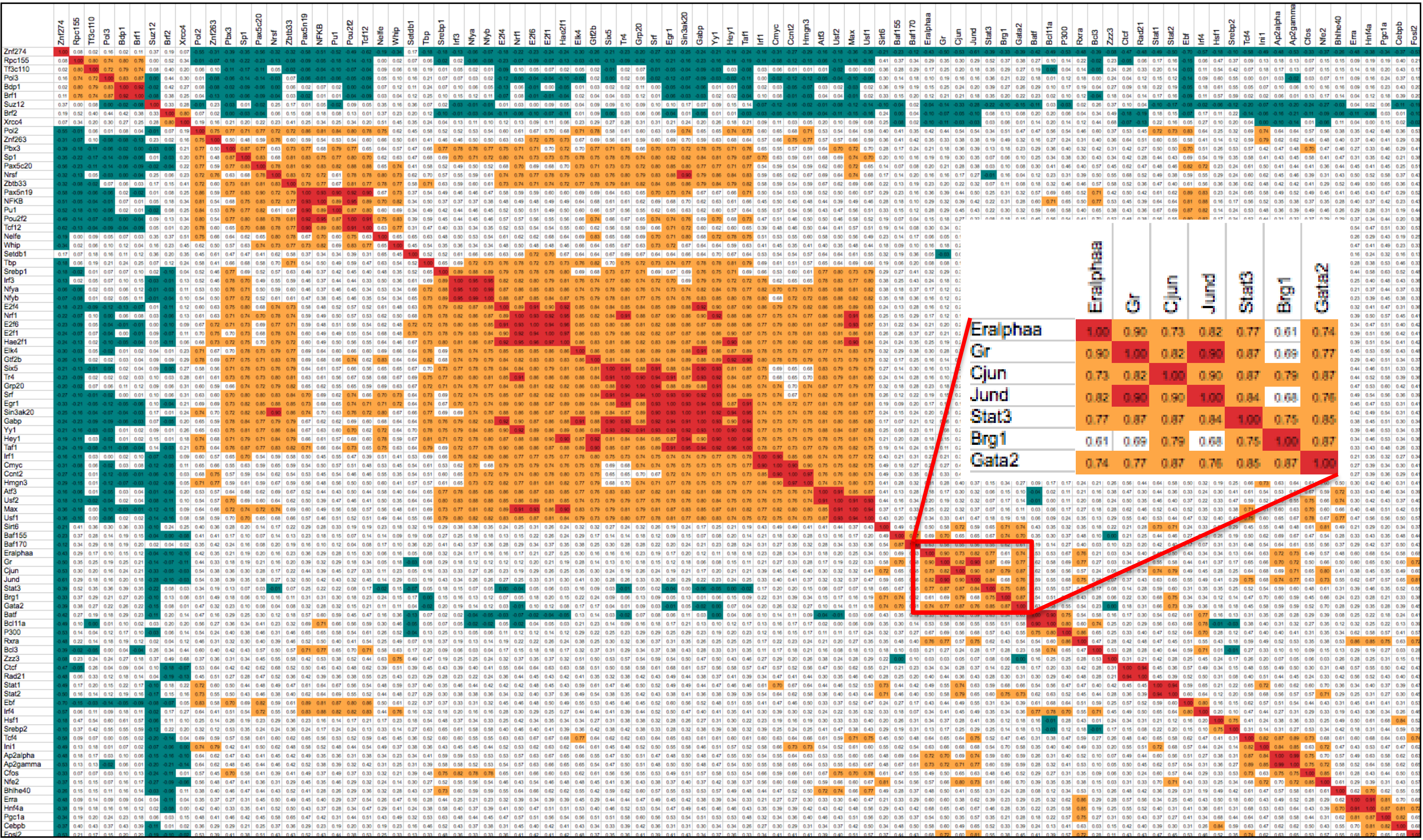
Cell type-specific Transcription Factor Co-associations:

Partner TF Importance & Combinatorial Associations

Anshul Kundaje & Manoj Hariharan
with Michael Snyder

Genome-wide Co-association Matrix

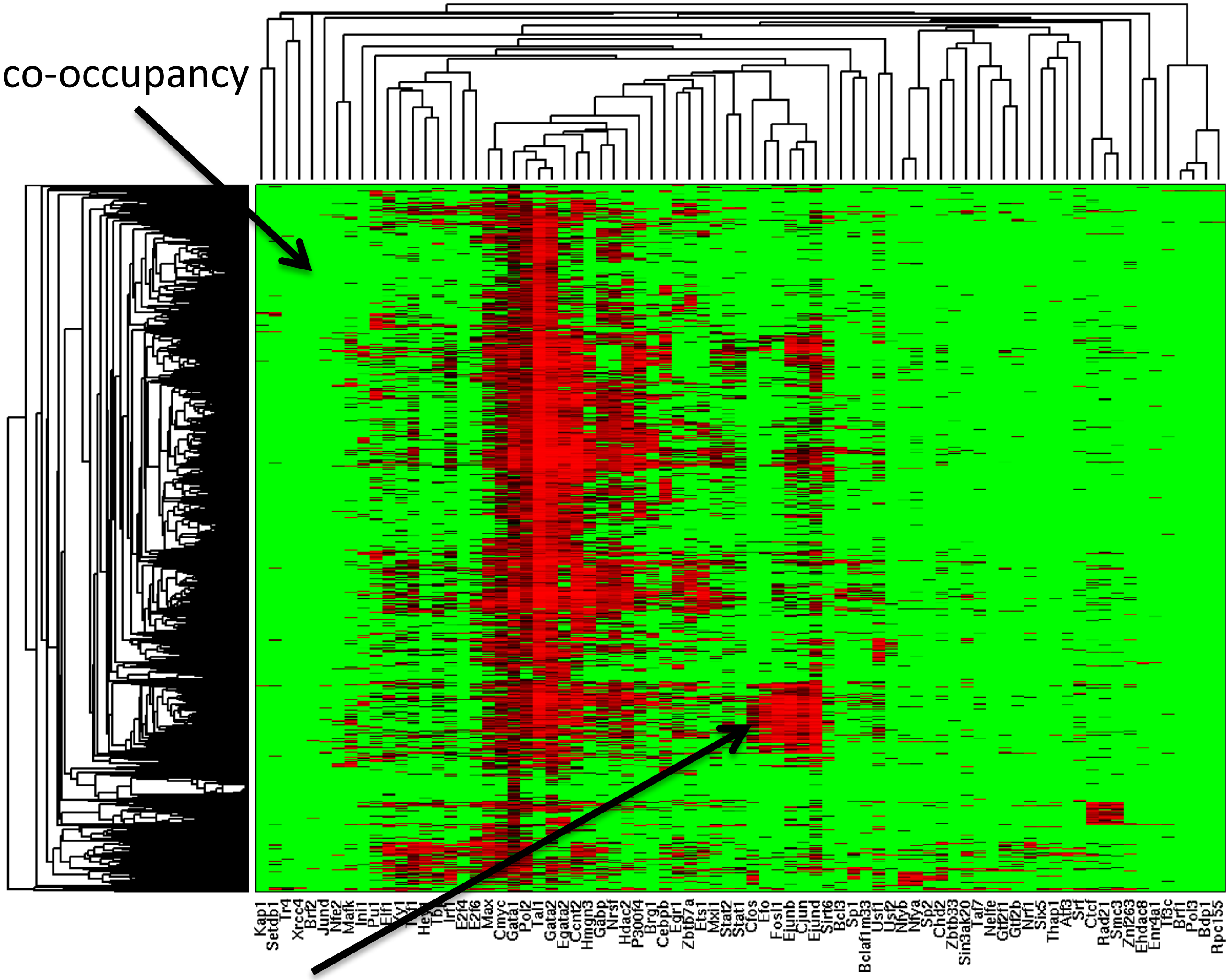


	Eralphaa	Gr	Cjun	Jund	Stat3	Brg1	Gata2
Eralphaa	1.00	0.90	0.73	0.82	0.77	0.61	0.74
Gr	0.90	1.00	0.82	0.90	0.87	0.69	0.77
Cjun	0.73	0.82	1.00	0.90	0.87	0.79	0.87
Jund	0.82	0.90	0.90	1.00	0.84	0.68	0.76
Stat3	0.77	0.87	0.87	0.84	1.00	0.75	0.85
Brg1	0.61	0.69	0.79	0.68	0.75	1.00	0.87
Gata2	0.74	0.77	0.87	0.76	0.85	0.87	1.00

