



# BIO 724D

# Data Lunch

Kayla Wilhoit

University Program in Genetics and  
Genomics

Magwene Lab



# About Me!



- Dallas, Texas
- Interest in biology and genetics
- No computational skills :(

# About Me!



- Dallas, Texas
- Interest in biology and genetics
- No computational skills :(



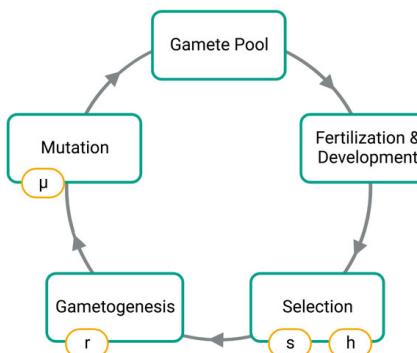
# About Me!



- Dallas, Texas
- Interest in biology and genetics
- No computational skills :(



- Texas A&M University
  - Biomedical Sciences Major
  - Genetics Minor
- R, RStudio, experimental design, basic statistics
- Population genetics simulations in R



# About Me!



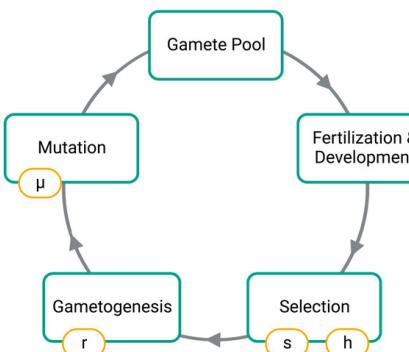
- Dallas, Texas
- Interest in biology and genetics
- No computational skills :(



MINECRAFT



- Texas A&M University
  - Biomedical Sciences Major
  - Genetics Minor
- R, RStudio, experimental design, basic statistics
- Population genetics simulations in R



- Duke UPGG Rotations
  - Wray lab - Intro to Unix, command line, genomics tools
  - Magwene lab - More Unix, python, snakemake,
  - Lowe lab - More snakemake, slurm, bash scripting
- Joined Magwene lab in Spring 2024

