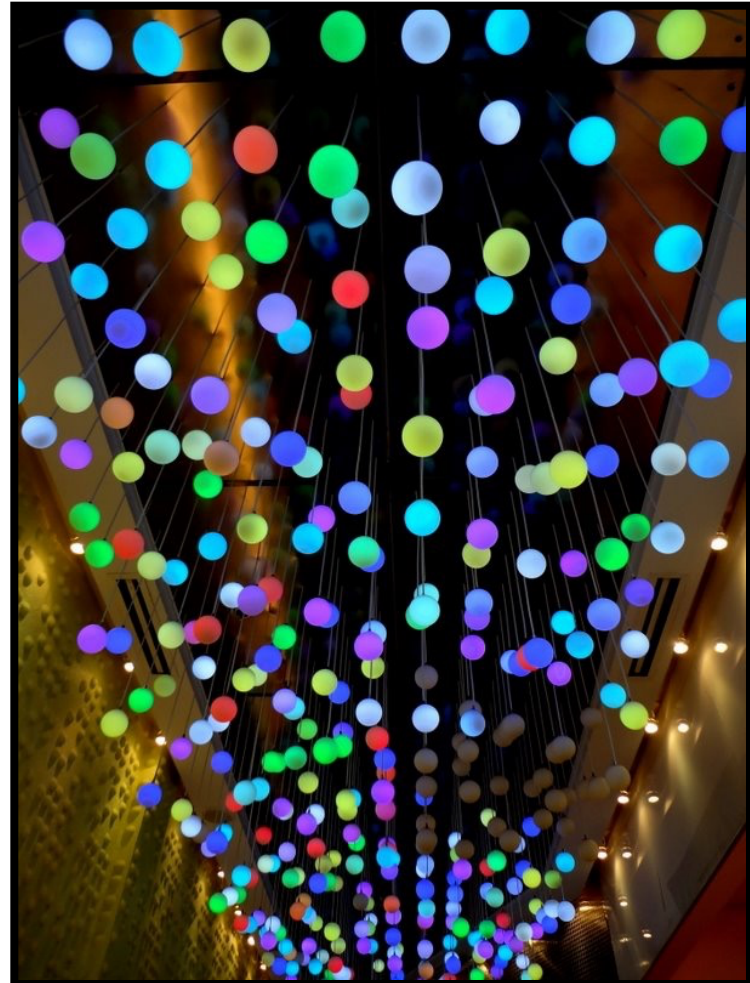
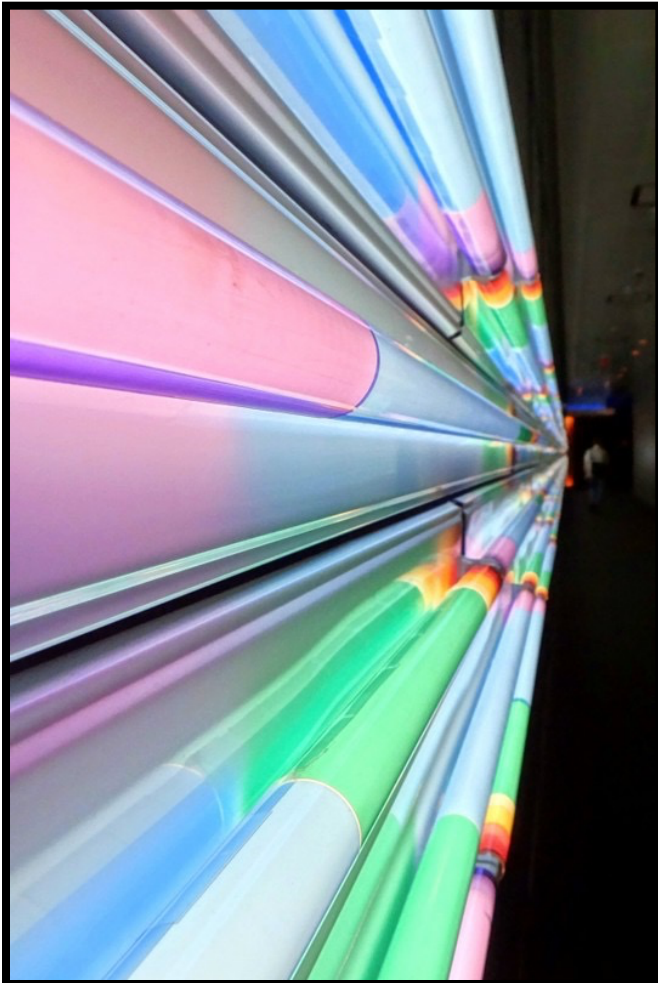


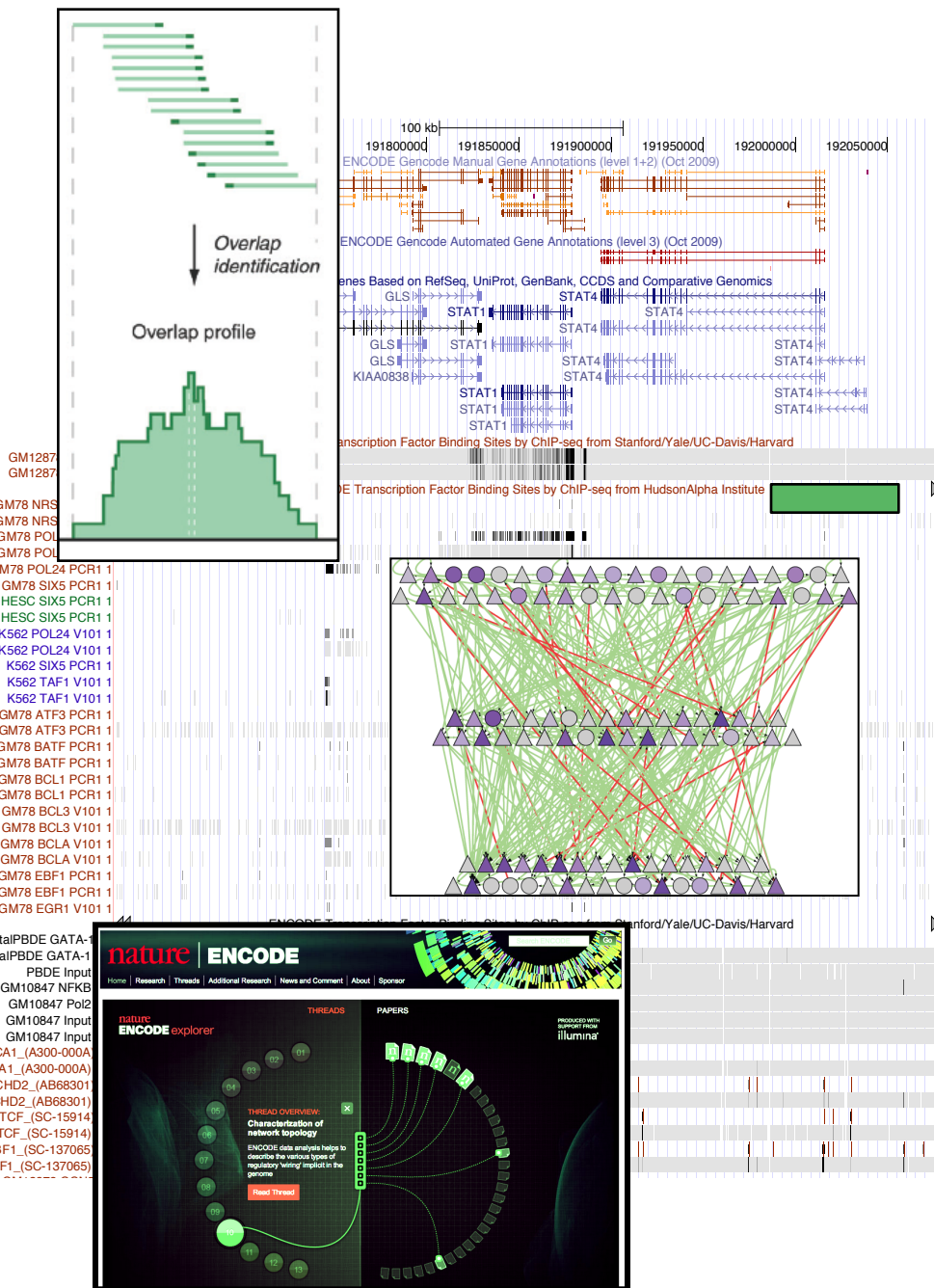
Human Genome Analysis: Progressive summarization of large-scale data, to interpret mutations & dis-regulation in cancer

Mark Gerstein, Yale

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



Organizing Genomic "Big Data" through a Hierarchy of Progressive Summarization



- Raw data (reads) at bottom
- Progressive Processed Summaries
 - Signals
 - Site locations
 - Networks, states & models
- At top are linked publications documenting everything, forming metadata for the summaries
- Using the summary to interpret a new dataset

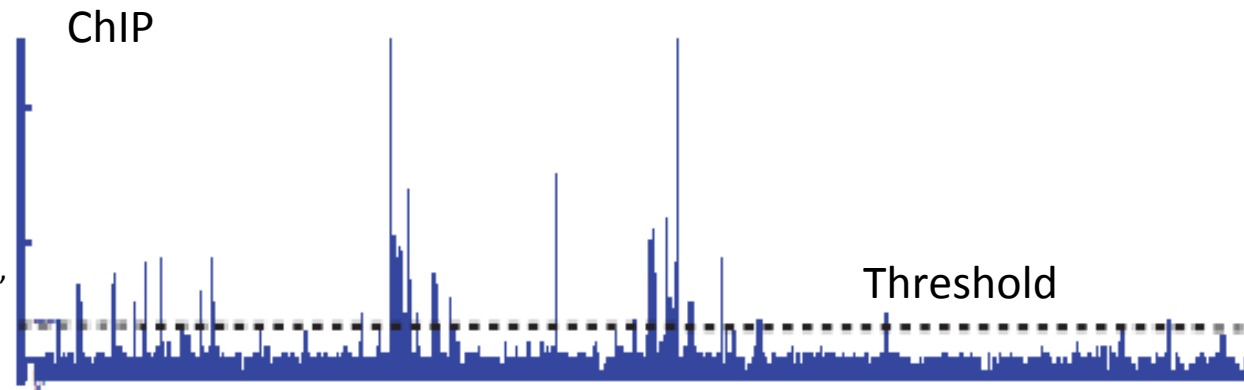
[Nature 489: 208]

Human Genome Analysis:
Progressive summarization of large-scale data,
to interpret mutations & dis-regulation in cancer

- **Summarizing Large-scale Genomic Information**
 - **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 Summaries to Interpret Alterations in Cancer**
 - **FunSeq software tool for mutation prioritization**
 - Systematically weighting all the features, for non-coding prioritization
 - Summarizing large data context into simple "Core Score File"
 - **Loregic: Logic-gate analysis of regulation**
 - Recasting the regulatory network as a collection of gates
 - Different gate structure in cancer, dominated by particular driver TFs

