

MotifMap: Genome-Wide Map of Regulatory Binding Sites

Kenny Daily

June 11, 2010



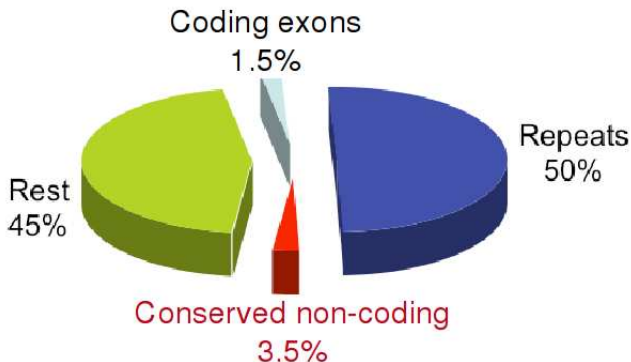
Bren School for Information and Computer Sciences
Institute for Genomics and Bioinformatics
NIH Biomedical Informatics Training Program

Motivation for MotifMap

- ▶ Understand non-coding, “genomic dark matter”
- ▶ Build maps of gene regulatory networks
- ▶ **First step: build complete maps of all TF binding sites in all genomes!**
- ▶ Applications
 - ▶ Fundamental biology
 - ▶ Medicine/Pharmacology
 - ▶ Many diseases (e.g. cancer) are complex diseases of regulation

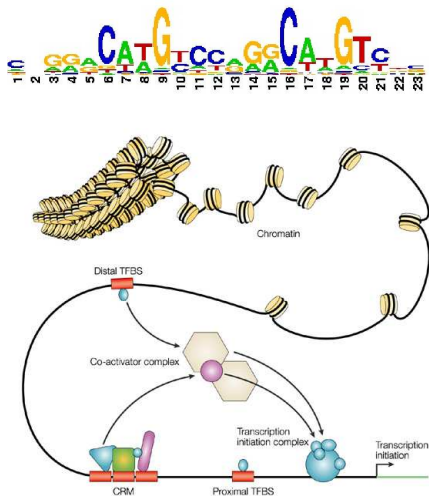
Genomic Dark Matter - Conserved Non-Coding Elements

One putative function of conserved, non-coding elements that are under positive selection are **regulatory binding sites**.



Transcription Factors and Regulation

Tumor Suppressor Protein 53 (p53) - The guardian of the cell
Mutations in p53 protein involved in most cancers



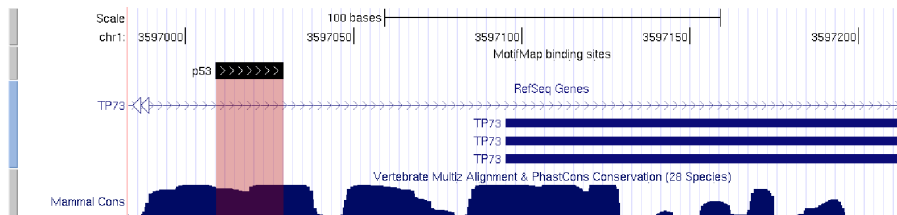
MotifMap Pipeline

1. **Collect data** for motifs from TRANSFAC, JASPAR
2. **Search** for motif binding sites across a genome
3. **Score** using the PSSM representation of the binding site
4. **Assess conservation** using multiple alignments and phylogeny of relationships between species (BBLs)
5. **Assess quality** using False Discovery Rate, experimental data (ChIPSeq)
6. **Integrate** with other biological data
7. **Disseminate data** for public use

Example: p53 Binding Site

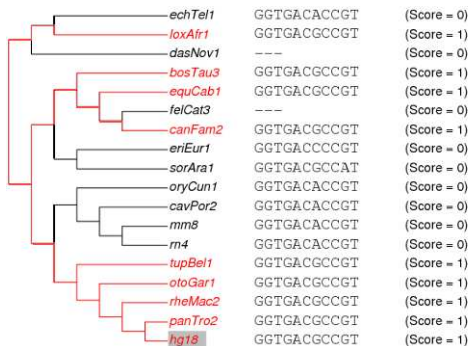


Found binding site in promoter (p-value $\leq 1 \times 10^{-5}$)



