Separating Metagenomic Short Reads into Genomes via Clustering

Tao Jiang

(joint work with Olga Tanaseichuk and James Borneman)

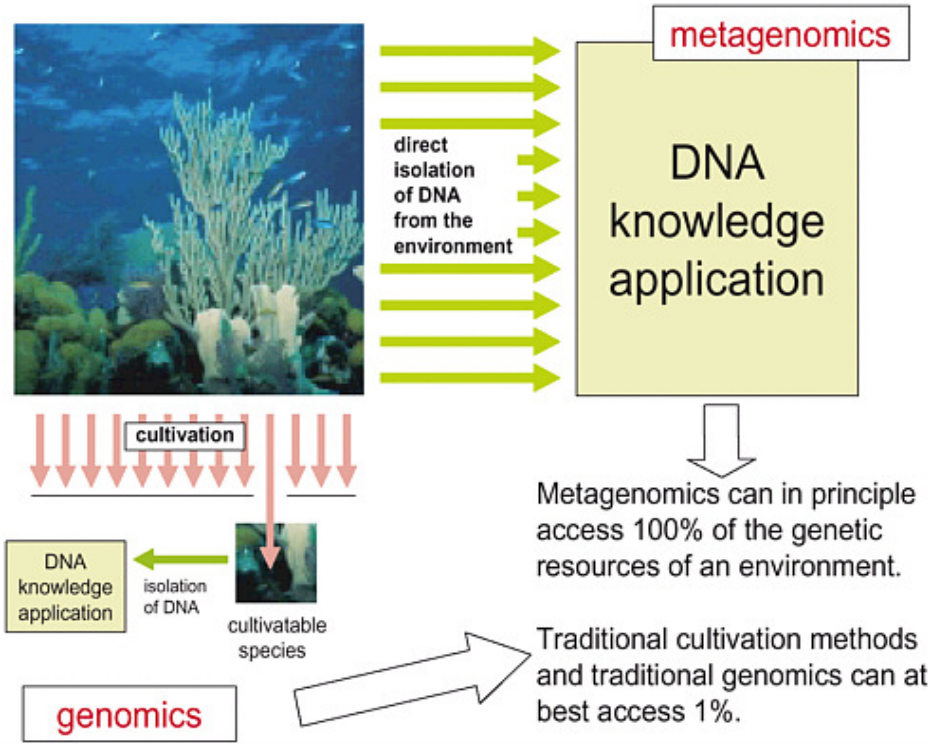
2013

Outline

- Metagenomics and DNA Sequencing
- Problem Formulation
- ~~Related Work~~
- Our Method
  - Overview
  - Observations and Intuition
  - ~~Details of the Algorithm~~
- Experimental Results
- Implementation and Conclusions

Metagenomics

- Genomics
  - Study of an organism's genome
  - Relies upon cultivation and isolation
  - > 99% of bacteria cannot be cultivated



- Metagenomics
  - Study of all organisms in an environmental sample by simultaneous sequencing of their genomes
  - Makes it possible to study organisms that can't be isolated or difficult to grow in a lab

Metagenomic Projects

The Acid Mine Drainage Project

The Tinto River in Spain (Credit - Carol Stoker)

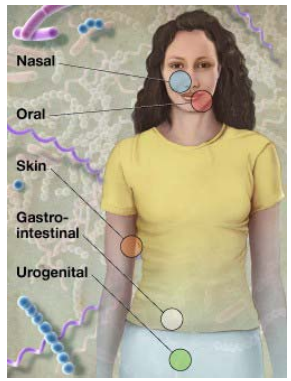
- Motivation: to understand mechanisms by which the microbes tolerate the extremely acid environments
- Simple community: 5 dominant species (3 bacteria and 2 archaea)

The Sargasso Sea Project

A coral reef off the coast of Malden Island in Kiritibati

- A large scale sequencing in an environmental setting
- Identified >1 million of putative genes (10 times > than in all databases at that time)
- ~1800 species

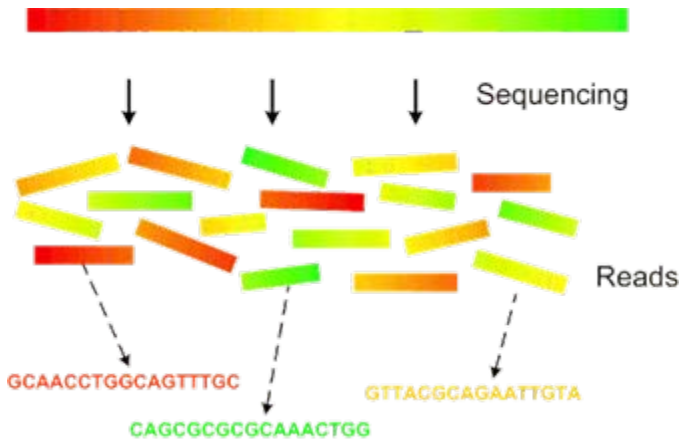
The Human-Microbiome Project



- Microbial community living in a host
- 100 trillion microbes
- 100 times more microbial than human genes
- Is there a core human microbiome?
- How changes in microbiome correlate with human health?

DNA Sequencing

- Sanger sequencing
- Next generation sequencing (NGS)
  - High-throughput
  - Cost- and time-effective
  - No cloning (reduced clonal biases)
  - Shorter read length compared to Sanger reads (~1000 bps)
    - Roche/454 (~450 bps)
    - Illumina/Solexa (35-100 bps)
    - ABI SOLiD (35–50 bps)
  - Due to rapid progress, sequencing lengths will increase



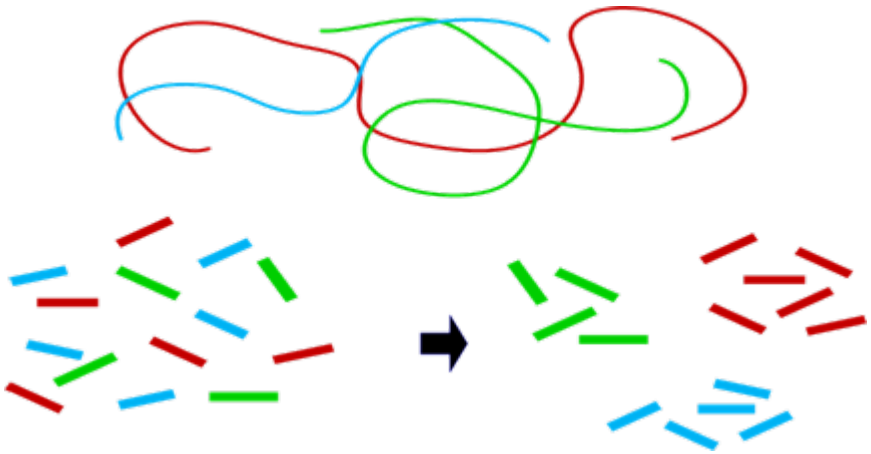
Goals of Metagenomics

- Phylogenetic diversity
- Metabolic pathways
- Genes that predominate in a given environment
- Genes for desirable enzymes
- **Comparative metagenomics ???**
- ...

