Big Data Analytics in Metagenomics:

Integration, Representation, Management, and Visualization

Yang Lu

USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Microbes are everywhere on the earth

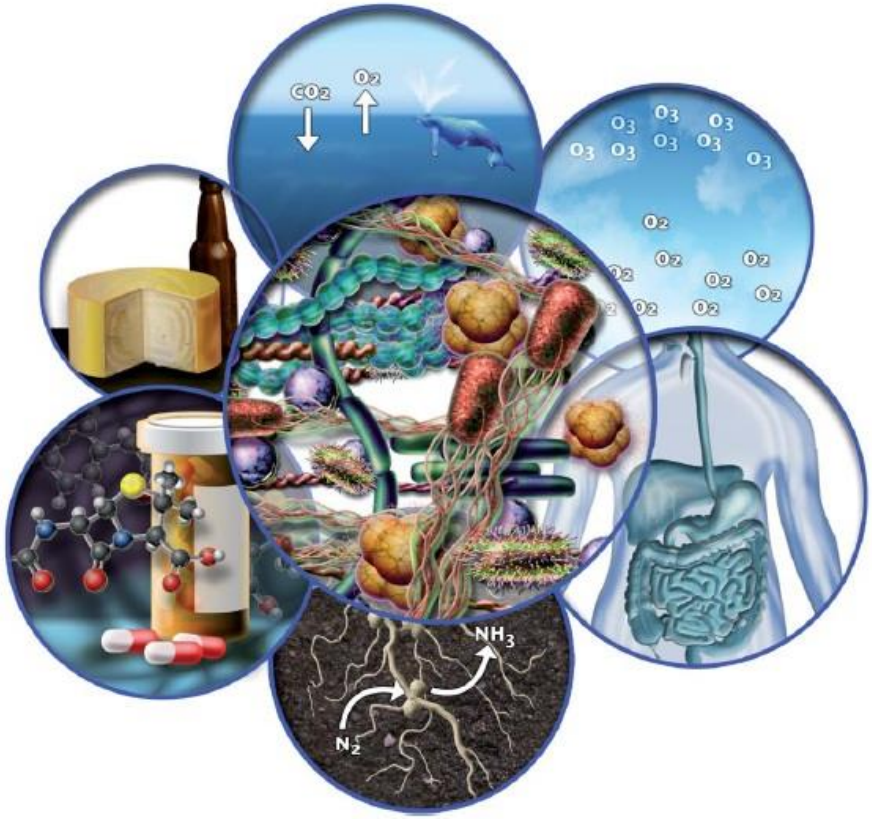
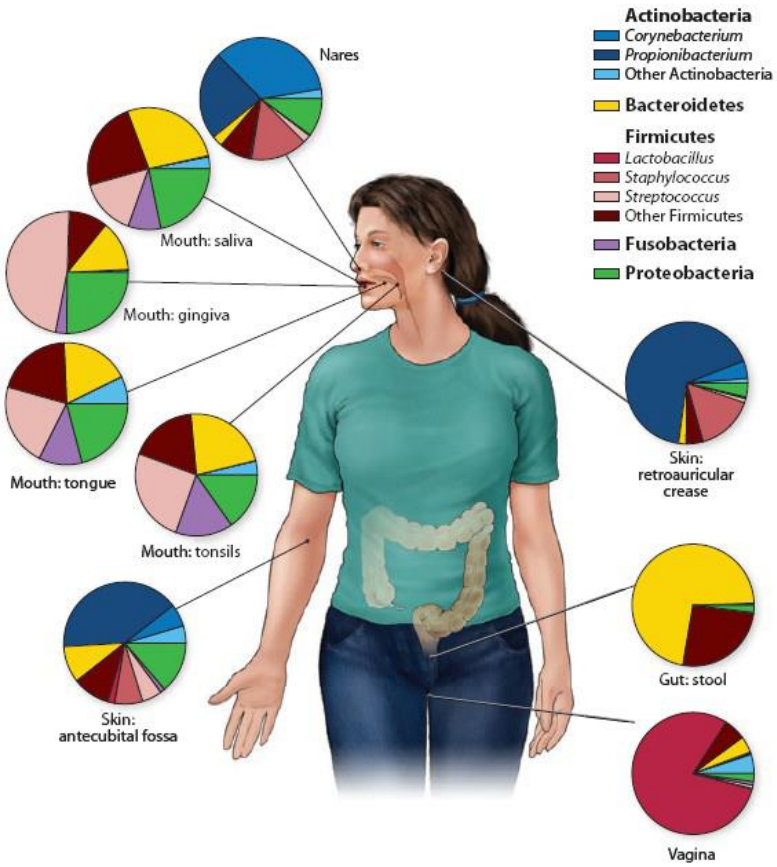


Image Source: Encyclopedia of Life Sciences

Microbes are everywhere in human body

- Microbiome as extended human genome
- 10^{13} human cells vs. 10^{14} bacterial cells
- $\geq 3 \times 10^{6}$ genes provided by gut microbiome
- Understanding Diseases
  - Obesity
  - Diabetes
  - neurodevelopmental disorders
  - …



Li (2015)

Microbiome study

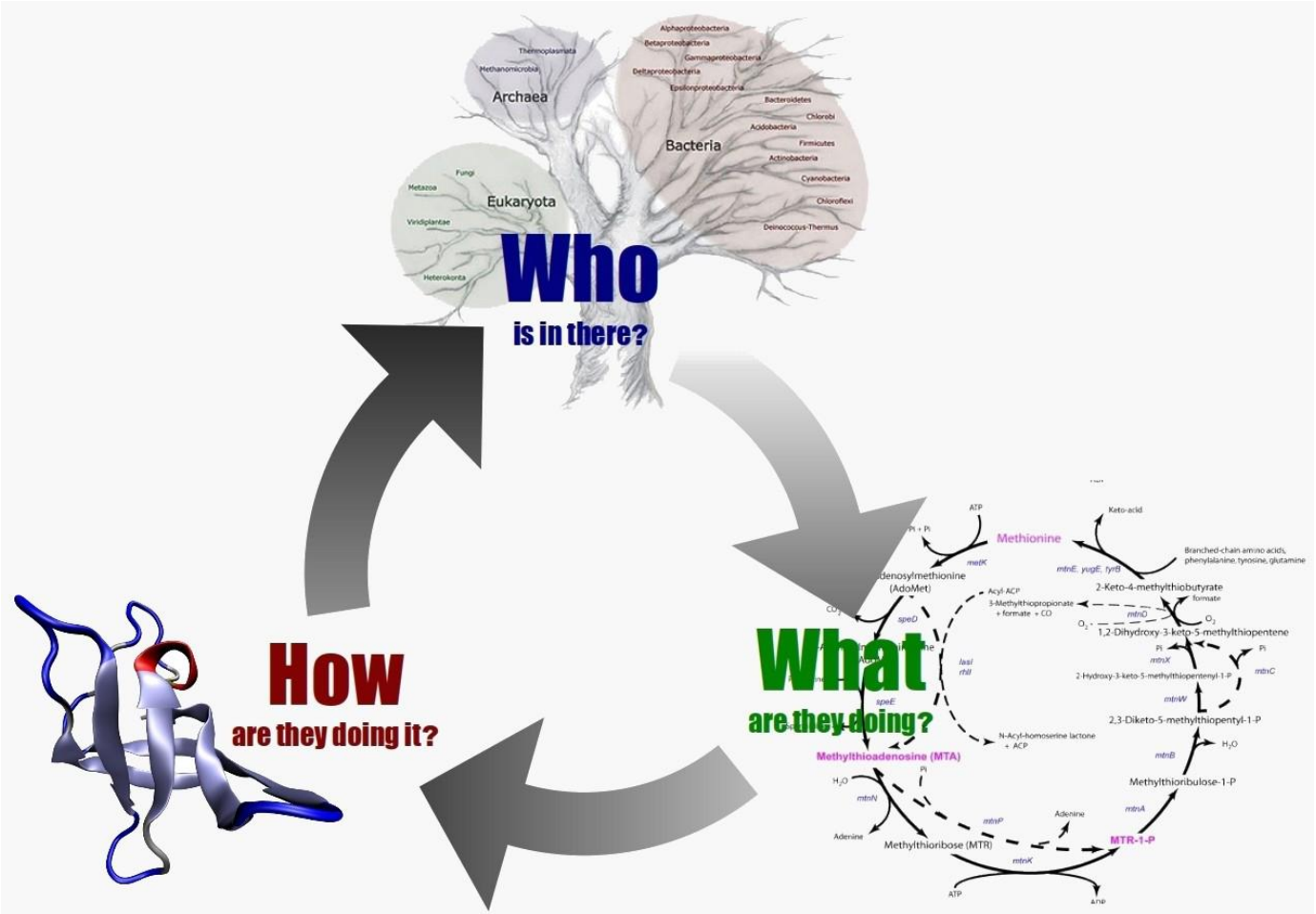
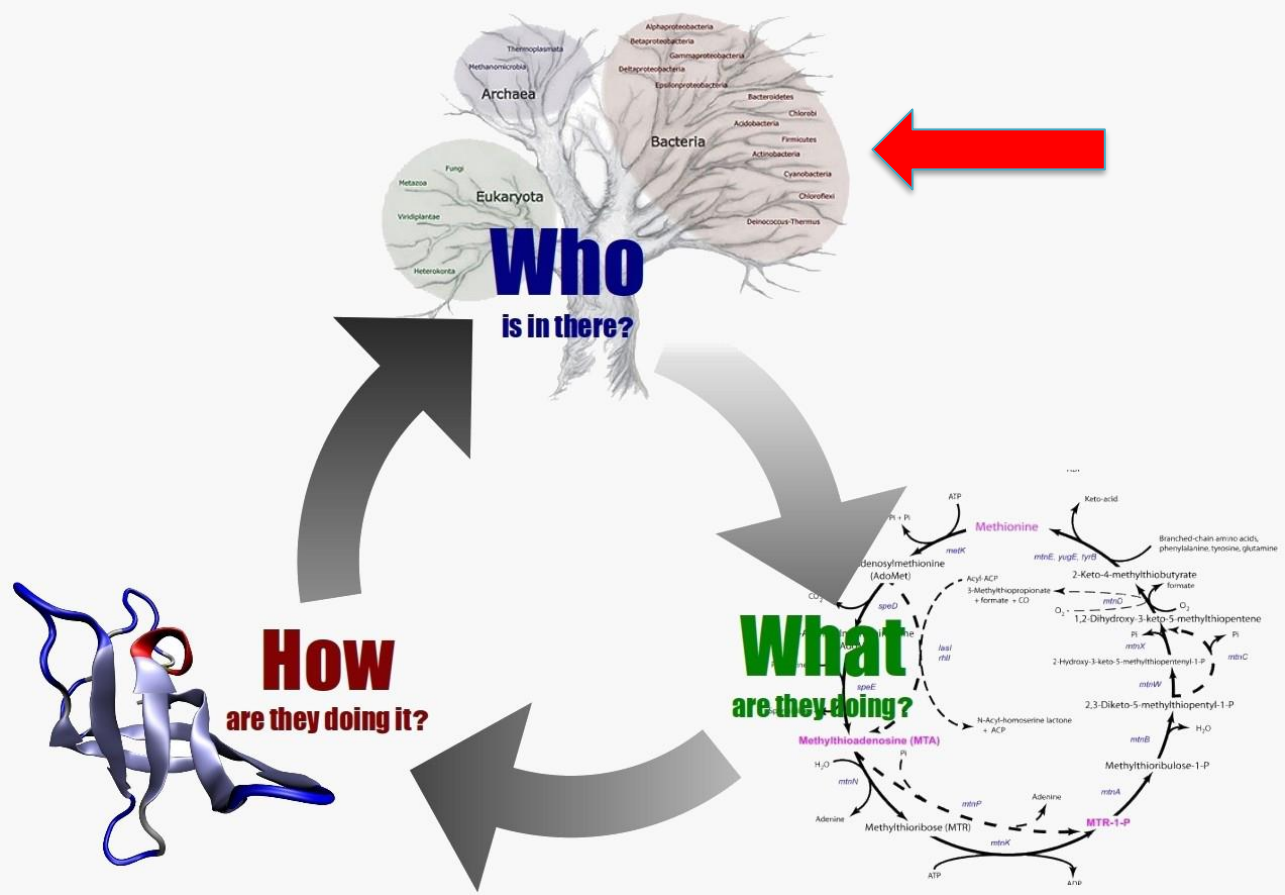


Image Source: http://users.metu.edu.tr/bicgen/research/envg.html

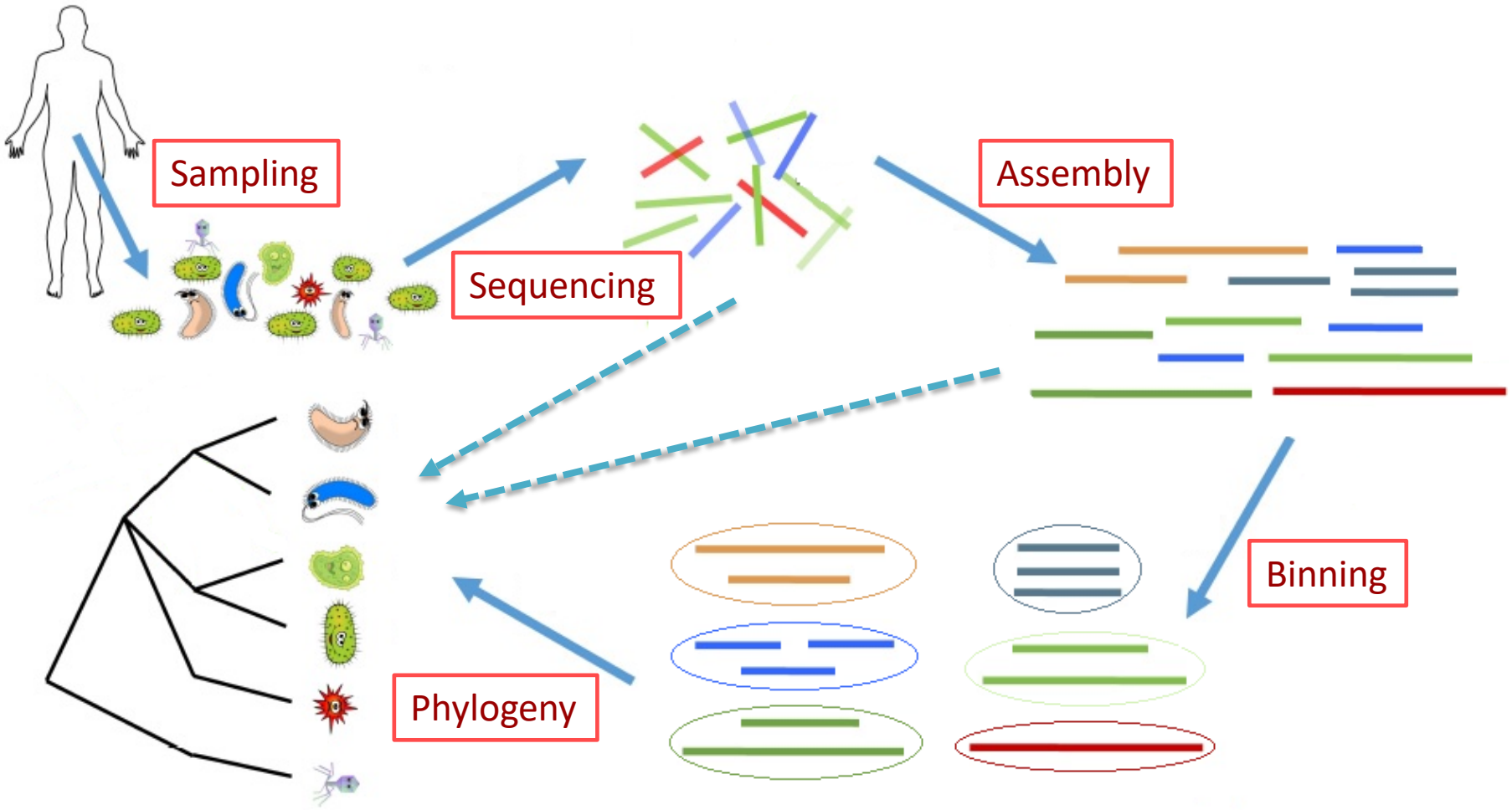
Motivation of my research



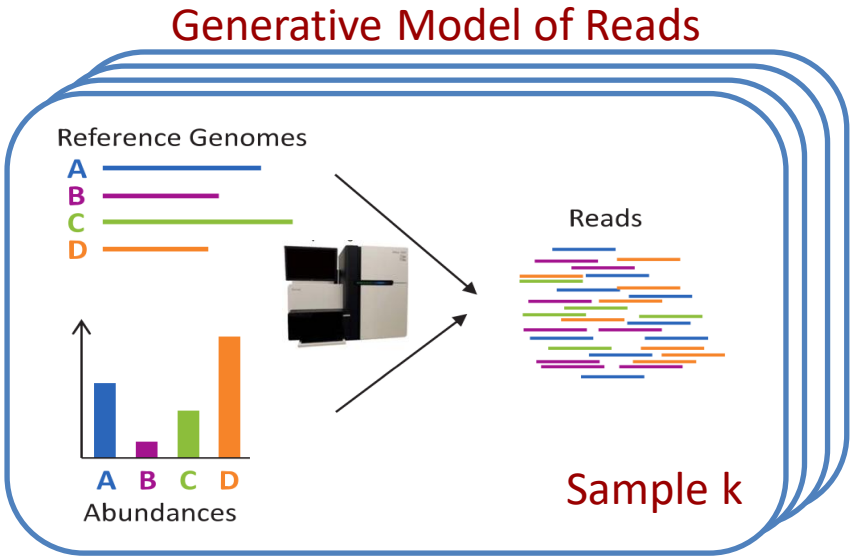
USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Image Source: http://users.metu.edu.tr/bicgen/research/envg.html

Pipeline



Shotgun Sequencing



Two types of approaches:

