Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

SINGLE-CELL SEQUENCING: UNRAVELING GENETIC HETEROGENEITY

Waleed Faleh Marzouq Albalawi1, Khaled Hassan Alfaifi2, Mohammed Saeed Alqhtani3 And Fahad Dukhi Qismi Alanazi4

1 Corresponding Author, MEDICAL LABORATORY TECHNOLOGY, weelly14@gmail.com, KING SALMAN MILITARY HOSPITAL

2 MEDICAL LABORATORY TECHNOLOGY, khfkhaled@gmail.com, KING SALMAN MILITARY HOSPITAL

3 MEDICAL LABORATORY TECHNOLOGY, mohammedsq11@gmail.com, KING SALMAN MILITARY HOSPITAL

4 Laboratory specialist, Fahad_dokhi@hotmail.com, KING SALMAN MILITARY HOSPITAL

Abstract

Cellular heterogeneity is a fundamental characteristic of multicellular organisms as well as some models of artificial intelligence. It contributes to the normal development of tissue variety and functional stability. However, in the context of disease, cellular homogeneity is often disrupted, resulting in genomic, epigenomic, transcriptomic, and phenotypic diversity among individual cells in a population. This genomic and phenotypic diversity is the driving force behind tumorigenesis, metastasis, drug resistance, and treatment failure in many cancers (Ye et al., 2016). Currently, the bulk population assay remains the gold standard in characterizing genomic and phenotypic diversity. However, population-based experiments can only provide averaged information and fail to represent the entire spectrum of cellular states. As a consequence, an understanding of oncogenic cellular diversity and the ability to accurately predict and intercept cancer progression are beyond reach. Whole-genome amplification (WGA) was first demonstrated in 1992 using degenerate oligonucleotide-primed PCR (DOP-PCR) on a few nanograms of input DNA. During the late 1990s, several other WGA protocols were developed, some of which are still in use today. Scheduling the polymerase chain reaction (PCR) and isothermal amplification at various temperatures can minimize bias due to local differences in GC content. In general, PCR-based WGA methods are simple, rapid, and inexpensive. Isothermal amplification is becoming increasingly popular, particularly for low-input DNA (C. Macaulay & Voet, 2014).

Keywords

Single-Cell Sequencing Genetic Heterogeneity Keywords: single-cell sequencing, heterogeneity, scRNA-seq, NGS, RNA, single cells

The importance of diversity and cellular specialization is clear for many reasons, from population-level diversification, to enhanced resiliency to unforeseen stresses, to unique functions within metazoan organisms during development and differentiation (L. Goldman et al., 2019). However, the level of cellular heterogeneity is just now becoming clear through the integration of genome-wide analyses and more cost effective Next Generation Sequencing. With easy access to single-cell NGS, new opportunities exist to examine different levels of gene expression and somatic

mutational heterogeneity, but these assays can generate yottabyte scale data. Here, the importance of heterogeneity for large-scale analysis of scNGS data is modeled, with a focus on the utilization in oncology and other diseases, and a guide is provided to aid in sample size and experimental design.

1. Introduction to Single-Cell Sequencing

Single-cell sequencing (SCS) technology has been recently employed for analyzing the genetic polymorphisms of individual cells at the genome-wide level. This new biotechnology allows for the comprehensive descriptive analysis of genomic, transcriptomic, and epigenomic data from a single cell, and has important applications in research and the clinical arena. A defining characteristic of healthy tissues and organs is the cellular homogeneity of the individual cell type. However, it is becoming more well known that cellular heterogeneity is a characteristic of many cancers (Ye et al., 2016). The widely accepted cancer stem cell theory of tumorigenesis posits that some somatic stem cells, or progenitor cells with stem cell-like properties, become cancerous and aberrantly proliferate into different subgroups of cancer cells, with some possessing unexpected quiescent or highly proliferative phenotypic characteristics.

This likely explains a tumor's capability to undergo multilineage differentiation, progress, and metastasize, as well as immune evasion and resistance to chemotherapy. The most effective current cancer therapies appear to be correlated with high degrees of cellular homogeneity (Qin et al., 2022). For instance, the good prognosis associated with luminal subtypes of breast tumors is correlated with strictly homogeneous ER/PR positive protein expression, with high levels of mRNA transcripts. Minimally differentiated acute myeloid leukemias (AMLs) are associated with poor prognosis, as these tightly clustered tumors are comprised of diverse, but predominantly CD34 positive, progenitor cells that are resistant to differentiation-inducing therapies. However, for most other cancers, protein expression appears to be significantly heterogeneous, limiting the efficacy of novel targeted therapies.

1.1. Definition and Importance

High-productivity sequencing technologies have enabled the characterization of a plethora of biological systems from the population-level down to the single-molecule level. In particular, the recent advent of single-cell sequencing technologies offers new possibilities to examine the genetic make-up of individual cells, addressing important biological questions that are either intractable or difficult to address at the population-level. As multicellular organisms, many biological systems exhibit a diverse range of cellular phenotypes. The importance of diversity and cellular specialization is clear for many reasons, from population-level diversification to unique functions within metazoan organisms during development and differentiation (L. Goldman et al., 2019). At a more subtle level, even apparently homogeneous cellular populations maintain a dynamically regulated balance between specialization and plasticity.