A fundamental step: complete genomic sequences

Problem Formulation

- Given metagenomic reads, separate reads from different species (or groups of related species)



Difficulties

- Repeats in genomic sequences
- Sequencing errors

genomics

- Unknown number of species and abundance levels
- Common repeats in different genomes due to homologous sequences

metagenomics

Approaches

- Similarity-Based
  - Similarity search against databases of known genomes or genes/proteins
- Composition-Based
  - Binning based on conserved compositional features of genomes
- Abundance-Based
  - Separate genomes by abundance levels

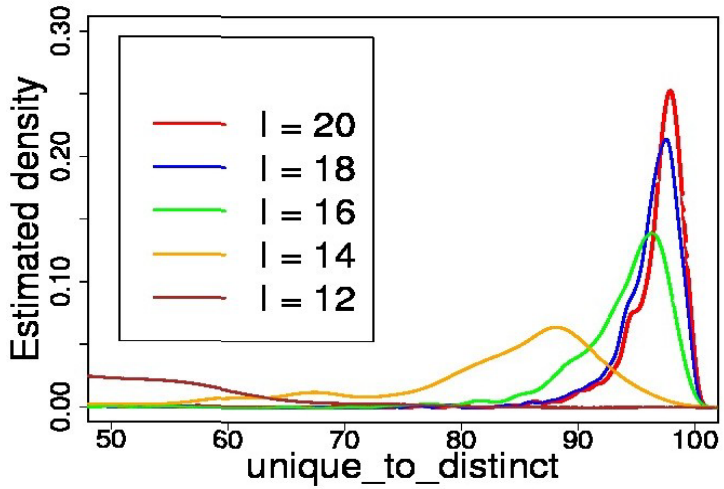
Our Algorithm: Overview

- Purpose: separating short paired-end reads from different genomes in a metagenomic dataset
- Two-phase heuristic algorithm
  - based on *l*-mers
  - similar abundance levels
  - arbitrary abundance levels (in combination with AbundanceBin [Wu and Ye, RECOMB, 2010])

Algorithm: Definitions and Observations

Unique *l*-mers (occur only once)

Repeated *l*-mers (occur > once)

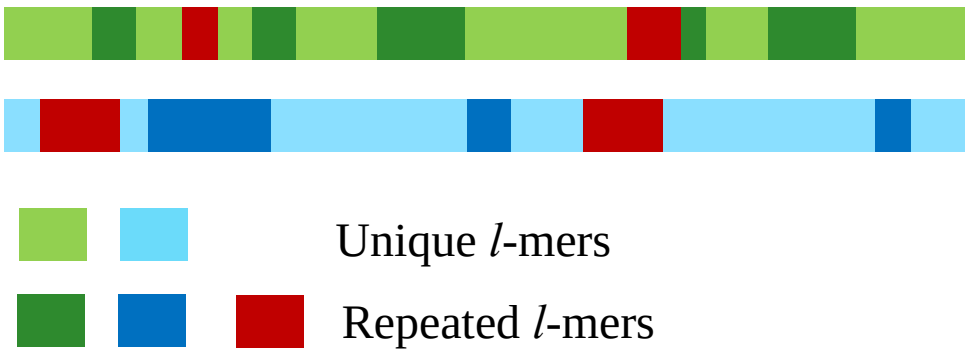


The ratio of unique *l*-mers to distinct *l*-mers

Observation 1: Most of the *l*-mers in a bacterial genome are unique

l ~ 20, for most of complete genomes

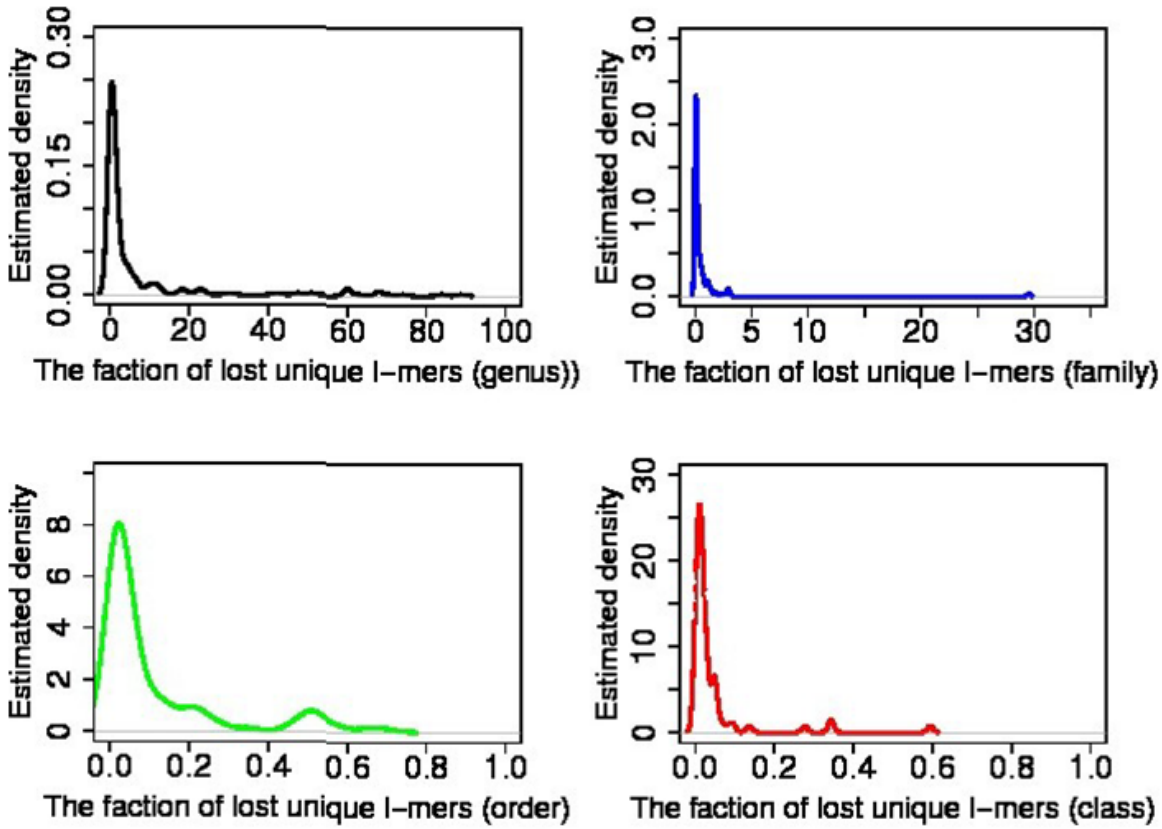
Algorithm: Definitions and Observations



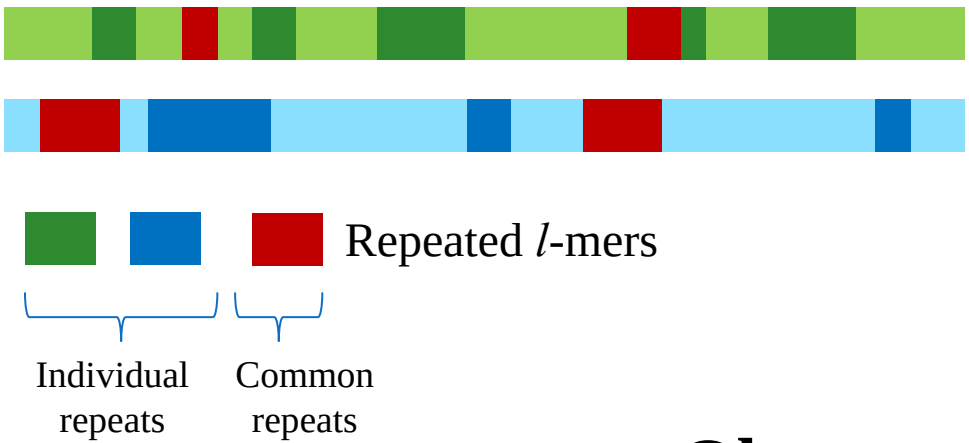
Observation 2: Most *l*-mers in a metagenome are unique