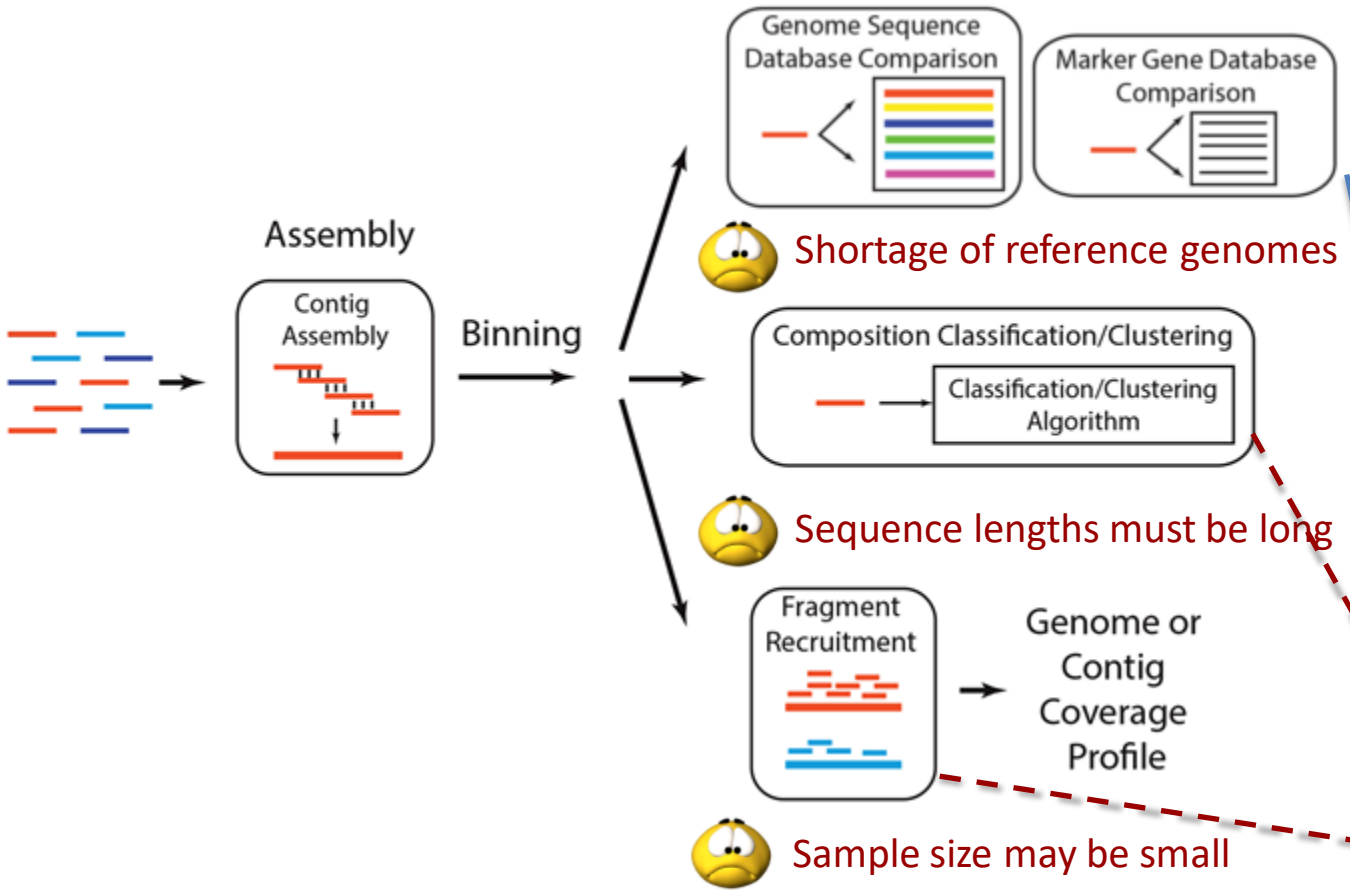
- Sequencing only specific marker genes
  - e.g. 16S rDNA gene in the bacterial genomes
  - Low sensitivity in the species and strain levels

- Sequencing all genomes
  - Demand high sequencing depth to detect rare taxa

Chen and Li (2013); Zhang et al. (2017)

Metagenomics Binning

Group contigs into Operational Taxonomic Units (OTUs).



- MEGAN
- MG-RAST
- AMPHORA
- PhyloSift
- MetaPhlAn
- Kraken
- Phymm

- Self Organizing Maps
- CompositBin

- AbundanceBin

- CONCOCT
- GroopM
- MetaBAT
- MaxBin

Sharpton (2014)

Part I

COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge

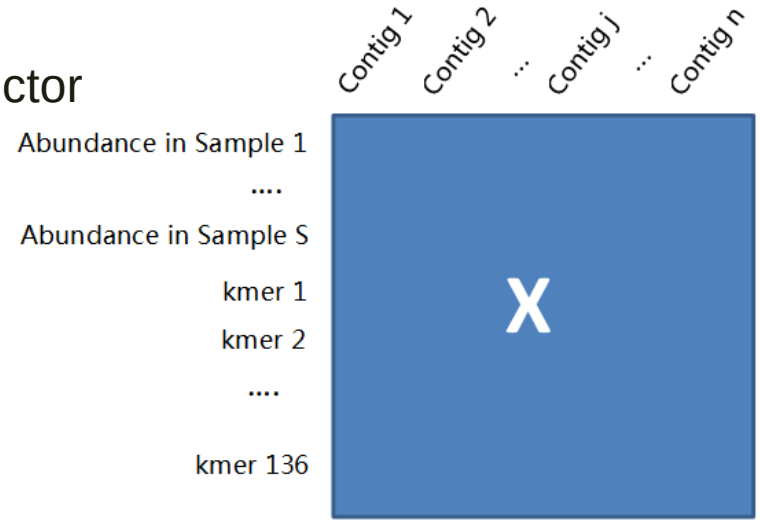
Availability: https://github.com/younglululu/COCACOLA

Publication: Lu et al. (2017) Bioinformatics

Problem Formulation

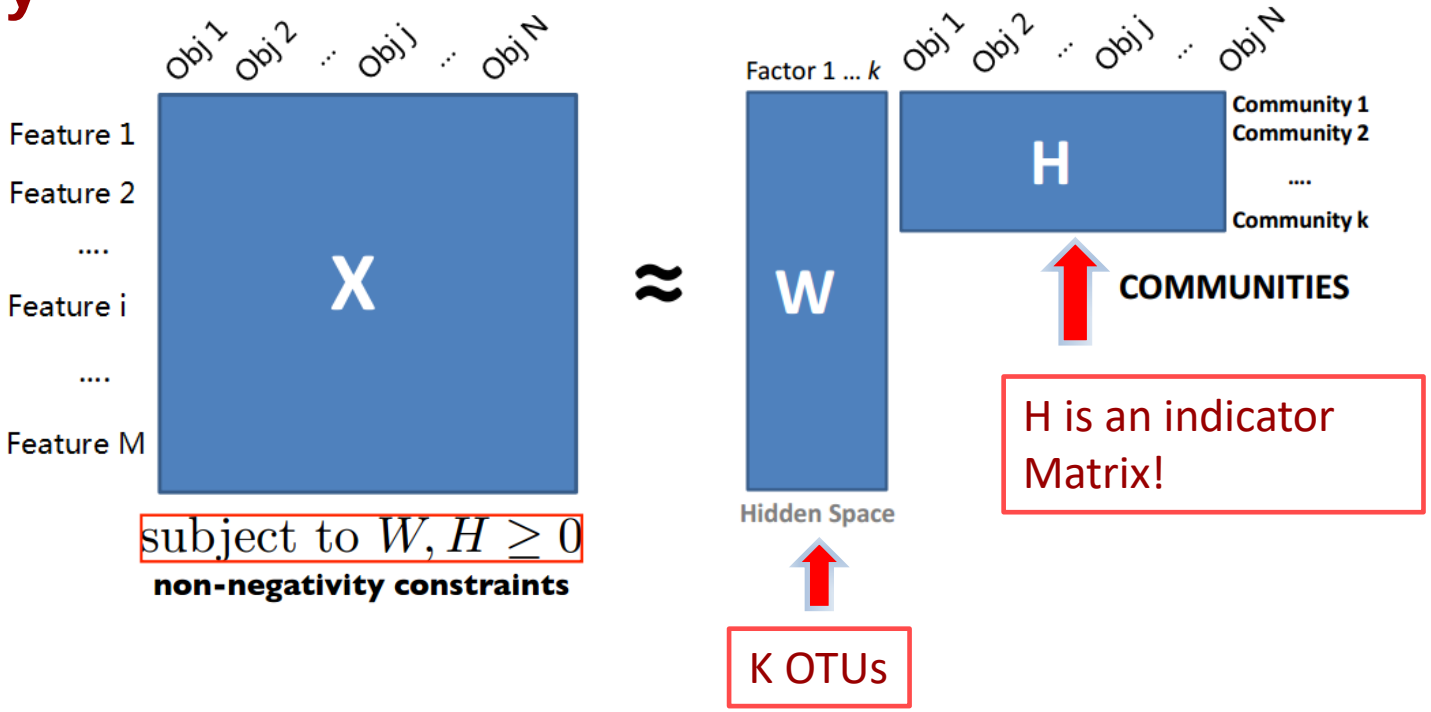
- There are co-assembled N contigs

- Each contig j is represented by a feature vector
  - Abundance profile
  - Tetramer Composition profile
  - Denoted as $x_{.j}$



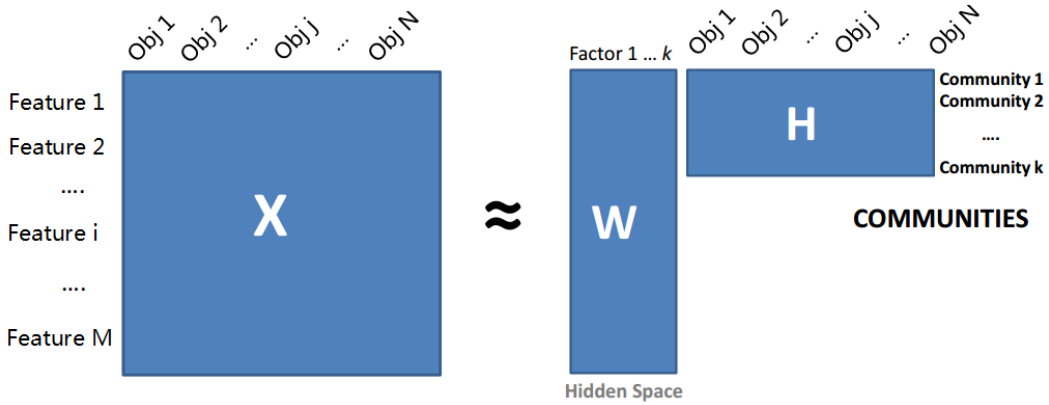
- Assume there are K OTUs
  - Each OTU k can also be represented as a corresponding feature vector (Latent)
  - Denoted as $w_{.k}$

Ideally



Whether contig j belongs to OTU k is denoted as an indicator $\mathbb{h}_{kj}$

$$\mathbf{x}_{.j} = \mathbb{h}_{1j}\mathbf{w}_{.1} + \mathbb{h}_{2j}\mathbf{w}_{.2} + \mathbb{h}_{3j}\mathbf{w}_{.3} + \cdots + \mathbb{h}_{kj}\mathbf{w}_{.k}$$



$$\mathbf{x}_{\cdot 1} = \mathbb{h}_{11}\mathbf{w}_{\cdot 1} + \mathbb{h}_{21}\mathbf{w}_{\cdot 2} + \mathbb{h}_{31}\mathbf{w}_{\cdot 3} + \cdots + \mathbb{h}_{k1}\mathbf{w}_{\cdot k}$$

$$\mathbf{x}_{\cdot 2} = \mathbb{h}_{12}\mathbf{w}_{\cdot 1} + \mathbb{h}_{22}\mathbf{w}_{\cdot 2} + \mathbb{h}_{32}\mathbf{w}_{\cdot 3} + \cdots + \mathbb{h}_{k2}\mathbf{w}_{\cdot k}$$

$$\cdots$$

$$\mathbf{x}_{\cdot N} = \mathbb{h}_{1N}\mathbf{w}_{\cdot 1} + \mathbb{h}_{2N}\mathbf{w}_{\cdot 2} + \mathbb{h}_{3N}\mathbf{w}_{\cdot 3} + \cdots + \mathbb{h}_{kN}\mathbf{w}_{\cdot k}$$

$$X \approx W\mathbb{H}$$

$$s.t.\ \ W \geq 0,\ \ \mathbb{H} \in \{0,1\}^{K\times N}, \|\mathbb{H}_{\cdot j}\|_0 = 1\ for\ j = 1, 2, \cdots, N$$

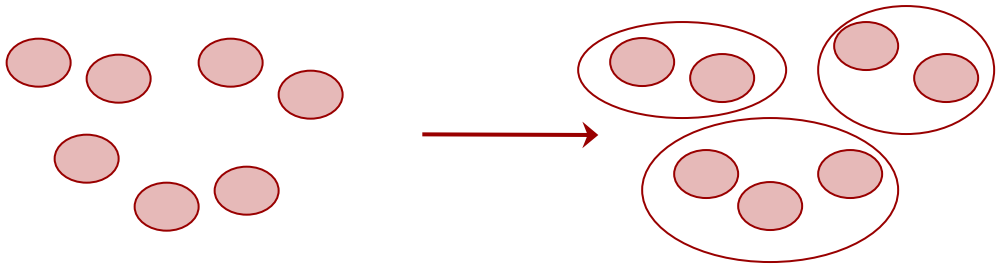
Relaxation. But…

$$\arg\min_{\substack{W\geq 0 \\ \mathbb{H}\in\{0,1\}^{K\times N}}} \|X - W\mathbb{H}\|_F^2$$

Hard to Solve!
Need Relaxation!

Hard Clustering

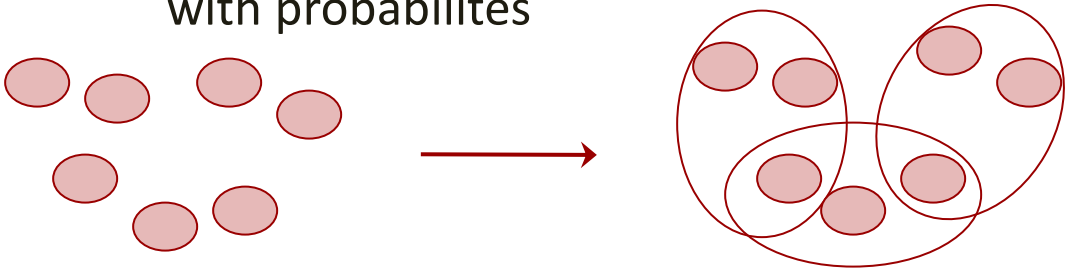
- Every contig belong to exactly one cluster.



$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2$$

Soft Clustering

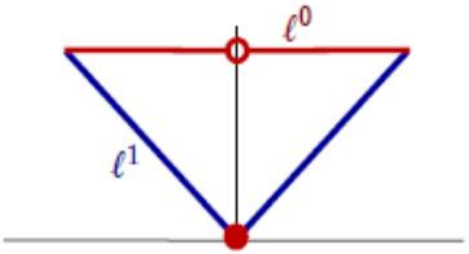
- Contigs may belong to several clusters with probabilites



To facilitate “hard clustering”-like behavior

$$\arg \min_{W,H \geq 0} \|X - WH\|_F^2 + \alpha \sum_{j=1}^{N} \|H_{\cdot j}\|_1^2$$

Sparse Non-negative Matrix Factorization

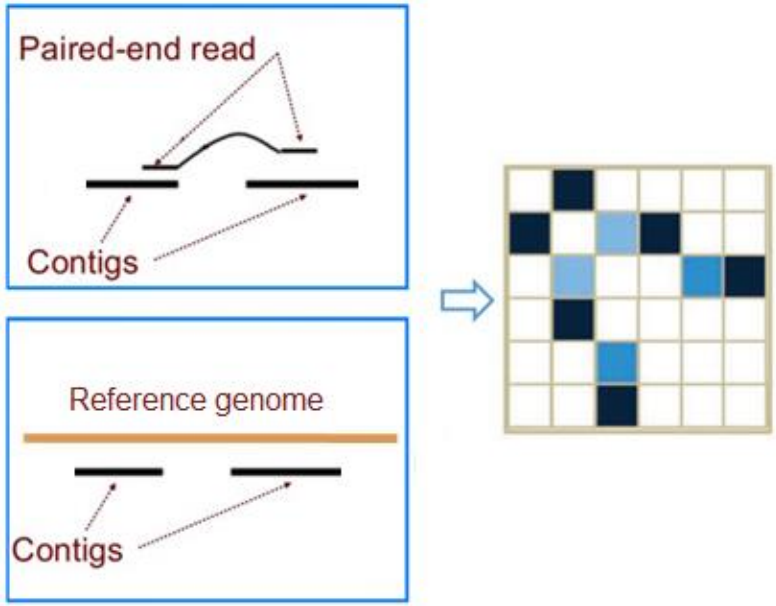


Kim and Park (2008)

Incorporating Additional Information

$$R_g = \sum_{n,n'=1}^{N} \|H_{.n} - H_{.n'}\|^2 \mathbf{A}_{nn'} = Tr(H\mathbf{L}H^T)$$

- $\boldsymbol{A_{ij}}$: the belief of two contigs i and j are in the same cluster
- $\boldsymbol{L}$: graph laplacian



Objective Function

$$\arg \min_{W,H \geq 0} \|X - WH\|_F^2 + \alpha \sum_{n=1}^{N} \|H_{\cdot n}\|_1^2 + \beta \, Tr(H \mathcal{L} H^T)$$

Solved by alternating nonnegative least squares

$$H \leftarrow \arg \min_{H \geq 0} \|X - WH\|_F^2 + \alpha \sum_{n=1}^{N} \|H_{\cdot n}\|_1^2 + \beta \, Tr(H \mathcal{L} H^T)$$

$$W \leftarrow \arg \min_{W \geq 0} \left\|X^T - H^T W^T\right\|_F^2$$

Block Coordinate Descent

$$\arg\min_{H\geq 0} \|X - WH\|_F^2 + \alpha \sum_{n=1} \|H_{\cdot n}\|_1^2 + \beta\, Tr(H\mathcal{L}H^T)$$

$$\approx \arg\min_{H\geq 0} \sum_{n=1}^{N} \left(\|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha \|H_{\cdot n}\|_1^2 + \beta\, H_{\cdot n}^T (H_{\cdot n} - 2 \sum_{n'=1}^{N} \mathcal{A}_{nn'} H_{\cdot n'}^{old}) \right)$$

