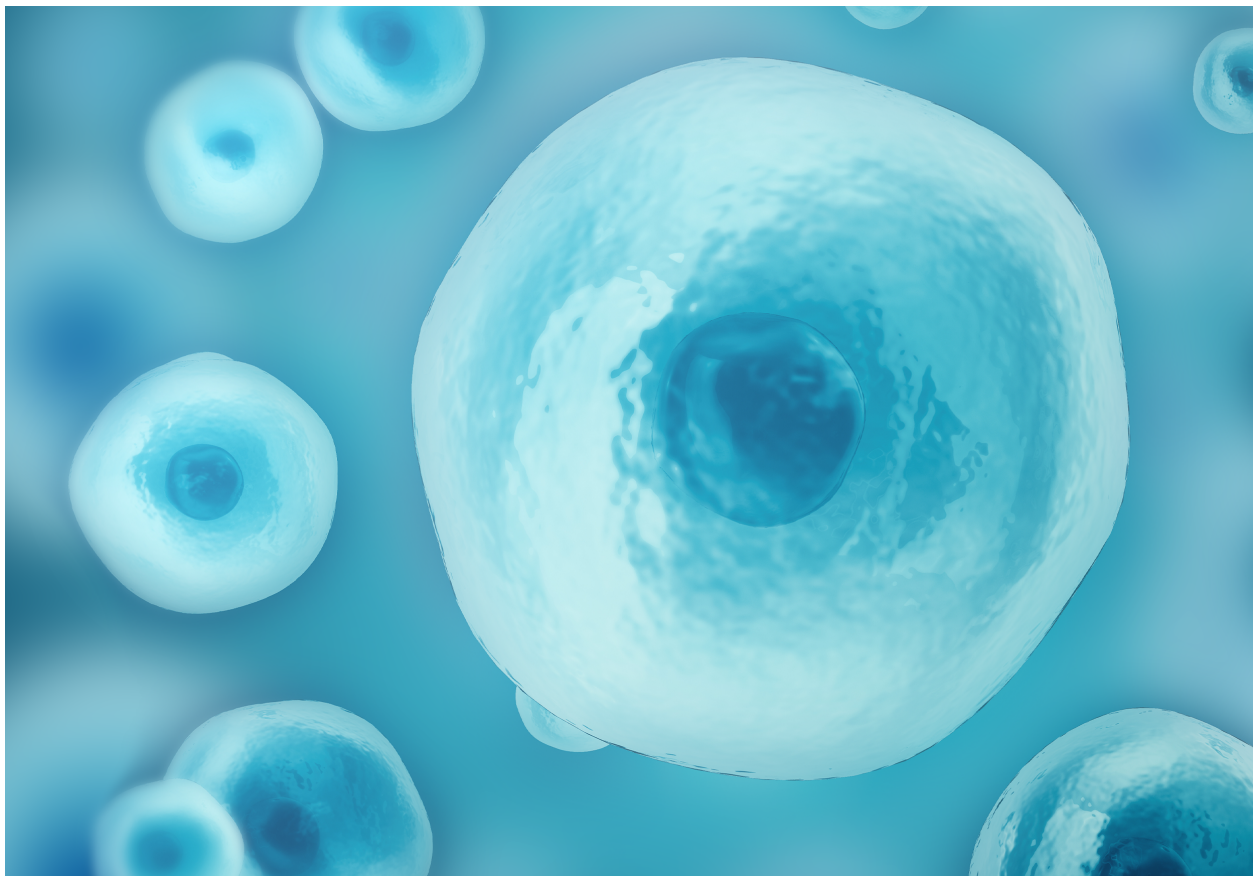# Single-Cell Sequencing Workflow: Critical Steps and Considerations

Explore every step of the single-cell sequencing workflow and learn valuable insights to ensure experimental success.

Learn more about single-cell sequencing

Table of Contents

1 Introduction

Extensive microscopic study by Anton van Leeuwenhoek and Robert Hooke in the mid seventeenth century resulted in the discovery of cells in 1665. This pioneering work eventually led to establishment of the scientific discipline of cellular biology and development of the cell theory in 1839. This historic scientific theory states that living organisms are composed of one or more cells, the cell is the basic unit of structure and organization of living organisms, and all cells arise from preexisting cells. Major advances in cellular and molecular biology, genetics, and other fields have revealed the highly complex composition of multicellular organisms and have enabled the study of biology at the resolution of its fundamental unit.
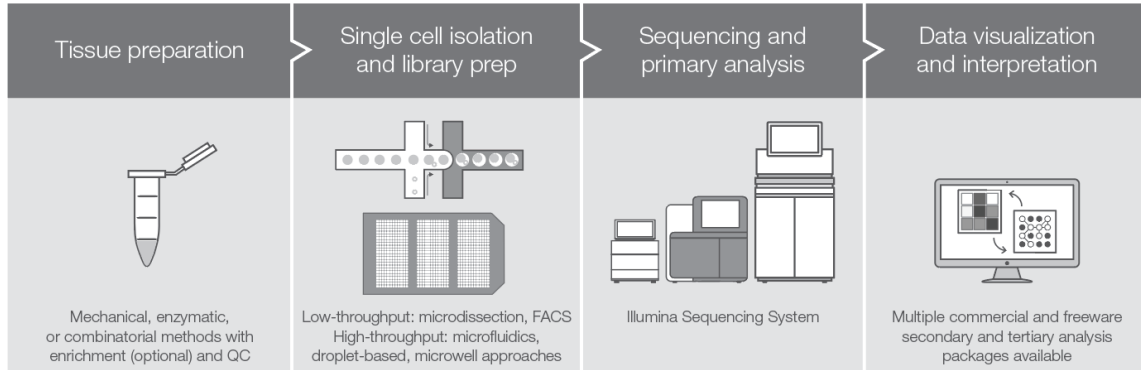
Living tissues are made up of an extensive variety of cell types, each with a distinct lineage and unique function that contribute to tissue and organ biology, and, ultimately, define the biology of the organism as a whole. The lineage and developmental stage of each cell determine how they respond to other cells and to their microenvironment. In addition, subpopulations of cells of the same type are often genetically heterogeneous from each other as well as from other cell types due to stochastic changes over time. Due to this complexity, gaining insights into cellular function through bulk analyses of tissues or cells presents significant challenges, highlighting the need to isolate individual cells for characterization.[1]

Various methods have been developed for isolation and analysis of single cells, but the invention of fluorescence activated cell sorting (FACS) in the late 1960s was a significant breakthrough that has impacted hematology, immunology, cancer research, and more.[2] This technology provides qualitative and quantitative measurement of cellular characteristics such as size, internal complexity, DNA/RNA content, and a wide range of membrane-bound and intracellular proteins (via detection of autofluorescence or fluorochrome-conjugated antibodies) and enables isolation of cells based on differential expression patterns. Isolated cells can be input into downstream analyses, *in vitro* culture experiments, or *in vivo* transplantation studies.

In addition to its widespread adoption in bulk analysis, quantitative PCR (qPCR) has been a preferred method for downstream analyses of single cells, given its wide dynamic range, familiar workflow, and lack of need for specialized instrumentation.[3] However, qPCR can only interrogate a small number of targets with known sequences, and the workflow can be cumbersome for large numbers of samples. The high accuracy and specificity of next-generation sequencing (NGS) technology makes it ideal for single-cell sequencing. NGS provides higher discovery power to detect novel genes, without prior knowledge of sequence information, and higher sensitivity to quantify rare variants and transcripts, making it the preferred method of single-cell analysis over qPCR, especially for higher throughput studies. Illumina has a large installed base of sequencing systems around the world, with a robust third-party ecosystem around our products that support an expanding offering of single-cell solutions to meet your research goals.

The single-cell sequencing workflow

The single-cell sequencing workflow includes four crucial steps: 1) initial tissue preparation, 2) single-cell isolation and library preparation, 3) sequencing and primary analysis, and 4) data visualization and interpretation (Figure 1). There are experimental considerations and critical steps throughout the workflow that can impact results and determine the success of a study. A well-planned and executed experiment is important to ensure accurate data and draw insightful conclusions.[4]

Tissue preparation	Single cell isolation and library prep	Sequencing and primary analysis	Data visualization and interpretation
Mechanical, enzymatic, or combinatorial methods with enrichment (optional) and QC	Low-throughput: microdissection, FACS High-throughput: microfluidics, droplet-based, microwell approaches	Illumina Sequencing System	Multiple commercial and freeware secondary and tertiary analysis packages available

Figure 1: The single-cell sequencing workflow — A single-cell sequencing workflow proceeds from initial tissue preparation through single cell isolation and library prep, sequencing and primary analysis, and data visualization and interpretation.

2 Step 1: Tissue preparation

Introduction

Most single-cell isolation platforms require a viable, monodispersed sample prior to compartmentalization or fixation. Type of tissue, species, and age of animal can all influence isolating live single-cells from tissues. This chapter presents some of the key considerations in the preparation of viable single-cell suspensions.

Dissociation

The process of single-cell preparation is a significant source of variability in any single-cell study.[1] Samples in which cells are adhering in clumps or have high rates of cell death can confound data and lead to misinterpretation. Nonadherent cells, such as peripheral blood mononuclear cells, are often more amenable to single-cell processing than adherent cells or cells isolated from tissue. Tissues can vary significantly in extracellular matrix (ECM) composition and cellularity, and protocols should be optimized for a specific tissue of interest.[2] Conventional protocols for tissue dissociation include mechanical dissection, enzymatic ECM breakdown, and combinatorial protocols (Table 1).

