Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Computational approaches for inferring tumor evolution from single-cell genomic data

Hamim Zafar[1,2], Nicholas Navin[2,3], Luay Nakhleh[1] and Ken Chen[2]

Abstract

Genomic heterogeneity in tumors results from mutations and selection of high-fitness single cells, the operational components of evolution. Precise knowledge about mutational heterogeneity and evolutionary trajectory of a tumor can provide useful insights into predicting cancer progression and designing personalized treatment. The rapidly advancing field of single-cell genomics provides an opportunity to study tumor heterogeneity and evolution at the ultimate level of resolution. In this review, we present an overview of the state-of-the-art single-cell DNA sequencing methods, technical errors that are inherent in the resulting large-scale datasets, and computational methods to overcome these errors. Finally, we discuss the computational and mathematical approaches for understanding intratumor heterogeneity and cancer evolution at the resolution of a single cell.

Addresses

[1] Department of Computer Science, Rice University, Houston, TX, USA
[2] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3] Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding author: Chen, Ken (kchen3@mdanderson.org)

Current Opinion in Systems Biology 2018, **7**:16–25

This review comes from a themed issue on **Genomics and epigenomics (2018)**

Edited by **Raul Rabadan**

For a complete overview see the Issue and the Editorial

Available online 6 December 2017

https://doi.org/10.1016/j.coisb.2017.11.008

Keywords

Single-cell, Genomics, DNA sequencing, Variant detection, Phylogenetics, Intra-tumor heterogeneity, Tumor evolution.

Introduction

Cancer is a disease emerging from a single cell in the somatic tissue and is driven by a complex interplay of somatic mutations, copy number alterations (CNAs) and chromosomal rearrangements [1,2]. As a tumor progresses, diverse genomic aberrations give rise to genetically heterogeneous subpopulations (clones) of cells interacting with each other in a Darwinian framework of mutations, fitness and selection [3–5]. Intratumor heterogeneity (ITH) complicates the diagnosis and treatment of cancer patients and causes relapse and drug resistance [6–8]. The emergence of next-generation sequencing (NGS) technologies enabled a thorough analysis of tumor heterogeneity through the generation of large-scale quantitative genomic datasets [9–11]. However, despite these advances, a comprehensive understanding of ITH has proved elusive thus far [12,13].

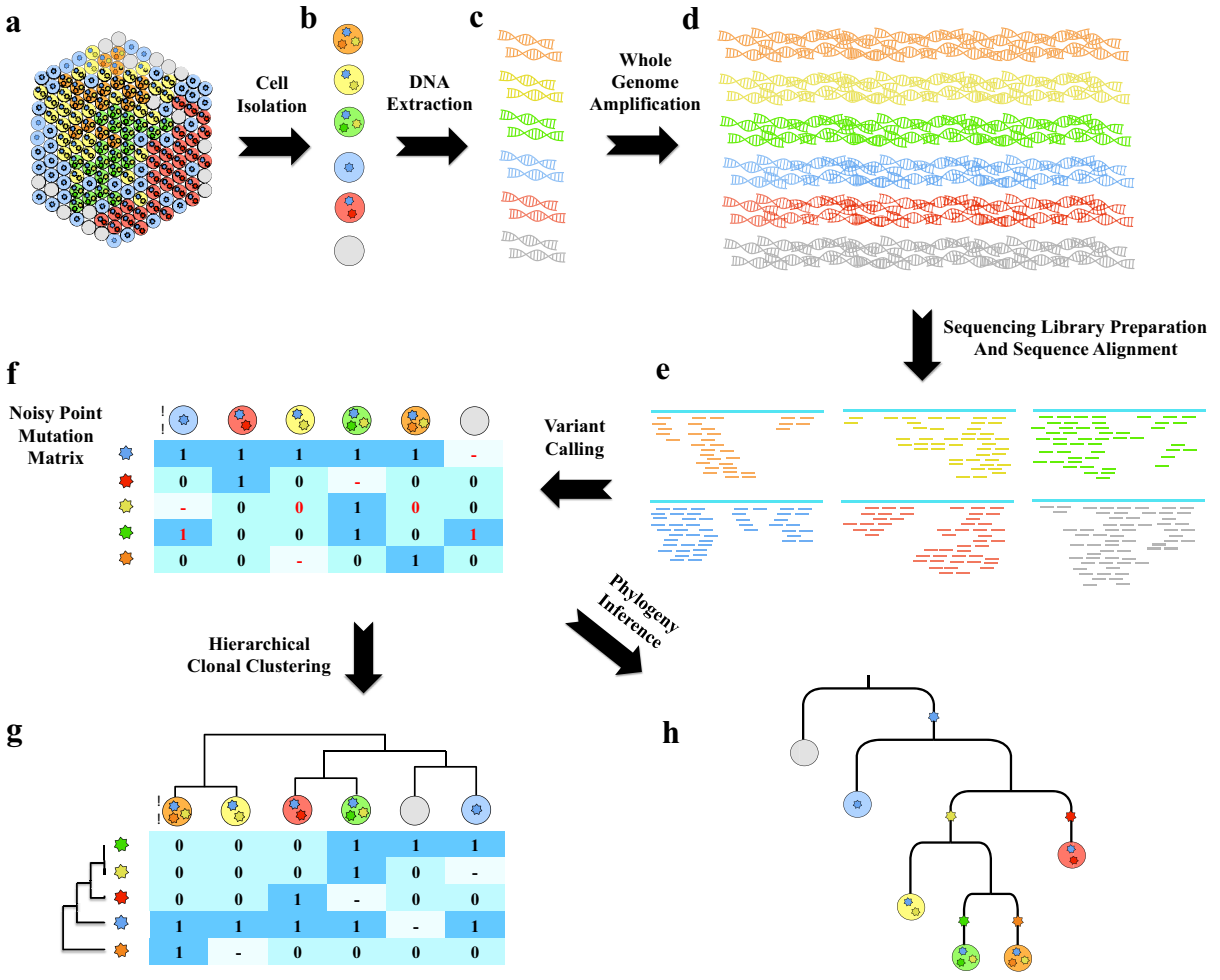
Bulk high-throughput sequencing has been the technology of choice for studying heterogeneity and tumor evolution [14,15]. Subpopulations are computationally inferred [16–22] from variant allele frequencies (VAFs) of mutations detected in bulk DNA that consists of an admixture of DNA from millions of cells in a cancer tissue. VAFs, however, provide a noisy signal for deconvoluting heterogeneity [23,24] and cannot reliably reconstruct rare subclones, or subclones having similar frequencies in the tumor mass. The single-sample approach of bulk sequencing is augmented in multi-region sequencing through which multiple samples obtained from different geographical regions of a tumor are analyzed [25–28]. Although multi-region sequencing can reveal geographically segregated subpopulations, resolving spatially intermixed subclones remains difficult and this approach still relies on deconvolution of subclones for phylogeny inference [29].

The emergence of single-cell DNA sequencing (SCS) technologies has enabled sequencing of individual cancer cells, providing the highest-resolution of the mutational histories of cancer [23,30]. SCS aims to further our knowledge of different aspects of cancer biology including resolving clonal substructure, tracing tumor evolution, identifying rare subclones and understanding the role of cancer microenvironment in tumor progression [23,24,31]. In this review, we discuss the state of the art of SCS technologies, technical challenges and computational approaches to overcome those, and finally, approaches for understanding ITH and tumor evolution from SCS data.

An overview of single-cell DNA sequencing methods

Figure 1 illustrates the steps of a single-cell DNA sequencing study. The first step in producing high-quality SCS data is the isolation of individual cells. Early experiments used techniques such as serial [32] or microwell dilution [33], micropipetting [34], laser-

Figure 1



Overview of single-cell DNA sequencing analysis: from isolation of single cells from a tissue to inference of subclones and phylogeny. (a) Illustration of a heterogeneous tumor tissue, different colors of the cells signify the membership of the cells to different subclones. The mutations present in each cell are represented by the small stars in it. (b–h) steps performed to conduct a heterogeneity or phylogeny analysis of single-cell DNA sequencing data. (b) First, single cells are isolated from tissue and (c) DNA is extracted from each cell. (d) Whole genome amplification (WGA) is performed on DNA extracted from each cell to produce the amount of DNA required for constructing a sequencing library for each individual cell. Various WGA methods are summarized in Refs. [23,75]. (e) Whole-exome or targeted sequencing libraries are constructed depending on the need of the study and sequenced reads are aligned to a reference genome. (f) Variant-calling [82,84,88] is performed on the sequencing library of single cells. Only single-nucleotide variant (SNV) profiles are illustrated here. Copy-number profiles for each cell can also be obtained and utilized in the subsequent steps. (g) Subclones are inferred by clustering [40,93] the cells into different populations. (h) Tumor phylogeny is inferred computationally [108,109,115] from the SNV profiles of single cells. The phylogeny also shows the order of mutations during the evolutionary history of the tumor.

capture microdissection (LCM) [35] to isolate cells from a solid tissue. Several methods [36,37] opted for isolation of single nuclei that remain intact in frozen samples. Later, flow-assisted cell sorting (FACS) [38,39] and microfluidics-based approaches [40] resulted in higher throughput. Scalability to thousands of cells came from barcoding methods [41,42] and single-nucleus DNA repair enabled sequencing of formalin-fixed paraffin-embedded (FFPE) tumor samples [43]. Commercial systems such as CellSearch [44], Magsweeper [45], DEP-Array system [46], CellCelector [47] have been used for the more challenging task of isolating circulating tumor cells (CTCs) and disseminated tumor cells (DTCs).

SCS was made possible by the development whole-genome amplification (WGA) methods that can amplify the 6 pg of DNA in a single-cell genome with a factor of 10^3 to 10^9 [48] to meet the amount of DNA (nanograms-micrograms) required for constructing a sequencing library. Three broad categories (PCR-based, isothermal and hybrid) of WGA methods exist with different advantages and limitations [48–50] (Table 1).

The technical artifacts associated with WGA methods limit the application of SCS. Use of G2/M cells [63] or performing cell lysis and DNA denaturation on ice [64] has improved some of these technical problems. Subsequently, multiplexing approaches coupled with