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

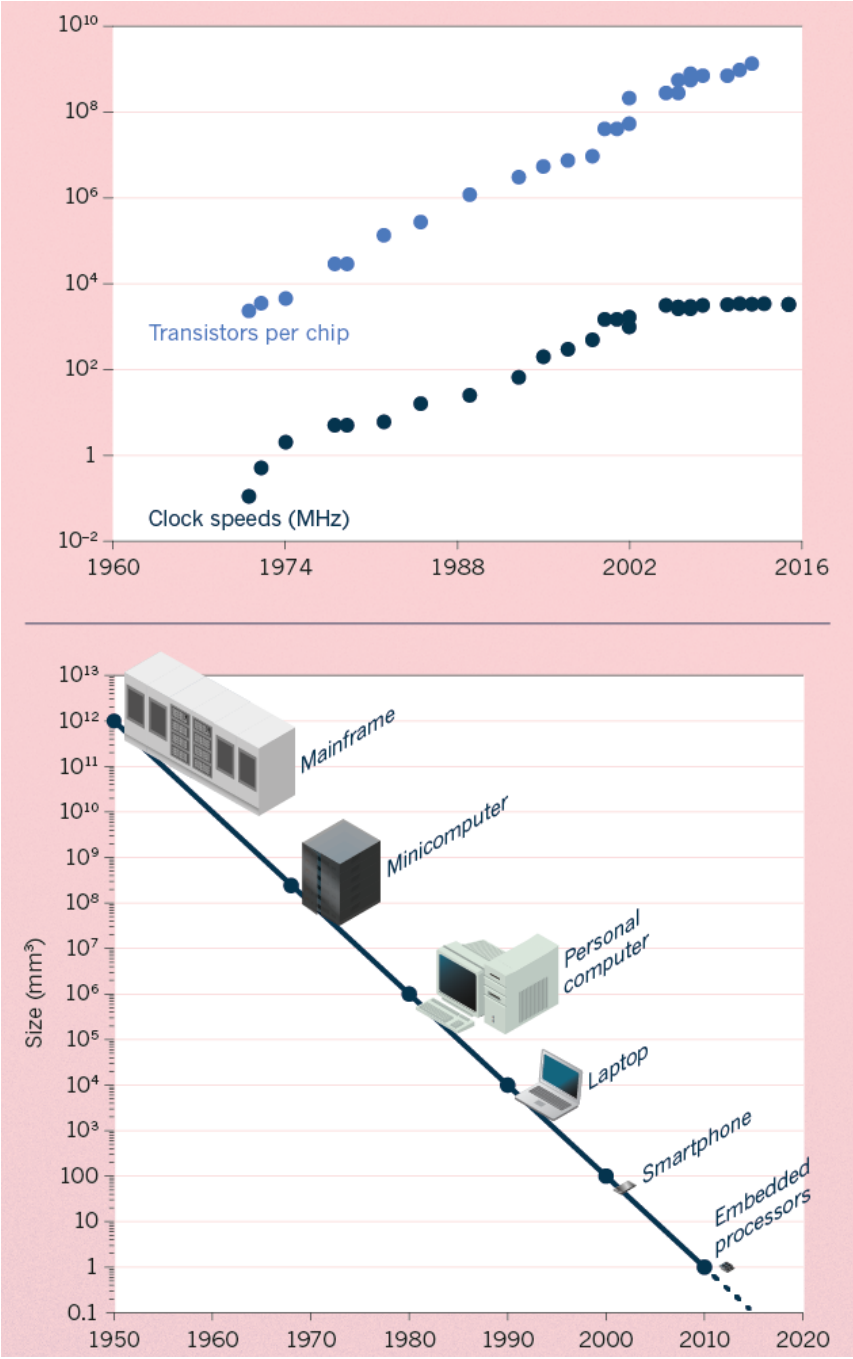
Mark Gerstein, Yale

Slides freely downloadable from Lectures.GersteinLab.org

& "tweetable" (via @markgerstein).
See last slide for more info.

Moore's Law: Exponential Scaling of Computer Technology

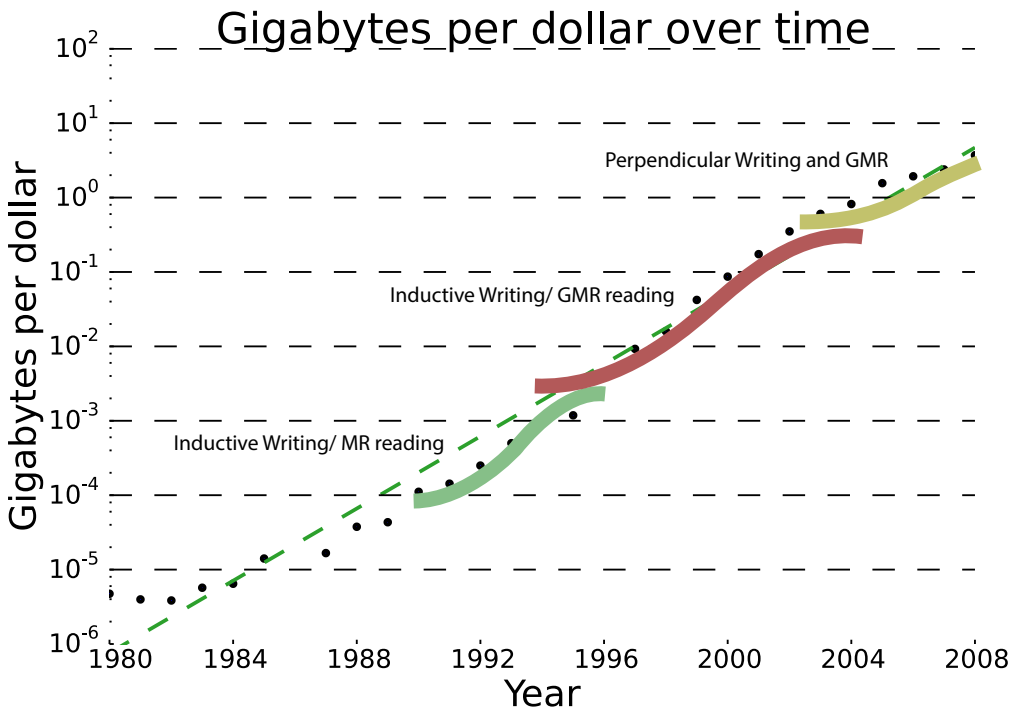
- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.

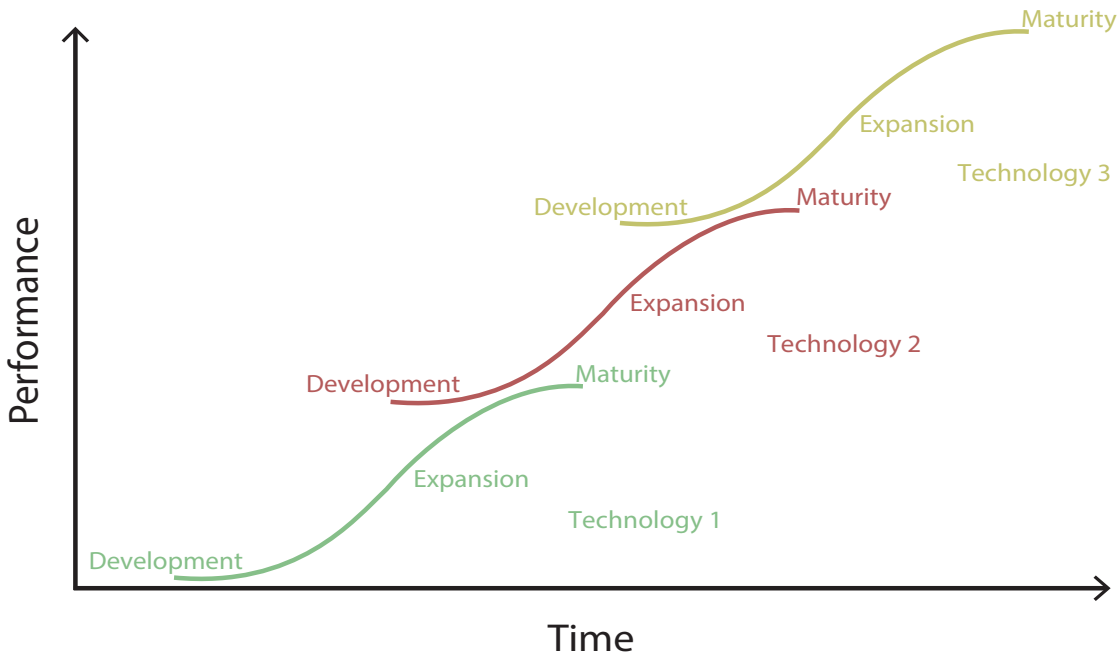


[Waldrop ('15) Nature]

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

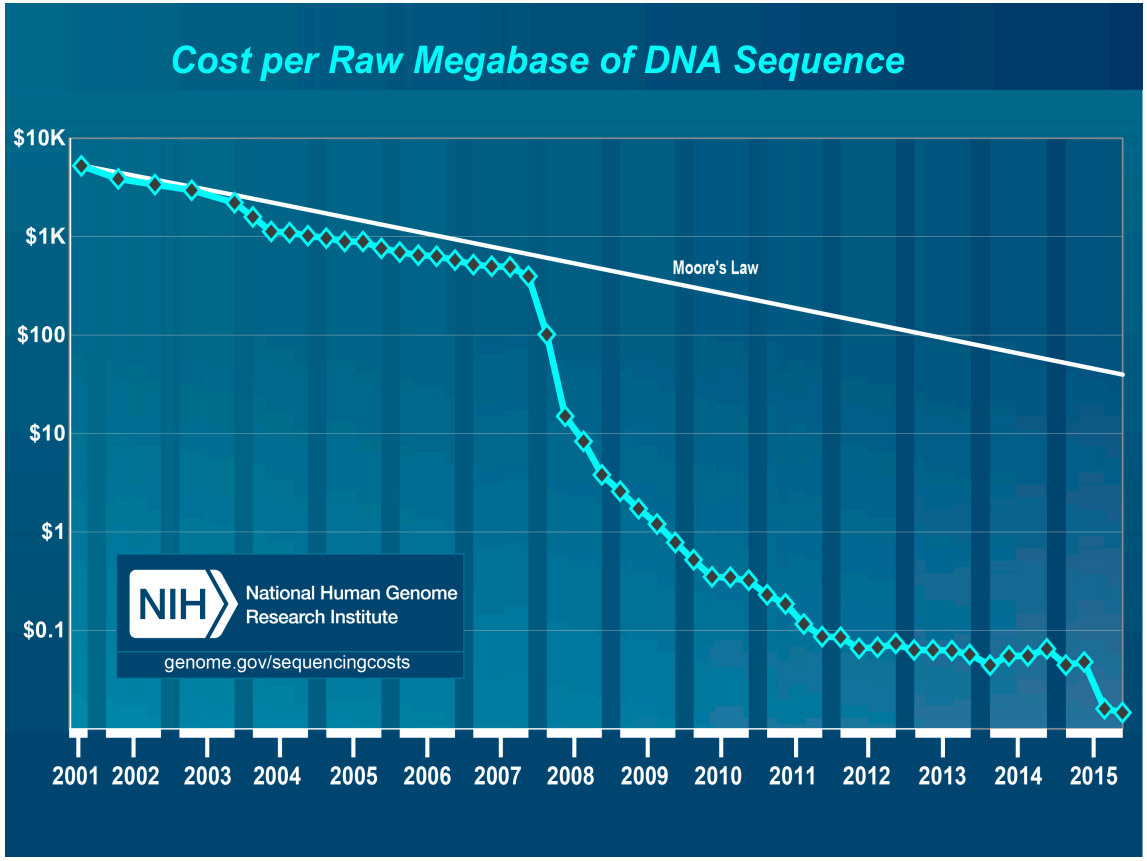




[Muir et al. ('15) GenomeBiol.]

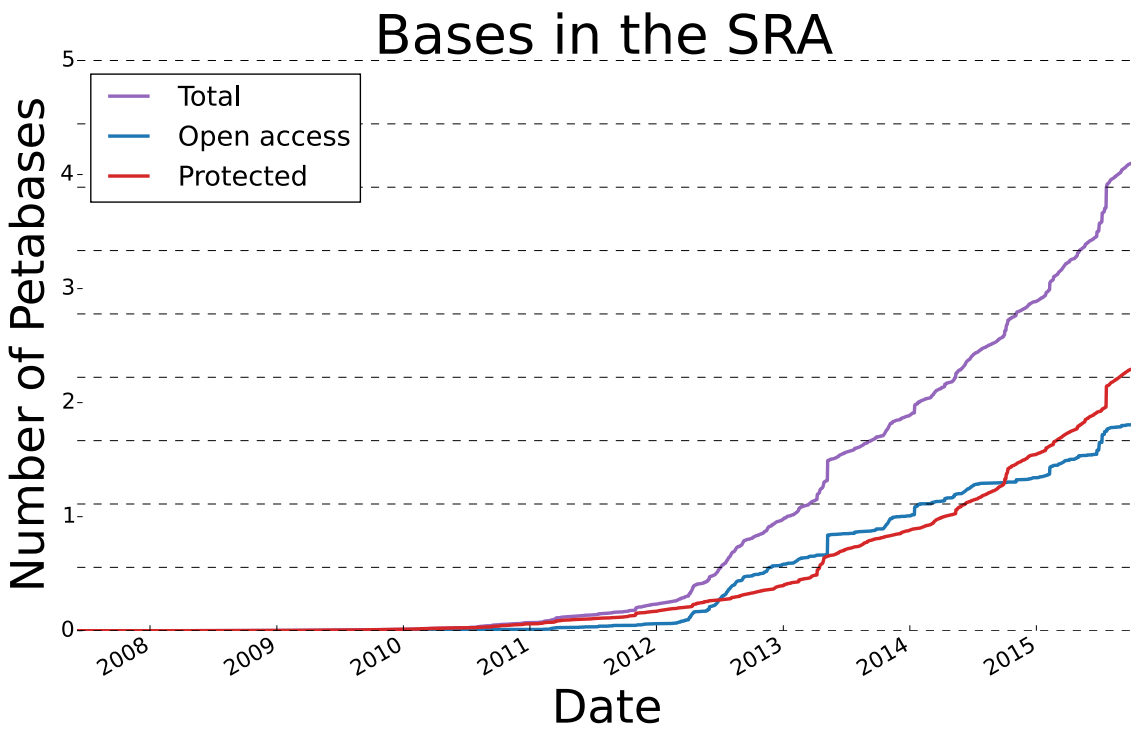
Sequencing Data Explosion: Faster than Moore's Law for a Time (or a S-curve)

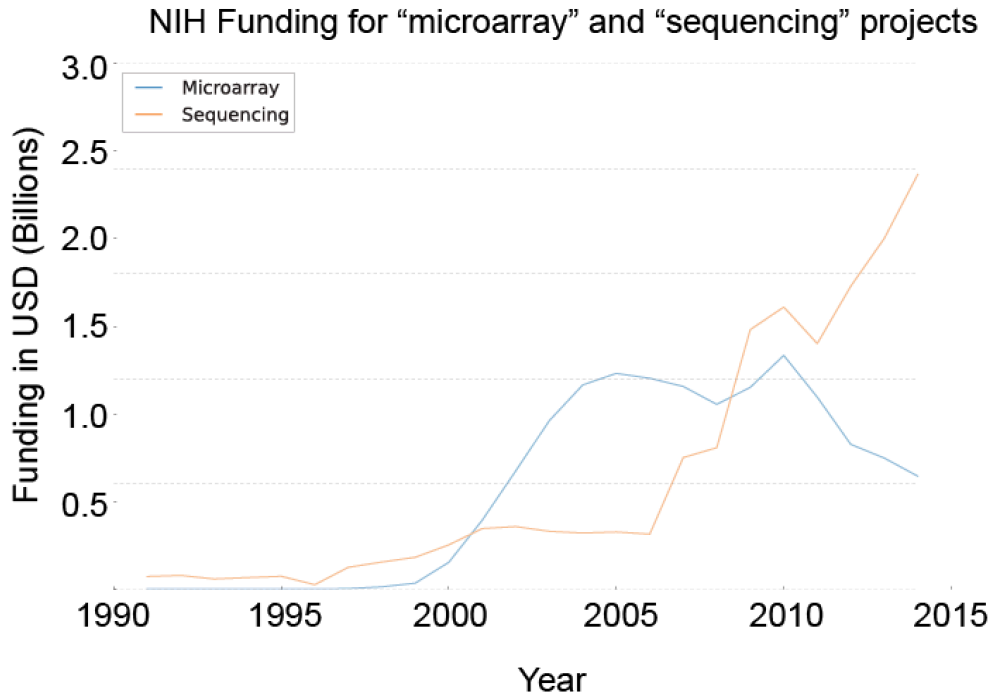
- DNA sequencing has gone through technological S-curves
  - In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
  - The advent of NGS was a shift to a new technology with dramatic decrease in cost).



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.





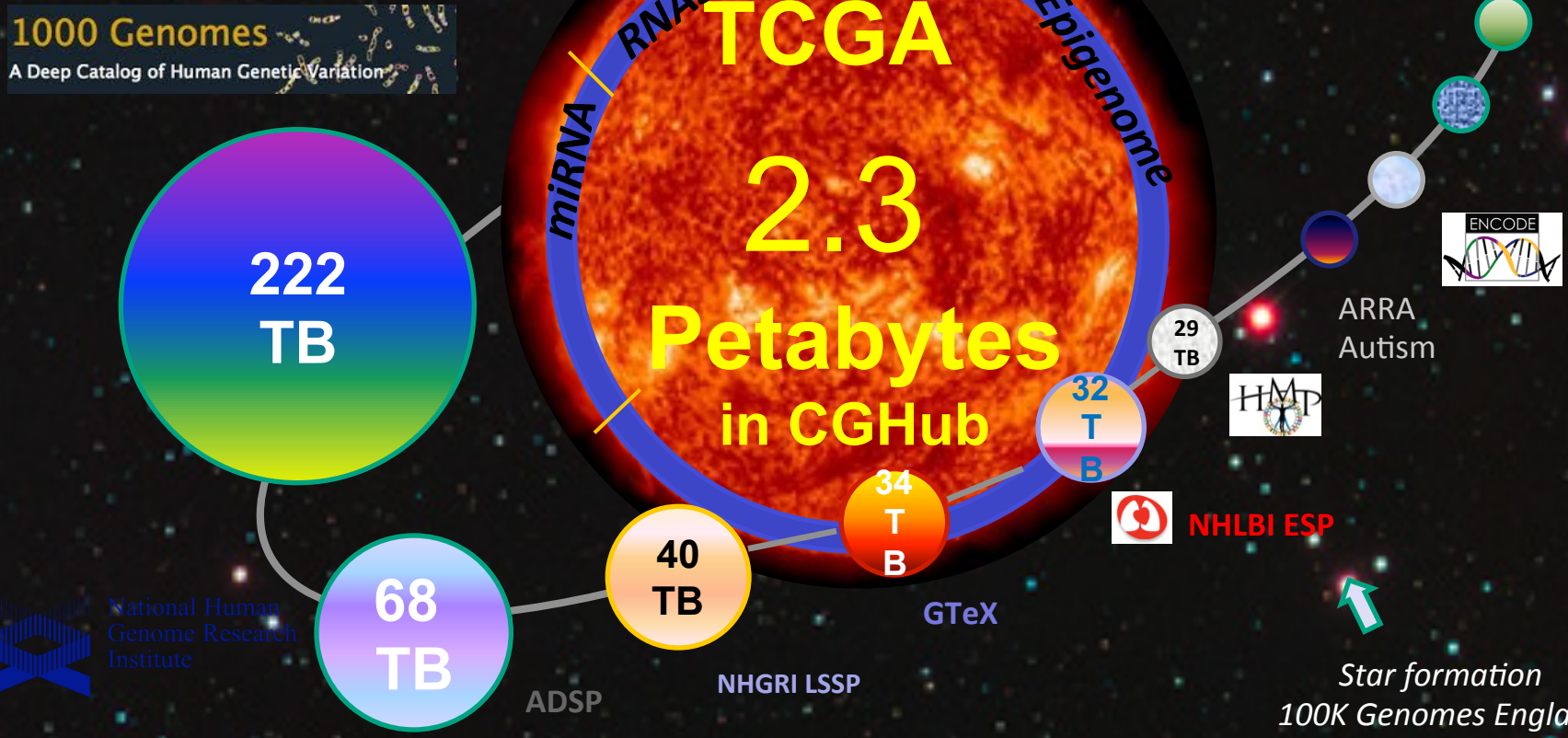
[Muir et al. (‘15) GenomeBiol.]

Sequence Universe

SRA ~1 petabyte

TCGA endpoint: ~2.5 Petabytes

- **~1.5 PB exome**
- **~1 PB whole genome**

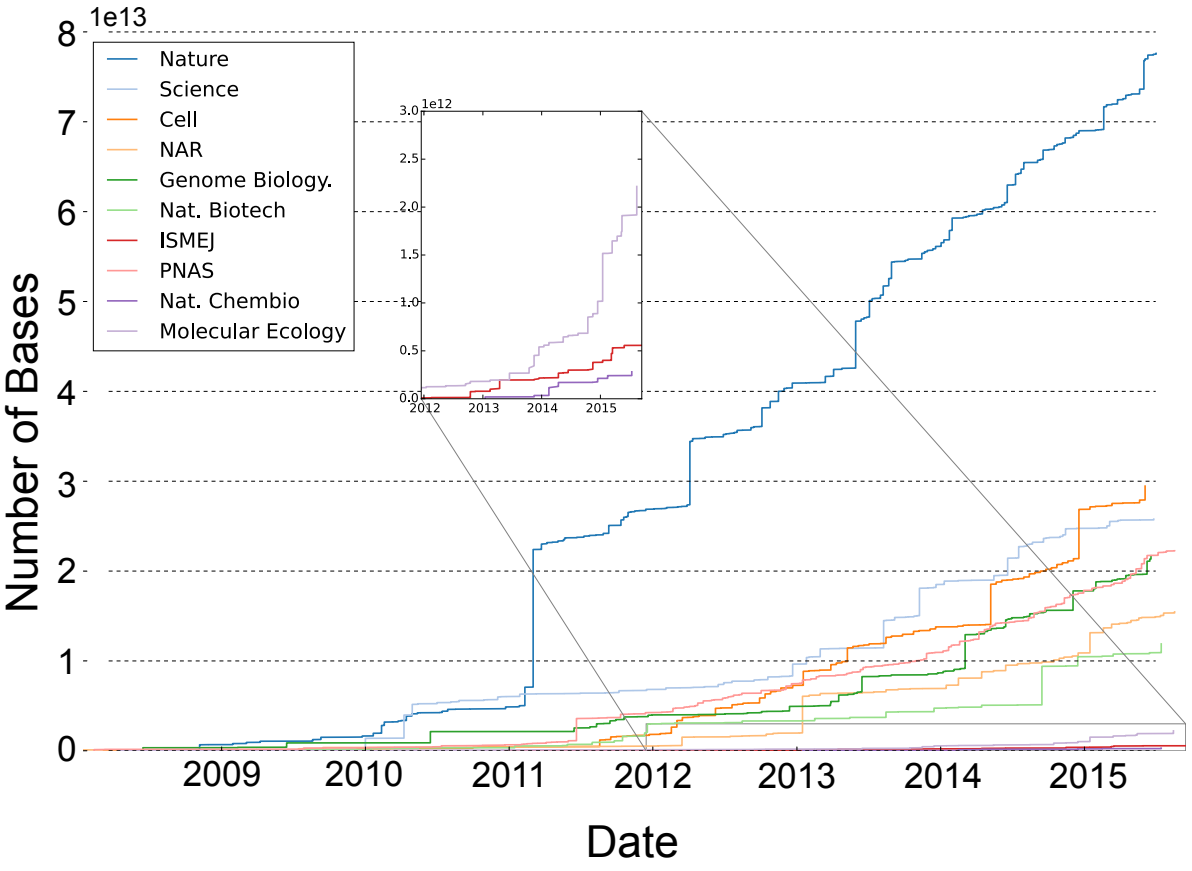


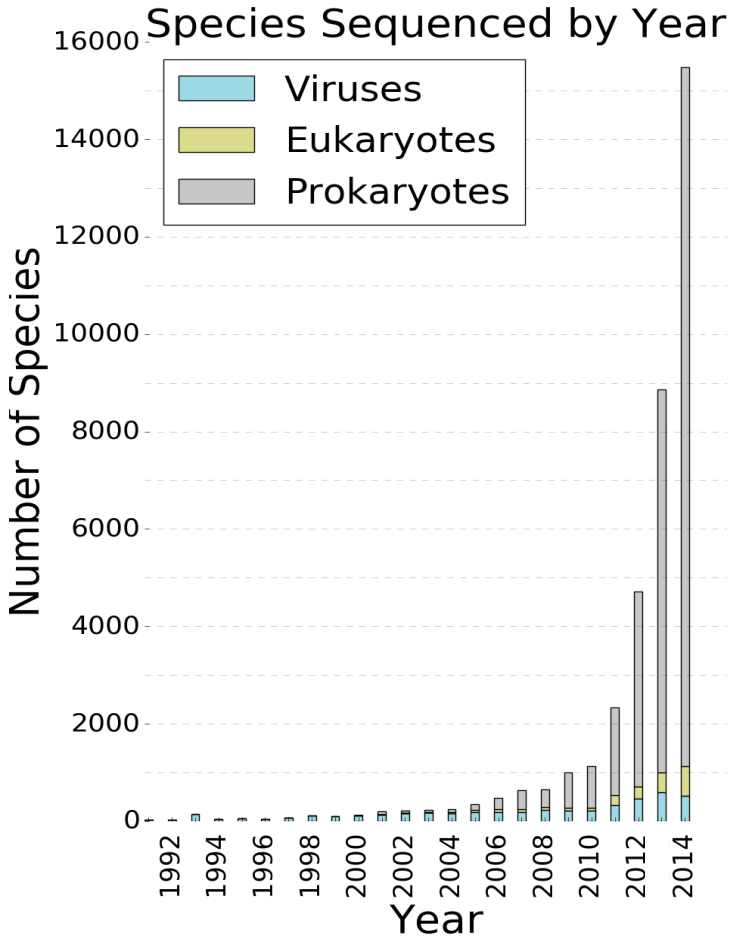
Heidi Sofia, 7-16-15

JESS3

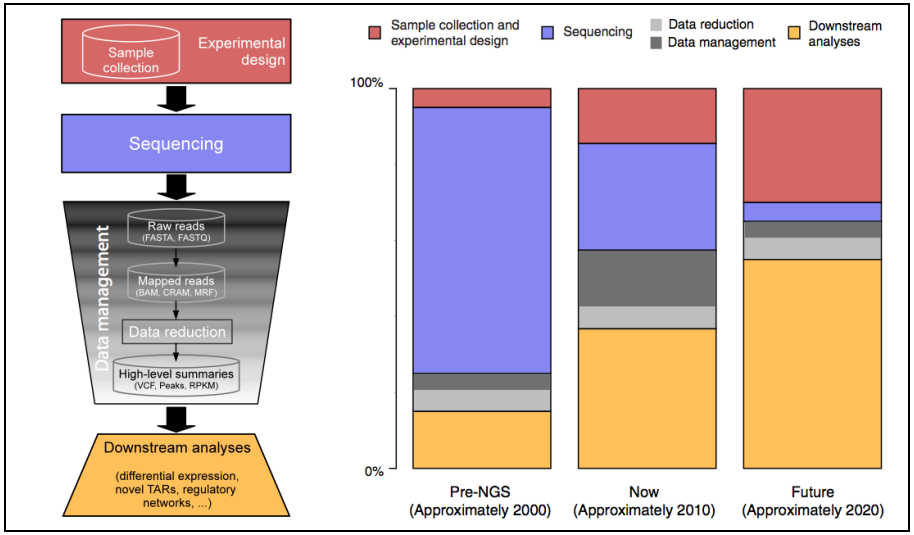
Increasing diversity in sequence data sources

