9-2015

Bioinformatics Approaches to Single-Cell Analysis in Developmental Biology

Dicle Yalcin
University of Nebraska-Lincoln, dyalcin2@unl.edu

Zeynep M. Hakguder
zhakguder2@unl.edu

Hasan H. Otu
University of Nebraska-Lincoln, hotu2@unl.edu

Bioinformatics Approaches to Single-Cell Analysis in Developmental Biology

Dicle Yalcin[*], Zeynep M. Hakguder[*], Hasan H. Otu[§]

Department of Electrical & Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0511 USA

[§]Correspondence address. E-mail: hotu2@unl.edu

[*]These authors should be regarded as joint First Authors.

Abstract

Individual cells within the same population show various degrees of heterogeneity, which may be better handled with single-cell analysis to address biological and clinical questions. Single-cell analysis is especially important in developmental biology as subtle spatial and temporal differences in cells have significant associations with cell fate decisions during differentiation and with the description of a particular state of a cell exhibiting an aberrant phenotype. Biotechnological advances, especially in the area of microfluidics, have led to a robust, massively parallel and multi-dimensional capturing, sorting, and lysis of single-cells and amplification of related macromolecules, which have enabled the use of imaging and omics techniques on single-cells. There have been improvements in computational single-cell image analysis in developmental biology regarding feature extraction, segmentation, image enhancement, and machine learning, handling limitations of optical resolution to gain new perspectives from the raw microscopy images. Omics approaches, such as transcriptomics, genomics, and epigenomics, targeting gene and small RNA expression, single nucleotide and structural variations, and methylation and histone modifications, rely heavily on high-throughput sequencing technologies. Although there are well-established bioinformatics methods for analysis of sequence data, there are limited bioinformatics approaches which address experimental design, sample size considerations, amplification bias, normalization, differential expression, coverage, clustering, and classification issues, specifically applied at the single-cell level. In this review, we summarize biological and technological advancements, discuss challenges faced in the aforementioned data

acquisition and analysis issues, and present future prospects for application of single-cell analyses to developmental biology.

Keywords: single-cell bioinformatics, normalization, clustering, missing data, differential expression

Introduction

The majority of life science research is performed on a set of cells assuming homogeneous behaviour among the cells. However, it is well established that cells that are even at close proximity to each other may exhibit heterogeneity at various levels, such as structure, transcription, and epigenetics (Buettner *et al.*, 2015). Therefore, there is a need for a fundamental paradigm shift towards analysis of single cells both at computational and experimental levels. Such an effort creates a number of challenges, including cell isolation, tracking, labeling, imaging, macromolecule amplification, measurement, and data analysis. The answers to these challenges are often intertwined, e.g., the need for new computational approaches accounting for amplification bias due to the distinctive experimental procedures used for single-cells (Pinard *et al.*, 2006).

Single-cell analysis becomes especially important in developmental biology as a small number of cells are usually available for analysis and minute spatial and temporal differences lead to significant changes in cell behaviour by virtue of the inherent

differentiation process. Single-cell analysis comprises three stages: (i) biotechnological and microfluidics approaches that deal with the experimental phase; (ii) imaging, sequencing, microarray, spectrometry, and other platforms for data acquisition; and (iii) data analysis. In this review, we briefly describe the current techniques, issues, and approaches for the first two stages, and focus on the bioinformatics analysis of single-cells within the context of developmental biology.

In its most general setting, bioinformatics methods are blind to the source of the data implying that techniques developed for a certain type of biological data analysis are not affected if the measurements belong to single cells or bulk cells. Nevertheless, contrary to this notion, single-cell data sets bring about unique properties that require specific attention and there is an increasing interest in developing analysis methods for single-cell bioinformatics (Ning *et al.*, 2014; Roach *et al.*, 2009). Some of the peculiar features specific to single-cell analysis that warrant specific bioinformatics approaches are: low volume, nonlinear amplification issues (Wu *et al.*, 2014); unconventional use of spike-ins for normalization due to expression bias (Katayama *et al.*, 2013); contamination from neighbouring cells (Harrington *et al.*, 2010); the need to account for subtle changes that are more likely to be seen in spatial/temporal separation of single-cells which are inherently related by potentially having originated from the same progenitor cell (Buettner and Theis, 2012); models to account for missing data, which is more likely to be seen in single-cell experiments due to insufficient starting material (Buettner *et al.*, 2014); and structure identification in low dimensional data (Feigelman *et al.*, 2014). The last of these features is particularly interesting as it presents a data analysis challenge

that is in between the very low dimensional space of the past (e.g., data sets with a handful of gene measurements) and modern-day, high-throughput data sets (e.g., a typical transcriptomic study with tens of thousands of gene measurements). Due to low initial material, nonlinear amplification, contamination, and background noise, single-cell experimental approaches often resort to techniques where one-to-a-few hundred reliable data points are generated. Such data sets require methods that are on neither the very low- or high-throughput end of the data size spectrum.

Experimental Techniques

Techniques such as polymerase chain reaction (PCR), fluorescence microscopy, microarrays, sequencing, and mass spectrometry have traditionally been successfully applied to a collection of cells (Kalisky and Quake, 2011). Adaptation of these techniques to single-cells is crucial in generating reliable data for bioinformatics analysis. PCR is used for amplification, detection, and quantification of DNA and RNA. Quantitative reverse transcription PCR (qRT-PCR) is sensitive enough to detect and measure a single mRNA molecule. Multiple genes can be quantified by multiplexing PCR (Stahlberg and Kubista, 2014). Microfluidic chips can be used to increase the number of quantitative PCR (qPCR) reactions. The samples and gene detectors are mixed combinatorially enabling thousands of reactions to be run in parallel on a single chip (Marcus *et al.*, 2006). Microfluidic chips can also be used to facilitate single-cell isolation by automating the process and to increase the efficiency of DNA/RNA purification and amplification (Roach *et al.*, 2009; White *et al.*, 2011).

In addition to techniques like laser capture microdissection (LCM) (Emmert-Buck *et al.*, 1996), fluorescence-activated cell sorting (FACS) is commonly used for single-cell isolation. FACS is used for single-cell characterization based on features such as size, granularity, and expression levels of proteins that are located on the cell membrane surface. FACS was used to identify hematopoietic stem cells (Spangrude *et al.*, 1988) and to decode the regulatory networks of hematopoietic development (Moignard *et al.*, 2015). Using DNA binding dyes, FACS can be used to investigate the DNA content of cells to detect genetic abnormalities and to identify stages of the cell cycle in individual cells (Trask, 2002). FACS can measure expression levels of up to a few tens of surface markers with expression above a certain threshold but it loses the spatial information about cells in their tissue context after sorting (Kalisky *et al.*, 2011).

RNA sequencing is used to estimate gene expression levels by mapping the reads to the coding region of genes and counting the mapped reads (Wang *et al.*, 2009). RNA-seq can also be used to detect miRNAs, transcript isoforms, and discover previously unknown transcripts and markers requiring a few µg of starting material, rendering a significant amplification challenge in single-cell studies. In a comparative study, a PCR-based amplification method was proposed for total mRNA amplification from an individual mouse blastomere, and sequencing of amplified RNA resulted in identification of 75% more genes than microarrays and more than a thousand unknown splice junctions (Tang *et al.*, 2009). Application of the same technique to investigate transcriptome changes during embryonic stem cell (ESC) formation from inner cell mass

(ICM) in blastocysts resulted in identification of transcript isoforms and miRNAs (Tang *et al.*, 2010). A recent improvement has been fluorescent in-situ sequencing (FISSEQ), in which amplification of transcripts and fluorescence imaging of the resulting amplicons take place in situ (Lee *et al.*, 2014). Expression levels measured by FISSEQ were shown to have good correlation with RNA-seq. Although FISSEQ generates fewer reads than RNA-seq, it mainly detects informative genes that represent cell type and function. Moreover, quantifying RNA expression within the cell provides further biological insights, such as spatial organization of transcripts and live observation of transcript abundance.

Sequencing has been the emerging method for single-cell RNA and DNA analysis (Baslan and Hicks, 2014). However, single-cell DNA analysis has been more challenging than RNA analysis as the raw material is scarcer and requires a higher degree of amplification. Several PCR-based methods, including primer extension preamplification (PEP) (Xu *et al.*, 1993; Zhang *et al.*, 1992) and degenerate oligonucleotide-primed PCR (DOP-PCR) (Telenius *et al.*, 1992; Wilton *et al.*, 2001), have been established and evaluated. These methods have limitations, such as limited yield, strong bias, and low genome coverage (Cheung and Nelson, 1996; Coskun and Alsmadi, 2007; Kittler *et al.*, 2002). Single-cell specific amplification protocols, such as the multiple annealing and looping-based amplification cycles (MALBAC), have been described (Zong *et al.*, 2012). A non-PCR based whole genome amplification method, called multiple displacement amplification (MDA) has been introduced (Dean *et al.*, 2002). MDA shows some unique advantages over PCR-based whole genome

amplification (WGA), including better fidelity (less error rate), higher average yield from a single-cell (Handyside *et al.*, 2004; Jiang *et al.*, 2005; Spits *et al.*, 2006), larger amplified DNA fragments, and more uniform representation of sequences. However, MDA can generate a high rate of chimeric sequences (1 per 10kb) (Dean *et al.*, 2002; Rodrigue *et al.*, 2009) and may lead to the amplification of even small quantities of contaminating DNA as well as dimerized primer pairs since random primers are used to initiate polymerization (Binga *et al.*, 2008; Raghunathan *et al.*, 2005; Zhang *et al.*, 2006). Contamination problems can be addressed by UV treatment of reagents (Zhang *et al.*, 2006) and reducing the amplification volume to a nanoliter scale (Marcy *et al.*, 2007; Wu *et al.*, 2014). Despite these limitations, the single-cell genomics approach has enabled researchers to: determine population level microheterogeneities (Blainey *et al.*, 2011), study cell-to-cell interactions (Yoon *et al.*, 2011), improve phylogenetic resolution of microbial diversity (Heywood *et al.*, 2011), reclassify an organism (Fleming *et al.*, 2011), and even study single viral genomes (Allen *et al.*, 2011; Tadmor *et al.*, 2011). A recent approach is the microwell displacement amplification system (MIDAS), which is a massively parallel polymerase cloning method (Gole *et al.*, 2013). Single-cells are distributed into thousands of nanoliter wells and their DNA is amplified for shotgun sequencing. It has been shown that MIDAS can reduce the amplification bias as the cloning step occurs in physically isolated nanoliter-scale reactors. Isolation and amplification of single chromosomes from individual cells is also possible. A microfluidic device was developed to separate and amplify homologous chromosomes from an individual human cell in independent chambers. Using this device, alleles of the homologous chromosomes were studied independently (Fan *et al.*, 2011).

Imaging

The viewing proteins and cellular components has provided much of the progress in cell biology since the invention of the microscope. Antibody staining has been the common method for visualizing proteins in fixed cells despite issues challenging its reliability (McDonough *et al.*, 2015). With the development of genetically encoded fluorescent proteins, proteins can be localized and their movement can be monitored within a single-cell. Fluorescent in-situ hybridization (FISH) is a method that utilizes fluorescently tagged oligonucleotides to analyze DNA and/or RNA molecules. Compared to other single-cell analysis methods, shape and position of the cell or the tissue that is being studied is better preserved in microscopy, generally by using fixation, which helps in understanding the spatial relationships between cells or cellular parts and the effect of spatial organization on gene expression.

Live cell fluorescence microscopy is one of the most commonly used techniques to track, visualize and quantify dynamic cellular processes in living cells at a molecular level (Chalfie *et al.*, 1994). Fluorescence correlation spectroscopy (FCS) is used for measuring molecular movement where detection is achieved at single molecule level using a focused laser beam across a minute, defined volume (Singh and Wohland, 2014). For investigating the quantitative measurements of molecular mobility, kinetics, and translocation mechanisms of target proteins and their subtypes in distinct cellular compartments, imaging techniques such as fluorescence recovery after photobleaching

(FRAP) (Staras *et al.*, 2013), inverse-FRAP, and fluorescence loss in photobleaching (FLIP) are used (Ishikawa-Ankerhold *et al.*, 2012; Shav-Tal *et al.*, 2004).

The dynamic structure of a living cell and the biochemical events taking place in real time provide insights to the spatiotemporal and biophysical state of the cell. Segmentation (in combination with surface rendering) and tracking are used for quantitative image analysis and further analysis of kinetic measurements (Gebhard *et al.*, 2002). To track individual particles that travel independent of one another, single particle tracking methods are preferred (Eils and Athale, 2003). For the determination of complex movement, optical flow and image registration methods are commonly employed. Optical flow methods estimate the local motion directly from local intensity value changes in image sequences (Amat *et al.*, 2013; Delpiano *et al.*, 2012). Image registration aims to combine different data sets by projecting them on the same reference coordinate set (Wang *et al.*, 2014). This helps to identify the local dynamics within a cell by rectifying translational and rotational movements over time. To evaluate diffusion, binding and trafficking in live cells, concentration changes by FRAP and FLIP are generally used as standard methods. In Figure 1, we summarize the experimental and imaging workflows used in single-cell analysis.

Data Analysis

Normalization

One of the first issues in single-cell bioinformatics analysis is the need for normalization due to amplification biases introduced by scarce amounts of starting RNA/DNA material. This challenge should be addressed using a combination of experimental and computational methods. In a study by Wu *et. al.* (Wu *et al.*, 2014), amplification methods for RNA-seq were compared using 102 cultured HCT116 single-cell samples. Single-cell RNA-seq data were compared against bulk-cell RNA-seq and multiplexed quantitative PCR data. The results suggest that amplification bias in single-cell RNA-seq is reduced, and high quality data is produced when sample preparation is performed in nanoliter-scale reaction volumes using a microfluidics device. Single-cell specific RNA-seq protocols also exist, such as Smart-seq (Ramskold *et al.*, 2012), Quartz-Seq (Sasagawa *et al.*, 2013), Strt-Seq (Islam *et al.*, 2011), and Cel-seq (Hashimshony *et al.*, 2012), and have significantly improved transcriptome coverage and data quality. Some of these and other similar methods were tested successfully on single mouse oocytes and single mouse embryonic stem cells (Tang *et al.*, 2009). Other experimental techniques to address amplification bias include the use of External RNA Control Consortium (ERCC) synthetic spike-in molecules (Jiang *et al.*, 2011) and “unique molecular identifiers” based barcoding to estimate the number of transcribed molecules (Islam *et al.*, 2014).

