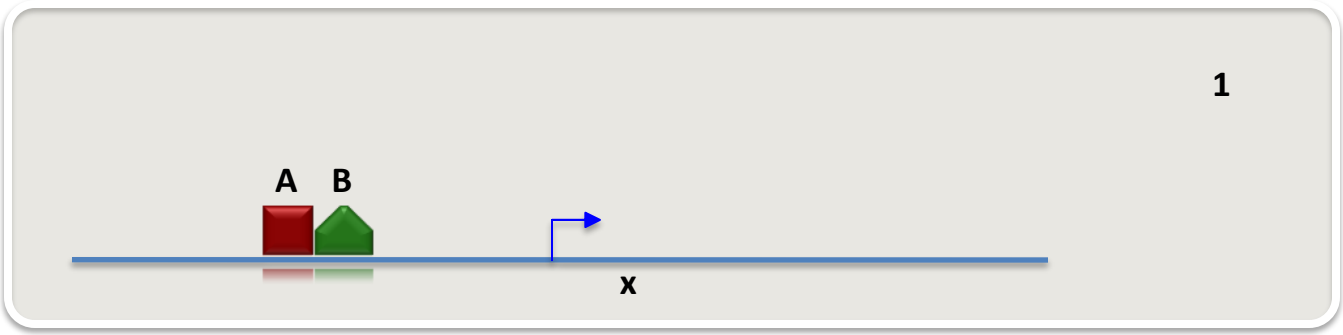
Transcription Factor Coassociation & its Effect

Anshul Kundaje

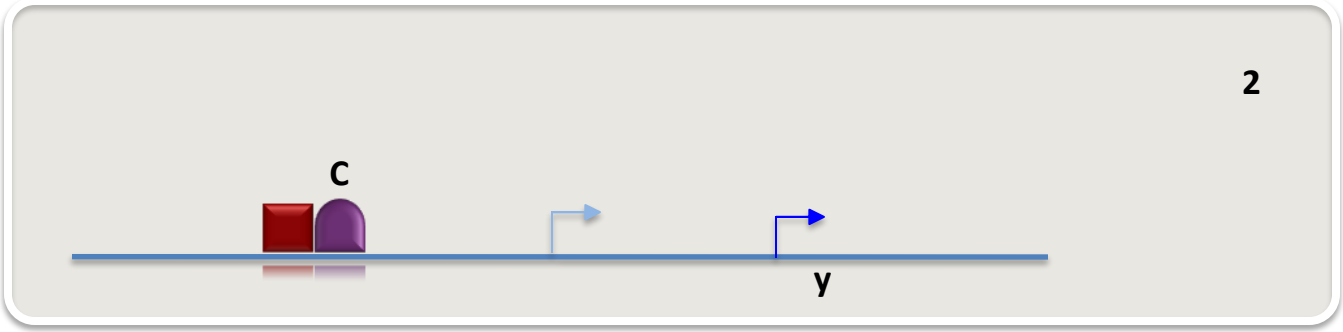
Manoj Hariharan

- Background
- Earlier Work
- Value Addition
- Current – Feb 2011
- Future – March 2011