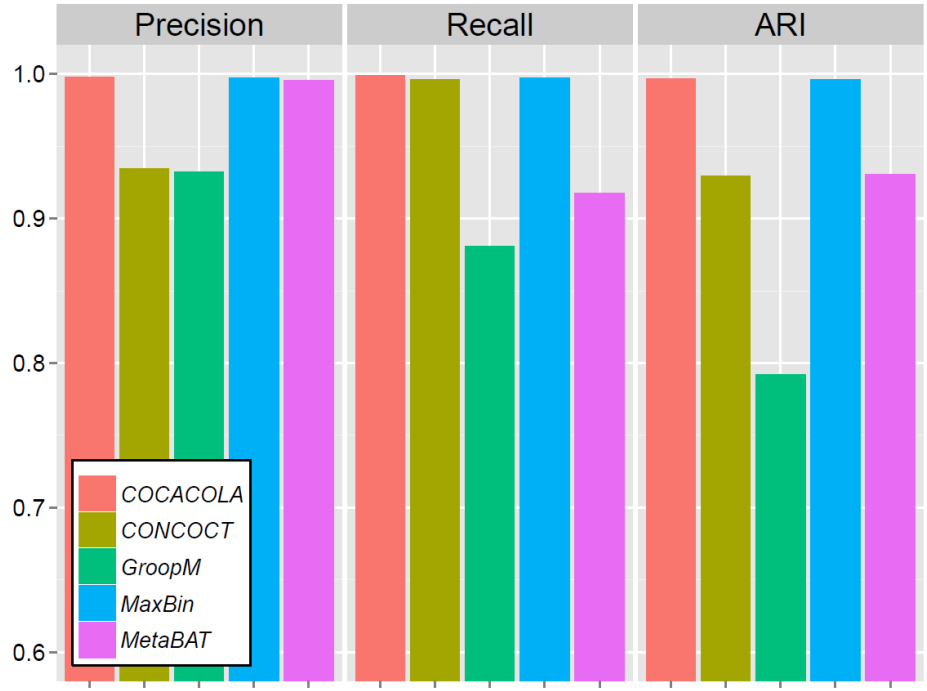
$$= \arg\min_{H\geq 0} \sum_{n=1}^{N} \left(\|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha \|H_{\cdot n}\|_1^2 + \beta \left\| H_{\cdot n} - \sum_{n'=1}^{N} \mathcal{A}_{nn'} H_{\cdot n'}^{old} \right\|_2^2 \right)$$

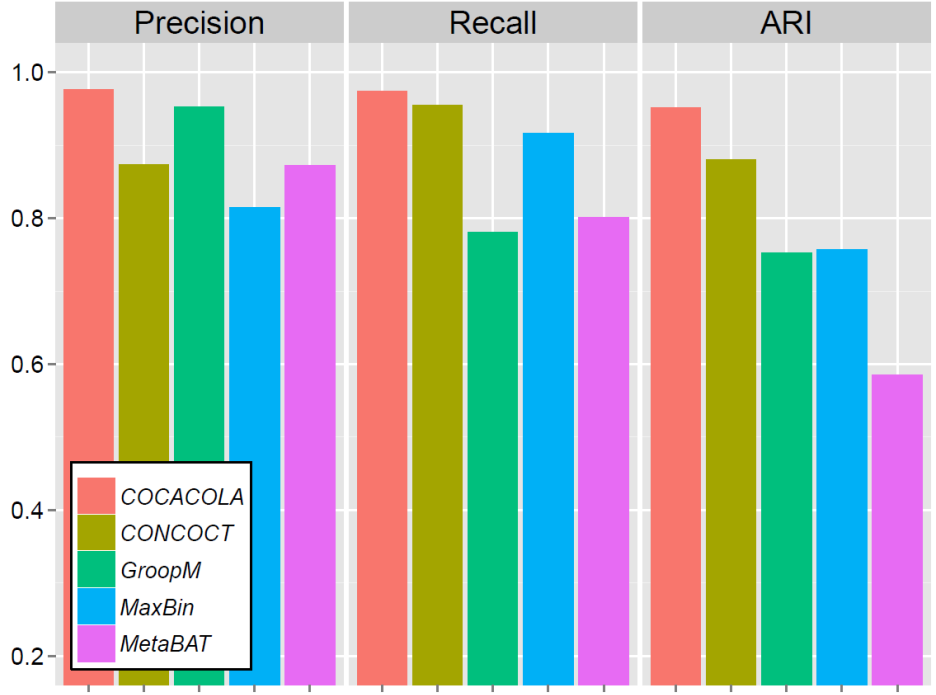
$$= \arg\min_{H} \left\| \begin{pmatrix} X \\ 0_{1\times N} \\ \sqrt{\beta} H^{old} \mathcal{A} \end{pmatrix} - \begin{pmatrix} W \\ \sqrt{\alpha} e_{1\times K} \\ \sqrt{\beta} I_K \end{pmatrix} H \right\|_F^2$$

Experiments

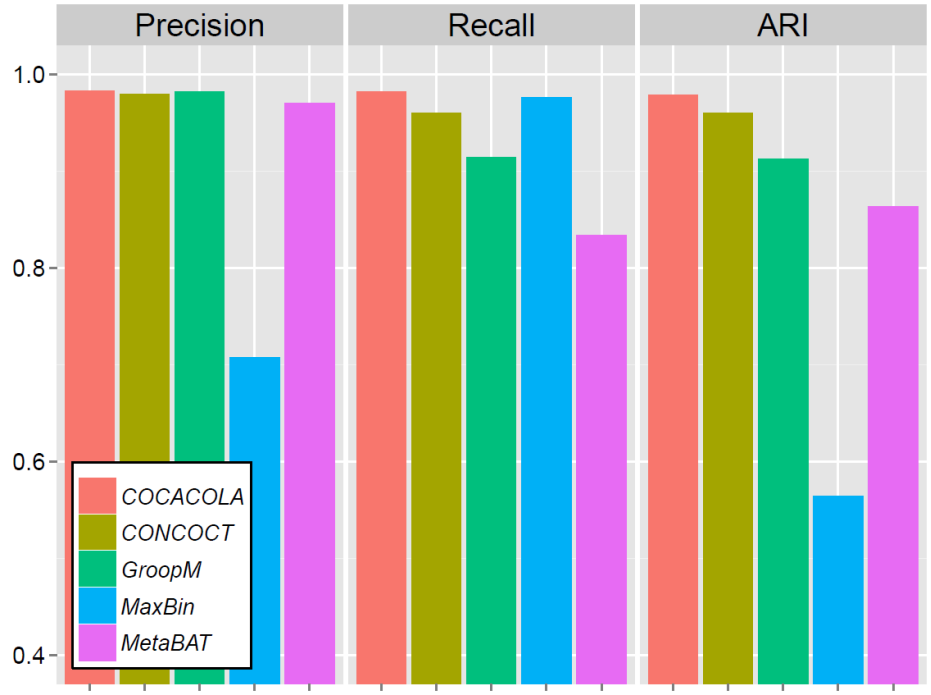
- Synthetic Datasets
  - Species Mock Community
    - 101 Species, 37,628 contigs, 96 Samples,
  - Strain Mock Community
    - Mixture of E. coli strains, five Bacteroides species, five Clostridium genera, five other typical gut bacteria
    - 9,417 contigs, 64 Samples
- Real Datasets
  - Sharon
    - 11 time-series samples from premature infant gut
    - 2,614 out of 5,579 contigs are labelled by TAXAassign
  - MetaHIT
    - 264 samples from MetaHIT consortium
    - 17,136 out of 192,673 contigs are labelled by TAXAassign

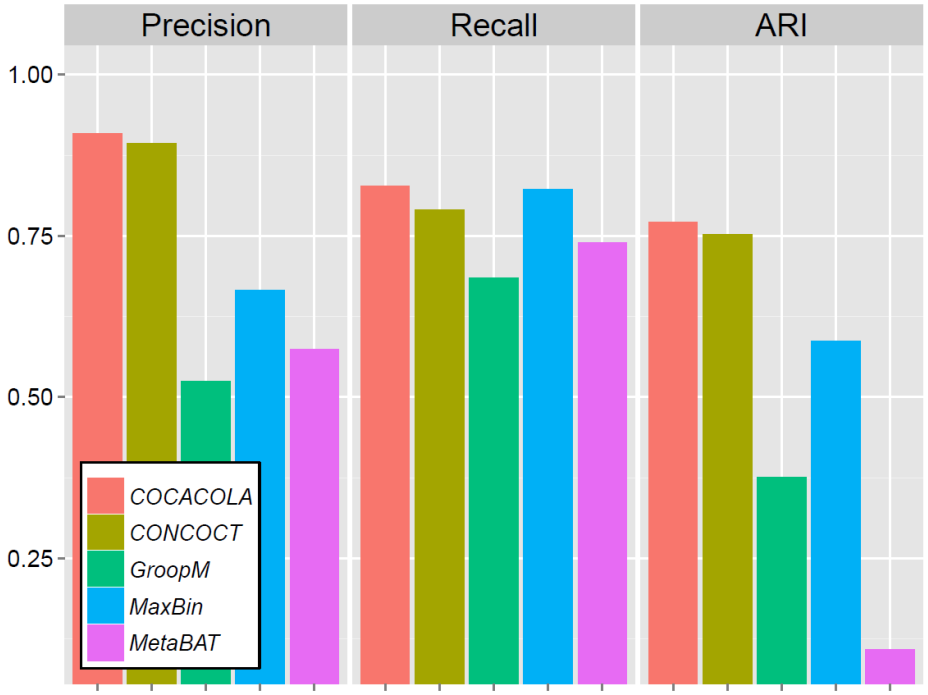
Synthetic “Species” and “Strain” Dataset



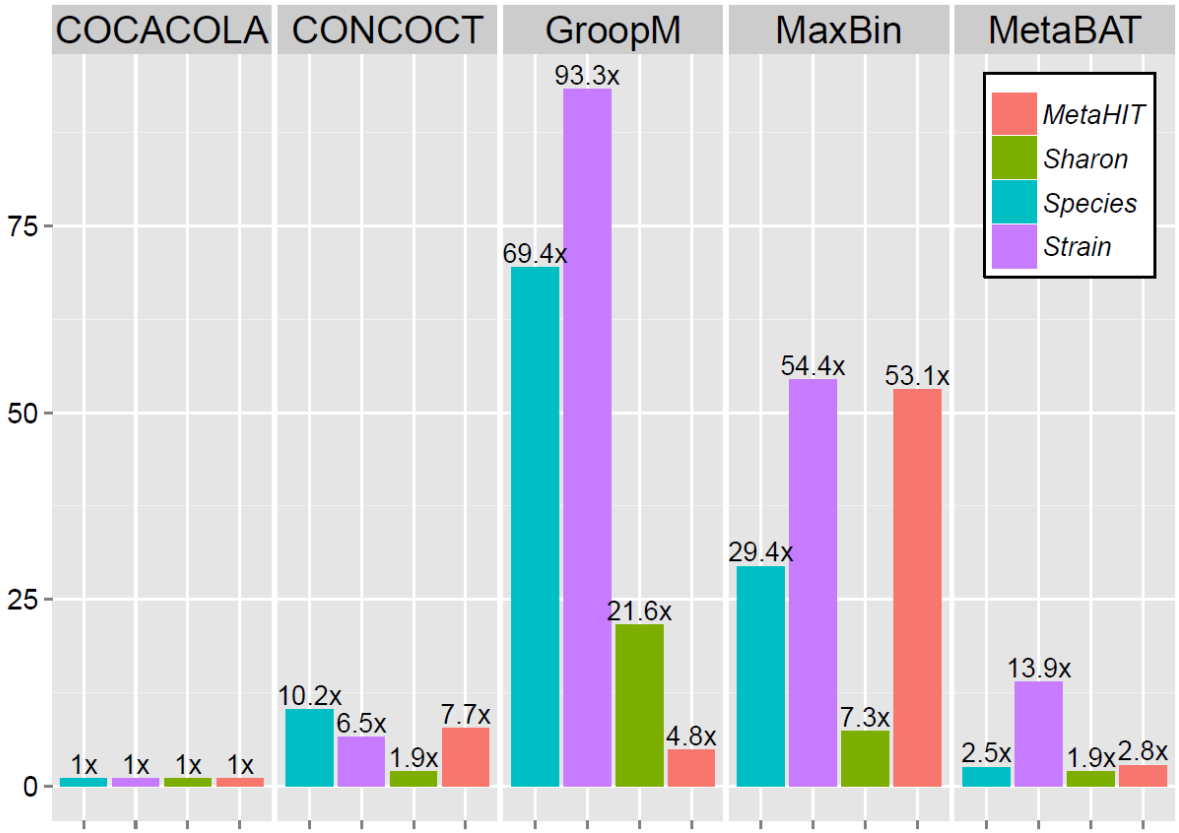


Real “Sharon” and “MetaHIT” Dataset

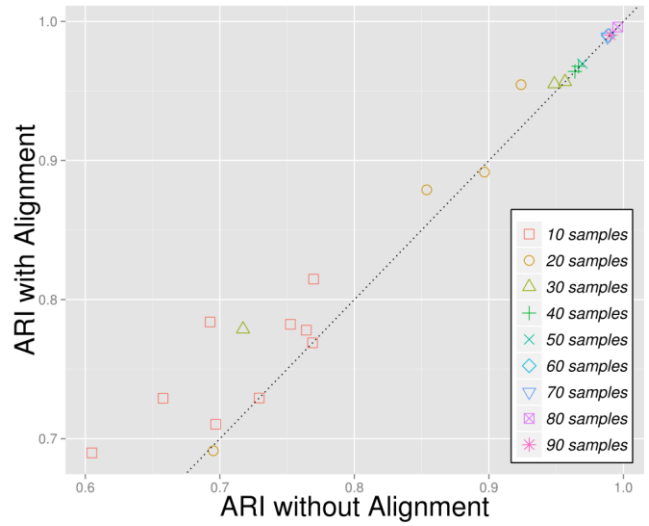


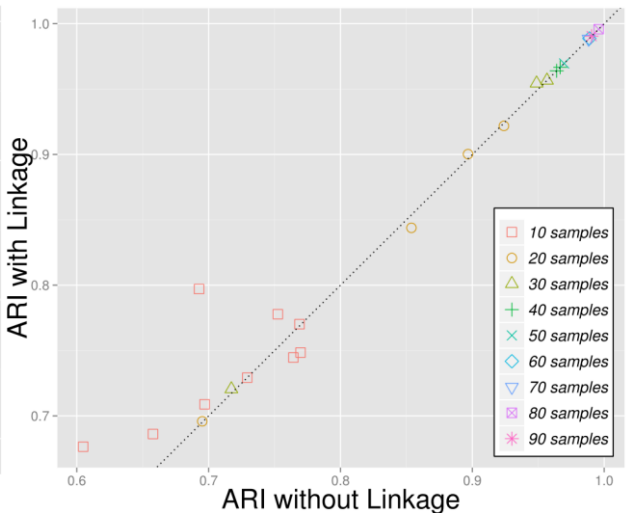


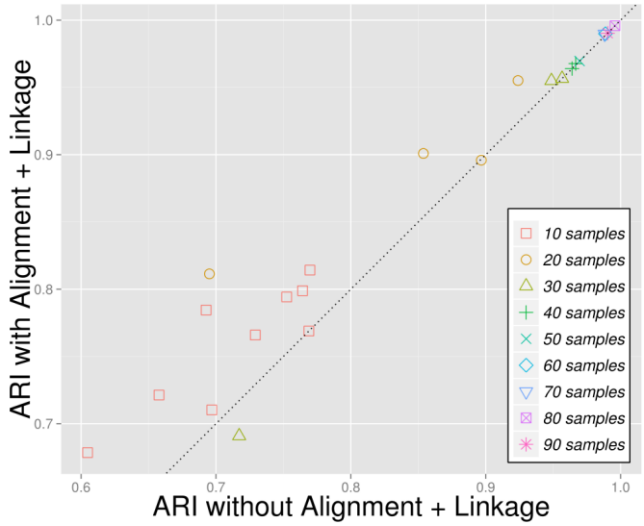
Speedup Ratio



Incorporating Additional Information



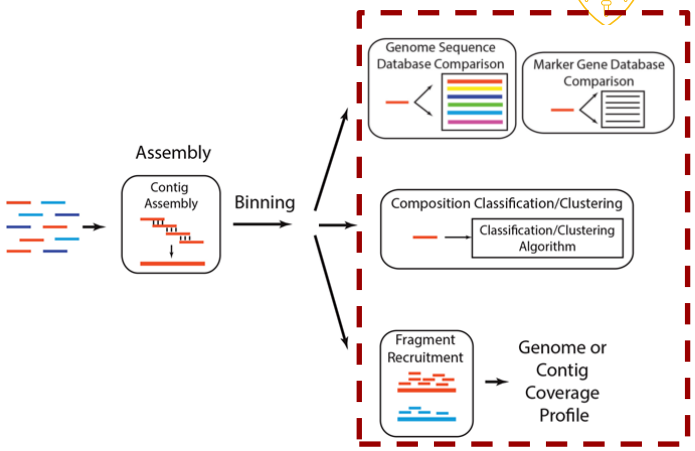




Summary so far

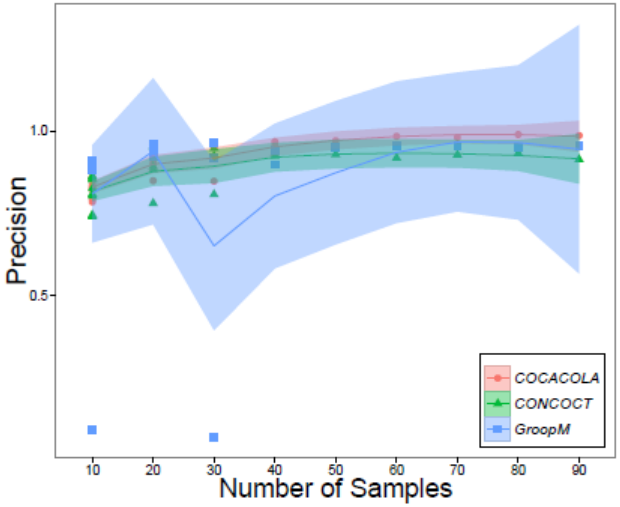
A metagenomics contigs binning framework

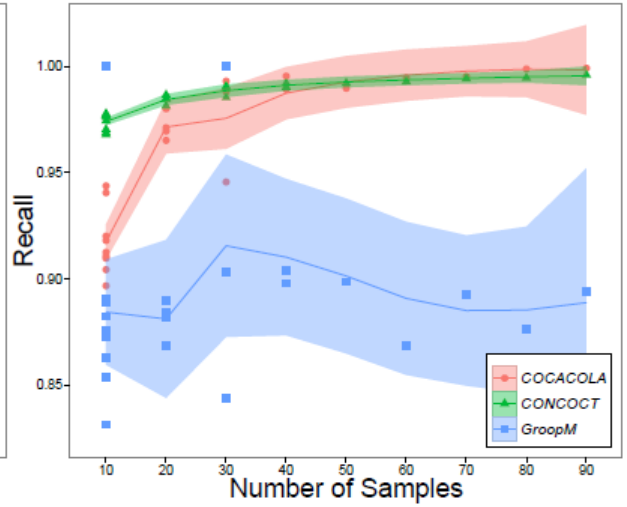
- Utilize all information available
- Embrace customized information
- Highly parallel and scalable
- However…

