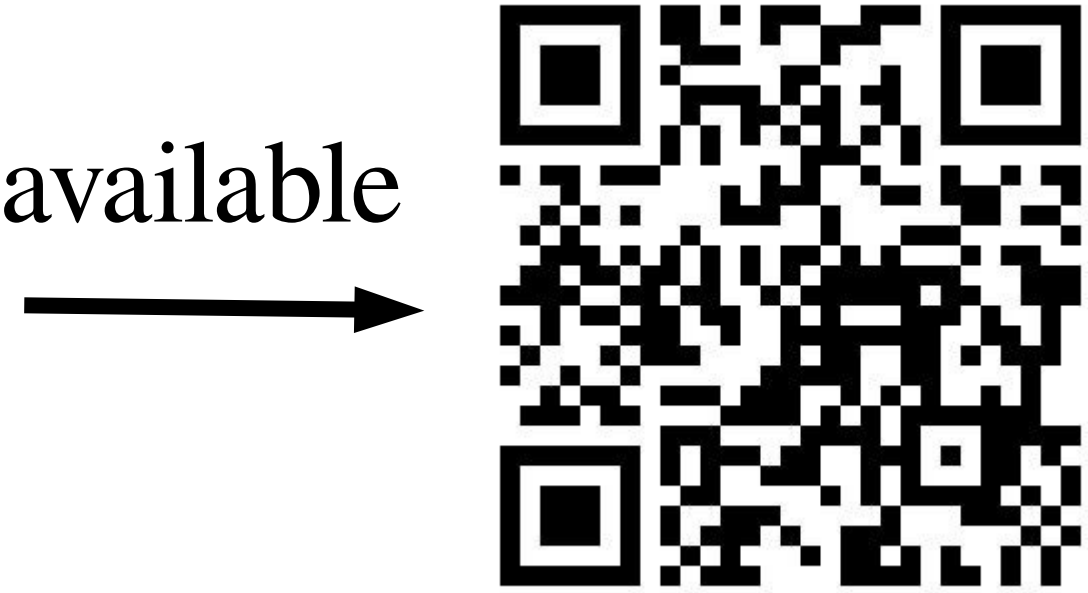
Application of tensor decomposition based unsupervised feature extraction to multi-omics data set

Y-h. Taguchi

Department of Physics, Chuo University,

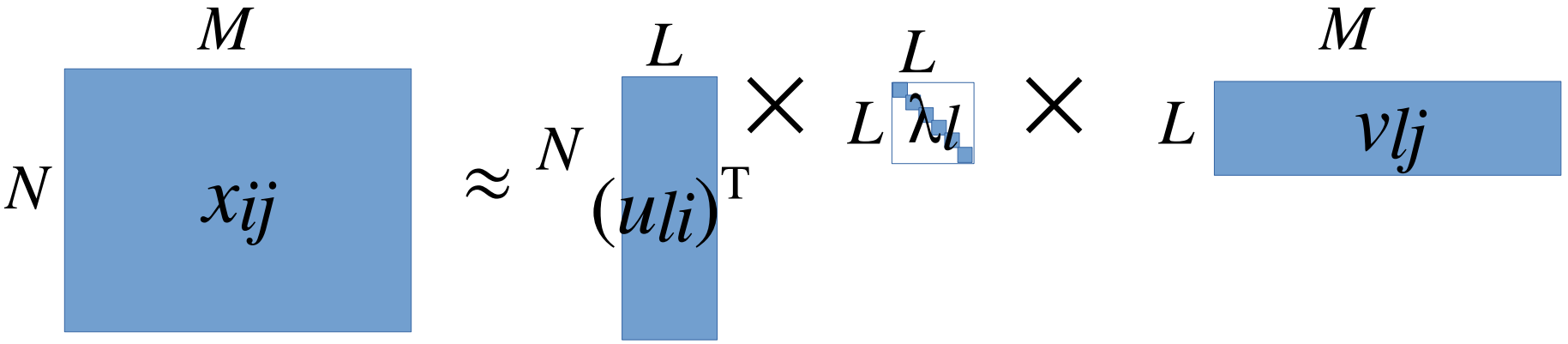
Tokyo 112-8551, Japan.

This presentation is available



This travel is supported by general citizens,
人見琢也 Takuya Hitomi
望月正弘 Masahiro Mochizuki
尾上辰徳 Tatsunori Onoe
堀内卓也 Takuya Horiuchi
那須川進一 Shinichi Nasukawa
片岡　昇 Noboru Kataoka
上原久幸 Hisayuki Uehara
阿部真也 Shinya Abe
増田一也 Kazuya MASUDA
藤根和穂 Dr. Kazuho Fujine
力丸 健太郎Kentaro Rikimaru,
through cloud founding managed by Academist
https://academist-cf.com/projects/122

Singular value decomposition



$$x_{ij} \simeq \sum_{l=1}^{L} u_{li}\,\lambda_l\,v_{lj}$$

Example

x_{ij}: gene expression

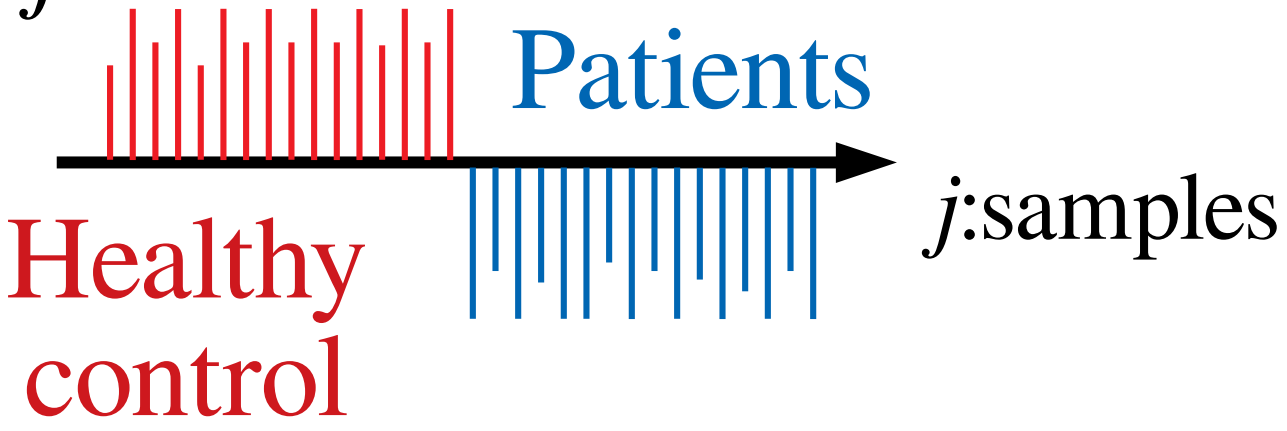
N: number of genes (i)

M: number of samples (j)

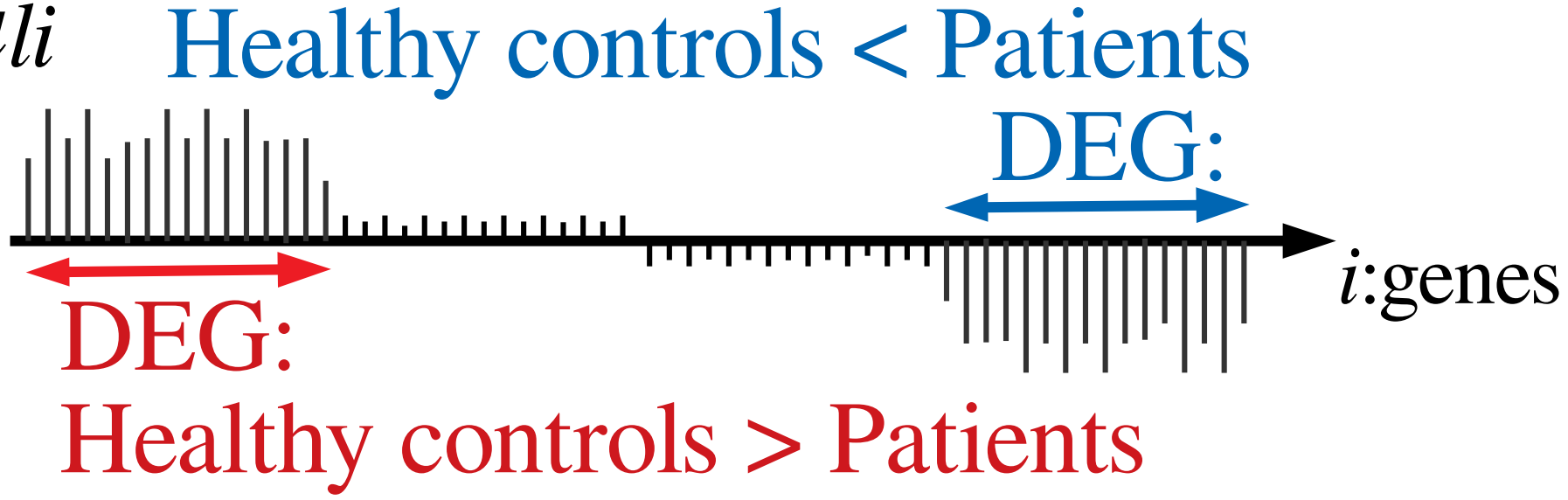
Interpretation…..