Logic of Transcription Factor Occupancy



A and B cooperatively activate gene x



A and C cooperatively activate gene y

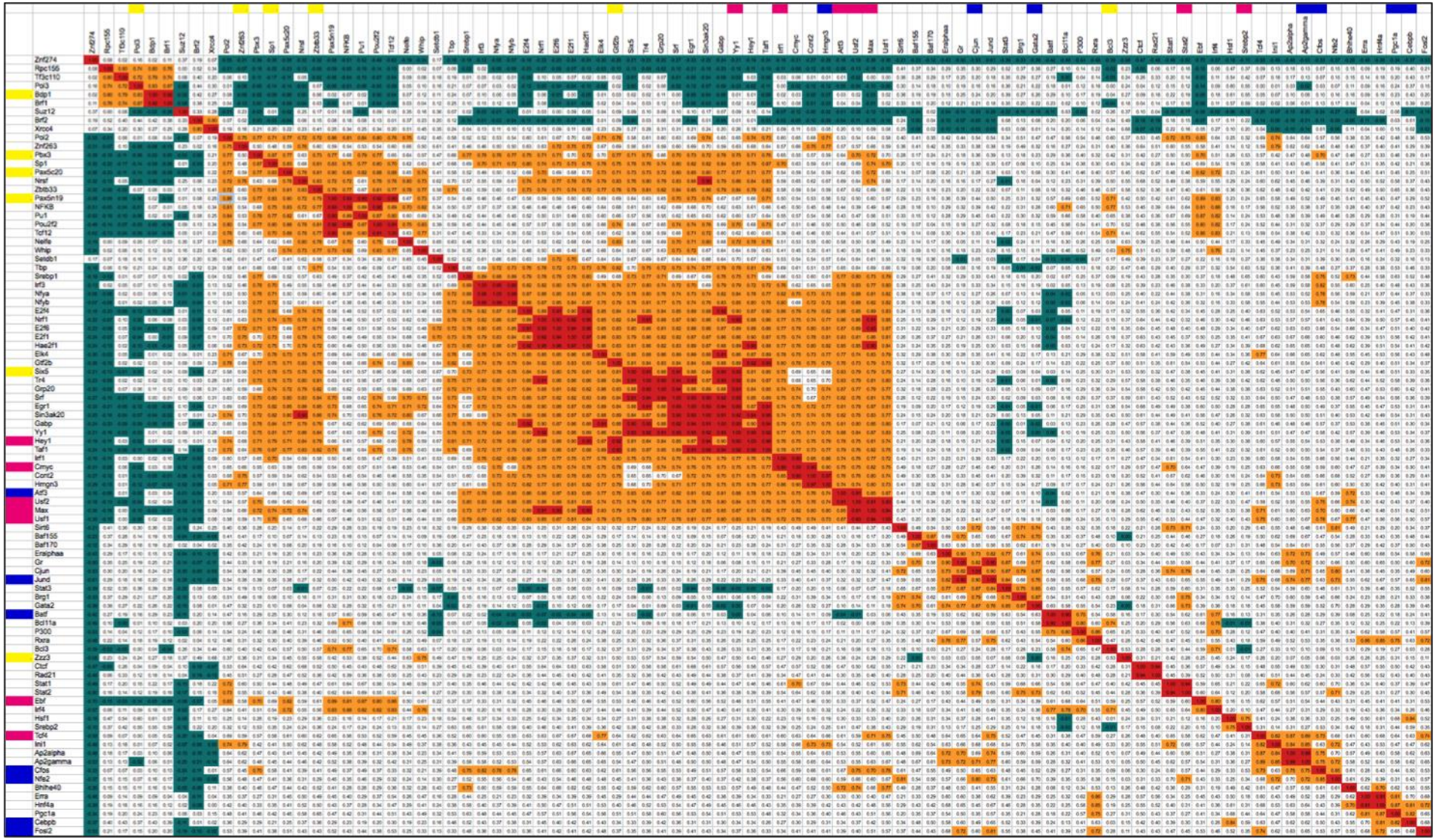
A and C cooperatively repress gene x

A and B cooperatively activate gene x

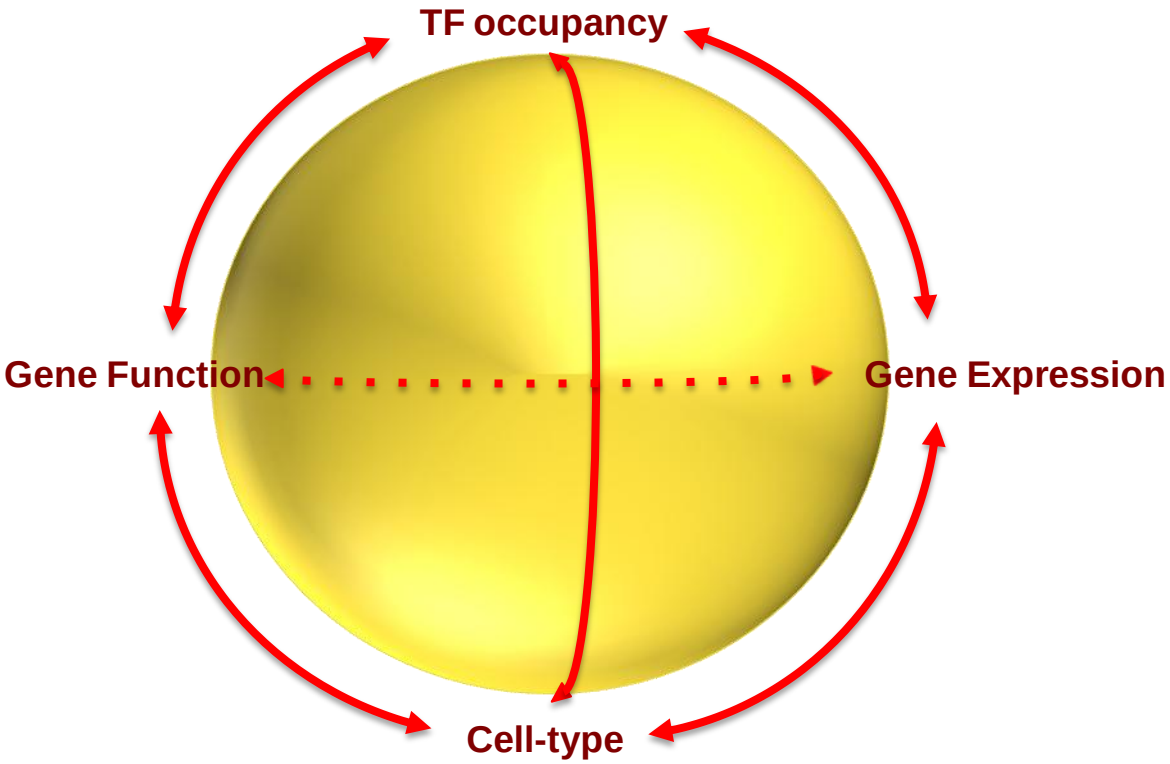
B and C compete

Earlier work

Correlation of co-association among TFs

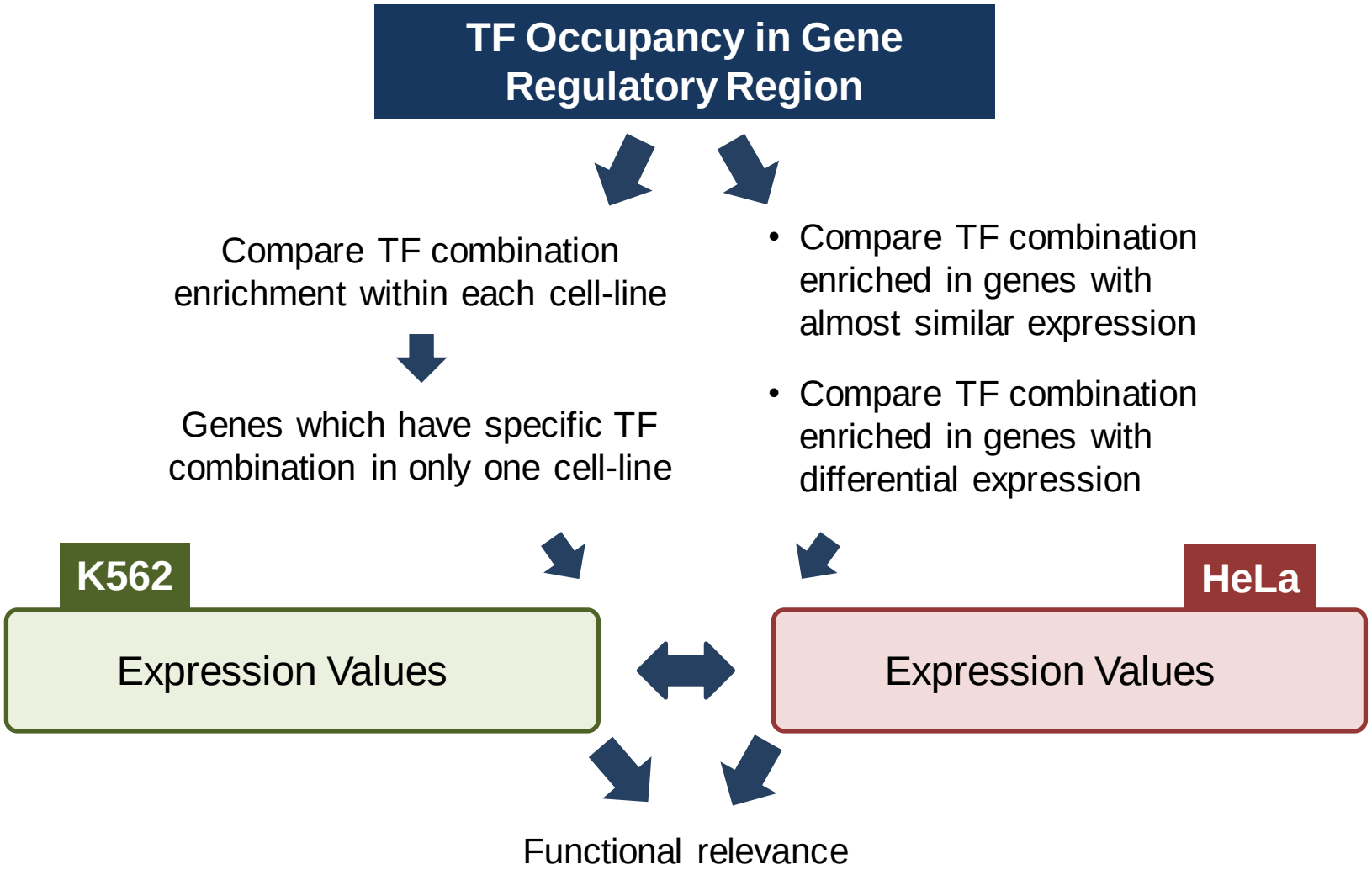


Transcription Factor Occupancy as a "predictor"



Defining the regulatory region for TF Occupancy

Earlier work



Earlier work

- ***E2F6- related combinations are enriched in K562***
- ***Myc, Max, Jun- related combinations are enriched in HeLa***

- ***There are combinations of TFs specific to cell-type***

nTF Combinations present in HeLa, but absent in K562: 985

nTF Combinations present in K562, but absent in HeLa: 983

Example:

The E2F4 – E2F6 combination is present upstream of **158 genes in K562**, but **absent** upstream of any gene in HeLa

In HeLa:

- 16 of 158 genes have E2F4-E2F6 combinations, but with other factors, mainly Max and Pol2
- 7 genes have E2F4, not E2F6 and in combination with other factors, mainly Pol2
- 52 genes have E2F6, not E2F4 and in combination with other factors, mainly Pol2
- 90 genes do not have both E2F4 or E2F6, other combination of factors are present

- Datasets K562: Then,– 42 TFs; Now: 78 TFs
- Use of ranked signals*
- Use of pre-normalization
- Use of biclustering
- Predictive – Discriminant analysis

Use of ranked signals – Why?

