Introduction to annotation

Jean Macklaim
Western University
19-Mar-2015

Annotation

- Genomics/transcripts: adding functional information to a predicted sequence (gene, ORF, transcript)
    - gene, ORF, transcript, 16S amplicon…
    - Table of differential features - what are they?
- Easy to do, easy to get wrong

How to do it

- Comparisons to what we know
    - Ideally: experimental evidence
        - Often: Circumstantial evidence by sequence similarity
- Databases for annotations
- Tools for annotations

Databases

- NCBI/EMBL
  - Whole and partial genomes
  - CDD, xfam (Pfam, Rfam), InterPro
  - GO, SEED, KEGG, IPA
  - Google…

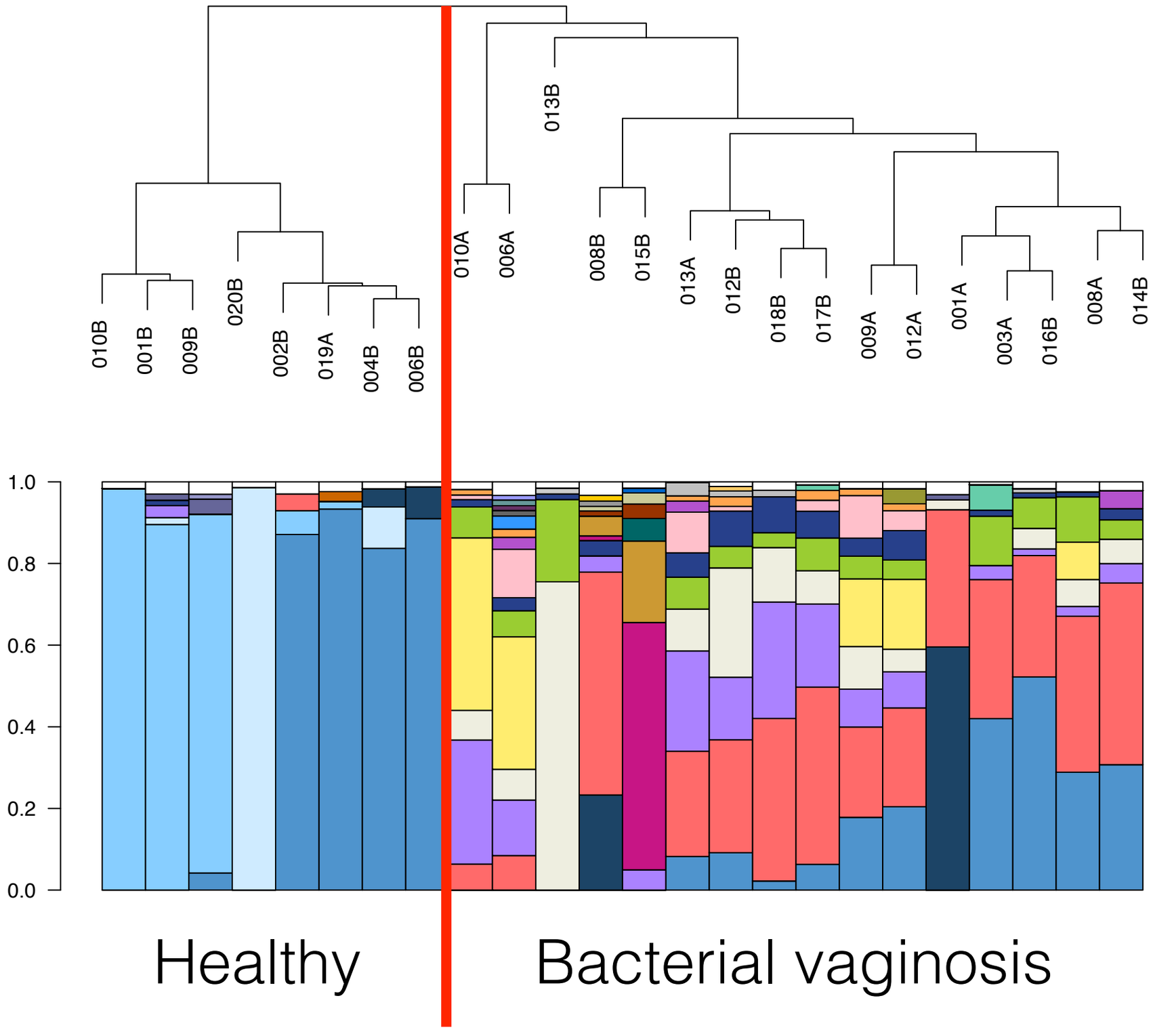
Tools

- Alignments: BLAST, HMMER (slow)
    - Assuming homology via sequence similarity
    - nucleotide vs amino acid comparisons
- W/out alignment: kmer (fast)
- Automated pipelines: MG-RAST, Galaxy, QIIME, DAVID
- Automatic, but need human validation
- Large amounts of data/atypical organisms (probably) require some scripting and command line

Example: bacterial meta RNAseq experiment

- Comparing transcript abundance from bacterial communities in vagina under normal conditions vs bacterial vaginosis

- 24 samples for RNAseq, paired with metabolomics (Amy McMillan)



