Binning Metagenomic Reads

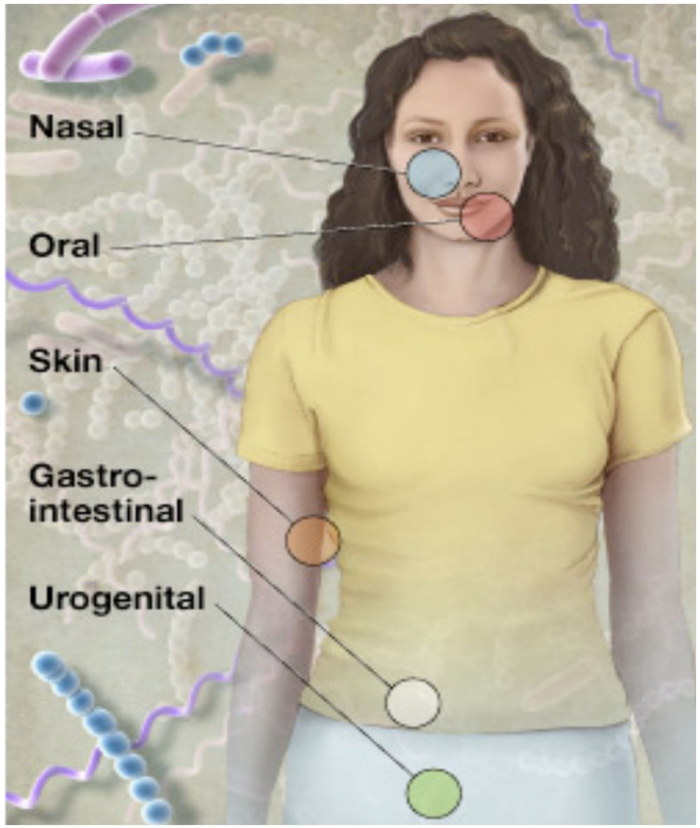
Microbes are important

http://npic.orst.edu/envir/soil.html

http://www.rimpro-india.com/articles1/waste-water-treatment-technologies-and-techniques.html

http://ocean.nationalgeographic.com/ocean/

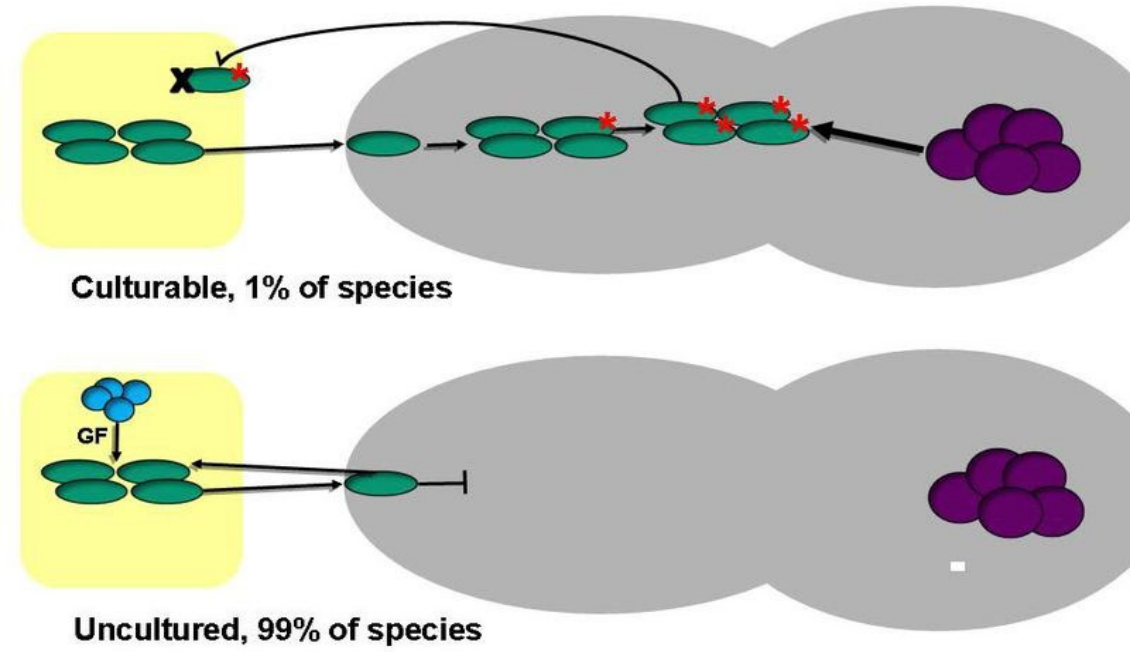
https://en.wikipedia.org/wiki/Desert



https://www.bcm.edu/departments/molecular-virology-and-microbiology/research/the-human-microbiome-project

Difficulty in study of microbes

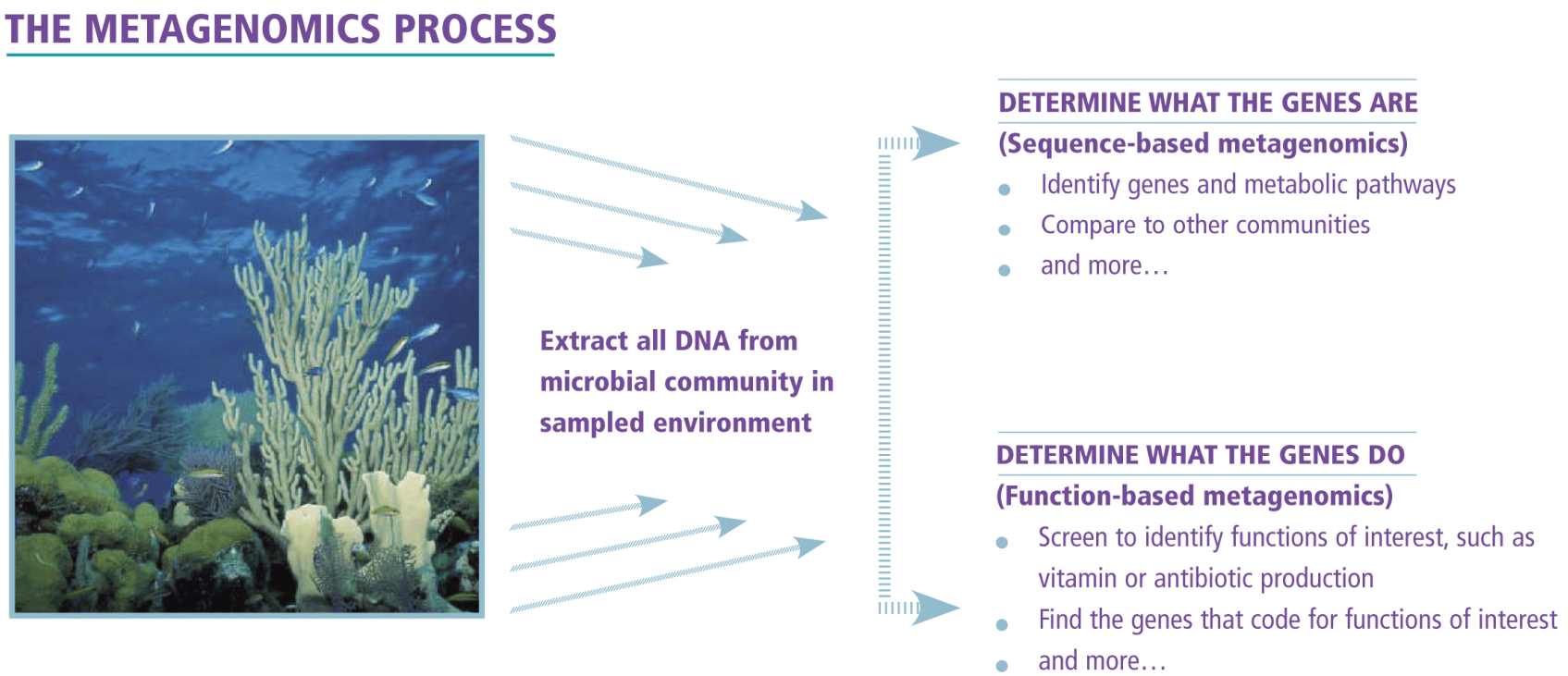
more than 99% of organism genomes in the environment are uncultured