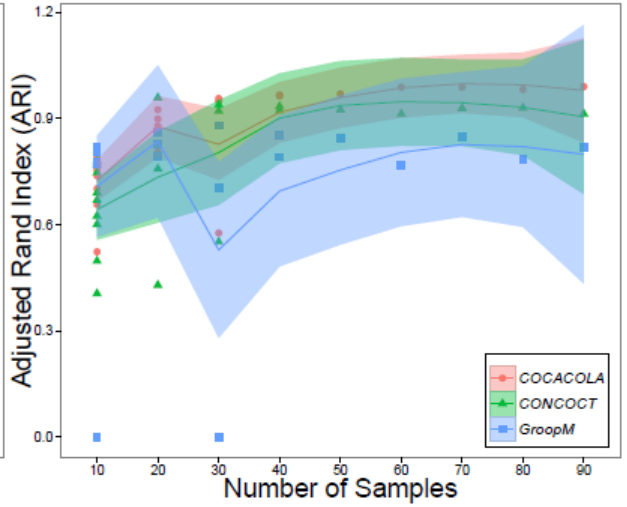


Observation

- Performance degrades when sample size shrinks







Part II

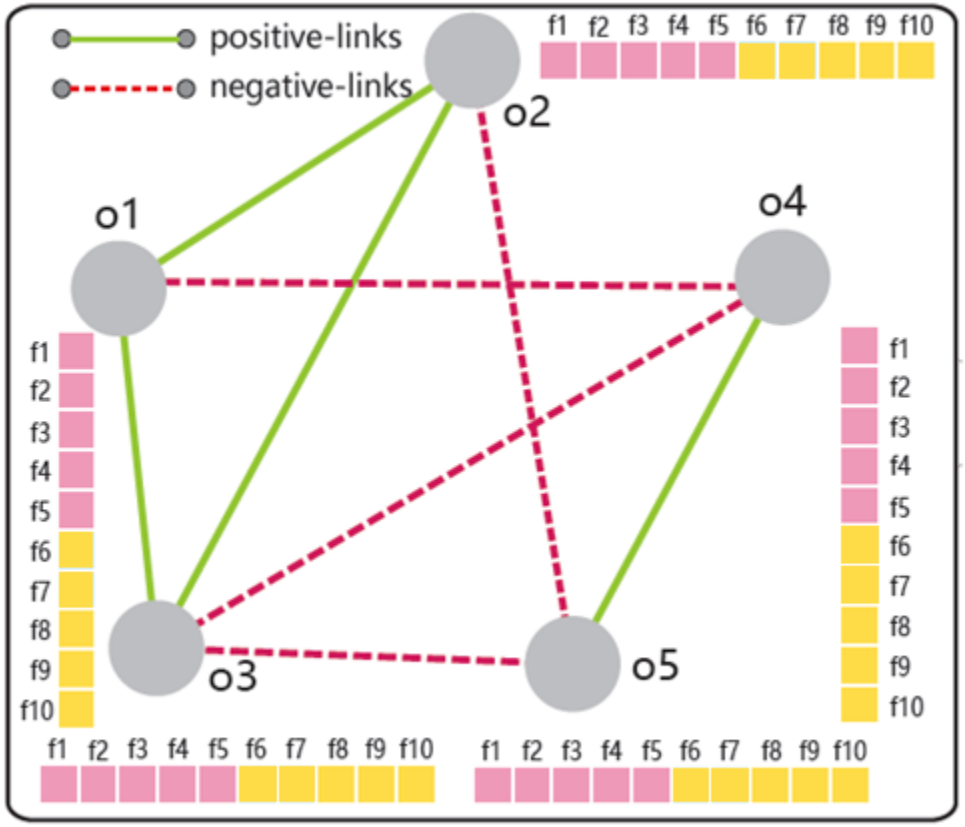
Hetero-RP: Towards Enhanced and Interpretable Clustering/Classification in Integrative Genomics

Availability: https://github.com/younglululu/Hetero-RP

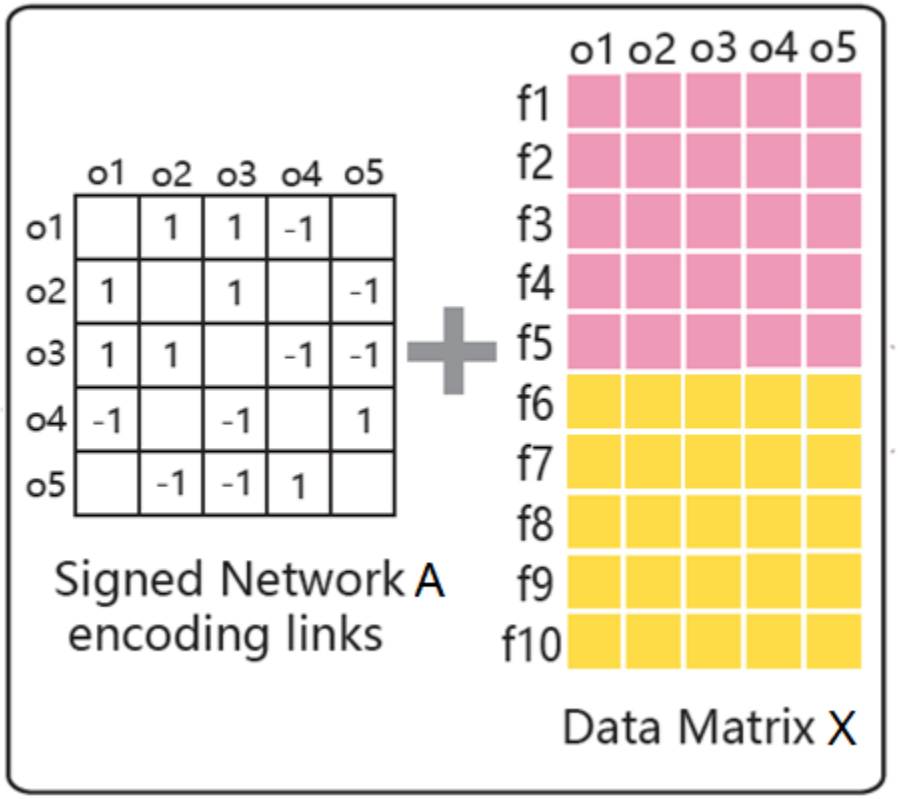
Publication: Lu et al. (2017) Nucleic Acid Research

Recall the metagenomic contig binning problem

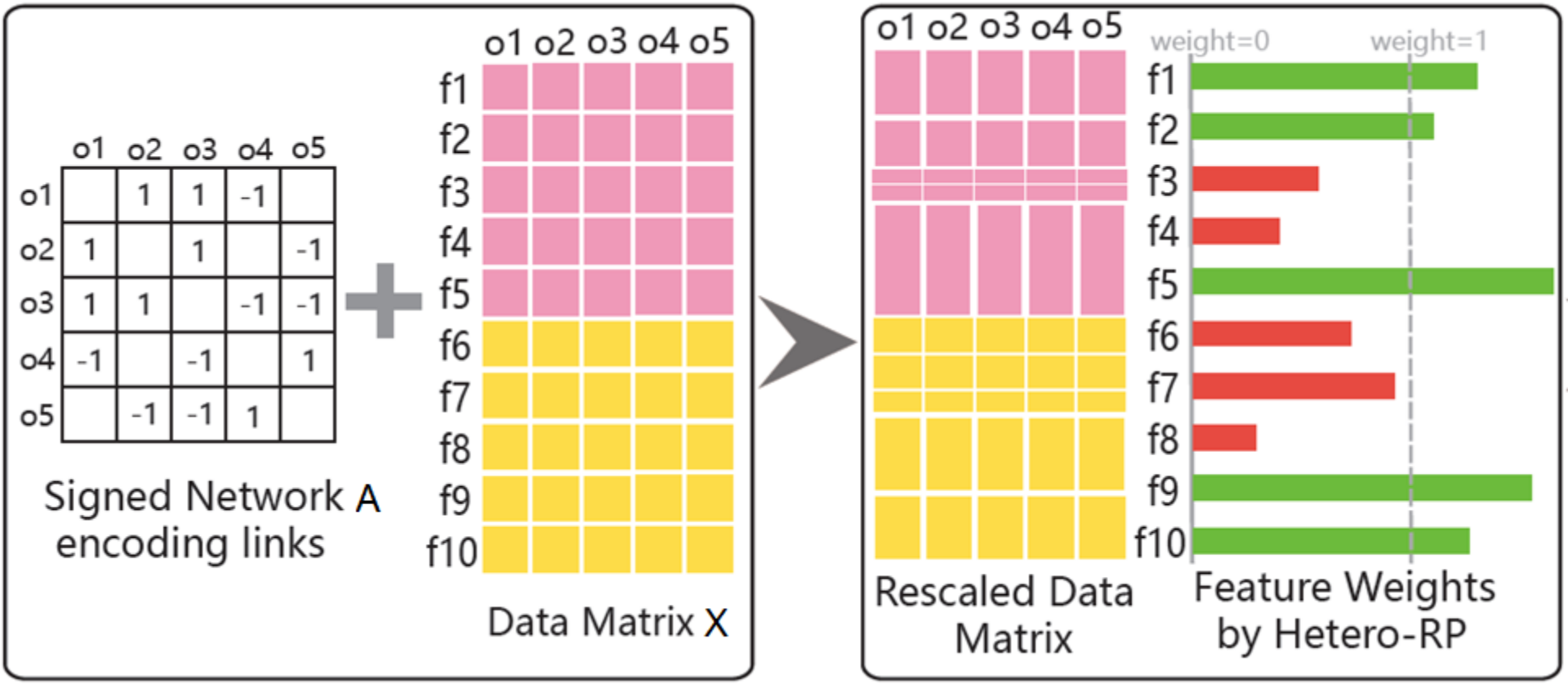
Illustation of Input



Re-form the Input



Goal


o1 o2 o3 o4 o5
o1 1 1 -1
o2 1 1 -1
o3 1 1 -1 -1
o4 -1 -1 1
o5 -1 -1 1
Signed Network A encoding links
o1 o2 o3 o4 o5
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
Data Matrix X
o1 o2 o3 o4 o5
Rescaled Data Matrix
weight=0
weight=1
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
Feature Weights by Hetero-RP

Problem Formulation

- Assume there are p features
- Find a p-dim weight vector W, to minimize the inconsistency between the signed graph and the feature-wise rescaled data matrix $\mathrm{diag}(W)X$

$$\min_W L(W) = \sum_{i,j} A_{ij} \left\| \mathrm{diag}(W)X_{\cdot i} - \mathrm{diag}(W)X_{\cdot j} \right\|^2$$

$$= \mathrm{tr}(\mathrm{diag}(W)XLX^T\mathrm{diag}(W)),$$

$$\text{s.t.} \quad W \geq 0, \quad \text{and} \quad \sum_i W_i = p$$

Meaningless if negative

To avoid trivial solutions

Hetero-RP assumes:

- The majority of features are useful.
- Among useful features:
  - Some are more or less informative: $w \neq 1$
  - The rest are neutral: $w = 1$

Compare to conventional feature selection

- Features are either informative ($w = 1$) or non-informative ($w = 0$)

To guarantee validity

- Check the multi-modality of each feature beforehand
- dip test

Hartigan and Hartigan (1985)

New formulation in quadratic programming

$$\min_W L(\Delta W) = \mathrm{tr}(\mathrm{diag}(1+\Delta W) X L X^T \mathrm{diag}(1+\Delta W)) + \lambda \|\Delta W\|^2$$

$$= \sum_i Y_i (\Delta W_i + 1)^2 + \lambda \Delta W_i^2, \qquad \Leftarrow \quad Y_i = (XLX)_{ii}$$

$$\text{s.t.} \qquad \Delta W_i \geq -1, \quad \text{and} \quad \sum_i \Delta W_i = 0$$

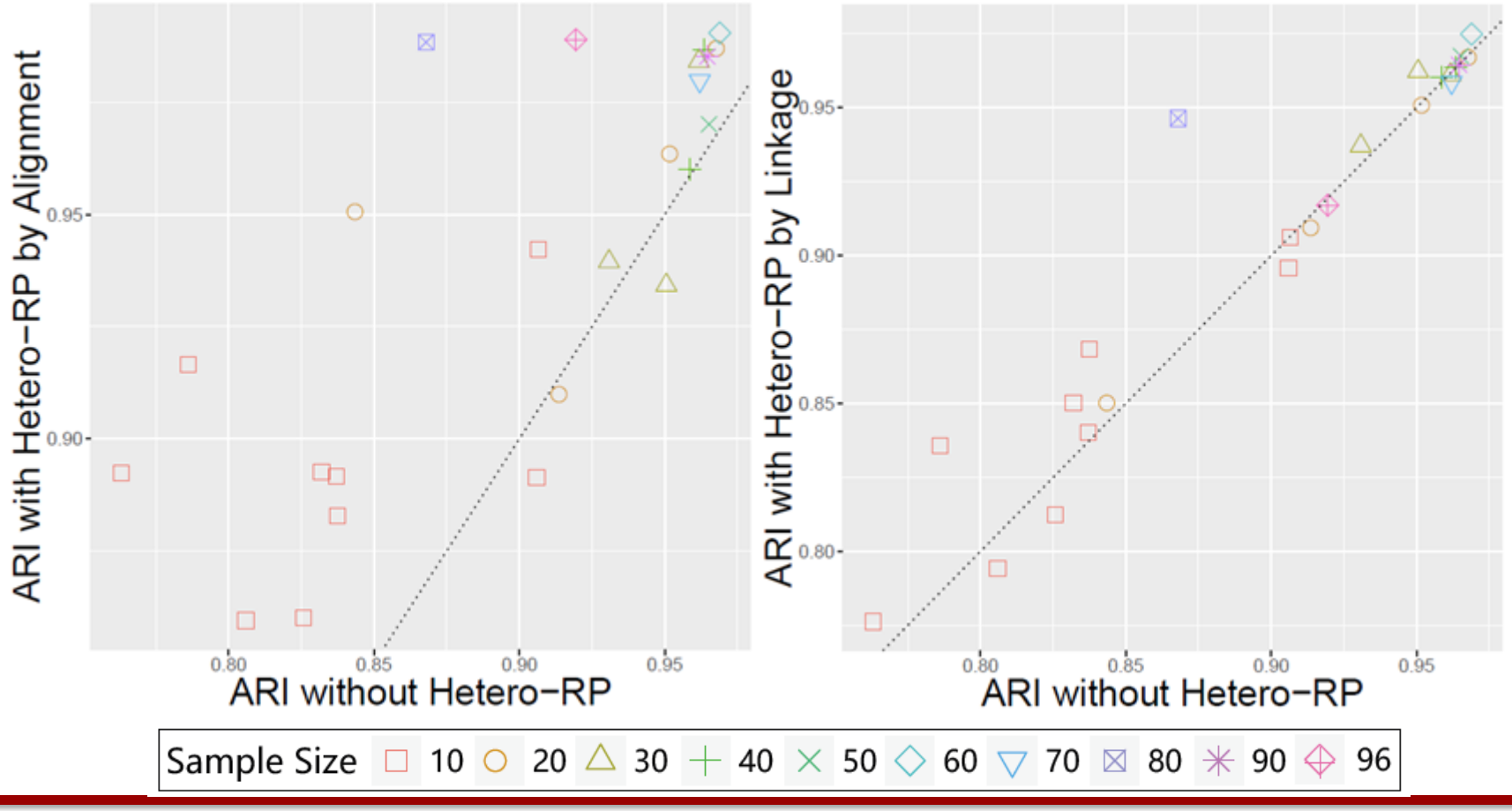
Parameter selection

$$\Delta\widehat{W} \leftarrow \arg \min_{\substack{\Delta W \geq -1 \\ \sum_i \Delta W_i = 0}} \sum_i Y_i(\Delta W_i + 1)^2 + 2p\lambda_0\widehat{\sigma}\,\|\Delta W\|^2\,,$$

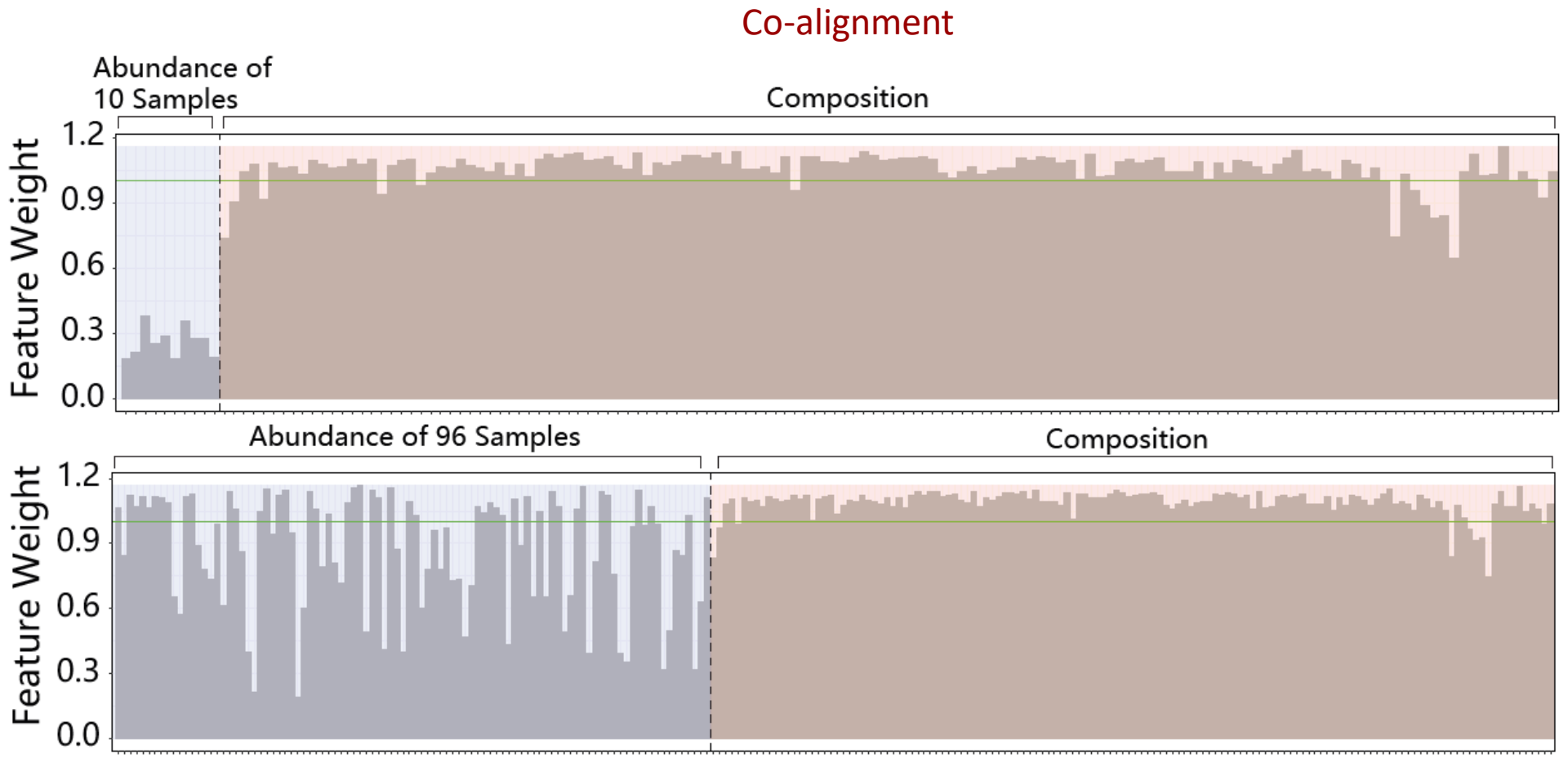
$$\widehat{\sigma} \leftarrow \sqrt{\frac{1}{p}\sum_i Y_i(\Delta W_i + 1)^2},$$

