Leading Edge
Primer

Cell

Design and Analysis of Single-Cell Sequencing Experiments

Dominic Grün[1,2,3] and Alexander van Oudenaarden[1,2,*]
[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), 3584 CT Utrecht, the Netherlands
[2]University Medical Center Utrecht, Cancer Genomics Netherlands, 3584 CX Utrecht, the Netherlands
[3]Max Planck Institute of Immunobiology and Epigenetics, D-79108 Freiburg, Germany
*Correspondence: a.vanoudenaarden@hubrecht.eu
http://dx.doi.org/10.1016/j.cell.2015.10.039

Recent advances in single-cell sequencing hold great potential for exploring biological systems with unprecedented resolution. Sequencing the genome of individual cells can reveal somatic mutations and allows the investigation of clonal dynamics. Single-cell transcriptome sequencing can elucidate the cell type composition of a sample. However, single-cell sequencing comes with major technical challenges and yields complex data output. In this Primer, we provide an overview of available methods and discuss experimental design and single-cell data analysis. We hope that these guidelines will enable a growing number of researchers to leverage the power of single-cell sequencing.

Introduction

Understanding the development and function of an organ requires the knowledge of its constituents, i.e., of all the different cell types the organ is composed of. It is still common practice to distinguish cell types based on a small set of marker genes. These can be used to isolate sub-populations of cells, e.g., by fluorescence-activated cell sorting (FACS), which can then be characterized by population-based assays such as next-generation sequencing. This approach is inherently constrained, since a pre-selection of marker genes limits the resolution and variability within a marker-gene-expressing sub-population of cells cannot be resolved. Moreover, even cells of the same type can show substantial gene expression variability leading to phenotypic variation (Eldar and Elowitz, 2010; Munsky et al., 2012; Snijder and Pelkmans, 2011). The ideal approach to profile the cell type composition of an organ or to explore transcriptome heterogeneity across cells of the same type is a separate analysis of individual cells randomly drawn from a sample. Single-cell analysis of a small number of genes can be performed with imaging-based methods such as single-molecule fluorescence in situ hybridization (Raj et al., 2008) or by flow cytometry, exploiting cell surface markers or fluorescent reporter proteins. Single-cell transcriptome analysis, on the other hand, is an experimental approach to obtain an unbiased view of all mRNAs present in a cell. Already by 1992 the expression of selected genes in individual neurons had been quantified by Southern blotting after amplifying the entire pool of mRNAs from a cell (Eberwine et al., 1992). Single-cell transcriptome sequencing had initially been applied by the Surani laboratory in 2009 (Tang et al., 2009). Over the last five years, a number of single-cell mRNA-sequencing methods with improved sensitivity and reduced technical noise have been introduced (Hashimshony et al., 2012; Islam et al., 2011, 2014; Picelli et al., 2013; Ramsköld et al., 2012; Sasagawa et al., 2013). These methods have been used to discriminate cell types in healthy tissues (Jaitin et al., 2014; Zeisel et al., 2015), to study differentiation dynamics (Treutlein et al., 2014), to discover rare cell types (Grün et al., 2015), to investigate the transcriptome response upon external signals (Shalek et al., 2013, 2014), or to profile tumor heterogeneity (Patel et al., 2014).

The genotypic variation that underlies cell-to-cell differences can be explored by single-cell genomics. In a landmark study, sequencing of the genomic DNA from single-tumor cell nuclei was employed to profile chromosome copy numbers in order to elucidate clonal expansion and tumor evolution (Navin et al., 2011). Subsequently, a number of improved methods have been published permitting the detection of genomic copy number variations and other structural rearrangements with increasing spatial resolution (Falconer et al., 2012; Gole et al., 2013; Wang et al., 2012; Zong et al., 2012).

In this Primer, we give an overview of the available techniques for genome and transcriptome sequencing, discuss the specific aspects and limitations of each method, and propose guidelines for designing single-cell sequencing experiments. Since any single-cell sequencing technique is based on amplification of minute amounts of material leading to substantial technical noise (Brennecke et al., 2013; Grün et al., 2014), data processing and analysis require extra care. We will discuss in depth all necessary steps for data acquisition, filtering, and analysis, with a focus on single-cell transcriptomics.

Isolating Single Cells for Sequencing

To perform any kind of single-cell sequencing assay, individual cells first have to be isolated from the system of interest. The method of choice to purify thousands of single cells is FACS. With unrestricted sorting gates, random samples of cells can be purified. Alternatively, sorting gates can be set based on scatter properties reflecting the morphology and composition of a cell. Fluorescently labeled antibodies against cell surface markers provide another strategy to purify sub-groups of cells.

Current technology permits the simultaneous measurement of up to 20 parameters per cell and thus highly specific sub-groups of cells can be isolated by FACS (Chattopadhyay and Roederer, 2012). These can be sorted directly into 96- or 384-well plates amenable to subsequent single-cell sequencing. Importantly, the parameter information can be allocated to each well. However, flow cytometry requires a large starting volume, and sorting errors can lead to wells with cell doublets or empty wells.

Micromanipulation provides an alternative approach when only a few cells are available and visual inspection of a cell is desired prior to sequencing. Here, cells are aspirated with a glass micropipette under the microscope. However, this method is very laborious and not well suited to high-throughput single-cell analysis.

More recently, microfluidic devices became available that enable sorting single cells into individual compartments where cells can be visually monitored and further processed. This Fluidigm C1 autoprep system is particularly suited to single-cell sequencing (Islam et al., 2014; Pollen et al., 2014). A shortcoming of this method is the fixed chip architecture that limits the selection of cells to a certain size window. A more detailed discussion of single-cell isolation methods has been published recently (Saliba et al., 2014).

Comparison of Whole-Genome Amplification Techniques

Being able to sequence the genome of individual cells permits the investigation of many relevant questions. Over the lifetime of an organism, cells undergo multiple rounds of division. During each cell division, DNA replication errors can escape the DNA repair machinery with a small probability and can lead to so-called somatic mutations, which can give rise to cancer (Alexandrov and Stratton, 2014) and other diseases (Biesecker and Spinner, 2013). Moreover, a surprisingly high frequency of chromosomal abnormalities has been observed during mammalian germline development (Nagaoka et al., 2012). All types of germline and somatic genome mutations, comprising substitutions, insertions and deletions (indels), copy number variations (CNV) and structural rearrangements, can in principle be detected by DNA sequencing. Moreover, genetic inheritance can be studied by quantifying maternal and paternal allele frequencies based on single-nucleotide polymorphisms (SNPs). However, a single mammalian cell contains less than 10 pg of DNA, necessitating whole-genome amplification (WGA) prior to sequencing or microarray-based analysis. Currently available WGA principles are based on polymerase chain reaction (PCR), multiple displacement amplification (MDA), or a combination of the two. PCR-based strategies initiate amplification by either priming with random oligonucleotides (Cheung and Nelson, 1996; Zhang et al., 1992) or by universal adaptors that are ligated to DNA fragments after enzymatic digestion (Klein et al., 1999). MDA utilizes isothermal amplification by a DNA polymerase with strand displacement activity, typically ϕ29, initiated by random priming of denatured DNA (Dean et al., 2002). The polymerase possesses high processivity and generates DNA amplicons up to 10 kb in length. Upon contact between the 3′ end of an amplicon and the 5′ end of an adjoining amplification product during synthesis, the latter gets displaced, liberating the strand for further amplification. All available methods introduce technical artifacts originating from non-uniform genome coverage, in particular due to biased amplification of sequence rich in cytosine and guanosine (GC-bias), preferential allelic amplification or allelic dropout, base copy errors, and chimeric DNA molecules (Macaulay and Voet, 2014). Since the prevalence of a particular type of error depends on the method, the experimental technique should be selected based on the desired readout. In general, random primed PCR-based methods achieve a highly uniform amplification but yield only sparse coverage of the genome and are therefore well suited for low-resolution copy number variant detection down to a length scale of 60 kb (Möhlendick et al., 2013). Due to the high processivity in combination with the strand displacement activity, a much better genome coverage can be achieved with MDA. Together with the high fidelity of the ϕ29 polymerase, this method is better suited for SNP calling. On the other hand, MDA yields highly non-uniform amplification, and the observed biases are only partially explained by the GC bias. This implies the risk of false positives if MDA is used for CNV detection. Moreover, both PCR- and MDA-based techniques produce chimeric DNA molecules, introducing artifacts that can be interpreted as indels or structural rearrangements (Voet et al., 2013).

