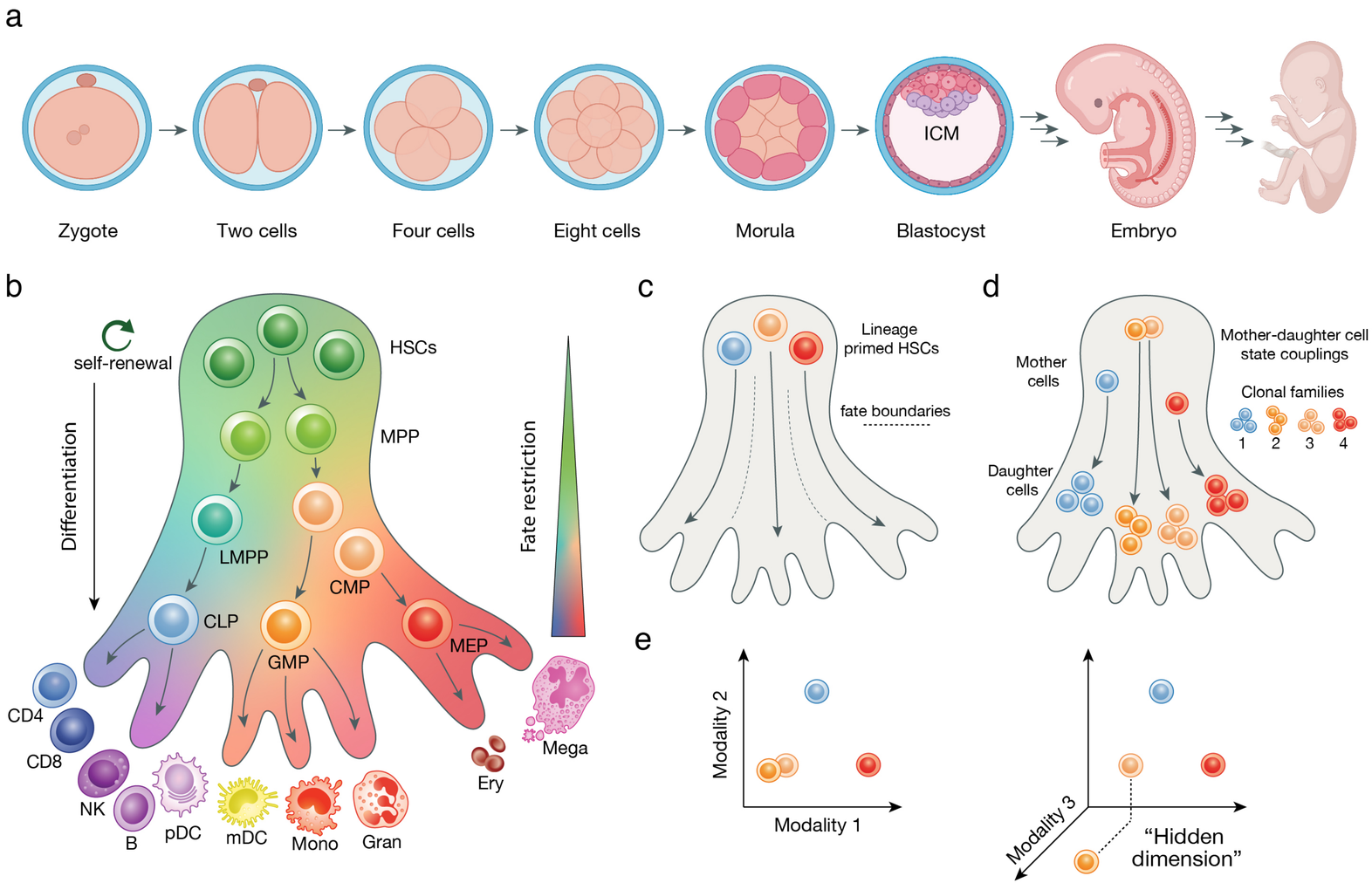
Systems Biology: Computational Analysis and Interpretation of High-Throughput Single-Cell Data

SESSION 1.2: INTRO

LALEH HAGHVERDI

Same DNA in all cells but different gene activities and cell types



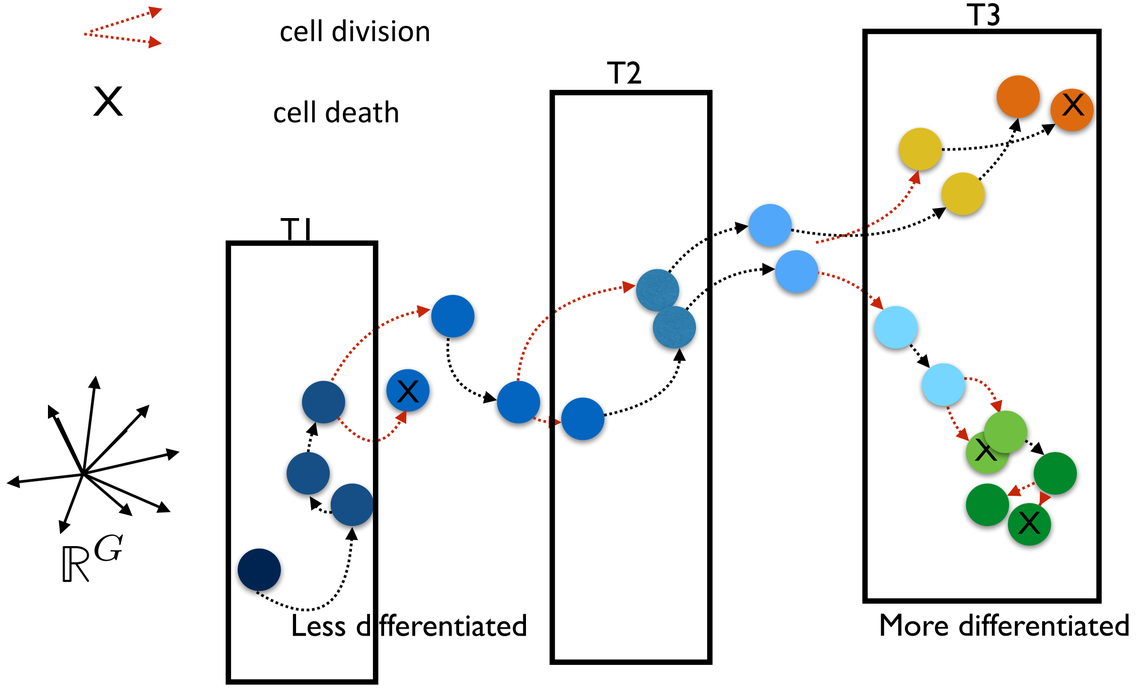
Understanding cell fate decision-making

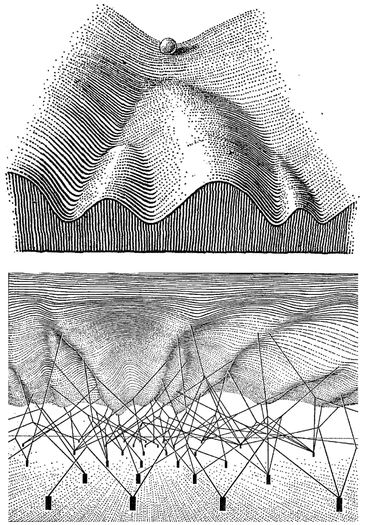
Stem Cell Reports
Perspective
ISSCR
OPEN ACCESS

Single-cell multi-omics and lineage tracing to dissect cell fate decision-making

Laleh Haghverdi[1,3,*] and Leif S. Ludwig[1,2,3,*]
[1]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany
[2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[3]These authors contributed equally
*Correspondence: laleh.haghverdi@mdc-berlin.de (L.H.), leif.ludwig@bih-charite.de (L.S.L.)
https://doi.org/10.1016/j.stemcr.2022.12.003

- Dynamical Inference
  - Pseudotime
  - Optimal Transport
  - Cell state velocities
  - Parallel methods as computational validation
- Mathematically derived relations in Supplemental Notes
  - Drift (needs energy)
  - Diffusion (stochastic)
  - Birth-Death
- Data Integration and latent spaces
- Regulatory Networks: perturbation —> control of cell fates





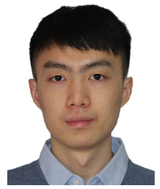
Haghverdi lab

MAX DELBRÜCK CENTER

- Diffusion maps for high-dimensional single-cell analysis of differentiation data, L. Haghverdi et al. Bioinformatics (2015)
- Diffusion pseudotime robustly reconstructs lineage branching, L. Haghverdi et al. Nature methods (2016)
- *destiny*: diffusion maps for large-scale single-cell data in R, P. Angerer et al. (2016)
- Towards reliable quantification of cell state velocities, V. Marot-Lassauzaie et al. PLoS Computational Biology (2022)
- Single-cell time series analysis reveals the dynamics of in vivo HSPC responses to inflammation, B. J. Bouman et al. Life Science Alliance (2023)
- Single-cell multi-omics and lineage tracing to dissect cell fate decision-making, L. Haghverdi & L. S. Ludwig, Stem Cell Reports (2023)




Dynamical Inference


Regulatory Networks

- Decoding the regulatory network of early blood development from single-cell gene expression measurements, V. Moignard et al. Nature biotechnology (2015)
- Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data, A. Ocone et al. Bioinformatics (2015)
- Seyed Amir Malekpour, et al, Single-cell multi-omics analysis identifies context-specific gene regulatory gates and mechanisms, Briefings in Bioinformatics (2024)

Data integration & latent spaces




- Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, L. Haghverdi et al. Nature biotechnology (2018)
- Adjustments to the reference dataset design improve cell type label transfer, C. Moelbert & L. Haghverdi Frontiers in Bioinformatics (2023)
- Colin G Cess, Laleh Haghverdi, Compound-SNE: comparative alignment of t-SNEs for multiple single-cell omics data visualization, Bioinformatics (2024)
- Valérie Marot-Lassauzaie, Sergi Beneyto-Calabuig, Benedikt Obermayer, Lars Velten, Dieter Beule, Laleh Haghverdi, Identifying cancer cells from calling single-nucleotide variants in scRNA-seq data, Bioinformatics(2024)

HELMHOLTZ


DISCOVERY
FOR
TOMORROW'S
MEDICINE
MDC
BIMSB
MAX-DELBRÜCK-CENTRUM
FÜR MOLEKULARE MEDIZIN

How does this relate to Physics?

- Methods from physics, e.g. Fokker-Plank equation, optimization, control theory, etc.
- Physics as quantification science
- Relating observation to theory so that we can do something useful with it
- Outside academia (for doing useful things ;)) you always need to work interdisciplinary
- Need to be able to communicate with other experts (biologists, programmers, mathematicians, etc)
- You need to understand about what you are talking about, e.g. the interpretation of every bit in an equation, its relation to the experiment, its relation with the computer code..

