Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

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

Mark Gerstein

Yale
Where is Waldo?
(Using Annotation to Prioritize the ~3M Germline variants & ~5K Somatic Mutations in a Tumor Sample)
Sources of Annotation: Comparative & Functional
1000G FIG

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

2) Non-coding
   Khurana et al. *Science* ('13)
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

- Finding **ultra-sensitive non-coding regions & disruptive mutations** (e.g., motif breakers)
  - Using 1000G & ENCODE to characterize natural patterns of **SNPs** in regulatory elements
  - Finding similar but not identical patterns for **indels & SVs**

- Also, annotation based on **network connectivity**
  - Prioritize hubs

- Building a **practical workflow & software tool** for cancer genomes
  - Identifying drivers as somatic variants breaking natural patterns
Gene categories with known phenotypic effects

- Decreasing tolerance to mutation

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

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

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

From Liao et al, PNAS, 2008
Enrichment of rare SNPs as a metric for Negative Selection

- **Initial Metrics for selection**
  - **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

[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 sub-categories

[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

[Khurana et al., Science ('13)]
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

• Finding **ultra-sensitive non-coding regions** & **disruptive mutations** (e.g. motif breakers)
  – Using 1000G & ENCODE to characterize natural patterns of **SNPs** in regulatory elements
  – Finding similar but not identical patterns for **indels & SVs**

• Also, annotation based on **network connectivity**
  – Prioritize hubs

• Building a **practical workflow** & software tool for cancer genomes
  – Identifying drivers as somatic variants breaking natural patterns
Indels and larger SVs show largely consistent patterns to SNPs

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

- **CDS** (538 SVs): % Enrichment
- **5'UTR only** (300 SVs)
- **3'UTR only** (308 SVs)
- **intron only** (5861 SVs)
- **partial** (5992 SVs)
- **whole** (90 SVs)

### SV Mechanisms
- **NAHR**: non-allelic homologous recombination (198 SVs)
- **VNTR**: variable number of tandem repeats (9 SVs)
- **NH**: non-homologous events (399 SVs)
- **TEI**: transposable element insertions (34 SVs)

**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.

Khurana et al, Science 2013
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

- Finding ultra-sensitive non-coding regions & disruptive mutations (e.g., motif breakers)
  - Using 1000G & ENCODE to characterize natural patterns of SNPs in regulatory elements
  - Finding similar but not identical patterns for indels & SVs

- Also, annotation based on network connectivity
  - Prioritize hubs

- Building a practical workflow & software tool for cancer genomes
  - Identifying drivers as somatic variants breaking natural patterns
Gene centralalities in networks reflect their selection constraints

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

Multinet – the ultimate hairball!

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.

Size of nodes scaled by total degree.

[Khurana et al., *PLOS Comp. Bio.* '13]
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

- Finding **ultra-sensitive non-coding regions** & disruptive mutations (eg motif breakers)
  - Using 1000G & ENCODE to characterize natural patterns of **SNPs** in regulatory elements
  - Finding similar but not identical patterns for **indels & SVs**

- Also, annotation based on **network connectivity**
  - Prioritize hubs

- Building a **practical workflow** & software tool for cancer genomes
  - Identifying drivers as somatic variants breaking natural patterns
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
  - but TERT promotor mutation
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

[1000 Genomes screen

[Functional annotation

[Sensitive

[Disruptive

[Network connectivity

[Enhancer / Promoter

[Candidate driver

[SNV \ Indel

[Non-coding annotation

[Degree of negative selection

[Motif disruptive score

[Degree of network centrality

[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)
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

![Sanger sequencing of FAM48A binding site (~570 bp) in WDR74 promoter from 19 additional samples](Khurana et al., Science ('13))
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

- Finding ultra-sensitive non-coding regions & disruptive mutations (e.g., motif breakers)
  - Using 1000G & ENCODE to characterize natural patterns of SNPs in regulatory elements
  - Finding similar but not identical patterns for indels & SVs

- Also, annotation based on network connectivity
  - Prioritize hubs

- Building a practical workflow & software tool for cancer genomes
  - Identifying drivers as somatic variants breaking natural patterns
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

~50 people
~1000 “authors”

Cornell
Steven Lipkin, Jishnu Das, Robert Fragoza, Xiaomu Wei, Haiyuan Yu
Andrea Sboner, D Chakravarty, N Kitabayashi, Vaja Liluashvili, Zeynep Gümüş, Mark Rubin

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