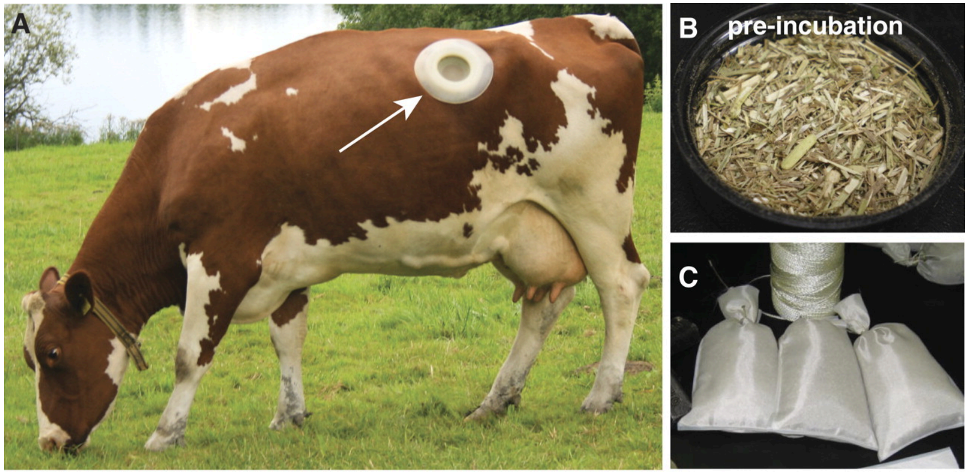
JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

