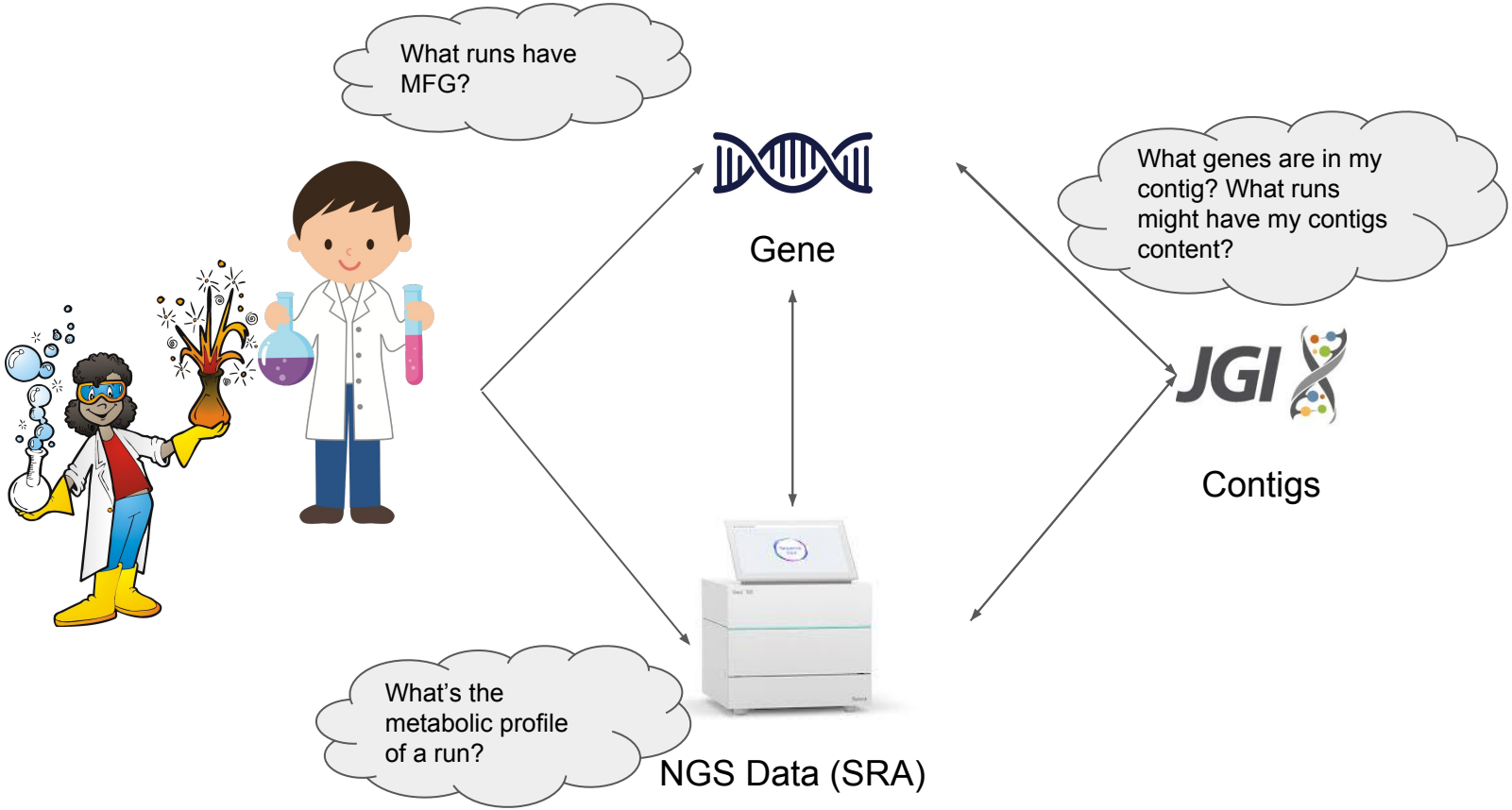
Bytes to Biology: Petabyte Scale Search of the Sequence Read Archive

What's in my tube? What tubes have MFG?

Workflow and Use-Cases

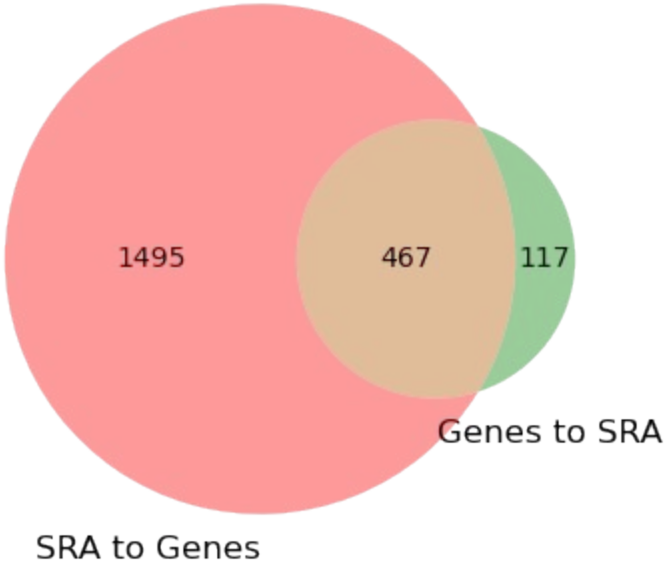


Background

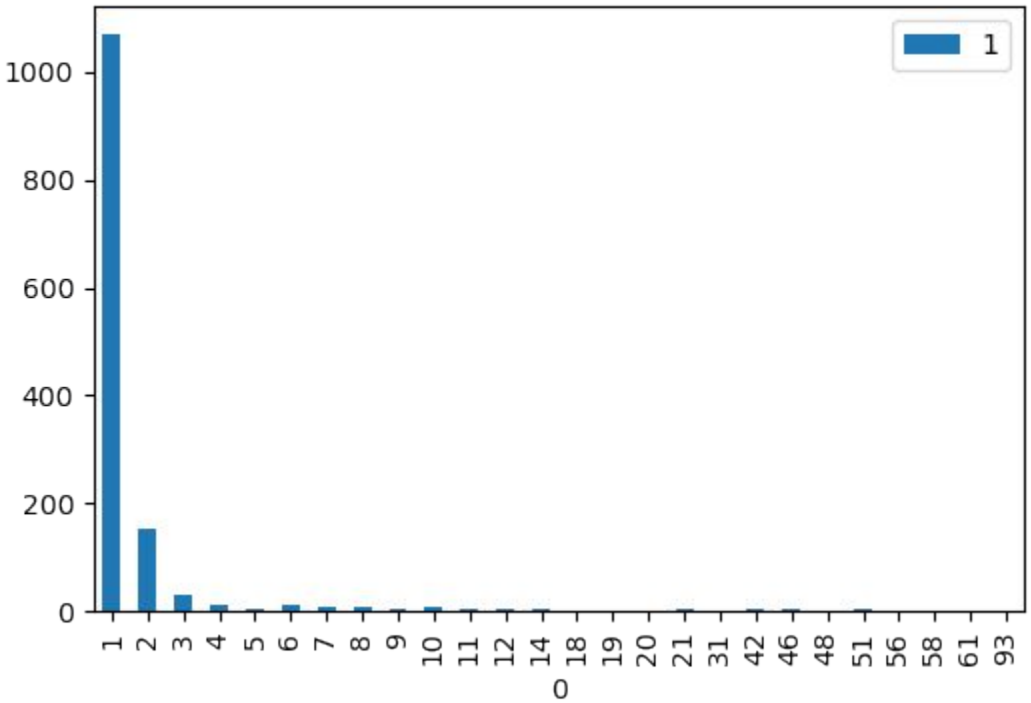
- NCBI Pebblescout
  - Decompose sequence queries to search a kmer-based index of SRA runs (metagenomic only)
- JGI Contigs
  - Assembled from metagenomic samples using diverse methods
- The Dataset
  - 24 runs with ~83M JGI contigs
  - 3 PacBio runs, others are Illumina
  - Mapped to 552 Genes (of over 2kb)

Results

- Overlap between Gene>SRA and SRA>Gene results
  - 1 Acc (SRR5165157), Bacterial CDS > 2kb

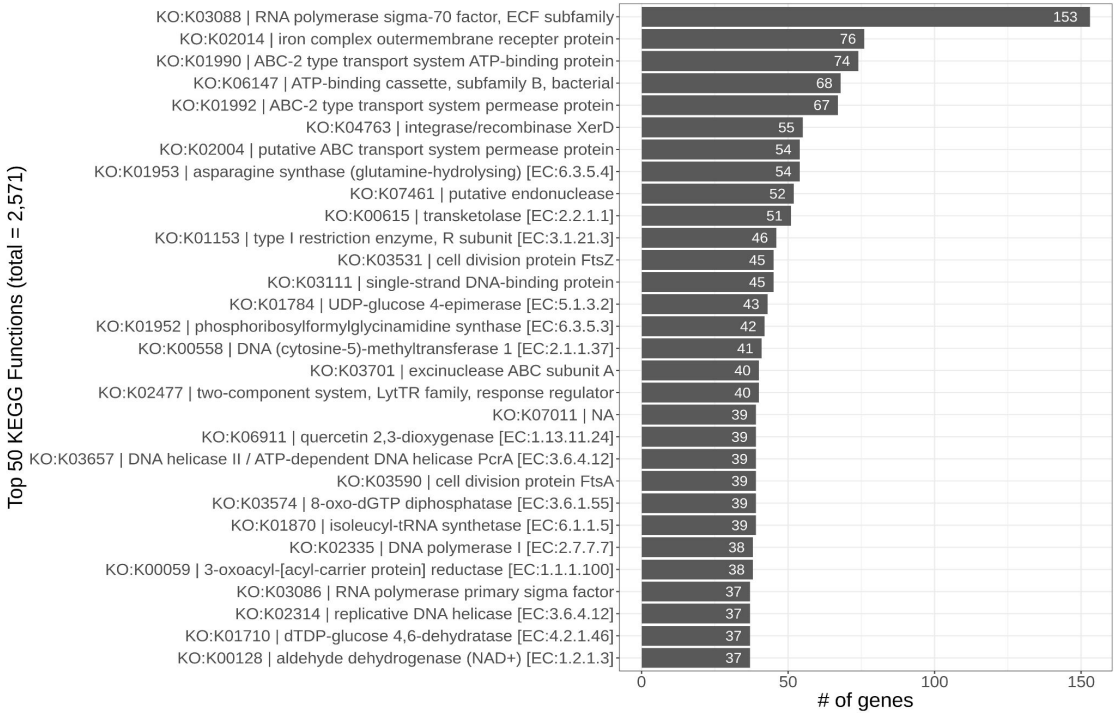


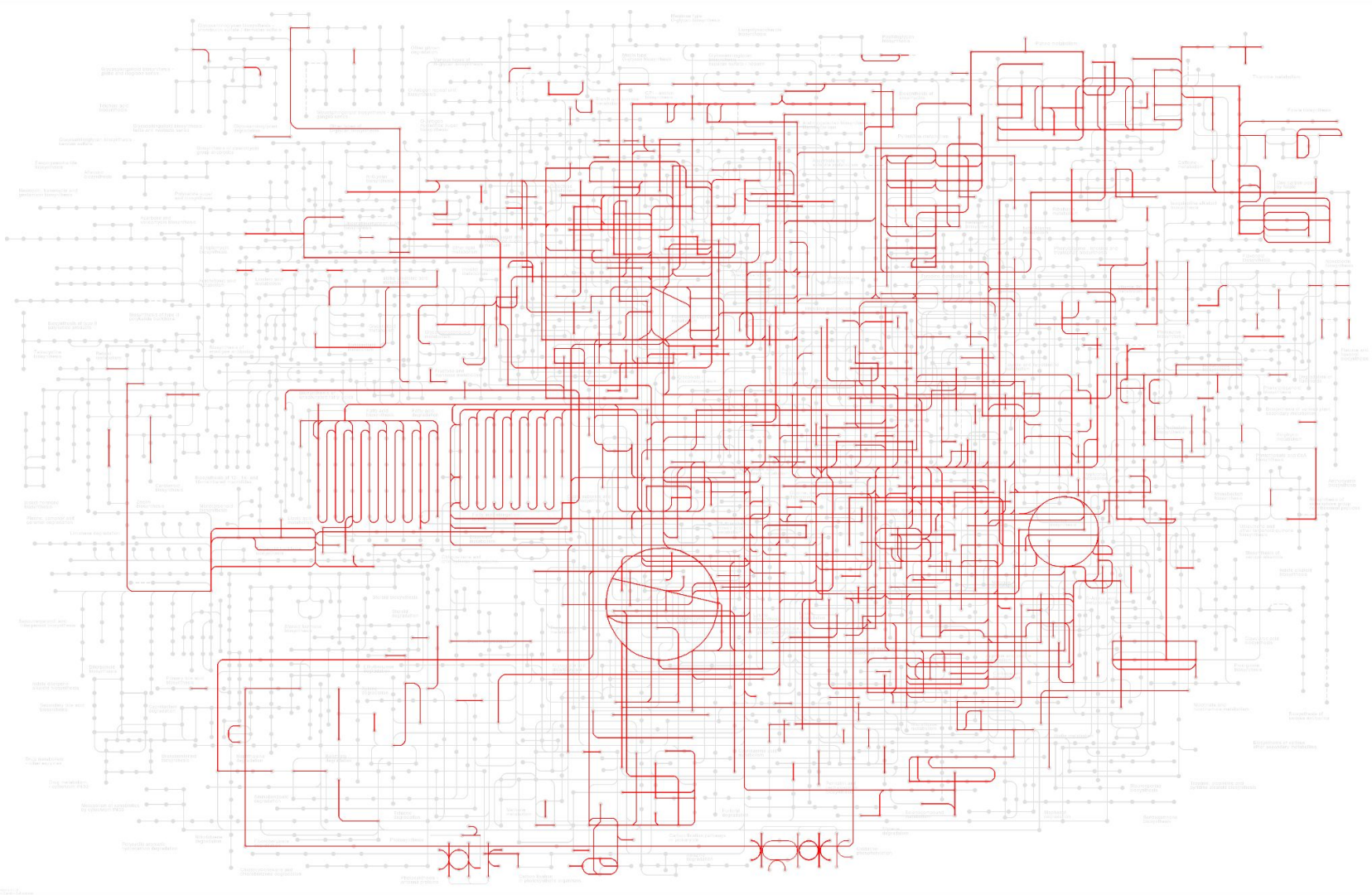
Contigs to CDS Mapping



Results

Metabolic Analysis





Future Directions

- Run on a larger dataset
- Investigate discrepancies (venn diagram)
- Collab some more!