

Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

Mark Gerstein, Yale

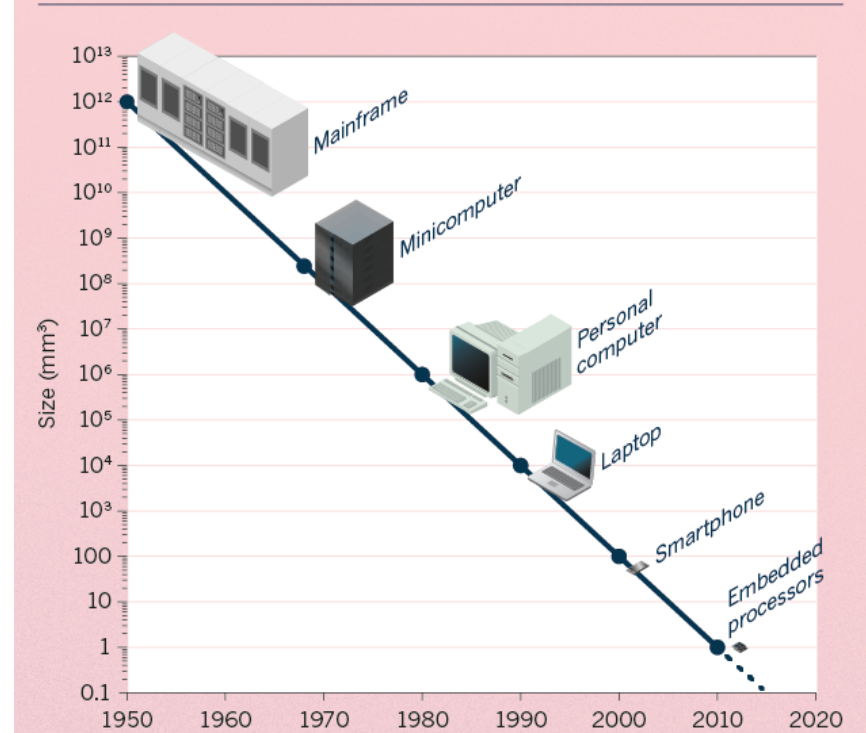
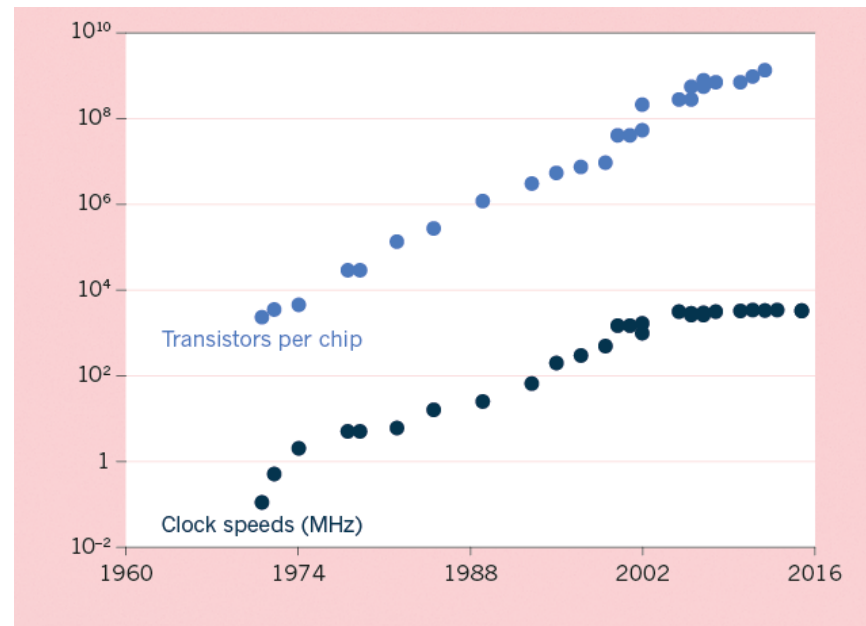
**Slides freely downloadable from
Lectures.GersteinLab.org**

& “tweetable” (via [@markgerstein](https://twitter.com/markgerstein)).

See last slide for more info.

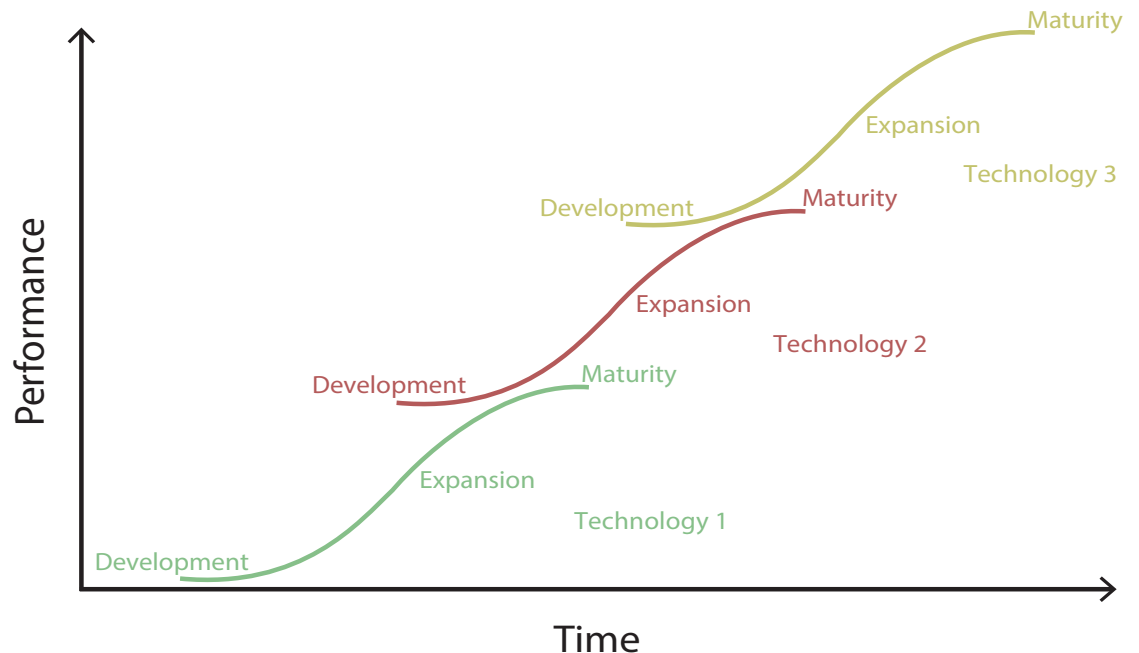
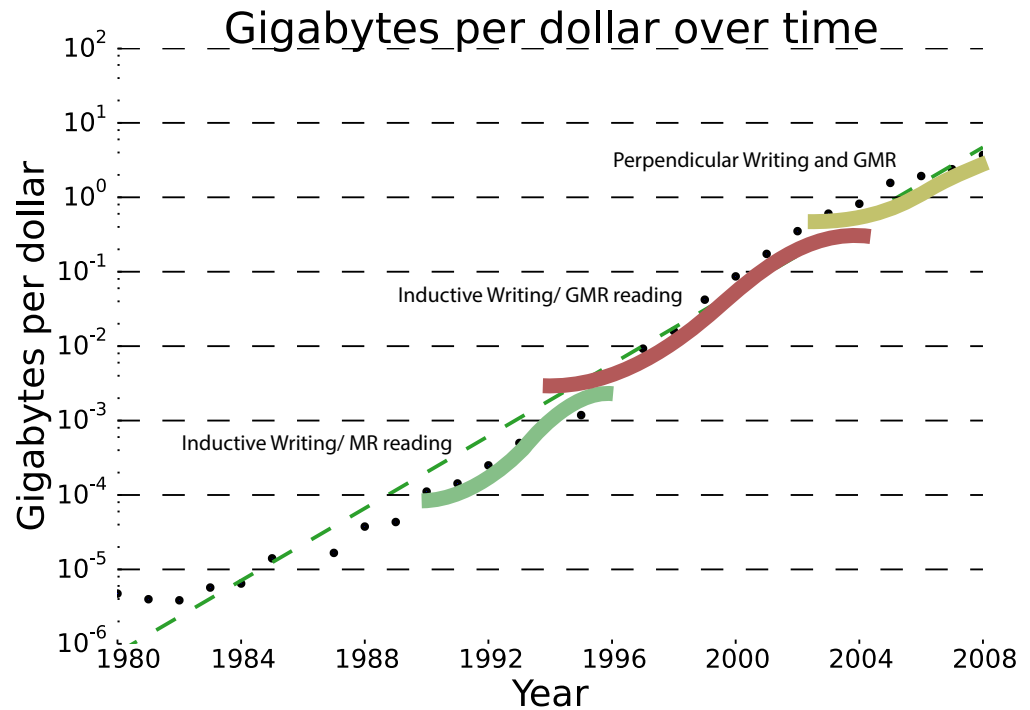
Moore's Law: Exponential Scaling of Computer Technology

- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.



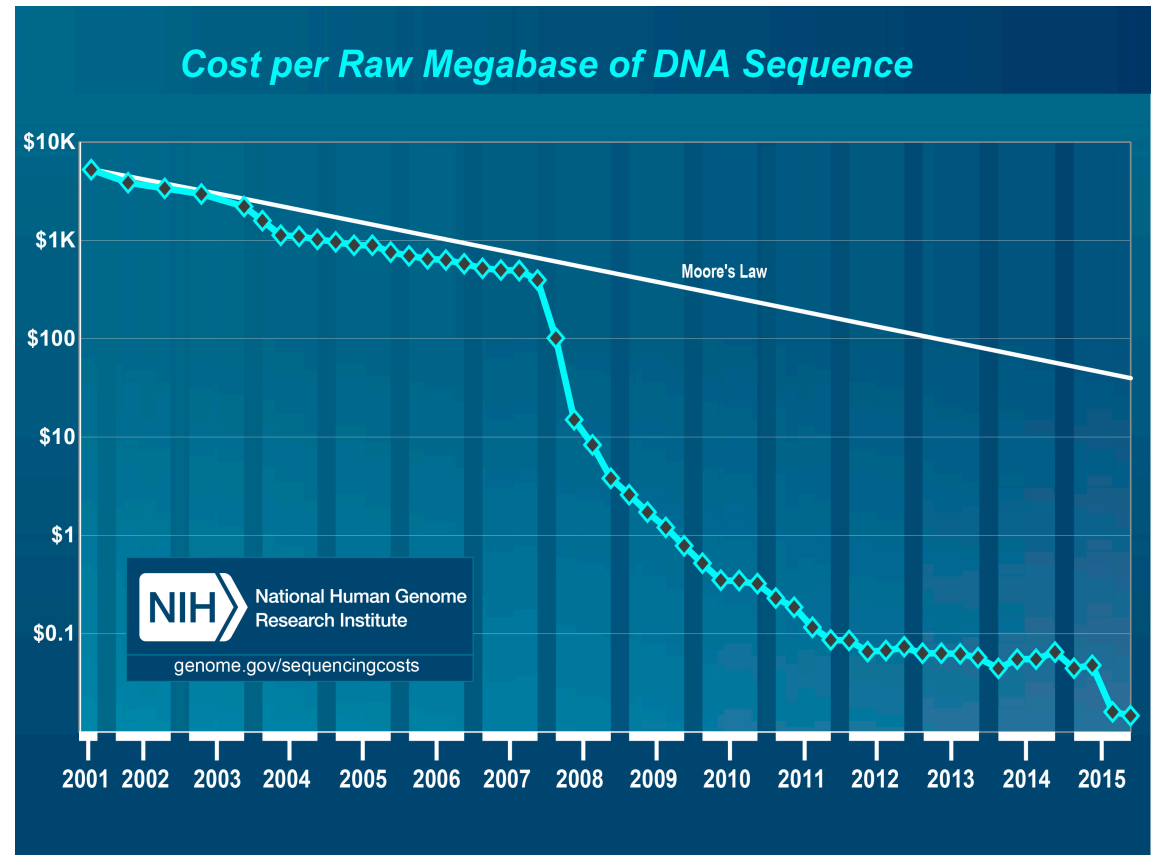
Kryder's Law and S-curves underlying exponential growth

- Moore's & Kryder's Laws
 - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies



Sequencing Data Explosion: Faster than Moore's Law for a Time (or a S-curve)

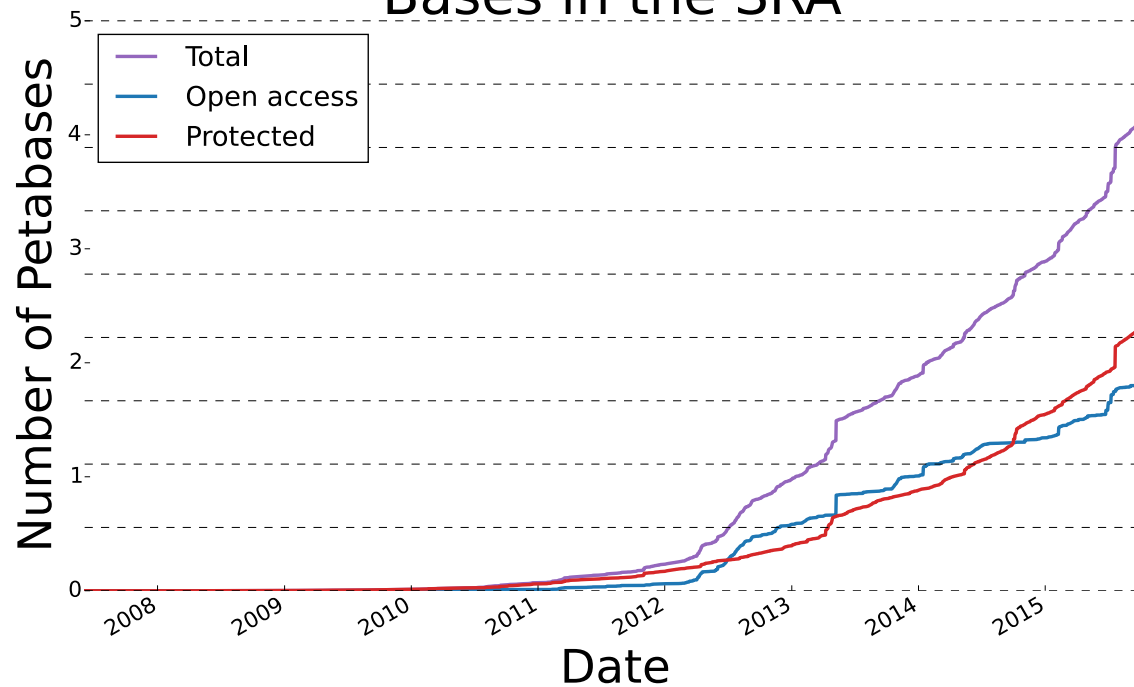
- DNA sequencing has gone through technological S-curves
 - In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
 - The advent of NGS was a shift to a new technology with dramatic decrease in cost).



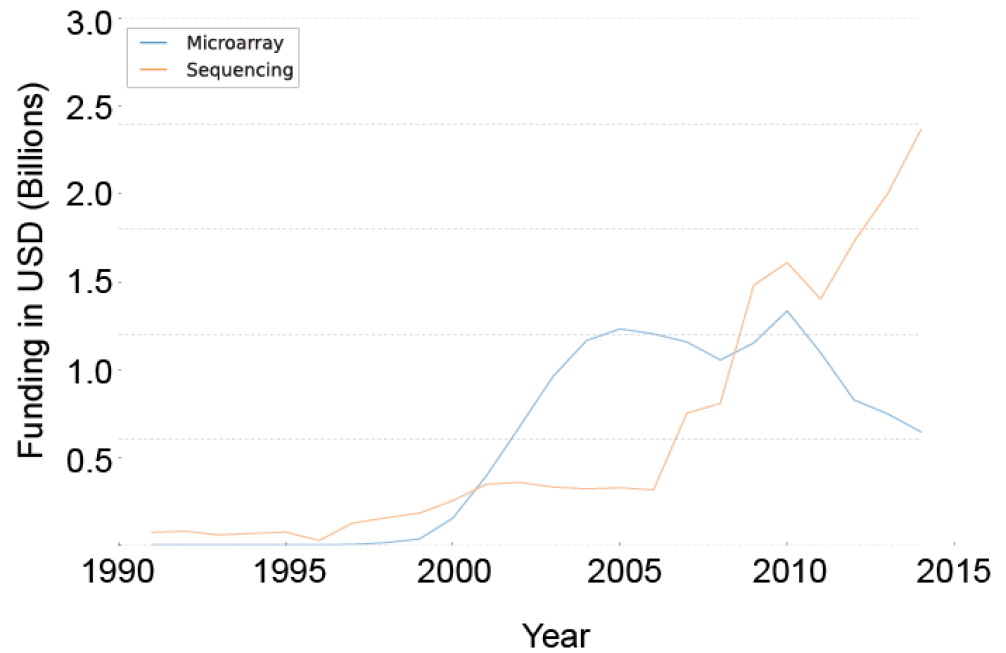
Sequencing cost reductions have resulted in an explosion of data

- The type of sequence data deposited has changed as well.
 - Protected data represents an increasing fraction of all submitted sequences.
 - Data from techniques utilizing NGS machines has replaced that generated via microarray.

Bases in the SRA



NIH Funding for “microarray” and “sequencing” projects



Sequence Universe

SRA ~1 petabyte

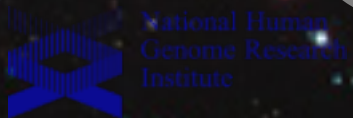
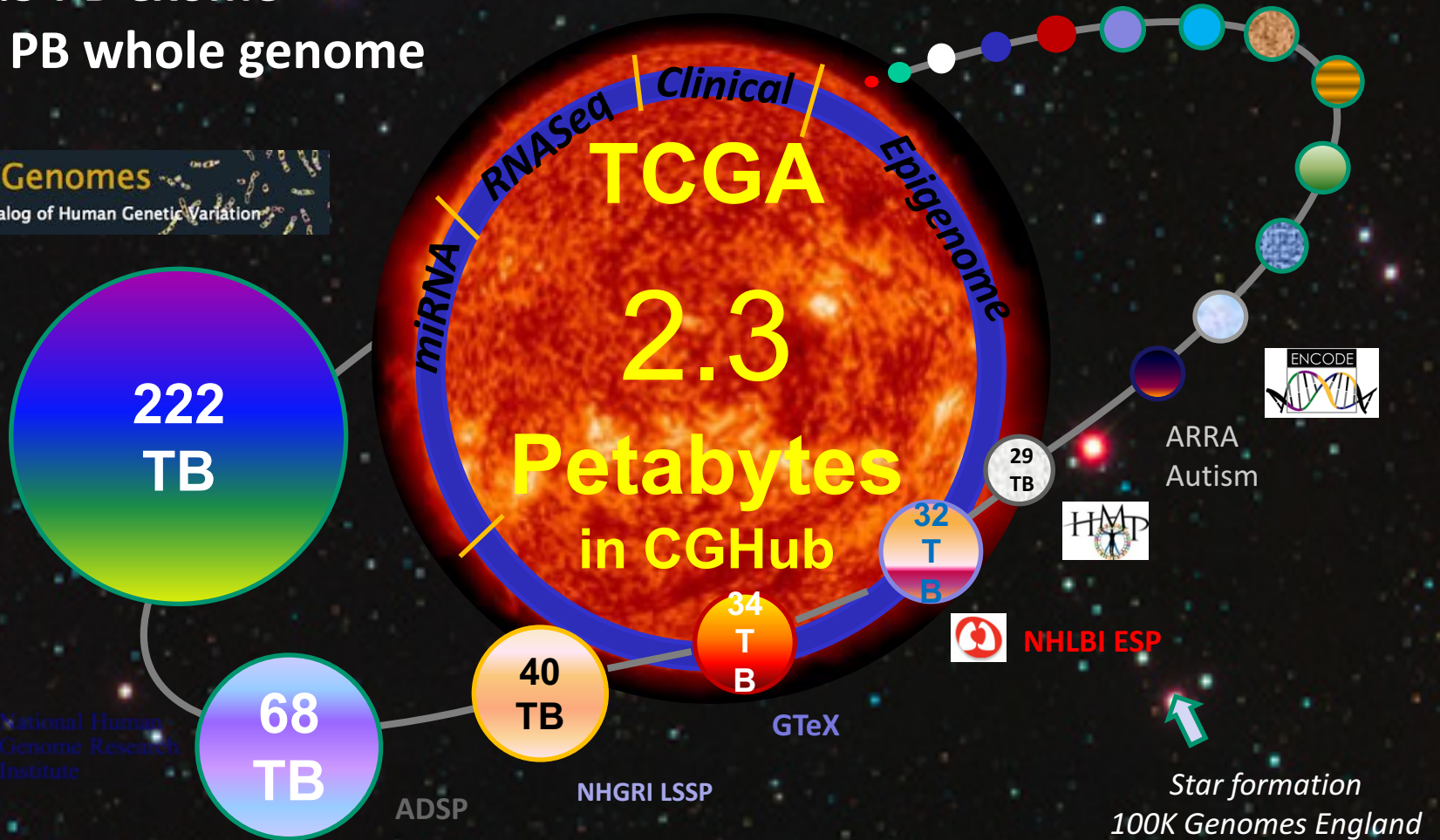
TCGA endpoint: ~2.5 Petabytes

~1.5 PB exome

~1 PB whole genome

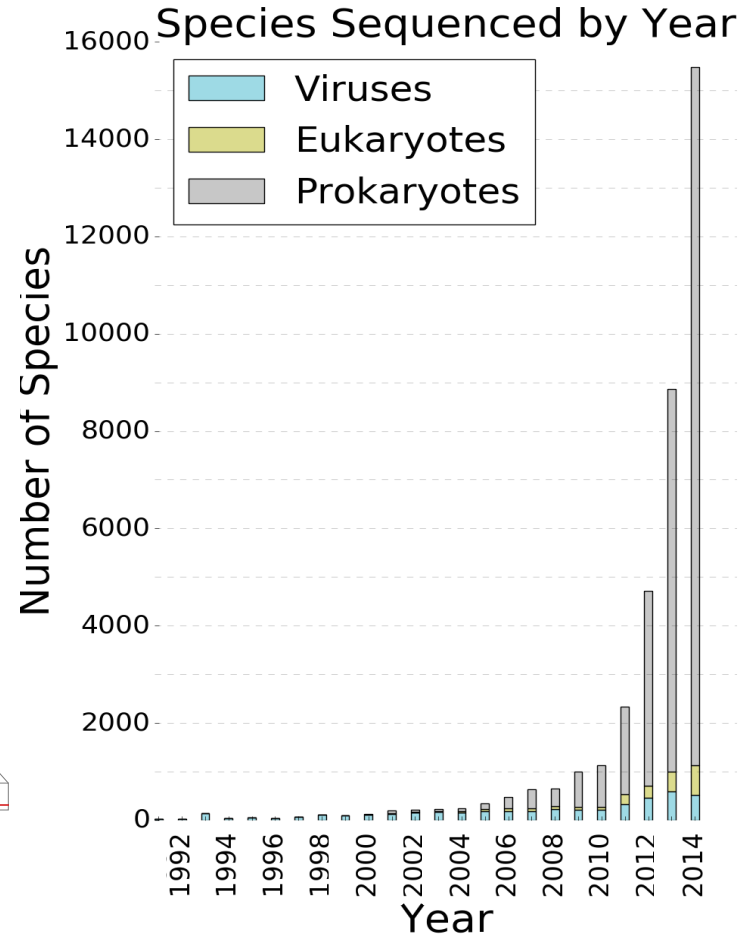
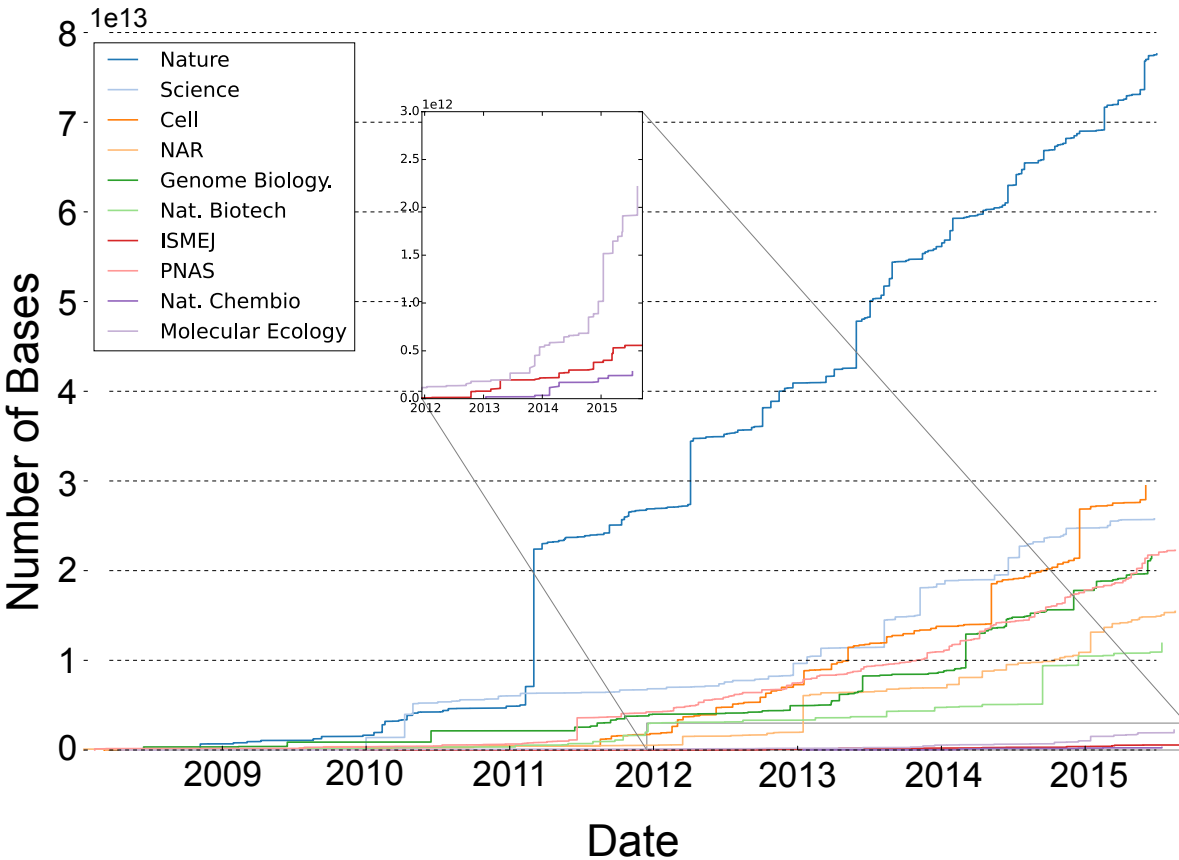
1000 Genomes

