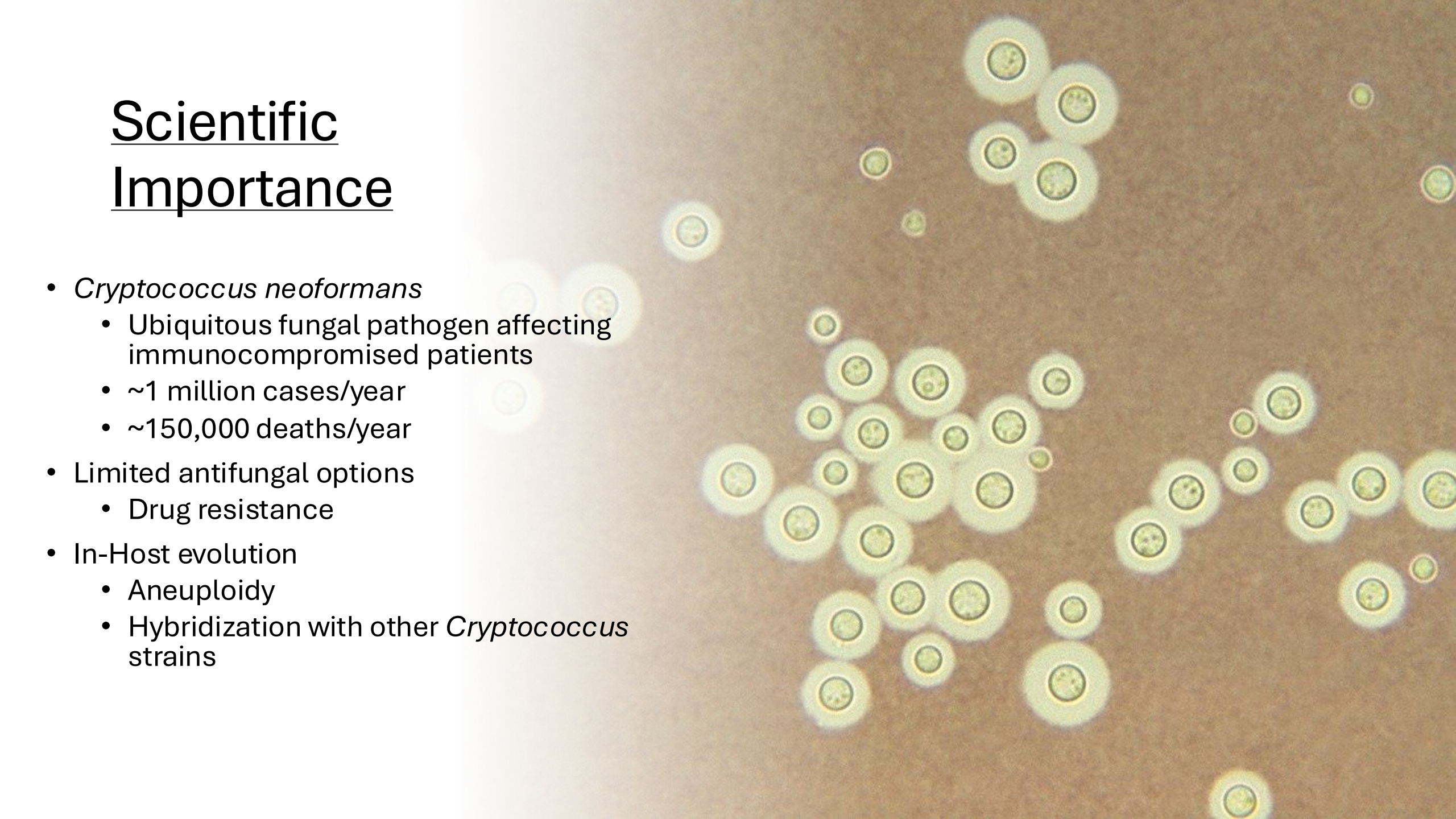
Gene expression workflow for Cryptococcus hybrid samples

