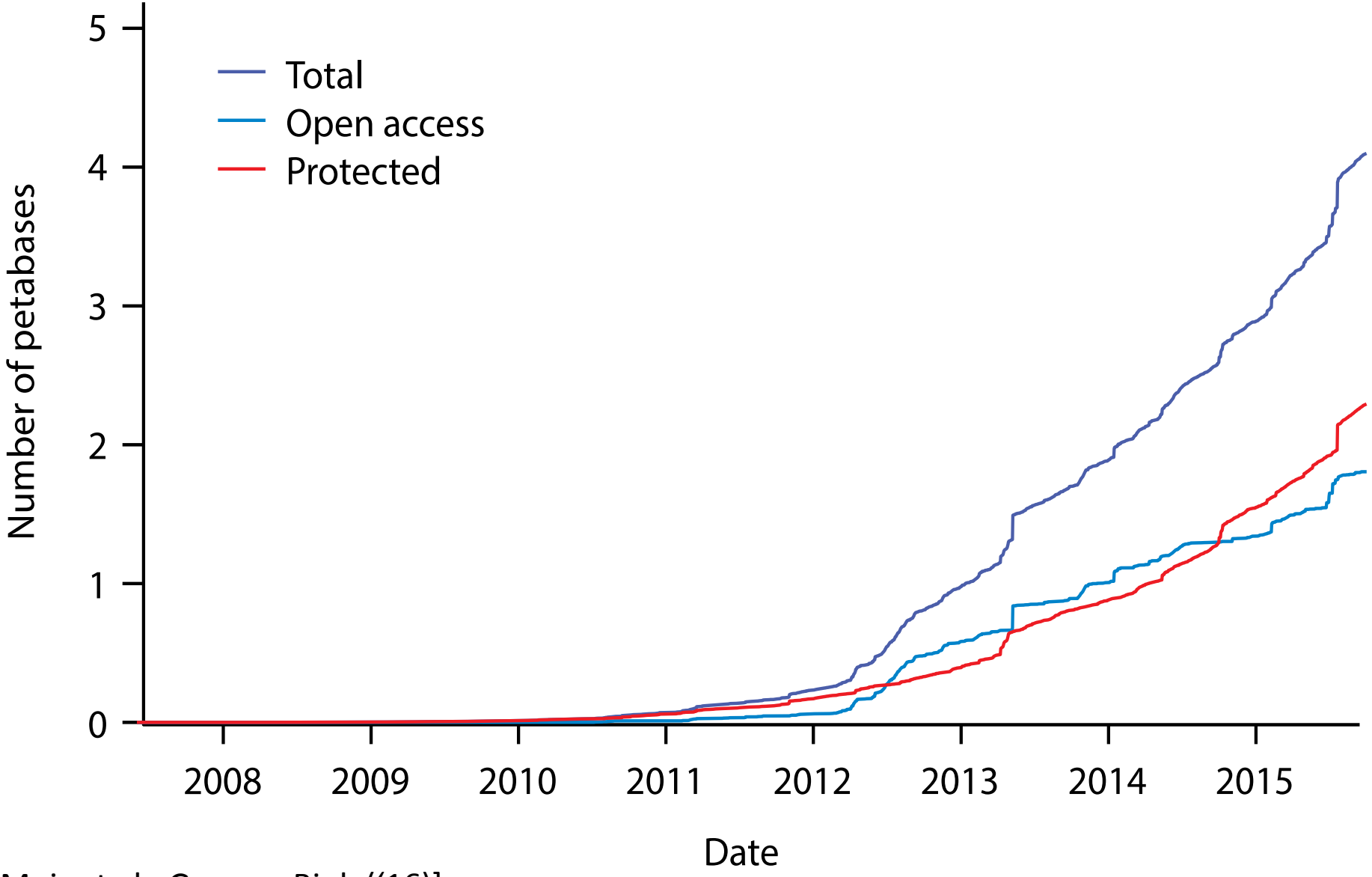
Personal Genomics:

Prioritizing High-impact Rare & Somatic Variants

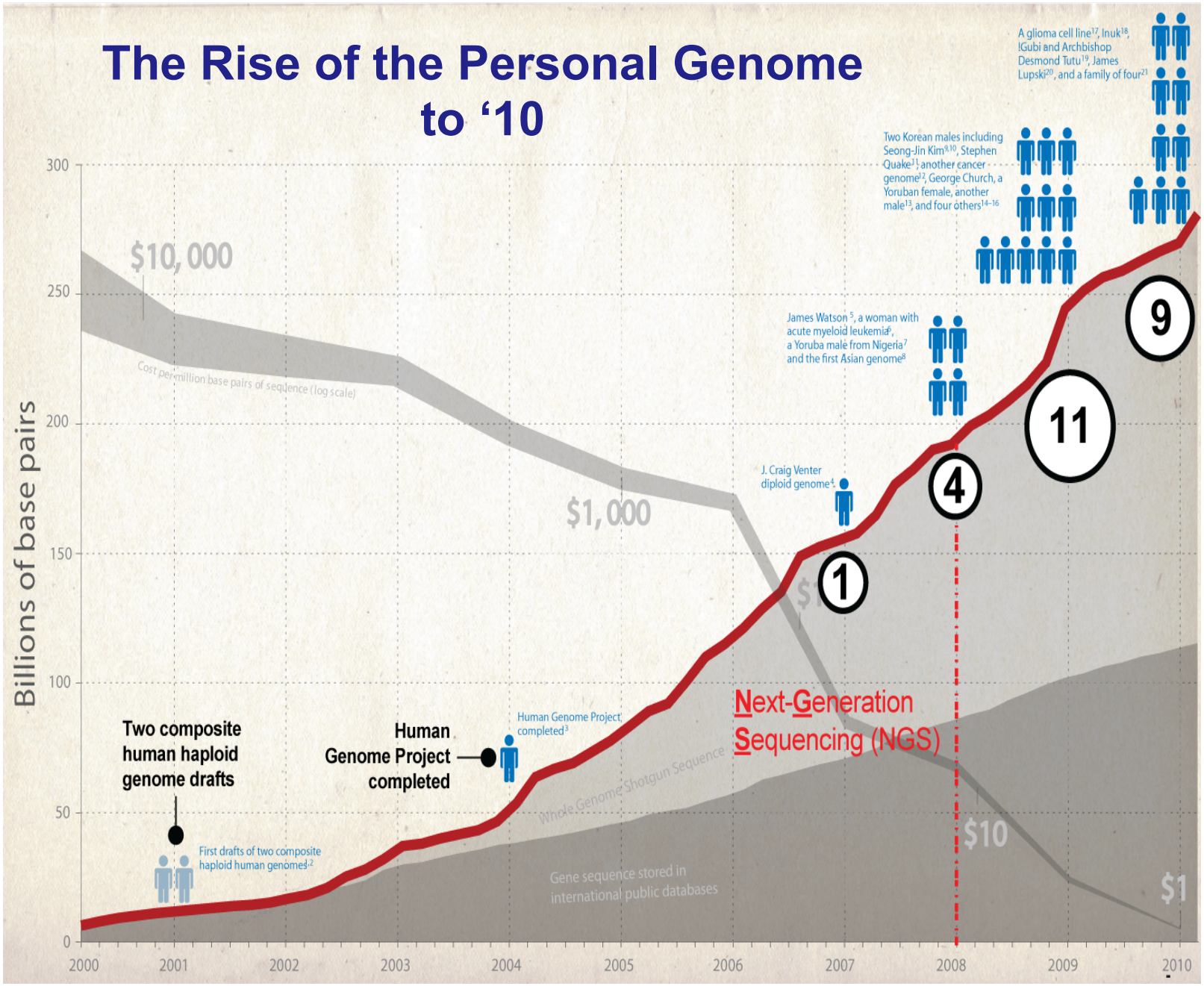
Mark Gerstein, Yale

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

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

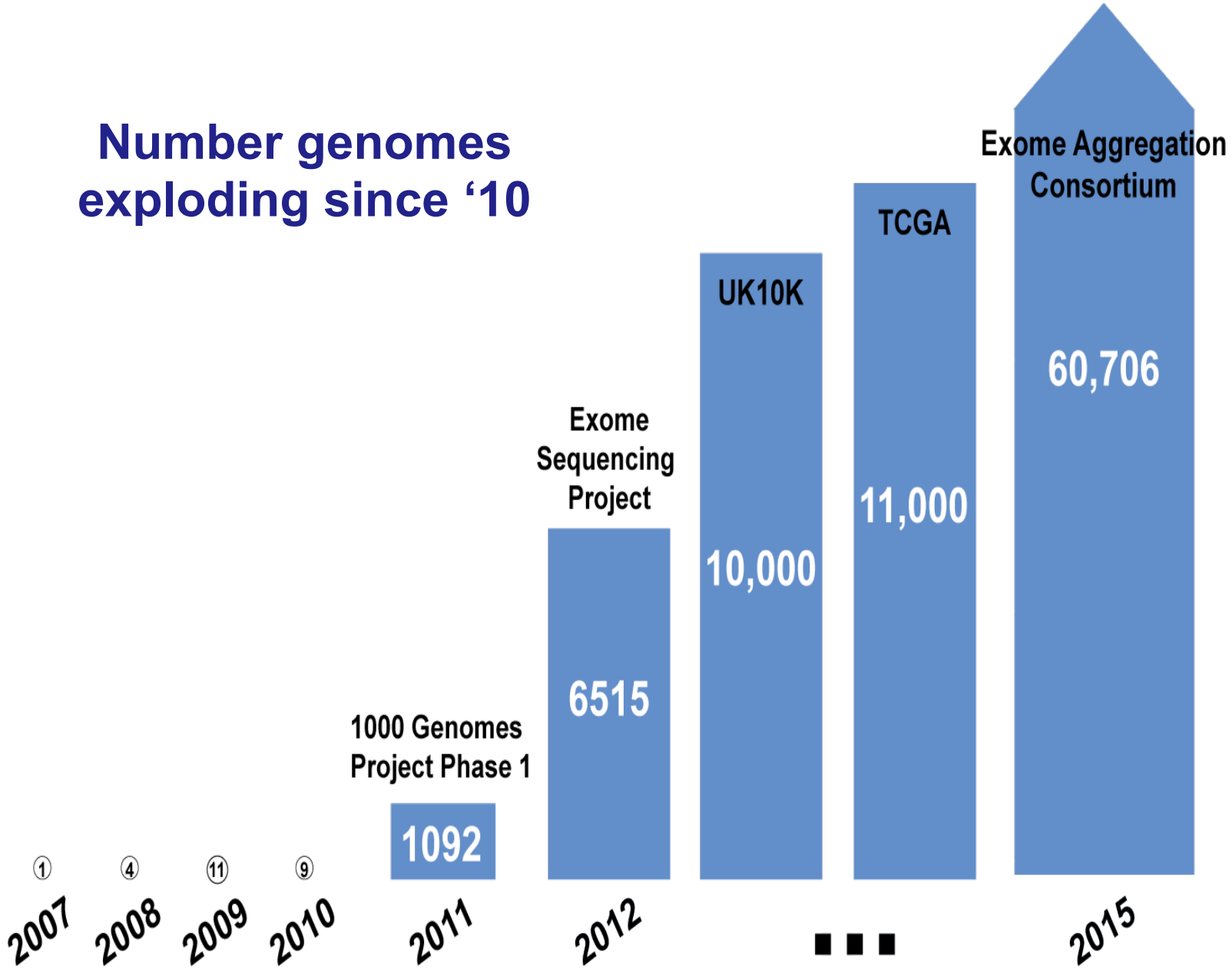


[Muir et al., GenomeBiol. (‘16)]