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

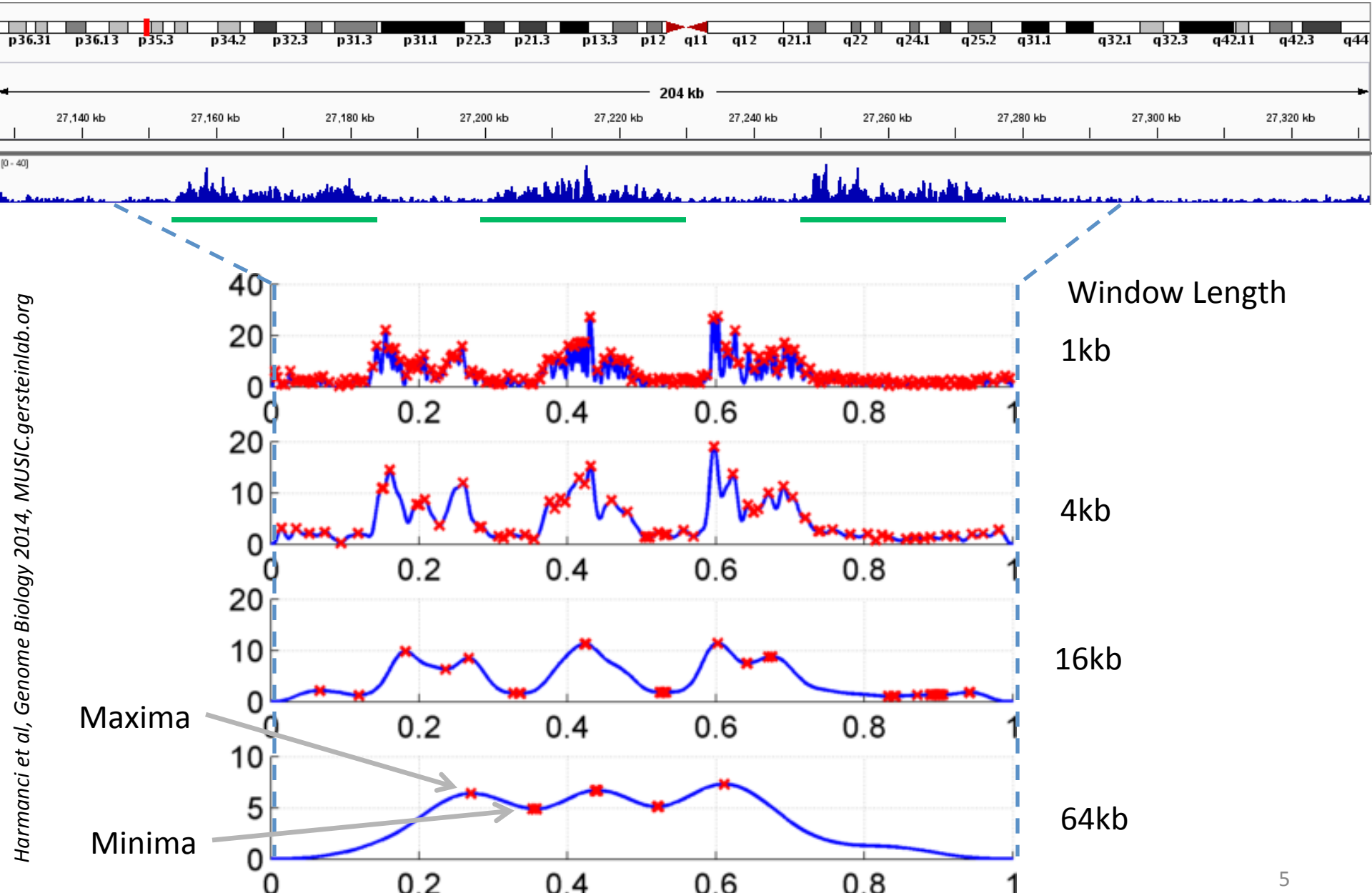


Significantly Enriched targets



Now an update: "PeakSeq 2" => MUSIC

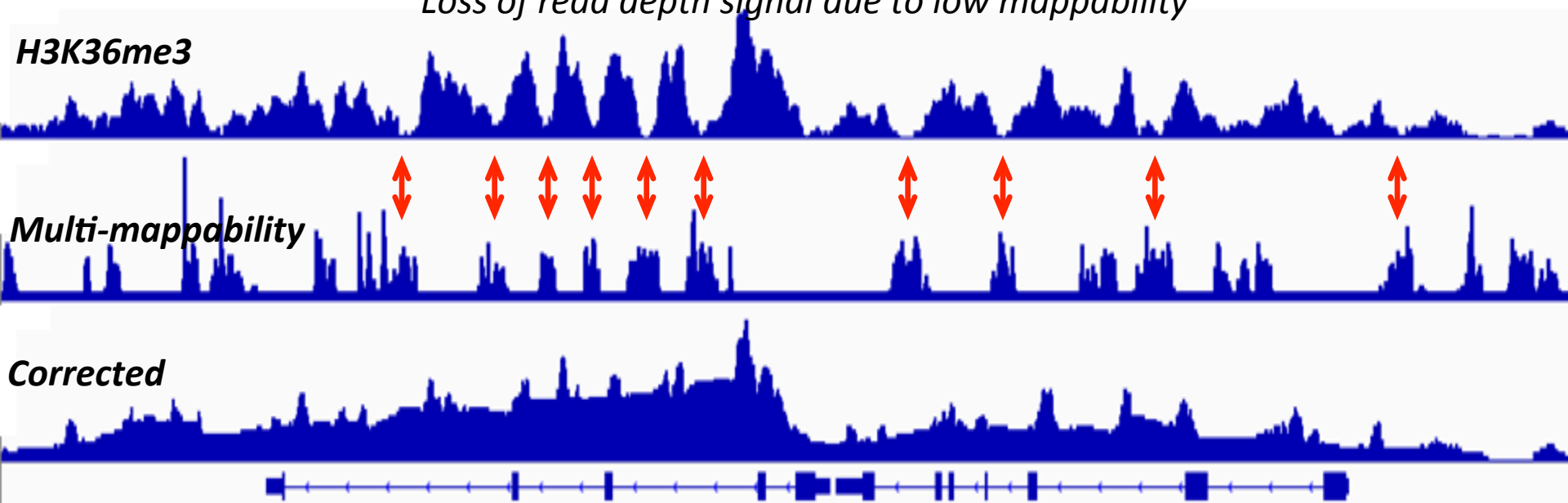
Multiscale Analysis, Minima/Maxima based Coarse Segmentation



Multi-mappability based Correction

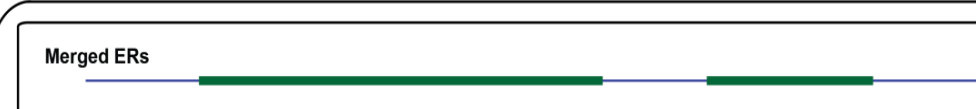
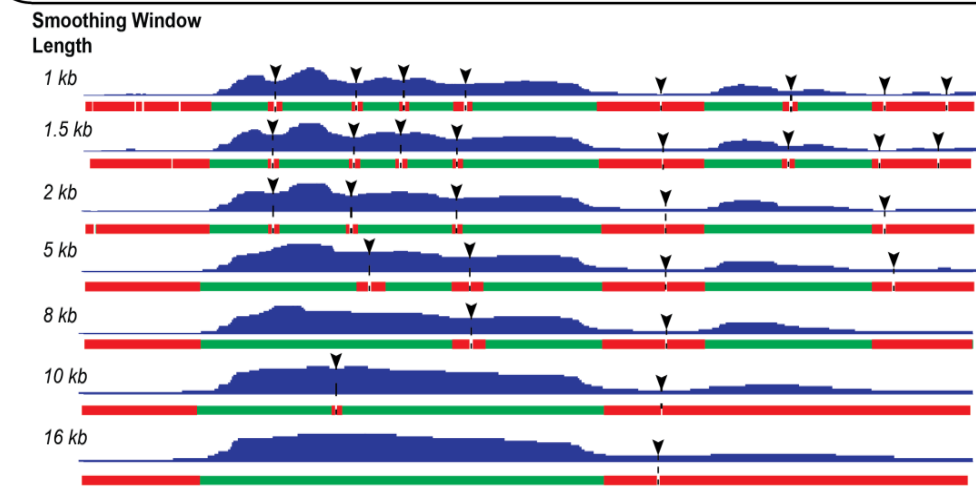
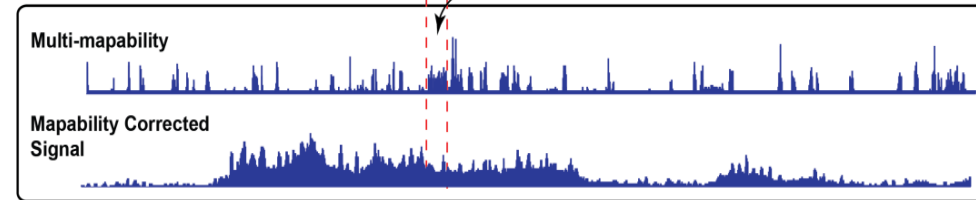
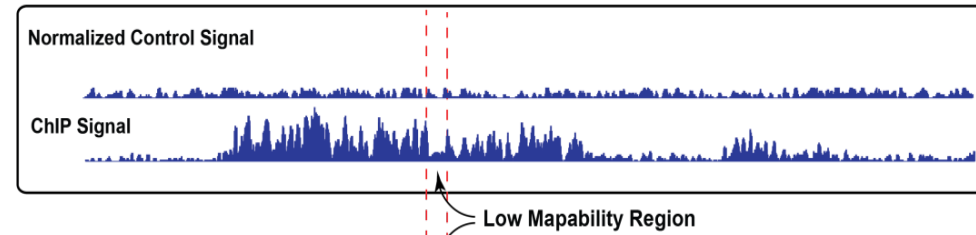
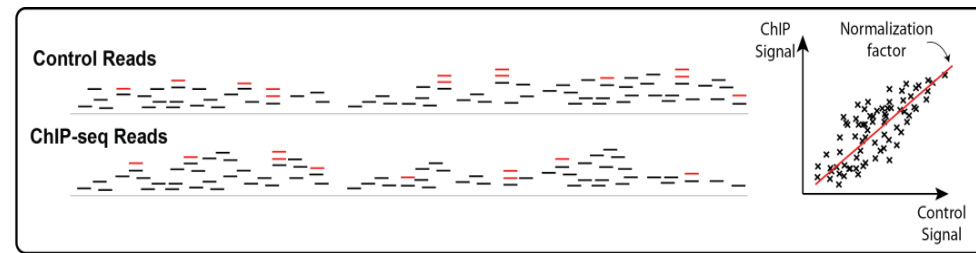
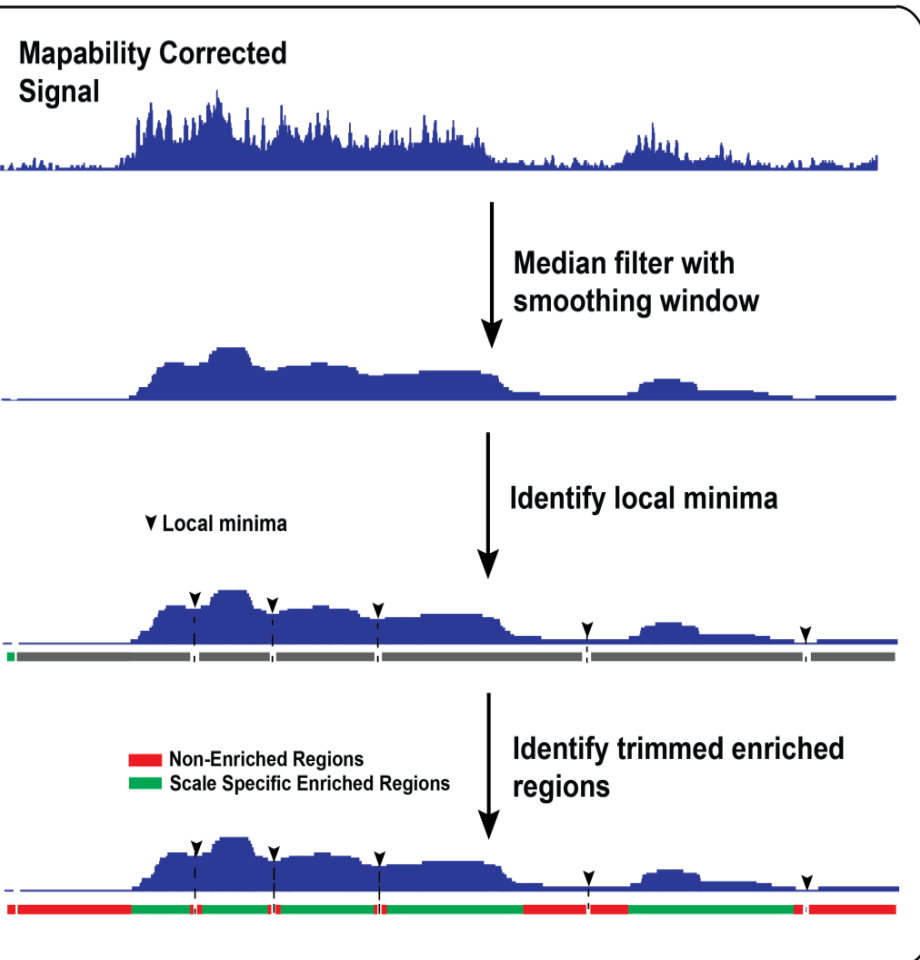
- Low mappability regions cause loss of signal and introduce burst-like noise
- To characterize the mappability of the genome, we build the **multi-mappability profile**
 - High multi-mappability signal \leftrightarrow Low mappability
- Correction Procedure:
 - “Whenever there is a lowly mappable position,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

