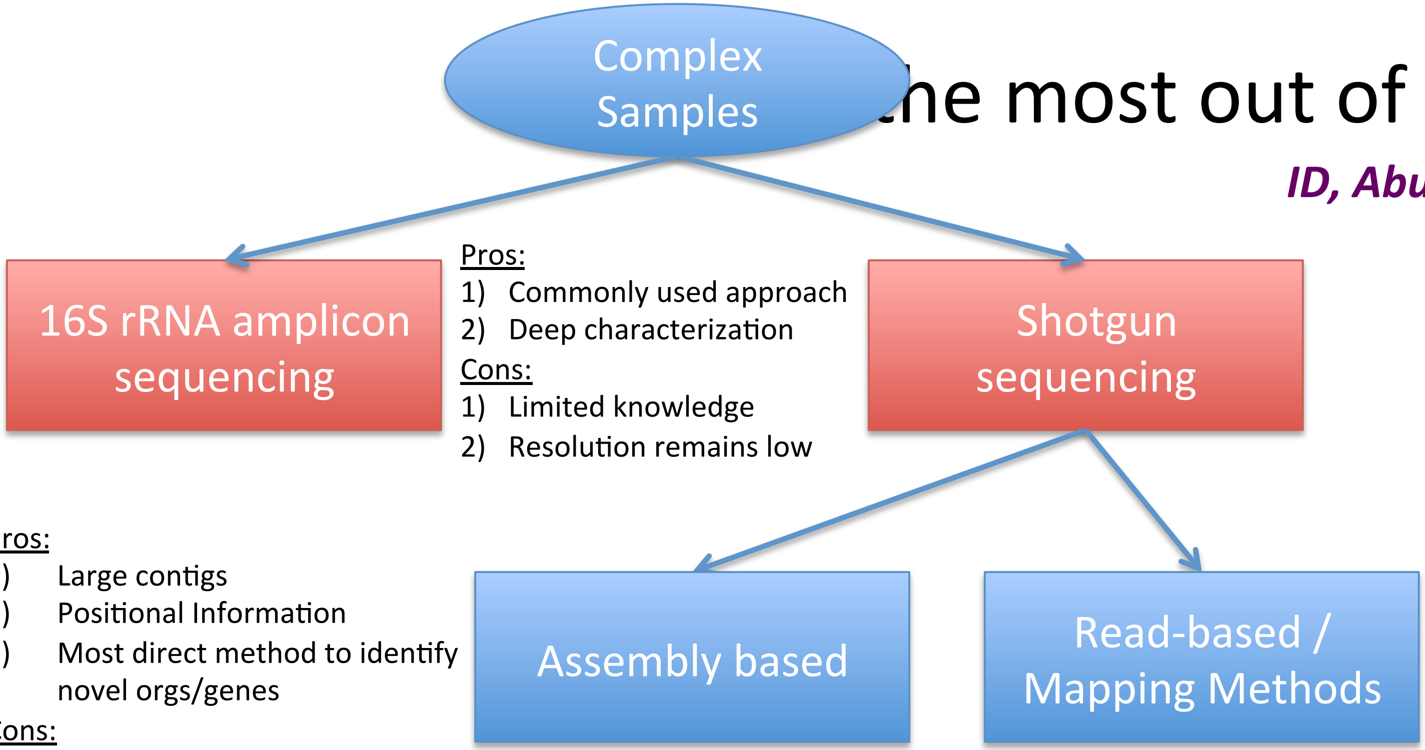
Metagenome Binning

What do we want out of metagenomes?

the most out of your data

ID, Abundance, Function



Pros:
1) Commonly used approach
2) Deep characterization

Cons:
1) Limited knowledge
2) Resolution remains low

Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Pros:
1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

Cons:
1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy - chimeras

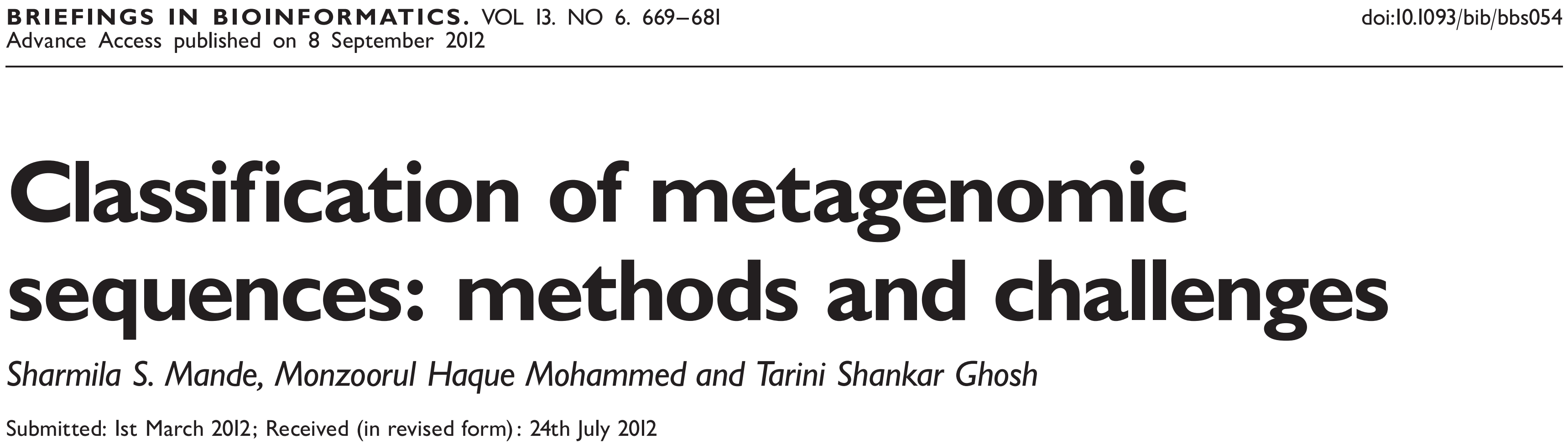
Cons:
1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

- Fragmented
- Difficult to produce full-length genomes

Patrick Chain

How can we separate these fragments?

Metagenome binning



BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 6. 669–681
Advance Access published on 8 September 2012

doi:10.1093/bib/bbs054

Classification of metagenomic sequences: methods and challenges

Sharmila S. Mande, Monzoorul Haque Mohammed and Tarini Shankar Ghosh

Submitted: 1st March 2012; Received (in revised form): 24th July 2012

- Binning, a process conceptually similar/analogous to established machine learning techniques, involves classifying and/or clustering reads into specific bins.

Binning is …

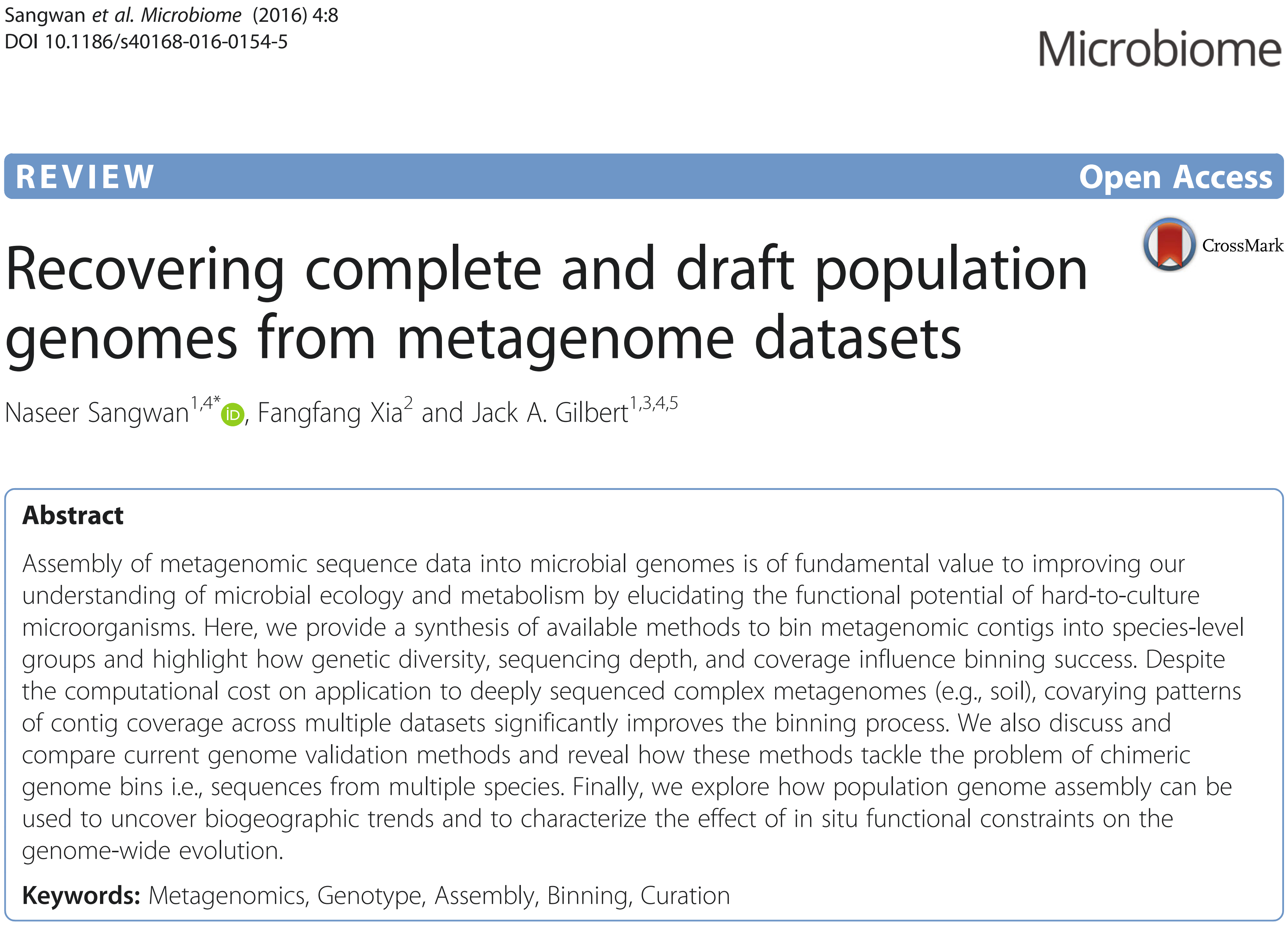
Clustering

&

Classification

General binning approaches

- Reference based approach
  - aka. supervised approach
- Unsupervised approach



Sangwan *et al. Microbiome* (2016) 4:8
DOI 10.1186/s40168-016-0154-5

Microbiome

REVIEW Open Access

Recovering complete and draft population genomes from metagenome datasets

Naseer Sangwan[1,4*], Fangfang Xia[2] and Jack A. Gilbert[1,3,4,5]

Abstract

Assembly of metagenomic sequence data into microbial genomes is of fundamental value to improving our understanding of microbial ecology and metabolism by elucidating the functional potential of hard-to-culture microorganisms. Here, we provide a synthesis of available methods to bin metagenomic contigs into species-level groups and highlight how genetic diversity, sequencing depth, and coverage influence binning success. Despite the computational cost on application to deeply sequenced complex metagenomes (e.g., soil), covarying patterns of contig coverage across multiple datasets significantly improves the binning process. We also discuss and compare current genome validation methods and reveal how these methods tackle the problem of chimeric genome bins i.e., sequences from multiple species. Finally, we explore how population genome assembly can be used to uncover biogeographic trends and to characterize the effect of in situ functional constraints on the genome-wide evolution.

Keywords: Metagenomics, Genotype, Assembly, Binning, Curation

Sangwan, N. et al, Microbiome 2016 4:8

MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities

Dongwan D. Kang[1,2], Jeff Froula[1,2], Rob Egan[1,2] and Zhong Wang[1,2,3]

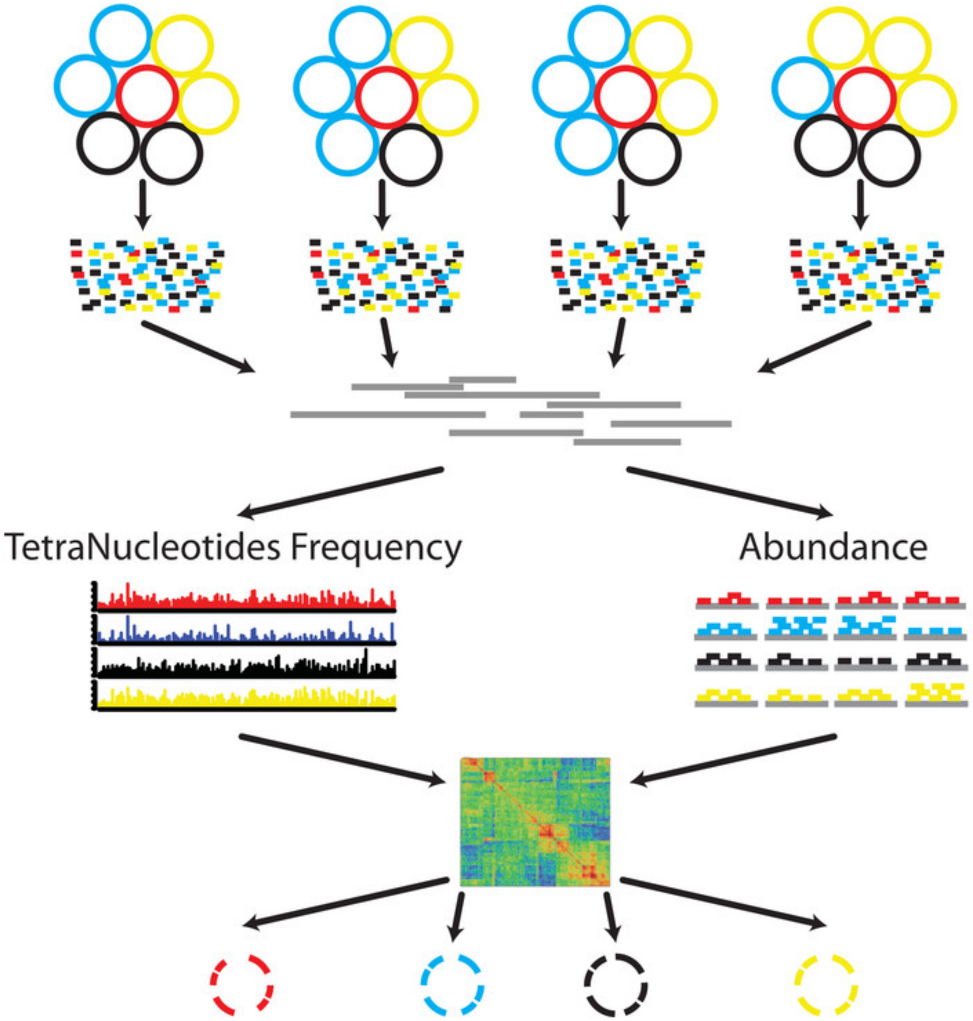
[1] Department of Energy Joint Genome Institute, Walnut Creek, CA, USA
[2] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3] School of Natural Sciences, University of California at Merced, Merced, CA, USA

ABSTRACT

Grouping large genomic fragments assembled from shotgun metagenomic sequences to deconvolute complex microbial communities, or metagenome binning, enables the study of individual organisms and their interactions. Because of the complex nature of these communities, existing metagenome binning methods often miss a large number of microbial species. In addition, most of the tools are not scalable to large datasets. Here we introduce automated software called MetaBAT that integrates empirical probabilistic distances of genome abundance and tetranucleotide frequency for accurate metagenome binning. MetaBAT outperforms alternative methods in accuracy and computational efficiency on both synthetic and real metagenome datasets. It automatically forms hundreds of high quality genome bins on a very large assembly consisting millions of contigs in a matter of hours on a single node. MetaBAT is open source software and available at https://bitbucket.org/berkeleylab/metabat.

Subjects Bioinformatics, Computational Biology, Genomics
Keywords Metagenome binning, MetaBAT

MetaBAT process



Preprocessing

1 Samples from multiple sites or times

2 Metagenome libraries

3 Initial de-novo assembly using the combined library

MetaBAT

4 Calculate TNF for each contig

5 Calculate Abundance per library for each contig

6 Calculate the pairwise distance matrix using pre-trained probabilistic models

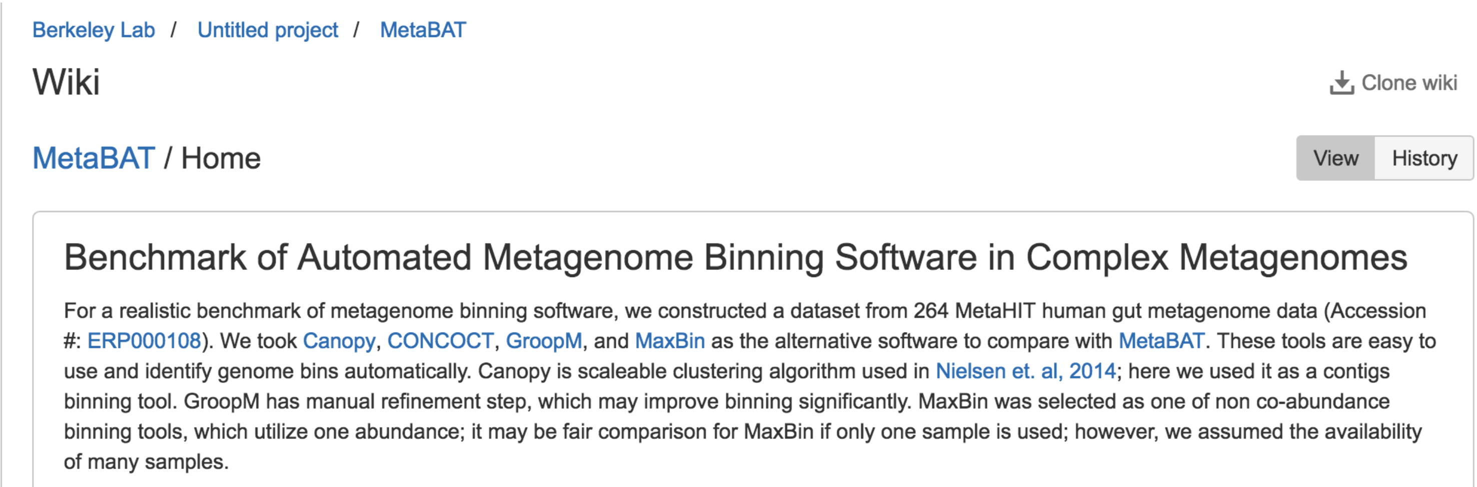
7 Forming genome bins iteratively

MetaBAT methodology

- Requires pre-trained models
  - Tetra nucleotide patterns
  - Genome abundance profile
  - Multiple samples
- Binning algorithm
  - Modified k-medoids
  - Heuristic
- Input
  - Assembled contigs (fasta file)
    - >2.5 kb worked the best
  - Mapped reads (sorted bam file)

Reproducible benchmarking

- **https://bitbucket.org/berkeleylab/metabat/wiki/Home**



Berkeley Lab / Untitled project / MetaBAT

Wiki

MetaBAT / Home

Benchmark of Automated Metagenome Binning Software in Complex Metagenomes

For a realistic benchmark of metagenome binning software, we constructed a dataset from 264 MetaHIT human gut metagenome data (Accession #: ERP000108). We took Canopy, CONCOCT, GroopM, and MaxBin as the alternative software to compare with MetaBAT. These tools are easy to use and identify genome bins automatically. Canopy is scaleable clustering algorithm used in Nielsen et. al, 2014; here we used it as a contigs binning tool. GroopM has manual refinement step, which may improve binning significantly. MaxBin was selected as one of non co-abundance binning tools, which utilize one abundance; it may be fair comparison for MaxBin if only one sample is used; however, we assumed the availability of many samples.

Other tools use oligonucleotide frequency and coverage for binning

- CONCOCT
- GroupM
- MaxBin
- Databionic ESOM tools