Biological Networks: Next Generation Annotation for the Human Genome

Mark Gerstein
Yale

Slides at Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
GersteinLab.org Overview & Relationship to Today’s Talk

- Personal Genomics & Human Variation
- Human Genome Annotation
- Networks of Genes & Proteins
- Macromolecular Structures & Motions
- Biological Knowledge Representation & Literature Mining
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: The Trace archive, started in 2000, houses raw sequence data and currently holds 1.6 trillion base pairs.

- Cost per million base pairs of sequence (log scale):
  - $10,000
  - $250
  - $200
  - $150
  - $100
  - $50
  - $10

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

- Whole Genome Shotgun Sequence:

- Gene sequence stored in international public databases:
  - $1,000
  - $10

- THE HUMAN GENOME PROJECT
- Navigenics
- 23andMe
- deCODEme
- International HapMap Project


- THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence 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.
Personal genomes will become a commonplace part of cancer research & treatment. They will brings biological science to general public.
Functional Interpretation of Personal Genomes

Interpreting each variant in molecular terms. Towards a personal annotation, giving a functional view of an individual’s genome.
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 human regulatory network as a hierarchy with information flow bottlenecks
  – Seeing the impact of variation and constraint on the network
  – “Parts” to Systems
ENCODE Production Project

Consortium
Comprises
~50 Labs

Subprojects:
Transcriptome
+ Chromatin
+ TFs
**Types of Annotation:** Comparative & Functional

- Comparative:
  - Within the human reference genome
  - Within the human population
  - Between closely related mammalian genomes

- Functional:

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 (ΨG)
  – 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 ($\psi$RPL21)

Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes

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

~14K total

Δ2-way
1,907

Level 1
7,186

Level 2
4,054

Surveyed Set
11,216

Polymorphic Pseudogenes
24

1000G

ENCODE

Pseudogene Decoration Resource psiDR

[Pei et al., GenomeBiology ('12, submitted)]
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

Processed/Duplicated

<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>0/1</td>
<td>-</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>
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

Processed/Duplicated

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>GPI</td>
<td>-</td>
<td>-</td>
<td>1/0</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>
</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>
</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>
</tr>
<tr>
<td><strong>GAPDH</strong></td>
<td><strong>62</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>
</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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>Total</td>
<td>97</td>
<td>91</td>
<td>422</td>
<td>463</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

GAPDH
Distribution of 62 human GAPDH pseudogenes

[Liu et al. BMC Genomics (’09, in press)]
Dating of GAPDH pseudogenes, highlights their creation ~50 Mya

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

Burst of Retrotranspositional Activity
• Unprocessed pseudogenes with no functional counterparts in the same genome

76 Unitary pseudogenes

[Diagram showing evolutionary relationships and gene names such as ART2B, AYTL1B, FETA, PRAME, SLC7A15, TAAR4, and more, along with timelines indicating millions of years ago (MYA).]
11 Polymorphic pseudogenes: Ex of SerpinB11

G_{aa} (E) => T_{aa} (stop)

Freq. of this Mutation in Various Human Populations

[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>total</td>
<td>CEU</td>
<td>CHB+JPT</td>
</tr>
<tr>
<td>stop</td>
<td>565</td>
<td>26.2 (5.2)</td>
<td>27.4 (6.9)</td>
</tr>
<tr>
<td>splice</td>
<td>267</td>
<td>11.2 (1.9)</td>
<td>13.2 (2.5)</td>
</tr>
<tr>
<td>frameshift</td>
<td>337</td>
<td>38.2 (9.2)</td>
<td>36.2 (9.0)</td>
</tr>
<tr>
<td>indel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>large deletion</td>
<td>116</td>
<td>28.3 (6.2)</td>
<td>26.7 (5.9)</td>
</tr>
<tr>
<td>total</td>
<td>1285</td>
<td>103.9 (22.5)</td>
<td>103.5 (24.3)</td>
</tr>
</tbody>
</table>

![Proportion of variant types](image)
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

Simple & Complex
Ex of Pseudogene Transcription

Parent: ENSG00000176444.13

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

[Pei et al., Genome Biology ('12, submitted)]
Partial Pseudogene Activity

Transcribed with Additional Activity

Partially Active

Transcription

Layered H3K4Me1
Layered H3K4Me3
Layered H3K27Ac
DNase Clusters
Txn Factor ChIP
Duke Uniq 35

Transcription

Layered H3K4Me1
Layered H3K4Me3
Layered H3K27Ac
DNase Clusters
Txn Factor ChIP
Duke Uniq 35
- 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
    (16% from sample of 167)
- Signature for transcribed pgenes using 1000G polymorphisms

![Diagram of segmental duplication and nucleotide substitutions](image)

- Comparison of nucleotide substitutions per site (N) in ψgenes and SDs
  - $N(\psi\text{gene}) < N(\text{SD})$
    - negative selection
  - $N(\psi\text{gene}) \approx N(\text{SD})$
    - neutral evolution
  - $N(\psi\text{gene}) > N(\text{SD})$
    - positive selection

[Khurana et al, NAR ('10)]

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

[Pei et al., GenomeBiology ('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

[UCSC genome browser]


[Chiu et al. Trends in Cell Biol, 16:144]

Connecting systems biology to genomics
Network pathology & pharmacology

Interactome networks

Breast Cancer
Alzheimer’s Disease
Parkinson’s Disease
Multiple Sclerosis

[Adapted from H Yu]
Data Flow: Chip-seq expts. to co-associating peaks

119 TFs from 458 ChIP-Seq experiments

Signal Tracks

7M Peaks from Uniform Peak Calling

TF1

TF2

TF119

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

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

~500K Edges

~26K Edges

[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
   - Pos. examples
   - Neg. examples
   - Chromatin features
   - Machine learning
   - Prediction
   - BAR scores
   - Thresholding
   - BARs
   - Filtering
   - DRMs
   - Gencode genes
   - Predicted genes

3. Filter close to genes to get enhancer list

Finding potential target genes

- HM signals
- Expression levels
- Cell lines
- Gene 1
- Gene 2
- Gene 3
- Gene 4
- Gene 5
- ...

