http://GREAT.stanford.edu/:
Genomic Regions Enrichment of Annotations Tool

Gill Bejerano
Dept. of Developmental Biology &
Dept. of Computer Science
Stanford University

http://bejerano.stanford.edu
Human Gene Regulation

$10^{13}$ different cells in an adult human.

All these cells have the same Genome.

20,000 Genes encode how to make proteins.

1,000,000 Genomic “switches” determine which and how much proteins to make.

Hundreds of different cell types.
Most Non-Coding Elements likely work in cis…

“IRX1 is a member of the Iroquois homeobox gene family. Members of this family appear to play *multiple roles* during pattern formation of vertebrate embryos.”
Many non-coding elements tested are cis-regulatory
Lab Focus: Human Cis-Regulation

GENOME

DEVELOPMENT

EVOLUTION

DISEASE

http://bejerano.stanford.edu
2,000 different proteins can bind specific DNA sequences.

A regulatory region encodes 3-10 such protein binding sites. When all are bound by proteins the regulatory region turns “on”, and the nearby gene is activated to produce protein.
ChIP-Seq: first glimpses of the regulatory genome in action

1. Cell Nucleus
2. Crosslink Protein and Shear DNA
3. Add Protein-Specific Antibody
4. Immunoprecipitate and purify complexes
5. Reverse Crosslinks, Purify DNA and prepare for sequencing
6. Sequence DNA fragment and map to genome

ACTGGTGACAGGACG

Cis-regulatory peak

http://bejerano.stanford.edu
What is the transcription factor I just assayed doing?

• Collect known literature of the form
  • Function A: Gene1, Gene2, Gene3, ...
  • Function B: Gene1, Gene2, Gene3, ...
  • Function C: ...
• Ask whether the binding sites you discovered are preferentially binding (regulating) any one or more of the functions listed above.
• Form hypothesis and perform further experiments.
Example: inferring functions of Serum Response Factor (SRF) from its ChIP-seq binding profile

![Gene transcription start site](image)

![SRF binding ChIP-seq peak](image)

- ChIP-seq identified SRF binding profile in human Jurkat cells\(^1\)
- 2,429 binding peaks genome-wide
- SRF is known as a “master regulator of the actin cytoskeleton”

---

Example: inferring functions of Serum Response Factor (SRF) from its ChIP-seq binding profile

Existing, **gene-based** method to analyze enrichment:

- Ignore distal binding events.
- Count affected genes.
- Rank by enrichment hypergeometric p-value.

\[ P = \Pr(k_\pi \geq 1 \mid n=2, K_\pi=3, N=8) \]

Gene transcription start site
SRF binding ChIP-seq peak
\( \pi \) Ontology term (e.g. ‘actin cytoskeleton’)

http://bejerano.stanford.edu
We have (reduced ChIP-Seq into) a gene list!
What is the gene list enriched for?

Pro: A lot of tools out there for the analysis of gene lists.
Cons: These tools are built for microarray analysis.
Does it matter ??

http://bejerano.stanford.edu
Gene-based enrichment results

- Original authors can only state: “basic cellular processes, particularly those related to gene expression” are enriched\(^1\)

<table>
<thead>
<tr>
<th>Term</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>nucleus</td>
<td>$5.18 \times 10^{-70}$</td>
</tr>
<tr>
<td>protein binding</td>
<td>$2.16 \times 10^{-50}$</td>
</tr>
<tr>
<td>cytoplasm</td>
<td>$6.67 \times 10^{-27}$</td>
</tr>
<tr>
<td>transcription</td>
<td>$4.13 \times 10^{-26}$</td>
</tr>
<tr>
<td>nucleotide binding</td>
<td>$1.04 \times 10^{-23}$</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>$1.92 \times 10^{-22}$</td>
</tr>
<tr>
<td>zinc ion binding</td>
<td>$5.76 \times 10^{-20}$</td>
</tr>
<tr>
<td>RNA binding</td>
<td>$3.38 \times 10^{-18}$</td>
</tr>
</tbody>
</table>

Where’s the signal? Top “actin” term is ranked #28 in the list.

\(^1\) Valouev A. et al., *Nat. Methods*, 2008
Associating only proximal peaks loses a lot of information

Restricting to proximal peaks often leads to complete loss of key enrichments
Associating distal peaks in a gene-based context is statistically inappropriate

- Associating SRF peaks with the nearest gene biases toward isolated genes

Example: In the human genome, 14% of genes annotated with GO term ‘multicellular organismal development’. But 33% of base pairs have a gene annotated with the term as its nearest up- or down-stream gene!
Associating distal peaks in a gene-based context is statistically inappropriate

- SRF ChIP-seq set has ~2,000 binding events in the genome.
- Gene-based enrichments from a set of 2,000 regions scattered randomly throughout the genome:

<table>
<thead>
<tr>
<th>Term</th>
<th>Bonferroni corrected p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>nervous system development</td>
<td>5x10^{-9}</td>
</tr>
<tr>
<td>system development</td>
<td>8x10^{-9}</td>
</tr>
<tr>
<td>anatomical structure development</td>
<td>7x10^{-8}</td>
</tr>
<tr>
<td>multicellular organismal development</td>
<td>1x10^{-7}</td>
</tr>
<tr>
<td>developmental process</td>
<td>2x10^{-6}</td>
</tr>
</tbody>
</table>

"Members of the Iroquois homeobox gene family play multiple roles during pattern formation of vertebrate embryos."

Many false positive terms for input sets typical of ChIP-seq (1-50k peaks)

http://bejerano.stanford.edu
Gene-based tools are unfit for assessing functional enrichments of cis-regulatory data

- Associating only proximal genomic regions to genes ignores a large fraction of binding data
  - Many associations are missed
  - Can lead to only general enrichments

- Associating distal genomic regions to genes biases toward isolated genes
  - Leads to spurious enrichments
GREAT accurately assesses statistical significance of non-coding enrichments

- GREAT’s genomic region-based statistical test accounts for variation in gene distribution
- For each genomic region in dataset, probability of hitting term $\pi$ calculated as the fraction of the genome that associates with a gene having that term

Since 33% of base pairs have a gene annotated with ‘multicellular organismal development’ as nearest up- or down-stream gene, GREAT assigns probability of each genomic region hitting the term as 33%.

http://bejerano.stanford.edu

GREAT accurately assesses statistical significance of non-coding enrichments

\[ P = \Pr_{\text{binom}}(k_{\pi} \geq 5 \mid n=6, p_{\pi}=0.44) \]

\[ p_{\pi} = 0.44 \text{ of genome annotated with } \pi \]
\[ n = 6 \text{ genomic regions} \]
\[ k_{\pi} = 5 \text{ genomic regions hit annotation} \]
GREAT accurately assesses statistical significance of non-coding enrichments

Randomly distributed sets of all sizes give negligible false positive enrichments using the genomic region-based test

http://bejerano.stanford.edu
How does GREAT know how to assign distal binding peaks to genes?

**Future:** High-throughput assays based on chromosome conformation capture (3C) methods will elucidate complex regulation mechanisms.

**Currently:** Flexible computational definitions allow assignment of peaks to nearest gene, nearest two genes, etc.
- Default: each gene has a “basal regulatory domain” of 5 kb up- and 1 kb downstream of transcription start site, extends to basal domain of nearest genes within 1 Mb.
- Though some associations may be missed or incorrect, in general signal richness and robustness is greatly improved by associating distal peaks.

http://bejerano.stanford.edu
Ensuring statistically significant enrichments are also “enrichments of significant effect”

• Consider a coin flipping experiment:
  • Flip 50,000 times \( \rightarrow \) 26,000 heads, 24,000 tails
  • One-sided p-value with null hypothesis of fair coin: \( 2 \times 10^{-19} \)
  • Fold enrichment = observed/expected = 26000/25000 = 1.04

• Consider another coin flipping experiment:
  • Flip 100 times \( \rightarrow \) 90 heads, 10 tails
  • One-sided p-value with null hypothesis of fair coin: \( 2 \times 10^{-17} \)
  • Fold enrichment = 90/50 = 1.80

GREAT requires statistical significance from binomial test, but also filters on fold enrichment to identify enrichments of large effect

http://bejerano.stanford.edu
GREAT infers many specific functions of SRF from its binding profile

### Top GREAT enrichments of SRF

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Term</th>
<th># Genes</th>
<th>Binomial P-value</th>
<th>Experimental support*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Ontology</td>
<td>actin cytoskeleton</td>
<td>30</td>
<td>7x10^{-9}</td>
<td>Miano et al. 2007</td>
</tr>
<tr>
<td></td>
<td>actin binding</td>
<td>31</td>
<td>5x10^{-5}</td>
<td>Miano et al. 2007</td>
</tr>
<tr>
<td>Pathway Commons</td>
<td>TRAIL signaling</td>
<td>32</td>
<td>5x10^{-7}</td>
<td>Bertolotto et al. 2000</td>
</tr>
<tr>
<td></td>
<td>Class I PI3K signaling</td>
<td>26</td>
<td>2x10^{-6}</td>
<td>Poser et al. 2000</td>
</tr>
<tr>
<td>TreeFam</td>
<td>FOS gene family</td>
<td>5</td>
<td>1x10^{-8}</td>
<td>Chai &amp; Tarnawski 2002</td>
</tr>
<tr>
<td>TF Targets</td>
<td>Targets of SRF</td>
<td>84</td>
<td>5x10^{-76}</td>
<td>Positive control</td>
</tr>
<tr>
<td></td>
<td>Targets of GABP#</td>
<td>28</td>
<td>4x10^{-9}</td>
<td>Natesan &amp; Gilman 1995</td>
</tr>
<tr>
<td></td>
<td>Targets of YY1</td>
<td>44</td>
<td>1x10^{-6}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Targets of EGR1</td>
<td>23</td>
<td>2x10^{-4}</td>
<td></td>
</tr>
</tbody>
</table>

* Known from literature – as in function is known, SOME of the genes are known, and the binding sites highlighted are NOT.

# Independent ChIP-seq of GABP shows 29% of SRF peaks co-localize with it

http://bejerano.stanford.edu

GREAT infers many specific functions of SRF from its binding profile

• By including distal binding events and using a genomic region-based test, GREAT identifies both specific known functions and novel hypotheses of SRF functions.

• GREAT tests of ChIP-seq profiles of other sequence-specific transcription factors (GABP, NRSF, Stat3) also identify specific known functions of the factors as enriched.

• Can GREAT also work on more general factors (e.g. the transcription-associated protein p300)?
DNA bound by p300 in limbs act as limb-specific enhancers of gene expression

p300 is a transcriptional co-activator
Thought to facilitate interactions between distal enhancers and target genes

2,105 limb peaks

http://bejerano.stanford.edu

[adapted from Visel et al., Nature, 2009]
GREAT infers specific functions of enhancers bound by p300 that gene-based tests do not

DAVID\(^1\) enrichments

<table>
<thead>
<tr>
<th>Enrichment Score: 6.31</th>
<th>Count</th>
<th>P_Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription regulation</td>
<td>43</td>
<td>4.6E-10</td>
</tr>
<tr>
<td>dna-binding</td>
<td>44</td>
<td>6.2E-10</td>
</tr>
<tr>
<td>Transcription</td>
<td>43</td>
<td>0.80E-10</td>
</tr>
<tr>
<td>nucleus</td>
<td>74</td>
<td>1.5E-9</td>
</tr>
<tr>
<td>nucleus</td>
<td>86</td>
<td>3.7E-9</td>
</tr>
<tr>
<td>transcription regulator activity</td>
<td>41</td>
<td>6.4E-9</td>
</tr>
<tr>
<td>regulation of cellular process</td>
<td>80</td>
<td>1.6E-8</td>
</tr>
</tbody>
</table>

GREAT enrichments

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Term Name</th>
<th>Binom FDR</th>
<th>Binom Fold Enrichment</th>
<th>Gene Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Biological Process</td>
<td>embryonic limb morphogenesis</td>
<td>1.1230e-27</td>
<td>3.9271</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>limb morphogenesis</td>
<td>2.9129e-26</td>
<td>3.5370</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>embryonic morphogenesis</td>
<td>4.9287e-26</td>
<td>2.7864</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>limb development</td>
<td>8.0925e-26</td>
<td>3.3920</td>
<td>50</td>
</tr>
</tbody>
</table>

Mouse Phenotype

<table>
<thead>
<tr>
<th>Term Name</th>
<th>Enrichment Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>abnormal craniofacial morphology</td>
<td>8.2588e-49</td>
</tr>
<tr>
<td>abnormal axial skeleton morphology</td>
<td>3.2434e-45</td>
</tr>
<tr>
<td>abnormal limbs/digits/tail morphology</td>
<td>2.8133e-43</td>
</tr>
<tr>
<td>abnormal skull morphology</td>
<td>5.5066e-41</td>
</tr>
<tr>
<td>abnormal craniofacial bone morphology</td>
<td>5.3360e-40</td>
</tr>
<tr>
<td>abnormal limb morphology</td>
<td>7.4781e-39</td>
</tr>
<tr>
<td>abnormal paw/hand/forearm morphology</td>
<td>8.5767e-36</td>
</tr>
</tbody>
</table>

MGI Expression: Detected

<table>
<thead>
<tr>
<th>Term Name</th>
<th>Enrichment Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS19_limb</td>
<td>7.4628e-49</td>
</tr>
<tr>
<td>TS19_forelimb bud</td>
<td>5.0643e-38</td>
</tr>
<tr>
<td>TS22_upper jaw</td>
<td>7.9650e-36</td>
</tr>
<tr>
<td>TS20_limb</td>
<td>1.0863e-35</td>
</tr>
<tr>
<td>TS16_hindlimb bud</td>
<td>2.4503e-34</td>
</tr>
</tbody>
</table>

[1] Huang et al., NAR, 2007

Peaks near these genes may be the most promising candidates for individual characterization of limb function.

http://bejerano.stanford.edu
Comparing association rules for p300 limb enriched terms emphasizes importance of distal binding

- If distal p300 limb peaks are accurately mapped to their target genes, we expect enrichment for limb-related signals when compared to simulation
- Simulation definition:
  - Fix positions of p300 limb proximal binding peaks (within 2 kb of TSS)
  - Shuffle all distal binding peaks to random spots in the genome
  - See how many limb-related genes and peaks are identified in random shuffle

Actual p300 peaks: 5 peaks hit 2 limb morphogenesis genes
Simulation 1: 2 peaks hit 1 limb morphogenesis gene

http://bejerano.stanford.edu
Comparing association rules for p300 limb enriched terms emphasizes importance of distal binding

- Actual p300 peaks: 5 peaks hit 2 limb morphogenesis genes
  - Simulation 1: 2 peaks hit 1 limb morphogenesis gene
  - Simulation 2: 1 peak hit 1 limb morphogenesis gene

- If distal p300 limb peaks are accurately mapped to their target genes, we expect enrichment for limb-related signals when compared to simulation

- Simulation definition:
  - Fix positions of p300 limb proximal binding peaks (within 2 kb of TSS)
  - Shuffle all distal binding peaks to random spots in the genome
  - See how many limb-related genes and peaks are identified in random shuffle

- Compare actual p300 limb data to average results of 1,000 random shuffles

http://bejerano.stanford.edu
Comparing association rules for p300 limb enriched terms emphasizes importance of distal binding

The excess of limb-related genes and peaks picked by actual p300 limb data over random shuffles strongly suggests the distal binding peaks are functionally relevant to their assigned genes.

Gene Ontology: Embryonic limb morphogenesis

http://bejerano.stanford.edu
GREAT analysis of ChIP-seq datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Author/DAVID Proximal promoter</th>
<th>GREAT basal-extension* Up to 1,000 kb</th>
<th>“Gene-based GREAT” Proximal promoter 2 kb</th>
<th>GREAT basal+extension* Up to 50 kb</th>
<th>GREAT two nearest genes Up to 1,000 kb</th>
<th>GREAT single nearest gene Up to 1,000 kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRF</td>
<td>Table 1a</td>
<td>Table 1b</td>
<td>Sup. Table 7</td>
<td>Sup. Table 9</td>
<td>Sup. Table 10</td>
<td>Sup. Table 11</td>
</tr>
<tr>
<td>p300 Limb</td>
<td>Sup. Table 12a</td>
<td>Sup. Table 12b</td>
<td>Sup. Table 13</td>
<td>Sup. Table 14</td>
<td>Sup. Table 15</td>
<td>Sup. Table 16</td>
</tr>
<tr>
<td>p300 Forebrain</td>
<td>Sup. Table 17a</td>
<td>Sup. Table 17b</td>
<td>Sup. Table 18</td>
<td>Sup. Table 19</td>
<td>Sup. Table 20</td>
<td>Sup. Table 21</td>
</tr>
<tr>
<td>p300 Midbrain</td>
<td>Sup. Table 22a</td>
<td>Sup. Table 22b</td>
<td>Sup. Table 23</td>
<td>Sup. Table 24</td>
<td>Sup. Table 25</td>
<td>Sup. Table 26</td>
</tr>
<tr>
<td>p300 mESC</td>
<td>Sup. Table 27a</td>
<td>Sup. Table 27b</td>
<td>Sup. Table 28</td>
<td>Sup. Table 29</td>
<td>Sup. Table 30</td>
<td>Sup. Table 31</td>
</tr>
<tr>
<td>Stat3</td>
<td>Sup. Table 32a</td>
<td>Sup. Table 32b</td>
<td>Sup. Table 33</td>
<td>Sup. Table 34</td>
<td>Sup. Table 35</td>
<td>Sup. Table 36</td>
</tr>
<tr>
<td>NRSF</td>
<td>Sup. Table 37a</td>
<td>Sup. Table 37b</td>
<td>Sup. Table 38</td>
<td>Sup. Table 39</td>
<td>Sup. Table 40</td>
<td>Sup. Table 41</td>
</tr>
<tr>
<td>GABP</td>
<td>Sup. Table 42a</td>
<td>Sup. Table 42b</td>
<td>Sup. Table 43</td>
<td>Sup. Table 44</td>
<td>Sup. Table 45</td>
<td>Sup. Table 46</td>
</tr>
</tbody>
</table>

- GREAT reveals strong enrichments for experimentally-validated and novel, testable functions\(^1\)
- Distal binding events are essential to recover known functions
  - Restricting regulatory domain extension to 50 kb retains many enriched terms but omits roughly half of both the binding events and the genes implicated
- Exact distal association rule not critical: all behave similarly

http://bejerano.stanford.edu  
GREAT data integrated

- Twenty ontologies spanning broad categories of biology
- 44,832 total ontology terms tested in each GREAT run

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene Ontology</strong></td>
<td><strong>Gene Expression Data</strong></td>
</tr>
<tr>
<td>GO Molecular Function</td>
<td>MGI Expression: Detected</td>
</tr>
<tr>
<td>(2,800 terms)</td>
<td>(6,700)</td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>MGI Expression: Not Detected</td>
</tr>
<tr>
<td>(5,215)</td>
<td>(3,079)</td>
</tr>
<tr>
<td>GO Cellular Component</td>
<td>MSigDB Perturbation</td>
</tr>
<tr>
<td>(834)</td>
<td>(911)</td>
</tr>
<tr>
<td><strong>Phenotype Data and Human Disease</strong></td>
<td><strong>Regulatory Motifs</strong></td>
</tr>
<tr>
<td>Mouse Phenotype</td>
<td>MSigDB Predicted Promoter Motifs</td>
</tr>
<tr>
<td>(5,781)</td>
<td>(615)</td>
</tr>
<tr>
<td>MSigDB Cancer Neighborhood*</td>
<td>Transcription Factor Targets</td>
</tr>
<tr>
<td>(427)</td>
<td>(19)</td>
</tr>
<tr>
<td>MSigDB Cancer Modules*</td>
<td>MSigDB miRNA Motifs</td>
</tr>
<tr>
<td>(456)</td>
<td>(222)</td>
</tr>
<tr>
<td><strong>Pathway Data</strong></td>
<td>miRNA Targets</td>
</tr>
<tr>
<td>PANTHER Pathway</td>
<td>(9)</td>
</tr>
<tr>
<td>(150)</td>
<td></td>
</tr>
<tr>
<td>Pathway Commons</td>
<td></td>
</tr>
<tr>
<td>(1,253)</td>
<td></td>
</tr>
<tr>
<td>BioCyc Pathway</td>
<td></td>
</tr>
<tr>
<td>(288)</td>
<td></td>
</tr>
<tr>
<td>MSigDB Pathway</td>
<td></td>
</tr>
<tr>
<td>(706)</td>
<td></td>
</tr>
</tbody>
</table>