A technique for obtaining broad coverage of the genome together with uniform amplification was recently developed that combines pre-amplification by a polymerase with strand-displacement activity and amplification by PCR (Zong et al., 2012). The method, termed multiple annealing and looping-based amplification cycles (MALBAC), pre-amplifies DNA with a strand-displacement polymerase and generates amplicons with complementary ends. This complementarity induces loop formation and prevents the amplicon from being used as a template during subsequent cycles to attain close-to-linear amplification. After five cycles of pre-amplification, the material is amplified exponentially by PCR. Sequencing of MALBAC-amplified material from a single cell yielded 93% genome coverage at an average 25× sequencing depth. Due to the improved uniformity and a substantially lower allele dropout rate in comparison to MDA (~1% for MALBAC versus ~31%–65% for MDA [Leung et al., 2015; Zong et al., 2012]), MALBAC shows higher detection efficiency for SNPs and CNVs. The residual false-positive rate of MALBAC ($\sim 4 \times 10^{-5}$) is due to the relatively low fidelity of the polymerase and could be reduced by sequencing two or three daughter cells derived from the same mother cell. MALBAC is therefore well suited for the simultaneous characterization of SNPs and CNVs.

Another strategy to eliminate amplification biases and alleviate non-uniformity of genome coverage inherent to MDA is the reduction of the reaction volume, for instance, by using nano-liter reaction wells (Gole et al., 2013). This method, termed micro-well displacement amplification system (MIDAS), reduces reaction volume by ~1,000-fold in comparison to conventional MDA, thereby increasing the effective template concentration and reducing contamination. Traditional whole-genome amplification requires extensive purification in order to reduce environmental contamination. In another study, a nano-liter reaction volume was obtained by applying microfluidics to WGA, thereby minimizing amplification error and yielding an extremely low error rate of 4×10^{-9} (Wang et al., 2012). A more detailed comparison

of the available WGA methods has been presented elsewhere (Macaulay and Voet, 2014).

Following WGA, quantification can be performed either by DNA microarrays or by next-generation sequencing. Microarrays can resolve larger CNVs, down to less than 100 kb (Möhlendick et al., 2013), and SNP arrays have been used to infer genome-wide haplotypes from a single human cell with high accuracy (Fan et al., 2011). Moreover, family-based phasing approaches were successfully applied for haplotyping human embryos (Ottolini et al., 2015). Next-generation sequencing offers the advantage that every amplified base of the DNA is quantified with digital precision and thus enables detection of all types of anomalies, while the microarray readout is constrained by the probe library. Moreover, paired-end sequencing provides additional information since the mapped loci of the two ends together with the fragment size distribution can reveal structural rearrangements within the genome. Of note, sequencing the genome of a single cell with the Strand-seq protocol retains the strand information and allows the derivation of sister chromatid exchange (Falconer et al., 2012). This method provides valuable information for de novo genome assembly or the revision of existing assemblies.

Although substantial progress has been made toward attaining high coverage and uniformity of WGA, there is room for improvement of existing methods, as recently demonstrated by the development of MALBAC (Zong et al., 2012) or by scaling down the reaction volume in order to reduce amplification (Gole et al., 2013; Wang et al., 2012).

Analysis of Single-Cell Genome Sequencing Data

The first step in the data analysis after obtaining a file with sequencing reads is mapping to a reference genome. The genomic DNA sequence for most model organisms can be readily obtained from various online databases, such as the UCSC genome browser (Meyer et al., 2013) or www.ensembl.org (Cunningham et al., 2015). Prior to mapping, it is advisable to inspect the read quality and trim low-quality bases as well as remaining adaptor sequences at the end of the reads. However, if the remaining read length is too short, reads should be discarded in order to avoid erroneous mappings. Furthermore, it is recommended to remove PCR duplicates. After the mapping is performed, reads that map to more than a single locus should be discarded or counted with reduced uniform weight for each locus, such that the weights of each read add up to one. Subsequent processing depends on the type of analysis. To determine CNVs, local variability in read coverage can be alleviated by segmenting the genome into bins. After correcting the number of reads within each bin for GC bias CNV breakpoints can be determined based on a comparison of the change in read number between adjacent bins to a background model (Venkatraman and Olshen, 2007; Zhang et al., 2013). For instance, the circular binary segmentation algorithm (Venkatraman and Olshen, 2007) uses t-statistics with a permutation reference distribution to infer p values for breakpoints. Another study employed a hidden Markov model for CNV detection, with the hidden states corresponding to the local copy number (Zong et al., 2012). Abnormal copy numbers in a cancer cell were inferred after eliminating the amplification bias with a normalization factor derived from a non-cancer cell. The emission probabilities of this model correspond to binary vectors indicating whether the cancer cell had higher copy number than the normal cell. The numerous published methods for CNV detection using next-generation sequencing were discussed in a recent review (Zhao et al., 2013).

The genome analysis toolkit GATK comprises a bundle of methods for processing of next-generation sequencing data and variant calling (McKenna et al., 2010). In particular, it contains a Bayesian framework that can be used for SNP detection. For each locus, the genotype with the highest posterior probability is emitted if its log odds ratio exceeds a defined threshold. A comprehensive overview and a comparative analysis of existing software tools for SNP calling from next-generation sequencing data can be found in the literature (Nielsen et al., 2011). An advanced method for the detection of structural rearrangements utilizes paired-end read information by creating a bona fide list of discordantly mapped read pairs and identifies candidate rearrangements supported by more than one pair from this list (Voet et al., 2013).

Although correction of GC bias is possible (Baslan et al., 2012; Voet et al., 2013; Zhang et al., 2013), other confounding factors such as allelic dropout or preferential allelic amplification cannot be easily corrected for and may introduce false positives in SNP and CNV detection. Random sequencing errors represent another source of uncertainty for SNP detection. To increase confidence, repeated detection of a given anomaly in more than a single daughter of the same cell is required (Zong et al., 2012). Finally, another confounding factor can be the cell-cycle phase since replication domains of cells in S phase can be mistaken as genuine structural aberrations (Van der Aa et al., 2013). This problem can be avoided by using only nuclei in G1 or G2/M phase. Limiting the analysis to G2/M phase comes with the additional advantage of having duplicated material after replication of the entire genome (Wang et al., 2014).

Comparison of Single-Cell Transcriptome Sequencing Techniques

Measuring gene expression in populations of cells with microarrays or RNA sequencing masks the true distribution of gene expression levels across cells, and it is therefore crucial to quantify gene expression in individual cells. The major hurdle is to obtain sufficient material from an individual cell that can be sequenced with standard next-generation sequencing protocols. Different methods for the amplification of the sub-picogram amount of mRNA from a single cell have been developed and are discussed in detail below. The main problem with any of these methods is the presence of amplification bias, which can distort the relative abundances of mRNAs from different genes.

