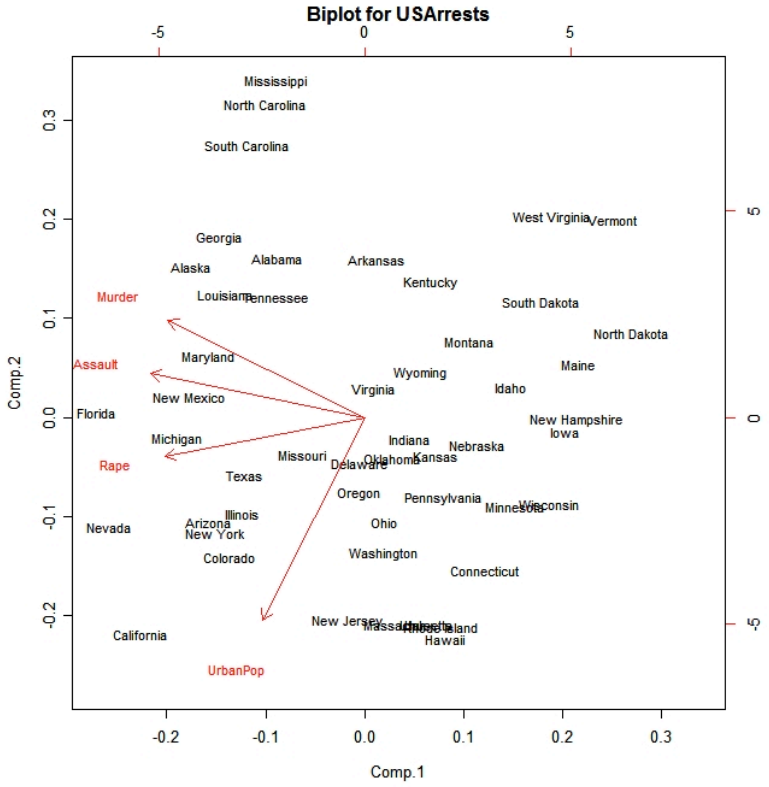
Exploratory analysis of the ENCODE TF binding data using biplot

Zhengdong Zhang
Chao Cheng, Pedro Alves
Mark Gerstein

4 June 2009

Yale University

Introduction



- A biplot is a low-dimensional (usually 2D) representation of a data matrix $\mathcal{A}$.
  - A point for each of the *m* observation vectors (rows of $\mathcal{A}$)
  - A line (or arrow) for each of the *n* variables (columns of $\mathcal{A}$)

1 Intuitive toy ex.
2 Worked out ex. on ENCODE pilot
3 Rough version on current ENCODE data matrix

PCA

TFs: a, b, c...

Genomic Sites: 1,2,3...

A

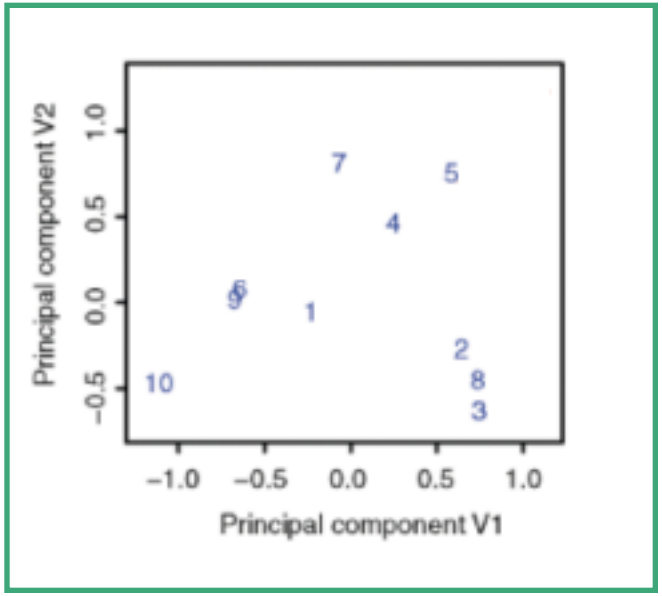
	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

A^T

	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

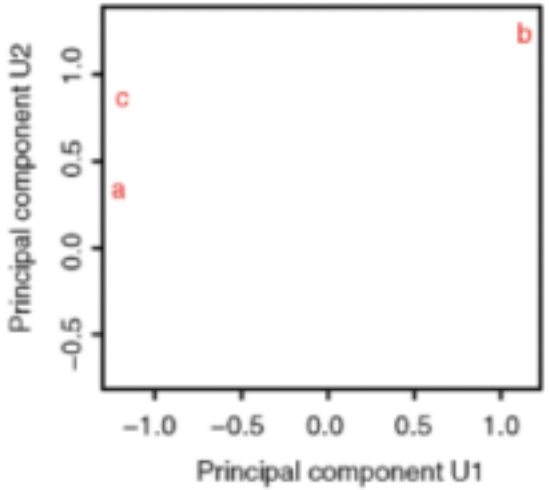
	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA (TF-TF corr.)



