Computational Biology

Genetics ≠ Genomics

talking about individual genes.

CRISPR - CAS9

LV0 → too much abstraction of biological situation.
→ very minimalist system.

LV1 → Using general published tools to understand particular situation.

LV2 →

LV3 → Computer comes up with hypothesis itself.

Sequences. → of nucleotides that form DNA and RNA

PDB – datasets – protein sequences.

SRB (NCBI) – Genome raw sequences.

limitations of sequencing.

→ short reads
→ imperfect reads
→ Biased reads.

→ Microarray

Connectivity Map -> MIT Project for research of gene expression under different environment.

level-1 Utilizing already published datasets using established analysis methods to point out defects.

BLAST- basic local alignment search tool

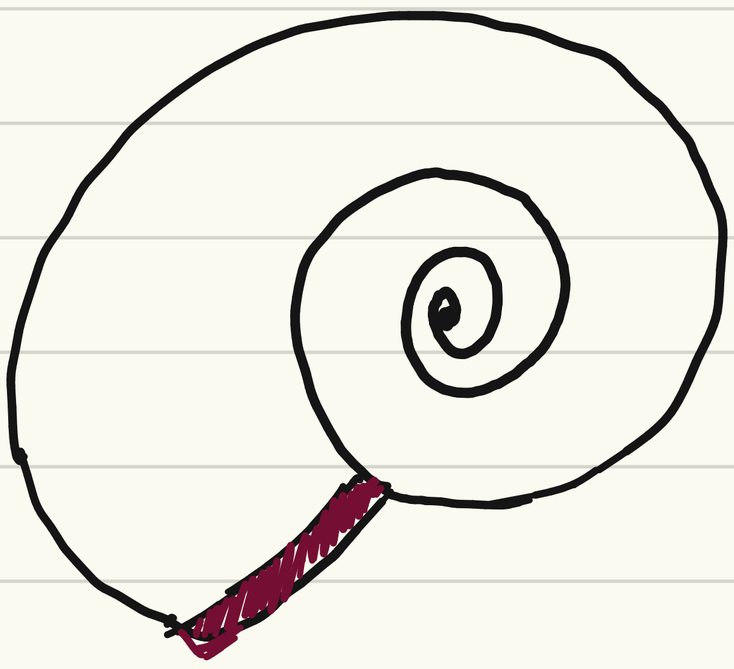
↑ online repository for finding matching sequences from the natural world.

level3- "Make biological discoveries using Comp."

looking for micro-structure in genomic data to infer things like, cell stress

Human reference genome → first generation sequencing by Human Genome project.

25,000 genes ⟶ 1,00,000 Proteins.



Single gene gives rise to multiple proteins

→ not all exons are completely present in mRNA structure.

DNA

| transcription.

pre-mRNA —Alternative splicing→ isoform / transcript

introns exons.

mRNA 1, mRNA 2, mRNA 3...

| ↑ wesequence small bits of these.

RNA sequencing

bulk

average gene exp per tissue

single-cell

individual values in single cell data.

2016 ten years of nex-gen sequencing

→ Paradigms (new)

Short read sequencing
- lower cost
- higher accuracy
- population lv. research

long read sequencing
- research (genome assembly)

Research field since 10 years
Latest → - Database generation
- Routine clinical sequencing
- Pathogen DNA monitoring
- Population-level studies

CoNS → - Error rates are still high
- Ethical and ownership issues.

Pourpose of research – how genome sequence variation underline phenotype / disease

→ Read lengths affect the ability to resolve complex information.

→ Sequence depth affects the ability to detect low-frequency variants.

→ high error rates can give false-calling

→ There are also inherent biasing in sequences.

→ there are often Annotation errors in datasets

→ open data standards are used.

→ high quality genome-databases have comprehensive metadata

Lecture - 9/8

mRNA single cell sequencing exp.

→ in 2015 Drop-seq was discovered which led to 100x increase in thoughtput.

→ **Droplet based scRNA-seq.**

barcodes are unique sequences that are added to each RNA molecule of a cell.

AMPlification (PCR)
increases amount of the genetic material to work on.

UMI-count - total no. of unique molecules of a particular mRNA molecule.

greater transcriptome coverage in a cell is more important than no. of cells sequenced.

there are tools to find out how many cells one needs to sequence to find a particular rare gene.

technical replicates vs. Biological replicals

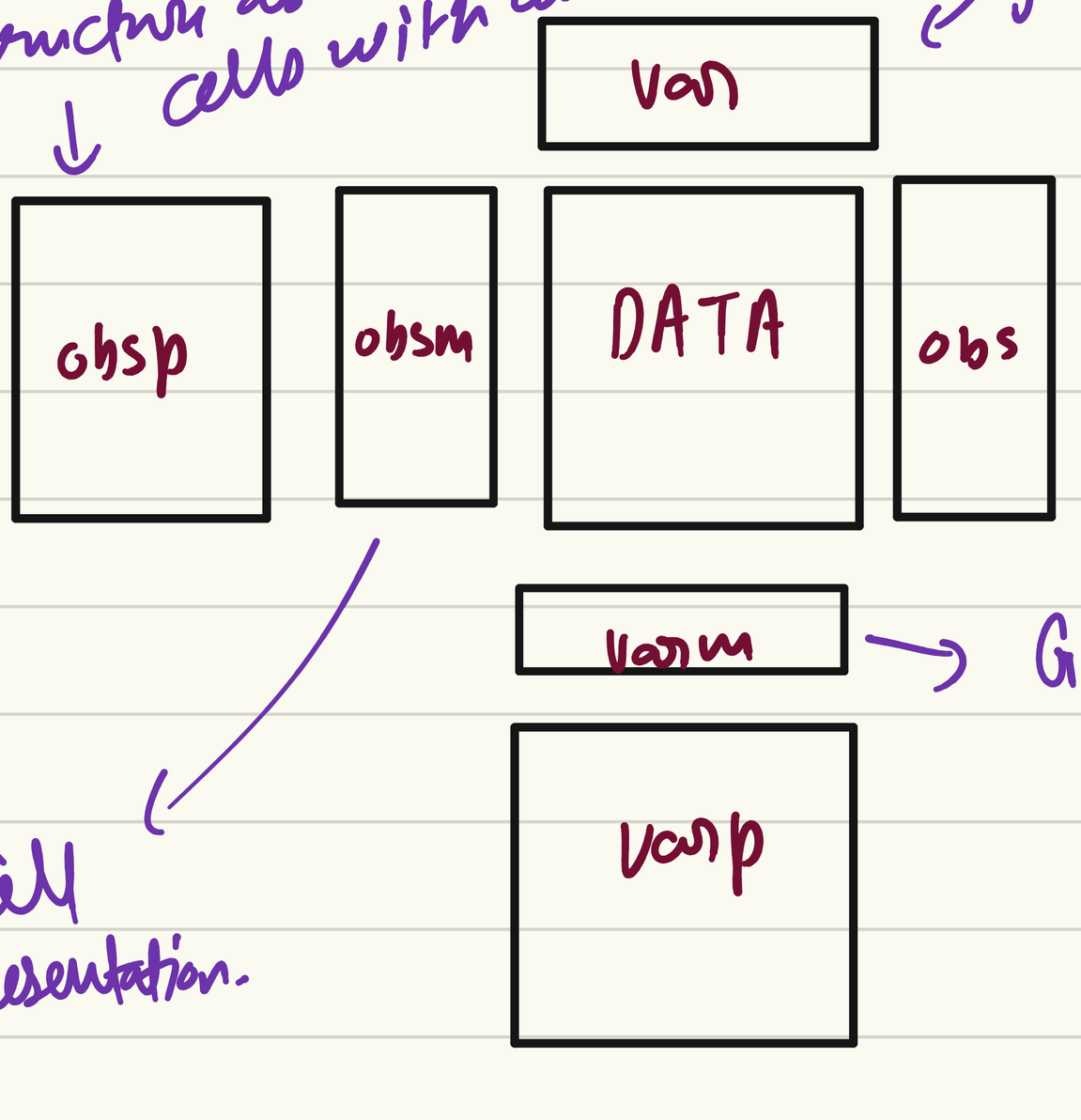
Sequencing many times so that accuracy increases

biological variation in the incoming material -

1 Dimension Reduction on cell gene matrix
2 Clustering
3 trajectory infern
4 Annotation.
5 Differential Gene exp.
6 Gene regulatory inference.
7 Cell-cell communication infere.
8 gene expression progen int
9 Pathway analysis.

SCANPy ← Storing unimodal data with AnnData

Graph like structure associating cells with cells ↓



← gene annotations.

← cell annotations.

→ Gene representations

cell Representation.

$\{uns\}$

↑ unstructured data.

quality control → Normalization

↓

feature selection. (more influenced by Biologics.)

