Biological Networks: Next Generation Annotation for the Human Genome

Mark Gerstein
Yale

Slides at
Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
Visiting Canada
2 Days, 2 Scheduled Events, 1.5 Talks:
Same Theme but Different Modules

CCBR
• Traditional Annotation: Predictive Models of Gene Expression from His. Mods & TF binding
• Next-Gen: Architecture of the Human Regulatory Net

OICR (2 pm, 18 May, TMDT 4-204)
• Traditional Annotation: Pseudogenes in the Human Genome
• Next-Gen: Architecture of the Human Regulatory Net
• A Different View of Variation & Networks from Model Organisms
THE SEQUENCE EXPLOSION

- Automated Sanger Sequencing: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

- Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

- 454 Pyrosequencing: Released in 2006, this technique produces billions of base pairs in a single run.

- A Thousand Genomes: Released in 2010, this project aims to sequence genomes from 1,092 individuals from 26 populations.


- The cost per million base pairs of sequence (log scale) has dramatically decreased over time.

- International HapMap Project: Sequencing the genomes of individuals from various populations to understand genetic variations.

- Whole Genome Shotgun Sequence: Gene sequence stored in international public databases.

- Third-Generation Sequencing: Companies such as Helicos BioSciences already read sequences from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

- Sequencing by Ligation: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

- 23andMe and deCODEme: Commercial genetic testing companies.

- Science: "The human genome" article.

- Nature: "A thousand genomes" article.

- James Watson, a woman with acute myeloid leukemia, a Yoruba male from Nigeria, and the first Asian genome.
Functional Interpretation of Personal Genomes

Interpreting each variant in molecular terms
Annotation giving global view of an individual’s genome

2000

nature

2012
Human Genome Analysis: Classic Approach v Future Direction

• A classic annotation story
  - Identifying Pseudogenes in intergenic DNA and potentially ascribing regulatory function to them
  - ENCODE & 1000G approach to current Human analysis
  - “Junk to Part”

• A networks view on the genome
  - Understanding the regulatory network as a hierarchy with information flow bottlenecks
  - Seeing the impact of variation and constraint on the network
  - “Parts” to Systems
Types of Annotation: Comparative & Functional

- Large-scale sequence similarity comparison
  - Identify large blocks of repeated and deleted sequence:
    - Within the human reference genome
    - Within the human population
    - Between closely related mammalian genomes
  - Identify smaller-scale repeated blocks using statistical models

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks
Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes ($\Psi G$)
  - Inheritable
  - Homologous to a functioning element
  - Non-functional
    - No selection pressure so free to accumulate mutations
      - Frameshifts & stops
      - Small Indels
      - Inserted repeats (LINE/Alu)
  - What does this mean? no transcription, no translation?…

[Mighell et al. FEBS Letts, 2000]
Identifiable Features of a Pseudogene (ψRPL21)