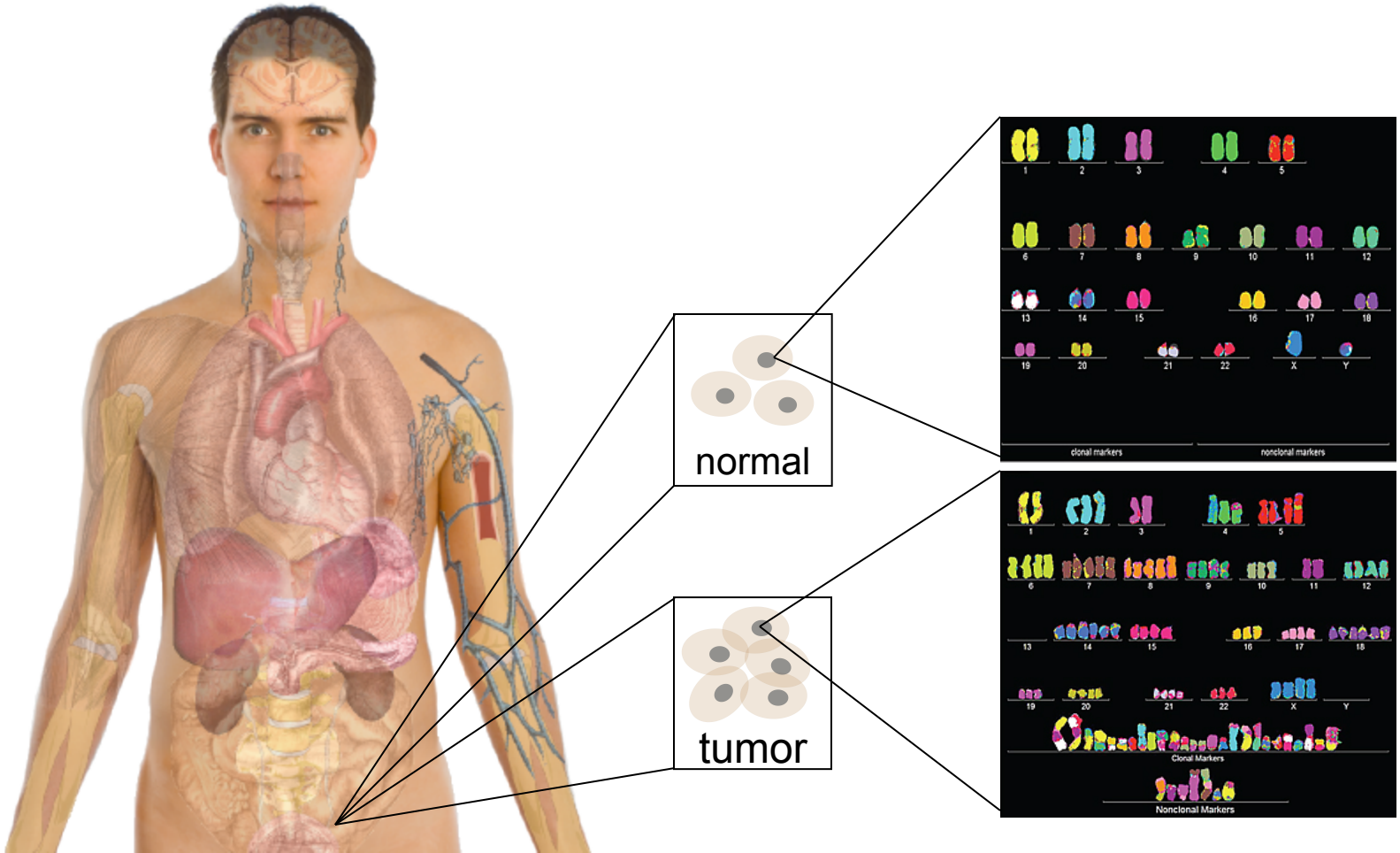
Adapted from *Nature* 2010

Number genomes exploding since ‘10



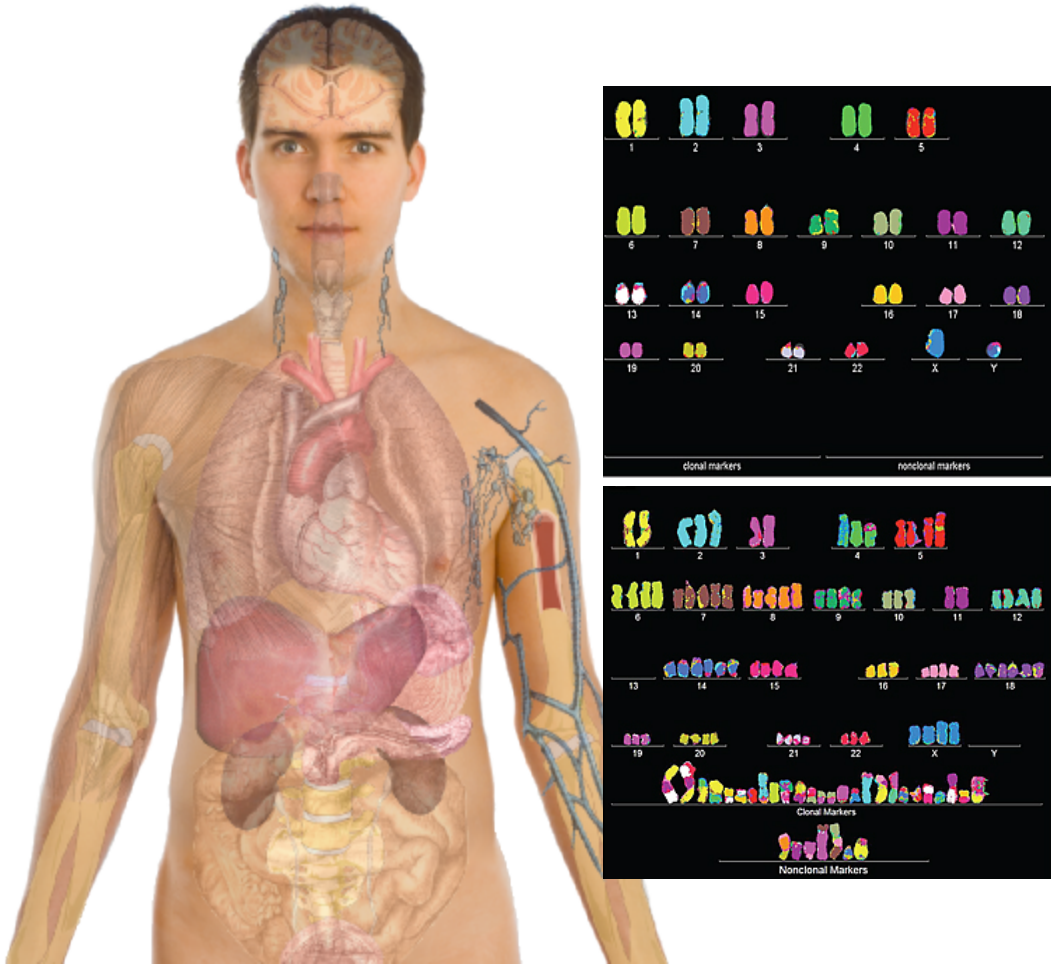
Personal Genomics as a Gateway into Biology

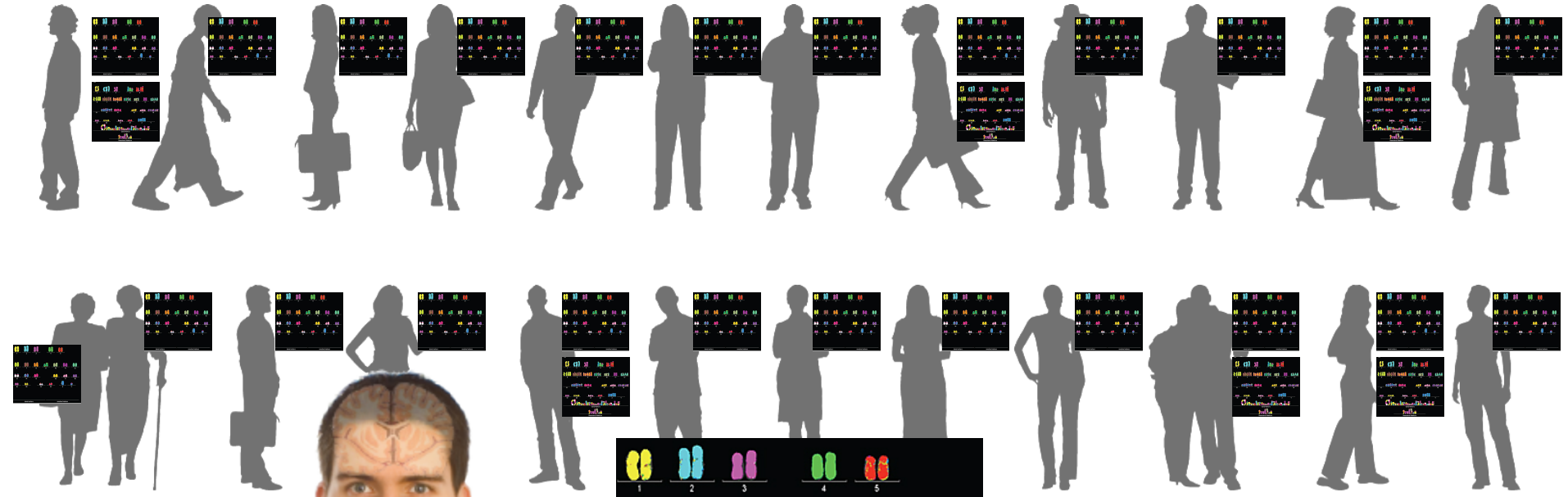
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

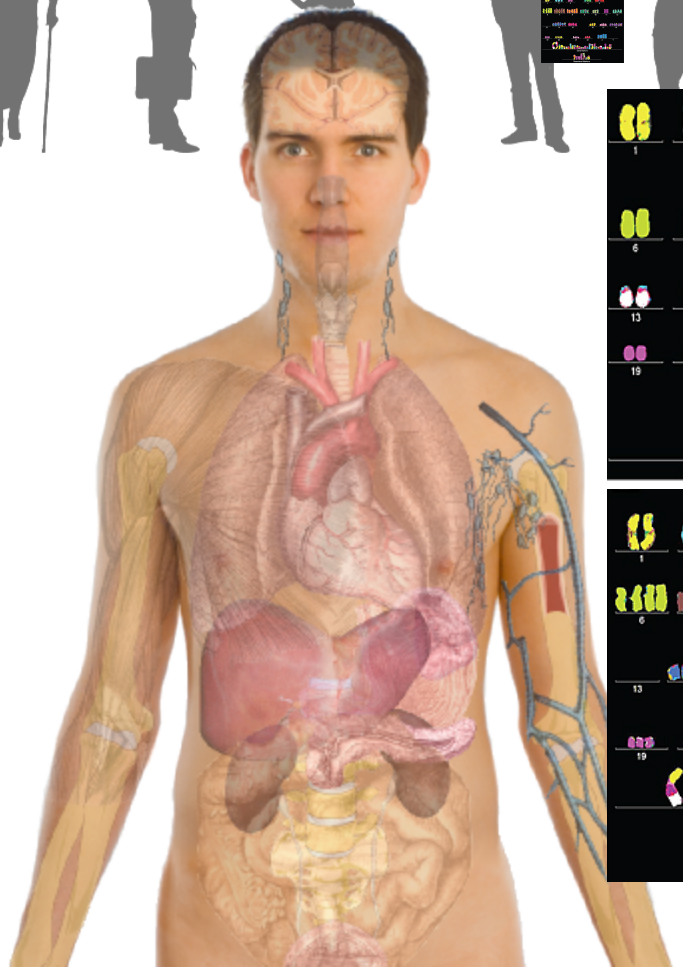


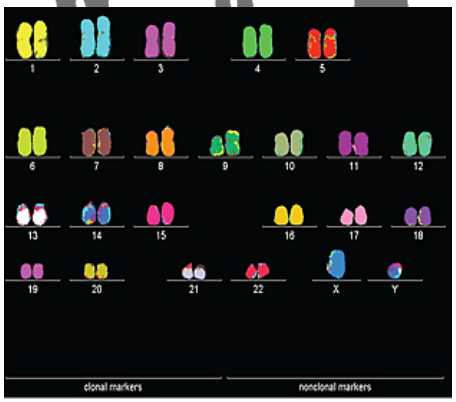
Personal Genomics as a Gateway into Biology

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

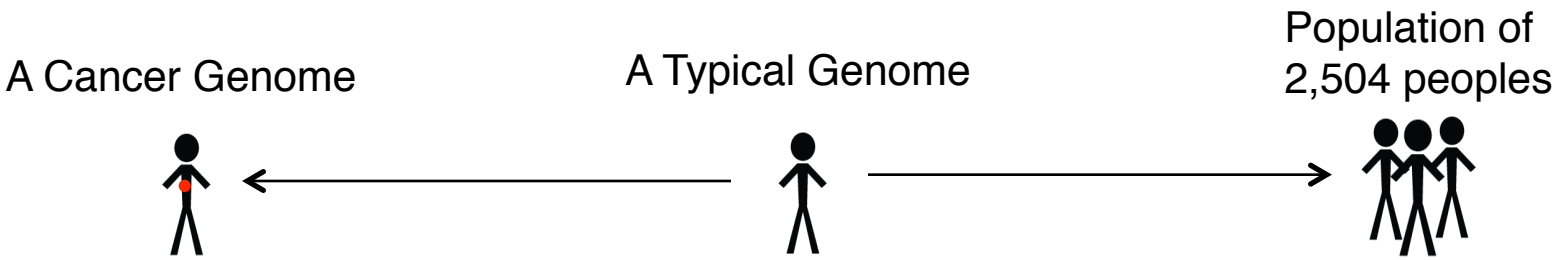






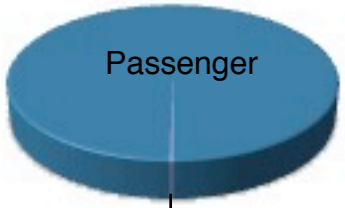


Human Genetic Variation



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

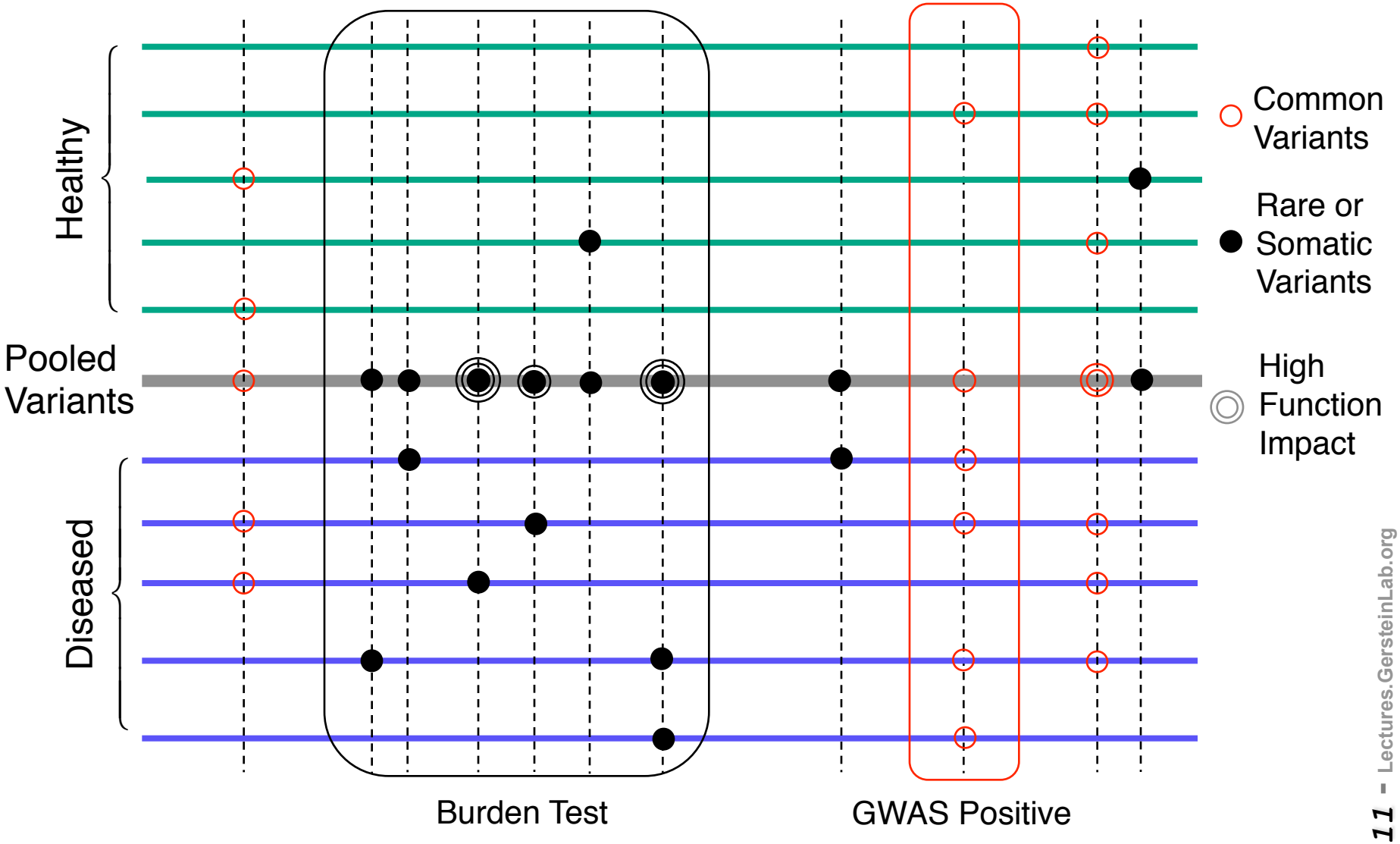
Somatic

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: Prioritizing High-impact Rare & Somatic Variants