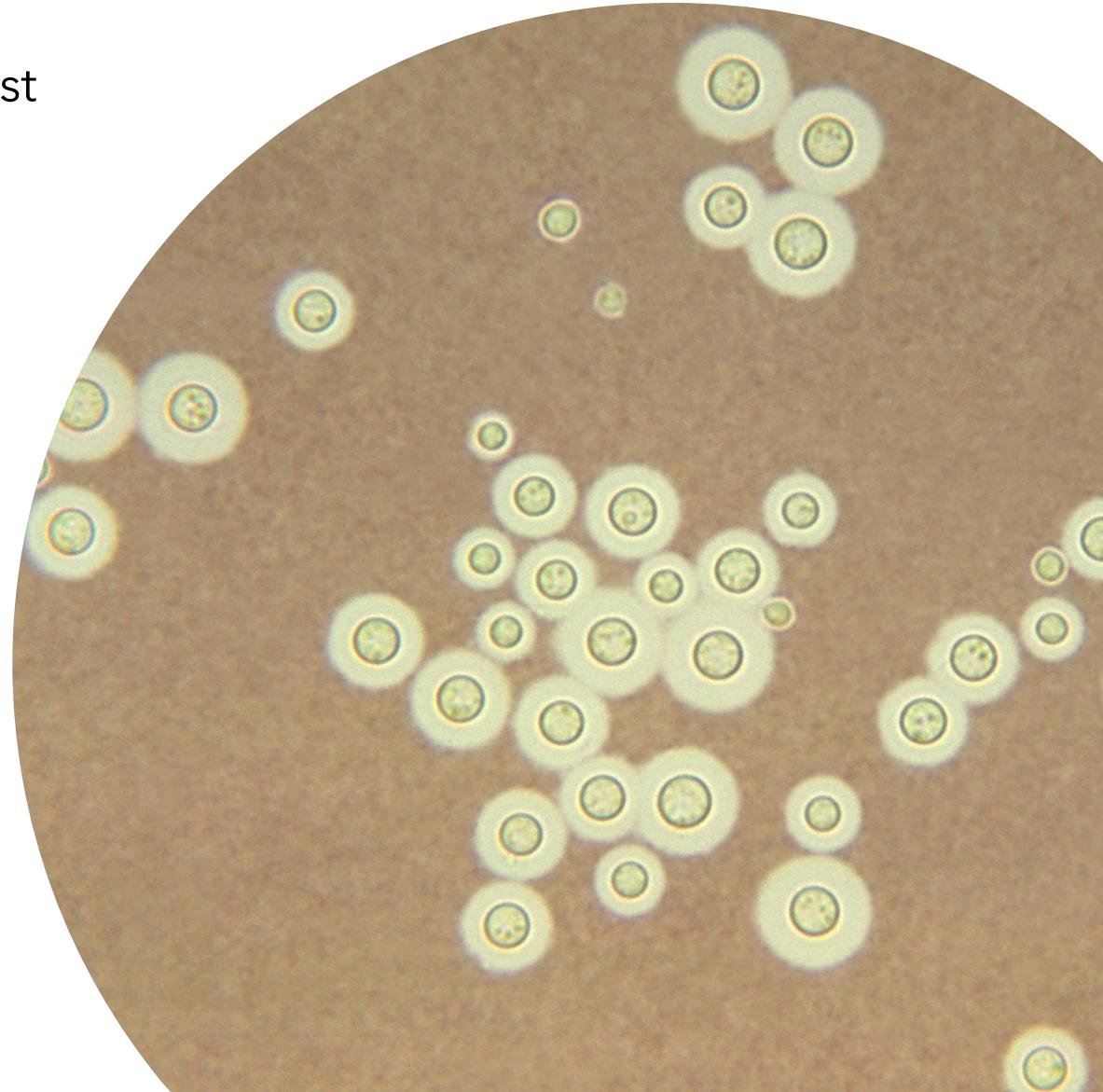
Duke



University Program in  
Genetics and Genomics

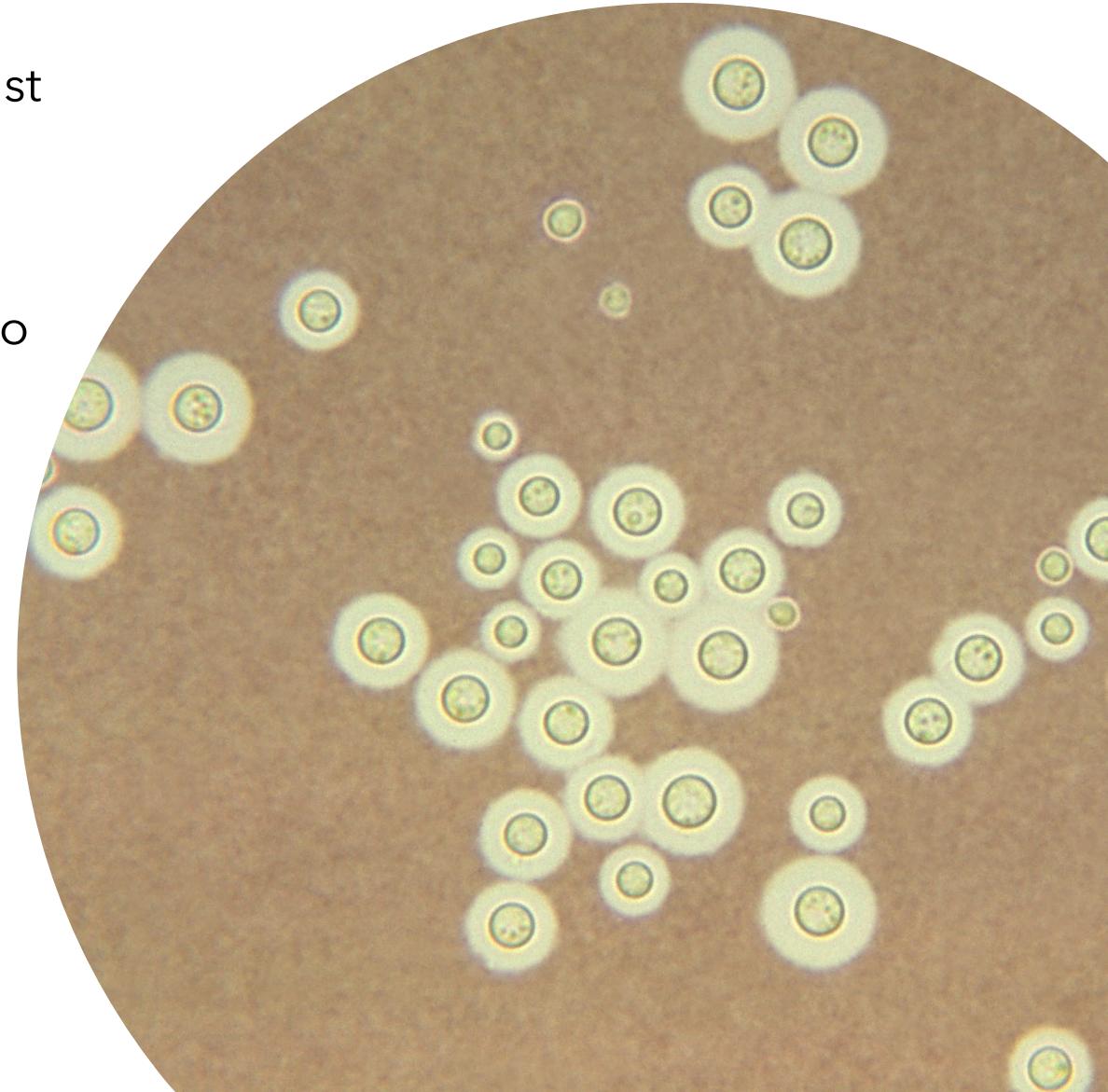
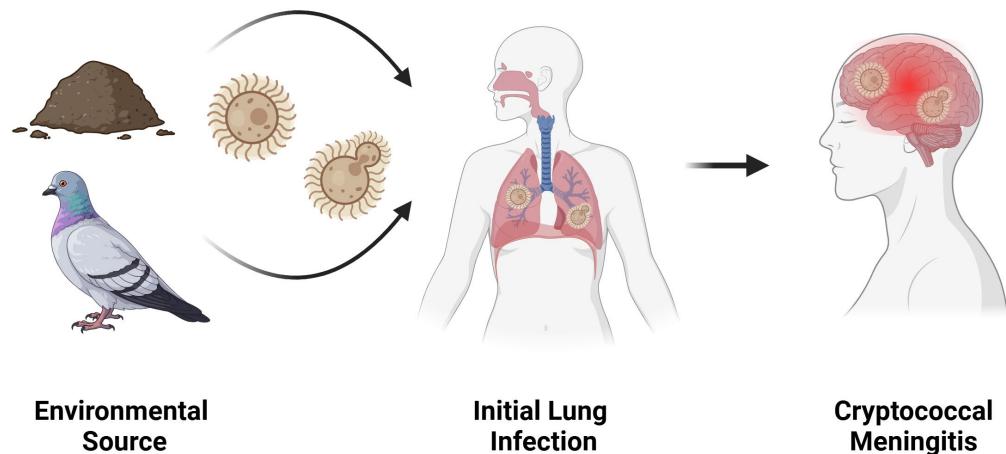
# Research Interests

- *Cryptococcus neoformans* is a pathogenic fungal yeast that causes over 180,000 deaths annually
  - Primarily affects immunocompromised patients
  - Current antifungal treatments are often ineffective



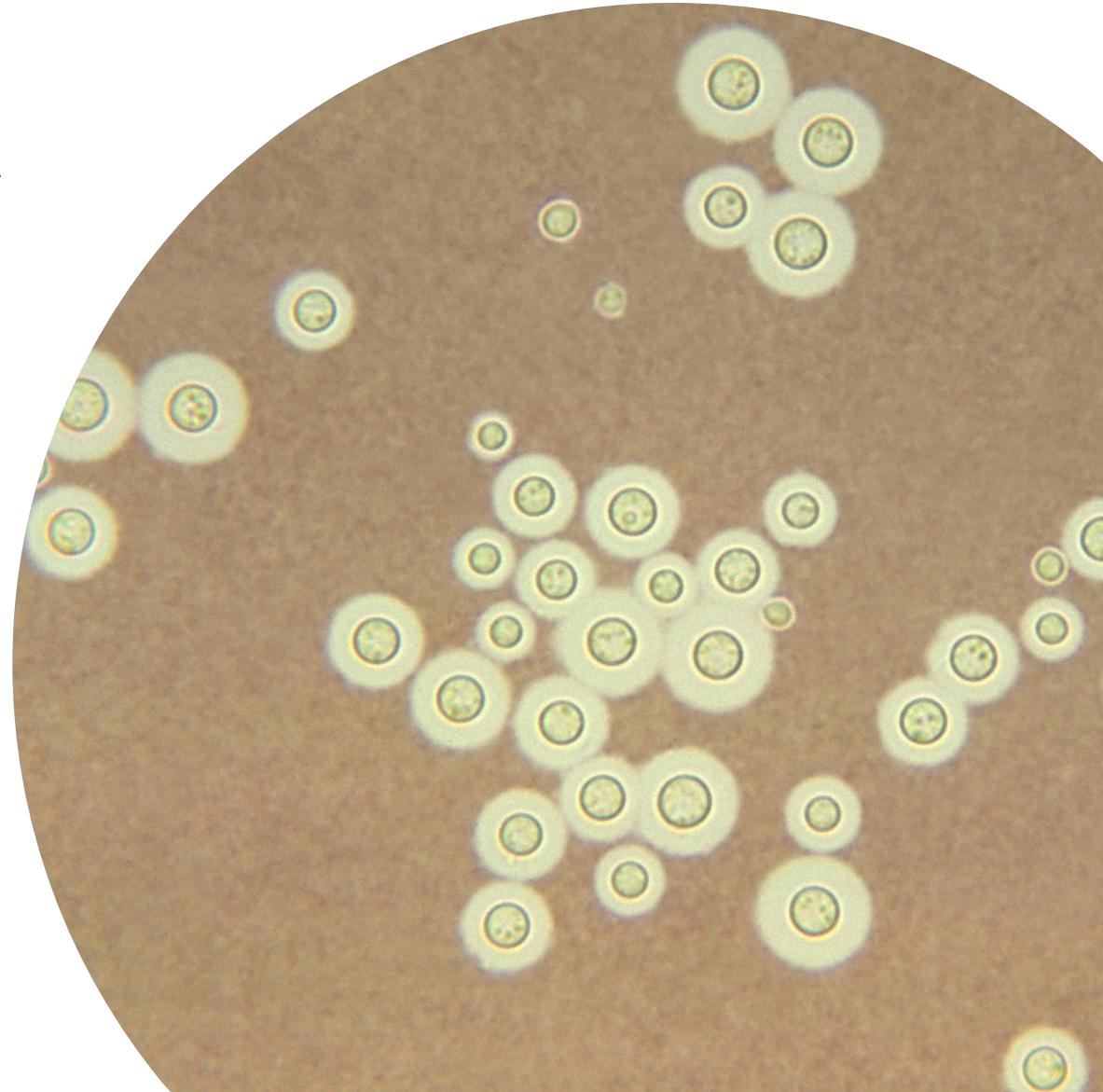
# Research Interests

- *Cryptococcus neoformans* is a pathogenic fungal yeast that causes over 180,000 deaths annually
  - Primarily affects immunocompromised patients
  - Current antifungal treatments are often ineffective
- Environmental sources include soil and pigeon guano
- High genetic plasticity and variation
  - 14 chromosomes, usually haploid
  - ~19 Mb genome

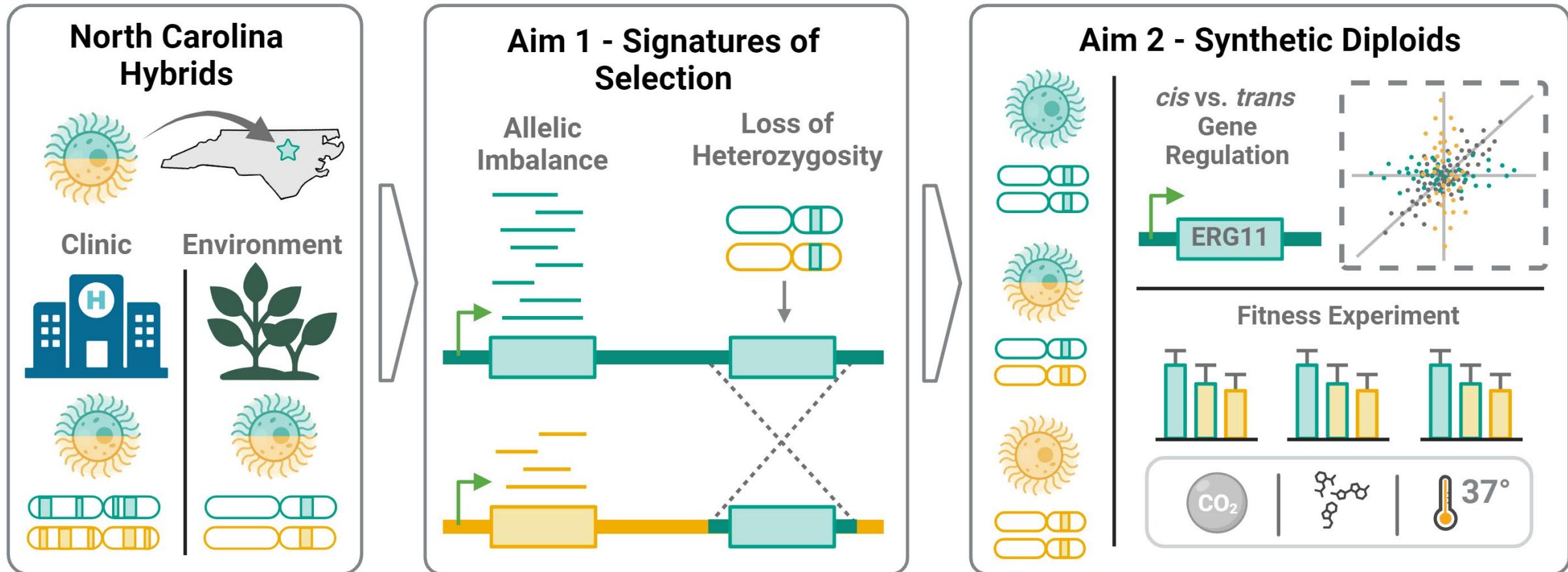


# Research Interests

- Evolution of and selection on fungal hybrids in *Cryptococcus neoformans*
- Population dynamics of clinical and environmental *Cryptococcus* in North Carolina
- Mechanisms of genome regulation in diploid and hybrid *Cryptococcus*



# Example 1 - Project Background



# North Carolina Samples

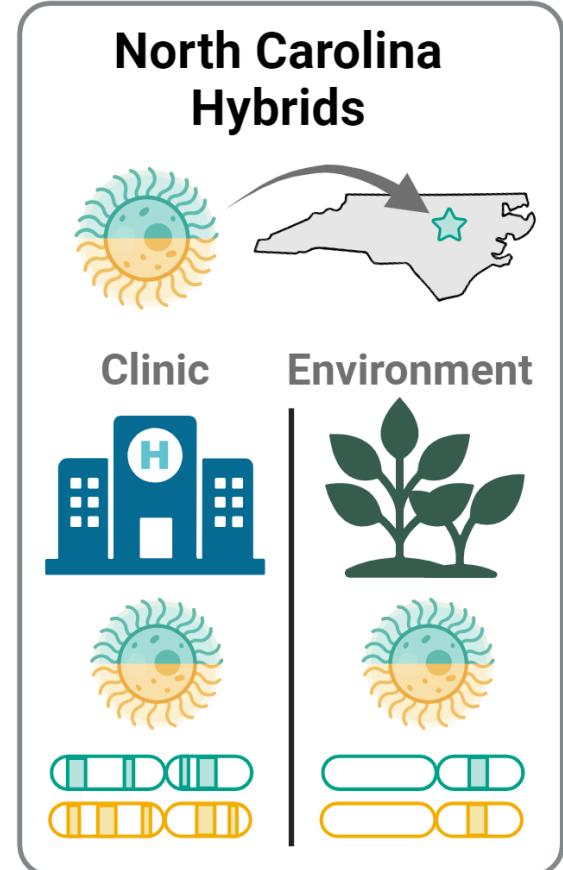
- **326** sequenced clinical isolates
  - + more from Duke Hospital
  - **40** diploid hybrids (13%)
- **203** sequenced environmental isolates
  - **53** diploid hybrids (45%)
- **Are there genomic differences between clinical and environmental isolates?**



