FunSeq3: design, plan and results

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

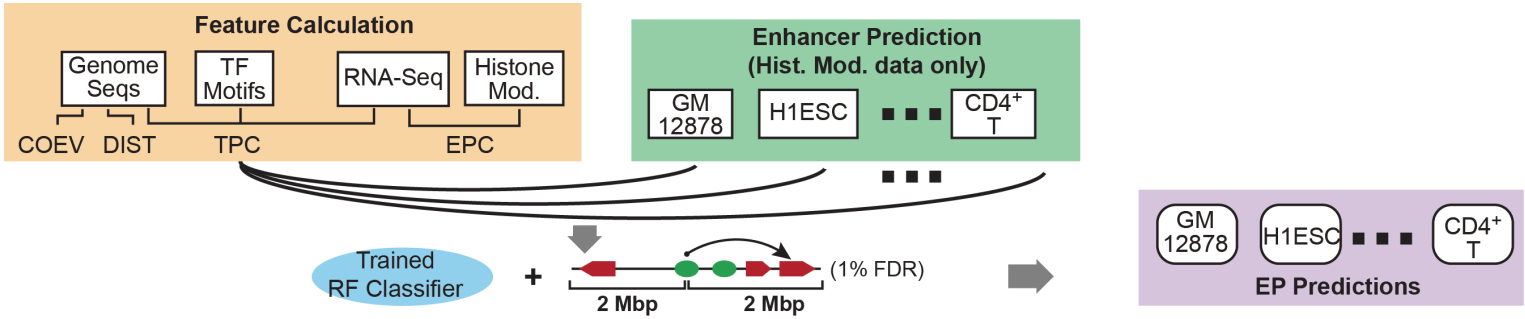
loushaoke@gmail.com

July 29, 2015

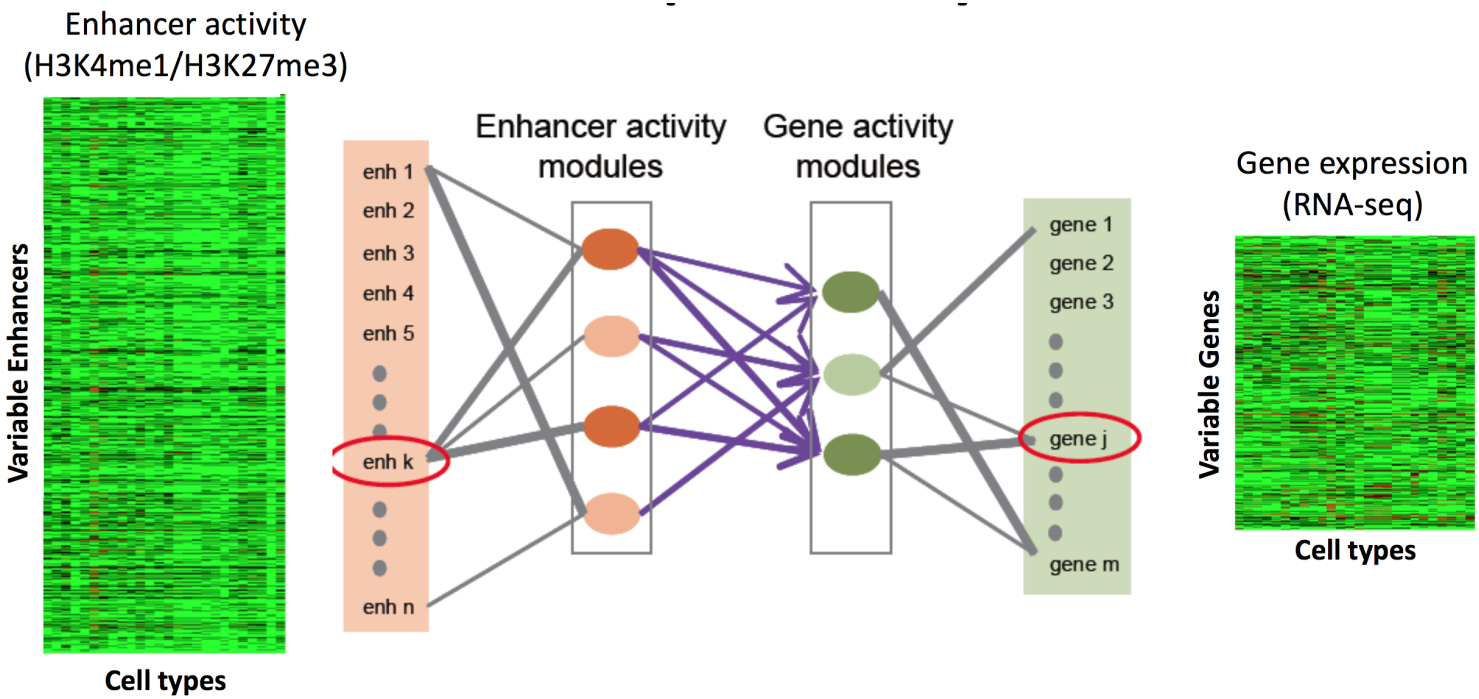
Yale

State-of-The-Art

- Linear correlation between enhancer marks and gene expression
- Random Forest, using ChIA-PET data



- mixed-membership problem



Yale

Enhancer-gene linkages

- why it forms 3D struct?
- All distal interaction need a 3D structure
- how to form 3D structure? Two factors:
  - capability of chromatin structure, not too loose and too tight. or special chromatin structure?
  - mediator and cofactors.