A Deep Catalog of Human Genetic Variation

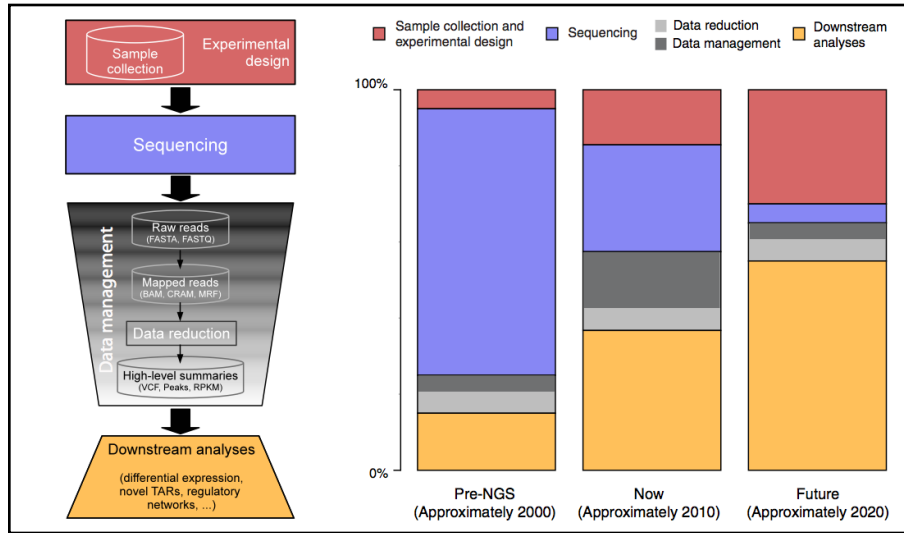


Increasing diversity in sequence data sources

[Muir et al. ('15) GenomeBiol.]

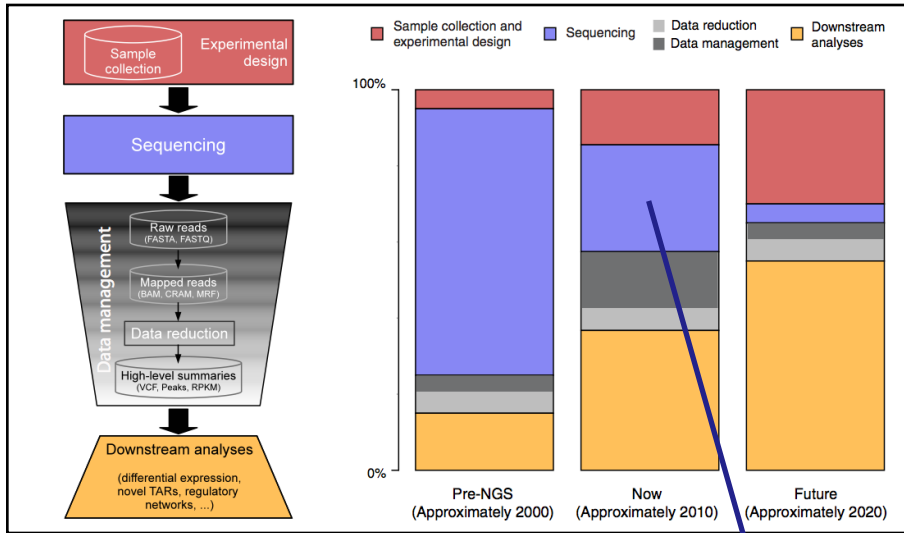


The changing costs of a sequencing pipeline

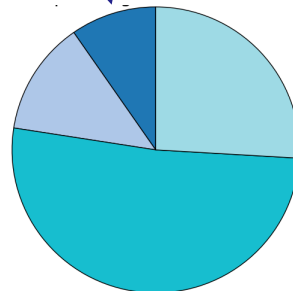
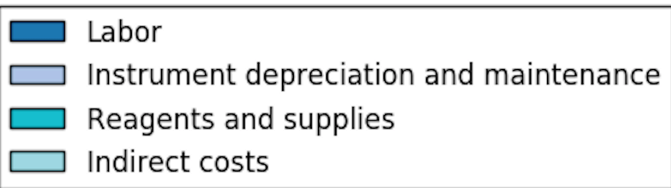


From '00 to ~' 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

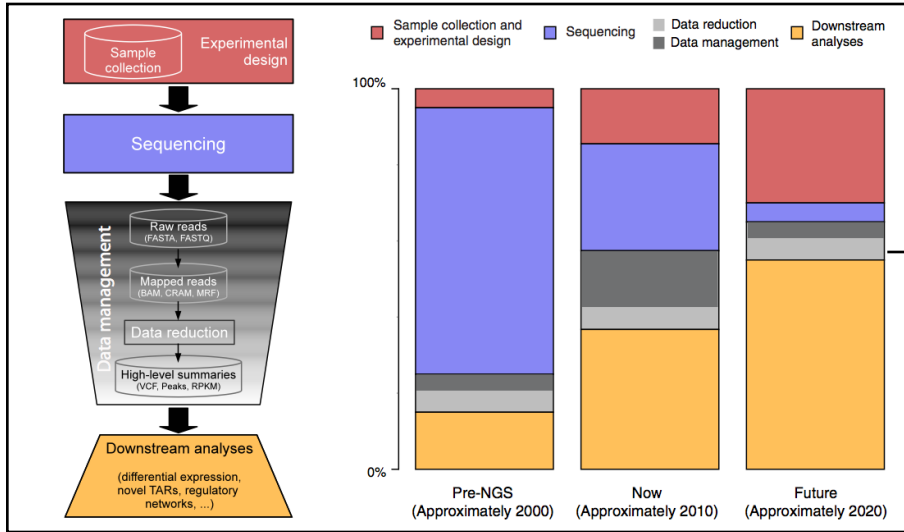
The changing costs of a sequencing pipeline



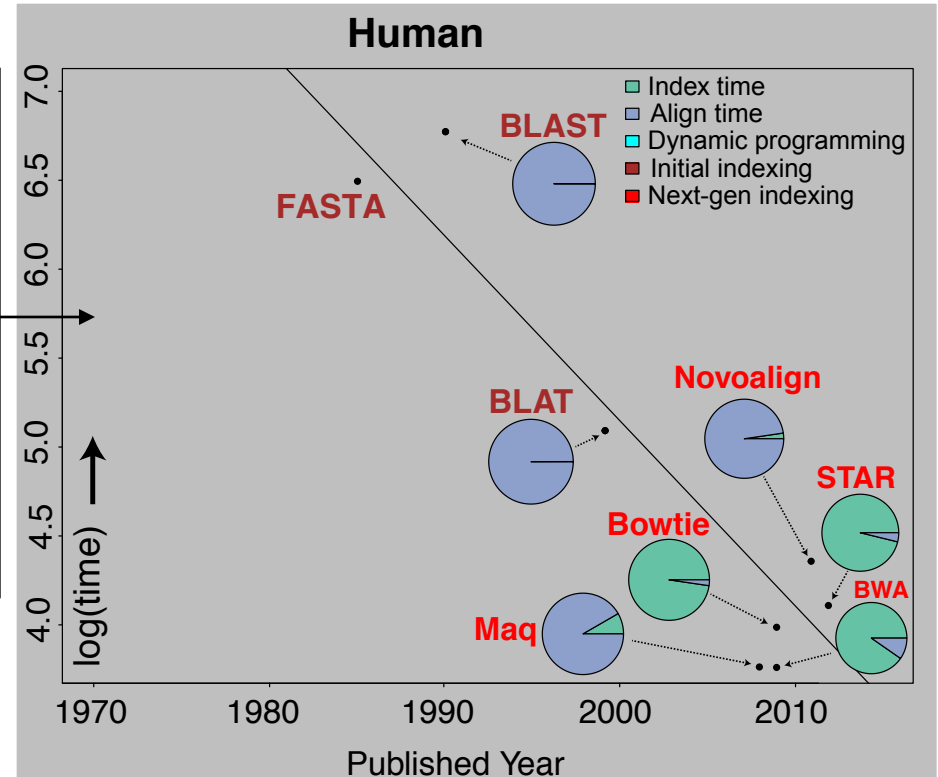
From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



The changing costs of a sequencing pipeline

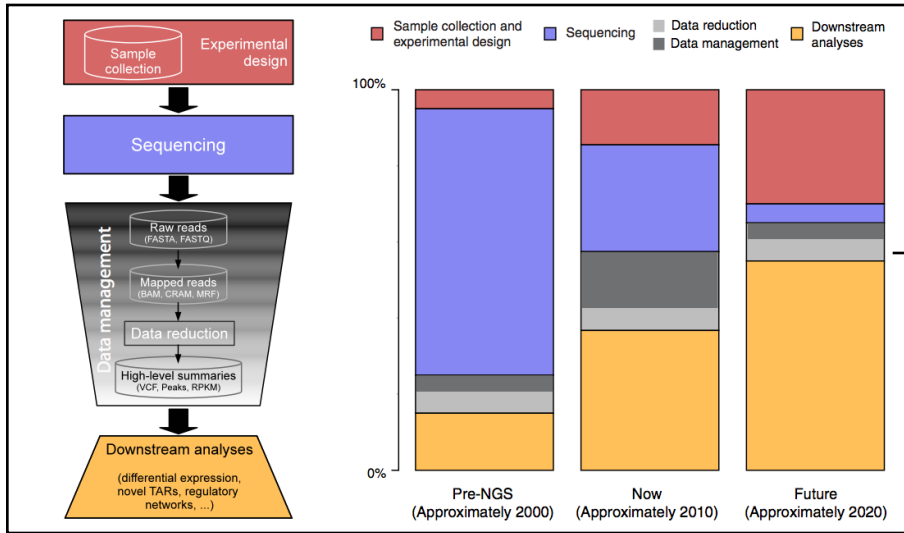


From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

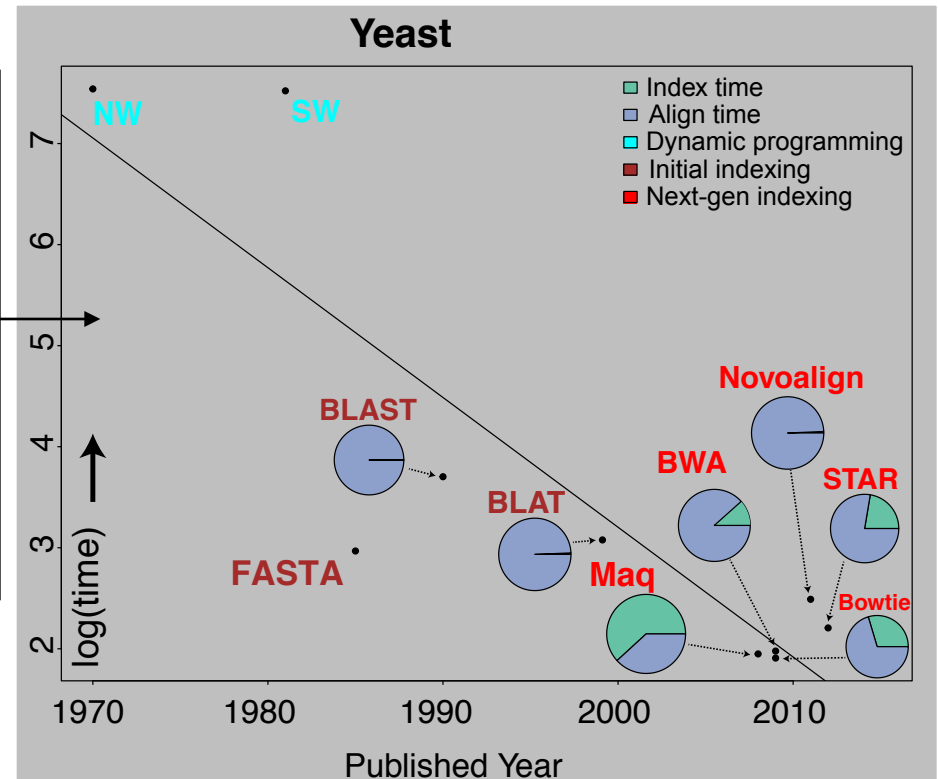


Alignment algorithms scaling to keep pace with data generation

The changing costs of a sequencing pipeline

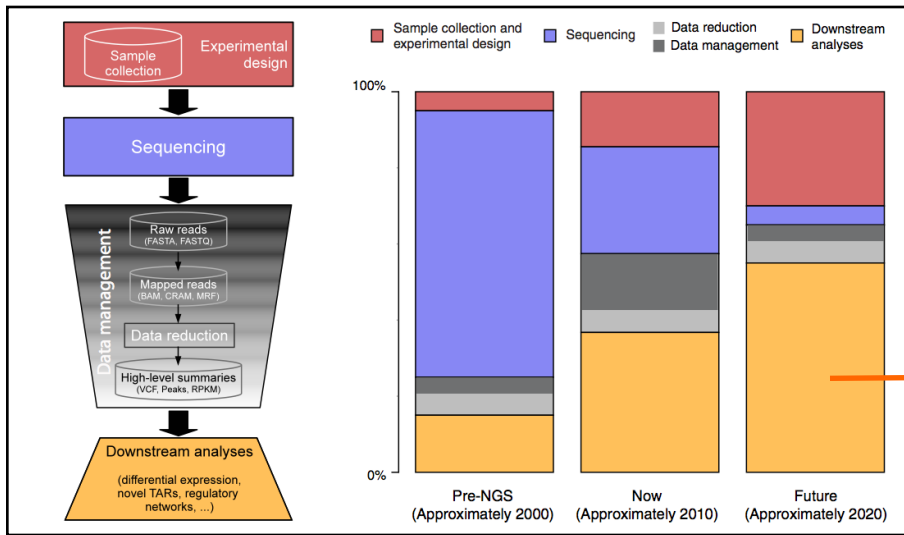


From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

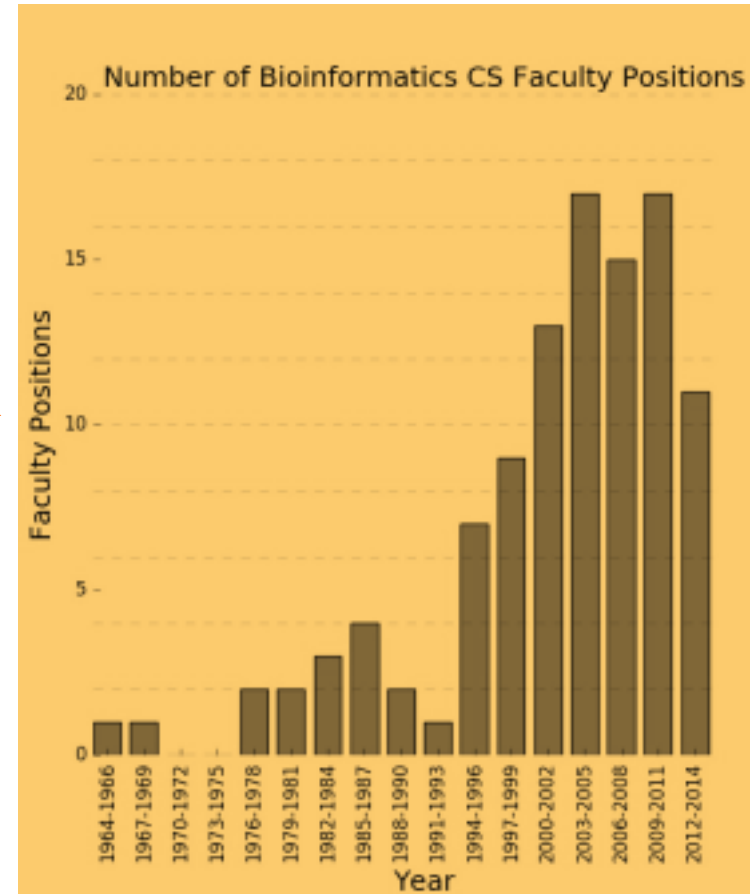


Alignment algorithms scaling to keep pace with data generation

The changing costs of a sequencing pipeline



From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with STRESS cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized Frustration to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with MUSIC
 - Prioritizing rare variants with “sensitive sites” (human-conserved)
- Putting it together in Workflows
 - Using LARVA to do burden testing on non-coding annotation
 - Need to correct for co-variables & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using FunSeq to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

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

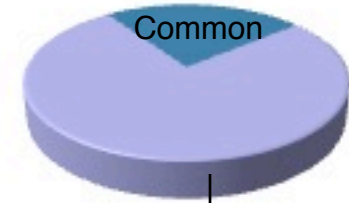
SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

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.

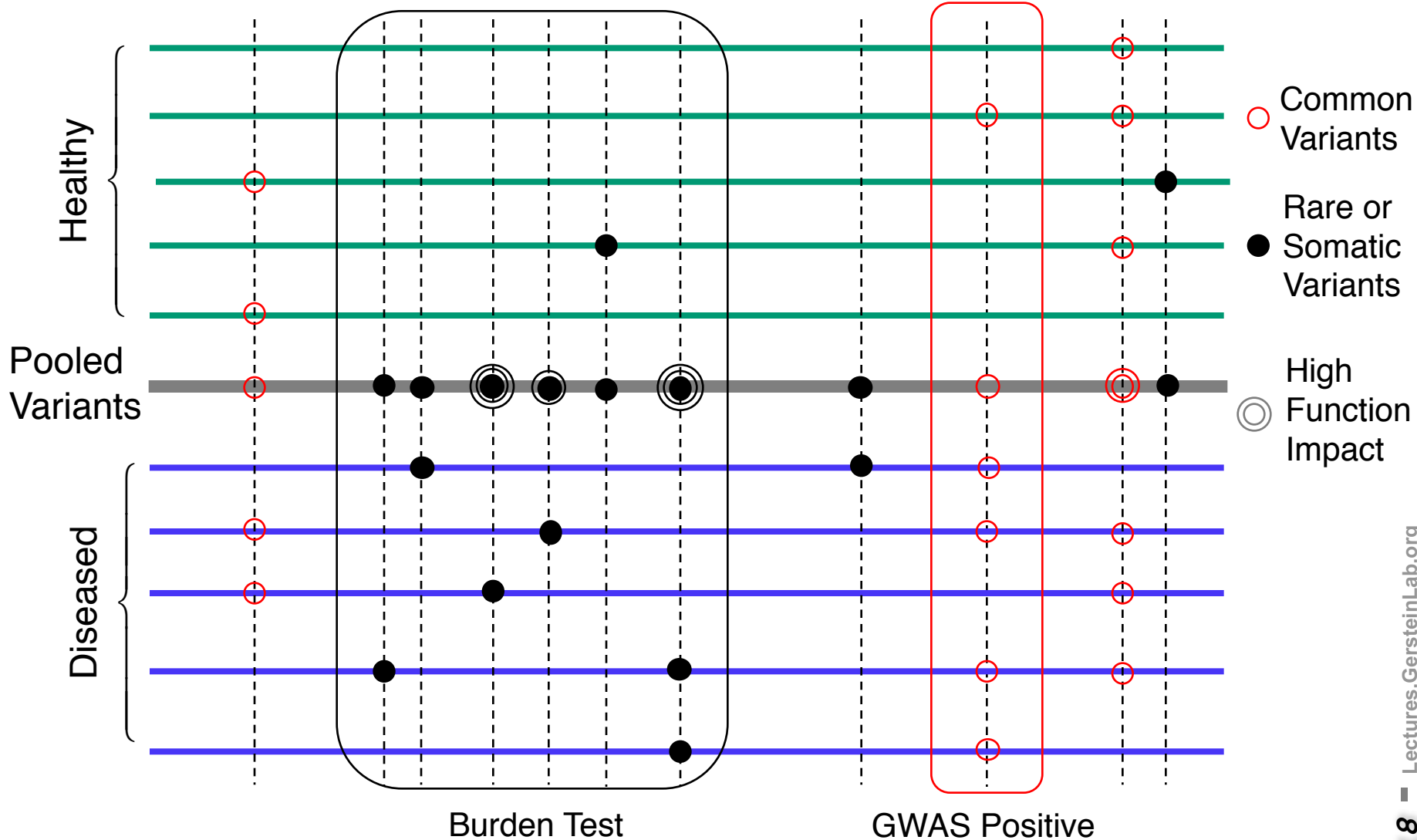
• Drivers

- 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



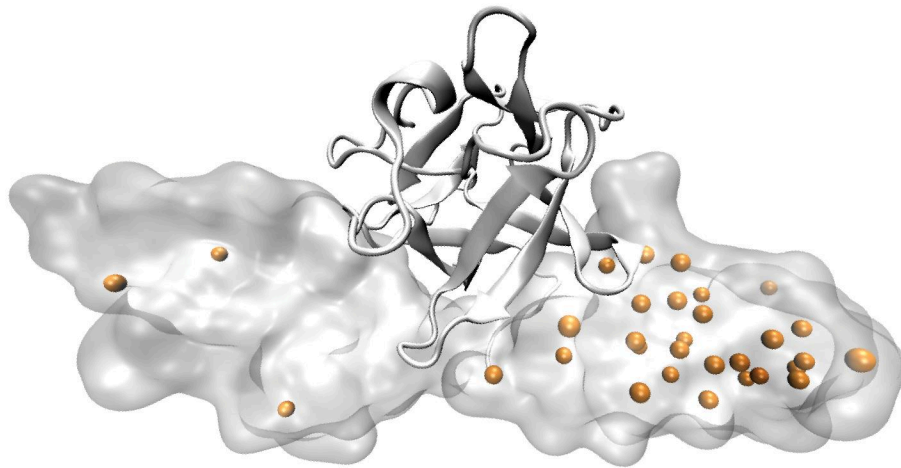
Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

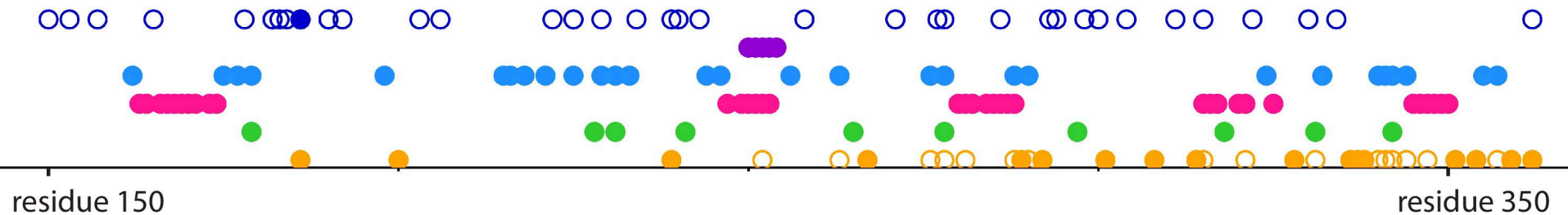
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



