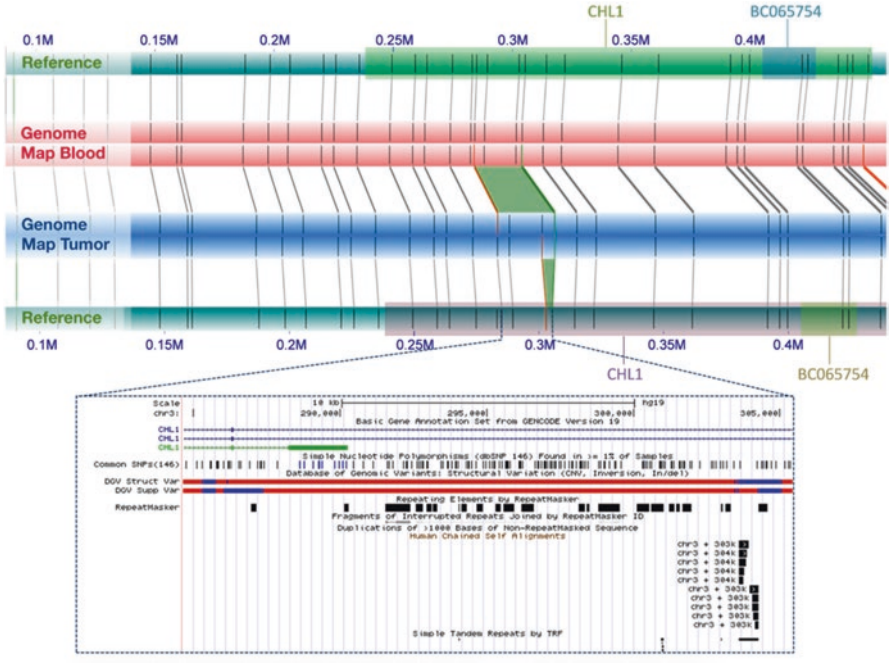

Fig. 10 A 4-kbp somatic insertion within the CHL1 gene on chromosome 3 identified in the prostate tumor of UP2153 using Bionano Mapping. The tumor map (blue track) shows a 2.5-kbp insertion (Chr3: 302.9–305.4 kbp) relative to hg19 (blue track), defined by a tandem repeat interval (inset). However, direct comparison of the tumor to genome maps derived from blood of the same patient (red track) found a larger 4-kbp insertion. (Jaratlerdsiri et al. 2017)

excision of two sites flanking the target region and isolation of the excised DNA segment by pulsed field gel electrophoresis (Gabrieli et al. 2017). The isolated gel fragment is then used in Bionano's standard DNA isolation and labeling workflow. The result is a highly enriched sample that can be mapped with Bionano or sequenced. In addition, analysis is performed directly on native genomic DNA that retains genetic and epigenetic composition without amplification bias. This method enables detection of mutations and structural variants as well as detailed analysis by generation of hybrid scaffolds composed of Bionano maps and sequencing data at a fraction of the cost of whole genome sequencing.

Immune Repertoire Mapping

The MHC region of the genome has a higher density of genes and of identified disease-causing variants than any other part of the human genome. It is prone to rearrangements, and sequencing based methods are unable to correctly identify and phase the structure of this region. In a Nature Biotech paper, the authors describe constructing Bionano maps covering the 4.7 Mbp MHC region from two individuals and performing *de novo* sequence assembly using NGS reads (Lam et al. 2012). The maps and NGS contigs were then compared to the reference sequences reported by the MHC Haplotype Consortium as confirmation and to uncover potential differences.

Employing this method, the study found and confirmed a number of interesting genomic features, including a 4 kb error in one reference sequence, anchoring and gap sizing of four NGS contigs, identification of misassembled NGS contigs, differentiation of the two HLA-DRB1 variants, and definition of numerous structural variants, such as a 5 kb insertion and 30 kb tandem duplication.

A second team studied the MHC region and other complex parts of the genome, in the YH reference genome (Cao et al. 2014). They used Bionano maps to comprehensively discover genome-wide SVs and characterize complex regions of the YH genome using long single molecules. They analyzed the structure of some complex regions of the human genome, including MHC also called Human Leukocyte Antigen (HLA), Killer-cell Immunoglobulin-like Receptor (KIR), IGL/IGH. The YH genome had Asian-specific structural variants in each of these regions. In addition to the MHC region, we also detected Asian/YH-specific structural differences in KIR (Fig. 11), compared to the reference genome.

Other Applications

Ultra-Long Range Epigenetic Pattern Mapping

In a recent prepublication (Grunwald et al. 2017), a team from Tel Aviv University working with Bionano scientists present a method to fluorescently label DNA molecules based on their methylation patterns. Using a methylation sensitive methyltransferase M.TaqI, a green fluorescent dye is attached to megabase size DNA when

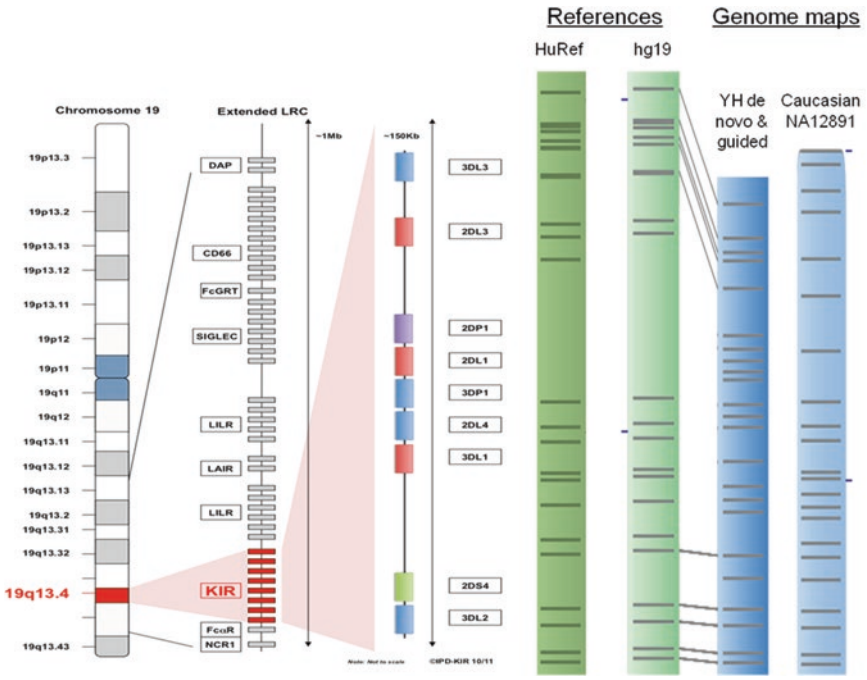

Fig. 11 Consensus genome maps compared to hg19 in the KIR region. The green bars represent the hg19 in silico motif map; the blue bars represent consensus genome maps. The YH genome map shows a huge variation relative to hg19 and HuRef human reference sequences. KIR: killer cell immunoglobulin-like receptor. (Cao et al. 2014)

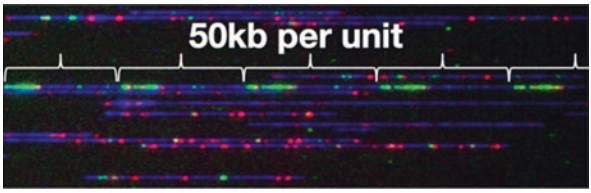

Fig. 12 Individual DNA molecules (blue) are stretch horizontally in NanoChannel arrays. Sequence motifs are labeled red, unmethylated sequences are labeled green, showing a repeating sequence with 50 kbp spacing. (Grunwald et al. 2017)

the enzyme's recognition sequence is present without CpG methylation. Bionano's standard nickase is then used with a red dye to allow for identification of the molecules and for assembly of the genome. In Fig. 12, a green signal is repeated every 50 kbp – this is an unmethylated CpG island in a 50 kbp repeat.

This technology opens up an entirely new field of research: we can now study if the methylation status of the promotor of a gene influences that of another promotor hundreds of kbp away on single molecule. This compares extremely favorably to the standard methylation analysis methods, in which DNA is chemically converted

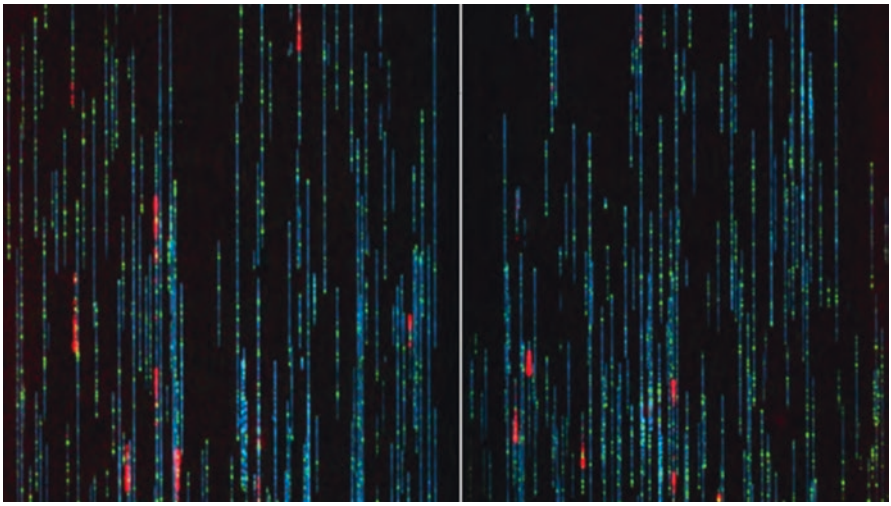

Fig. 13 Individual DNA molecules (blue) are stretch vertically in NanoChannel arrays. Sequence motifs are labeled green, DNA replication origins are shown in red. (Klein et al. 2017)

using sodium bisulfite, followed by array hybridization or sequencing. Bisulfite conversion damages the DNA, and only very fragmented DNA molecules can be isolated and single molecule methylation patterns can be measured over no more than a few hundred basepairs at best.

The proof of concept study presented here demonstrates that we can now read the genome wide methylation profile of cells on long, single molecules while simultaneously mapping major structural variation on these same molecules.

Dynamic Mapping of Genome Functions – Replication Imaging

Cell replication is essential to life, and uncontrolled replication of cells is the cause of cancer. Exactly where eukaryotic cells initiate replication is hard to analyze. Studies looking into replication origins have largely focused on simple organisms with smaller genomes. Observing this process in large genomes is difficult because eukaryotic cells have up to 50,000 replication start points per cell per cycle, and even the most commonly observed replication origin in the genome functions as such in just 10% of cells. Several groups have demonstrated visualization of these replication origins on Saphyr in the bacteriophage Lambda (De Carli et al. 2017) and in human cells (Klein et al. 2017). Synchronized and arrested HeLa cells are transfected with red fluorescent nucleotides, the cell cycle is allowed to resume and DNA prepared using Bionano's standard workflow. Sequence motifs are then labeled green using a Bionano's NLRS or DLS kits and imaged on Saphyr. The green signal is used to assemble and align the molecules to the reference, the red signal shows where on those molecules the replication originated. The resulting images (Fig. 13) are stunning and the 290x coverage of the genome allows the team to identify early-firing human replication origins that occur in as few as 1% of cells.

