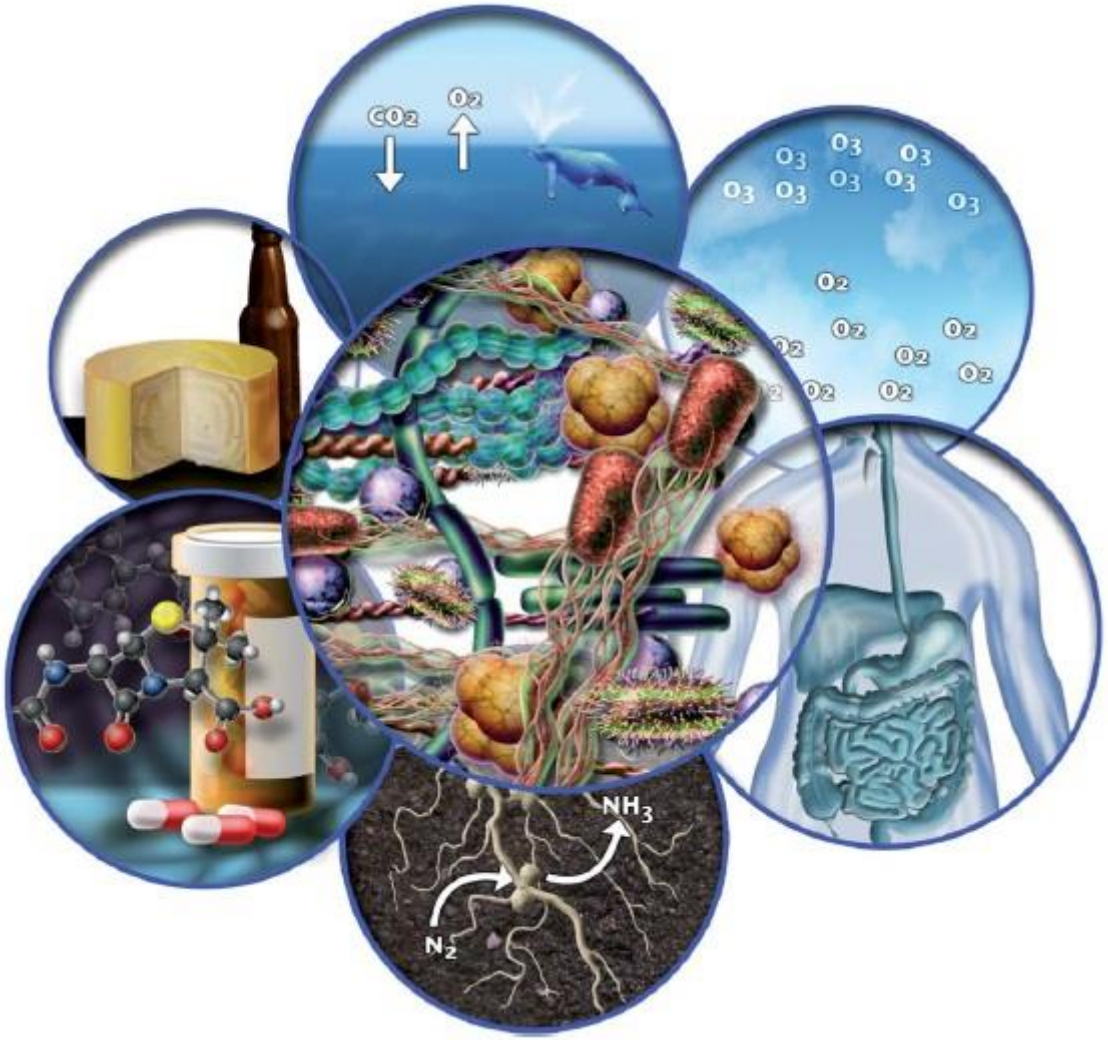
Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge

Yang Lu @ Prof. Fengzhu Sun's Lab
University of Southern California

JSM 2016

Microbes are Everywhere

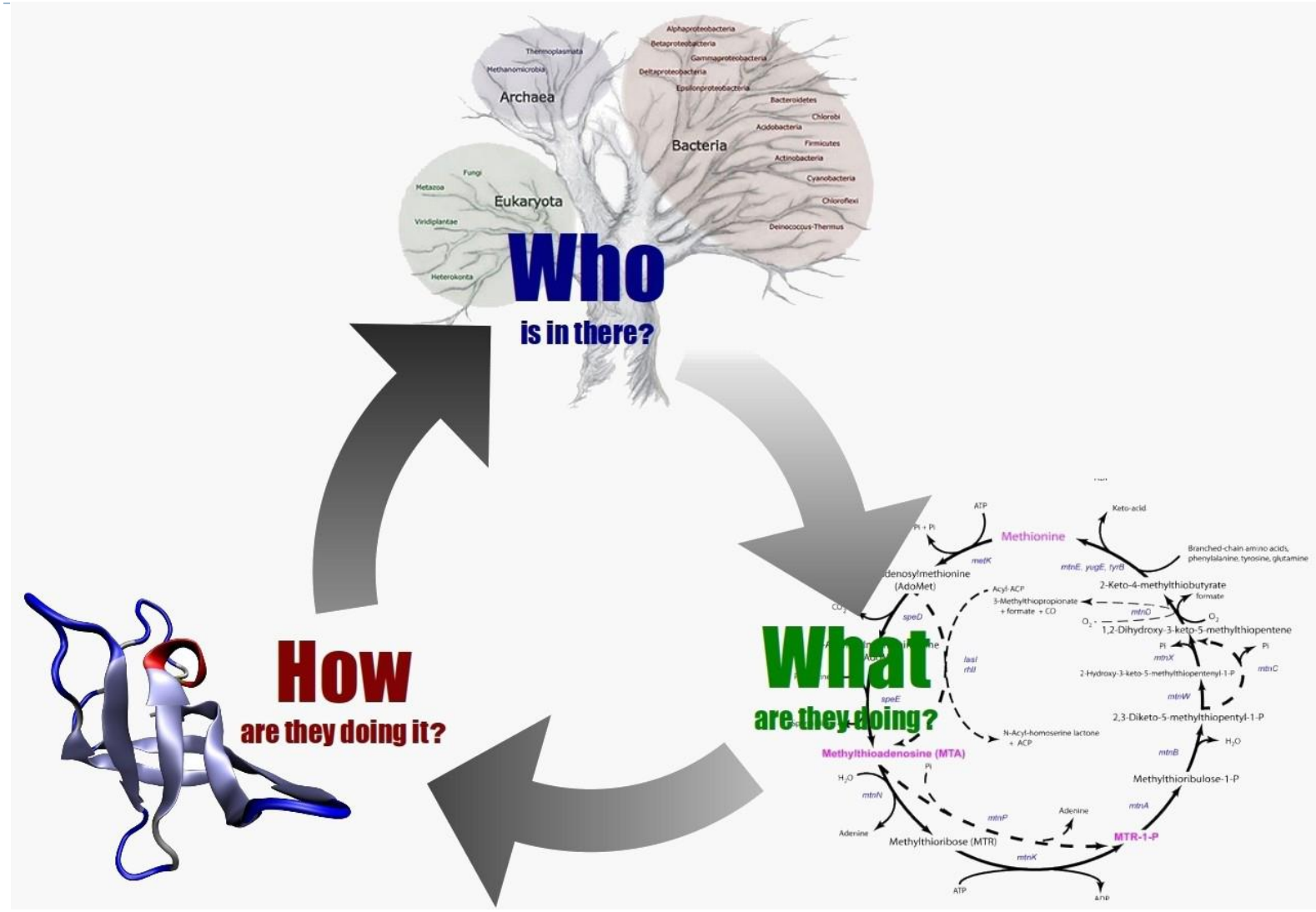


http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0020367

What is Metagenomics?