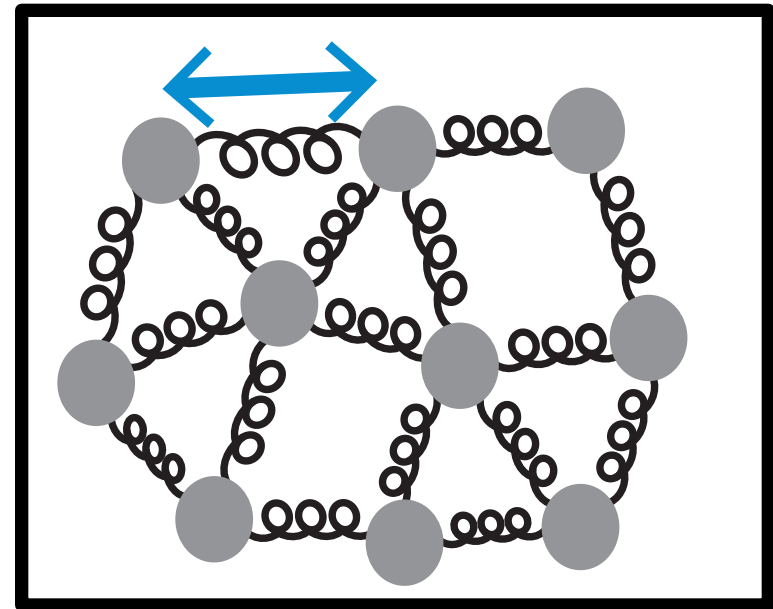
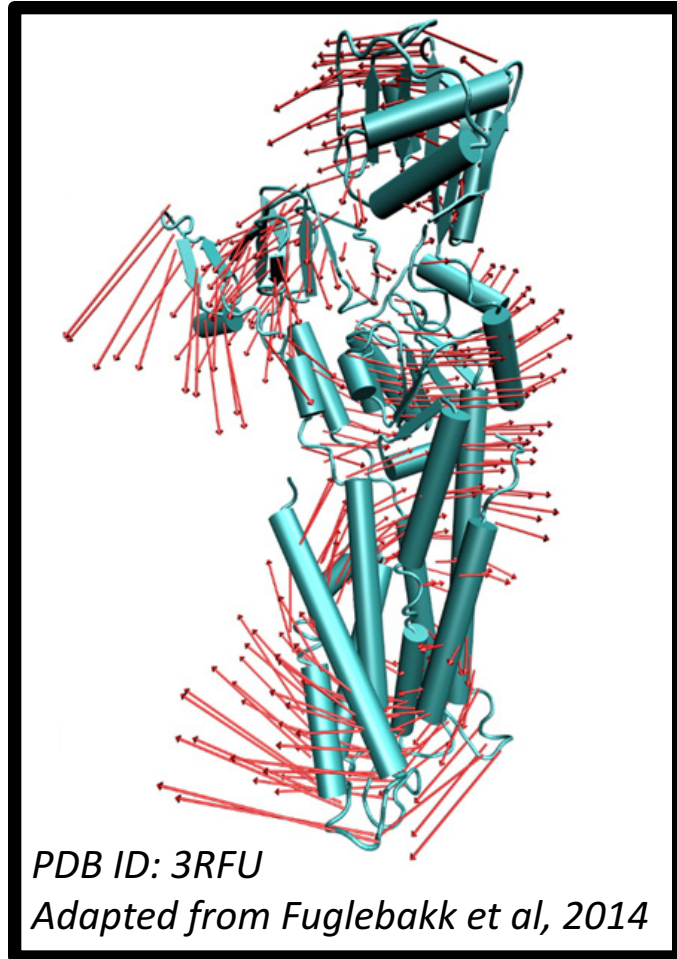
- ○ 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

Fibroblast growth factor receptor 2 (pdb: 1IIL)



Models of Protein Conformational Change

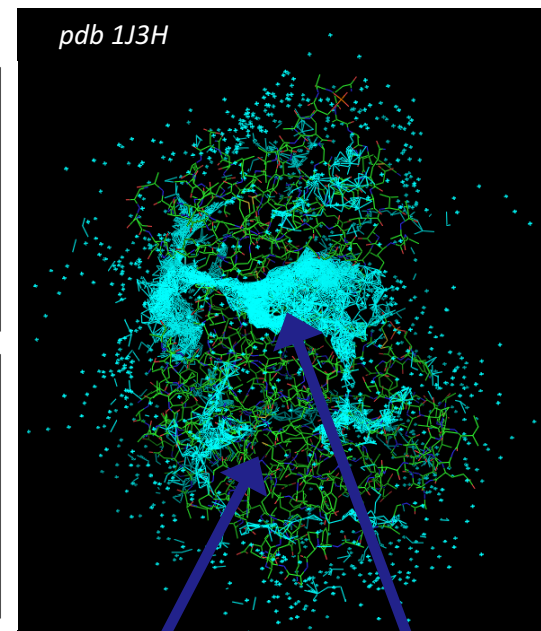
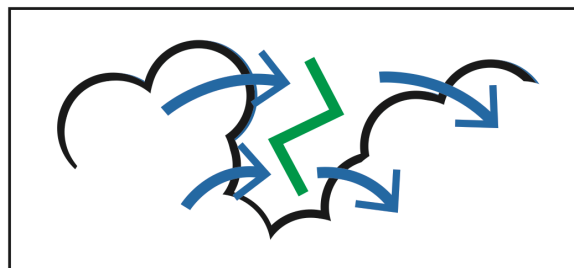
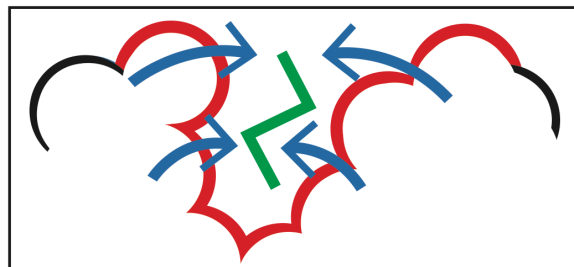
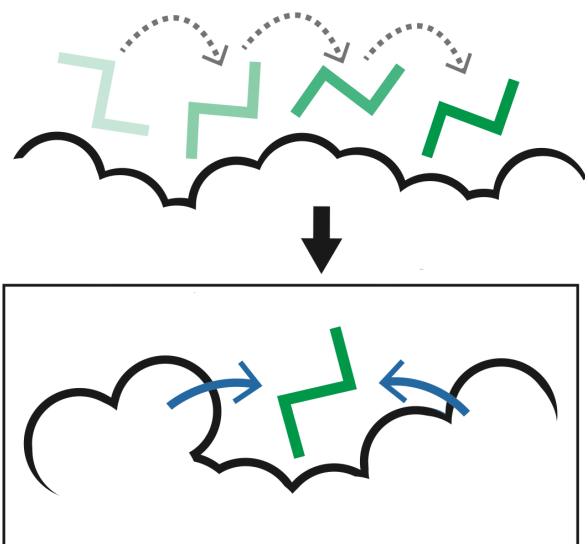
Motion Vectors from Normal Modes (ANMs)



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



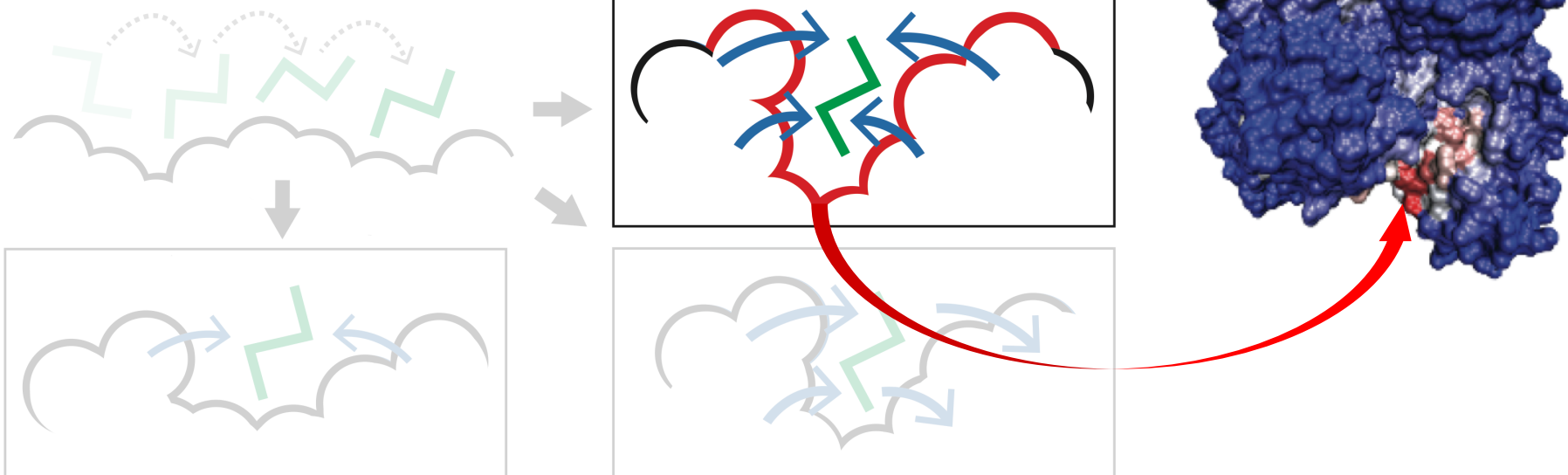
Surface region with high density of candidate sites

Surface region with low density of candidate sites

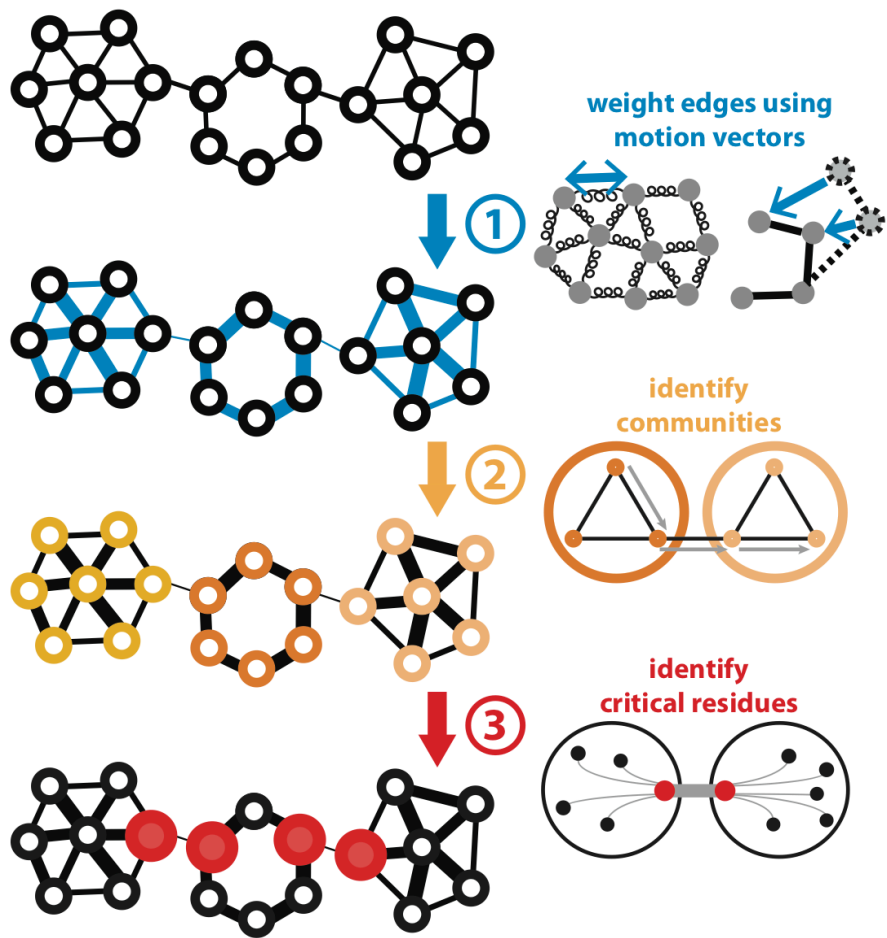
$$\text{binding leverage} = \sum_{m=1}^{10} \left(\sum_i \sum_j \Delta d_{ij}^2 \right)$$

Predicting Allosterically-Important Residues at the Surface

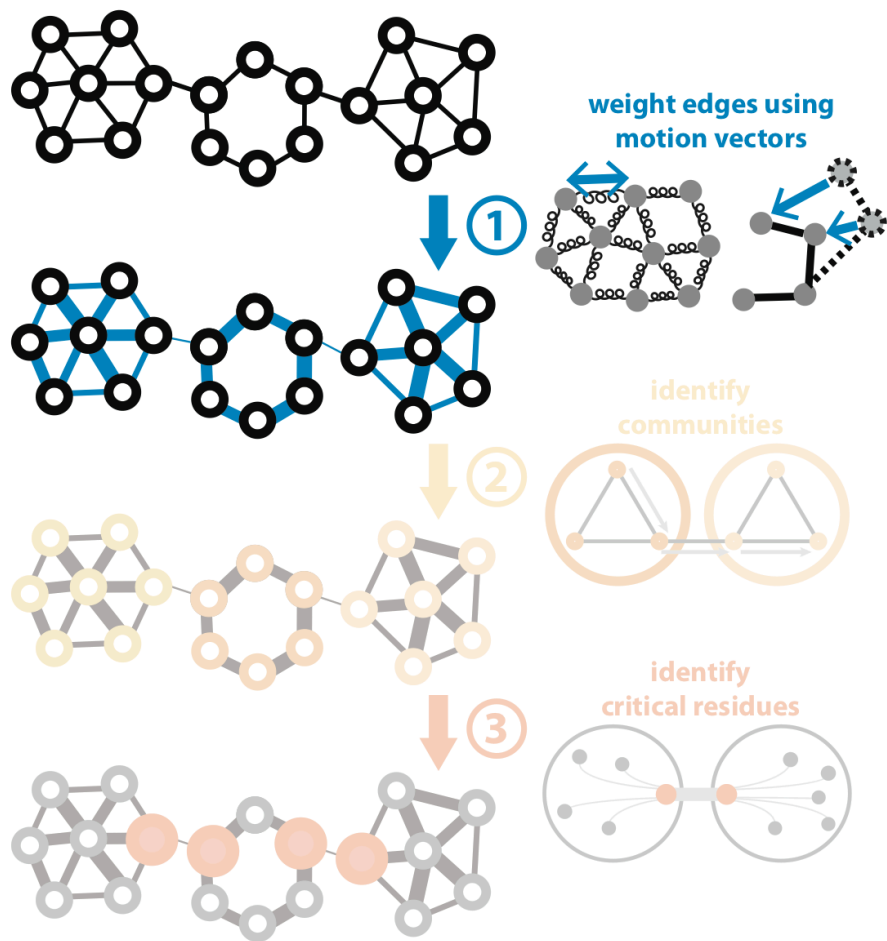
PDB: 3PFK



Predicting Allosterically-Important Residues within the Interior



Predicting Allosterically-Important Residues within the Interior

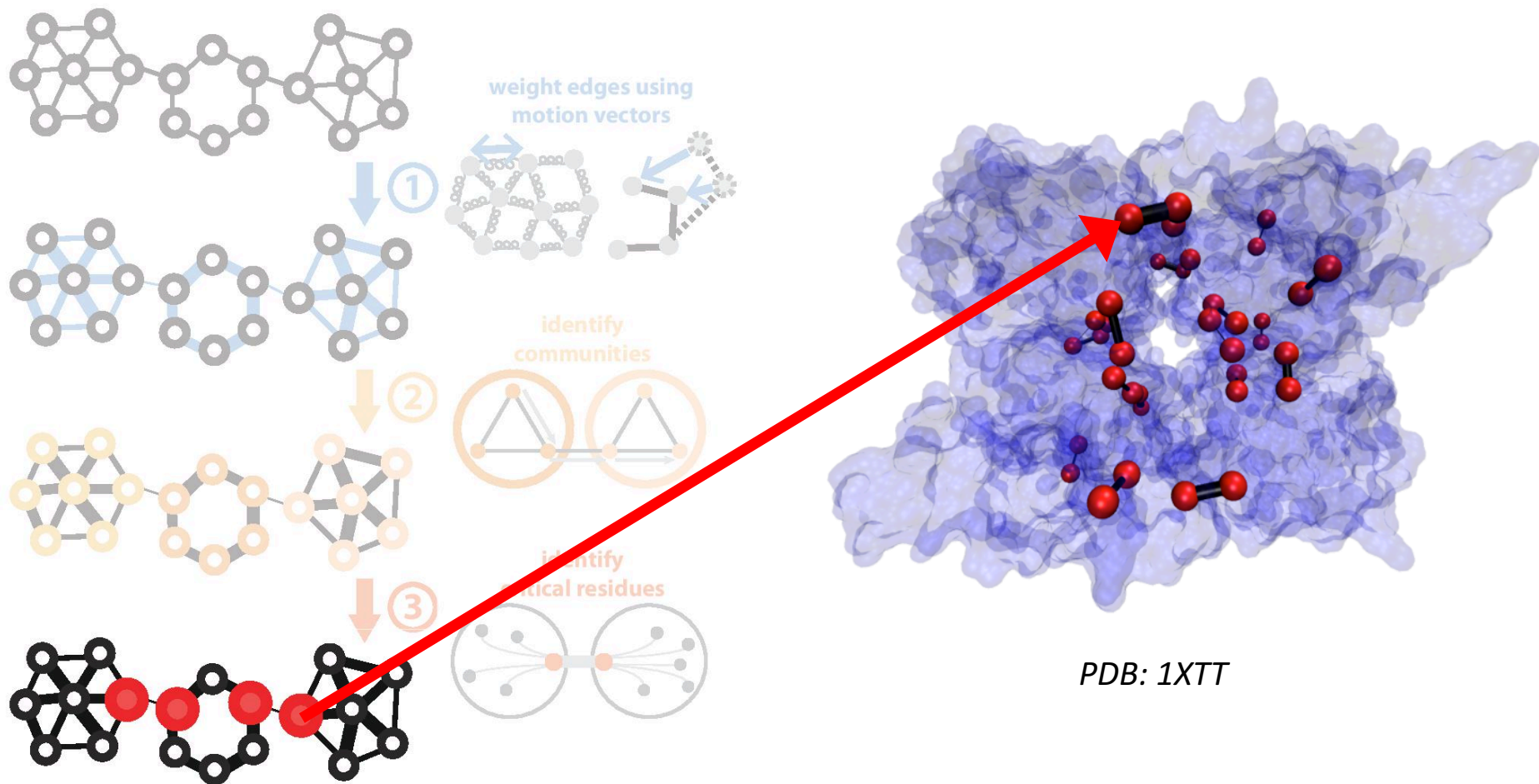


$$Cov_{ij} = \langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle$$

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

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

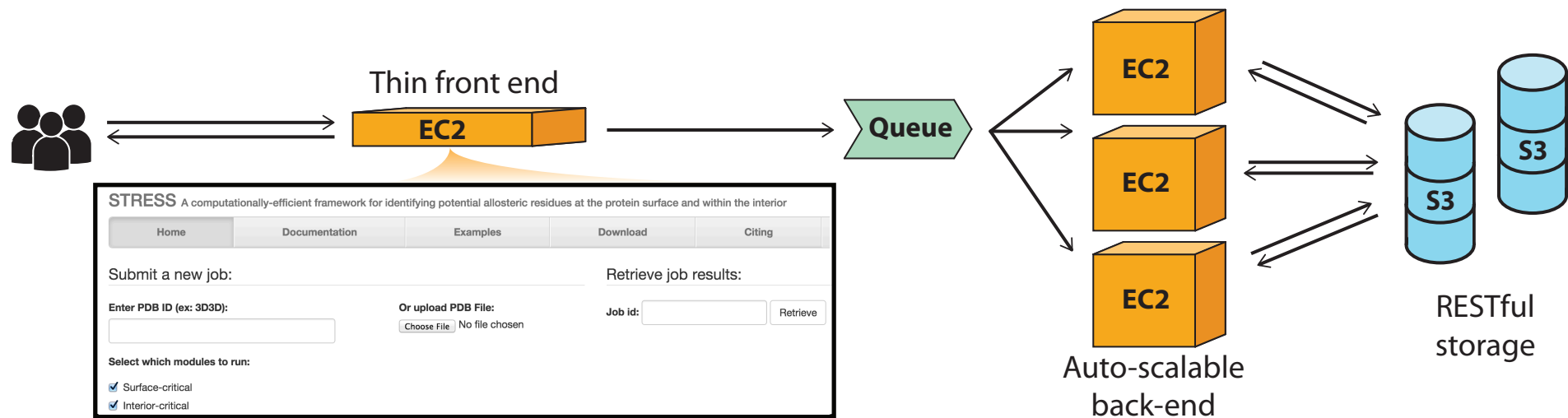
Predicting Allosterically-Important Residues within the Interior



PDB: 1XTT

STRESS Server Architecture: Highlights

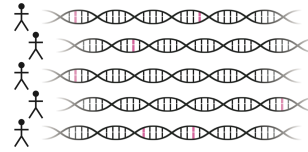
stress.molmovdb.org



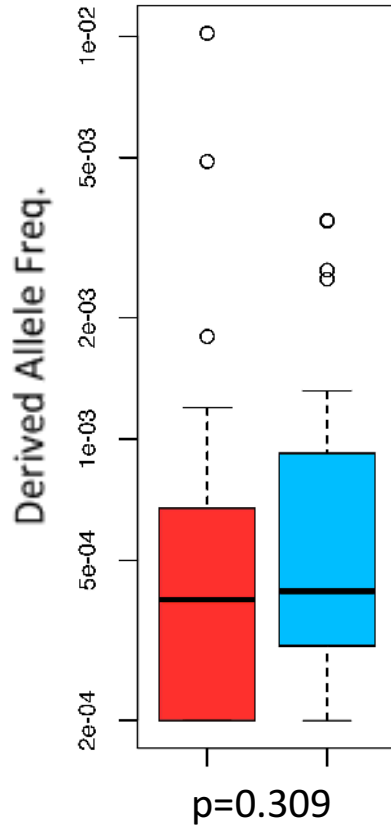
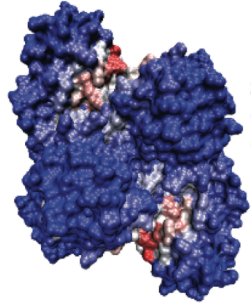
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

Intra-species conservation of predicted allosteric residues

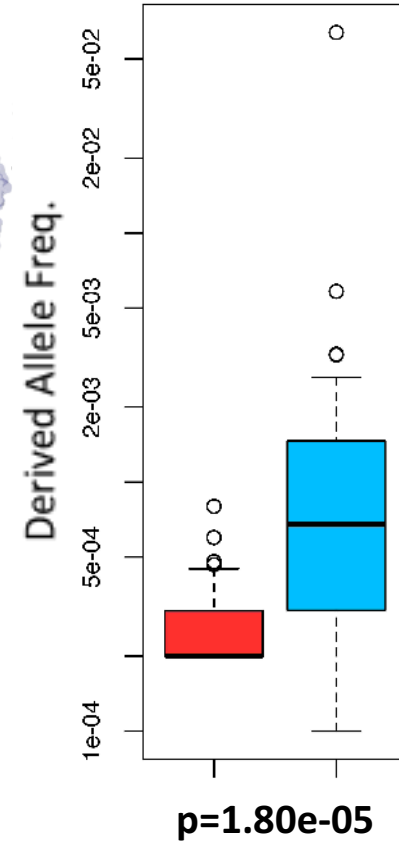
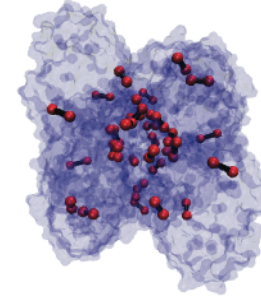
1000 Genomes



Surface



Interior



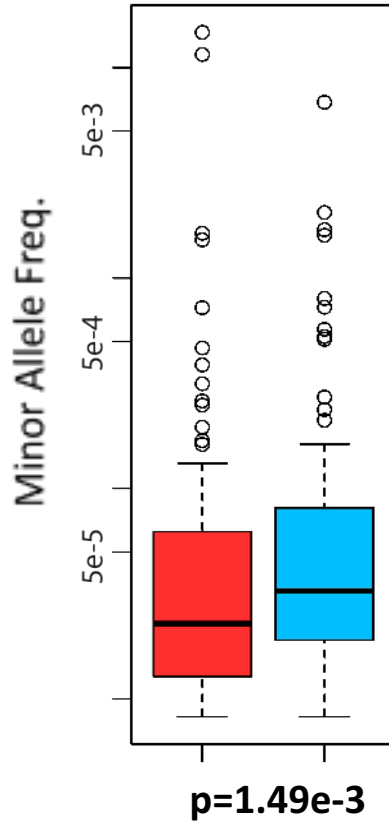
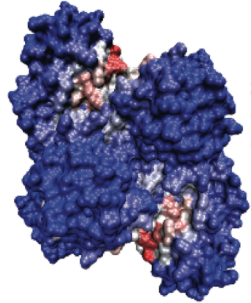
critical
non-critical

Intra-species conservation of predicted allosteric residues

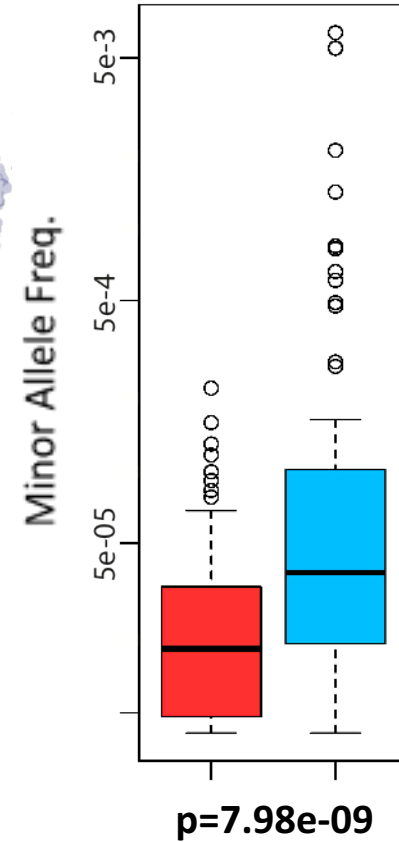
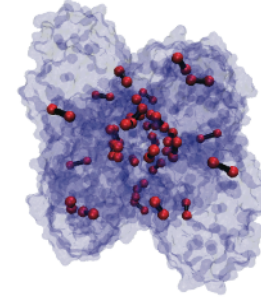
ExAC



Surface



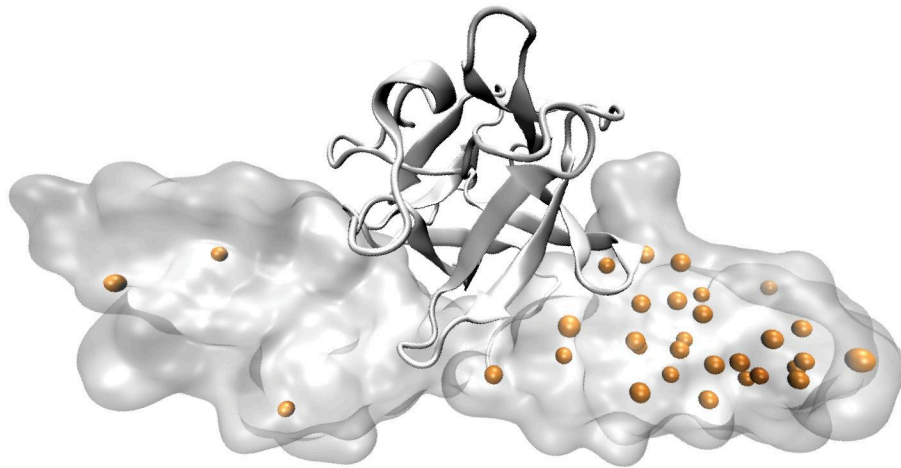
Interior



critical
non-critical

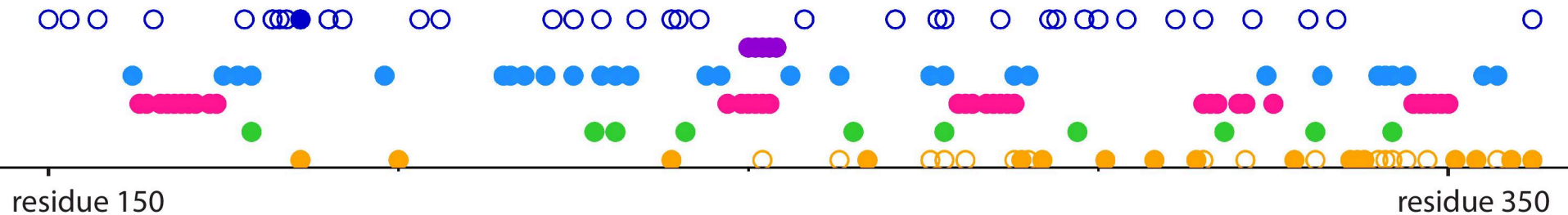
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



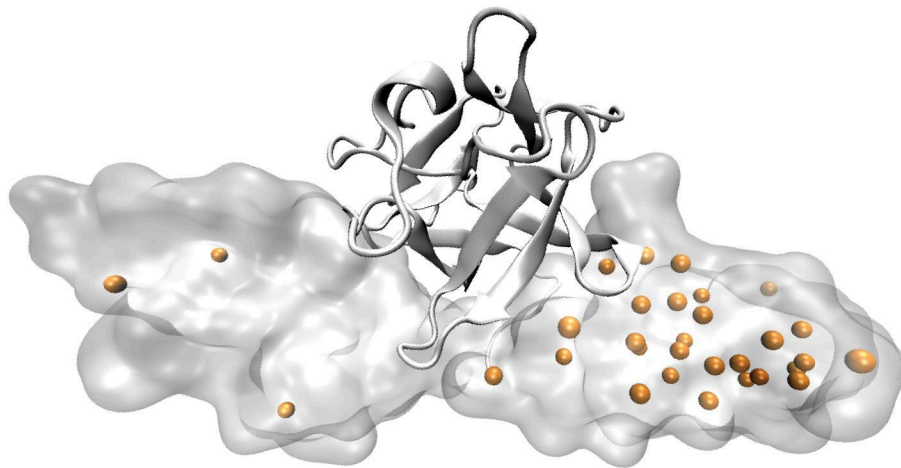
- ○ 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

Fibroblast growth factor receptor 2 (pdb: 1IIL)



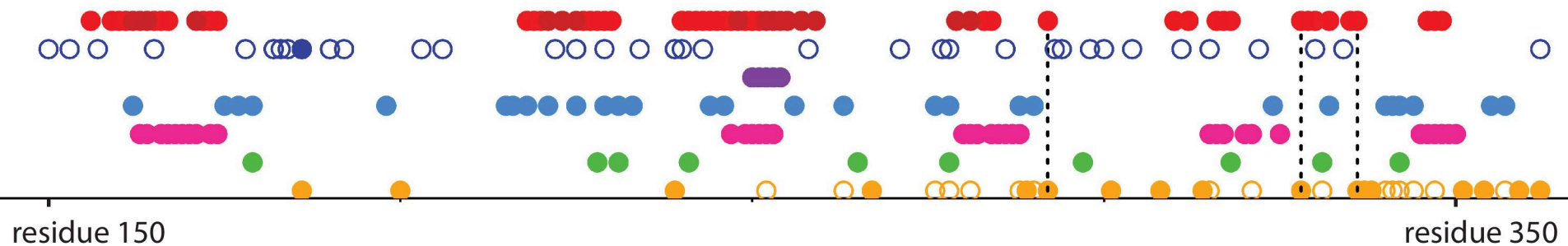
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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)

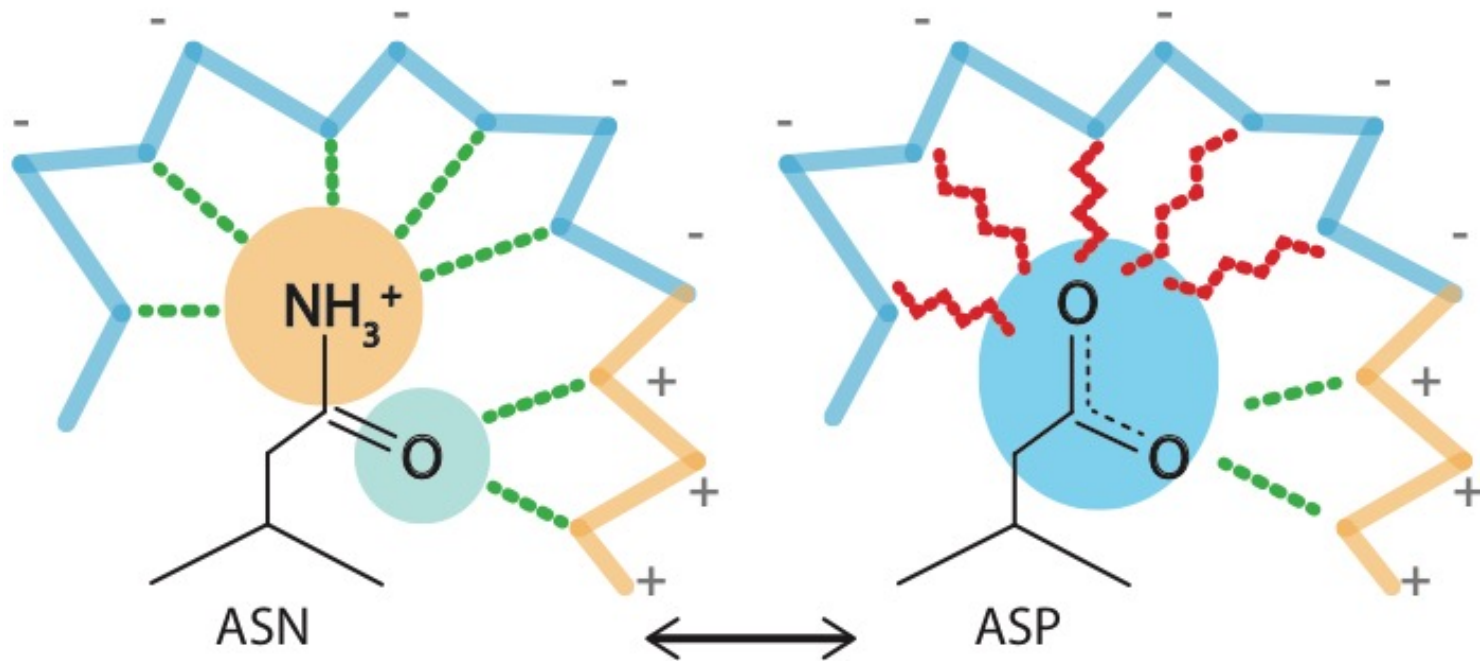


Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

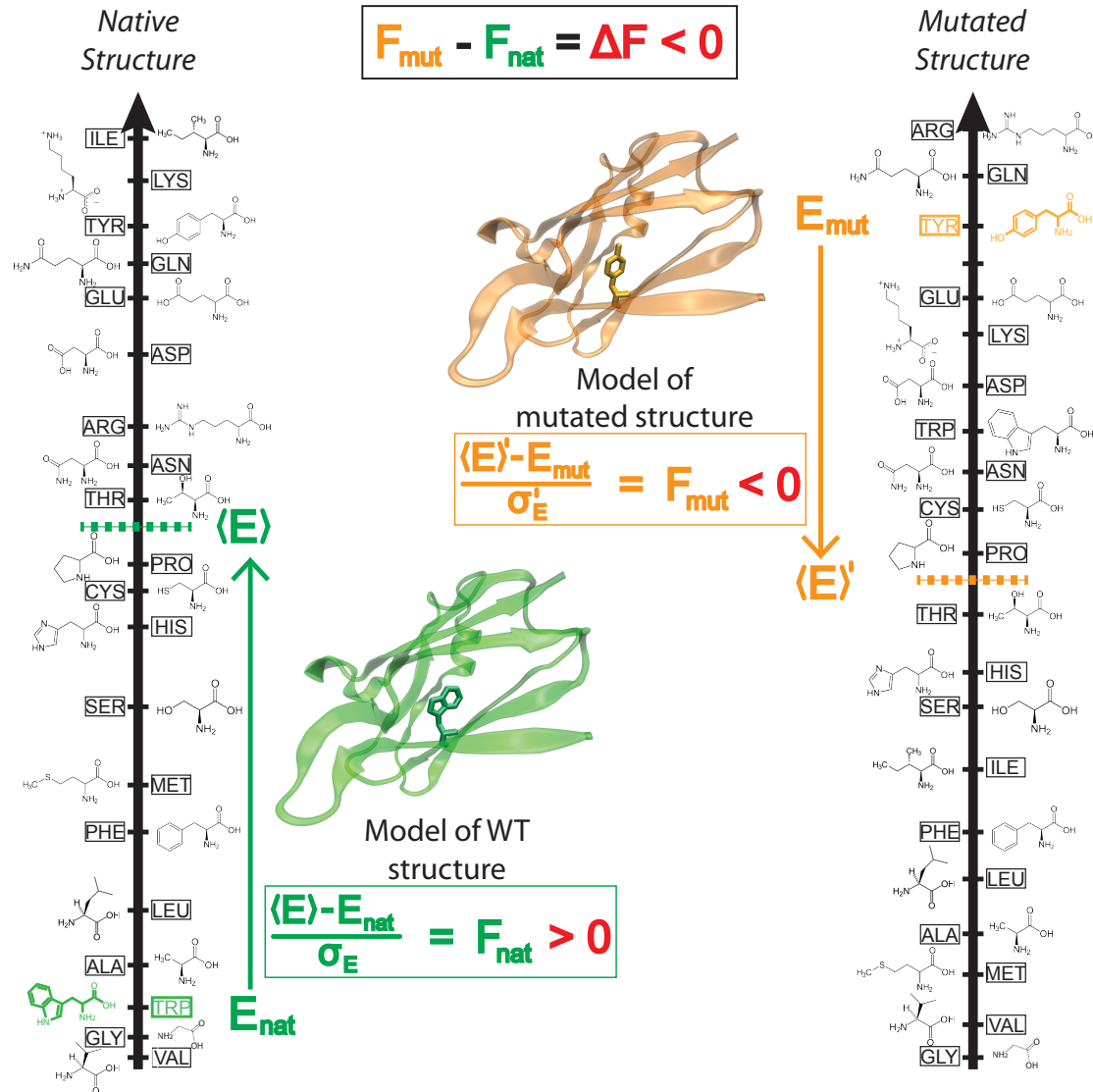
- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

Schematic illustration of localized frustration

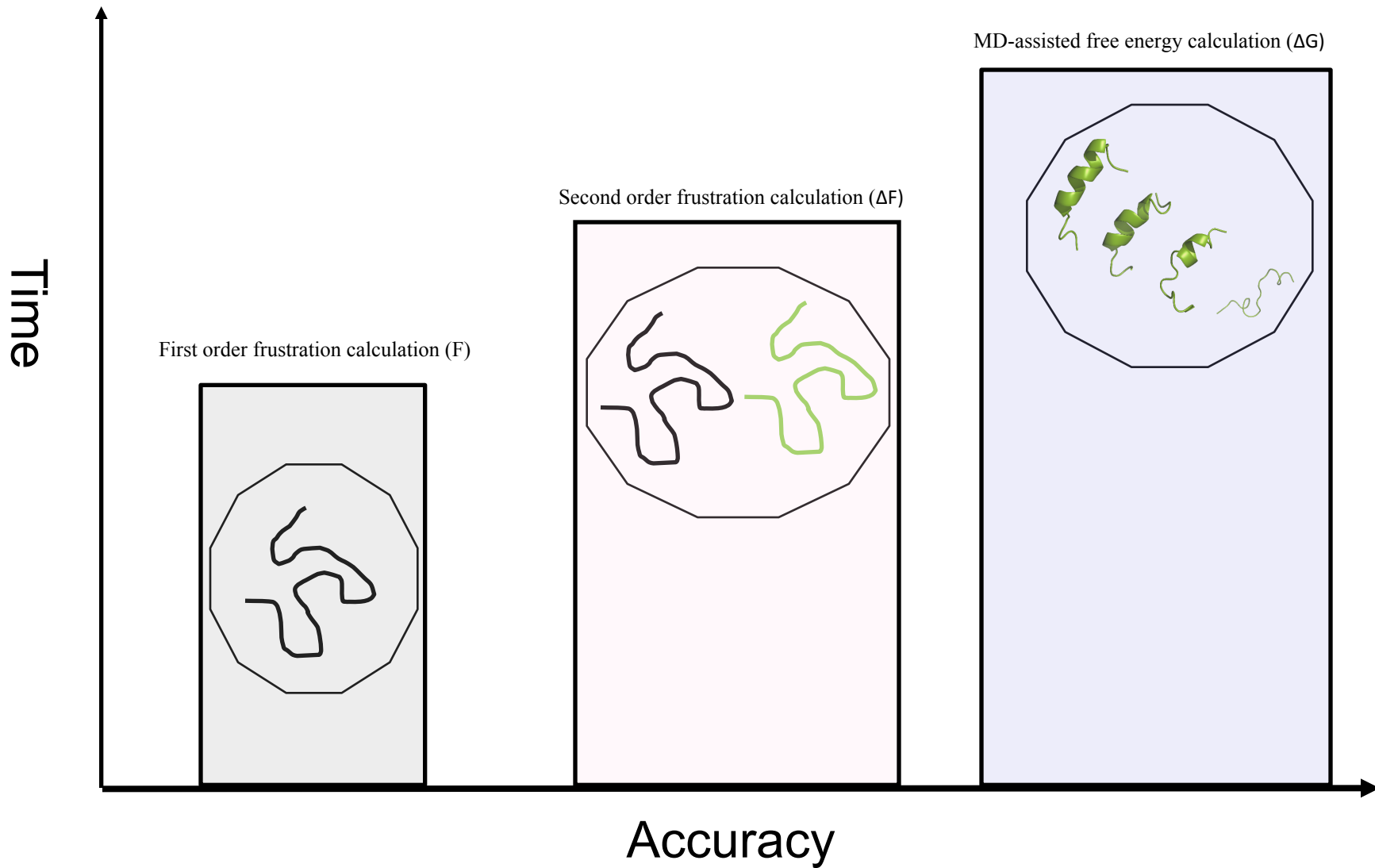


[Ferreiro et al., *PNAS* ('07)]

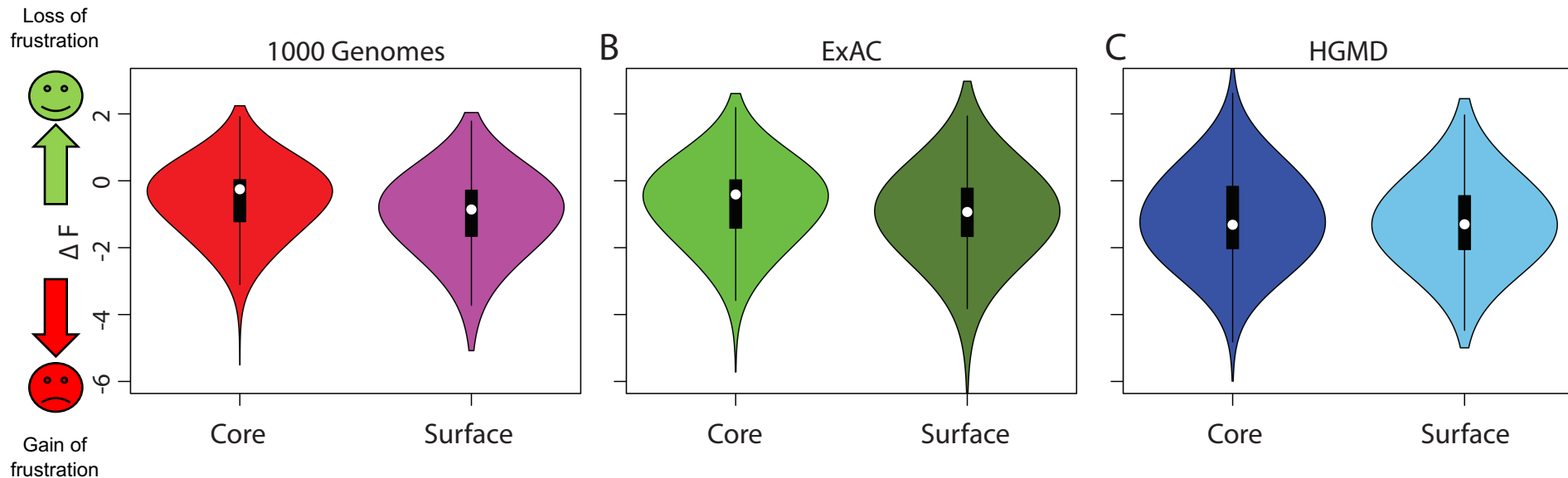
Workflow for evaluating localized frustration changes (ΔF)



Striking a balance: the complexity of the second order frustration calculation



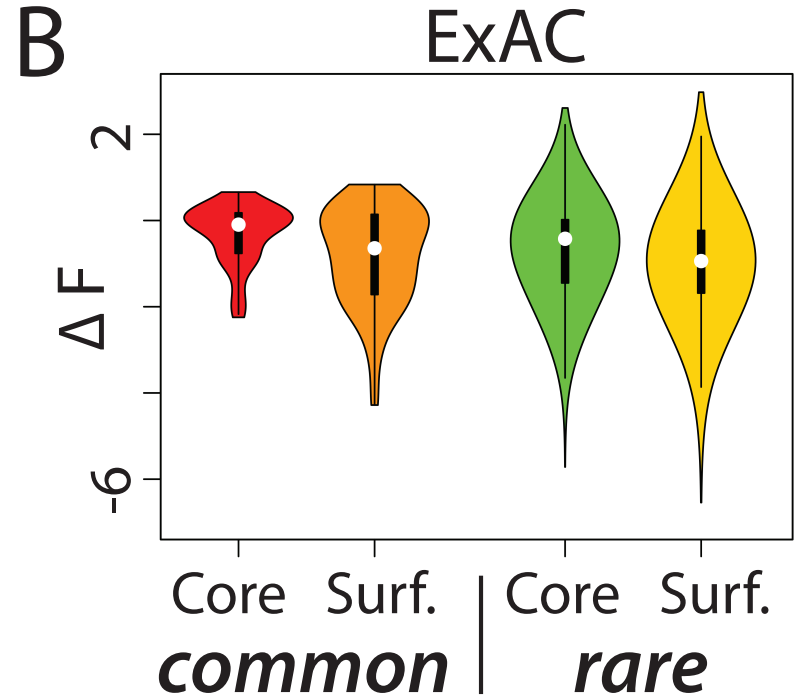
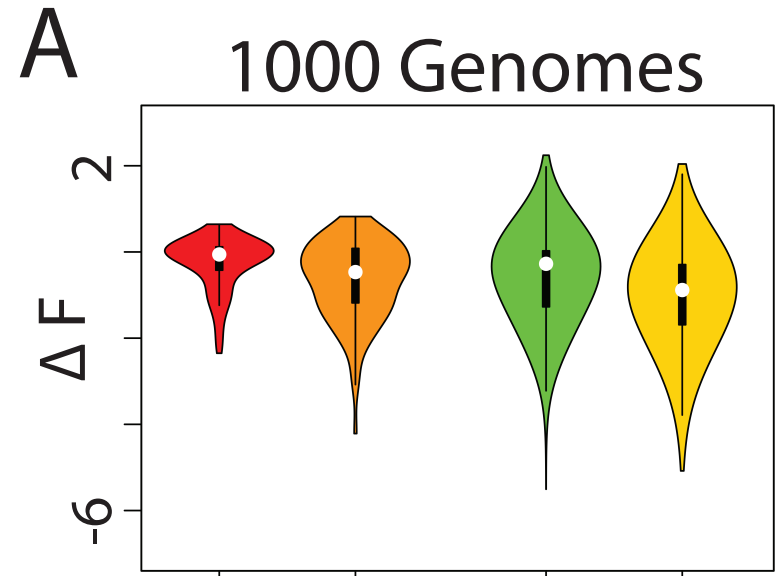
Comparing ΔF values across different SNV categories: Normal v disease



