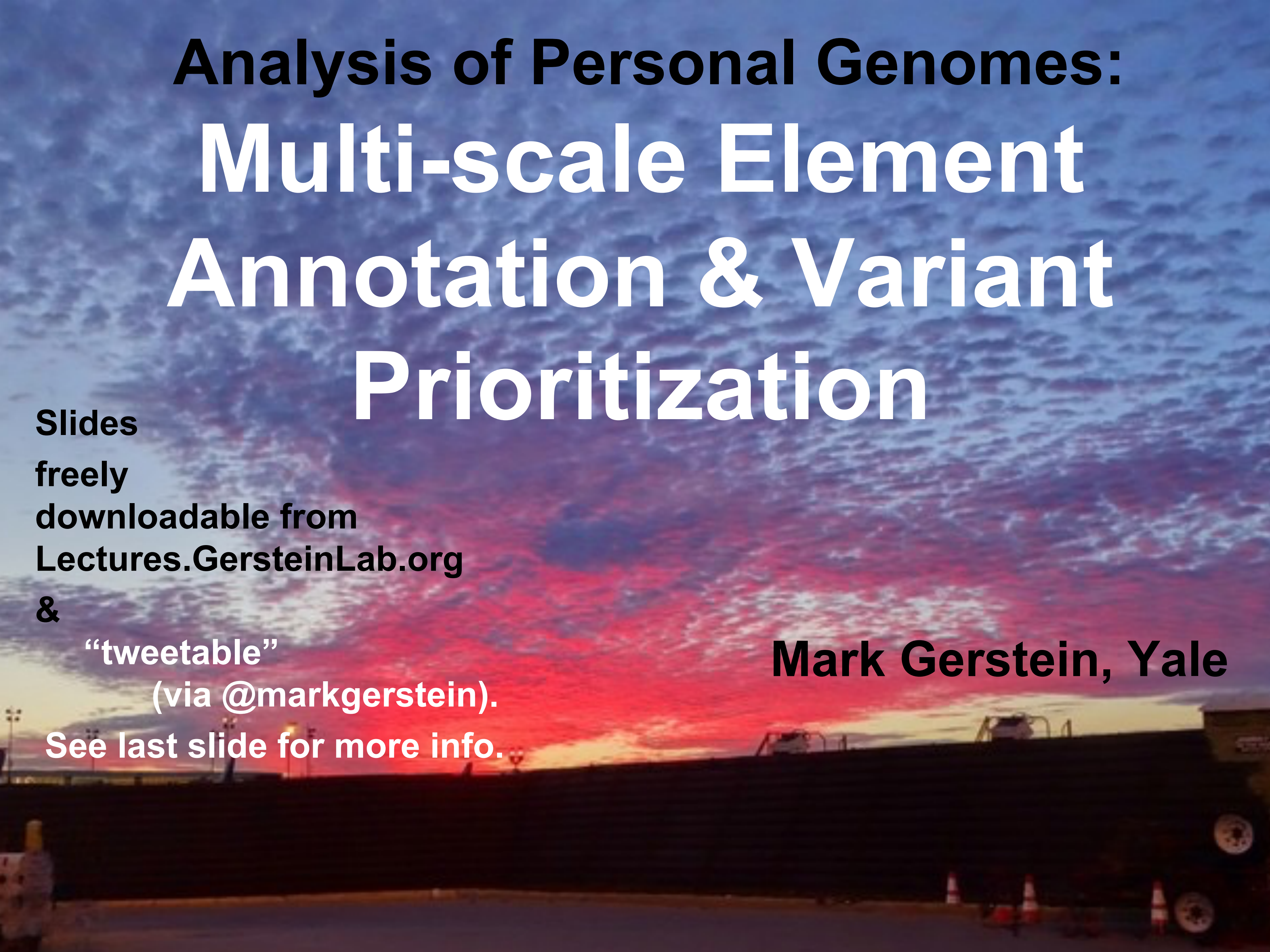


Analysis of Personal Genomes:

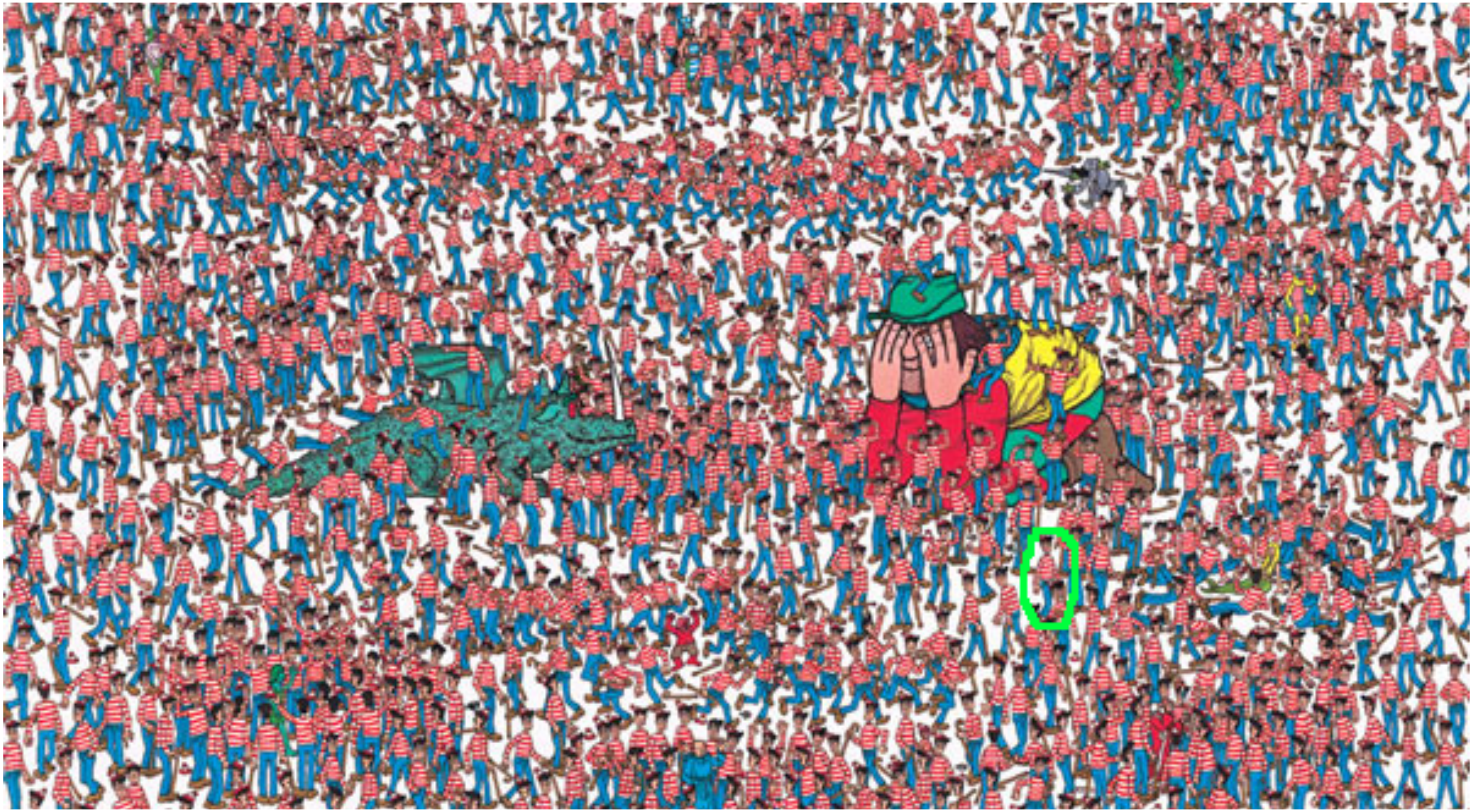
Multi-scale Element Annotation & Variant Prioritization

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

Mark Gerstein, Yale

Where is Waldo?

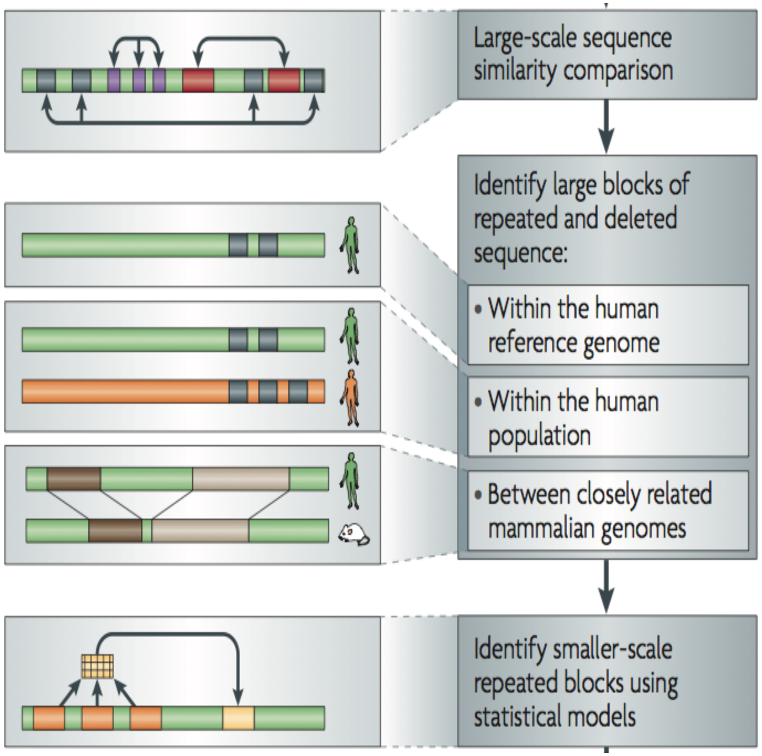
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



Non-coding Annotations: Overview

Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**



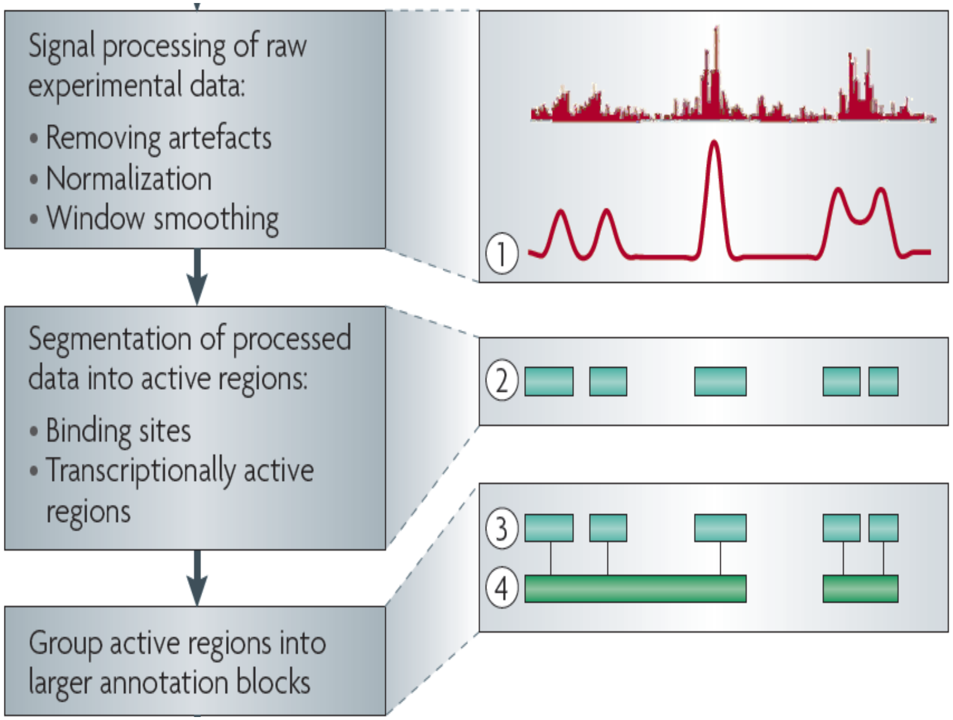
Large-scale sequence similarity comparison

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

Multi-scale Element Annotation & Variant Prioritization

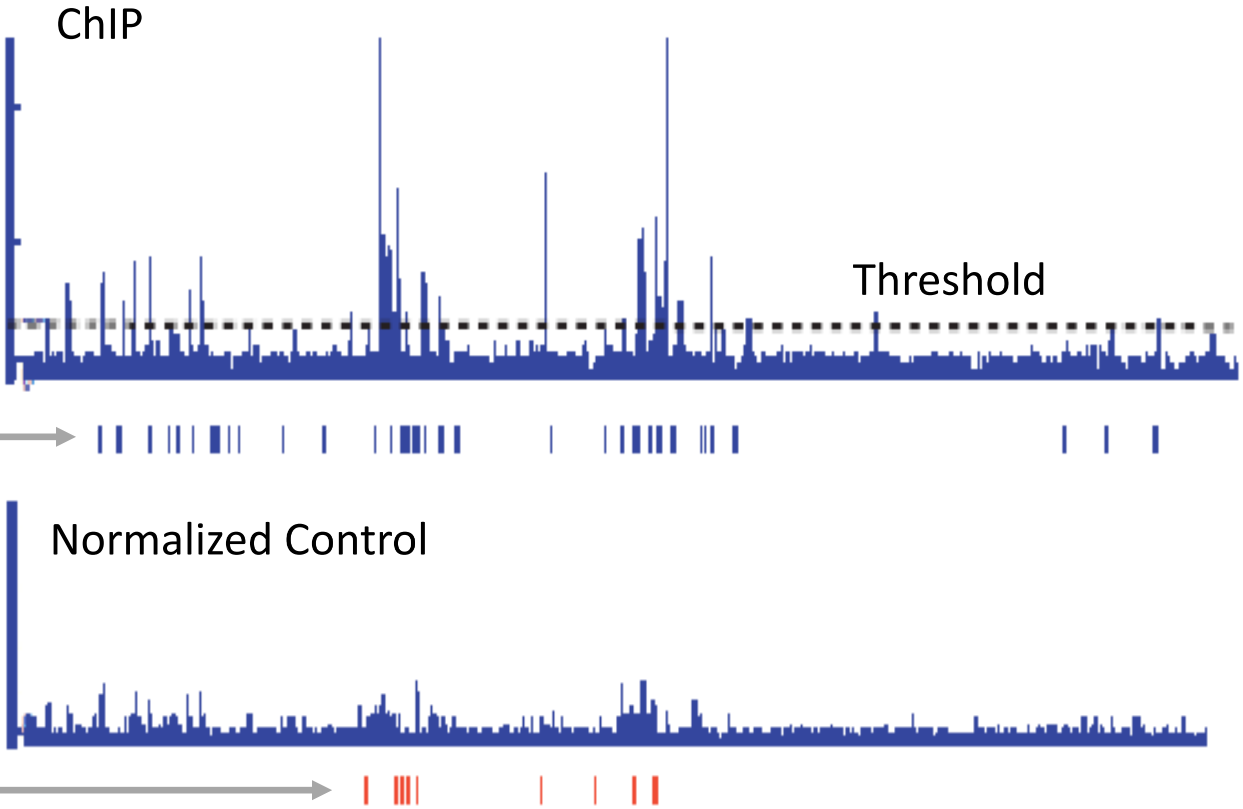
- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization

Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant 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