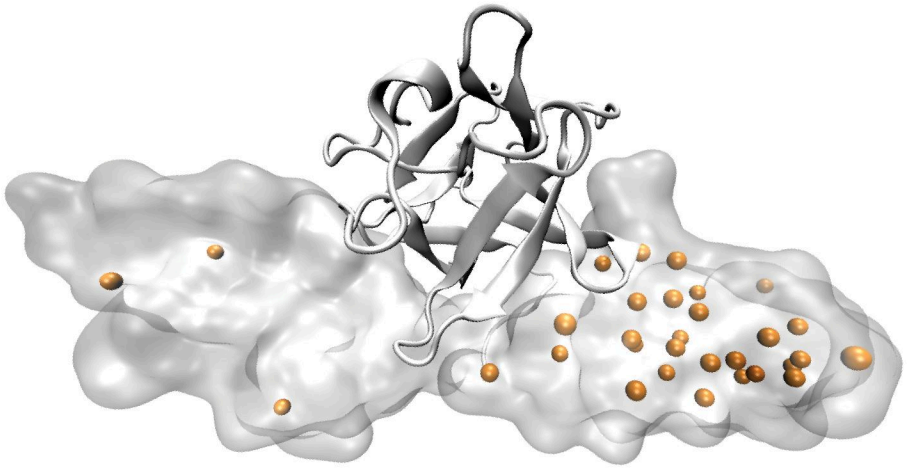
- **Introduction: the landscape of variants in personal genomes**
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
    - On surface & in interior bottlenecks
- **Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)**
- **Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
  - **Key technical Issue: Need to build personal genomes**
    - Assessing their quality via read mapping
- **Putting it together in workflows**
  - **Integrating evidence on non-coding variants with FunSeq**
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - **Using Larva to do burden testing on non-coding annotation**
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

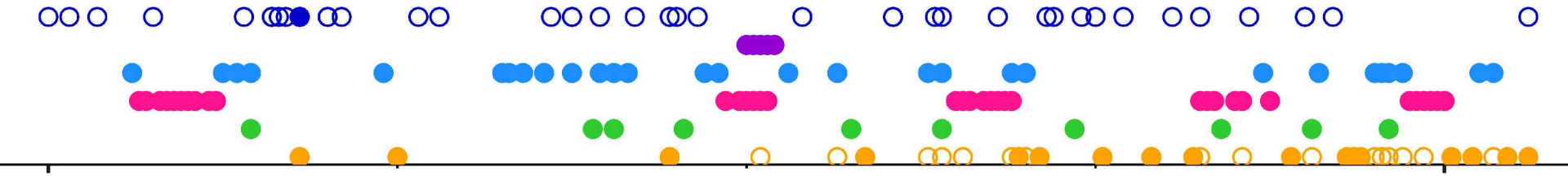
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

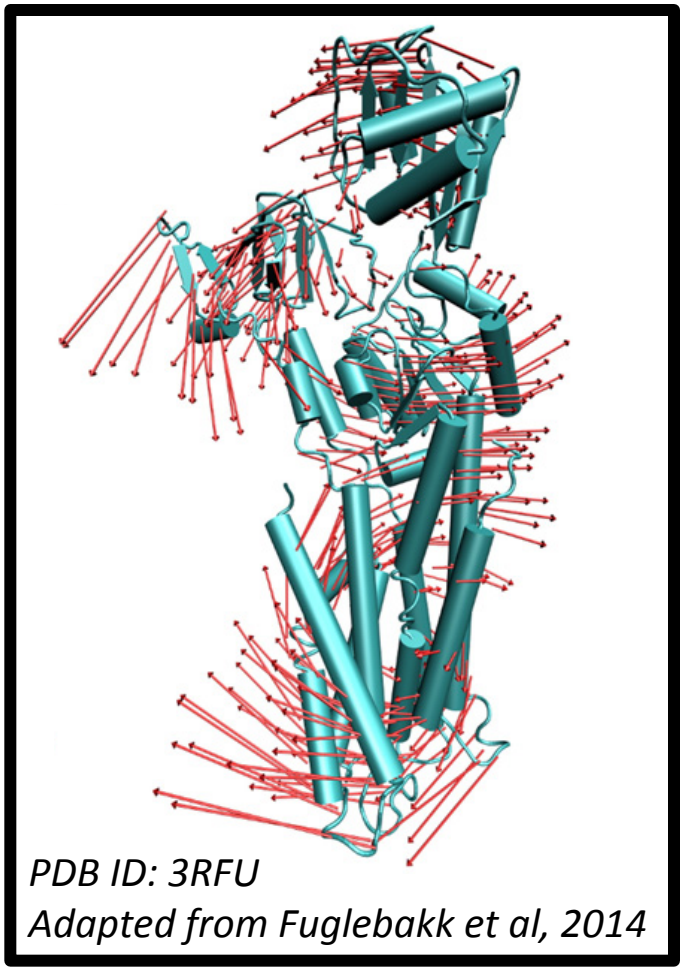
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



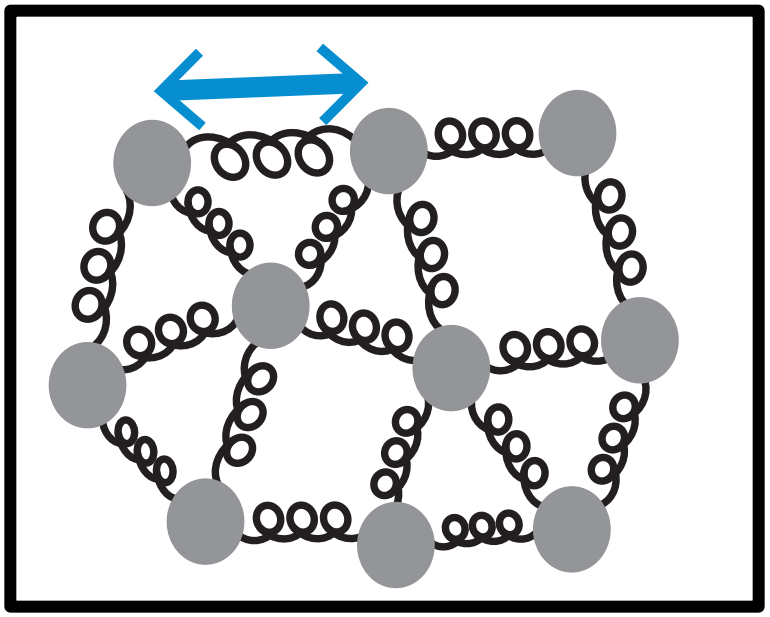
[Sethi et al. COSB ('15)]

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



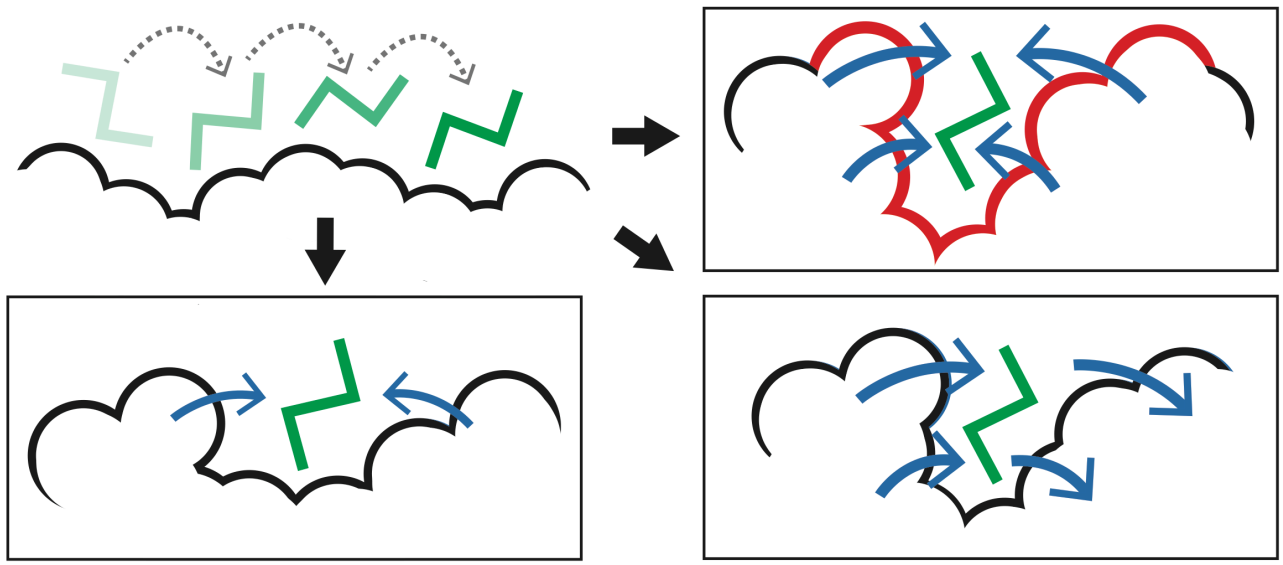
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



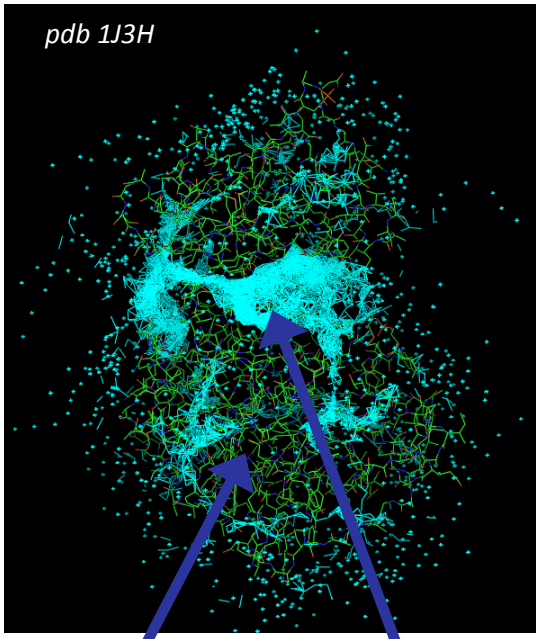
Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion

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



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

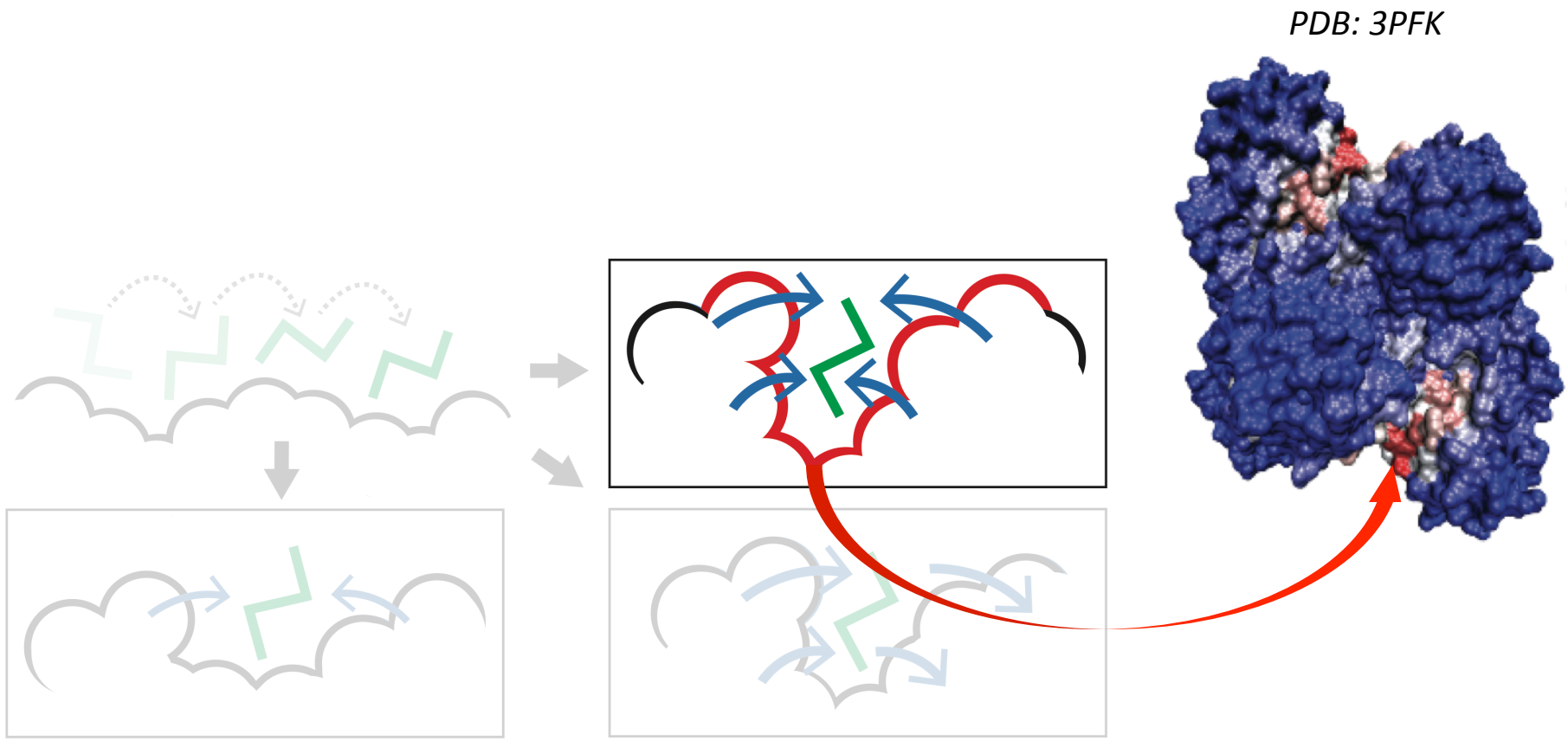


Surface region with high density of candidate sites

Surface region with low density of candidate sites

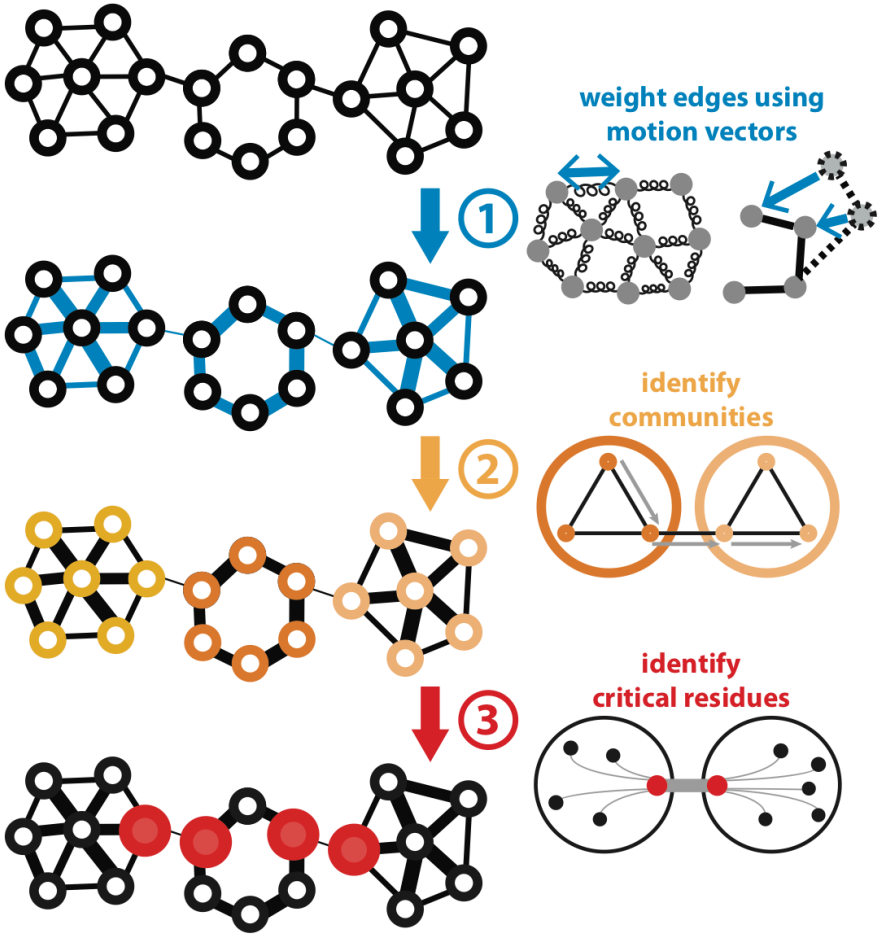
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface



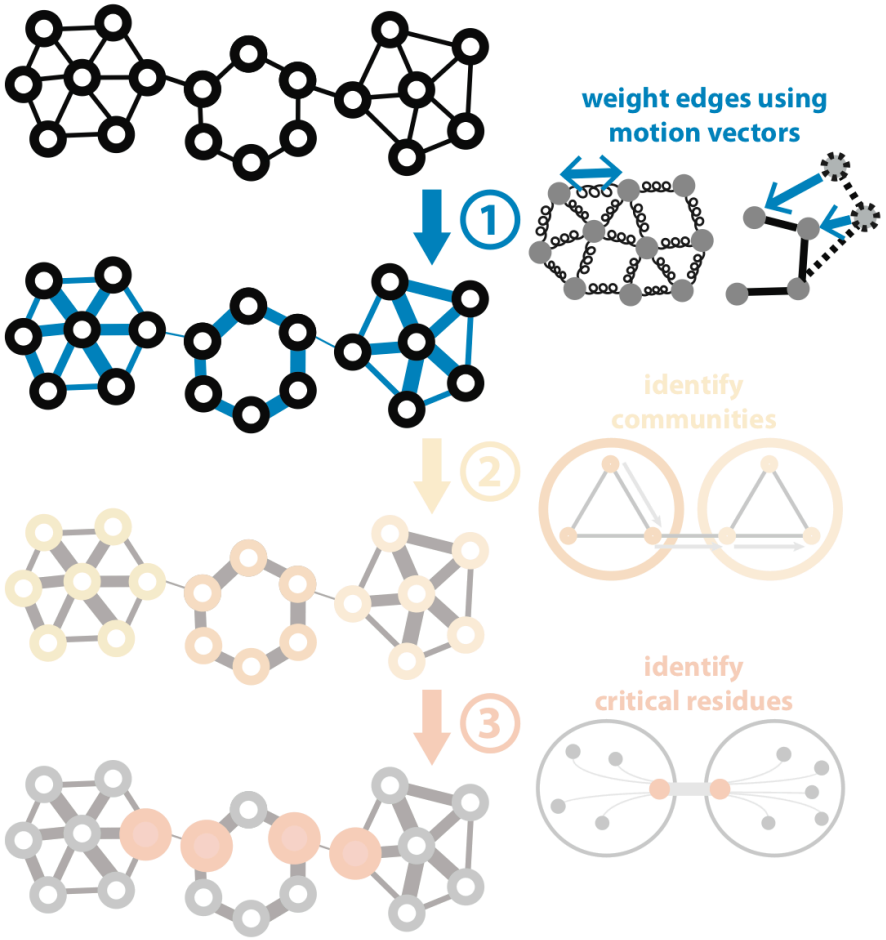
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



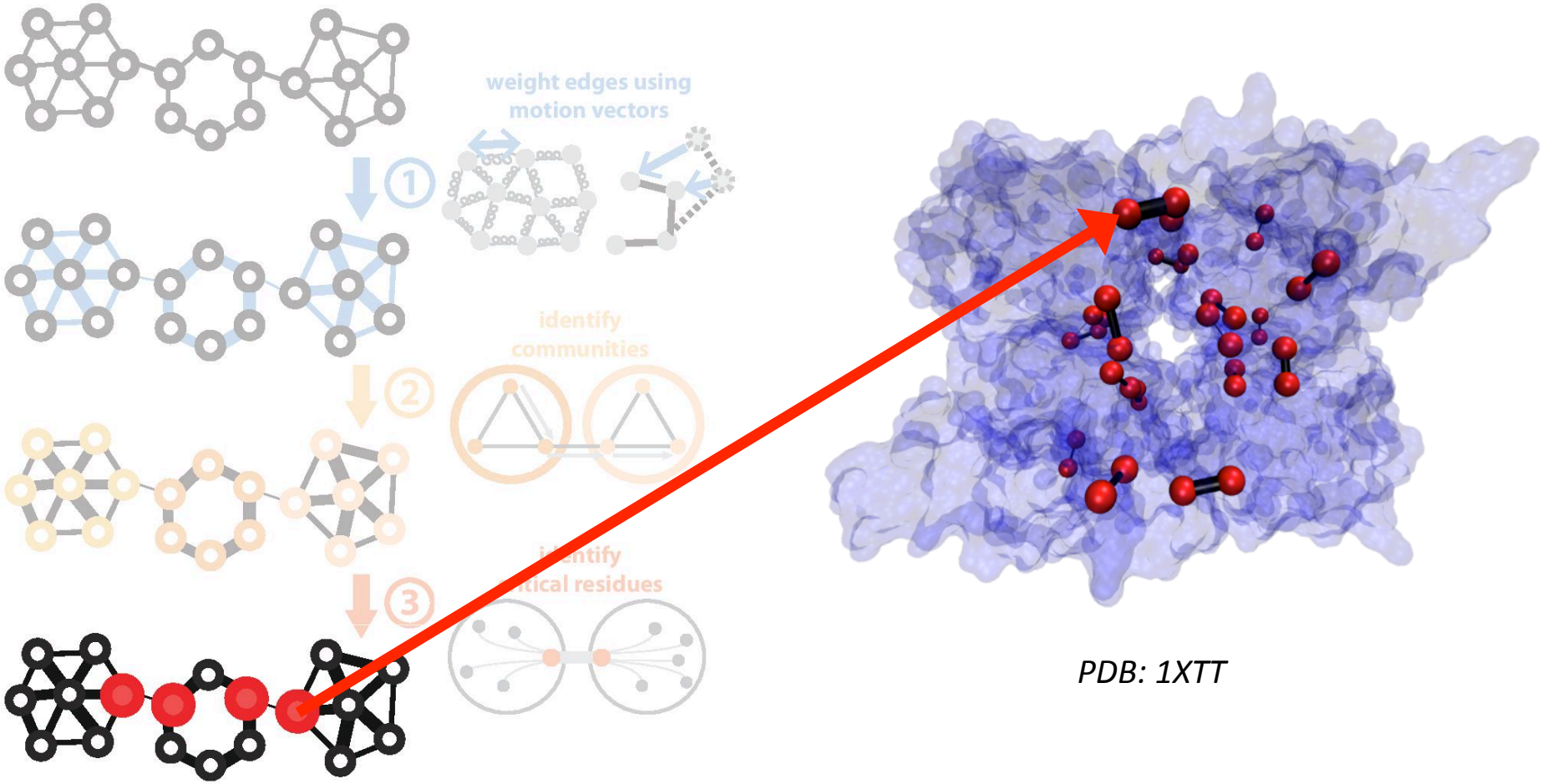
$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

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

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

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

Predicting Allosterically-Important Residues within the Interior

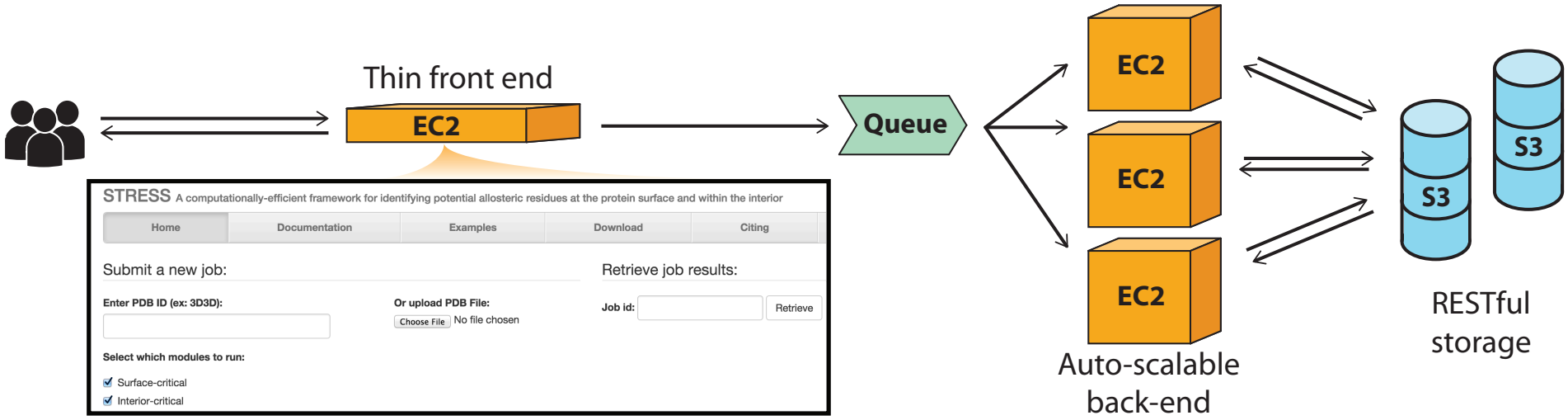


PDB: 1XTT

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

STRESS Server Architecture: Highlights

stress.molmovdb.org

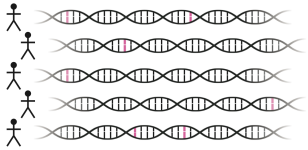


- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

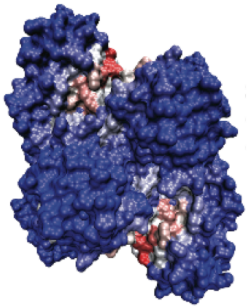
Adapted from Clarke*, Sethi*, et al *(in press)*

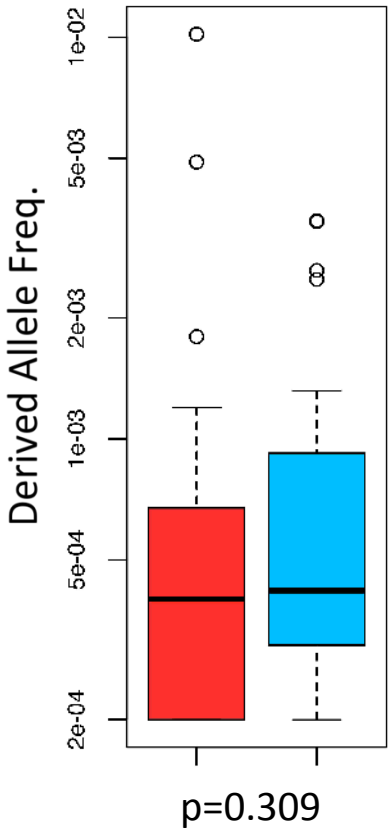
Intra-species conservation of predicted allosteric residues

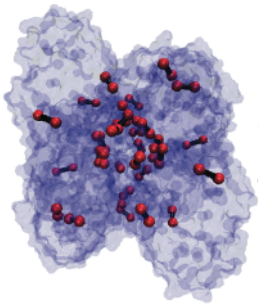
1000 Genomes



Surface



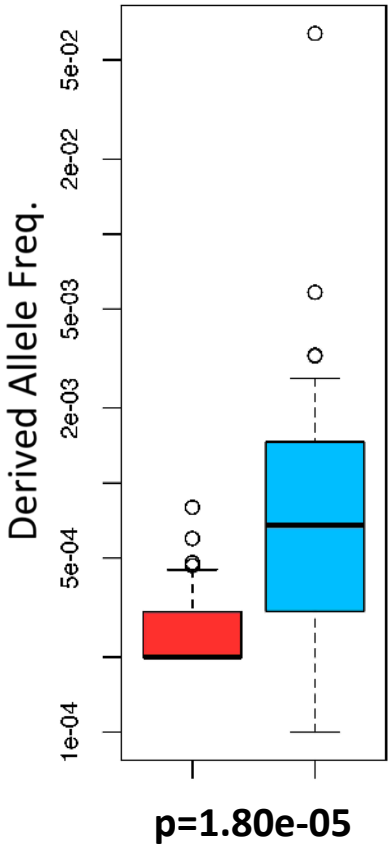




critical

non-critical

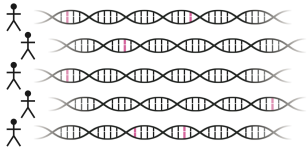
Interior

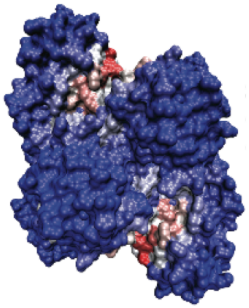


Adapted from Clarke*, Sethi*, et al *(in press)*

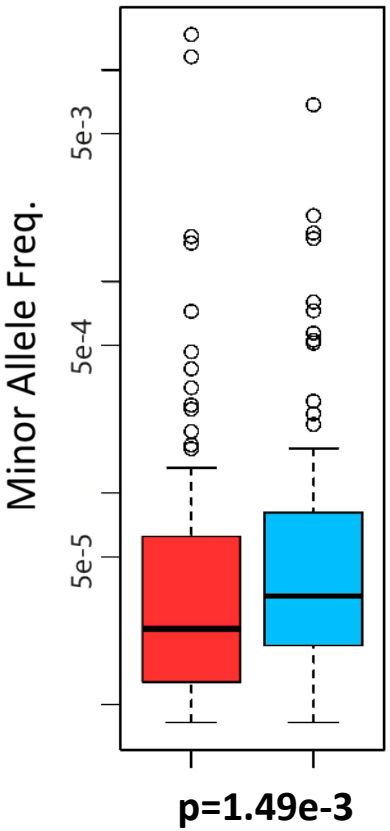
Intra-species conservation of predicted allosteric residues

ExAC

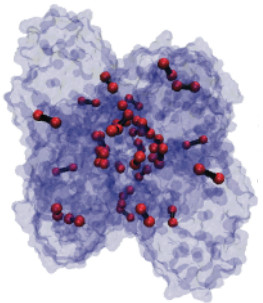




Surface



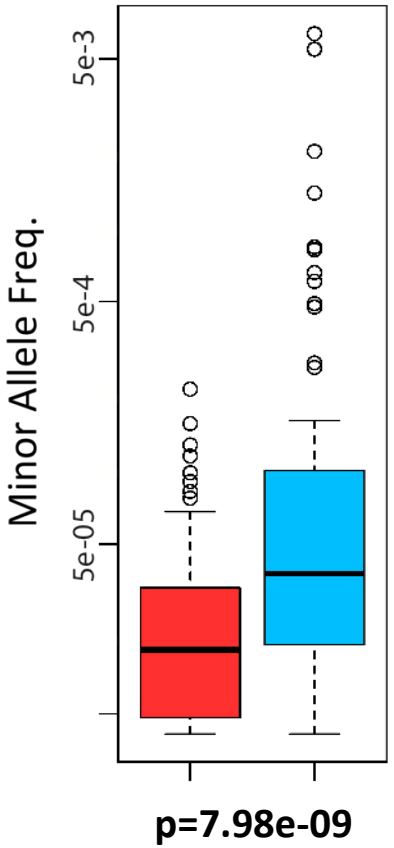
p=1.49e-3



critical

non-critical

Interior

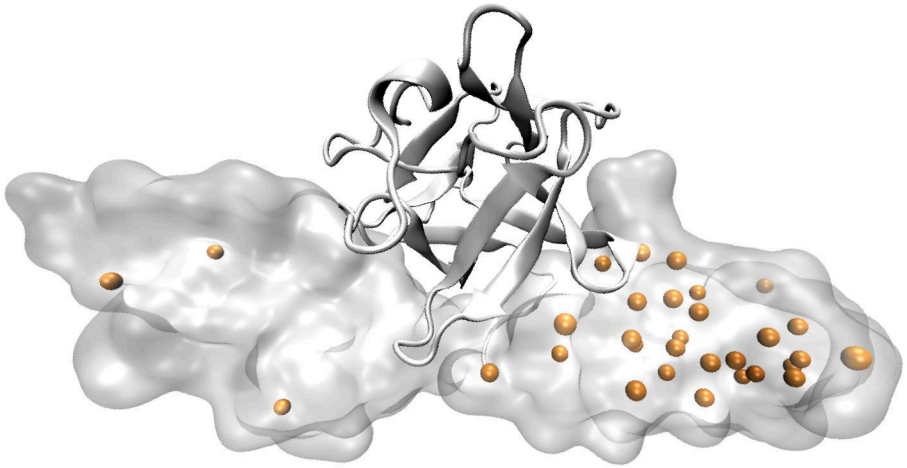


p=7.98e-09

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

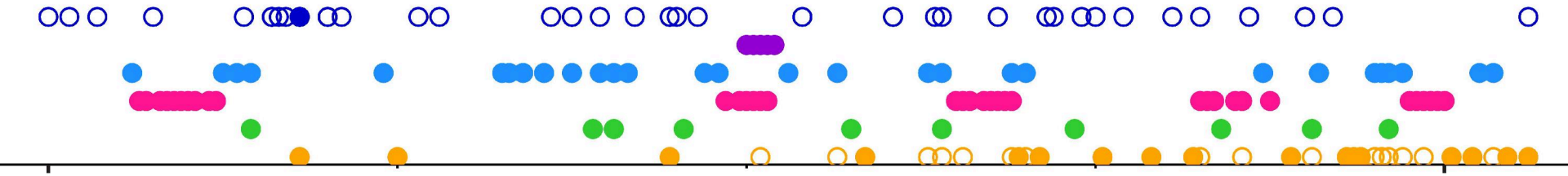
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

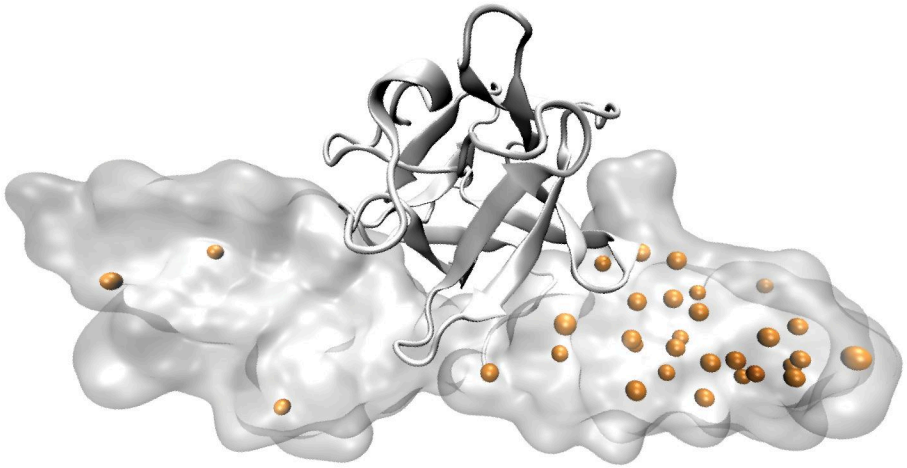
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

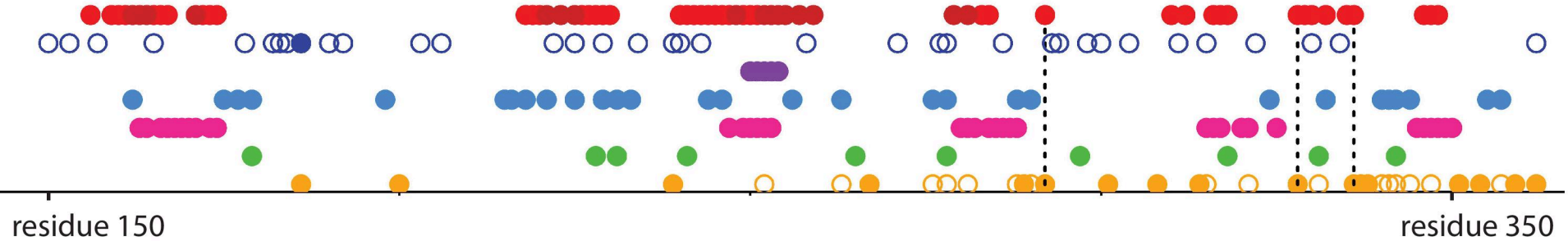
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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)]

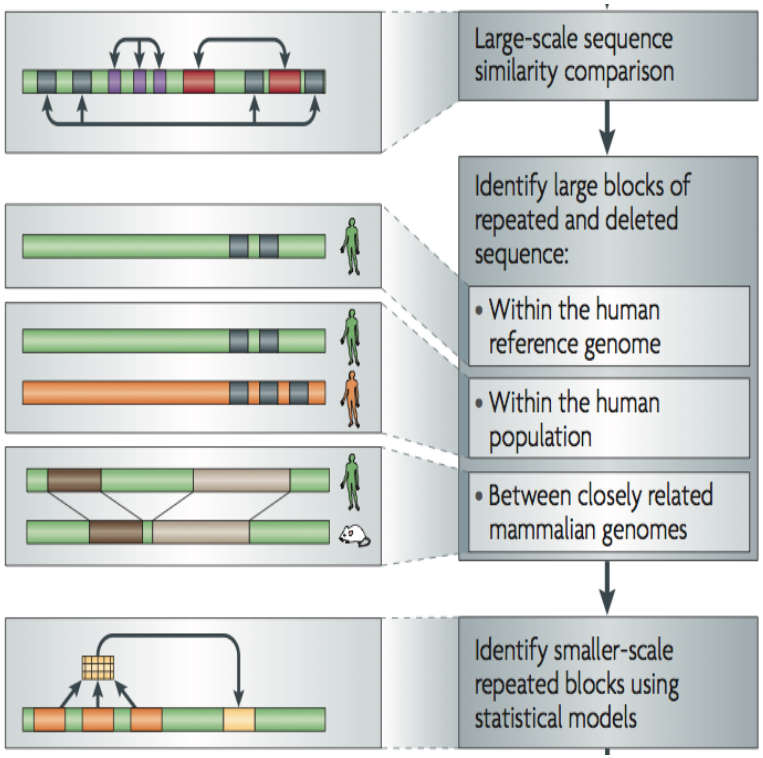
Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

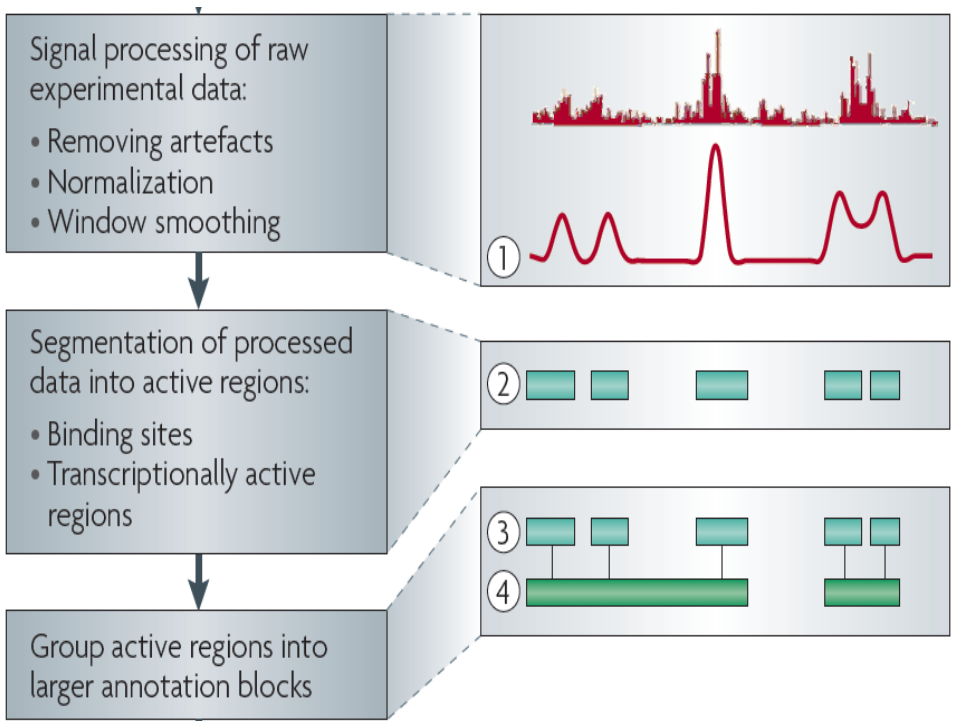
- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

Non-coding Annotations: Overview

Sequence features, incl. **Conservation**

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

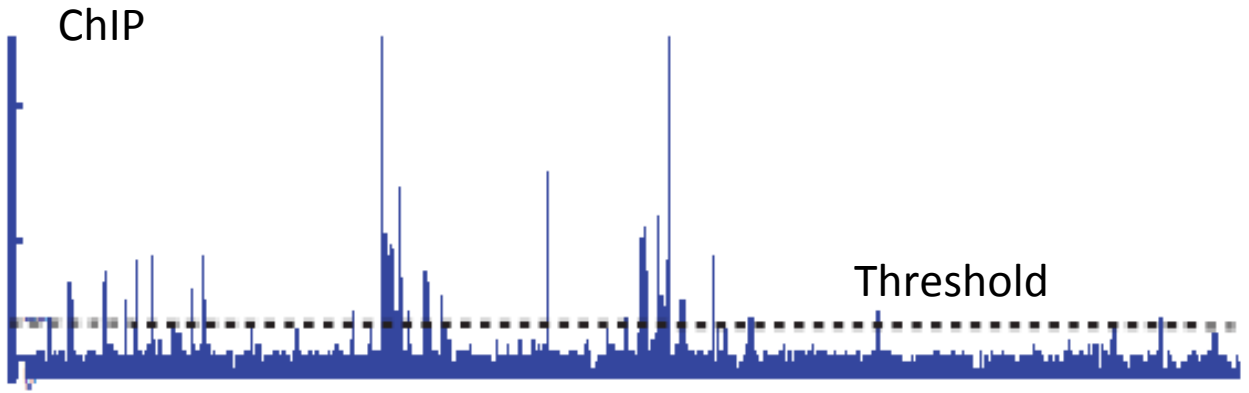




[Alexander et al., *Nat. Rev. Genet.* ('10)]