↓

Dimentionality reduction. (PCA? +)

← Cell-Cell distance. Computation

(cells as nodes
values as edges)

↓

Unsupervised Clustering.

Quality control

→ lot of entries are zero
→ performing QC also interferes with biologics.

Ensure that all cellular barcode data correspond to viable cells.

lostinfo

Automatic thresholding

→ MAD (Mean absolute deviation)

Lecture - 11/8

Ambient RNA Problem

Some of RNA is expressed so much that they end up in every droplet.

→ Ambient RNA is treated as noise.

→ we seperate this noise on distribution basis.

Doublets (two cells with same barcode)

→ We introduce two cell's doublets artificially, we then carry out the principal component analysis.

→ doublets mostly end up in one particular group in PCA.

→ doublet detection is still an area of research.

NORMALIZATION

→ make profile of each cell comparable.

Assumptions: → all cells in dataset initially contain equal no. of RNA molecule.
→ Count depth difference is only due to sampling.

Scaling is the solution of this

$$CPM_{ij} = \frac{n_{ij}}{R_j} \times sf$$

(n_{ij} ← gene UMI count; R_j ← total UMI count.)

(counts per Million)

→ logarithmic transformation is also done.

$$\log(CPM_{ij} + 1)$$

→ Regressing out biological effects.
effect of different life cycles stages of cell on data.

→ We take cell populations at different stages of its life and check difference in PCA.

Feature selection.

Selecting suitable features.

→ takes more time to process more features.
→ we have to remove bad features
→ Redundance is also removed.
→ dimentionality reduces.

Feature selection aims to exclude uninformative genes which do not represent meaningful biological variation.

Genes are binned by their mean expression and the genes with highest

→ highly informative genes across cells will have a high deviance value.

Dimentionality Reduction.

→ Visualizing the dataset.

Why? → data - high dim.
→ Noisy
→ non-trivial dist.
→ Non-linear Manifold.

PCA → linear combinations of existing dimentions to create new dimentions (axes) that have highest variations.

16/8

Curse of Dimentionality.

as the Number of genes increase the variance that can be explained by any one principal component reduces.

Batch effect - Cluster Variation due to environmental factors.

Biological challenge. -

Biology messing with cell identity.

for cancer research biological research is important.

how to validate clustering from computation

-> Experimental differentiation. seems to be the best option.
-> care has to be taken to mentain quality and information of samples.

Clustering pitfall : overclustering
Cluster spliting

Clustering approach.

1. K-means : CS771 — only K- clusters. use distances from K-means to classify.
2. hierarchical clustering:
   We're going to look for 2 datapoints which are closest to each other. then we repeat the process.
   → gives a tree structure.
   → cutting the tree at any one level gives the clusters.
3. Graph based clustering.
   → very high data dimentionality.
   → may not be simple Gaussian distribution.

Using a Graph → { V, E }
vertex and edges.

→ Adjadency Matrix
→ sparse adjadency Matrix
→ data matrix is also sparse.

sparse x sparse data is most useful for graph based clustering.

dimentionality : V x V

→ Dabate still exists that what is the best measure of distance.
→ learn Manifold from data and then get distance in that manifold

Graph type

→ K-nearest neighbors.
→ Shared-nearest neighbours graph.
  → obtained from K-nn
  → similarity is based on nearest shared neighbours.

slides have more information.

distance ⟷ Rank

#shared neighbors.

edge => atleast one shared neighbors in K-NN

edge strength => Rank

how to define a Community

Clique - a maximum complete subgraph

quasi-Clique - Close enough but not necessarily complette.
(γ-dense)

$$\gamma = \frac{E_s}{V_s(V_s-1)/2}$$

E_s ← actual no. of edges.

$V_s(V_s-1)/2$ ← total possible edges.

SNN-Cliq Algoritm.

→ extract local maximal quasi cliques.

→ merge quasi-cliques to get communities.

Graph Cuts

Cut the graph at edges to define a community

→ Change objective fn. to account Size and Volume of the community.

→ Ratio cut
number based

→ Normalized Cut.
volume based.

→ finding Minim on normalized is NP-hard.
ratio cut

measure of structure.
density of connection within a module

Modularity optimization

SEURAT (R package)

shared overlap in local neighborhood

K-nearest neighbour ⟶ Jaccard distance.

↑ very noisy.

↓ modularity optimization

LOUVAIN ALGORITM

---

weight of connection (P,q)

modularity

$$Q = \frac{1}{2m} \sum_{p,q} \left[A_{pq} - \frac{k_p k_q}{2m} \right] \delta(C_p, C_q)$$

probability of a connection by chance.

$$k_p = \sum_q A_{pq}$$

$C_p \leftarrow$ cluster to which p is assigned.

$$\delta(C_p, C_q) = 1 \quad \text{if} \quad C_p = C_q$$
$$\text{otherwise} \quad 0$$

$$m = \frac{1}{2} \sum_{pq} A_{pq}$$

LEIDEN ALGORITM

measures of similarity.

- ARI -
- NMI -

CellSIUS ← clustering for rare cell population.

Cell type annotations

→ process of labeling clusters on basis of phenotype.

Associate in each clusters a set of genes which are highly expressed in this cluster but not otherwise.

Scanpy is a Python package for single-cell RNA sequencing (scRNA-seq) analysis. It provides a user-friendly interface for preprocessing, analyzing, and visualizing scRNA-seq data. Scanpy offers a wide range of functionalities, including quality control, normalization, dimensionality reduction, clustering, differential expression analysis, trajectory inference, and cell type annotation. It also integrates with other popular scRNA-seq analysis tools, such as Seurat and Cell Ranger. Scanpy is widely used in the single-cell genomics community and has become a standard tool for scRNA-seq analysis.
Scanpy is a software tool that helps scientists analyze data from single-cell RNA sequencing experiments. This type of experiment allows researchers to study the gene expression of individual cells, which can provide valuable insights into how cells function and interact with each other.

Scanpy makes it easier for scientists to process and analyze this data by providing a user-friendly interface. It offers a variety of useful features, such as checking the quality of the data, normalizing it to make comparisons easier, reducing its complexity to identify patterns, grouping similar cells together, identifying genes that are differentially expressed between cell types, predicting cell development paths, and labeling cell types.

In addition, Scanpy can work together with other popular tools used in single-cell RNA sequencing analysis, such as Seurat and Cell Ranger. This makes it even more powerful and flexible for researchers.

Overall, Scanpy is widely used in the scientific community and has become a standard tool for analyzing single-cell RNA sequencing data. It helps researchers make sense of complex data and gain a better understanding of how cells work.

Alphafold - Deepminds.

Date.
Page.

simulating protein folding using AI

→ Physics based modelling tool - Rossetta.
→ DATA - World wide Protein databank.
(All known protein structure).

Alphafold - 1

Multiple sequence alignment

"3D structure of a protein is much more stable ~~than~~ overtime than sequence that code for protein"

Same protein
(different species)
|| from similar evolutionary history
different ~~species~~ sequence
but same function and structure.

Distogram - 2D array of distance between every pair pair of residues in protein.
(independent of rotations / translations)

Sequence +MSA features → DNN → Distogram → Gradient descent on protein specific potential and prediction scores.

Alphafold - 2

Sequence → MSA → Evoformer.

Principal Component Analysis

→ learning dimentions that capture maximum variance.

→ change in axis which we represent the data.

→ implement w_1 such that the variance

$w_1^T S w_1$ is maximised.

$\underset{w_1}{\text{argmax}} (w_1^T S w_1)$ ← objective function

→ Find out : first principal component.

d=20 ⇒ take top-20 (etc.) Principal components. (eigenvalues)

PCA is identical to classical scaling.

↑

Preserve distance (Euclidian.) between higer and lower dimention

→ PCA retains only large pairwise distances, but local distances are lost.