Yale

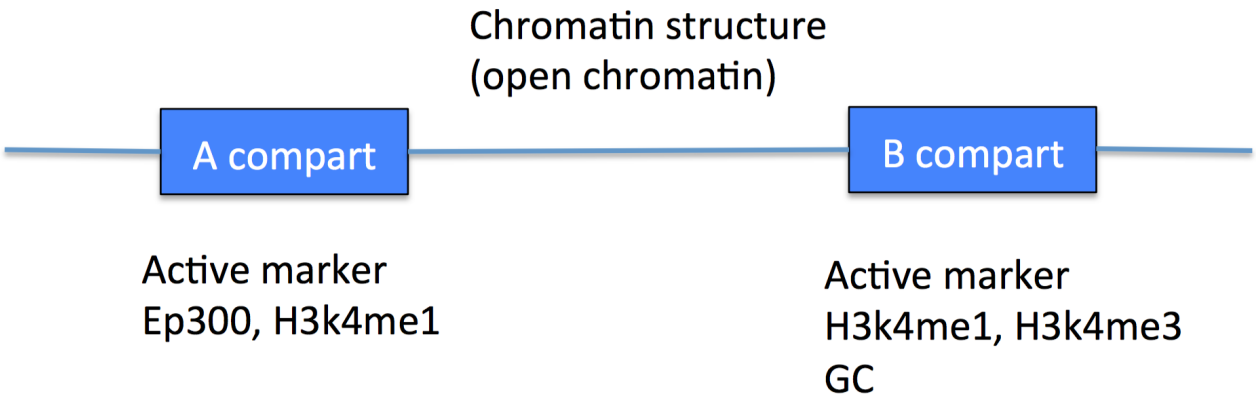
Hypothesis

1. enhancer-gene linkage doesnot randomly happpened. The set of linkages is a collection of linkages from all cell lines/tissues.
2. tissue-specific effect follow the same mechanism to form 3D struct: some tissue don't have the same linkage because the changes of associated factors.
3. the linkage can have positive or negative effect (activation or repression)

How to find the associated features?

Yale

Features

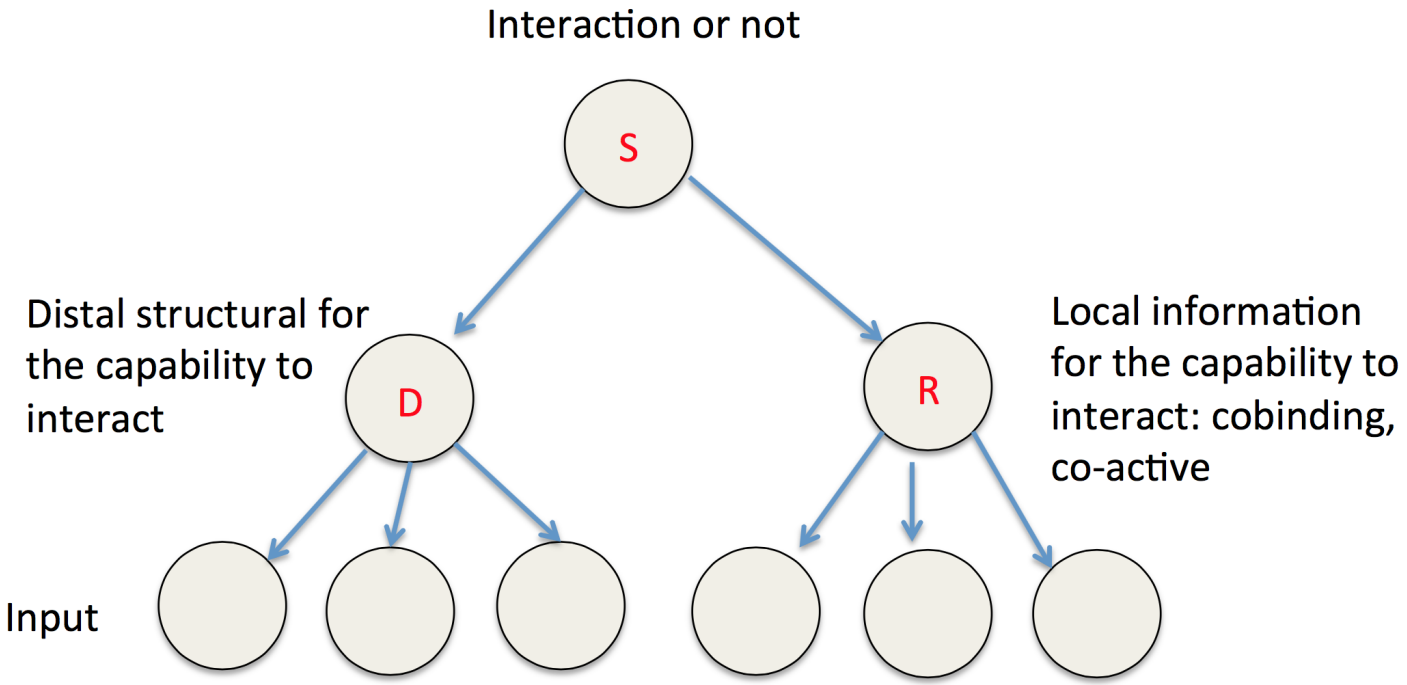


A compartment: The leftmost interaction region
B compartment: The rightmost interaction region
Distal(determine whether can fold): structural information (DNase signal)
Local (determine whether can co-bind by mediators/factors): H3k4me1, GC contents, Co-evolution(sequence and synteny), TF motif, footprint data?
Gene expression