MetaBAT: An Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities

Don Kang

Metagenomics at Gene or Pathway Levels



20,000 New Cellulase **Genes**

Expression of Methanogenesis **Pathways**

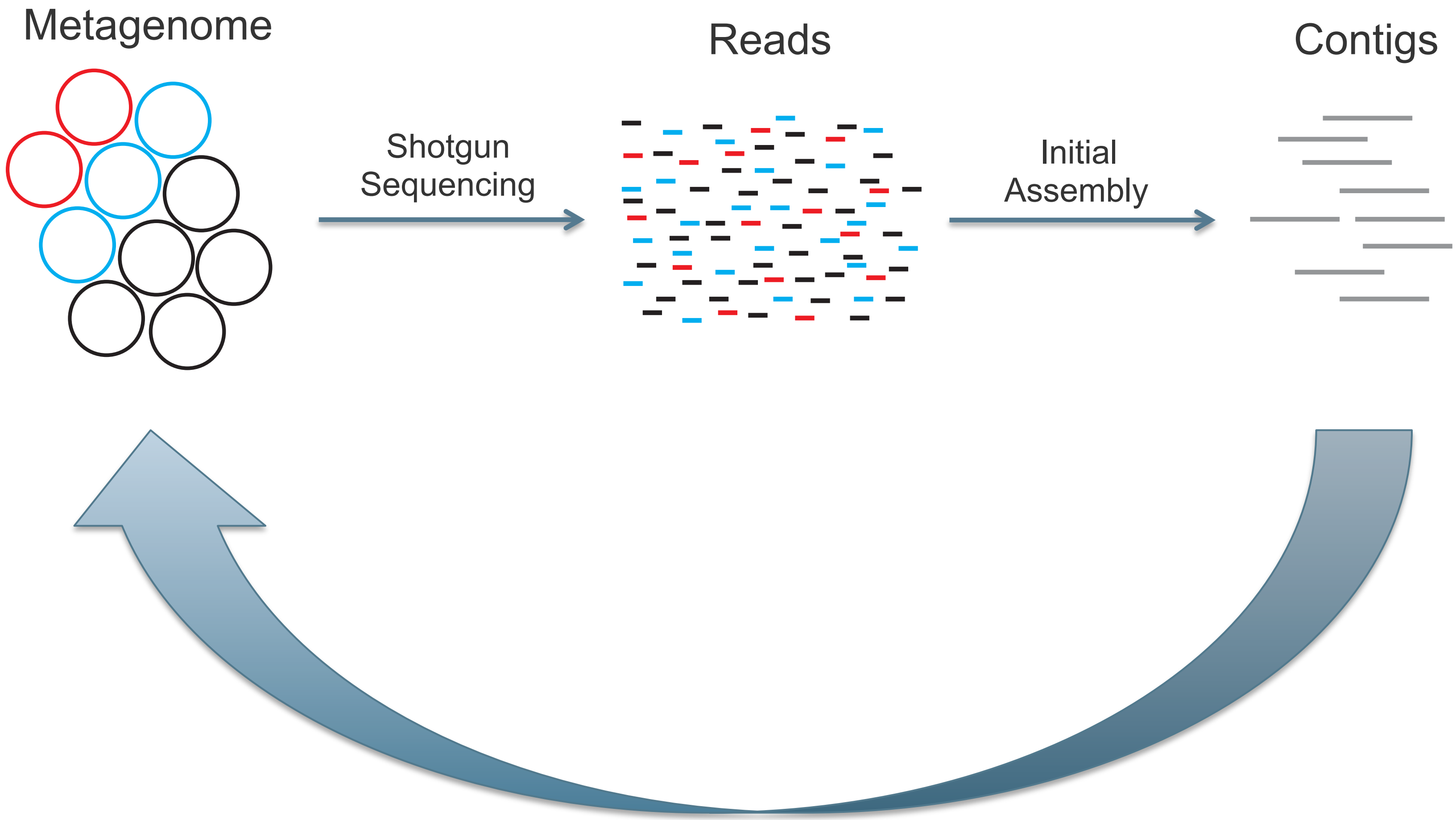
Hess et al, 2011, Shi et al, 2014

But, Genomes would be Better!