for $l \sim 20$ and genomes separated by sufficient phylogenetic distances

Algorithm: Definitions and Observations



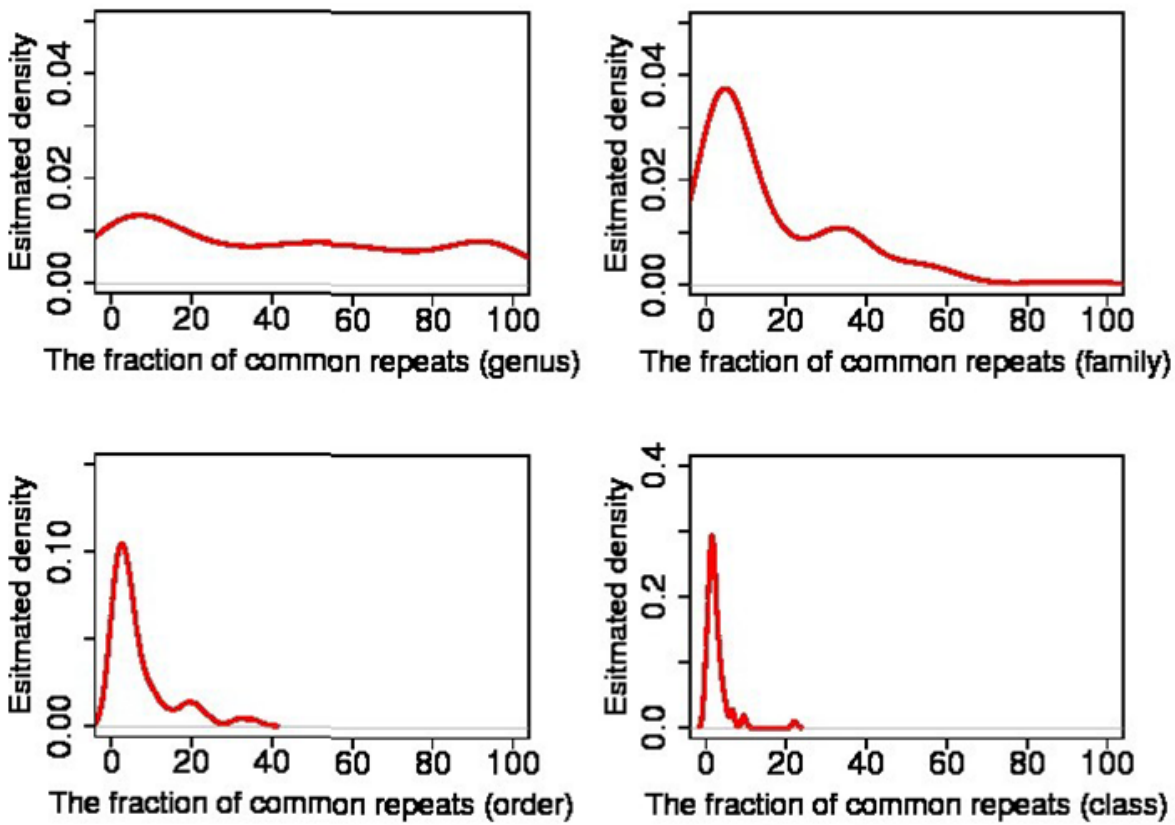
Algorithm: Definitions and Observations



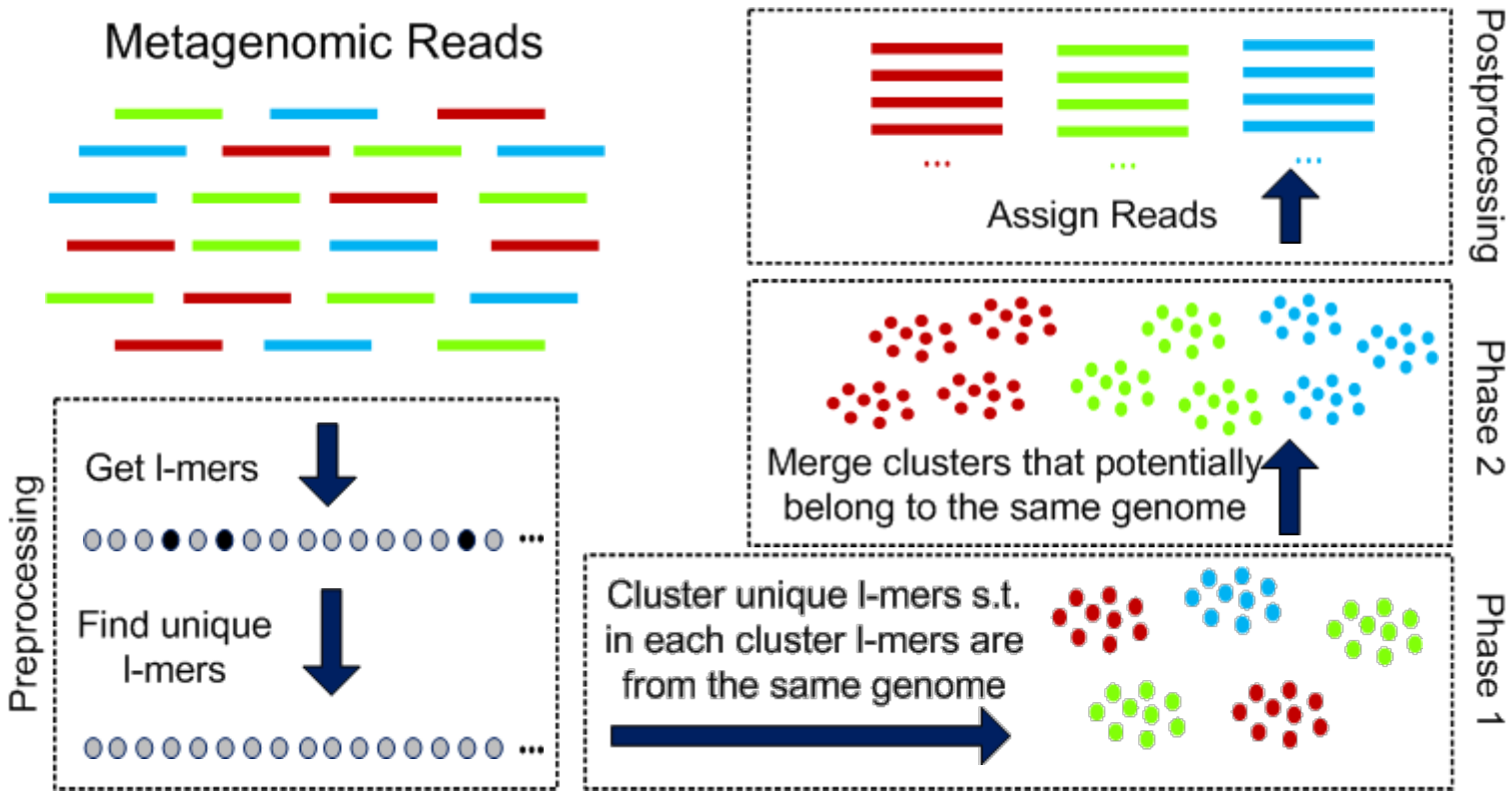
Observation 3: Most of the repeats in a metagenome are individual

for $l \sim 20$ and genomes separated by sufficient phylogenetic distances

Algorithm: Definitions and Observations

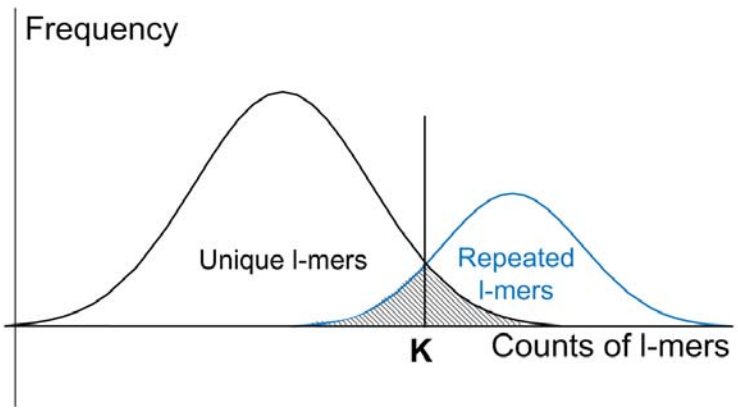


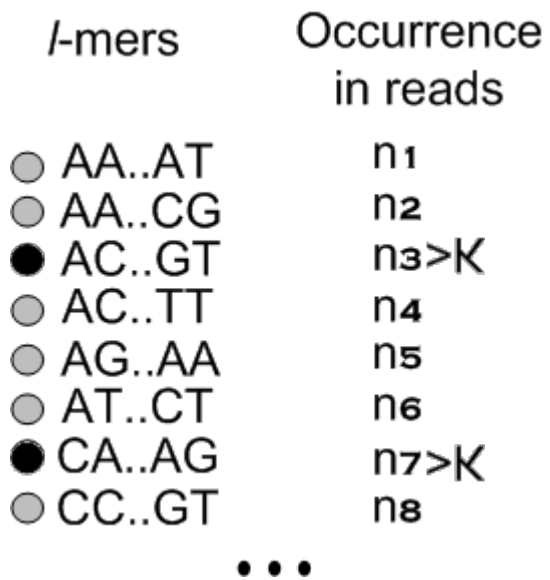
Flowchart


Metagenomic Reads
Preprocessing
Get l-mers
Find unique l-mers
Cluster unique l-mers s.t. in each cluster l-mers are from the same genome
Phase 1
Merge clusters that potentially belong to the same genome
Phase 2
Assign Reads
Postprocessing

