Human Genome Analysis:

from Constructing Informative Regulatory Networks,
to Relating them to Variation,
to Practical Cancer Tools

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).
See last slide for references & more info.
Big data is transforming science

- Genomics - DNA sequencer
- Neuroscience - The Human Connectome Project
- High energy physics - Large Hadron Collider
- Ecology - Fluxnet
- Astronomy - Sloan Digital Sky survey
- Knowledge of knowledge - Meta-data of scientific documents
- Computational social science - Online communities
Sequencing Data Explosion in last 5 years

From ‘00 to ~’20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. ('11) Genome Biology]
What do people do with Big Data?

- Fundamental goal is general understanding & answering specific Qs: modeling & making predictions

- Explicit Description of Data not Important -- Fast query, hiding underlying structure (e.g. Google Search)

- Explicit Description of Data Important – Organization highlighting underlying structure (e.g. Google Maps)

[ Nature 489: 208 ]

Higgs Boson: Searching Through Many Events for a Few Needles

"Golden" Events
One $H \rightarrow 4 \ell / \text{Billion}$
Making Intuitive Maps, Highlighting Large-scale Structure of Stars & the Earth


[SDSS.org]
Human Genome Annotation – a non-intuitive map

- Large-scale organization, providing an overview of the genome
- Integration of heterogeneous data

Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
Sources of Annotation: Comparative & Functional

Identification of large blocks of repeated and deleted sequence:
- Within the human reference genome
- Within the human population
- Between closely related mammalian genomes

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
Networks occupy a midway point in terms of level of understanding

~2D: Network Wiring Diagram of a Molecular System

3D & 4D: Detailed structural understanding of cellular machinery

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

[UCSC genome browser]


[Chiu et al. Trends in Cell Biol, 16:144]
Networks as a common language in many scientific contexts

- Internet
- Food Web
- Electronic Circuit
- Neural Network
- Disease Spread
- Protein Interactions
- Social Network
Combining networks forms an ideal way of integrating diverse information.

- Metabolic pathway
- Transcriptional regulatory network
- Physical protein-protein Interaction
- Co-expression Relationship
- Genetic interaction (synthetic lethal)
- Signaling pathways

Part of the TCA cycle
Network pathology & pharmacology

Breast Cancer

Alzheimer’s Disease

Parkinson’s Disease

Multiple Sclerosis

Interactome networks

[Adapted from H Yu]
Exploiting Network Analogies to Gain Intuition

Guilt by association

Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
Data Flow: Chip-seq expts. to co-associating peaks

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

- 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

1. **Peak Calling**
2. **Assigning TF binding sites to targets**
3. **Filtering high confidence edges & distal regulation**

- Potential Distal Edge
- Strong Proximal Edge

Based on a statistical model combining signal strength and location relative to typical binding.

- ~500K Edges
- ~26K Edges

Cheng et al., Bioinfo. ('11); Gerstein et al., Nature (in press, '12); Yip et al., GenomeBiology (in press, '12)
Target Identification with a Probabilistic model (TIP)

$S_i(g)$: the signal of a TF at nucleotide $i$ of gene $g$

--- TF binding signal at different position contribute differently

(1) Distance from TSS

(2) binding preference of TFs

 Avg binding signal across all genes

Calculate regulatory score and $p$-value for all genes

Cheng et al. 2011, Bioinformatics
‘Supervised’ enhancer prediction

- Identifying Potential Enhancer-like Elements from Discriminative Model

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

Human genome in 100bp bins

Positive examples

Negative examples

Features

Machine learning

Prediction

Filtering

Get enhancer list away to genes

Strong H3K4me1 & H3K27ac signal

~130K enhancer-like elements

Yip et al., Genome Biology (2012)
Associate enhancers with target genes

- Idea: Histone modifications to predict gene expression.