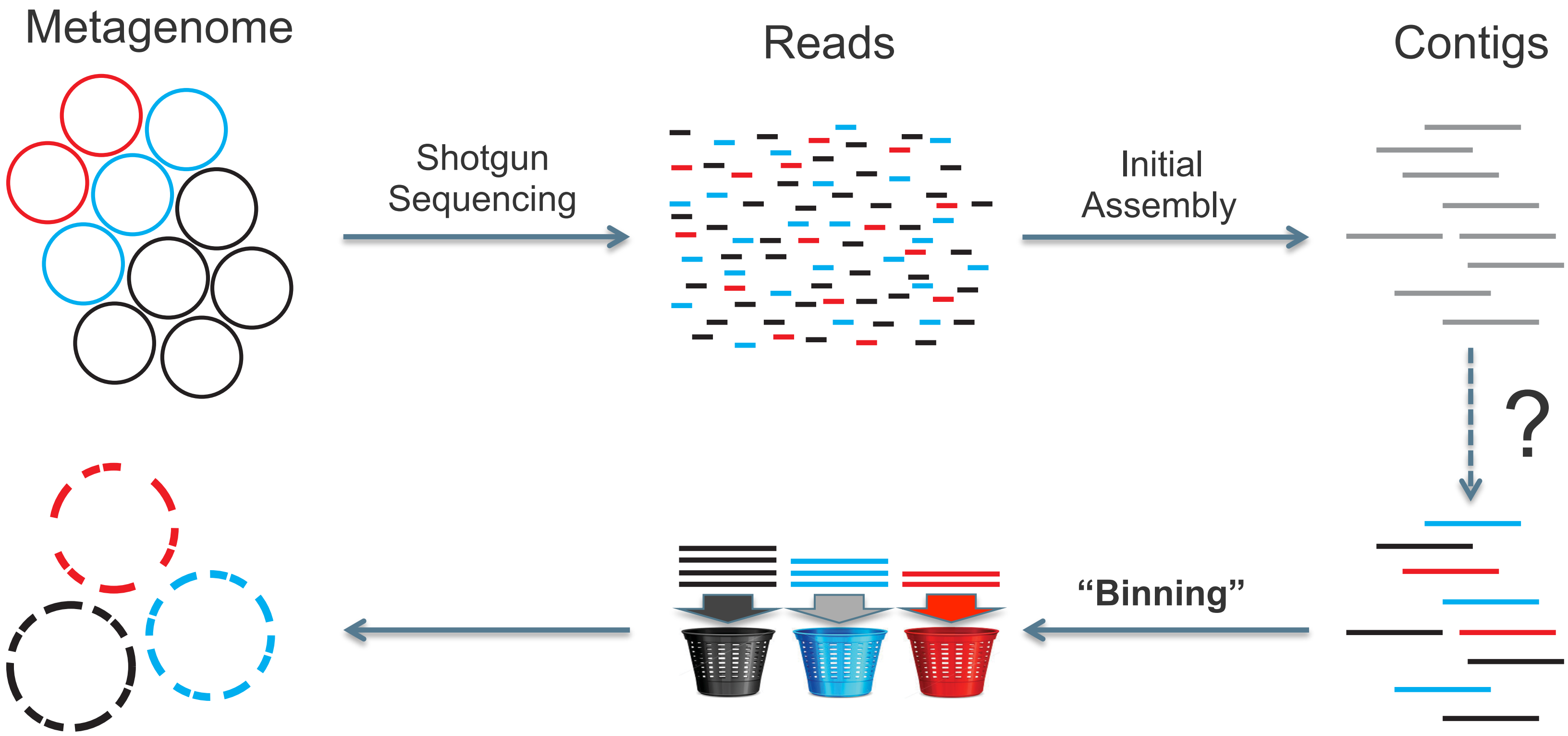
- **Able to get a full picture of metabolic capacity of an individual member of the community**
- **Study genome dynamics of individual members**
  - Genome-wide sweep, gene gain/loss analysis
- **Understanding inter-species interaction**

How can we construct single genomes from metagenomic data?

Genome Reconstruction from Metagenomic Data



Genome Reconstruction from Metagenomic Data



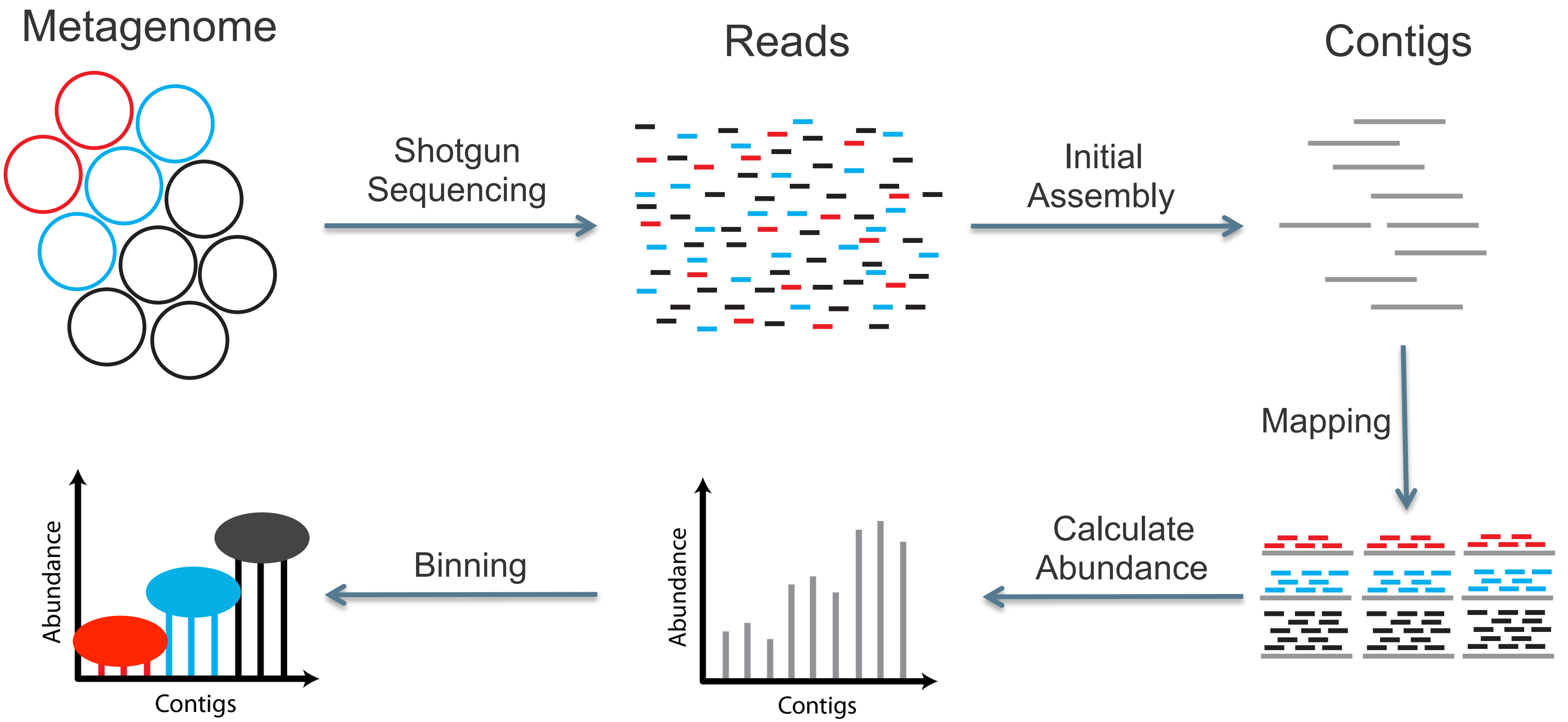
Existing Binning Methods

- **Reference Based Binning**
  - Phylogeny based
- **De novo Binning**
  - Sequence composition
  - Abundance
  - Both

➢ Inaccurate for complex metagenomes
➢ Manual
➢ Not scalable for many samples

Co-Abundance (coverage covariance) Binning

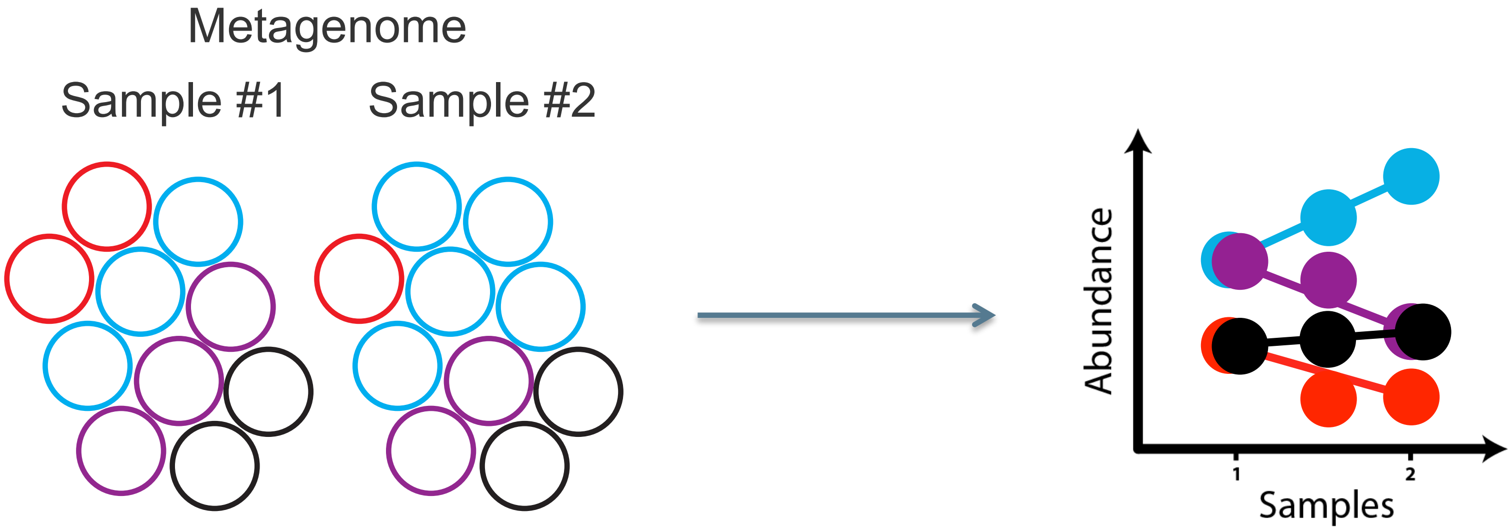
Abundance (Coverage) Binning



Ideally, contigs from the same genome should have the same coverage.

But, single abundance cannot differentiate multiple genomes of similar abundance?

Co-abundance Binning



Multiple samples (libraries) help to differentiate the similar abundance in single sample (library).