4. Find correlated enhancer-target pairs

~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

1. TF peaks from ChIP-seq
   Human genome in 100bp bins
   - TF A
   - TF B
   - ... (DNase I, FAIRE, H3K4me3)

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
   - DRM

3. Filter close to genes to get enhancer list
   - Genes:
   - Gencode genes
   - Predicted genes

4. TF A, TF B, ...

5. Find TFs binding enhancer-like elt. in cell lines with strong HMs
   - GM12878
   - H1-hESC
   - HeLa-S3
   - Hep-G2
   - K562
   - H3K4me1
   - H3K27ac

6. Draw distal edges from TFs to targets
   - Expression levels
   - Scale: Strong, Weak

~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 enhancers

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

- 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
  - DRM

- Filter close to genes to get enhancer list
  - Genes:
    - Gencode genes
    - Predicted genes

- TF A, TF B, ...

- Find TFs binding enhancer-like elt. in cell lines with strong HMs
  - GM12878
  - H1-hESC
  - HeLa-S3
  - Hep-G2
  - K562
  - H3K4me1
  - H3K27ac

- Draw distal edges from TFs to targets
  - Expression levels
  - Scale: Strong, Weak

~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.)
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

Sig. corr. w/ TF
hubbiness
(.24 & .62)

# regulating miRNAs & # regulated miRNAs

Avg. values

16
630
24
593
10
321

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

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

Cooperation of Mid-level Regulators

Co-binding & co-regulation

From looking at expr of shared v unshared targets

Co-association  Cooperativity  Physical Interaction
Network Motif Analysis: Enrichment of FFLs

<table>
<thead>
<tr>
<th>Frequency N</th>
<th>3-node motifs</th>
<th>6 pairs of toggle switches</th>
</tr>
</thead>
<tbody>
<tr>
<td>868 (0.84)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>490 (0.81)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>729 (0.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 (0.62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 (0.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 (0.91)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 (1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 (2.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 (5.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 (4.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>122 (1.4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

[ Rozowsky et al. MSB ('11) ]
Maternal & Paternal Personal Regulatory Networks: 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.

More "allelic" components under weaker selection.

TF target in-degree

Neg. corr. with

(SCC=-.2, P<0.5)

dN/dS

(from chimp alignments)

TF target in-degree & TF out-degree

Neg. corr. with

ns SNP density, pN/pS, avg. DAF

TFs at the Top Under Stronger Negative Selection

SNP dens. $\times 10^{-3}$

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 kind 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:**
  \[ \rho = -0.06, \ p < 1.2 \times 10^{-6} \]
  - Independent gene expression bias
  - Update w. 1000G Phase I
    SNP dens. v. centrality: \[ \rho = -0.1, \ P < 2.2 \times 10^{-16} \]

- **Sequence variation v. centrality**
  - Nonsyn / synonymous SNPs v. deg. centrality
    \[ \rho = -0.1, \ P < 4.0 \times 10^{-4} \]
    \[ \rho = -0.3, \ P < 2.2 \times 10^{-16} \] (updated to 1000G phase I)
  - # overlapping SDs of each gene v. betweenness centrality: \[ \rho = -0.04, \ P < 3.3 \times 10^{-3} \]
  - # 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 (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

- `kswpdp`
- `shrink_slab`
- `cond_resched`
- `put_swap_token`
- `shrink_zone`
- `zone_watermark_int`
- `refrigerator`
- `throttle_vma_writeout`
- `shrink_inactive_list`
- `shrink_active_list`
- `recalc_smppending`
- `spin_lock_irq`
- `congestion_wait`
- `get_dirty_limits`
- `finish_wait`
- `prepare_to_wait`
- `io_schedule_timeout`
- `boll_waitout_fraction`
- `chp_poll_dirty_limit`
- `tsk_dirty_limit`
- `determine_dirtyable_memory`
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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Linux 0.1</th>
<th>Linux 1.0</th>
<th>Linux 1.1</th>
<th>Linux 1.2</th>
<th>Linux 1.3</th>
<th>Linux 1.4</th>
<th>Linux 2.0</th>
<th>Linux 2.4</th>
<th>Linux 2.6</th>
<th>Linux 3.x</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

From Wikipedia
E. coli transcriptional regulatory network

Linux call graph

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

<table>
<thead>
<tr>
<th>Role</th>
<th>% in E. coli regulatory network</th>
<th>% in Linux call graph</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>

Pyramidal vs Top-heavy

Degree distribution
Roles of hubs

out-degree hubs e.g. “crp”

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

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

Persistent genes evolve slowly

E. coli

Linux

Two classes of persistent functions

- E. coli: dnaA
- Linux: strlen, random32, counter_set, mem_read, release_dev, swap_free

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

Reusable components need to be constantly fine-tuned

Spearman correlation

$r=0.26, P<10^{-75}$

[Yan et al., *PNAS* ‘10]
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 (e.g. with 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 (Linux call graph provides an illustrative contrast)
Acknowledgements

ENCODE, 1000G

Pseudogene.org

GENCODE (Not Linked): B Pei, C Sisu
Acknowledgements

(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