Algorithm: Preprocessing

- Finding unique *l*-mers
  - Count occurrence of *l*-mers in reads
  - Find threshold *K* for counts of *l*-mers to separate unique *l*-mers and repeats
    - Unique *l*-mers: counts < *K*.
    - Repeats: counts > *K*.

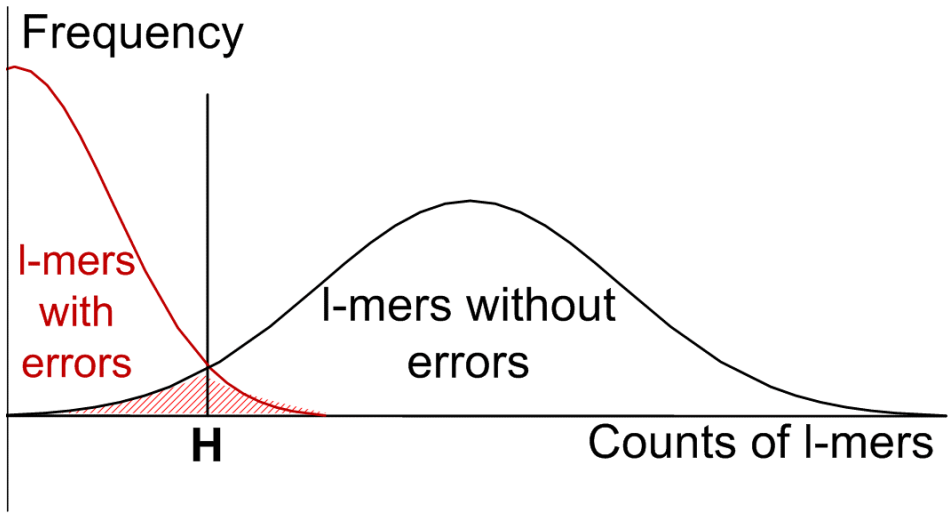




Choice of *K*: Observed frequency of the count = 2 * (expected frequency of the count in unique *l*-mers)

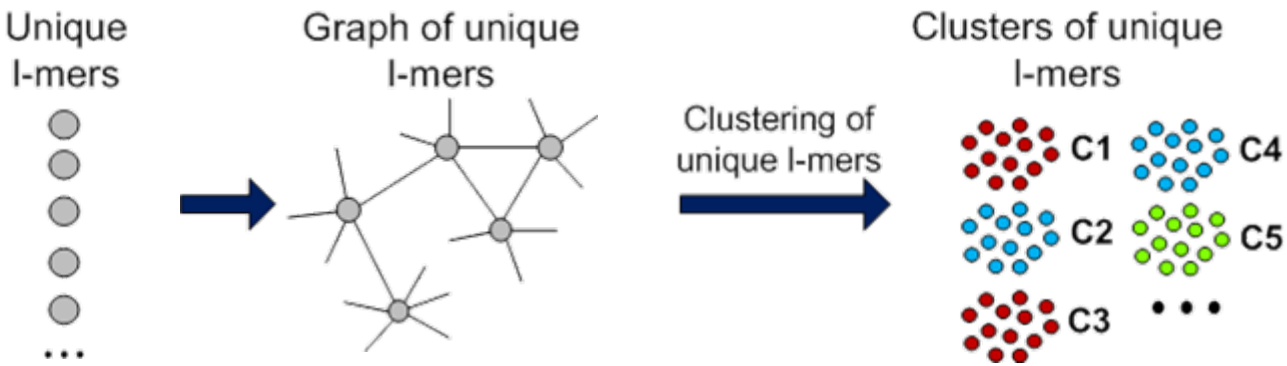
Algorithm: Preprocessing

- Finding *l*-mers with errors
  - Threshold *H* for counts of *l*-mers to separate *l*-mers with and without errors



Algorithm: Phase I

- Goal:
  - *l*-mers in each cluster are from the same genome
  - Each genome may correspond to several clusters



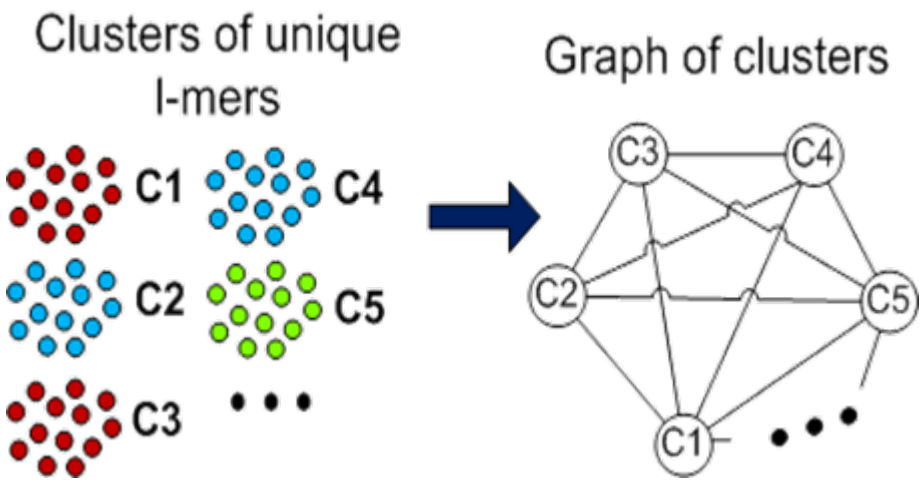
- Graph of unique *l*-mers:
  - Nodes – unique *l*-mers
  - Edge (*u*,*v*) iff *u* and *v* occur in the same read

Algorithm: Phase I

- Cluster initialization
  - l-mers of an unclustered read
- Cluster expansion
  - Add nodes with at least T neighbors
  - Stop if more than $2(L-(l+T)+1)$ l-mers are to be added
    - It means that repeated l-mers (wrongly classified as unique) were added at a previous step. L is read length.
  - Choose T s.t. the expected number of gaps in coverage by $(l+T)$-mers < 1