In the past, amplified RNA from single cells was quantified with microarrays (Iscove et al., 2002). More recently, a number of single-cell sequencing techniques with improved sensitivity were developed. The first protocol for single-cell sequencing was published in 2009 by the Surani laboratory (Tang et al., 2009) and was subsequently used to trace the derivation of mouse embryonic stem cells from the inner cell mass with single-cell resolution (Tang et al., 2010). The amplification method is based on pull-down and reverse transcription of polyadenylated RNA using a poly(T) primer with a specific anchor sequence. Thereafter, the

single-stranded cDNA is polyadenylated and second-strand synthesis is performed using a poly(T) primer with another anchor sequence. The double-stranded cDNA is then PCR amplified from primers against the two anchor sequences, and the resulting material is fragmented prior to library preparation. Although SOLiD sequencing was applied initially, the protocol is compatible with Illumina sequencing, which has become the prevalent method for single-cell sequencing. An initial method that leveraged the integration of DNA barcodes to allow pooling of the material extracted from different cells along with preservation of strand information was termed single-cell tagged reverse transcription (STRT) (Islam et al., 2011). This technique exploits the template-switching property of the reverse transcriptase to tag the 5′ end of polyadenylated mRNA molecules. Following PCR amplification, the tagged ends are pulled down and sequenced, yielding a strong 5′ end bias of the sequencing read. A complementary method termed cell expression by linear amplification and sequencing (CEL-seq) amplifies polyadenylated mRNA linearly from a T7 promoter introduced during cDNA synthesis, thereby reducing amplification bias and alleviating the need for a template switch. Here, only fragments derived from the 3′ end of the mRNA are sequenced. CEL-seq and STRT-seq integrate a barcode into the sequencing primer, a stretch of eight nucleotides that uniquely labels all mRNAs from the same cell. In order to robustly assign mRNAs to different cells, each pair of barcodes should differ in at least two positions. To obtain read coverage along the entire transcript, the Smart-seq and Smart-seq2 methods are a more recent alternative (Picelli et al., 2013; Ramsköld et al., 2012). Similar to STRT, this approach reverse transcribes polyadenylated RNA and exploits the template-switching capacity of the reverse transcriptase. However, using the Nextera technology, the Tn5 transposase simultaneously fragments the cDNA and ligates sequencing adaptors to all fragments, yielding sequencing reads derived from the entire transcript. Another more recent method that yields read coverage of the entire gene body is the Quartz-seq method, which is similar to the approach developed by the Surani laboratory (Tang et al., 2009) but achieves higher sensitivity and reproducibility (Sasagawa et al., 2013). Moreover, two whole-transcript sequencing methods for low starting material have been published, exploiting either Φ29 DNA polymerase or semi-random-primed PCR based amplification (Pan et al., 2013).

To reduce amplification bias, unique molecular identifiers (UMI) (Kivioja et al., 2012) have been integrated into some of the single-cell sequencing protocols. UMIs are stretches of four to ten random nucleotides integrated into a sequencing primer and serve as a random barcode for each mRNA molecule. Upon binding of the sequencing primer, each mRNA is uniquely labeled with a random barcode and the labeled end of the mRNA is amplified along with the barcode. After sequencing, the amplification bias can be eliminated by counting each label only once instead of the reads derived from all amplicons. The number of UMIs can thus be directly translated into the number of sequenced molecules from a cell after application of a mathematical correction to account for the effect of random counting statistics (Grün et al., 2014; Kivioja et al., 2012).

UMIs can only be used for methods that sequence a single tag derived from a given mRNA and have been integrated, for example, into the STRT protocol (Islam et al., 2014) and into modified versions of CEL-seq (Grün et al., 2014; Jaitin et al., 2014). It has been shown that counting UMIs instead of reads leads to a 2-fold reduction of technical noise (Grün et al., 2014).

An overview of three common single-cell sequencing methods is given in Figure 1. In order to select the appropriate sequencing technology, one has to consider the goal of the experimental study. For example, in order to investigate gene expression heterogeneity between cells, the technical variability should be minimized and a technology that allows integration of UMIs should be chosen. However, if information along the entire transcript is required, for instance, to examine splicing patterns, a technology that yields whole-transcript coverage should be selected. Moreover, methods that sequence either the 5′ or 3′ end of a transcript provide single-cell information on the transcriptional start site or polyadenylation site usage, respectively. Another aspect to consider is ease of the experimental procedure and sequencing cost per cell. An increasing number of protocols can be conveniently performed on the Fluidigm C1 multifluidic auto-prep system. This device permits the isolation and processing of the cells, with the important benefit that each cell is imaged. This allows controlling for multiple cells per well and empty wells. However, sequencing-chips that can be used in this device come in fixed geometries and preferentially select cells of particular sizes. Moreover, this technology is relatively expensive. A massively parallel RNA single-cell sequencing framework termed MARS-seq (Jaitin et al., 2014) has been developed based on the CEL-seq technology and employs automated processing of single cells sorted into 384-well plates.

Recently, two advanced droplet-based microfluidic methods, termed Drop-seq (Macosko et al., 2015) and inDrop sequencing (Klein et al., 2015), were published that can dramatically increase the throughput to thousands of cells and at the same time minimize the sequencing costs. Both of these methods rely on the separation of cells into nanoliter-sized aqueous droplets in an oil-water emulsion, which contains sequencing primers with unique cell barcodes and UMIs. The co-occurrence of multiple cells in the same droplet is avoided by a low cell-loading rate into the droplets. In Drop-seq cDNA is PCR amplified, while inDrop sequencing amplifies cDNA by in vitro transcription akin to CEL-seq. In terms of technical noise and sensitivity, these methods compare favorably to previous protocols. Drop-seq was used to characterize mouse retinal cells, while inDrop sequencing was applied to explore cellular heterogeneity during mouse embryonic stem cell differentiation. However, the set-up for neither of these methods is commercially available, and the user is required to build a microfluidic device based on the information provided by the authors.

Although there has been much progress in increasing throughput and lowering costs of single-cell sequencing, there has been only a moderate improvement of the sequencing sensitivity during the last 3 years. The most common method to quantify sensitivity is the usage of external spike-in RNA of known concentration. The spike-in concentration should be chosen such that spike-in RNA contributes 1%–5% of the number of mRNA molecules (Hashimshony et al., 2012). Most of the recently published sensitivity estimates are derived from a set of 92 spike-in RNAs designed by the External RNA Controls