Genome-wide Annotation of Pseudogenes

PseudoPipe

Protein Sequence  Reference Genome

Six-frame blast

Eliminate redundant hits
Remove hits overlapping exon

Merge hits and identify parents

FASTA re-alignment

Processed Pseudogenes  Duplicated Pseudogenes

Pseudogene Information Pool

- 18,046 PseudoPipe
- 13,644 RetroFinder
- 11,240 HAVANA

2-way consensus
9,093

$\Delta$2-way consensus
1,907

Level 1
7,186

Level 2
4,054

Surveyed Set
11,216

Polymorphic Pseudogenes
24

1000G

ENCODE

psiDR

[Pei et al., GenomeBiology ('12, submitted)]
Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimp</th>
<th>Mouse</th>
<th>Rat</th>
<th>Chicken</th>
<th>Zebrafish</th>
<th>Pufferfish</th>
<th>Fruitfly</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>1/0</td>
<td>1/2</td>
<td>0/1</td>
<td>-</td>
<td>0/2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPI</td>
<td>-</td>
<td>-</td>
<td>1/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PFK</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ALDO</td>
<td>1/1</td>
<td>1/1</td>
<td>11/0</td>
<td>7/0</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TPI</td>
<td>3/0</td>
<td>2/1</td>
<td>6/1</td>
<td>3/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>GAPDH</strong></td>
<td><strong>60/2</strong></td>
<td><strong>47/3</strong></td>
<td><strong>285/46</strong></td>
<td><strong>329/35</strong></td>
<td><strong>0/1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGK</td>
<td>1/1</td>
<td>1/2</td>
<td>2/0</td>
<td>12/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGM</td>
<td>12/0</td>
<td>13/1</td>
<td>9/0</td>
<td>3/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENO</td>
<td>1/0</td>
<td>1/2</td>
<td>12/1</td>
<td>36/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PK</td>
<td>2/0</td>
<td>3/0</td>
<td>10/3</td>
<td>4/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LDH</td>
<td>10/2</td>
<td>9/1</td>
<td>27/7</td>
<td>25/4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>97</strong></td>
<td><strong>91</strong></td>
<td><strong>422</strong></td>
<td><strong>463</strong></td>
<td><strong>4</strong></td>
<td><strong>1</strong></td>
<td><strong>0</strong></td>
<td><strong>0</strong></td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>
Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.
Aproximate Age of GAPDH pseudogenes

Age calculated based on Kimura-2 parameter model of nucleotide substitution

[Bai et al. BMC Genomics (09)]
• Unprocessed pseudogenes with no functional counterparts in the same genome

76 Unitary pseudogenes

[Diagram showing evolutionary relationships and gene trees]
### 11 Polymorphic pseudogenes

<table>
<thead>
<tr>
<th>CDS-disrupted gene</th>
<th>GPR33</th>
<th>SERPINB11</th>
<th>TAAR9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disruptive mutation</strong></td>
<td>Cga (R) → Tga</td>
<td>Gaa (E) → Taa</td>
<td>Aaa (K) → Taa</td>
</tr>
</tbody>
</table>

**Allele frequency**

- **GPR33**: 0.285 ($P = 0.867$)
- **SERPINB11**: 8.659 ($P = 0.013$)
- **TAAR9**: 0.071 ($P = 0.965$)

---

[Zhang et al. (’10) GenomeBiology]
LOF variants: creation of “pseudogenes” from annotated genes

[Balasubramanian et al., Genes Dev., '11]
## Number of LOF Variants in an Individual

[MacArthur et al., *Science* ('12, in press)]

<table>
<thead>
<tr>
<th>variant type</th>
<th>total</th>
<th>1000G low-coverage average per individual</th>
<th>NA12878 (high coverage European)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CEU</td>
<td>CHB+JPT</td>
</tr>
<tr>
<td>stop</td>
<td>565</td>
<td>26.2</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.2)</td>
<td>(6.9)</td>
</tr>
<tr>
<td>splice</td>
<td>267</td>
<td>11.2</td>
<td>13.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.9)</td>
<td>(2.5)</td>
</tr>
<tr>
<td>frameshift</td>
<td>337</td>
<td>38.2</td>
<td>36.2</td>
</tr>
<tr>
<td>indel</td>
<td></td>
<td>(9.2)</td>
<td>(9.0)</td>
</tr>
<tr>
<td>large deletion</td>
<td>116</td>
<td>28.3</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.2)</td>
<td>(5.9)</td>
</tr>
<tr>
<td>total</td>
<td>1285</td>
<td>103.9</td>
<td>103.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(22.5)</td>
<td>(24.3)</td>
</tr>
</tbody>
</table>

### Proportion Distribution

- **High-confidence LoF**
- **Candidate LoF**
- **Missense**
- **Synonymous**

![Proportion Distribution Chart](chart.png)
Impact of Polymorphic Pseudogenes & LoF Events on Gene Sets

- What should be the reference gene set?
  - Single individual
  - Ancestral individual
  - Current reference
  - Union of genes found in any individual
  - Intersection of genes found in everyone

[Balasubramanian et al., *Genes Dev.*, '11]
IncRNA: Identification of many candidate ncRNAs through evidence integration

- No single feature (e.g. expr. expts., conservation, or sec. struc.) finds all known ncRNAs => combine features in stat. model
- 90% PPV, 13 of 15 tested validate
Pseudogene Transcription

Total transcribed pseudogene: 863

RTPCR Experimental validated: 57 of 76

[Pei et al., GenomeBiology ('12, submitted)]
Chromatin Around Active Pseudogenes

[Diagram showing signal profiles for DNase I hypersensitivity and H3K4me3 across genomic position, with peaks at TSS (Transcription Start Site) for transcribed and non-transcribed pseudogenes.]