RNA Sequencing

From the computational end, algorithms deal with unequal sequencing depths and total transcript numbers coupled with amplification bias. Single-cell RNA-seq data obtained from MCF7 cells amplified using in-vitro transcription (IVT)-based linear amplification (Morris *et al.*, 2011) were compared against the corresponding bulk-cell RNA-Seq data (Vassou *et al.*, 2015). The use of LOWESS (LOcally WEighted polynomial regreSSion) (Cleveland, 1981) and housekeeping-genes-based normalization approaches have been shown to improve the data quality. However, as the use of housekeeping genes requires careful selection of stable expression across samples, it may be better to use the ERCC spike-ins instead. A recently described method, called “remove unwanted variation” (RUV), (Risso *et al.*, 2014) adjusts for technical effects (e.g., disproportion between spike-in read counts and concentrations) by using factor analysis on a subset of suitable control genes (e.g., spike-in or housekeeping) or samples (e.g., technical replicates). The RUV normalization approach has been shown to result in an improved fold change and differential expression analysis. Improvements have been proposed for existing bulk-cell RNA-seq normalization methods, such as SAMstrt, which is tested on mouse embryonic stem cells and fibroblasts that have ~100-fold sequencing depth differences (Katayama *et al.*, 2013).

Typical RNA-seq normalization methods calculate signal values often represented as fragments per kilobase of transcript per million mapped reads (FPKM), which aim to represent transcript concentrations. A recent method suggests a novel use of ERCC

spike-ins for single-cell RNA-seq data by using the FPKM values to model known spike-in concentrations (Ding *et al.*, 2015). This reverse approach is applied by fitting a gamma regression model (GRM) between sequencing reads (e.g., FPKM) and spike-in ERCC concentrations. For each run, the fitted model built using known concentrations is applied to the remaining transcripts to estimate corresponding concentrations. GRM was applied to an RNA-seq data set of four developmental stages (E14.5, n=45; E16.5, n=27; E18.5, n=80; adult, n=46) of individual mouse lung cells. Significant improvements in sample correlations and clustering of individual groups were achieved. Another peculiarity of single-cell RNA-seq normalization arises from the estimated transcript length. In bulk-cell RNA-seq approaches, full-length transcripts may be used to calculate FPKM values, as this likely represents the mode of the transcript length distribution across the cells. However, in single-cell transcriptomics, expression levels should be normalized using coverage lengths (Ning *et al.*, 2014) as the transcript length is likely to be fixed within the cell.

There are also approaches that incorporate noise models to account for gene expression variability in single-cell transcriptomics. It has been shown that technical noise is higher in genes with low expression levels and a statistical method is proposed to remove this noise to identify biological variation with greater success (Brennecke *et al.*, 2013). In another study performed on mouse ESCs (Grun *et al.*, 2014), two types of technical noise were described: random sampling (Poissonian) noise and variability due to sequencing efficiency affecting lowly and highly expressed genes, respectively.

Models to quantify and eliminate both noise types have been proposed and the role of culture conditions in expression variability has been established.

Understanding and characterizing the noise sources in single-cell data are still very challenging, and this formed the premise for a study that used highly expressed genes to build a Poisson-beta model to infer the kinetics of gene expression in single-cell RNA-seq (Kim and Marioni, 2013). In this paper, the transitions of genes from “on” and “off” states, as well as transcription bursts were modeled for mouse ESC data. The resulting kinetics was confirmed by measuring consistency with PolII binding and chromatin modification. The algorithm Monocle was also developed to infer gene expression kinetics from single-cell RNA-seq data (Trapnell *et al.*, 2014). In this approach, high-dimensional transcriptomic data is reduced to a lower dimension using independent component analysis. A minimum spanning tree is built using cells as the nodes and the longest path in this tree is considered as the most viable trajectory, which is used to infer expression kinetics and reveal the dynamics of cell fate decisions. For a more in-depth coverage of single-cell transcriptomics, we refer the reader to two recent review articles (Kolodziejczyk *et al.*, 2015; Stegle *et al.*, 2015).

DNA Sequencing

Amplification bias in DNA sequencing affects bioinformatics approaches that deal with sequence assembly and algorithms that call single nucleotide polymorphisms (SNP), copy number variations (CNV), and structural variants (SV). Although methods like MDA

and MALBAC offer improvements, single-cell genome coverage is still too low (~25x, 75% coverage) compared to its bulk-cell counterpart (~4x, 90% coverage) (Zong *et al.*, 2012). An important statistical inference is estimating the coverage in single-cell whole genome sequencing. In a recent method, a compound Poisson model for sequencing followed by an empirical Bayes estimator for coverage was proposed (Daley and Smith, 2014). The proposed method can be used prior to deep sequencing to estimate the coverage performance of the intended experimental workflow with shallow sequencing. Another challenge in single-cell whole-genome sequencing is posed due to a phenomenon called “allele dropout” (ADO), which is defined as loss of heterozygosity due to amplification failure of one of the two alleles. ADO rates can be as high as 60% for single-cell DNA sequencing studies and specifically affect variant-calling algorithms (Ren *et al.*, 2007). Although there are no specific algorithms for SNP calling for single-cell DNA sequencing, bulk-cell SNP-calling algorithms are used in conjunction with microarray-based SNP detection to improve fidelity (Ling *et al.*, 2009). The current false positive rate for single-cell SNP calling is estimated at around 5% (Ning *et al.*, 2014). There are, however, single-cell specific CNV-calling algorithms, which generally increase the bin size to a few kb (as opposed to a few hundred bp seen in bulk-cell sequencing) and use varying bin sizes (Baslan *et al.*, 2012; Navin *et al.*, 2011).

Assembly of whole genome sequencing has received less attention than variant calling in developmental biology, as the reference genomes of the model organisms are already well established. Single-cell genome assembly is challenging due to the highly non-uniform coverage. Techniques exist to address low-coverage regions by using a

dynamic cut-off to prune contigs from the de Bruijn graph of individual assemblies (Chitsaz *et al.*, 2011) and tree-based decision systems that choose the best workflow through combinatorial testing of different stages of single-cell genome assembly (Harrington *et al.*, 2010). Another algorithmic question arises from the need to construct a phylogeny-like similarity tree that exhibits genomic mutational changes along different temporal and lineage groups of single cells. Although more relevant in areas such as cancer than in development, a recent method provides an evolutionary mutation tree based on single-cell sequencing data (Kim and Simon, 2014). Using a likelihood function to incorporate ADO, mutations between pairs of samples are obtained. A Bayesian approach is applied to identify mutation ordering, and finally a minimum spanning tree algorithm is used to find the final tree, which is the maximum likelihood tree depicting the order and estimated time of mutations along its branches. The algorithm was successfully applied to data from exome sequencing of 58 single cells of an essential thrombocythemia tumour (Hou *et al.*, 2012) but is extendable to other genomic variation measurements, such as CNV.

Comparative Analysis

In a single-cell sequencing project, genomic and/or transcriptomic information on dozens of individual cells is obtained. Comparative studies aim to identify structural variants or transcripts that are differentially abundant in different cells or cell populations. In development applications, single-cell sequencing has been utilized to investigate: the relationships between different stem cell stages (Tang *et al.*, 2009), the

transcriptome changes from oocyte to morula in human and mouse embryos (Xue *et al.*, 2013), the derivation of embryonic stem cells from the inner cell mass using mRNA and miRNA expression (Tang *et al.*, 2010), the relationship between cell fate decisions and gene expression going from zygote to blastocyst (Guo *et al.*, 2010), the heterogeneity of human-induced pluripotent stem cells (Narsinh *et al.*, 2011), and the character of stem cells and early embryos (Liu *et al.*, 2014). To assess differential expression, techniques developed for high-throughput data, such as microarrays and RNA-seq (Durinck, 2008; Rapaport *et al.*, 2013; Soneson and Delorenzi, 2013), are generally adapted to single-cell analysis. However, one of the challenges in single-cell comparative analysis is the identification of the classes of cells that exhibit homogeneous expression as the cell populations exhibit heterogeneous behaviour (Martinez Arias and Brickman, 2011; Narsinh *et al.*, 2011). Therefore, in single-cell analysis, comparative analysis goes hand-in-hand with clustering approaches to identify groups for differential analysis (Roach *et al.*, 2009). Alternatively, there exist single-cell bioinformatics methods that infer gene regulatory networks (GRN) to compare different biological states. Applied to RNA-seq data from single-cell mouse preimplantation embryo blastomeres (Taher *et al.*, 2015), network biology tools, like the PluriNetWork (Som *et al.*, 2010) and ExprEssence (Warsow *et al.*, 2010), were used to infer GRNs for different cell stages. Recently, a genetic algorithm-based GRN inference method for single-cell transcriptomic data was proposed (Chen *et al.*, 2015). GRNs are modeled as probabilistic Boolean networks and a guide tree representing cell lineage structure is used to incorporate the cell development dynamics. The approach has successfully identified GRNs governing cell fate decisions transitioning from the 16-cell stage into the trophectoderm and ICM

states, as well as from the ICM into primitive endoderm and epiblast, using 1- to 64-cell stage mouse transcriptomic data.

When comparing experimental data from two separate measurements, it is essential to account for different hidden factors, such as the cell-cycle state that might result in gene expression heterogeneity in single-cells, which are not observed in bulk-cells, as an average profile is measured. In a study by Buettner *et al.* (Buettner *et al.*, 2015), the authors described a computational approach that uses single-cell latent variable models (scLVM) to reconstruct the hidden factors from the observed data. The model is used to assess the variance in expression explained separately by the biological, technical, and hidden factors. Using in-house generated and existing (Sasagawa *et al.*, 2013) mouse embryonic stem cell data, the scLVM method identified physiologically meaningful subpopulations, which otherwise would be disregarded. When nonlinear principal components analysis (PCA) was applied to “cell-cycle corrected data,” accounting for cell cycle related variation, two clear subpopulations of cells that correspond with physiologically distinct subsets emerge. Application of this approach to additional single-cell RNA-seq datasets, from 34 human embryonic stem cells (hESC) and a set of 90 cells from human preimplantation embryos (Yan *et al.*, 2013), verified that cell cycle explains most of the variability in expression and correlates with different cell populations. Correcting this attribute as a confounder uncovers hidden structures that would otherwise go undetected.

The heterogeneity seen in single-cell populations is more subtle than well-defined, bulk-cell phenotypes, as single-cell analyses generally aim to identify the differences between cells that are considered to exhibit a homogeneous behaviour. With high dimensional data, such as RNA-seq, standard distance measures, such as the Euclidean distance between data points, fail to resolve the true clustering due to the small distance between measurement profiles. A more refined distance between data points is defined as the shared nearest neighbor (SNN). Using a generic distance measure (e.g., Euclidean) and a fixed-sized neighbourhood, SNN considers the intersection of these neighbourhoods between two data points (Huttenhower *et al.*, 2007). A similarity graph is constructed where nodes represent data points, and a link between two nodes represents the overlap of the neighbourhoods of the two nodes. A quasi-clique-based graph clustering algorithm (Zhang *et al.*, 2009) was applied to SNN-based similarity graphs obtained using RNA-seq data (Xu and Su, 2015). The proposed algorithm, SNN-Cliq, automatically determines the number of clusters and identifies clusters of different densities and shapes. When applied to single-cell RNA-seq data regarding human oocytes and human (Yan *et al.*, 2013) and mouse (Deng *et al.*, 2014) early embryonic development stages, SNN-Cliq has identified clusters based on cell stages, embryo, and library preparation protocols. Moreover, clusters of genes that describe the embryonic development and maternal to zygotic transitions in both organisms were identified.

Due to low starting material, dropout events are common in single-cell transcriptomics which means that an existing transcript will not be sequenced. To account for this, a

mixture model is proposed to separately model measured and dropout transcripts (Kharchenko *et al.*, 2014). Measured transcripts are modeled using negative binomial distribution and the dropout rate is approximated with logistic regression. Resulting error models are used in the single-cell differential expression (SCDE) method where a Bayesian framework is used to estimate the likelihood of gene expression and fold change.

Low-dimensional Analysis

Due to problems such as amplification bias and background noise in high-throughput, single-cell experiments, it is common to resort to low-dimensional measurements such as the qPCR and FACS methods. A quality control and comparative analysis method has been developed addressing single-cell multiplexed qPCR data (McDavid *et al.*, 2013). In this approach, a z-transform-based measure of positive expression values is used to filter outliers and has been proposed to replace the generic qPCR normalization methods. This provides an alternative solution as the dichotomous nature of single-cell expression (the “off” state of genes), which is not observed in bulk cells, hinders the use of generic normalization approaches. For differential expression analysis, a likelihood ratio test that simultaneously tests for differences in both means and proportions of gene expression across samples is proposed. Compared to other common methods (e.g., t-test) for differential expression, the proposed method identifies differentially expressed genes that are superior both in quantity (for fixed false discovery rates) and relevance.

Similar to high-throughput, single-cell data, there is a need to identify the subpopulations in single cells based on low-dimensional expression data. Multiresolution correlation analysis (MCA) was developed for just such data to visualize the correlations of data subsets of all sizes, thereby enabling regions with robust correlations that may indicate distinct subpopulations to be distinguished (Feigelman *et al.*, 2014). MCA estimates deteriorate with small sample size or a large number of variables, which also makes it difficult to generate all possible MCA plots due to the increase in dimension. When MCA was used to analyze qPCR single-cell transcriptomic data from mouse embryonic stem cells (Hayashi *et al.*, 2008; Trott *et al.*, 2012), new biologically relevant subpopulations were discovered and previously identified subgroups were confirmed.

For data sizes of similar dimensionality, a Gaussian process latent variable model (GPLVM) based nonlinear probabilistic generalization of PCA was proposed (Buettner and Theis, 2012). The proposed method was applied to qPCR expression data of 48 genes from 442 single mouse cells at different developmental stages (zygote to blastocyst) (Guo *et al.*, 2010). A linear PCA-based method can distinguish between the trophectoderm, endoderm, and epiblast cell types at the 64-cell stage but fails to find distinguishing characteristics at the 2-, 4-, and 8-cell stages. On the other hand, the GPLVM-based dimension reduction approach successfully separates all cell types and all cell stages using a nonlinear, probabilistic 2D embedding of the higher-dimensional expression data. Another dimension reduction method, called viSNE, has been

developed using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Amir el *et al.*, 2013). Originally developed for mass cytometry data, viSNE projects the high-dimensional single-cell data on to two dimensions, by minimizing the difference in the ensemble pairwise distance observed in high- and low-dimensional space, and has successfully been applied to leukemic human bone marrow data.

Also developed using mass cytometry data, the algorithm Wanderlust constructs a trajectory of cell lineages predicting the developmental path (Bendall *et al.*, 2014). Assuming the developmental process is serial with no branching, the algorithm was applied to human B cell lymphopoiesis, ordering the cells according to their developmental chronology from hematopoietic stem cells to naïve B cells. Nonlinearity of the distance between the measured parameters of cells in different stages is overcome by a graph representation where nodes represents cells and links connects a cell to the ones most similar to it. Such a graph representation is reduced to linear trajectories by placing a cell on the trajectory using its shortest path to the user-defined start cell. The trajectories are used to identify expression kinetics and key molecular and cellular events during development. Another method has been introduced to infer signaling cascades in single-cell mass cytometry data using a protein-based representation instead of considering the relationship between cell states. (Krishnaswamy *et al.*, 2014). In this approach, conditional-density based analysis has been applied to determine the mutual information between pairs of proteins to determine the influence between protein pairs. Using temporal data, the protein-protein

interaction dynamics is calculated and was found to exhibit a change between naïve and antigen-exposed $CD4^+$ T-cells of B6 mice.

Low-dimensional multiplexed single-cell qPCR data provides reliable measurements but has a limit of detection below which gene activity cannot be quantified. This, in turn, requires censored data analysis, which is not as commonly seen in bulk-cell measurements. In qPCR analysis, non-detected values are either removed, or substituted by a constant, or imputed. The first two methods result in information loss while data imputation models are heavily dependent on expression distributions, which are unknown. Moreover, the effect of such remedies on downstream analysis steps, such as clustering and classification, are not immediately clear. In order to address these issues, a noise model based on the probit function is introduced to handle the censored data. After a Gaussian approximation is found for the noise model, nonlinear probabilistic PCA using GPLVM is applied to identify the subpopulations in the data. The proposed approach was shown to better separate known cell types and identifies subpopulations not discovered using standard censoring and PCA approaches using mouse stem cell data (Guo *et al.*, 2010).

In Table I, we list the bioinformatics algorithms developed specifically for single-cell analysis, noting the accessibility and problem/solution summary of the algorithm.

Discussion

As a discipline functioning at the intersection of life and computational sciences, bioinformatics approaches do not have the luxury of being blind to the biological characteristics of the underlying data. Single-cell analysis provides a new venue for bioinformatics, as bulk-cell data analysis methods may not be directly applicable to single-cell data. In this review, we listed the challenges posed by single-cell data and summarized methods that address these challenges. Single-cell approaches have been widely used, especially in development, as the spatiotemporal organization of the cells vastly affects their characteristics. In addition to imaging data analysis, the bulk of the problems are rooted in omics-based approaches, which are dominated by transcriptomic profiling (e.g., RNA-seq, qPCR) and genomic approaches, addressing assembly, SNP, CNV, and SV calling. In bulk-cell data, the measurements target the output from an ensemble of cells generating a data matrix that is not sparse. In single-cell experiments, factors such as scarce input material, amplification/coverage bias, lack of observation for a significant number of data points due to the “off” state of DNA/RNA molecules, low dimensionality of high quality data, and subtle, biologically meaningful heterogeneity seen in well-defined phenotypes require specific attention.

The approaches geared to single-cell analysis roughly fall into six categories: normalization approaches accounting for highly prevalent amplification, coverage, sequencing depth, and input material biases; methods functioning at the presence of missing data; algorithms focused on low-dimensional, semi-high-throughput data sets;

clustering methods aimed at identifying subtle heterogeneities to discover well-characterized populations; specific noise and signal models for differential expression analysis; and identification of genome level variations. For SNP calling algorithms and downstream prediction, functional, network/pathway-based approaches, the tendency has been to resort to existing approaches. Therefore, there is room for improvement in these analysis areas to develop algorithms accounting for single-cell data characteristics. It is also desirable to analyze DNA/RNA measurements from the same cell in parallel to relate genomic variations with expression profiles. Although there are some initial attempts (Dey *et al.*, 2015; Macaulay *et al.*, 2015), a more integrated approach, possibly including the epigenome and the proteome is needed for a more comprehensive view of the single cell. An important challenge lies in spatial mapping of individual cells given experimental data (Achim *et al.*, 2015; Satija *et al.*, 2015). This often requires incorporating existing external knowledge in the mapping strategy, which is not readily available for different organisms, organs, or cell types. One area that might expedite the advances in this venue as well as in others is the barcoding of individual cells that enables high-throughput sequencing using droplets (Klein *et al.*, 2015; Macosko *et al.*, 2015). There is also a need to define technological standards and gold data sets to accurately assess the performance of different bioinformatics algorithms. Data management is likely to be another challenge for single-cell bioinformatics as the amount of data generated far surpasses its bulk-cell counterpart. The scientific community would greatly benefit from single-cell-specific bioinformatics tools with workflows that address the aforementioned issues and provide modules covering each step of the data-analysis phase.

Tables

Table I. List of Bioinformatics algorithms developed for single-cell analysis.

Figure Legends

Figure 1. Schematic overview of single cell analysis. Individual cells are captured and isolated from the collected tissue or environmental sample using techniques such as FACS, FISH, and LCM. Imaging, particle tracking, and in-vivo biomolecular interaction assessment can be done using fluorescence recovery after photobleaching (FRAP) or fluorescence resonance energy transfer (FRET) based methods. Isolated cells further go into lysis and separation procedures, preferably using microfluidics approaches where lysis occurs in the device. External, spike-in controls and unique barcodes are used during amplification to later normalize for amplification bias. DNA/RNA sequencing, cDNA/oligo microarrays, and multiplexed qPCR are most common methods for generation of data, which are passed on to the Bioinformatics phase for analysis. ERCC: External RNA Control Consortium. IVT: in-vitro Transcription. MDA: Multiple Displacement Amplification. MALBAC: Multiple Annealing and Looping-based Amplification Cycles. MIDAS: Microwell Displacement Amplification System. FISH: Fluorescent in-situ hybridization. qPCR: Quantitative Polymerase Chain Reaction.

Acknowledgments

We thank Michele Boiani and Jose Cibelli for the opportunity to contribute to this Special Issue. We thank the anonymous reviewers for their comments, which have improved the quality of the manuscript.

Authors' Roles

All authors participated in the conception, design, drafting, revision, and final approval of the manuscript.

Funding

No external funding was used for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* 2015; **33**: 503-9.

Al-Gubory KH, Houdebine LM. In vivo imaging of green fluorescent protein-expressing cells in transgenic animals using fibred confocal fluorescence microscopy. *Eur J Cell Biol* 2006; **85**: 837-45.

Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ. Single virus genomics: a new tool for virus discovery. *PLoS One* 2011; **6**: e17722.

Amat F, Myers EW, Keller PJ. Fast and robust optical flow for time-lapse microscopy using super-voxels. *Bioinformatics* 2013; **29**: 373-80.

Amir el AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 2013; **31**: 545-52.

Baslan T, Hicks J. Single cell sequencing approaches for complex biological systems. *Curr Opin Genet Dev* 2014; **26**: 59-65.

Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B *et al.* Genome-wide copy number analysis of single cells. *Nat Protoc* 2012; **7**: 1024-41.

Bendall SC, Davis KL, Amir el AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 2014; **157**: 714-25.

Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* 2008; **2**: 233-41.

Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR. Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS One* 2011; **6**: e16626.

Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013; **10**: 1093-5.

Buettner F, Moignard V, Gottgens B, Theis FJ. Probabilistic PCA of censored data: accounting for uncertainties in the visualization of high-throughput single-cell qPCR data. *Bioinformatics* 2014; **30**: 1867-75.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 2015; **33**: 155-60.

Buettner F, Theis FJ. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* 2012; **28**: i626-i32.

Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Green fluorescent protein as a marker for gene expression. *Science* 1994; **263**: 802-5.

Chen H, Guo J, Mishra SK, Robson P, Niranjan M, Zheng J. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. *Bioinformatics* 2015; **31**: 1060-6.

Cheung VG, Nelson SF. 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 U S A* 1996; **93**: 14676-9.

Chiang MK, Melton DA. Single-cell transcript analysis of pancreas development. *Dev Cell* 2003; **4**: 383-93.

Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA *et al.* Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol* 2011; **29**: 915-21.

Cleveland WS. LOWESS - A program for smoothing scatterplots by robust locally weighted regression. *American Statistician* 1981; **35**: 54-.

Coskun S, Alsmadi O. Whole genome amplification from a single cell: a new era for preimplantation genetic diagnosis. *Prenat Diagn* 2007; **27**: 297-302.

Daley T, Smith AD. Modeling genome coverage in single-cell sequencing. *Bioinformatics* 2014; **30**: 3159-65.

Danuser G, Tran PT, Salmon ED. Tracking differential interference contrast diffraction line images with nanometre sensitivity. *J Microsc* 2000; **198**: 34-53.

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

Delpiano J, Jara J, Scheer J, Ramirez OA, Ruiz-del-Solar J, Hartel S. Performance of optical flow techniques for motion analysis of fluorescent point signals in confocal microscopy. *Machine Vision and Applications* 2012; **23**: 675-89.

Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014; **343**: 193-6.

Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* 2015; **33**: 285-9.

Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W. Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 2015.

Durinck S. Pre-processing of microarray data and analysis of differential expression. *Methods Mol Biol* 2008; **452**: 89-110.

Eils R, Athale C. Computational imaging in cell biology. *J Cell Biol* 2003; **161**: 477-81.

Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA. Laser capture microdissection. *Science* 1996; **274**: 998-1001.

Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* 2011; **29**: 51-7.

Feigelman J, Theis FJ, Marr C. MCA: Multiresolution Correlation Analysis, a graphical tool for subpopulation identification in single-cell gene expression data. *BMC Bioinformatics* 2014; **15**: 240.

Fleming EJ, Langdon AE, Martinez-Garcia M, Stepanauskas R, Poulton NJ, Masland ED, Emerson D. What's new is old: resolving the identity of Leptothrix ochracea using single cell genomics, pyrosequencing and FISH. *PLoS One* 2011; **6**: e17769.

Gebhard M, Eils R, Mattes J (2002) Segmentation of 3D objects using NURBS surfaces for quantification of surface and volume dynamics. International Conference on Diagnostic Imaging and Analysis (ICDIA). Shanghai, China, pp. 125–30.

Gole J, Gore A, Richards A, Chiu YJ, Fung HL, Bushman D, Chiang HI, Chun J, Lo YH, Zhang K. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* 2013; **31**: 1126-32.

Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014; **11**: 637-40.

Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* 2010; **18**: 675-85.

Handyside AH, Robinson MD, Simpson RJ, Omar MB, Shaw MA, Grudzinskas JG, Rutherford A. Isothermal whole genome amplification from single and small numbers of cells: a new era for preimplantation genetic diagnosis of inherited disease. *Mol Hum Reprod* 2004; **10**: 767-72.

Harrington ED, Arumugam M, Raes J, Bork P, Relman DA. SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics* 2010; **26**: 2979-80.

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

Hayashi K, Lopes SM, Tang F, Surani MA. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* 2008; **3**: 391-401.

Heywood JL, Sieracki ME, Bellows W, Poulton NJ, Stepanauskas R. Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J* 2011; **5**: 674-84.

Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 2012; **148**: 873-85.

Huttenhower C, Flamholz AI, Landis JN, Sahi S, Myers CL, Olszewski KL, Hibbs MA, Siemers NO, Troyanskaya OG, Coller HA. Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* 2007; **8**: 250.

Ishikawa-Ankerhold HC, Ankerhold R, Drummen GP. Advanced fluorescence microscopy techniques--FRAP, FLIP, FLAP, FRET and FLIM. *Molecules* 2012; **17**: 4047-132.

Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. 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, Lonnerberg P, Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014; **11**: 163-6.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011; **21**: 1543-51.

Jiang Z, Zhang X, Deka R, Jin L. Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Res* 2005; **33**: e91.

Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet* 2011; **45**: 431-45.

Kalisky T, Quake SR. Single-cell genomics. *Nat Methods* 2011; **8**: 311-4.

Katayama S, Tohonen V, Linnarsson S, Kere J. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 2013; **29**: 2943-5.

Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods* 2014; **11**: 740-2.

Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* 2013; **14**: R7.

Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* 2014; **15**: 27.

Kittler R, Stoneking M, Kayser M. A whole genome amplification method to generate long fragments from low quantities of genomic DNA. *Anal Biochem* 2002; **300**: 237-44.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015; **161**: 1187-201.

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015; **58**: 610-20.

Krishnaswamy S, Spitzer MH, Mingueneau M, Bendall SC, Litvin O, Stone E, Pe'er D, Nolan GP. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. *Science* 2014; **346**: 1250689.

Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clin Chem* 2005; **51**: 1973-81.

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* 2014; **343**: 1360-3.

Lichtman JW, Livet J, Sanes JR. A technicolour approach to the connectome. *Nat Rev Neurosci* 2008; **9**: 417-22.

Ling J, Zhuang G, Tazon-Vega B, Zhang C, Cao B, Rosenwaks Z, Xu K. Evaluation of genome coverage and fidelity of multiple displacement amplification from single cells by SNP array. *Mol Hum Reprod* 2009; **15**: 739-47.

Liu N, Liu L, Pan X. Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell Mol Life Sci* 2014; **71**: 2707-15.

Livet J, Weissman TA, Kang H, Draft RW, Lu J, Bennis RA, Sanes JR, Lichtman JW. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 2007; **450**: 56-62.

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* 2015; **12**: 519-22.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015; **161**: 1202-14.

Marcus JS, Anderson WF, Quake SR. Microfluidic single-cell mRNA isolation and analysis. *Anal Chem* 2006; **78**: 3084-9.

Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* 2007; **104**: 11889-94.

Martinez Arias A, Brickman JM. Gene expression heterogeneities in embryonic stem cell populations: origin and function. *Curr Opin Cell Biol* 2011; **23**: 650-6.

McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, Roederer M, Gottardo R. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 2013; **29**: 461-7.

McDonough AA, Veiras LC, Minas JN, Ralph DL. Considerations when quantitating protein abundance by immunoblot. *American journal of physiology Cell physiology* 2015; **308**: C426-33.

Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M *et al.* Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proc Natl Acad Sci U S A* 2008; **105**: 1579-84.

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E *et al.* Decoding the regulatory network of early

blood development from single-cell gene expression measurements. *Nat Biotechnol* 2015; **33**: 269-76.

Morris J, Singh JM, Eberwine JH. Transcriptome analysis of single cells. *J Vis Exp* 2011.

Narsinh KH, Sun N, Sanchez-Freire V, Lee AS, Almeida P, Hu S, Jan T, Wilson KD, Leong D, Rosenberg J *et al.* Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *J Clin Invest* 2011; **121**: 1217-21.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* 2011; **472**: 90-4.

Ning L, Liu G, Li G, Hou Y, Tong Y, He J. Current challenges in the bioinformatics of single cell genomics. *Front Oncol* 2014; **4**: 7.

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006; **7**: 216.

Ponti A, Vallotton P, Salmon WC, Waterman-Storer CM, Danuser G. Computational analysis of F-actin turnover in cortical actin meshworks using fluorescent speckle microscopy. *Biophys J* 2003; **84**: 3336-52.

Racine V, Sachse M, Salamero J, Fraisier V, Trubuil A, Sibarita JB. Visualization and quantification of vesicle trafficking on a three-dimensional cytoskeleton network in living cells. *J Microsc* 2007; **225**: 214-28.

Raghunathan A, Ferguson HR, Jr., Bornarth CJ, Song W, Driscoll M, Lasken RS. Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* 2005; **71**: 3342-7.

Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression underlies incomplete penetrance. *Nature* 2010; **463**: 913-8.

Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012; **30**: 777-82.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013; **14**: R95.

Ren Z, Zhou C, Xu Y, Deng J, Zeng H, Zeng Y. Mutation and haplotype analysis for Duchenne muscular dystrophy by single cell multiple displacement amplification. *Mol Hum Reprod* 2007; **13**: 431-6.

Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014; **32**: 896-902.

Roach KL, King KR, Uygun BE, Kohane IS, Yarmush ML, Toner M. High throughput single cell bioinformatics. *Biotechnol Prog* 2009; **25**: 1772-9.

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* 2009; **4**: e6864.

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

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015; **33**: 495-502.

Shav-Tal Y, Singer RH, Darzacq X. Imaging gene expression in single living cells. *Nat Rev Mol Cell Biol* 2004; **5**: 855-61.

Singh AP, Wohland T. Applications of imaging fluorescence correlation spectroscopy. *Curr Opin Chem Biol* 2014; **20**: 29-35.

Som A, Harder C, Greber B, Siatkowski M, Paudel Y, Warsow G, Cap C, Scholer H, Fuellen G. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* 2010; **5**: e15165.

Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013; **14**: 91.

Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. *Science* 1988; **241**: 58-62.

Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, Sermon K. Optimization and evaluation of single-cell whole-genome multiple displacement amplification. *Hum Mutat* 2006; **27**: 496-503.

Stahlberg A, Kubista M. The workflow of single-cell expression profiling using quantitative real-time PCR. *Expert Rev Mol Diagn* 2014; **14**: 323-31.

Staras K, Mikulincer D, Gitler D. Monitoring and quantifying dynamic physiological processes in live neurons using fluorescence recovery after photobleaching. *J Neurochem* 2013; **126**: 213-22.

Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015; **16**: 133-45.

Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S *et al.* Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proc Natl Acad Sci U S A* 2005; **102**: 4453-8.

Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* 2011; **333**: 58-62.

Taher L, Pfeiffer MJ, Fuellen G. Bioinformatics approaches to single-blastomere transcriptomics. *Mol Hum Reprod* 2015; **21**: 115-25.

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* 2010; **6**: 468-78.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009; **6**: 377-82.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 1992; **13**: 718-25.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014; **32**: 381-6.

Trask BJ. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nat Rev Genet* 2002; **3**: 769-78.

Trott J, Hayashi K, Surani A, Babu MM, Martinez-Arias A. Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol Biosyst* 2012; **8**: 744-52.

Vassou D, Arhondakis S, Kalantzaki K, Zervakis M, Kafetzopoulos D. Towards Single-Cell Gene Expression Profiling: Assessing Amplification Biases. *6th European Conference of the International Federation for Medical and Biological Engineering* 2015; **45**: 598-601.

Viollier PH, Thanbichler M, McGrath PT, West L, Meewan M, McAdams HH, Shapiro L. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc Natl Acad Sci U S A* 2004; **101**: 9257-62.

Wang CW, Ka SM, Chen A. Robust image registration of biological microscopic images. *Sci Rep* 2014; **4**: 6050.

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

Warsow G, Greber B, Falk SS, Harder C, Siatkowski M, Schordan S, Som A, Endlich N, Scholer H, Repsilber D *et al.* ExprEssence--revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* 2010; **4**: 164.

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL. High-throughput microfluidic single-cell RT-qPCR. *Proc Natl Acad Sci U S A* 2011; **108**: 13999-4004.

Wilton L, Williamson R, McBain J, Edgar D, Voullaire L. Birth of a healthy infant after preimplantation confirmation of euploidy by comparative genomic hybridization. *N Engl J Med* 2001; **345**: 1537-41.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* 2014; **11**: 41-6.

Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 2015.

Xu KP, Tang YX, Grifo JA, Rosenwaks Z, Cohen J. Primer extension preamplification for detection of multiple genetic-loci from single human blastomeres. *Human Reproduction* 1993; **8**: 2206-10.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 2013; **500**: 593-7.

Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 2013; **20**: 1131-9.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 2011; **332**: 714-7.

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 2006; **24**: 680-6.

Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A* 1992; **89**: 5847-51.

Zhang S, Xu M, Li S, Su Z. Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res* 2009; **37**: e72.

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.

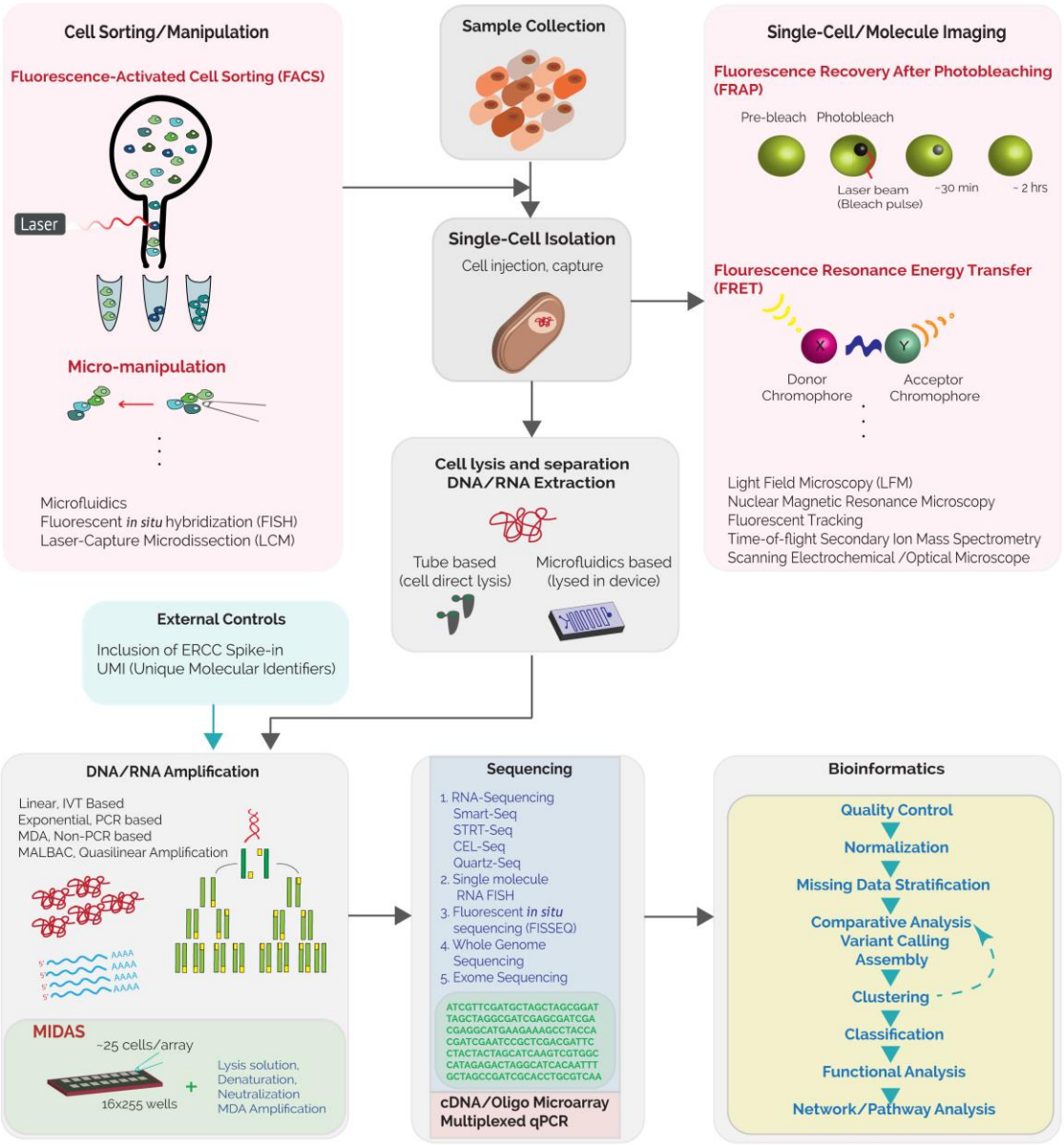
Cell Sorting/Manipulation
Fluorescence-Activated Cell Sorting (FACS)
Laser
Micro-manipulation
Microfluidics
Fluorescent in situ hybridization (FISH)
Laser-Capture Microdissection (LCM)
Sample Collection
Single-Cell Isolation
Cell injection, capture
Cell lysis and separation
DNA/RNA Extraction
Tube based
(cell direct lysis)
Microfluidics based
(lysed in device)
External Controls
Inclusion of ERCC Spike-in
UMI (Unique Molecular Identifiers)
DNA/RNA Amplification
Linear, IVT Based
Exponential, PCR based
MDA, Non-PCR based
MALBAC, Quasilinear Amplification
MIDAS
~25 cells/array
16x255 wells
Lysis solution,
Denaturation,
Neutralization
MDA Amplification
Sequencing
1. RNA-Sequencing
Smart-Seq
STRT-Seq
CEL-Seq
Quartz-Seq
2. Single molecule
RNA FISH
3. Fluorescent in situ
sequencing (FISSEQ)
4. Whole Genome
Sequencing
5. Exome Sequencing
cDNA/Oligo Microarray
Multiplexed qPCR
Single-Cell/Molecule Imaging
Fluorescence Recovery After Photobleaching
(FRAP)
Pre-bleach
Photobleach
Laser beam
(Bleach pulse)
~30 min
~ 2 hrs
Flourescence Resonance Energy Transfer
(FRET)
Donor
Chromophore
Acceptor
Chromophore
Light Field Microscopy (LFM)
Nuclear Magnetic Resonance Microscopy
Fluorescent Tracking
Time-of-flight Secondary Ion Mass Spectrometry
Scanning Electrochemical /Optical Microscope
Bioinformatics
Quality Control
Normalization
Missing Data Stratification
Comparative Analysis
Variant Calling
Assembly
Clustering
Classification
Functional Analysis
Network/Pathway Analysis

Table I. List of Bioinformatics algorithms developed for single-cell analysis.

Reference	Algorithm	Data	Description	Availability
Vassou *et al.,* 2015	Norm.	Microarray	Use of LOWESS (LOcally WEighted polynomial regreSSion) and housekeeping gene selection and application.	Upon request from the authors
Risso *et al.,* 2014	Norm.	RNA-seq	Remove unwanted variation (RUV) uses factor analysis on subset of control genes (e.g., spike-ins) and/or samples (replicates).	http://www.bioconductor.org/RUVSeq
Katayama *et al.*, 2013	Norm.	RNA-seq	SAMstrt uses spike-in controls to normalize and estimate transcript numbers per cell; tolerates variations in sequencing depth.	https://github.com/shka/R-SAMstrt
Ding *et al.*, 2015	Norm.	RNA-seq	Uses External RNA Control Consortium (ERCC) reads and concentrations to build a gamma regression model to estimate RNA concentrations from read counts.	http://wanglab.ucsd.edu/star/GRM
Daley and Smith, 2014	Coverage	DNA-seq	Estimates gain in coverage with increased sequencing depth from initial shallow sequencing using Bayes Poisson models.	http://smithlabresearch.org/preseq
Baslan *et al.*, 2012	CNV	DNA-seq	Varbin uses variable bin sizes to call copy number variations (CNV).	Journal website
Navin *et al.*, 2011	CNV	DNA-seq	Uses variable bin sizes to call copy numbers.	Upon request from the authors
Chitsaz *et al.*, 2011	Assembly	DNA-seq	Addresses low-coverage regions by using de Bruijn graphs with a dynamic cut-off.	http://bix.ucsd.edu/singlecell/
Harrington *et al.*, 2010	Assembly Annotation	DNA-seq	SmashCell uses a tree with branches representing different choice of algorithm or parameters, mostly used in metagenomics.	http://asiago.stanford.edu/SmashCell
Kim and Simon, 2014	Evolutionary tree	Exome-seq	Likelihood function for allele dropouts (ADOs), Bayesian approach for mutation ordering, temporal relationships among mutation sites.	https://sites.google.com/site/kyungin2013
Buettner *et al.*, 2015	Clustering Diff. exp.	RNA-seq	Single-cell latent variable model (scLVM) estimates proportion of variation associated with hidden factors to identify subpopulations.	https://github.com/PMBio/scLVM
Xu *et al.*, 2015	Clustering	RNA-seq	SNN-Cliq uses shared nearest neighbor based similarity graphs. Partitioning of the graphs automatically identifies subgroups of cells.	http://bioinfo.uncc.edu/SNNCliq
McDavid *et al.*, 2013	Norm. Diff. exp.	qPCR	A z-transform-based measure of positive expression is used to filter outliers for normalization and a likelihood ratio tests for differences.	http://github.com/RGLab/SingleCellAssay
Buettner *et al.*, 2012	Clustering Diff. exp.	qPCR	Gaussian process latent variable model (GPLVM) based nonlinear probabilistic Principal Components Analysis (PCA).	http://github.com/SheffieldML/vargplvm
Feigelman *et al.*, 2014	Clustering	qPCR	Multiresolution correlation analysis (MCA) uses local correlation between data of different sizes to visually identify subpopulations.	Upon request from the authors
Buettner *et al.*, 2014	Censored Clustering	qPCR	Models noise using probit function for censored data and applies nonlinear probabilistic PCA with GPLVM to identify subpopulations.	http://icb.helmholtz-muenchen.de
Chen *et al.*, 2014	Network Analysis	qPCR	SingCellNet models Gene Regulatory Networks (GRNs) as probabilistic Boolean networks using a tree representing cell lineage.	Upon request from the authors
Kharchenko *et al.*, 2014	Clustering Diff. exp.	RNA-seq	Single-cell differential expression (SCDE) uses a separate model for dropouts and a Bayesian model for diff. expr.	pklab.med.harvard.edu/scde/index.html
Grun *et al.*, 2014	Noise Model	RNA-seq	Two technical noise sources: random sampling (Poissonian) noise and variability due to sequencing efficiency characterization.	Upon request from the authors
Bendall *et al.*, 2014	Trajectory Analysis	mass cytom.	Using graph depiction of cells, the shortest distance to the user-defined start cell defines a cell's position in the cell lineage trajectory.	c2b2.columbia.edu/danapeerlab
Trapnell *et al.*, 2014	Expression Kinetics	RNA-seq	Monocle uses independent component analysis for dimension reduction and minimum spanning tree for cell ordering.	monocle-bio.sourceforge.net