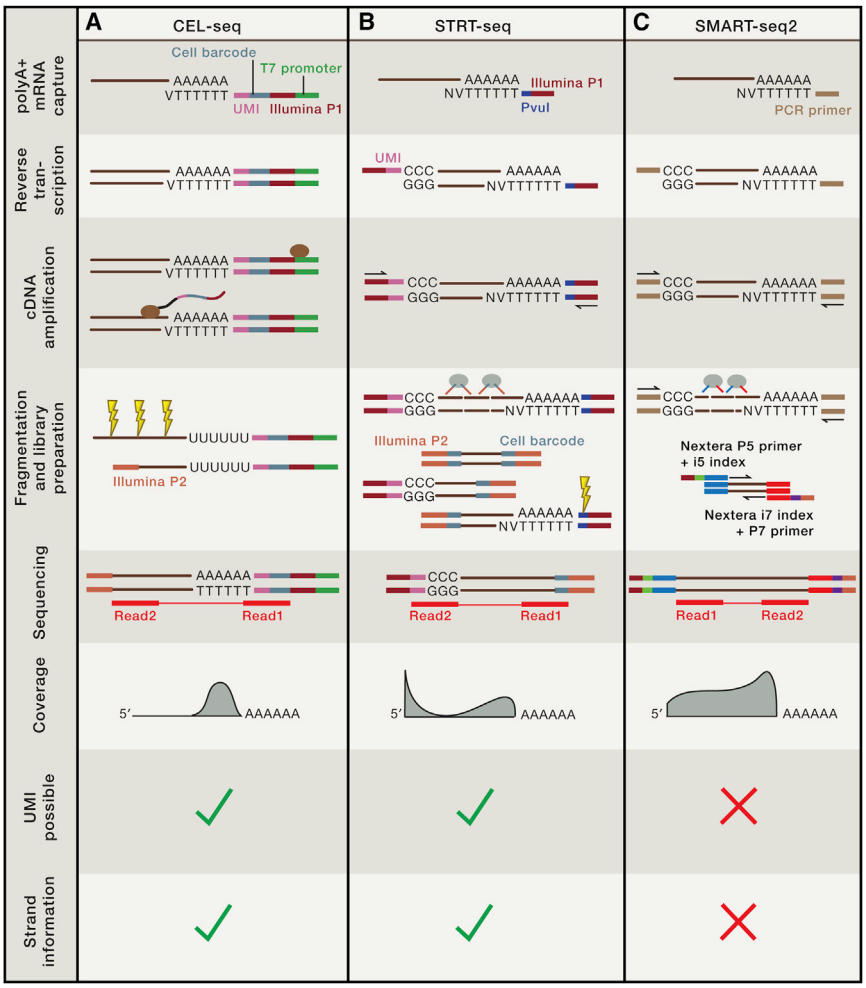

Figure 1. Three Common Experimental Protocols for Single-Cell Sequencing

(A) CEL-seq. Polyadenylated mRNA is reverse transcribed from an Oligo dT primer containing the Illumina P1 adaptor, a cell barcode, and a T7 promoter. The sequencing primer can, in principle, also accommodate a UMI. Following second-strand synthesis, the cDNA is amplified by in vitro transcription from the T7 promoter, and the Illumina P2 adaptor is ligated after fragmentation. The sequencing reads are thus derived from the mRNA 3′ end.

(B) STRT-seq. Polyadenylated RNA is reverse transcribed from an Oligo-dT pimer containing the Illumina P1 adaptor and a PvuI restriction site. After full-length reverse transcription, a template-switching oligo with another Illumina P1 adaptor and the UMI is added to the 5′ end of the transcript. Following second-strand synthesis, the cDNA is then PCR amplified using primers complementary to the Illumina P1 adaptor. Fragmentation and ligation of the Illumina P2 adaptor and the cell barcode are performed simultaneously utilizing the Tn5 transposase. To retain only 5′ ends for sequencing, the 3′ ends are digested by the PvuI restriction enzyme.

(C) Smart-seq2. Polyadenylated RNA is reverse transcribed from an Oligo dT with a PCR primer. The same PCR primer is part of the template-switching oligo added to the 5′ end of the cDNA upon reverse transcription. After PCR amplification, the cDNA is fragmented by tagmentation using the Tn5 transposase. Simultaneously, Tn5 ligates different 5′ and 3′ primers to the fragments. Another round of PCR introduces Nextera-sequencing primers to the ends of the fragments, enabling sequencing with full-length read coverage. However, Smart-seq2 does not allow for the integration of UMIs.

Consortium (ERCC) (Baker et al., 2005) and cover a wide range, from 5% to 40%. However, independent methods such as imaging-based molecule counting in single cells by single-molecule fluorescent in situ hybridization (smFISH) (Raj et al., 2008) yield deviating estimates (Grün et al., 2014). Moreover, the absolute number of transcripts per cell is comparable when sequencing cells of the same type with different methods. The ERCC spike-in RNAs are relatively short in comparison to mammalian genes, have short poly(A) tails, and lack a 5′ cap. It is unclear how much these differences between external and cellular RNA—as well as the fact that the external RNA is not spiked directly into the cell—affect the relative sequencing efficiencies of cellular and spike-in RNA.

Data Analysis of Single-Cell Transcriptome Data

Preprocessing and Read Mapping

In order to retrieve the maximum information from single-cell mRNA sequencing data, a careful experimental design is required (see Box 1). Following sequencing, a number of data processing and filtering steps are recommended to reduce the impact of technical noise. The first analysis step is usually a quality filtering or trimming of the sequencing reads prior to mapping the reads to a reference database. Standard tools, e.g., fastqc, permit a quality analysis of the sequenced library, and standard mapping tools, such as bwa (Li and Durbin, 2010), allow trimming of low-quality bases from the end of the reads. However, a minimum remaining read length (> 35 bp for mouse or human) after trimming should be required in order to avoid false-positive hits. For the mapping, available standard tools developed for bulk RNA-seq analysis can be used (Garber et al., 2011). However, sequenced cell barcodes, UMIs, and other primer-derived sequences have to be removed from the remaining read to be mapped to the reference database. Usually, one read of a pair contains all of the index information, while the other one can be mapped to the gene models (see Figure 1). In general, reads can be mapped to the genome followed by expression quantification via intersecting the read coverage of the genome with gene model annotations. However, this can lead to a larger number of reads mapping to multiple loci, for instance, due to the existence of inactive pseudogenes. Using the transcriptome as a reference reduces the sequence space and increases the fraction of unique reads. Since non-unique reads can introduce spurious correlations between different genes across single cells, it is advisable to discard these reads prior to analysis.

Box 1. Design of Single-Cell Sequencing Experiments

The power of single-cell sequencing crucially depends on two parameters: the number of cells and the sequencing complexity. These parameters can be controlled by the experimental design and should be chosen according to the goal of the study. The size of the dataset, i.e., the number of cells is important for profiling the cell composition of a sample with high sensitivity. Typically, several hundreds of cells have to be sequenced in order to capture not only abundant, but also rare, cell types. Possible biases that might occur during purification of the single-cell sample due to cell size or other factors have to be considered. Moreover, one should incorporate an estimate for the success rate, since a number of single-cell samples will likely yield only little or no material due to RNA degradation or low amplification efficiency. This estimate can be derived from trial experiments. The second parameter is the library complexity. Since the efficiency of single-cell mRNA sequencing is still limited, it is important to sequence each single cell with sufficient sequencing depth. If transcripts are counted with UMIs, the sequencing depth should be adjusted such that every transcript is sequenced at least three to four times. This ensures that even lowly expressed genes can be quantified and do not drop out due to sampling noise. To determine how many cells can be sequenced at once, e.g., on a single lane of an Illumina sequencing machine, the fraction of reads that can be mapped to the transcriptome has to be taken into account. This fraction is typically lower than 50%, since in most protocols additional abundant contributions can originate from sequencing products containing only primer or adaptor sequences (Grün et al., 2014). For example, assuming that ~10,000 transcripts per cell have been amplified and 50% of the reads can be mapped to the transcriptome, about 2,500 cells can be sequenced on a single lane of an Illumina NextSeq machine with 200 million reads. A fraction of those, typically around 10% to 20%, will not pass the quality filtering. Microfluidic devices like the Fluidigm C1 further provide an image of each cell being processed and allow filtering of wells containing no or more than a single cell.

To avoid batch effects, one should follow general guidelines applicable for bulk sequencing. For instance, single-cell libraries corresponding to different conditions should not be sequenced on separate lanes but, rather, distributed in equal fractions across the same set of lanes.