Table 1: Tissue dissociation protocols

Method	Description	Example protocol/provider
Mechanical	Tissue is mechanically sheared and disrupted through cutting, dicing, pipetting, etc	Isolation of various hematopoietic lineages from bone marrow, spleen, or lymph nodes
Enzymatic	Tissues are incubated with various enzymes such as collagenase, trypsin, dispase, elastase, etc to cleave protein bonds	Worthington Biochemical Corporation
Combinatorial	Mechanical and enzymatic methods can be performed sequentially or simultaneously, with the aid of automated systems, for more extensive dissociation	Miltenyi gentleMACS

Enrichment

Enrichment of specific cell populations, or removal of unwanted cell populations, including dead cells, is an optional, but often critical, step in single-cell preparation, especially when working with rare cells or precious samples. Various methods are available that require optimization for each specific tissue type (Table 2). Manual isolation of cells based on size, shape, and density can be achieved through differential/density gradient centrifugation and filtration. For example, mononuclear cells can be isolated from peripheral blood or bone marrow by centrifugation through various density gradient media.[5] Various fluorescent dyes are available to label and separate live cells from dead or apoptotic cells using FACS (Table 3). For enrichment of cell subpopulations/rare cell types, antibody labeling for positive/negative selection can be combined with FACS or magnetic bead-based isolation. Ultimately, the method chosen will be driven by a combination of factors, including sample type, antibody availability, and experimental design.

Table 2: Enrichment methods

Method	Description	Available protocol/provider
Centrifugation	Cell populations of interest are enriched based on size, shape, or density by centrifugation through a density gradient medium	Sigma-Aldrich
Bead-based enrichment	Cell populations of interest (including live cells) are enriched by positive/negative selection with magnetic bead–conjugated antibodies	Miltenyi Biotec
FACS	Cell populations of interest (including live cells) are enriched by positive/negative selection with fluorophores/fluorochrome-conjugated antibodies	Beckman Coulter Becton Dickinson BioLegend Bio-Rad
Microfluidic cell sorting	Cell populations of interest are enriched using low-pressure microfluidics based on positive and negative selection with fluorophores/fluorochrome-conjucated antibodies	NanoCellect

Table 3: Reagents for separating live and dead cells

Reagent	Mechanism	Pros	Cons
Classic DNA dyes	Membrane impermeant dyes (eg, PI, 7-AAD) that bind DNA will be excluded by live cells	Inexpensive, easy to use	Not compatible with intracellular staining
Amine dyes	Membrane impermeant dyes that bind amine groups of proteins will be excluded by live cells	Compatible with intracellular staining, wide selection of dyes available	More expensive than other dyes, labeling must be done in absence of free protein
Vital dyes	Membrane permeable dye that becomes fluorescent only when cleaved by metabolically active (live) cells	Inexpensive, easy to use	Challenging to use with intracellular staining

Abbreviations: PI, propidium iodide; 7-AAD, 7-aminoactinomycin D

Learn more at expertcytometry.com/3-reagents-for-identifying-live-dead-and-apoptotic-cells-by-flow-cytometry/

Quality control

Single-cell sequencing experiments represent a significant investment of time, money, sample material, and resources. Several simple quality control (QC) measures throughout can ensure a high-quality experiment before proceeding with cell isolation, library preparation, and sequencing.

Visual inspection

Visual inspection of the cell suspension under a microscope is valuable as it enables quick identification of debris, cell doublets, and cell aggregates that can complicate downstream steps (Figure 2). Importantly, accurate cell counts are critical to achieve target cell throughputs in subsequent single-cell isolation procedures. Cell counts can be determined manually by combining microscopy with a hemocytometer. Automated systems are also available that provide accurate cell counts, capture brightfield images of the cell suspension, and generate histograms for more detailed inspection based on cell characteristics such as size, brightness, and circularity. Examples of commercially available automated cell counters include the Countess II Automated Cell Counter (Thermo Fisher Scientific), the TC20 Automated Cell Counter (Bio-Rad), and the Auto 1000 Bright Field Cell Counter (Nexcelom Bioscience).

Cell aggregates   < 10% doublets
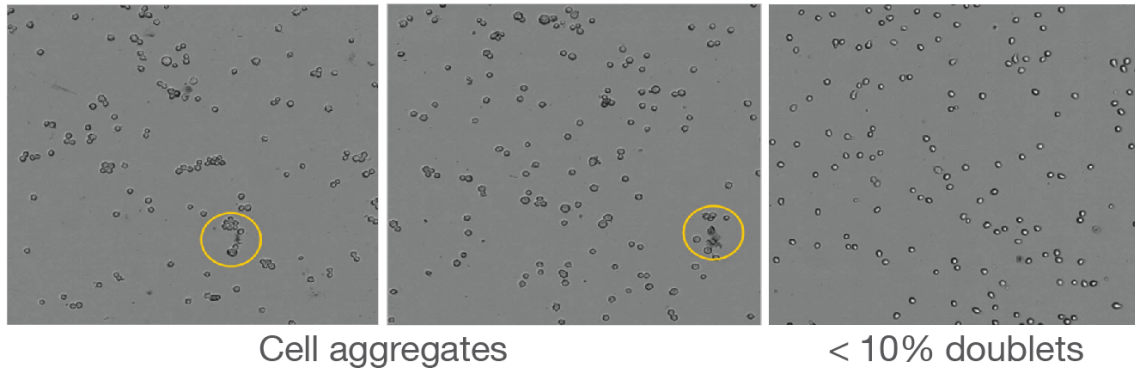
Figure 2: Visual inspection of cell suspensions — Visual inspection of prepared cell suspensions after tissue dissociation by brightfield microscopy reveals debris, cell doublets, and larger cell aggregates (yellow circles) present in the samples. A sample with < 10% doublets is shown (right).

Flow cytometry

Flow cytometry is a valuable QC tool, as multiple metrics can be assessed simultaneously, including cell size, viability, and presence of doublets or aggregates. Also, antibody labeling can be included as part of the analysis to evaluate whether cell populations of interest are present and maintained at the appropriate frequency.

Key metrics

Several key QC metrics that indicate successful preparation of a monodispersed cell suspension include:
- **Cell viability**: Dead or damaged cells can release nucleic acids into cell suspension that remain through subsequent steps, possibly impacting results. Cell viability levels > 85% are recommended.
- **Cell size distribution**: Histogram plots can be inspected for presence of multiple peaks indicating cellular fragments (smaller peaks), doublets/aggregates (peak at twice nominal cell size), or large debris (larger peaks).
- **Cell concentration**: The ideal loading concentration for cells depends on the isolation method. Optimal concentration is critical as underloading or overloading can cause issues with single-cell isolation or data quality.

Summary

Harnessing the potential of NGS to investigate complex biological systems at the level of individual cells requires that tissues are properly dissociated to monodispersed suspensions of viable cells. A wide selection of methods is available, and the specific protocol should be selected and optimized based on the tissue of interest. Consideration should be given to the inclusion of an enrichment step and key QC metrics to ensure a high yield of single cells while maintaining viability. After a tissue preparation protocol has been optimized, researchers can proceed with confidence to single-cell isolation and library preparation.

Tissue preparation

3 Step 2: Single-cell isolation and library preparation

Introduction

Rapid advances in both methods and commercial products available to profile RNA, DNA and protein in the same cell resulted in single-cell sequencing being named Method of the Year for a second time.[6] Ultimately, the method chosen to capture and isolate single cells will determine library preparation, sequencing, and downstream analyses. This chapter discusses options available for single cell isolation and highlights techniques used for global characterization of isolated cells.

Cell isolation methods and platforms

Cell isolation methods can be distinguished by throughput. Low-throughput methods include mechanical manipulation or cell sorting/partitioning technologies (eg, FACS) and are able to process dozens to hundreds to a few thousand cells per experiment (Table 4). Advances in microfluidic technologies have enabled high-throughput single cell profiling where researchers can examine hundreds to tens of thousands of cells per experiment in a cost-effective manner (Table 5).[7]

Table 4: Low-throughput single-cell isolation approaches

Method	Description	Advantages	Disadvantages	Commercial offering/ Example methods
Serial dilution	Serial dilution of cell suspension down to one cell per well	Simple approach; does not require specialized equipment	Time-consuming, probability of isolating multiple cells	Corning Serial Dilution Protocol
Mouth pipetting	Isolation of single cells with glass pipettes	Simple approach	Technically difficult, random	N/A
Robotic micromanipulation	Isolation of single cells with robotic micropipettes	Positional placement of cells	Requires specialized equipment	An automated system for high-throughput single cell-based breeding. Single cell deposition and patterning with a robotic system.
Laser capture microdissection	Dissection of single cells from tissue sections using a laser	Spatial context is preserved	Technically challenging, potential UV damage to DNA/RNA	Laser capture microdissection of single cells from complex tissues.
FACS	Isolation of microdroplets containing single cells using electric charge	Accurate selection of cell types by size, morphology, internal complexity, and protein expression by antibody labeling	Requires expensive, specialized equipment, cells exposed to high pressure	Beckman Coulter Becton Dickinson Bio-Rad
Microfluidics circuits	Microfluidic chips isolate cells in flow channels	Highly sensitive chemistry, compatible with small volumes, flexible with custom reagents	Requires uniform cell size, expensive consumables	Fluidigm C1 System Fluidigm Polaris System

Table 5: High-throughput single-cell isolation approaches

Method	Description	Advantages	Disadvantages	Commercial offering/ Example methods
Droplet fluidics platforms[8-11]	Compartmentalization of individual cells in droplets using a microfluidics device followed by lysis and capture of target DNA/RNA	Unique molecular identifiers (UMIs) and cell barcodes enable cell and gene-specific identification, low cost per cell, wide menu of commercial applications	Requires specialized equipment, can be technically challenging	1CellBio inDrop System 10X Genomics Chromium Controller Bio-Rad ddSEQ Single-Cell Isolator Instrument Dolomite Bio Nadia Instrument Mission Bio Tapestri Platform
Microwells [12,13]	Capture of individual cells in microwells of fabricated arrays	Supports imaging and short-term culture of cells, ideal for adherent cells, UMIs enable cell and gene-specific identification	Requires specialized equipment, can be technically challenging	BD Rhapsody Single-Cell Analysis System CellMicrosystems CellRaft AIR System Bio-Rad Celsee Genesis System Takara ICELL 8 cx Single-Cell System
Combinatorial indexing[14-16]	Intact nuclei are tagged with unique barcodes via two rounds of random distribution into microwells and labeling via transposases	Low-cost approach to profile large number of cells and compatible with multiomic methods, no specialized equipment available	Limited commercial solutions	Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing.