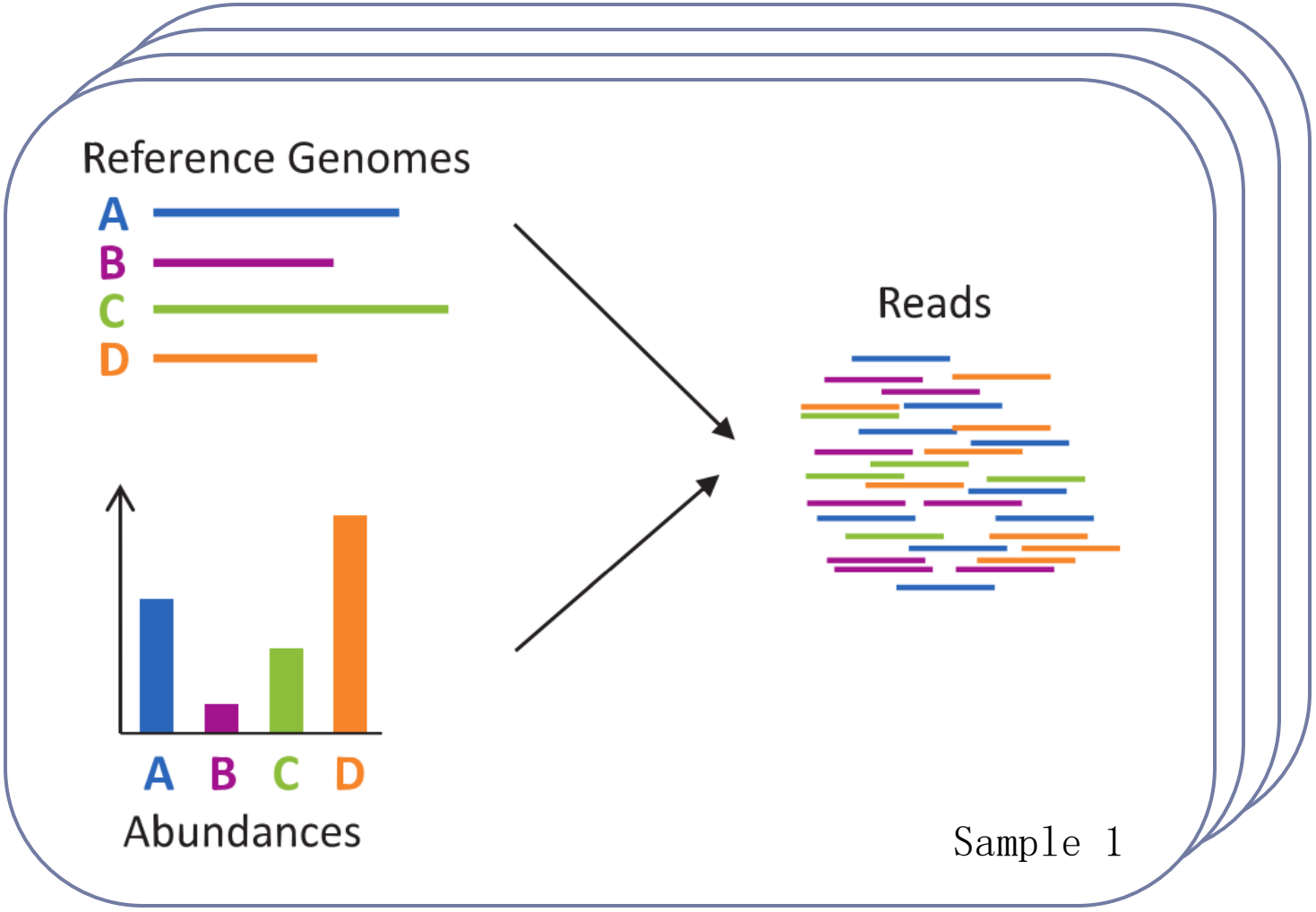
- **Metagenomics** is the study of genetic material recovered directly from environmental samples.
- Many Organisms in one sample
- Many samples from the same environment

Who is in there?

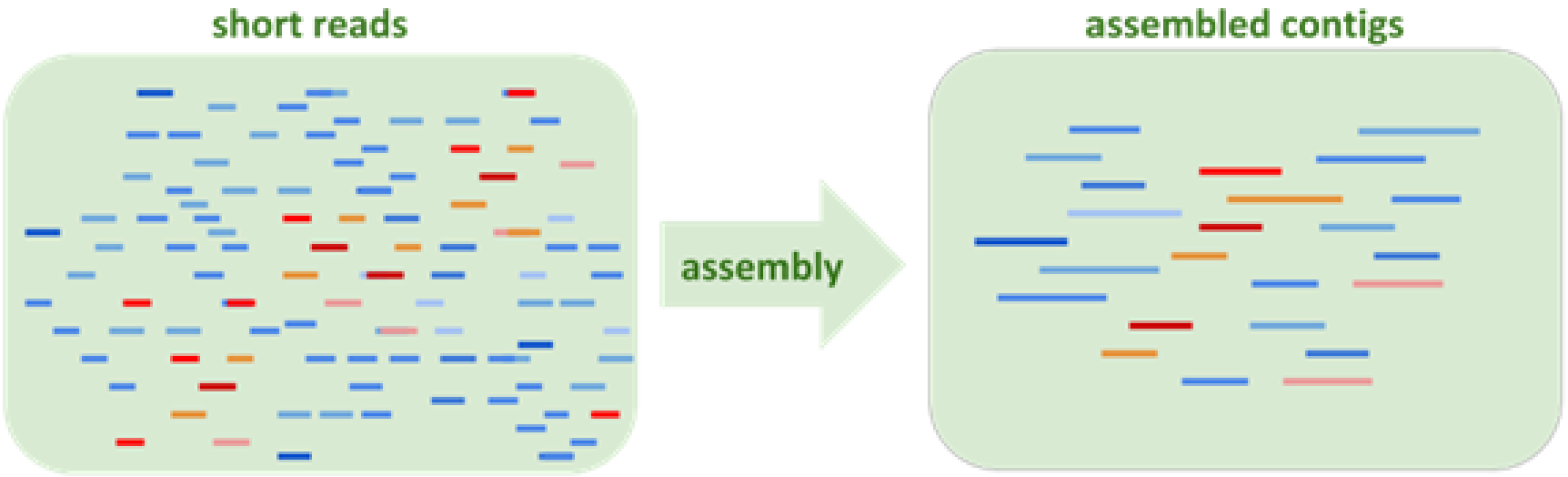


http://www.csirnotes.com/2015/12/metagenomics.html

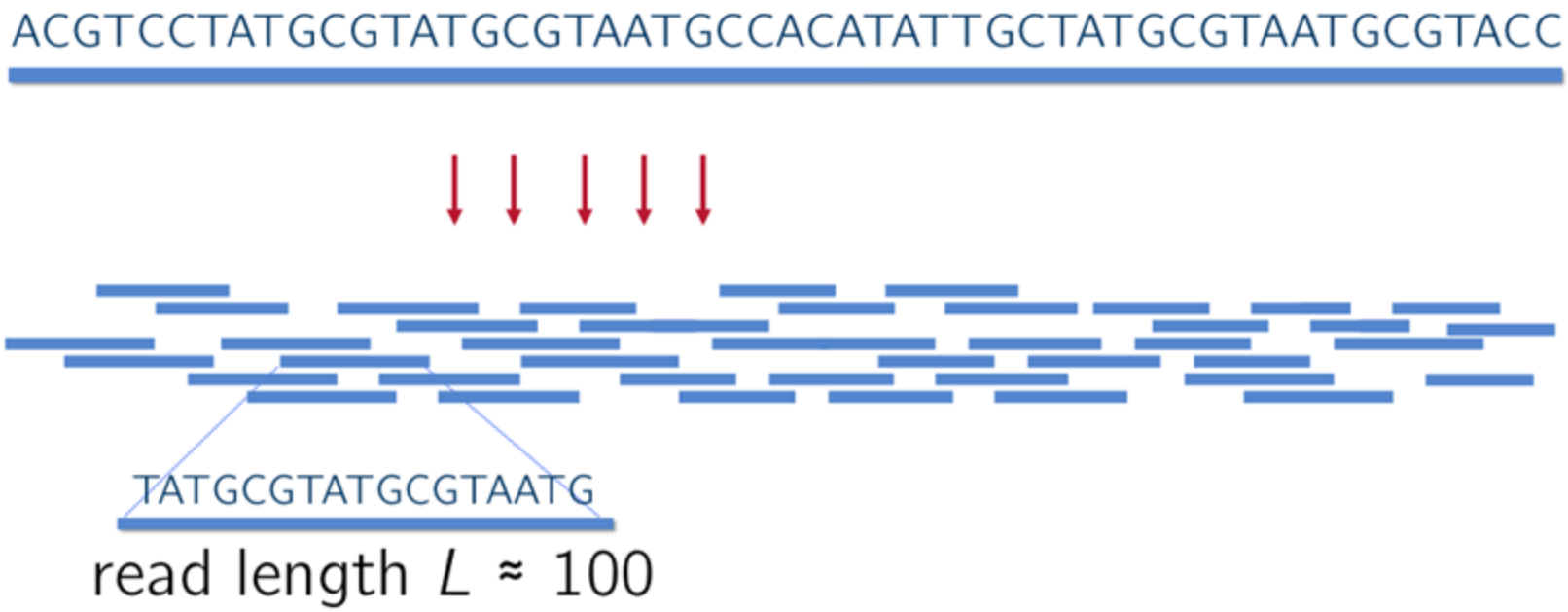
Generative Model of Whole Genome Sequencing