**Marhiah Montoya, PhD**



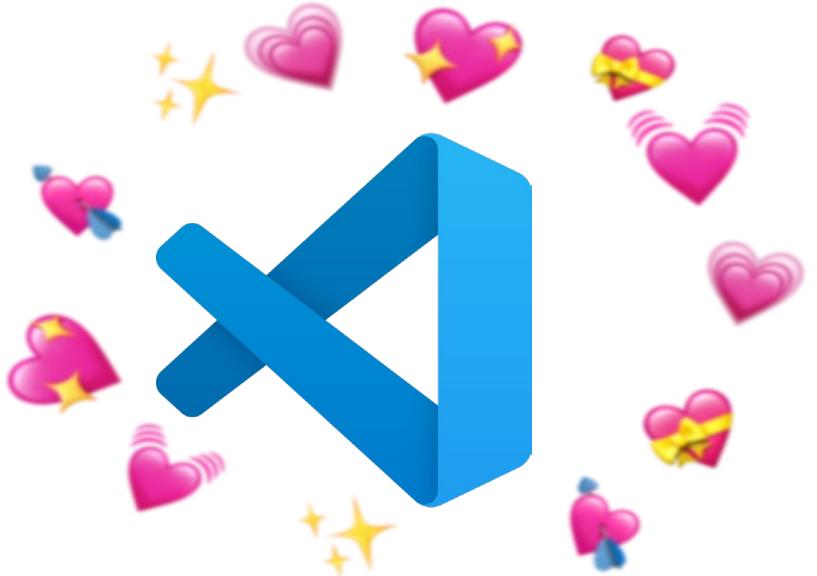
**Anastasia Litvintseva, PhD**



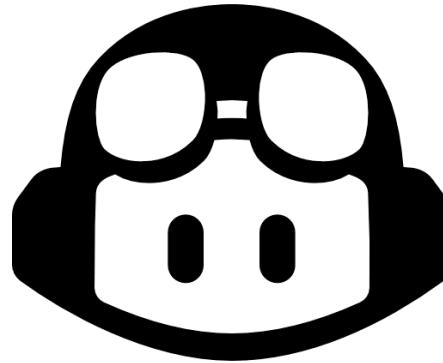
# Computational Problems

- Lots of big genome files!
  - Naming conventions
  - Location consistency
  - Subsetting groups of samples for efficient analysis
- Dealing with hybrid genomes
- Integrating multiple genomics tools into one pipeline
- Producing useful intermediate files
- Visualizing results quickly

# Favorite Tools



**Visual Studio Code**  
**<3**



**GitHub Copilot**



**Anaconda**



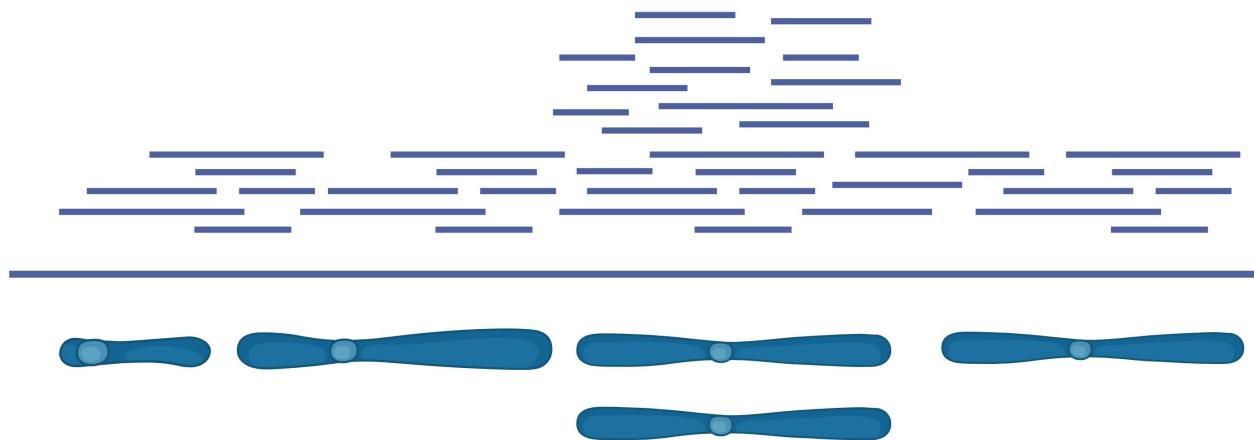
**Windows Subsystem for  
Linux (Ubuntu)**



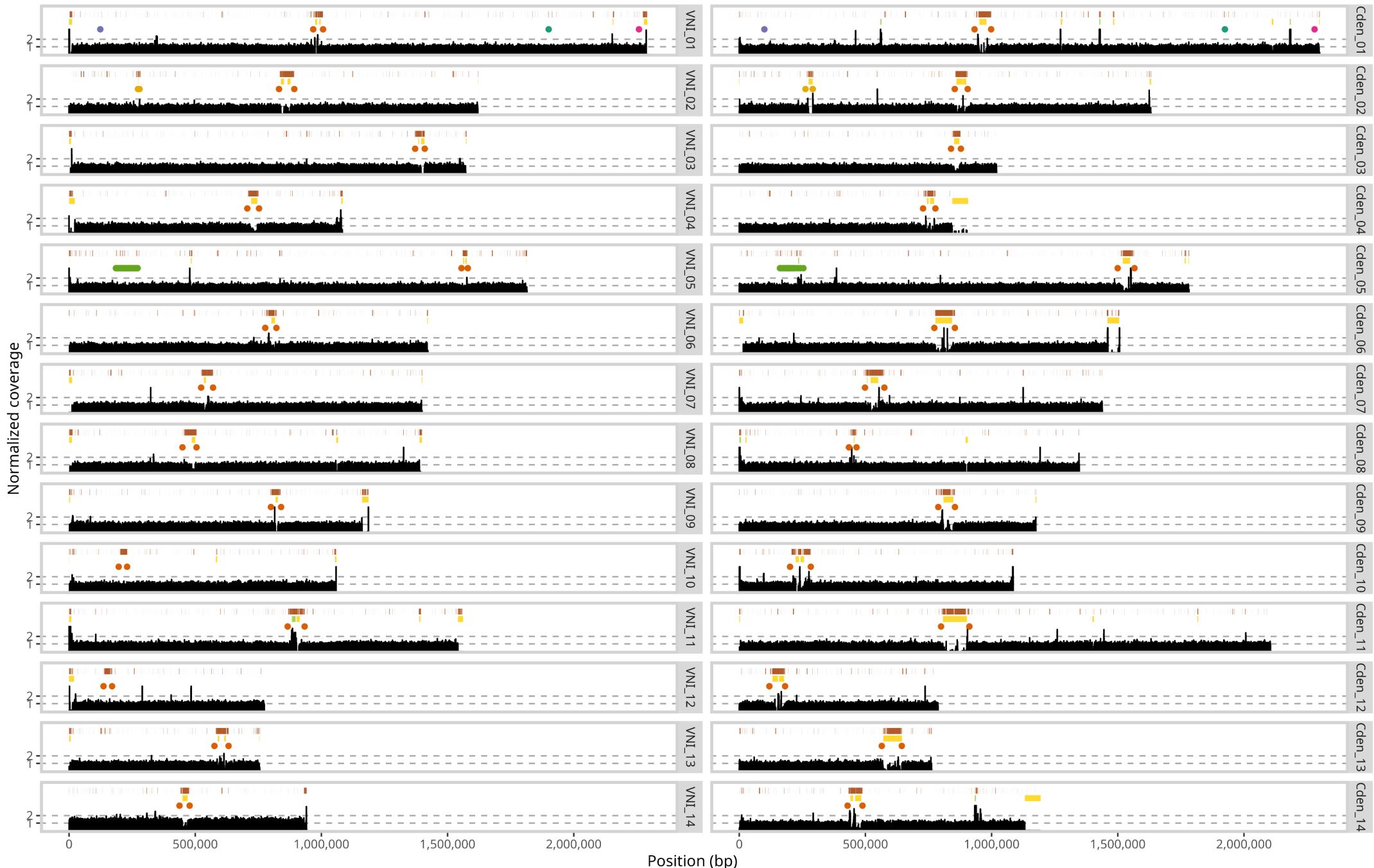
**Snakemake**

# Read Depth Plots

- Align Illumina short-reads to concatenated reference genome
- Normalize reads to background
- Quantify read depth across the genome



