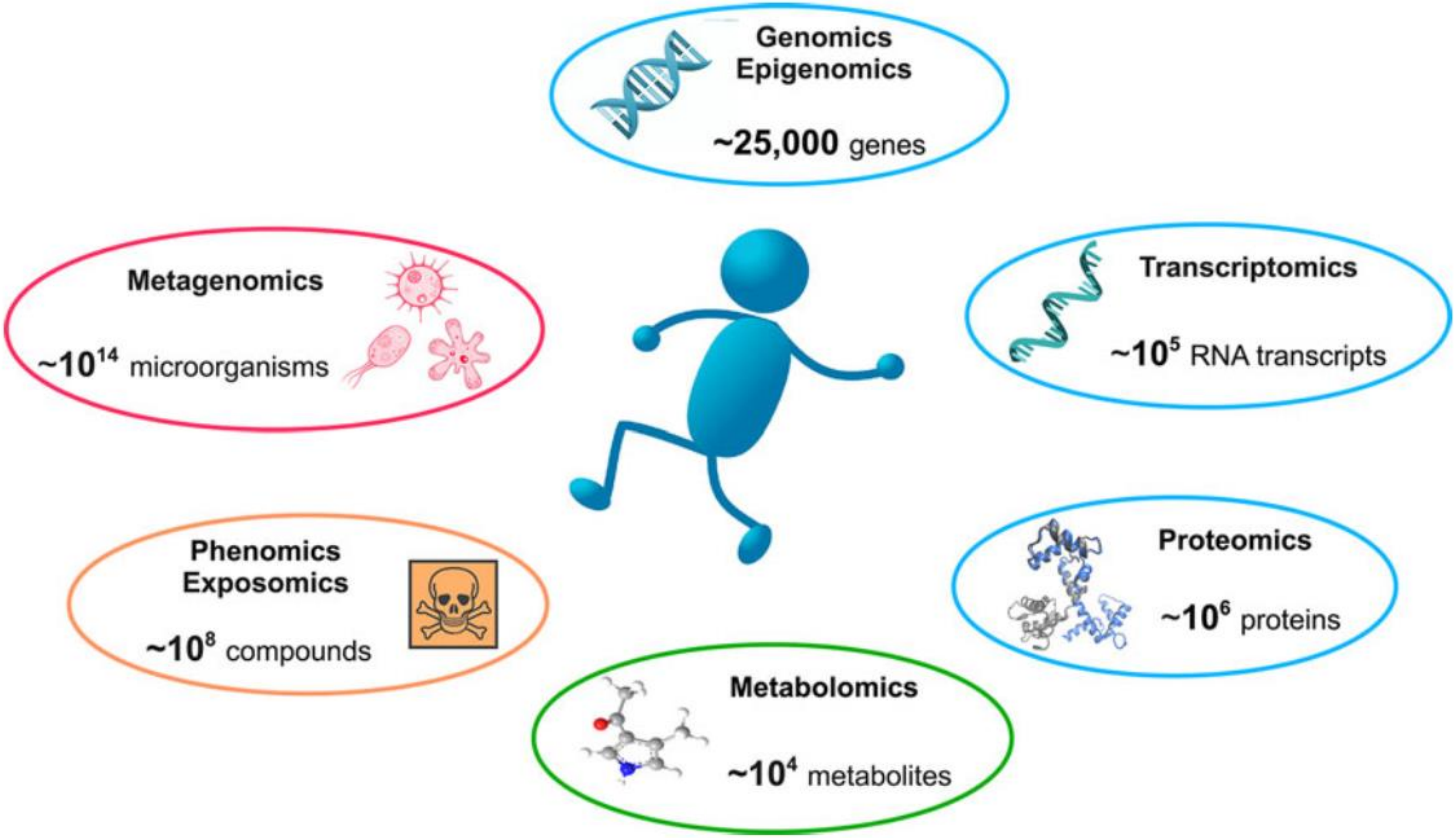
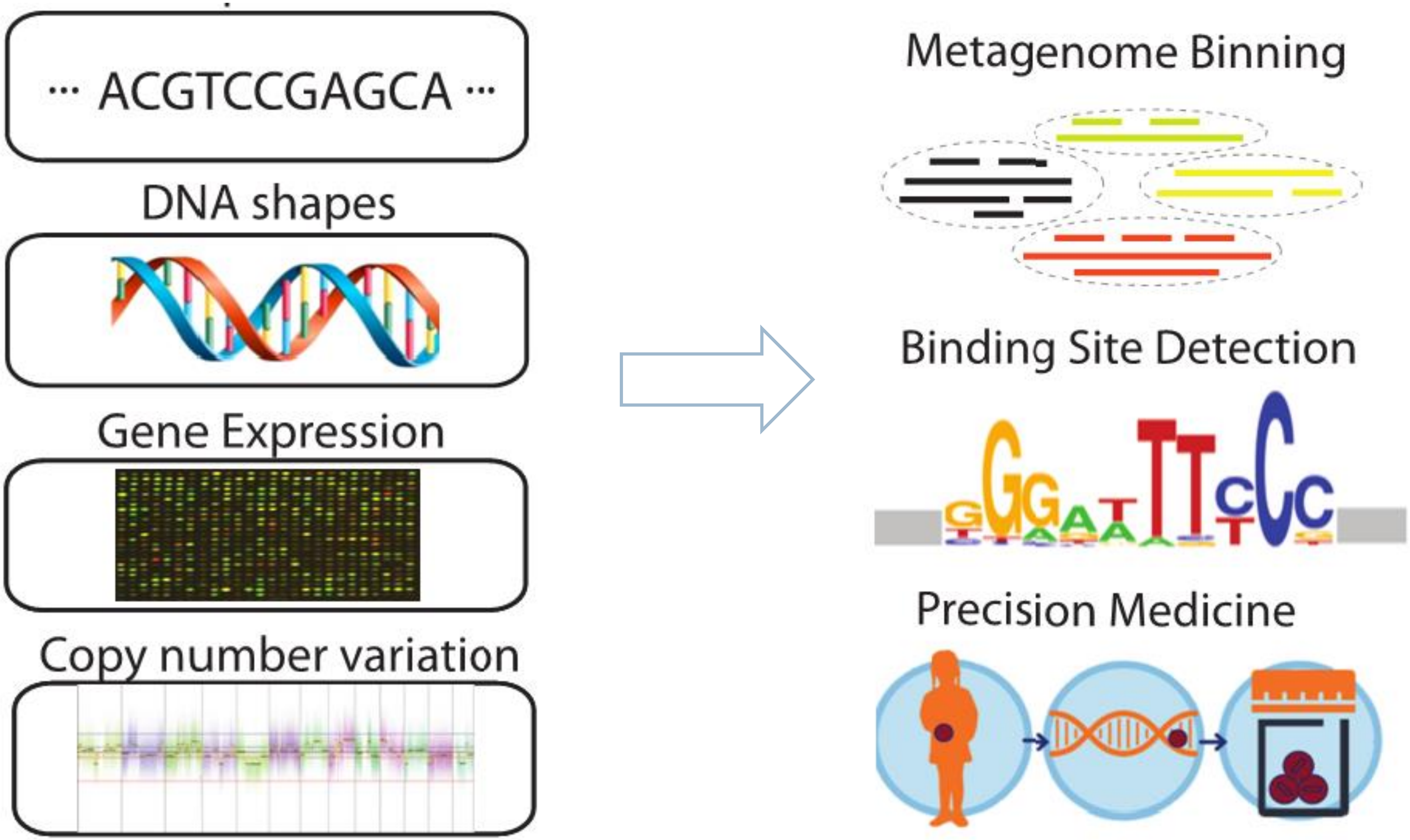
Heterogeneous Feature Weighting Improves Clustering and Classification in Integrative Genomics