Loss of read depth signal due to low mappability

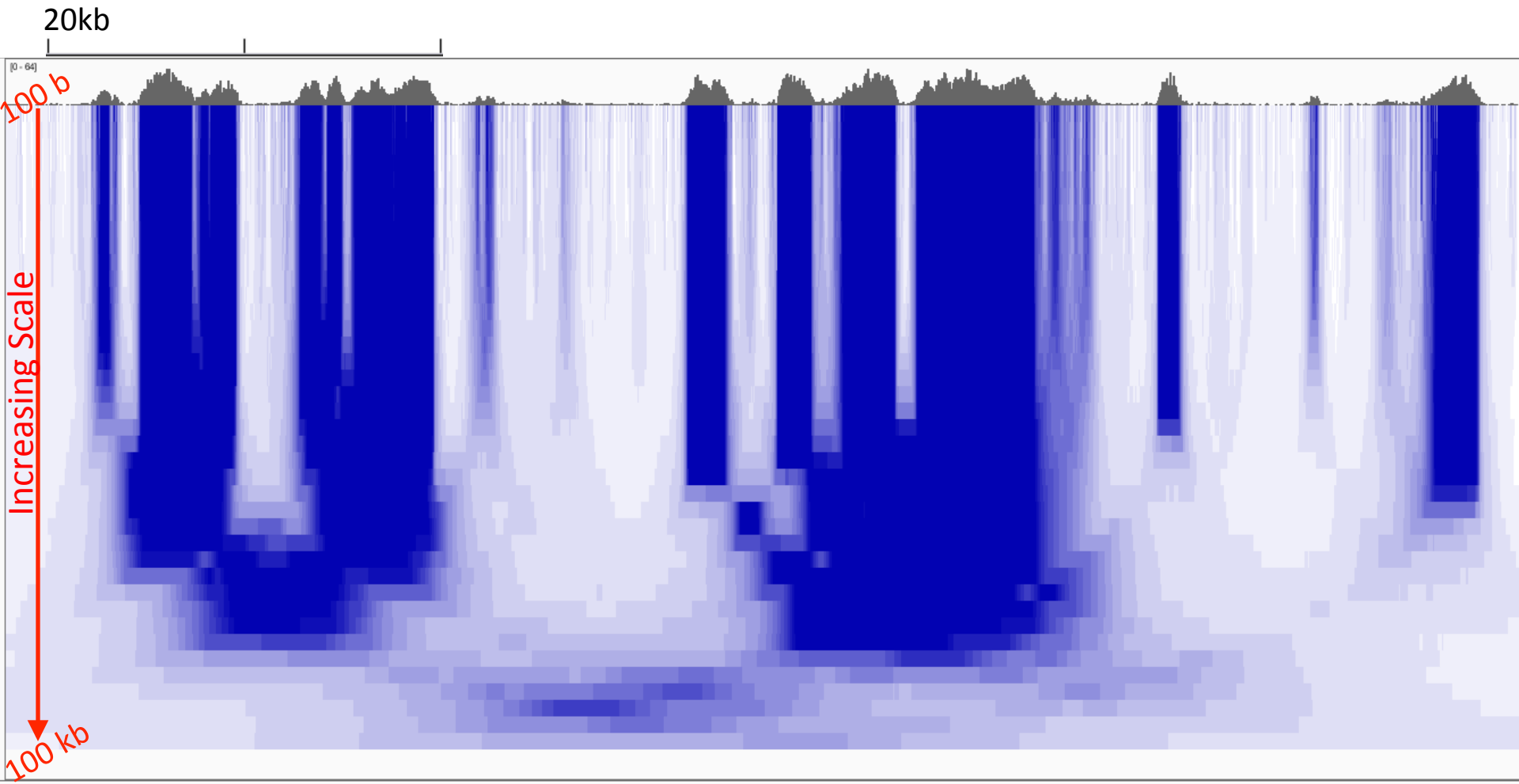


MUSIC.gersteinlab.org

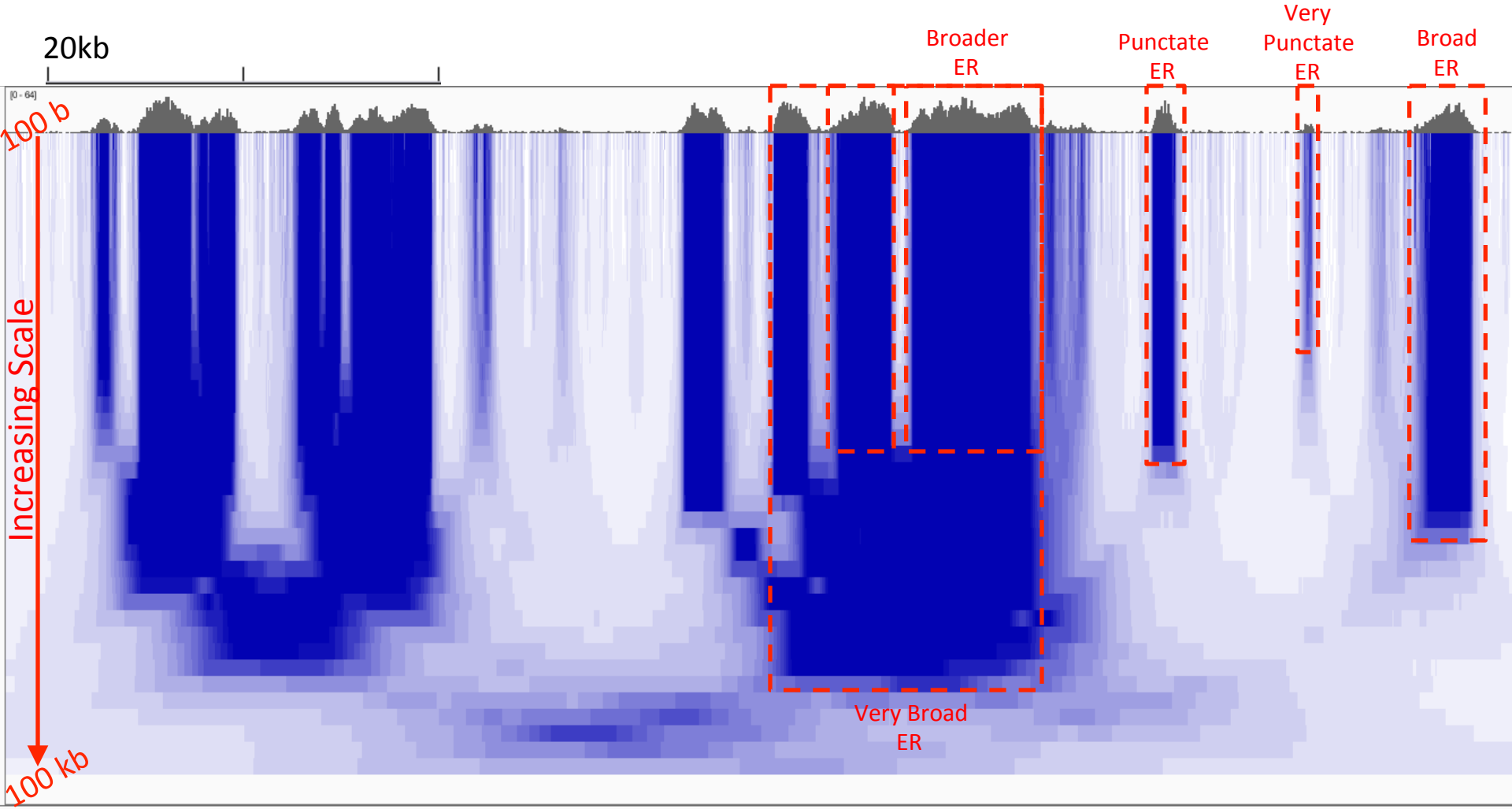
Algorithm



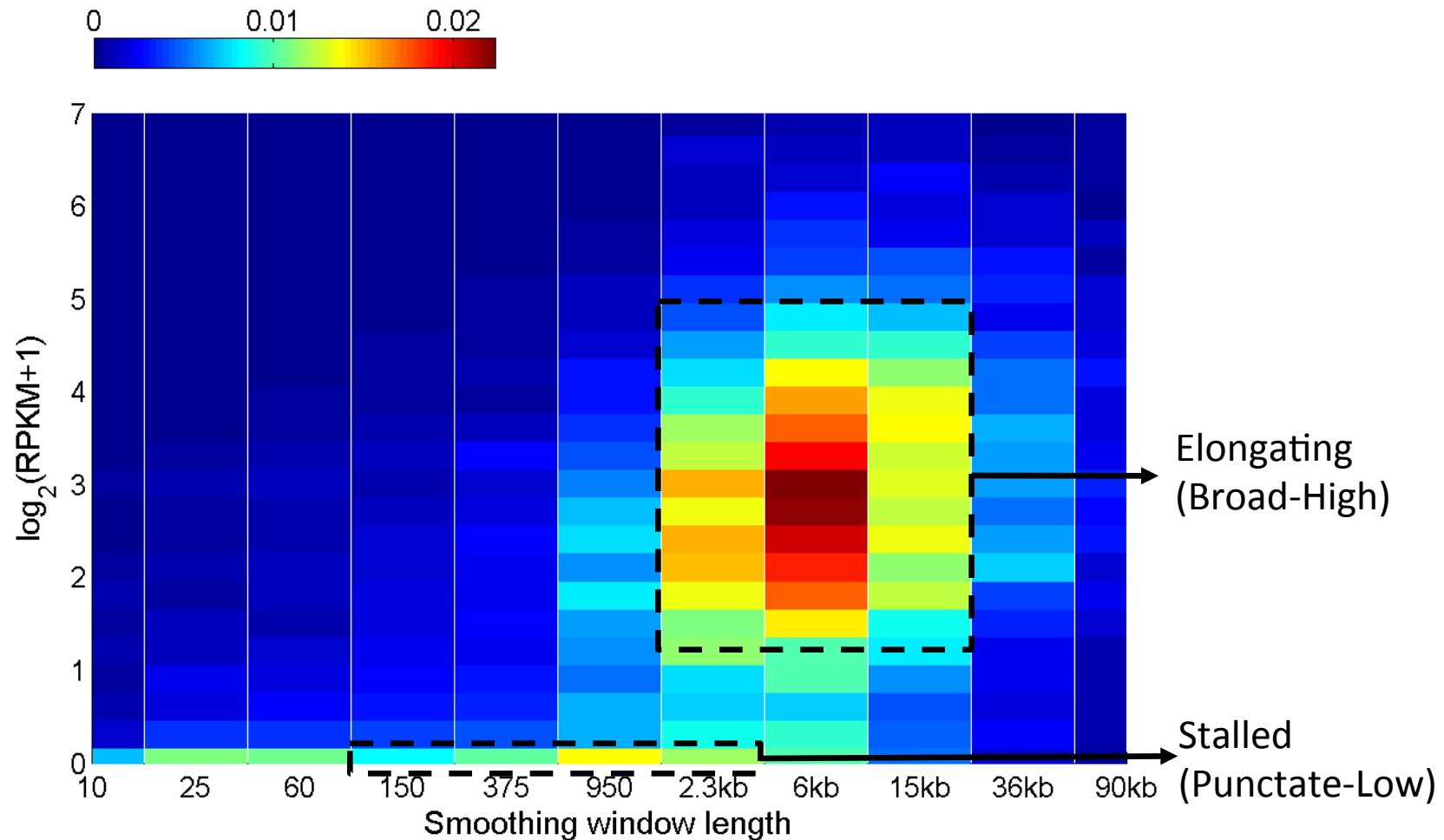
Multiscale Decomposition



Multiscale Decomposition

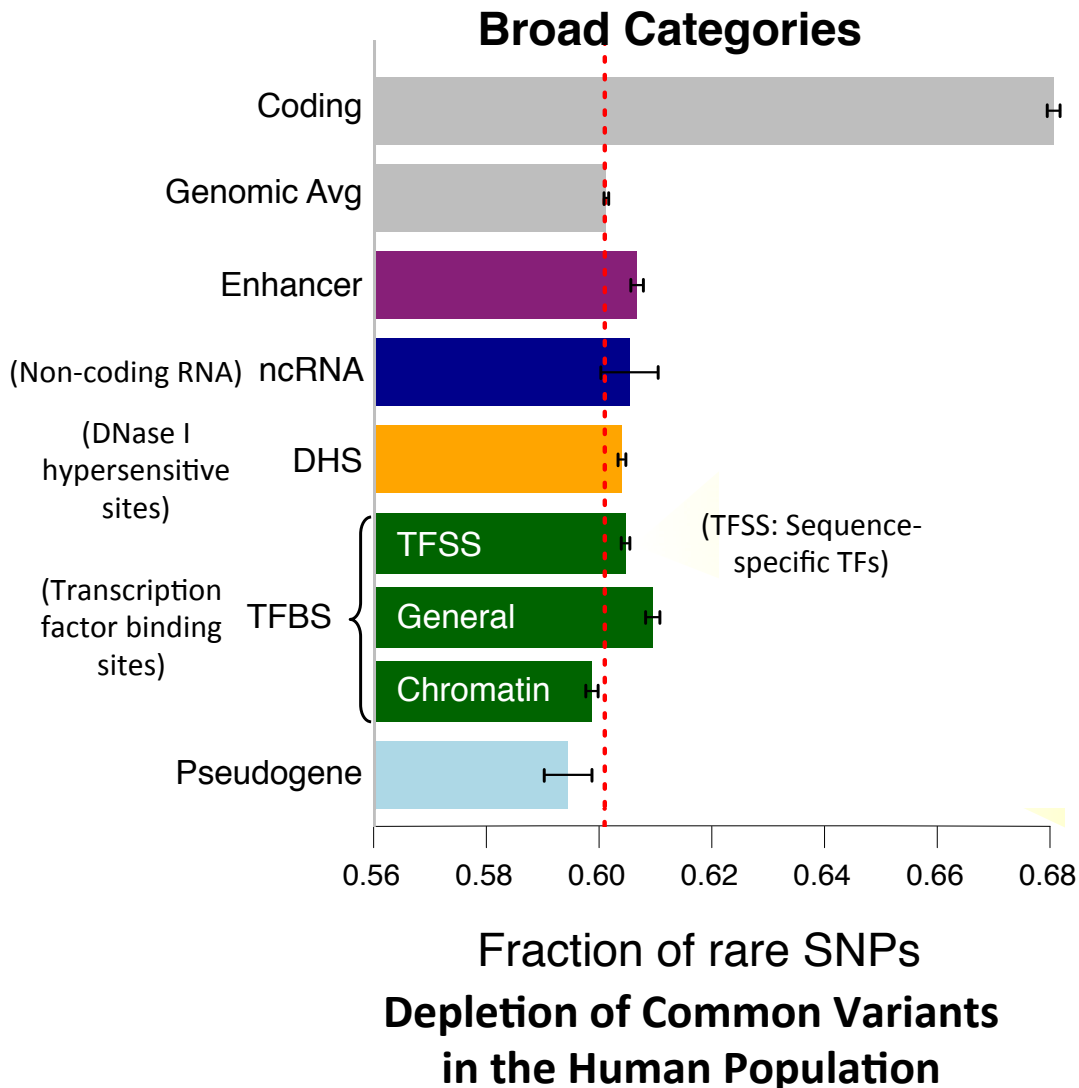


Pol2 Scale Spectrum over Protein Coding Genes Reveal Different Patterns of Gene Activity



Finding "Conserved" Sites:

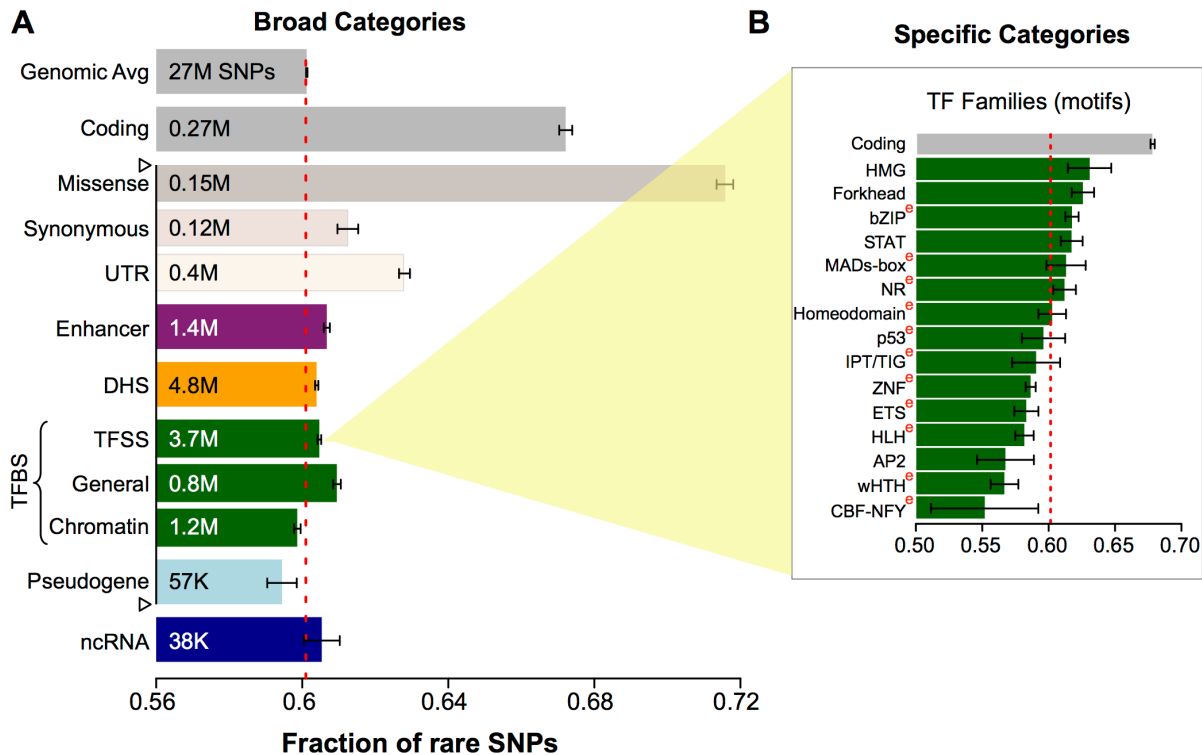
Negative selection in non-coding elements



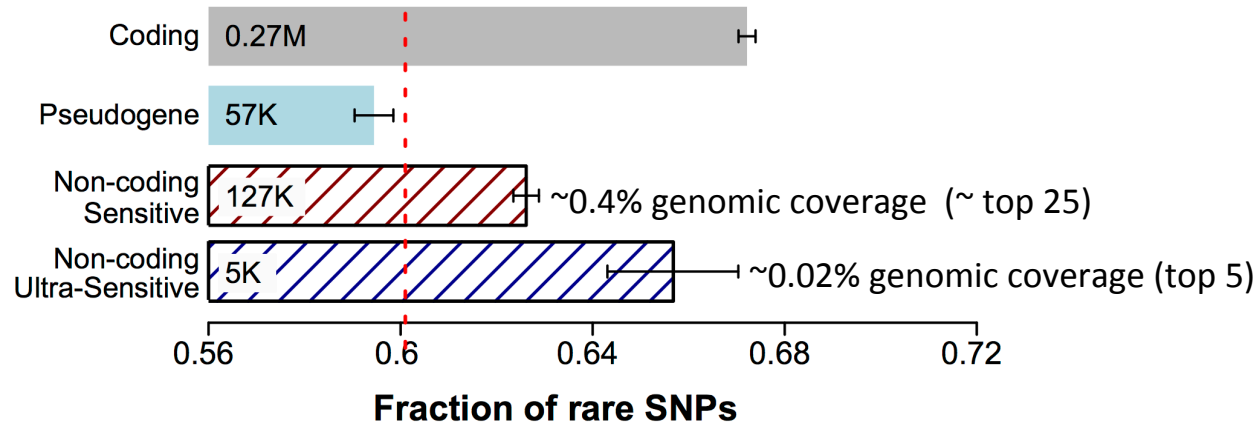
- Broad categories of regulatory regions under negative selection
- Consistent with previous studies

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

Differential selective constraints among sub-categories



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

Human Genome Analysis:
Progressive summarization of large-scale data,
to interpret mutations & dis-regulation in cancer

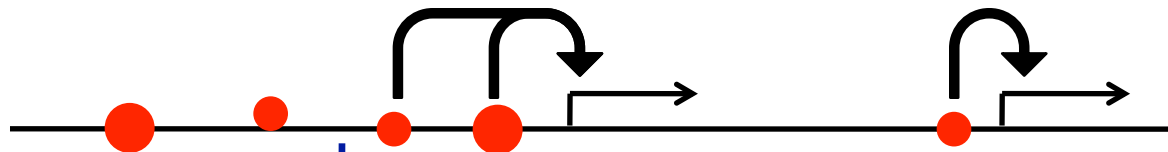
- **Summarizing Large-scale Genomic Information**
 - **1st Level Linear Annotation: Regulatory Sites**
 - Multi-scale "site" calling (with Music)
 - Finding small number of sites particularly sensitive to mutations
 - **2nd Level Network Annotation**
 - Building a network from the linear annotation
 - More connectivity = more constraint => highlights hubs
- **Using Summaries to Interpret Alterations in Cancer**
 - **FunSeq software tool for mutation prioritization**
 - Systematically weighting all the features, for non-coding prioritization
 - Summarizing large data context into simple "Core Score File"
 - **Loregic: Logic-gate analysis of regulation**
 - Recasting the regulatory network as a collection of gates
 - Different gate structure in cancer, dominated by particular driver TFs

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

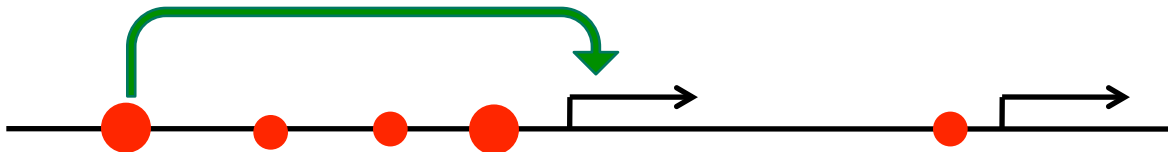
Regulatory elements



Assigning proximal sites to target genes

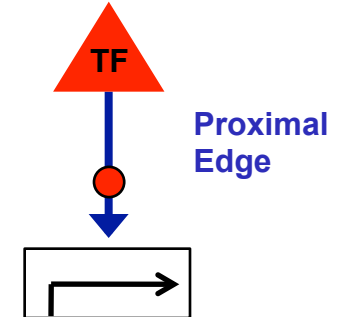


Assigning distal sites to targets

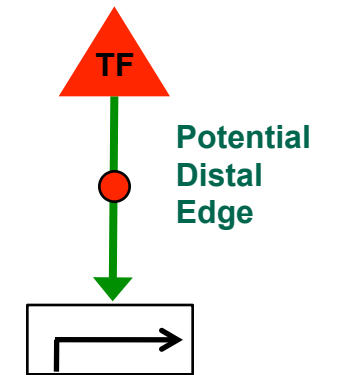


~500K Edges

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

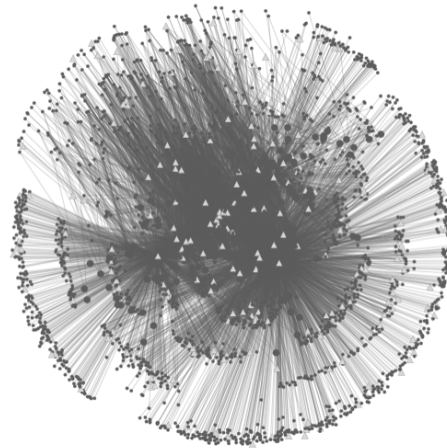


Proximal Edge

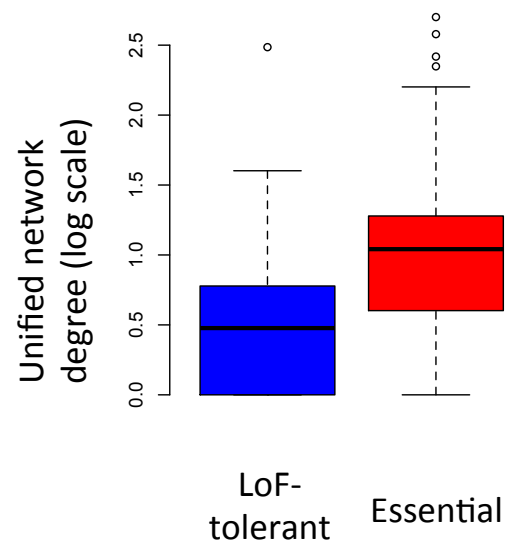


Potential Distal Edge

~26K Edges; ~
5K per cell line

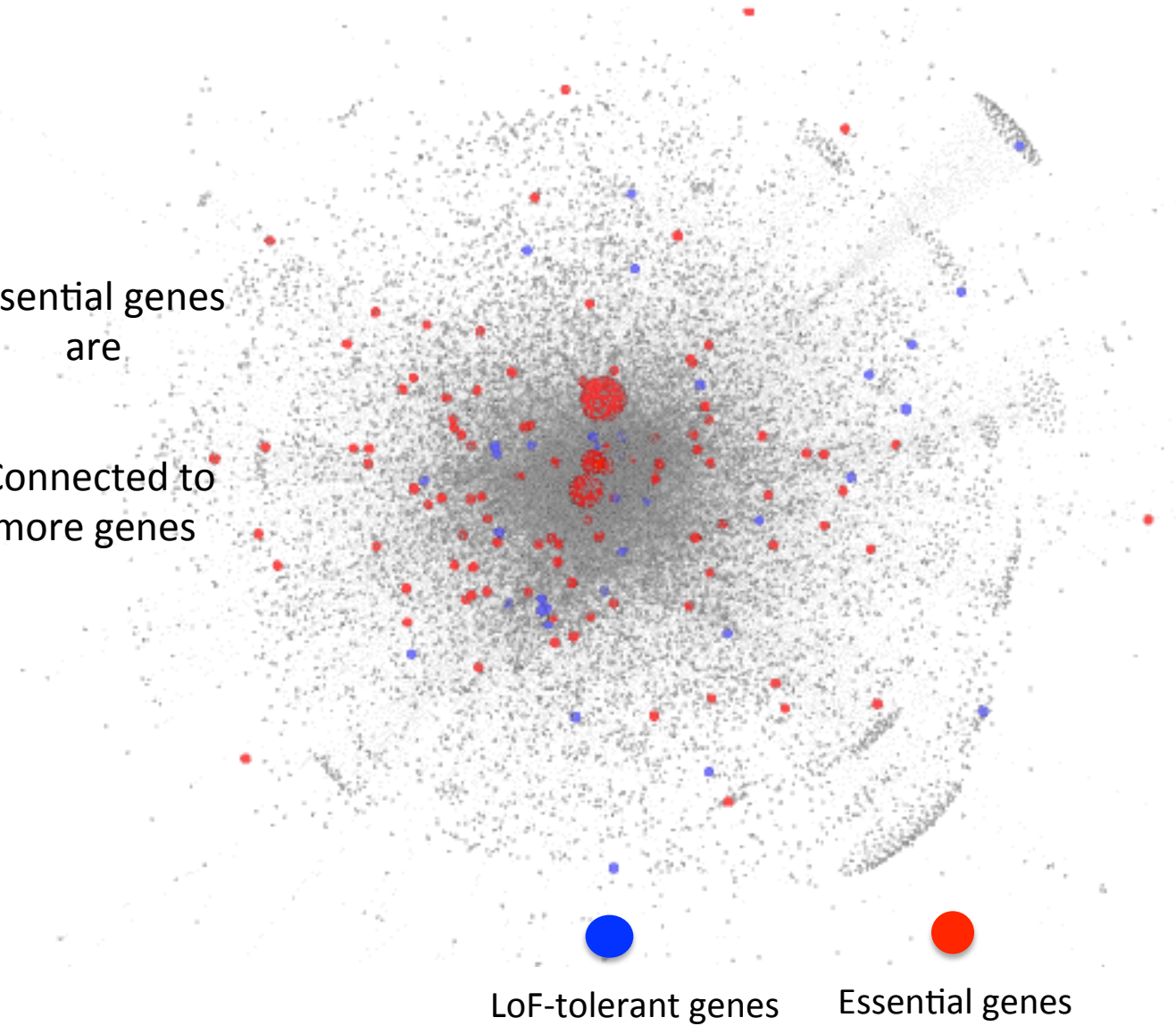


Hubs are conserved



Essential genes are

Connected to more genes



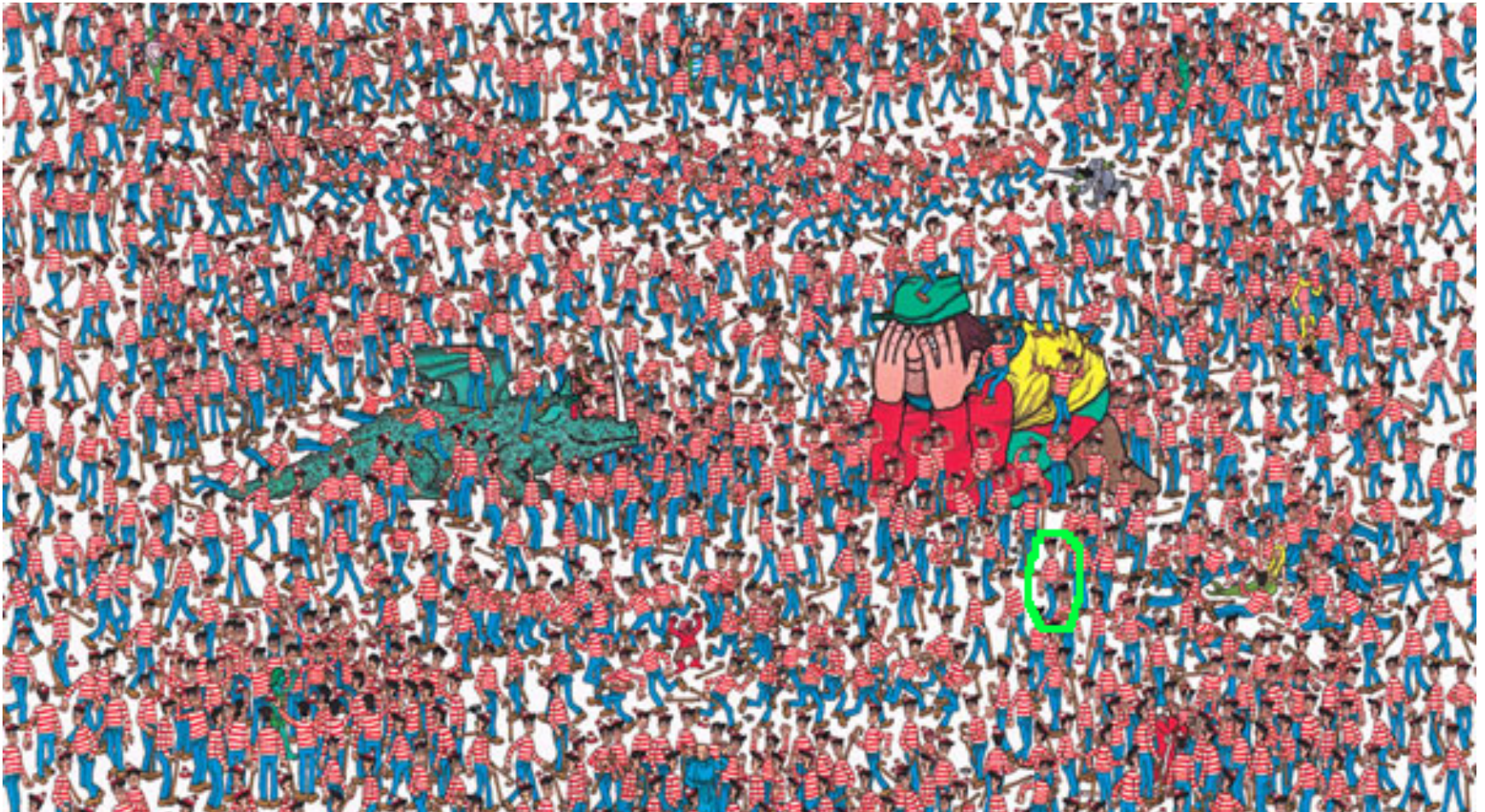
(More Connectivity = More Constraint)

Human Genome Analysis:
Progressive summarization of large-scale data
to interpret mutations & dis-regulation in cancer

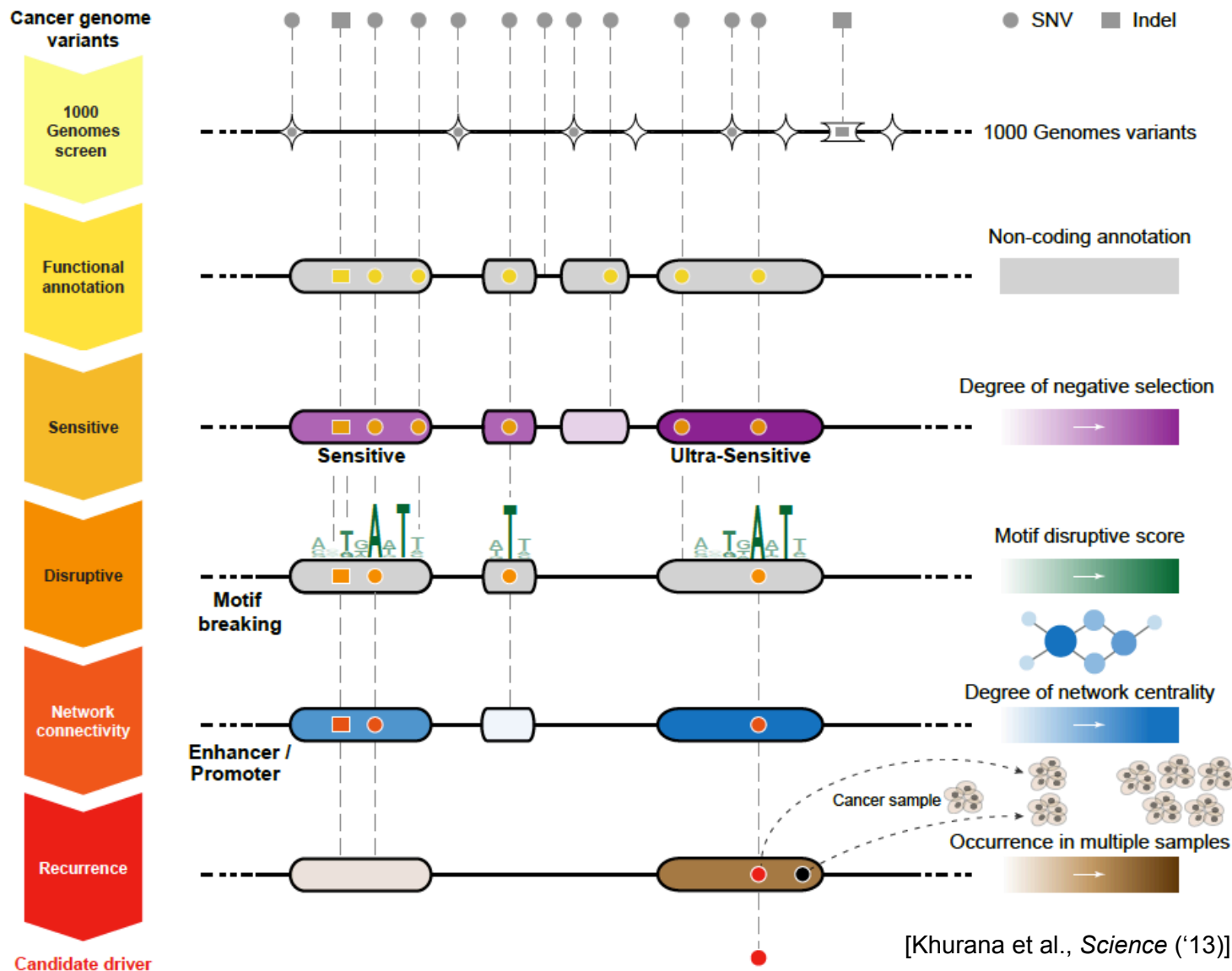
- **Summarizing Large-scale Genomic Information**
 - **1st Level Linear Annotation: Regulatory Sites**
 - Multi-scale "site" calling (with Music)
 - Finding small number of sites particularly sensitive to mutations
 - **2nd Level Network Annotation**
 - Building a network from the linear annotation
 - More connectivity = more constraint => highlights hubs
- **Using Summaries to Interpret Alterations in Cancer**
 - **FunSeq software tool for mutation prioritization**
 - Systematically weighting all the features, for non-coding prioritization
 - Summarizing large data context into simple "Core Score File"
 - **Loregic: Logic-gate analysis of regulation**
 - Recasting the regulatory network as a collection of gates
 - Different gate structure in cancer, dominated by particular driver TFs

Where is Waldo?

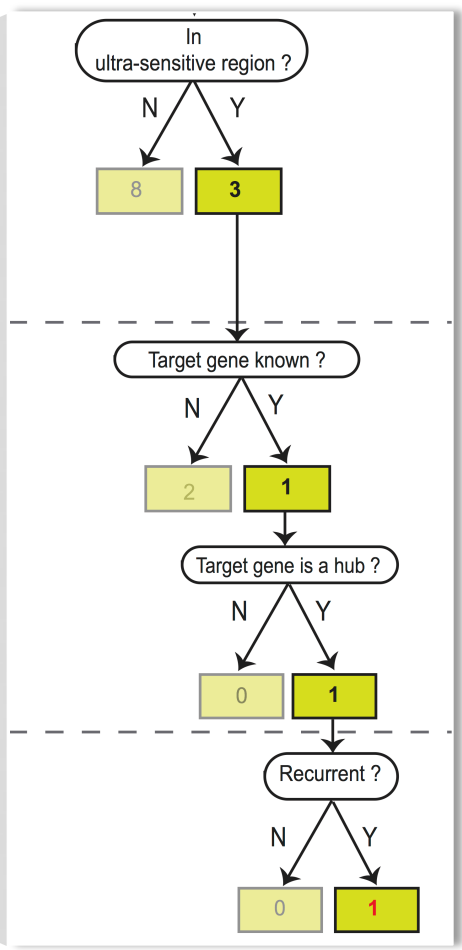
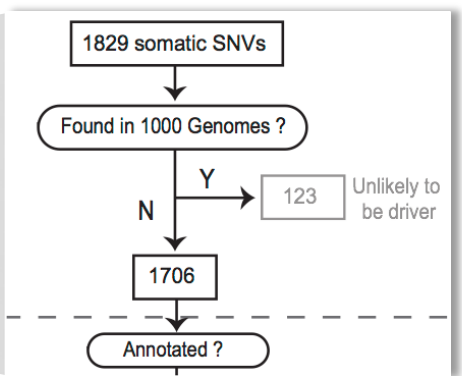
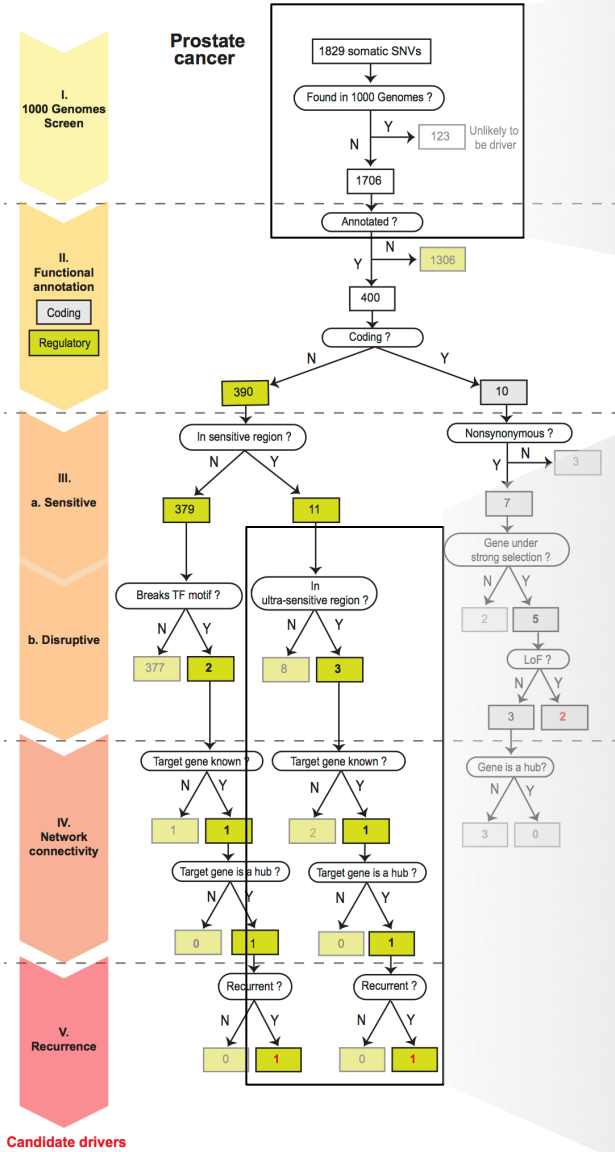
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



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



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





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

Variant Prioritization

Weighted scoring scheme

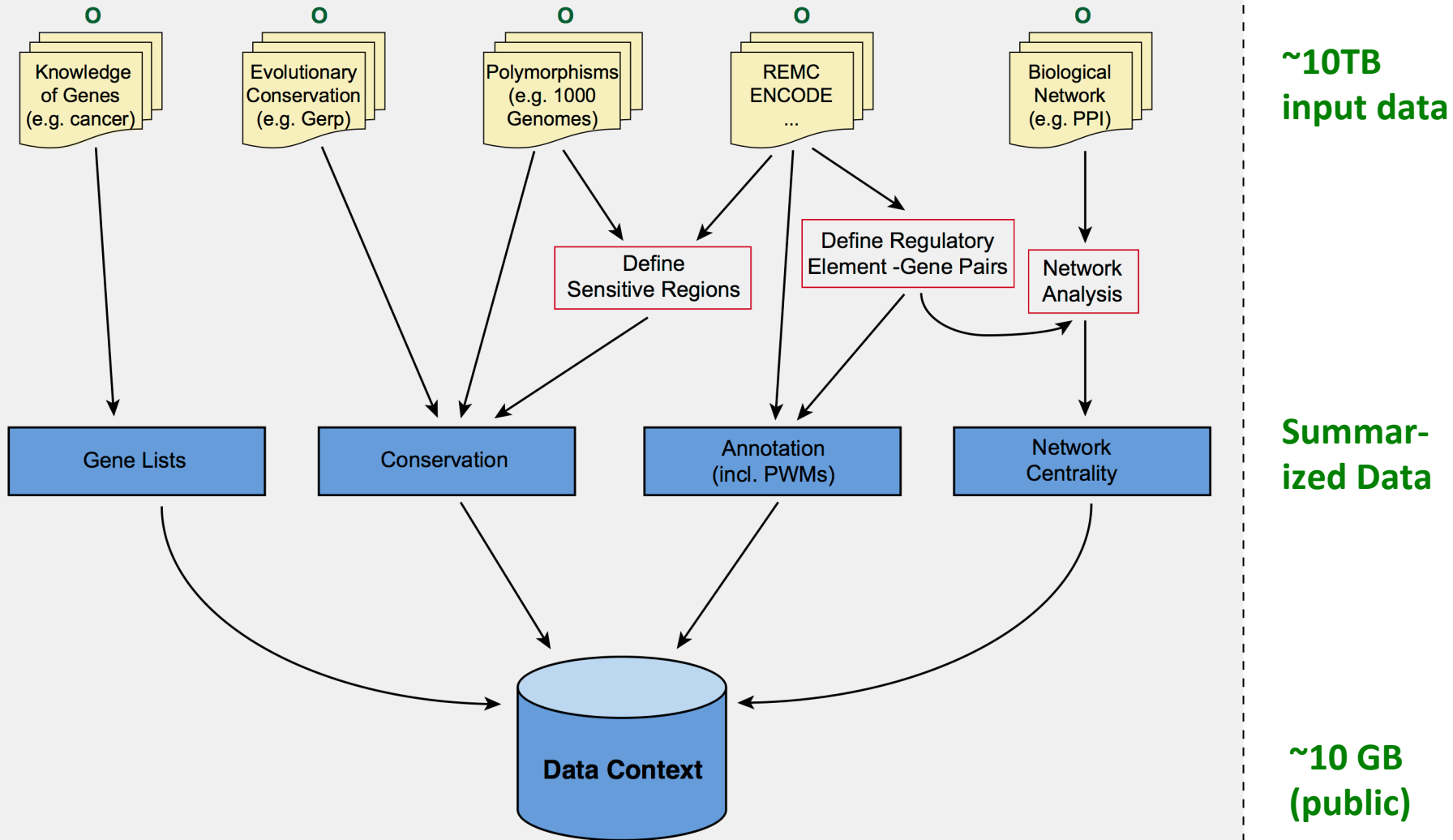
Highlighting variants

User Cancer Variants

Variant Reports

FunSeq.gersteinlab.org

Data context



Data Context (~10Gb, public)

Variant Prioritization

Weighted scoring scheme

Functional annotations

Regulatory regions
HOT regions

Conservation

Evolutionary
Human-specific

Nucleotide-level analysis

Motif-breaking
Motif-gaining

Network analysis

Linking regulatory elements with genes
Centrality

Core scores

Recurrence module

Regulatory elements

Recurrence DB *

Recurrent elements

Variants in recurrent elements

Final scores

Highlighting variants with additional features

Knowledge of genes

Cancer genes
DNA repair genes
Differentially expressed genes
...

User annotations

Sample-specific epigenetic / open chromatin profiles
...

Compute variant prioritization for every possible substitution in the genome (3 subs * 3 Gb). Not specific to any study (or cancer) & not private

"Core Score File"

User Variants (small, ~5000 & pot. private)

Weighted scoring scheme

Functional annotations

Regulatory regions
HOT regions

Conservation

Evolutionary
Sensitive/Ultra-sensitive regions

Network study

Linking regulatory elements with genes
Hubs

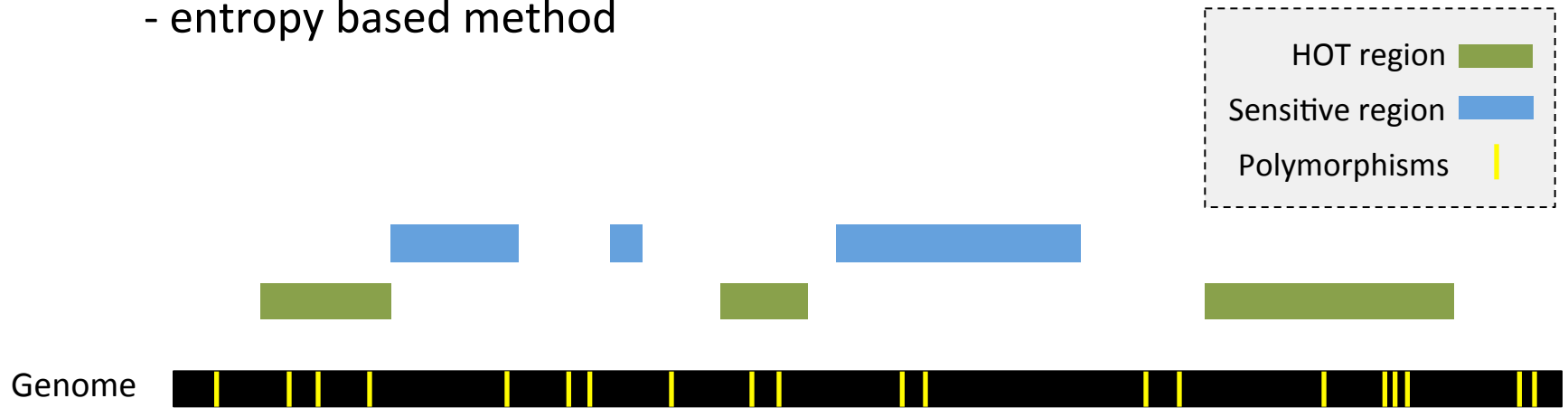
Nucleotide-level analysis

Motif-breaking
Motif-gaining

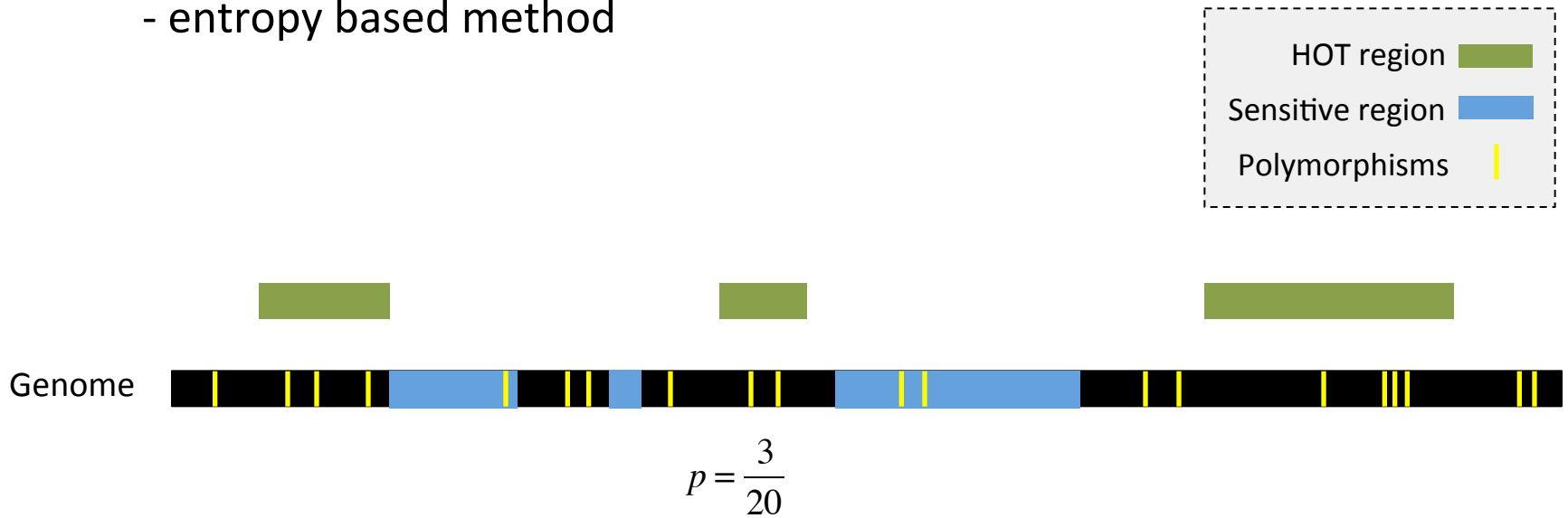
↓
Variant core scores

●----- **"Core Score File"**

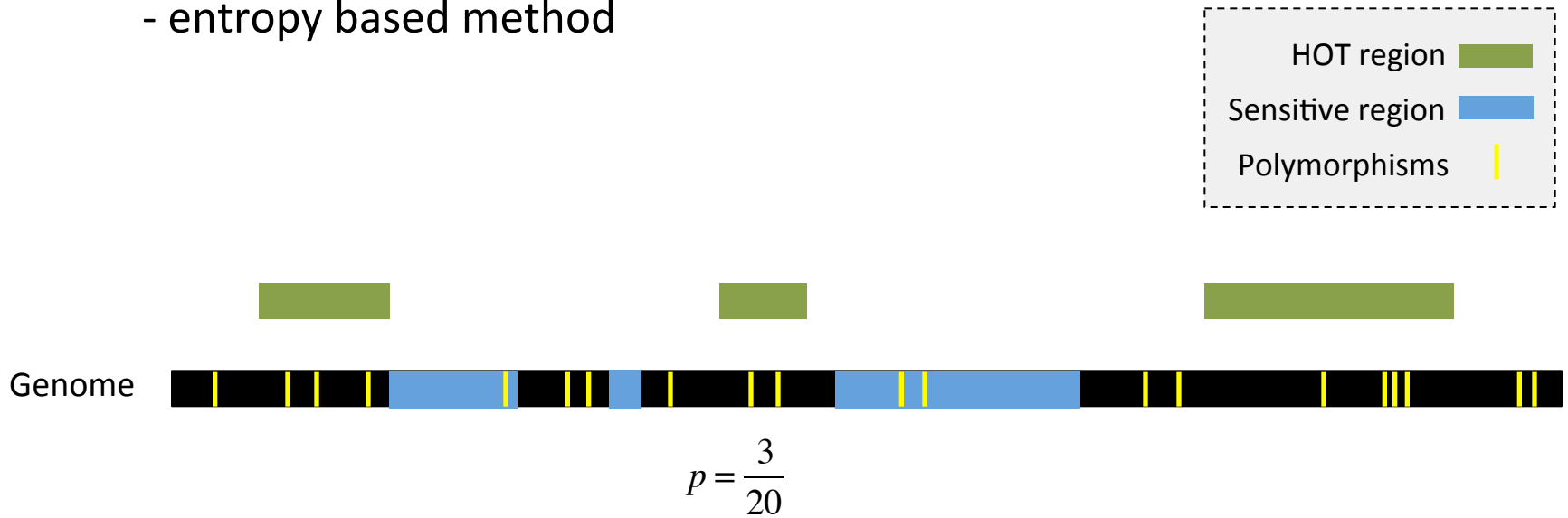
- Feature weight
 - Weighted with mutation patterns in natural polymorphisms (features frequently observed weight less)
 - entropy based method