Lineage: VNI\_VNIV Sample: PMY3800



Type of repetitive sequence

- Simple repeat
- Others

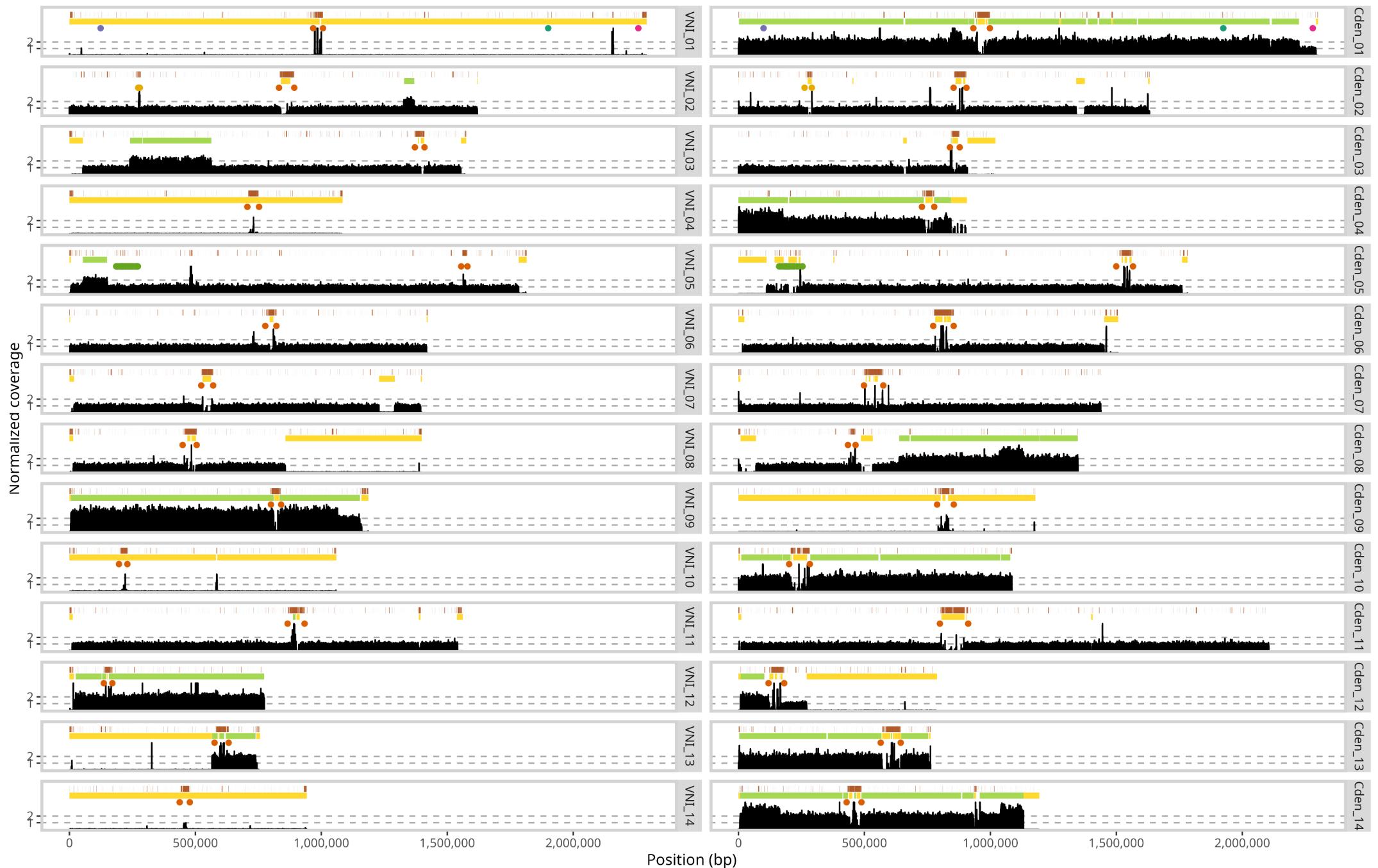
Structural variant

- DELETION
- DUPPLICATION

Loci

- AFR1
- Centromeres
- ERG11
- ERG2
- MAT
- rRNA

Lineage: VNI\_VNIV Sample: PMY3476



Type of repetitive sequence

- Simple repeat
- Others

Structural variant

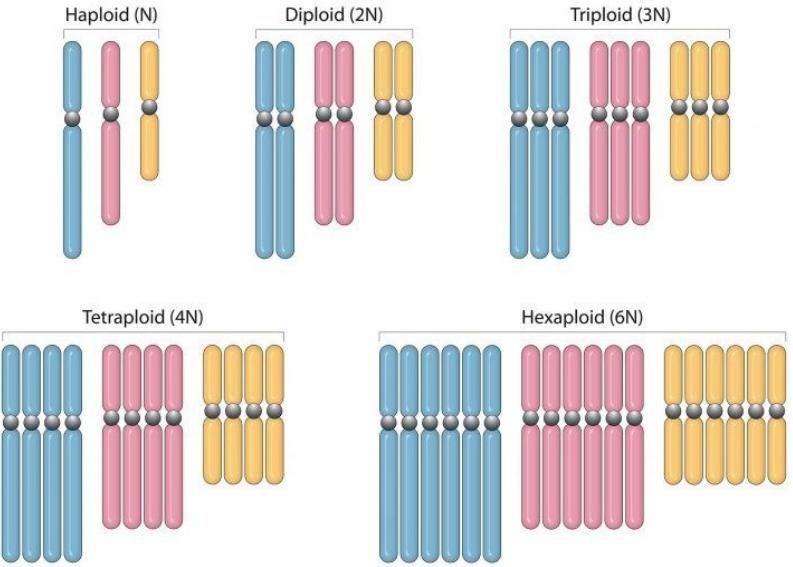
- DELETION
- DUPPLICATION

Loci

- AFR1
- Centromeres
- ERG11
- ERG2
- MAT
- rRNA

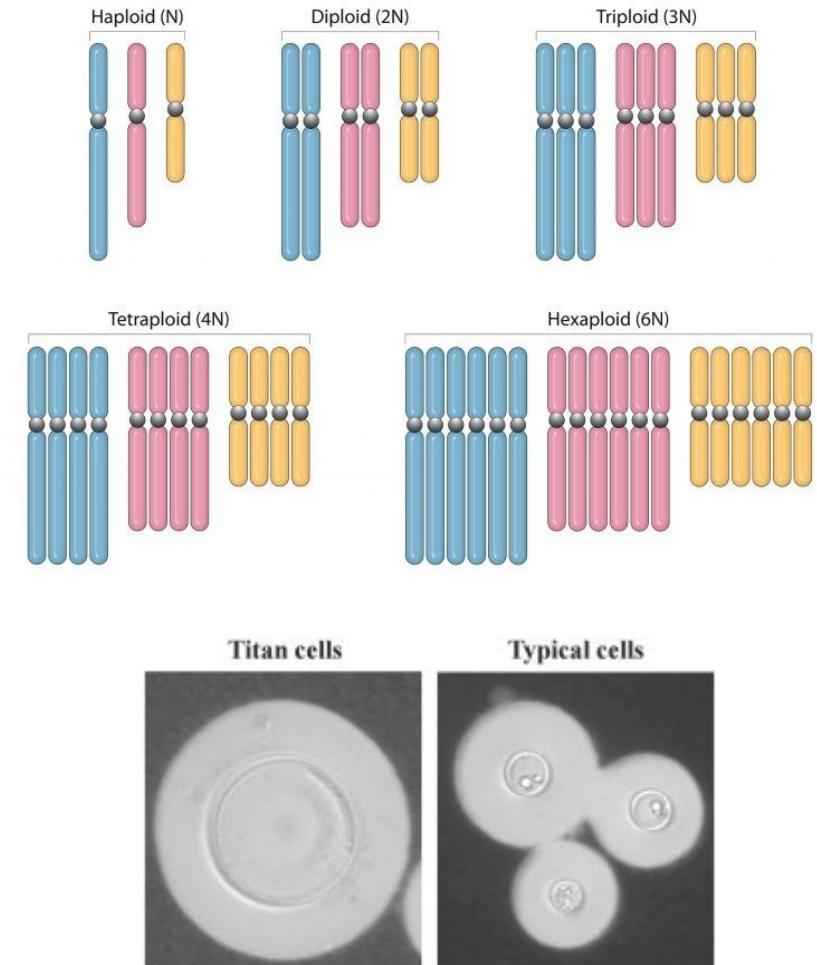
# Example 2 – Project Background

- *Cryptococcus neoformans* is typically haploid and clonal
  - Traditional crossing-over recombination and meiosis only occurs during specific mating conditions



# Example 2 – Project Background

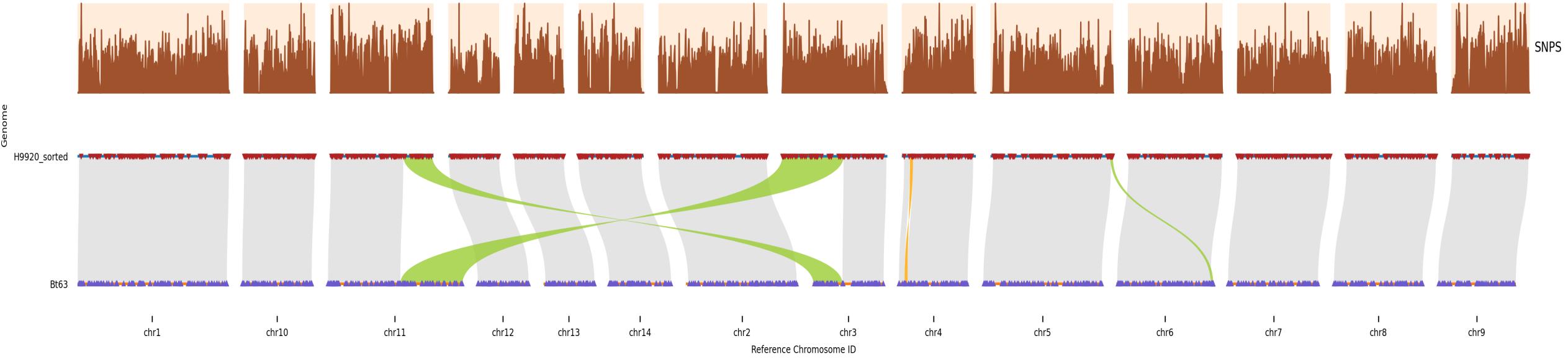
- *Cryptococcus neoformans* is typically haploid and clonal
  - Traditional crossing-over recombination and meiosis only occurs during specific mating conditions
- Some *C. neoformans* cells increase in ploidy and become massive under stress (titanization)
  - These polyploid titan cells can shed off diploid daughter cells that look normal
- **Is recombination occurring in titan cells?**



