FunSeq3: design, plan and results

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

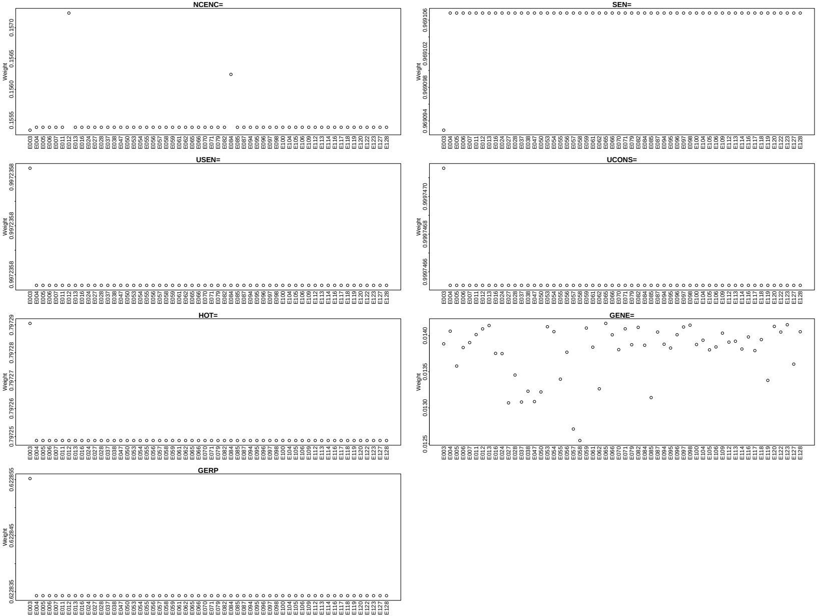
July 22, 2015

Method

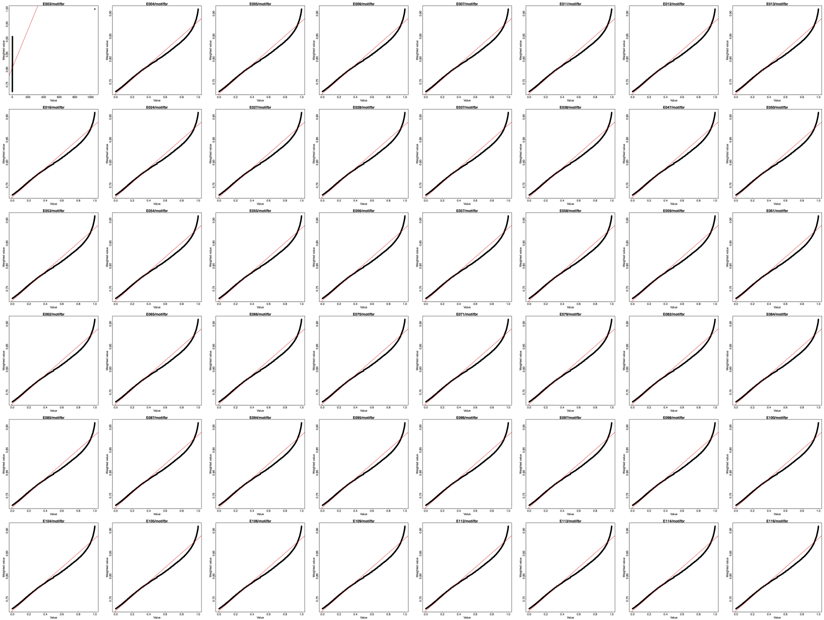
Tissue-specific enhancer-gene linkage:

