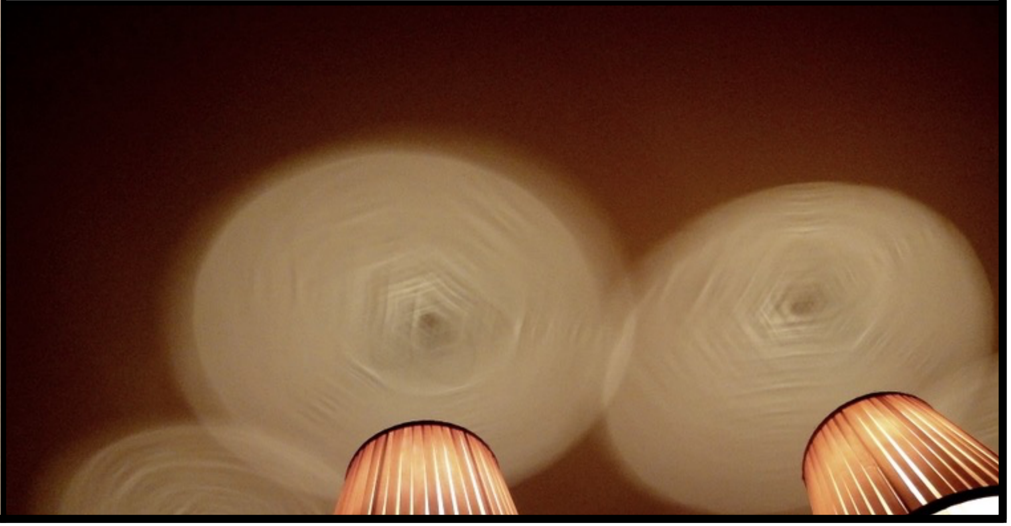
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

Mark Gerstein, Yale

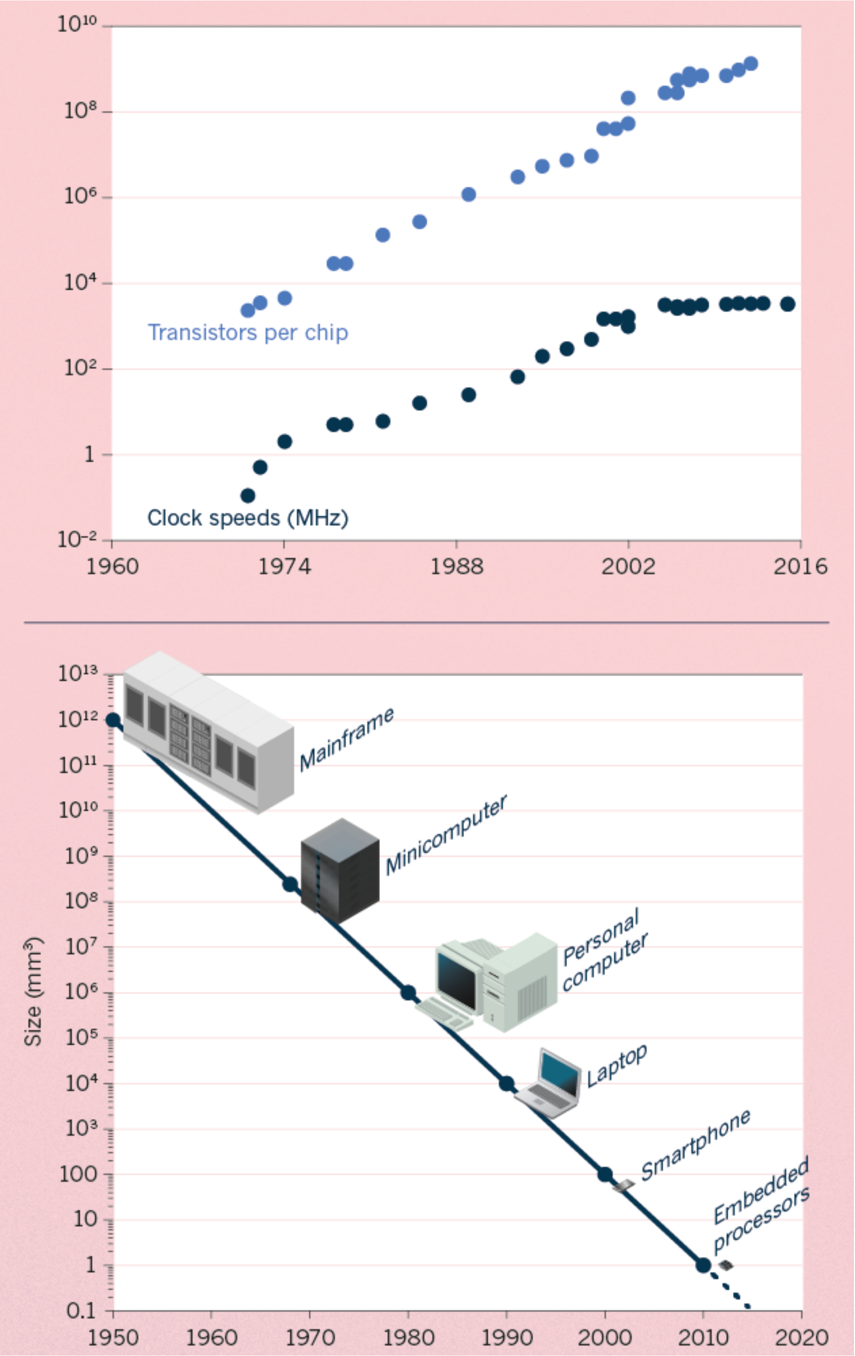
Slides freely downloadable from Lectures.GersteinLab.org

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



Moore’s Law: Exponential Scaling of Computer Technology

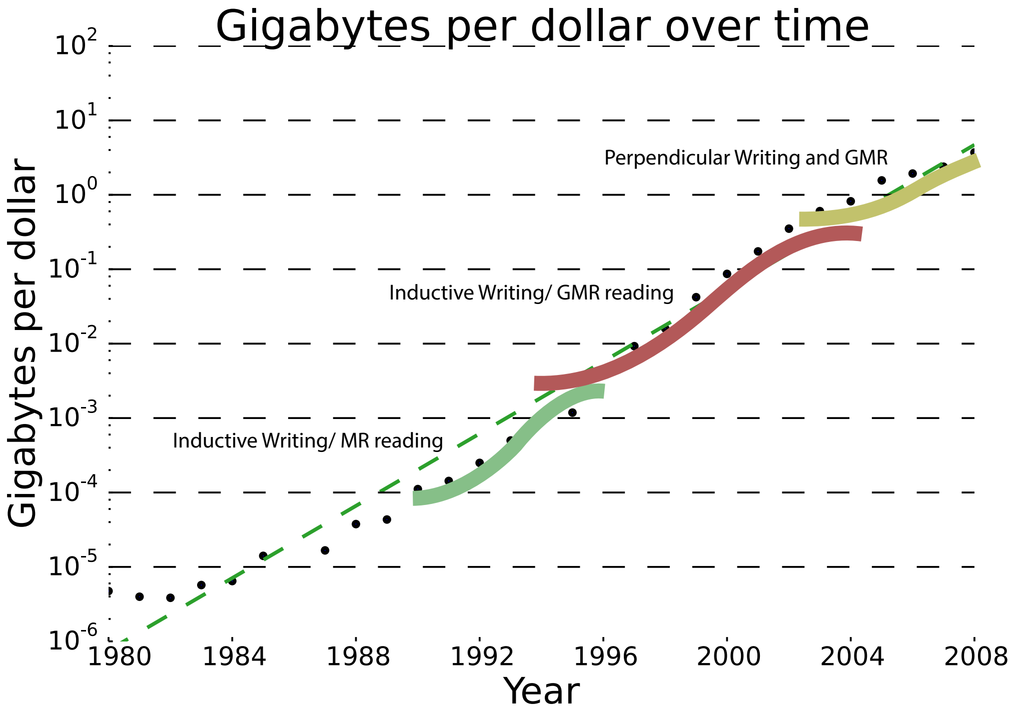
- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.

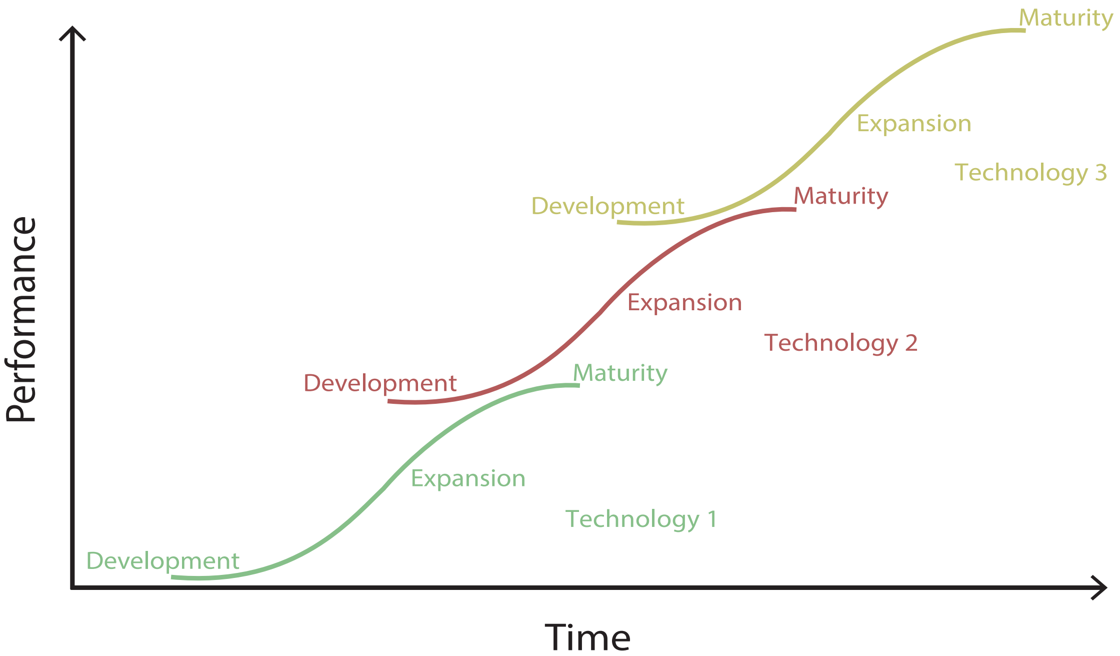


[Waldrop (‘15) Nature]

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

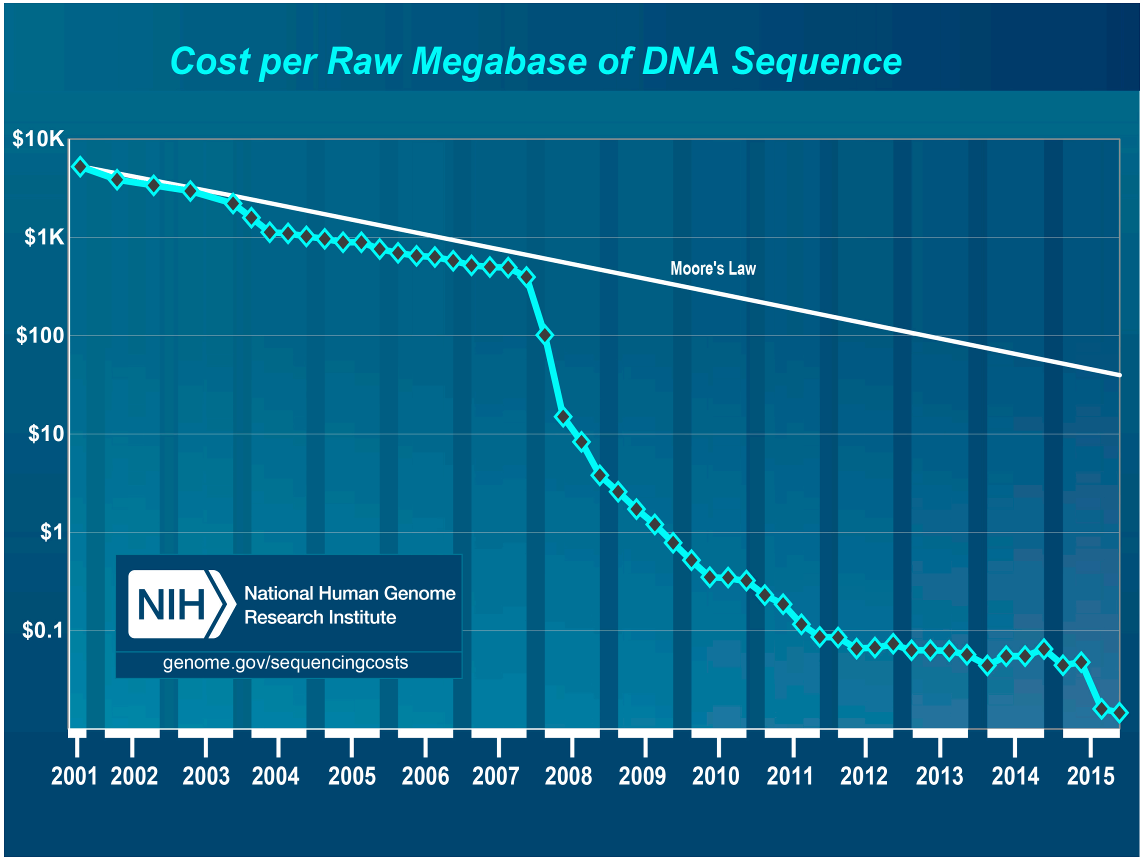




[Muir et al. ('15) GenomeBiol.]

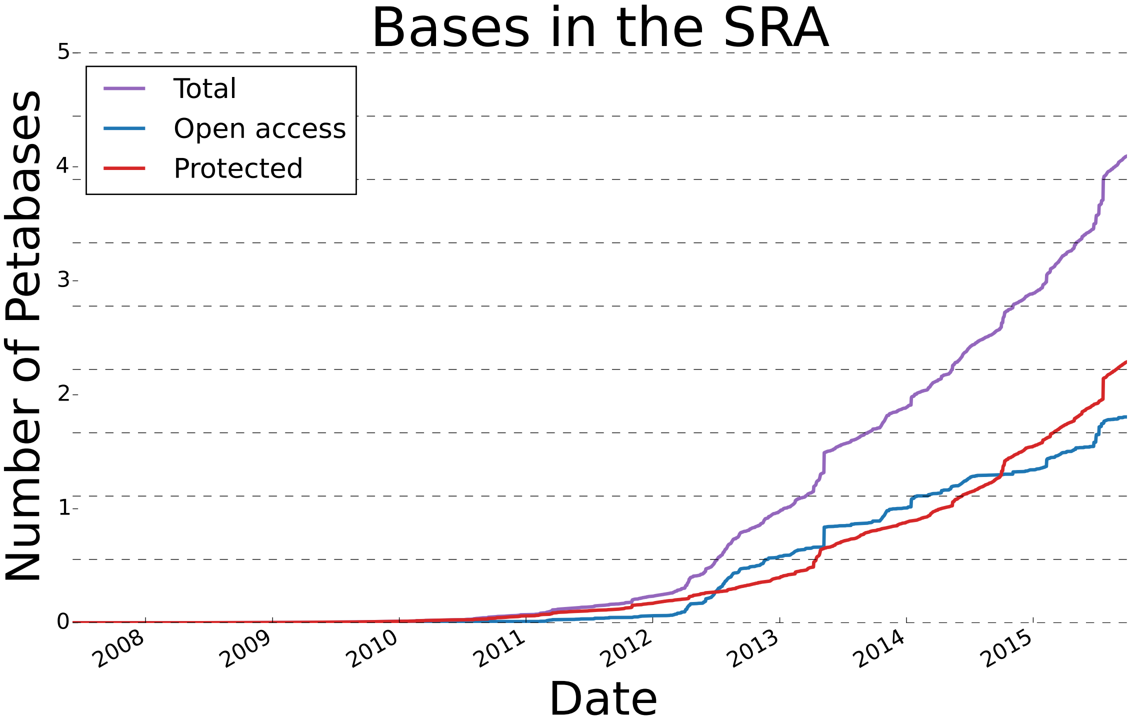
Sequencing Data Explosion:
Faster than Moore's Law for a Time (or a S-curve)

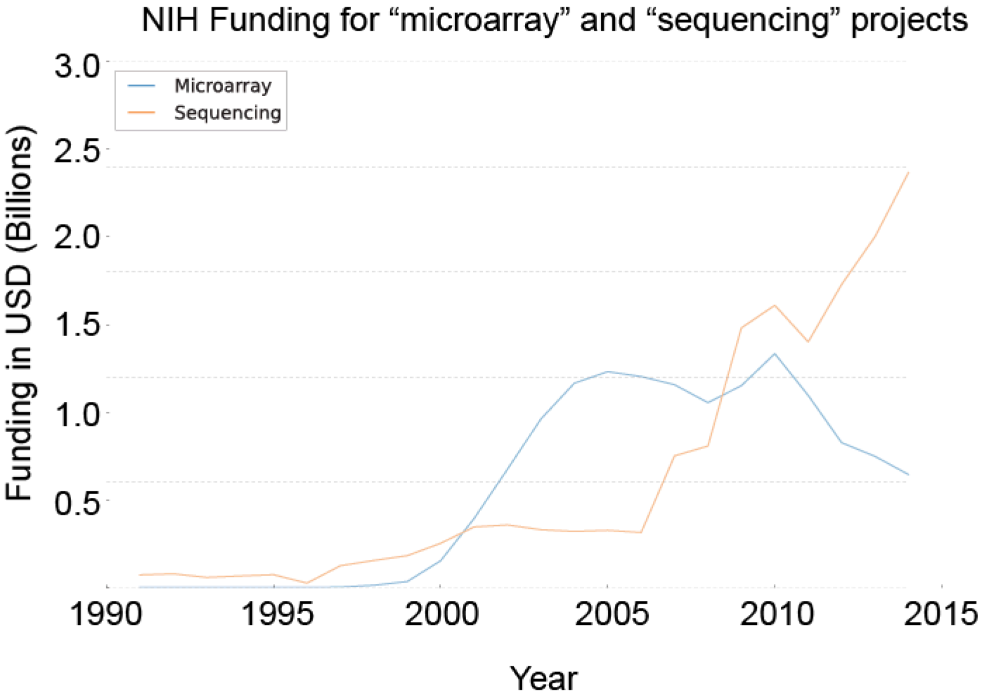
- DNA sequencing has gone through technological S-curves
  - In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
  - The advent of NGS was a shift to a new technology with dramatic decrease in cost).



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.]

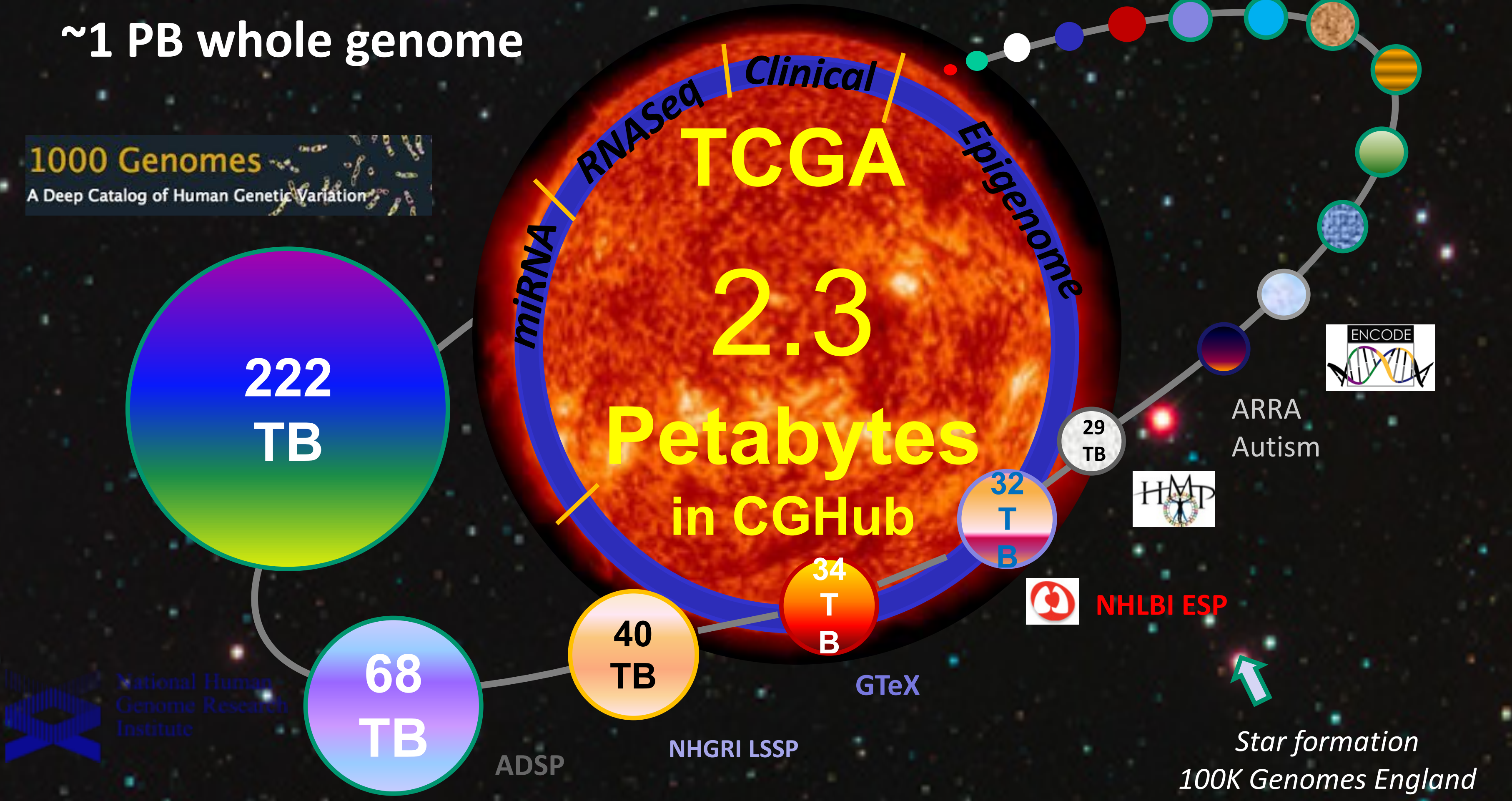
Sequence Universe

SRA ~1 petabyte

TCGA endpoint: ~2.5 Petabytes

~1.5 PB exome

~1 PB whole genome



1000 Genomes
A Deep Catalog of Human Genetic Variation

TCGA

2.3 Petabytes in CGHub

Clinical

RNASeq

miRNA

Epigenome

222 TB

68 TB

ADSP

40 TB

NHGRI LSSP

34 TB

GTeX

32 TB

NHLBI ESP

HMP

29 TB

ARRA Autism

ENCODE

Star formation
100K Genomes England

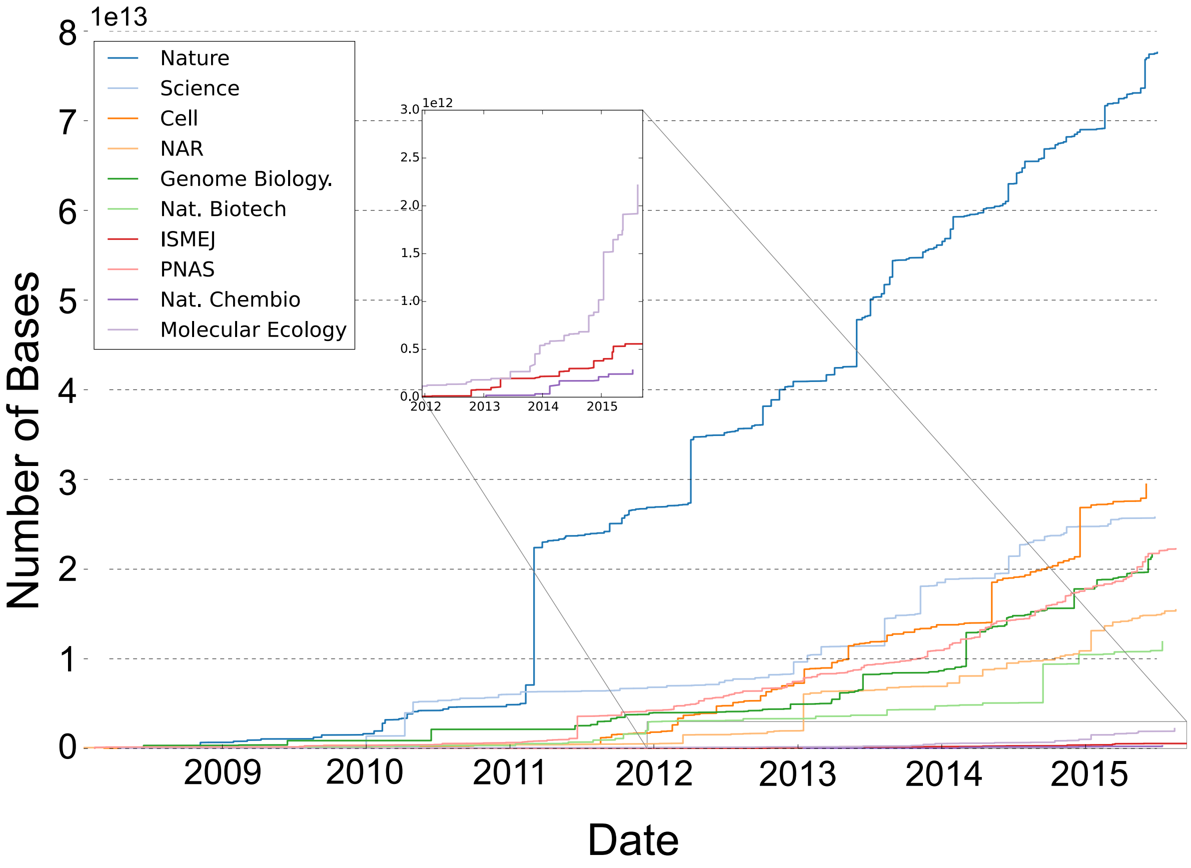
National Human Genome Research Institute

