Human Genome Analysis: Building the Regulatory Network and Using it to Interpret Cancer Mutations

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein). See last slide for references & more info.
Personal genomes 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 as a Gateway into Biology

Personal genomes 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

Sources

Comparative & Functional
1000G FIG

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

2) Non-coding
   Khurana et al. Science ('13)
Seq Universe

TCGA endpoint: ~2.5 Petabytes
~1.5 PB exome
~1 PB whole genome

SRA >1 petabyte

1000 Genomes
A Deep Catalog of Human Genetic Variation

TCGA
910 Terabytes in CGHub

220 TB
46 TB
32 TB
31 TB
19 TB
16 TB
9 TB

National Human Genome Research Institute
ADSP
NHLBI ESP
NHGRI LSSP
ARRA Autism

Star formation
100K Genomes England

[from Hedi Sofia, NHGRI]
Human Genome Analysis: Building the Regulatory Network and Using it to Interpret Cancer Mutations

• 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**
  – **Allelic effects** in the regulatory network

• 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
  – Seeing related patterns from **indels & SVs**
  – Building a **practical workflow** & software 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

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)]
Network Stats to Identify Bottlenecks & Hubs

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

Middle level has highest betweenness, creating info. flow bottlenecks

Strongest Proximal Regulatory Edges Can be Arranged into a Hierarchy

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

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

Avg. # of PPI for each TF

Integration of TF hierarchy with other ‘omic information: more influential & connected TFs on the top
Human Genome Analysis: Building the Human Regulatory Network and Using it to Interpret Cancer Mutations

- 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
  - **Allelic effects** in the regulatory network

- 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
  - Seeing related patterns from **indels & SVs**
  - Building a **practical workflow** & software tool for disease genomes
TFs at the Top Under Stronger Negative Selection

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

1.0

3.1

3.8

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

(SCC=-.2, P<0.5)

• 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
Human Genome Analysis: Building the Human Regulatory Network and Using it to Interpret Cancer Mutations

• 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
  – Allelic effects in the regulatory network

• 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
  – Seeing related patterns from indels & SVs
  – Building a practical workflow & software 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 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]

**More Connectivity, More Constraint : A theme borne out in many studies**

- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

Result for Yeast PPI [Xia et al. PLOS CB (‘09)]
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

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

Multinet – the ultimate hairball!

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

Essential genes are
Connected to more genes
Involved in more networks

Size of nodes scaled by total degree

Z Gumus iCAVE movie

[Khurana et al., PLOS Comp. Bio. ’13]
Human Genome Analysis:
Building the Human Regulatory Network and
Using it to Interpret Cancer Mutations

- 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
  - Allelic effects in the regulatory network

- 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
  - Seeing related patterns from indels & SVs
  - Building a practical workflow & software tool for disease genomes
Inferring Allele Specific Binding/Expression using Actual Sequence Reads

RNA/ChIP-Seq Reads
ACTTTGATAGCGTCAA{T}G
CTTTGATAGCGTCAA{T}GC
CTTTGATAGCGTCAACGC
TTGACAGCGTCAA{T}GCAC
TGATAGCGTCAA{T}GCACG
ATAGCGTCAA{T}GCACGTC
TAGCGTCAA{T}GCACGTCG
CGTCAACGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAA{T}GCACGTCGGGAGTT
AA{T}GCACGTCGGGAGTTG
TGCACGTTGGGAGTTGGC

10 x T
2 x C

Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants
Many Technical Issues in Determining ASE/ASB:
Reference Bias
(naïve alignment against reference)

ASE/ASB Example:
...GTCAATGCAC
...GTCAATGCACG
...GTCAATGCACGTC
...GTCAATGCACGTCG
...GTCAACGCACGTCGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGGGAGTT
AATGCACGTCGGAGTTG

Null Example:
ACTTTGATAGCGTCAA
CTTTGATAGCGTCAACG
TTGACAGCGTCAAACAC
ATACGTCACGTCAACGT...
TAGCGTCACGTCACGT...
CGTCACGTCACGT...
CAATGCACG...
AATGCACG...

Binomial Null Distribution
(no allele-specific behavior)
Construction of a Personal Diploid Genome & Transcriptome

Reference Genome

Equivalence map

Personal Genome

Equivalence map

Paternal Haplotype

Equivalence map

Maternal Haplotype

vcf2diploid

Genotyping, Phasing, Filtering

Personal Variants

SVs

Indels

SNPs

Reference

Deletion

SNP

Insertion

Personal Haplotype

[alleleseq.gersteinlab.org]

[Rozowsky et al., MSB ('11)]
Allele-Specific Behavior in the Regulatory Network

- In GM12878, determine ASB for ~50 TFs & ASE using RNA-Seq
  - ~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

[ Rozowsky et al. MSB ('11) ]
Maternal & Paternal Personal Regulatory Networks: combinatorial coordination of ASE & ASB

[ 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 "allelic" components under weaker selection

(from 1000G pilot & phase I)

[Khurana et al. Science ('13)]
[Gerstein et al. Nature ('12)]
• **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
  – Allelic effects in the regulatory network

• **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
  – Seeing related patterns from indels & SVs
  – Building a practical workflow & software tool for disease genomes
Negative selection in non-coding elements

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

Mu et al, *NAR*, 2011

[Khurana et al., *Science* ('13)]
~700 specific sub-categories of broad non-coding categories; Possible to study now using 1000G Phase 1

- Divide broad categories
  - ncRNA: snRNA, snoRNA, miRNA, lincRNA
  - Motifs & binding sites of different TF families
  - TFBSs divide into proximal vs distal and cell-line–specific vs – non-specific

- Large sample size: 1,092 humans compared to pilot ~180

[Khurana et al., *Science* ('13)]
Differential selective constraints among subcategories

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

[Khurana et al., Science ('13)]
Negative selection and tissue-specificity of coding and non-coding regions

- Ubiquitously expressed genes and bound regions show stronger selection
- Differences in constraints amongst tissues
- Constraints in coding genes and regulatory genes are correlated across tissues

[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 know disease-causing mutations from Human Gene Mutation database

[Khurana et al., Science ('13)]
Human Genome Analysis: Building the Regulatory Network and Using it to Interpret Cancer Mutations

- A networks view on organizing regulatory annotation
  - Constructing the human **regulatory network as a hierarchy**
  - The **impact of variation** on the network

- Prioritizing cancer mutations from regulatory annotation
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  - **Prioritizing based on general network** connectivity
  - Finding **ultra-sensitive non-coding regions** & **disruptive mutations** (e.g. motif breakers) from SNPs
  - Seeing related patterns from **indels & SVs**
  - Building a **practical workflow** & software tool for disease genomes
Indels and larger SVs show largely consistent patterns to SNPs

Figure 4

Structural variants are generally depleted for functional elements

[Khurana et al., Science ('13)]
TF binding sites have a complex relationship with SVs, depending on their mechanisms

- **% Enrichment**
  - CDS: 538
  - 5’UTR only: 300
  - 3’UTR only: 308
  - intron only: 5861
  - partial: 5992
  - whole: 90

- **# SVs intersecting elem.**
  - gene: all: 6082
  - pseudogene: 519
  - all TF motifs: 2098

- **Mechanism**
  - enhancer: NAHR: 198, VNTR: 9, NH: 399, TEI: 34

NAHR: non-allelic homologous recombination
VNTR: variable number of tandem repeats
NH: non-homologous events
TEI: transposable element insertions
TF: transcription factor

Khurana et al, Science 2013
Histone modifications at SV breakpoints also differs depending on mechanism

Two chromatin states: transcriptionally active and structurally accessible; transcriptionally repressive and structurally condensed. H3K4me1 marks the active state. H3K27me3 marks the repressive state.

Supports "looping conjecture" for NAHR
Human Genome Analysis: Building the Regulatory Network and Using it to Interpret Cancer Mutations

• A networks view on organizing regulatory annotation
  – Constructing the human regulatory network as a hierarchy
  – The impact of variation on the network

• Prioritizing cancer mutations from regulatory annotation
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Prioritizing based on general network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Seeing related patterns from indels & SVs
  – Building a practical workflow & software 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 promotor mutation – exception proves the rule!
Germline vs somatic variants

- Somatic mutations do not follow patterns of natural polymorphisms
- Those deviating the most from these patterns are most likely to be cancer drivers providing selective advantage to the tumor cells (confirmed for protein-coding genes)
- Look for mutations in elements under strong negative selection

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

[Cancer genome variants]

1. 1000 Genomes screen
2. Functional annotation
3. Sensitive
4. Disruptive
5. Network connectivity
6. Recurrence
7. Candidate driver

[SNV \ Indel]

[Non-coding annotation]

[Degree of negative selection]

[Motif disruptive score]

[Degree of network centrality]

[Enhancer / Promoter]

[Occurrence in multiple samples]

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

[Diagram of the flowchart with detailed steps for identifying candidate drivers based on various genomic parameters such as 1000 Genomes Screen, Functional annotation, Sensitive, Disruptive, Network connectivity, and Recurrence.]
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

\begin{itemize}
  \item \textbf{WDR74} shows increased expression in tumor samples
\end{itemize}

\[\text{ACGGR\_TC\_CC\_GT\_A\_A\_A\_ATAGA...}\]

\[\text{chr11: 62,609,084}\]

\[\text{chr11: 62,609,138}\]
Human Genome Analysis: Building the Regulatory Network and Using it to Interpret Cancer Mutations

• A networks view on organizing regulatory annotation
  – Constructing the human regulatory network as a hierarchy
  – The impact of variation on the network

• Prioritizing cancer mutations from regulatory annotation
  (i.e. using 1000G & ENCODE to characterize natural patterns of variation in regulatory elements & identifying drivers as somatic mutations breaking these patterns.)
  – Prioritizing based on general network connectivity
  – Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers) from SNPs
  – Seeing related patterns from indels & SVs
  – Building a practical workflow & software tool for disease genomes
Acknowledgements

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

Networks/Elements (~60 participants):


Alleleseq.gersteinlab.org:
J Rozowsky, A Abyzov, J Wang, P Alves, D Raha, A Harmanci, J Leng, R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj, M Rubin, M Snyder

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

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