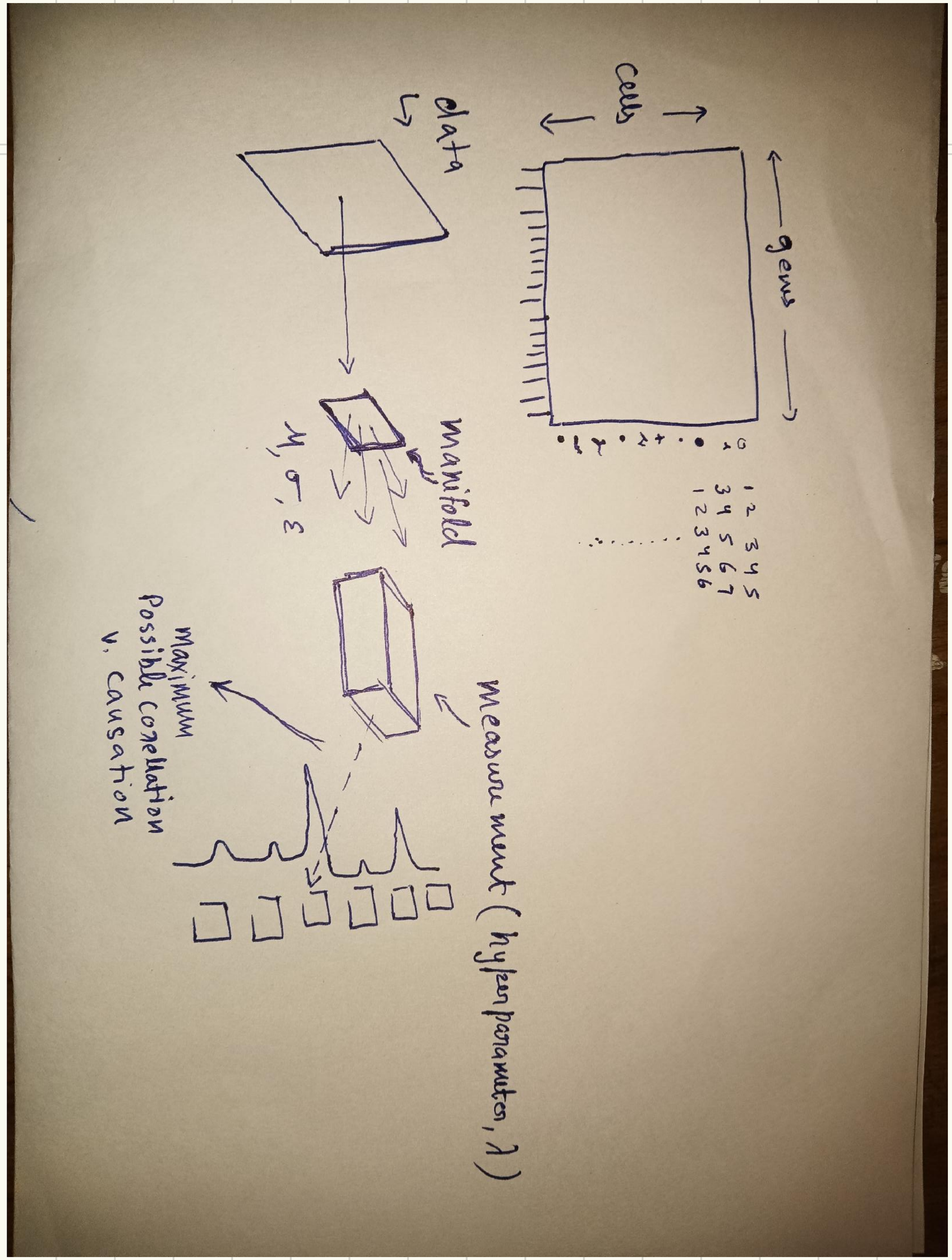
ISOMAP → preserve pairwise geodesic distances.

↑

distance measured over a manifold.

→ first compute k-NN graph using euclidean distance.

→ the apply preservance of pairwise geodesic distances.



CS690 (1 Sept)

Batch effect (additive effect)

Mutual nearest neighbour

Mutually similar expression profile between different batches.

pair specific batch correction vector = vector diff b/w expression profiles of paired cells

↓

Cell-specific batch correction

Seurat V3

Canonical Correlation analysis (CCA)

→ Correlated sources of variations between two datasets

L2 Normalization → Correcting in differences.
* in global scale
Vector normalization.

if there is a particular type of cell that does not has any correspondence with the other batch MNN should not classify these.

CCA
↓
L2
↓
Anchors
(high scoring)

Once the data is properly integrated and clustered we can use the ~~o~~ marker genes ~~based~~ to annotate cells.

LABEL transfer. (Surat accuracy pretty good)
transfer learning,

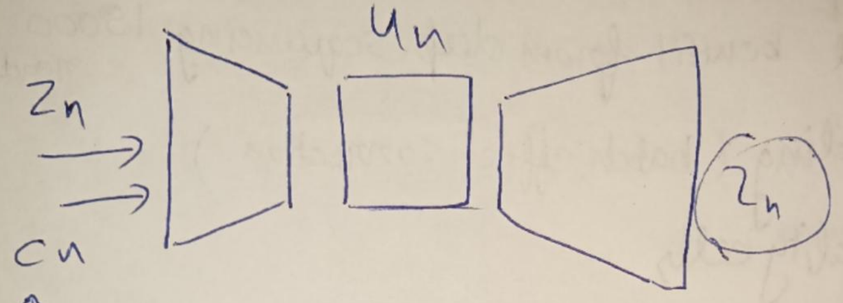
Reference data → Query

Harmony (integration method)
→ fuzzy clustering

* for simple integration ~~pro~~ problems harmony ~~does~~ well
Clusters + Multiple batches → diversity within cluster (max)

Supervised integration (SCANVI)

z_n => latent space corresponding to cell



(observed from some cells)

* Performs both annotation an integration.

→ Any integration method has to ensure that biological ~~trajectory~~ trajectory

KBET - K nearest neighbor batch effect.
→ if well mixed data
→ then if you choose one neighborhood. the cell

UMI - Unique molecular identifier.

Sparse gene expression matrix
→ substantial benifit from deep sequencing 15000 reads/cell
→ Surface labeling (batch effect correction.)
→ low quality cells

8 CS690 (8 sept) (1)

NGS tech for spatial data

→ 10x Visium (Most used)
→ Sterio-seq (Most advanced)

General problem → image segmentation + gene ID + Very Sparce dataset (missing values)

Spatial transcriptomics

Basic Unit → Spots

many cells

Imaging based (using Microscope) Eg: FISH (Less genes)

Sequencing based.

Spot size → 10x Visium (2019) = 50 µm (~10 cells)
→ Slide-seq = 10 µm (~ 1-2 cells)

Spatially Variable Genes ← genes selectively expressed based on spatial params.

Cell type Deconvolution → Supervised (SC-RNA-data)
→ Unsupervised (topic modelling / VAE / HMRF)

Principle

Tissue → SC (Single cell) + ST (spatial) (2)

Spot gene expression signature = linear Combination of (Single cell signature)

RCTD (method)

i - spots
j - genes.
k - cell types.

$\rightarrow y_{ij} = \text{Poisson}(N_i, \lambda_{ij})$

$\log(\lambda_{ij}) = \alpha_i + \log(\sum_{k=1}^{K} \beta_{ik} u_{kj}) + \gamma_j$

α_i ← Spot Variation

u_{kj} ← Comes from se-seq data.

Cell 2 loc (Method) (supervised.)
→ Probablistic (likelyhood, Prior)
↓
Posterior for (LV) ← Cell type Proportions

Dest VI (latent variable model (LVM)) (3)
→ depending on spatial context & reference signature can vary.
→ Sharing of latent space in SC and ST

Cell type Sig (SC) → E → γ_n → D

Spatial Trans Criptom → E → γ_B^c, β^c

Spice Mix - if spatially spots are co-located then their biological state should also be similar.

HMRF ← Hidden markov random field.

there is some sort of function that relates the hidden biological state to spatial information.

Markov ← Current state is only dependent on neighbouring state.

