7

Single Cell Expression Data

In multicellular organisms cells can have different genomes, and distinct cell types have different expression profiles. Humans, for instance, are composed of more than 40 billion cells [53] forming distinct organs, tissues, and cell types. This genetic and transcriptomic variability has important phenotypic consequences. An example of genetic variability is evident in some of our immune cells, T-cells and B-cells, which rearrange and mutate sections of their genome. These mutations and rearrangements lead to a large repertoire of B- and T-cell receptors providing the means to fight the gamut of potential pathogens. Our gametes contain half of the genomic material of somatic cells after carrying out meiotic recombination. Even for two cells that share the same genome, the expression profile can vary dramatically. The changes in expression from a stem cell to terminally differentiated cells are the result of a carefully orchestrated program of cellular differentiation.

As cells transit through different states of differentiation and the cell cycle, different transcription programs are activated and deactivated. A population of cells contains, in general, a representation of a diverse set of transcriptional programs, and expression profiles from these cells represent an average that may not correctly represent the underlying diversity. Single cell RNA sequencing provides the opportunity to accurately map these transcriptional states. In single cell RNA-seq experiments, each cell can be represented by a point in a very high-dimensional space, whose dimension is typically the number of expressed genes (several thousands). Due to the high-throughput nature of the data (measures involving tens of thousands of genes and thousands of cells), single cell analysis requires methods that are able to deal with large amounts of very high-dimensional data. In addition, these methods should preserve the continuous character of the data, as cellular differentiation can be thought of as a

biological continuous process: there is usually a continuous set of states interpolating between a stem-like state and any of the fully differentiated states descending from it.

We will argue in this section that the condensed representations produced by Mapper, applied in the previous chapter to the analysis of cancer cross-sectional data, satisfy these two requirements precisely.

7.1 Introduction to Single Cell Technologies

Recently, single cell sequencing has emerged as a new high-throughput method to access the genome, the epigenome, and the transcriptome of hundreds or thousands of individual cells. These technical developments have been the confluence of several techniques, including the following.

- **Single cell isolation methods**. The first step in sequencing RNA or DNA from single cells is to generate a suspension of single cells, which can be challenging for particular tissues and cell types. Once in suspension, individual cells are isolated by serial dilution, micropipetting, optical tweezers, etc. Although these techniques are effective in isolating single cells, they are not scalable for isolating thousands of single cells. Scalable techniques for single cell isolation remains an active area of research where the most popular techniques include fluorescent activated cell sorting (FACS) and microfluidic devices.
- **Methods for amplification of DNA and RNA from single cells**. A variety of methods have been described to amplify genomic material from single cells, including polymerases from different organisms. For RNA, one of the common techniques for single cell RNA-seq is Smart-Seq, which amplifies full transcripts using a retroviral reverse transcriptase, a switching mechanism at the $5'$ end of the RNA transcript, and then amplifies the CDNA [419]. CEL-Seq uses in vitro transcription as an amplification protocol, avoiding some of the exponential amplification artifacts from PCR [232]. Drop-seq and inDROP are two related but independently developed methods based on micro droplets [302, 338]. Each micro droplet contains a cell barcode and primers together with a captured single cell. The approach allows study of the transcriptome of thousands of cells. Several techniques have been described to amplify DNA material from single cells. PCR based methods (degenerative oligonucleotide PCR, or DOP-PCR) use random or degenerate primers, providing low coverage of whole genomic regions. More popular methods are based on multiple displacement amplification (MDA) using DNA polymerases from a phage ($\Phi 29$)

or, more rarely, from a thermophilic bacterium, *Bacillus stearothermophilus* (Bst polymerase).

There is a frenzied competition in the development of methods for isolating single cells, amplifying RNA and DNA, and sequencing. In the next few years, this will result in dramatic changes in the type of data available, the quality of the data, and the throughput. The availability of single cell data has led to the development of computational methods to study diverse biological processes. Among other things, single cell transcriptomics has enabled more detailed studies of cellular differentiation processes in developmental biology [47, 408, 431, 505] and cancer biology [401, 499]. Single cell analysis has the power to identify different types of minority cells that are eclipsed within larger populations, to identify transitions between different states to draw transcriptional trajectories, and to find specific markers and transcription factors for the different cell types and states.

