

Model based clustering in RNA-seq and Phylogenetic data

Kushal K Dey

Advisor:
Matthew Stephens



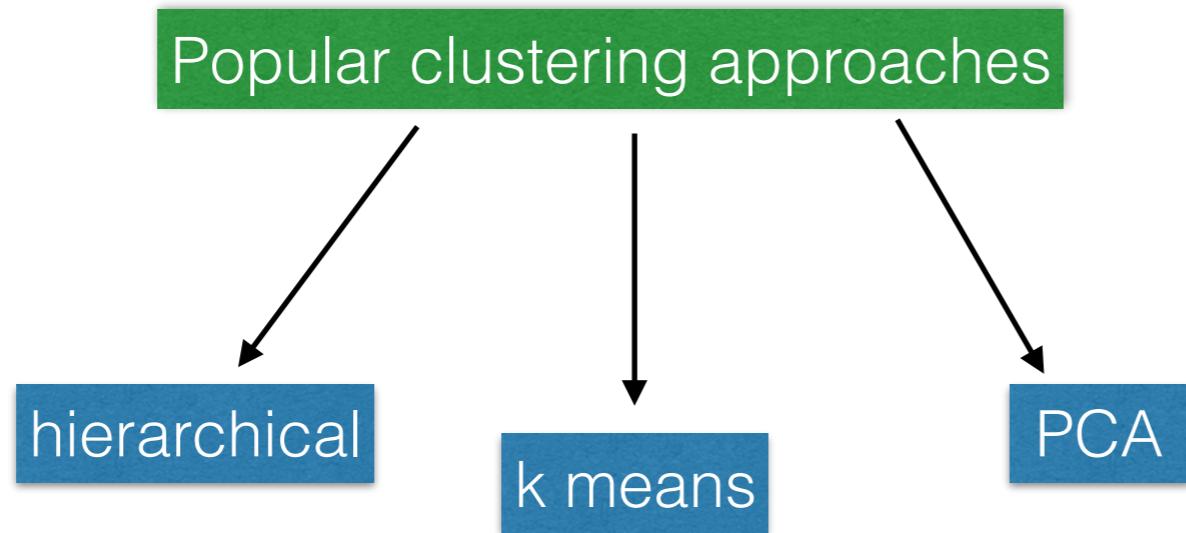
In collaboration with :
Alex White
Trevor Price
Dhananjai Mohan



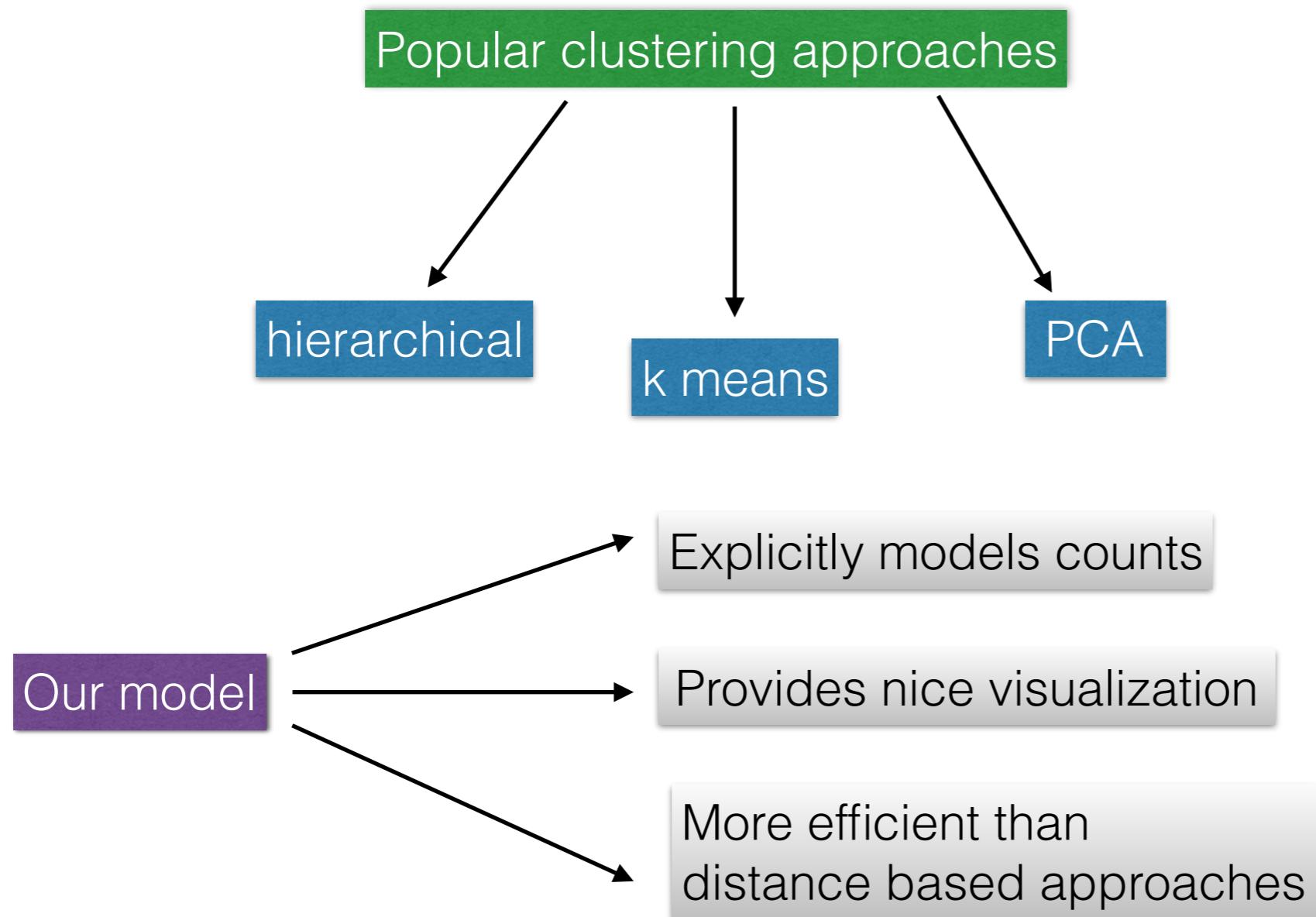
*HG Work in Progress Seminar,
11.18.2015*

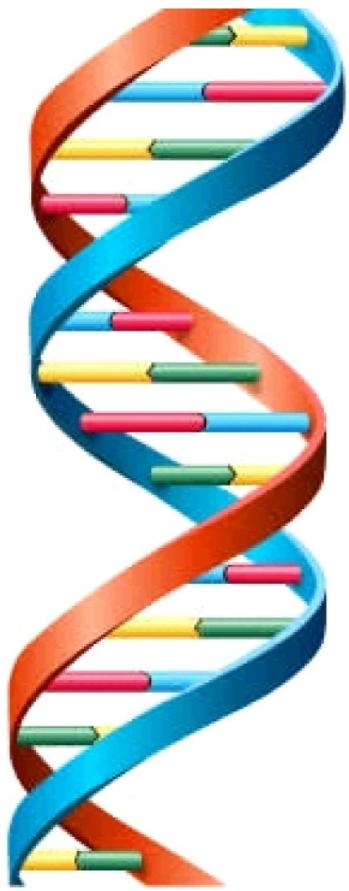
Focus on data that are discrete counts:

Focus on data that are discrete counts:



Focus on data that are discrete counts:

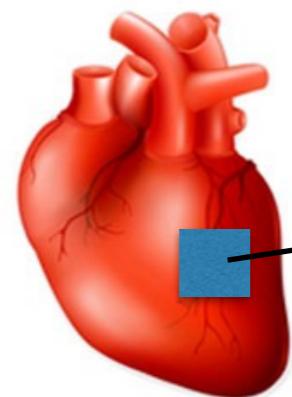




A
T
C
G

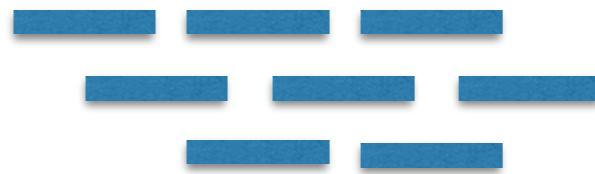
RNA-seq Data Modeling

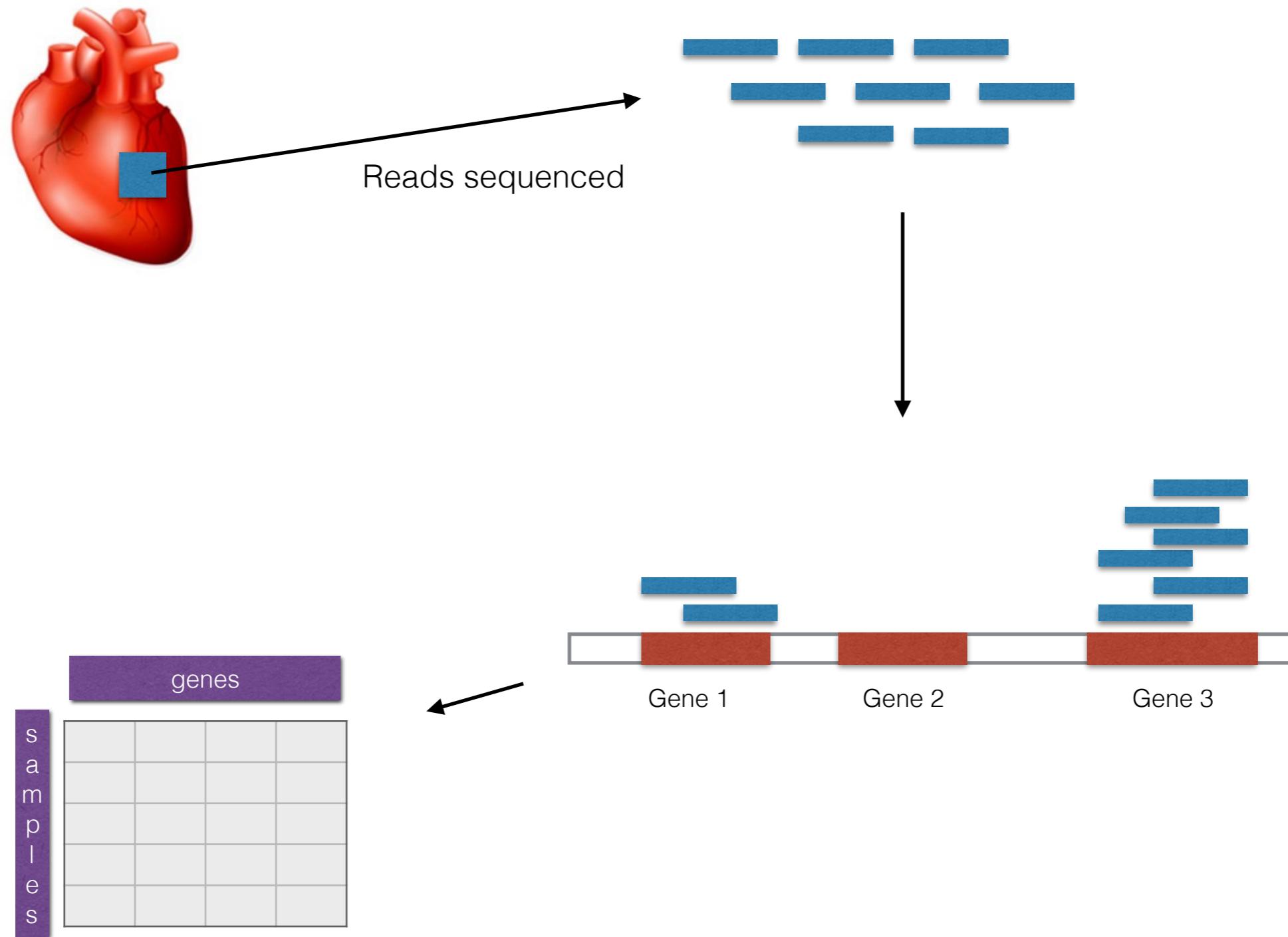
RNA- seq: Overview

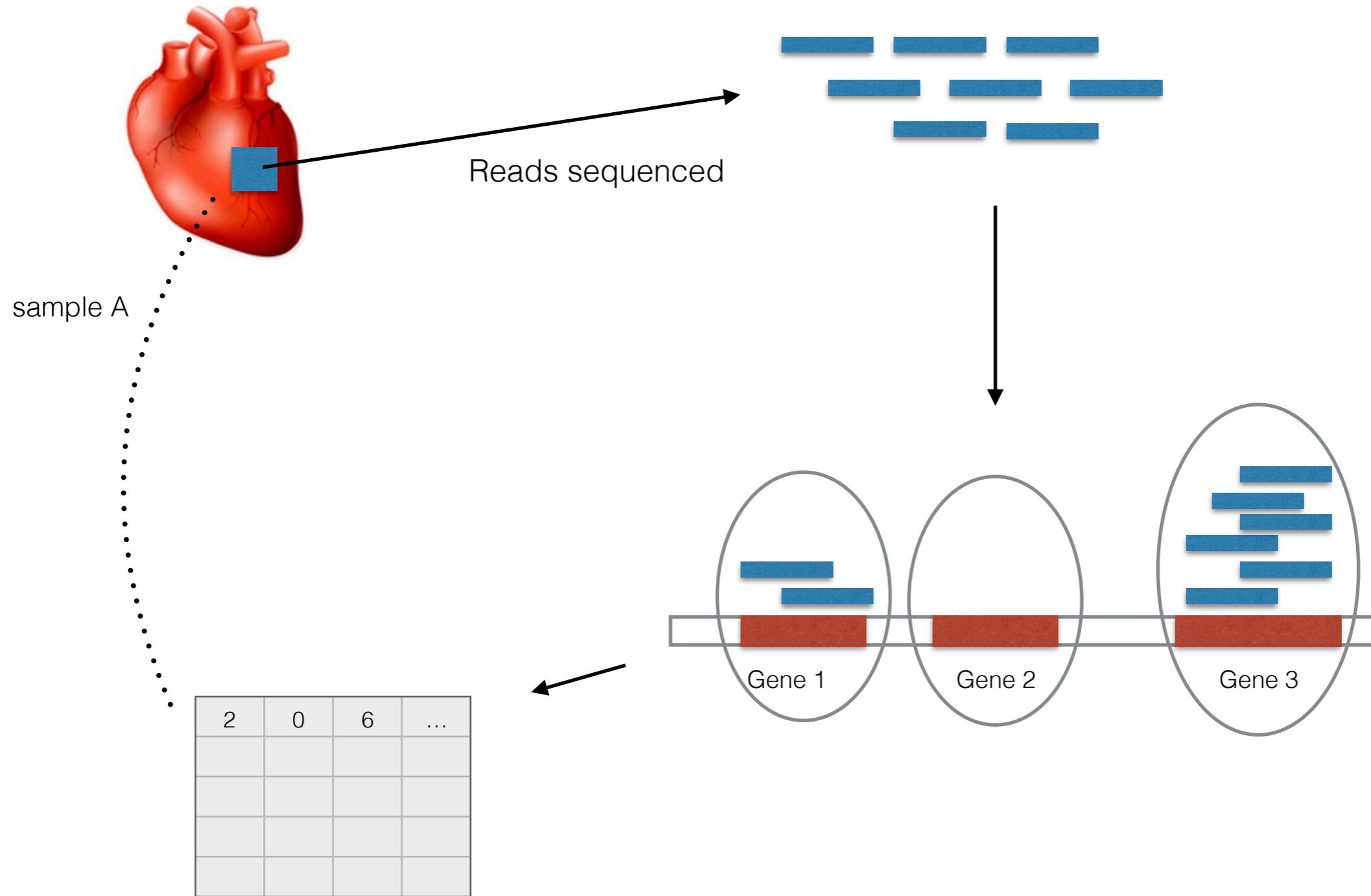


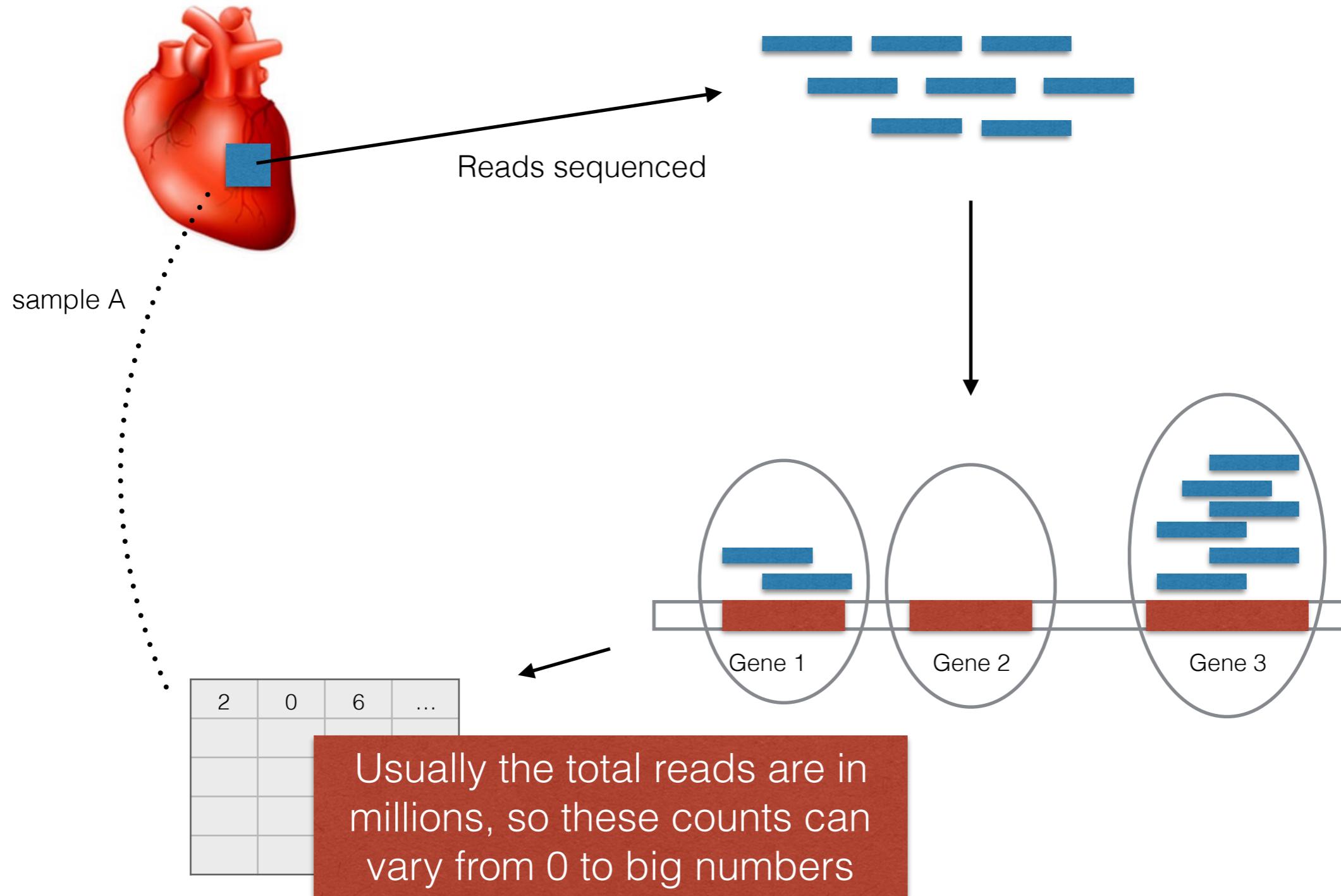
Reads sequenced

Take a sample from a
tissue (heart here)

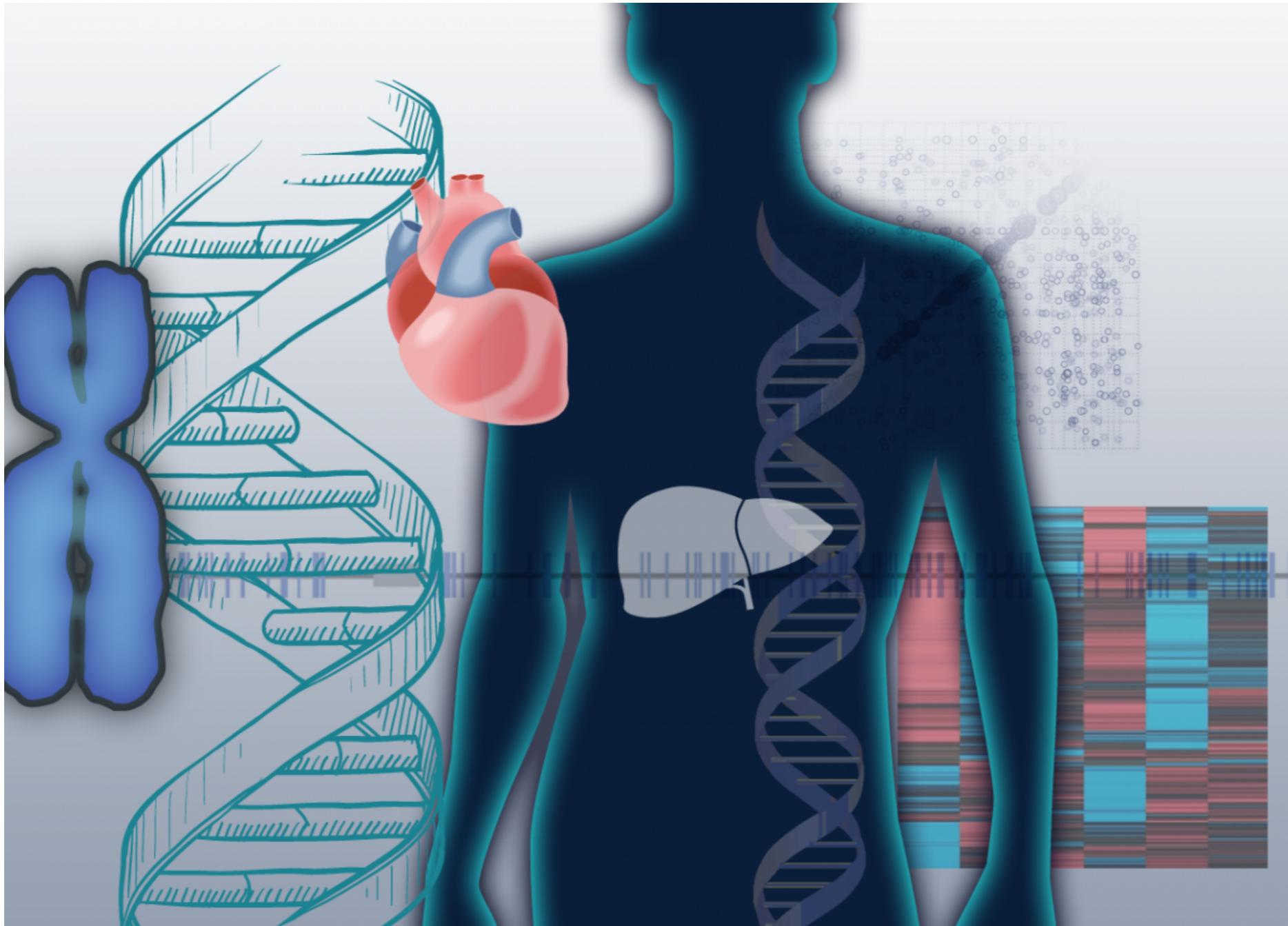


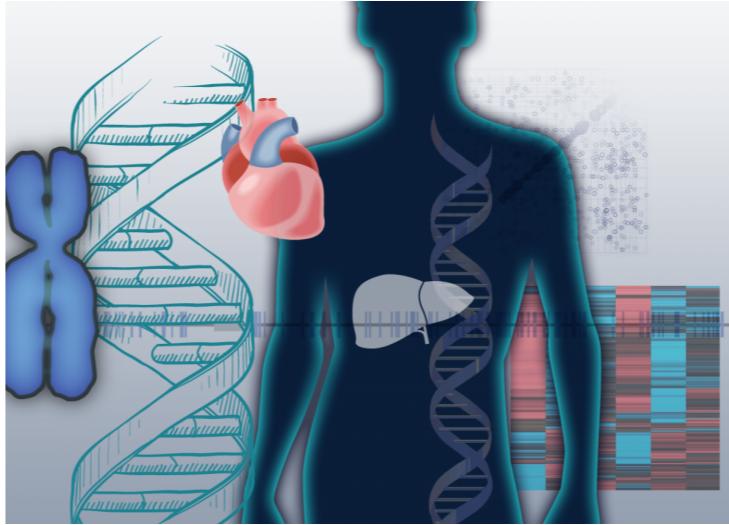






Genotype Tissue Expression (GTEx) Project





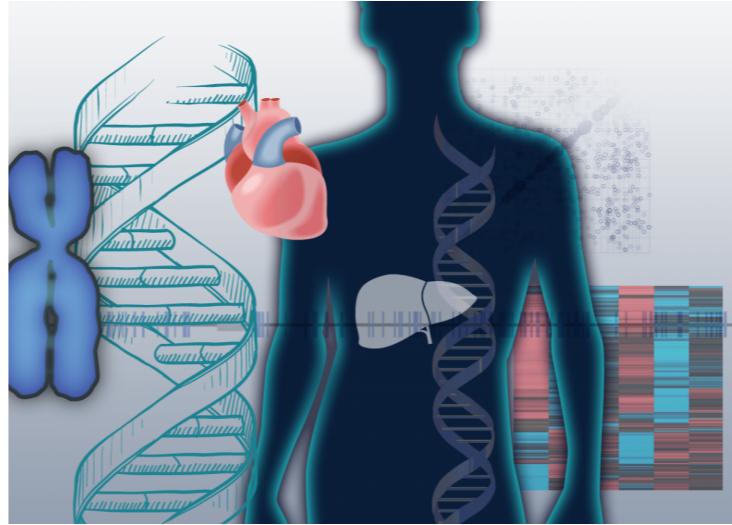
has data on 8555 tissue samples taken from 900 donors
(Phase 6) for around 56,000 genes.



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.



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

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

Model Overview

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

$$(c_{n1}, c_{n2}, \dots, c_{nG}) \sim Mult(c_n, p_{n1}, p_{n2}, \dots, p_{nG})$$

$$p_{ng} = \sum_{k=1}^K \omega_{nk} \theta_{kg}$$

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

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

$$(c_{n1}, c_{n2}, \dots, c_{nG}) \sim Mult(c_n, p_{n1}, p_{n2}, \dots, p_{nG})$$

$$p_{ng} = \sum_{k=1}^K \omega_{nk} \theta_{kg}$$

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

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

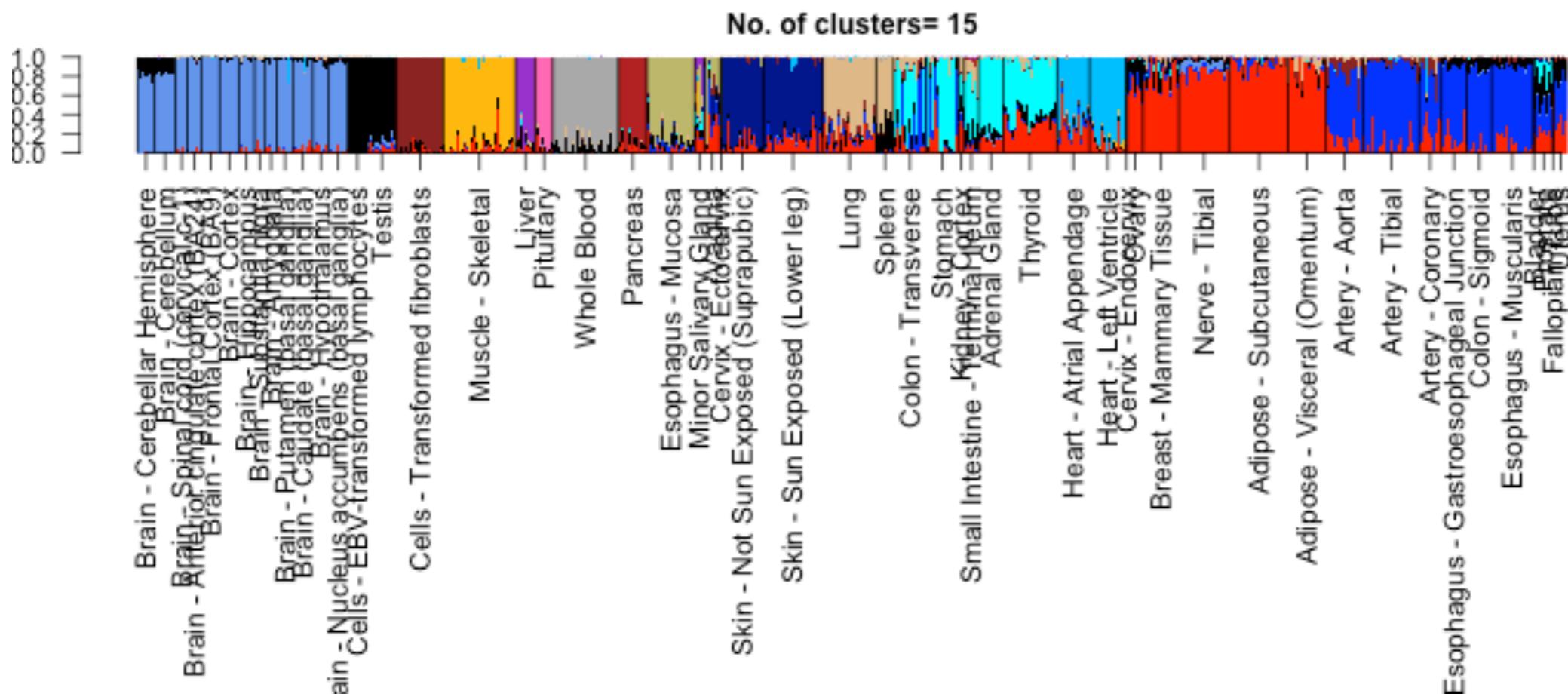
θ_{k*} : relative gene expression profile
for cluster k

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

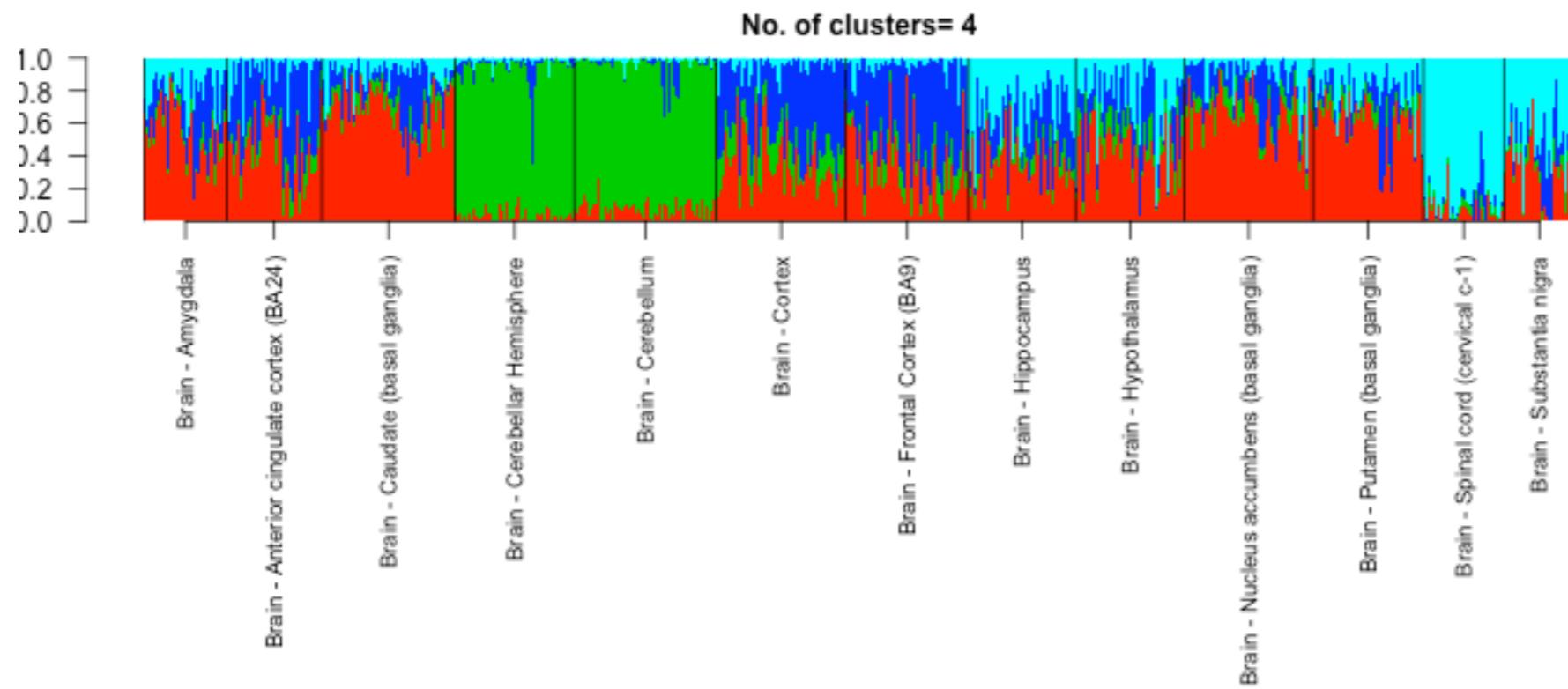
Cluster Post-processing

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 mutant beta globin causes sickle cell anemia, absence of
hemoglobin, beta beta chain/ reduction in beta globin leads to thalassemia.

ENSG00000188536 deletion of alpha genes may lead to alpha
hemoglobin, alpha 2 thalassemia.

ENSG00000206172 deletion of alpha genes may lead to alpha
hemoglobin, alpha 1 thalassemia.

Brown cluster (Pancreas): driving genes

ENSG00000204983 secreted by pancreas, associated with
protease, serine 1 pancreatitis.

ENSG00000091704 secreted by pancreas, linked to pancreati-
carboxypeptidase A1 tis and pancreatic cancer

ENSG00000175535 encodes a carboxyl esterase that hydrolyzes insoluble,
pancreatic lipase emulsified triglycerides, and essential for the efficient di-
gestion 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)?

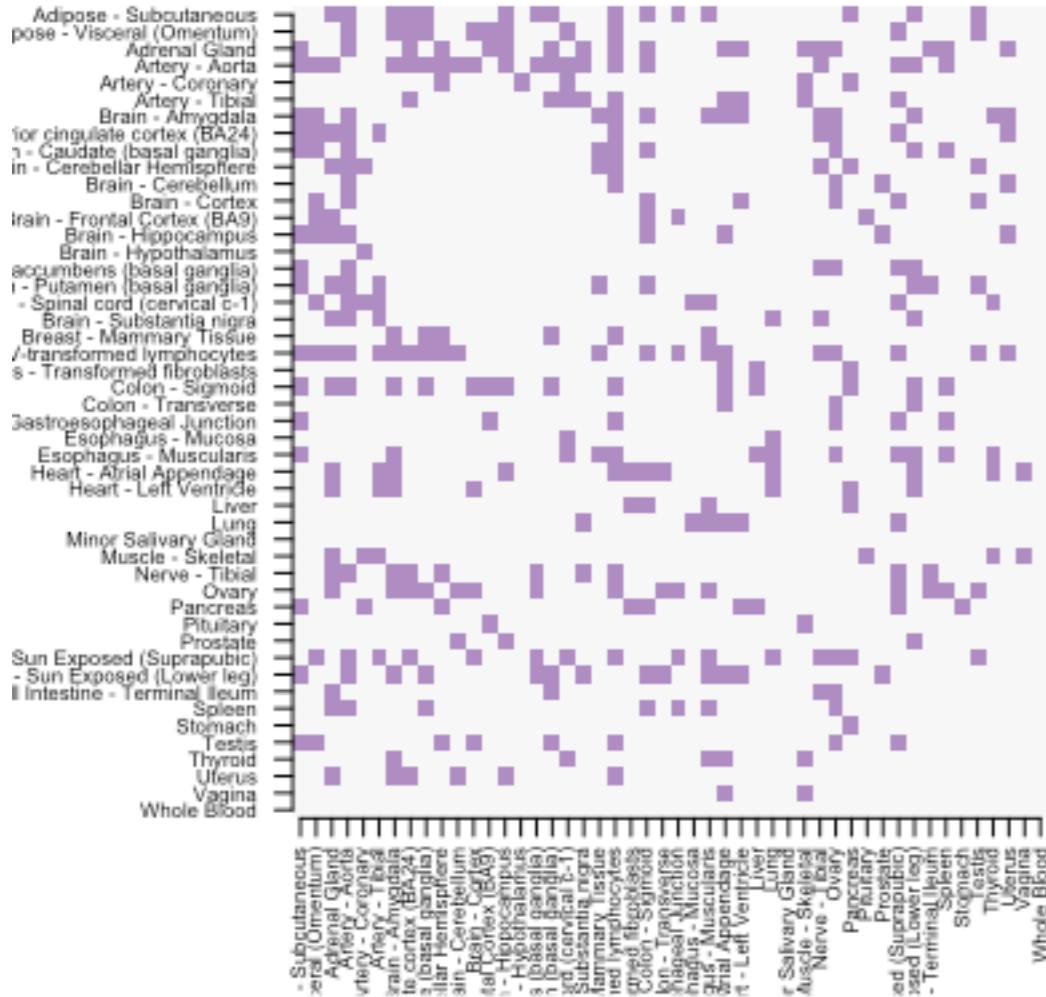
Which method separates tissues better?

separates

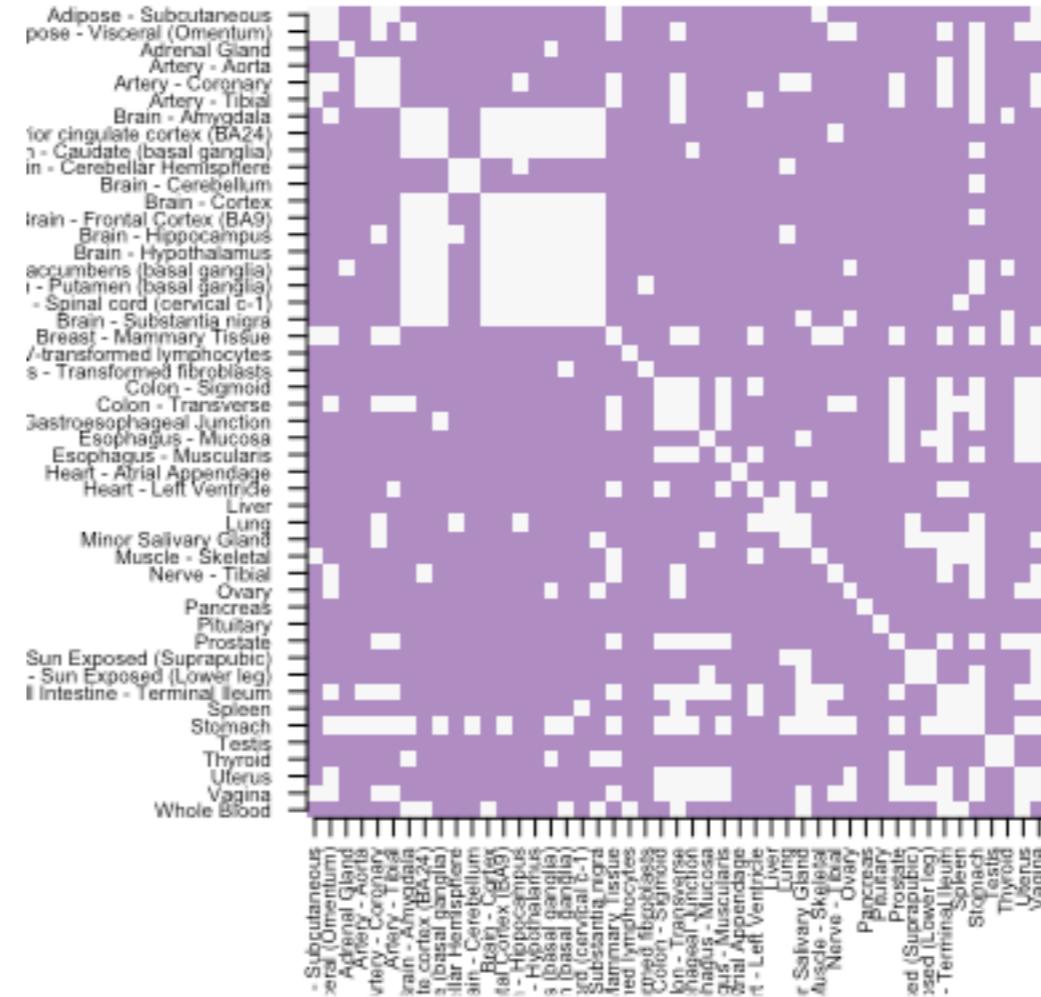


does not separate

Hierarchical



Our model

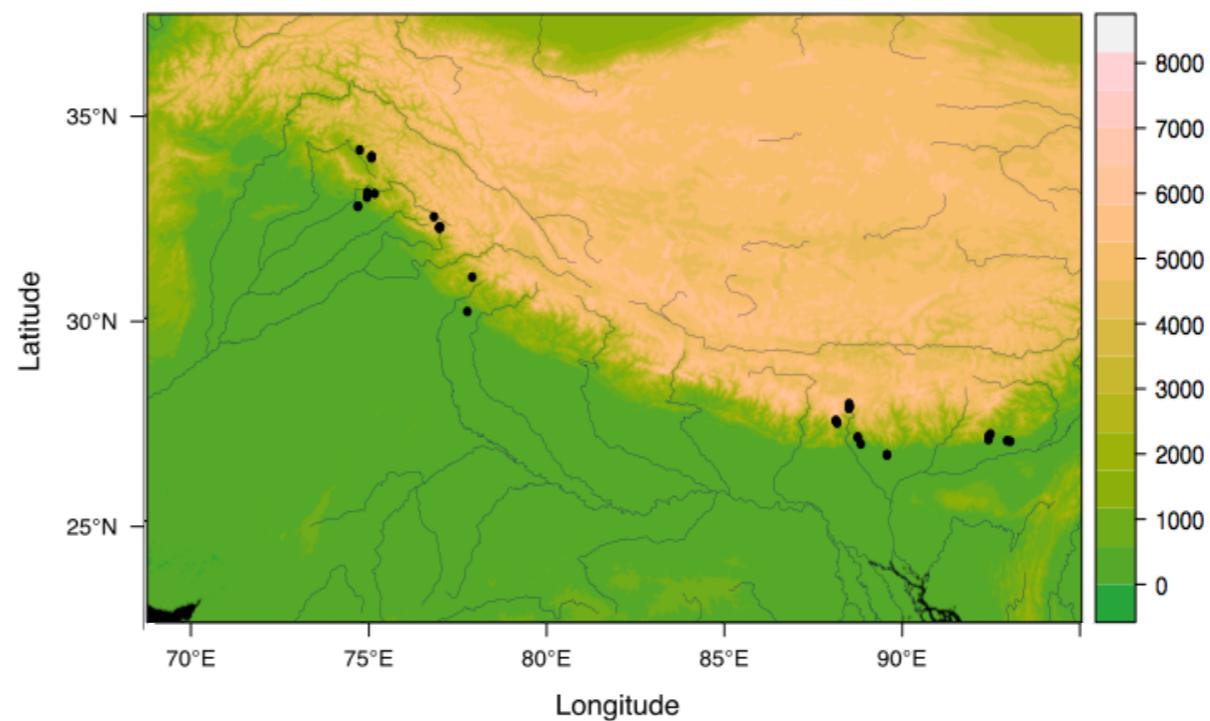
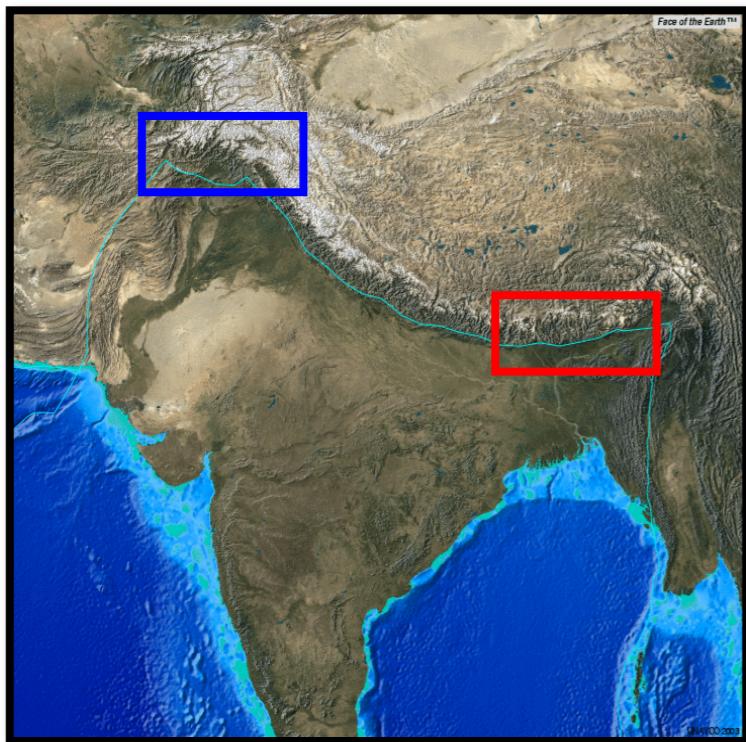


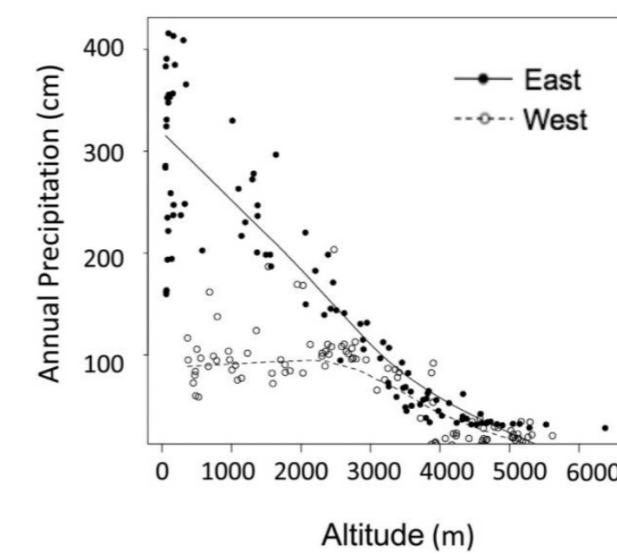
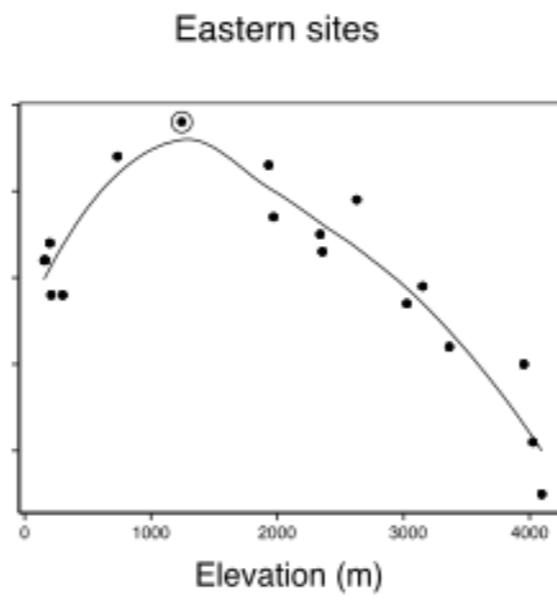
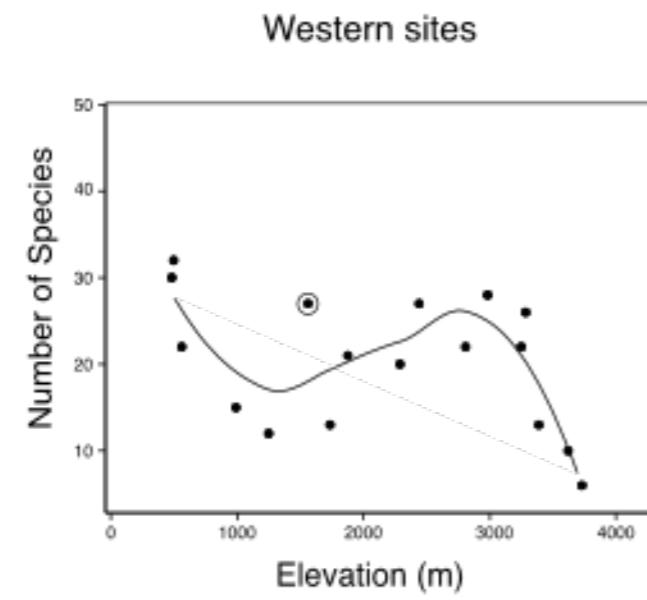
Road Ahead

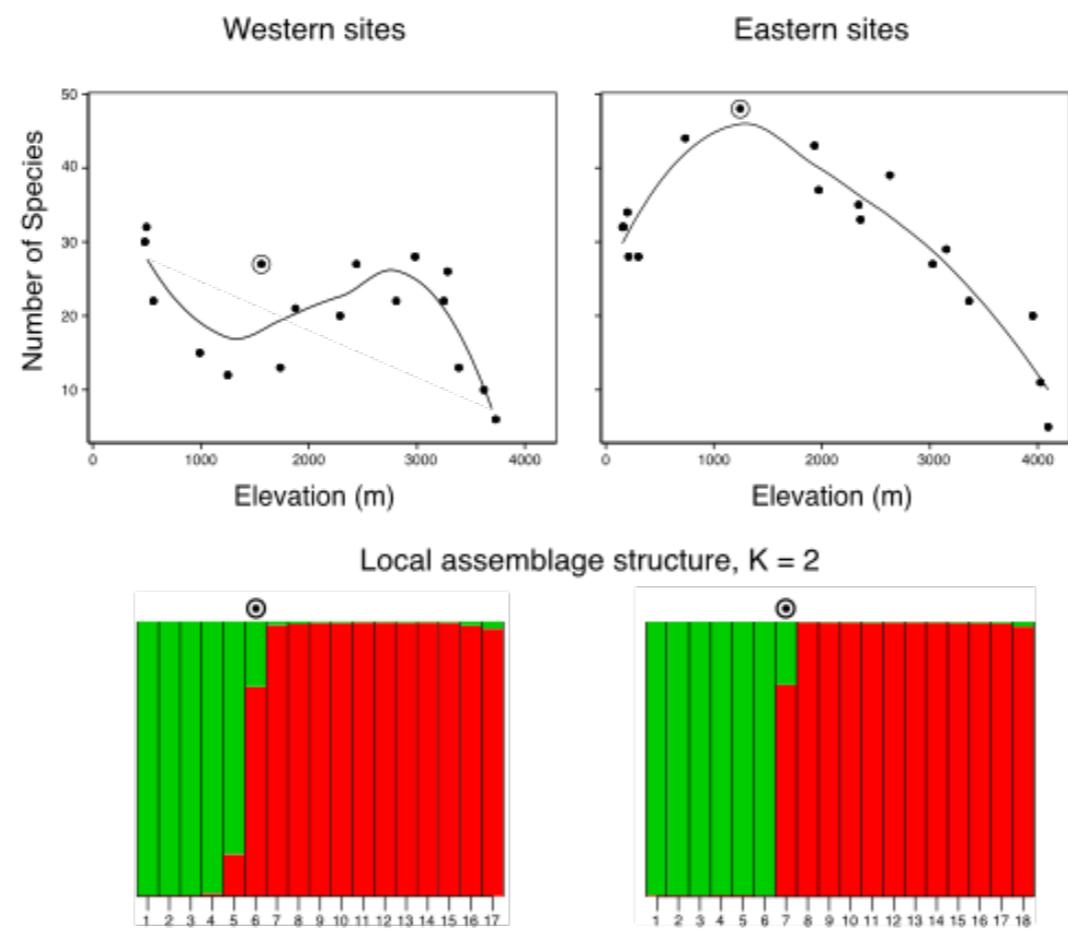
1. Address the issue of how to make the clusters more representative of cell types.
2. Variable selection or gene selection preprocessing step to eliminate genes not driving clustering.
3. Pathway analysis of the cluster expression profiles of genes.