* Ontology only supported in human.

http://bejerano.stanford.edu  Michael Hiller  30
GREAT implementation

- Can handle datasets of hundreds of thousands of genomic regions
- 60-100x speedup over early prototype
  - Testing a single ontology term now takes ~1 ms
  - Enables real-time calculation of enrichment results for all 20 of our ontologies

http://bejerano.stanford.edu

Cory McLean
GREAT web app: input page

http://great.stanford.edu

Genomic Regions Enrichment of Annotations Tool

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT, the Genomic Regions Enrichment of Annotations Tool, assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions, including those identified by chromatin immunoprecipitation experiment or by a computational screen, such as conservation in a given clade.

Input choices

Species assembly
- Human: NCBI build 36.1 (UCSC hg18, Mar/2006)
- Mouse: NCBI build 37 (UCSC mm9, Jul/2007)

Can I use a different species or assembly?

Test regions
- BED file:
- BED data:

Pick a genome assembly

Input BED regions of interest

What should my test regions file contain?

How can I create a test set from a UCSC Genome Browser annotation track?

http://bejerano.stanford.edu
GREAT documentation on detailed wiki

Mailing lists:
great-announce@lists.stanford.edu
great-users@lists.stanford.edu

Welcome
The Bejerano Lab at Stanford University developed GREAT and hosts the GREAT web server.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT, the Genomic Regions Enrichment of Annotations Tool, assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions, including those identified by chromatin immunoprecipitation experiment or by a computational screen, such as conservation in a gene clade.

GREAT Help

General
- Overview - What is GREAT useful, and for which users should I prefer it to other annotation tools?
- Citation - How do I cite GREAT?
- Version History
- About Us
- Contact Us

Input
- Genome Assemblies - Which genome assemblies does GREAT support, and can I use other assemblies or species?
- File Formats - What should my test regions and background regions files contain?
- Background Sets - Why should I use a background set in evaluating enrichments?

Output
- Statistics - How does GREAT calculate enrichments, and how should I interpret my results?
- Output - What output does GREAT provide?
- Genes - Which set of genes does GREAT use, and how does GREAT determine a single transcription start site for a gene?
- Ontologies - What is an ontology, and what data do the ontologies in GREAT provide?

http://bejerano.stanford.edu

Aaron Wenger
GREAT web app: 
(Optional): alter association rules

http://great.stanford.edu

Three association rule choices

Associate genomic regions with genes

GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two step process. First, each gene is assigned a regulatory domain. Then, a genomic region is associated with all genes whose regulatory region it overlaps.

Note that all distances are measured from the transcription start site of the gene's canonical isoform.

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

Three association rule choices

- Basal plus extension
- Two nearest genes
- Single nearest gene

Include curated regulatory domains

Submit  Reset

http://bejerano.stanford.edu  [adapted from Spitz, Gonzalez, & Duboule, Cell, 2003]  34
Additional ontologies, term statistics, multiple hypothesis corrections, etc.

Ontology-specific enrichments

Display filters

Multiple hypothesis correction options

Data export options

http://bejerano.stanford.edu
GREAT web app: term details page

Genes annotated as “actin binding” with associated genomic regions

Genomic regions annotated with “actin binding”

Frame holding http://www.geneontology.org definition of “actin binding”

http://bejerano.stanford.edu
GREAT web app: export data

Global Controls

Export
- Export All Tables
- Export All Tables
- Shown data in all tables as HTML
- Shown data in all tables as .tsv