Can be arbitrary and complicated

$h_n \mid h_{n-1}$ → Markov property

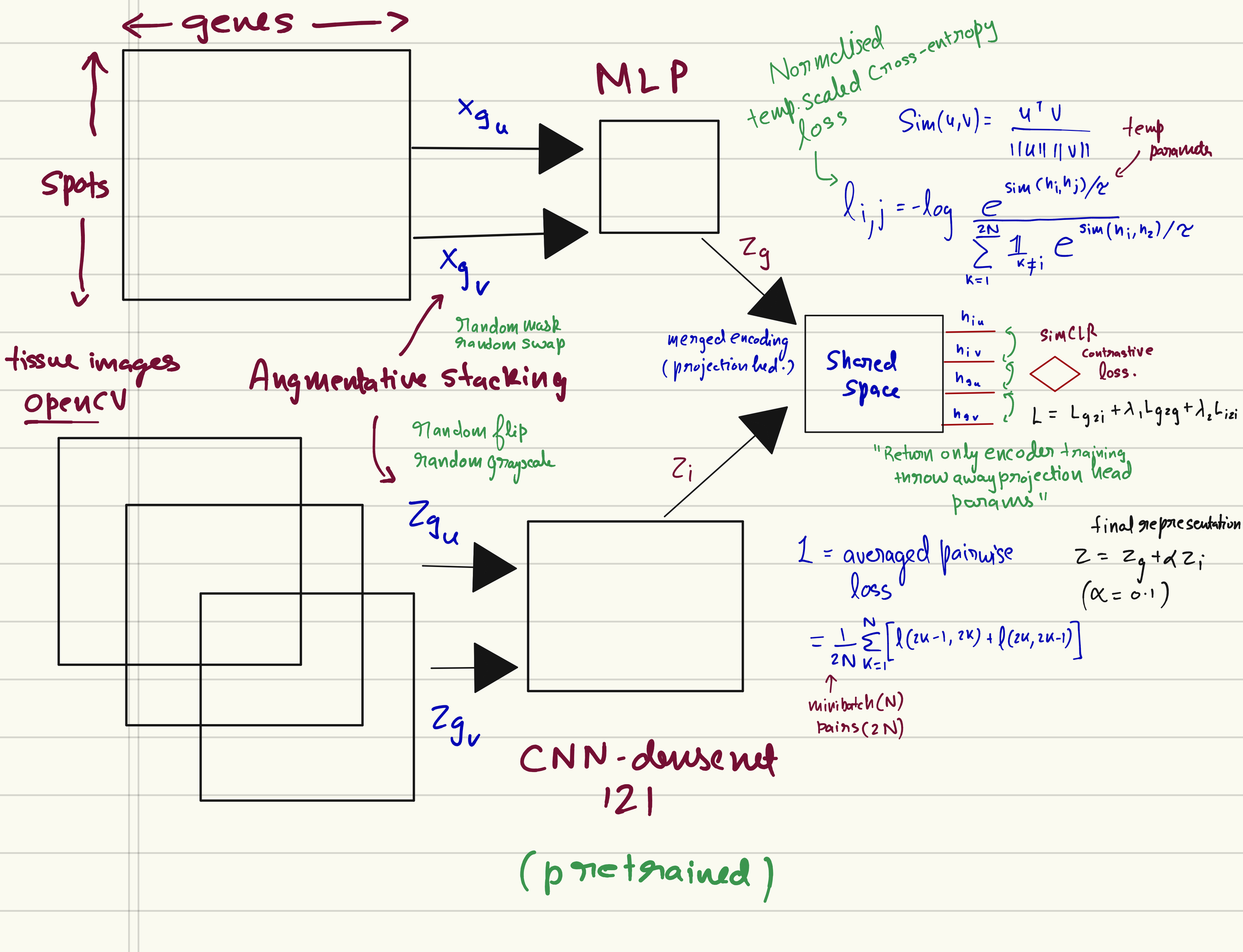
$O_n \mid h_n$ → markov property.

Assignment - 2

Clustering and classification using ConGI

Contrastive learning in latent space

spot associated gene matrix



Shared representation = positive learning

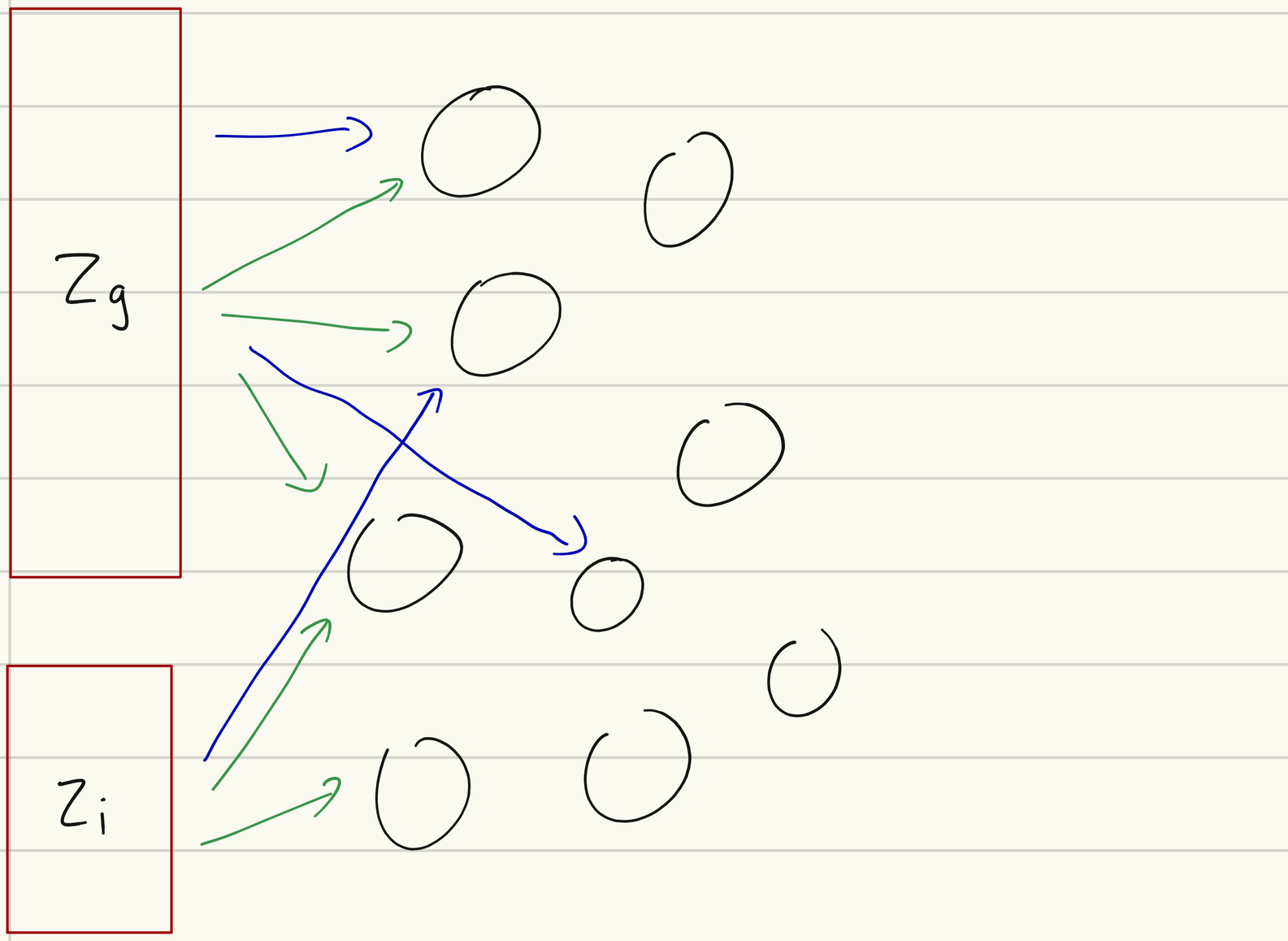
Unshared = negative

Optimizer = LARS (layerwise adaptive rate scaling)

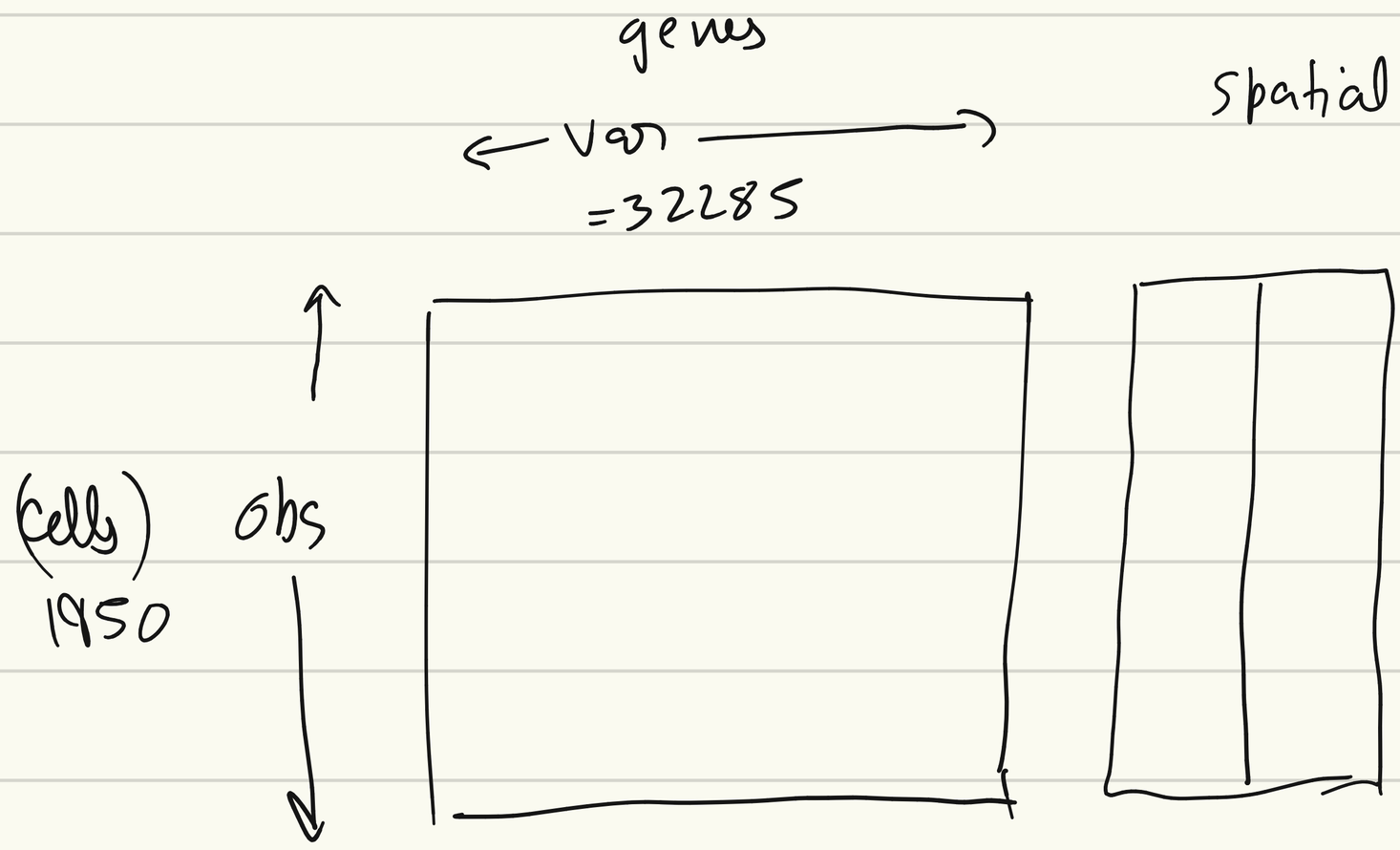
"Magnitude of update depends on weights norms for better training"

"LARS uses seperate learning rate for each layer and not each weight."

Spatial domains



Clustering + Refinement
(mclust, leiden) (against SpaGCN)
optional



Counts

position

x = pos(x)
y = pos(y)

labels (dropna)

label_ids(x x y)

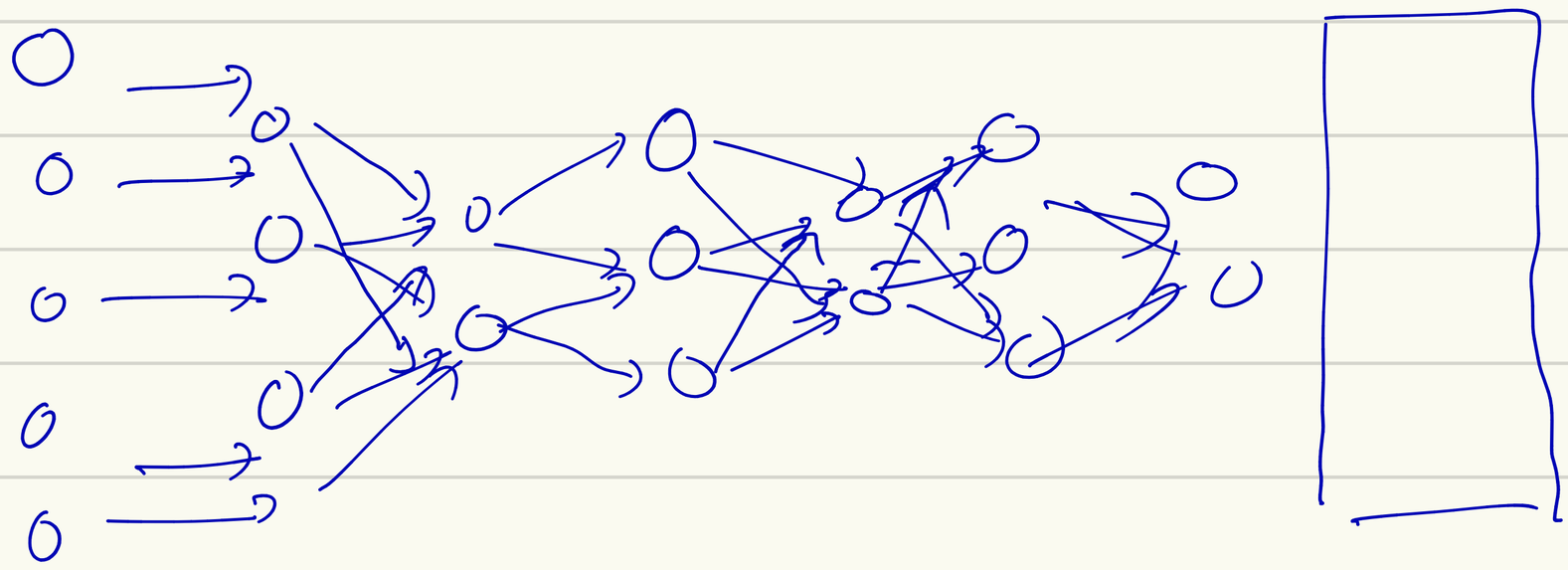
lbl_x = lbl_x + 0.5
lbl_y = lbl_y + 0.5

meta_pos ← joined (counts + posids)
meta_label

adata obsm(spatial)
adata . obs (label)

GPCR- g-protein coupled receptor.

(Everything is protein)



→ gene expression leads on method.

LIANA +

FASTA – text format for representing nucleotide sequence.

FASTQ – Sequence + Base quality.

NCBI SRA – database in FASTQ format.

$$Q = -10 \log_{10} p$$

Q ← base quality score

p ← probability that base call is incorrect.

Read alignment

Genome scale index structures.

genome scale index

1) k-mer

Suffix tree

→ Different biological properties learned by different levels of latent variables.

VAE

NVAE,
ladder VAE

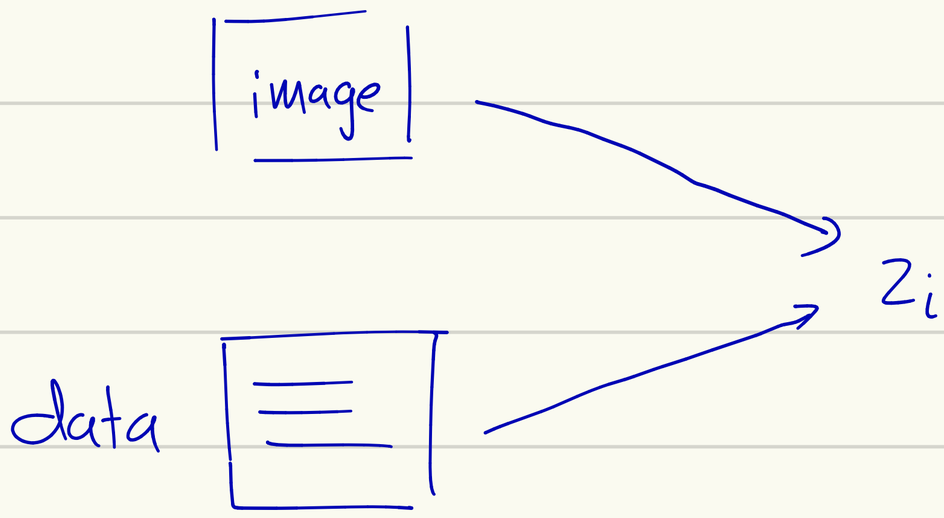
"Contrastive learning"

information in
latent space.

Project 1:

Mosaic - Using "Uncommon features" to learn latent representation.

Project 3: Multimodal VAE



Project 4:

Graph neural network.

→ do not have very good imputation methods for multiomics data.

→ dataset

↑

graph provided ← "Missing data" / Smooth-ness prob.

→ MAGIC - "diffusion map approach"

MAGIC
+
GNN
⟶ GREP

→ Read literature

→ Base model ready (Comparable to State of Art)

→ Better model

1st week Nov, Review lecture

use github ~~~ - Every week discussion (friday)

Suffix Array.

