

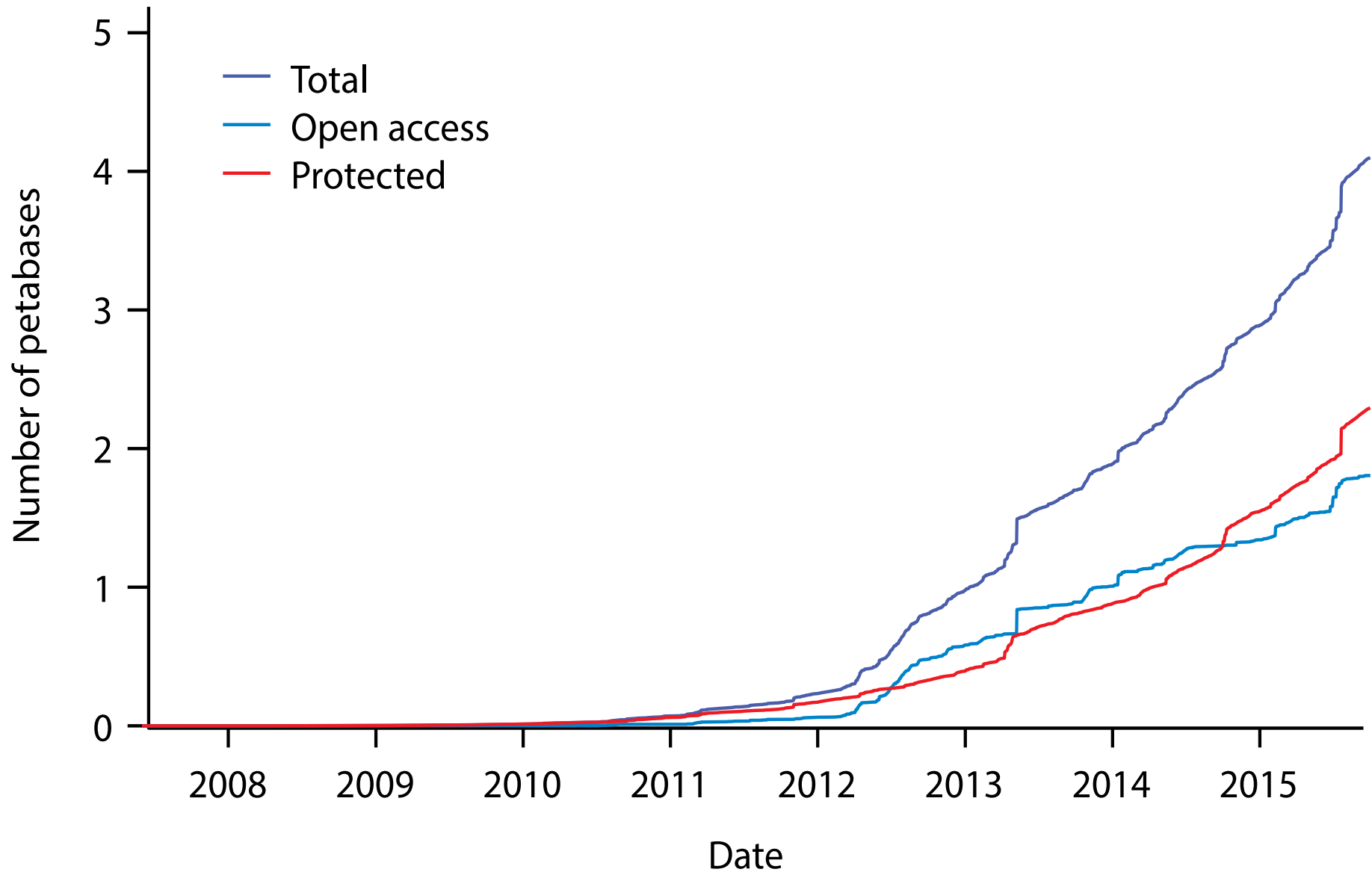
# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants



Mark Gerstein, Yale

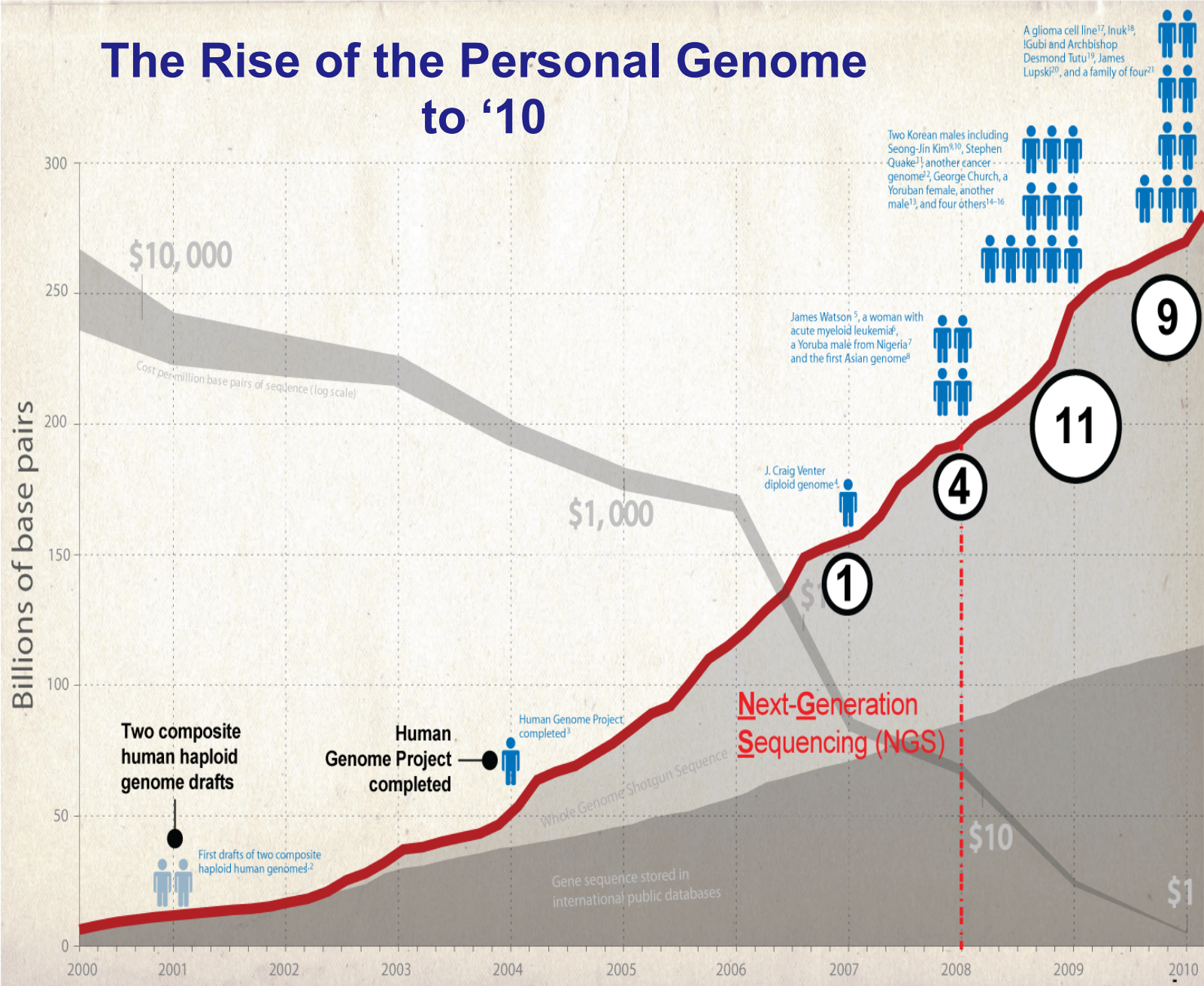
Slides freely downloadable from [Lectures.GersteinLab.org](http://Lectures.GersteinLab.org)  
& “tweetable” (via @markgerstein). See last slide for more info.

## Increase in number of bases in SRA, Peta-scale after '10



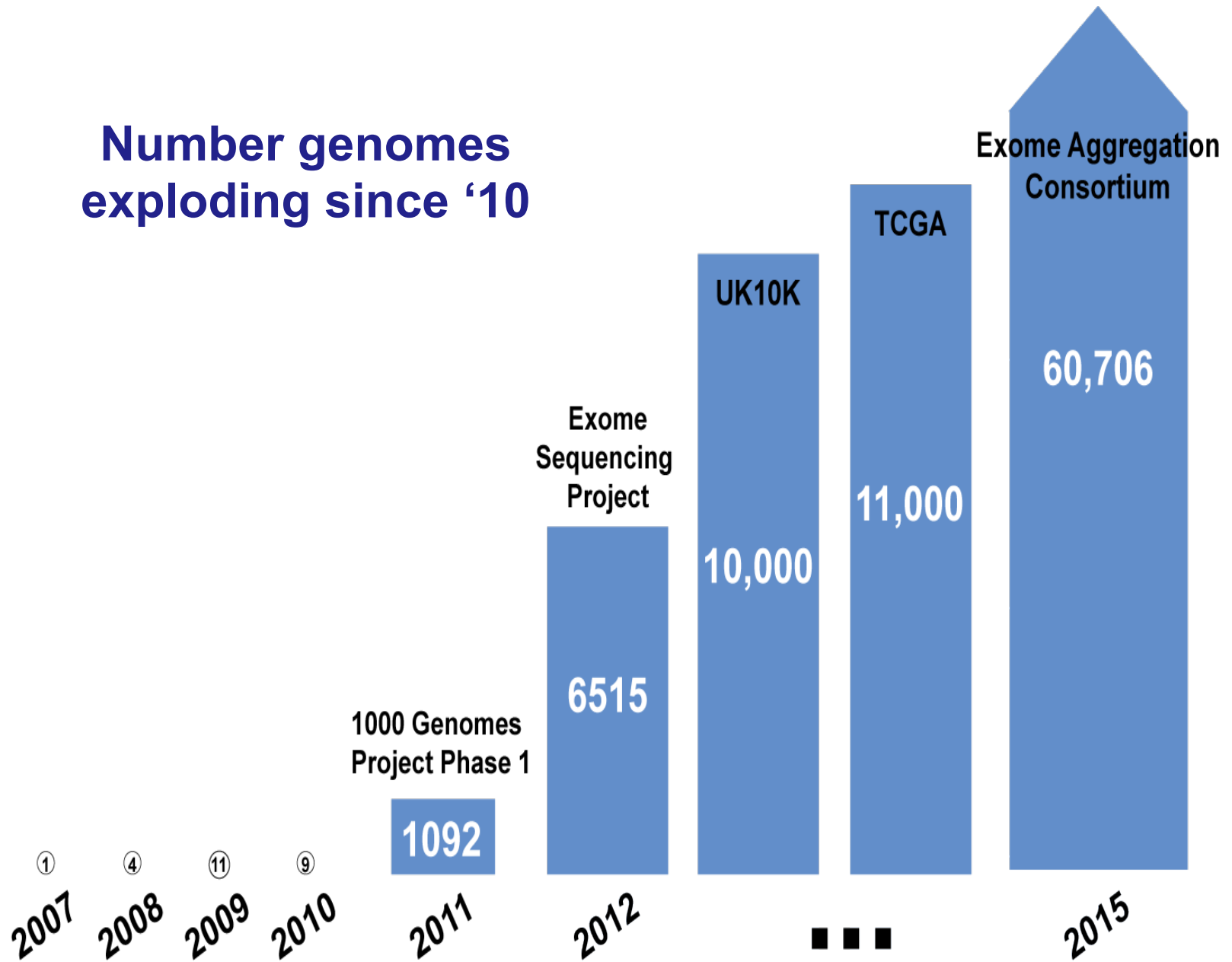


# The Rise of the Personal Genome to '10



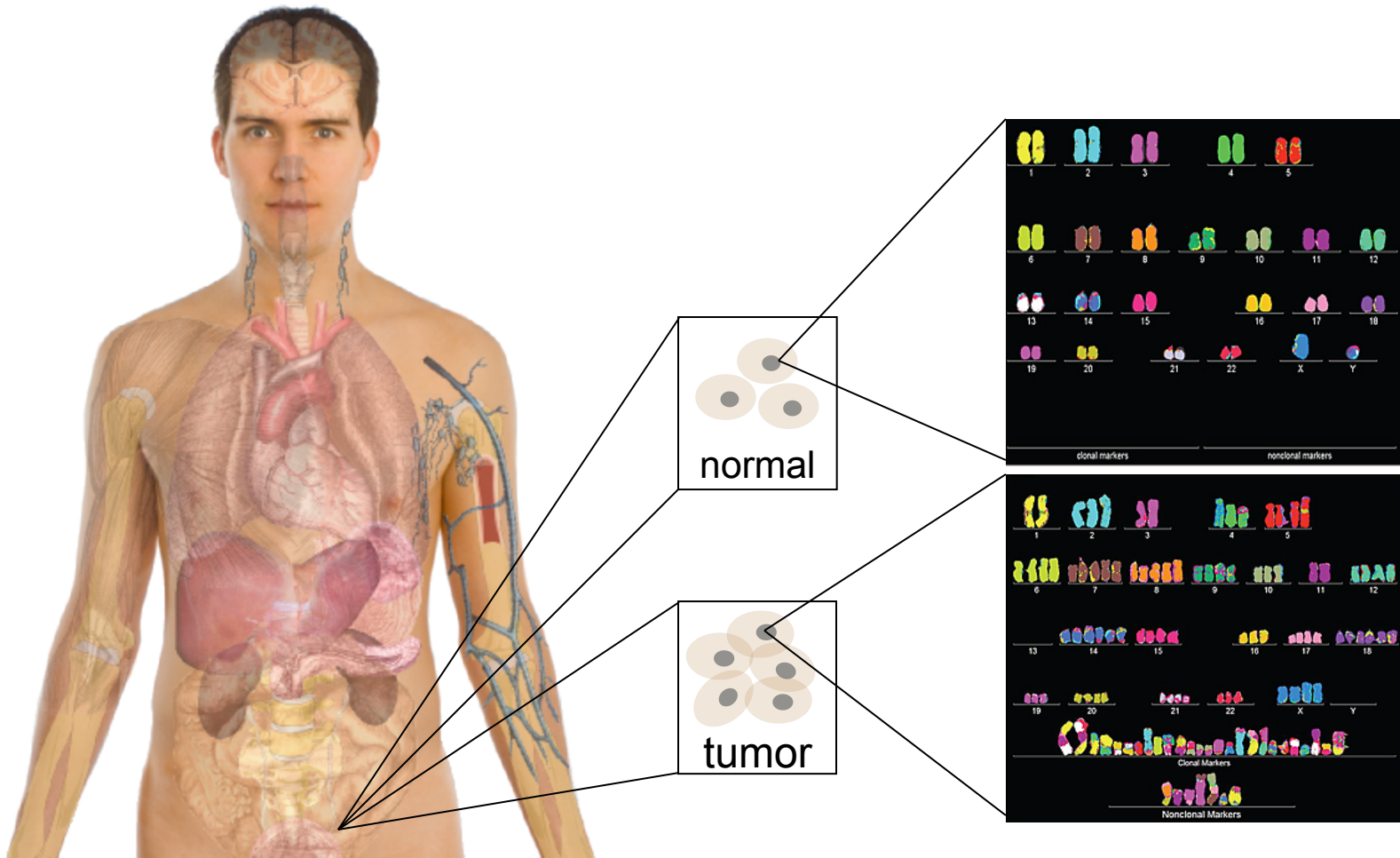
Adapted from *Nature* 2010

# Number genomes exploding since '10



# Personal Genomics as a Gateway into Biology

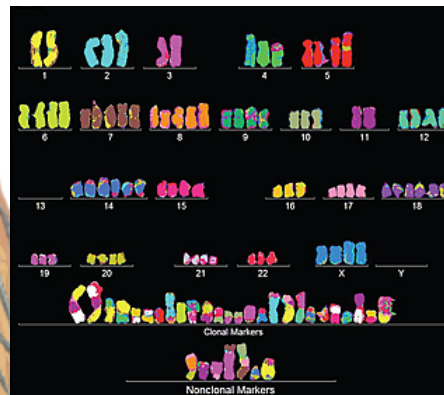
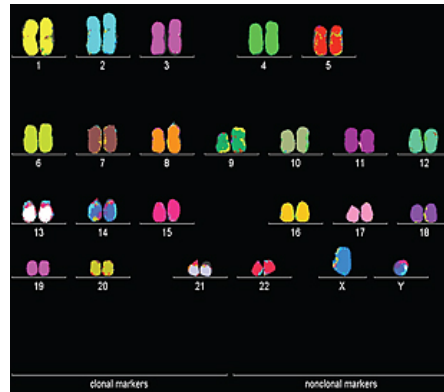
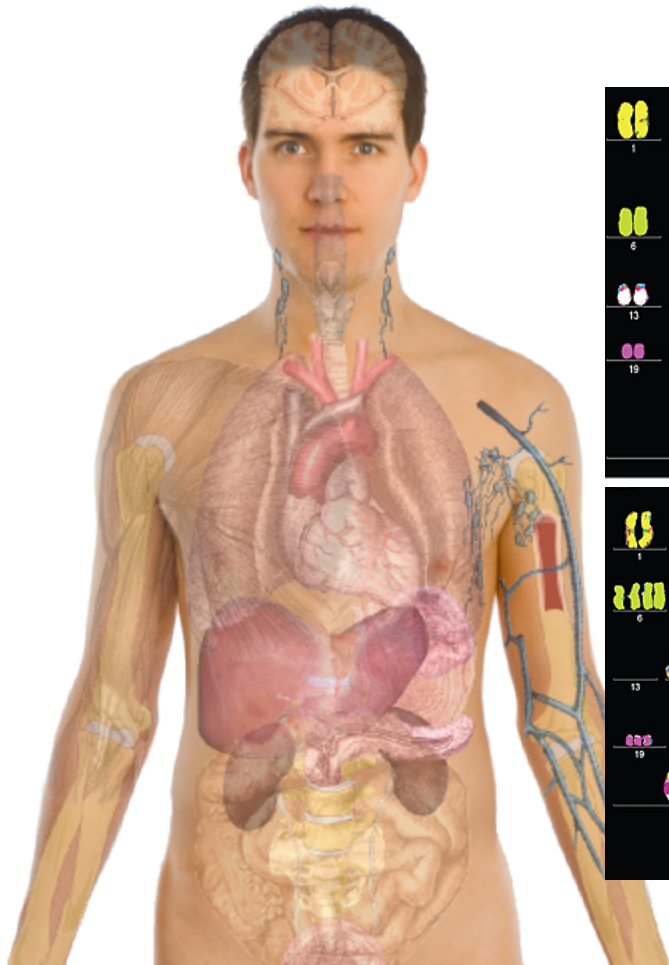
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.



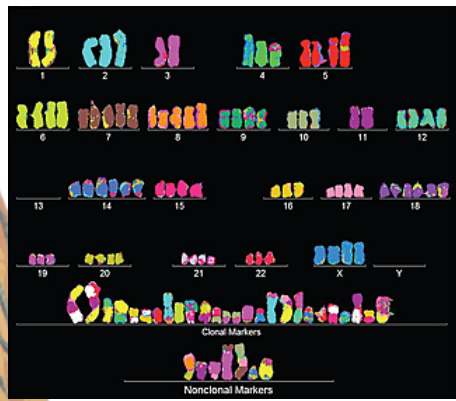
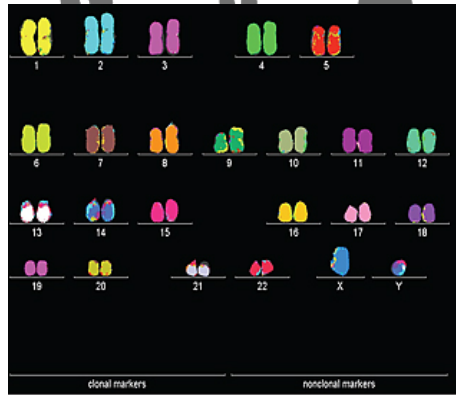
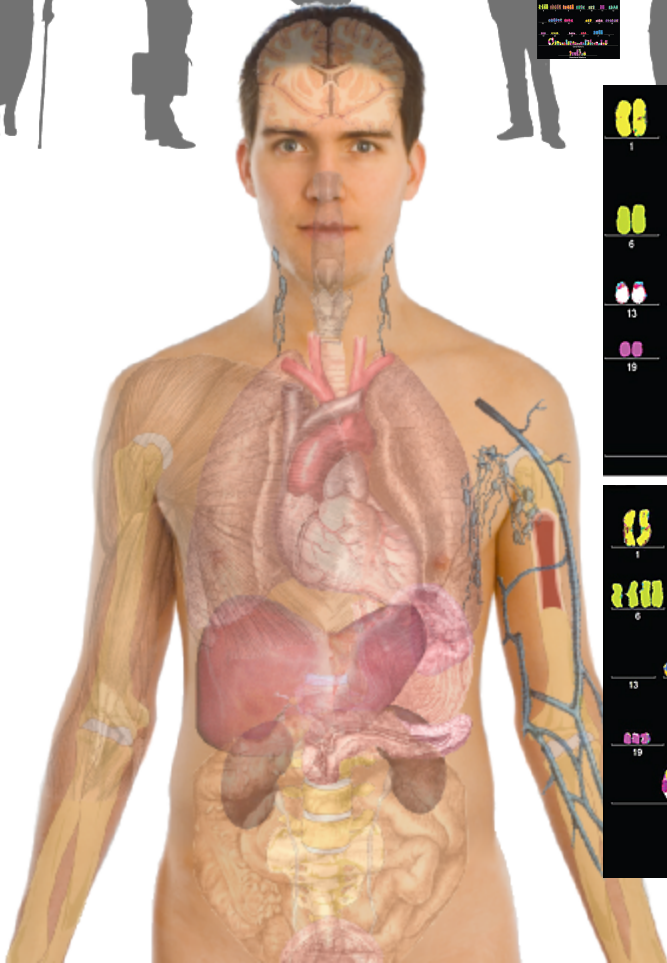
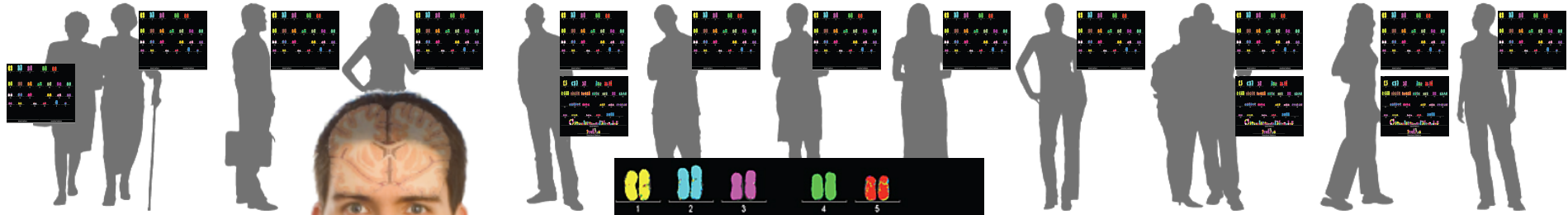
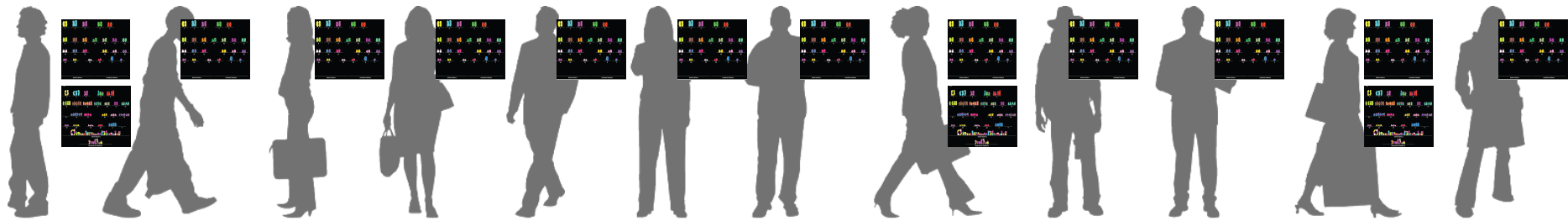


# Personal Genomics as a Gateway into Biology

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.







# Human Genetic Variation

A Cancer Genome



A Typical Genome



Population of 2,504 people



Origin of Variants

	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K

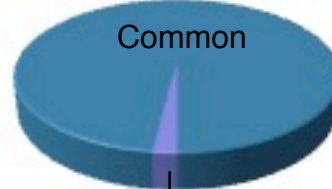


Driver (~0.1%)

Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants



Rare\* (1-4%)

SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M



Rare (~75%)

\* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

## CAN YOU FIND THE PANDA?

# Finding Key Variants

## Germline



- **Common variants**

- Can be associated with phenotype (ie disease) via a Genome-wide Association Study (GWAS), which tests whether the frequency of alleles differs between cases & controls.
- Usually their functional effect is weaker.
- Many are non-coding
- Issue of LD in identifying the actual causal variant.

- **Rare variants**

- Associations are usually underpowered due to low frequencies.
- They often have larger functional impact
- Can be collapsed in the same element to gain statistical power (burden tests).
- In some cases, causal variants can be identified through tracing inheritance of Mendelian subtypes of diseases in large families.

## CAN YOU FIND THE PANDA?



# Finding Key Variants

## Somatic

### • Overall

- Often these can be conceptualized as very rare variants
- A challenge to identify somatic mutations contributing to cancer is to find driver mutations & distinguish them from passengers.