For some specific l

v_{lj}



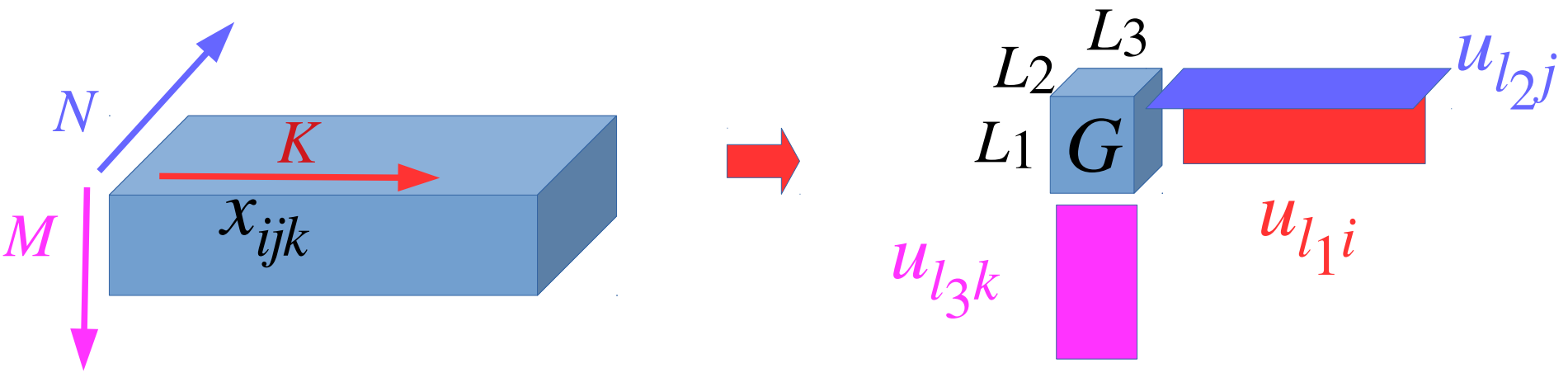
u_{li}



DEG: Differentially Expressed Genes

Extension to tensor…..

HOSVD (Higher Order Singular Value Decomposition)



$$x_{ijk} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} G(l_1 l_2 l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

Example

x_{ijk}: gene expression

N: number of genes (i)

M: number of samples (j)

K: number of tissues (k)

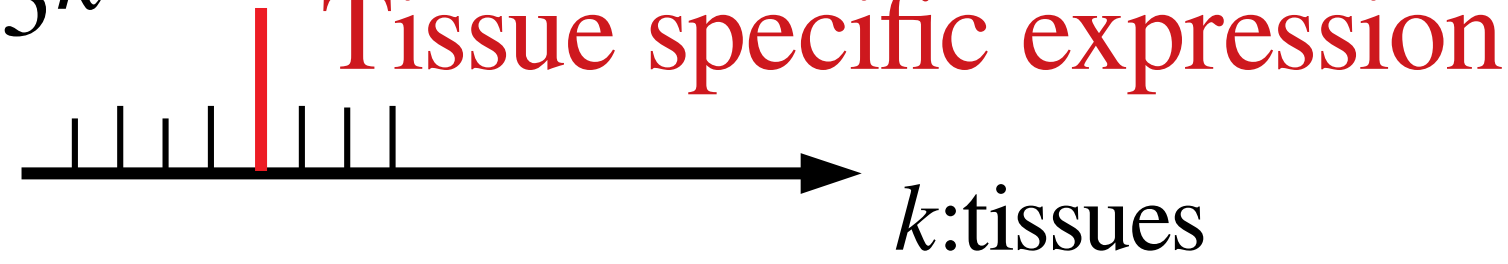
Interpretation…..

For some specific l_2

$u_{l_2 j}$

For some specific l_3

$u_{l_3 k}$

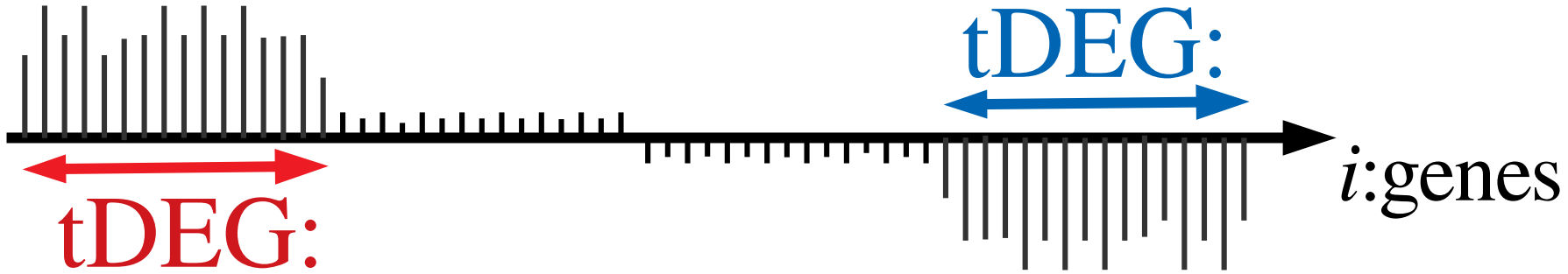


For some specific l_1 with max $|G(l_1 l_2 l_3)|$

Fixed

If $G(l_1 l_2 l_3)>0$

$ul_1 i$ Healthy controls < Patients



Healthy controls > Patients

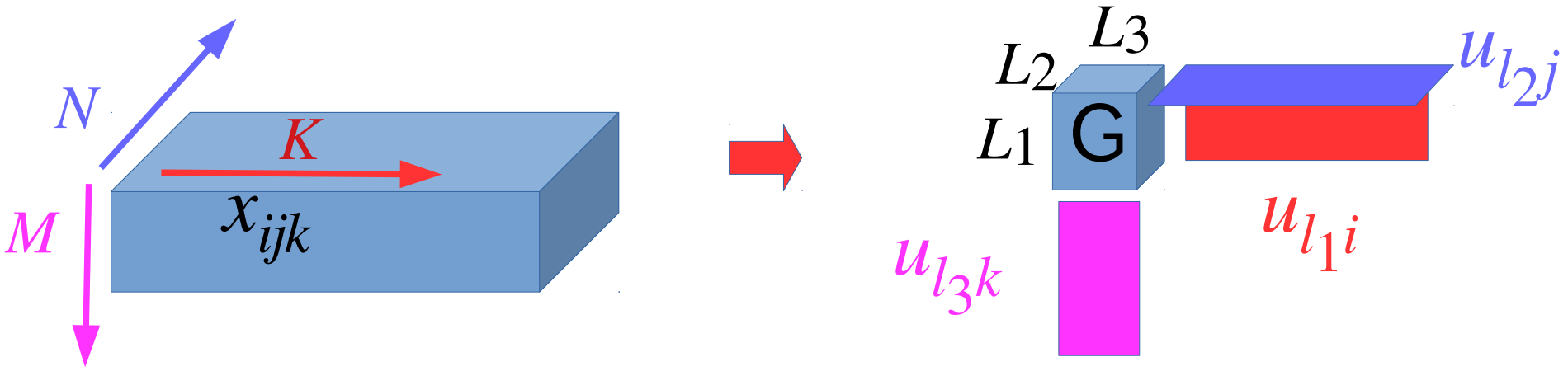
tDEG:
tissue specific Differentially Expressed Genes