- New physics is also biology, economics, AI; our everyday life and needs

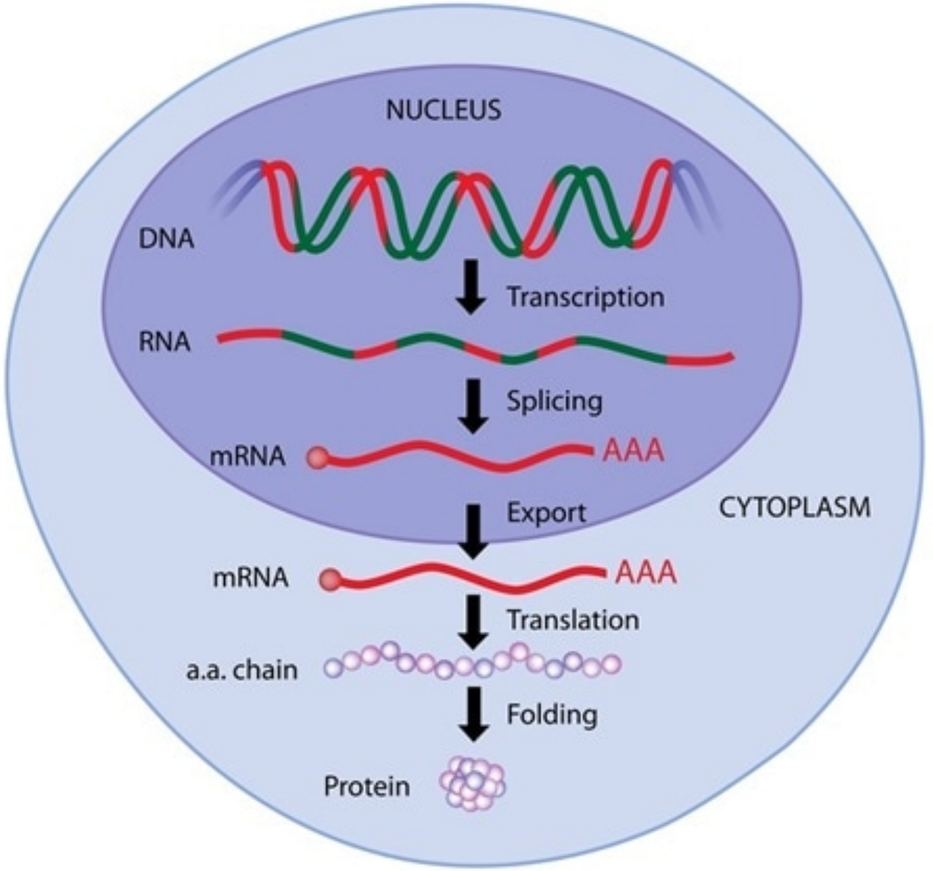
- Physicists right and wrong feeling that they can involve in every matter :->, quite often they overlook observation and are biased towards existing theory

Topics

- Intro, single-cell NGS measurements and data quantification
- Preprocessing
- Dimension reduction
- Dynamical inference (Langevine, Fokker-planck, OT, etc.)
- Gene regulatory networks
- Learning theory (optimisation, identifiability, etc.)
- Maximum Likelihood and Expectation Maximisation algorithm
- Variational methods (PPCA, VAE, etc.) and generative models

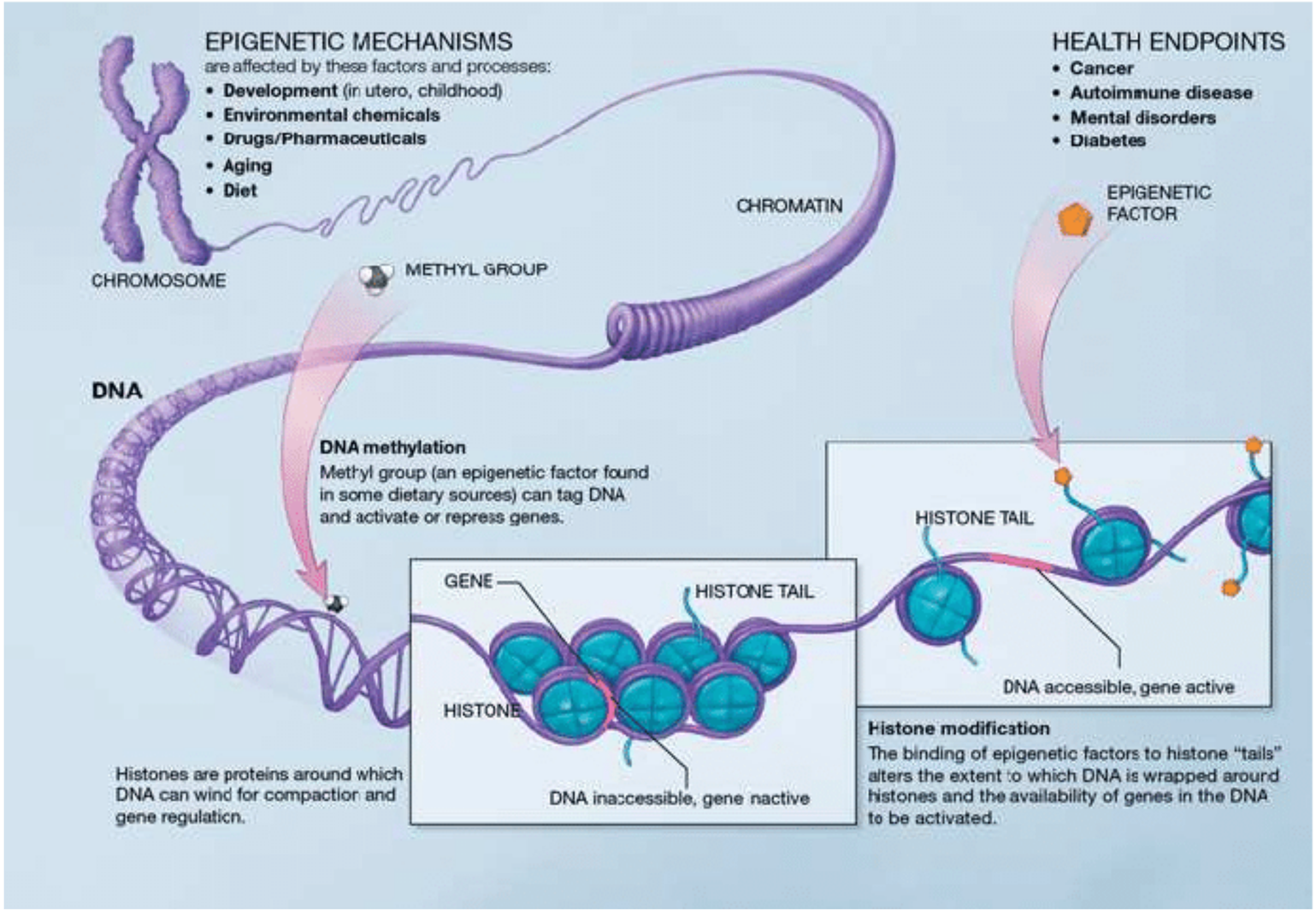
DNA —> nascent mRNA —> spliced mRNA —> protein

- A defect in any step may lead to a disease e.g.
  - Mutations in the gene body —> broken protein
  - Mutations in other regulatory regions —> too little or too much production of an mRNA or protein
- Human and mouse ~20k genes
- Only ~10% of the DNA is genes



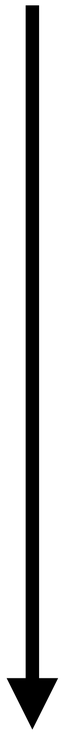
Genome organisation: epigenetic

- Some diseases arise because of incorrect chromatin unfolding and folding



Omic data types

Information maturation

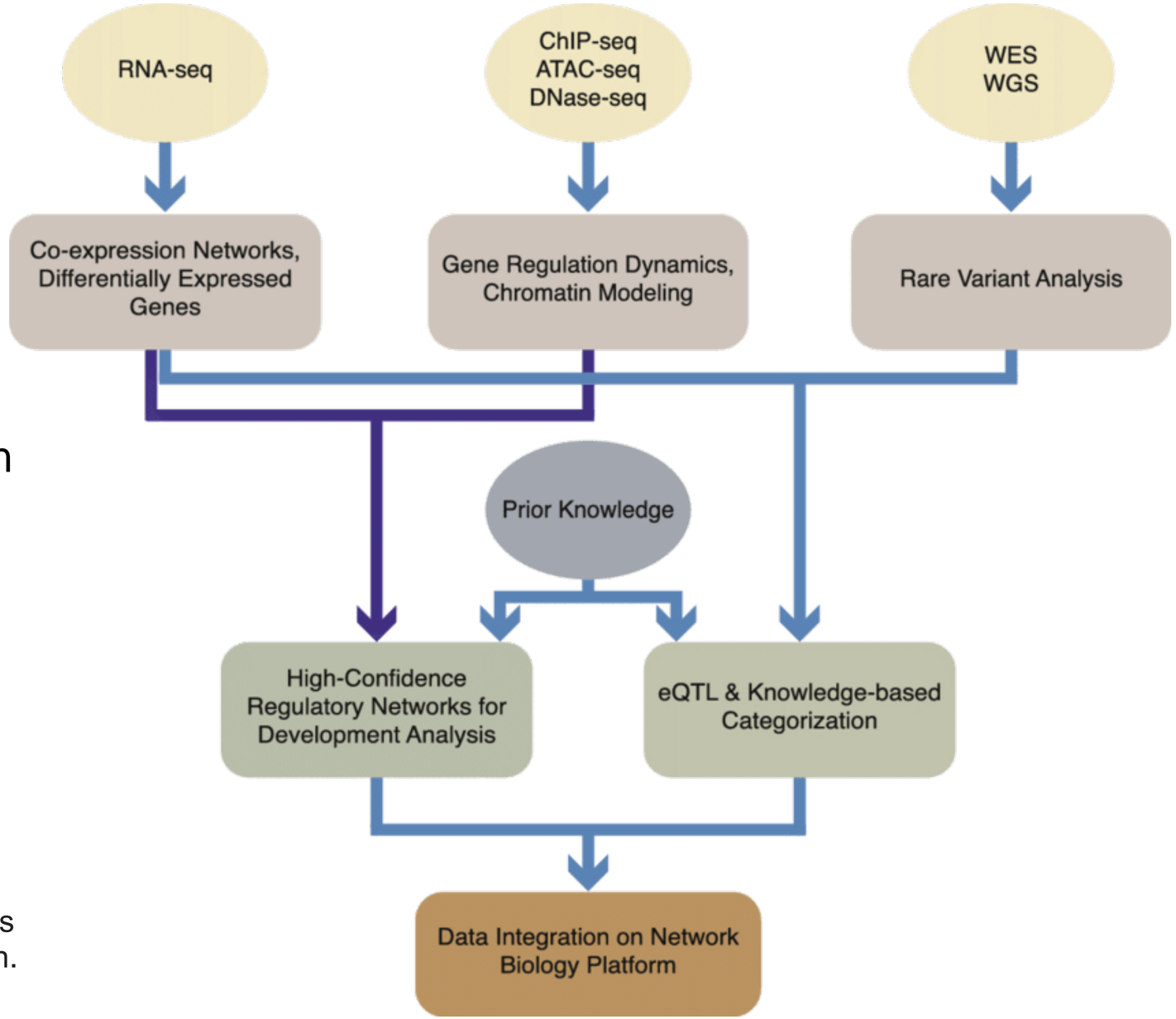


- Genomics
- Epigenomics (Methylation, Chromatin accessibility, etc)
- Transcriptomics
- Proteomics

A lot of inter- and intra-layer interactions

Next-Generation Sequencing

- Any sequence of nucleic acids (from DNA or RNA) can be sequenced



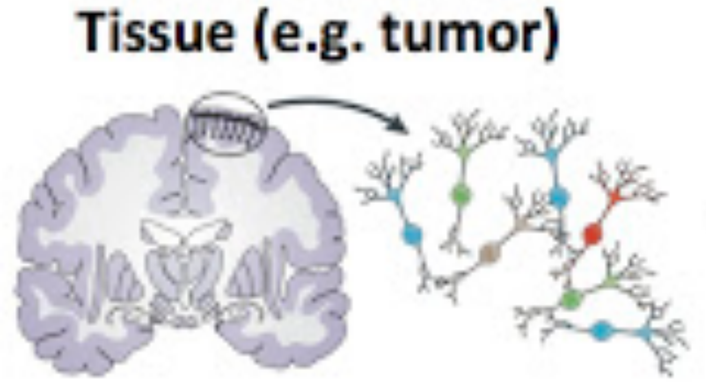
Chaitankar V, et al. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Progress in retinal and eye research. 2016 Nov 1;55:1-31.

