Human Genome Annotation
(Stories Contrasting Traditional Annotation with Newer Network Approaches: Regulatory Networks v Track Models)

Slides at
Lectures.GersteinLab.org
(See Last Slide for References & More Info.)

Mark Gerstein
Yale
THE SEQUENCE EXPLOSION

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

Cost per million base pairs of sequence (log scale)

$10,000

$250

$200

$150

$100

$50

$1,000

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Billions of base pairs

Audience Sequencing:
This technique employed in SOLiD and
Polonator instruments uses a different
chemistry from previous technologies
and samples every base twice, reducing
the error rate.
Personal genomes will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the 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
Consortium Comprises ~50 Labs

Subprojects:
- Transcriptome
- Chromatin
- TFs
How might we annotate a human text?

**The Semicolon Wars**

Brian Hayes

If you want to be a thoroughgoing world traveler, you need to learn 6,912 ways to say “Where is the toilet, please?” That’s the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot, you also have quite a challenge ahead of you, learning all the ways to say:

```c
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we’ve been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

Every programmer knows there is one true programming language. A new one every week

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect; they’ll concede, but it’s built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will concede which end of a boiled egg to crack.

This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that’s not what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift’s Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, “On holy wars and a plea for peace.” The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, `in x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C
Sources of Annotation: Comparative & 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

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
Human Genome Analysis:
Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr : Using His. Mods. to predict gene expression
  - TFs=>Expr : Comparing this with TF binding
  - HMs=>TFs : Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
    • Practical Application: Finding disease genes
Modeling Transcription: Connecting Inputs & Outputs

- Models connecting various types of high-level data
  - HMs+TFs => gene expression
  - HMs => TFs
Histone Modification (HM) model

Chromatin features: Histone modifications

Predictors

RNA-Seq data

Prediction target: Gene expression level

[Cheng et al. (11), Genome Biol. 12: R15]
His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion.

H3K4me2

Transcript Level

Autosomal

top 20%
bottom 20%

[Science 330:6012] [Related work: Ouyang et al. ('09) PNAS; Karlic et al. ('10) PNAS]
Integrate all histone modifications to predict gene expression levels

Classify H/L genes (SVM)

Predict expression values

Magnitude of Prediction from a “bin” around the TSS
HM models are tissue specific

HM model-- Best prediction is achieved by using histone modification and expression data from the same developmental stage

[Cheng et al. ('11) Genome Biol. 12: R15]
Scale up to Human

Pearson's $r = 0.9$ (p-value < $2.2 \times 10^{-16}$)
RMSE = 1.9
Classification: AUC = 0.95
Regression: $r = 0.77$ (RMSE = 2.3)
Application of chromatin model in 5 species: Consistent Performance

>50% of variation of expression levels can be explained by HMs
Human Genome Analysis: Classic Approach v Future Direction

- Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr: Using His. Mods. to predict gene expression
  - TFs=>Expr: Comparing this with TF binding
  - HMs=>TFs: Using His. Mods to predict TF binding

- A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    - Particularly with network analogies
    - Practical Application: Finding disease genes
Doing a Model with TFs:
Positive and negative regulators from correlating TF signal at TSS with gene expression

[Cheng et al. ('11) PLOS CB]
Predictor v2: 2-levels, now with TFs

[Cheng et al. NAR ('11)]
Mouse ESC Models Illuminates Different Regions of Influence for TFs vs HMs

- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.’08; Mikkelsen et al. ’07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!

A Model with only a Few of the Thousands of Mouse TFs is able to Predict Well

[Cheng et al. NAR ('11, in press)]
Scale up to Human: TFs

Pearson’s r=0.81; RMSE=2.57
Classification: AUC = 0.89
Regression: r = 0.62; RMSE = 3.06

Relative importance of TFs

Different types of TFs have different correlation with gene expression

Prediction of Differential Expression