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

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



Potential Targets →

- Score against the control

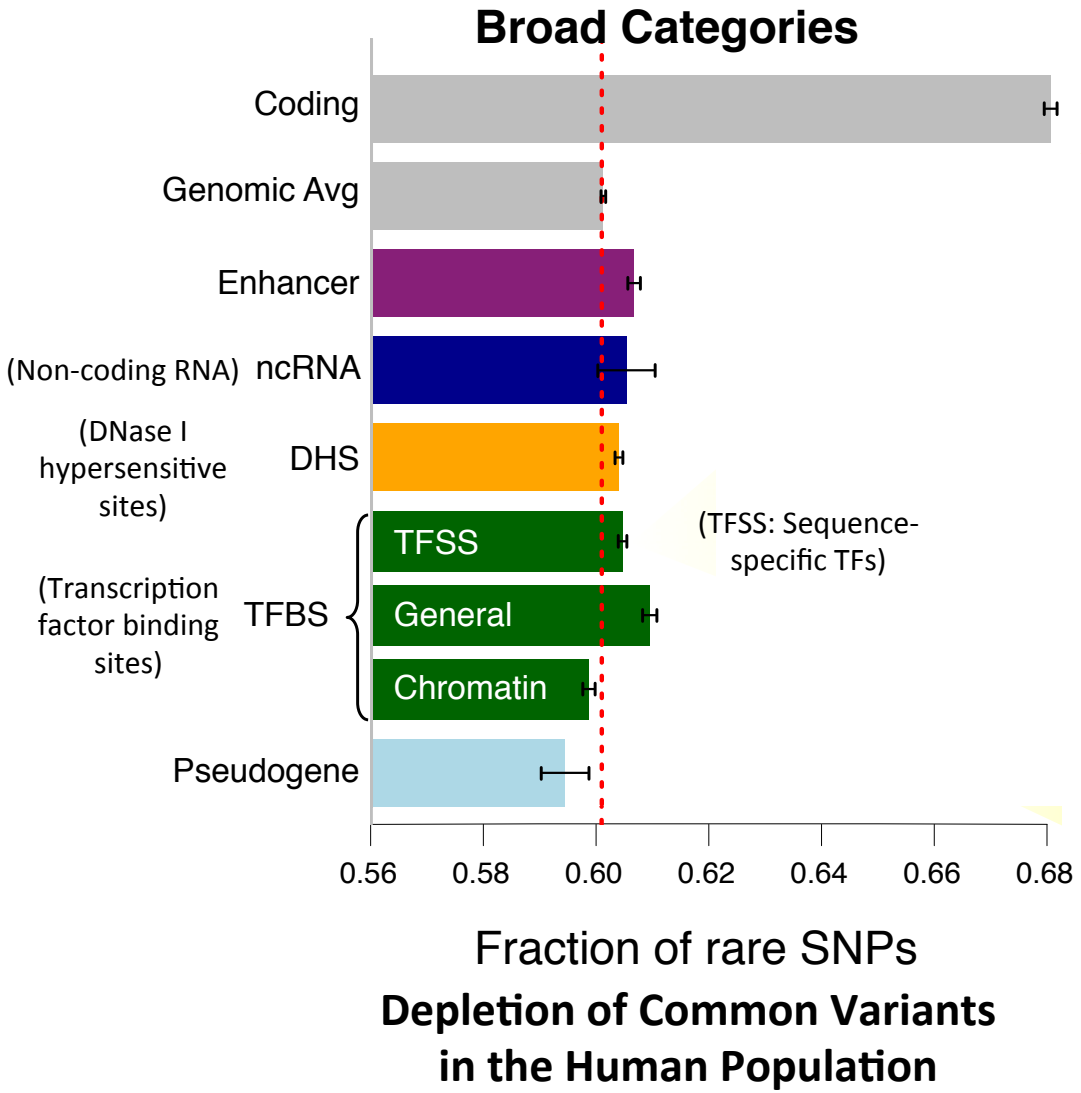
Significantly Enriched targets →

Now an update: "PeakSeq 2" => MUSIC

[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

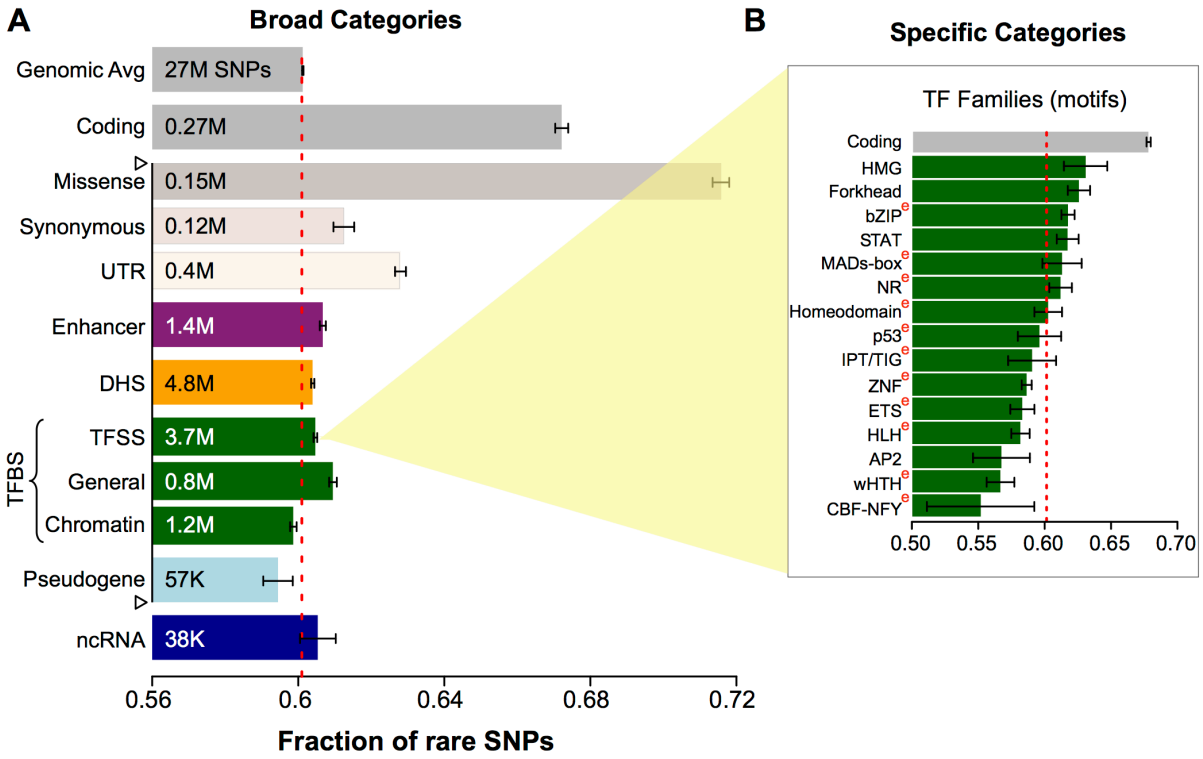


- Broad categories of regulatory regions under negative selection
- Related to:

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

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

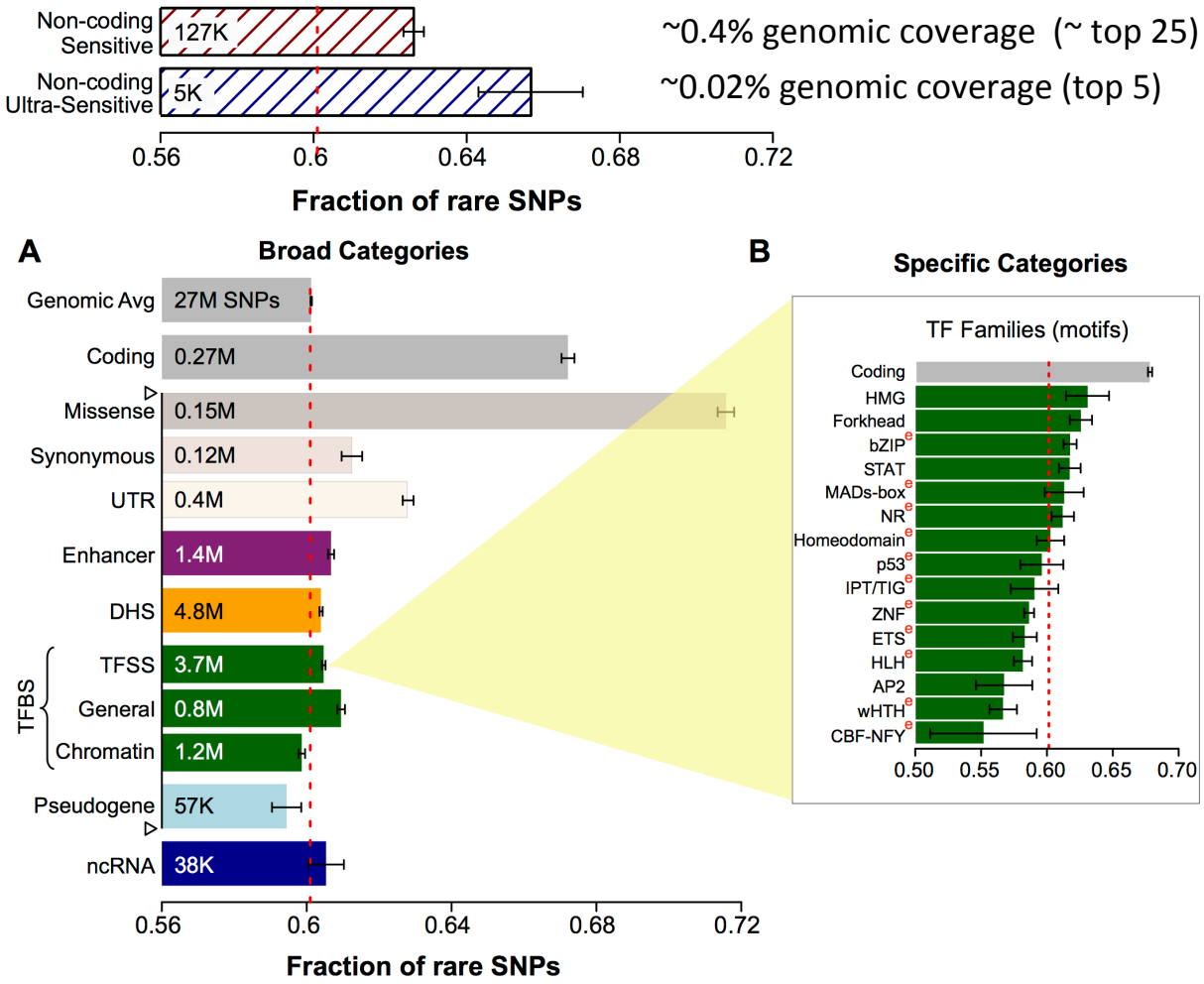
Differential selective constraints among specific sub-categories



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)

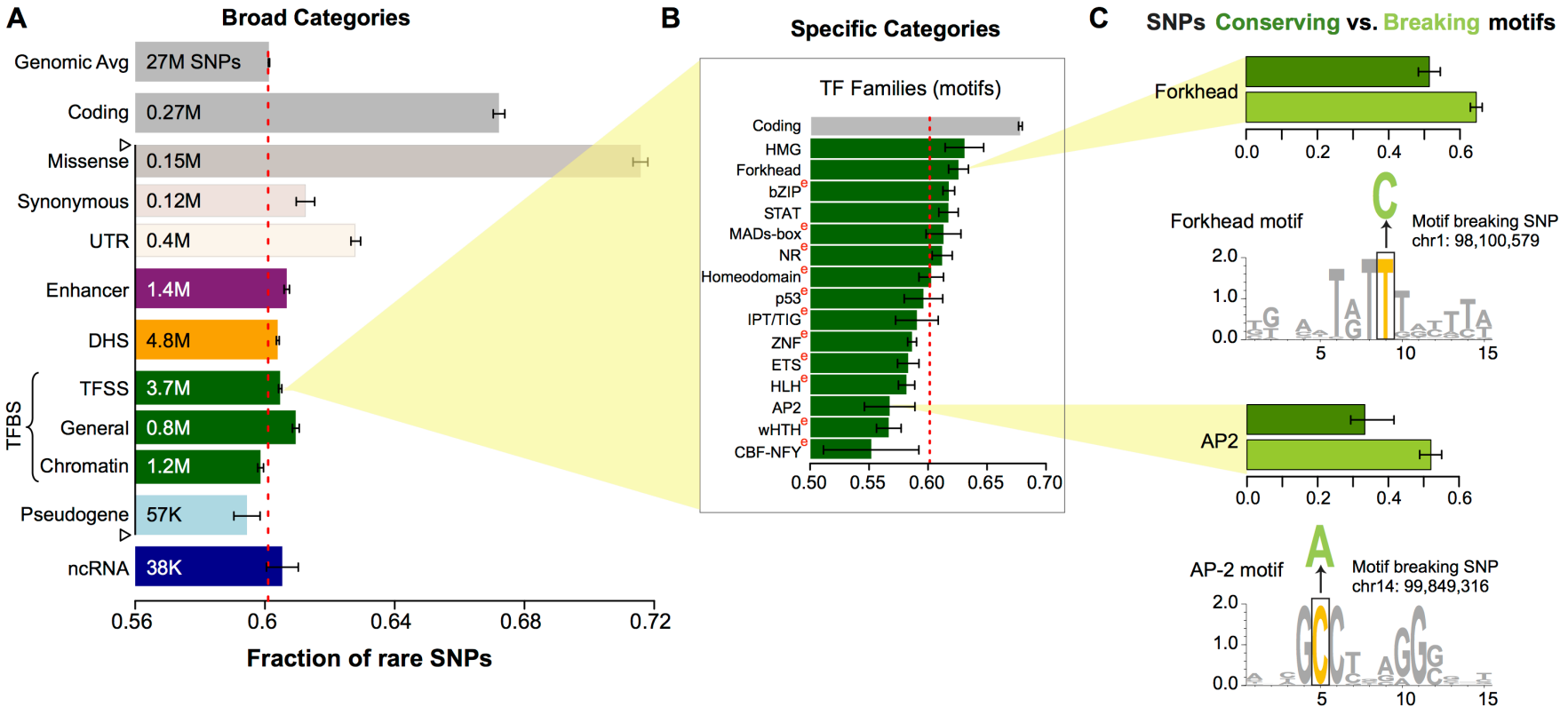
~0.02% genomic coverage (top 5)

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

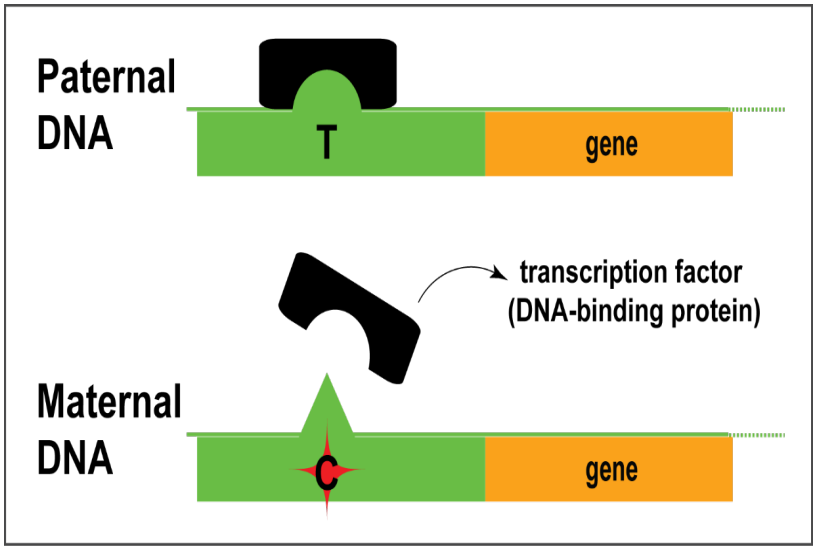


[Khurana et al., *Science* ('13)]

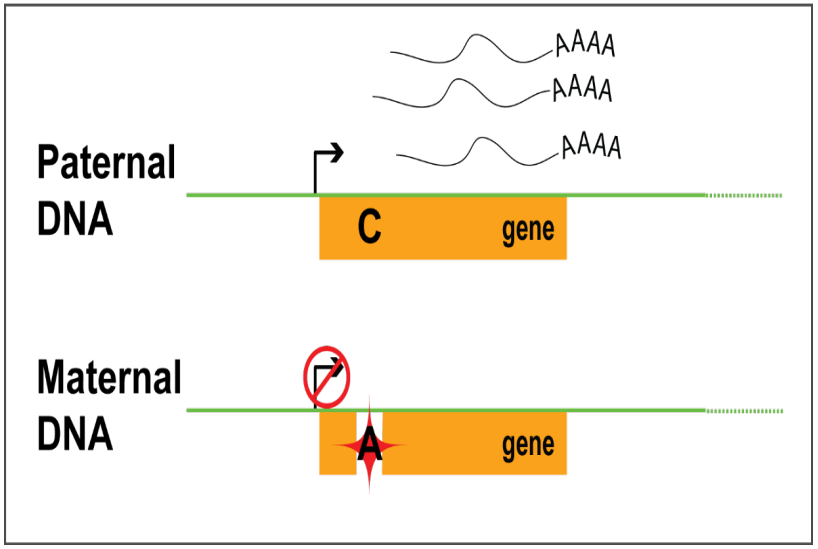
Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

Allele-specific binding and expression



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



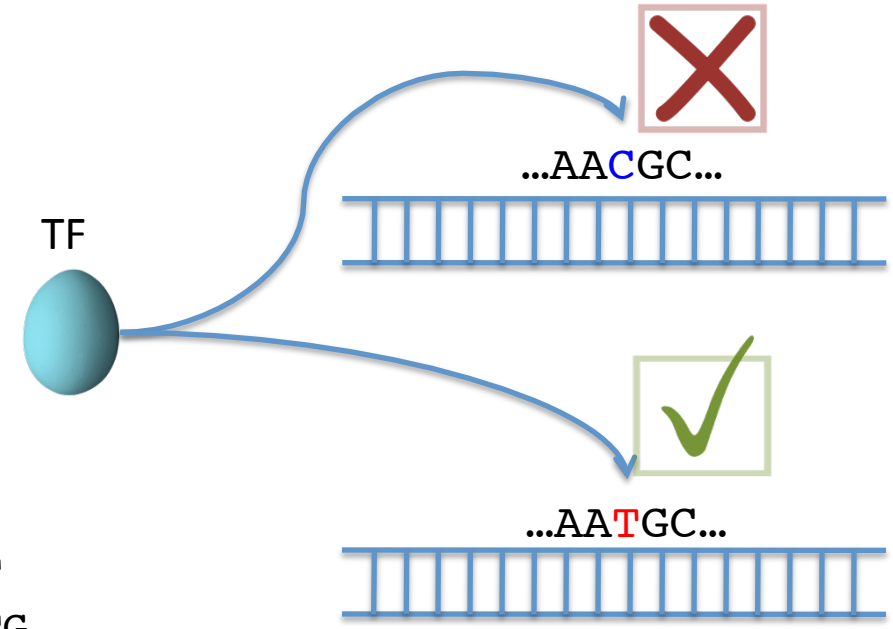
e.g. allele-specific expression (ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

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

```
10 x T
 2 x C
```

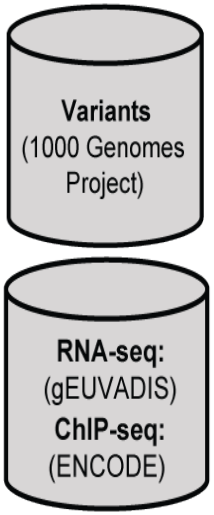


Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants

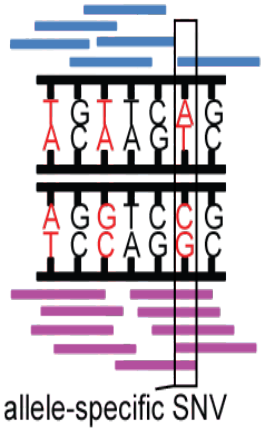
AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes



2. Align ChIP-seq & RNA-seq reads

3. Detect allele-specific variants via a series of filters and tests

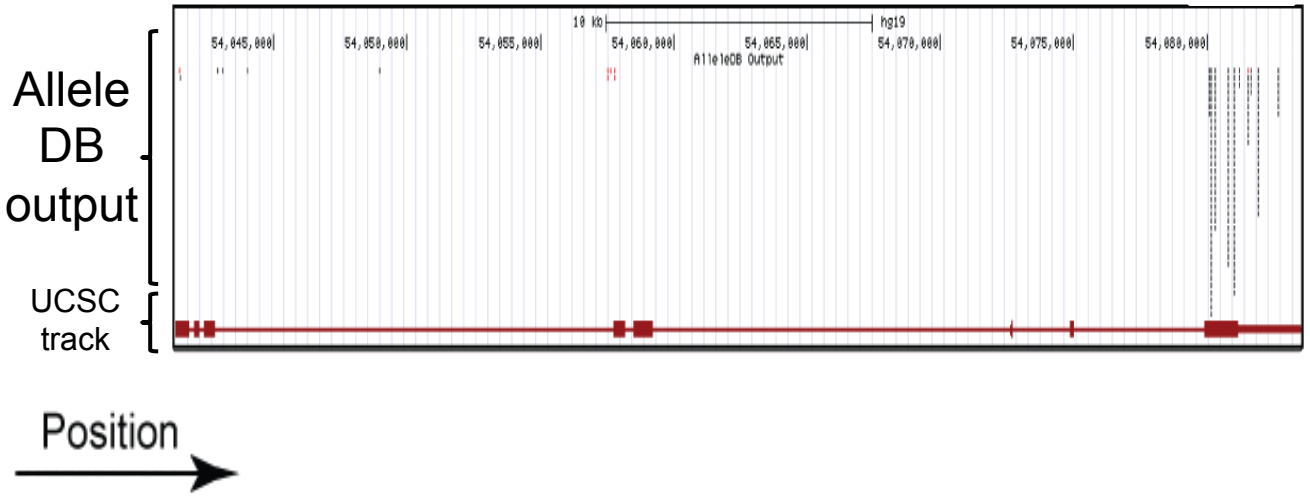


Many Technical Issues: Reference bias, Ambiguous mapping bias, Over-dispersed (non binomial null)