Not all peaks are equally “good”
Read-depth varies between datasets

Preview: Initial analysis

Rank Normalize every peak in all datasets based on Signal

Merge datasets where cell-line and TF are same

Get average of the normalized rank, if datasets are merged

Assign peaks to gene regulatory region & get expression value: avRPKM

Clustering – Hierarchical & Biclustering

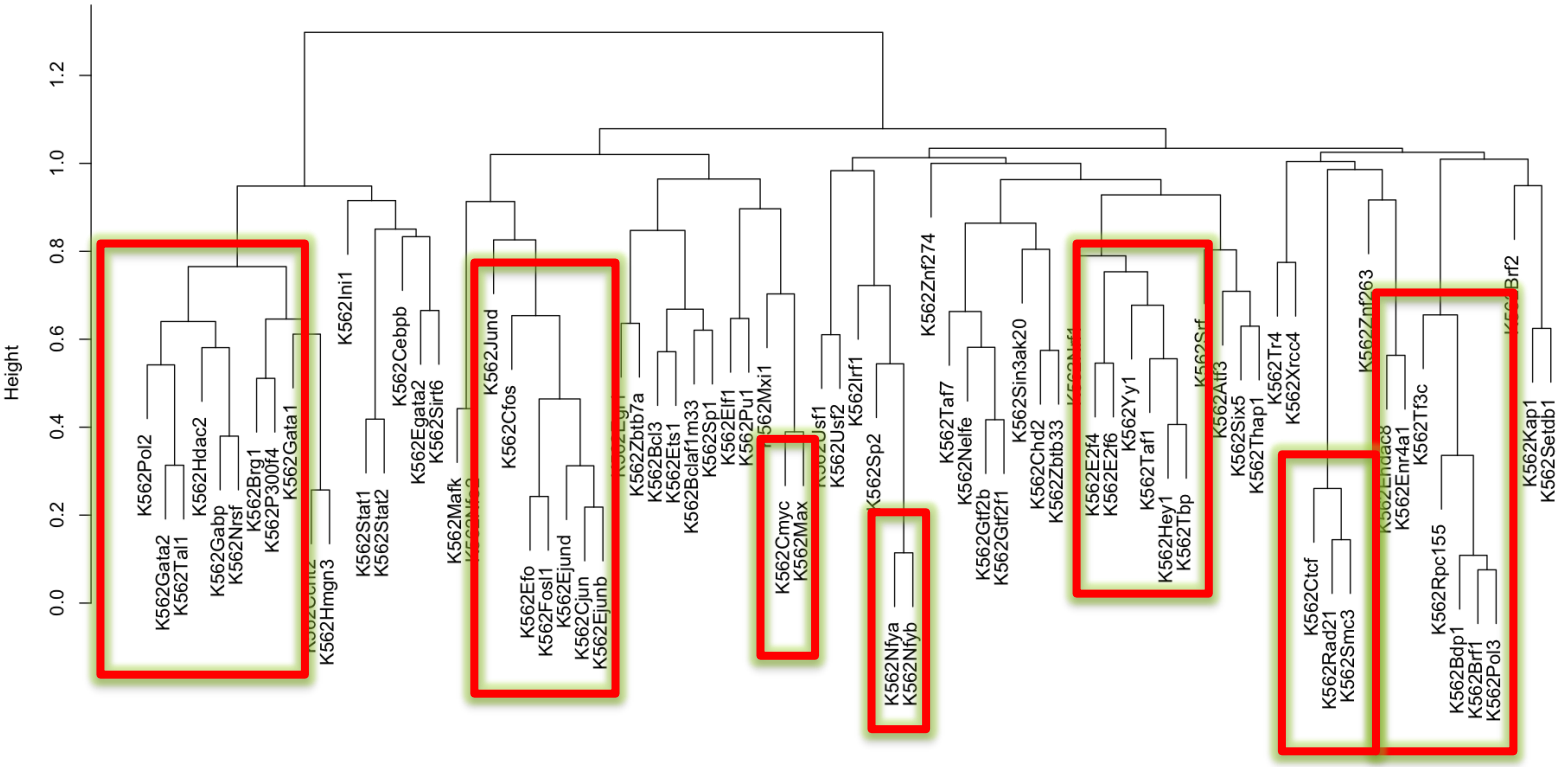
Gata1 – Hierarchical Clustering

First TF-wise, then Peak-wise

a) TF Cluster members

K562Atf3	1	K562Bcl3	2	K562Cfos	6
K562Chd2	1	K562Bclaf1m	2	K562Cjun	6
K562E2f4	1	K562Cmyc	2	K562Efo	6
K562E2f6	1	K562Egr1	2	K562Ejunb	6
K562Gtf2b	1	K562Elf1	2	K562Ejund	6
K562Gtf2f1	1	K562Ets1	2	K562Fosl1	6
K562Hey1	1	K562Max	2	K562Jund	6
K562Nelfe	1	K562Mxi1	2	K562Mafk	6
K562Nrf1	1	K562Pu1	2	K562Nfe2	6
K562Sin3ak2	1	K562Sp1	2		
K562Six5	1	K562Zbtb7a	2	K562Ctcf	7
K562Srf	1			K562Rad21	7
K562Taf1	1	K562Brg1	5	K562Smc3	7
K562Taf7	1	K562Ccnt2	5		
K562Tbp	1	K562Cebpb	5	K562Ehdac8	8
K562Thap1	1	K562Egata2	5	K562Enr4a1	8
K562Yy1	1	K562Gabp	5	K562Znf263	8
K562Zbtb33	1	K562Gata1	5		
		K562Gata2	5	K562Irf1	9
K562Bdp1	3	K562Hdac2	5	K562Nfya	9
K562Brf1	3	K562Hmgn3	5	K562Nfyb	9
K562Pol3	3	K562Ini1	5	K562Sp2	9
K562Rpc155	3	K562Nrsf	5	K562Usf1	9
K562Tf3c	3	K562P300f4	5	K562Usf2	9
		K562Pol2	5		
K562Brf2	4	K562Sirt6	5	K562Tr4	10
K562Kap1	4	K562Stat1	5	K562Xrcc4	10
K562Setdb1	4	K562Stat2	5		
		K562Tal1	5	K562Znf274	11