### • Drivers

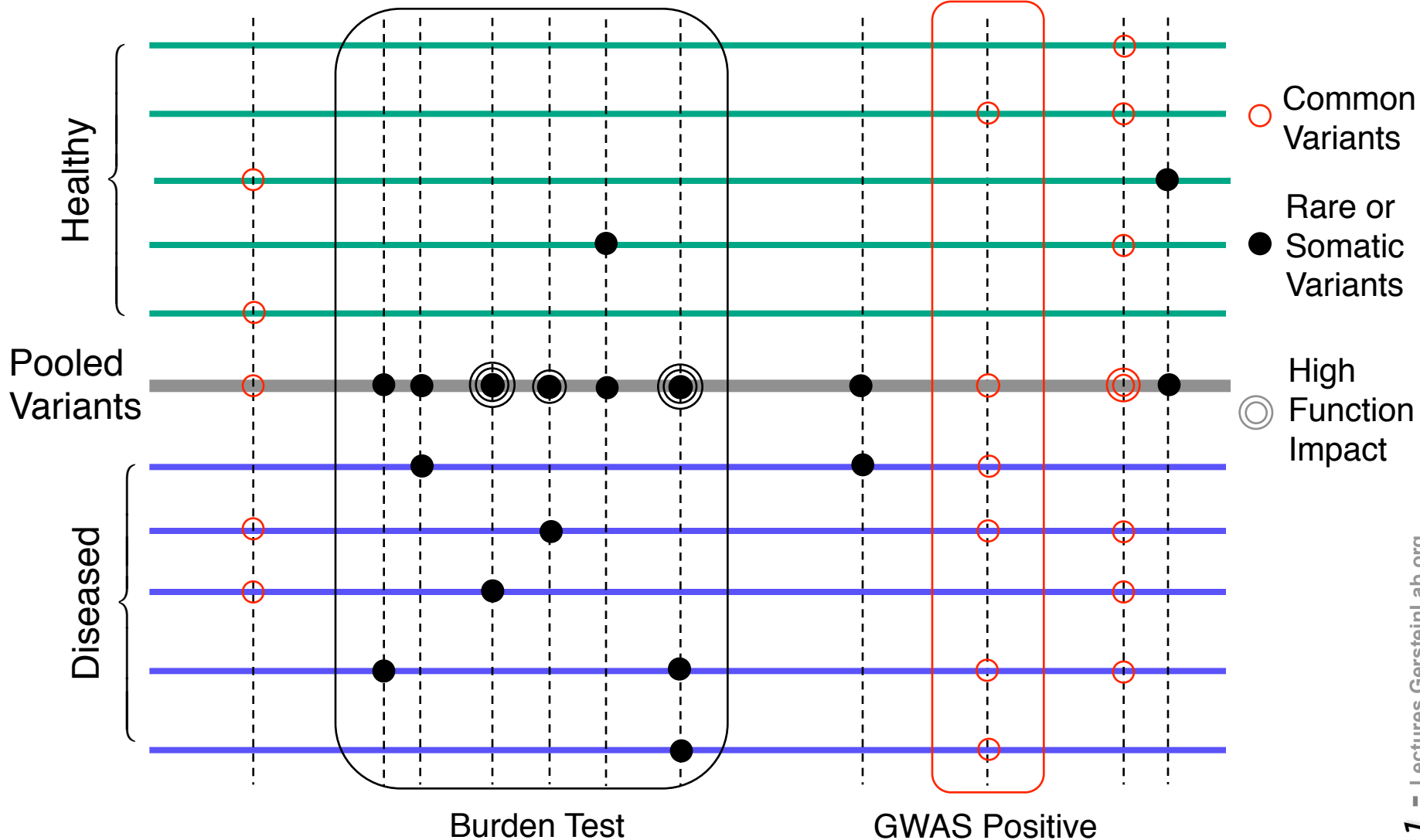
- Driver mutation is a mutation that directly or indirectly confers a selective growth advantage to the cell in which it occurs.
- A typical tumor contains 2-8 drivers; the remaining mutations are passengers.

### • Passengers

- Conceptually, a passenger mutation has no direct or indirect effect on the selective growth advantage of the cell in which it occurred.



# Association of Variants with Diseases



# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

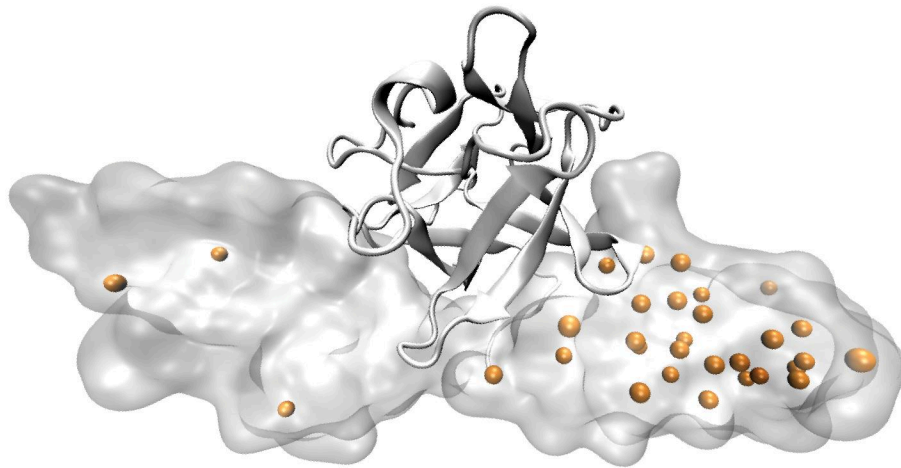
- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with **STRESS** cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with **MUSIC**
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using **AlleleDB** in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

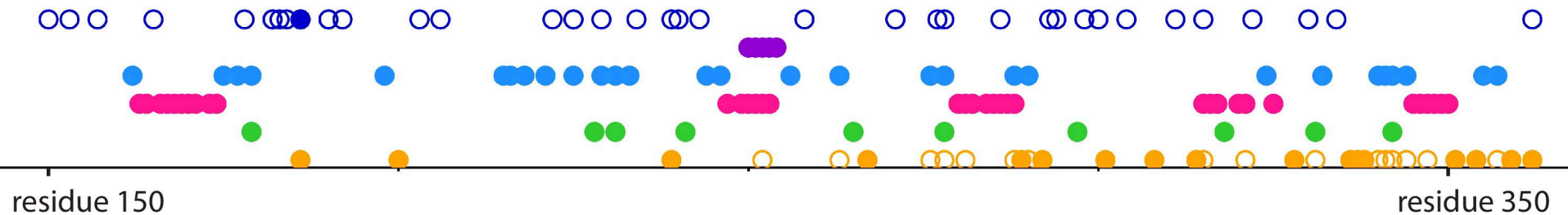
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated



- ○ 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

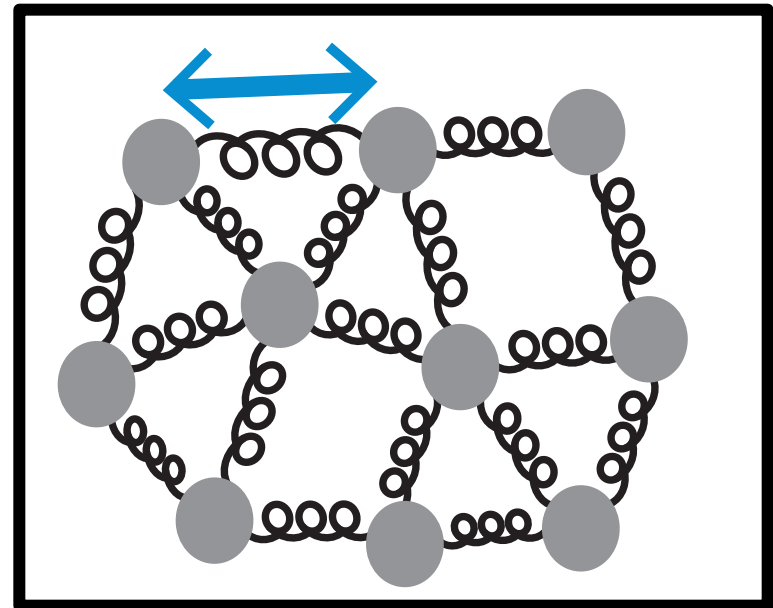
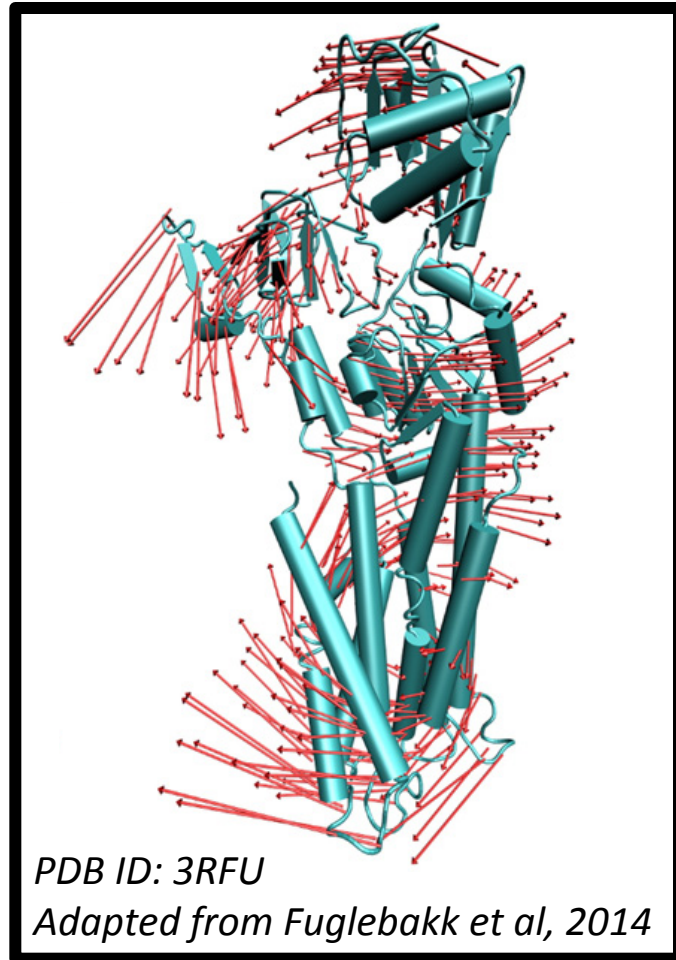
*Fibroblast growth factor receptor 2 (pdb: 1IIL)*





# Models of Protein Conformational Change

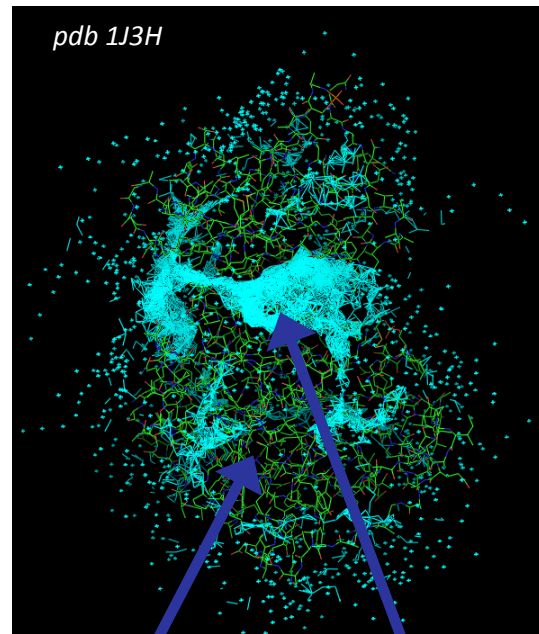
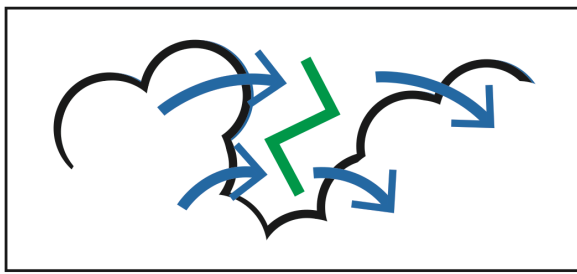
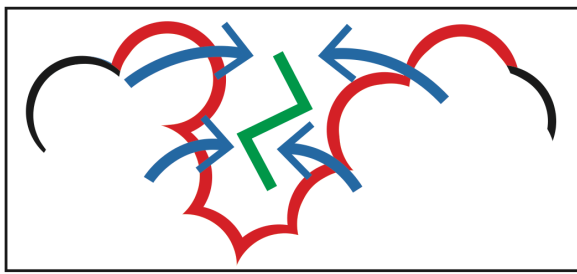
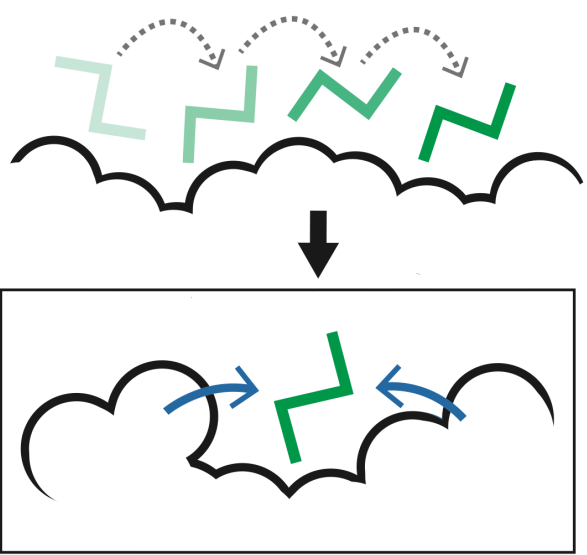
## Motion Vectors from Normal Modes (ANMs)



Characterizing uncharacterized variants  
<= Finding Allosteric sites  
<= Modeling motion

# Predicting Allosterically-Important Residues at the Surface

1. MC simulations generate a large number of candidate sites
2. Score each candidate site by the degree to which it perturbs large-scale motions
3. Prioritize & threshold the list to identify the set of high confidence-sites



Surface region with high density of candidate sites

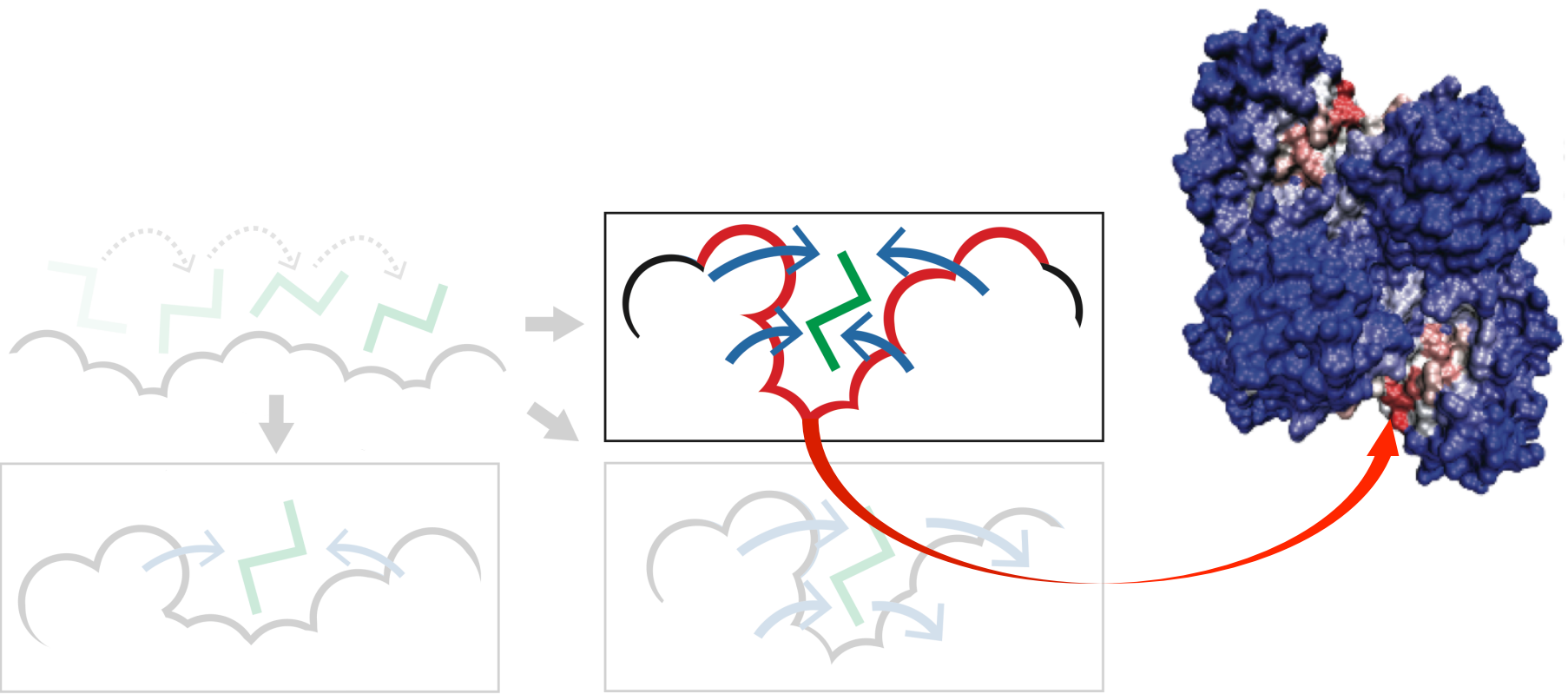
Surface region with low density of candidate sites

$$\text{binding leverage} = \sum_{m=1}^{10} (\sum_i \sum_j \Delta d_{ij(m)}^2)$$

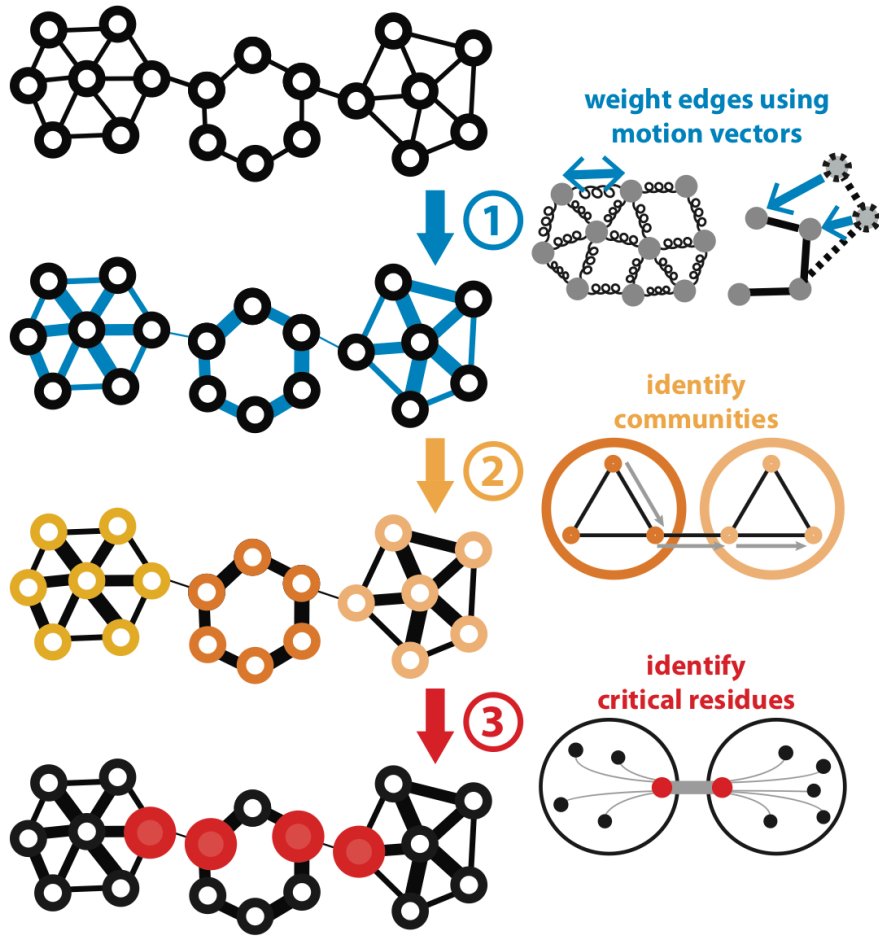
Adapted from Clarke\*, Sethi\*, et al (in press)

# Predicting Allosterically-Important Residues at the Surface

PDB: 3PFK

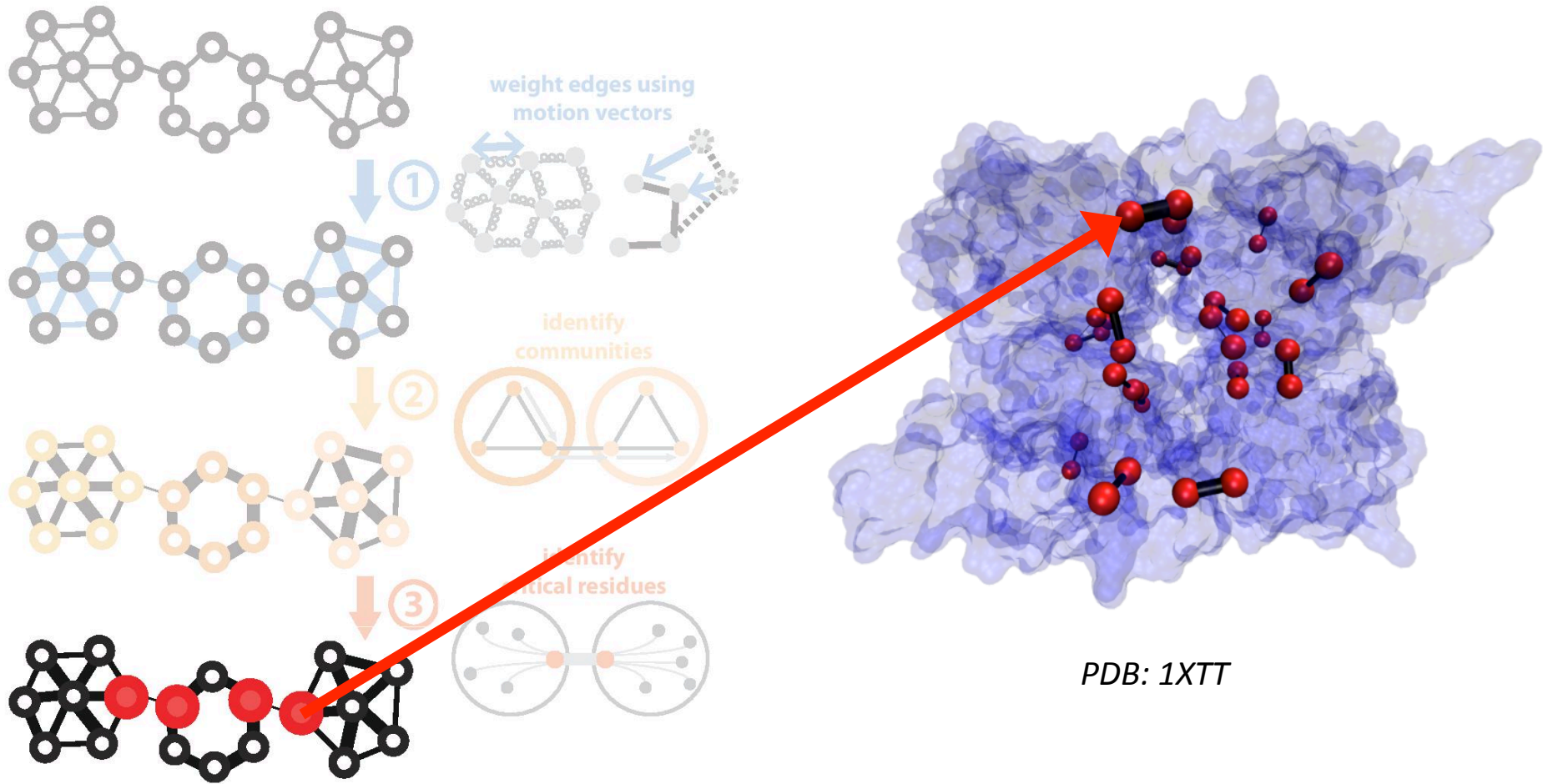


# Predicting Allosterically-Important Residues within the Interior



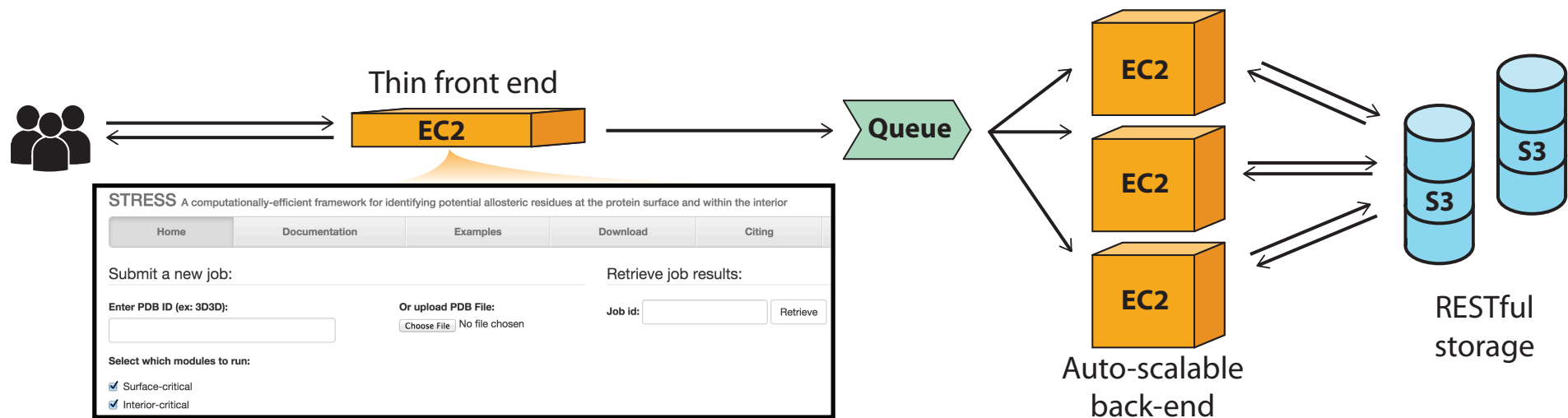


# Predicting Allosterically-Important Residues within the Interior



# STRESS Server Architecture: Highlights

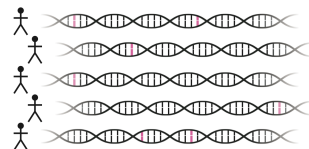
stress.molmovdb.org



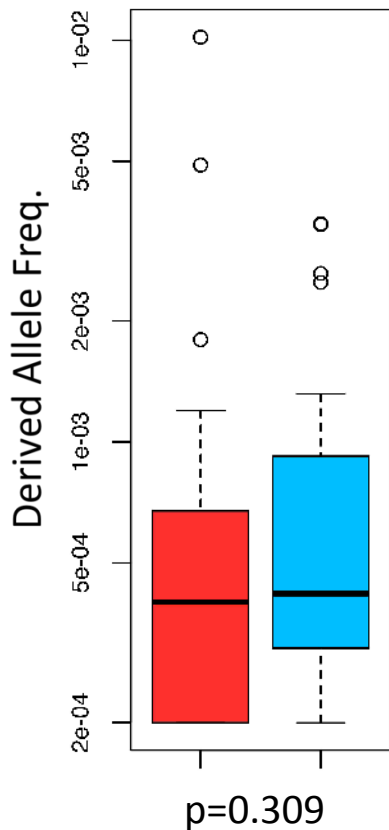
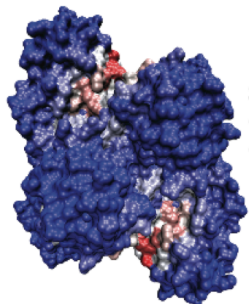
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are stored in S3 buckets.