References

Barseghyan H, et al. Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. Genome Med. 2017;9:90.

Bickhart DM, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.

Cao H, et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience. 2014;3:34.

De Carli F, Menezes N, Berrabah W, Barbe V, Genovesio A, Hyrien O. High-throughput optical mapping of replicating DNA. Small Methods. 2017;2(9):1800146. https://doi.org/10.1101/239251.

de Koning AP, et al. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7(12):e1002384.

Gabrieli T, et al. Cas9-assisted targeting of CHromosome segments (CATCH) for targeted nanopore sequencing and optical genome mapping. 2017. https://doi.org/10.1101/110163.

Grunwald A, et al. Reduced representation optical methylation mapping (R2OM2). 2017. https://doi.org/10.1101/108084.

Huddleston J, Eichler EE. An incomplete understanding of human genetic variation. Genetics. 2016;202(4):1251–4.

Huddleston J, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27(5):677–85.

Jaratlerdsiri W, et al. Next generation mapping reveals novel large genomic rearrangements in prostate cancer. Oncotarget. 2017;8:23588–602.

Jiao Y, Peluso P, Ware D. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(22):524–7. https://doi.org/10.1038/nature22971.

Klein K, et al. Genome-wide identification of early-firing human replication origins by optical replication mapping. 2017. https://doi.org/10.1101/214841.

Lam ET, et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012;30:771–6.

Lee H, et al. Clinical exome sequencing for genetic identification of rare mendelian disorders. JAMA. 2014;312(18):1880–7.

Mak A, et al. Genome-wide structural variation detection by genome mapping on NanoChannel arrays. Genetics. 2016;202:351–62.

McCaffrey J, et al. CRISPR-CAS9 D10A nickase target-specific fluorescent labeling of double strand DNA for whole genome mapping and structural variation analysis. Nucleic Acids Res. 2016;44:e11.

McCaffrey J, et al. High-throughput single-molecule telomere characterization. Genome Res. 2017;27:1904–15.

Miller DT, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749–64.

Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.

Informatics for PacBio Long Reads

Yuta Suzuki

Abstract In this article, we review the development of a wide variety of bioinformatics software implementing state-of-the-art algorithms since the introduction of SMRT sequencing technology into the field. We focus on the three major categories of development: read mapping (aligning to reference genomes), *de novo* assembly, and detection of structural variants. The long SMRT reads benefit all the applications, but they are achievable only through considering the nature of the long reads technology properly.

Advances in SMRT Biology and Challenges in Long Read Informatics

In 2011, advent of the PacBio RS sequencer and its SMRT (single molecule real-time) sequencing technology revolutionized the concept of DNA sequencing. Longer reads are promised to generate de novo assembly of much higher contiguity, and the claim was proved by several assembly projects (Steinberg et al. 2014; Pendleton et al. 2015; Seo et al. 2016). The lack of sequencing bias was proved to be able to read regions which are extremely difficult for NGS (Next Generation Sequencers) (Loomis et al. 2013).

None of these achievement, however, was just straightforward application of conventional informatics strategy developed for short read sequencers; the virtue of the long reads was not free at all. As many careful skeptics claimed in the early history of PacBio sequencing, the long reads seemed too noisy. Base accuracy was around ~85% for single raw read, that is, ~15% of bases were wrong calls, and

Y. Suzuki (✉)
Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
e-mail: yuta_suzuki@edu.k.u-tokyo.ac.jp

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_8

indels consisted most of the errors. The higher error rate made it inappropriate to apply informatics tools designed for much accurate short read technologies.

Even the higher error rate is properly handled by sophisticated algorithms, the length of the reads itself can pose another problem. Computational burden of many algorithms depends on the read length L. When only the short reads are assumed, it may be considered as constant, e.g., L = 76, 150, etc. The emergence of long read sequencer changed the situation drastically by improving the read length by orders of magnitudes, to thousands of bases, and to tens of thousands of bases by now. Besides the ongoing innovations for longer reads, there is a large variation in length of sequencing reads even in the same sequencing run. Therefore, the assumption that the read length is constant is not valid anymore, and one must have a strategy to handle (variably) long reads in reduced time (CPU hours) and space (memory footprint) requirement.

Availability of long read opened a door to the set of problems which were biologically existing in real but implicitly ignored by studies using short read sequencing. For example, we had to realize that a non-negligible fraction of reads could cover SVs (structural variants), requiring a new robust mapping strategy other than simply masking the known repetitive regions.

Consequently, many sophisticated algorithms had to be developed to resolve these issues; how to mitigate higher error rate, and how it can be done efficiently for long reads. The rest of this article covers some important innovations achieved and ongoing efforts in informatics area to make the most of long reads data.

Aligning Noisy Long Reads with Reference Genome

When one aligns long reads against reference sequence, one must be aware that the variations between reads and reference stems from two conceptually separate causes. On one hand, there are sequencing errors in its simple sense, which is discrepancy between a read observed and actual sequence being sequenced. On the other hand, we expect a sample sequenced would have slightly different sequence than a reference sequence (otherwise there is no point in doing sequencing), and those difference are usually called variants. Though sequencing errors and sequence variants are conceptually different, however, they both appears just as “errors” to us unless they have some criteria to distinguish them. The next two examples are for understanding why the distinction between two classes of “error” is relevant here.

Let’s consider we have some *noisy* reads. Clearly, we cannot call sequence variants specific to the sample unless the frequency of sequencing errors is controlled to be sufficiently low compared to the frequency of variants. This is the reason why it is difficult for noisy reads to detect small nucleotide variants such as point mutations and indels.

Next, assume we have *long* reads. Then, there are more chances that the reads span the large variations such as structural variations (SVs) between a reference genome and the sample sequenced. This situation is problematic for aligners who

considered any possible variation between reads and reference to be sequencing errors, for such aligners would fail to detect correct alignment as they need to introduce too much errors for aligning these sequences. Some aligners try to combat the situation by employing techniques such as chaining and split alignment. Some aligners (NGMLR, Minimap2) explicitly introduce an SV-aware scoring scheme such as a two-parts concave gap penalty, which reflects the two classes of variations between read and reference.

Sequence alignment is so fundamental in sequence analysis that it finds its application everywhere. For example, mapping sequencing reads to reference genome is the very first step of resequencing studies. Accuracy of mapping can directly be translated into the overall reliability of results. Also, mapping is often one of the most computationally intensive steps. Therefore, accurate and faster mapping software would benefit the whole area of resequencing studies. In the context of de novo assembly pipeline, it is used for detecting overlap among long reads. Of noted, desired balance of sensitivity and specificity of overlap detection is controlled differently from mapping to reference, and it could often be very subtle.

Though it is more or less subjective to make distinction between standalone aligners and aligners designed as a module of assembly pipeline or SV detection pipeline, we decided to cover some aligners in other sections. MHAP will be introduced in relation with Canu in the section devoted to assembly tools. Similarly, NGMLR will be detailed together with Sniffle in the section for SV detection.

BWA-SW and BWA-MEM

Adopting the seed-and-extend approach, BWA-SW (Li & Durbin 2010) builds FM-indices for both query and reference sequence. Then, DP (dynamic programming) is applied to these FM-indices to find all local matches, i.e., seeds, allowing mismatches and gaps between query and reference. Detected seeds are extended by Smith-Waterman algorithm. Some heuristics are explicitly introduced to speed up alignment of large-scale sequencing data and to mitigate the effect of repetitive sequences. BWA-MEM (Li 2013) inherits similar features implemented in BWA-SW such as split alignment, but is found on a different seeding strategy using SMEM (supermaximal exact matches) and *reseeding* technique to reduce mismapping caused by missing seed hits.

BLASR

BLASR (Chaisson & Tesler 2012) (Basic Local Alignment with Successive Refinement) is also one of the earliest mapping tools specifically developed for SMRT reads. Like BWA-MEM, it is probably the most widely used one to date. Bundled with official SMRT Analysis, it has been the default choice for the

mapping (overlapping) step in all protocols such as resequencing, de novo assembly, transcriptome analysis, and methylation analysis. In the BLASR's paper, the authors explicitly stated it was designed to combine algorithmic devices developed in two separate lines of studies, namely, a coarse alignment method for whole genome alignment and a sophisticated data structure for fast short read mapping. Proven to be effective for handling noisy long read, the approach of successive refinement, or seed-chain-align paradigm, has become a standard principle.

BLASR first finds short exact matches (anchors) using either suffix array or FM index (Ferragina & Manzini 2000). Then, the regions with clustered anchors aligned colinearly are identified as candidate mapping locations, by global chaining algorithm (Abouelhoda & Ohlebusch 2003). The anchors are further chained by sparse dynamic programming (SDP) within each candidate region (Eppstein et al. 1992). Finally, it gives detailed alignment using banded DP (dynamic programming) guided by the result of SDP. BLASR achieved tenfold faster mapping of reads to human genome than BWA-SW algorithm at comparable mapping accuracy and memory footprint.

DALIGNER

DALIGNER (Myers 2014) is specifically designed for finding overlaps between noisy long reads, though its concept can also be adopted for a generic long read aligner, as implemented in DAMAPPER (https://github.com/thegenemyers/DAMAPPER). Like in BLASR, DALIGNER also performs filter based on short exact matches. Instead of using BWT (FM index), it explicitly processes k-mers within reads by thread-able and cache coherent implementation of radix sort. Detected k-mers are then compared via block-wise merge sort, which reduces memory footprint to a constant depending only on the block size. To generate local alignment, it applies O(ND) diff algorithm between two candidate reads (Myers 1986). DALIGNER achieved 22 ~ 39-fold speedup over BLASR at higher sensitivity in detecting correct overlaps (Myers 2014). DALIGNER is supposed to be a component for read overlap (with DAMASKER for repeat masking, DASCRUBBER for cleaning up low quality regions, and a core module for assembly) of DAZZLER de novo assembler for long noisy reads, which will be released in future.

Minimap2

Minimap2 (Li 2017) is one of the latest and state-of-the-art alignment program. Minimap2 is general-purpose aligner in that it can align short reads, noisy long reads, and reads from transcripts (cDNA) back to a reference genome. Minimap2 combines several algorithmic ideas developed in the field, such as locality-sensitive

hashing as in Minimap and MHAP. For accounting possible SVs between reads and genome, it employs concave gap cost as in NGMLR, and it is efficiently computed using formulation proposed by Suzuki & Kasahara (2017). In addition to these features, the authors further optimized the algorithm, by transforming the DP matrix from row-column coordinate to diagonal-antidiagonal coordinate for better concurrency in modern processors. According to the author of Minimap2, it is supposed to replace BWA-MEM, which is in turn a widely used extension of BWA-SW.

De novo Assembly

As Lander-Waterman theory (Lander & Waterman 1988) would assert, the longer input reads are quite essential in achieving a high-quality genome assembly for repetitive genomes. Therefore, developing a *de novo* assembler for long read is naturally the most active area in the field of long read informatics.

To our knowledge, almost all assemblers published for long read take an overlap-layout-consensus (OLC) approach, where the overall task of assembly can be divided into the three steps. (1. Overlap) The overlaps between reads are identified as candidate pairs representing the same genomic regions, and the overlap graph is constructed to express these relations. (2. Layout) The graph is transformed to generate linear contigs. The step often starts by constructing the string graph (Myers 2005), a string-labeled graph which encodes all the information in reads observed, and eliminates edges containing redundant information. (3. Consensus) The final assembly is polished. To eliminate errors in contigs, consensus is taken among reads making up the contigs.

Though we do not cover tools for the consensus step here, there are many of them released to date including official Quiver and Arrow bundled in SMRT Analysis (https://github.com/PacificBiosciences/GenomicConsensus), another official tool pbdagcon (https://github.com/PacificBiosciences/pbdagcon), Racon (Vaser et al. 2017), and MECAT (Xiao et al. 2017). Of note, quality of a polished assembly can be much better than a short-read-based assembly due to the randomness of sequencing errors in long reads (Chin et al. 2013; Myers 2014).

FALCON

FALCON (Chin et al. 2016) is designed as a diploid-aware *de novo* assembler for long read. It starts by carefully taking consensus among the reads to eliminate sequencing errors while retaining heterozygous variants which can distinguish two homologous chromosomes (FALCON-sense). For constructing a string graph, FALCON runs DALIGNER. The resulted graph contains "haplotype-fused" contigs and "bubbles" reflecting variations between two homologous chromosomes. Finally, FALCON-unzip tries to resolve such regions by phasing the associated long reads

and local re-assembly. The contigs obtained are called “haplotigs”, which are supposed to be faithful representation of individual alleles in the diploid genome.

Canu (& MHAP)

MHAP (Berlin et al. 2015) (Min-Hash Alignment Process) utilized MinHash for efficient dimensionality reduction of the read space. In MinHash, H hash functions are randomly selected, each of them maps k-mer into an integer. For a given read of length L, only the minimum values over the read are recorded for each of H hash functions. The k-mers at which the minimum is attained are called *min-mers*, and resulted representation is called a *sketch*. The sketch serves as a locality sensitive hashing of each read, for the similar sequences are expected share similar sketches. Because the sketch retains the data only on H min-mers, its size is fixed to H, independent of read length L.

Built on top of MHAP, Canu (Koren et al. 2017) extends best overlap graph (BOG) algorithm (Miller et al. 2008) for generating contigs. A new “bogart” algorithm estimates an optimal overlap error rate instead of using predetermined one as in original BOG algorithm. This requires multiple rounds of read and overlap error correction, but eventually enables to separate repeats diverged only by 3%. Though BOG algorithm is greedy, the effect is mitigated in Canu by inspecting non-best overlaps as well to avoid potential misassemblies.

HINGE

While there is no doubt that obtaining more contiguous (i.e., higher contig N50) assembly is a major goal in genome assembly, the quest just for longer N50 may cause misassemblies if the strategy gets too greedy. Being aware that danger, HINGE (Kamath et al. 2017) aims to perform the optimal resolution of repeats in assembly, in the sense that the repeats should be resolved if and only if it is supported by long read data available. To implement such a strategy is rather straightforward for de Bruijn graphs. In de Bruijn graph, its k-mers representing nodes are connected by edges when they co-occur next to each other in reads. In ideal situation, the genome assembly is realized as an Eulerian path, i.e., trail which visits every edge exactly once, in the de Bruijn graph. However, de Bruijn graphs are not robust for noisy long read, so overlap graphs are usually preferred for long read. One of the key motivations of HINGE is to give such a desirable property of de Bruijn graphs, to overlap graphs which is more error-resilient. To do so, HINGE enriches string graph with additional information called “hinges” based on the result of the read overlap step. Then, assembly graph with optimal repeat resolution can be constructed via a hinge-aided greedy algorithm.

Miniasm (& Minimap)

Minimap (Li 2016) adopts a similar idea as MHAP, it uses *minimizers* to represents the reads compactly. For example, Minimap uses a concept of (w,k)-minimizer, which is the smallest (in the hashed value) k-mer in w consecutive k-mers. To perform mapping, Minimap searches for colinear sets of minimizers shared between sequences. Miniasm (Li 2016), an associated assembly module, generates assembly graph without error-correction. It firstly filters low-quality reads (chimeric or with untrimmed adapters), constructs graph greedily, and then cleans up the graph by several heuristics, such as popping small bubbles and removing shorter overlaps.

Detection of Structural Variants (SVs)

Sequence variants are called *structural* when they are explained by the mechanisms involving double-strand breaks, and are often defined to be variants larger than certain size (e.g., 50 bp) for the sake of convenience. They are categorized into several classes such as insertions/deletions (including presence/absence of transposons), inversion, (segmental) duplication, tandem repeat expansion/contraction, etc. While some classes of SVs are notoriously difficult to detect via short reads (especially long inversions and insertions), long reads have promise to detect more of them by capturing entire structural events within sequencing reads.

PBHoney

PBHoney (English, Salerno & Reid 2014) implements combination of two methods for detecting SVs via read alignment to reference sequence. Firstly, PBHoney exploits the fact that the alignment of reads by BLASR should be interrupted (giving soft-clipped tails) at the breakpoints of SV events. PBHoney detects such interrupted alignments (piece-alignments) and clusters them to identify individual SV events. Secondly, PBHoney locates SVs by examining the genomic regions with anomalously high error rate. Such a large discordance can signal the presence of SVs because sequencing errors within PacBio reads are supposed to distribute rather randomly.

Sniffles (& NGMLR)

NGMLR (Sedlazeck et al. 2017) is a long-read aligner designed for SV detection, which uses two distinct gap extension penalties for different size range of gaps (i.e., concave gap penalty) to align entire reads over the regions with SVs. Intuitively, the

concave gap penalty is designed so that it can allow longer gaps in alignment while shorter gaps are penalized just as sequencing errors. Adopting such a complicated scoring scheme makes the alignment process computationally intensive (Miller et al. 1988), but NGMLR introduces heuristics to perform faster alignment. Then, an associated tool to detect SVs, Sniffles scans the read alignment to report putative SVs which are then clustered to identify individual events and evaluated by various criteria. Optionally, Sniffle can infer genotypes (homozygous or heterozygous) of detected variants, and can associate "nested SVs" which are supported by the same group of long reads.

SMRT-SV

SMRT-SV (Huddleston et al. 2017) is a SV detection tool based on local assembly. It firstly maps long reads to reference genome, against which SVs are called. Then it searches signatures of SVs within alignment results, and 60 kbp regions around the detected signatures are extracted. The regions are to be assembled locally from those reads using Canu, then SVs are called by examining the alignment between assembled contigs and reference. Local assembly is performed for other regions (without SV signatures) as well to detect smaller variants.

Beyond DNA – Transcriptome Analysis and Methylation Analysis

SMRT sequencing has been found its applications outside DNA analysis as well. When it is applied to cDNA sequencing, long read would be expected to capture the entire structures of transcripts to elucidate expressing isoforms comprehensively. IDP (Isoform Detection and Prediction) (Au et al. 2013) and IDP-ASE (Deonovic et al. 2017) are tools dedicated to analyze long read transcriptome data. To detect expressing isoforms from long read transcriptome data, IDP formulates it in the framework of integer programming. To estimate allele-specific expression both in gene-level and isoform-level, IDP-ASE then solves probabilistic model of observing each allele in short read RNA-seq. Both IDP and IDP-ASE effectively combines long read data for detection of overall structure of transcripts, and short read data for accurate base-pair level information.

In methylation analysis, official *kineticsTools* in SMRT Analysis has been widely used to detect base modification sites and to estimate sequence motives for DNA modification (see (Flusberg et al. 2010) for the principle of detection). Detecting 5-methyl-cytosines (5mC), which is by far the dominant type of DNA modification in plants and animals, is challenging due to their subtle signal. Designed for detecting 5mC modifications in large genomes within practical sequencing depth, AgIn

(Suzuki et al. 2016) exploits the observation that CpG methylation events in vertebrate genomes are correlated over neighboring CpG sites, and tries to assign the binary methylation states to CpG sites based on the kinetic signals under the constraint that a certain number of neighboring CpG sites should be in the same state. Making the most of high mappability of long read, AgIn has been applied to observe diversified CpG methylation statuses of centromeric repeat regions in fish genome (Ichikawa et al. 2017), and to observe allele-specific methylation events in human genomes.

Concluding Remarks

We have briefly described some innovative ideas in bioinformatics for an effective use of long read data. As concluding remarks, let me mention a few prospects for the future development in the field. By now, it is evident the quest for complete genome assembly is almost done, but the remaining is the most difficult part such as extremely huge repeats, centromeres, telomeres. While many state-of-the-art assemblers take the presence of such difficult regions into account and can carefully generate high quality assembly for the rest of genomes, it is remained open how to tackle these difficult part of the genome, how to resolve its sequence, not escaping from them.

Base modification analysis using PacBio sequencers may also have huge potential to distinguish several types of base modifications and to detect them simultaneously in the same sample (Clark et al. 2011), but only the limited number of modification types (6 mA, 4mC, and 5mC) are considered for now. This is mainly due to the technical challenge to alleviate noise in kinetics data to distinguish each type of modifications and unmodified bases from each other.

That said, it will be no doubt that the field would be more attractive than ever, as the use of long read sequencer becomes a daily routine in every area of biological research, or maybe even in clinical practice.

Acknowledgements I'd like to thank Yoshihiko Suzuki, Yuichi Motai and Dr./Prof. Shinichi Morishita for insightful comments on the draft.

References

Abouelhoda MI, Ohlebusch E. A local chaining algorithm and its applications in comparative genomics. International workshop on algorithms in bioinformatics. Berlin/Heidelberg: Springer; 2003.

Au KF, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci. 2013;110(50):E4821–30.

Berlin K, et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33(6):623–30.

Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13(1):238.
Chin C-S, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–9.
Chin C-S, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–4.
Clark TA, et al. Direct detection and sequencing of damaged DNA bases. Genome Integr. 2011;2(1):10.
Deonovic B, et al. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. Nucleic Acids Res. 2017;45(5):e32.
English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014;15(1):180.
Eppstein D, et al. Sparse dynamic programming I: linear cost functions. J ACM (JACM). 1992;39(3):519–45.
Ferragina P, Manzini G. Opportunistic data structures with applications. Foundations of computer science, 2000. Proceedings. 41st annual symposium on. IEEE, 2000.
Flusberg BA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7(6):461–5.
Huddleston J, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27(5):677–85.
Ichikawa K, et al. Centromere evolution and CpG methylation during vertebrate speciation. Nat Commun. 2017;8(1):1833.
Kamath GM, et al. HINGE: long-read assembly achieves optimal repeat resolution. Genome Res. 2017;27(5):747–56.
Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988;2(3):231–9.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv. 2013:1303.3997.
Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32(14):2103–10.
Li H. Minimap2: versatile pairwise alignment for nucleotide sequences. arXiv. 2017:1708.
Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589–95.
Loomis EW, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 2013;23(1):121–8.
Miller W, Myers EW. Sequence comparison with concave weighting functions. Bull Math Biol. 1988;50(2):97–120.
Miller JR, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24(24):2818–24.
Myers EW. An O (ND) difference algorithm and its variations. Algorithmica. 1986;1(1):251–66.
Myers EW. The fragment assembly string graph. Bioinformatics. 2005;21(Suppl_2):ii79–85.
Myers G. Efficient local alignment discovery amongst noisy long reads. International workshop on algorithms in bioinformatics. Berlin/Heidelberg: Springer; 2014.
Pendleton M, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015;12(8):780–6.
Sedlazeck FJ, et al. Accurate detection of complex structural variations using single molecule sequencing. bioRxiv. 2017:169557.
Seo J-S, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538:243–7.
Steinberg KM, et al. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. 2014;24(12):2066–76.

Suzuki H, Kasahara M. Acceleration of nucleotide semi-global alignment with adaptive banded dynamic programming. bioRxiv. 2017:130633.
Suzuki Y, et al. AgIn: measuring the landscape of CpG methylation of individual repetitive elements. Bioinformatics. 2016;32(19):2911–9.
Vaser R, et al. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
Xiao C-L, et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods. 2017;14(11):1072–4.

Challenges of Single-Molecule DNA Sequencing with Solid-State Nanopores

Yusuke Goto, Rena Akahori, and Itaru Yanagi

Abstract A powerful DNA sequencing tool with high accuracy, long read length and high-throughput would be required more and more for decoding the complicated genetic code. Solid-state nanopore has attracted many researchers for its promising future as a next-generation DNA sequencing platform due to the processability, the robustness and the large-scale integratability. While the diverse materials have been widely explored for a solid-state nanopore, silicon nitride (Si_3N_4) is especially preferable from the viewpoint of mass production based on semiconductor fabrication process. Here, as a nanopore sensing mechanism, we focused on the ionic blockade current method which is the most developed technique. We also highlight the main challenges of Si_3N_4 nanopore-based DNA sequencer that should be addressed: the fabrication of ultra-small nanopore and ultra-thin membrane, the modulation of DNA translocation speed and the detection of base-specific signals. In this chapter, we discuss the recent progress relating to solid-state nanopore DNA sequencing, which helps to provide a comprehensive information about the current technical situation.

Introduction

Essential information required for living organisms is encoded into their genomes, and deciphering this genetic code via DNA sequencing is the first step in further understanding organisms. DNA sequencing is a powerful tool for disclosing genetic variations at a molecular, biological level, such as single-nucleotide polymorphism, copy number variation, gene fusion, and insertion/deletion, and these genetic variations are related to various diseases, including cancer (Shendure et al. 2017). The significance of DNA sequencing in elucidating disease mechanisms and improving

Y. Goto (✉) · R. Akahori · I. Yanagi
Center for Technology Innovation – Healthcare, Research & Development Group, Hitachi Ltd., Kokubunji-shi, Tokyo, Japan
e-mail: yusuke.goto.bo@hitachi.com

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_9

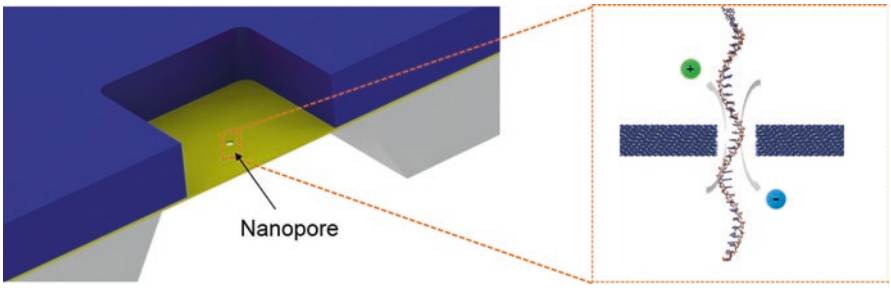

Fig. 1 Schematic of a solid-state nanopore embedded in an ultra-thin, insulating Si_3N_4 membrane supported by a silicon chip. A nanopore is fabricated in the ultra-thin Si_3N_4 membrane, and the DNA molecule is typically electrophoretically driven through the nanopore. The DNA sequence is read out at the narrowest part of the nanopore during translocation

genetic diagnoses cannot be overlooked. Due to enormous support and investments from governments and the private sector, the scientific and economic growth in this field shows no signs of slowing down over the last decade.

DNA sequencing techniques should have a high accuracy, long read length, high throughput and low cost (Goodwin et al. 2016). Additionally, emerging technologies, which are represented by single-molecule sequencing, are expected to reduce the PCR-derived bias and allow longer reads. Furthermore, these single-molecule techniques can detect methylation and footprinting information. One of the most promising approaches with the above attributes is the use of DNA translocation across nanopores, which is called nanopore sequencing (Branton et al. 2008).

Nanopore sequencing is a label-free, single-molecule approach in which DNA molecules are driven through an ultra-tiny one-dimensional channel and produce electrical signals corresponding to the DNA sequence (Venkatesan and Bashir 2011). When negatively charged DNA is driven by an external electric field through a nanopore in an ultra-thin membrane between two electrolytes, the target DNA molecule temporarily occupies the nanopore (Fig. 1). This occupancy results in a change in the electrical signal. For example, because the pore size is close to the molecular diameter of DNA, each base (A, G, C, and T) produces its own characteristic blockage current, which is similar to a fingerprint. This approach, i.e., detection of the blockage current, allows DNA to be sequenced with a single-base resolution and without the need for fluorescent labels.

Presently, nanopores are categorized as two types: biological nanopores and solid-state nanopores. Biological nanopores rely on the use of transmembrane proteins (called porins), such as alpha-hemolisin (α-HL) (Clarke et al. 2009) or Mycobacterium smegmatis porin A (MspA) (Derrington et al. 2010), with a nanometer-scale hole. Recently, great achievements, including *de novo* assembly of the human genome (Jain et al. 2018), have been achieved for practical DNA sequencing with biological nanopores. However, biological nanopores may still be limited by short lifetimes and the intrinsic instability of natural proteins, and therefore, they are not especially favorable for long-term operations. Solid-state nanopores are

typically fabricated by top-down lithography, and they are expected to offer a number of practical advantages over their biological counterparts. They have superior mechanical, thermal and chemical stabilities (Dekker 2007), the potential for large-scale integration with on-chip electronics, and size tunability (Kwok et al. 2014; Yanagi et al. 2014). These advantages endow solid-state nanopore applications with the potential for commercial expansion. Therefore, over the past decade, many studies have been conducted to realize DNA sequencing based on solid-state nanopores (Lindsay 2016). This review briefly introduces recent progress in DNA sequencing with a solid-state nanopore. Further detailed issues are discussed in a large number of comprehensive reviews and books (Dekker 2007; Branton et al. 2008; Iqbal and Bashir 2011; Wanunu 2012; Edel and Albrecht 2013; Carson and Wanunu 2015; Lindsay 2016).

Solid-State Nanopore DNA Sequencing Based on Ionic Current Detection

Nanopore DNA sequencing procedures strongly rely on the sensing mechanism. Based on the type of signal, the detection methods can be roughly classified into two types: electrical readout and optical detection. In particular, electrical detection is preferable due to its potential to reduce the cost and size of the instruments required for the procedures. Electrical sensing methods can be categorized into three types: ionic blockade current (Howorka et al. 2001), tunneling current (Tsutsui et al. 2010) and capacitance variation signal (Sigalov et al. 2008). Among them, the most widely developed technique is measuring ionic currents during DNA translocation through a nanopore (Iqbal and Bashir 2011; Edel and Albrecht 2013). In this concept (Fig. 2), the bases produce unique ionic current patterns as the DNA traverses across the nanopore, and these patterns are decoded into the original DNA sequence (Pennisi 2012). When two chambers containing electrolytic fluid are separated with an

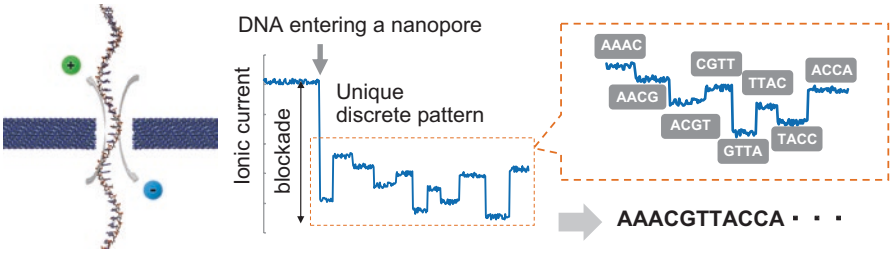

Fig. 2 Schematic concept of the sensing mechanism for the ionic blockade current from nanopore DNA sequencing. When a DNA molecule is driven through a nanopore, the steady ionic current flux is blocked by the geometric exclusion due to DNA. Since this blockade depends on a sequence of several nucleotides, the DNA translocation generates unique discrete patterns. The original DNA sequence can be determined by deciphering the ionic current pattern

ultra-thin membrane containing a nanometer-scale hole, the application of a voltage bias across the membrane can generate a stable ionic transmembrane current. A single-stranded DNA (ssDNA) translocating through the nanopore blocks the ion fluxes induced by geometrical exclusion, and the blockade current pattern corresponding to the DNA sequence is recorded as a function of time. The core concept of this sequencing method is based on the assumption that the changes in the ionic currents strongly correlate with the sequence of several nucleotides located at the narrowest part of the nanopore during the translocation (Manrao et al. 2012; Laszlo et al. 2014).

In this scheme, strict requirements must be simultaneously satisfied to achieve actual DNA sequencing. In particular, the device must be fabricated considering three key factors: an ultra-small nanopore, an ultra-thin membrane and control of the DNA speed. First, creating an ultra-small nanopore with a diameter size similar to that of DNA is essential to prevent multiple or folded DNA strands from entering the nanopore. Considering that the diameter of a single-stranded DNA is approximately 1.4 nm, the preferred nanopore diameter is less than 2 nm. Second, the fabrication of an ultra-thin membrane is equally important. The number of nucleotides located at the nanopore during DNA translocation increases as the membrane thickness increases. As a result, detecting the unique signal from each nucleotide becomes more difficult. Consequently, a thinner membrane is desirable because the membrane determines the spatial resolution of the nanopore sequencing. The third issue is controlling the DNA motion across the nanopore. The DNA quickly translocates through the nanopore, i.e., typically >1 μs per nucleotide (μs/nt) (Akahori et al. 2014), and recording ultra-fast signals with a low noise is difficult using the currently available amplifiers. Thus, tremendous efforts have been devoted to resolving these challenging issues.

Fabrication of a Single Nanopore with a 1–2 nm Diameter in a Solid-State Membrane

Solid-state nanopore technology has advantages in terms of its robustness and possible large-scale integration. However, this technology has a serious drawback, i.e., the nanopore fabrication process. Fabricating solid-state nanopores requires dimensional control at the sub-nm scale, and this can be successfully achieved by means of focused-electron beam etching via transmission electron microscopy (TEM). A TEM beam can be condensed to a diameter of less than 1 nm and used to successfully fabricate a small nanopore with a diameter of less than 2 nm (Storm et al. 2003; Larkin et al. 2013; Venta et al. 2013). For mass production, however, TEM-beam etching is not suitable because of its high cost, low throughput, and complexity. This labor-intensive, low-throughput, non-scalable, high-cost, and sequential fabrication process also requires the constant presence of a machine operator, which

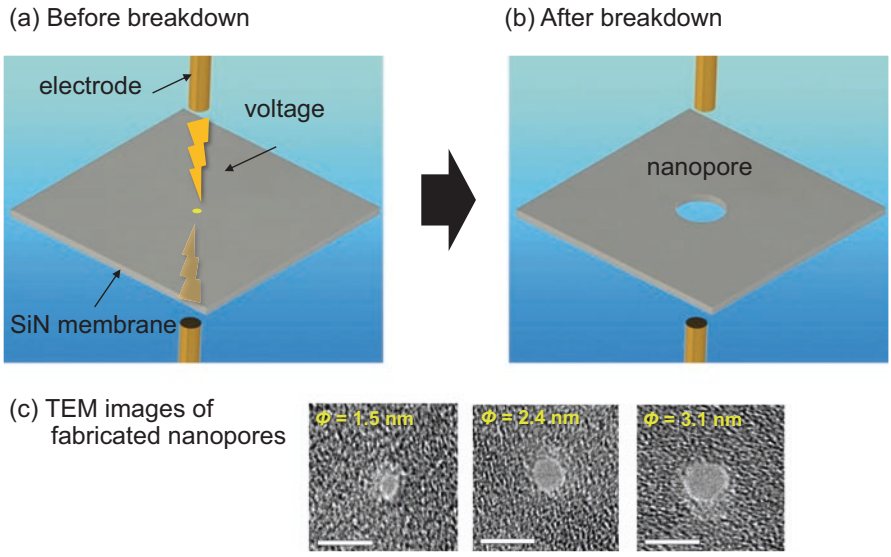

Fig. 3 Nanopore fabrication via CBD. (**a**) The voltage is applied against an insulating Si_3N_4 membrane. (**b**) This strong electric field induces the dielectric breakdown of the insulating membrane, generating a single nanopore with an ultra-small diameter. (**c**) Typical TEM images of fabricated nanopores with diameters of 1–3 nm

significantly limits the yield of nanopore devices. Therefore, an alternative nanopore fabrication method is strongly desired.

Recently, promising alternative fabrication methods based on controlled breakdown (CBD) have been proposed and demonstrated as shown in Fig. 3a, b (Kwok et al. 2014; Yanagi et al. 2014). Dielectric breakdown is local failure due to the electron-trap density reaching a critical value (Briggs et al. 2015), and this is induced by producing a strong electric field across an insulating membrane with two Ag/AgCl electrodes. A voltage-induced dielectric breakdown event in ultra-thin membranes results in the formation of single nanopores with diameters as small as 1 nm (Fig. 3c). The dielectric breakdown is caused by the strong electric field produced by two conventional Ag/AgCl electrodes, and no special setup is required to fabricate nanopores. For example, when using the multilevel pulse voltage injection technique, the size of the generated nanopores can be calculated by cycle-monitoring the ionic current that flows through the newly formed pore. Once the size reaches a preset value, the breakdown process can be stopped through feedback control under a low monitoring voltage. Thereafter, the diameter of the nanopore can be increased by applying short pulse voltages.

CBD has several advantages over conventional methods. First, CBD is an easy, simple and low-cost method that does not require special instruments. Second, if the method uses the appropriate parameters for single nanopore creation, it can precisely fabricate nanopores with a sub-nanometer accuracy under the appropriate conditions (Briggs et al. 2014; Goto et al. 2016). Third, this method also enables *in*

situ fabrication of nanopores in an arrayed device with multiple, ultra-thin membranes (Yanagi et al. 2016). Finally, the method can prepare a fresh nanopore just before the DNA measurement, which allows long-term storage of the devices. Accordingly, due to its advantages, CBD has the potential allow a greater number of researchers to use solid-state nanopore technology.

Fabrication of Ultra-Thin Solid-State Membranes at the Wafer Scale

The membrane thickness determines the spatial resolution of a nanopore sensor. Because the distance between neighboring nucleotides in DNA is quite short (approximately 0.5 nm) at the molecular level, the fabrication of ultra-thin membranes is important for highly accurate discrimination of each nucleotide in DNA (Fig. 4a). One promising approach is to utilize two-dimensional materials such as graphene (Schneider et al. 2010), molybdenum disulfide (MoS_2) (Liu et al. 2014) and boron nitride (Liu et al. 2013). Although these atomically thin materials are quite attractive due to their atomic scale sensitivity (Heerema and Dekker 2016), stable mass production and control over their surface conditions are still challenges. Another approach is to thin membranes with semiconductor-related materials such as Si_3N_4 and hafnium oxide (HfO_2). Recently, HfO_2 membranes with 3–8-nm thicknesses have been fabricated using atomic layer deposition (Larkin et al. 2013). For Si_3N_4 membranes, thinning a membrane using reactive ion etching (Venta et al. 2013) or helium ion beam (Carlsen et al. 2014) etching has been demonstrated, and the thickness of the fabricated membrane is approximately 5 nm. In addition, a method for transferring a Si_3N_4 membrane to a quartz substrate has been proposed to fabricate 5-nm-thick Si_3N_4 membranes (Lee et al. 2014).

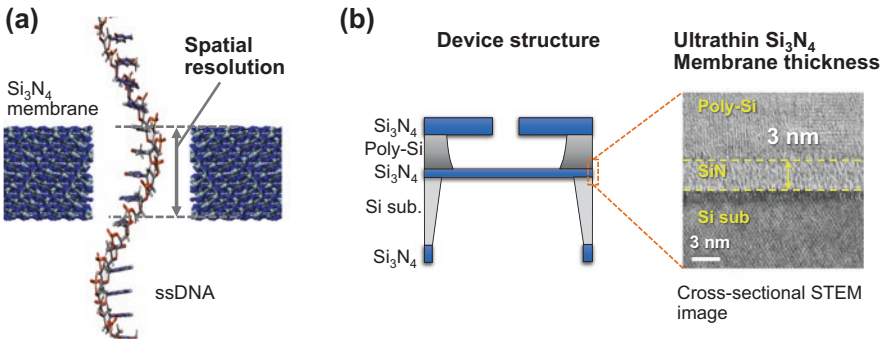

Fig. 4 Schematic image of a device with an ultra-thin insulating membrane for solid-state nanopores. (**a**) The thickness of the membrane with a nanopore determines the spatial resolution of the DNA sequencing. (**b**) Typical device structure with an ultra-thin Si_3N_4 membrane and a Si wafer. The right image shows a cross-sectional STEM image of a 3-nm-thick Si_3N_4 membrane

Si_3N_4 is a traditional semiconductor-related material that is highly compatible with the semiconductor process. Therefore, using Si_3N_4 as a membrane material for solid-state nanopores is highly desirable. Recently, wafer-scale fabrication of Si_3N_4 membranes with thicknesses of less than 5 nm can be successfully fabricated by employing a polycrystalline-Si (poly-Si) sacrificial layer as shown in Fig. 4b (Yanagi et al. 2015). Because this process significantly minimizes the damage to the membrane, Si_3N_4 membranes with thicknesses of 3 nm were stably fabricated with small variations. These devices with ultra-thin Si_3N_4 membranes enabled a detailed investigation of nanopore generation mechanisms via dielectric breakdown, especially the thickness dependence (Yanagi et al. 2017).

These technologies, including CBD, would allow the mass production of nanopore devices with an ultra-thin membrane at the wafer scale and contribute to realizing practical solid-state nanopore sensor applications.

Controlling the Speed of DNA Translocation Across a Nanopore

The next primal challenge for DNA sequencing with solid-state nanopores is controlling the translocation speed of DNA through a nanopore. When DNA passes through a nanopore via an electric field in an ionic solution, the typical dwell time of DNA in the nanopore is less than 1 μs/nt (Fig. 5a). This duration time is too short to detect the ionic current signal derived from each nucleotide using commercially available amplifiers (Wanunu et al. 2008; Venkatesan and Bashir 2011). Ideally, the dwell time of DNA in a nanopore should be more than 100 μs/nt to enable a sufficient recording of the signal from each nucleotide.

To reduce the DNA translocation speed through a nanopore, numerous strategies have been proposed (Carson and Wanunu 2015). For example, changing the electrolyte solutions has been tested as a simple method. The DNA translocation speed can

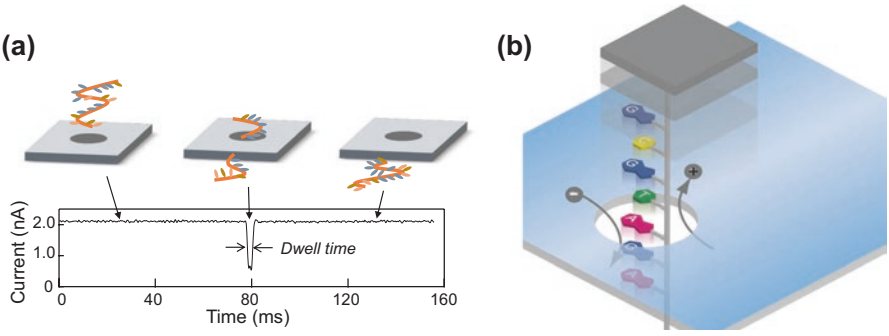

Fig. 5 (**a**) Typical example of an ultra-fast DNA translocation event for 5 k-mer ssDNA across a solid-state nanopore. This dwell time is less than 1 μs per nucleotide. (**b**) Schematic image of external control of DNA motion with a solid-state nanopore

be reduced by adding ethylene glycol to the ionic solution, and the resulting speed is reduced by as much as sixfold compared to that observed without ethylene glycol (Fologea et al. 2005). The DNA translocation speed in a LiCl aqueous solution has been reported to be approximately tenfold slower than the speed observed using a KCl aqueous solution (Kowalczyk et al. 2010). Several researchers have used a strategy in which the membrane is coated with various obstacles to decelerate the DNA translocation, such as a nanofibre mesh (Squires et al. 2013), polyethylene oxide (PEO)-filled nano-cylindrical domains (Yoshida et al. 2016), amine-functionalized beads (Goto et al. 2015), or a hydrophilic self-assembled monolayer (Wang et al. 2014). Using that approach, the dwell time of DNA in the nanopore can be increased to approximately 10–100 μs/nt without any external instruments.

As shown in Fig. 5b, other speed control approaches utilize a DNA-immobilized atomic force microscopy (AFM) probe or bead. Control of the translocation speed of the immobilized DNA through the nanopore is achieved by controlling the motion of the probe or bead using an actuator or optical potential (Keyser et al. 2006; Nelson et al. 2014). DNA immobilized on the probe can be inserted into and removed from the nanopore using a piezo-actuator (Akahori et al. 2017). In this method, the dwell time of DNA is greater than 100 μs/nt. However, at present, every reported control method for solid-state nanopores cannot prevent dwell time variability. This variation is larger than one order of magnitude and is thought to be caused by the interaction between DNA and the surface of the nanopore. For the realization of actual DNA sequencing, this variance must be resolved.

Toward Solid-State Nanopore DNA Sequencing with Single-Nucleotide Discrimination Based on External DNA Motion Control

Based on ionic current detection, several studies have reported the possibility of identifying each nucleotide. Three types of ssDNA homopolymers (poly(dA), poly(dC) and poly(dT)) can be identified using Si_3N_4 nanopores with a 5-nm thickness (Venta et al. 2013). Similarly, Si_3N_4 nanopores have been reported to discriminate three types of nucleotide homopolymers in block-copolymer DNA (poly(dT)-poly(dC)-poly(dA)) (Akahori et al. 2017). Molybdenum disulfide nanopores enable the discrimination of all four types of ssDNA homopolymers and monomers using room-temperature ionic liquids (Feng et al. 2015). Recently, we reported that a Si_3N_4 nanopore with a 5-nm thickness can also discriminate all four types of ssDNA homopolymers, even in an aqueous salt solution, by unfolding the G-quadruplex complex derived from a guanine homopolymer (Goto et al. 2018). These positive results indicate that solid-state nanopores have the potential to detect each nucleotide in a DNA strand similar to a biological one. Interestingly, the order of the signal magnitude with a solid-state strand (A > G > T > C) is different from that of a biological strand (MspA; T > C > G > A) (Derrington et al. 2010). Since the sensing mechanism for each nucleotide with a nanopore cannot be explained by

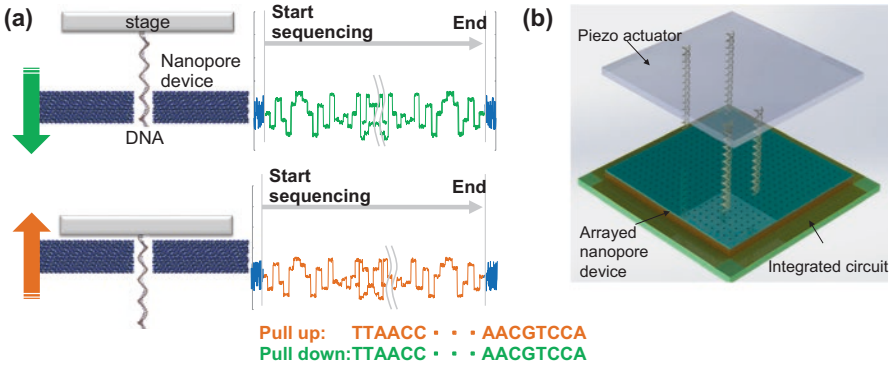

Fig. 6 (**a**) Schematic of a solid-state nanopore DNA sequencing system with a bidirectional DNA motion control unit. (**b**) Conceptual system drawing with a piezo-actuator, an arrayed ultra-thin nanopore and an integrated circuit (amplifier and A/D converter)

a simple size-exclusion effect, further investigation is required to reveal the blockade mechanism. However, the detection of signals at the single-nucleotide scale with a solid-state nanopore has not yet been demonstrated. Nanopore sequencing relies on the sensitivity of the ultra-thin membrane and the technique to precisely control DNA motion to acquire single-base resolution.

Recently, we developed a solid-state nanopore system with bidirectional DNA motion control for accurate, long-read DNA sequencing as shown in Fig. 6 (Goto et al. 2017). As mentioned above, precise semiconductor processes can fabricate high-quality solid-state nanopores with a low cost. Solid-state nanopore-based sequencing for reading DNA at a single-nucleotide sensitivity by detecting the blockage of ionic currents, however, is still under development. Our technology enables the fabrication of a precisely controlled ultra-thin Si_3N_4 nanopore at the wafer scale (Yanagi et al. 2015). Nanopores with diameters of a few nanometers, thicknesses of several nanometers, and accuracies at the sub-nanometer level can be electrically produced using a simple *in situ* process (Yanagi et al. 2014). Our system consists of a nanopore unit and a bidirectional DNA motion control unit that can provide single-nucleotide sensitivity and reversible DNA translocation.

We confirmed the proof-of-concept data that demonstrated single-nucleotide discrimination in a large tandem repeat DNA molecule. We observe sequence-dependent blockage current levels that corresponded to di-, tri- and tetranucleotide repeats. This result clearly indicated that our ultra-thin Si_3N_4 nanopore can read ssDNA with a single-nucleotide sensitivity. We estimated a resolution of 4 nucleotides for the nanopore, which contributed to the fine reading accuracy. The bidirectional DNA translocation was driven by a piezo-actuator with nanometer-scale accuracy. The ionic blockade signature followed the translocation speed of the piezo-actuator at an average speed of up to 100 bases per second. The bidirectional DNA measurement allows a highly accurate sequence to be obtained from multiple passes of a single DNA molecule. We believe that these encouraging proof-of-concept data contribute to the auspicious future of solid-state nanopore DNA sequencers.

Future Prospects

Solid-state nanopore DNA sequencing has the potential to become a label-free, rapid, long-read and low-cost sequencing technology. Solid-state nanopores provide several advantages, including mass production and large-scale integration. However, significant challenges remain to be resolved. A key limitation of high-accuracy DNA sequencing is the requirement of ultra-sensitive DNA detection with ultra-precise DNA motion control. Therefore, a single-base recognition sensitivity and precise control of the DNA speed are still the principal issues. However, solid-state nanopore technology can significantly impact the DNA sequencing field and the future of molecular biology and precision medicine.

Acknowledgments The authors would like to express the utmost thanks to all co-workers for their dedication to Hitachi's solid-state nanopore DNA sequencer project.

References

Akahori R, Haga T, Hatano T, Yanagi I, Ohura T, Hamamura H, Iwasaki T, Yokoi T, Anazawa T. Slowing single-stranded DNA translocation through a solid-state nanopore by decreasing the nanopore diameter. Nanotechnology. 2014;25(27):275501.

Akahori R, Yanagi I, Goto Y, Harada K, Yokoi T, Takeda K. Discrimination of three types of homopolymers in single-stranded DNA with solid-state nanopores through external control of the DNA motion. Sci Rep. 2017;7:9073.

Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26(10):1146–53.

Briggs K, Kwok H, Tabard-Cossa V. Automated fabrication of 2-nm solid-state nanopores for nucleic acid analysis. Small. 2014;10(10):2077–86.

Briggs K, Charron M, Kwok H, Le T, Chahal S, Bustamante J, Waugh M, Tabard-Cossa V. Kinetics of nanopore fabrication during controlled breakdown of dielectric membranes in solution. Nanotechnology. 2015;26(8):084004.

Carlsen AT, Zahid OK, Ruzicka J, Taylor EW, Hall AR. Interpreting the conductance blockades of DNA translocations through solid-state nanopores. ACS Nano. 2014;8(5):4754–60.

Carson S, Wanunu M. Challenges in DNA motion control and sequence readout using nanopore devices. Nanotechnology. 2015;26(7):074004.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4):265–70.

Dekker C. Solid-state nanopores. Nat Nanotechnol. 2007;2(4):209–15.

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH. Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A. 2010;107(37):16060–5.

Edel JB, Albrecht T. Engineered nanopores for bioanalytical applications. Amsterdam: Elsevier Science; 2013.

Feng J, Liu K, Bulushev RD, Khlybov S, Dumcenco D, Kis A, Radenovic A. Identification of single nucleotides in MoS_2 nanopores. Nat Nanotechnol. 2015;10(12):1070–6.

Fologea D, Uplinger J, Thomas B, McNabb DS, Li J. Slowing DNA translocation in a solid-state nanopore. Nano Lett. 2005;5(9):1734–7.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.

Goto Y, Haga T, Yanagi I, Yokoi T, Takeda K. Deceleration of single-stranded DNA passing through a nanopore using a nanometre-sized bead structure. Sci Rep. 2015;5:16640.
Goto Y, Yanagi I, Matsui K, Yokoi T, Takeda K. Integrated solid-state nanopore platform for nanopore fabrication via dielectric breakdown, DNA-speed deceleration and noise reduction. Sci Rep. 2016;6:31324.
Goto Y, Akahori R, Matsui K, Yanagawa Y, Aoki M, Yanagi I, Nara Y, Yoshida M, Yokoi T, Takeda K. Solid-state nanopore DNA sequencing: single-nucleotide discrimination and bidirectional DNA translocation. In: Advances in genome biology and technology (AGBT) The General Meeting. Hollywood: The Diplomat Beach Resort; 13–16 February, 2017.
Goto Y, Yanagi I, Matsui K, Yokoi T, Takeda K. Identification of four single-stranded DNA homopolymers with a solid-state nanopore in alkaline CsCl solution. Nanoscale. 2018;10(44):20844–50.
Heerema SJ, Dekker C. Graphene nanodevices for DNA sequencing. Nat Nanotechnol. 2016;11(2):127–36.
Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotechnol. 2001;19(7):636–9.
Iqbal SM, Bashir R. Nanopores: sensing and fundamental biological interactions. Heidelberg: Springer; 2011.
Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
Keyser UF, Koeleman BN, van Dorp S, Krapf D, Smeets RMM, Lemay SG, Dekker NH, Dekker C. Direct force measurements on DNA in a solid-state nanopore. Nat Phys. 2006;2(7):473–7.
Kowalczyk SW, Tuijtel MW, Donkers SP, Dekker C. Unraveling single-stranded DNA in a solid-state nanopore. Nano Lett. 2010;10(4):1414–20.
Kwok H, Briggs K, Tabard-Cossa V. Nanopore fabrication by controlled breakdown. PLoS One. 2014;9(3):e92880.
Larkin J, Henley R, Bell DC, Cohen-Karni T, Rosenstein JK, Wanunu M. Slow DNA transport through nanopores in hafnium oxide membranes. ACS Nano. 2013;7(11):10121–8.
Laszlo AH, Derrington IM, Ross BC, et al. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol. 2014;32(8):829–33.
Lee M-H, Kumar A, Park K-B, Cho S-Y, Kim H-M, Lim M-C, Kim Y-R, Kim K-B. A low-noise solid-state nanopore platform based on a highly insulating substrate. Sci Rep. 2014;4:7448.
Lindsay S. The promises and challenges of solid-state sequencing. Nat Nanotechnol. 2016;11(2):109–11.
Liu S, Lu B, Zhao Q, et al. Boron nitride nanopores: highly sensitive DNA single-molecule detectors. Adv Mater. 2013;25(33):4549–54.
Liu K, Feng J, Kis A, Radenovic A. Atomically thin molybdenum disulfide nanopores with high sensitivity for DNA translocation. ACS Nano. 2014;8(3):2504–11.
Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol. 2012;30(4):349–53.
Nelson EM, Li H, Timp G. Direct, concurrent measurements of the forces and currents affecting DNA in a nanopore with comparable topography. ACS Nano. 2014;8(6):5484–93.
Pennisi E. Search for pore-fection. Science. 2012;336(6081):534–7.
Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW, Vandersypen LM, Dekker C. DNA translocation through graphene nanopores. Nano Lett. 2010;10(8):3163–7.
Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345–53.
Sigalov G, Comer J, Timp G, Aksimentiev A. Detection of DNA sequences using an alternating electric field in a nanopore capacitor. Nano Lett. 2008;8(1):56–63.
Squires AH, Hersey JS, Grinstaff MW, Meller A. A nanopore-nanofiber mesh biosensor to control DNA translocation. J Am Chem Soc. 2013;135(44):16304–7.