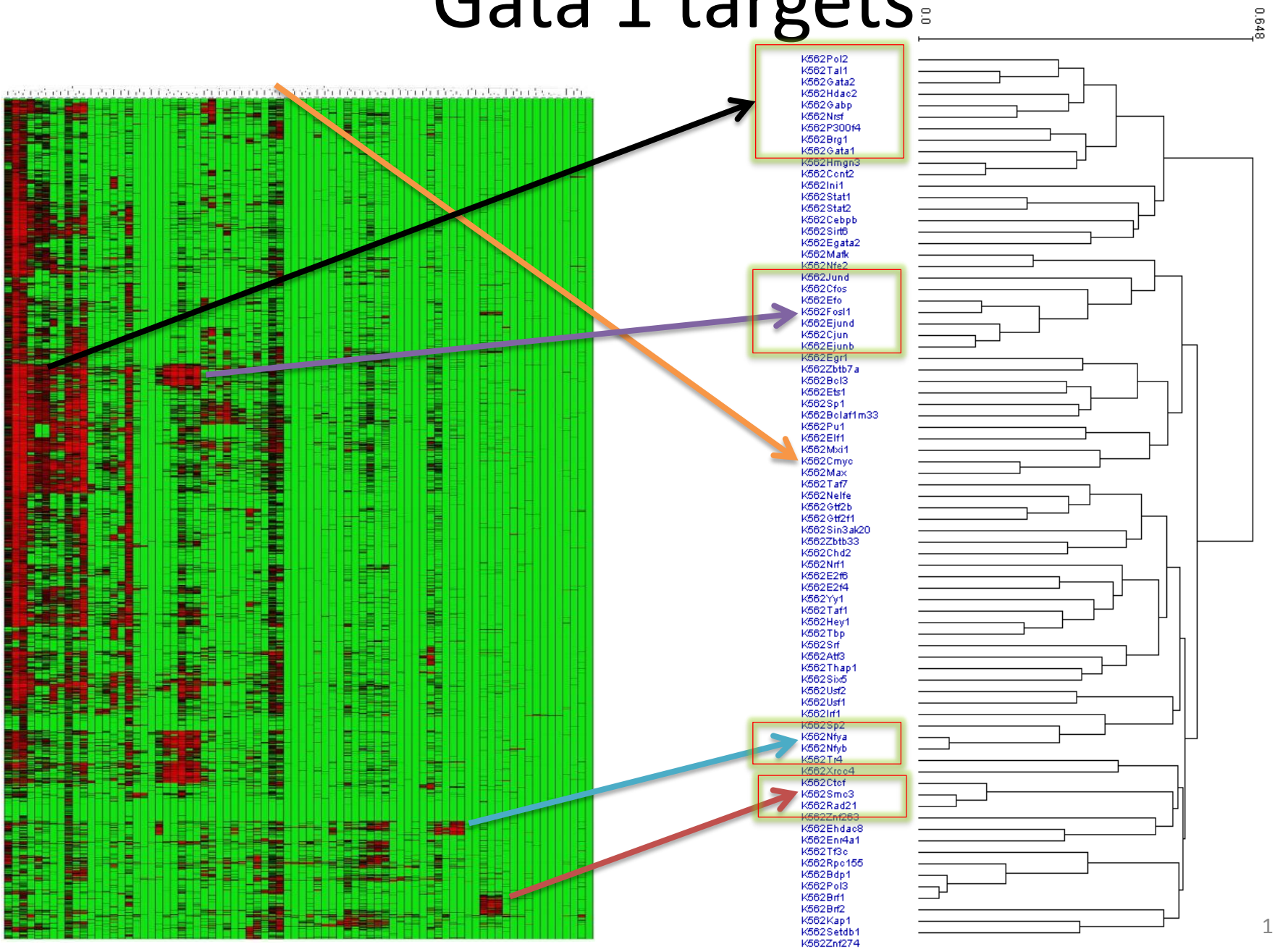
b) Cluster Dendrogram



c) Clustered Peaks

Cluster	nPeaks
1	883
2	1207
3	108
4	93
5	90
6	241
7	93
8	33
9	20
10	5
11	9

Gata 1 targets



- **Map nearest gene to peak in the cluster**
- **Average Expression**
- **Functional Annotation**

Cluster	nPeaks	nGenes	Av(avRPKM)
1	883	352	11.1
2	1207	462	9.8
3	108	47	11.8
4	93	35	21.0
5	90	36	9.5
6	241	74	5.1
7	93	78	27.0
8	33	30	18.9
9	20	12	23.3
10	5	4	11.2
11	9	6	32.0

Nrsf Effect ?

Gata1 – Sub-cluster 1 – functional categories

Ontology	# Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
GO Molecular Function	No results meet your chosen criteria.												
GO Biological Process	negative regulation of striated muscle cell differentiation	50	2.4584e-4	3.5253e-2	9.2318	5	0.57%	9	3.3987e-2	12.3615	4	4	0.28%
GO Cellular Component	No results meet your chosen criteria.												
MSigDB Cancer Neighborhood	Neighborhood of MAP2K3	5	1.2013e-6	1.0259e-4	3.9934	18	2.04%	1	1.9815e-2	2.8165	18	79	1.27%
	Neighborhood of SPTA1	14	2.0302e-5	6.1922e-4	3.3729	17	1.93%	4	4.9138e-2	2.4435	17	86	1.20%
	Neighborhood of RAD23A	15	2.2888e-5	6.5153e-4	3.4985	16	1.81%	3	2.7850e-2	2.7094	16	73	1.13%
	Neighborhood of CDC27	20	7.0909e-5	1.5139e-3	3.5069	14	1.59%	2	1.9017e-2	3.1466	14	55	0.98%
PANTHER Pathway	PDGF signaling pathway	2	8.5285e-5	5.9700e-3	2.1509	31	3.51%	2	2.2399e-2	2.1703	23	131	1.62%
	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	3	1.0816e-4	5.0476e-3	2.1867	29	3.28%	1	3.5081e-4	2.7593	25	112	1.76%
	Histamine H1 receptor mediated signaling pathway	4	2.3269e-4	8.1442e-3	3.1205	14	1.59%	4	5.4404e-2	2.9432	10	42	0.70%
	Alpha adrenergic receptor signaling pathway	6	1.0870e-3	2.5364e-2	3.6140	9	1.02%	3	6.0072e-2	3.9332	7	22	0.49%
Pathway Commons	S1P1 pathway	11	8.2322e-5	1.0477e-2	3.0050	17	1.93%	14	7.1682e-3	3.2048	14	54	0.98%
	Sphingosine 1-phosphate (S1P) pathway	12	1.3601e-4	1.5868e-2	2.2699	26	2.94%	23	3.2436e-2	2.1814	21	119	1.48%
	Signaling events mediated by HDAC Class II	23	4.3617e-4	2.6550e-2	3.0773	13	1.47%	26	4.5173e-2	3.1696	10	39	0.70%
BioCyc Pathway	No results meet your chosen criteria.												