TF model– differential TF binding signals are predictive of differential expression levels between two human cell lines

TFs active in a particular cell type are the strongest predictors in that cell

Human Genome Analysis: Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr : Using His. Mods. to predict gene expression
  - TFs=>Expr : Comparing this with TF binding
  - HMs=>TFs : Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    - Particularly with network analogies
    - Practical Application: Finding disease genes
Chromatin model: link histone modification patterns to TF binding

Predictors (HM)

|  | HM1 | HM2 | HM3 | ...
|---|-----|-----|-----|---
| Bin1 | 0.2 | 0.4 | 0.6 | ...
| Bin2 | 0.3 | 0.3 | 0.2 | ...
| Bin3 | 0   | 0.2 | 2.1 | ...
| Bin4 | 0.4 | 0.4 | 0   | ...
| Bin5 | 0.3 | 1.2 | 0.5 | ...
| Bin6 | 1.2 | 3.1 | 2.1 | ...
| Bin7 | 3.4 | 2.4 | 0.8 | ...
| Bin8 | 1.5 | 1.2 | 0.9 | ...

......

TF binding site?

|  | TF1 | TF2 | ...
|---|-----|-----|---
| Bin1 | 0   | 0   | ...
| Bin2 | 0   | 0   | ...
| Bin3 | 1   | 0   | ...
| Bin4 | 1   | 0   | ...
| Bin5 | 0   | 1   | ...
| Bin6 | 0   | 0   | ...
| Bin7 | 0   | 1   | ...
| Bin8 | 1   | 0   | ...

......

Machine Learning Method (SVM et al.)

ROC curve

AUC

FPR (1-specificity)

TPR (sensitivity)
Data for Predicting TF targets from HMs

- **Worm** (same as for gene expr.)
  - Chip-seq: 22 TFs + Pol2 in a variety of stages
  - Chromatin Chip-chip: >12 HMs mostly in EE & L3
- **Yeast**
  - TF binding data:
    - ChIP-chip for 203 yeast TFs
  - Chromatin modification profile data
    - ChIP-chip, Pokholok et al. 2005
    - 14 profiles: H3(YPD), H4(YPD), H3(H2O2), H3K9AC/H3(YPD), H3K14AC/H3(YPD), H3K14AC/WCE(YPD), H3K14AC/H3(H2O2), H4AC/H3(YPD), H4AC/H3(H2O2), H3K4ME1/H3(YPD), H3K4ME2/H3(YPD), H3K4ME3/H3(YPD), H3K36ME3/H3(YPD), H3K79ME3/H3(YPD)
  - Positional weighted matrices for yeast TFs

Relating the Chromatin Model to TF Binding Sites in worm

• Model Predicts Sites Fairly Accuracy
• Accuracy improved when coupled with PWM

[Gerstein et al., Science ‘10]
HIS+PWM leads to higher accuracy for predicting SOME yeast TF target genes – the His-sensitive ones

[Cheng et al. Genome Biol. ('11) 12:R111]
Histone-sensitive vs. -insensitive TFs

[Cheng et al. Genome Biol. (‘11) 12:R111]
Human Genome Analysis: Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr : Using His. Mods. to predict gene expression
  - TFs=>Expr : Comparing this with TF binding
  - HMs=>TFs : Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
    • Practical Application: Finding disease genes
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]
Network pathology & pharmacology

Breast Cancer

Alzheimer’s Disease

Parkinson’s Disease

Multiple Sclerosis

Interactome networks

[Adapted from H Yu]
Networks as a common language in many scientific contexts

- Internet
- Food Web
- Electronic Circuit
- Neural Network
- Disease Spread
- Protein Interactions
- Social Network
Exploiting Network Analogies to Gain Intuition

Guilt by association

Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]
**Data Flow: Chip-seq expts. to co-associating peaks**

**119 TFs** from 458 ChIP-Seq experiments (2 Tb tot.)

- 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)


2785 GATA1 (focus-factor) peak locations
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

Potential Distal Edge

Strong Proximal Edge

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

Finding potential target genes

1. TF peaks from ChIP-seq from
   - GM12878
   - H1-hESC
   - HeLa-S3
   - Hep-G2
   - K562
   - ...
2. Use peaks as examples to learn chromatin features of binding active regions
   - HM signals
   - Expression levels
   - Gene 1
   - Gene 2
   - Gene 3
   - Scale
   - Strong
   - Weak
3. Find correlated enhancer-target pairs
4. Find correlated enhancer-target pairs

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
   - BARs
   - Filtering

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

4....

5. Find TFs binding enhancer-like elt. in cell lines with strong HMs
   - GM12878
   - H1-hESC
   - HeLa-S3
   - Hep-G2
   - K562
   - ...
   - Gene 1
   - Gene 2
   - Gene 3
   - Scale
     - Strong
     - Weak

6. Draw distal edges from TFs to targets
   - Expression levels

~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. GenomeBiology (in press, ’12)]
Network Stats to Identify Bottlenecks & Hubs

[Yu et al., PLOS CB (2007)]
Hierarchy Height Statistic = (normalized TF Out deg. – In deg.)

Network Stats to Identify Hierarchy
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

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

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

Cooperation of Mid-level Regulators

From looking at expr of shared v unshared targets