Yang Lu

Multi-omics Data are Available



Integrative Genomics are Pervasive



··· ACGTCCGAGCA ···

DNA shapes

Gene Expression

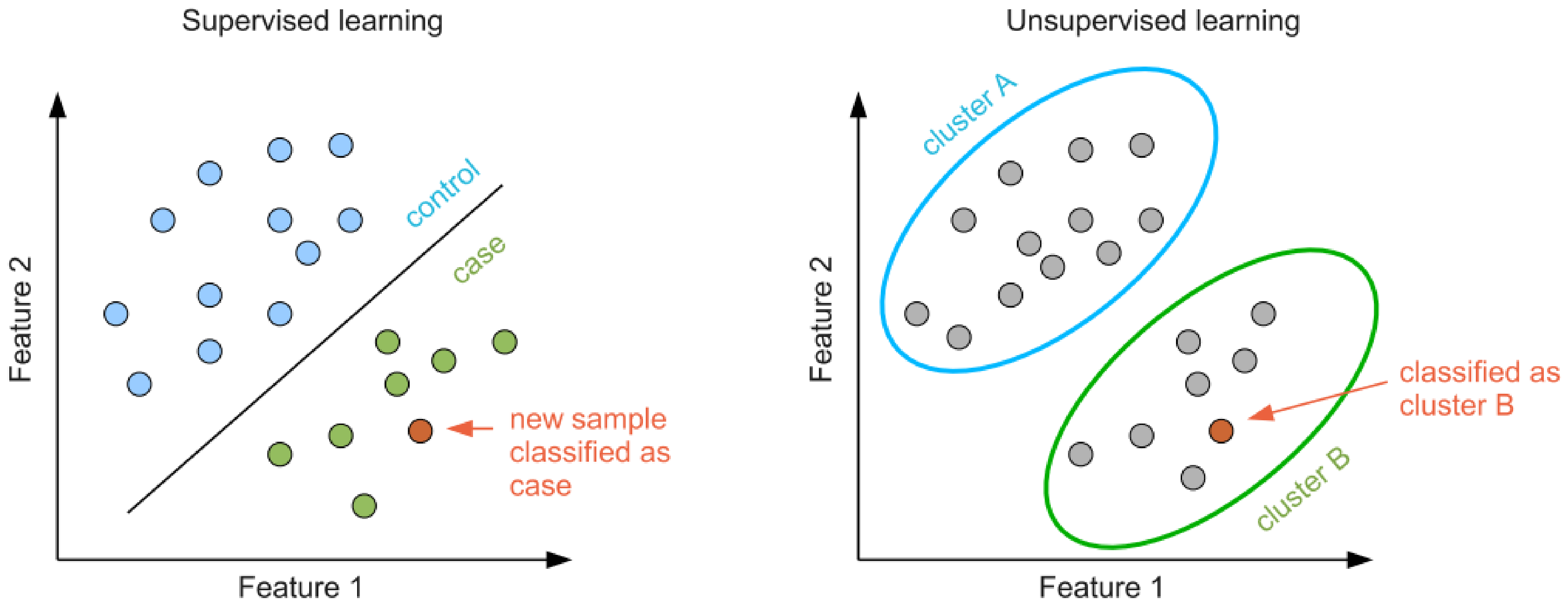
Copy number variation

Metagenome Binning

Binding Site Detection

Precision Medicine

Two Main ML Techniques



Challenges of Integration

- curse of dimensionality
  - Large p vs. small n
- data heterogeneity
  - different omics data vary in data distribution
- unbalanced scales
  - uneven sizes across different types
- noise, redundancy and disagreement among data

Current Solution

- curse of dimensionality
  - Large p vs. small n

  Solution: Feature Selection, sparsity, etc.
- data heterogeneity
  - different omics data vary in data distribution

  Solution: Parameter estimation, etc.
- unbalanced scales
  - uneven sizes across different types

  Solution: Normalization, scaling, etc.
- noise, redundancy and disagreement among data

  Solution: cleaning, consensus analysis, etc.

Can we do better?

Feature Weighting as Preprocessing

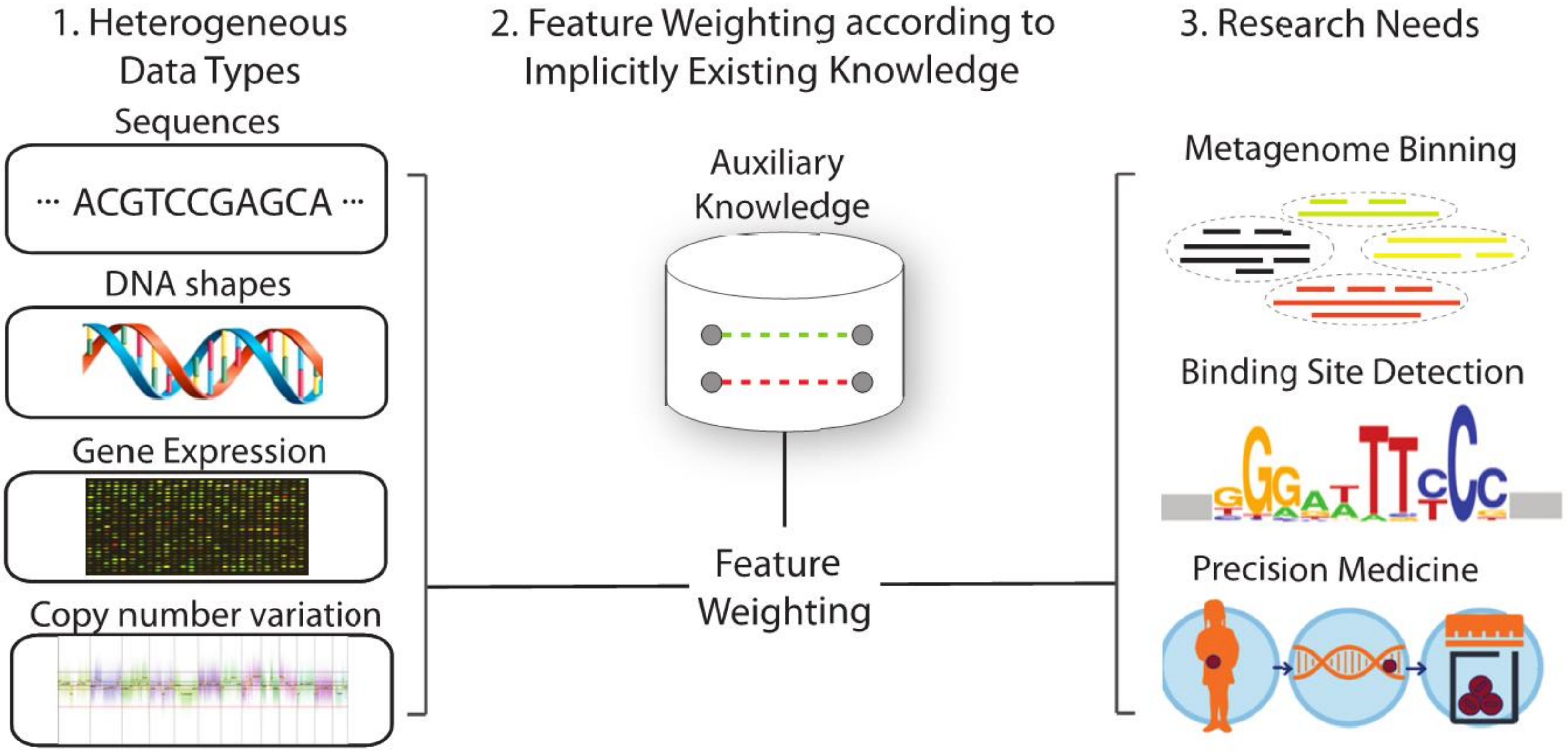
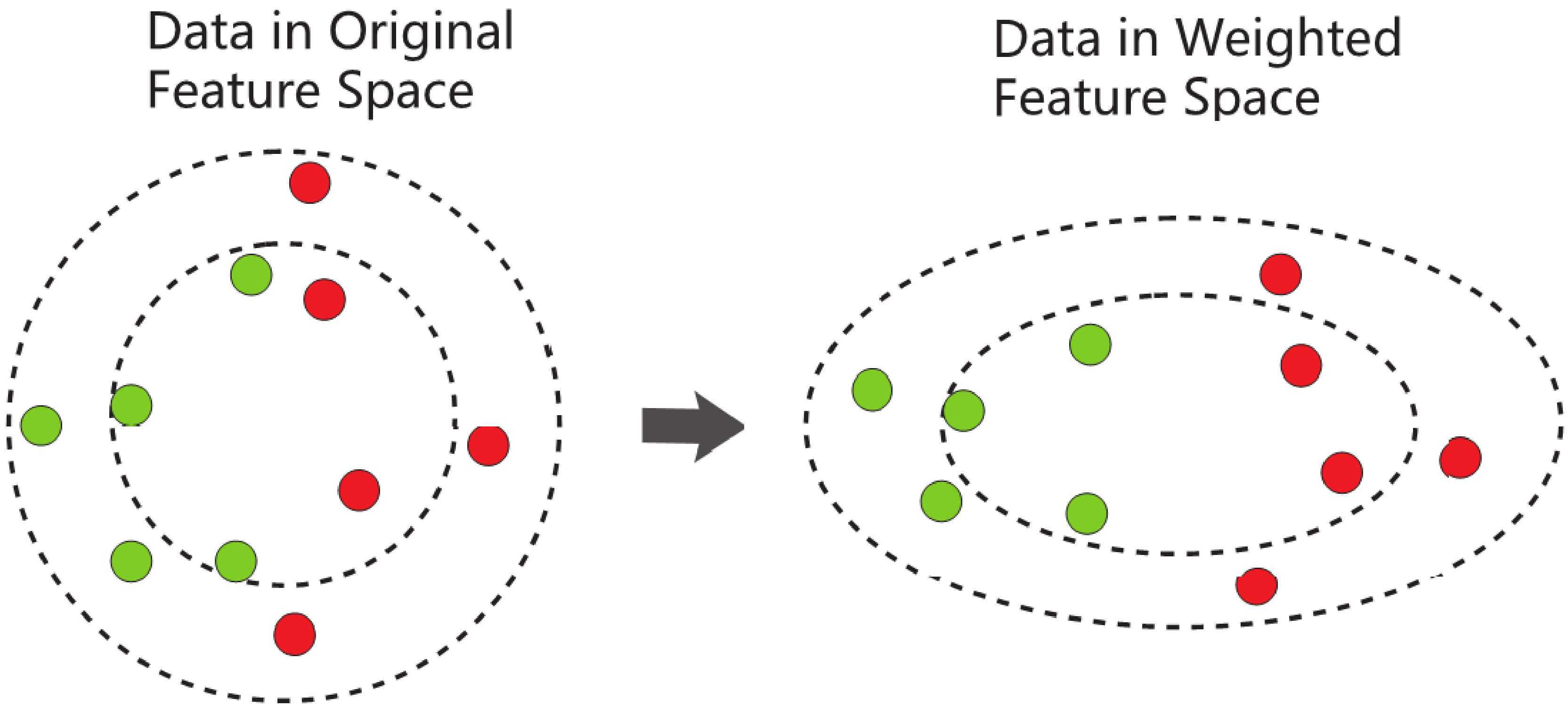