Kayla Wilhoit

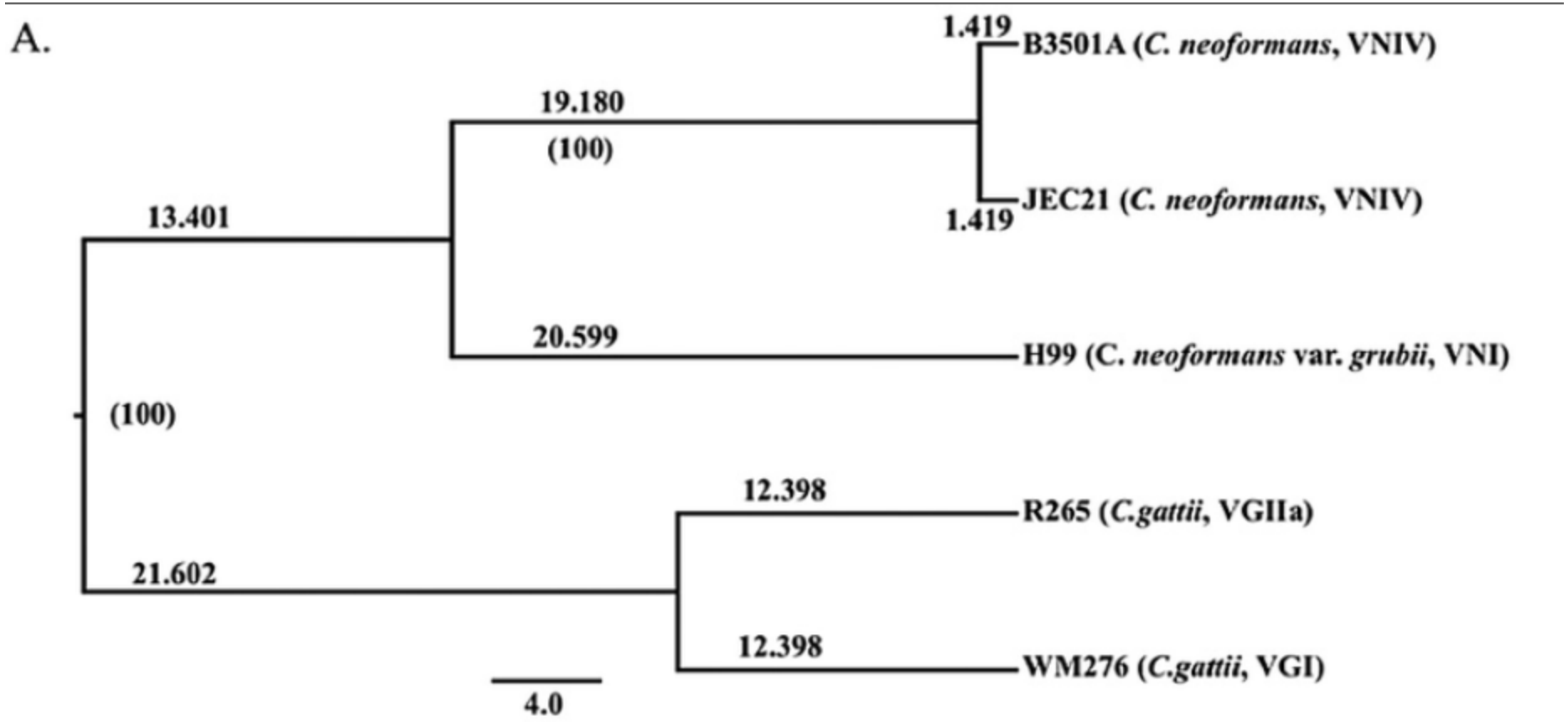
4/15/2024

Scientific Importance

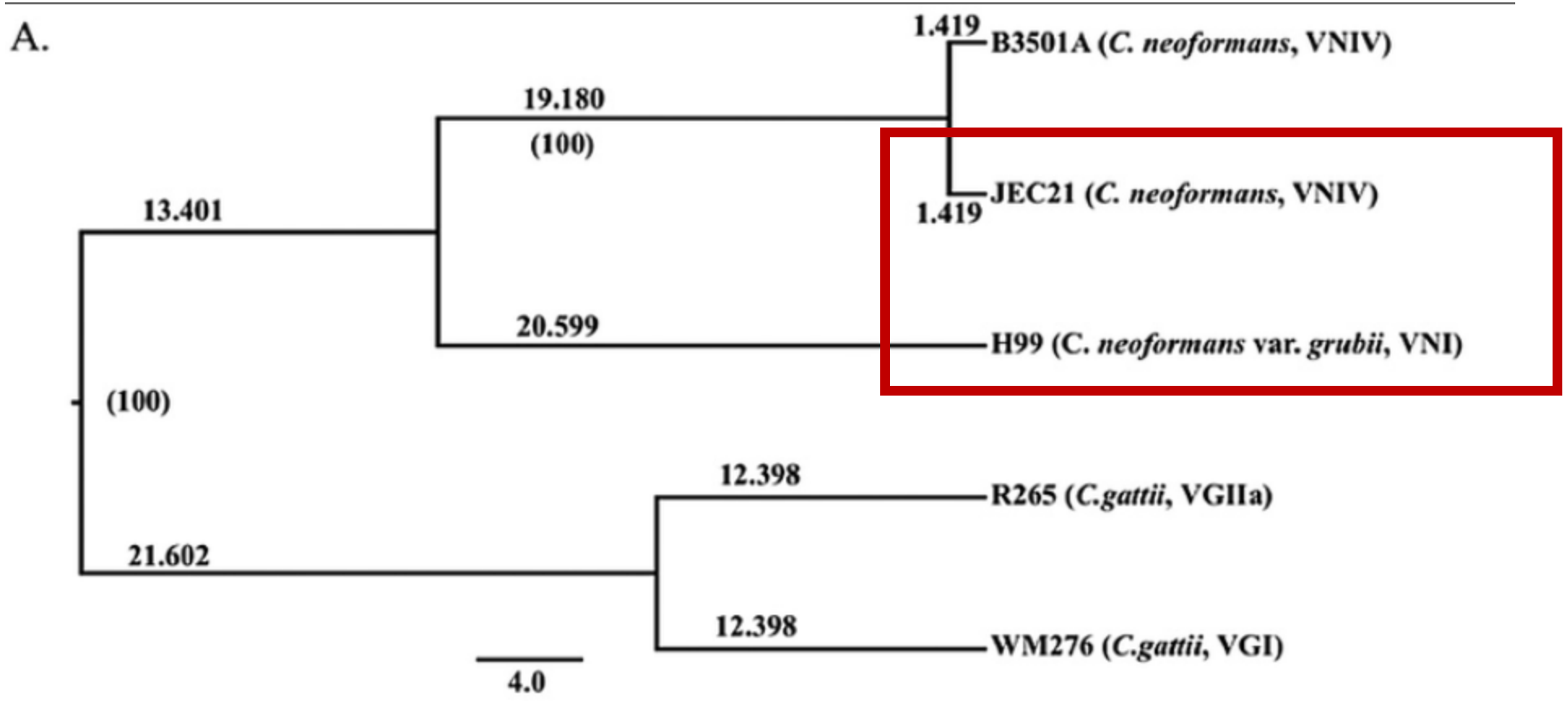
- *Cryptococcus neoformans*
  - Ubiquitous fungal pathogen affecting immunocompromised patients
  - ~1 million cases/year
  - ~150,000 deaths/year
- Limited antifungal options
  - Drug resistance
- In-Host evolution
  - Aneuploidy
  - Hybridization with other *Cryptococcus* strains



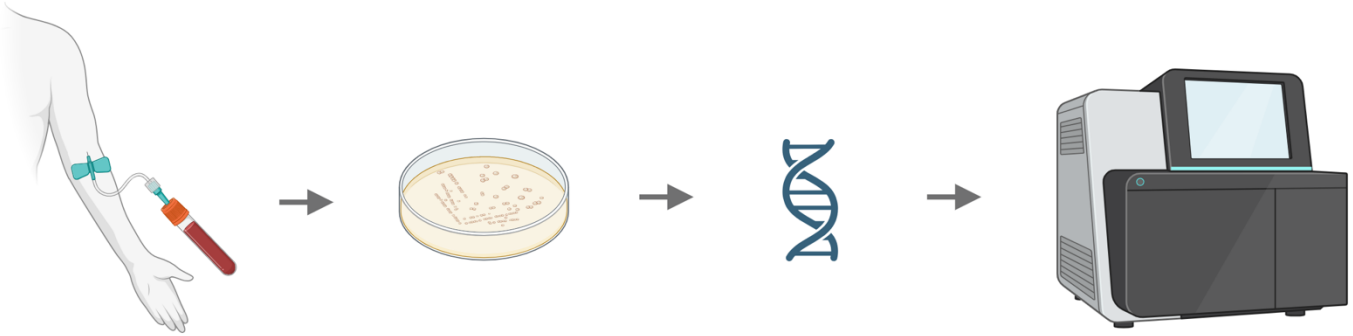
Cryptococcal Hybrids



Cryptococcal Hybrids



Previous analysis of clinical *Cryptococcus* samples

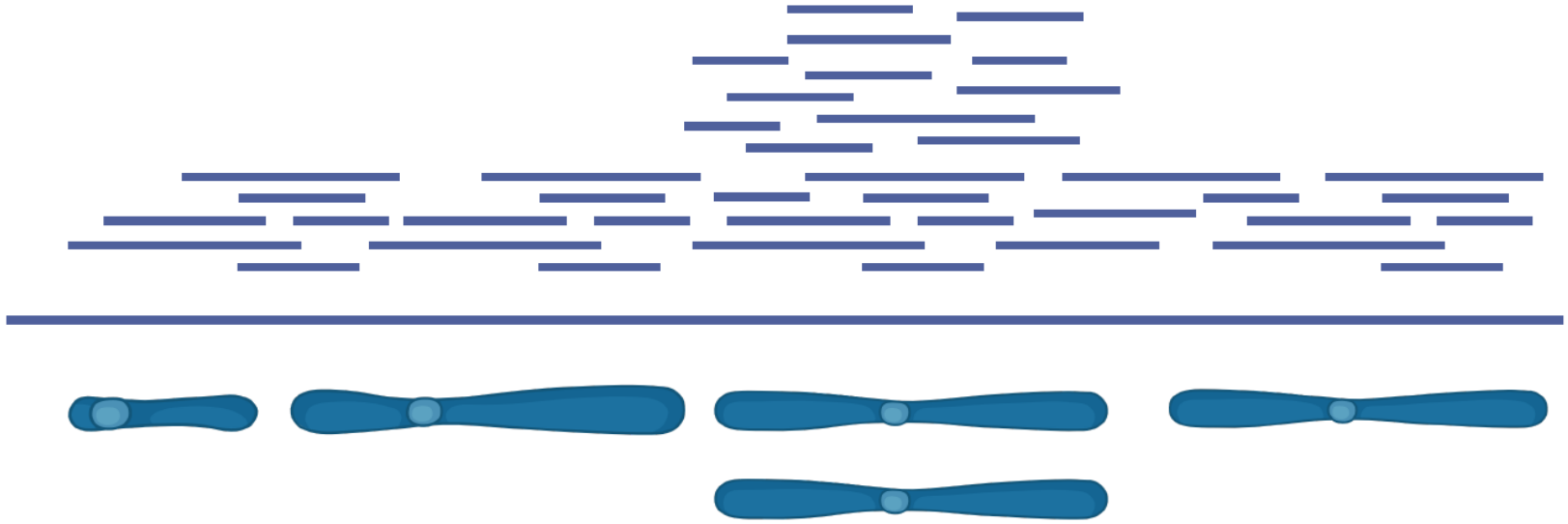


- Preserved cultures isolated and sequenced with **Illumina short-read sequencing**

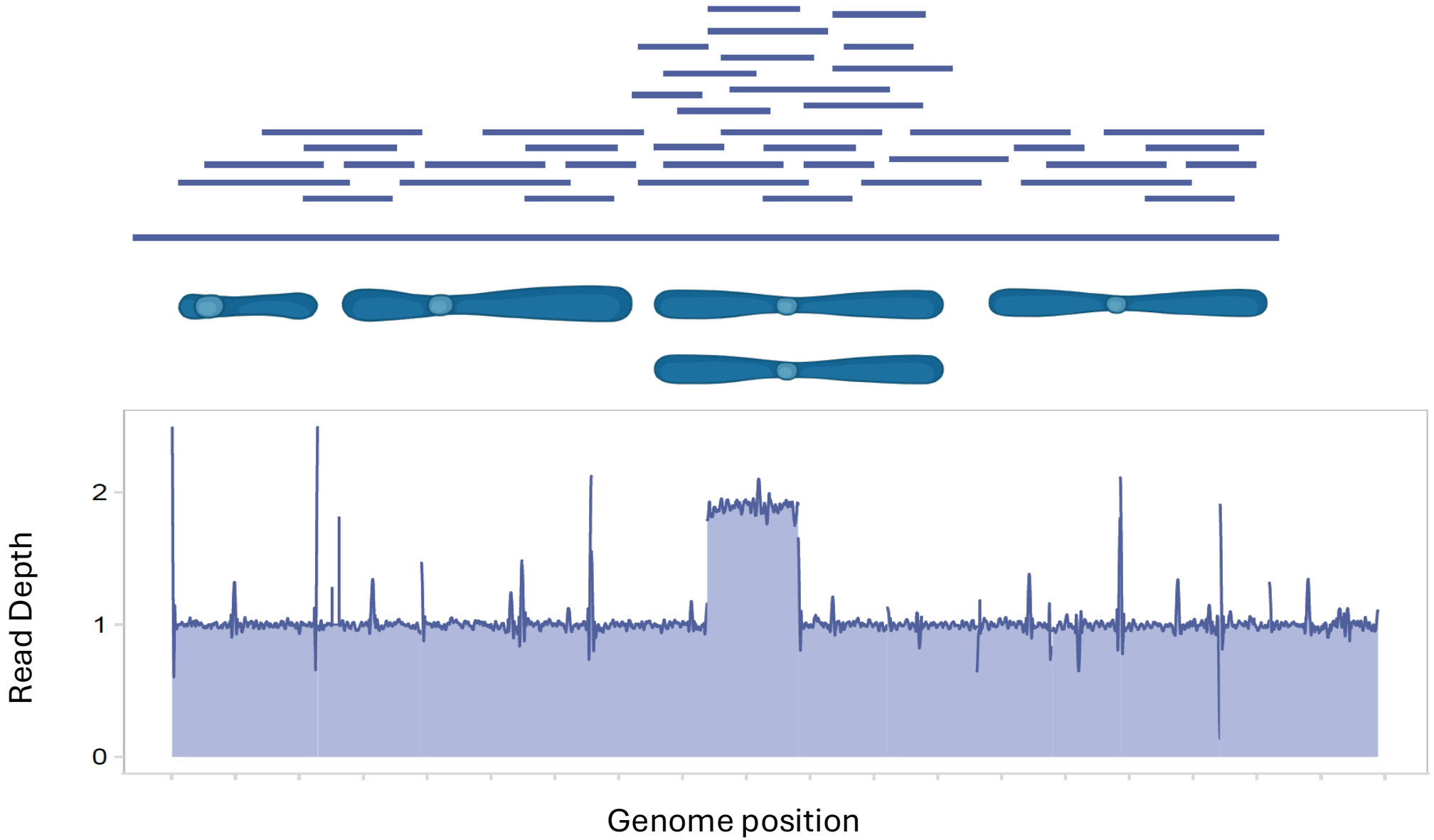
- Sample reads mapped to genomes of all sequenced *Cryptococcus* strains

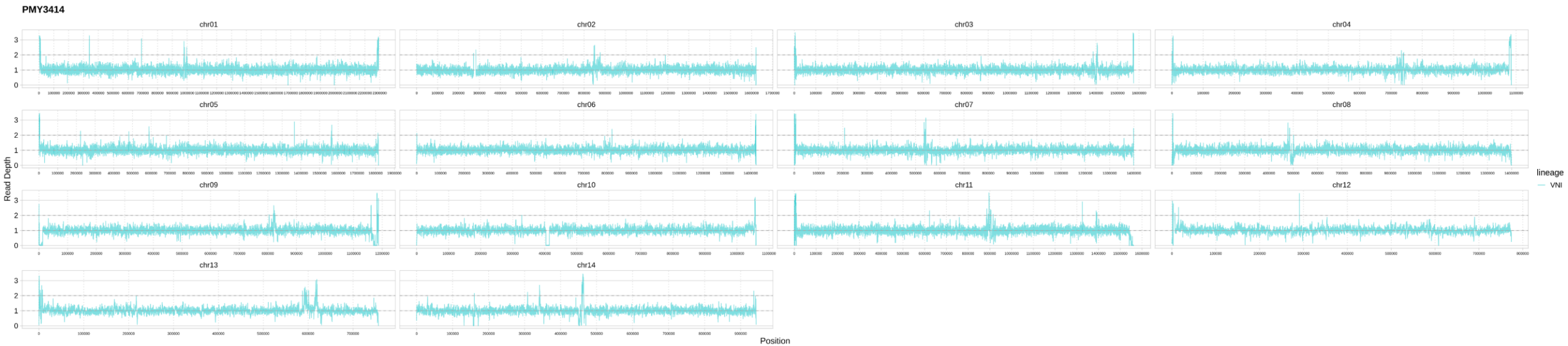
- **Read depth analysis** to find structural variants

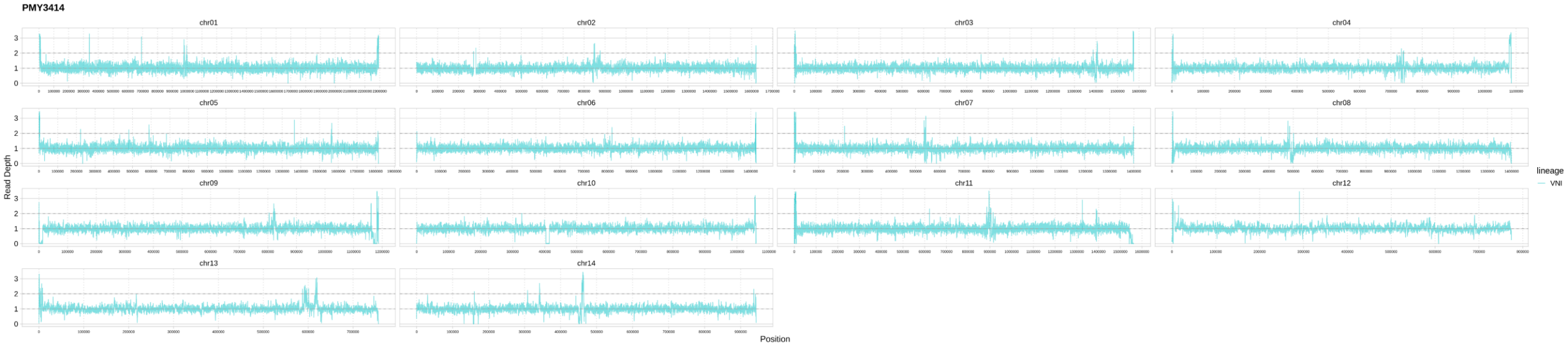
Genomic read depth analysis



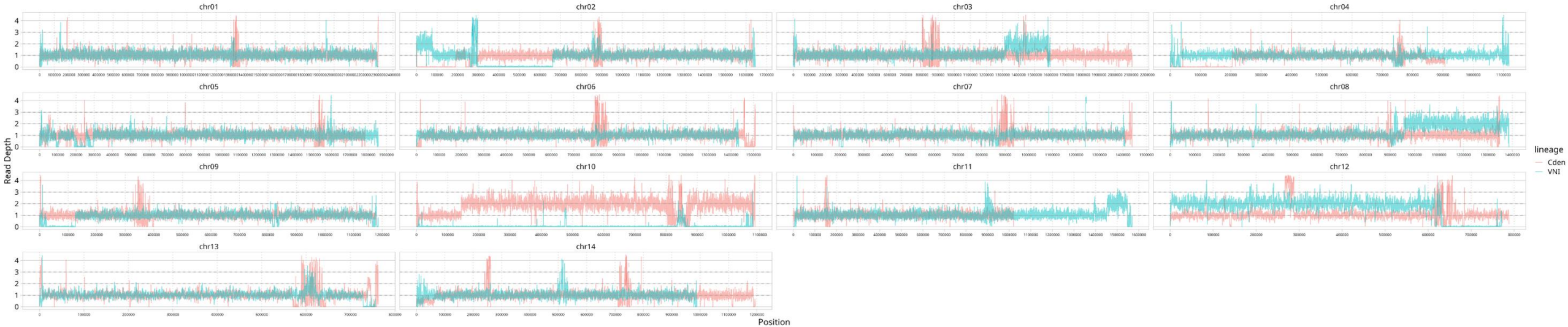
Genomic read depth analysis



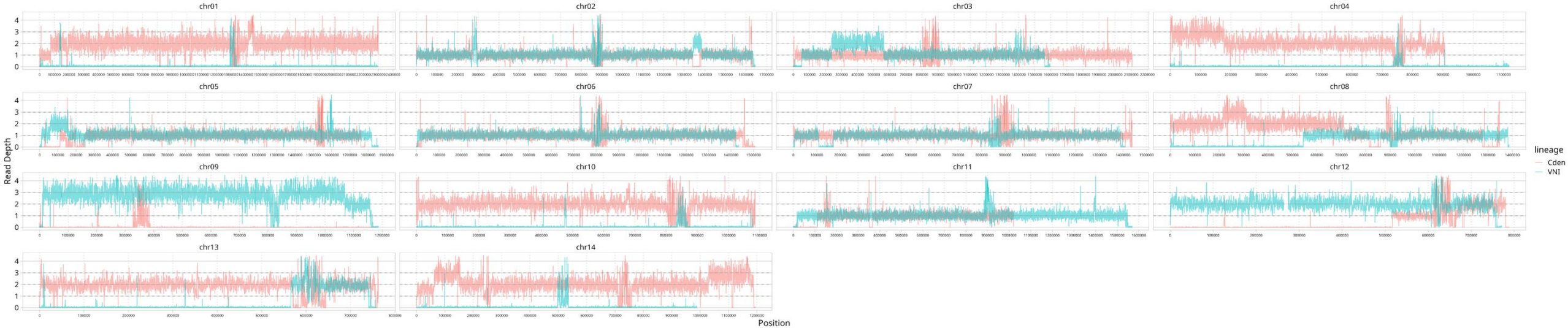

PMY3414
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
Read Depth
Position
lineage
VNI


PMY3414
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
Read Depth
Position
lineage
VNI

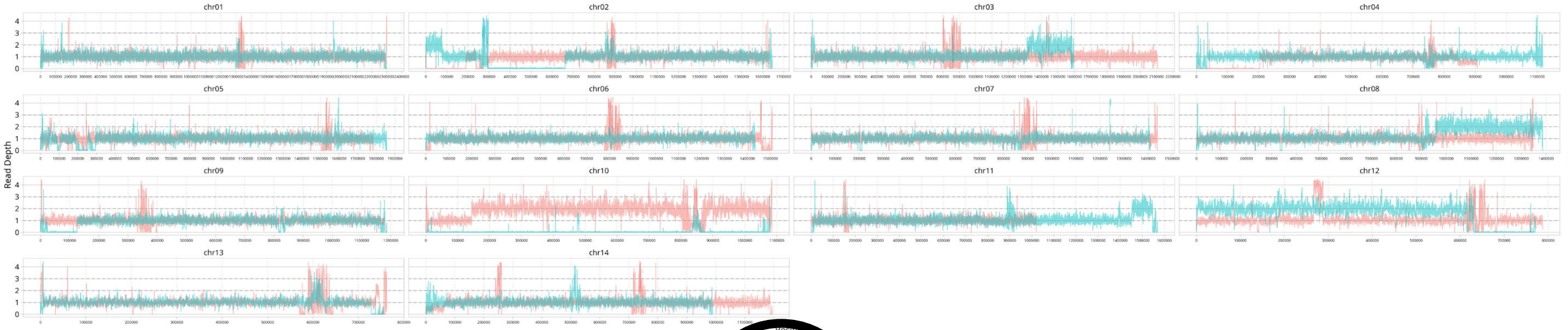
PMY3439



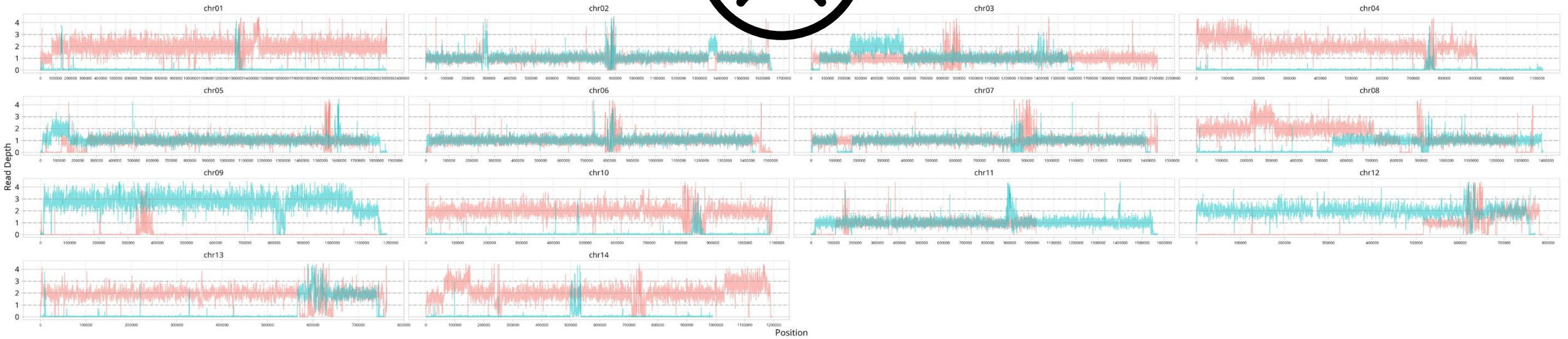
PMY3483



PMY3439

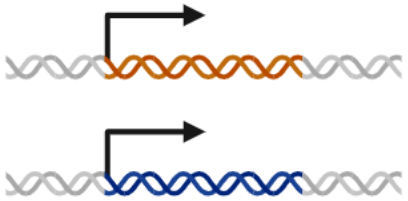


PMY3483

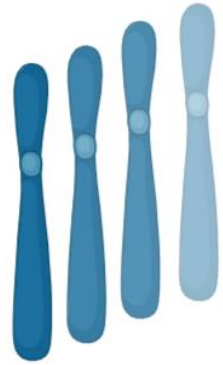


Main Questions

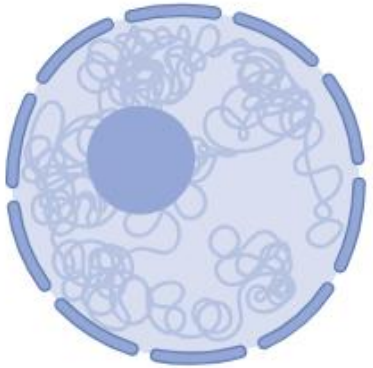
- 1) How are hybrid cryptococcus samples behaving transcriptionally?
  - Are parental haplotypes separately regulated?
  - Cis/trans effects?
  - Differences from synthetic hybrids?
  - Isoforms?



- 2) How do ploidy changes affect transcriptional regulation in clinical cryptococcus samples?
  - Linear increase in expression?
  - Differences in hybrids/nonhybrids?

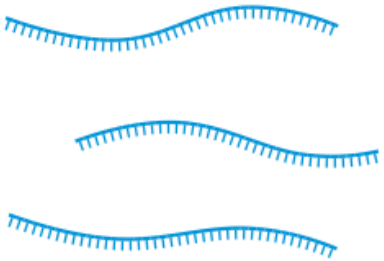


- 3) How do chromosomal rearrangements affect transcriptional regulation in hybrid samples?
  - Cis/trans changes?
  - Physical interactions?

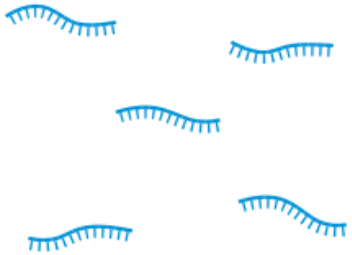


RNA-sequencing overview

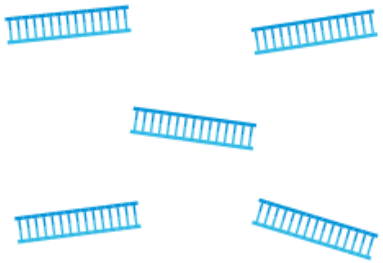
1 **Isolate RNA from samples**



2 **Fragment RNA into short segments**

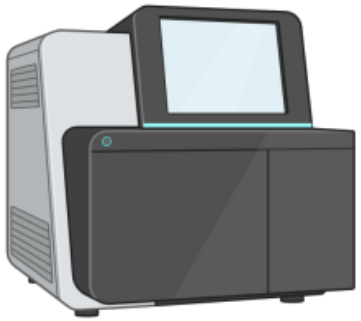


3 **Convert RNA fragments into cDNA**

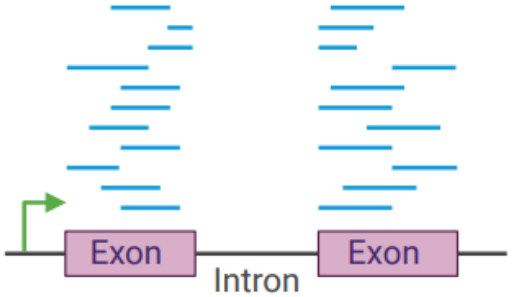


4 **Ligate sequencing adapters and amplify**

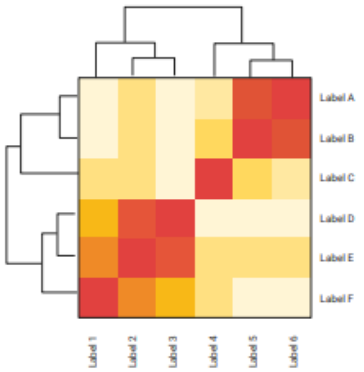
5 **Perform NGS sequencing**



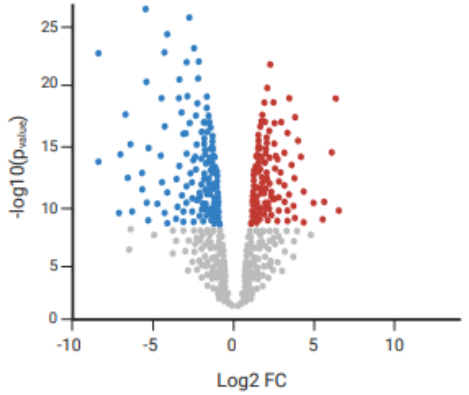
6 **Map sequencing reads to the transcriptome/genome**



7 **Correct and normalize transcript counts**

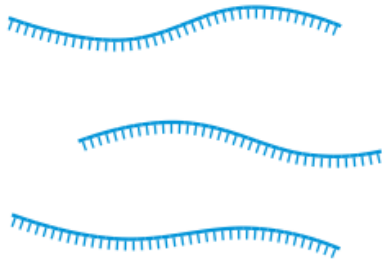


8 **Identify differential gene expression**

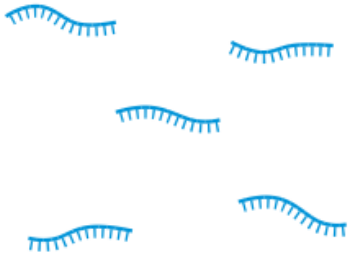


RNA-sequencing overview

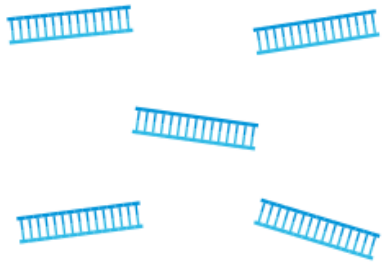
1 **Isolate RNA from samples**



2 **Fragment RNA into short segments**

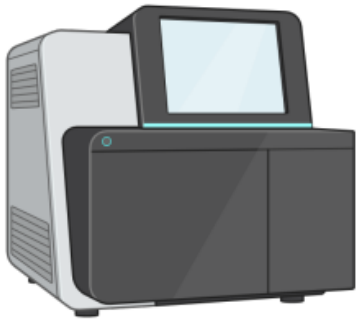


3 **Convert RNA fragments into cDNA**

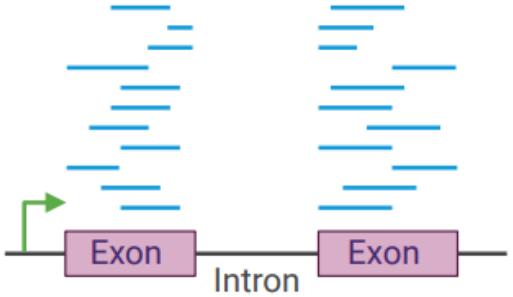


4 **Ligate sequencing adapters and amplify**

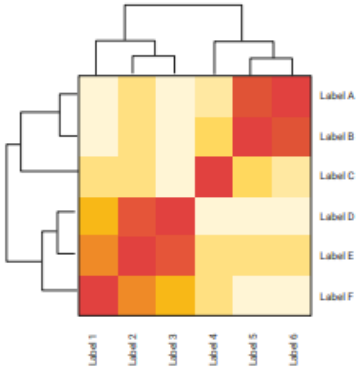
5 **Perform NGS sequencing**



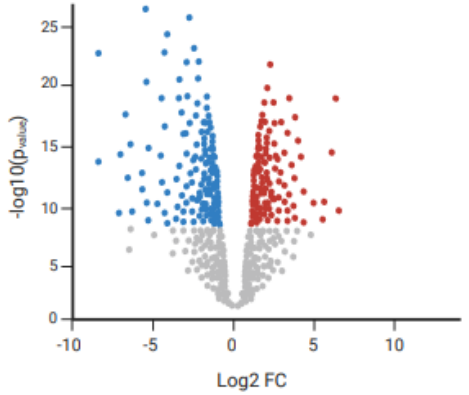
6 **Map sequencing reads to the transcriptome/genome**



7 **Correct and normalize transcript counts**



8 **Identify differential gene expression**



Analysis tools

- 1) How are hybrid cryptococcus samples behaving transcriptionally?
  - Quality control: **FastQC, Trimmomatic, Cutadapt, Skewer**
  - Read mapping/alignment: **STAR, Stampy, Bowtie2**
  - Differential expression: **DESeq2, edgeR**

- 2) How do ploidy changes affect transcriptional regulation in clinical cryptococcus samples?
  - Normalization: **RSEM, edgeR**

- 3) How do chromosomal rearrangements affect transcriptional regulation in hybrid samples?
  - de novo assembly: **Trinity, Trans-ABySS, SOAPdenovo-Trans, BUSCO, DETONATE**

Data formats

- 1) How are hybrid cryptococcus samples behaving transcriptionally?
  - RNA-sequencing
    - **fastq** raw reads and **fasta** reference genome
    - **SAM/BAM** files – alignment mapping
    - **GFF/GTF** files – gene annotations, isoform information
    - Output = **Count matrix (csv)**

- 2) How do ploidy changes affect transcriptional regulation in clinical cryptococcus samples?
- 3) How do chromosomal rearrangements affect transcriptional regulation in hybrid samples?
  - **Metadata (csv/tsv)** – hybrid status, ploidy estimates, known translocations
  - **Count matrix** and **Normalized count matrix**
  - Differential gene expression output = **Volcano plots, etc.**
  - Hi-C / methylation information?

Computing needs and workflow

- Many small genomes (19 Mb) Many intermediate files that can be very large (BAM/SAM, BED, GFF)
  - Both temporary and long term storage needed

- Workflow structure in **Snakemake**
  - **Conda** environments for software version consistency
  - Modular workflow + small genomes = not as much working memory needed
  - Already tested and works well on local cluster

Decisions

- Isoforms
  - May depend on sequencing platform and cost
    - Long + short reads?
  - Likely to be complicated in messy hybrids
  - May need preliminary data first

- Alignment of reads
  - Mostly reference-based
  - May also create de novo genome assemblies

- Differential expression
  - Looking for specific genes or general patterns?
  - What are the significance cutoffs?

Statistical challenges

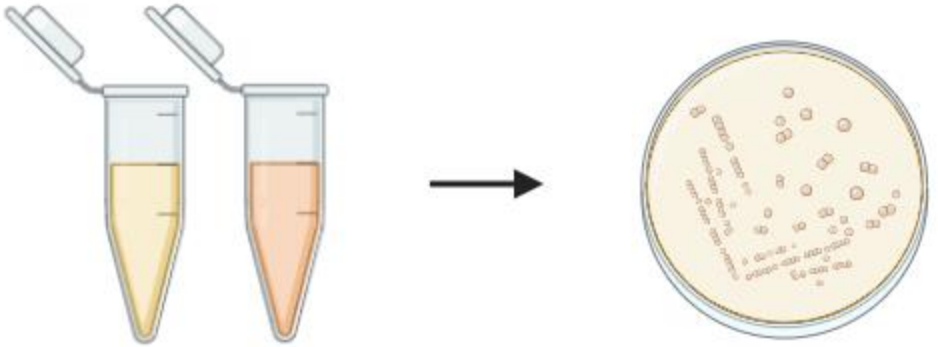
- Normalization and differential expression
  - How to deal with ploidy changes?

> "The **overall expression** of diploids was remarkably similar to that of the haploids in the same condition (**r=0.94, 0.93** for S. cerevisiae and S. paradoxus, respectively) and the **between-species differences** were also highly correlated between the haploids and diploids (**r=0.82**), suggesting that the results are not affected by ploidy level"

- Recombination + structural variation
  - Unknown area of research
  - May vary by strain and mating type

Statistical challenges

- Synthetic hybrids vs. clinical hybrids
  - Creating *in vitro* hybrids = better cis/trans calling, compare early/established hybrids
    - Both parents known and available
    - May not be stable or 'realistic'
  - Clinical strains may not have clear parental strains available
    - Could split haplotypes and treat as parents?
    - Could use closest sequenced strain to each haplotype?



Future Plans

- Testing on example data
  - Choosing software tools
  - Testing with/without parents included
  - Testing with imputed parents from offspring haplotypes

- Finding more information on current software
  - Isoform information
  - Cis/trans calling
  - Short + long read sequencing
  - Hybrid organisms

- Narrowing goals
  - Looking for specific genes (drug resistance)
  - Patterns temporally/geographically