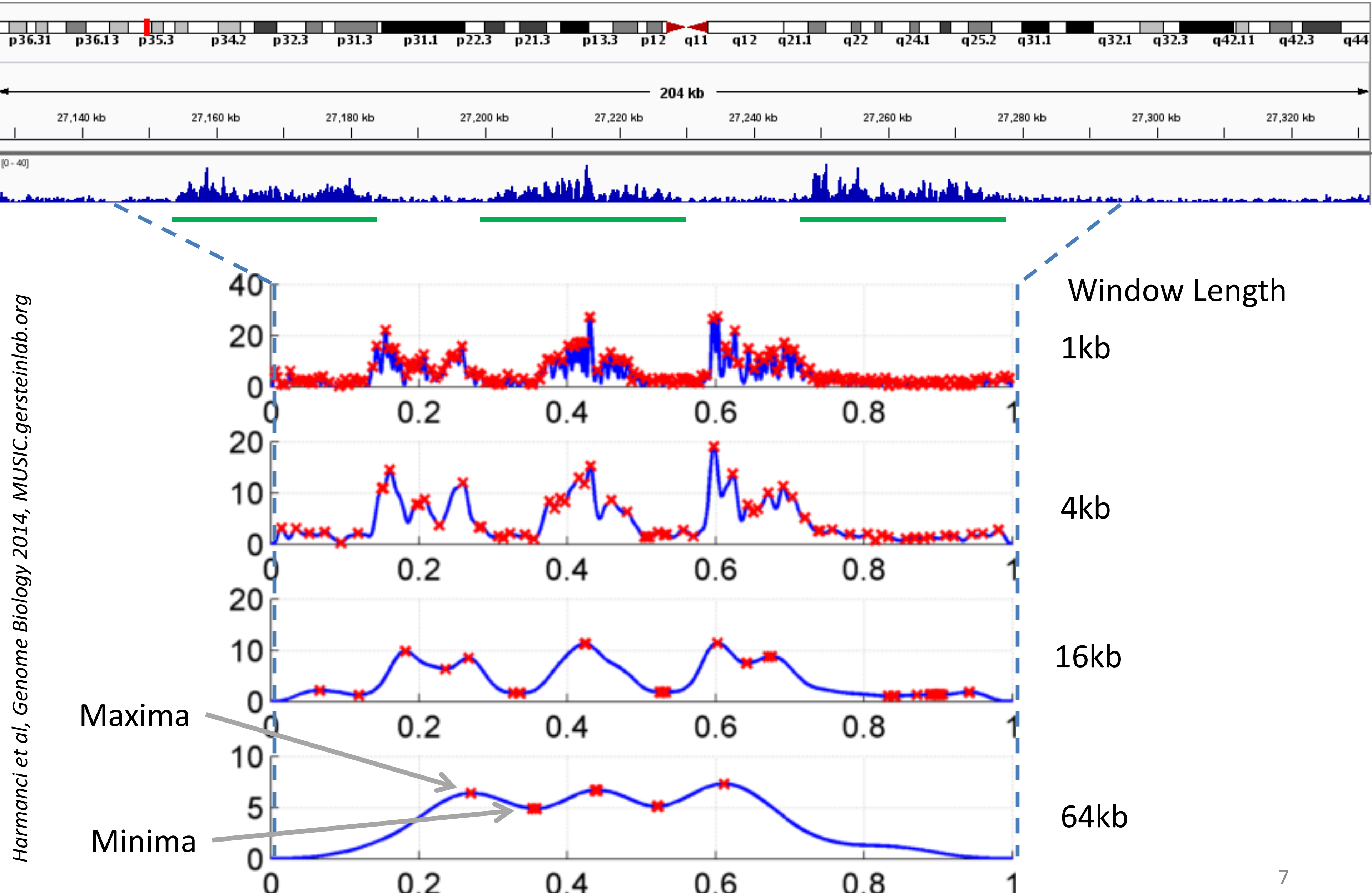
- Score against the control

Significantly Enriched targets

Now an update: "PeakSeq 2" => MUSIC

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

Multiscale Analysis, Minima/Maxima based Coarse Segmentation



Window Length

1kb

4kb

16kb

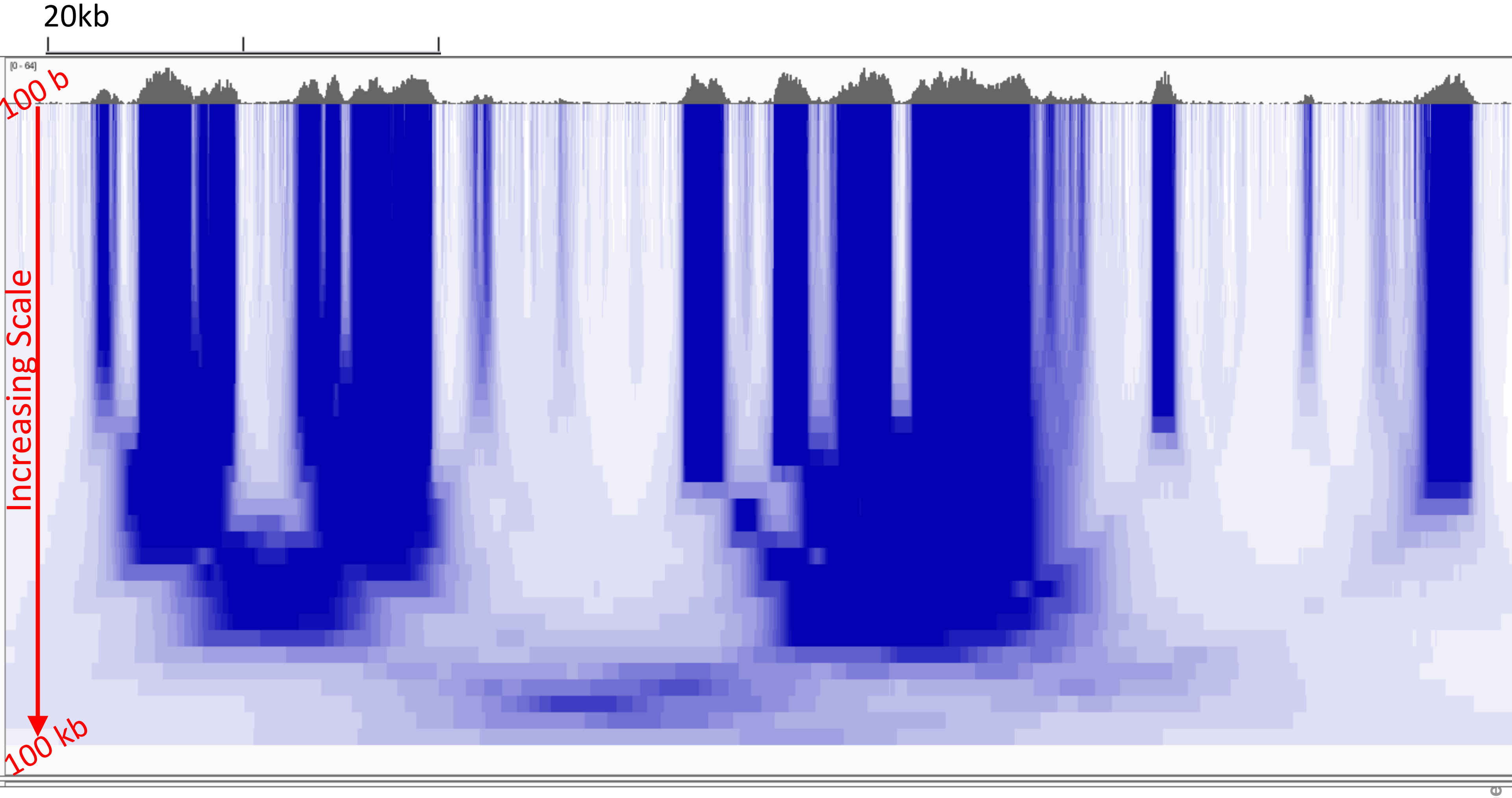
64kb

Maxima

Minima

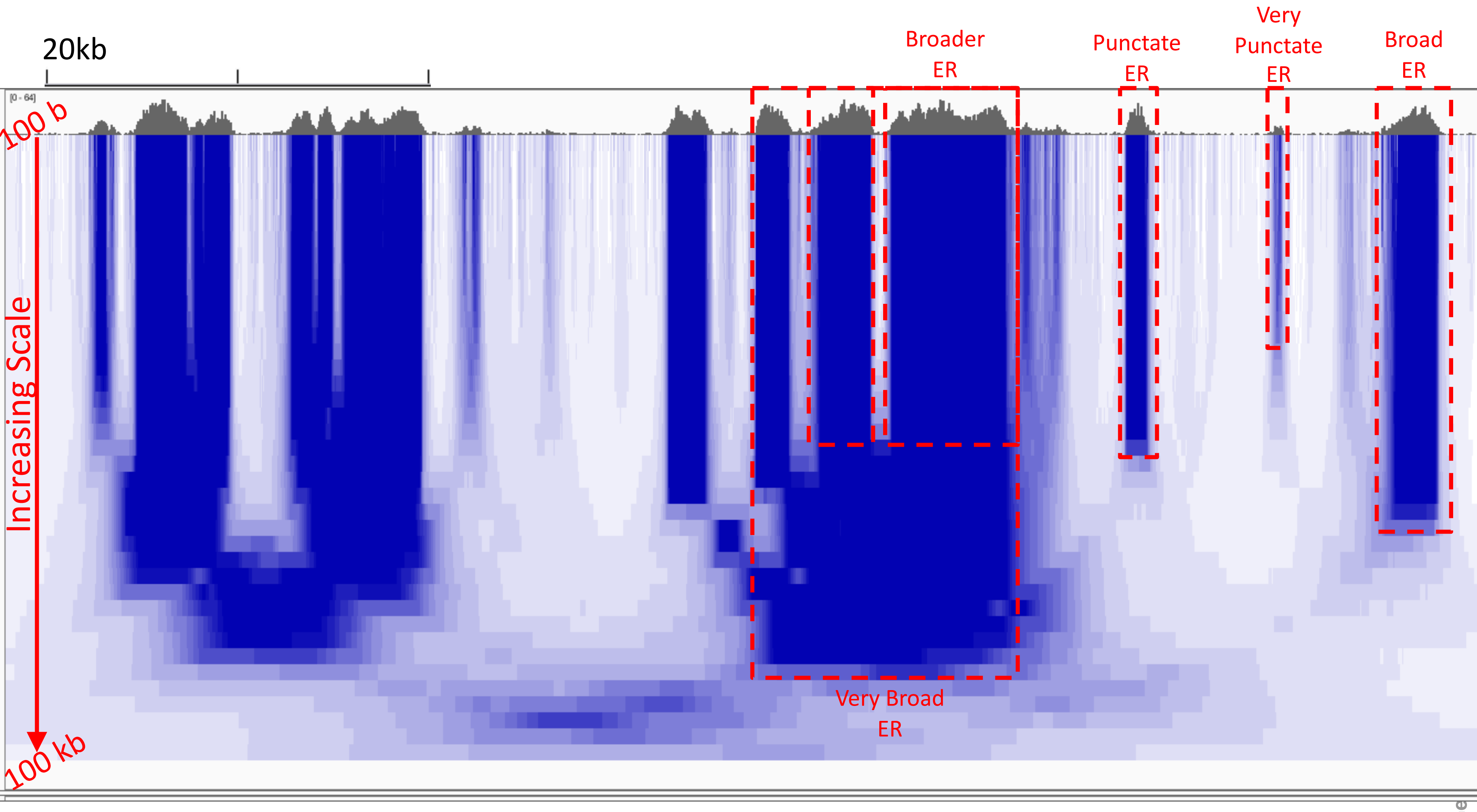
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



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

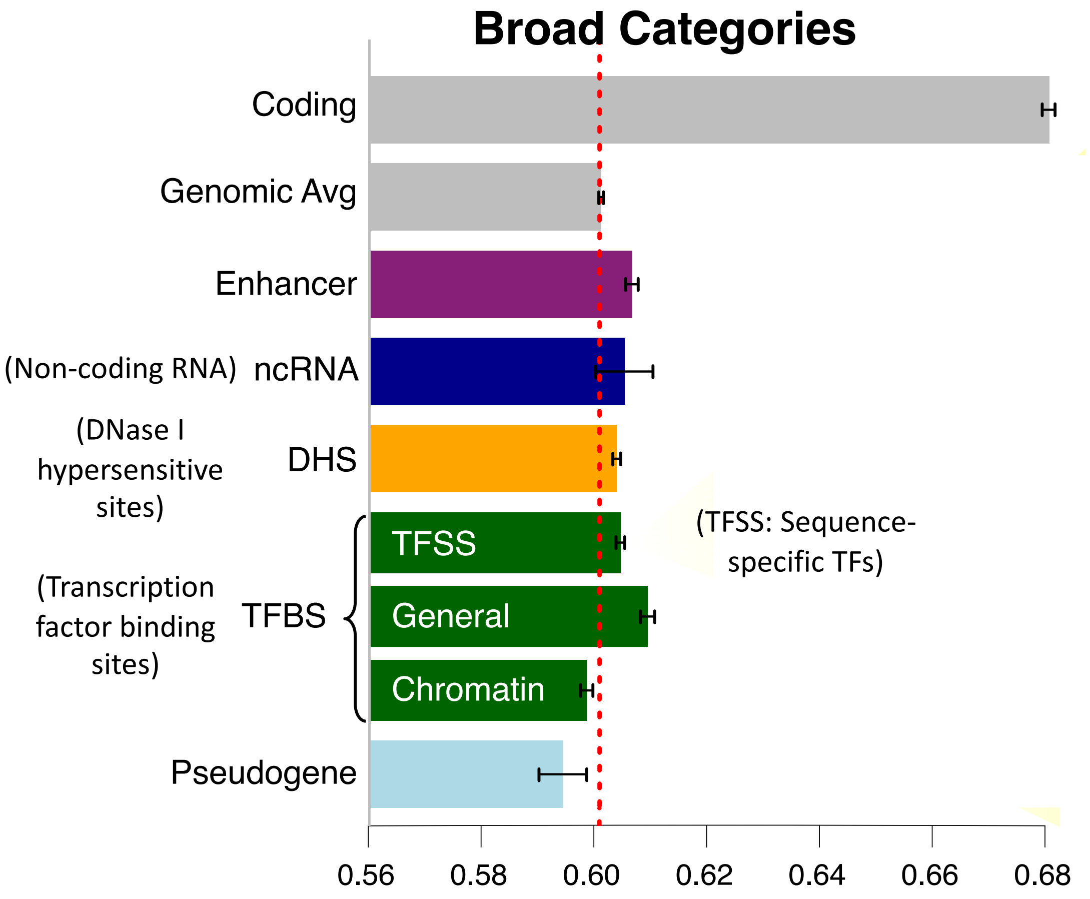
Multiscale Decomposition



[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



Fraction of rare SNPs

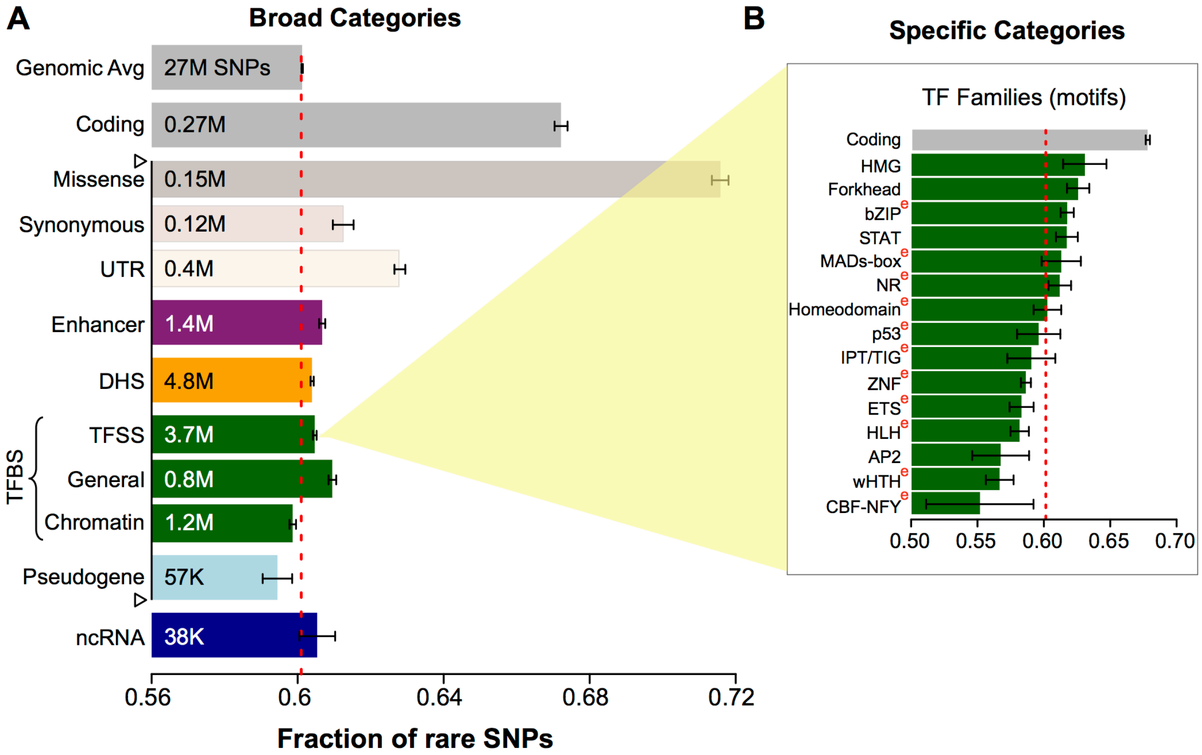
Depletion of Common Variants in the Human Population