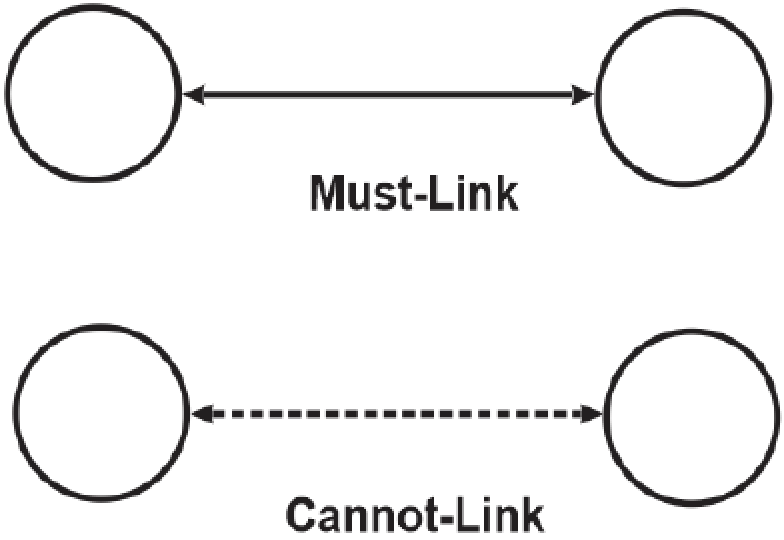
Illustration by Toy Example



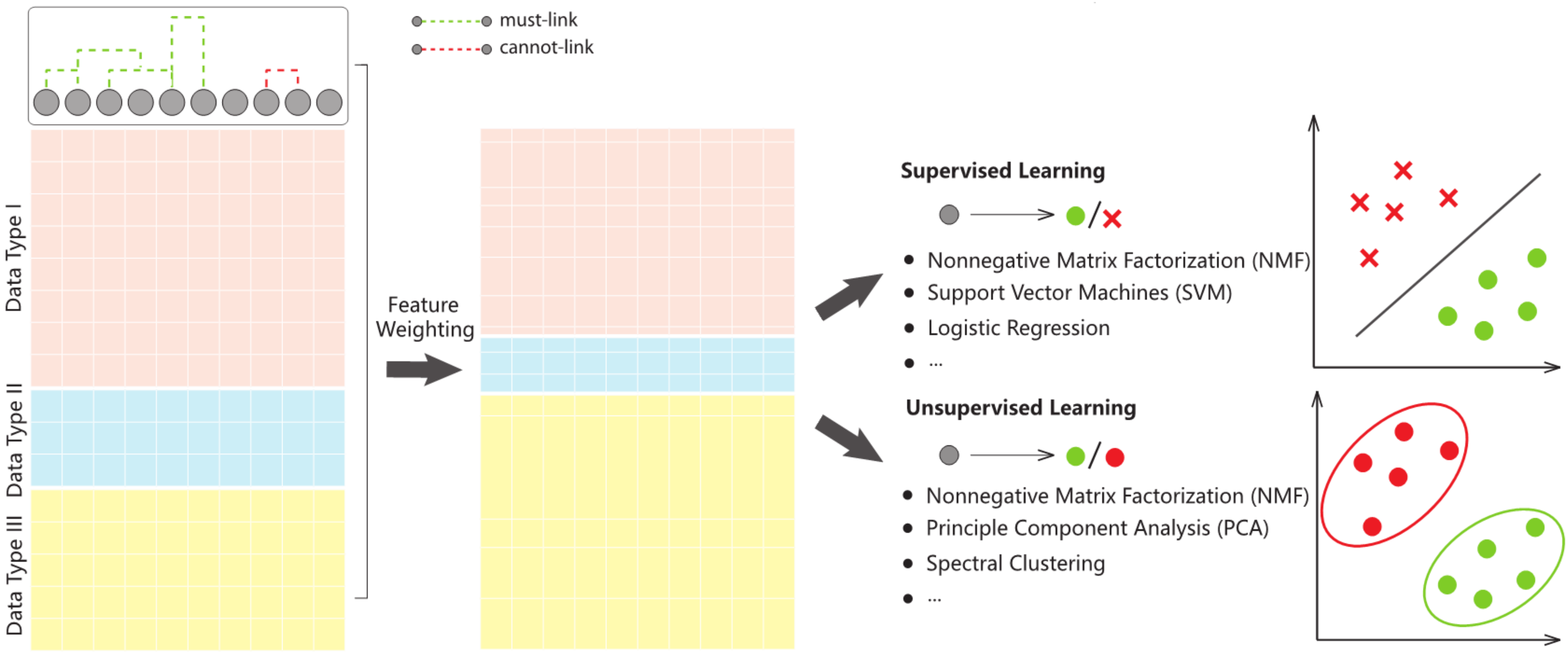
Data in Original Feature Space

Data in Weighted Feature Space

Auxiliary Knowledge Format



Workflow



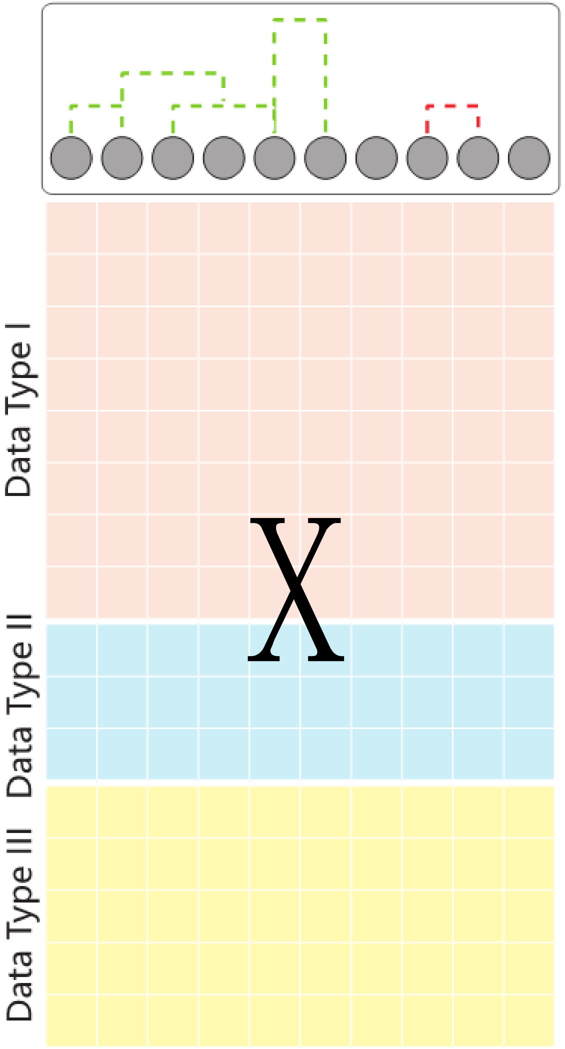
Supervised Learning

- Nonnegative Matrix Factorization (NMF)
- Support Vector Machines (SVM)
- Logistic Regression
- ...