TF-centric regulatory modules

Find biclusters of a focus-TF’s peaks (e.g. GATA1) that co-associate with distinct combinations of other TFs

Low co-occupancy



GATA1 peaks in K562

High co-occupancy

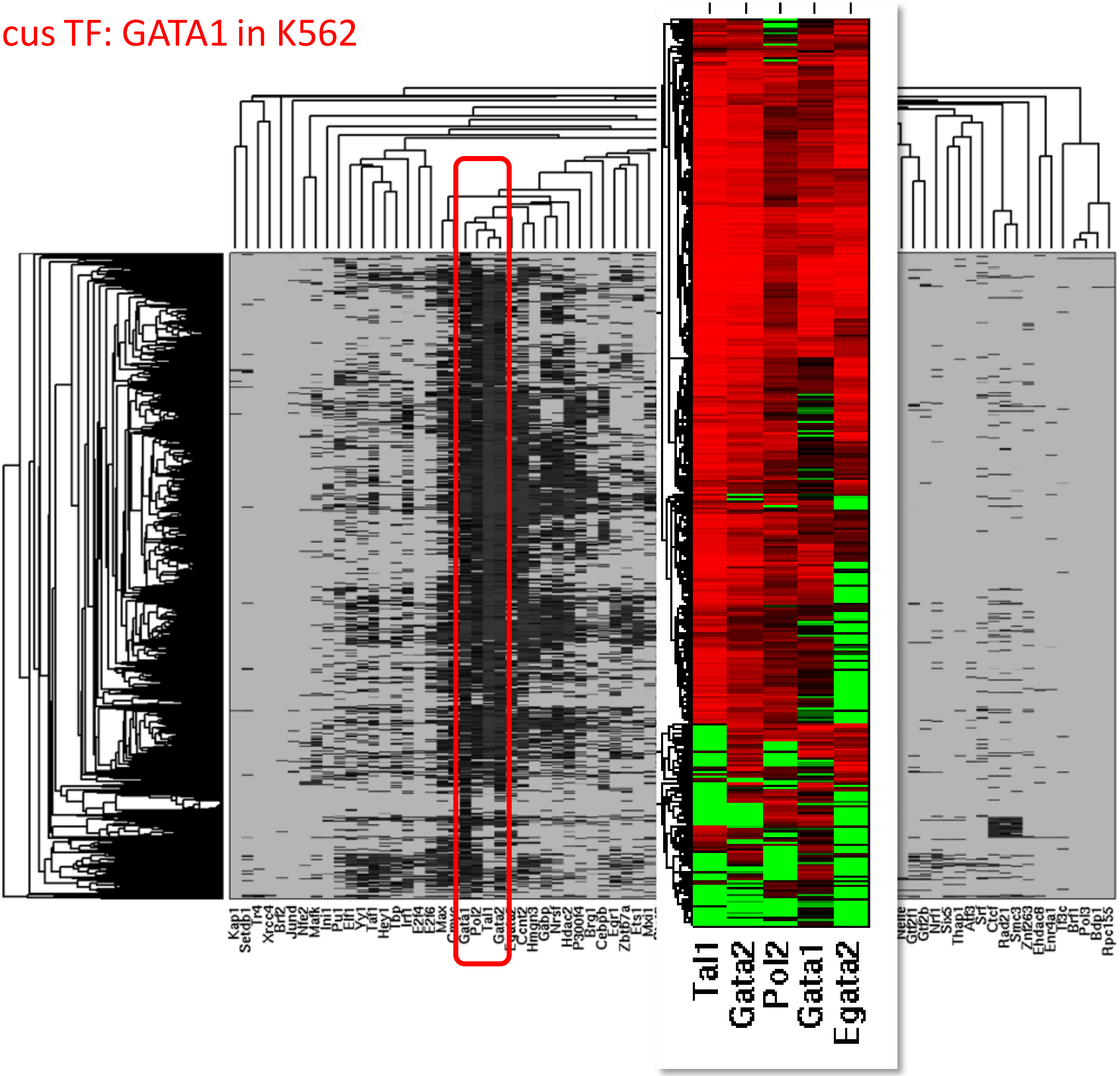
(> 1 bp peak overlap ~800 bp)