Yale

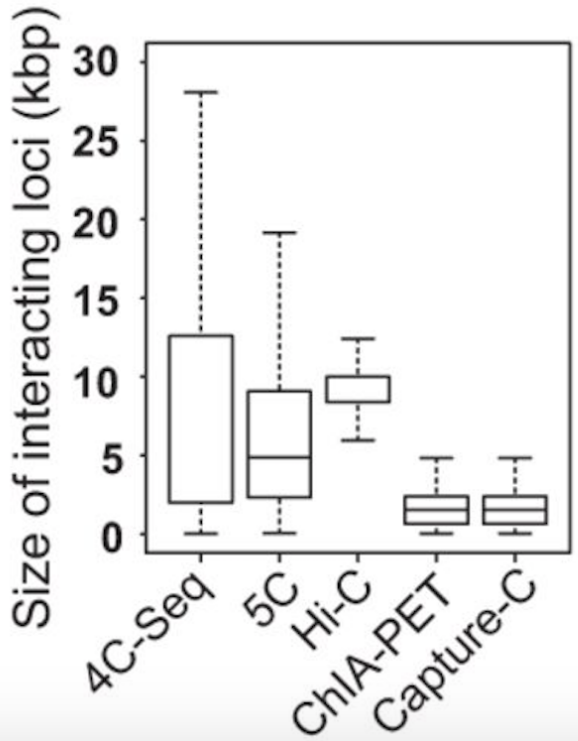
Model

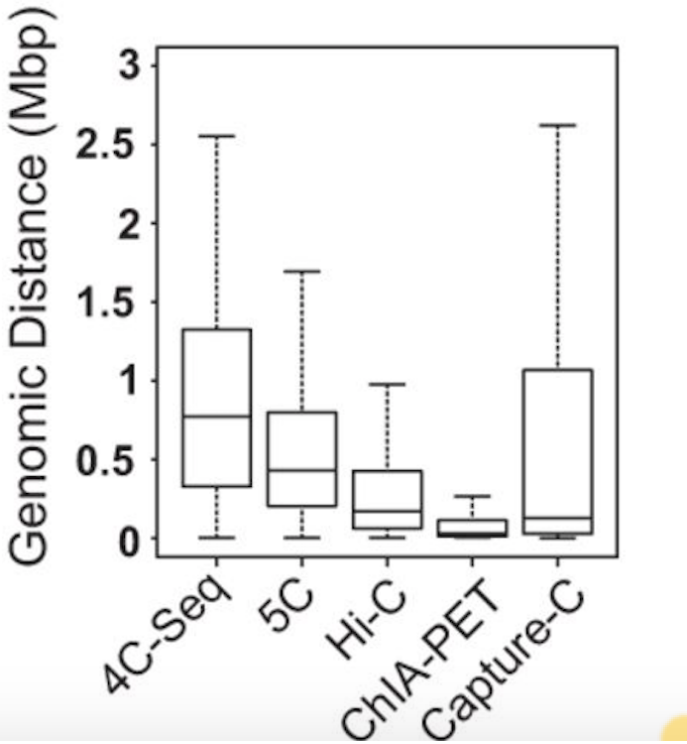
(Not a real model)



Dataset

Interaction map:
4D map: http://4dgenome.int-med.uiowa.edu/Download.html





The number of interactions for HiC, ChIA-PET and 5C

GM12878	H1ESC	HELA	IMR90	K562	MCF7
1177	1221	1849	1114640	66516	64487

Use all predicted enhancer-gene pairs from Wang et al.
ChIA-PET data for: MCF7 K562

Yale

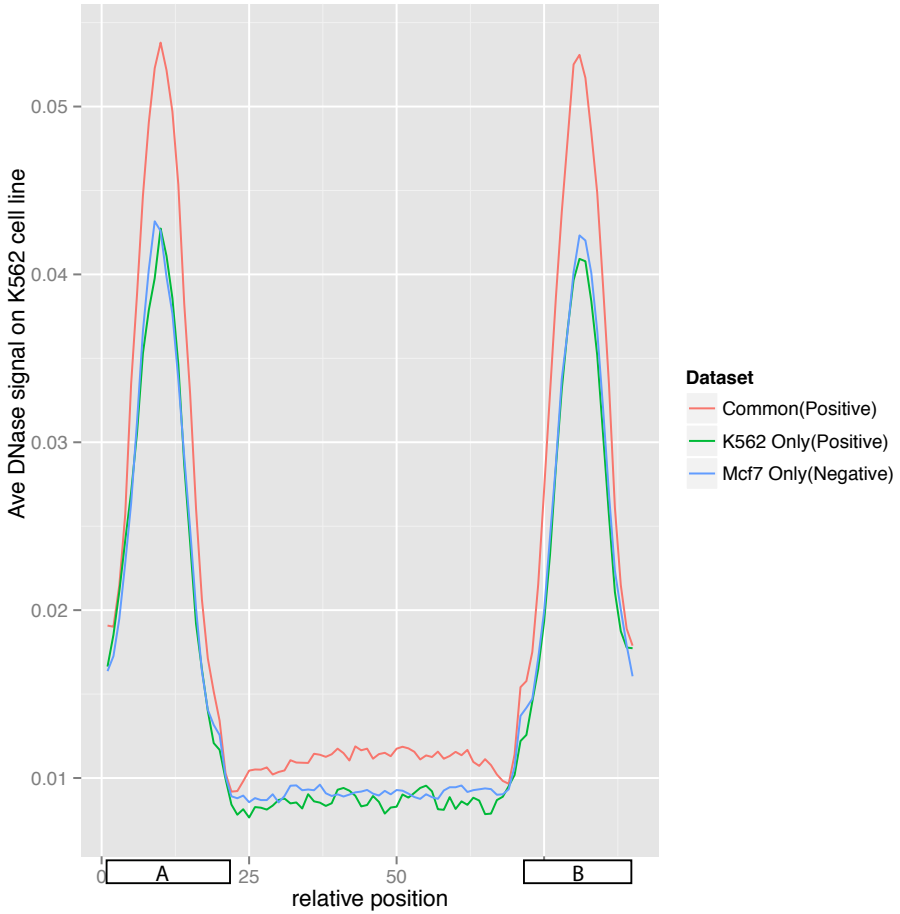
Define the positive and negative set

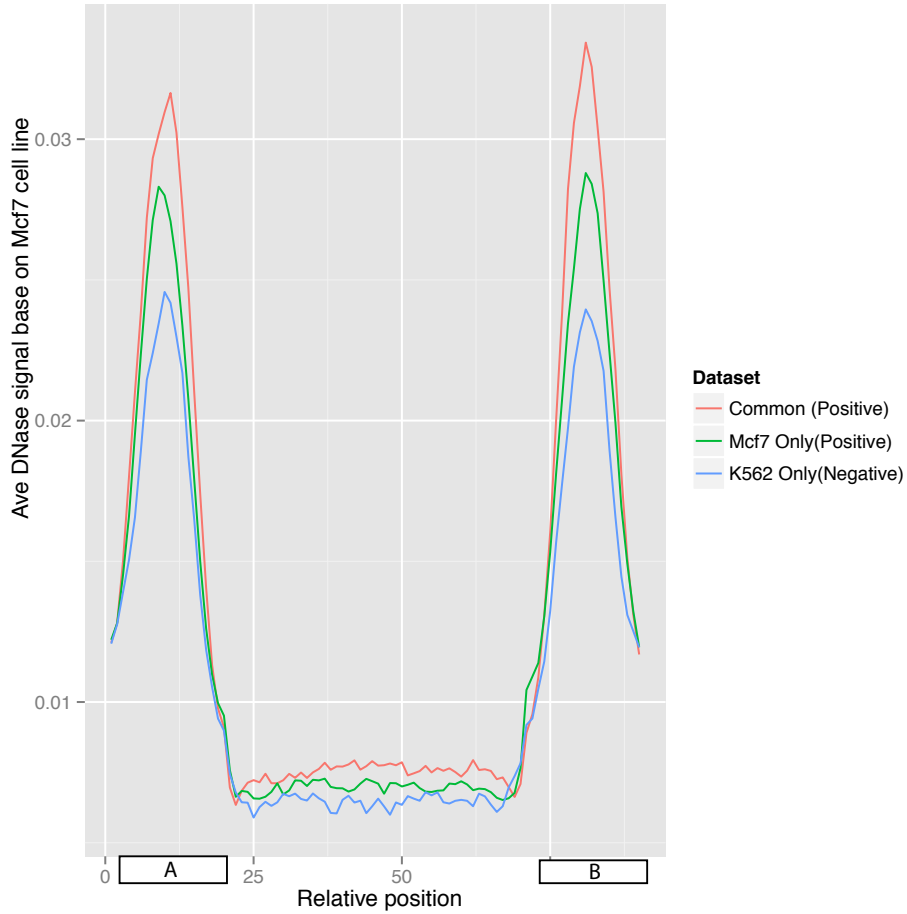
1. Take the intersection of predicted enhancer-gene linkage and ChIA-PET interactions to define positive set
2. The linkage prediction algorithm should not consider 3d interaction: linear correlation or mix-membership relationship
3. Use ChIA-PET for Mcf7 and K562
4. Define datasets:
   Mcf.K562.com(Positive set) : Interaction overlaps between K562 and Mcf7
   Mcf or K562 only: Only have interaction in one cell cell, as Positive set for itself, Negative set for the other.
   For example: Mcf7-only interaction is positive for Mcf7 cell but to be treated as negative set for K562.

Yale

Chromatin openness

A,B are group into 20 equal-size bins; The intra-region between AB are split into 50 equal-size bins, and then compare the average signal for Postive , cell-specific interaction(Positive) and negative set. (K562 specific interactions are thought as Mcf7-Negative, vice versa)





The negaive set might contain positive set because of the ChIA-PET sensitivity.

Future works

- extract other related features, especially for enhancer-gene local co-binding features
- build models: bayesian network model, and also try deep learning and other data mining algorithms

Any question or advices?

Yale

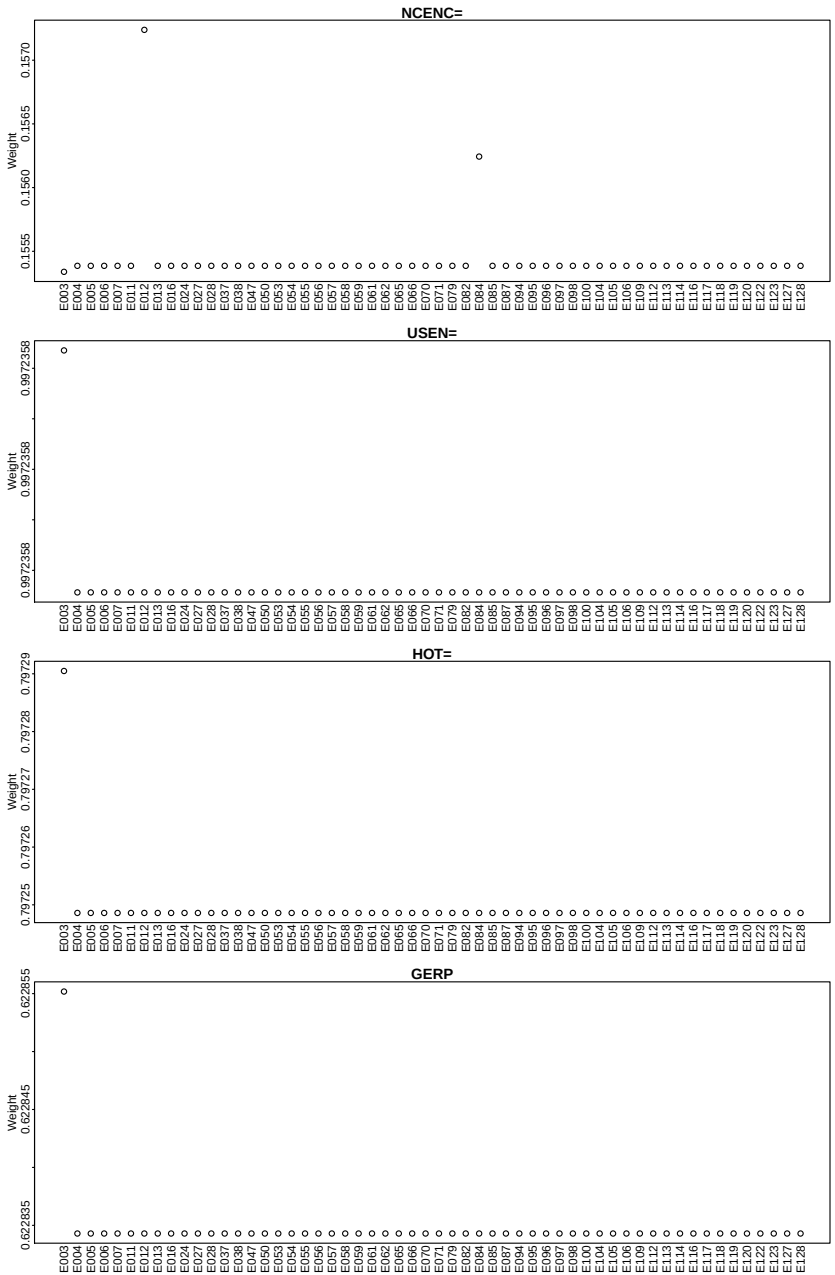
Method

Tissue-specific enhancer-gene linkage:

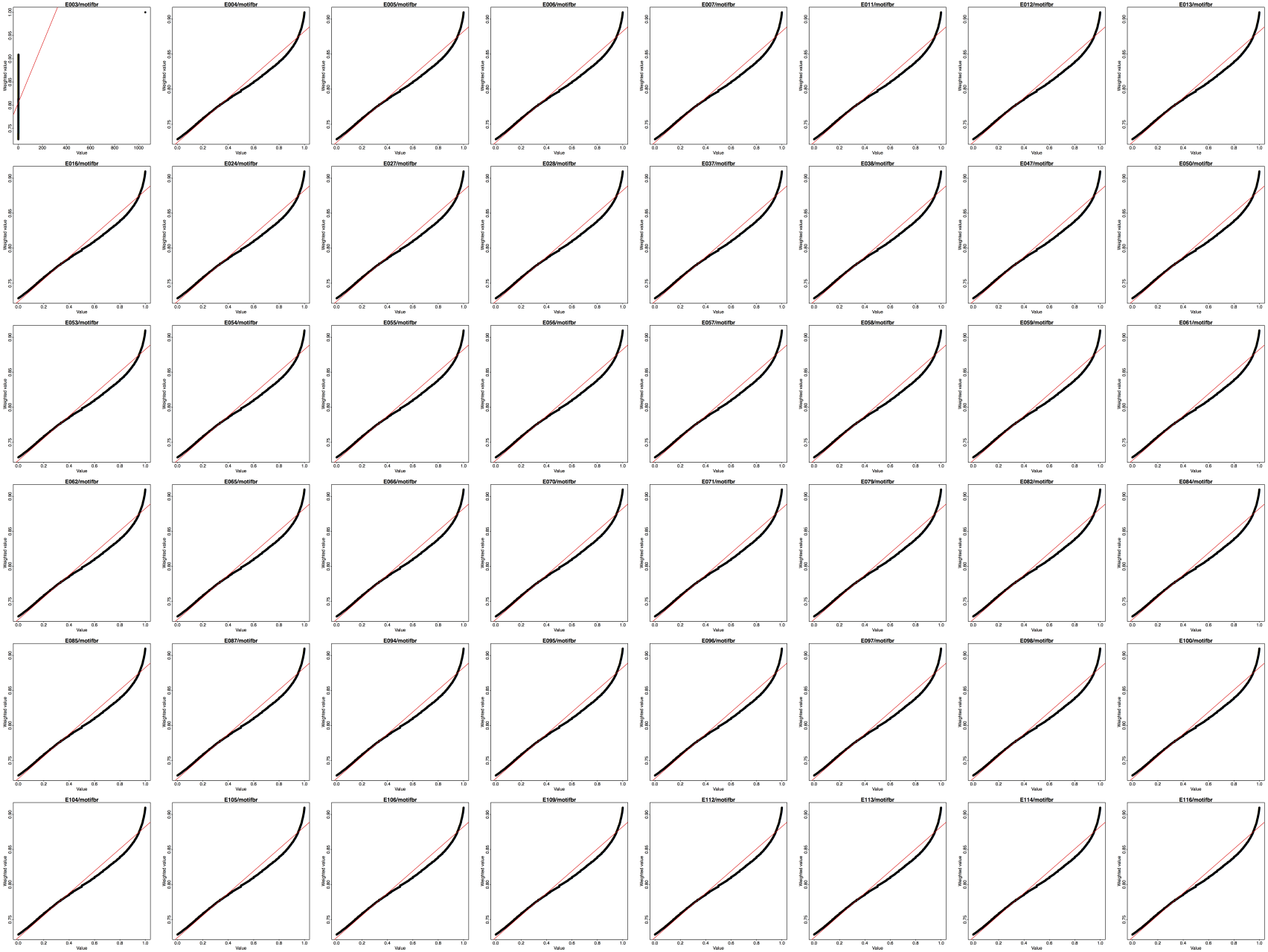
- Data from "Learning three-dimensional regulation of gene expression" (Jianrong Wang, Manolis et al), contains enhancer-gene linkage from 56 tissues.
- Variants Sampling from 1KG
- FunSeq 2.1.2

Yale

Weights for discrete features

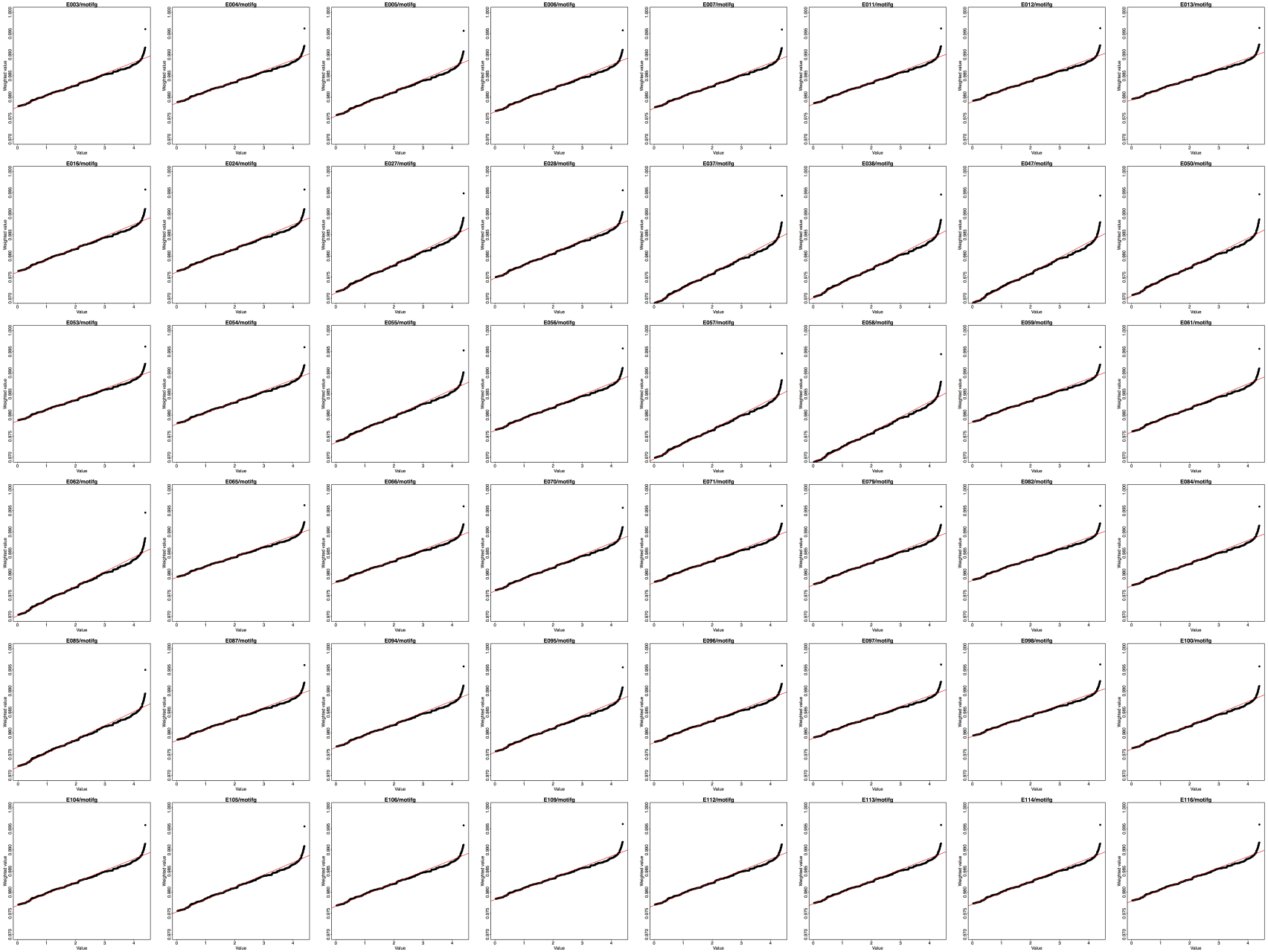


Motif breaking score fitting



Yale

Motif gain score fitting



Yale

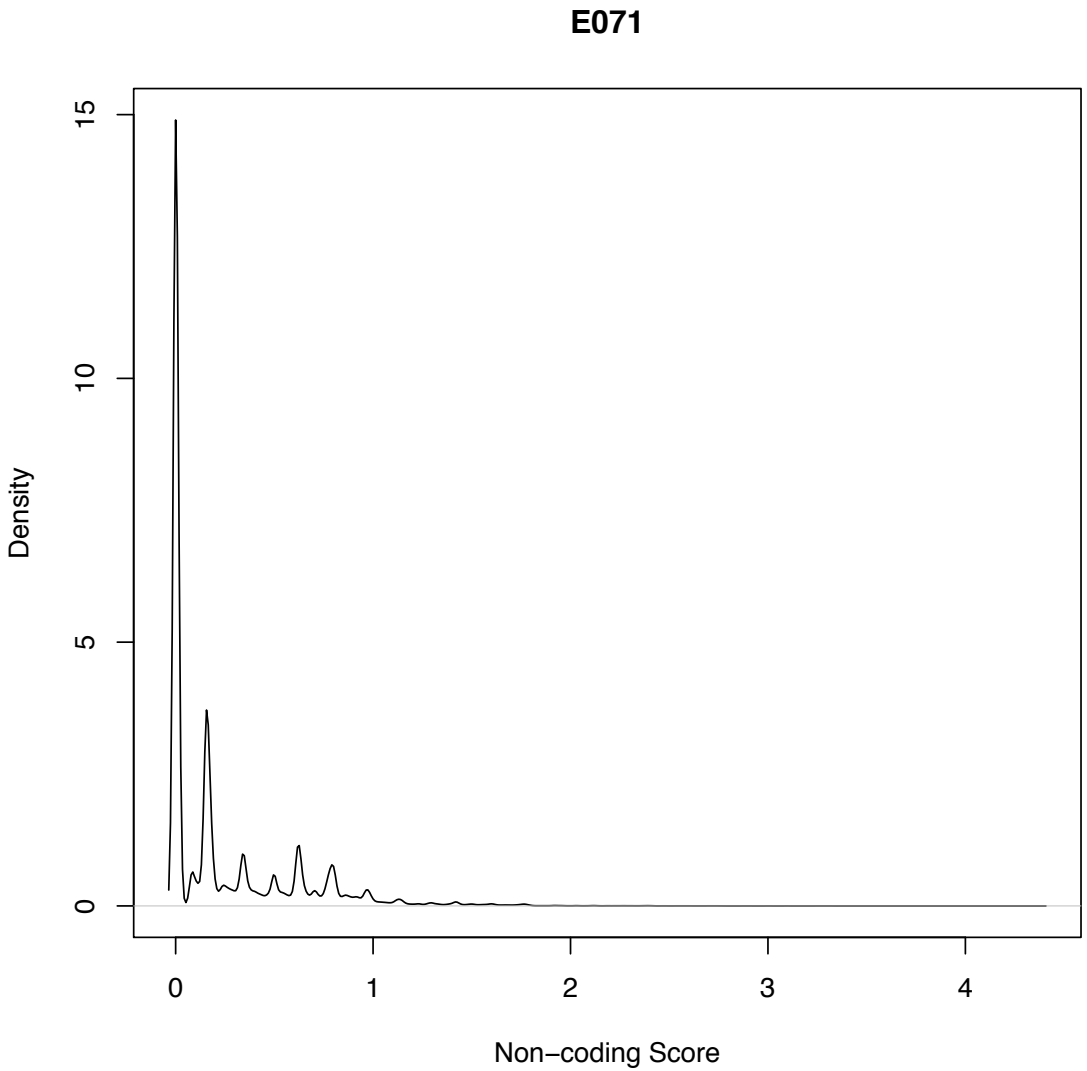
Case study: Glioblastoma

1. exome somatic mutation (public data)
2. test on three set of gene-enhancer linkages:

Original enhancer-gene linkage;
E070 Brain_Germinal_Matrix
E071 Brain_Hippocampus_Middle
Results from different linkages set are the same

Yale

Distribution of Noncoding Score



Yale

Enrichment analysis

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_TISSUE	Epithelium	RT		43	35.5	2.9E-9	3.4E-7
	GNF_U133A_QUARTILE	Cardiac Myocytes_3rd	RT		86	71.1	5.0E-8	3.9E-6
	KEGG_PATHWAY	Pathways in cancer	RT		18	14.9	6.7E-8	5.5E-6
	GNF_U133A_QUARTILE	PB-CD14+Monocytes_3rd	RT		97	80.2	2.1E-7	8.3E-6
	KEGG_PATHWAY	Small cell lung cancer	RT		10	8.3	3.3E-7	1.3E-5
	SMART	HLH	RT		10	8.3	1.2E-6	8.3E-5
	KEGG_PATHWAY	Chronic myeloid leukemia	RT		9	7.4	1.6E-6	4.4E-5
	CGAP_SAGE_QUARTILE	vascular_high-grade comedo DCIS endothelium_3rd	RT		23	19.0	1.9E-6	6.7E-4
	CGAP_SAGE_QUARTILE	stem cell_null_3rd	RT		22	18.2	5.1E-6	8.8E-4
	CGAP_SAGE_QUARTILE	stem cell_null_3rd	RT		21	17.4	1.5E-5	1.7E-3
	CGAP_SAGE_QUARTILE	brain__3rd	RT		23	19.0	3.2E-5	2.8E-3
	CGAP_SAGE_QUARTILE	kidney_normal epithelium_3rd	RT		25	20.7	3.6E-5	2.5E-3
	GNF_U133A_QUARTILE	leukemiapromyelocytic(hl60)_3rd	RT		43	35.5	3.9E-5	1.0E-3
	GNF_U133A_QUARTILE	Whole Brain_3rd	RT		51	42.1	4.9E-5	9.5E-4
	CGAP_SAGE_QUARTILE	stem cell_null_3rd	RT		22	18.2	5.0E-5	2.9E-3
	KEGG_PATHWAY	Prostate cancer	RT		8	6.6	5.8E-5	1.2E-3
	CGAP_SAGE_QUARTILE	stem cell_null_3rd	RT		18	14.9	6.8E-5	3.4E-3
	CGAP_SAGE_QUARTILE	stem cell_null_3rd	RT		20	16.5	6.9E-5	3.0E-3
	CGAP_SAGE_QUARTILE	vascular_normal liver_3rd	RT		21	17.4	7.0E-5	2.7E-3

Yale

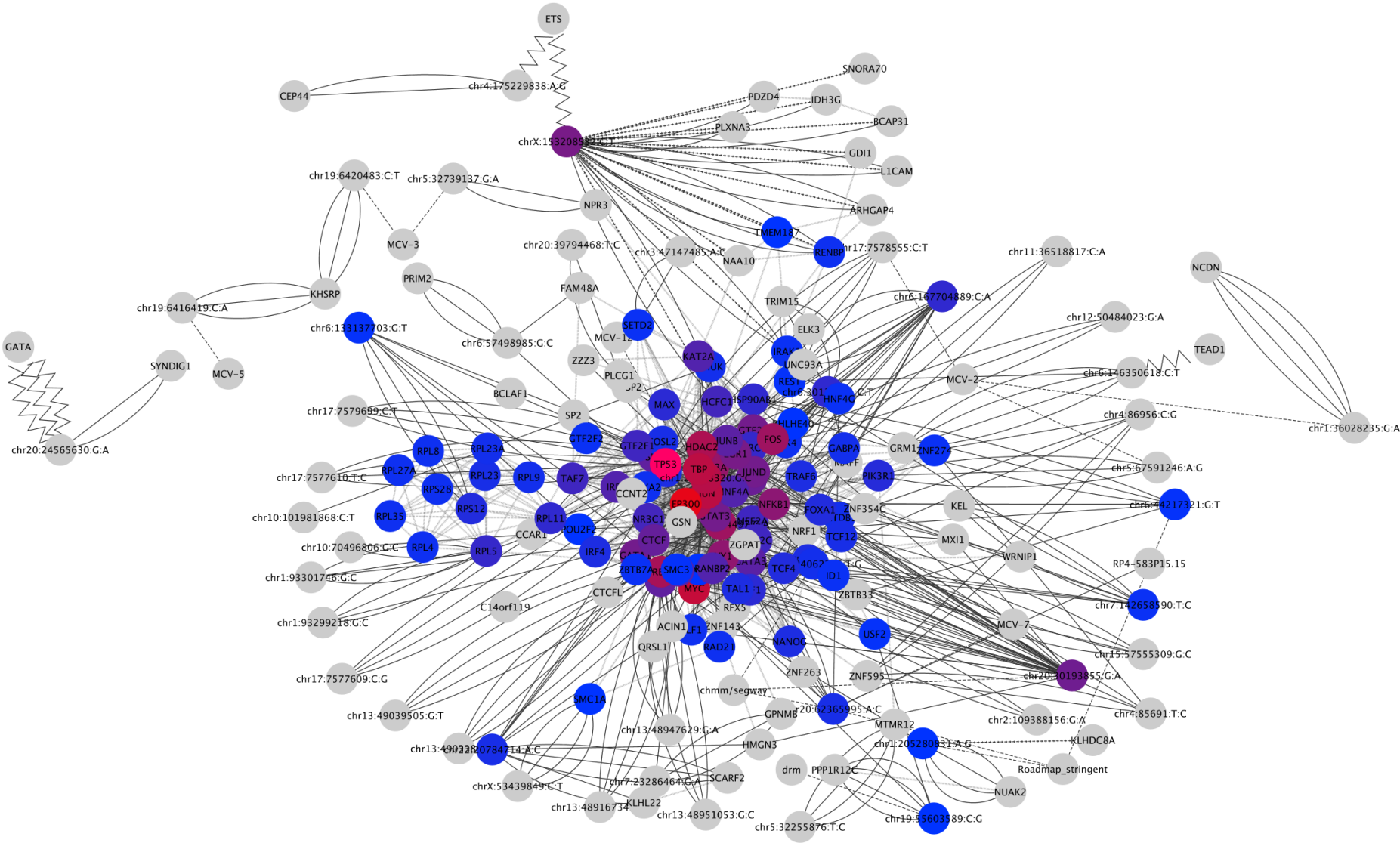
Motif Gain

Variant	TF	diff-score
chr20:24565630:G:A	GATA	3.819
chr4:175229838:A:G	ETS	3.271
chr6:146350618:C:T	TEAD1	4.394
chr6:30131441:C:T	ZNF354C	0.811
chrX:153208532:C:T	ETS	4.395

ETS high expression lead to glioma proliferation, and we found the motif gain of ETS followed by highly expression of ETS.
TEAD1, MAPK14 and SERP1 promote glioblastoma progression.
GATA, ZNF354C

Yale

Network visulization



Yale

Relative importance of features

Features are independantuse hirachical rule-based scoring to decrease the dependancy of different features, such as: motif gain/break versus gene-link(promoter, enhancer) etc however,the annotation and linkage is quite important, annotation currently cannot
Todo:

1. redefine enhancer-gene linkage
2. code rewrite
3. flexible on-the-fly weight updating, weight schema and relative importance.
4. Add more features, as indicated in ncvar grants
5. Functional annotation by enhanced gene-based network/visualization

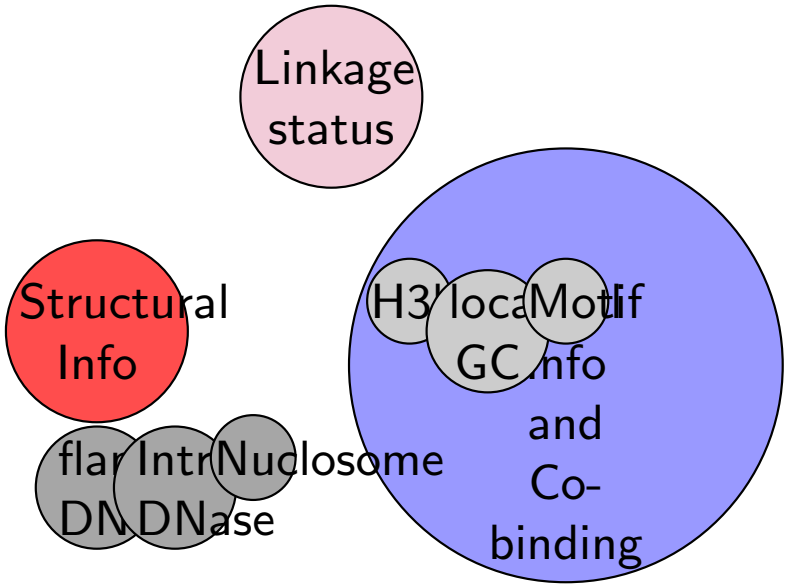
Yale

Function annotation

Extend the current output to a network-like view:
construct gene-based annotation and network, integrate with network analysis for the functional annotation of non-coding variants.
Motif break/gain; GENE-link,promoter, distal; hot/encode region;

Yale

bak


Linkage status
Structural Info
H3
loca
Moti
if
GC
nfo
and
Co-
binding
flar
Intr
Nuclosome
DN
DNase

Yale