Ideally, one would like methods for studying single cell transcriptomic data that do not rely on previously known information and thereby allow the discovery of potentially novel biology. The number of cells in these experiments is on the order of thousands which is frequently comparable to the number of genes studied. Recall from the discussion in Chapter 3, to get a good sample of a truly high-dimensional object (here the dimension is the number of genes), one needs a number of points (cells) exponential in the dimension. This is one of the reasons that most approaches to the study of single cell data are based on dramatically reducing the dimensionality of the space through the selection of a few known markers, applying standard dimensionality reduction techniques (e.g., using PCA or *t*-SNE [13]), or looking for specific low-dimensional features (such as reconstructing trajectories or bifurcating points [223, 452]).

An alternative strategy based on ideas from topological data analysis tries to derive a low-dimensional space that can capture some of the biologically interesting properties (such as number of cell types, or trajectories); we use a Reeb graph to describe the data. Recall from Section 1.12 that a Reeb graph is a one-dimensional object that can capture some of the low-dimensional features of the data. As discussed in Section 2.8, Reeb graphs can be approximately inferred from the data using the Mapper algorithm. A generic pipeline for analyzing single cell expression data begins by filtering out low quality cells, based on standard criteria such as the ratio of mapped to unmapped reads, and normalizing the data to account for differences in the length of the transcripts and the total amount of RNA sequenced, as determined by spikes in reads or other methods [483]. The remaining high quality cells are represented as points in a high-dimensional space, of dimension given by the number of different transcripts present in the samples. This

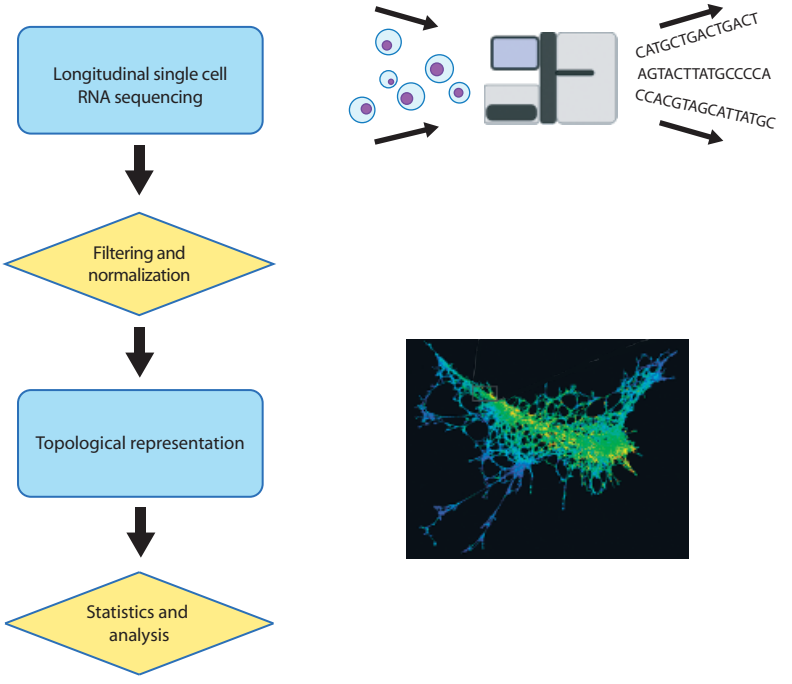

Figure 7.1 Processing of single cell transcriptomic data.

space is endowed with a metric in a standard way, for instance using Pearson's correlation as a measure of similarity. Applying Mapper with various choices of filter function then produces a graph representation; this yields a low-dimensional condensed representation that tries to preserve salient local relations between cells in the high-dimensional space (Figure 7.1).

7.2 Identifying Distinct Cell Subpopulations in Cancer

Our first example of the use of single cell genomic data is in cancer (see Figure 7.2). As we previously discussed, cancers are (among other factors) the result of the accumulation of somatic mutations and epigenetic changes that lead to uncontrolled cell growth. Not all cells in a tumor share the same genetic, transcriptional, epigenetic, morphological, and phenotypic profile, a fact that is usually described as tumor heterogeneity. Two populations of cells that share a dominant clone could have very different phenotypes, as minor populations can be incentivized to grow or become resistant to specific therapies, leading to the long time evolution of these tumors.