- 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)]

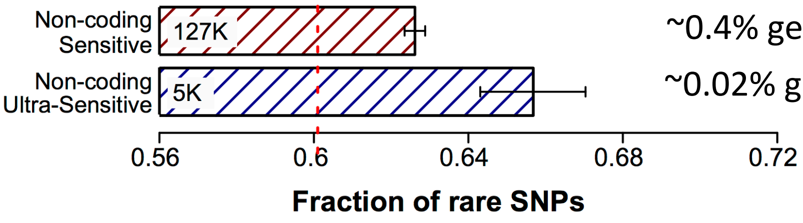
Differential selective constraints among specific sub-categories



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

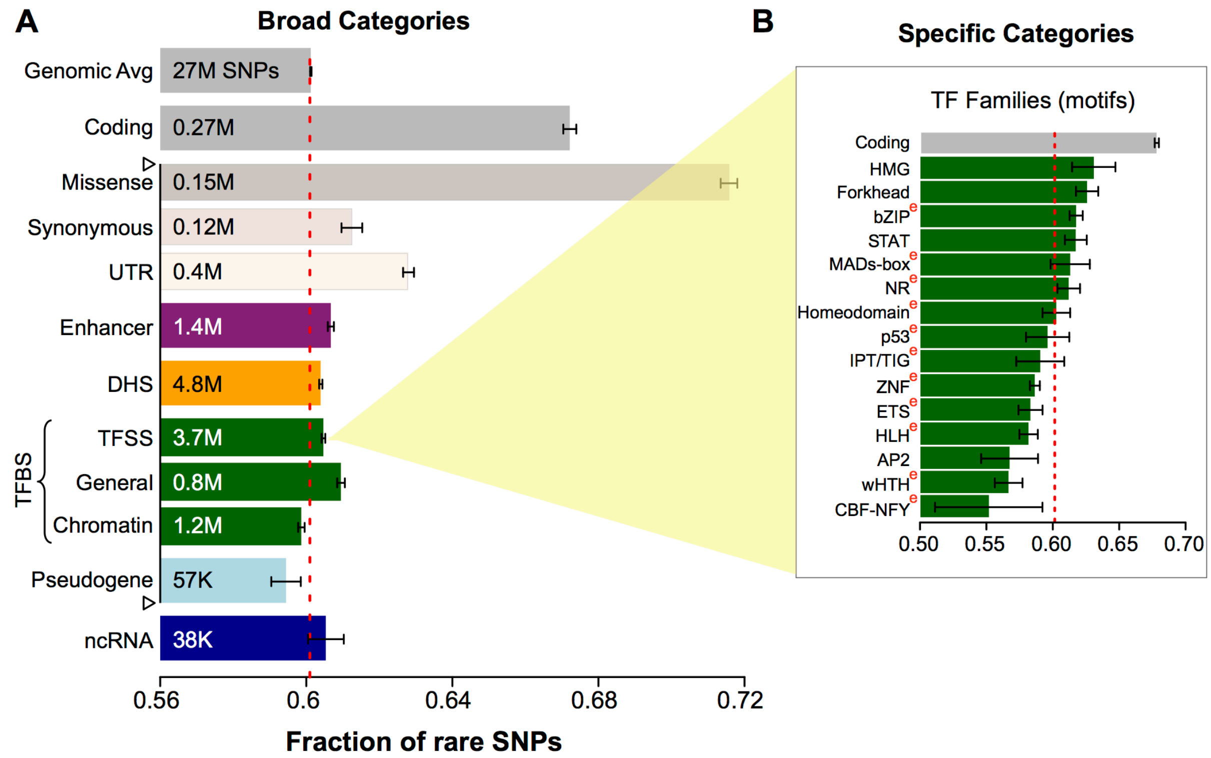
[Khurana et al., *Science* ('13)]

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)]

Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation

- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization

3D organization of genome

"We finished the genome map, now we can't figure out how to fold it."

image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

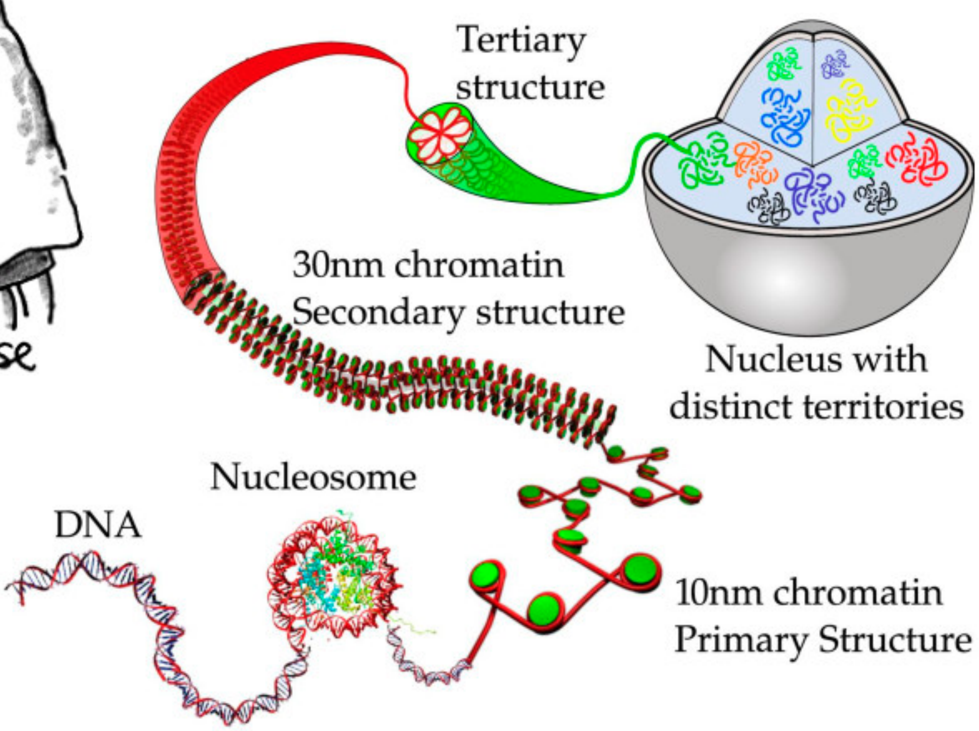
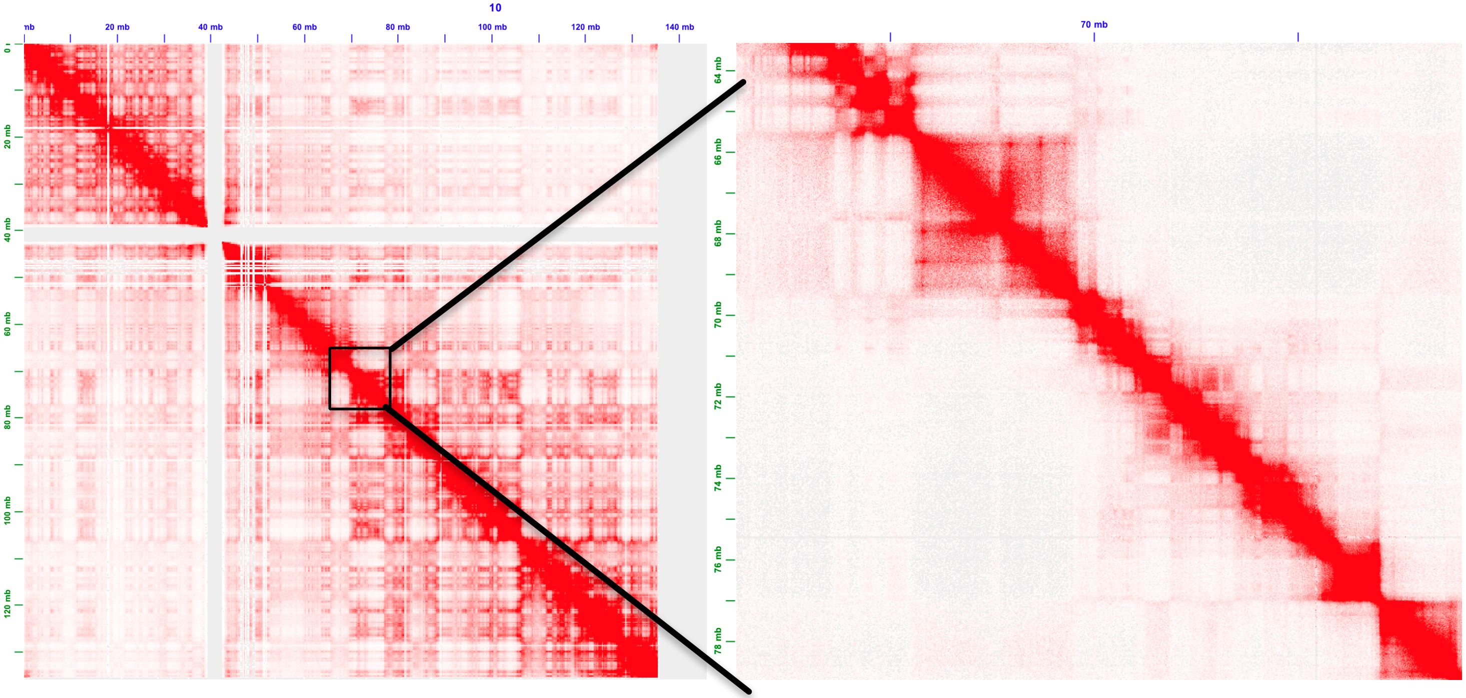

Tertiary structure
30nm chromatin Secondary structure
Nucleus with distinct territories
Nucleosome
DNA
10nm chromatin Primary Structure

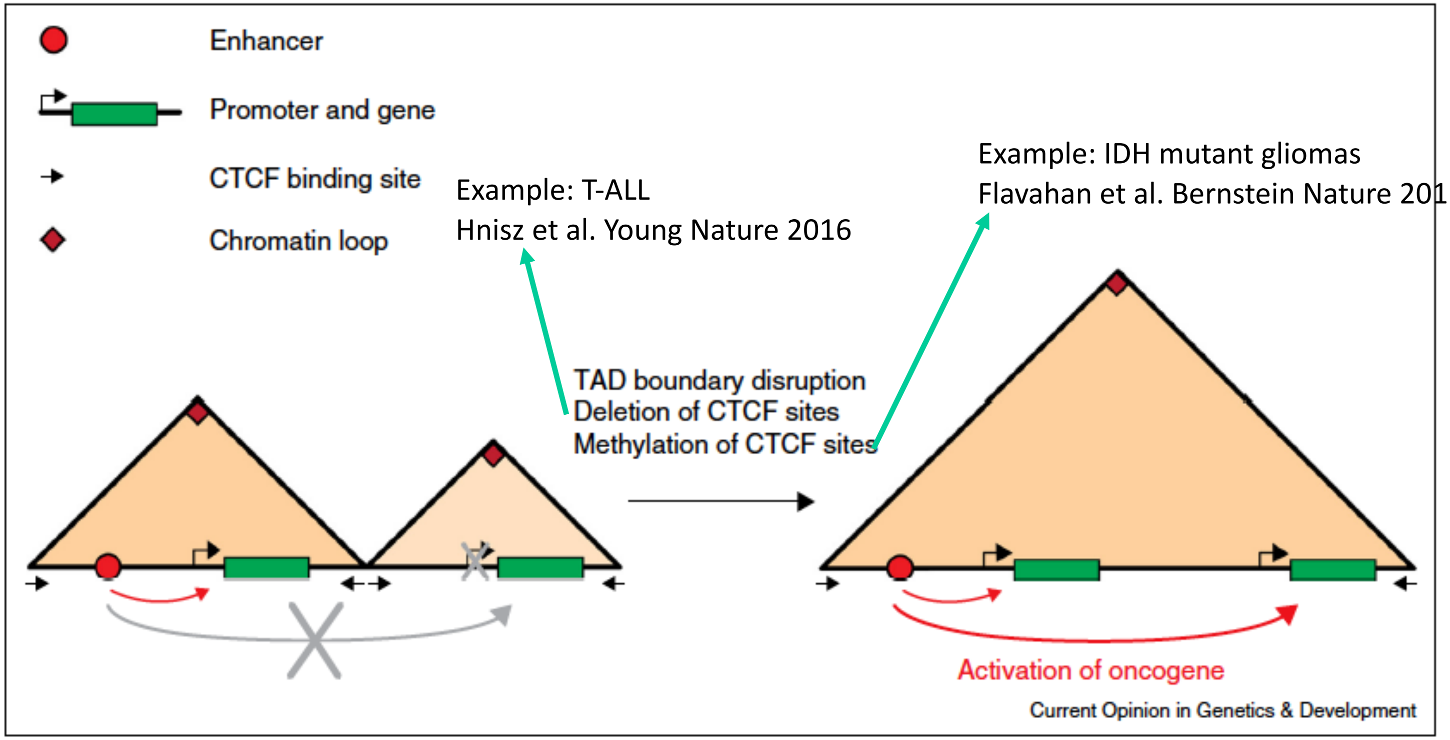
image credit: Iyer et al. BMC Biophysics 2011

Topologically associating domains (TADs)



TADs have apparent hierarchical organization

Local TAD boundary disruption activates oncogene



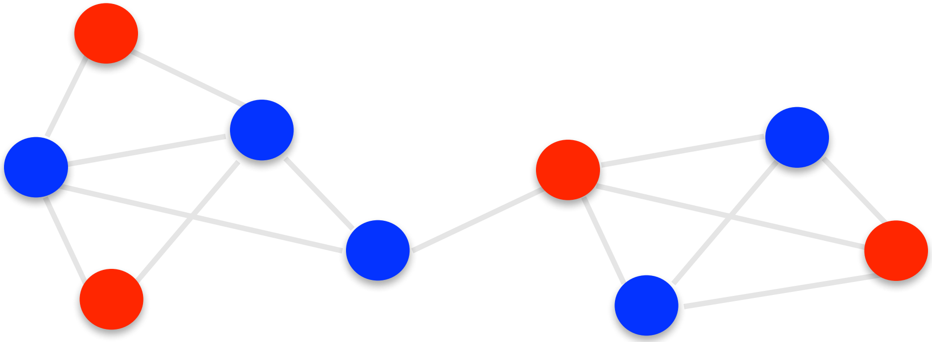
Example: T-ALL
Hnisz et al. Young Nature 2016

Example: IDH mutant gliomas
Flavahan et al. Bernstein Nature 2016

Current Opinion in Genetics & Development

Valton and Dekker Curr. Opin. Genetics and Development 2016

Network modularity



$$Q \approx 0$$

degree of i

adjacency matrix

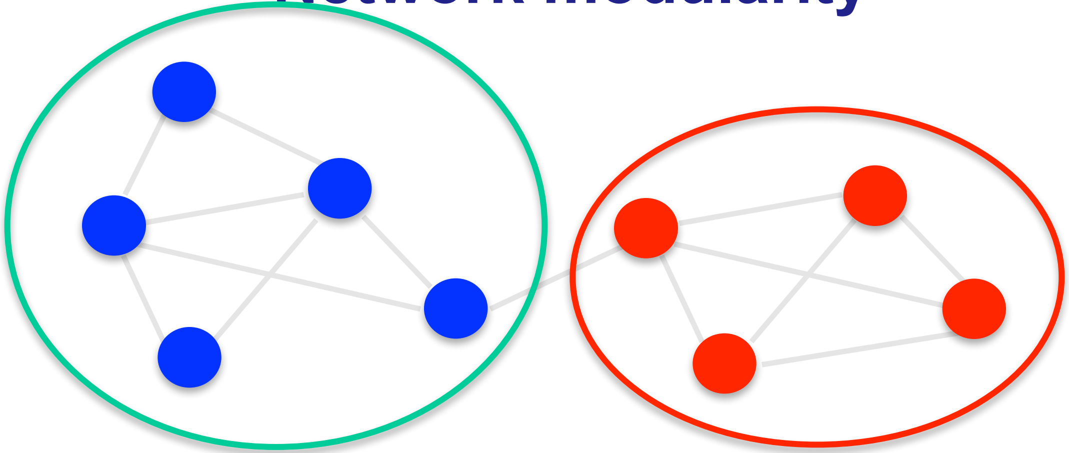
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Network modularity



$$Q = Q_{max}$$

Optimization problem

degree of i

adjacency matrix

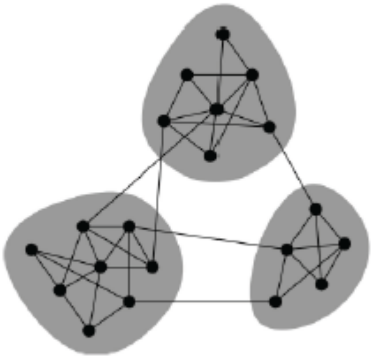
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

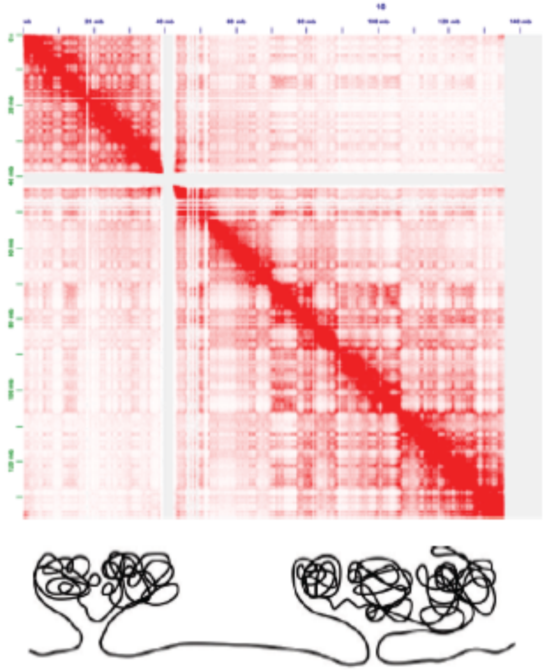
Identifying TADs in multiple resolutions



Modularity maximization

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

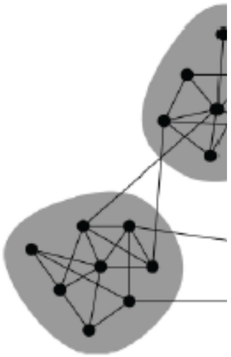
network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

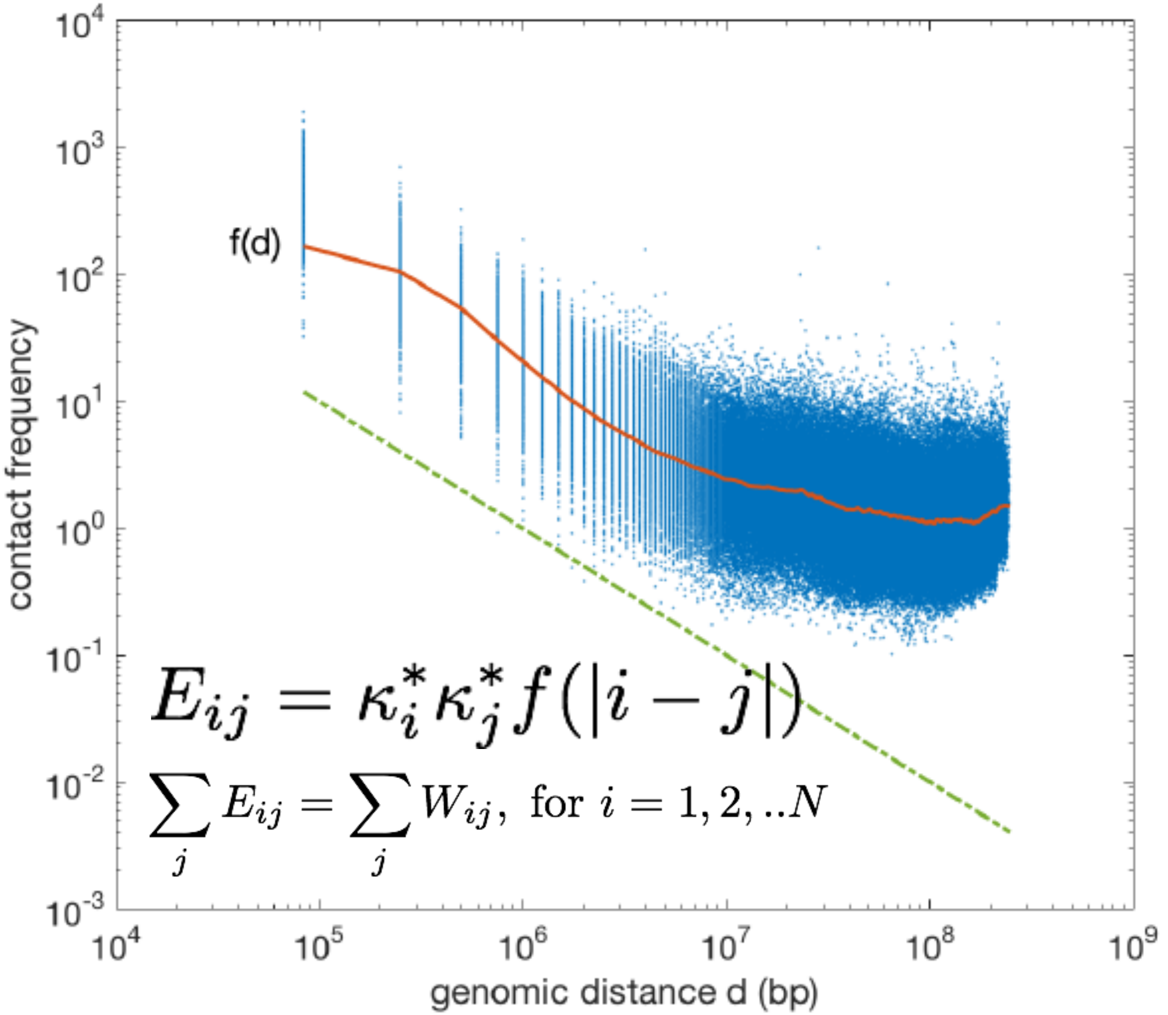
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



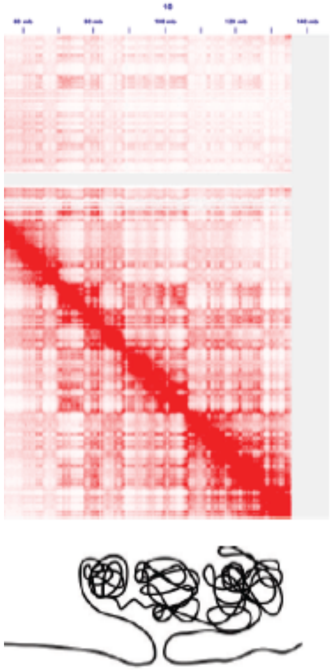
Modularity m

$$Q = \frac{1}{2m}\sum_{i,j}$$



$$E_{ij} = \kappa_i^* \kappa_j^* f\left(|i - j|\right)$$

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$

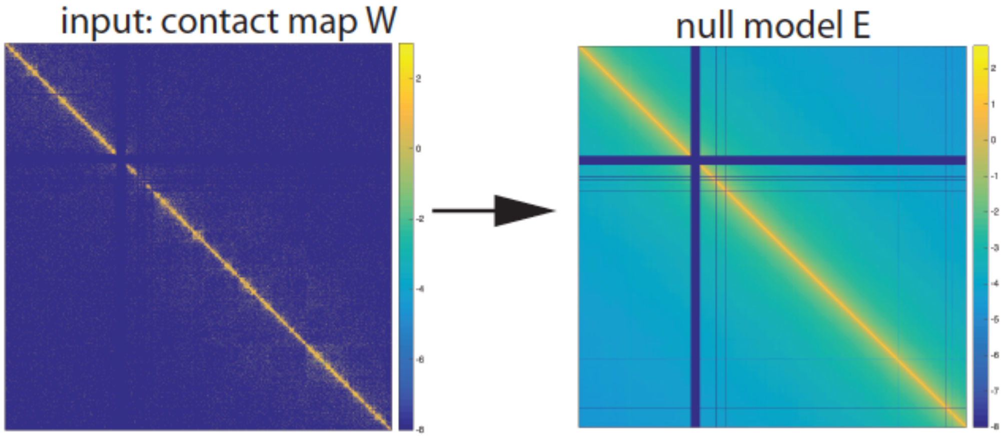


: adapted from ref. [2]

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

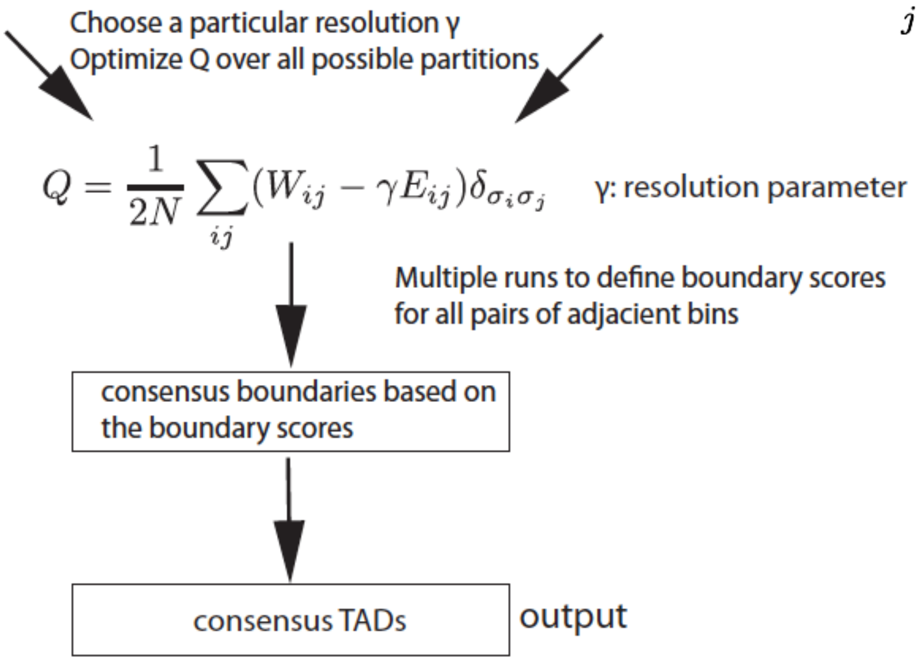
Identifying TADs in multiple resolutions



$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|)$$

Numerically solve for κ_i^* in equations

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$



Choose a particular resolution γ

Optimize Q over all possible partitions

$$Q = \frac{1}{2N} \sum_{ij} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i \sigma_j}$$

γ: resolution parameter

Multiple runs to define boundary scores for all pairs of adjacent bins

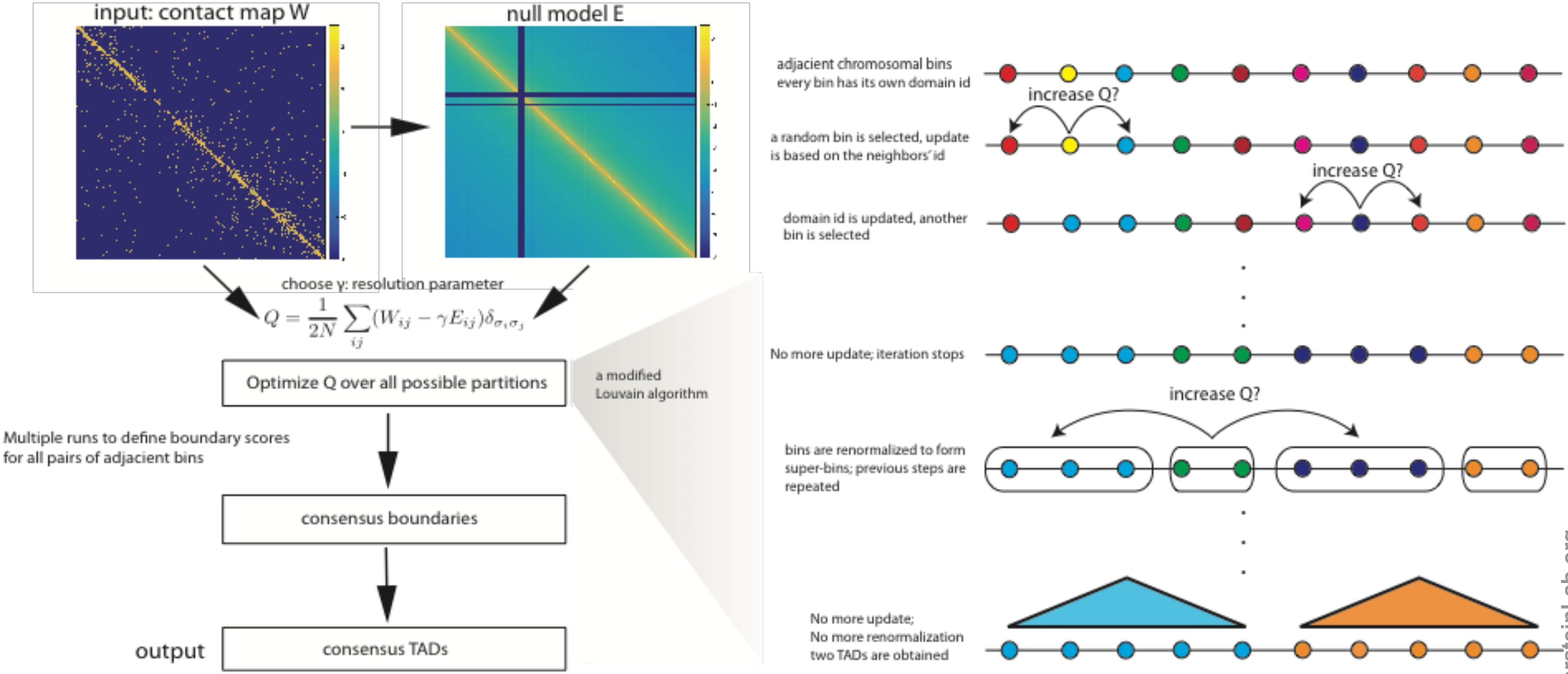
consensus boundaries based on the boundary scores

consensus TADs

output

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

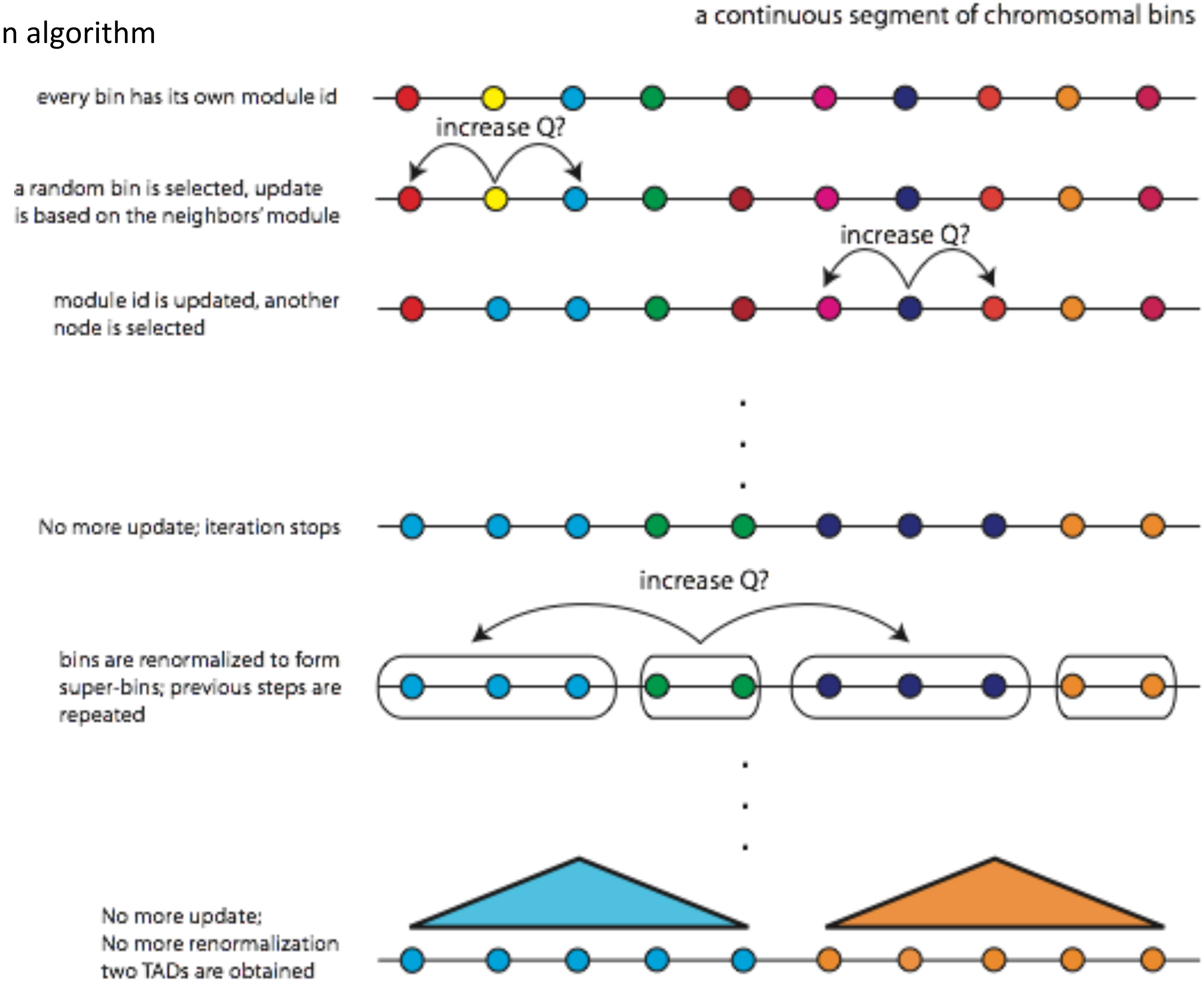
Identifying TADs in multiple resolutions


input: contact map W
null model E
choose γ: resolution parameter
$Q = \frac{1}{2N}\sum_{ij}(W_{ij} - \gamma E_{ij})\delta_{\sigma_i,\sigma_j}$
Optimize Q over all possible partitions
a modified Louvain algorithm
Multiple runs to define boundary scores for all pairs of adjacent bins
consensus boundaries
output
consensus TADs
adjacent chromosomal bins every bin has its own domain id
increase Q?
a random bin is selected, update is based on the neighbors' id
increase Q?
domain id is updated, another bin is selected
No more update; iteration stops
increase Q?
bins are renormalized to form super-bins; previous steps are repeated
No more update; No more renormalization two TADs are obtained

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions

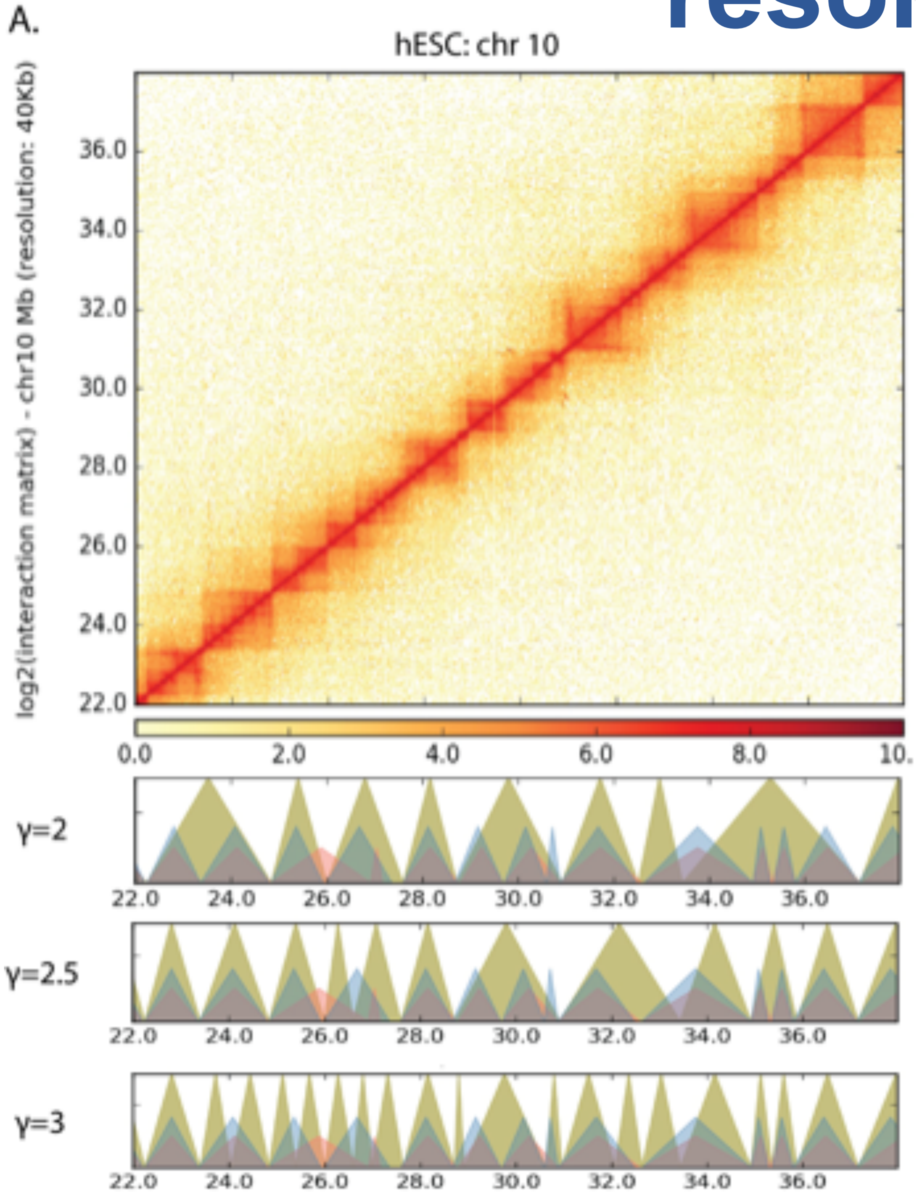
a modified Louvain algorithm



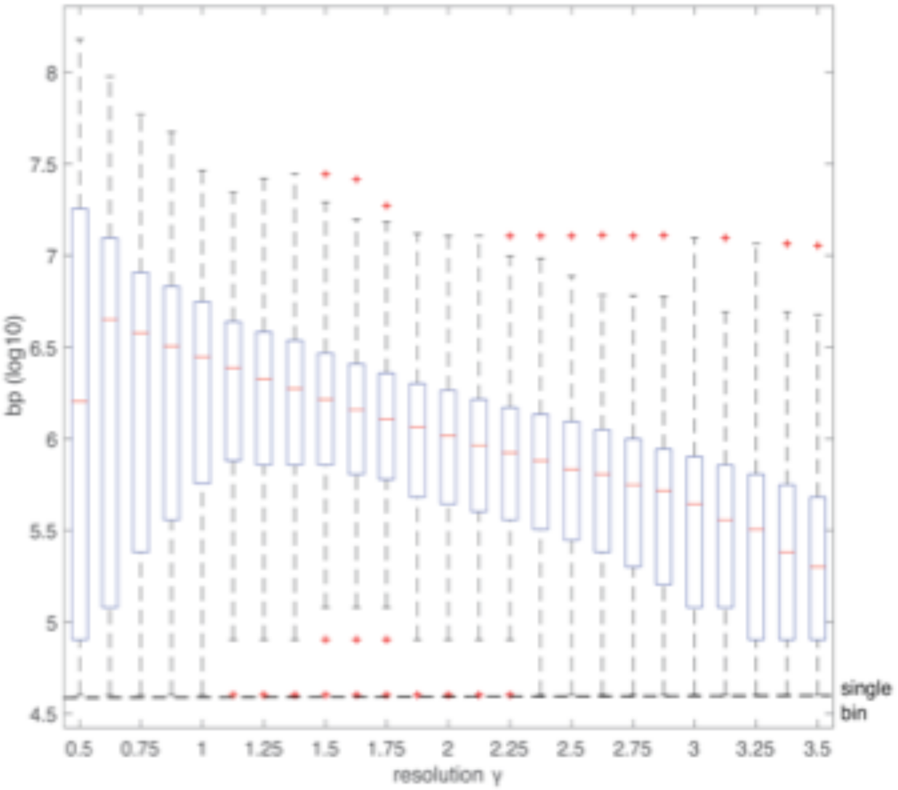
a continuous segment of chromosomal bins
every bin has its own module id
increase Q?
a random bin is selected, update is based on the neighbors' module
increase Q?
module id is updated, another node is selected
No more update; iteration stops
increase Q?
bins are renormalized to form super-bins; previous steps are repeated
No more update;
No more renormalization
two TADs are obtained

Identifying TADs in multiple resolutions

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]



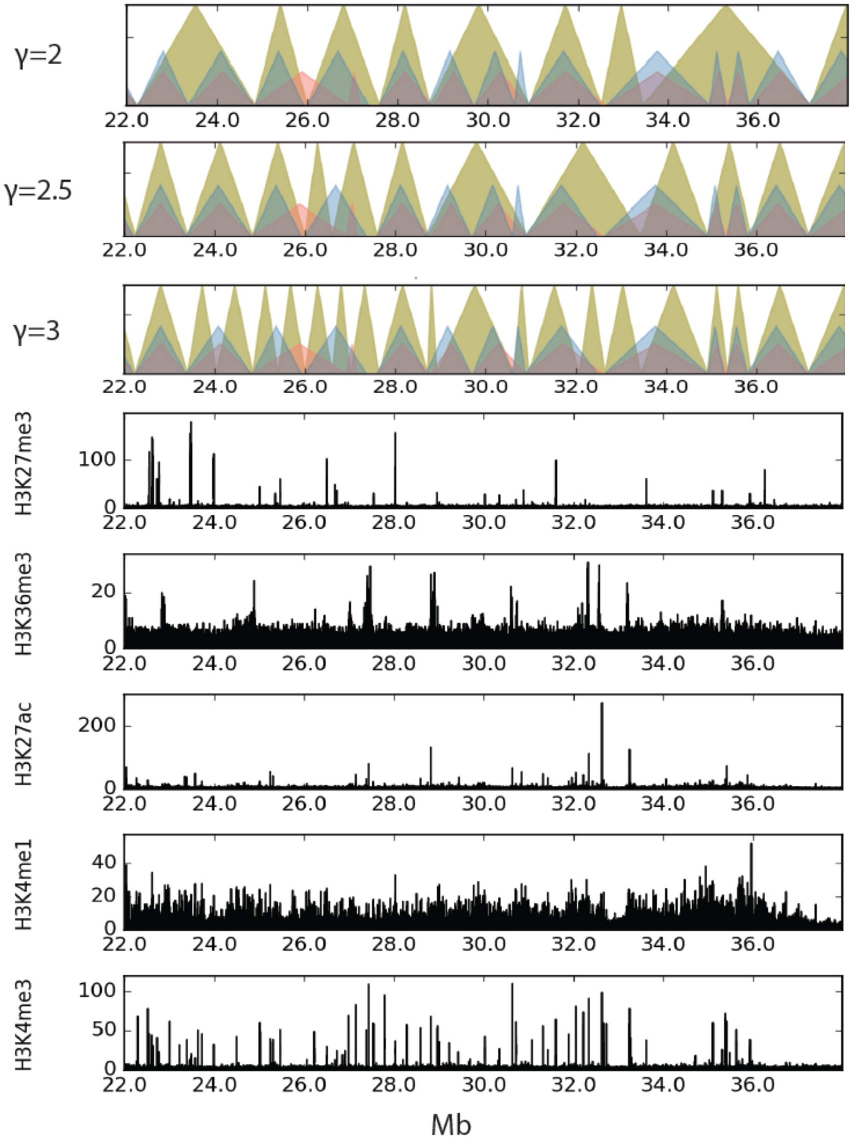
smaller TADs but are detected as the resolution increases

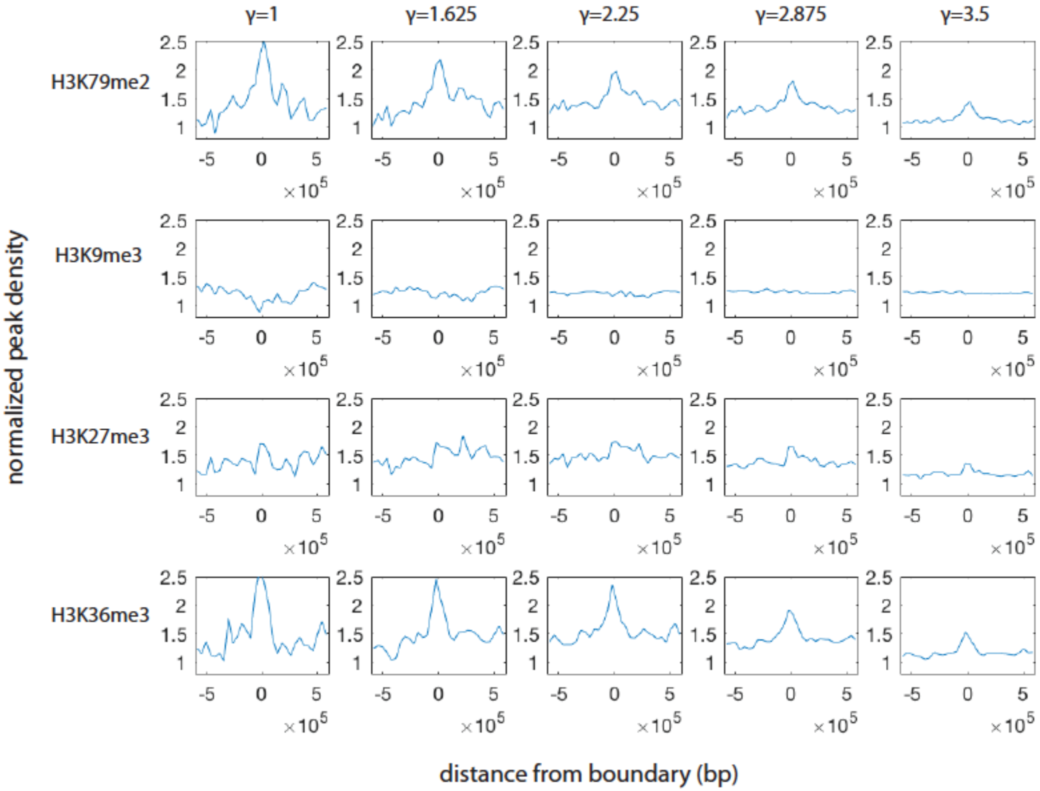


Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization

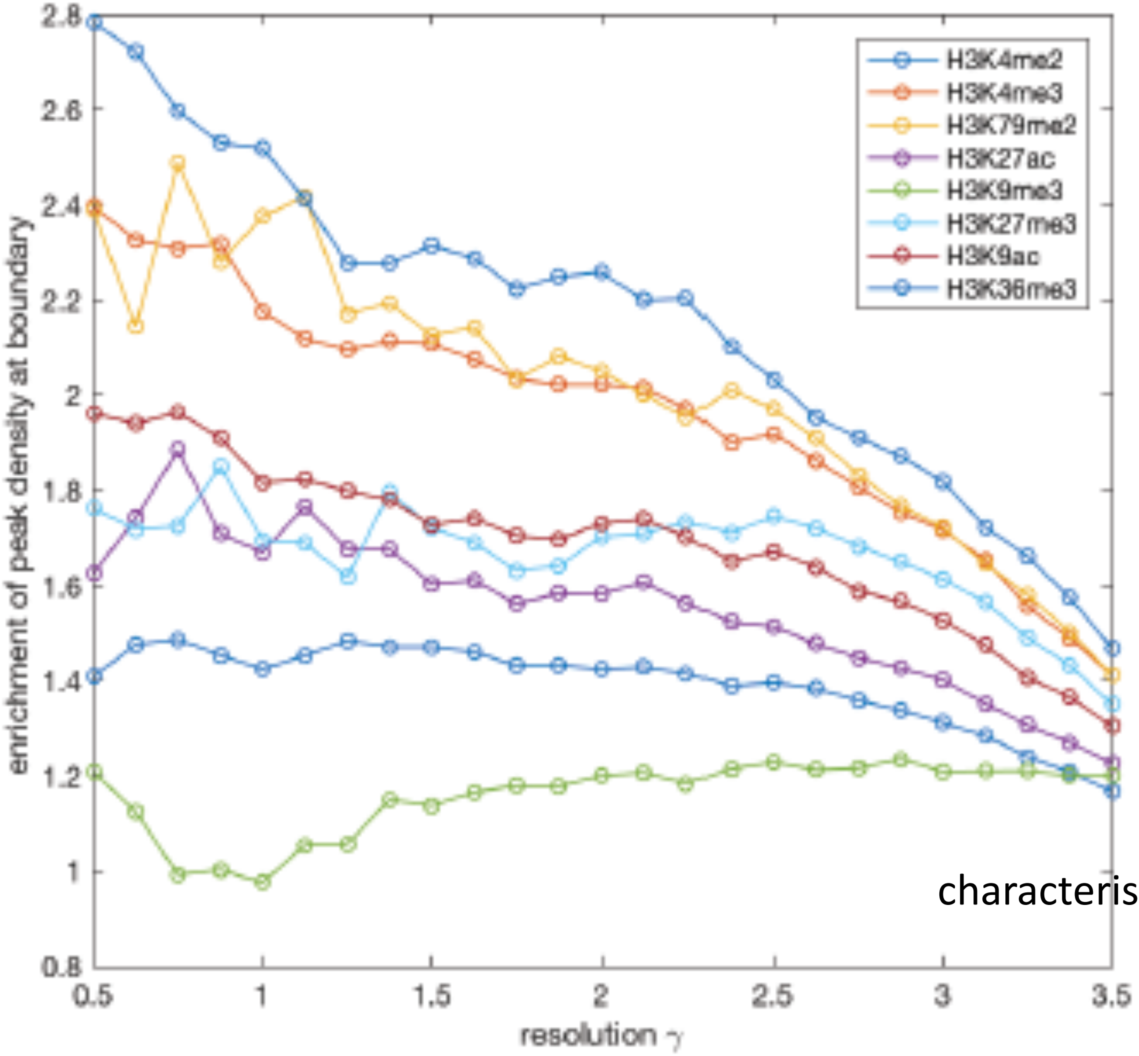
Enrichment of histone features at different resolution





[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

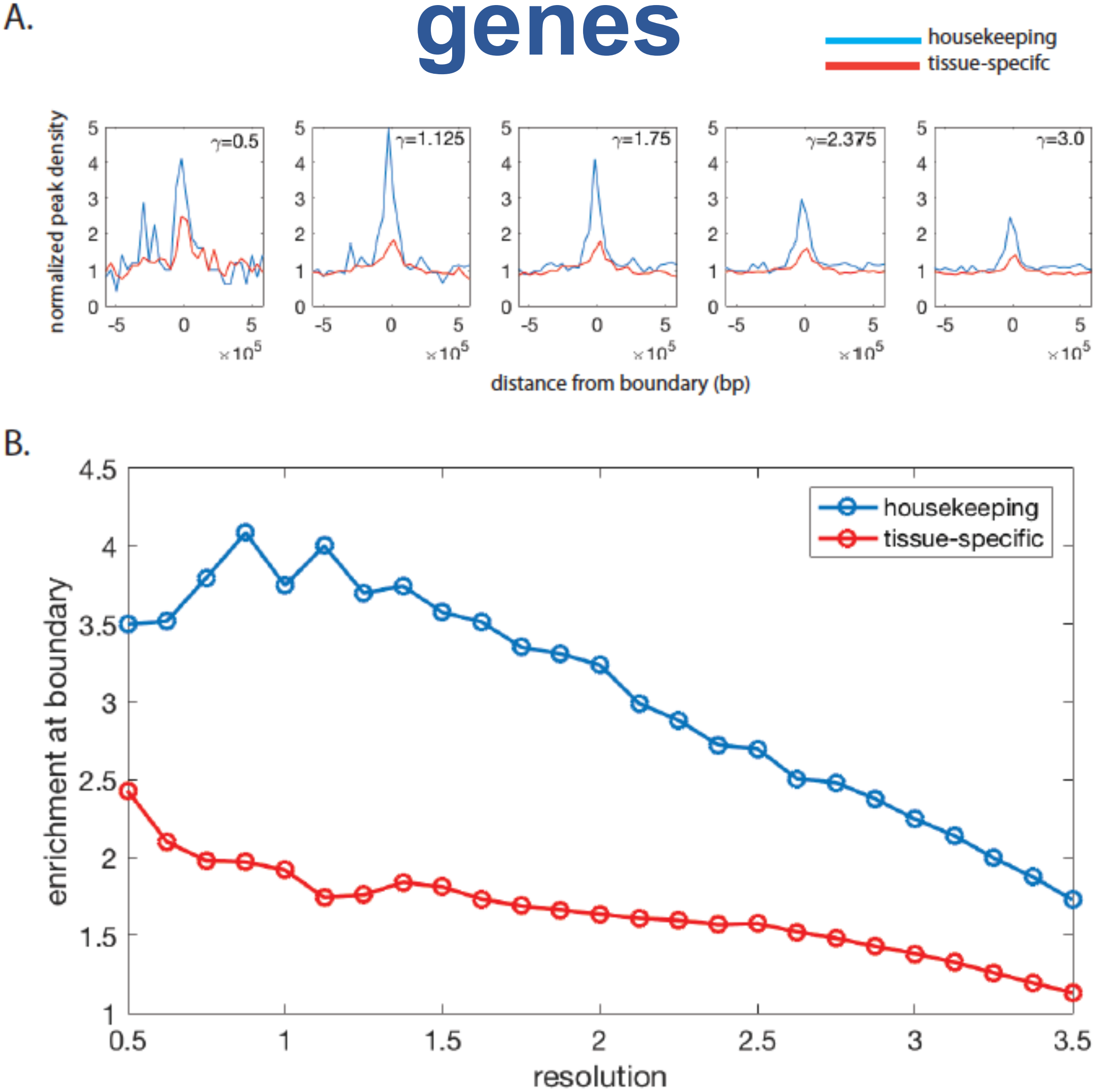
Enrichment of histone features at different resolution



characteristic length scale

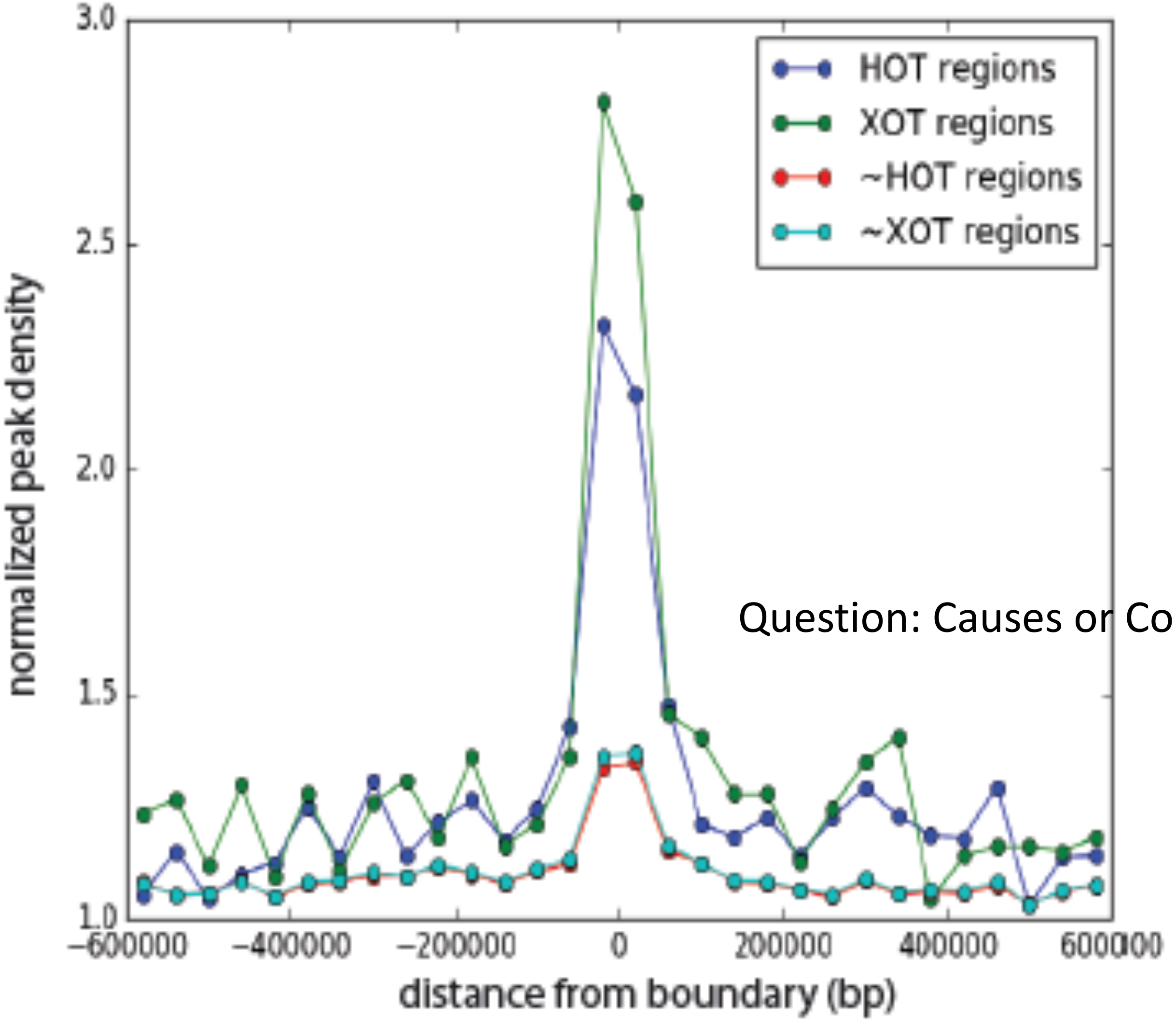
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

House-keeping vs tissue-specific genes



[Yan et al., *PLOS Comp. Bio.* (in revision, '17); bioRxiv 097345]

Enrichment of TF binding sites near boundaries

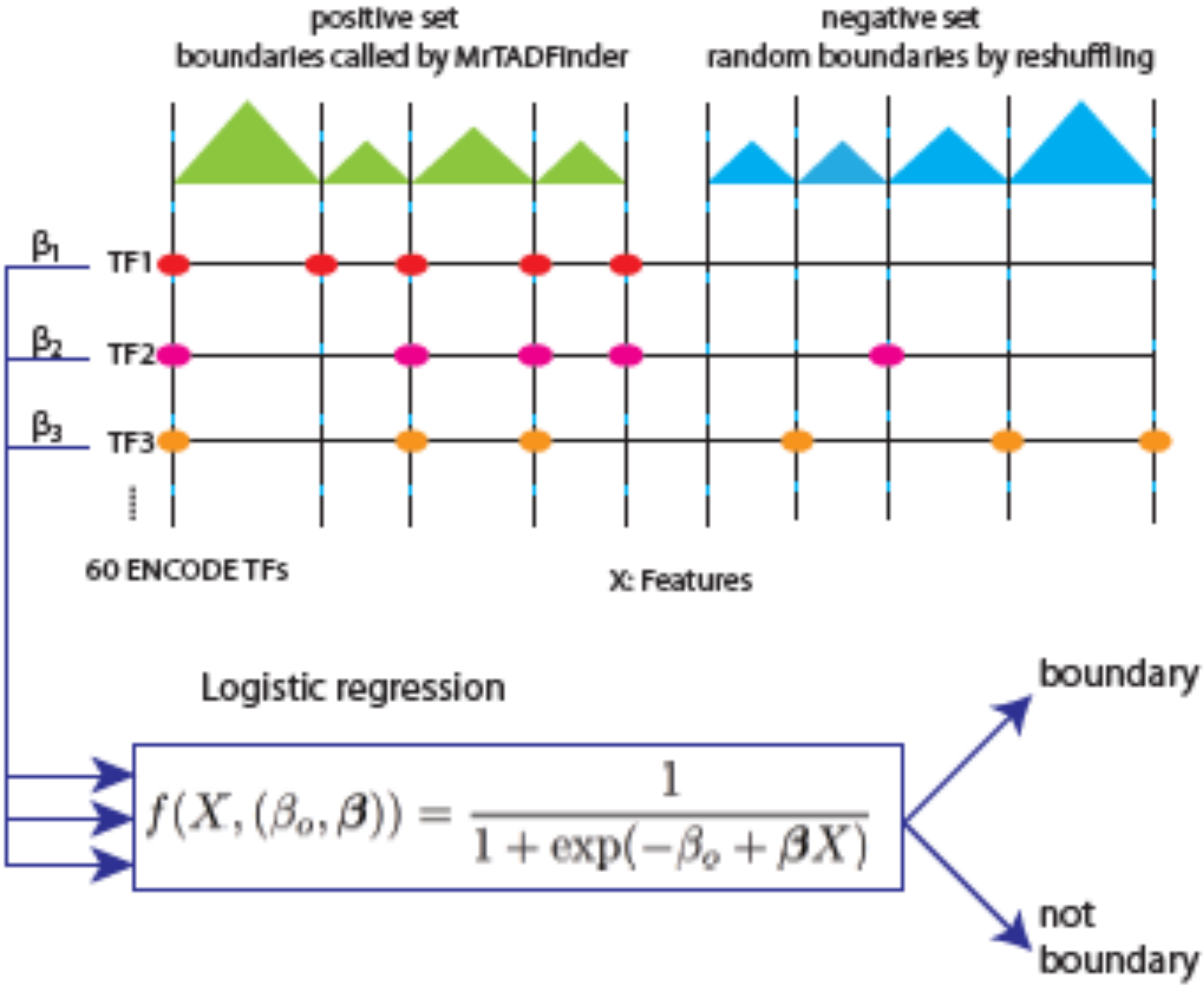


Question: Causes or Consequences?

[Yan et al., *PLOS Comp. Bio.* (in revision, '17); bioRxiv 097345]

Predicting TAD boundaries using TFs binding pattern

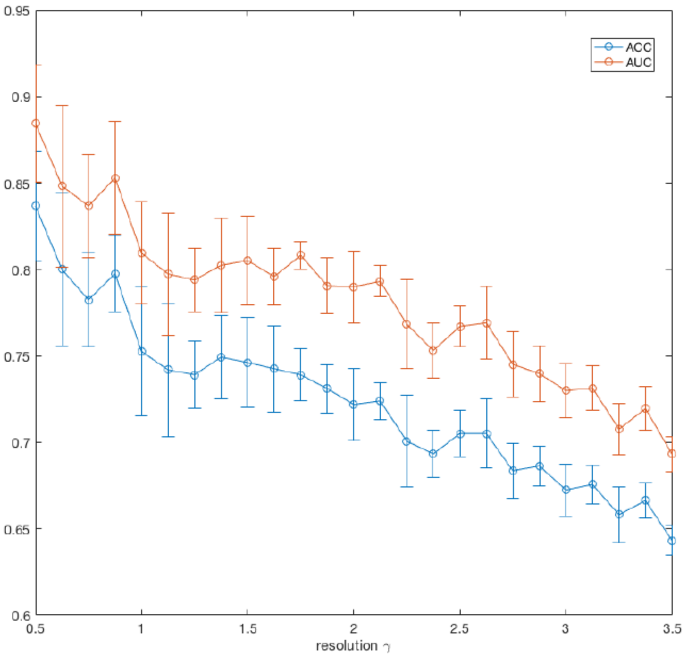
Classification problem:



$$f(X,(\beta_o,\beta)) = \frac{1}{1+\exp(-\beta_o+\beta X)}$$

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

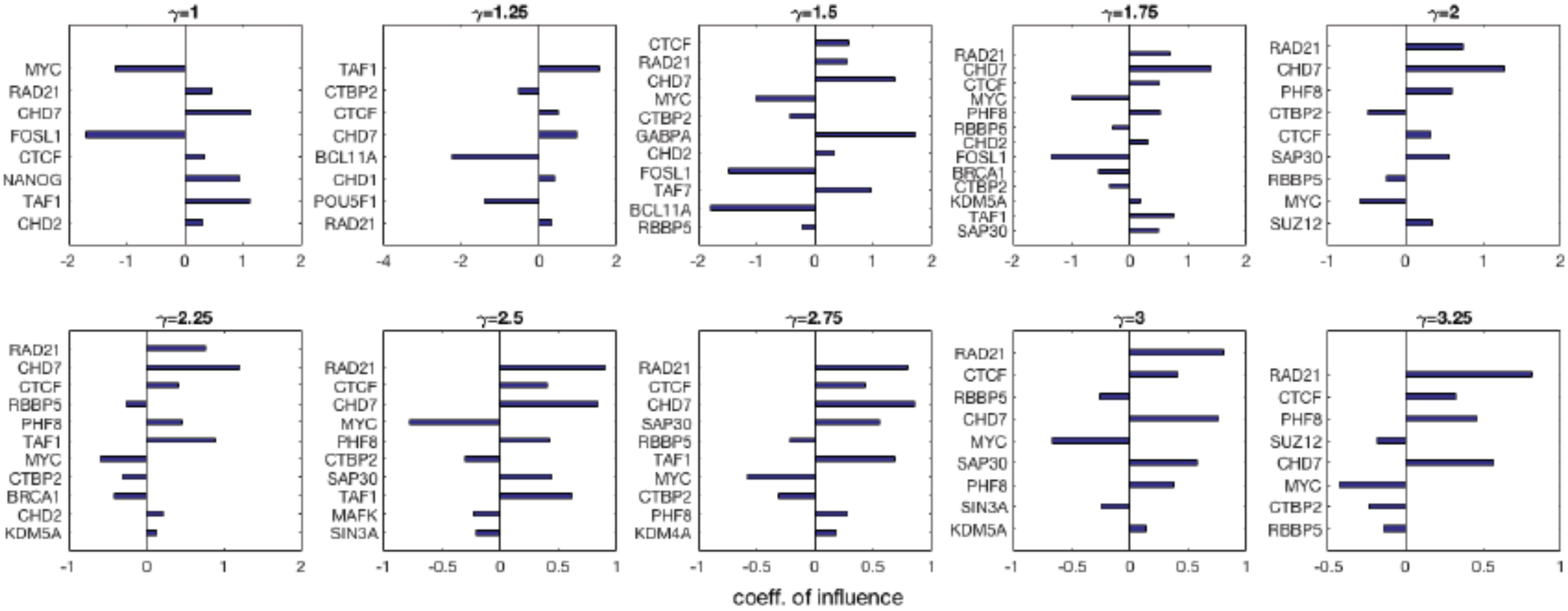
model performance



Predicting TAD boundaries using chromatin features

Which transcription factors play a role in border formation?

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]



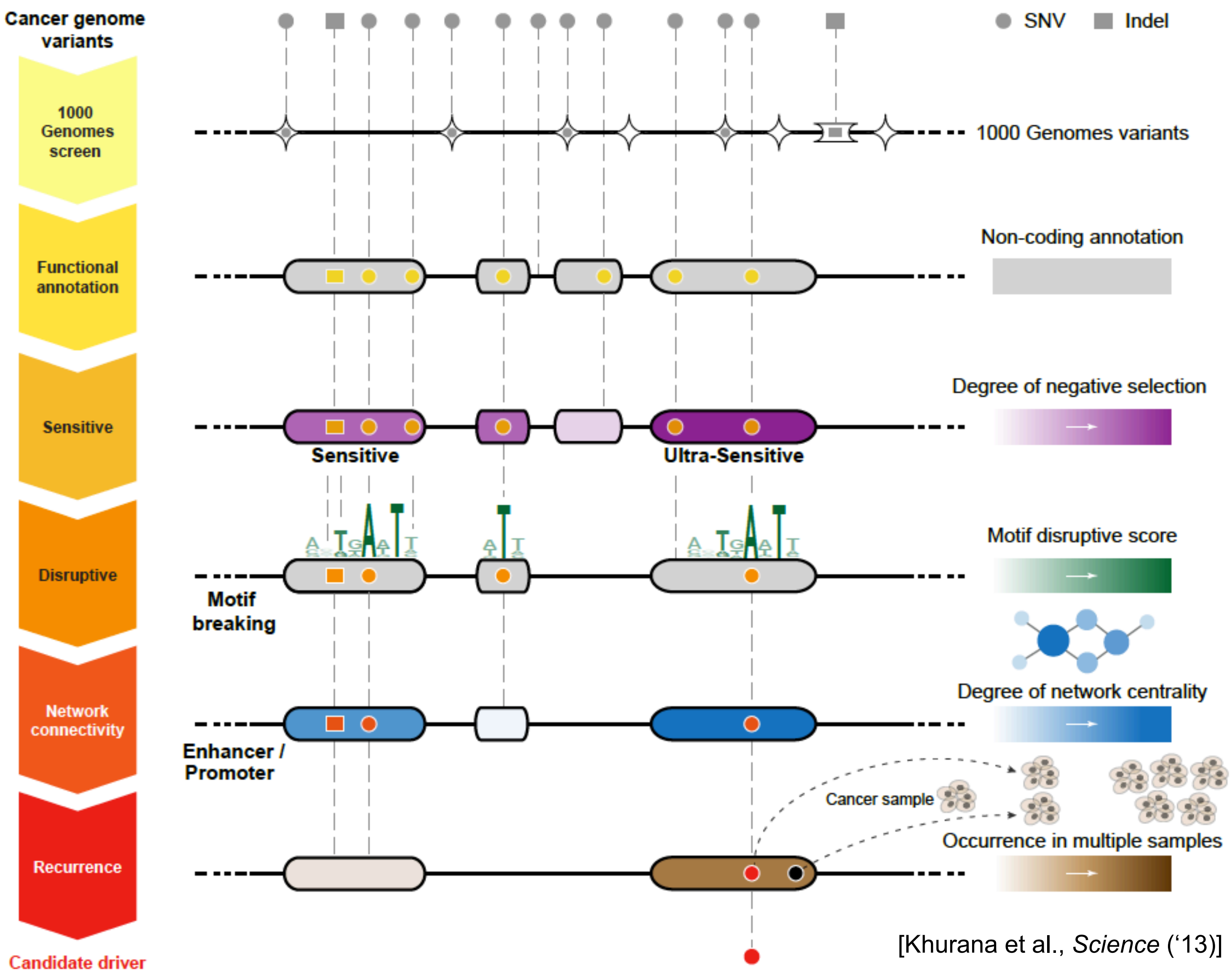
contribution of individual factors

Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation

- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization

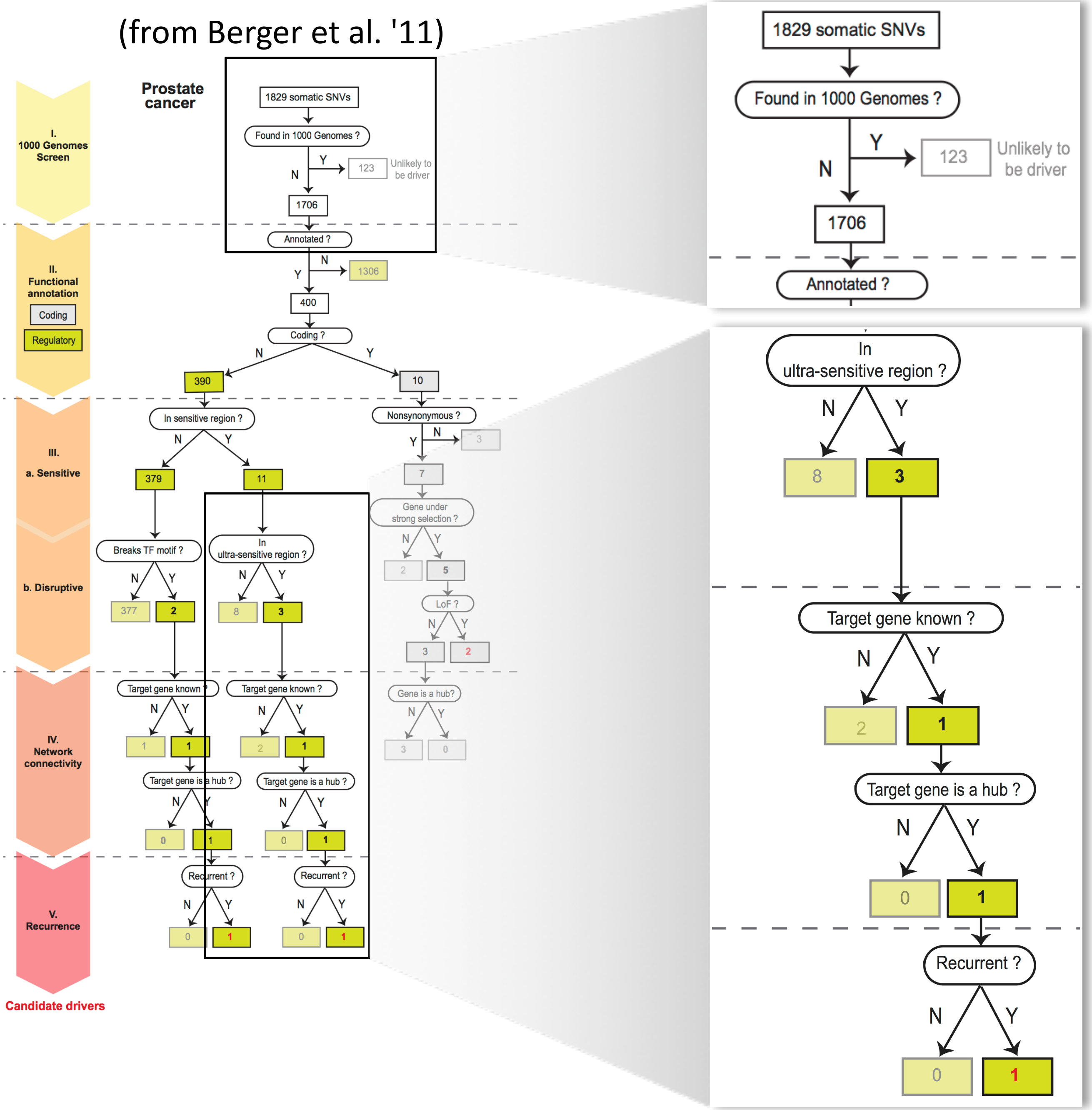
Identification of non-coding candidate drivers amongst somatic variants: Scheme



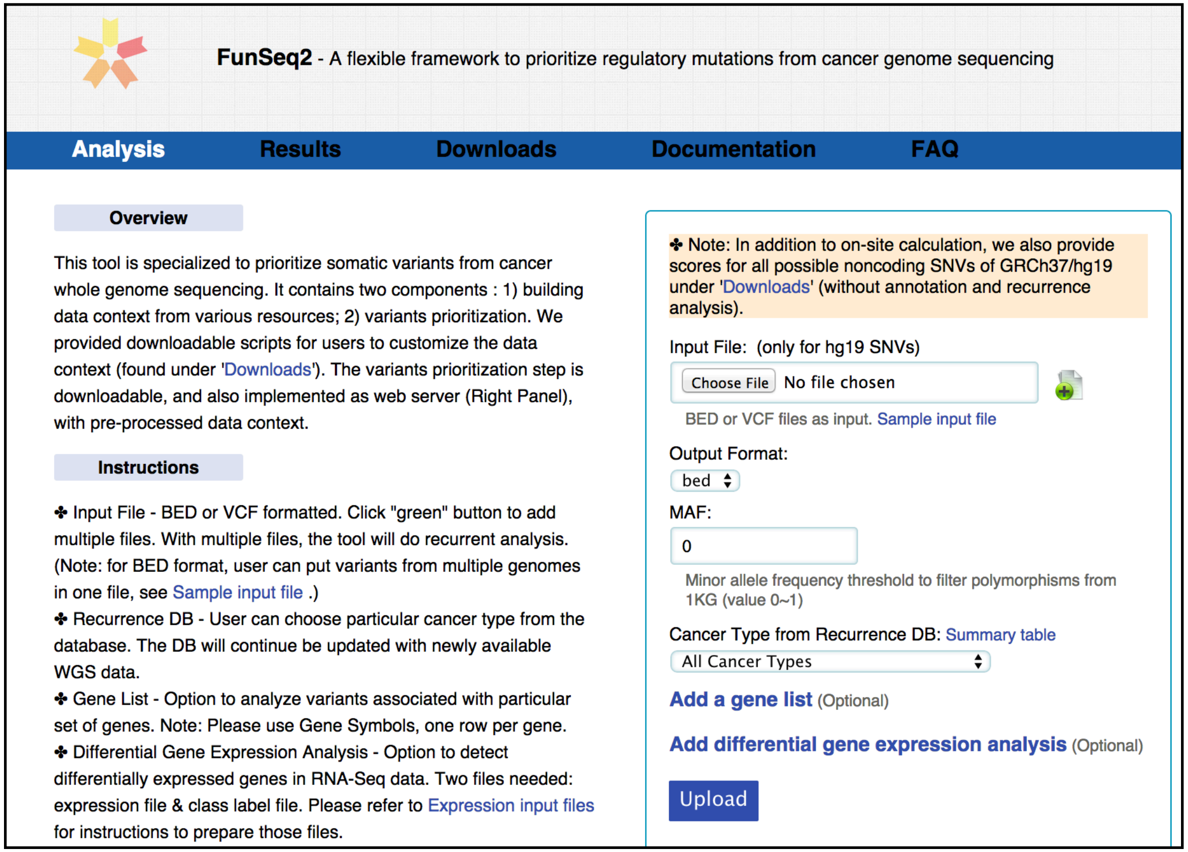
[Khurana et al., *Science* ('13)]

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)

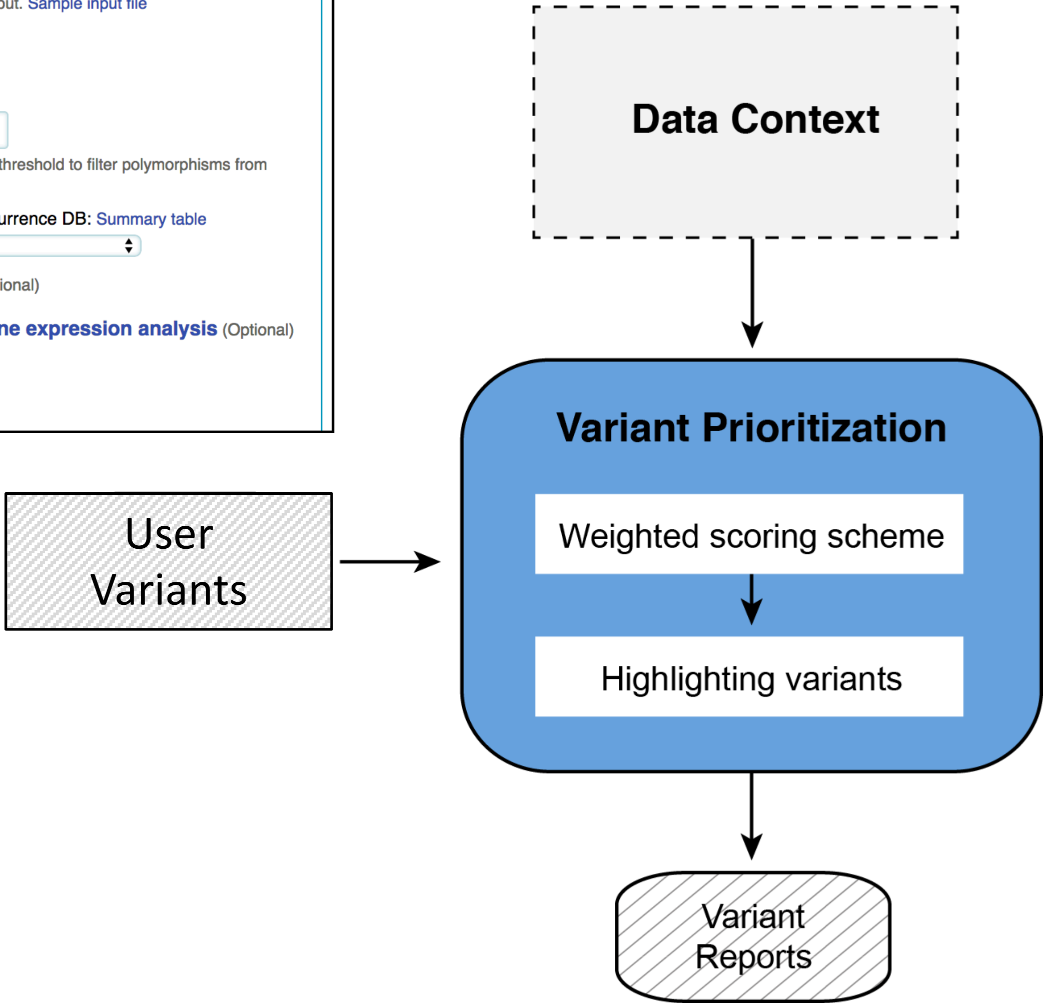
Cancer Type from Recurrence DB: Summary table