"Spatially more efficient and longer search times"

→ Suffix array to store whole genome - 12 Gb

→ (text) T ⟶ [0, m] m = len(T)

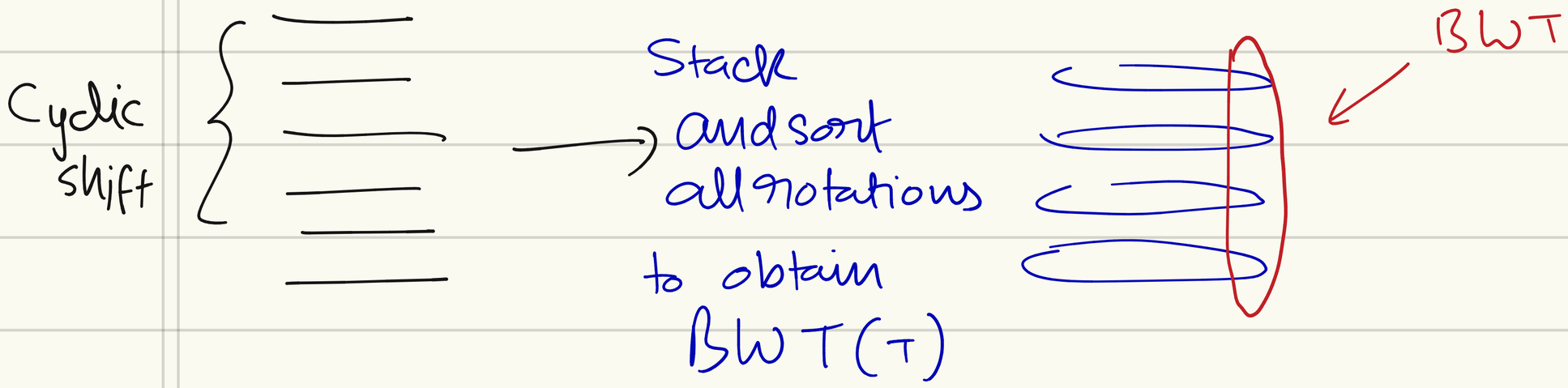
take all suffixes ⟶ sort lexicographically.

→ "multi key quick sort"

→ space efficient linear time construction of SA

new paper.

Burrow's wheeler transform



LF mapping ← last to first mapping.

Imputation + Data cleaning.

→ BERT based pattern recognition and generation.
→ Graph representation learning with IGRAM
→ Casually aware imputation
→ Vertex and Edges based Friend network reinforcement
  → tracking closely related observations togather.
  → optimizing friend networkand observations togther.

CONGI

→ folder containing the data has now been properly passed through library.

→ Empty images in patches of spatial('obsm')
have been sorted.
Cell positions wRT image more clear.
→ trained on 10 Epochs obtained
Combined Embedding structure.
var.

$$z = x_g + \alpha x_i$$

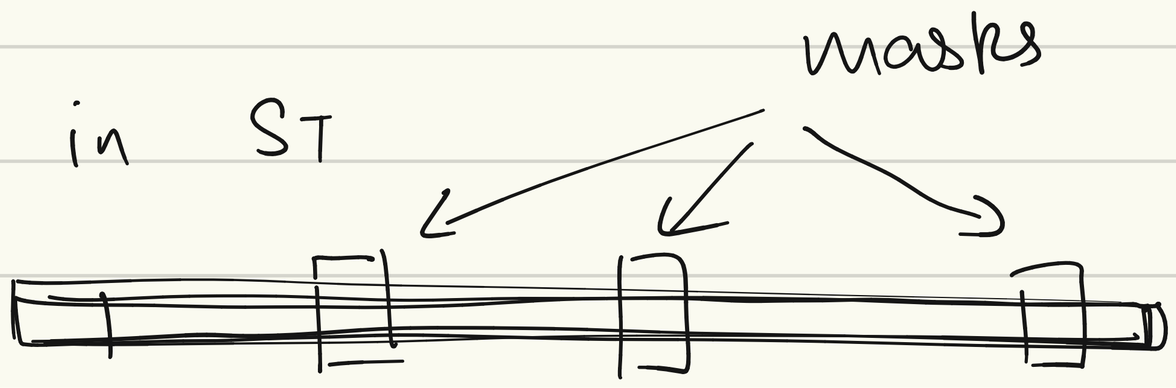
training takes some time
10-15 min
for 10 epochs

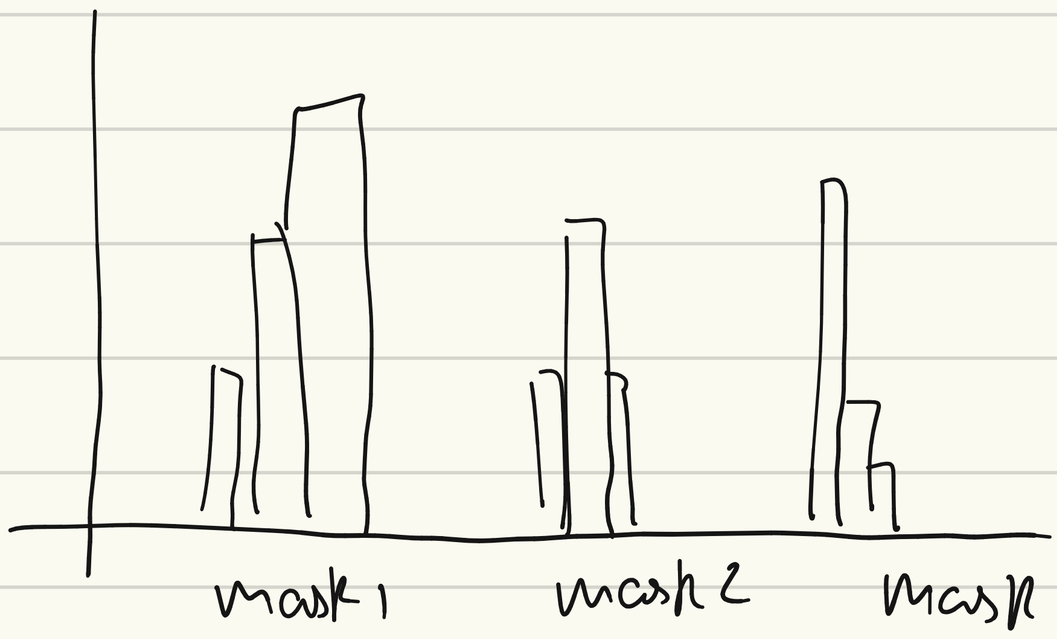
z - has structure
that can be studied
for plotting.

→ mclust using R-py2 package is unable to load

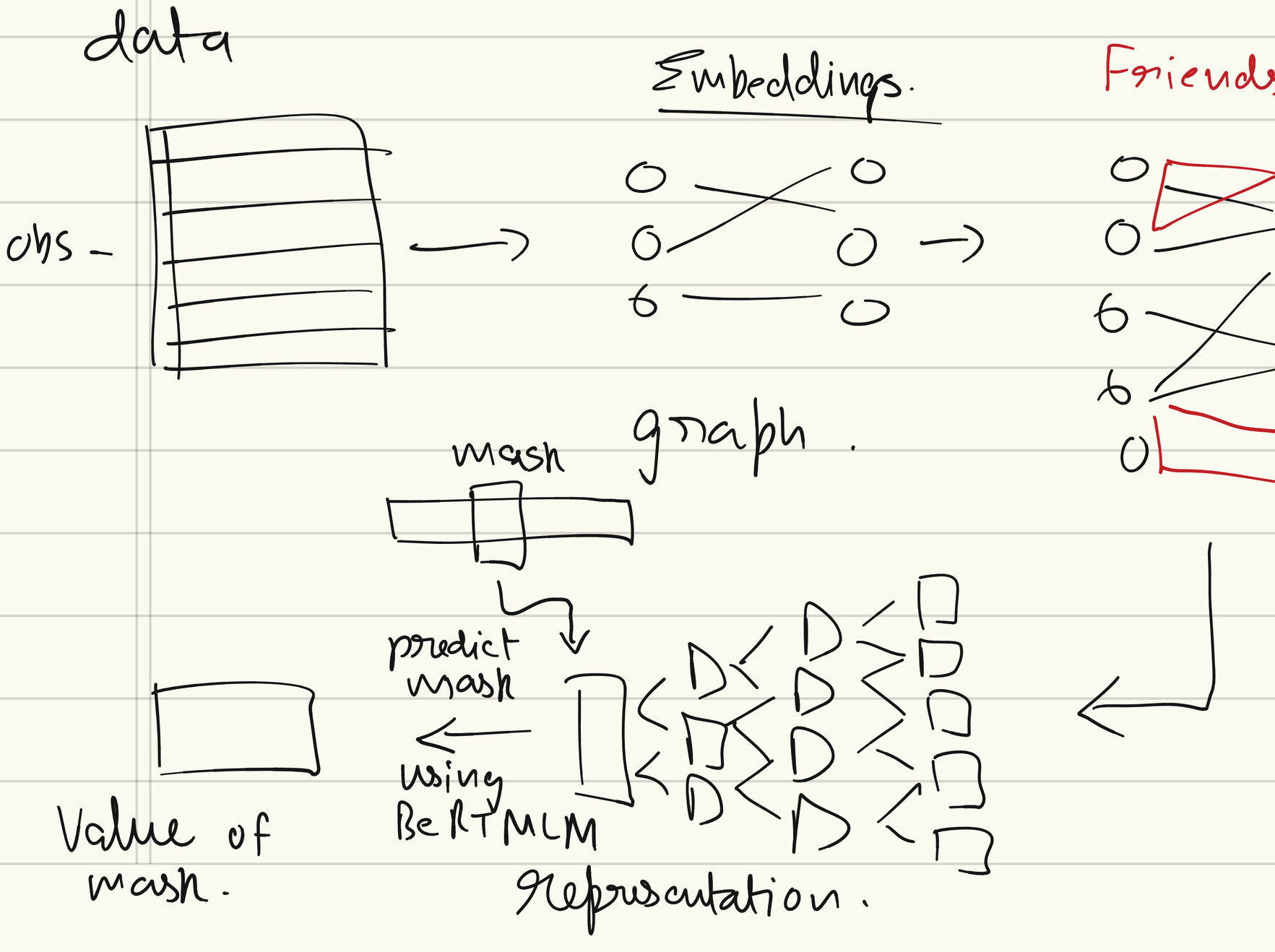
→ thus ari and cluster prediction's are unable to learn, No PLOTS.