[Pei et al., Genome Biology (2012, submitted)]
Partial Pseudogene Activity

Transcribed with Additional Activity

Partially Active
Signatures of Selection on Some Pseudogenes

- Signature for Ka/Ks for all pseudogenes or just transcribed
- Signature for SD pgenes:
  - Intersect with UW SD DB [She et al. Nature '04 431:927]
  - Some candidates under selection (28 [16\%] from sample of 167)
- Signature for transcribed pgenes using 1000G polymorphisms

Comparison of nucleotide substitutions per site (N) in ψ genes and SDs

- N (ψgene) < N (SD) → negative selection
- N (ψgene) ≈ N (SD) → neutral evolution
- N (ψgene) > N (SD) → positive selection

167

[Mu et al. NAR 39: 7058]
Summarizing Partial Pseudogene Activity

[Pei et al., Genome Biology ('12, submitted)]
Examples & speculation on the function of pseudogene ncRNAs:

Regulating their parents

- via acting as endo-siRNAs
  [Recent ex. in fly & mouse, ‘08 refs.]
- via acting as miRNA decoys
  [PTEN]
- via inhibiting degradation of parent’s mRNA
  [makorin]

Poliseno et al. Nature 465:1033 (’10)
Part 2 : Networks
Current Annotation: 1D Browser Tracks
Current Annotation: 1D Browser Tracks

Will this scale to 1000+ tracks? What will next-gen annotation look like?...
Networks occupy a midway point in terms of level of understanding.

1D: Complete Partslist ("Elements" in genomic tracks)

~2D: Network Wiring Diagram of a Molecular System

3D & 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

[UCSC genome browser]


[Chiu et al. Trends in Cell Biol, 16:144]
More coming in ENCODE Production Project

Consortium
Comprises ~50 Labs

Transcriptome Chromatin
TFs
Data Flow: Chip-seq expts. to co-associating peaks

119 TFs from 458 ChIP-Seq experiments

- Mostly in Tier 1 cell lines
  - K562, GM12878, H1h-ESC...
- Matching RNA-Seq data in all cell-lines
- SPP & PeakSeq
- thresholding w. IDR (replicas)

Signal Tracks

7M Peaks from Uniform Peak Calling

TF1

TF2

TF119

Data Flow: peaks to proximal & distal networks

Peak Calling

Assigning TF binding sites to targets

Filtering high confidence edges & distal regulation
Based on stat. model combining signal strength & location relative to typical binding

Distal Regulation
Proximal Regulation

[Cheng et al., Bioinfo. ('11); Gerstein et al. Nature (in press, '12); Yip et al., GenomeBiology (in press, '12)]
Identifying potential enhancers

1. TF peaks from ChIP-seq
   Human genome in 100bp bins
   - TF A
   - TF B
   - ...

2. Use peaks as examples to learn chromatin features of binding active regions
   - Chromatin features
   - Machine learning
   - Prediction
   - Thresholding
   - BAR scores
   - BARs
   - Filtering
   - DRMs
   - Predicted genes
   - Gencode genes

3. Filter close to genes to get enhancer list
   - Enhancer list

4. Find correlated enhancer-target pairs

Finding potential target genes

- HM signals
  - H3K4me1
  - H3K27ac
- Expression levels
  - Gene 1
  - Gene 2
  - Gene 3
  - Scale

~20K distal edges tot. from ~130K enhancer-like elements

(Related to but more “targeted” than enhancer “states” from unsupervised segmentation,
M Hoffman et al. & J Ernst et al.)

Identifying Potential Enhancer-like Elements from Discriminative Model & then Linking these to Targets (via cell-line correlations) to Create Distal Edges

[ Yip et al. Genome Biology (in press, '12) ]
Identifying potential enhancers

1. TF peaks from ChIP-seq

   *Human genome in 100bp bins*

   - TF A
   - TF B
   - ...