Unsupervised Learning

- Nonnegative Matrix Factorization (NMF)
- Principle Component Analysis (PCA)
- Spectral Clustering
- ...

Problem Formulation



$$X = [X_1; X_2; \cdots; X_m]$$

must-link set $\mathcal{M}$

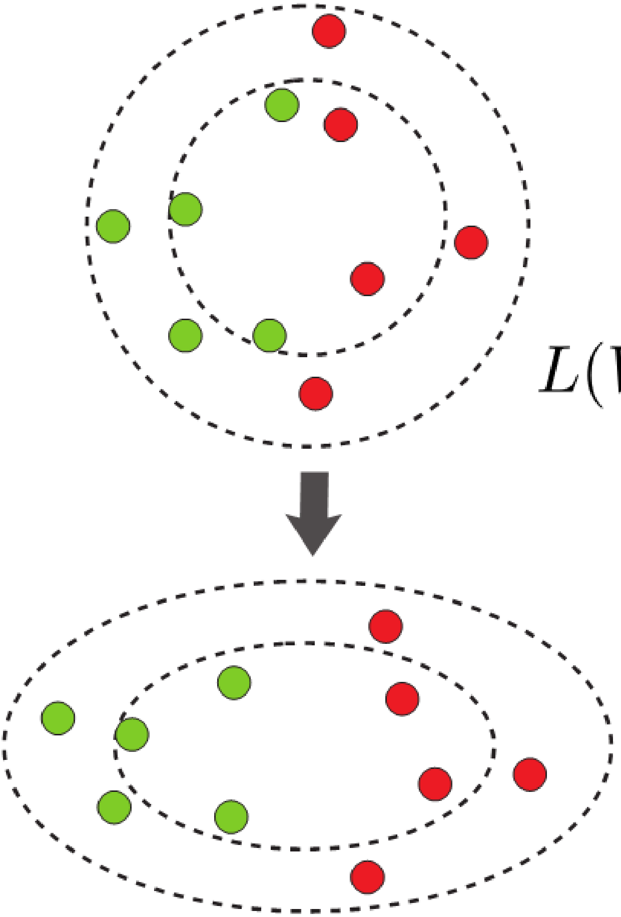
cannot-link set $\mathcal{C}$

inconsistency between X and $\mathcal{M}$

$$\sum_{i,j} A^{\mathcal{M}}_{ij} \|X_{.i} - X_{.j}\|^2 = tr(XL^{\mathcal{M}}X^T)$$

$A^{\mathcal{M}}$ where $A^{\mathcal{M}}_{ij} = 1$ for $(i, j) \in \mathcal{M}$

Feature Weighting mitigate inconsistency



$$L(W) = \sum_{i,j} A^{\mathcal{M}}_{ij} \|diag(W)X_{\cdot i} - diag(W)X_{\cdot j}\|^2$$

$$= tr(diag(W)XL^{\mathcal{M}}X^T diag(W))$$

$$W \text{ satisfy} \left[\begin{array}{l} \text{nonnegativity} \quad W_i \geq 0 \\ \text{conservation,} \ \sum_i W_i = p. \end{array}\right.$$

Homogeneity Assumption

- 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 Selection:
  - Majority of features are useless (weight=0)
  - Only small amount of features are important (weight=1)

- Let $\Delta W = W - 1$
  satisfying $\sum_i \Delta W_i = 0 \text{ and } \Delta W_i \geq -1$

Minimize the Objective Function

$$L(\Delta W) = tr(diag(1+\Delta W) X L^{\mathcal{M}} X^T diag(1+\Delta W)) + \lambda \left\|\Delta W\right\|^2$$

$$= \|Z + Zdiag(\Delta W)\|_F^2 + \lambda \left\|\Delta W\right\|^2$$

where $L^{\mathcal{M}} = UU^T$

$Z = U^T X^T$

$\lambda > 0$

Automatic Coefficient Selection

Iterate until convergence:

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

$$\widehat{\sigma} \leftarrow \frac{1}{\sqrt{p}} \left\| Z + Z diag(\widehat{\Delta W}) \right\|_F$$

Implementation Tricks

Solve the Equivalent Quadratic Programming:

$$\begin{aligned} L(\Delta W) &= \sum_i Y_i(\Delta W_i + 1)^2 + \lambda \Delta W_i^2 \\ &= \Delta W^T diag(Y + \lambda)\Delta W + 2Y^T \Delta W + const \end{aligned}$$

where $Y_i = (X_{i.} L^{\mathcal{M}}) X_{i.}^T$

Extension 1

- Sparse must-link set
  - under-determined, infinite solution
  - Add a k-nearest neighbor graph as local embedding

$$A = A^{\mathcal{M}} + \gamma A^{\mathcal{X}}$$

where $$A^{\mathcal{X}}_{ij} = \exp\left\{ -\frac{\|X_{\cdot i} - X_{\cdot j}\|^2}{2\sigma^2} \right\}$$

Extension 2

- Both must-link and cannot-link set available

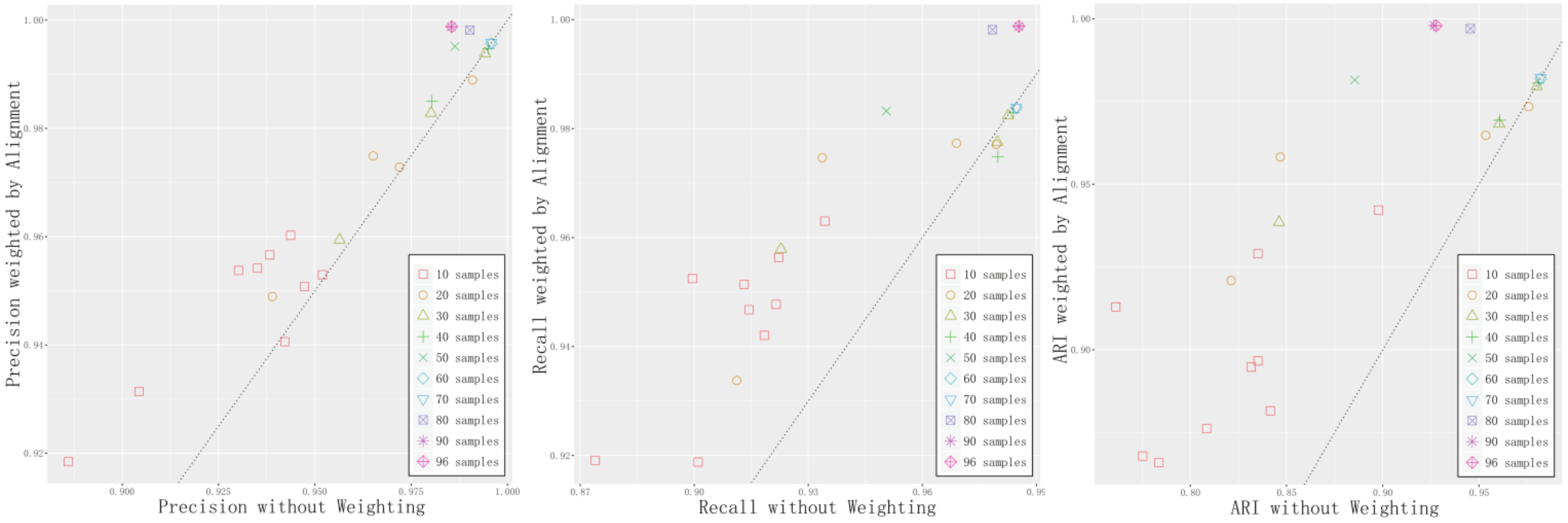
$$
\begin{aligned}
L(W) &= tr(\left[diag(W)XL^{\mathcal{M}}X^{T}diag(W) - \eta diag(W)XL^{\mathcal{C}}X^{T}diag(W)\right]_{+}) \\
&= tr(diag(W)\left[XL^{\mathcal{M}}X^{T} - \eta XL^{\mathcal{C}}X^{T}\right]_{+} diag(W))
\end{aligned}
$$