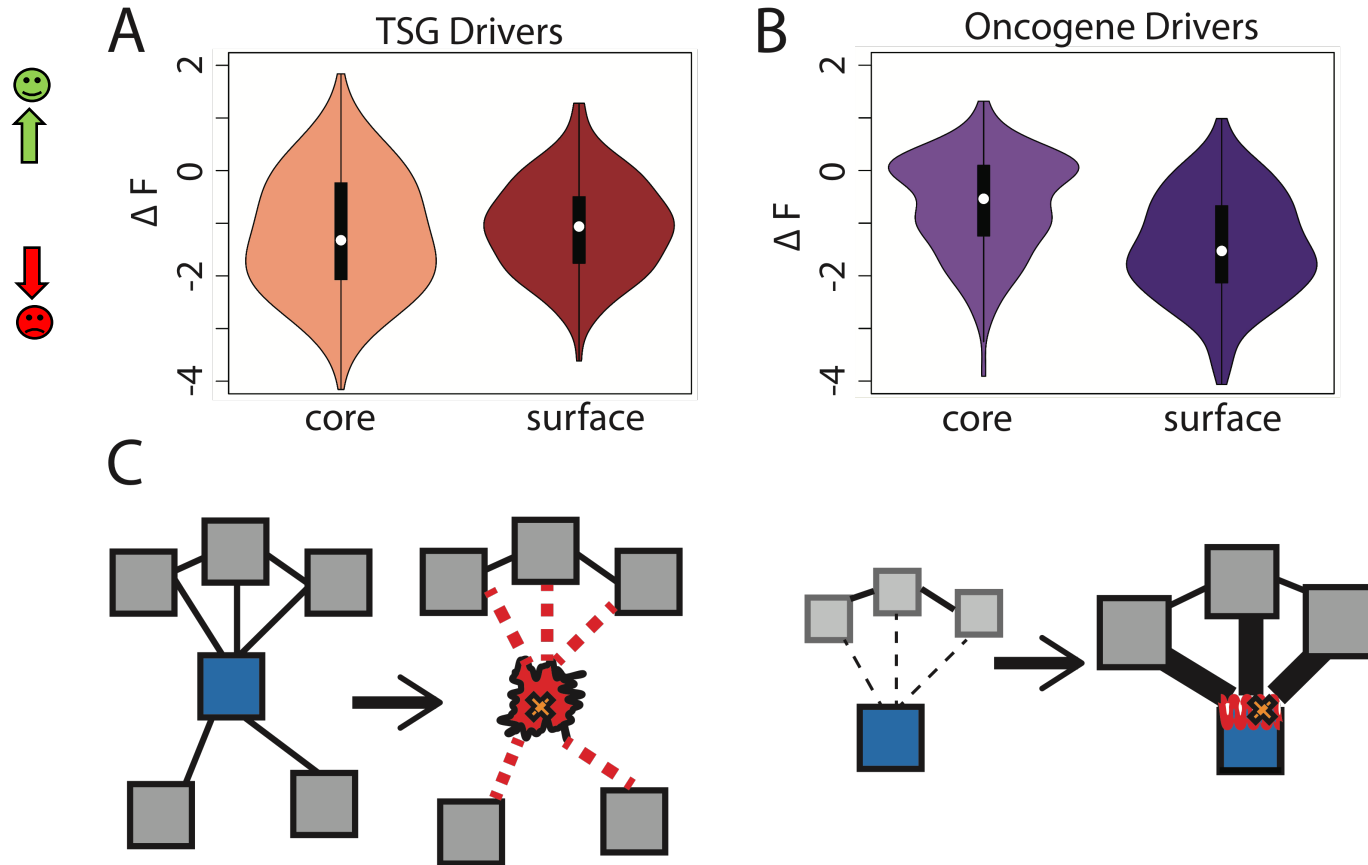
Normal mutations (1000G) tend to unfavorably frustrate (less frustrated) surface more than core, but for disease mutations (HGMD) no trend & greater changes

ΔF distributions among rare v. common SNVs

Rare mutations cause more unfavorable frustration change than common ones



Comparison between ΔF distributions: TSGs v. oncogenes



SNVs in TSGs change frustration more in core than the surface, whereas those associated with oncogenes manifest the opposite pattern. This is consistent with differences in LOF v GOF mechanisms.

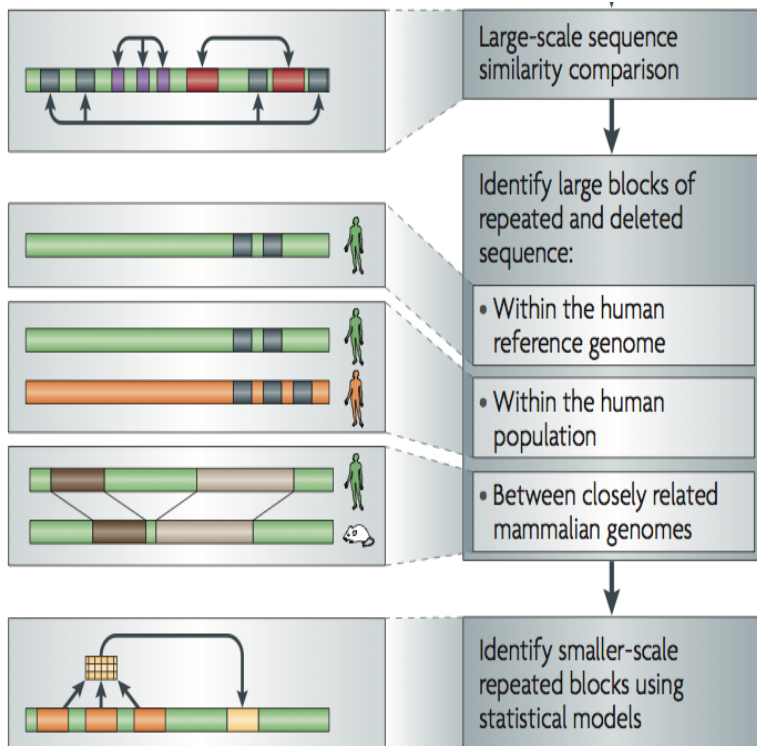
Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

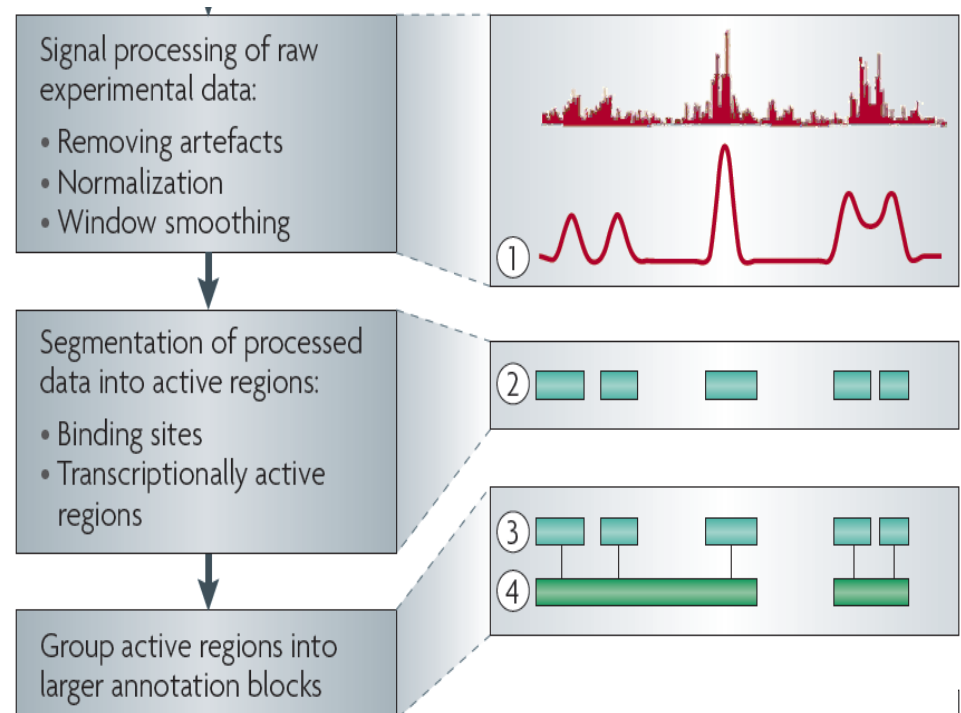
Non-coding Annotations: Overview

Sequence features, incl. Conservation



Functional Genomics

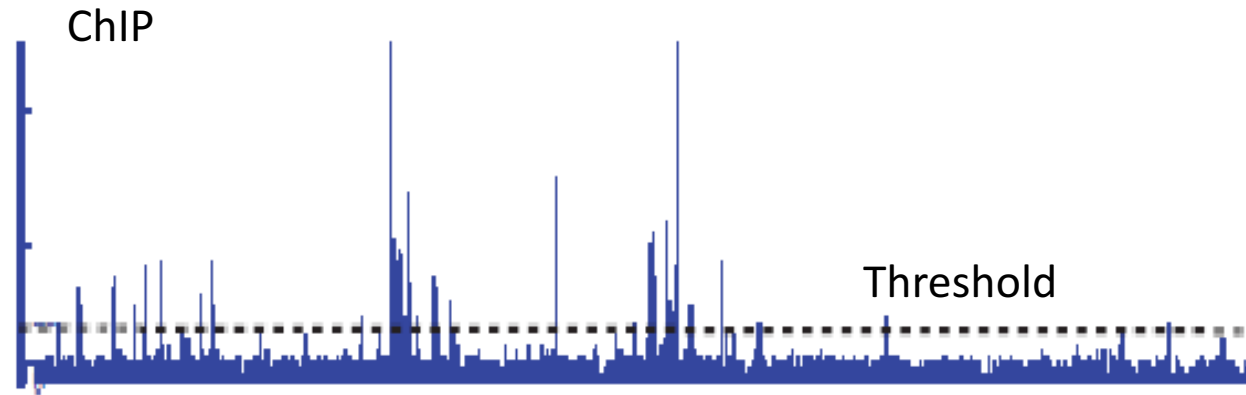
Chip-seq (Epigenome & seq. specific TF)
and ncRNA & un-annotated transcription



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

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

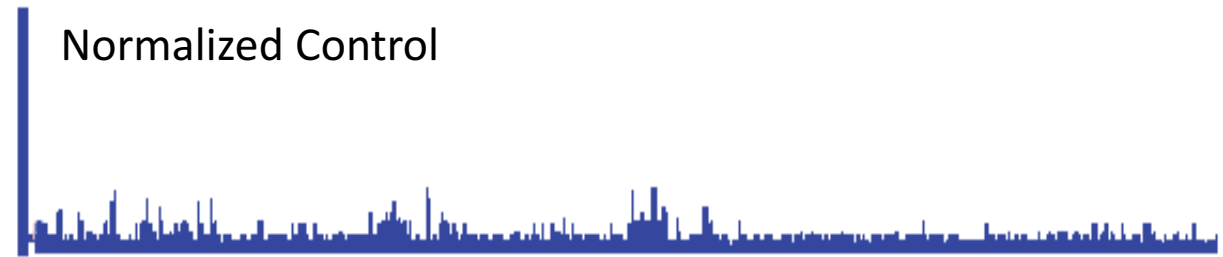
- Generate & threshold the signal profile to identify candidate target regions
 - Simulation (PeakSeq),
 - Local window based Poisson (MACS),
 - Fold change statistics (SPP)



Potential Targets



- Score against the control

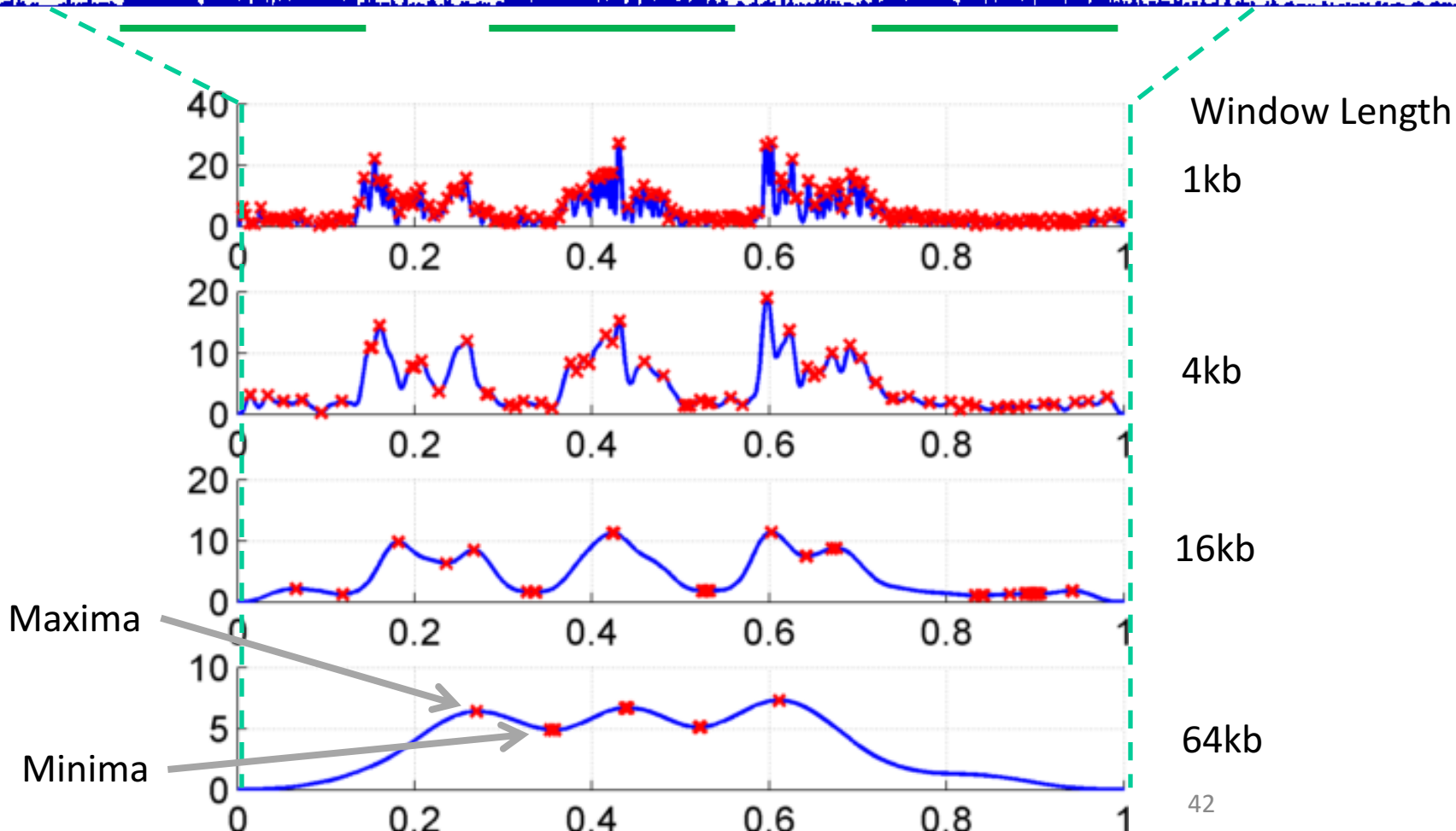
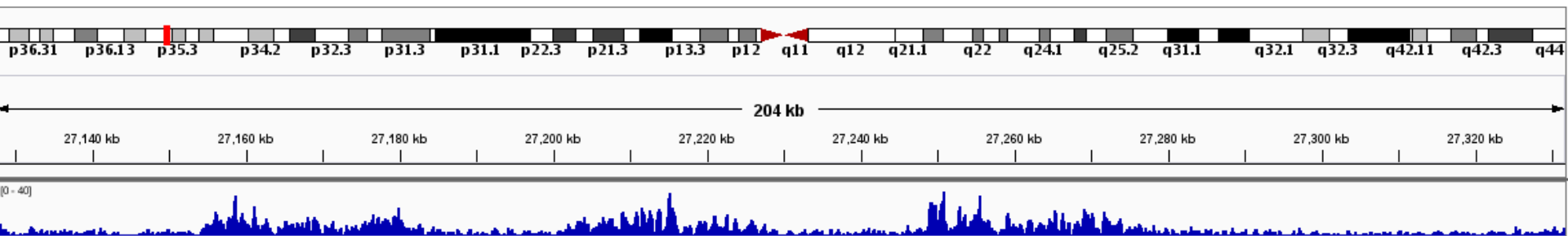


Significantly Enriched targets

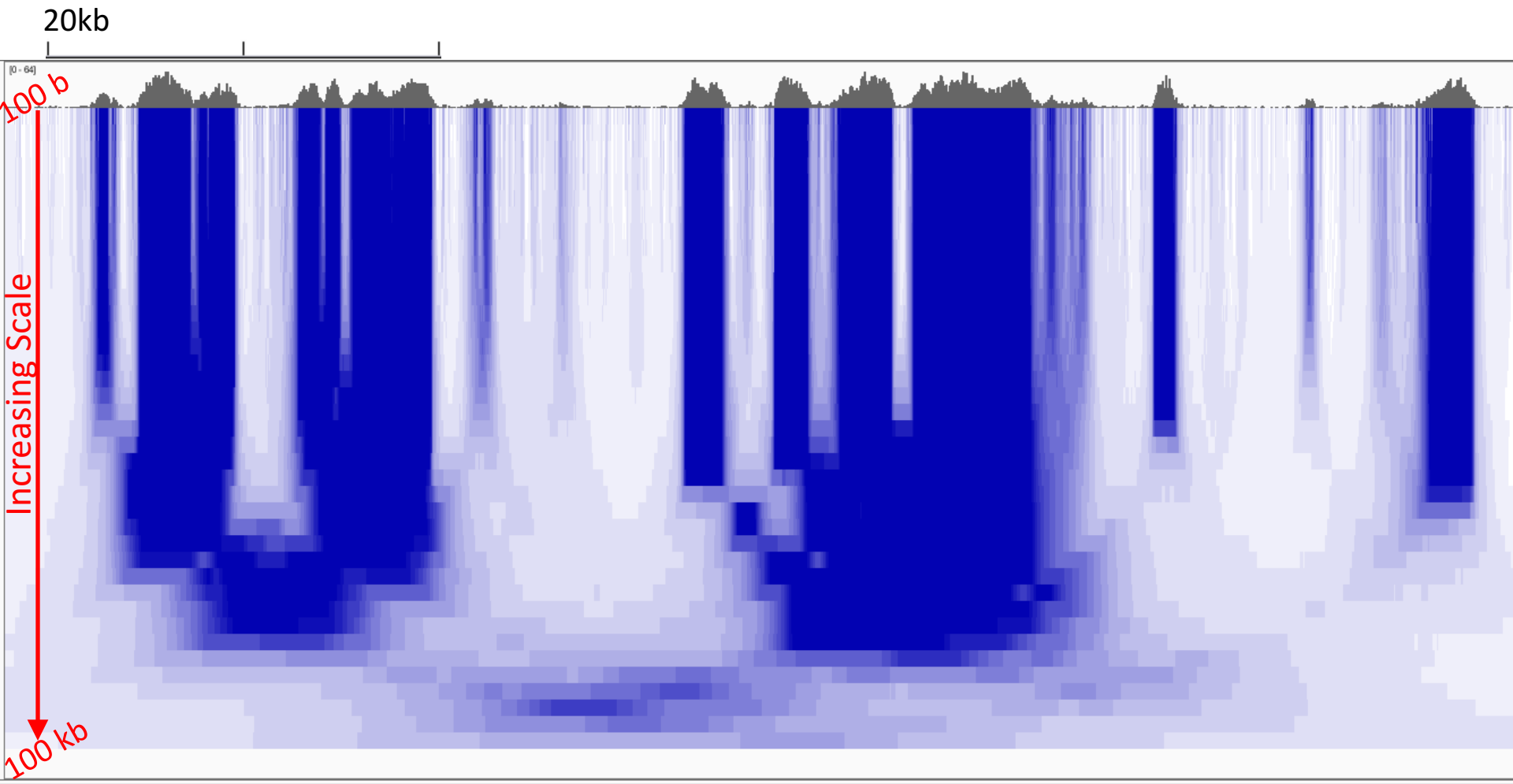


Now an update: "PeakSeq 2" => MUSIC

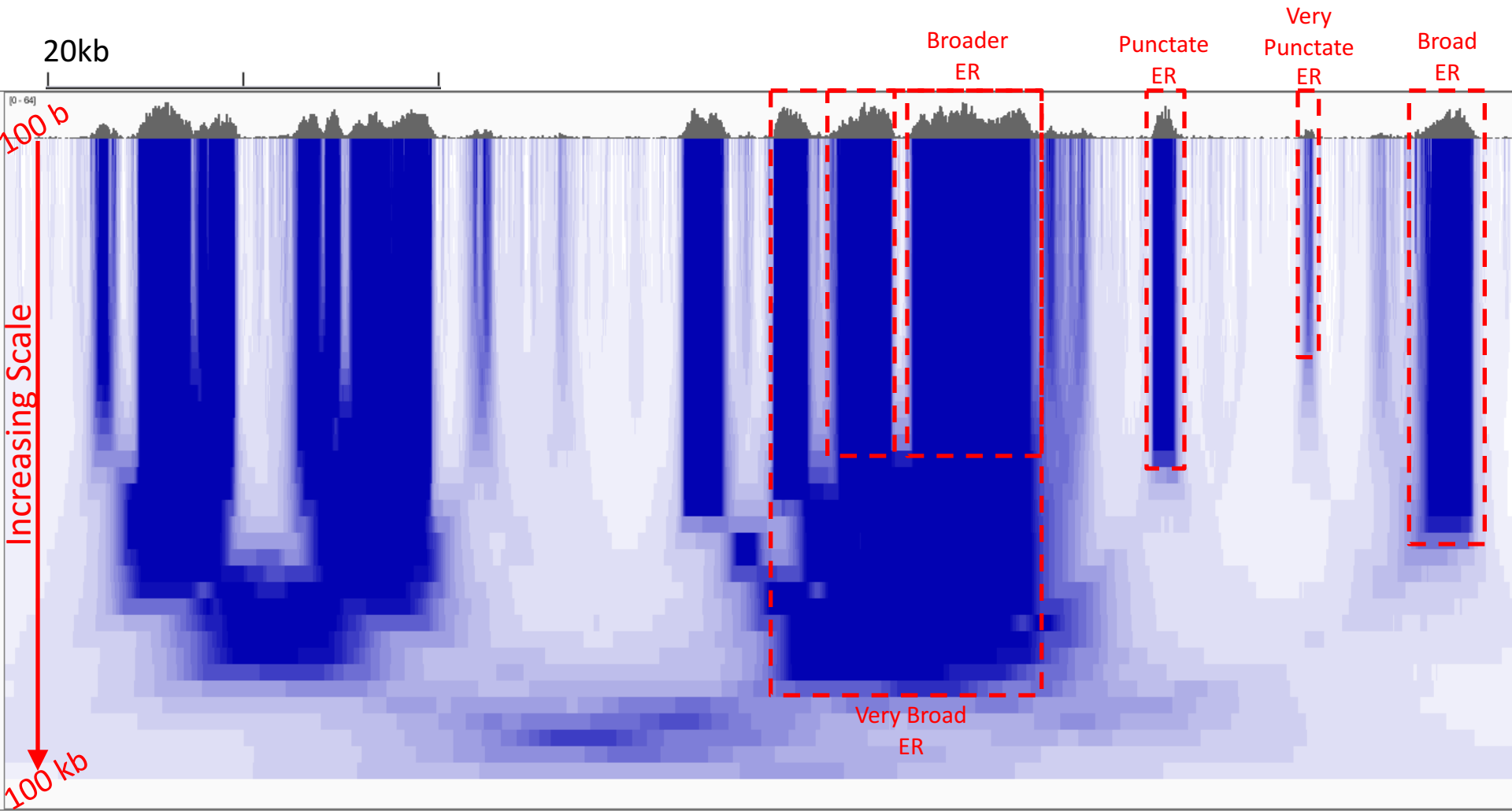
Multiscale Analysis, Minima/Maxima based Coarse Segmentation



Multiscale Decomposition



Multiscale Decomposition



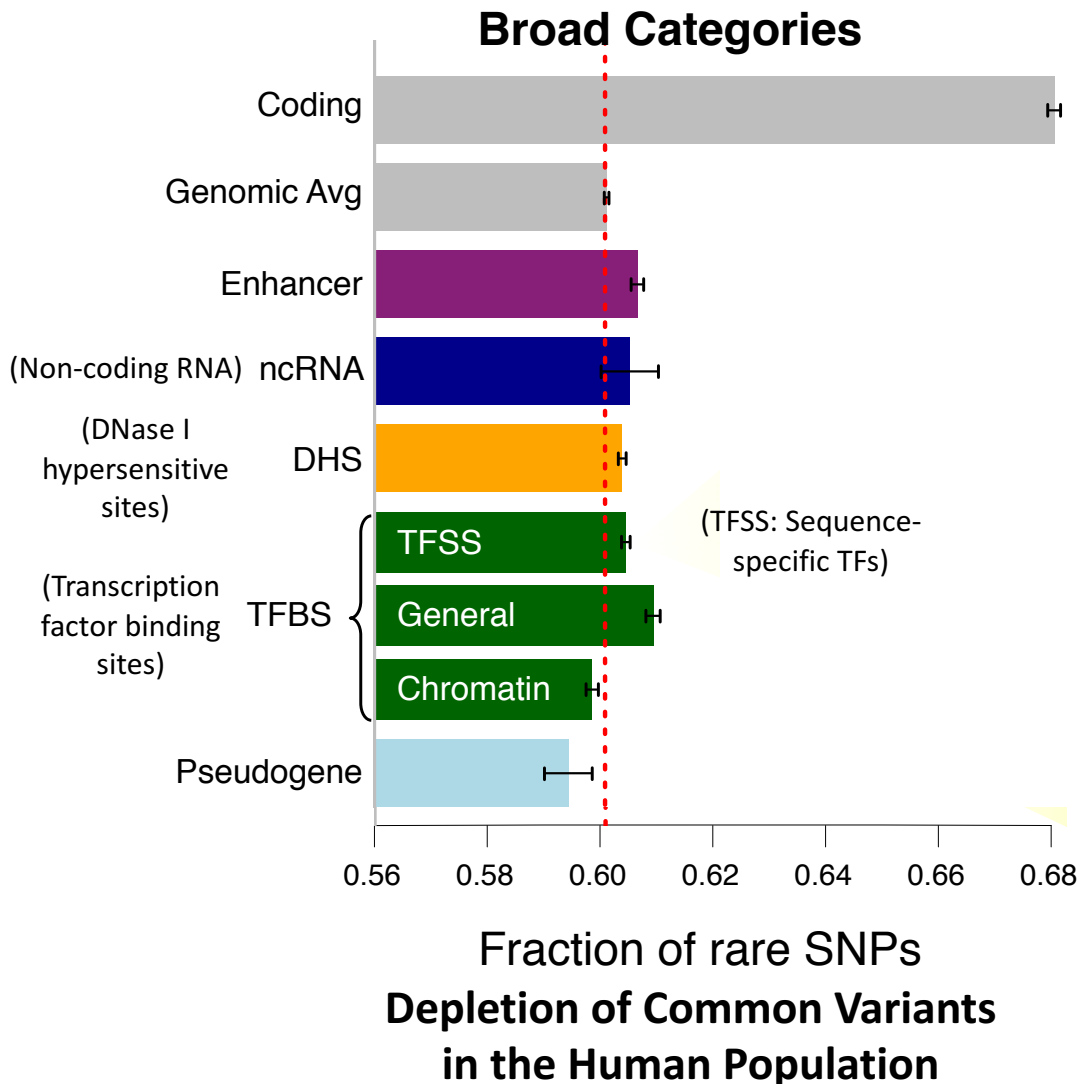
Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

Finding "Conserved" Sites in the Human Population:

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

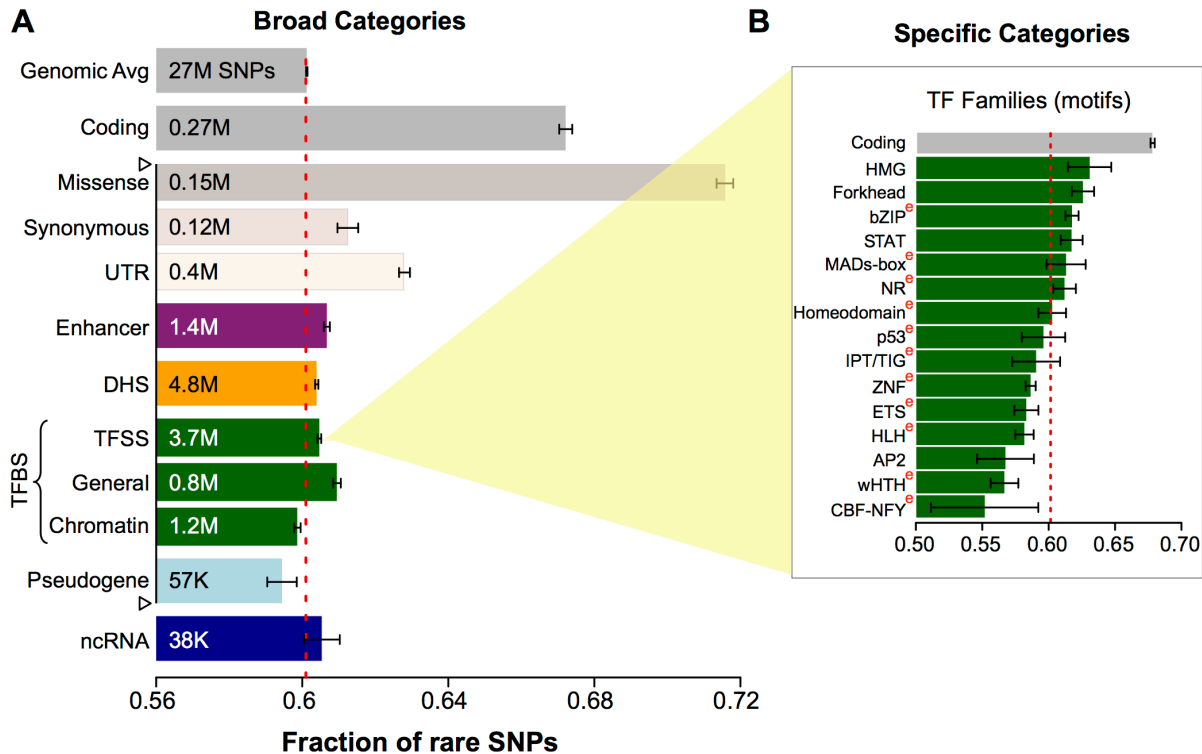


- Broad categories of regulatory regions under negative selection

• Related to:

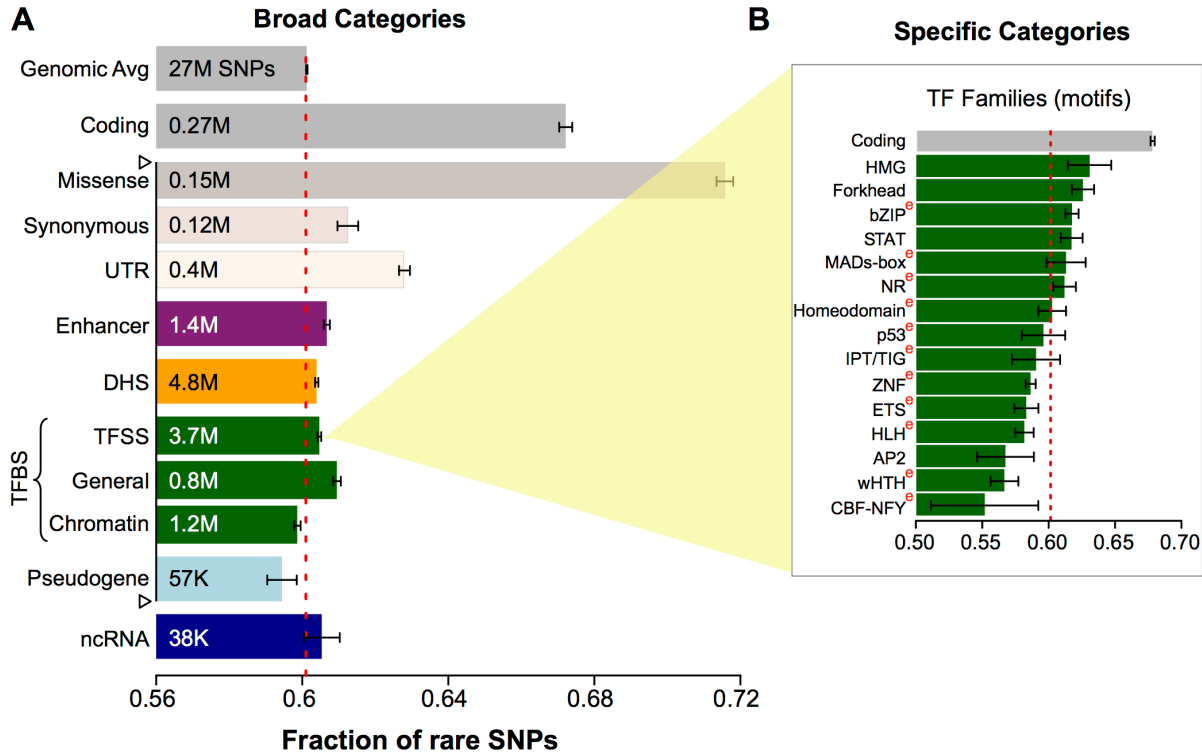
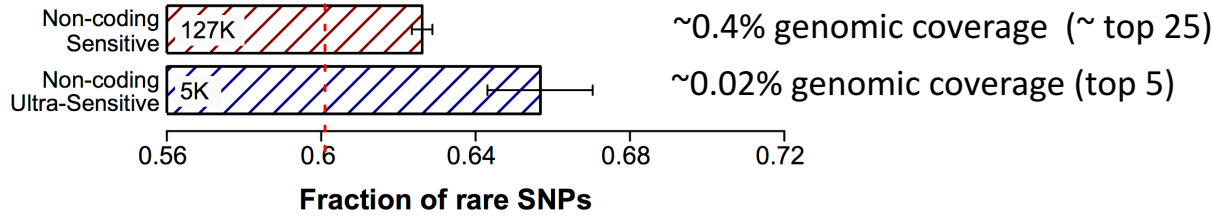
ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

Differential selective constraints among specific sub-categories



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

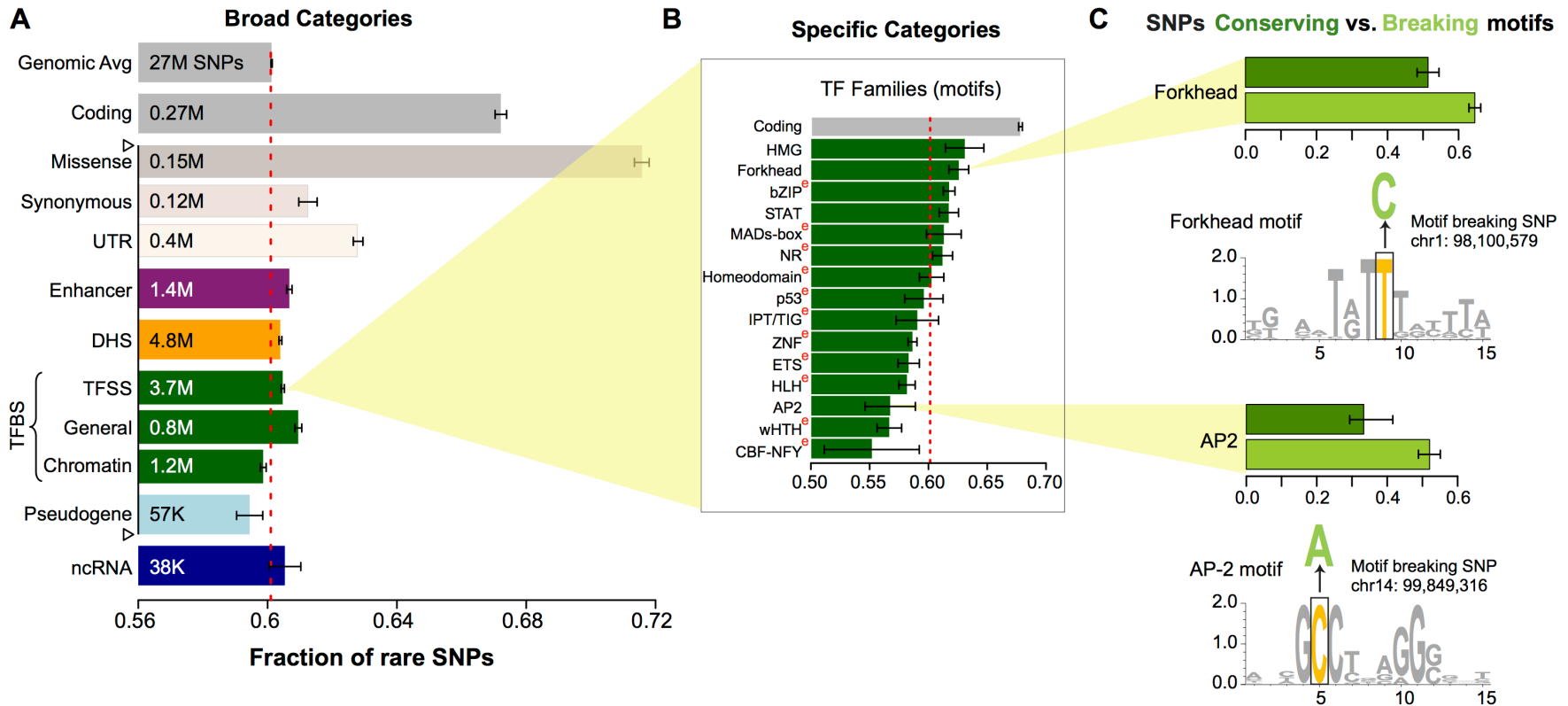
Defining Sensitive non-coding Regions



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

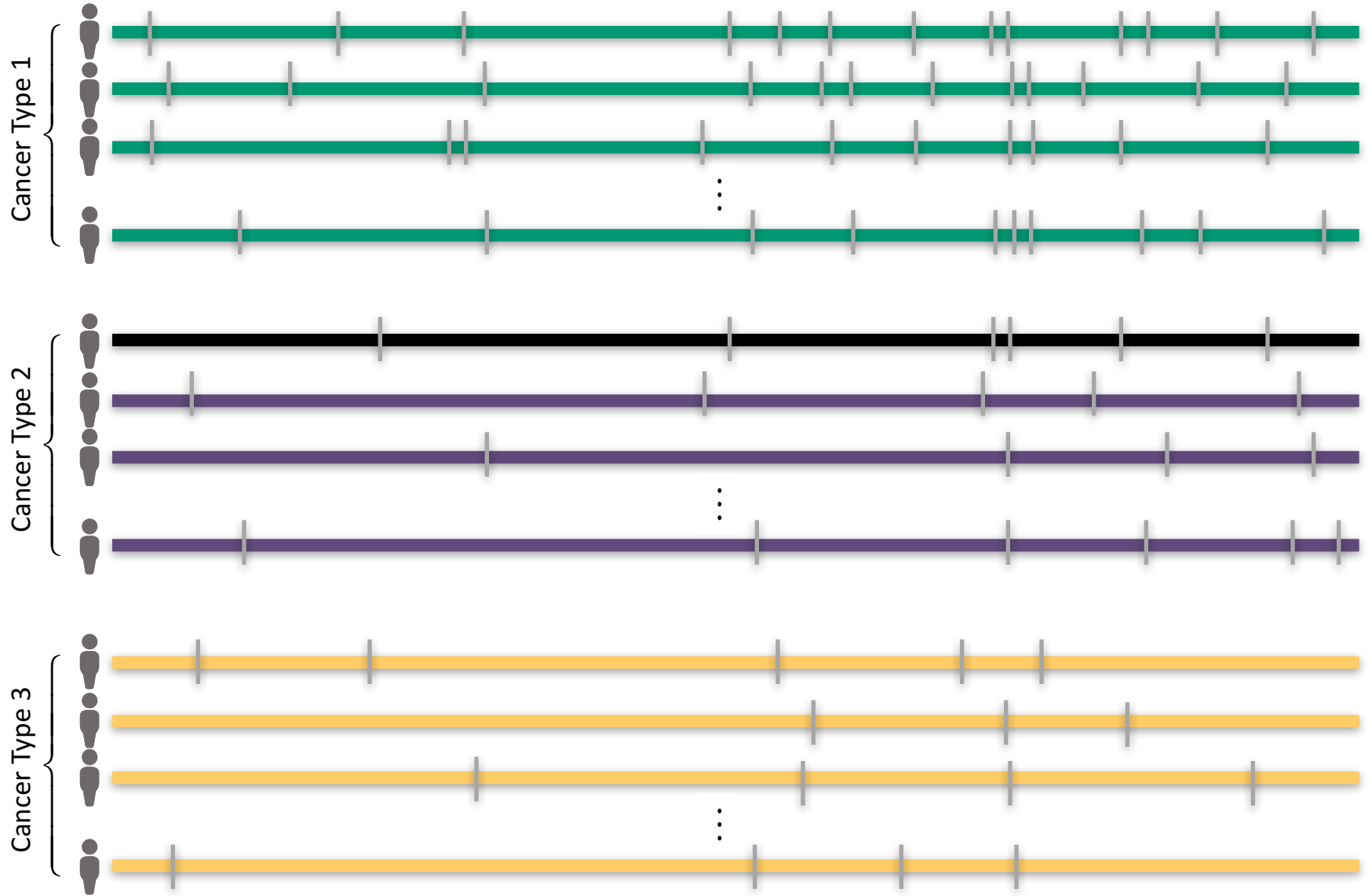


Personal Genomics:

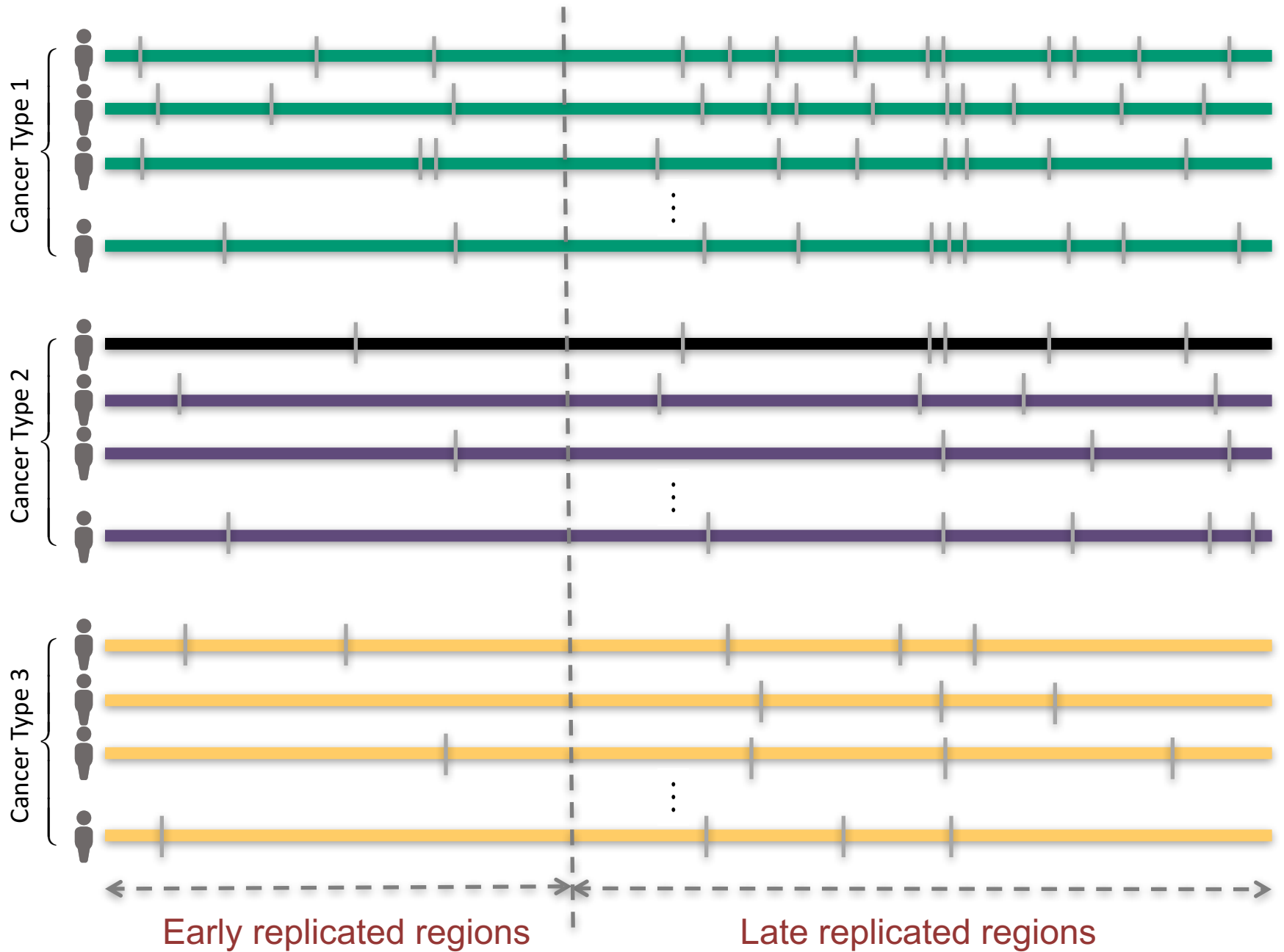
Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

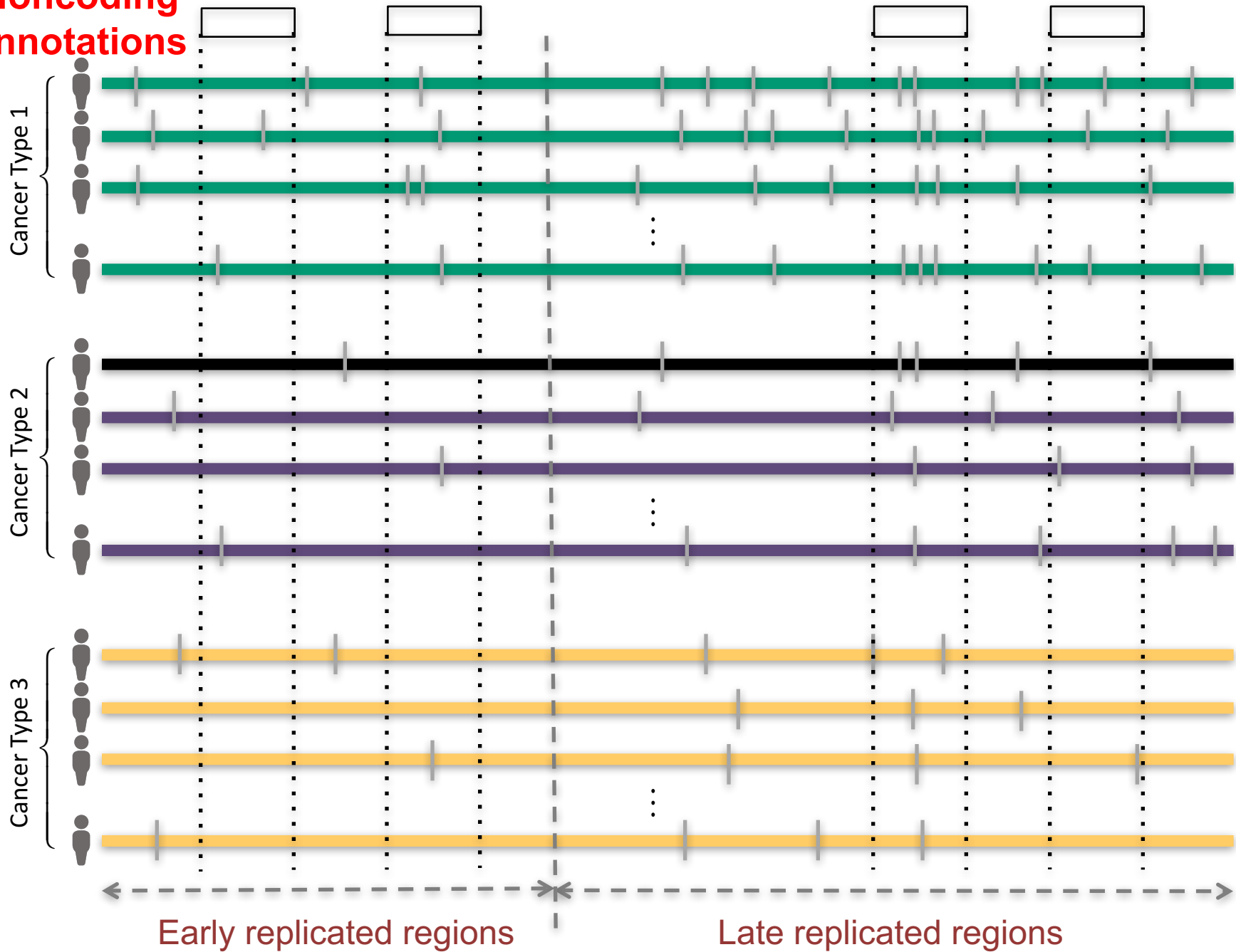
Mutation recurrence



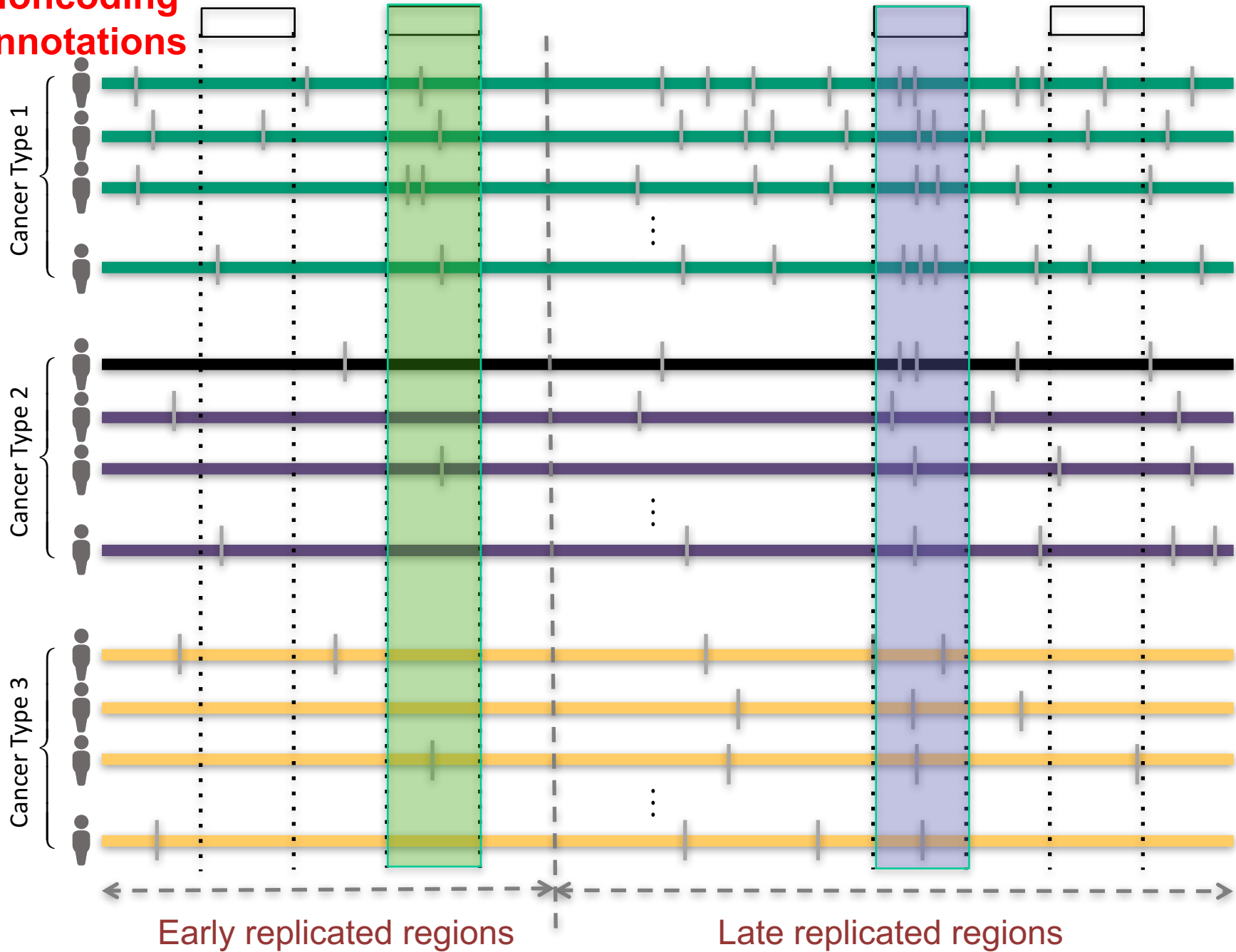
Mutation recurrence



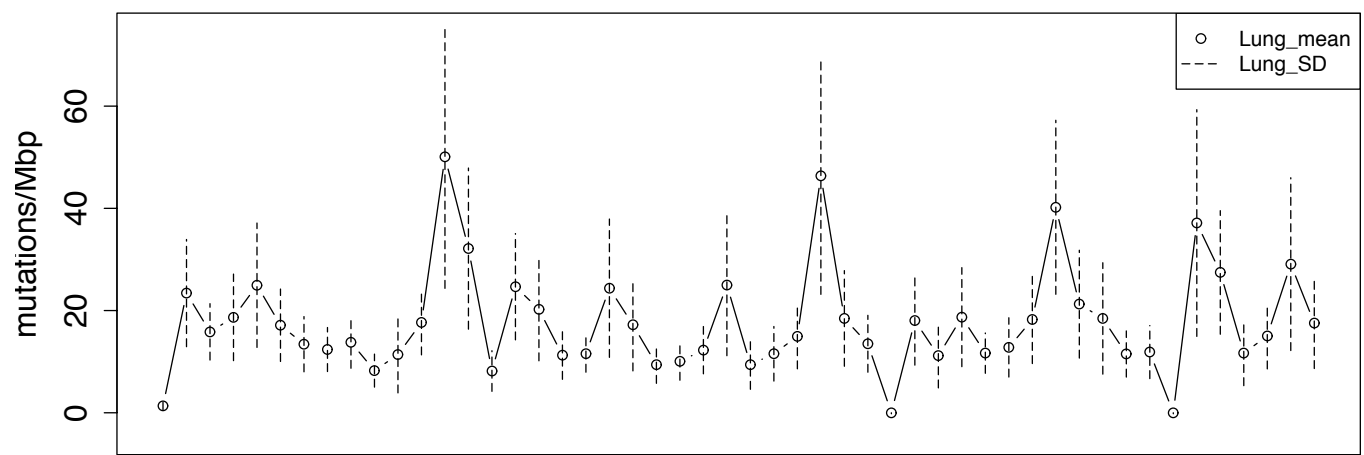
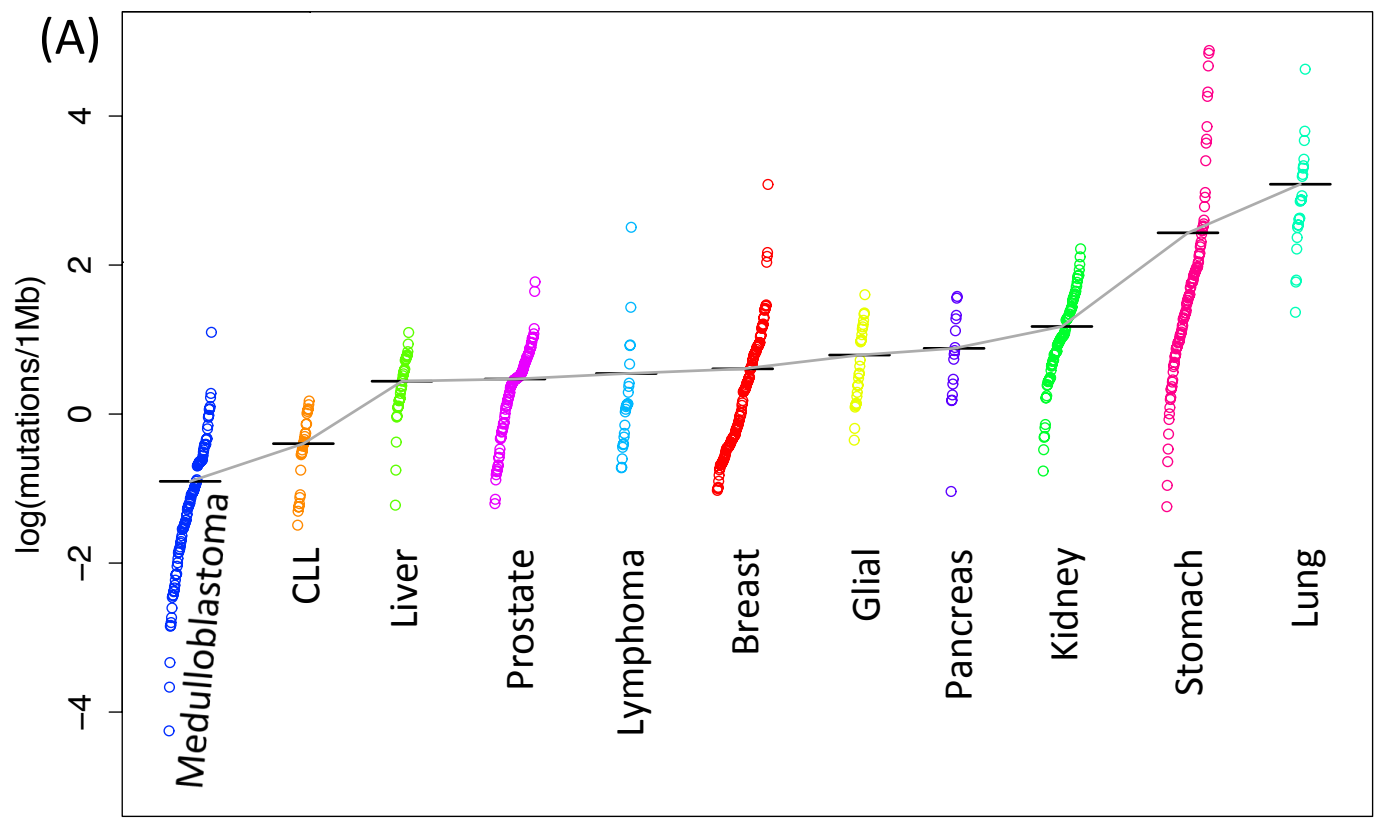
Noncoding annotations



Noncoding annotations



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



1 Mbp genome regions (locations chosen at random)

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

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 : 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

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

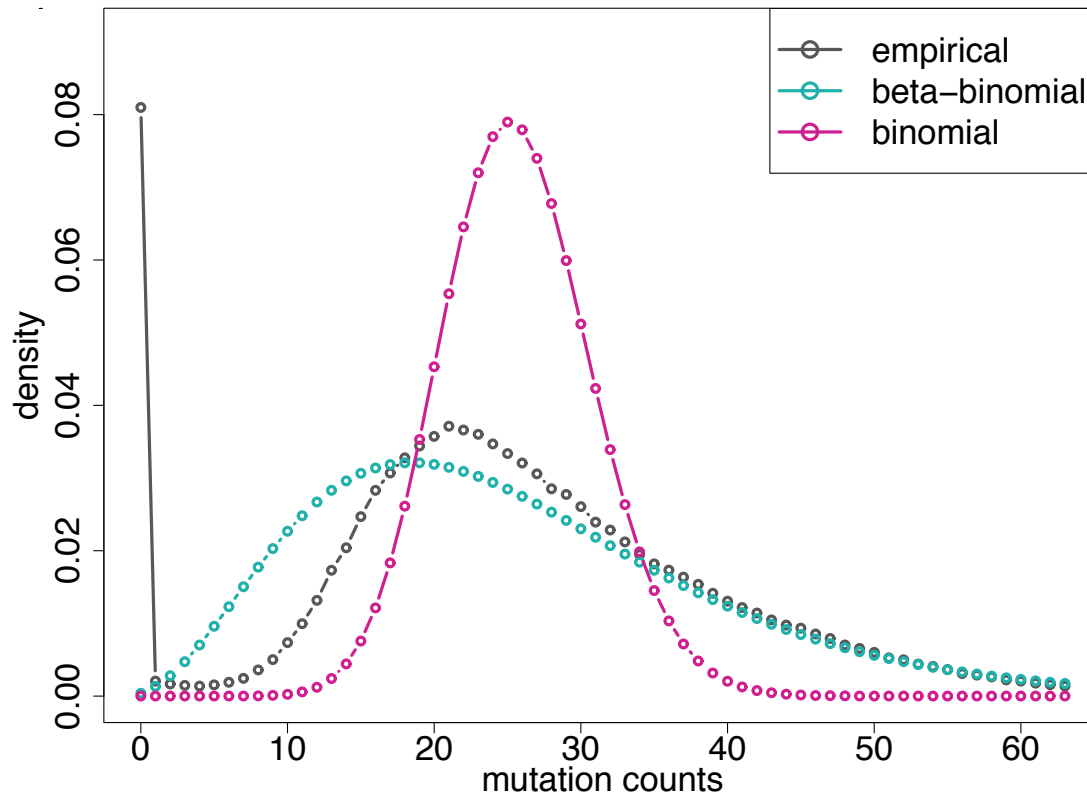
$$p_i : \text{Beta}(\mu | \mathbf{R}, \sigma | \mathbf{R})$$

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

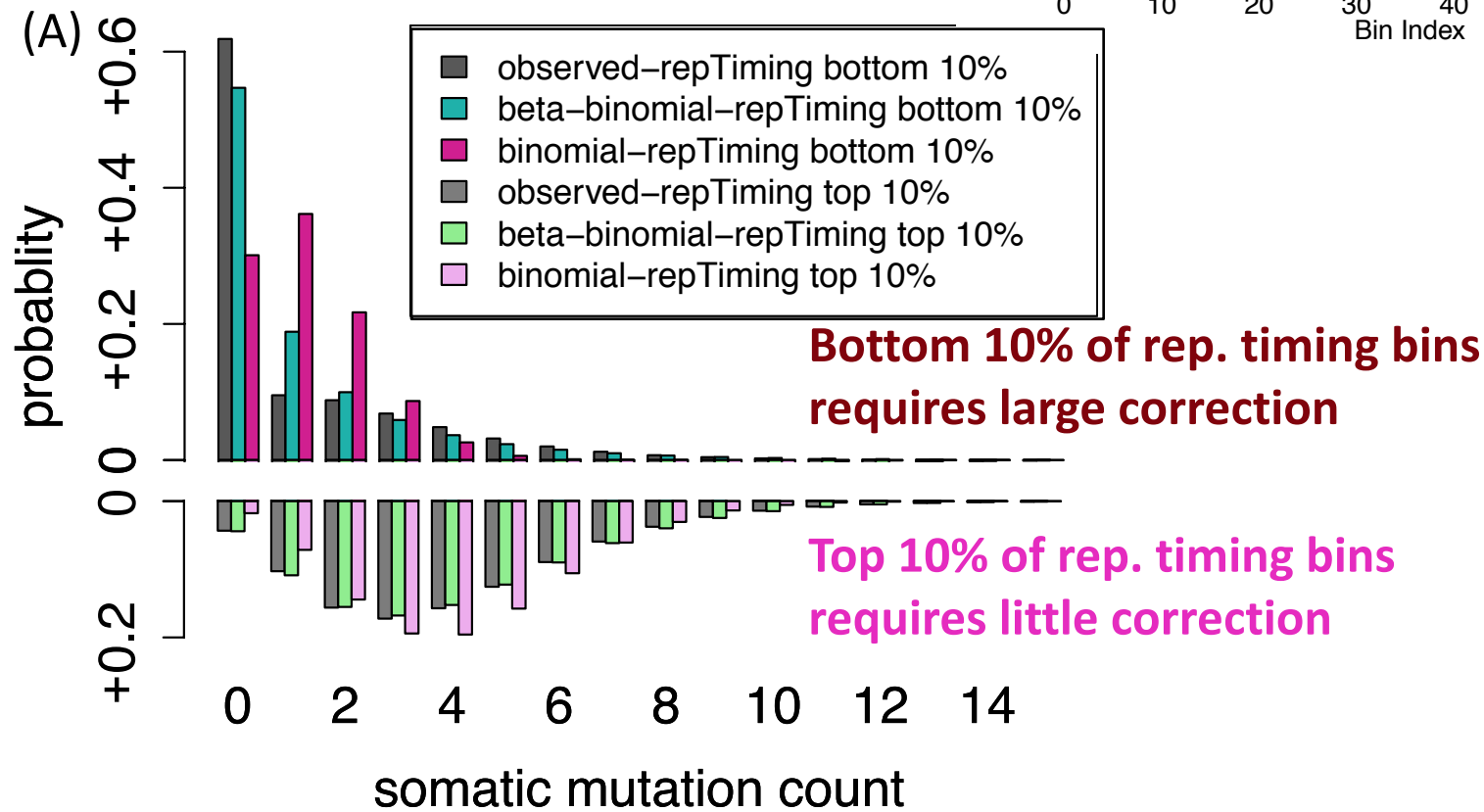
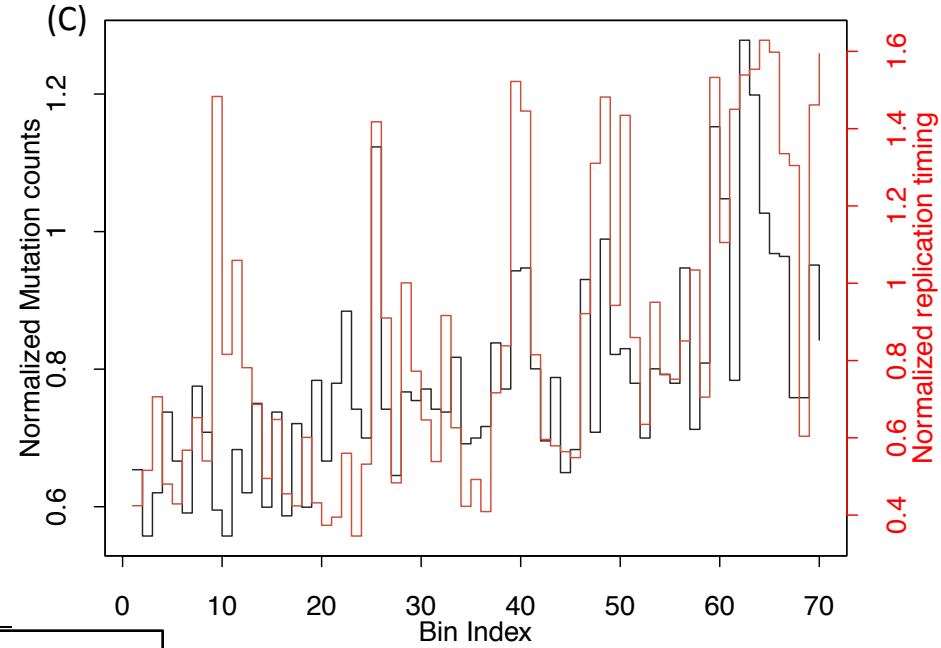
[Lochovsky et al. *NAR* ('15)]

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

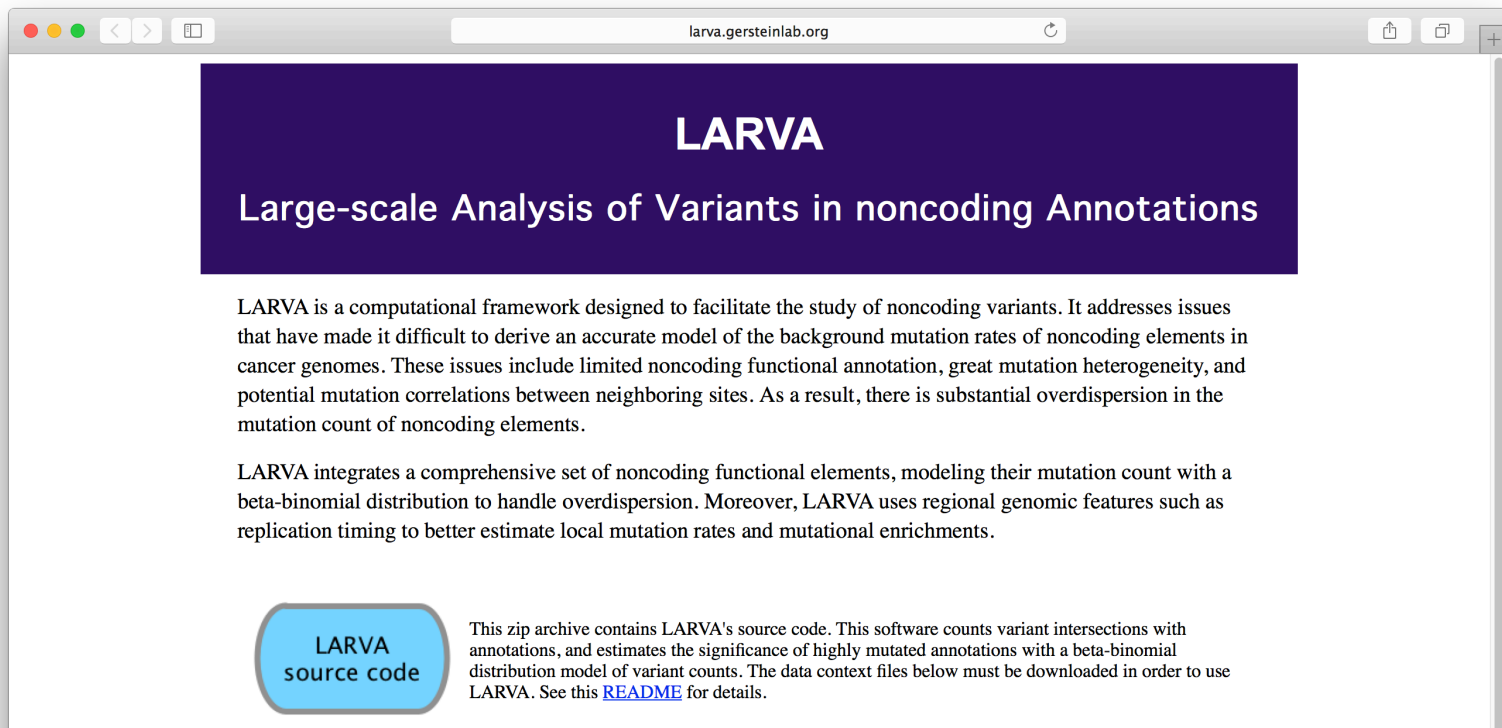


Adding DNA replication timing correction further improves the beta-binomial model



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. The main heading is "LARVA" in white text on a dark purple background, followed by the subtitle "Large-scale Analysis of Variants in noncoding Annotations". Below this, there is a paragraph of text describing the framework's purpose and a second paragraph detailing its methodology. 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.