# Intra-species conservation of predicted allosteric residues

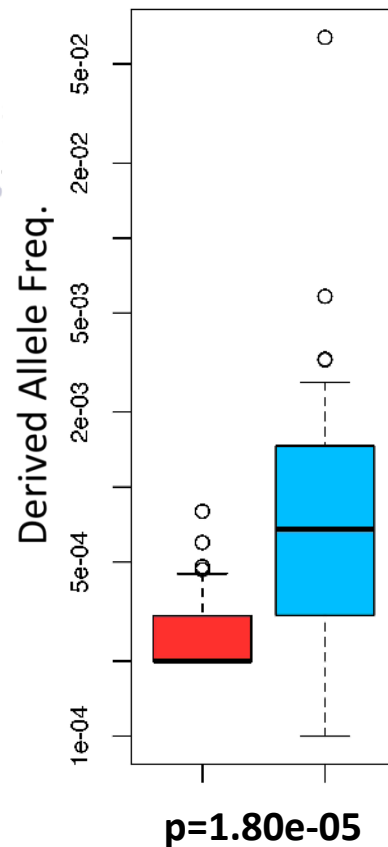
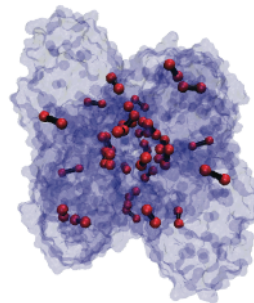
## 1000 Genomes



### Surface

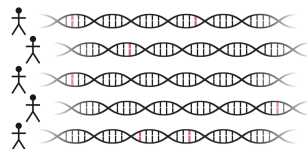


### Interior

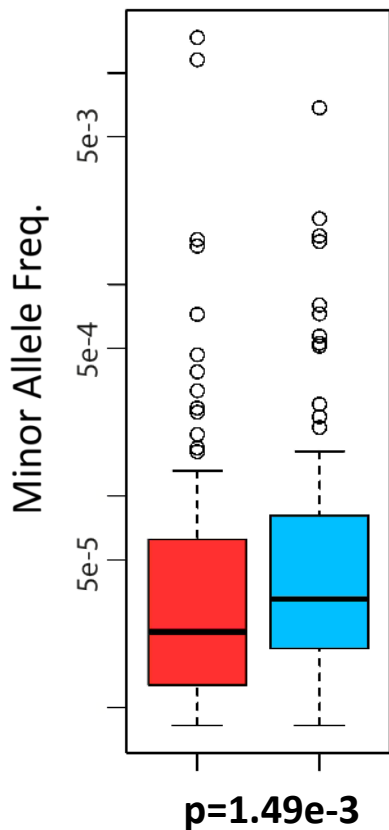
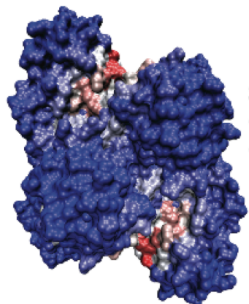


# Intra-species conservation of predicted allosteric residues

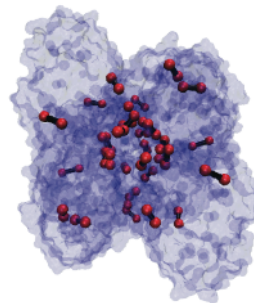
*ExAC*



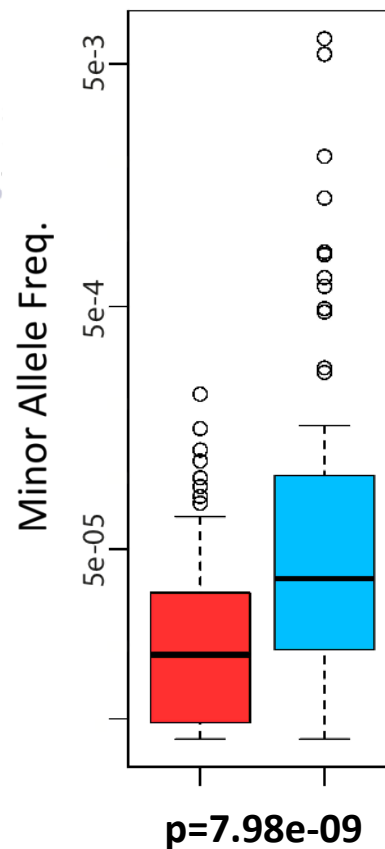
Surface



Interior



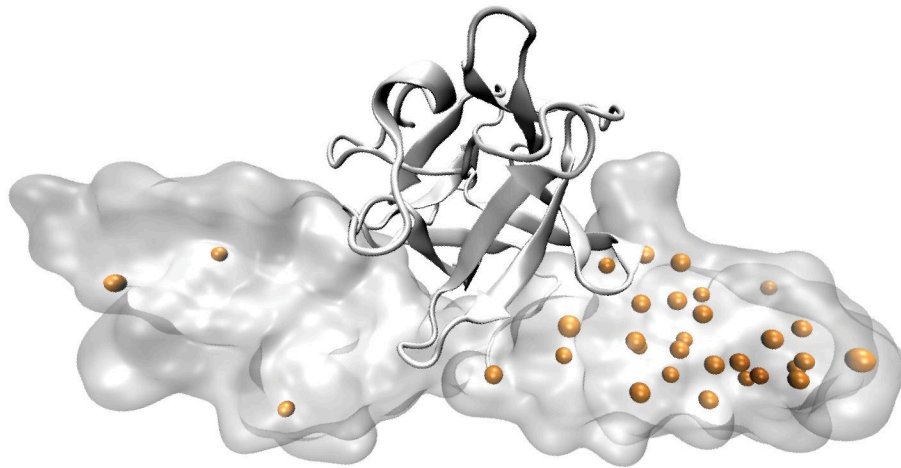
■ critical  
■ non-critical





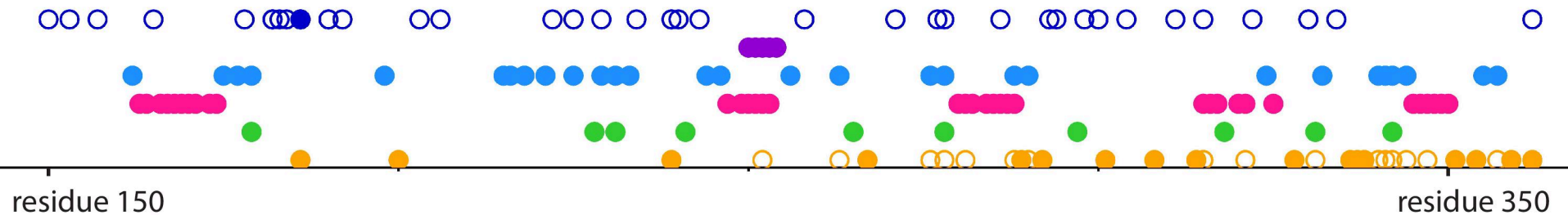
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated



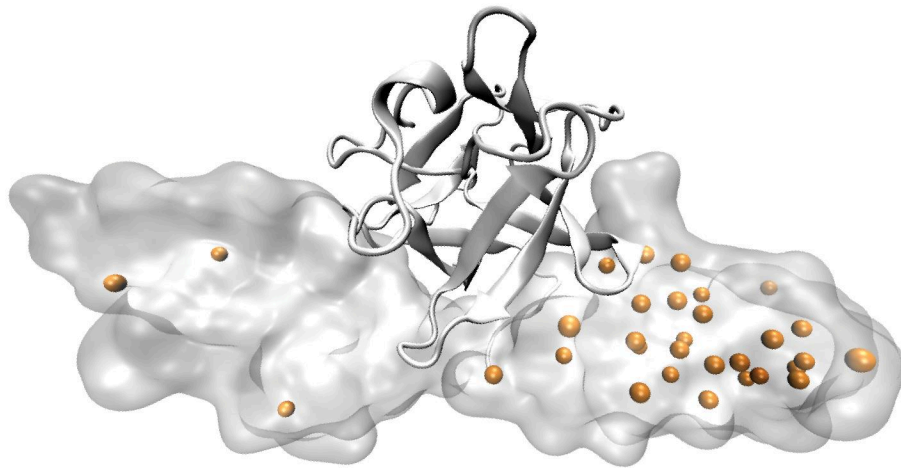
- ○ 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

*Fibroblast growth factor receptor 2 (pdb: 1IIL)*



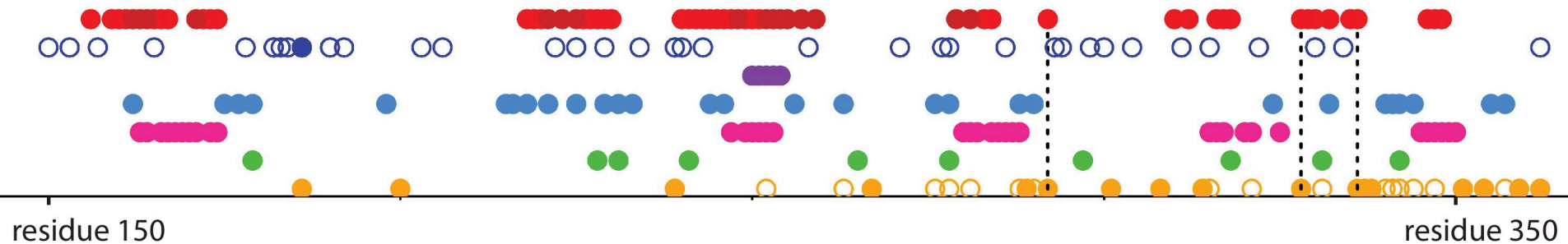
# Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

Rationalizing disease variants in the context of allosteric behavior with allostery as an added annotation



- Predicted allosteric (surface | interior)
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

*Fibroblast growth factor receptor 2 (pdb: 1IIL)*

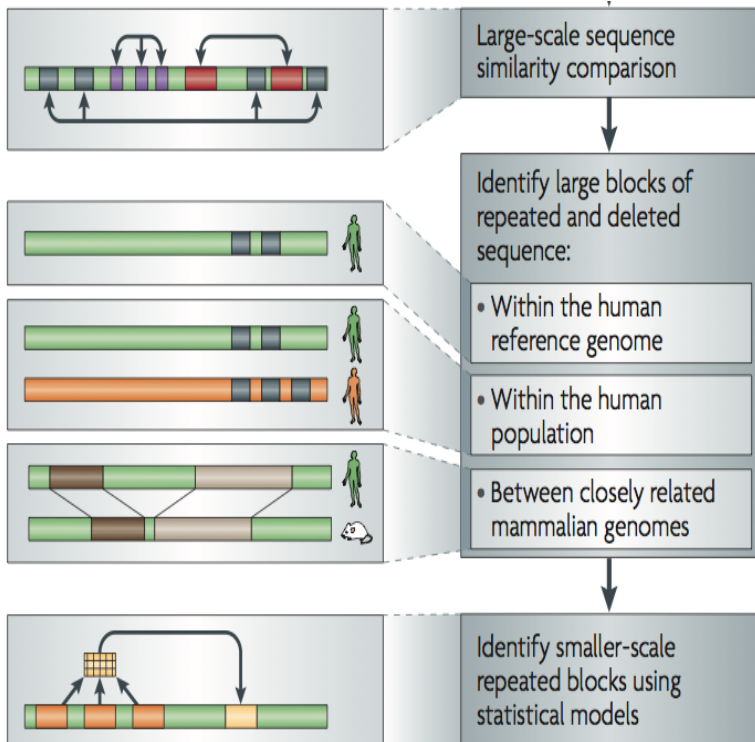


# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

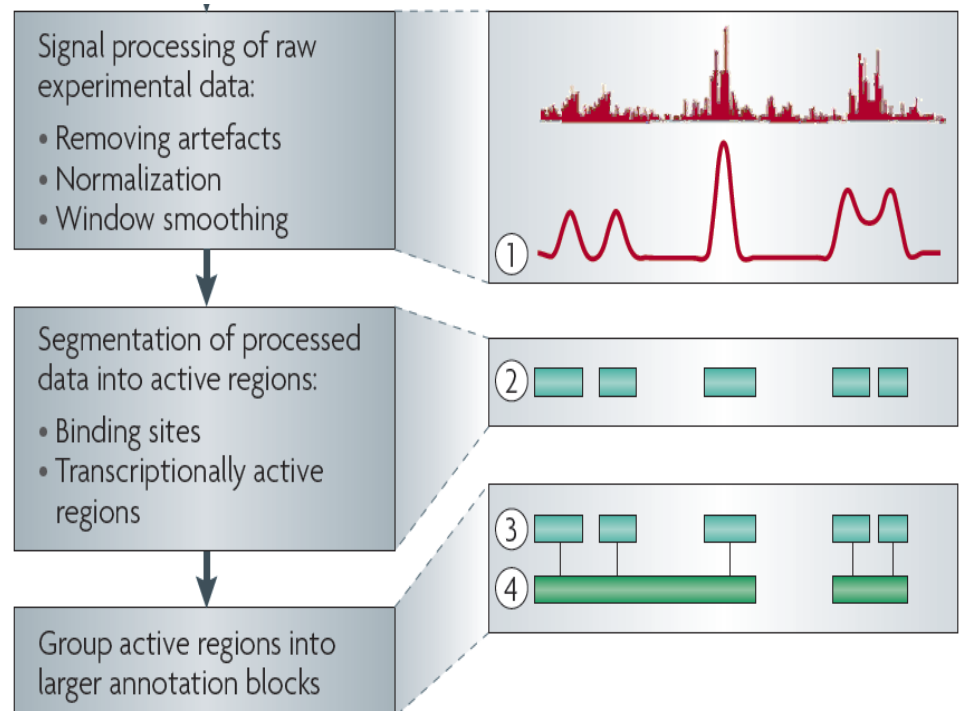
# Non-coding Annotations: Overview

## Sequence features, incl. Conservation



## Functional Genomics

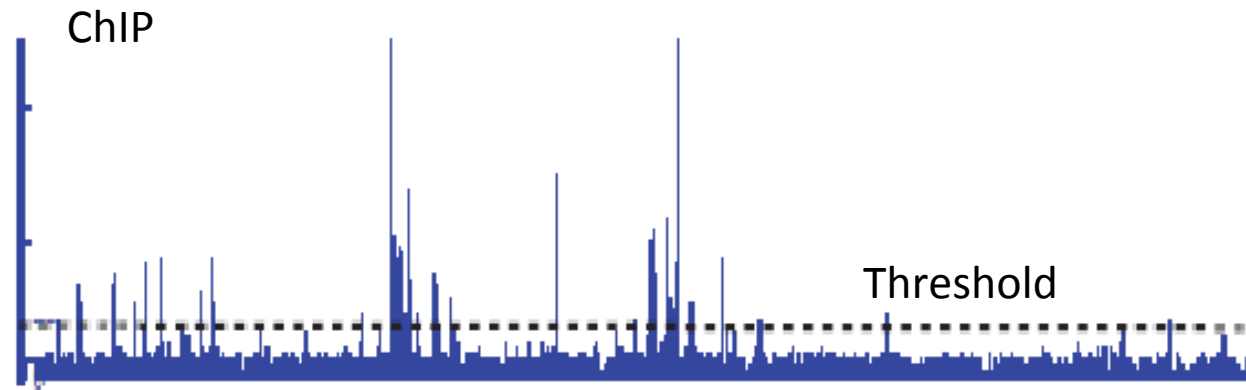
Chip-seq (Epigenome & seq. specific TF)  
and ncRNA & un-annotated transcription



[Alexander et al., *Nat. Rev. Genet.* ('10)]

# Summarizing the Signal: "Traditional" ChipSeq Peak Calling

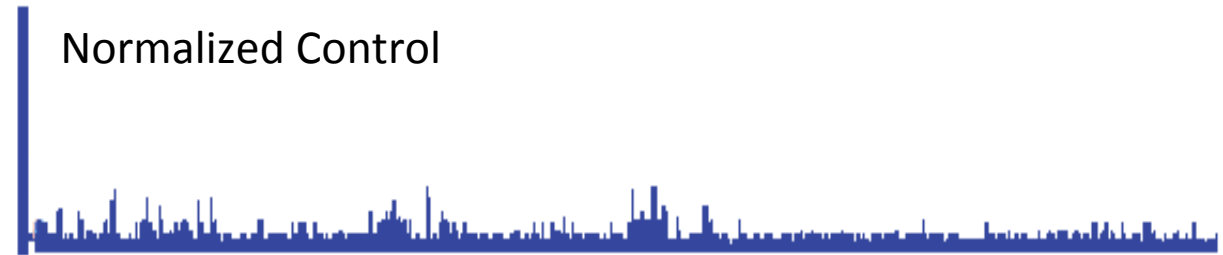
- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



Potential Targets



- Score against the control



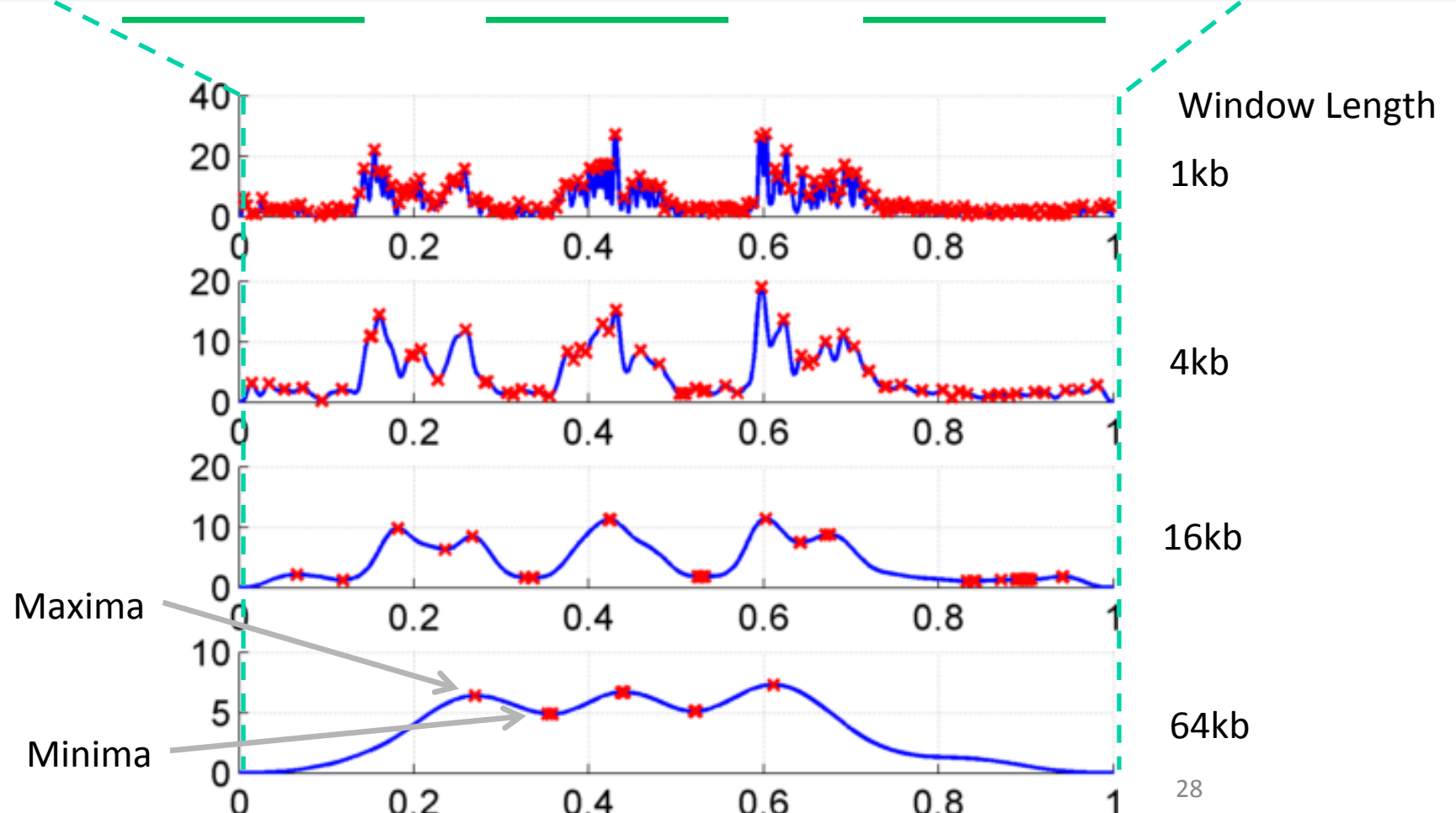
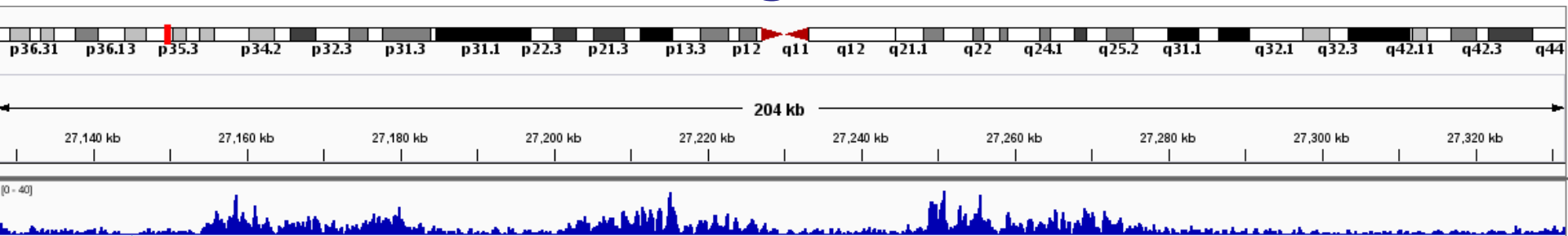
Significantly Enriched targets



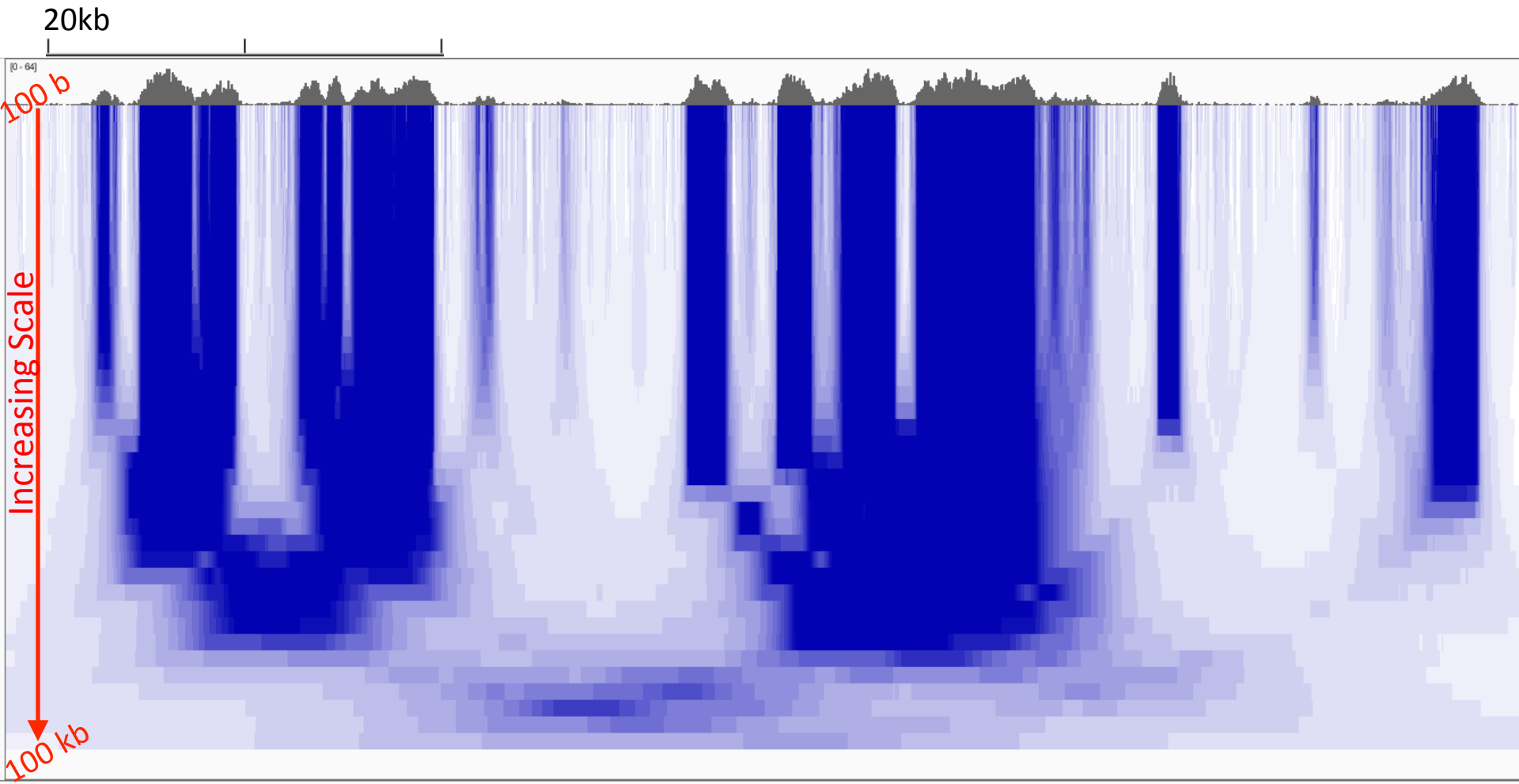
Now an update: "PeakSeq 2" => MUSIC



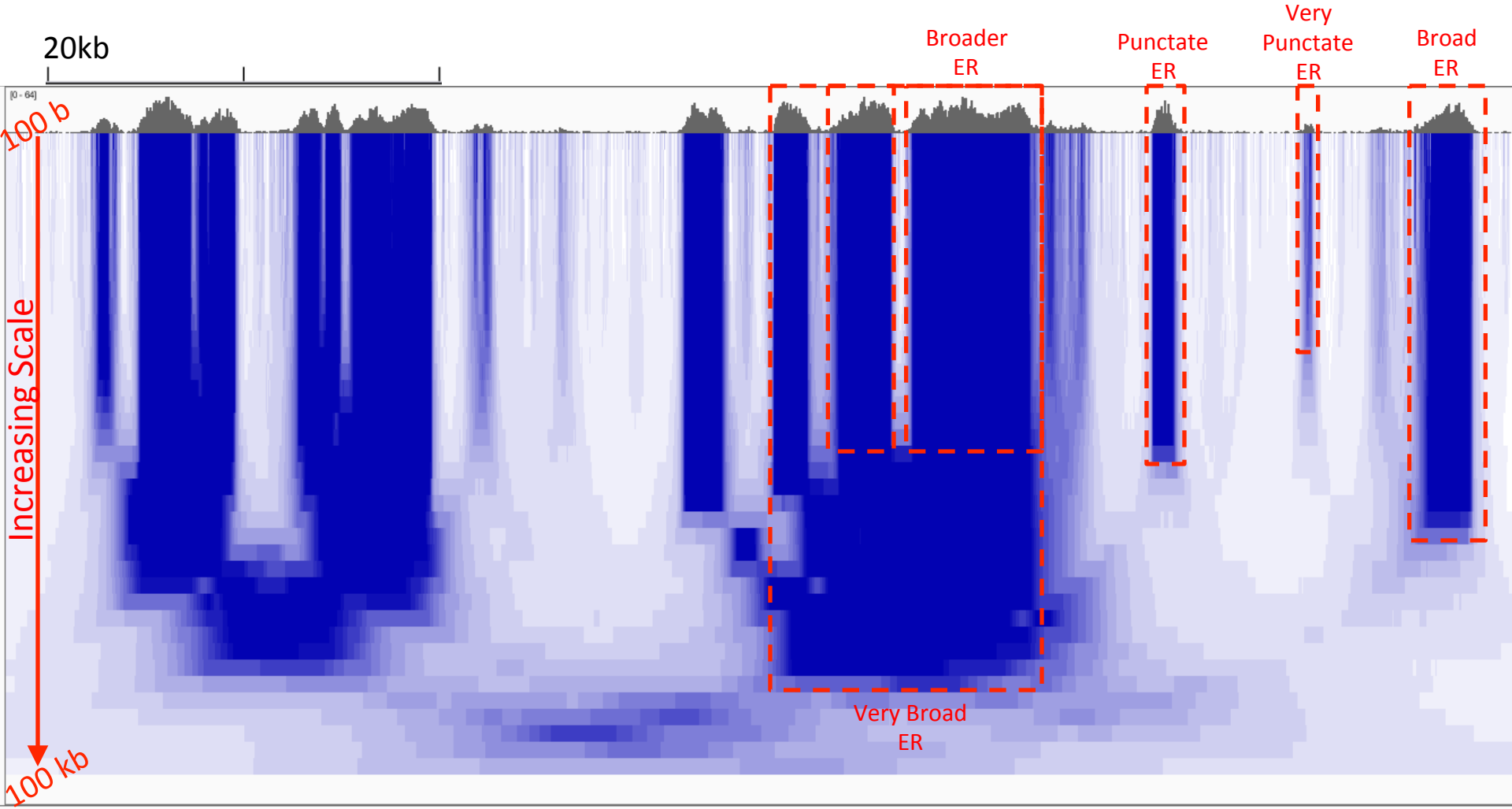
# Multiscale Analysis, Minima/Maxima based Coarse Segmentation