where $[x]_{+} = \max(0, z)$

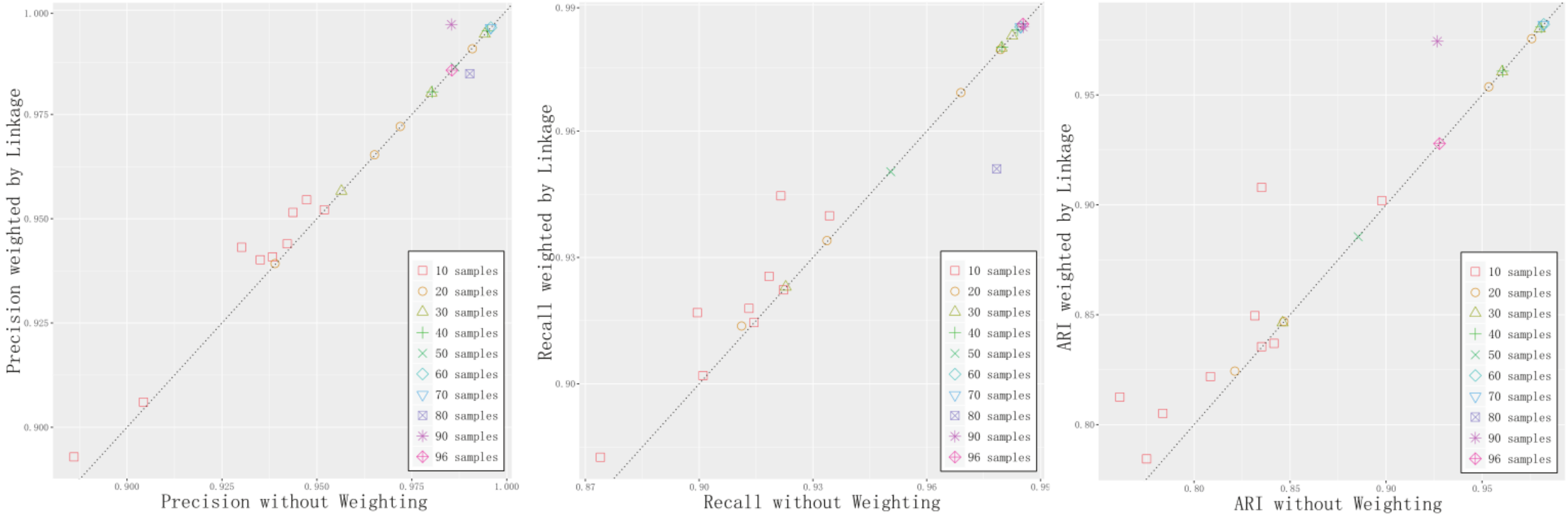
Results

- Metagenomic Contig Binning
  - Features: abundance and composition profiles
  - Must-link: co-alignment and linkage
  - Dataset: simulated “SpeciesMock” dataset and real “MetaHIT” dataset

