Model based clustering in RNA-seq and Phylogenetic data

Kushal K Dey

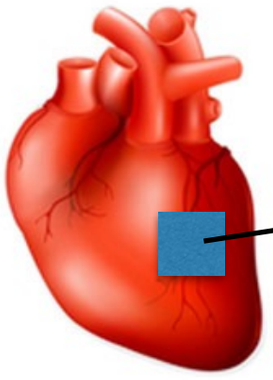
Advisor:
Matthew Stephens

In collaboration with :
Alex White
Trevor Price
Dhananjai Mohan

HG Work in Progress Seminar,
11.18.2015

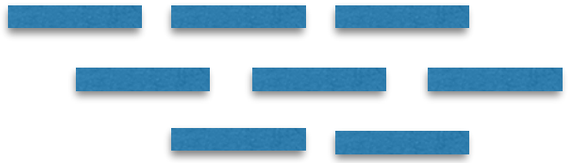
RNA-seq Data Modeling

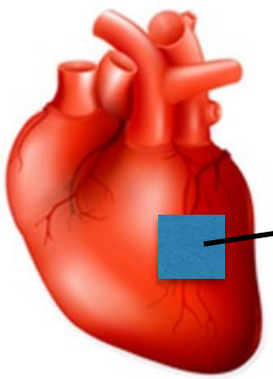
RNA- seq: Overview



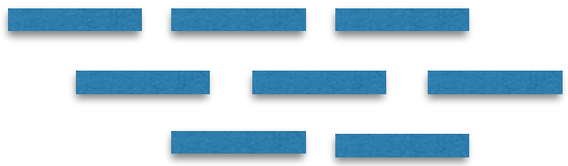
Take a sample from a tissue (heart here)

