Personal Transcription Factor Binding Site Mutations Point to Personal Medical Histories

Gill Bejerano

http://bejerano.stanford.edu

Genome = Genes + Gene Regulation

<table>
<thead>
<tr>
<th>Type</th>
<th># in genome</th>
<th>% of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>genes</td>
<td>20,000</td>
<td>2%</td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cis elements</td>
<td>1,000,000</td>
<td>&gt;10%</td>
</tr>
</tbody>
</table>

Atomic event – transcription factor binding

Disease Associated tag SNPs

- Over 15,000 distinct tag SNPs in the GWAS Catalog
- 80-90% far away from (linkage with) gene exons
- Are most gene cis regulatory?
- Are they near genes with common functionality?

GWAS Catalog Growth

Cis-reg enrichments: GREAT.stanford.edu

½ million job submissions, 700+ references, established defaults

<table>
<thead>
<tr>
<th>Gene transcription start site</th>
<th>Gene regulatory domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function (‘abnormal cardiac output’)</td>
<td>Cis-reg rich region set</td>
</tr>
</tbody>
</table>

GREAT = Genomic Regions Enrichment of Annotations Tool

\[ P = P_{\text{GREAT}}(k \geq 5 \mid n=6, p_0=0.33) \]

[McLean et al, Nature Biotech, 2010]

Cis-reg enrichment: GREAT.stanford.edu

½ million job submissions, 700+ references, established defaults

Advantages of GREAT:
1. Accounts for both proximal and distal binding sites
2. Variable length gene regulatory domains
3. Multiple hits next to same gene add significance
4. Extensive body of knowledge (16,000 functions)

\[ P = P_{\text{GREAT}}(k \geq 5 \mid n=6, p_0=0.33) \]

[McLean et al, Nature Biotech, 2010]

Unlinked GWAS SNPs \(\rightarrow\) GREAT
**Unlinked GWAS SNPs → GREAT**

http://bejerano.stanford.edu

**Individual Genomes → GREAT?**

http://bejerano.stanford.edu

**TF Motif Library (PBM+ChIP+SELEX)**

http://bejerano.stanford.edu

**Predict conserved binding sites**

http://bejerano.stanford.edu

**Compare everything to shuffled motifs & weed!**


**PRISM vs. ChIP-seq → GREAT**

http://bejerano.stanford.edu

---

3/15/2016
**PRISM vs. ChIP-seq → GREAT**

**Term**
- actin cytoskeleton
- structural constituent of muscle
- dilated heart ventricles
- regulation of insulin secretion

**SRF**
- PRISM: Known
- ChIP-seq: Novel

**Actin cytoskeleton**
- PRISM: Known
- ChIP-seq: Known

**Every known function is supported by dozens or hundreds of novel binding sites.**

Note: Sensitivity vs Specificity

---

**GWAS SNPs: Predict upstream regulator**


---

**Personal Deleterious Binding Sites**

**COBELs** = Conserved Binding Site Eroding Loci

Individuals w public medical records
Randomize COBELs

- Replace every CoBEL with a random binding site prediction for the same transcription factor of same affinity and similar cross-species conservation.

- Using 10,000 random control sets, the likelihood of obtaining the functions reported in Table 1 as top prediction due to bias in the distribution of binding sites in the genome is low (Quake P = 3 x 10^-4, Church P = 5.7 x 10^-3, Angrist P = 4.8 x 10^-3, Gill P = 1 x 10^-4, Lupski P = 1.9 x 10^-3, and combined P = 1.6 x 10^-15).

- Significance remains high when we relax the requirement to recover each exact same term with matching any one of a broader group of 12–60 related functions as a top prediction (Quake P = 1.1 x 10^-3, Church P = 1.3 x 10^-2, Angrist P = 7.7 x 10^-3, Gill P = 7.4 x 10^-3, Lupski P = 6.5 x 10^-3, and combined P = 5.2 x 10^-12).
Randomize Medical Histories

- Define an association matrix linking enrichment and medical history, with the phenotypes observed in the five individuals as rows, and top enriched terms in all as columns. A cell in the matrix would be marked “true” only where the enriched term (of any individual) is thought to be related to the etiology of the phenotype (of any individual).
- One instance of this matrix was filled by a medical doctor based on their medical knowledge and training and another instance was independently filled using a literature survey. The objective was to compute the chance of associating a set of five individuals with random medical histories with the observed enrichments using one of the two association matrices as the “gold” association.
- We generated 1,000 sets of five individuals with random medical histories composed of similar disease profiles and assessed the likelihood of being able to associate them with enrichments. Successfully linking five random individuals with enrichments was highly significant using the association matrix generated by the medical doctor ($P = 3.0 \times 10^{-3}$) and by the matrix generated by literature survey ($P = 3.0 \times 10^{-2}$) suggesting our links between enrichment and medical histories are not just a function of the listed histories.

COBELs ≠ GWAS SNPs or HGMD

Most Predictive COBELs are Private

Contributions from Common & Rare

Summary

- We define likely deleterious events as personal variants that erode the affinity of human conserved binding sites.
- When the set of all such events is probed for lying next to gene sets of particular function or phenotype, we repeatedly get a solid match between top genomic prediction and self reported medical summary.
- Top genomic predictions are eroded at both gene and gene set level.
- The variants we highlight appear to be part of the mutational load predisposing individual lineages to different diseases.

Other Lab Interests: 1) Solve Patient
2) Automate Patient Solving

3) Discover Mammalian Adaptations

Kudos

COBELS: (PLoS Comp Bio, 2016)
Harendra Guturu, Sandeep Chinchali, Shoa Clarke

PRISM: (Genome Research, 2013)
Aaron Wenger, Shoa Clarke, Harendra Guturu, Jenny Chen, Bruce Schaar, Cory McLean

GREAT: (Nature Biotechnology, 2010)
Cory McLean, Dave Bristor, Michael Hiller, Shoa Clarke, Bruce Schaar, Craig Lowe, Aaron Wenger

Bejerano Lab past & present The Organizers