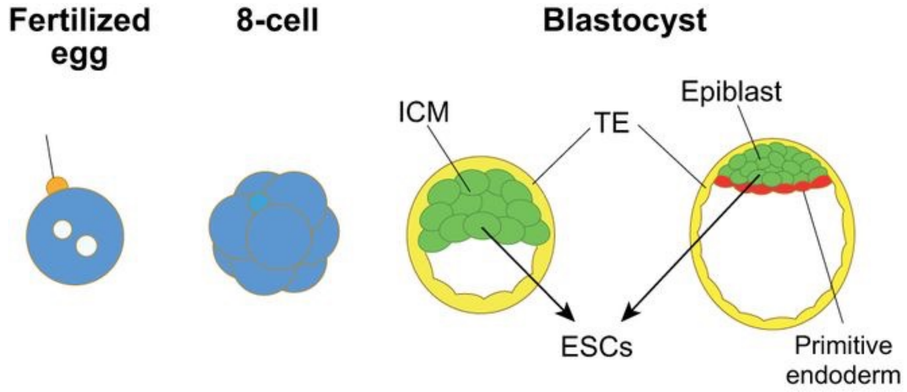
Bulk (~1990) vs. single-cell (~2010) omics

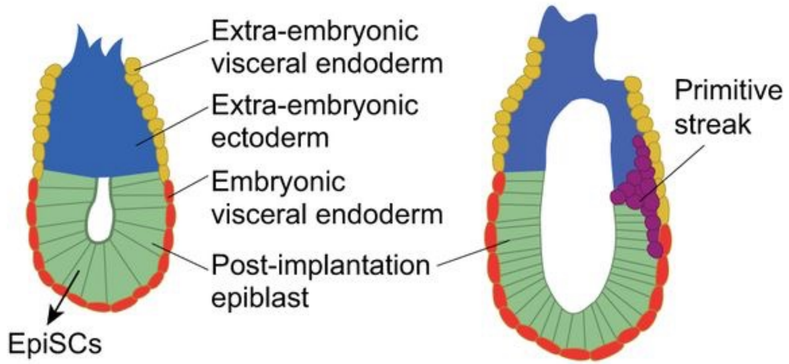
- Single cell isolation is a recent technology
- Eukaryote cell diameter ~10-100 nano-meter
- Low number of molecules per cell —> amplification

Single-cell measurements

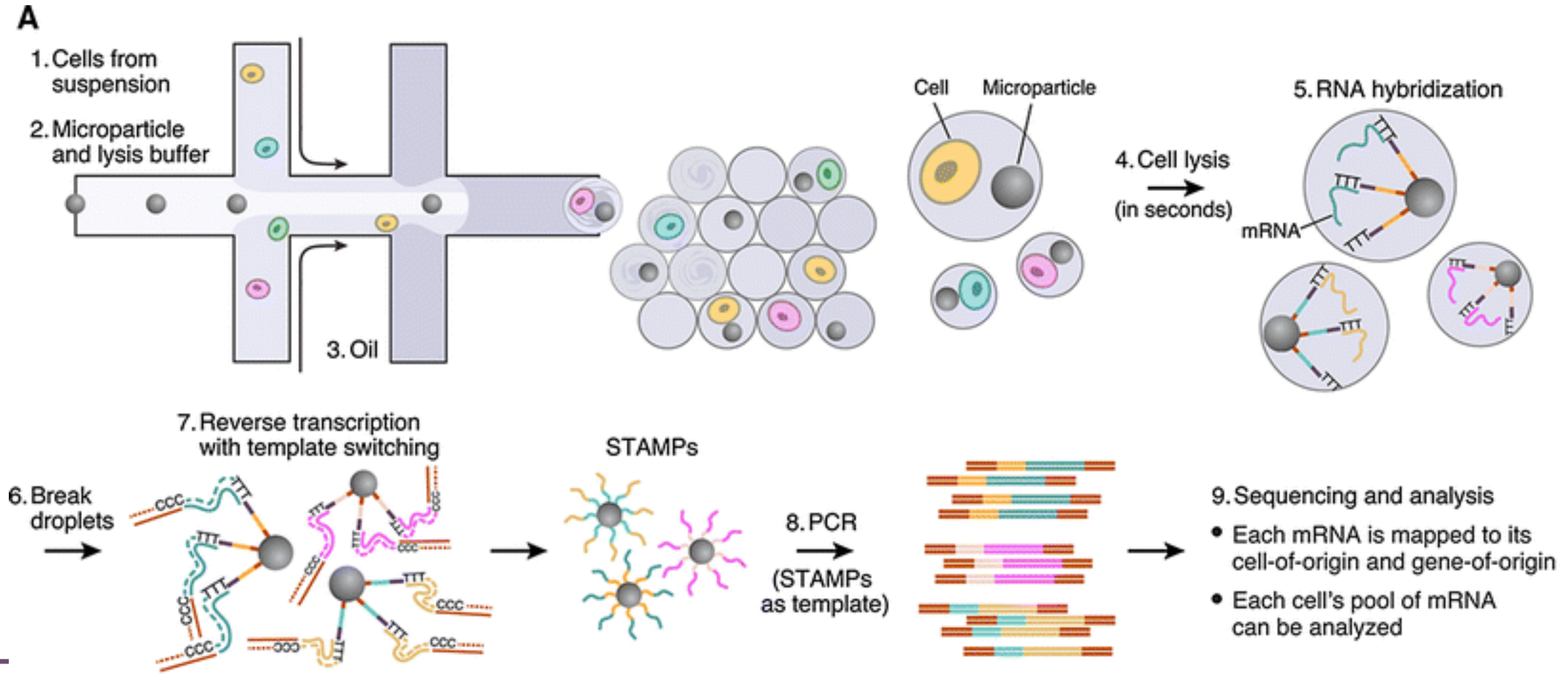
- Heterogenous cell populations:
  - Cancer and tumour cell populations
  - Rare cell types
  - Several seemingly homogenous cell populations
  - Transitory and asynchronous states e.g. in development

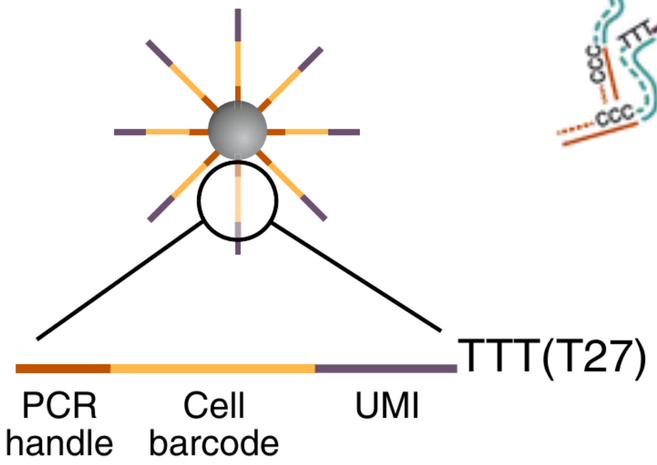






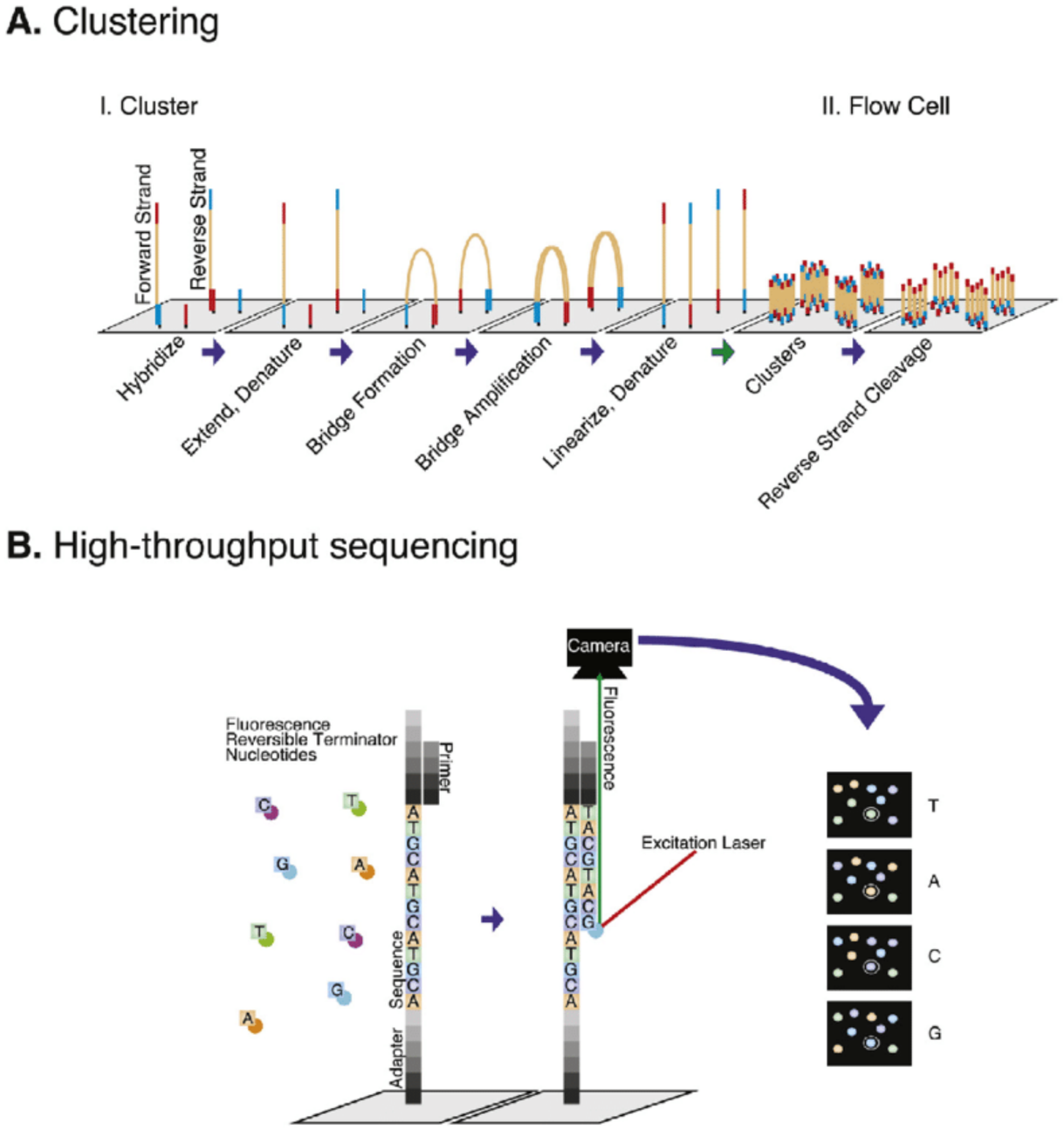
Single-cell omics



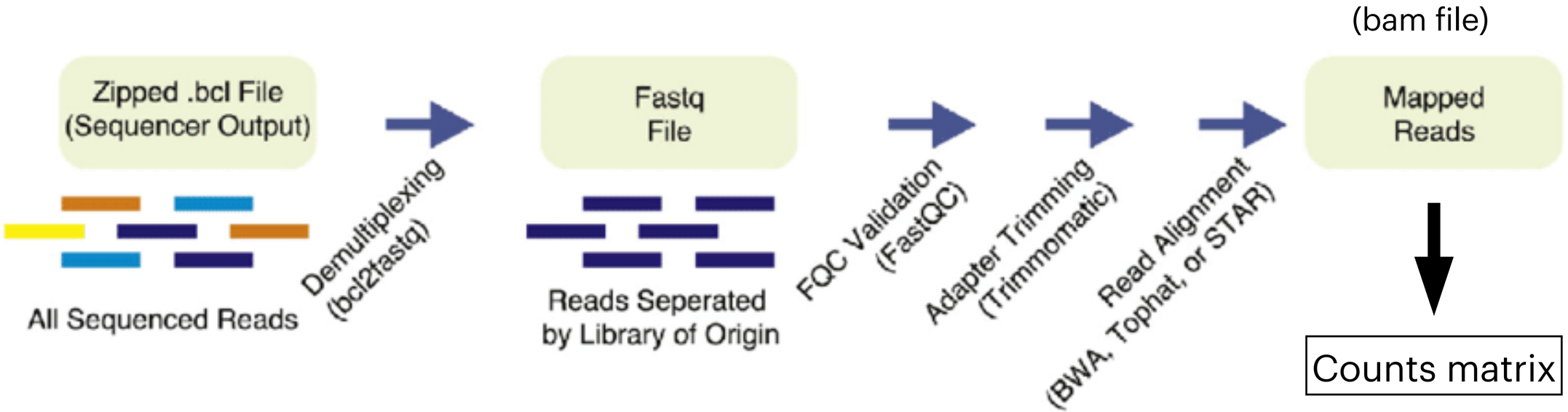


Choi JR, Yong KW, Choi JY, Cowie AC. Single-cell RNA sequencing and its combination with protein and DNA analyses. Cells. 2020 May 4;9(5):1130.

