Human Genome Analysis:
Large-scale Non-coding Annotation & Application to Cancer Genomics

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 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.
Non-coding Annotations: Overview

There are several collections of information "tracks" related to non-coding features.

Sequence features, incl. **Conservation**

**Functional Genomics**
Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription

Functional Genomics Annotations

A) PEAKS
1. DNase peaks at the UCSC genome browser {on many cell lines}
2. The regulation track at the UCSC genome browser, with compilation of TF ChIP-seq peaks from uniform processing (individual peaks are annotated with TF and cell line)
3. Blacklist Regions

B) RNA BASICS [~Regulatory]
4. A matrix of expression data of known genes (or exons) for protein-coding genes & known ncRNAs {on many cell lines}
5. Novel RNA contigs track, i.e., possible novel transcripts. "Transcriptionally Active Regions (TARs)"
6. Novel junctions

C) PROMOTERS
Annotated GENCODE TSSes (also, TSSes with FANTOM CAGE support)

D) ENHANCERS (Supervised)
- Yip et al., Ren et al. &c

E) UNSUPERVISED SEGMENTATIONS, INCLUDING ENHANCERS
- ChromHMM, SegWay, HiHMM....

F) HOT/LOT REGIONS

G) CONNECTIVITY
7. Enhancer-target gene connection
8. TF-target network connectivity
9. TADs: Topologically Associated Domains.

H) Models

I) Motifs for TF binding

J) OTHER [~Regulatory]
List of Allelic SNPs & Regions
Human Genome Analysis: Large-scale Non-coding Annotation & Application to Cancer Genomics

- **Noncoding Annotation**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale Peak calling (with Music) & enhancer finding
    - Sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Constructing the human regulatory network
    - More connectivity = more constraint

- **Application: Prioritization of cancer mutations**
  - Building a conceptual workflow for prioritization
  - **FunSeq software tool**
    - Summarizing large data context
  - Validating the tool
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

• Mostly in Tier 1 cell lines
  – K562, GM12878, H1h-ESC…
• Matching RNA-Seq data in all cell-lines
• SPP & PeakSeq
• thresholding w. IDR (replicas)


94 partner-factors

2785 GATA1 (focus-factor) peak locations
Peak Calling

- Generate and threshold the signal profile and identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)

- Score against the control

- New Challenges:
  - Different shapes and scales of the ChIP-Seq profiles determines the mode of study.
  - Non-uniform mappability of the genome tends to fragment large peaks, e.g. histone marks.
Mulyscale Analysis, Minima/Maxima based Coarse Segmentation

Maxima

Minima

Harmanci et al, Genome Biology 2014, In Press
Multi-mappability based Correction

- Low mappability regions cause loss of signal and introduce burst-like noise
- To characterize the mappability of the genome, we build the **multi-mappability profile**
  - High multi-mappability signal ↔ Low mappability
- Correction Procedure:
  - “Whenever there is a lowly mappable position, use the surrounding regions with high mappability to correct the value”

![Graph showing loss of read depth signal due to low mappability](H3K36me3)

[Harmanci et al, Genome Biol. ('14, in press). MUSIC.gersteinlab.org]
MUSIC Algorithm

[Harmaneci et al, Genome Biol. ('14, in press). MUSIC.gersteinlab.org]
‘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)
• 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
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)]
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)]
Loss- and gain- of motif mutations

Loss-of-motif

Gain-of-motif
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: Large-scale Non-coding Annotation & Application to Cancer Genomics

• Noncoding Annotation
  – 1st Level Linear Annotation: Regulatory Sites
    • Multi-scale Peak calling (with Music) & enhancer finding
    • Sites particularly sensitive to mutations
  – 2nd Level Network Annotation
    • Constructing the human regulatory network
    • More connectivity = more constraint

• Application: Prioritization of cancer mutations
  – Building a conceptual workflow for prioritization
  – FunSeq software tool
    • Summarizing large data context
  – Validating the tool
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

~26K Edges

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

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

• Among ~ 2.2M regulatory elements, we identified ~ 703K significant associations between ~ 332K regulatory elements and ~ 17K genes
• Median number of regulatory elements per gene: 26
• Median number of genes per regulatory element: 1

[Fu et al., GenomeBiology ('14, in revision)]
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**
  - Neg. corr. with
    - dN/dS (from chimp alignments)

- **TF target in-degree & TF out-degree**
  - Neg. corr. with
    - ns SNP density, pN/pS, avg. DAF

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 and involved in more networks compared to LoF-tolerant genes.

Size of nodes scaled by total degree.

[Khurana et al., PLOS Comp. Bio. '13]
Human Genome Analysis:
Large-scale Non-coding Annotation & Application to Cancer Genomics

• **Noncoding Annotation**
  - **1st Level Linear Annotation: Regulatory Sites**
    • Multi-scale Peak calling (with Music) & enhancer finding
    • Sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    • Constructing the human regulatory network
    • More connectivity = more constraint

• **Application: Prioritization of cancer mutations**
  - Building a conceptual workflow for prioritization
  - **FunSeq software tool**
    • Summarizing large data context
  - Validating the tool
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!
Where is Waldo?
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)
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

[Khurana et al., Science (’13)]
Human Genome Analysis:
Large-scale Non-coding Annotation & Application to Cancer Genomics

• Noncoding Annotation
  – 1st Level Linear Annotation: Regulatory Sites
    • Multi-scale Peak calling (with Music) & enhancer finding
    • Sites particularly sensitive to mutations
  – 2nd Level Network Annotation
    • Constructing the human regulatory network
    • More connectivity = more constraint

• Application: Prioritization of cancer mutations
  – Building a conceptual workflow for prioritization
  – FunSeq software tool
    • Summarizing large data context
  – Validating the tool
FunSeq2 - A flexible framework to prioritize regulatory mutations from cancer genome sequencing

**Overview**

This tool is specialized to prioritize somatic variants from cancer whole genome sequencing. It contains two components: 1) building data context from various resources; 2) variants prioritization. We provided downloadable scripts for users to customize the data context (found under 'Downloads'). The variants prioritization step is downloadable, and also implemented as web server (Right Panel), with pre-processed data context.

**Instructions**

- **Input File** - BED or VCF formatted. Click "green" button to add multiple files. With multiple files, the tool will do recurrent analysis. (Note: for BED format, user can put variants from multiple genomes in one file, see Sample input file.)
- **Recurrence DB** - User can choose particular cancer type from the database. The DB will continue be updated with newly available WGS data.
- **Gene List** - Option to analyze variants associated with particular set of genes. Note: Please use Gene Symbols, one row per gene.
- **Differential Gene Expression Analysis** - Option to detect differentially expressed genes in RNA-Seq data. Two files needed: expression file & class label file. Please refer to Expression input files for instructions to prepare these files.

**Note:** In addition to on-site calculation, we also provide scores for all possible noncoding SNVs of GRCh37/hg19 under 'Downloads' (without annotation and recurrence analysis).

**Input File:** (only for hg19 SNVs)

| Choose file | No file chosen |

**Output Format:**

- **bed**
- **MAF:**

0

- **Minor allele frequency threshold to filter polymorphisms from 1KG (value 0–1)**

**Cancer Type from Recurrence DB:**

- **Summary table**
- **All Cancer Types**

**Add a gene list** (Optional)

**Add differential gene expression analysis** (Optional)

**Upload**
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)

Gene Lists
Conservation
Annotation (incl. PWMs)
Network Centrality

Define Sensitive Regions
Define Regulatory Element - Gene Pairs
Network Analysis

~10TB

~10 GB

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

• Core score
  – Each feature gets a weight

User-variants

Functional annotations
  - Regulatory regions
  - HOT regions

Conservation
  - Evolutionary
  - Sensitive/Ultra-sensitive regions

Network study
  - Linking regulatory elements with genes
  - Hubs

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

Variant core scores

[Fu et al., GenomeBiology ('14, in revision)]
• Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weighed less)
  - Entropy based method

Feature weight: \( w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d) \)

\( p \) \text{ up} \quad \downarrow w_d \quad p = \text{probability of the feature overlapping natural polymorphisms}

For a variant: \( \text{Score} = \sum w_d \) of observed features

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

- Recurrent elements
  - regulatory elements mutated in >= 2 samples
- Recurrence database
  - 670 whole-genome sequenced samples
  - 11 cancer types
  - also contains COSMIC dataset

[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

[Sakabe et al., Genome Biology, 2012]

[Fu et al., Genome Biology ('14, in revision)]
Human Genome Analysis:
Large-scale Non-coding Annotation & Application to Cancer Genomics

• Noncoding Annotation
  – 1st Level Linear Annotation: Regulatory Sites
    • Multi-scale Peak calling (with Music) & enhancer finding
    • Sites particularly sensitive to mutations
  – 2nd Level Network Annotation
    • Constructing the human regulatory network
    • More connectivity = more constraint

• Application: Prioritization of cancer mutations
  – Building a conceptual workflow for prioritization
  – FunSeq software tool
    • Summarizing large data context
  – Validating the tool
Recurrent variants have higher core 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 core 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)

- E.g. one Medulloblastoma sample with 2,183 somatic SNVs
Human Genome Analysis:
Large-scale Non-coding Annotation & Application to Cancer Genomics

• Noncoding Annotation
  – 1st Level Linear Annotation: Regulatory Sites
    • Multi-scale Peak calling (with Music) & enhancer finding
    • Sites particularly sensitive to mutations
  – 2nd Level Network Annotation
    • Constructing the human regulatory network
    • More connectivity = more constraint

• Application: Prioritization of cancer mutations
  – Building a conceptual workflow for prioritization
  – FunSeq software tool
    • Summarizing large data context
  – Validating the tool
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

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

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
Acknowledgements

- MUSIC.gersteinlab.org
  - A Harmanci, J Rozowsky
- encodenets.gersteinlab.org/metatracks
  - K Yip, C Cheng, N Bhardwaj, JB Brown, J Leng, A Kundaje, J Rozowsky, E Birney, P Bickel, M Snyder
- FunSeq2.gersteinlab.org
  - Y Fu, Z Liu, S Lou, J Bedford, X Mu, K Yip, E Khurana
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