# Multiscale Decomposition



# Multiscale Decomposition

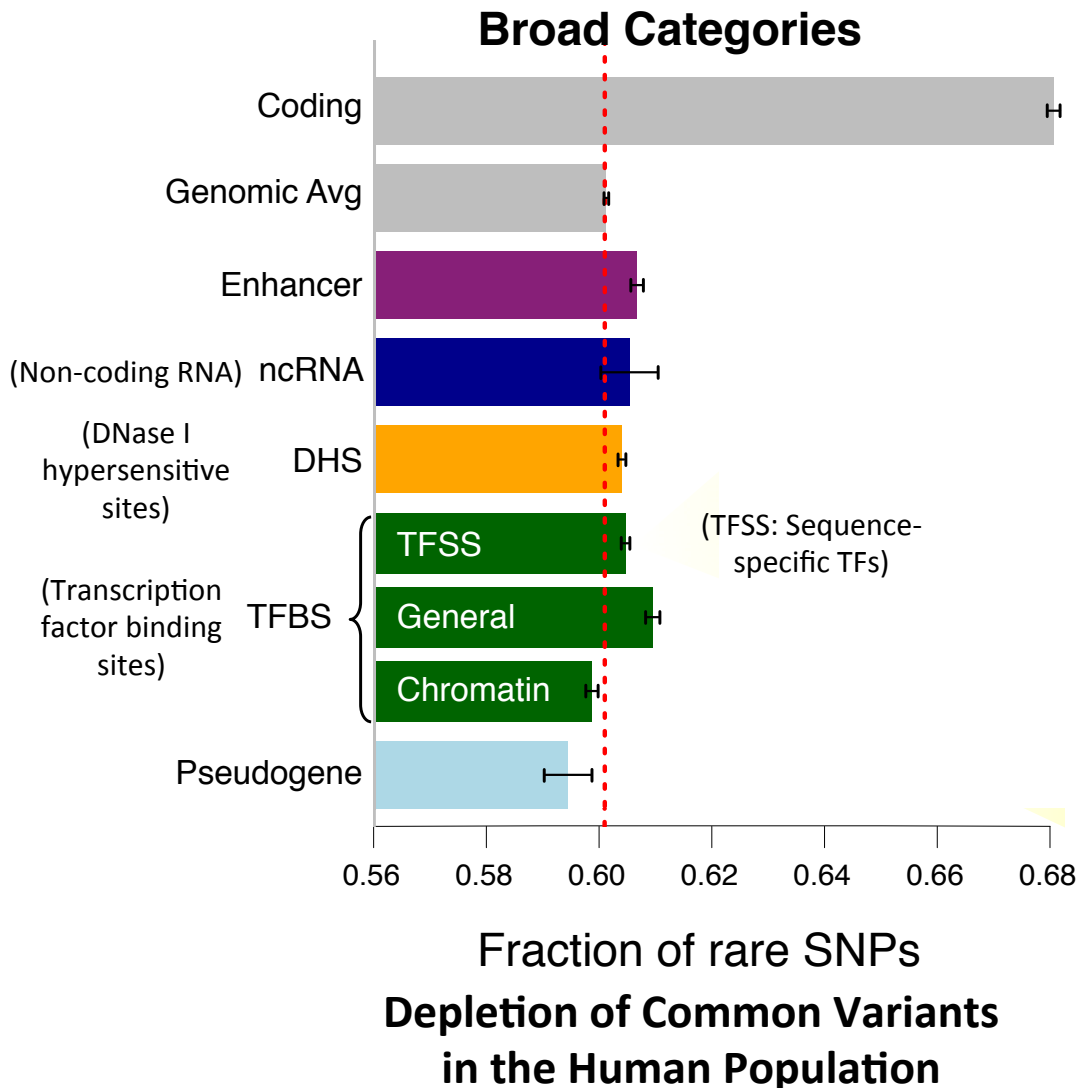


# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

# Finding "Conserved" Sites in the Human Population:

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

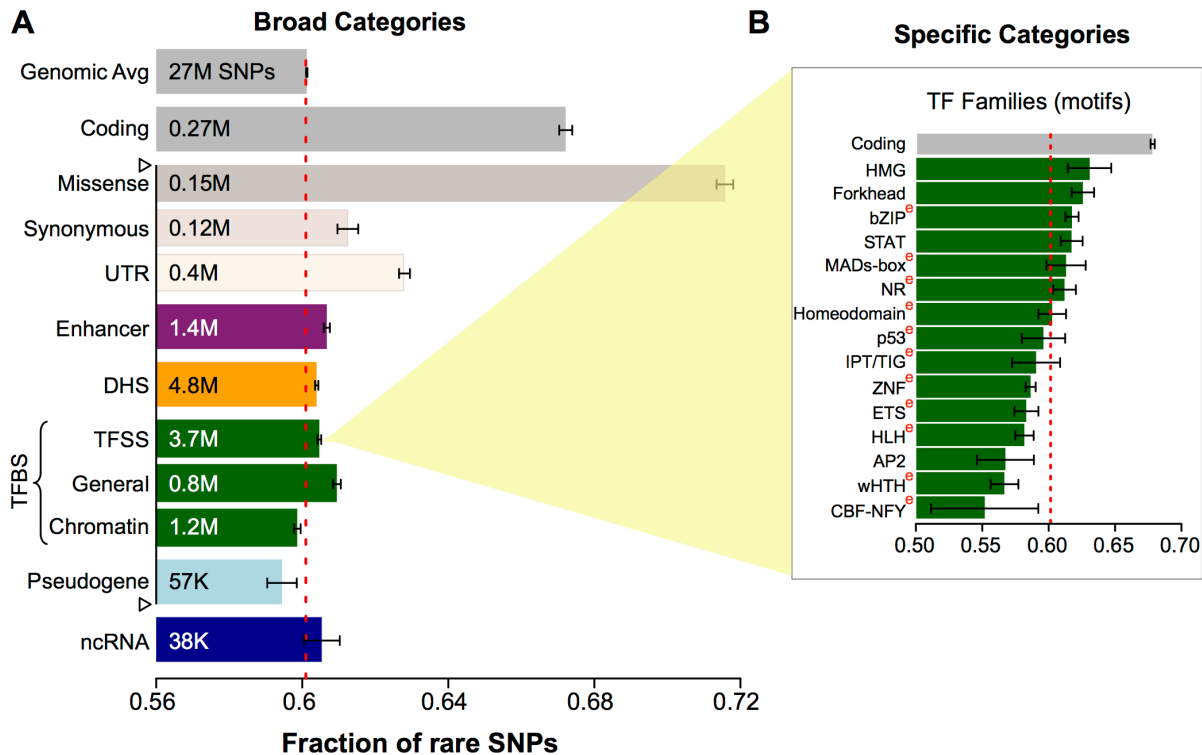


- Broad categories of regulatory regions under negative selection
  - Related to:

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

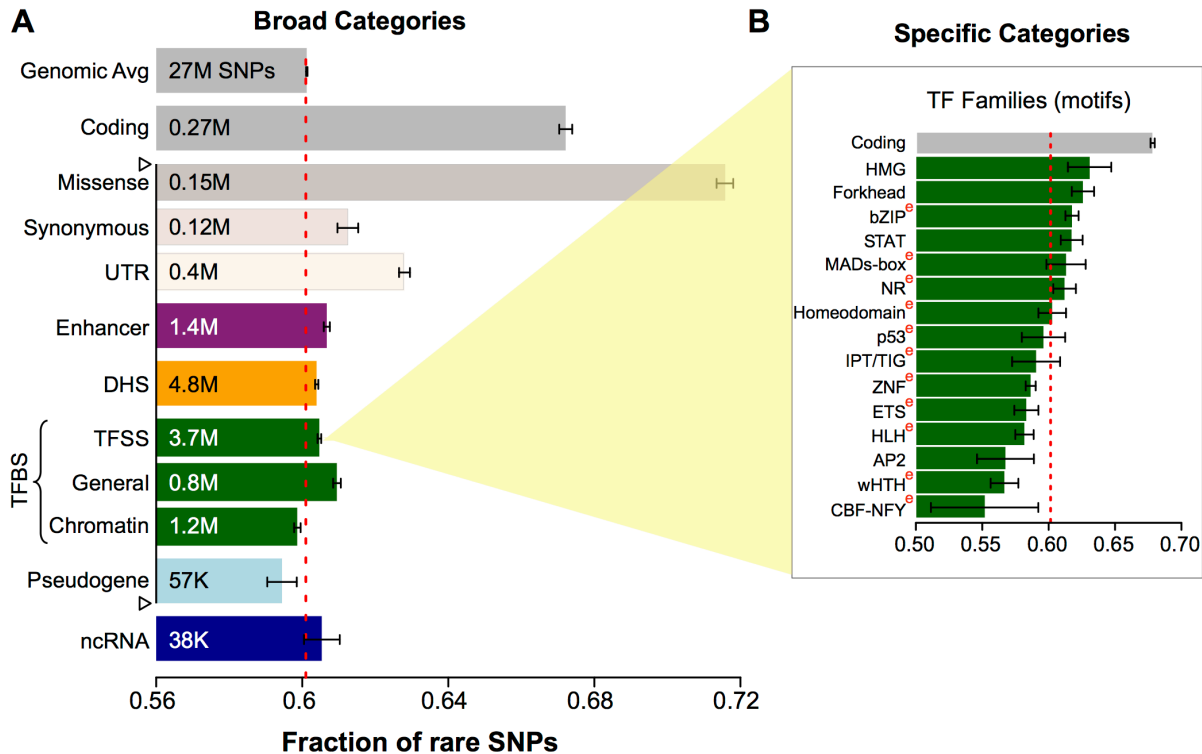
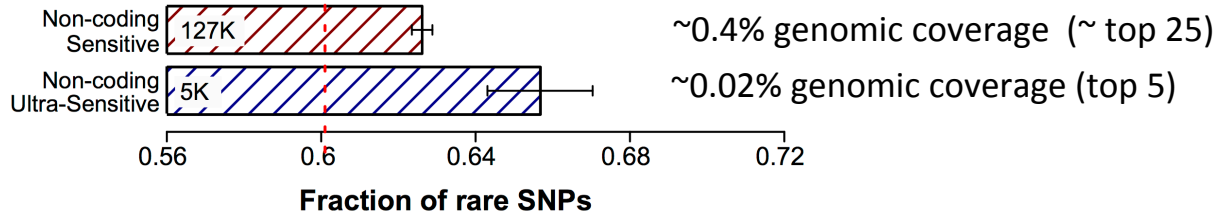


# Differential selective constraints among specific sub-categories



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

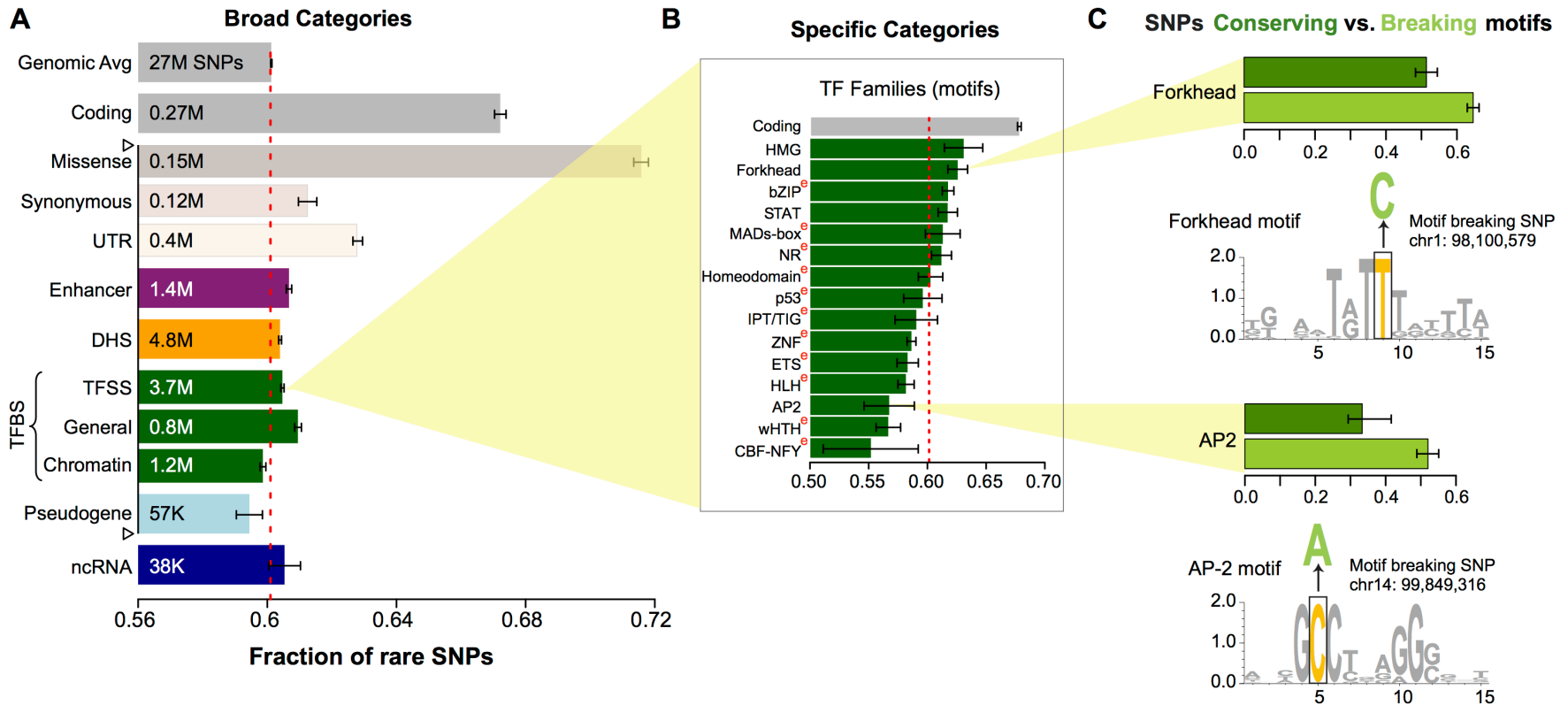
# Defining Sensitive non-coding Regions



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

# SNPs which break TF motifs are under stronger selection

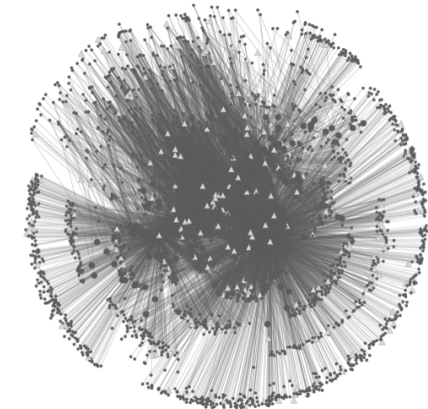
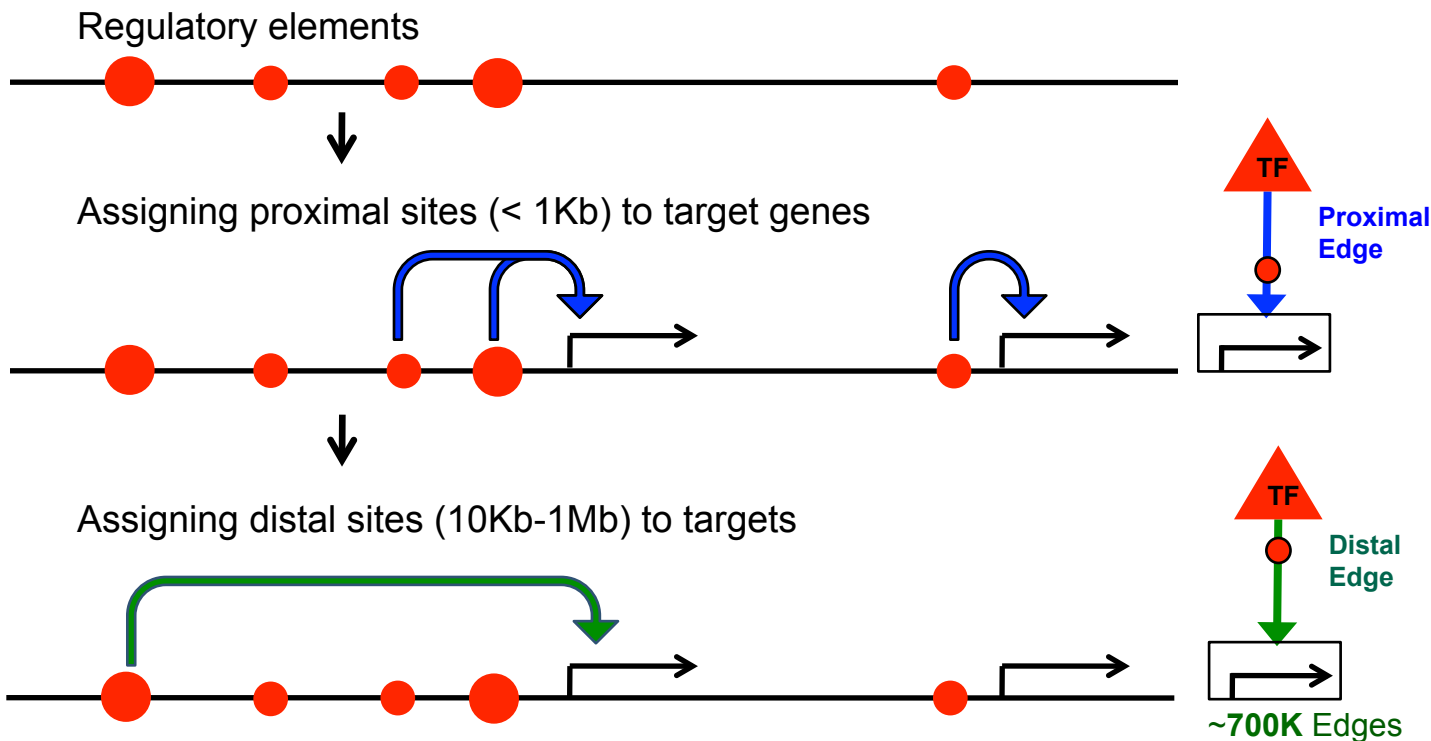


# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

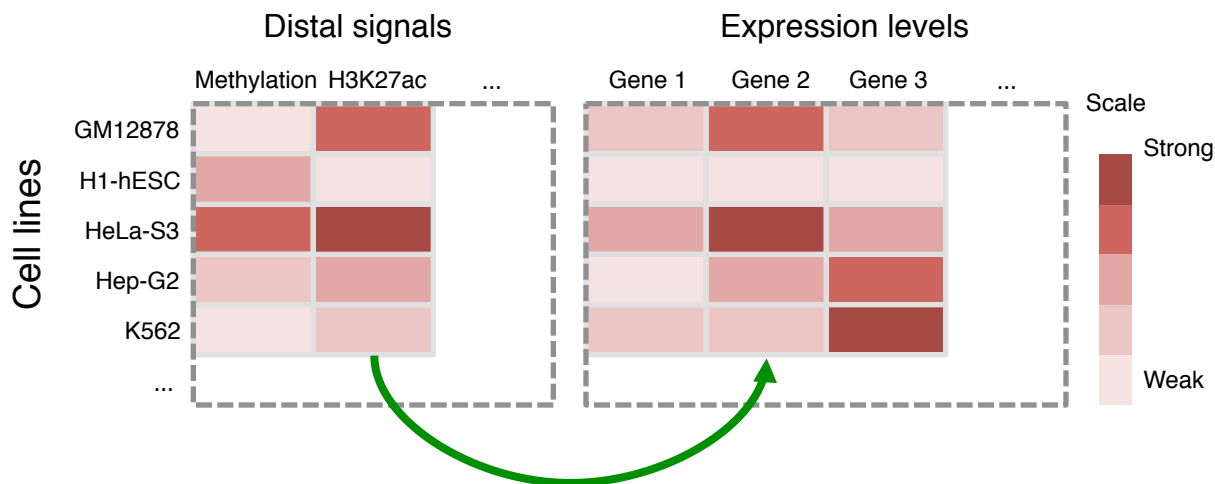
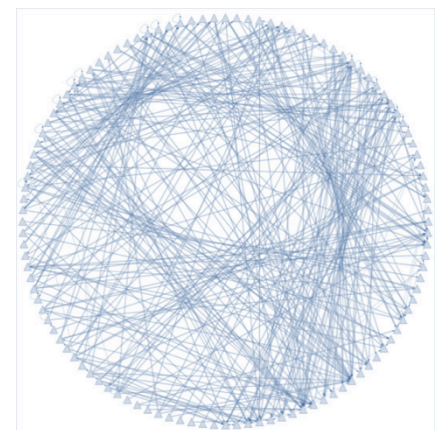
- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

# 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) ]



Filtering  
~26K

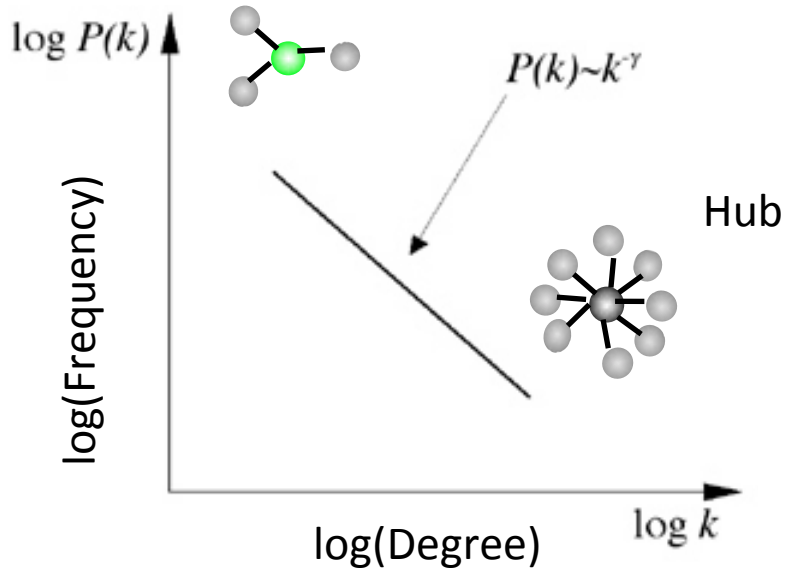


Connecting Distal Elements via **Activity Correlations**.