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





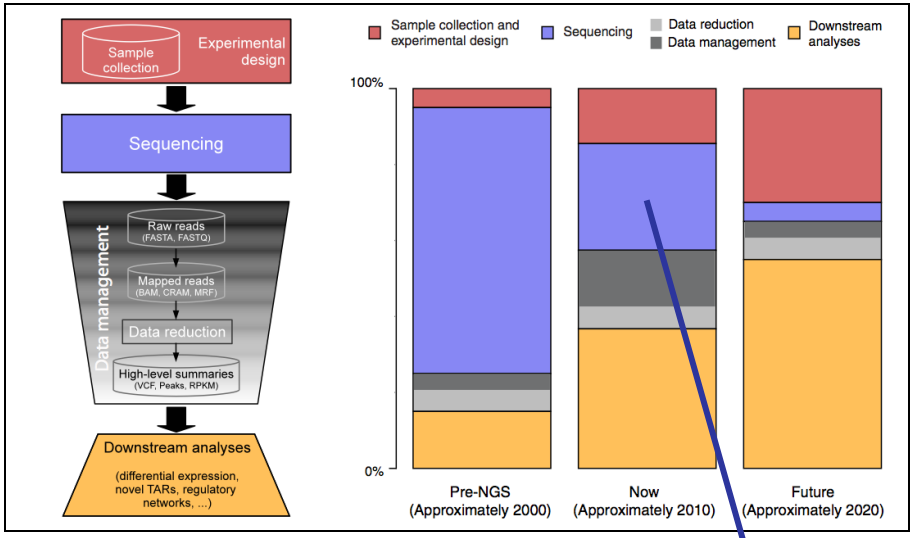
The changing costs of a sequencing pipeline



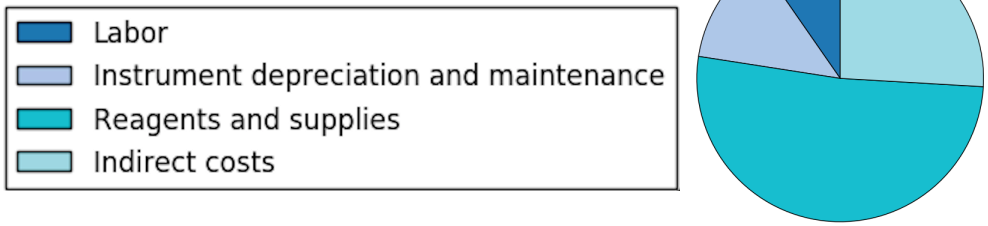
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

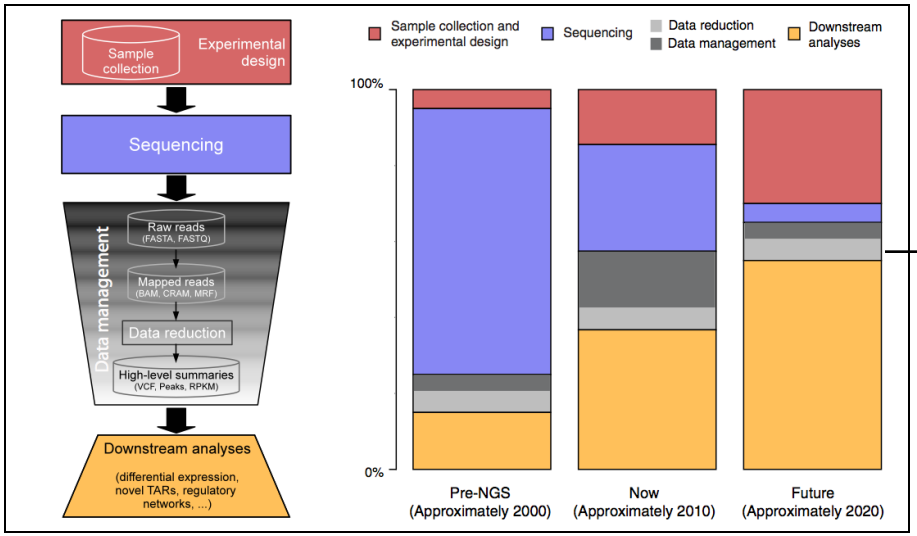


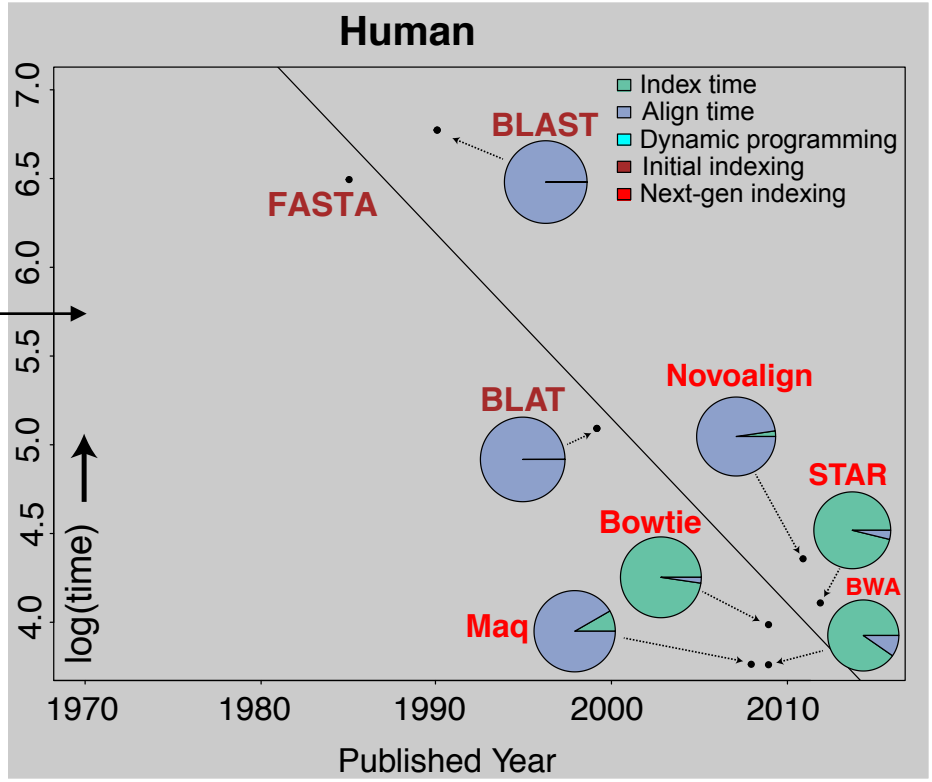
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



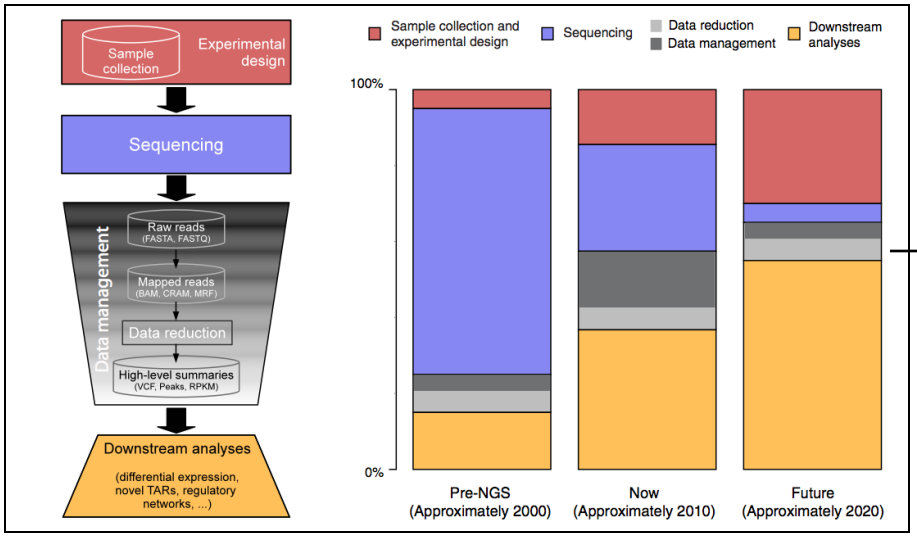


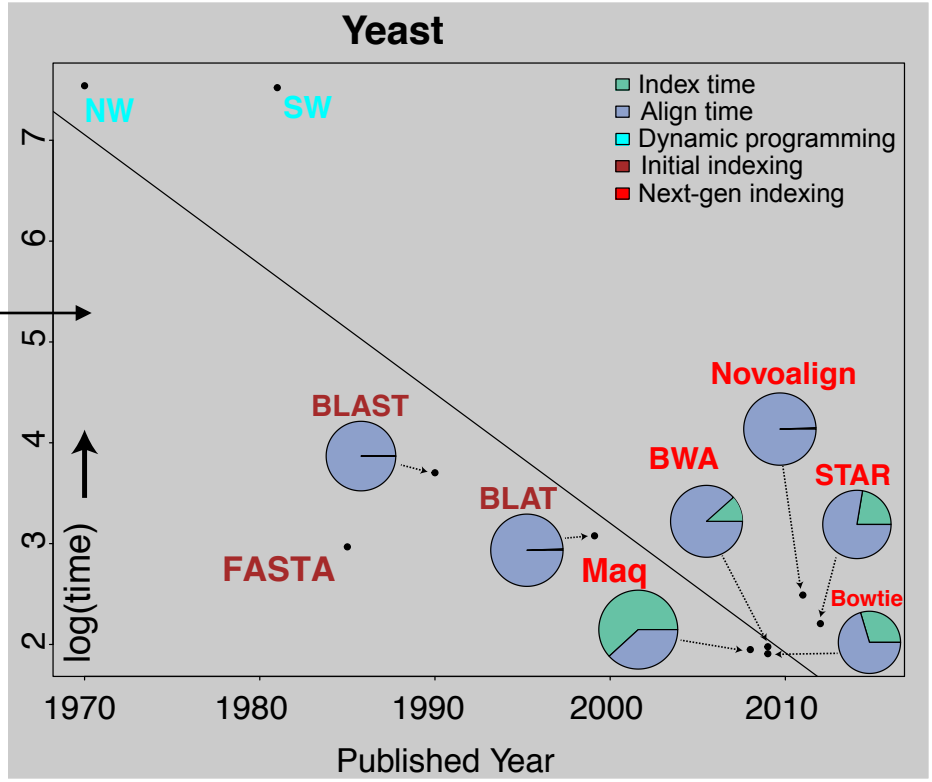
From '00 to ~' 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

The changing costs of a sequencing pipeline



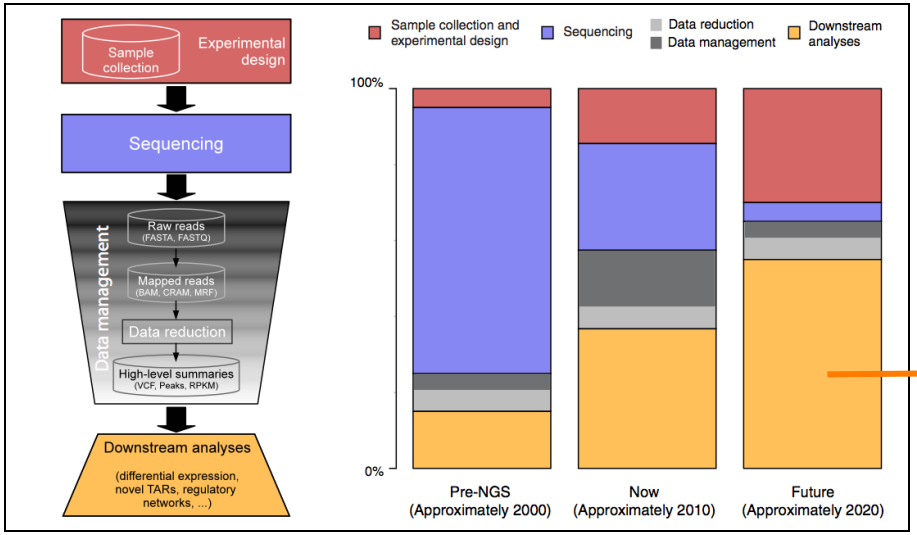


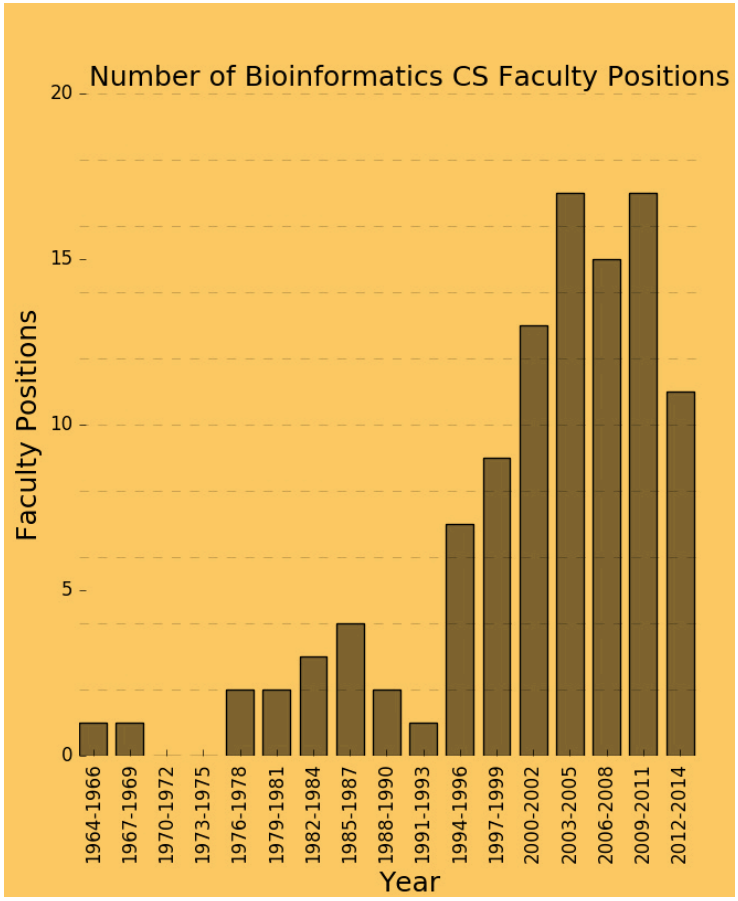
From '00 to ~' 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

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

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

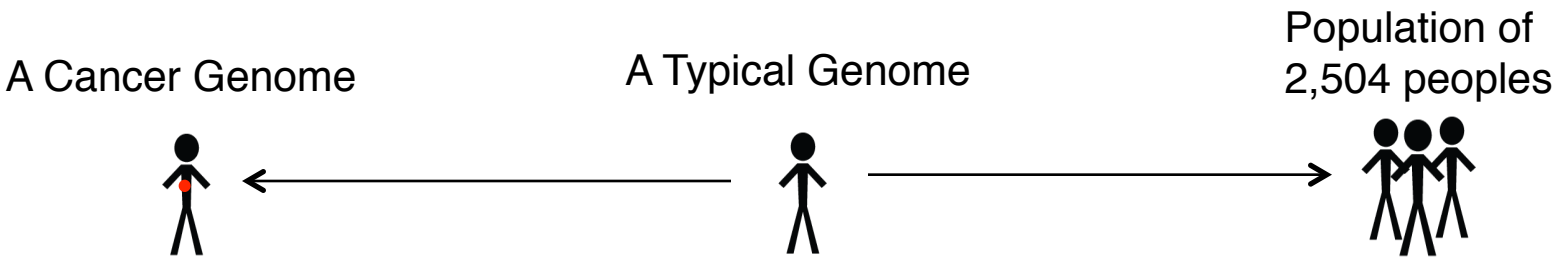
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

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

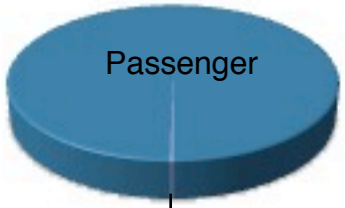
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

Human Genetic Variation



Origin of Variants

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



Class of Variants

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

Prevalence of Variants

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

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

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Finding Key Variants

Germline

CAN YOU FIND THE PANDA?

- **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.

McCarthy, M. et al. Nat. Rev. Genet. 2008. 9, 356-369, Zuk, O. et al. PNSA. 2014. Vol. 11, no. 4, MacArthur DG et al. Nature 2014. 508:469-476

Finding Key Variants

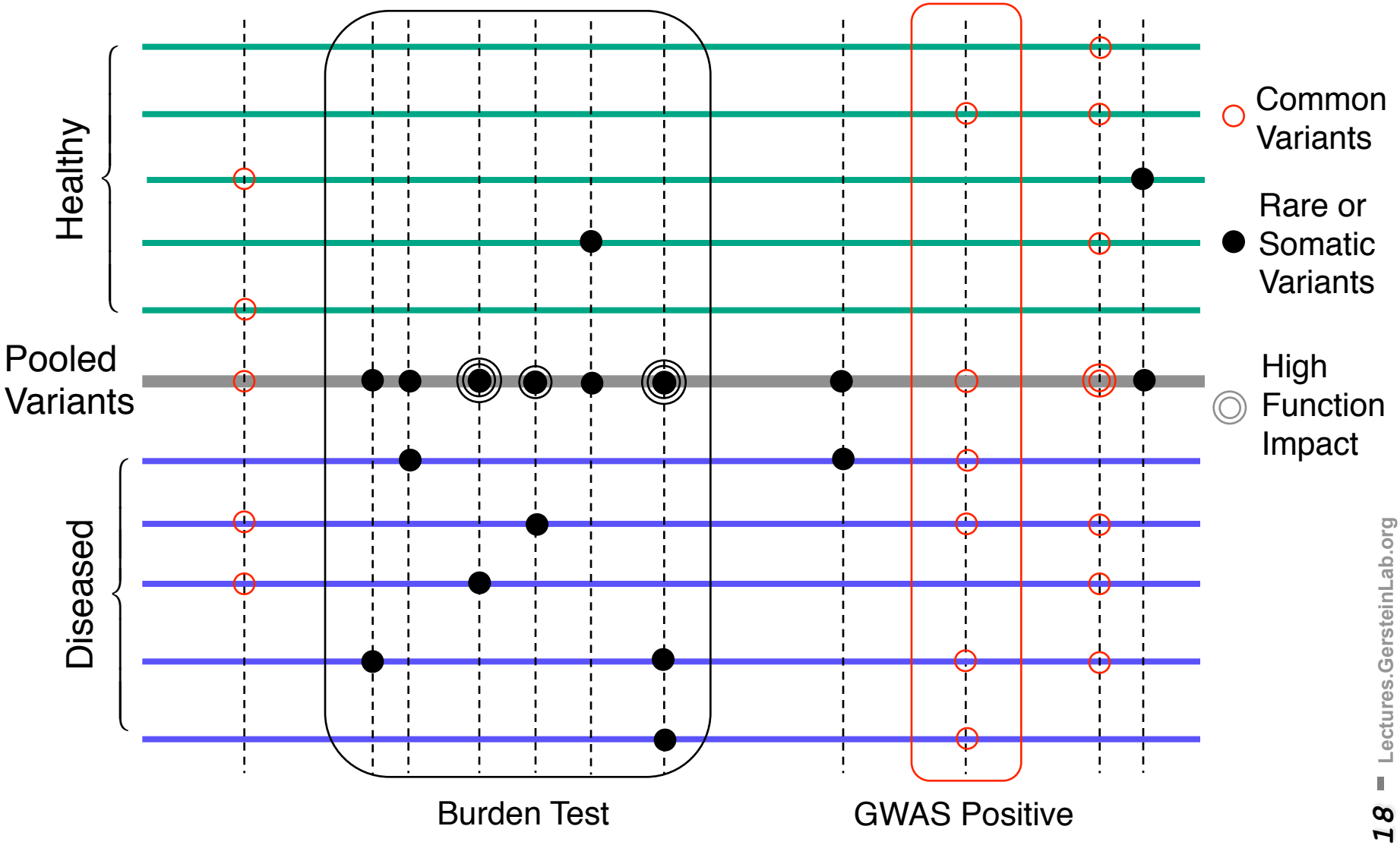
Somatic

CAN YOU FIND THE PANDA?

- **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.

Vogelstein B. Science 2013. 339(6127):1546-1558

Association of Variants with Diseases

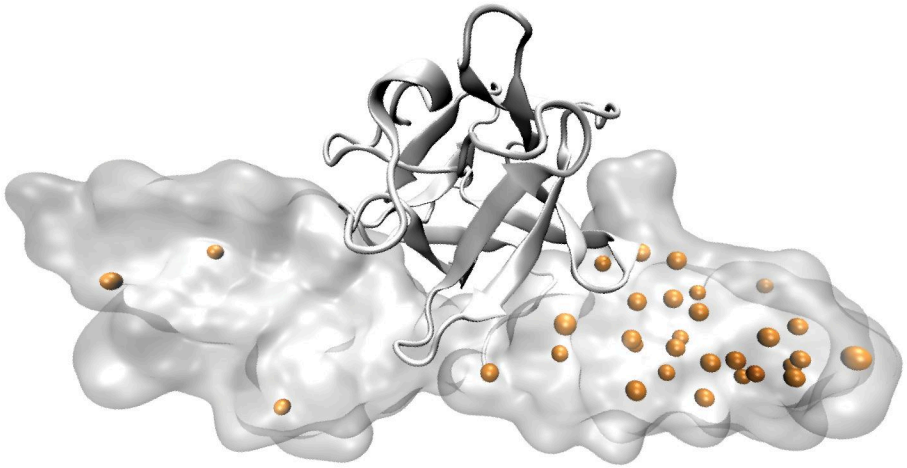


Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

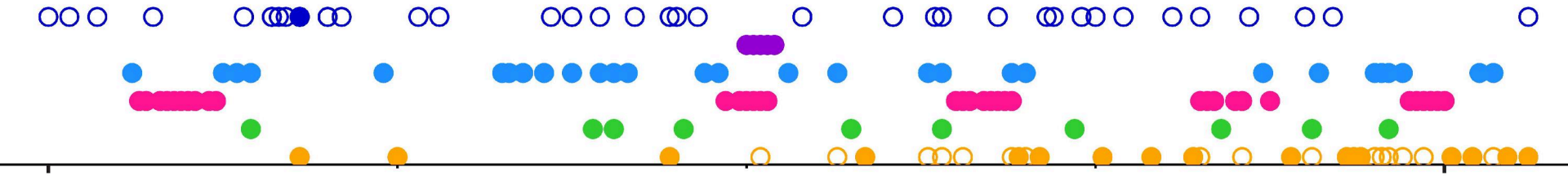
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

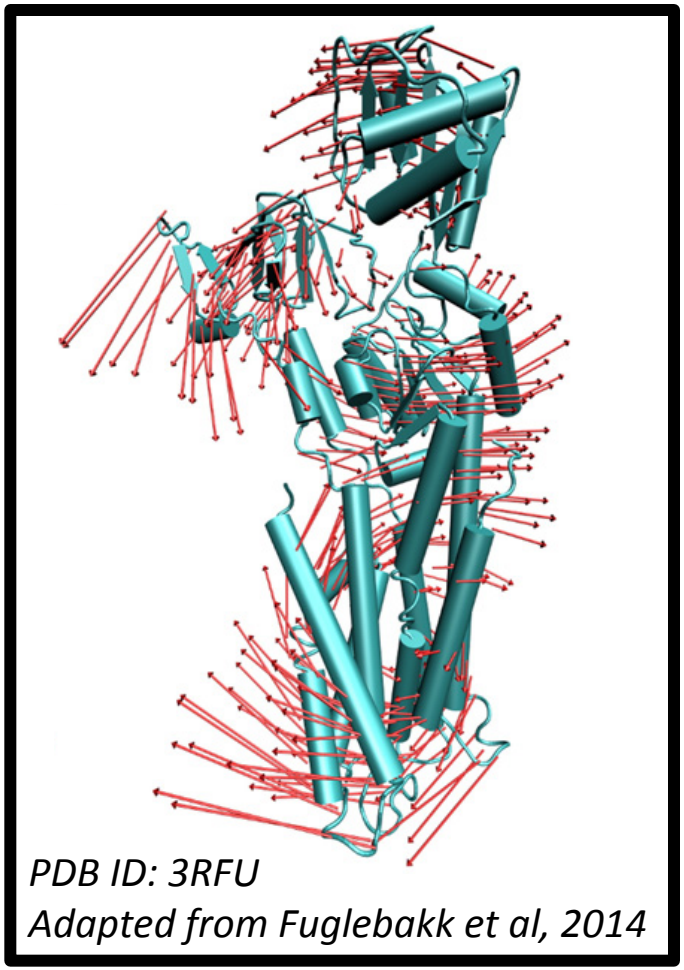
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



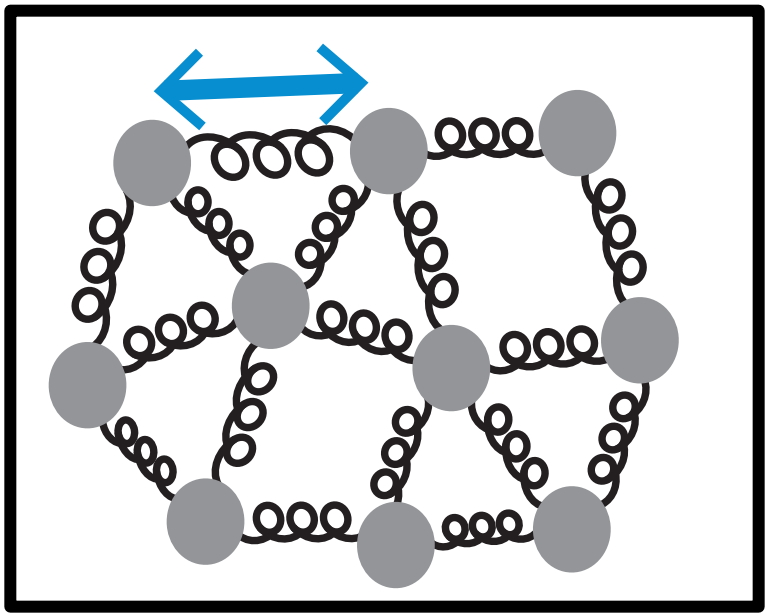
[Sethi et al. COSB ('15)]

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



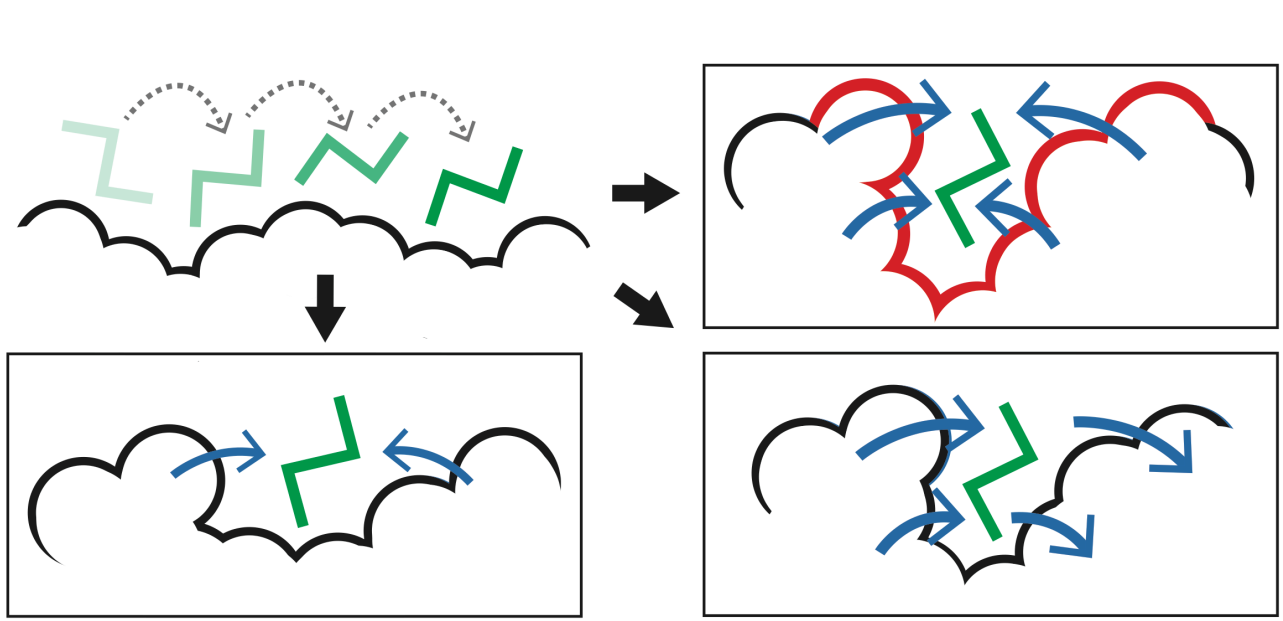
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



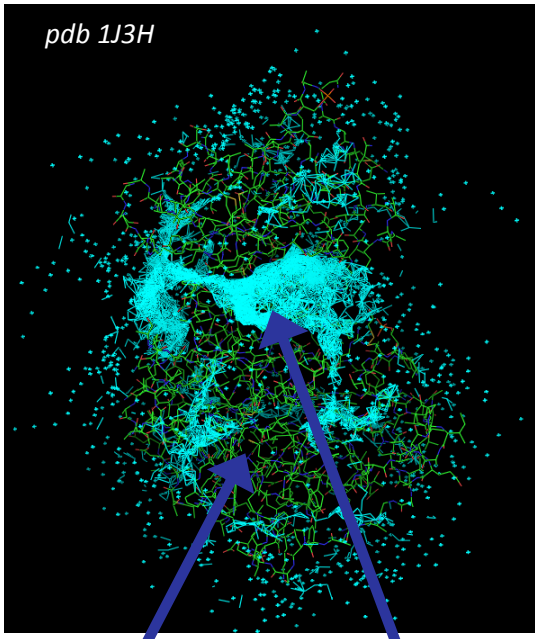
Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion

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

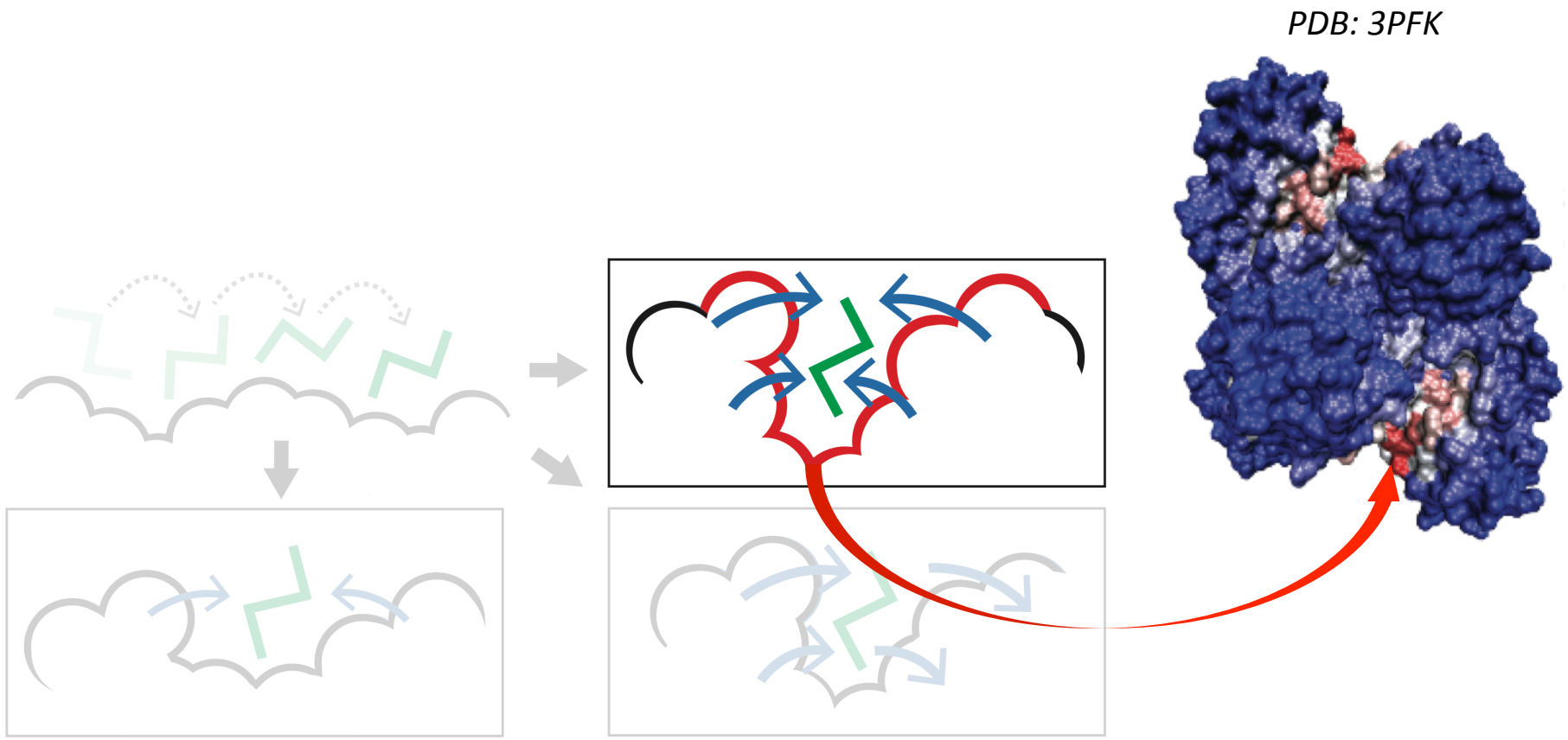


$$\textit{binding leverage} = \sum_{m=1}^{10}\left(\sum_{i}\sum_{j}\Delta d_{ij(m)}^{2}\right)$$



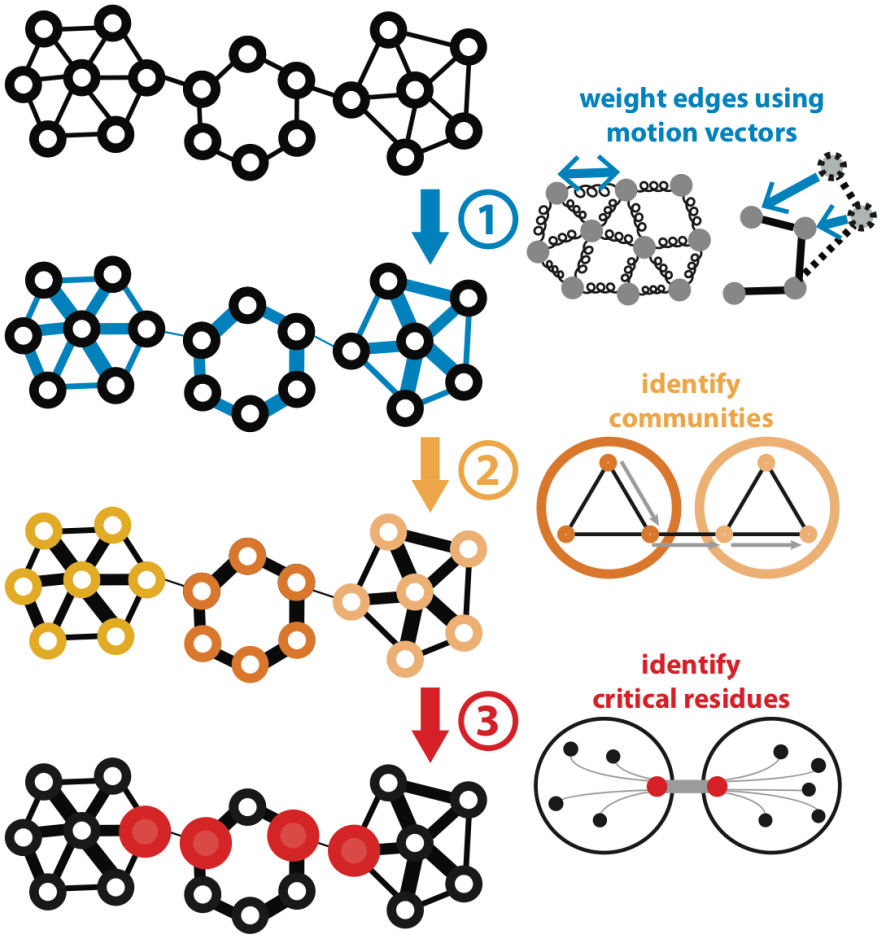
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface



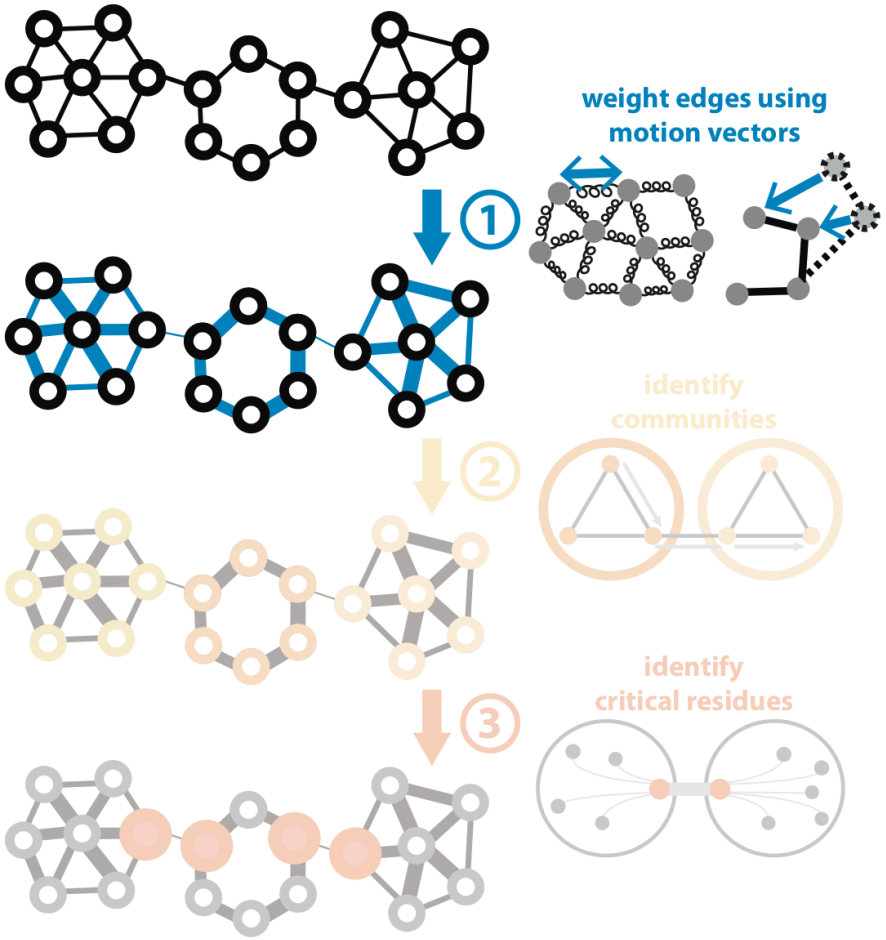
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



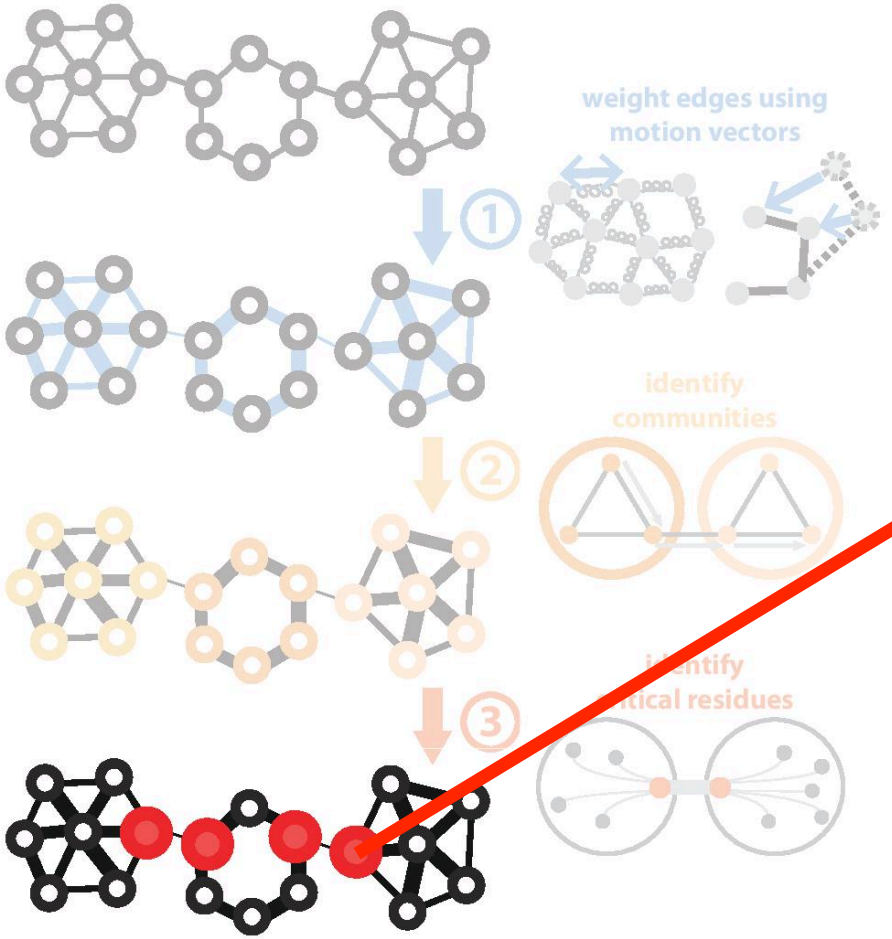
$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

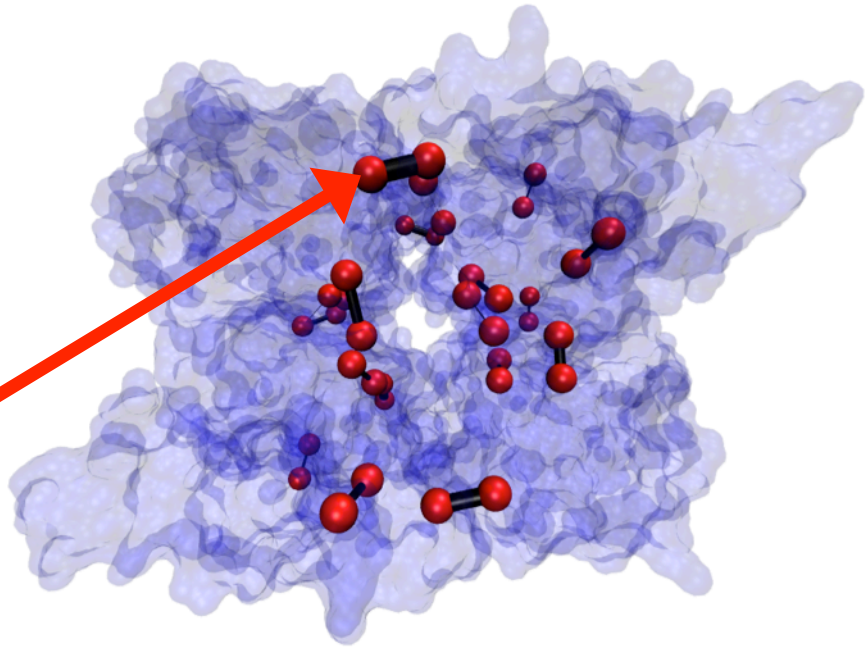
$$C_{ij} = Cov_{ij} \; / \; \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

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

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

Predicting Allosterically-Important Residues within the Interior



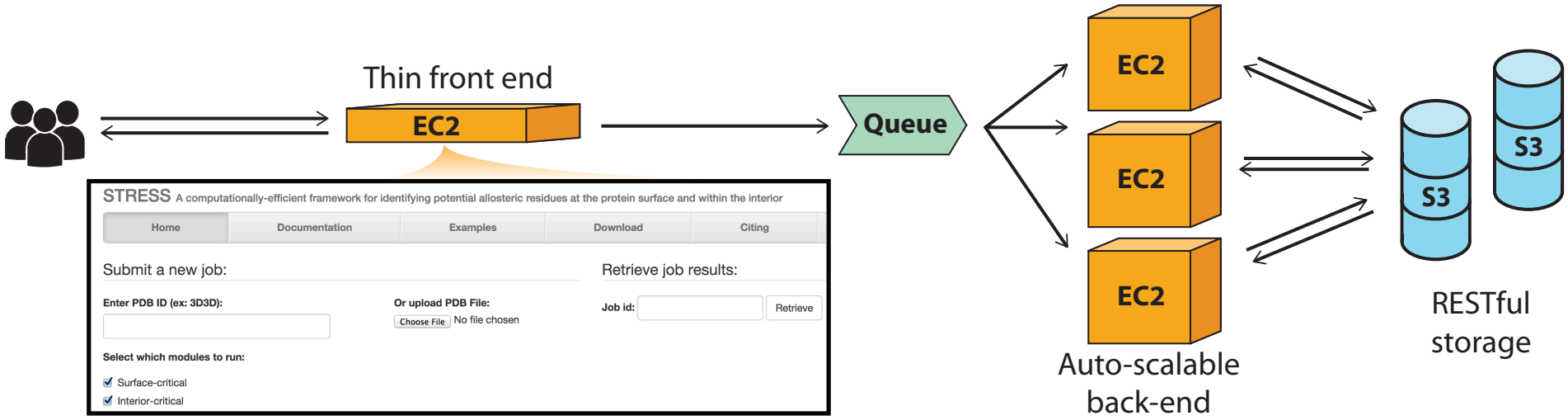


PDB: 1XTT

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

STRESS Server Architecture: Highlights

stress.molmovdb.org

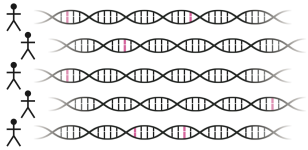


- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

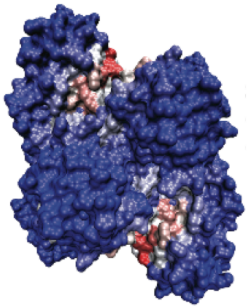
Adapted from Clarke*, Sethi*, et al *(in press)*