10x Illumina technology

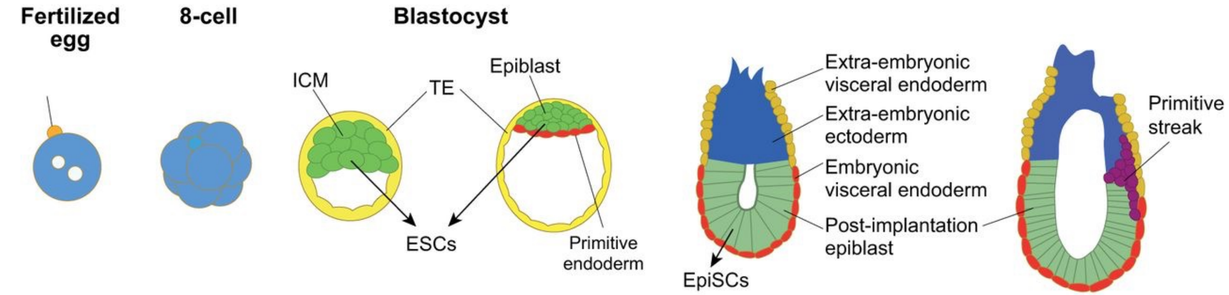


Sequence reads data processing

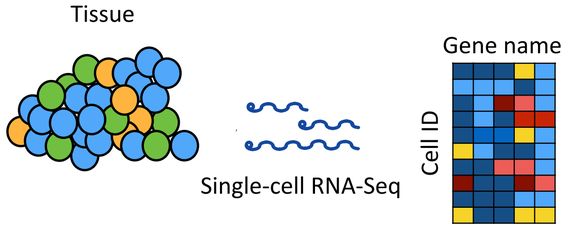


- Tools such as Cell Ranger:
  - End-to-end pipeline (from raw reads to expression matrices, clustering, and visualization).
  - Barcode demultiplexing and UMI correction tailored for 10x Genomics.
  - Pre-built reference datasets and multi-modal data processing (e.g., CITE-seq, ATAC-seq).
  - Integrated quality control metrics and automatic filtering of empty droplets and doublets.
  - Built-in visualizations (UMAP, t-SNE) and differential expression analysis.
  - Optimized for and supported by 10x Genomics for seamless integration with their hardware.

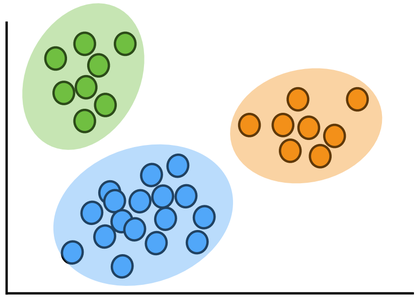
Single-cell counts data processing

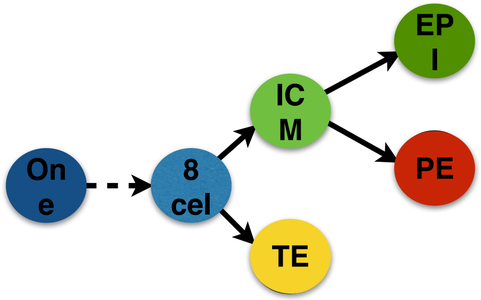


(i) Preprocessing of expression profiles

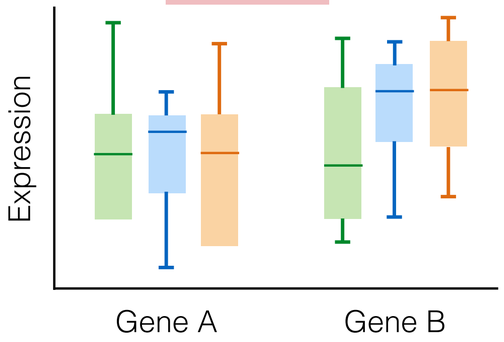


(ii) Clustering and visualisation



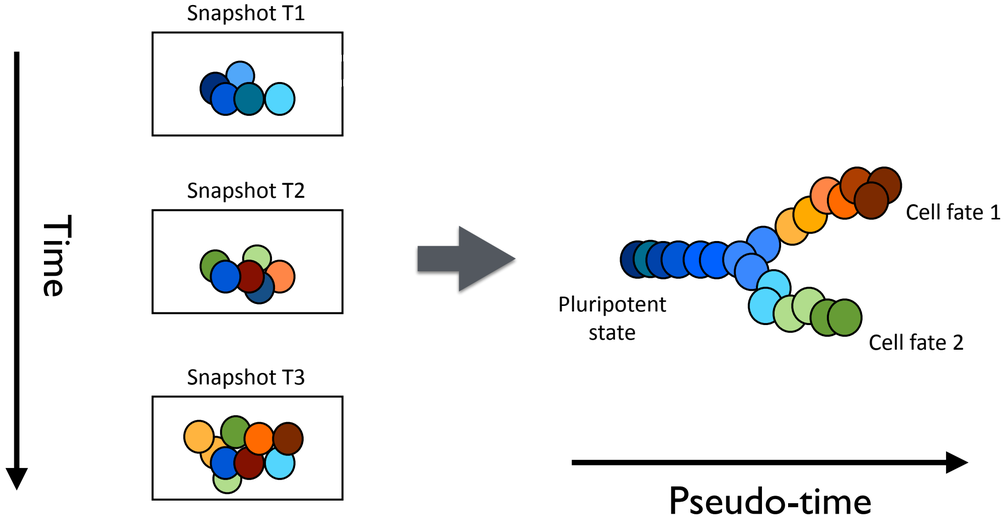


(iii) Differential comparison



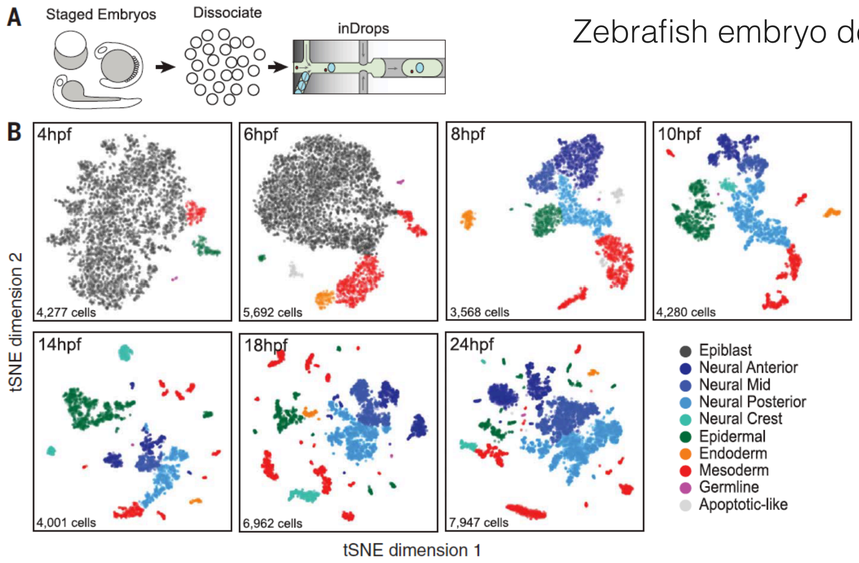
(iv) Integration with biological knowledge (phenotype)

- **T cell**
- **B cell**
- **Unknown**

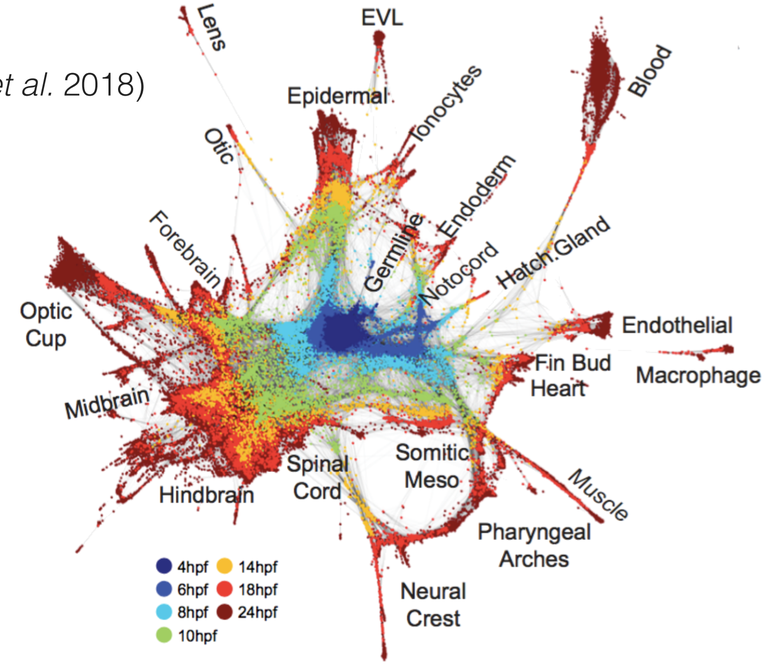


Typical scRNA-seq data sizes and properties

- ~10^6 cells * ~20k genes
- Sparse and noisy data
- The volume of 2-state genes space 2^20k >> number of sampled cells
- Even with non-zero genes only in a dataset 2^5k >> !0^6 cells
- The higher the number of dimensions, the more severe artefacts of curse of dimensionality



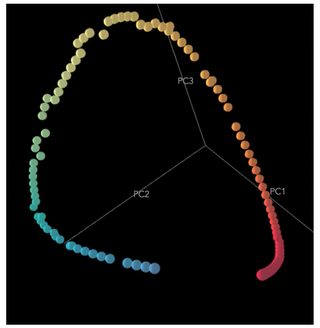
Zebrafish embryo development (Wagner *et al.* 2018)



Curse of dimensionality

- Expansion of sampling space and vagueness of distance measures (curse of dimensionality), some algorithms e.g. minimum spanning trees (MST) won't work properly

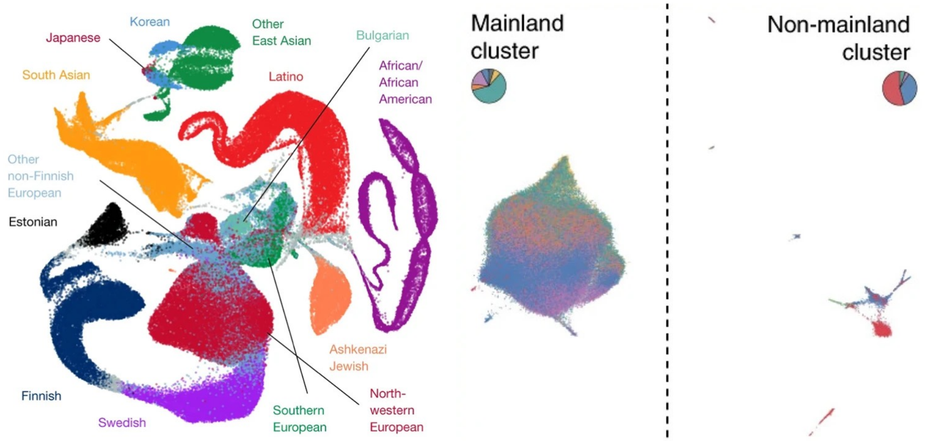
$$\lim_{d\to\infty} E\left(\frac{\mathrm{dist}_{\max}(d) - \mathrm{dist}_{\min}(d)}{\mathrm{dist}_{\min}(d)}\right) \to 0.$$



- Horse-shoe effect in PCA of random points D >>> N
- Weird string shapes in UMAP and t-SNE, overfitting
- especially in population genetics, no. variants >>> no. patients

Diaz-Papkovich, et al. A review of UMAP in population genetics. *J Hum Genet* (2021)

The Genome Aggregation Database (gnomAD, left) and Biobank Japan (BBJ, right) visualized using UMAP.



(a) gnomADv3 data visualized using UMAP.

(b) BBJ data visualized using UMAP.