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



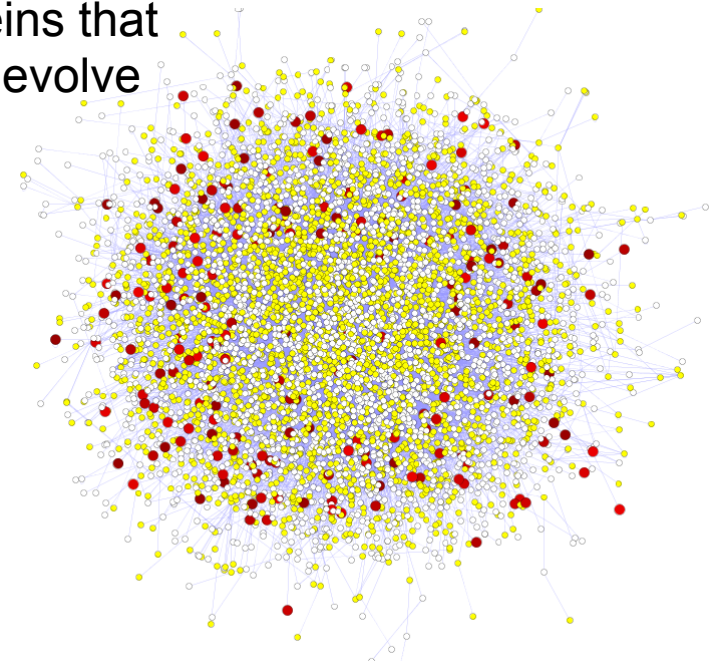
## Power-law distribution



## Hubs Under Constraint: A Finding from the Network Biology Community

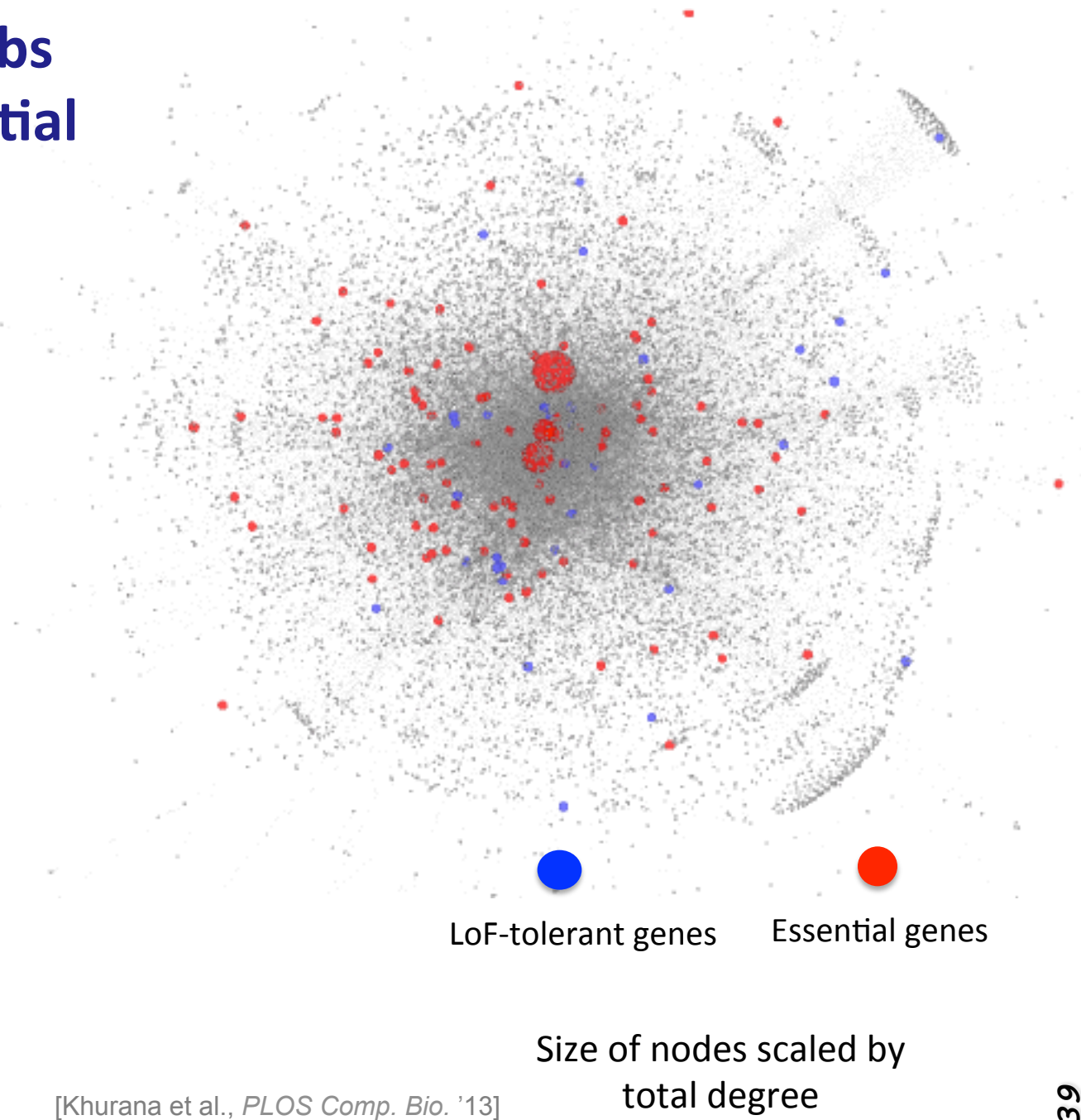
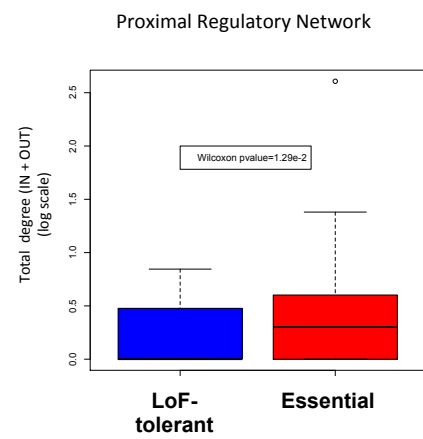
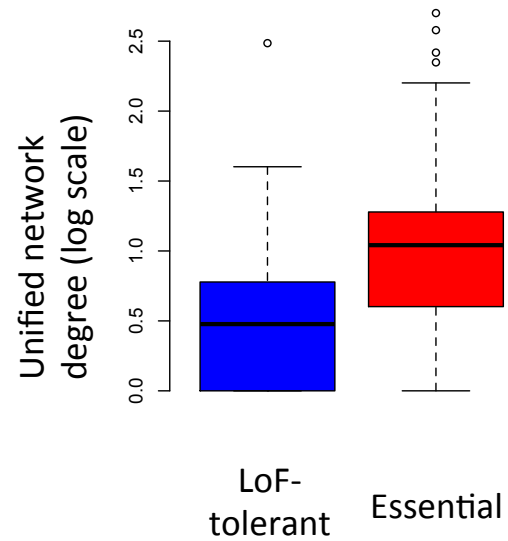
- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

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



- 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*

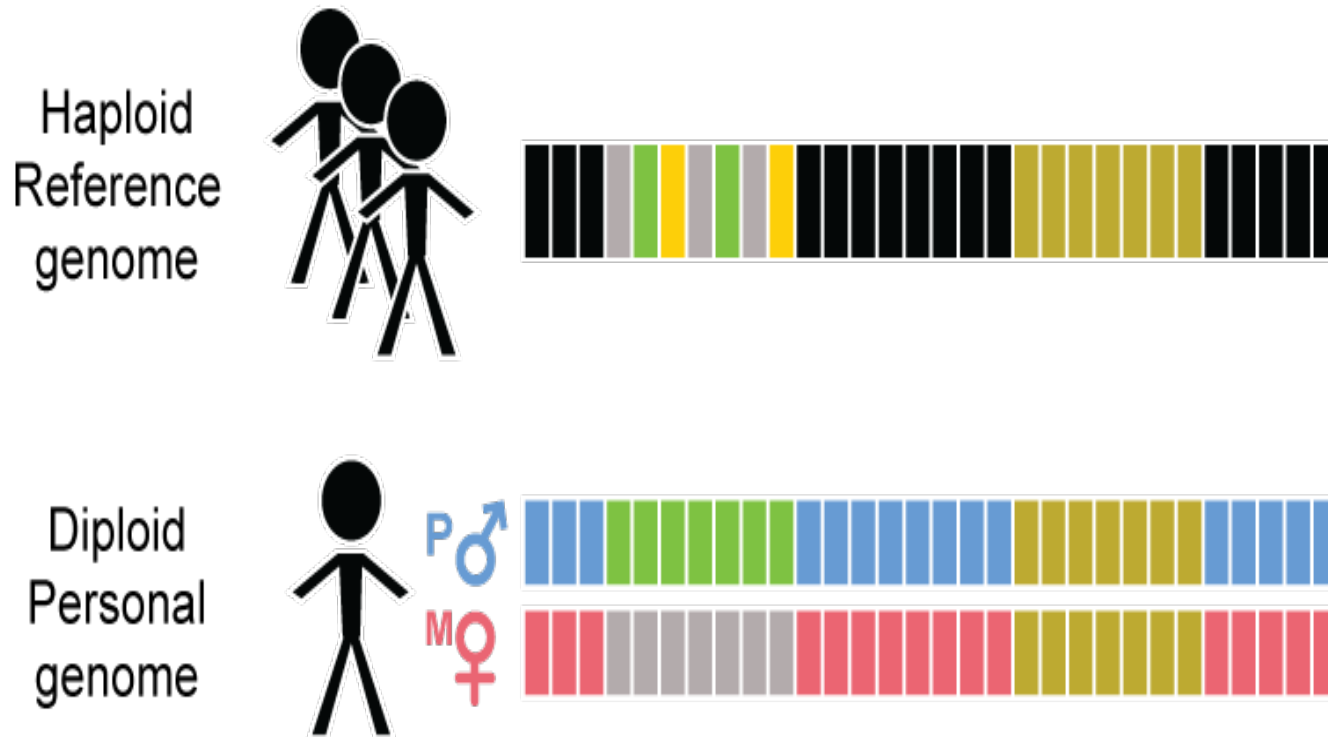
# Regulatory Hubs are more Essential



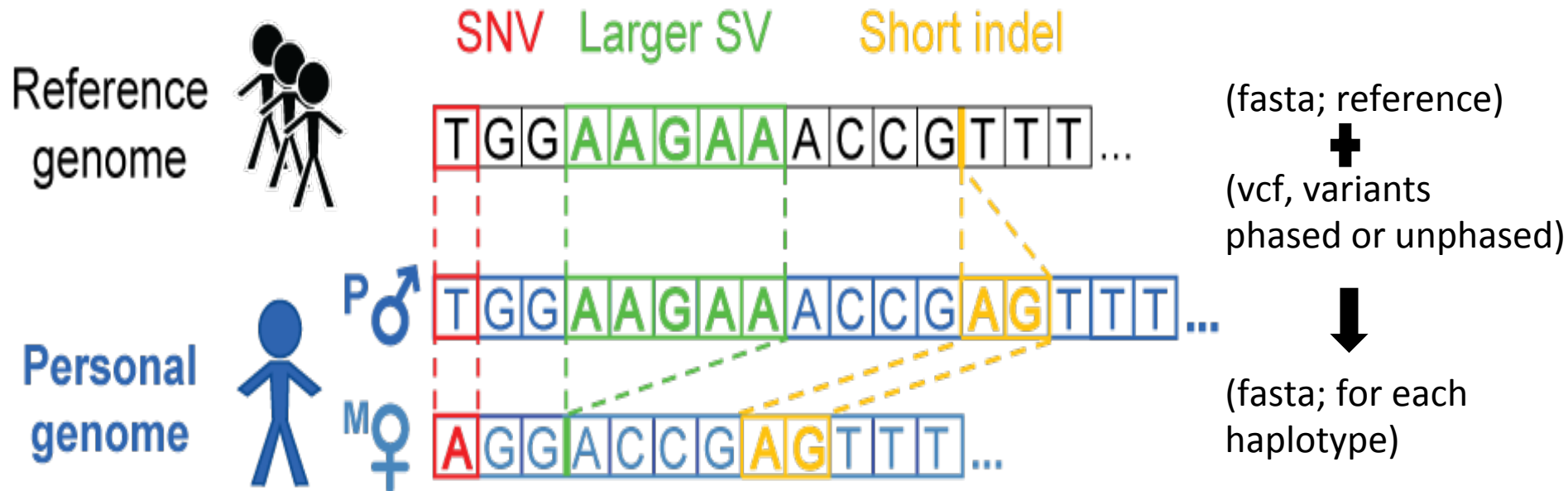
# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

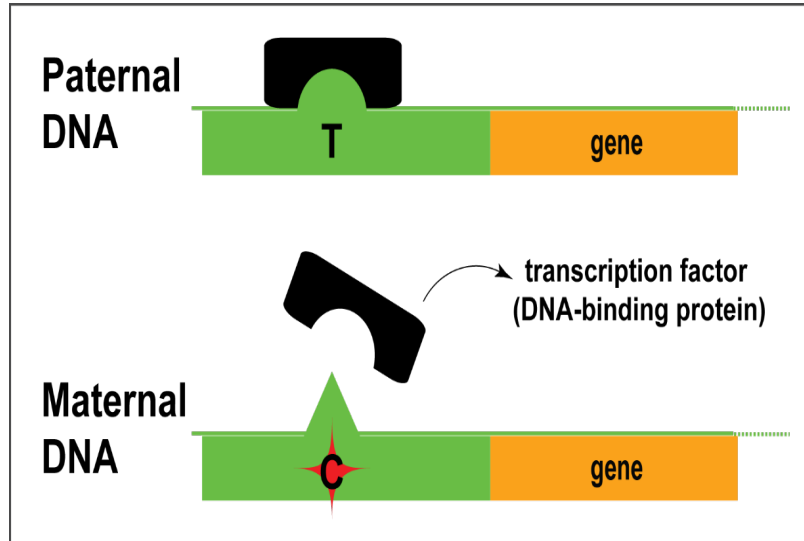
# Diploid personal genome



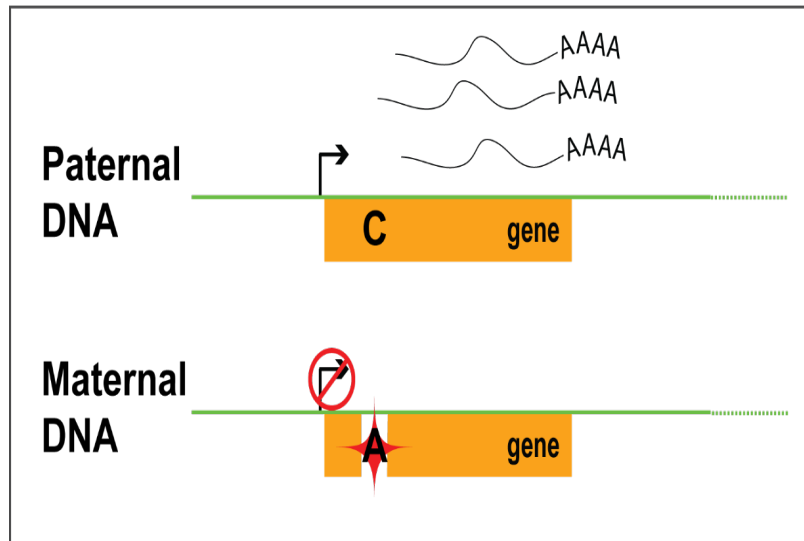
# How to build a personal genome



# Allele-specific binding and expression



Genomic variants affecting allele-specific behavior e.g. allele-specific binding (ASB)



e.g. allele-specific expression (ASE)



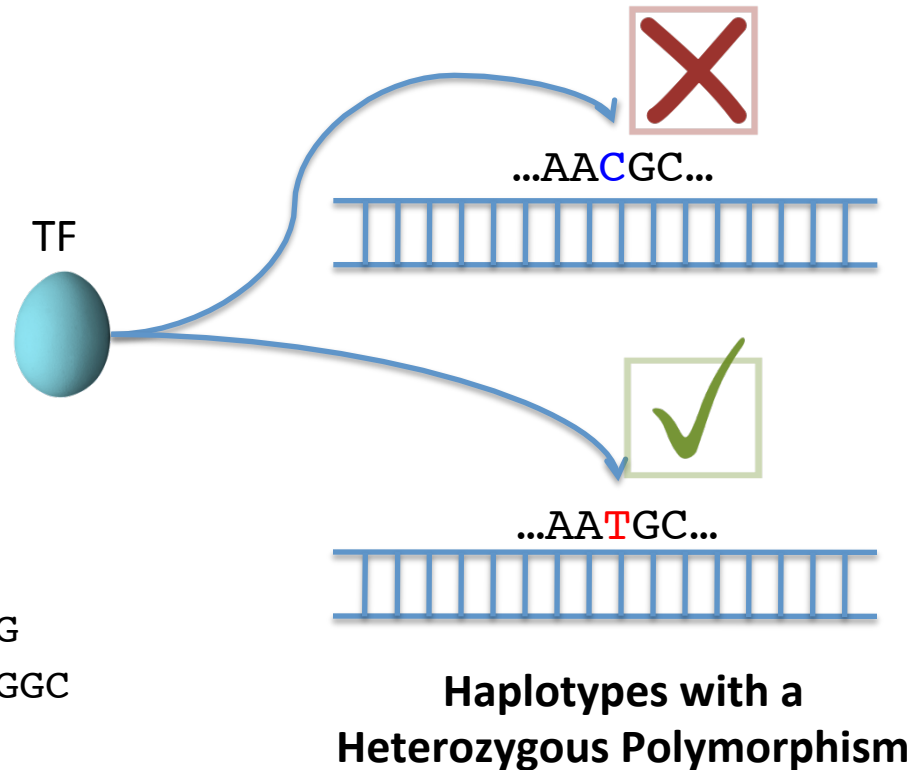
# Inferring Allele Specific Binding/Expression using Sequence Reads

## RNA/ChIP-Seq Reads

ACTTTGATAGCGTCAATG  
 CTTTGATAGCGTCAATGC  
 CTTTGATAGCGTCAACGC  
 TTGACAGCGTCAATGCAC  
 TGATAGCGTCAATGCACG  
 ATAGCGTCAATGCACGTC  
 TAGCGTCAATGCACGTCG  
 CGTCAACGCACGTCGGGA  
 GTCAATGCACGTCGAGAG  
 CAATGCACGTCGGGAGTT  
 AATGCACGTCGGGAGTTG  
 TGCACGTTGGGAGTTGGC

10 x T

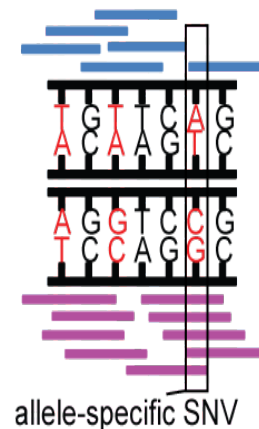
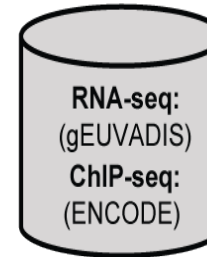
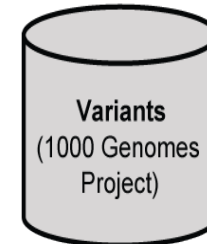
2 x C



Interplay of the annotation and individual sequence variants

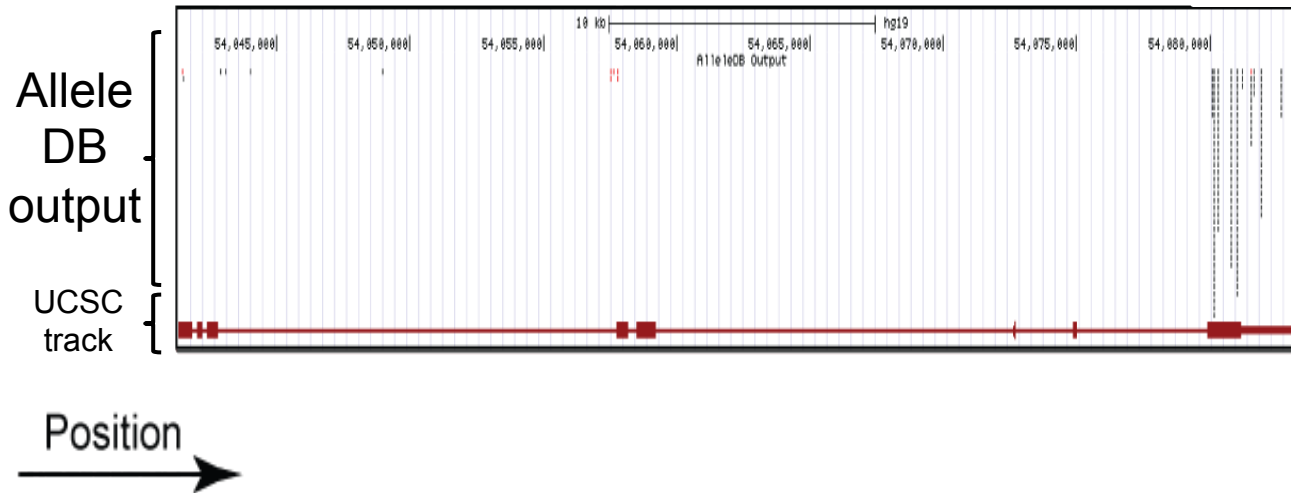
# AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes
2. Align ChIP-seq & RNA-seq reads
3. Detect allele-specific variants via a series of filters and tests



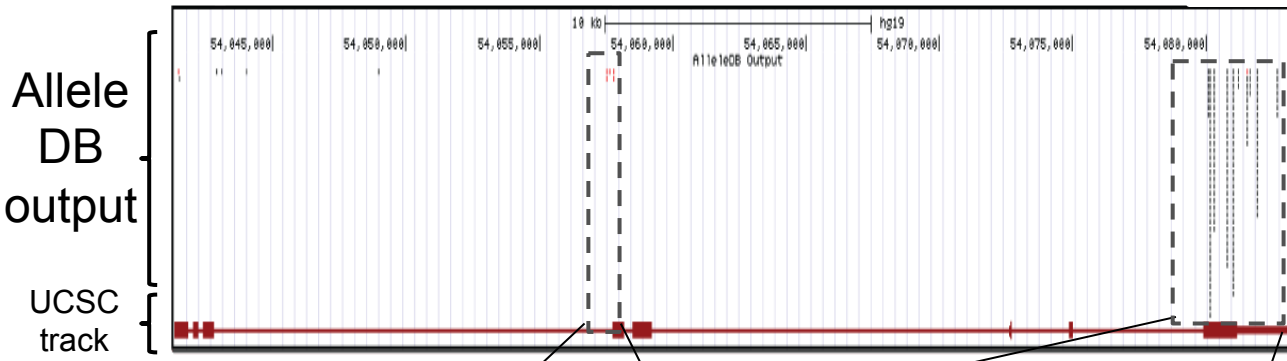
**Many Technical Issues:  
Reference bias, Ambiguous  
mapping bias, Over-dispersed  
(non binomial null)**

# AlleleDB: Annotating rare & common allele-specific variants over a population

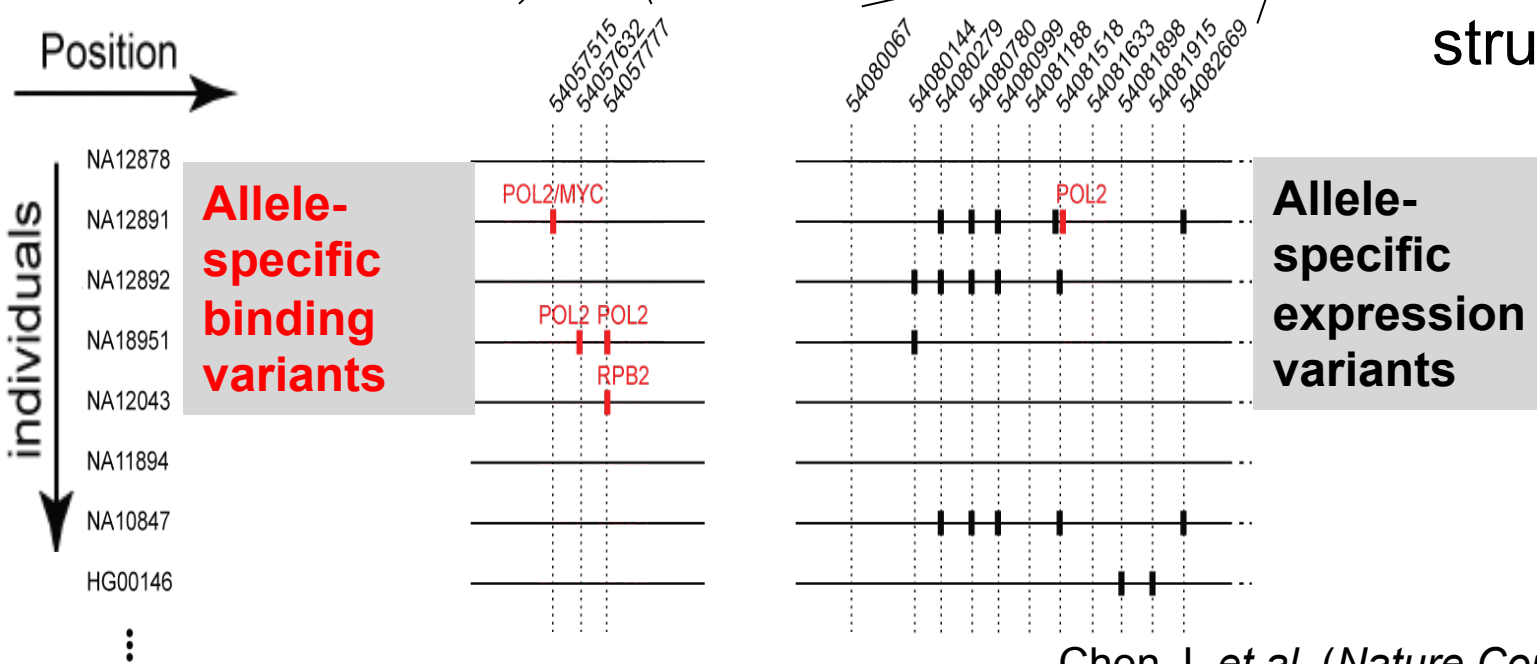


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

# AlleleDB: Annotating rare & common allele-specific variants over a population



- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

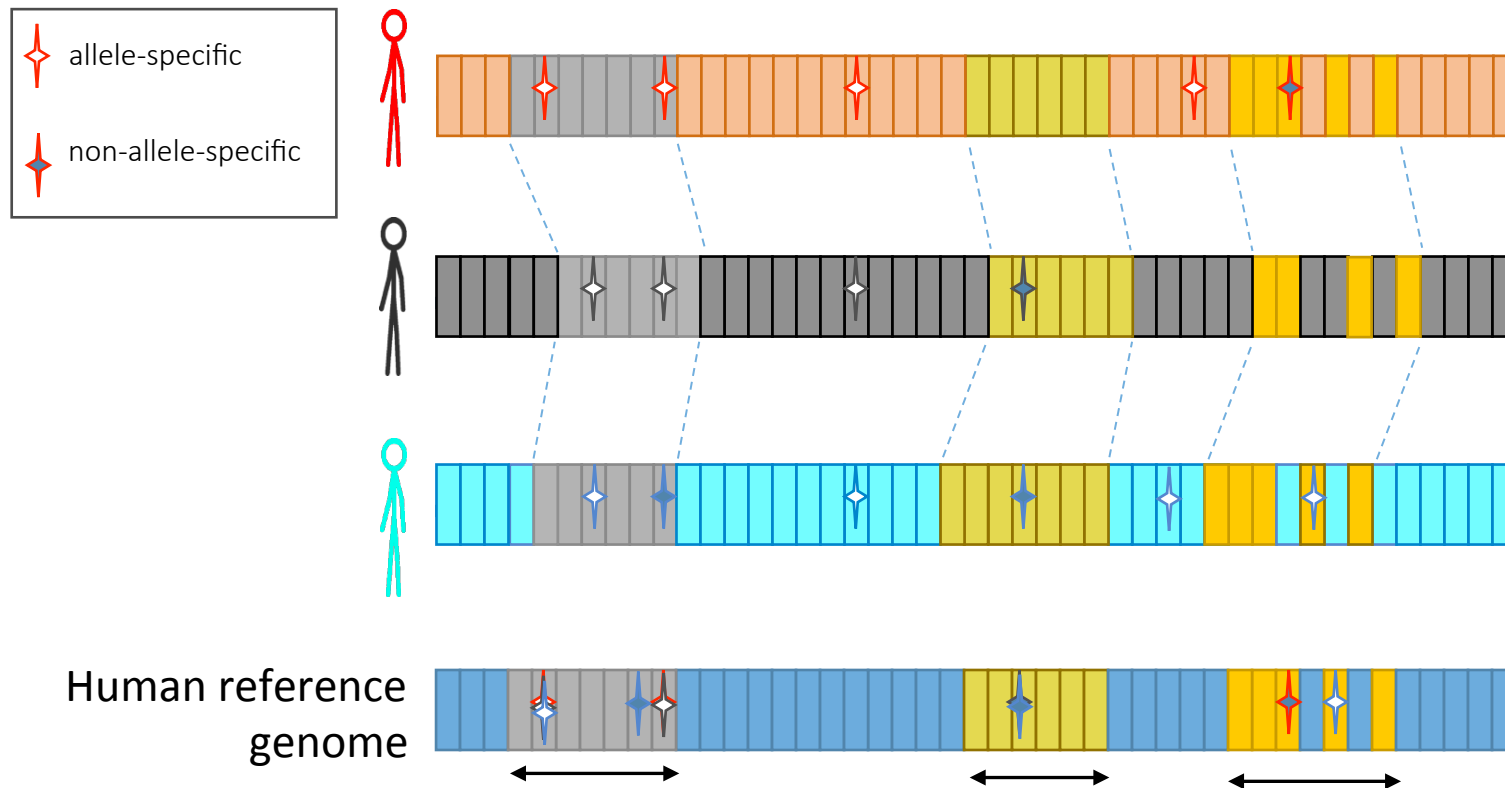


Chen J. et al. (*Nature Commun*, in press)