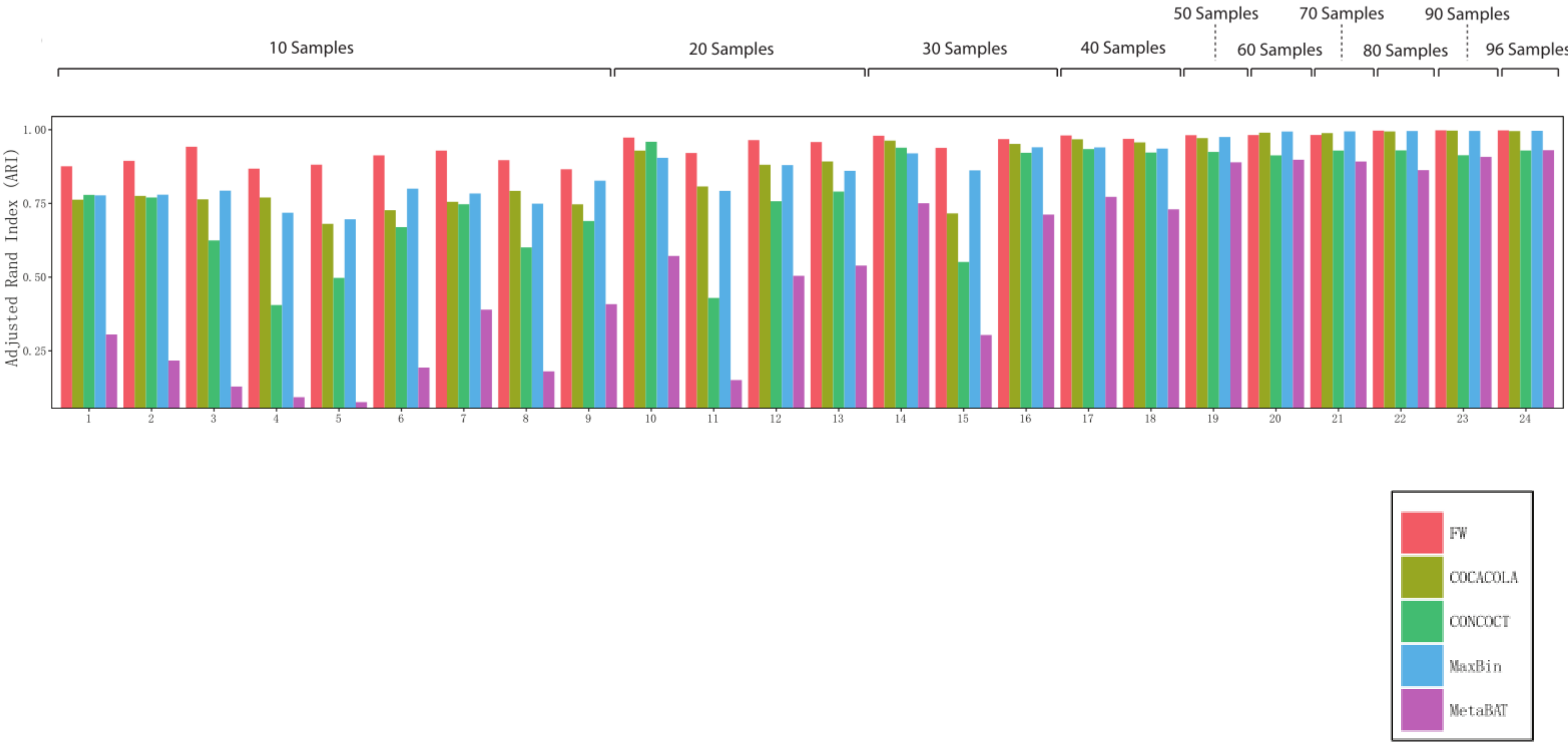
Metagenomic Contig Binning



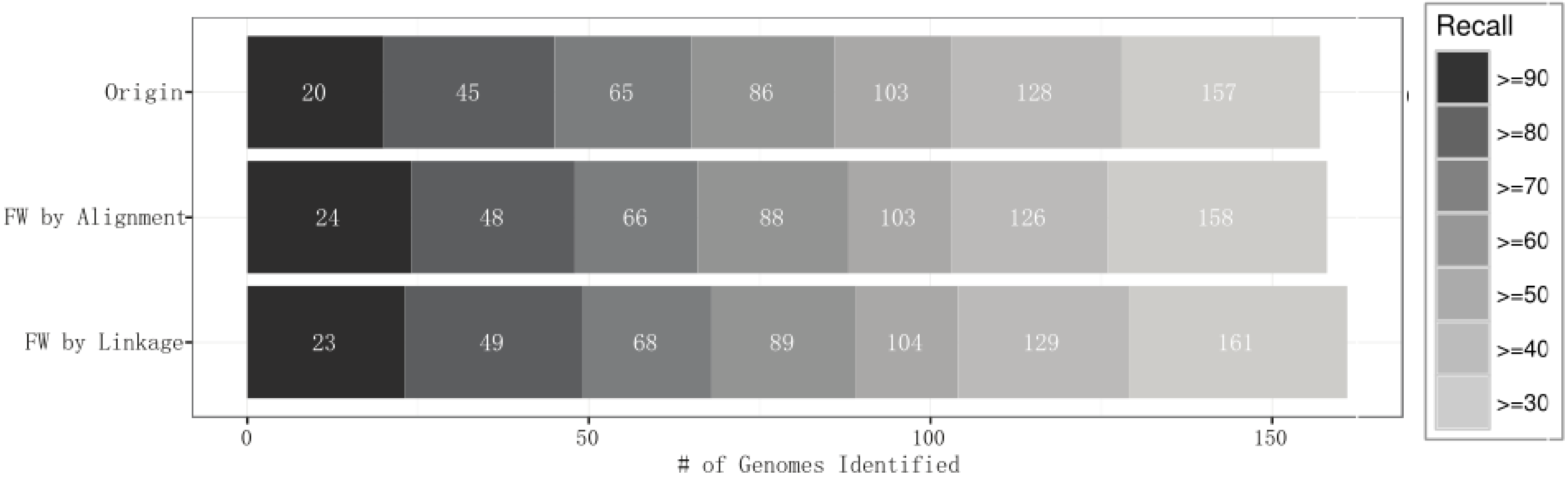
Metagenomic Contig Binning



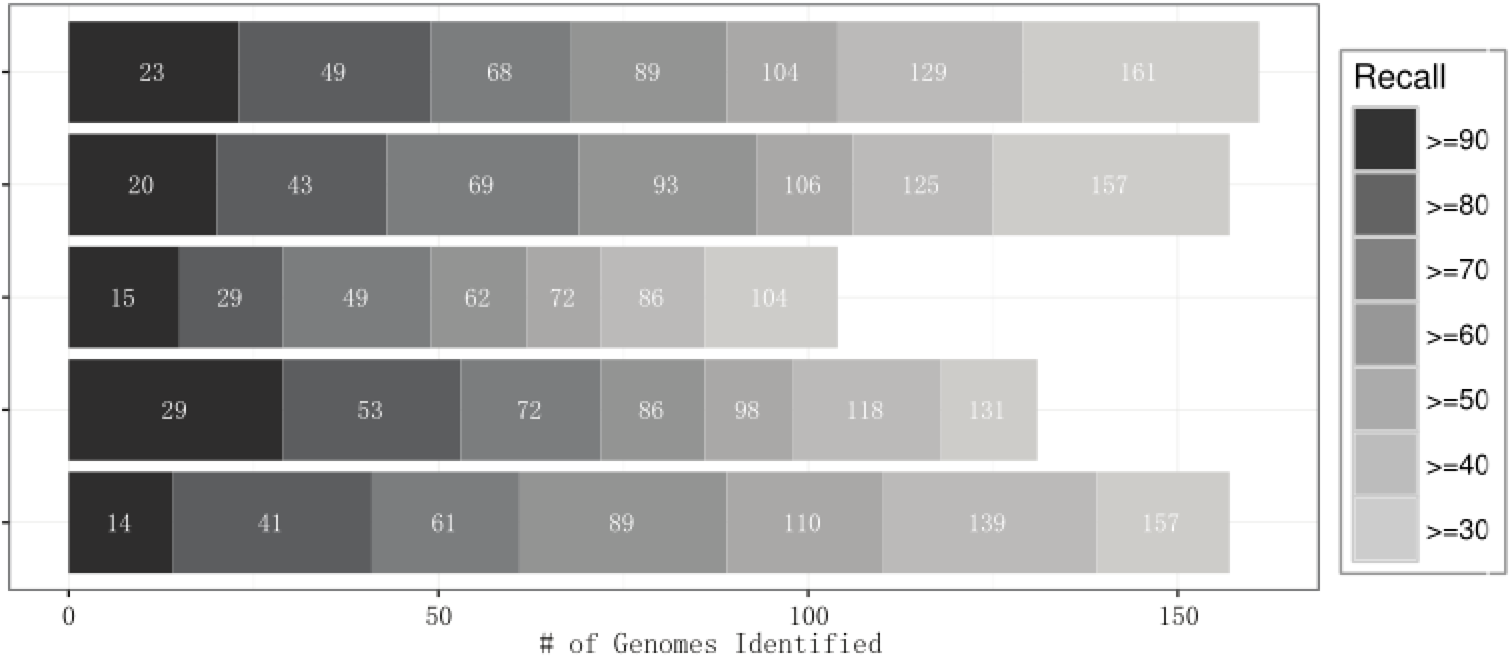
Metagenomic Contig Binning



Metagenomic Contig Binning



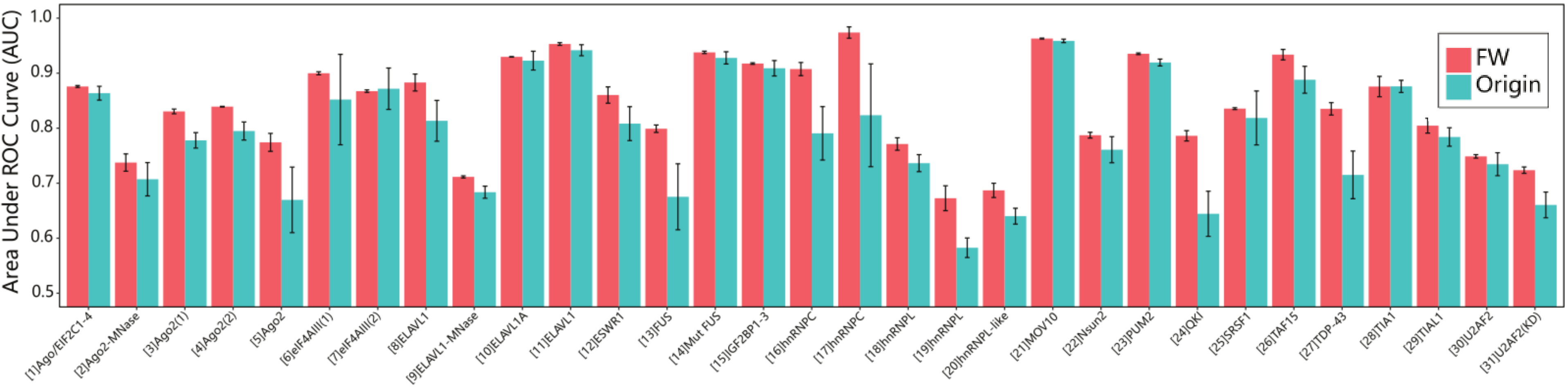
Metagenomic Contig Binning



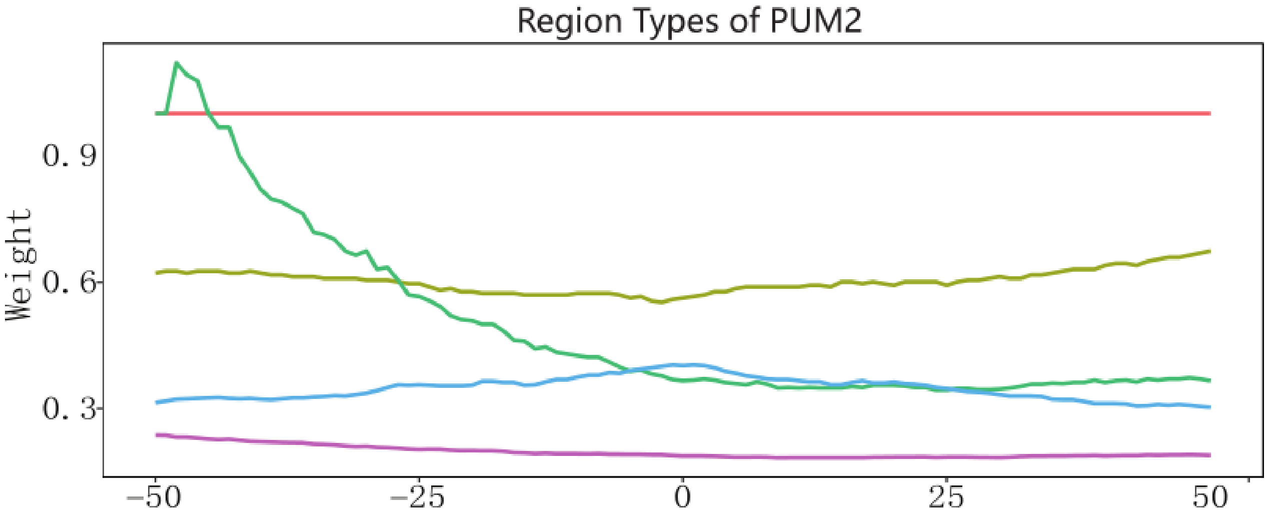
Results

- RBP(RNA binding protein) Binding Site Prediction
  - Features: RNA tetra-mer composition, RNA secondary structure, surrounding region types, co-binding profiles associated with other RBPs and Gene Ontology (GO) terms.
  - Must-link and cannot-link: labels in training set
  - Dataset: 19 distinct RBPs with one or multiple experimental replicates, in 31 published CLIP experiments

