Human Genome Analysis:

Building the Regulatory Network, Using it to Interpret Cancer Mutations & Practical Software Tool

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein). See last slide for references & more info.
Personal genomics soon 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.
Personal genomics soon 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.
Sources of Annotation: Comparative & Functional

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

Identify smaller-scale repeated blocks using statistical models

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

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

Group active regions into larger annotation blocks
1000G FIG

1) coding (LoF)
   MacArthur et al. Science ('12)

2) Non-coding
   Khurana et al. Science ('13)
Human Genome Annotation & Predicting Disease Drivers

• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy

• The impact of variation on networks
  – Constraint associated with centrality in the regulatory network
  – Impact is even stronger when considering multiple networks

• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Building a conceptual workflow for prioritization
  – Practical software tool for application to cancer genomes
  – Evaluation of tool's performance
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)

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

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

~500K Edges

Potential Distal Edge

Strong Proximal Edge

~26K Edges

[Cheng et al., Bioinfo. ('11); Gerstein et al. Nature (in press, '12); Yip et al., GenomeBiology (in press, '12)]
‘Supervised’ enhancer prediction

- Identifying Potential Enhancer-like Elements from Discriminative Model

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

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

Positive: Overlapping with TF peaks

Strong H3K4me1 & H3K27ac signal

~130K enhancer-like elements

Yip et al., Genome Biology (2012)
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.)
• '14 Enhancer update

• Form distal regulatory networks (we extend to edges with ~17k genes)
• Median number of regulatory elements per gene: 2
• Median number of genes per regulatory element: 22

[Fu et al., GenomeBiology ('14, in revision)]
Hierarchy Height Statistic = (normalized TF Out deg. – In deg.)

Network Stats to Identify Hierarchy
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.
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

Avg. #
of PPI for
each TF

Human Genome Annotation & Predicting Disease Drivers

- A networks view on regulatory annotation
  - Constructing the human regulatory network as a hierarchy
- The impact of variation on networks
  - Constraint associated with centrality in the regulatory network
  - Impact is even stronger when considering multiple networks
- Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  - Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  - Building a conceptual workflow for prioritization
  - Practical software tool for application to cancer genomes
  - Evaluation of tool's performance
TFs at the Top Under Stronger Negative Selection

SNP dens. $x10^{-3}$

1.0

3.1

3.8

• 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.

‘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
More connected components ("hubs") have less variation

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

This shows:

**TF target in-degree** & **TF target in-degree**

**Neg. corr.** with

Neg. corr. with

\( (SCC=-.2, P<0.5) \)

\( dN/dS \) (from chimp alignments)

ns SNP density, pN/pS, avg. DAF

Human Genome Annotation & Predicting Disease Drivers

• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy

• The impact of variation on networks
  – **Constraint associated with centrality** in the regulatory network
  – Impact is even stronger when considering multiple networks

• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding **ultra-sensitive non-coding regions** & disruptive mutations (eg motif breakers) from SNPs
  – Building a conceptual **workflow** for prioritization
  – Practical **software tool** for application to cancer genomes
  – **Evaluation** of tool's performance
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 PPI]; Cheng et al. ('09), BMC Genomics [miRNA nets]

- **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)

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 PPI]; Cheng et al. ('09), BMC Genomics [miRNA nets]

Gene categories with known phenotypic effects

Decreasing tolerance to mutation

- LoF-tol
- Neutral
- GWAS (common disease-assoc. 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

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

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

Essential genes are

Connected to more genes

Involved in more networks

Z Gumus iCAVE movie

LoF-tolerant genes
Essential genes

Size of nodes scaled by total degree

[Khurana et al., PLOS Comp. Bio. ’13]
Human Genome Annotation & Predicting Disease Drivers

• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy
• The impact of variation on networks
  – Constraint associated with centrality in the regulatory network
  – Impact is even stronger when considering multiple networks
• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Building a conceptual workflow for prioritization
  – Practical software tool for application to cancer genomes
  – Evaluation of tool's performance
Negative selection in non-coding elements

- Broad categories of regulatory regions under negative selection
- Consistent with previous studies

[Khurana et al., Science ('13)]

**Broad Categories**

- Coding
- Genomic Avg
- Enhancer
- ncRNA
- DHS
- TFSS
- General
- Chromatin
- Pseudogene

Fraction of rare SNPs

**TF Families (motifs)**

- Homeodomain
- MADs-box
- ZNF274(Prox.)
- Forkhead
- IFT/TIG
- AP2
- CBF-NFY
- ETS
- AP-2 motif
- Forkhead motif
- HMG
- p53
- NR
- AP2
- BRF
- ETS
- NF
- AP2
- BRF
- ETS

**Image**

- chr14: 99849316
- chr1: 98100579

**Motif breaking SNP**
Differential selective constraints among sub-categories

[A graph showing the fraction of rare SNPs across different categories.]