	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



Biplot to Show Overall Relationship of TFs & Sites

TFs: a, b, c...

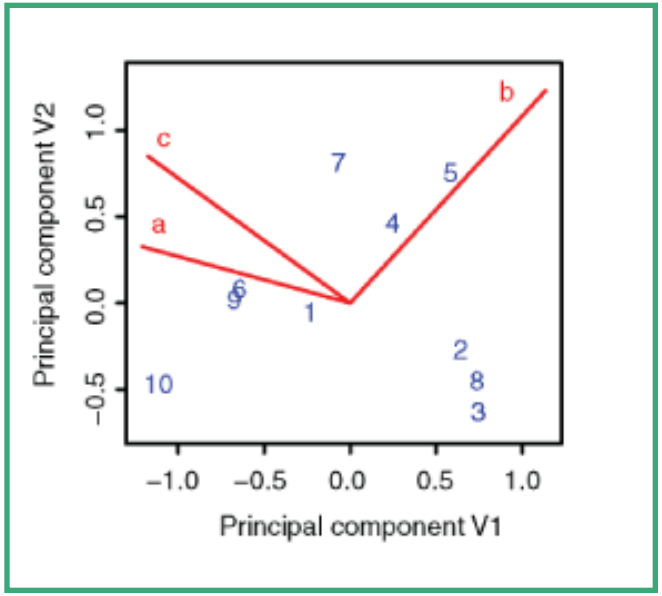
Genomic Sites: 1,2,3...

$A=USV^T$

	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA (TF-TF corr.)

