Metagenome assembly… in the cloud!!!

C. Titus Brown

Adina Howe

Michigan State University

Why assembly?

1. Significantly increases signature for homology searches (BLAST, HMMER).

2. Can dramatically reduce data set size.

3. Not dependent on nearby reference.

4. Long-range correlations (operons, etc.)

   (Eventually, whole genomes from metagenome WGS?)

Why *not* assembly?

1. Fairly strict coverage cutoff (below ~2-5x little assembles)

2. Unknown effect of strain variation on sensitivity of assemblies.

Apart from that, **with our tools**, we get sensitive recovery of spiked-in genomes and highly specific contigs.

The practical barrier - memory

For even relatively small data sets, metagenomic assemblers scale poorly.

Memory usage ~ number of errors

Number of errors ~ size of data set

Size of data set == big!!

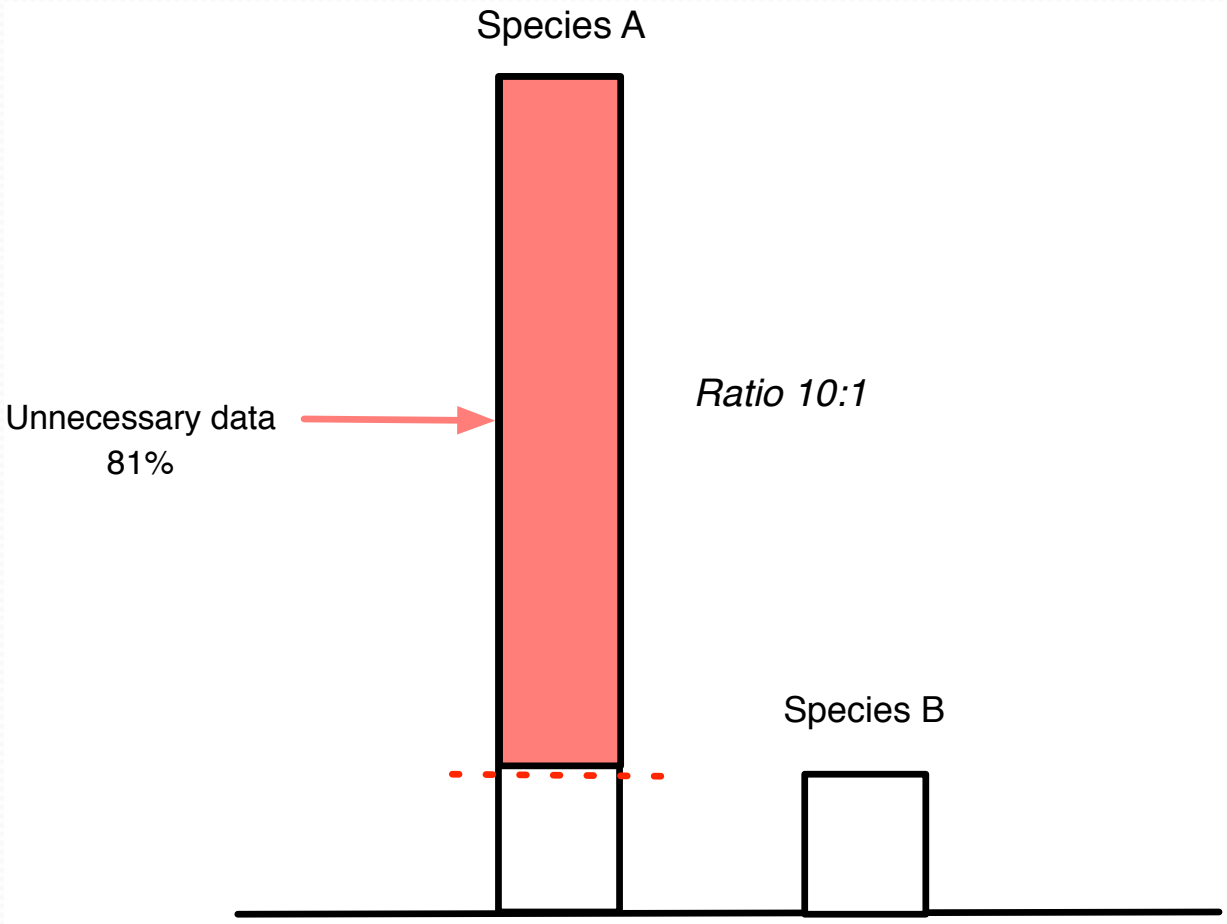
This is the problem that we have (mostly) solved.

Two basic techniques

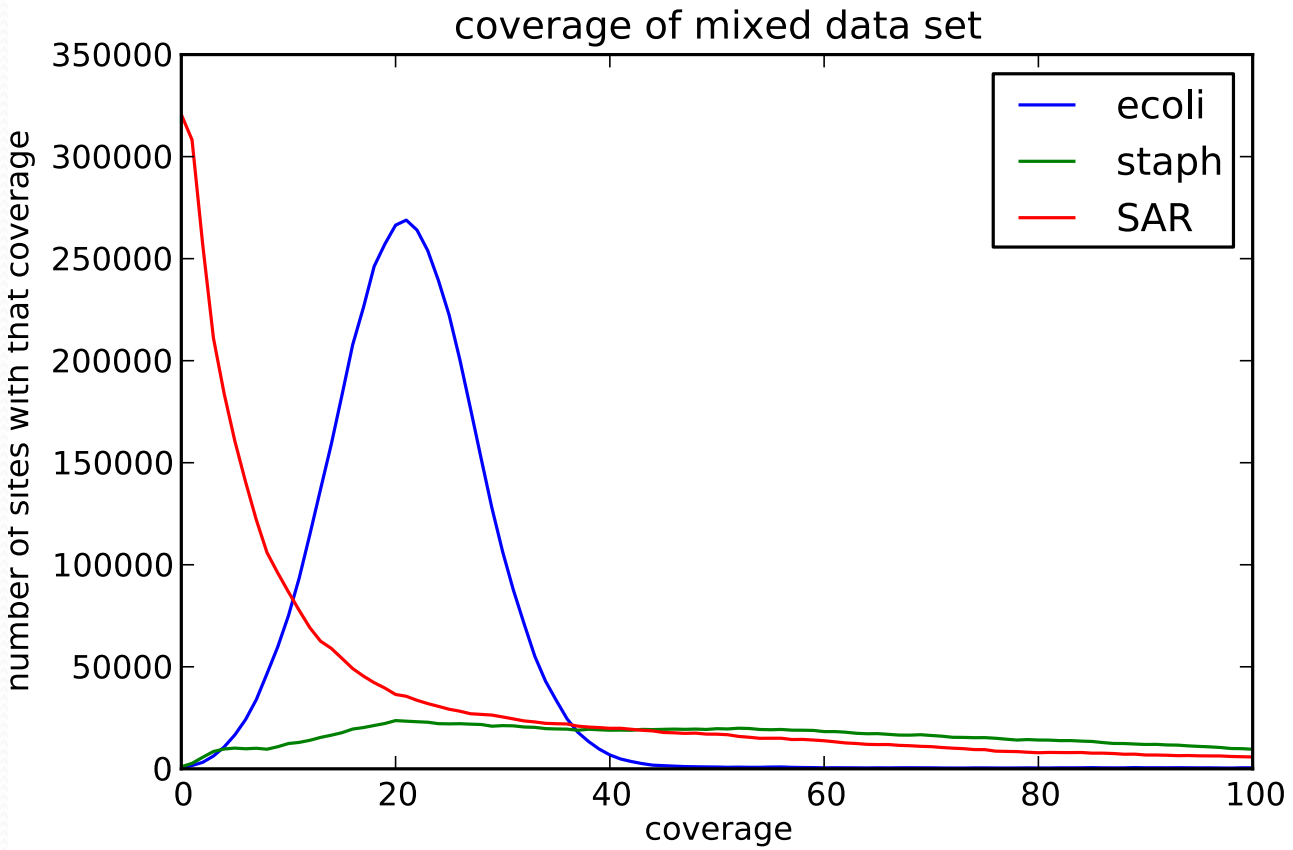
Digital normalization

Partitioning

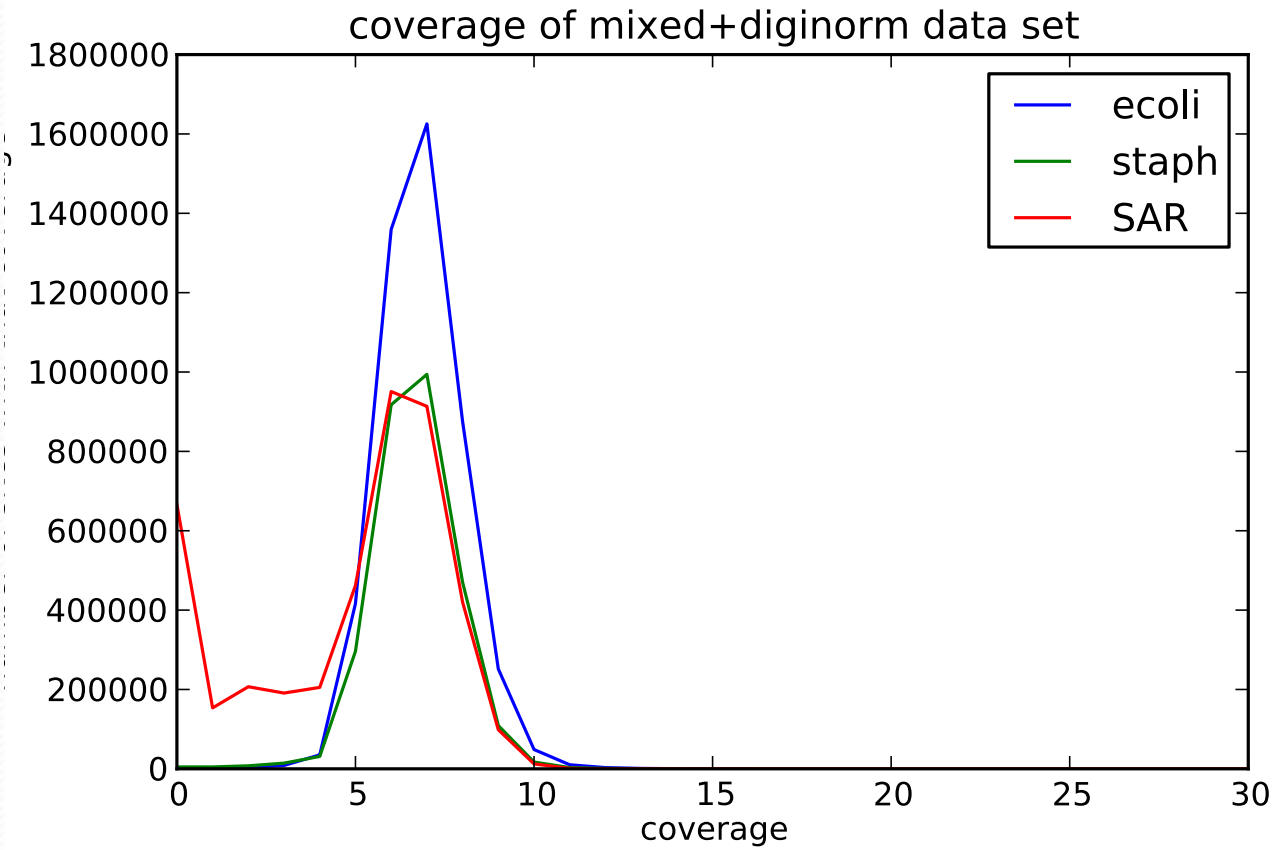
Digital normalization



Coverage before normalization

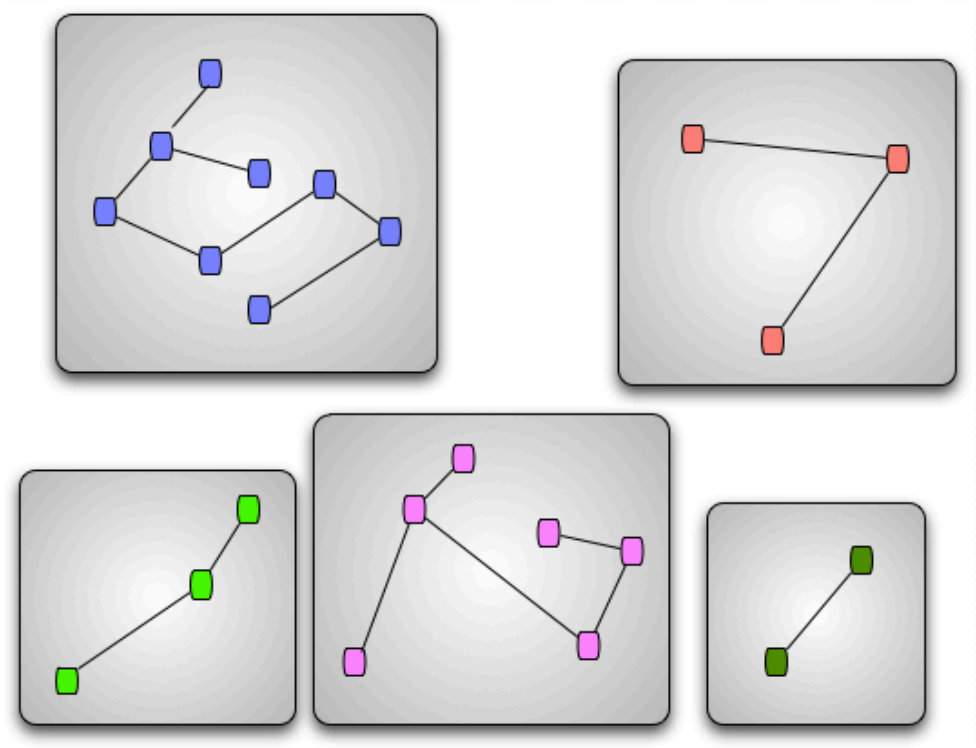


Coverage after normalization



Partitioning

Split reads into "bins" belonging to different source species.



Diginorm + partitioning make small-memory assembly possible.

- Digital normalization is fixed memory, single pass (streaming, online), scales with diversity of underlying sample.

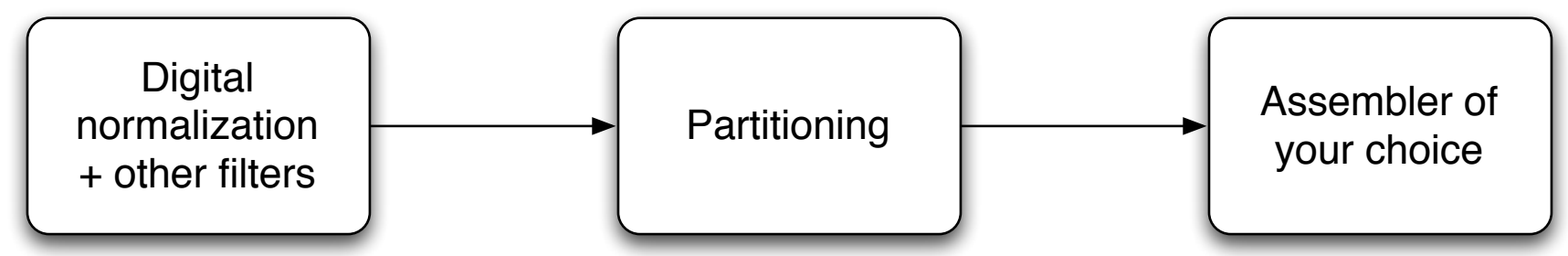
  Note: Diginorm applies equally well to (meta)transcriptome, genome, MDA single-cell.

- Partitioning is ~20x lower memory usage than assembly, and following assembly steps are guaranteed to be <= unpartitioned assembly. After partitioning, remaining steps are pleasantly parallel & small memory.

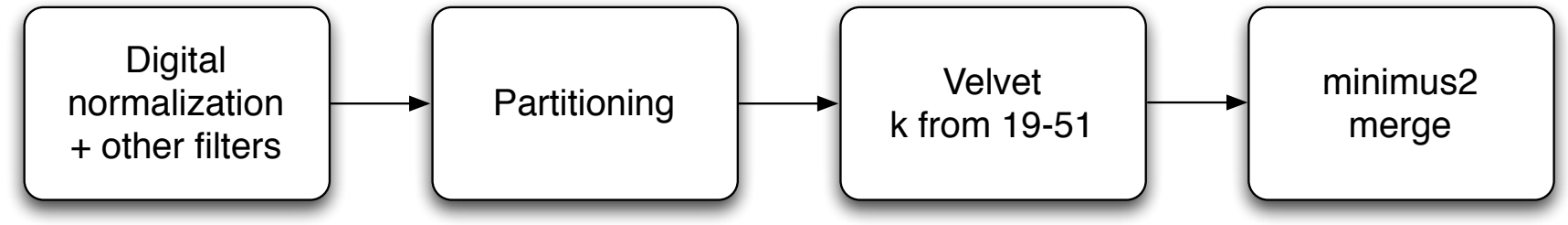
- We (Adina :) can assemble 3bn soil reads in < 300 GB of RAM; requires 3 TB of RAM using Velvet.

- We know how to scale metagenome assembly *arbitrarily*: "just" engineering at this point.

Pipeline options



Below is what *we* do:



Example

Dethlefsen data set / Relman lab

251 m reads / 16gb FASTQ gzipped

~ 24 hrs, < 32 gb of RAM for full pipeline

(reads => final assembly + mapping)

Assembly stats:

58,224 contigs > 1000 bp (average 3kb)

summing to 190 mb genomic

~38 microbial genomes worth of DNA

~65% of reads mapped back to assembly

Why would you want to assemble?

Some use cases:

- Look for large-scale variation from reference – pathogenicity islands, etc.
- Assemble new “reference”.
- Discriminate between different members of gene families.
- Discover operon assemblages & annotate on co-incidence of genes.