Sun and Zhang (2012); Fan and Lv (2016)

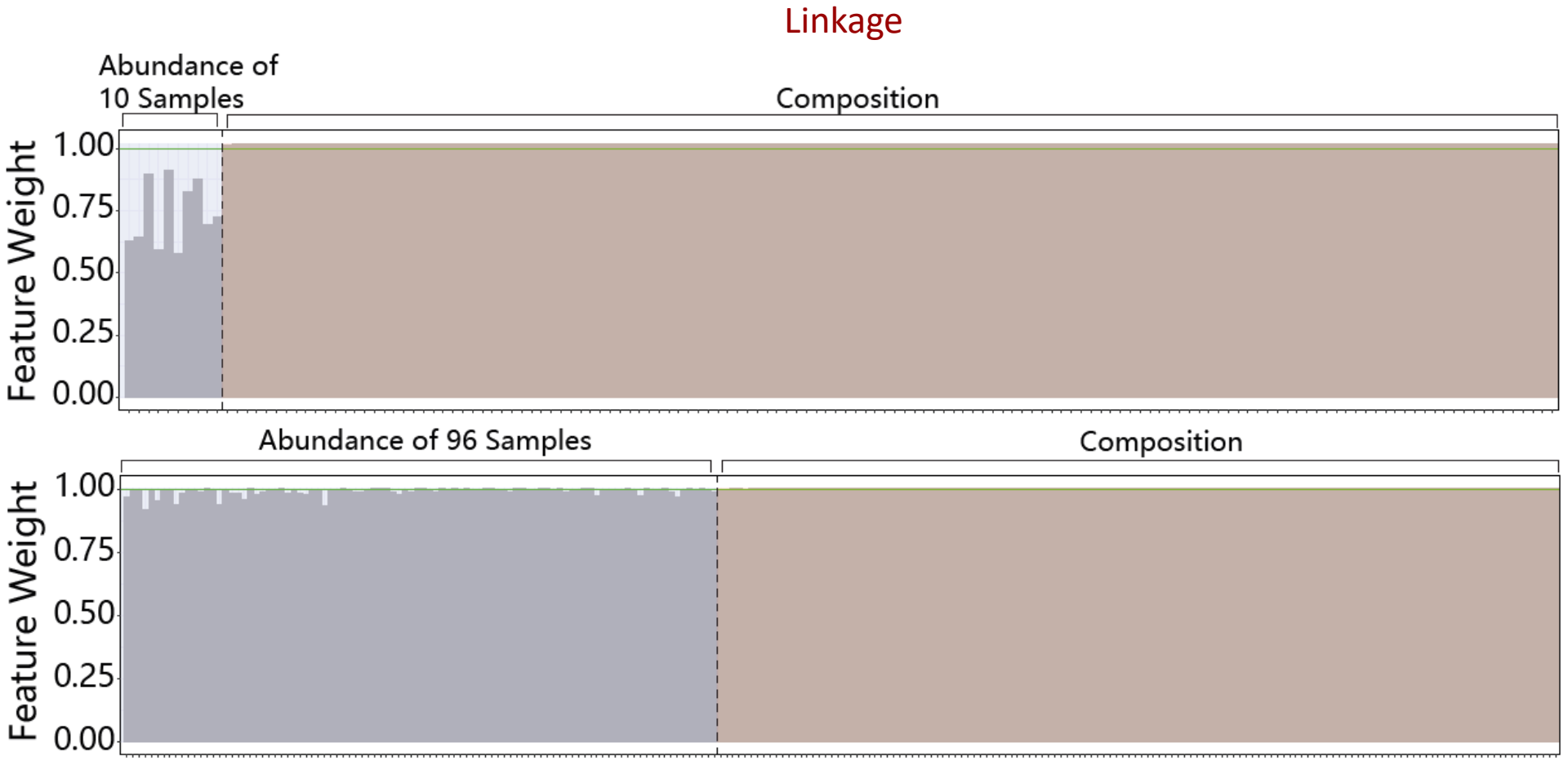
Experiments on Synthetic “Species” Dataset



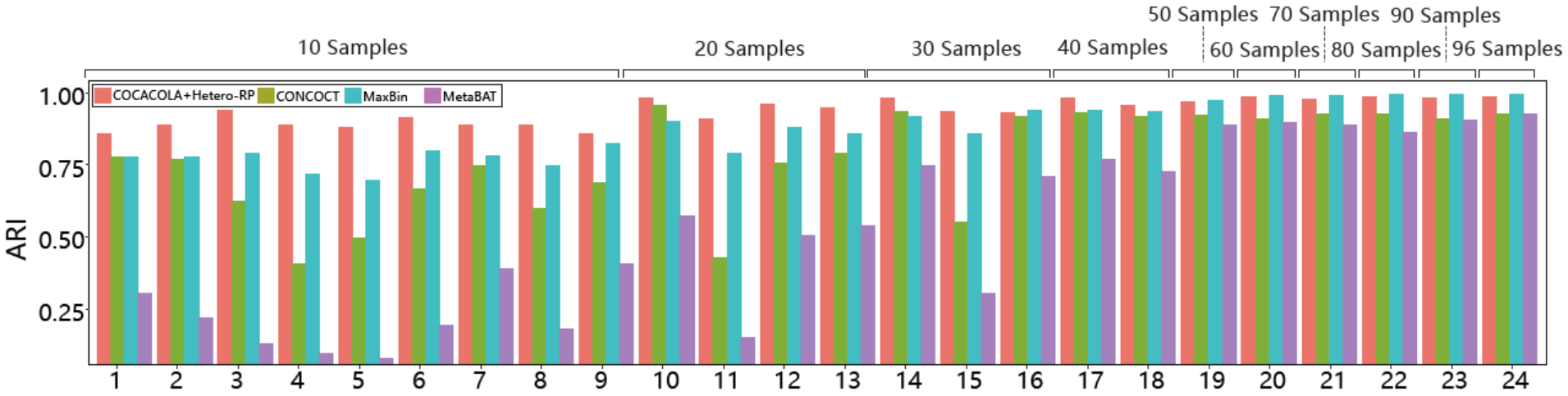
What feature weights look like?



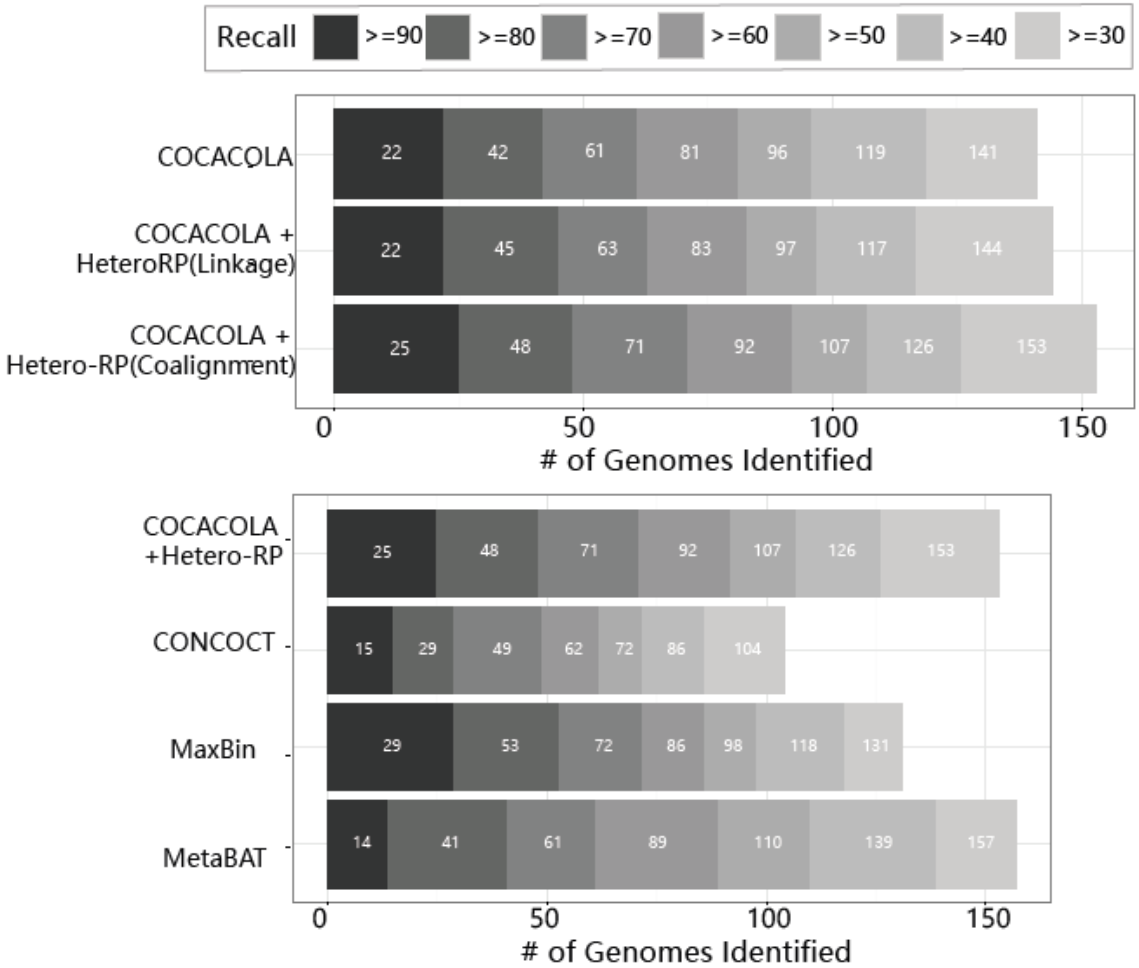
What feature weights look like?



Compare to state-of-the-arts



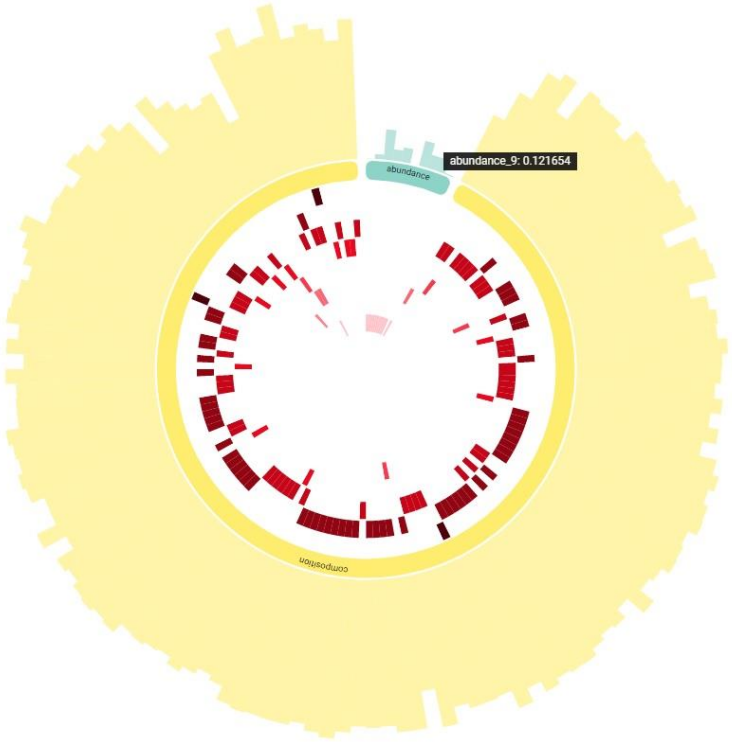
Experiments on Real "MetaHIT" Dataset



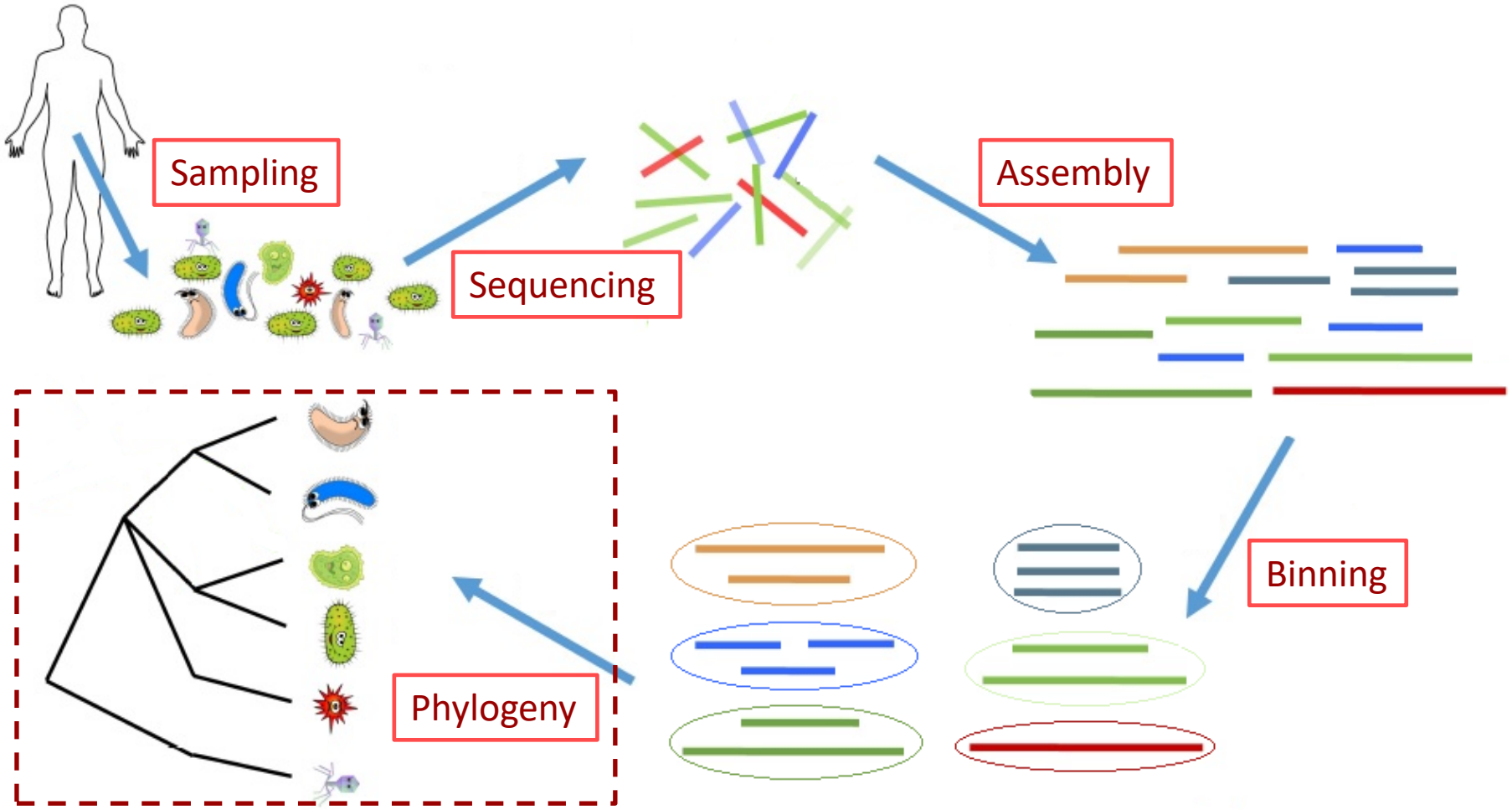
Summary so far

A integration framework for different types of data

- Weigh important features more highly than less important ones
- Scalable and tuning-free
- Visualization tool is provided to view the features
- Not limited to Metagenomic Binning
- Not limited to Clustering



Recall



Construct phylogeny by sequence comparison

Alignment-based approaches

Drawbacks:

- Slow
- Not designed for shotgun reads

- BLAST
- BLAT
- …

Alignment-free approaches

- Compare sequences using k-mer counts/frequencies
- State-of-the-art: CVTree, d_2^*, d_2^S
  - Basic idea: use the centralized k-mer counts
  - Removing the background noise enhances the true signal
- Drawback:
  - Still slow

- Manhattan
- Euclidean
- Chebyshev
- Cosine
- Pearson
- Jensen-Shannon
- FFP
- Co-phylog
- CVTree
- d_2^*
- d_2^S
- …

Song et al. (2014)

Part III

CAFE: aCcelerated Alignment-FrEe sequence analysis

Availability: https://github.com/younglululu/CAFE

Publication: Lu et al. (2017) Nucleic Acid Research

Highlights

- Integrate 28 alignment-free dissimilarity measures
  - 3 measures based on background adjusted k-mer counts
  - 10 conventional measures based on k-mer counts
  - 15 measures based on presence/absence of k-mers

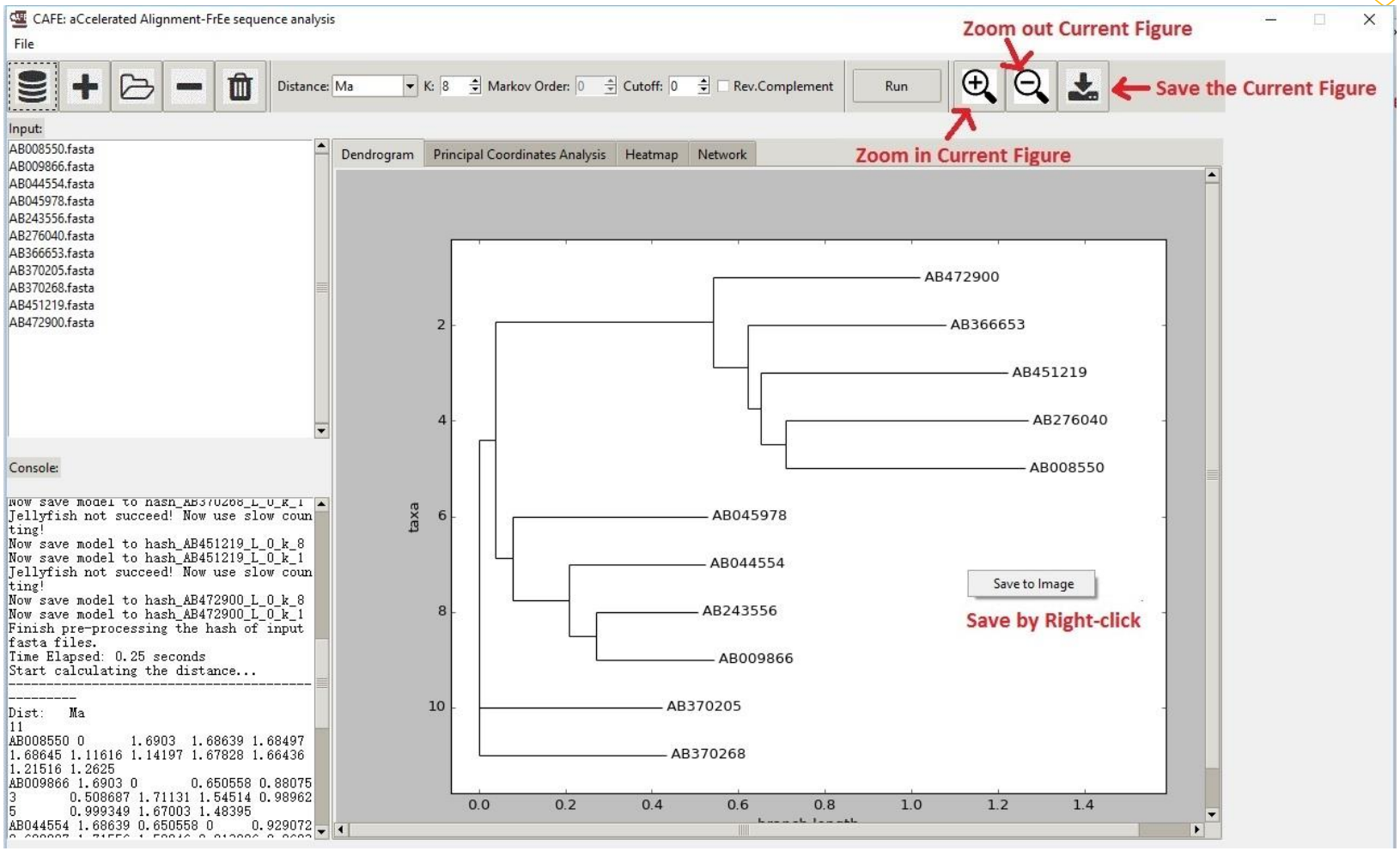
- Novel data structure and powerful engineering skills to speed up CVTree, d_2^*, d_2^S