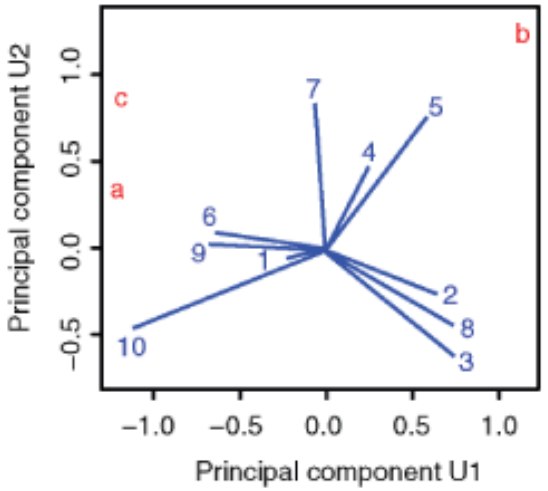


A^T

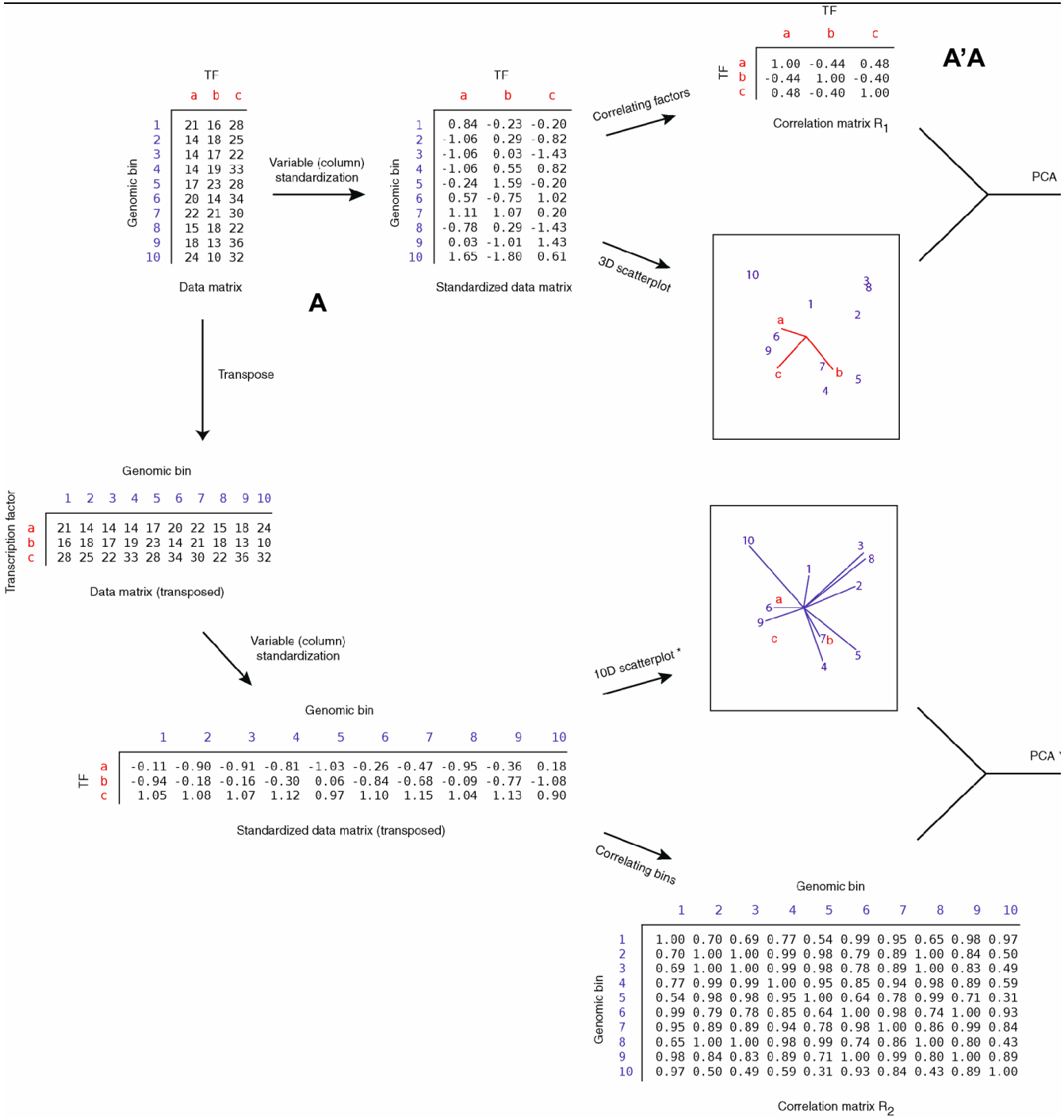
	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



Biplot Ex

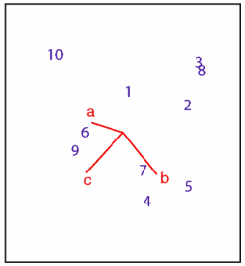

TF
a b c
Genomic bin
1 21 16 28
2 14 18 25
3 14 17 22
4 14 19 33
5 17 23 28
6 20 14 34
7 22 21 30
8 15 18 22
9 18 13 36
10 24 10 32
Data matrix
Variable (column) standardization
TF
a b c
1 0.84 -0.23 -0.20
2 -1.06 0.29 -0.82
3 -1.06 0.03 -1.43
4 -1.06 0.55 0.82
5 -0.24 1.59 -0.20
6 0.57 -0.75 1.02
7 1.11 1.07 0.20
8 -0.78 0.29 -1.43
9 0.03 -1.01 1.43
10 1.65 -1.80 0.61
Standardized data matrix
A
Transpose
Genomic bin
1 2 3 4 5 6 7 8 9 10
Transcription factor
a 21 14 14 14 17 20 22 15 18 24
b 16 18 17 19 23 14 21 18 13 10
c 28 25 22 33 28 34 30 22 36 32
Data matrix (transposed)
Variable (column) standardization
Genomic bin
1 2 3 4 5 6 7 8 9 10
TF
a -0.11 -0.90 -0.91 -0.81 -1.03 -0.26 -0.47 -0.95 -0.36 0.18