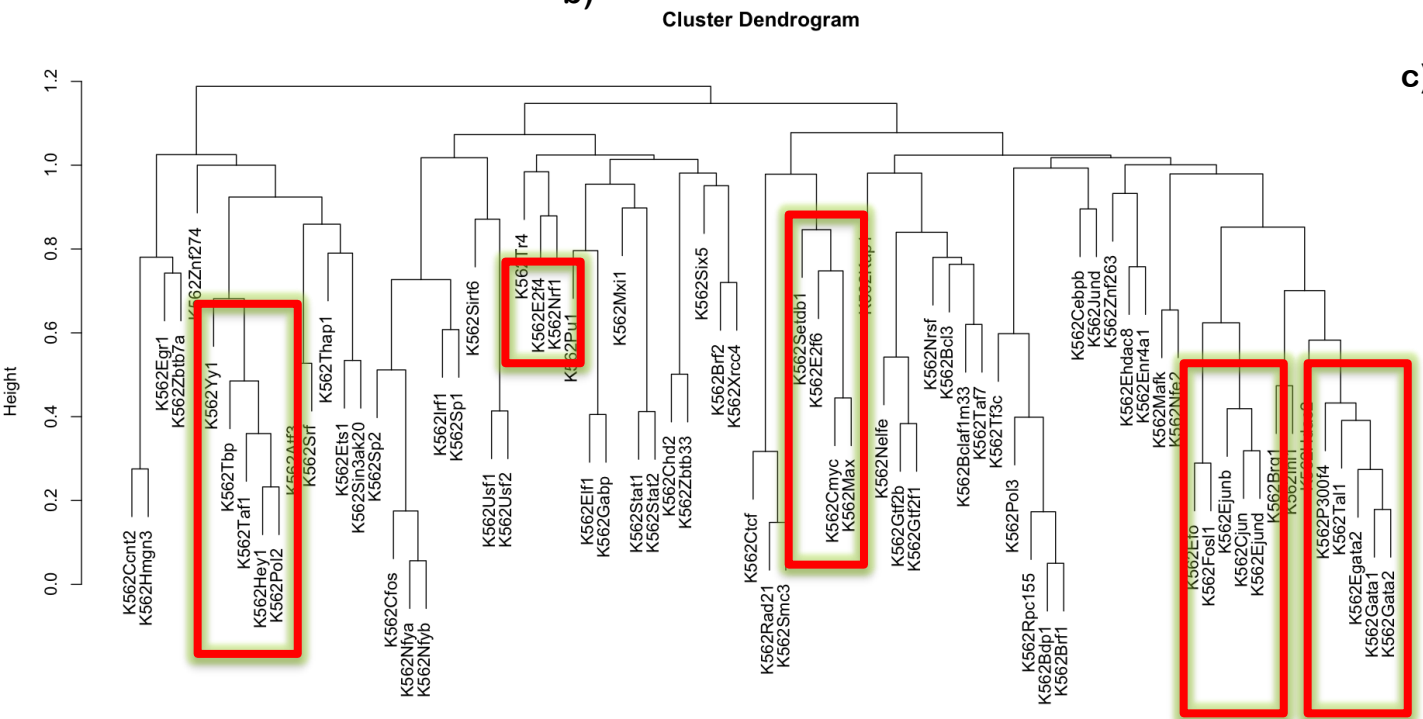
E2f4 – Hierarchical Clustering

First TF-wise, then Peak-wise

a) TF Cluster members

K562Atf3	1	K562Bcl3	2	K562Bdp1	3
K562Ets1	1	K562Bclaf1m	2	K562Brf1	3
K562Hey1	1	K562Gtf2b	2	K562Cebpb	3
K562Pol2	1	K562Gtf2f1	2	K562Jund	3
K562Sin3ak2	1	K562Kap1	2	K562Pol3	3
K562Srf	1	K562Nelfe	2	K562Rpc155	3
K562Taf1	1	K562Nrsf	2	K562Tf3c	3
K562Tbp	1	K562Taf7	2		
K562Thap1	1			K562Ccnt2	6
K562Yy1	1	K562Brg1	5	K562Egr1	6
K562Znf274	1	K562Cjun	5	K562Hmgn3	6
		K562Efo	5	K562Zbtb7a	6
K562Brf2	4	K562Egata2	5		
K562Chd2	4	K562Ehdac8	5	K562Cmyc	8
K562Six5	4	K562Ejunb	5	K562Ctcf	8
K562Xrcc4	4	K562Ejund	5	K562E2f6	8
K562Zbtb33	4	K562Enr4a1	5	K562Max	8
		K562Fosl1	5	K562Rad21	8
K562Cfos	7	K562Gata1	5	K562Setdb1	8
K562Irf1	7	K562Gata2	5	K562Smc3	8
K562Nfya	7	K562Hdac2	5		
K562Nfyb	7	K562Ini1	5	K562E2f4	9
K562Sp1	7	K562Mafk	5	K562Nrf1	9
K562Sp2	7	K562Nfe2	5	K562Tr4	9
		K562P300f4	5	K562Elf1	10
K562Sirt6	11	K562Tal1	5	K562Gabp	10
K562Usf1	11	K562Znf263	5	K562Mxi1	10
K562Usf2	11			K562Pu1	10
				K562Stat1	10
				K562Stat2	10

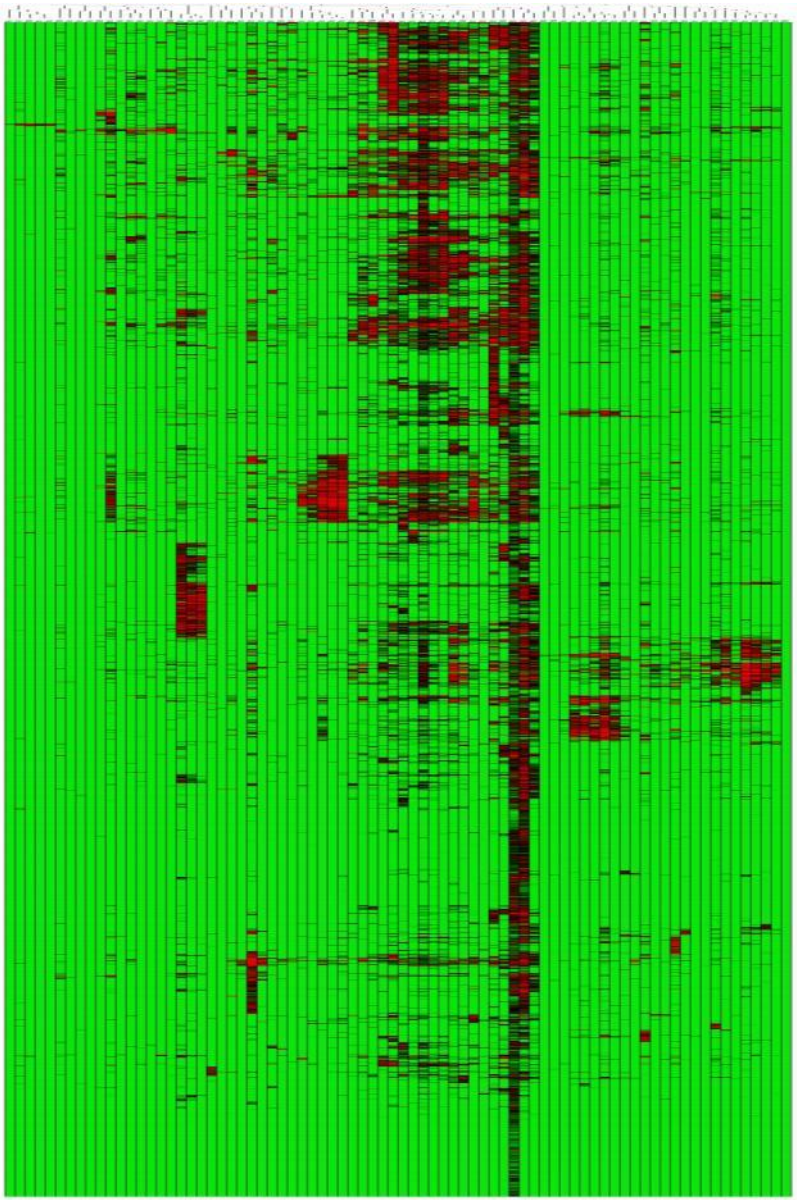
b)