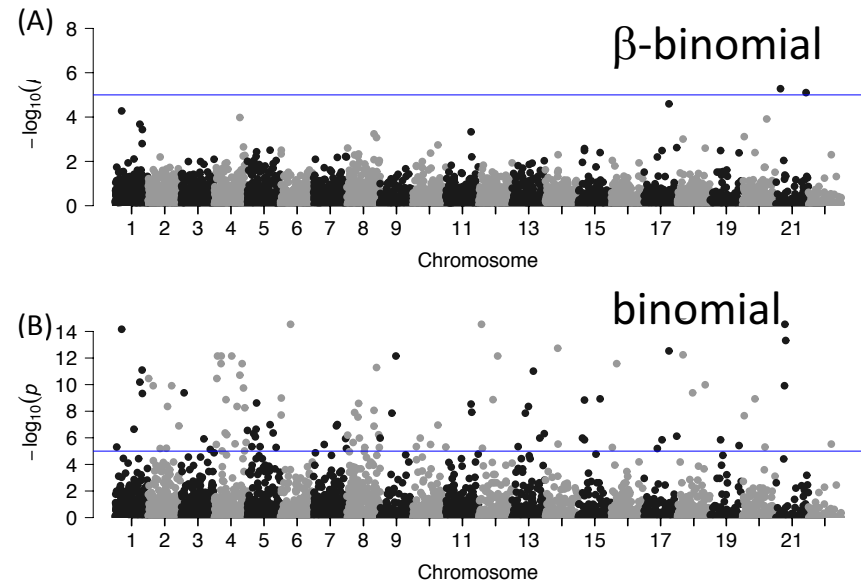
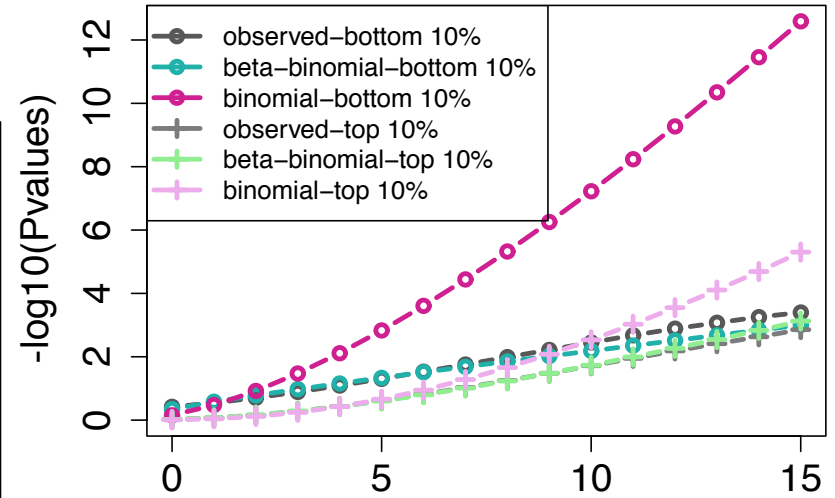
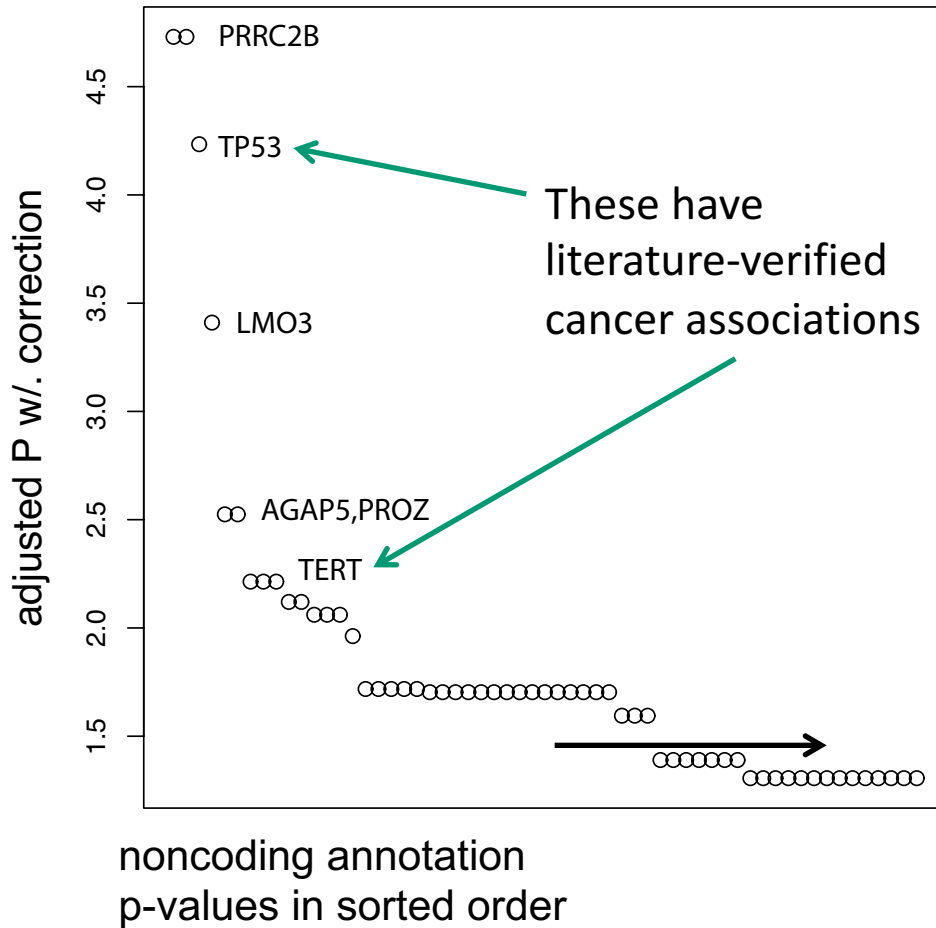
LARVA integrates a comprehensive set of noncoding functional elements, modeling their mutation count with a beta-binomial distribution to handle overdispersion. Moreover, LARVA uses regional genomic features such as replication timing to better estimate local mutation rates and mutational enrichments.

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

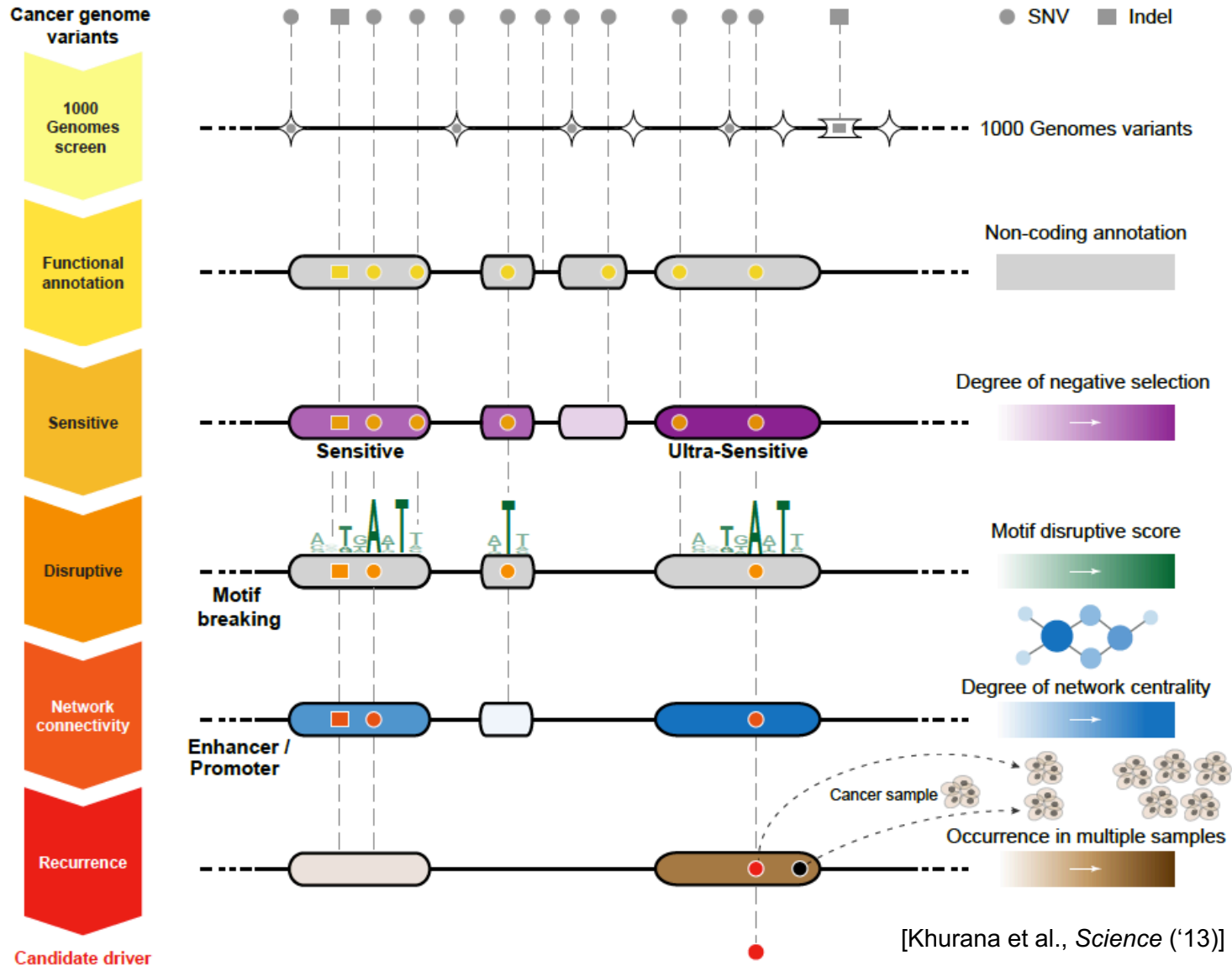


Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

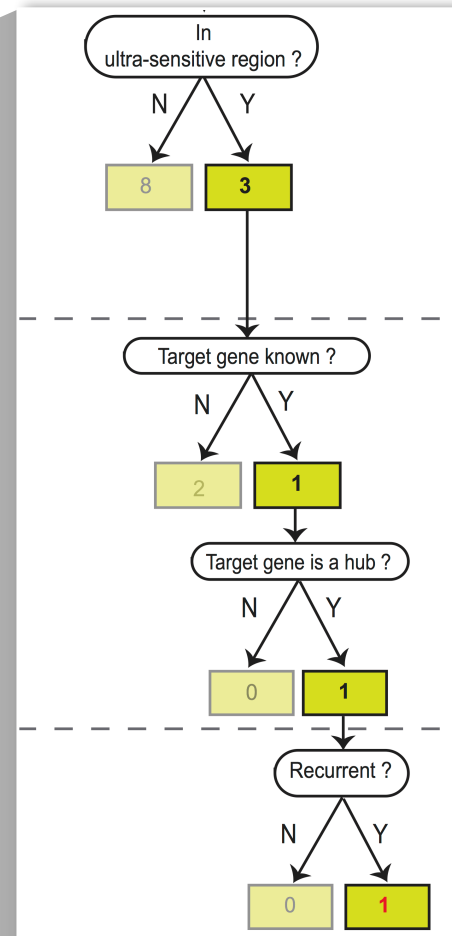
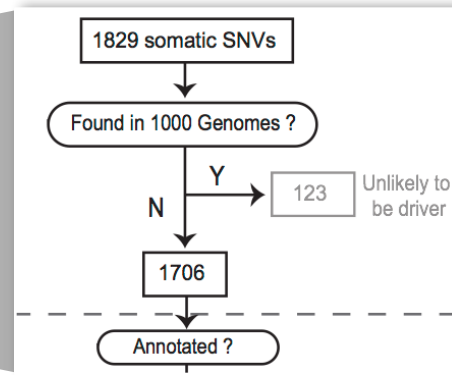
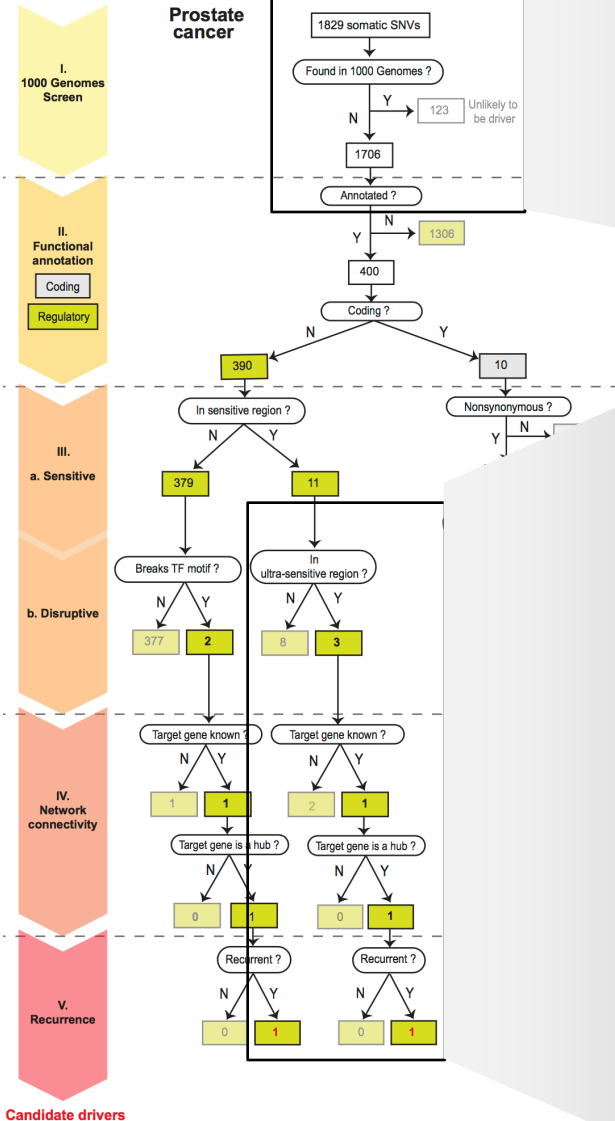
- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

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

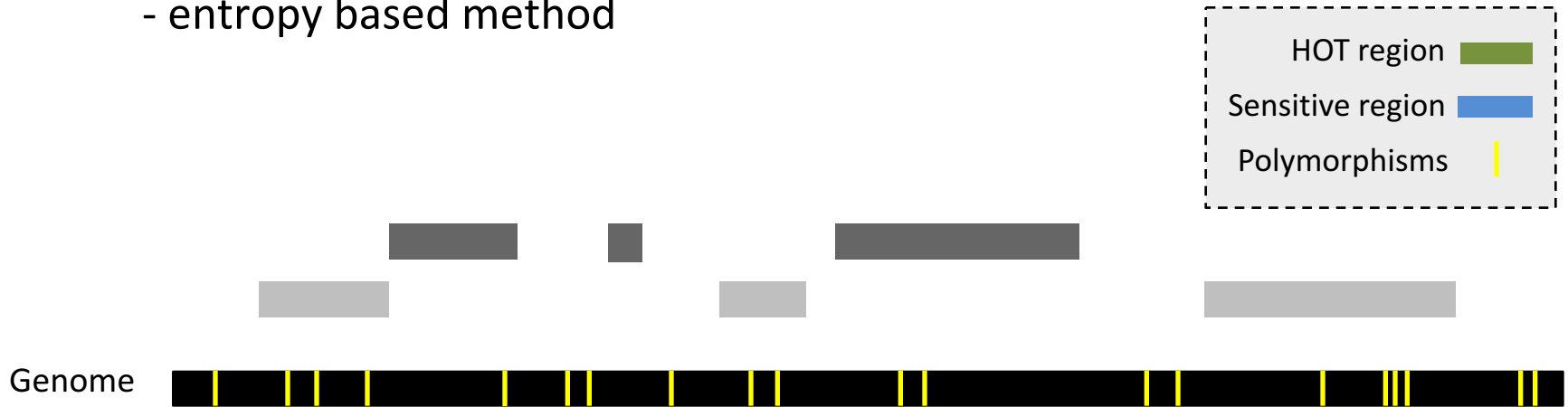
Highlighting variants

User Variants

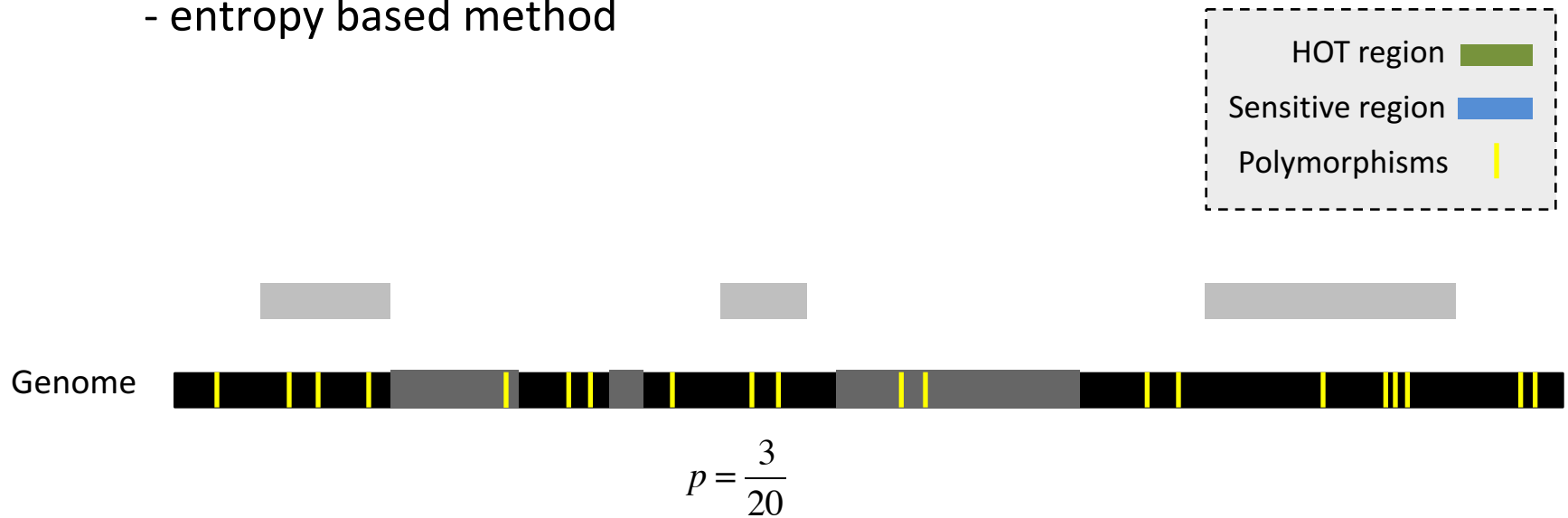
Variant Reports

FunSeq.gersteinlab.org

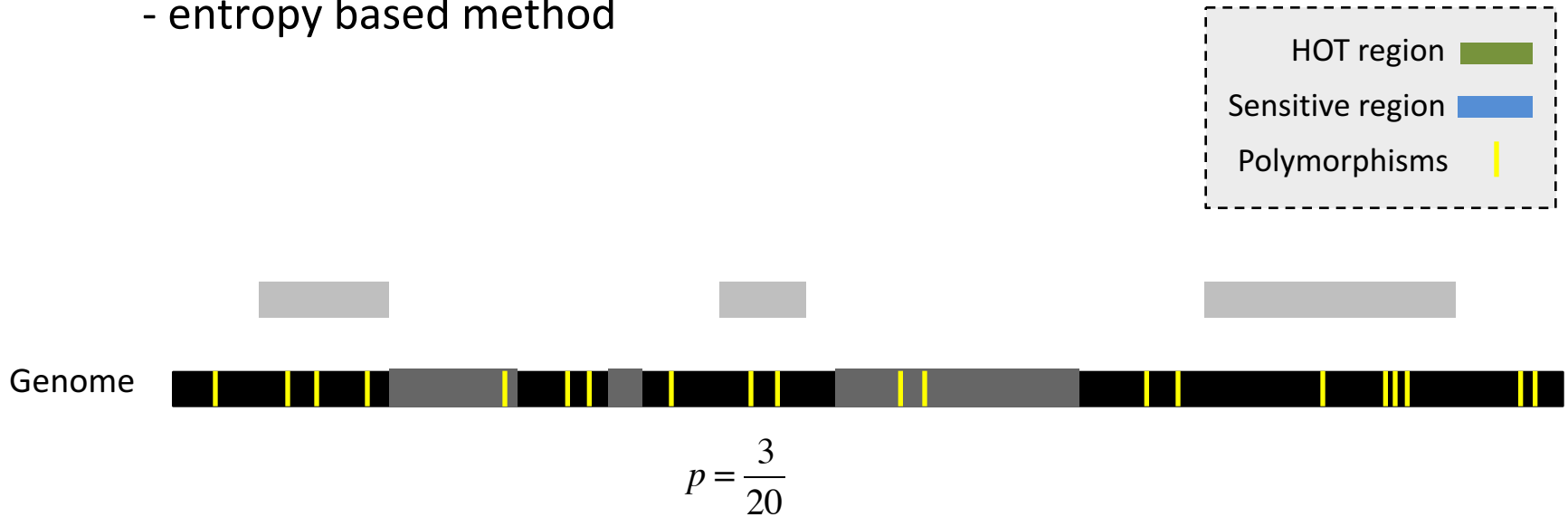
- Feature weight
 - Weighted with mutation patterns in natural polymorphisms
(features frequently observed weight less)
 - entropy based method



- 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

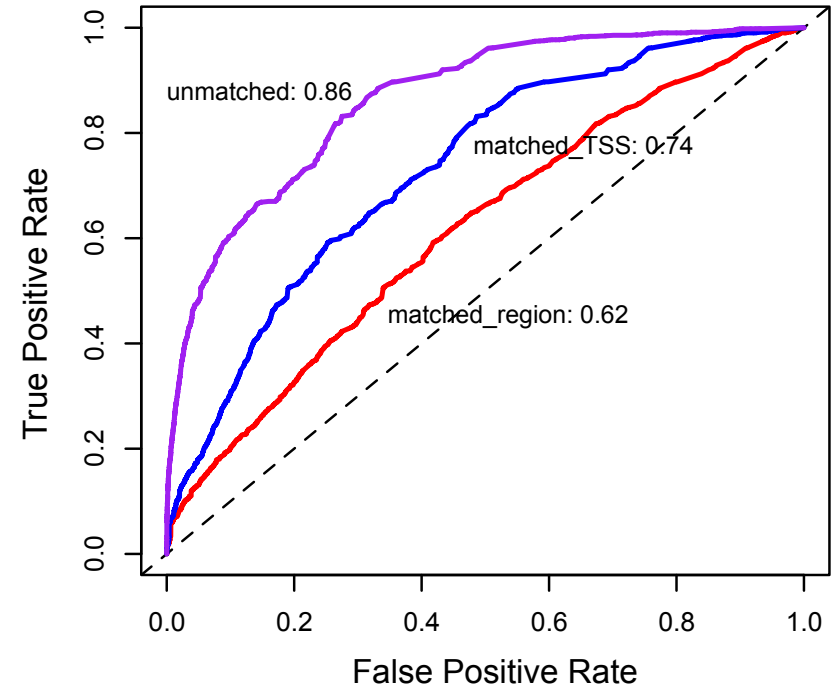
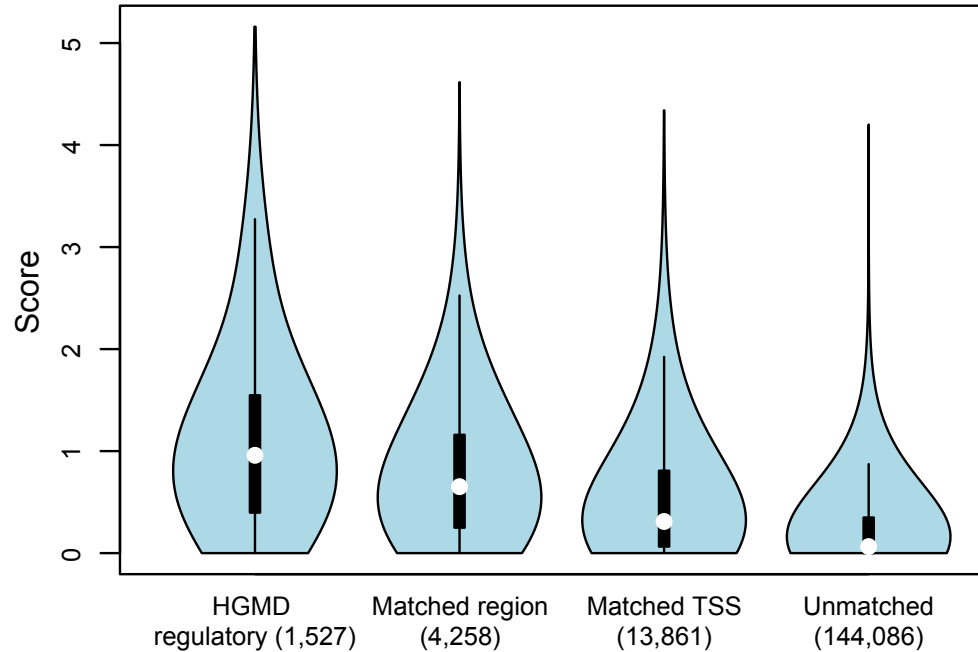


Feature weight: $w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d)$

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

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

Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with **STRESS** cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized **Frustration** to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with **MUSIC**
 - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
- Putting it together in Workflows
 - Using **LARVA** to do burden testing on non-coding annotation
 - Need to correct for co-variates & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using **FunSeq** to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

Personal Genomics:

Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
 - The exponential scaling of data generation & processing
 - The landscape of variants in personal genomes suggests prioritizing a few is an efficient interpretation strategy
- Characterizing Rare Variants in Coding Regions
 - Identifying with STRESS cryptic allosteric sites
 - On surface & in interior bottlenecks
 - Using changes in localized Frustration to find further sites sensitive to mutations
 - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
 - Annotating non-coding regions on different scales with MUSIC
 - Prioritizing rare variants with “sensitive sites” (human-conserved)
- Putting it together in Workflows
 - Using LARVA to do burden testing on non-coding annotation
 - Need to correct for co-variables & over-dispersion mutation counts
 - Parameterized according to replication timing
 - Using FunSeq to integrate evidence on variants
 - Systematically weighting all the features
 - suggesting non-coding drivers
 - Prioritizing rare germline variants

github.com/gersteinlab/**Frustration**

S **Kumar**, D Clarke

MUSIC.gersteinlab.org

A **Harmanci**, J Rozowsky

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

STRESS.molmovdb.org

D **Clarke**, A **Sethi**,

S Li, S Kumar, R W.F. Chang,
J Chen

Hiring Postdocs. See gersteinlab.org/jobs

Acknowledgments

Extra



Info about content in this slide pack

- General PERMISSIONS
 - This Presentation is copyright Mark Gerstein, Yale University, 2016.
 - Please read permissions 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 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>