Library preparation

The next critical step in the single-cell sequencing workflow is library preparation. The cell profiling approach and specific sequencing method chosen are important considerations, as various options are available (Table 6). The particular method chosen will largely be determined by the experimental question.

Cell isolation and library prep

Table 6: Amplification techniques for single-cell profiling

Method	Description	Commercial offering/Example method
Transcriptome		
Full-length RNA-Seq	Switching Mechanism at 5' end of RNA Template (SMART) technology enables amplification of full-length cDNA	Takara SMARTer cDNA Synthesis Kits
mRNA end-tag amplification (3' WTA or 5' WTA)	Capture of mRNA by 3' polyadenylated (poly(A)) tails enables sequencing of the coding transcriptome with strand-specific information	10X Genomics Chromium Single Cell Gene Expression Solution (3' WTA) 10X Genomics Chromium Single Cell Immune Profiling Solution (5' WTA) SureCell WTA 3' Library Prep Kit for the ddSEQ System
Targeted panels	Various predesigned single-cell targeted RNA sequencing panels enable IR, T-cell, and breast cancer profiling, and more.	BD Rhapsody Single-Cell Analysis
IR-Seq	Immune repertoire sequencing (IR-Seq) is a targeted sequencing method used to quantify the composition of B or T-cell antigen receptor repertoires.	10X Genomics Chromium Single Cell Immune Profiling Solution
Genome		
MALBAC	Multiple Annealing and Looping Based Amplification Cycle (MALBAC) enables quasilinear amplification of the whole genome from single cells.	Single cell transcriptome amplification with MALBAC Yikon Genomics
DOP-PCR	Degenerate Oligonucleotide-Primed PCR (DOP-PCR) uses oligos of partially degenerate sequence for whole genome amplification.	Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer Whole-genome amplification by degenerate oligonucleotide primed PCR (DOP-PCR)
Targeted Panels	Various predesigned single-cell targeted DNA sequencing panels enable profiling of hematologic malignancies, solid tumors, copy number variation (CNV), and more.	10X Genomics Chromium Single Cell CNV Solution MissionBio Tapestri Designer for Custom Single-Cell DNA Panels Mission Bio Tapestri Single-Cell DNA Panels
Epigenome		
ATAC-Seq	Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-Seq) assesses chromatin accessibility genomewide by using a transposase to insert sequencing adapters into regions of open chromatin.	10X Genomics Chromium Single Cell ATAC Solution Abcam ATAC-Seq protocol Bio-Rad SureCell ATAC-Seq Library Prep Kit
HiC	HiC combines chromosome conformation capture (3C) with NGS to enable unbiased identification of chromatin interactions across the genome.	Hi-C: a comprehensive technique to capture the conformation of genomes. Comprehensive mapping of long-range interactions reveals folding principles of the human genome.
Protein detection		
AbSeq	DNA-tagged antibodies enable protein profiling by NGS	BD Abseq antibody-oligonucleotide conjugates
CITE-Seq	Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) uses oligonucleotide-labeled antibodies to convert protein detection into a quantitative assay by NGS	Simultaneous epitope and transcriptome measurement in single cells. cite-seq.com

Cell isolation and library prep

QC of Prepared Libraries

Accurate assessment of both quality and quantity of prepared libraries is important to maximize sequencing data quality and output. The Fragment Analyzer (Agilent Technologies) is a proven solution for simultaneous qualification and quantification of DNA and RNA during library preparation for Illumina sequencing workflows (Figure 3).[17] The Bioanalyzer (Agilent Technologies) is another option for library QC. Both instruments have advantages over traditional methods such as qPCR including: accurate and sensitive quantitation of DNA/RNA, size measurement of fragments, and detection of possible contaminants.[18]

Regardless of the method chosen for library quantification and quality assessment, only high-quality libraries should be subject to sequencing to ensure generation of reliable, high-quality data. Calculation of the Genomic Quality Number (GQN) can assess genomic DNA samples as they relate to a user-defined, application-specific threshold for "good quality DNA". Similarly, the RNA Quality Number (RQN), or the equivalent RNA Integrity Number (RIN), are two metrics broadly accepted for assessing quality of RNA samples.[17] After libraries have been assessed for quality and quantified, the appropriate amount can be loaded for sequencing, which depends on the sequencing platform and flow cell.
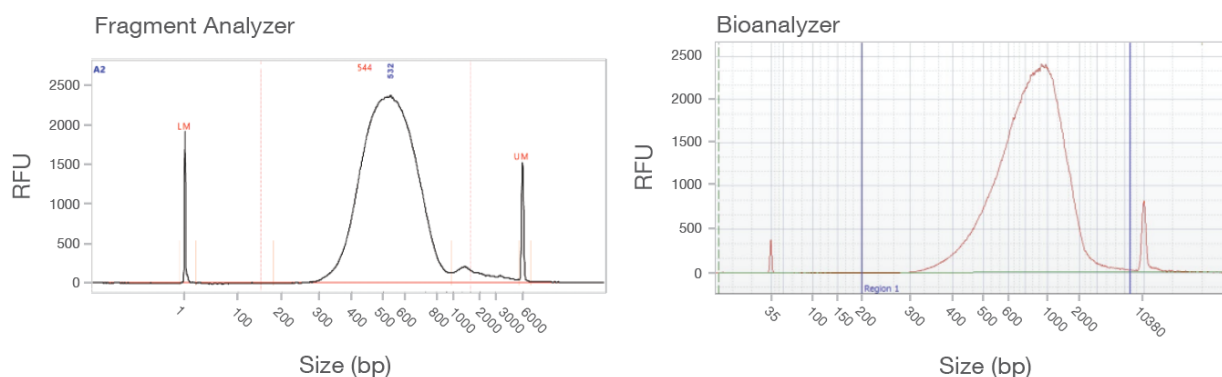
Figure 3: Library QC—Library QC traces on the Fragment Analyzer (left) and Bioanalyzer (right) showing high-quality sequencing libraries. RFU = relative fluorescence units.

Summary

A critical step in the single-cell sequencing workflow is isolation of individual cells. A wide selection of methods is available, and the specific protocol should be selected and optimized based on the experimental question. Consideration should be given to the inclusion of QC measurement of prepared libraries. After high-quality single-cell libraries have been prepared, researchers can proceed with confidence to sequencing. If you would like to discuss various single cell sequencing methods and how they can be integrated with your research, contact your local Illumina representative.

Cell isolation and library prep

4 Step 3: Sequencing

Introduction

After viable single cells have been isolated and the genetic material of interest has been extracted and libraries have been prepared, the crucial step of sequencing can be performed. All Illumina sequencing platforms use sequencing by synthesis (SBS) chemistry, responsible for generating more than 90% of the world's sequencing data.[19] Illumina SBS chemistry is a proprietary method that detects single bases as they are incorporated into growing DNA strands in a massively parallel fashion and has been referenced in over 250,000 independent publications. Illumina sequencing systems can deliver data output ranging from 300 kilobases up to multiple terabases in a single run, depending on instrument type and configuration. This chapter presents the Illumina sequencing systems that are appropriate for single-cell studies and discusses important considerations to ensure a successful sequencing run.

Compatible sequencing systems

Although all Illumina sequencing systems are capable of sequencing single-cell libraries, the sequencing system chosen for a single-cell sequencing experiment will be determined largely by the research question and scale of the study (Table 7, Figure 4).

Table 7: Overview of compatible sequencing systems

	NextSeq 1000 and NextSeq 2000 Systems	NextSeq 550 System	NovaSeq 6000 System	iSeq 100 System
Description	The NextSeq 1000 and NextSeq 2000 Systems offer 2.5× the throughput of the NextSeq 550 System, empowering deeper investigation, larger studies, and higher resolution . On-board informatics streamline analysis, making NGS accessible to novice and expert users.		The NovaSeq 6000 System represents the most powerful, scalable, and reliable high-throughput Illumina sequencing platform to date, producing outstanding data quality.	The iSeq 100 System provides accurate data with fast turnaround times in the smallest and most affordable Illumina sequencing system.
Configuration	Multiple flow cell types and run configurations, from 400M reads with NextSeq 1000/2000 S2 Reagents to 1B reads with NextSeq 2000 S3 Reagents.	Scales from 130M reads with Mid-Output kits up to 400M reads with High-Output kits.	Multiple flow cell types and run configurations, from 800M reads with the SP flow cell to 10B reads with the S4 flow cell.	4M reads per flow cell
Advantages	Power of high-throughput sequencing with the simplicity and affordability of a benchtop system		Unprecedented output and throughput	Speed and smaller scale
Ideal for	Mid- to high-throughput sequencing applications and average scale single-cell sequencing studies, such as studies to profile cell function in both development and disease.		Extensive screening studies, such as pharmaceutical screens and cell atlas studies.	Library QC by sequencing single-cell libraries at shallow depths to characterize key metrics before a high-depth NGS run.
Learn more	NextSeq 2000 System	NextSeq 550 System	NovaSeq 6000 System	iSeq 100 System

| iSeq 100 System | NextSeq 550 System | NextSeq 1000 and NextSeq 2000 System | NovaSeq 6000 System |

Figure 4: Compatible Illumina sequencing systems for single-cell sequencing — Illumina NGS systems deliver high-accuracy data with flexible throughput and simple, streamlined workflows compatible with single-cell sequencing experiments of any scale.

Considerations for sequencing

Experiment planning

Read depth

Sequencing coverage for traditional or bulk samples describes the average number of reads that align to, or "cover," known reference bases. NGS coverage level often determines whether variant discovery can be made with a certain degree of confidence at particular base positions. Sequencing coverage requirements vary by application (Table 8). At higher levels of coverage, each base is covered by a greater number of aligned sequence reads, or a greater "depth," so base calls can be made with a higher degree of confidence.[20]

Table 8: Recommended reads for different single-cell sequencing applications

| Method	Recommended no. of reads[a]
3′ gene expression	15K–50K reads per cell
5′ gene expression	50K reads per cell
Antibody sequencing	100 reads per antibody/cell
scATAC-Seq	50K reads per nuclei
5′ TCR/BCR	5K reads per cell
Takara SMARTer	1M–2M reads per cell (> 300,000 reads per cell)

The recommended number of reads is based upon manufacturer recommendations (see Table 6).