Due to the low read coverage of the gene body in single-cell sequencing experiments, isoform quantification with standard methods such as Cufflinks (Trapnell et al., 2010) can be problematic. If isoform information is not essential for the study, an ideal strategy is to merge all isoforms of a given gene into a so-called gene locus and quantify the expression of these gene loci. Independent of the reference, it is important to consider specific aspects of the experimental strategy. If sequencing protocols are used that enrich for the 5′ or 3′ end of an mRNA, the quality of the gene annotation can have a huge impact on the sensitivity. Gene models tend to be less reliable at both ends of a transcript, and an experimental strategy for improving 5′ or 3′ end annotation might be beneficial, in particular, for non-standard model organisms. For example, Junker et al. applied an amended CEL-seq protocol to sequence longer reads at low depths on bulk material in order to accurately detect 3′ polyadenylation sites for zebrafish embryos (Junker et al., 2014). Finally, the reference database has to be augmented by sequences representing any spike-in RNA added to the samples.

Expression Quantification and Filtering

In order to arrive at expression levels for all genes, PCR duplicates should be removed. Next, the cell of origin is determined based on the sequenced cell barcode (Figure 2A). If the base-calling quality is not sufficiently high at the cell barcode position within the read, an error-tolerant assignment scheme can be applied by aggregating all barcodes up to a single mismatch away from the perfect sequence. In order to apply this scheme, however, each pair of cell barcodes has to differ in at least two positions. If UMIs are available, the number of different UMIs per gene in each cell has to be converted into a transcript count estimate (Figure 2A) by applying a statistical correction to account for sampling effects (Grün et al., 2014; Kivioja et al., 2012).

Once read or transcript counts have been determined for all cells, it is recommended to filter out cells of low yield (Figure 2B). These samples can arise already prior to or during isolation of the cells, e.g., due to stress or apoptosis, or can occur due to incomplete lysis, RNA degradation, or low sequencing efficiency of a particular cell. The total number of reads or UMI-derived transcript counts per cell is a first proxy for the sample quality. Applying a threshold to discard cells in the lower tail of the distribution of read or transcript counts, respectively, safeguards against artifacts arising from low-quality cells. The expression of spike-in RNA can be utilized to identify and discard samples of low sequencing efficiency. Since the number of spike-in RNA should be identical for all samples, the identification of low yield samples is straightforward (Figure 2B). On the other hand, a relatively large ratio between transcript or read counts, respectively, of spike-in and cellular RNA reveals cells that contribute little material, e.g., due to RNA degradation or incomplete cell lysis (Figure 2B). The described strategies are only guidelines for filtering, and the exact method strongly depends on the dataset under examination. For example, if the cell volume varies substantially within a dataset, the total transcript count should only be subject to mild filtering, while the transcript count of the spike-in RNA is still a good proxy for the sequencing efficiency and can be used to discard low yield samples.

Data Normalization

For subsequent analysis, an appropriate normalization of the expression data is necessary. In the case of read-based quantification, normalization to transcripts per one million reads (TPM)—if reads are only generated from one end of the transcript—or transcript per one million reads per kilobase of transcript (RPKM)—if reads cover the entire transcript—is appropriate. Alternatively, standard quantification methods like Cufflinks (Trapnell et al., 2010) yield normalized expression values. More refined normalization schemes have been developed for bulk RNA-seq data (Anders and Huber, 2010). Here, derivation of a size factor for each replicate accounts for variability in sequencing depth between replicates, and a similar method can be applied to normalize single-cell data (Brennecke et al., 2013). If transcripts are counted with UMIs, cell-to-cell differences in transcript numbers are to a certain extent biologically meaningful and indicative of variations in the RNA content of a

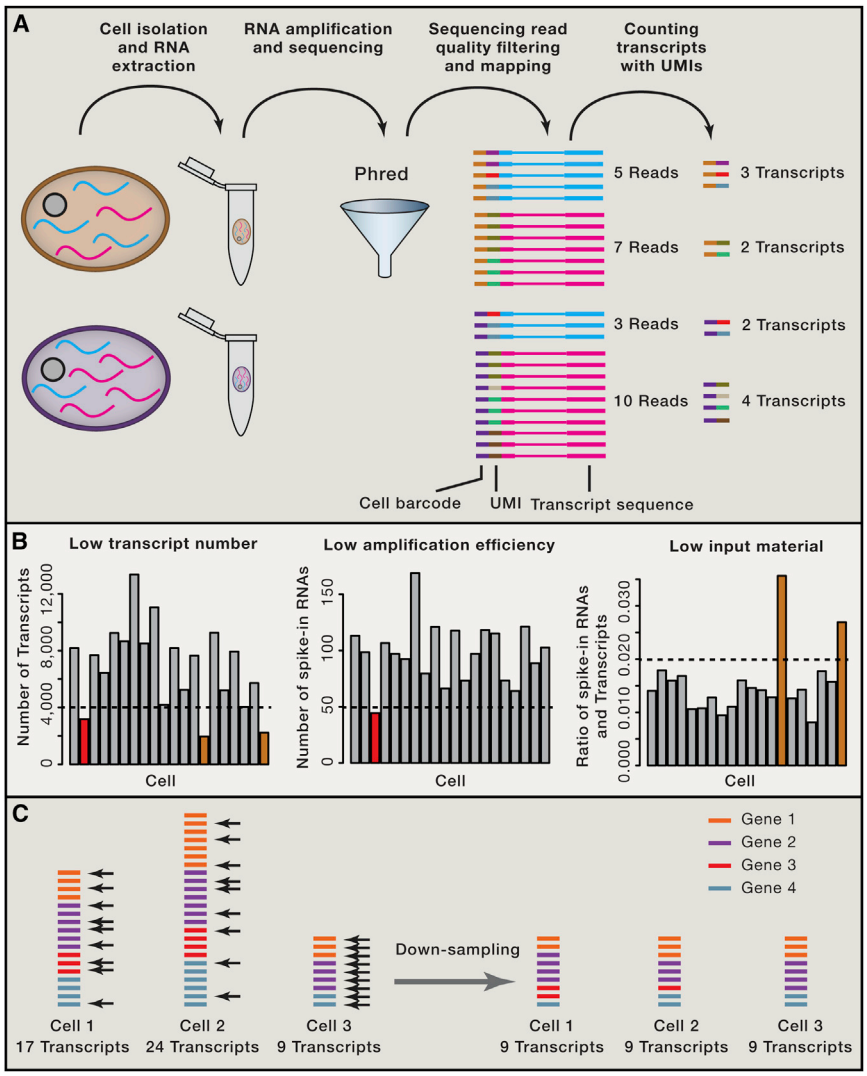

Figure 2. Quantification of mRNA Expression with UMIs

(A) For single-cell sequencing, RNA is isolated from individual cells and, after labeling with cellular barcodes, amplified by PCR or in vitro transcription. Sequencing reads are subject to quality filtering and trimming before mapping to reference sequences representing all genes of the organism. In (A), only two cells with two different genes are shown for simplicity. Amplification bias can distort the relative expression of the two genes and can be eliminated by counting the number of UMIs per genes instead of sequencing reads.

(B) Cells with low yield due to RNA degradation or low sequencing efficiency should be discarded. These cells can be identified based on low total transcript counts (left), which can be explained by low-amplification efficiency (red bar, middle) or low-input material (orange bar, right). The middle panel depicts the total number of spike-in RNA, which should theoretically be the same in all cells. Variations are due to variability in sequencing efficiency. The right panel shows the ratio between spike-in RNA and transcripts of cellular genes. High ratios correspond to reduced amounts of cellular RNA.