2. Use peaks as examples to learn chromatin features of binding active regions

   - Pos. examples
   - Neg. examples

   *Chromatin features → Machine learning → Prediction → BAR scores → Thresholding → FILTERING → DRMs*

3. Filter close to genes to get enhancer list

   - Gencode genes
   - Predicted genes

4. Use peaks as examples to learn chromatin features of binding active regions

5. Find TFs binding enhancer-like elt. in cell lines with strong HMs

6. Draw distal edges from TFs to targets

   *Cell lines → Cell lines *

   *Gene 1 → Gene 2 → Gene 3 *

   *Expression levels → H3K4me3 → H3K27ac → Scale*

   *Strong → Weak*

   *GM12878 → H1-hESC → HeLa-s3 → Hep-G2 → K562 → ...

   ~20K distal edges tot. from ~130K enhancer-like elements

   (Related to but more “targeted” than enhancer “states” from unsupervised segmentation, M Hoffman et al. & J Ernst et al.)

**Identifying Potential Enhancer-like Elements from Discriminative Model & then Linking these to Targets (via cell-line correlations) to Create Distal Edges**

*Yip et al. Genome Biology (in press, ’12)*
Network Stats to Identify Bottlenecks, Hubs & Hierarchies

Hierarchy Height Statistic (normalized Out - In)

[Yu et al., PLOS CB (2007)]
Strongest Proximal Regulatory Edges Can be Arranged into a Hierarchy

Hierarchy height distribution approximated by 3 levels

Optimally arrange TFs into 3 levels by sim. annealing, maximizing downward-pointing edges

Middle level has highest betweenness, creating info. flow bottlenecks

Comparing Proximal & Distal Networks

Integration of TF hierarchy with other ‘omic information:
more influential & connected TFs on the top

Integration of TF hierarchy with other ‘omic information:
more influential & connected TFs on the top
Integration of TF hierarchy with other ‘omic information: more influential & connected TFs on the top


Avg. correlation betw. binding signal of TF & gene expr. of its target

.62

.63

.56
Non-influential TFs are collectively influential (based on TF model, in K562)

\[ X(i) = \text{Correlation between binding of TF } i \text{ and gene expression of its targets} \]

Corr. between \( X(i) \) & degree centrality of TF \( i \) = .42

Cooperation of Mid-level Regulators

Co-binding & co-regulation

From looking at expr of shared v unshared targets

Co-association  Cooperativity  Physical Interaction

Enrichment or Depletion (Observed/Expected)

Network Motif Analysis: Enrichment of FFLs

3-node motifs

<table>
<thead>
<tr>
<th>Frequency N</th>
<th>3-node motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>868 (0.84)</td>
<td>![3-node motif 1]</td>
</tr>
<tr>
<td>490 (0.81)</td>
<td>![3-node motif 2]</td>
</tr>
<tr>
<td>729 (0.72)</td>
<td>![3-node motif 3]</td>
</tr>
<tr>
<td>26 (0.62)</td>
<td>![3-node motif 4]</td>
</tr>
<tr>
<td>0 (0)</td>
<td>![3-node motif 5]</td>
</tr>
<tr>
<td>0 (0)</td>
<td>![3-node motif 6]</td>
</tr>
<tr>
<td>8 (0.70)</td>
<td>![3-node motif 7]</td>
</tr>
<tr>
<td>64 (0.91)</td>
<td>![3-node motif 8]</td>
</tr>
<tr>
<td>6 (1.8)</td>
<td>![3-node motif 9]</td>
</tr>
<tr>
<td>6 (2.9)</td>
<td>![3-node motif 10]</td>
</tr>
<tr>
<td>2 (5.6)</td>
<td>![3-node motif 11]</td>
</tr>
<tr>
<td>16 (4.8)</td>
<td>![3-node motif 12]</td>
</tr>
<tr>
<td>122 (1.4)</td>
<td>![3-node motif 13]</td>
</tr>
</tbody>
</table>