- Form distal regulatory networks (~20k distal edges in ENCODE rollout; we extend to edges with ~17k genes)

- (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

[Yu et al., PLOS CB (2007)]
Hierarchy Height Statistic = 
(normalized TF Out deg. – In deg.)
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

- **Why Networks?**
  - A compelling approach to organize big data

- **A networks view on organizing regulatory annotation**
  - Motivating & constructing the human **regulatory network**
  - Understanding it as a **hierarchy** with information flow bottlenecks
  - Greater mid-level **collaboration** in more complex organisms

- **Analyzing the impact of variation on the network**
  - Node Variation: **more connectivity = more constraint**
  - Useful analogies to **designed systems**
  - **Edge rewiring**

- **Practical App: Network prioritization of cancer mutations**
  - Prioritizing based on **multiple network connectivity**
  - Finding **ultra-sensitive non-coding regions** & disruptive mutations (eg motif breakers)
  - Building a **practical workflow** & tool for disease genomes
Optimally arrange TFs into 3 levels by simulated annealing, maximizing downward-pointing edges.

Hierarchy height distribution approximated by 3 levels.
Probing direction framed as an optimization problem.
Strongest Proximal Regulatory Edges Can be Arranged into a Hierarchy

Global wiring pattern of TFs

Middle level has highest betweenness, creating info. flow bottlenecks
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

16

630

24

593

10

321

Cooperation of Mid-level Regulators

From looking at expr of shared v unshared targets

Co-binding & co-regulation

Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
Yeast Regulatory Hierarchy:
Middle-managers Rule

[Yu et al., PNAS ('06)]
Transcriptional Regulatory (Tr) Networks available in some form for:

- **Ec** - *E. coli* (most complete)
- **Sc** - *S. cerevisiae* (most complete)
- **Hs** - *H. Sapiens* (early pre-encode draft)
- **Mm** - *M. Musculus* (draft)
- **Rr** - *R. rattus* (draft)
- **Ce** - *C. eegans* (modencode w/ miRNAs)
- **Dm** - *D. melanogaster* (modencode, w/ miRNAs)

Other Networks (e.g. phosphorylation [Ph] & modification [Mo]) available in some form for *S. cerevisiae* & *H. sapiens*

[Bhardwaj et al., PNAS '10; Negre et al. Nature '11]
Different kinds of Hierarchies

- Well-defined levels and a clear chain of command
- A military hierarchy

- Without well-defined levels & with more co-regulatory partnerships
- A club or a scientific collaboration network

- High degree of co-regulation and can be organized into hierarchies
- A law firm

[Bhardwaj et al., PNAS (2010), in press]
Collaborative Nature of the Nodes

Degree of Collaboration

1/6 = 0.16

4/4 = 1

More Collaborative: Democratic
More Autonomous: Autocratic

[Bhardwaj et al., PNAS (2010), in press]
Most collaboration involves middle level

[Bhardwaj et al., PNAS (2010), in press]
Higher species have more collaborative nodes

[Bhardwaj et al., PNAS '10]
Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers

Governmental hierarchy of a representative city (Macao)
Middle Managers Interact the Most in Efficient Corporate Settings


[Bhardwaj et al., PNAS (2010), in press]
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  - Motivating & constructing the human regulatory network
  - Understanding it as a hierarchy with information flow bottlenecks
  - Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  - Node Variation: more connectivity = more constraint
  - Useful analogies to designed systems
  - Edge rewiring

• Practical App: Network prioritization of cancer mutations
  - Prioritizing based on multiple network connectivity
  - Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  - Building a practical workflow & tool for disease genomes
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)]
More connected components ("hubs") have less variation

Integrate TFs & their binding sites with 1000G variation data & primate alignments (GERP score).

This shows:

<table>
<thead>
<tr>
<th>TF target in-degree &amp; TF out-degree</th>
<th>Neg. corr. with ns SNP density, pN/pS, avg. DAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg. corr. with dN/dS (from chimp alignments)</td>
<td></td>
</tr>
</tbody>
</table>
TFs at the Top Under Stronger Negative Selection

SNP density

$\times 10^{-3}$

1.0

3.1

3.8

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

[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

% in *E. coli* regulatory network | % in Linux call graph
---|---
master regulator | 4.6 | 29.6
middle manager | 5.1 | 58.2
workhorse | 90.2 | 12.3

[Yan et al., *PNAS* ('10)]

out-degree hubs e.g. “crp”

in-degree hubs e.g. “printk”, “spin_lock”
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 1.2</th>
<th>Linux 1.3</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>[Graph showing line of code progression]</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>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>
<tr>
<td></td>
<td>20 years</td>
</tr>
</tbody>
</table>
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

e.g. strlen
srandom32

counter_set

Higher reuse

e.g. mem_read
release_dev
swap_free

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

$\text{Spearman correlation} \quad r=0.26, \quad P<10^{-75}$

[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: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
Interologs & Regulogs

Target Organism (e.g. Fly)

Source Organism (e.g. Yeast)

Target Organism (e.g. Yeast)

Source Organism (e.g. Worm)

Orthologs

Protein-DNA Interologs

Same Sequence

Orthologs

Regulogs

Orthologs with identity $I_A$

Orthologs with identity $I_B$

$J_{AB} = \sqrt{I_A \times I_B}$

Interacting Proteins

Protein-Protein

Interologs with joint identity

[Yu et al. Gen. Res. (‘04)]
High likelihood ratios from interolog mapping (for very similar sequences)

LR cutoff $\sim 1000$

More similar Pairs of Sequences

[Yu et al, Genome Res 14: 1107-18]

[Yu et al. Gen. Res. ('04)]
Current network data

- Systematic study of biological network rewiring
  - Use Interologs & Regulogs
  - Quantify rewiring in a unified way for all types of networks
  - Compare network rewiring to sequence change

<table>
<thead>
<tr>
<th>Species</th>
<th>scer</th>
<th>calb</th>
<th>spom</th>
<th>other fungi</th>
<th>worm</th>
<th>fly</th>
<th>mouse</th>
<th>human</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Genetic</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>TF</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>miRNA</td>
<td>Not Exist</td>
<td></td>
<td></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Phospho</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Metabolic</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Rewiring rate analysis

[Shou et al. PLOS CB (‘11)]
Define rewiring rate

• Given two networks with certain time divergence

\[
\text{Rewiring rate} = \frac{R}{C \times \text{Time divergence}}
\]

\( R \): total number of rewired edges

\( C \): total number of possible rewiring edges (number of edges in a complete full network)

• Pros
  - Normalized by the size of two comparing networks
  - Considering overlap of nodes
  - Unified measure applicable to all types of biological networks

[Shou et al. *PLOS CB* (‘11)]
Number of total possible changes

A
\[
\frac{\text{CNs} \times (\text{CNs} - 1) + \text{GNs} \times (\text{GNs} - 1) + \text{LNs} \times (\text{LNs} - 1)}{2} + \text{CNs} \times (\text{GNs} + \text{LNs})
\]

B
\[
\frac{\text{Reg CNs} \times (\text{Reg CNs} - 1) + \text{Reg GNs} \times (\text{Reg GNs} - 1) + \text{Reg LNs} \times (\text{Reg LNs} - 1)}{2} + \\
\text{Reg CNs} \times \text{Tar CNs} + \text{Reg GNs} \times \text{Tar GNs} + \text{Reg LNs} \times \text{Tar LNs} + \\
\text{Reg CNs} \times (\text{Tar GNs} + \text{Tar LNs}) + \text{Tar CNs} \times (\text{Reg GNs} + \text{Reg LNs})
\]

C
\[
\text{Reg CNs} \times \text{Tar CNs} + \text{Reg GNs} \times \text{Tar GNs} + \text{Reg LNs} \times \text{Tar LNs} + \\
\text{Reg CNs} \times (\text{Tar GNs} + \text{Tar LNs}) + \text{Tar CNs} \times (\text{Reg GNs} + \text{Reg LNs})
\]

[Shou et al. PLOS CB ('11)]
Rewiring rate bands

[Shou et al. PLOS CB ('11)]
Comparing the 3 groups of molecular network rewiring to Social network rewiring

Fast
TF reg.
&
Phosphorylation

Moderate
PPI,
Gen. Int.,
&
miRNA reg.

Slow
Metabolic Pathways

The Ronald Reagan family poses in the Red Room for a family portrait on his inaugural, January 20, 1981.
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human **regulatory network**
  – Understanding it as a **hierarchy** with information flow bottlenecks
  – Greater mid-level **collaboration** in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: **more connectivity = more constraint**
  – Useful analogies to **designed systems**
  – **Edge rewiring**

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on **multiple network connectivity**
  – Finding **ultra-sensitive non-coding regions** & disruptive mutations (eg motif breakers)
  – Building a **practical workflow** & tool for disease genomes
Where is Waldo?
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)
Gene categories with known phenotypic effects

Decreasing tolerance to mutation

- LoF-tol
- Neutral
- GWAS (common disease-associating variants)
- HGMD (rare disease-causing variants)
- Essential

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

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

From Liao et al, PNAS, 2008
Recasting Association of Connectivity with Constraint in terms of Gene Essentiality Categories: Application to Regulatory & Protein Interaction Networks

Regulatory Network

Essential genes

Higher Centrality

More interaction interfaces

Khurana et al., PLoS Comp. Bio., 2013
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

Edges shown in gray

[Khurana et al., PLOS Comp. Bio. ’13]
Gene properties in Multinet

Essential genes are connected to more genes. Involved in more networks.

Unified network degree (log scale)

Num of networks

LoF-tolerant genes
Essential genes

Size of nodes scaled by total degree

[Khurana et al., PLOS Comp. Bio. '13]
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

AUC=0.91

Predict higher scores for disease-causing genes, GWAS genes & cancer drivers

[Khurana et al., *PLOS Comp. Bio.* ’13]
• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  - Motivating & constructing the human regulatory network
  - Understanding it as a hierarchy with information flow bottlenecks
  - Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  - Node Variation: more connectivity = more constraint
  - Useful analogies to designed systems
  - Edge rewiring

• Practical App: Network prioritization of cancer mutations
  - Prioritizing based on multiple network connectivity
  - Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  - Building a practical workflow & tool for disease genomes
Annotating the non-coding genome: Using FunSeq

‘Conservation’
- Typically defined by comparison across species

*Khurana et al. 2013 Science*

How do we define this within human population?
- where 99.9% of the genome is conserved between people
- a depletion of common variants/an enrichment of rare variants (more rare variants)
‘Conservation’
- Typically defined by comparison across species

How do we define ‘sensitivity’ within human population?
- a depletion of common variants/
an enrichment of rare variants
‘Conservation’
- Typically defined by comparison across species

How do we define ‘sensitivity’ within human population?
- a depletion of common variants/
an **enrichment** of rare variants
‘Conservation’
- Typically defined by comparison across species

How do we define ‘sensitivity’ within human population?
- a depletion of common variants/
  an enrichment of rare variants
‘Conservation’
- Typically defined by comparison across species

How do we define ‘sensitivity’ within human population?
- a depletion of common variants/
an enrichment of rare variants
Depletion of Common SNPs as a metric for Negative Selection

- Metrics for selection
  - Evolutionary conservation (GERP)
  - SNP density (confounded by mutation rate)
- Depletion of common polymorphisms for regions under selection
  - Alternatively, negative selection restricts the allele frequency of deleterious mutations.

LOF-tol (Loss-of-function tolerant): least negative selection
Cancer: most selection

[Khurana et al., Science ('13)]
Can we identify which non-coding elements are under very strong “coding-like” selection?

- Start 677 high-resolution non-coding categories; Rank & find those under strongest selection
- Binding peaks of some general TFs (eg FAM48A)
- Core motifs of some TF families (eg JUN, GATA)
- DHS sites in spinal cord and connective tissue

Enrichment of known disease-causing mutations from Human Gene Mutation database

[Khurana et al., Science ('13)]
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data

• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms

• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring

• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
Noncoding cancer variants from whole-genome sequencing

- 64 prostate cancer (Berger et al., Nature, 2011; Baca et al., Cell, 2013)
  ~1500 to 18,000 per sample
- 21 breast cancer (Nik-Zainal et al., Cell, 2012)
  ~2000 to 80,000 per sample
- 3 medulloblastoma (Rausch et al., Cell 148, 2012).
  ~1600 to 2000 per sample

- ~99% of somatic SNVs occur in non-coding regions, including TFBSs, ncRNAs and pseudogenes
  - Cancer sequencing has been very exome focused
  - Publicity for TERT promoter mutation – exception proves the rule!
Identification of non-coding candidate drivers amongst somatic variants: Scheme

[Khurana et al., Science ('13)]
Flowchart for 1 Prostate Cancer Genome
(from Berger et al. '11)

[Khurana et al., Science ('13)]
Identification of non-coding candidate drivers amongst somatic variants: Examples

Validation of a candidate driver identified in prostate cancer sample in \textit{WDR74} gene promoter

- Sanger sequencing in 19 additional samples confirms the recurrence

- \textit{WDR74} shows increased expression in tumor samples

[Khurana et al., Science ('13)]
Human Genome Analysis: from Constructing Informative Regulatory Networks, to Relating them to Variation, to Practical Cancer Tools

• Why Networks?
  A compelling approach to organize big data
• A networks view on organizing regulatory annotation
  – Motivating & constructing the human regulatory network
  – Understanding it as a hierarchy with information flow bottlenecks
  – Greater mid-level collaboration in more complex organisms
• Analyzing the impact of variation on the network
  – Node Variation: more connectivity = more constraint
  – Useful analogies to designed systems
  – Edge rewiring
• Practical App: Network prioritization of cancer mutations
  – Prioritizing based on multiple network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
  – Building a practical workflow & tool for disease genomes
TopNet – an automated web tool

Topology of Networks

(Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al)

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)
FunSeq.GersteinLab.org: webserver & code download

This site can be used to automatically score and annotate disease-causing potential of SNVs, particularly the non-coding ones. It can be used on cancer and personal genomes. It also contains a downloadable tool (found under 'Downloads').

Function based Prioritization of Sequence Variants

Under 'Analysis', an online version of the tool is available, where a personal or cancer genome variant file (VCF or BED) can be uploaded and analysed. Additionally, the tool can also detect recurrent annotation elements in non-coding regions when running with multiple genomes.

Copyright © 2013, GersteinLab@Yale
Acknowledgements

Networks.gersteinlab.org
Pseudogene.org

ENCODE
modENCODE
1000G

Not Linked:
B Pei
C Sisu
Y Fu

MG

E Khurana

H Yu K Yip

N Bhardwaj

K Yan R Alexander

C Cheng

G Fang C Shou
RNAi: Birth of a Field in the Literature Culminating in the 2006 Nobel

PubNet.GersteinLab.org
Acknowledgements

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

Networks/Elements (~60 participants):


Hiring Postdocs. See gersteinlab.org/jobs!
Cancer Prioritization

Acknowledgements

Yale
Ekta Khurana, Yao Fu, Jieming Chen, Xinmeng Mu, Lucas Lochovsky, Arif Harmanci, Alexej Abyzov, Suganthi Balasubramanian, Cristina Sisu, Declan Clarke, Mike Wilson

Sanger
Vincenza Colonna, Yali Xue, Chris Tyler-Smith

US, UK, Switzerland….
Hyun Min Kang, Tuuli Lappalainen, Kathryn Beal, Daniel Challis, Yuan Chen, Laura Clarke, Fiona Cunningham, Emmanouil T. Dermitzakis, Uday Evani, Paul Flicek, Erik Garrison, Javier Herrero, Yong Kong, Kasper Lage, Daniel G. MacArthur, Gabor Marth, Donna Muzny, Tune H. Pers, Graham R. S. Ritchie, Jeffrey A. Rosenfeld, Fuli Yu, Richard Gibbs

Cornell
Steven Lipkin, Jishnu Das, Robert Fragoza, Xiaomu Wei, Haiyuan Yu
Andrea Sboner, Dimple Chakravarty, Naoki Kitabayashi, Vaja Liluashvili, Zeynep H. Gümüş, Mark A. Rubin

Functional Interpretation Subgroup

~50 people
~1000 “authors”

Hiring Postdocs. See gersteinlab.org/jobs!
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