Recover by Assembly



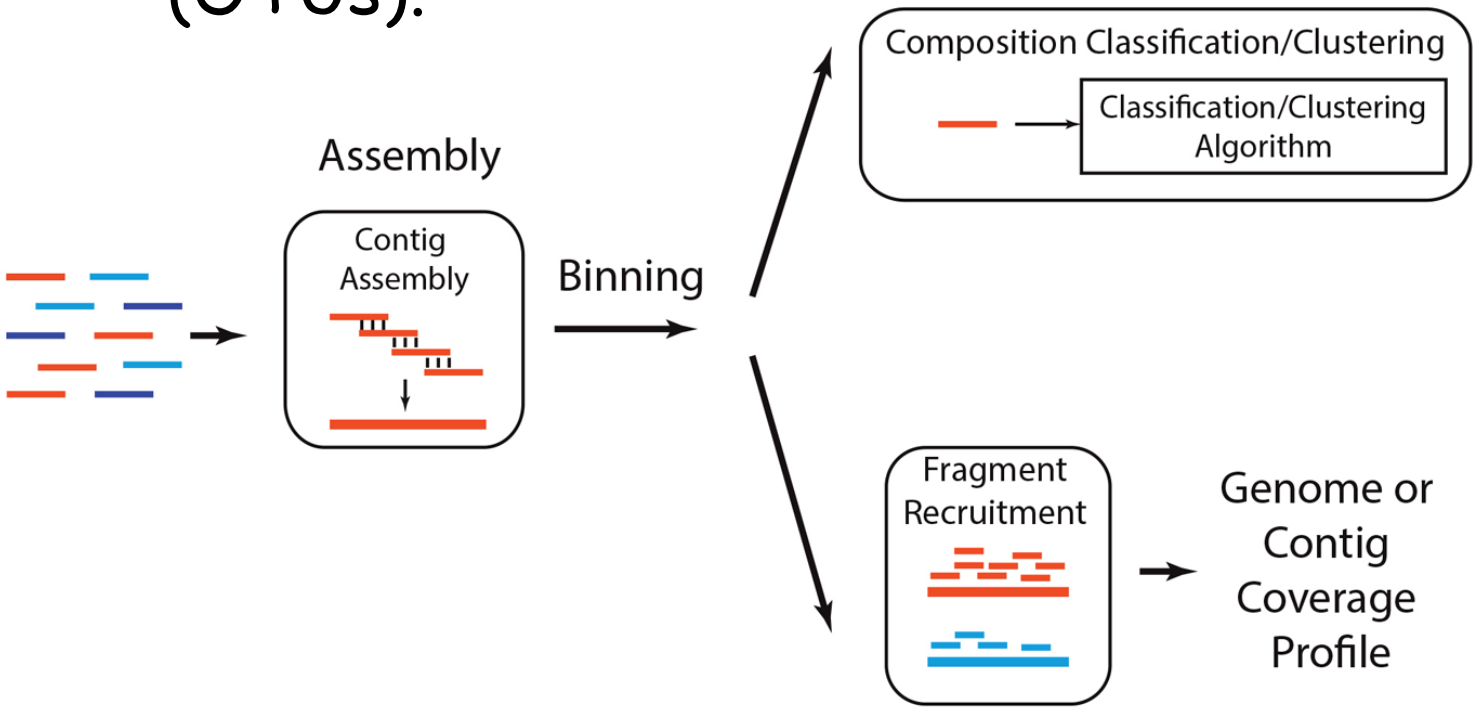
Assembly as Huge Jigsaw Puzzle

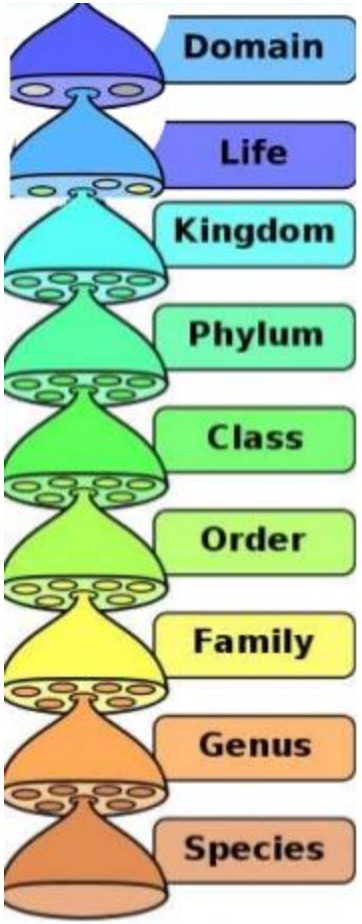


- Assembly is Error-prone
  - Sequencing error rate by technology limitation
  - Strain-level variation by environment complexity
  - Repetitive regions within and across genomes

Metagenomics Binning

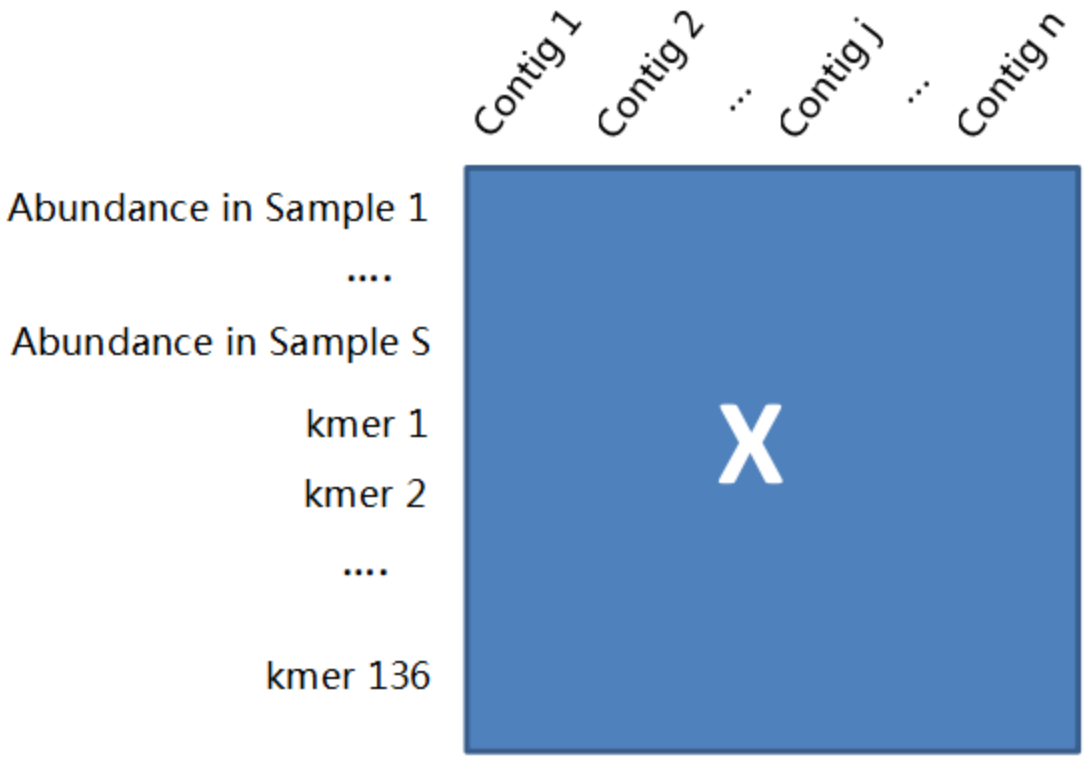
- Group contigs into Operational Taxonomic Units (OTUs).



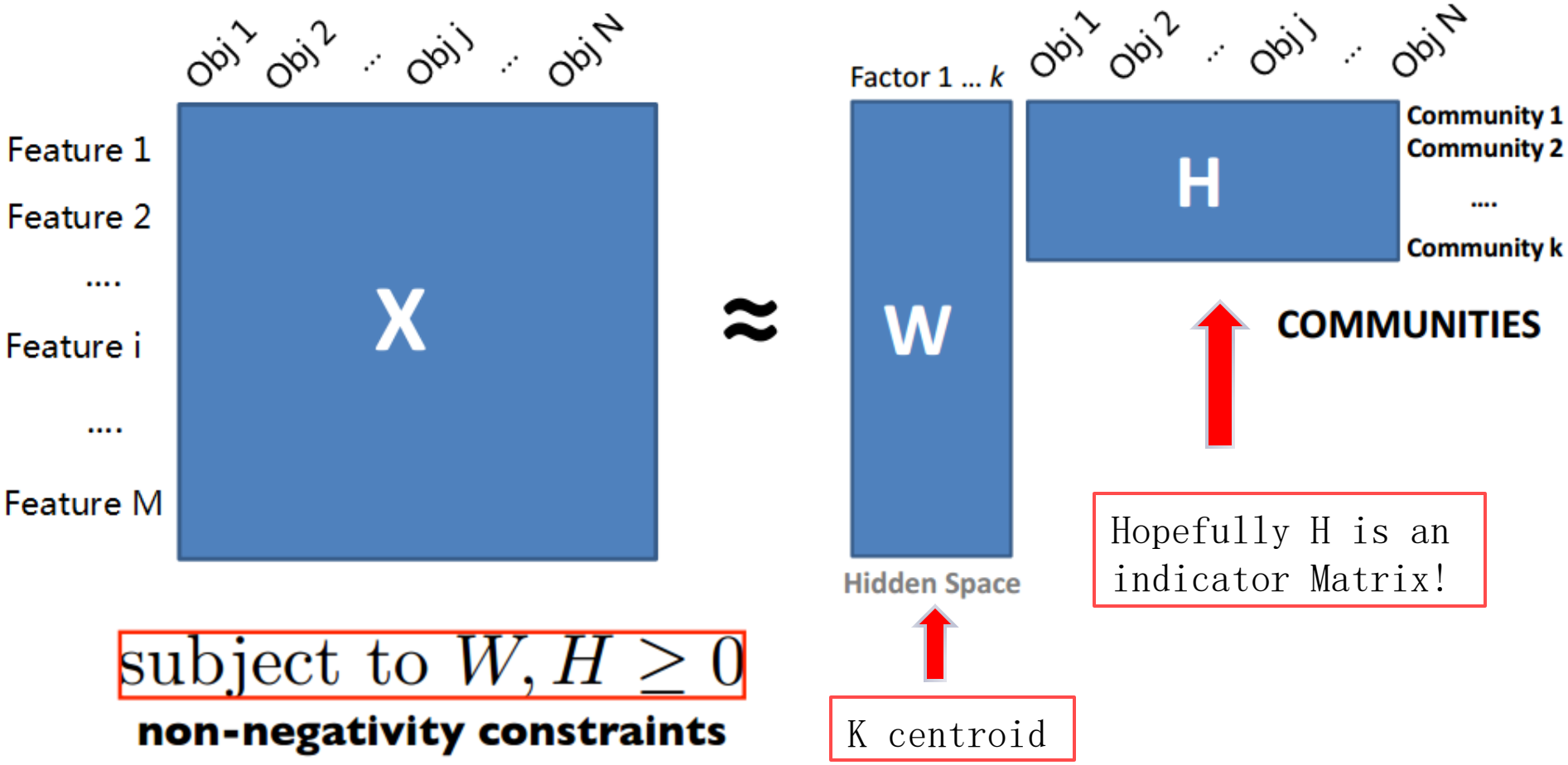


- Representative Methods:
  - MetaBAT, CONCOCT, MaxBin, GroopM.