6 pairs of toggle switches

Allele-Specific Behavior in the Regulatory Network

- In GM12878, determine ASB for ~50 TFs & ASE using RNA-Seq
  - AlleleSeq pipeline using personal diploid genome & annotation constructed for NA12878 (to prevent reference bias)
  - ~20% of expressed genes show ASE
  - ~10% of binding sites show ASB
- GM12878 Allele-Specific Regulatory Network
  - Just proximal edges with ASB
  - Just target nodes with ASE

RNA/ChIP-Seq Reads

[ Rozowsky et al. MSB ('11) ]
Allele-Specific Regulatory Network: combinatorial coordination of ASE & ASB

[ Rozowsky et al. MSB ('11) ]
More "allelic" components under weaker selection

(from 1000G pilot & phase I)
More generally, more connected components ("hubs") have less variation

TF target in-degree

Neg. corr. with (SCC=-.2, P<0.5)

dN/dS
(from chimp alignments)

More "allelic" components under weaker selection

TF target in-degree &
TF out-degree

Neg. corr. with

ns SNP density, pN/pS, avg. DAF

(from 1000G pilot & phase I)
TFs at the Top Under Stronger Negative Selection

SNP dens. $x10^{-3}$

1.0
3.1
3.8

A Different View on Selection
More Connectivity, More Constraint: A theme borne out in many studies

Proteins that have a more central position evolve more slowly and are more likely to be essential. This phenomenon is observed in many organisms and different kinds of networks: Fraser et al. (’02) Science, (’03) BMC Evo. Bio. [yeast PPI]; Butland et al. (’04) Nature [E coli. PPI]; Hahn et al. (’05), MBE [worm, fly]; Cheng et al. (’09), BMC Genomics [miRNA nets].

Rapid Evolution in the interaction network takes place at the periphery

- Pos. sel. v. deg. centrality:
  \[ \rho = -0.06, p < 1.2e-6 \]
  - Independent gene expression bias
  - Update w. 1000G Phase I
  SNP dens. v. centrality: \[ \rho = -0.1, P < 2.2e-16 \]

- Sequence variation v. centrality
  - Nonsyn / synonymous SNPs v. deg. centrality
    \[ \rho = -0.1, P < 4.0e-4 \]
    \[ \rho = -0.3, P < 2.2.0e-16 \] (updated to 1000G phase I)
  - # overlapping SDs of each gene
    v. betweenness centrality : \[ \rho = -0.04, P < 3.3e-3 \]
  - # CNVs intersecting a gene
    v. network centrality: \[ \rho = -0.03, P < 0.002 \]

E. Coli Transcriptional regulatory network vs Linux call graph

<table>
<thead>
<tr>
<th>Basic properties of systems</th>
<th>E. coli transcriptional regulatory network</th>
<th>Linux call graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>Genes (TFs &amp; targets)</td>
<td>Functions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(subroutines)</td>
</tr>
<tr>
<td>Edges</td>
<td>Transcriptional regulation</td>
<td>Function calls</td>
</tr>
<tr>
<td>External constraints</td>
<td>Natural environment</td>
<td>Hardware architecture, customer requirements</td>
</tr>
<tr>
<td>Origin of evolutionary changes</td>
<td>Random mutation &amp; natural selection</td>
<td>Designers’ fine tuning</td>
</tr>
</tbody>
</table>

A subnetwork in the Linux call graph from CodeViz

[Yan et al., PNAS ’10]
The Linux Kernel Evolves!

We can track the evolution of a function as the "rate of evolution of a function" — the number of times it got revised.

From Wikipedia

<table>
<thead>
<tr>
<th></th>
<th>E. coli transcriptional regulatory network</th>
<th>Linux call graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>1378</td>
<td>12391</td>
</tr>
<tr>
<td>Number of persistent nodes</td>
<td>72* (5%)</td>
<td>5120 (41%)</td>
</tr>
<tr>
<td>Number of edges</td>
<td>2967</td>
<td>33553</td>
</tr>
<tr>
<td>Number of modules</td>
<td>64</td>
<td>3665</td>
</tr>
<tr>
<td>Number of comparative references</td>
<td>200 bacterial genomes</td>
<td>24 versions of kernels</td>
</tr>
<tr>
<td>Years of evolution</td>
<td>Billions years</td>
<td>20 years</td>
</tr>
</tbody>
</table>
E. coli transcriptional regulatory network

master regulator

middle manager

workhorse

Linux call graph

[Yan et al., PNAS (2010), in press]
Comparison: hierarchical organization

Pyramidal vs Top-heavy

Degree distribution
Roles of hubs

out-degree hubs e.g. “crp”

in-degree hubs e.g. “printk”, “spin_lock”