MLM- in ST





Deep Generative Models.



GRAPE

+

MAGIC

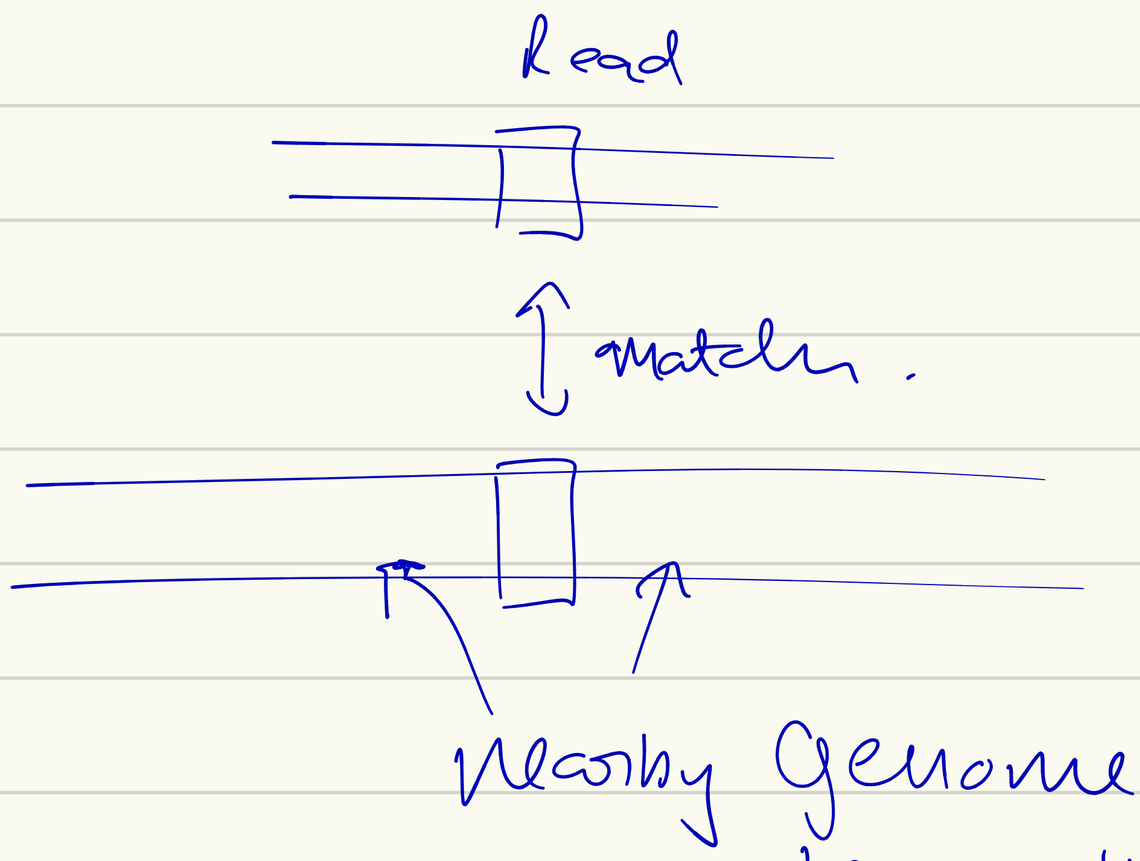
scBERT

2 Graphs. → Label

→ space - based.

BWT based. Alignment-

Read alignment



BWT and LF mapping.

→ GoTcMethod Ultrafast human genome alignment
isk cite.

→ quality aware back tracking

if there is a multiple candidate subs.

→ K-mismatch occourences of P and T

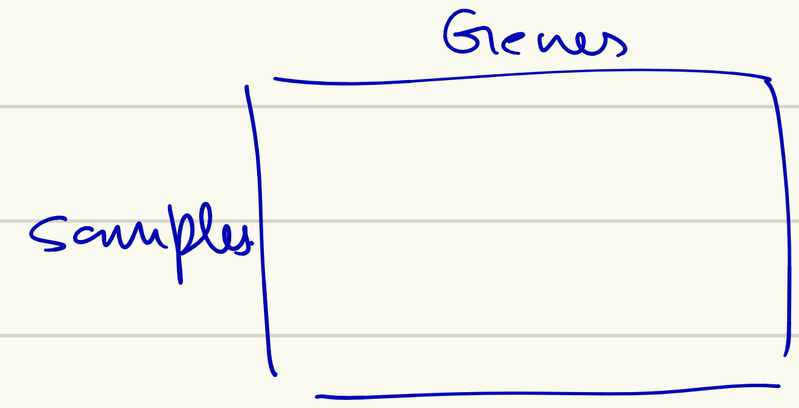
RNA Alignment

- STAR (suffix map)
- HISAT

RSEM

Gene alignment, Fragment alignment,
sampling from latent space.

Transcript quantification:-

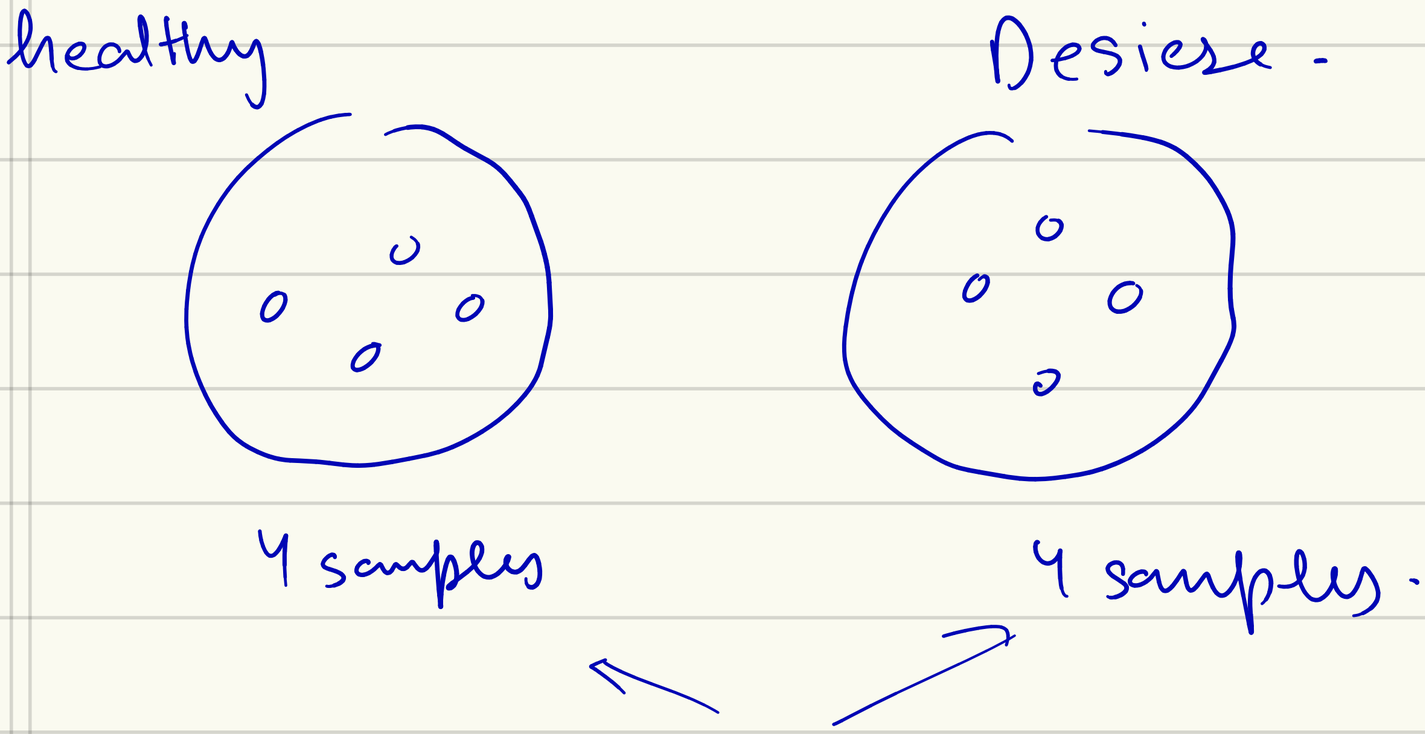


3 Nov

Scribe lecture

Differential Expression.

if we have samples from different conditions.