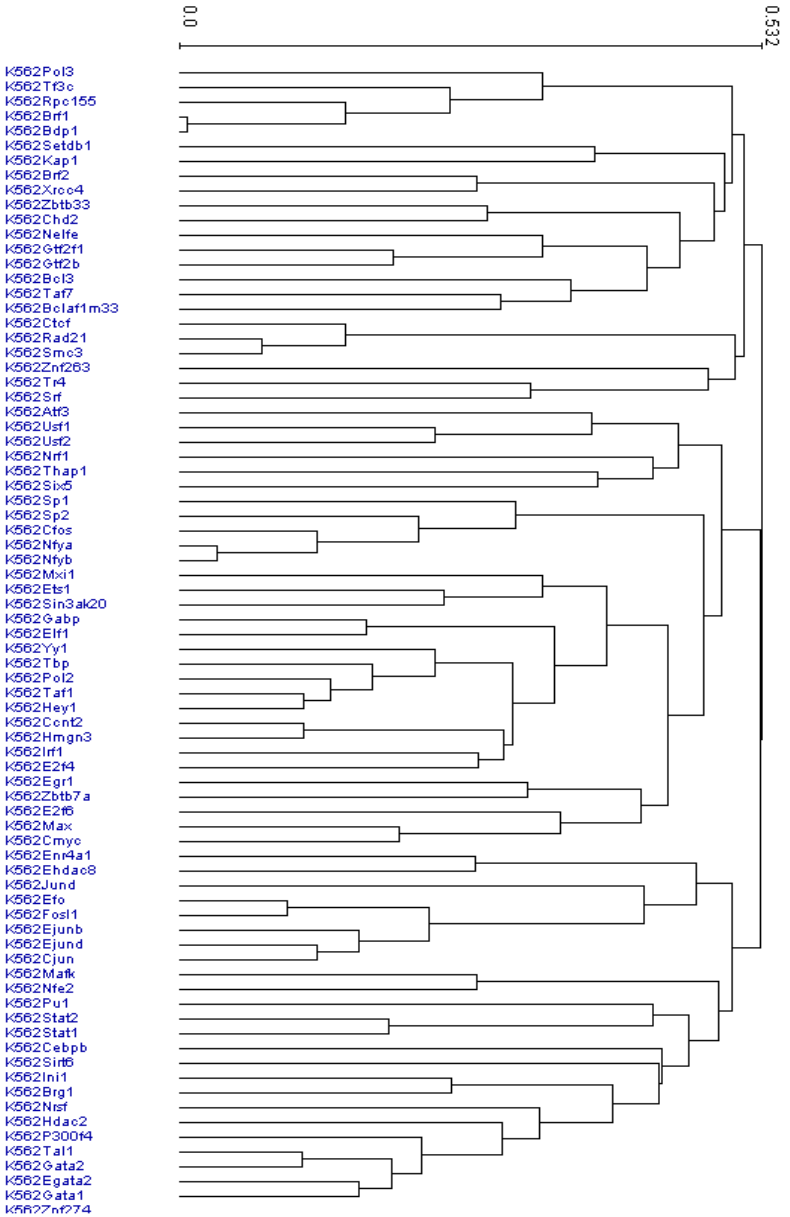


c) Clustered Peaks

Cluster	nPeaks
1	836
2	2402
3	1131
4	143
5	75
6	125
7	64
8	30
9	25
10	29
11	3

E2f6 targets





- **Map nearest gene to peak in the cluster**
- **Average Expression**
- **Functional Annotation**

Nrsf Effect ?

Cluster	nPeaks	nGenes	Av(av(RPKM))
1	836	729	20.3
2	2402	2055	15.5
3	1131	815	11.6
4	143	131	24.1
5	75	47	30.2
6	125	51	13.6
7	64	61	13.8
8	30	23	49.9
9	25	7	7.3
10	29	22	29.2
11	3	2	8.6