(C) Data normalization by down-sampling. The same number of transcripts is randomly picked from each cell. Shown is a toy example with three cells and four different genes.

cell. However, cell-to-cell variability in sequencing efficiency and other sources of technical noise contribute to the observed variability. In principle, the technical cell-to-cell variability could be deconvolved with the help of spike-in RNA. The ratio of the number of sequenced spike-in molecules over the number of spike-in molecules added to the cell extract yields a conversion factor. In theory, dividing the number of sequenced transcripts by this conversion factor yields an estimate of the actual number of transcripts in a cell. However, as already discussed, commonly used ERCC spike-in RNA does not provide a good standard for absolute quantification. For most applications, the relative contribution of each gene to the transcriptome will be the relevant readout, and in these cases, simpler normalization schemes apply. One possibility is the normalization of the total transcript count in each cell to the median of the total transcript count across all cells. Alternatively, subsampling of the same number of transcripts from each cell, termed down-sampling (Figure 2C), is more efficient in eliminating technical variability but comes with a loss in complexity since all cells are down-sampled to a number lower or equal than their actual transcript count. However, for most applications, this approach is preferable since technical artifacts such as batch effects are efficiently eliminated.

Biological Insights from Single-Cell Transcriptome Data

Identification of Cell Types

Perhaps the most important application of single-cell mRNA sequencing is the identification of cell types in a complex mixture. The transcriptome of a cell can be interpreted as a fingerprint revealing its identity. An unbiased screening of randomly sampled cells from a mixture, such as an organ, could therefore reveal the cellular composition of this sample. A number of studies could recover known cell types and identify novel marker genes in diverse systems, for example in the spleen (Jaitin et al., 2014), the lung epithelium (Treutlein et al., 2014), or the retina (Macosko et al., 2015). Another recent landmark paper revealed the complex cellular composition of the mouse hippocampus and uncovered novel cell types (Zeisel et al., 2015). Although these studies convincingly demonstrated that single-cell mRNA sequencing is a powerful method for cell type identification, computational methods to leverage the full complexity within single-cell transcriptome data are just beginning to emerge. Distinguishing cell types in a mixture corresponds to a typical unsupervised learning problem in which data points, in this case given by single-cell transcriptomes, are grouped into clusters reflecting subsets of data points that are more similar to each other than to the remaining data points (Figure 3A). A commonly applied

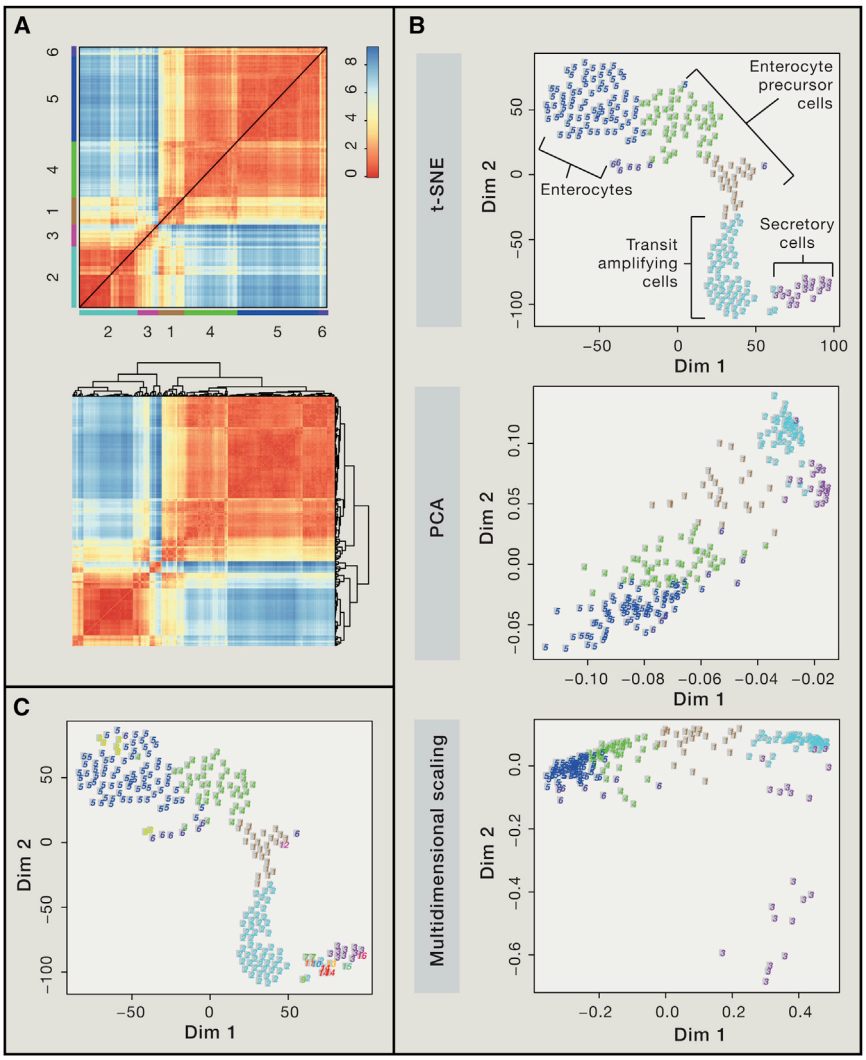

Figure 3. Single-Cell Sequencing Allows the Inference of Cell Type Composition

(A) Unsupervised learning can be used to distinguish different cell types in a random sample of sequenced cells from a complex mixture. K-means clustering with a cluster number estimated by gap statistics (top) or hierarchical clustering (bottom) based on transcriptome similarity can be used to identify different abundant cell types. All data shown in the figure are derived from 238 random cells isolated from mouse intestinal organoids (Grün et al., 2015).

(B) Dimensional reduction algorithms can be applied for data visualization. The t-SNE method (top) resolves the local structure of the data but tends to group outliers together by their dissimilarity to bigger clusters. PCA (middle) allows visual inspection of data separation along the main axis of variability but can be inconvenient if a larger number of principal components contribute substantial variability. Classical multidimensional scaling achieves dimensional reduction with well-conserved point-to-point distances. Outliers are well separated, but dense clusters tend to be condensed. K-means clusters (A) are highlighted in all of the maps, and intestinal cell types are indicated in the t-SNE map.

(C) The RaceID algorithm identifies rare cell types within more abundant groups separated by k-means clustering. The algorithm can detect cell types represented by only a single cell in the mixture.

visual method is principal component analysis (PCA), which converts a set of correlated variables into a set of orthogonal uncorrelated variables, termed principal components. These principal components are ordered by the fraction of the total variance they explain, and usually only the first two or three principal components are analyzed. Visual inspection of a scatterplot showing the first two principal components can already reveal the main subgroups in the data, i.e., the abundant cell types (Pollen et al., 2014; Shalek et al., 2014; Treutlein et al., 2014). Moreover, a number of algorithms for dimensional reduction exist that can be used to obtain an approximate visualization of the data in two dimensions (Figure 3B). These algorithms take a matrix with all pairwise distances of data points as input and project these points onto a low-dimensional space, trying to preserve the original pairwise distances as much as possible. For example, classical multidimensional scaling was used to visualize intratumoral heterogeneity in glioblastoma (Patel et al., 2014), and t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) beautifully visualized heterogeneity within the retina (Macosko et al., 2015) or the hippocampus (Zeisel et al., 2015). These methods take arbitrary similarity measures as input. The most common choices are the Euclidean distance between vectors with expression values for all genes or a correlation-based distance between these vectors, e.g., 1 − Pearson's correlation coefficient.

To identify cell types more systematically, conventional clustering methods can be applied. For instance, hierarchical clustering was used, alone or in combination with PCA, to explore cellular heterogeneity (Patel et al., 2014; Pollen et al., 2014; Treutlein et al., 2014). On the other hand, more sophisticated algorithms have been specifically adjusted for cell type profiling. Jaitin et al. utilized hierarchical clustering to initialize a probabilistic mixture model for cell type classification (Jaitin et al., 2014). Zeisel et al. developed a clustering method based on sorting points into neighborhoods (SPIN) (Tsafrir et al., 2005). In an iterative procedure, an optimal splitting of the cell-to-cell correlation matrix is determined after ordering the expression matrix by cells and genes using SPIN.