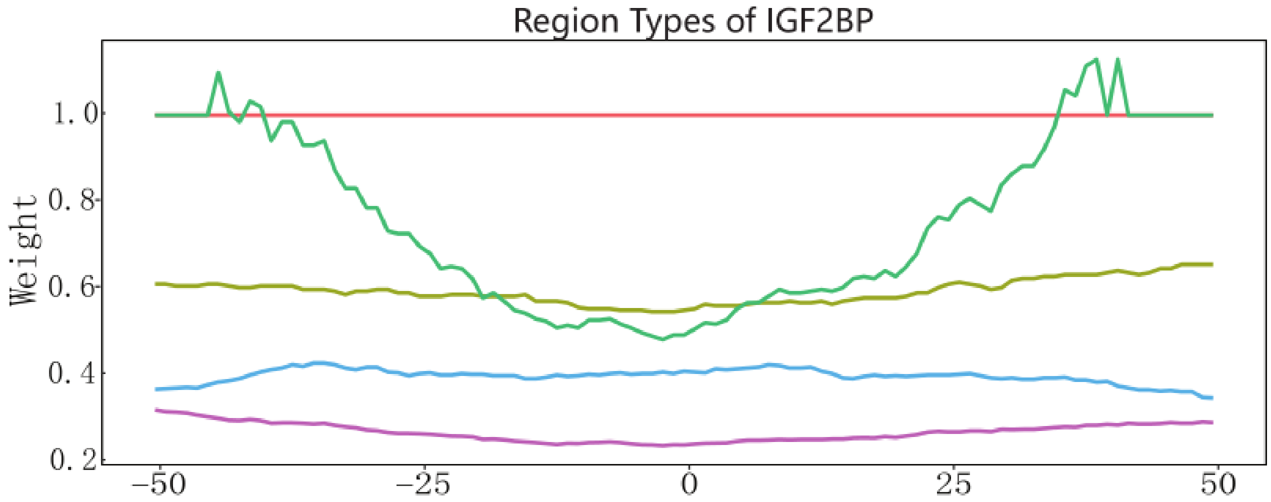
RBP Binding Site Prediction



RBP Binding Site Prediction

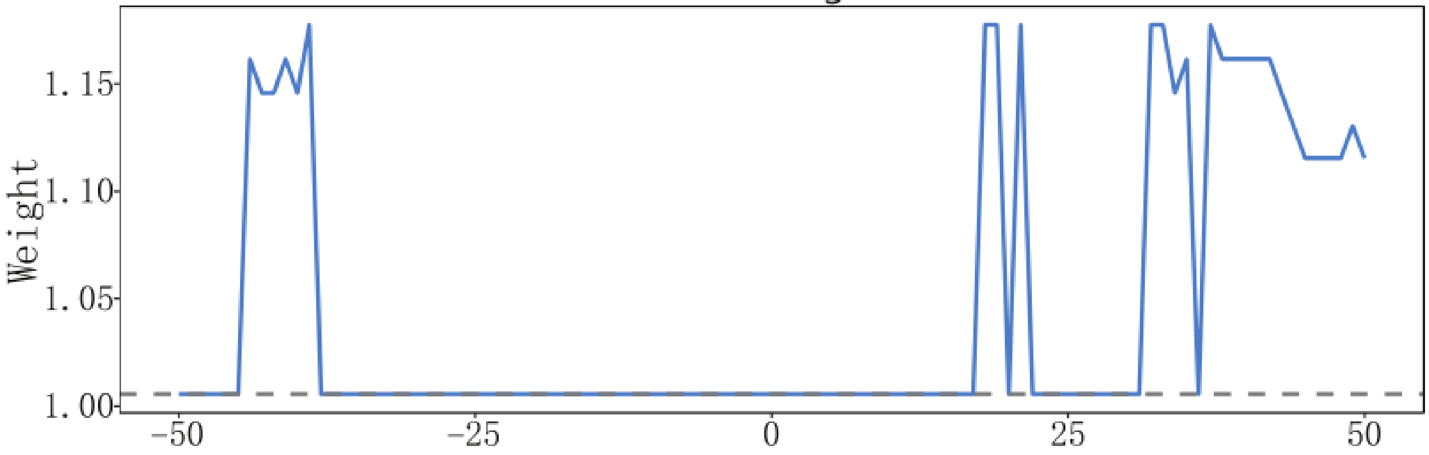


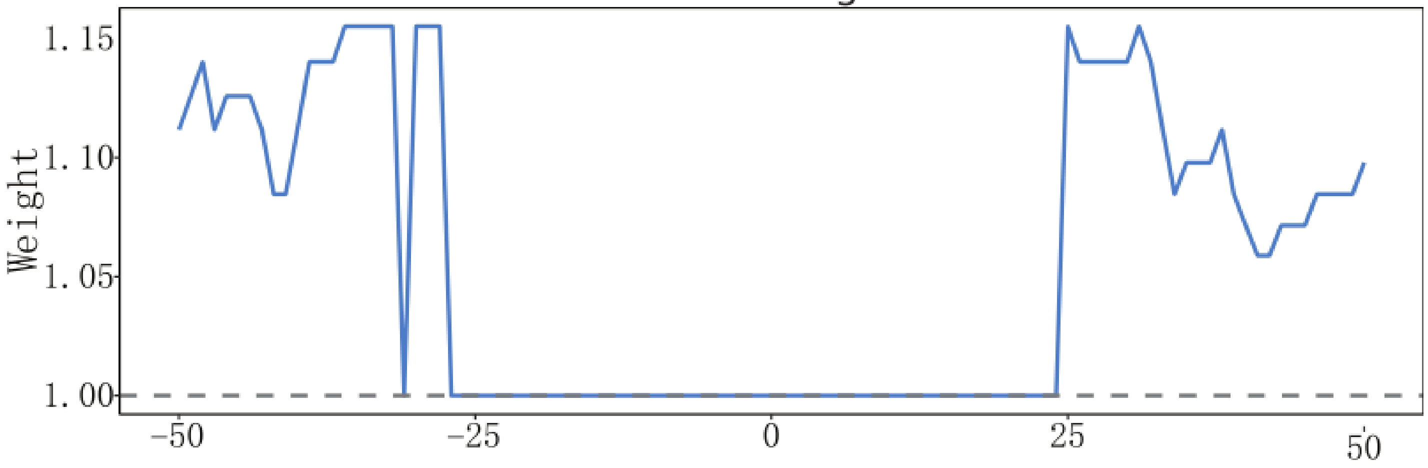
Region Types of PUM2



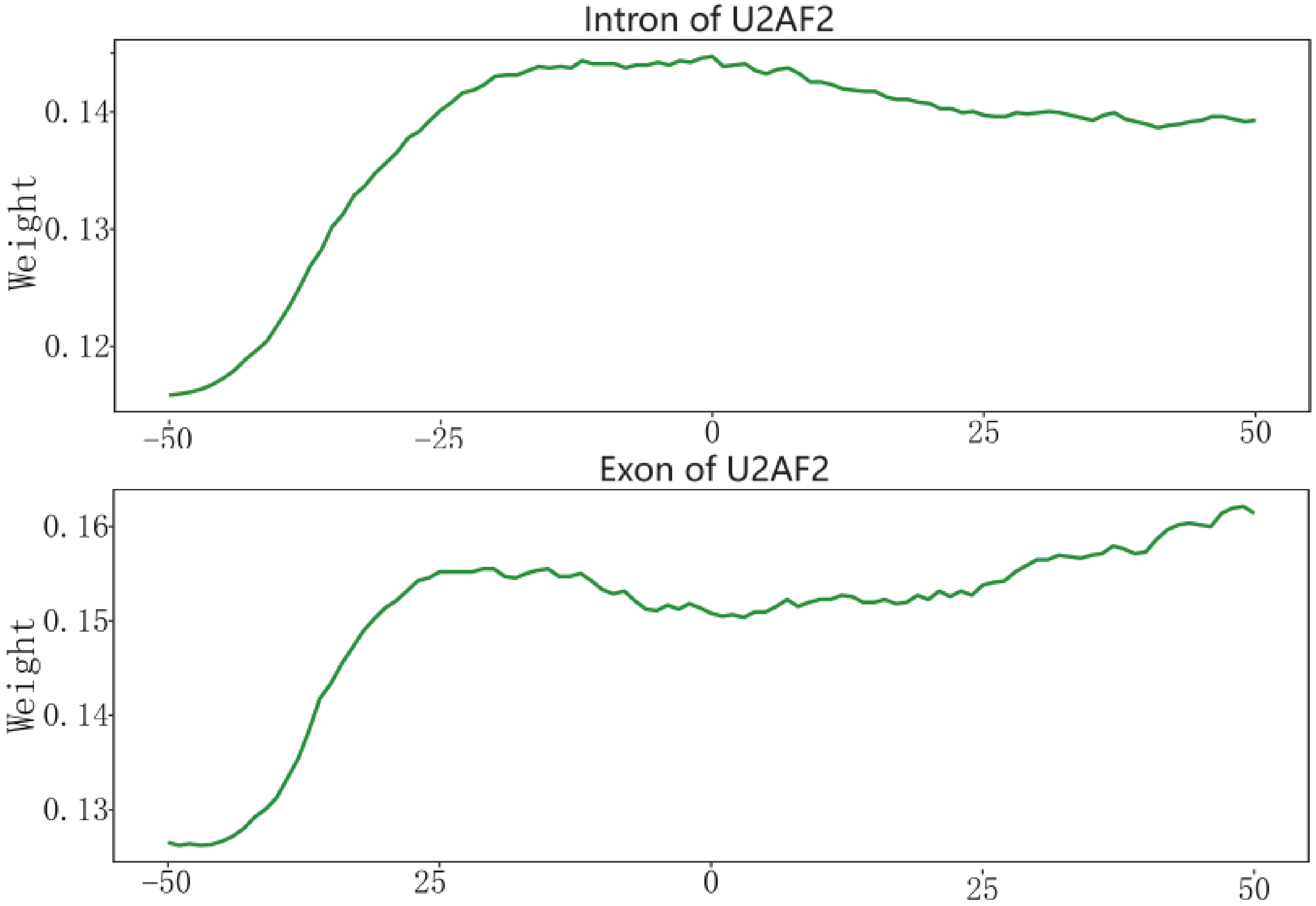
Region Types of IGF2BP

RBP Binding Site Prediction


U2AF2 Cobinding to hnRNPC


hnRNPC Cobinding to U2AF2

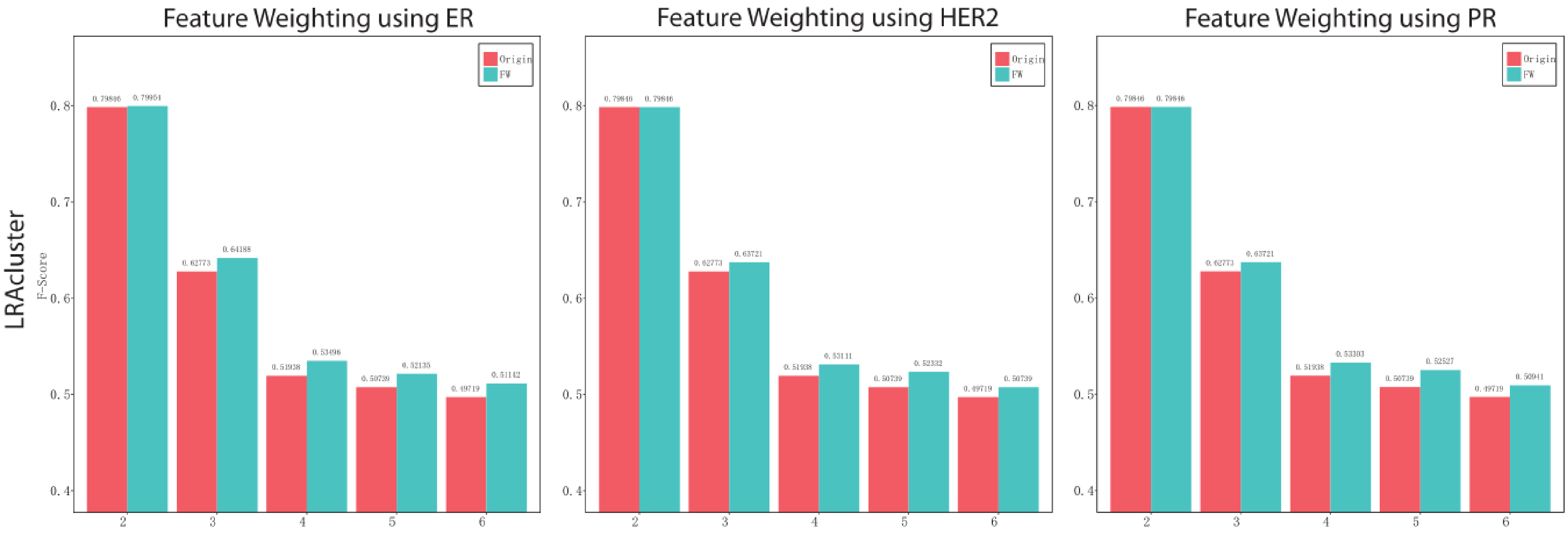
RBP Binding Site Prediction



Results

- Cancer subtyping
  - Features: gene expression, DNA methylation, copy number variation, somatic mutation.