## Network Motif Analysis: Enrichment of FFLs

### 3-node motifs

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

### Diagram

- 6 pairs of toggle switches

### Bar Chart

- Number of FFLs

### References

Human Genome Analysis: Classic Approach v Future Direction

- Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr: Using His. Mods. to predict gene expression
  - TFs=>Expr: Comparing this with TF binding
  - HMs=>TFs: Using His. Mods to predict TF binding

- A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    - Particularly with network analogies
    - Practical Application: Finding disease genes
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 "Difference" Network
  - Just proximal edges with ASB
  - Just target nodes with ASE

RNA/ChIP-Seq Reads

ACTTTGATAGCGTCAATG
CTTTGATAGCGTCAATGC
CTTTGATAGCGTCAACGC
TTGACAGCGTCAATGCACT
TGATAGCGTCAATGCACT
ATAGCGTCAATGCACT
TAGCGTCAATGCACT
CGTCAACGCACGTGGGA
GTCAATGCACTGAGAG
CAATGCACTGAGAG
AATGCACTGAGAG
TGCACGTTGGGAGTTGGC

[ Rozowsky et al. MSB ('11) ]
Many Technical Issues in Determining ASE/ASB: Reference Bias
(naïve alignment against reference)

[Diagram showing frequency distribution of ASE SNPs]

Reference Allele

Alt Counts / (Alt Counts + Ref Counts)

Alternate Allele

[Rozowsky et al., MSB ('11)]
Construction of a Personal Diploid Genome & Transcriptome

Reference Genome

vtc2diploid

Genotyping, Phasing, Filtering

Personal Variants

SVs  Indels  SNPs

Paternal Haplotype

Personal Genome

Maternal Haplotype

[Reference: TGGAAGAAGACCGTTT...]
[Deletion: TGGAAGAAGACCGTTT...]
[SNP: TGGAAGAAGACCGTTT...]
[Insertion: TGGAAGAAGACCGAGTTT...]
[Personal Haplotype: TGGAAGCCGAGTTT...]

[Rozowsky et al., MSB (in press,'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

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

TFs at the Top Under Stronger Negative Selection

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

1.0

3.1

3.8

TopNet – an automated web tool

Encodenets.gersteinlab.org
(tYNA: Normal website + Downloaded code (JAVA) + Web service (SOAP) with Cytoscape plugin)

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]
Human Genome Analysis: Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr: Using His. Mods. to predict gene expression
  - TFs=>Expr: Comparing this with TF binding
  - HMs=>TFs: Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
    • Practical Application: Finding disease genes
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 PPI]; Cheng et al. (’09), BMC Genomics [miRNA nets] [Nielsen et al. PLoS Biol. (2005), HPRD, Kim et al. PNAS (2007)]

- **Sequence variation v. centrality**
  - Nonsyn / synonymous SNPs v. deg. centrality
    \[ \rho = -.1, P < 4.0e-4 \]
    \[ \rho = -.3, P < 2.2.0e-16 \]
  (updated to 1000G phase I)
# E. Coli Transcriptional regulatory network vs Linux call graph

## Basic properties of systems

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

---

![Diagram of E. Coli transcriptional regulatory network vs Linux call graph](image)

- **A subnetwork in the Linux call graph from CodeViz**

[Yan et al., PNAS ’10]
E. coli transcriptional regulatory network

master regulator

middle manager

workhorse

Linux call graph

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

<table>
<thead>
<tr>
<th></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>

out-degree hubs e.g. “crp”

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

[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>Linux 0.1</th>
<th>Linux 1.0</th>
<th>Linux 1.1</th>
<th>Linux 2.0</th>
<th>Linux 2.2</th>
<th>Linux 2.4</th>
<th>Linux 2.6</th>
<th>Linux 3.x</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.1.0</td>
<td>2.2.0</td>
<td>2.2.1</td>
<td>2.2.4</td>
<td>2.3.0</td>
<td>2.4.0</td>
<td>2.4.2</td>
</tr>
</tbody>
</table>

Line of code: 6 million
Line of code: 4 million

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

Updated 26/01/2012.
Distribution of Evolutionary Rates of components in E. coli vs Linux

Persistent genes evolve slowly
- e.g. dnaA

Two classes of persistent functions
- e.g. strlen
- srandom32
- counter_set
- swap_free

Higher reuse
- e.g. mem_read
- release_dev
- swap_free

[E. coli] [Linux]

[Yan et al., PNAS '10]
Centrality is correlated with variation in Linux, opposite of the trend of constraint with centrality in biological networks.

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
Human Genome Analysis: Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  - HMs=>Expr: Using His. Mods. to predict gene expression
  - TFs=>Expr: Comparing this with TF binding
  - HMs=>TFs: Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
    • Practical Application: Finding disease genes
Practical application of network properties: predicting disease genes