b -0.94 -0.18 -0.16 -0.30 0.06 -0.84 -0.68 -0.09 -0.77 -1.08
c 1.05 1.08 1.07 1.12 0.97 1.10 1.15 1.04 1.13 0.90
Standardized data matrix (transposed)
Correlating factors
TF
a b c
A'A
a 1.00 -0.44 0.48
b -0.44 1.00 -0.40
c 0.48 -0.40 1.00
Correlation matrix R1
PCA
3D scatterplot
10D scatterplot *
PCA '
Correlating bins
Genomic bin
1 2 3 4 5 6 7 8 9 10
Genomic bin
1 1.00 0.70 0.69 0.77 0.54 0.99 0.95 0.65 0.98 0.97
2 0.70 1.00 1.00 0.99 0.98 0.79 0.89 1.00 0.84 0.50
3 0.69 1.00 1.00 0.99 0.98 0.78 0.89 1.00 0.83 0.49
4 0.77 0.99 0.99 1.00 0.95 0.85 0.94 0.98 0.89 0.59
5 0.54 0.98 0.98 0.95 1.00 0.64 0.78 0.99 0.71 0.31
6 0.99 0.79 0.78 0.85 0.64 1.00 0.98 0.74 1.00 0.93
7 0.95 0.89 0.89 0.94 0.78 0.98 1.00 0.86 0.99 0.84
8 0.65 1.00 1.00 0.98 0.99 0.74 0.86 1.00 0.80 0.43
9 0.98 0.84 0.83 0.89 0.71 1.00 0.99 0.80 1.00 0.89
10 0.97 0.50 0.49 0.59 0.31 0.93 0.84 0.43 0.89 1.00
Correlation matrix R2

Biplot Ex #2

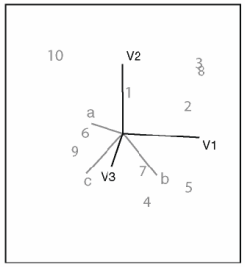
TF

TF	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

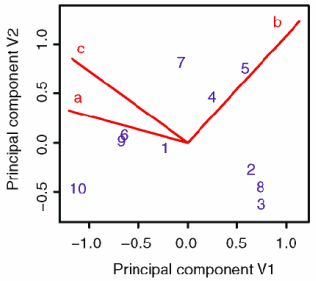
Correlation matrix R_1



PCA *



Projection *

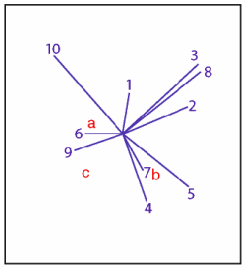


$$A^T A = V S^2 V^T$$

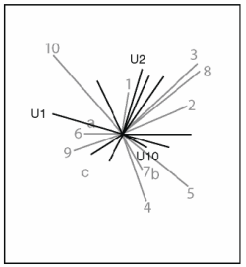
The same rank-2 approximation of the original data matrix

$$A v_j = u_j s_j \;\&\; A^T u_j = v_j s_j$$

$$A = (U S^r)(V S^{1-r})^T$$

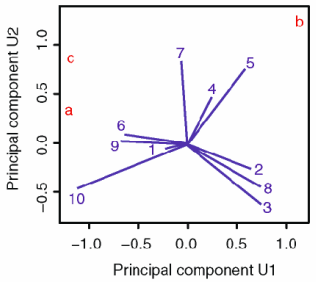


PCA *



$$A A^T = U S^2 U^T$$

Projection *

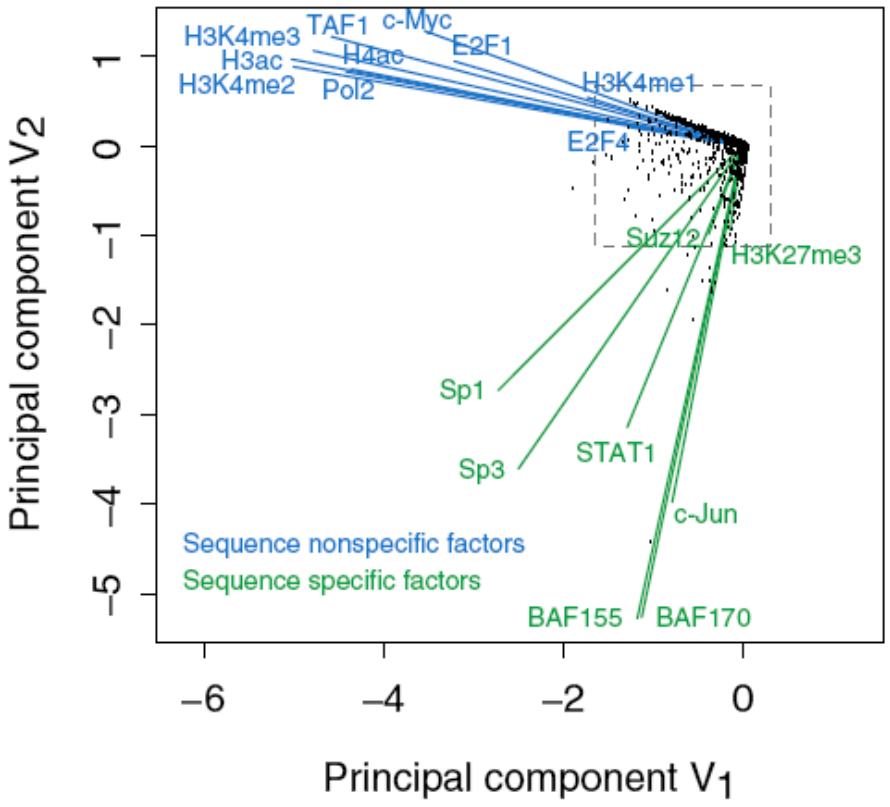


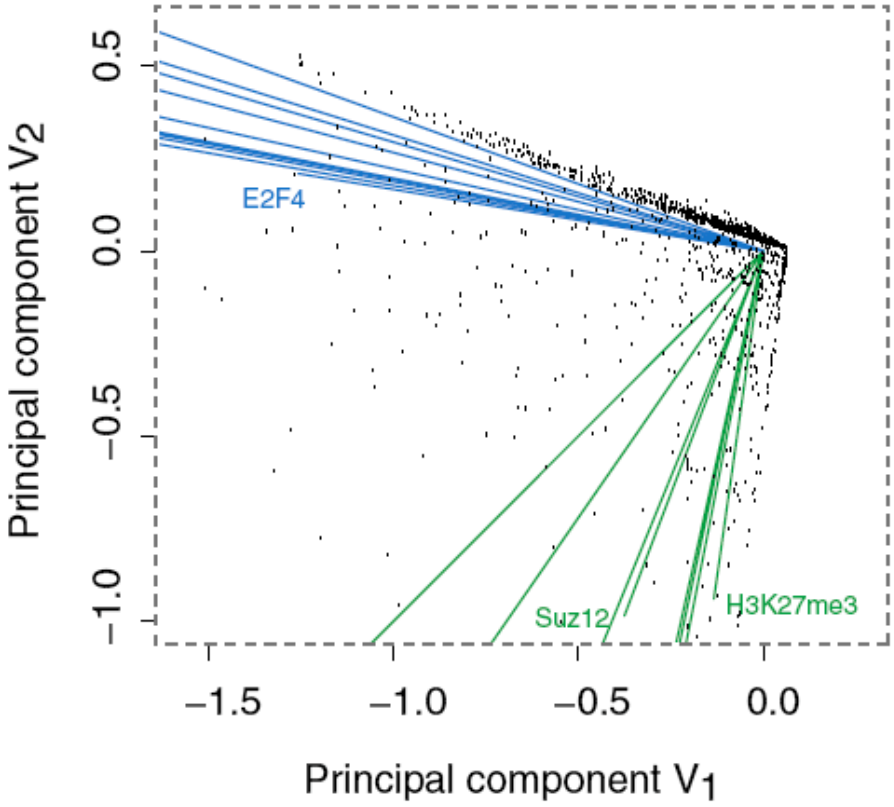
Genomic bin

	2	3	4	5	6	7	8	9	10
	.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
	.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
	.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
	.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
	.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
	.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
	.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
	.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
	.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
	.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

Correlation matrix R_2

* 10D scatterplots are used here for illustrative purpose only.
PCA: the correlation matrix is eigen-decomposed; then the principal components are added to the original space.
Projection: the points and axes in the original space are projected onto the plane defined by the top two principal components.