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



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



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

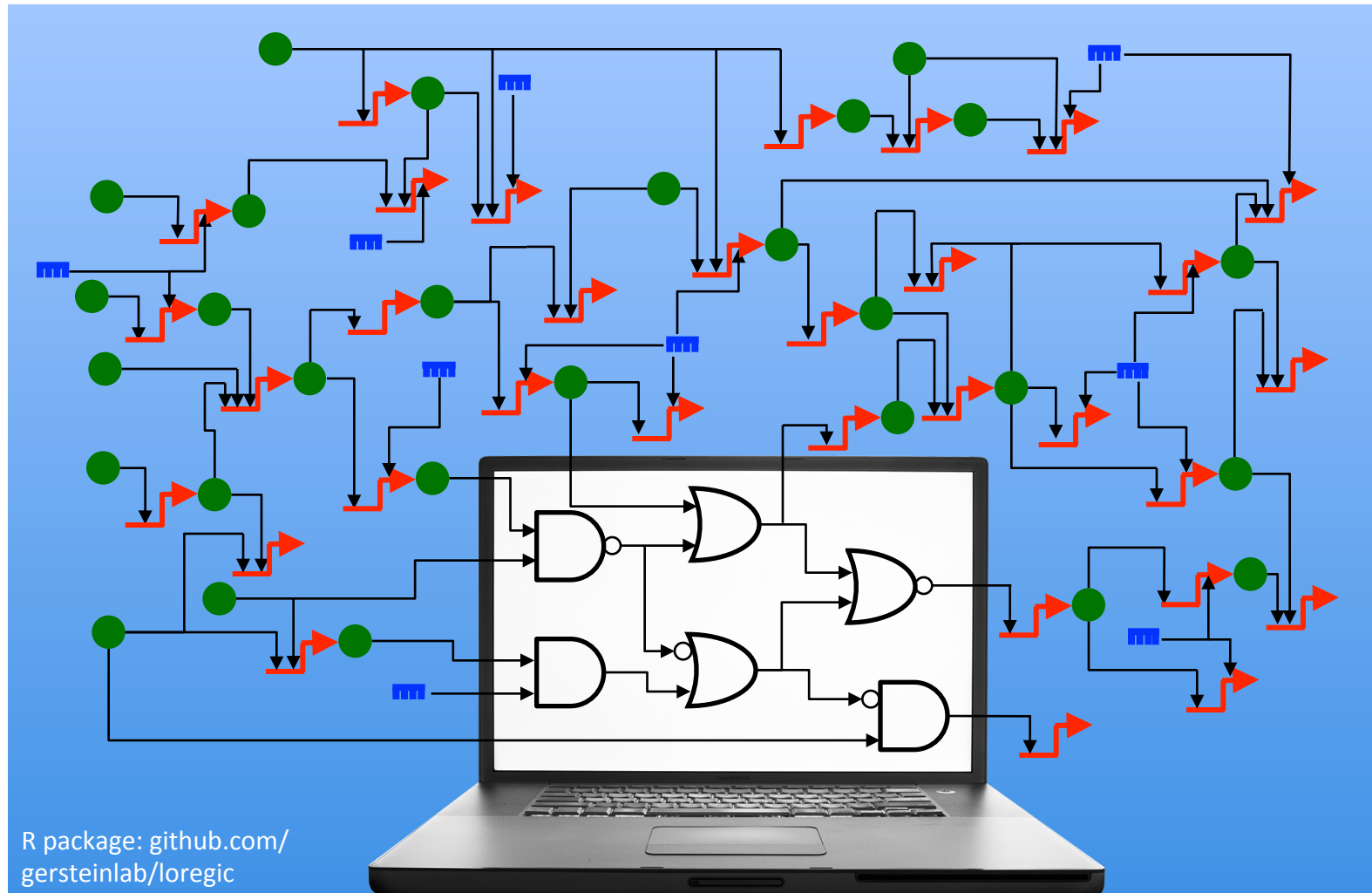
$p \uparrow$ $w_d \downarrow$ $p = \text{probability of the feature overlapping natural polymorphisms}$

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

Human Genome Analysis:
Progressive summarization of large-scale data
to interpret mutations & dis-regulation in cancer

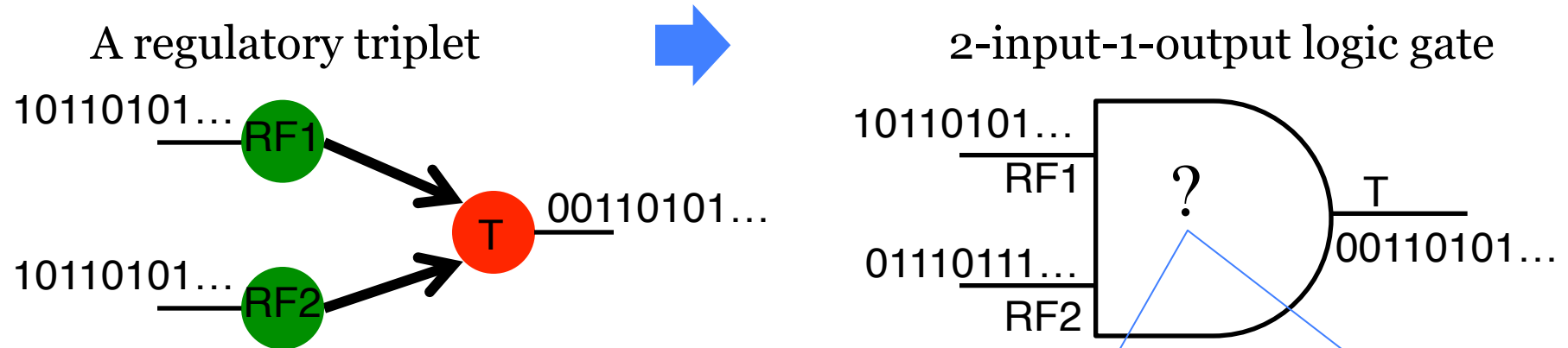
- **Summarizing Large-scale Genomic Information**
 - **1st Level Linear Annotation: Regulatory Sites**
 - Multi-scale "site" calling (with Music)
 - Finding small number of sites particularly sensitive to mutations
 - **2nd Level Network Annotation**
 - Building a network from the linear annotation
 - More connectivity = more constraint => highlights hubs
- **Using Summaries to Interpret Alterations in Cancer**
 - **FunSeq software tool for mutation prioritization**
 - Systematically weighting all the features, for non-coding prioritization
 - Summarizing large data context into simple "Core Score File"
 - **Loregic: Logic-gate analysis of regulation**
 - Recasting the regulatory network as a collection of gates
 - Different gate structure in cancer, dominated by particular driver TFs

Loregic: A method to characterize the cooperative logic of regulatory factors



Wang, et al., *PLoS Computational Biology*, 2015

Modeling cooperativity between RFs to target gene using logic gates



0 – gene off
 1 – gene on
 after binarizing gene
 expression data*

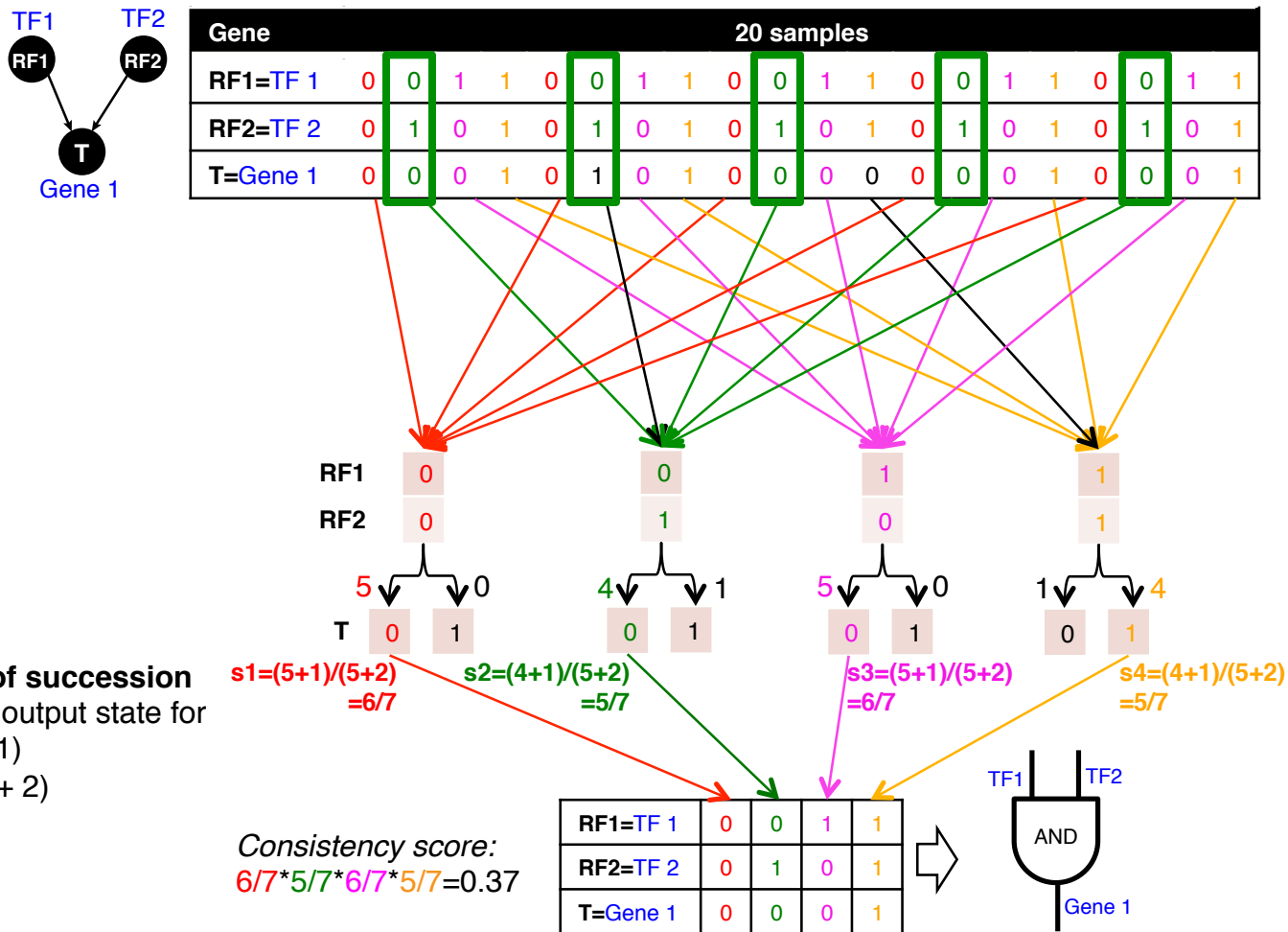
Input type (RF1, RF2)	RF1	0	0	1	1	} Binarized expression
	RF2	0	1	0	1	
Output	T	X	X	X	X	

X can be 0 or 1, so there are $2^4=16$ possible output combinations, each of which corresponds to a unique 2-input-1-output logic gate



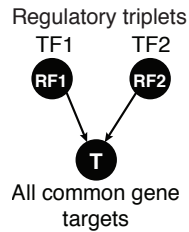
*BoolNet, R package

An example: selection of the best-matched logic gate



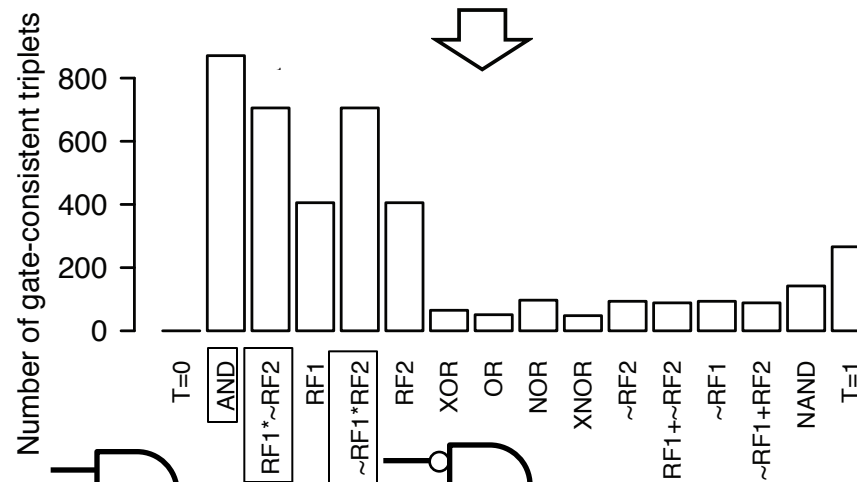
Wang, et al., *PLoS Computational Biology*, 2015