A general problem for cell type classification is the presence of confounding factors due to technical and biological variability. The result of any clustering routine has to be carefully examined for batch effects leading to unwanted clustering by experimental batch, sequencing library, or other technical factors. Batch effects can be reduced by normalization strategies such as down-sampling that eliminate differences in complexities between libraries.

However, additional confounding factors can arise from biological heterogeneity such as cell-to-cell differences in the cell-cycle phase. If only cells of a similar size are analyzed, cell sorting can be used to purify cells within a given cell-cycle phase. Otherwise, computational approaches can be used to deconvolve cell-cycle-related variability. A recently published approach utilizes latent variable models to account for the cell cycle and other hidden factors (Buettner et al., 2015). On the other hand, normalization schemes that eliminate absolute cell-to-cell differences in transcript count are often sufficient.

A major challenge for any cell type inference method is the identification of rare cell types. With a frequency of ~1% or less in a sample of sequenced cells from a complex mixture, these cell types typically occur as outliers. Although unsupervised learning methods for outlier identification exist, these approaches oftentimes cannot capture the full complexity of the data. For instance, classifying a variety of different rare cell types in an organ cannot be achieved by these methods (Grün et al., 2015). In a recent study, an algorithm for rare cell type identification (RaceID) was introduced (Grün et al., 2015) that first infers abundant cell types by k-means clustering followed by a systematic outlier screening (Figure 3C). In this step, the cell-to-cell variability of every gene is compared to a background model that accounts for technical and biological noise within a cluster. Cells exhibiting transcript counts with a low p value according to this background model are identified as outliers and are used as new cluster seeds. RaceID was shown to identify rare mouse intestinal cell types with high sensitivity and specificity and discovered novel rare subtypes of the enteroendocrine lineage.

Identification of Marker Genes

Once cell types can be delineated, the data can be mined for specific marker genes to better characterize a cell type and, with the help of cell surface markers or fluorescent reporter genes, allow the purification of a cell type. The discovery of a marker gene requires the identification of differentially expressed genes between the cell type of interest and the remaining cells. For this task, available methods for modeling over-dispersed count statistics in bulk sequencing data, such as DESeq (Anders and Huber, 2010), can be applied. Another probabilistic method, which was developed specifically for single-cell sequencing data, accounts for the relatively high rate of dropout events in these data, i.e., transcripts that escaped reverse transcription and therefore could not be sequenced (Kharchenko et al., 2014).

Inference of Differentiation Dynamics

Related to the cell type inference is the application of single-cell transcriptomics to reveal differentiation pathways. A comparison of single-embryo transcriptomes collected at sub-sequent stages of nematode embryonic development has already revealed insights into gene expression changes underlying the emergence of the three germ layers (Hashimshony et al., 2015). More generally, if a sample is analyzed that contains all differentiation stages of a given cell lineage, a pseudo-temporal ordering of single-cell transcriptomes can be inferred. For example, such a sample can be composed of cells collected at different time points during in vitro differentiation or can be a random sample of a mitotic adult stem cell differentiation system such as the intestinal epithelium. The general idea is that differentiation is accompanied by continuous temporal changes in gene expression and that ordering of single-cell transcriptomes by similarity reflects the succession of these changes, yielding a pseudo-temporal ordering of single-cell transcriptomes. One existing method termed Monocle combines dimensional reduction with the construction of a minimum spanning tree (Trapnell et al., 2014). Monocle is an unsupervised approach that can infer branching into multiple lineages and was used to elucidate gene expression dynamics during differentiation of primary human fibroblasts. Another more recent method relies on the use of diffusion maps to define differentiation trajectories, incorporating the idea that the movement of a cell within the transcriptional landscape follows diffusion-like dynamics (Haghverdi et al., 2015).

Finally, by defining links between gene pairs, e.g., based on the significance of correlation, a variety of network analysis methods can be applied (Ocone et al., 2015).

There is certainly room for further development of computational methods to infer cell lineages. This inference is particularly challenging if the lineage tree segregates into multiple branches, since technical and biological gene expression noise can confound the assignment of a cell to a particular lineage. The single-cell perspective will yield exciting new insights into the impact of gene expression noise on lineage commitment and on the regulation of gene expression noise during differentiation.

Measuring Gene Expression Noise

Another application of single-cell mRNA sequencing is the investigation of biological gene expression variability, or gene expression noise, in a population of cells. Current models of transcriptional dynamics describe promoter bursting (Figure 4A), where the promoter of a gene switches between an active and an inactive state and, once activated, initiates transcript production at a constant rate (Raj et al., 2006; Raser and O'Shea, 2004). These dynamics imply a variance in transcript levels exceeding the lower limit of pure sampling, i.e., Poissonian noise. Single-cell mRNA sequencing is a suitable method to infer the biological noise and investigate transcriptional parameters on a genome-wide level in a cell population of interest. However, technical noise due to sampling of transcripts to be sequenced from each cell and due to global cell-to-cell variability in sequencing efficiency (Figure 4B) has a substantial contribution to the measured noise levels (Brennecke et al., 2013; Grün et al., 2014). The technical noise component can be quantified, for instance, based on sample-to-sample variability in spike-in RNA levels. After fitting a technical noise model that incorporates sampling noise and global sample-to-sample variability in sequencing efficiency, the technical noise component can be deconvolved from the total noise in order to infer the biological noise component (Figure 4C) (Grün et al., 2014). This approach has been shown to yield precise noise estimates consistent with single-molecule FISH, a highly sensitive imaging-based method for transcript counting (Raj et al., 2008), and can be used, for instance, to measure changes in biological noise between different conditions. Furthermore, given a model of transcriptional bursting, kinetics model parameters such as burst size and burst frequency can be derived from the biological noise estimates (Grün et al., 2014).

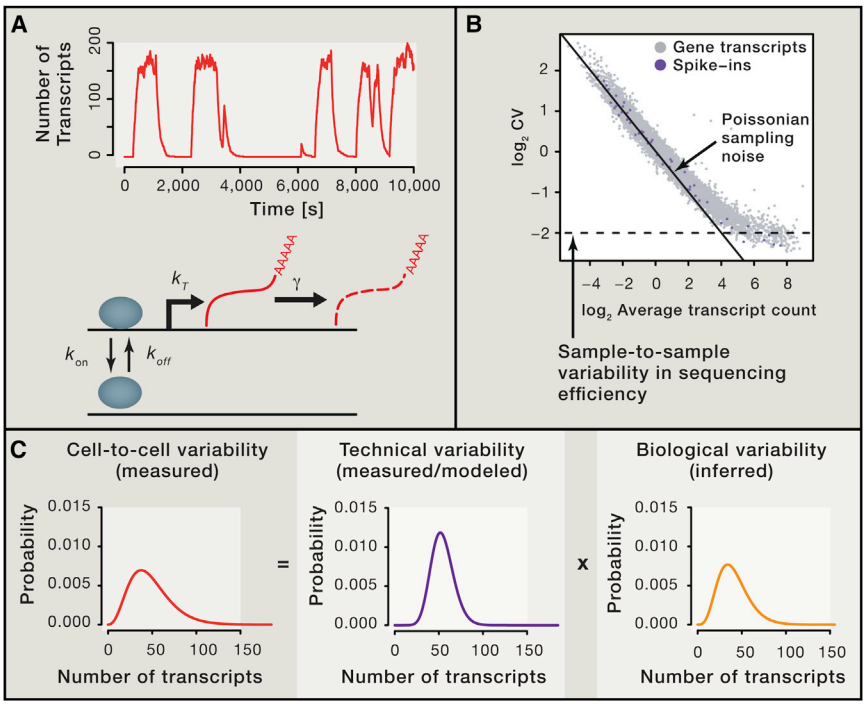

Figure 4. Single-Cell Sequencing Reveals Biological Gene Expression Noise

(A) Transcription is not a time-continuous process. Switching of a gene promoter between an active and an inactive state leads to transcriptional bursting. The kinetic parameters can be estimated from burst size and burst frequency, which can be derived from biological noise estimates measured with single-cell sequencing.

