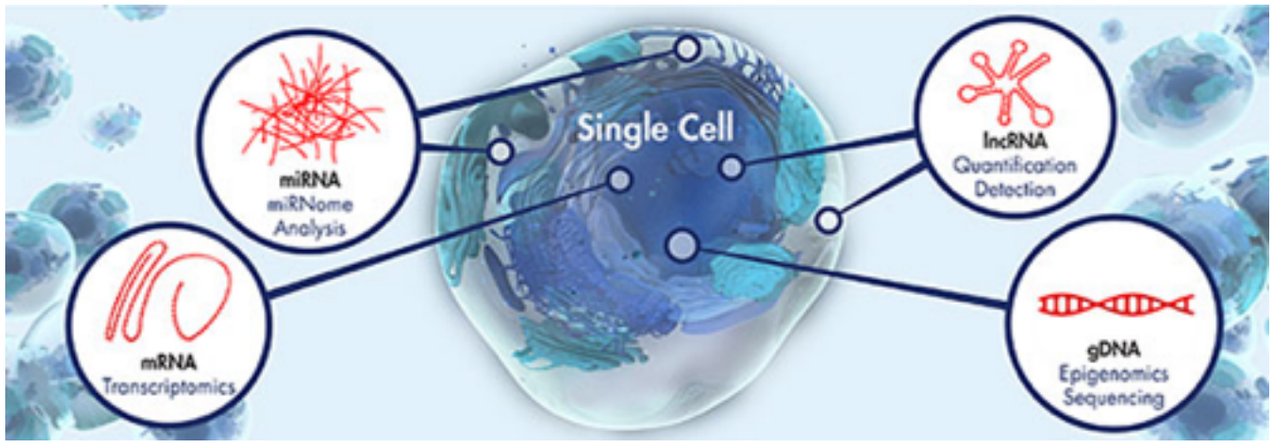
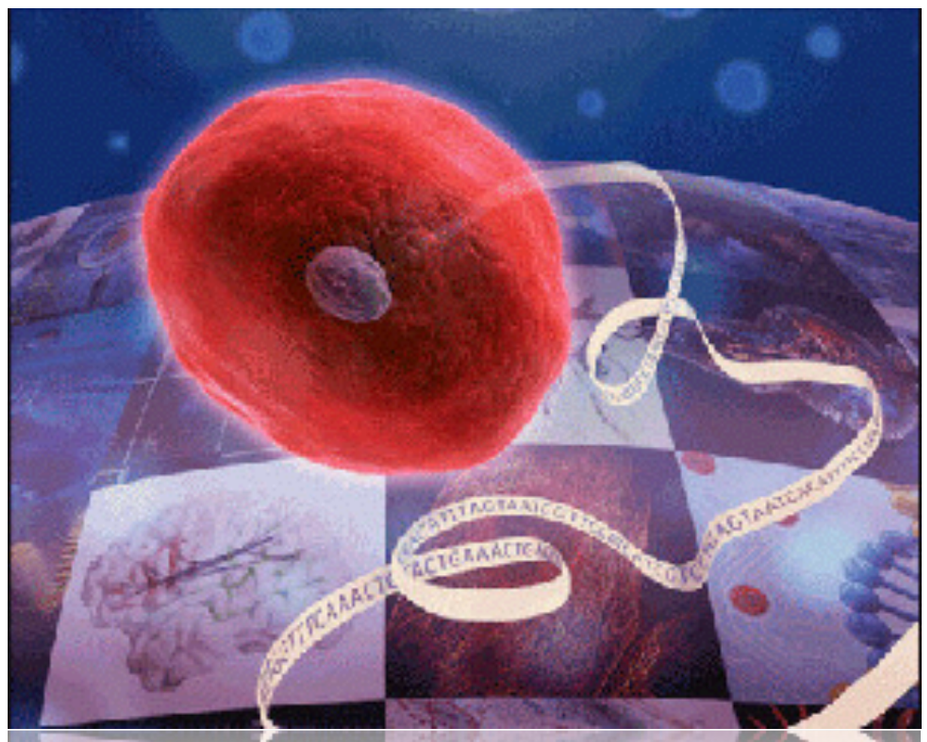
CS262 Winter 2016

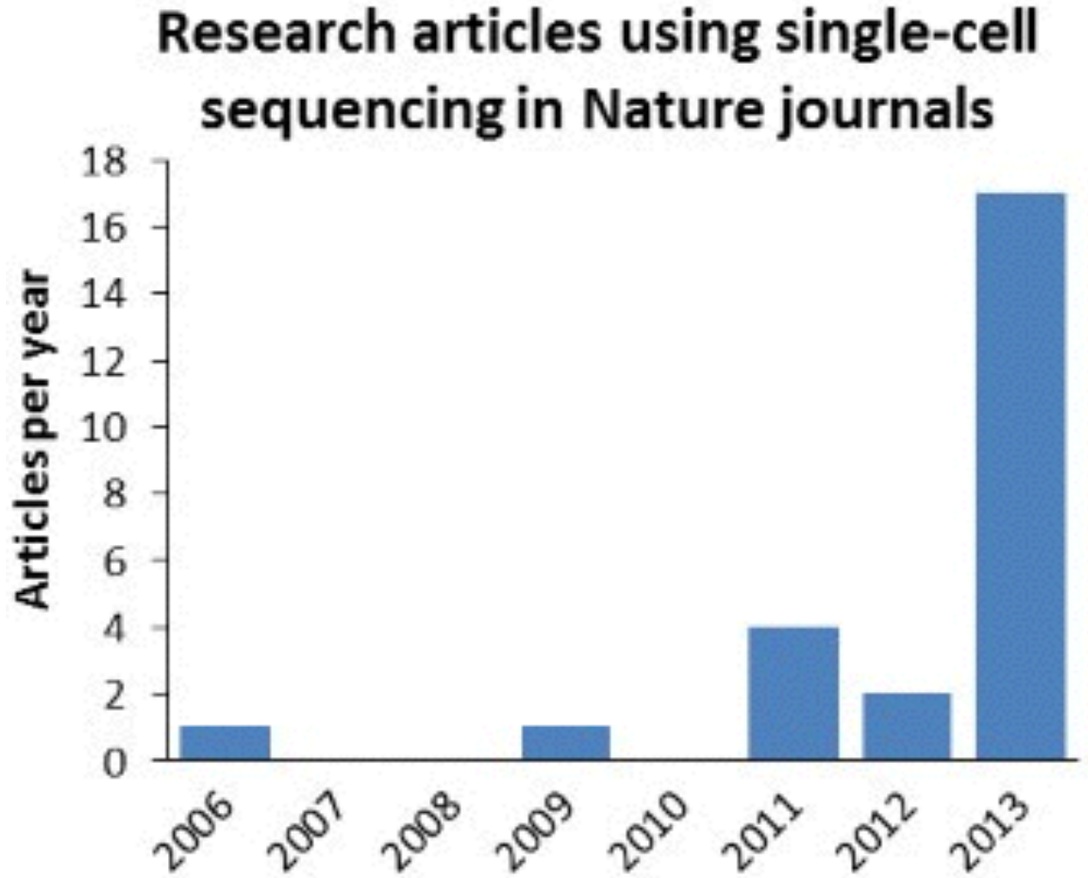
Single Cell Sequencing

Background

NATURE METHODS | METHAGORA







significant increase in publications and data in the last two years

Background

Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing

Mei-Chong Wendy Lee[a,1], Fernando J. Lopez-Diaz[b,1], Shahid Yar Khan[a,2], Muhammad Akram Tariq[a,3], Yelena Dayn[c], Charles Joseph Vaske[d], Amie J. Radenbaugh[a], Hyunsung John Kim[a], Beverly M. Emerson[b,4], and Nader Pourmand[a,4]

FOCUS

CANCER

Cell

Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation

Aleksandra A. Kolodziejczyk,[1,2,5] Jong Kyoung Kim,[1,5] Jason C.H. Tsang,[2] Tomislav Ilicic,[1,2] Johan Henriksson,[1] Kedar N. Natarajan,[1,2] Alex C. Tuck,[1,3] Xuefei Gao,[2] Marc Bühler,[3] Pentao Liu,[2] John C. Marioni,[1,2,4,*] and Sarah A. Teichmann[1,2,*]

Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations

Nicola K. Wilson,[1,9] David G. Kent,[1,9] Florian Buettner,[2,9] Mona Shehata,[7] Iain C. Macau
Manuel Sánchez Castillo,[1] Caroline A. Oedekoven,[1] Evangelia Diamanti,[1] Reiner Schult
Thierry Voet,[3,6] Carlos Caldas,[7] John Stingl,[7] Anthony R. Green,[1] Fabian J. Theis,[2,8] and

ARTICLE

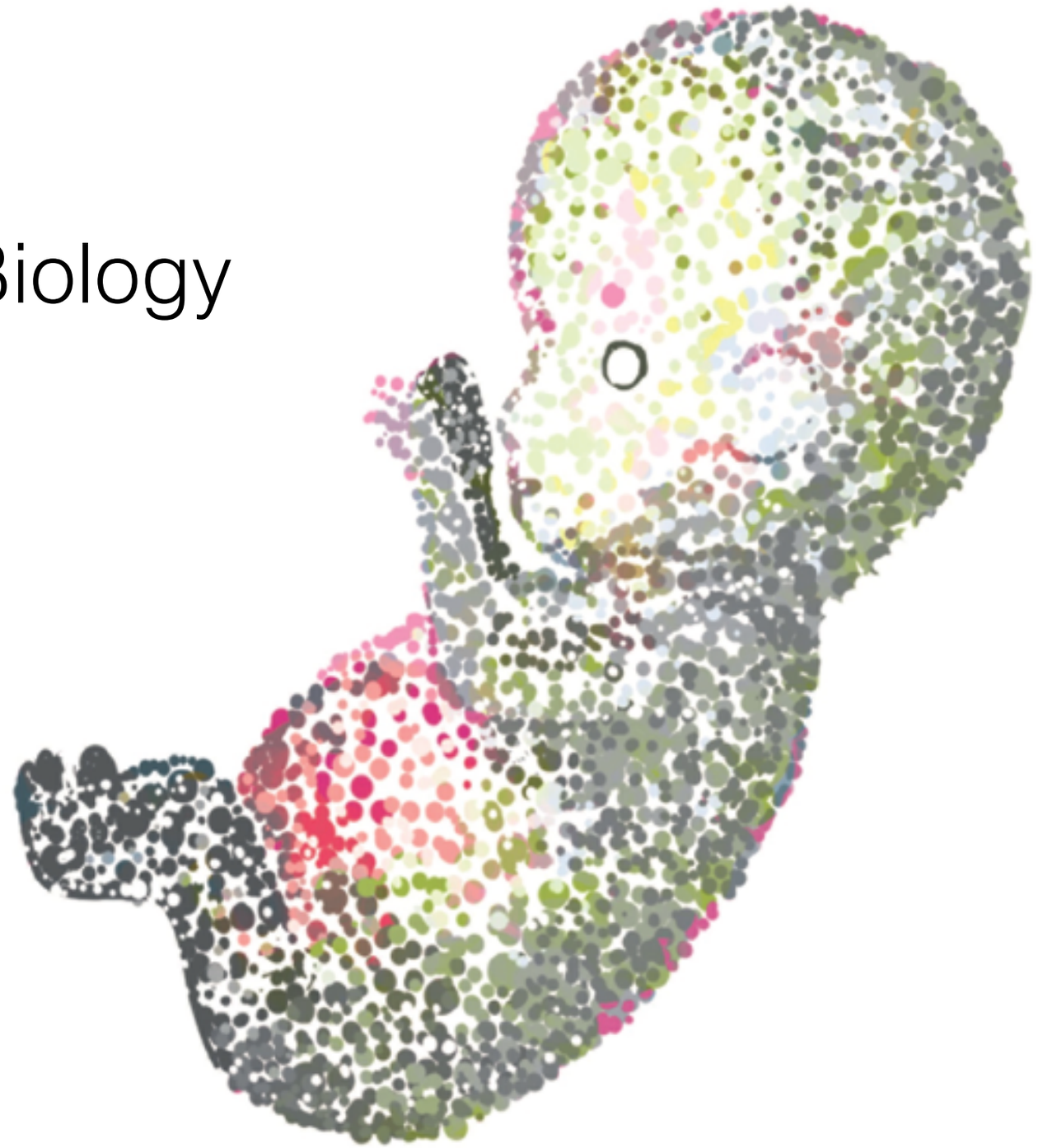
NATURE

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

Amit Zeisel,[1*] Ana B. Muñoz-Manchado,[1*] Simone Codeluppi,[1] Peter Lönnerberg,[1] Gioele La Manno,[1] Anna Juréus,[1] Sueli Marques,[1] Hermany Munguba,[1] Liqun He,[2] Christer Betsholtz,[2,3] Charlotte Rolny,[4] Gonçalo Castelo-Branco,[1] Jens Hjerling-Leffler,[1†] Sten Linnarsson[1†]

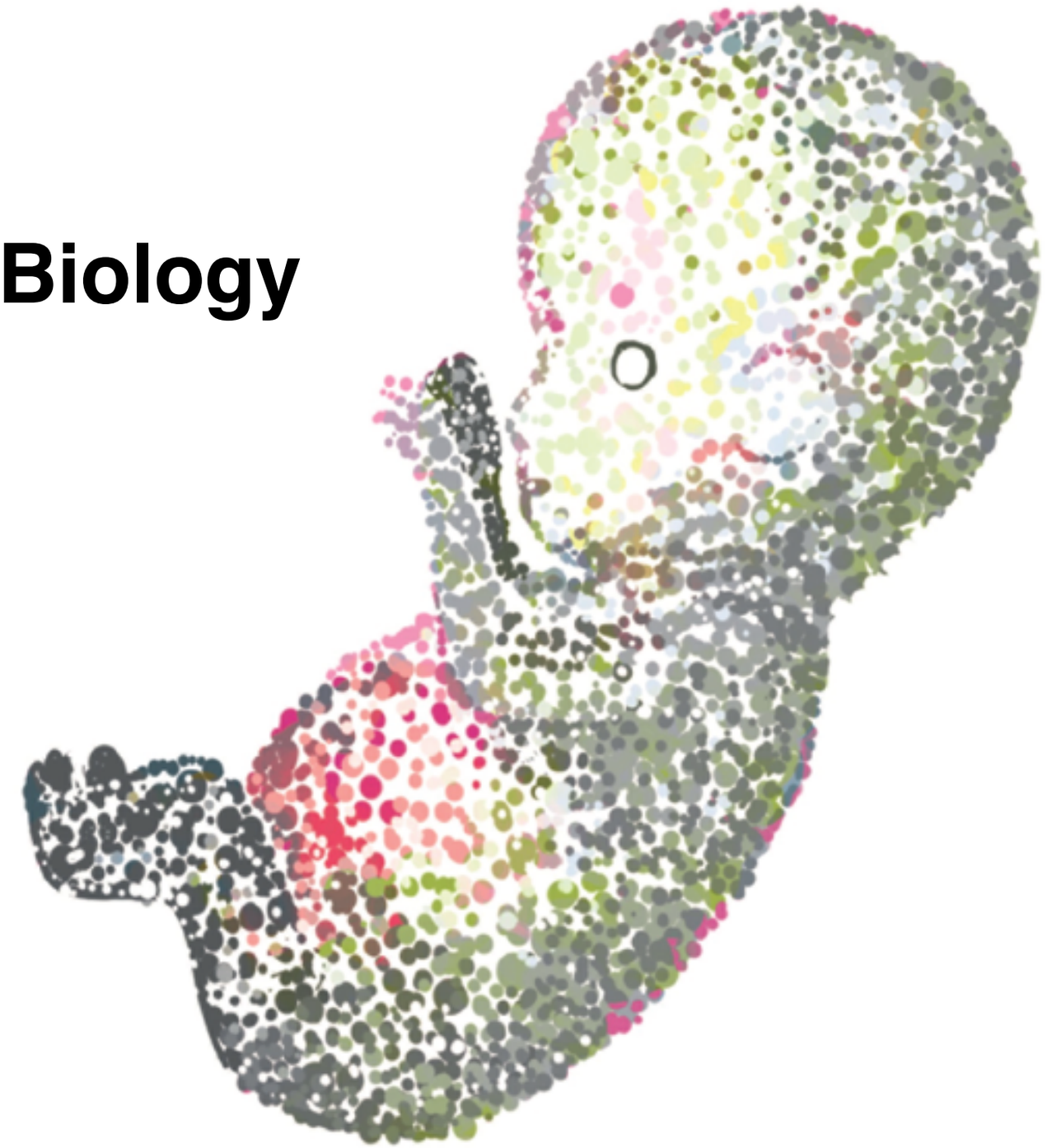
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology



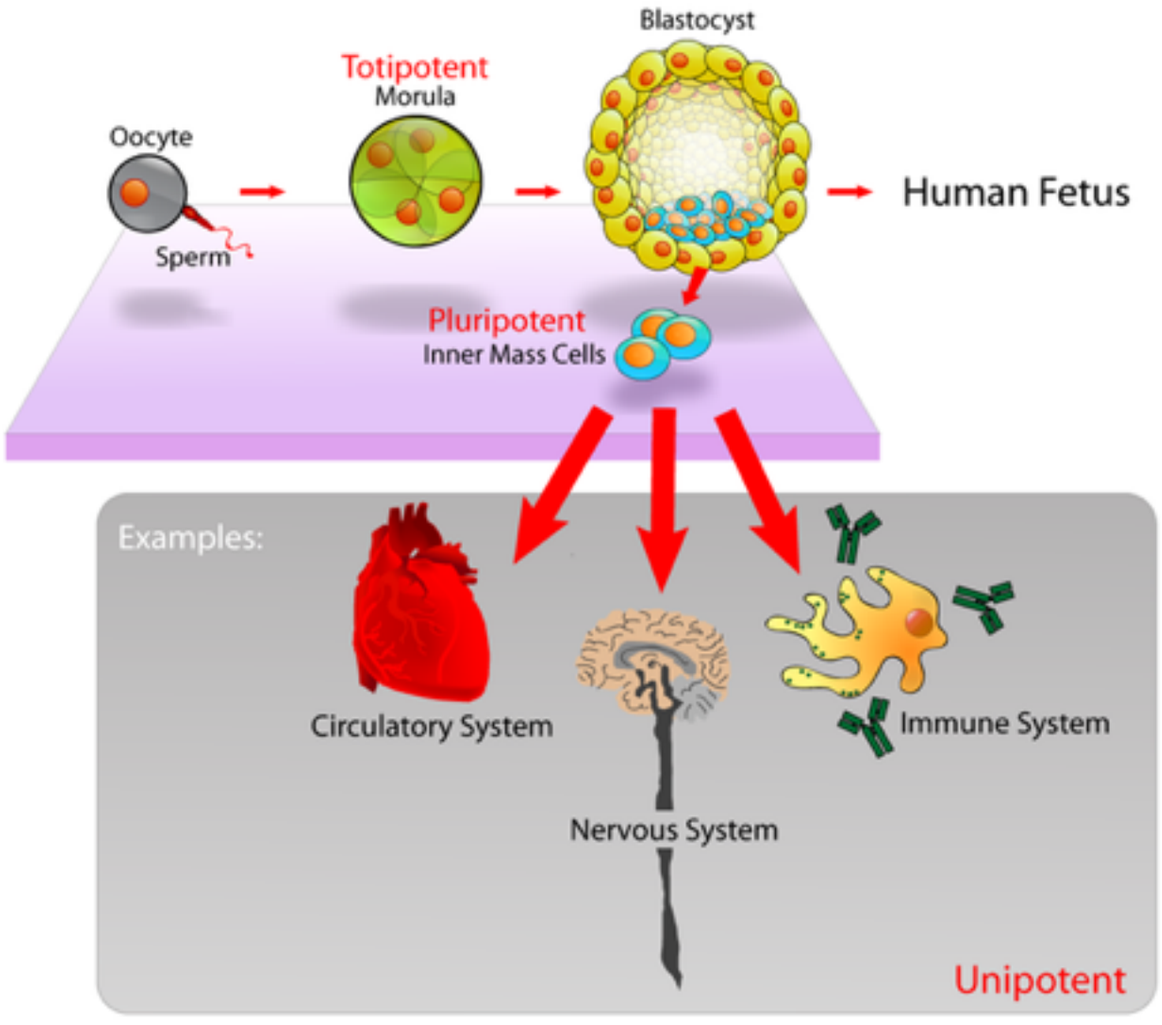
Applications

- **Developmental Biology**
- Cancer Biology
- Microbiology
- Neurology

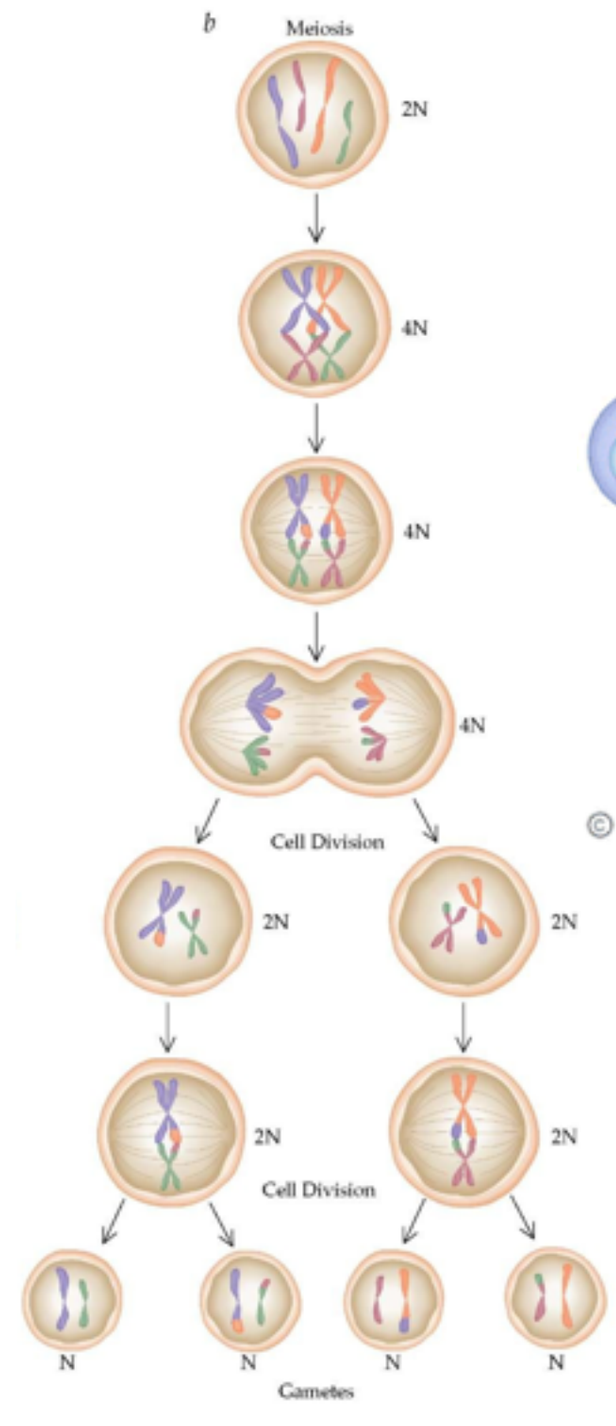


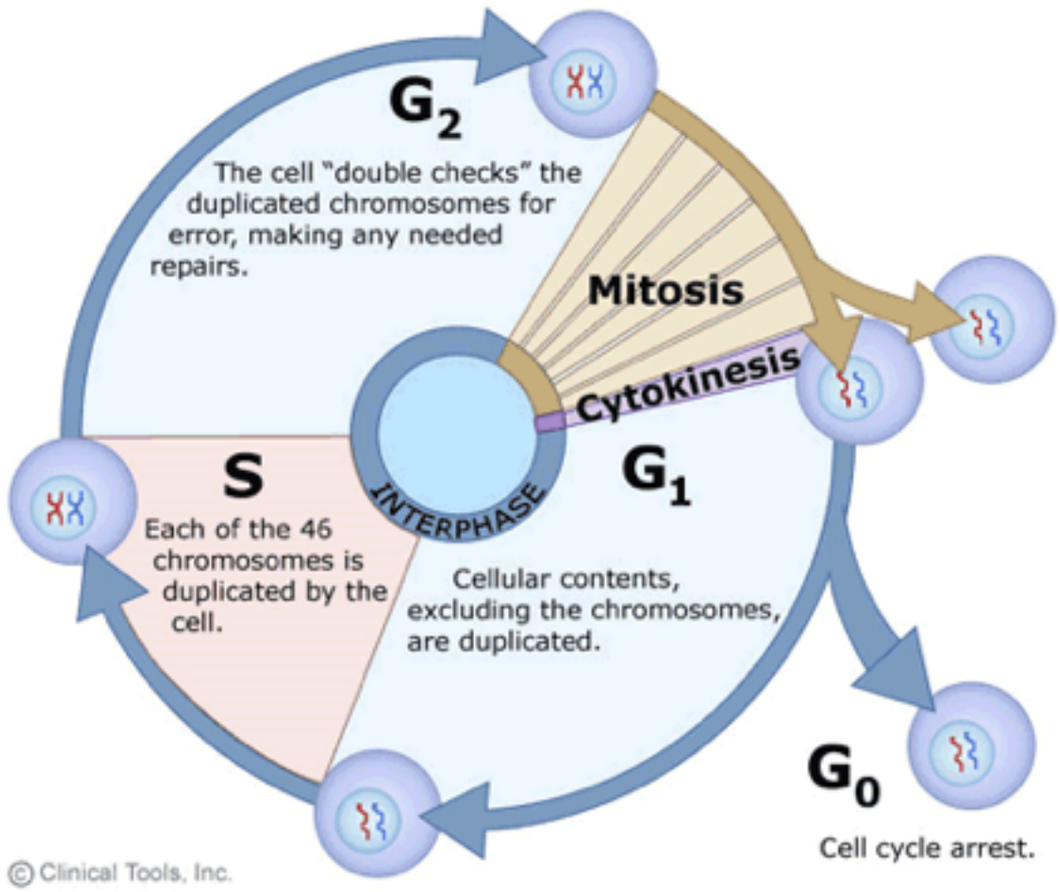
Developmental Biology

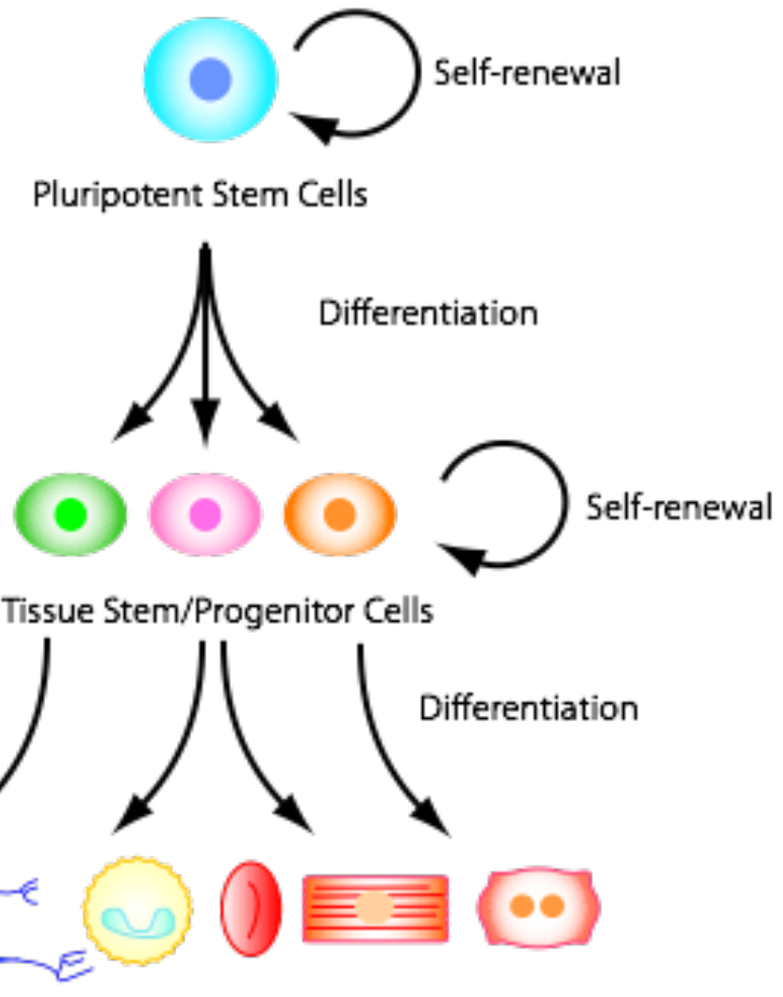
How do animals grow and develop from a single cell?

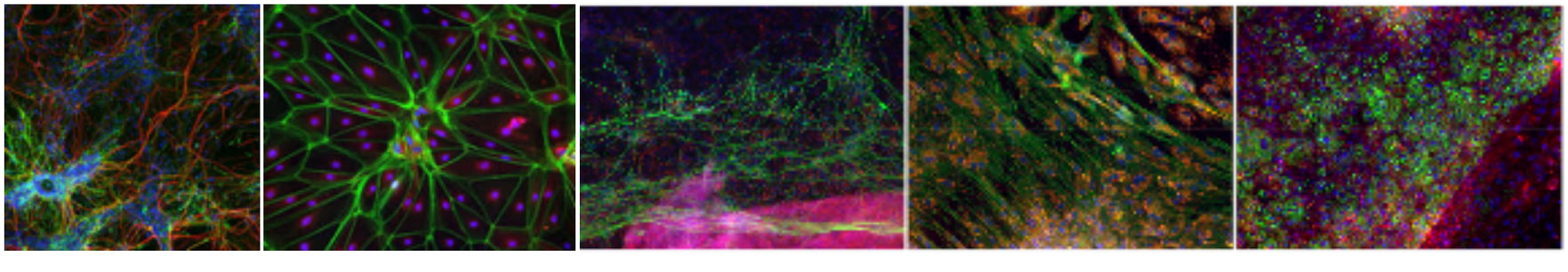


Developmental Biology









Various cell types

Developmental Biology

We need single-cell resolution to:

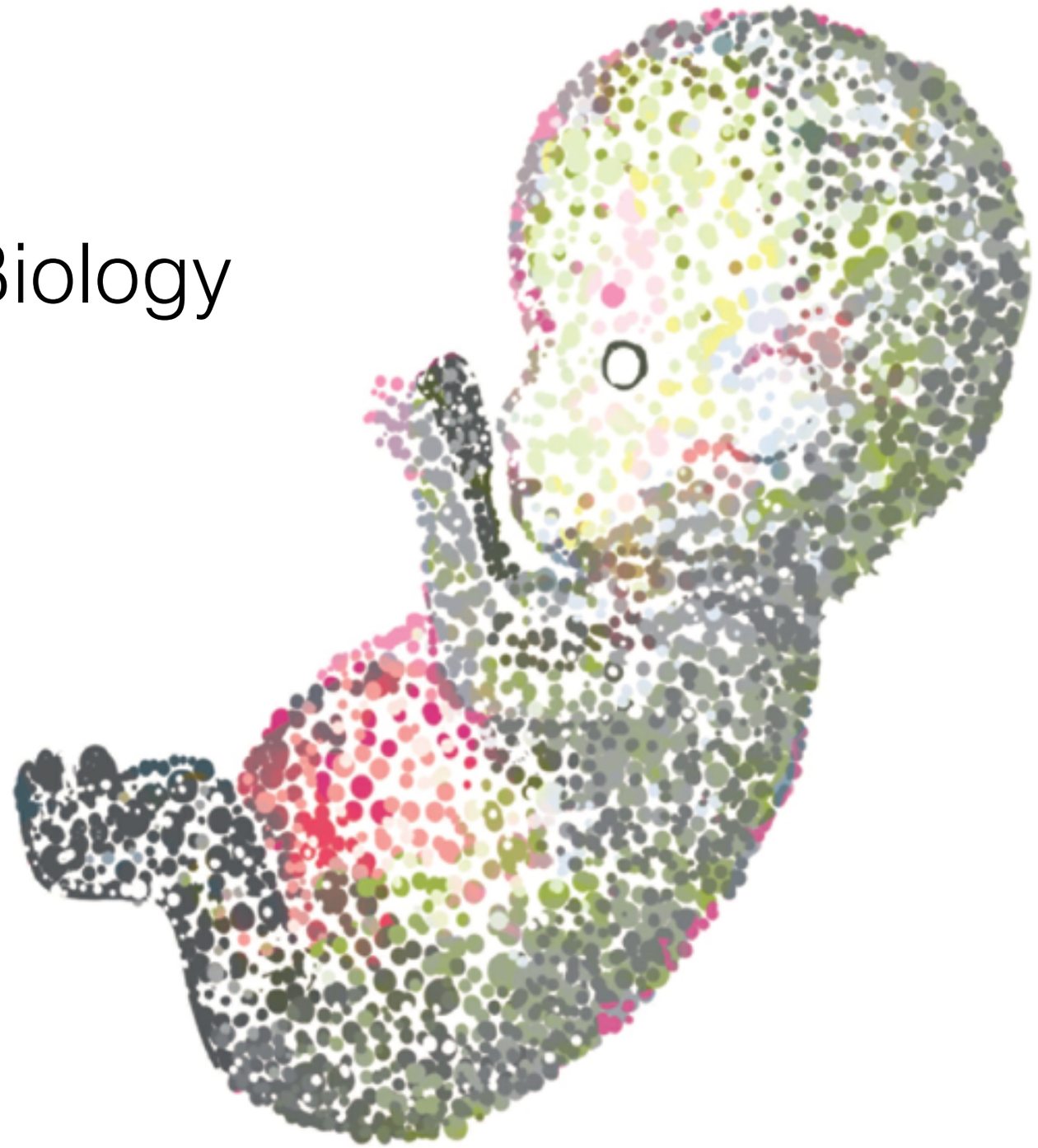
- Discover more complicated mechanisms in cellular development

- Confirm the distinct gene expression signatures across different cell types

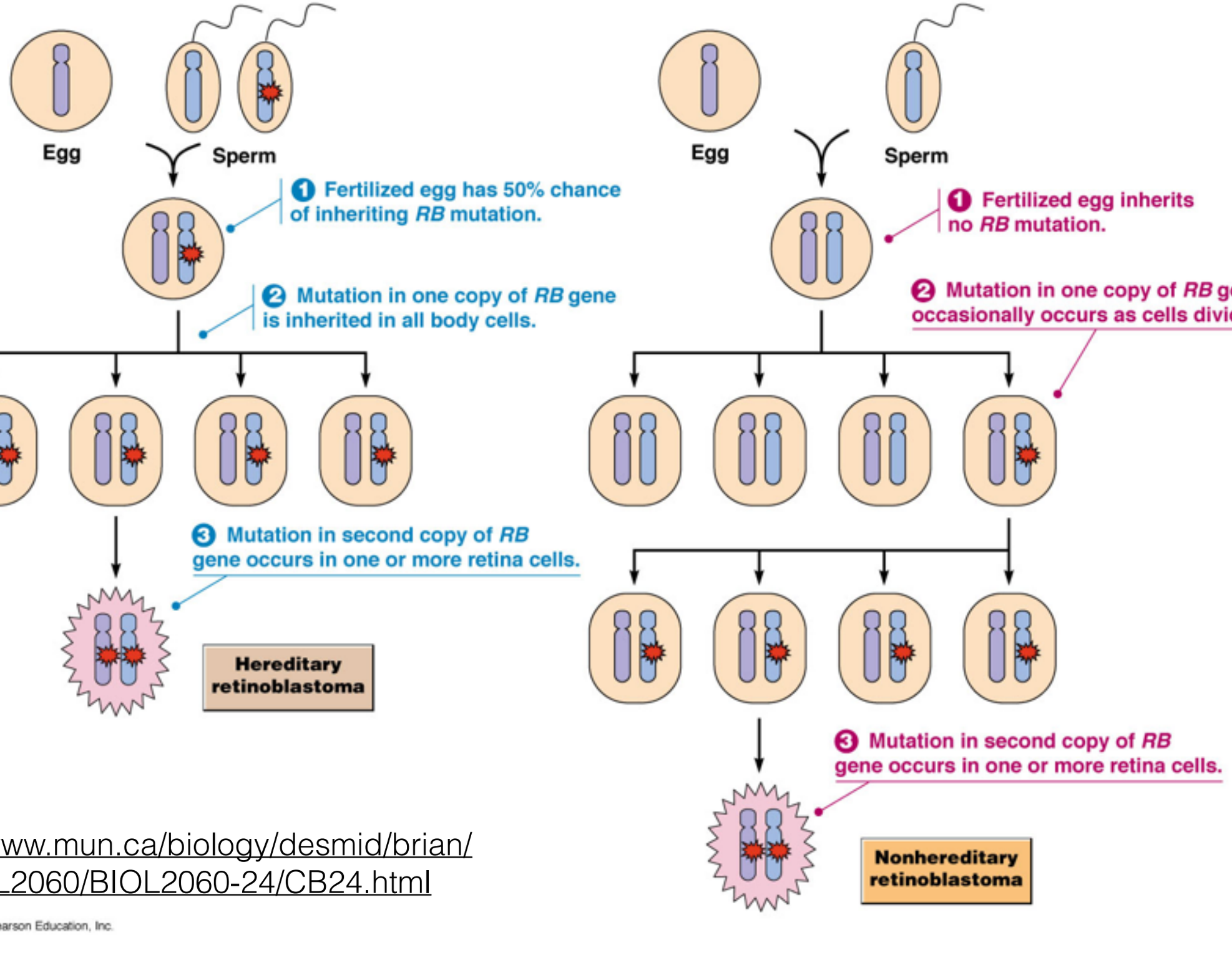
- Identify functional differences among the same cell cell type

Applications

- Developmental Biology
- **Cancer Biology**
- Microbiology
- Neurology



Cancer Biology

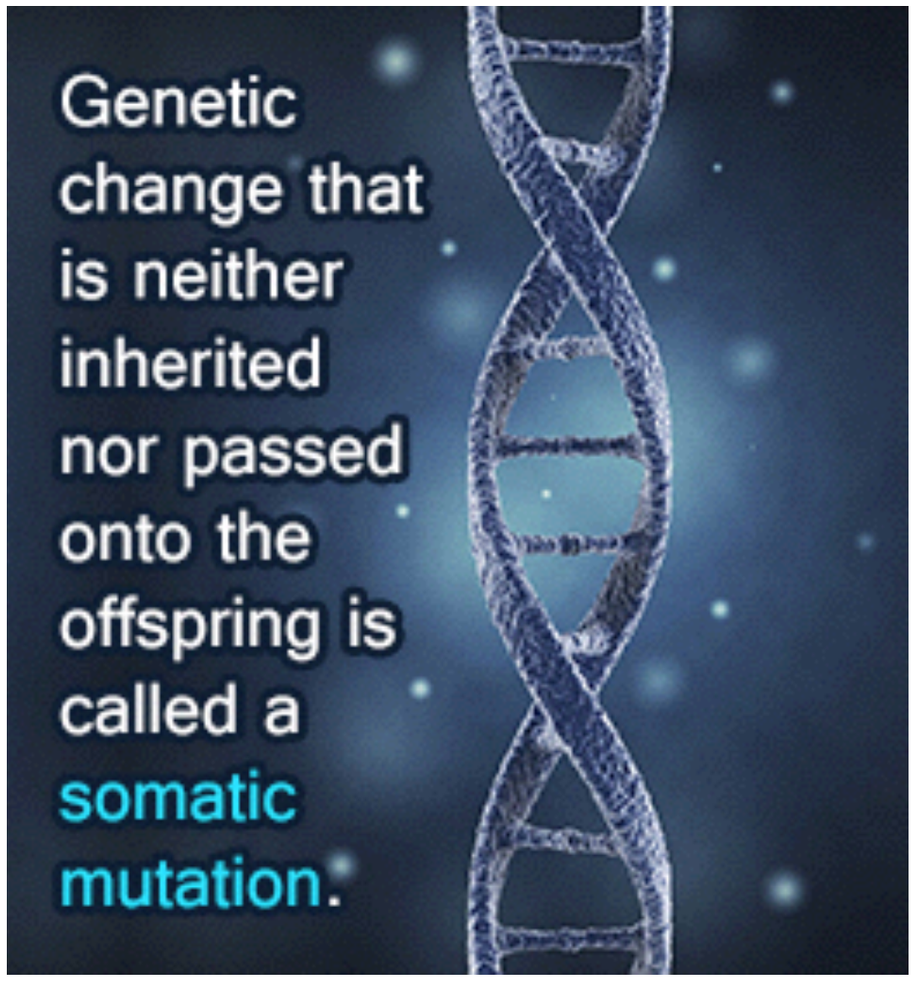


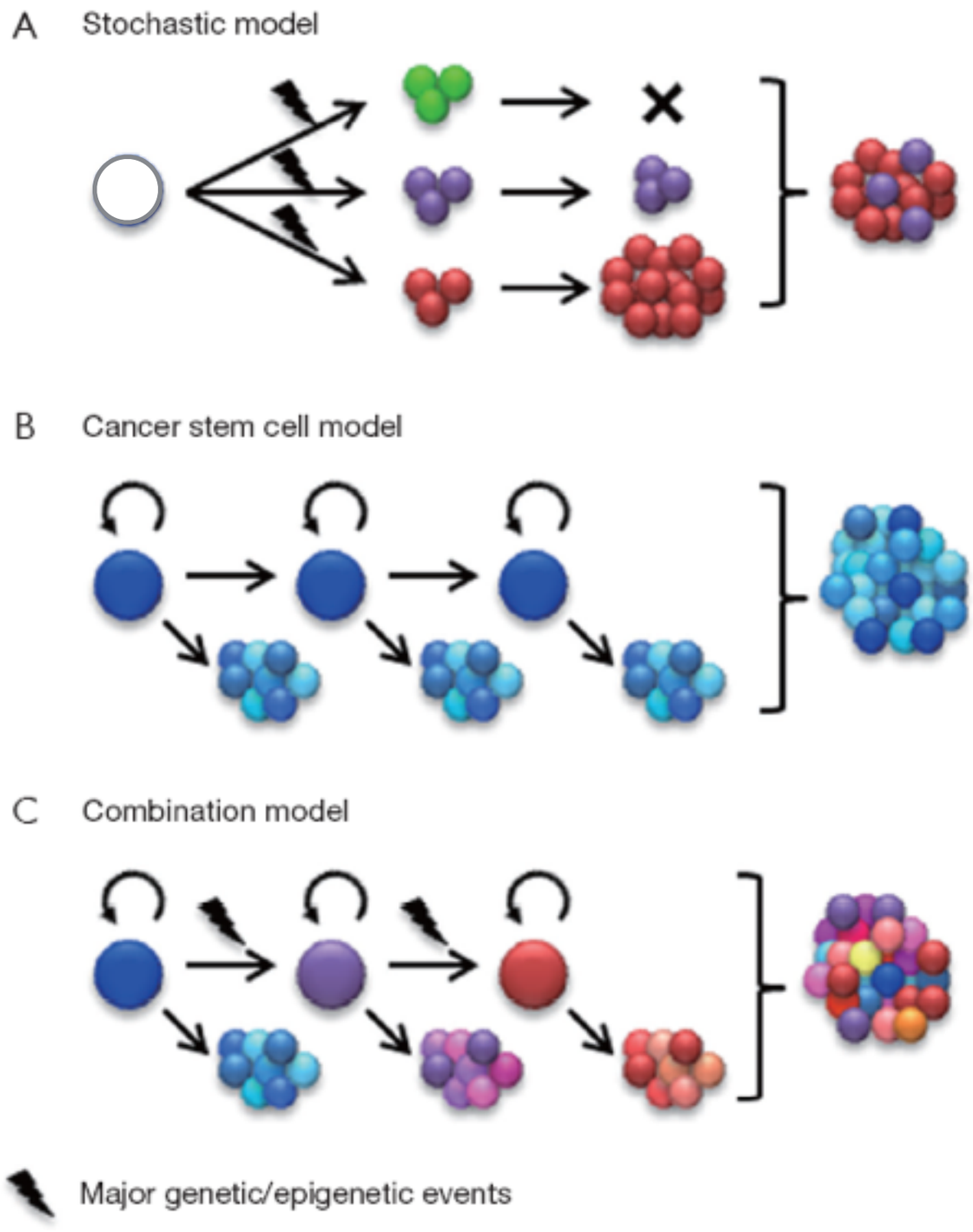
http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-24/CB24.html

© 2012 Pearson Education, Inc.

Cancer Biology

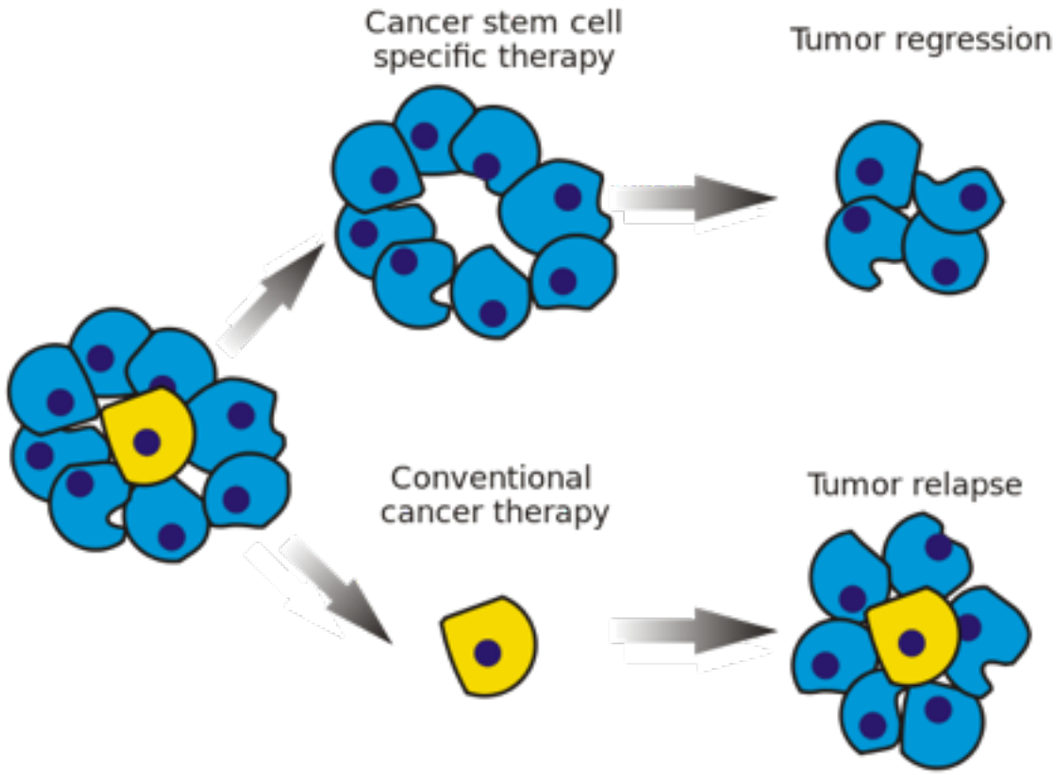
Tumors are composed of genetically and phenotypically **heterogeneous** clones

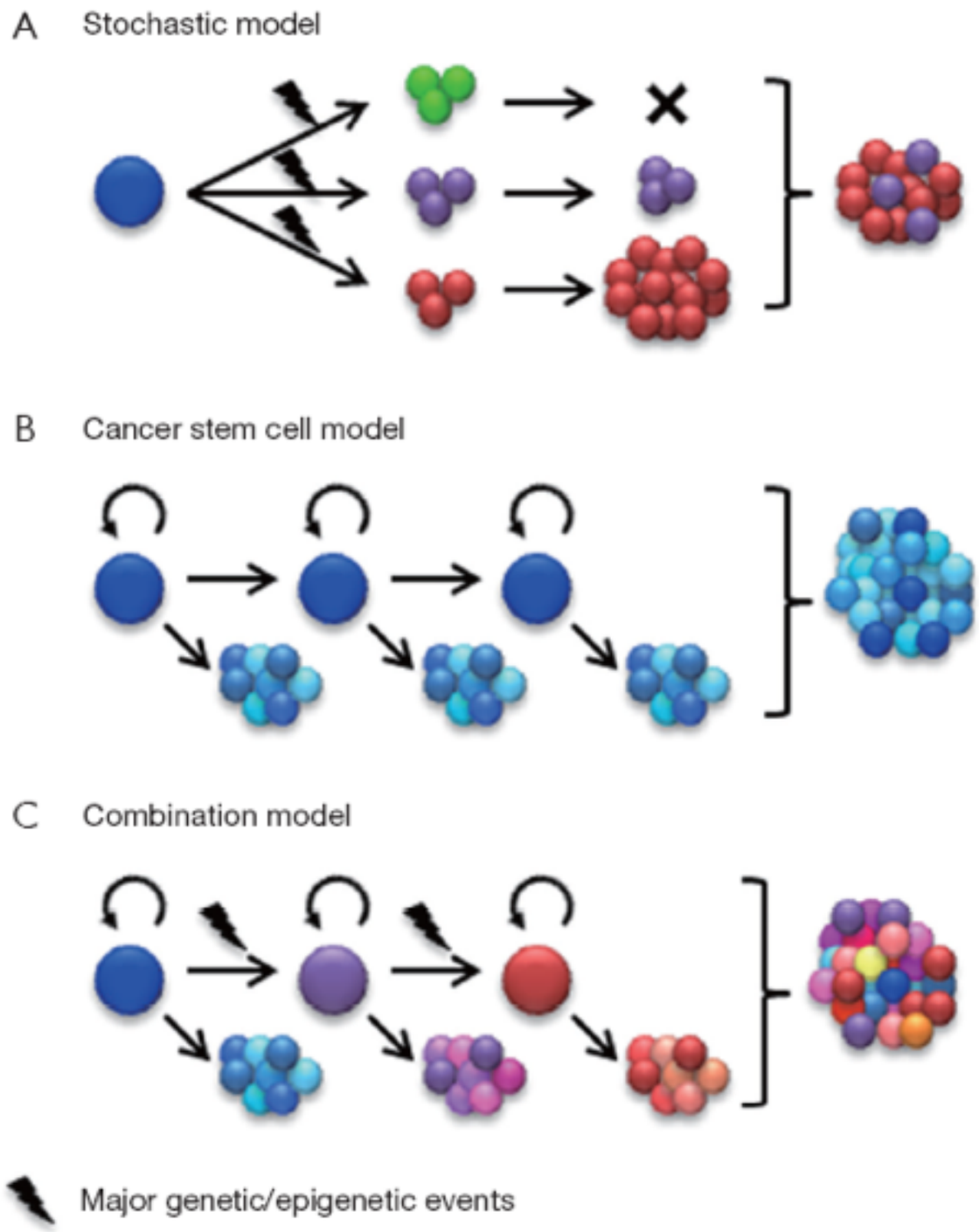




http://www.thetcr.org/article/viewFile/1415/html/10439

Cancer Biology





Deep (bulk) sequencing can only capture 1% of the cell population (excluding some types such as circulating tumor cells).

http://www.thetcr.org/article/viewFile/1415/html/10439

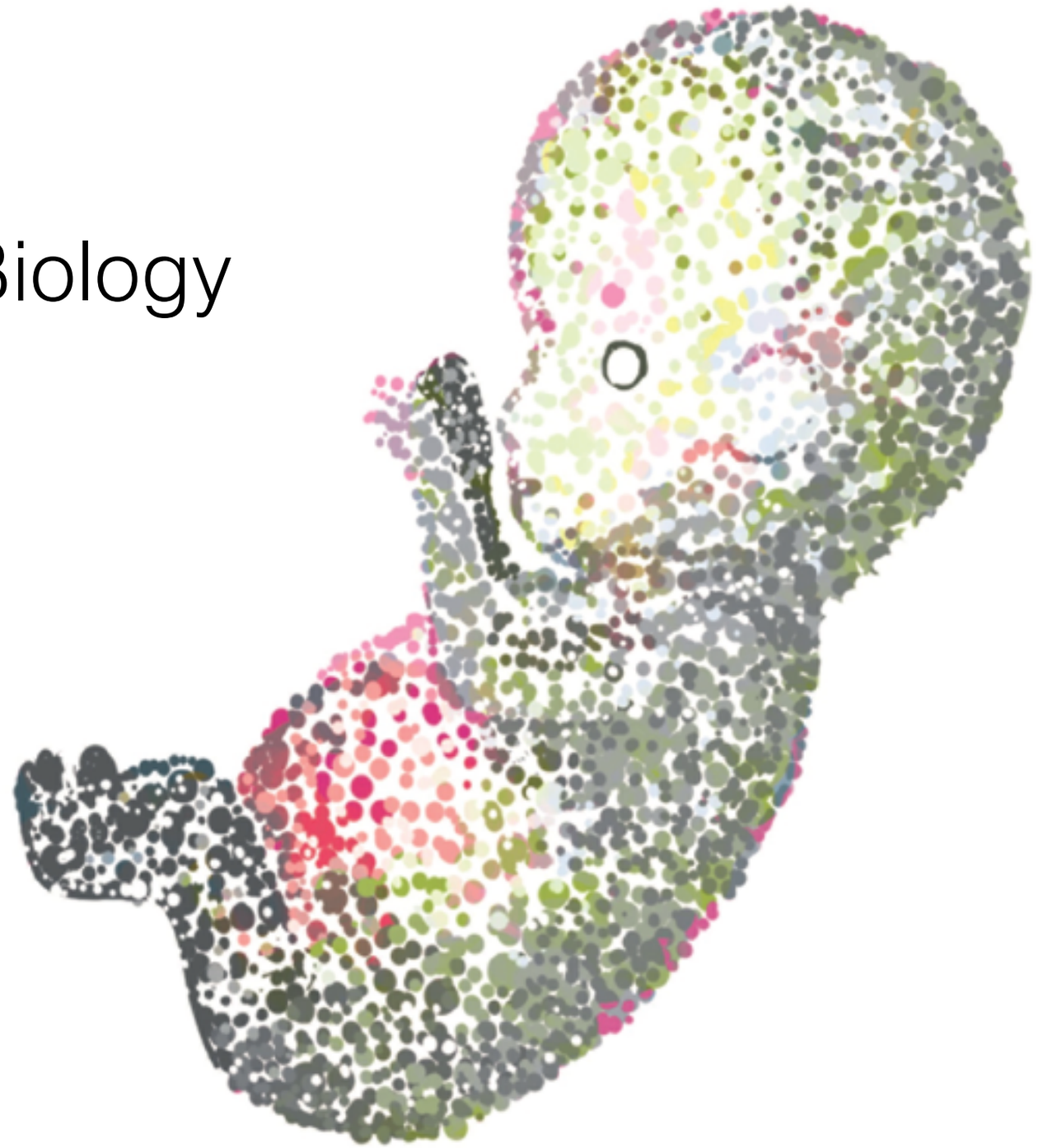
Cancer Biology

We need single-cell resolution to:

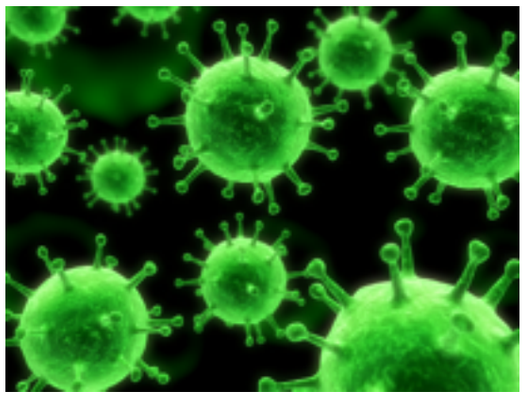
- Find evidence for models of cancer
- Infer timing of mutations and the drivers
- Evaluate effectiveness of targeted therapy

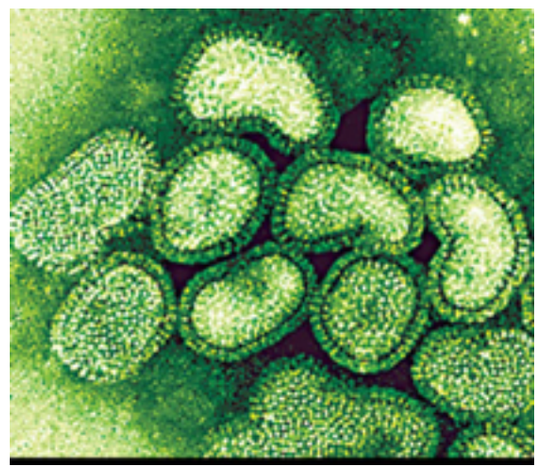
Applications

- Developmental Biology
- Cancer Biology
- **Microbiology**
- Neurology

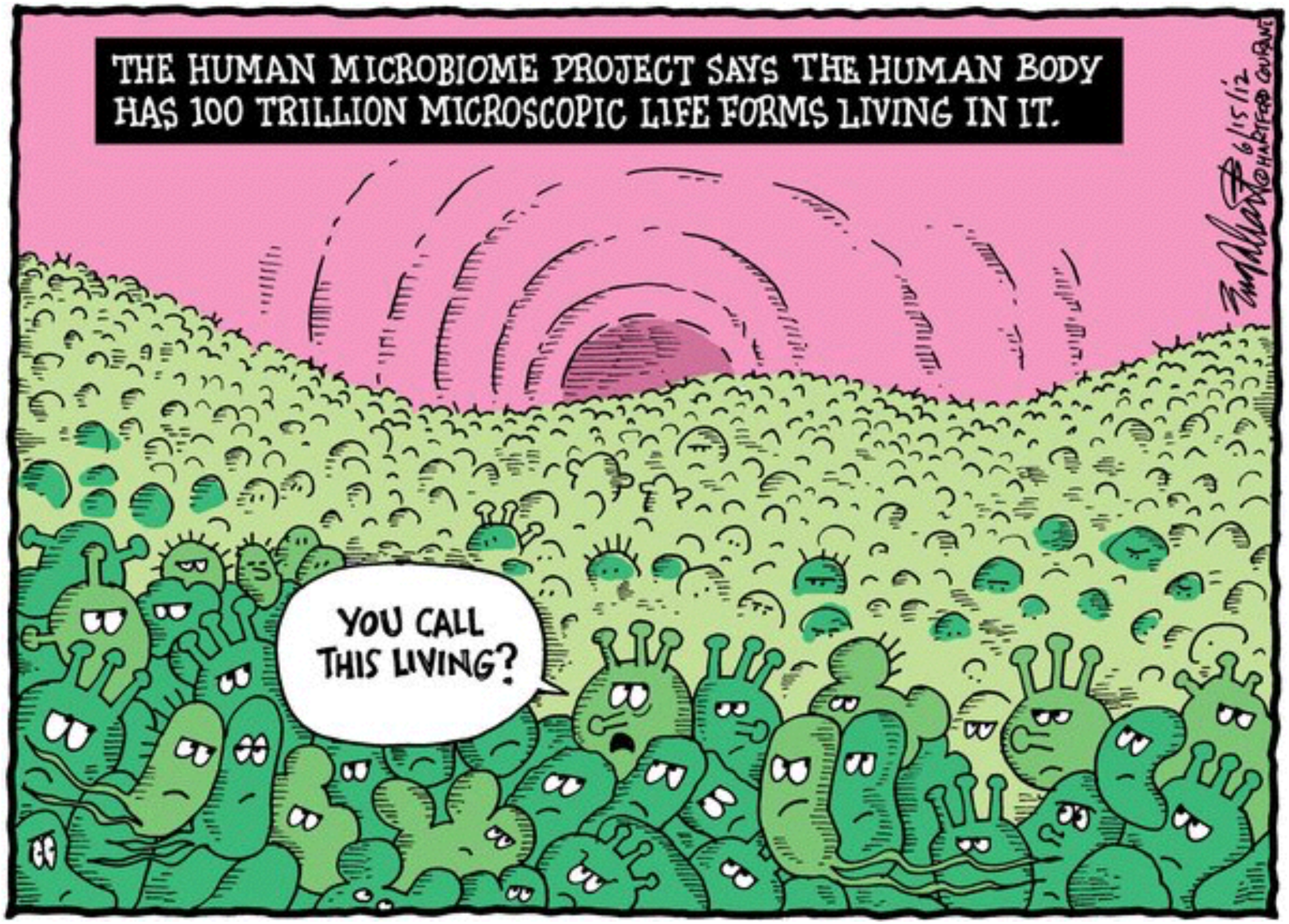


Microbiology

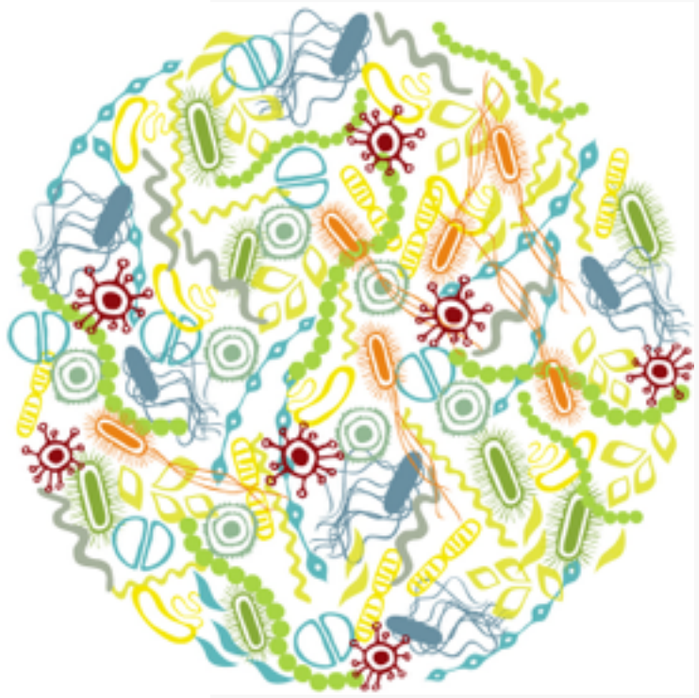


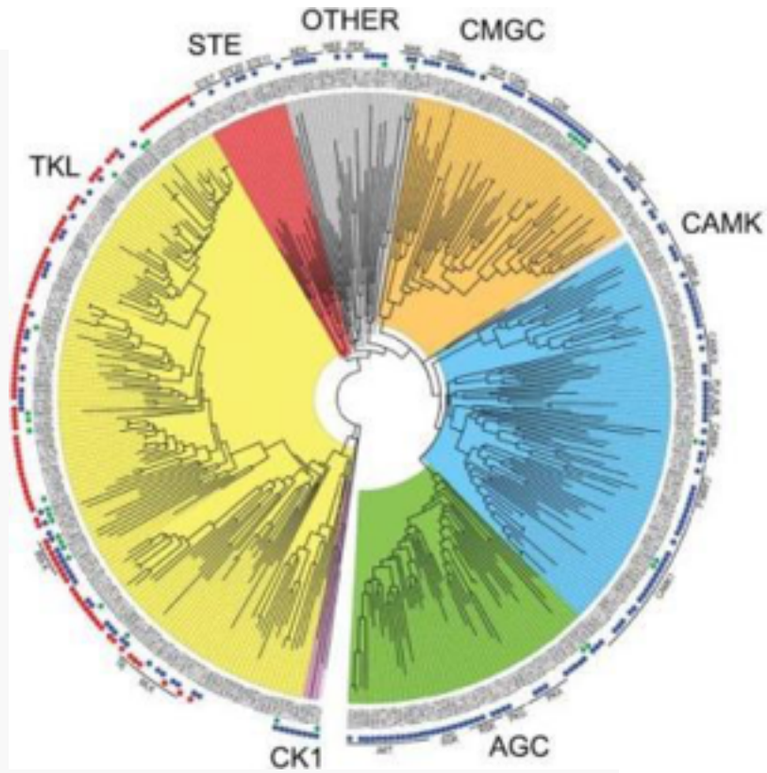


Microbiology



Microbiology





http://www.cbs.dtu.dk/researchgroups/metagenomics/mg.jpg

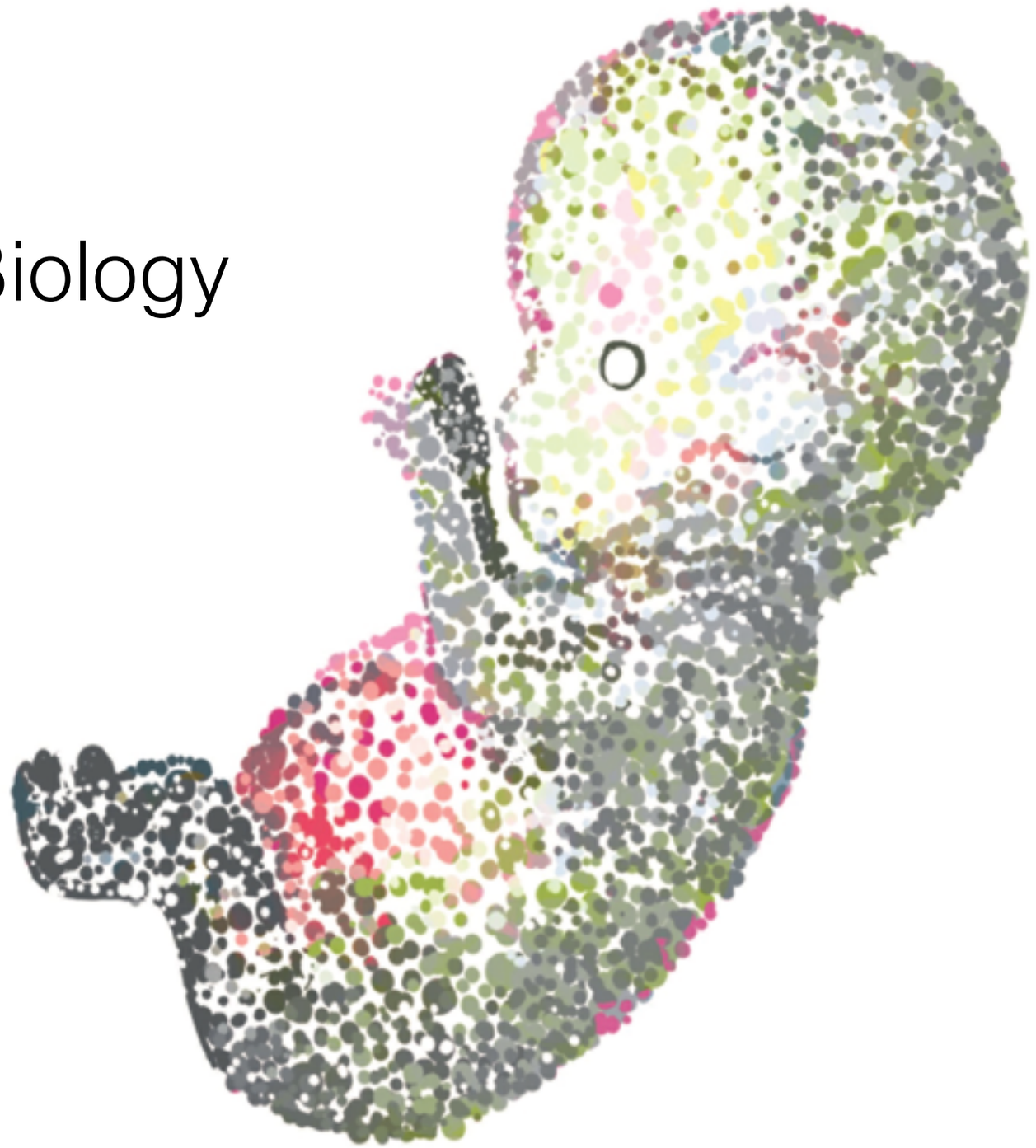
Microbiology

We need single-cell resolution to:

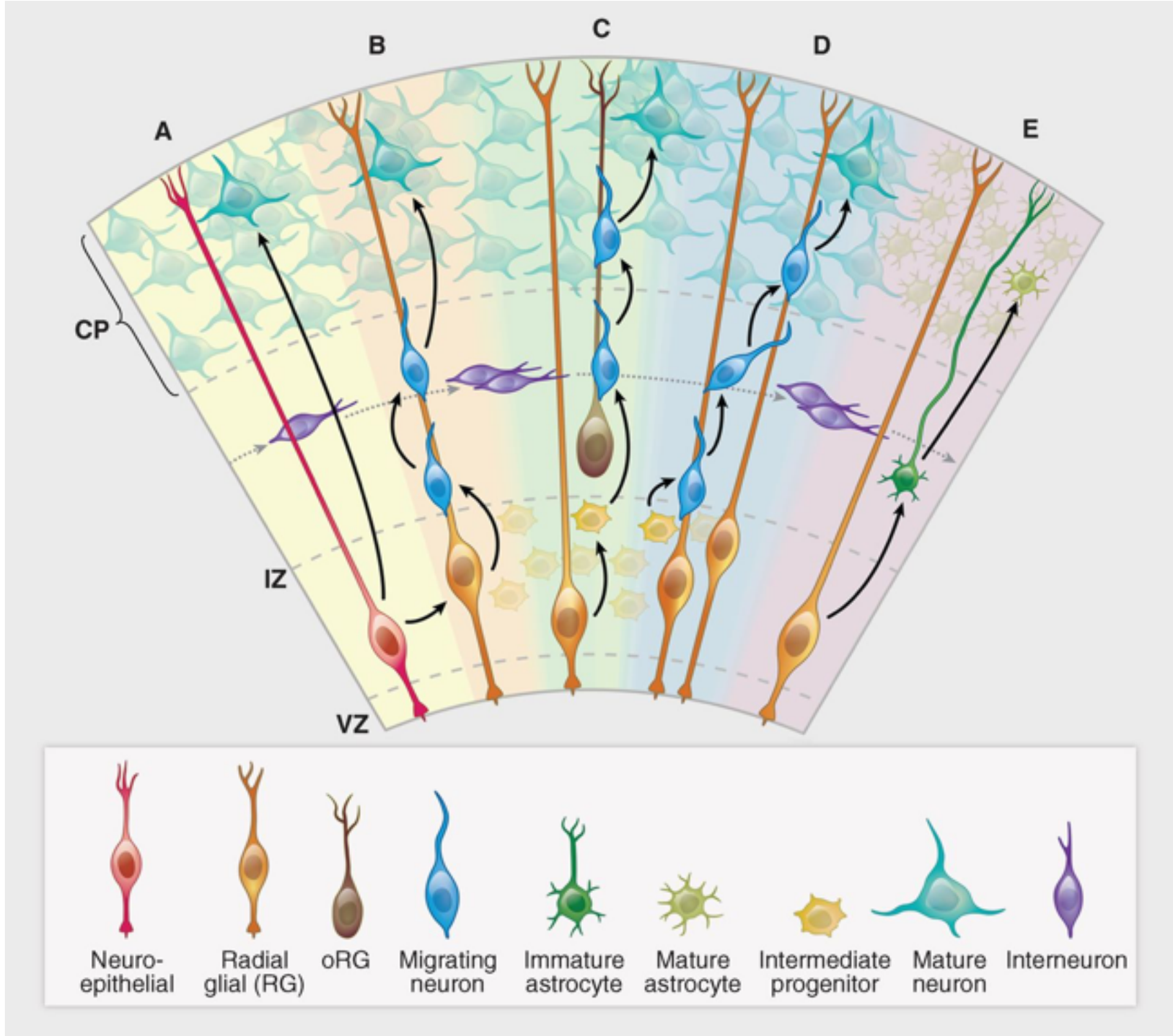
- Discover low-abundance species that are are difficult to culture in vitro
- Monitor transcriptional gene activation mechanisms for functional annotation

Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- **Neurology**

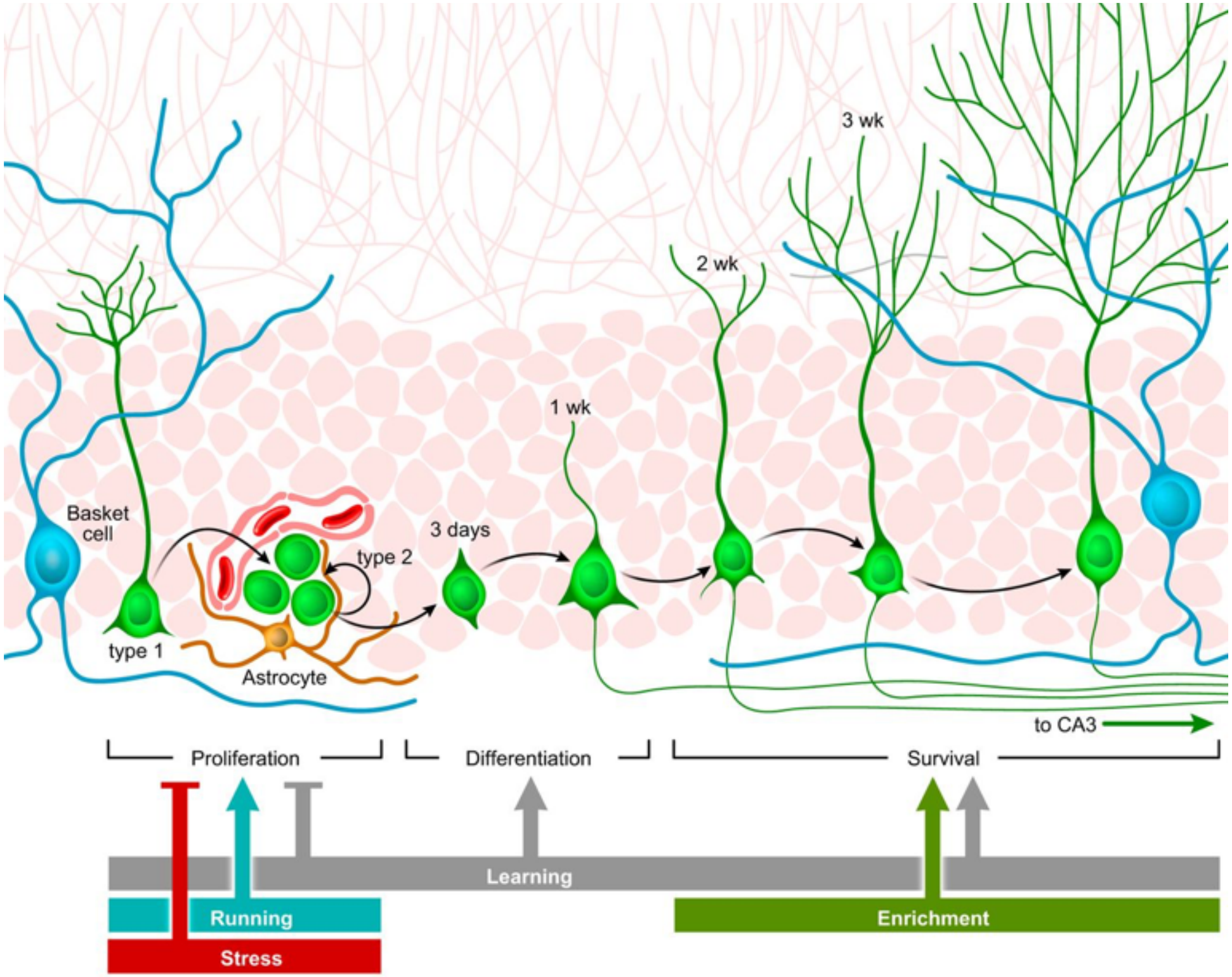


Neurology



https://www.sciencemag.org/content/341/6141/1237758/embed/inline-graphic-1.gif

Neurology


3 wk
2 wk
1 wk
3 days
type 2
Basket cell
type 1
Astrocyte
to CA3
Proliferation
Differentiation
Survival
Learning
Running
Stress
Enrichment

Neurology

We need single-cell resolution to:

- 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

NOW THAT'S IRONIC!

26

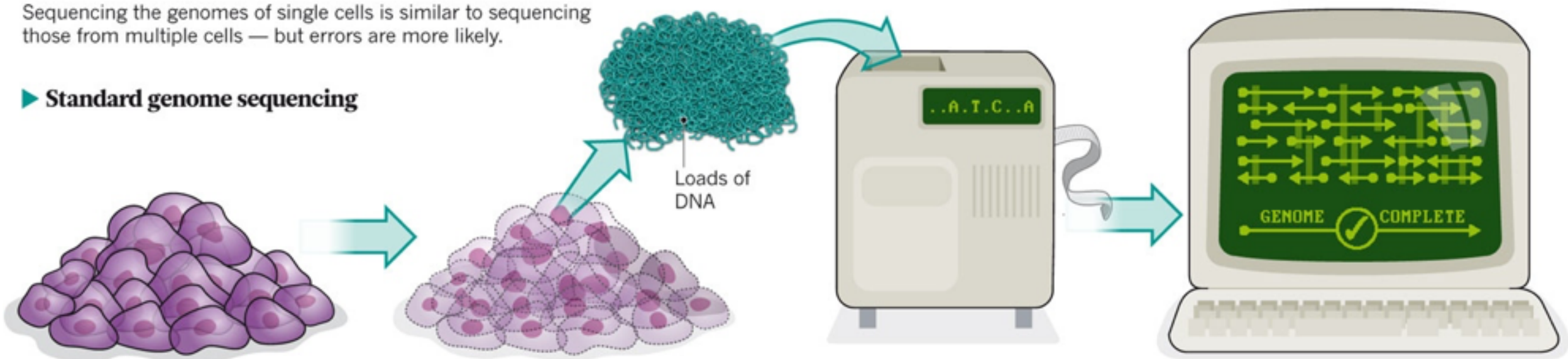
©2010 CarrieGlennStudios.com

Traditional v.s. Single-cell

ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

▶ Standard genome sequencing



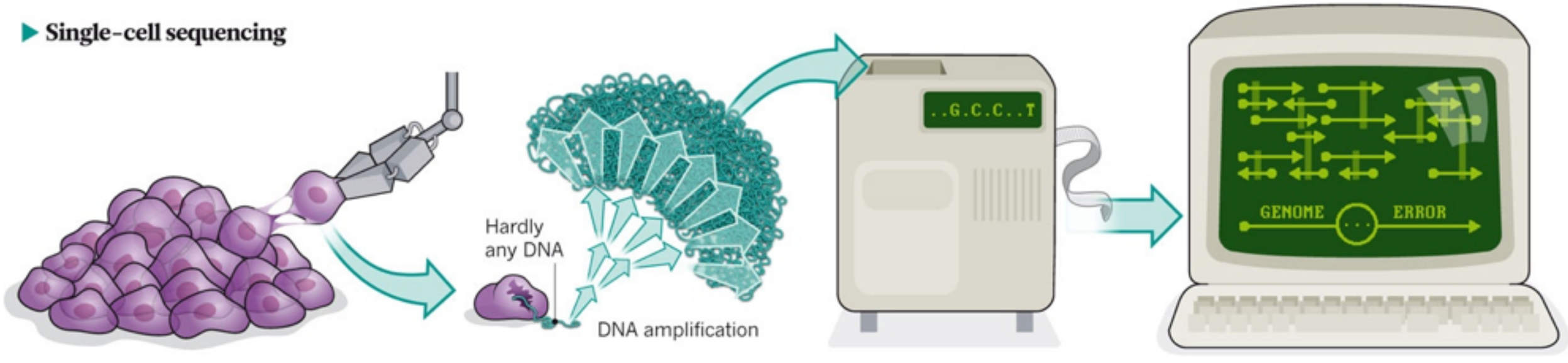
A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

▶ Single-cell sequencing



A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

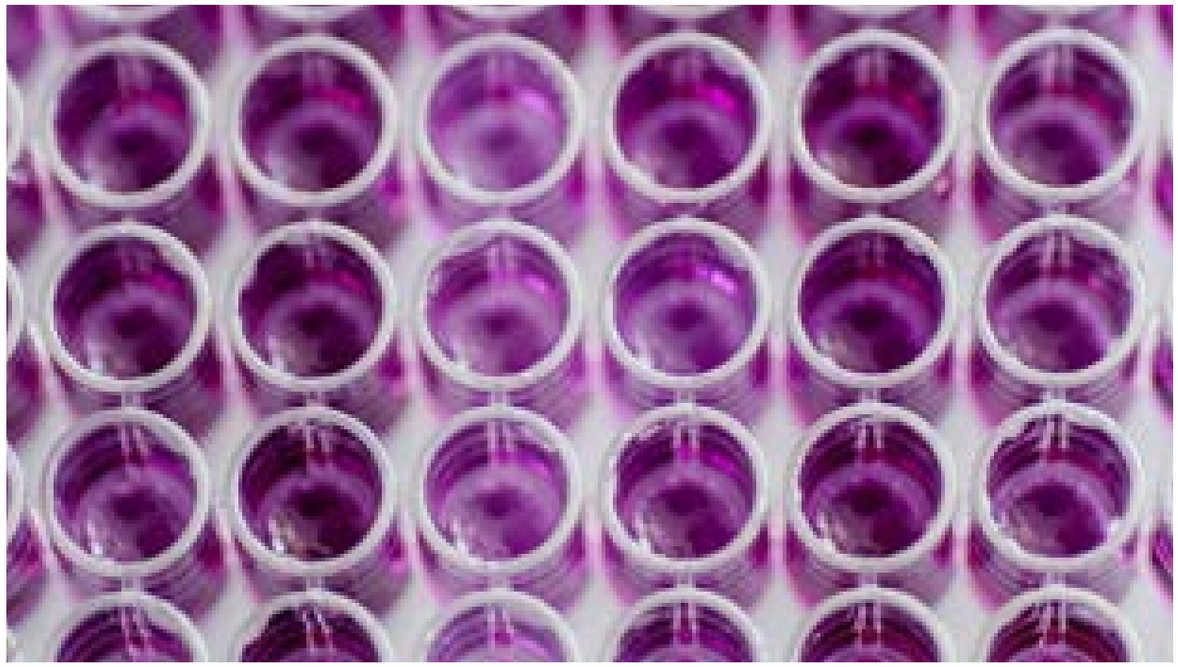
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

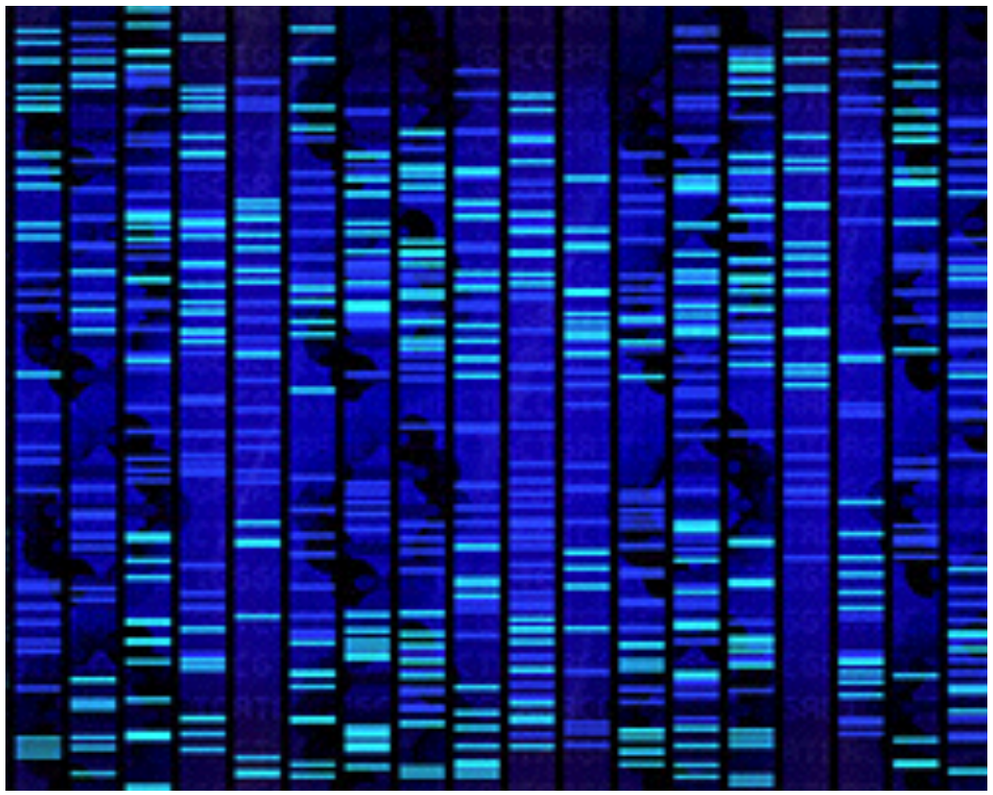
Single-Cell Technologies

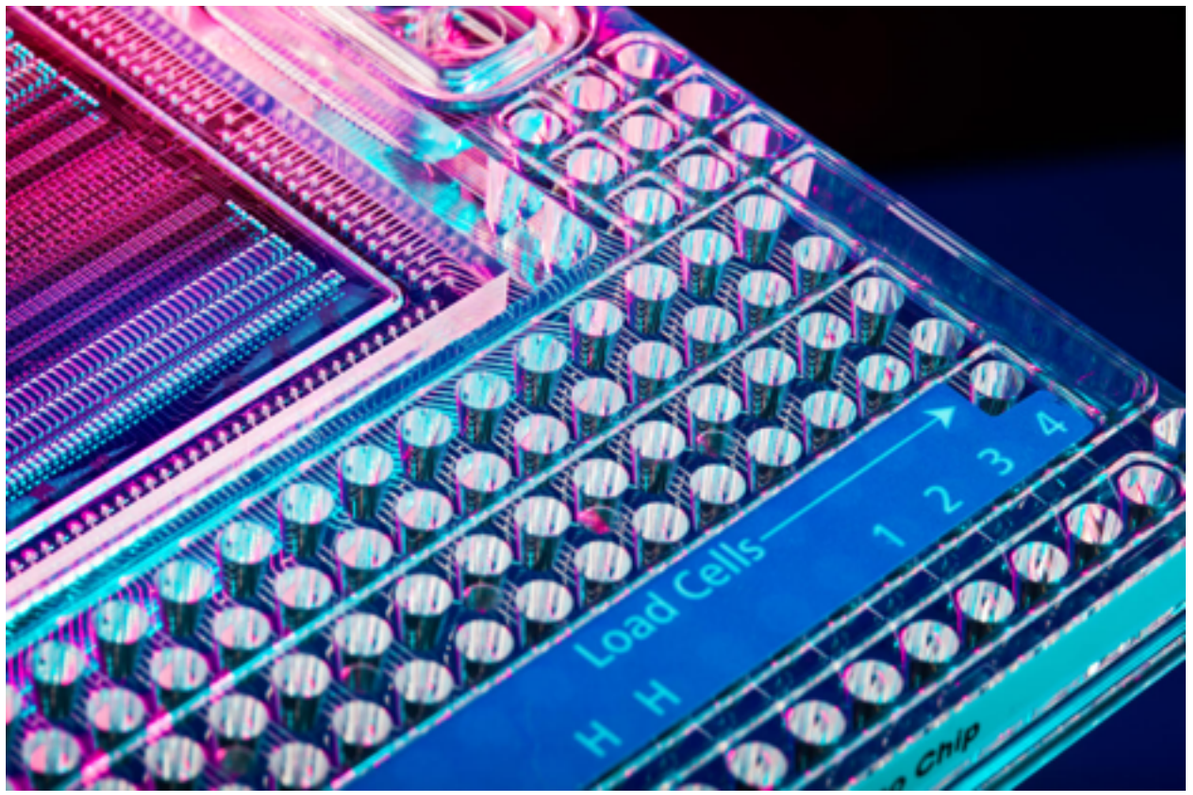
(i) isolate single cells

(ii) amplify genome efficiently

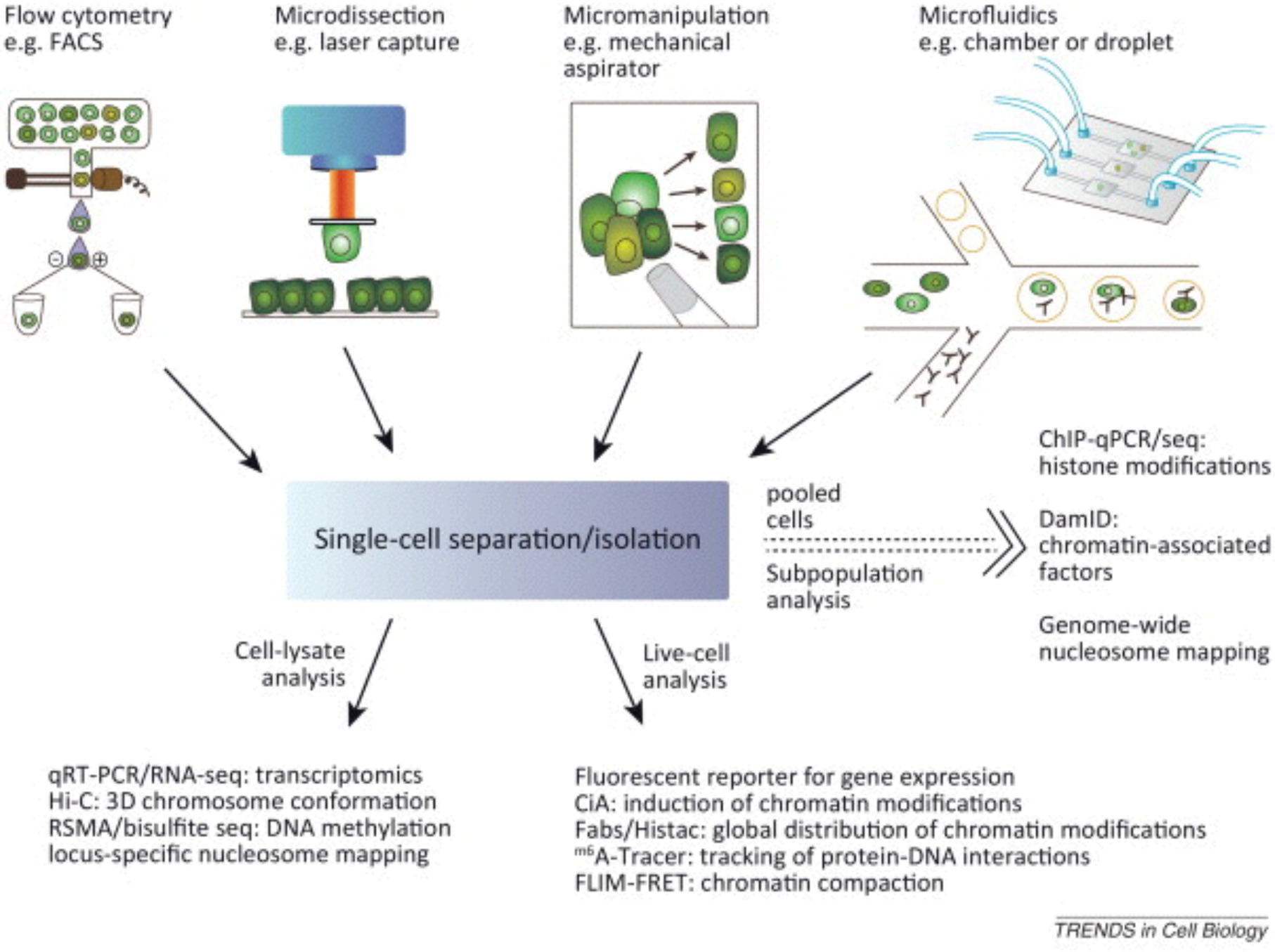
(iii) sequence DNA







Single-Cell Technologies

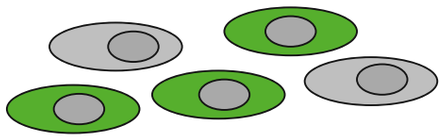


TRENDS in Cell Biology

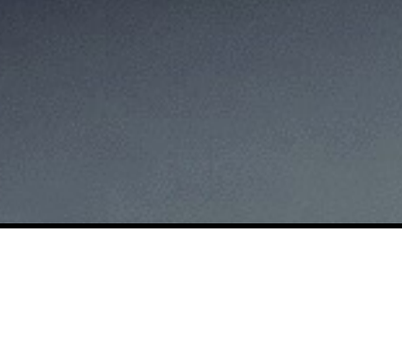
Cell Sorting

Manual

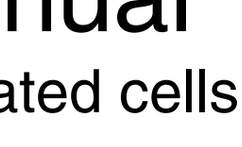
dissociated cells

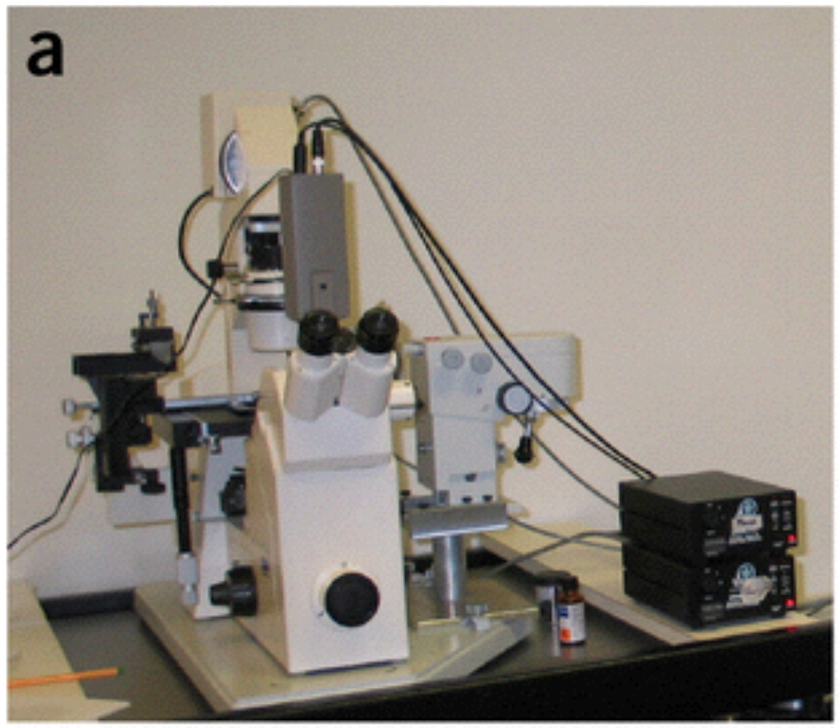


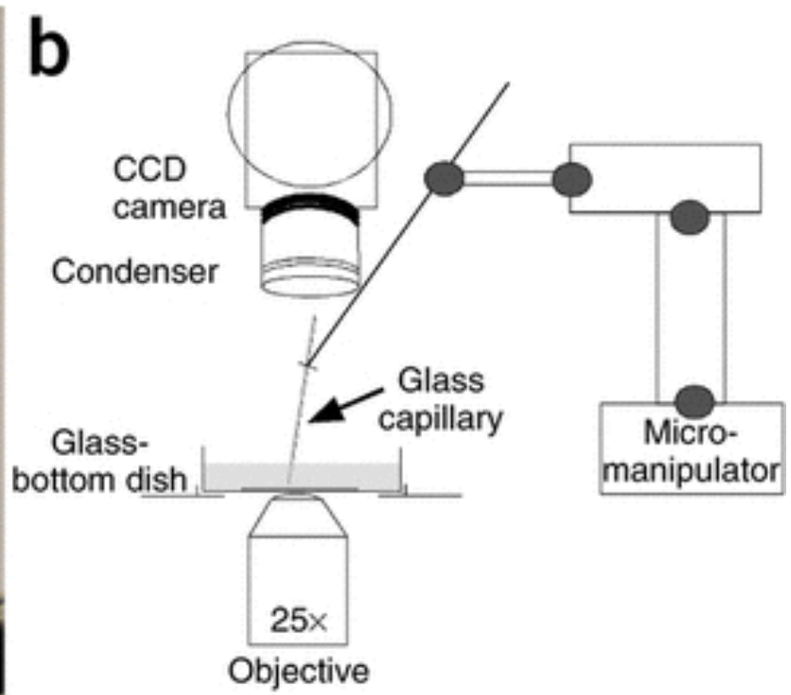
fluorescence microscope

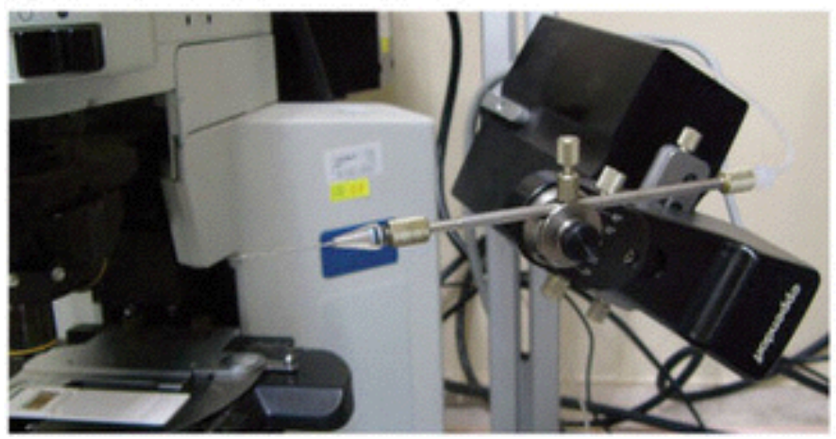


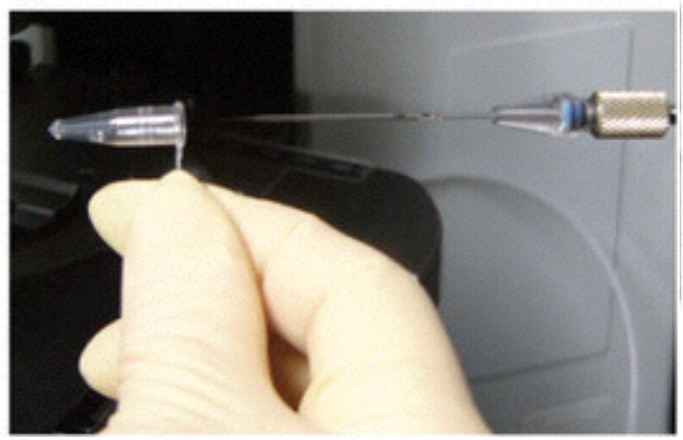
~100 cells





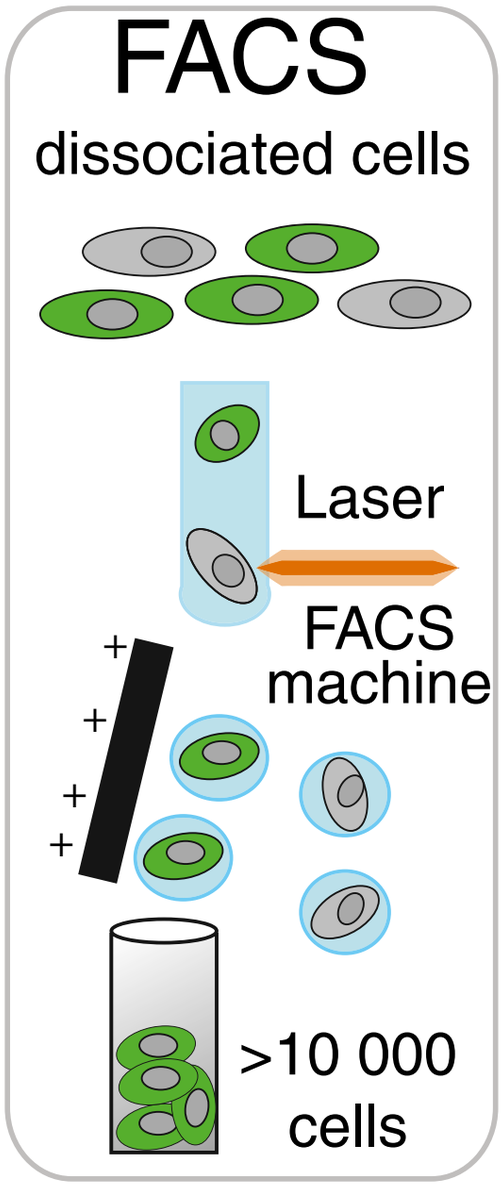




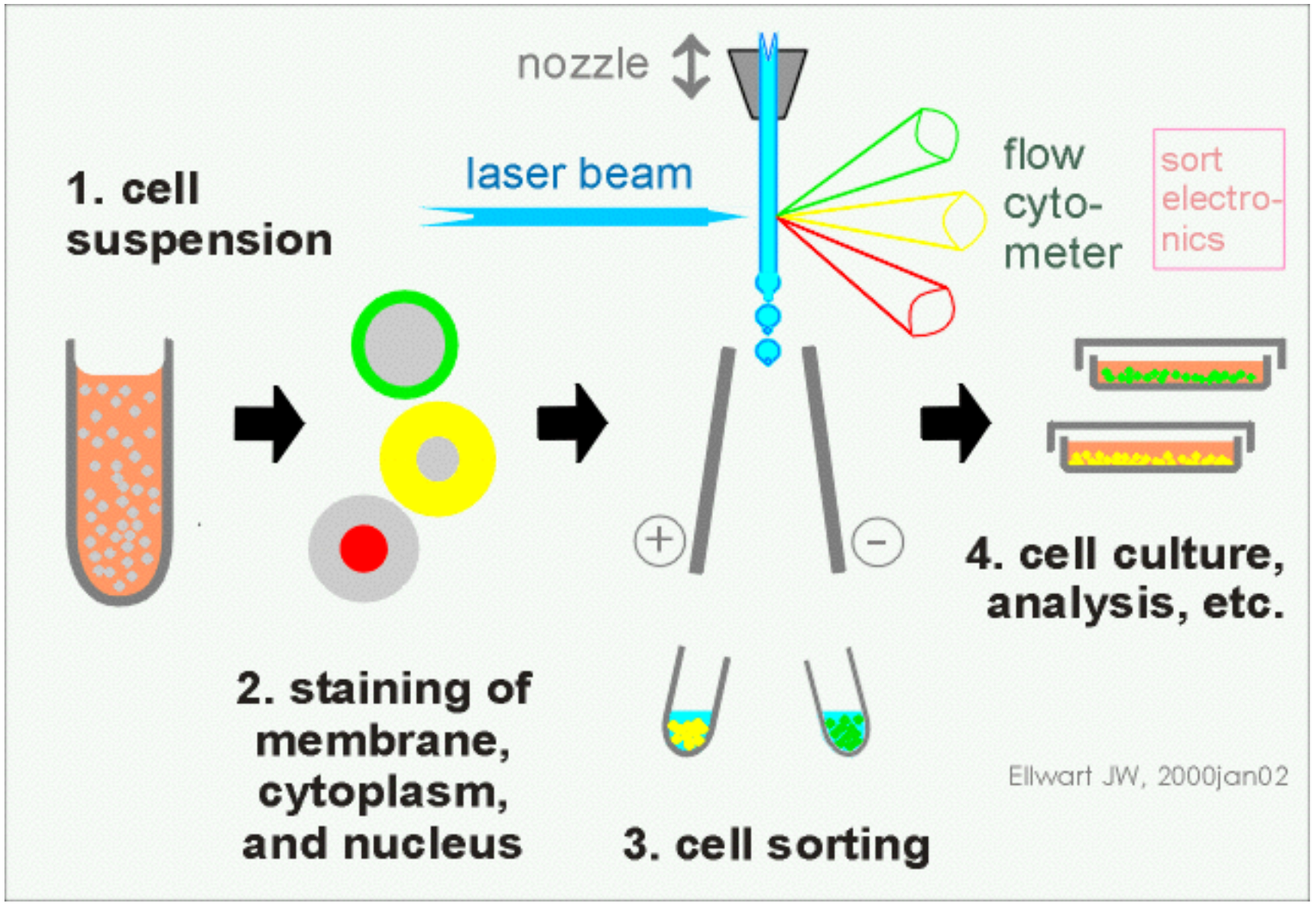


http://www.nature.com/nprot/journal/v6/n5/images_article/nprot.2011.322-F2.jpg

Cell Sorting

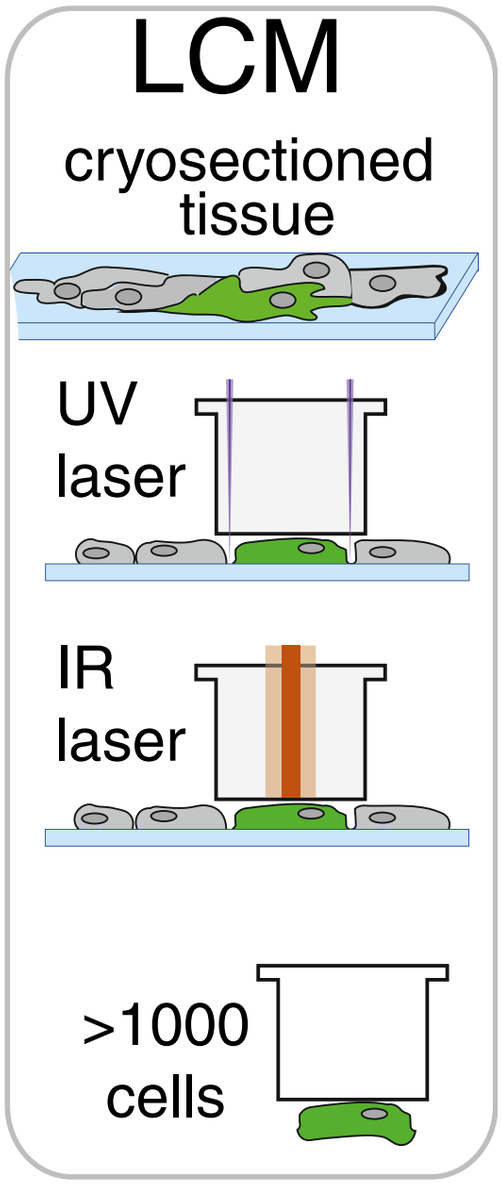


FACS: fluorescence activated cell sorting

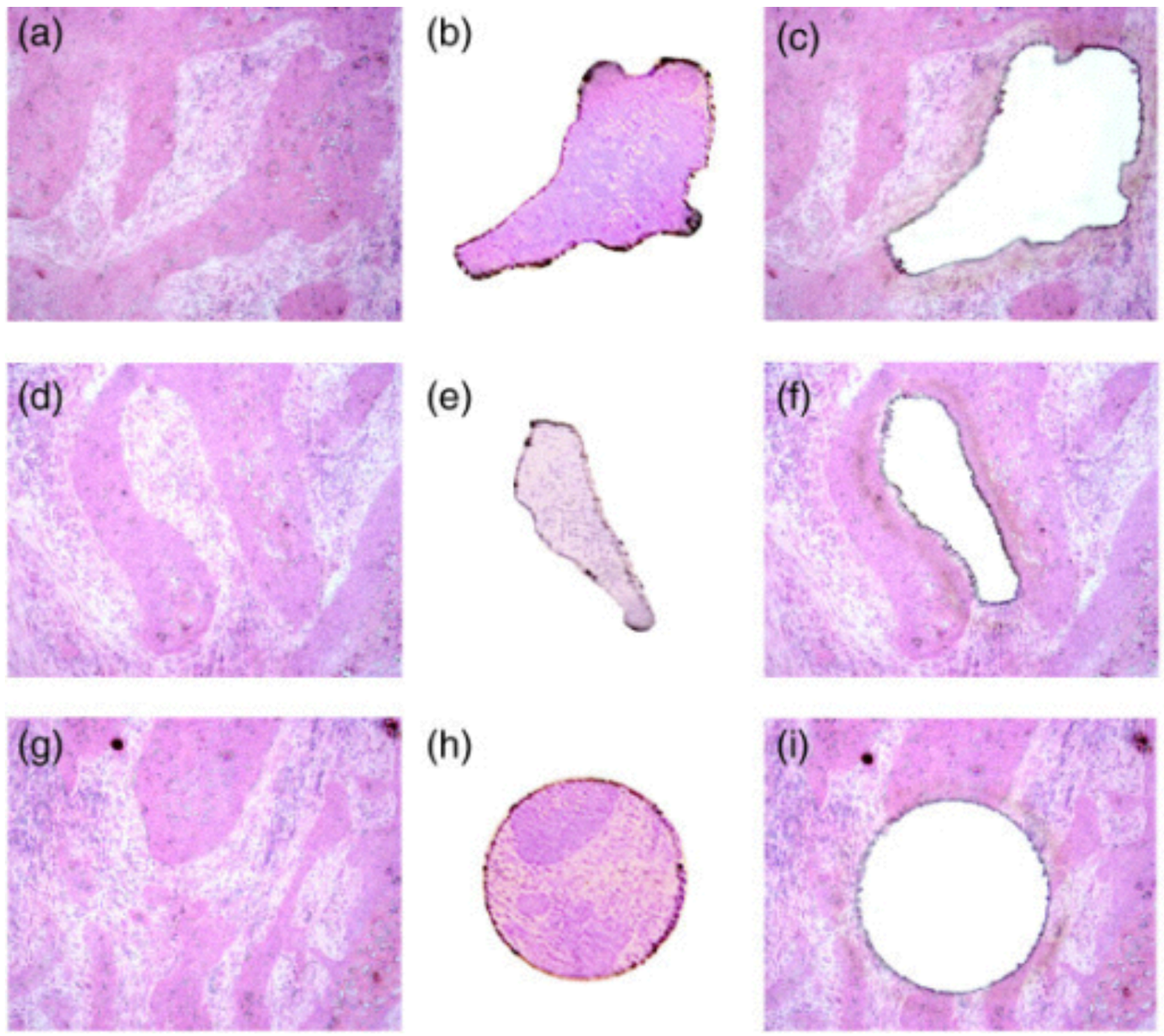


http://www.flowlab-childrens-harvard.com/yahoo_site_admin/assets/images/principle123.285181420_std.gif

Cell Sorting



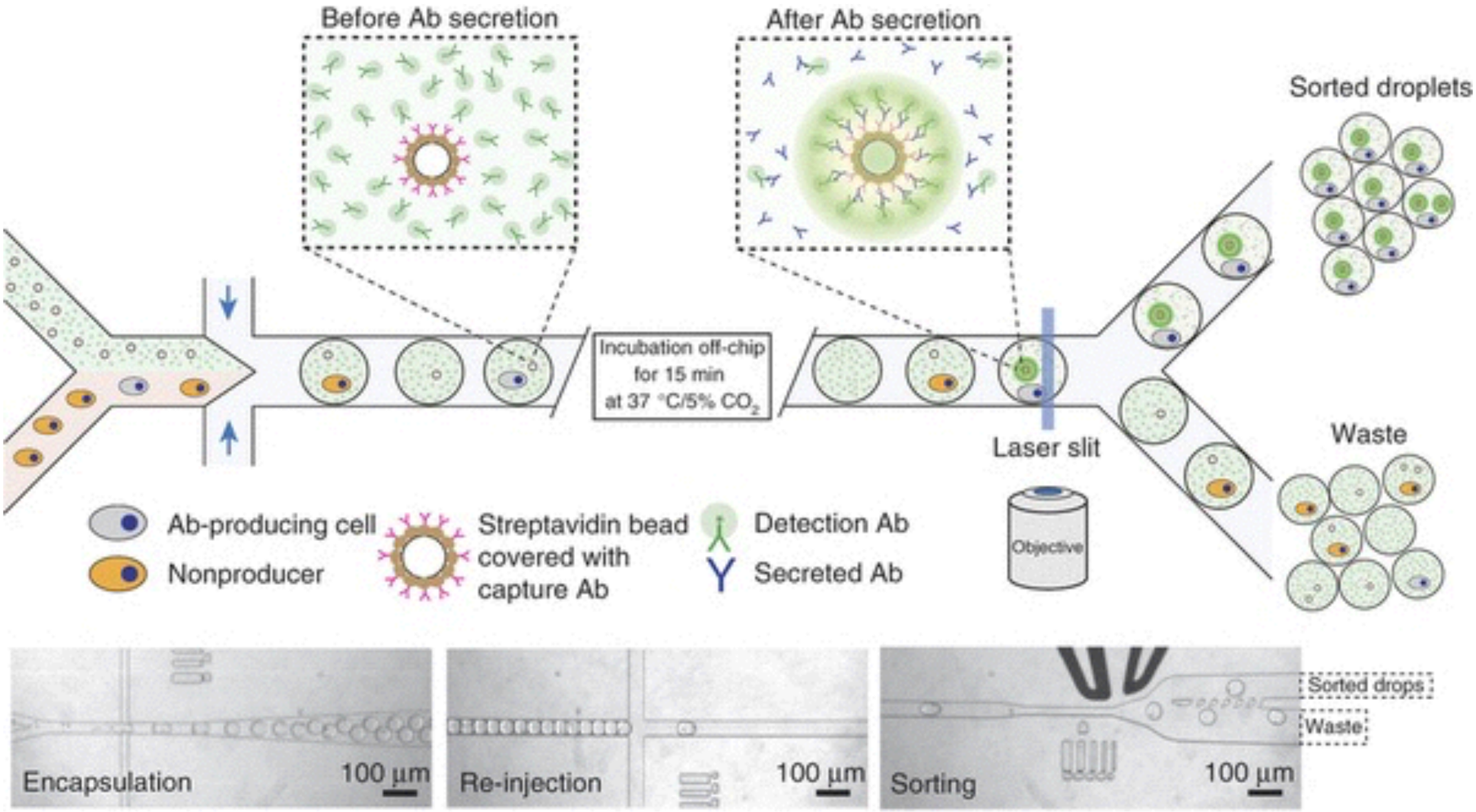
LCM: laser capture microdissection



http://www.genomemedicine.com/content/figures/gm247-2-l.jpg

Cell Sorting

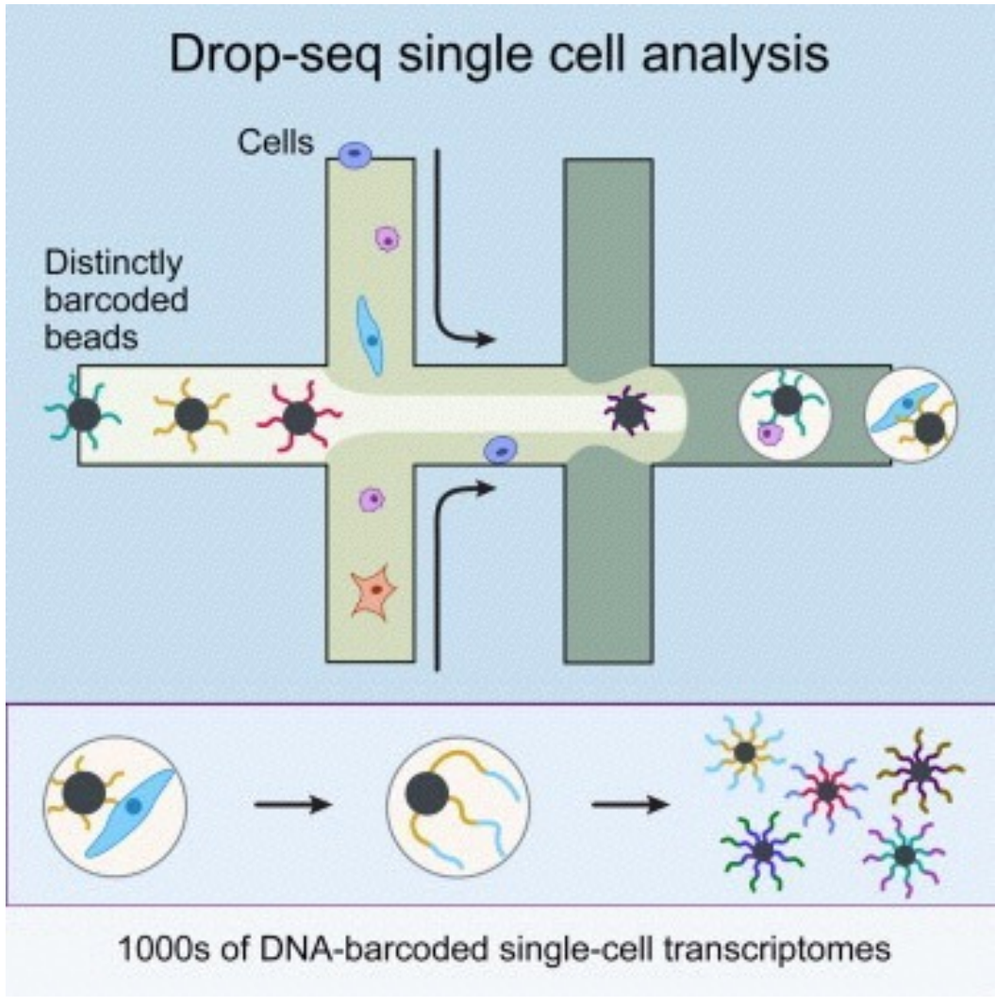
Microfluidics: can isolate rare circulating cells



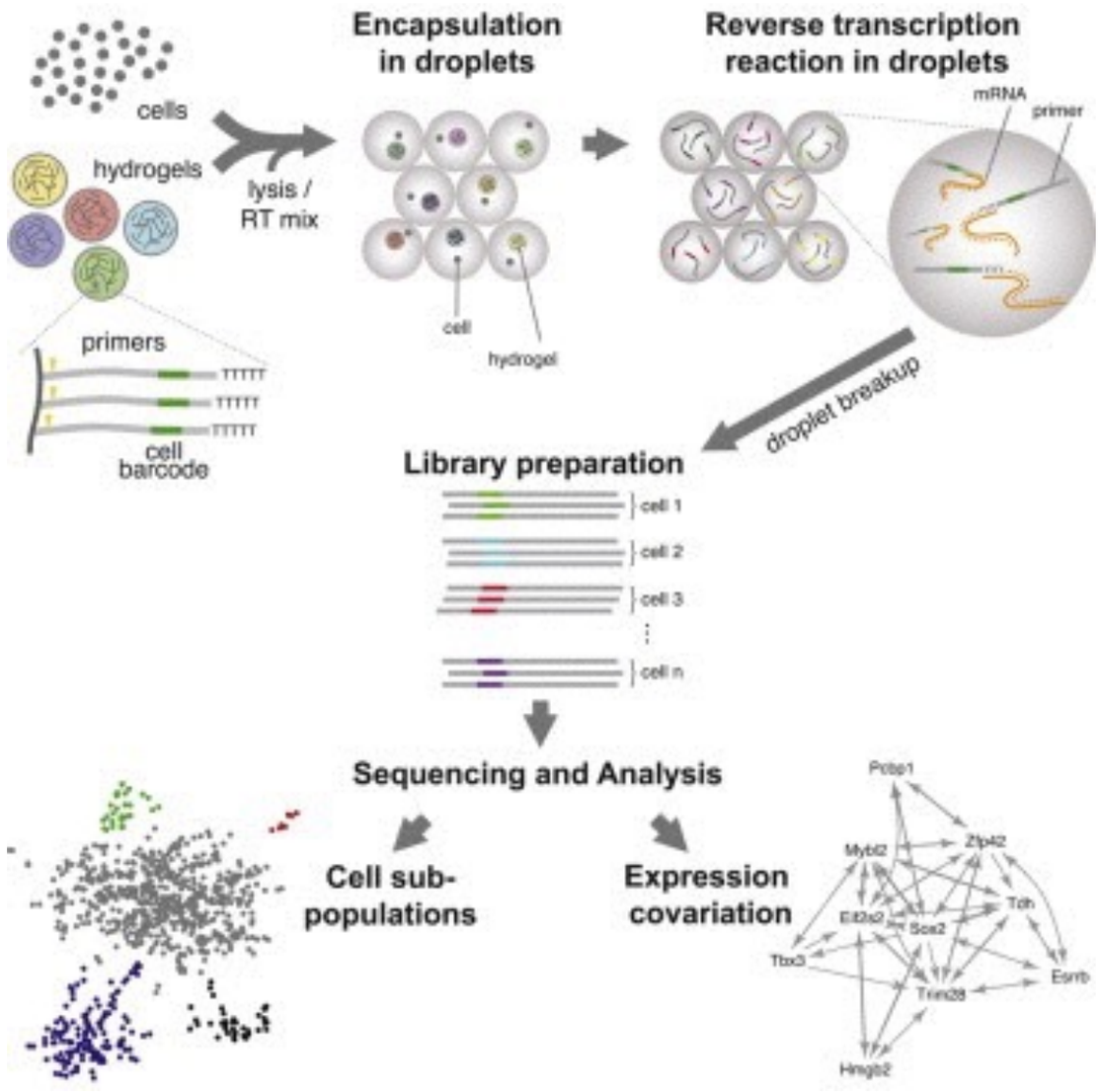
http://www.nature.com/nprot/journal/v8/n5/images_article/nprot.2013.046-F4.jpg

Cell Sorting

High-throughput (~100,000 cells)



Drop-seq



inDrop

http://www.cell.com/abstract/S0092-8674(15)00549-8

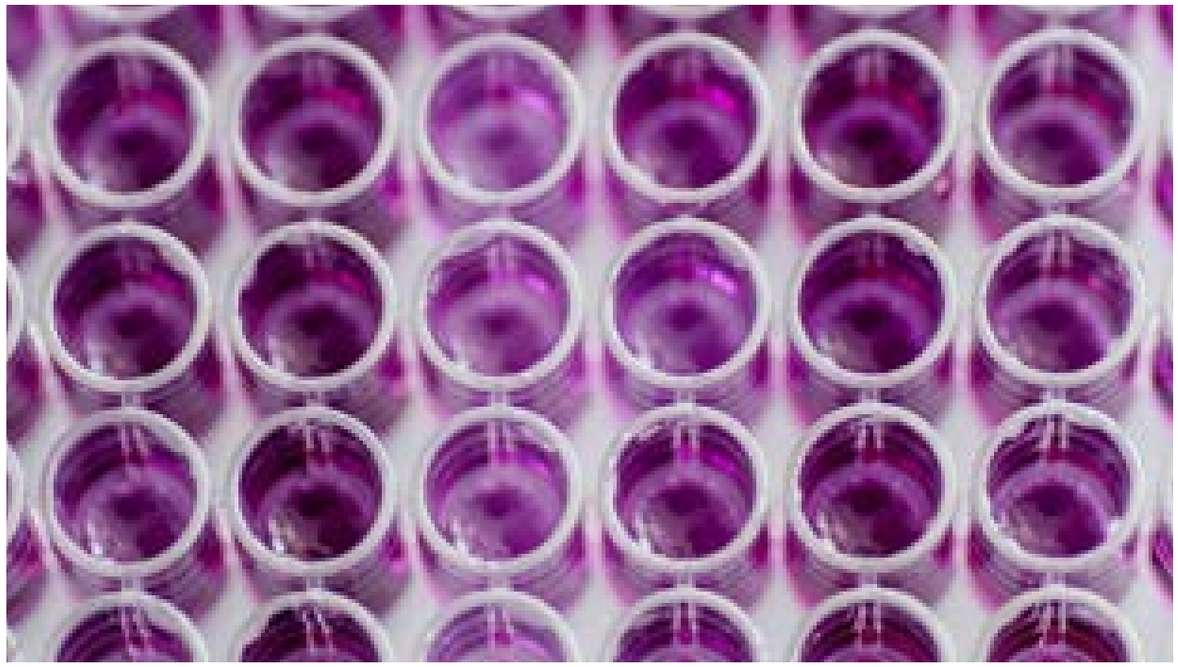
http://www.cell.com/cell/abstract/S0092-8674(15)00500-0

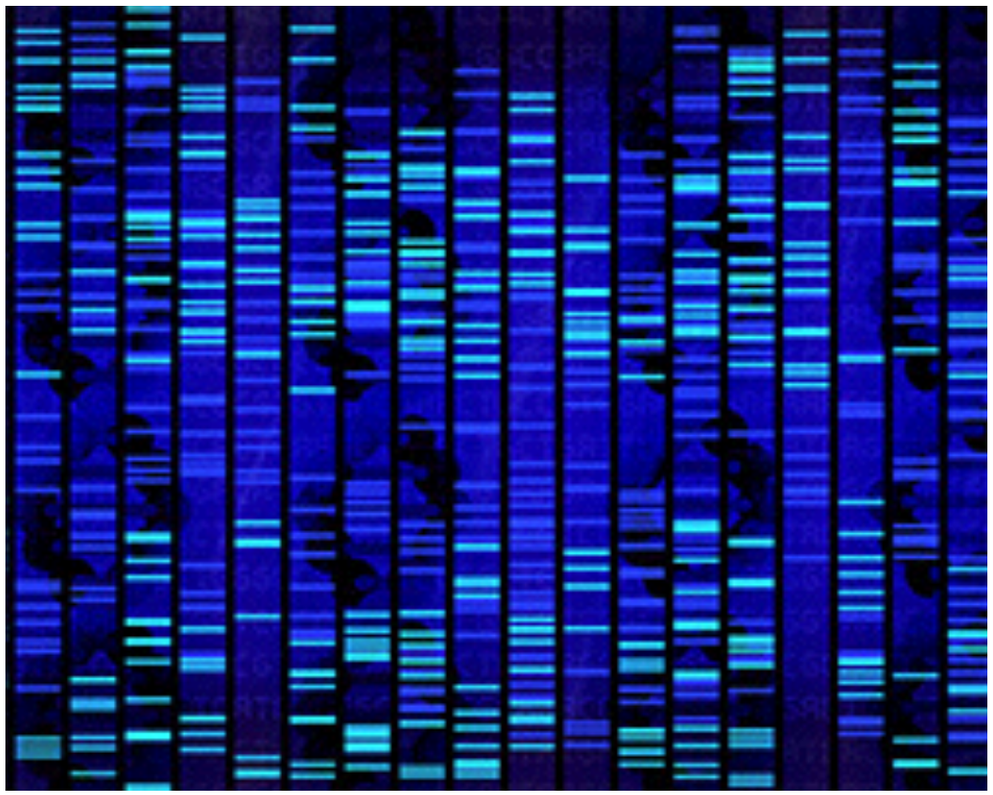
Single-Cell Technologies

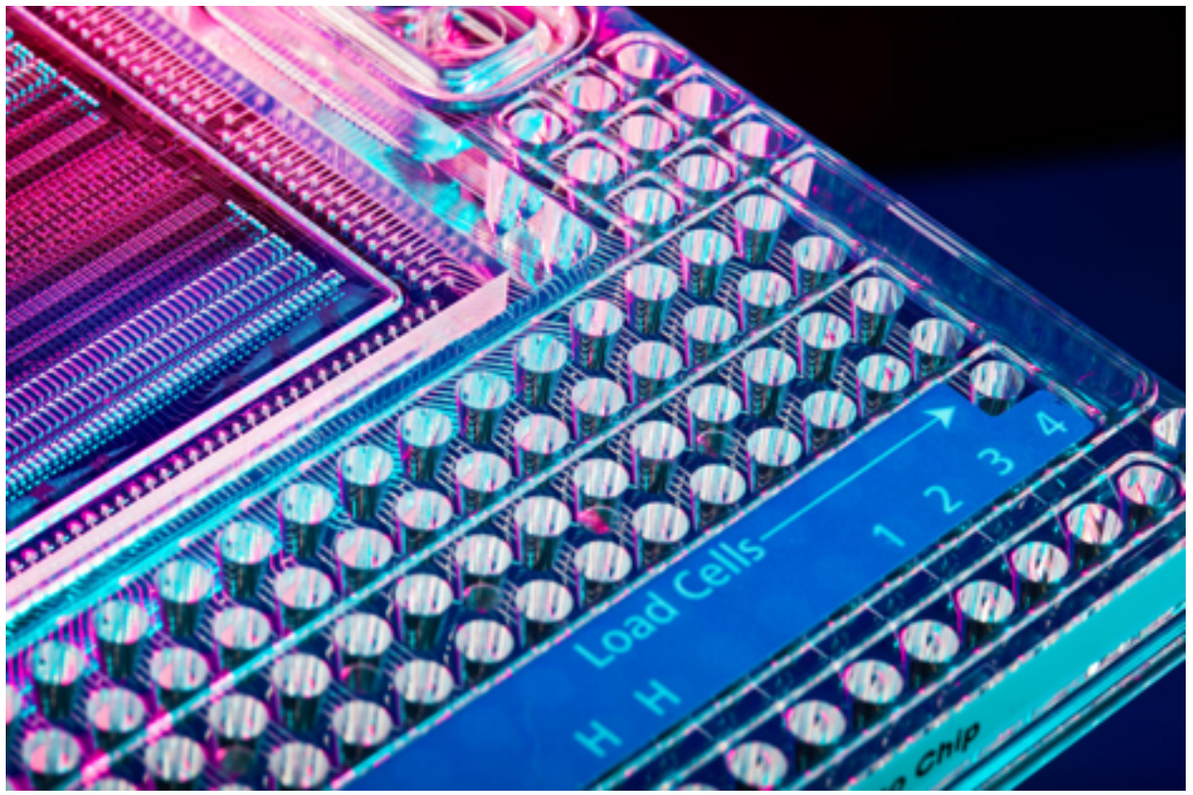
(i) isolate single cells

(ii) amplify genome efficiently

(iii) sequence DNA







Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

illumina

Roche

life technologies

PACIFIC BIOSCIENCES

Helicos BioSciences Corporation

ion torrent

Oxford NANOPORE Technologies

HiSeq2000 / 2500

MiSeq

Ion Pronton

Ion torrent

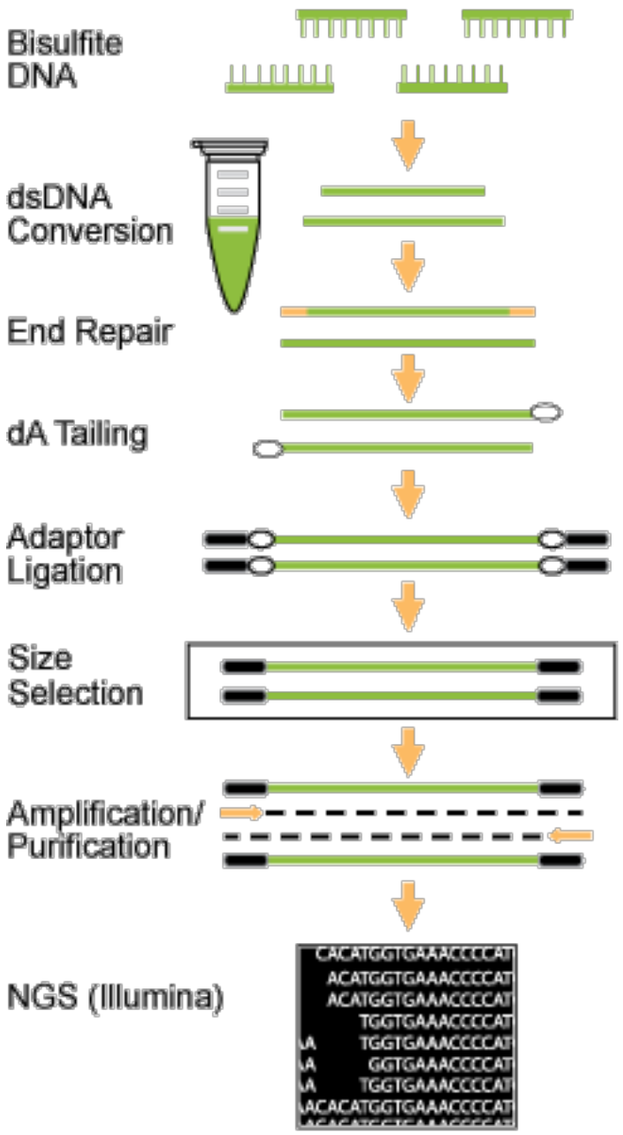
GS-FLX

Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

https://www.illumina.com

library preparation

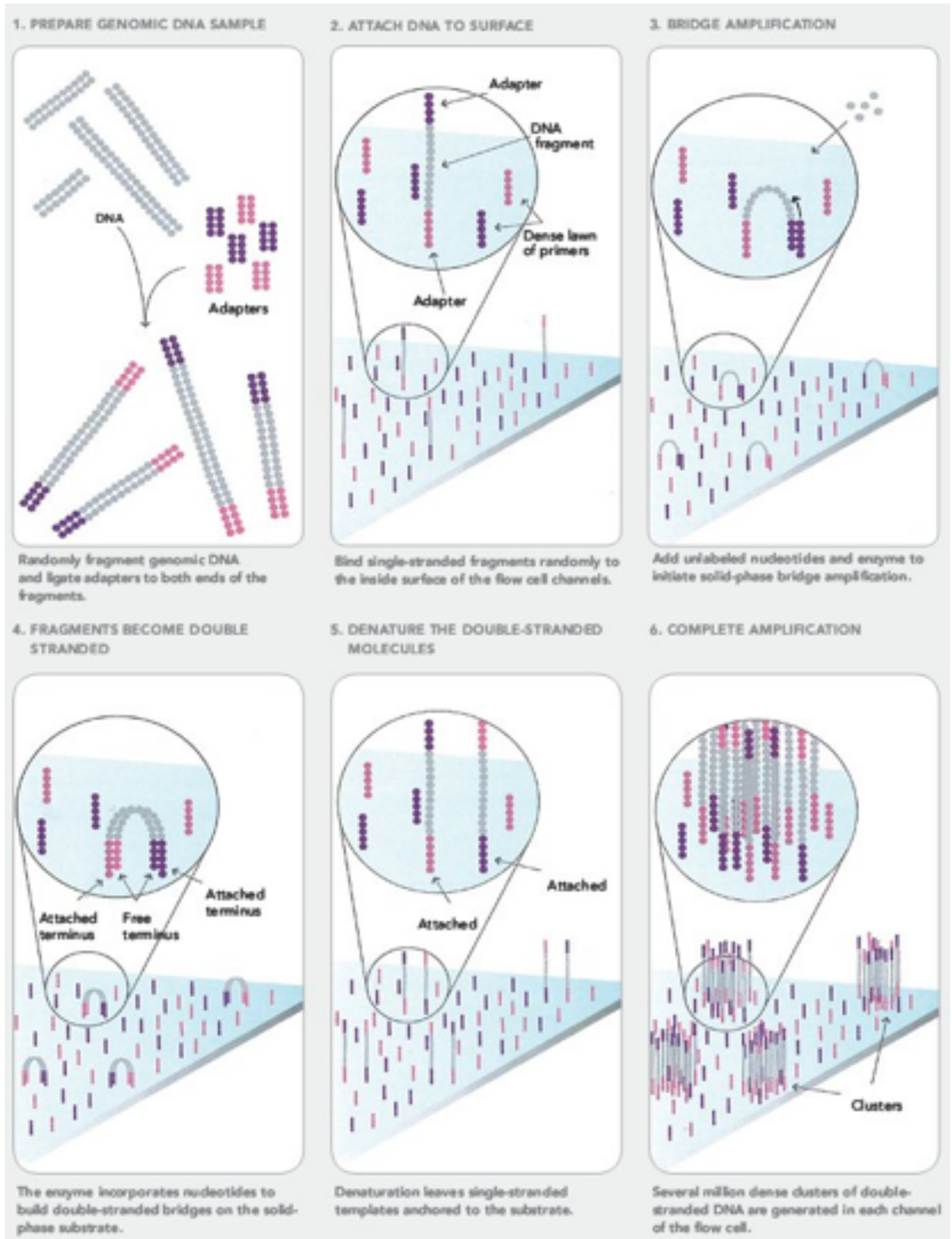


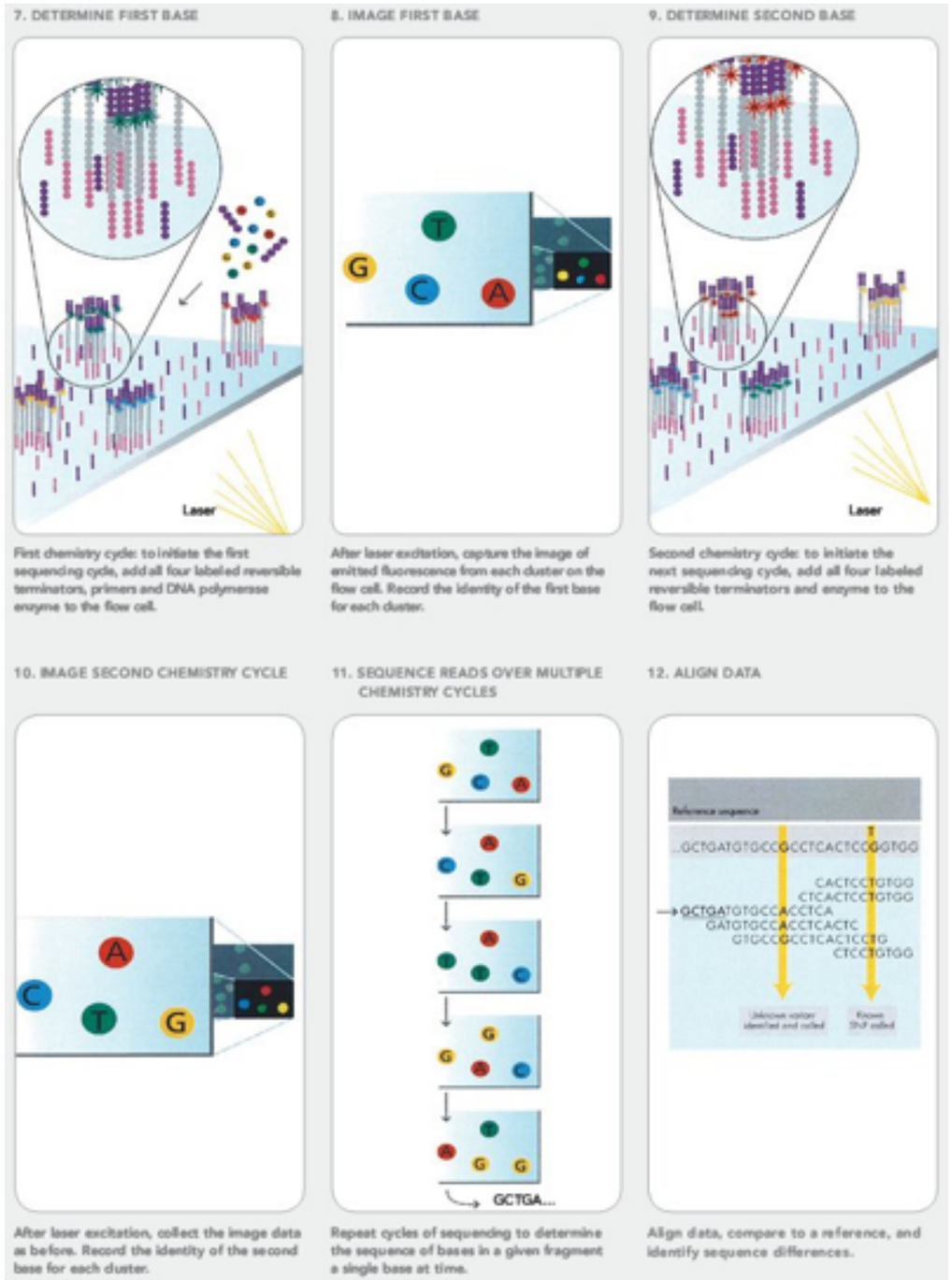
©2011, Illumina Inc. All rights reserved.

Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

https://www.illumina.com





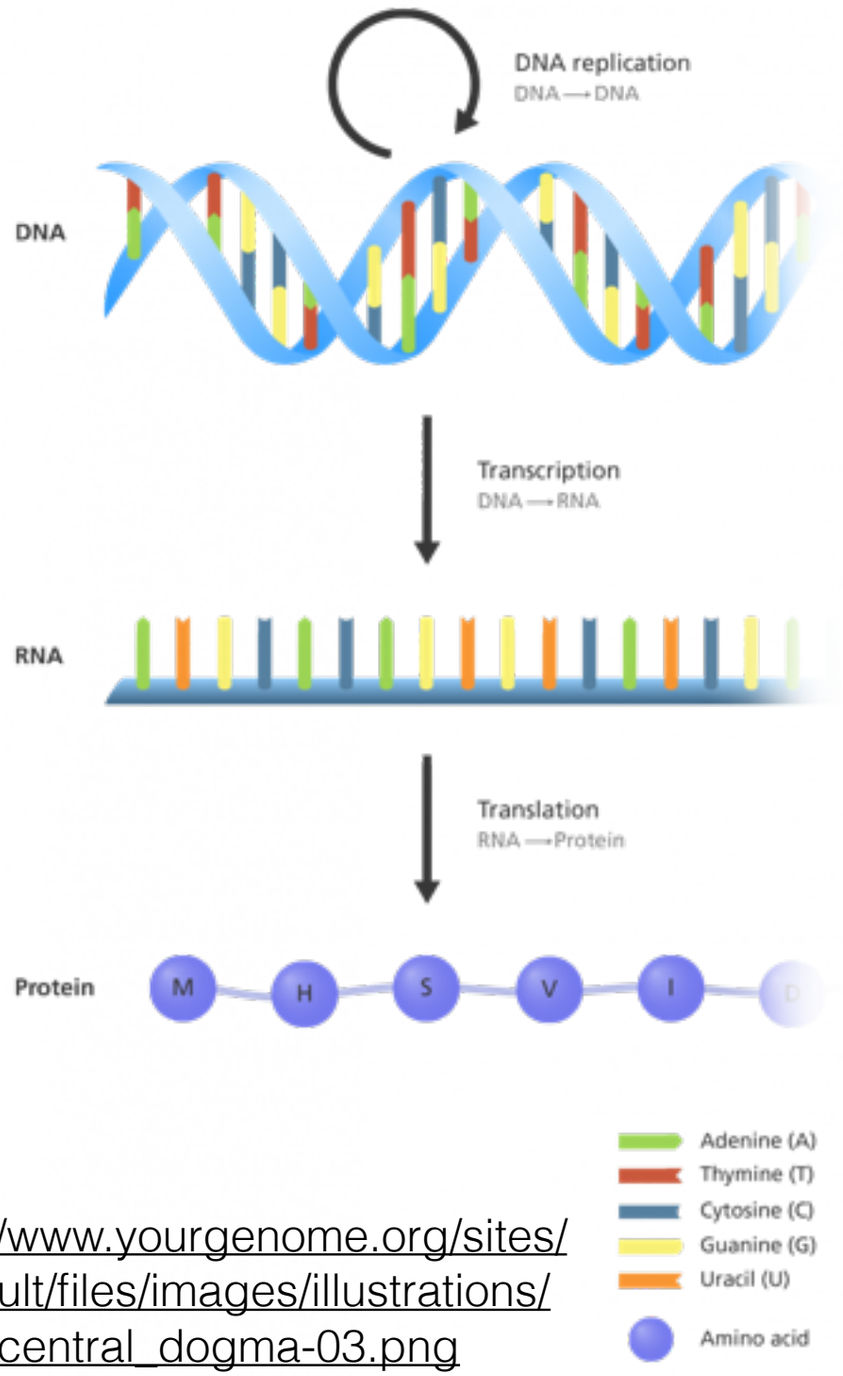
Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

https://www.illumina.com

Single-cell Amplification

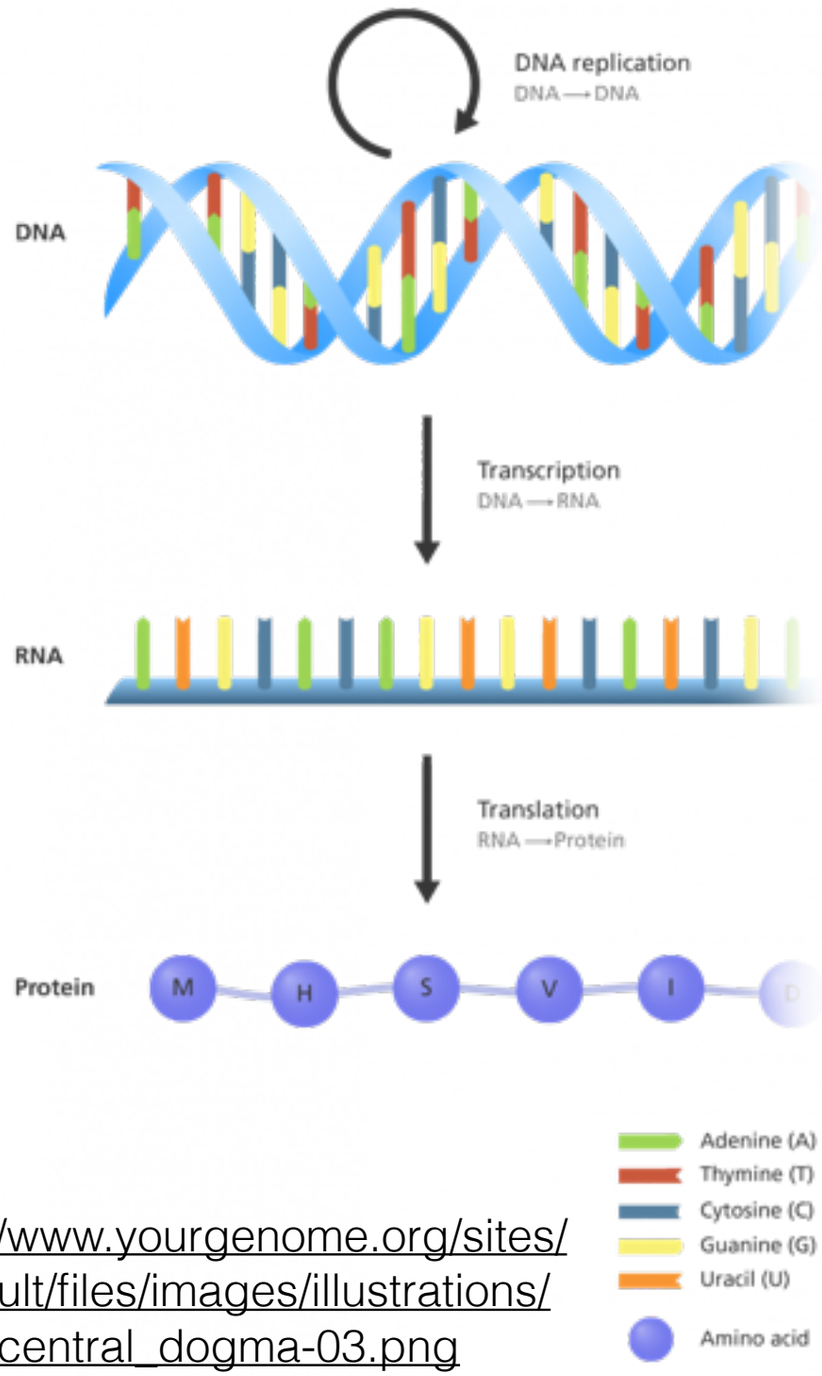
Review: RNA-Sequencing



http://www.yourgenome.org/sites/default/files/images/illustrations/central_dogma-03.png

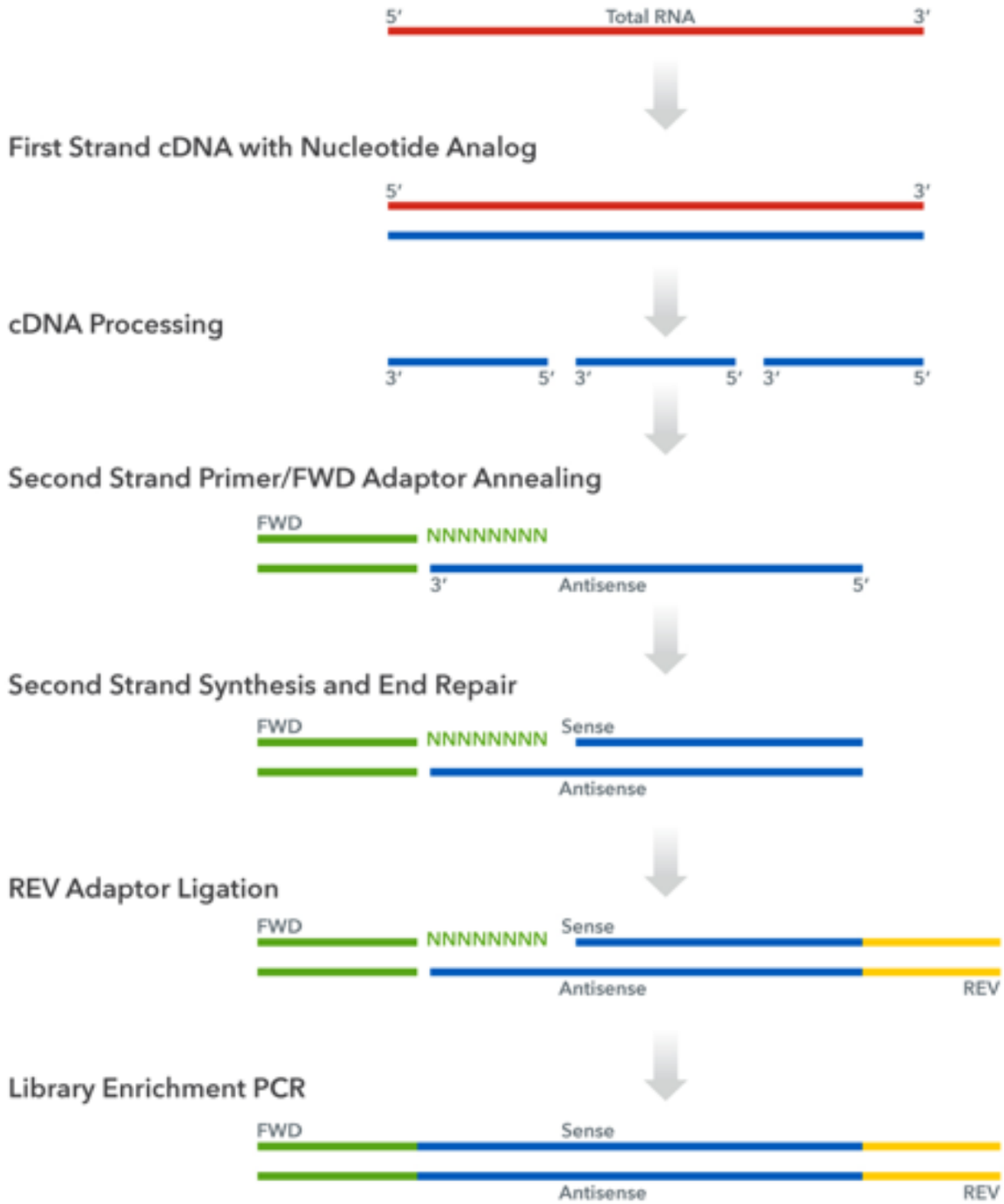
Single-cell Amplification

Review: RNA-Sequencing



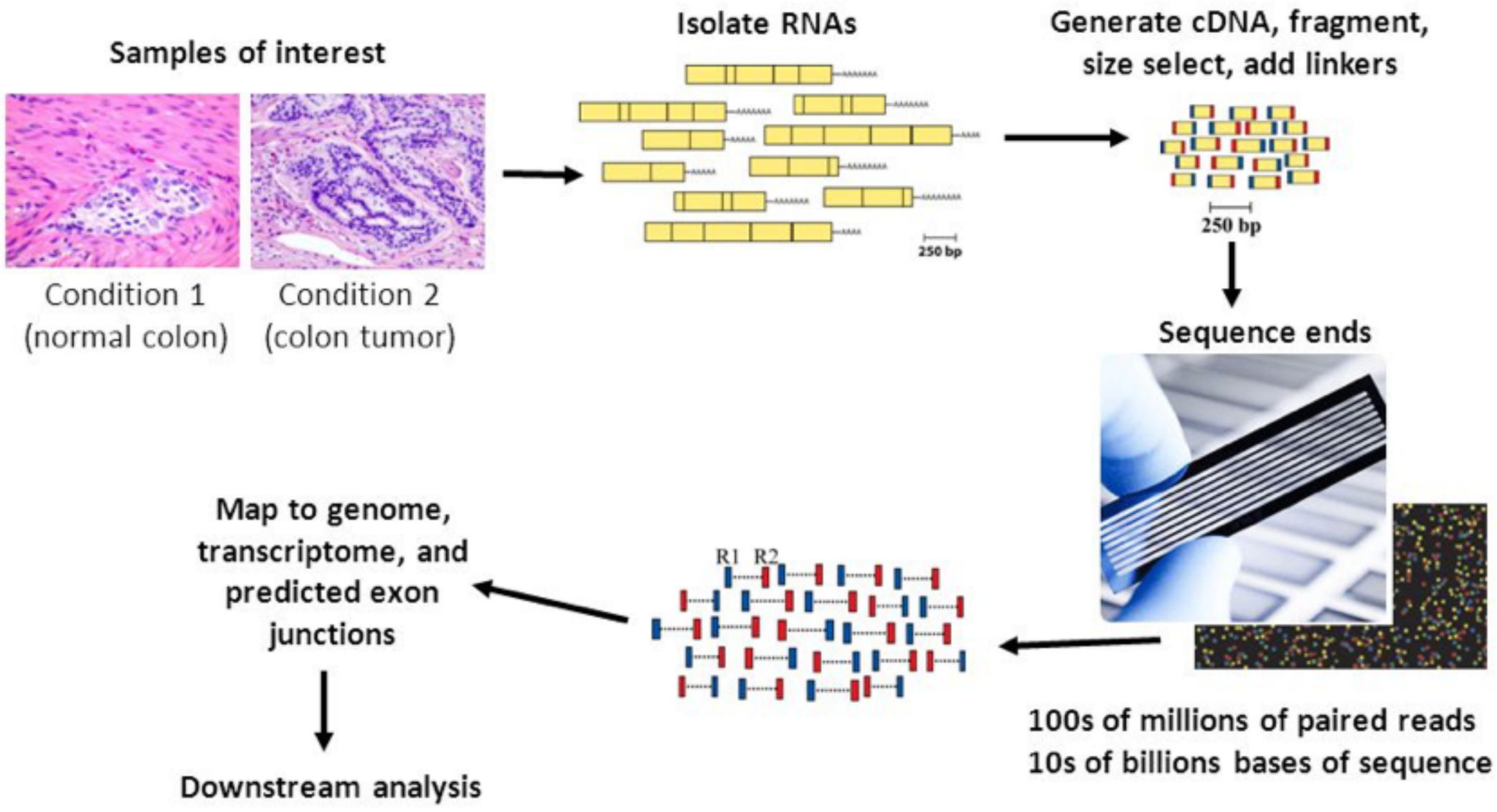
http://www.yourgenome.org/sites/default/files/images/illustrations/central_dogma-03.png

library preparation



Amplification and Sequencing

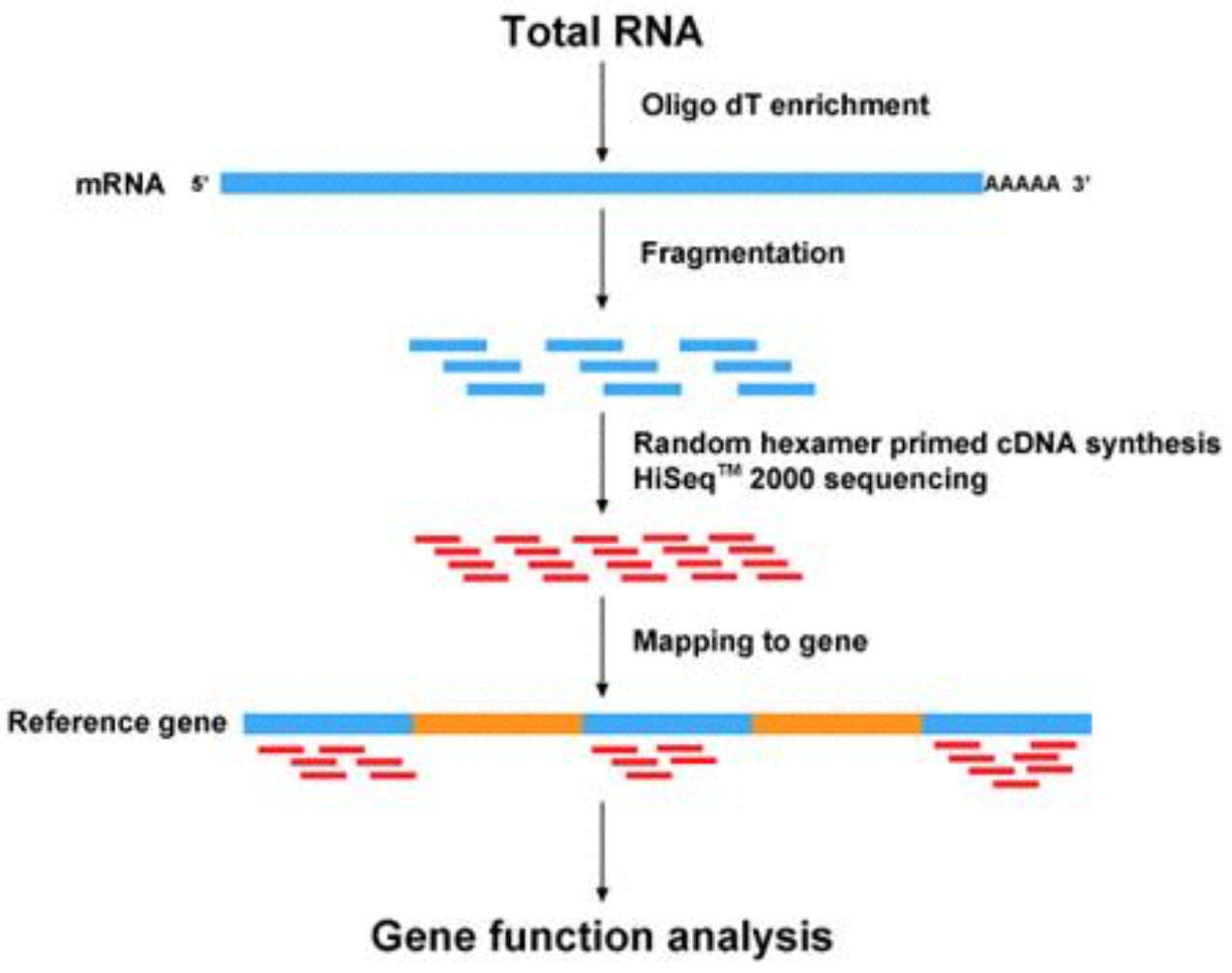
Review: RNA-Sequencing



http://images.slideplayer.com/4/1455873/slides/slide_4.jpg

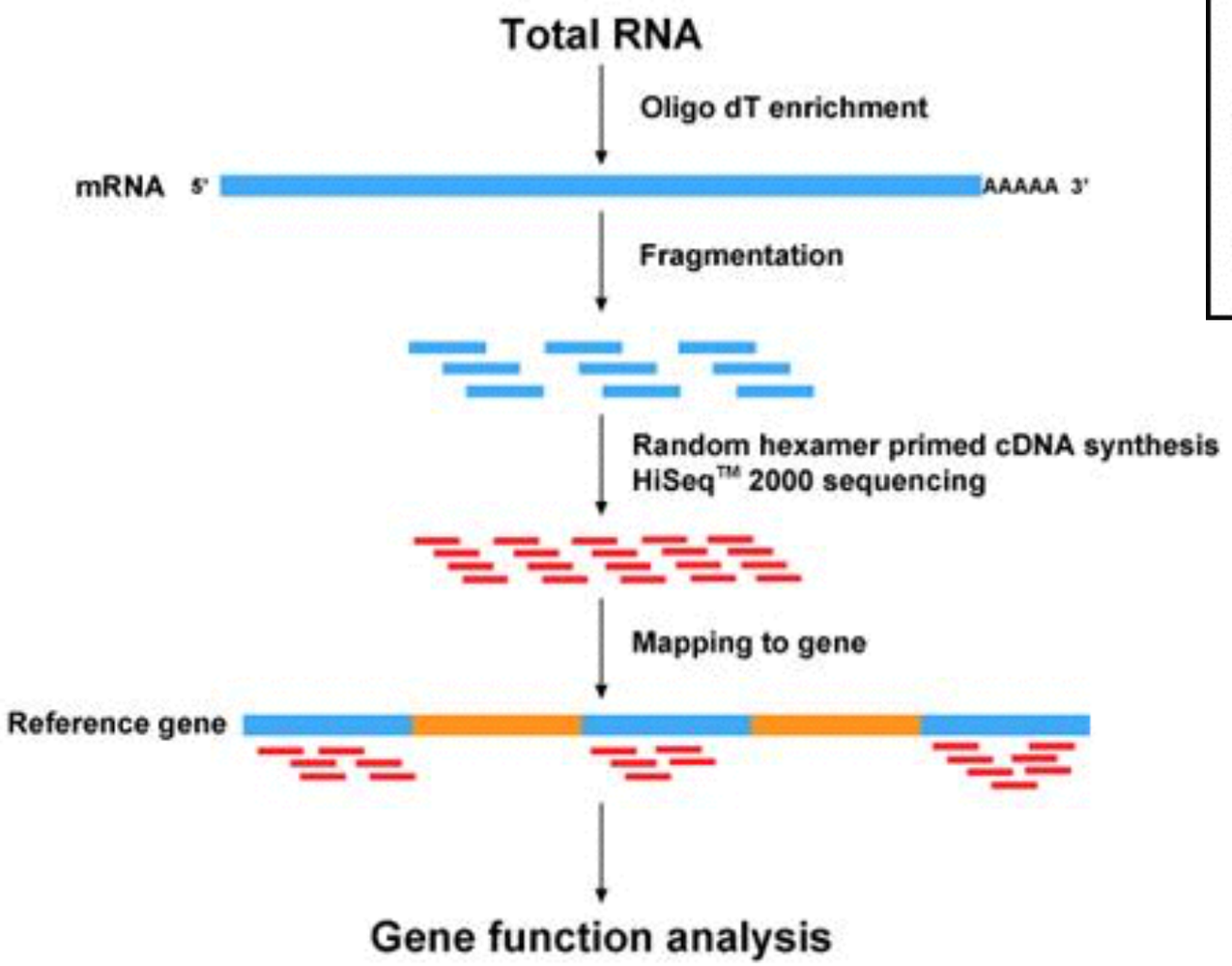
Amplification and Sequencing

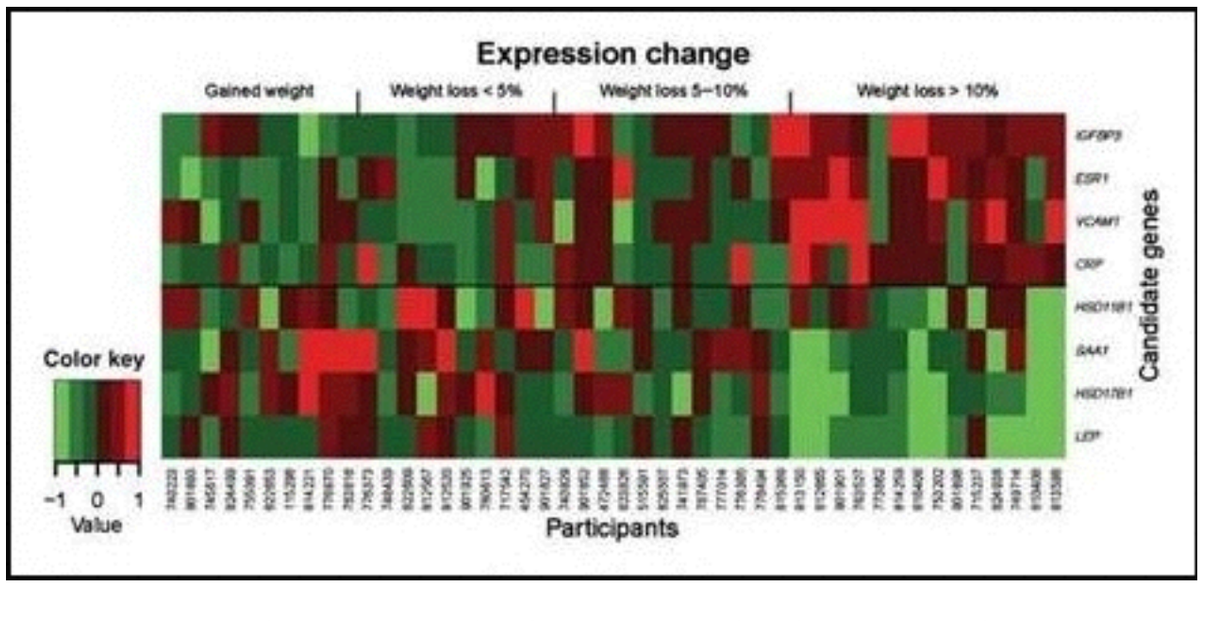
Review: RNA-Sequencing

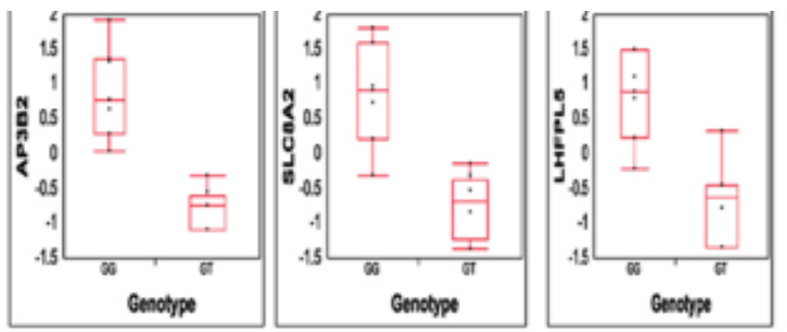


Amplification and Sequencing

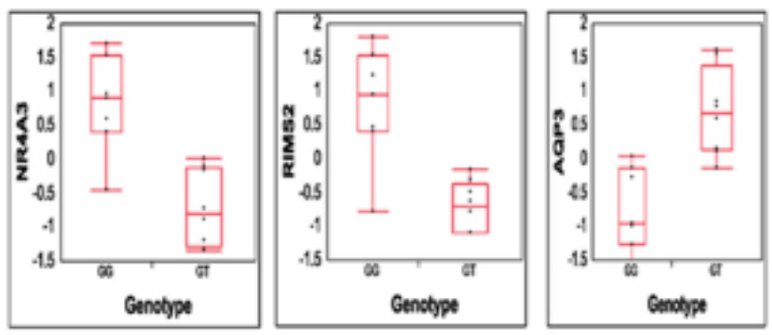
Review: RNA-Sequencing





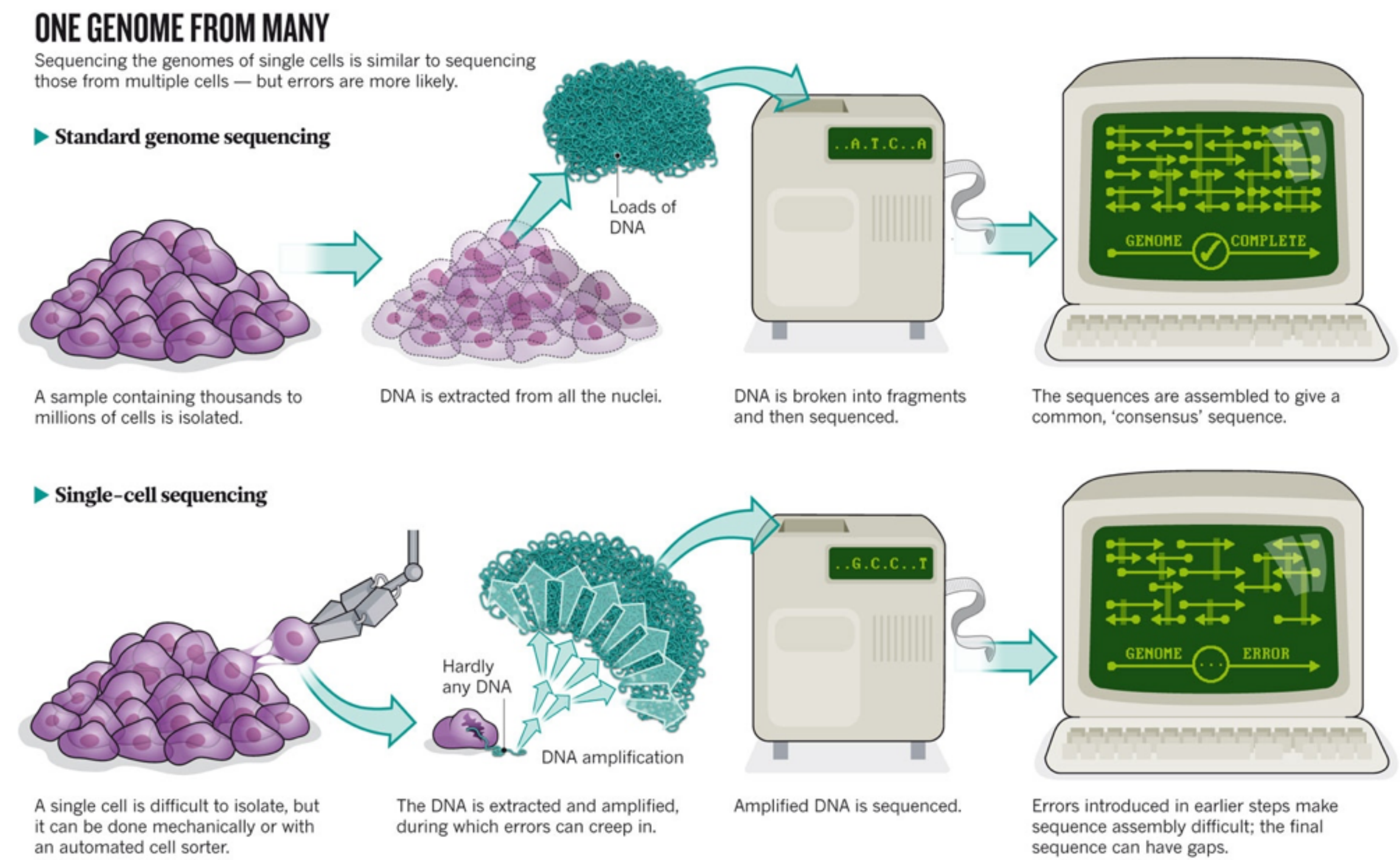


AP3B2 SLC8A2 LHPL5



NR4A3 RIMS2 AQP3

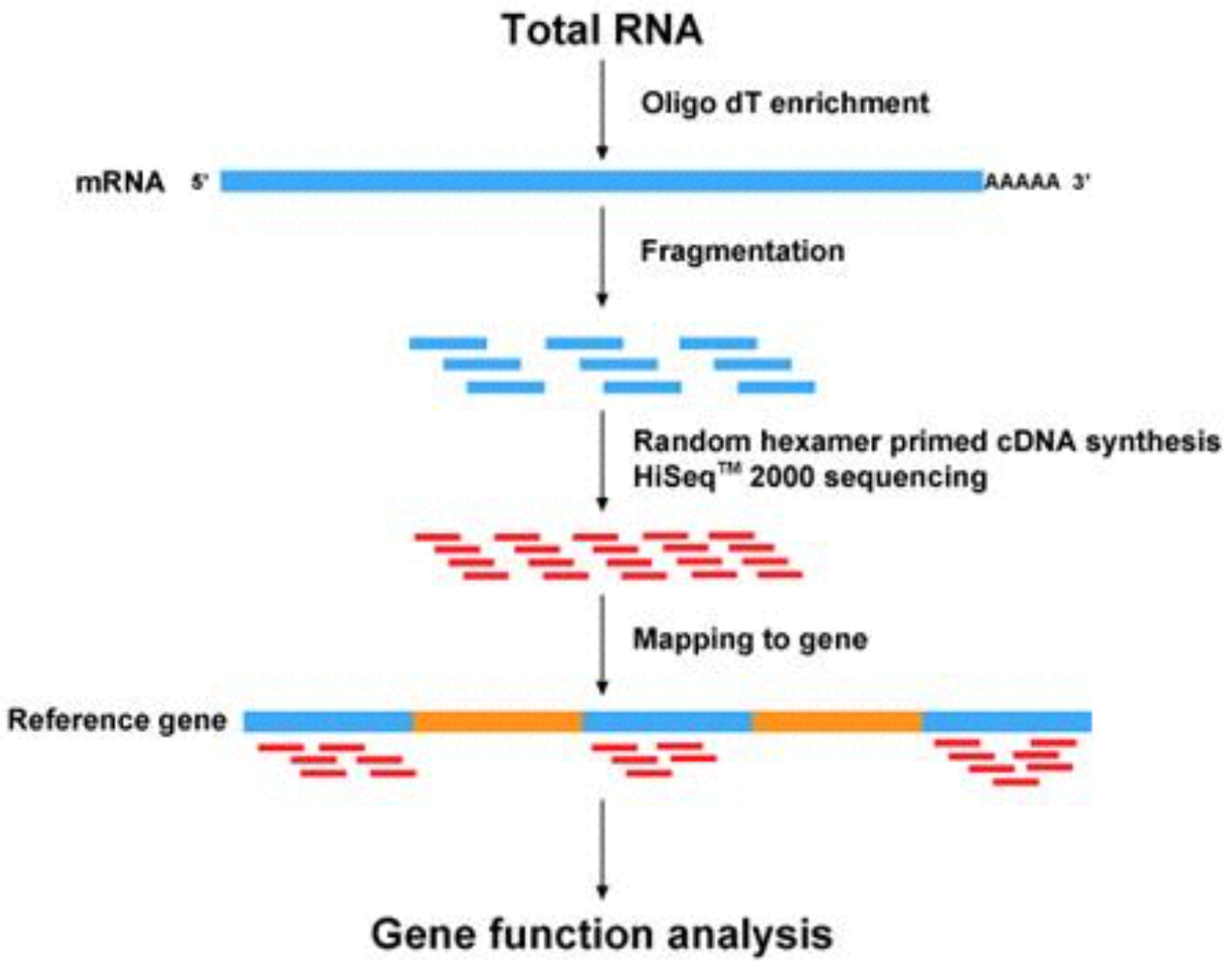
Single-cell Amplification

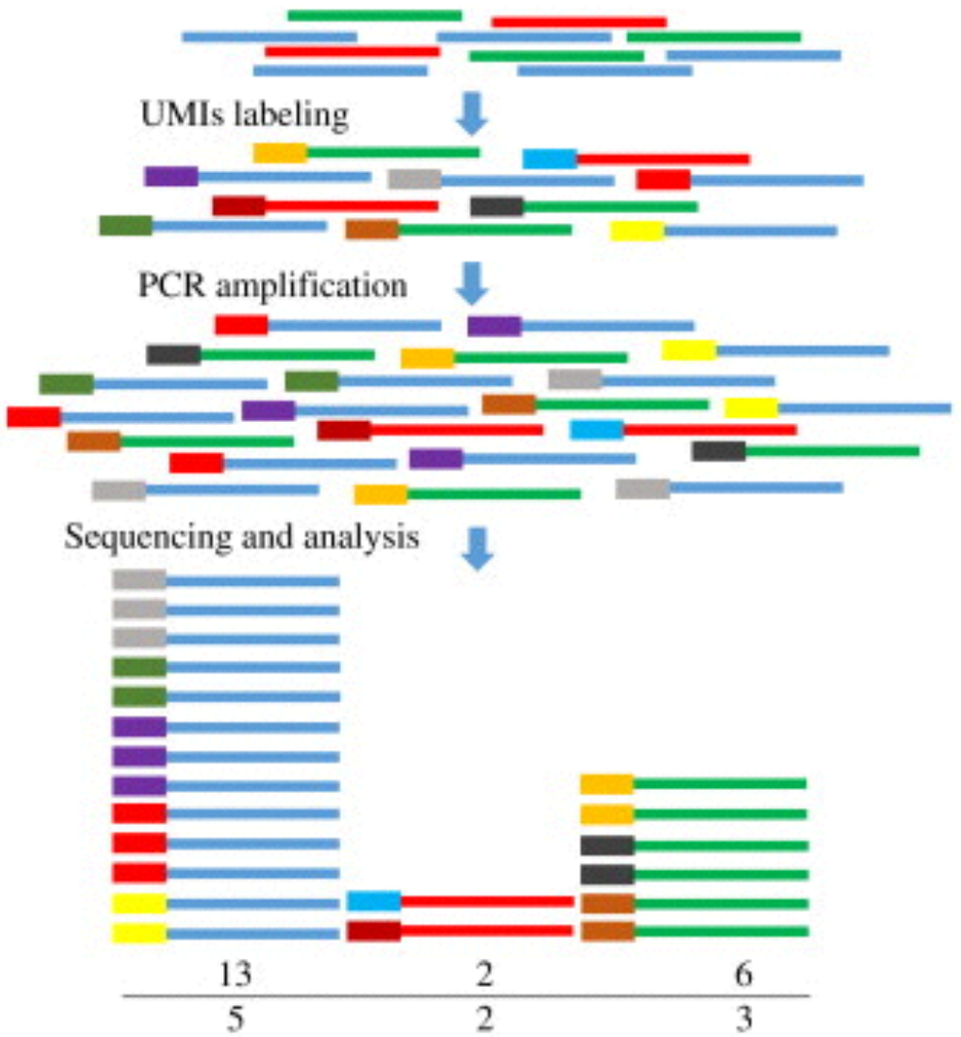


http://scitechdaily.com/images/one-genome-from-many.jpg

Single-cell Amplification

Digital Expression Matrix: counting unique molecules





http://www.nature.com/nmeth/journal/v9/n3/images_article/nmeth.f.355-F1.jpg

Short Summary

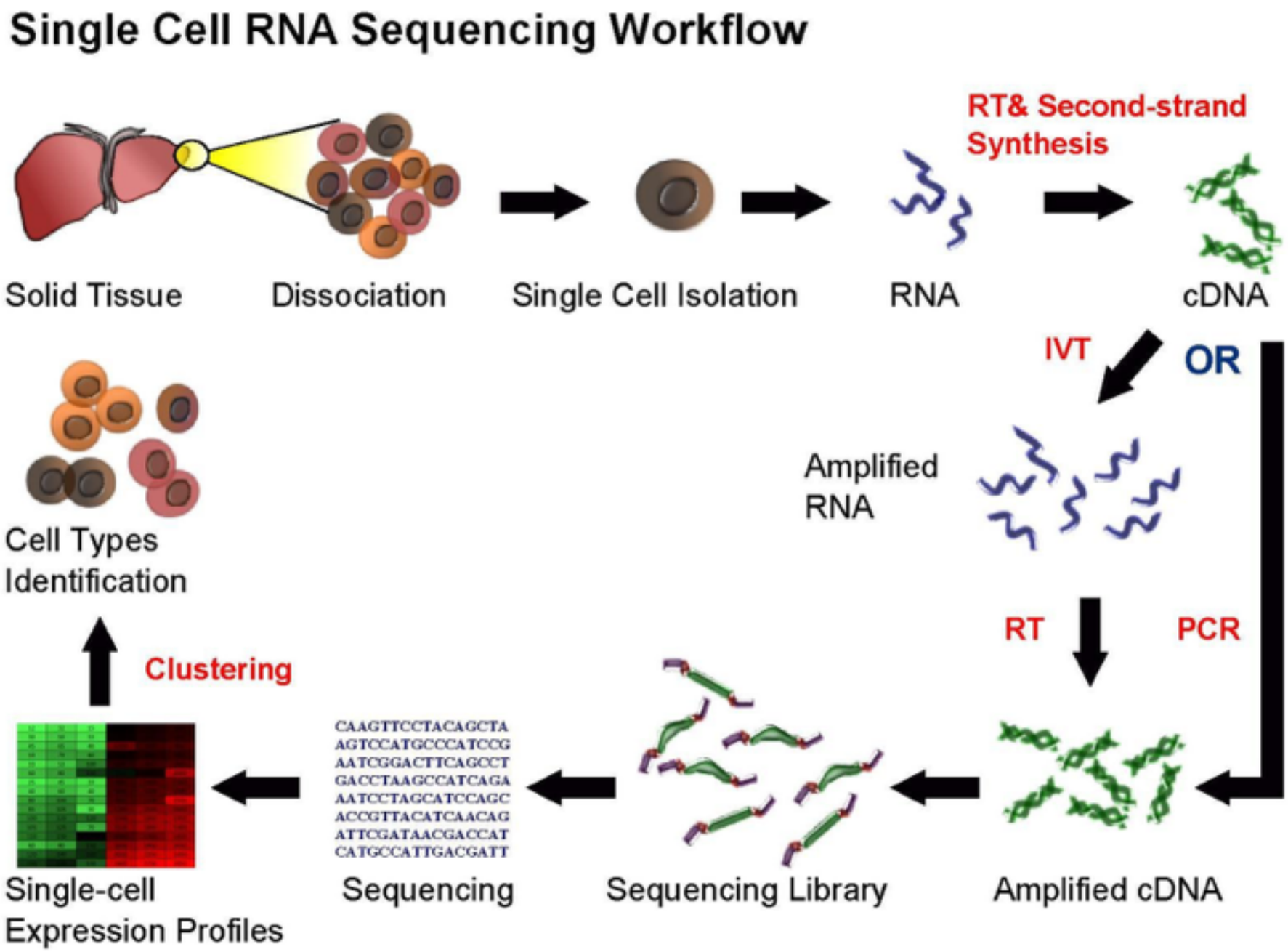
<u>DNA sequencing:</u>

- new amplification methods other than PCR
- statistical methods for SNPs/CNV calling

<u>RNA sequencing:</u>

- standards created for quality control
- can achieve high sequencing depth
- high cell throughput methods arising

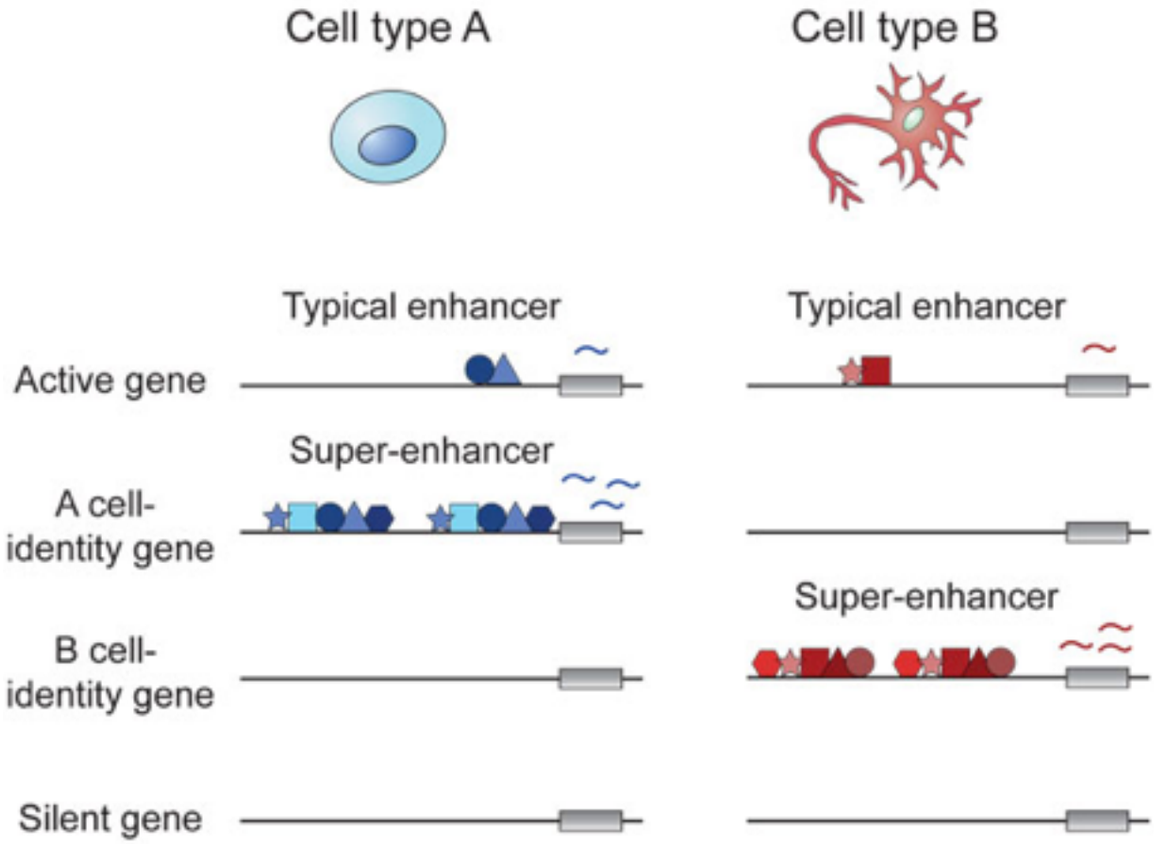
Short Summary

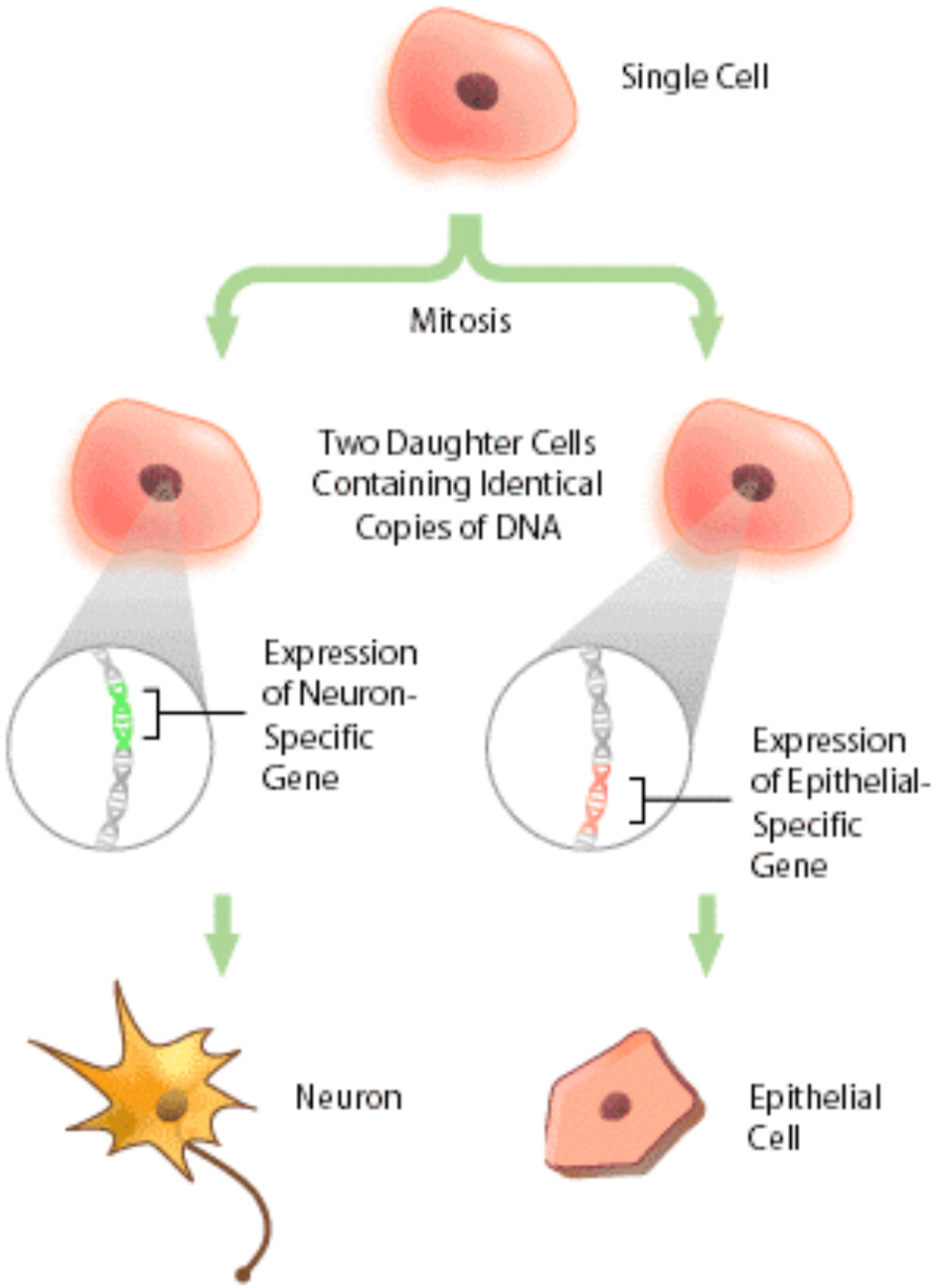


en.wikipedia.org

Downstream Analysis

Supervised Analysis





http://scienceblogs.com/clock/2006/12/07/from-two-cells-to-many-cell-di/

Cell Population Identification

Unsupervised Analysis

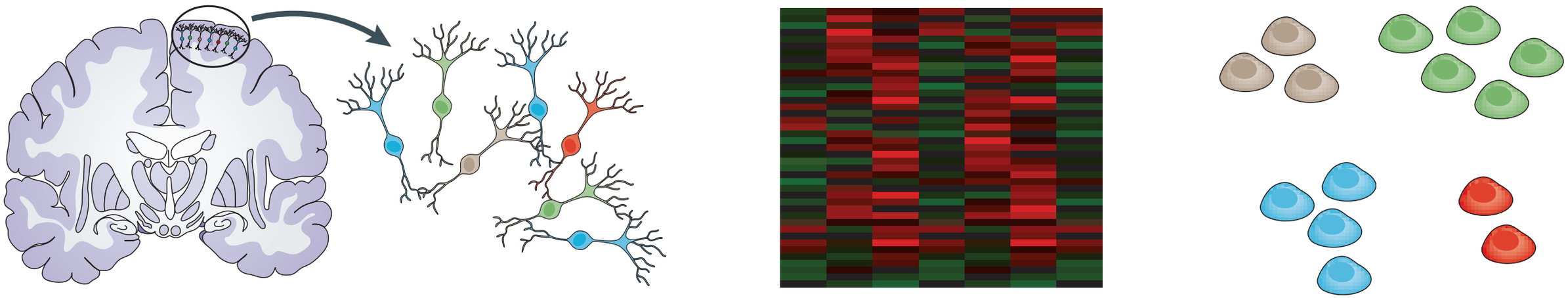


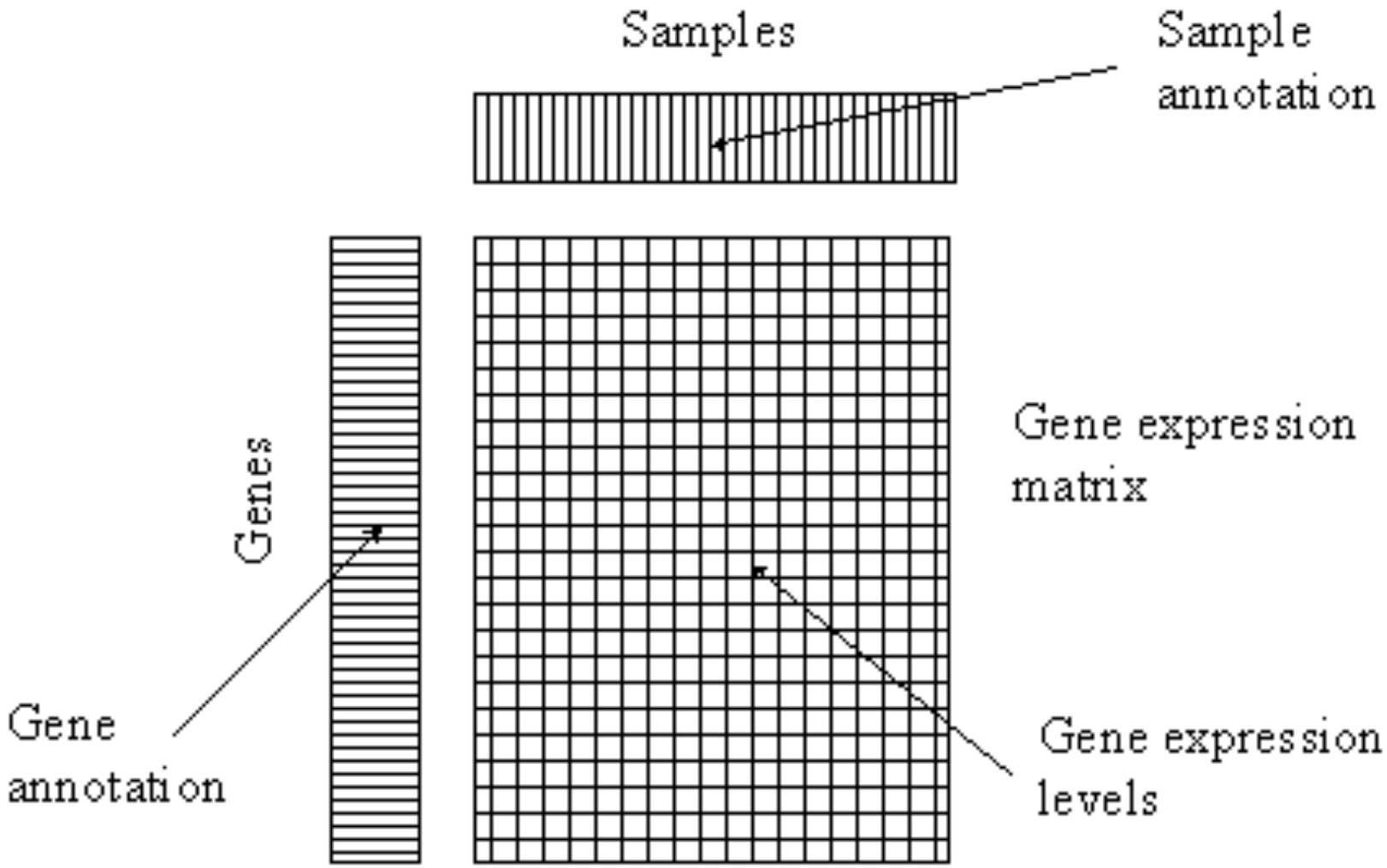
Figure 3 | **Cell-type discovery by unbiased sampling and transcriptome profiling of single cells.**

Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. "Single-cell sequencing-based technologies will revolutionize whole-organism science." *Nature Reviews Genetics* 14.9 (2013): 618-630.

Downstream Analysis

How do cell types differ from each other?

Is there any addition diversity in the same cell type?

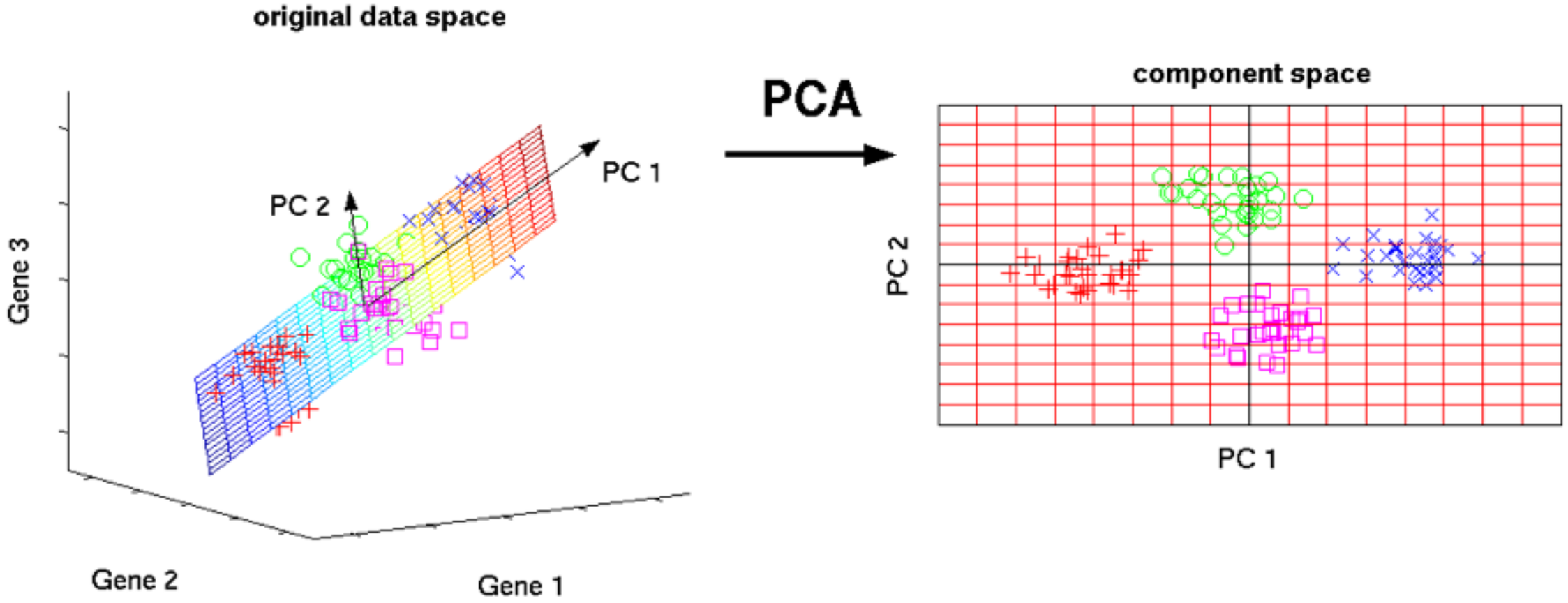


http://www.people.vcu.edu/~mreimers/OGMDA/gene_expression_matrix.gif

Dimension Reduction

Principle Component Analysis (PCA)

e.g., visualizing the samples in a smaller subspace



http://www.nlpca.org/fig_pca_principal_component_analysis.png

PCA

Probability and Linear Algebra Review

Variance / Standard Deviation: measure of the spread of the data
(Calculation: average distance from the mean of the data)

Covariance: measure of how much each of the dimensions vary from the mean with respect to each other; measured between 2 dimensions to see if there is a relationship between the 2 dimensions

* The covariance between one dimension and itself is the variance.

PCA

Probability and Linear Algebra Review

E.g. for 3 dimensions, consider random vector (x,y,z):

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}$$

Diagonal is the variances of x, y and z

cov(x,y) = cov(y,x) hence matrix is symmetrical about the diagonal

N-dimensional data will result in nxn covariance matrix

PCA

Probability and Linear Algebra Review

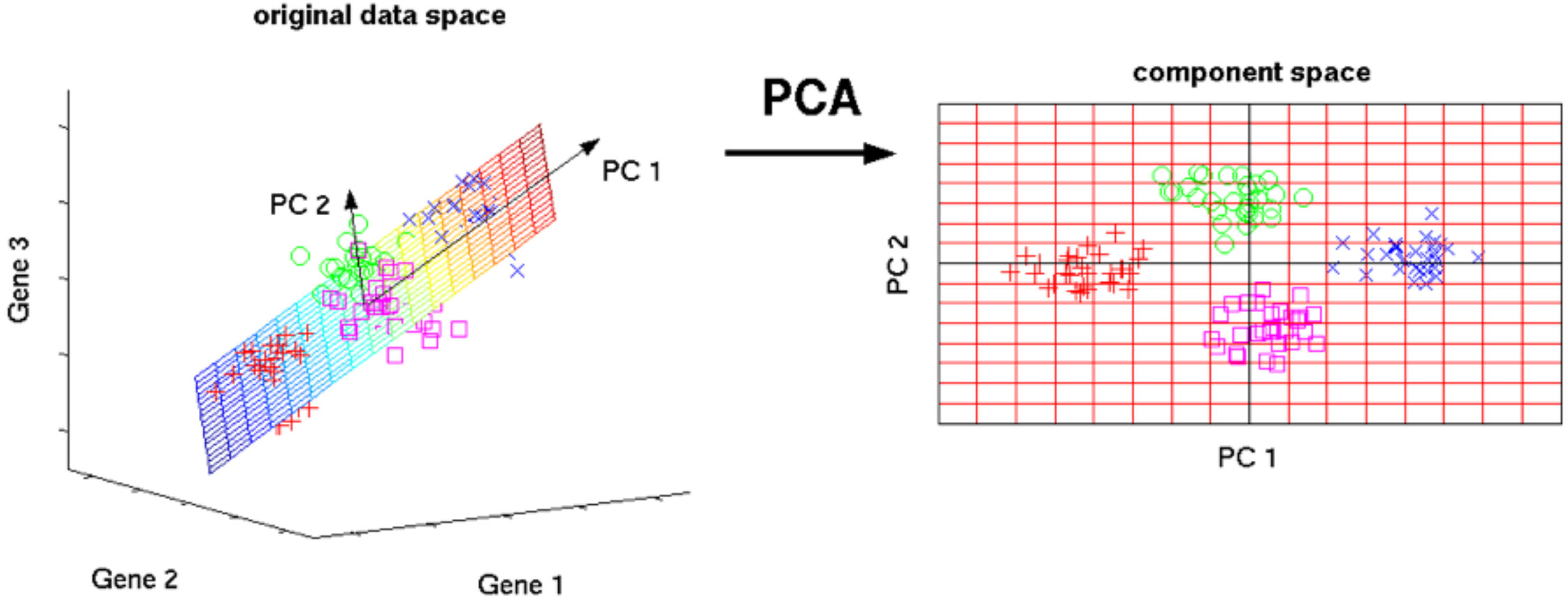
- The eigenvalue problem is any problem having the following form:

$$A \cdot v = \lambda \cdot v$$

  - A: n x n matrix
  - v: n x 1 non-zero vector
  - λ: scalar
- Any value of λ for which this equation has a solution is called the eigenvalue of A and vector v which corresponds to this value is called the eigenvector of A.

Dimension Reduction

Principle Component Analysis (PCA)



http://www.nlpca.org/fig_pca_principal_component_analysis.png

PCA

Principal component analysis (PCA) converts a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components.

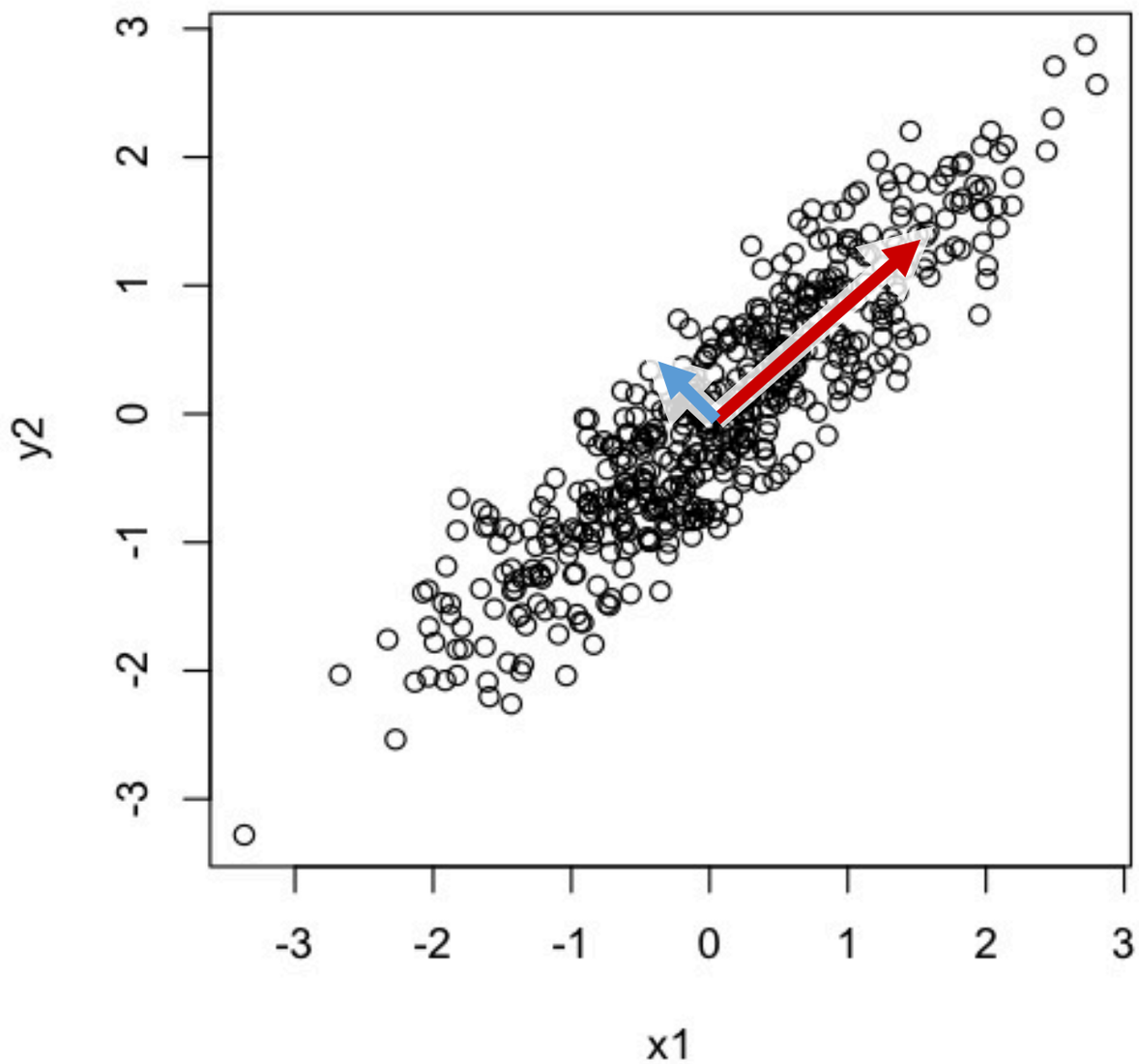
The first principal component is the projection of the data into a single dimension that has as high a variance as possible (that is, accounts for as much of the variability in the data as possible); each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components.

Therefore the PCs provide a view on the structure of the data that best explains its variance.

cf. Wikipedia: Principal component analysis

PCA

The example data is two-dimensional, but most of the information is contained along a dimension shown here by the red vector.
We could thus restrict our analysis to a projection along that vector.



PCA

PCA process –STEP 1

- Subtract the mean

  from each of the data dimensions. All the x values have x subtracted and y values have y subtracted from them. This produces a data set whose mean is zero.

  Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.

PCA

PCA process –STEP 1

DATA:

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

ZERO MEAN DATA:

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA

PCA process –STEP 2

- Calculate the covariance matrix

$$\text{cov} = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

- since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

PCA

PCA process –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

PCA

PCA process –STEP 4

- Reduce dimensionality and form *feature vector* the eigenvector with the *highest* eigenvalue is the *principle component* of the data set.

- In our example, the eigenvector with the larges eigenvalue was the one that pointed down the middle of the data.

- Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance.

PCA

PCA process –STEP 4

- Now, if you like, you can decide to *ignore* the components of lesser significance

- You do lose some information, but if the eigenvalues are small, you don’t lose much

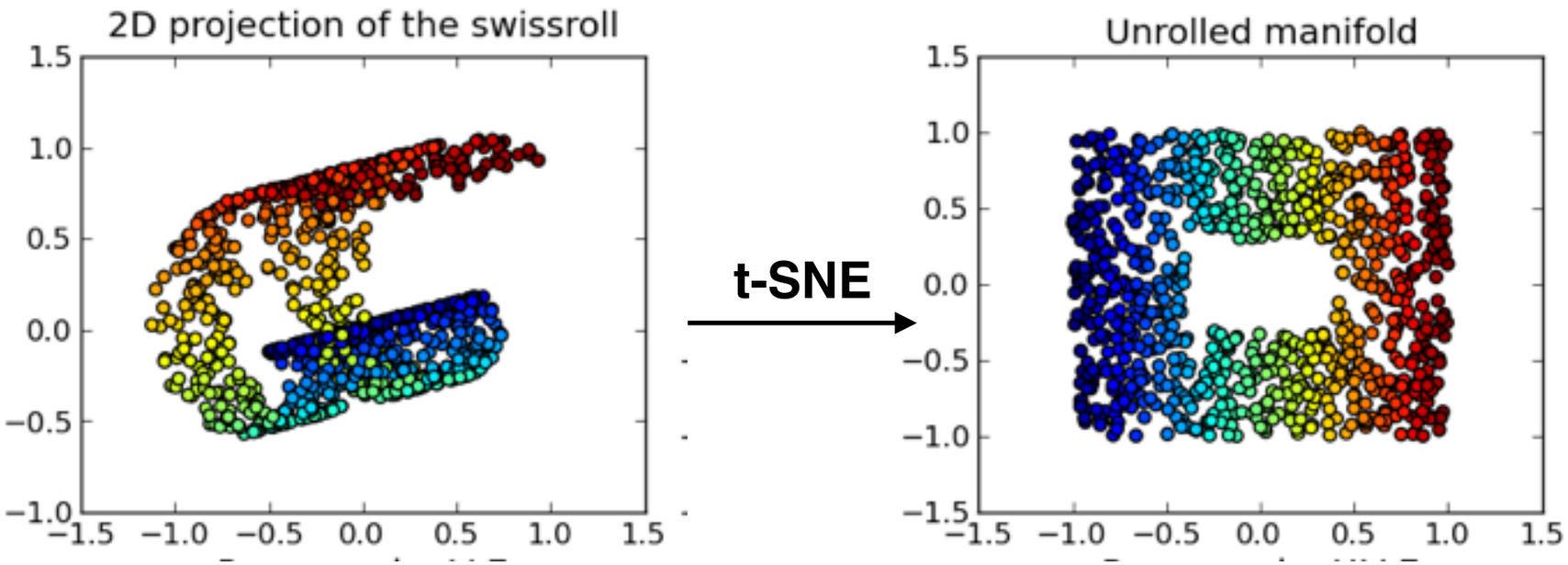
  - n dimensions in your data
  - calculate n eigenvectors and eigenvalues
  - choose only the first p eigenvectors
  - final data set has only p dimensions.

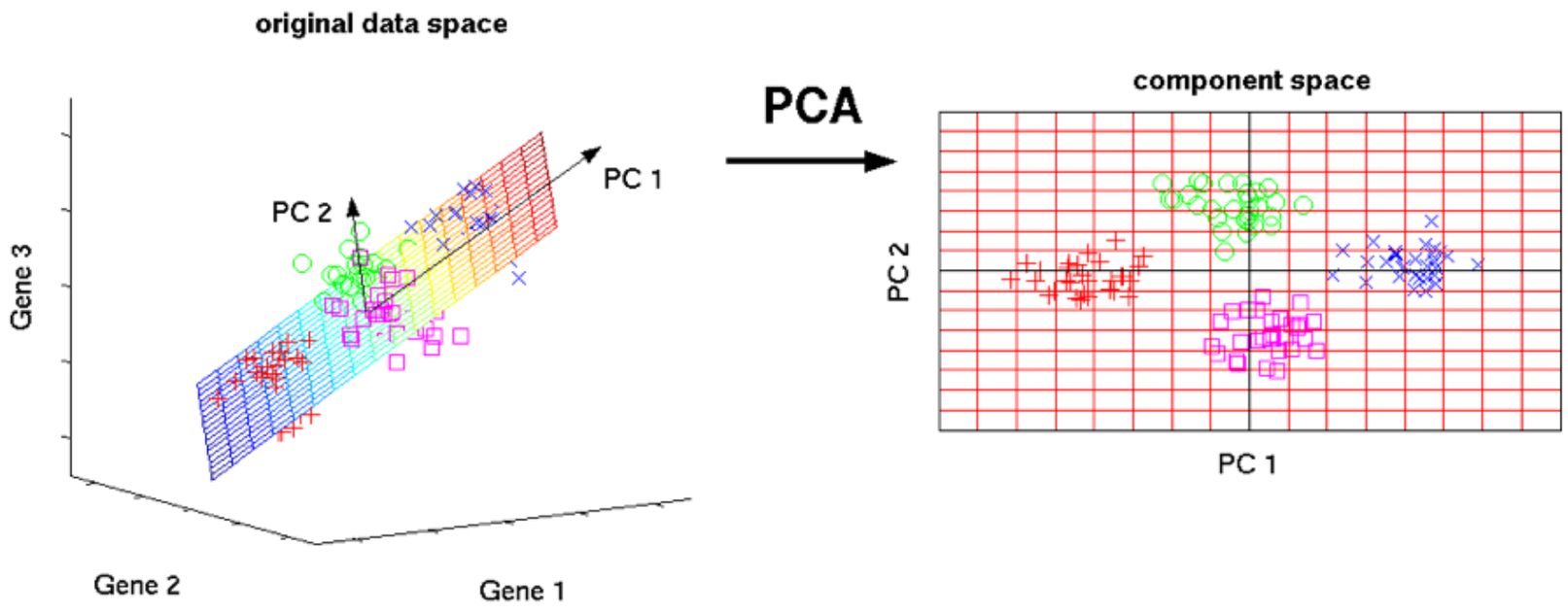
Dimension Reduction

Principle Component Analysis (PCA)

- linear multivariate statistical analysis
- understand underlying data structures
- identify bias, experimental errors, batch effects
- visualize the samples in a smaller subspace (dimension reduction)
- visualize the relationship between variables (correlation analysis)

t-SNE


2D projection of the swissroll
Unrolled manifold
t-SNE
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5


original data space
PCA
component space
PC 1
PC 2
Gene 1
Gene 2
Gene 3

t-SNE

Key quantities

high-dimensional
joint distribution

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Kullback–Leibler divergence
(to be minimized)

low-dimensional
joint distribution

Cluster Analysis

Cluster: a collection of data objects

- Similar to the objects in the same cluster (Intraclass similarity)
- Dissimilar to the objects in other clusters (Interclass dissimilarity)

Cluster analysis

- Statistical method for grouping a set of data objects into clusters
- A good clustering method produces high quality clusters with high intraclass similarity and low interclass similarity

Clustering is an unsupervised classification method

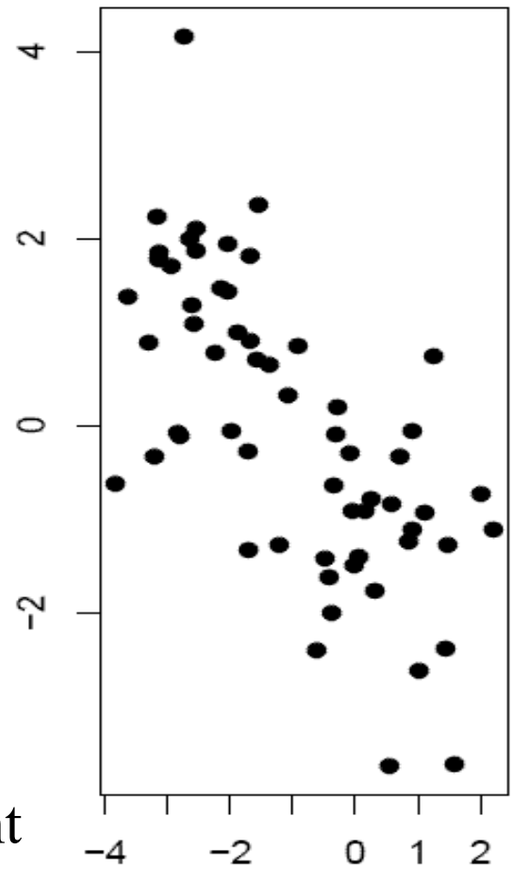
Can be a stand-alone tool or as a preprocessing step for other algorithms

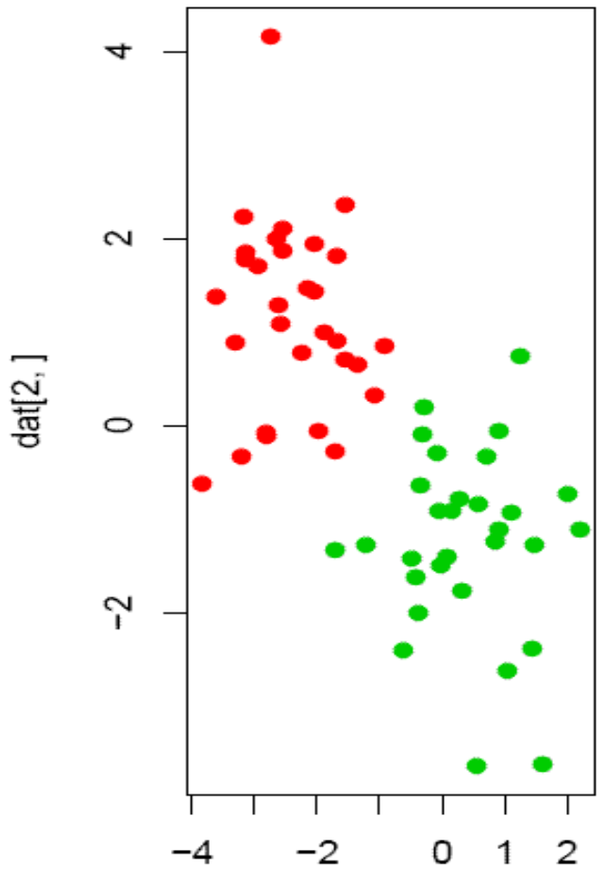
Cluster Analysis

Group objects according to their similarity

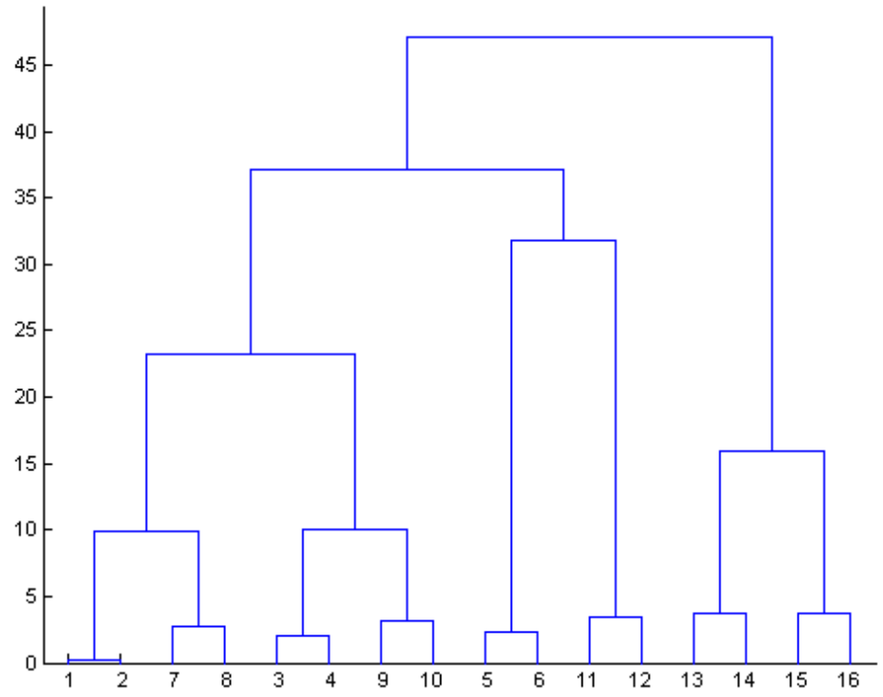
Cluster:
a set of objects that are similar to each other and separated from the other objects.

Example: green/ red data points were generated from two different normal distributions





Hierarchical Clustering



- This produces a binary tree or ***dendrogram***
- The final cluster is the root and each data item is a leaf
- The height of the bars indicate how close the items are

Hierarchical Clustering

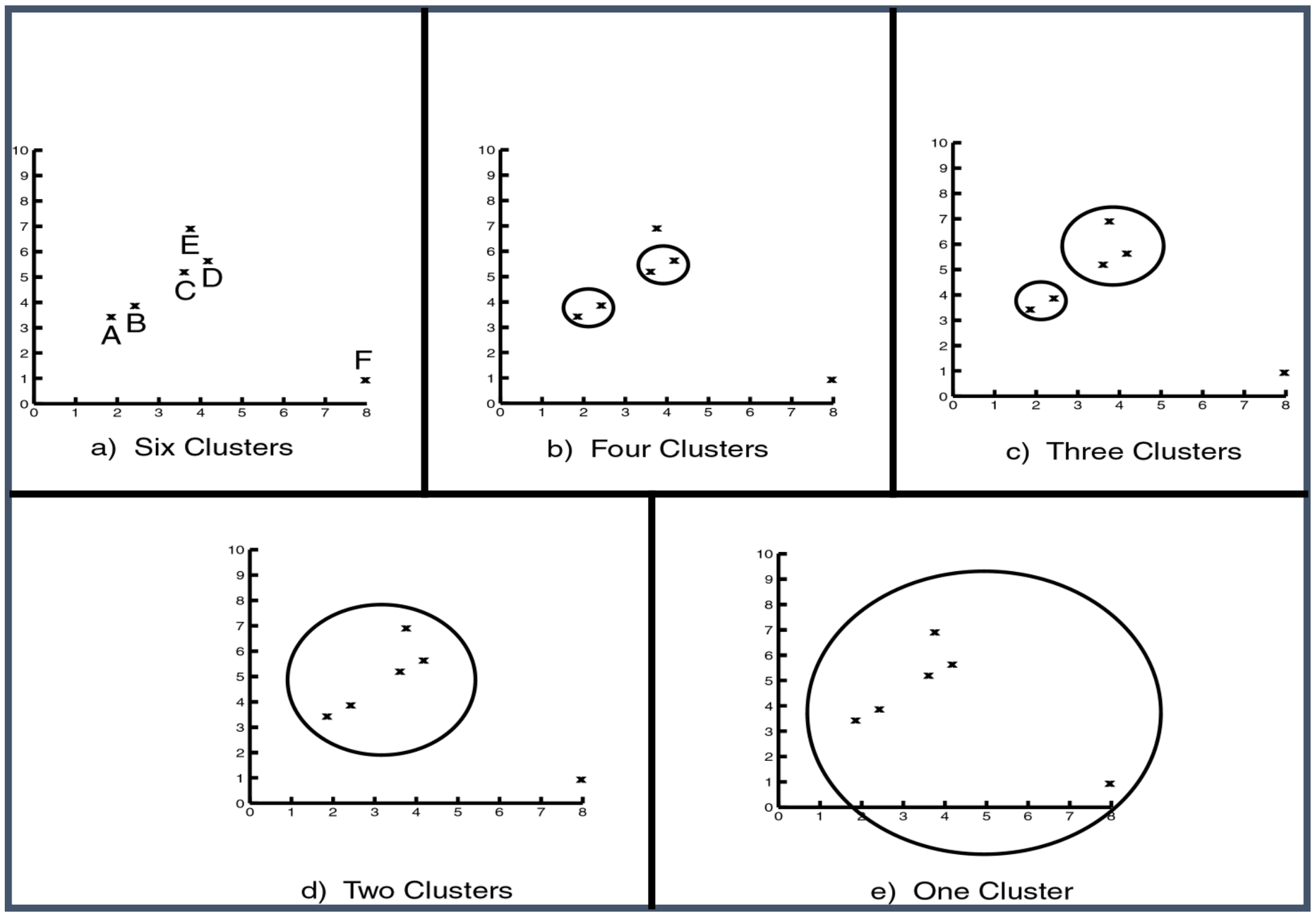
Start with every data point in a separate cluster

Keep merging the most similar pairs of data points/clusters until we have one big cluster left

This is called a bottom-up or agglomerative method

Hierarchical Clustering

Levels of Clustering



Hierarchical Clustering

Linkage in Hierarchical Clustering

We already know about distance measures between data items, but what about between a data item and a cluster or between two clusters?

We just treat a data point as a cluster with a single item, so our only problem is to define a ***linkage*** method between clusters

As usual, there are lots of choices…

Hierarchical Clustering

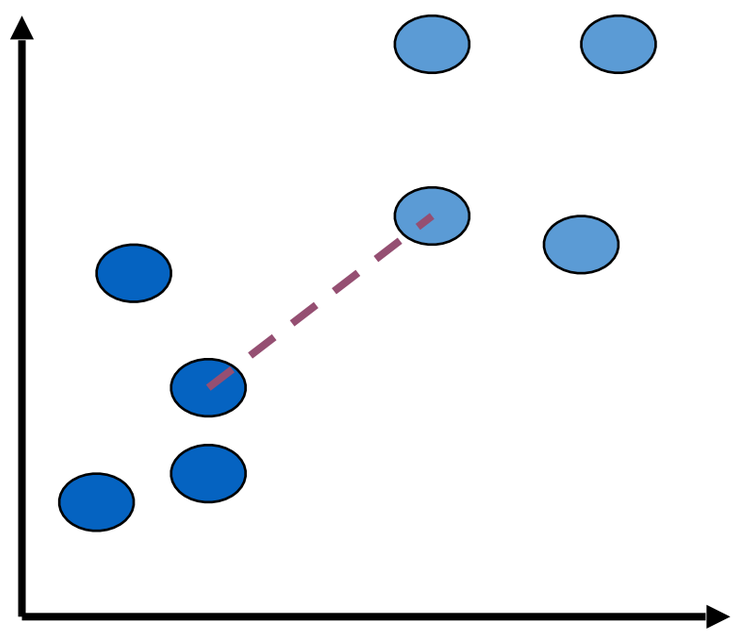
Average Linkage

- Definition
  - Each cluster *ci* is associated with a mean vector μi which is the mean of all the data items in the cluster
  - The distance between two clusters *ci* and *cj* is then just $d(\mu i , \mu j)$
- This is somewhat non-standard – this method is usually referred to as centroid linkage and average linkage is defined as the average of all pairwise distances between points in the two clusters

Hierarchical Clustering

Single Linkage

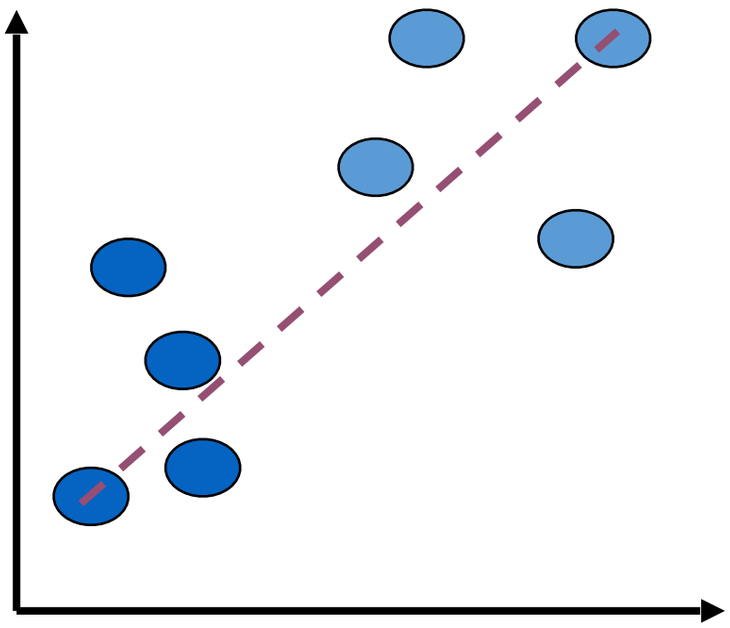
- The minimum of all pairwise distances between points in the two clusters
- Tends to produce long, “loose” clusters



Hierarchical Clustering

Complete Linkage

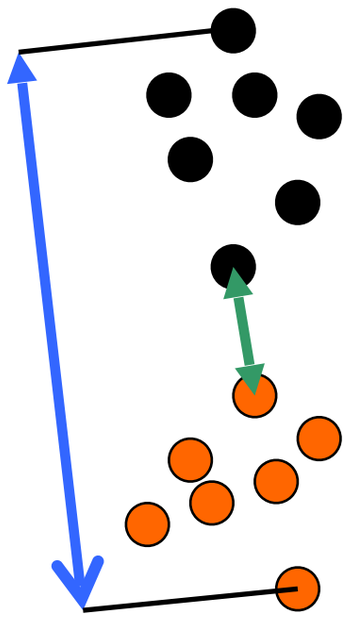
- The maximum of all pairwise distances between points in the two clusters
- Tends to produce very tight clusters



Hierarchical Clustering

Distances between clusters (summary)

- Calculation of the distance between two clusters is based on the pairwise distances between members of the clusters.
  - **Complete linkage**: largest distance between points
- **Average linkage**: average distance between paris of points
- **Single linkage**: smallest distance between points
- ***Centroid:*** distance between centroids



Complete linkage gives preference to compact/spherical clusters. Single linkage can produce long stretched clusters.

Hierarchical Clustering

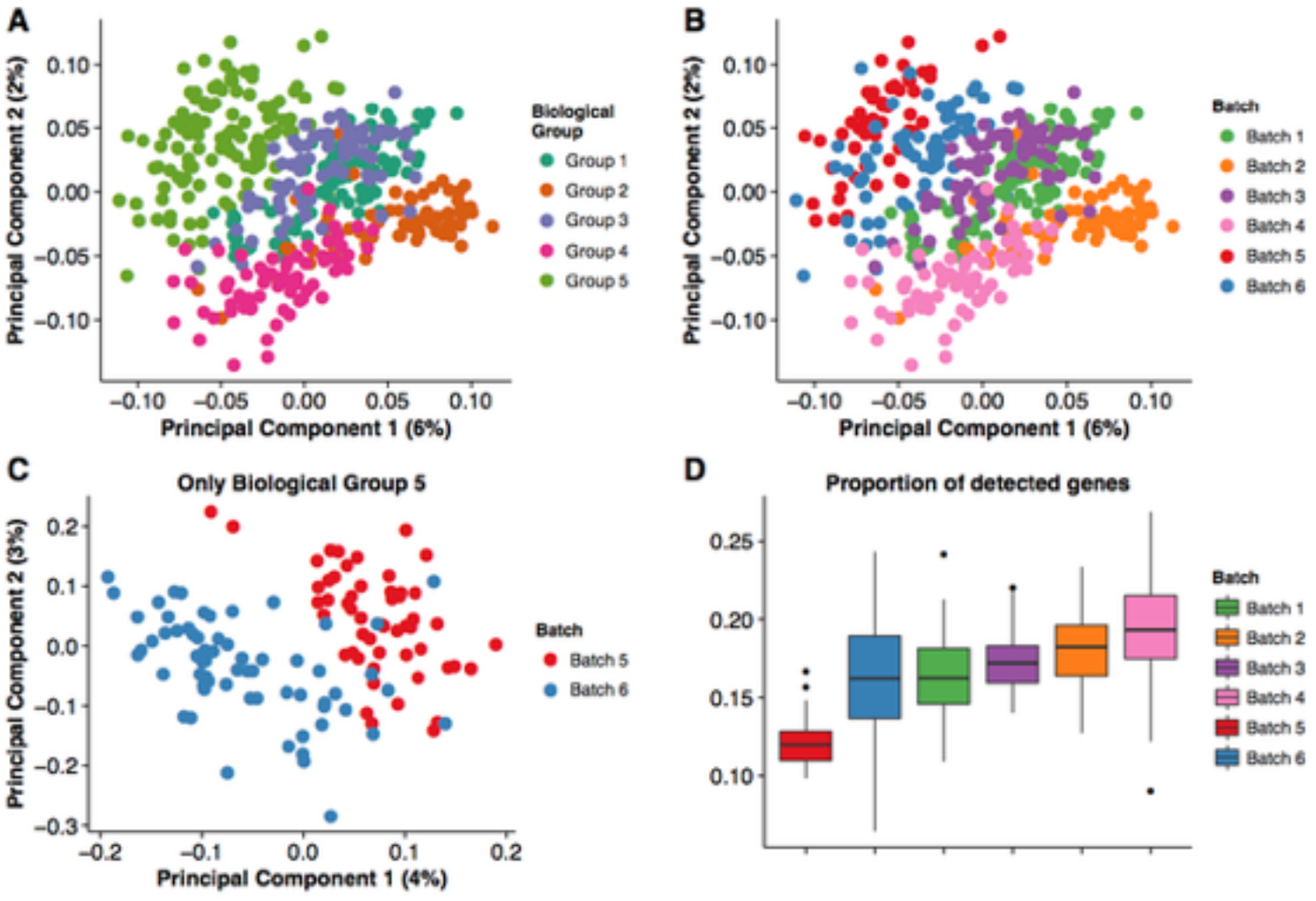
- Major advantage
  - Conceptually very simple
  - Easy to implement → most commonly used technique

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo what was done previously → high likelihood of getting stuck in local minima

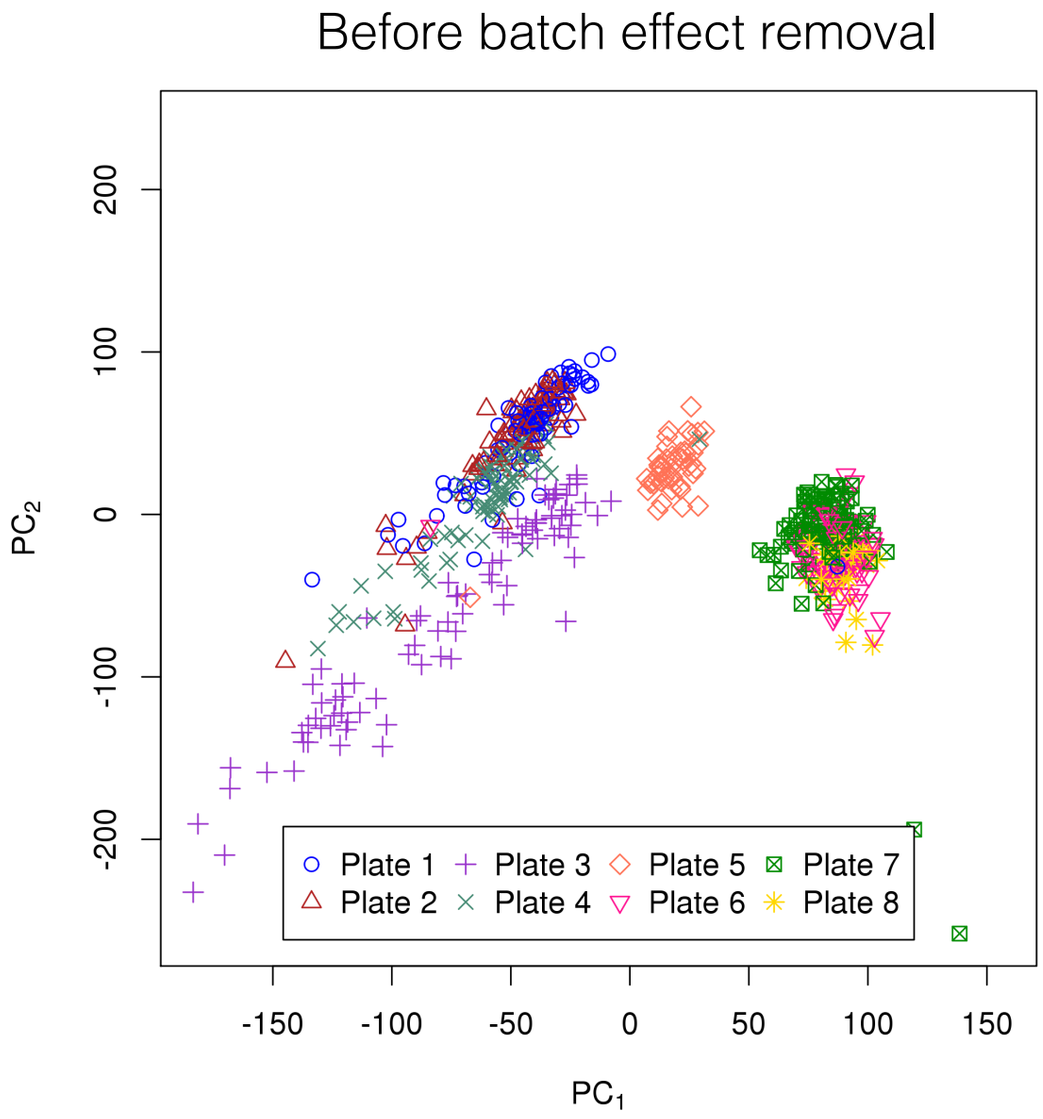
Other Challenges

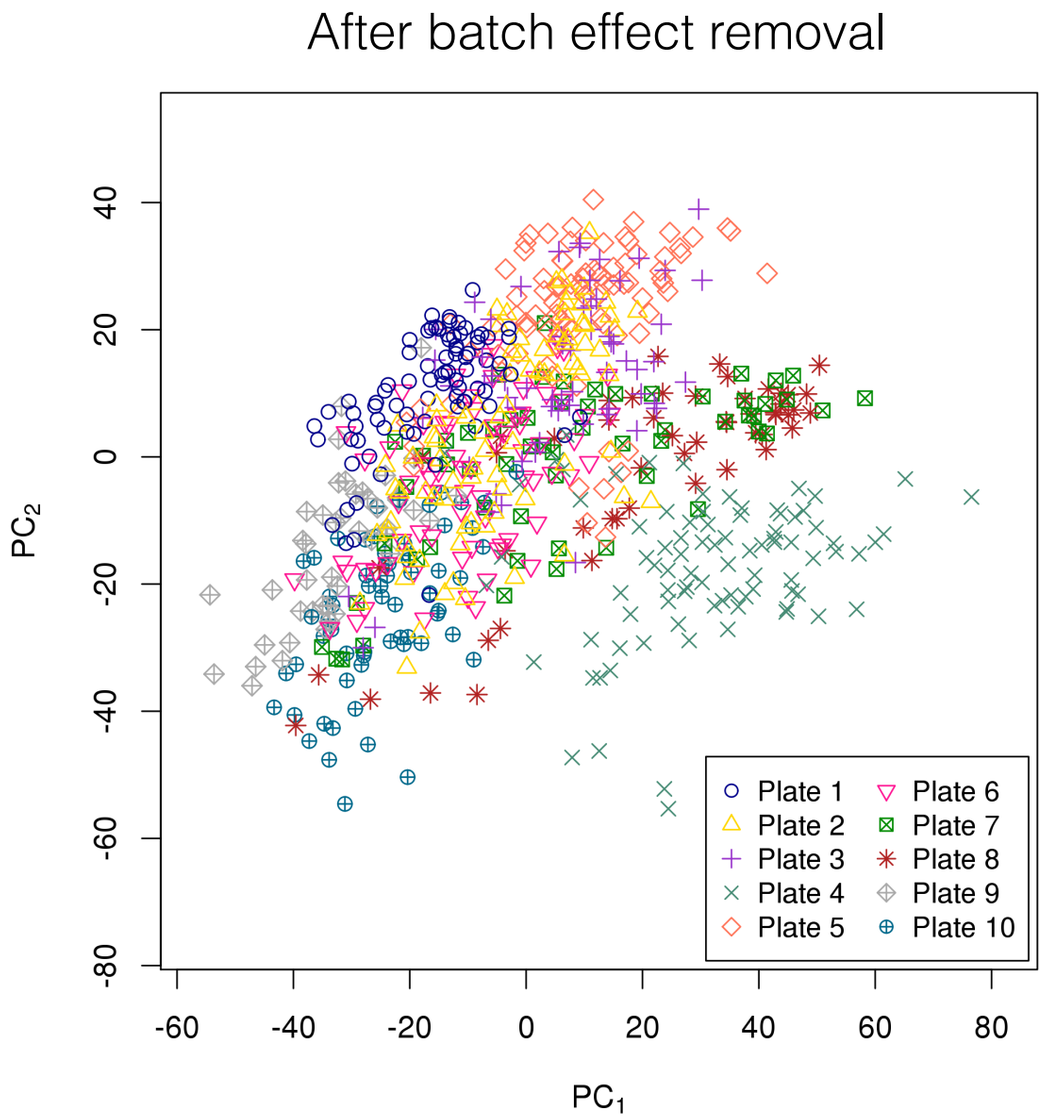
Batch Effects Occur

the batch effect represents the systematic technical differences when samples are processed and measured in different batches and which are unrelated to any biological variation recorded



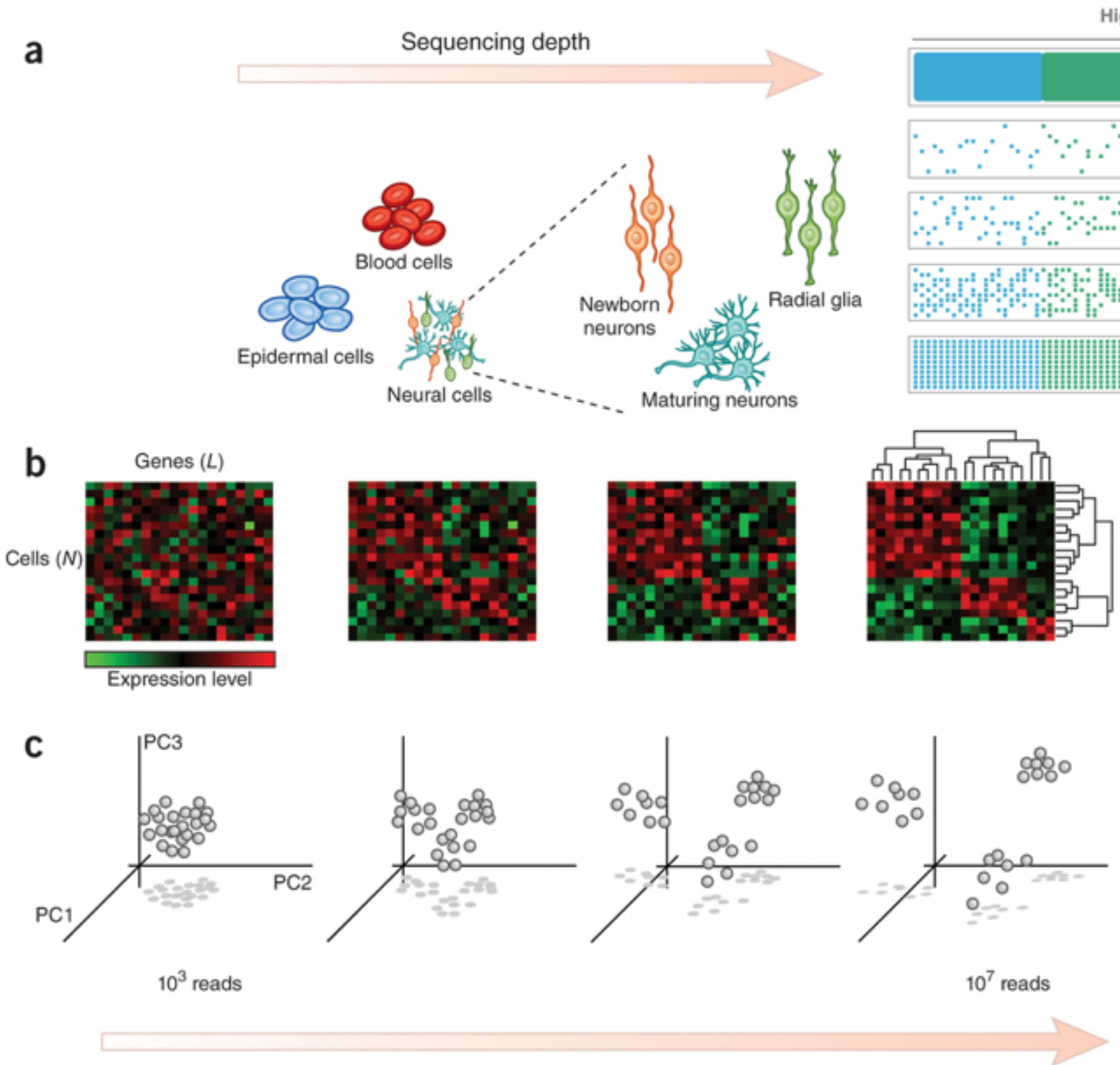
Batch Effects Occur





http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=4179&context=etd

Sequencing Depth



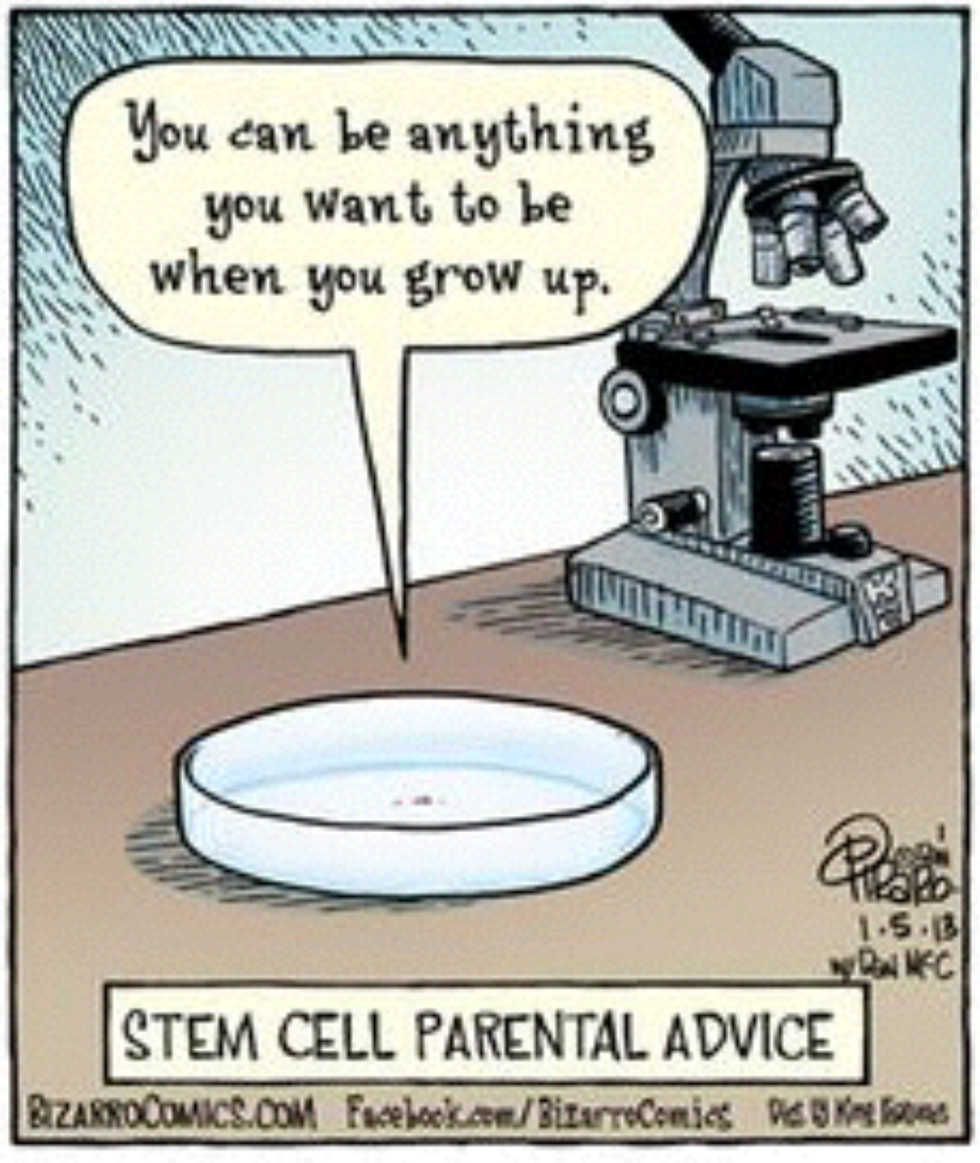
Pollen AA et al. (2014) **Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex.** Nat Biotechnol 32(10):1053-1058.

Streets AM, Huang Y. (2014) **How deep is enough in single-cell RNA-seq?** Nat Biotechnol 32(10): 1005-1006.

https://www.lab7.io/test/wp-content/uploads/2013/09/rna-sample-depth1.png

Biological Effects

- Cancer: cell lineage
- Metagenomics: cis/trans mechanisms
- Stem Cells: cellular phenotypes
- Immunology: cell type identification
- Neurology: somatic mutations

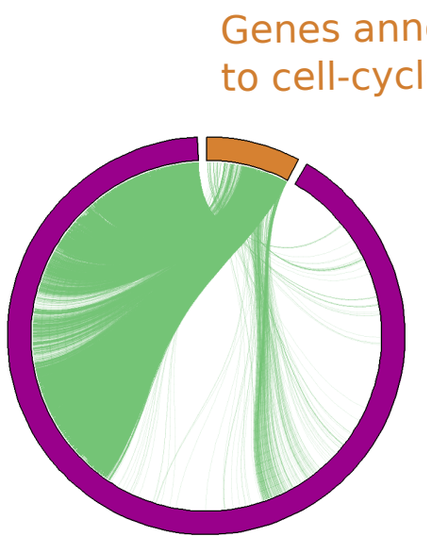


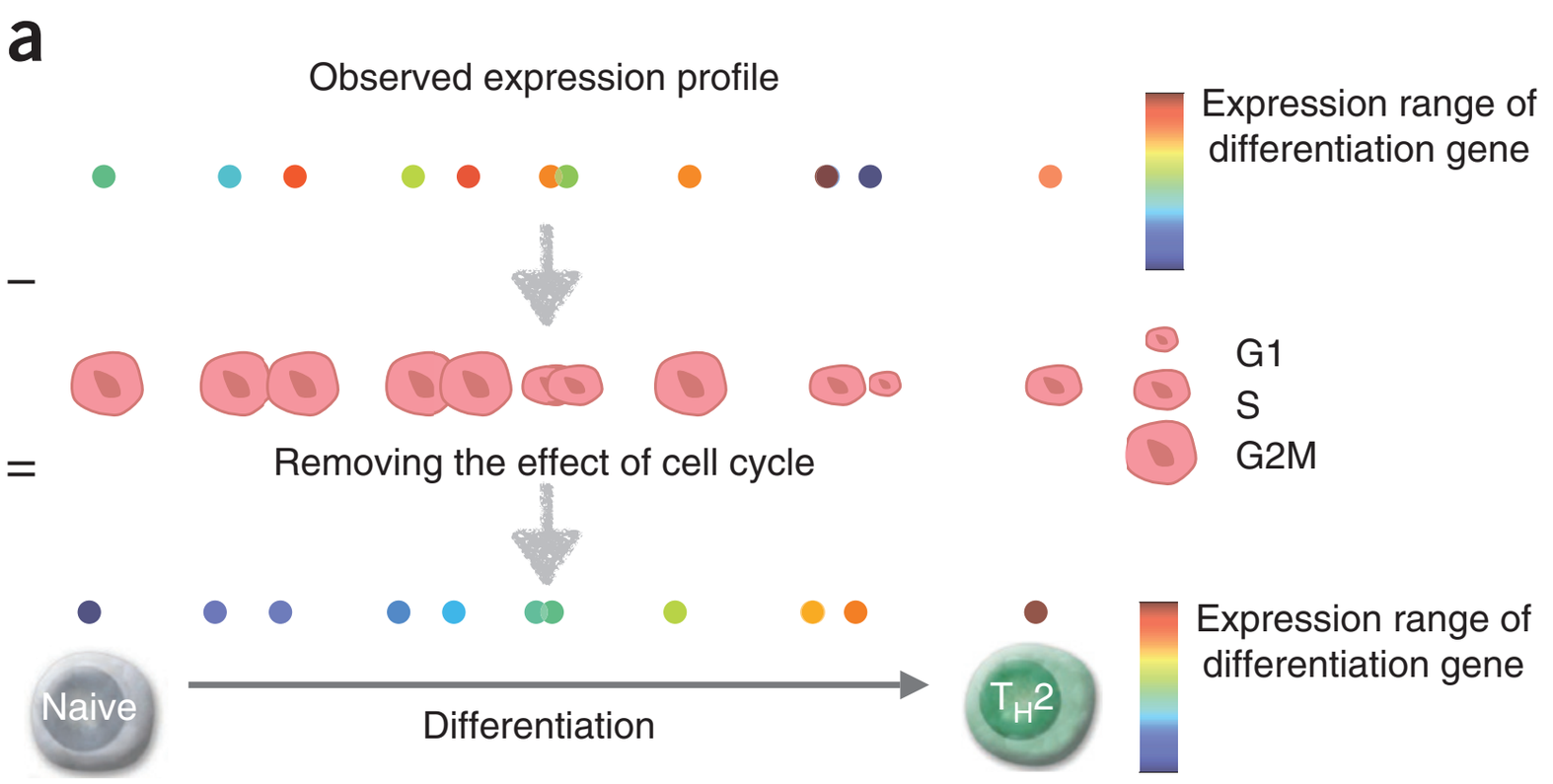
https://s-media-cache-ak0.pinimg.com/736x/c4/f2/d4/c4f2d4cd1b03f73b34ce38be28edca18.jpg

Example

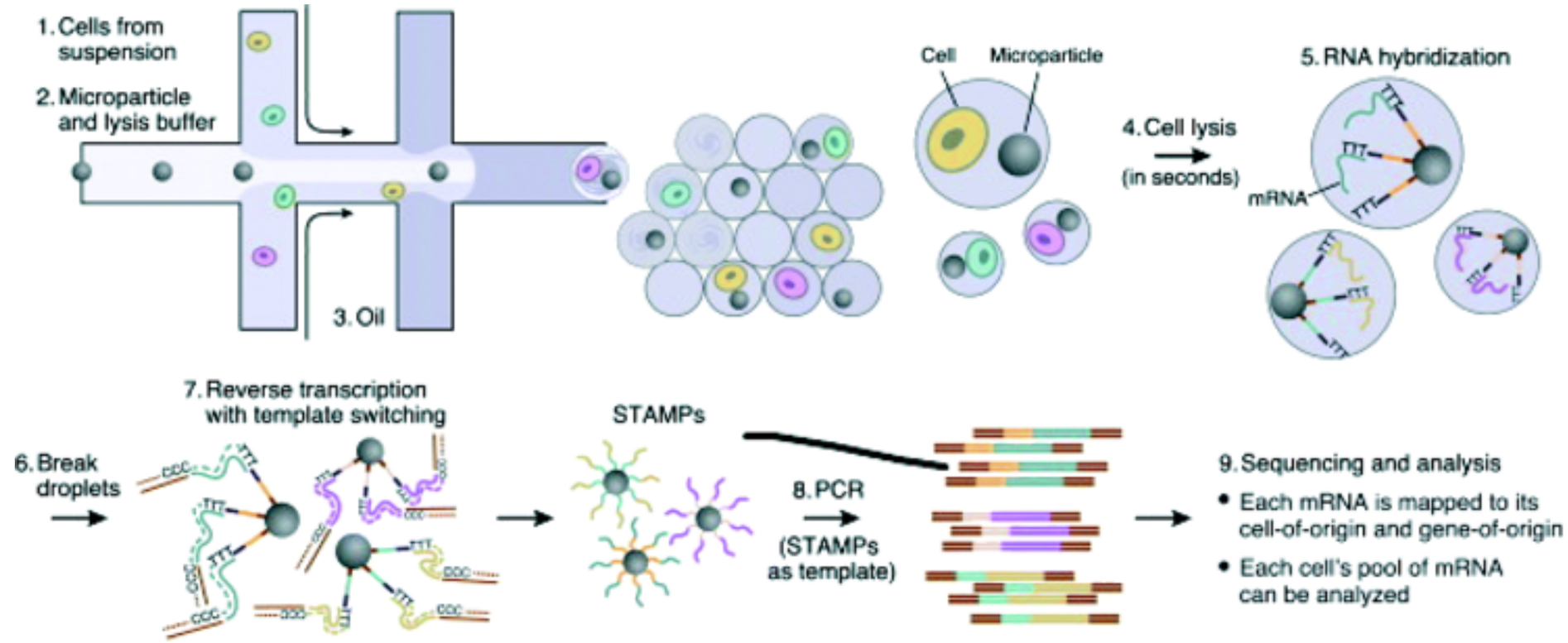
Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Florian Buettner[1,2,5], Kedar N Natarajan[2,3,5], F Paolo Casale[2], Valentina Proserpio[2,3], Antonio Scialdone[2,3], Fabian J Theis[1,4], Sarah A Teichmann[2,3], John C Marioni[2,3] & Oliver Stegle[2]



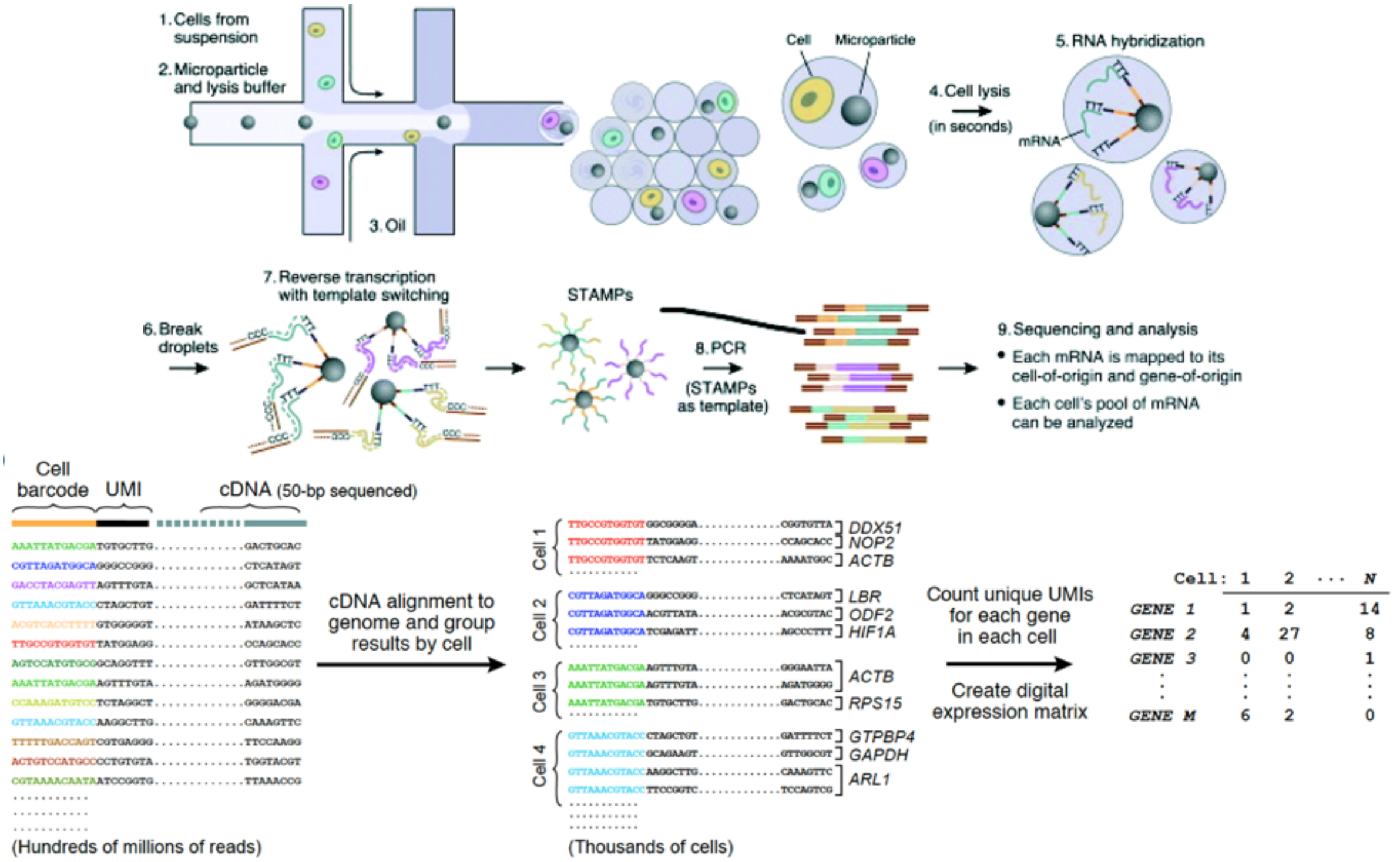


Full example: DropSeq



Macosko, Evan Z., et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161.5 (2015): 1202-1214.

Full example: DropSeq



Macosko, Evan Z., et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161.5 (2015): 1202-1214.