CS262 Lecture 12 Notes
Single Cell Sequencing
Jan. 11, 2016

Background
A typical human cell consists of ~6 billion base pairs of DNA and ~600 million bases of mRNA. It is time-consuming and expensive to sequence by traditional Sanger sequencing, so deep sequencing on a single cell yields a more effective low-cost approach. In 2013, Nature named single cell sequencing as the method of the year in 2013, and publications and data using single-cell sequencing have increased rapidly in the past two years.

Applications
- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology

Developmental Biology
Life begins from a single cell, which – through subsequent mitotic cellular divisions – generates a population of cells that makes up the human body.

Use of single-cell sequencing:
- Elucidate how a stem cell responds to its environment and how it chooses a specific lineage during differentiation.
- Shed light on the genetic, epigenetic, and expression variations unique to specific cells that elicit a unique phenotype.
- Uncovers transcription factors, RNA species, receptors and ligands that direct lineage specification during development in a sequential, tissue-specific order, identifying and tracing back cell lineages.

Cancer Biology
Tumor progression occurs through driver mutations that undergo Darwinian selection for successive clonal expansion of tumor subclones. Thus, advanced tumors may contain a number of clones where ach clone has a unique set of mutations, pathology, and drug response. Conventional deep sequencing of the bulk tumor can only capture about 1% of the cell population.

Use of single-cell sequencing:
- Find evidence for models of cancer.
- Infer timing of mutations and the drivers.
- Evaluate effectiveness of targeted therapy.

Microbiology
Single cell resolution provides the ultimate resolution of a microbial population since it allows for the identification of low-abundance species that may be missed in population sequencing.

Use of single-cell sequencing:

- Characterize organisms that are difficult to culture in vitro.
- Improve detection and analysis of outbreaks, antibiotic-resistant strains, food-borne pathogens, and microbial diversities in the environment or the gut.

Neurology

Recent advances in research have highlighted the mosaic genomes of individual neurons.

Use of single-cell sequencing:

- Study the mosaic genomes of individual neurons and compositions in the brain.
- Follow genetic variations during fetal development.
- Develop targeted therapy for neurological diseases for specific cell types.

Single-Cell Technologies

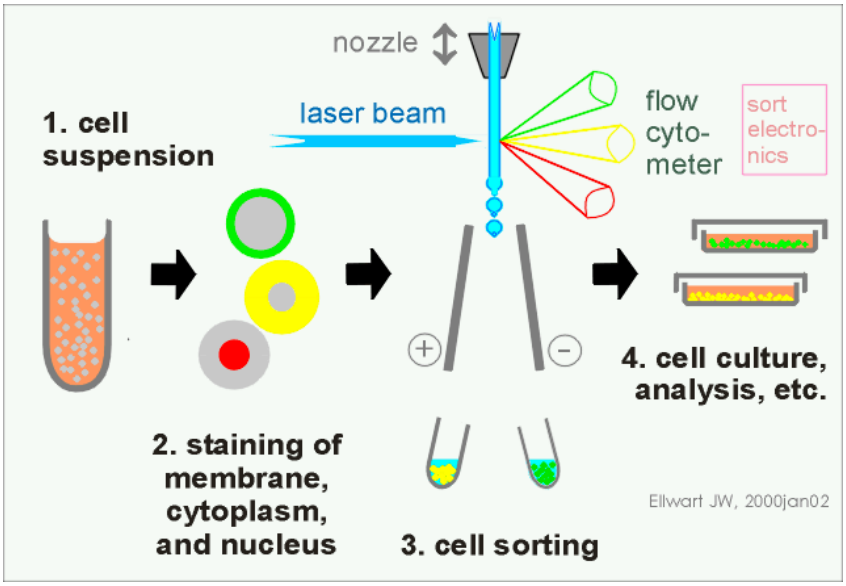
i. Isolate single cells
ii. Amplify genome efficiently
iii. Sequence DNA

Isolation of single cells

No current standardized technique for single-cell isolation.

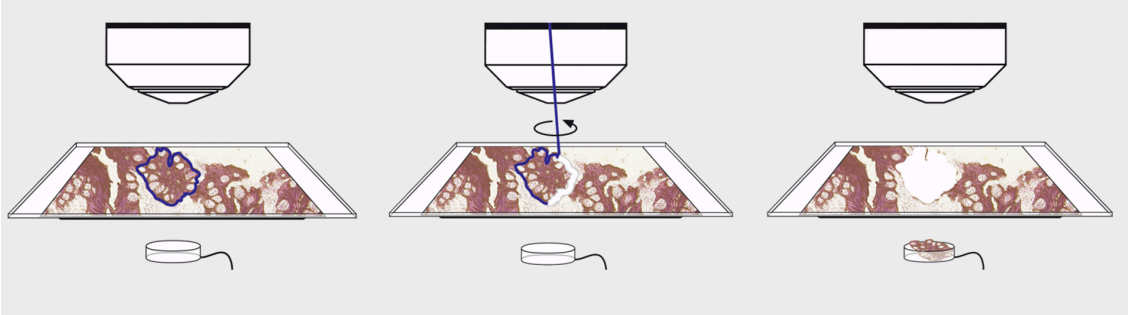
Methods for cell sorting:

i. FACS (fluorescence-activated cell sorting)
   a. Method for sorting a heterogeneous mixture of biological cells into two or more containers, one cell at a time, based on the specific light scattering and fluorescent characteristics of each cell.



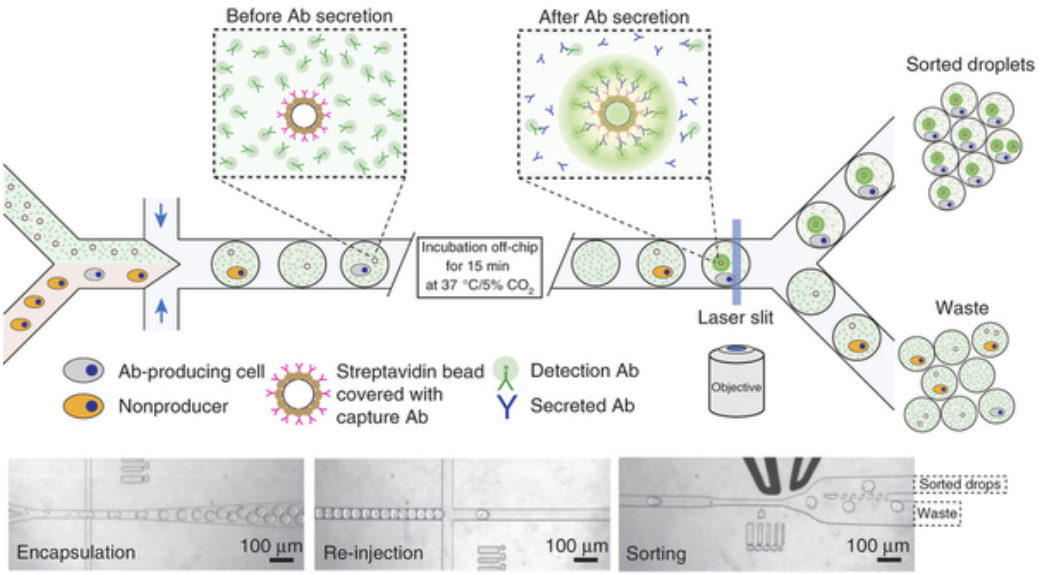
ii. LCM (laser-capture microdissection)

a. Method to procure subpopulations of tissues cells under direct microscopic visualization from a wide variety of tissue samples.