For various single-cell sequencing applications, read depth is discussed not in the number of reads per base, but in the number of reads per cell. The required sequencing depth for a single-cell sequencing run will depend on several factors, including sample type, the number of cells to be analyzed, experimental objectives, and more. For single-cell RNA-Seq, it has been reported that unbiased cell-type classification within a mixed population of distinct cell types can be achieved with as few as 10,000 to 50,000 reads per cell.[21] Such lower read depth can be practical and economical if the experimental objective is to identify rare cell populations or to scan cells for presence of mixed populations. However, this read depth may not be

Sequencing

sufficient when more homogeneous cell populations are studied, and it is unlikely to provide detailed information on gene expression within any given cell. In such cases deeper sequencing may be required for improving cell identification and detection of genes with low expression. Indeed, it has been reported that 500,000 reads per cell are sufficient to detect most genes expressed in a cell, and 1,000,000 reads per cell approaches sequencing saturation, enabling the estimation of the mean and variance of gene expression.[22,23] Ultimately, the required sequencing depth will largely depend sample type and experimental objective and will need to be optimized for each study.

Paired-end vs. single-read sequencing

Single-read sequencing involves sequencing DNA from only one end and is the simplest way to utilize Illumina sequencing. Single-read sequencing delivers large volumes of high-quality data, faster and cheaper than paired-end sequencing.[24] Single-read runs can be a good choice for certain methods such as small RNA-Seq or chromatin immunoprecipitation sequencing (ChIP-Seq). In contrast, paired-end sequencing involves sequencing both ends of DNA fragments in a library and aligning the forward and reverse reads as read pairs. This results in better alignment of reads, especially across repetitive, difficult-to-sequence regions. All Illumina NGS systems are capable of paired-end sequencing (Figure 5).
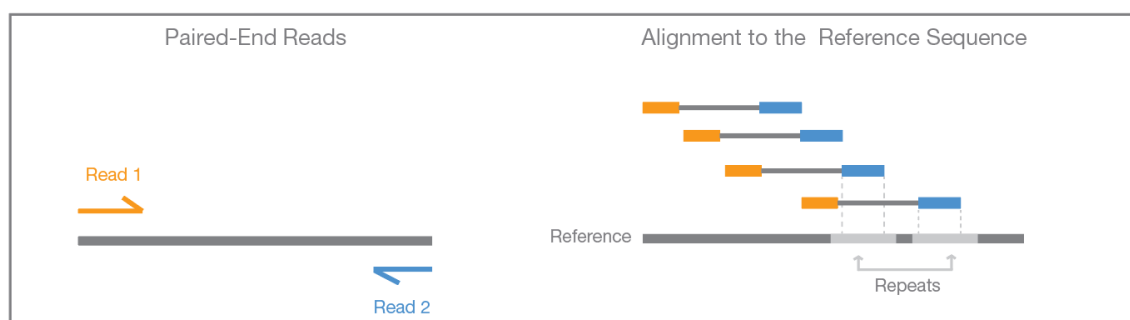

Figure 5: Paired-end sequencing and alignment — Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely than single-read sequencing.

In addition to producing twice the number of reads for the same time and effort in library preparation, sequences aligned as read pairs enable detection of insertion-deletion (indel) variants, which is not possible with single-read data.[25] Furthermore, paired-end sequencing facilitates detection of genomic rearrangements such as insertions, deletions, and inversions. Paired-end RNA sequencing enables discovery applications such as detecting gene fusions, novel transcripts, and novel splice isoforms.[26]

Cluster density

The massively parallel nature of Illumina sequencing is enabled by cluster generation on the surface of flow cells. Historically, during cluster generation, adapter-ligated library elements hybridized to complementary oligonucleotides on the surface of a flow cell. Each attached library fragment acted as a "seed" and, through a process called bridge amplification, was amplified to generate a clonal cluster containing thousands of identical fragments (Figure 6). After cluster generation was complete, the flow cell contained millions to billions of clusters on its surface.

Sequencing

Cluster Amplification
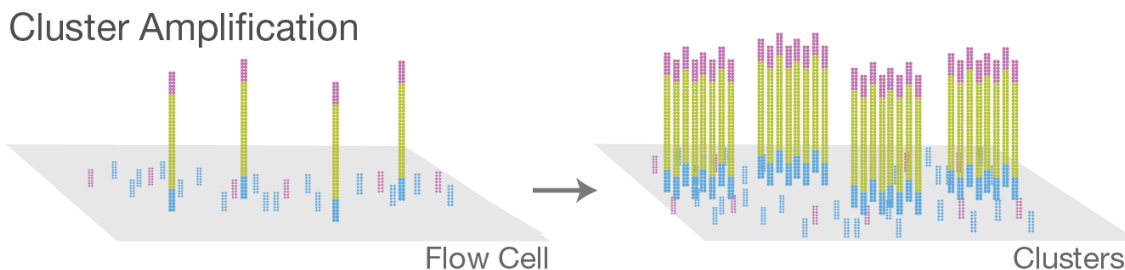
Flow Cell → Clusters

Figure 6: Cluster generation—Library fragments are loaded onto a flow cell and hybridize to the flow cell surface. Each bound fragments is amplified into a clonal cluster through bridge amplification.

Ideally, clusters are of similar size and spaced well apart from each other to achieve accurate resolution during imaging. In reality, DNA clusters are randomly distributed across these "nonpatterned" flow cells with many clusters in close proximity to neighboring clusters, especially if the sample is overloaded, making it difficult to discern individual clusters from each other. The NextSeq™ 550 System uses nonpatterned flow cells.

To make more effective use of the flow cell surface space, Illumina created patterned flow cell technology. Patterned flow cells feature patterned nanowells etched into the surface. Each nanowell contains DNA probes used to capture prepared DNA strands for amplification during cluster generation. The area between the nanowells is devoid of DNA probes. This process ensures that DNA clusters only form within the nanowells, providing even, consistent spacing between adjacent clusters and allowing accurate resolution of clusters during imaging. The result is maximal use of the flow cell surface leading to overall higher clustering.[27] The iSeq™ 100, NextSeq 1000, NextSeq 2000, and NovaSeq™ 6000 Systems use patterned flow cells.

The density of clusters on a flow cell, particularly with nonpatterned flow cells, significantly impacts data quality and yield from a run and is a critical metric for measuring sequencing performance. It influences run quality, reads passing filter, Q30 scores, and total data output. Performing a run at optimal cluster density involves finding a balance between underclustering and overclustering. The goal is to sequence at a high enough density to maximize total data output, while maintaining a low enough density to avoid overclustering. The recommended cluster density for the NextSeq 550 System is 170–220 K/mm^2.[28] Because patterned flow cells provide optimal cluster density, they are less susceptible to underclustering and overclustering. However, libraries should still be loaded at recommended concentrations for optimal performance. Most commercial single-cell library preparation kits provide cluster density recommendations for each Illumina sequencing system.

Sequencing

Run QC

Percent passing filter

Percent passing filter (%PF) is an important sequencing QC metric that refers to the number of clusters that have passed a filter and will be retained for downstream analysis. With nonpatterned flow cells, Real-Time Analysis software evaluates clusters during image analysis early in the template generation stage of the sequencing run. Any dim or low-quality clusters are removed, effectively acting as a prefiltration step, resulting in relatively high %PF values. With patterned flow cells, fixed cluster locations eliminate the need for template generation, so there is no prefiltration of underperforming clusters. Instead, suboptimal clusters are filtered out during the later stage of chastity filtration. Chastity is defined as the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities. Clusters "pass filter" if no more than one base call has a chastity value below 0.6 in the first 25 sequencing cycles. This filtration process removes the least reliable clusters from the image analysis results. Consequently, for patterned flow cells the %PF metric will be lower (than that for nonpatterned flow cells), but it will not affect performance or data quality (Figure 7).[29]
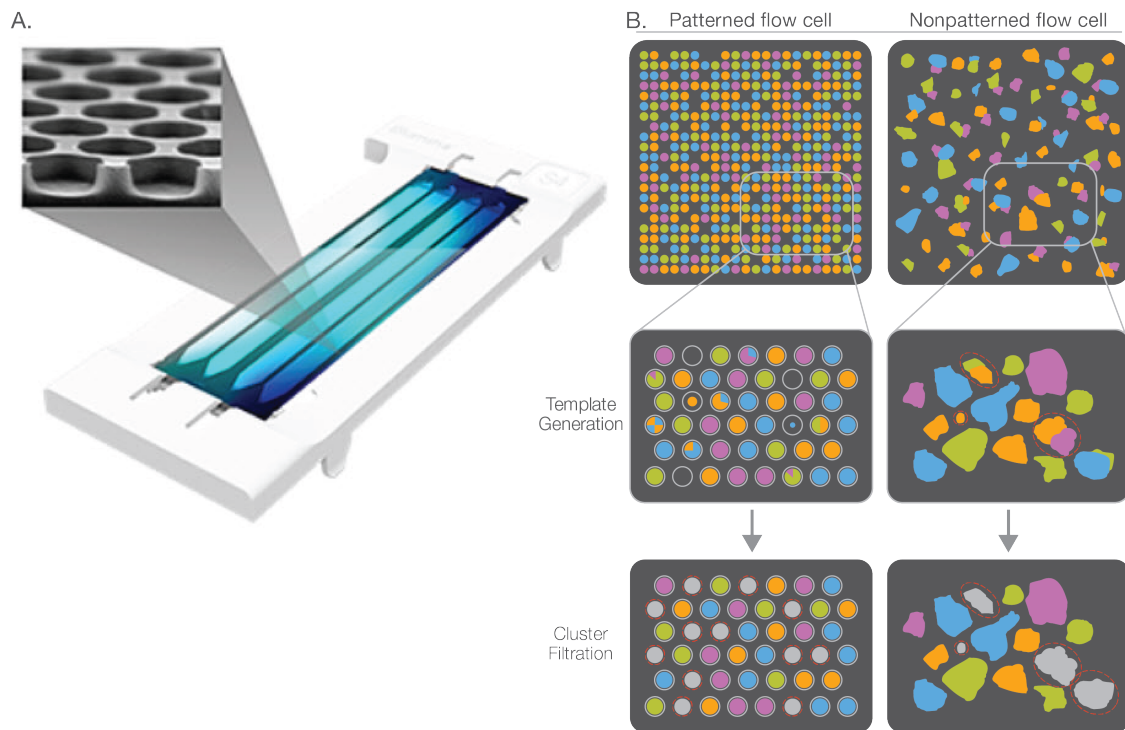
Figure 7: Clusters passing filter on patterned and nonpatterned flow cells — (A) A patterned flow cell with nanowells etched into its surface (B). With nonpatterned flow cells, poor quality or dim clusters are filtered during template generation. With patterned flow cells, empty wells and suboptimal clusters are filtered during the later stage of chastity filtration, which leads to a lower %PF metric (B).