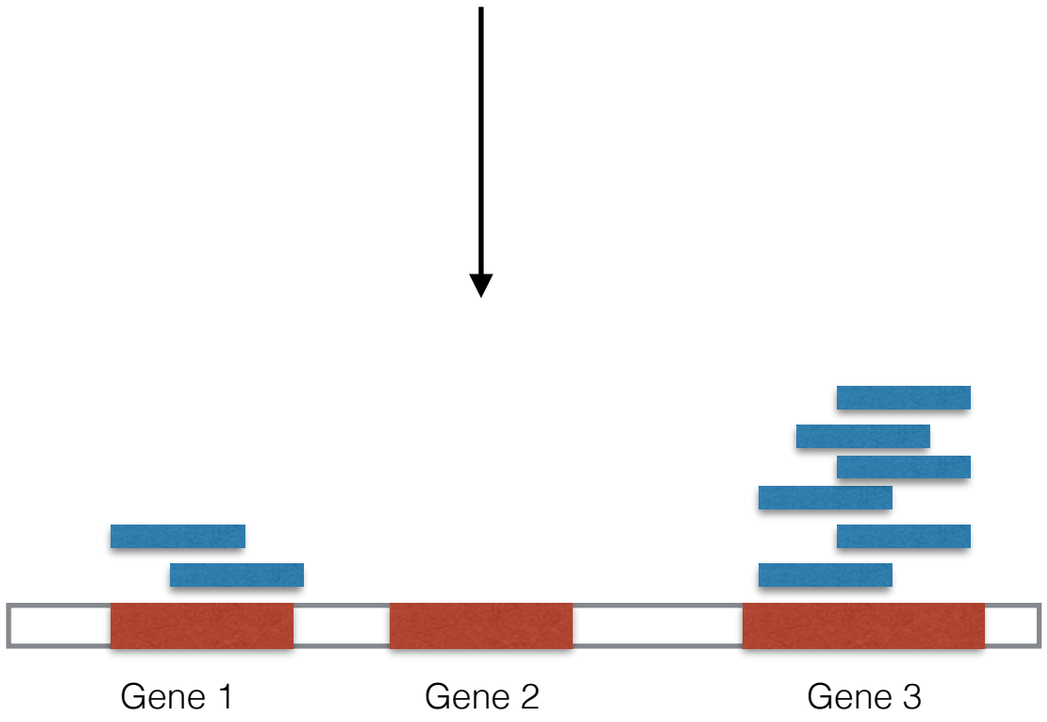
Reads sequenced

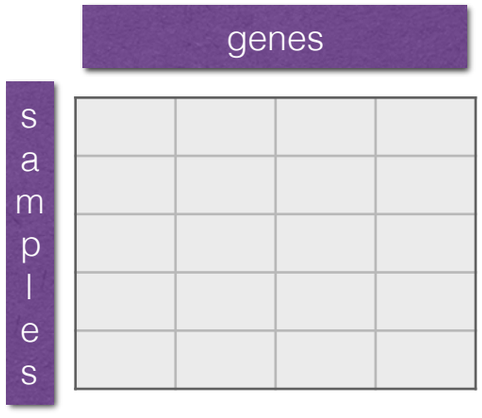


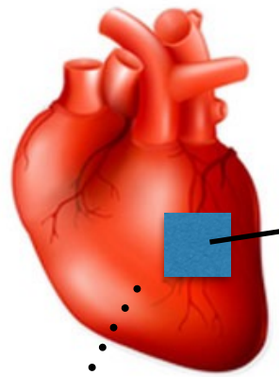


Reads sequenced

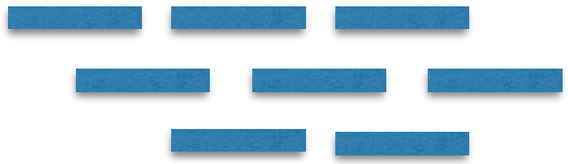


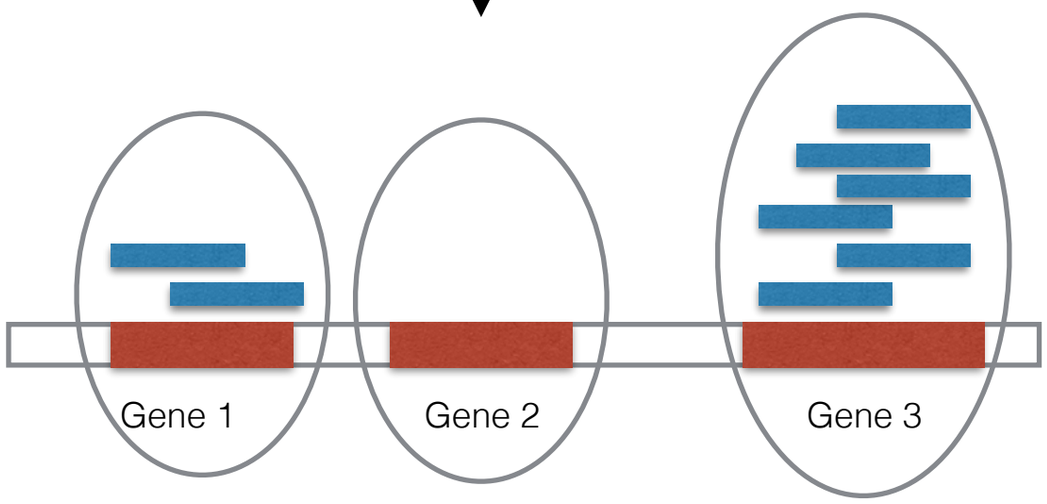






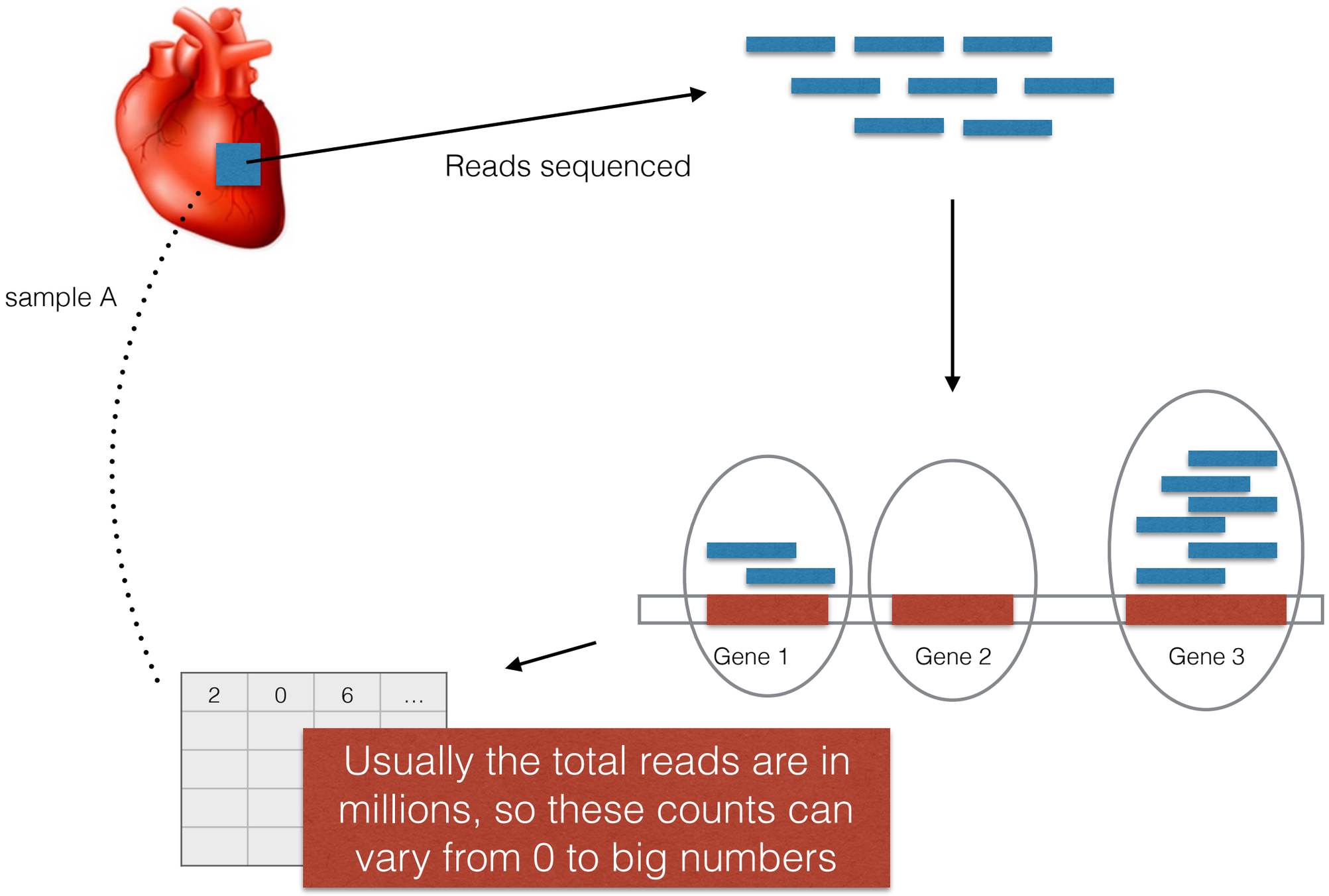
Reads sequenced



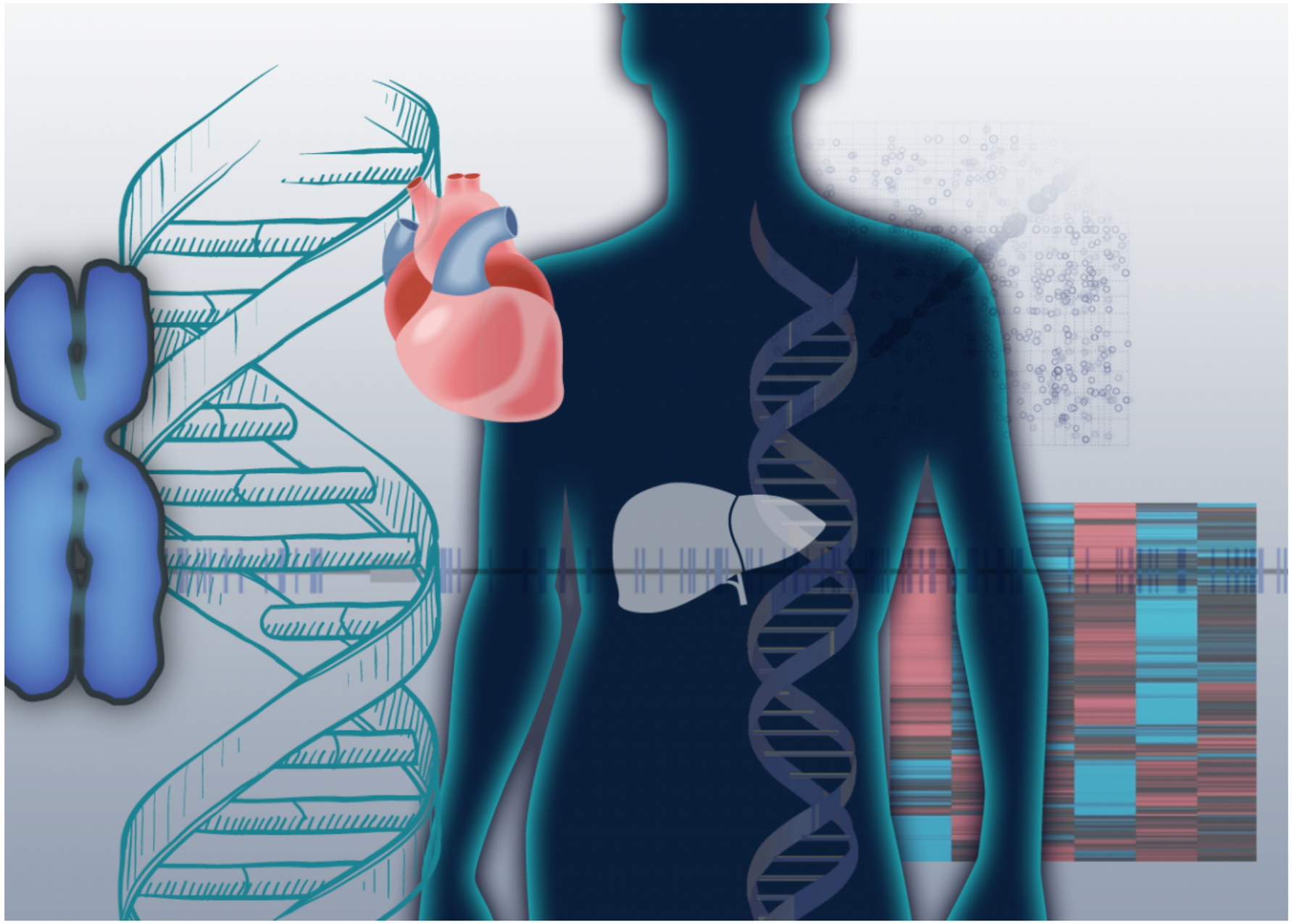


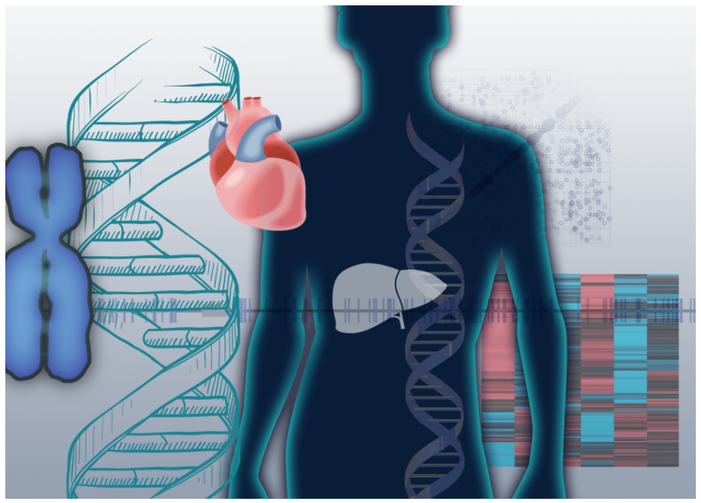
sample A

2	0	6	…

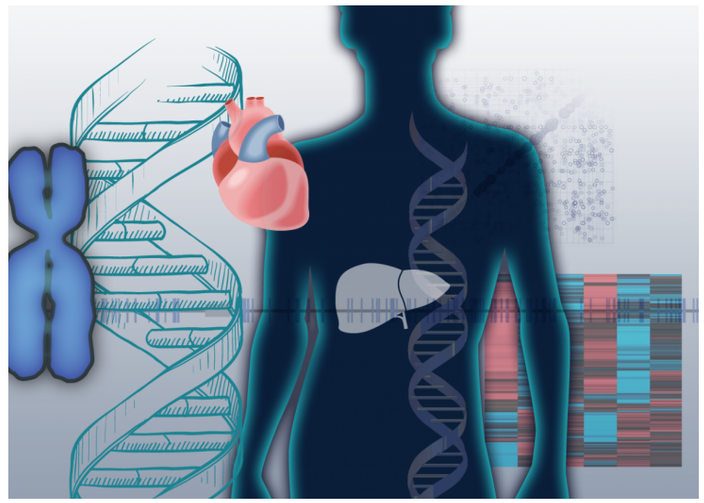

Reads sequenced
sample A
Gene 1
Gene 2
Gene 3
2
0
6
…
Usually the total reads are in millions, so these counts can vary from 0 to big numbers

Genotype Tissue Expression (GTEX) Project



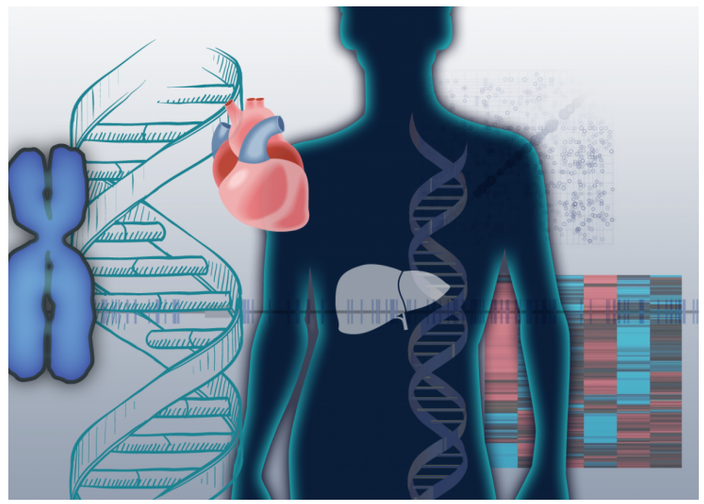


has data on 8555 tissue samples taken from 900 donors (Phase 6) for around 56,000 genes.



has data on 8555 tissue samples taken from 900 donors
(Phase 6) for around 56,000 genes.

The tissue samples came from a total of 53 different tissues/ sub-tissues
(e.g. - brain sub tissues, whole blood etc)



has data on 8555 tissue samples taken from 900 donors
(Phase 6) for around 56,000 genes.

The tissue samples came from a total of 53 different tissues/ sub-tissues
(e.g. - brain sub tissues, whole blood etc)

We filtered out around 16,069 genes that passed QC tests.

Model Overview

c_{ng} : counts for tissue sample *n* and gene *g*

$$c_{ng} \sim Mult\left(\sum_{k=1}^{K} \omega_{nk}\theta_{kg}\right)$$

$$\sum_{k=1}^{K} \omega_{nk} = 1 \;\forall\; n \qquad \sum_{g=1}^{G} \theta_{kg} = 1 \;\forall\; k$$

ω_{n*} : The membership proportion vector of sample *n* in the K clusters

θ_{k*} : relative gene expression profile for cluster *k*

This model is similar to *admixture model* in population genetics and *topic model* in natural language processing

Model Overview

c_{ng} : counts for tissue sample *n* and gene *g*

$$c_{ng} \sim Mult\left(\sum_{k=1}^{K} \omega_{nk}\theta_{kg}\right)$$

$$\sum_{k=1}^{K} \omega_{nk} = 1 \;\forall\; n \qquad \sum_{g=1}^{G} \theta_{kg} = 1 \;\forall\; k$$