Integrated analysis of multiple matrices and/or tensors

x_{ij} : expression of gene i of sample j

x_{kj}: methylaion of region k of sample j

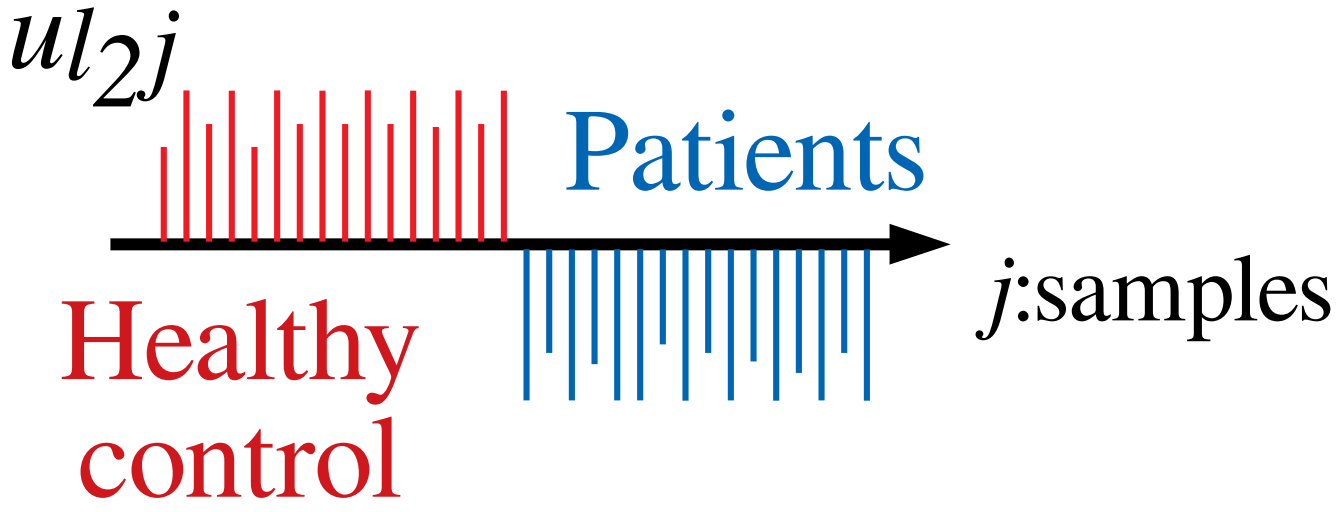
$$x_{ijk} \equiv x_{ij} \times x_{kj}$$



$$x_{ijk} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} G(l_1 l_2 l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

Interpretation…..

For some specific l_2

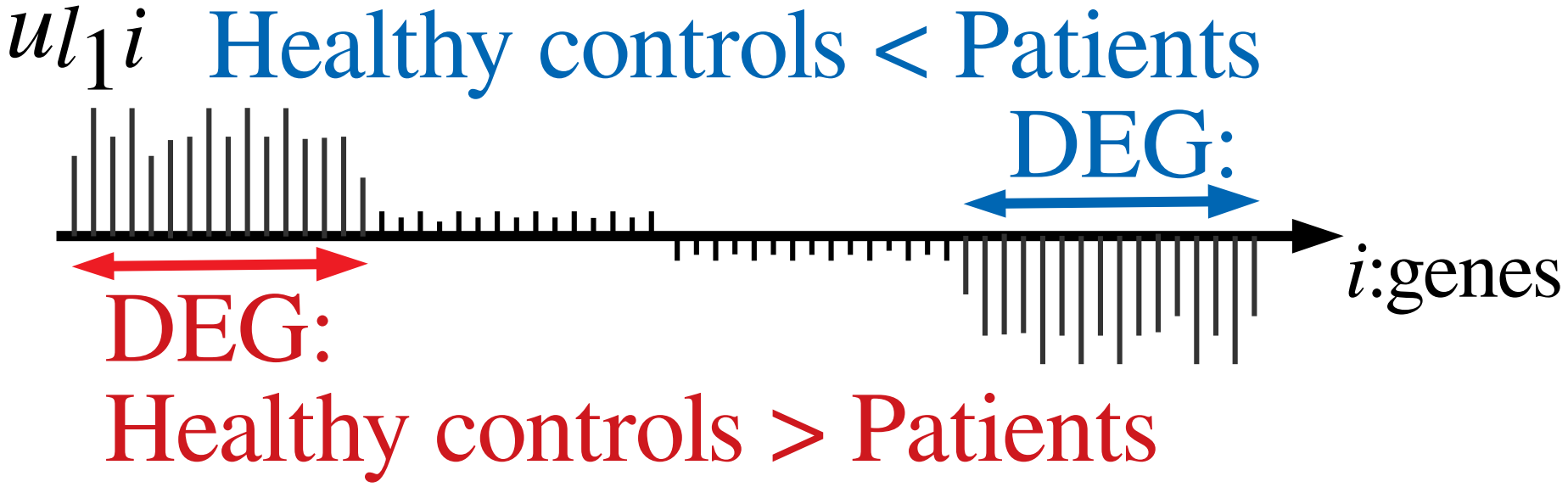


For some specific l_1, l_3 with max $|G(l_1 l_2 l_3)|$

Fixed

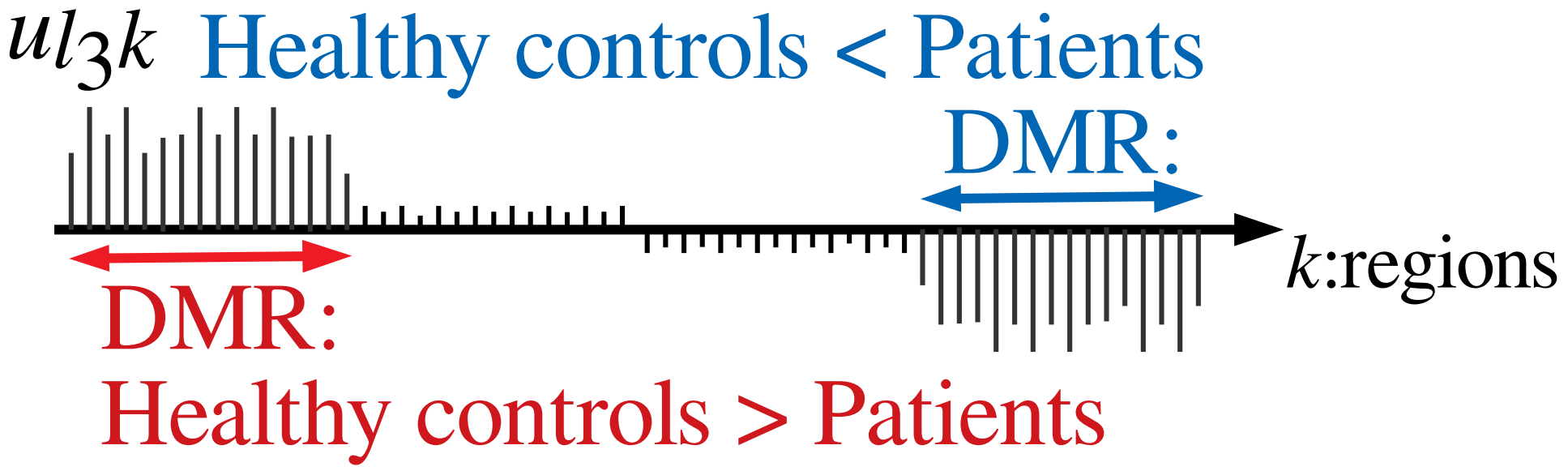
If $G(l_1 l_2 l_3)>0$

For gene expression



DEG:
Differentially Expressed Genes

For methylation



DMR:
Differentially Methylated Regions

Application example No.1

“Multiomics Data Analysis Using Tensor Decomposition Based Unsupervised Feature Extraction –Comparison with DIABLO–”

Y-h. Taguchi
in De-Shuang Huang Vitoantonio Bevilacqua Prashan Premaratne (Eds.), Intelligent Computing Theories and Application, 15th International Conference, ICIC 2019 Nanchang, China,
August 3–6, 2019 Proceedings, Part I, pp.565-574
https://doi.org/10.1007/978-3-030-26763-6_54
Preprint: https://doi.org/10.1101/591867

Taken from mixOmics package in bioconductor https://bioconductor.org/packages/release/bioc/html/mixOmics.html

```
## $mRNA
## [1] 150 samples × 200 mRNAs
##
## $miRNA
## [1] 150  samples ×184 miRNAs
##
## $proteomics
## [1] 150 samples ×142 proteins
```

Three cell lines

```
## Basal  Her2  LumA
##    45    30    75
```

x_{ij}:expression of ith mRNA of jth sample

x_{kj}:expression of kth miRNA of jth sample

x_{pj}:expression of pth protein of jth sample

tensor : $x_{ikpj}=x_{ij} \cdot x_{kj} \cdot x_{pj}$

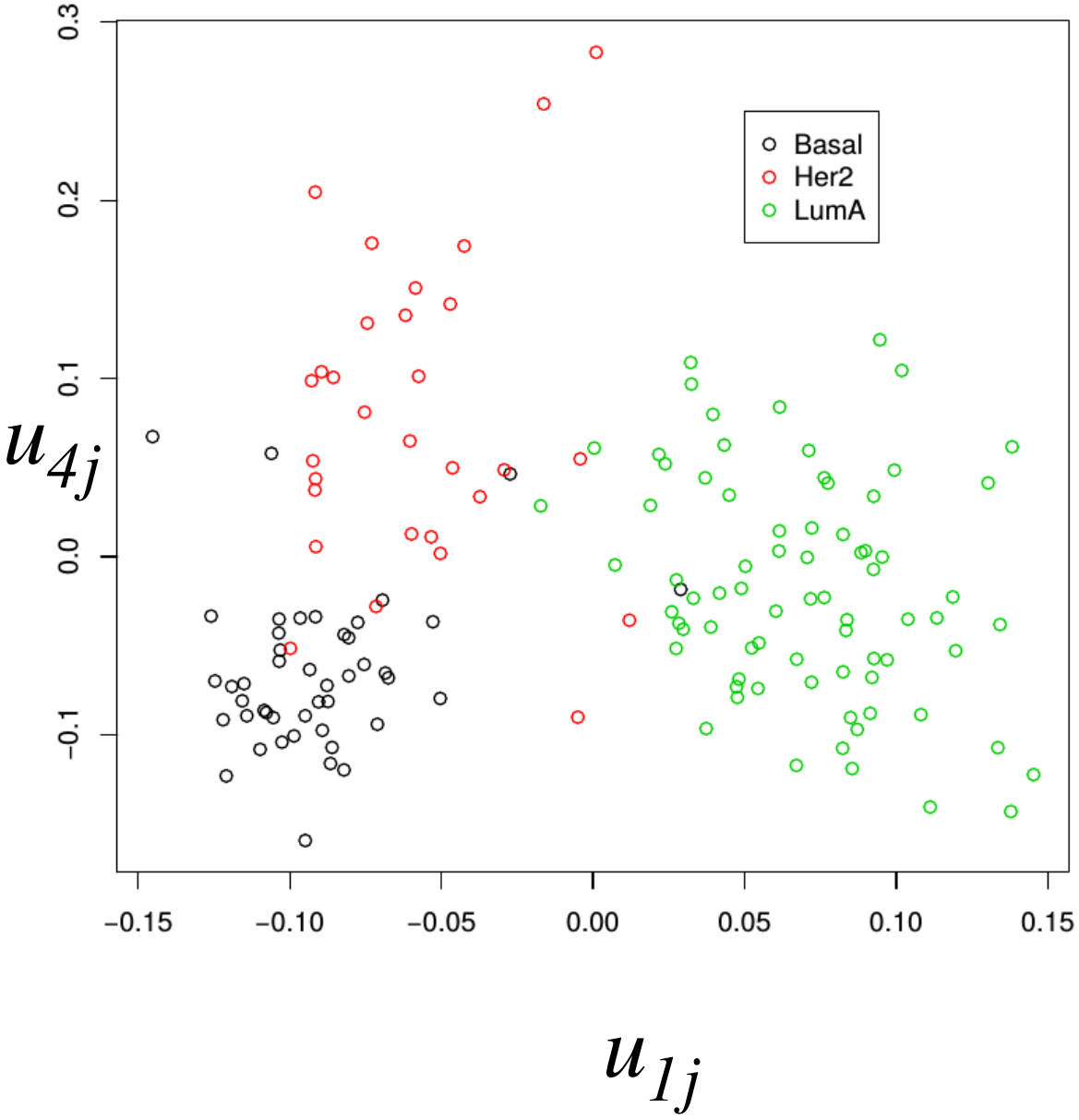
Apply tensor decomposition (tensor version of singular vallue decomposition)

$$x_{ikpj} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} \sum_{l_4=1}^{L_4} G(l_1 l_2 l_3 l_4) u_{l_1 i} u_{l_2 k} u_{l_3 p} u_{l_4 j}$$

ul_1i: mRNA, ul_2k: miRNA

ul_3p: proteome, ul_4j: sample

Linear discriminant analysis
Leave One Out Cross Validation



predict	Real: Basal	Her2	LumA
Basal	**42**	4	0
Her2	2	**25**	2
LumA	1	1	**73**

Error 6.5%

Descending order of $|G(l_1,l_2,l_3,l_4)|$ with l_4=1,4

rank	ℓ_1	ℓ_2	ℓ_3	ℓ_4	$G(\ell_1,\ell_2,\ell_3,\ell_4)$
1	1	1	1	1	-407857.582
2	1	1	4	4	-209720.615
3	2	1	1	4	-20452.480
4	2	1	3	1	-11677.505
5	2	1	4	1	-10428.742
6	2	1	2	1	10157.467
7	1	1	2	1	-8973.774
8	1	2	1	4	8360.976
9	2	1	5	4	-6628.467
10	1	1	3	4	6623.046

$1 \leqq l_1 \leqq 2$, mRNA

$1 \leqq l_2 \leqq 2$, miRNA

$1 \leqq l_3 \leqq 4$, proteome

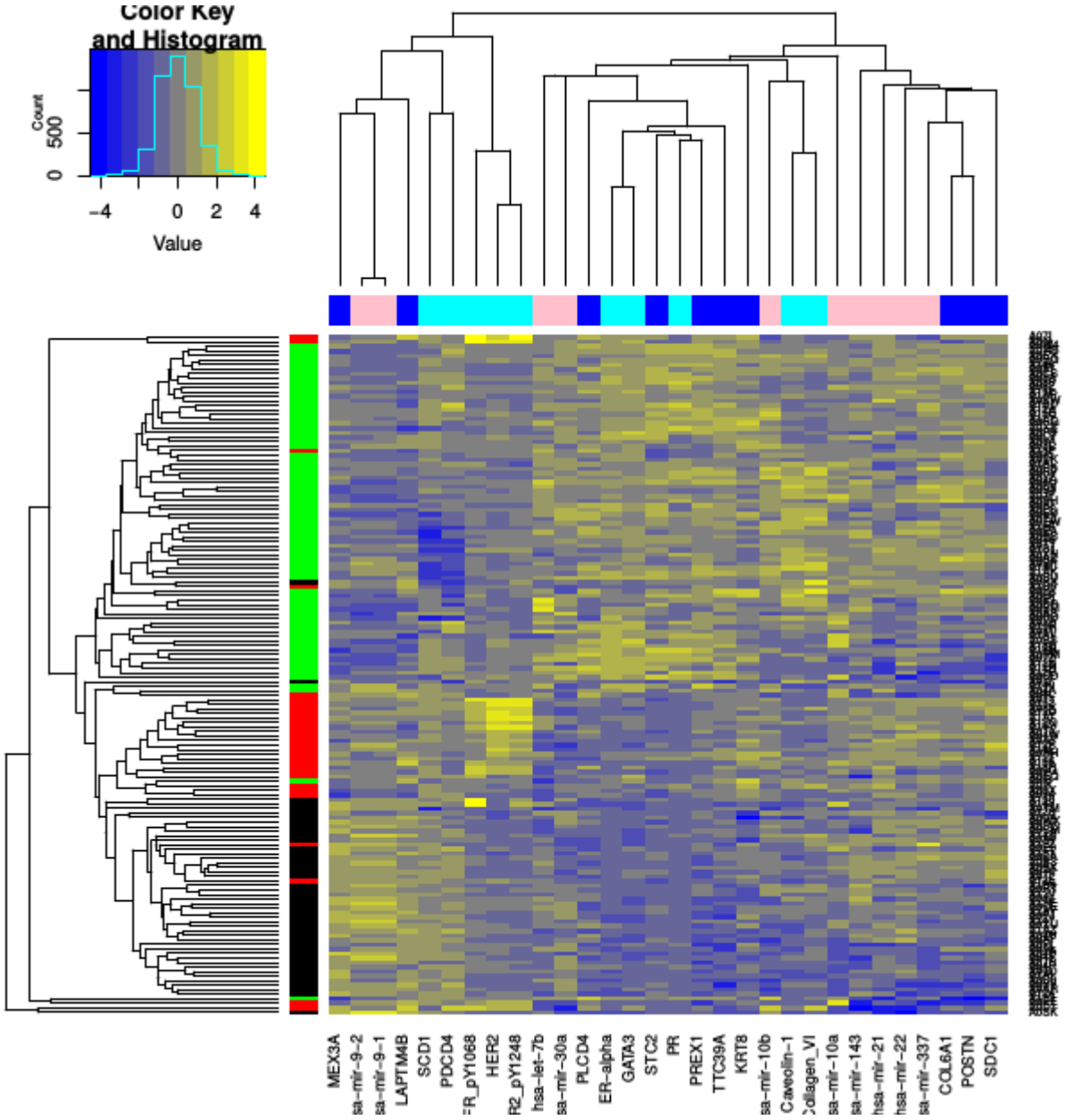
Selecting 10 top ranked mRNAs, miRNAs and proteins based upon squared sum of singular value vectors