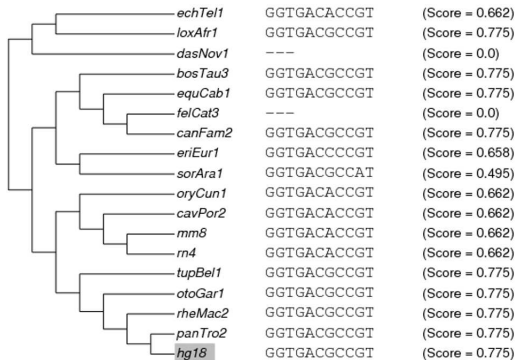
Assessing Conservation - Branch Length Score [2]

BLS = Sum the length of the branches of the subtree to most recent common ancestor



Assessing Conservation - Bayesian Branch Length Score [3]

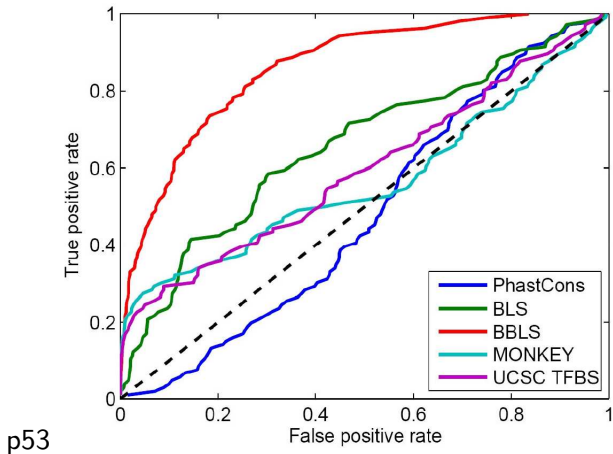
$$BBLS(p_V) = \sum_{\sigma_V} P(\sigma_V) BLS(\sigma_V)$$



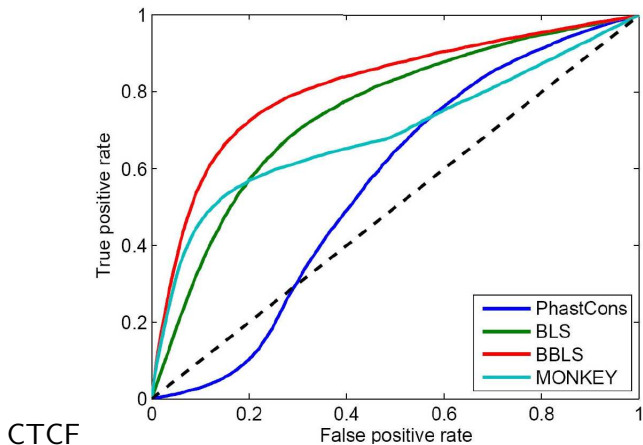
$P(\sigma_V)$ = probability of a motif (score from PSSM)

Recursive implementation of algorithm is $O(n)$, n = number of species

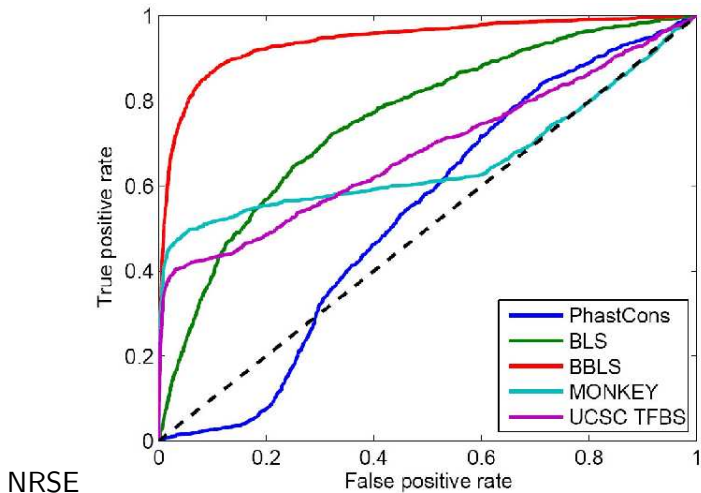
Validation - BBLs Outperforms Other Methods - p53 [3]



Validation - BBLS Outperforms Other Methods - CTCF [3]



Validation - BBLS Outperforms Other Methods - NRSE [3]



Integration With Other Data Sources

- ▶ HT data (GEO Microarray, ChipSeq, RNASeq)
- ▶ Gene Ontology (GO)
- ▶ **Genome-wide SNP Datasets (GWAS)**
- ▶ Protein-protein Interaction Networks (PPI)
- ▶ Epigenetic Interactions (Methylation, etc.)
- ▶ Sets of Homologous/Orthologous Genes (Homologene)

SNP Data Integration

The genetic basis of some diseases (or phenotypes) is not necessarily due to changes in the coding sequences of proteins/genes, but disruption of regulator elements (motif binding sites).

- ▶ Hypertension
- ▶ Schizophrenia
- ▶ Accelerated growth in flies

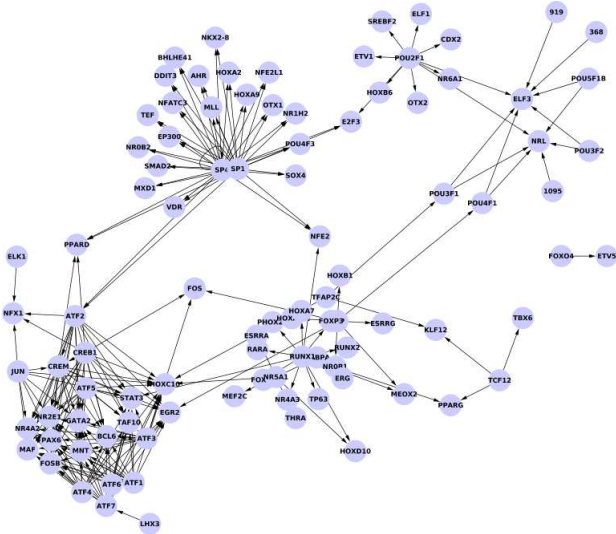
SNPs and Schizophrenia

- ▶ 1,879 genes (literature, GWAS, and domain knowledge)
- ▶ Illumina Human CNV370 (64 chronic schizophrenia subjects, and 74 matched controls)

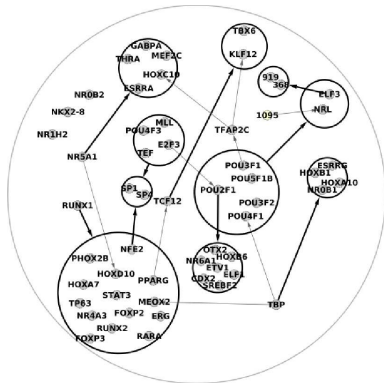
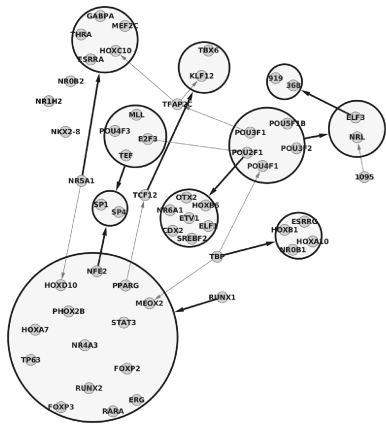
Significance of these 302,783 SNPs (Z-scores) determined using blood oxygen level in the brain during an Item Recognition Test

Regulatory Networks of Transcription Factor Genes

What binding sites are found near a genes which are themselves transcription factors?



Regulatory Networks of Transcription Factor Genes



MotifMap Web Server

- ▶ Access to data (biologists in particular)

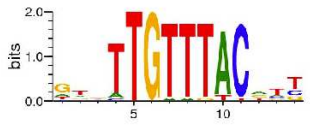
<http://motifmap.ics.uci.edu/>

- ▶ Answer common questions:
 - ▶ Where are the binding sites for some binding protein that are highly conserved?
 - ▶ What binding sites are nearby my favorite gene?
 - ▶ In which species is the binding site for p53 binding protein near the p73 gene conserved across?
 - ▶ I want to visualize all the binding sites for X and Y binding sites across the genome.
 - ▶ I know the transcription of genes X and Y are somehow correlated. What binding sites do they share?

Select Genome Track: **Human (hg18 multiz28way_placental)** species (build, alignment)

Select Motif:

Motif ID	Human (hg18 multiz28way_placental)	Length	
M00474	Human (hg18 multiz28way_placental Version 1)	14	
M00476	Human (hg18 multiz44way_placental)	14	
M00477	Fly (dm3 multiz15way_filesonly)	14	
M00478	Yeast (sacCer2 multiz7way)	14	
M00492	FOXO3	HNBTTGTTTACDWTW	14
	Cdc5	GNKTTAACRTAD	12
	STAT1	BNNTTCCS	8



Parameters: ZScore: 3.72 BLS: 0 BBLS: 0 FDR: 0.1 Distance: 1000

[Preview](#) [Add to Genome Browser View](#)

Location	+/-	NLOD	Z-score	BLS	BBLS	FDR	Gene	Distance (bp)	Description
chr7:113841746..1138417...	A +	0.912	4.590	2.945	1.741	0.000	FOXP2	-541	forkhead box P2 isoform II
chr2:174968639..1749686...	A -	0.925	4.704	2.546	1.696	0.000	CIR1	-36	CBF1 interacting corepressor
chr5:64955831..64955845	A +	0.881	4.332	2.847	1.664	0.000	TRIM23	-112	ADP-ribosylation factor domain protein 1 isoform
chr1:87567626..87567640	A -	0.930	4.744	2.136	1.495	0.000	LMO4	902	LIM domain only 4
chr1:152197762..1521977...	A +	0.875	4.281	2.945	1.489	0.000	CRTC2	-95	CREB regulated transcription coactivator 2
chr18:30545110..30545124	A -	0.929	4.735	2.586	1.465	0.000	DTNA	931	dystrobrevin alpha isoform 4
chr1:172684492..1726845...	A +	0.906	4.540	2.945	1.440	0.000	GPR52	658	G protein-coupled receptor 52
chr5:131160175..1311601...	A -	0.919	4.652	2.182	1.435	0.000	FNIP1	-466	folliculin interacting protein 1 isoform 1
chr18:37354308..37354322	A +	0.896	4.453	2.945	1.434	0.000	KC6	-251	
chr2:196230266..1962302...	A +	0.942	4.850	1.872	1.406	0.000	SLC39A10	170	solute carrier family 39 (zinc transporter),

Select Genome Track: **Human (hg18 multiz28way_placental)** species (build, alignment)

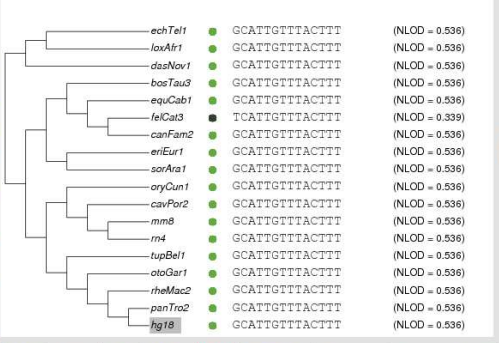
Select Motif:

Motif ID	Name
M00474	FOXO1
M00476	FOXO4
M00477	FOXO3
M00478	Cdc5
M00492	STAT1

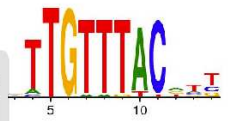
Parameters: ZScore: BLI

Location	
chr7:113841746..1138417...	A +
chr2:174968639..1749686...	A -
chr5:64955831..64955845	A +
chr1:87567626..87567640	A -
chr1:152197762..1521977...	A +
chr18:30545110..30545124	A -
chr1:172684492..1726845...	A +
chr5:131160175..1311601...	A -
chr18:37354308..37354322	A +
chr2:196230266..1962302...	A +

chr7:113841746..113841760



0.942	4.850	1.872	1.406	0.000	SLC39A10	170	solute carrier family 39 (zinc transporter),
-------	-------	-------	-------	-------	----------	-----	--



orm II
pressor
or domain protein 1 isoform
cription coactivator 2
oform 4
eptor 52
rotein 1 isoform 1

Human (hg18 multiz28way_placental)

NM_003106|SOX2 x

select species (build, alignment)

enter gene name or ID.

Location		Strand	NLOD	Z-score	BLS	BBLS	FDR	MotifId	TF Name	Distance (bp)	Region
chr3:182912155..182912161	A	+	1.000	4.127	1.113	1.113	0.030	M00271	AML1a	-260	Upstream
chr3:182912291..182912301	A	-	1.000	5.110	1.472	1.472	0.026	M00931	Sp1	-124	Upstream
chr3:182912290..182912303	A	-	0.997	5.243	1.472	1.455	0.009	M00932	Sp1	-125	Upstream
chr3:182912990..182912996	A	+	1.000	4.127	2.087	1.985	0.022	M00271	AML1a	575	Downstream
chr3:182913163..182913171	A	+	1.000	3.971	2.296	2.258	0.010	M00471	TBP	748	Downstream
chr3:182913368..182913376	A	+	0.971	4.421	1.870	1.381	0.002	M00698	HEB	953	Downstream

Options

Scores:

NLOD: 0.8

BBLS: 0.5

BLS: 0

FDR: 0.1

Distance:

Upstream: 1000

Downstream: 1000

Search!

Reset

Download:

GFF3

BED

CSV

Show in UCSC

Human (hg18 multiz2Bway_placental) ▾

select species (build, alignment)

M00172_AP-1 x

enter Transcription Factor name or Tansfac ID.

Search!

Motif Options

dbSNP Options

near-gene-5 ▾

Class - All ▾

HapMap SNP Options

Download Options

Refresh

Reset

Summary **M00172_AP-1**

Motif Information

Length of Motif:
 Total # of sites:
 Sites with BBLs > 0.5:
 Sites with BBLs > 1:
 Sites with BBLs > 2:

HapMap SNP Density (SNPs/Kb DNA)

- All Sites:
 - BBLs > 0.5:
 - BBLs > 1:
 - BBLs > 2:

dbSNP Density (SNPs/Kb DNA)

- All Sites:
 - BBLs > 0.5:
 - BBLs > 1:
 - BBLs > 2:
 # of sites with overlapping SNPs:

dbSNP (version 130) data

10 record(s) found

Location	Strand	NLOD	Z-score	BLS	BBLs	FDR	Motif	SNP - rs#	SNP Start	Observed	Function	Class
chr1:15482903	+	0.958	4.356	1.756	1.325	0.000	M00172	rs2949417	76165	G/T	near-gene-5	single
chr1:15489792	+	0.982	4.539	1.824	1.242	0.000	M00172	rs5772017	77148	-/T	near-gene-5	insertion
chr9:9095606..	+	0.953	4.321	2.461	1.390	0.000	M00172	rs28693927	847038	A/G	near-gene-5	single
chr9:16820707	+	0.969	4.439	1.742	1.167	0.000	M00172	rs59192718	847515	C/G	near-gene-5	single
chr9:22107593	+	0.960	4.371	1.670	1.077	0.000	M00172	rs71509446	849563	C/G	near-gene-5	single
chr9:22149120	+	0.946	4.270	2.262	1.426	0.000	M00172	rs71509446	849563	C/G	near-gene-5	single

HapMap3 data

10 record(s) found

Location	Strand	NLOD	Z-score	BLS	BBLs	FDR	Motif	SNP - rs#	SNP Start
chr1:2306417..2306	+	0.969	4.439	1.348	1.018	0.000	M00172	rs11046090	21549797
chr1:2377776..2377	+	1.000	4.672	1.534	1.153	0.000	M00172	rs1490388	126877347
chr1:3000522..3000	+	0.952	4.312	1.841	1.005	0.000	M00172	rs1183756	105230013
chr1:4766062..4766	+	0.994	4.628	1.650	1.276	0.000	M00172	rs16916114	90886909

Thank You!

Funding

- ▶ NIH Biomedical Informatics Training Grant

Baldi Lab

- ▶ Pierre Baldi, Advisor
- ▶ Paul Rigor
- ▶ Sholeh Forouzan
- ▶ Michael Zeller
- ▶ Yimeng Dou
- ▶ Vishal Patel
- ▶ Jacob Biesinger

Collaborators

- ▶ Xiaohui Xie
- ▶ Olivier Cinquin
- ▶ Suzanne Sandmeyer
- ▶ Tony Long
- ▶ Fabio Macciardi
- ▶ Ken Cho



John S. Mattick, Ryan J. Taft, and Geoffrey J. Faulkner.

A global view of genomic information moving beyond the gene and the master regulator.

Trends in Genetics, 26(1):21–28, January 2010.



Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, Leopold Parts, Joseph W. Carlson, Madeline A. Crosby, Matthew D. Rasmussen, Sushmita Roy, Ameya N. Deoras, Graham G. Ruby, Julius Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Emily Hodges, Angie S. Hinrichs, Anat Caspi, Benedict Paten, Seung-Won W. Park, Mira V. Han, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Stein Aerts, Jacques van Helden, Bassem Hassan, Donald G. Gilbert, Deborah A. Eastman, Michael Rice, Michael Weir, Matthew W. Hahn, Yongkyu Park, Colin N. Dewey, Lior Pachter, James J. Kent, David Haussler, Eric C. Lai, David P. Bartel, Gregory J. Hannon, Thomas C. Kaufman, Michael B. Eisen, Andrew G. Clark, Douglas Smith, Susan E. Celniker, William M. Gelbart, and Manolis Kellis.

Discovery of functional elements in 12 drosophila genomes using evolutionary signatures.

Nature, 450(7167):219–232, November 2007.



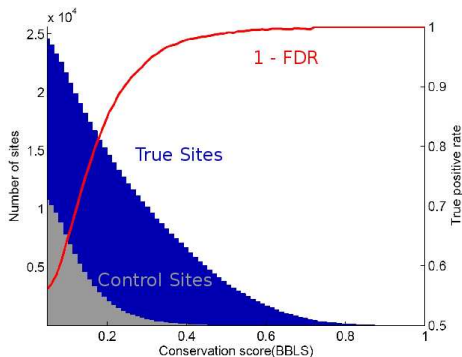
Xiaohui Xie, Paul Rigor, and Pierre Baldi.

Motifmap: a human genome-wide map of candidate regulatory motif sites.

Bioinformatics, 25(2):167–174, January 2009.

Validation - Estimating False Discovery Rate [3]

1. Randomize original TF binding site PSSM (permute the nucleotide positions only)
2. Apply the same MotifMap pipeline to find matches to the control motifs



Data sources

- ▶ TF Binding matrices
 - ▶ TRANSFAC database
 - ▶ JASPAR database
 - ▶ Biologists
- ▶ Genomes
 - ▶ UC Santa Cruz Genome Browser
 - ▶ Species-specific databases
 - ▶ Size: 10MB-4GB/species
- ▶ Multiple Sequence Alignments/Phylogeny
 - ▶ UC Santa Cruz Genome Browser
 - ▶ Biologists
 - ▶ Size: 100MB-400GB

Currently Available Species:

Species	# of matrices	Sites identified	Conserved sites
Human	400	61,078,732	35,706,731
Chimpanzee	400	72,129,675	39,834,020
Mouse	400	63,075,690	17,128,081
Fly	150	12,937,527	6,234,280
Yeast	150	12,415,229	8,080,077
Nematode	9	2,319,239	314,351

Other Current Projects

- ▶ Fly Growth - Find TF binding sites with SNPs that change binding affinity (Anthony Long)
- ▶ Nematode Germline Development - Examining DNA and RNA binding motifs to identify candidate genes (Olivier Cinquin)
- ▶ Yeast Transposons - finding sites of Pol III binding to BoxA and BoxB elements (Suzanne Sandmeyer)
- ▶ Genome wide analysis to find genes with multiple, common binding sites (*cis*-regulatory modules)
- ▶ Genome wide comparison of human and chimpanzee to find genes with common and different binding sites