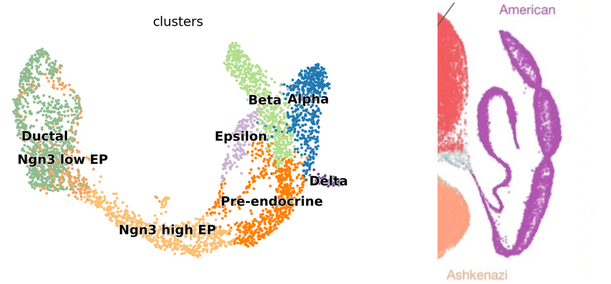
~36 million singleton variants from 3560 whole-genome sequences
Carlson, J. *et al. Nat Commun* (2018)

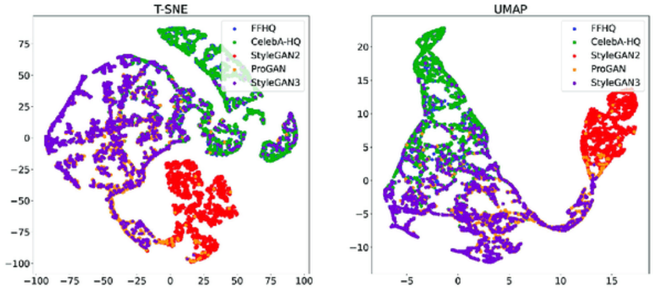
Hollow volumes in presence of noise (randomness)

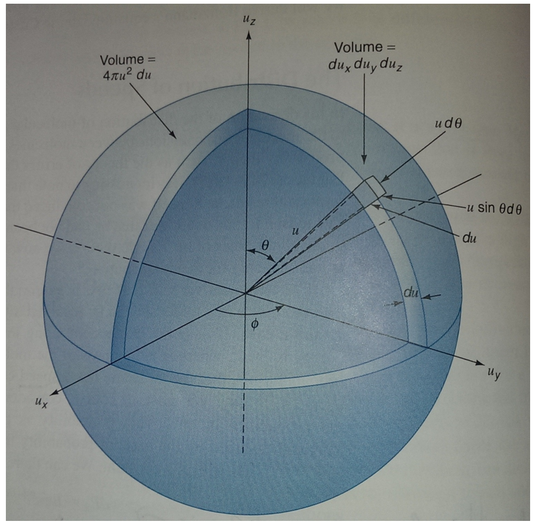
- With too many features noise dominates the signal
- Hollow spheres (and volumes in case of some non-random dimensions)

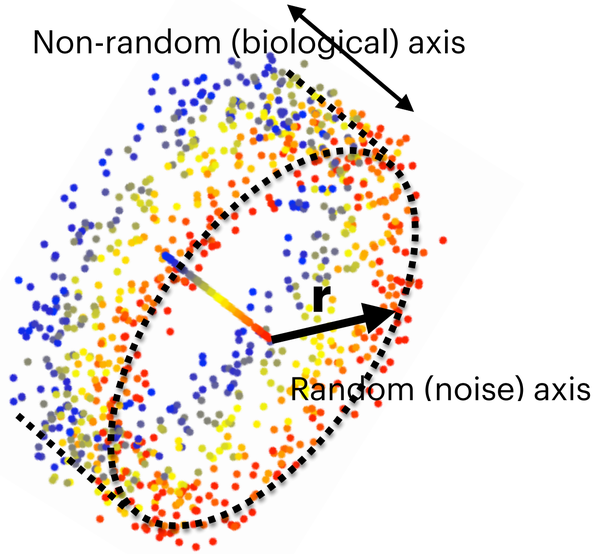
$$\text{number of data points} \propto \text{volume of the shell} \propto 2\pi r^{D-1}$$

- Bad similarity measure (due to curse of dimensionality, bad similarity kernel, etc.)







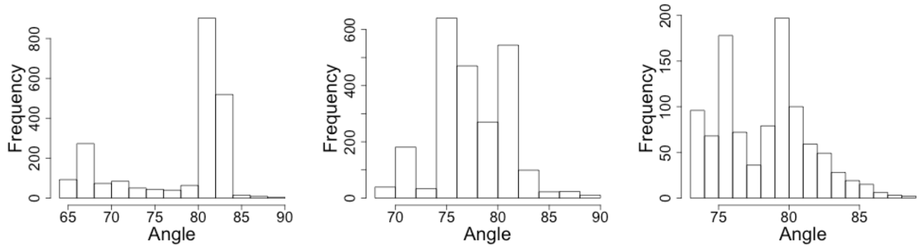


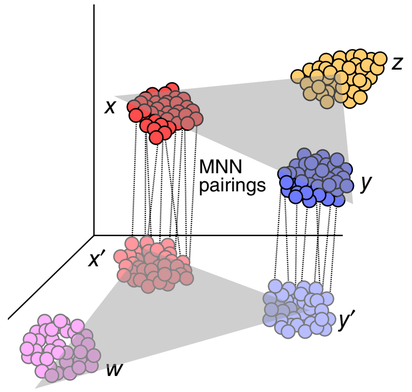
We can make use of these “weird” properties, they are not always bad!

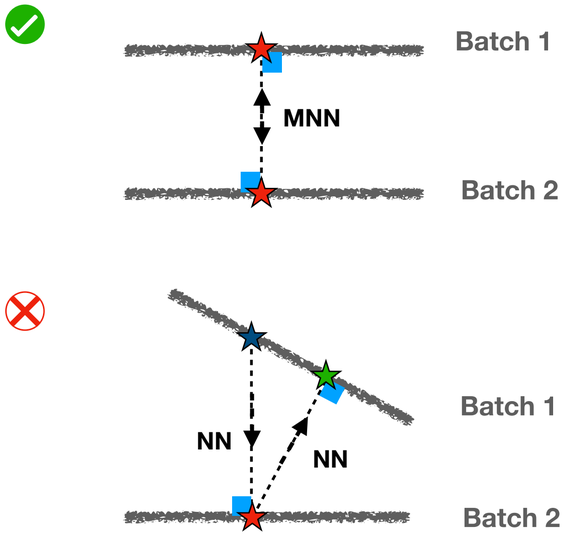
- Correlated biological hyperplanes
- Batch effect vectors are (almost) orthogonal to the biological hyper-plane in high dimensions (central limit theorem)
- In high-D there exists no correlation by chance (e.g. between batch effects), i.e., random vectors are just orthogonal!

Central limit theorem on summation of n (many) random values

$$\vec{X} \cdot \vec{Y} = \sum_{i=1}^{D} x_i y_i \;\; \text{—> 0} \quad \text{,for large D}$$







Data integration by matching Mutual Nearest Neighbours
Haghverdi,et al., Nature biotechnology 2018

3. Feature selection and scRNA-seq noise models

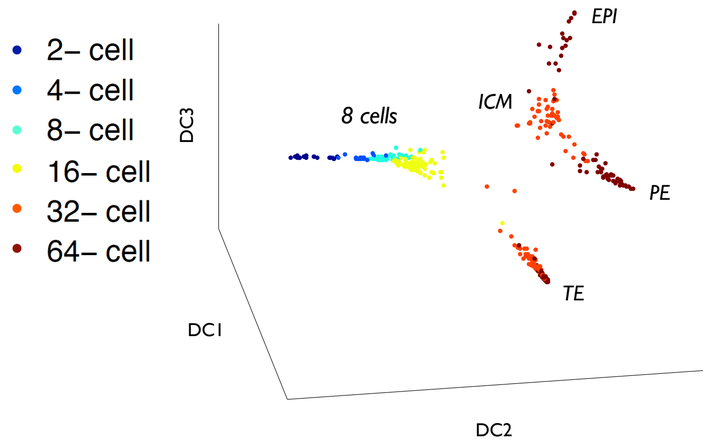
Feature selection

- Avoid high-D weird effects
- So a major goal of Highly Variable Genes selection is just feature reduction
- Even randomly selecting fewer features helps
- But also keeping more interesting features, e.g. for better separation of specific clusters (cell type? Or patients? Or …) or for capturing trajectories
    - DELVE: feature selection for preserving biological trajectories in single-cell data. Ranek JS, Stallaert W, Milner JJ, Redick M, Wolff SC, Beltran AS, Stanley N, Purvis JE. Nature Communications. 2024
- For any high-d data, different features indicate different properties, not only in biology!
- Example of good selection of features:

Haghverdi, et al. Bioinformatics (2015)

sc-qPCR ~500*48 mouse embryonic stem cells data set

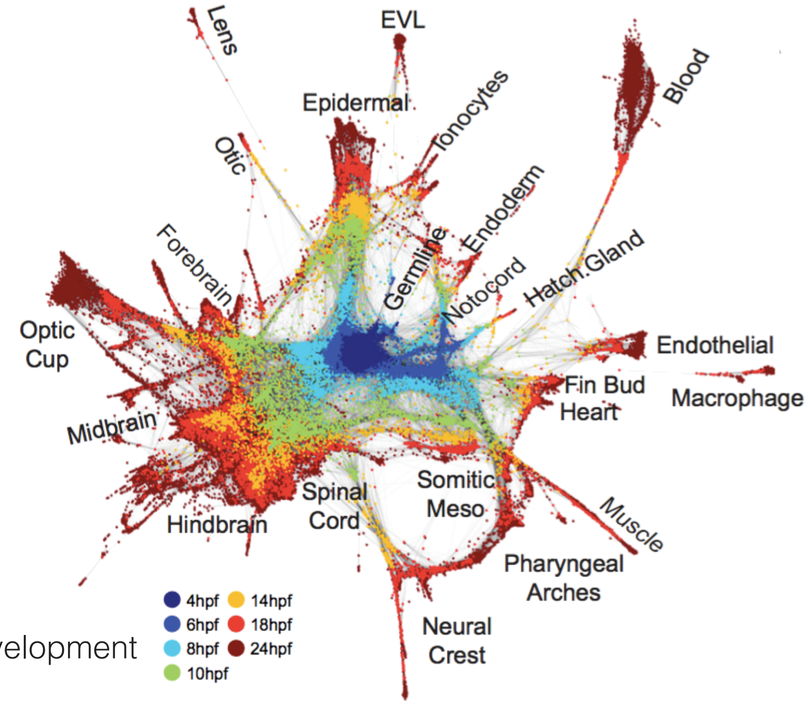
Guo *et al. Developmental cell 2010*



HVG selection for reducing the number of features

- ~10^6 cells * ~20k genes
- So a major goal of Highly Variable Genes selection is feature reduction
- Even randomly selecting “HVG”s helps! That’s why even bad hvg selection methods work

- Standard analysis pipelines~ 2000 HVGs; who said this is a good number?!
  - N prop. 2*pi* r^D —> log(N) prop. D
- So with fewer cells pick also fewer HVGs



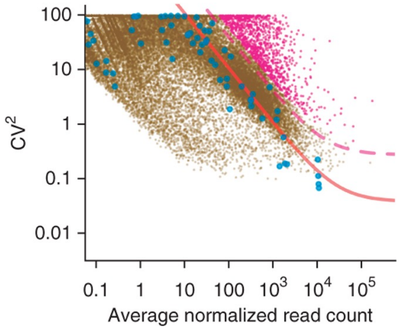
Zebrafish embryo development (Wagner *et al.* 2018)

HVG methods based on specific noise models

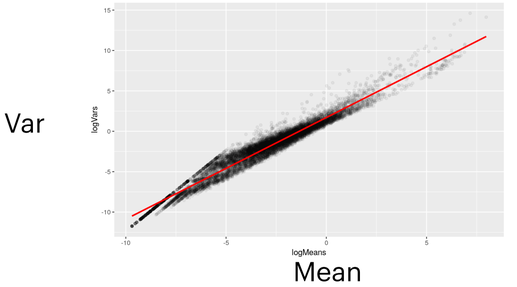
- Mean-variance interdependency is difficult to get rid of. This challenges some analysis (e.g. noise quantification, HVG selection etc.)

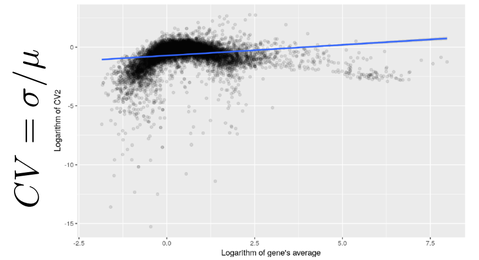
- Brennecke's method doesn't resolve this problem! (Still selects high mean genes, also some theoretical discrepancies, e.g. why CV2?! It only reverses the trend)

- NB or Poisson fit only raw counts data, WITHOUT normalised and log transform!



Brennecke et al. Nature Methods 2013





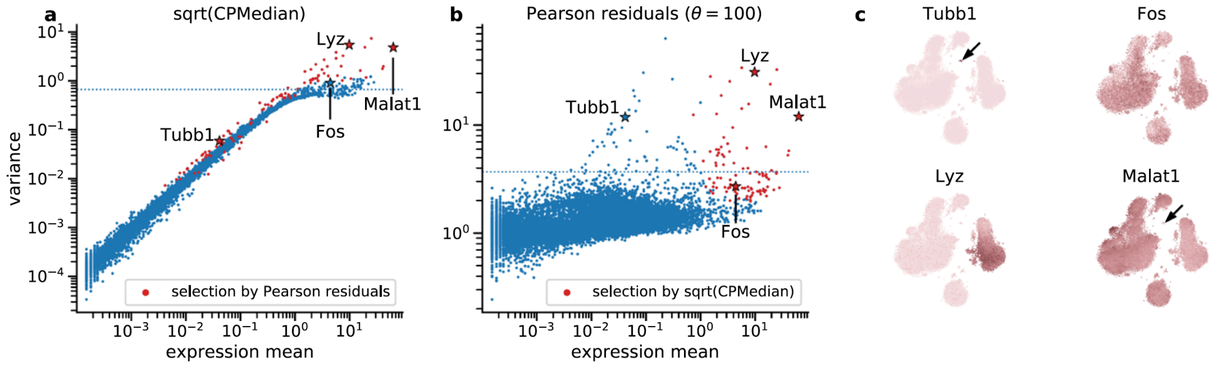
- Pearson residuals looks better (i.e. more mean independent)

- On raw counts!

  - For cell c and gene g, define Pearson residual as:

$$Z_{cg} = \frac{X_{cg} - \hat{\mu}_{cg}}{\sqrt{\hat{\mu}_{cg} + \hat{\mu}_{cg}^2/\theta}}, \quad \hat{\mu}_{cg} = \frac{\sum_j X_{cj} \cdot \sum_i X_{ig}}{\sum_{i,j} X_{ij}}, \quad \theta = 100.$$

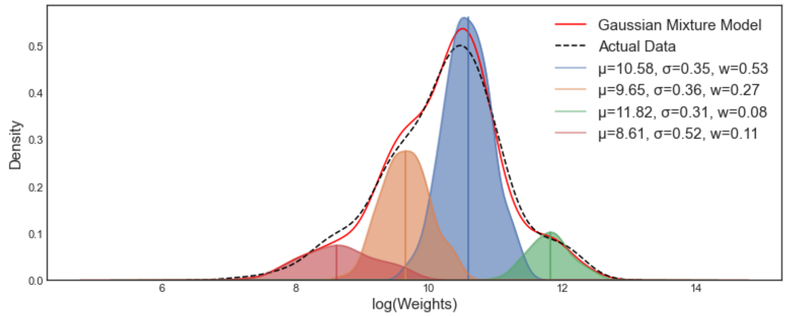
  - Plot Variance of Z over cells



Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome biology. 2021

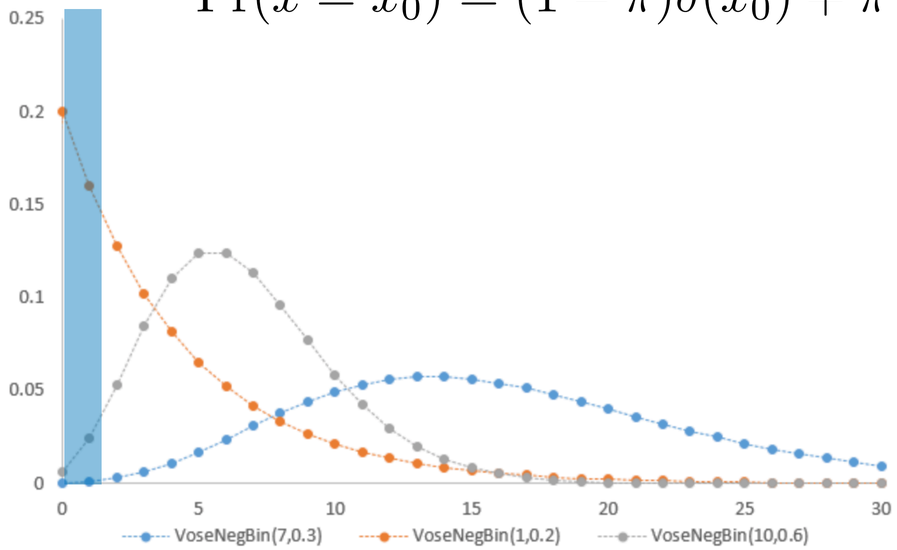
Zero-inflated NB (ZINB) model for heterogeneous cell populations

- Mechanistic Poisson and Gamma-Poisson (~NB) process (gene expression x sampling)
- Gaussian mixture —> multi-modal
- Poisson mixture —> NB $\sigma^2 = \mu$
- NB mixture —> NB $\sigma^2 = \mu + \frac{\mu^2}{\phi}$



- How about including cell populations which completely do not express that gene? —> Zero inflated model is necessary:

$$\Pr(x = x_0) = (1 - \pi)\delta(x_0) + \pi \ \mathrm{NB}(x_0|\mu, \phi)$$



- Some misleading debates are around! E.g., Svensson V. Droplet scRNA-seq is not zero-inflated. Nature Biotechnology. 2020

Poisson  **Variance=Mean**

NB  **Variance = Mean+ a Mean^2**

Counts matrix processing steps

Standard pipelines (e.g. Seurat, scanpy)

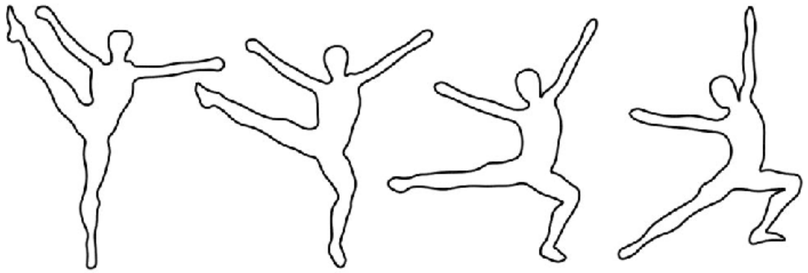
~Data transformation

1. (Library) size factor normalisation
2. Feature selection,
    - Uninteresting genes (mitochondrial, ribosomal, etc) filtering + cells filtering (doublets etc)
    - HVG selection
    - Other, depending on the question e.g. about transitions or separate clusters?
3. Log transform (UMI vs. non-UMI)
4. Metric (Euclidean, Cosine distance, etc) choice
5. Dimension Reduction and visualisation
6. Data integration, (Unsupervised) Clustering and/or Cell type label transfer

✳ Order of steps (1-3)?!

4. Data transformation and metrics

Several data transformations can keep your “interesting” data properties



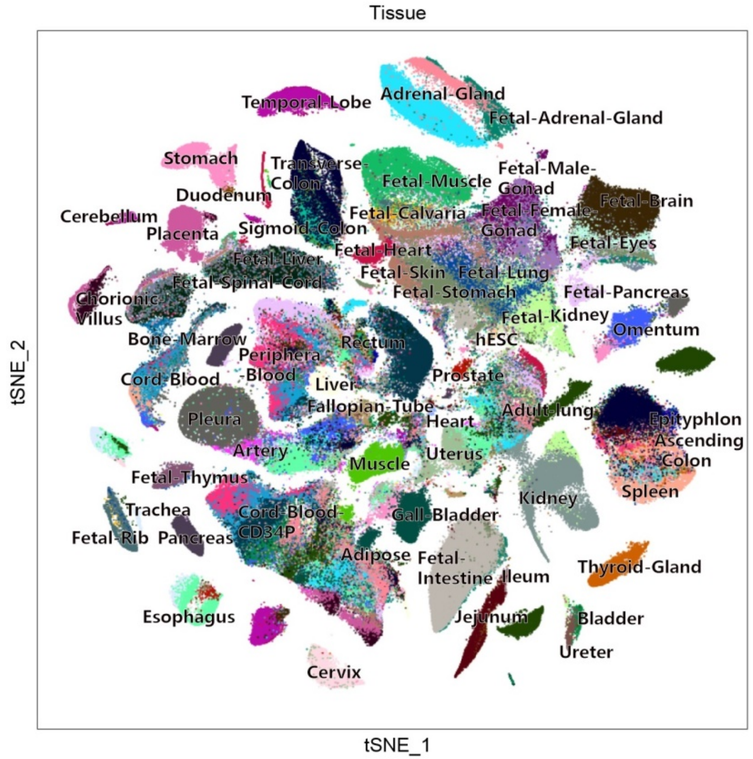
Why data transformation?

- Remove uninteresting features and effects
- Make different datasets comparable, e.g. across different batches
- Stabilisation and easier data handling (e.g. knowing its range, distribution)
- Examples:
  - Log or sqrt transform: make data distribution more Gaussian like
  - Standardisation: mean <— 0 , var <— 1
  - Size factor normalisation: removes different cell sizes and capture efficiency

  - Human Cell Atlas Consortium for standardising single-cell processing (like an SI system)

HCA consortium for standardising data collection and processing

- An SI system for omics data



Impact of the Human Cell Atlas

Diagnostics

Drug development

HUMAN CELL ATLAS

Regenerative medicine

Cellular changes in disease

Reference for molecular and spatial characteristics of cells

Diffeomorphism transformations

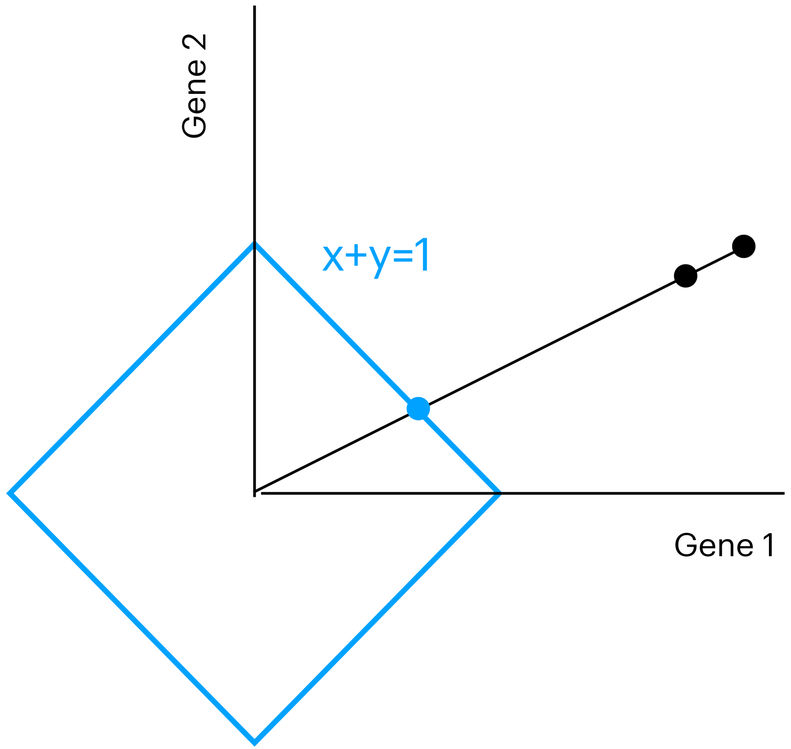
- A **diffeomorphism** provides a smooth one-to-one correspondence between two smooth manifolds
- F: M—>N is diffeomorphism if:
  - F is smooth (i.e., differentiable)
  - F is invertible (i.e., one-to-one mapping for both M—>N and N—>M)
  - F^{-1} is smooth

- Examples:
  - TRUE: Sigmoid (or logistic) function
  - FALSE: Step function
  - TRUE: A data rotation is diffeomorphism since it is smooth itself as well as its inverse
  - FALSE: Log(x) near zero (x —>0)

Size factor normalisation: use only genes' relative expression

- Different cell sizes
- But also different batches and efficiencies
- Absolute count values are not really meaningful
    - Different technologies, different read lengths, amplification and counting procedure
- Only the ratios among my measurements on different genes matters

- Even this is far from perfect:
    - Different genes may have different capture efficiencies in different technologies
    - Dropout effects
    - Distinction of Differentially Expressed genes and fold changes
    - DE analysis is still usually done on raw counts, not transformed data)



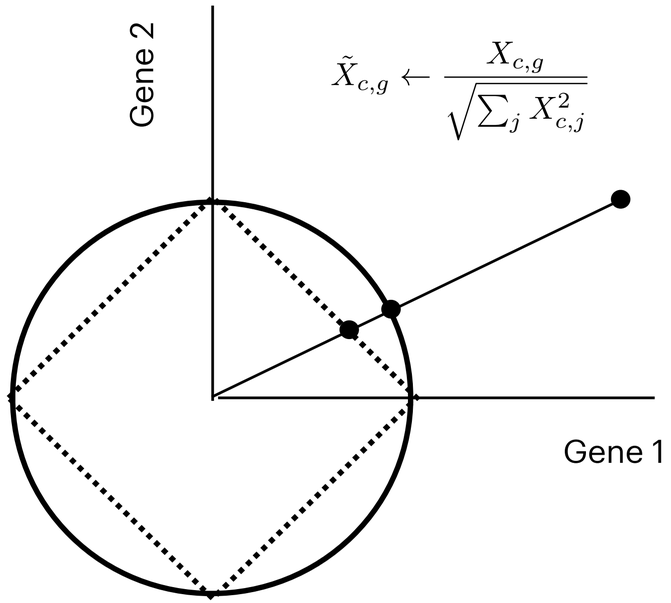
$$\tilde{X}_{c,g} \leftarrow \frac{X_{c,g}}{\sum_j X_{c,j}}$$

Cosine or Euclidean distance for expression data?