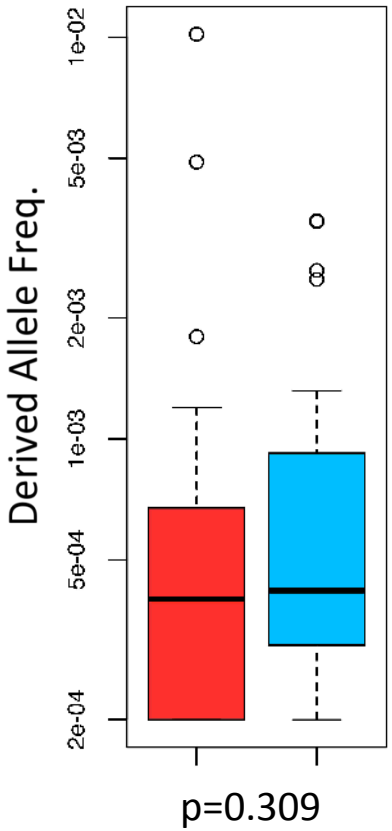
Intra-species conservation of predicted allosteric residues

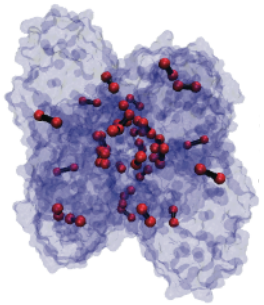
1000 Genomes



Surface



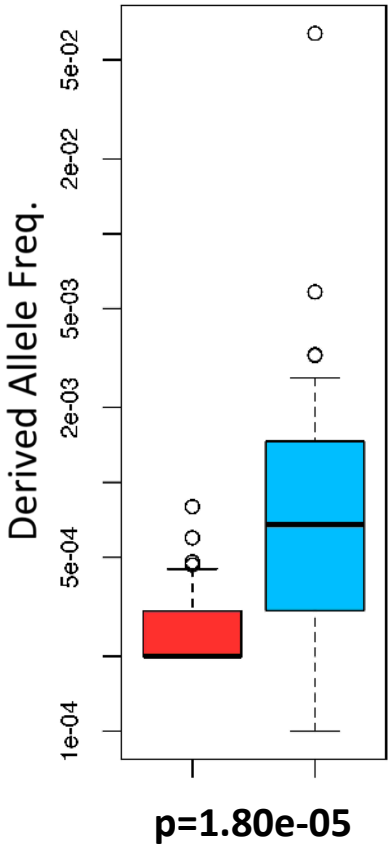




critical

non-critical

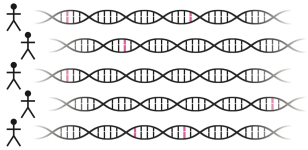
Interior

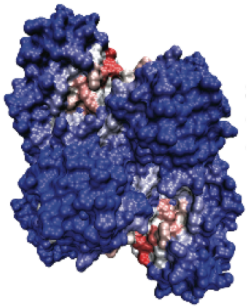


Adapted from Clarke*, Sethi*, et al *(in press)*

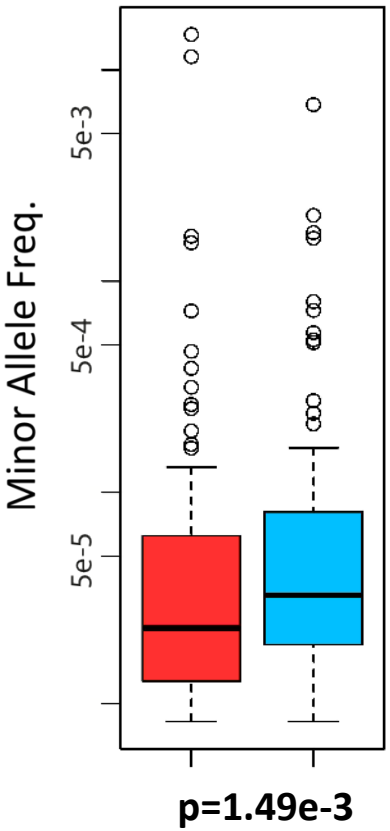
Intra-species conservation of predicted allosteric residues

ExAC

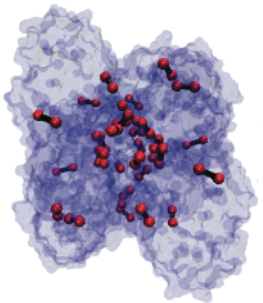




Surface



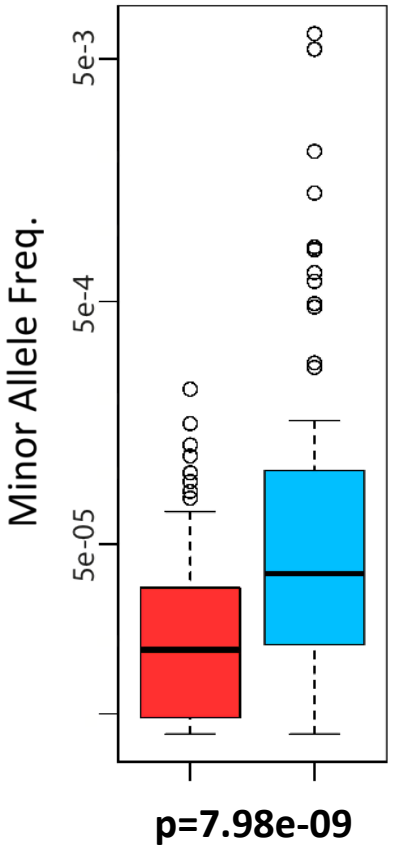
p=1.49e-3



critical

non-critical

Interior

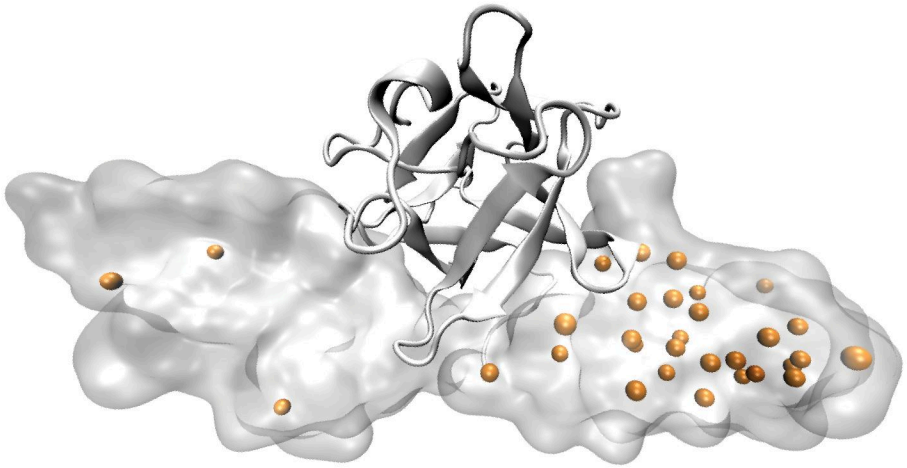


p=7.98e-09

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

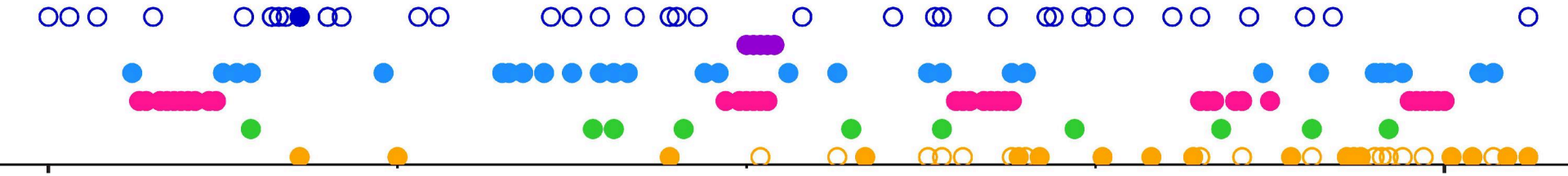
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

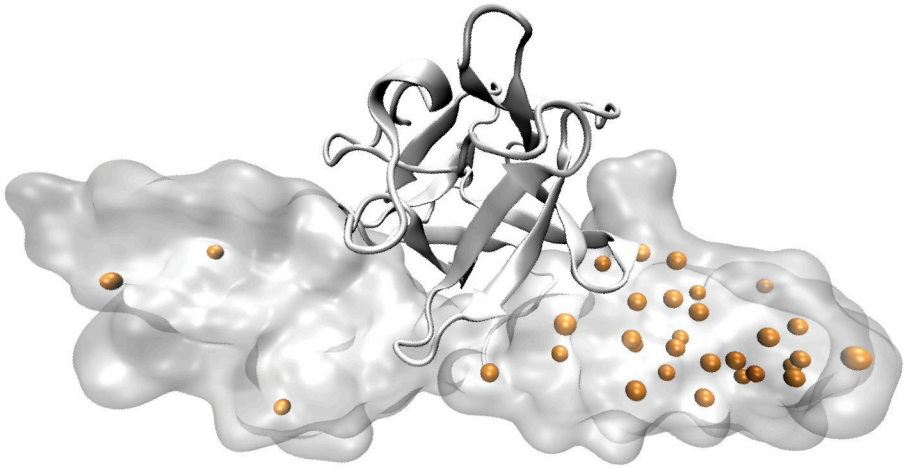
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

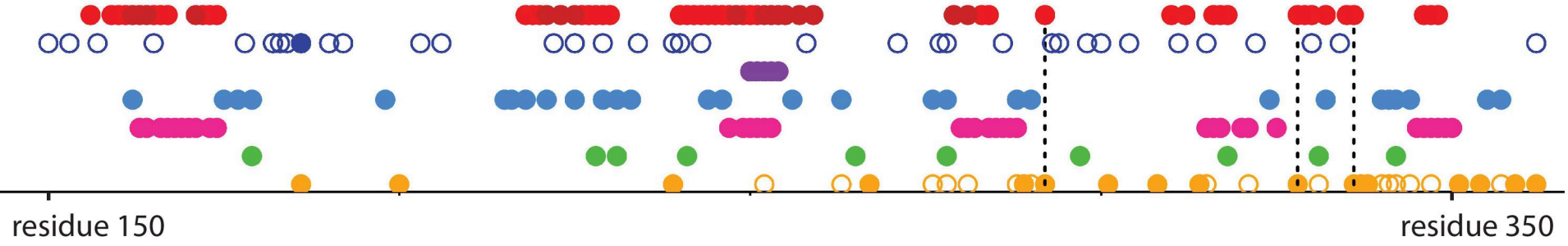
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

- 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)



[Sethi et al. COSB ('15)]

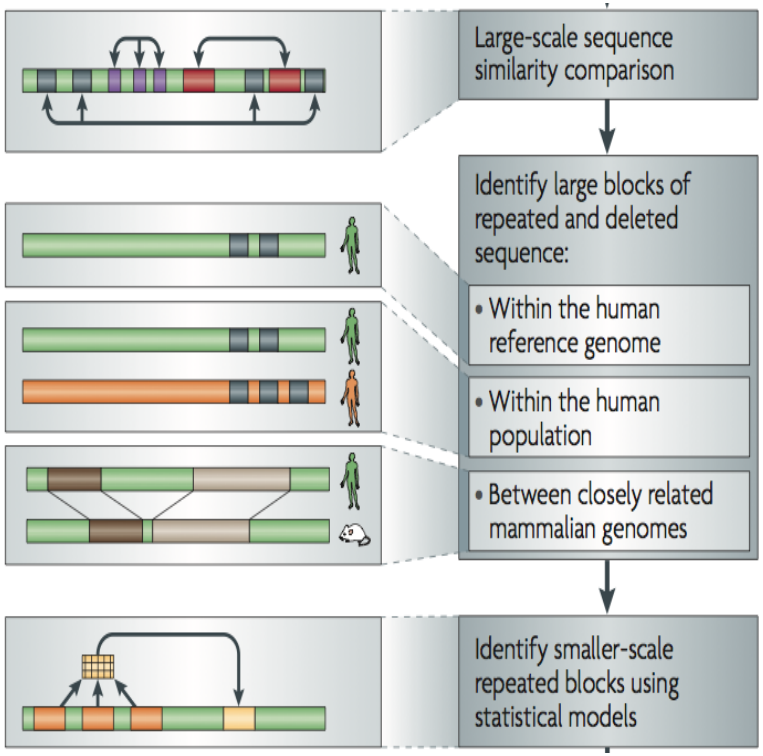
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

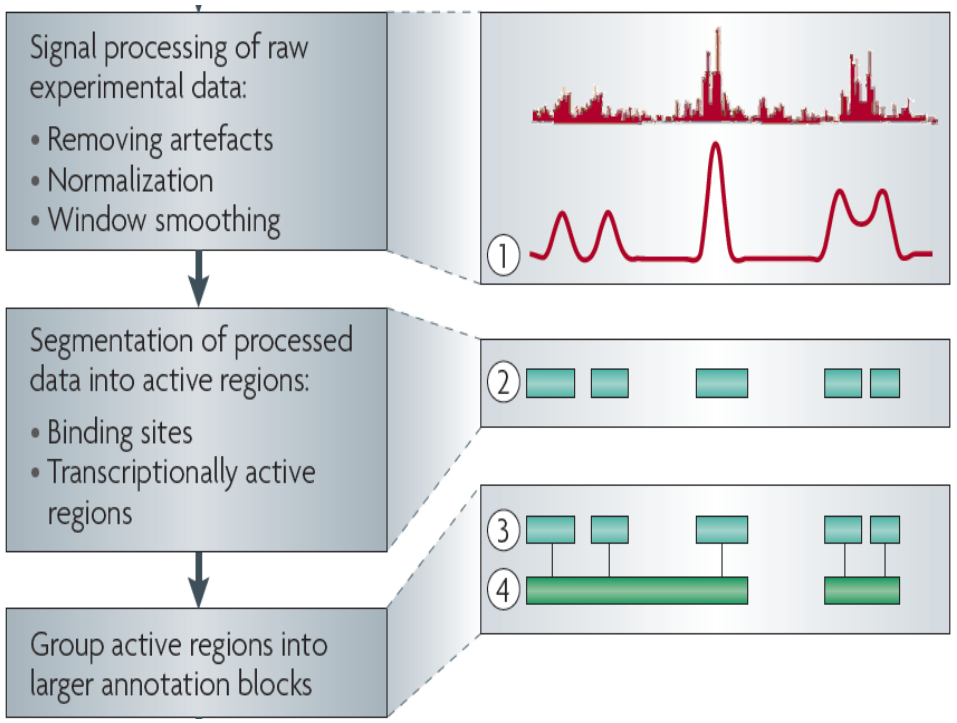
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

Non-coding Annotations: Overview

Sequence features, incl. **Conservation**

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

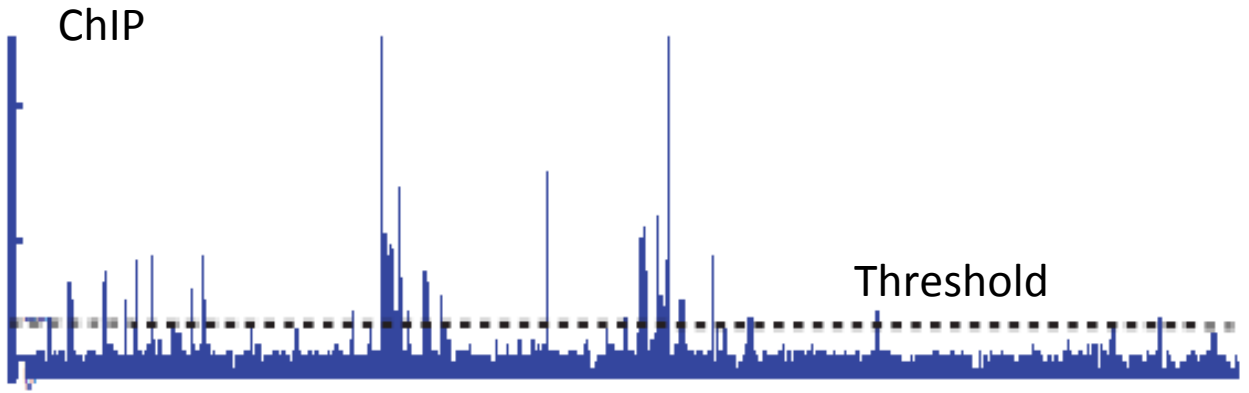




[Alexander et al., *Nat. Rev. Genet.* ('10)]

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 →

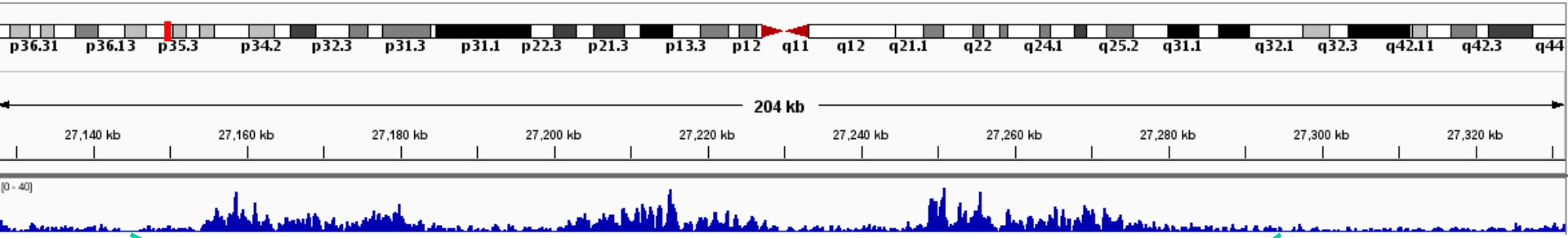
- Score against the control

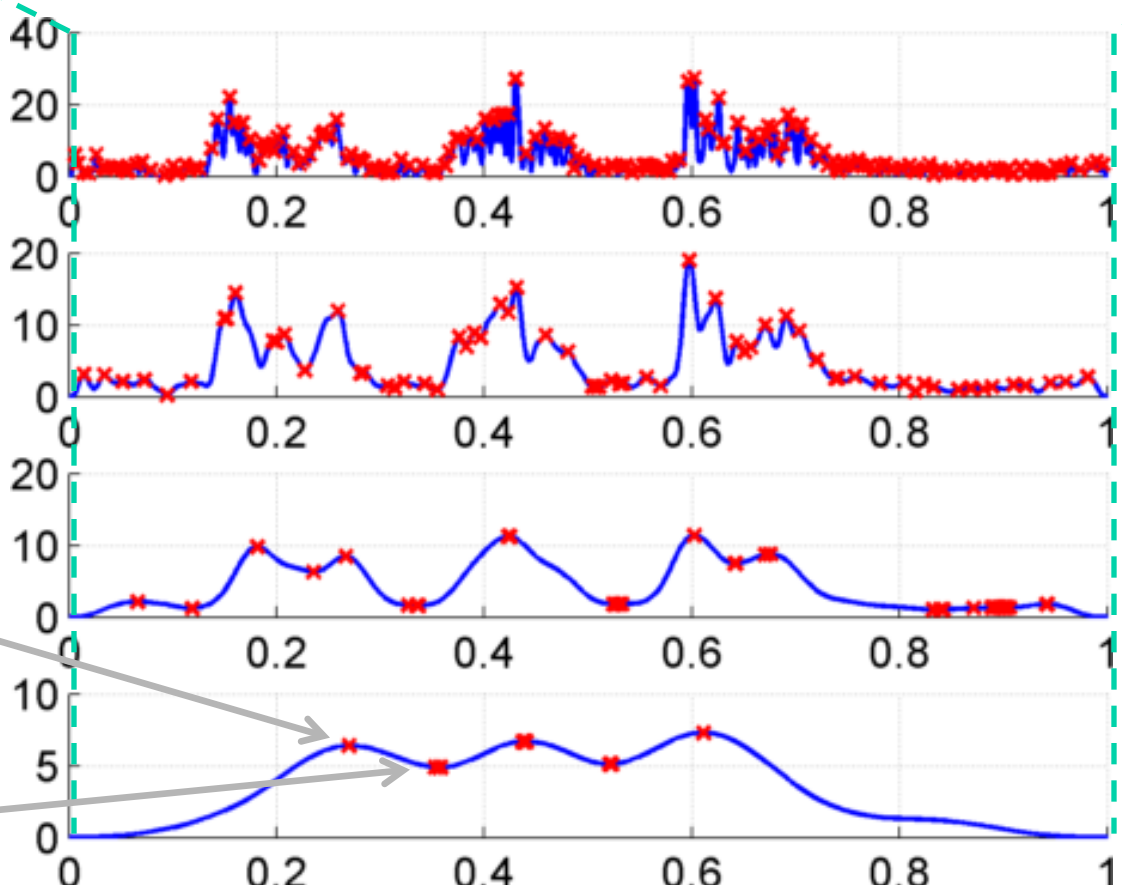
Significantly Enriched targets →

Now an update: "PeakSeq 2" => MUSIC

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

Multiscale Analysis, Minima/Maxima based Coarse Segmentation





Window Length

1kb

4kb

16kb

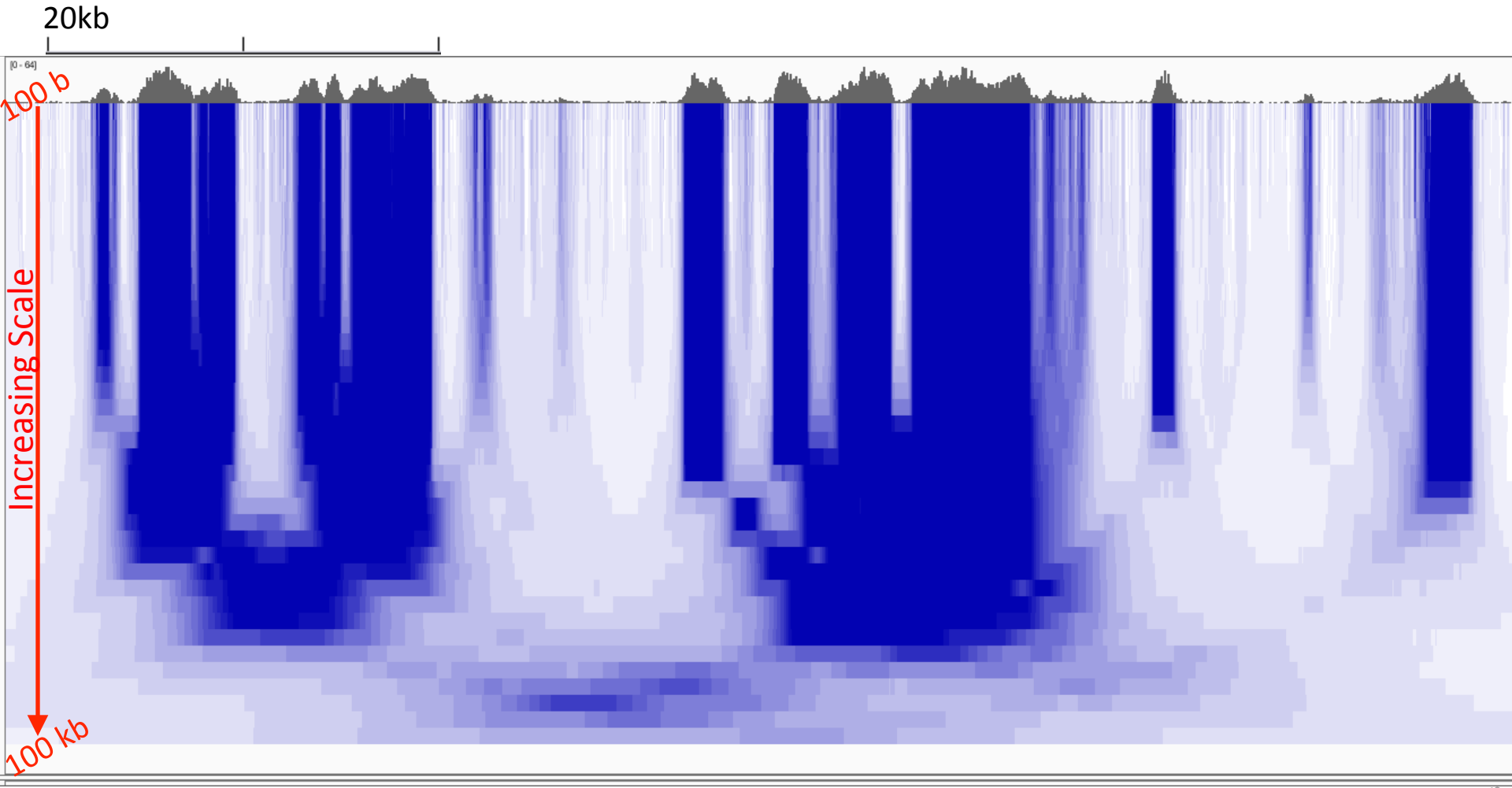
64kb

Maxima

Minima

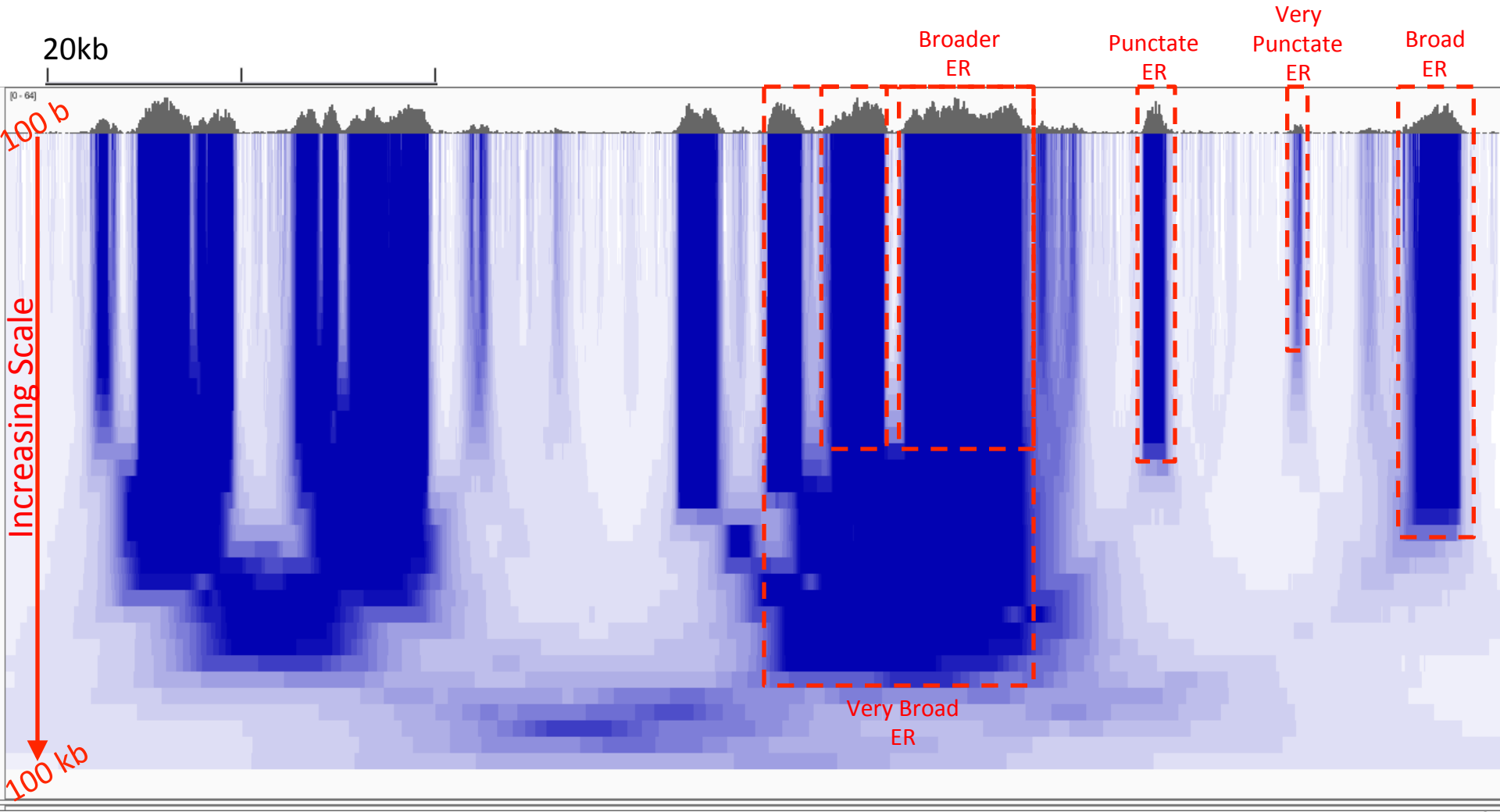
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



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

Multiscale Decomposition



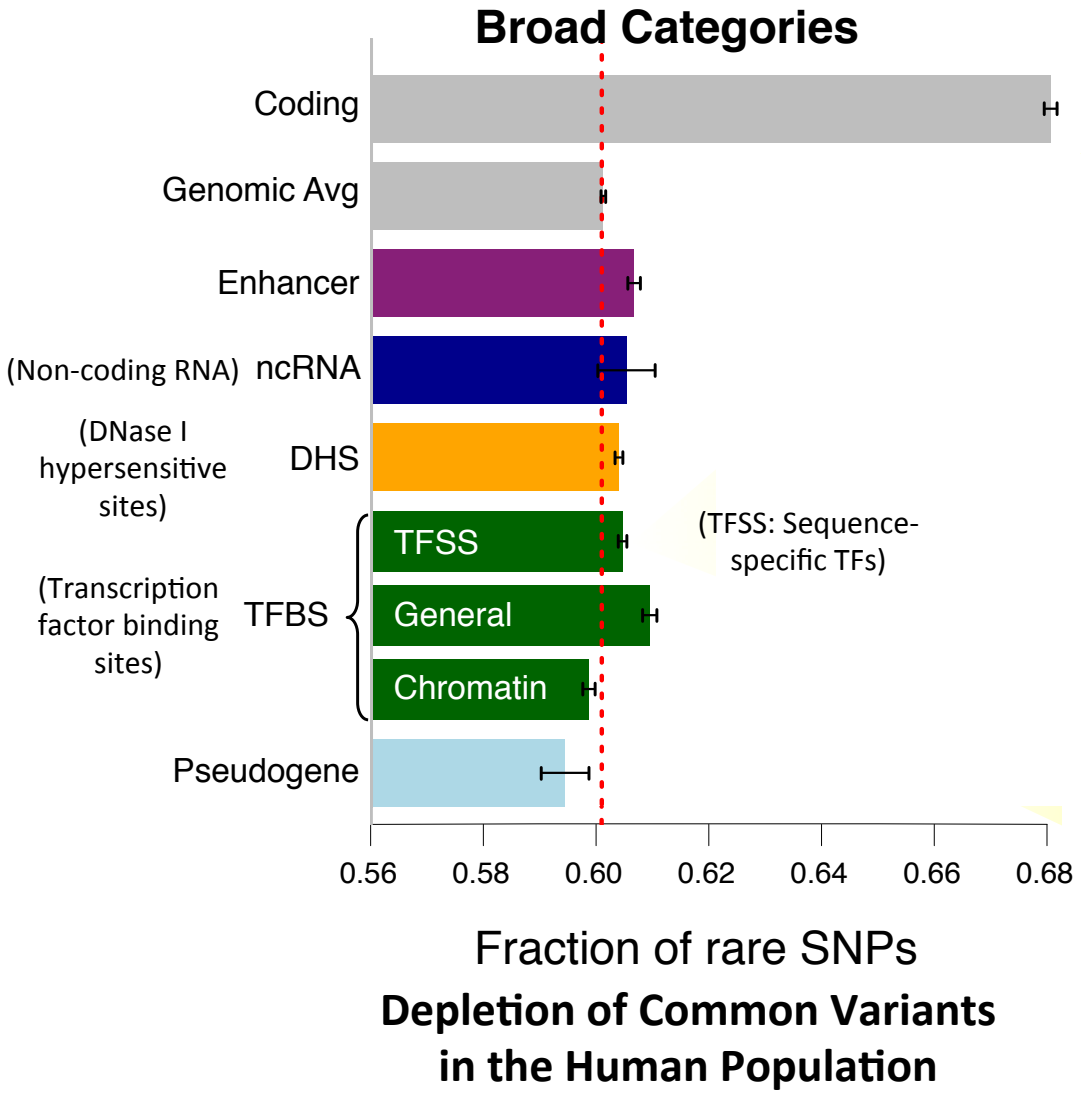
[Harmanci *et al, Genome Biol.* ('14)]

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 finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

Finding "Conserved" Sites in the Human Population:

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

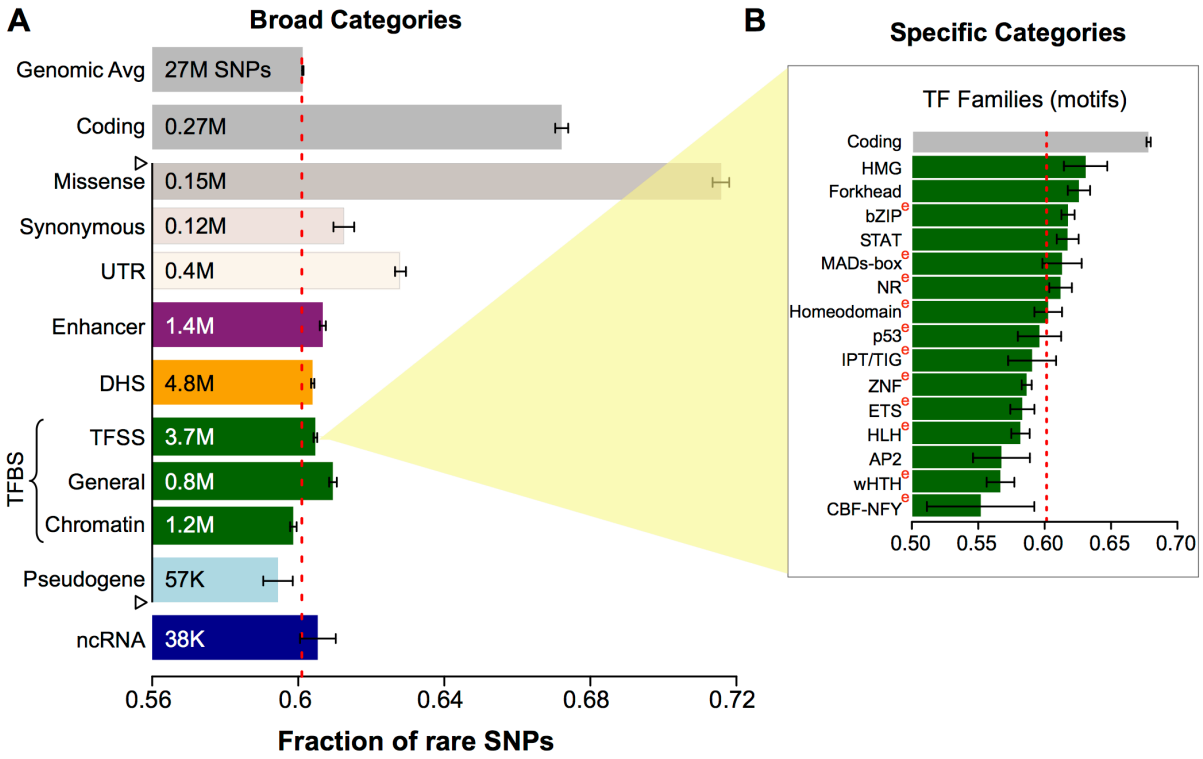


- Broad categories of regulatory regions under negative selection
- Related to:

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

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

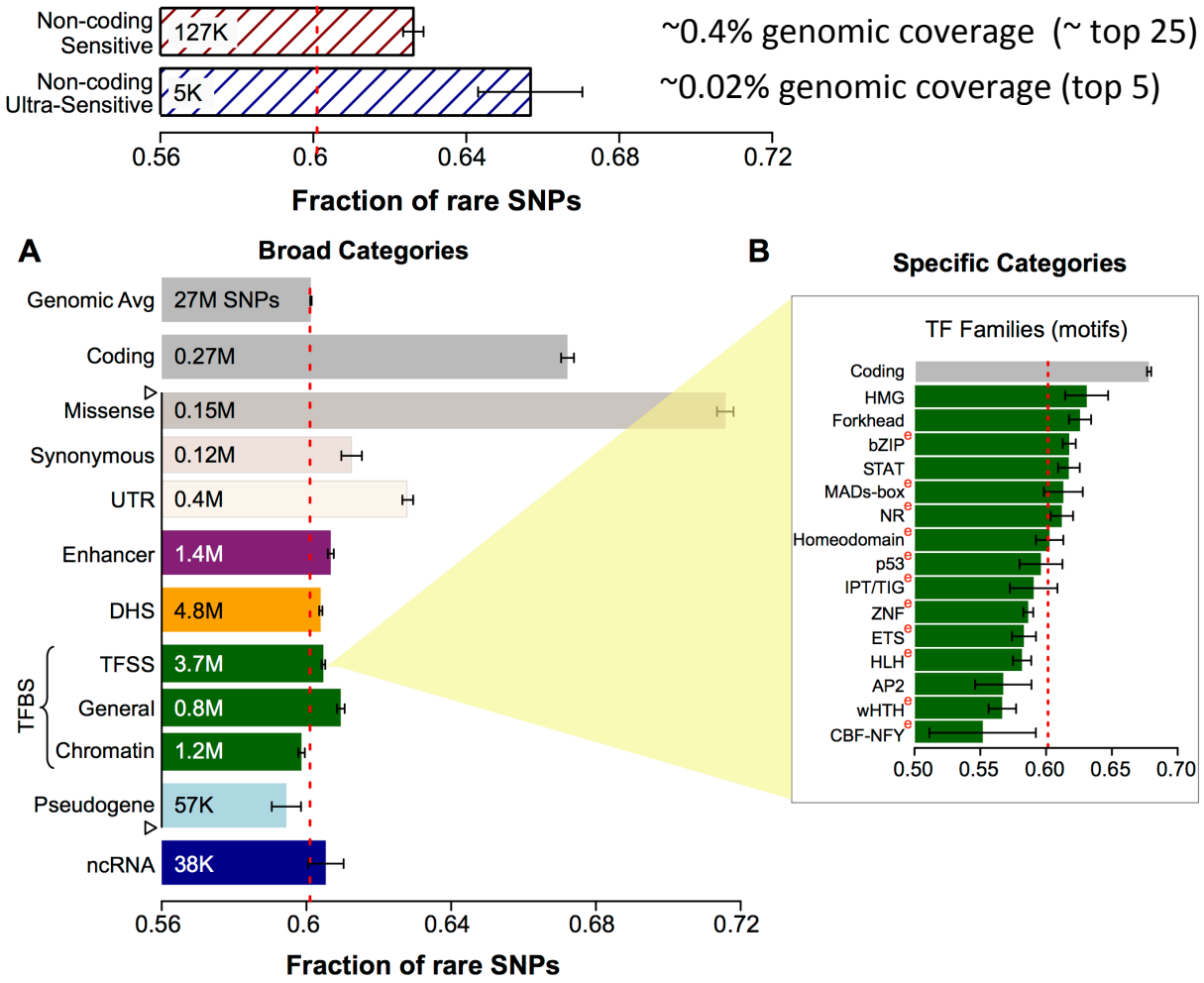
Differential selective constraints among specific sub-categories



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

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

Defining Sensitive non-coding Regions



~0.4% genomic coverage (~ top 25)

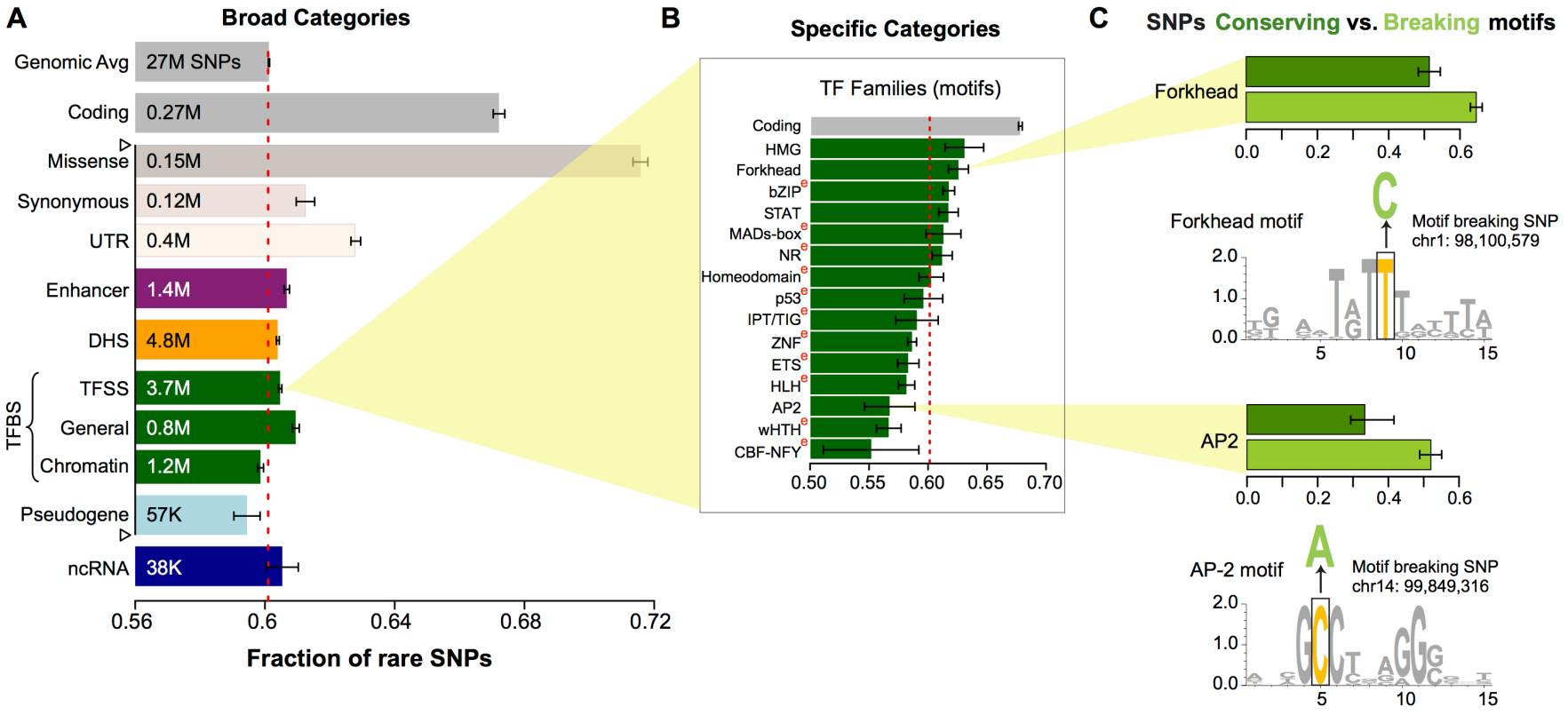
~0.02% genomic coverage (top 5)

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

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

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

SNPs which break TF motifs are under stronger selection



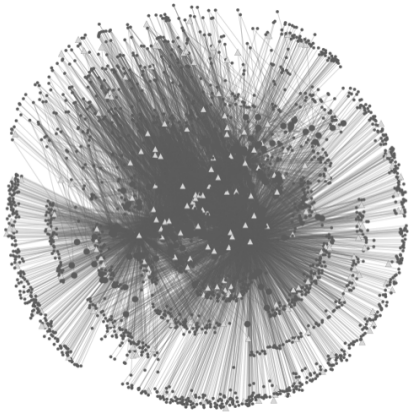
[Khurana et al., *Science* ('13)]

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 finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

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

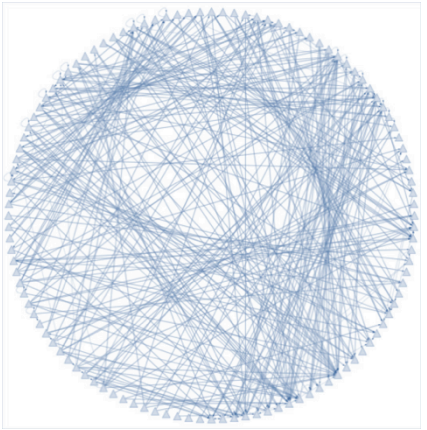
[Cheng et al., *Bioinfo.* ('11),
Gerstein et al., *Nature* ('12) ,
Yip et al., *GenomeBiology* ('12),
Fu et al., *GenomeBiology*('14)]



~500K Prox. Edges

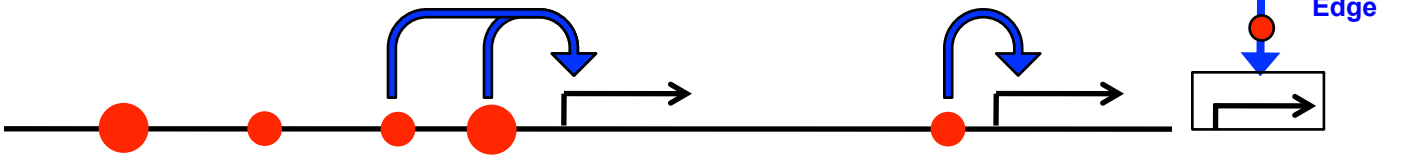
Filtering

~26K

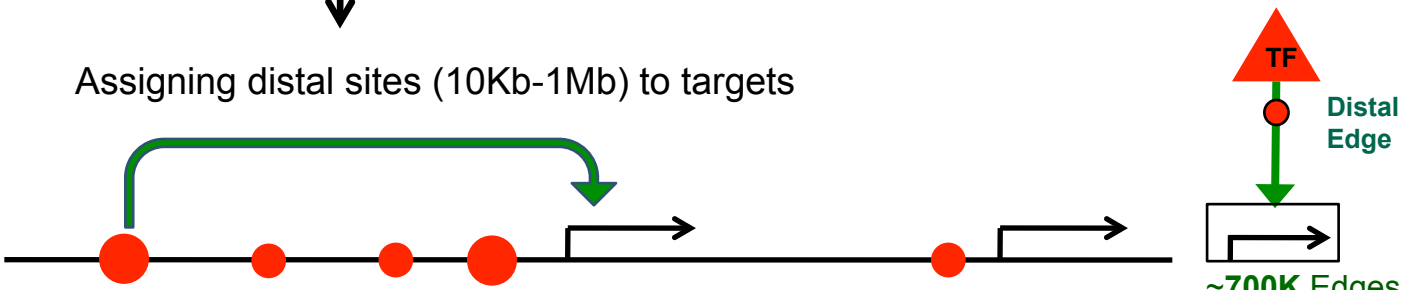


Regulatory elements

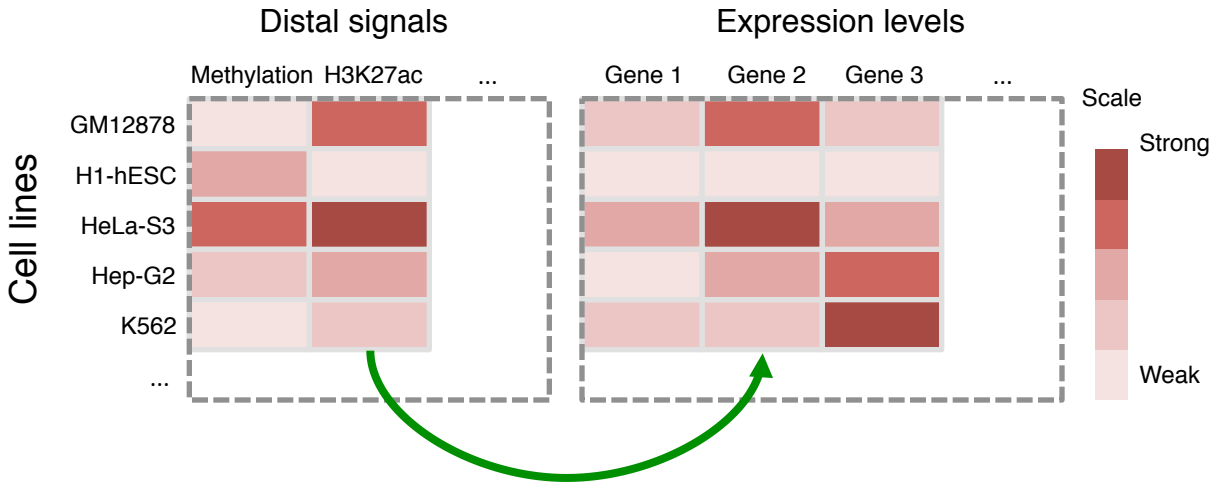
Assigning proximal sites (< 1Kb) to target genes



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



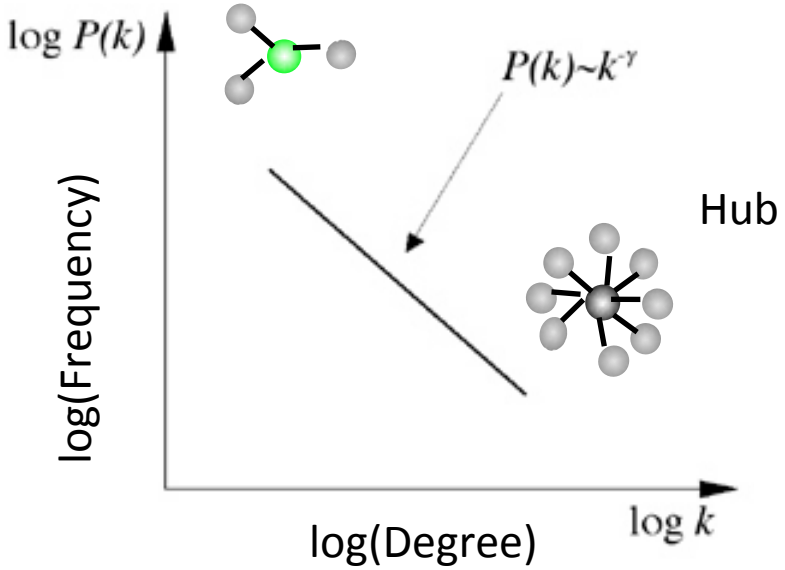
~700K Edges



Connecting Distal Elements via **Activity Correlations**.

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

Power-law distribution

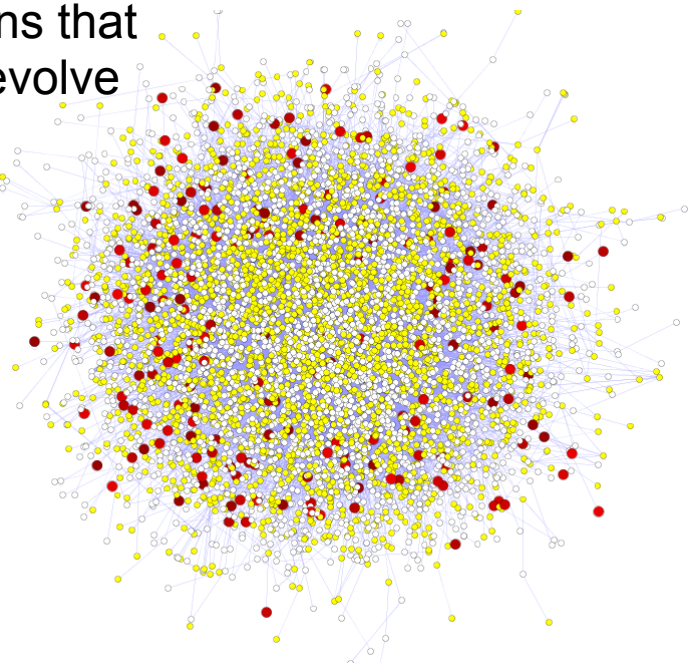


Hubs Under Constraint: A Finding from the Network Biology Community

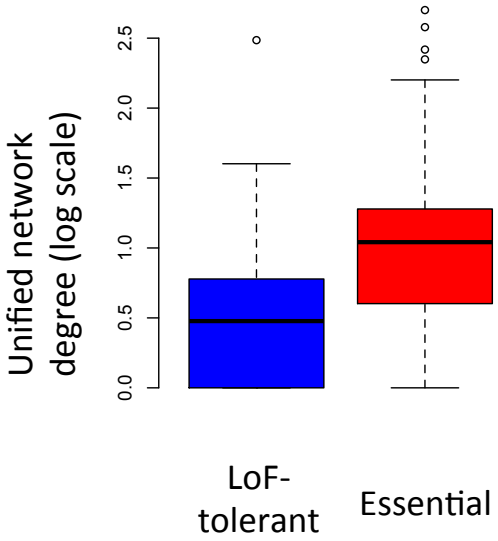
- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

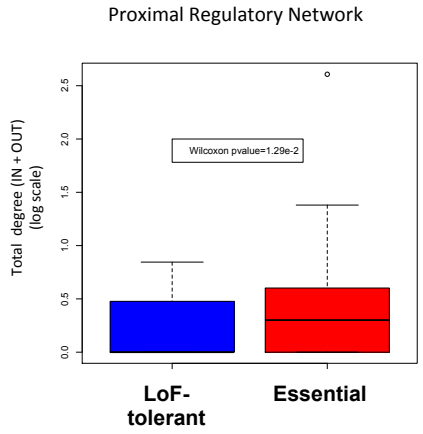
[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

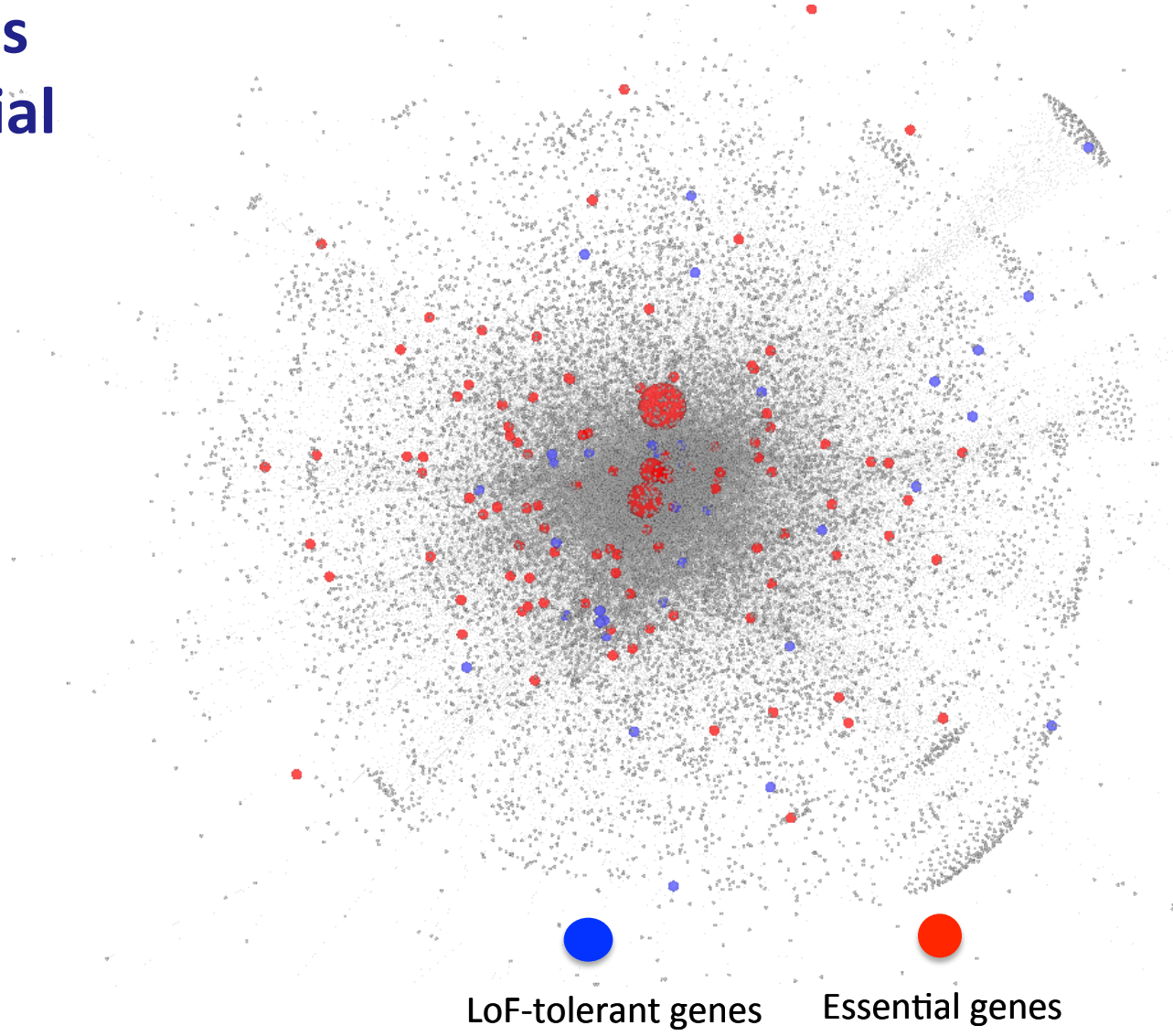
- More Connectivity, More Constraint: Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
  - **yeast PPI** - Fraser et al ('02) Science, ('03) 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







Z Gumus iCAVE movie

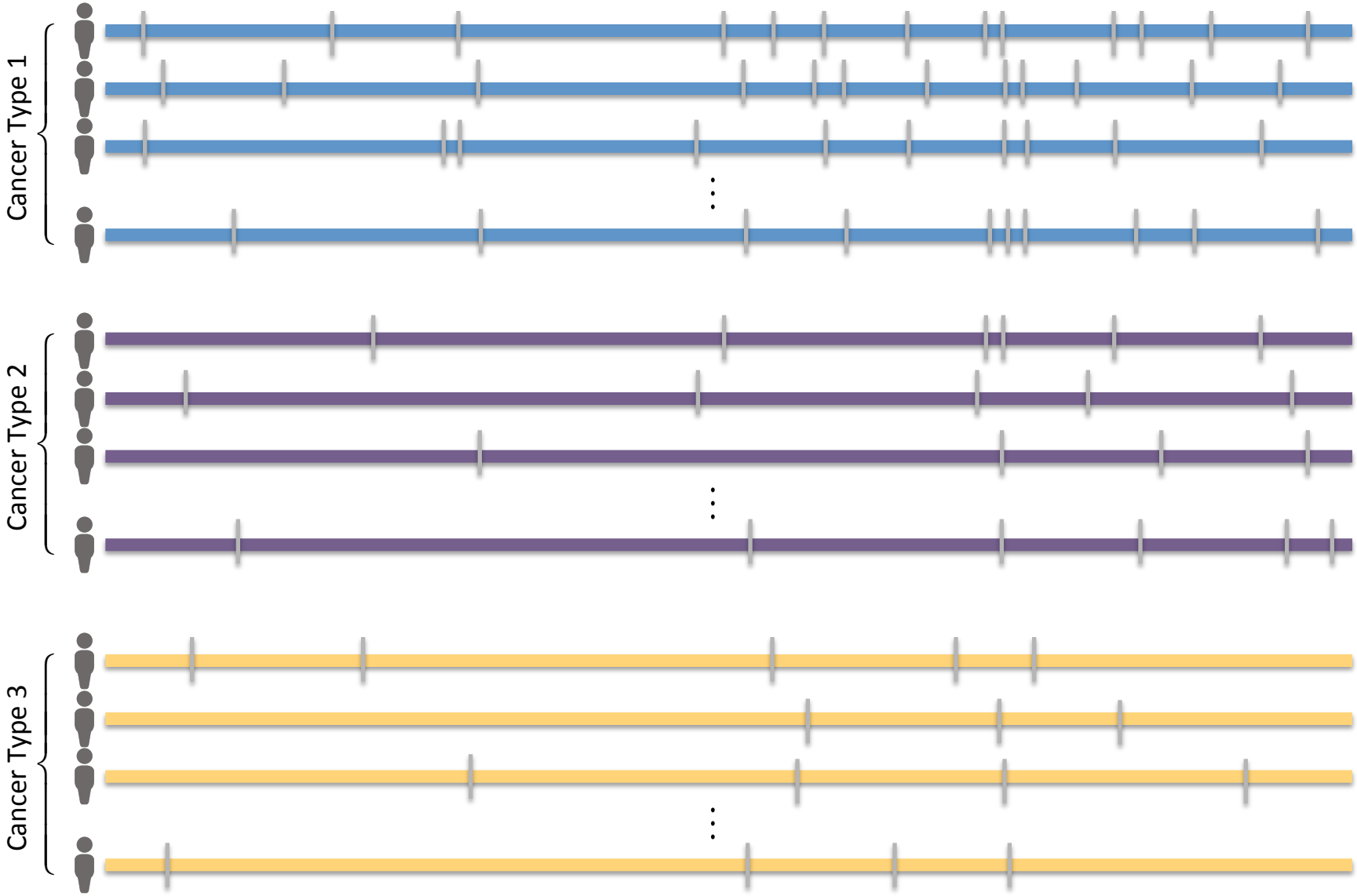
[Khurana et al., *PLOS Comp. Bio.* '13]

Size of nodes scaled by total degree

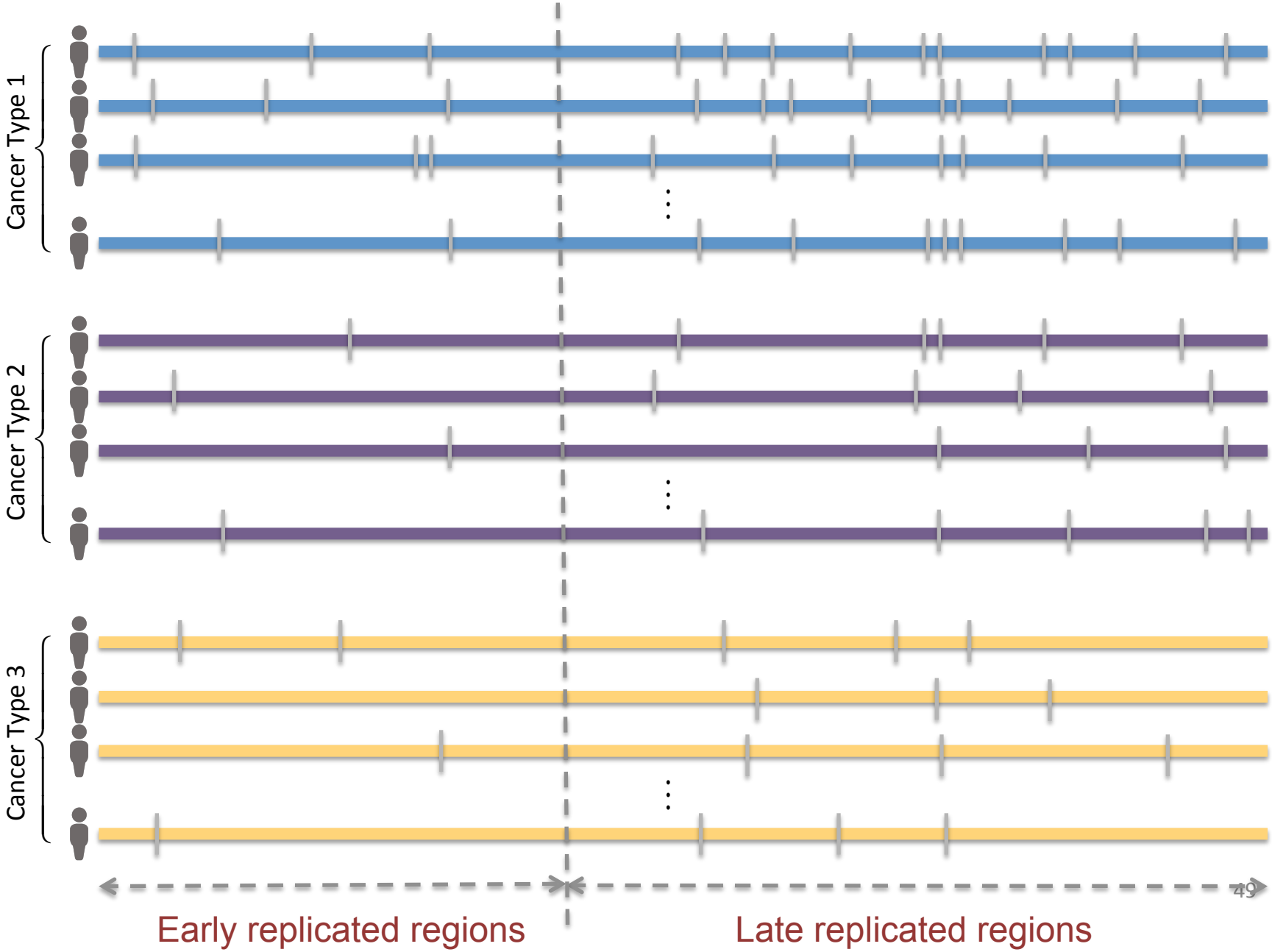
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

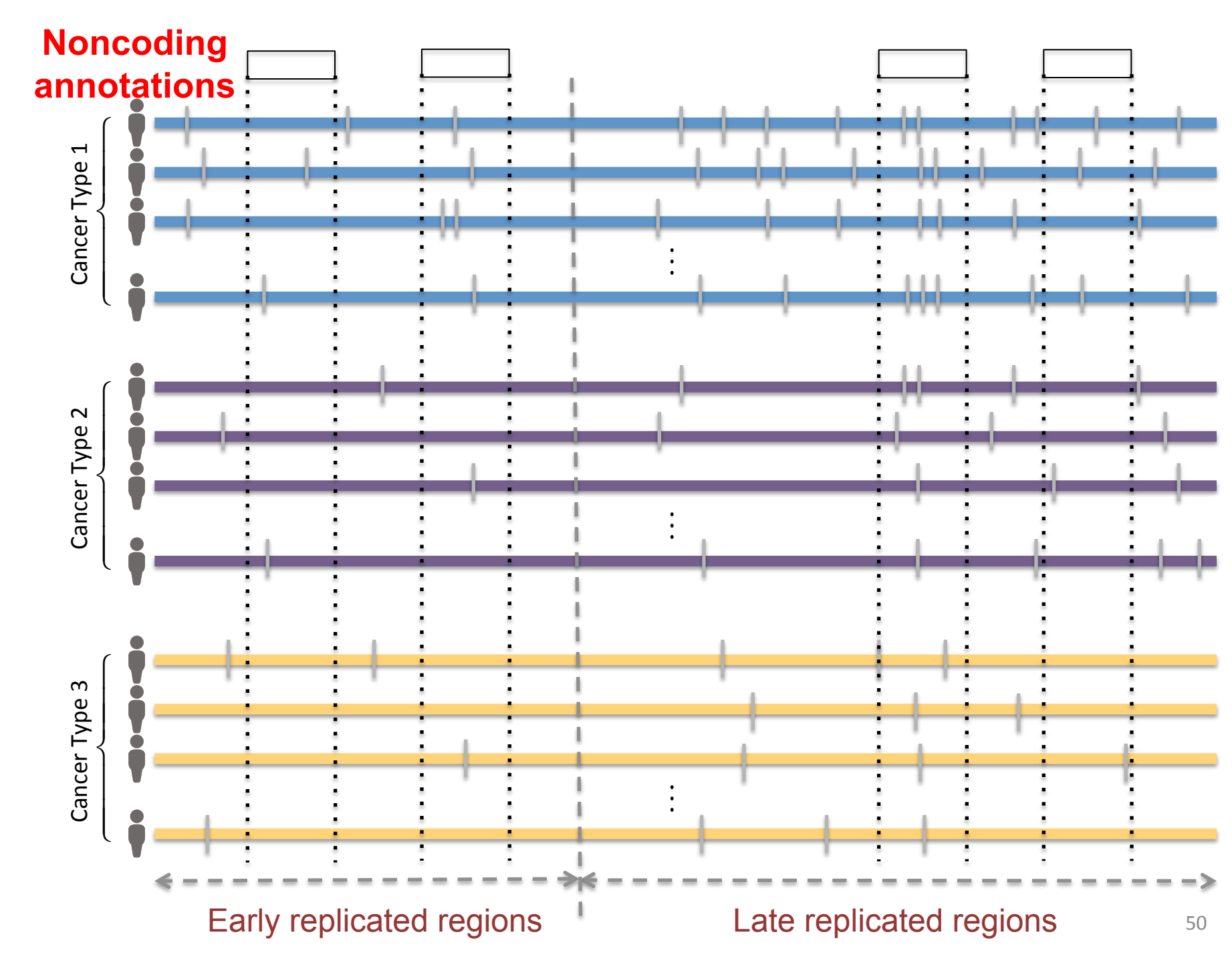
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

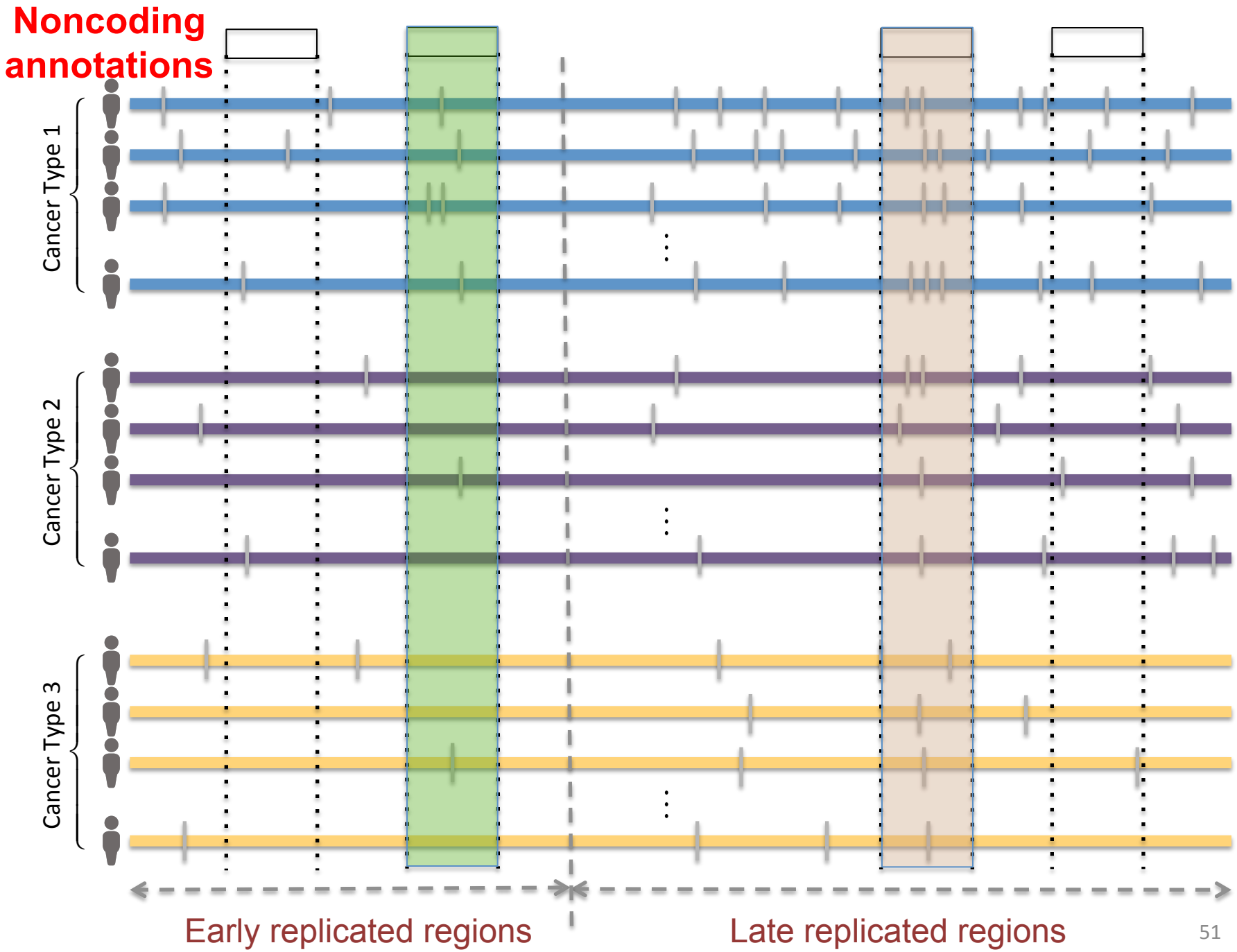
Mutation recurrence



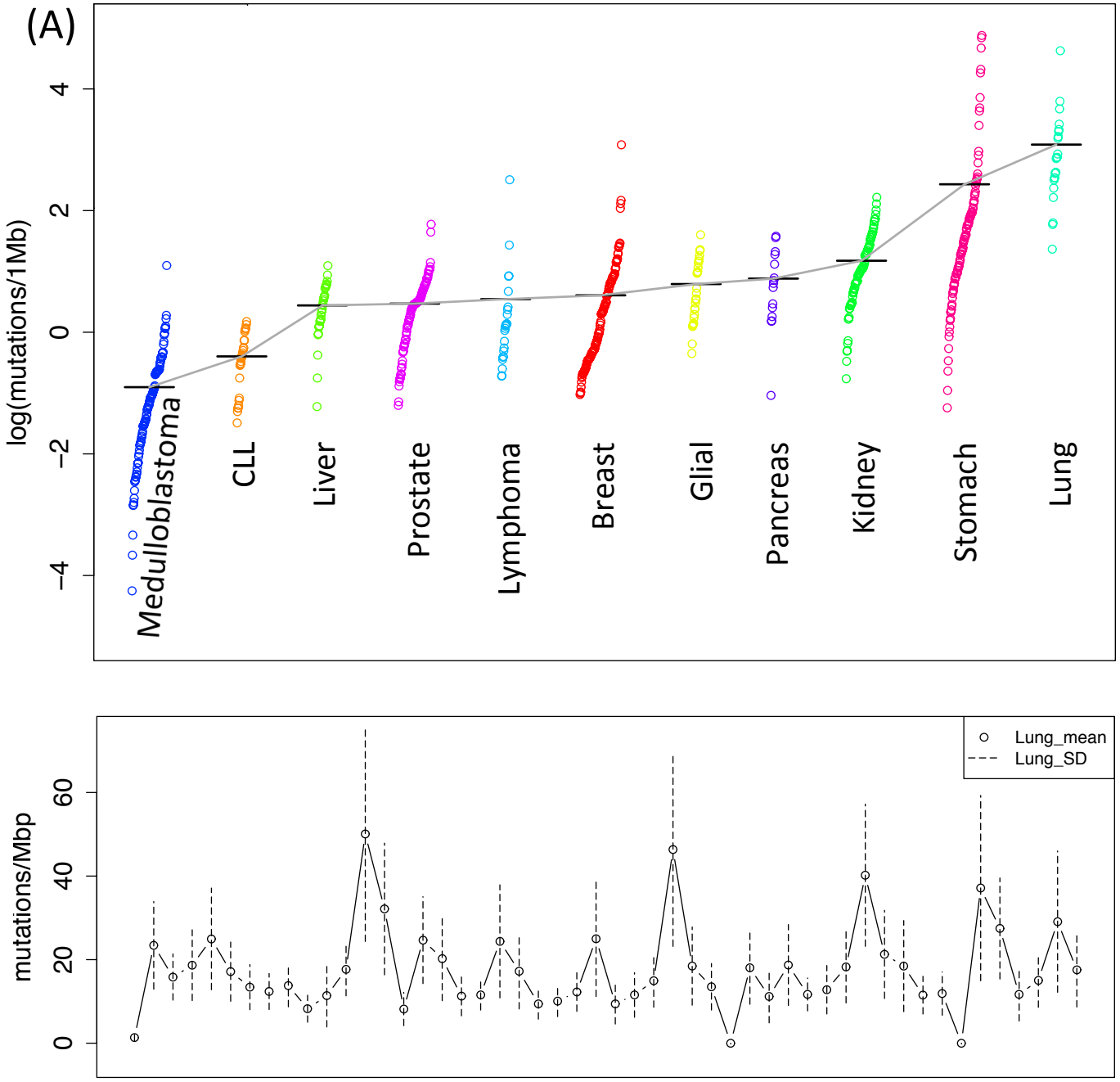
Mutation recurrence







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



[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)

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

Model 2: Varying Mutation Rate

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

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

Model 3: Varying Mutation Rate with Replication Timing Correction

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

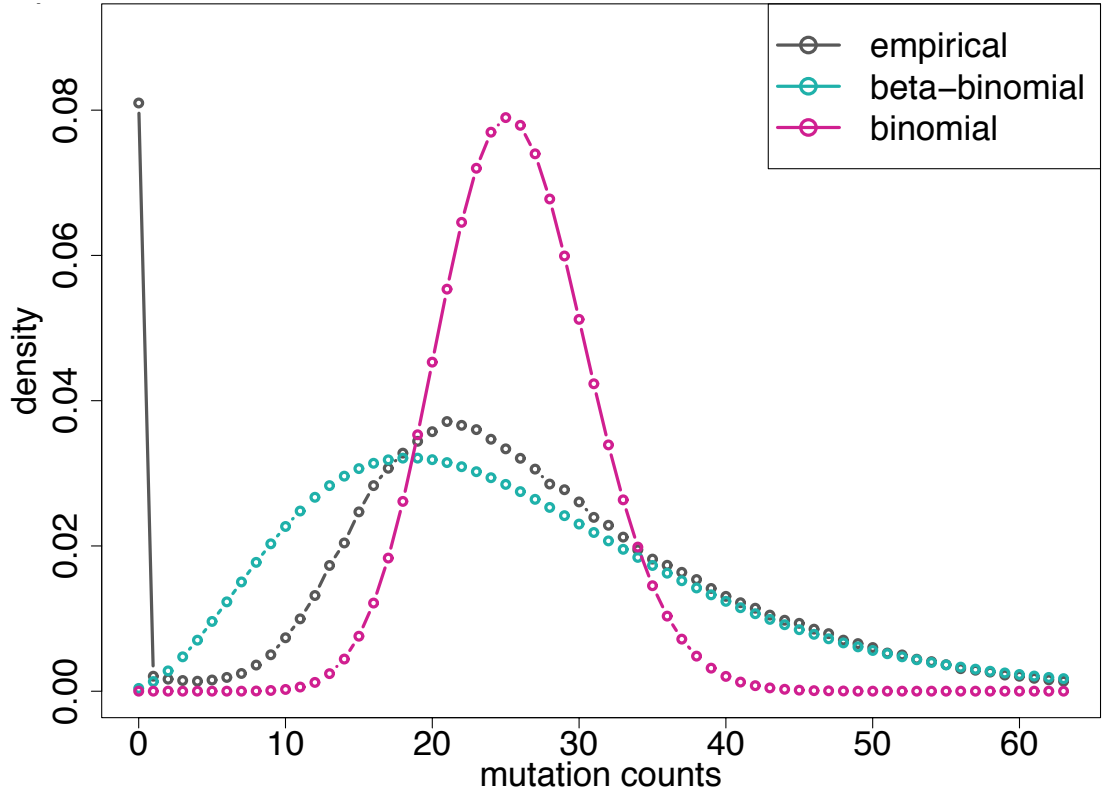
$$\mathbf{p_i} : Beta(\mu|\mathbf{R}, \sigma|\mathbf{R})$$

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

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

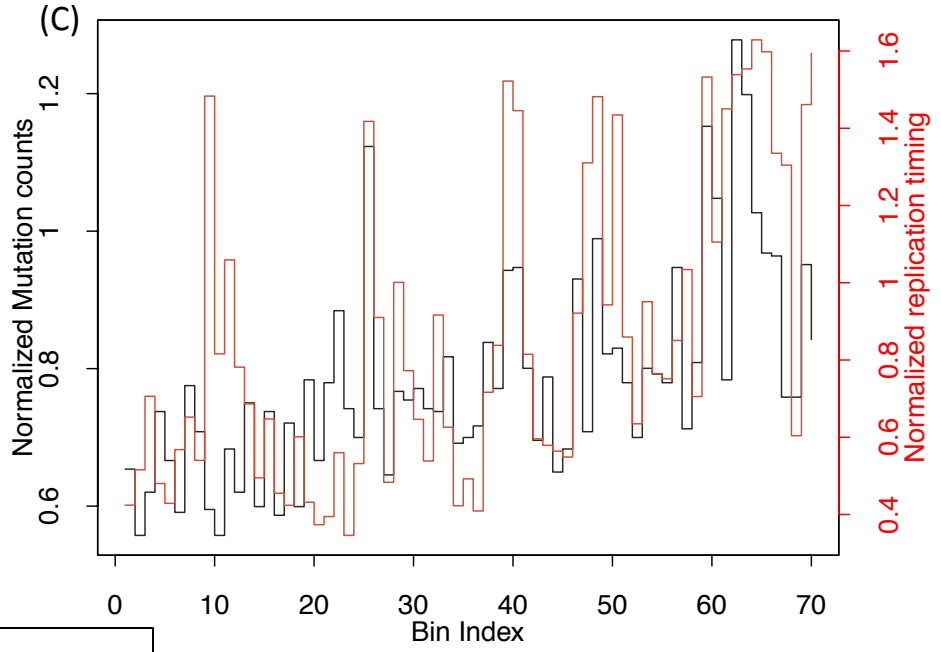
LARVA Model Comparison

- 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

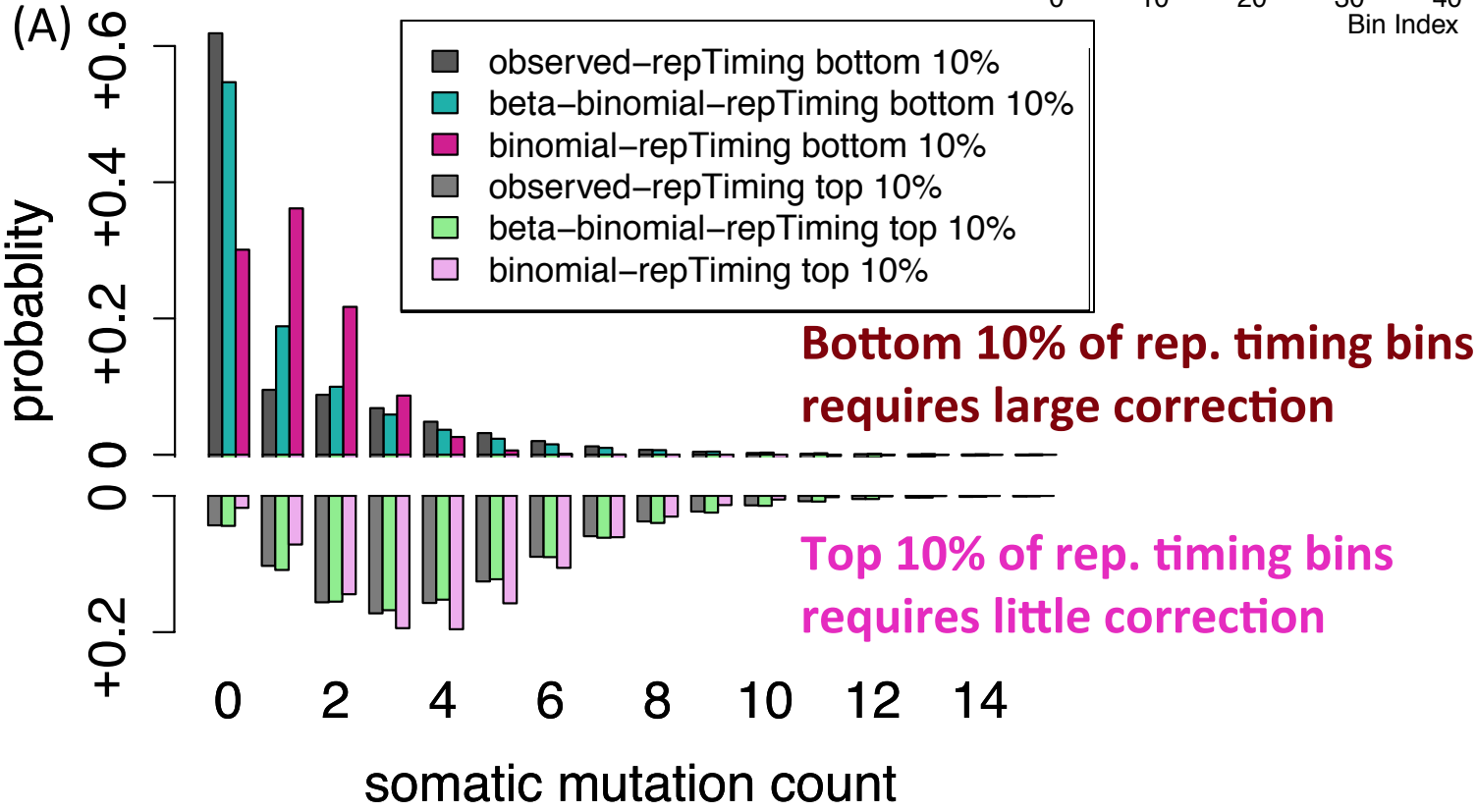


[Lochovsky et al. *NAR* ('15)]

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



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

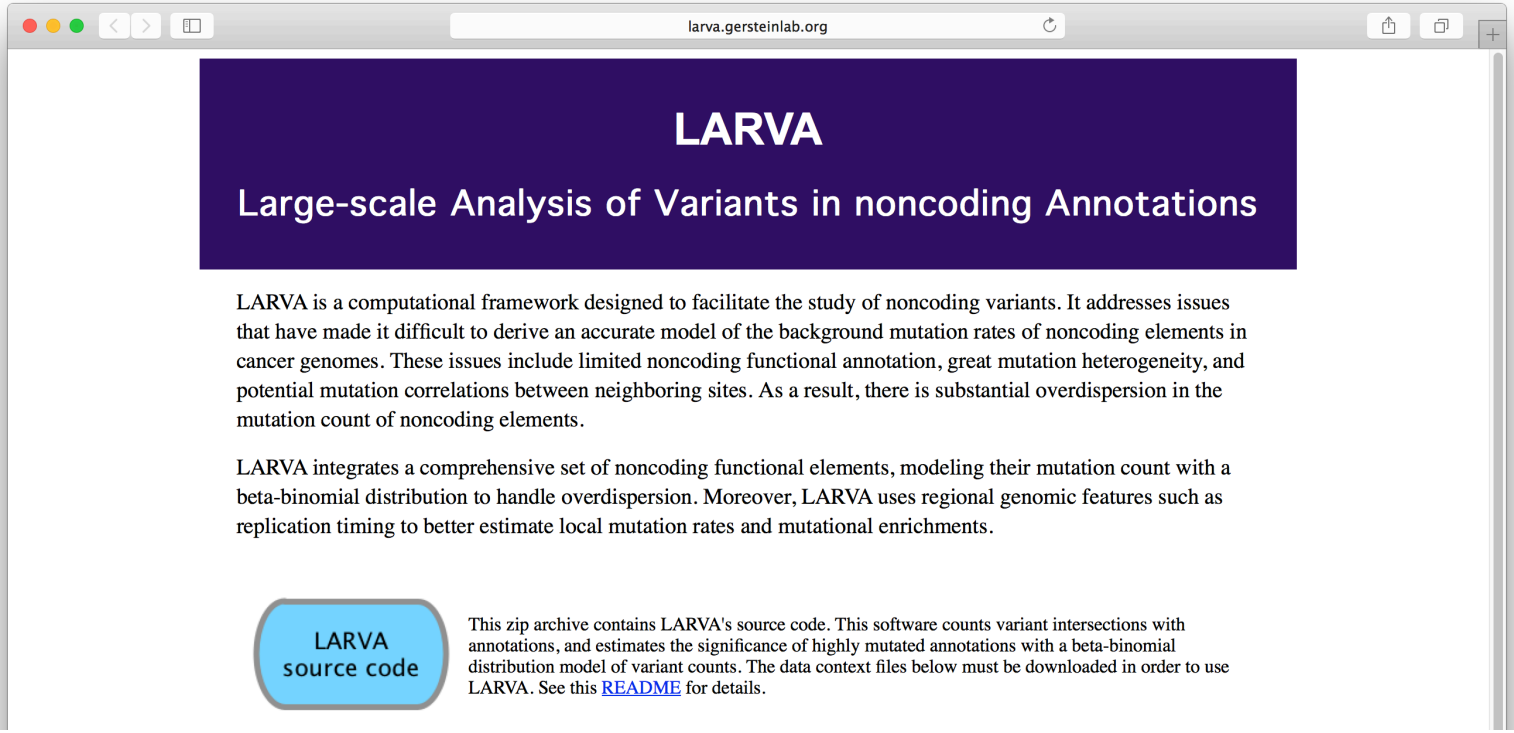


Bottom 10% of rep. timing bins requires large correction

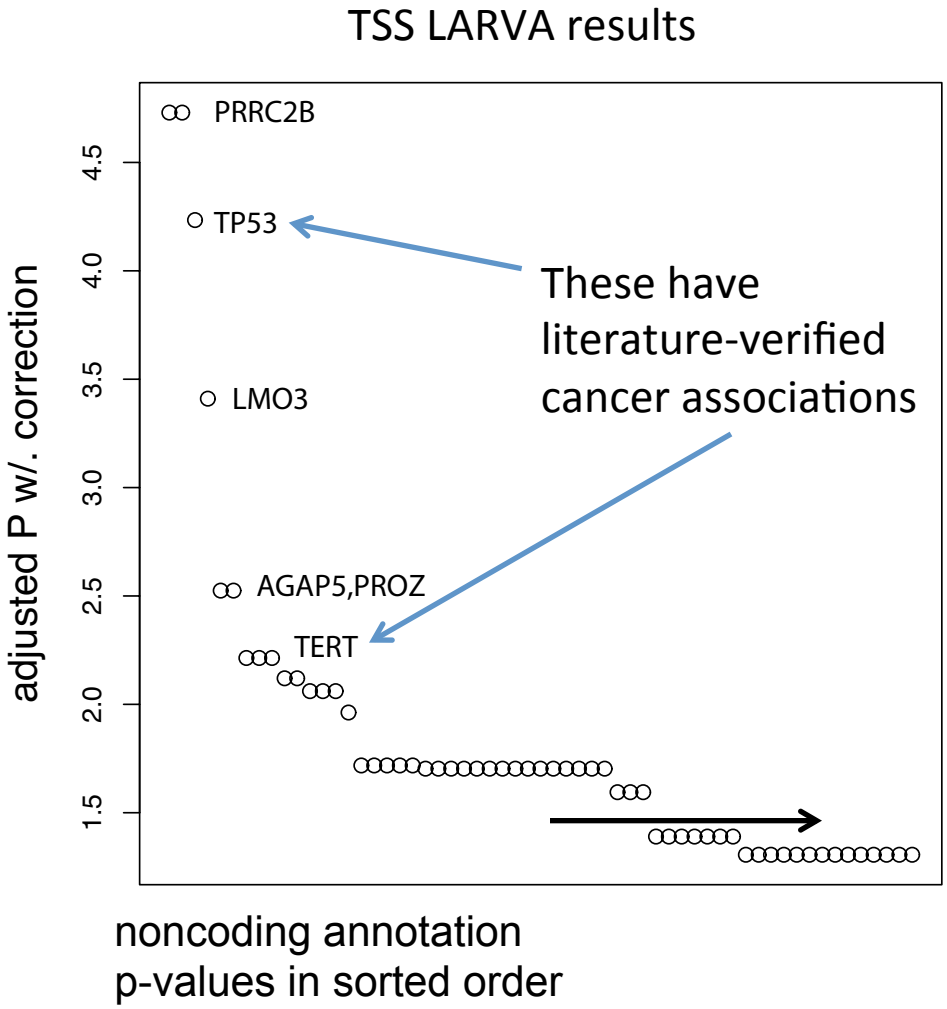
Top 10% of rep. timing bins requires little correction

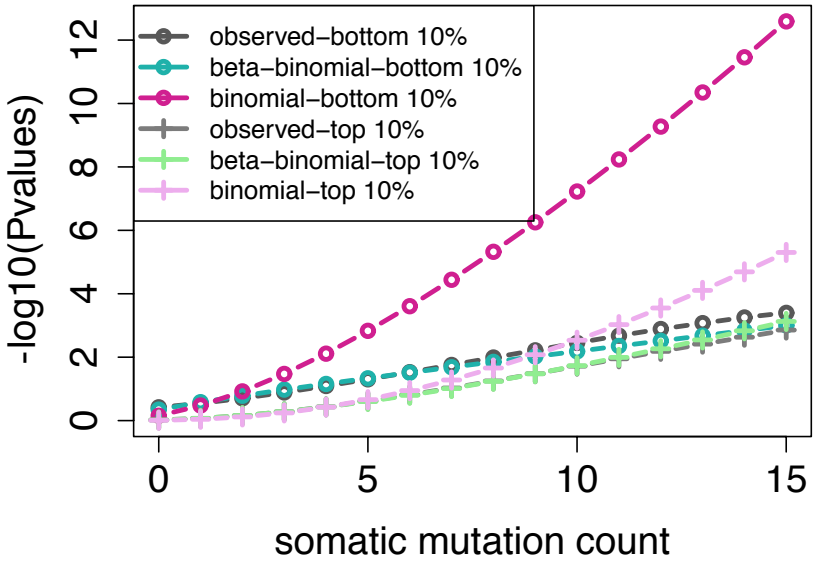
LARVA Implementation

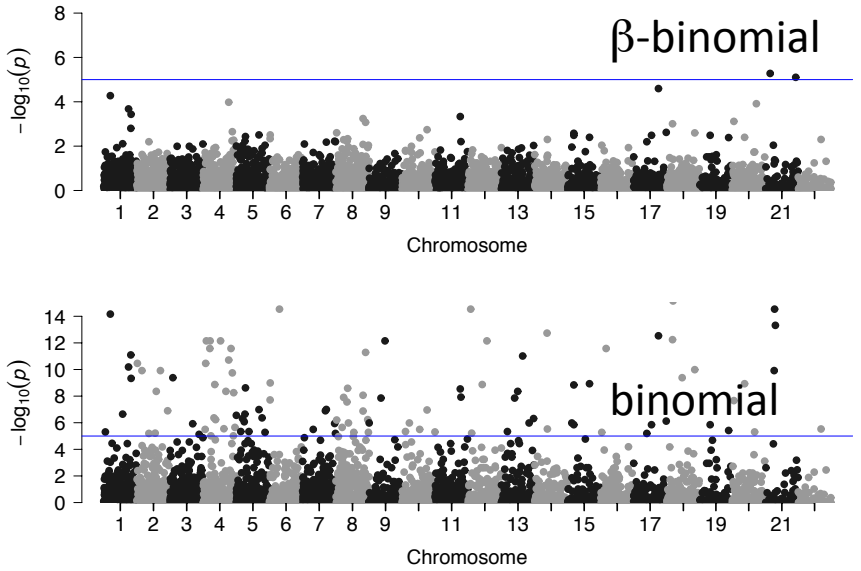
- http://larva.gersteinlab.org/
- Freely downloadable C++ program
  - Verified compilation and correct execution on Linux
- A Docker image is also available to download
  - Runs on any operating system supported by Docker
- Running time on transcription factor binding sites (a worst case input size)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



LARVA Results





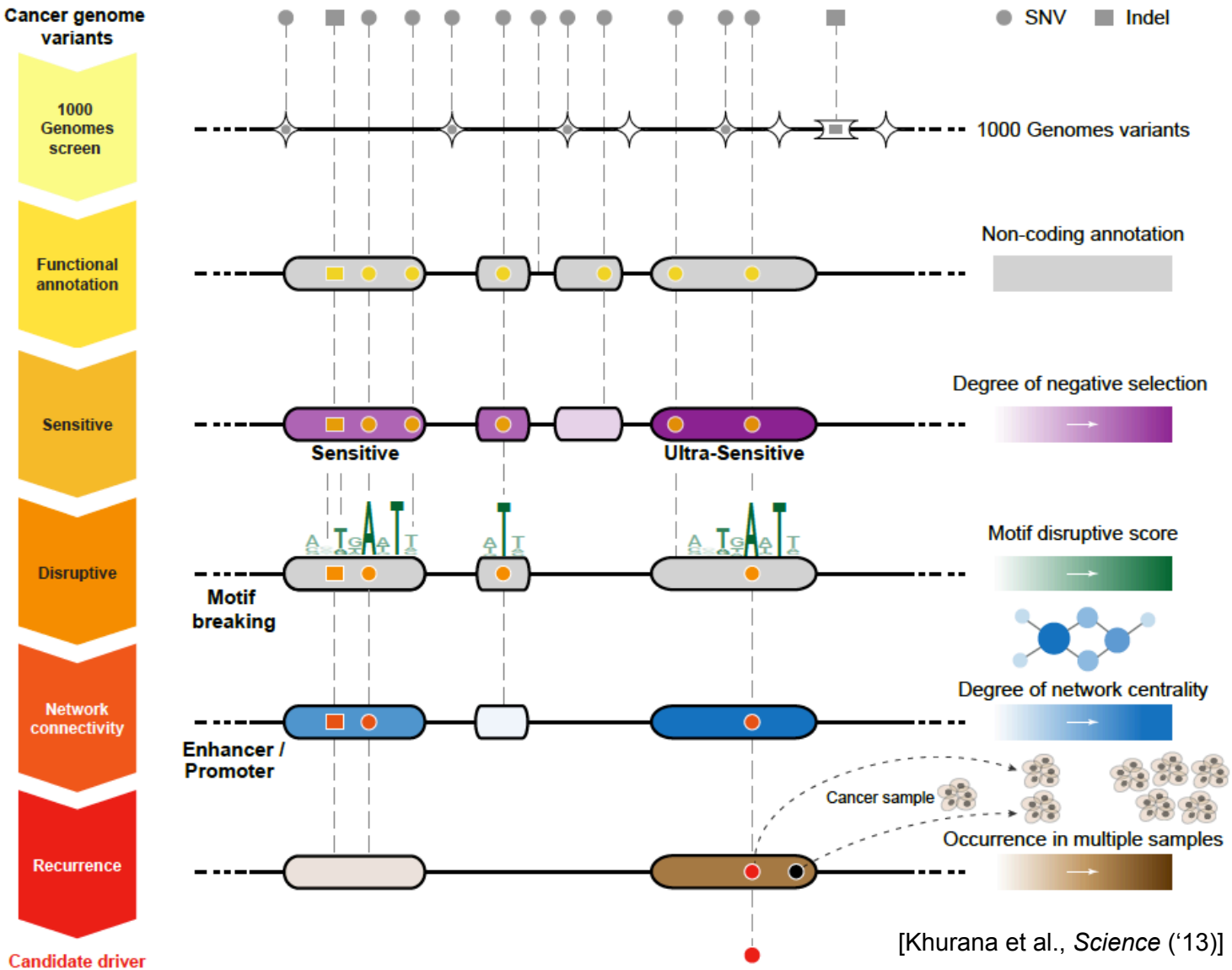


[Lochovsky et al. *NAR* ('15)]

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

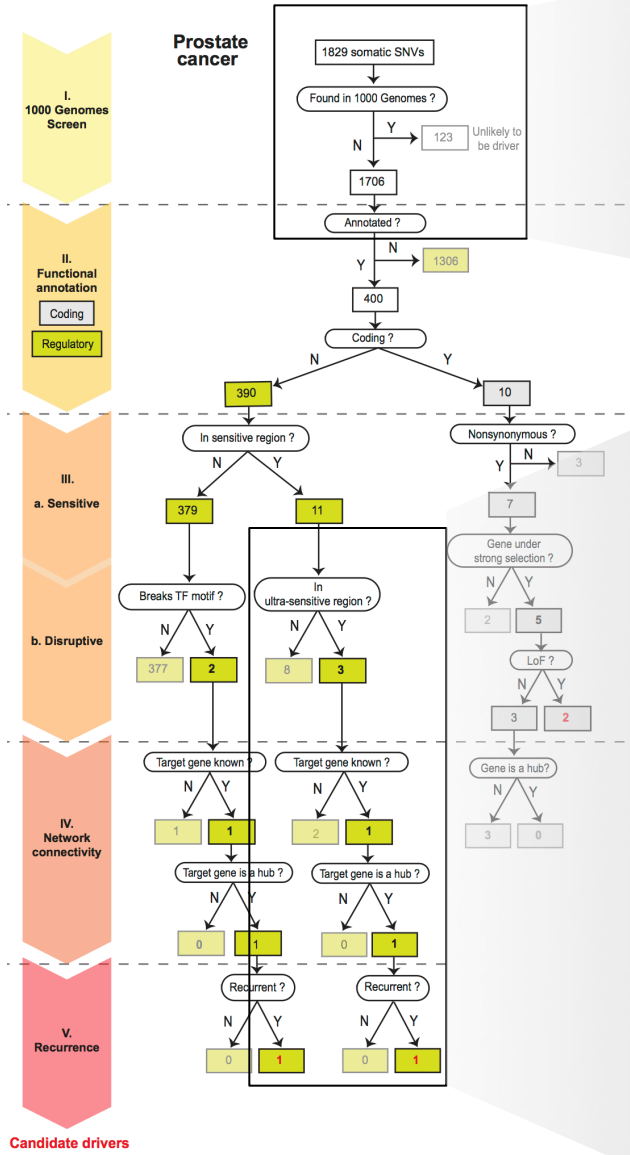
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

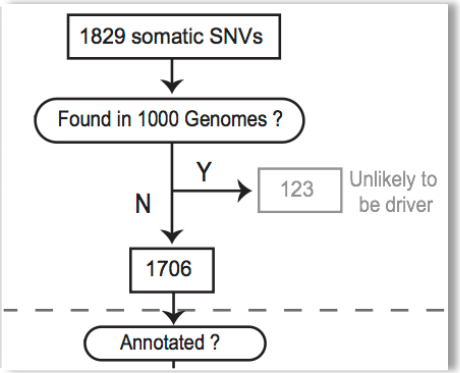
Identification of non-coding candidate drivers amongst somatic variants: Scheme

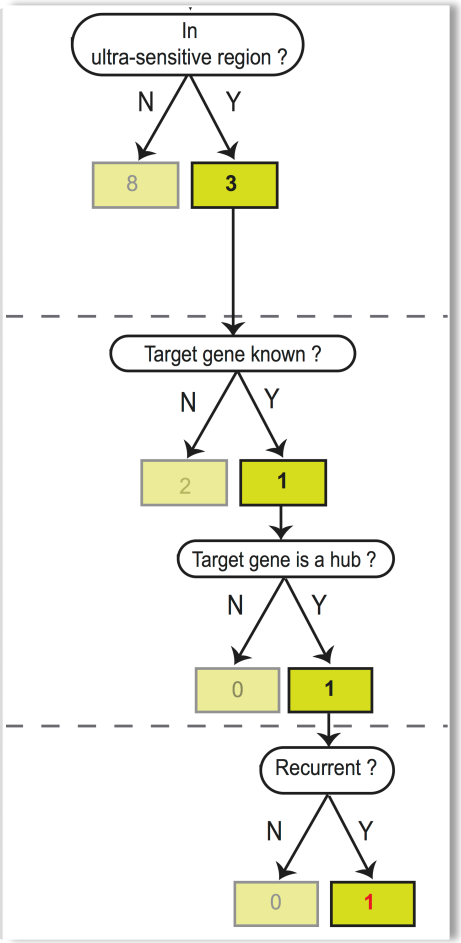


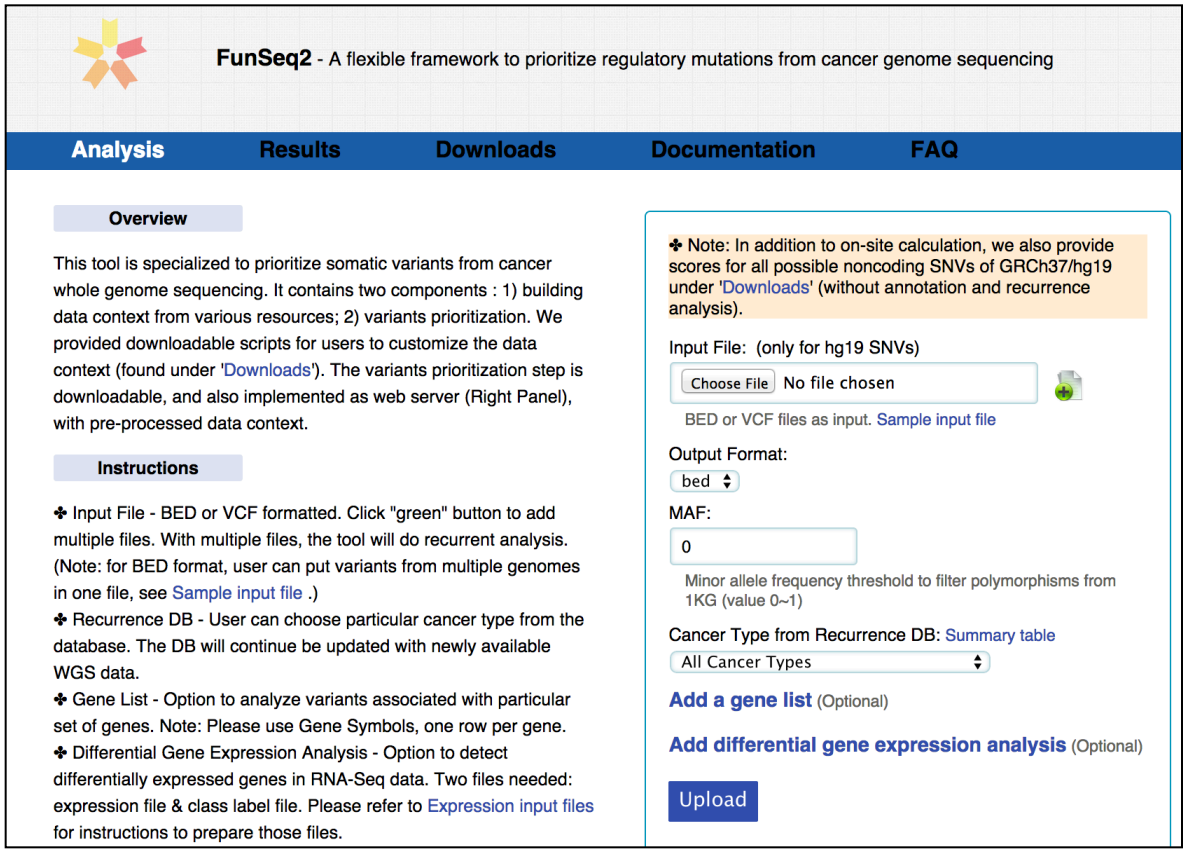
[Khurana et al., *Science* ('13)]

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



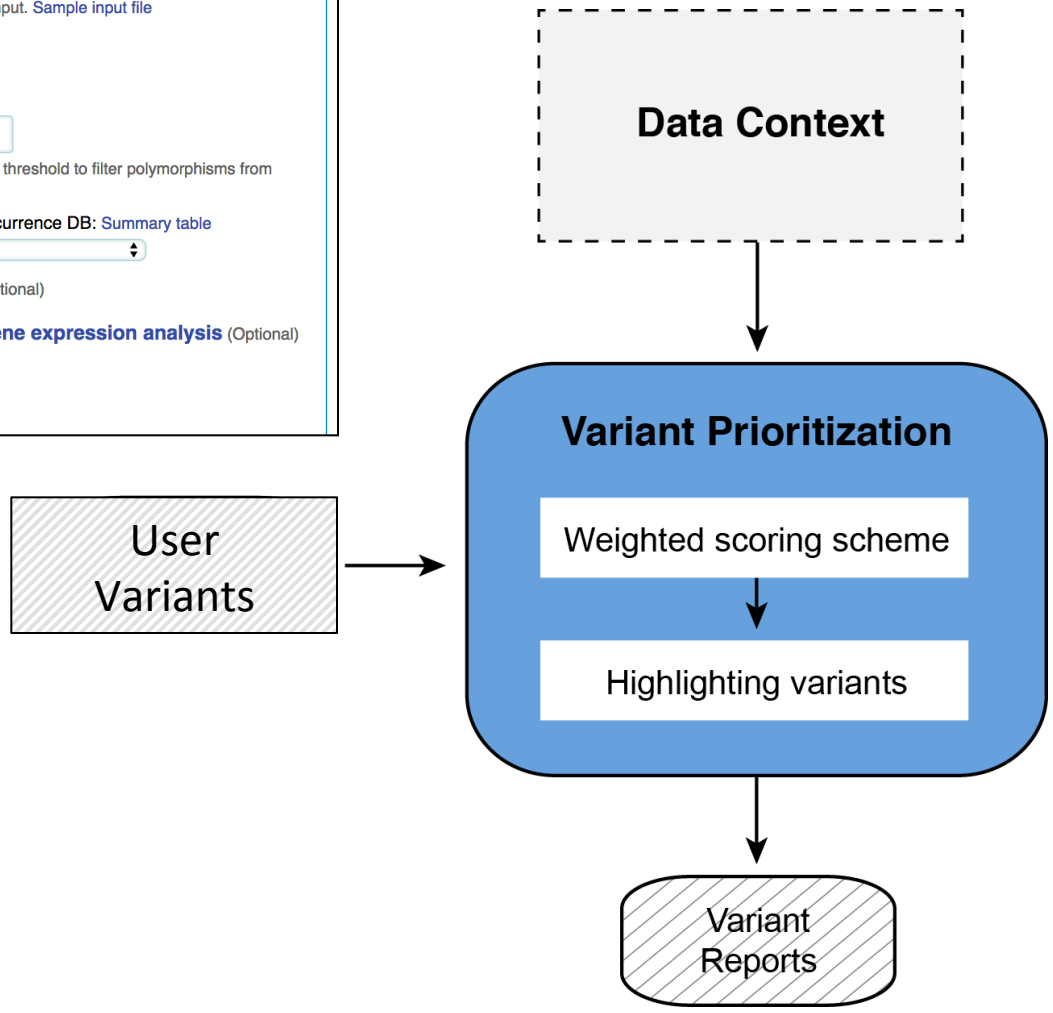






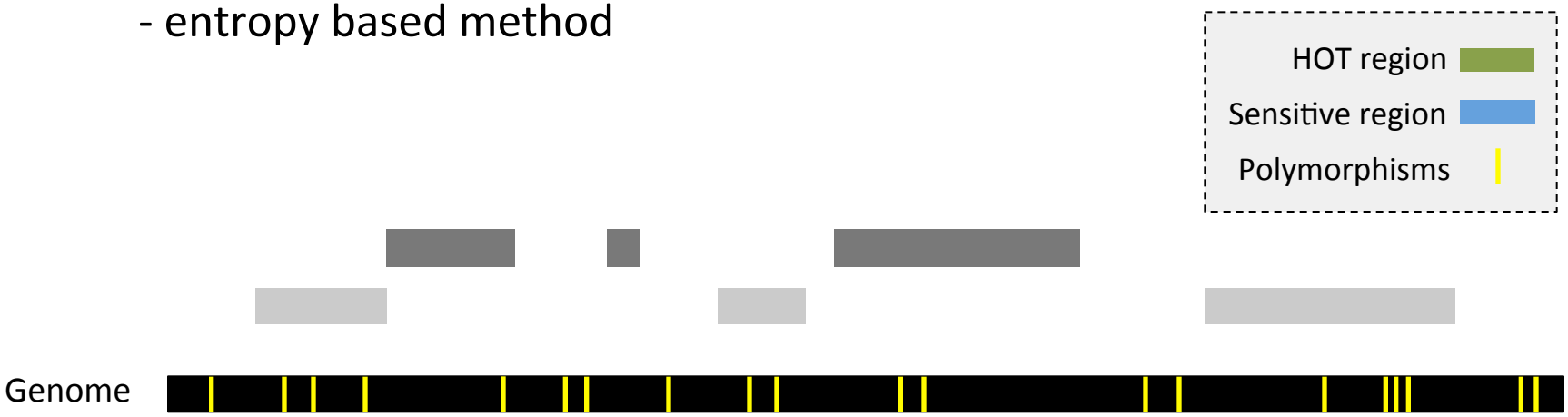
Site integrates user variants with large-scale context

FunSeq.gersteinlab.org



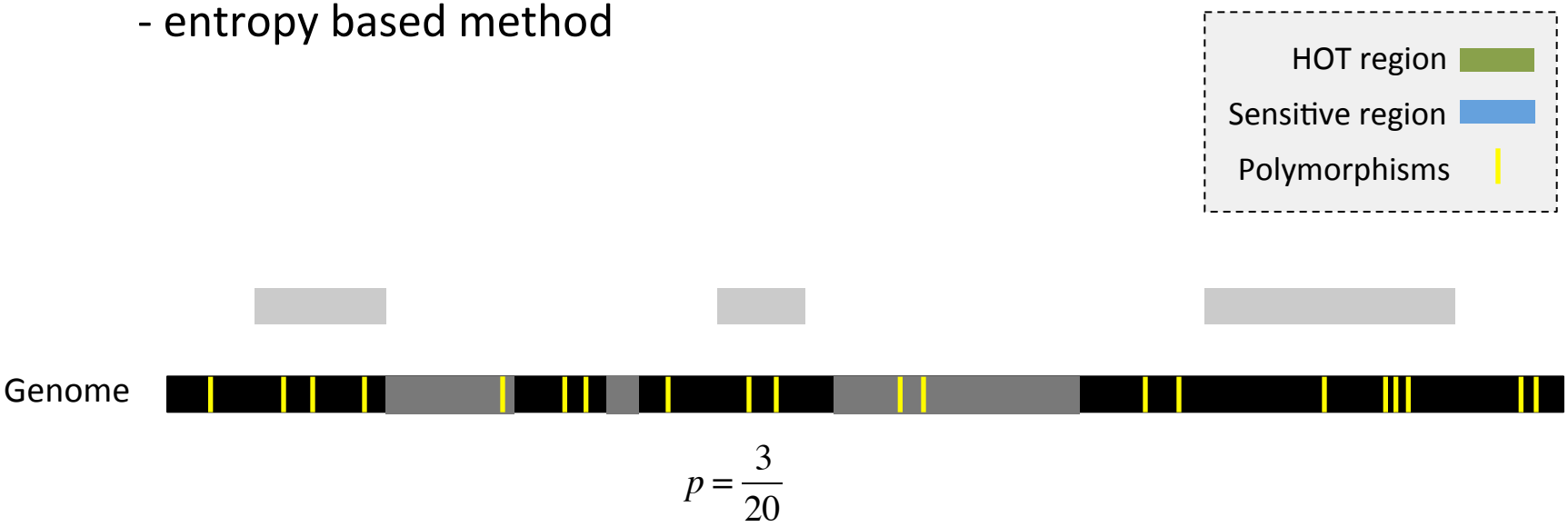
[Fu et al., GenomeBiology ('14)]

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



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

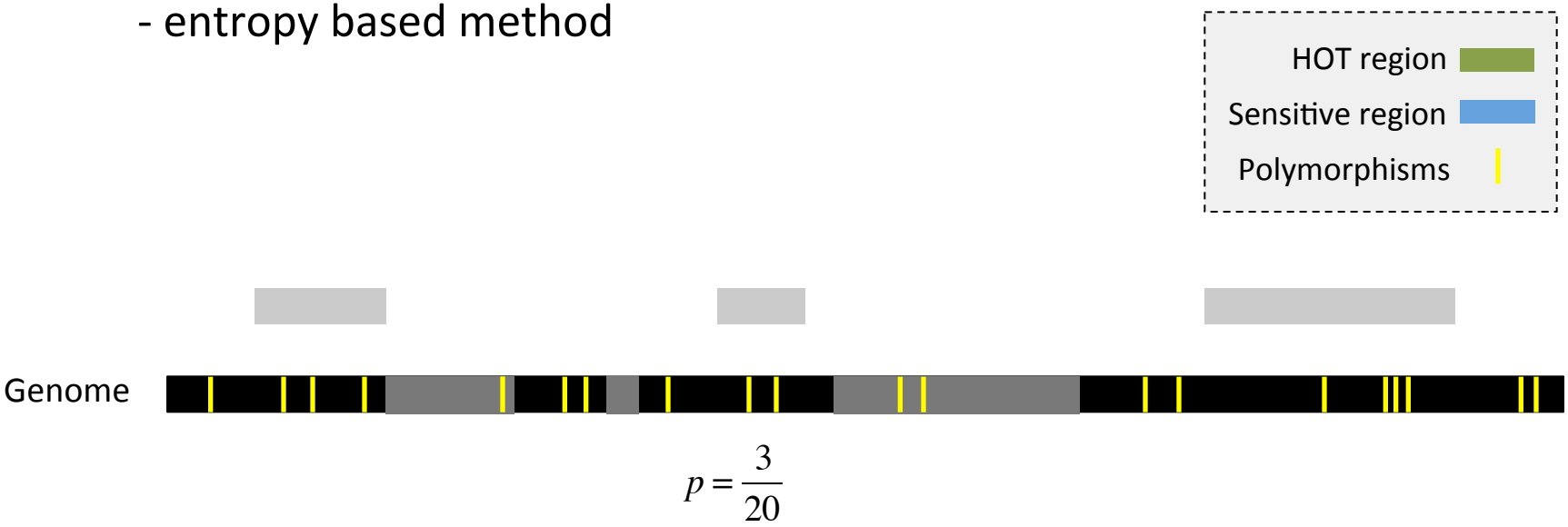
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



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

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

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



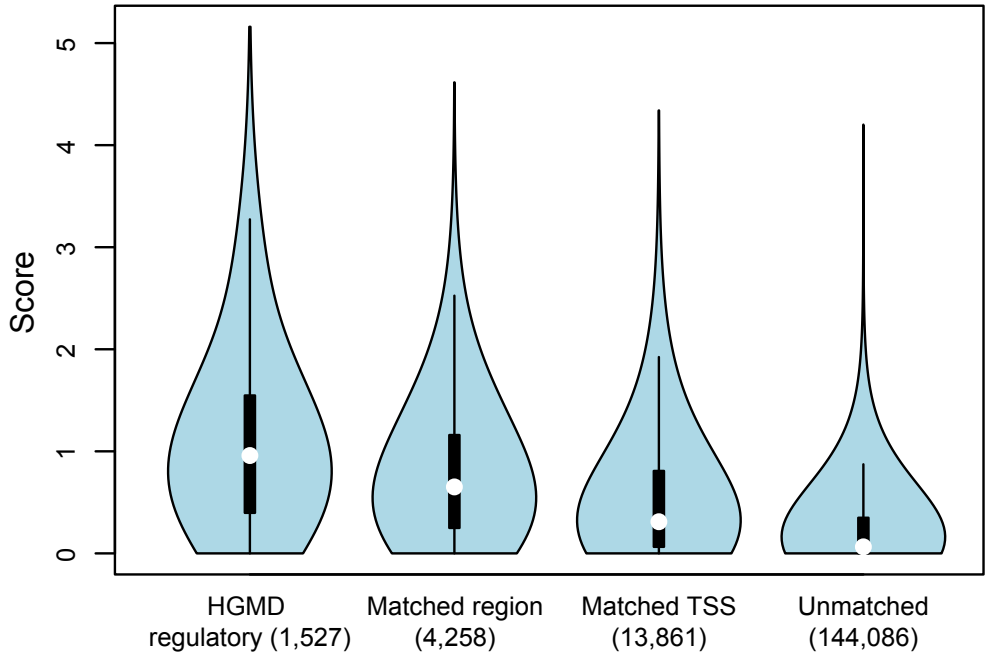
Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

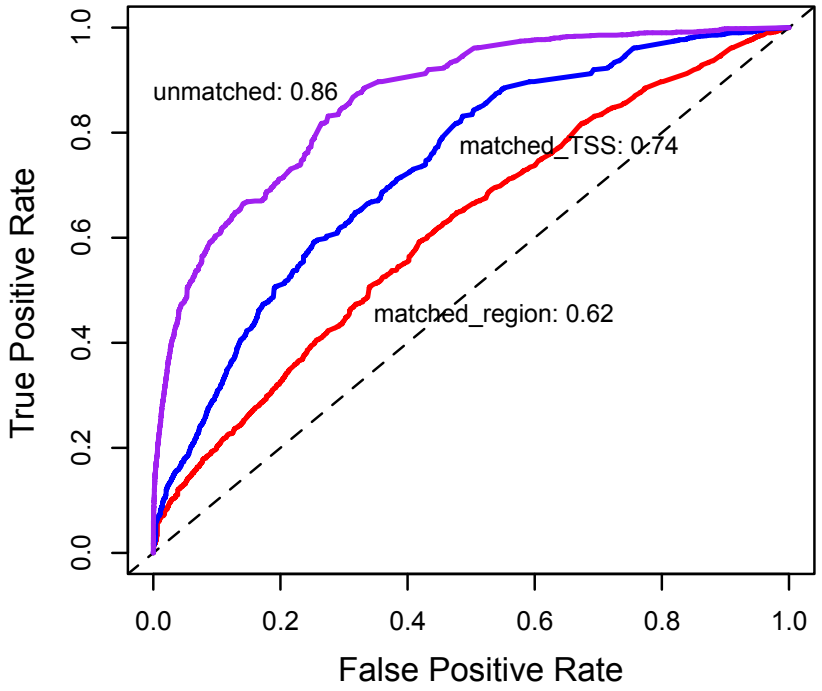
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

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

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

Germline pathogenic variants show higher core scores than controls





3 controls with natural polymorphisms (allele frequency >= 1%)

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

Ritchie et al., Nature Methods, 2014

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

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 finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing 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 finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- 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)
  - Prioritizing in terms of network connectivity (eg hubs)
- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for 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
    - Prioritzing rare germline variants

MUSIC.gersteinlab.org
A **Harmanci**,
J Rozowsky

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

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

Acknowledgments

Hiring Postdocs. See gersteinlab.org/jobs

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