Application 1 – transcription factor cooperativity in Yeast cell cycle

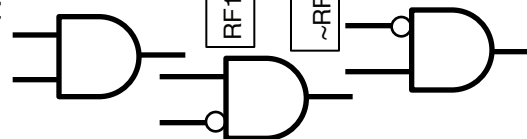


Target gene	2464
TF	176
Triplet	39,011
Time point	59

Triplet ID	RF1	RF2	Common Target Gene (T)	Matched logic gate
1	YHR084W	YBR083W	YBR082C	AND
2	YKL112W	YIL131C	YMR198W	OR
...
39011	YOR113W	YBL103C	YDR042C	XOR



AND-like gates



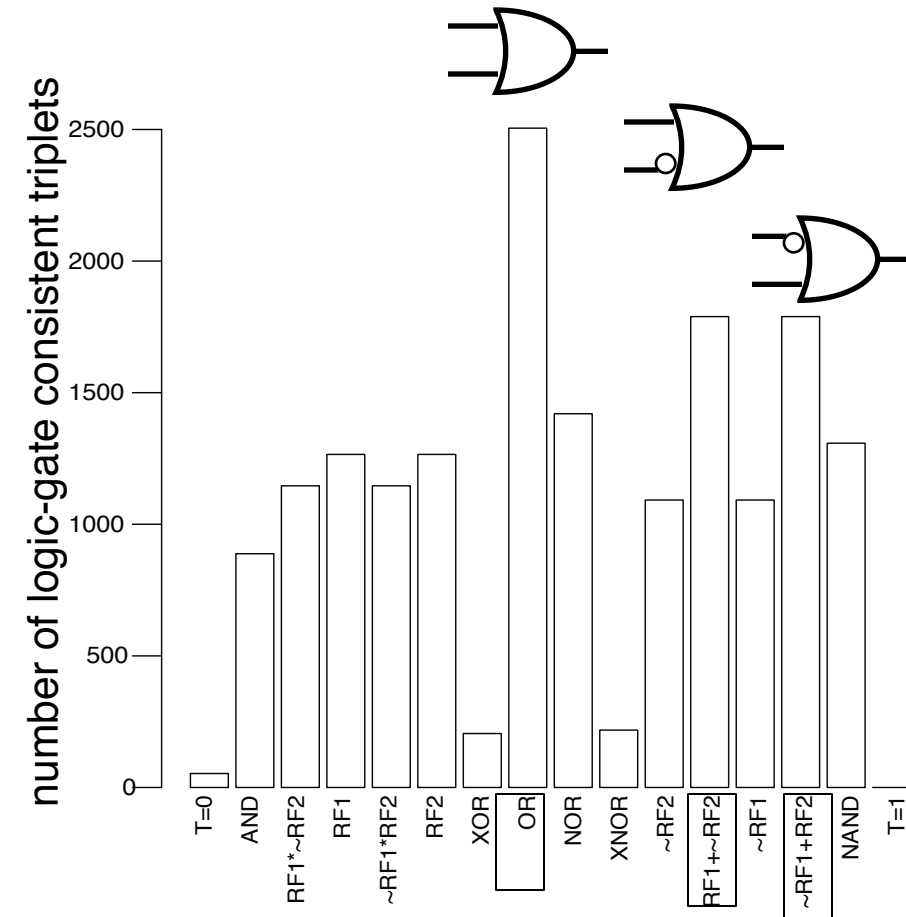
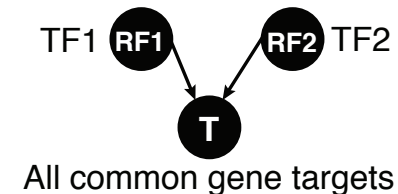
Wang, et al., *PLoS Computational Biology*, 2015

Application – transcription factor cooperativity in Acute Myeloid Leukemia (AML)

Target gene	1824
TF (ENCODE)	70
Triplet	50,865
Patient (TCGA)	197

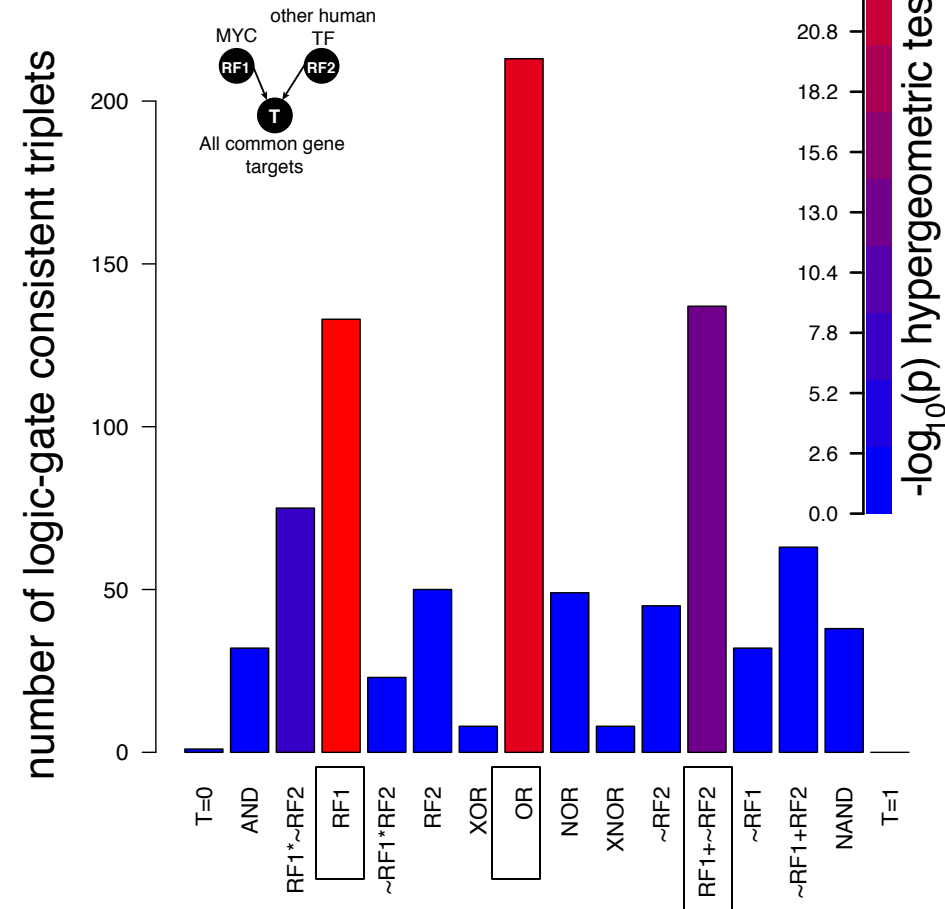
Human TF-TF-target



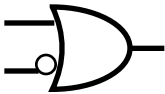
RF1	RF2	Common Target Gene (T)	Matched logic gate
ATF3	BDP1	YPEL1	AND
MYC	BCL3	BCR	T=RF1
ATF3	BRF2	AIF1L	AND
...



Cancer-related TF, MYC universally amplifies target expression

2,153 (RF1=MYC, RF2=other TFs, T=all common targets) triplets



- RF1 
- **OR**(RF1, RF2) 
- **OR**(RF1, **NOT** RF2) 



High expression of MYC is sufficient for high target gene expression

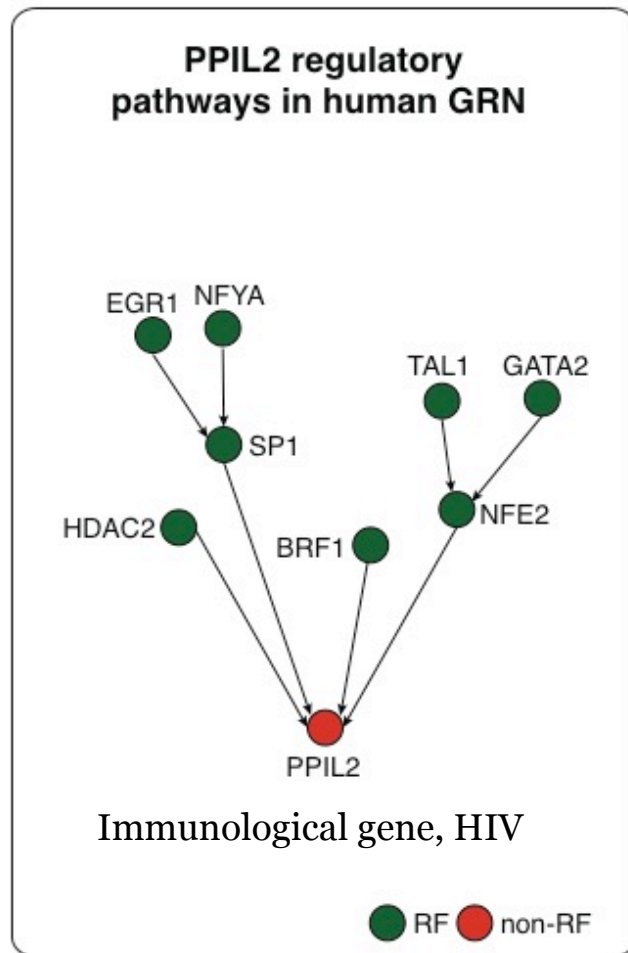
c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells



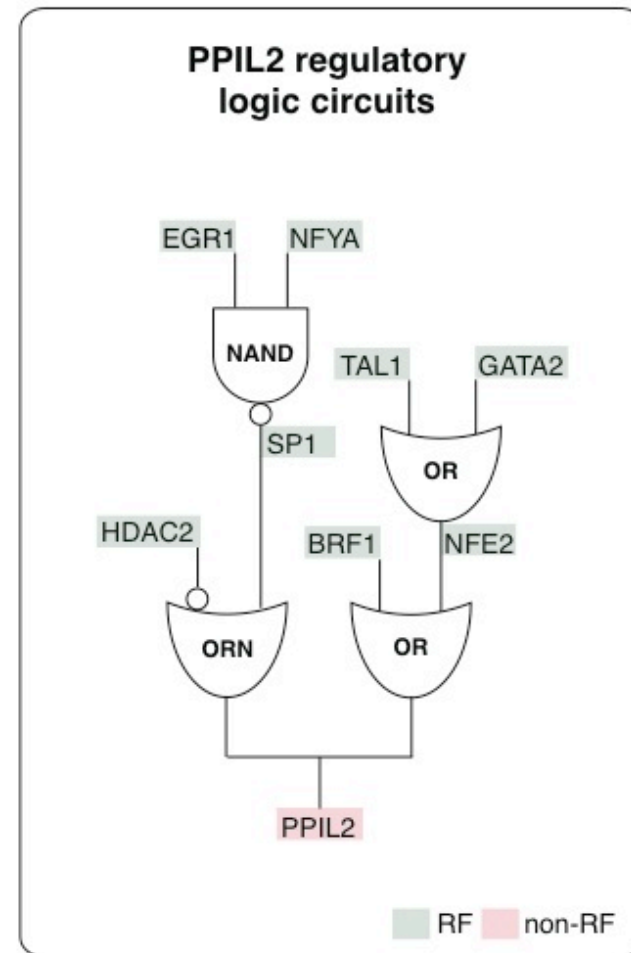
Zuqin Nie,^{1,6} Gangqing Hu,^{2,6} Gang Wei,² Kairong Cui,² Arito Yamane,³ Wolfgang Resch,³ Ruoning Wang,⁴ Douglas R. Green,⁴ Lino Tessarollo,⁵ Rafael Casellas,³ Keji Zhao,^{2,*} and David Levens^{1,*}

Wang, et al., *PLoS Computational Biology*, 2015

Gene regulatory pathways have logic-circuit behaviors



Loregic →



Wang, et al., *PLoS Computational Biology*, 2015

Human Genome Analysis:
Progressive summarization of large-scale data,
to interpret mutations & dis-regulation in cancer

- **Summarizing Large-scale Genomic Information**
 - **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 Summaries to Interpret Alterations in Cancer**
 - **FunSeq software tool for mutation prioritization**
 - Systematically weighting all the features, for non-coding prioritization
 - Summarizing large data context into simple "Core Score File"
 - **Loregic: Logic-gate analysis of regulation**
 - Recasting the regulatory network as a collection of gates
 - Different gate structure in cancer, dominated by particular driver TFs

Cancer Prioritization Acknowledgements

← ~50 people ← ~1000 “authors”

Functional
Interpretation
Subgroup



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. Dermizakis,
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



Hiring postdocs, see
GersteinLab.org/jobs



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
- github.com/gersteinlab/lorepic
 - D **Wang**, KK Yan, C Sisu, C Cheng, J Rozowsky, W Meyerson



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>

MUSIC makes music

- -get_multiscale_music: Generates a .wav file using the aggregate multiscale decomposition
- Listen to K562 H3K36me3 chromosome 1:
<http://archive.gersteinlab.org/proj/MUSIC/music/H3K36me3.mp3>
 - Telomeres are vocal, centromeres (46:00-53:00) are silent
- Listen K562 H3K4me3 chromosome 1:
<http://archive.gersteinlab.org/proj/MUSIC/music/H3K4me3.mp3>
 - More “clicky” than H3K36me3 with more punctate enriched regions