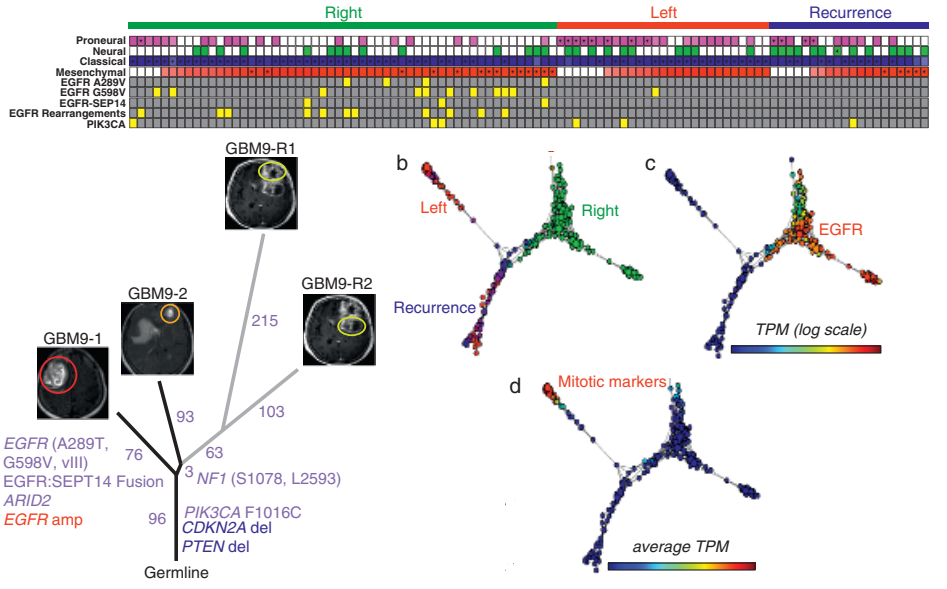

Figure 7.2 Single cell RNA-seq allows the spatial and temporal study of the structure of tumors. This is a particular case of a patient with two focal glioblastomas, on the left and right hemispheres. After surgery and standard treatment, the tumor reappeared on the left side. Genomic analysis (on the left) shows that the initial tumors were seeded by two independent, but related clones. The recurrent tumor was genetically similar to the one on the left. The expression profiles from single cells from the two foci at diagnosis and the relapse recapitulate the clonal history. Transcriptionally and genetically, the recurrence resembles the left parental tumor. A small subset of the cells in the initial left tumor show a similar transcription profile as the recurrent tumor, suggesting that the resistant population originated from a subclonal population in the original tumor. Source: [320]. From Jin-Ku Lee et al., Spatiotemporal genomic architecture informs precision oncology in glioblastoma, *Nature Genetics* 49.4 (2017): 594-599. © 2017. Reprinted with permission from Springer Nature.

Single cell techniques provide the means to study heterogeneous cell populations. The following example studies the mutational and transcriptional profile of a multicentric glioblastoma. Multicentric glioblastomas represent tumors that occur in multiple discrete areas in the brain. In this particular case, at diagnosis, the tumor presented two focal points, on the left and on the right brain frontal lobes. After surgery, chemoradiotherapy, and *EGFR* targeted therapy, the tumor recurred on the left side. Different samples were taken from the initial left and right loci and two samples at recurrence. The history of this tumor was then reconstructed using genomic sequencing from each of the biopsies. The genetic characterization shows that the right tumor shares most but not all genetic alterations with the left tumor, indicating a common origin for the two clones that seeded the left and right tumors.

The two loci at diagnosis show distinct clonal and subclonal alterations, indicating that there were two independent founding clones for each location. The recurrence samples were genetically similar to the original tumor in the same side.

Although the recurrent tumor shared many alterations with the parental tumor in the left section, the recurrent tumor had also acquired other alterations in the course of the progression.

To study this case in further detail, single cell RNA-seq was performed on cells from the two primary tumors and the recurrent tumor. The current standard for classification of glioblastomas based on expression identifies four subtypes, neural, proneural, classical and mesenchymal [515]. When single cells are classified into these four types, the heterogeneity becomes evident, with the right initial tumor being composed of a majority of classical cells. Both the left initial and recurrence tumors showed more heterogeneous cell populations involving three different subtypes (classical, proneural and mesenchymal). This classification does not provide any information on how related the cells responsible for the relapse are to any of the original tumors. All three cell populations show a minority of cells in active cellular division, as indicated by the upregulation of mitotic genes.

Using Mapper, one can appreciate a more continuous structure that recapitulates the clonal and genetic history. The tumor on the right appears to be transcriptionally distinct from the left tumor and the recurrence tumor. Expression profiles from cells in the recurrence tumor resembled the originating initial tumor. This is an important finding, as it shows a continued progression at the expression level, with a few cells at diagnosis having a similar pattern as cells at relapse. It also shows that *EGFR* mutation is a subclonal event, occurring only in the tumor at diagnosis that is not responsible for the relapse. This observation illustrates the problem of clonal heterogeneity for targeted therapies: tumors with heterogeneous populations of cells containing different alterations are less sensitive to specific therapies which target a subpopulation.

7.2.1 Clonal Heterogeneity from Single Cell Tumor Genomics

The recent development of single cell transcriptomics and genomics is providing an opportunity to study the role of clonal heterogeneity in tumors [159, 378, 401] and to identify small, previously uncharacterized cell populations [214]. The single cell approach to studying complex populations brings with it new challenges associated with the large number of sampled genomes. Another rapidly maturing technology in modeling tumor population dynamics is that of patient-derived xenografts. Patient-derived xenografts, or PDX, are generated by transplanting tumor tissue into immunodeficient mice. With different rates of success depending on tumor type and specific samples, these tumors are able to proliferate in the mice, and

they can be passed from one animal to another. While not completely recapitulating tumors in humans with an intact immune system, they capture many in vivo properties of tumors, allowing tumor evolution studies with and without therapy.

Single cell genomics provide the opportunity to understand clonal dynamics in PDX models, connecting different cell populations that are established at different times. Subclones are selected to set up different passages. Eirew et al. studied single-nucleus deep-sequencing from different passages of breast cancer PDX [159]. This study collected single cell data from 55 informative sites from a primary breast cancer tumor and three subsequent mouse passages. These sites were selected using the union of bulk DNA sequencing data across different samples, which excludes, of course, specific alterations in single cells.

Since single cell data from the primary tumor was not available, we generated eight cluster-representative sequences using 27, 36, and 27 single nuclei from the first, second, and fourth passages. From these, we subsampled 3000 trees from all possibilities and projected the data into $\mathbb{P}\Sigma_4$. First, we included trees relating the germline sequence, a randomly selected cluster-representative sequence from the primary tumor, and randomly selected single-nucleus sequences from the initial two xenograft passages. Then, in the second analysis, we included trees relating a randomly selected cluster-representative sequence from the primary tumor and randomly selected single-nucleus sequences from three consecutive xenograft passages.

The results (Figure 7.3) showed consistent linear evolution from primary tumor through the first two xenograft passages. However, significant heterogeneity of tumor clones is observed upon the fourth mouse passage. The first time window (purple) is completely contained within the topology corresponding to linear evolution, unlike the second (gold) which is centered on the origin and extends into all three possible topologies. The point cloud for the second time window displays a higher standard deviation than the first (10.49 versus 8.69), and its centroid is essentially a star tree. The high degree of genotypic heterogeneity giving rise to the second time window distribution is suggestive of a clonal replacement event between the time points of Xenograft 2 (X2) and Xenograft 4 (X4). Many of the prevalent alterations before X4 disappear during the final passage, and many new mutations rise to dominance. This raises interesting questions about the long-term fidelity of PDX vehicles to the genetics of their ancestral primary tumors, which theoretically they serve to mimic.