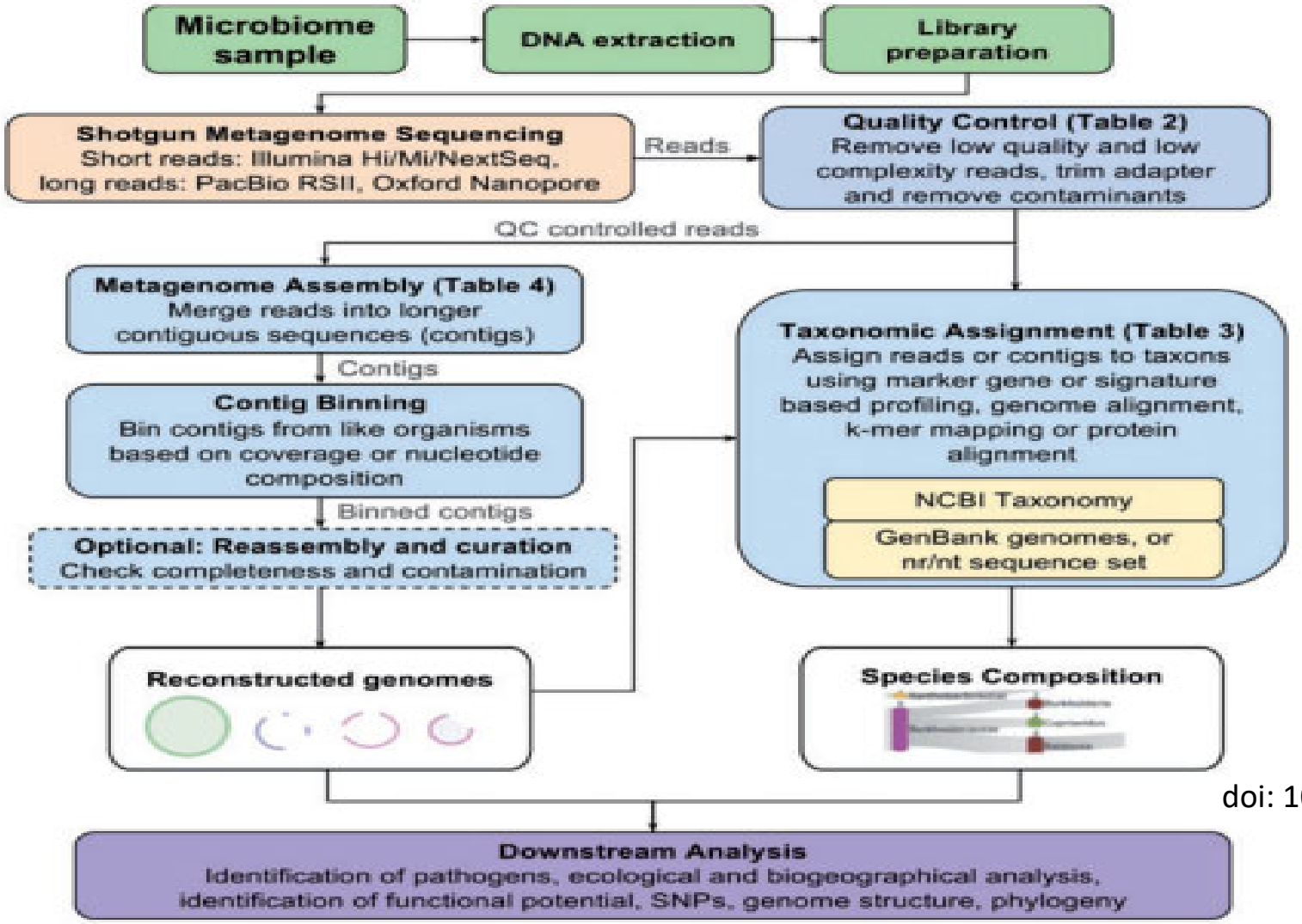
http://schaechter.asmblog.org/schaechter/2010/07/the-uncultured-bacteria.html

What is Metagenomics?

Metagenomics is the study of genetic material recovered directly from environmental samples.