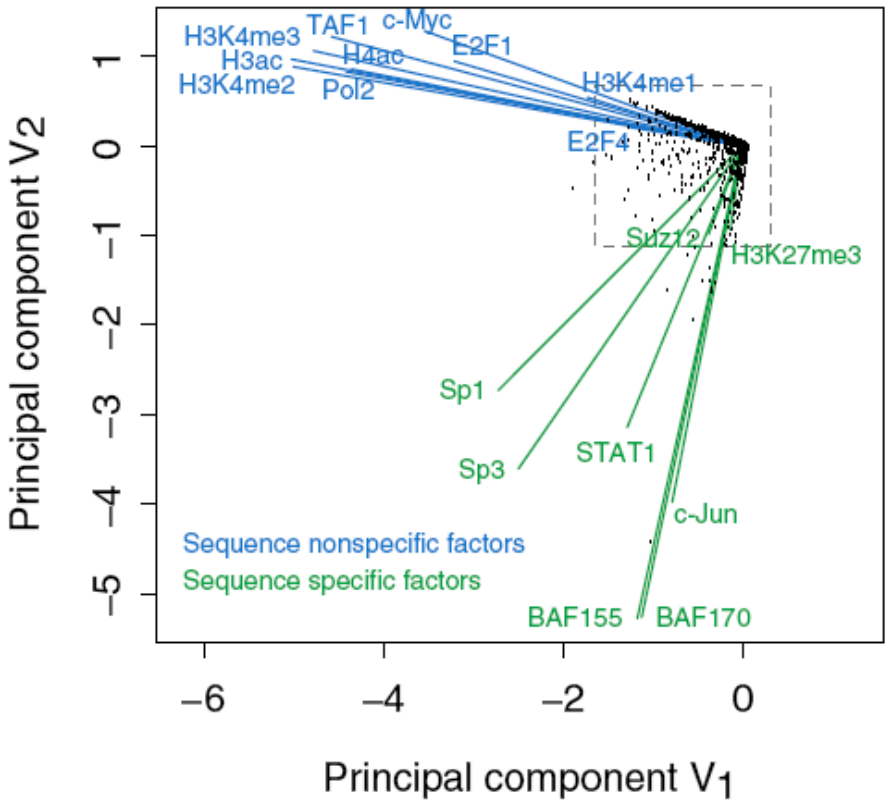


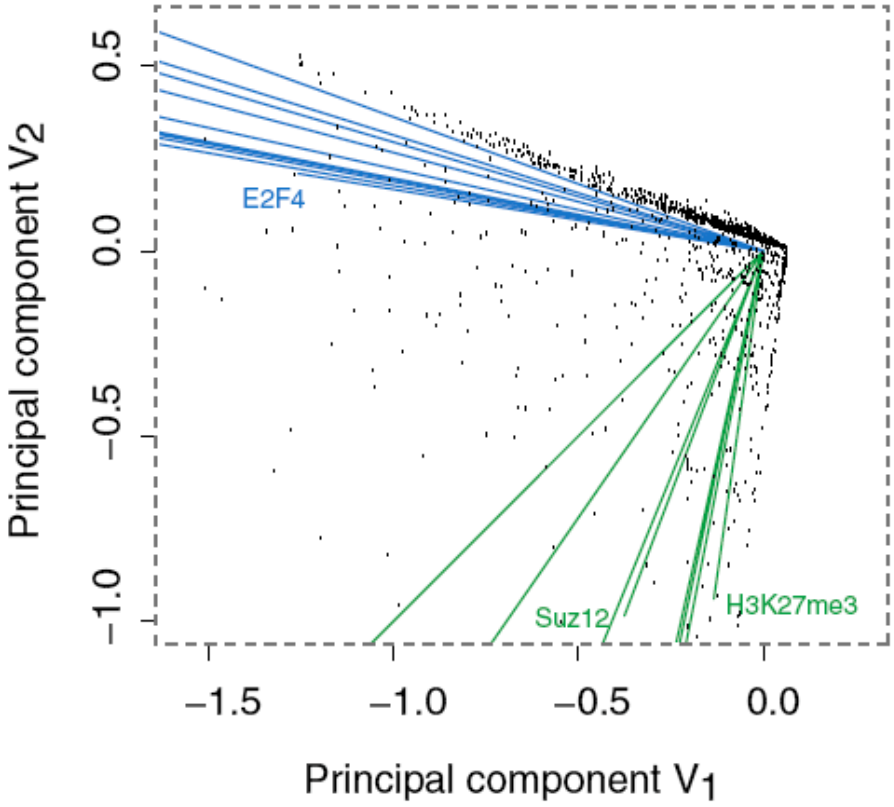


Results of Biplot

- Pilot ENCODE (1% genome): 5996 10 kb genomic bins (adding all hits) + 105 TF experiments → biplot
- Angle between TF vectors shows relation b/w factors
- Closeness of points gives clustering of "sites"
- Projection of site onto vector gives degree to which site is assoc. with a particular factor

Zhang et al. (2007) Gen. Res.