7.3 Asynchronous Differentiation Processes

One of the most interesting applications of topological data analysis is related to single cell expression profiles along a particular differentiation process. For instance, during the process of differentiation, one can observe how stem cells

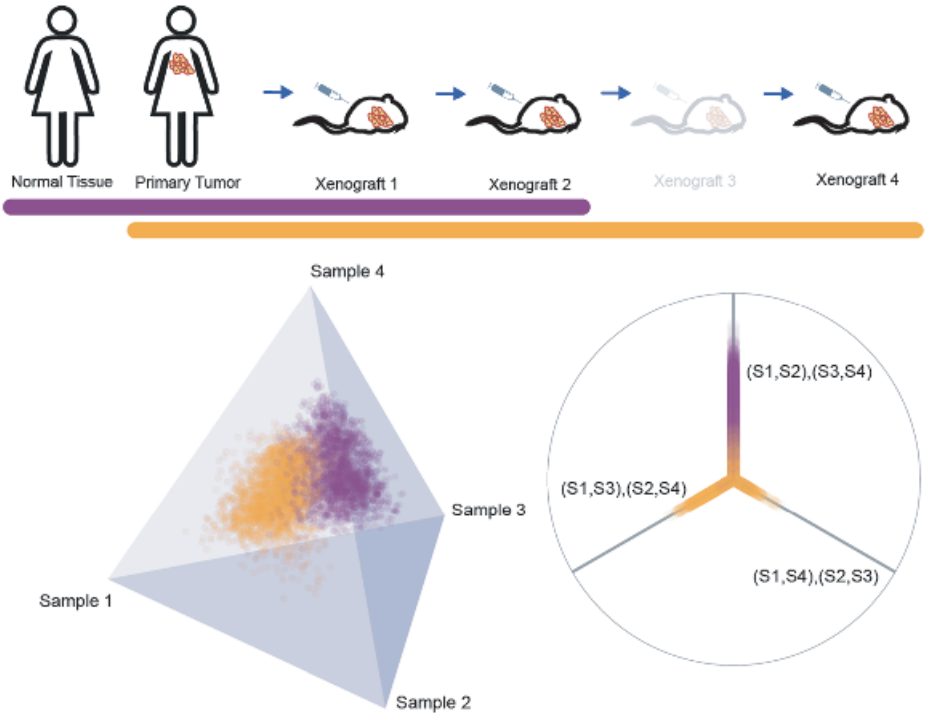

Figure 7.3 Emerging clonal heterogeneity in patient-derived xenograft. Single cell analysis of tumor evolution in a breast cancer derived xenograft model. Single-nucleus deep-sequence data obtained from passages 1, 2, and 4. Single cell data from the primary tumor was not available, however, Eirew et al. identified eight distinct clusters. These data were used to generate two $\mathbb{P}\Sigma_4$ spaces. Source: [545]. From Zairis et al., Genomic data analysis in tree spaces, arXiv: 1607.07503 [9-bio.GN].

evolve into multiple differentiated cells [431]. Single cell RNA-seq samples the transcriptional programs of cells moving along differentiation trajectories. But of course, not all cells move at the same time, and while some retain the characteristics of the original state, others quickly differentiate into final states. In experiments where time information is available, one can organize the process and assign a pseudo-time, so that the transcription data correlates with time. This pseudo-time information is extremely useful as it can organize different transcriptional programs along the differentiation process.

Ideally, one would like to reconstruct evolutionary trajectories in the high-dimensional expression space, and provide a representation that preserves the high-dimensional similarity. One of these representations can be obtained using Mapper. Once the Mapper representation has been established, one can associate a time to different states along the graph (Figure 7.4). We can define a root node as

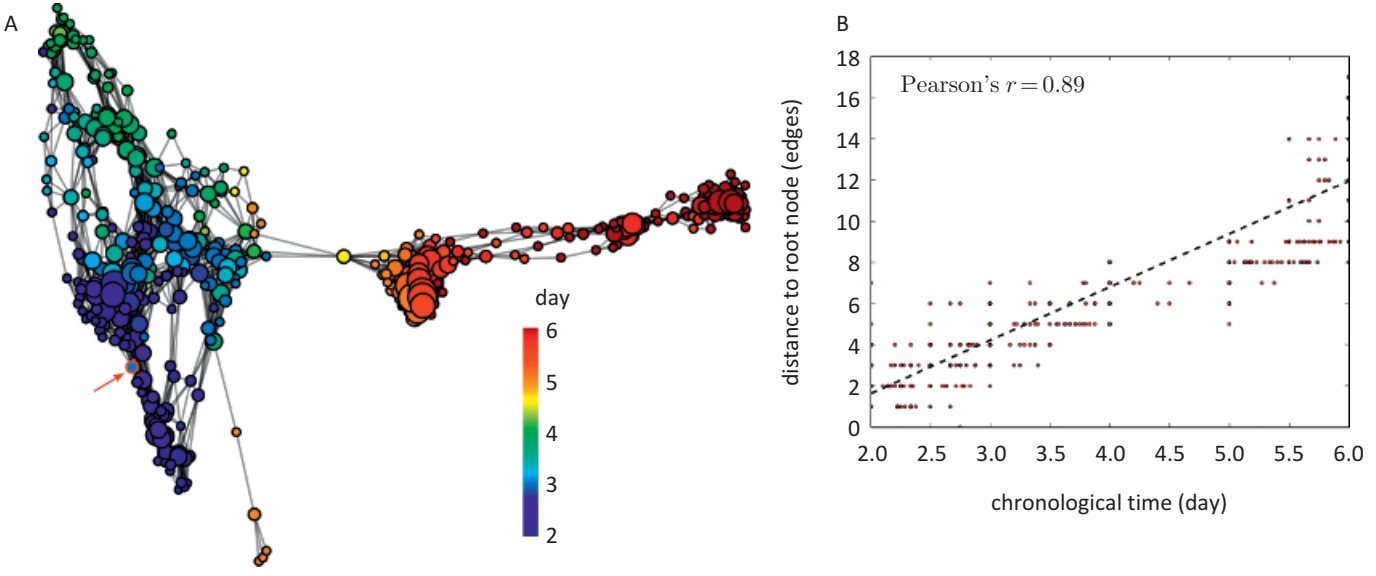

Figure 7.4 When single cells are obtained at different time points one can assign a pseudo-time to each transcriptional state. The pseudo-time orders transcriptional states for the root node to the most differentiated states. Source: [431]. From Abbas H. Rizvi et al., Nature Biotechnology 35, 551-560 (270). © 2017 Nature. Reprinted with Permission from Springer Nature.

the node that maximizes the correlation between the distance in the Mapper graph and time. A pseudo-time can then be inferred by calculating the distance in the graph. In Figure 7.4, the representation is marked with a red arrow. In differentiation, this node corresponds to the most undifferentiated transcriptional state. As expected, the distance along the graph from the root node is associated with the differentiation state.

Different genes are expressed at different stages while others are not expressed or not particularly associated with the progression. One can define the centroid of the expression of a particular gene in the representation to quantify the measure of dispersion of its expression. A relatively simple way to do this while matching to the experimental time is to fit a linear relation between the distance in the representation from the root node to a particular node β, d_β, and the average time of sampling cells associated with that node, $\langle t_\beta \rangle$, (right of Figure 7.4):

$$d_\beta \sim a_0 + a_1 \langle t_\beta \rangle.$$

From there one can define the centroid μ_i for the expression e_i of a particular gene i, measured in time units as:

$$\mu_i = \frac{1}{a_1}\left(\frac{\Sigma_\beta d_\beta e_{i,\beta}}{\Sigma_\beta e_{i,\beta}} - a_0\right).$$

In a similar way, one can define the dispersion, σ_i, as

$$\sigma_i = \frac{1}{a_1}\left(\sqrt{\frac{\Sigma_\beta (d_\beta - a_1\mu_i - a_0)^2 e_{i,\beta}}{\Sigma_\beta e_{i,\beta}}}\right).$$

Centroids and dispersions are a way to assign different transcriptional programs to different differentiation states. In the following, we show how topological data analysis could be used for studying single cell transcriptomic data in differentiation processes.

7.4 Differentiation in Human Preimplantation Embryos

One of the most fascinating biological processes is the development of a metazoon from a single cell: an exquisitely orchestrated organization of transcriptional programs that gives rise to different tissues and cell types, in a particular spatiotemporal fashion. Fertilization occurs with the fusion of parental gametes (egg with a sperm) which creates a zygote. Before the implantation of the embryo in the mother's uterus (six days after fertilization in humans), the original zygote cell undergoes successive replications (Figure 7.5). In the first stages, the zygote divides exponentially (2, then 4, then 8 cells, etc.) to generate the morula. During this process the preimplantation embryo is surrounded by a protein shell called the zona pellucida, that precludes premature attachment to the oviduct walls, where the

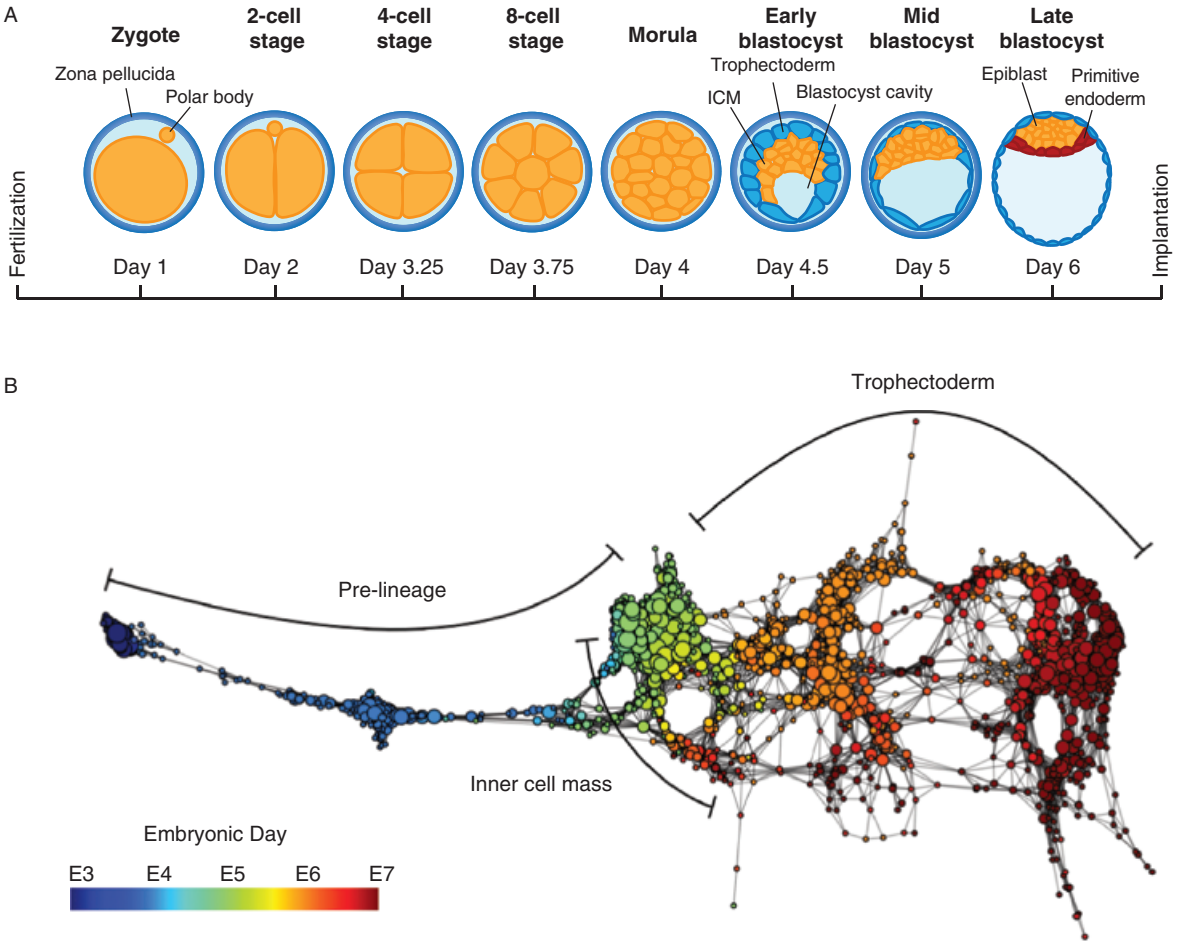

Figure 7.5 Differentiation states of preimplantation human embryos. (A) The development in the preimplantation period in humans lasts a week, where the zygote divides exponentially to generate the morula, which further differentiates into a blastocyst. (B) Mapper graph from single cell data. Source: [431]. From Abbas H. Rizvi et al., *Nature Biotechnology* 35, 551–560 (270). © 2017 Nature. Reprinted with Permission from Springer Nature.