# Process

Annotations  
Syntenic  
Inversion  
Translocation

Syri - Synteny and Rearrangement Identifier  
<https://github.com/schneebergerlab/syri>



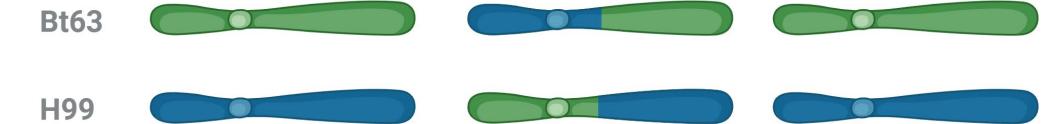
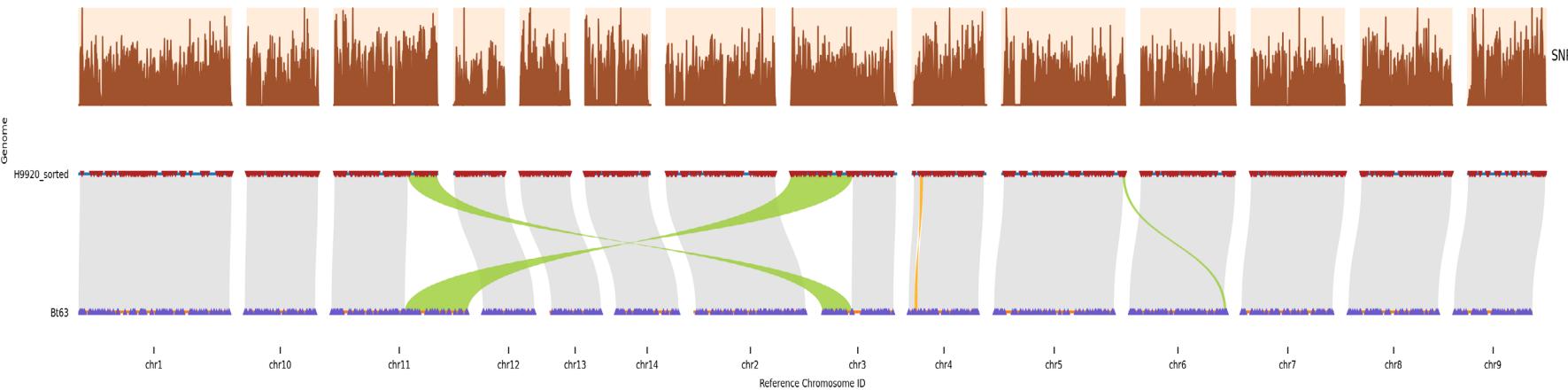
- Leverage SNPs distributed across the genome to identify origin of individual reads

# Process

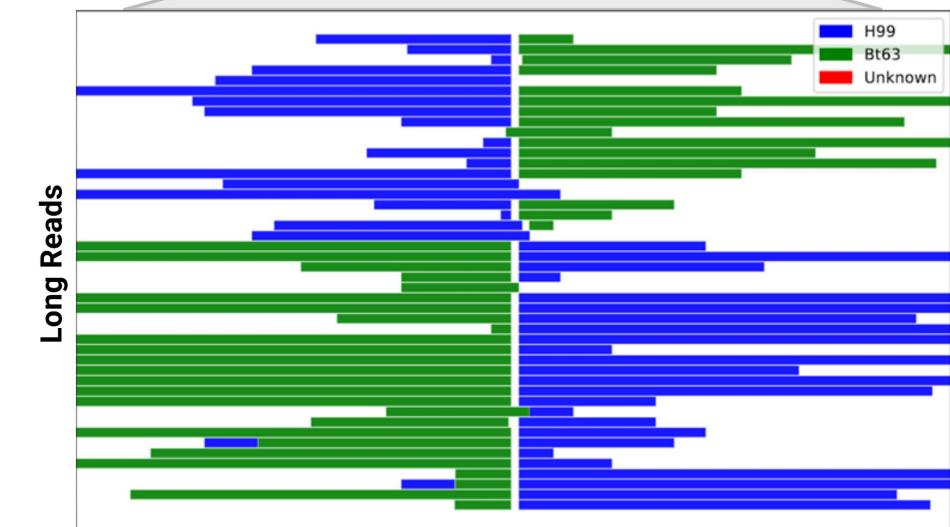
Annotations  
Syntenic  
Inversion  
Translocation

Syri - Synteny and Rearrangement Identifier

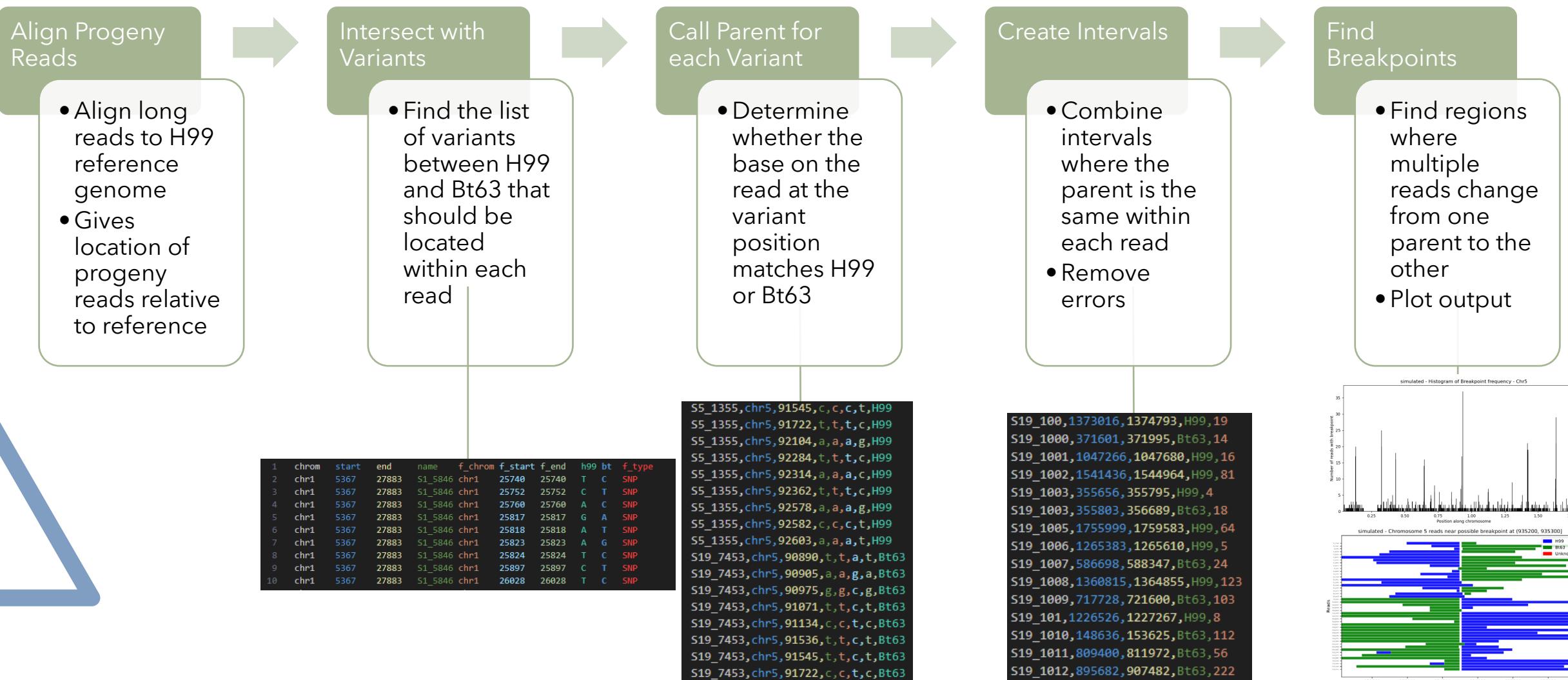
<https://github.com/schneebergerlab/syri>