ω_{n*} : The membership proportion vector of sample *n* in the K clusters

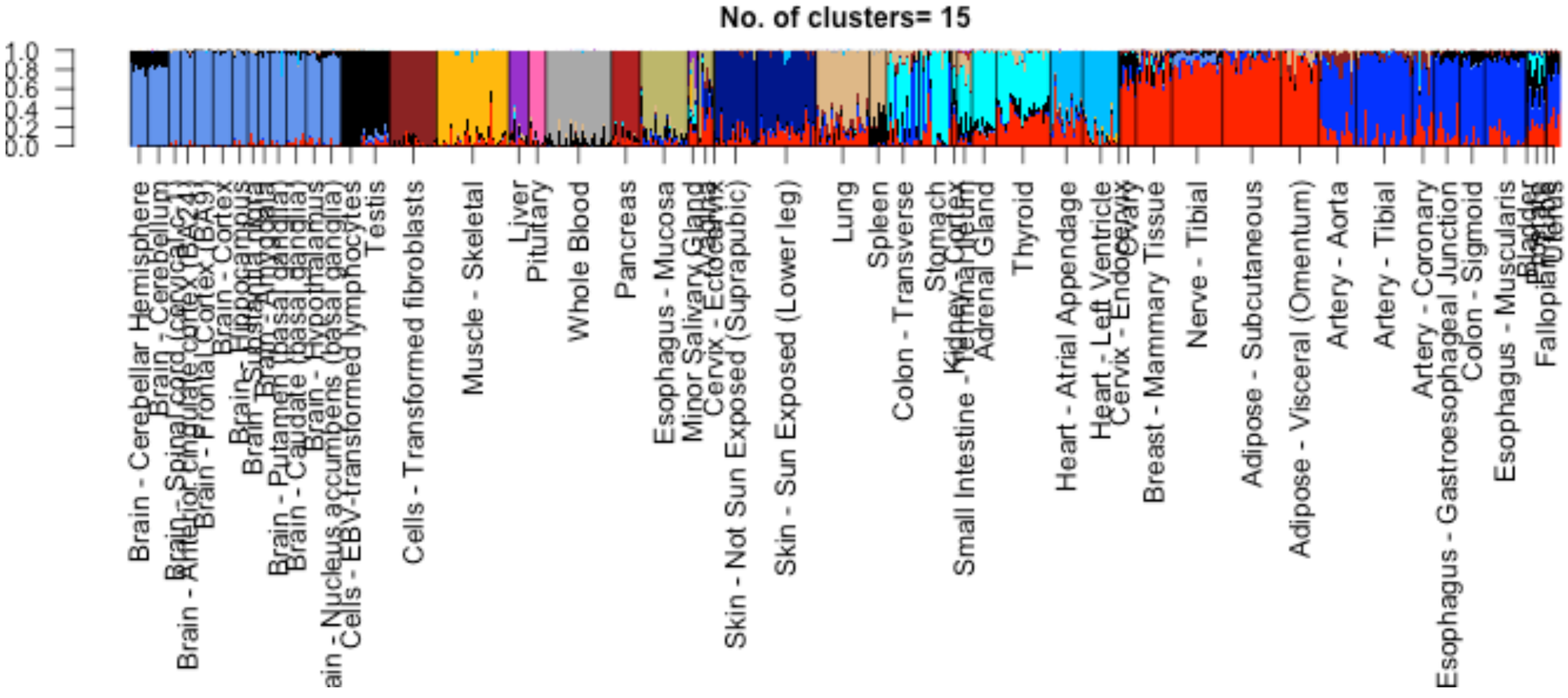
θ_{k*} : relative gene expression profile for cluster *k*

We use the R package **maptpx** (due to Matt Taddy, UChicago Booth School) which uses a EM type algorithm to estimate model parameters

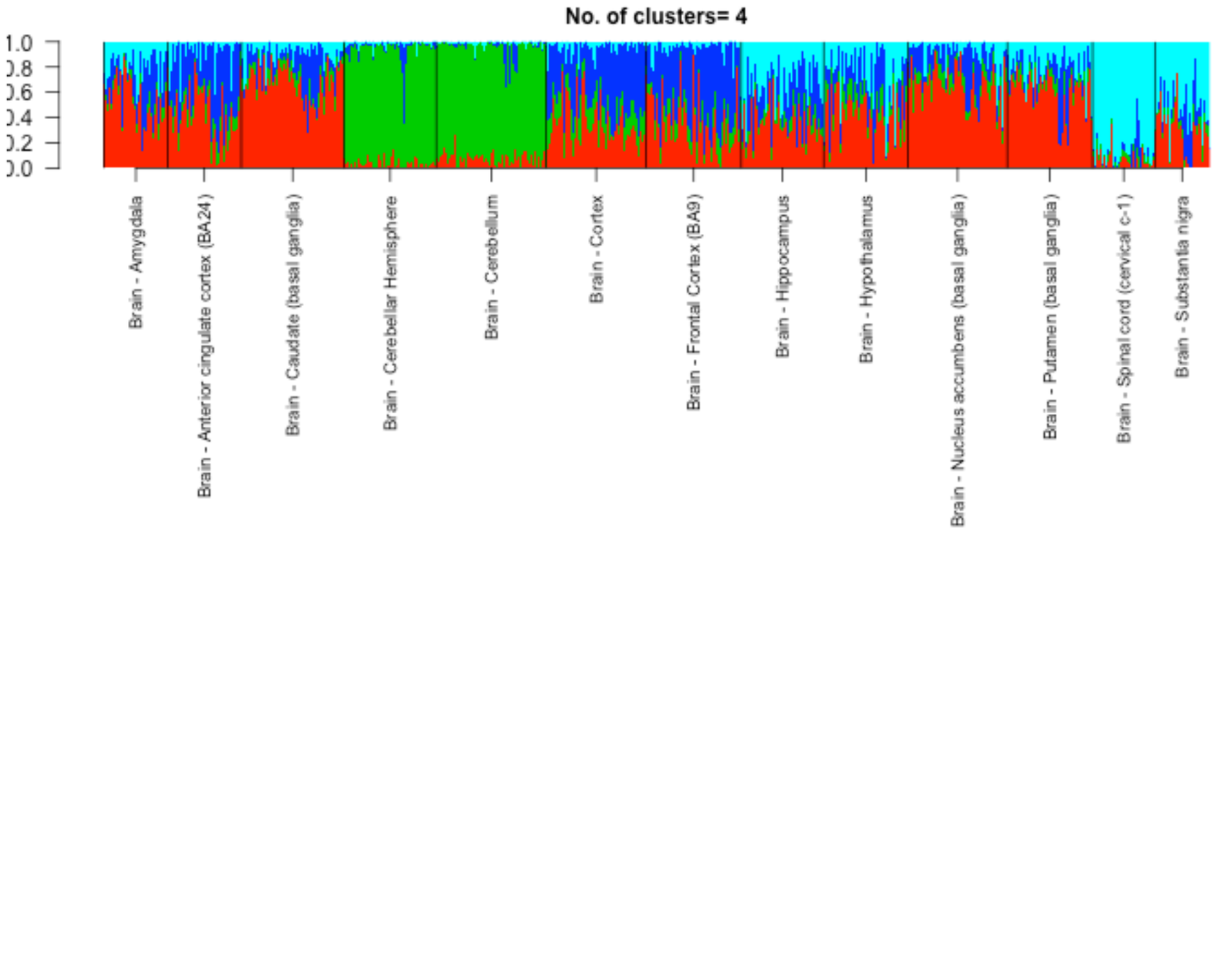
Cluster Post-processing

Visualization: Stacked bar chart plot of the ω_{n*}

GTEx Tissue Admixture plot- All tissues



GTEx Tissue Admixture plot- Brain tissues



Cluster Post-processing

Visualization: Stacked bar chart plot of the ω_{n*}

Gene Annotations : For each cluster *k*, derive a list of genes that have markedly higher or lower relative expression profile compared to other clusters.