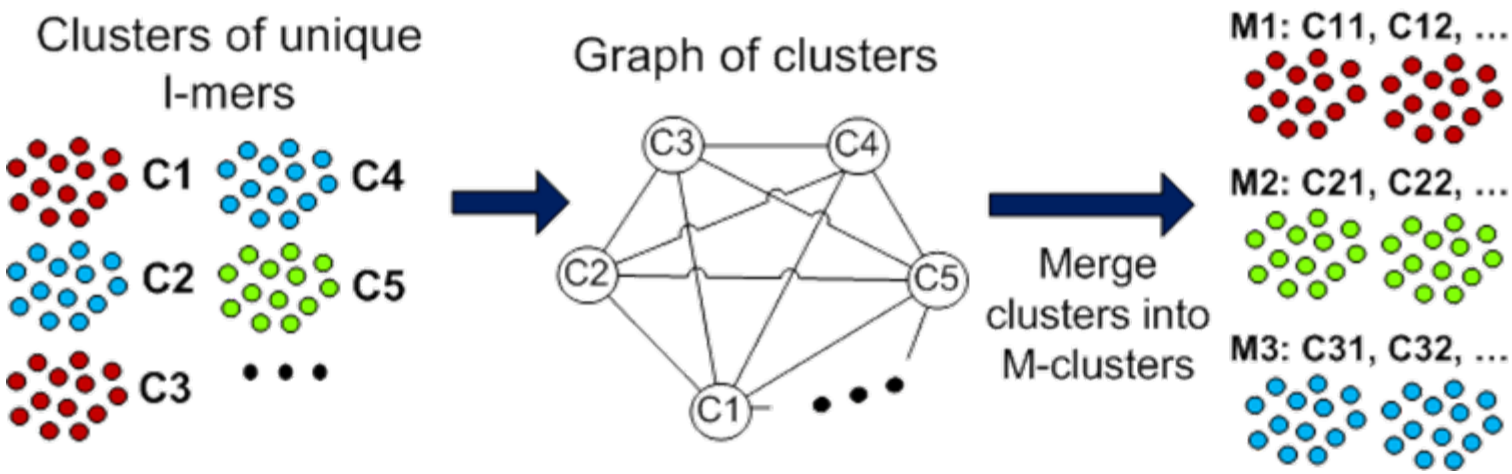
Algorithm: Phase II

- Goal: merge clusters from the same genome
- Weighted graph
  - For every cluster C_i construct set R_i that contains:
    - Repeats in reads assigned to C_i
    - Repeats in mate-pairs of reads assigned to C_i
  - Nodes – clusters R_i
  - Weights: $w(i,j) = R_i \cap R_j$



Algorithm: Phase II

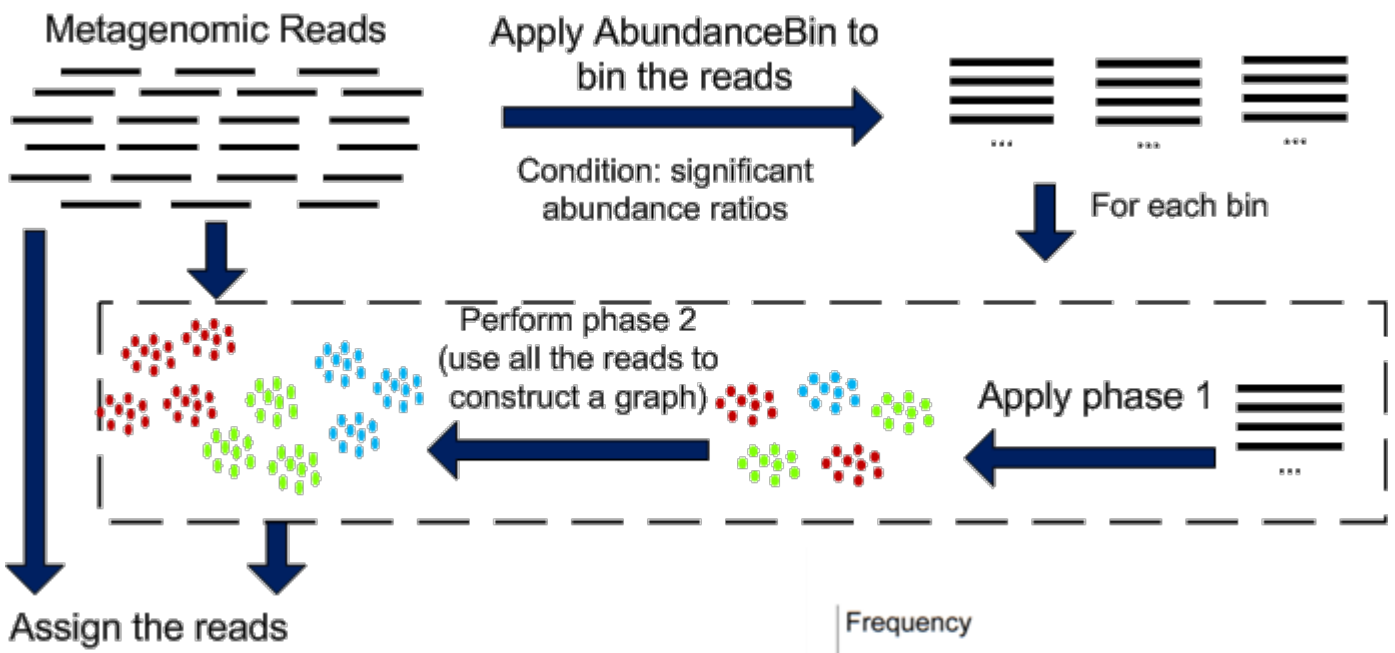
- MCL algorithm [van Dongen, PhD Thesis, 2000]
  - For clustering sparse weighted graphs
  - Parameter P ~ granularity
  - We use an iterative algorithm to find the best P



Algorithm: Postprocessing

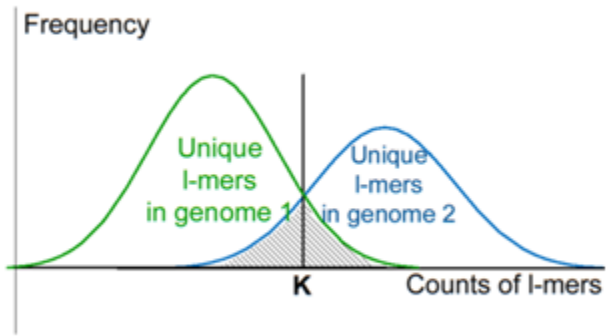
- Assign a read to a cluster if >50% of its *l*-mers correspond to the same cluster
- Unassigned reads: iteratively assigned using mates

Arbitrary Abundance Levels



- Significant abundance ratios is defined by the expected misclassification rate (>3%)

$$l_2 \sum_{i=1}^{K-1} \frac{\lambda_2^i e^{-\lambda_2}}{i!} + l_1 \sum_{i=K}^{Max} \frac{\lambda_1^i e^{-\lambda_1}}{i!}$$



Experimental Results: Overview

- Lack of NGS metagenomic benchmarks
- Most binning algorithms in the literature are concerned with Sanger reads
- Datasets
  - Tests on variety of synthetic datasets with different number of genomes, phylogenetic distances and abundance ratios
  - Performance on a real metagenomic dataset from gut bacteriocytes of a glassy-winged sharpshooter
- Comparison
  - We modify the Velvet assembler [Zerbiono and Birney, Renome Research, 2008] to work as a genome separator (clusters in Phase I are replaced by sets of *l*-mers from the Velvet contigs)
  - With CompostBin [Chatterji et al., RECOMB, 2008] on Sanger reads
  - With MetaCluster on short NGS reads [Wang et al., Bioinformatics, 2012]