Sequencing

Percent ≥ Q30

Sequencing quality scores measure the probability that a base is called incorrectly. With SBS chemistry, each base in a read is assigned a quality score by a phred-like algorithm,[31,32] similar to that originally developed for Sanger sequencing experiments. The sequencing quality score of a given base, Q, is defined by the following equation:

$$Q = -10\log_{10}(e)$$

Where e is the estimated probability of the base call being wrong. A higher Q-score indicates a smaller probability of error. Illumina SBS chemistry delivers high accuracy, with a vast majority of bases scoring Q30 and above (% ≥ Q30). However, the Q-score may not be the most appropriate metric for assessing sequencing results for single-cell libraries, as significant variability in Q30 scores can be observed due to differences in library chemistry, barcode design, and sample preparation. Most commercial single-cell library preparation kits provide guidance on key metrics to assess a high-quality experiment including Q30 scores, valid barcodes, estimated number of cells, fractions of reads in cells, and total genes detected.[33,34]

For more information on sequencing quality scores, read the following technical notes:
- Quality Scores for Next-Generation Sequencing
- Understanding Illumina Quality Scores

Instrument control software

Instrument control software is preinstalled on all Illumina sequencing systems. Control software guides users through the steps to load the flow cell and reagents, and provides an overview of quality statistics for monitoring a sequencing run as it progresses. The software can also perform image analysis, base calling, and base call quality automatically.
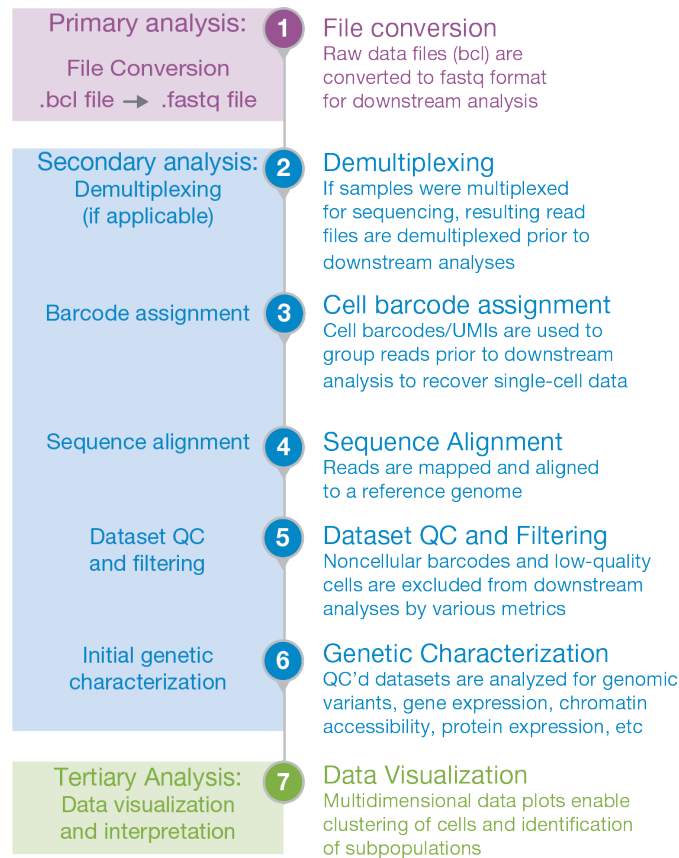
Summary

Illumina sequencing systems offer high data accuracy with flexible throughput to deliver a proven NGS solution for single-cell sequencing studies, regardless of scale. It is important to consider read depth, and whether single-read or paired-end sequencing is required before committing to a sequencing run. This will help to balance cost with sequencing parameters best suited to meet experimental objectives. Additional sequencing metrics, such as cluster density, %PF, and % ≥ Q30 (and alternative metrics), should be considered before, and evaluated after, sequencing is performed to help ensure successful results. After high-quality, reliable sequencing data are obtained, researchers can proceed with data visualization, analysis, and interpretation.

Sequencing

5

Step 4: Data analysis, visualization, and interpretation

Introduction

After the single-cell sequencing run is complete, downstream analysis can be performed. Generally, the analysis pipeline for single-cell sequencing experiments involves three phases: primary analysis (base calling), secondary analysis (demultiplexing, alignment, and genetic characterization), and tertiary analysis (data visualization and interpretation) (Figure 8). There is no one, correct way to carry out an analysis pipeline for single-cell sequencing experiments. Many approaches and software programs are available for each step in the pipeline.[30] The research objective, single-cell isolation platform, and general lab considerations will largely determine the specific pipeline used. This chapter outlines the steps involved in single-cell sequencing analysis and some of the tools available.

Primary analysis:
File Conversion
.bcl file → .fastq file

1 File conversion
Raw data files (bcl) are converted to fastq format for downstream analysis

Secondary analysis:
Demultiplexing
(if applicable)

2 Demultiplexing
If samples were multiplexed for sequencing, resulting read files are demultiplexed prior to downstream analyses

Barcode assignment

3 Cell barcode assignment
Cell barcodes/UMIs are used to group reads prior to downstream analysis to recover single-cell data

Sequence alignment

4 Sequence Alignment
Reads are mapped and aligned to a reference genome

Dataset QC and filtering

5 Dataset QC and Filtering
Noncellular barcodes and low-quality cells are excluded from downstream analyses by various metrics

Initial genetic characterization

6 Genetic Characterization
QC'd datasets are analyzed for genomic variants, gene expression, chromatin accessibility, protein expression, etc

Tertiary Analysis:
Data visualization and interpretation

7 Data Visualization
Multidimensional data plots enable clustering of cells and identification of subpopulations

Figure 8: Example single-cell sequencing analysis pipeline—An example of an analysis pipeline for single-cell sequencing experiments from initial file conversion through primary, secondary, and tertiary analysis.

Data analysis

Primary analysis: file conversion

*.bcl file format

Illumina sequencing systems generate raw data files in binary base call (BCL) format. This sequencing file format contains both the base call and the quality of that base call for each cluster on a per-cycle basis. While the BCL file format is efficient for the sequencing system, it requires conversion to FASTQ format for use with user-developed or third-party data analysis tools.

*.fastq file format

FASTQ is a text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems, and can be used as input for a wide variety of secondary data analysis solutions.

bcl2fastq conversion software

bcl2fastq software coverts BCL files to FASTQ files for downstream analysis. This process can begin this process as soon as the first read has been completely sequenced. If samples were multiplexed, the first step in FASTQ file generation is demultiplexing. Multiplexed sequencing enables multiple individual samples to be run in a single lane of a flow cell, greatly increasing a systems output. Demultiplexing assigns clusters to a sample, based on the cluster's index sequence. After demultiplexing, the assembled sequences are written to FASTQ files per sample. If samples were not multiplexed, the demultiplexing step does not occur, and, all clusters on the flow cell are assigned to a single sample.[35]

Secondary analysis: demultiplexing, alignment, and QC

Single-cell sequencing data can be instantly transferred, stored, and analyzed securely in BaseSpace™ Sequence Hub, the Illumina cloud-based genomics computing environment. BaseSpace Sequence Hub provides a large collection of BaseSpace Apps. These Apps feature intuitive push-button user interfaces designed to be used without the need for bioinformatics expertise. Commercial and open-source tools support a range of common data analysis needs such as alignment, variant calling, and more.

Read mapping and alignment to a reference genome is often the first step in data analysis. Various software applications are available, including the Burrows-Wheeler Alignment (BWA)[36] algorithm, used in the BWA Aligner BaseSpace App, and Spliced Transcripts Alignment to a Reference (STAR)[37] algorithm, included in the RNA-Seq Alignment BaseSpace App (Table 9).

Data analysis

Table 9: Primary and secondary analysis BaseSpace Apps

BaseSpace App	Description
BWA Aligner	The BWA Aligner App aligns samples (consisting of FASTQ files) using the BWA-MEM aligner to a reference genome, including a custom reference genome created from imported FASTA files.
RNA-Seq Alignment	The RNA-Seq Alignment workflow performs the following: read mapping using the STAR aligner, quantification of reference genes and transcripts using salmon, variant calling (SNVs and small indels) using the Strelka Variant caller, fusion calling with Manta, and QC metrics from Picard and other sources.
SureCell RNA Single-Cell	The SureCell RNA Single Cell app is designed to analyze samples prepared using the SureCell Whole Transcriptome Analysis 3' Library Preparation kit. This app performs cell and gene counting, filtering, and calculates and reports metrics for the Illumina Bio-Rad Single-Cell Sequencing Solution.