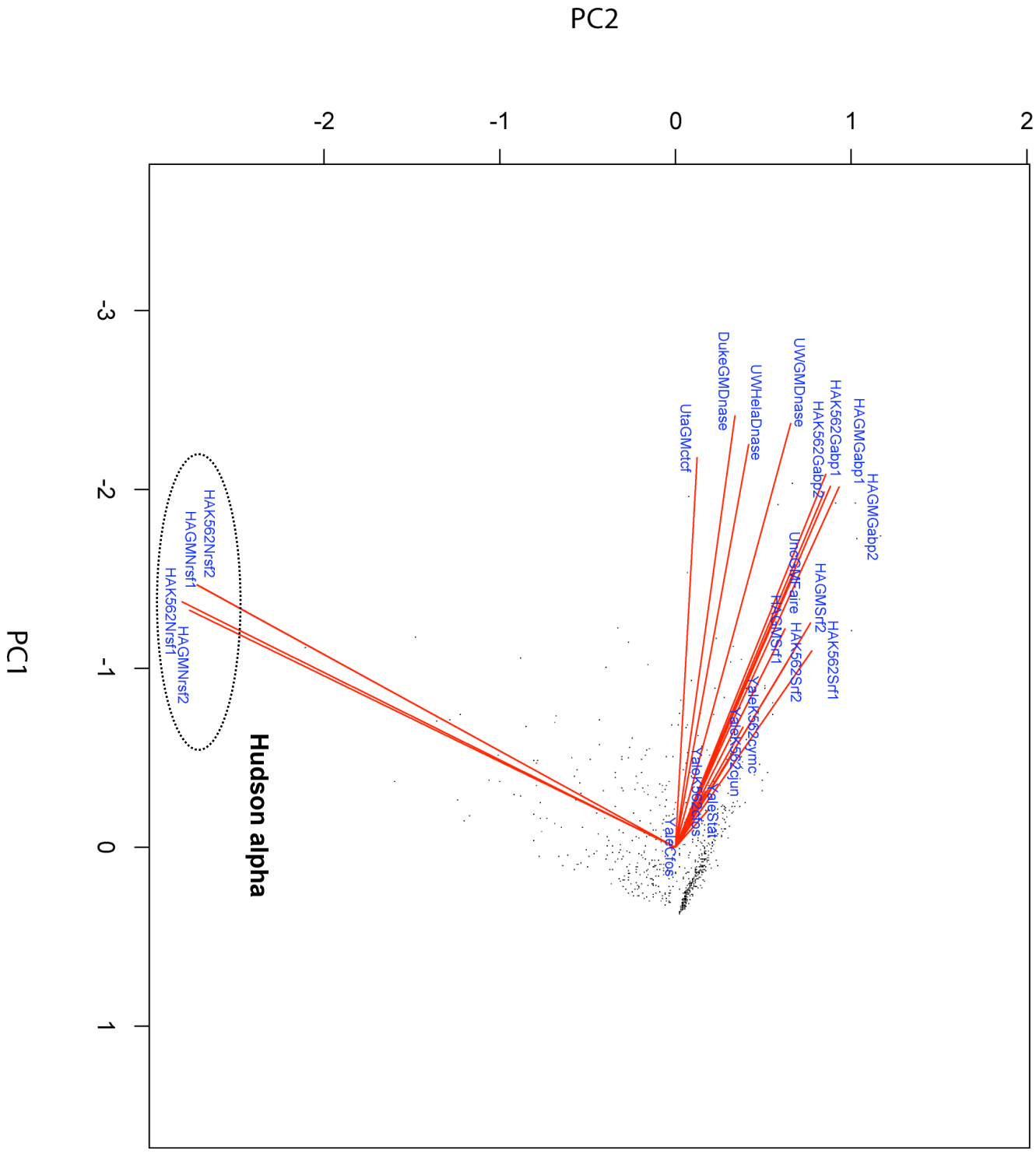
Results of Biplot

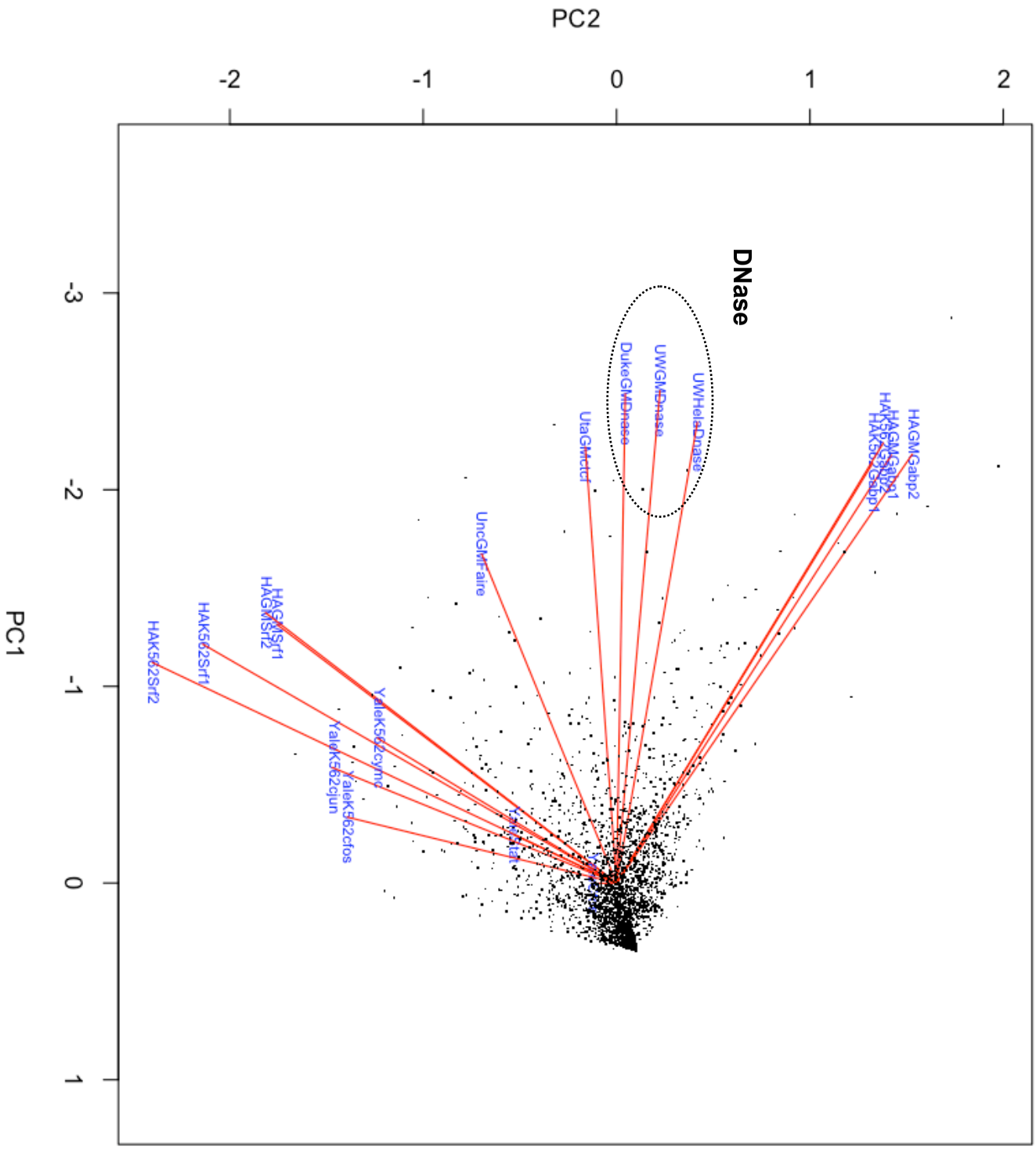
- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - c-Myc may behave more like a sequence-nonspecific TF.
  - H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion each bin is "annotated" by closest TF cluster

Zhang et al. (2007) Gen. Res.

Biplot Application

- Data: the raw ENCODE data matrix (397706 x 22)
- Procedure:
  - Observation reduction: (2903 x 22)
  - Standardization: transform the columns into unit vectors: mean=0, sd=1
  - Perform SVD
- Calculation can be easily done in MatLab. But interpretation may not be simple…


PC2
-2
-1
0
1
2
PC1
-3
-2
-1
0
1
HAK562Nrsf2
HAGMNrsf1
HAK562Nrsf1
HAGMNrsf2
Hudson alpha
DukeGMDnase
UWHelaDnase
UWGMDnase
HAK562Gabp1
HAK562Gabp2
HAGMGabp1
HAGMGabp2
UtaGMctcf
HAGMSrf2
HAK562Srf1
HAGMSrf1
HAK562Srf2
YaleK562cjun
YaleStat
YaleC1q6


PC2
-2
-1
0
1
2
PC1
-3
-2
-1
0
1
DNase
DukeGMDnase
UWGMDnase
UWHelaDnase
UtaGMctcf
UncGMFaire
HAGMGabp2
HAGMSrf2
HAGMSrf1
HAK562Srf1
HAK562Srf2
YaleK562cjun
YaleK562cfos