Diverse cellular phenotypes mitigate risk and promote robustness in the face of environmental fluctuations. Cellular specialization enables population-level persistence through the allocation of different functions to different cell-types. The level of cellular heterogeneity is just now becoming clear through the integration of genome-wide analyses and more cost-effective Next Generation Sequencing. With easy access to single-cell NGS, new opportunities exist to examine different levels of gene expression and somatic mutational heterogeneity. Here, the importance of heterogeneity for large-scale analysis of scNGS data is modeled, with a focus on the utilization in

oncology and other diseases, providing a guide to aid in sample size and experimental design. Despite being ostensibly homogeneous, cellular populations often beget, or contain, sub-populations that differ with respect to one or many characteristics.

1.2. Technological Advancements

The realization that biological systems are inherently heterogeneous as well as a response to the need for lower-cost, higher-resolution analysis platforms prompted an explosion of interest in single-cell genomic technologies. Most early efforts focused on DNA analysis, but the recent advent of transcriptomics and other "-omics" platforms has expanded the range of possible applications (C. Macaulay & Voet, 2014). Techniques for the isolation of single cells, cultivation and manipulation in discrete niches or compartments, amplification of cell-derived genome or transcript copies, and genome-wide analysis platforms—primarily microarray and next-generation sequencing devices—have all been developed at progressively higher levels of precision, resolution and throughput in recent years.

Analysis from one cell reveals a wealth of information and complexity that is obscured in population measurements. Assessment of variation in time or space is generally impossible without being able measure the relevant quantity at high resolution in a single entity. In biological systems, as in many others, "the whole is greater than the sum of its parts" and emergent properties arise from complex interactions between individual components, some of which are hidden from view in population measurements. Complex systems generally fail to conform to simplistic or ideal mathematical models, presenting a challenge to the quantification and understanding of how they operate and evolve.

As a consequence of the "-omic" technology and data analysis revolutions, biologists now have access to high-throughput platforms for measuring a range of genomic, epigenomic, transcriptomic and proteomic properties for individual cells (Hwang et al., 2018). For a number of analyses, single-cell approaches are either preferable or essential. Some cell types are so rare—typically comprising a fraction of less than one part in a million of the total cell population—that without enrichment and single-cell characterisation, the chance of detection approaches zero.

2. Genetic Heterogeneity in Cellular Populations

As cellular and genomic technologies advance complimentarily, novel services emerge for basic biology and human health. Genomes can be sequenced cheaply and thoroughly; however, metagenomic and transcriptomic studies reveal the need for comprehending environmental effects, species distinction, and population-wide diversity. Ecological studies show the importance of "ex-lab" experiments. Single-Cell Jackie Chan sequencing (?cc-seq) combines orthogonal ex-situ measurements of cellular transcription, genotype, and phenotype with imaging and analysis pipelines to characterize, isolate, and sequence individual eukaryotic microbes in bulk samples, enhancing understanding of different ecosystems. Soils are complex, multilayered, and heterogeneous matrices that support diverse life and biogeochemical cycling. Most soil microbes remain uncultivated, leading to gaps in understanding soil ecology and function. This new service enables the post-agricultural sheen of genomic insight into individual soil eukaryotes, unraveling diversity, ecological roles, and biogeochemical functions beyond expectations. Urbanization drastically modifies soils, altering physical and chemical properties, and affecting microbial communities (L. Goldman et al., 2019). Agricultural land use also shapes soils and microbial

metagenomes, impacting ecosystem services essential for food security and human health. Future sequencing in other biomes will uncover new diversity and ecology paradigms.

2.1. Causes of Genetic Heterogeneity

Genetic heterogeneity, generally defined as the coexistence of multiple genotypes within an isogenic population, can arise from both genetic and epigenetic causes contributing to transcriptional variability. A genetic origin of heterogeneity is provided by mutations in genomic sequence. Mutations can affect gene coding regions or regulatory sequences and modulate gene expression either by directly altering protein function or by changing the activity of regulatory elements (L. Goldman et al., 2019). Transcription can also be affected by mutations in non-coding regions, such as those that alter chromatin structure or modify transcription factor binding sites. Genomic mutations are often deleterious to a cell population, as in the case of single point mutations causing oncogenic transformation. However, some mutations can provide a short-term advantage to a subpopulation in a changing environment, as in the case of antibiotic resistance mutations in bacteria. In such scenarios, heterogeneous cell populations may arise due to the random distribution of mutations during replication.

Cellular behavior is controlled by a complex network of interactions among proteins, RNAs and metabolites. Each cellular component can affect the activity or fate of others, and these interactions can lead to the emergence of non-deterministic outcome in cellular responses to perturbations, even in the presence of a well-defined input. A class of mathematical models, broadly referred to as “nonequilibrium stochastic systems”, has been widely used to describe the emergence of cellular heterogeneity due to nonlinear interactions among intracellular biochemical processes. In the simplest formulation, such models consider the dynamics of a single variable affected by deterministic changes and random fluctuations. The deterministic part accounts for the average cellular behavior, while randomness is explicitly taken into account by adding stochasticity to model equations.