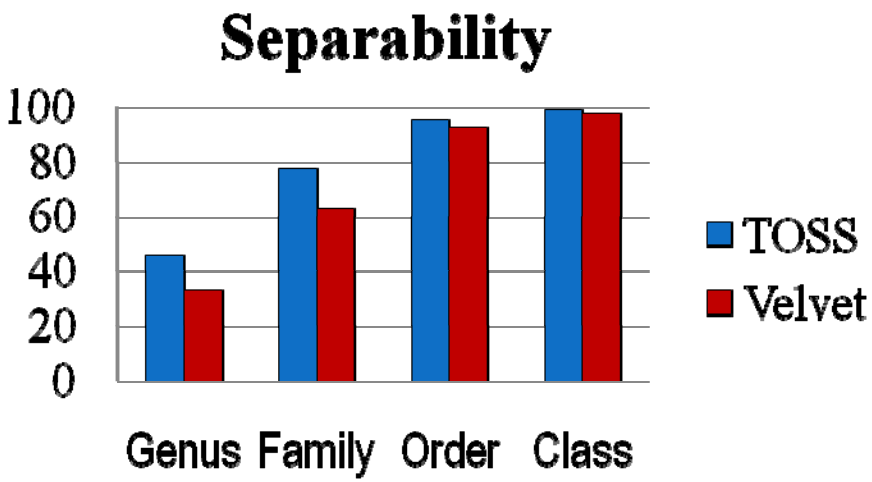
Experimental Results: Evaluation

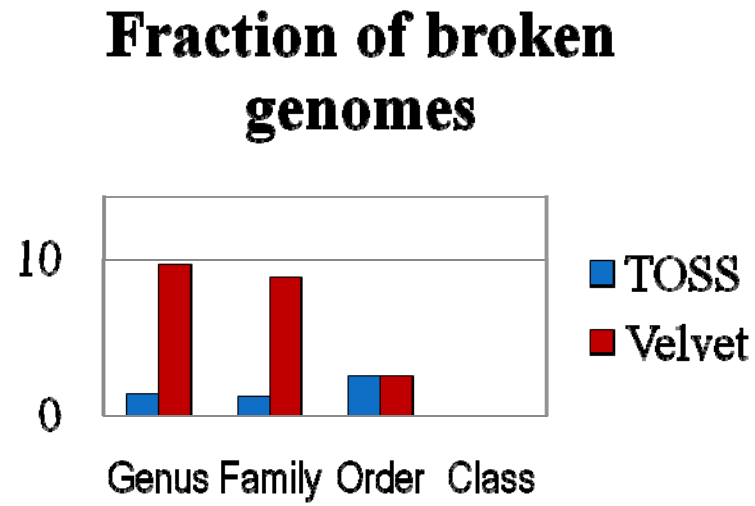
- Genomes are assigned by majority of reads (at least 50%)
- Several genomes may correspond to one cluster
- Evaluation factors
  - Broken genomes (not assigned)
  - Separability (percent of separated pairs)
- Sensitivity
  - (# true positives)/(# all reads from the genomes assigned to the cluster)
- Precision
  - (# true positives)/(# reads in a cluster)

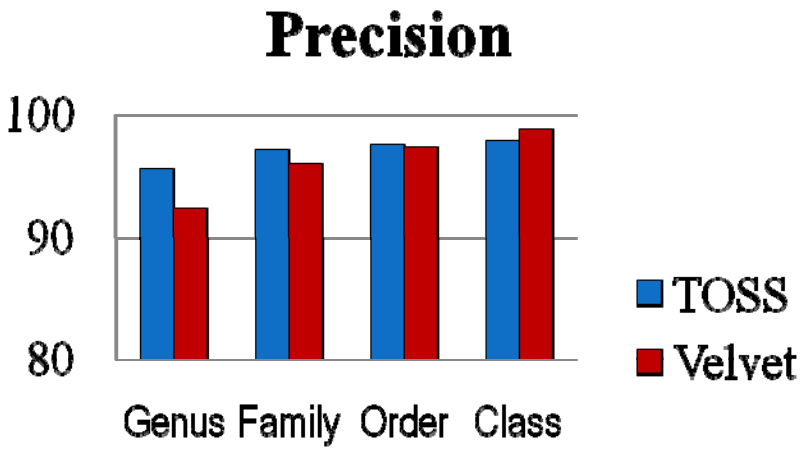
Experimental Results

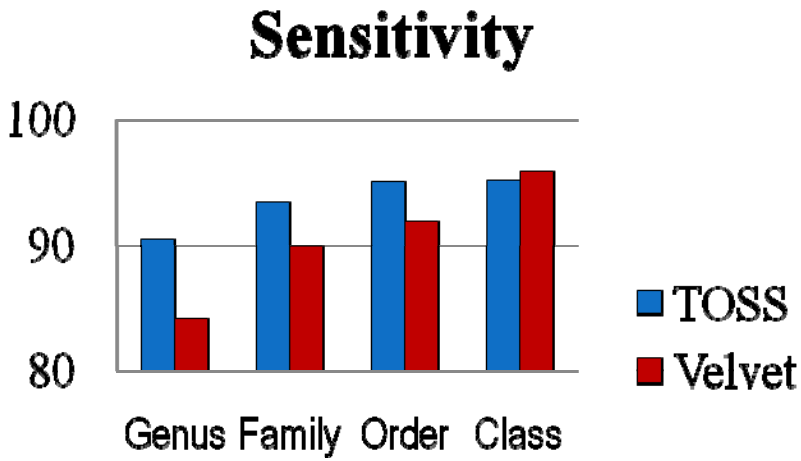
- 182 synthetic datasets of 4 categories
  - 79 experiments for the same genus
  - 66 – same family
  - 29 – same order
  - 8 – same class
- Read length: 80 bps
- Coverage depth: ~15-30
- Equal abundance levels
- 2-10 genomes in each dataset
- Simulation: Metasim [Richter et al., PloS ONE, 2008]
- Phylogeny: NCBI taxonomy