[Khurana et al., Science ('13)]
SNPs which break TF motifs are under stronger 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

~0.4% genomic coverage (~ top 25)
~0.02% genomic coverage (top 5)

~400-fold
~40-fold

[Khurana et al., Science ('13)]
• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy

• The impact of variation on networks
  – Constraint associated with centrality in the regulatory network
  – Impact is even stronger when considering multiple networks

• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Building a conceptual workflow for prioritization
  – Practical software tool for application to cancer genomes
  – Evaluation of tool's performance
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)
Identification of non-coding candidate drivers amongst somatic variants: Examples

Validation of a candidate driver identified in prostate cancer sample in *WDR74* gene promoter

- Sanger sequencing in 19 additional samples confirms the recurrence

- *WDR74* shows increased expression in tumor samples

![Diagram showing the validation process and data visualization](image)
Human Genome Annotation & Predicting Disease Drivers

• **A networks view on regulatory annotation**
  – Constructing the human regulatory network as a hierarchy

• **The impact of variation on networks**
  – **Constraint associated with centrality** in the regulatory network
  – Impact is even stronger when considering multiple networks

• **Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality**
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Building a conceptual workflow for prioritization
  – Practical software tool for application to cancer genomes
  – Evaluation of tool's performance
FunSeq.GersteinLab.org: webserver & code download (update FunSeq2.GersteinLab.org)

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
FunSeq Structure

0. Data Context

1. Weighted scoring scheme
   1.1 Core score
   1.2 recurrence module

2. Highlighting variants with additional features

[Fu et al., GenomeBiology ('14, in revision)]
Data context

- Knowledge of Genes (e.g. cancer)
- Evolutionary Conservation (e.g. GERP)
- Polymorphisms (e.g. 1000 Genomes)
- REMC Encode
- Biological Network (e.g. PPI)

- Define Sensitive Regions
- Define Regulatory Element - Gene Pairs
- Network Analysis

- Gene Lists
- Conservation
- Annotation (incl. PWMs)
- Network Centrality

Data Context
Variant Prioritization

Weighted scoring scheme

Functional annotations
- Regulatory regions
- HOT regions

Conservation analysis
- Evolutionary
- Human-specific

Nucleotide-level analysis
- Motif-breaking
- Motif-gaining

Network analysis
- Linking regulatory elements with genes
- Centrality

Variant core scores

Recurrence module

Recurrence DB

Regulatory elements

Recurrent elements

Variants in recurrent elements

Variant final scores

Highlighting variants with additional features

Knowledge of genes
- Cancer genes
- DNA repair genes
- Differentially expressed genes

User annotations
- Sample-specific epigenetic / open chromatin profiles

Variant reports

[Fu et al., GenomeBiology ('14, in revision)]
Weighted scoring scheme

- Core score
  - Each feature gets a weight

[Fu et al., GenomeBiology ('14, in revision)]
Gain of motif for *TERT* promoter mutations

- high-recurrent *TERT* promoter mutations

<table>
<thead>
<tr>
<th>Mutation Position</th>
<th>Gain of Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5 1295250</td>
<td>Ets_known10#129524#1295252#++#4#5.743#2.472</td>
</tr>
<tr>
<td>chr5 1295228</td>
<td>Ets_known10#129522#1295229#++#5#5.743#1.893</td>
</tr>
</tbody>
</table>
• Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weighed less)
  - entropy based method

\[ w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d) \]

\[ \text{Score} = \sum w_d \]

[Fu et al., GenomeBiology ('14, in revision)]
Weighted scoring scheme (2)

- Recurrence module
  - Recurrence elements – regulatory elements mutated in >= 2 samples

- Variant final score = core score + recurrence score

[Fu et al., GenomeBiology ('14, in revision)]
Recurrence database

- Identify recurrent elements in 570 whole-genome sequencing cancer samples and data from COSMIC noncoding mutations

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th># Samples</th>
<th># Somatic Mutations (SNVs)</th>
<th># Recurrent Elements</th>
<th>Genes</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML</td>
<td>7</td>
<td>271~1068</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>119</td>
<td>1043~67347</td>
<td>69,140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLL</td>
<td>28</td>
<td>522~3338</td>
<td>709</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td>88</td>
<td>1348~25131</td>
<td>74,144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Adeno</td>
<td>24</td>
<td>9284~297569</td>
<td>162,165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lymphoma B cell</td>
<td>24</td>
<td>1502~37848</td>
<td>4,233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medulloblastoma</td>
<td>100</td>
<td>44~47440</td>
<td>2,793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreas</td>
<td>15</td>
<td>1096~14998</td>
<td>2,591</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilocytic Astrocytoma</td>
<td>101</td>
<td>2~926</td>
<td>58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>64</td>
<td>1430~18225</td>
<td>36,327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COSMIC recurrent regulatory mutations</td>
<td>-</td>
<td>-</td>
<td>10,041</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Fu et al., GenomeBiology ('14, in revision)]
Highlighting variants with additional features

• Cancer genes, DNA-repair genes....
• Differentially expressed genes - a module to detect those genes from RNA-Seq.
• Other user annotations (optional), e.g. sample-specific epigenetic modifications

[ Fu et al., GenomeBiology ('14, in revision)]
Human Genome Annotation & Predicting Disease Drivers

• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy

• The impact of variation on networks
  – **Constraint associated with centrality** in the regulatory network
  – Impact is even stronger when considering multiple networks

• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding **ultra-sensitive non-coding regions & disruptive mutations** (eg motif breakers) from SNPs
  – Building a conceptual **workflow** for prioritization
  – Practical **software tool** for application to cancer genomes
  – **Evaluation** of tool's performance
Recurrent variants have higher scores than non-recurrent ones

- COSMIC noncoding variants
- Variants from 119 breast cancer samples

[Fu et al., GenomeBiology ('14, in revision)]
Germline pathogenic variants show higher scores than controls

3 controls with natural polymorphisms (allele frequency >= 1%)
1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

Ritchie et al., Nature Methods, 2014

[Fu et al., GenomeBiology ('14, in revision)]
Apply to individual tumor genomes

Cancer samples with TERT promoter mutation (chr5:1295228)

- 2 Medulloblastoma samples
- 5 Liver samples
chr 5 : 1295228 G -> A (TERT Promoter)
0.64% ( 14 / 2183)

chr 5 : 1295228 G -> A (TERT Promoter)
10.3% ( 224 / 2183)

[Fu et al., GenomeBiology ('14, in revision)]
Human Genome Annotation & Predicting Disease Drivers

• A networks view on regulatory annotation
  – Constructing the human regulatory network as a hierarchy

• The impact of variation on networks
  – Constraint associated with centrality in the regulatory network
  – Impact is even stronger when considering multiple networks

• Prioritizing cancer mutations from regulatory annotation & relationship of variation to network centrality
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Finding ultra-sensitive non-coding regions & disruptive mutations (e.g. motif breakers) from SNPs
  – Building a conceptual workflow for prioritization
  – Practical software tool for application to cancer genomes
  – Evaluation of tool's performance
Network Acknowledgements

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

A Kundaje, M Hariharan, S Landt, K Yan, C Cheng, X Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, A Boyle, P Cayting, A Charos, D Chen, Y Cheng, D Clarke, C Eastman,
G Euskirchen, S Frietze, Y Fu, J Gertz, F Grubert, A Harmanci, P Jain,
M Kasowski, P Lacroute, J Leng, J Lian, H Monahan, H O'Geen, Z Ouyang,
E Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, T Reddy, B Reed, M Shi, T Slifer,
J Wang, L Wu, X Yang, K Yip,
G Zilberman-Schapira, S Batzoglou, A Sidow, P Farnham, R Myers,
S Weissman, M Snyder

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,
Sugan thi Balasubramanian, Cristina Sisu,
Declan Clarke, Mike Wilson

Sanger
Vincenza Colonna, Yali Xue,
Chris Tyler-Smith

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

Functional Interpretation Subgroup

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

Also, in V2
Zhu Liu, Shaoke Lou, Jason Bedford, Kevin Yip

Hiring Postdocs: see gersteinlab.org/jobs!
Info about content in this slide pack

• General PERMISSIONS
  – This Presentation is copyright Mark Gerstein, Yale University, 2014.
  – Feel free to use slides & 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.

• For SeqUniverse slide, please contact Heidi Sofia, NHGRI

• 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