Design Goals for Binning Software

- **Automated Unsupervised Co-abundance Binning**
  - Integration of tetranucleotide frequency (TNF) and (or) abundance (ABD) as features
  - Handling of multiple ABDs from samples
- **Highly Efficient**
  - A couple of hours to bin millions of contigs having thousands of samples
  - Runnable in a single node (<20G memory)
- **Reproducible and Reliable**
  - Robust to noise in contigs or samples
  - Designed to have high specificity than sensitivity
- **Flexible**
  - Handle any number of samples
  - Adjustable parameter setting to change sensitivity and specificity
- **Simple**
  - Easy to run and fully automated

Run MetaBAT!

```
runMetaBat.sh  assembly.fasta  *.bam
```

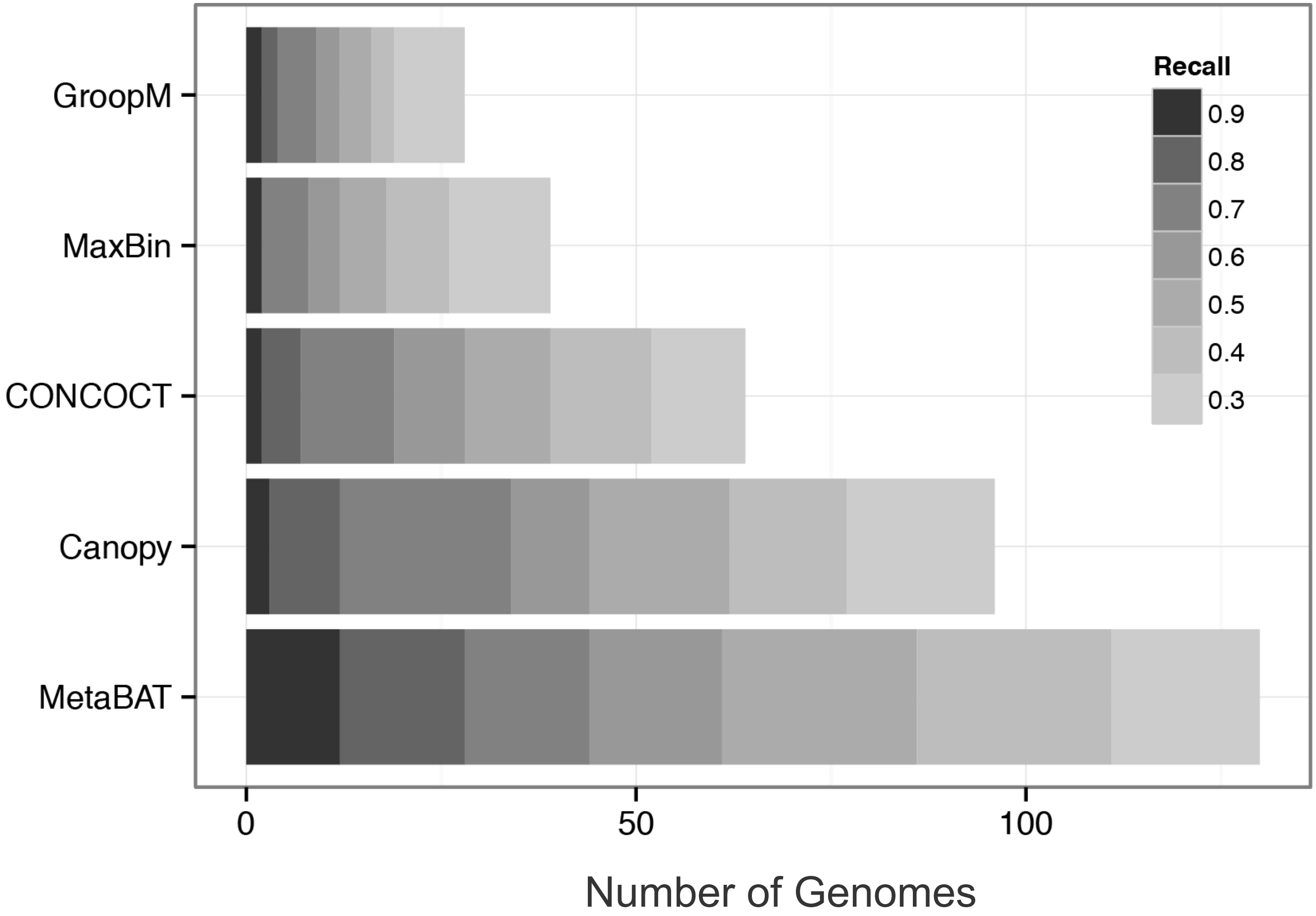
Benchmarks of Automated Metagenome Binners With A Medium Sized Data Set

- **5 binning methods**
- **264 human gut metagenomic samples (ERP000108)**
  - Assembled into 200K contigs
  - Used a method (CheckM) to estimate completeness and precision based on single copy genes