GENE CATEGORIES

LoF-tol | Neutral | GWAS (common disease-assoc. variants) | HGMD (rare disease-causing variants) | Essential

Decreasing tolerance to mutation
Use properties of genes at the ends of spectrum

LoF-tol

- Homozygous inactivation in at least one healthy 1000 Genomes individual
- Weak selection constraints

Essential

- Homozygous inactivation leads to clinical features of death before puberty or infertility
- Very strong selection constraints

From Liao et al, PNAS, 2008
Gene essentiality and regulatory network

Functionally essential genes tend to be more central in regulatory network than LoF-tolerant genes

Results similar to previous studies in protein-protein interaction network (Kim et al, PNAS, 2007)

[Khurana et al., PLOS Comp. Bio. ’13 (in press)]
Genes participate in many networks and no single network captures the global picture of gene interactions.

Combine regulatory interactions with other networks: physical protein-protein, signaling, metabolic, phosphorylation and genetic to create a unified network (Multinet).

Nodes: ~15,000 genes
Edges: ~110,000 interactions

[Khurana et al., PLOS Comp. Bio. ’13 (in press)]
Genes participate in many networks and no single network captures the global picture of gene interactions.

[Khurana et al., PLOS Comp. Bio. '13 (in press)]
Gene properties in Multinet

Essential genes are Connected to more genes

Involved in more networks

LoF-tolerant genes   Essential genes

Edges shown in gray
Size of nodes scaled by total degree

[Khurana et al., PLOS Comp. Bio. ’13 (in press)]
Other properties of LoF-tolerant and Essential genes

Essential genes are

- Functionally essential genes
  - High evolutionary conservation
  - Reduced genetic variability in modern humans
  - Participate in many networks
  - High connectivity in most networks
  - Top level in TF-TF hierarchical network

Highly conserved across species

Have more interfaces to interact with other proteins

Can we combine all these features to predict tolerance of each gene to deleterious mutation?

[Khurana et al., PLOS Comp. Bio. ’13 (in press)]
Integration of network properties to predict systems-level effects of deleterious mutations

Train logistic regression model using network and evolutionary properties

Can distinguish between LoF-tolerant and Essential genes with high accuracy

Application of the model on all genes

Predicted score

AUC=0.91

True positive rate

False positive rate

[Khurana et al., PLOS Comp. Bio. '13 (in press)]
Human Genome Analysis: Classic Approach v Future Direction

• Classic view: Large-scale Interrelation of Linear Elements through Statistical Models
  – HMs=>Expr: Using His. Mods. to predict gene expression
  – TFs=>Expr: Comparing this with TF binding
  – HMs=>TFs: Using His. Mods to predict TF binding

• A Networks View on Large-scale Organization of Genomic Elements
  – Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  – Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
    • Practical Application: Finding disease genes
Models of Transcription Relating Various Genome Tracks of Information – His. Mods, TF Binding & Gene Expression

- Predictive models of gene expression
  - Applicable in many contexts
  - Work for miRNAs as well as genes
  - Show variable importance of regions around genes for chromatin & TFs
  - Show TF & HM signals are redundant for ‘predicting’ gene expression
  - Surprisingly, a few TFs, particularly TFSSs, are quite predictive

- Predictive models of TF Binding
  - Open chromatin model + PWM does better than features individually
  - TFs in yeast divide into 2 groups: His-sens & His-insens
  - His-sens TFs easier to predict & higher expression + more PPIs
Networks as Next-Gen Annotation: 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)
• Maternal & Paternal personal regulatory networks manifest combinatorial coordination
• More connected components are under greater constraint
  • Understanding this analogies + Practical Applications
TF-v-expr:

worm-HM:
**Cheng C**, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C

ENCODEx:
**Chao Cheng**, Roger Alexander, Renqiang Min, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder

Models Acknowledgements
Acknowledgements

(11 Main Projects, ~50 labs, >700 substantial contributors + NHGRI)

Networks/Elements (~60 participants):


+ (for enhancers) E Birney, P Bickel, & B Brown + (for variation) J Chen
Default Theme

- Default Outline Level 1
  - Level 2
Info about content in this slide pack

• PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2013. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
• Paper references in the talk were mostly from Papers.GersteinLab.org.
• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt