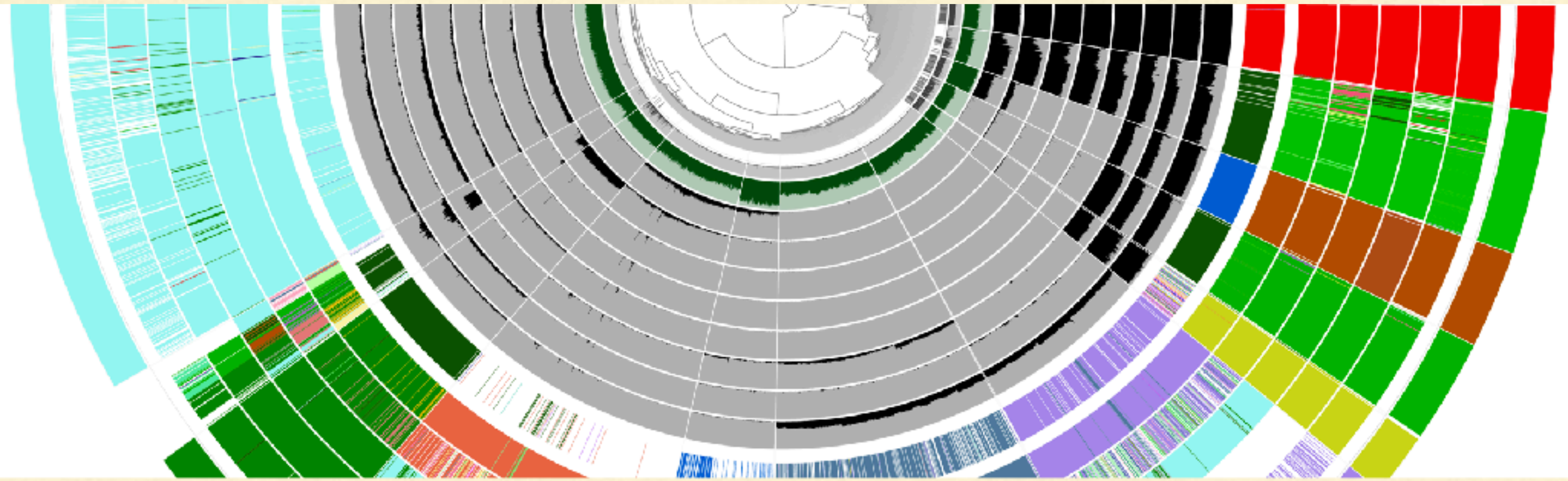


MODELING SHOTGUN DATA

Joint work with Sam Minot, Fred Hutch

Joint work with Meren Lab, U Chicago

Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

Photo credit: *anvi'o* development team

WGS DATA

- Metagenomic data gives you information about the complete genome of the organisms in a sample
  - Allows us to look at *function* rather than *just taxonomy*
- Perhaps ~100,000 genes per sample…
  - …many unique to a single environment!

DEPTH

- You can summarize metagenomic data into the *coverage* of each gene in each sample
  - Most depths are zero
- With so many more genes than samples, you are guaranteed to find "false discoveries”
  - Linear algebra exercise: 7 covariates (full rank), you can predict 7 observations with 100% accuracy

DIMENSION REDUCTION

- Statisticians *love* high dimensional problems
  - Statistics in high dimensions is not intuitive!
  - “Shrinkage”/LASSO ideas
- Statisticians know how to deal with high dimensional problems…

DIMENSION REDUCTION

- Step 1: Dimension reduce via biology
  - Cluster genes: genes are not independent biological observations, but rather are linked via physical pieces of DNA
  - Intelligent clustering reduces dimension

DIMENSION REDUCTION

- Step 2: Dimension reduce via statistics
  - If goal is to detect a difference between the average coverage in one group of samples versus another (e.g. pre/post treatment), can use regression-type approach
- Critical: multiple comparison adjustment
  - False discovery rate control: control expected % of "false" "significant" genes/clusters

DIMENSION REDUCTION

- Step 3: External validation

- A publicly available data is an amazing resource

- Apply the clustering algorithm from step 1 to new data, and see if results from model in step 2 persist

External validation is critical in high dimensional problems

RESOURCES

- For a baby R package and workflow, check out
  - **github.com/adw96/ShotgunSeq**
- Development version: *For advanced R users*
- Key: locally parallelise and trade off model complexity to optimize both computational and memory efficiency

COMING SOON-ISH

- Lots to do
- Adjust for different sampling intensities
- Adjust for short-read bias
- Extend to complex designs
  - Challenge: maintain speed

ENRICHMENT

- A few words about modeling in pangenomics
  - Pangenomics: comparing multiple genomes
  - Pangenomics looks at evolution of genomes, mutation rates, functional enrichment

Specifically, I'll say some words about modeling *enrichment* of genes: presence of a gene in one group of genomes vs another group of genomes

ENRICHMENT

- Suppose we have
- n_1 genomes from one group; n_2 genomes from another group
- X_1 genomes with the gene in group1; X_2 genomes with the gene in group 2
- If samples the genomes came from were observed independently, the "enrichment score" is

$$\frac{X_1/n_1 - X_2/n_2}{\sqrt{\left(\frac{X_1+X_2}{n_1+n_2}\right)\left(1-\frac{X_1+X_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

ENRICHMENT

- If samples the genomes came from were observed independently, the "enrichment score" is

$$\frac{X_1/n_1 - X_2/n_2}{\sqrt{\left(\frac{X_1+X_2}{n_1+n_2}\right)\left(1-\frac{X_1+X_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

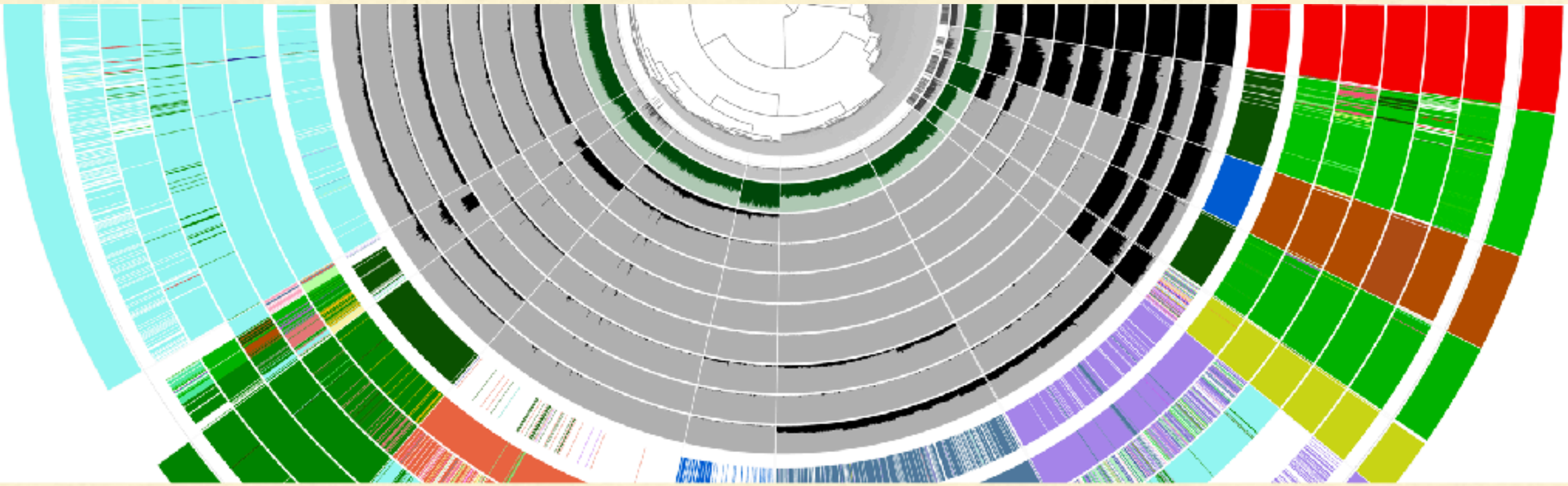
- Key points: adjusts for different numbers of genomes in each group; allows valid hypothesis testing & false discovery control

- Coming soon to anvi'o

ENRICHMENT

- If some genomes are more correlated than others, you need generalized mixed model *to do hypothesis testing*

    - Complex experimental design: take data out of anvi'o and use statistical software for modeling



MODELING SHOTGUN DATA

Joint work with Sam Minot, Fred Hutch

Joint work with Meren Lab, U Chicago

Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

Photo credit: *anvi'o* development team