Gray cluster (Whole Blood): driving genes

ENSG00000244734 hemoglobin, beta	mutant beta globin causes sickle cell anemia, absence of beta chain/ reduction in beta globin leads to thalassemia.
ENSG00000188536 hemoglobin, alpha 2	deletion of alpha genes may lead to alpha thalassemia.
ENSG00000206172 hemoglobin, alpha 1	deletion of alpha genes may lead to alpha thalassemia.

Brown cluster (Pancreas): driving genes

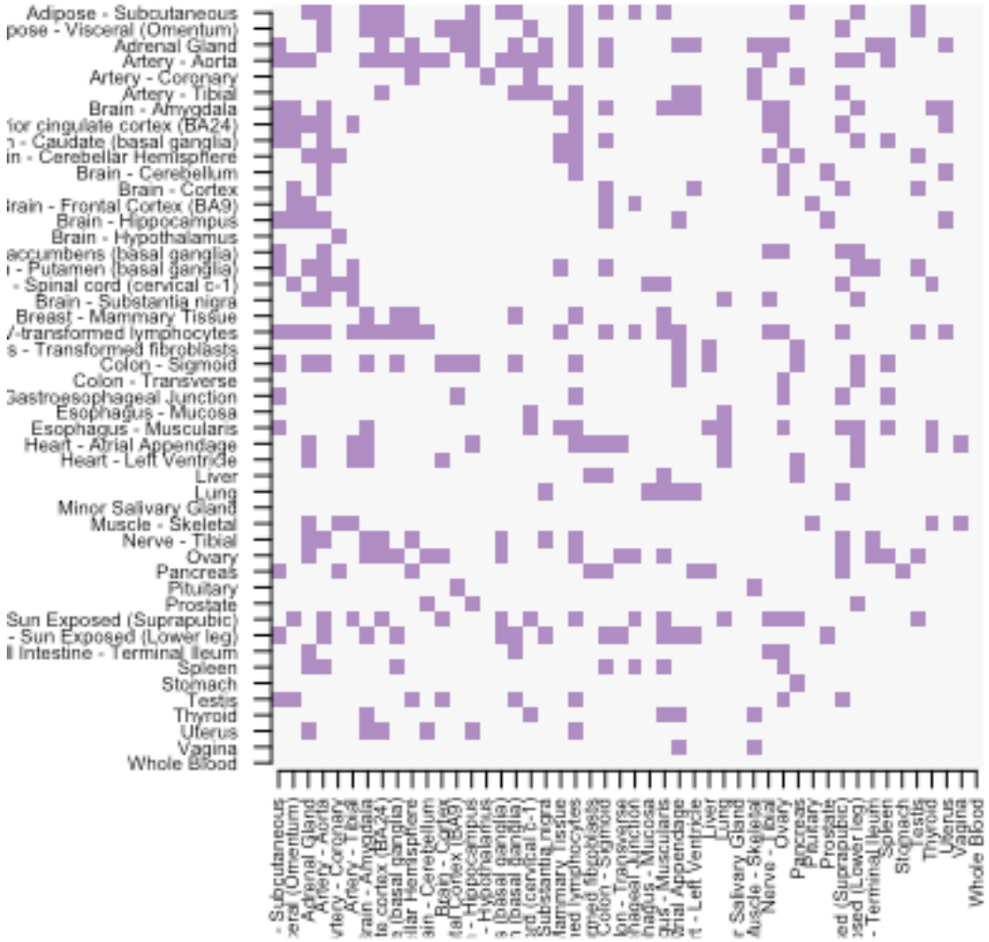
ENSG00000204983 protease, serine 1	secreted by pancreas, associated with pancreatitis.
ENSG00000091704 carboxypeptidase A1	secreted by pancreas, linked to pancreatitis and pancreatic cancer
ENSG00000175535 pancreatic lipase	encodes a carboxyl esterase that hydrolyzes insoluble, emulsified triglycerides, and essential for the efficient digestion of dietary fats. This gene is expressed mainly in the pancreas.

Is this model based approach recommended over distance based approaches (for instance, hierarchical clustering)?

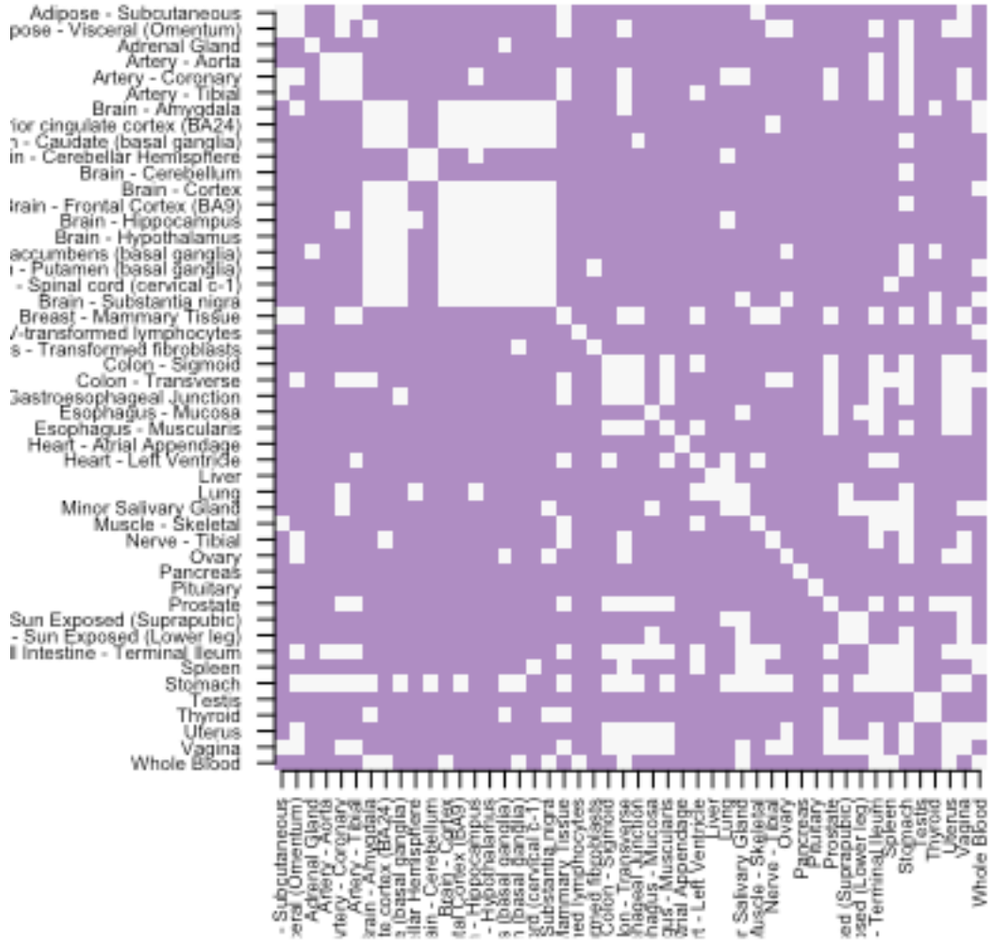
Which method separates tissues better?