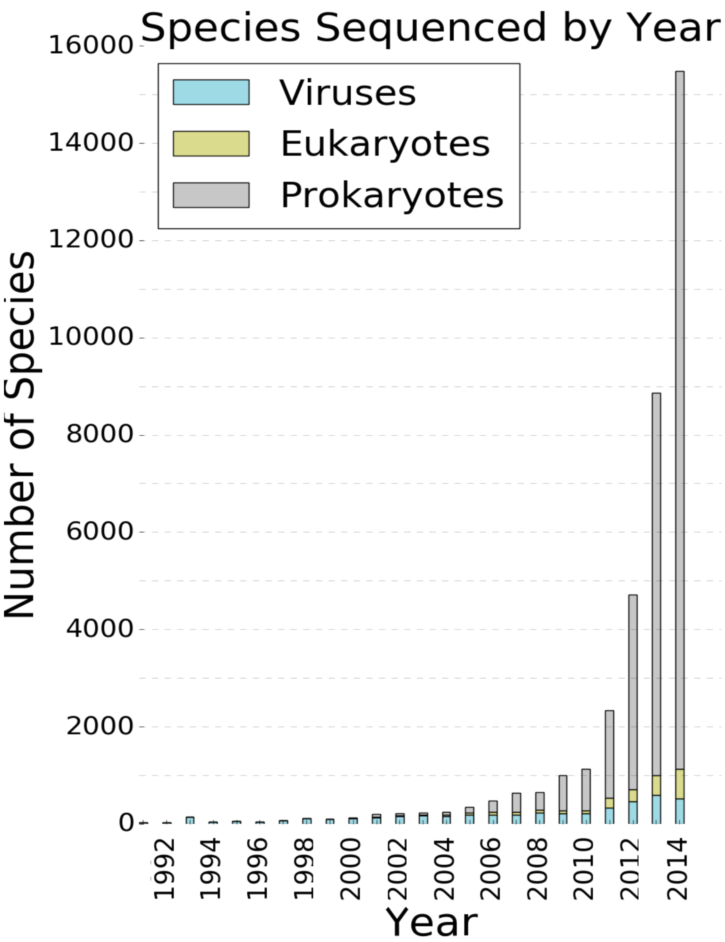
Heidi Sofia, 7-16-15

JESS3

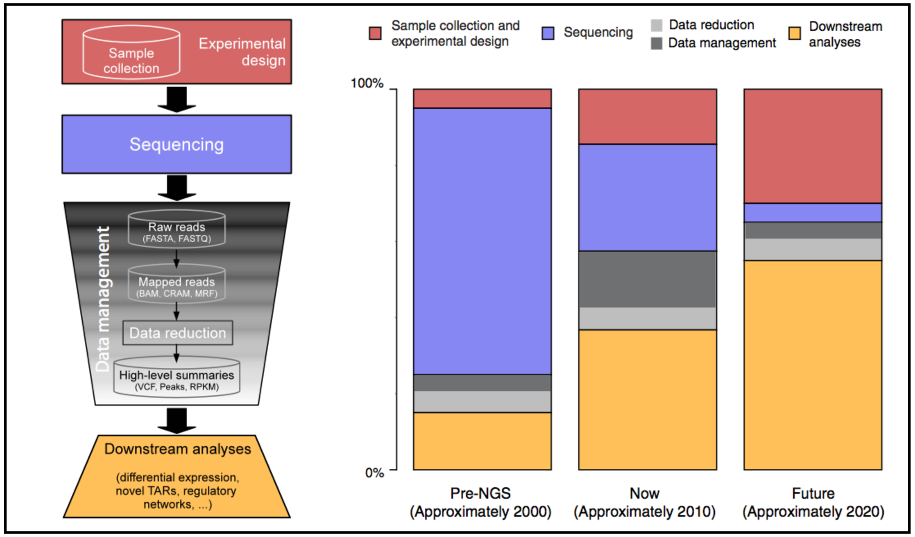
Increasing diversity in sequence data sources

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





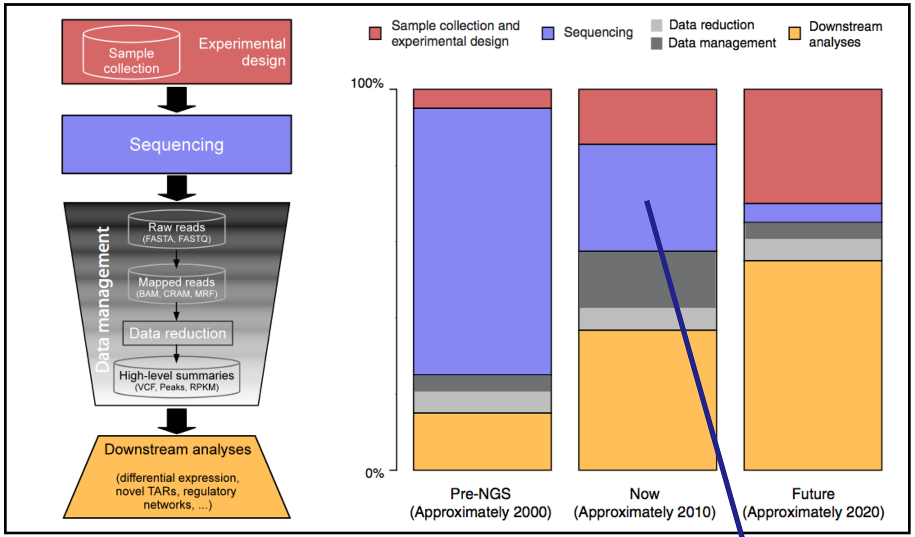
The changing costs of a sequencing pipeline



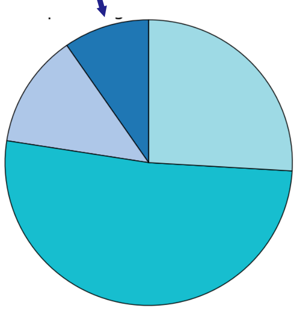
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

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

The changing costs of a sequencing pipeline

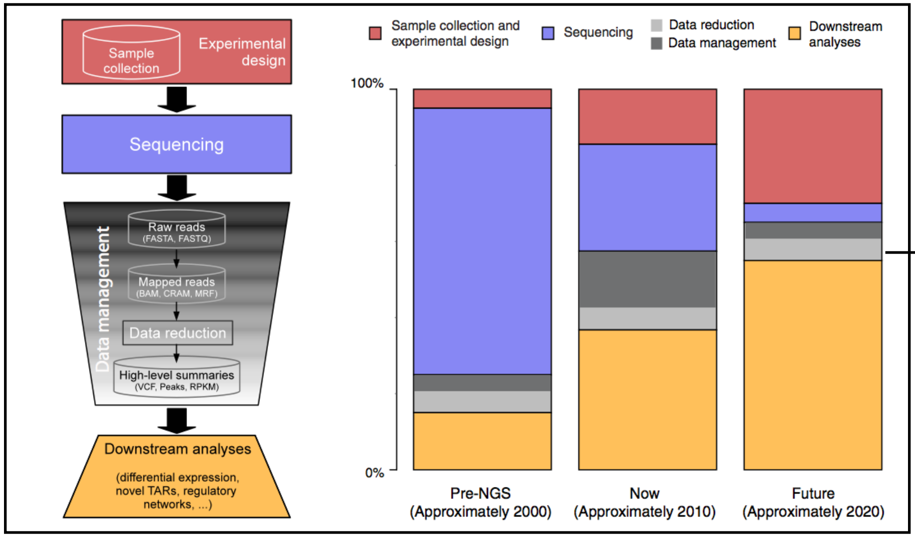


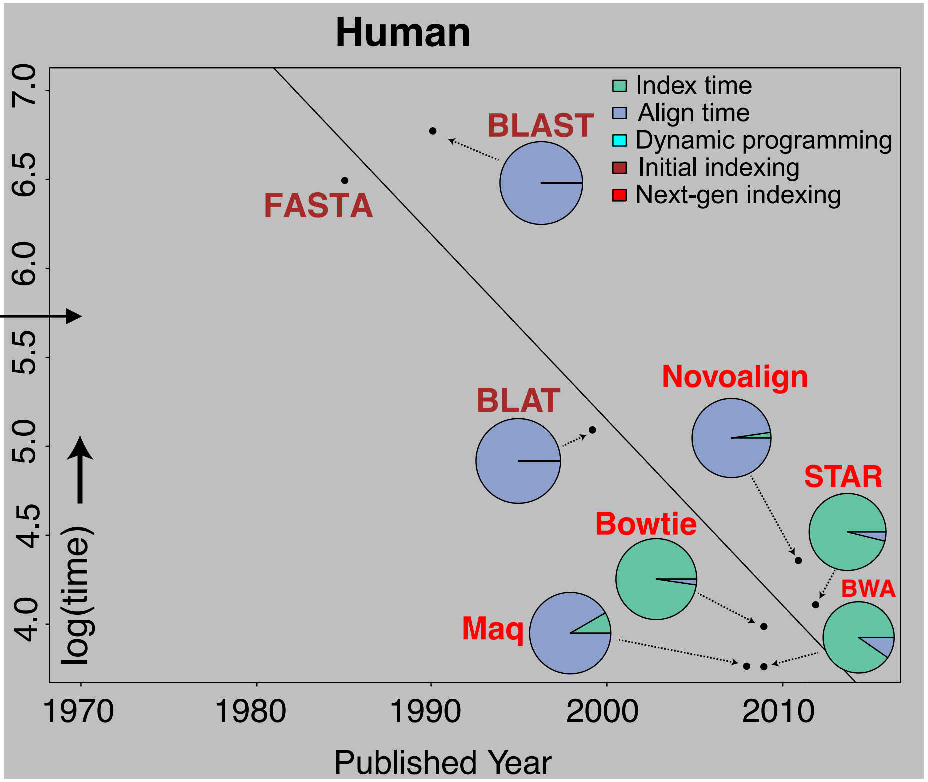
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



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

The changing costs of a sequencing pipeline



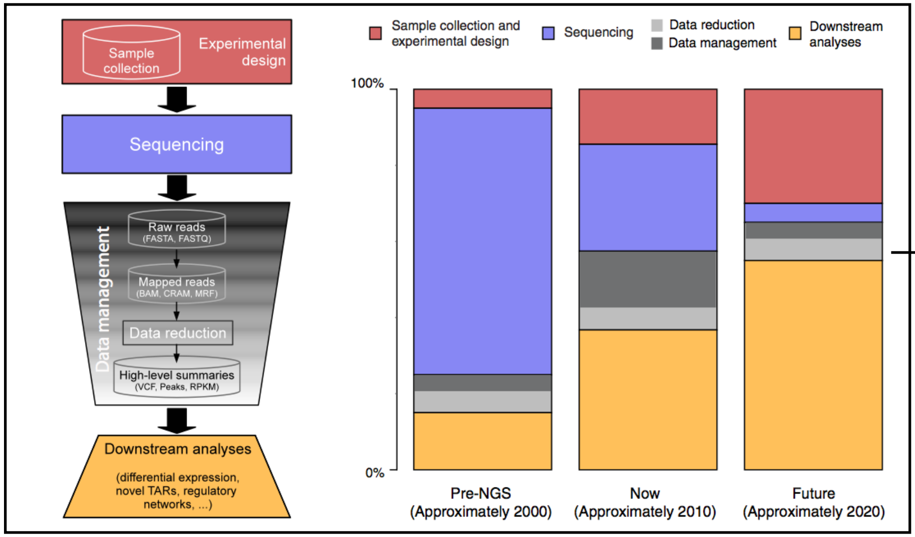


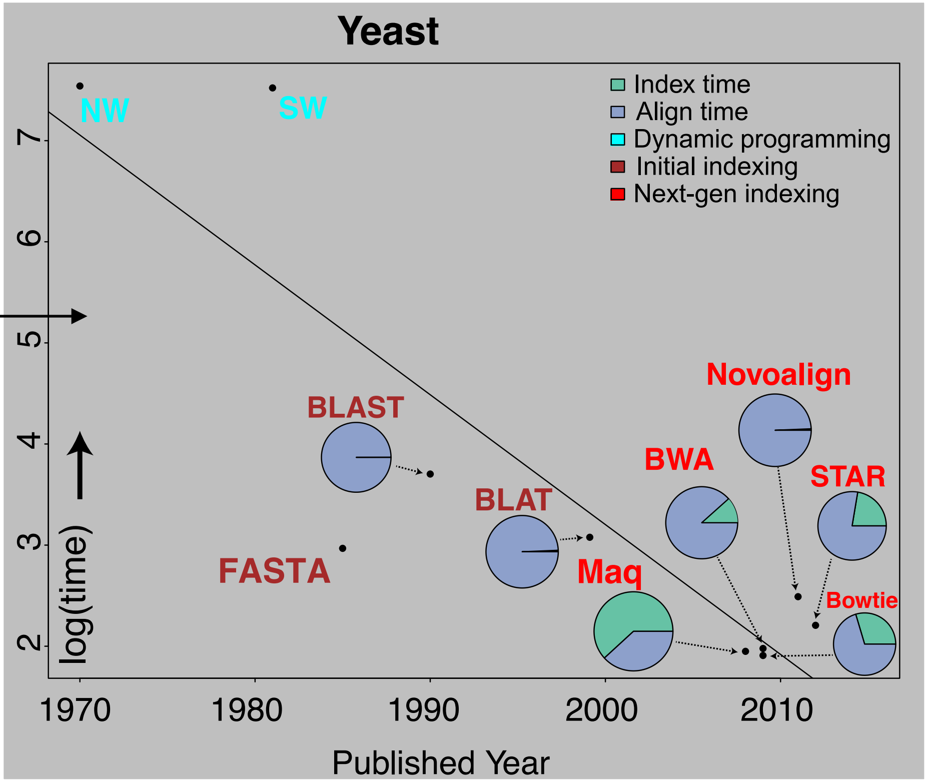
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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



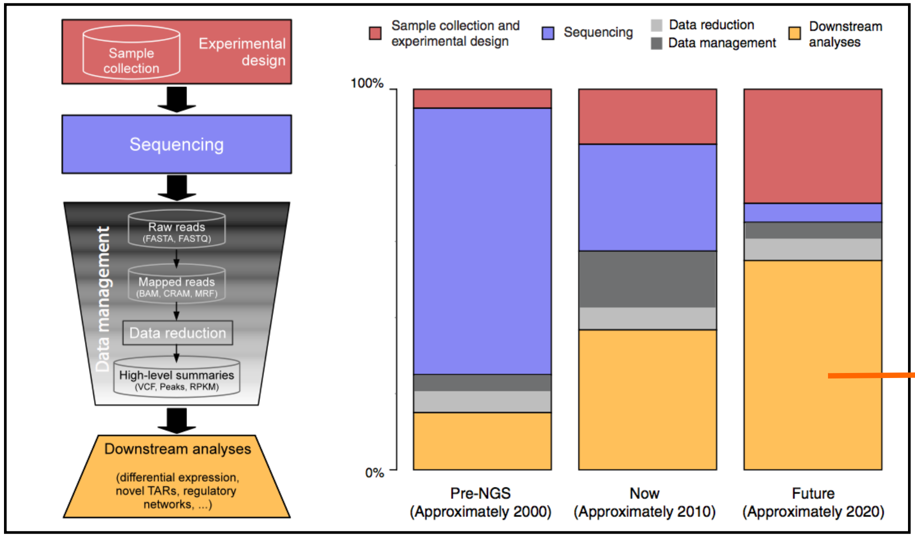


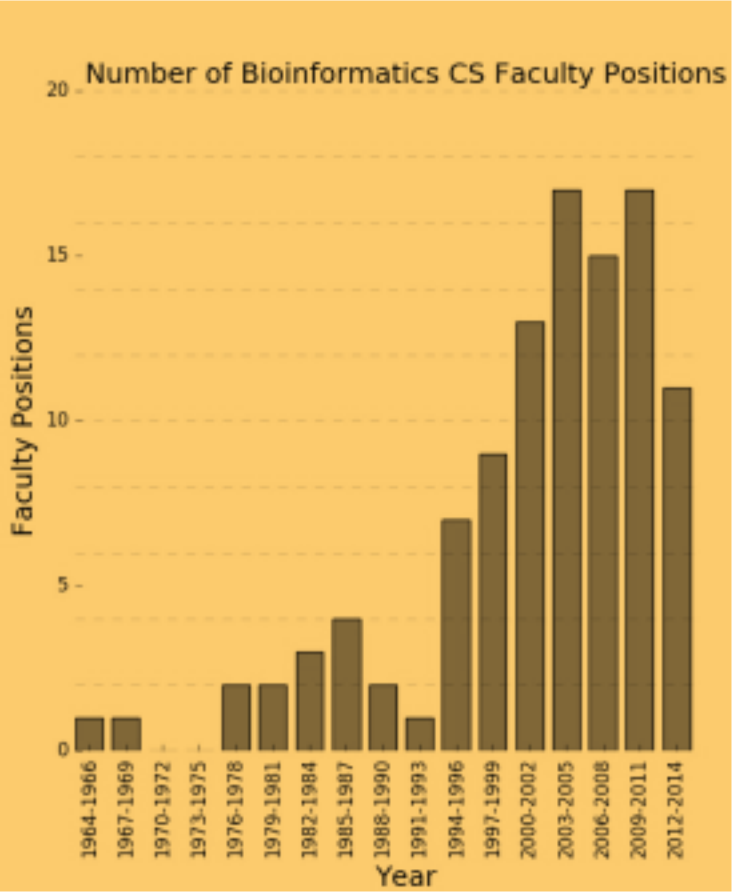
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

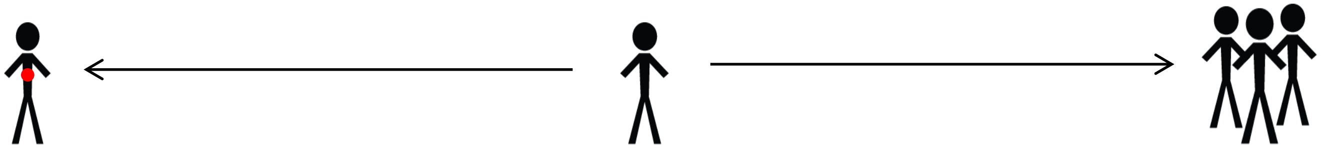
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

A Cancer Genome

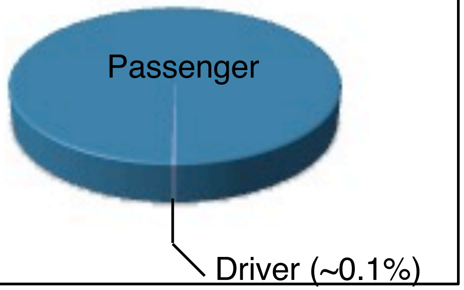
A Typical Genome

Population of
2,504 peoples



Origin of Variants

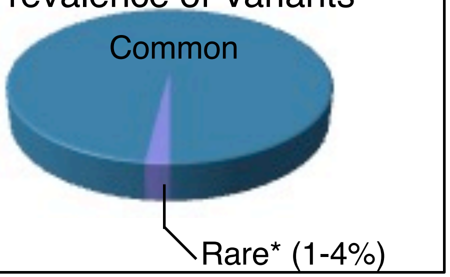
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



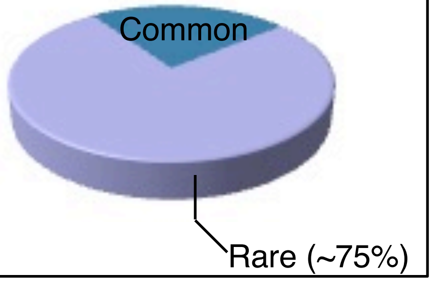
Class of Variants

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

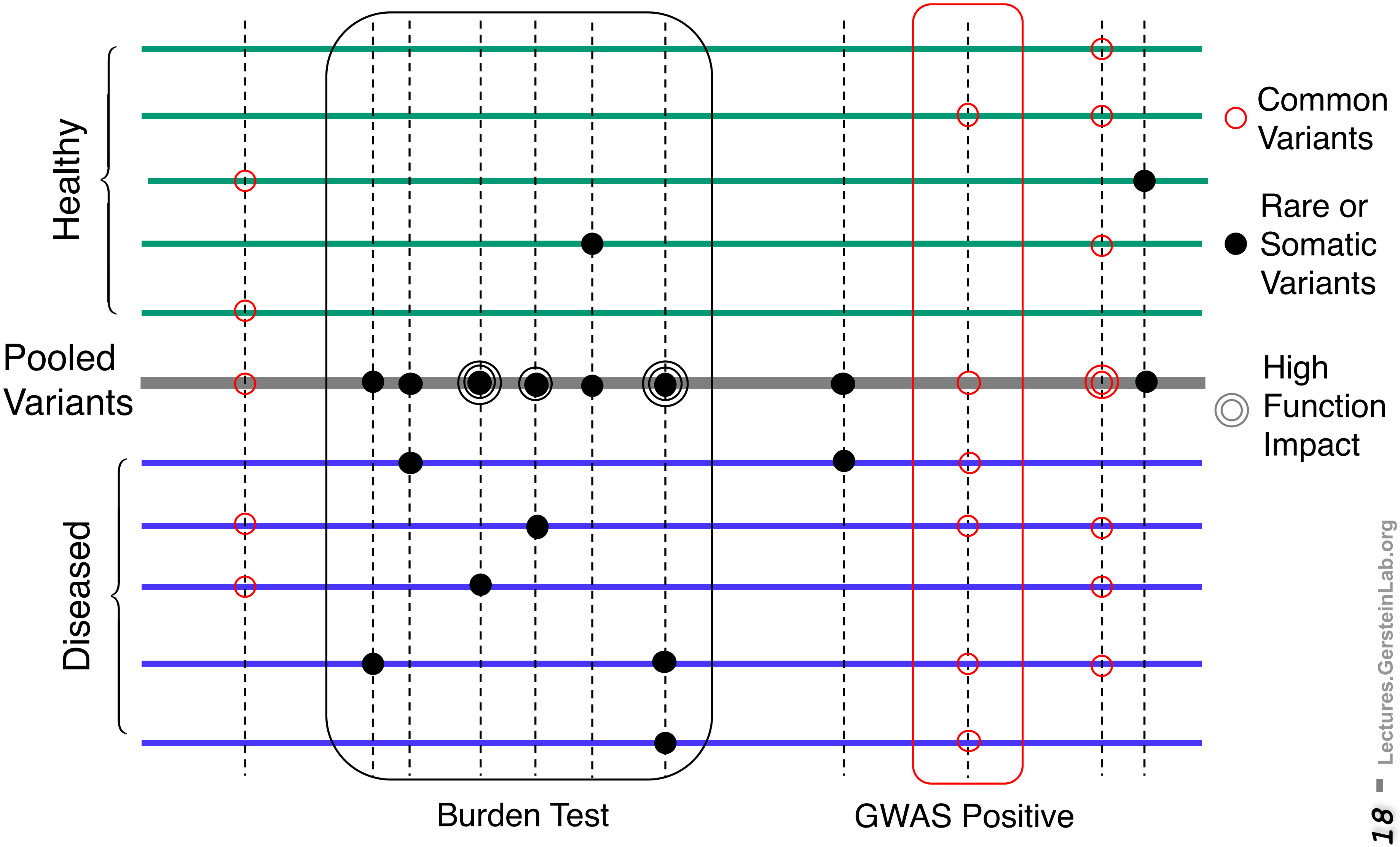
Somatic

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

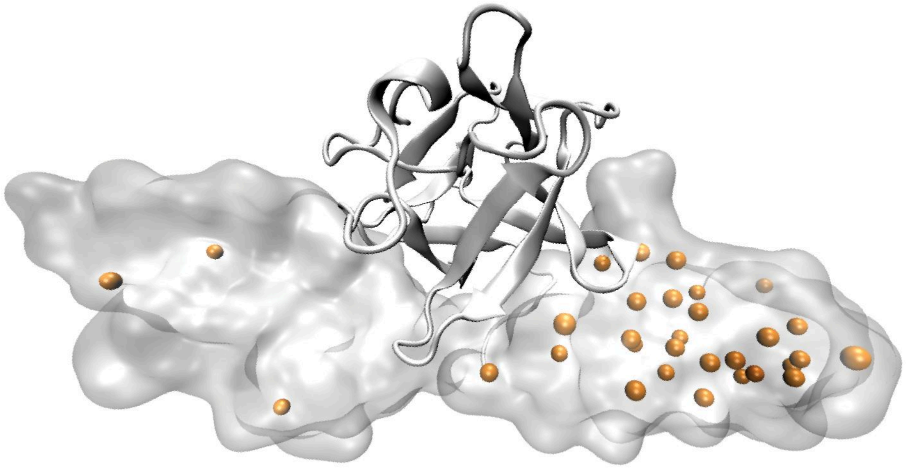


Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

- 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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

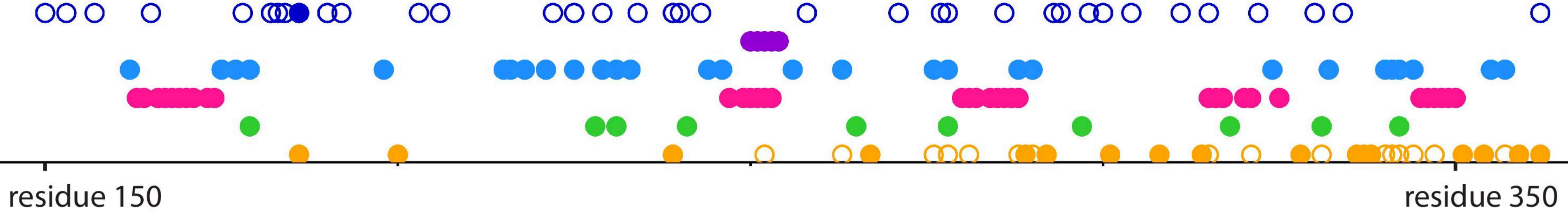
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

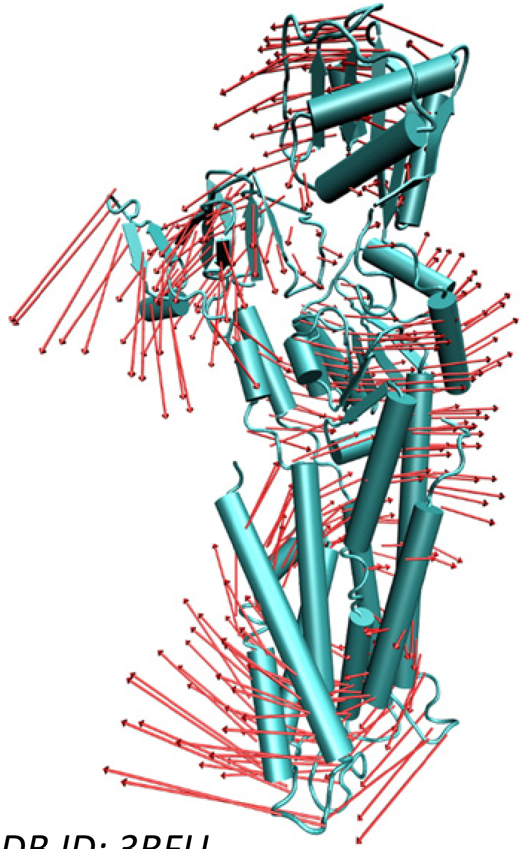
- ● ○ 1000G & ExAC SNVs (common | rare)
- ● Hinge residues
- ● Buried residues
- ● Protein-protein interaction site
- ● Post-translational modifications
- ● HGMD site (w/o annotation overlap)
- ○ HGMD site (w/annotation overlap)



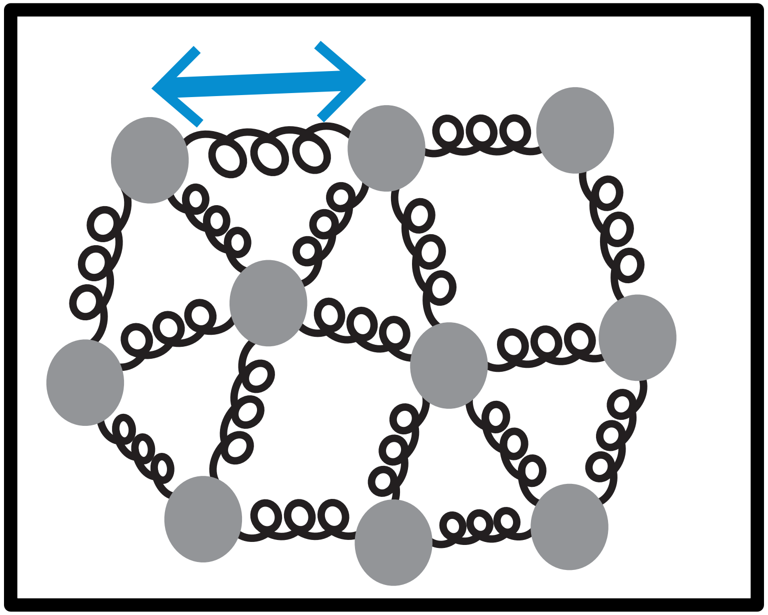
[Sethi et al. COSB ('15)]

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



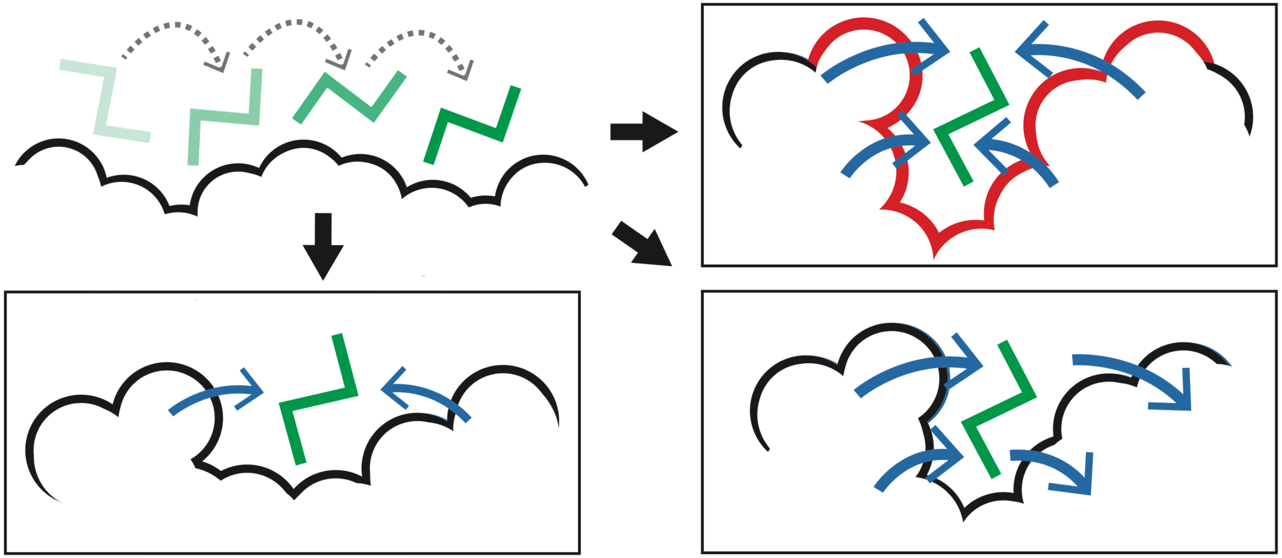
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



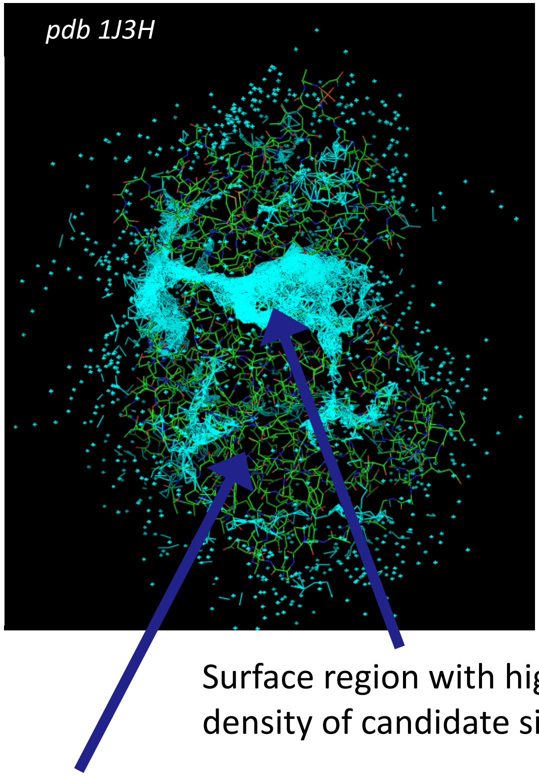
Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion

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



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

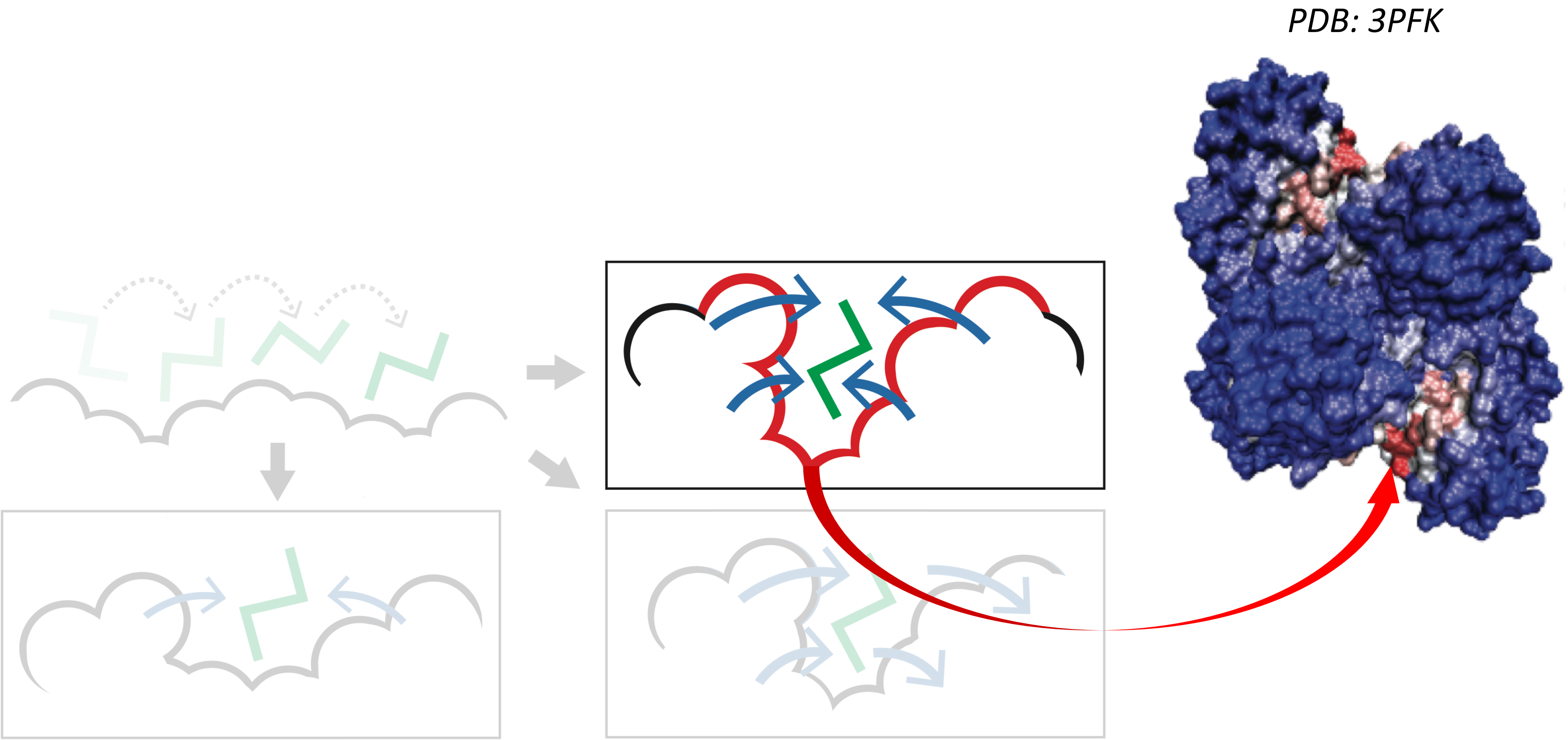


Surface region with high density of candidate sites

Surface region with low density of candidate sites

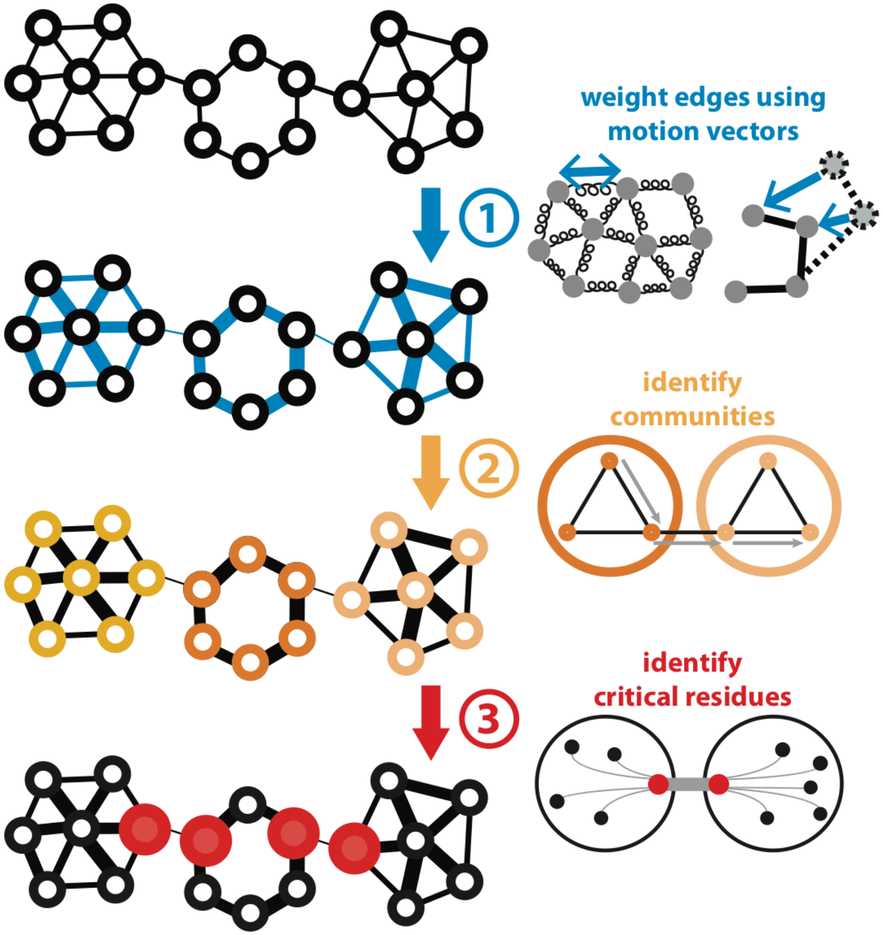
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface



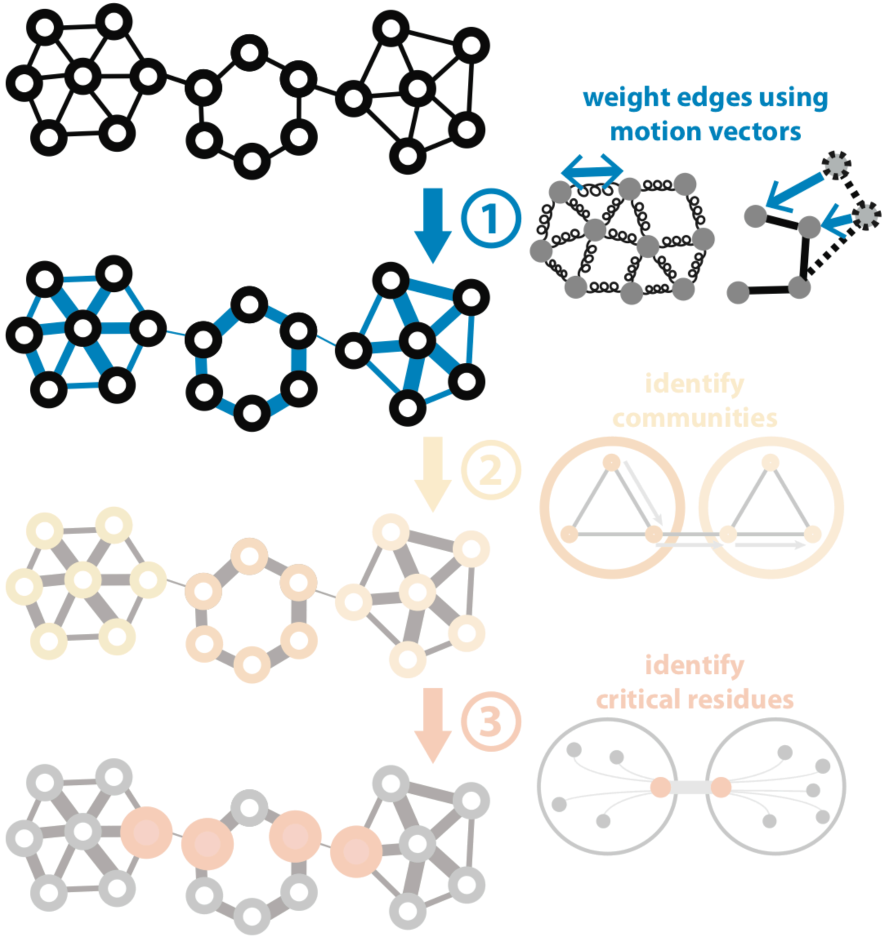
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



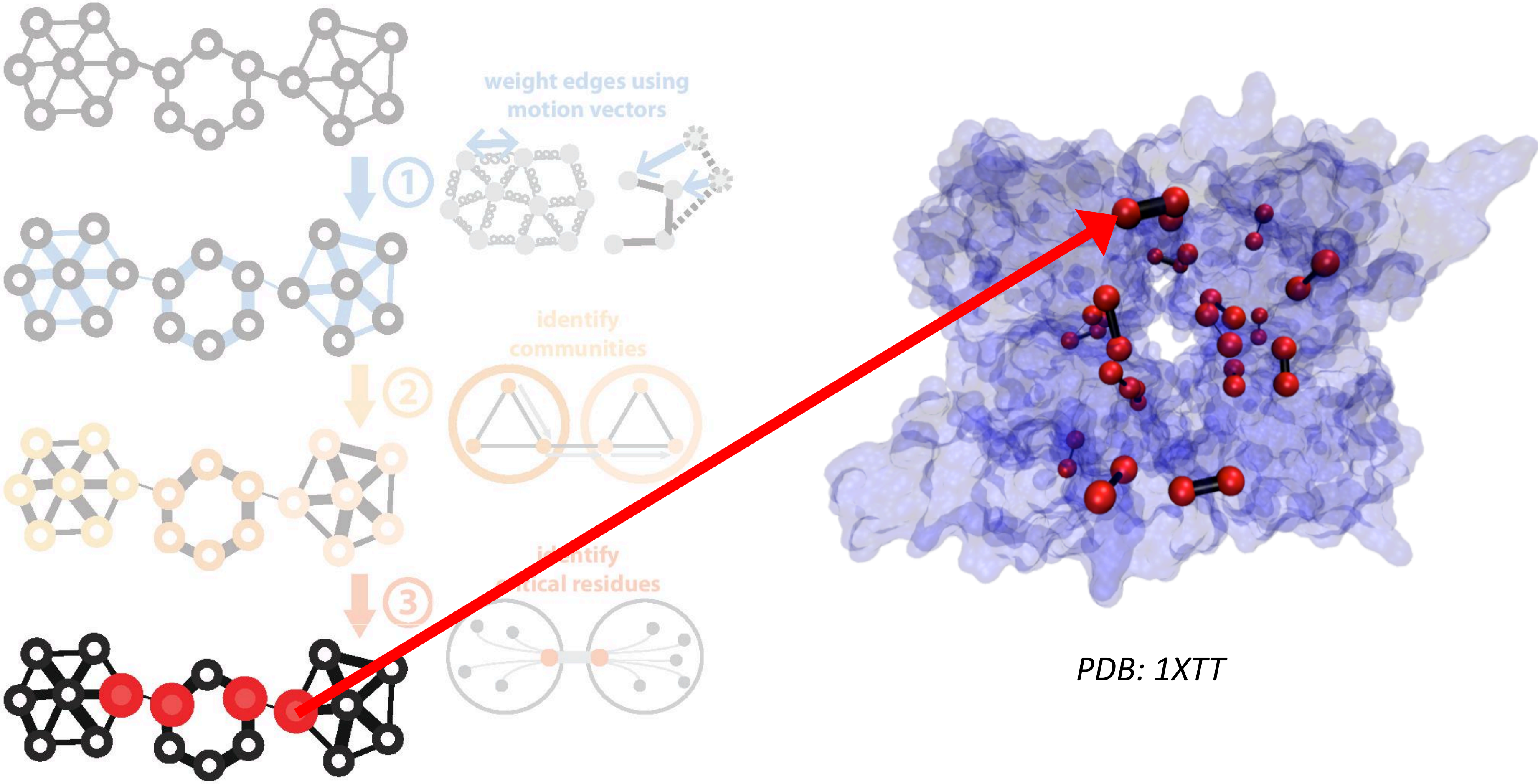
$$Cov_{ij} = \langle \mathbf{r}_\mathrm{i} \bullet \mathbf{r}_\mathrm{j} \rangle$$

$$C_{ij} = Cov_{ij} \ / \ \surd(\langle \mathbf{r}_\mathrm{i}{}^2\rangle\langle \mathbf{r}_\mathrm{j}{}^2\rangle)$$

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

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

Predicting Allosterically-Important Residues within the Interior

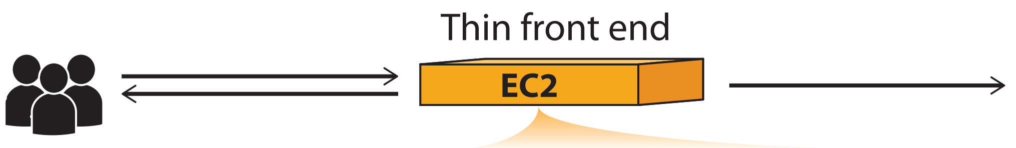


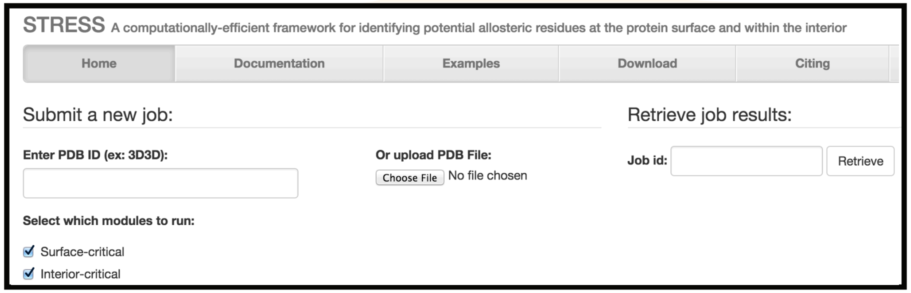
PDB: 1XTT

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

STRESS Server Architecture: Highlights

stress.molmovdb.org





STRESS A computationally-efficient framework for identifying potential allosteric residues at the protein surface and within the interior

Submit a new job:

Enter PDB ID (ex: 3D3D):

Or upload PDB File: Choose File No file chosen

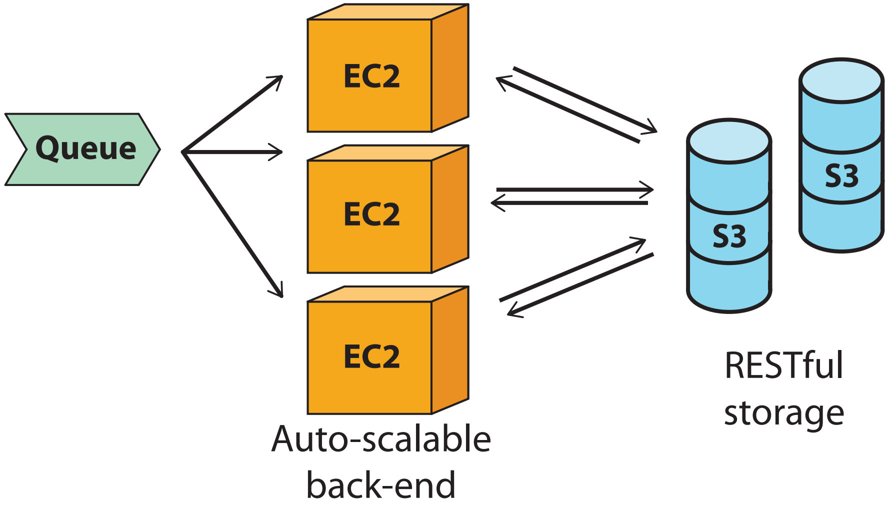
Select which modules to run:

☑ Surface-critical

☑ Interior-critical

Retrieve job results:

Job id: Retrieve



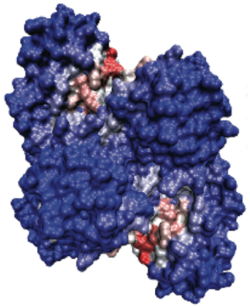
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

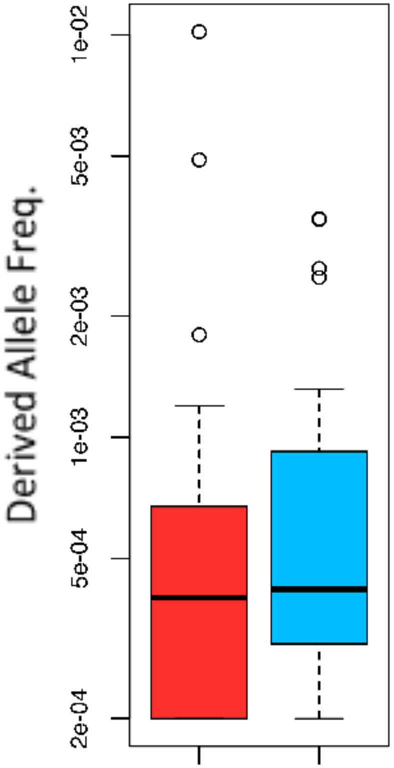
Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

1000 Genomes

Surface



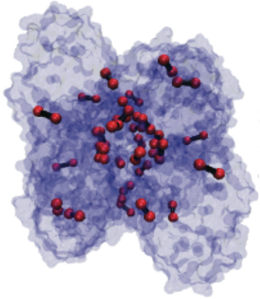


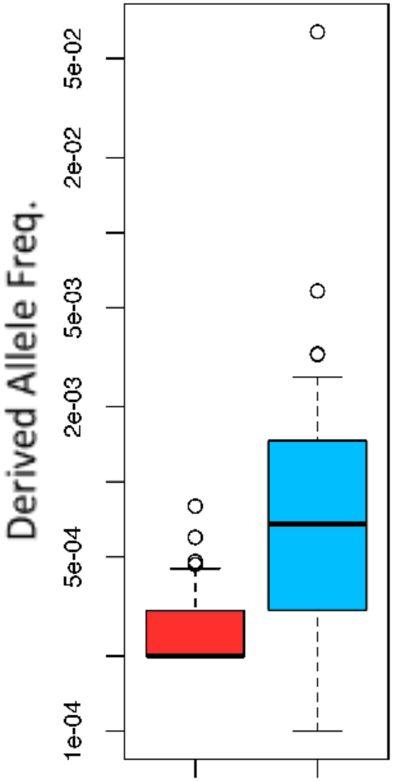
p=0.309

🟥 critical

🟦 non-critical

Interior





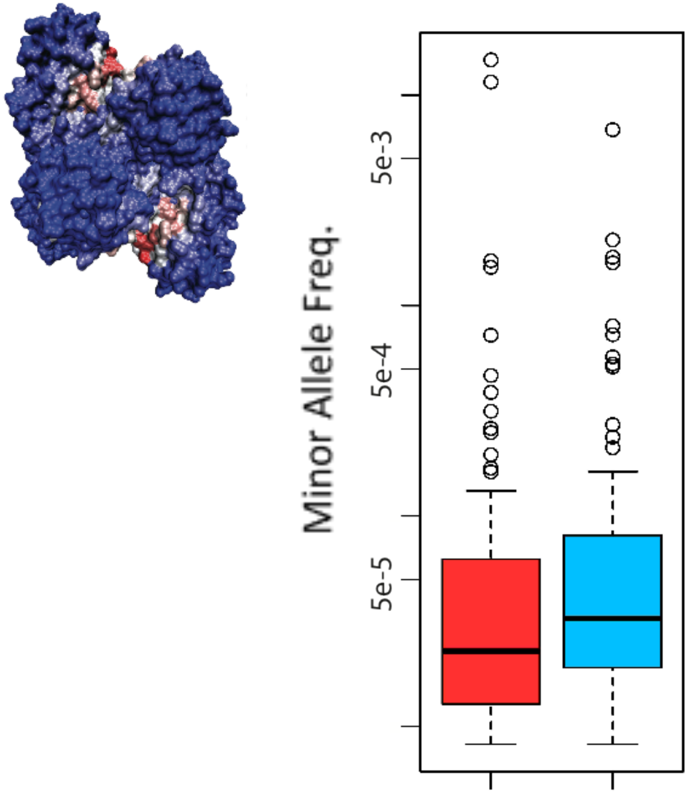
p=1.80e-05

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

Intra-species conservation of predicted allosteric residues

ExAC

Surface

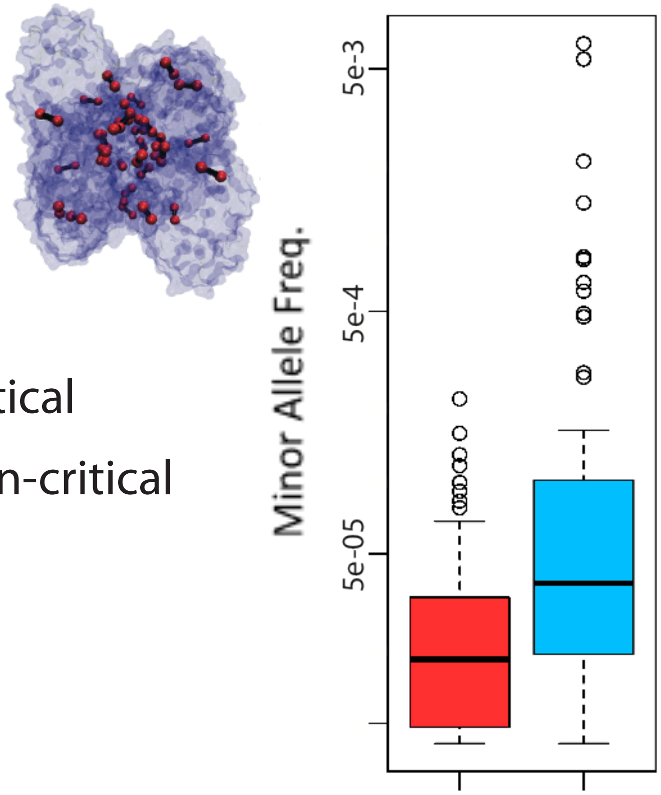


p=1.49e-3

🟥 critical

🟦 non-critical

Interior

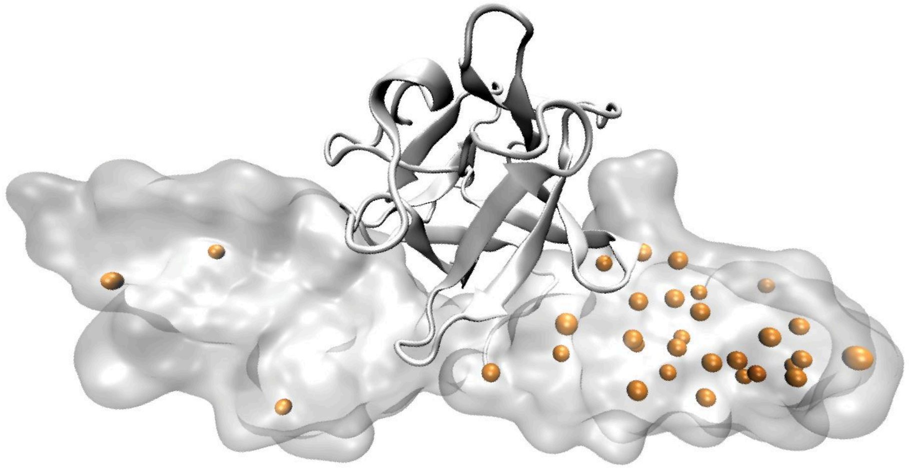


p=7.98e-09

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

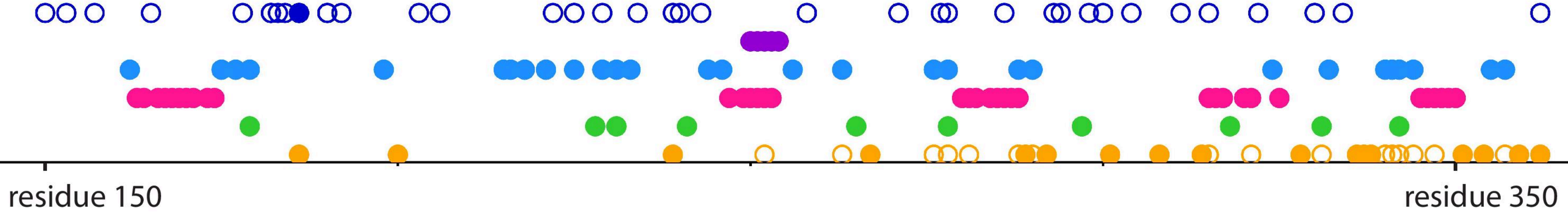
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

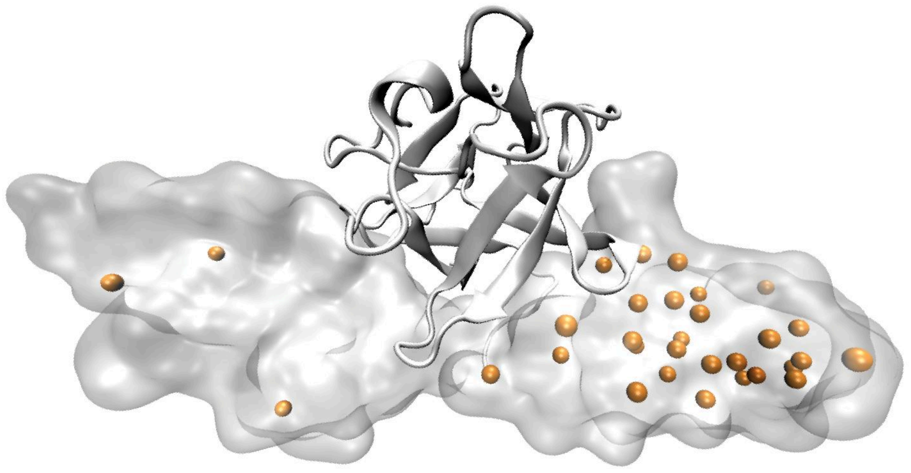
●○ 1000G & ExAC SNVs (common | rare)
● Hinge residues
● Buried residues
● Protein-protein interaction site
● Post-translational modifications
● HGMD site (w/o annotation overlap)
○ HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

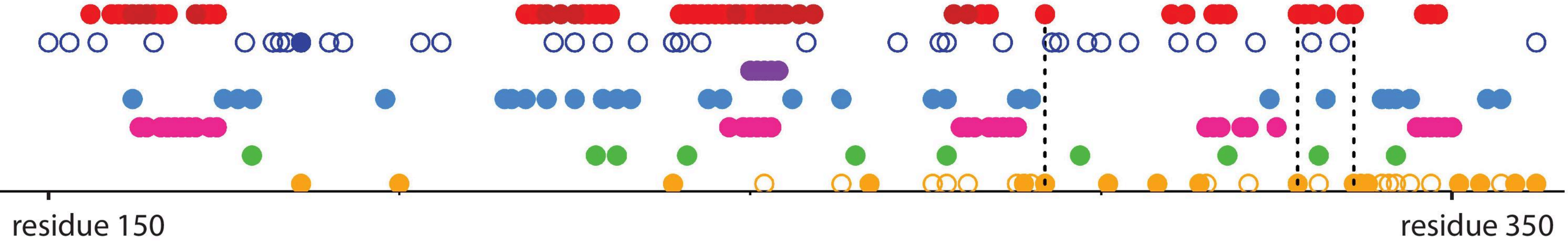
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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)

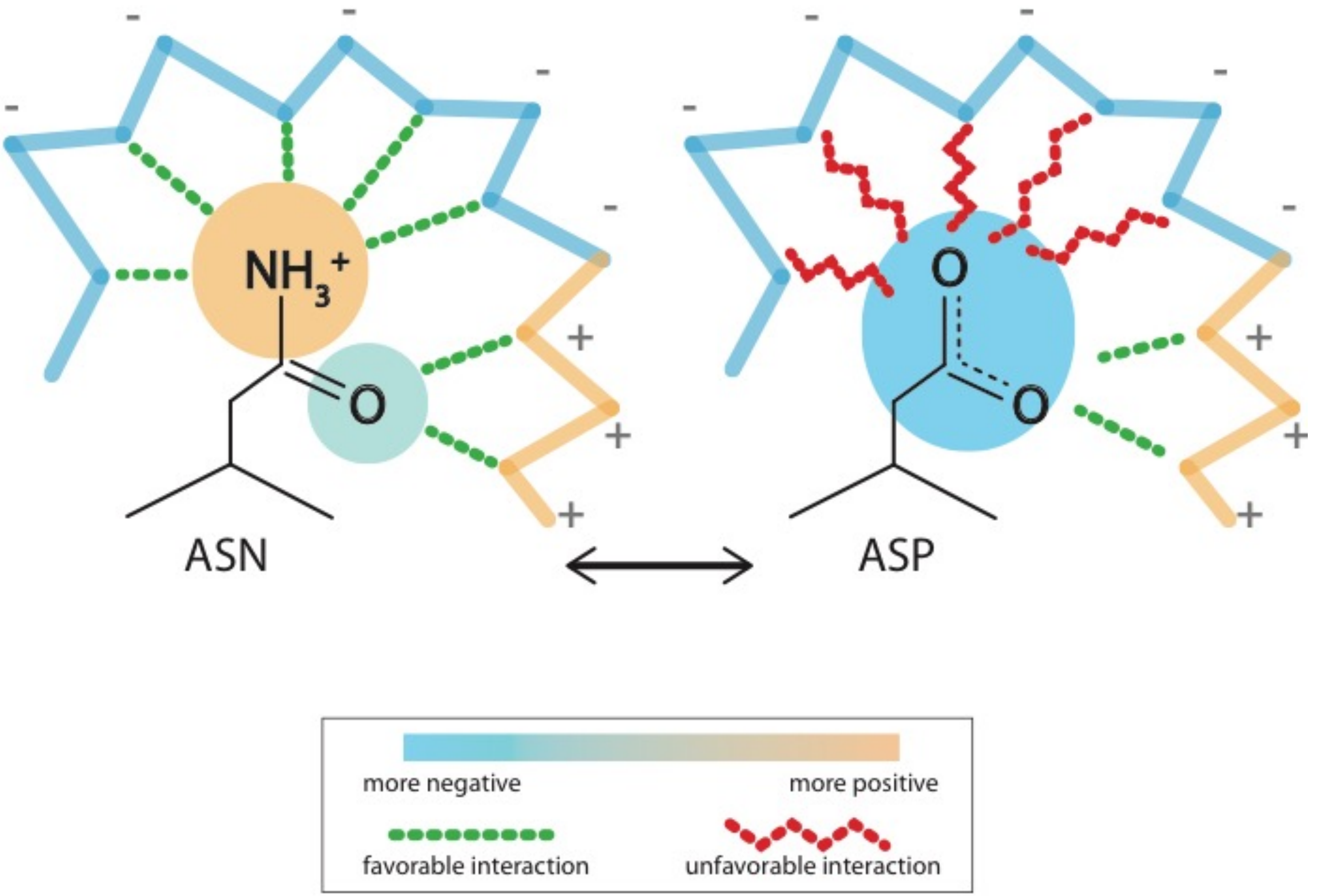


[Sethi et al. COSB ('15)]

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

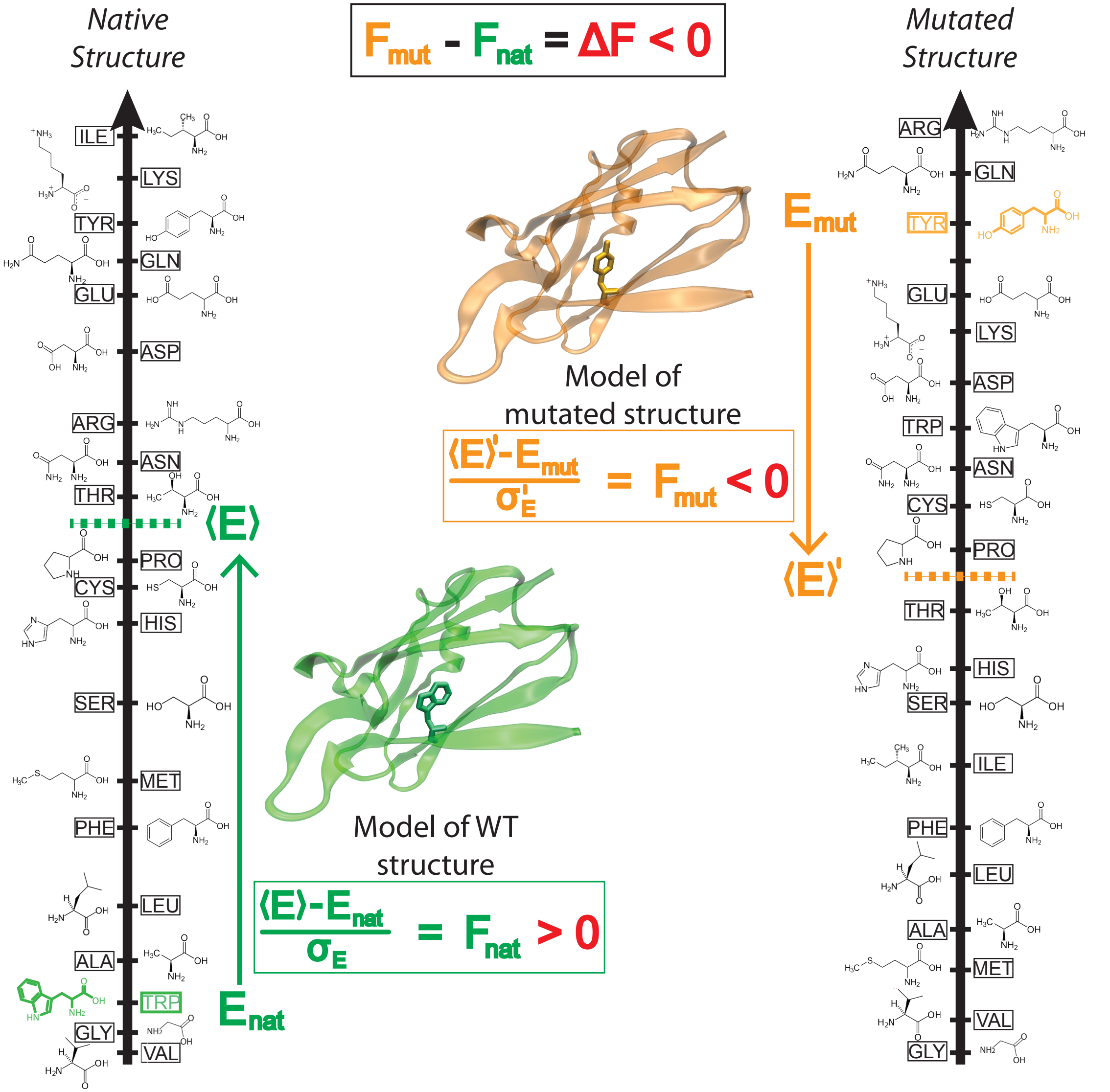
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

Schematic illustration of localized frustration



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

Workflow for evaluating localized frustration changes (ΔF)



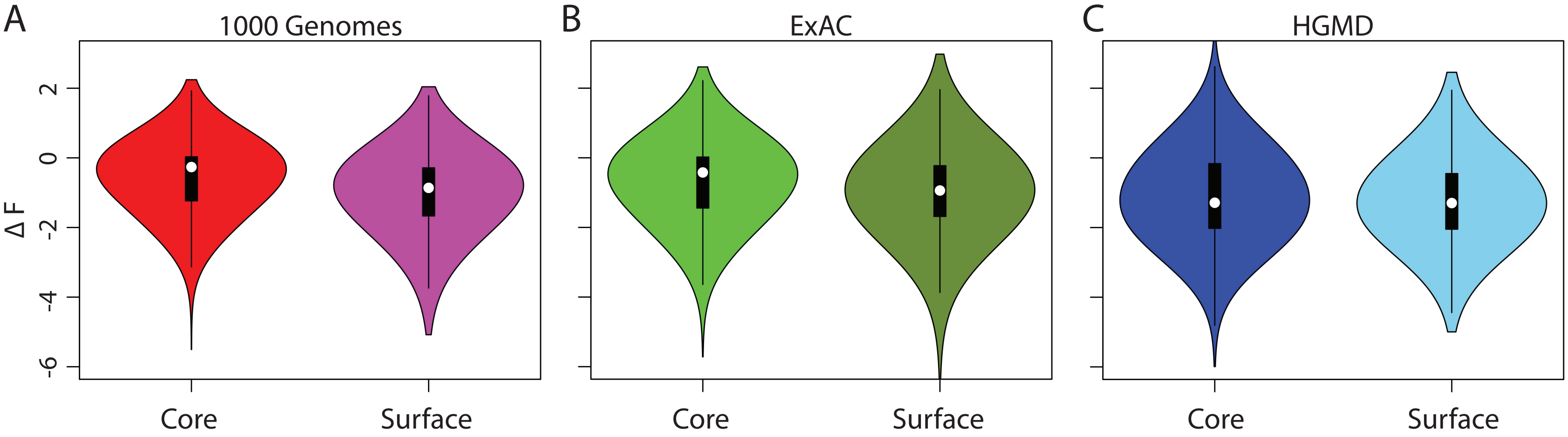
$F_{mut} - F_{nat} = \Delta F < 0$

$$\frac{\langle E\rangle' - E_{mut}}{\sigma'_E} = F_{mut} < 0$$

$$\frac{\langle E\rangle - E_{nat}}{\sigma_E} = F_{nat} > 0$$

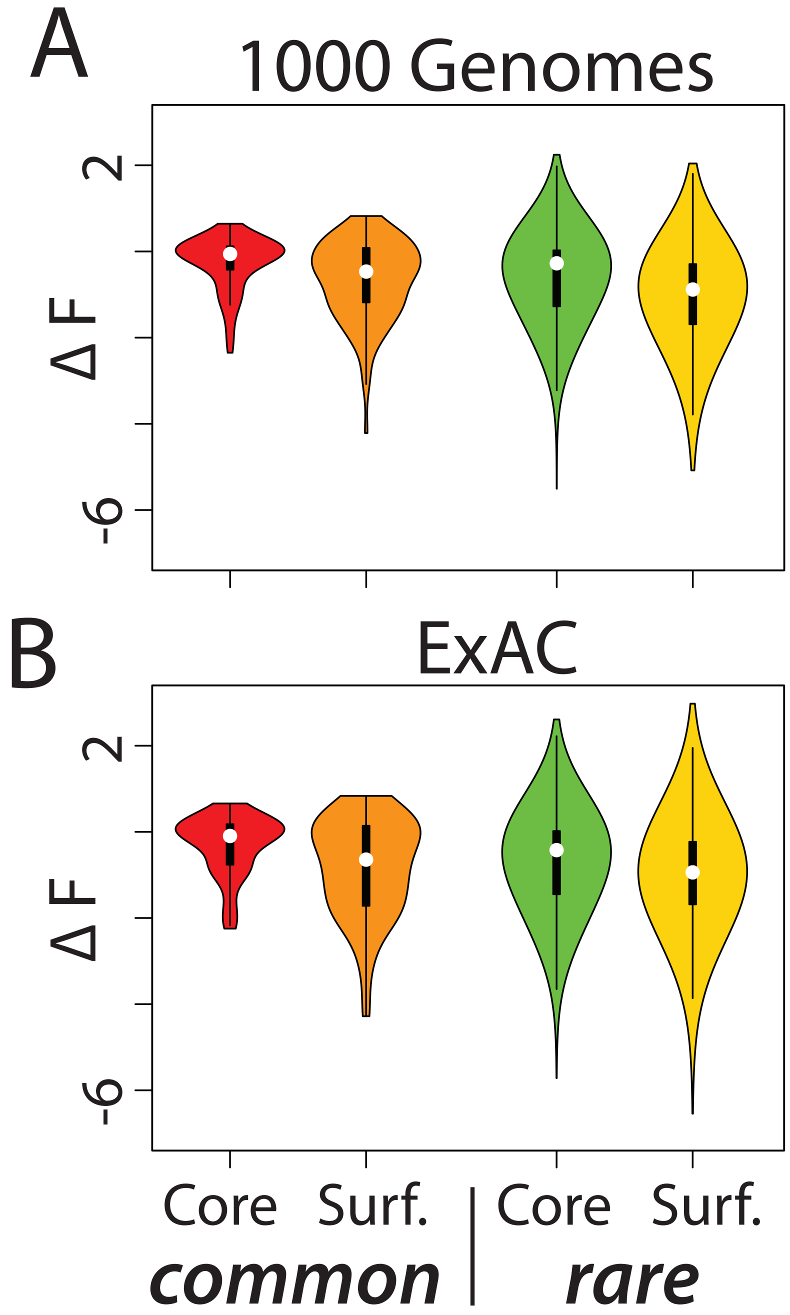
[Kumar et al. *NAR* (in press); *biorxiv* 052027]

Comparing Frustration (ΔF values) across different SNV categories



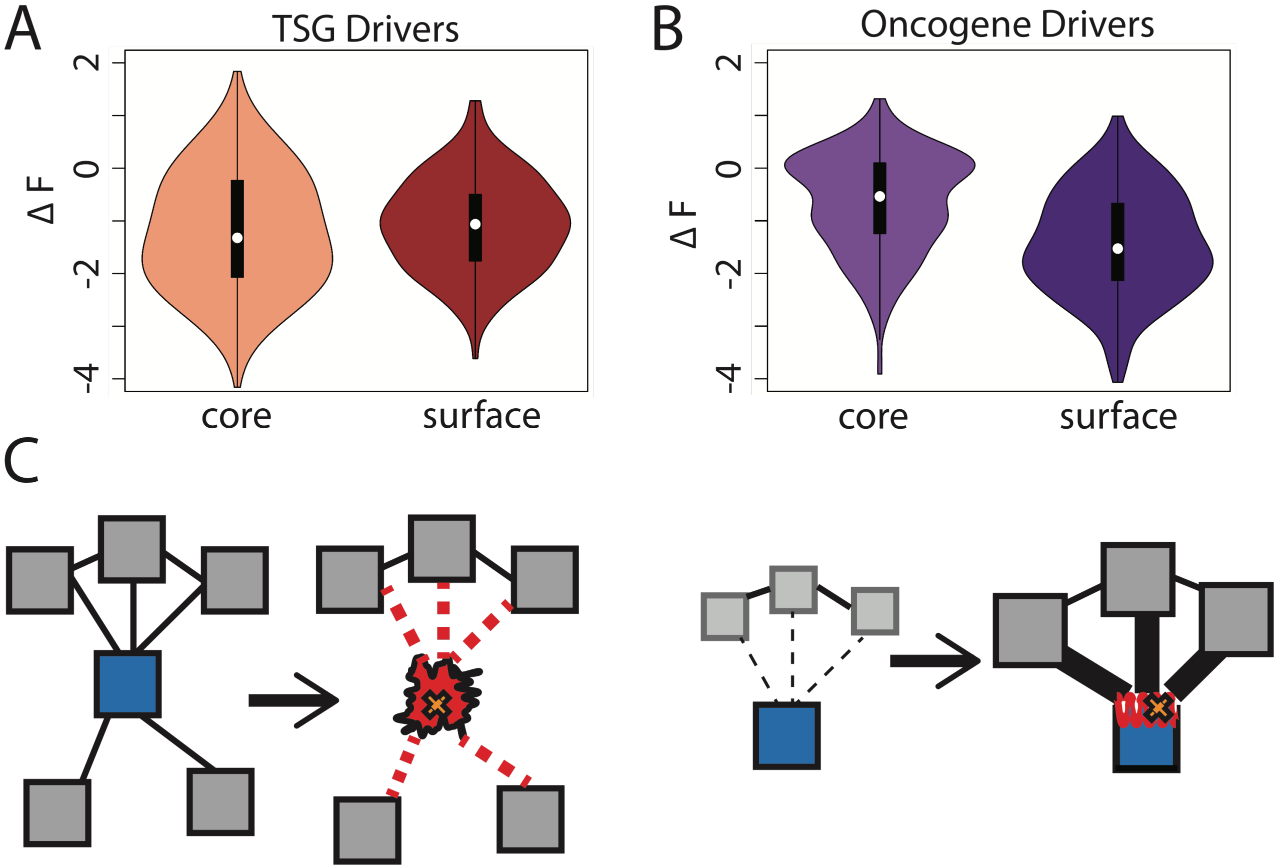
[Kumar et al. *NAR* (in press); *biorxiv* 052027]

∆F distributions among rare and common SNVs



[Kumar et al. *NAR* (in press); *biorxiv* 052027]

Comparison between ΔF distributions: TSGs vs. oncogenes



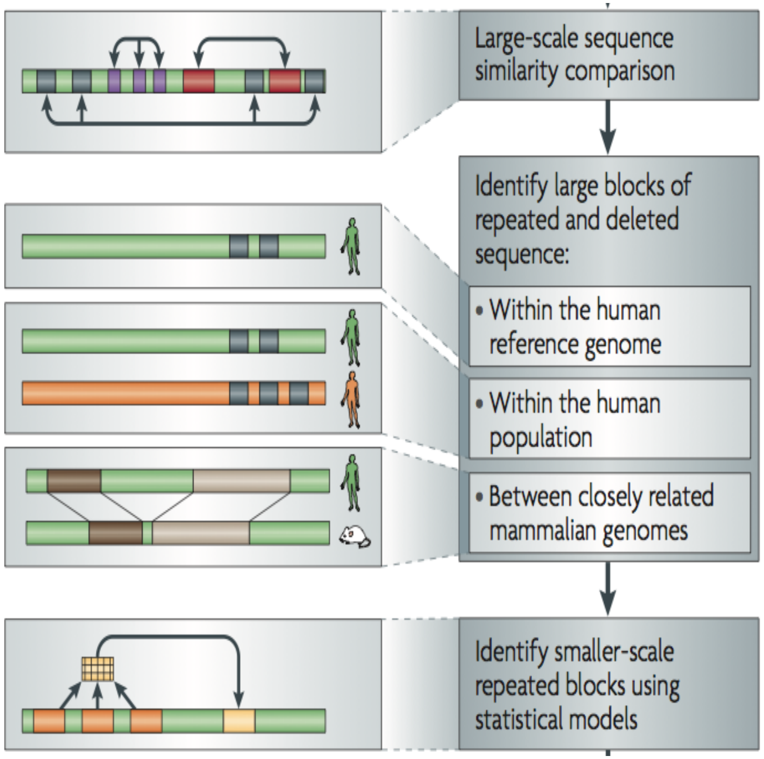
[Kumar et al. *NAR* (in press); *biorxiv* 052027]

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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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**



Large-scale sequence similarity comparison

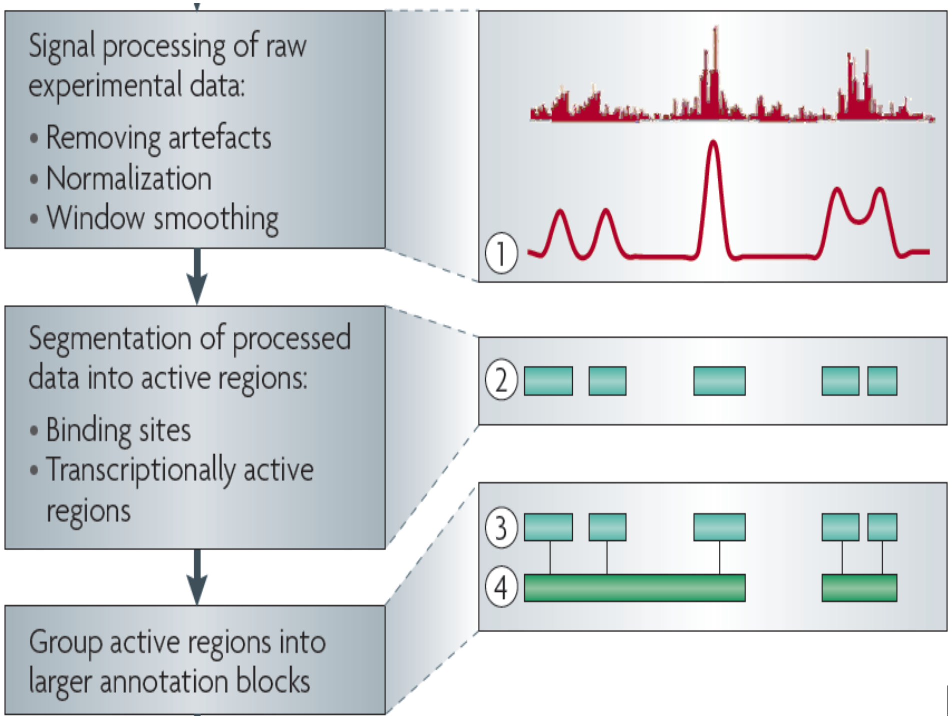
Identify large blocks of repeated and deleted sequence:

- Within the human reference genome
- Within the human population
- Between closely related mammalian genomes

Identify smaller-scale repeated blocks using statistical models

Functional Genomics

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



Signal processing of raw experimental data:

- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:

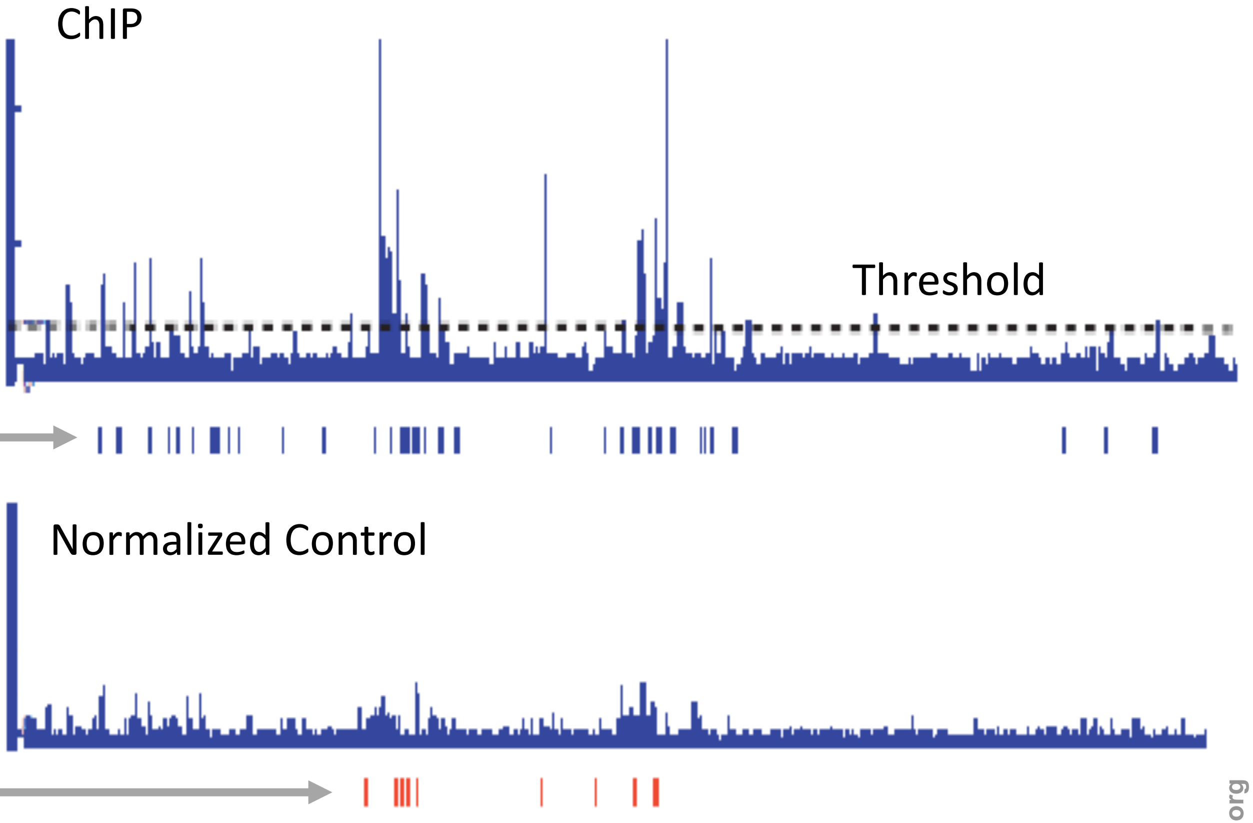
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

[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

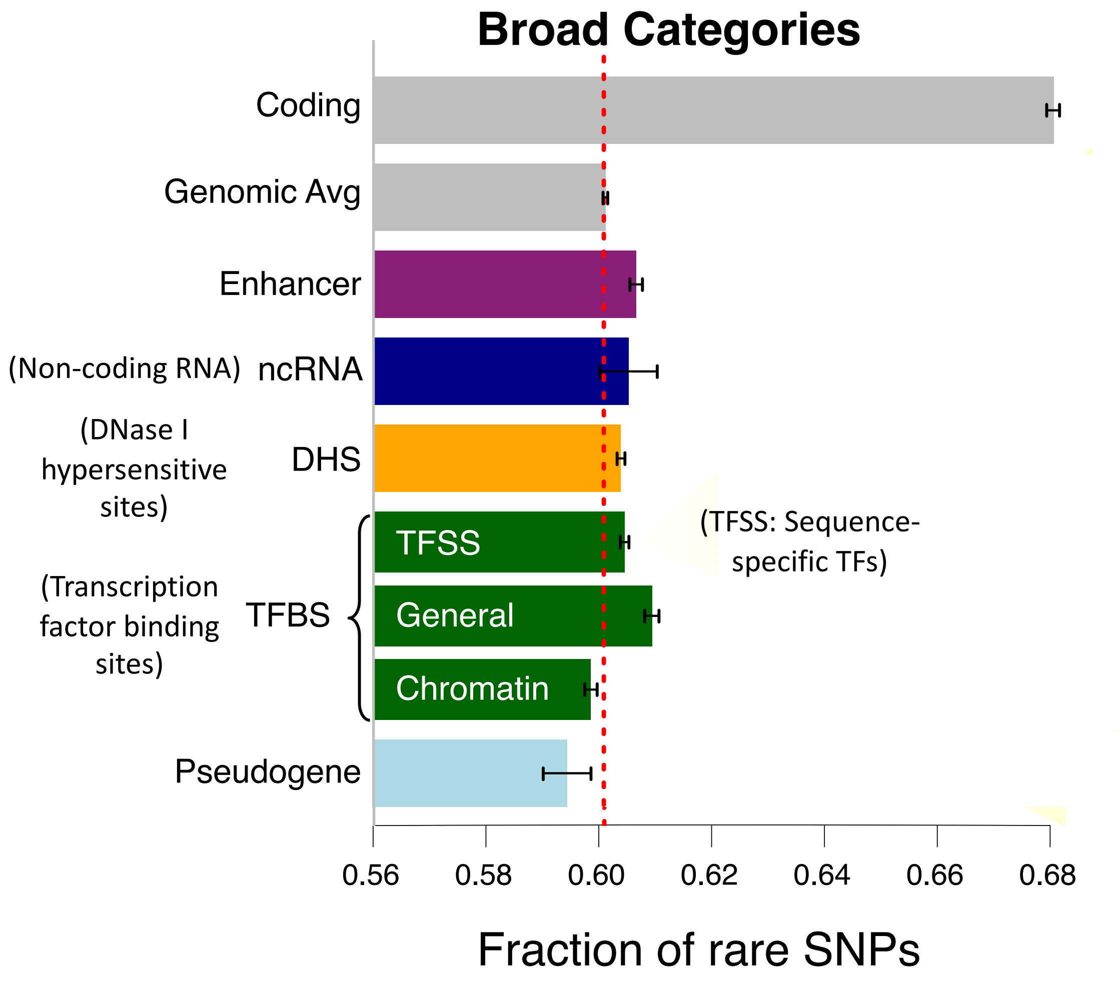
Significantly Enriched targets

Now an update: "PeakSeq 2" => MUSIC

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

Finding "Conserved" Sites in the Human Population:

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



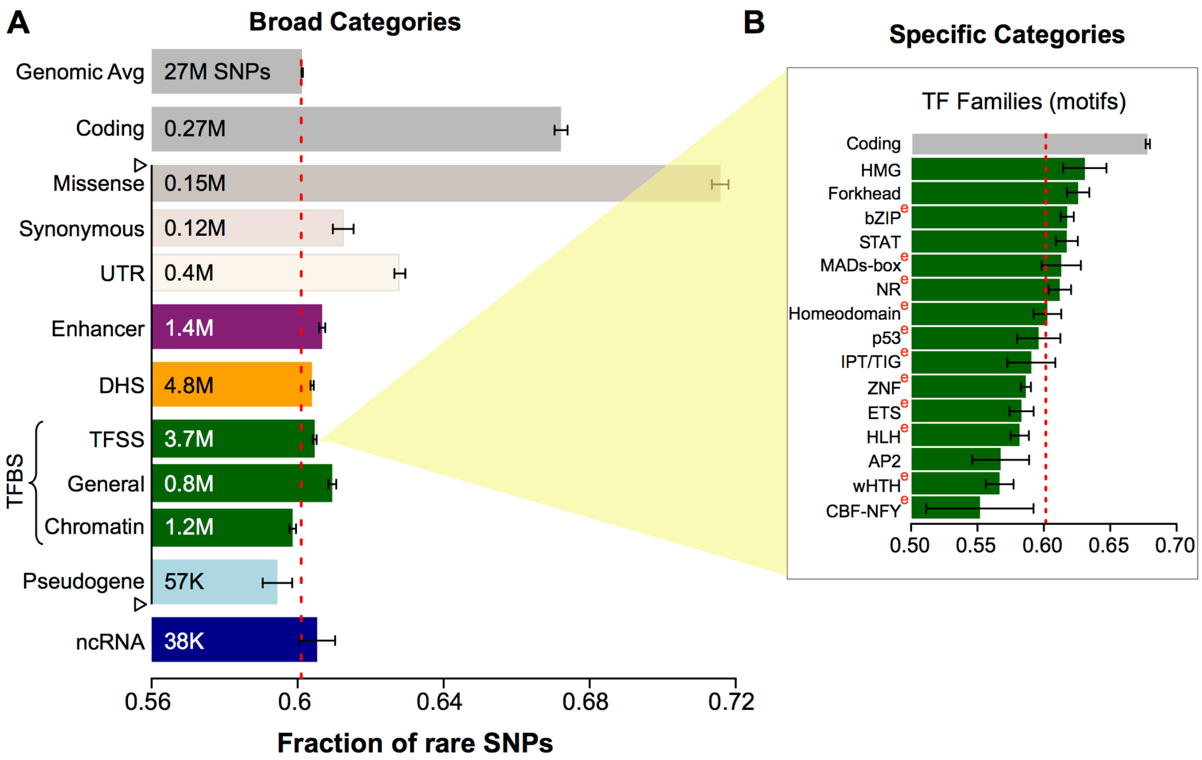
Depletion of Common Variants in the Human Population

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

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

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

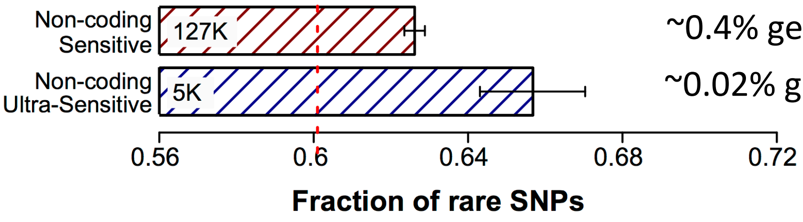
Differential selective constraints among specific sub-categories



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

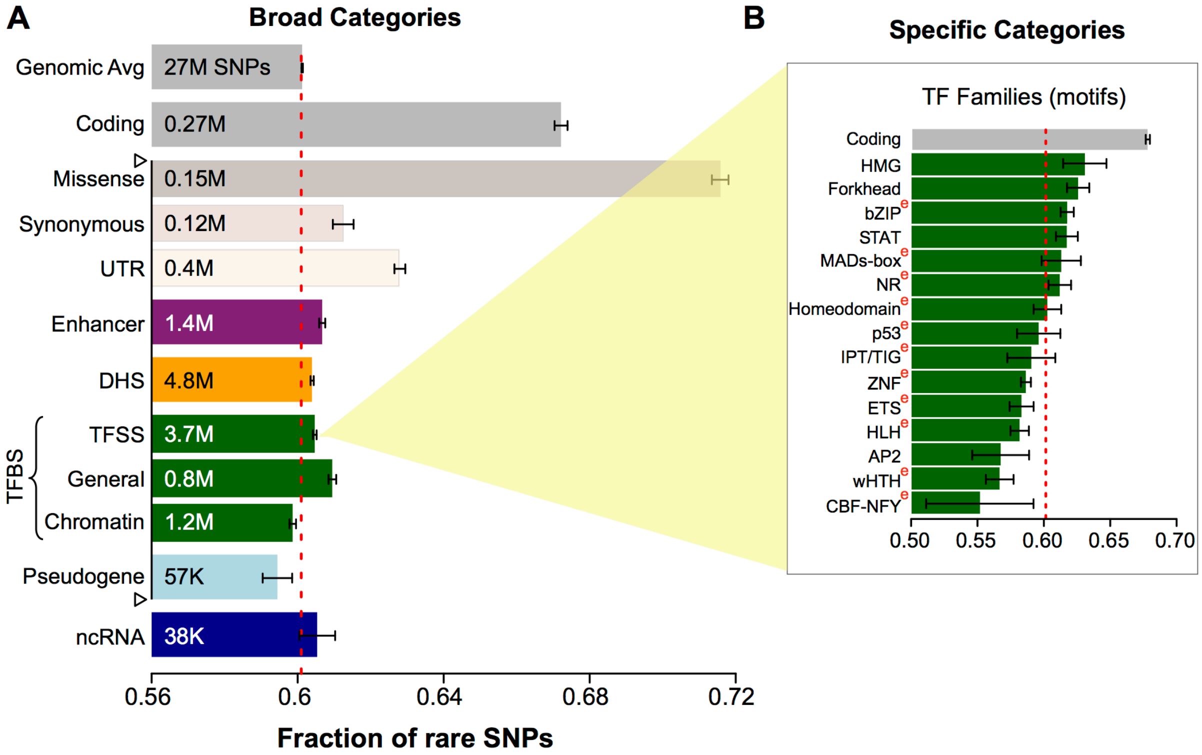
[Khurana et al., *Science* ('13)]

Defining Sensitive non-coding Regions



~0.4% genomic coverage (~ top 25)

~0.02% genomic coverage (top 5)



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

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

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

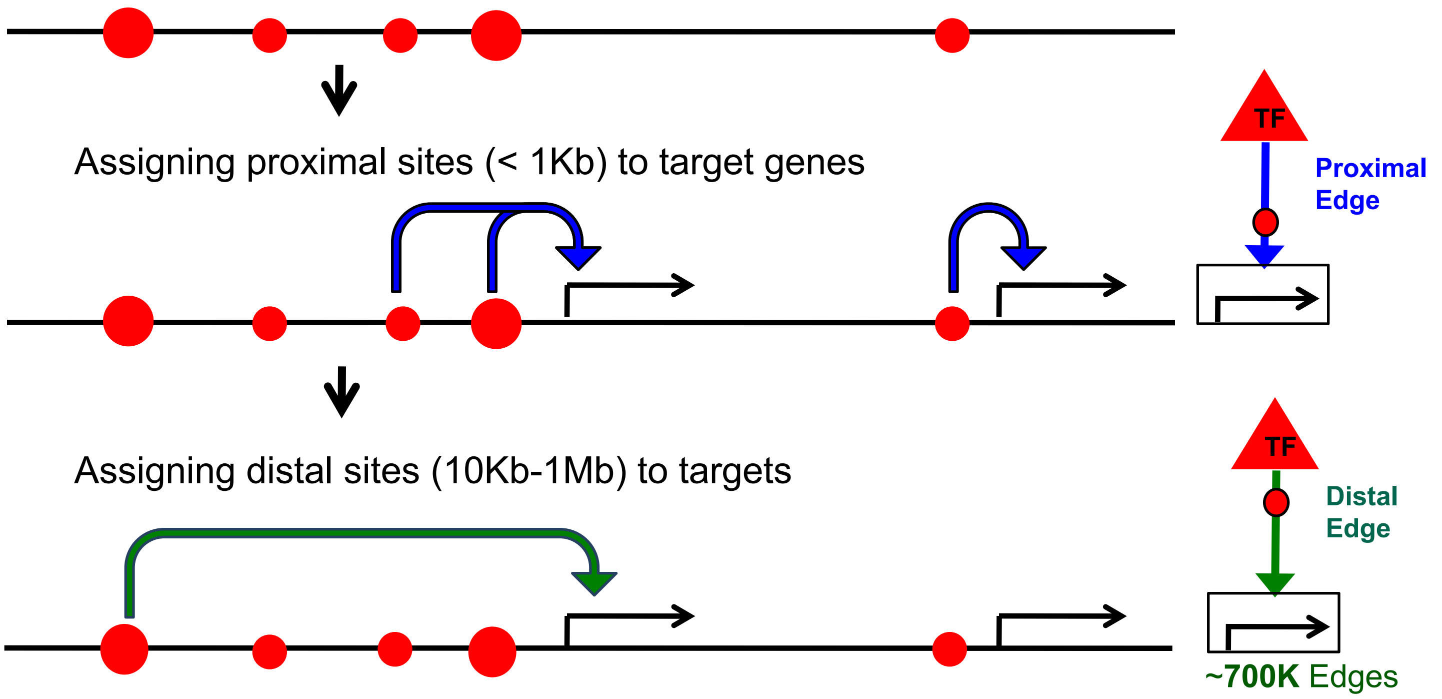
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

- 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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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)]

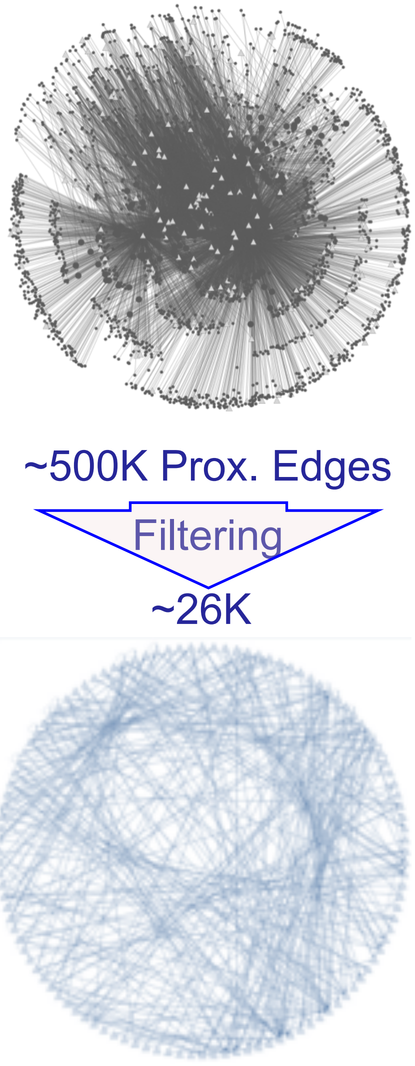
Regulatory elements

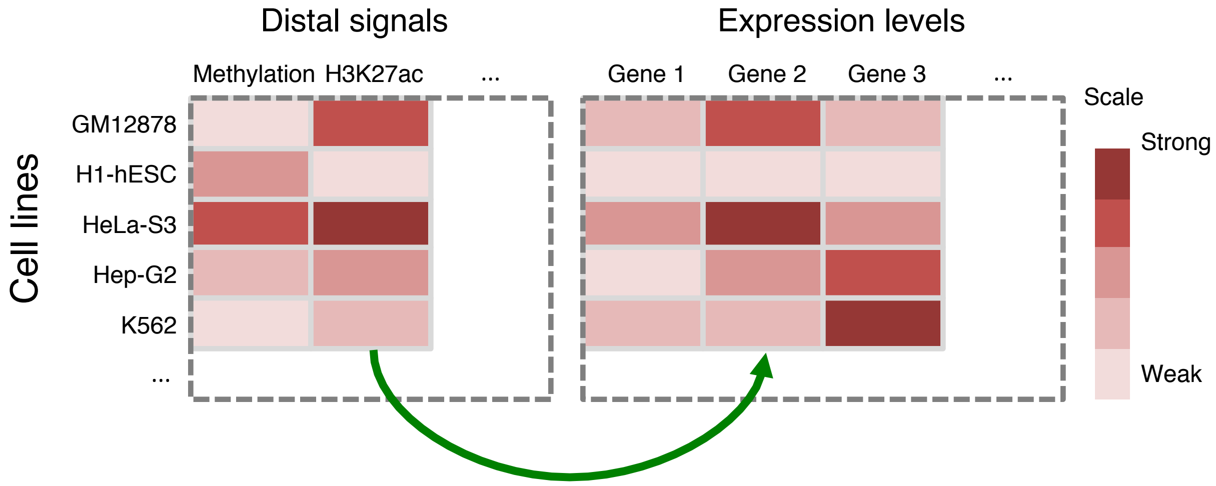


Assigning proximal sites (< 1Kb) to target genes

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

~700K Edges

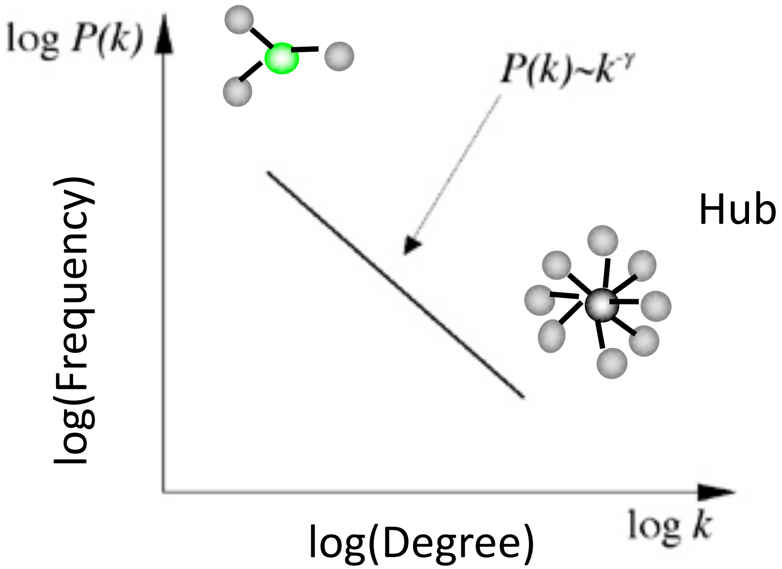




Connecting Distal Elements via **Activity Correlations**.

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

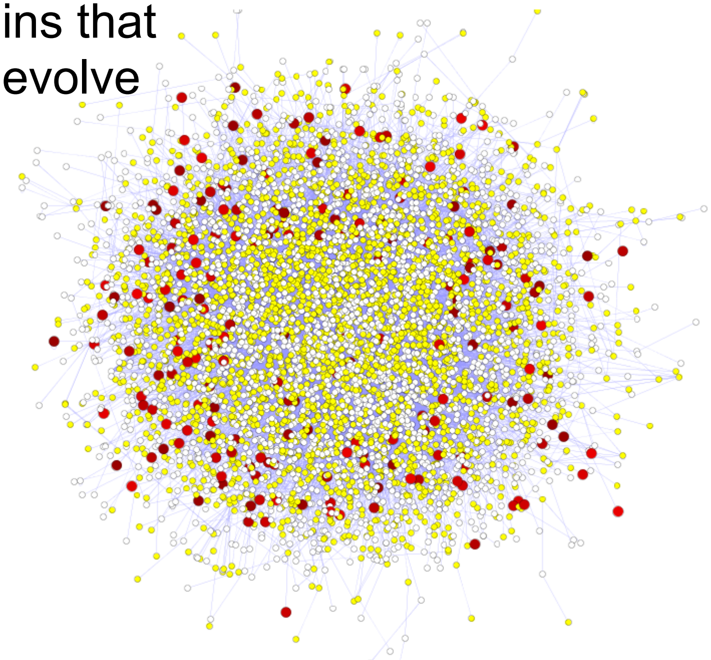
● **High likelihood of positive selection**

● **Lower likelihood of positive selection**

● **Not under positive selection**

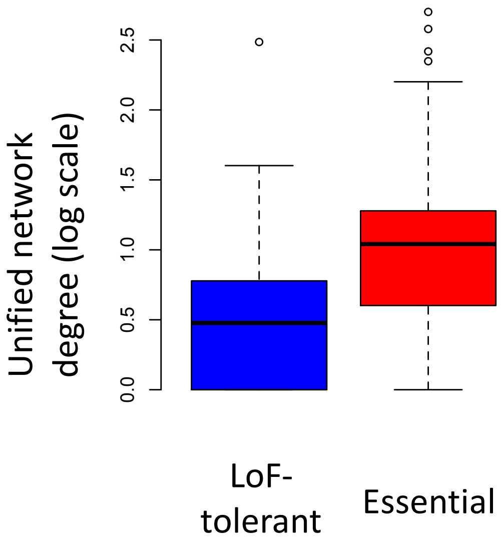
○ **No data about positive selection**

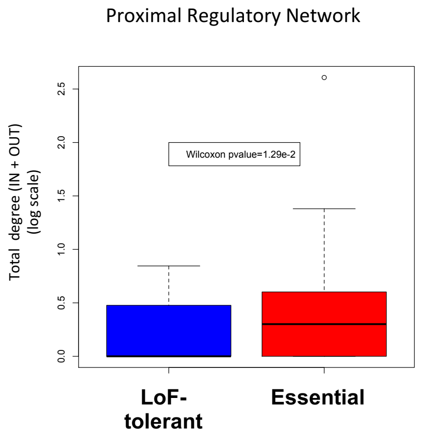
[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

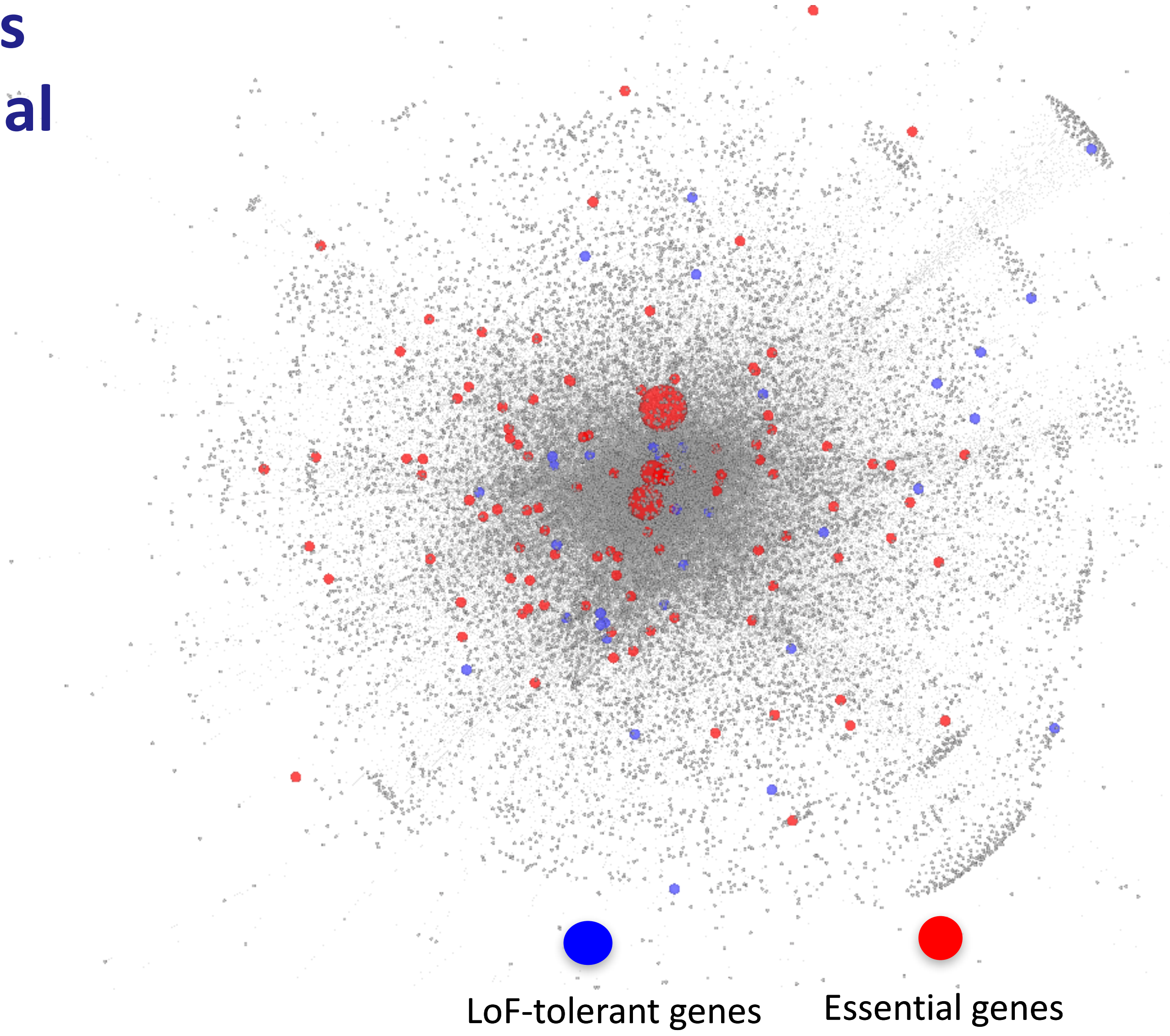


- <u>More Connectivity, More Constraint:</u> Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
  - **yeast PPI** - Fraser et al ('02) Science,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

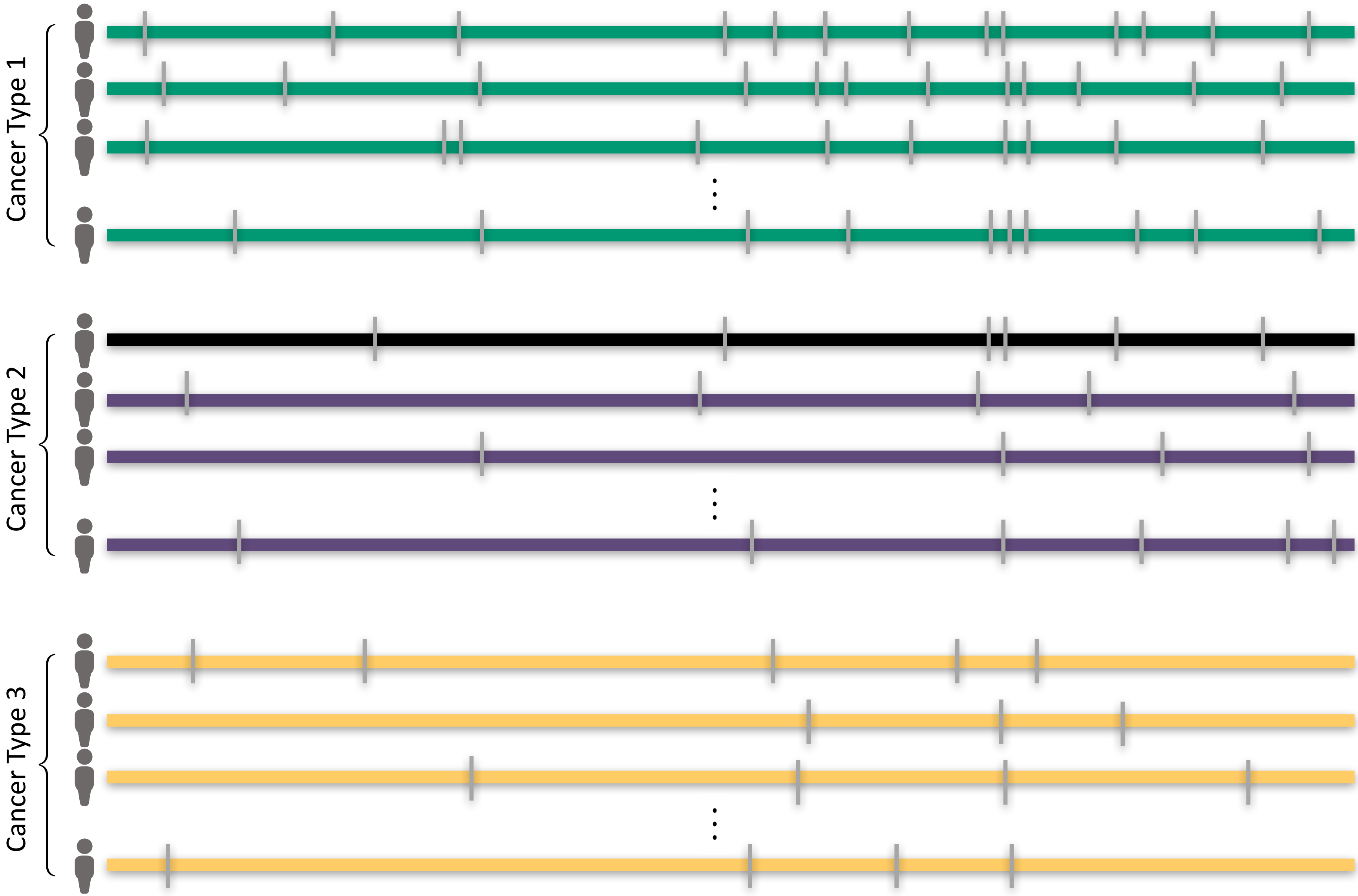
[Khurana et al., *PLOS Comp. Bio.* '13]

Size of nodes scaled by total degree

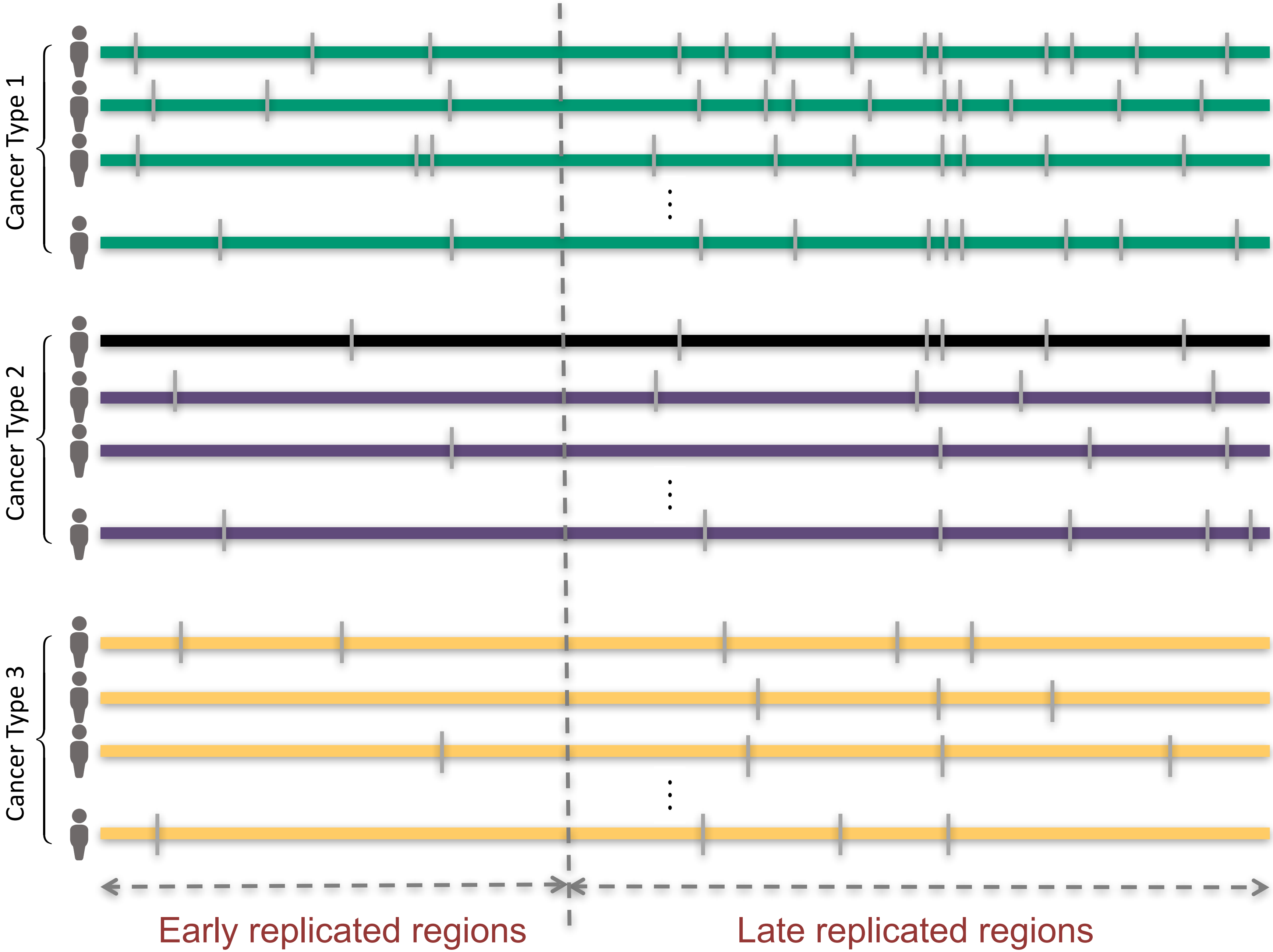
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

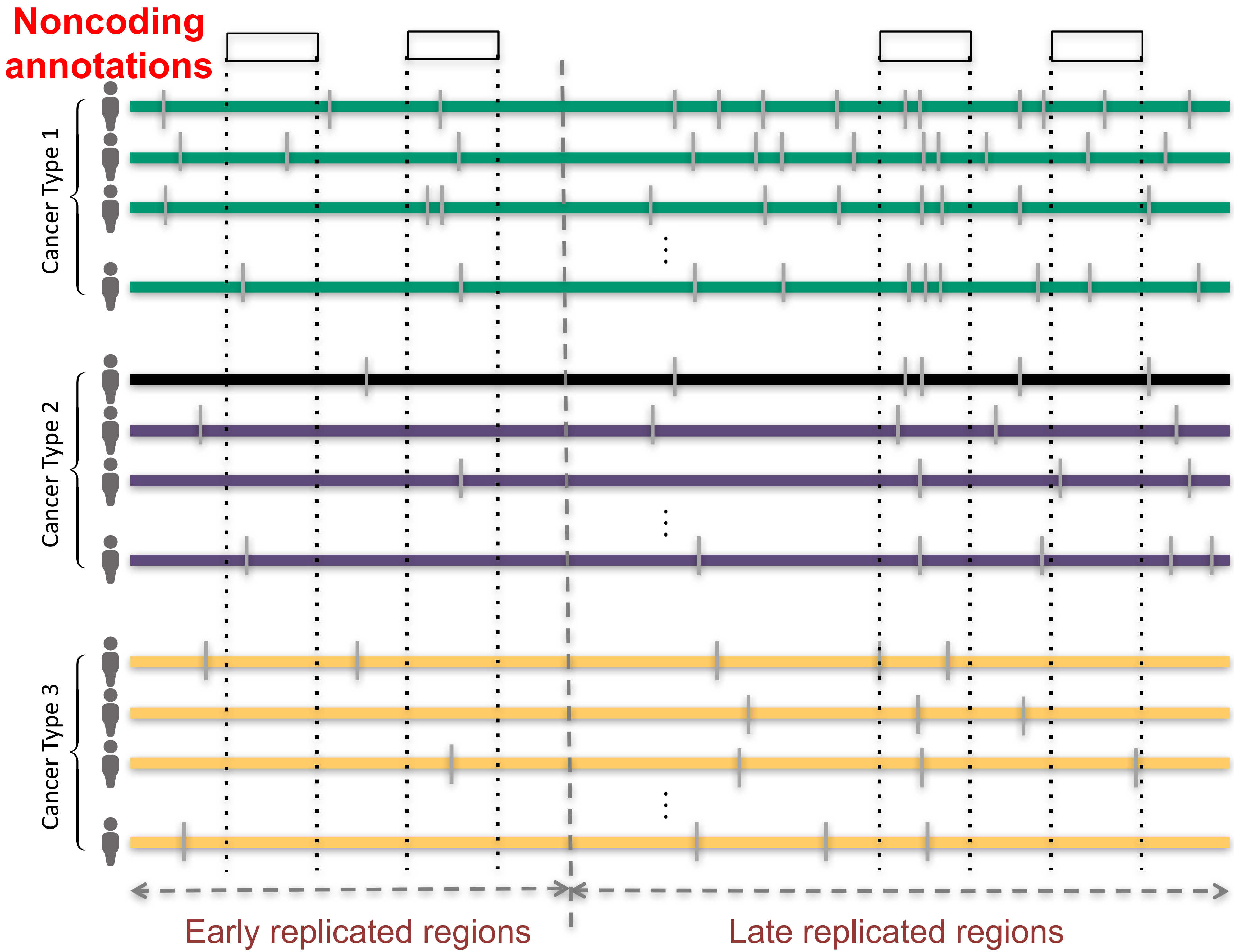
- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

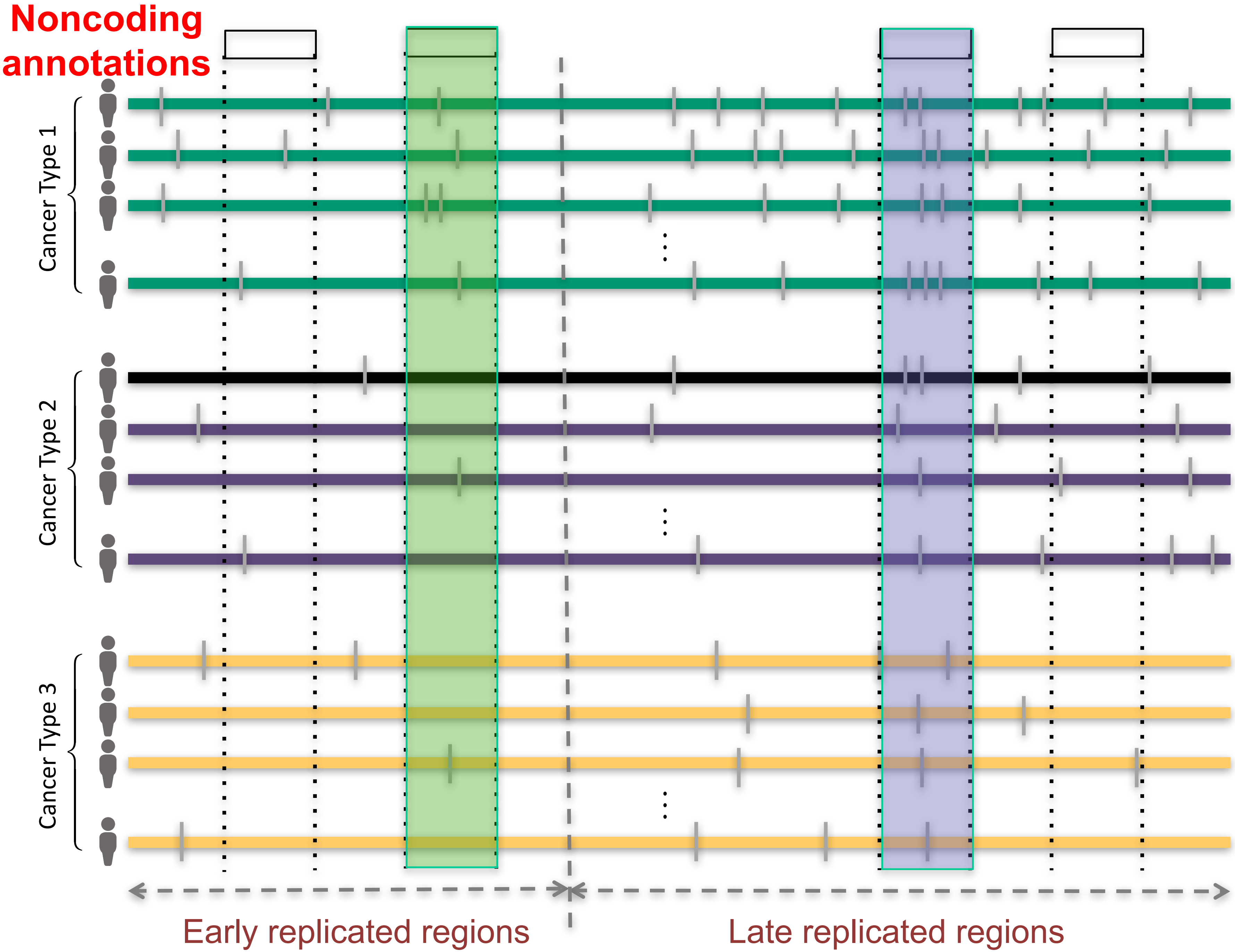
Mutation recurrence



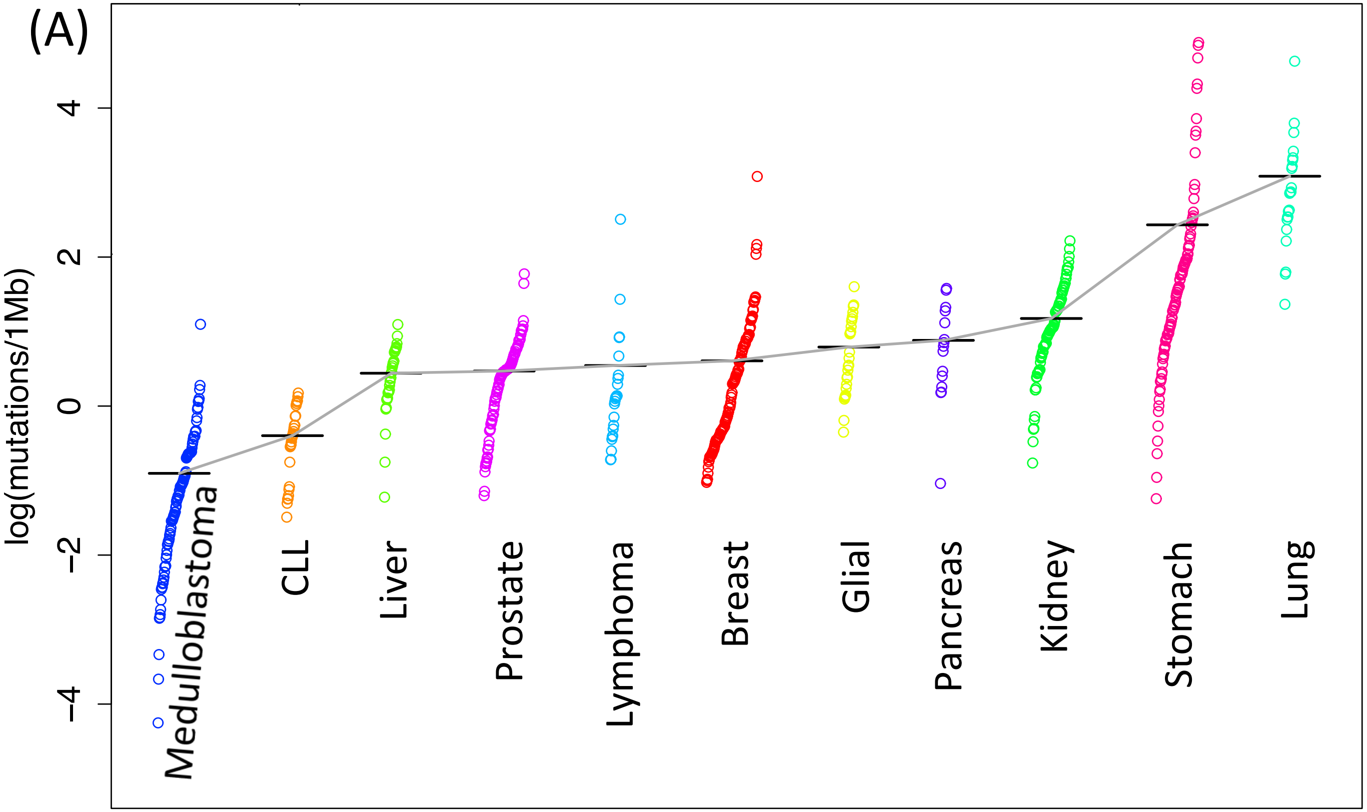
Mutation recurrence

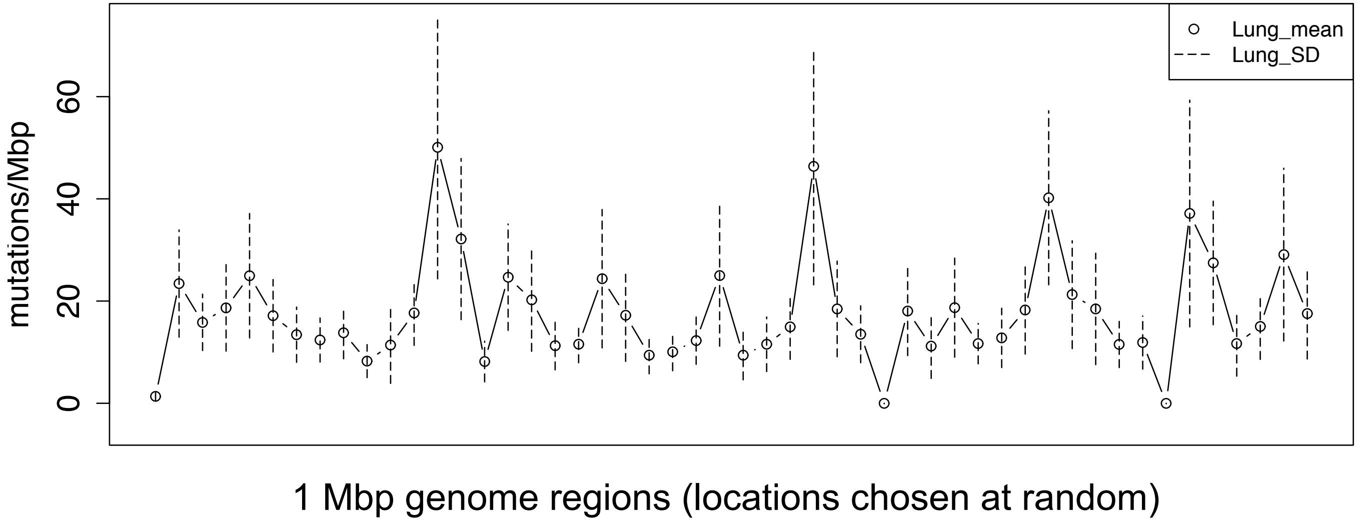



Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions

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

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

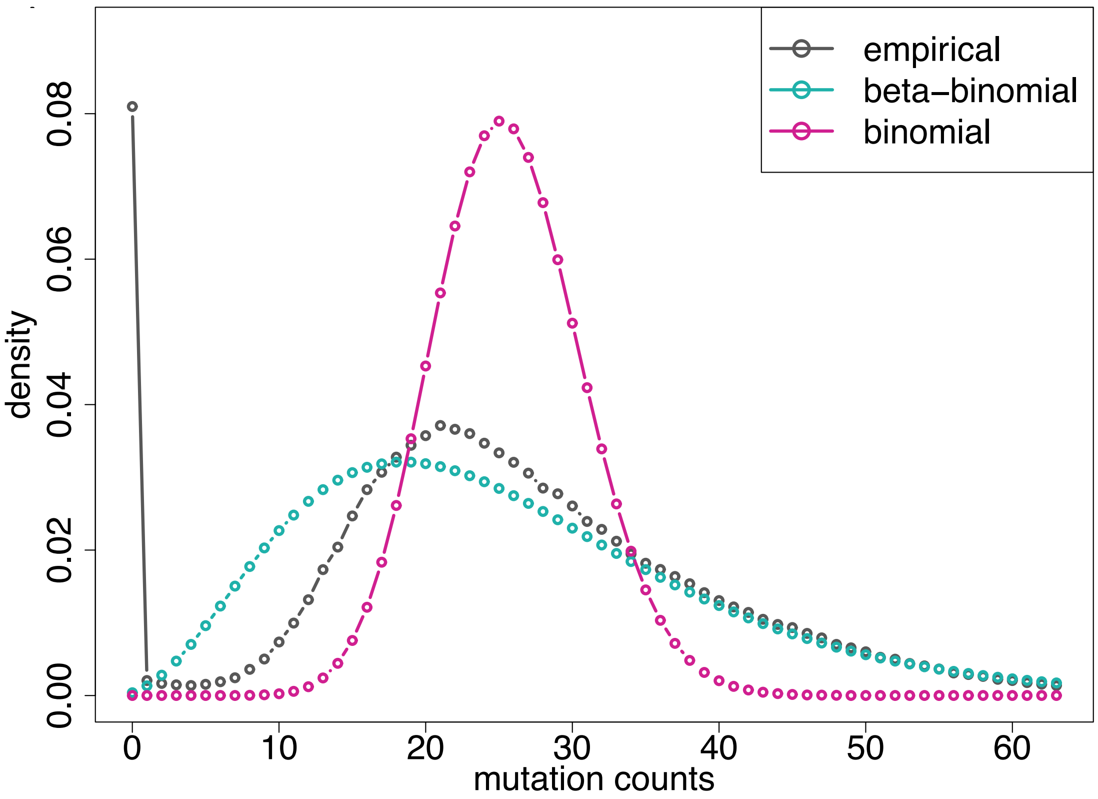
$$\mathbf{p_i} : Beta(\mu|\mathbf{R}, \sigma|\mathbf{R})$$

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

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

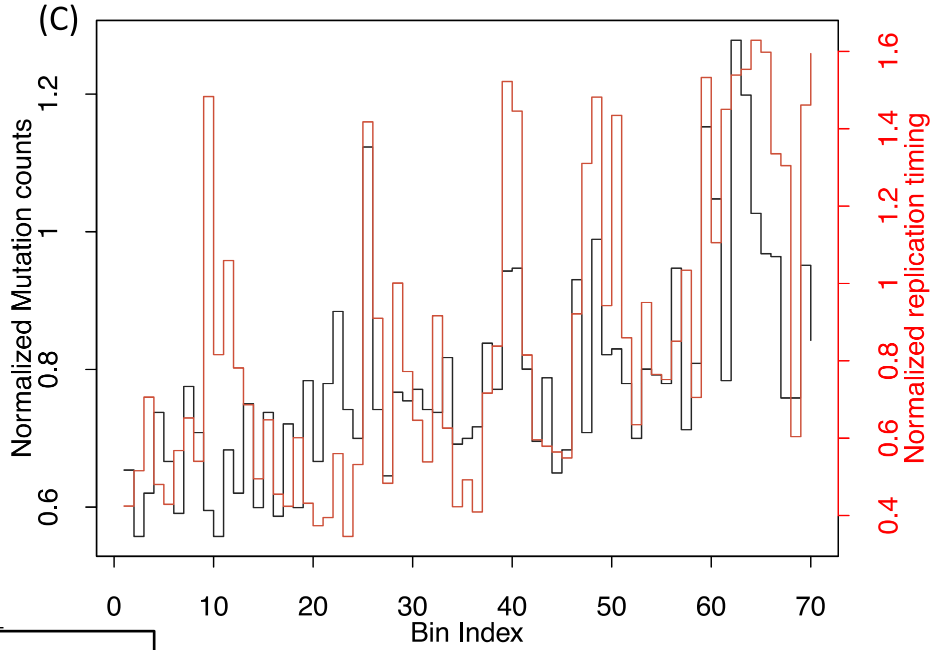
LARVA Model Comparison

- 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

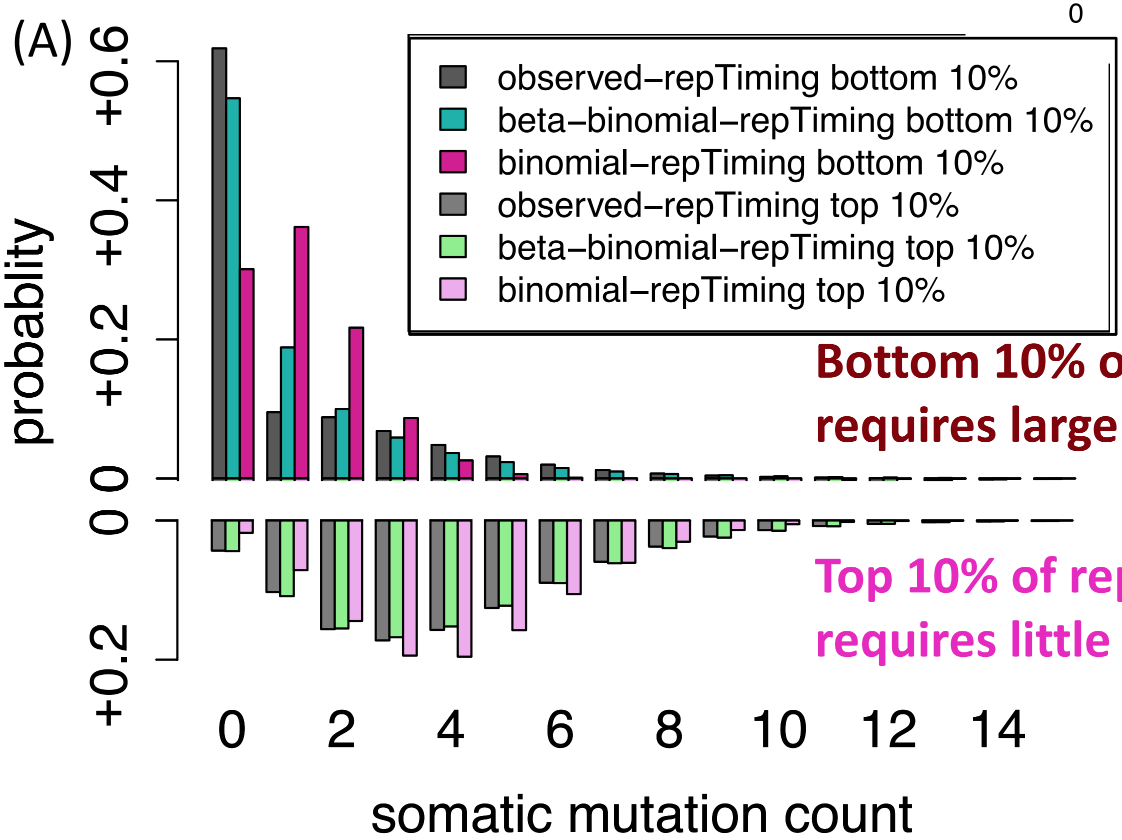


[Lochovsky et al. *NAR* ('15)]

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



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

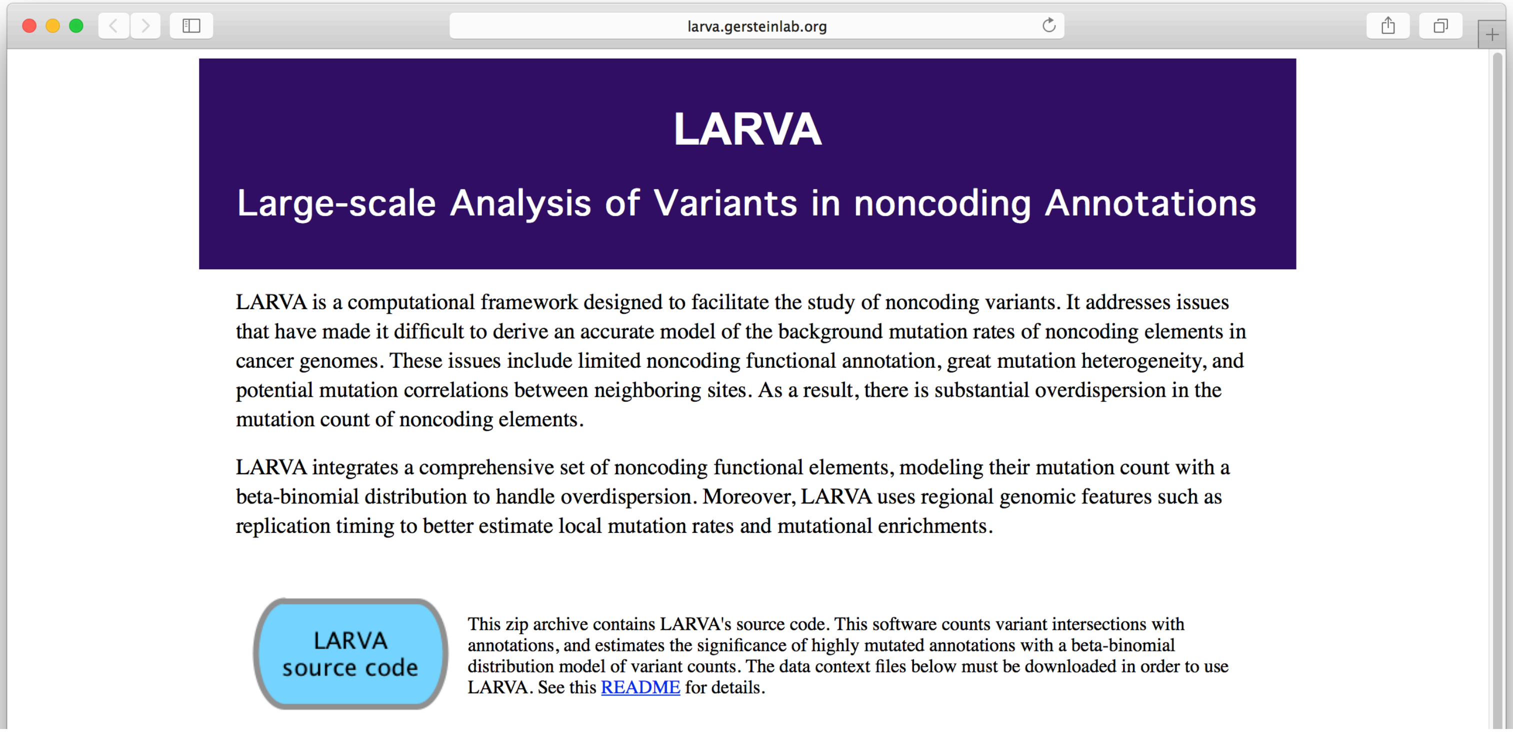


Bottom 10% of rep. timing bins requires large correction

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

Large-scale Analysis of Variants in noncoding Annotations

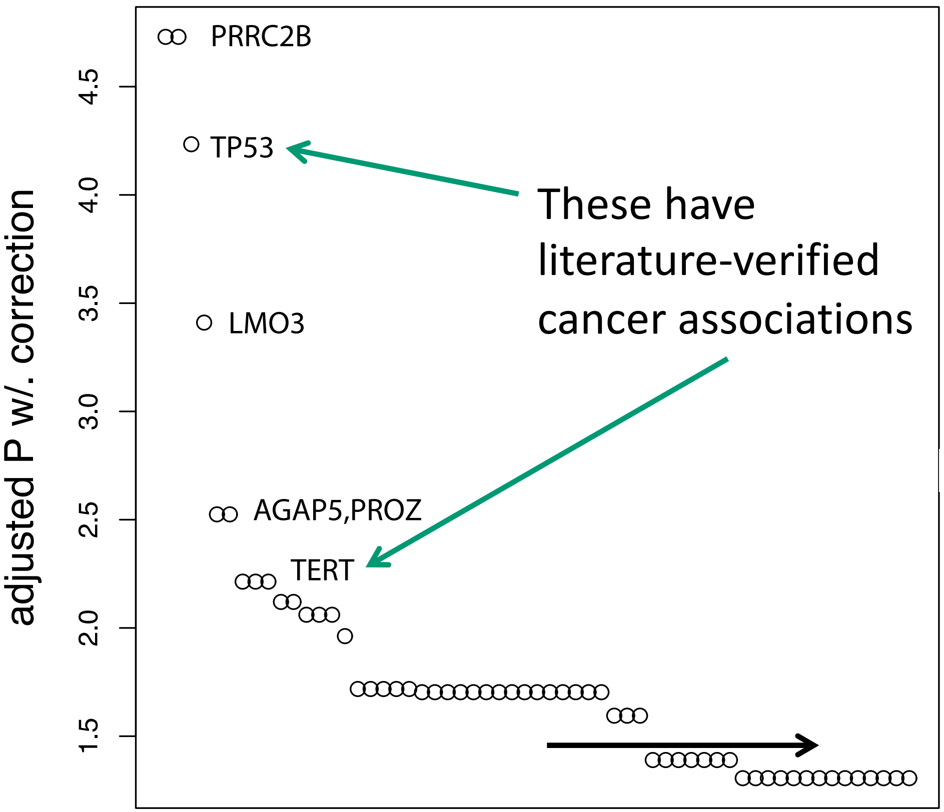
LARVA is a computational framework designed to facilitate the study of noncoding variants. It addresses issues that have made it difficult to derive an accurate model of the background mutation rates of noncoding elements in cancer genomes. These issues include limited noncoding functional annotation, 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.

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.

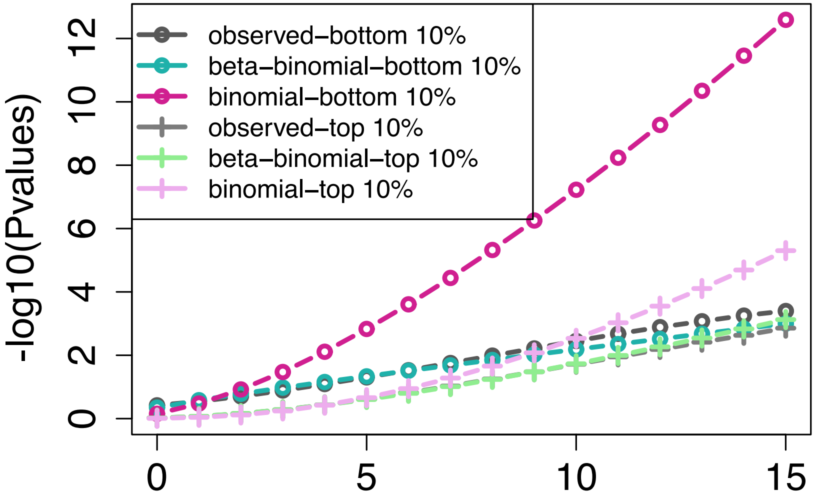
LARVA Results

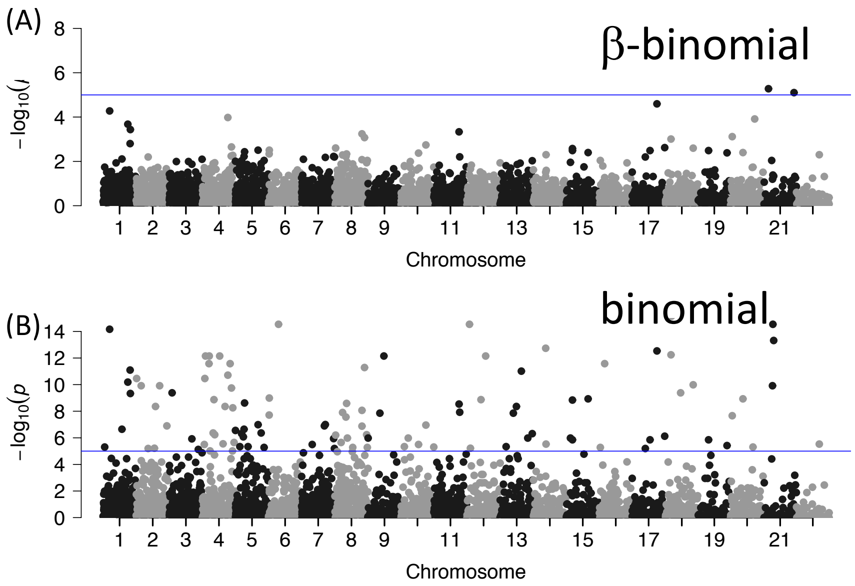
TSS LARVA results



These have literature-verified cancer associations

noncoding annotation p-values in sorted order





β-binomial

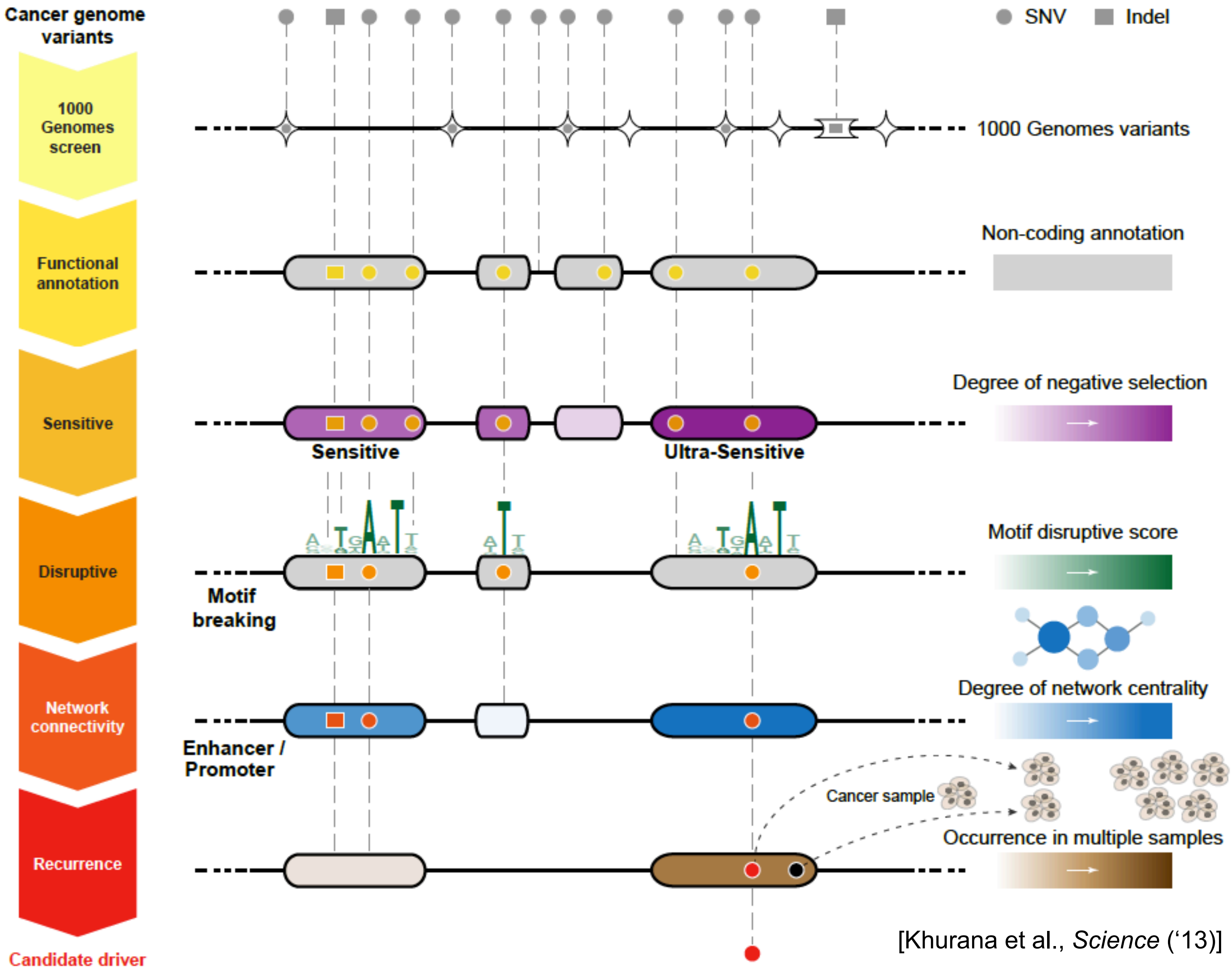
binomial

[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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

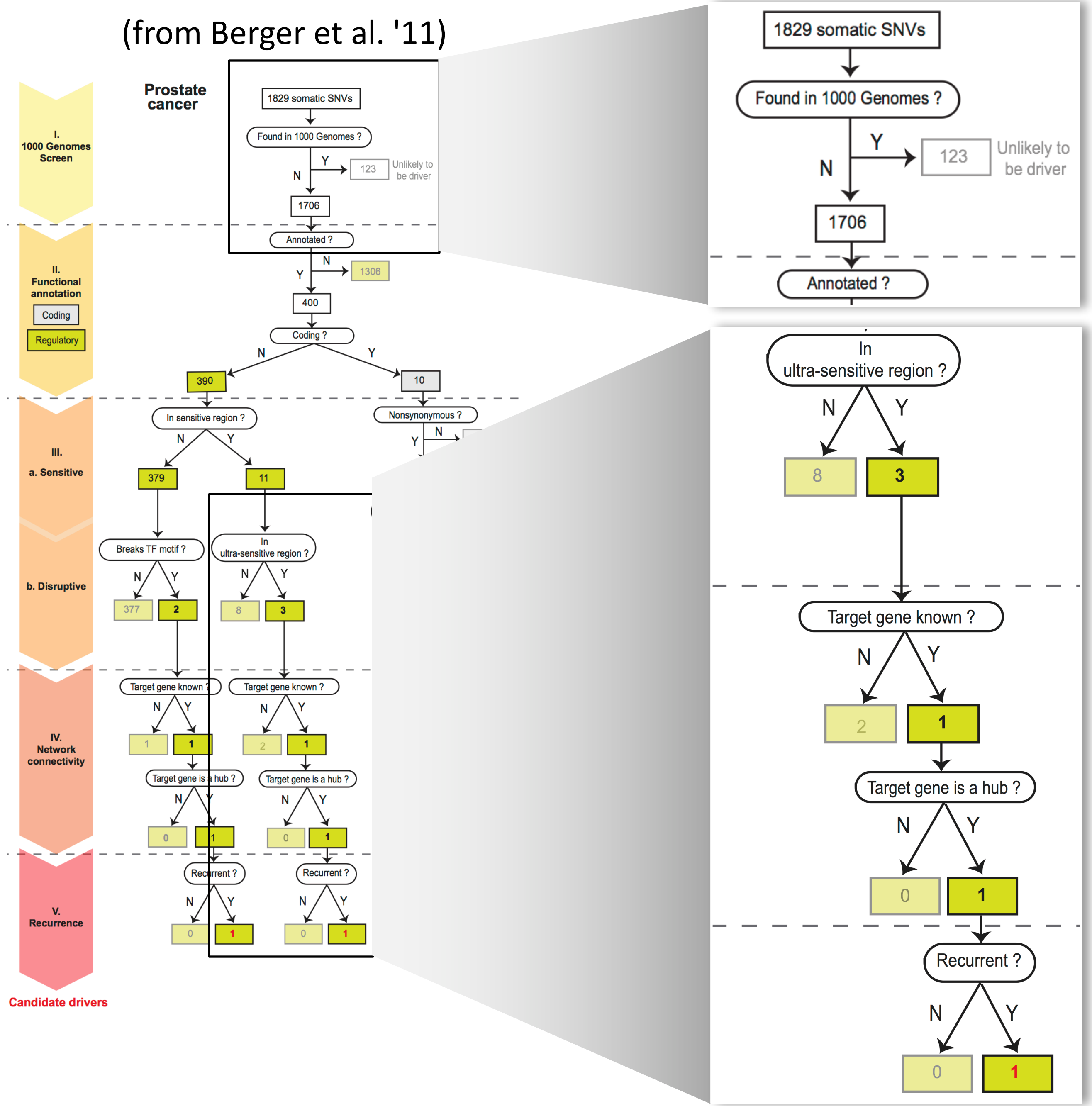
Identification of non-coding candidate drivers amongst somatic variants: Scheme



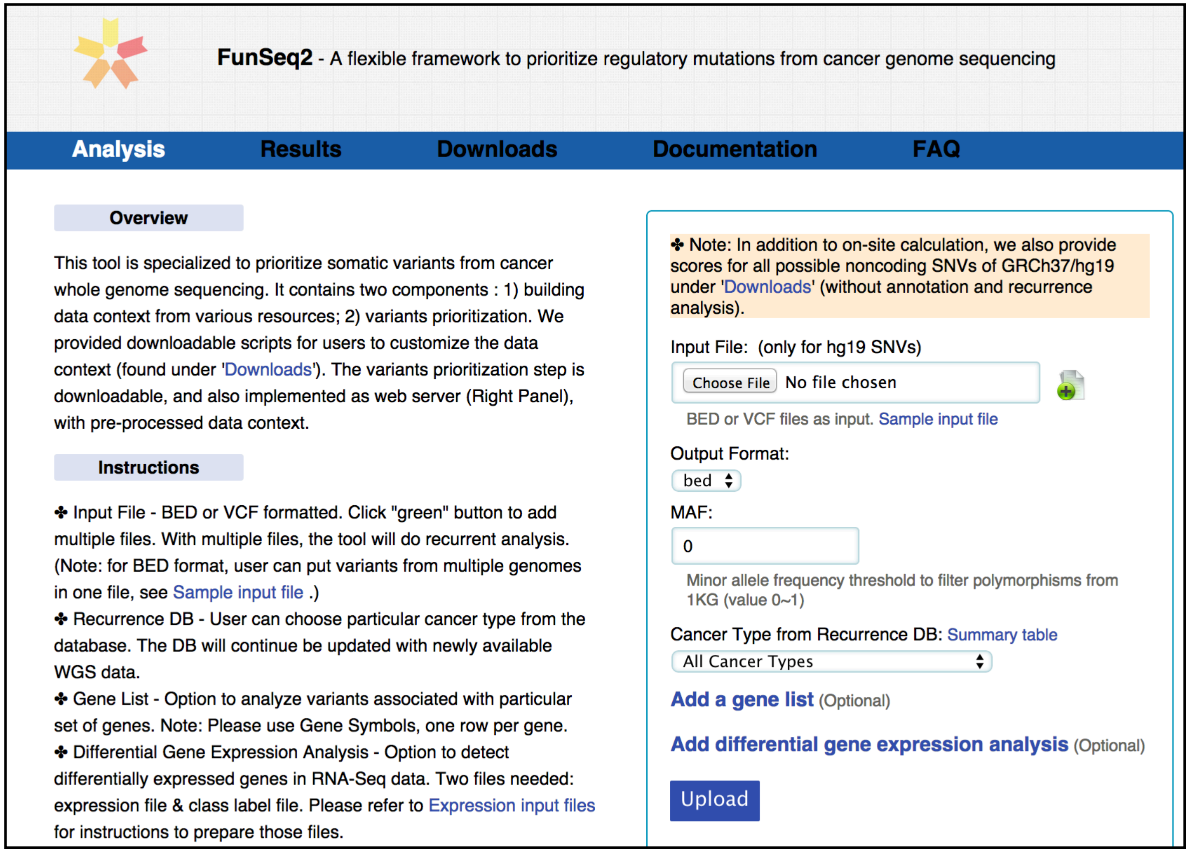
[Khurana et al., *Science* ('13)]

Flowchart for 1 Prostate Cancer Genome

(from Berger et al. '11)



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



FunSeq2 - A flexible framework to prioritize regulatory mutations from cancer genome sequencing

Analysis | Results | Downloads | Documentation | FAQ

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

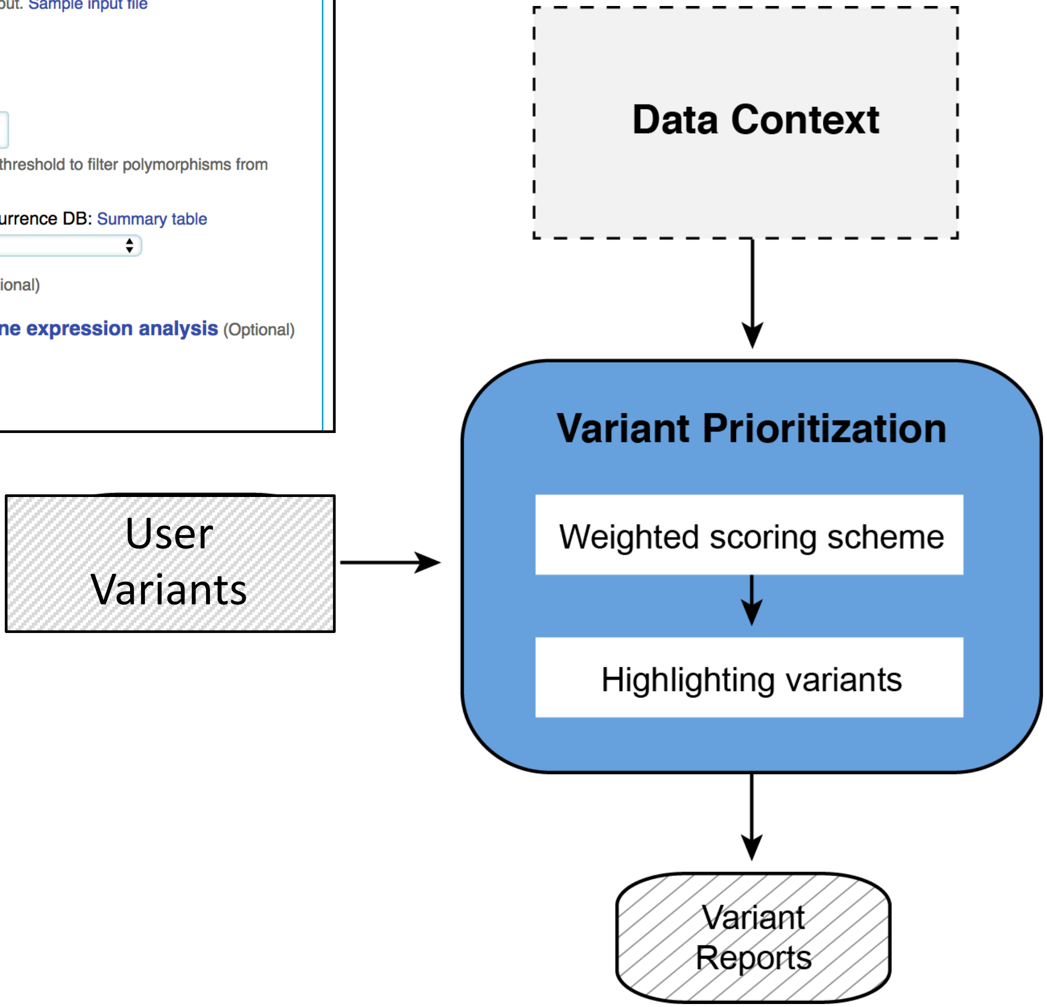
All Cancer Types

Add a gene list (Optional)

Add differential gene expression analysis (Optional)

Upload

Site integrates user variants with large-scale context

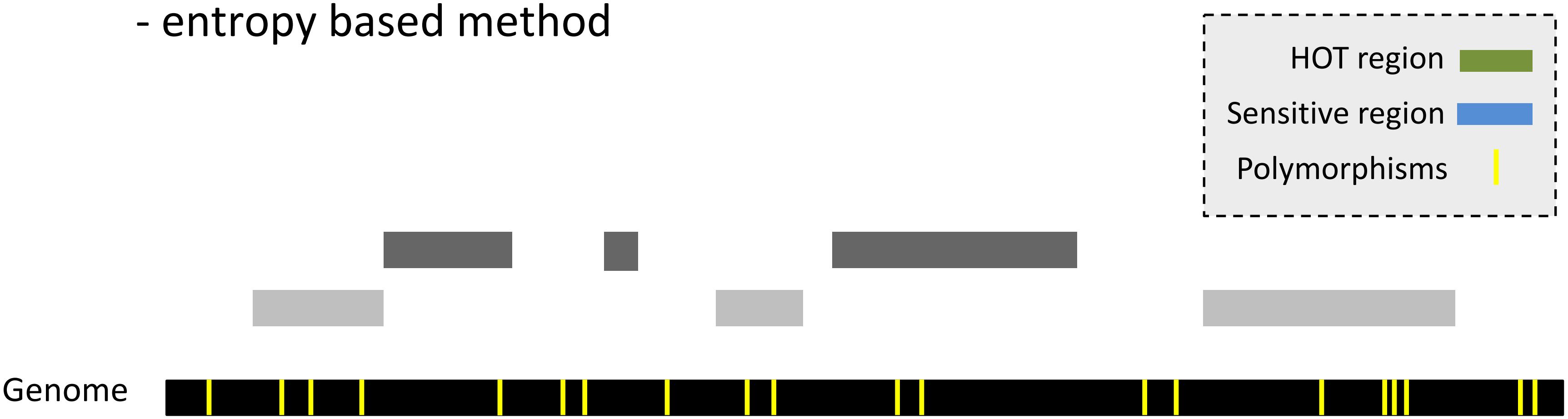


FunSeq.gersteinlab.org

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

- Feature weight

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

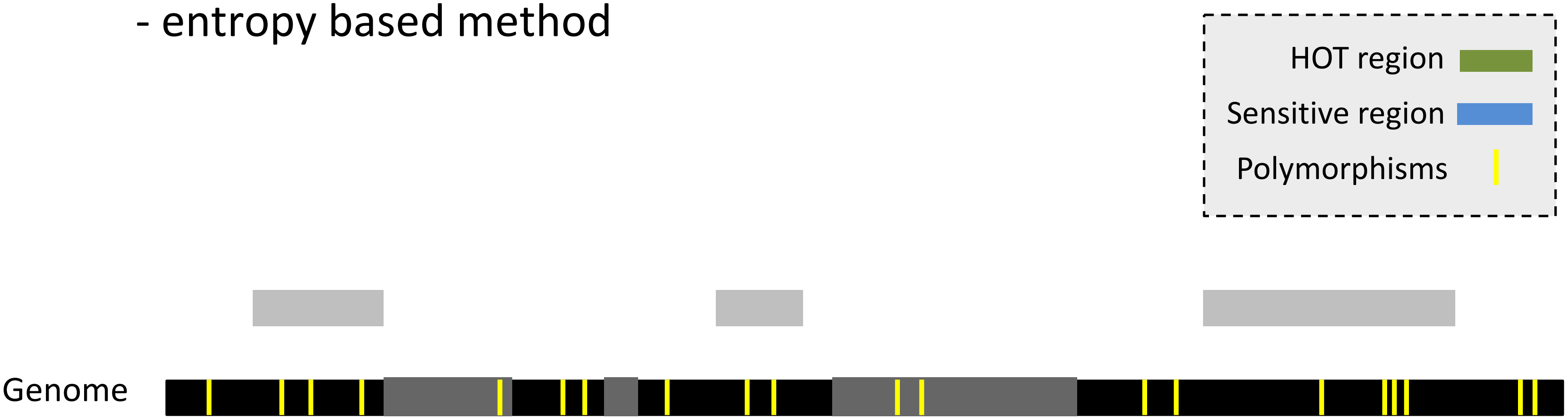


[Fu et al., GenomeBiology ('14)]

- Feature weight

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

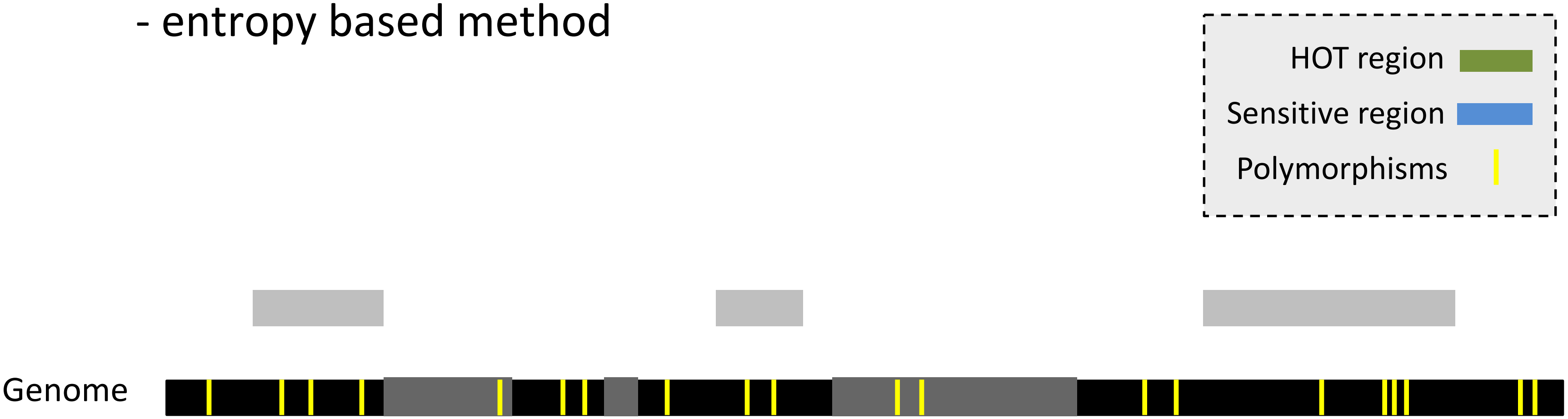
  - entropy based method



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

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

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



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

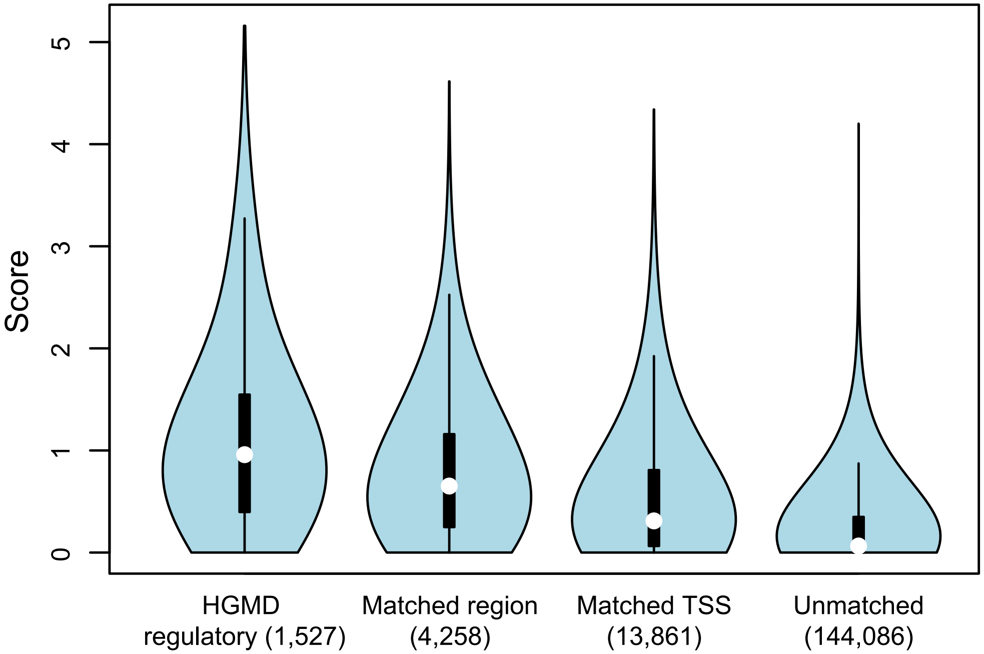
Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

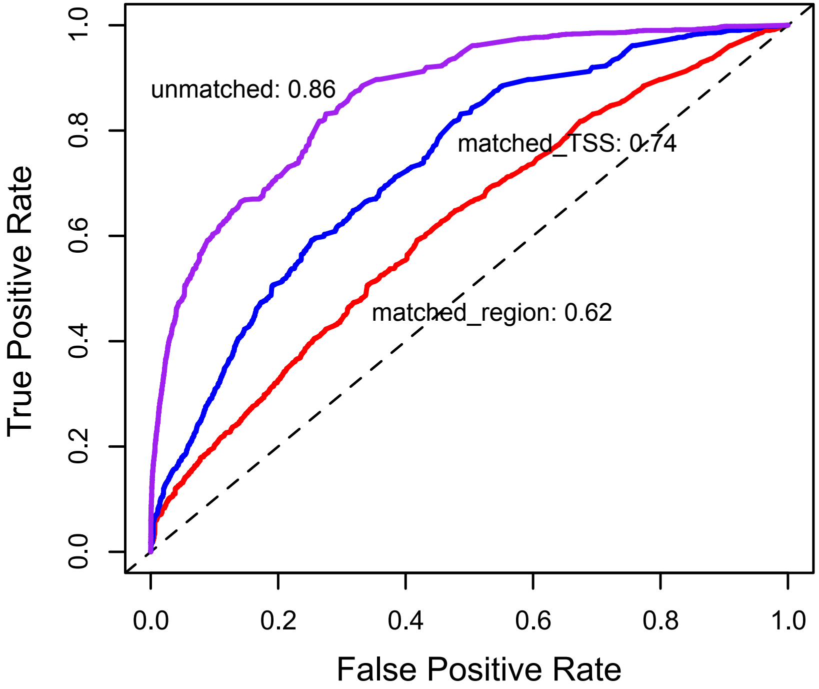
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

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

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

Germline pathogenic variants show higher core scores than controls





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

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

Ritchie et al., Nature Methods, 2014

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

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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes
- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - 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

github.com/gersteinlab/

Frustration

S **Kumar**,
D Clarke

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

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