The Contestants

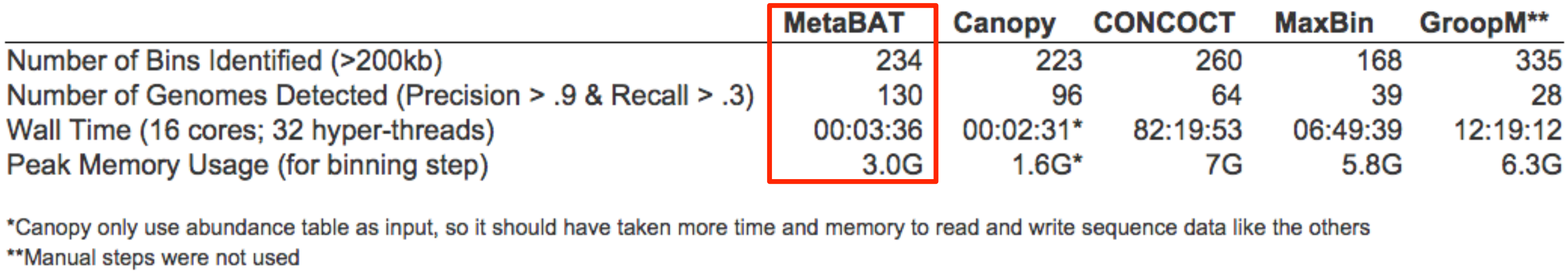
- **MetaBAT**
  - Sequence composition (TNF) + Co-abundance
- **CONCOCT**
  - Sequence composition + Co-abundance
- **GroopM**
  - Sequence composition + Co-abundance
  - Optional manual steps
- **MaxBin**
  - Sequence composition + Abundance
- **Canopy**
  - General purpose clustering algorithm
  - Co-abundance only

MetaBAT found the most genomes



For details, refer to *https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT*

MetaBAT runs very efficiently



	MetaBAT	Canopy	CONCOCT	MaxBin	GroopM**
Number of Bins Identified (>200kb)	234	223	260	168	335
Number of Genomes Detected (Precision > .9 & Recall > .3)	130	96	64	39	28
Wall Time (16 cores; 32 hyper-threads)	00:03:36	00:02:31*	82:19:53	06:49:39	12:19:12
Peak Memory Usage (for binning step)	3.0G	1.6G*	7G	5.8G	6.3G

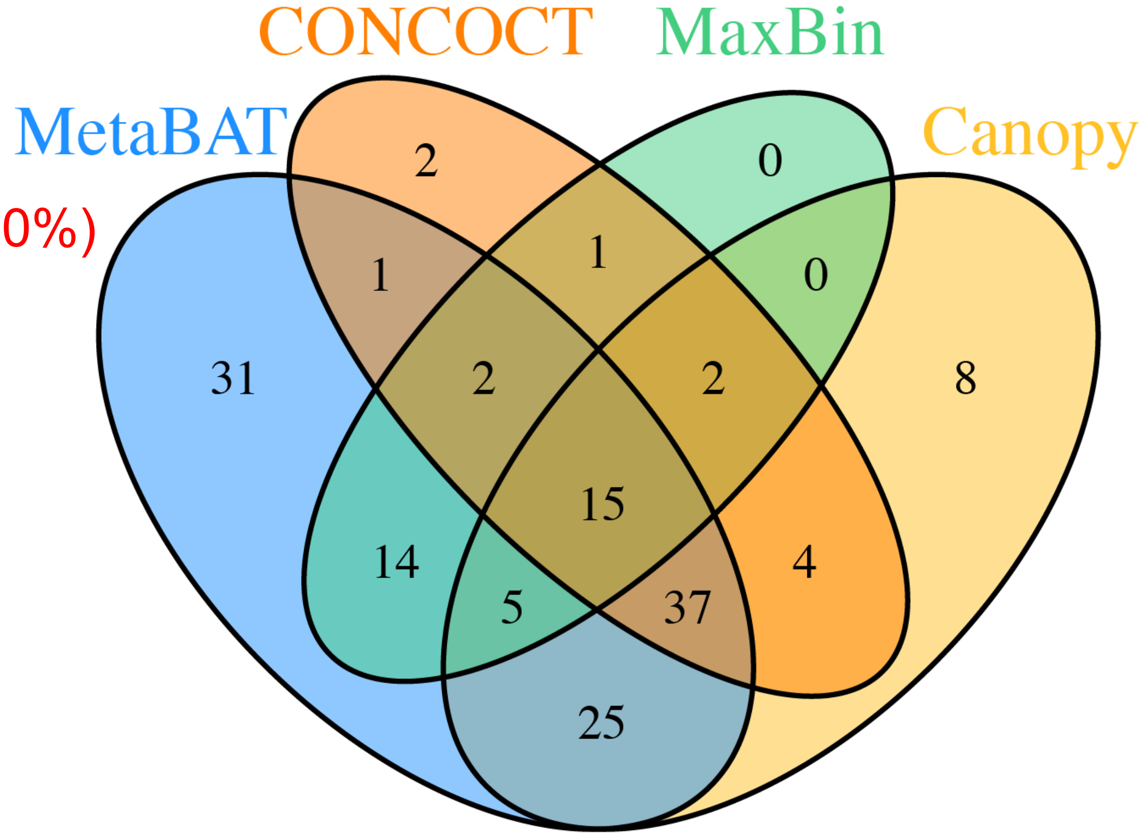
*Canopy only use abundance table as input, so it should have taken more time and memory to read and write sequence data like the others

**Manual steps were not used

For details, refer to *https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT*

Binners complement each other

130/144 (90%)

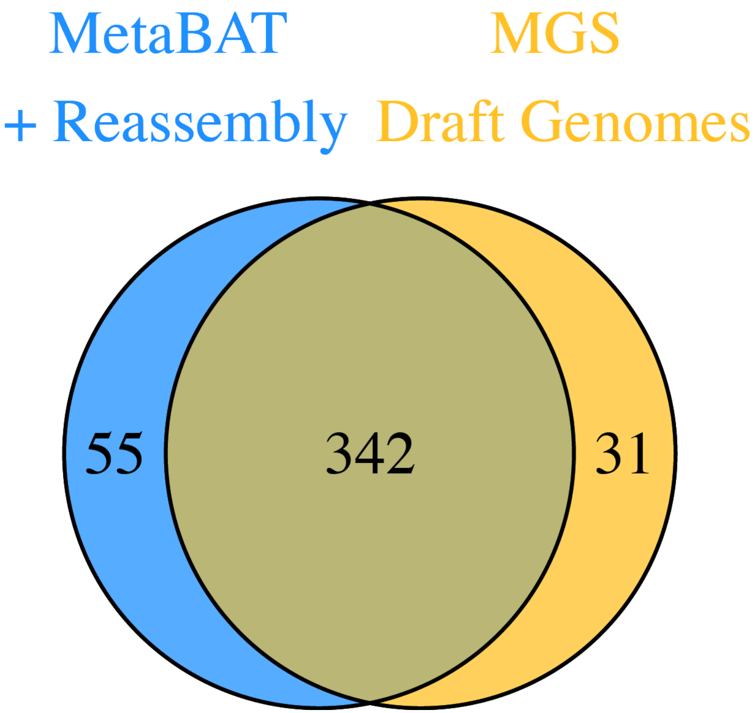


For details, refer to *https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT*

Can MetaBAT Scale to Huge Data Set?

- 1704 human gut metagenomic samples (ERP002061)
- >1M contigs over 1kb
- Only MetaBAT and Canopy was able to handle the amount of data
- 3 hours in a single node (with 32 threads using 17G memory)
- MetaBAT produced 790 (out of 1634) genome bins with >30% completeness and <5% contamination
- **Using genome bins as seeds, we recruited & reassembled reads to improve the quality of bins.**

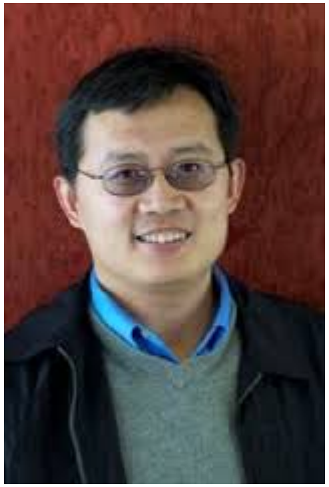
The Quality of Genome Bins Approximates High Quality Draft Genomes



342/373 (92%)

Acknowledgement

JGI JOINT GENOME INSTITUTE



Zhong Wang

Rob Egan

Jeff Froula

NERSC

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY