OUR FIELD IS
THE BIG
WINNER
PAST & PRESENT GROUP

PhD Students
Miriam Hirshberg
Chris Lee
Britt Park
Dave Hinds
Enoch Huang
Jerry Tsai
Yu Xia
Michael Sykes
Rachel Kolodny
Nizar Batada
Sergio Moreno

Dahlia Weiss
Gaurav Chopra
Adelene Sim

Post Docs
S. Subbiah
Valerie Daggett
Peter David
Mark Gerstein
Steven Brenner
Boris Fain
Chen Keasar

Golan Yona
Ram Samudrala
Tanya Raschke
Patrice Koehl
Erik Sandelin
Chris Summa
Erik Lindahl
Julian Gough
Peter Minar
Gunnar Schroeder
Karine Bastard
Avraham Samson
Xuhui Huang
Julie Bernauer
Alena Shmygelska
Mitul Saha
Andrea Scaiwitz
Jenelle Bray
Marie Brut
Junjie Zhang
Leonid Pereyaslavets
Yana Gofman
Ivan Ufimtsev

Canada
China
France
Germany

Hungary
India
Israel
Pakistan

Russia
Singapore
Spain
Sweden

UK
Ukraine
Uruguay
USA

Recent Research Support: NIH, NSF, HFSP
Human Genome Analysis:

Prioritizing Non-coding Mutations in Cancer

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein). See last slide for references & more info.
• Worked as postdoc in Michael's lab '93 to '96
• Focused mostly on molecular mechanics of water & proteins

• Got excited by sequencing of first bacterial genome in '95

• Given lots of freedom to explore... & did so

• Where has this taken me in ~15 years...
Where is Waldo?
(Using Annotation to Prioritize the ~3M Germline variants & ~5K Somatic Mutations in a Tumor Sample)
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

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)
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
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
‘Conservation’
- Typically defined by comparison across species

- 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.

How do we define ‘sensitivity’ within human population?
- a depletion of common variants/ an enrichment of rare variants
Enrichment of rare SNPs as a metric for Negative Selection

LOF-tol (Loss-of-function tolerant): least negative selection
Cancer: most selection

[rare=derived allele freq < 0.5%]

[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)]
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)]
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)]
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

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.

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

- **gene**: 6082
- **pseudogene**: 519
- **all TF motifs**: 2098

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

**Abbreviations**
- NAHR: non-allelic homologous recombination
- VNTR: variable number of tandem repeats
- NH: non-homologous events
- TEI: transposable element insertions
- TF: transcription factor
Two chromatin states: transcriptionally active and structurally accessible; transcriptionally repressive and structurally condensed. H3K4me1 marks the active state. H3K27me3 marks the repressive state.

Histone modifications at SV breakpoints also differs depending on mechanism.

Supports "looping conjecture" for NAHR
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
Recasting Association of Connectivity with Constraint in terms of Gene Essentiality Categories: Application to Regulatory & Protein Interaction Networks

Regulatory Network

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).

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

Essential genes are connected to more genes.

Involved in more networks.

[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 promoter mutation
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.

Copyright © 2013, GersteinLab@Yale
Flowchart for 1 Prostate Cancer Genome  
(from Berger et al. '11)

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

Yale
Ekta Khurana, Yao Fu, Jieming Chen,
Xinmeng Mu, Lucas Lochovsky,
Arif Harmanci, Alexej Abyzov,
Suganathi 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

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

US, UK, Switzerland….

Functional Interpretation Subgroup
~50 people
~1000 “authors”

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