- 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

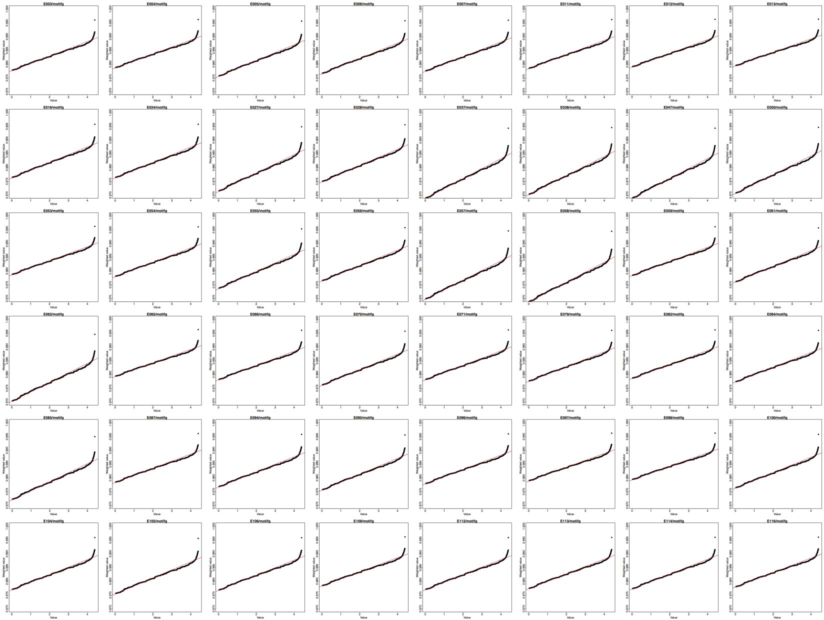
Weights for discrete features



Motif breaking score fitting



Motif gain score fitting

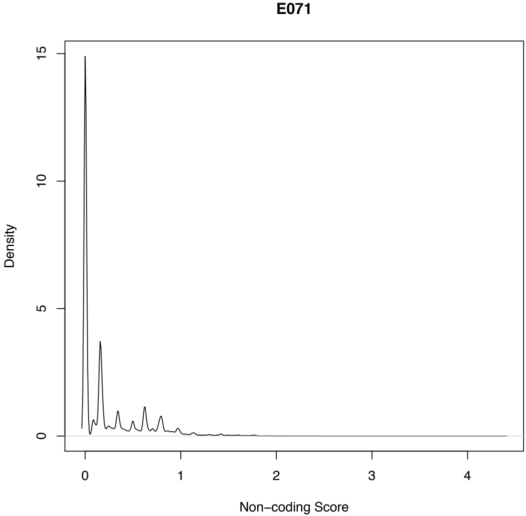


Case study: Glioblastoma

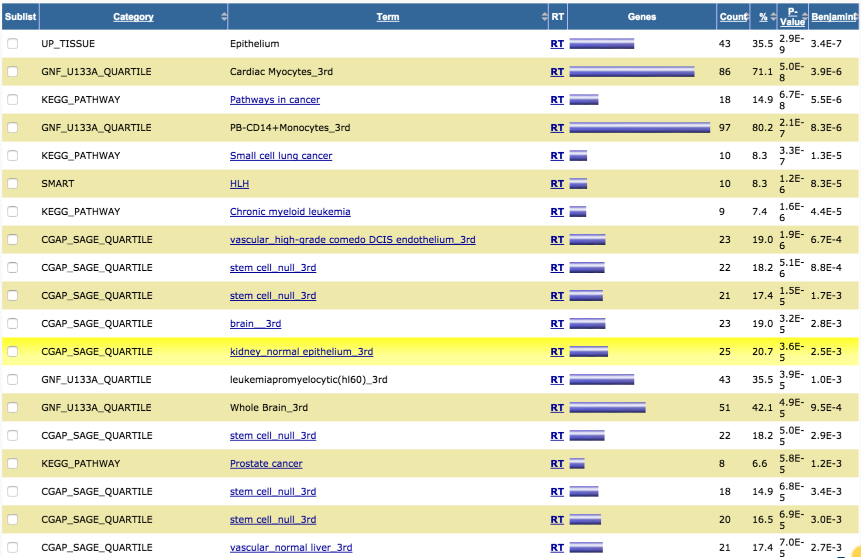
1. exome somatic mutation (public data)
2. test on three set of gene-enhancer linkages:

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

Distribution of Noncoding Score



Enrichment analysis



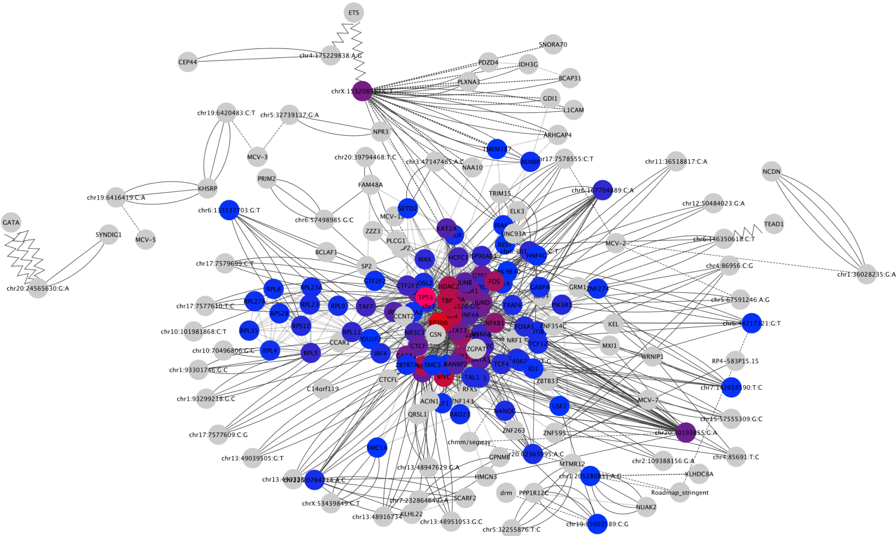
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
☐	UP_TISSUE	Epithelium	RT	▬	43	35.5	2.9E-9	3.4E-7
☐	GNF_U133A_QUARTILE	Cardiac Myocytes_3rd	RT	▬	86	71.1	5.0E-8	3.9E-6
☐	KEGG_PATHWAY	Pathways in cancer	RT	▬	18	14.9	6.7E-8	5.5E-6
☐	GNF_U133A_QUARTILE	PB-CD14+Monocytes_3rd	RT	▬	97	80.2	2.1E-7	8.3E-6
☐	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

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

Network visulization



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

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;