75 TFs in K562

Manoj H, Anshul K

GATA1-centric regulatory modules

Focus TF: GATA1 in K562

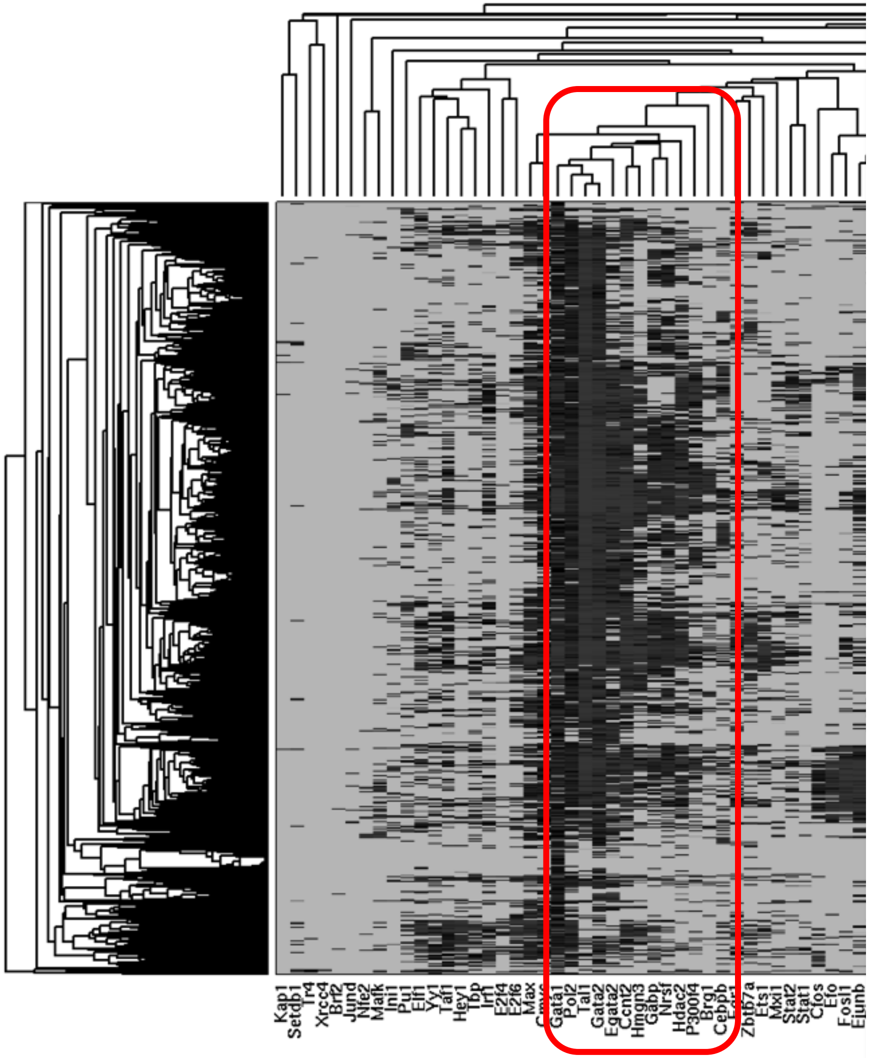


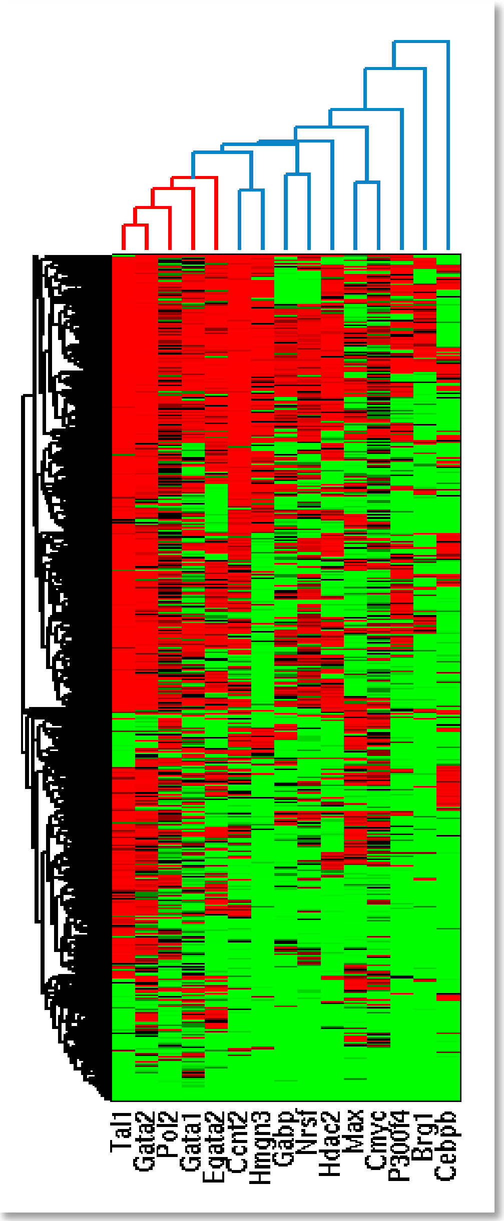
Core Module

Manoj H, Anshul K

GATA1-centric regulatory modules

Focus TF: GATA1 in K562





Core

+

Ccnt2/Hmgn3

Gabp/Nrsf/Hdac2

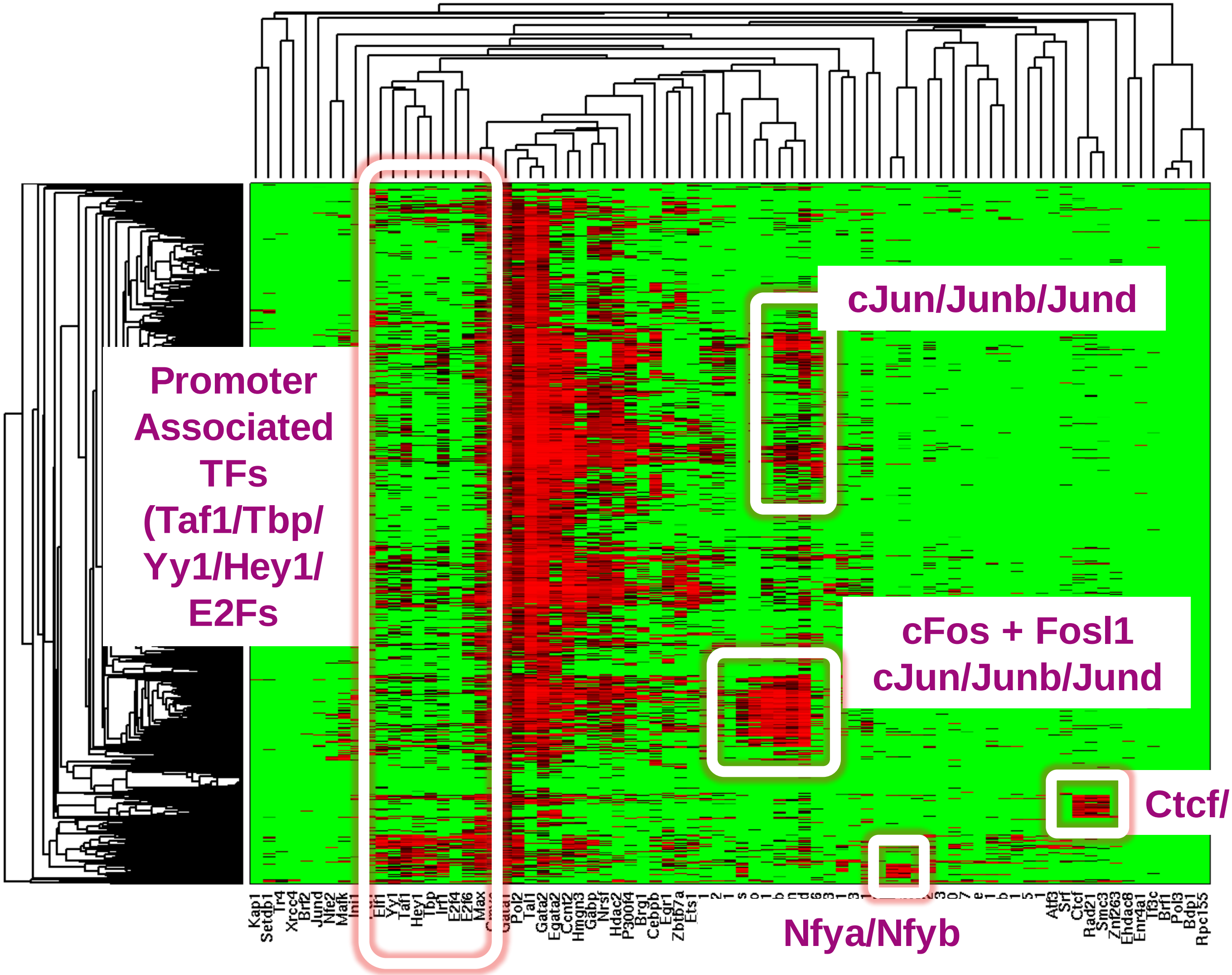
Myc/Max

P300/Cebp/Brg1

Manoj H, Anshul K

GATA1-centric regulatory modules

Focus TF: GATA1 in K562



Promoter Associated TFs (Taf1/Tbp/ Yy1/Hey1/ E2Fs

cJun/Junb/Jund

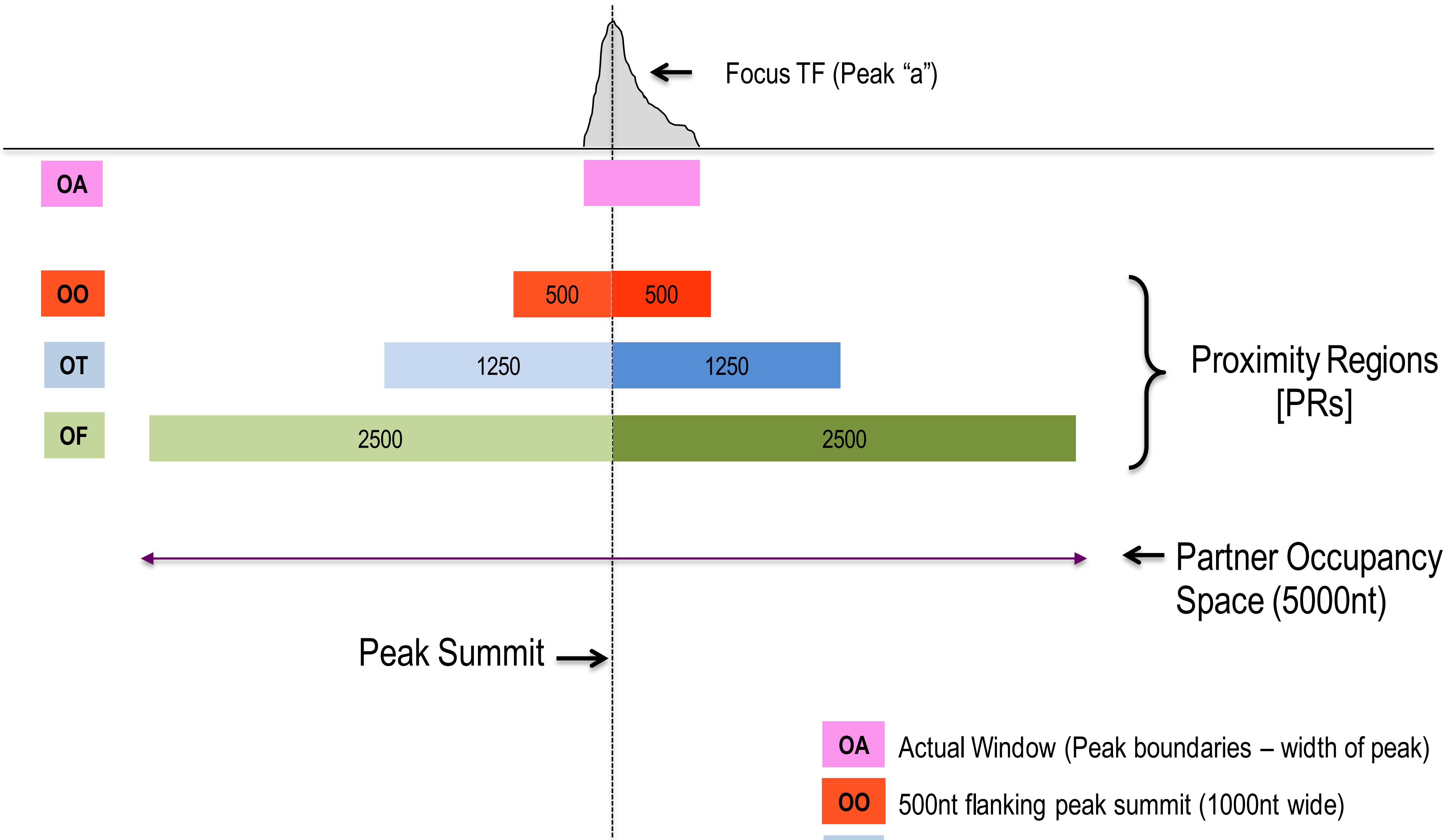
cFos + Fosl1 cJun/Junb/Jund

Ctcf/Rad21/Smc3

Nfya/Nfyb

Manoj H, Anshul K

Defining Proximity Regions



Focus TF (Peak "a")

OA

OO

500 500

OT

1250 1250

OF

2500 2500

Proximity Regions
[PRs]

Partner Occupancy
Space (5000nt)

Peak Summit

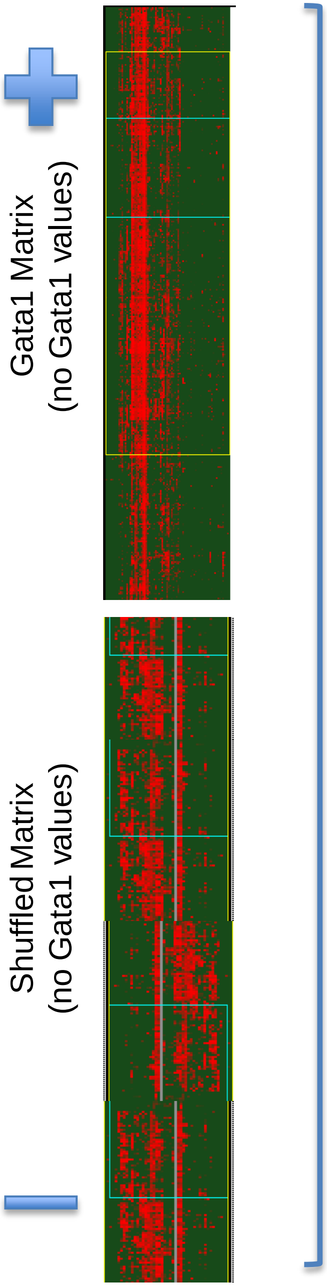