- Support both assembled genome sequences and unassembled shotgun reads

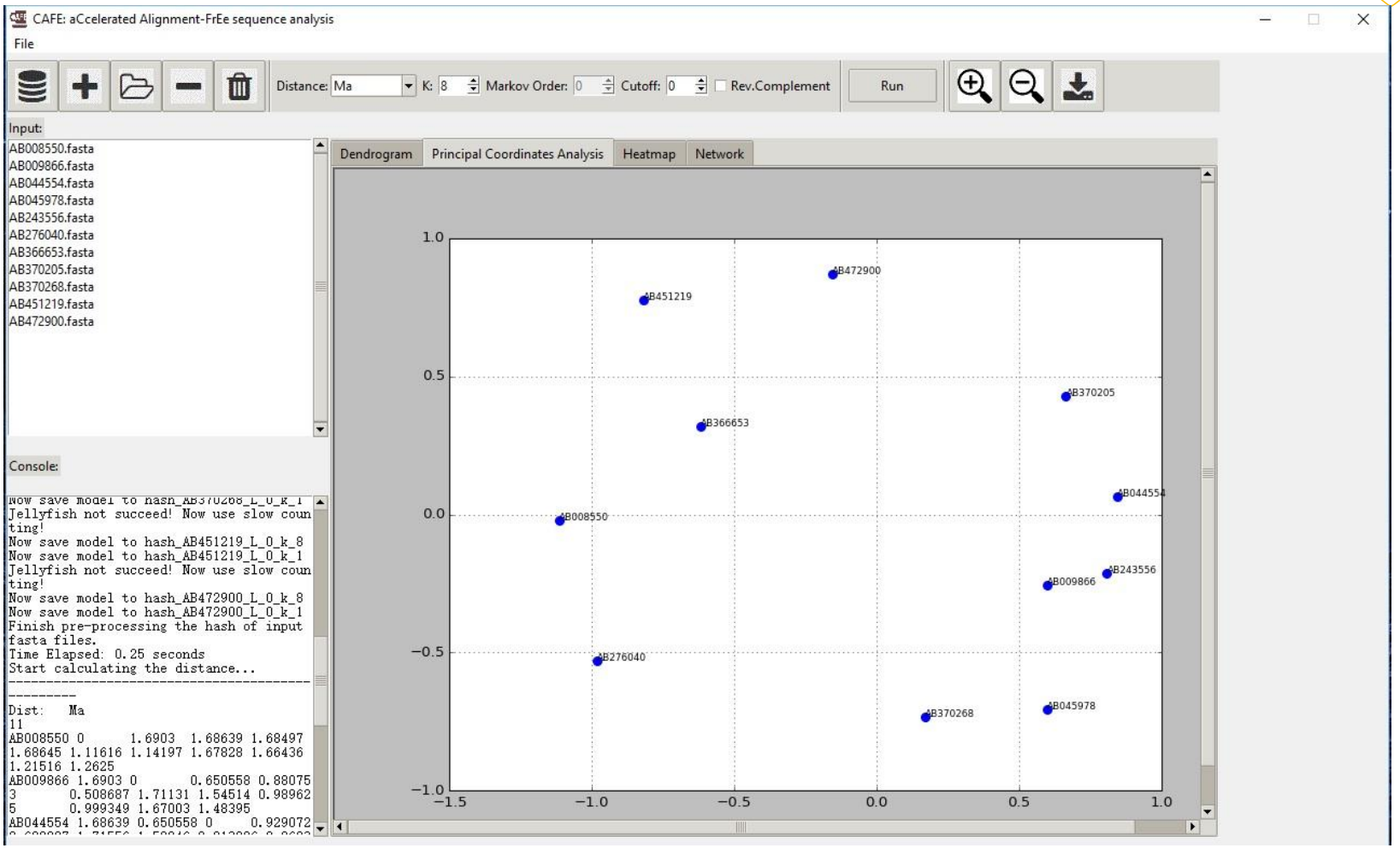
- Provide interactive visualized tool for downstream analyses
  - Dendrograms
  - Heatmap
  - Principal coordinate analysis (PCoA)
  - Network display

- Design for extensibility and reusability
  - Allow adding customized dissimilarity measures as plug-ins

Screenshot


CAFE: aCcelerated Alignment-FrEe sequence analysis
File
Distance: Ma
K: 8
Markov Order: 0
Cutoff: 0
Rev.Complement
Run
Zoom out Current Figure
Zoom in Current Figure
Save the Current Figure
Input:
AB008550.fasta
AB009866.fasta
AB044554.fasta
AB045978.fasta
AB243556.fasta
AB276040.fasta
AB366653.fasta
AB370205.fasta
AB370268.fasta
AB451219.fasta
AB472900.fasta
Console:
Jellyfish not succeed! Now use slow counting!
Now save model to hash_AB451219_L_0_k_8
Now save model to hash_AB451219_L_0_k_1
Jellyfish not succeed! Now use slow counting!
Now save model to hash_AB472900_L_0_k_8
Now save model to hash_AB472900_L_0_k_1
Finish pre-processing the hash of input fasta files.
Time Elapsed: 0.25 seconds
Start calculating the distance...
Dist: Ma
11
AB008550 0 1.6903 1.68639 1.68497 1.68645 1.11616 1.14197 1.67828 1.66436 1.21516 1.2625
AB009866 1.6903 0 0.650558 0.88075 3 0.508687 1.71131 1.54514 0.98962 5 0.999349 1.67003 1.48395
AB044554 1.68639 0.650558 0 0.929072
Dendrogram
Principal Coordinates Analysis
Heatmap
Network
AB472900
AB366653
AB451219
AB276040
AB008550
AB045978
AB044554
AB243556
AB009866
AB370205
AB370268
taxa
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Save to Image
Save by Right-click

Screenshot


CAFE: aCcelerated Alignment-FrEe sequence analysis
File
Distance: Ma
K: 8
Markov Order: 0
Cutoff: 0
Rev.Complement
Run
Input:
AB008550.fasta
AB009866.fasta
AB044554.fasta
AB045978.fasta
AB243556.fasta
AB276040.fasta
AB366653.fasta
AB370205.fasta
AB370268.fasta
AB451219.fasta
AB472900.fasta
Console:
Jellyfish not succeed! Now use slow counting!
Now save model to hash_AB451219_L_0_k_8
Now save model to hash_AB451219_L_0_k_1
Jellyfish not succeed! Now use slow counting!
Now save model to hash_AB472900_L_0_k_8
Now save model to hash_AB472900_L_0_k_1
Finish pre-processing the hash of input fasta files.
Time Elapsed: 0.25 seconds
Start calculating the distance...
Dist: Ma
11
AB008550 0 1.6903 1.68639 1.68497 1.68645 1.11616 1.14197 1.67828 1.66436 1.21516 1.2625
AB009866 1.6903 0 0.650558 0.88075
3 0.508687 1.71131 1.54514 0.98962
5 0.999349 1.67003 1.48395
AB044554 1.68639 0.650558 0 0.929072
Dendrogram
Principal Coordinates Analysis
Heatmap
Network
AB472900
AB451219
AB370205
AB366653
AB044554
AB008550
AB243556
AB009866
AB276040
AB370268
AB045978
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0

Experiments

Primate and Vertebrate Genomic Sequences

- 28 vertebrate and 21 primate species
- K=14, Markov order=12
- Investigate the relationship between the pairwise dissimilarity measure with evolutionary distance

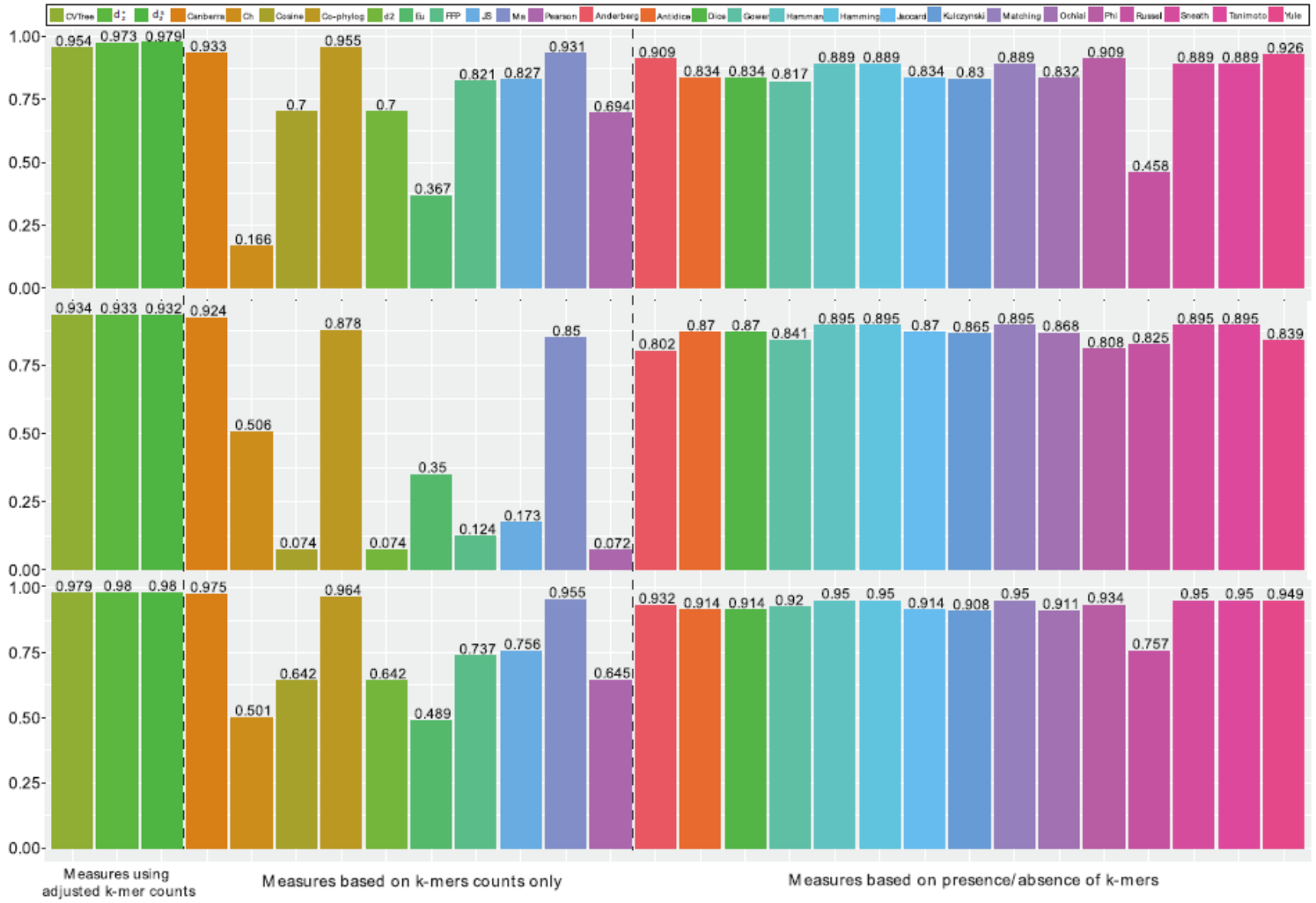
Microbial Genomic Sequences

- 27 E.coli and Shigella genomes, with 6 E.coli reference (ECOR) groups
- Investigate whether the UPGMA-constructed tree can identify the groups

Metagenomic Samples

- 28 samples of mammalian gut, short reads
- 3 groups: 8 hindgut-fermenting herbivores, 13 foregut-fermenting herbivores, and 7 simple-gut carnivores
- Investigate whether the UPGMA-constructed tree can identify the groups

Primate and Vertebrate Genomic Sequences

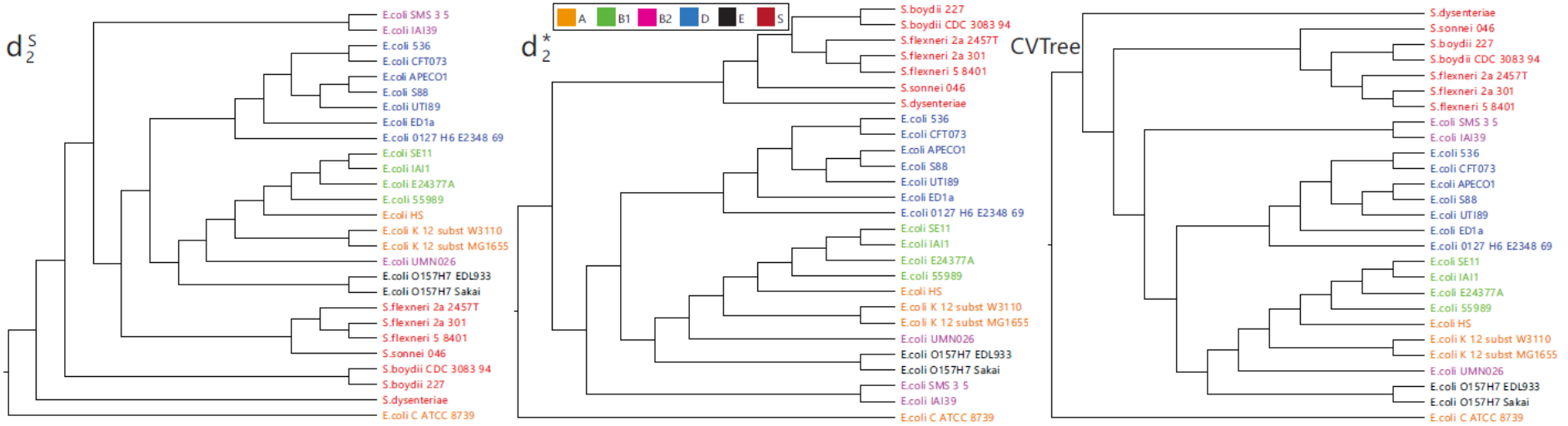


USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

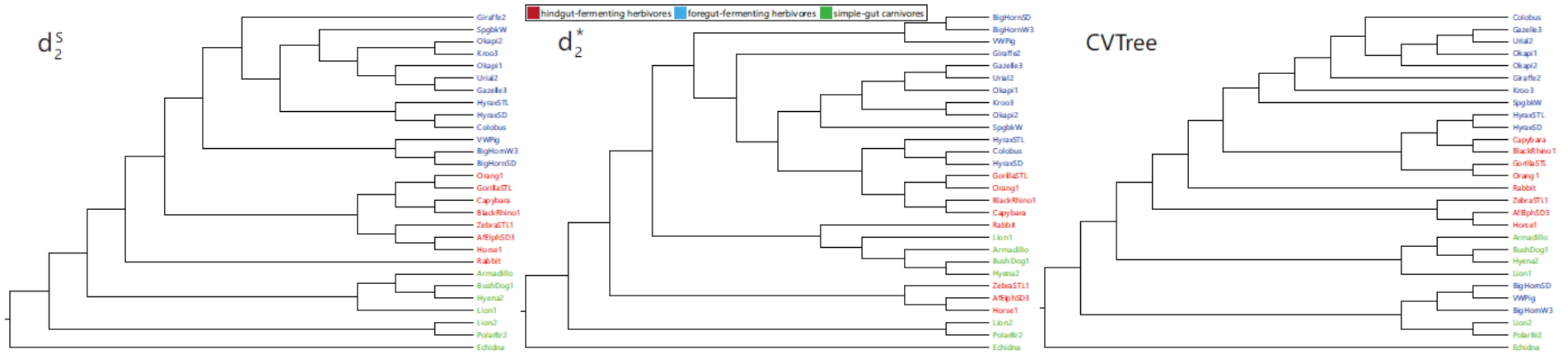
Wall time, peak memory usage, and speedup ratio

Sequence Model	Original Implementation		CAFE		
	Wall time	Peak memory	Wall time	Speedup	Peak memory
order=0	0:42′32″	64.0G	0:6′09″	6.9x	31.1G
order=1	1:44′18″	64.0G	0:6′13″	16.8x	31.1G
order=2	2:11′32″	64.0G	0:6′12″	21.2x	31.1G
order=3	2:34′28″	62.4G	0:5′05″	30.4x	24.8G
order=4	2:34′11″	62.3G	0:6′10″	25.0x	31.1G
order=5	3:24′43″	64.0G	0:5′08″	39.9x	24.8G
order=6	2:53′08″	63.9G	0:5′14″	33.1x	24.8G
order=7	2:40′04″	64.0G	0:6′29″	24.7x	31.1G
order=8	2:33′19″	64.0G	0:6′08″	25.0x	48.1G
order=9	2:37′50″	64.2G	0:6′19″	25.0x	48.2G
order=10	2:22′18″	64.7G	0:5′15″	27.1x	48.5G
order=11	2:05′55″	60.4G	0:6′29″	19.4x	49.6G
order=12	1:53′40″	74.6G	0:6′39″	17.1x	37.0G