2.2. Implications in Disease and Development

Under both normal and pathological conditions, tissues become a complex mixture of cell types and states as they form and mature. This diversity and associated specialization in functions are essential for health, survival, and adaptability to environmental changes. At the same time, the understanding of cell population composition, transition, and fate decisions is crucial for modeling development, physiology, and disease (L. Goldman et al., 2019). Transcriptional profiling of cells acquired over space and time can address these questions. Technologies for examining the transcriptome of individual cells have recently emerged in response to the demand for better measurements and algorithms for interpreting high-dimensional datasets. With these tools becoming widely accessible, single-cell transcriptomics is expected to transform biology from basic research to drug discovery and development.

The first transcriptomic profiling of individual cells used microdissection to isolate cells followed by amplification of the collected mRNAs. While pioneering, this proof-of-principle study was limited to modeling and validating cell types in a fixed tissue and demonstrated the technical difficulties of single-cell studies. Single-cell sequencing (scSeq) overcame many limitations of prior methods for transcriptomic profiling. Like bulk sequencing, scSeq is library-based and relies on reverse transcription to convert polyadenylated mRNAs into cDNA libraries for downstream amplification and sequencing. However, in contrast to bulk, scSeq uses individual-cell-level barcoding to uniquely tag and preserve sequences representative of each cell across a population.

Here, development and applications of scSeq are summarized, with a focus on transcriptomic methods. High-throughput technologies that resolve the spatiotemporal states of cells and developmental trajectories inferred from transcriptomic data are also reviewed.

3. Applications of Single-Cell Sequencing

Single cell analysis has great promises in basic biomedical research and clinical applications. It helps to study the dynamic changes of the cells and the intricate networks formed by the interactions among cells, and distinguishes cell populations with different functions, states, or prognoses. In basic research, single cell measurements can reveal cell to cell variability and heterogeneous differences in the individual cells that cannot be captured by bulk measurements. With clinical samples, single cell analysis can be used to detect and characterize rare cells that are of great importance for disease diagnosis and prognosis. For example, in cancer, a single tumor cell can be studied to infer the evolution history of the tumor and to detect the mutations that confer drug resistance. On the other hand, a few circulating tumor cells can be isolated from a large volume of blood and analyzed to get information about the metastatic tumors (Hu et al., 2016). A combination of high-throughput and multiparameter approaches is used in single cell analysis. By profiling a certain type of biomolecule of the single cells, the cells are interrogated in parallel and the perturbations are multiplexed, which can reduce the experimental cost and time. Currently the single cell analysis tools can be divided into three groups according to the biomolecules being analyzed: genomics, transcriptomics, and proteomics. For each group, a number of techniques and platforms have been developed or are under development, which can accommodate various experimental requirements. The development of efficient single cell analysis methods still requires attention in instrumentation, algorithms, and applications. Single cell genome sequencing allows us to identify chromosomal variations, such as copy number and single-nucleotide variations. It also allows us to study tumor evolution, gamete genesis, and somatic mosaicism, which is reflected in the genomic heterogeneity among a population of cells (Navin & Hicks, 2011).

3.1. Cancer Research

Heterogeneity is a complex characteristic of biological systems, which is of great importance in many areas of life sciences and biomedical studies. At the cellular level, cellular heterogeneity means the irregularities of cells in terms of some characteristics, such as size, shape, growth rate, and physiological behavior. Cellular heterogeneity is a fundamental characteristic of many cancers (Ye et al., 2016). The proliferating cancer cells acquire various genetic alterations during tumorigenesis, which give rise to a subpopulation of cells with diverse phenotypes. Such a lack of cellular homogeneity profoundly contributes to the great difficulty in designing effective targeted oncological therapies.

Therefore, the development of novel methods to determine and characterize the oncological cellular heterogeneity is a critical next step in the development of novel cancer therapies. Single-cell sequencing (SCS) technology has been recently employed for analyzing the genetic polymorphisms of individual cells at the genome-wide level (Müller & Diaz, 2017). Based on high-throughput next-generation sequencing, single-cell sequencing (SCS) technology can systematically characterize genetic, transcriptomic, or epigenomic alterations of individual cells. SCS technology comprises three basic steps: (i) precise isolation of the single cell of interest, (ii) isolation and amplification of the genetic material of the single cell, and (iii) descriptive analysis of the genomic, transcriptomic, and epigenomic data of the single cell. In solid tumors, most studies have focused on characterizing bulk tumors, which generally ignore the cellular

heterogeneity in cancers. SCS technology may be applied to circulating tumor cells, which may greatly aid in predicting the tumor progression and metastasis.