OA Actual Window (Peak boundaries – width of peak)

OO 500nt flanking peak summit (1000nt wide)

OT 1250nt flanking peak summit (2500nt wide)

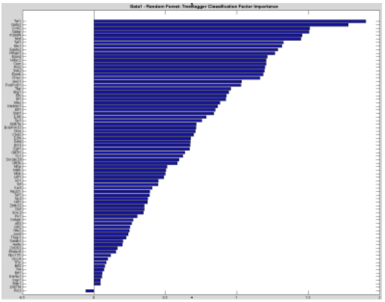
OF 2500nt flanking peak summit (5000nt wide)

Classification/Regression models for associations



Random Forest /
Boosted trees /
RuleFit

Factor Importance



Gata1 Partners

Itemset Mining

```
Gata1 <- Egata2 Gata2  (70.3, 99.9)
Gata1 <- Ccnt2 Pol2  (70.5, 99.9)
Gata1 <- Ccnt2 Gata2  (70.6, 99.9)
Gata1 <- Tal1 Cmyc  (70.4, 99.9)
Gata1 <- Tal1 Cmyc  (70.4, 99.9)
Gata1 <- Nrsf Pol2  (60.7, 99.9)
Gata1 <- Nrsf Gata2  (60.8, 99.9)
Gata1 <- Max Cmyc  (60.8, 99.9)
Gata1 <- Nrsf Pol2 Gata2  (60.0, 99.9)
Gata1 <- Ccnt2 Tal1 Cmyc  (60.1, 99.9)
```

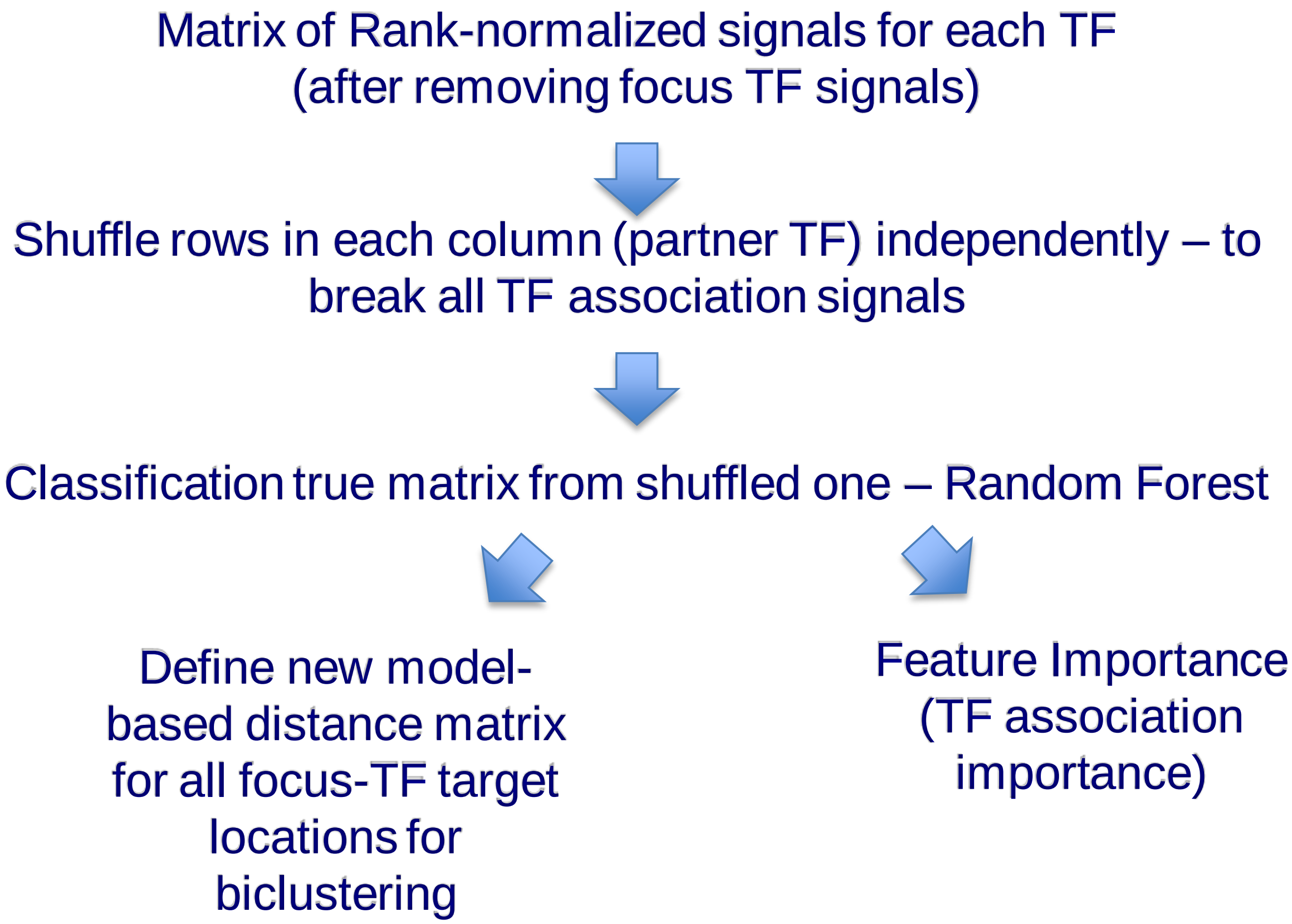
- Quantify importance of individual associated TFs and frequently occuring sets of TFs – by Random Forest/Boosted trees/RuleFit