Feature-Object Matrix Representation



Illustration



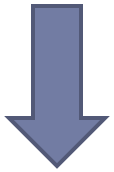
Problem Formulation

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

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

$$\cdots$$

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



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

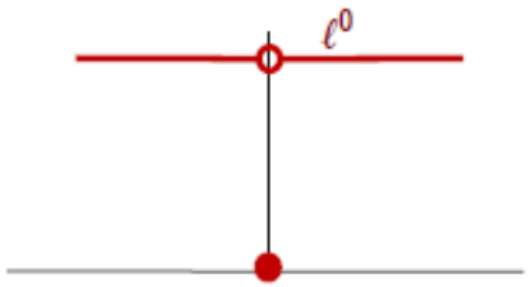
$$s.t.\ \ W \geq 0,\ \ \mathbb{H} \in \{0,1\}^{K\times N}, \|\mathbb{H}_{\cdot j}\|_0 = 1\ for\ j = 1, 2, \cdots, N$$

Relaxation

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

$$\mathbb{H} \in \{0,1\}^{K\times N}$$

Hard to Solve!
Need Relaxation!



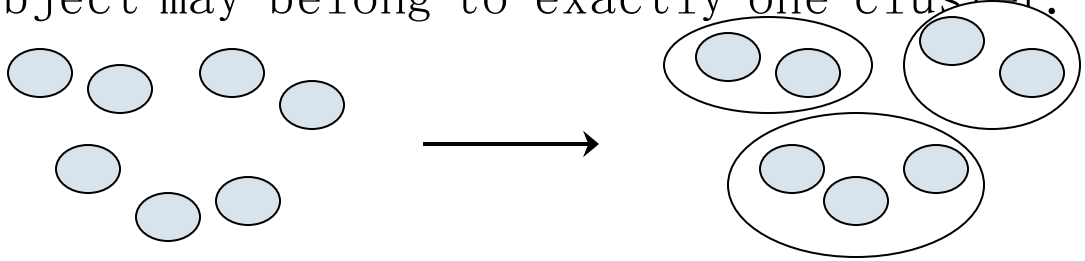
$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2$$

Drawback

- “hard clustering” to “soft clustering”

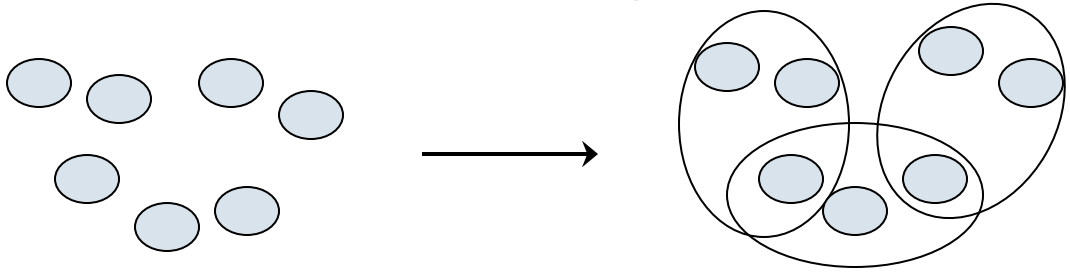
Hard Clustering

- Every object may belong to exactly one cluster.



Soft Clustering

- The membership is fuzzy - Objects may belong to several clusters with a fractional degree of membership in each.

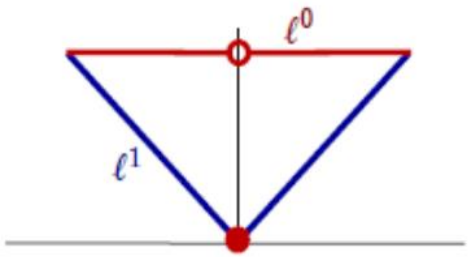


Sparsity comes to rescue

- To facilitate "hard clustering" -like behavior

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

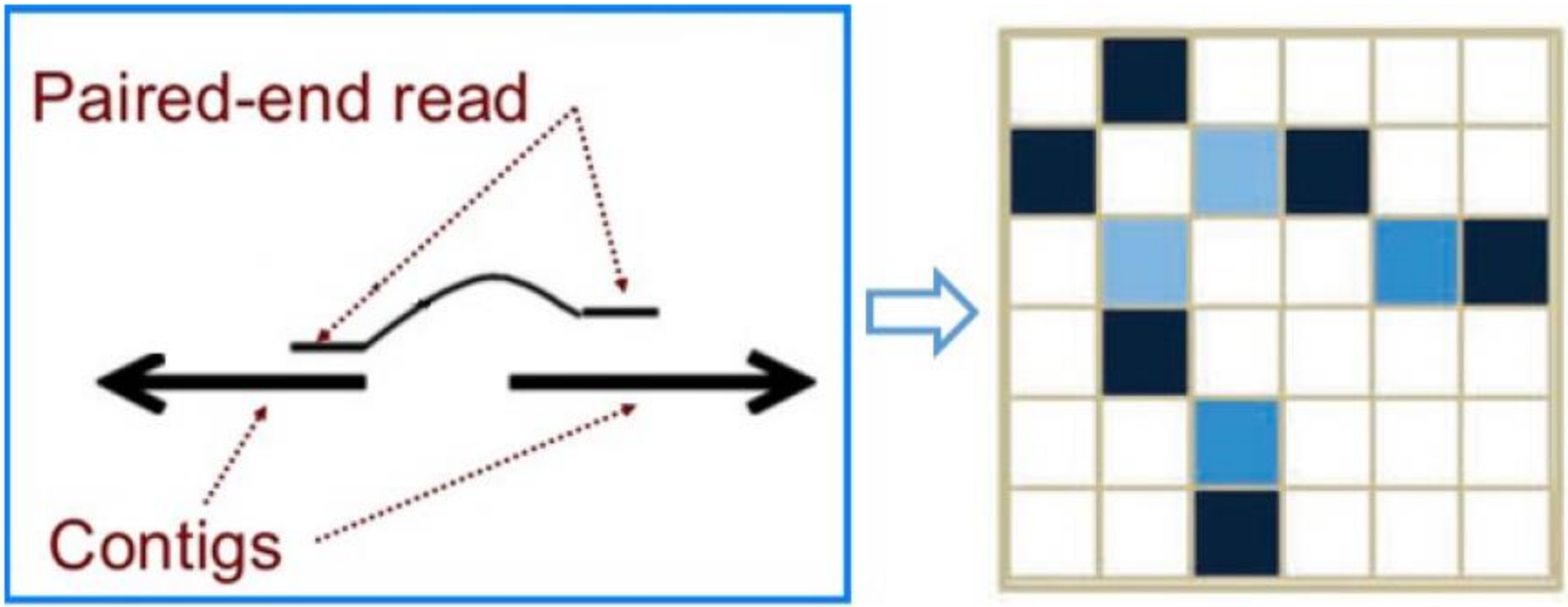
Sparse Non-negative Matrix Factorization



Kim and Park (2008). Sparse nonnegative matrix factorization for clustering. Technical Report

Incorporating Side Information

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



Optimization

- By Alternating Nonnegative Least Squares

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

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

Block Coordinate Descent

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

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

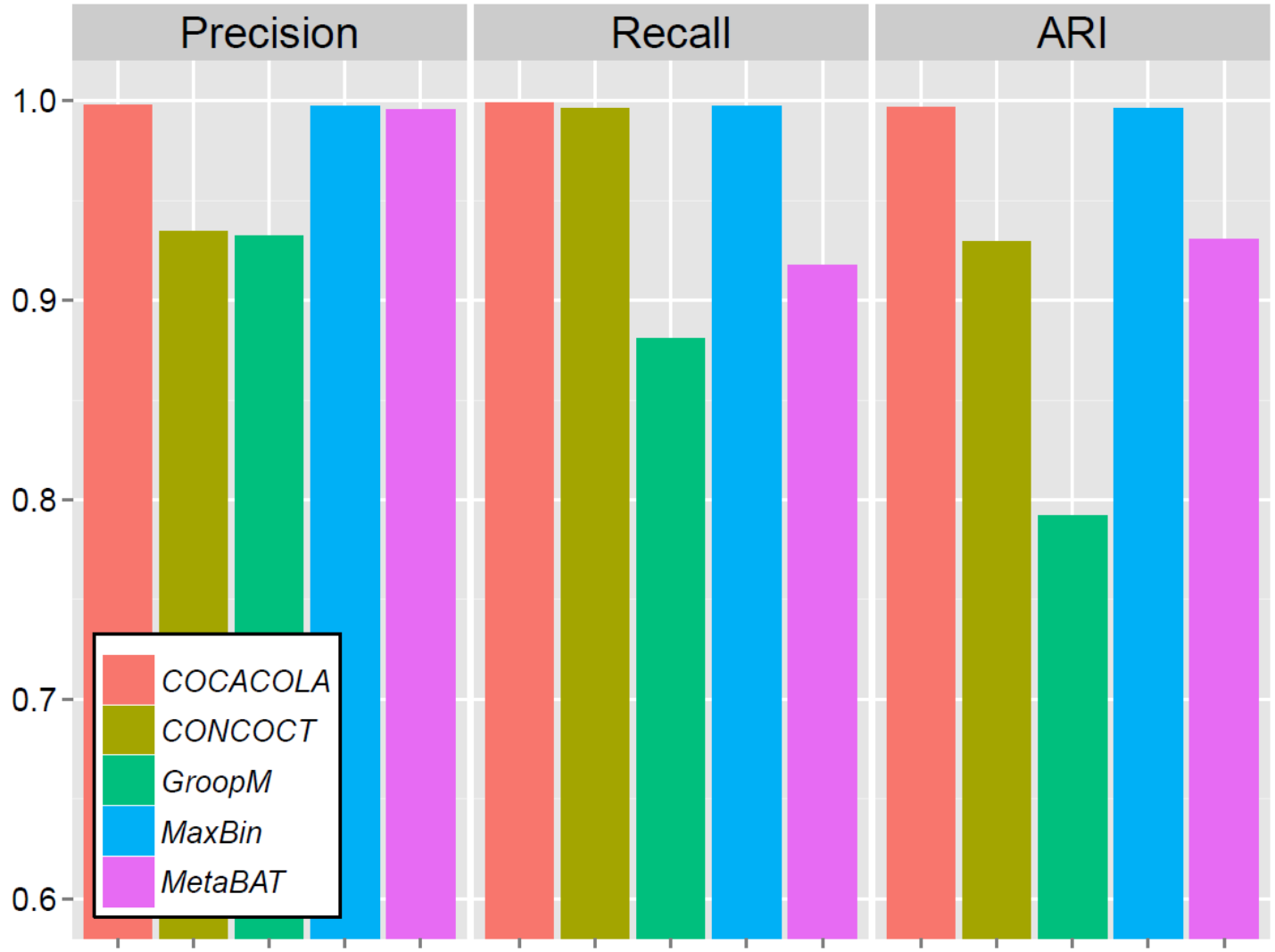
$$= \arg\min_{H\geq 0} \sum_{n=1}^{N} \left(\|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha \|H_{\cdot n}\|_1^2 + \beta \left\| H_{\cdot n} - \sum_{n'=1}^{N} \mathcal{A}_{nn'} H_{\cdot n'}^{old} \right\|_2^2 \right)$$

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

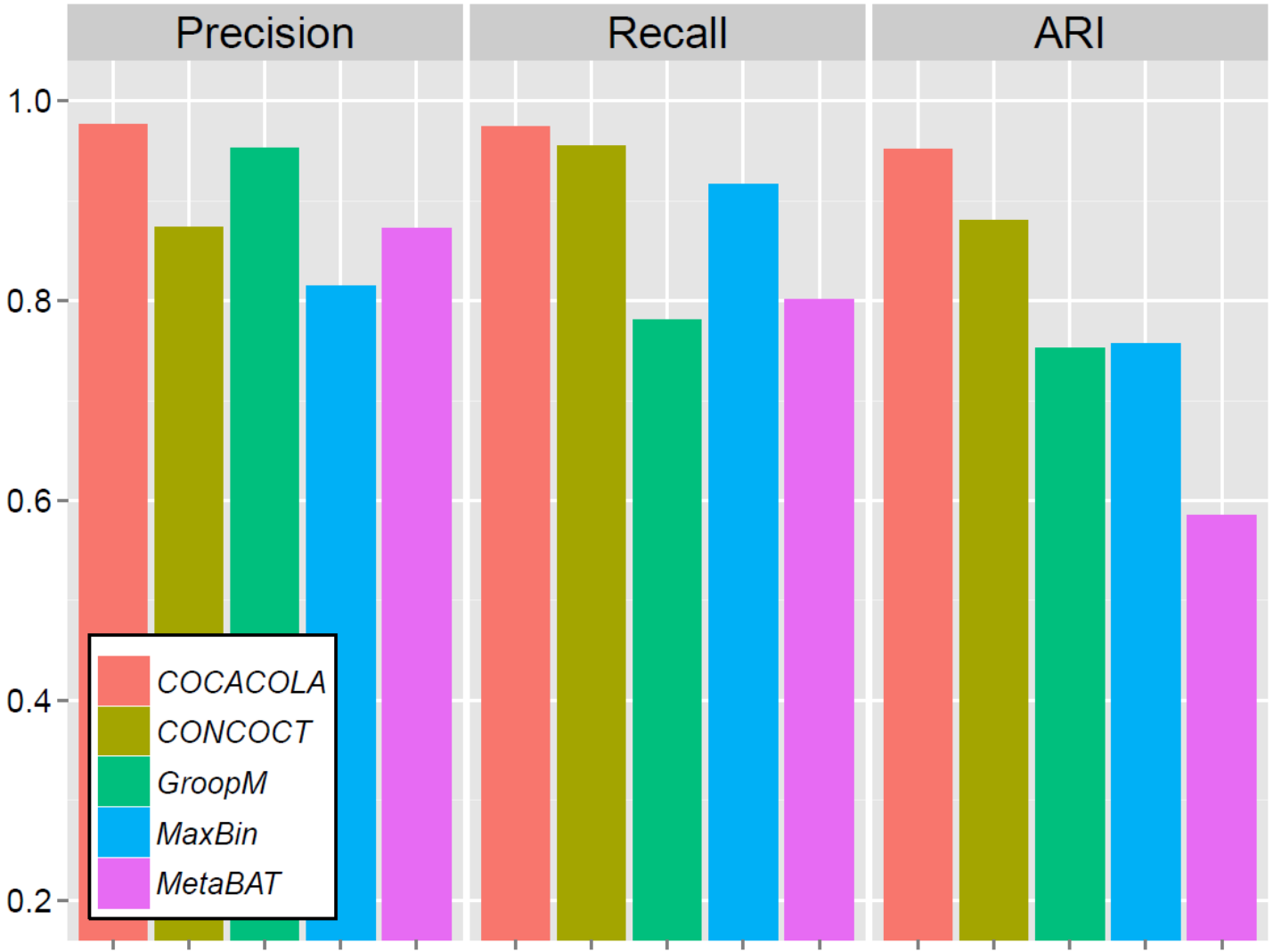
Experiments

- 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

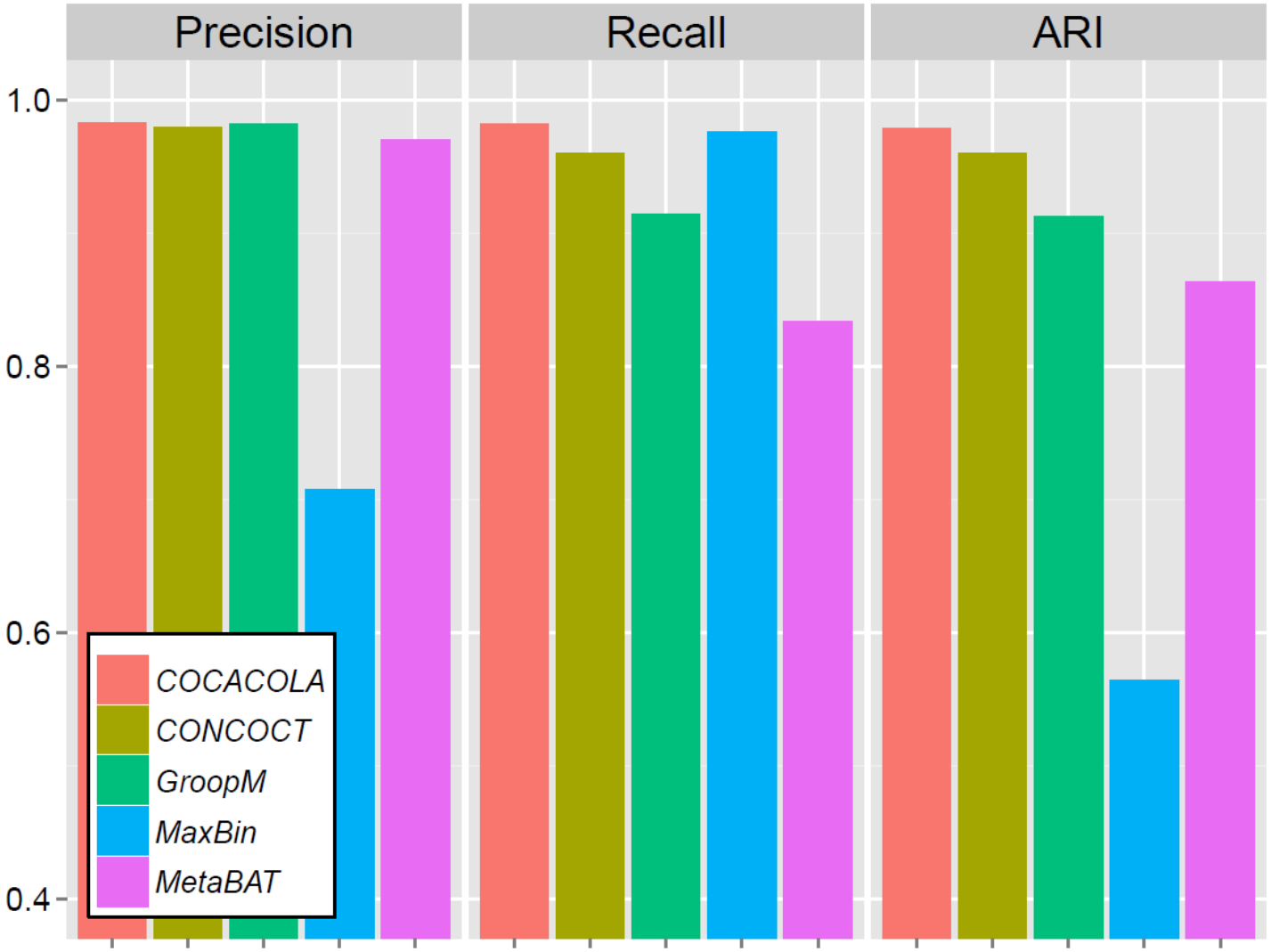
Synthetic "Species" Dataset



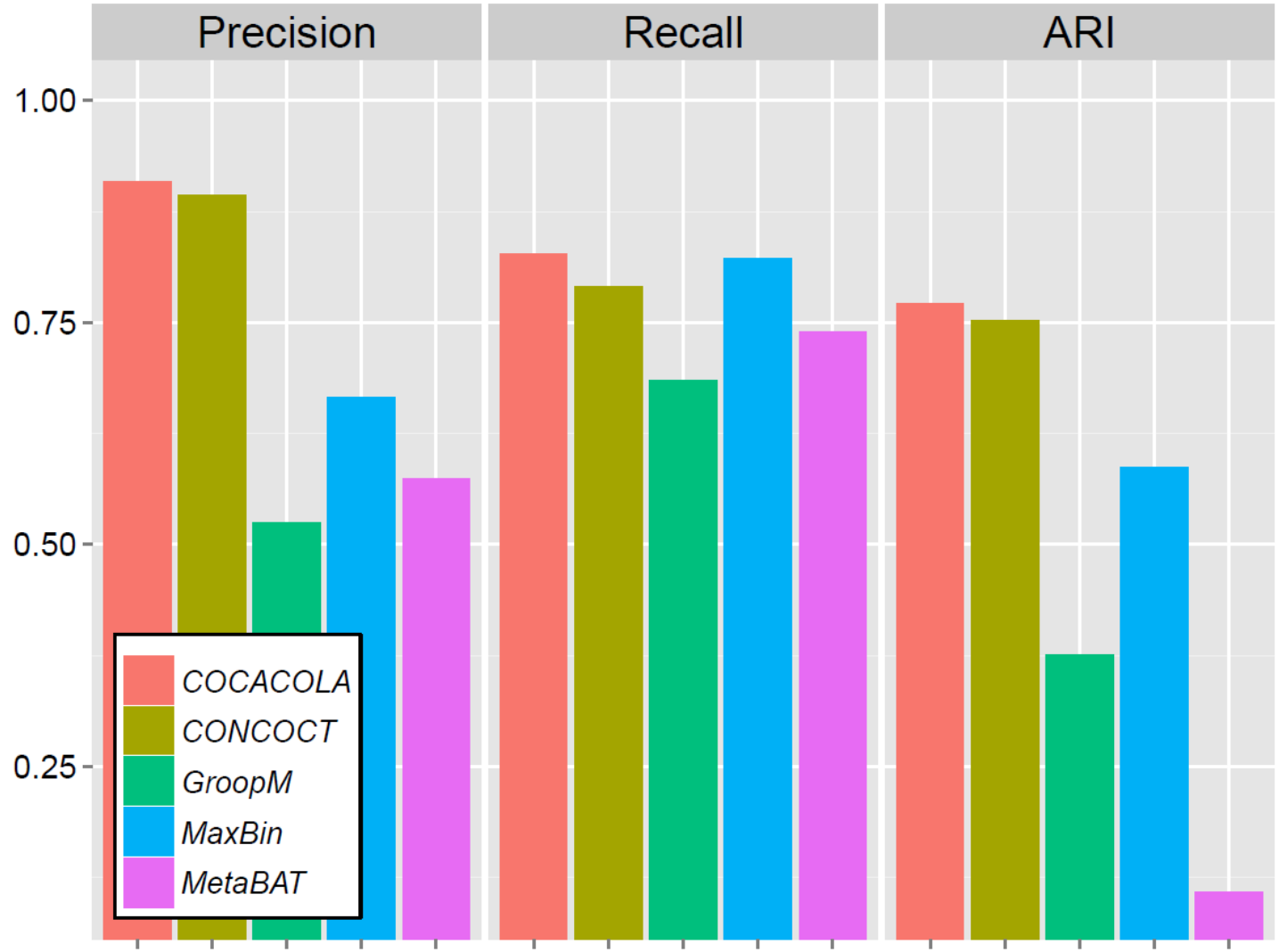
Synthetic "Strain" Dataset



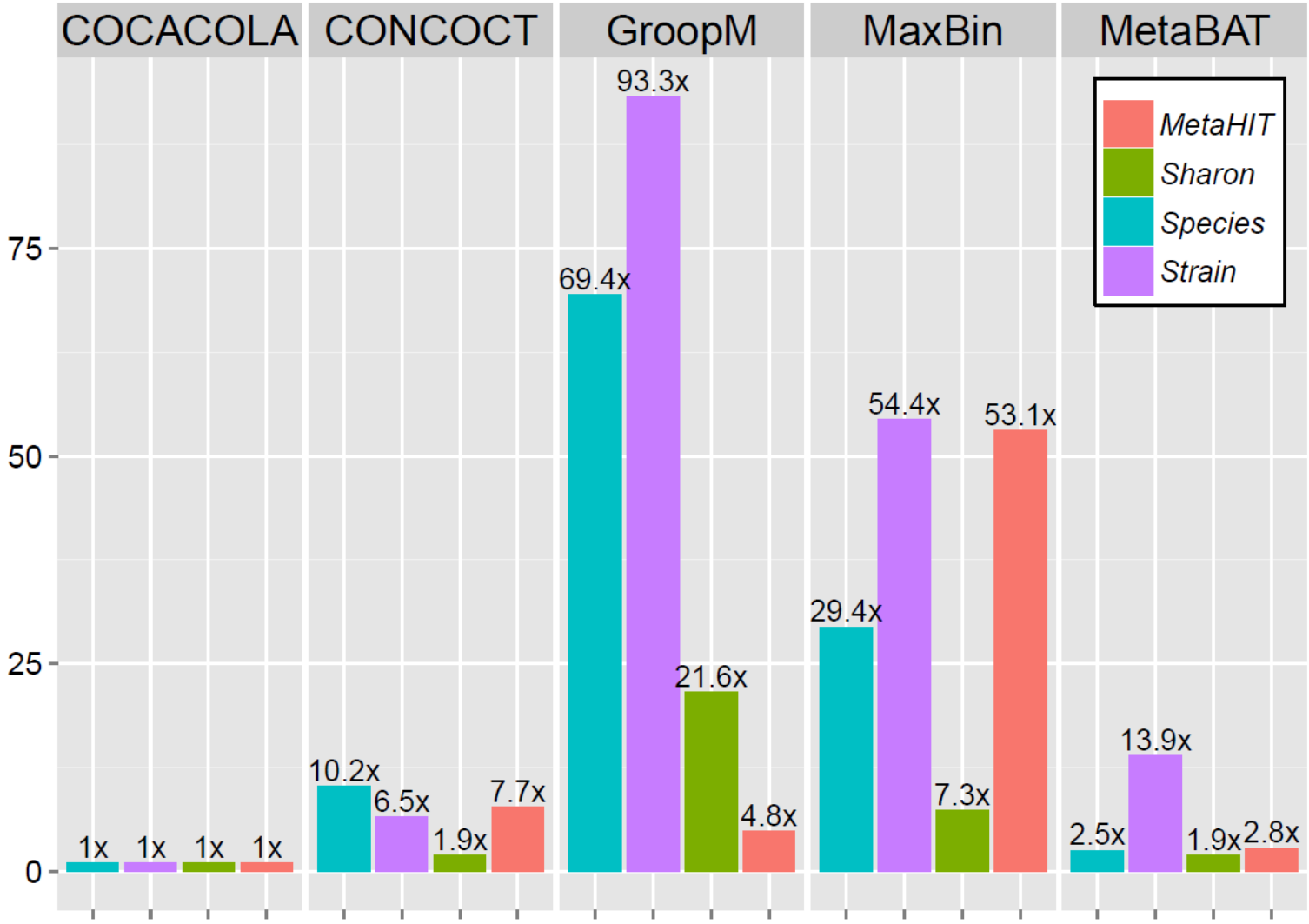
Real "Sharon" Dataset



Real "MetaHIT" Dataset



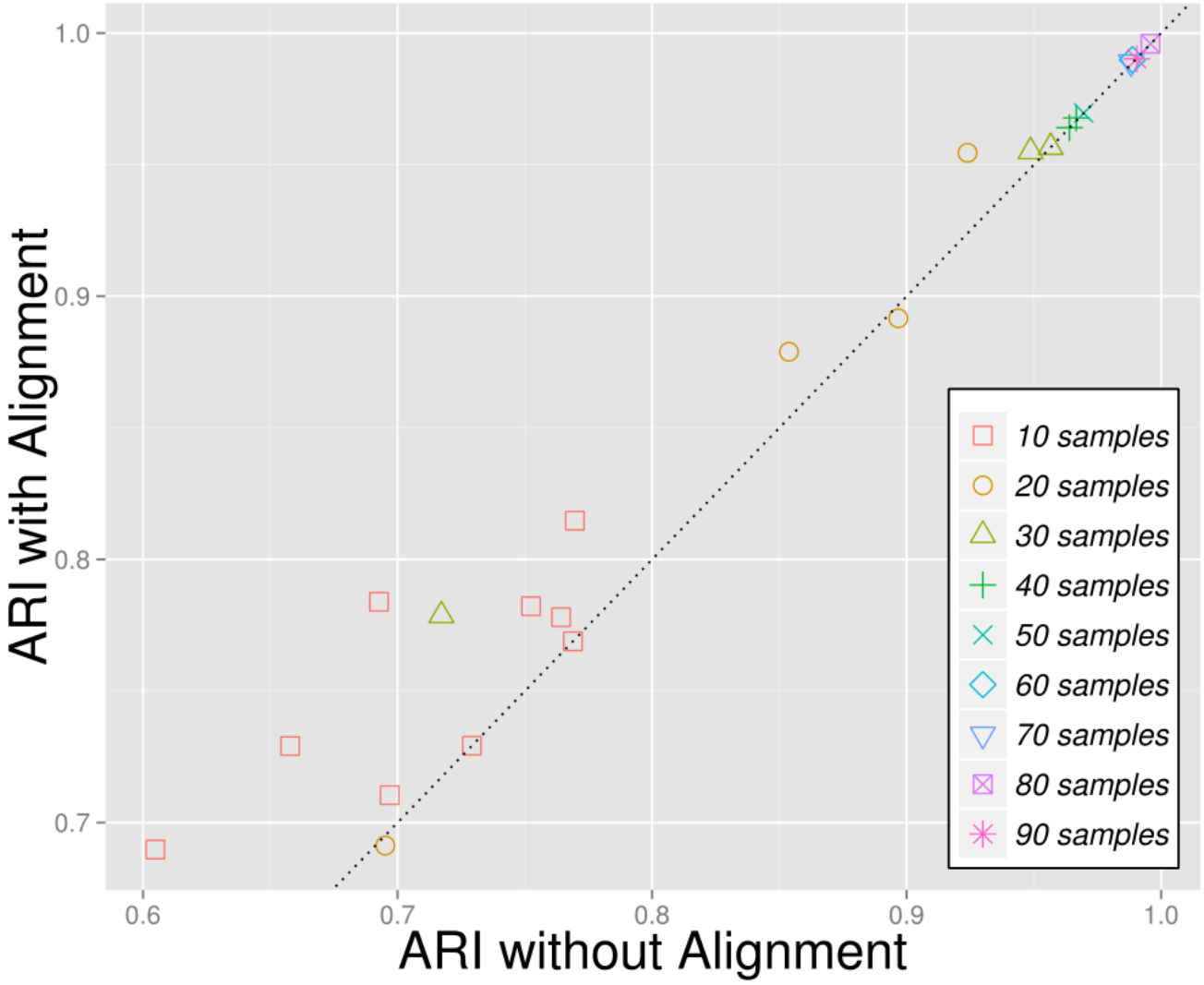
Speedup Ratio



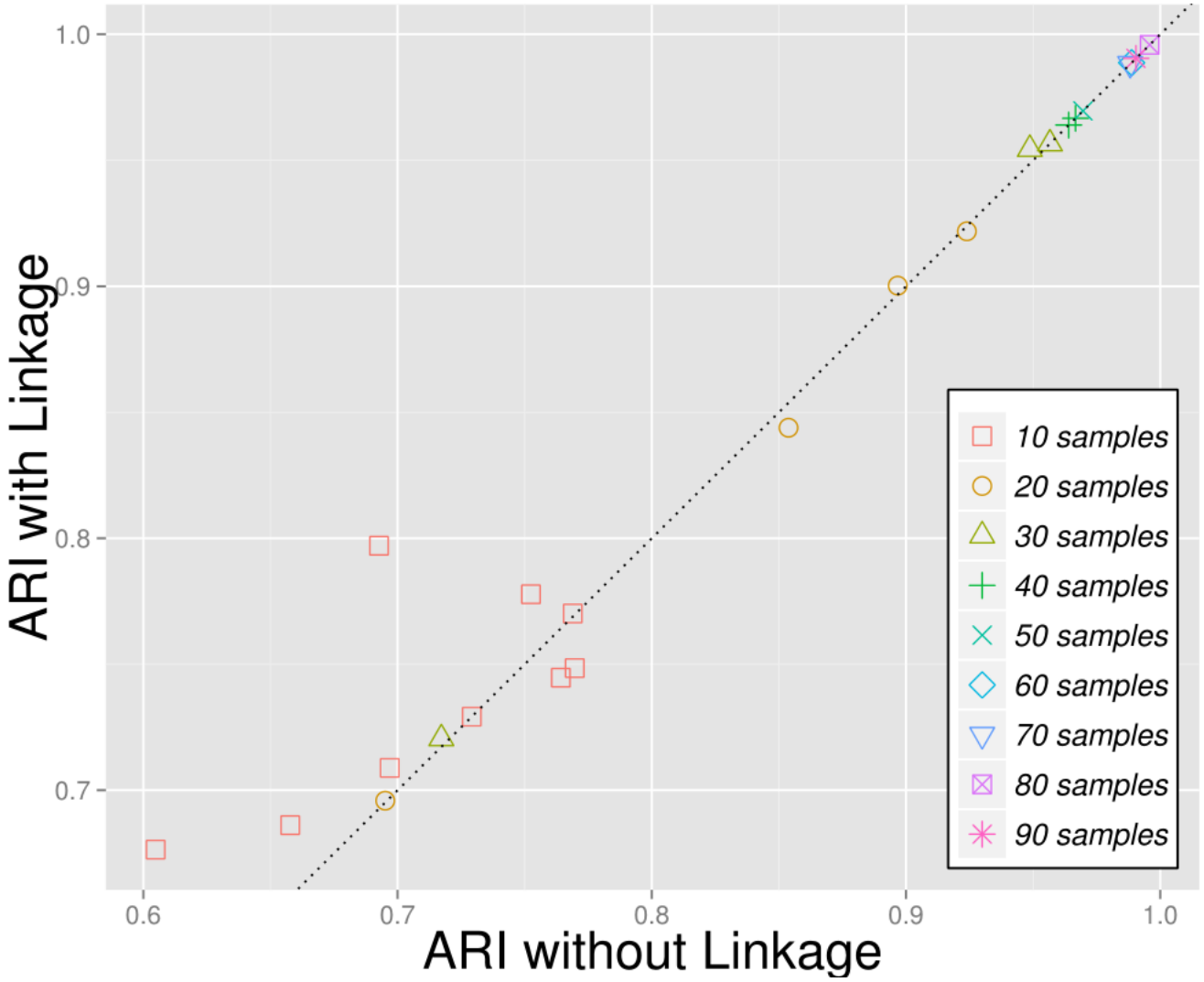
Is Side Information Useful?

- Co-alignment to the same reference genome
- Paired-end reads linkage
  - Minimum samples support = 2
- Ensemble of both with equal weight

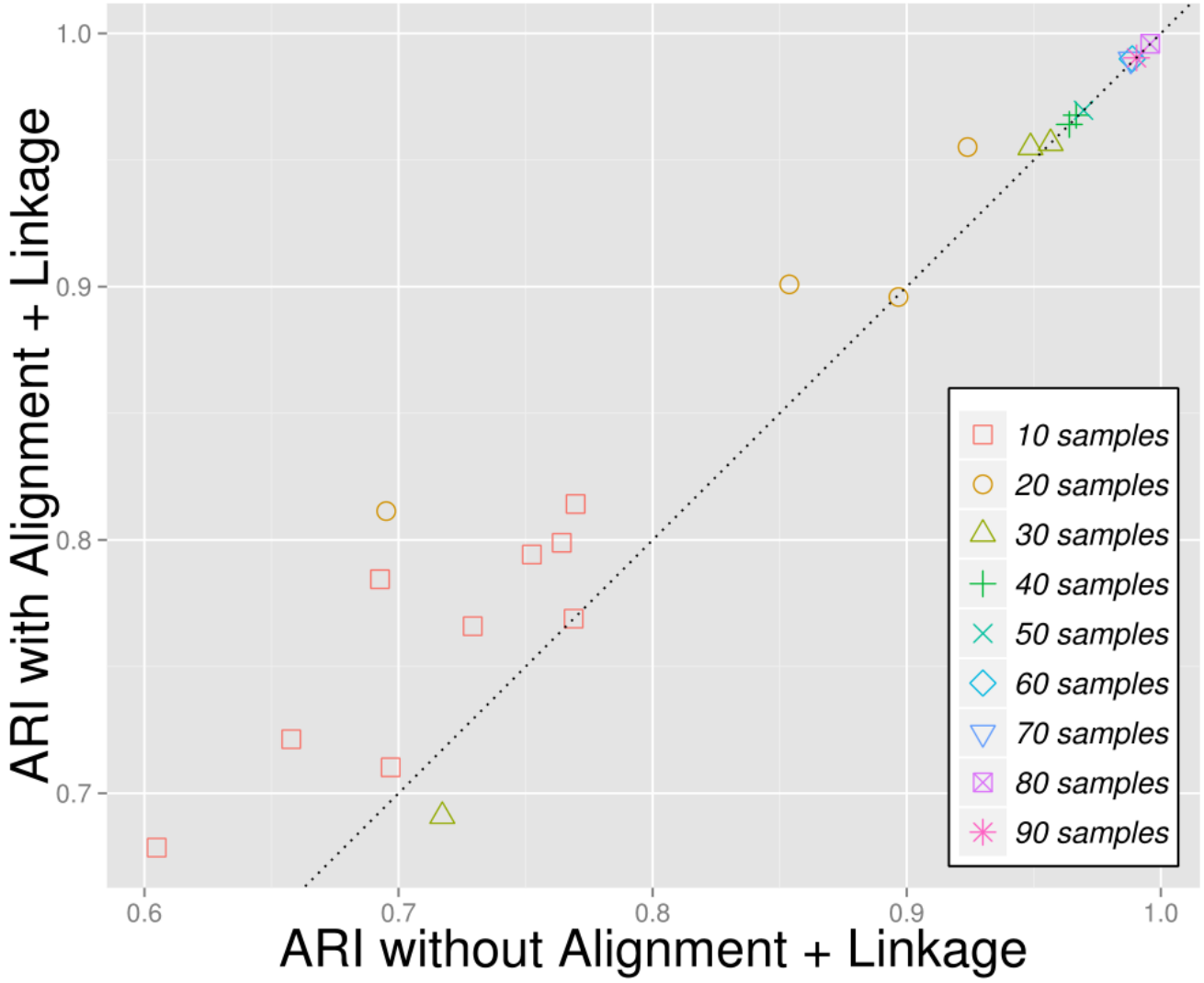
Co-Alignment as Side Information



Linkage as Side Information



Ensemble of Both



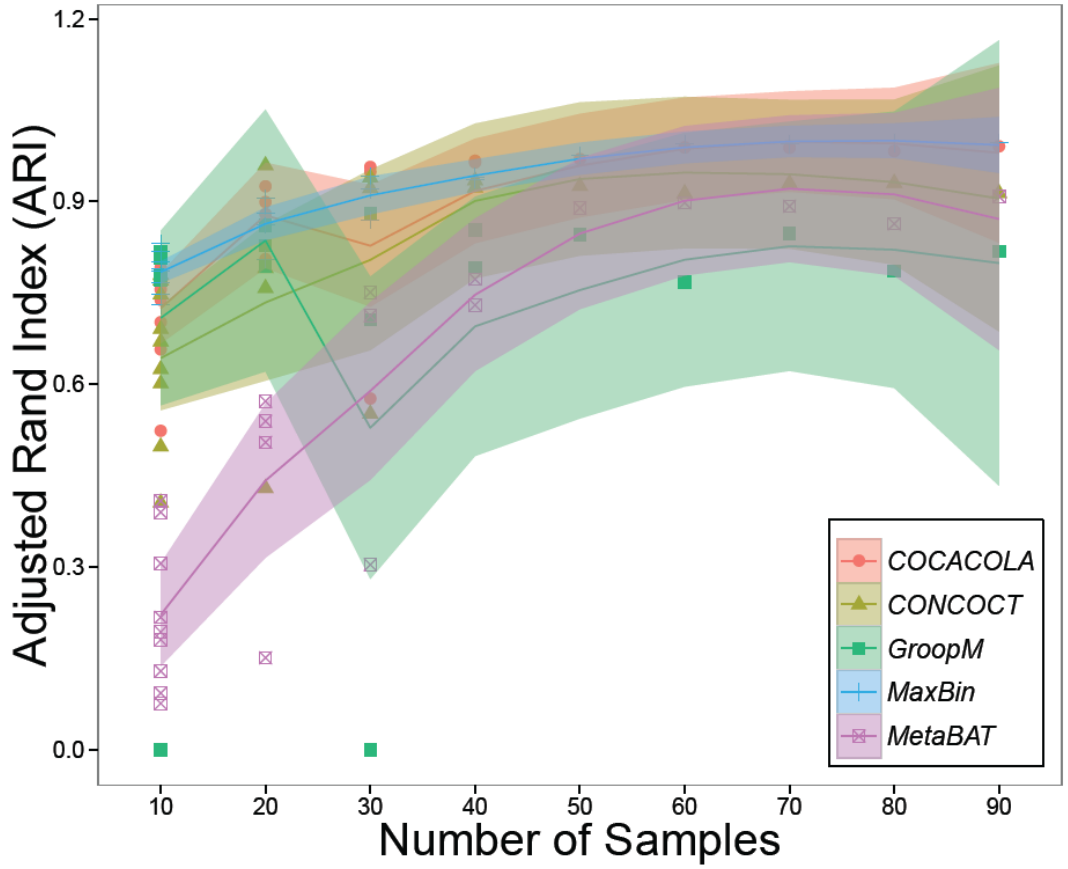
Summary so far

- A metagenomics contigs binning framework
  - Utilize abundance profile and sequence composition
  - Embrace additional information such as co-alignment, linkage, even customized information
  - Highly parallel and scalable

- What's next?

Limitation of Current Binning Approaches

- Observation:
  - When samples size is small, binning is unstable



Re-weight the Input Features

- Assumption:
  - Majority of features are neutral, i.e. with weight 1
  - Only small amount of features are either very good (weight >1) or very bad (weight <1)
- Different from Feature Screening:
  - Majority of features are useless (weight=0)
  - Only small amount of features are important (weight=1)
- For each feature
  - Tested by Multimodality dip test

Maechler (2013). Package 'diptest.

Re-weighting Needs Side Information

- Let A_1 be the KNN matrix of data using heat kernel, symmetrized
- Let A_2 be the side information matrix
- Let $A = A_1 + \gamma A_2$ where $\gamma = tr(A_1' A_2)/(A_1' A_1)$ so that

$$\arg\min_\gamma \|A_1 - \gamma A_2\|_F^2$$

- Objective Function

$$L(W, A) = \sum_{i,j=1}^{N} \|diag(W)X_{.i} - diag(W)X_{.j}\|^2 a_{ij}$$

$$= tr(diag(W)XLX^T diag(W))$$

Objective Function

- Equivalent to a simple version of Mahalanobis distance Learning Formulation
- Doesn't work!