All Cancer Types

Add a gene list (Optional)

Add differential gene expression analysis (Optional)

Site integrates user variants with large-scale context

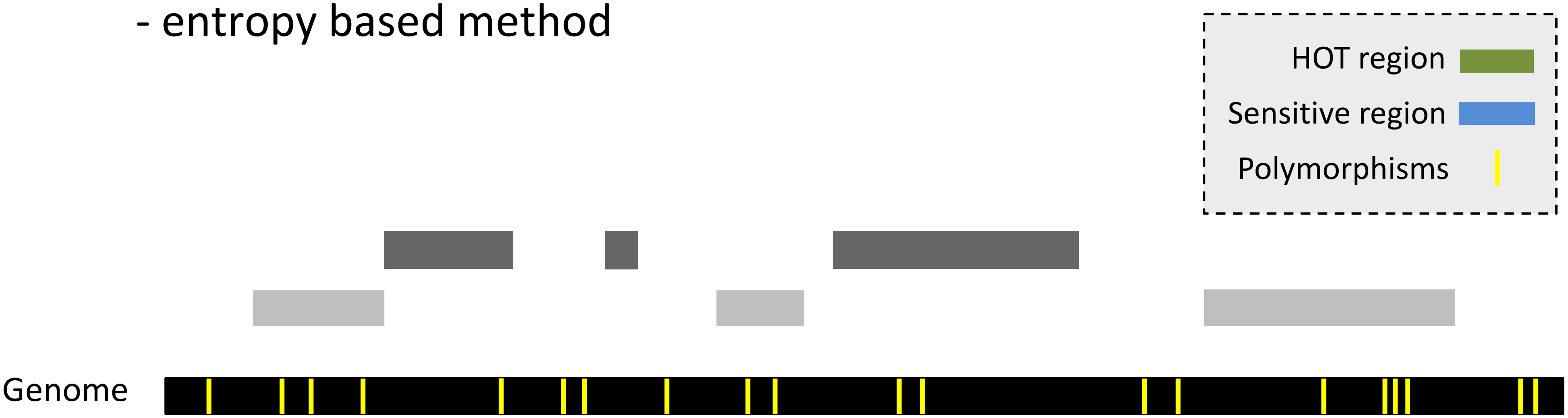


FunSeq.gersteinlab.org

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

- Feature weight
  - Weighted with mutation patterns in natural polymorphisms

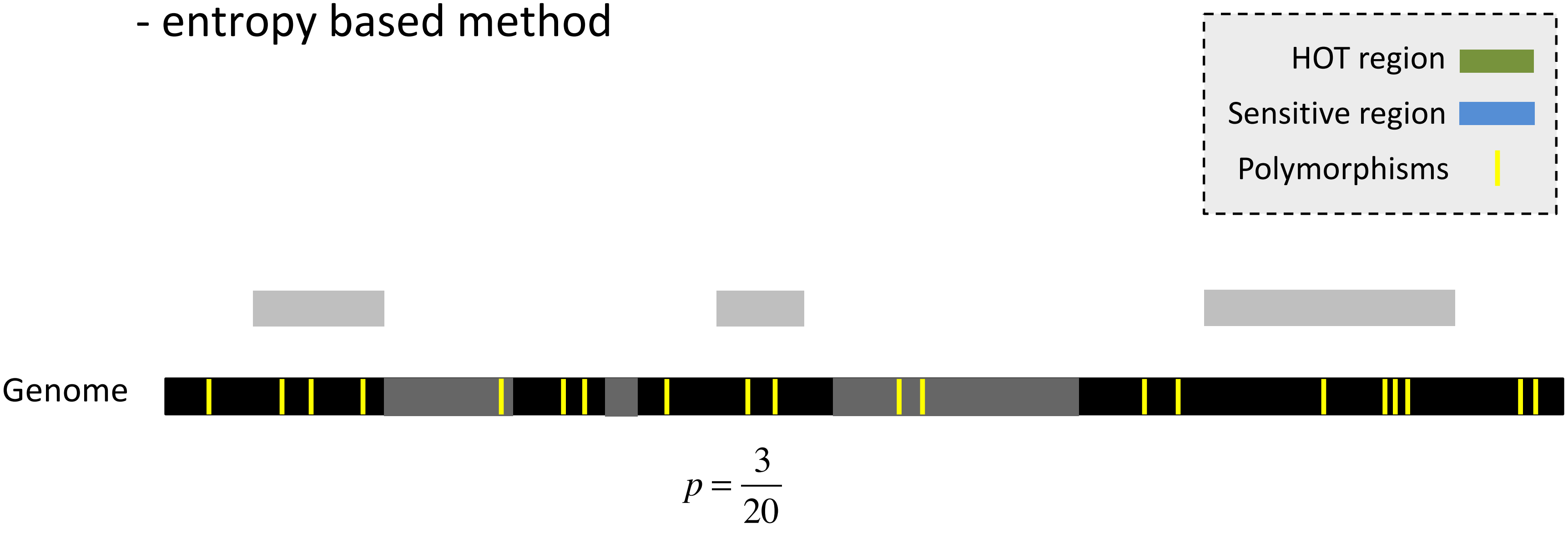
    (features frequently observed weight less)
  - entropy based method



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

- Feature weight

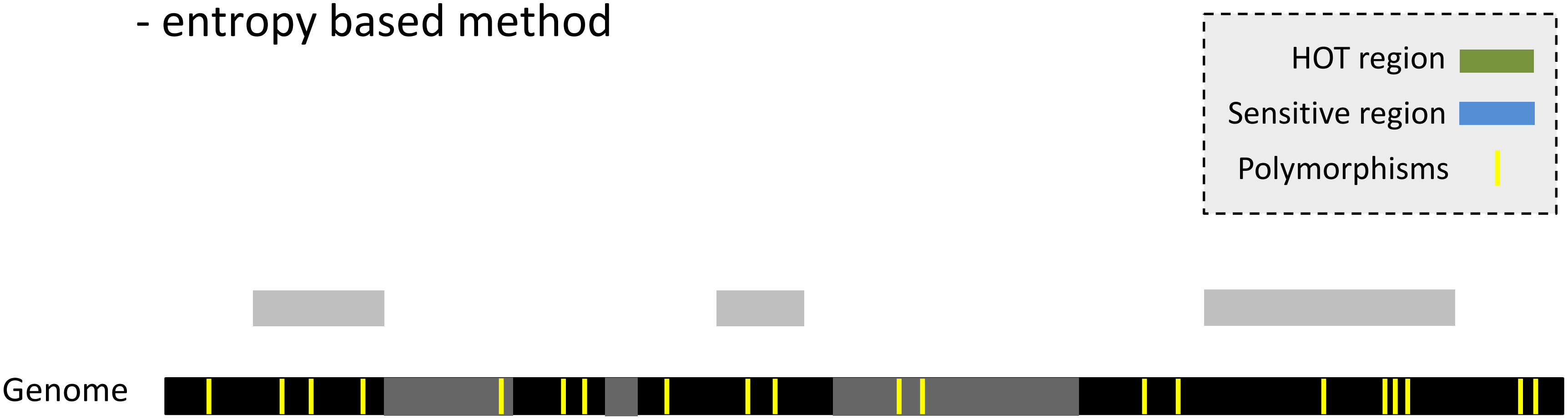
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



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

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

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



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

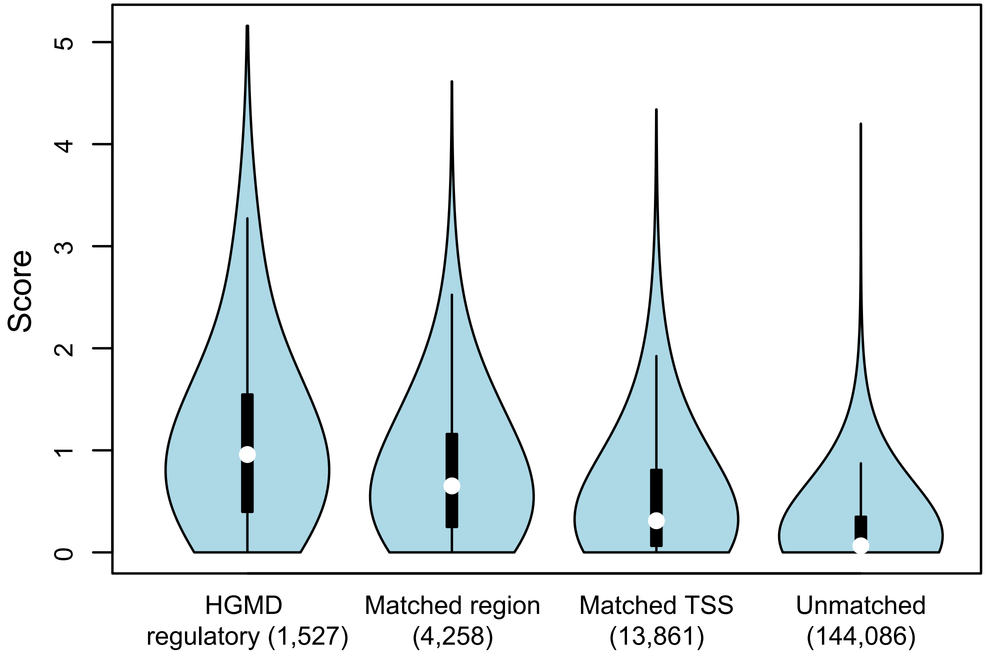
Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

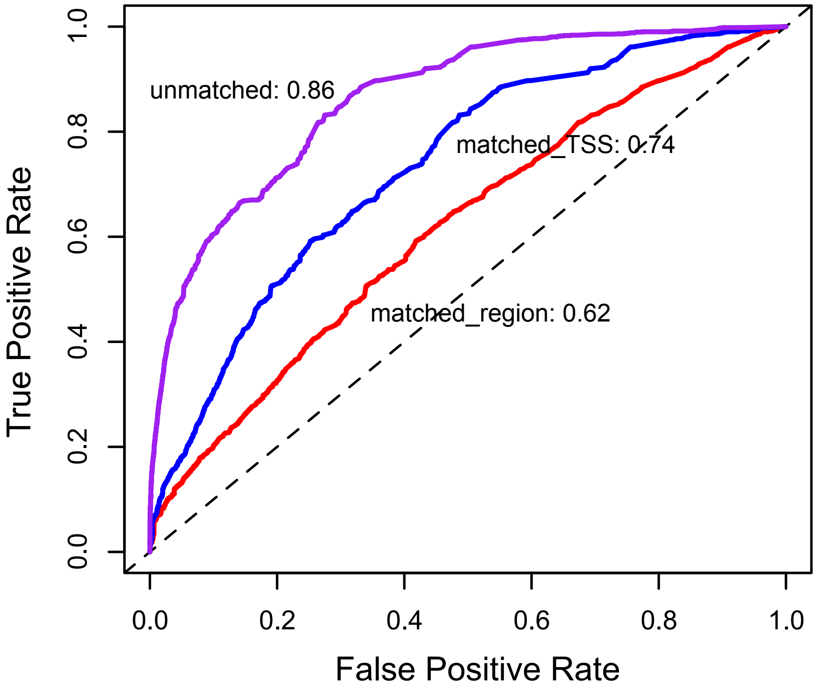
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

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

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

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)]

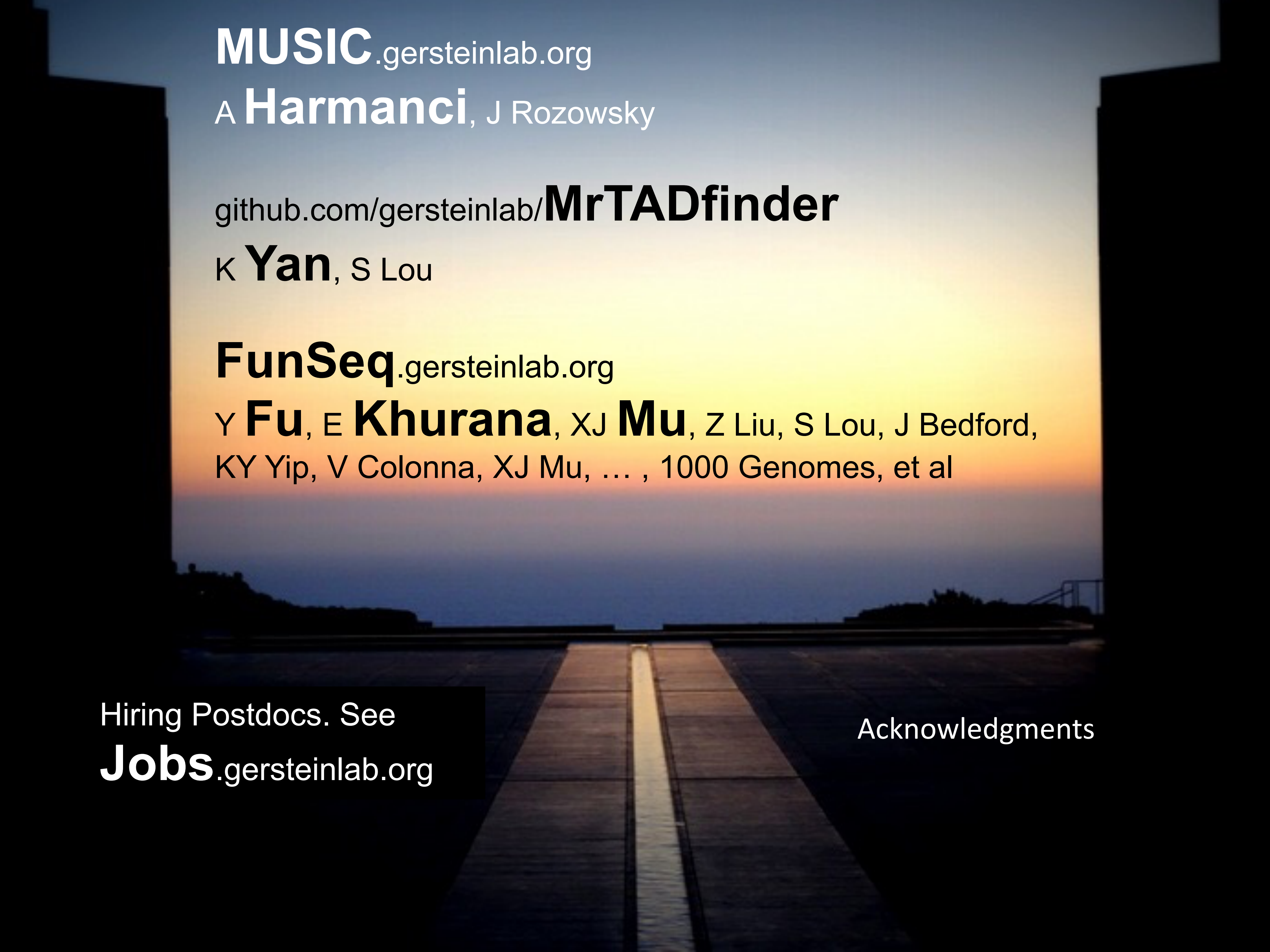
Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation

- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization

Multi-scale Element Annotation & Variant Prioritization

- Characterizing Regulatory Sites at Multiple Scales
  - Multi-scale "site" calling (with Music)
  - Using high resolution conservation information to find sensitive sites
- Characterizing TADs at Multiple Scales
  - Using modularity for identification
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
- FunSeq Software Tool for Variant Prioritization
  - Systematically weighting all the features, for non-coding prioritization



Acknowledgments

MUSIC.gersteinlab.org

A **Harmanci**, J Rozowsky

github.com/gersteinlab/**MrTADfinder**

K **Yan**, S Lou

FunSeq.gersteinlab.org

Y **Fu**, E **Khurana**, XJ **Mu**, Z Liu, S Lou, J Bedford, KY Yip, V Colonna, XJ Mu, … , 1000 Genomes, et al

Hiring Postdocs. See
Jobs.gersteinlab.org

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read statement at 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 are 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 streams.gerstein.info .