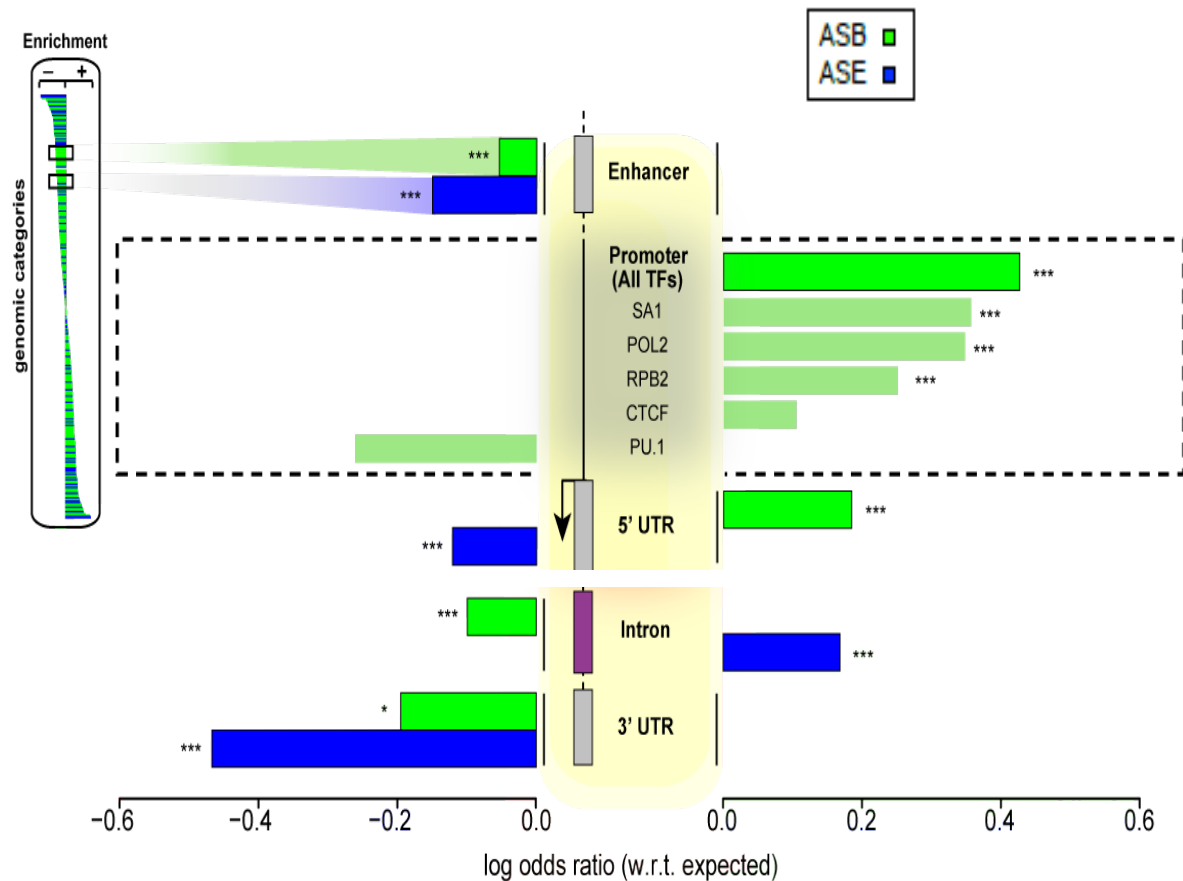
# Collecting ASE/ASB variants into allele-specific genomic regions

Does a particular genomic element have a higher tendency to be allele-specific?

Fisher's exact test, for the **enrichment** of allele-specific variants in the element (with respect to non-allele-specific variants that could potentially be called as allelic)



# Groups of elements that are enriched or depleted in allelic activity



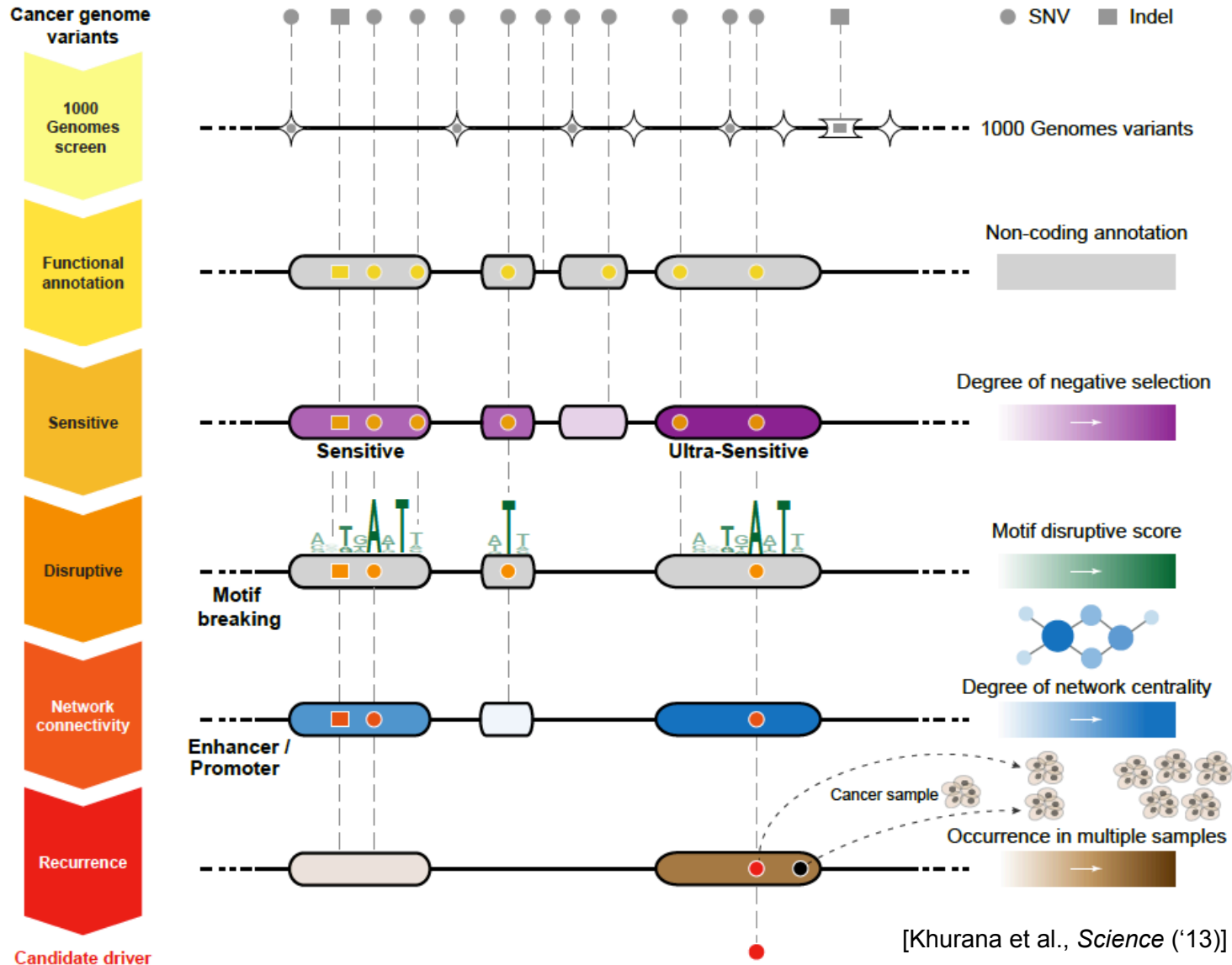
Chen J. *et al.* (*Nature Commun*, in press)



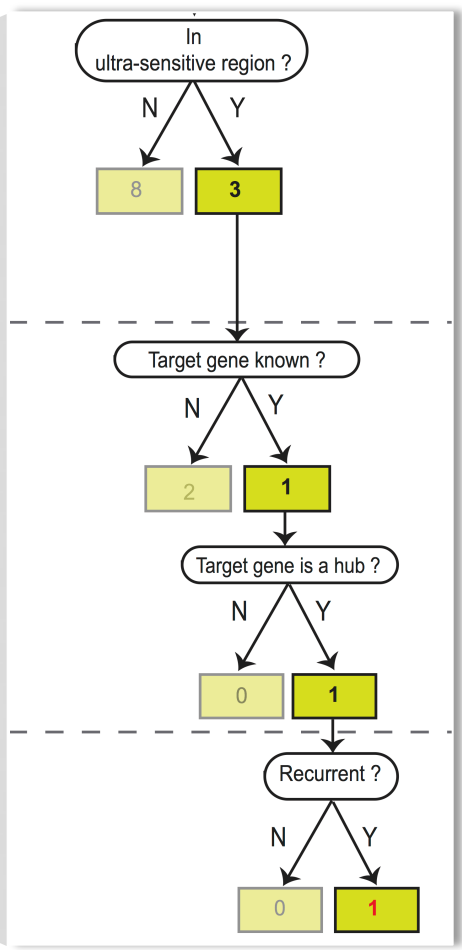
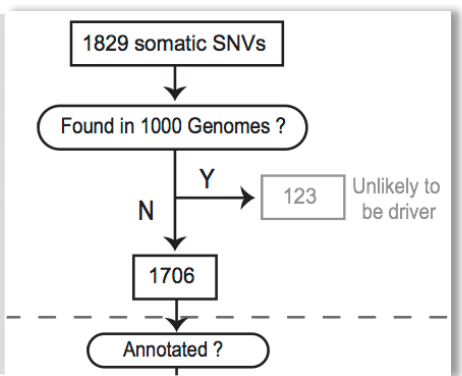
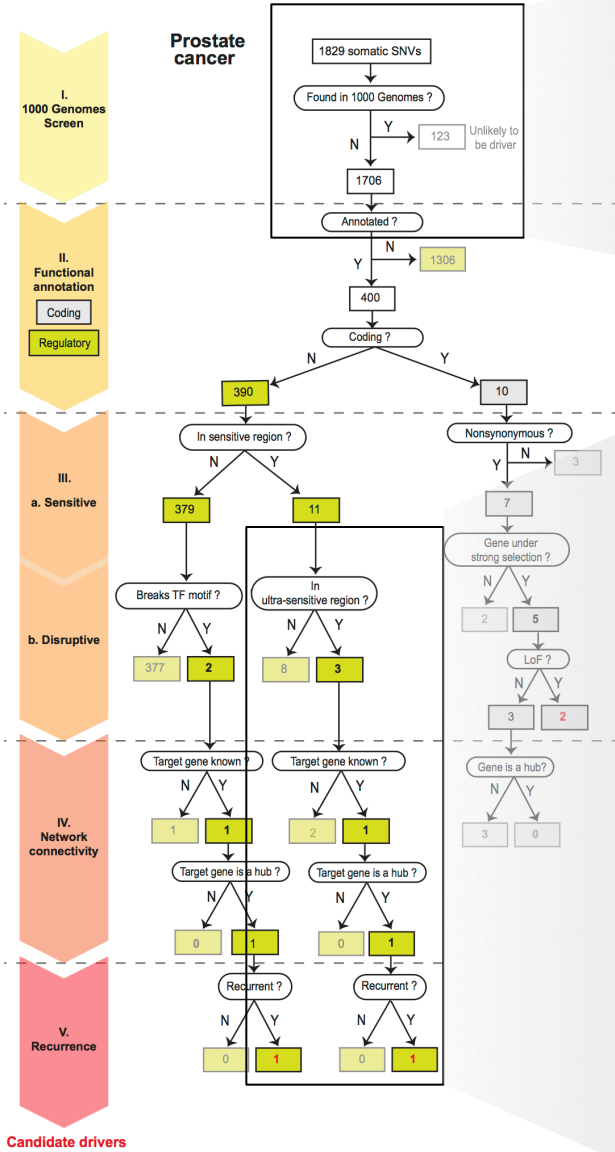
# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

# 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

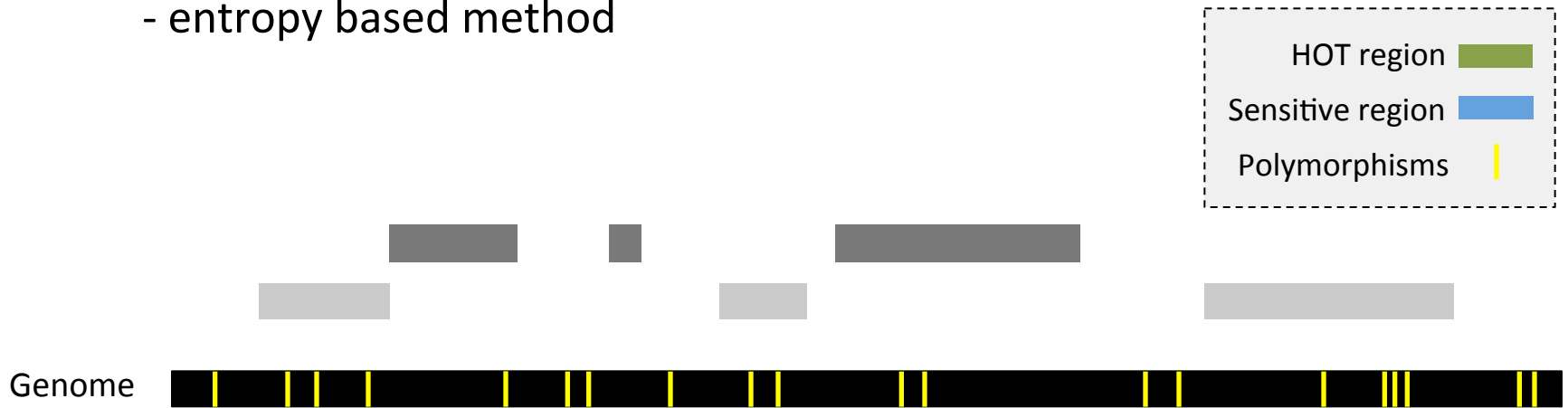
Highlighting variants

User Variants

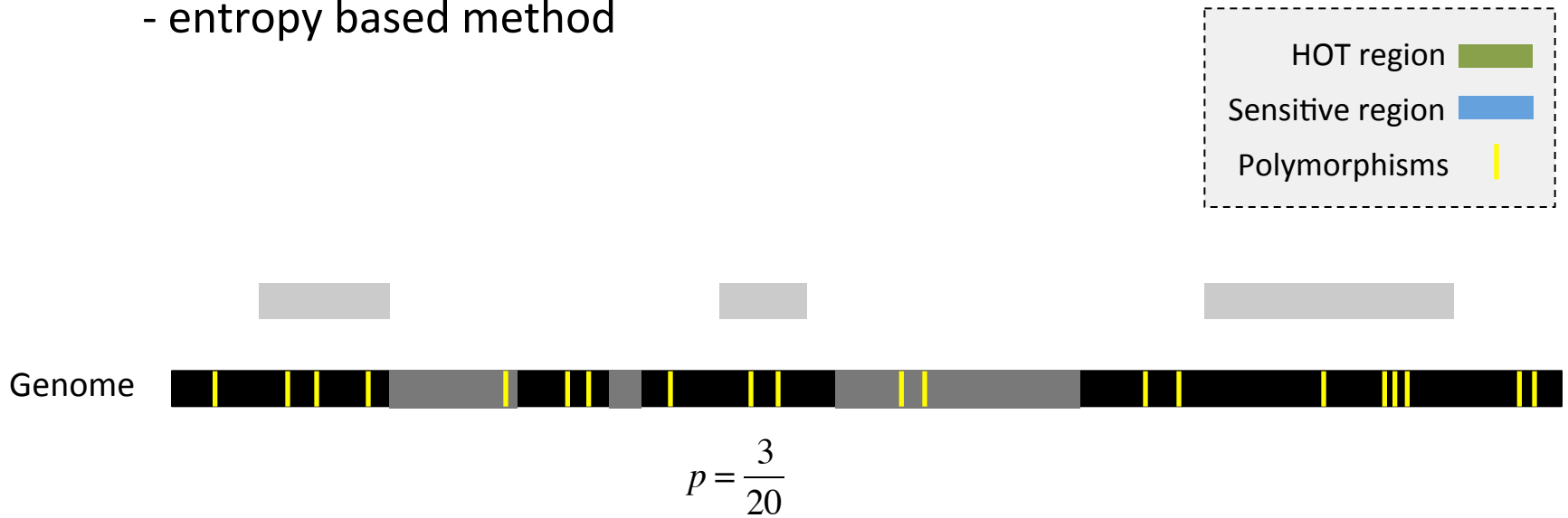
Variant Reports

FunSeq.gersteinlab.org

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

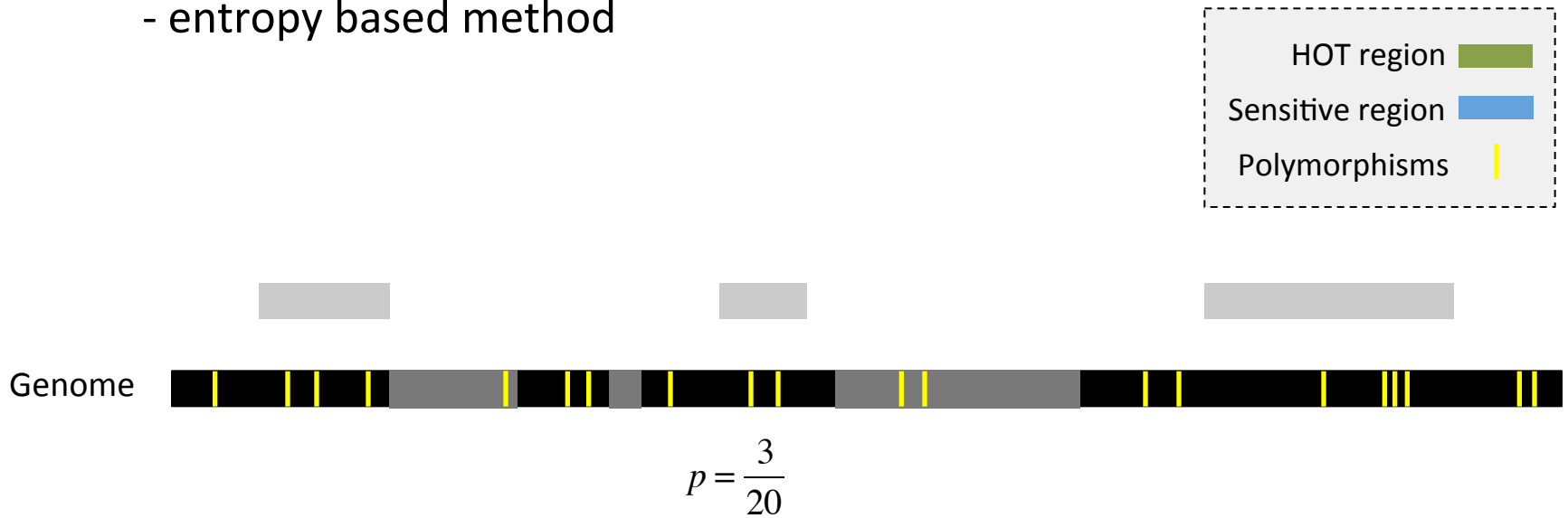


- Feature weight
  - Weighted with mutation patterns in natural polymorphisms  
(features frequently observed weight less)
  - entropy based method





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

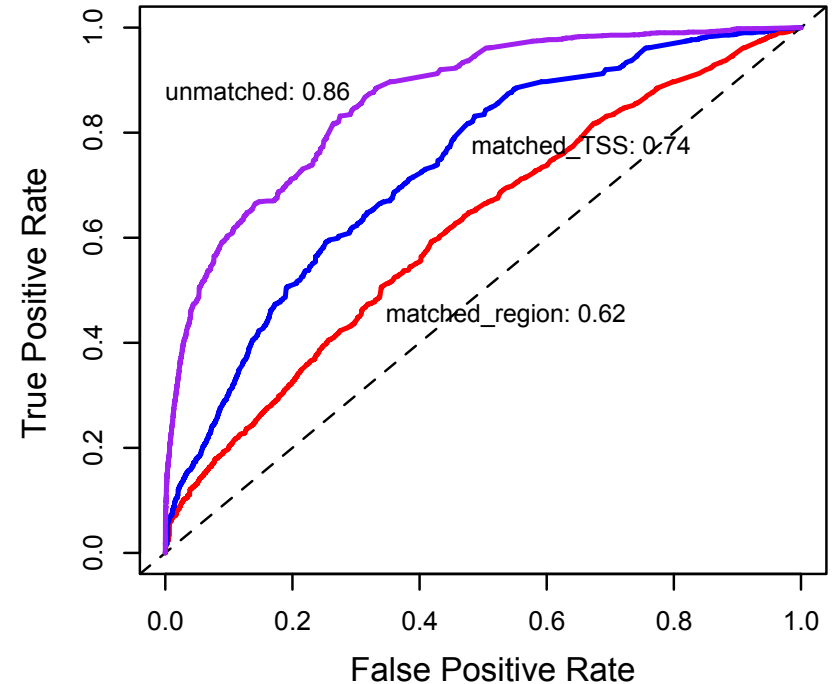
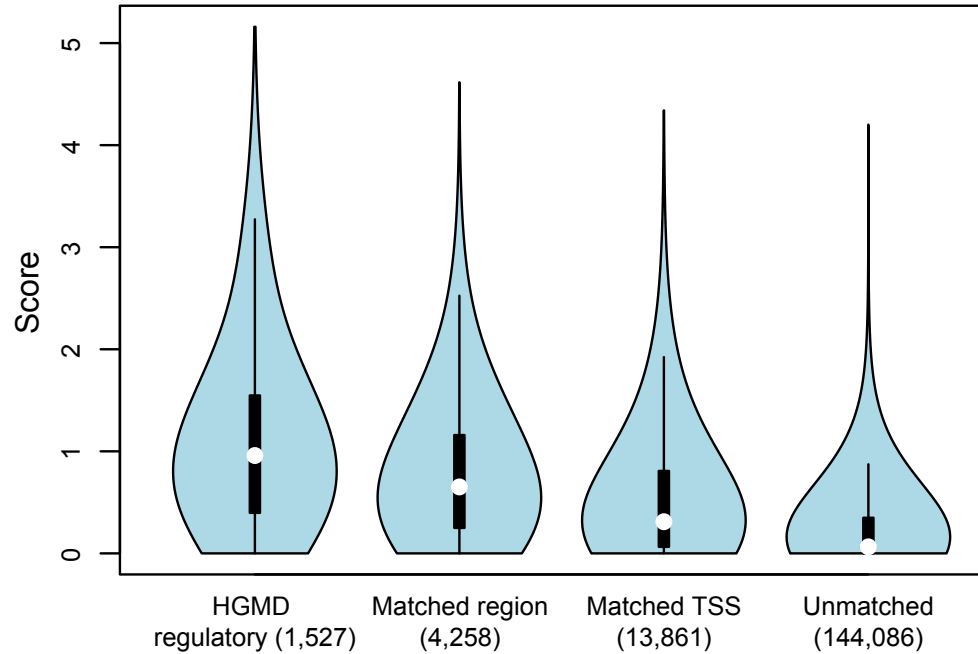


Feature weight:  $w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d)$

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

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

# Germline pathogenic variants show higher core scores than controls



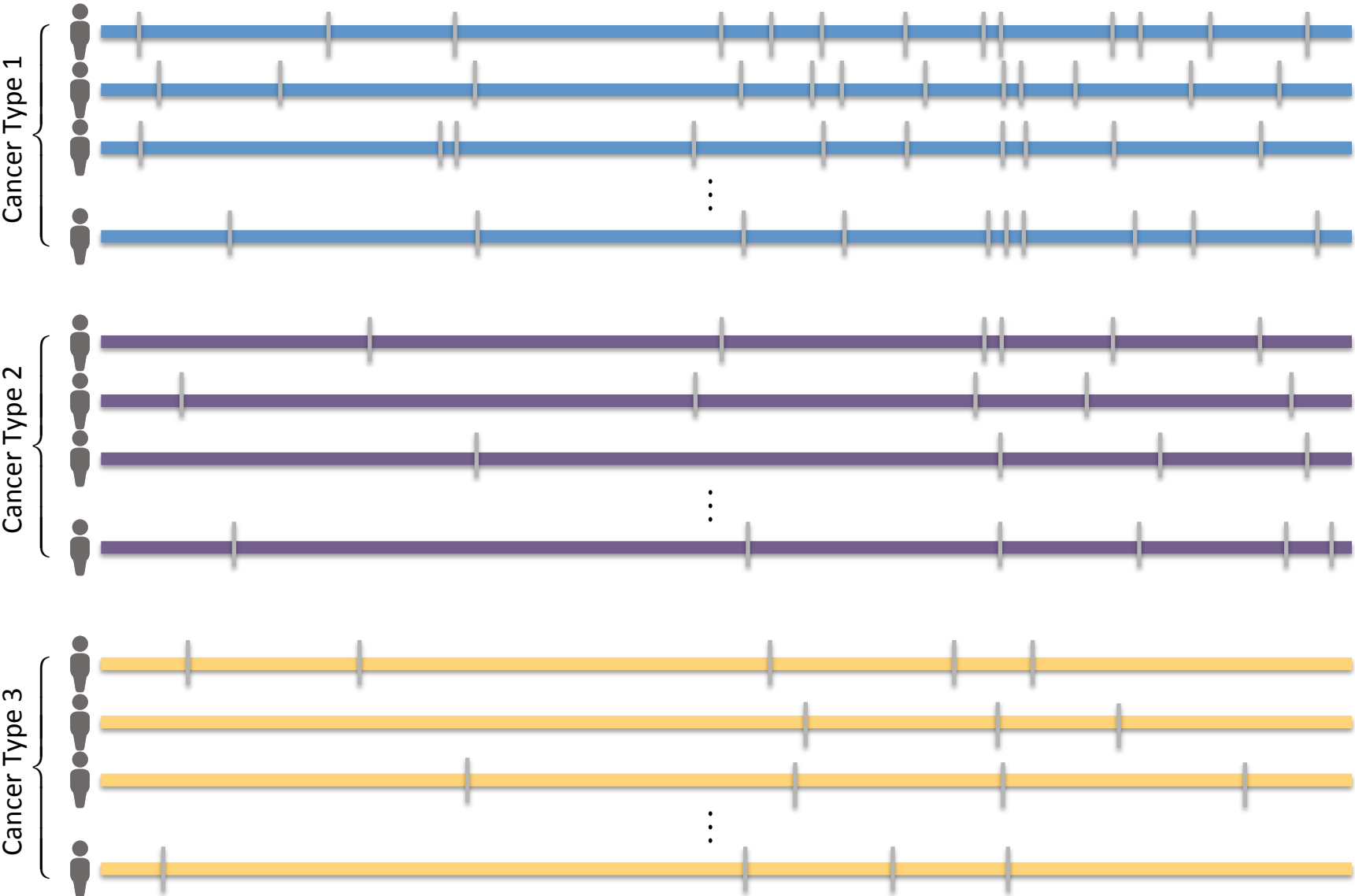
3 controls with natural polymorphisms (allele frequency  $\geq 1\%$  )