% in E. coli regulatory network % in Linux call graph

<table>
<thead>
<tr>
<th>Role</th>
<th>E. coli</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>master regulator</td>
<td>4.6</td>
<td>29.6</td>
</tr>
<tr>
<td>middle manager</td>
<td>5.1</td>
<td>58.2</td>
</tr>
<tr>
<td>workhorse</td>
<td>90.2</td>
<td>12.3</td>
</tr>
</tbody>
</table>

[Yan et al., PNAS (2010), in press]
Comparison: organization of modules

Modules are labeled by master regulators: TFs, high-level starting functions

TRN: modules overlap little, components are less generic: “ompF”

Call graph: modules overlap, Functions are highly reused (generic): “printk”

<table>
<thead>
<tr>
<th></th>
<th>E. Coli TRN</th>
<th>Linux call graph</th>
</tr>
</thead>
<tbody>
<tr>
<td># of modules</td>
<td>64</td>
<td>3665</td>
</tr>
<tr>
<td>Average overlap</td>
<td>4.3%</td>
<td>80.7%</td>
</tr>
<tr>
<td>Maximum node reuse</td>
<td>15.6%</td>
<td>87.5%</td>
</tr>
<tr>
<td>Average node reuse</td>
<td>3.5%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

[Yan et al., PNAS (2010), in press]
Distribution of Evolutionary Rates of components in E. coli vs Linux

E. coli

Persistent genes evolve slowly

- e.g. dnaA

Linux

Two classes of persistent functions

- purifying selection
- adaptive evolution

Higher reuse

- e.g. mem_read
- release_dev
- swap_free

- e.g. strlen
- srandom32
- counter_set

[Yan et al., PNAS '10]
Centrality is Correlated with Variation in Linux

Reusable components need to be constantly fine-tuned

Spearman correlation

$r=0.26$, $P<10^{-75}$
Perspectives on Random Change v Intelligent Design

• Central points = hubs & bottlenecks
• If changes random, best not to put them in central pts.
• If changes made rationally, can put them into central pts.
  – Moreover, good to do this, as these more often used
  – i.e more efficient
  – Why there’s so much GWB construction
Summary of Traditional Annotation Effort on Human Pseudogenes

- Converting what was considered “junk” into potentially functional elements on the genomic “parts” list

- Large scale assignment of pgenes (~11K hi qual. of ~14K tot.)
  - Polymorphic pgenes & ~100 LOF events per person call into question ref. gene set

- Overlap with functional annotation
  - Many pgenes are transcribed ncRNAs (~900)
  - Many are under selection but not as proteins
  - Partial activity be signature of a dying gene or of a regulatory ncRNA
Conclusions on the structure of the human regulatory network

- Reformulating peak lists into simplified network hierarchies, with "downward" proximal regulation
- Distal regulation has a very different connectivity pattern
- TFs on the top are more "influential" and well-connected (eg w/ miRNAs)
- Info.-flow bottlenecks in middle, ameliorated by co-operation of mid-level regulators
  - Cooperation evident in network motifs (e.g. prevalence of noise-buffering FFLs)
- Network connectivity shows coordinated binding to Mat. or Pat. alleles
- More connected components are under greater constraint
Summary of Analysis of Model Organism Regulatory Networks

- Linux call graph provides a contrast: Top-heavy hierarchy emphasizes reuse of components vs robustness
  - More changes to highly connected components only possible given “designed” changes
Acknowledgements

ENCODEx, 1000G

Pseudogene.org

GENCODE (Not Linked): B Pei, C Sisu
(11 Main Projects, ~50 labs, >350 substantial contributors, Analysis Lead E Birney)

NHGRI, particularly the “encode team”


+ (for enhancers) E Birney, P Bickel, & B Brown