3.2. Neuroscience

Neurons with identical genomes and basic functions exhibit a unique physiological and morphological variability. This "phenotypic mosaicism" provides a basis for learning, memory and recovery after injuries but also is a substrate for emergence of neurodegenerative phenotype(s). Genetic and environmental factors cumulatively affect cells' genomes, transcriptomes, epigenomes and proteomes causing diverse neuropsychiatric phenotypes and diseases. The emergence of new technologies such as next-generation sequencing (NGS) and microarrays for in situ hybridization and immunocytochemistry allow large scale gene/protein expression analyses at the single cell level. Novel solid-state nanotechnologies may further simplify miniaturization and multiplexing of single cell analyses (Y Iourov et al., 2012). Single cell genomic/epigenomic/proteomic analyses of neuronal cells using current NGS technologies can be challenged and new scientific opportunities created by the complexity of cellular and molecular "neuropathways", uniqueness of cellular/circuital patterns and functional variability of neuronal cells. Brain regions consist of dozens to hundreds of different neuronal cell types. As a rule, neurobiologists studying brain functions/neuropsychiatric diseases focus on large populations of one or a few types of "classically" recognized neuronal cells.

3.3. Developmental Biology

During organismal development, the zygote divides to form distinct cell types and lineages, establishing axes of symmetry and differentiating structures such as limbs, eyes, and hearts. This remarkable feat is both robust and sensitive, as millions of cells assemble into complex biological structures. Perturbations can lead to significant differences in diagnostic outcomes, from healthy to diseased states such as cancer or developmental disorders. Thus, the ability to resolve changes in health and disease states at the single-cell level is paramount. Developmental biology studies both robustness and sensitivity in gene regulation, cell fate decisions, and embryonic patterning, focusing on how cells respond to spatial and temporal patterns of signaling molecules (L. Goldman et al., 2019). In developmental biology model systems, sensitivity is often examined using perturbations such as knockout of a signaling component. This approach has been applied to single-cell sequencing data to infer the presence of a developmental signaling pathway. However, many questions remain unanswered, particularly concerning self-organizing models of embryogenesis. Here, a actively proliferating multi-species reaction-diffusion model is considered where mature cells secrete chemical signals that re-pattern the underlying field. The significance of this model is twofold: it provides a mathematical foundation for studying sensitivity in complex systems, and it predicts that developmental patterns emerge from both macroscopic and microscopic phenomena, which is relevant in the context of developmental biology.

4. Techniques and Methodologies

Sequencing and Pre-amplification. Experimental considerations including cell capture and lysis, reverse transcription, amplification, library preparation, multiplexing and bioinformatics analysis are discussed for instruments ranging from simple to complex. A focus is on scRNA-seq — the sequencing of RNA from individual cells and currently one of the most widely used single-cell genomics applications (H. Nguyen et al., 2018). Other single-cell applications, methods, or considerations can be examined through the general framework here.

As sequencing costs decrease, single-cell genomics expands into new applications and fields. Comprehensive characterizations of systems at the single-cell level can yield important insights into cellular diversity, state and lineage as well as the effects of perturbations. For example, the large-scale single-cell analysis of a biological system might address how the system is organized, how it changes over time or in response to perturbations and the identification of rare or difficult-to-observe states. Single-cell approaches can recover information that is lost in traditional observations of ensembles (V. H. Hornung et al., 2023). For example, measurements on populations can only characterize averages, which can obscure important biological features such as sub-population distributions or dynamic trajectories through state space.

4.1. Single-Cell Isolation Techniques

Successful single-cell isolation is a primary step for subsequent chemical and biological analyses of single cells (Zhang et al., 2014). Currently, four main approaches are available for single-cell isolation: serial dilution, micromanipulation, fluorescence-activated cell sorting, and laser-capture microdissection. These strategies are plagued by time spent, operational complexity, limited efficiency, deterioration of cell viability, incompetence in the isolation of single cells into nanoliter liquids, inability to select single adherent cells with distinct phenotypes, and/or the requirement of expensive instruments. Recently, extensive efforts have been devoted to developing new techniques for rapid, efficient, and high-throughput single-cell isolation.

The size and morphology of cells restrict the choices of available materials for fabrication of single-cell manipulators and the induced force (Hu et al., 2016). Typically, hydrophilic polydimethylsiloxane (PDMS) microdevices are designed for single-cell manipulation by pressure-driven flow in microchannels. The predominated mechanism for hydrophilic PDMS microdevices to trap cells is the balance between flow drag force and trapping force. As the trapping force is determined by the shape of the trapping area, microstructures with different geometries have been implemented to trap cells. Various geometries of microstructures have been designed to improve the efficiency of cell trapping including a step structure, cross-junction structures, straight microchannels with micro-posts, dumbbell shaped microchannels, and spiral microchannels. Although innovative designs of microstructures have been widely reported to enhance the efficiency of trapping cells, this improvement is still limited.

4.2. Library Preparation and Sequencing

Each technique describes a different method for single-cell RNA sequencing, a process used to analyze the transcriptomes of individual cells. The first technique involves a series of equipment and kits, which captures and lyses cells, and performs reverse transcription and amplification using microfluidic chips and primers with unique barcodes and poly(A) tails. The second technique details a different approach that uses gel bead emulsions to partition cells and add Hash Barcodes for multiplexing, with reverse transcription occurring in a thermocycler. The third technique employs droplets generated by a commercial microfluidics system to partition cells and reagents for complementary DNA synthesis, with cell-specific barcodes included in reverse transcription primers, and post-reaction cleanup using paramagnetic beads. Finally, the fourth technique isolates single cells using a micromanipulator, lyses them, and synthesizes cDNA with RNA polymerase and a specific primer containing a cell-specific tag, followed by biotinylated primer-directed amplification, which allows capture on streptavidin beads and generates libraries with 5′ adaptors.