1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

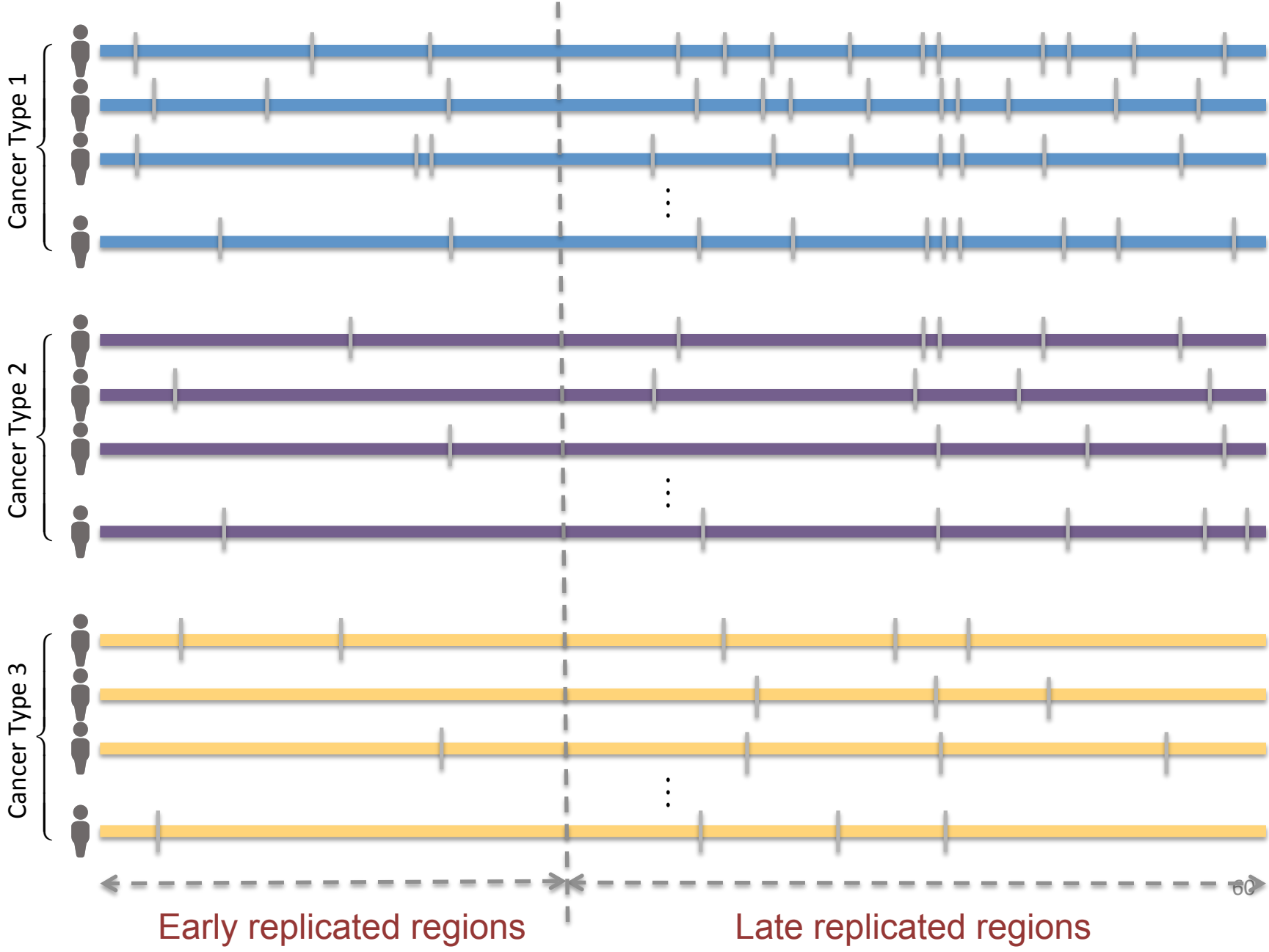
# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

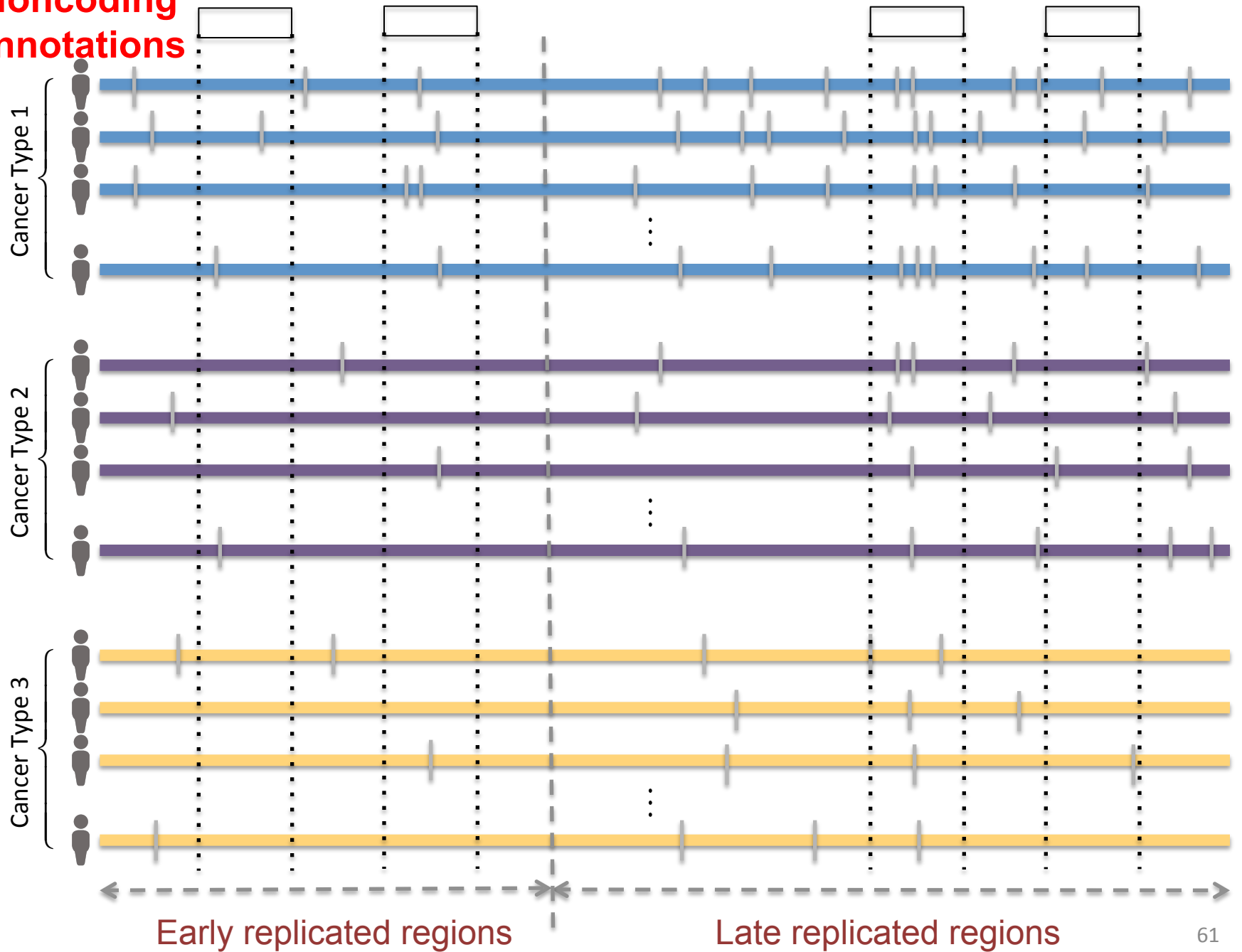
# Mutation recurrence



# Mutation recurrence

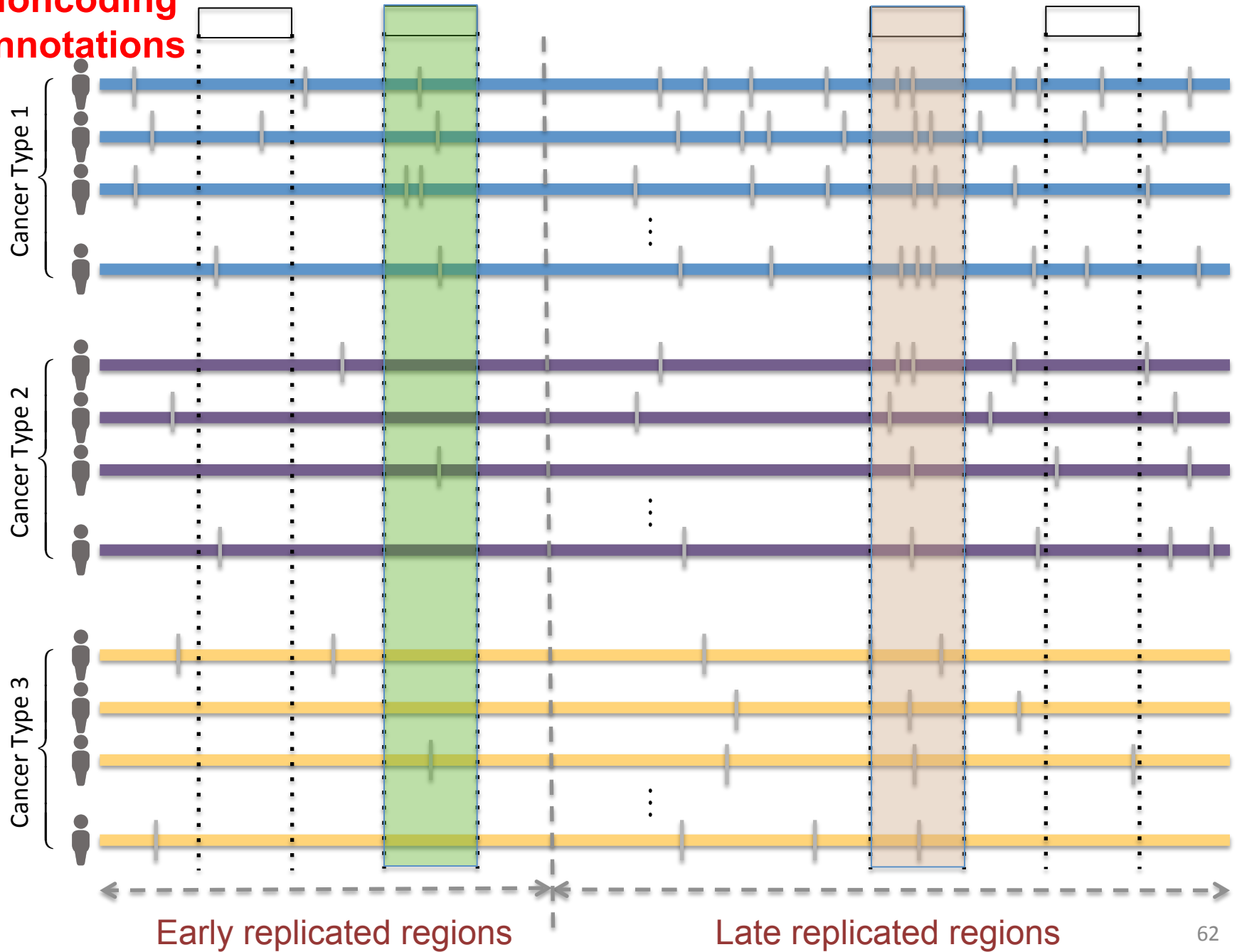


# Noncoding annotations

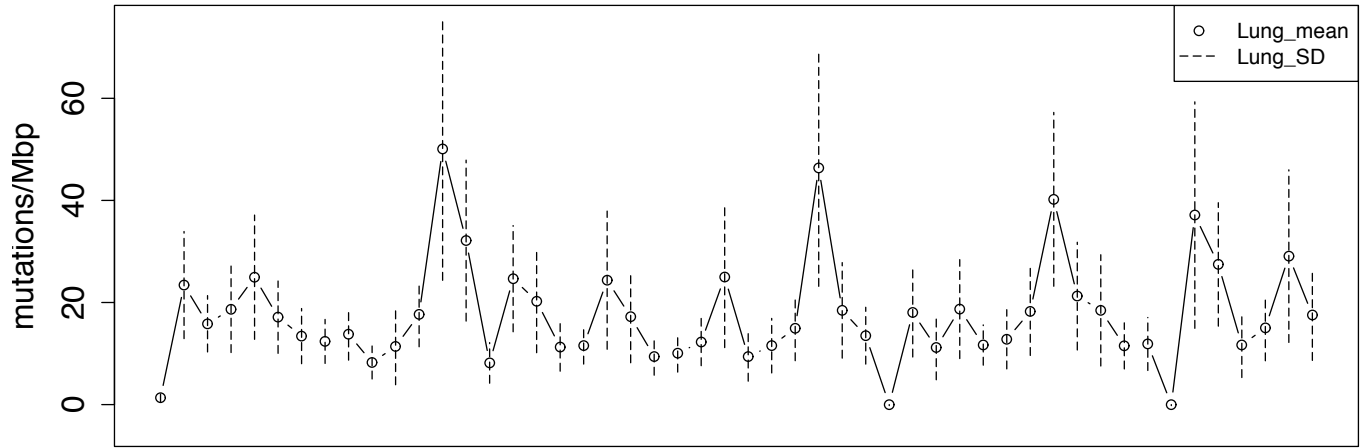
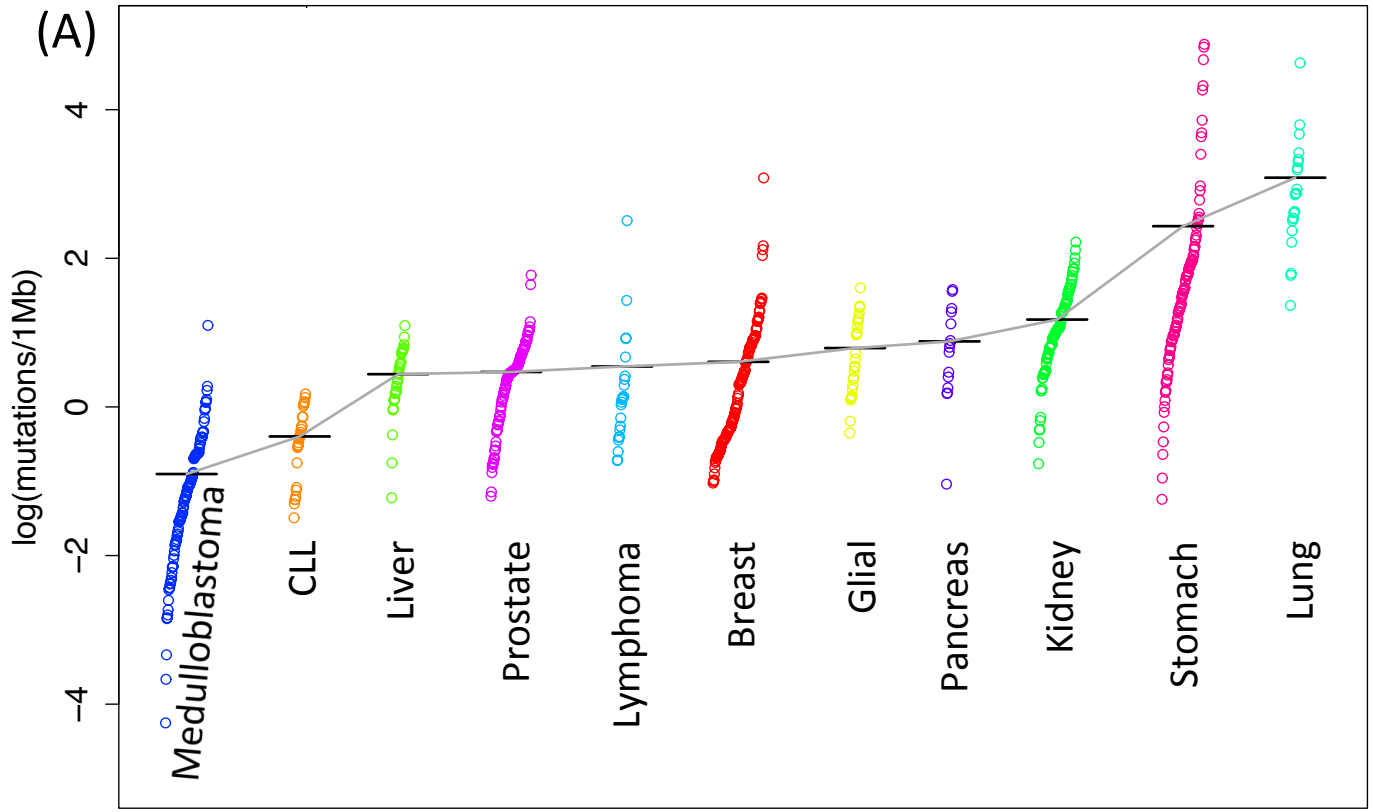




# Noncoding annotations



# Cancer Somatic Mutational Heterogeneity, across cancer types, samples & regions



1 Mbp genome regions (locations chosen at random)

# Cancer Somatic Mutation Modeling

- 3 models to evaluate the significance of mutation burden
- Suppose there are  $k$  genome elements. For element  $i$ , define:
  - $n_i$ : total number of nucleotides
  - $x_i$ : the number of mutations within the element
  - $p_i$ : the mutation rate
  - $R$ : the replication timing bin of the element

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

$$x_i : \text{Binomial}(n_i, p)$$

**Model 2: Varying Mutation Rate**

$$x_i | p_i : \text{Binomial}(n_i, p_i)$$

$$p_i : \text{Beta}(\mu, \sigma)$$

**Model 3: Varying Mutation Rate with Replication Timing Correction**

$$x_i | p_i : \text{Binomial}(n_i, p_i)$$

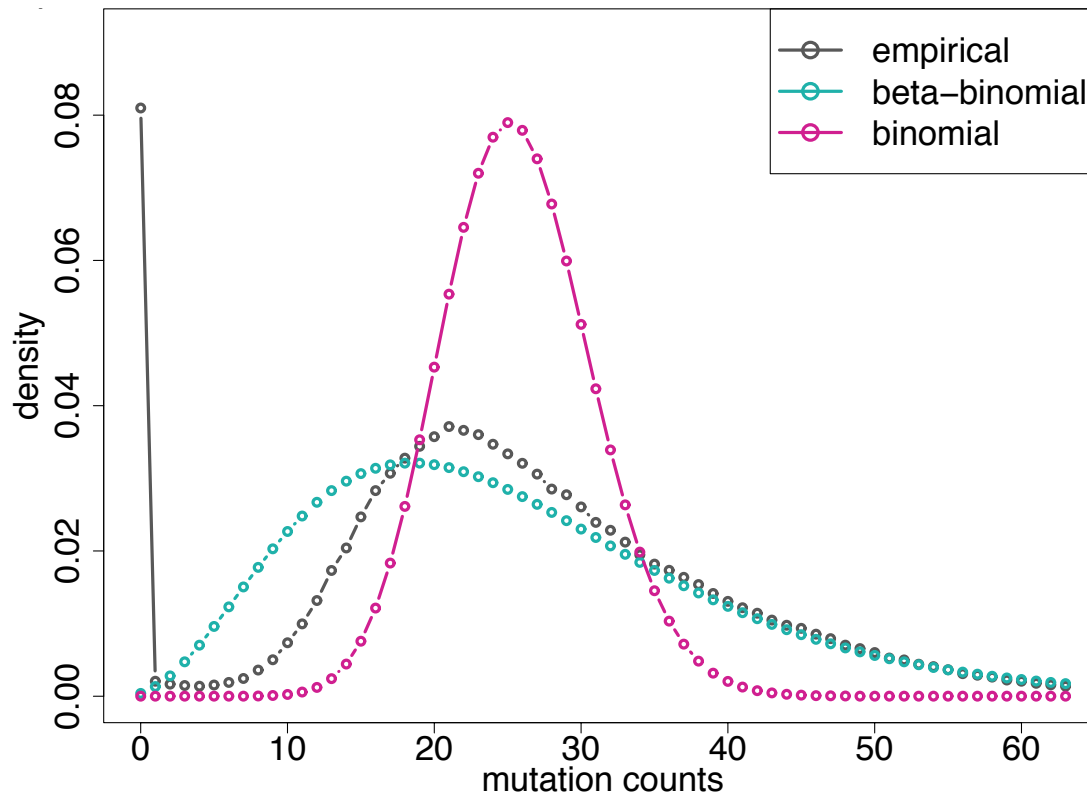
$$p_i : \text{Beta}(\mu | \mathbf{R}, \sigma | \mathbf{R})$$

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

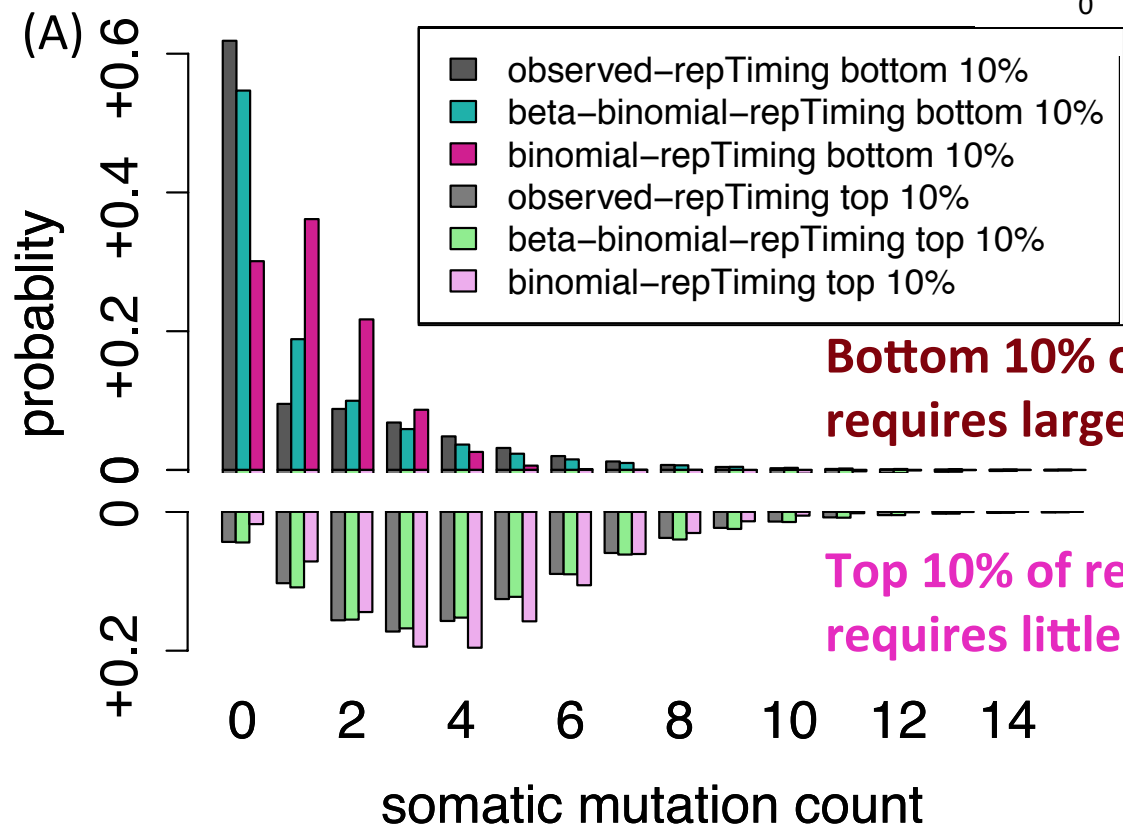
[Lochovsky et al. *NAR* ('15)]