Microbial Genomic Sequences



Metagenomic Samples



Summary so far

An interactive and visualized platform for alignment-free analysis

- 28 alignment-free dissimilarity measures
- ~24x speedup ratio

Limitation of alignment-free approaches

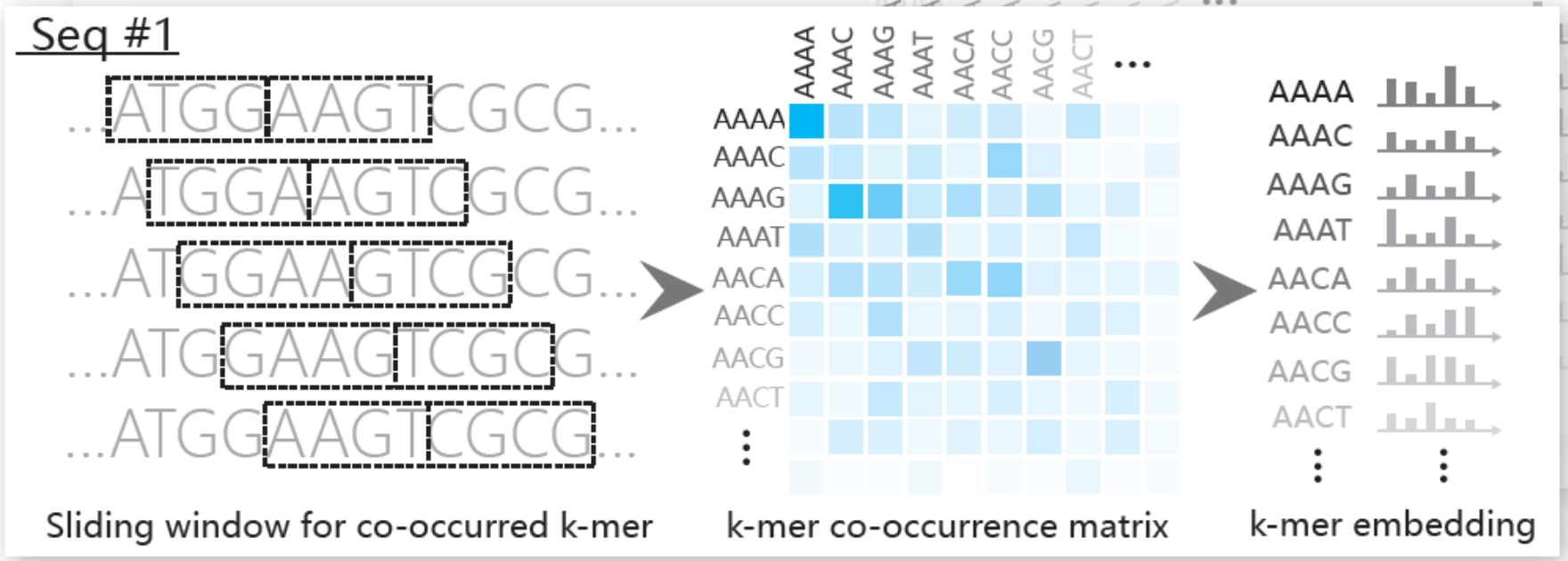
- The size of k-mer frequency vector is $\Theta(4^k)$ for each sequence
  - K=14, ~1GB
- Impractical for storage, sharing, and transmit
- NCBI RefSeq Database
  - 92651 sequences
  - 840.6 GB

Part IV

CRAFT: Compact genome Representation towards large-scale Alignment-Free daTabase

Manuscript in preparation

Illustrative Example



Equivalent to extract 8-mers

Store in $4^4 \times 4^4$ matrix instead of 4^8 vector

Reduce $4^4 \times 4^4$ matrix to $4^4 \times d$ matrix

Low-dimensional spaces for storage

- Let X be the co-occurrence matrix

- Goal: to find a latent, low-dimensional space of each k-mer,
  - Preserving the co-occurrences between pairwise k-mers
  - Reduce the bias of each k-mer originated from sequence background noise

- Given the low-dimension d, for each k-mer i, to find:
  - Biase $b_i \in \mathbb{R}$
  - d-dimensional vectors $w_i, \tilde{w}_i \in \mathbb{R}^d$

$w_i, \tilde{w}_i$ are latent space for k-mer i in preceding and succeeding of the pair

$$b_i + b_j + \langle w_i, \tilde{w}_j \rangle \approx \log X_{ij}$$

- Latent Semantic Analysis
- Factorization Machines
- Distributed Representations
- …

Pennington et al. (2014)

Recover the conditional probability matrix

$$P_{ij}(w, \tilde{w}, b) = \frac{\exp\left(b_i + b_j + \langle w_i, \tilde{w}_j \rangle\right)}{\sum_j \exp\left(b_i + b_j + \langle w_i, \tilde{w}_j \rangle\right)}$$

softmax

Dissimilarity measures

$$Eu = \sqrt{\sum_i \sum_j \left(P_{ij}^{(1)} - P_{ij}^{(2)}\right)^2}$$

$$Ma = \sum_i \sum_j \left|P_{ij}^{(1)} - P_{ij}^{(2)}\right|$$

$$JS = h\left(\frac{P^{(1)} + P^{(2)}}{2}\right) - \frac{1}{2}h\left(P^{(1)}\right) - \frac{1}{2}h\left(P^{(2)}\right)$$

$$h(P^{(i)}) = -\sum_i \sum_j P_{ij}^{(i)} \log P_{ij}^{(i)}$$

Experiments (The same as CAFE)

Primate and Vertebrate Genomic Sequences

- 28 vertebrate and 21 primate species
- K=14, Markov order=12
- Investigate the relationship between the pairwise dissimilarity measure with evolutionary distance

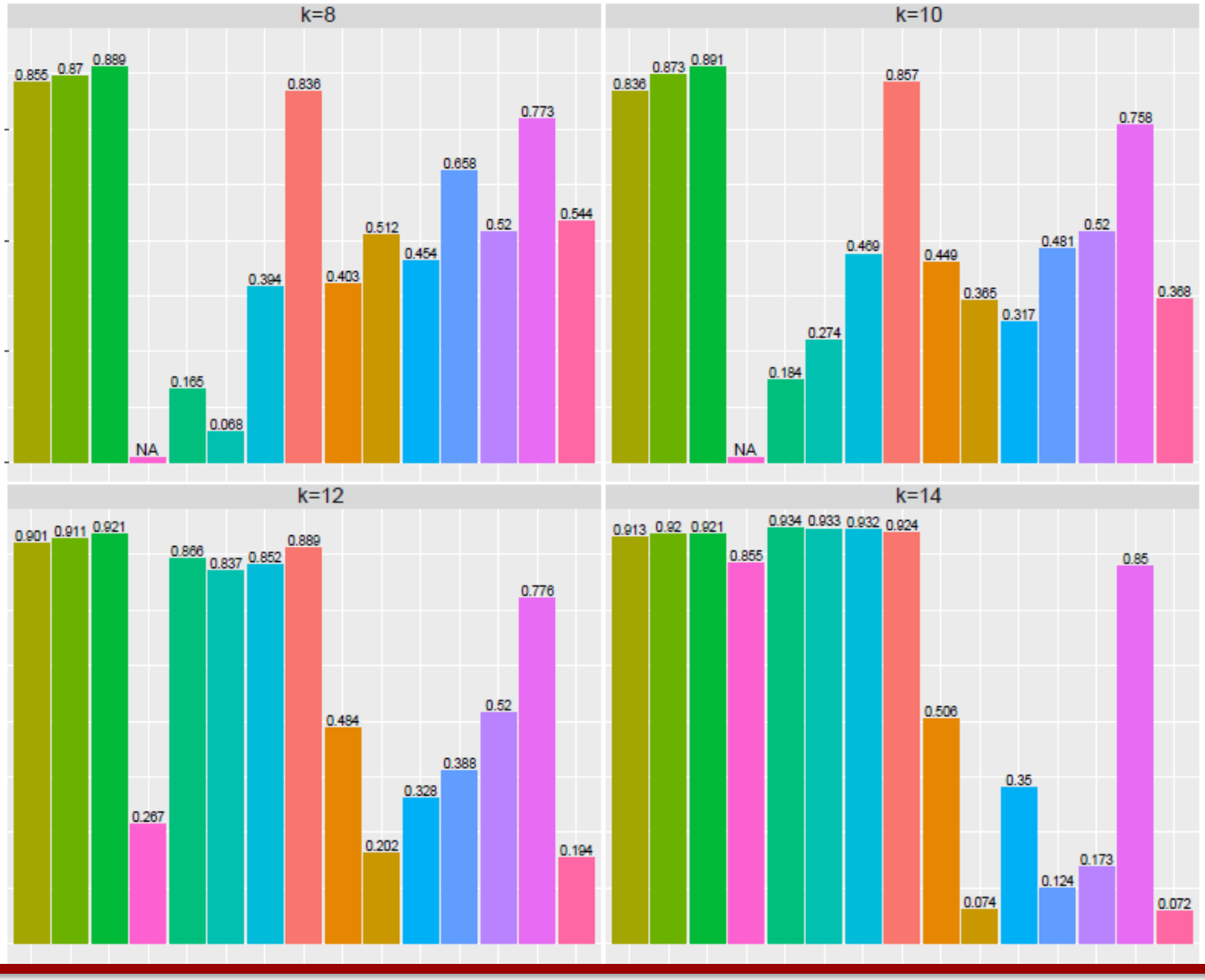
Microbial Genomic Sequences

- 27 E.coli and Shigella genomes, with 6 E.coli reference (ECOR) groups
- Investigate the UPGMA-constructed tree can identify the groups

Metagenomic Samples

- 28 samples of mammalian gut, short reads
- 3 groups: 8 hindgut-fermenting herbivores, 13 foregut-fermenting herbivores, and 7 simple-gut carnivores
- Investigate the UPGMA-constructed tree can identify the groups

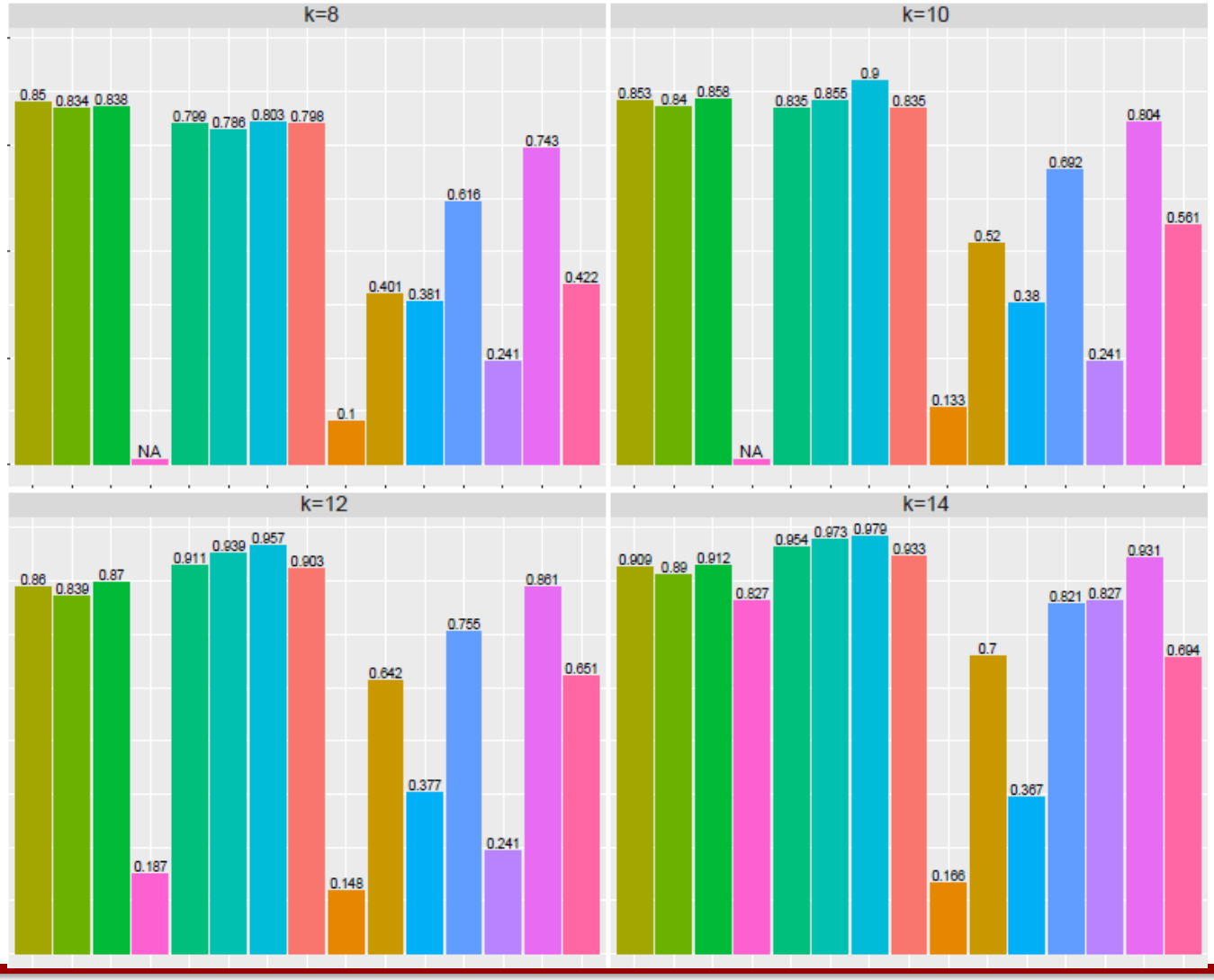
28 vertebrate species



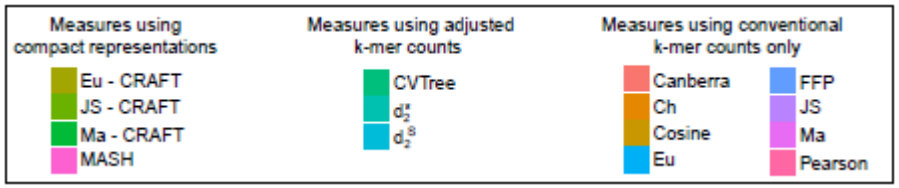
USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Measures using compact representations	Measures using adjusted k-mer counts	Measures using conventional k-mer counts only	
Eu - CRAFT	CVTree	Canberra	FFP
JS - CRAFT	d_2^*	Ch	JS
Ma - CRAFT	d_2^S	Cosine	Ma
MASH		Eu	Pearson

