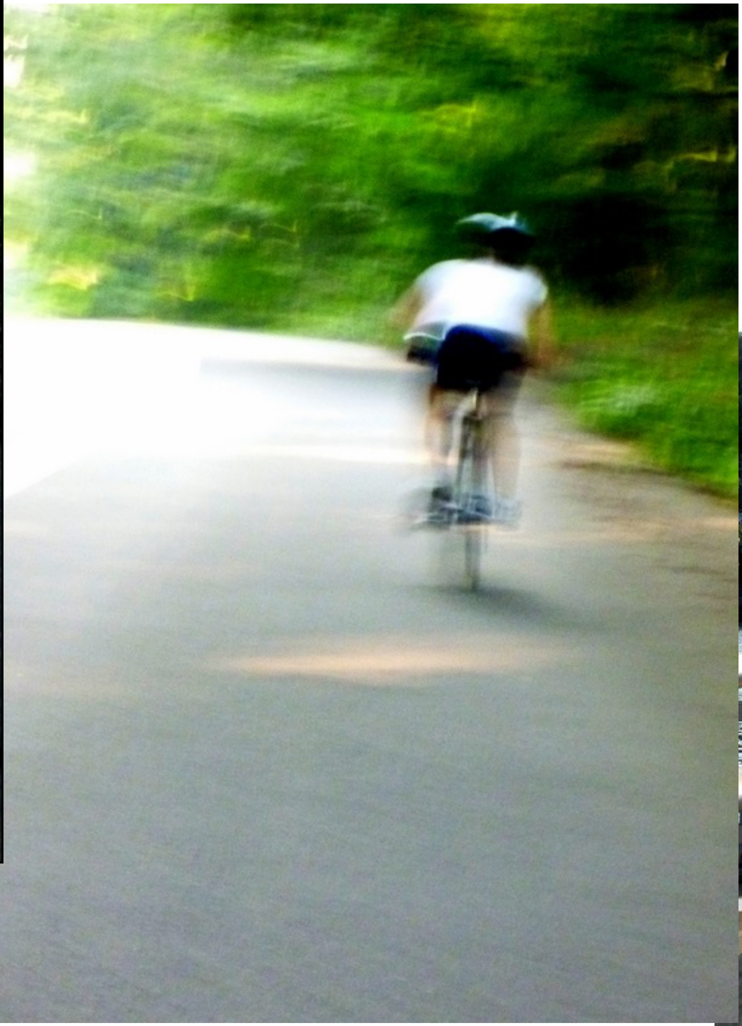
Explorations in Summer Camp in CT:
Prioritizing non-coding mutations as potential cancer drivers

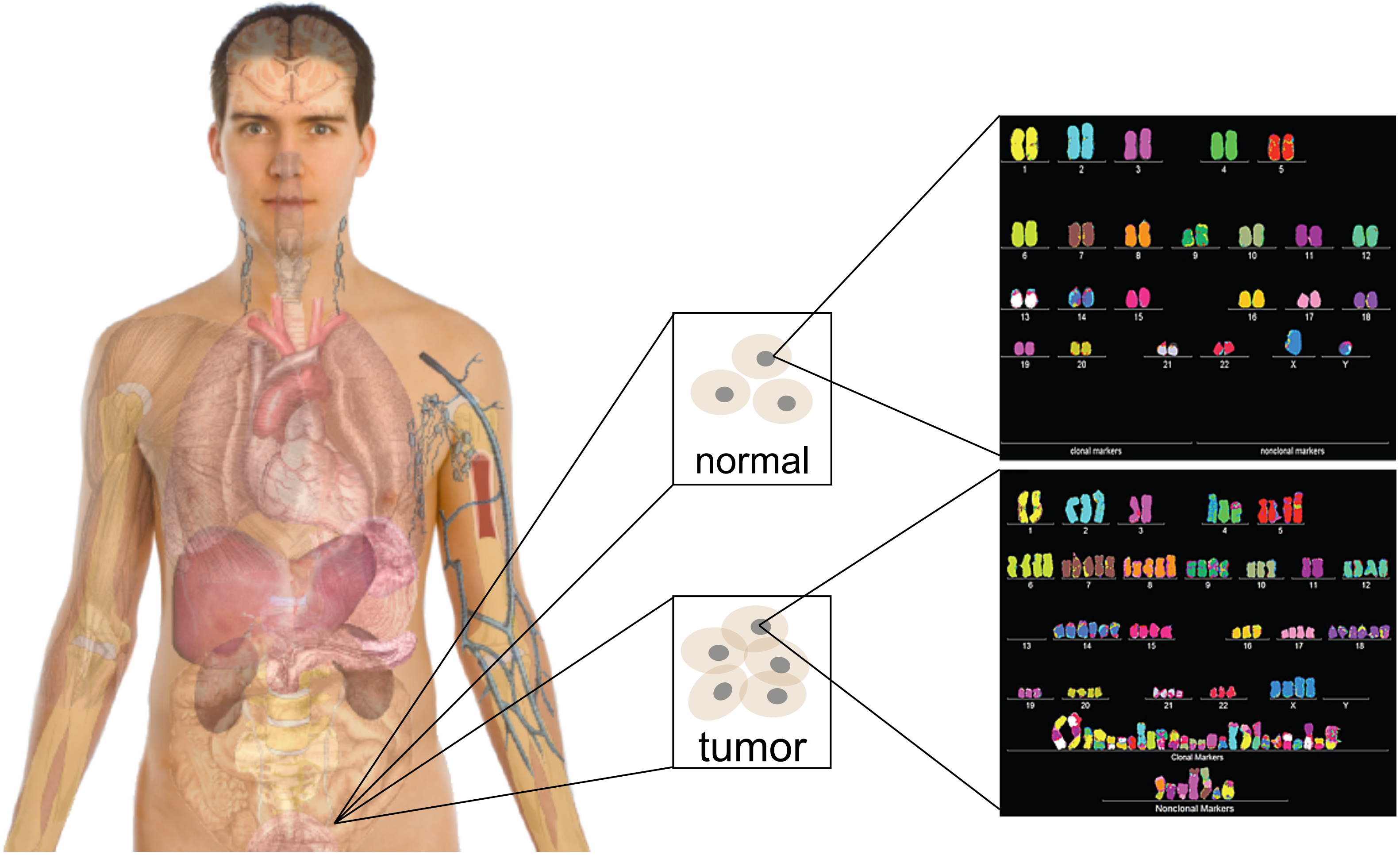
Mark
Gerstein



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

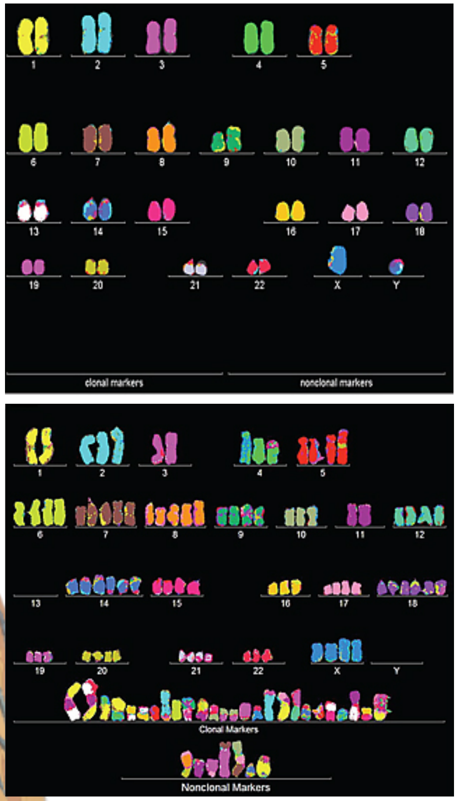
Personal Genomics as a Gateway into Biology

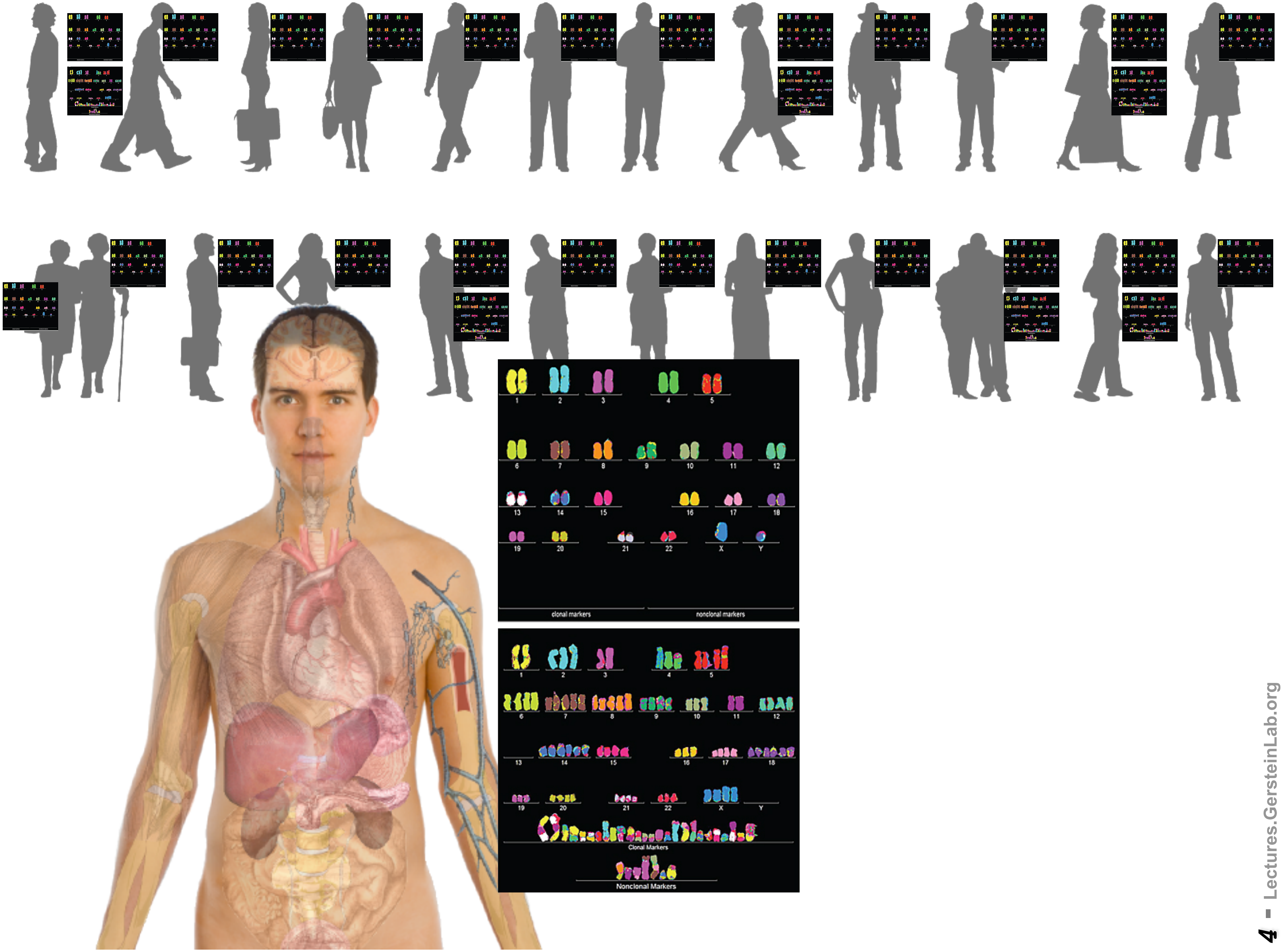
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.



Personal Genomics as a Gateway into Biology

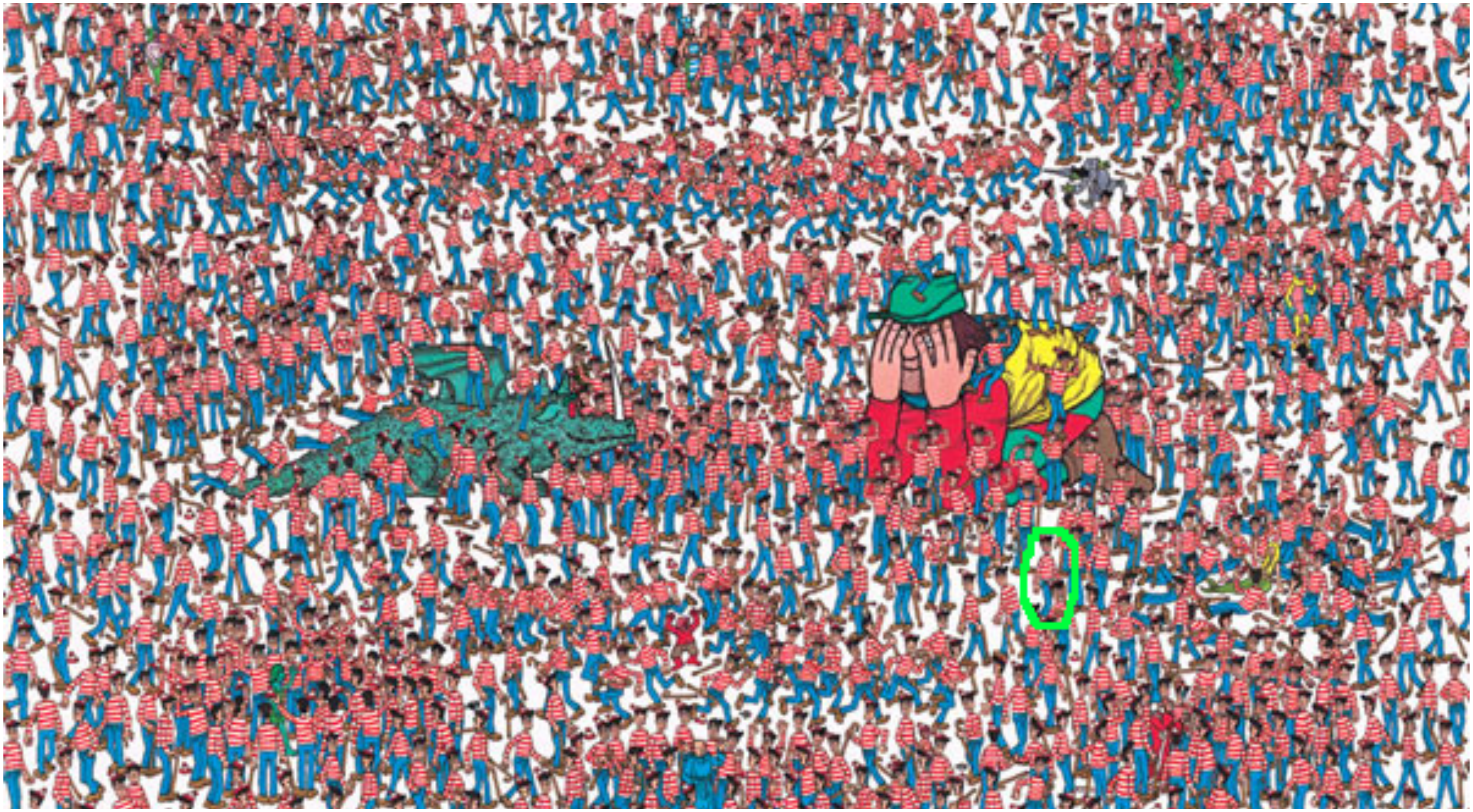
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
clonal markers
nonclonal markers
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Clonal Markers
Nonclonal Markers

Where is Waldo?

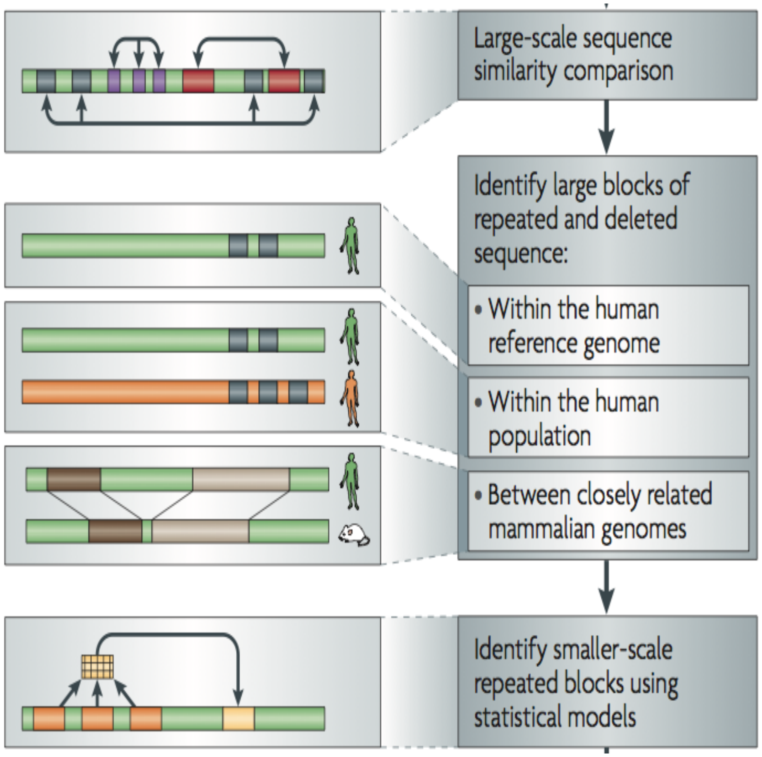
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



Non-coding Annotations: Overview

Most of cancer genomics has focused on mutations in non-coding regions – ie the exome
There are several collections of information "tracks" related to non-coding features, perhaps of use

Sequence features, incl. **Conservation**



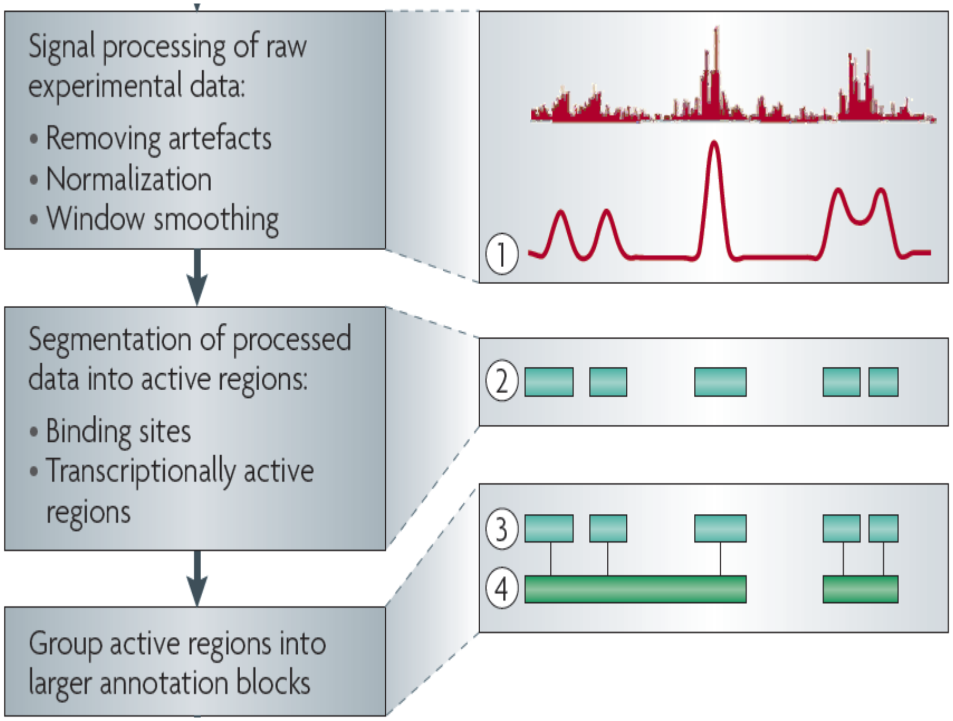
Large-scale sequence similarity comparison

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

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



Signal processing of raw experimental data:

- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:

- Binding sites
- Transcriptionally active regions

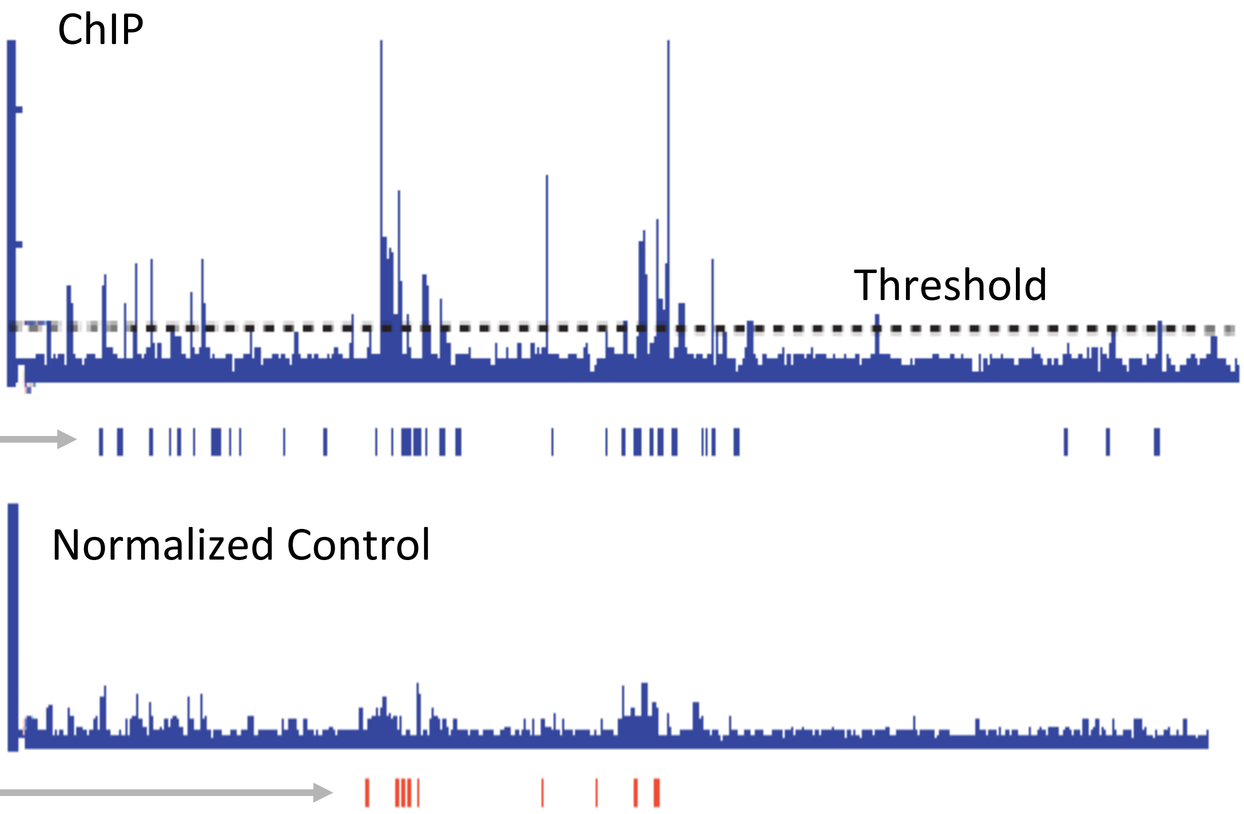
Group active regions into larger annotation blocks

Summer Camp Explorations:
Prioritizing non-coding mutations as potential cancer drivers

- **Finding Non-coding Regions Sensitive to Mutations**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale "site" calling (with Music)
    - Finding small number of sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Building a network from the linear annotation
    - More connectivity = more constraint => highlights hubs
- **Using this to Interpret Alterations in Cancer**
  - **LARVA: to find recurrently mutated annotations**
    - Need to correct for overdispersion in bionomial
    - Use beta-bin parameterized according to replication timing
  - **FunSeq software tool for mutation prioritization**
    - Systematically weighting all the features, for non-coding prioritization

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

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



Potential Targets

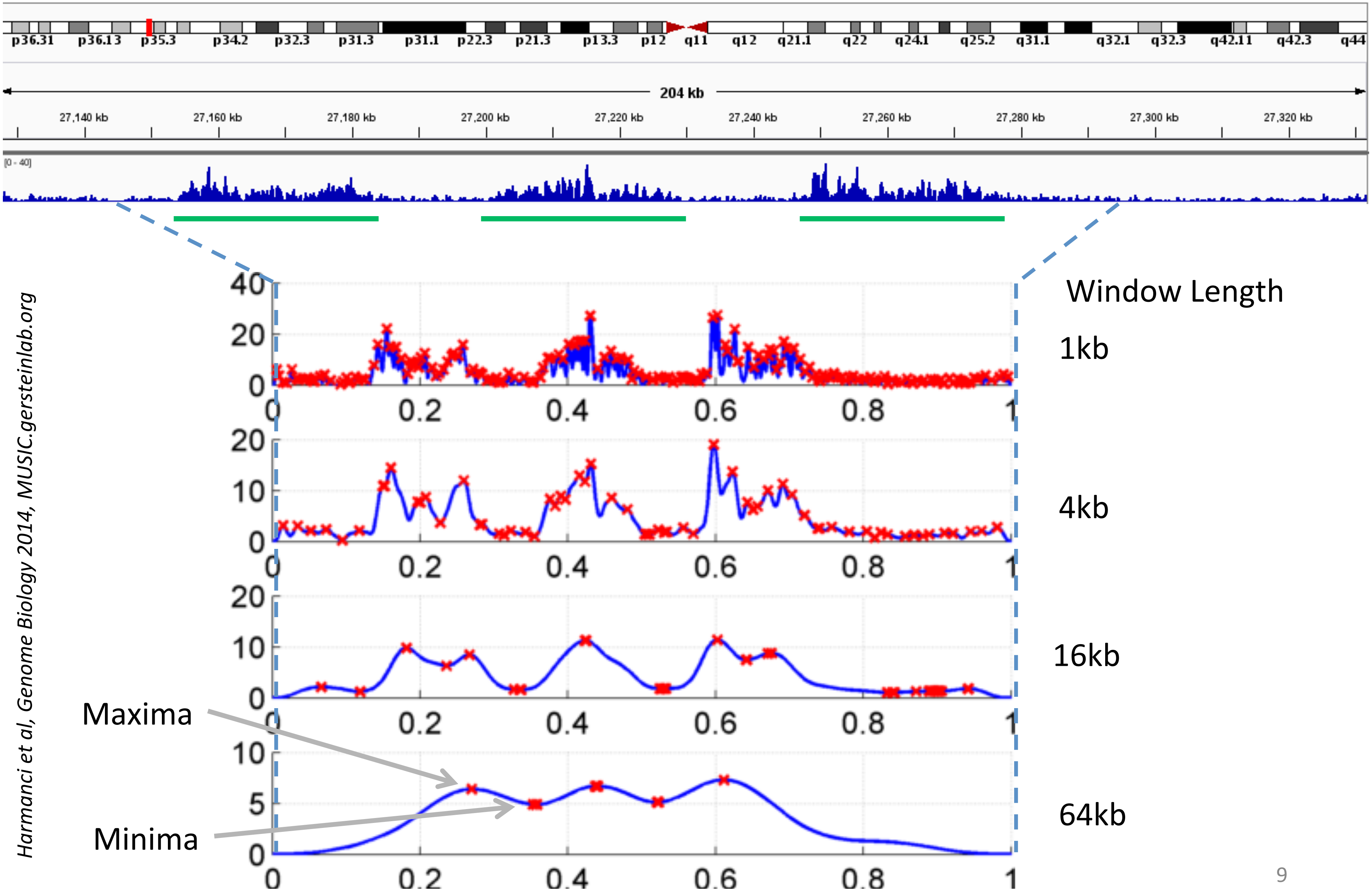
- Score against the control

Significantly Enriched targets

Now an update: "PeakSeq 2" => MUSIC

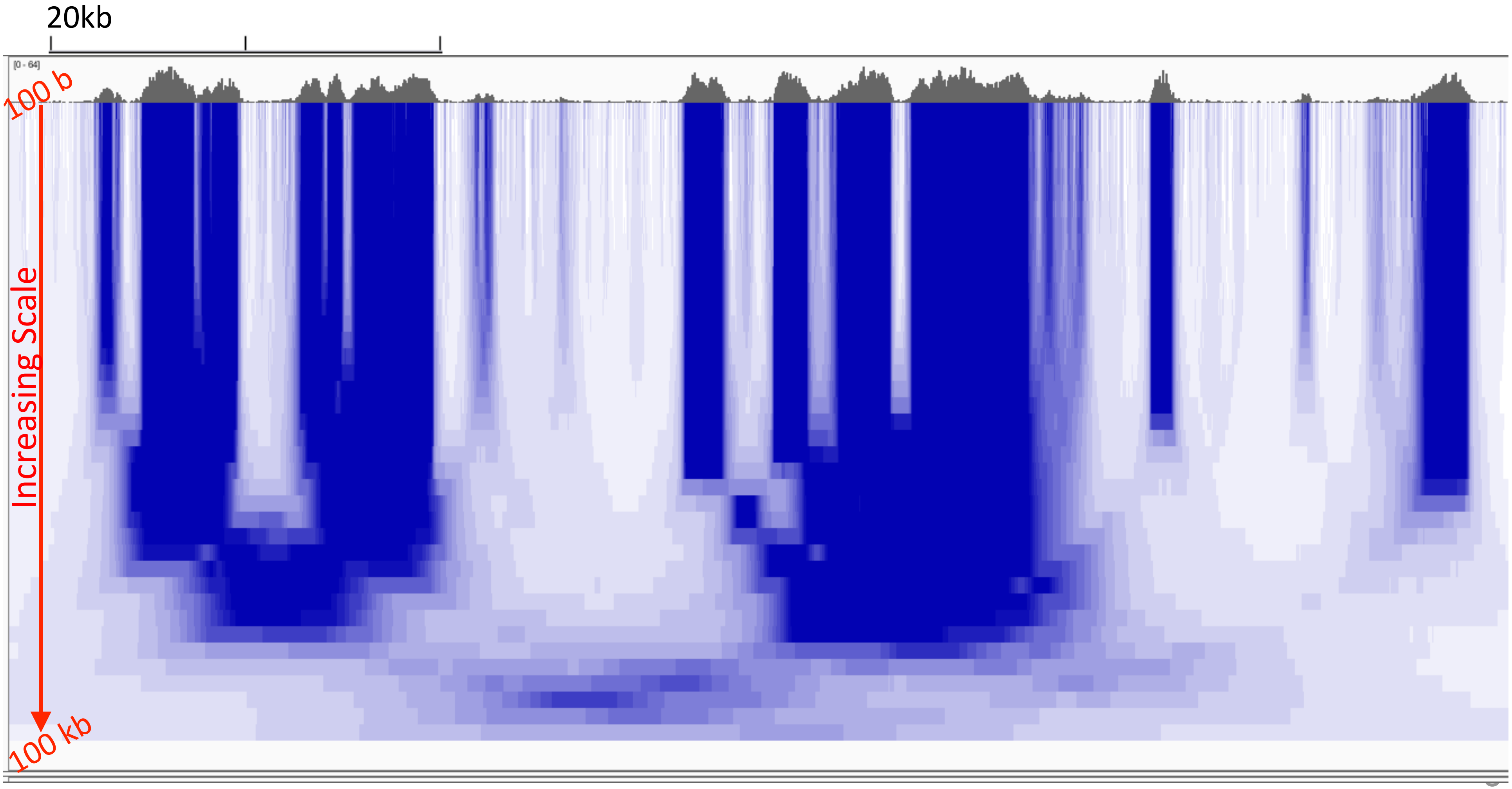
[Rozowsky et al. ('09) *Nat Biotech*]

Multiscale Analysis, Minima/Maxima based Coarse Segmentation



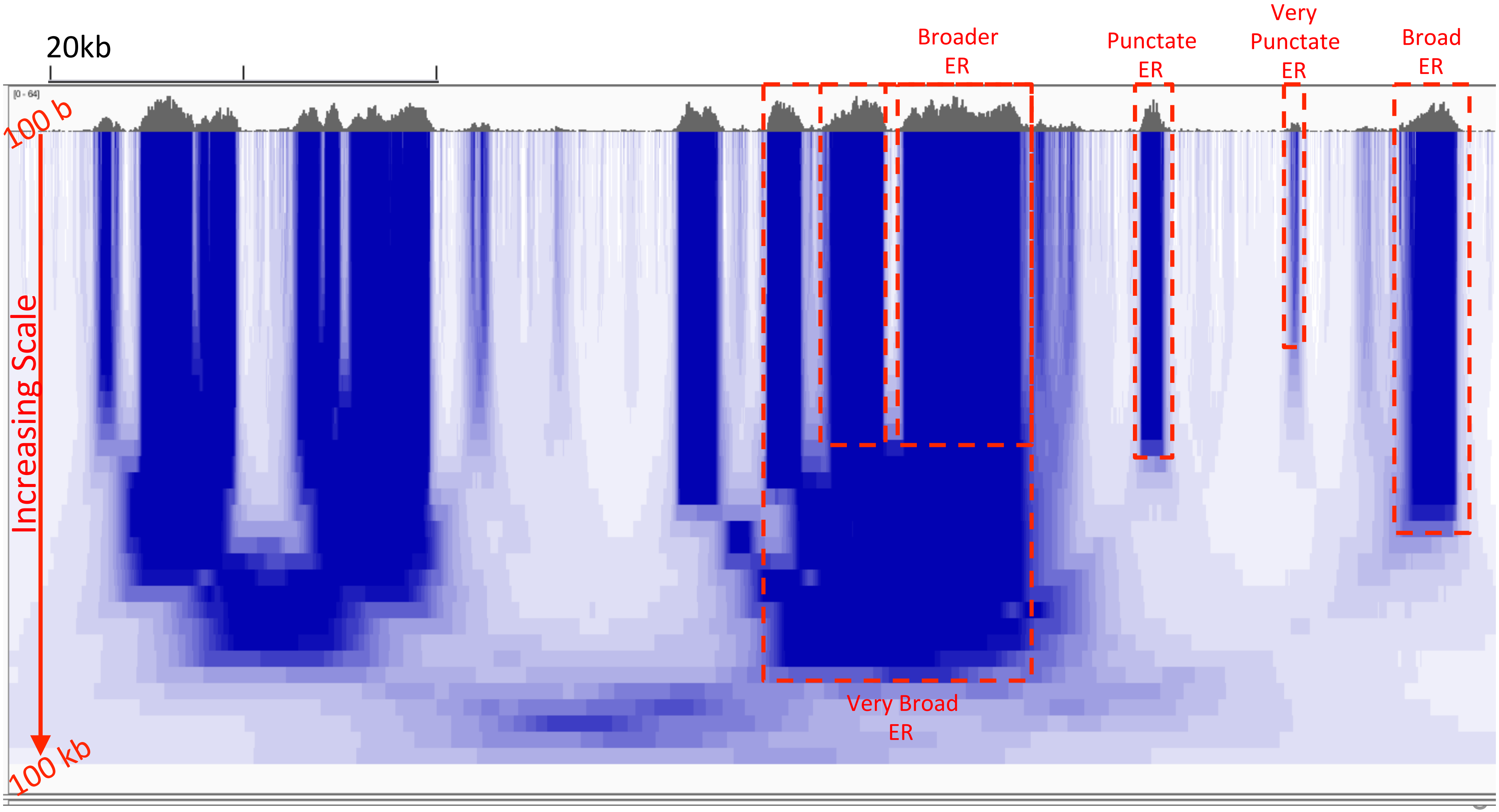
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



[Harmanci *et al, Genome Biol.* ('14)]

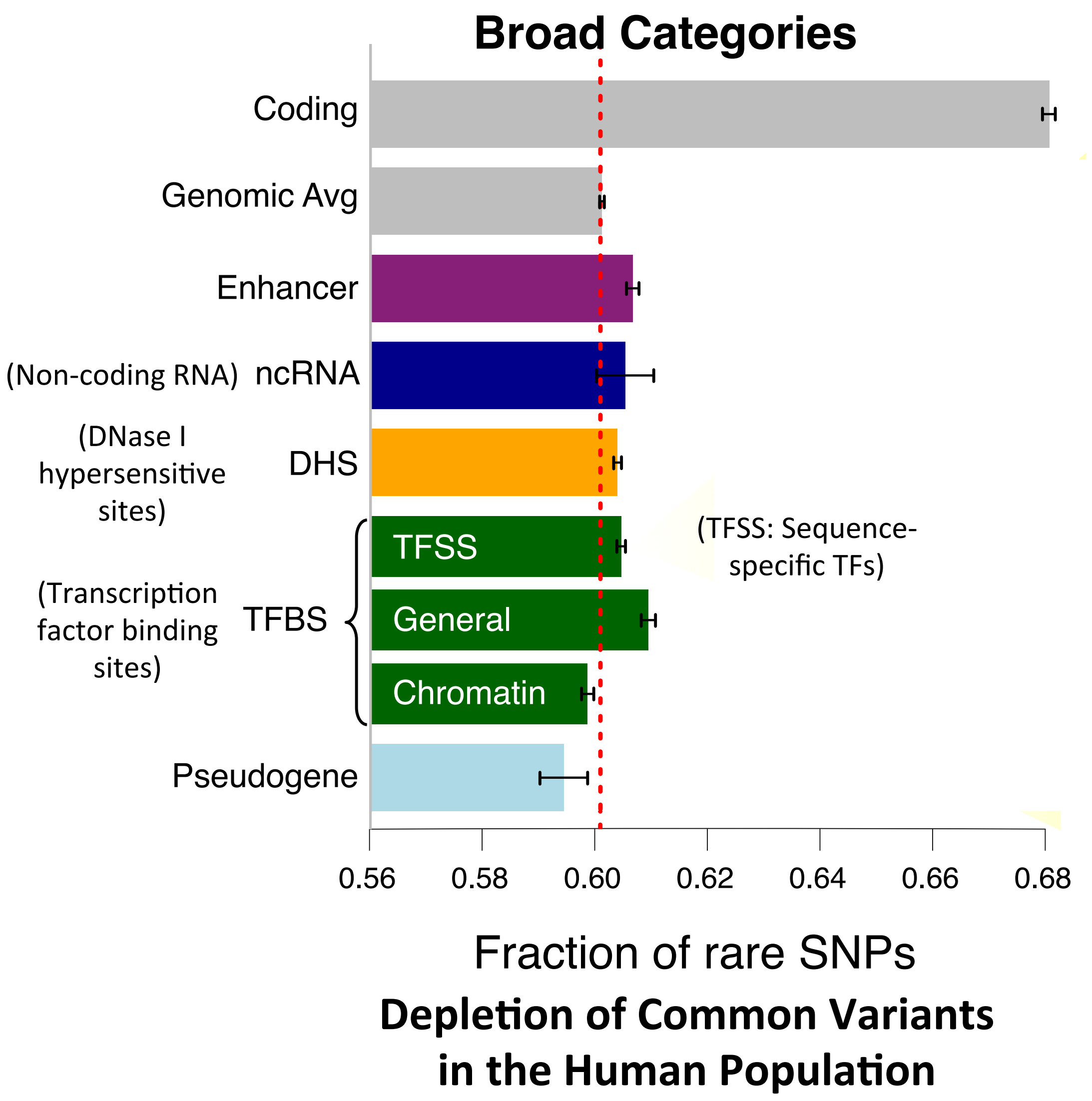
Multiscale Decomposition



[Harmanci *et al, Genome Biol.* ('14)]

Finding "Conserved" Sites in the Human Population:

Negative selection in non-coding elements based on Production ENCODE & 1000G Phase 1



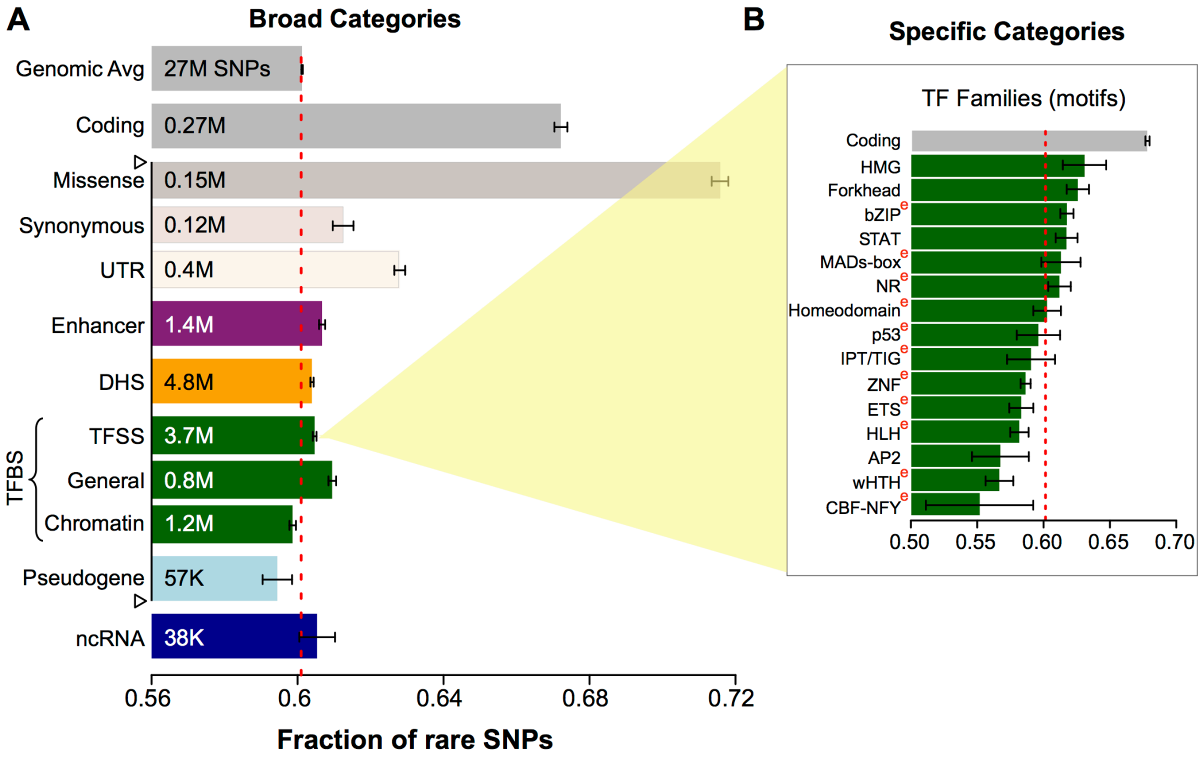
Depletion of Common Variants in the Human Population

- Broad categories of regulatory regions under negative selection
- Related to:

ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

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

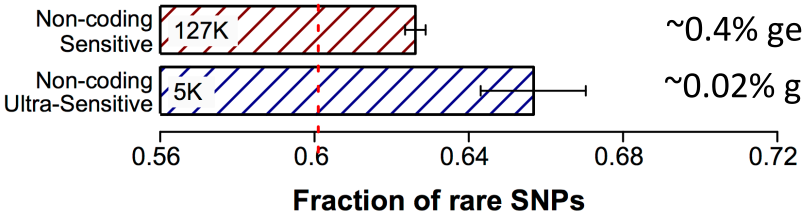
Differential selective constraints among specific sub-categories



Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

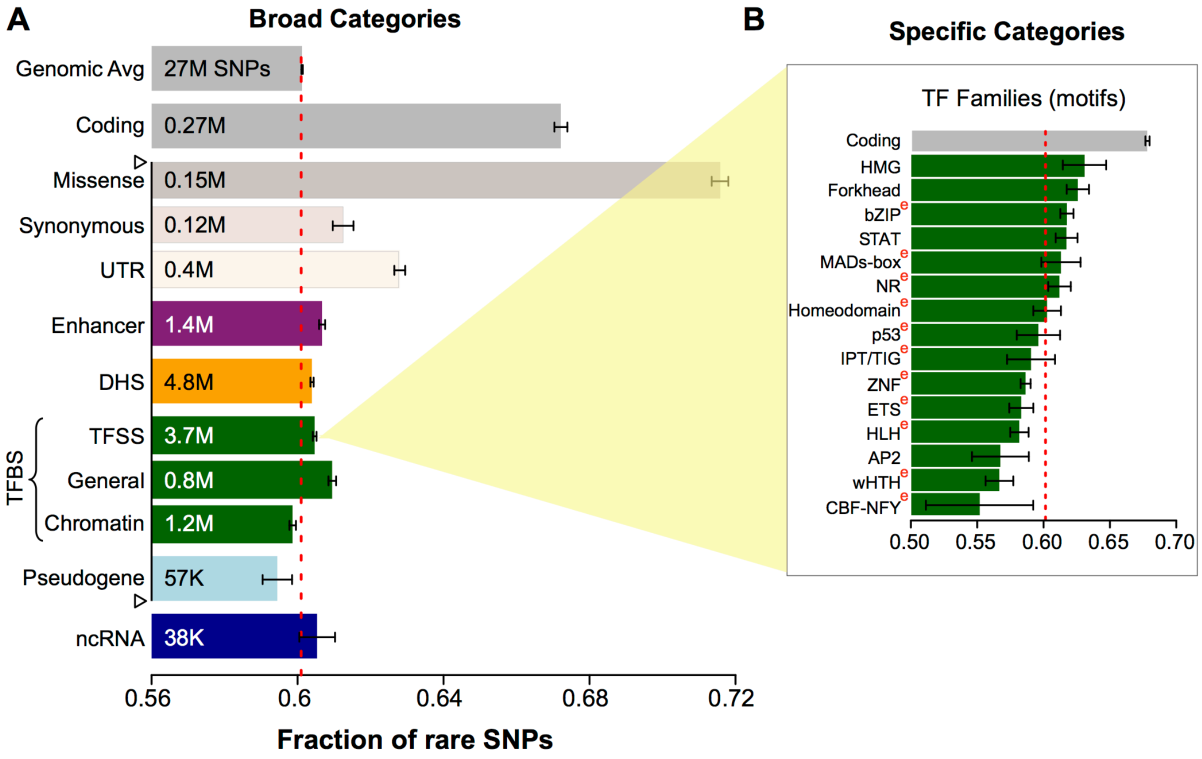
[Khurana et al., *Science* ('13)]

Defining Sensitive non-coding Regions



~0.4% genomic coverage (~ top 25)

~0.02% genomic coverage (top 5)



Start 677 high-resolution non-coding categories; Rank & find those under strongest selection

Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

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

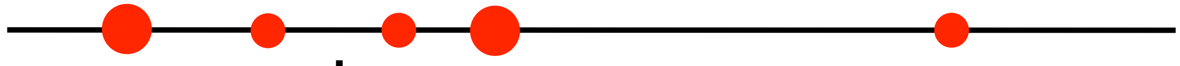
Summer Camp Explorations:
Prioritizing non-coding mutations as potential cancer drivers

- **Finding Non-coding Regions Sensitive to Mutations**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale "site" calling (with Music)
    - Finding small number of sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Building a network from the linear annotation
    - More connectivity = more constraint => highlights hubs
- **Using this to Interpret Alterations in Cancer**
  - **LARVA: to find recurrently mutated annotations**
    - Need to correct for overdispersion in bionomial
    - Use beta-bin parameterized according to replication timing
  - **FunSeq software tool for mutation prioritization**
    - Systematically weighting all the features, for non-coding prioritization

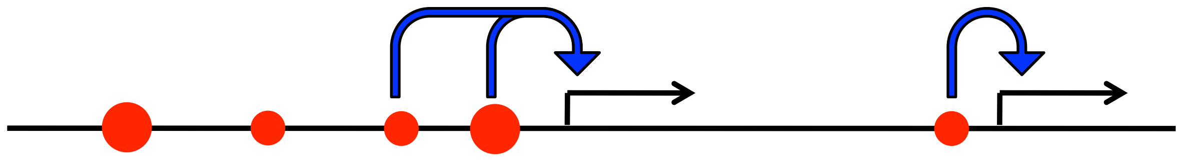
Relating Non-coding Annotation to Protein-coding Genes via Networks

[Cheng et al., *Bioinfo.* ('11), Gerstein et al., *Nature* ('12) , Yip et al., *GenomeBiology* ('12), Fu et al., *GenomeBiology*('14)]

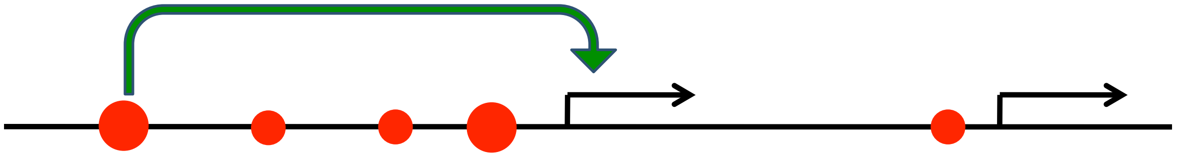
Regulatory elements

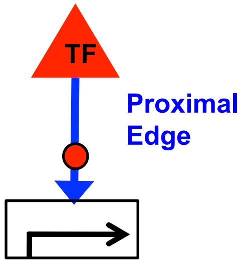


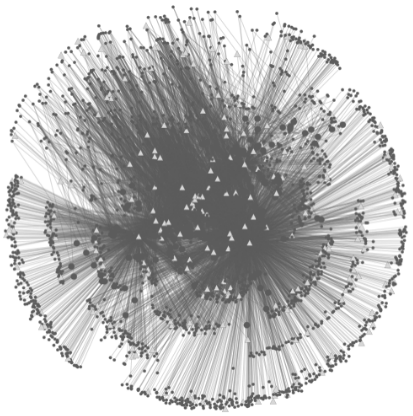
Assigning proximal sites (< 1Kb) to target genes



Assigning distal sites (10Kb-1Mb) to targets



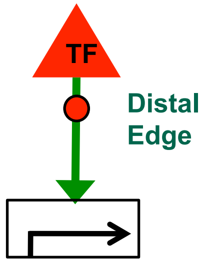




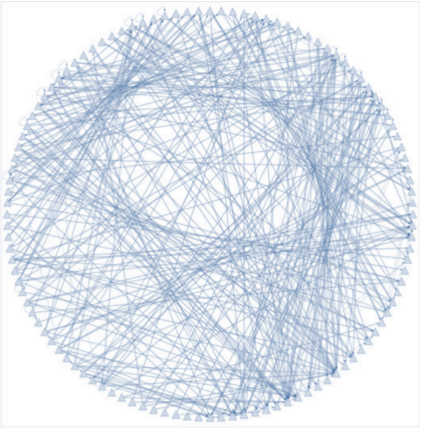
~500K Prox. Edges

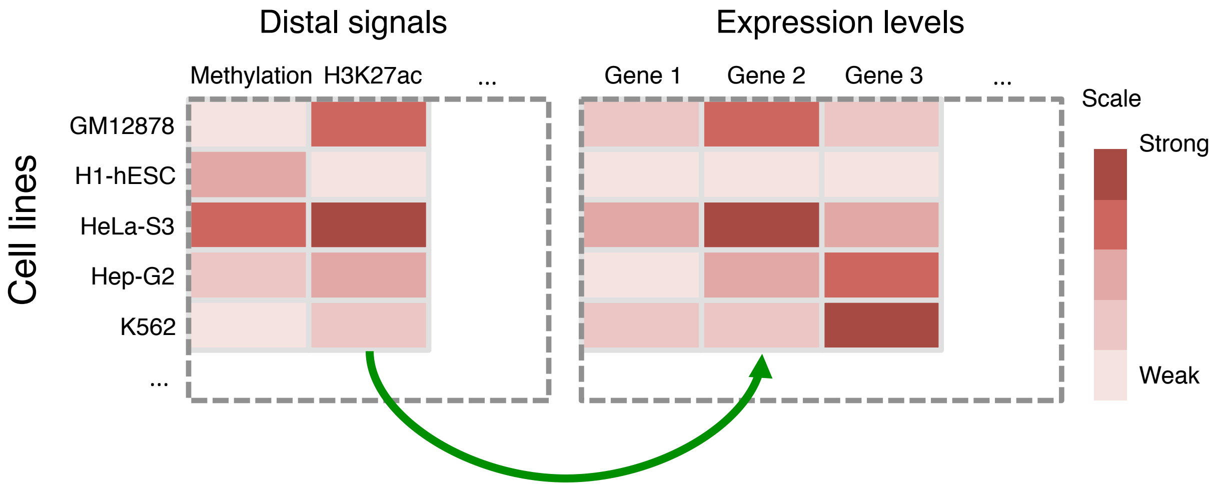
Filtering

~26K



~700K Edges



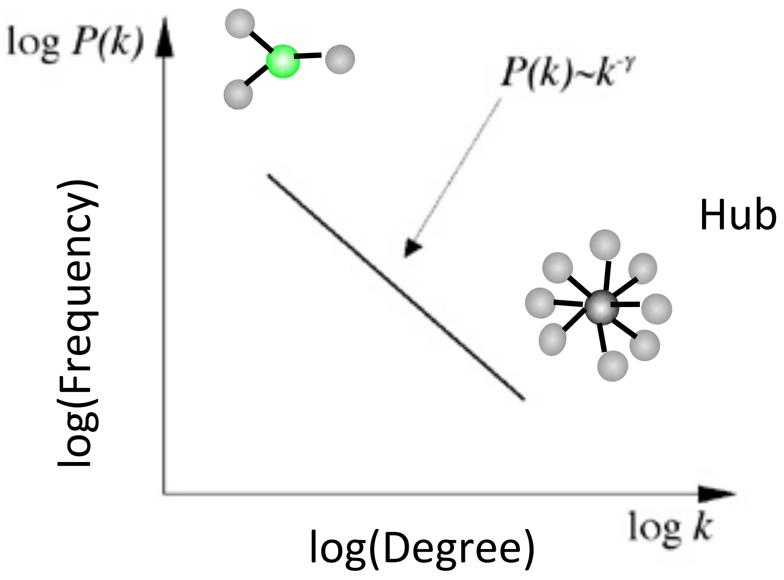


Connecting Distal Elements via **Activity Correlations**.

Other strategies to create linkage incl. eQTL and Hi-C**.** Much in recent Epigenomics Roadmap.

Hubs Under Constraint: A Finding from the Network Biology Community

Power-law distribution



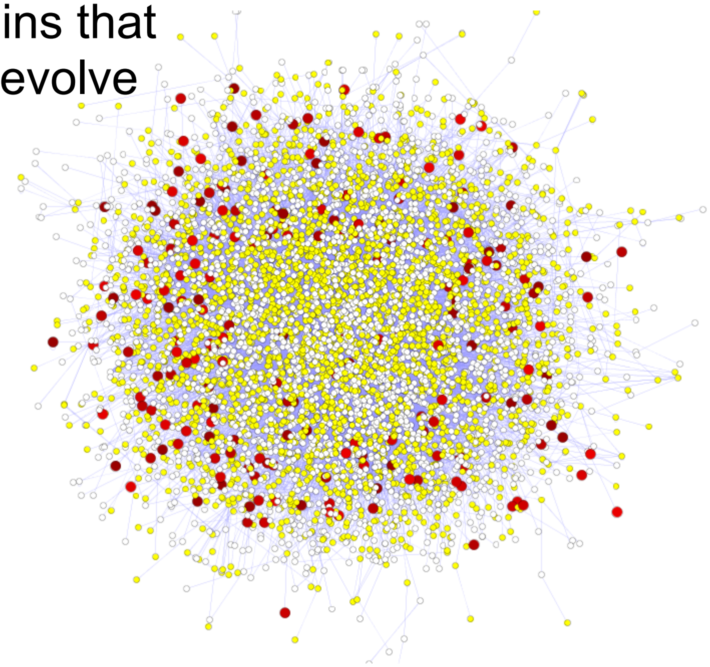
● High likelihood of positive selection

● Lower likelihood of positive selection

● Not under positive selection

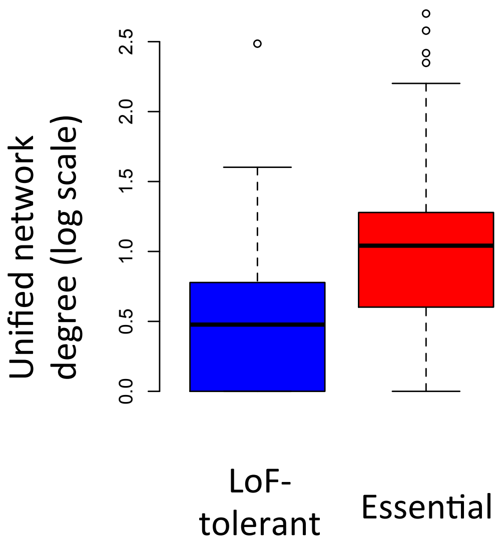
○ No data about positive selection

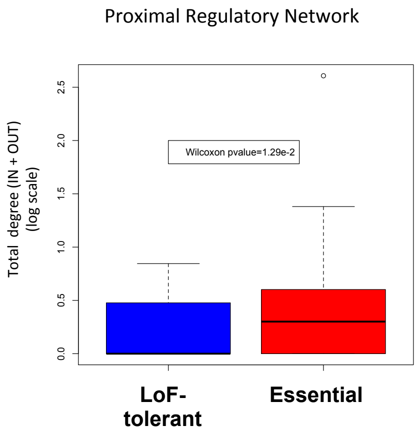
- More Connectivity, More Constraint: Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
  - **yeast PPI** - Fraser et al ('02) Science, ('03) BMC Evo. Bio.
  - **Ecoli PPI** - Butland et al ('04) Nature
  - **Worm/fly PPI** - Hahn et al ('05) MBE
  - **miRNA net** - Cheng et al ('09) BMC Genomics

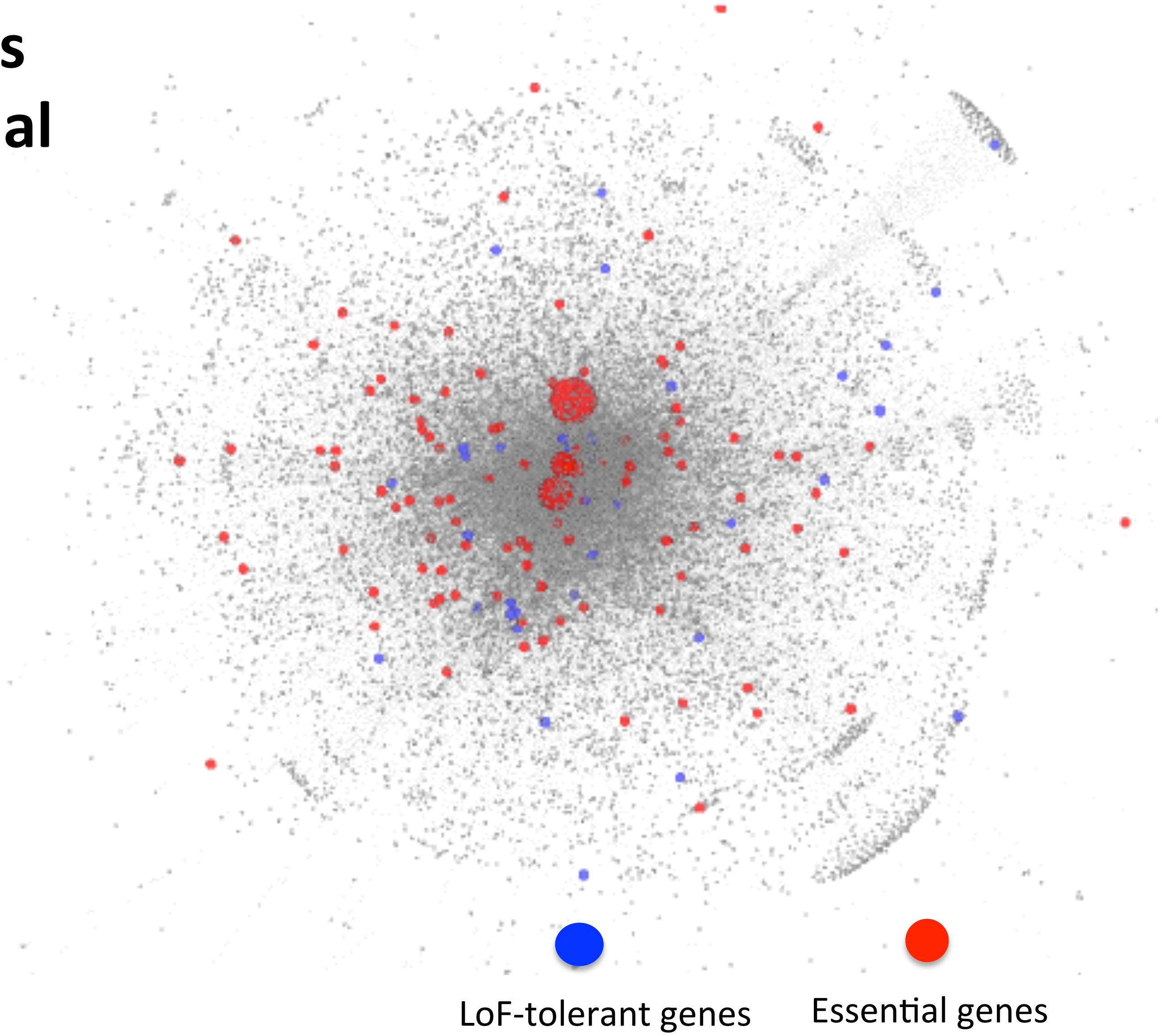


[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

Regulatory Hubs are more Essential







Size of nodes scaled by total degree

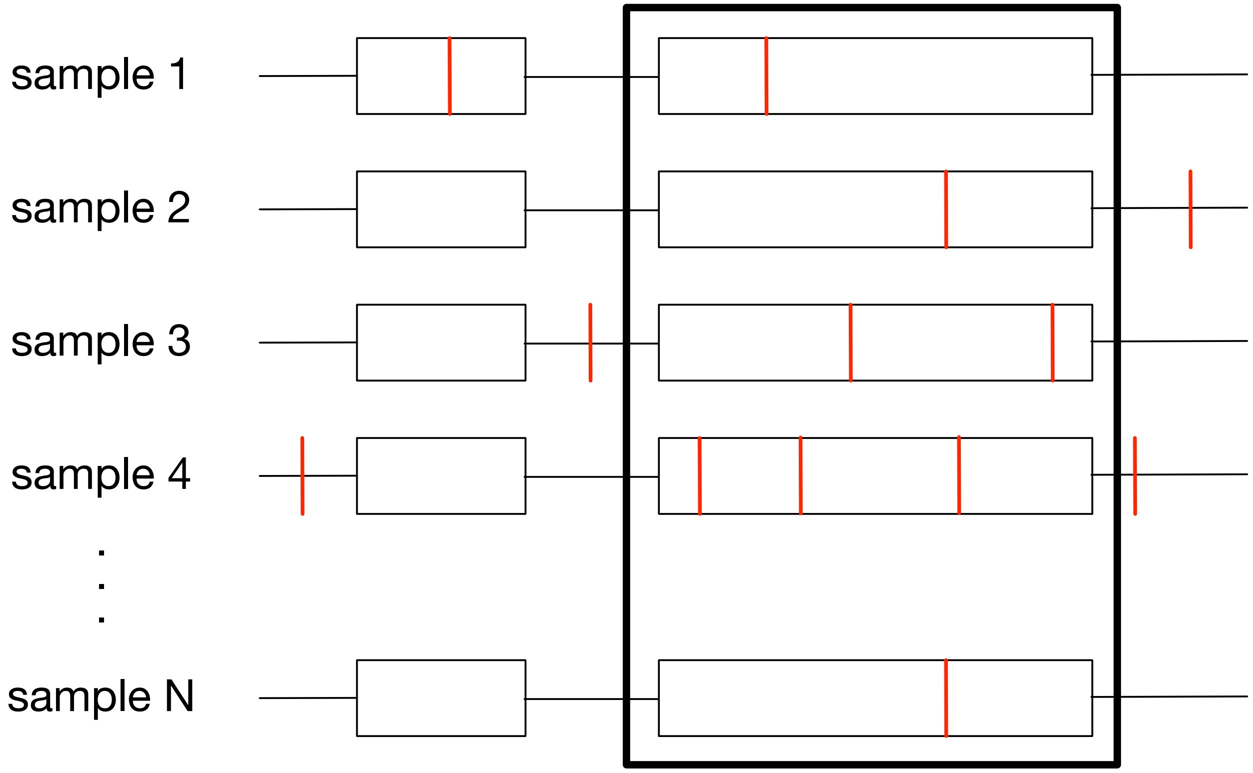
[Khurana et al., *PLOS Comp. Bio.* ’13]

Summer Camp Explorations:
Prioritizing non-coding mutations as potential cancer drivers

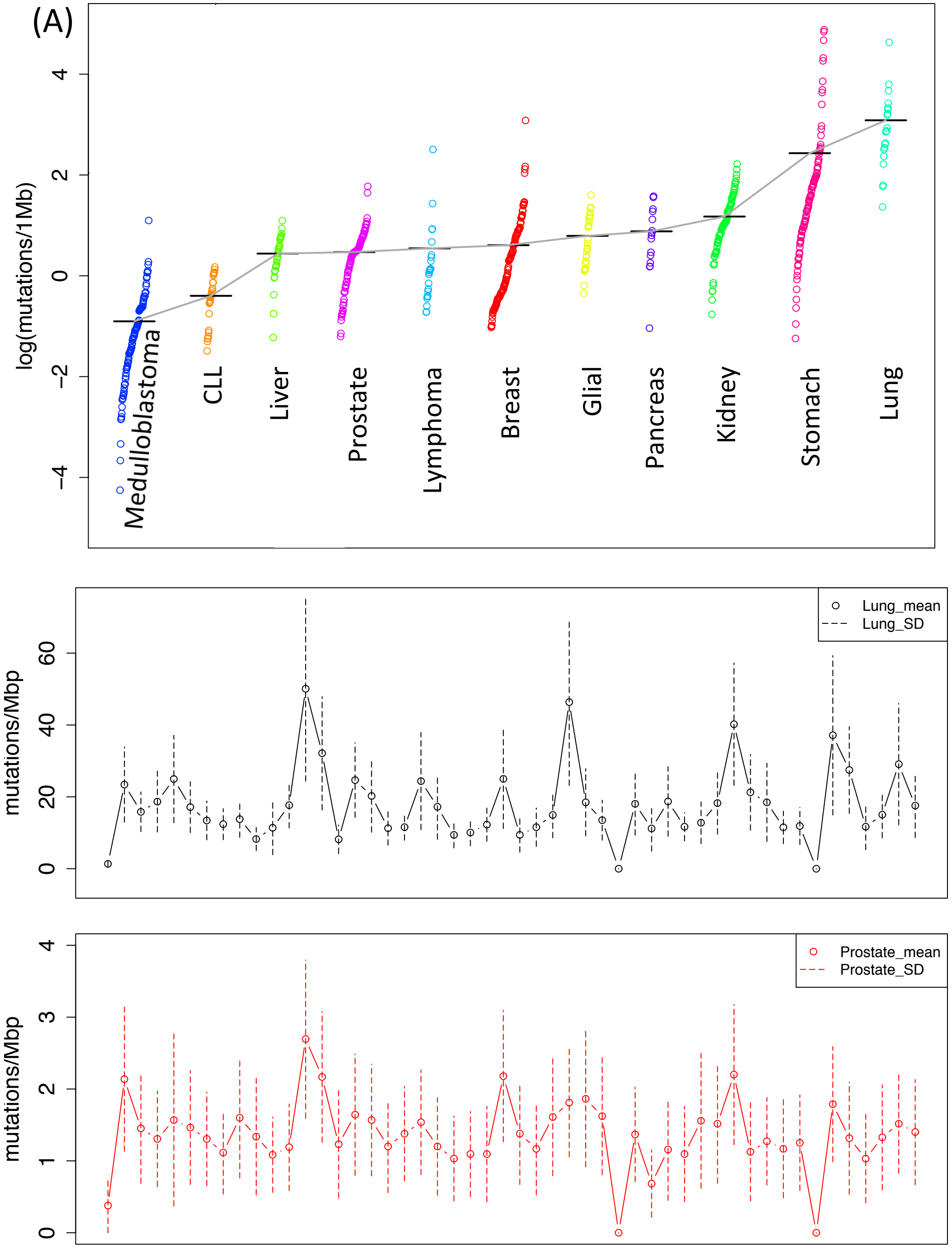
- **Finding Non-coding Regions Sensitive to Mutations**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale "site" calling (with Music)
    - Finding small number of sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Building a network from the linear annotation
    - More connectivity = more constraint => highlights hubs
- **Using this to Interpret Alterations in Cancer**
  - **LARVA: to find recurrently mutated annotations**
    - Need to correct for overdispersion in bionomial
    - Use beta-bin parameterized according to replication timing
  - **FunSeq software tool for mutation prioritization**
    - Systematically weighting all the features, for non-coding prioritization

LARVA

- Somatic single nucleotide variant (SNV) data
  - Which functional noncoding genome elements are hotspots for SNVs across multiple samples?
    - Are they mutated more than expected from neutral mutation processes?



[Lochovsky et al. NAR ('15, in press)]



Cancer Somatic Mutational Heterogeneity

- The distribution of variants throughout the genome indicates high mutation rate heterogeneity between samples of the same cancer type, and on many other levels
- **Goal:** Develop a model for the whole genome background somatic mutation distribution in cancer to identify potential noncoding cancer driving elements
- LARVA: Large-scale Analysis of Recurrent Variants in noncoding Annotations

[Lochovsky et al. NAR ('15, in press)]

Cancer Somatic Mutation Modeling

- We tested 3 models evaluating the significance of a mutation burden of a genome element
- Suppose there are *k* genome elements. For element *i*, define:
  - n_i: total number of nucleotides in *i*
  - x_i: the number of mutation within element *i*
  - *p*: the probability of observing a mutation in each position
  - *R*: The replication timing tenth percentile of *i*

Model 1: Constant Background Mutation Rate (Model from Previous Work[1])

$$x_i : Binomial(n_i, p)$$

Model 2: Varying Mutation Rate

$$x_i | p : Binomial(n_i, p)$$

$$p : Beta(\mu, \sigma)$$

Model 3: Varying Mutation Rate with Replication Timing Correction

$$x_i | p : Binomial(n_i, p)$$

$$p : Beta(\mu | R, \sigma | R)$$

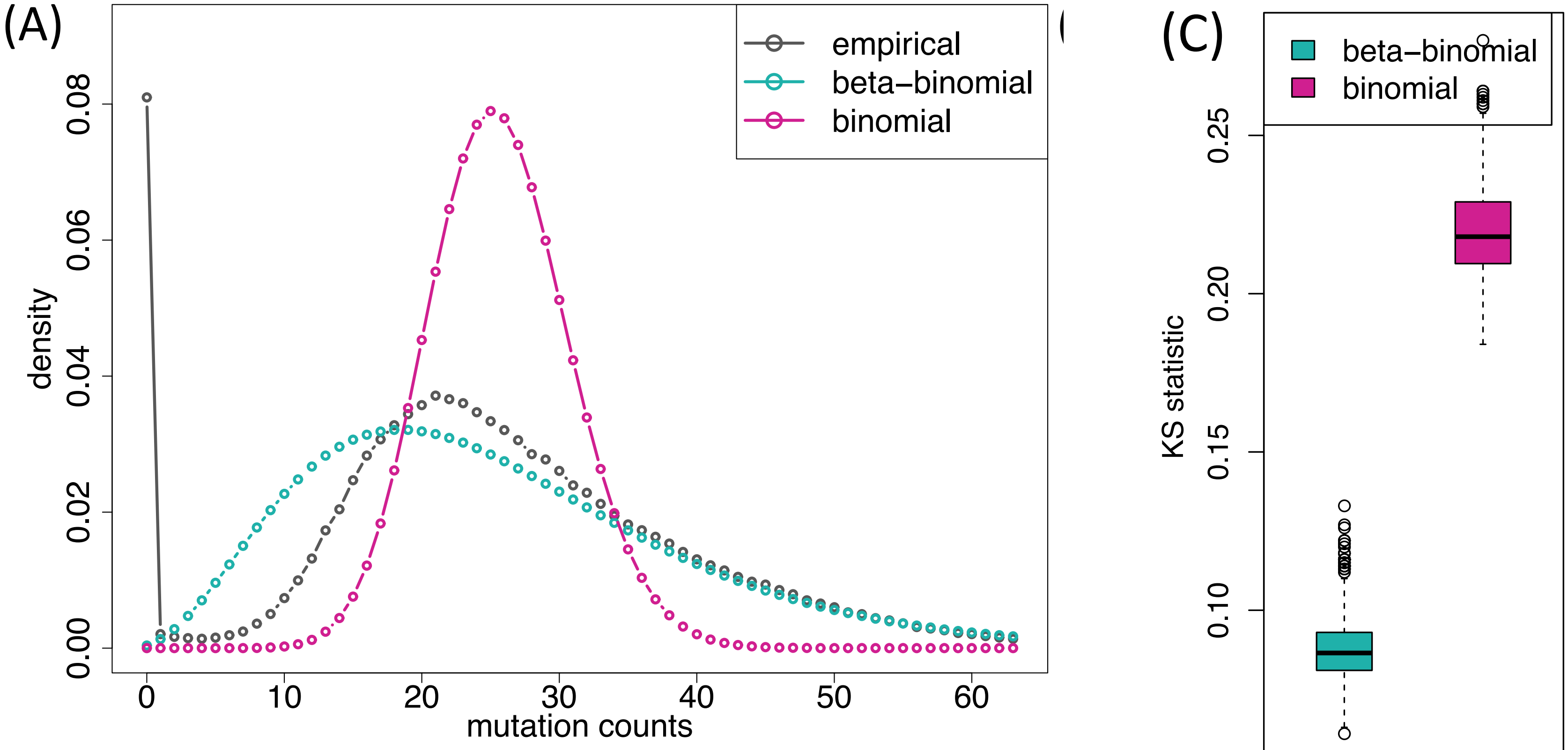
$$\mu | R, \sigma | R : \text{constant within the same } R \text{ bin}$$

[Lochovsky et al. NAR ('15, in press)]

[1] Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics* **46,** 1160–1165 (2014).

LARVA Model Comparison

- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution

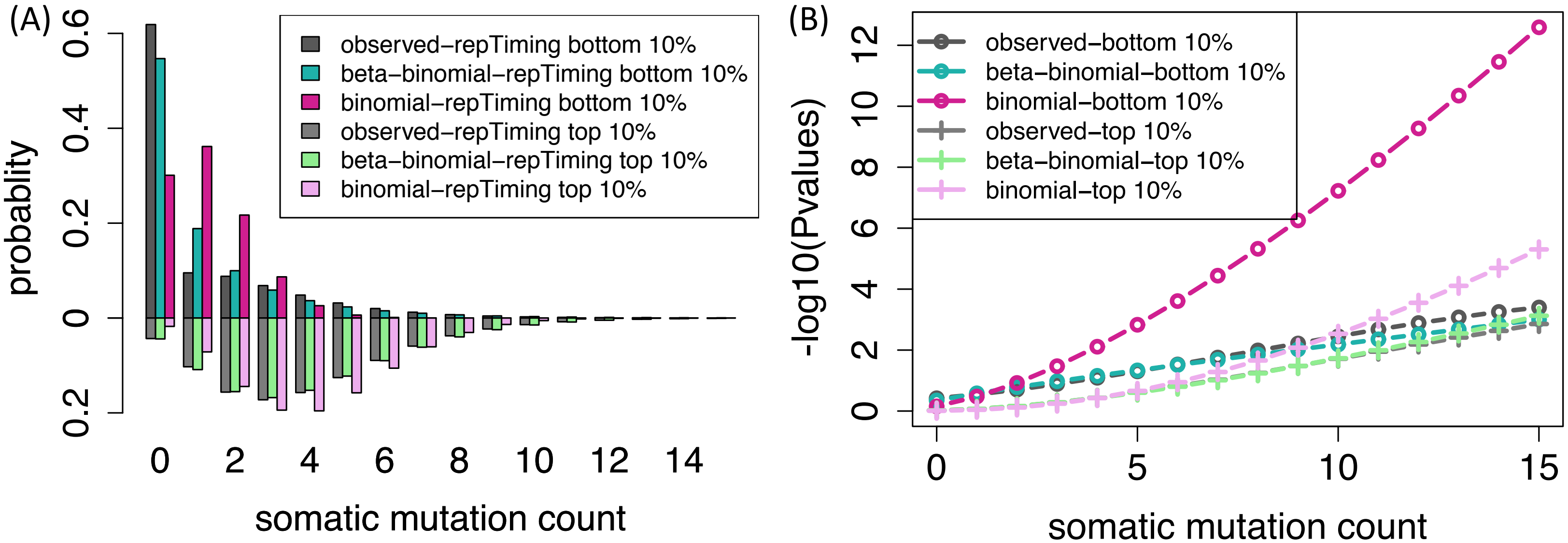


[Lochovsky et al. NAR ('15, in press)]

LARVA Model Comparison

- Demonstrate that adding the DNA replication timing correction (model 3) further improves the beta-binomial model (model 2)
- Top 10% of replication timing bins requires little correction
- Bottom 10% of replication timing bins requires massive correction

- Demonstrate that the number of significant p-values is inflated under the binomial model
- Neither the empirical or beta-binomial models exhibit this inflation

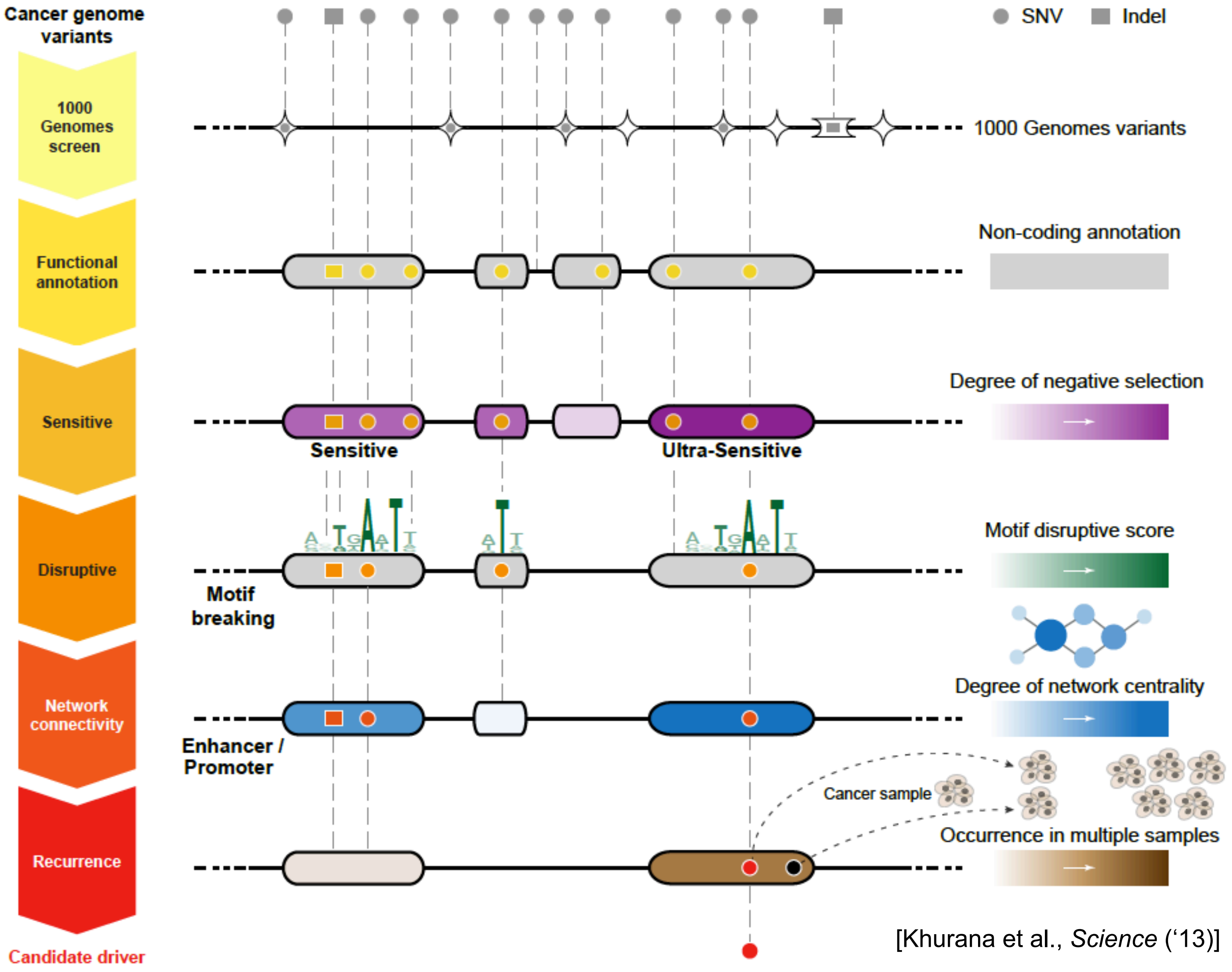


[Lochovsky et al. NAR ('15, in press)]

Summer Camp Explorations:
Prioritizing non-coding mutations as potential cancer drivers

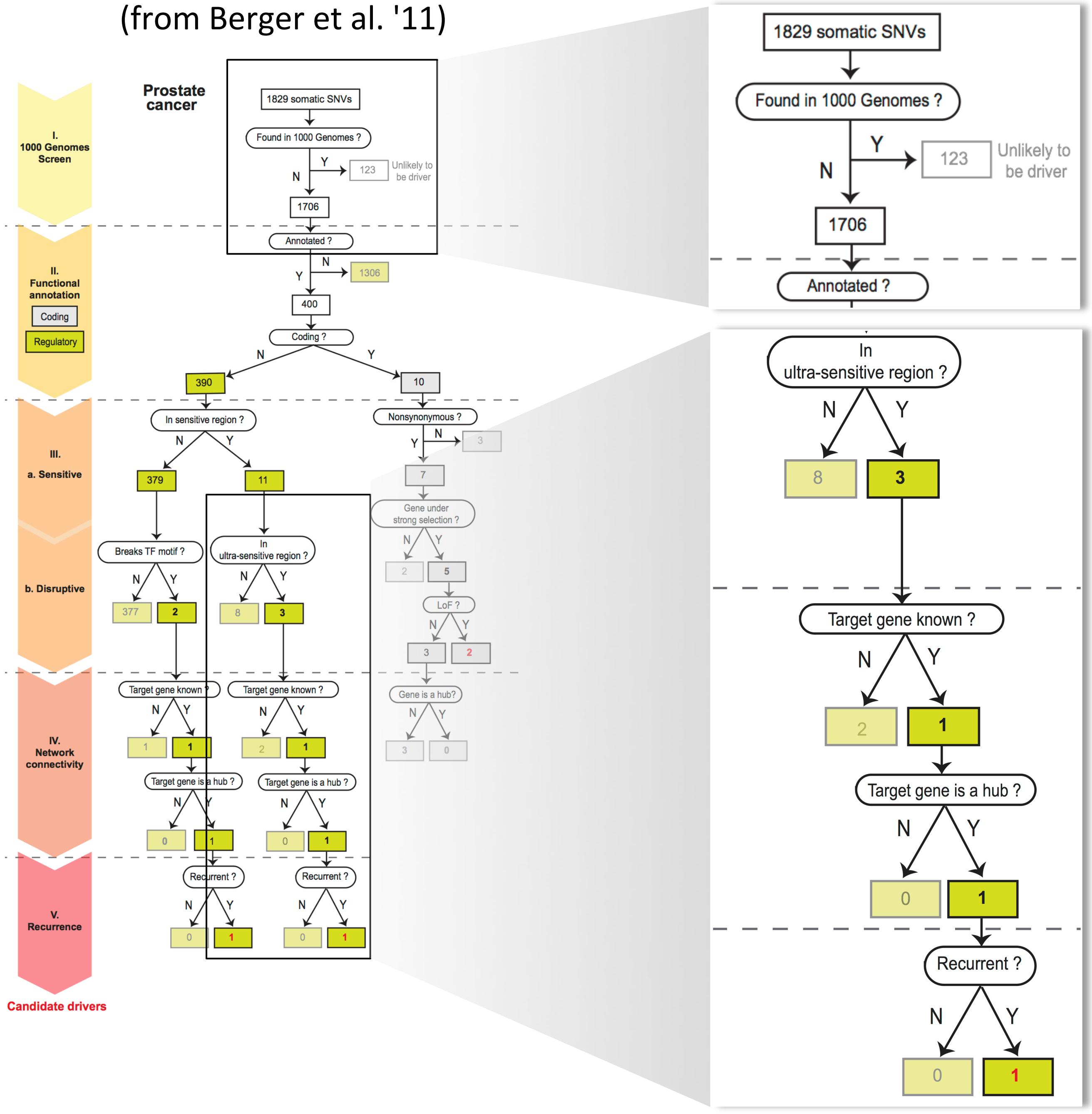
- **Finding Non-coding Regions Sensitive to Mutations**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale "site" calling (with Music)
    - Finding small number of sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Building a network from the linear annotation
    - More connectivity = more constraint => highlights hubs
- **Using this to Interpret Alterations in Cancer**
  - **LARVA: to find recurrently mutated annotations**
    - Need to correct for overdispersion in bionomial
    - Use beta-bin parameterized according to replication timing
  - **FunSeq software tool for mutation prioritization**
    - Systematically weighting all the features, for non-coding prioritization

Identification of non-coding candidate drivers amongst somatic variants: Scheme

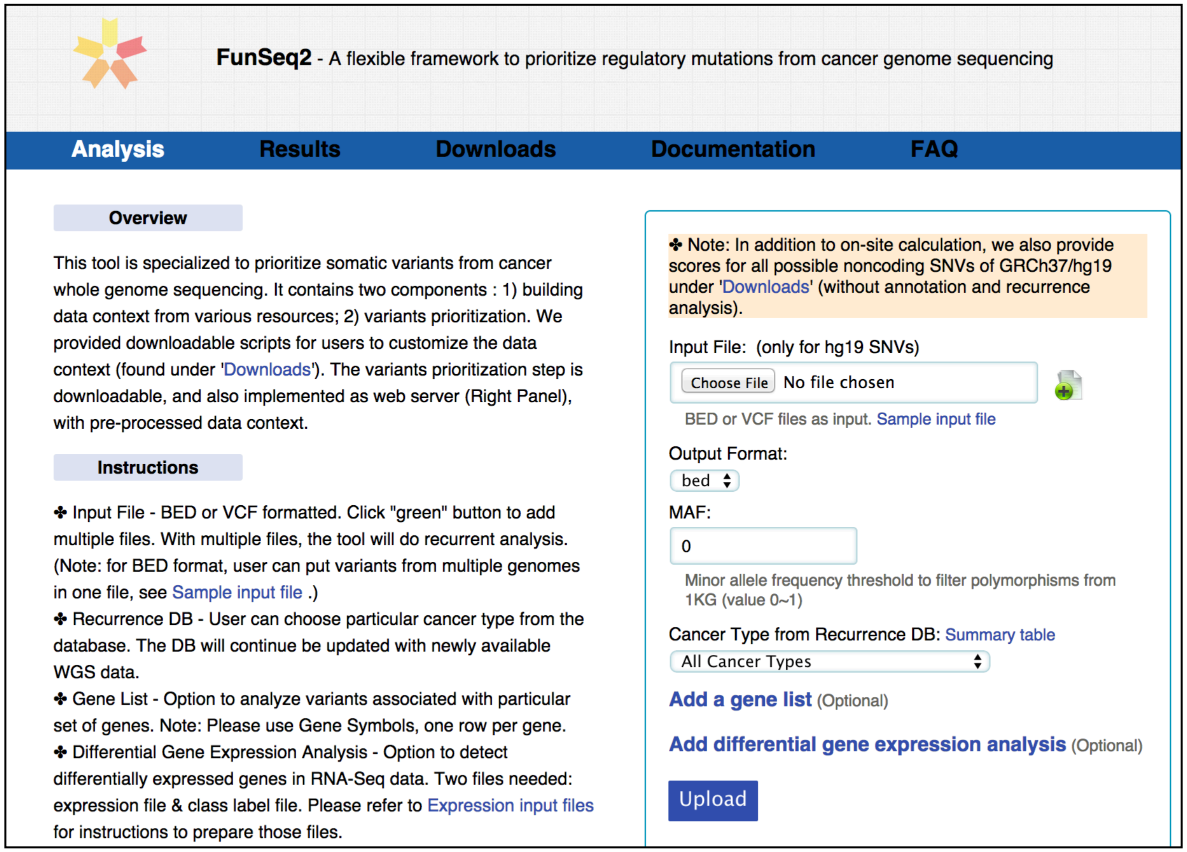


[Khurana et al., *Science* ('13)]

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



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



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 those 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

BED or VCF files as input. Sample input file

Output Format:

bed

MAF:

0

Minor allele frequency threshold to filter polymorphisms from 1KG (value 0~1)

Cancer Type from Recurrence DB: Summary table

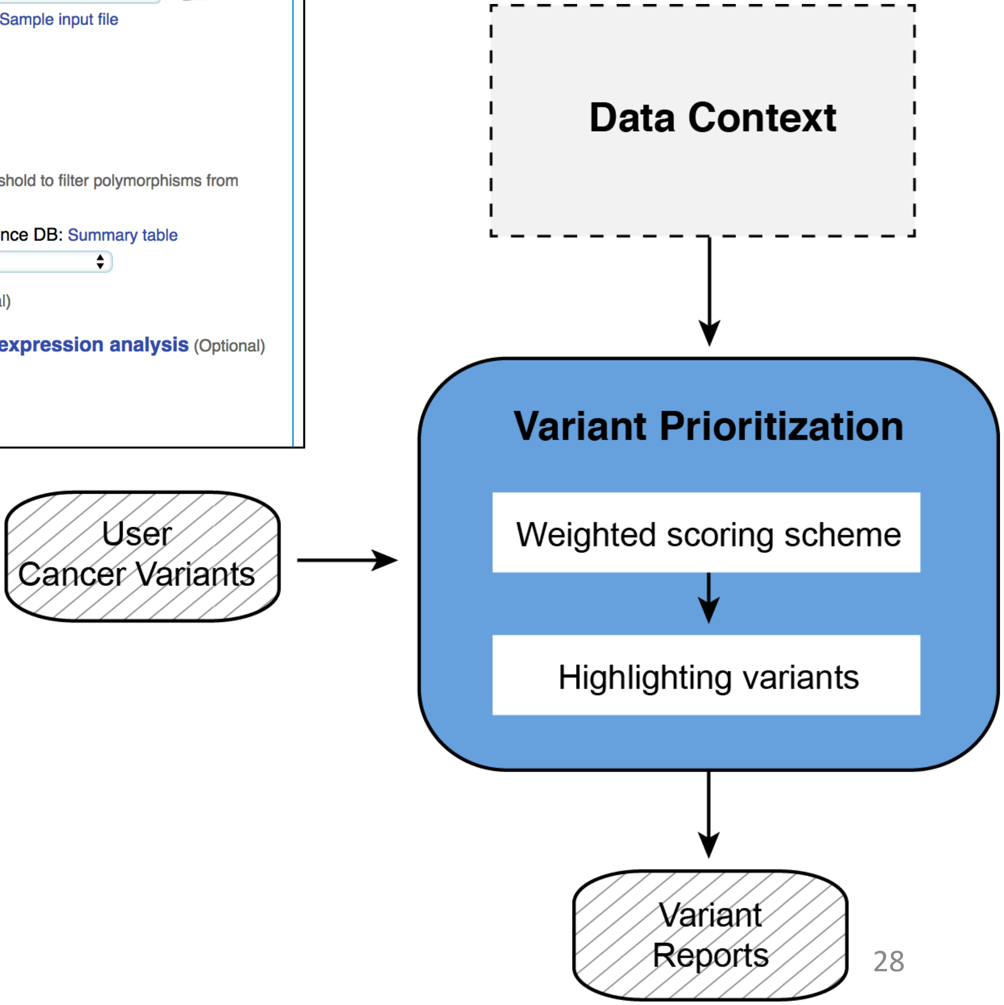
All Cancer Types

Add a gene list (Optional)

Add differential gene expression analysis (Optional)

Upload

Site integrates user variants with large-scale context



Data Context

User Cancer Variants

Variant Prioritization

Weighted scoring scheme

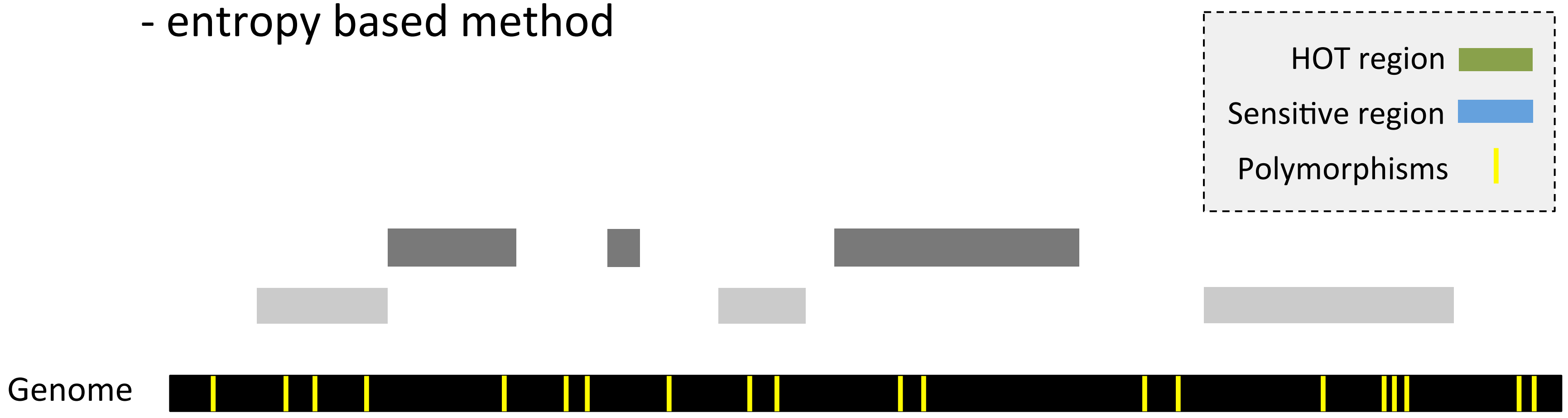
Highlighting variants

Variant Reports

FunSeq.gersteinlab.org

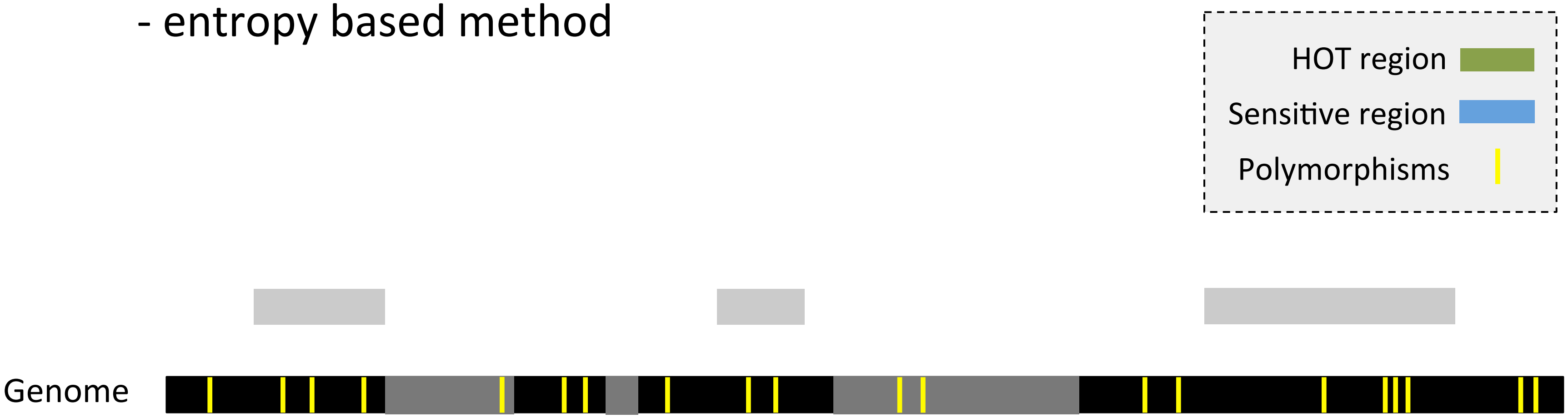
[Fu et al., GenomeBiology ('14)]

- Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weight less)
  - entropy based method



[Fu et al., GenomeBiology ('14)]

- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method

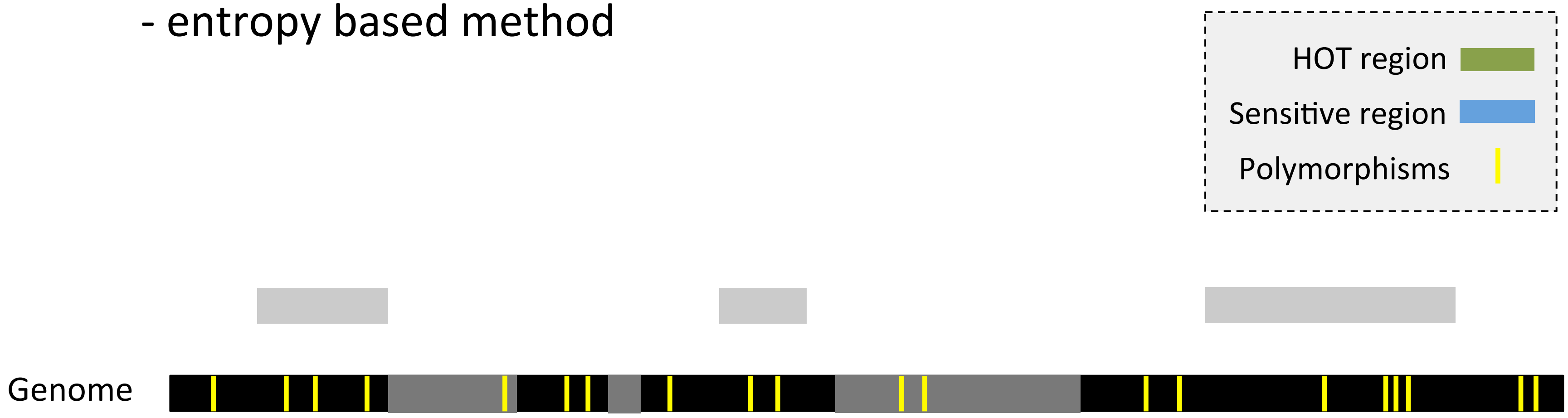


$$p = \frac{3}{20}$$

[Fu et al., GenomeBiology ('14)]

- Feature weight

  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



$$p = \frac{3}{20}$$

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

$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

For a variant: $Score = \sum w_d \ \ of\ observed\ features$

[Fu et al., GenomeBiology ('14)]

Summer Camp Explorations:
Prioritizing non-coding mutations as potential cancer drivers

- **Finding Non-coding Regions Sensitive to Mutations**
  - **1st Level Linear Annotation: Regulatory Sites**
    - Multi-scale "site" calling (with Music)
    - Finding small number of sites particularly sensitive to mutations
  - **2nd Level Network Annotation**
    - Building a network from the linear annotation
    - More connectivity = more constraint => highlights hubs
- **Using this to Interpret Alterations in Cancer**
  - **LARVA: to find recurrently mutated annotations**
    - Need to correct for overdispersion in bionomial
    - Use beta-bin parameterized according to replication timing
  - **FunSeq software tool for mutation prioritization**
    - Systematically weighting all the features, for non-coding prioritization

Cancer Prioritzation Acknowledgements

← ~50 people ← ~1000 "authors"

Functional Interpretation Subgroup

A THOUSAND GENOMES

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

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
- **FunSeq2**.gersteinlab.org
  - Y **Fu**, Z Liu, S Lou, J Bedford, X Mu, K Yip, E Khurana
- **LARVA**.gersteinlab.org
  - L **Lochovsky***, J **Zhang***, Y Fu, E Khurana

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2014.
  - Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html .
  - 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