Experimental Results

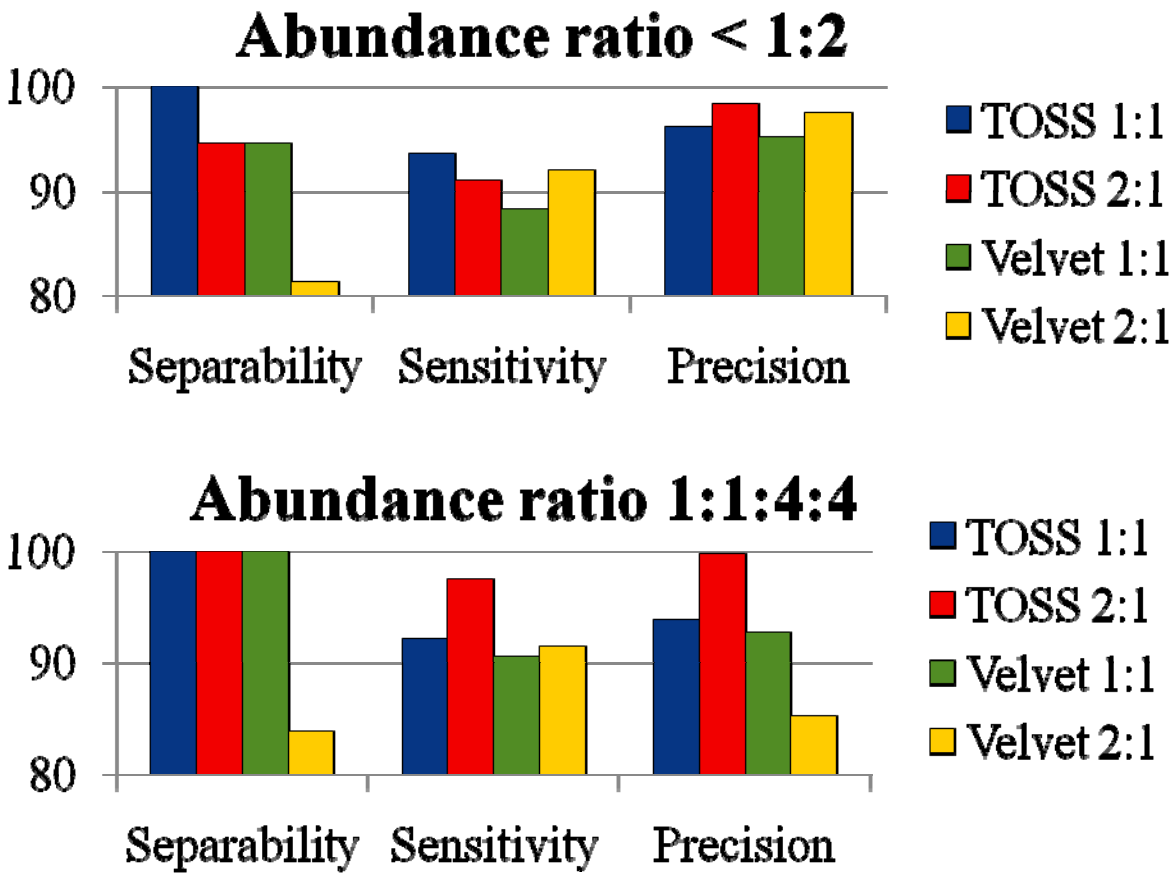








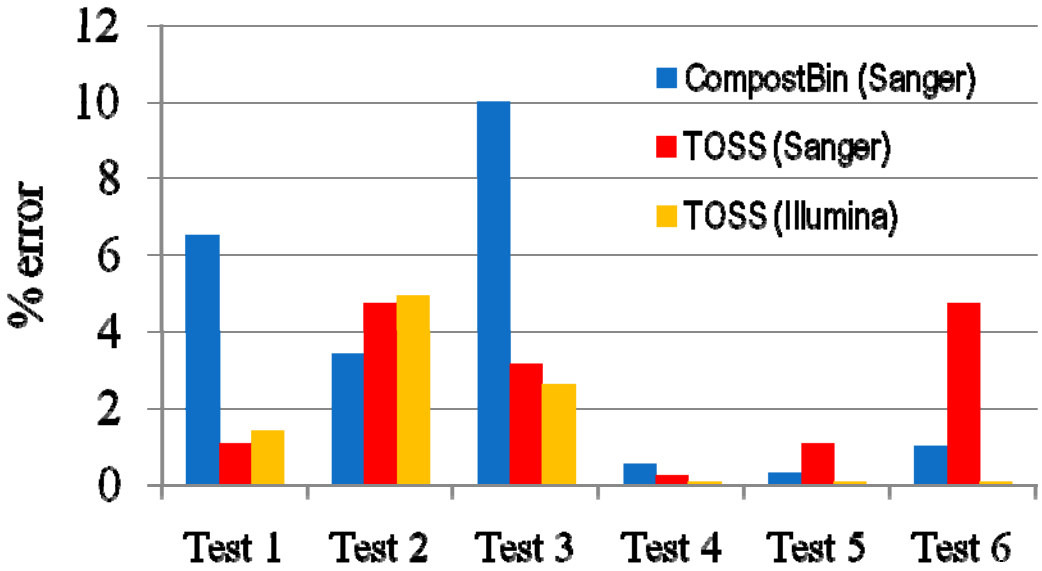
Experimental Results: Genomes with Different Abundance Levels

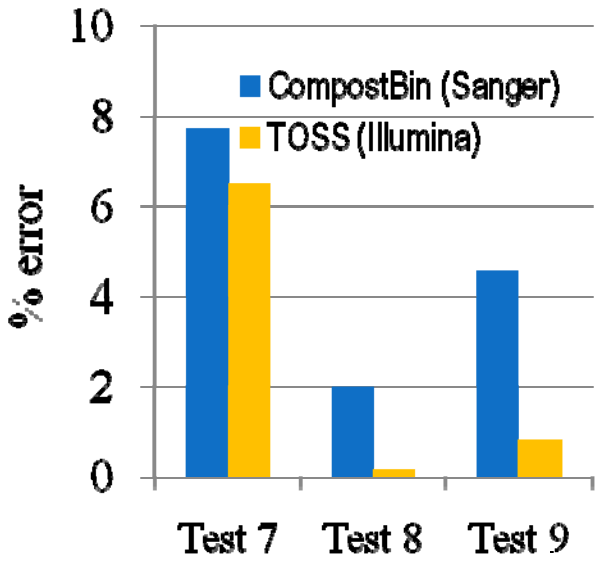


Experimental Results: Comparison with CompostBin

- Simulated paired-end Sanger reads from [Chatterji et al., RECOMB, 2008]
  - Handling longer reads (1000 bps)
    - Cut long reads into short reads of 80 bps
    - Linkage information is recovered in Phase II
  - Handling lower coverage depth (~3-6)
    - Choose higher threshold K to separate repeats and unique l-mers in preprocessing
- Simulated paired-end Illumina reads
  - 80 bps, high coverage depth (~15-30)

Experimental Results: Comparison with CompostBin





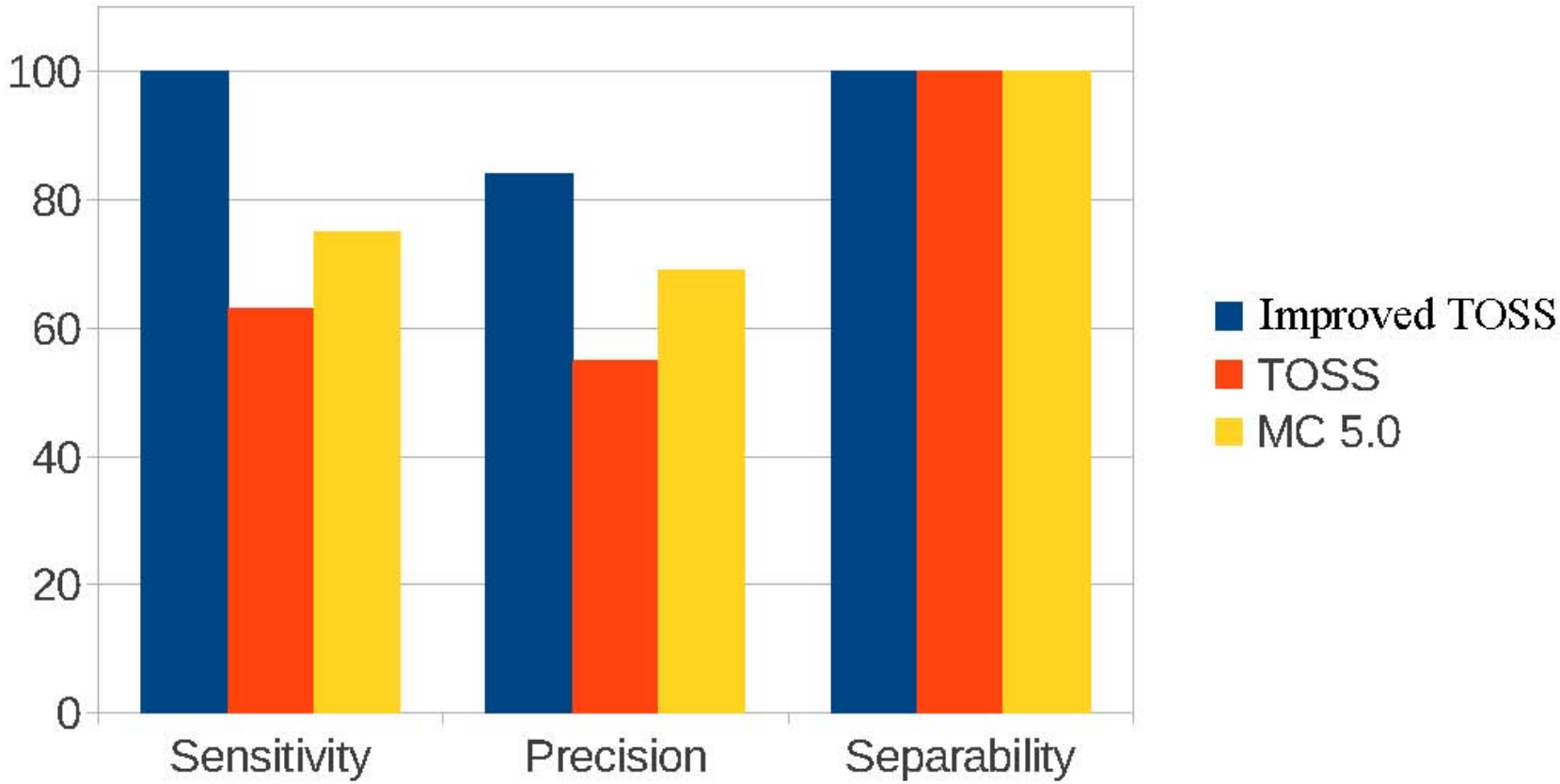
	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9
Abundance ratio	1:1	1:1	1:1	1:1	1:1	1:1	1:1:8	1:1:8	1:1:1:1:2:14
Phylogenetic distance	Species	Genus	Genus	Family	Family	Order	Family Order	Order Phylum	Species, Order, Family Phylum, Kingdom

Experimental Results: Real Dataset

- Gut bacteriocytes of glassy-winged sharpshooter, *Homalodisca coagulata*
  - Consists of reads from:
    - *Baumannia cicadellinicola*
    - *Sulcia muelleri*
    - Miscellaneous unclassified reads
- Sanger reads
- Performance is measured on the ability to separate reads from *B.cicadellinicola* and *S.muelleri*
- Performance
  - TOSS: Sensitivity: ~92%, error rate ~1.6%
  - CompostBin: error rate: ~9%

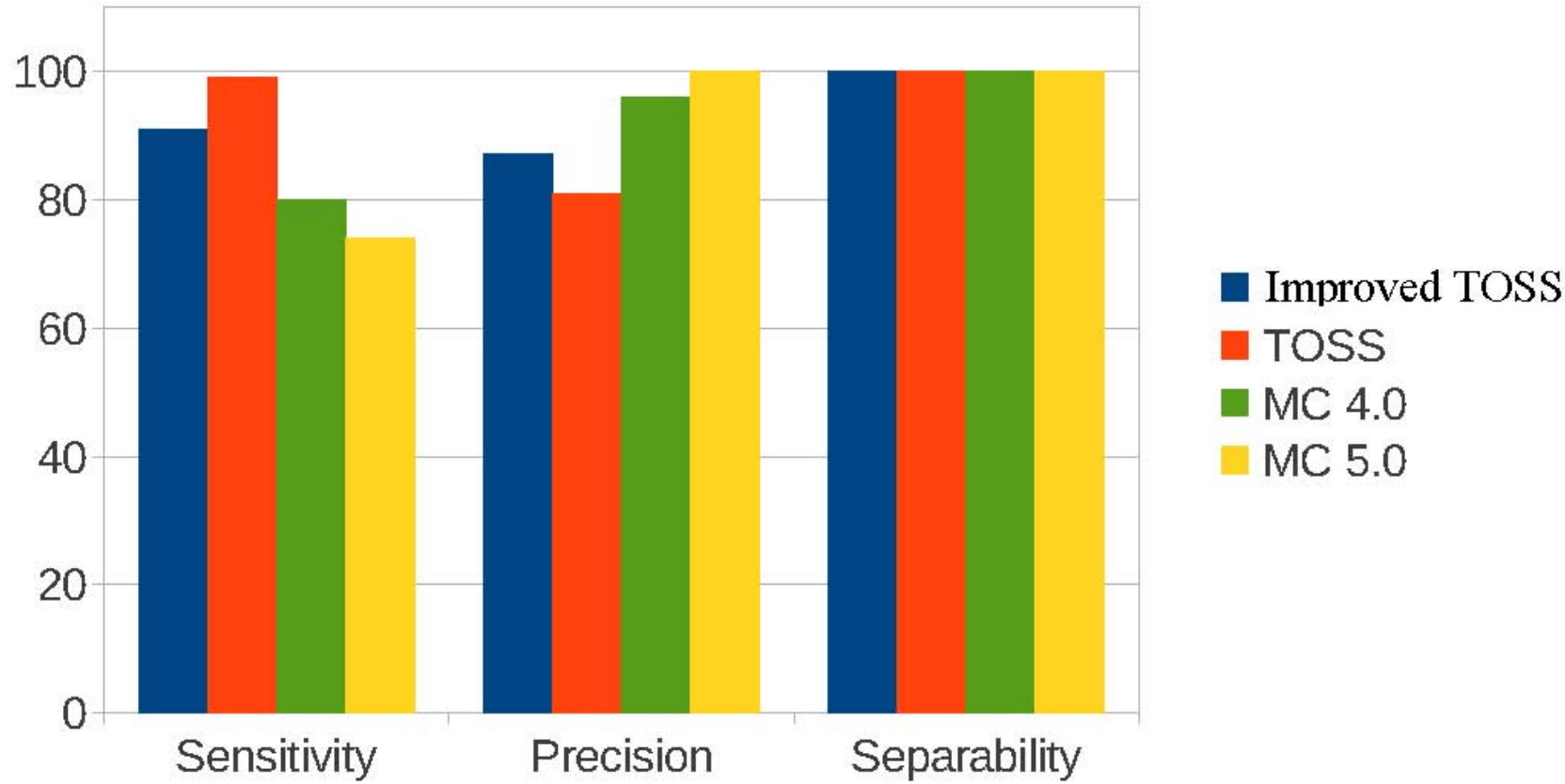
Performance of the Improved TOSS

by modeling *l*-mer frequency distribution more carefully,
taking into account sequencing errors



4 genomes, coverages 4 and 10

Performance of the Improved TOSS



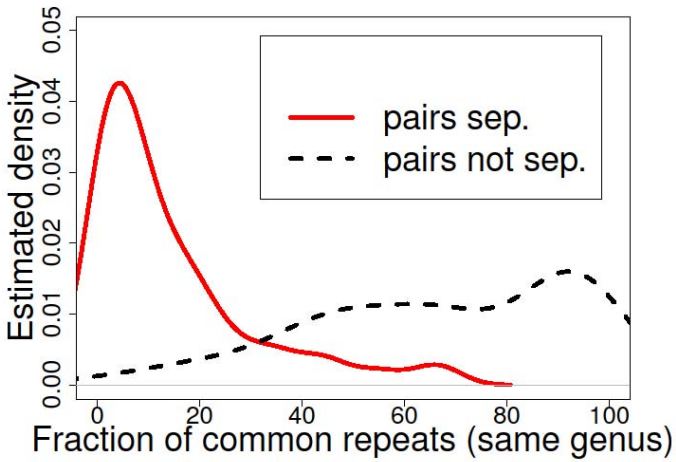
10 genomes, coverages 1.5 and 9

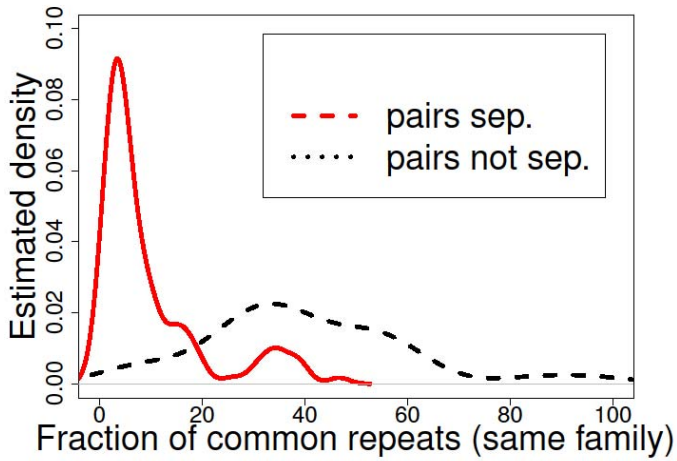
Implementation of TOSS

- Implemented in C
- Running time and memory depend on
  - Number and length of reads
  - Total length of the genomes
- For 80 bps reads -- 0.5 GB of RAM per 1 Mbps
  - 2-4 genomes, total length 2-6 Mbps – 1-3 h, 2-4 GB of RAM
  - 15 genomes, total length 40 Mbps – 14 h, 20 GB of RAM

Conclusion

- Genomes can be separated if the number of common repeats is small compared to the number of all repeats.





Fraction of common repeats to all repeats in evaluated datasets tests

- Additional information (such as compositional properties) could be added to improve separability in Phase II.