whole process takes place. When there are about 32 cells, the blastomeres generate a cavity by accumulating fluid in the intercellular space. The generation of this cavity generates the blastocyst. Subsequently, the cells on the outside of the blastocyst differentiate into the trophectoderm, induced by the expression of a combination of transcription factors. The cells in the interior of the morula form the inner cell mass that further differentiates into the epiblast and the primitive endoderm. The late blastocyst is composed of three different cell types: the trophectoderm, primitive endoderm, and epiblast. The cells in the trophectoderm lead to the development of and interaction with the placenta, the primitive endoderm forms the amniotic sac where the embryo resides during pregnancy, and the epiblast further differentiates into the three germ layers (endoderm, mesoderm, and ectoderm). Finally, the blastocyst growth disrupts the zona pellucida, leading to the implantation of the zygote into the uterine wall.

To study this process, our final example is a single cell RNA sequencing data set of 1529 cells collected from 88 preimplantation human embryos [408]. The data set captures the process of differentiating embryonic cells at different times and characterizes the segregation between trophectoderm and the inner cell mass lineages. In these examples, multidimensional scaling (MDS) was used as the auxiliary filter function for the condensed representation and Pearson's correlation distance was used as the metric. Analysis of the Mapper graph shows how the cells progress from a highly homogeneous expression pattern corresponding to the morula formation to an intermediate state; this is followed by the establishment of specific transcriptional programs of expression of lineage-specific genes, coinciding with the blastocyst formation. The inner mass cells present a more homogeneous transcriptional program with high expression of embryonic-specific growth factors and receptors (such as TDGF1 and PDGFRA), while the trophectoderm is associated with GATA transcription factor genes expression [431].

This is a nice example of how topological data analysis applied to single cell expression data can recapitulate the history of the first stages of human differentiation. By studying the specific cell populations, one can hope to recover the successive combinatorial transcriptional programs that define this process.

7.5 Summary

The application of topological based approaches to single cell data is at a nascent stage. The technology, methods and many of the ideas reviewed here will be rapidly evolving in the next few years.

- Genomic technologies have recently been applied to single cells to study a diversity of biological problems, including heterogeneity in cancer, mapping transcriptional programs along development, and identification of rare species.

- Evolution in cancer occurs by changes in the genome and the transcriptional states. The spatial and temporal diversity maps to transcriptional states.
- In the differentiation processes, expression profiles of single cells can capture transition states, bridging the differences between the undifferentiated and differentiated populations.
- Topological data analysis methods, such as Mapper, can identify transcriptional profiles and infer the continuous relationship between related states. This is of particular relevance to the study of transition states.
- Topological data analysis can be complemented with temporal information of the biological processes, allowing the identification of different transcriptional states.

7.6 Suggestions for Further Reading, Databases, and Software

- Single cell genomic and transcriptomic studies are relatively recent, and dramatic developments appear almost every other month in both the technology and analysis sectors. Recent reviews worth noting include one by Yong Wang and Nicholas E. Navin [524] and one by Stephen Quake and colleagues [192].
- On the computational side, different approaches for dimensionality reduction have been applied, including multidimensional scaling (MDS), independent component analysis (ICA), and *t*-distributed stochastic neighbor embedding (*t*-SNE). A nice review of these techniques can be found in [483]. As is often the case, the computational techniques are developed within particular applications, including resolving spatial/expression structures [445], studying B-cell development [47], transcriptome dynamics of skeletal myoblasts during differentiation [504], and early development of mouse embryos [340], among many others.
- Interesting applications of single cell genomic and transcriptomic technologies beyond the few examples described in this chapter can be found in lineage decision making [451], understanding tumor heterogeneity in cancer [377], the discovery of new species in the tree of life [430], etc.

The software (and documentation) for analyzing time evolution using single cell data can be found at `http://github.com/RabadanLab/SCTDA`. An online database and exploration tool for some results in neuronal development can be found at `http://rabadan.c2b2.columbia.edu/motor_neurons_tda`.