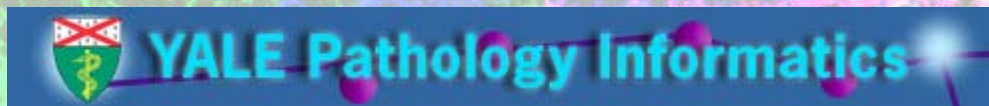
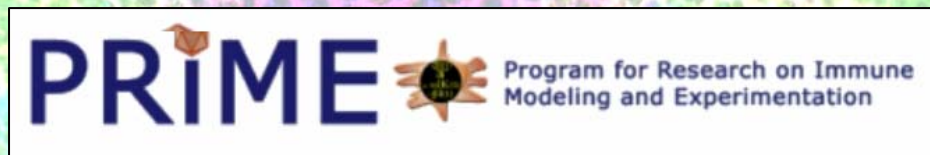


Promoter Analysis & Gene Set Enrichment

Steven H. Kleinstein



Department of Pathology
Yale University School of Medicine
steven.kleinstein@yale.edu

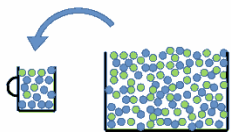


May 6, 2010

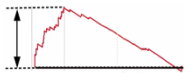
Lecture & Lab Outline



- Promoter analysis



- Over-representation analysis



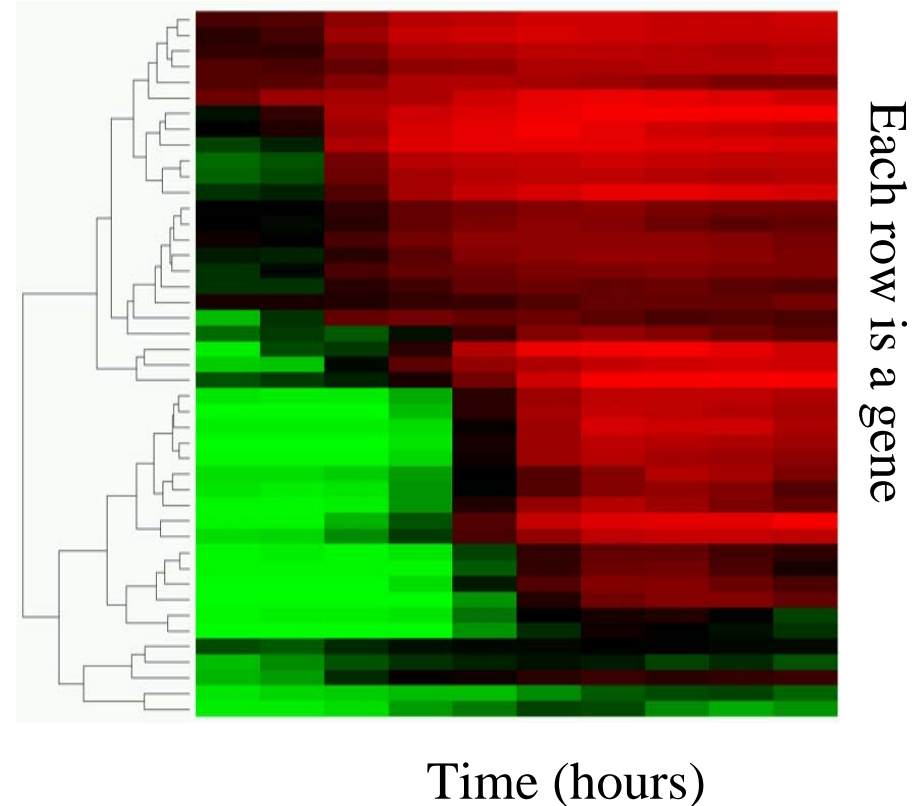
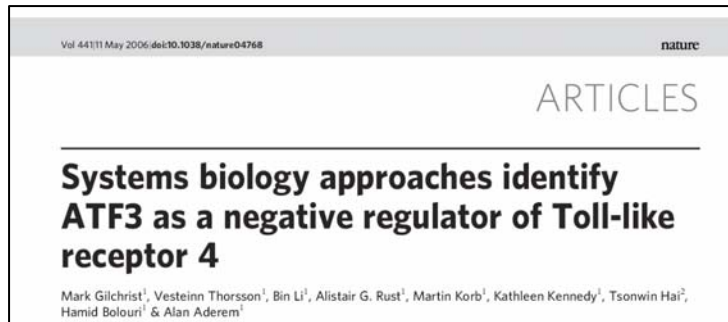
- Gene set enrichment analysis

Lab section by Uri Hershberg

Illustrate some general approaches and concepts

Identifying regulators of TLR responses

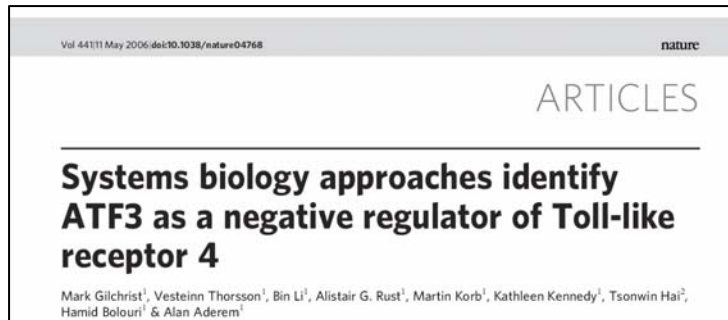
Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)



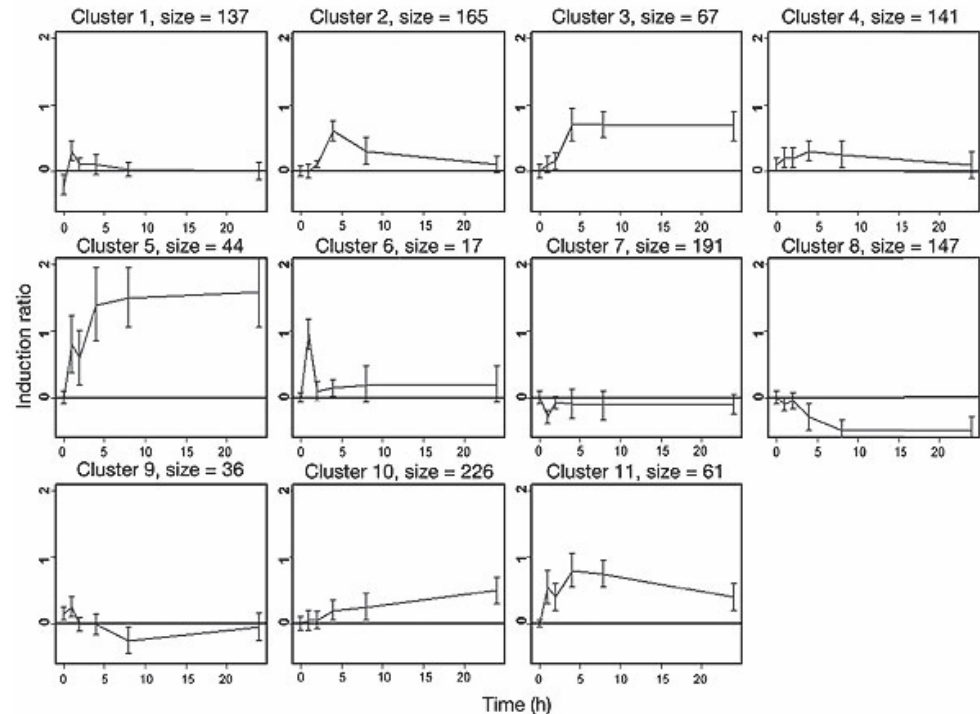
Hypothesize that genes with similar temporal kinetics are co-regulated and that they share regulators

Identifying regulators of TLR responses

Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)



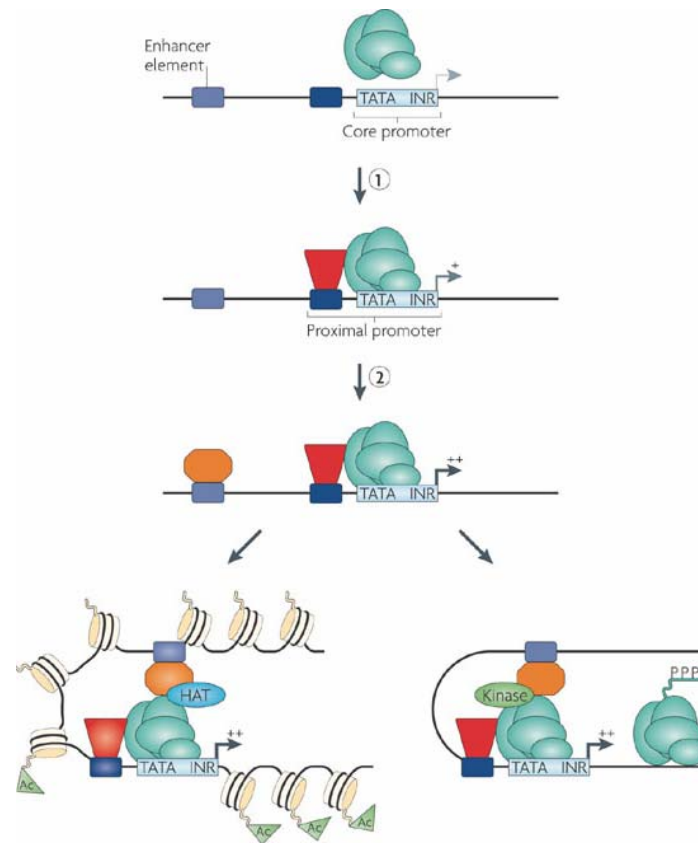
K-means clustering defined 11 groups of genes comprising regulated 'waves' of transcription



Hypothesize that clustered genes are co-regulated and that they share cis-regulatory elements

Transcriptional regulation by promoters and enhancers

General transcription factors (green ovals) bind to core promoter regions through recognition of common elements such as TATA boxes and initiators (INR)



Nature Reviews | Genetics

(Farnham, Nature Reviews Genetics, 2009)

Promoter activity can be altered by site-specific DNA-binding factors (red trapezoid) interacting with cis elements (dark blue box)

DNA Sequence Motifs for TF Binding Sites

Short, recurring patterns in DNA with presumed biological function

Nature Biotechnology **24**, 423 - 425 (2006)

a

| | | |
|-------|---------------|--------------------------------------|
| HEM13 | CCCATTGTTCTC | } Collection of binding sites (ROX1) |
| HEM13 | TTTCTGGTTCTC | |
| HEM13 | TCAATTGTTTAG | |
| ANB1 | CTCATTGTTGTC | |
| ANB1 | TCCATTGTTCTC | |
| ANB1 | CCTATTGTTCTC | |
| ANB1 | TCCATTGTTTCGT | |
| ROX1 | CCAATTGTTTGT | |

b YCHATTGTTCTC ← Consensus sequence

c

| | | |
|---|--------------|--------------------|
| A | 002700000010 | } Frequency Matrix |
| C | 464100000505 | |
| G | 000001800112 | |
| T | 422087088261 | |



For prediction of new sites, need to account for conservation

Measuring Conservation in the Binding Site

Information content measures conservation at each site

Measure of conservation
at each position i :

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

ATG

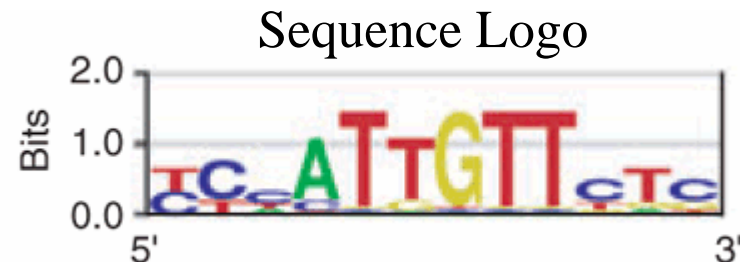
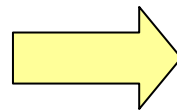
ATC

AAT

AAA

210

Information content



Total information content related to probability of finding motif in 'random' DNA sequence

<http://weblogo.berkeley.edu/>

The TRANSFAC Database

Eukaryotic transcription factors and their genomic binding sites

TRANSFAC MATRIX TABLE, Release 12.1 - licensed - 2008-03-31, (C) Biobase GmbH

Statistics Number of binding factors 3
Number of references 1

[Accession Number](#) M00513

[Identifier](#) V\$ATF3_Q6

[Created](#) 06.11.2001 by [rio](#).

[Updated](#) 11.03.2003 by [dtc](#).

Copyright Copyright (C), Biobase GmbH.

[Name](#) ATF3

[Factor Description](#) activating transcription factor 3

[Binding factors](#) [T01095](#); ATF3; Species: rat, Rattus norvegicus.

[T01313](#); ATF3; Species: human, Homo sapiens.

[T04850](#); ATF3; Species: mouse, Mus musculus.

[Binding Matrix](#)

| A | C | G | T | Consensus |
|----|---|----|----|-----------|
| 1 | 3 | 1 | 1 | C |
| 0 | 2 | 2 | 3 | B |
| 1 | 4 | 2 | 0 | C |
| 0 | 0 | 0 | 11 | T |
| 0 | 0 | 10 | 1 | G |
| 10 | 1 | 0 | 0 | A |
| 0 | 9 | 1 | 1 | C |
| 1 | 0 | 7 | 0 | G |
| 0 | 0 | 0 | 10 | T |
| 0 | 9 | 1 | 0 | C |
| 8 | 0 | 0 | 2 | A |
| 1 | 2 | 2 | 3 | N |
| 0 | 5 | 1 | 1 | C |
| 0 | 3 | 4 | 0 | S |

 TGAcGTCA

TRANSFAC has public (older version)
and commercial (more features) versions

Other (free) possibility:



The high-quality transcription factor binding profile database

Current version contains 834 matrices (601 vertebrate)

The TRANSFAC Database

Eukaryotic transcription factors and their genomic binding sites

TRANSFAC MATRIX TABLE, Release 12.1 - licensed - 2008-03-31, (C) Biobase GmbH

| Statistics | Number of binding factors 3 Number of references 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--------------------|---|----|----|-----------|---|-----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|---|---|---|----|---|---|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Accession Number | M00513 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Identifier | V\$ATF3_Q6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Created | 06.11.2001 by rio | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Updated | 11.03.2003 by dtc | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Copyright | Copyright (C), Biobase GmbH | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Name | ATF3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Factor Description | activating transcription factor 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Binding factors | T01095 ; ATF3; Species: rat, Rattus norvegicus. T01313 ; ATF3; Species: human, Homo sapiens. T04850 ; ATF3; Species: mouse, Mus musculus. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Binding Matrix | <table border="1"> <thead> <tr> <th>A</th> <th>C</th> <th>G</th> <th>T</th> <th>Consensus</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td><td>1</td><td>1</td><td>C</td></tr> <tr><td>0</td><td>2</td><td>2</td><td>3</td><td>B</td></tr> <tr><td>1</td><td>4</td><td>2</td><td>0</td><td>C</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>11</td><td>T</td></tr> <tr><td>0</td><td>0</td><td>10</td><td>1</td><td>G</td></tr> <tr><td>10</td><td>1</td><td>0</td><td>0</td><td>A</td></tr> <tr><td>0</td><td>9</td><td>1</td><td>1</td><td>C</td></tr> <tr><td>1</td><td>0</td><td>7</td><td>0</td><td>G</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>10</td><td>T</td></tr> <tr><td>0</td><td>9</td><td>1</td><td>0</td><td>C</td></tr> <tr><td>8</td><td>0</td><td>0</td><td>2</td><td>A</td></tr> <tr><td>1</td><td>2</td><td>2</td><td>3</td><td>N</td></tr> <tr><td>0</td><td>5</td><td>1</td><td>1</td><td>C</td></tr> <tr><td>0</td><td>3</td><td>4</td><td>0</td><td>S</td></tr> </tbody> </table> | A | C | G | T | Consensus | 1 | 3 | 1 | 1 | C | 0 | 2 | 2 | 3 | B | 1 | 4 | 2 | 0 | C | 0 | 0 | 0 | 11 | T | 0 | 0 | 10 | 1 | G | 10 | 1 | 0 | 0 | A | 0 | 9 | 1 | 1 | C | 1 | 0 | 7 | 0 | G | 0 | 0 | 0 | 10 | T | 0 | 9 | 1 | 0 | C | 8 | 0 | 0 | 2 | A | 1 | 2 | 2 | 3 | N | 0 | 5 | 1 | 1 | C | 0 | 3 | 4 | 0 | S |
| A | C | G | T | Consensus | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 3 | 1 | 1 | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 2 | 2 | 3 | B | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 4 | 2 | 0 | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 11 | T | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 0 | 10 | 1 | G | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | 1 | 0 | 0 | A | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 9 | 1 | 1 | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 0 | 7 | 0 | G | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 10 | T | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 9 | 1 | 0 | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 0 | 0 | 2 | A | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 2 | 3 | N | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 5 | 1 | 1 | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | 3 | 4 | 0 | S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

MATCH Score

Information Vector
(higher for conserved positions)

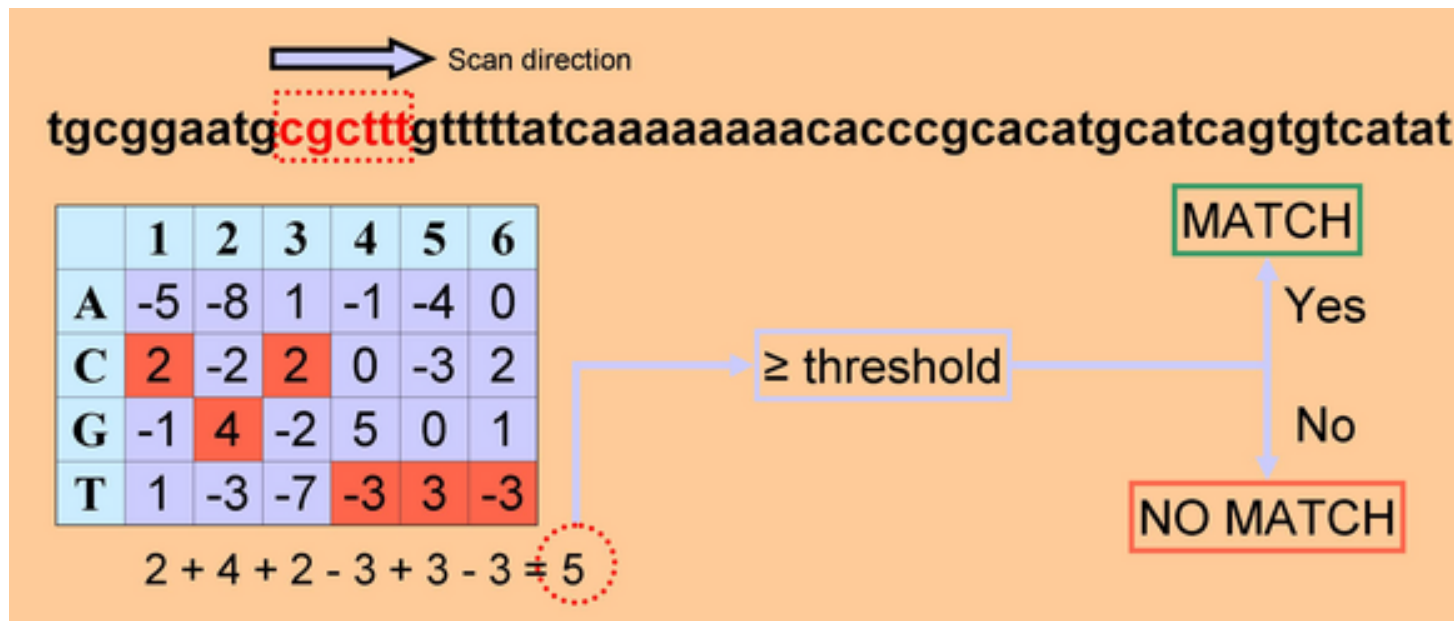
$$\sum_{i=1}^L I(i) f_{i,b_i}$$

Frequency of nucleotide b_i to occur at the position i of the matrix ($B \in \{A, T, G, C\}$)

Assumes positions are independent

Identifying putative TF binding sites

Search by scanning the promoter region

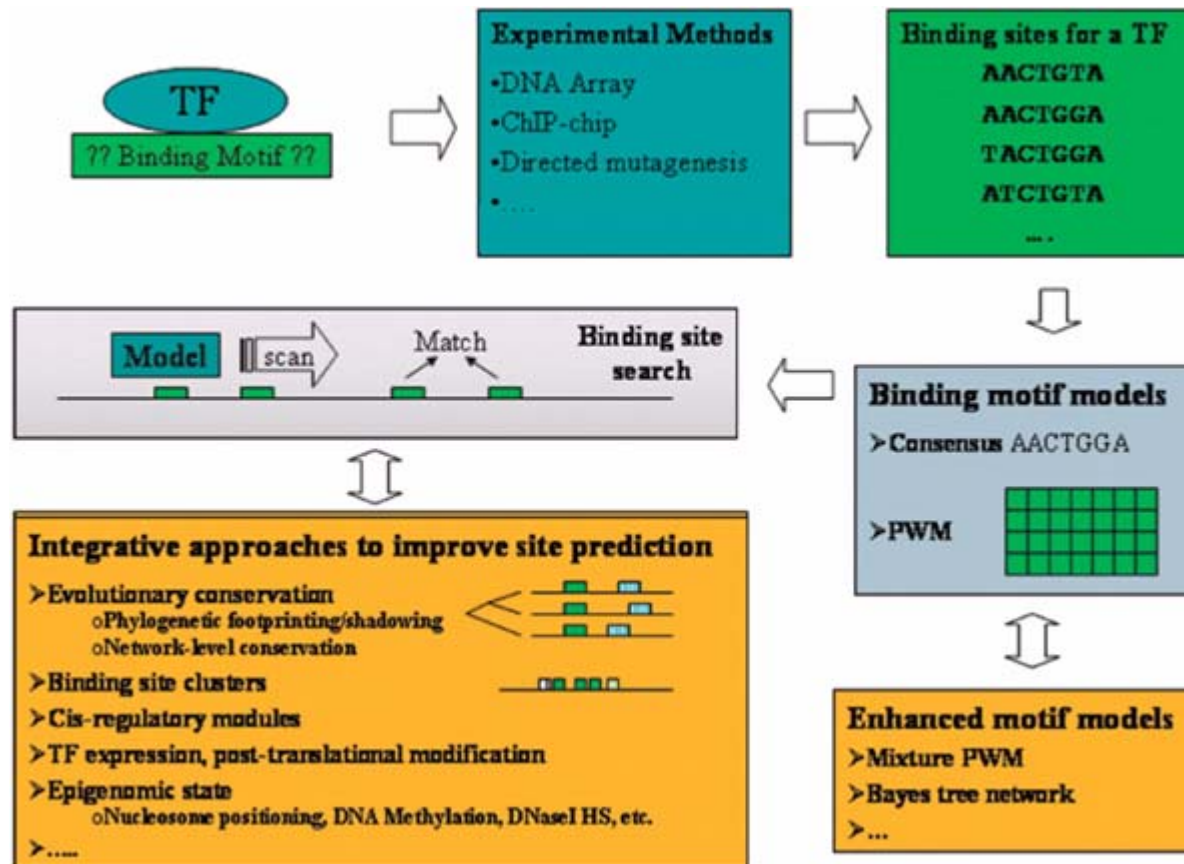


MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2: e36.

Threshold can be determined by looking at “random” DNA

Identifying putative TF binding sites

Integrative approaches improve predictions – active research area

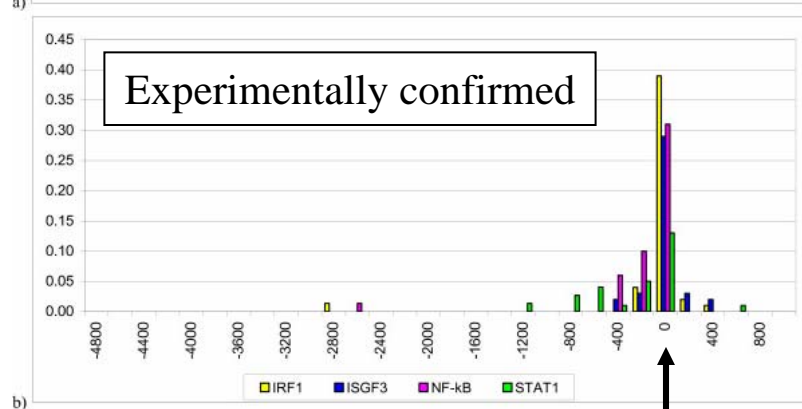
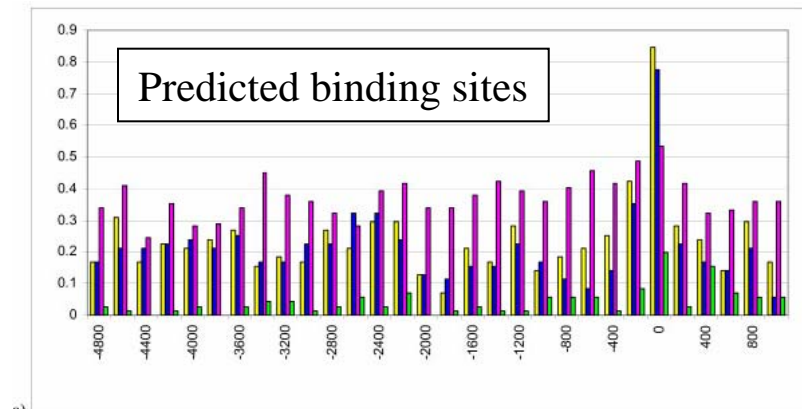


(Hannenhalli, Bioinformatics, 2008)

‘Gene Sets’ of target genes for each transcription factor

Focus on proximal promoter regions

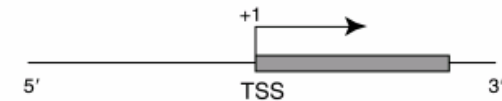
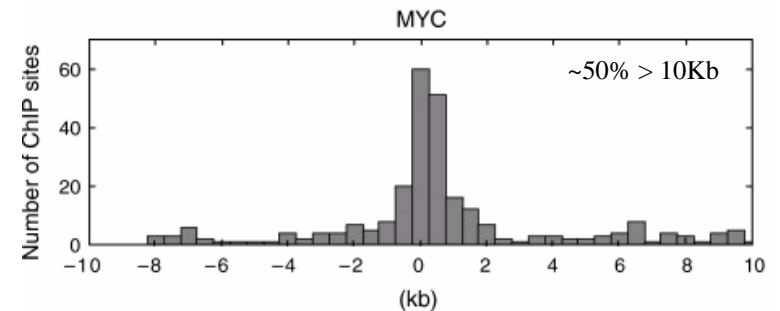
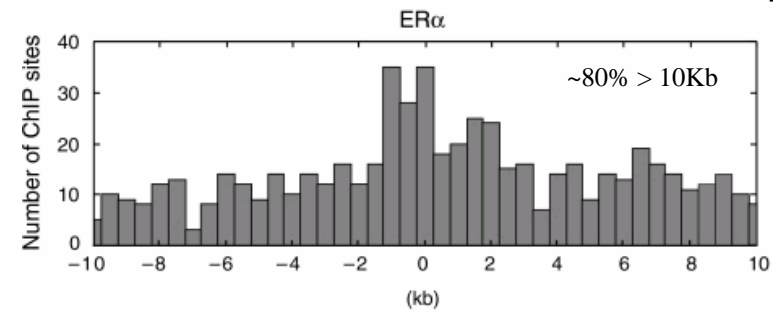
Common practice to consider 1-2Kb region around TSS



(Ananko et al, BMC Bioinformatics, 2007)

TSS

ChIP-chip data is mixed



(Hua et al, MSB, 2008)

Recent genome-wide data calls this into question

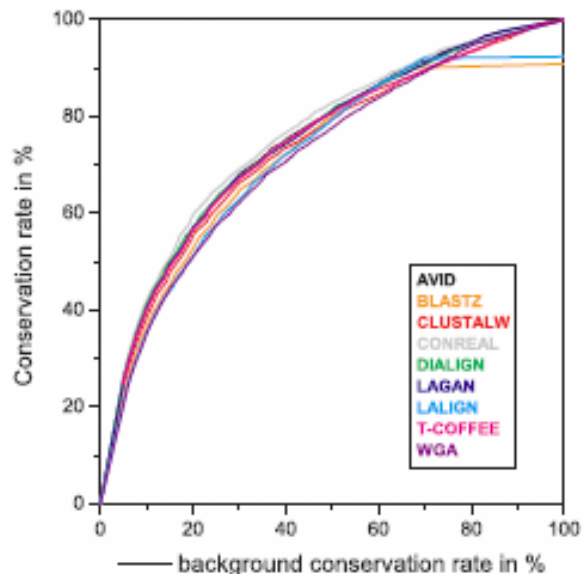
Focus on evolutionarily conserved regions

98% experimentally defined sequence-specific binding sites of skeletal-muscle-specific TFs confined to 19% of human sequences most conserved in rodent

(Wasserman et al., Nat Genet. 2000)

**Sequence identity >65% identifies
72% of the known TFBSs**

(Sauer et al, Bioinformatics. 2006)



Evolutionary conservation excludes known sites

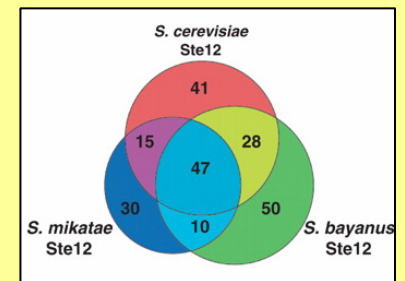
32-40% of functional human binding sites
are not functional in rodents

(Dermitzakis and Clark, Mol Biol Evol., 2002)

Divergence of Transcription Factor Binding Sites Across Related Yeast Species

Anthony R. Borneman,^{1*} Tara A. Gianoulis,² Zhengdong D. Zhang,³
Haiyuan Yu,² Joel Rozowsky,³ Michael R. Seringhaus,² Lu Yong Wang,⁴
Mark Gerstein,^{2,3,5} Michael Snyder^{1,2,3†}

SCIENCE VOL 317 10 AUGUST 2007



Requiring human–mouse–rat genomic alignments provided a 44-fold increase in the specificity of TRANSFAC predictions (Rat Genome Sequencing Project, Nature, 2004)

Variation in TF binding across individuals

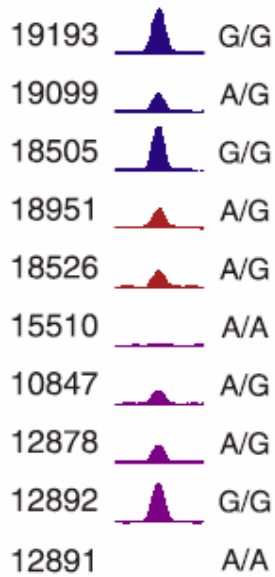
6% of binding regions within 1 kb of transcription start sites (TSSs) of RefSeq genes differed significantly across individuals

Variation in Transcription Factor Binding Among Humans

Maya Kasowski,^{1,2*} Fabian Grubert,^{1,2*} Christopher Heffelfinger,¹ Manoj Hariharan,^{1,2} Akwasi Asabere,¹ Sebastian M. Waszak,^{3,4} Lukas Habegger,⁵ Joel Rozowsky,⁶ Minyi Shi,^{1,2} Alexander E. Urban,^{1,7} Mi-Young Hong,⁷ Konrad J. Karczewski,² Wolfgang Huber,³ Sherman M. Weissman,⁷ Mark B. Gerstein,^{5,6,8} Jan O. Korbel,^{3,9†} Michael Snyder^{1,2†}