Various BaseSpace applications and third-party programs are available for cell identification and counting and mapping of the genetic component of interest, including:

- **RNA-Seq**: cell barcodes and UMIs (if used) are demultiplexed to build a matrix of genes expressed in each cell
- **ATAC-Seq**: adapter sequences inserted into areas of open chromatin are identified in each cell
- **CITE-Seq**: antibody, UMI, and cell barcodes are demultiplexed to map protein expression in each cell
- **Targeted DNA sequencing**: cell barcodes are demultiplexed to build a matrix of genomic variants by each cell

For RNA-Seq, the SureCell RNA Single-Cell App supports data analysis for the Illumina Bio-Rad Single-Cell Sequencing Solution. The SureCell RNA Single-Cell App enables streamlined data analysis and includes sequencing QC metrics, assignment of unique transcripts to single cells, and options for identifying subpopulations and differentially expressed genes.

Analysis QC metrics

Before downstream analyses, several QC metrics should be performed to help determine the quality of a single-cell sequencing data set and filter out poor quality data points/cells.

Library QC with the iSeq 100 System

Accurate characterization of multiplexed library composition is a vital step in optimizing library loading, particularly for single-cell gene expression profiling (sc-GEX) libraries. Important metrics of sc-GEX libraries, such as estimated cell counts, intergenic/intronic/ exonic content, and fraction of reads in cells cannot be assessed before sequencing using traditional quantitation methods like Qubit or qPCR. To maximize the efficiency of high-throughput sc-GEX experiments, sequencing these libraries first at shallow depths on the iSeq 100 System enables characterization of key metrics and subsequent rebalancing before a high-depth NGS run. Library QC not only saves time and money, but also leads to more consistent sequencing results, which can simplify data analysis and interpretation.

🔗 Read the QC and rebalancing of single-cell gene expression libraries using the iSeq 100 System application note

Data analysis

## Expected library size and number of expressed genes

All cell types have an expected library size and a typical number of expressed genes. Cells that fall outside of that expected range (too low or too high) may represent low-quality "cells" to be excluded from downstream analyses or unusual cells of interest warranting further investigation (Figure 9).
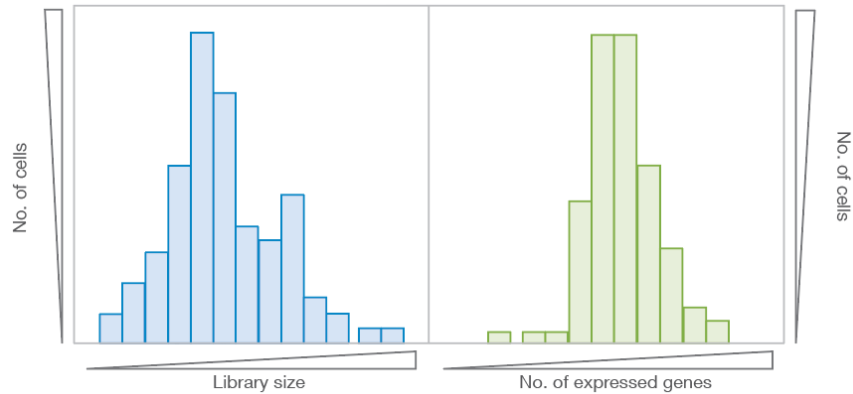
Figure 9: Filtering by library size or number of expressed genes — Cell distribution plots by library size and no. of expressed genes. Cells outside of the expected range for each cell type may be poor quality or unusual cells of interest.

## Proportion of reads aligning to mitochondria/ribosomes

Another QC metric is the proportion of reads that map to genes in the mitochondrial genome or reads that map to ribosomal RNAs (Figure 10). High mitochondrial and ribosomal proportions are indicative of poor-quality cells, possibly because of increased apoptosis, which can be excluded from downstream analyses.
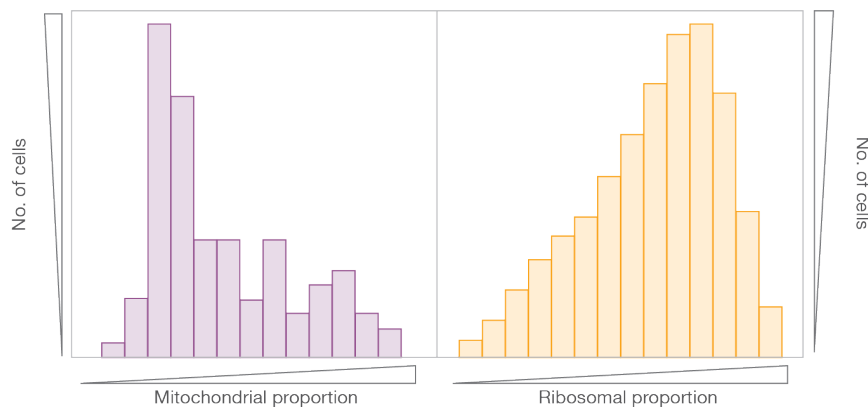
Figure 10: Filtering by mitochondrial or ribosomal proportion — Cell distribution plots by proportion of reads mapping to the mitochondrial genome and ribosomes. Cells with high proportions are likely poor quality.

Data analysis

Knee plot

Plotting genic UMI counts in descending order against cell barcodes enables statistical identification of "true" cells and exclusion of noncellular barcodes (Figure 11). Cell barcodes above the threshold (left of the knee) have genic UMI that represent true cells, while those below the threshold (to the right of the knee) have genic UMI counts below what is expected for that particular cell.
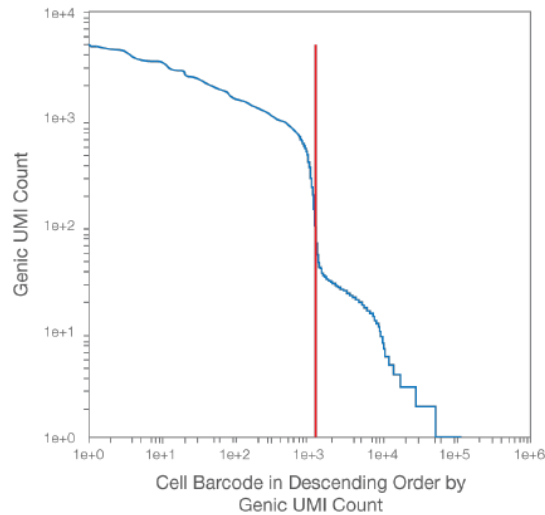

Figure 11: Filtering out noncellular barcodes — Cell barcodes to the left of the threshold (vertical red line) have genic UMI counts in the thousands, representing true cells. Cell barcodes to the right of the threshold have genic UMI counts of 1–100, typically below what is expected for live, intact cells, representing empty beads.

Evaluating doublets

Number of genes per cell

For any given cell type, there is a typical, expected number of expressed genes. Historically, this has been used to detect and exclude doublets from downstream analyses.[34] However, while using gene number per cell can be useful for single-cell sequencing experiments with a homogenous cell population, eg, cultured cell lines, it can be problematic with complex heterogeneous tissues. Indeed, while a majority of viable single cells may fall in a natural distribution around an expected number of expressed genes, (n), cells observed outside that distribution, eg, with roughly twice that number, ($2n$), may represent cells of interest that warrant further investigation and characterization, eg, circulating cancer cells in a blood sample (Figure 12). Ultimately, given the lack of a credible computational method for detecting doublets, researchers should minimize doublet rates by experimental design.[38]

Data analysis
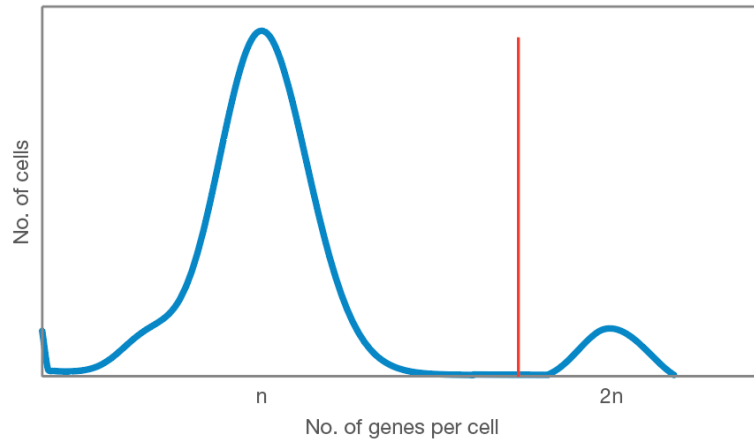
Figure 12: Doublet exclusion by gene content — Representation of distribution plot of cells by gene content. Cells to the right of the threshold (vertical red line) have twice the expected number of genes per cell and are likely doublets.

Cross-species analysis

Crosstalk represents the percentage of doublet cells in droplets or microwells across a given experiment. An effective way to determine cellular crosstalk is by mixing cells from two different species in one sample at a 1:1 ratio. When analyzing sample types that include cells from two different species, any cells detected with UMIs from both species represent doublets (Figure 13).
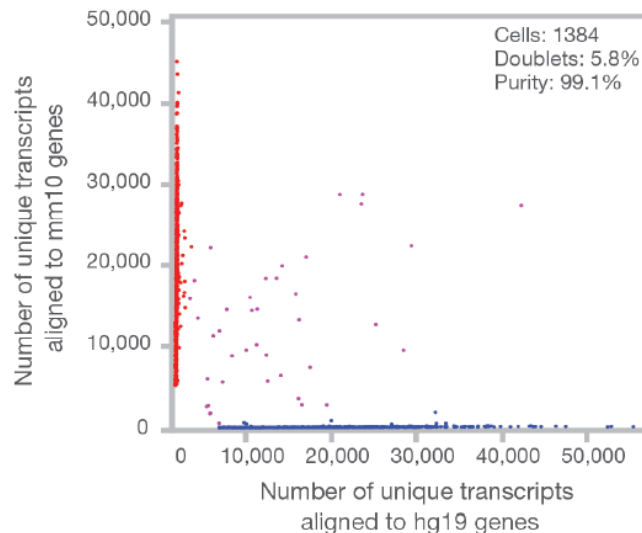
Figure 13: Doublet exclusion by species-specific UMIs — In a cell mixing experiment with cells from two different species, detection of cells with UMIs mapping to both species (purple dots) represent doublets.
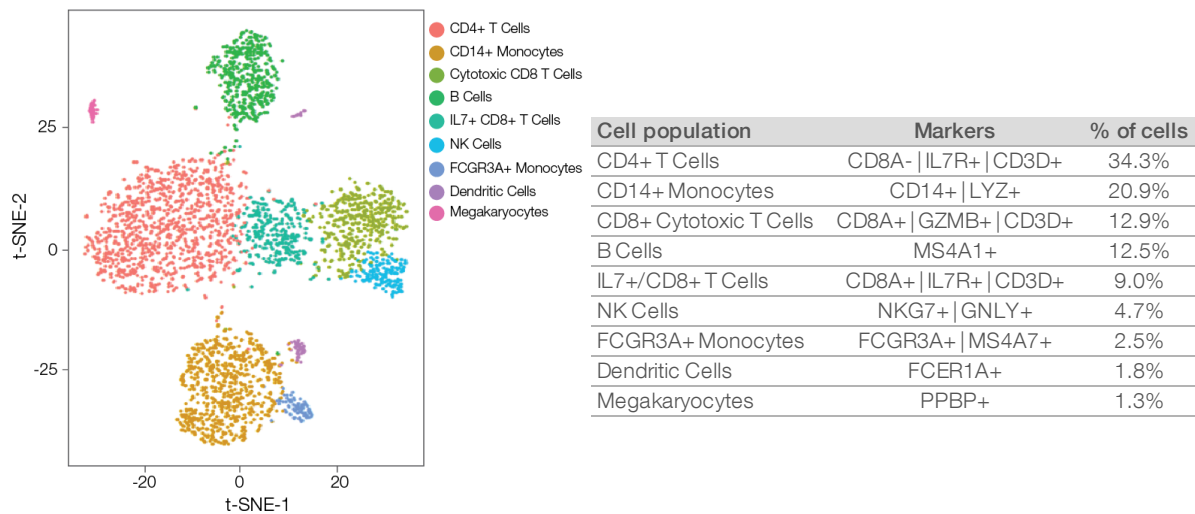
Data analysis

Tertiary analysis: data visualization and interpretation

After reads have been aligned to a reference genome and secondary analysis has been performed, including data QC to remove noncellular barcodes and/or poor-quality cells, a good quality data set can be visualized and explored to gain insights into the biology of the cells being studied. There are many free and commercially available software programs available (Table 10 and Table 11).

Seurat

Seurat is an R-based single-cell RNA-Seq analysis software designed to assess cellular heterogeneity with a number of tools such as normalization, dimensionality reduction approaches, plots, heatmaps, and data integration tools.[39] Seurat uses dimensionality reduction to make multidimensional data (eg, thousands of cells, each with thousands of expressed genes) understandable by mathematically reducing the number of dimensions into a two- or three-dimensional representation. The resulting clustering of cells into groups corresponds to particular cell states or types with characteristic features. (Figure 14).

Cell population	Markers	% of cells
CD4+ T Cells	CD8A- | IL7R+ | CD3D+	34.3%
CD14+ Monocytes	CD14+ | LYZ+	20.9%
CD8+ Cytotoxic T Cells	CD8A+ | GZMB+ | CD3D+	12.9%
B Cells	MS4A1+	12.5%
IL7+/CD8+ T Cells	CD8A+ | IL7R+ | CD3D+	9.0%
NK Cells	NKG7+ | GNLY+	4.7%
FCGR3A+ Monocytes	FCGR3A+ | MS4A7+	2.5%
Dendritic Cells	FCER1A+	1.8%
Megakaryocytes	PPBP+	1.3%

Figure 14: Unbiased cluster analysis of PBMCs in Seurat—Nine cell clusters were identified with 3354 cells down-sampled to approximately 70,000 reads per cell, with a resolution setting of 0.80 and 100 genes as a cutoff. Cells identified with Seurat are listed in the table.

Advanced data visualization with SeqGeq software

SeqGeq Software is a desktop application for advanced data analysis, exploration, and visualization of single-cell gene expression data developed by FlowJo, LLC (now part of BD Biosciences). SeqGeq Software features powerful data reduction and population identification tools. Direct integration with BaseSpace Sequence Hub enables visualization and analysis of expression data with statistic color-mapping of individual cells, summary heat maps, and drag-and-drop report editors (Figure 15).

Data analysis
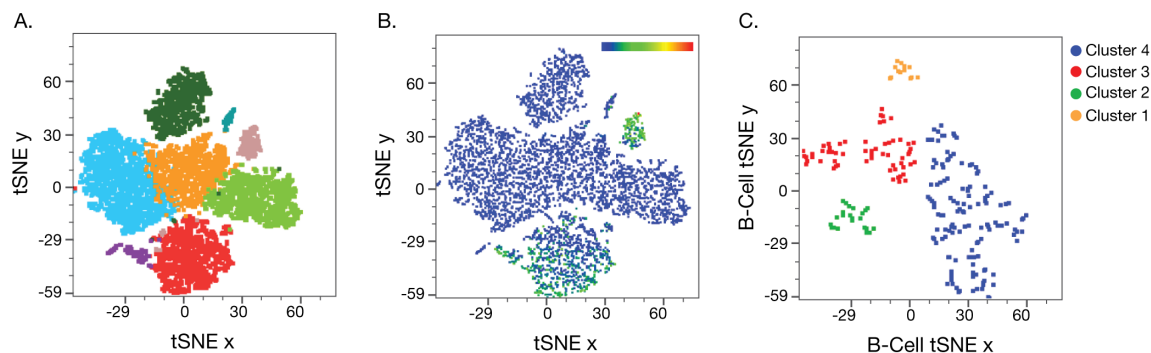
Figure 15: Simplified clustering analysis in SeqGeq — (A) Unbiased clustering analysis of PBMCs based on differential expression of cell-specific genes/markers. (B) Example of using gene set enrichment to identify B cells (green) based on expression of B-cell genes (blue corresponds to low expression and red to high in heat map). (C) B cells identified in (B) undergo further unbiased clustering using PCA guided t-SNE to identify subpopulations of B cells.

Table 10: Open-source tertiary analysis software

Software	Provider	Description
Seurat	Satija Lab satijalab.org/seurat	Seurat is an R package designed for QC, analysis, and exploration of single-cell RNA-Seq data.[33] Seurat aims to enable users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data.
t-SNE	Van der Maaten Lab lvdmaaten.github.io/tsne	t-distributed stochastic neighborhood embedding (t-SNE) is a computational technique that visualizes high dimensional data by giving each data point a location in a two- or three-dimensional map.[40] t-SNE is commonly used to visualize subpopulations with single-cell sequencing data.
UMAP	Git Hub github.com/lmcinnes/umap rdrr.io/cran/Seurat/man/RunUMAP.html	Uniform manifold approximation and projection (UMAP) is an algorithm for analysis of high dimensional data and an alternative to t-SNE, offering faster computation times.[41]
Monocle	Trapnell Lab cole-trapnell-lab.github.io/monocle-release	Monocle is an R-based single cell RNA-Seq analysis software designed to determine cell developmental trajectory. Monocle is ideal for experiments where there are known beginning and terminal cell states.
Human Cell Atlas	Broad Institute www.humancellatlas.org	The Human Cell Atlas is a consortium effort that will curate a data coordination platform intended to provide four key components: intake services for data submission, synchronized data storage across multiple clouds, standardized secondary analysis pipelines, and portals for data access, tertiary analysis, and visualization.

Table 11: Commercially available tertiary analysis software

Software	Provider	Description
SeqGeq	BD Biosystems	SeqGeq Software is a desktop application for advanced data analysis, exploration, and visualization of single-cell gene expression data. SeqGeq offers powerful data reduction and population identification tools.
Partek Flow	Partek	Partek Flow is a software analysis solution for NGS data applications. It has robust statistical algorithms, information-rich visualizations, and cutting-edge genomic tools enabling researchers of all skill levels to confidently perform data analysis.
CytoBank Platform	CytoBank, Inc.	Cytobank is a cloud-based platform designed for analysis and visualization of multiple single-cell data sets simultaneously.
Loupe Cell Browser	10x Genomics	The Loupe Cell Browser is designed to enable users to quickly and interactively find significant genes, cell types, and substructure within single cell data.
Tapestri Insight	MissionBio	Tapestri Insight is a software solution for single-cell DNA analysis. It includes sequence import, data analysis, and visualization.

Data analysis

For more information about steps in the single-cell sequencing analysis pipeline, read the Single-Cell RNA Data Analysis Workflow Technical Note

To see an example of secondary and tertiary analysis in a single-cell sequencing experiment, read the Single-Cell Sequencing of Peripheral Blood Mononuclear Cells Application Note

Summary

The precise analysis pipeline used for a single-cell sequencing experiment is variable and can be customized based on the research objectives of the study. Generally, this pipeline includes primary, secondary, and tertiary phases, in which sequences are aligned, genetic components are characterized, and data are visualized and explored, respectively. If you would like to discuss various single cell sequencing analysis options and how they can be integrated with your research, contact your local Illumina representative.

Data analysis

6 Summary

Over the past decade there has been significant advancement in the area of single-cell characterization and study, with development of new technologies for cell isolation and new methods and applications for single-cell sequencing. These advances have stimulated the launch of numerous, accessible commercial solutions for every step of the single-cell sequencing workflow, from tissue preparation through data analysis. With increasing options for single-cell isolation and interrogation, there has been a remarkable diversification of experimental protocols, each with inherent strengths and weaknesses. Researchers, therefore, face decisions about cell throughput, sequencing depth, required transcript length, whether epigenetic or protein-level measurements should be included, and more when designing an experiment.

To fully harness the potential of single-cell sequencing to elucidate complex biological systems, careful experimental design and optimization of every step of the workflow is critical. Researchers must have clearly defined biological objectives and a rational experimental design to make informed decisions about the optimal approach for their research question. Here, we have outlined every step of the single-cell sequencing workflow and discussed important considerations and potential challenges for each, presented commercial offerings, and offered advice for designing and executing a successful single-cell study. Illumina is committed to harnessing the power of NGS for single-cell sequencing, to build a deeper understanding of cellular and molecular biology, complex diseases, and environmental impacts on human health.