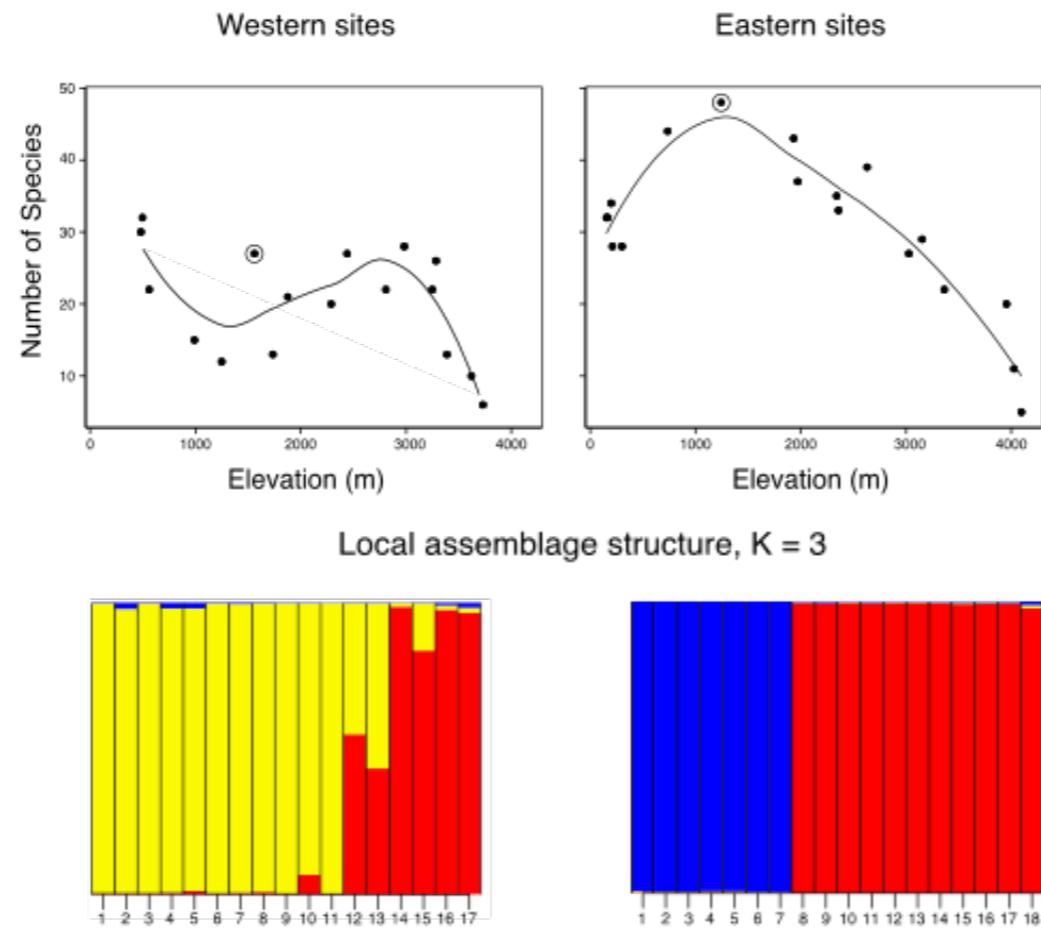


Phylogenetic Data Modeling

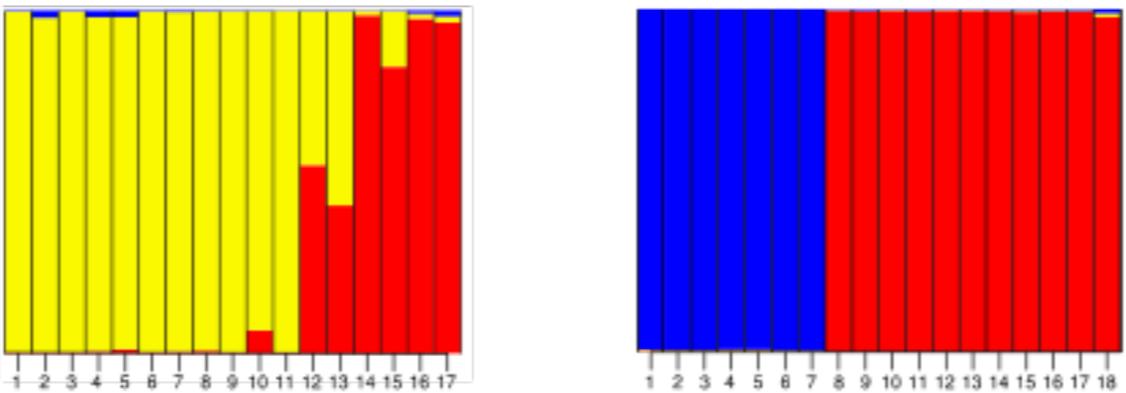




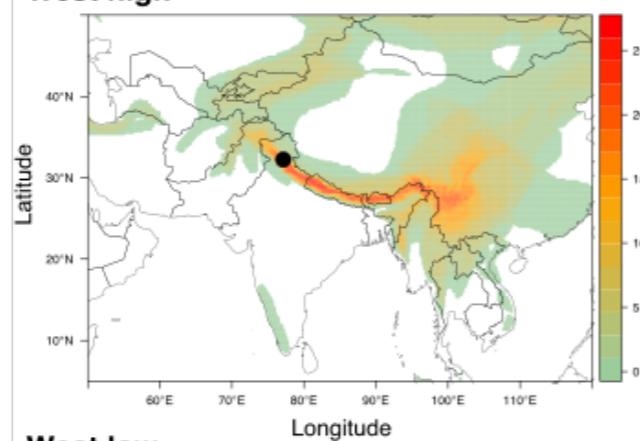




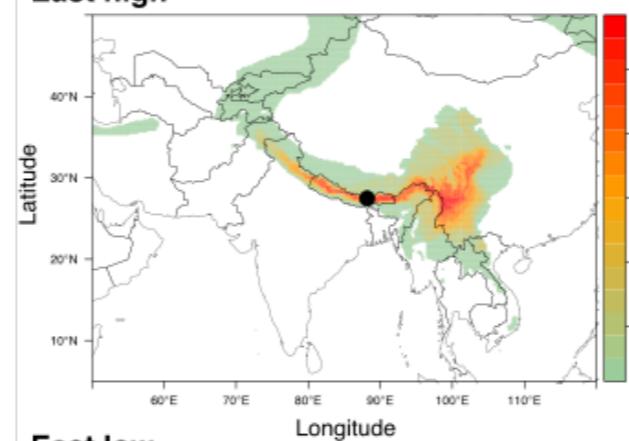
Local assemblage structure, K = 3



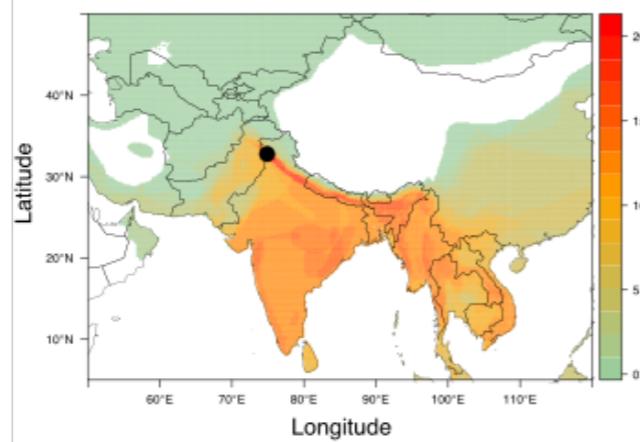
West high



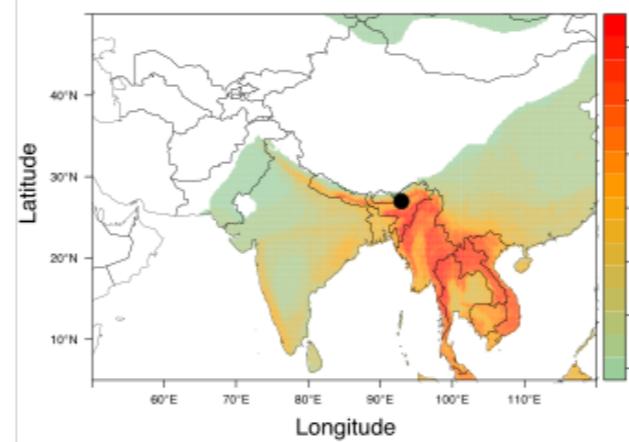
East high



West low



East low



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

Implementation:
R package **CountClust**
<https://github.com/kkdey/CountClust>

Acknowledgements

Matthew Stephens
Alex White
Trevor Price
Dhananjai Mohan

Matt Taddy
Gao Wang
Sarah Urbut
Raman Shah