(B) The CV as a function of the mean expression for spike-in RNA or fixed aliquots of cellular RNA reveals sources of technical noise. While sampling noise dominates at low expression, global variability of sequencing efficiency is the major contribution for highly expressed genes.

(C) Technical noise can be modeled and deconvolved from the transcript count distribution measured in cells, yielding good estimates of the actual biological noise (Grün et al., 2014).

Another method allows the identification of highly variable genes by assigning a p value to each gene reflecting to what extent the biological noise exceeds technical variability (Brennecke et al., 2013). This method also relies on technical noise estimates derived from external spike-in RNA.

Investigating Allelic Expression

Single-cell sequencing offers the possibility to study allelic expression on a genome-wide level. If the two alleles of a gene differ by a sufficient number of single-nucleotide polymorphisms (SNPs), transcripts derived from the two alleles can be distinguished by single-cell mRNA sequencing. However, this analysis is highly sensitive to technical noise, i.e., spurious differences in allele frequencies due to sampling effects and stringent controls are required to infer actual biological differences. By analyzing mouse embryos of mixed genetic background, this approach has revealed the presence of abundant random monoallelic expression during preimplantation development and has demonstrated de novo inactivation of the paternal X chromosome (Deng et al., 2014).

Concluding Remarks

The power of single-cell sequencing as a method to characterize the state of a cell across multiple molecular layers has been demonstrated by a number of beautiful studies published during the last few years. Most of the previous research was focused on the investigation of single-cell genomes and transcriptomes. While experimental protocols have improved rapidly, sophisticated computational methods are just beginning to emerge, and in this Primer, we have summarized a number of state-of-the-art methods along with general guidelines covering all analysis stages. We hope that this overview will enable a growing number of researchers to leverage the maximum out of their single-cell sequencing data. The field of single-cell sequencing will keep developing rapidly in the near future and will reveal exciting insights into the regulatory mechanisms that determine the identity of a cell.

REFERENCES

Alexandrov, L.B., and Stratton, M.R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr. Opin. Genet. Dev. *24*, 52–60.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al.; External RNA Controls Consortium (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., et al. (2012). Genome-wide copy number analysis of single cells. Nat. Protoc. *7*, 1024–1041.

Biesecker, L.G., and Spinner, N.B. (2013). A genomic view of mosaicism and human disease. Nat. Rev. Genet. *14*, 307–320.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.

Chattopadhyay, P.K., and Roederer, M. (2012). Cytometry: today's technology and tomorrow's horizons. Methods *57*, 251–258.

Cheung, V.G., and Nelson, S.F. (1996). Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc. Natl. Acad. Sci. USA *93*, 14676–14679.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA *99*, 5261–5266.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193–196.

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. Proc. Natl. Acad. Sci. USA *89*, 3010–3014.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167–173.

Falconer, E., Hills, M., Naumann, U., Poon, S.S.S., Chavez, E.A., Sanders, A.D., Zhao, Y., Hirst, M., and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods *9*, 1107–1112.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nat. Biotechnol. *29*, 51–57.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods *8*, 469–477.

Gole, J., Gore, A., Richards, A., Chiu, Y.-J., Fung, H.-L., Bushman, D., Chiang, H.-I., Chun, J., Lo, Y.-H., and Zhang, K. (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nat. Biotechnol. *31*, 1126–1132.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255. http://dx.doi.org/10.1038/nature14966.

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

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

Hashimshony, T., Feder, M., Levin, M., Hall, B.K., and Yanai, I. (2015). Spatio-temporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature *519*, 219–222.

Iscove, N.N., Barbara, M., Gu, M., Gibson, M., Modi, C., and Winegarden, N. (2002). Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. Nat. Biotechnol. *20*, 940–943.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Junker, J.P., Noël, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P., Berezikov, E., Bakkers, J., and van Oudenaarden, A. (2014). Genome-wide RNA Tomography in the zebrafish embryo. Cell *159*, 662–675.

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

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74.

Klein, C.A., Schmidt-Kittler, O., Schardt, J.A., Pantel, K., Speicher, M.R., and Riethmüller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc. Natl. Acad. Sci. USA *96*, 4494–4499.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Leung, M.L., Wang, Y., Waters, J., and Navin, N.E. (2015). SNES: single nucleus exome sequencing. Genome Biol. *16*, 55.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Macaulay, I.C., and Voet, T. (2014). Single cell genomics: advances and future perspectives. PLoS Genet. *10*, e1004126.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Möhlendick, B., Bartenhagen, C., Behrens, B., Honisch, E., Raba, K., Knoefel, W.T., and Stoecklein, N.H. (2013). A robust method to analyze copy number alterations of less than 100 kb in single cells using oligonucleotide array CGH. PLoS ONE *8*, e67031.

Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. Science *336*, 183–187.

Nagaoka, S.I., Hassold, T.J., and Hunt, P.A. (2012). Human aneuploidy: mechanisms and new insights into an age-old problem. Nat. Rev. Genet. *13*, 493–504.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Ocone, A., Haghverdi, L., Mueller, N.S., and Theis, F.J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics *31*, i89–i96.

Ottolini, C.S., Newnham, L.J., Capalbo, A., Natesan, S.A., Joshi, H.A., Cimadomo, D., Griffin, D.K., Sage, K., Summers, M.C., Thornhill, A.R., et al. (2015). Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. Nat. Genet. *47*, 727–735.

Pan, X., Durrett, R.E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., Marjani, S.L., Euskirchen, G., Ma, C., Lamotte, R.H., et al. (2013). Two methods for full-length RNA sequencing for low quantities of cells and single cells. Proc. Natl. Acad. Sci. USA *110*, 594–599.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. *32*, 1053–1058.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol. *4*, e309.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Saliba, A.-E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. *42*, 8845–8860.

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

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.

Snijder, B., and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. Nat. Rev. Mol. Cell Biol. *12*, 119–125.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell *6*, 468–478.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371–375.

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

Van der Aa, N., Cheng, J., Mateiu, L., Zamani Esteki, M., Kumar, P., Dimitriadou, E., Vanneste, E., Moreau, Y., Vermeesch, J.R., and Voet, T. (2013). Genome-wide copy number profiling of single cells in S-phase reveals DNA-replication domains. Nucleic Acids Res. *41*, e66.

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2570–2605.

Venkatraman, E.S., and Olshen, A.B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics *23*, 657–663.

Voet, T., Kumar, P., Van Loo, P., Cooke, S.L., Marshall, J., Lin, M.-L., Zamani Esteki, M., Van der Aa, N., Mateiu, L., McBride, D.J., et al. (2013). Single-cell paired-end genome sequencing reveals structural variation per cell cycle. Nucleic Acids Res. *41*, 6119–6138.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–412.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155–160.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

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

Zhang, C., Zhang, C., Chen, S., Yin, X., Pan, X., Lin, G., Tan, Y., Tan, K., Xu, Z., Hu, P., et al. (2013). A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. PLoS ONE *8*, e54236.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics *14* (*Suppl 11*), S1.

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