E2f4 – Sub-cluster 3 – functional categories

Ontology	# Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
GO Biological Process	one-carbon metabolic process	4	4.1643e-9	7.4645e-6	2.8805	41	3.63%	70	1.9329e-2	2.1262	26	150	1.81%
	covalent chromatin modification	5	2.0931e-8	3.0015e-5	2.7144	41	3.63%	77	3.3304e-2	2.0444	26	156	1.81%
	base-excision repair	39	2.6597e-4	4.8898e-2	4.4182	9	0.80%	64	1.5137e-2	4.2461	9	26	0.63%
GO Cellular Component	No results meet your chosen criteria.												
Mouse Phenotype	No results meet your chosen criteria.												
PANTHER Pathway	No results meet your chosen criteria.												
Pathway Commons	E2F transcription factor network	1	2.7202e-10	3.8083e-7	4.3268	28	2.48%	9	1.5470e-4	3.4786	19	67	1.33%
	HIV Life Cycle	2	5.7963e-10	4.0574e-7	3.9327	30	2.65%	8	1.7125e-4	2.8931	25	106	1.74%
	Reverse Transcription of HIV RNA	14	1.1310e-6	1.1310e-4	8.9664	9	0.80%	12	2.6507e-4	8.5866	7	10	0.49%
	Minus-strand DNA synthesis	14	1.1310e-6	1.1310e-4	8.9664	9	0.80%	12	2.6507e-4	8.5866	7	10	0.49%
	Plus-strand DNA synthesis	14	1.1310e-6	1.1310e-4	8.9664	9	0.80%	12	2.6507e-4	8.5866	7	10	0.49%
	Uncoating of the HIV Virion	14	1.1310e-6	1.1310e-4	8.9664	9	0.80%	12	2.6507e-4	8.5866	7	10	0.49%
	DNA Repair	18	1.1461e-6	8.9144e-5	2.5638	34	3.01%	29	1.8329e-3	2.1218	32	185	2.23%
	Nuclear import of Rev protein	25	1.4358e-6	8.0405e-5	5.4277	13	1.15%	26	1.9697e-3	4.0889	11	33	0.77%
	Rev-mediated nuclear export of HIV-1 RNA	25	1.4358e-6	8.0405e-5	5.4277	13	1.15%	26	1.9697e-3	4.0889	11	33	0.77%
	Interactions of Rev with host cellular proteins	25	1.4358e-6	8.0405e-5	5.4277	13	1.15%	26	1.9697e-3	4.0889	11	33	0.77%
	Mitotic G2-G2/M phases	28	1.5454e-6	7.7269e-5	2.0578	52	4.60%	35	5.0502e-3	1.7813	44	303	3.07%
	Mitotic G1-G1/S phases	28	1.5454e-6	7.7269e-5	2.0578	52	4.60%	35	5.0502e-3	1.7813	44	303	3.07%
	Mitotic M-M/G1 phases	28	1.5454e-6	7.7269e-5	2.0578	52	4.60%	35	5.0502e-3	1.7813	44	303	3.07%
	DNA Replication	28	1.5454e-6	7.7269e-5	2.0578	52	4.60%	35	5.0502e-3	1.7813	44	303	3.07%
	Basigin interactions	32	1.9929e-6	8.7188e-5	16.9544	6	0.53%	95	3.4723e-2	6.1333	4	8	0.28%
	Late Phase of HIV Life Cycle	33	2.2630e-6	9.6006e-5	3.3795	21	1.86%	53	1.3772e-2	2.3489	18	94	1.26%
	Resolution of AP sites via the single-nucleotide replacement pathway	36	3.2000e-6	1.2444e-4	11.6904	7	0.62%	91	3.5207e-2	4.0889	6	18	0.42%
	Base-free sugar-phosphate removal via the single-nucleotide replacement pathway	36	3.2000e-6	1.2444e-4	11.6904	7	0.62%	91	3.5207e-2	4.0889	6	18	0.42%
	Resolution of Abasic Sites (AP sites)	38	3.2959e-6	1.2143e-4	11.6373	7	0.62%	103	4.2359e-2	3.8737	6	19	0.42%
	Removal of DNA patch containing abasic residue	38	3.2959e-6	1.2143e-4	11.6373	7	0.62%	103	4.2359e-2	3.8737	6	19	0.42%
BioCyc Pathway	No results meet your chosen criteria.												

Value addition

Use of pre-processed signals – Why?

Not all peaks are equally “good”
Read-depth varies between datasets

Approaches for Pre-processing

1. Raw-signal
2. Rank Normalized Raw-signal
3. Z-score of Raw-signal
4. Robust Z-score of Raw-signal
5. Quantile Binning of Raw-signal

Raw-signal Matrix

Stage 1: Generate TF association rules

- Get n(peaks) of any TF overlapping with each peak of the candidate TF
- Generate matrix

Stage 2: Generate TF occupancy containing "raw" signal values

- Identify occupancy of TFs in promoter region, peak-wise
- For every TF peak in the region, obtain the normalized signal at 99^{th} quantile
- Aggregate signals if more than one peak of same TF is found (sum of signals)
- If none of the TFs have a peak in the region, assign signal of genomic region

Rank Normalized Raw-signal Matrix

- Convert all raw-signal values to lie between 0 and 1, TF-wise
- Generate matrix

Z-Score of Raw-signal Matrix

- Convert all raw-signal values to z-score values, TF-wise
- Generate matrix

Robust Z-Score of Raw-signal Matrix

MAD = median (| x – median |) ; MAD = Median Absolute Deviation
Robust z-score = (x – median) / MAD

- Convert all raw-signal values to robust z-score values, TF-wise
- Generate matrix

Quantile Binning of Raw-signal Matrix

- Get 10 bins based on quantile values of raw-signal, TF-wise
- Replace raw-signal values with bin number (0 to 10)
- Generate matrix

Biclustering – why?

Certain combination of TFs would be co-associated only in specific sets of peak regions

Predictive model for gene expression

- Build generalized sparse linear models OR non-linear boosted decision trees
  - Y = F(X)
  - Y = expression of target genes of a particular TF
  - X = all TFs binding affinity in the regulatory regions of above genes

- Can generate general rules for TF coassociation
  - eg., TF1 = 2*TF6 + 2*TF7 + 2*TF8 + 2*TF9 + 2*TF10 + 4*TF15 + 4*TF16 + 6*TF26

- Can build predictive model
  - eg., given a cell-type, predict expression of genes in specific functional classes

Acknowledgement

Michael Snyder
Mark Gerstein

"Elements" for inputs during earlier presentations