separates ■ ☐ does not separate

Hierarchical



Our model

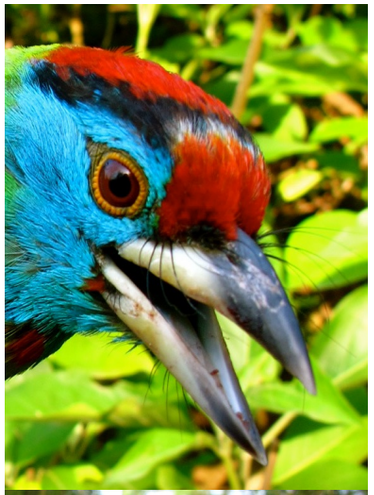


Road Ahead

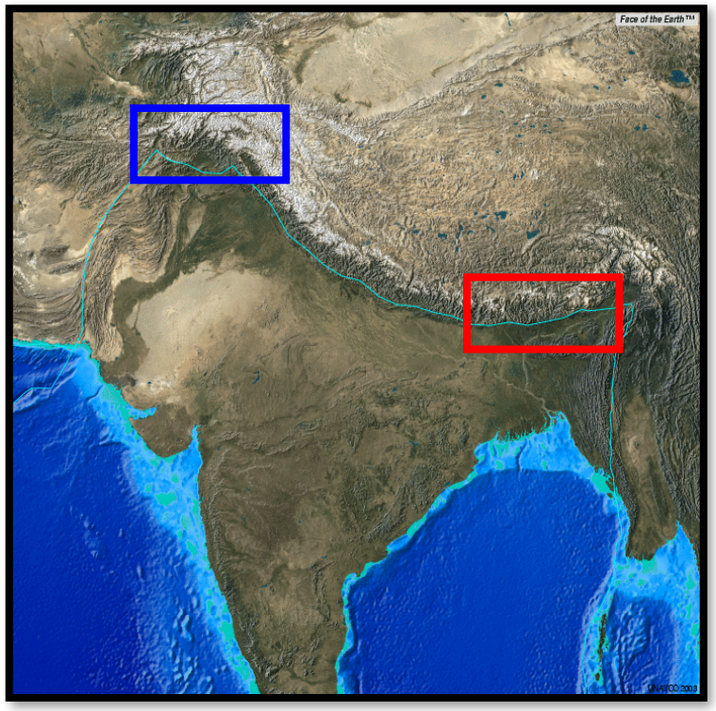
1. Address the issue of how to make the clusters more representative of cell types.

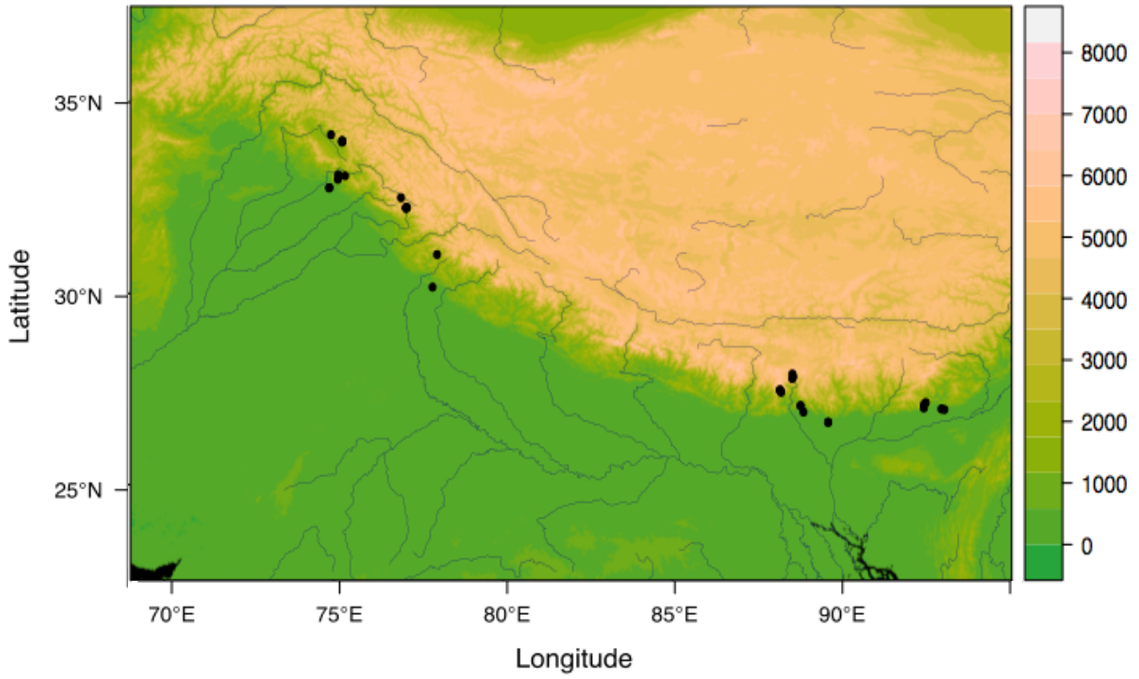
2. Variable selection or gene selection preprocessing step to eliminate genes not driving clustering.