HTML output displays all selected rows and columns

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Term Name</th>
<th>Binom FDR Q-Val</th>
<th>Binom Fold Enrichment</th>
<th>Observed Gene Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Biological Process</td>
<td>central nervous system development</td>
<td>8.2845e-35</td>
<td>2.3002</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>brain development</td>
<td>7.7036e-23</td>
<td>2.3490</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>forebrain development</td>
<td>2.4169e-25</td>
<td>2.6771</td>
<td>58</td>
</tr>
<tr>
<td>Mouse Phenotype</td>
<td>abnormal neurogenesis</td>
<td>2.7377e-35</td>
<td>2.4125</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>abnormal brain white matter morphology</td>
<td>3.3748e-33</td>
<td>2.8781</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>abnormal tract</td>
<td>8.9150e-32</td>
<td>2.6702</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>abnormal brain commissure morphology</td>
<td>3.5512e-30</td>
<td>2.9718</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>abnormal corpus callosum morphology</td>
<td>2.1780e-27</td>
<td>3.2697</td>
<td>38</td>
</tr>
<tr>
<td>NCI Expression: Detected</td>
<td>TS21_thalamus</td>
<td>5.7670e-49</td>
<td>2.4054</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>TS17_brain</td>
<td>1.2668e-48</td>
<td>2.0581</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>TS21_midbrain</td>
<td>1.7767e-45</td>
<td>2.5506</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>TS21_hypothalamus</td>
<td>6.0287e-45</td>
<td>2.4920</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>TS21_cerebral cortex</td>
<td>1.3484e-43</td>
<td>2.0736</td>
<td>164</td>
</tr>
</tbody>
</table>

Tab-separated values also available for additional postprocessing

http://bejerano.stanford.edu
**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

- **clade:** Mammal
- **genome:** Human
- **assembly:** Mar. 2006 (NCBI36/hg18)
- **group:** Comparative Genomics
- **track:** Ultra Conserved

**Table**:
- uc18
- describe table schema

**Region**:
- genome
- ENCODE Pilot regions
- position: chrX:151073054-151383975

**Identifiers (names/accessions)**:
- paste list
- upload list

**Filter**:
- create

**Intersection**:
- create

**Correlation**:
- create

**Output format**:
- all fields from selected table
- Send output to Galaxy

**Output file**:
- (leave blank to keep output in browser)

**File type returned**:
- plain text
- gzip compressed

**Buttons**:
- get output
- summary/statistics

To reset all user cart settings (including custom tracks), click here

http://bejerano.stanford.edu
Coming soon: submit from the UCSC Table Browser

Functional Prototype from Stanford Mirror Site

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

clade: Mammal  genome: Human  assembly: Mar. 2006 (NCBI38/hg18)

manage custom tracks
table: uc16  describe table schema

region:  genome  ENCODE Pilot regions  position chrX:151073054-151303576

identifiers (names/accessions):  paste list  upload list

filter: create
intersection: create
correlation: create

output format: all fields from selected table  Send output to Galaxy  GREAT

output file:  (leave blank to keep output in browser)

file type returned: plain text  gzip compressed

get output  summary/statistics

To reset all user cart settings (including custom tracks), click here

http://bejerano.stanford.edu
Coming soon: submit from the UCSC Table Browser

Functional Prototype from Stanford Mirror Site

We encourage integration from other tools via our well-documented interface.
Future work:
Cytoscape visualization of enrichments

Explore enriched terms against the structure of an ontology.

http://bejerano.stanford.edu
Summary

- Current technologies identify *cis*-regulatory sequences
- GREAT accurately assesses functional enrichments of *cis*-regulatory sequences using a genomic region-based approach
- Applying GREAT to eight transcription factors and transcription associated factors (including SRF & p300) reveals enrichments for known and novel functions [McLean et al., *Nat Biotechnol.*, 2010]
- Online tool available at [http://great.stanford.edu](http://great.stanford.edu)
Acknowledgments

GREAT developers

Cory McLean
Dave Bistor
Michael Hiller
Shoa Clarke
Craig Lowe (UCSC)
Aaron Wenger
Gill Bejerano

Contributors
Fah Sathira
Marina Sirota
Bruce Schaar
Terry Capellini
Christopher Meyer
Jennifer Hardee
Miles Davis
Sebastian Gutierrez

Others
Craig Mak & NBT reviewers
Ontology developers & annotators communities

http://great.stanford.edu

http://bejerano.stanford.edu