Biasies:
in RNA

1. Fragment length prefrence
   Some fragments are prefforably sampled.

2. Positional bias.
   when sampling, selecting. there are prefrenses. in positional.
   there's a higher probability of getting good fragments from end of transcript.

3. Sequence based bias.
   primers efficiency of binding to molecules differ.

4. Fragment GC bias.

GC - content of organism.

GC content high. - reads less likely to getting sequenced.

RSEM - already accounts for fragment-length and positional bias.

- specific variables to account them.

Transcript length:

property vs. prob. of getting sequenced.

→ give length → effective length.

→ soft assignment | probabilities change.

Sum of bias terms across transcript.

Formulate and test hypothesis.

→ identify genes which behave differently from one conditions to other.

Two diff. samples.
two diff genes expression. } NULL hypothesis.

Single cell data - No problem. Each cell gives a lot of samples. (replicates)

Single gene - lookout for variations in other genes also.

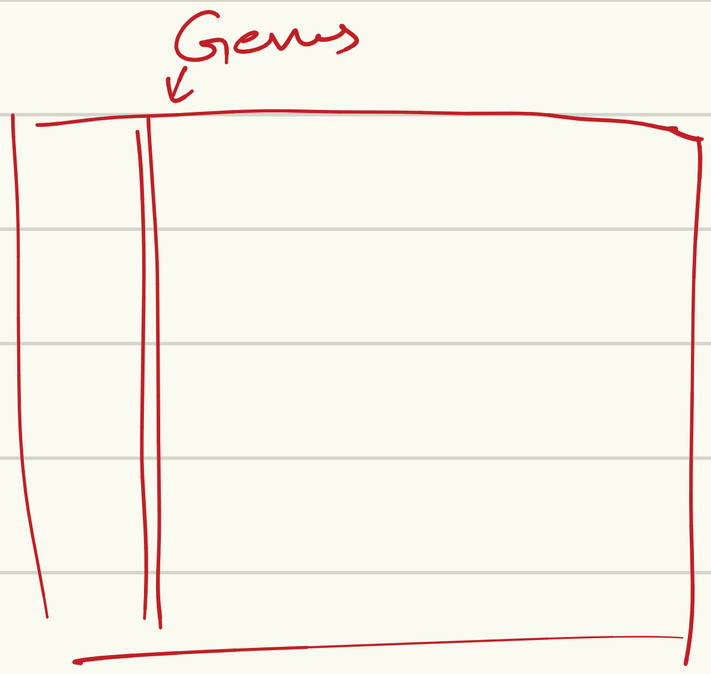
Normalization - (library sized normalization) for DGE

Bulk seq. data: -

Sequencing depth might be different

Normalization Assumption.

different genes may be assigned.



Rank invariance -

-> there are some genes which will not change across samples.

-> House keeping genes.

Quantile Norm.

no difference across diffrent trials.

Scale Factor Norm.

invariant SET Norm.

↳ Scale Fac Norm.

Same distribution

but different scale factor.

↓ eg lib size.

Assumption - total mRNA quant, same for all samples.

Mean Normalization -

$$\hat{y}' = \hat{y} - M^j + M$$

results in comparable datasets.

Variance normalization :-

$$\hat{y}_j^i = \frac{(y_i^j - M^j)\sqrt{v}}{\sqrt{v^j}}$$

τ_i ← eff. no. of trans. per gene (i)

↑

$(\times 10^6)$ = TPM ↩

Not the only way

RPKM ← Comparability b/w diff genes.

↑

$$RPKM(G_1) = \frac{R}{L}\frac{10^6 \cdot 10^3}{M}$$

15 year old

↓

FPKM (Fragments)

Count No. of fragments that are coming directly from the experimental data.

$$FPKM(G_1) = \frac{C}{L}\frac{10^6 10^3}{M}$$

{Set of genes} ← highly expressed in one experimental conditions.

library size same ← { Sample A (500 genes)
                    { Sample B

↑ build statistical model.

problem.
→ Exp. conditions.

Gene

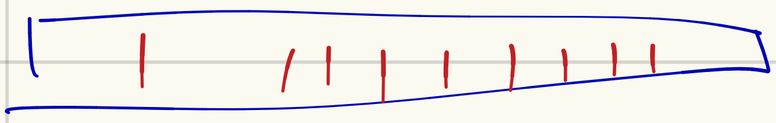
A

20Mill

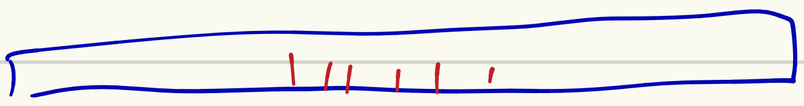
B

20milli

50 genes. ← same → 50 genes.
50 genes ← diff → 0 genes.



100 reads spread.



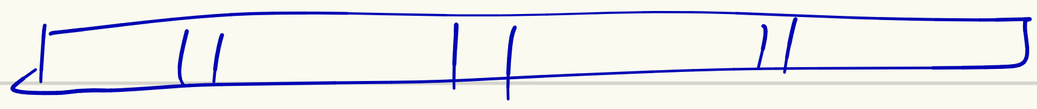
spread across 50 genes.

TRIM mean of M (TMM)

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

$$M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk}/N_{k'}}$$

A_g



Trim of high variability genes. (A_g) and outlier mean value. then normalise on the basis of the remaining gens.

Spiking ← molecules whose exact sequence an strength is known.
(calibration uses)

Differential Expression

↳ for different genes, when the expression levels are statistically significantly different.

→ Statistical tests: performed to determine DE.

Example: If we have ~~~~ genomic data from cells under different conditions (Eg. disease / heat stress)
↳ "How would we, by looking at the expression data, find out what genes are expressing differently?"

→ DE genes could tell us about cells & samples as well. (Eg — classifying diseased cells).

→ DE genes can help us understand the mechanisms behind certain biological processes.

→ DE gene expression could be because of ARTIFACTS
— experimental / technological.

→ Calculate p-values to see if expression data distribution could have been achieved by "accident".

→ Example of A vs B

→ Simpler statistical tests are useful when you have a large number of replicates of the dataset.

→ Alas, ~~for~~ experiments are expensive, and we have 2-3 replicates.

→ Closely spaced here, we want to find out if any variation in read counts is "natural" or "by chance".

Candidates for Read Counts:

↳ Poisson: Mean == Variance == λ.

↳ Underestimates variations
→ Overdispersion (Real data: Variance >> Mean)

Note: In some samples the count value might be 100, in some others it might be 1.

Negative Binomial:

↳ How many losses before 2 ~~successes~~ needs?

↳ better fit for datasets with greater mean expression levels.

$$P(X=k) = {}^{n-1}C_{r-1}\,(1-p)^k p^r$$

→ Mean and Variance can be modeled using a dispersion parameter.

↳ μ ↳ $\boxed{\mu + \alpha\mu^2}$

→ DESeq: $p(j) \rightarrow$ Conditions

$i \rightarrow$ gene

true concentration of treatments for

Here i is an unknown (latent) variable

μ_{ij} → Shot noise

$s_j^2 v_{i,p(j)}$ → Raw variance.

↳ pools information from genes with similar expression levels.

Parameters → Size factors

→ Expression strength parameters.

Dependence of raw variance on expected mean expression.

Simple Assumption:

If gene i is not differentially expressed.

↳ The count of reads from i should be proportional only to the SIZE factors.

⇒ A few DE genes might have an inflated effect.

Common scale: denominator is the estimated size factor.

⇒ v_p → first we calculate sample variations, normalized by size factors.

⇒ for a small number of replicates, we use a local regression for the smooth dependence function.

NFPBT (No DE genes)

→ H_0: gene expression is constant.

p-value → Probability of our data being generated under our null hypothesis.

→ Depends strongly on our underlying model/distribution

In our case, the test statistic is the count in each condition.

Log. fold change?

Multiple test p-value problem:

↳ DESeq2 Edger DIMMA

Volcano Plot Explanation:

MA plot

↳ Log fold change vs Mean Expression

Pathway Analysis Methods:

↳ GSEA DAVID

Cuffdiff → Gene Ontology area