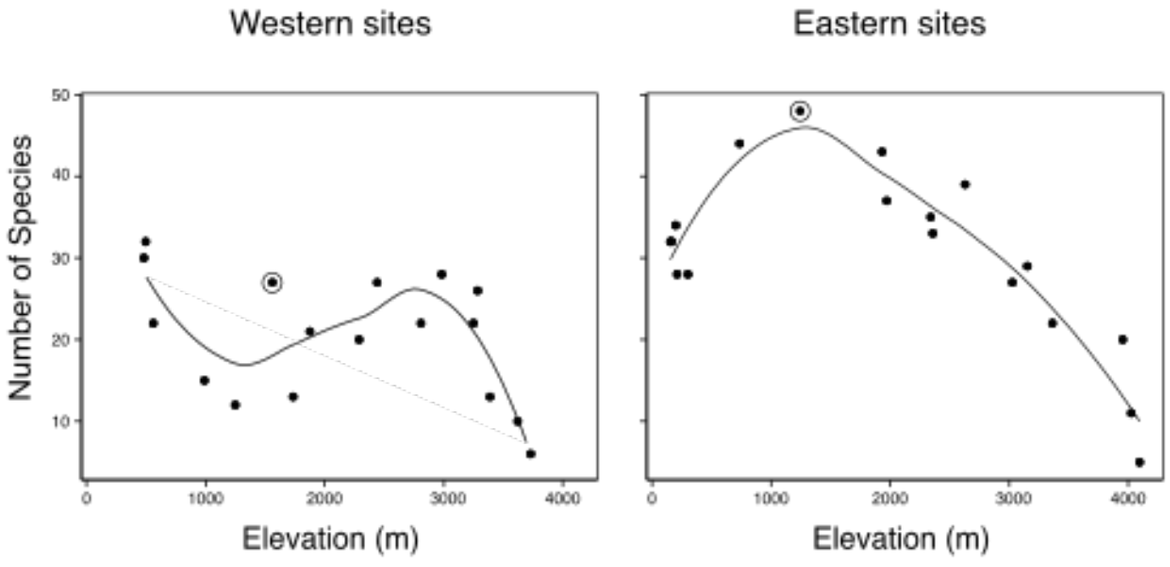
3. Pathway analysis of the cluster expression profiles of genes.

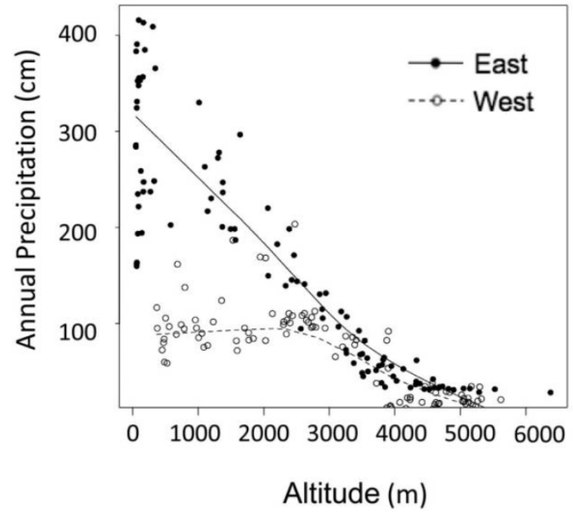


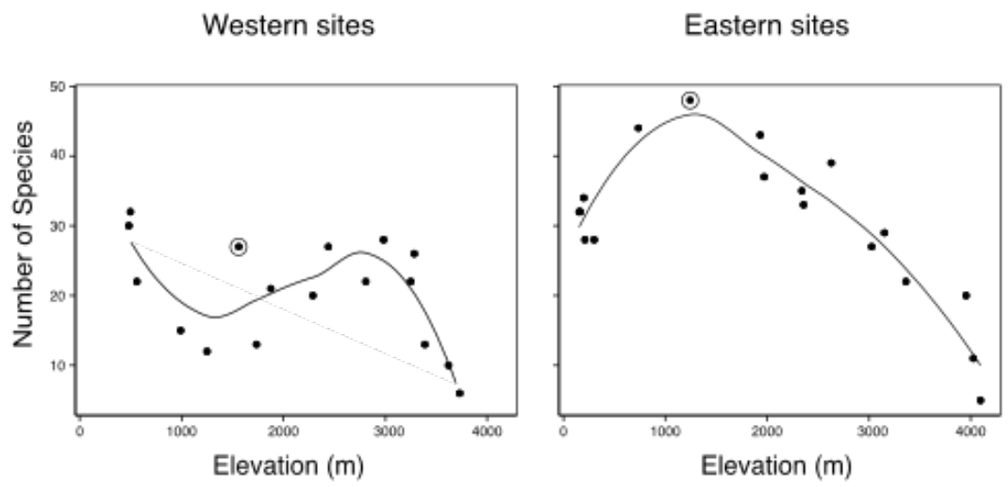

Phylogenetic Data Modeling

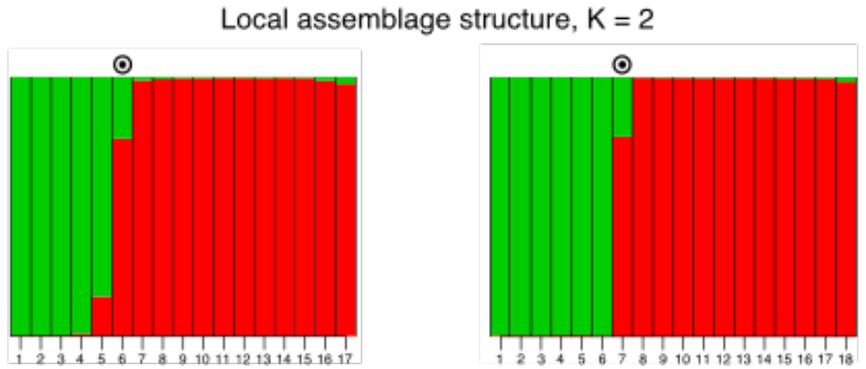

Face of the Earth™

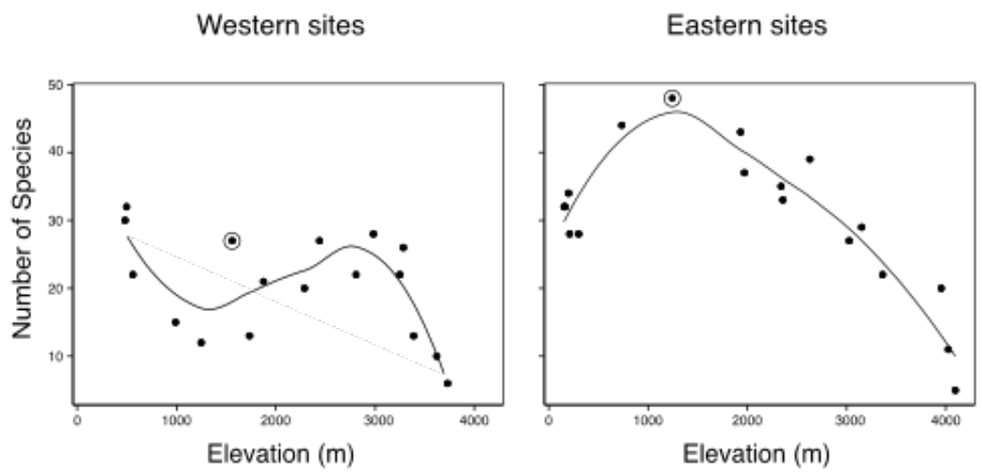

35°N
30°N
25°N
Latitude
70°E
75°E
80°E
85°E
90°E
Longitude
8000
7000
6000
5000
4000
3000
2000
1000
0

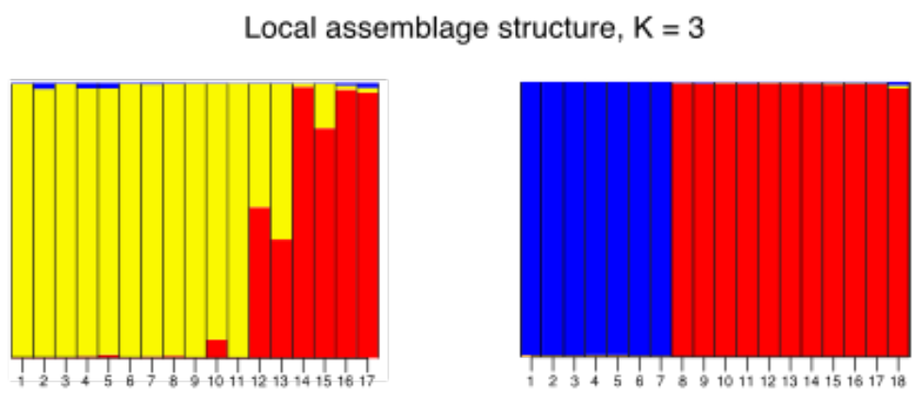

Western sites
Eastern sites
Number of Species
Elevation (m)
Elevation (m)


East
West
Annual Precipitation (cm)
Altitude (m)

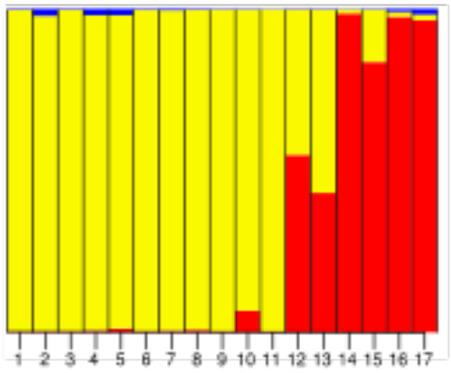

Western sites
Eastern sites
Number of Species
Elevation (m)
Elevation (m)

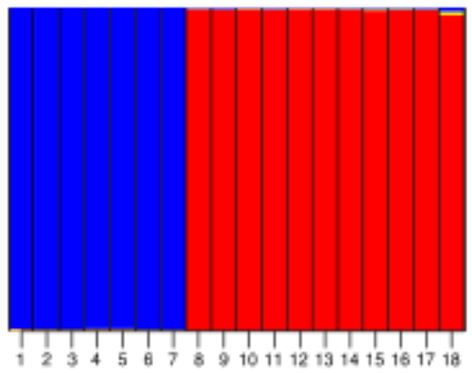

Local assemblage structure, K = 2

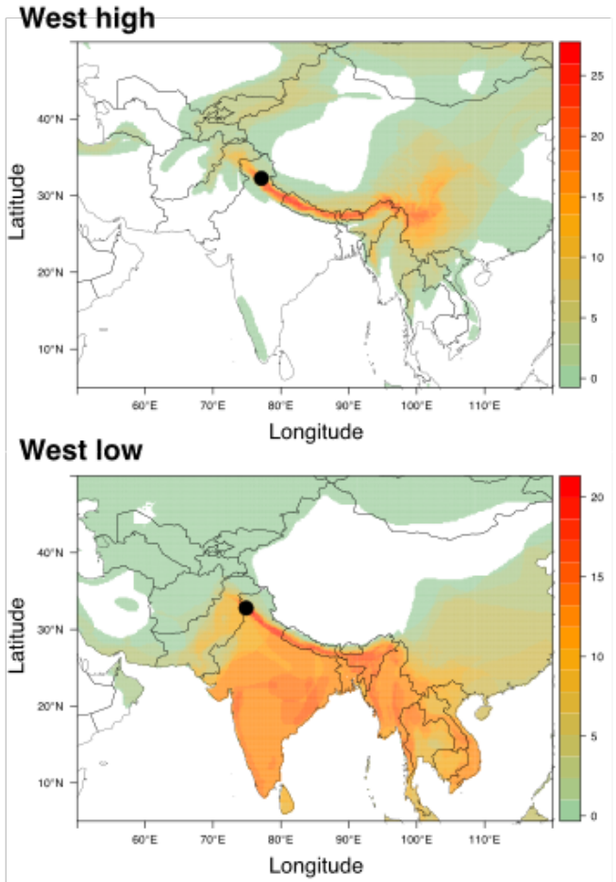

Western sites
Eastern sites
Number of Species
Elevation (m)
Elevation (m)

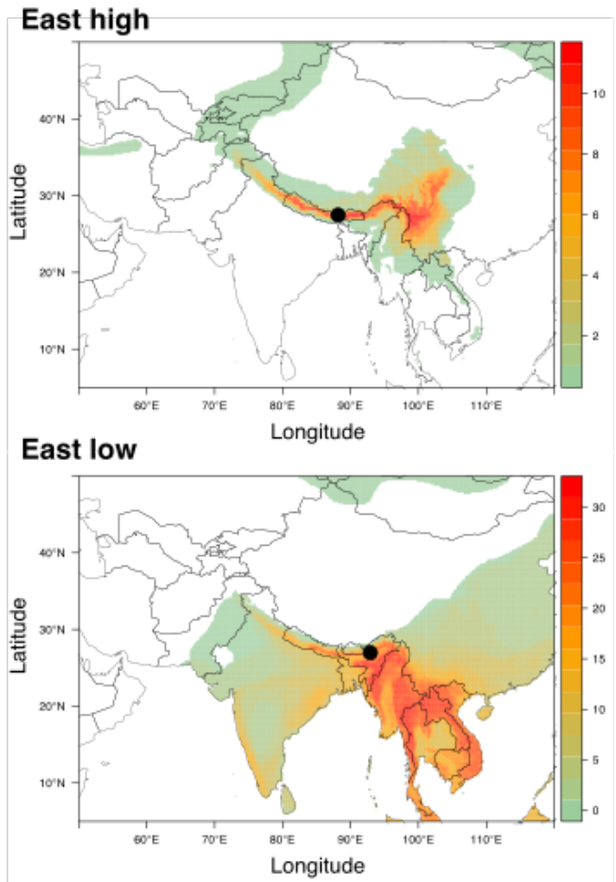

Local assemblage structure, K = 3

Local assemblage structure, K = 3









Road Ahead

1. Check if we can incorporate bird metadata, for example bird body mass, as additional information to drive the clusters.

2. Pull in other diversity measures like phylogenetic beta diversity etc to drive the clusters

Acknowledgements

Matthew Stephens
Alex White
Trevor Price
Dhananjai Mohan

Gao Wang
Sarah Urbut
Raman Shah