iii. Microfluidics

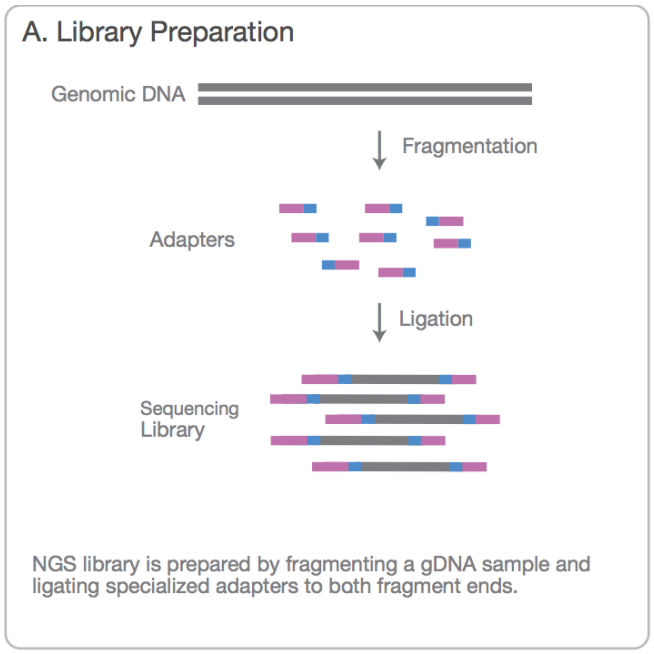
a. High throughput system for sorting; can isolate rare circulating cells.



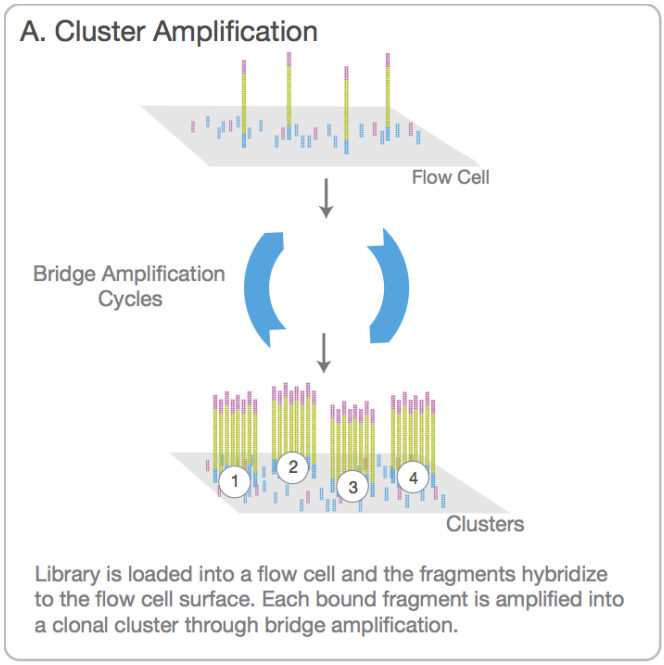
Amplification and Sequencing

Next Generation Sequencing (NGS):

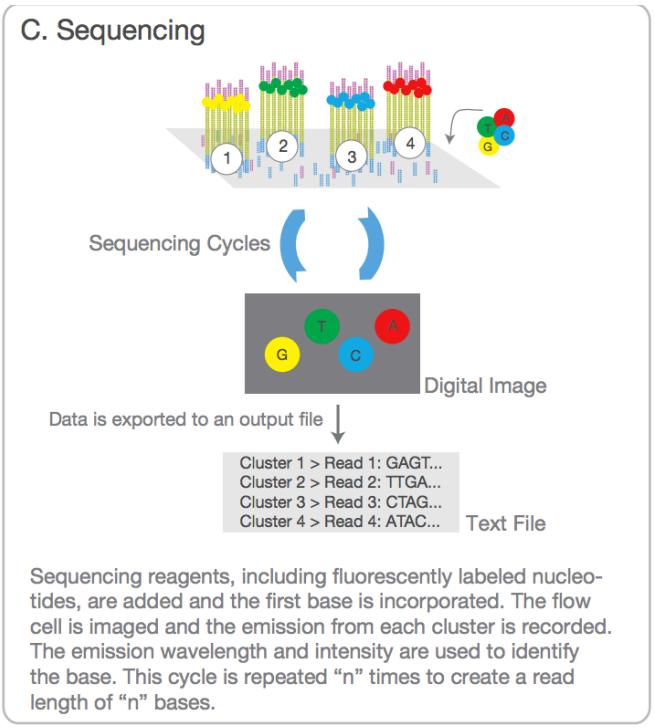
- DNA polymerase catalyzes the incorporation of fluorescently labeled dNTPs into a DNA template strand during sequential cycles of DNA synthesis. During each cycle, the nucleotides are identified using fluorophore excitation. Instead of sequencing a single DNA fragment, NGS extends this process across millions of fragments in massive parallel runs.
- Illumina sequencing by synthesis (SBS) chemistry is the most widely adopted chemistry in the industry.
- Illumina NGS workflow:
    1. Library preparation



2. Cluster generation



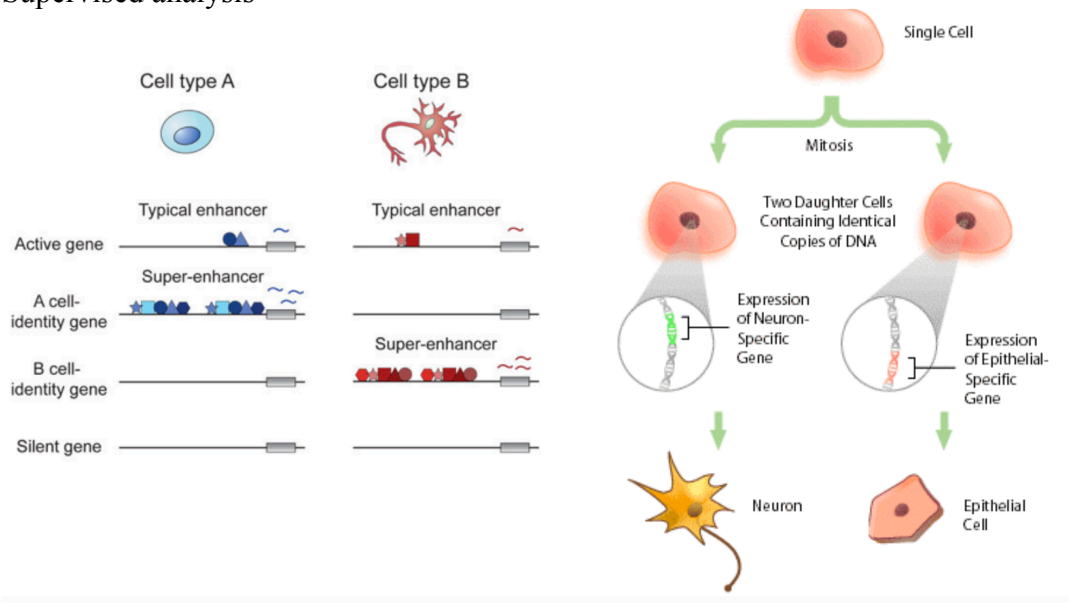
3. Sequencing



4. Data analysis

Cell population identification

- Supervised analysis



- Unsupervised analysis

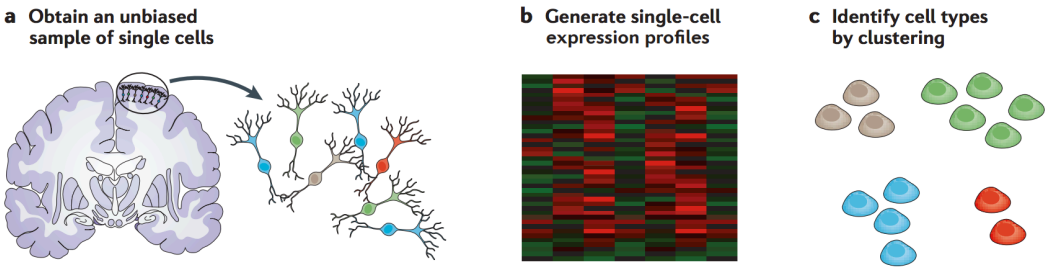


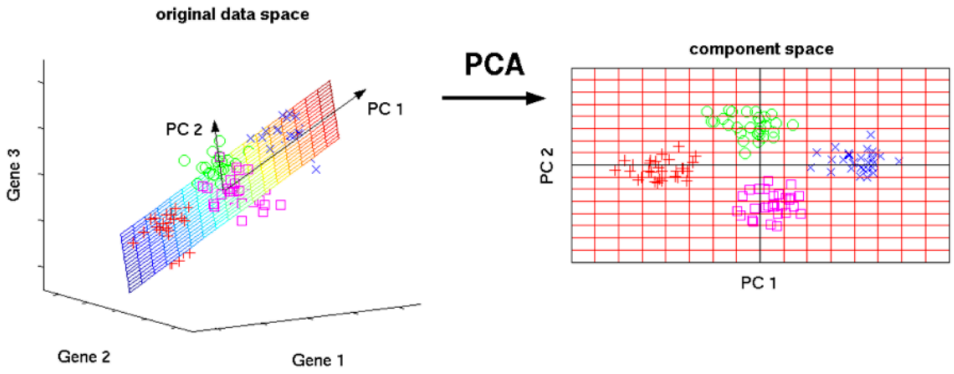
Figure 3 | **Cell-type discovery by unbiased sampling and transcriptome profiling of single cells.**

- **Cluster Analysis**
  - Group objects in "clusters" according to their similarity
  - **Hierarchical clustering (connectivity model)**
    - Build a hierarchy of clusters
    - Agglomerative "bottoms up" approach: each data point starts in its own separate cluster and pairs of clusters are gradually merged until one big cluster is left.
    - Distance metric: i.e. Euclidean distance
    - Linkage criterion: specifies the dissimilarity of clusters as a function of the pairwise distances of observations
      - Complete linkage
        - The maximum of all pairwise distances between points in the two clusters.
      - Single linkage

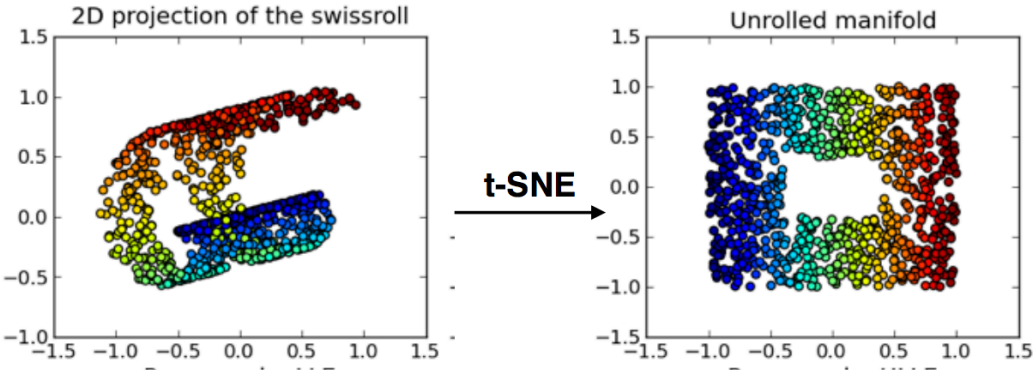
        - The minimum of all pairwise distances between points in the two clusters.
    - Average linkage
        - Each cluster c_i is associated with a mean vector μ_i, which is the mean of all the data items in the cluster. The distance between two clusters c_i and c_j is then just $d(\mu_i, \mu_j)$.
- **K-means algorithm (centroid model)**

Dimension Reduction

- Feature extraction: create a subset of new features by combinations of the existing features.
- **Principal Components Analysis (PCA)** – for linear feature extraction
    - Objective: perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible.
    - The optimal approximation of a random vector $x \in R^N$ by a linear combination of M (M < N) independent vectors is obtained by projecting the random vector x onto the eigenvectors φ_i corresponding to the largest eigenvalues λ_i of the covariance matrix $\sum_x$.
    - Main limitation: PCA does not consider class separability since it does not take into account the class label of the feature vector.



- **t-distributed stochastic neighbor embedding (t-SNE)**
    - Nonlinear dimensionality reduction that typically embeds high-dimensional data into a space of two or three dimensions.
    - Two main stages of the t-SNE algorithm:
        1. Construct a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked.
        2. Define a similar probability distribution over the points in the low-dimensional map, and minimizes the Kullback-Leibler divergence between the two distributions with respect to the locations of the points in the map.
            a. $Kullback - Leiber\ divergence: KL(P||Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}}$



References

llumina. An Introduction to Next-Generation Sequencing Technology. http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

Illumina. Single Cell Research. https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/single-cell-sequencing-research-review.pdf.