# 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

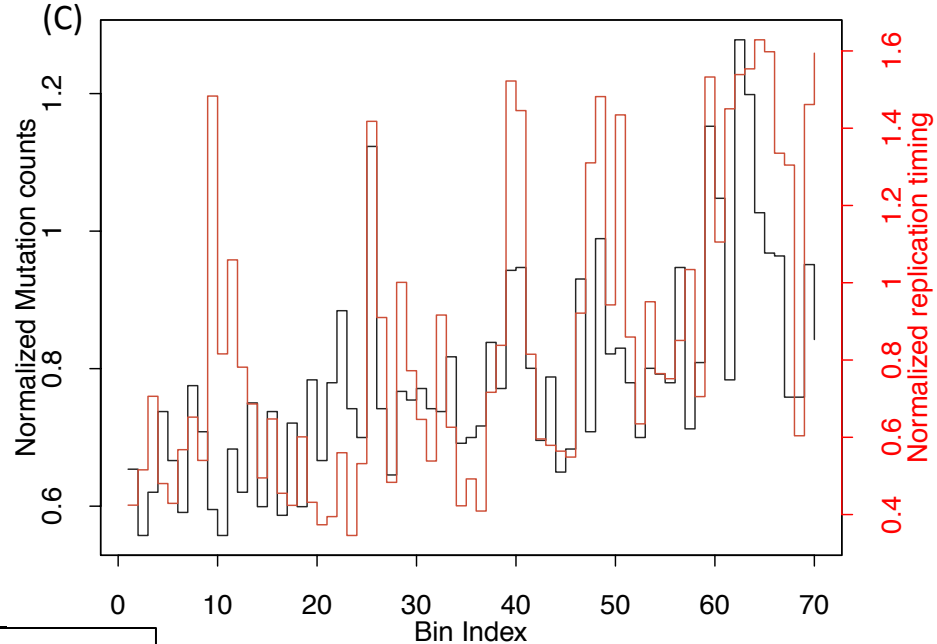


# Adding DNA replication timing correction further improves the beta-binomial model



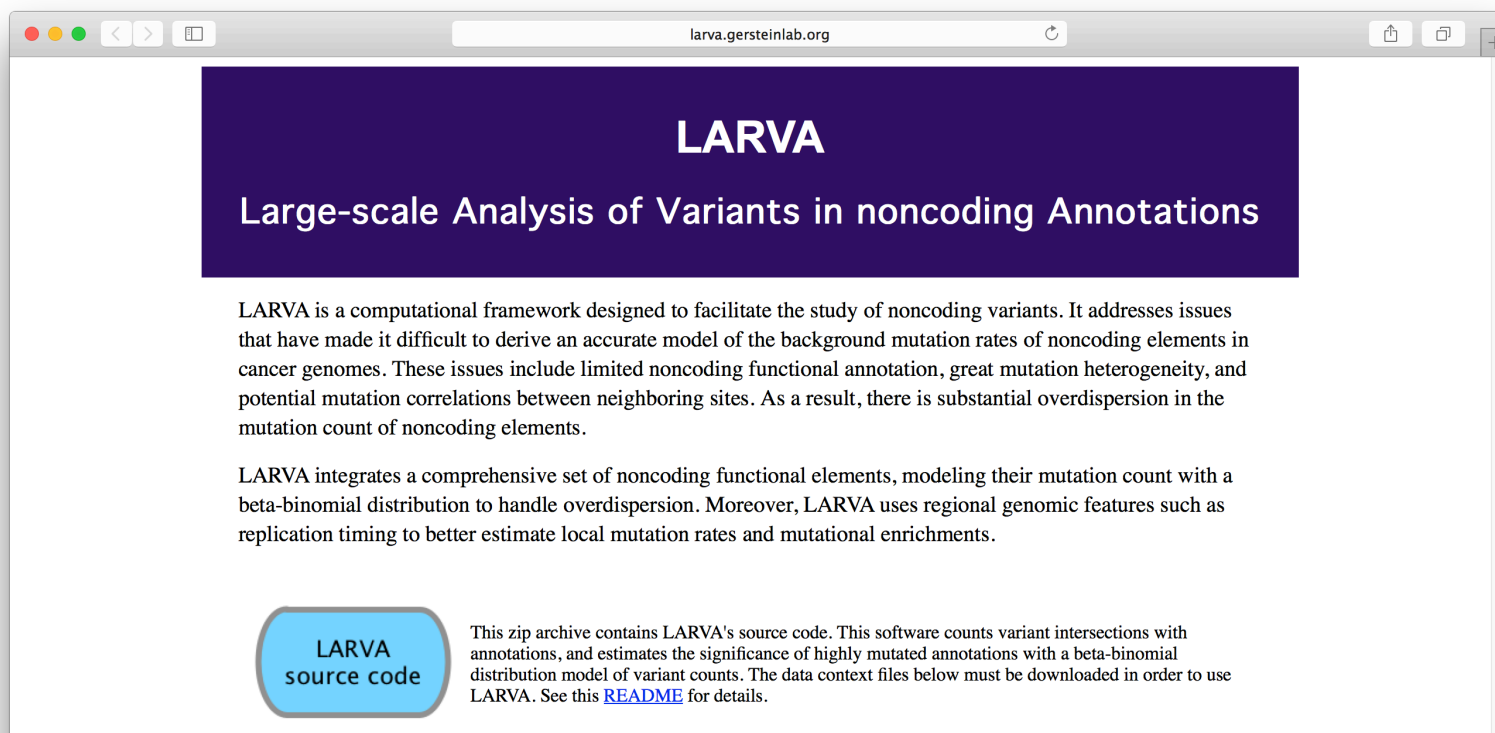
**Bottom 10% of rep. timing bins requires large correction**

**Top 10% of rep. timing bins requires little correction**



# LARVA Implementation

- <http://larva.gersteinlab.org/>
- Freely downloadable C++ program
  - Verified compilation and correct execution on Linux
- A Docker image is also available to download
  - Runs on any operating system supported by Docker
- Running time on transcription factor binding sites (a worst case input size) is ~80 min
  - Running time scales linearly with the number of annotations in the input



The screenshot shows a web browser window with the URL [larva.gersteinlab.org](http://larva.gersteinlab.org/). The page features a dark purple header with the text "LARVA" in white, followed by the subtitle "Large-scale Analysis of Variants in noncoding Annotations" in white. Below the header, there is a paragraph of text describing LARVA as a computational framework for studying noncoding variants. A second paragraph explains how LARVA integrates noncoding functional elements and regional genomic features. At the bottom, there is a blue button labeled "LARVA source code" and a link to a README file.

**LARVA**  
Large-scale Analysis of Variants in noncoding Annotations

LARVA is a computational framework designed to facilitate the study of noncoding variants. It addresses issues that have made it difficult to derive an accurate model of the background mutation rates of noncoding elements in cancer genomes. These issues include limited noncoding functional annotation, great mutation heterogeneity, and potential mutation correlations between neighboring sites. As a result, there is substantial overdispersion in the mutation count of noncoding elements.

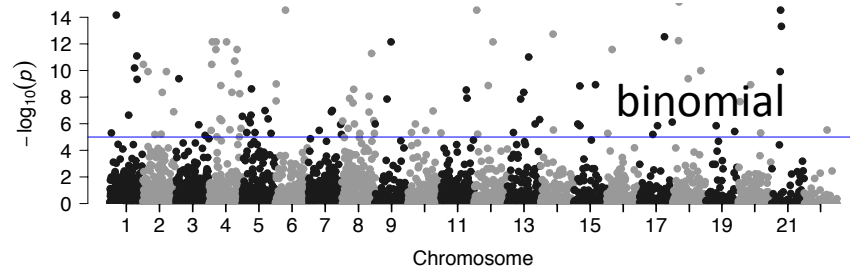
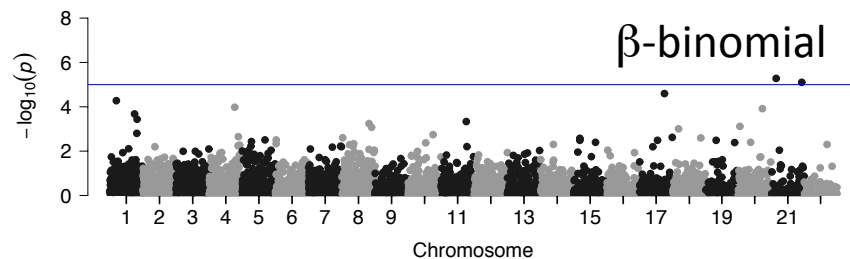
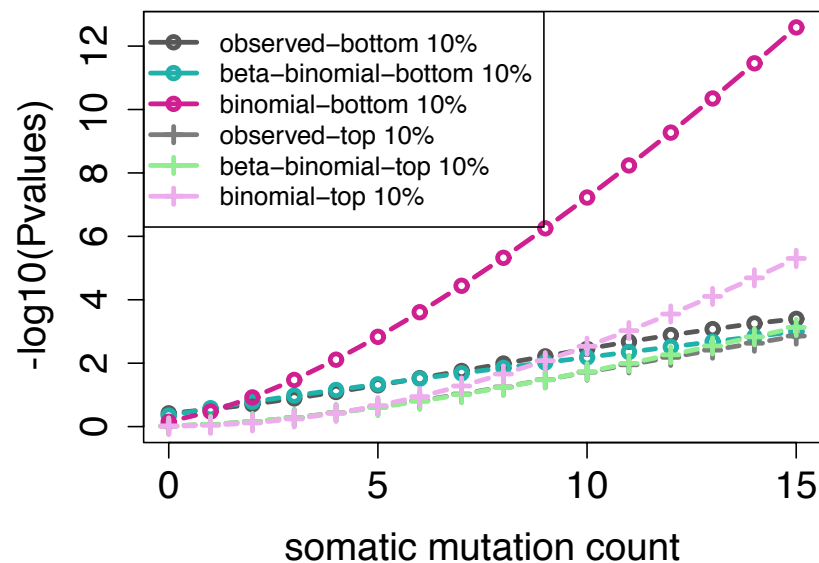
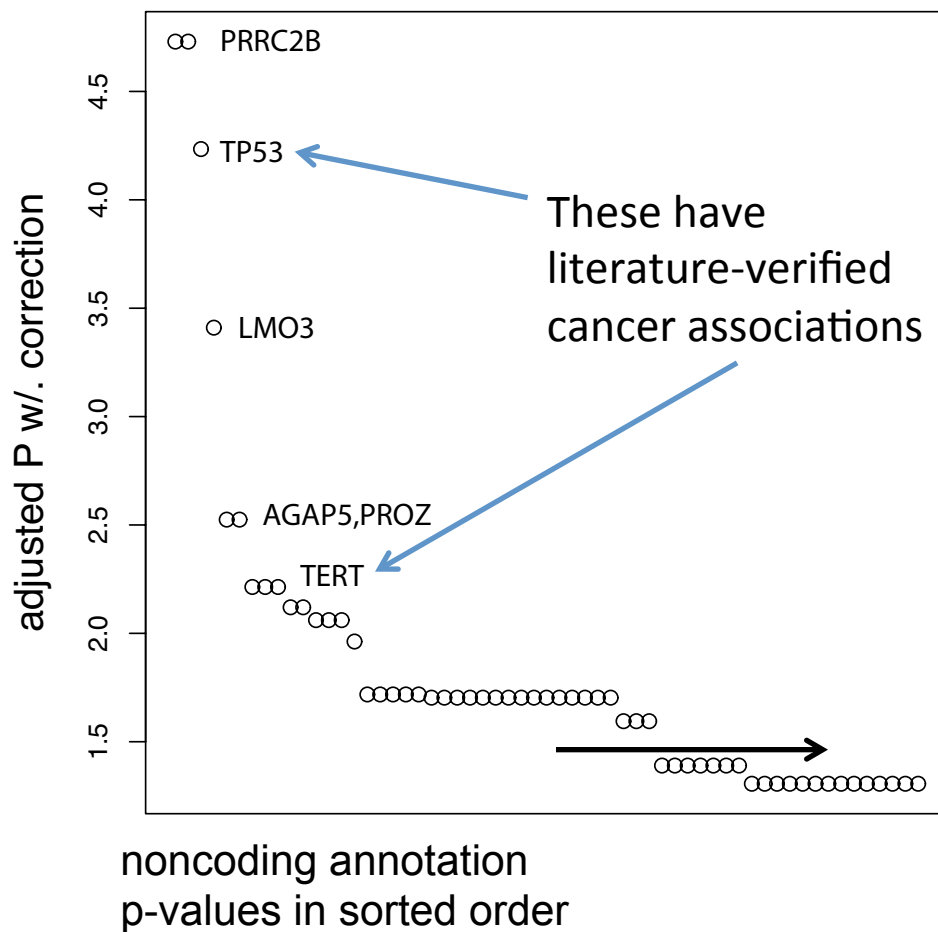
LARVA integrates a comprehensive set of noncoding functional elements, modeling their mutation count with a beta-binomial distribution to handle overdispersion. Moreover, LARVA uses regional genomic features such as replication timing to better estimate local mutation rates and mutational enrichments.

**LARVA source code**

This zip archive contains LARVA's source code. This software counts variant intersections with annotations, and estimates the significance of highly mutated annotations with a beta-binomial distribution model of variant counts. The data context files below must be downloaded in order to use LARVA. See this [README](#) for details.

# LARVA Results

## TSS LARVA results





# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

# Analyzing Personal Genomes: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with **STRESS** cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1
  - Annotating non-coding regions on different scales with **MUSIC**
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
- Non-coding Variants #2
  - Prioritizing using **AlleleDB** in terms of allelic elements
    - Having observed difference in molecular activity in many contexts
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in binomial
    - Parameterized according to replication timing

**AlleleDB**.gersteinlab.org

J **Chen**, J Rozowsky,  
TR Galeev, A Harmanci,  
R Kitchen, J Bedford,  
A Abyzov, Y Kong, L Regan

**CostSeq2**

P **Muir**, S Li, S Lou,  
D Wang, DJ Spakowicz,  
L Salichos, J Zhang, F Isaacs,  
J Rozowsky

**FunSeq**.gersteinlab.org

- & -

**FunSeq2**.gersteinlab.org

Y **Fu**, E **Khurana**, Z Liu,  
S Lou, J Bedford, XJ Mu, KY  
Yip, V Colonna, XJ Mu, ... ,

**1000 Genomes**

**Project** Consortium, et al

**LARVA**.gersteinlab.org

L **Lochovsky**,  
J **Zhang**, Y Fu,  
E Khurana

**MUSIC**.gersteinlab.org

A **Harmanci**,  
J Rozowsky

archive.gersteinlab.org/proj/

**netsnp**

E **Khurana**, Y **Fu**,  
J Chen

**STRESS**.molmovdb.org

D **Clarke**, A **Sethi**, S Li,  
S Kumar, R W.F. Chang,  
J Chen



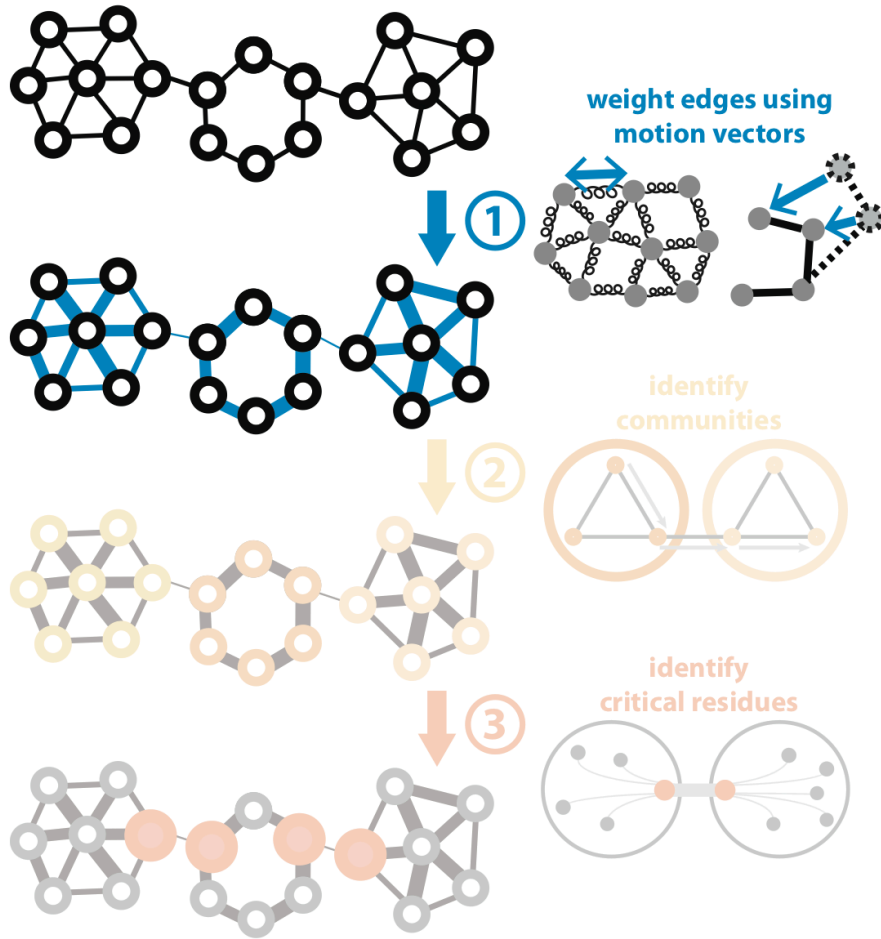
**Acknowledgments**

Hiring Postdocs. See [gersteinlab.org/jobs](http://gersteinlab.org/jobs)

**Extra**



# Predicting Allosterically-Important Residues within the Interior



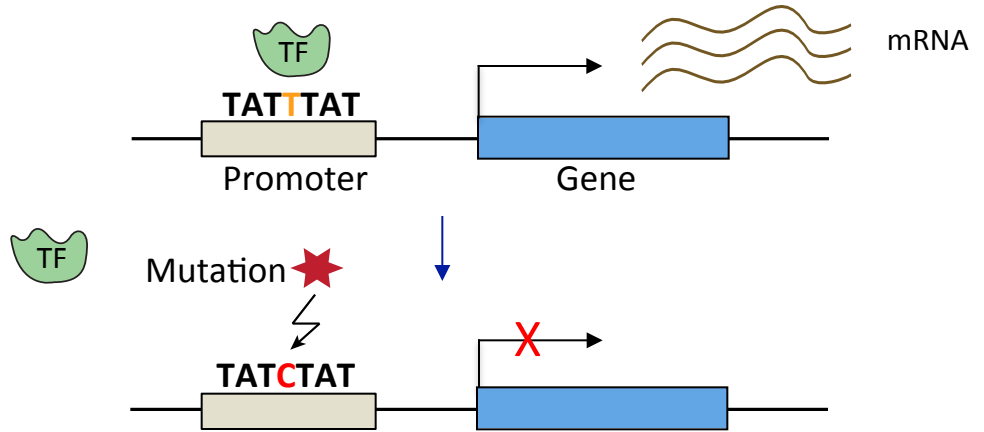
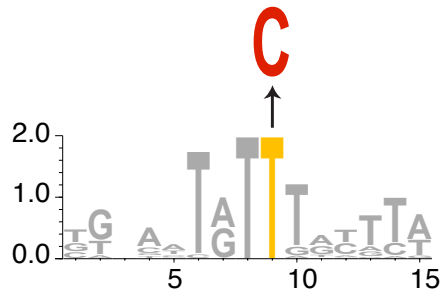
$$Cov_{ij} = \langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle$$

$$C_{ij} = Cov_{ij} / \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

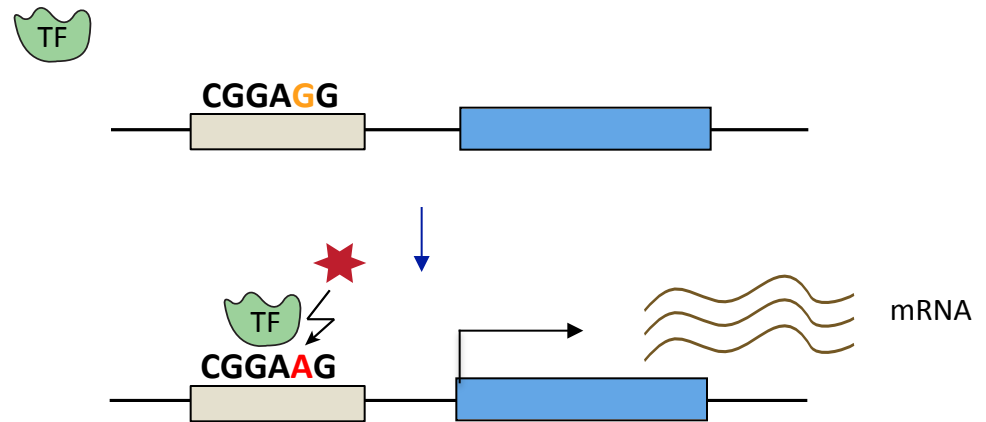
$$D_{ij} = -\log(|C_{ij}|)$$

# Loss- and gain- of motif mutations

## Loss-of-motif



## Gain-of-motif



# Many Technical Issues in Determining ASE/ASB: Reference Bias (naïve alignment against reference)

ASE/ASB Example:

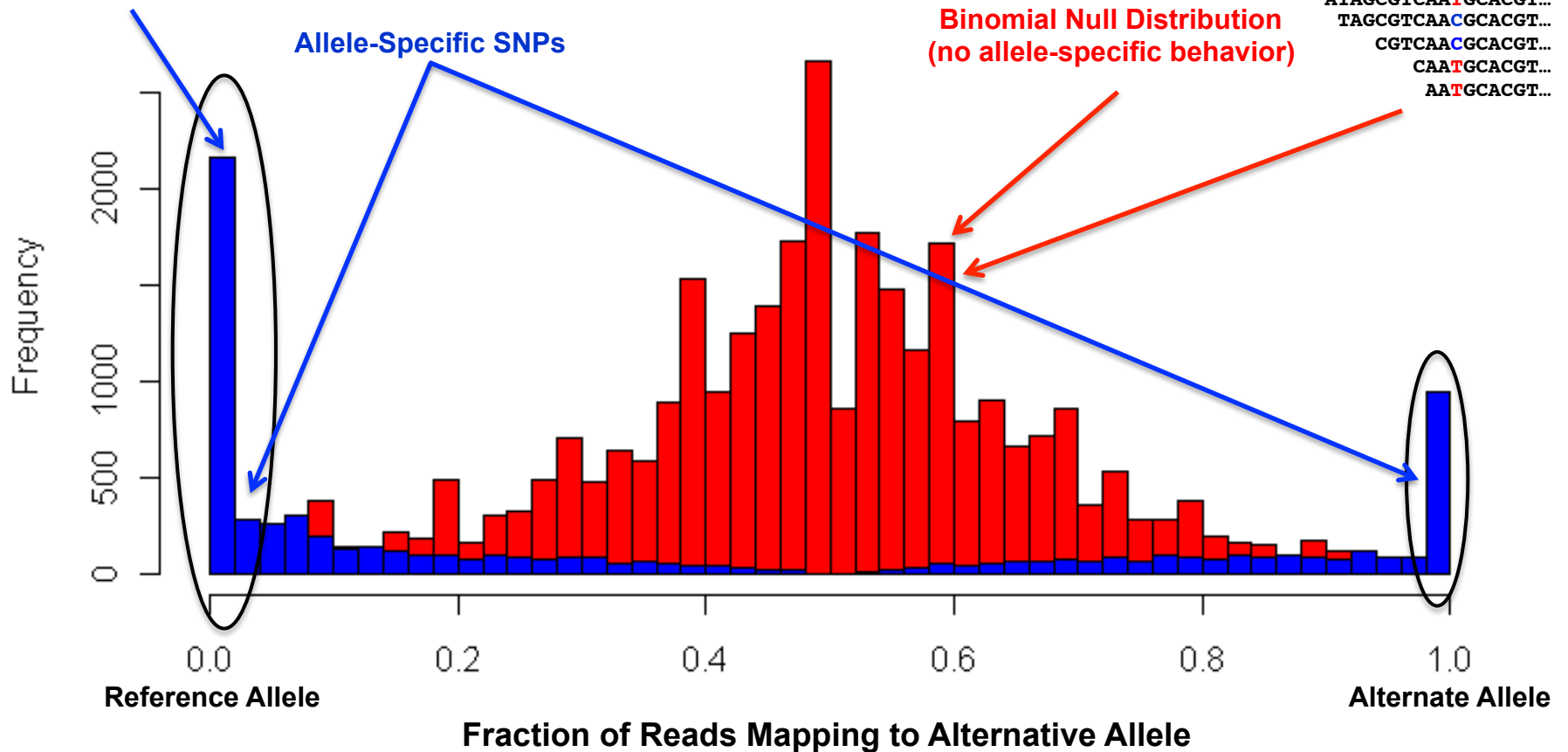
```

...GTCAATGCAC
...GTCAATGCACG
...GTCAATGCACGTC
...GTCAATGCACGTCG
...GTCAATGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGGGAGTT
AATGCACGTCGGGAGTT
    
```

Null Example:

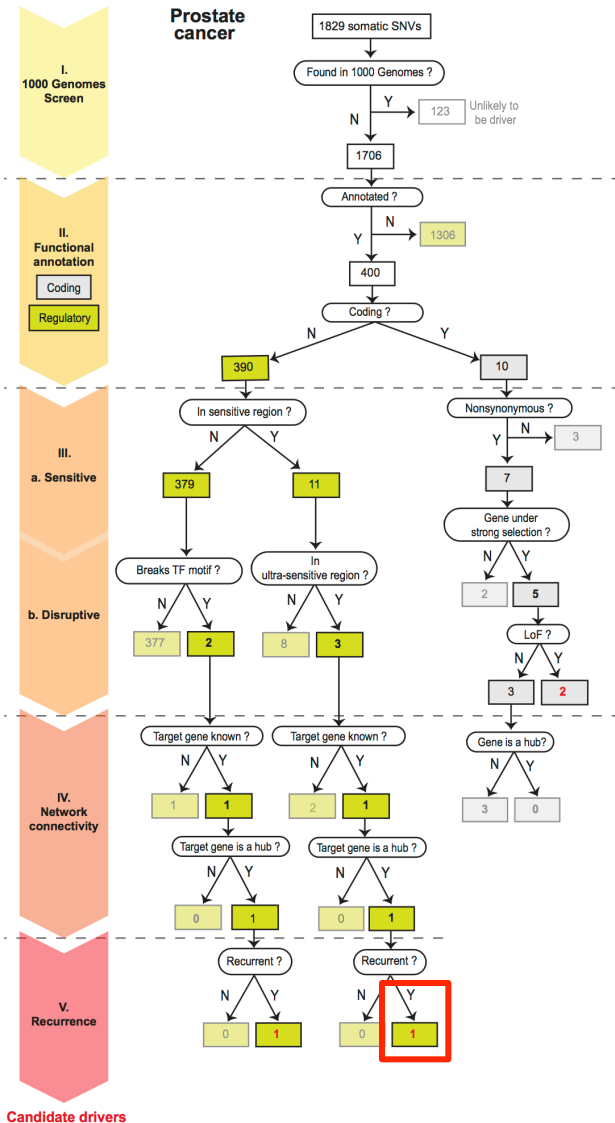
```

ACTTTGATAGCGTCAATG
CTTTGATAGCGTCAACGC
TTGACAGCGTCAATGCAC
ATAGCGTCAATGCACGT...
TAGCGTCAACGCACGT...
CGTCAACGCACGT...
CAATGCACGT...
AATGCACGT...
    
```





# 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

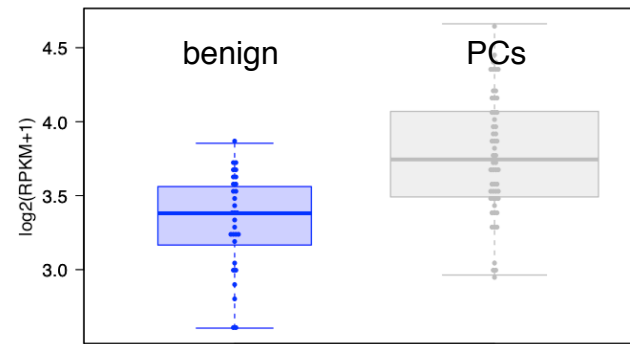
Sanger sequencing of *FAM48A* binding site (~570 bp) in *WDR74* promoter from 19 additional samples

..ACGGT...Tc|c|T|CC...GT|G|A|GA...ATAGA..

— chr11: 62,609,084

— chr11: 62,609,138

- ❑ *WDR74* shows increased expression in tumor samples



# Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - Please read permissions statement at [www.gersteinlab.org/misc/permissions.html](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.
- 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>