  - Must-link : surface receptors ER/HER2/PR status
  - Dataset: breast cancer from TCGA

Cancer subtyping



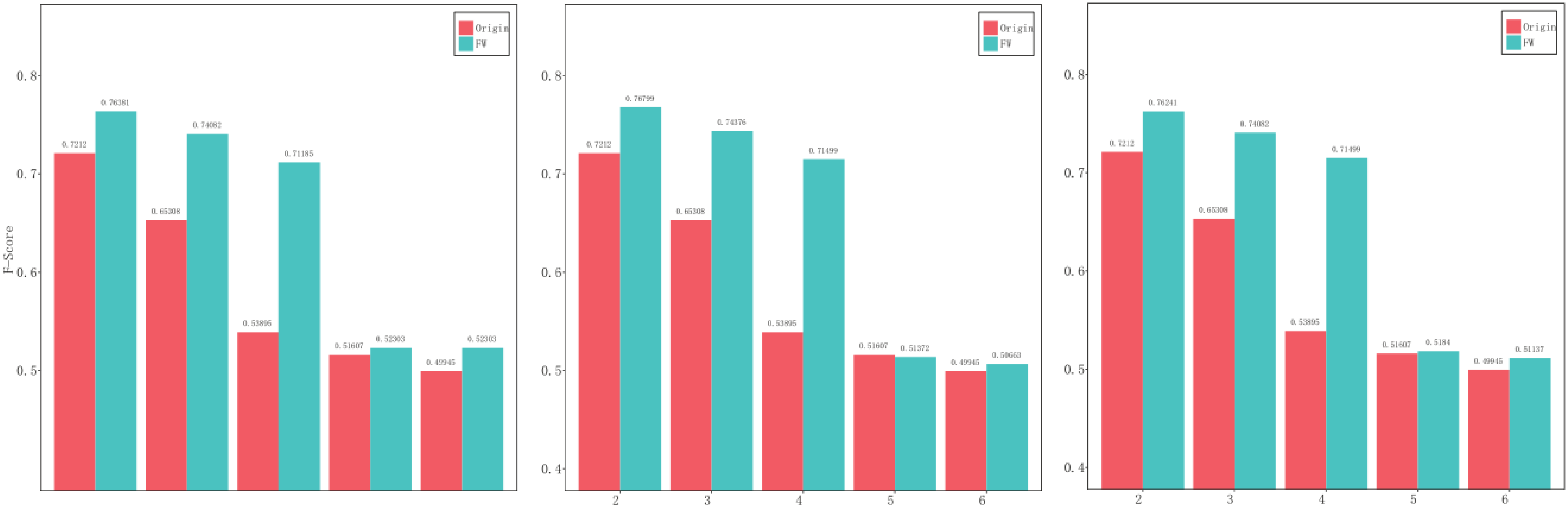
Feature Weighting using ER

Feature Weighting using HER2

Feature Weighting using PR

Cancer subtyping

Network-based Stratification (NBS)



Future Direction

- Deal with kernel matrix
- Deal with more general auxiliary knowledge
  - Relative comparison
  - Weighted kmer distance
- Deal with iterative weighting and screening

Questions?

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