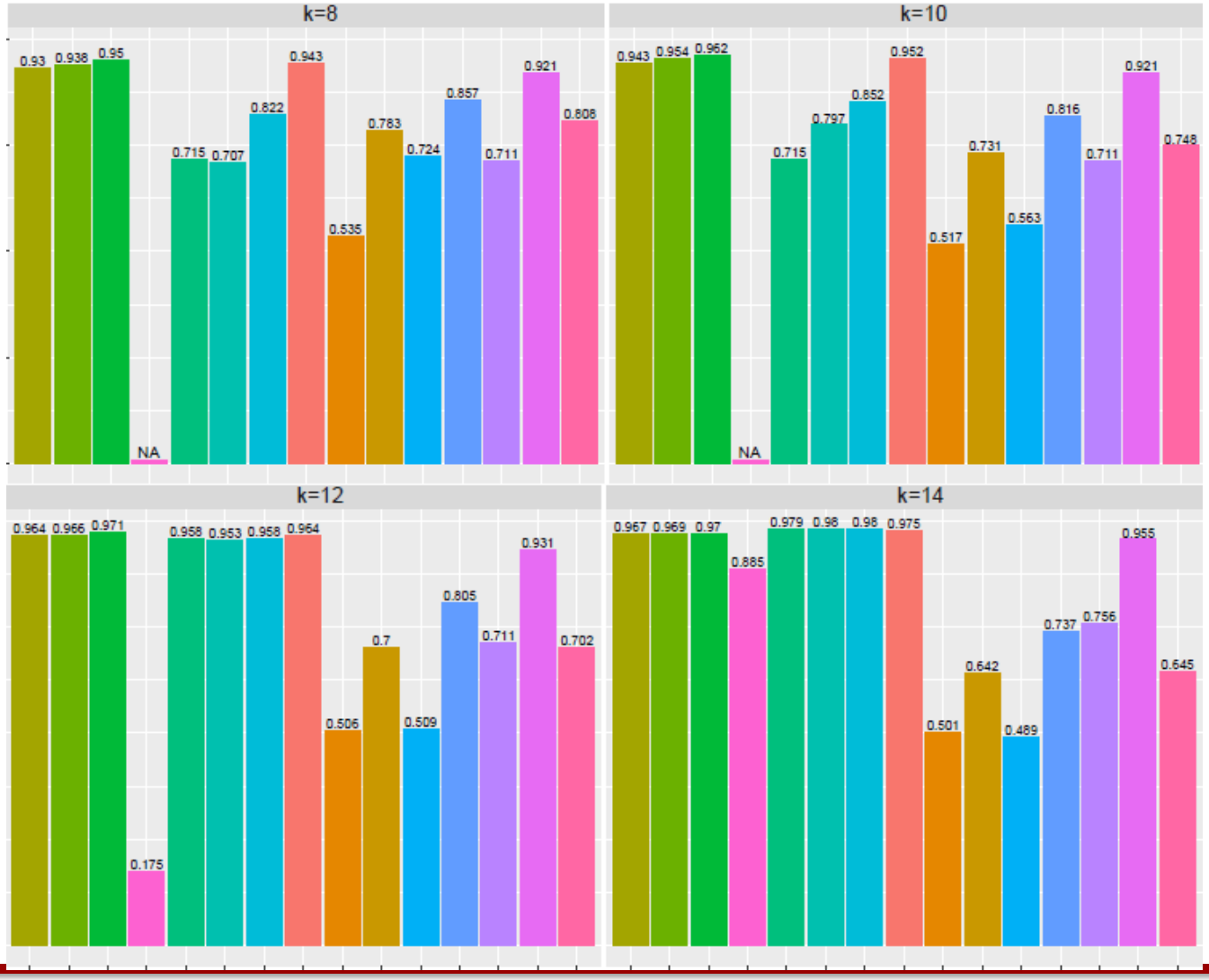
21 primate species



USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences



21 primate and 28 vertebrate species

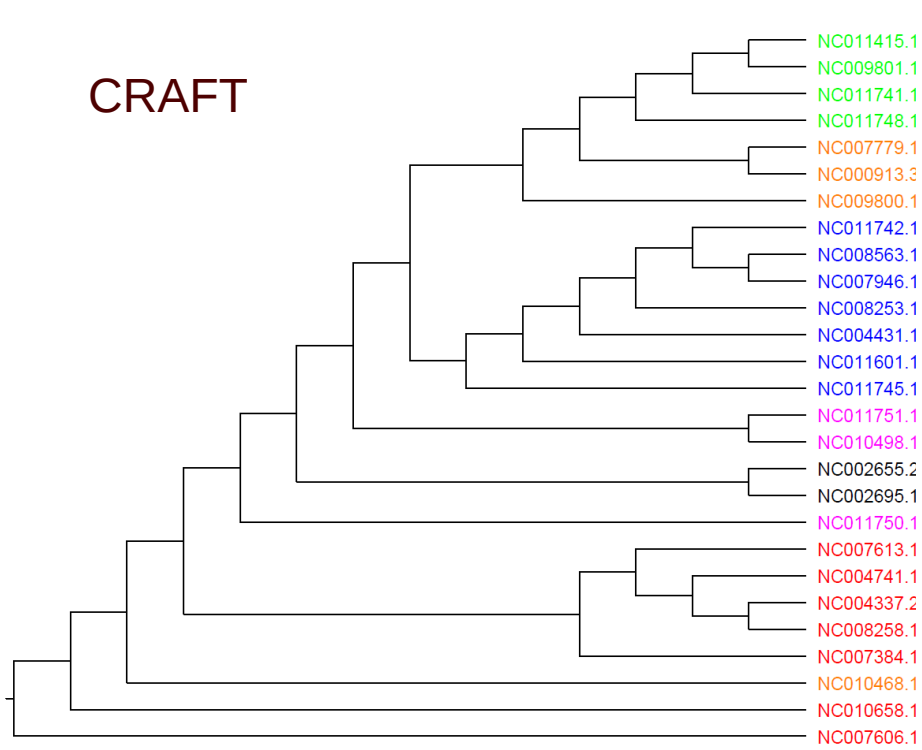


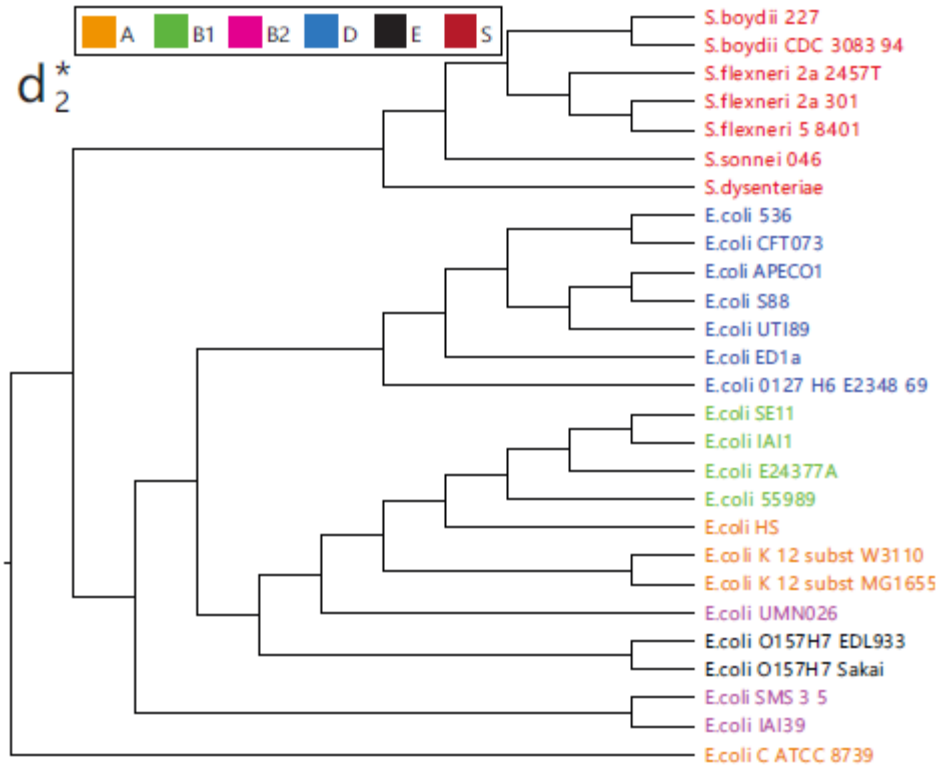
USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Measures using compact representations	Measures using adjusted k-mer counts	Measures using conventional k-mer counts only	
Eu - CRAFT	CVTree	Canberra	FFP
JS - CRAFT	d_2^*	Ch	JS
Ma - CRAFT	d_2^S	Cosine	Ma
MASH		Eu	Pearson

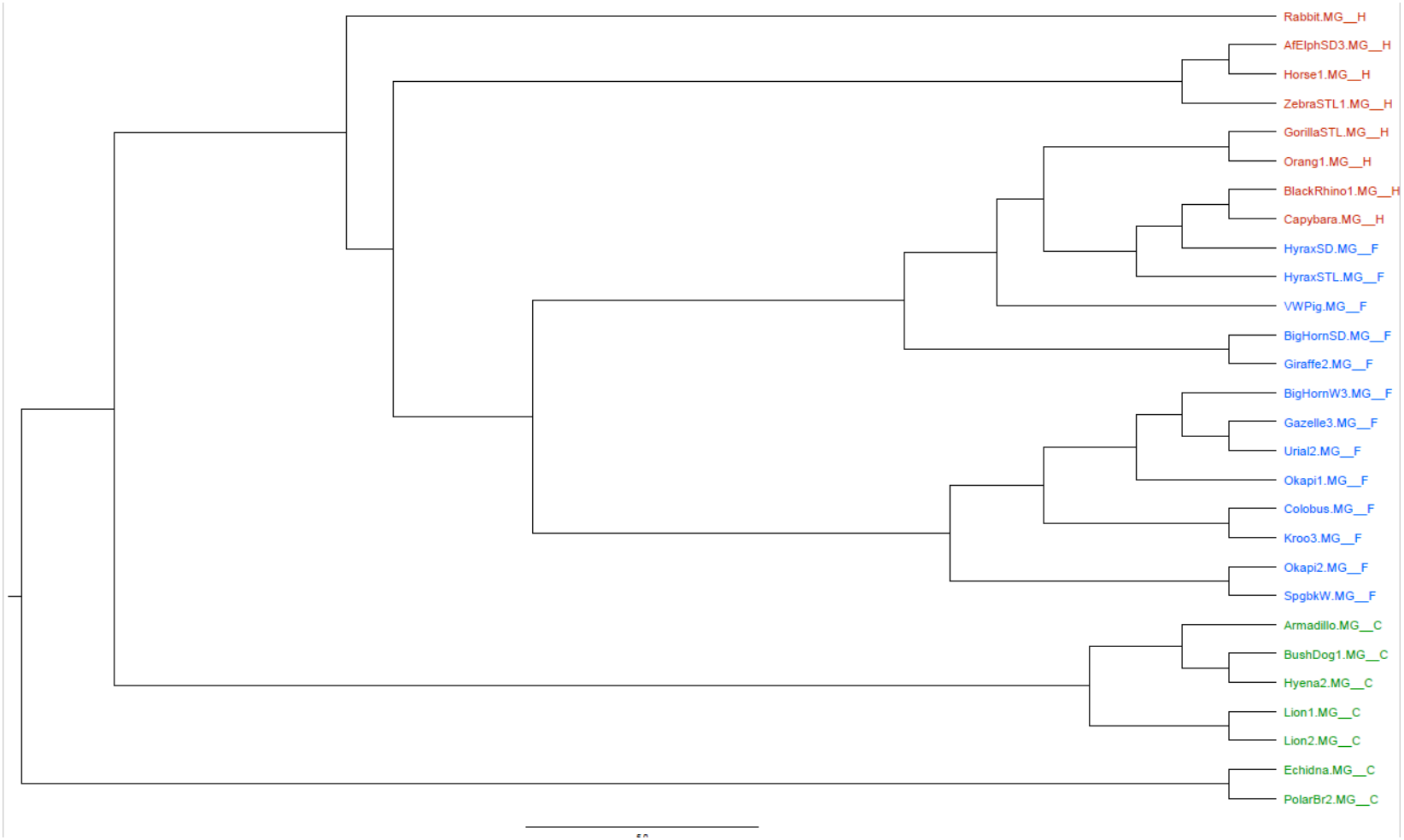
Microbial Genomic Sequences

CRAFT



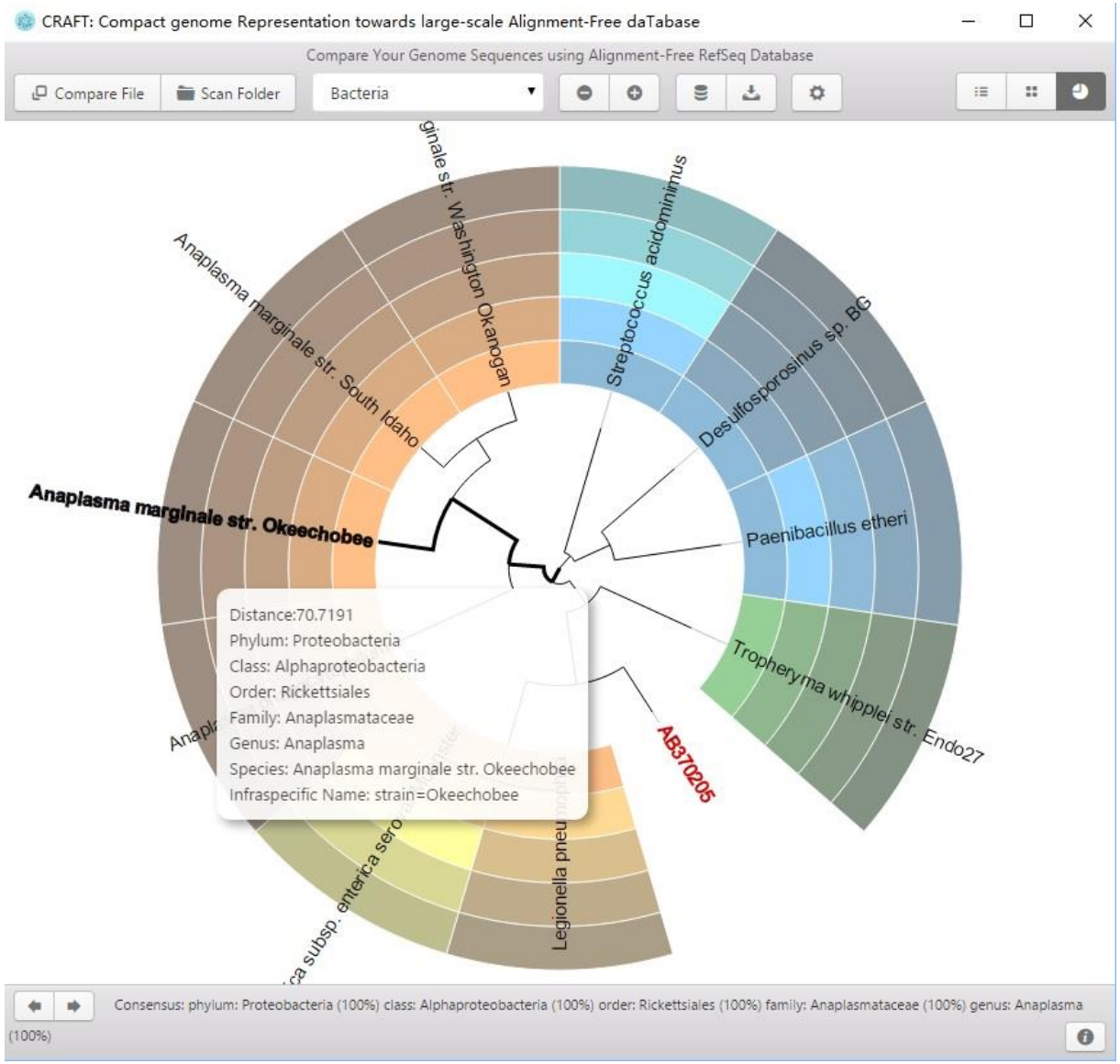


Metagenomic Samples



Compress NCBI RefSeq Database

- NCBI RefSeq Database
  - 92651 sequences
  - 840.6 GB
- Size after CRAFT compression
  - 1.86G
- Search time for the whole database
  - 2m32s
- Interactive visualized tool for database management

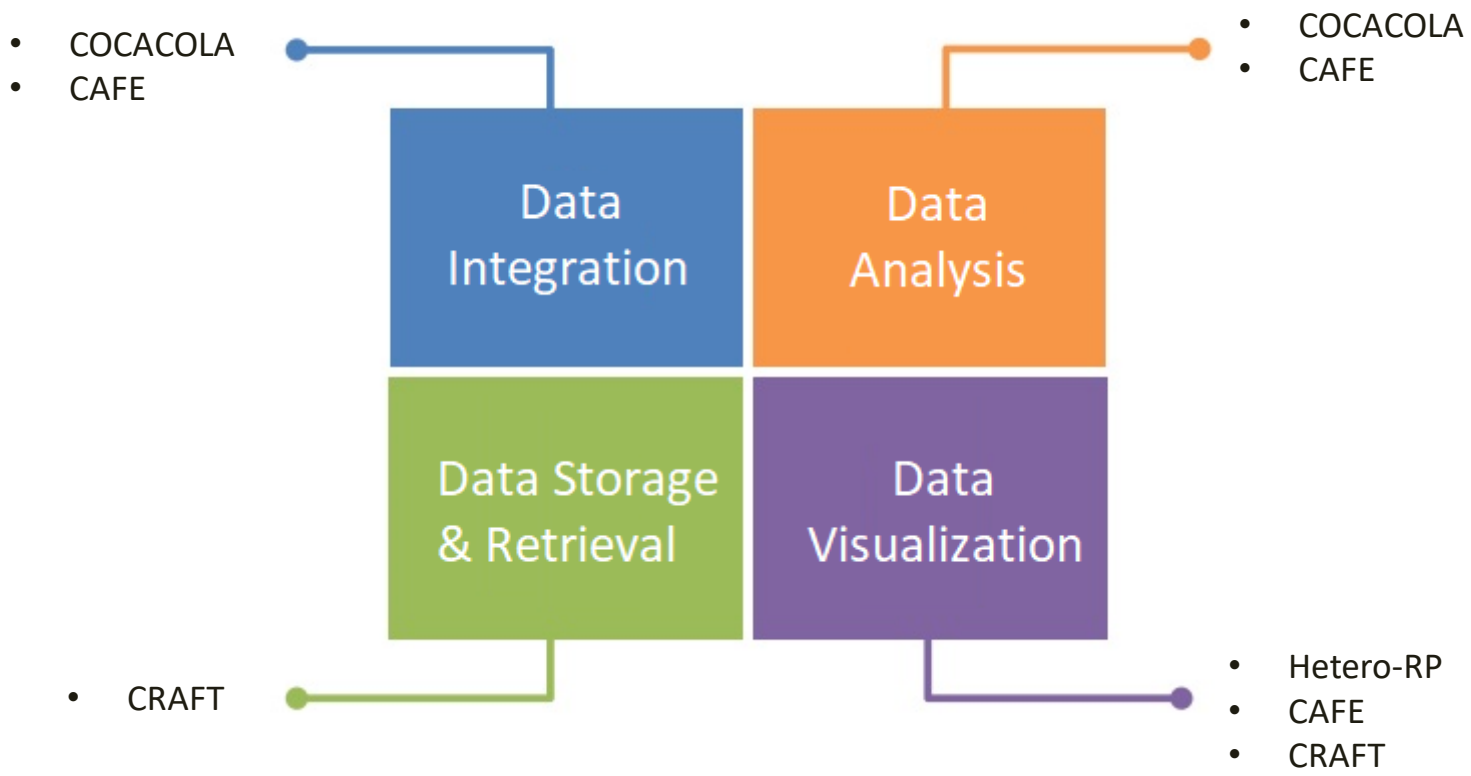

CRAFT: Compact genome Representation towards large-scale Alignment-Free daTabase
Compare Your Genome Sequences using Alignment-Free RefSeq Database
Compare File
Scan Folder
Bacteria
Anaplasma marginale str. Okeechobee
Anaplasma marginale str. South Idaho
Streptococcus acidominimus
Desulfosporosinus sp. BG
Paenibacillus etheri
Tropheryma whipplei str. Endo27
AB370205
Distance:70.7191
Phylum: Proteobacteria
Class: Alphaproteobacteria
Order: Rickettsiales
Family: Anaplasmataceae
Genus: Anaplasma
Species: Anaplasma marginale str. Okeechobee
Infraspecific Name: strain=Okeechobee
Consensus: phylum: Proteobacteria (100%) class: Alphaproteobacteria (100%) order: Rickettsiales (100%) family: Anaplasmataceae (100%) genus: Anaplasma (100%)

Summary so far

A compact representation for alignment-free database

- K=4 is good enough
- Support both genome sequences and unassembled shotgun reads
- Tailored for large-scale storage, sharing, and transmit
- User-friendly database management tool

Conclusion


COCACOLA
CAFE
COCACOLA
CAFE
Data Integration
Data Analysis
Data Storage & Retrieval
Data Visualization
CRAFT
Hetero-RP
CAFE
CRAFT

Future Directions

- Deep Learning based Heterogeneous Data Integration
- Constructing varying-length k-mer dictionary purely from genome data
- Theory behind empirical works

Acknowledgement

- Advisor in academy and mentor in life:
  - Prof. Fengzhu Sun

- Co-advisor:
  - Prof. Jinchi Lv

- Qualification and dissertation committee:
  - Prof. Ting Chen
  - Prof. Jed A. Fuhrman
  - Prof. Andrew D. Smith
  - Prof. Shang-Hua Teng
  - Prof. Michael S. Waterman

- All friends and colleagues in CBB program and in USC!

- My parents, my wife, and forthcoming son

Questions?

Image Source: http://safe4work.org/wp-content/uploads/2011/06/smile-in-the-sky.jpg

References

- Li, Hongzhe. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics and Its Application, 2:73-94, 2015.
- Chen, Jun, and Hongzhe Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. The Annals of Applied Statistics 7:1, 2013.
- Zhang, Xuegong, Shansong Liu, Hongfei Cui, and Ting Chen. Reading the Underlying Information From Massive Metagenomic Sequencing Data. Proceedings of the IEEE, 105(3):459-473, 2017.
- Thomas J. Sharpton. An Introduction to the Analysis of Shotgun Metagenomic Data. Plant Genetics and Genomics, 209(5), 2014.
- Lu, Yang Young, Ting Chen, Jed A Fuhrman, and Fengzhu Sun. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge. Bioinformatics, 33(6):791-798, 2017.
- Lu, Yang Young, Kujin Tang, Jie Ren, Jed A Fuhrman, Michael S Waterman, and Fengzhu Sun. CAFE: accelerated alignment-free sequence analysis. Nucleic Acids Research, gkx351, 2017.
- Kim, Jingu, and Haesun Park. Sparse nonnegative matrix factorization for clustering. Georgia Institute of Technology, 2008.
- Hartigan, John A and PM Hartigan. The Dip Test of Unimodality. The Annals of Statistics, pages 70{84, 1985.
- Sun, Tingni and Cun-Hui Zhang. Scaled sparse linear regression. Biometrika, 99:879-898, 2012.
- Fan, Yingying and Jinchi Lv. Innovated scalable efficient estimation in ultra-large Gaussian graphical models. The Annals of Statistics, 44:2098-2126, 2016.
- Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S Waterman, and Fengzhu Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics, 15(3):343-353, 2014.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. EMNLP. Vol. 14., 2014.