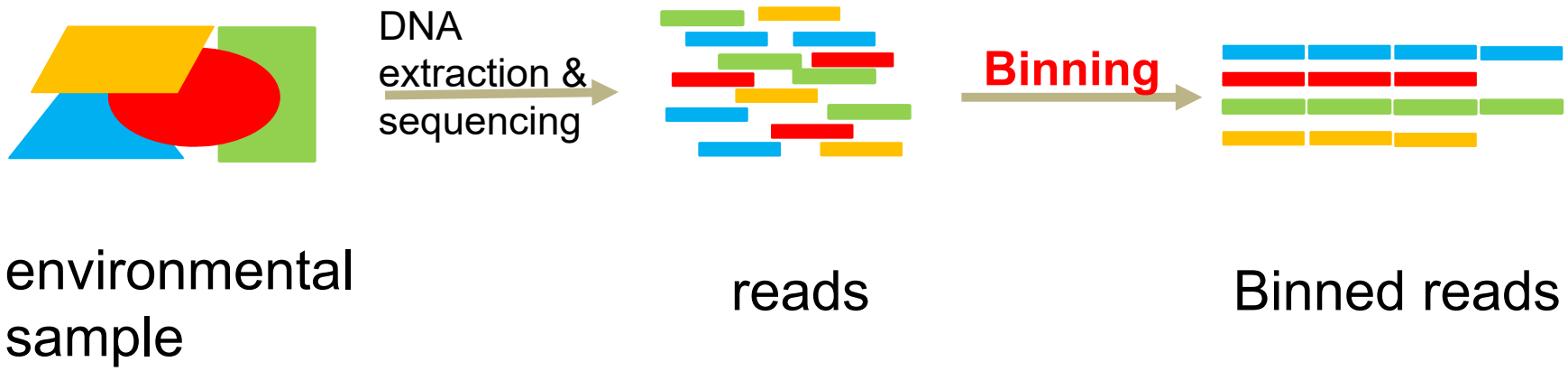


http://blog.openhelix.eu/?p=431

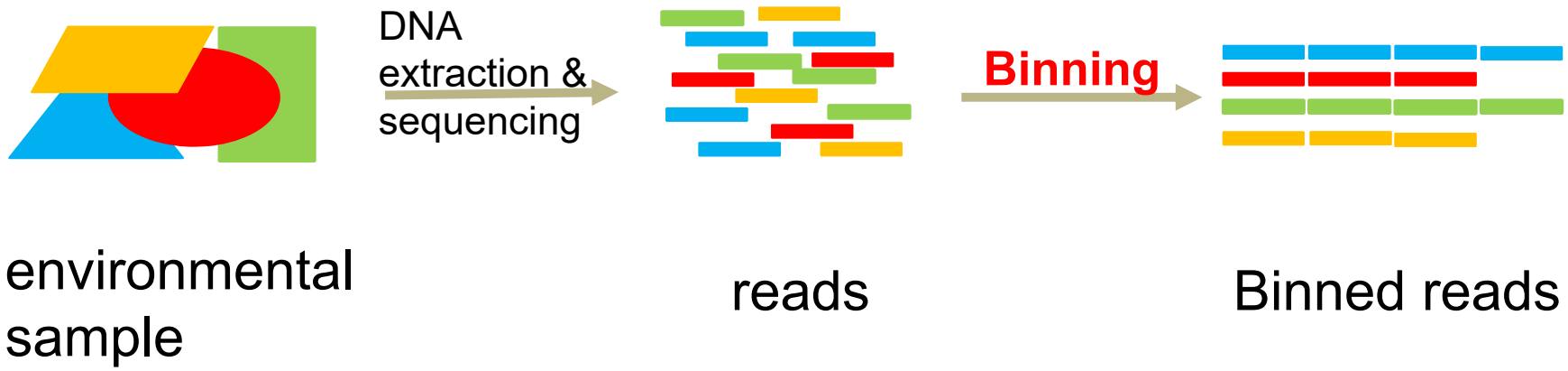


doi: 10.1093/bib/bbx120

Metagenomic analysis--Binning

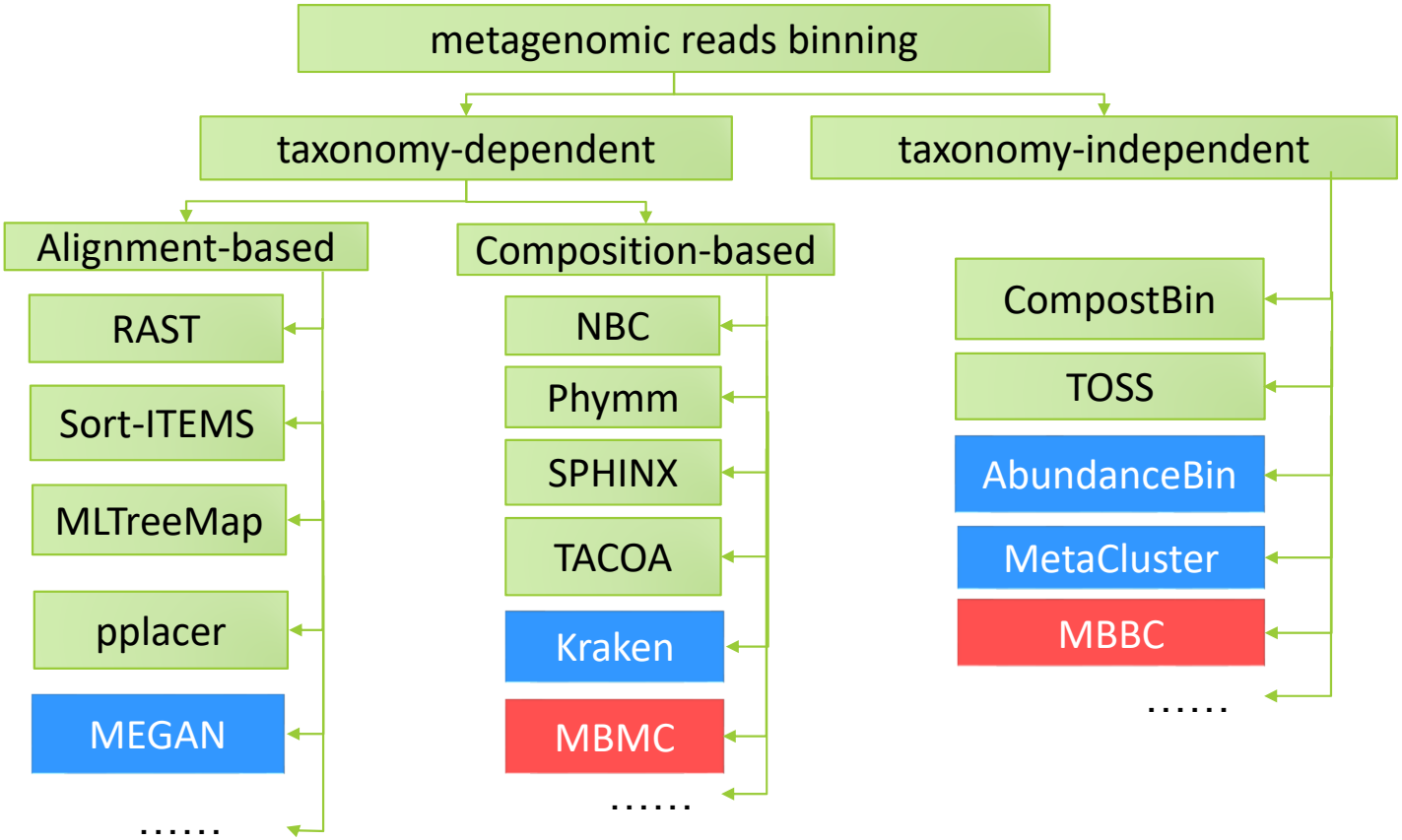


Metagenomic analysis--Binning



- Short reads length
- Sequencing errors
- A huge number of unknown genomes

classical approaches



Taxonomy-dependent methods:

Alignment-based binning (**MEGAN**)

http://ab.inf.uni-tuebingen.de/software/welcome.html/megan5

Composition-based binning (**Kraken**)

Kraken
Taxonomic Sequence Classification System

https://ccb.jhu.edu/software/kraken/

Problems in taxonomy-dependent methods

MBMC works better for datasets that contain unknown species

Taxonomy-independent methods

AbundanceBin, MetaCluster

the difference of k-mer (short sequence with length k) frequencies of different microbes in the environmental samples

Observations:

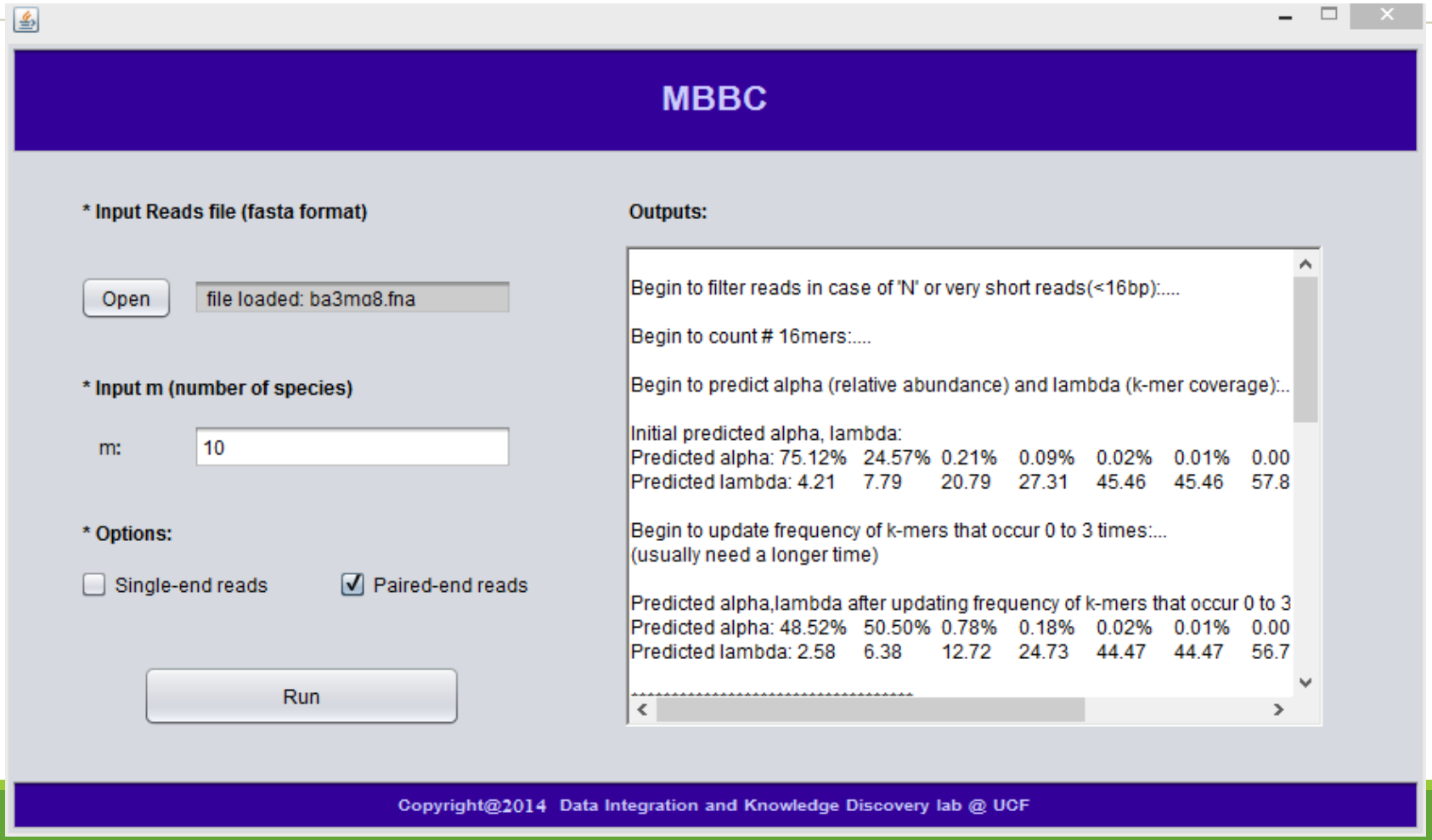
- k-mer frequency ⟺ genome's coverage
- long k-mers are usually unique in each genome
- k-mer frequency distribution from the same genome are similar

Problems in taxonomy-independent methods

- AbundanceBin/MetaCluster: only utilize the k-mers frequency.

→MBBC: improve the method to utilize the k-mer frequency;
utilized Markov properties shared by a group of reads

MBBC-(Metagenomic Binning Based on Clustering)

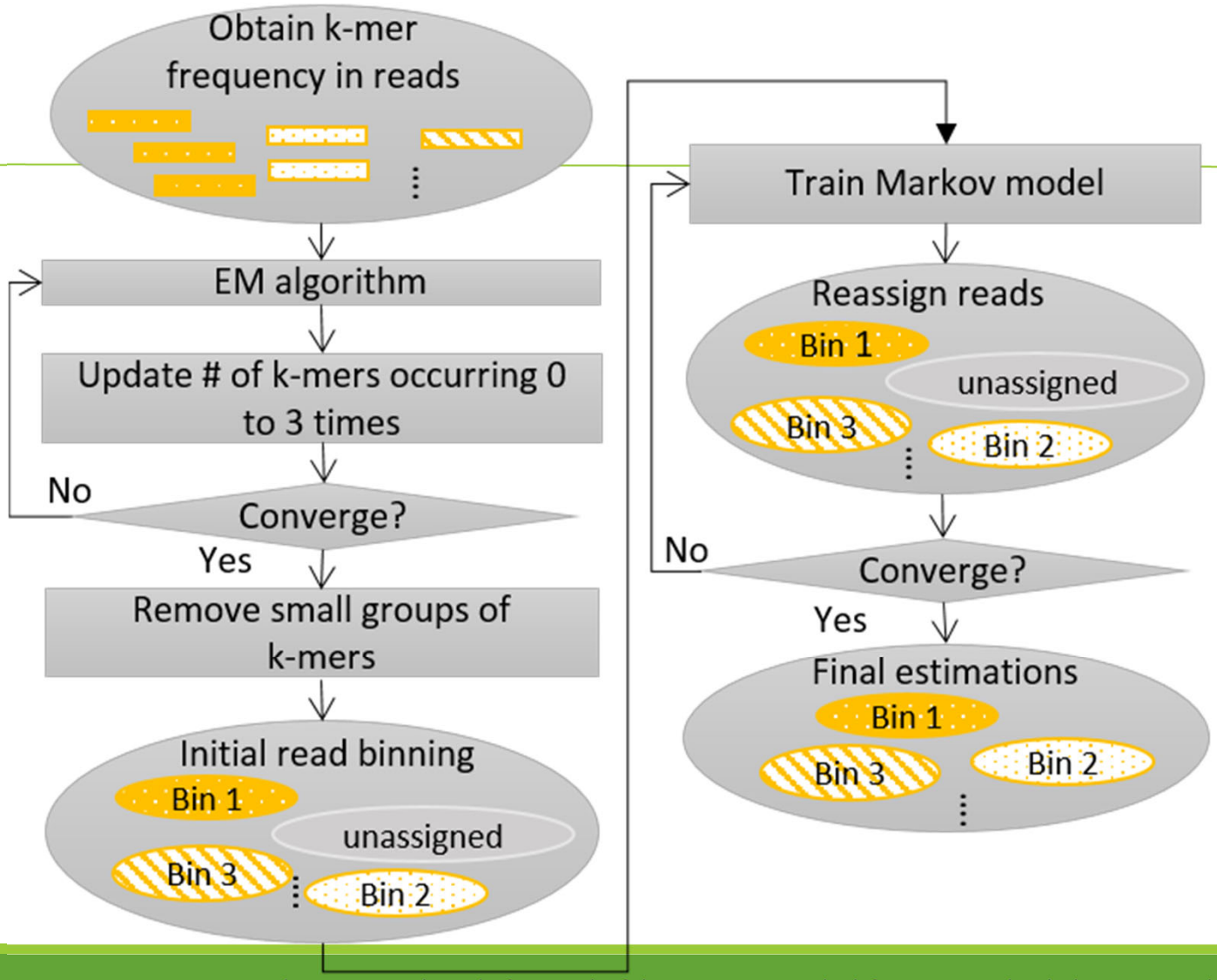


MBBC

Taxonomy-independent method, bin reads based on

- k-mer frequency
- Markov properties

Wang Y, Hu H, Li X. MBBC: an efficient approach for metagenomic binning based on clustering. BMC Bioinformatics. 2015 Feb 5; 16(1):36.



The procedure of read clustering in MBBC. The output on the right from each of the main steps on the left is connected with the corresponding steps.

Wang *et al. BMC Bioinformatics* 2015 **16**:36 doi:10.1186/s12859-015-0473-8

Mixture Poisson distribution

- Frequency of k-mers in all reads: x_i follows a mixture of m Poisson distribution

- if x_i is from j-th Poisson distribution, then $P(x_i = \mathrm{x}) = \alpha_j p_j(\lambda_{i,} x) = \alpha_j \frac{\lambda_j^x}{x!} e^{-\lambda_j}$, j:1,2,…,m

α_j : relative abundance

λ_j : k-mer coverage

EM algorithm

$$Z_{ij} = \frac{\alpha_{j*}p_j(\lambda_j, x_i)}{\sum_{r=1}^{m} \alpha_{r*}p_r(\lambda_r, x_i)}$$

M-step

$$\alpha_j = \frac{\sum_{i=1}^{n} Z_{ij}}{n}$$

$$\lambda_j = \frac{\sum_{i=1}^{n} z_{ij}x_i}{\sum_{i=1}^{n} z_{ij}}$$

E-step

- Initialize α_j=1/m, λ_j=j*10+10, m=10
- Iterate E-step and M-step until converge (the difference between updated α_j, λ_j<1e-05)

Estimate the number of k-mers occur 0 to 3 times

$$\sum_{j=1}^{m} \frac{p_{\mathrm{j}}\left(\lambda_{\mathrm{j}}, x\right) \sum_{i=1 \& x_i \geq 4}^{n} Z_{\mathrm{ij}}}{1-\sum_{s=0}^{3} p_{\mathrm{j}}\left(\lambda_{\mathrm{j}}, s\right)}$$

- x_i for i<4 are inaccurate because of the existence of low abundance species and sequencing errors.

Estimate species number

Genome size $$g_{\mathrm{j}} = \frac{\sum_{i=1}^{n'} Z_{\mathrm{ij}} * x_i}{\lambda_{\mathrm{j}}}$$

- Delete small groups if $g_{\mathrm{j}} <= 400{,}000$, the size of the sequenced smallest genome of living organisms

Assign reads confidently

x: the median frequency of k-mers in each read

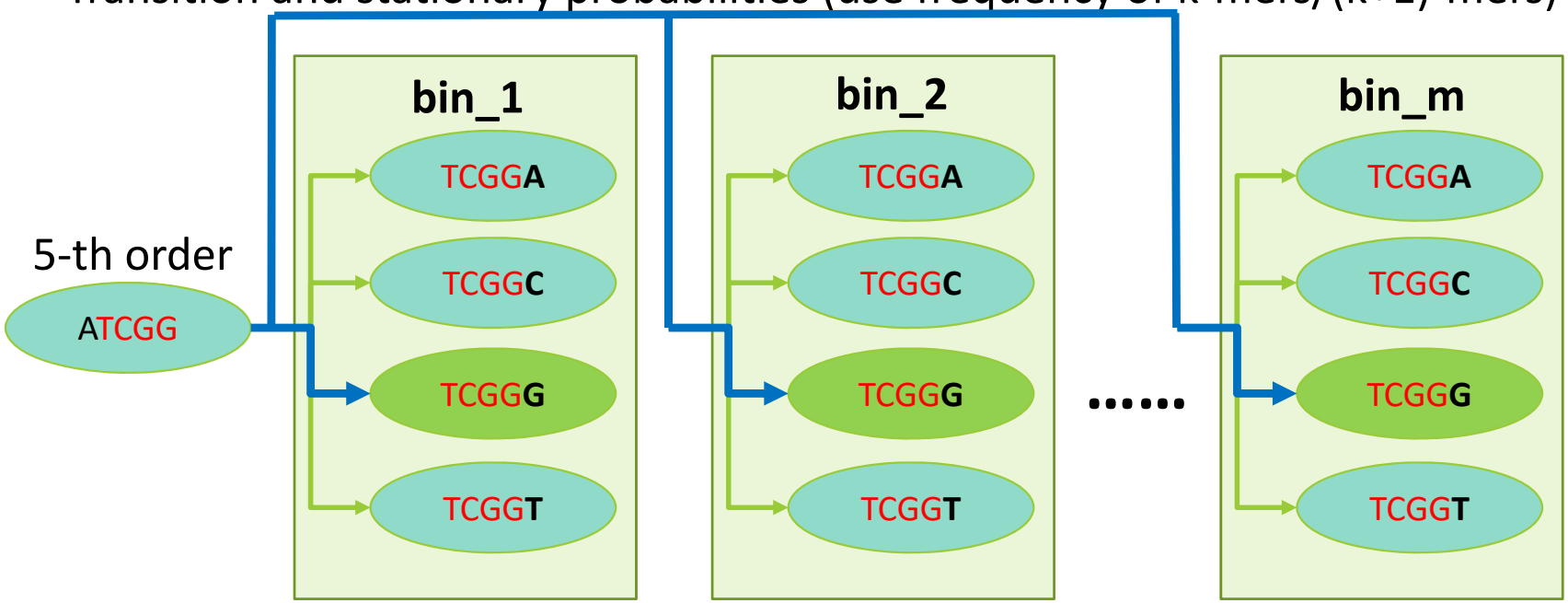
Calculate: $p_j(\lambda_j, x)$, j=1,2,..,m

For a read, only if its largest probability minus the second largest probability is larger than a cutoff C (C=0.5), this read will be assigned

-> unassign reads …

Assign reads to trained Markov chains

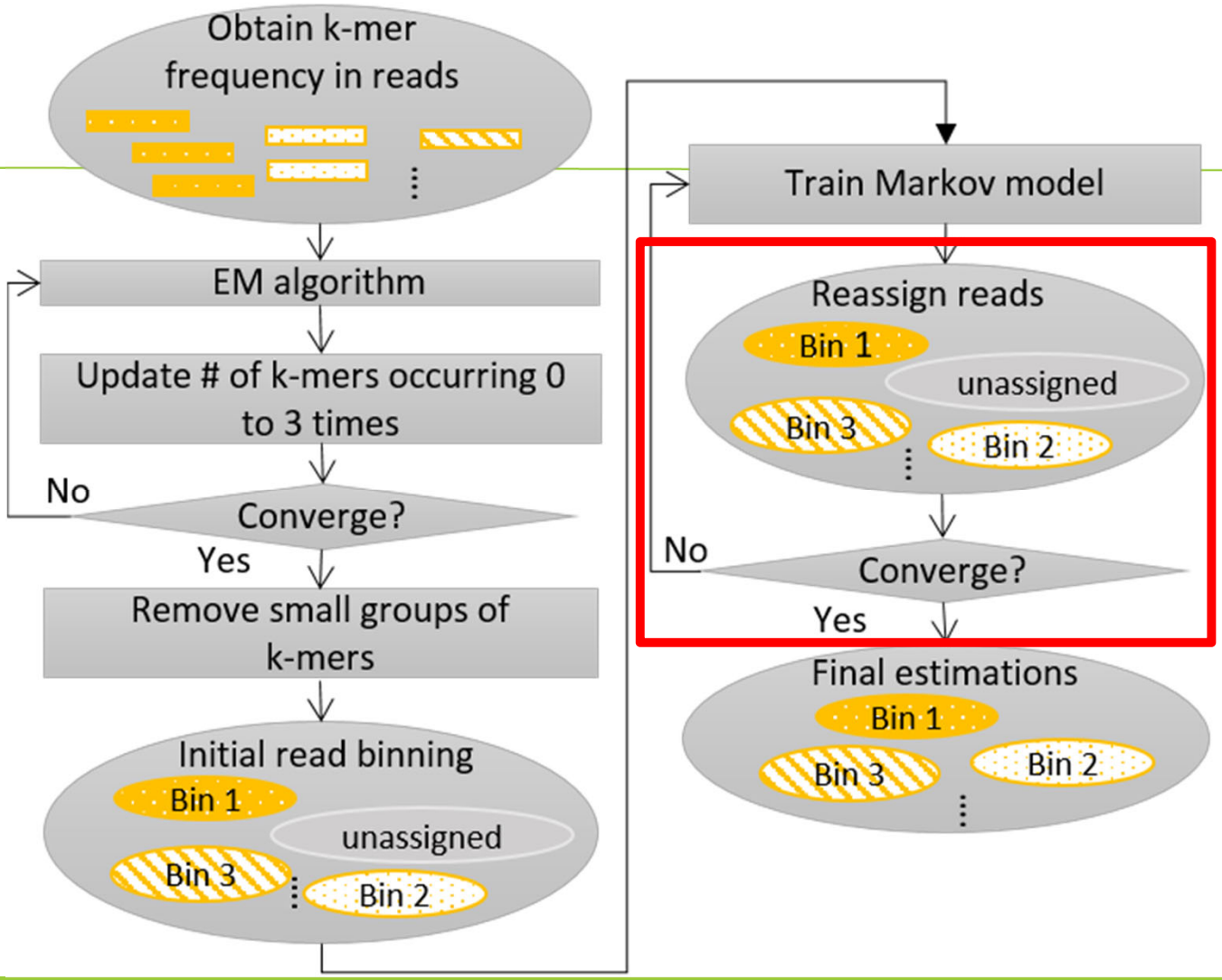
Transition and stationary probabilities (use frequency of k-mers/(k+1)-mers)



Assignment score

$$S(a_1 a_2 \dots a_k) * T(a_{k+1} | a_1 a_2 \dots a_k) * T(a_{k+2} | a_2 a_3 \dots a_{k+1}) * \cdots * T(a_n | a_{n-k} a_{n-k-1} \dots a_{n-1})$$

a_i->nucleotide at the i-th position of this read, n ->the length of read
S -> transition probability, T -> stationary probability



The procedure of read clustering in MBBC. The output on the right from each of the main steps on the left is connected with the corresponding steps.
Wang *et al. BMC Bioinformatics* 2015 **16**:36 doi:10.1186/s12859-015-0473-8

Results1 MBBC reliably estimates the species number, genome sizes, relative species abundances, and k-mer coverage

A. Initial prediction of α, λ										
Initial Species	1	2	3	4	5	6	7	8	9	10
α	43.30%	22.97%	11.07%	20.84%	1.16%	0.51%	0.12%	0.03%	0.00%	0.00%
λ	3.88	11.14	16.57	23.61	38.71	51.62	74.22	105.37	158.79	329.53

B. Prediction after updating #k-mers that occur 0 to 3 times										
α	31.59%	16.01%	25.79%	24.33%	1.38%	0.72%	0.14%	0.03%	0.01%	0.00%
λ	3.34	6.67	13.05	22.98	35.61	49.23	72.22	103.45	156.95	328.64

C. Prediction after removing small groups of k-mers										
Genome size	3009885	660737	1005524	948301	~~53786~~	~~27871~~	~~5352~~	~~1249~~	~~197~~	~~36~~
α	31.59%	16.01%	25.79%	24.33%	~~1.38%~~	~~0.72%~~	~~0.14%~~	~~0.03%~~	~~0.01%~~	~~0.00%~~
λ	3.34	6.67	13.05	22.98	~~35.61~~	~~49.23~~	~~72.22~~	~~103.45~~	~~156.95~~	~~328.64~~

D. Prediction after iteratively binning read based on Markov chains: Predicted (real data)				
Predicted Species	1	2	3	4
Genome size	1498994 (1160554)	825923 (945296)	1138156 (1107344)	1212248 (1075140)
α	9.42% (6.98%)	10.35% (11.36%)	27.91% (29.95%)	52.33% (51.70%)
λ	3.34 (3.49)	6.67 (5.83)	13.05 (12.48)	22.98 (20.52)

MBBC: an efficient approach for metagenomic binning based on clustering
BMC Bioinformatics. 2015;16(1):36.

Results1 MBBC reliably estimates the species number, genome sizes, relative species abundances, and k-mer coverage

A. Initial prediction of α, λ										
Initial Species	1	2	3	4	5	6	7	8	9	10
α	43.30%	22.97%	11.07%	20.84%	1.16%	0.51%	0.12%	0.03%	0.00%	0.00%
λ	3.88	11.14	16.57	23.61	38.71	51.62	74.22	105.37	158.79	329.53

B. Prediction after updating #k-mers that occur 0 to 3 times										
α	31.59%	16.01%	25.79%	24.33%	1.38%	0.72%	0.14%	0.03%	0.01%	0.00%
λ	3.34	6.67	13.05	22.98	35.61	49.23	72.22	103.45	156.95	328.64

C. Prediction after removing small groups of k-mers										
Genome size	3009885	660737	1005524	948301	~~53786~~	~~27871~~	~~5352~~	~~1249~~	~~197~~	~~36~~
α	31.59%	16.01%	25.79%	24.33%	~~1.38%~~	~~0.72%~~	~~0.14%~~	~~0.03%~~	~~0.01%~~	~~0.00%~~
λ	3.34	6.67	13.05	22.98	~~35.61~~	~~49.23~~	~~72.22~~	~~103.45~~	~~156.95~~	~~328.64~~

D. Prediction after iteratively binning read based on Markov chains: Predicted (real data)				
Predicted Species	1	2	3	4
Genome size	1498994 (1160554)	825923 (945296)	1138156 (1107344)	1212248 (1075140)
α	9.42% (6.98%)	10.35% (11.36%)	27.91% (29.95%)	52.33% (51.70%)
λ	3.34 (3.49)	6.67 (5.83)	13.05 (12.48)	22.98 (20.52)

MBBC: an efficient approach for metagenomic binning based on clustering
BMC Bioinformatics. 2015;16(1):36.

Results1 MBBC reliably estimates the species number, genome sizes, relative species abundances, and k-mer coverage

A. Initial prediction of α, λ										
Initial Species	1	2	3	4	5	6	7	8	9	10
α	43.30%	22.97%	11.07%	20.84%	1.16%	0.51%	0.12%	0.03%	0.00%	0.00%
λ	3.88	11.14	16.57	23.61	38.71	51.62	74.22	105.37	158.79	329.53

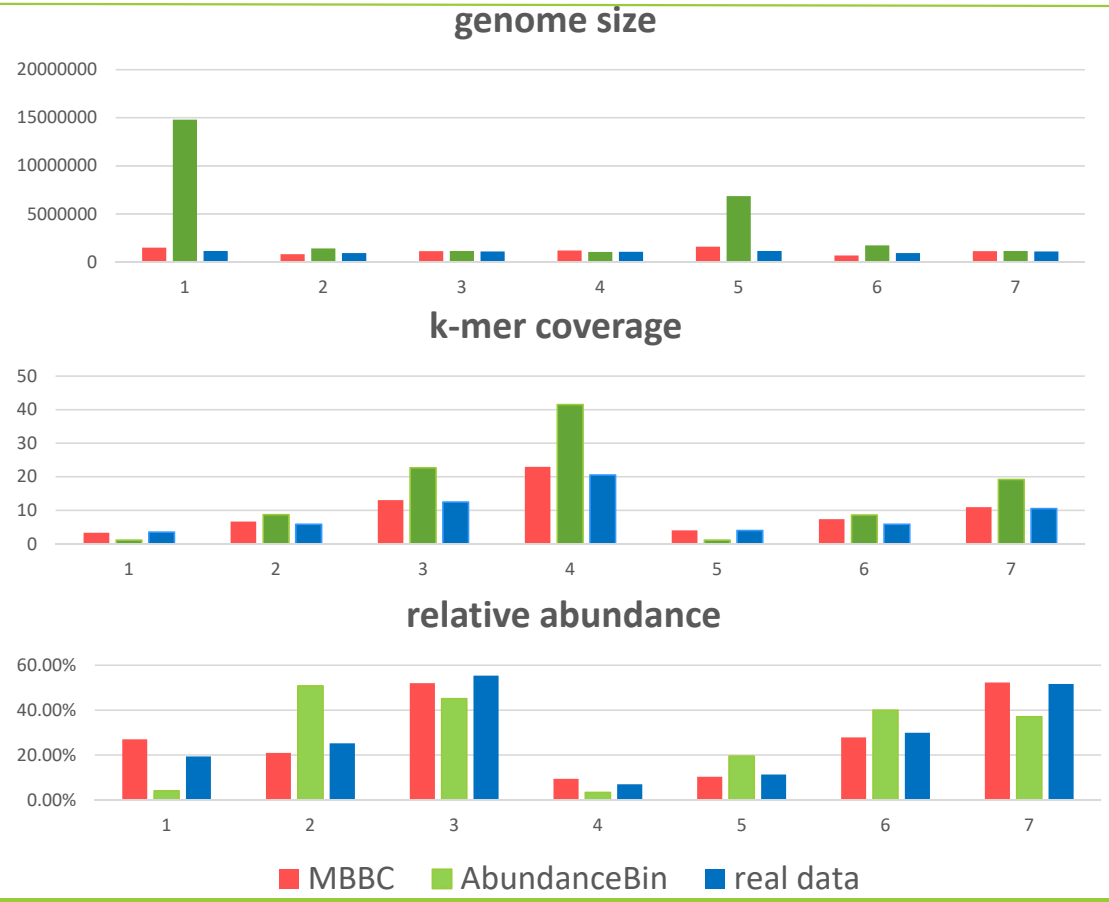
B. Prediction after updating #k-mers that occur 0 to 3 times										
α	31.59%	16.01%	25.79%	24.33%	1.38%	0.72%	0.14%	0.03%	0.01%	0.00%
λ	3.34	6.67	13.05	22.98	35.61	49.23	72.22	103.45	156.95	328.64

C. Prediction after removing small groups of k-mers										
Genome size	3009885	660737	1005524	948301	~~53786~~	~~27871~~	~~5352~~	~~1249~~	~~197~~	~~36~~
α	31.59%	16.01%	25.79%	24.33%	~~1.38%~~	~~0.72%~~	~~0.14%~~	~~0.03%~~	~~0.01%~~	~~0.00%~~
λ	3.34	6.67	13.05	22.98	~~35.61~~	~~49.23~~	~~72.22~~	~~103.45~~	~~156.95~~	~~328.64~~

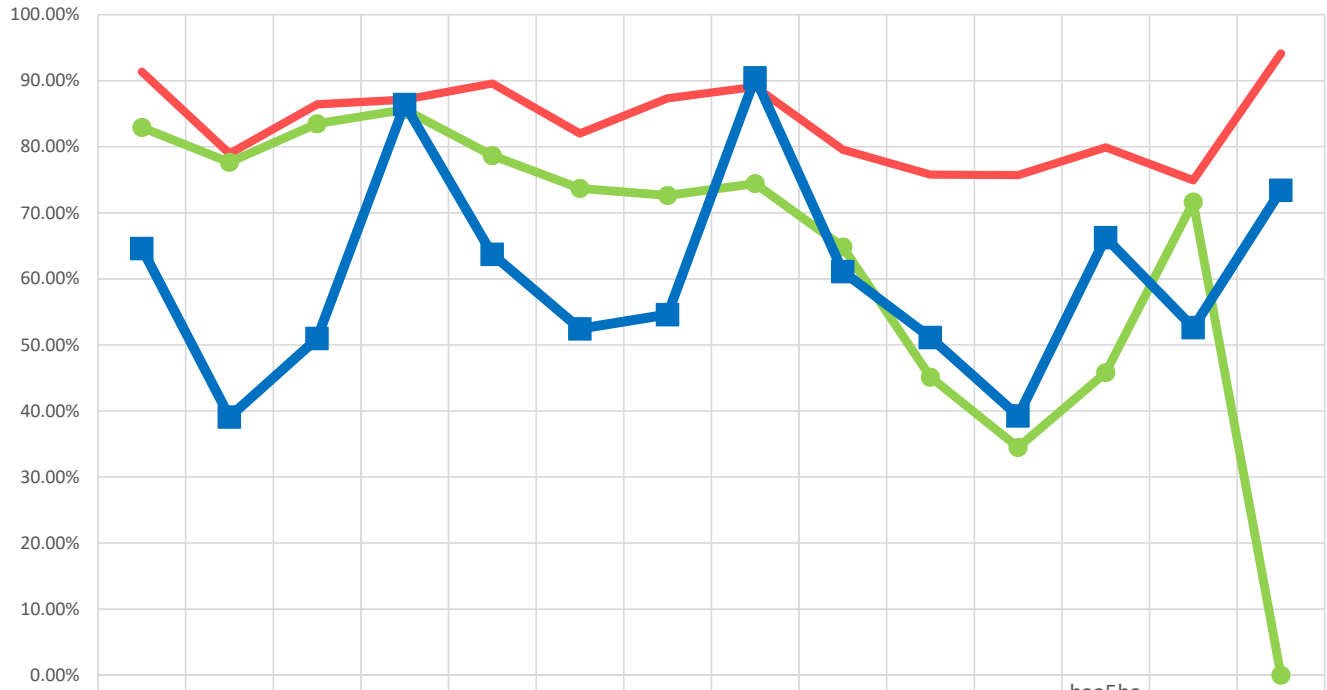
D. Prediction after iteratively binning read based on Markov chains: Predicted (real data)				
Predicted Species	1	2	3	4
Genome size	1498994 (1160554)	825923 (945296)	1138156 (1107344)	1212248 (1075140)
α	9.42% (6.98%)	10.35% (11.36%)	27.91% (29.95%)	52.33% (51.70%)
λ	3.34 (3.49)	6.67 (5.83)	13.05 (12.48)	22.98 (20.52)

MBBC: an efficient approach for metagenomic binning based on clustering
BMC Bioinformatics. 2015;16(1):36.

Results1 MBBC reliably estimates the species number, genome sizes, relative species abundances, and k-mer coverage



Results 2 MBBC reliably assigns reads

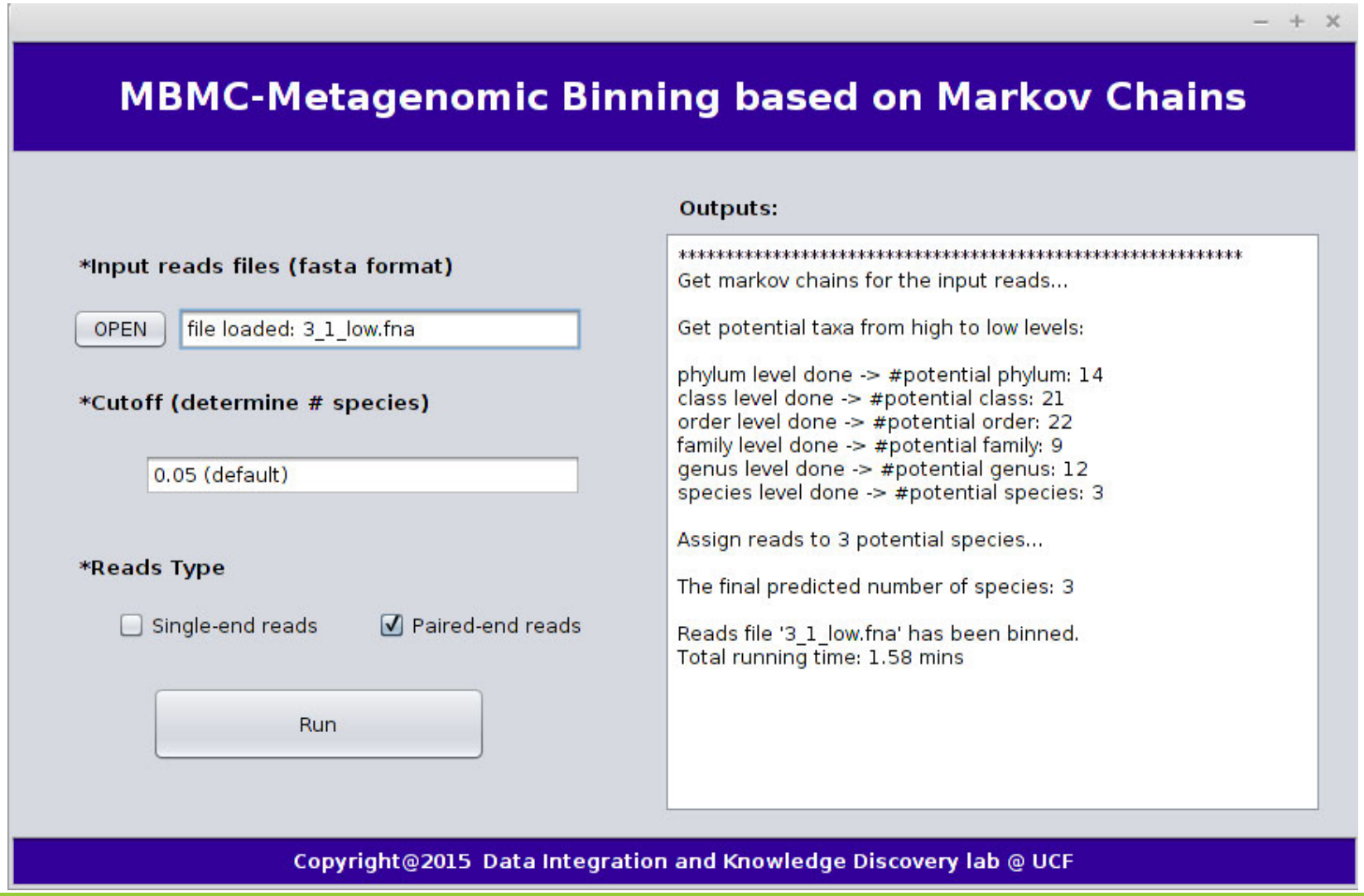


	lag5lar11las24	lag4lar7las12	laa4lag8lar15las30	laa4lag8lar15las30(no errors)	spa4spd9sps18	spa5spd8sps15	spa4spd8sps18spt32	spa4spd8sps18spt32(no errors)	baa3bab7bac15	baa6bab10bac18	baa5bab10bac18bah30	baa5bab10bac18bah30(no errors)	human gut dataset	AMD dataset
MBBC	91.34%	78.97%	86.43%	87.13%	89.58%	82.01%	87.35%	89.09%	79.55%	75.80%	75.71%	79.90%	74.94%	94.14%
MetaCluster	82.93%	77.66%	83.49%	85.64%	78.68%	73.71%	72.64%	74.43%	64.83%	45.12%	34.48%	45.82%	71.65%	0
AbundanceBin	64.60%	39.09%	50.98%	86.41%	63.73%	52.44%	54.60%	90.44%	61.11%	51.13%	39.25%	66.25%	52.63%	73.42%

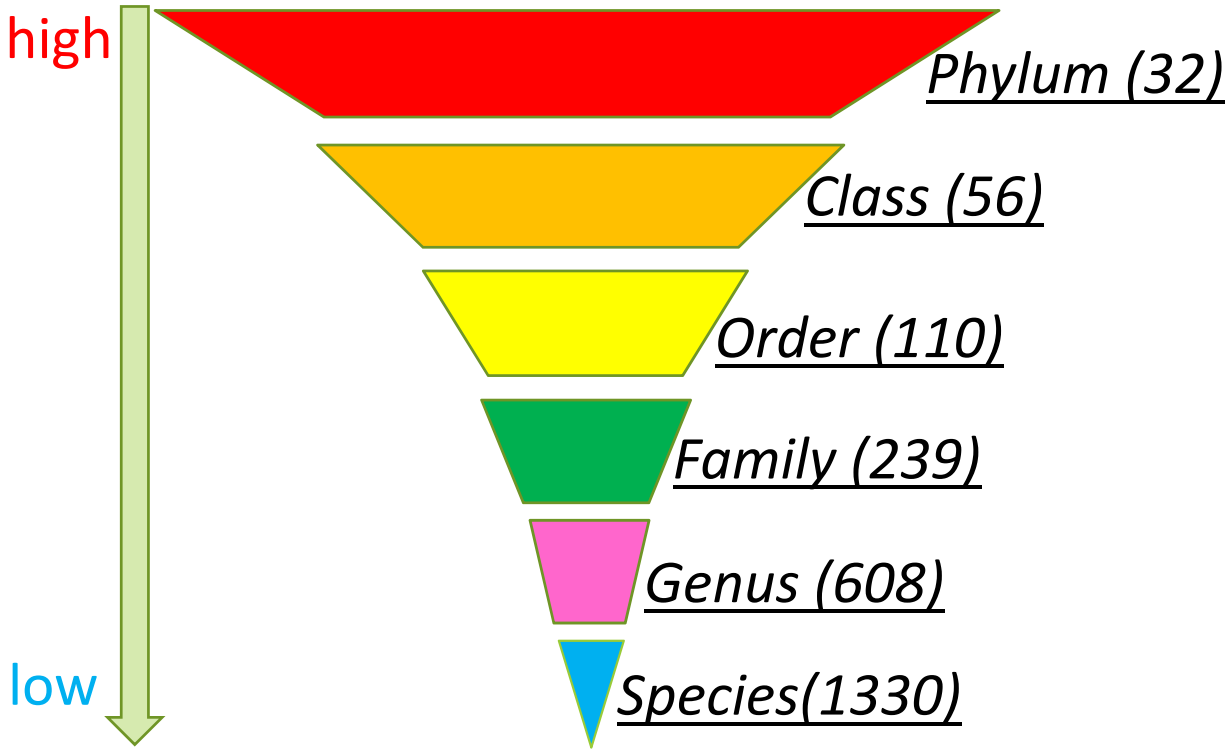
Problems in MBBC, MetaCluster, AbundanceBin

- Difficulty to bin reads with low abundances or similar abundances

→a common problem in taxonomy independent method


MBMC-Metagenomic Binning based on Markov Chains
*Input reads files (fasta format)
OPEN
file loaded: 3_1_low.fna
*Cutoff (determine # species)
0.05 (default)
*Reads Type
Single-end reads
Paired-end reads
Run
Outputs:
****
Get markov chains for the input reads...
Get potential taxa from high to low levels:
phylum level done -> #potential phylum: 14
class level done -> #potential class: 21
order level done -> #potential order: 22
family level done -> #potential family: 9
genus level done -> #potential genus: 12
species level done -> #potential species: 3
Assign reads to 3 potential species...
The final predicted number of species: 3
Reads file '3_1_low.fna' has been binned.
Total running time: 1.58 mins
Copyright@2015 Data Integration and Knowledge Discovery lab @ UCF

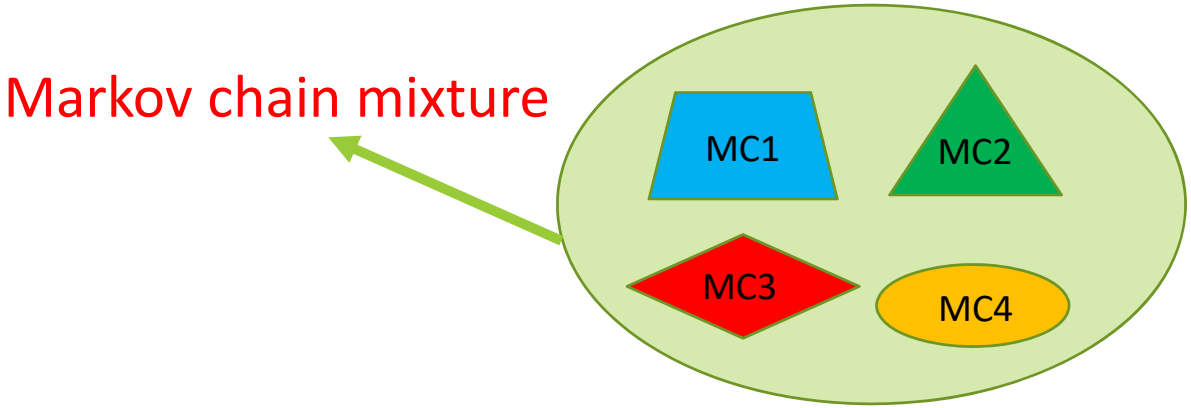
Reference database



each taxon was represented by a 9-th order Markov chain

Markov chain for input reads

- All input reads were represented by a 9-th order Markov chain



Estimate relative abundance

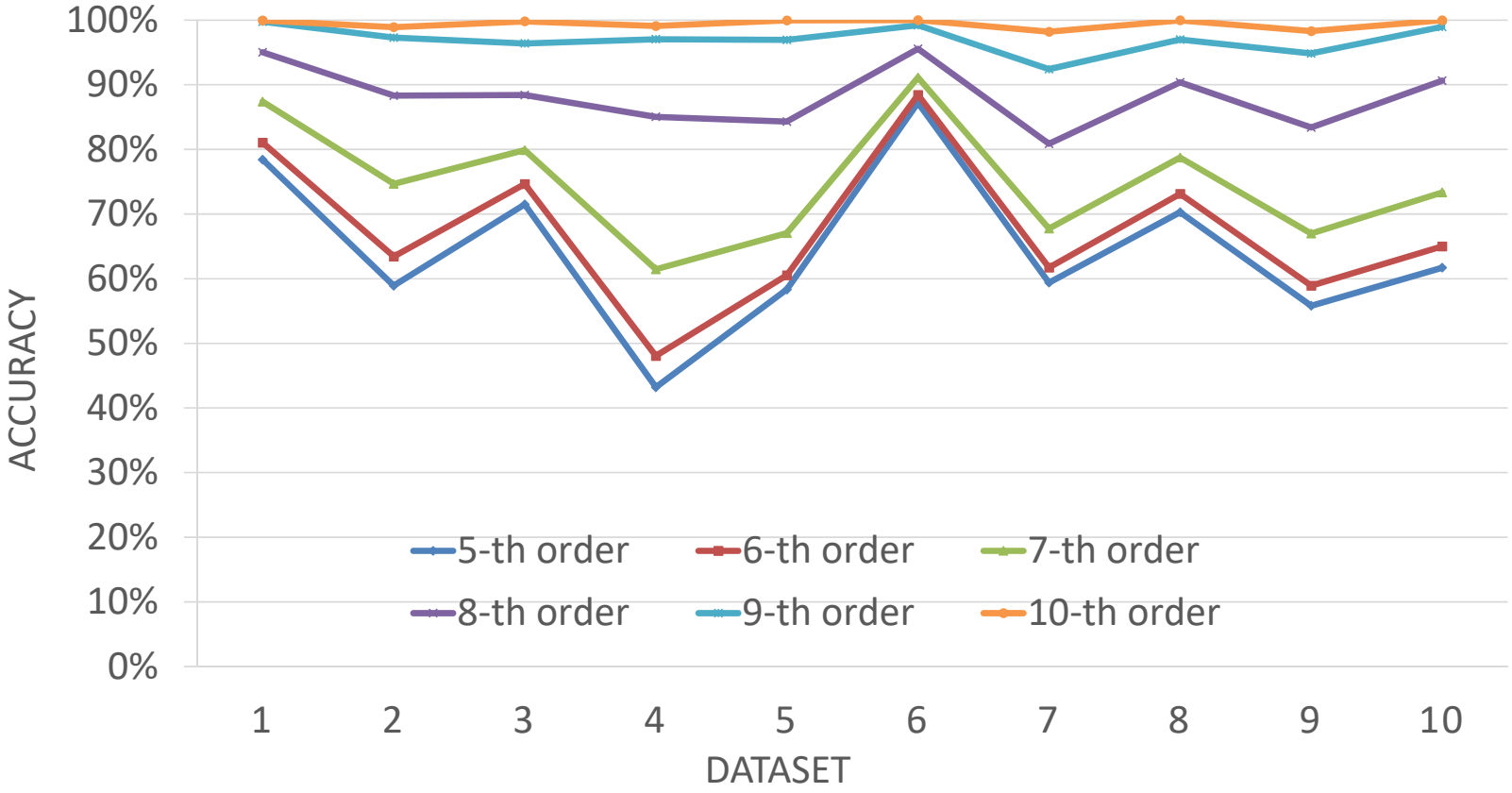
$$y = \mathbf{X}\beta + \epsilon$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, X = \begin{pmatrix} X_1^T \\ X_2^T \\ \dots \\ X_n^T \end{pmatrix} = \begin{pmatrix} x_{11} \dots x_{1p} \\ x_{21} \dots x_{2p} \\ \dots \\ x_{n1} \dots x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \dots \\ \boldsymbol{\beta_p} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_p \end{pmatrix}$$

β_j approximates the relative abundance of reads from a taxon

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Results 1 9-th Markov chain models are effective in representing microbial genomes



OMICS. 2016 Aug;20(8):470-9. doi: 10.1089/omi.2016.0081. Epub 2016 Jul 22.

Results 2 MBMC reliably predicted the species number and accurately grouped reads in simulated datasets

dataset[species #]	MBMC [m]	MetaCluster	Abundancebin	MEGAN5 [m]	Kraken [m]
1_1[5]	96.96%[5]	42.56%	na	94.19%[4]	92.00%[4]
1_2[6]	96.25%[6]	91.36%	na	97.02%[6]	90.57%[5]
2_1[5]	99.04%[5]	42.76%	34.48%	88.37%[5]	92.60%[5]
2_2[6]	95.26%[6]	89.38%	33.50%	97.71%[6]	89.16%[6]
3_1[5]	89.11%[5]	36.19%	29.55%	99.01%[5]	94.59%[5]
3_2[6]	94.37%[7]	87.20%	38.50%	93.25%[7]	93.32%[6]
4_1[5]	97.36%[5]	41.57%	28.07%	97.39%[5]	96.27%[5]
4_2[6]	86.01%[5]	90.42%	28.06%	98.82%[6]	96.43%[6]
5_1[5]	93.36%[5]	40.78%	28.28%	82.02%[6]	91.68%[5]
5_2[6]	69.39%[6]	51.02%	na	89.55%[6]	90.50%[7]

Results 3 MBMC worked well on datasets with unknown species

dataset[species #]	MBMC [m]	MetaCluster	Abundancebin	MEGAN5 [m]	Kraken [m]
0_1	65.39%[12]	63.60%	39.71%	0.00%[0]	0.00%[0]
0_1*	65.47%[13]	na	38.76%	0.00%[0]	0.00%[0]
0_2	68.55%[11]	60.16%	39.66%	0.00%[0]	0.00%[0]
0_2*	68.05%[12]	na	38.68%	0.00%[0]	0.00%[0]
0_3	79.15%[13]	61.56%	36.59%	0.00%[0]	0.00%[0]
0_3*	77.23%[12]	na	36.25%	0.00%[0]	0.00%[0]
0_4	63.87%[15]	58.38%	35.00%	0.00%[0]	0.00%[0]
0_4*	64.09%[15]	na	35.04%	0.00%[0]	0.00%[0]
0_5	75.02%[15]	57.15%	37.34%	0.00%[0]	0.00%[0]
0_5*	76.22%[14]	na	37.51%	0.00%[0]	0.00%[0]

Results 4 MBMC performed much better than other methods on experimental datasets

dataset(known species#[species#])	MBMC [m]	MetaCluster	Abundancebin	MEGAN5 [m]	Kraken [m]
SRS017080(1[4])	80.62%[8]	35.65%	41.02%	72.55%[5]	56.50%[5]
SRS013705(1[5])	42.63%[8]	14.70%	32.47%	13.85%[1]	10.70%[1]
HMP mock(6[6])	76.63%[5]	na	na	81.79%[6]	62.52%[6]
human gut(1[3])	83.16%[7]	71.77%	69.68%	14.72%[1]	11.38%[1]

Problems in MBMC

- MBMC tended to divide reads from an unknown species into multiple small bins.

→How do we know there are unknown species in the dataset?

→How to bin reads better for these unknown species?

Metagenomic binning based on clustering of Markov chains

Improve the method of MBMC:

- long k-mers
- Reads Assembly

Question

How to use the sequenced genomes across taxa to train a model to bin reads? Alternative approaches that extend the idea of MBMC.