- Cosine distance: care only about the angel between two data points
- So after size factor normalisation Euclidean or cosine distance don’t make any meaningful difference

- In fact after L2 normalisation, there Euclidean and Cosine distance become exactly the same

$$D^2(x,y) = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} - \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||}\right)^2 + \left(\frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 - 2\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} \cdot \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}$$
$$= 2\left(1 - \frac{\mathbf{Y_x} \cdot \mathbf{Y_y}}{||\mathbf{Y_x}||||\mathbf{Y_x}||}\right) = 2 \cdot \text{cosine.distance}(x,y)$$



Data integration by matching Mutual Nearest Neighbours
Haghverdi,et al., Nature biotechnology 2018 (Supplement!)

scRNA-seq transform Summary

- HVG and other feature selections
- (Library) size factor normalisation
- Log transform (UMI vs. non-UMI)
- Order of steps
- Metric (Euclidean, Cosine distance, etc) choice

- There are sometimes redundant discussions, or doing unnecessary things or doubling, or dirty non-optimal procedures…
- Several (available) works flows can give you valid results
- So don’t worry too much! Just know what are the important properties in your data / question and make sure those make sense

- Nevertheless:
  - We want a standard approach among all researchers for
    - Interpretability
    - Reproducibility
    - Communication
    - Data integration
  - Seurat, scanpy, Human Cell Atlas (HCA) consortium

Counts matrix processing steps

My suggestion (it is not the consensus!):

~Data transformation

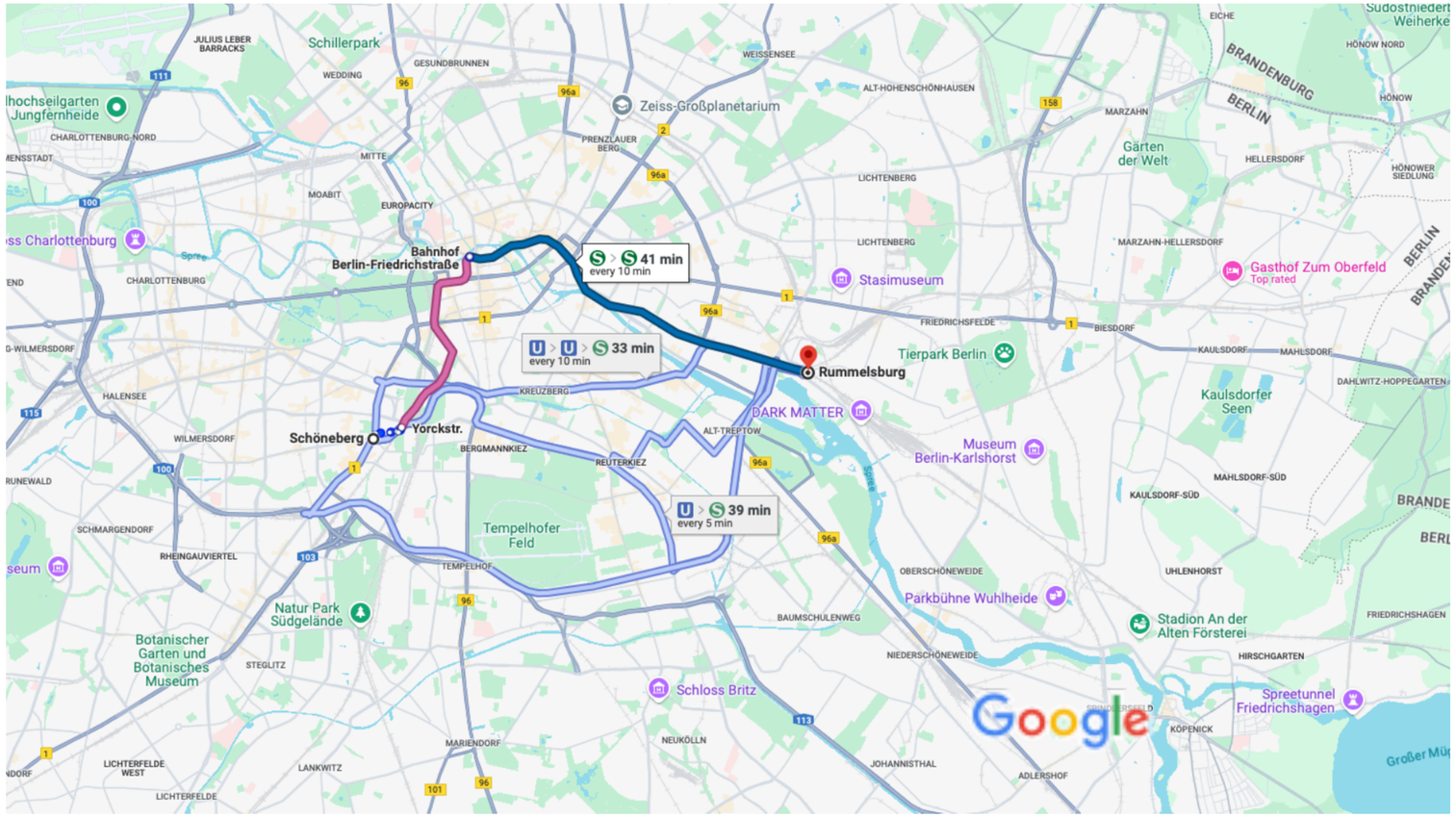
1. Feature selection (data and problem specific),
    1. Uninteresting genes (mitochondrial, ribosomal, etc) filtering + cells filtering (doublets etc)
    2. HVG selection (by Pearson residuals)
    3. Other, depending on the question e.g. about transitions or separate clusters?
2. Log transform (both UMI and non-UMI)
3. Size factor normalisation
4. Now Euclidean ~ Cosine distance
5. Dimension Reduction and visualisation
6. Data integration, (Unsupervised) Clustering and/or Cell type label transfer

Exercises:

a) What tests would you do to show a preprocessing pipeline is advantageous over another?

Different distance measures for different data/problems

- Example:
  - Euclidean distance
  - Great-circle (on earth) distance
  - S-bahn distance
  - S-+U-bahn distance
  - Drive distance



Metrics: properties and examples

- Valid metric (distance):

$$D(x,x) = 0$$
$$D(x,y) = D(y,x)$$
$$D(x,z) \leq D(x,y) + D(y,z)$$

- Distance between data points: Hamming distance, Euclidean, Cosine, geodesic, Diffusion distance,...
- Distance between distributions: Wasserstein (Earth movers distance), (application in Variational methods, image processing,...), Jensen-Shannon Divergence (JSD)

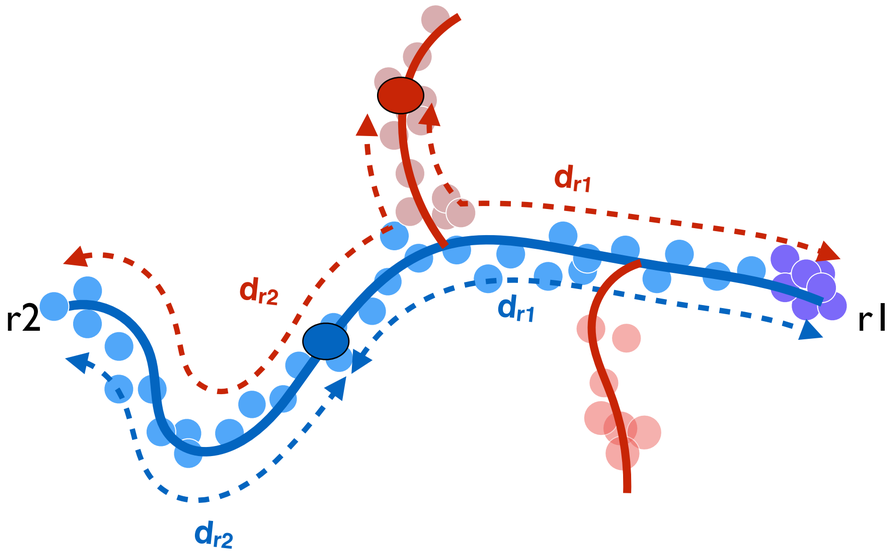
$$JSD(P,Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$
$$M = \frac{1}{2}P + Q$$

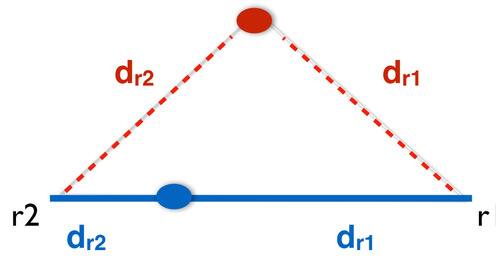
- Note Kullback–Leibler divergence is not a distance itself!

$$\mathrm{KL}(P||Q) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Trajectory analysis: pseudotime ordering and branch identification

- Pseudotime: on-manifold distance from the root cell
- Branch identification: generalised triangle inequality



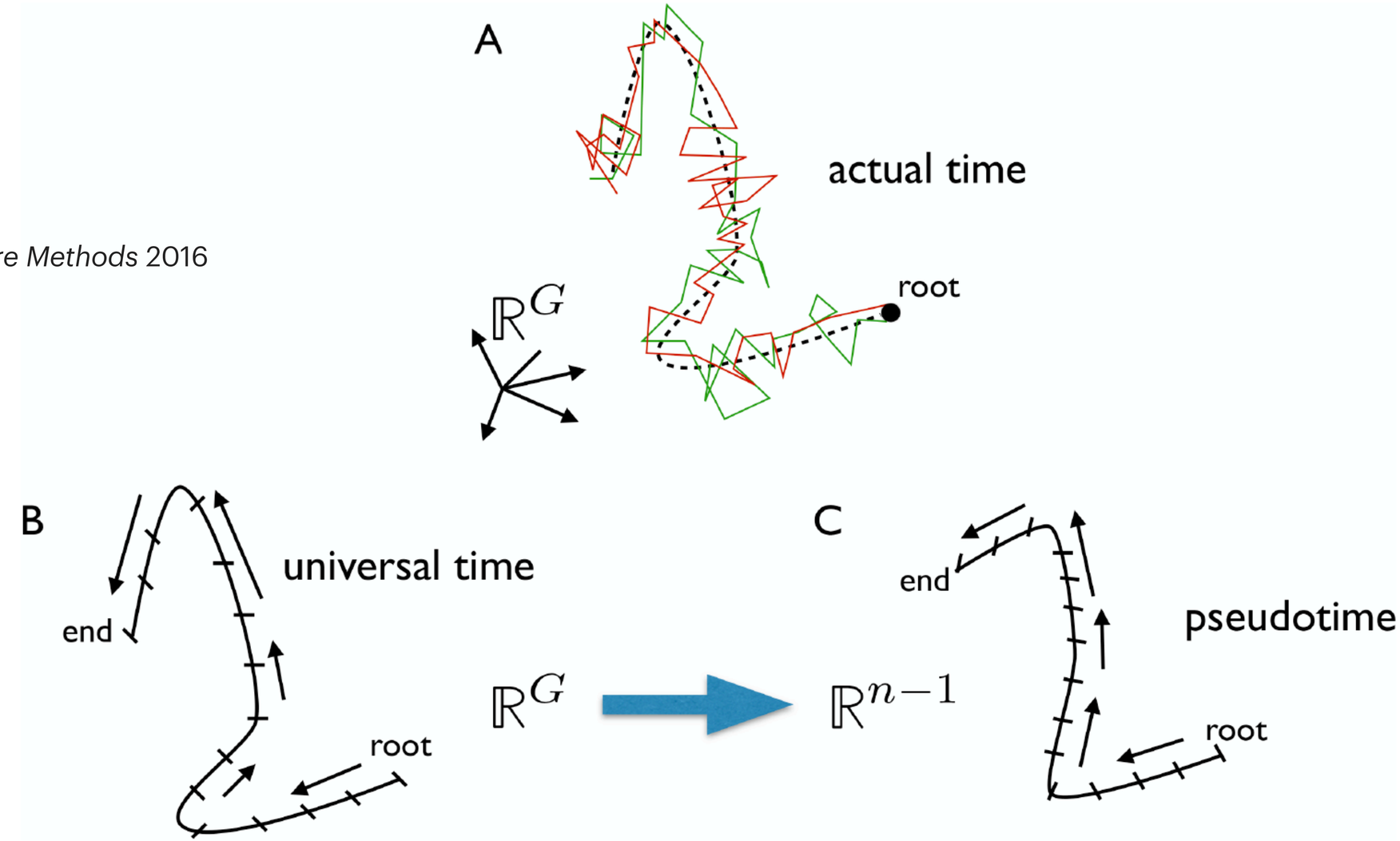


We need a continuity keeping DR method

Haghverdi *et al. Nature Methods* 2016 (Supplement)