$$\sum_{\ell_1=1}^{2} \left(u_{\ell_1 i_1}^{\mathrm{mRNA}}\right)^2$$

$$\sum_{\ell_2=1}^{2} \left(u_{\ell_2 i_2}^{\mathrm{miRNA}}\right)^2$$

$$\sum_{\ell_3=1}^{4} \left(u_{\ell_3 i_3}^{\mathrm{prot}}\right)^2$$

Discrimination performances using selected features

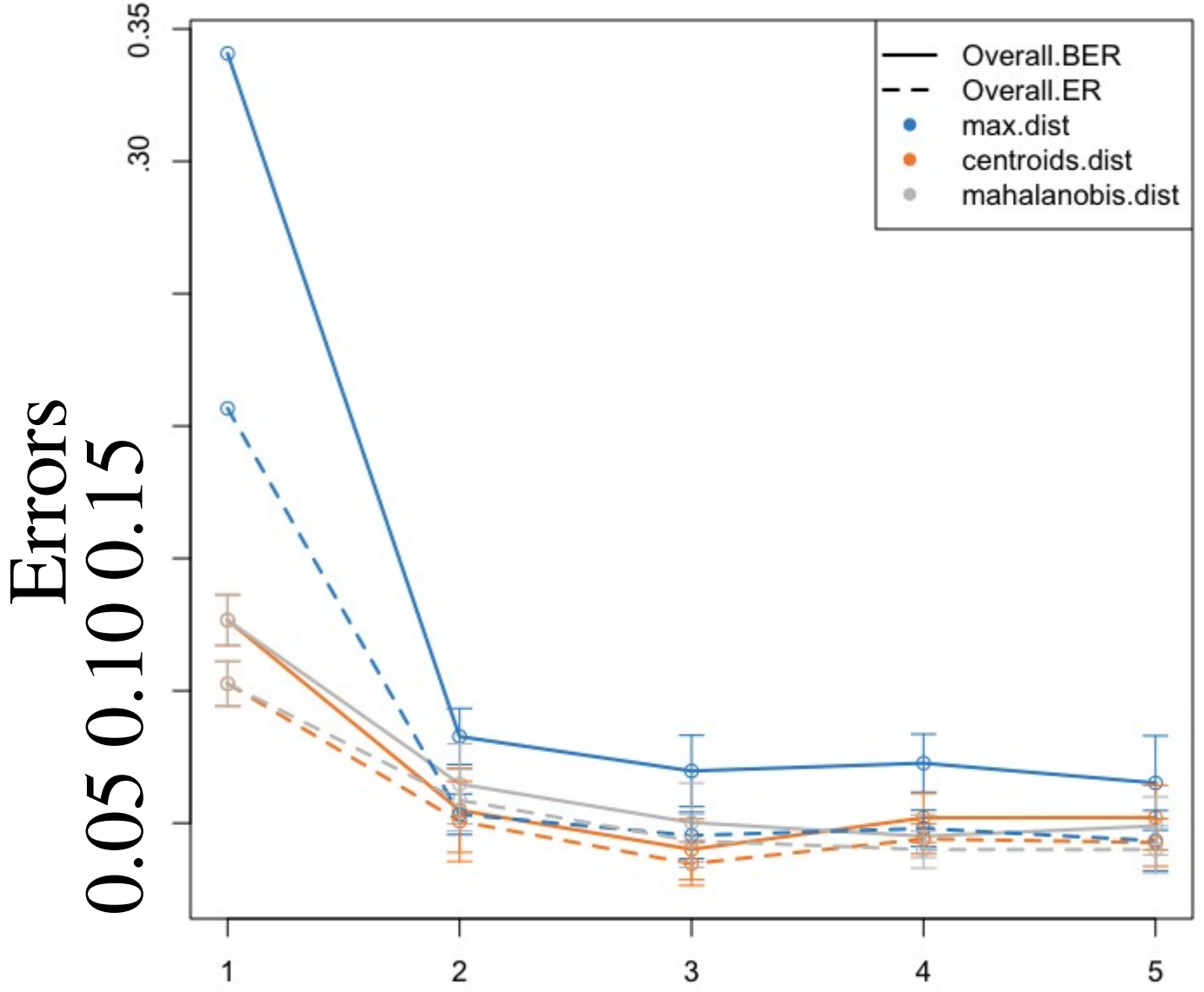


mRNA
miRNA
protein

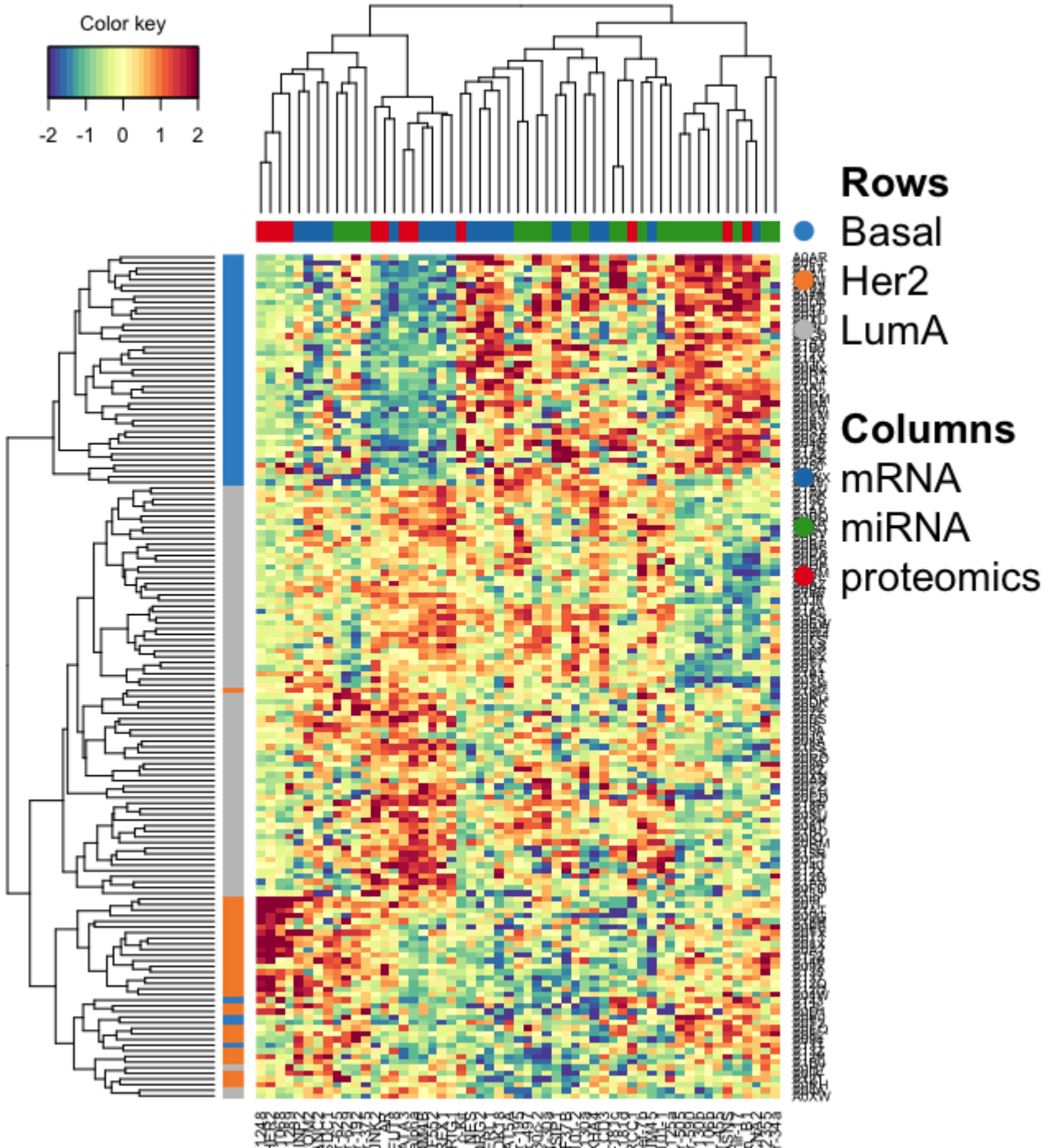
Basal Her2 LumA

Comparisons with DIABLO impremented in mixOmics

Discrimination performances using generated features



Discrimination performances using selected features



Pros and cons of TD based unsupervised FE

Pros:

Fast (because of no optimization)
Robust (independent of label information)
Unsupervised (no need to construct model in advance)

Cons:

No ways if it does not work
Need more memories:
$150 \times (200+184+142)$ vs $150 \times 200 \times 184 \times 142$

Application example No.2

Y-H. Taguchi & Ka-Lok Ng

Tensor Decomposition-based Unsupervised Feature Extraction for Integrated Analysis of TCGA Data on MicroRNA Expression and Promoter Methylation of Genes in Ovarian Cancer

Conf Paper: doi 10.1109/BIBE.2018.00045

Preprint: https://doi.org/10.1101/380071

Biologically, it is unlikely that promoter methylation of protein coding genes and miRNA expression is correlated.

Can our method detect this?

x_{ij}:methylation of ith gene of jth sample
x_{kj}:expression of kth miRNA of jth sample

tensor : $x_{ijk}=x_{ij} \cdot x_{kj}$

$$x_{ijk} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} G(l_1 l_2 l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

$ul_1 i$: gene promoter methylation
$ul_2 j$: samples
$ul_3 k$: miRNA expression

Datasets: Ovarian cancer from TCGA

i: 24906 **protein coding** genes to which promoter methylation is attributed
j: 8 normal vs 569 tumor samples = 577 samples
k: 732 **miRNAs** profiles

Tesnor: $x_{ijk} \in \mathbb{R}^{24906 \times 577 \times 732} \rightarrow$ too huge!

→ approximation (Y-h. Taguchi, PloS ONE, 2017)

$$x_{ik} = \sum_j x_{ijk} \in \mathbb{R}^{24906 \times 732} \rightarrow \text{computable}$$

$$u_{l_2 j}^{\text{miRNA}} = \sum_k u_{l_3 k} x_{kj}$$

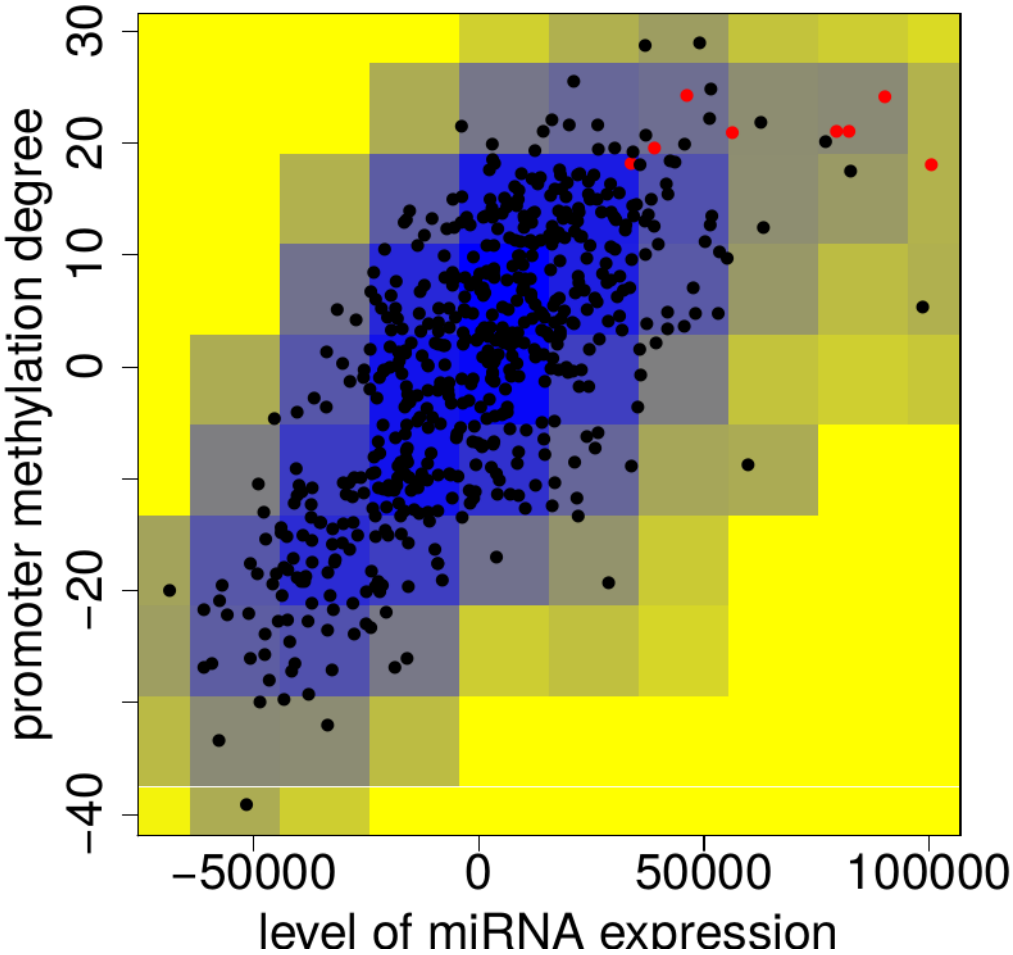
$$u_{l_2 j}^{\text{methyl}} = \sum_i u_{l_1 i} x_{ij}$$

Results

$u_{l_2 j}^{\text{miRNA}}$ and $u_{l_2 j}^{\text{methyl}}$ for $l_2 = 2$ are distinct between 8 normal tissues and 569 tumors.

$\rightarrow u_{l_2 j}^{\text{miRNA}}$ and $u_{l_2 j}^{\text{methyl}}$ are also significantly correlated.

COR=0.72 (P=10^{-9})



P values are computed using chi-squared distribution for $u2i$ and $u2k$ and gene and miRNAs associated with corrected P-values less than 0.01

$$P_i = P\left[> \left(\frac{u_{l_1 i}}{\sigma} \right)^2 \right] \qquad P_k = P\left[> \left(\frac{u_{l_3 k}}{\sigma} \right)^2 \right]$$

→ 7 **miRNAs** are selected using $u_{l_3=2,k}$ and 241 **protein coding** genes are selected using $u_{l_1=2,i}$.

We found that seven **miRNAs** and 241 **protein coding** genes are distinct between normal tissues and tumors.

1681 pairs = 7 **miRNAs** × 241 **protein coding** genes are highly correlated (P<0.01 after BH correction).

		negative correlation	
		T	F
positive correlation	T	0	985
	F	607	95

Most of pairs (94%) are correlated significantly.

Our method can identify promoter methylation of protein coding genes and miRNA expression that satisfy

Distinct between normal controls and tumors as well as mutually correlated between methylation and miRNA expression.

Can other methods do?

Comparisons with conventional methods

Selections of **miRNAs** and **protein coding** genes using t test (normal tissue vs tumors) P<0.01 after BH correction

→ 214 out of 732 **miRNAs** and 19395 out of 24906 **protein coding** genes

→ too many **miRNAs** and **protein coding** genes

Correlation between top 214 miRNAs and 19395 **protein coding** genes

		negative correlation	
		T	F
positive correlation	T	0	329896
	F	225495	3595139

Only 6% pairs are significantly correlated.

Correlation between top 7 **miRNAs** and top 241 **protein coding** genes by t test

		negative correlation	
		T	F
positive correlation	T	0	13
	F	28	1646

Poorer correlation than those selected by TD based unsupervised FE

Conversely, we might be able to first select pairs of **miRNAs** and **protein coding** genes with significantly correlation ($P<0.01$ after BH correction) and select those distinct between normal tissues and tumors….

		negative correlation	
		T	F
positive	T	0	608989
correlation	F	588783	16809266

Only 10% pairs are significantly correlated. Thus, limited number of pairs are selected succesfully. But…..

608989 positively correlated pairs and 588783 negatively correalted pairs include unfortunately all of **miRNAs** and **protein coding genes**…

→ useless for **miRNAs** and **protein coding genes** selection…..

Although we have also evaluated biological significance of seven **miRNAs** selected by TD (using DIANA-mirpath) and 241 **protein coding** genes selected by TD (using MSigDB), no time to report it. Basically, they are highly related to ovarian cancers.

Application example No.3

Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes

Y-h. Taguchi
BMC Bioinformatics volume 19,
Article number: 99 (2018)
Supposed to be presented at APBC2018
https://doi.org/10.1186/s12859-018-2068-7

Phenotype $\not\leftrightarrow$ genotype

Adult vs Child

Male vs female (not human, e.g., fish)

Same genome with distinct phenotype

Social insects with caste

- Ant
- Bee
- Termite

What causes distinction between worker and queen?
→ Epigenetics

http://pestworldforkids.org/pest-guide/bees/

https://www.terminix.com/blog/bug-facts/do-all-ants-bite/

https://www.terminix.com/blog/bug-facts/most-destructive-types-of-termites-and-areas-they-are-found

GEO ID : GSE59525

Gene expression and methylation profiles of

Polistes canadensis and *Dinoponera quadriceps*.

https://bugguide.net/node/view/1478279

https://alchetron.com/Dinoponera-quadriceps

Gene expression

4 Queens vs 6 workers

7 Queens vs 6 workers

Methylation Profiles

1 Control 3 Queens 3 workers

Purpose : Identification of genes associated with aberrant gene expression and methylation profiles between queens and workers simultaneously

Methods

Tensor decomposition based unsupervised feature extraction

x_{ij}:expression of ith gene of jth sample

x_{ik}:methylation of ith gene of kth sample

(methylation integrated over gene body, since it affects gene expression in insects)

tensor : $x_{ijk}=x_{ij} \bullet x_{ik}$

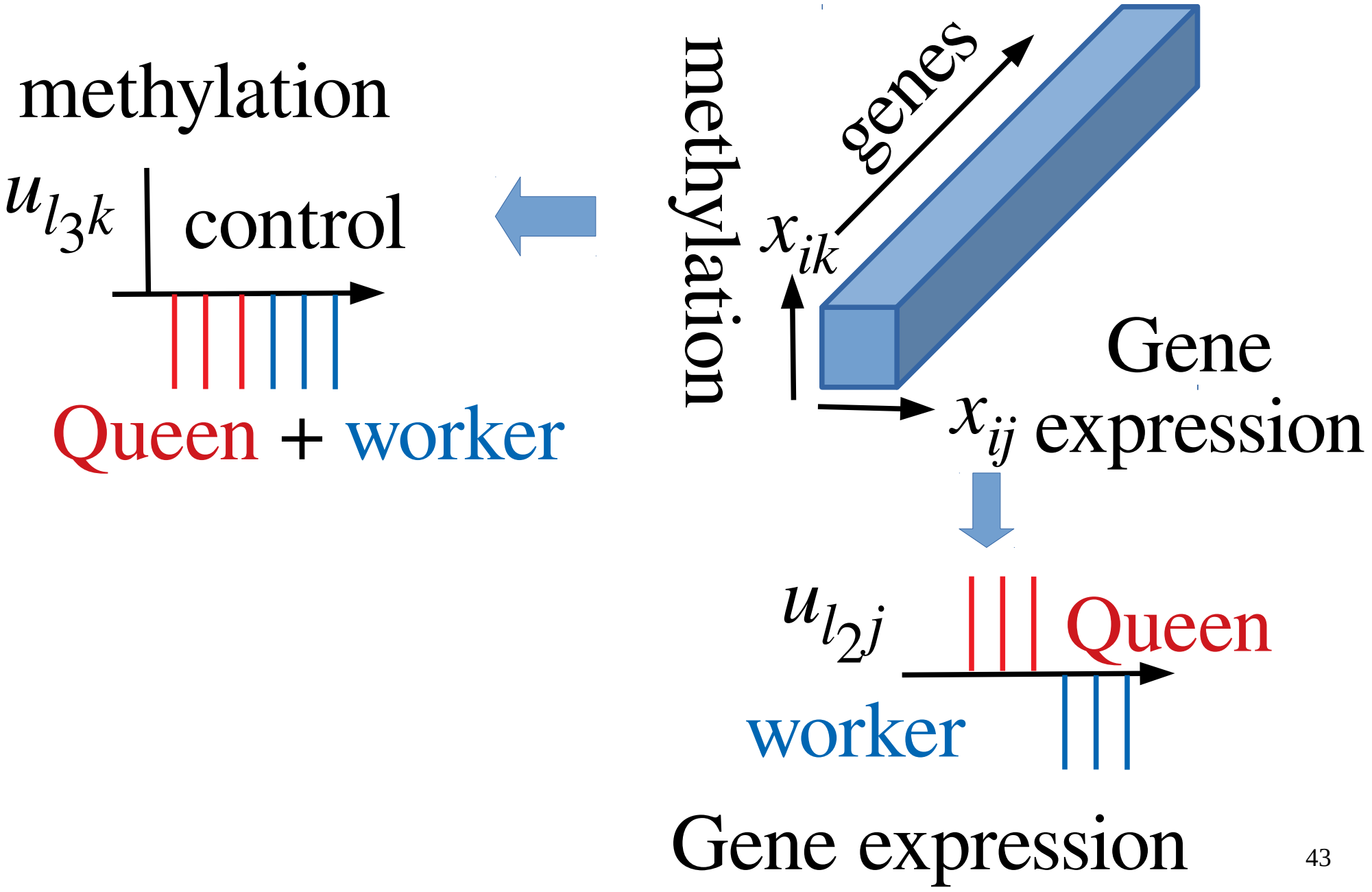
$$x_{ijk} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} G(l_1 l_2 l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

$ul_1 i$: gene

$ul_2 j$: mRNA samples

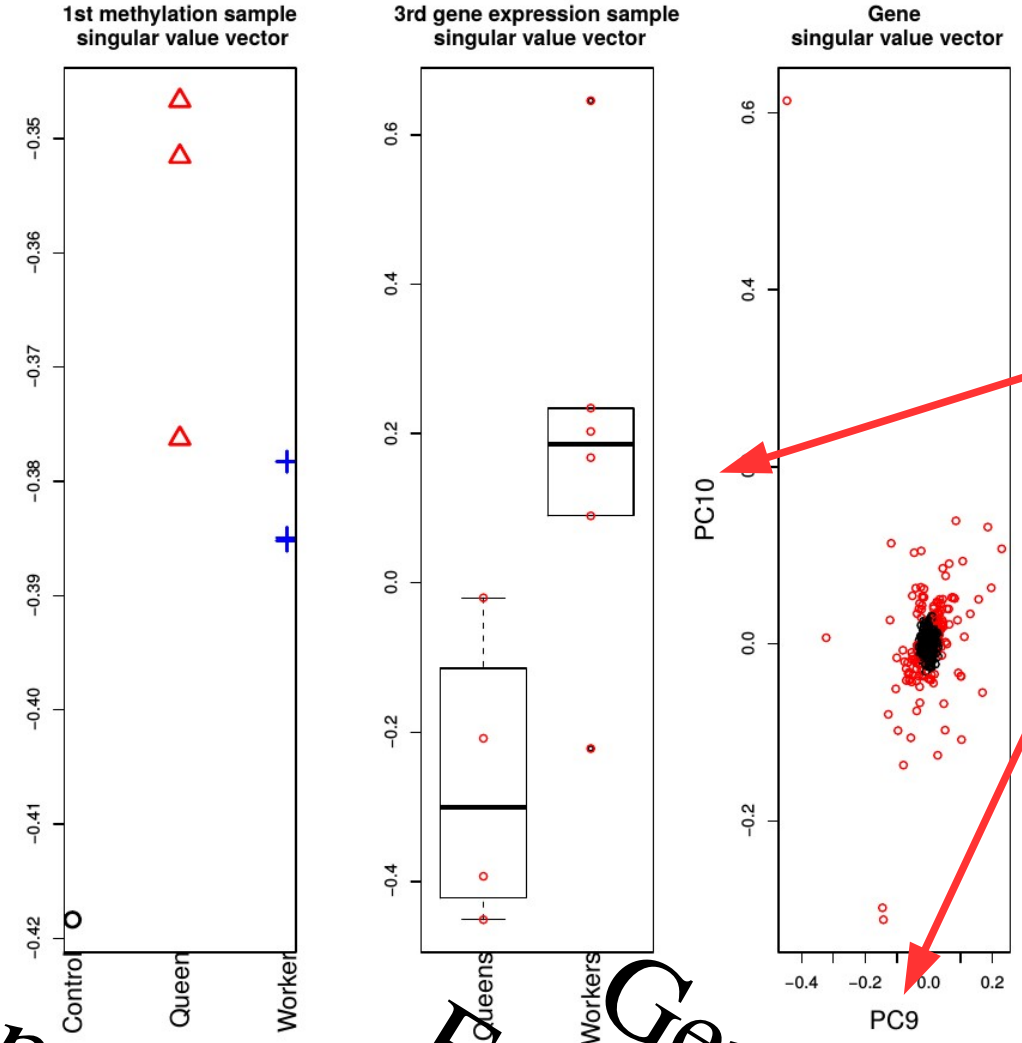
$ul_3 k$: methylation samples

Generating tensor by product $x_{ijk}=x_{ij}\cdot x_{ik}$



P. canadensis

u_{1k} u_{3j}



Gene Expression sample

methylation sample

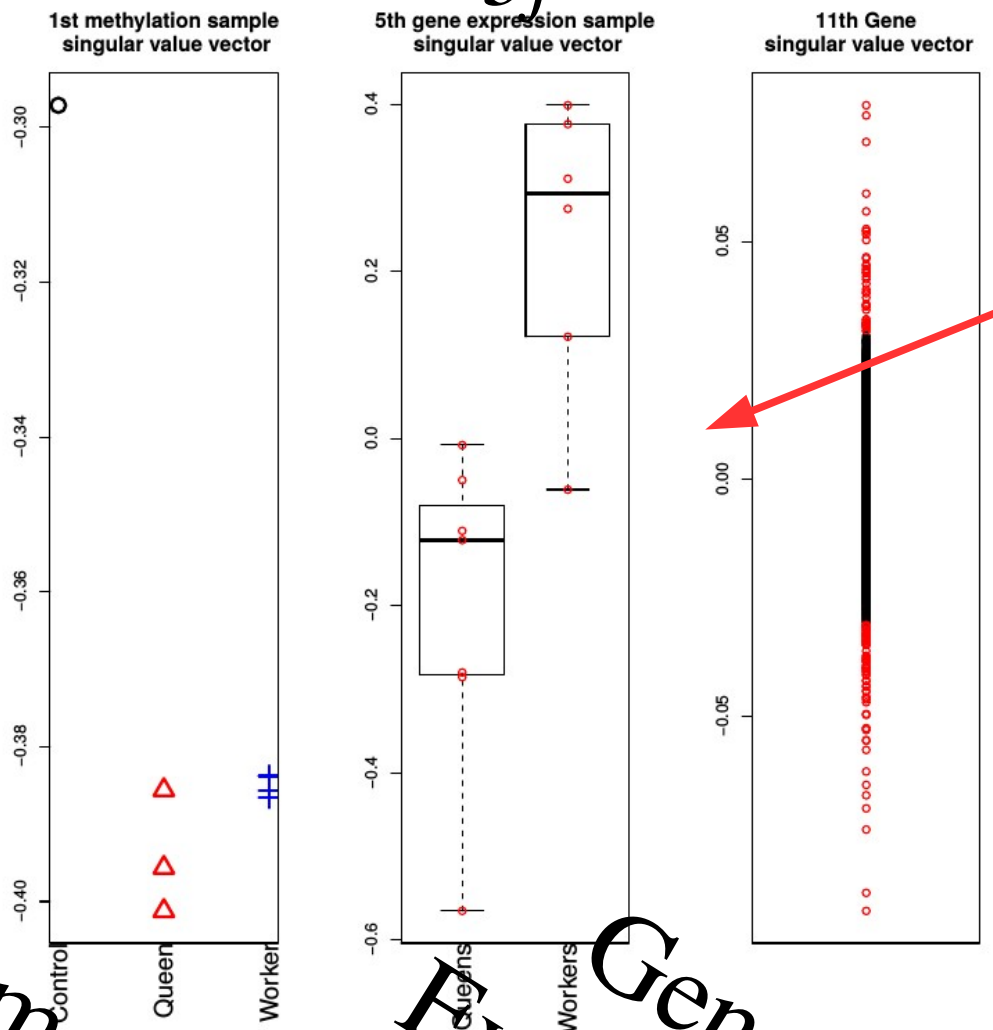
P. canadensis

$l1$	$G(l1, l2, l3)$ $(l2, l3) = (1,3)$
9	−79.8
10	75.4
7	−61.4
11	38.4
5	−23.4
4	−16.0
12	−11.9
1	−5.4
13	5.4
6	−4.5

Red: selected genes

D. quadriceps

u_{1k} u_{5j}



Gene Expression sample

methylation sample

D. quadriceps	
l_1	$G(\,l1, l2, l3\,)$ $(\,l2, l3\,) = (1,5)$
11	−54.8
12	4.1
25	3.4
2	−2.9
23	2.8
9	2.4
20	−2.2
8	2.2
10	−1.7
22	−1.4

Red: selected genes

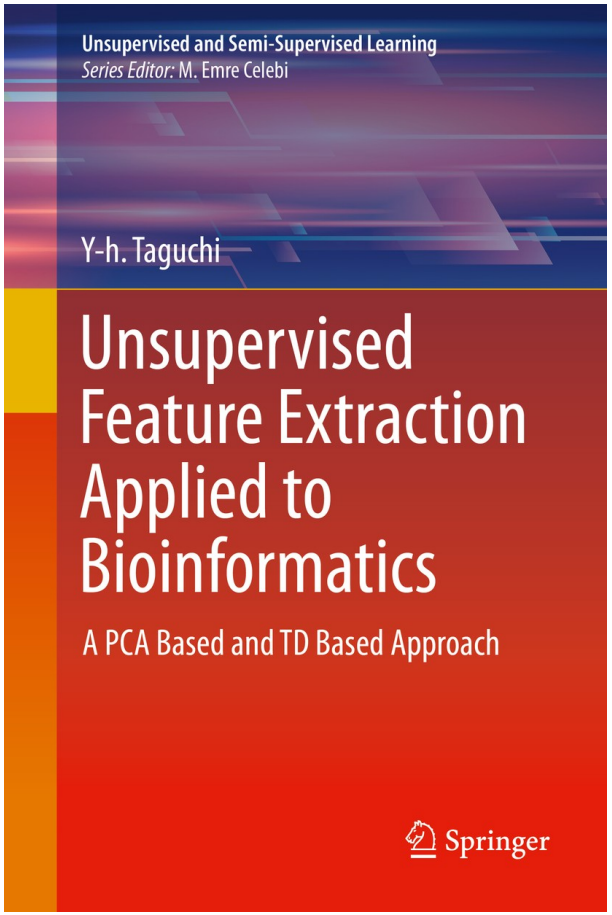
Are the selected genes methylated and/or expresseddistinctly between queens and worker?

		t	Wilcox	KS
P. canadensis	gene expression	1.71×10^{-3}	1.89×10^{-2}	0.08
	methylation	1.74×10^{-4}	5.06×10^{-3}	1.02×10^{-3}
D. quadriceps	gene expression	2.73×10^{-12}	9.05×10^{-12}	4.41×10^{-11}
	methylation	0.3757	0.7163	0.4413

Yes, the selected genes are expressed distinctly between queens and worker for both species,
but are methylated distinctly between queens and workers only for *P. canadensis*

Summary

We can select biologically reasonable genes with unsupervised methods using TD for multi-omics data analysis.



I have published a monograph from Springer. I am happy if you can but it, although it is extremely expensive.