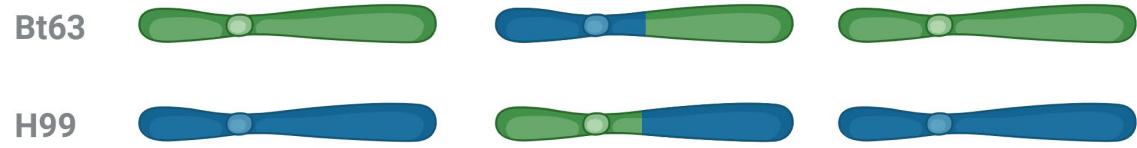
- Leverage SNPs distributed across the genome to identify origin of individual reads



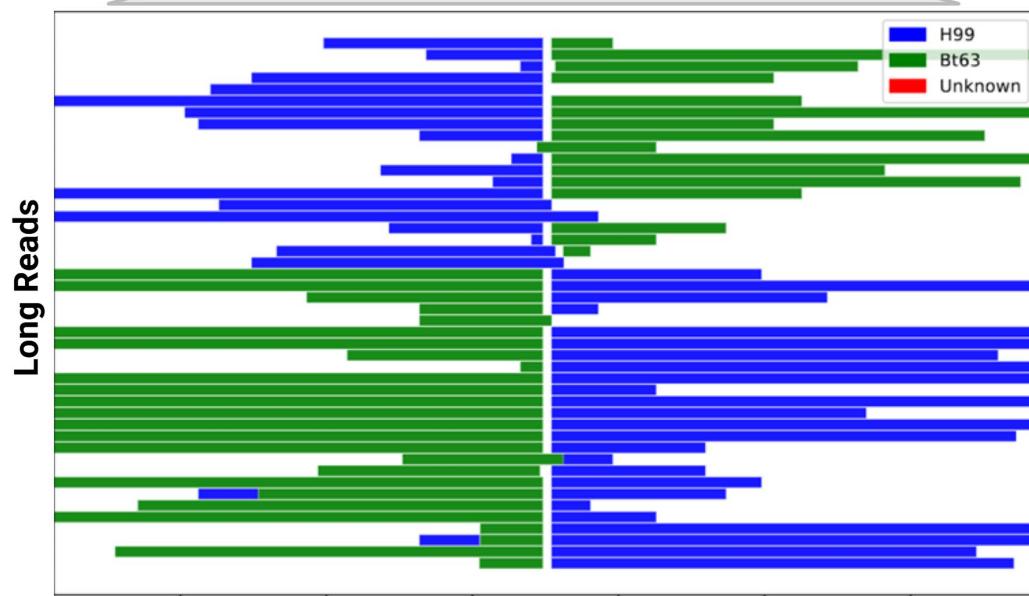
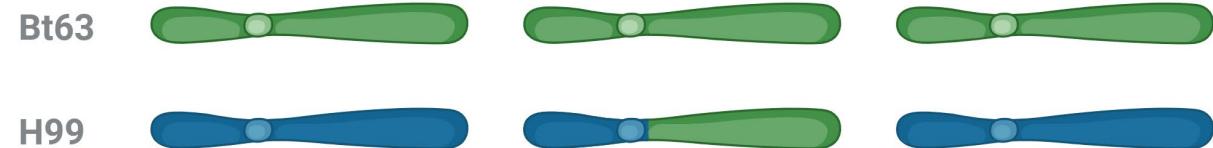
# Process



**Simulated Data  
(Crossing Over)**



**Actual Data  
(Loss of Heterozygosity)**



# Genomics Software and Packages

- Genome tools (command line)
  - **minimap2**
  - **mosdepth**
  - **seqkit**
  - **Ira**
  - **samtools**
  - **bedtools**
  - **wfmash**
  - **syri**
  - **snippy/snpeff**
- Python packages
  - **pandas**
  - **pysam**
  - **pybedtools**
  - **multiprocessing**
- R packages
  - **tidyverse**
  - **ggplot2 <3**
  - **patchwork**

# Useful VSCode Extensions

- Remote Development
  - **Remote - SSH: Editing Configuration Files**
  - **Remote Explorer**
  - **Remote - Tunnels**
    - For WSL
  - **Remote Repositories**
    - Copy/download files directly from repositories without cloning!
- Language-specific
  - **Jupyter (+ cell tags, renderers, slide show)**
  - **Python (+ Pylance, Debugger)**
  - **R (+ Debugger)**
  - **Snakemake Language**
- Quality of Life
  - **Rainbow CSV**
  - **Edit CSV**
    - Excel-like sheets interface
  - **Data Wrangler**
    - Powerful python-based data frame manipulation interface!
  - **SVG Preview**
  - **vscode-pdf**
  - **filesize**
    - Shows file size info + other metadata interface
- Copilot
  - **GitHub Copilot**
  - **GitHub Copilot Chat**

# Miscellaneous Tips

- Get on one of the Duke BioRender subscriptions for quick figures
  - Bio/EvAnth, Chemistry
- Get the free Adobe suite from Duke OIT
- Connect GitHub Copilot to VSCode
- Add your ssh key to Duke IT to avoid entering a password every time



**Thank you!**

[kayla.wilhoit@duke.edu](mailto:kayla.wilhoit@duke.edu)