5. Bioinformatics Analysis of Single-Cell Data

Bioinformatics high-throughput single-cell transcriptomic studies start with quality control. The raw data files in FASTQ format store the sequence information and quality scores of the single-cell transcripts. Pre-processing of single-cell RNA sequencing (scRNA-Seq) data files generally contains four steps: quality filtering, alignment, gene quantification, and outputting as count matrix. After initial quality control (QC) of sequencing reads, the pre-processed data is ready for downstream bioinformatics analyses (B. Poirion et al., 2016). The key downstream analyses of high-throughput scRNA-Seq studies start with normalizing the count matrix data. Normalization of the transcript reads is necessary to remove technical noise and to control the variability from confounding factors. After normalization, the bioinformatics analysis should make data exploratory analysis, detection of differentially expressed genes (DEG) and pathways, unsupervised cell clustering, and annotation cell types. After exploratory data analysis and cell clustering, it is possible to detect cell states or cell subpopulations that are close across the underlying pseudo-time developmental trajectory.

5.1. Data Preprocessing

With the development of sequencing technology, single-cell RNA sequencing (scRNA-seq) has become an important method in dissecting cell heterogeneity, cell lineage tracing, and identifying new cell types, states, and biomarkers. With its gradually reduced costs, growing applications in basic and clinical research have been seen, as well as rapid developments of bioinformatics tools for analyses of scRNA-seq data. For beginners, it is overwhelming to confront the rapidly growing toolkits and platforms. A systematic pipeline with guideline scripts and configuration files for using both command line and R shiny app is provided, beginning from the experimental design and considerations through checkups and analyses of the data quality using widely used tools, and standard downstream analyses on clustering, trajectory reconstruction, and differential expression of scRNA-seq data using Seurat, for illustrative examples. scRNA-seq has been extensively applied to uncover transcriptomic diversity across a variety of biological systems and diseases. Furthermore, an understanding of basic principles regarding considerations and potential pitfalls before and during scRNA-seq experiments is provided, as well as ways to deal with them (H. Nguyen et al., 2018).

A number of sequencing platforms have been widely used to generate scRNA-seq data. For poly-A enriched RNA, full-length scRNA-seq libraries can be prepared in 10-25 minutes. There was a certain proportion of low-quality data in raw data obtained by sequencing, which will cause a great interference with subsequent data analysis. Qualified sequencing data are a prerequisite for the reliability of subsequent data analysis results, and hence it is necessary to preprocess and evaluate the raw data. Preprocessing the raw data using self-developed open-source software to acquire the Clean Data for subsequent analysis (Xu et al., 2023). Then, the pre-processing analysis for the raw data of scRNA-seq uses that consists of four steps: (i) error correction of cell barcodes using predefined whitelists; (ii) use to align the reads to a reference genome; (iii) UMI correction and deduplication; (iv) obtain gene expression data. It is preferable to use genome reference because it allows for easier removal of captured “off-target” sequences. Eventually, each sample obtains gene expression matrices that can be further filtered and analyzed. Detailed and demonstrated filtering criteria will be provided for widely used preparations and sequencing platforms. Generally, if too many or too few genes are detected, the cell may have problems with high probability (e.g., doublets or empty droplet). In addition, there are special biochemical reasons why some cells have more mitochondrial genes than others, thus making this feature an efficient

filtering strategy to identify low-quality data. In general, the quality control analysis includes the following (i) check the basic information of scRNA-seq data; (ii) filter cells with low-quality reads; and (iii) filter out genes expressed in too few cells.

5.2. Clustering and Dimensionality Reduction

High-throughput single-cell sequencing technologies generate large scale gene expression matrices (genes × cells) with millions of noisy and heterogeneous measurements. For most applications, an essential first step in the analysis pipeline is to summarize and visualize a cohesive representation of this high-dimensional space. Clustering techniques facilitate the identification of transcriptionally similar groups of cells, while dimensionality reduction methods create low-dimensional embeddings to visualize the high-dimensional space. Clustering methods can be broadly categorized as being either deterministic or probabilistic. Deterministic methods assign each cell to exactly one cluster, including K-means, spectral clustering, graph-based clustering, and single-cell specific extensions such as RaceID, SC3, and CID. Probabilistic methods provide a soft clustering (or membership) scheme where a cell can belong to multiple clusters, including Dirichlet Process Mixture Models and single-cell specific extensions such as pNMF, SNN-Cliq, and DPT (L. Hsu & C. Culhane, 2020).

