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 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.
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

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

2) Non-coding  
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 (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.)

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)]
Network Stats to Identify Bottlenecks & Hubs

[Hub-bottleneck node]
[Non-hub-bottleneck node]
[Hub-non-bottleneck node]
[Non-hub-non-bottleneck node]

[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

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

TFs at the Top Under Stronger Negative Selection

SNP dens. $x 10^{-3}$

1.0

3.1

3.8

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
Gene categories with known phenotypic effects

Decreasing tolerance to mutation

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

- **Neutral**

- **GWAS** (common disease-assoc. variants)

- **HGMD** (rare disease-causing variants)

- **Essential**
  - 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

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.

LoF-tolerant genes
Essential genes

Z Gumus iCAVE movie

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

[Khurana et al., PLOS Comp. Bio. ’13]

AUC=0.91

Predict higher scores for disease-causing genes, GWAS genes & cancer drivers
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
Enrichment of rare 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)]
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 known 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 (eg 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

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

**Figure 4. Functional annotations of indels and SVs.**

- **A** Fraction of rare indels in coding-gene categories. Total number of indels shown.
- **B** Enrichment of SVs affecting functional annotations. Middle box shows genes, pseudogenes, and TF motifs; upper blow-out shows gene parts in different modes, and bottom blow-out shows enhancers with different formation mechanisms, i.e., NAHR, NH (non-homologous), TEI (transposable element insertion), and VNTR (variable number of tandem repeats). Asterisks indicate significant enrichment (green) or depletion (red) after multiple hypothesis correction.
- **C** Aggregation of histone signal around breakpoints of deletions formed by different mechanisms. Breakpoints centered at zero. Aggregation for upstream and downstream regions corresponds to negative and positive distance, respectively. Signals for an activating histone mark (H3K4me1) and a repressive mark (H3K27me3) are shown.

**Mechanism**

- 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.
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
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

[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.
Flowchart for 1 Prostate Cancer Genome (from Berger et al. '11)

[Diagram of the flowchart showing the process from identifying somatic SNVs to determining if a gene is a driver in the genome.]

[Reference to Berger et al., Science ('11)]
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: 
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):


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

~50 people ~1000 “authors”

Functional Interpretation Subgroup

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