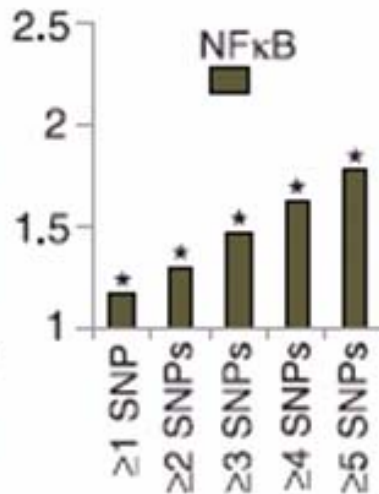
9 APRIL 2010 VOL 328 SCIENCE www.sciencemag.org

ChIP-Seq Analysis

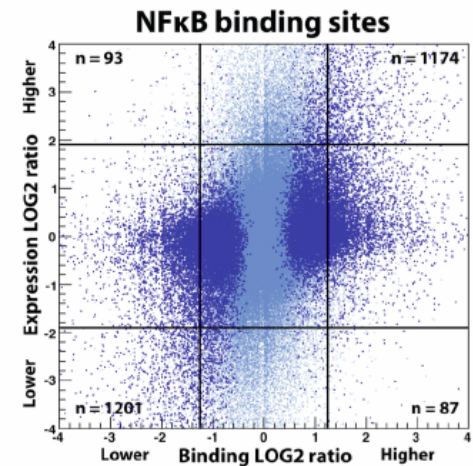


SNPs in motif predict binding sites

Enrichment in significant binding differences



Also correlated with match to consensus site

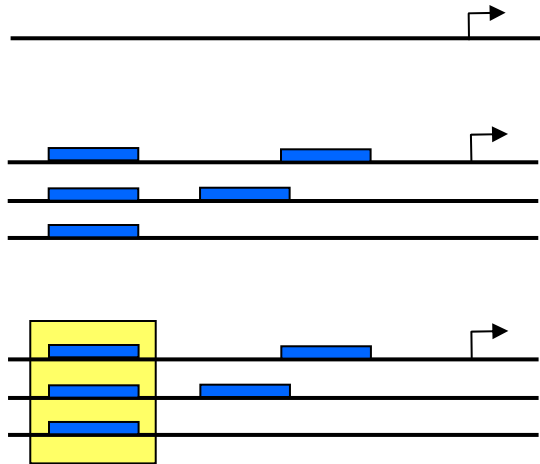


Binding and expression are correlated

PolII binding between humans and chimpanzee suggests extensive divergence

Identifying Transcription Factor Target Genes

Scan 2kb up-stream of transcription start site



1. Extract genomic sequence (-2kb of TSS)

2. Scan conserved regions for potential binding sites using TRANSFAC binding matrices

3. Identify conserved sites (Human/Chimp/Mouse)

| | TF 1 | TF 2 | ... | TF M |
|--------|------|------|-----|------|
| Gene 1 | √ | | √ | |
| Gene 2 | | | √ | |
| ... | √ | | | √ |
| Gene N | √ | √ | | |

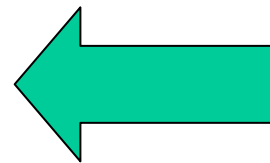


Table linking transcription factors and putative target genes

‘Gene Sets’ of target genes for each transcription factor

Gene Sets of Transcription Factor Targets

Molecular Signatures Database at Broad Institute
(<http://www.broad.mit.edu/gsea/msigdb>)

V\$NRSF_01 (Neuron Restrictive Silencing Factor)

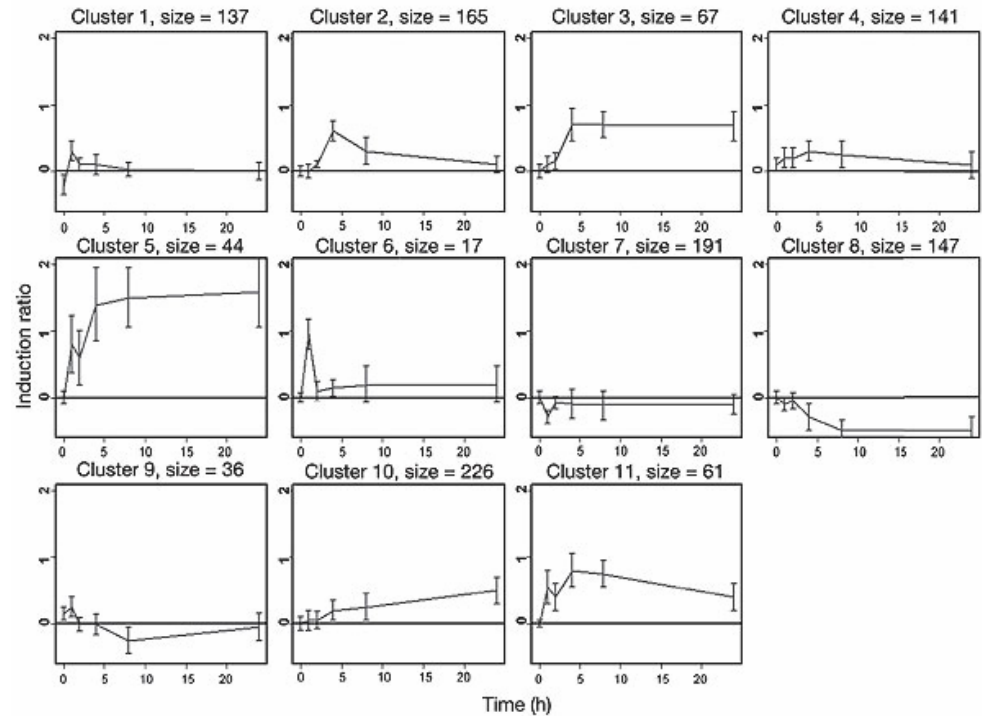
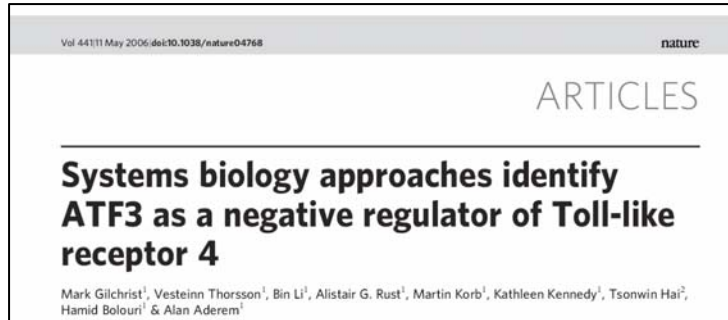
Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif
TTCAGCACCCACGGACAGMGCC which matches annotation for REST: RE1-silencing transcription factor

| | | | | | | |
|----------|---------|--------|----------|---------|----------|--------|
| ATP6V0A1 | RPIP8 | POU4F3 | FLJ42486 | L1CAM | SLC17A6 | TRIM9 |
| MAPK11 | DDX25 | SNAP25 | DRD3 | FGF12 | COL5A3 | SYT4 |
| BDNF | POMC | GABRB3 | TMEM22 | GRM1 | HES1 | |
| MGAT5B | TCF1 | PCSK2 | FLJ44674 | VIP | FLJ38377 | ZNF335 |
| GABRG2 | LHX3 | DNER | CHKA | NEFH | ZNF579 | CHAT |
| SCAMP5 | CDKN2B | SST | OGDHL | KCNH4 | SEZ6 | GLRA1 |
| HTR1A | RPH3A | PRG3 | NPPB | FGD2 | RNF13 | SYT6 |
| CHGA | SLC12A5 | ELAVL3 | KCNH8 | GDAP1L1 | HCN1 | DRD2 |
| HCN3 | PAQR4 | CALB1 | BARHL1 | SCN3B | CRYBA2 | TNRC4 |
| VEGF | RASGRF1 | NEF3 | OMG | KCNIP2 | CDK5R1 | ATP2B2 |
| HTR5A | PHYHIPL | SARM1 | GHSR | INA | PTPRN | DBC1 |
| CSPG3 | CHRN2 | GRIN1 | STMN2 | POU4F2 | APBB1 | GLRA3 |

Gene sets can also be defined manually

Which TFs are driving dynamics of each cluster?

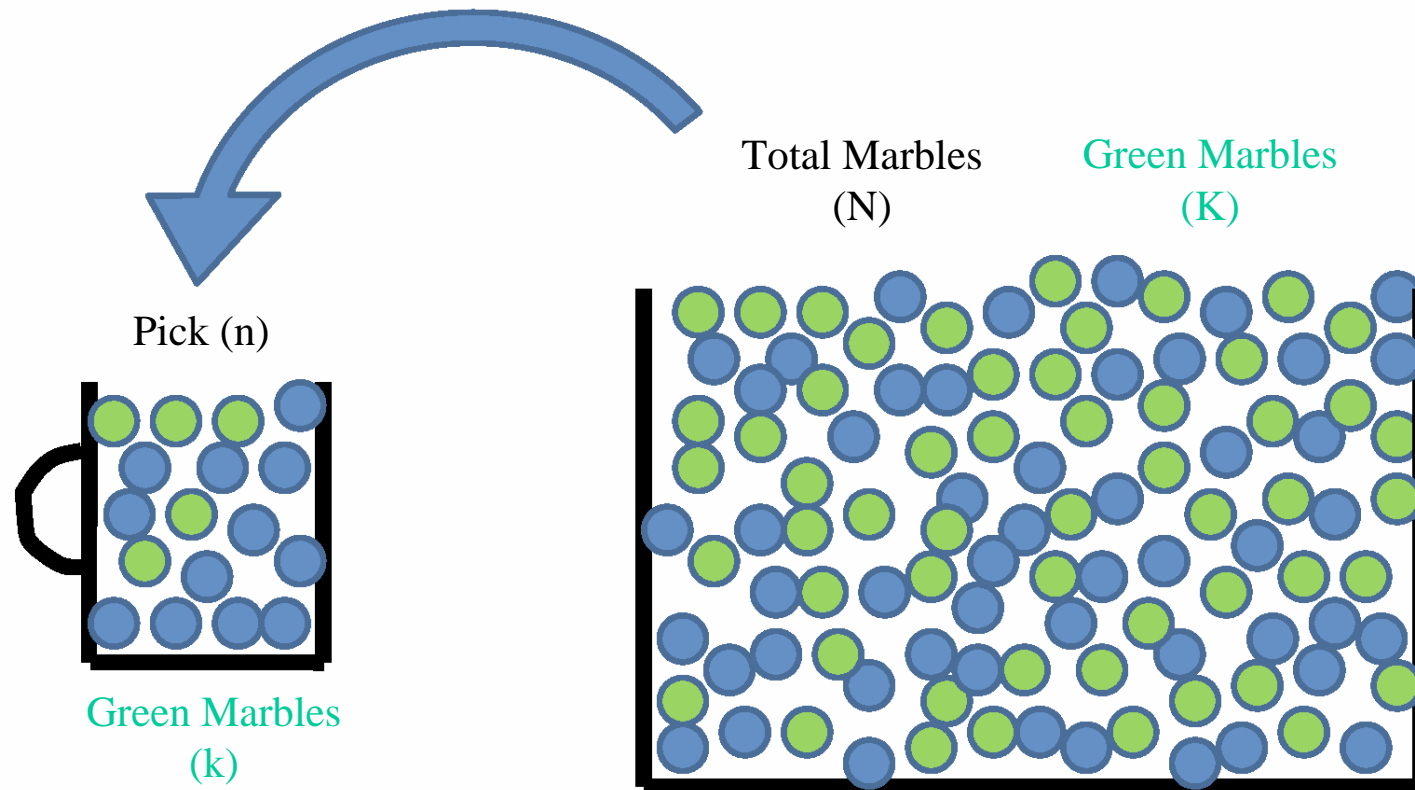
Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)



Look for TF targets that are 'over-represented' in a cluster

Over-Representation Analysis

If you draw n marbles at random, what is probability of k green ones?



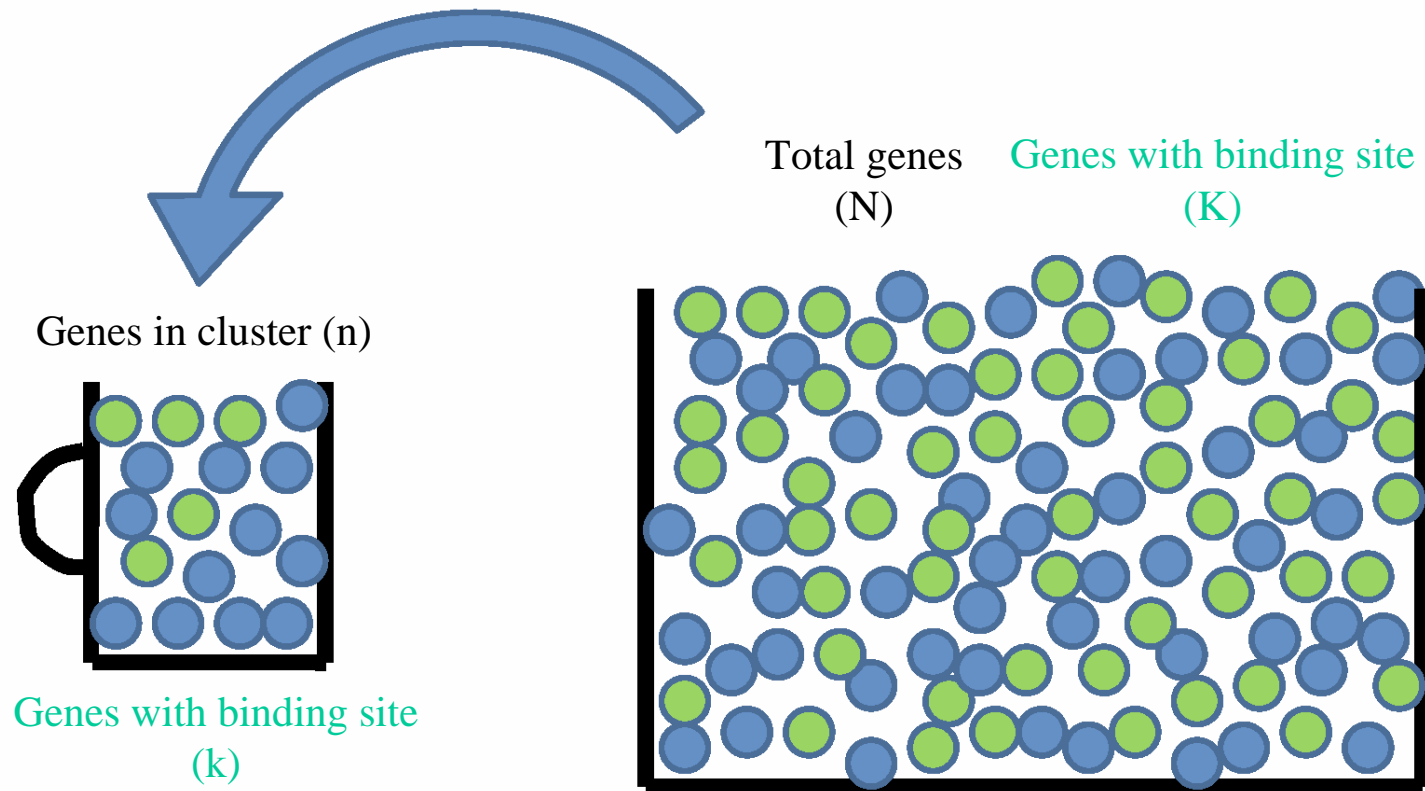
Adapted from Can (John) Bruce

Hypergeometric Distribution:
Probability of k green if n is random sample

$$P(k | n, K, N) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

Over-Representation Analysis

Is set of TF targets over-represented among genes in cluster?



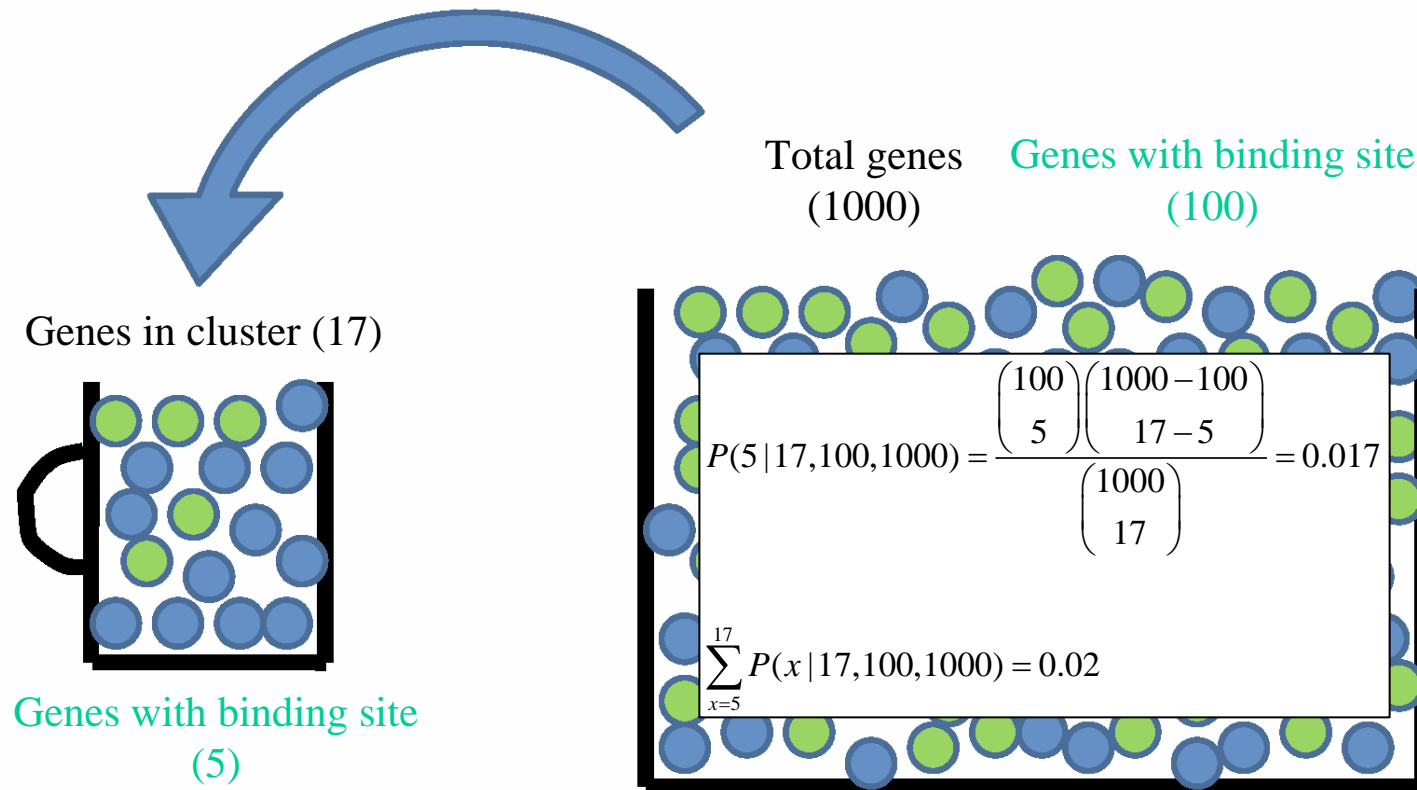
Adapted from Can (John) Bruce

Hypergeometric Distribution:

Probability of k TF targets if cluster is random sample

Over-Representation Analysis

If 17 genes in cluster, 5 with transcription factor binding site...

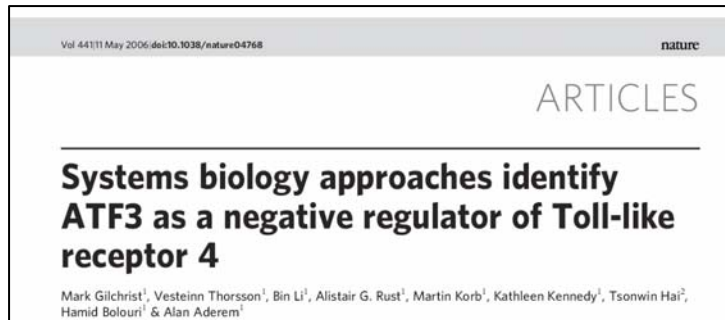


Adapted from Can (John) Bruce

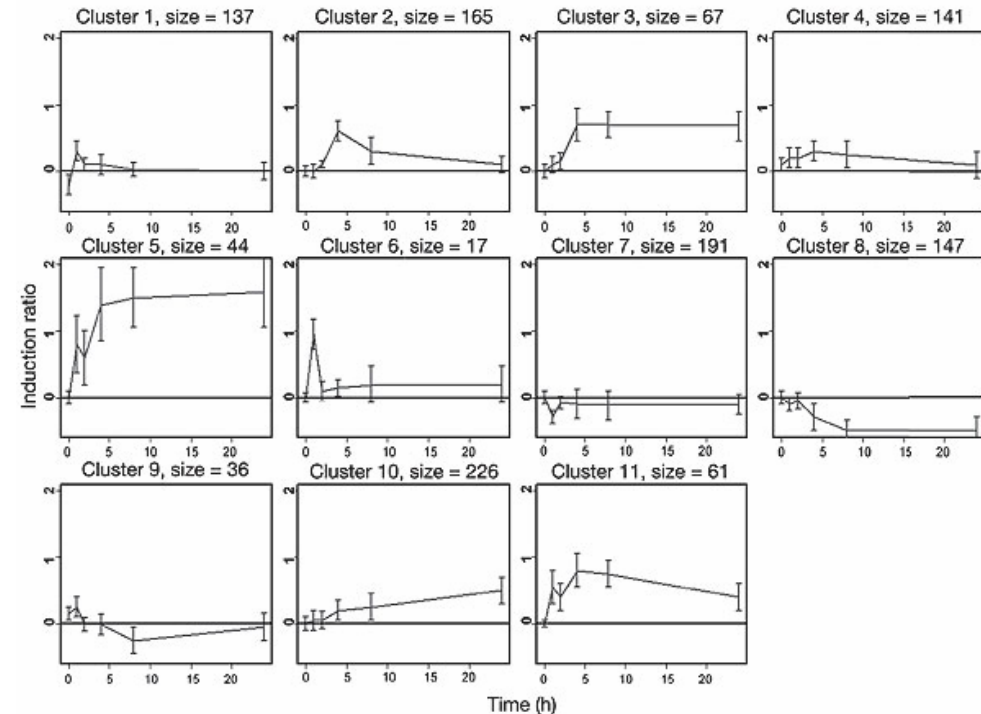
Must choose threshold to define “differential expression”

Identifying regulators of TLR responses

Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)



K-means clustering defined 11 groups of genes comprising regulated 'waves' of transcription

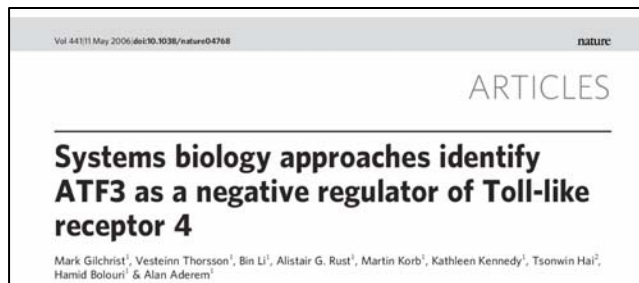


What is the role of ATF3?

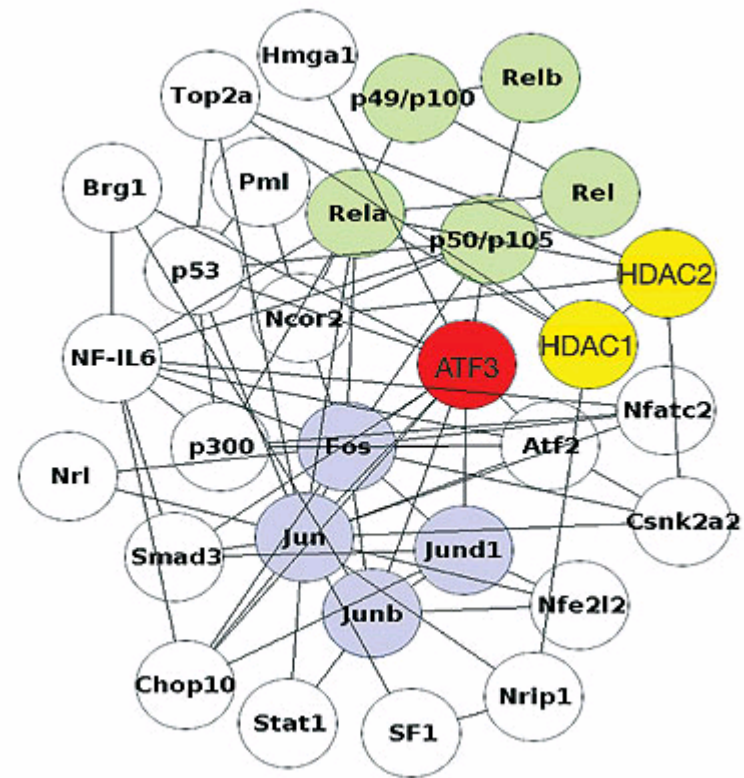
Network Analysis: role of ATF3?

“Guilt by association”

Highly connected proteins are likely to be functionally related



protein–protein interaction network

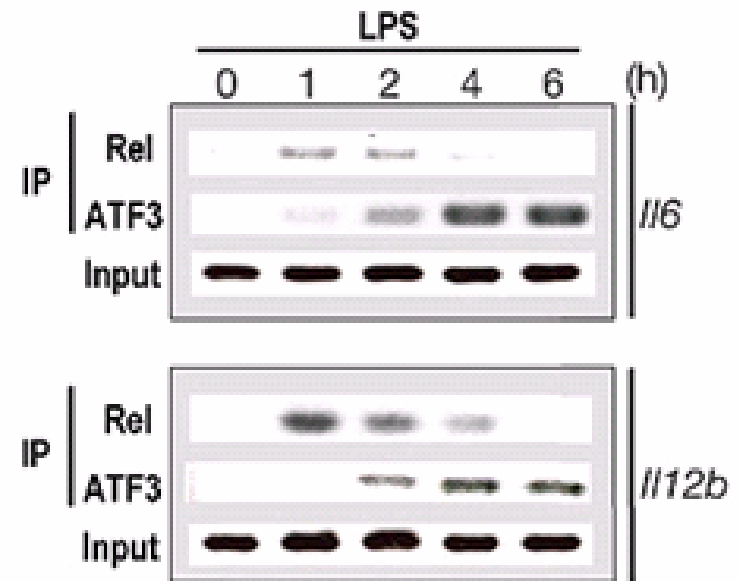
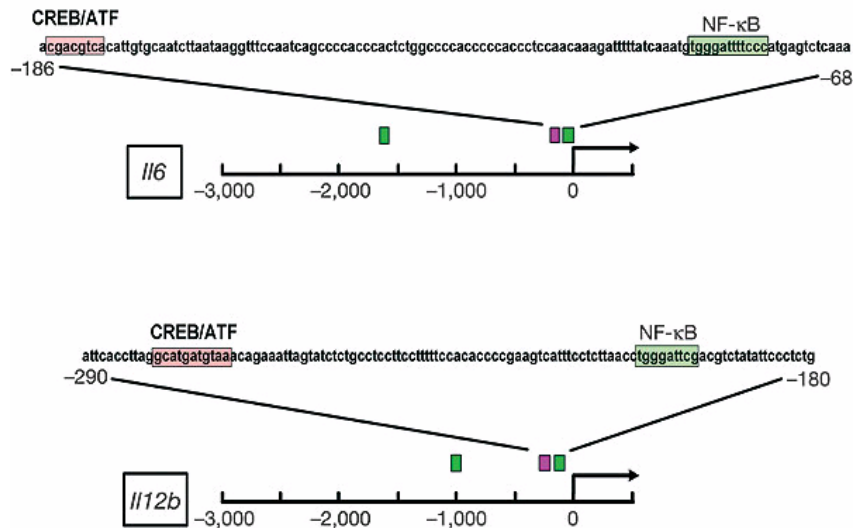


ATF3 (red) interacts with AP1 (light blue) and NF- B (light green) TF complexes

What is the role of ATF3?

Identified many target genes with nearby ATF3 and NFkB binding sites

Temporal recruitment of ATF3 and Rel to Il6 and Il12b promoters

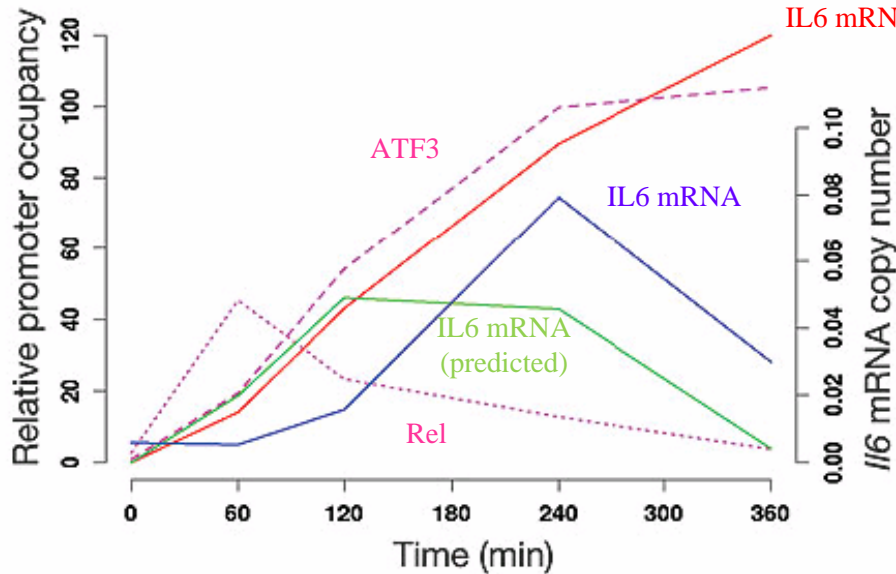


ChIP assays

How does ATF3 regulate IL6 and IL12b?

What is the role of ATF3?

Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)



Model used to predict IL6 mRNA as function of Rel and ATF3 binding

mRNA degradation

Influence on transcription

$$\tau \frac{d(Il6)}{dt} = -Il6 + g(\beta_{Rel}Rel + \beta_{ATF3}ATF3)$$

Change in IL6 mRNA

Predict

ATF3 is a negative regulator of IL6 and IL12b

Which TFs are driving dynamics of each cluster?

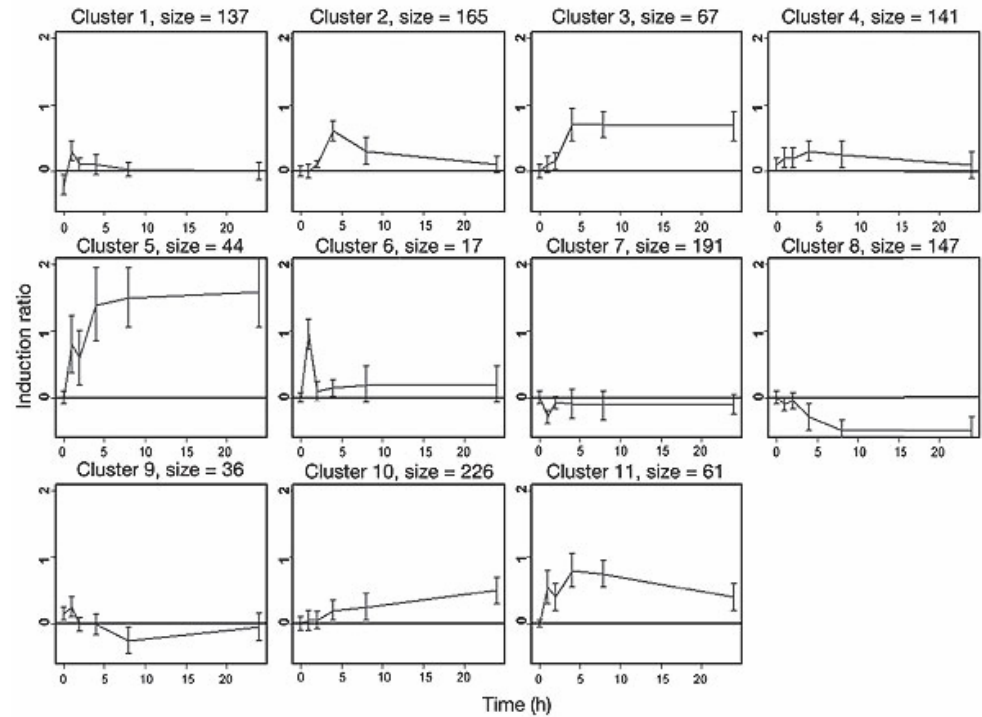
Temporal activation of macrophages by TLR4 agonist bacterial lipopolysaccharide (LPS)

Vol 441 | 11 May 2006 | doi:10.1038/nature04768 nature

ARTICLES

Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4

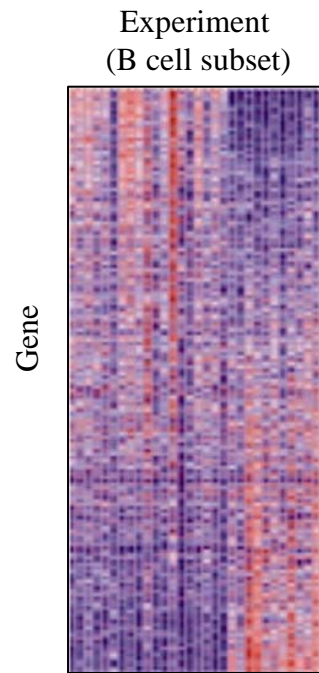
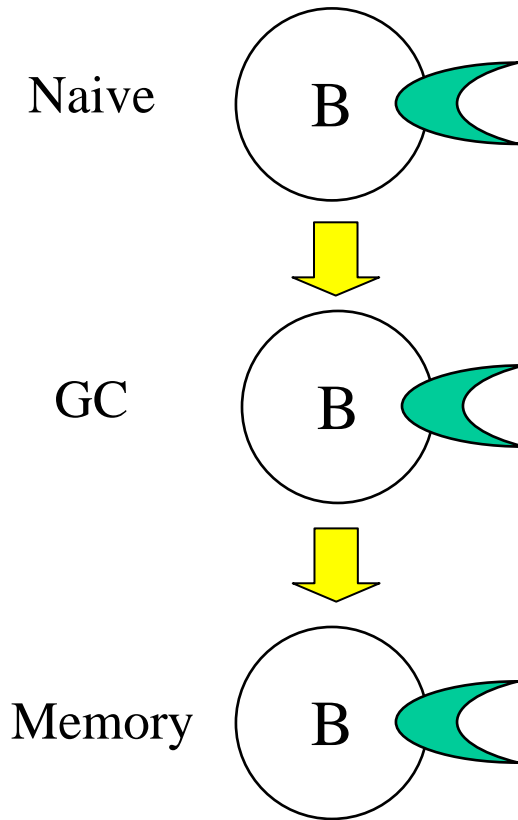
Mark Gilchrist¹, Vesteinn Thorsson¹, Bin Li¹, Alistair G. Rust¹, Martin Korb¹, Kathleen Kennedy¹, Tsonwin Hai², Hamid Bolouri¹ & Alan Aderem¹



Need to assign genes to single cluster

Can we identify TFs driving B cell differentiation?

Implicate TFs by analyzing behavior of target genes

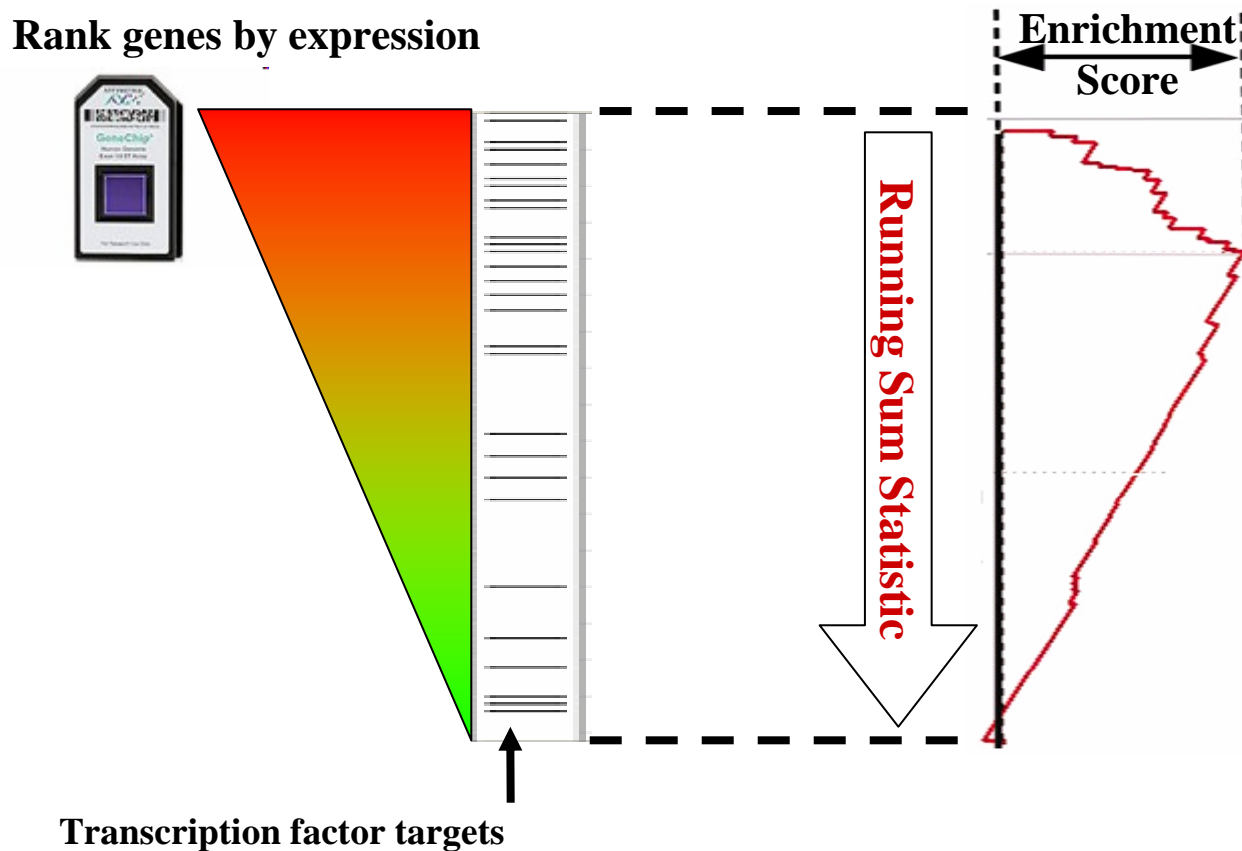


If genes targeted by particular transcription factor are differentially expressed, then the transcription factor is likely to play role

Need to identify which genes are differentially-expressed

Gene Set Enrichment Analysis (GSEA)

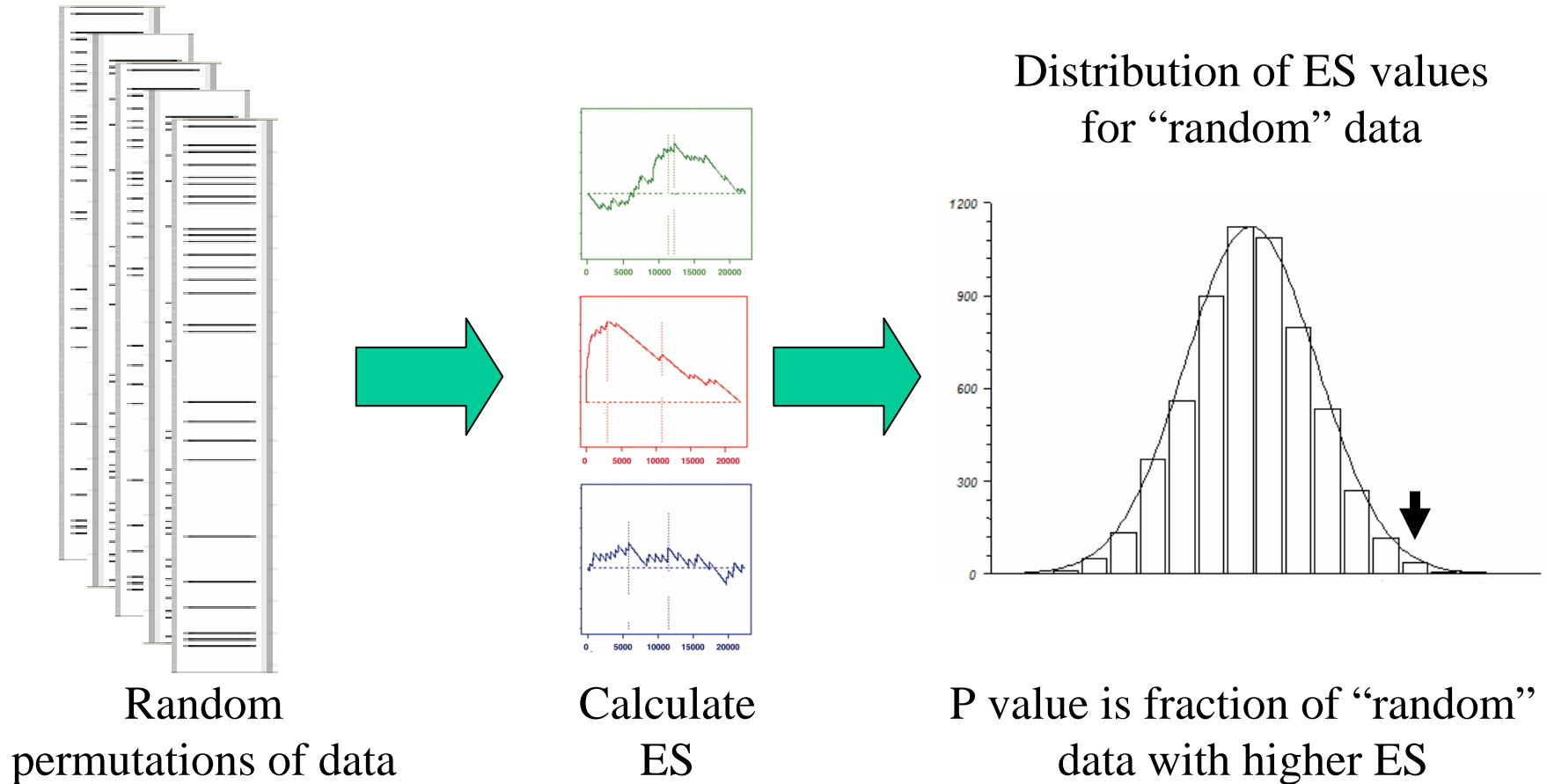
Are TF targets **enriched** among most differentially expressed genes?



Does not require a threshold for differential expression

Gene Set Enrichment Analysis (GSEA)

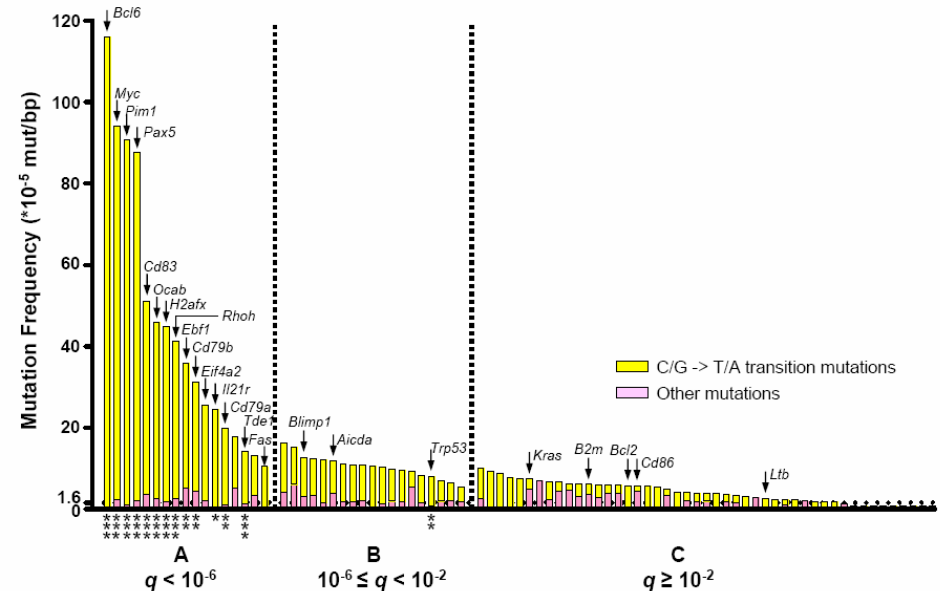
What is distribution for enrichment score (ES) under null hypothesis?



Permute class labels or genes to estimate null distribution

Can we identify TFs driving mutation targeting?

Are particular motifs enriched among the most mutated genes?



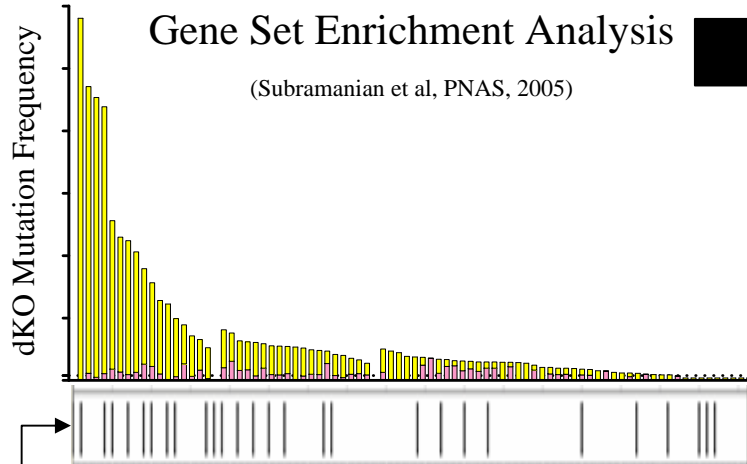
If genes targeted by particular transcription factor tend to be more mutated, then the transcription factor is likely to play role

Target genes identified by presence of binding sites

Does E2a influence AID targeting?

Are transcription factor target genes **enriched** among the most mutated?

Computational screen including E2a + all TRANSFAC transcription factors



Genes with binding sites (+/- 2Kb)
Found through computational screen



E2a binding sites top (and only) significant hits

| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val |
|----|--------------------------------------|-----------------------------|------|------|------|-----------|-----------|
| 1 | CAG_GC TG\$E2A MOTIF | Details ... | 28 | 0.35 | 1.84 | 0.017 | 0.155 |
| 2 | CAGGTG V\$E12_Q6 | Details ... | 20 | 0.35 | 1.66 | 0.028 | 0.196 |
| 3 | CAGCTG V\$AP4_Q5 | Details ... | 12 | 0.41 | 1.62 | 0.037 | 0.158 |
| 4 | CTTTGA V\$LEF1_Q2 | Details ... | 10 | 0.31 | 1.13 | 0.295 | 0.825 |
| 5 | TGGAAA V\$NFAT_Q4_Q1 | Details ... | 13 | 0.28 | 1.13 | 0.300 | 0.665 |
| 6 | AACTTT UNKNOWN | Details ... | 10 | 0.29 | 1.07 | 0.357 | 0.654 |
| 7 | GGGTGGRR V\$PAX4_Q3 | | | | | | |
| 8 | GGGCGGR V\$SP1_Q6 | | | | | | |
| 9 | V\$OCT1_B | | | | | | |
| 10 | TTGTTT V\$FOXO4_Q1 | | | | | | |
| 11 | GTGCCTT_MIR-506 | | | | | | |

Yes, E2a sites enriched among mutated genes in UNG/MSH2 dKO mice

Other Applications of Gene Set Enrichment Analysis

Molecular Signatures Database at Broad Institute

| collection | contents |
|---|--|
| c1: positional gene sets (view gene sets) | Gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. (Cytogenetic locations were parsed from HUGO, October 2006, and Unigene, build 197. When there were conflicts, the Unigene entry was used.) These gene sets are helpful in identifying effects related to chromosomal deletions or amplifications, dosage compensation, epigenetic silencing, and other regional effects. |
| c2: curated gene sets (view gene sets) | Gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. The gene set card for each gene set lists its source. details |
| CP: Canonical Pathways (view gene sets) | Gene sets from the pathway databases. Usually, these gene sets are canonical representations of a biological process compiled by domain experts. details |
| CGP: chemical and genetic perturbations (view gene sets) | Gene sets that represent gene expression signatures of genetic and chemical perturbations. A number of these gene sets come in pairs: an xxx_UP (xxx_DN) gene set representing genes induced (repressed) by the perturbation. The gene set card for each gene set lists the PubMed citation on which it is based. |
| c3: motif gene sets (view gene sets) | Gene sets that contain genes that share a <i>cis</i> -regulatory motif that is conserved across the human, mouse, rat, and dog genomes. The motifs are catalogued in Xie, et al. (2005, <i>Nature</i> 434, 338-345) and represent known or likely regulatory elements in promoters and 3'-UTRs. These gene sets make it possible to link changes in a microarray experiment to a conserved, putative <i>cis</i> -regulatory element. |
| TFT: transcription factor targets (view gene sets) | Gene sets that contain genes that share a transcription factor binding site defined in the TRANSFAC (version 7.4, http://www.gene-regulation.com/) database. Each of these gene sets is annotated by a TRANSFAC record. |
| MIR: miRNA targets (view gene sets) | Gene sets that contain genes that share a 3'-UTR microRNA binding motif. |
| c4: computational gene sets (view gene sets) | Gene sets defined by expression neighborhoods centered on 380 cancer-associated genes (Brentani, Caballero et al. 2003). This collection is identical to that previously reported in (Subramanian, Tamayo et al. 2005). details |

Gene sets can also be defined manually

Gene Ontology

Structured, controlled vocabularies (ontologies) that describe gene products in terms of associated biological processes, cellular components and molecular functions

Organization and functional annotation of molecular aspects of cellular system

GO: 0050864 regulation of B cell activation (a)

Any process that modulates the frequency, rate or extent of B cell activation

Term information Ancestor chart Ancestor table Child Terms **Protein Annotation** Statistics

Visual display of a section of the GO directed acyclic graph (DAG) for a single Go term. Terms in the GO are linked to parent (more general) terms and often to child (more specific) terms by one or more 'relationships'

Parent
is a
Term
part of
Child
regulates
Regulation
+ve regulates
+ve regulation
-ve regulates
-ve regulation

GO: 0050864 regulation of B cell activation (b)

Any process that modulates the frequency, rate or extent of B cell activation

| Symbol | GO ID | Reference | Ev | With | Taxon | From | GO Term name |
|--------|------------|-----------|-----|----------|-------|---------|---------------------|
| Il4 | GO:0048295 | 14988498 | IDA | | 10090 | MGI | Positive regulation |
| Fcgr2 | GO:0030889 | 8552190 | IMP | | 10090 | MGI | Negative regulation |
| PTPRC | GO:0030890 | 1793833 | IMP | | 9606 | UniProt | Positive regulation |
| Il7 | GO:0030890 | 12970760 | IDA | | 10090 | MGI | Positive regulation |
| Il7 | GO:0030890 | P13232 | ISS | P10168 | 9606 | UniProt | Positive regulation |
| Tcf3 | GO:0030890 | P15806-2 | ISS | P15923-2 | 10090 | UniProt | Positive regulation |
| TCF3 | GO:0030890 | 11509675 | IMP | | 9606 | UniProt | Positive regulation |

(Lovering et al, Immunology, 2008)

Annotations include evidence code (experimental and computational)

For more information:

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Message from ISCB

Getting Started in Computational Immunology

Steven H. Kleinstein*

Interdepartmental Program in Computational Biology and Bioinformatics, and Department of Pathology, Yale University School of Medicine, New Haven, Connecticut, United States of America



Or send me email at: steven.kleinstein@yale.edu