Types of positive/negative sets

- Classification scenarios
  - **True association matrix (+) vs Shuffled coassociation matrix that breaks associations (-)**
  - True association matrix (+) vs association matrix for peaks present in other cell-lines but NOT in target cell-line (-)
  - TSS-Proximal (+) vs TSS-Distal sites (-)
  - Peaks near highly expressed genes (+) vs. Peaks near low expressed genes (-)

- Regression scenarios
  - True association matrix -> focus TF binding strength
  - True association matrix -> gene expression

Discriminative/Regression models for associations



Matrix of Rank-normalized signals for each TF (after removing focus TF signals)

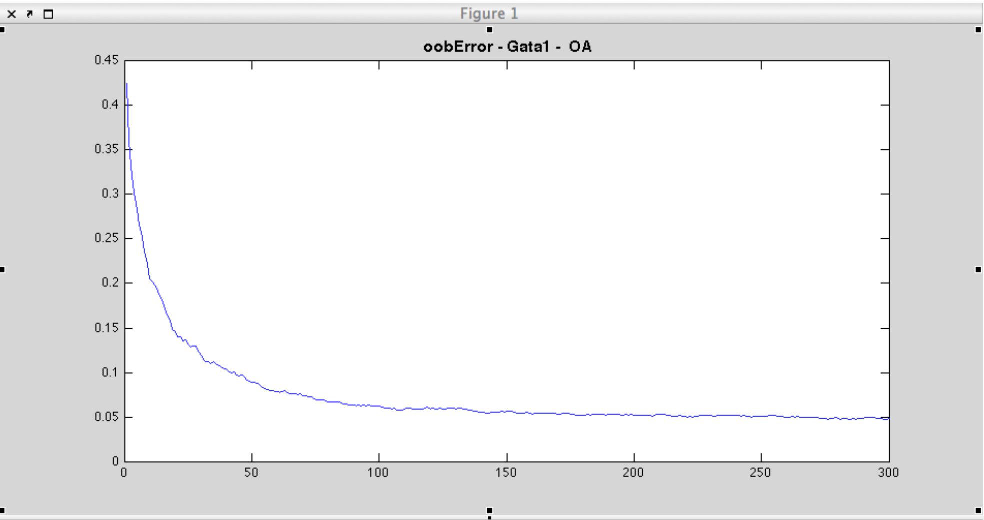
Shuffle rows in each column (partner TF) independently – to break all TF association signals

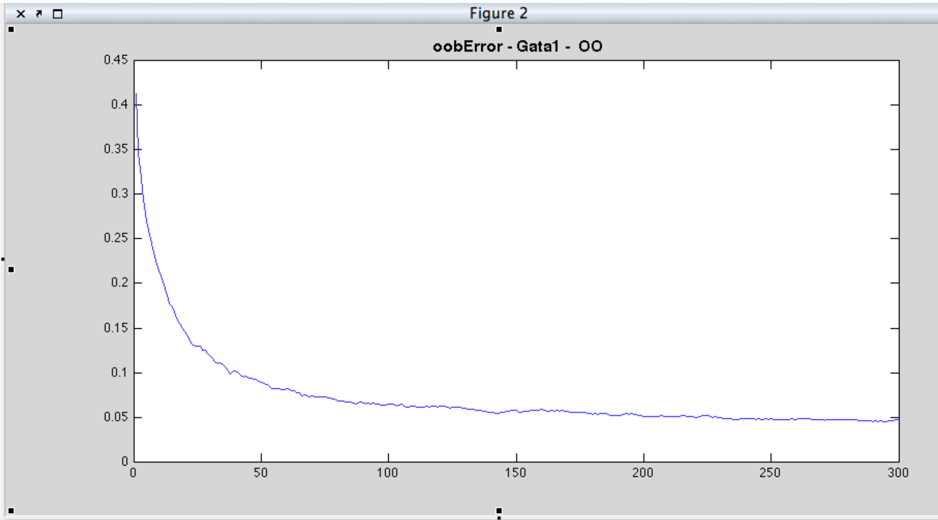
Classification true matrix from shuffled one – Random Forest

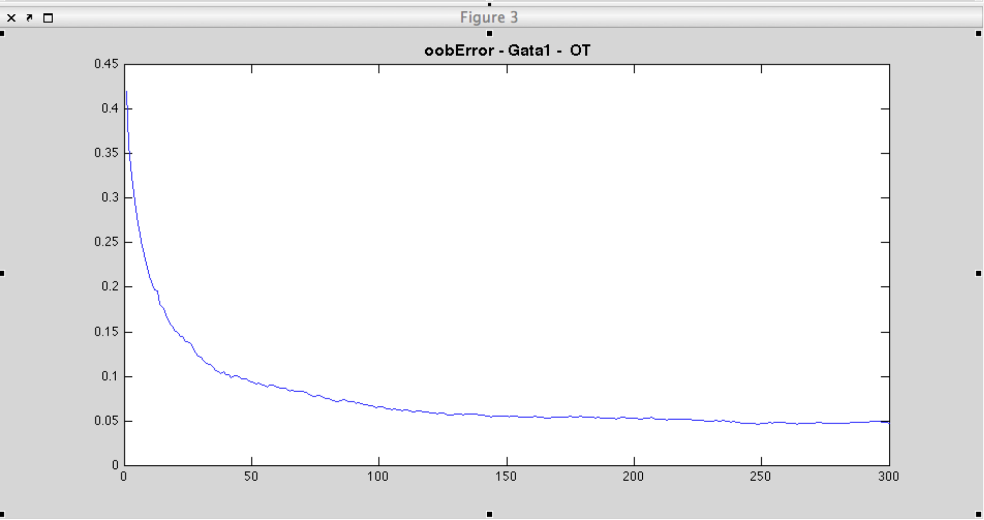
Define new model-based distance matrix for all focus-TF target locations for biclustering