7   Learn more

How do I get started with single-cell sequencing?

To start planning your single-cell sequencing experiment, take advantage of these resources:
- Buyer's Guide to Simple, Customized RNA-Seq Workflows
- Buyer's Guide to Next-Generation Sequencing Systems
- Single-Cell RNA Data Analysis Workflow Technical Note
- Library Prep and Array Kit Selector Tool

What if I need help during a sequencing run or with data analysis?

Whether you have basic data analysis questions that require immediate attention or you have advanced questions requiring in-depth consultations, Illumina can help. Beyond immediate phone and email support, Illumina customer service and support teams provide a full suite of expedient solutions from initial trainings, to instrument support, personalized consultation, and ongoing NGS education. Illumina customer support offerings include:

Illumina Technical Support

Global, 24/5 phone and email support in the Americas, Europe, and Asia-Pacific.

Illumina Technical Support specialists can perform desktop sharing with GoToAssist—a powerful tool for quick identification and diagnosis of issues over the phone with live desktop sharing. For faster case handling, enter your case number at the main phone menu to be routed directly to the Technical Support specialist handling your case.

Illumina University Training

- Instructor-Led Training at your chosen facility
- Instructor-Led Training at an Illumina Training Center
- On-Line courses
- Webinars

Illumina Consulting Services

- Proof-of-Concept Services for instrument and library preparation testing
- Concierge Custom Design Service for design assistance and product optimization
- Illumina Bioinformatics Professional Services for bioinformatics consultation and/or training
- Illumina Genomics IT Consulting Services for genomics IT solutions
- Installation Qualification (IQ), Operational Qualification (OQ), and Performance Qualification (PQ)

Who can I talk to for more information on single-cell sequencing?

To speak with an Illumina representative about single-cell sequencing solutions, call the Illumina Customer Solutions Center at 1.800.809.4566 (North America) or 1.858.202.4566 (Outside North America) and start planning your single-cell sequencing experiments today.

8   Glossary

chastity filtration: A process where suboptimal clusters on patterned flow cells are filtered and removed from image analysis. Chastity is defined as the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities. Clusters "pass filter" if no more than one base call has a chastity value below 0.6 in the first 25 sequencing cycles.

cluster density: The number and distribution of clusters on a flow cell. Cluster density is an important metric for sequencing performance, particularly with nonpatterned flow cells, as it can significantly impact data quality and yield from a sequencing run.

cluster generation: A process where libraries are loaded onto flow cells and fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is amplified into distinct, clonal clusters through bridge amplification. Each cluster contains up to 1000 sample strands, usually 120-170 base pairs in length

complementary metal-oxide semiconductor (CMOS) technology: CMOS technology enables one-channel sequencing chemistry, which supports lower sequencing costs in a compact system while maintaining high-accuracy data.

flow cell: A glass slide with one, two, or eight physically separated lanes, depending on the instrument platform. Each lane is coated with a lawn of surface bound, adapter-complimentary oligos. A single library or a pool of up to 96 multiplexed libraries can be run per lane, depending on application parameters.

fluorescence activated cell sorting (FACS): A technology that provides qualitative and quantitative measurement of cellular characteristics such as size, internal complexity, DNA/RNA content and a wide range of membrane-bound and intracellular proteins via detection of autofluorescence or fluorochrome-conjugated antibodies.

Genomic Quality Number (GQN): A calculation developed by Advanced Analytical Technologies, Inc. (AATI) for use with the Fragment Analyzer for assessing quality of DNA samples.

index/barcode/tag: A unique DNA sequence ligated to fragments within a sequencing library for downstream *in silico* sorting and identification.

multiplexing: A technique to increase throughput of sequencing systems where large numbers of libraries with unique indexes can be pooled together, loaded into one lane of a sequencing flow cell, and sequenced in the same run. Reads are later identified and sorted via bioinformatic software in a process called demultiplexing.

next-generation sequencing (NGS): A non-Sanger-based high-throughput DNA sequencing technology. Compared to Sanger sequencing, NGS platforms sequence as many as billions of DNA strands in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes.

patterned flow cell: A flow cell that contains billions of nanowells at fixed locations, providing even cluster spacing and uniform cluster size to deliver extremely high cluster densities.

paired-end sequencing: A process of sequencing from both ends of a DNA fragment in the same run and aligning the forward and reverse reads as read pairs.

percent passing filter (%PF): Percent passing filter (%PF) is an important sequencing QC metric that refers to the number of clusters that have passed a filter and will be retained for downstream analysis.

percent ≥Q30: Q30 is a quality score in which one base call in 1000 is predicted to be incorrect. Percent ≥Q30 refers to the percentage of bases that have a quality score of Q30 or above.

quality score (Q-score): A prediction of the probability of an error in base calling.

quantitative polymerase chain reaction (qPCR): An application that enables the measurement of nucleic acid quantities in samples. The nucleic acid of interest is amplified with the polymerase enzyme. The level of the amplified product accumulation during PCR cycles is measured in real time. These data are used to infer starting nucleic acid quantities.

read depth: See "sequencing coverage". Alternatively, in single-cell sequencing read depth is discussed not in the number of reads per base, but in the number of reads per cell.

RNA Integrity Number (RIN): An algorithm that assigns integrity values to RNA measurements based on electrophoretic RNA measurements and a combination of different features that contribute information about RNA integrity to obtain a more universal measure.

RNA Quality Number (RQN): A proprietary algorithm developed by AATI for use with the Fragment Analyzer for assessing quality of RNA samples, which is equivalent to the RIN.

sequencing by synthesis (SBS): SBS technology uses four fluorescently labeled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel. During each sequencing cycle, a single labeled dNTP is added to the nucleic acid chain. The nucleotide label serves as a "reversible terminator" for polymerization: after dNTP incorporation, the fluorescent dye is identified through laser excitation and imaging, then enzymatically cleaved to allow the next round of incorporation. Base calls are made directly from signal intensity measurements during each cycle.

sequencing coverage: The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times. Sequencing coverage can also be referred to as "read depth".

9 References

1. Tung PY, Blischak JD, Hsiao CJ, et al. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep*. 2017;7:39921.
2. Herzenberg LA, Parks D, Sahaf B, et al. The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin Chem. 2002;48(10):1819–1827.
3. Jones GM, Busby E, Garson JA, et al. Digital PCR dynamic range is approaching that of real-time quantitative PCR. *Biomol Detect Quantif*. 2016;10:31–33.
4. Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K. Experimental considerations for single-cell RNA sequencing approaches. *Front Cell Dev Biol*. 2018;6:108.
5. Sigma-Aldrich. Centrifugation. *Biofiles*. 2011;6(5):4–14.
6. Method of the Year 2019: Single-cell multimodal omics. *Nat Methods*. 2020;17(1):1.
7. Valihrach L, Androvic P, Kubista M. Platforms for Single-Cell Collection and Analysis. *Int J Mol Sci*. 2018;19(3):DOI 10.3390.
8. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161(5):1202–1214.
9. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161(5):1187–1201.
10. Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
11. Pellegrino M, Sciambi A, Yates JL, Mast JD, Silver C, Eastburn DJ. RNA-Seq following PCR-based sorting reveals rare cell transcriptional signatures. *BMC Genomics*. 2016;17:361.
12. Retig JR, Folch A. Large-scale single-cell trapping and imaging using microwell arrays. *Anal Chem*. 2005;77(17):5628–5634.
13. Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*. 2018;172(5):1091–1107.
14. Cao C, Lemaire LA, Wang W, et al. Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nat*. 2019;571:349-354.
15. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nat*. 2019;566(7745):496-502.
16. Srivatsan SR, McFaline-Figueroa JL, Ramani V, et al. Massively mutliplex chemical transcriptomics at single-cell resolution. *Science*. 2020;367(6473):45-51.
17. Illumina. (2017) Scalable Nucleic Acid Quality Assessments for Illumina Next-Generation Sequencing Library Prep. Accessed May 2019.
18. Agilent Technologies. (2000) Comparing the Agilent 2100 Bioanalyzer performance to traditional DNA analysis techniques.
19. Data calculations on file. Illumina, Inc., 2017.
20. Sims D, Sudbery I, Ilot NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121–132.
21. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 2017;9:75.
22. Streets AM, Huang Y. How deep is enough in single-cell RNA-seq? *Nat Biotechnol*. 2014;32(10):1005–1006.
23. Rizzetto S, Eltahla AA, Lin P, et al. Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Sci Rep*. 2017;7(1):12781.
24. National Genomics Infrastructure. Comparison of PE and SE for RNA Seq. *SciLifeLab*. 2016;1–3.
25. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013;8(10):e77910.
26. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63.
27. Illumina. (2015) Patterned Flow Cell Technology Technical Spotlight. Accessed May 2019.

28. Illumina. (2019) Cluster Optimization Overview Guide. Accessed May 2019.
29. Illumina. (2017) Calculating Percent Passing Filter for Patterned and Nonpatterned Flow Cells Technical Note. Accessed May 2019.
30. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;15 (6):e8746.
31. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8(3):175–185
32. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998;8(3):186–194
33. 10X Genomics. (2017) Chromium Single Cell 3' v2 Libraries – Sequencing Metrics for Illumina Sequencers. Accessed May 2019.
34. AlJanahi AA, Danielsen M, Dunbar CE. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol Ther Methods Clin Dev*. 2018;10:189–196.
35. Illumina. (2013) bcl2fastq Conversion User Guide. Accessed May 2019.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25 (14):1754–1760.
37. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
38. Lafzi A, Moutinho C, Picelli S, Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat Protoc*. 2018;13(12):2742–2757.
39. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotech*. 2018;36;411–420.
40. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J of Machine Learning Res*. 2008;9:2579–2605.
41. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotech*. 2019;37:38–44.

illumina®

For Research Use Only. Not for use in diagnostic procedures.