Storm AJ, Chen JH, Ling XS, Zandbergen HW, Dekker C. Fabrication of solid-state nanopores with single-nanometre precision. Nat Mater. 2003;2(8):537–40.
Tsutsui M, Taniguchi M, Yokota K, Kawai T. Identifying single nucleotides by tunnelling current. Nat Nanotechnol. 2010;5(4):286–90.
Venkatesan BM, Bashir R. Nanopore sensors for nucleic acid analysis. Nat Nanotechnol. 2011;6(10):615–24.
Venta K, Shemer G, Puster M, Rodríguez-Manzo JA, Balan A, Rosenstein JK, Shepard K, Drndić M. Differentiation of short, single-stranded DNA homopolymers in solid-state nanopores. ACS Nano. 2013;7(5):4629–36.
Wang D, Harrer S, Luan B, Stolovitzky G, Peng H, Afzali-Ardakani A. Regulating the transport of DNA through biofriendly nanochannels in a thin solid membrane. Sci Rep. 2014;4:3985.
Wanunu M. Nanopores: a journey towards DNA sequencing. Phys Life Rev. 2012;9(2):125–58.
Wanunu M, Sutin J, McNally B, Chow A, Meller A. DNA translocation governed by interactions with solid-state nanopores. Biophys J. 2008;95(10):4716–25.
Yanagi I, Akahori R, Hatano T, Takeda K. Fabricating nanopores with diameters of sub-1 nm to 3 nm using multilevel pulse-voltage injection. Sci Rep. 2014;4:5000.
Yanagi I, Ishida T, Fujisaki K, Takeda K. Fabrication of 3-nm-thick Si_3N_4 membranes for solid-state nanopores using the poly-Si sacrificial layer process. Sci Rep. 2015;5:14656.
Yanagi I, Akahori R, Aoki M, Harada K, Takeda K. Multichannel detection of ionic currents through two nanopores fabricated on integrated Si_3N_4 membranes. Lab Chip. 2016;16(17):3340–50.
Yanagi I, Fujisaki K, Hamamura H, Takeda K. Thickness-dependent dielectric breakdown and nanopore creation on sub-10-nm-thick SiN membranes in solution. J Appl Phys. 2017;121(4):045301.
Yoshida H, Goto Y, Akahori R, Tada Y, Terada S, Komura M, Iyoda T. Slowing the translocation of single-stranded DNA by using nano-cylindrical passage self-assembled by amphiphilic block copolymers. Nanoscale. 2016;8(43):18270–6.

On-Site MinION Sequencing

Lucky R. Runtuwene, Josef S. B. Tuda, Arthur E. Mongan, and Yutaka Suzuki

Abstract DNA sequencing has reached an unprecedented level with the advent of Oxford Nanopore Technologies' MinION. The low equipment investment, ease of library preparation, small size, and powered only by a laptop computer enable the portability for on-site sequencing. MinION has had its role in clinical, biosecurity, and environmental fields. Here, we describe the many facets of on-site sequencing with MinION. First, we will present some field works using MinION. We will discuss the requirements for targeted or whole genome sequencing and the challenges faced by each technique. We will also elaborate the bioinformatics procedures available for data analysis in the field. MinION has greatly changed the way we do sequencing by bringing the sequencer closer to the biodiversity. Although numerous limitations exist for MinION to be truly portable, improvements of procedures and equipment will enhance MinION's role in the field.

The arrival of MinION, a portable long read DNA sequencer, has changed many aspects of DNA sequencing. The low overhead cost, ease of library preparation, small size, and use of a USB port of a laptop computer as its power supply have placed DNA sequencing at the frontier of research in the field. Researchers now have the freedom to perform DNA sequencing in the field where they collect samples. MinION has proven to be very beneficial in the clinical, biosecurity, and environmental fields (Hardegen et al. 2018).

Since its availability on the market through the MinION Access Program (MAP) in 2015, MinION has significantly evolved. From the first iteration that produced only up to 1 GB of data, the most recent flow cell is able to produce up to10–20 GB of data. The accuracy of MinION has improved dramatically from 65 to 88% (Lu et al. 2016) to greater than 90% (Oxford Nanopore Technologies

L. R. Runtuwene (✉) · Y. Suzuki (✉)
Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan
e-mail: luckyruntuwene@edu.k.u-tokyo.ac.jp; ysuzuki@hgc.jp

J. S. B. Tuda · A. E. Mongan
Faculty of Medicine, Sam Ratulangi University, Manado, Indonesia

Y. Suzuki (ed.), *Single Molecule and Single Cell Sequencing*, Advances in Experimental Medicine and Biology 1129,
https://doi.org/10.1007/978-981-13-6037-4_10

2017). Many options are available in proprietary kits, ranging from a PCR-free quick preparation to high-accuracy PCR-based kits. Special purposes kits are also available, e.g., barcoding, low input, and ultralow input. Furthermore, many aspects of DNA sequencing have also been simplified with the prospective release of equipment, such as Zumbador (simplifying DNA extraction and library preparation), Voltrax (simplifying library preparation), MinIT (a portable CPU to substitute the whole laptop), and even further miniaturization of MinION itself through sequencing on a smartphone.

The MinION principle relies on the recognition of an electrical current disturbance in a nanopore protein. When a strand of DNA is inside the nanopore, the nucleotides induce a change in the basal electrical current. This electric current disturbance pattern is specific to kmers and is recognized by sensors and translated into a sequence of nucleotides by an algorithm. The exclusion of lasers and fluorescence, which are commonly used in traditional and high-throughput DNA sequencers, permits the sequencing technology to be portable and energy efficient. Base-calling can be performed on-the-go or after sequencing. An online connection is also optional; thus, sequencing can occur in seclusion.

One of the most powerful aspects of MinION is the ability to sample biodiversity that is not easily sampled or requires an unmanageable amount of time. Many of the planet's biological resources are locked in the countries inaccessible to DNA sequencers. Consequently, instead of taking the sample to a laboratory, a group of researchers decided to take DNA sequencing to a remote rainforest of Tanzania (Menegon et al. 2017). These researchers successfully identified a frog species based on its DNA barcode, i.e., the small subunit of ribosomal RNA (16s rRNA). Despite the high error rate of the first generation of flow cell, a generation of consensus sequence exhibited 97% similarity to the reference genome of the frog. Another group of researchers conducted on-site sequencing in the Ecuadorian Choco rainforest to sequence the DNA barcodes of a toad and three species of snakes (Pomerantz et al. 2018), two of which were previously undescribed. These researchers appreciated how mobile laboratories and portable sequencing can accelerate species identification in remote areas. Extreme conditions also complicate the transport of samples to a laboratory, e.g., in the Canadian high Arctic. In this setting, MinION has been utilized to identify microorganisms through 16s rRNA sequencing. These researchers were able to identify bacteria, archaea, and eukaryotes in the permafrost layer (Goordial et al. 2017). Despite the high error rate, the results were still comparable to Illumina sequencing. The region resembles the planet Mars and acts as a test area should the need to identify Martian microorganisms arise. Similarly, DNA sequencing has been performed in space at the International Space Station (ISS) (Castro-Wallace et al. 2017). Here, three types of DNA from *E. coli* K-12, bacteria phage, and BALB/C mice were sequenced over four MinION runs, and the results did not differ between the space and Earth, suggesting MinION makes sequencing in extreme environments possible.

The advantage of MinION's low overhead cost is welcomed by a group of researchers working with metagenomics in waste sludge (Hardegen et al. 2018). Sequencing is rarely used for this purpose partially due to the required investment.

Here, MinION has aided in the analysis of real-time changes in microorganisms in the biomass industry, increasing the efficiency of the production process. MinION has been implemented in food security by identifying begomoviruses causing cassava mosaic disease (Boykin et al. 2018), further protecting the loss of productivity for 800 million people worldwide. The researchers performed the experiments in Tanzania, Uganda and Kenya, and all of these location lack sequencing facilities. The benefit of real-time analysis allowed the identification of begomoviruses as early as 11 s after sequencing. MinION has been applied for genomic surveillance, e.g., yellow fever outbreak in Brazil. By sequencing 52 whole genome sequences of yellow fever virus (YFV), the researchers identified an early case of sylvatic transmission. Seventy percent of the new sequences were generated using MinION (N. Faria et al. 2018), further demonstrating MinION's ability to complement and potentially replace established techniques.

When working on-site to identify the causative agent of a disease, researchers are puzzled by the available methods to enrich pathogenic nucleic acids. Several methods might be employed in this setting. Direct metagenomics sequencing is preferable when the causative agent is not known. However, this method has a direct correlation with the concentration of pathogens in the sample (Houldcroft et al. 2017). A project in Brazil to improve the molecular surveillance and sequencing of Zika virus (ZIKV) reported that direct metagenomics is difficult to perform with MinION when the viral concentration in the sample is low (N. R. Faria et al. 2016). The sequencing projects were performed in a mobile laboratory across five federal states in Northern region of Brazil. A total of 1349 samples were tested in 16 days during this field trip. Due to the low virus concentration, the genomes generated were fragmented and exhibit less than 50% coverage. A low sample concentration must undergo repeated sequencing using several flow cells to reach the adequate depth, which is not feasible in the field where resources are limited. To overcome this problem, a PCR-amplified tiling sequencing has been developed specifically for ZIKV (Quick et al. 2017). This technique requires segmental amplification using multiple sets of primers to amplify the entire genome of a virus. This methodology can be applied to samples containing as few as 50 genome copies per reaction, and the result are available 1–2 days following clinical sample acquisition. This technique can also be applied in other virus genomes. For example, the same technique was employed in the Ebola outbreak in Africa on 2016. MinION was used to sequence the entire genome of 146 EBOV samples (Quick et al. 2016). Using reverse transcription and tiling amplification with PCR, the fastest confirmation of EBOV infection was attainable within 1 h. Further, by combining the results obtained in the field with a database, the rate of EBOV mutation is determined. Again, successful detection depends on the viral titer. While this technique is feasible for small-genome viruses, sequencing viruses with 20–50-kB genomes using this technique is impractical due to the number of primers needed. The third technique commonly used to enrich viral genomes is oligonucleotide hybridization. Specific or degenerate probes bound to a solid phase (e.g., streptavidin bead) bind to the virus nucleic acids and are pulled down. The captured virus is subsequently adapter ligated and low-cycle amplified prior to sequencing. However, this technique

Fig. 1 Portable PCR

has not been tested in an on-site sequencing context; this technique most consistently yields full genome results in the laboratory.

When full-genome sequencing of a pathogen is not likely, targeted sequencing is the method of choice. Parts of the genome can be amplified by multiple methods and subsequently sequenced. The use of this technique offers the advantage of cost-effectiveness by mixing amplicons together in one sequence run. The mixture can be either multiple target genes of a sample or multiple samples of the same gene. Due to the high output of the most recent flow cell, a researcher can sequence up to 96 amplicons and decrease the cost of a sequencing run to $5.2–9.4 per amplicon. Multiplexing can be achieved by using the proprietary kit (only available in 1D rapid sequencing mode) or the self-produced barcodes attached to the forward primer of target gene (for compatibility with 1D^2 sequencing mode). As a proof-of-concept, we have performed a multiplex sequencing of amplicons of genes related to drug resistance in *Plasmodium falciparum* in the field (Runtuwene et al. 2018). DNA was extracted using spin-columns, and the 1D^2 sequencing protocol for genomic DNA was employed. Amplification was conducted using the mobile thermal cycler (Fig. 1), which only requires a USB connection to a laptop computer. The flow cell contained a mixture of 11 amplicons of nine genes, and the sequencing was run for 48 h. Consensus sequence was created for each amplicon per sample, yielding an 84.56% accuracy for sequencing. This seemingly low accuracy despite the use of flow cell R9.4 (i.e., the recent version) is due to the abundance of AT-rich and homopolymer tracts in the parasite genome. Nonetheless, we determined the parasites' drug-resistance status in Manado (Indonesia) (Fig. 2 presents the remote laboratory where we performed the sequencing), Thailand, and Vietnam from blood samples that were either frozen or preserved in an FTA card.

An interesting application for the use of MinION to assist in an outbreak is the development of a system to detect all of the viruses causing hemorrhagic diseases (Brinkmann et al. 2017). This technique involves targeted sequencing using two pools of 285 and 256 primer pairs for the identification of 46 virus species. Target

Fig. 2 MinION in the remote laboratory

genes are amplified with PCR. Using clinical specimens, the panel enables characterization of the causative agent within 10 min of sequencing, and a definitive diagnosis can be procured in less than 3.5 h.

Currently, the required starting DNA concentration is approximately 1 μg for 1D^2 and 10–100 ng for 1D sequencing. Clinical samples rarely yield high concentrations for pathogen sequencing. In the event of the unavailability of a thermal cycler, PCR can be substituted with isothermal amplifications to enrich the input DNA. Many techniques ranging from targeted [e.g., loom-amplified isothermal amplification (LAMP)] to whole genome isothermal amplifications [e.g., multiple displacement amplification (MDA)] are possible. LAMP relies on a denaturing-DNA-polymerase, such as Bst polymerase, to amplify a genome segment flanked by two to three pairs of primers (Notomi et al. 2000). The denaturing nature of the polymerase (i.e., simultaneously denatures the double strand of DNA and elongates the primers) allows the LAMP reaction to be run in isothermal conditions, thus completely eliminating the need for a thermal cycler. Further, LAMP reagents can be dried to assure their stability upon transportation to the field (Hayashida et al. 2015). Applying the LAMP technique as the amplification method, we performed a genomic epidemiology study of dengue virus in Indonesia and Vietnam (Yamagishi et al. 2017). Up to 141 and 80 DENV-positive samples were amplified isothermally and sequenced with MinION. We were able to determine the infecting virus serotype and reported a successful detection rate of 79%. Serotype can also be determined despite the 74–80% identity. We also developed LAMP combined with MinION sequencing to detect and differentiate among five species of malaria parasites by sequencing 18 s rRNA genes.

One of the main bottlenecks of MinION's analysis (and all next-generation sequencing) is the requirement of a bioinfomatician to handle the magnitude of data generated. At least a basic knowledge in computer science is necessary to start navigating the information encoded in the output files. MinION's raw data are provided in an incomprehensible binary language that first must be converted to human

readable files. Conversion to such files called FASTQ files is performed through ONT's proprietary Metrichor, which has been replaced with Nanonet, Albacore, Guppy, and Scrappie, which are also from ONT. These software programs identify DNA sequences directly from raw data and subsequently enhance accuracy. The output FASTQ file is a set of four lines of readable characters that contains a read identifier, the read sequences, a plus sign, and its quality. At this point, data analysis depends on the computer skill of the researcher. Realizing that not all researchers are adept in computer science, ONT has released their suite software to assist with analysis using a graphical user interface. This program is accessible through EPI2MEAgent software and is currently a paid service to the members of the MinION community. *What's in My Pot* (WIMP) is an example of one of these software programs that takes the FASTQ file as input and maps sequence fragments to a database to provide a set of possible answers to the question: "What organisms are likely be sequenced?" A more specific approach is to map the 16s rRNA amplicon sequences to the 16 s rRNA database through *16s* workflow to know the possible bacteria genus (sometimes species) in the sequenced sample. These software programs are helpful for the crude identification of an organism, which is one of the strongest advantages of applying MinION in the workflow. However, the downside is that an internet connection is required to execute these cloud-based software programs. A more experienced researcher in bioinformatics will typically use offline software in a Linux environment.

Acclimatization to a Linux environment represents the other half of the on-site sequencing pipeline, especially in the field where it is difficult to access the internet. Linux provides many open-source software programs for biologists. Using these software programs, endless possibilities for MinION analysis are available. A decent and powerful laptop computer is required for most of these programs. Mapping software programs (or mappers) are the Swiss army knife in on-site sequencing. These programs map the input FASTQ files to a set of reference genomes. Burrows-Wheeler Aligner (BWA) is probably the most well-known mapper as it has served as a staple in complementing Illumina sequencers since the dawn of the next-generation sequencing. It is designed to be compatible with short read sequencing; however, it currently has a MinION-mode to use with MinION's long reads and somewhat noisy sequencing. LAST is also another mapper that is comparable to BWA in terms of processing speed. Minimap is a mapper with superior speed. The second iteration (Minimap2) is threefold and tenfold faster than BWA-MEM for mapping >100 bp short reads and >10 kb long reads, respectively (Li 2018).

Although organism identification is sufficiently achieved using a mapper alone, different tasks require different software tools. For example, variant detection in real-time Ebola surveillance in Africa used Nanopolish. The software directly detects variants using the event-level ('squiggle') data generated by the MinION to evaluate candidate variants found in the aligned reads. However, this technique is not compatible with non-standard genomes (e.g., *P. falciparum*) given that Nanopolish tends to yield false positive results. Noisy MinION sequencing can gain benefits from consensus sequence generation. When we employed this technique to

detect the SNPs in genes conferring drug resistance in *P. falciparum*, it improved accuracy in the obsolete flow cell R7 and the most recent flow cell R9 to 73.46% and 84.56%, respectively (Runtuwene et al. 2018). Nevertheless, these processes were not straightforward and required knowledge in programming language.

In summary, on-site sequencing has propelled the advance of genome research. In this field, MinION has greatly reduced the equipment investment cost, simplified library preparation, and improved the accessibility to biodiversity. Although numerous limitations prevent the complete adoption of MinION as a true portable device, accessories currently being developed for MinION. They will simplify everything from DNA extraction to laptop computers and even the sequencer itself, causing these devices more portable and cheaper in the foreseeable future.

References

Boykin LM, Ghalab A, Rossitto De Marchi B, Savill A, Wainaina JM, Kinene T, et al. Real time portable genome sequencing for global food security. bioRxiv. 2018:1–10.

Brinkmann A, Ergu K, Radoni A, Tufan ZK, Domingo C, Nitsche A. Development and preliminary evaluation of a multiplexed amplification and next generation sequencing method for viral hemorrhagic fever diagnostics. PLoS Negl Trop Dis. 2017;11(11):1–13.

Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins H, Mcintyre ABR, et al. Nanopore DNA sequencing and genome assembly on the international space station. Sci Rep. 2017;7(18022):1–12. https://doi.org/10.1038/s41598-017-18364-0.

Faria NR, Sabino EC, Nunes MRT, Carlos L, Alcantara J, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016;8(97):1–4. https://doi.org/10.1186/s13073-016-0356-2.

Faria N, Kraemer N, Hill S, Goes de Jesus J, de Aguiar R, Iani F, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. bioRxiv. 2018:1–35.

Goordial J, Altshuler I, Hindson K, Chan-yam K. In situ field sequencing and life detection in remote (79°26'N) Canadian high Arctic permafrost ice wedge microbial communities. Front Microbiol. 2017;8(2594):1–14. https://doi.org/10.3389/fmicb.2017.02594.

Hardegen J, Latorre-Perez A, Vilanova C, Thomas G, Simeonov C, Porcar M. Liquid co-substrates repower sewage microbiomes. bioRxiv. 2018.

Hayashida K, Kajino K, Hachaambwa L, Namangala B, Sugimoto C. Direct blood dry LAMP: a rapid, stable, and easy diagnostic tool for human African trypanosomiasis. PLoS Negl Trop Dis. 2015;9(3):e0003578. https://doi.org/10.1371/journal.pntd.0003578.

Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nat Rev Microbiol. 2017;15(3):183–92. https://doi.org/10.1038/nrmicro.2016.182.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:1–7. https://doi.org/10.1093/bioinformatics/bty191.

Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. Genomics, Proteomics Bioinformatics. 2016;14(5):265–79. https://doi.org/10.1016/j.gpb.2016.05.004.

Menegon M, Cantaloni C, Rodriguez-prieto A, Centomo C, Abdelfattah A, Rossato M, et al. On site DNA barcoding by nanopore sequencing. PLoS One. 2017;12(10):1–18.

Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, Hase T. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res. 2000;28(12):E63. https://doi.org/10.1093/nar/28.12.e63.

Oxford Nanopore Technologies. New basecaller now performs "raw basecalling", for improved sequencing accuracy. 2017. Retrieved from https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy

Pomerantz A, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Salazar-valenzuela D, et al. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. GigaScience. 2018;7:1–14. https://doi.org/10.1093/gigascience/giy033.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228–32. https://doi.org/10.1038/nature16996.

Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261–76. https://doi.org/10.1038/nprot.2017.066.

Runtuwene LR, Tuda JSB, Mongan AE, Makalowski W, Frith MC, Imwong M, et al. Nanopore sequencing of drug- resistance-associated genes in malaria parasites, Plasmodium falciparum. Sci Rep. 2018;8(8286):1–13. https://doi.org/10.1038/s41598-018-26334-3.

Yamagishi J, Runtu LR, Hayashida K, Mongan AE, Thi AN, Thuy LN, et al. Serotyping dengue virus with isothermal amplification and a portable sequencer, (January), 1–10. 2017. https://doi.org/10.1038/s41598-017-03734-5.