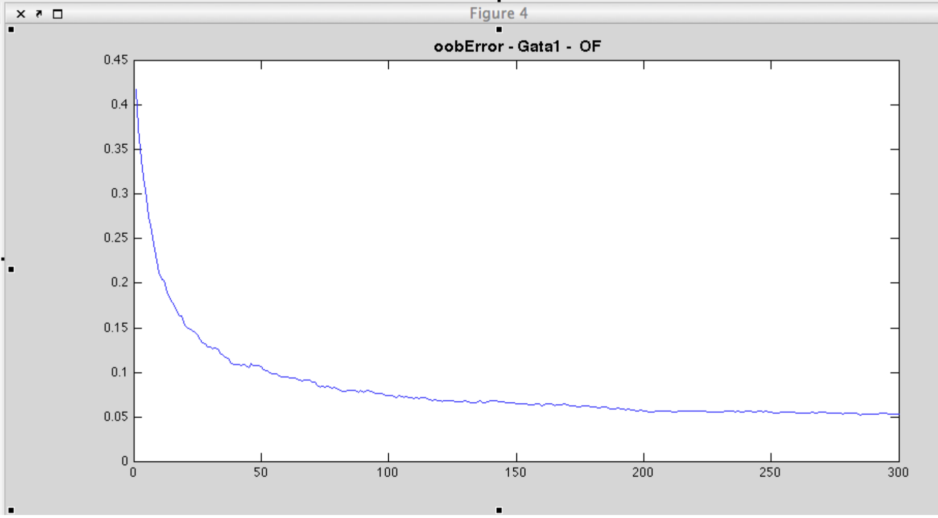
Feature Importance (TF association importance)

K562 Gata1 Partner TF OOBError Plot

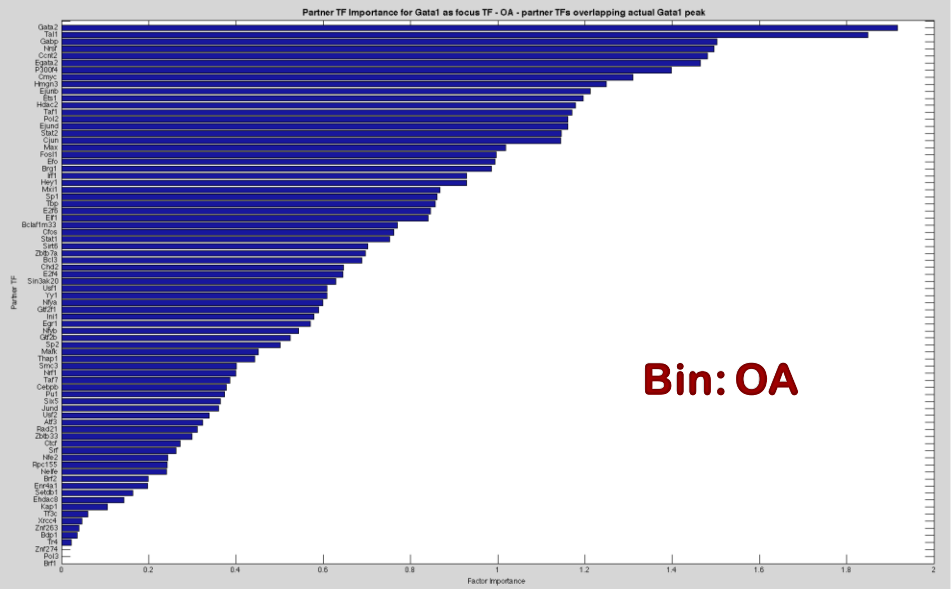


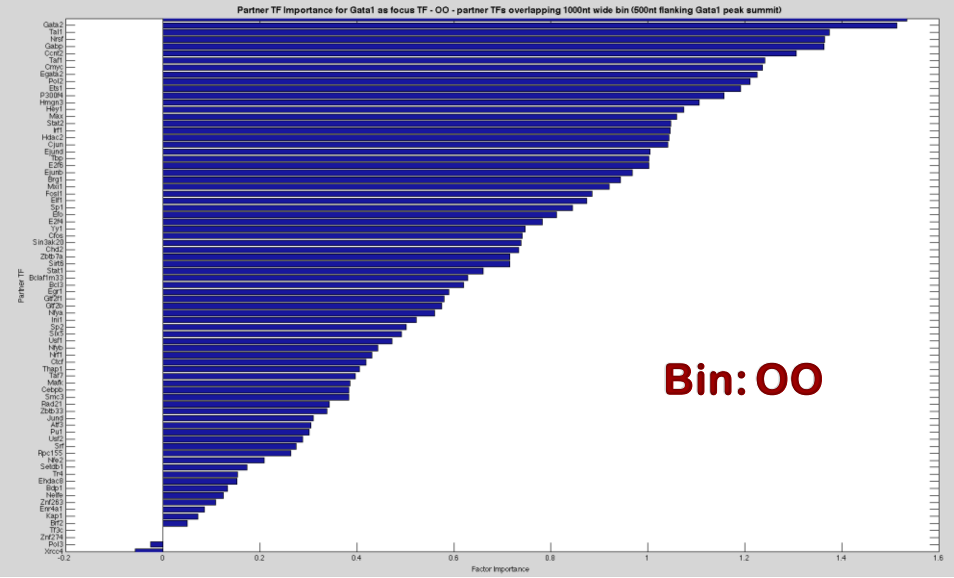


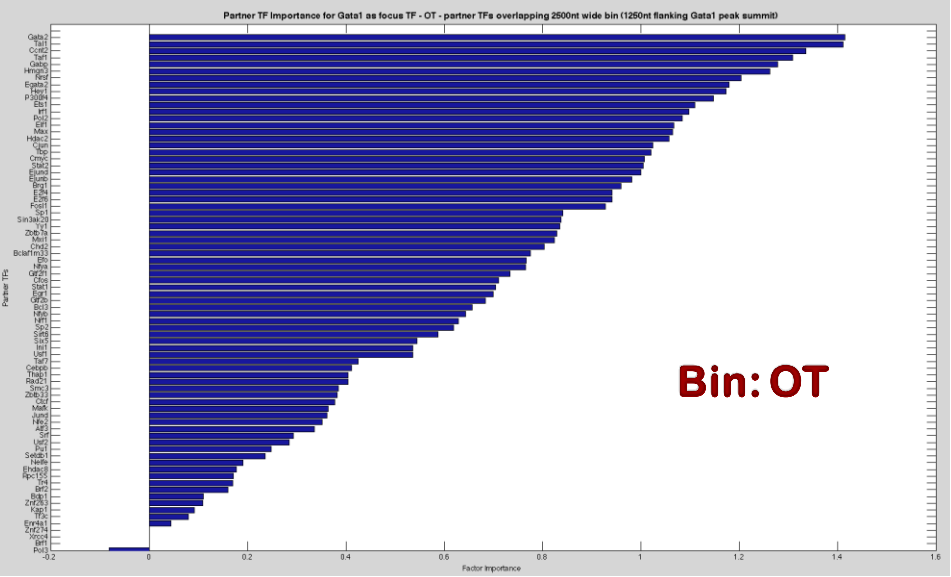


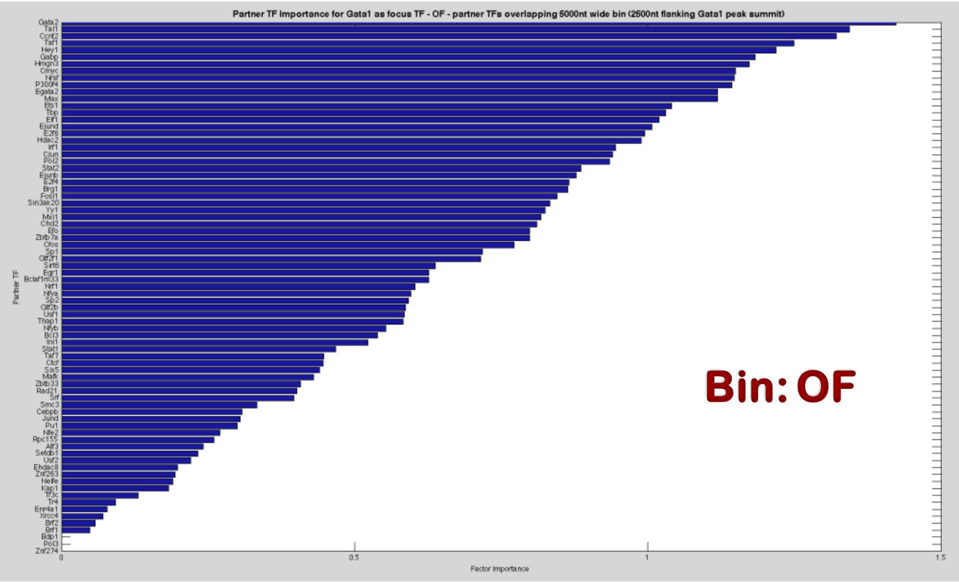


K562 Gata1 Partner TF OOB Feature Importance



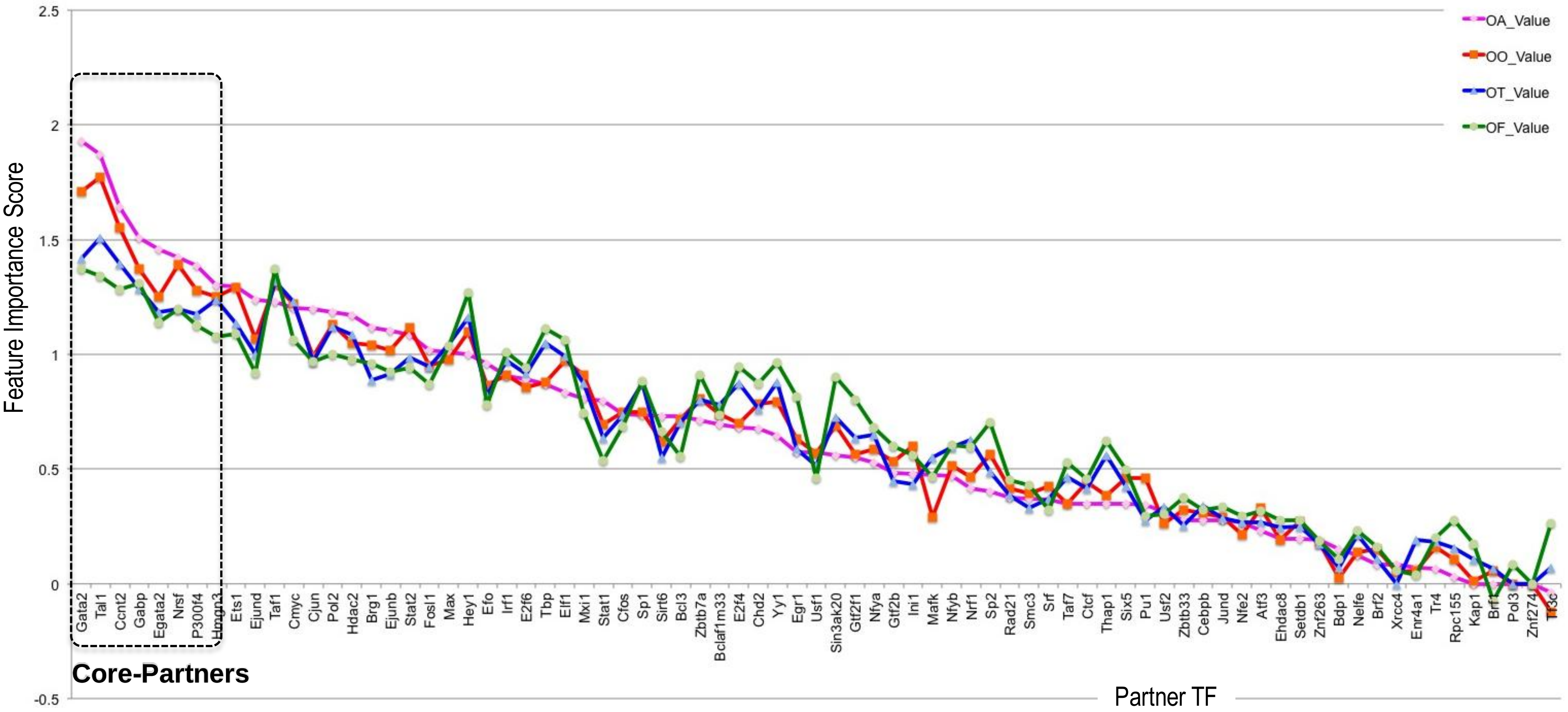






Change in Order of Partner TF Importance in Different Bins

based on Feature Importance scores of Random Forest Bagging

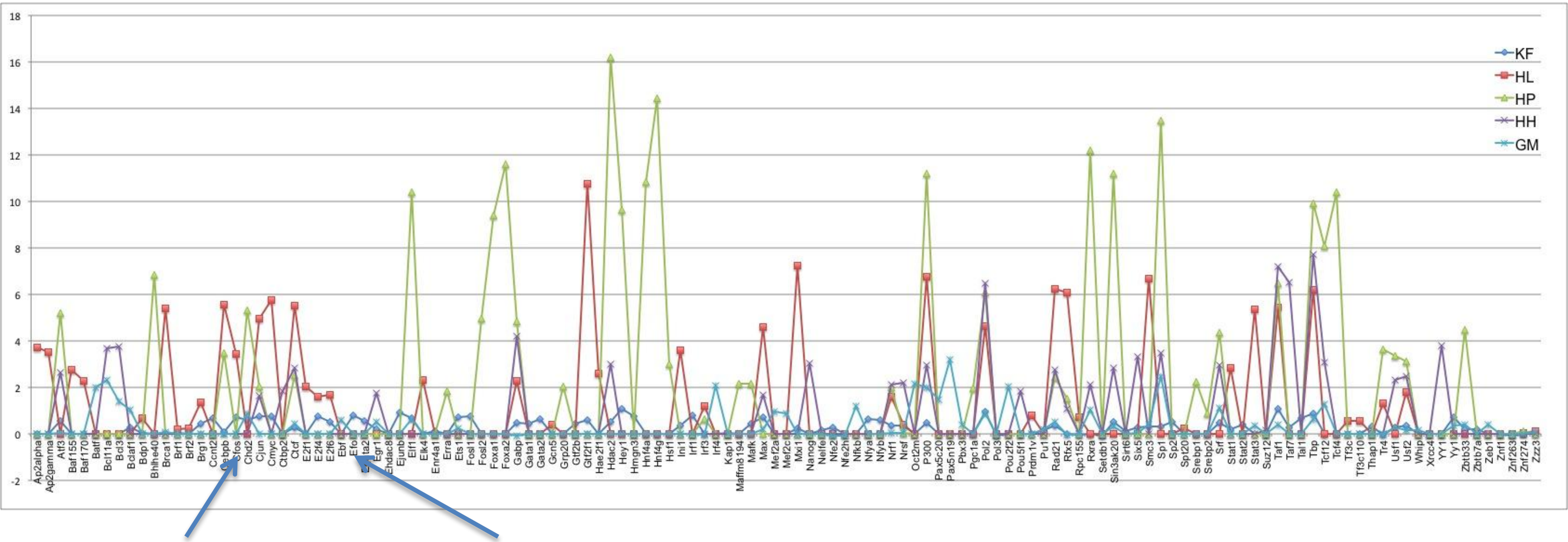


Locally associated Partners

OA	Actual Window (Peak boundaries – width of peak)
OO	500nt flanking peak summit (1000nt wide)
OT	1250nt flanking peak summit (2500nt wide)
OF	2500nt flanking peak summit (5000nt wide)

- **Core-partners get lower score in distant bins**
- **Long-range partners get higher score in distant bins**

Jund Partner TF Feature Importance – Effect of data quality



c-fos (Snyder)

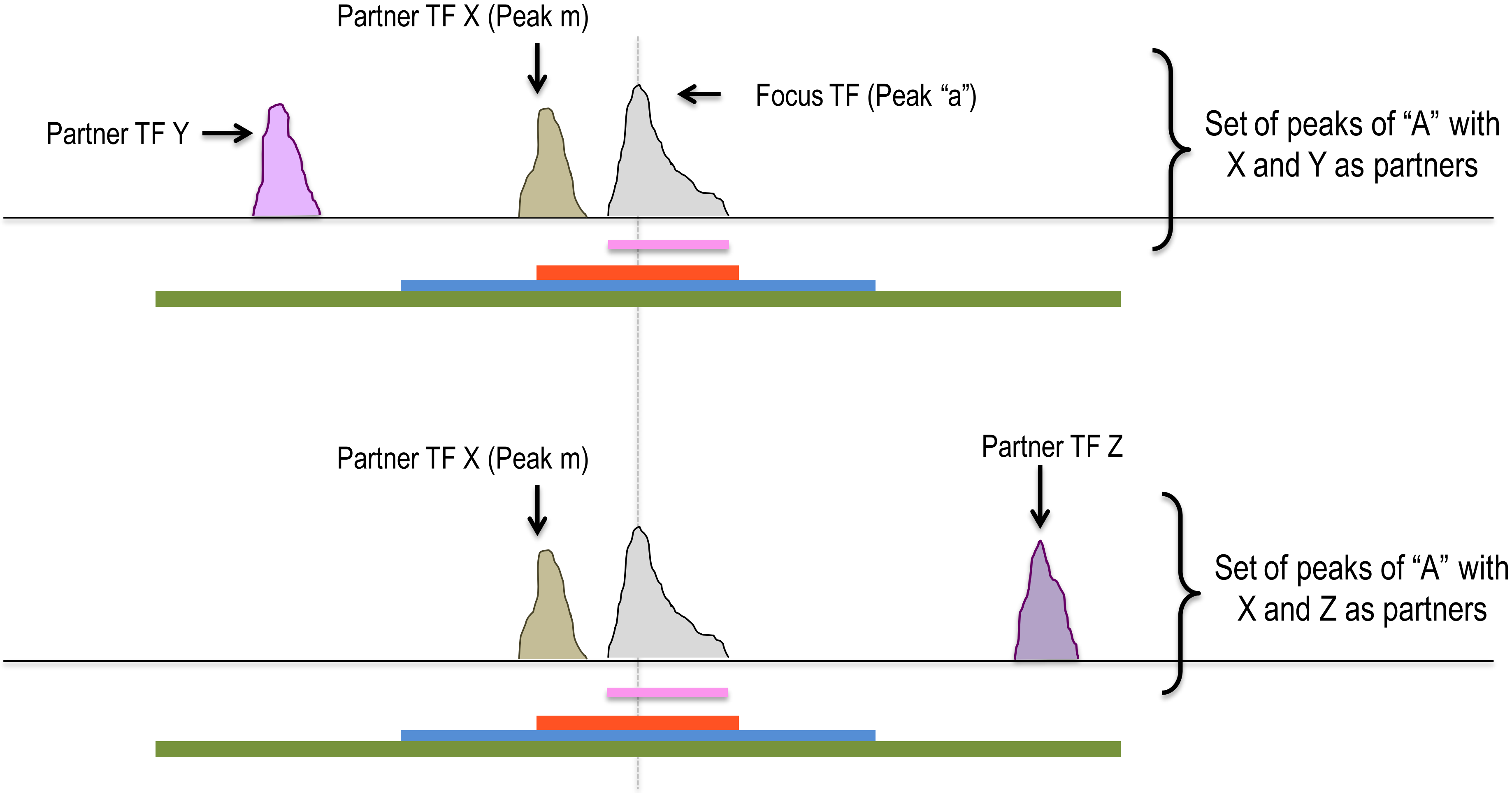
c-fos (Uchicago)

Why *fos* doesn't come up..

Partner	Cfos	Cfos	Fosl2	Cfos
	K562	HelaS3	HepG2	GM12878
OA	299	5444	8561	4
OO	330	5703	9119	4
OT	354	6294	10179	7
OF	374	7035	11436	14
nJund_Peaks	759	26074	22441	1715
nPartner_Peaks	5810	6236	19714	1744

Results can be made more robust to data quality by using predicted binding sites (using integrative TF binding models)

Combinatorial Coassociation


Partner TF X (Peak m)
Focus TF (Peak "a")
Partner TF Y
Set of peaks of "A" with X and Y as partners
Partner TF X (Peak m)
Partner TF Z
Set of peaks of "A" with X and Z as partners

- **Does signal from X independently do better than that from X and Y?**
- **Does signal from X and Y do better than X and Z?**

Frequent TF-set Mining

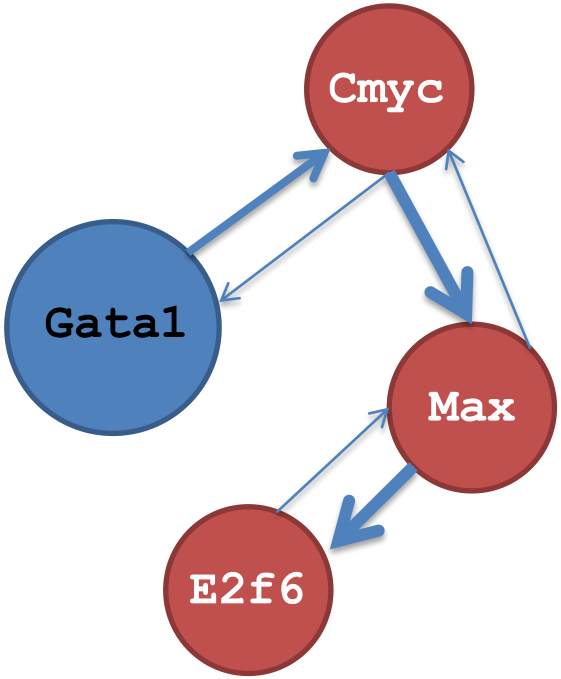
```
Gata1 <- Egata2 Gata2  (70.3, 99.9)
Gata1 <- Ccnt2 Pol2  (70.5, 99.9)
Gata1 <- Ccnt2 Gata2  (70.6, 99.9)
Gata1 <- Tal1 Cmyc  (70.4, 99.9)
Gata1 <- Tal1 Cmyc  (70.4, 99.9)
Gata1 <- Nrsf Pol2  (60.7, 99.9)
Gata1 <- Nrsf Gata2  (60.8, 99.9)
Gata1 <- Max Cmyc  (60.8, 99.9)
Gata1 <- Nrsf Pol2 Gata2  (60.0, 99.9)
Gata1 <- Ccnt2 Tal1 Cmyc  (60.1, 99.9)
```

```
Gata1 <- Tal1 Cmyc  (70.4, 99.9)
Gata1 <- Taf1 Nrsf Max Egata2 Tal1 Cmyc  (20.2, 99.8)
Gata1 <- Efos Cjun P300f4 Hdac2 Nrsf Tal1 Cmyc Pol2 Gata2  (5.8,
100.0)
```

We are also using TF item sets from

- **hierarchical clustering of columns of TF-specific association matrix**
- **Can also use significant global associations (post GSC)**

Regulatory network construction



- Feature importance can determine edge strength for each TF
- This gives bidirectional asymmetric network
- Highlight asymmetric edges & see if that is because of quality or master regulator vs cofactor effect
- Superimpose protein interaction data/annotations
- We can then extract cliques with strong edges as modules