6. Challenges and Limitations

Single-cell sequencing, while a revolutionary advancement in genomic research, faces numerous challenges and limitations. Notably, these hurdles include concerns over data handling, bioinformatics, regulatory policies, optimal study design, and cost management (H. Nguyen et al., 2018). Well-established technical protocols exist for library preparation, but the interpretation of single-cell sequencing data remains a relatively new exploration for most investigators. Consequently, processing and analysis methods can vary greatly across different studies, making replication and validation problematic. Reducing batch effects is particularly crucial for comparative studies, especially when multiple tissue types or samples are analyzed together. The choice of single-cell sorting methods is usually influenced by sample type and downstream analysis. Fluorescence-activated cell sorting (FACS) is the most flexible platform, allowing for simultaneous collection of multiple cell populations and an opportunity for further phenotyping. However, FACS requires prior knowledge of cell surface markers and may introduce challenges in obtaining high-quality single-cell RNA due to time delays and complex procedures. Other methods, like micromanipulation or laser capture microdissection, also target specific cell types, with the former offering greater flexibility in downstream implementation. Nevertheless, both approaches are low-throughput and can subject cells to extreme conditions due to exposure to heat or toxic solvents. Commercially available multi-well chip systems provide an alternative, high-throughput means of capturing individual cells, allowing for better control of experimental conditions. These systems use either nanoliter droplet emulsions or gel-embedded barcoded beads to tag cDNA with cell-specific barcodes, enabling post-hoc sequence demultiplexing and identification of cell types. However, both approaches may result in the loss of complex cellular signaling networks due to the dissolution of co-cultured cells (F. Wills & J. Mead, 2015).

6.1. Technical Challenges

As the cost of bulk sequencing continues to fall, it has been widely adopted in both laboratory and clinical settings, enabling the speedy generation of genomic data for a broad range of applications. Unfortunately, many experiments can only be analyzed on a population-wide level. Bulk sequencing averages molecular measurements across many cells, confounding the information

intrinsic to individual cells. This limitation is particularly stark in cancer applications, where the clonal architecture of tumors—stemmed from the evolutionary diversification of their constituent cells—holds important prognostic and therapeutic insights (F. Wills & J. Mead, 2015). Considerable effort has been devoted to the development of bulk sequencing-derived methodologies for inferring tumor architecture. However, the reliance on statistical models that are prone to inaccuracies and artifacts has limited their efficacy, particularly in low-depth sequencing scenarios resembling those common in a clinical context.

Single-cell sequencing overcomes this limitation by measuring molecular profiles in individual cells. Initially pioneered in transcriptomic applications, the accessibility and diversity of bulk -omic technologies have catalyzed the rapid expansion of analogous single-cell methods across the genome, epigenome, transcriptome, proteome, and metabolome. Together, these allow the comprehensive assessment of cell state and identity, as defined by the intersection of their genetic, architectural, and environmental features, across diverse biological systems. In particular, the advent of high-throughput and cost-effective single-cell sequencing platforms is catalyzing the exploration of previously intractable phenomena at the single-cell level, from development to homeostasis, aging, and disease, including the ever-evolving pathobiology of cancer (H. Nguyen et al., 2018).

6.2. Biological Interpretation

The advent of high-throughput, next-generation sequencing (NGS) technologies has ushered in a new era of genomic investigations, accelerating even more so with single-cell NGS (scNGS). From the Human Genome Project comprising 1,000 bulk samples spanning 13 years of sequencing effort, large-scale genomic analyses now routinely begin with thousands of single-cell sequencing experiments generating millions or billions of reads in just a few hours. There are numerous and diverse technological approaches to single-cell genomics, transcriptomics, epigenomics, and beyond. While biological interpretation of bulk NGS analysis focuses on reconstruction from an ensemble of similar cells, scNGS necessitates reconstruction from a population of differing cells. Analysis of scNGS data must account for two key issues not applicable to bulk NGS: batch effects due to technical noise and cell-to-cell differences in genome-wide data.

Significantly, there are biological advantages in view of scNGS experimental design—and perhaps in advocacy for single-cell analyses over bulk analyses—that emerge from the statistical framework. With proper consideration, scNGS data can be informative for interpreting the entirety of the underlying cellular system and reconstruction across different cellular states (L. Goldman et al., 2019). To promote understanding of the importance and advantages of scNGS over high-quality bulk NGS, a statistical model quantifying cellular signal and noise is presented, along with examples applied widely from oncology to developmental biology. Large-scale analysis of single-cell data offers great insight into cellular systems, while being nontrivial and prohibitive with traditional bulk sequencing approaches.

7. Future Directions and Emerging Technologies

Recent years have witnessed an explosion of interest in single-cell experimentation, driven by remarkable advances in sensitive cell-disruption, genetic-barcode labelling, and high-throughput sequencing technologies. Most early efforts directly examined the heterogeneity of 1–100 genes across thousands to millions of single cells, typically using targeted amplification of cloned-primed transcripts. Sequentially generating numerous shallow-sequencing libraries interrogated the same

pool of barcoded-lipid droplets, micro-chambers, or wells, and was highly cost-effective. Novel transcriptome- and genome-sequencing technologies produced a deluge of remarkable discoveries and extraordinary cell atlases. Systematic deep exploration of single-cell epigenomic heterogeneity across diverse organisms, tissues, and developmental stages has just begun and will reveal important insights about chromatin states, 3D organization, and cell-type evolution and plasticity (C. Macaulay & Voet, 2014).

Over the past decade, many single-cell sequencing techniques have been developed, making it possible to examine single cells in different dimensions—DNA, RNA, epigenome, and 3D genome—and in different contexts and environments. These technologies have been broadly applied to explore various biological questions. With the rapid progress of technology development, bioinformatics methods, and data integration approaches, single-cell sequencing will continue to generate fruitful biological insights. In addition to studying basic biological issues, single-cell sequencing technologies might lead to new clinical applications in medicine and healthcare. Continuous efforts are required to reduce the cost and improve the throughput.

7.1. Spatial Transcriptomics

Spatial transcriptomics enables the visualization and analysis of gene expression in tissue sections while preserving spatial information. Tissue sections are spatially barcoded or tagged, and genes are measured by sequencing or other approaches, liberating transcriptomic and spatial data at the same time. Spatial transcriptomic data can be generated by several ways, and in the early days, microfluidic-based methods were developed to measure gene expression in tissue sections. Another approach is to combine in situ sequencing with spatial barcoding to recover original spatial coordinates. In situ sequencing quantifies gene expression in fixed tissue by sequencing DNA in the same location as the target RNA. The inclusion of spatial information in transcriptomic data expands a range of possibilities for analysis and visualization (Liu et al., 2021). Understanding the spatial distribution of gene expression has helped to answer fundamental questions in various research areas. For example, in cancer research, insights into the architecture of the tumor microenvironment can reveal how cancer cells communicate with surroundings. In pathology, the spatial arrangement of immune cells in tissues helps to estimate faster and more accurately the severity of infection. In developmental biology, the spatiotemporal pattern of gene expression is crucial to understanding how organisms develop from a single cell.

Two widely used methods for gene expression quantification are fluorescent in situ hybridization (FISH) and next-generation sequencing. FISH uses fluorescently-labeled RNA sequences as probes to identify its naturally occurring complementary sequence while preserving the spatial location of the target sequences. By designing multiple probes for different genes, the identity of target genes can be revealed by the fluorescence distribution of the probes. In contrast, next-generation sequencing methods use a shotgun approach to quantify RNA molecules across the entire transcriptome. Because of the high background signal caused by nonspecific binding, FISH methods are generally limited to detecting dozens of genes at a time. Even with single-cell sequencing, spatial information of cells can only be inferred (Chen et al., 2023). Various approaches have been made to measure gene expression while preserving spatial information. For example, one of the earliest spatial transcriptomic techniques, referred to as tomo-seq, applied the principle of tomography to measure spatial transcriptomic information in three-dimensional (3D) tissue samples. In tomo-seq, tissue samples are sliced by cryosection and measured with RNA-seq on the slice plane. Each slice contains transcriptomic information of a specific z-coordinate and

reconstructing the transcriptomic information in 3D involves a complex tomographic reconstruction process.

7.2. Multi-Omics Integration

As a fundamental characteristic of organisms, cellular heterogeneity arises from differences in genetic background, environment, and cellular history that affect gene activities. Measurement of gene activity through mRNA abundance gives rise to the transcriptome, which varies across individual cells even in apparently uniform environments. Thus, the transcriptome can be seen as a readout of cell state and regulation, and deciphering its heterogeneity unveils complex regulatory networks underlying developmental and pathological processes. Fortunately, advances in sequencing technologies have transformed transcriptome profiling from a population-average "batch" approach into a "single-cell" one that resolves transcriptome heterogeneity across individual cells (Rui Xing et al., 2020). Realized by barcoding strategies combined with sequencing-by-synthesis, titration-enabled microfluidics, and droplet encapsulation, transcriptome profiling at the single-cell resolution can routinely be performed in massively parallel formats, measuring thousands of cells in a single experiment. Initially developed for the capture of polyadenylated mRNA in eukaryotes, single-cell transcriptional profiling methods have also been extended to bacteria, comparing active and total transcriptomes in single bacterial cells. Furthermore, a variety of experimental platforms have been established for batch mode single-cell RNA sequencing (scRNA-seq) to address specific technical requirements or biological questions. In addition to transcriptome, cellular contents such as genome, epigenome, and proteome can also be measured with single-cell resolution (C. Sierant & Choi, 2018). To achieve a holistic view of cell state, a number of single-cell multi-omics methods have been developed to simultaneously profile different cellular contents within the same individual cell, which is the utmost measure of cellular heterogeneity. This approach greatly expands the capability of single-cell technologies and has been transformative in studying complex systems.

8.Conclusion

Since the first scRNA-seq technology was published in 2009, the single-cell sequencing field has seen rapid development, with new platforms, methods, and applications released each year. Single-cell sequencing technologies can generate transcriptome, genome, and epigenome information from individual cells. With the decreasing cost of high-throughput sequencing and the development of large-scale single-cell analysis software, single-cell sequencing is now an affordable and accessible approach for many labs. There are a few considerations to keep in mind when planning single-cell sequencing experiments, as there are unique aspects of working with single cells compared to bulk sequencing. Careful consideration of these aspects will lead to more robust and reproducible results.

Cancer is a disease of the genome, caused by the accumulation of genetic alterations. However, not all cells in a tumor share the same genome. Recent studies have shown that a wide variety of somatic mutations can be found in individual cells from a tumor, and that some of these mutations affect the expression of proto-oncogenes and tumor suppressors, leading to heterogeneous proliferation rates in cancers that initially appeared histologically homogeneous (K A Sreenivasan et al., 2022). Cell-to-cell variability can also occur in non-genetic factors, such as differences in transcription factor activity, mRNA levels, and protein activities. The existence of such cellular heterogeneity can have profound implications for cancer progression and treatment. Since most targeted therapies attack cancer on a cellular level, differences in cellular responses have been

hypothesized to underlie therapy resistance. In addition to targetable mutations, untreated tumors can evolve to a therapy-resistant state through the acquisition of new mutations, as predicted by a "punctuated" evolutionary model. Recent studies have suggested that some cancers are initially able to resist treatment due to the presence of rare pre-existing resistant cells, in accordance with a "reserve tank" model. Understanding how cellular heterogeneity affects the emergence of therapy resistance will be critical for the design of next-generation cancer therapies (Ye et al., 2016).

References:

Ye, B., Gao, Q., Zeng, Z., M. Stary, C., Jian, Z., Xiong, X., & Gu, L. (2016). Single-Cell Sequencing Technology in Oncology: Applications for Clinical Therapies and Research. ncbi.nlm.nih.gov

C. Macaulay, I. & Voet, T. (2014). Single Cell Genomics: Advances and Future Perspectives. ncbi.nlm.nih.gov

L. Goldman, S., MacKay, M., Afshinnekoo, E., M. Melnick, A., Wu, S., & E. Mason, C. (2019). The Impact of Heterogeneity on Single-Cell Sequencing. ncbi.nlm.nih.gov

Qin, R., Zhao, H., He, Q., Li, F., Li, Y., & Zhao, H. (2022). Advances in single-cell sequencing technology in the field of hepatocellular carcinoma. ncbi.nlm.nih.gov

Hwang, B., Hyun Lee, J., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. ncbi.nlm.nih.gov

Hu, P., Zhang, W., Xin, H., & Deng, G. (2016). Single Cell Isolation and Analysis. ncbi.nlm.nih.gov

Navin, N. & Hicks, J. (2011). Future medical applications of single-cell sequencing in cancer. ncbi.nlm.nih.gov

Müller, S. & Diaz, A. (2017). Single-Cell mRNA Sequencing in Cancer Research: Integrating the Genomic Fingerprint. ncbi.nlm.nih.gov

Y Iourov, I., G Vorsanova, S., & B Yurov, Y. (2012). Single Cell Genomics of the Brain: Focus on Neuronal Diversity and Neuropsychiatric Diseases. ncbi.nlm.nih.gov

H. Nguyen, Q., Pervolarakis, N., Nee, K., & Kessenbrock, K. (2018). Experimental Considerations for Single-Cell RNA Sequencing Approaches. [PDF]

V. H. Hornung, B., Azmani, Z., T. den Dekker, A., Oole, E., Ozgur, Z., W. W. Brouwer, R., C. G. N. van den Hout, M., & F. J. van IJcken, W. (2023). Comparison of Single Cell Transcriptome Sequencing Methods: Of Mice and Men. ncbi.nlm.nih.gov

Zhang, K., Han, X., Li, Y., Yalan Li, S., Zu, Y., Wang, Z., & Qin, L. (2014). Hand-Held and Integrated Single-Cell Pipettes. ncbi.nlm.nih.gov

B. Poirion, O., Zhu, X., Ching, T., & Garmire, L. (2016). Single-Cell Transcriptomics Bioinformatics and Computational Challenges. ncbi.nlm.nih.gov

H. Nguyen, Q., Pervolarakis, N., Nee, K., & Kessenbrock, K. (2018). Experimental Considerations for Single-Cell RNA Sequencing Approaches. ncbi.nlm.nih.gov

Xu, L., Zhang, J., He, Y., Yang, Q., Mu, T., Guo, Q., Li, Y., Tong, T., Chen, S., & D. Ye, R. (2023). ScRNAPip: A systematic and dynamic pipeline for single-cell RNA sequencing analysis. ncbi.nlm.nih.gov

L. Hsu, L. & C. Culhane, A. (2020). Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data. ncbi.nlm.nih.gov

F. Wills, Q. & J. Mead, A. (2015). Application of single-cell genomics in cancer: promise and challenges. ncbi.nlm.nih.gov

Liu, B., Li, Y., & Zhang, L. (2021). Analysis and visualization of spatial transcriptomic data. [PDF]

Chen, T. Y., You, L., Angelito U. Hardillo, J., & Chien, M. P. (2023). Spatial Transcriptomic Technologies. ncbi.nlm.nih.gov

Rui Xing, Q., Omega Cipta, N., Hamashima, K., Liou, Y. C., Gee Koh, C., & Loh, Y. H. (2020). Unraveling Heterogeneity in Transcriptome and Its Regulation Through Single-Cell Multi-Omics Technologies. ncbi.nlm.nih.gov

C. Sierant, M. & Choi, J. (2018). Single-Cell Sequencing in Cancer: Recent Applications to Immunogenomics and Multi-omics Tools. ncbi.nlm.nih.gov

K A Sreenivasan, V., Balachandran, S., & Spielmann, M. (2022). The role of single-cell genomics in human genetics. ncbi.nlm.nih.gov