- Reformulate in terms of ΔW

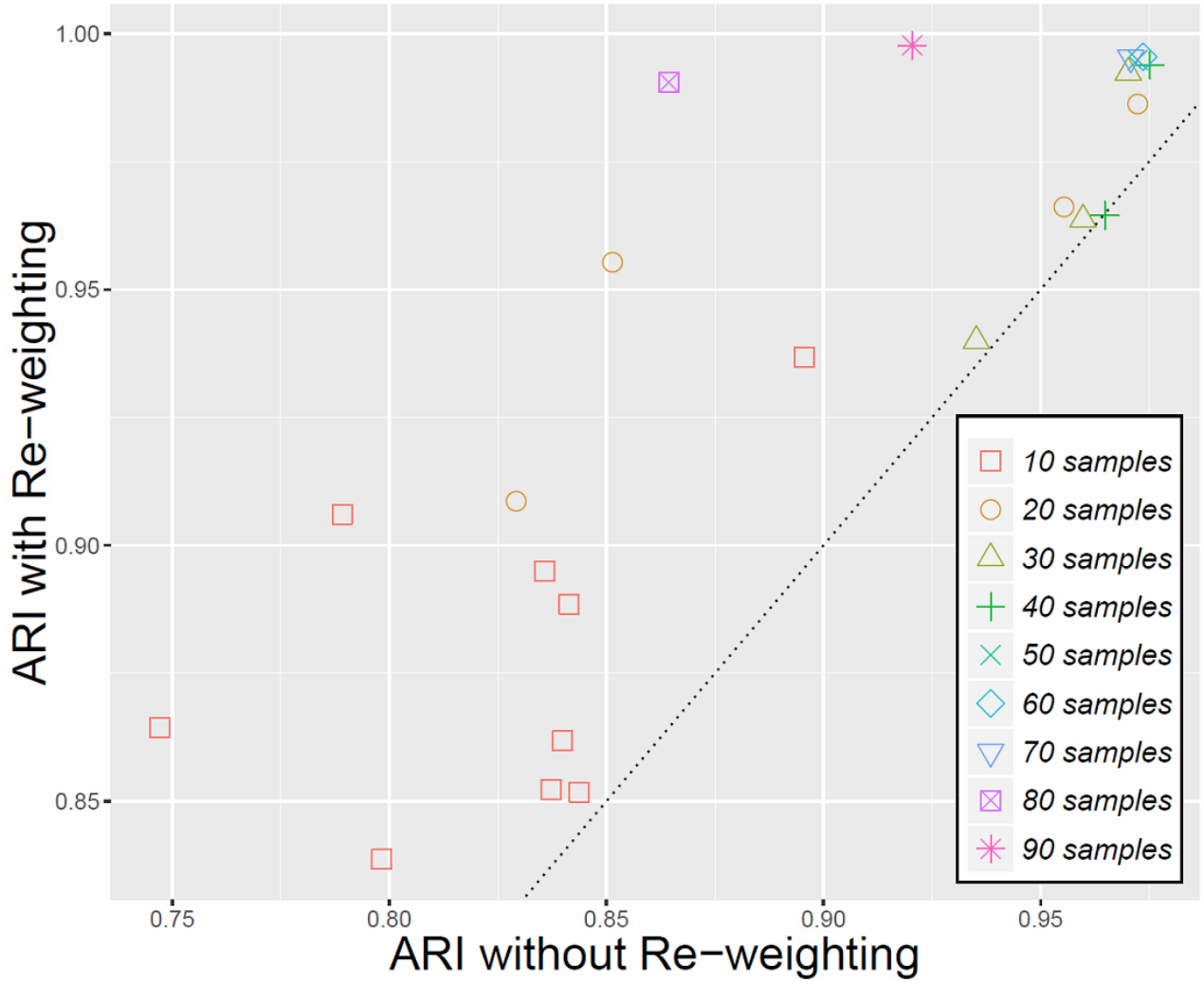
$$L(W) = tr(diag(W)XLX^T diag(W))$$
$$= \|Z diag(W)\|_F^2$$
$$= \|Z + Z diag(\Delta W)\|_F^2$$

Final Objective Function

$$\arg\min_{\Delta W} \left\{ \| Z + Z diag(\Delta W) \|_F^2 + \lambda \| \Delta W \|_1 + \lambda \| \Delta W \|^2 \right\}$$

$$s.t. \quad \sum_{i=1} \Delta W_i = 0 \ , \quad \Delta W_i \geq -1$$

Spectral Clustering after re-weighting the input



Ongoing Direction

- Encode Relative Comparison Information into regularization:
  - Contig A is closer to contig B (within the same species) than A is to C (within the same genus)
  - Incorporate the phylogenetic tree
- Feature-reweighting formulation works
  - Not only Metagenomic Binning scenario
  - Not only clustering scenario
  - Not only untransformed feature space scenario
  - More powerful combined with feature screening

Summary so far

- A metagenomics contigs binning framework
  - Utilize abundance profile and sequence composition
  - Embrace additional information such as co-alignment, linkage, even customized information
  - Highly parallel and scalable
- Feature-reweighting for input data enhancement
  - Different assumption compared to feature screening

Acknowledgement

- Research is partially supported by NSF DMS-1518001 and OCE 1136818.

Prof. Fengzhu Sun
@ Comp. Bio

Prof. Ting Chen
@ Comp. Bio

Prof. Jed Fuhrman
@Marine Bio

Prof. Jinchi Lv
@ DSO

Questions?

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