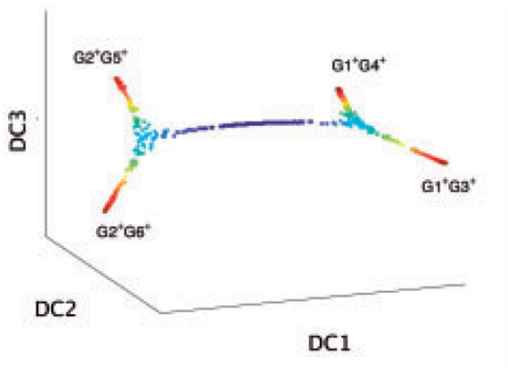
How pseudotime is defined

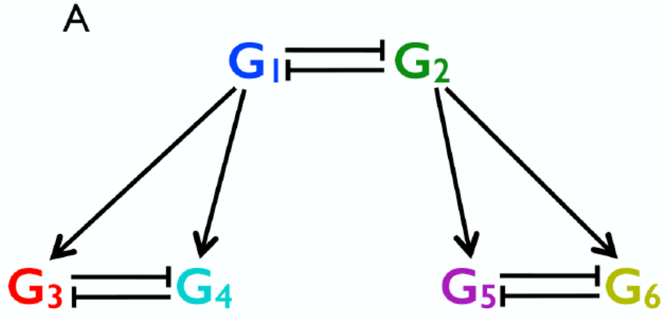
Haghverdi *et al*. *Nature Methods* 2016 (Supplement)

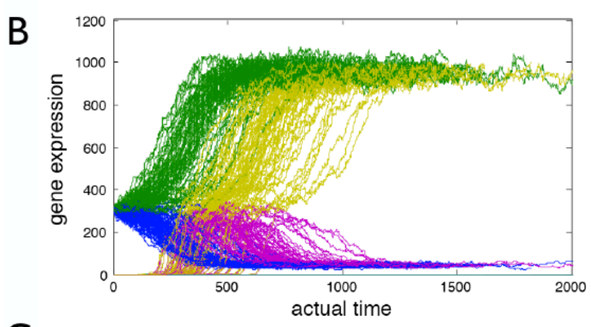


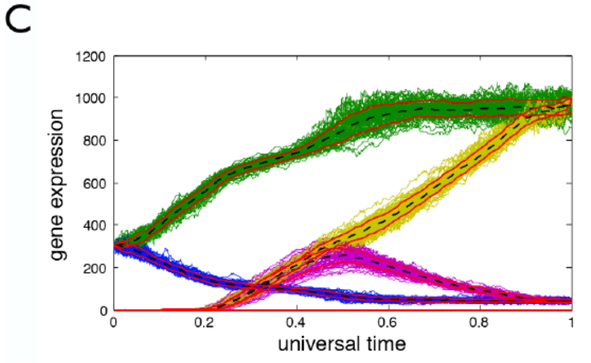
Supplementary Figure N7: A) Two (red and green) actual time single cell trajectories in gene expression space($\mathbb{R}^G$). Each jump on a trajectory happens in an (equidistant) unit of actual time. B) Universal time is defined as *arc* length on the data manifold starting from the root. This manifold $C \subset \mathbb{R}^G$ remains the same for several single cell trajectories, as well as for snapshot samples of single cells. C) Pseudotime (in

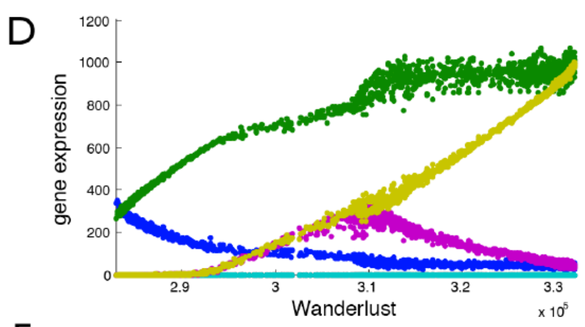
Actual time, universal time, pseudo time

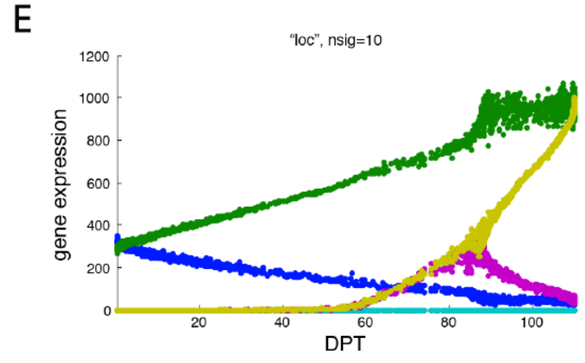


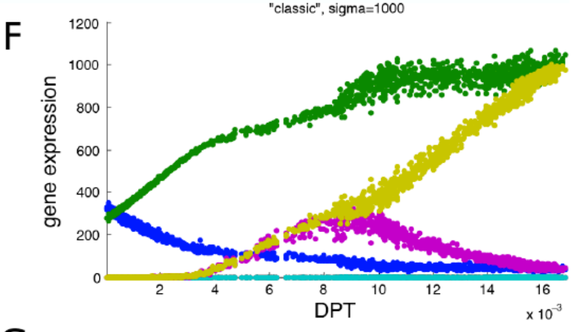


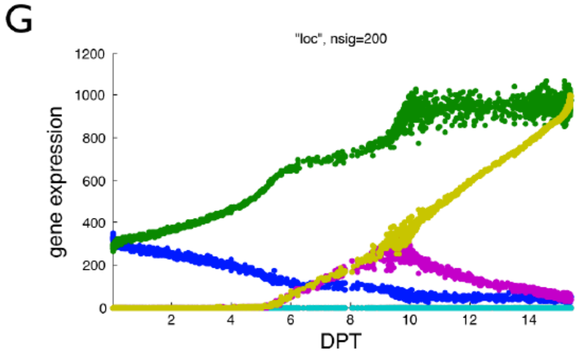












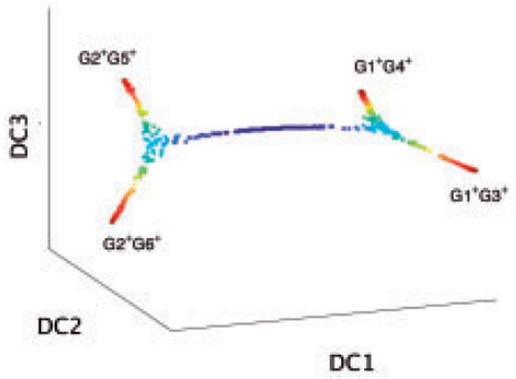
universal time

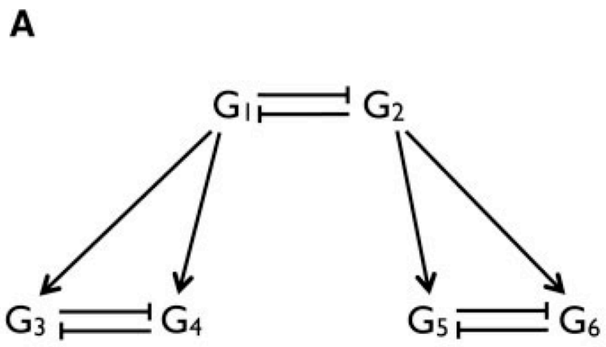
$$u(t) = \int_0^t \frac{1}{\rho(t')} dt'$$

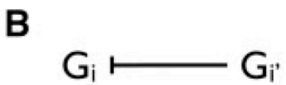
Exercises:

a) **Use the package BeeLine and BoolODE to simulate data and differentiation trajectory as below, then share the notebook with the class on Github. Specify the gene-gene interaction strength as well as the degradation rates. How can you increase the noise level? As we will need this simulation in future exercises as well, could you share your code with everyone?**

Haghverdi, et al. Bioinformatics (2015)

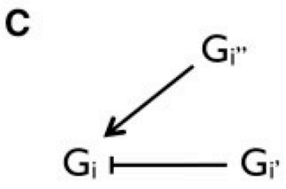






$$\varnothing \xrightarrow{\alpha_i} \text{Protein}_i \xrightarrow{\gamma} \varnothing$$

$$\alpha_i = k_i \frac{1}{1 + (\frac{\text{Protein}_{i'}}{k_{i'}})^{m_1}}$$



$$\varnothing \xrightarrow{\alpha_i} \text{Protein}_i \xrightarrow{\gamma} \varnothing$$

$$\alpha_i = k_i \frac{1}{1 + (\frac{\text{Protein}_{i'}}{k_{i'}})^{m_1}} \cdot \frac{1}{1 + (\frac{k_{i''}}{\text{Protein}_{i''}})^{m_2}}$$

```
decay          = 0.25;
synth          = 250;
K_d_inhibition = 200;
K_d_activation = 400;
n_inhibition   = 2;
n_activation   = 20;
init = [300 300 0 0 0 0]';
sigma   = 100e-0;
dt      = 1;
T       = 240;
a(:,1)  = init;
nRealiz = 400;
% ------------------------------------------------------------------------
% ------------------------------------------------------------------------
MM = 100000;
for i = 1:MM
    rp = randperm(2); rp = rp(1);
    if rp==1
        np = randn/8 + 0.22;
    else
        np = randn/8 + 0.77;
    end
    Np(i) = np;
end
k = 1;
for i = 1:length(Np)
    if (Np(i)<1)&&(Np(i)>0)
        SampleDistribution(k) = Np(i);
        k = k+1;
    end
end
figure, hist(SampleDistribution,100), title('Sampling strategy')
MM = length(SampleDistribution);
```

Matlab Code for toggle switch data simulation with 6 genes

Matlab Code for toggle switch data simulation with 6 genes

Page 2

```
% --------------------------------------------------------------------
% --------------------------------------------------------------------
for j = 1:nRealiz

    clear a
    a(:,1) = init;

    for i = 1:T-1

        a(1,i+1) = a(1,i) + dt*(synth*(1/(1+ (a(2,i)/K_d_inhibition)^n_inhibition)) - decay*a(1,i)) + sqrt(sigma)*randn;
        a(2,i+1) = a(2,i) + dt*(synth*(1/(1+ (a(1,i)/K_d_inhibition)^n_inhibition)) - decay*a(2,i)) + sqrt(sigma)*randn;

        a(3,i+1) = a(3,i) + dt*(synth*(1/(1+(K_d_activation/a(1,i))^n_activation))*(1/(1+ (a(4,i)/K_d_inhibition)^n_inhibition)) - decay*a(3,i)) + sqrt(sigma)*randn;
        a(4,i+1) = a(4,i) + dt*(synth*(1/(1+(K_d_activation/a(1,i))^n_activation))*(1/(1+ (a(3,i)/K_d_inhibition)^n_inhibition)) - decay*a(4,i)) + sqrt(sigma)*randn;

        a(5,i+1) = a(5,i) + dt*(synth*(1/(1+(K_d_activation/a(2,i))^n_activation))*(1/(1+ (a(6,i)/K_d_inhibition)^n_inhibition)) - decay*a(5,i)) + sqrt(sigma)*randn;
        a(6,i+1) = a(6,i) + dt*(synth*(1/(1+(K_d_activation/a(2,i))^n_activation))*(1/(1+ (a(5,i)/K_d_inhibition)^n_inhibition)) - decay*a(6,i)) + sqrt(sigma)*randn;

        if sum(a([3 4 5 6],i+1)>950)
            ind = ceil(i*SampleDistribution(round(rand*MM)));
            DataDiff(1:6,j) = a(1:6,ind);
            break
        end

    end

end

figure, plot(a'), title('Single realisation from toggle switch network')

% save diffusionInput DataDiff
```

Next session

Dimension reduction techniques

Grading system

- 10 points: (~5) Exercises and seminars
- 10 points: final project
  - Groups of 2-3 people
  - 5-6 pages report including Abstract, Intro, Methods, Results, Discussion
  - In the Abstract specify what problem you are addressing and based on which lecture(s)/exercise(s) it is
  - Share coding scripts and data (on GitHub) for reproducing your plots

Thank you for your attention!