- Estimate the population by 16S

How many ORFs

Reference genomes
51complete+250draft

500,640

Clusters
90%

300,311

Map reads
86,974

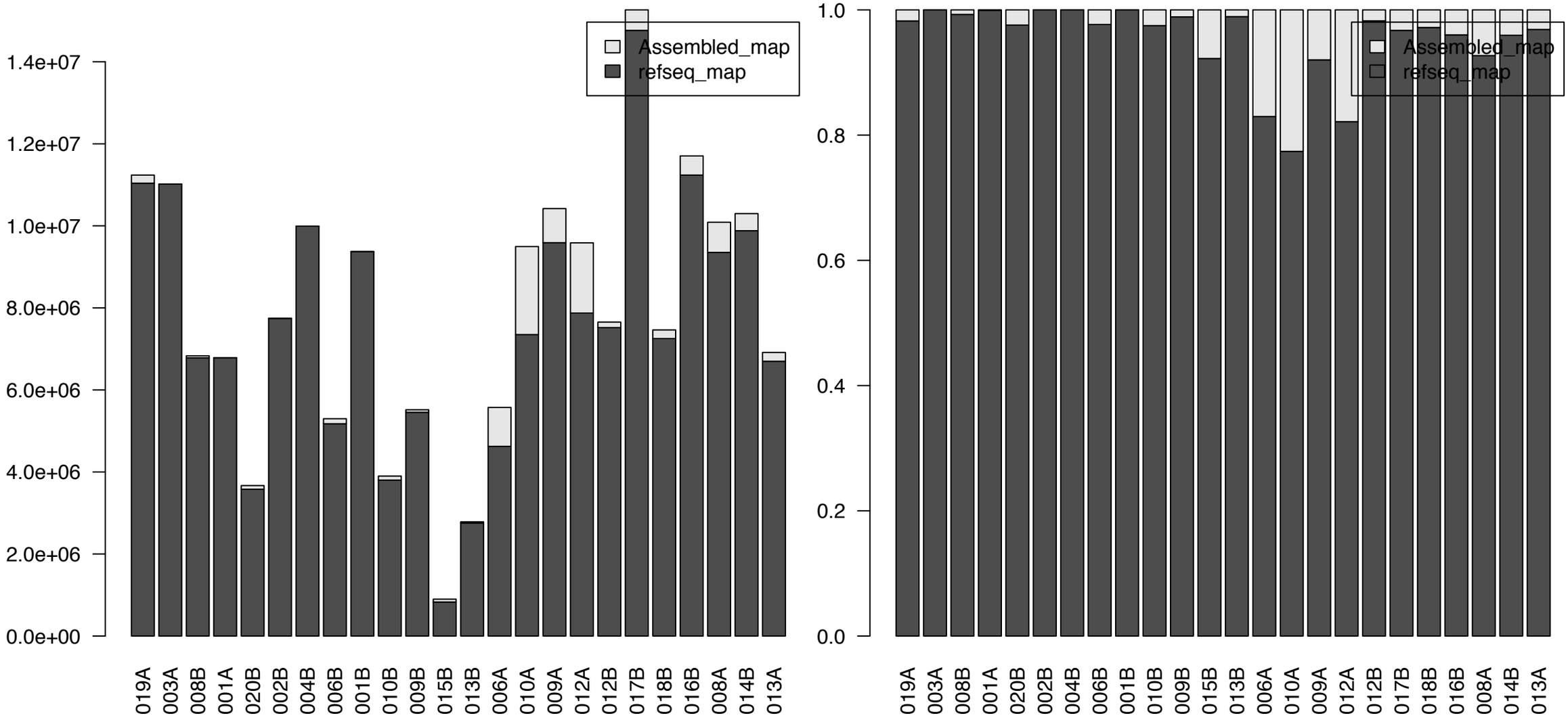
Assembled de novo
2,169

+

Annotated to SEED
47,716

- Reduce computation: remove redundancy

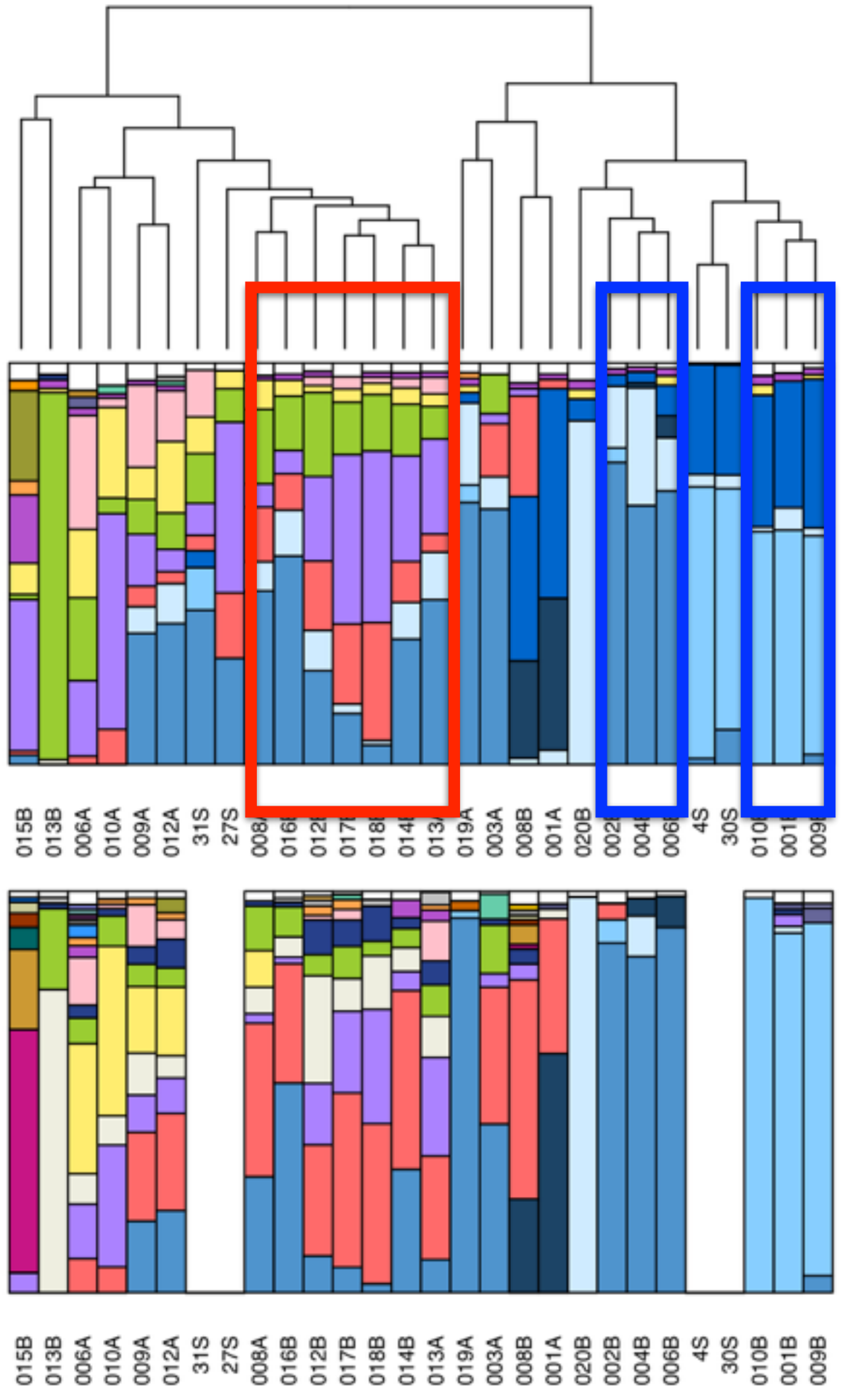
How many reads



- Total: 1 bil 200 mil (50mil per sample)
- Mapped: 7,519,347 +/- 3,167,315

mapped mRNA

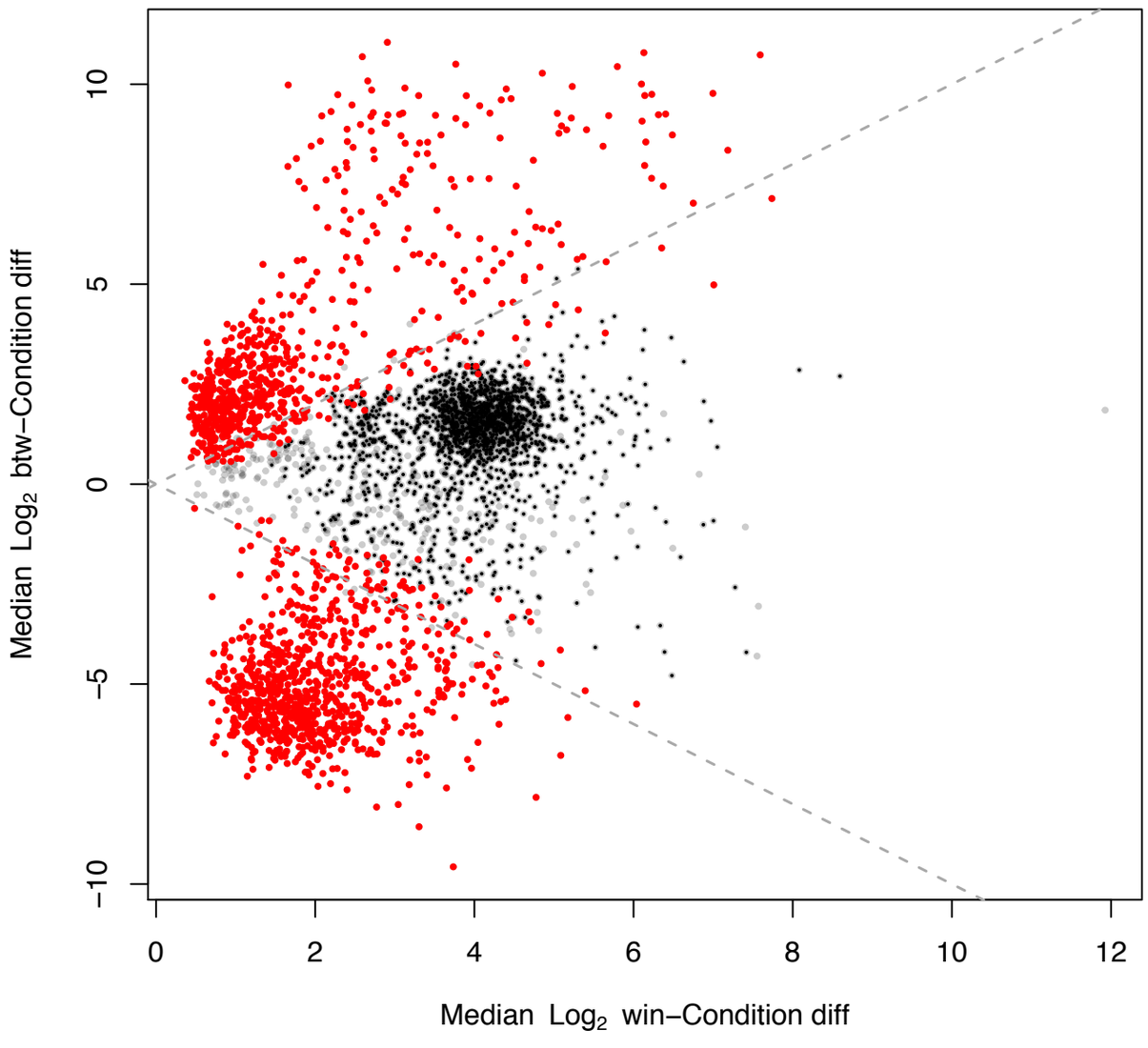
16S
(as before)

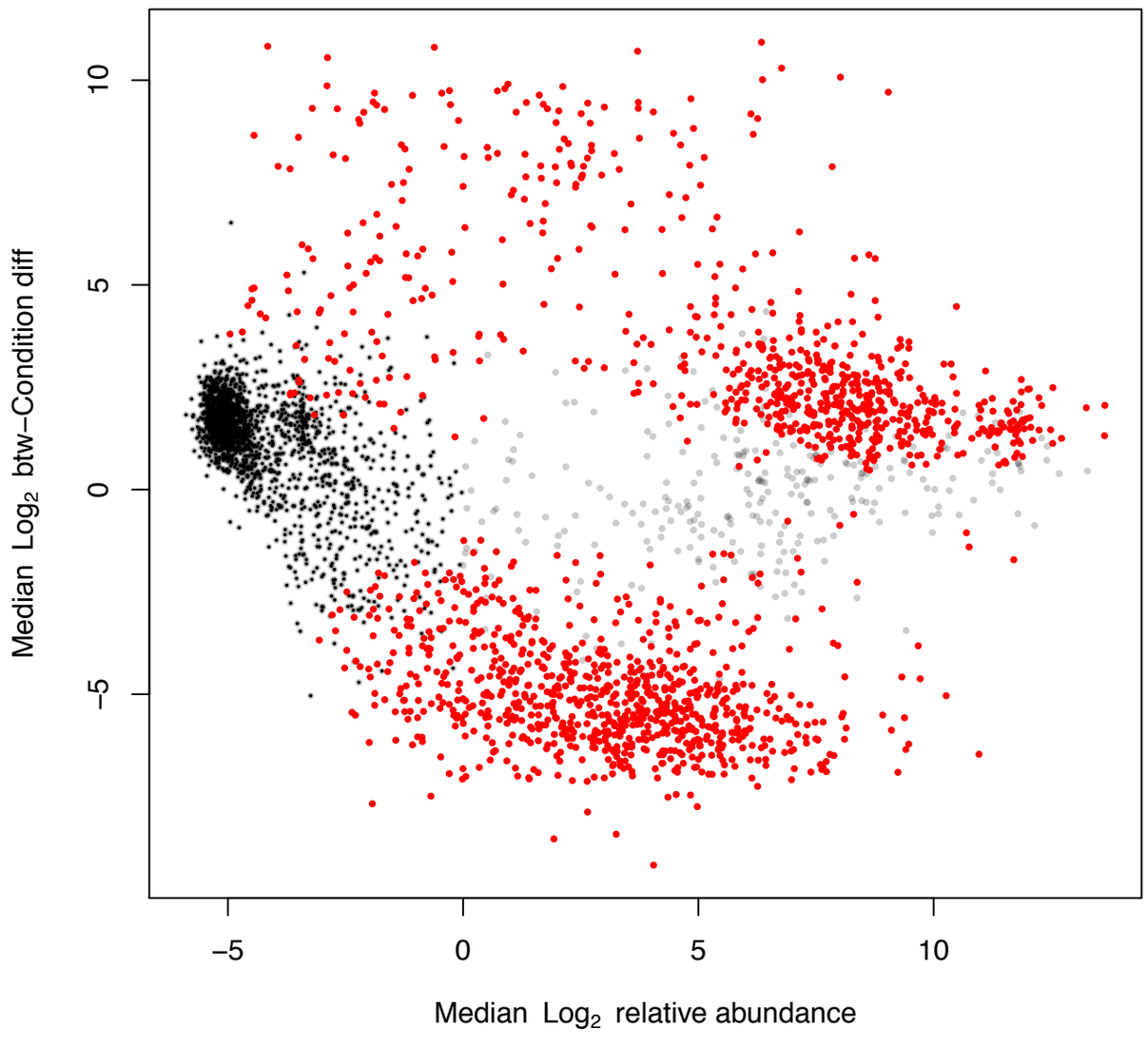


7x6
comparison

ALDEx2 output

- Have output, now what





Data: github.com/mmacklai

- Pick an online tool:
  - What's in the database you are comparing (and how did it get there)?
  - What method (brief explanation)?
  - What do you think the sequence is?

How to make the call

- Sorry…no hard rules. Err on conservative (10^{-6})
- Use multiples line of validation
  - New methods/databases will give different results
- Understand the evidence - don't take the output as gospel
- Curating one gene/coding sequence (or family of sequences) vs extrapolating to whole genomes/ metagenomes
  - Sacrificing quality for quantity

Grouping into categories

General

subsys1

subsys2

subsys3

Specific

subsys4
(SEED)

refseq
refseq
refseq
refseq

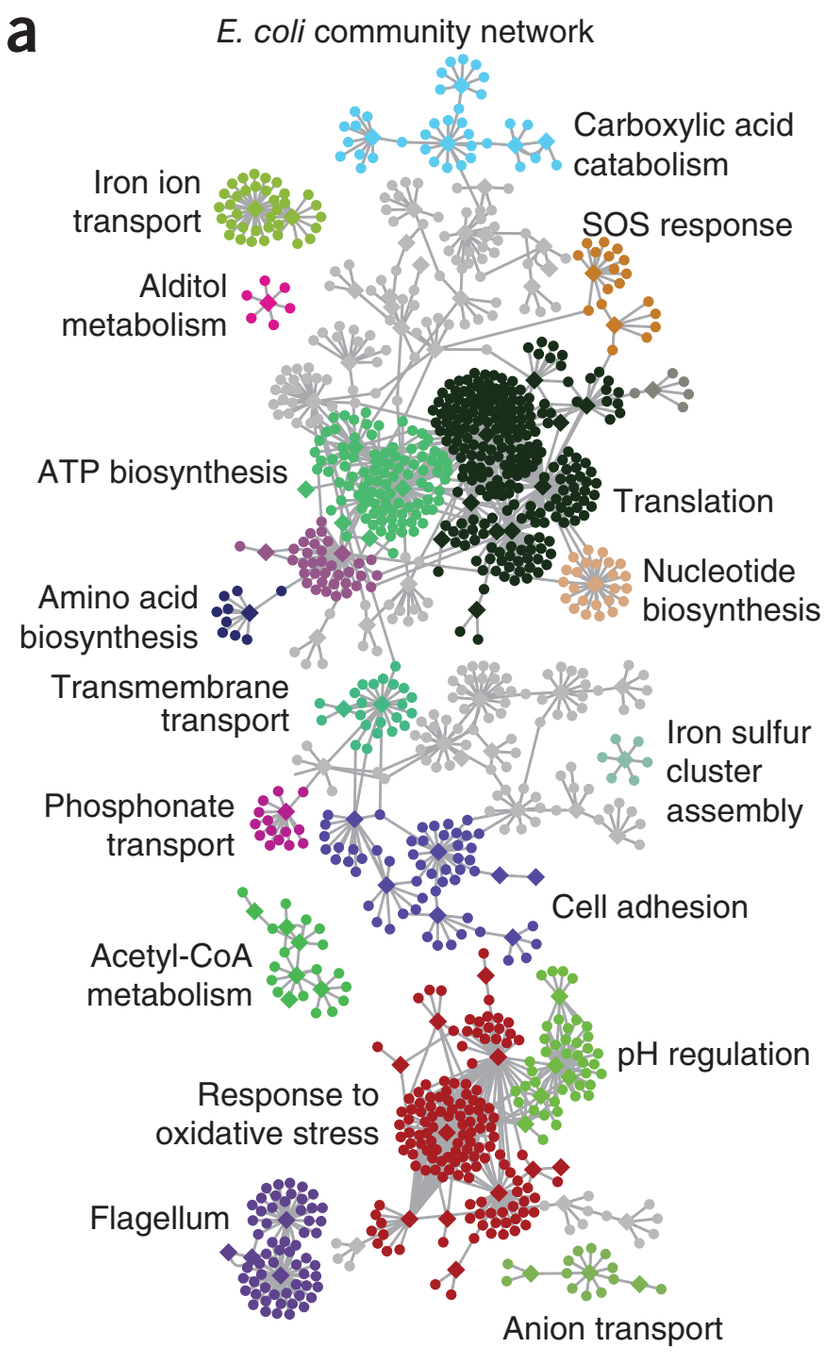
ORF
ORF
ORF
ORF
ORF
ORF

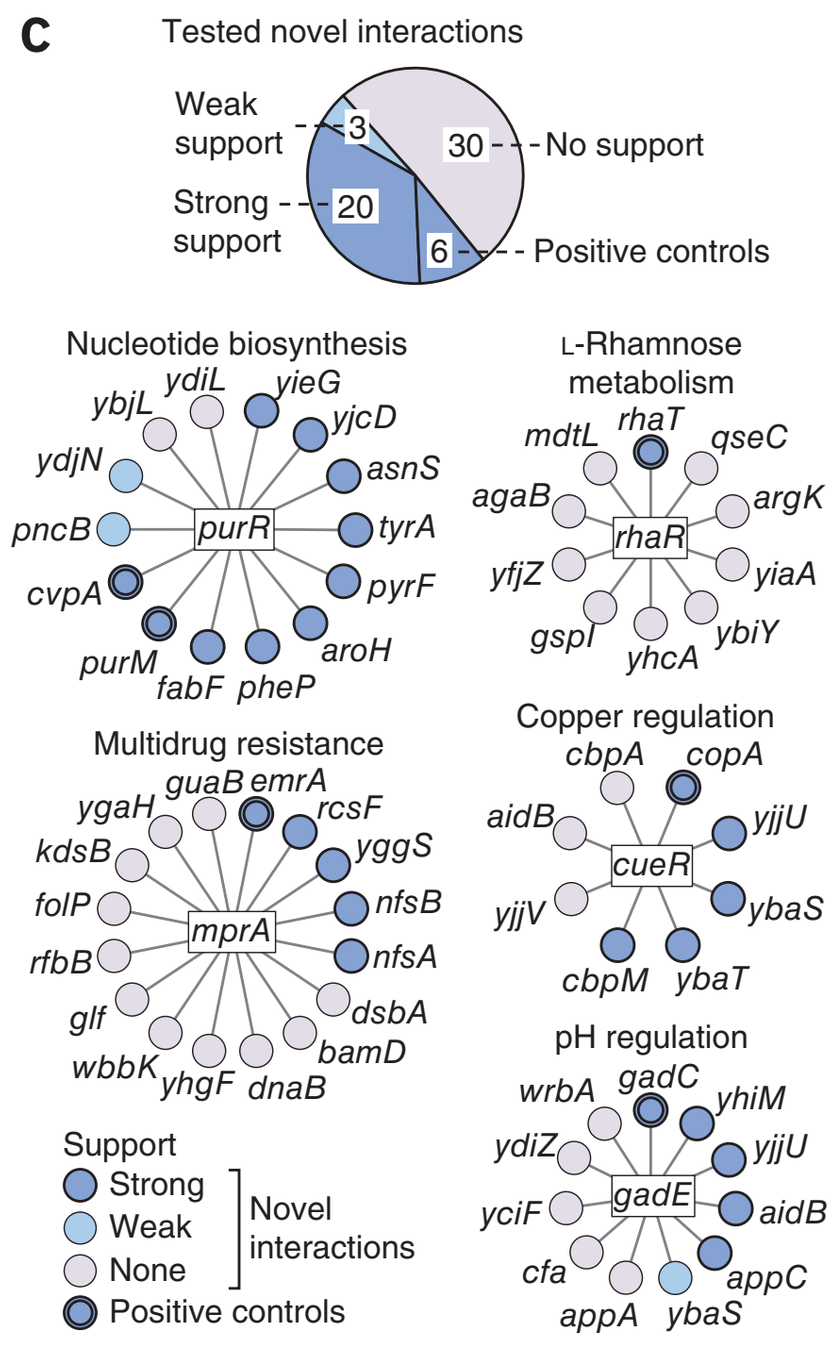
Genomes
+
de novo
assemblies

Exploratory analysis

- Summarizing and visualizing data (plotting)
- Enrichment analysis
- Network/correlations

Making connections





Nat Methods. 2012 Jul 15;9(8):796-804. doi: 10.1038/nmeth.2016.
Wisdom of crowds for robust gene network inference.

Exploratory analysis

- NOT quantitative, but descriptive - be careful about drawing conclusions you haven't explicitly tested
- Easy to mislead or be misled
    - Texas sharpshooter fallacy…

Reporting data

- Sequence data should be publicly available
  - Raw reads: SRA…?? QIIME-DB, MG-RAST
  - Reads count table
- Fully report methods: include all parameters, versions, cutoffs/thresholds
  - Ideally: post your script workflow (commands you used)

Any reader should be able to reproduce your results