alleledb.gersteinlab.org

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

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

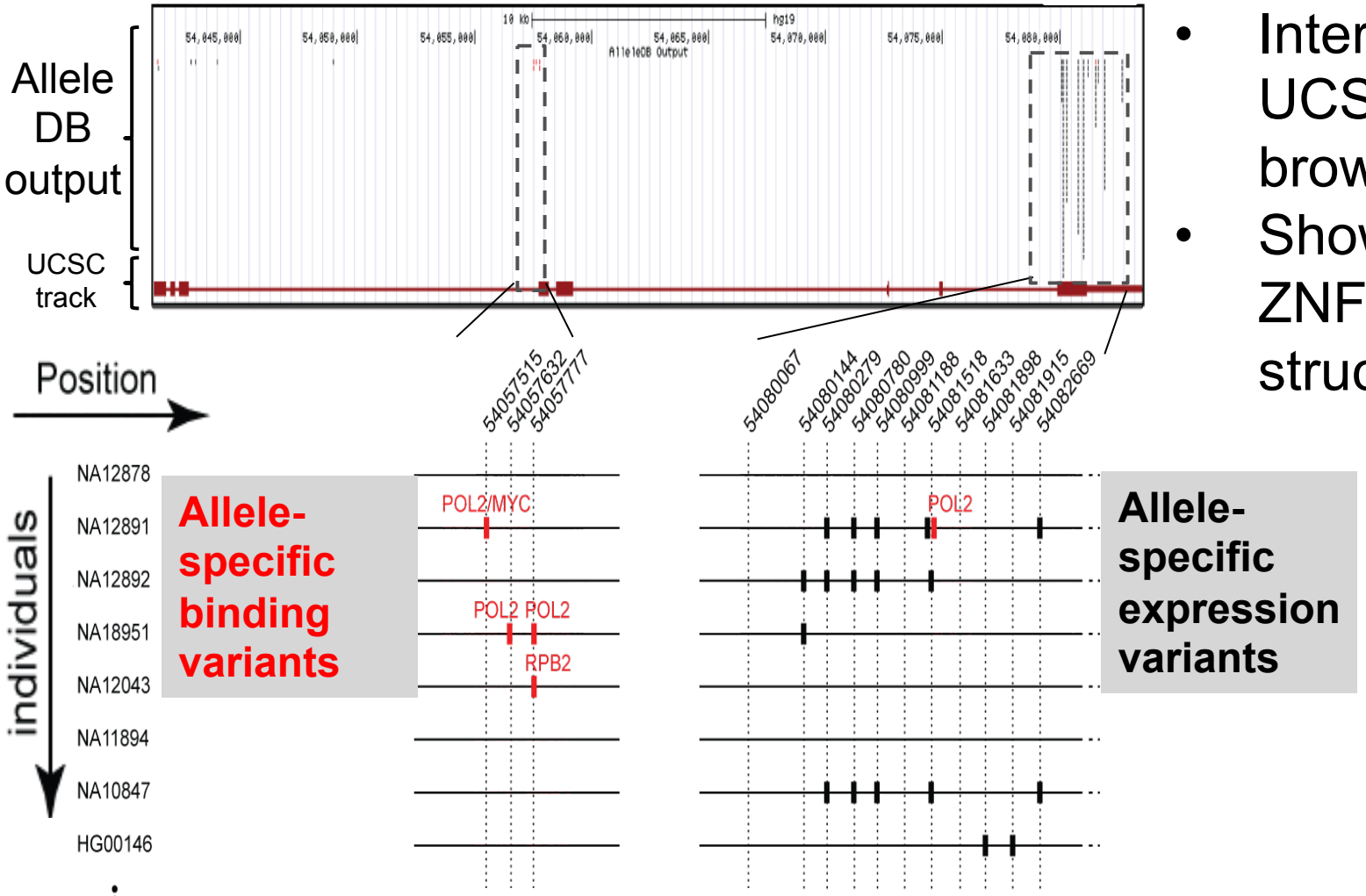


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

alleledb.gersteinlab.org

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

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

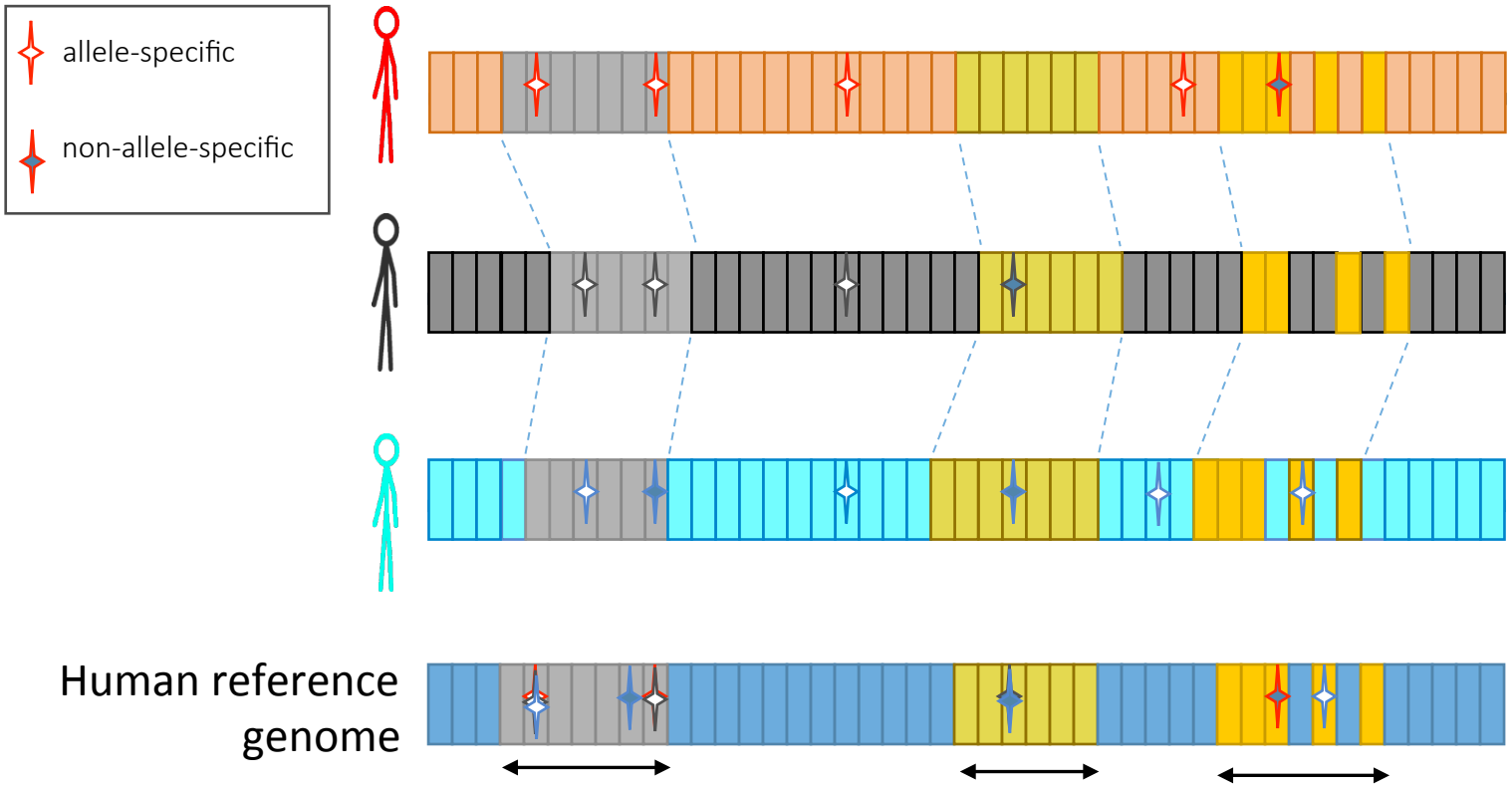


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

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

Collecting ASE/ASB variants into allele-specific genomic regions

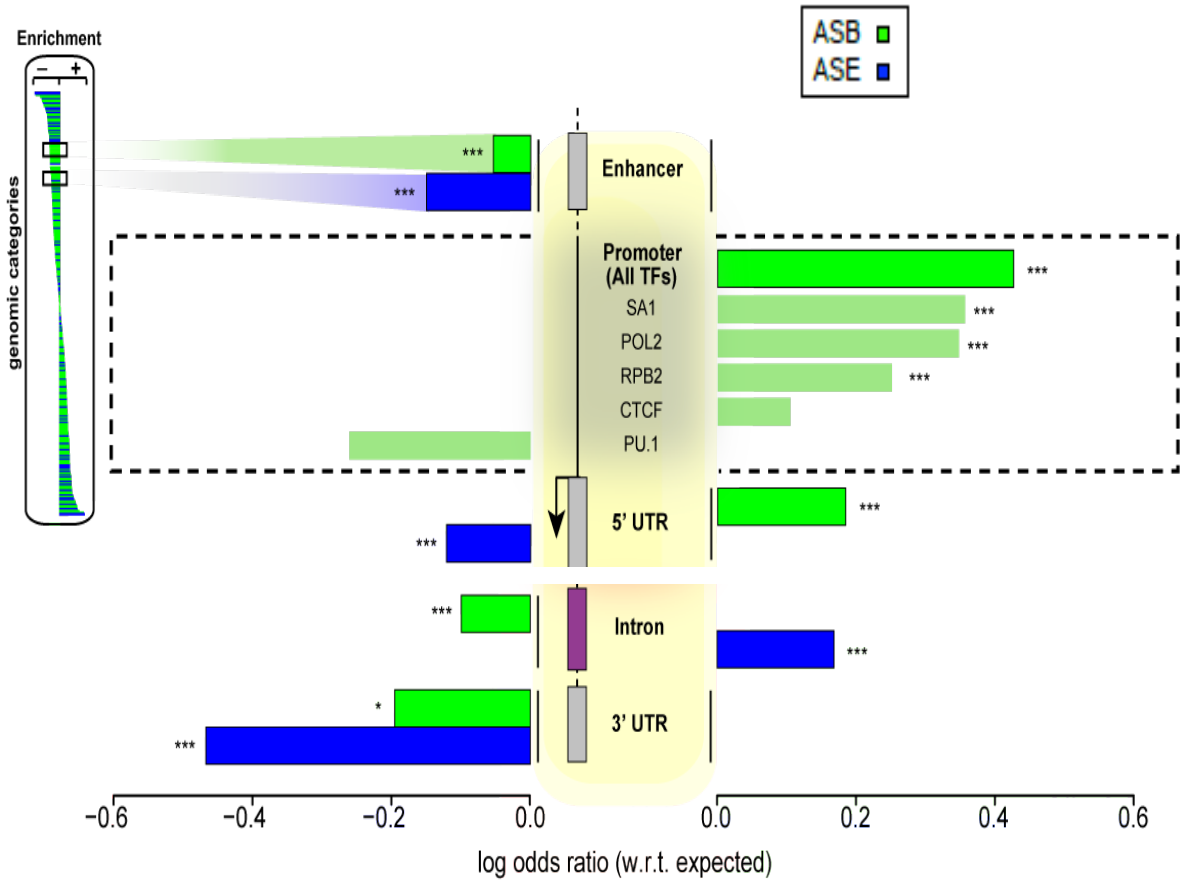
Does a particular genomic element have a higher tendency to be allele-specific? Fisher's exact test, for the **enrichment** of allele-specific variants in the element (with respect to non-allele-specific variants that could potentially be called as allelic)



alleledb.gersteinlab.org

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

Groups of elements that are enriched or depleted in allelic activity

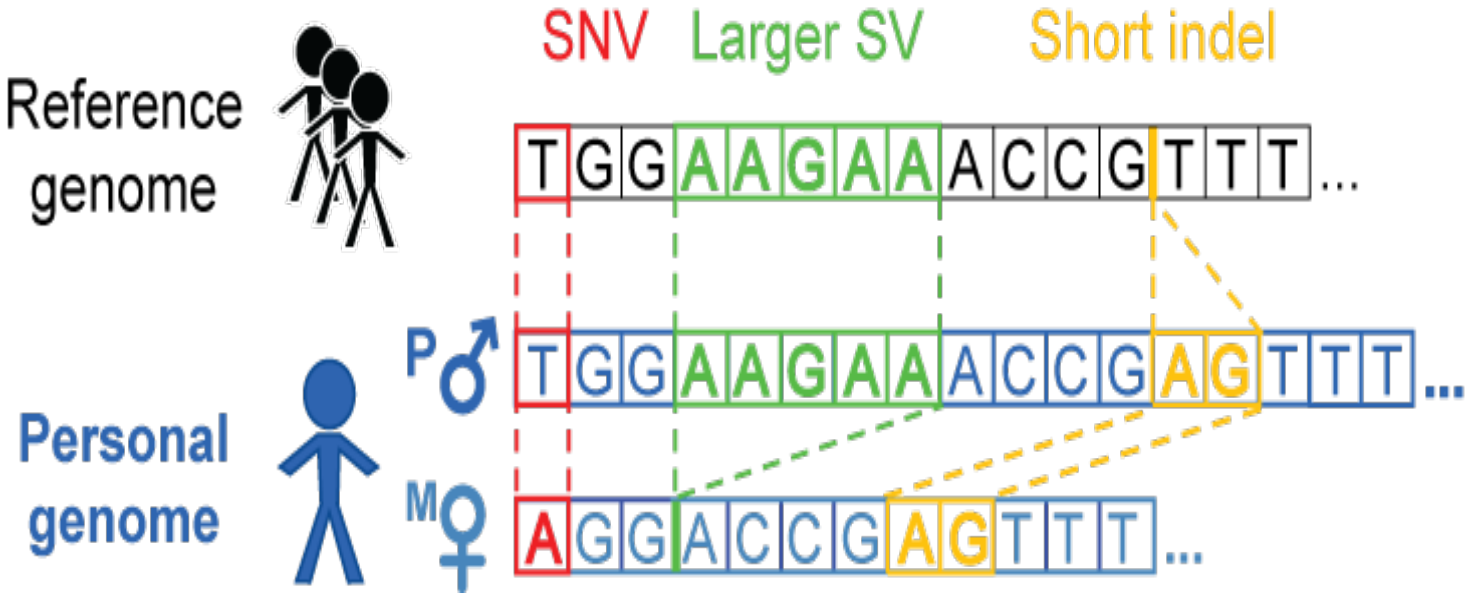


Chen J. *et al.* (*Nature Commun*, in press)

Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

How to build a personal genome



(fasta; reference)

+

(vcf, variants phased or unphased)

↓

(fasta; for each haplotype)

alleleseq.gersteinlab.org

Rozowsky *et al. Mol Syst Biol* (2011)

Why the personal genome (PG) should be the platform for functional genomics

1. **Diploid**
   --Ability to incorporate private variants of any size
   --exhibit phase information
2. **Scale easily with more samples (v graph genome) and improving sequencing technologies: longer reads and more accurate phase information**
3. **Very useful in functional genomic assay analyses**
   a) read alignment
   b) RNA-seq quantification
   c) allele-specific analyses

Some construction considerations

1. **Choice of call set(s)**

-- e.g. different versions of 1000GP call sets

2. **Choice of variants**

-- e.g. SVs or indels or SNVs only

3. **Choice of reference**

-- choose the reference genome in which the call set is derived from

4. **Assessment of call set quality**

-- e.g. analysis of Mendelian inconsistency in family data

NA12878 family of PGs we already have

	Source	Refgen	Depth	Variants
1	1000 Genomes Project (1000GP) pilot	hg18	60x	SNVs, indels, deletions (including 33 from fosmid sequencing)
2	GATK Best Practices v3 (UnifiedGenotype)	hg19	64x	SNVs, indels
3	GATK Best Practices v4 (HaplotypeCaller, PCR-free)	hg19	64x	SNVs, indels
4	1000GP Phase 3 SNVs-only	hg19	7.4x	SNVs
5	1000GP Phase 3 SNVs-indels	hg19	7.4x	SNVs, indels
6	1000GP Phase 3 SNVs-indels-SVs	hg19	7.4x	SNVs, indels, SVs

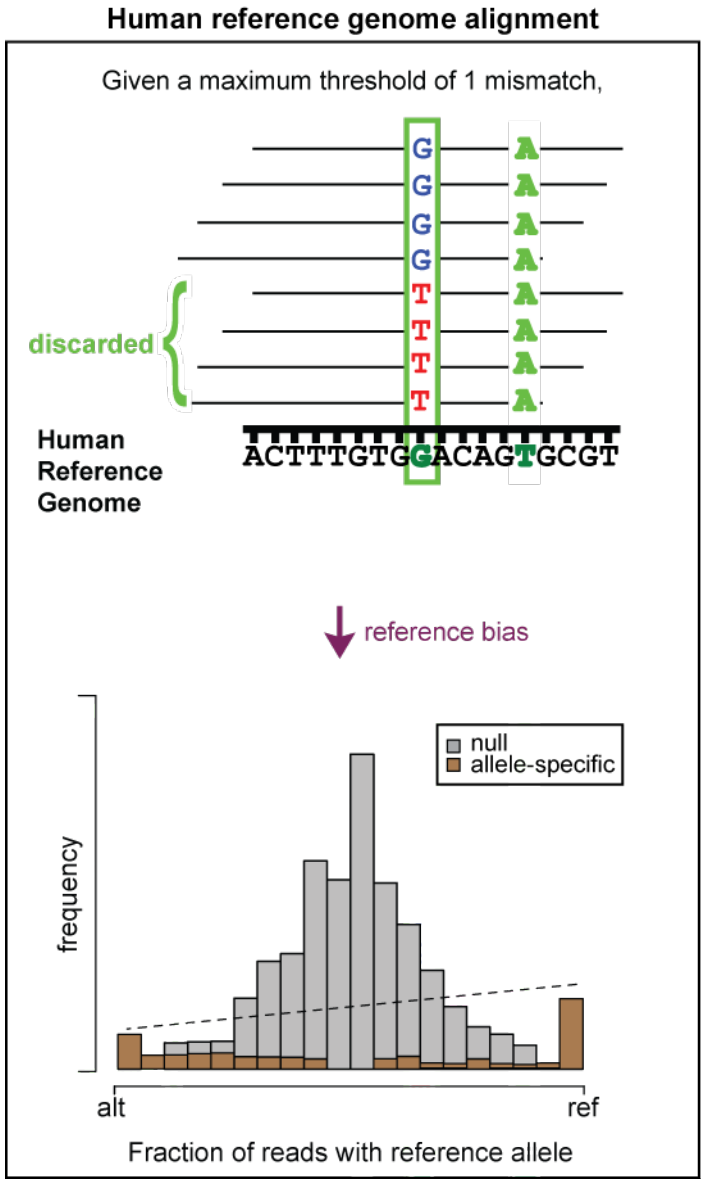
Alignment gets better as variant sets get more complete: NA12878 Pol2 ChIP-seq (ENCODE)

	Ref genome	Pgenome: SNVs only	Pgenome: SNVs + indels only	Pgenome: SNVs + indels + SVs
Reads processed	208,051,087			
# reads uniquely aligned	171,944,588 (82.65%)	172,591,380 (82.96%)	172,738,321 (83.03%)	172,743,175 (83.03%)
# reads that multimap	17,826,675 (8.57%)	17,795,258 (8.55%)	17,782,167 (8.55%)	17,779,800 (8.55%)

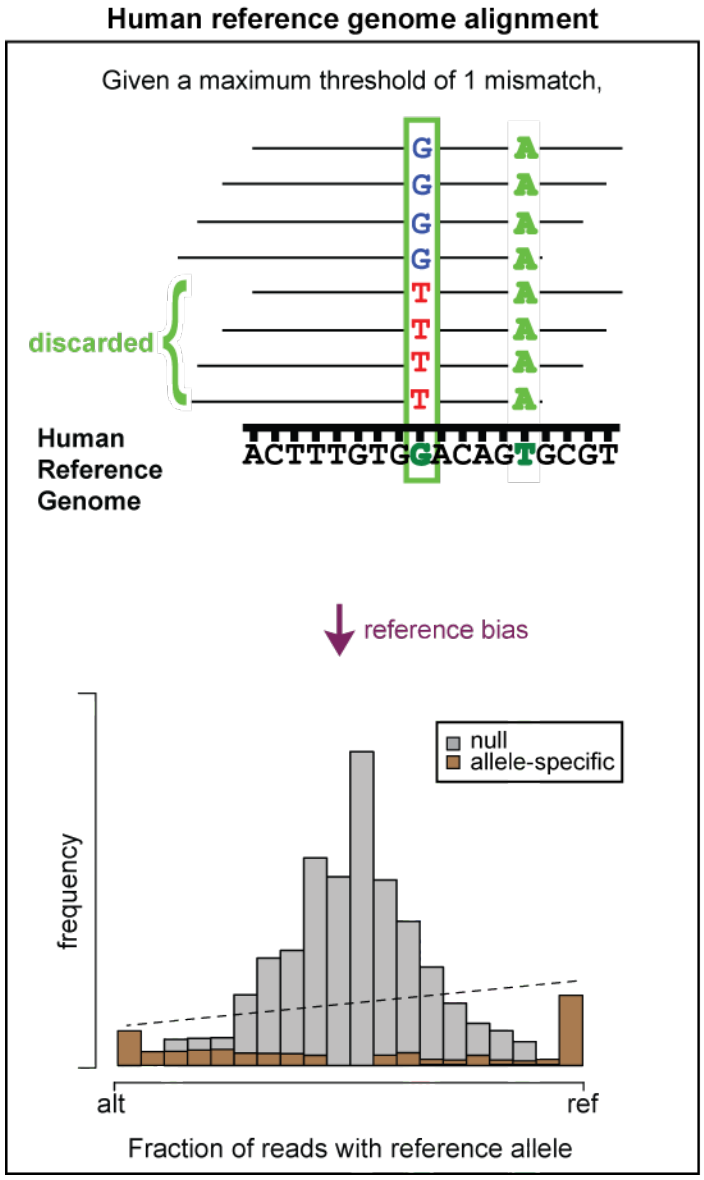
Alignment gets better as variant sets get more complete: NA12878 RNA-seq (Kilpinen *et al.* 2013)

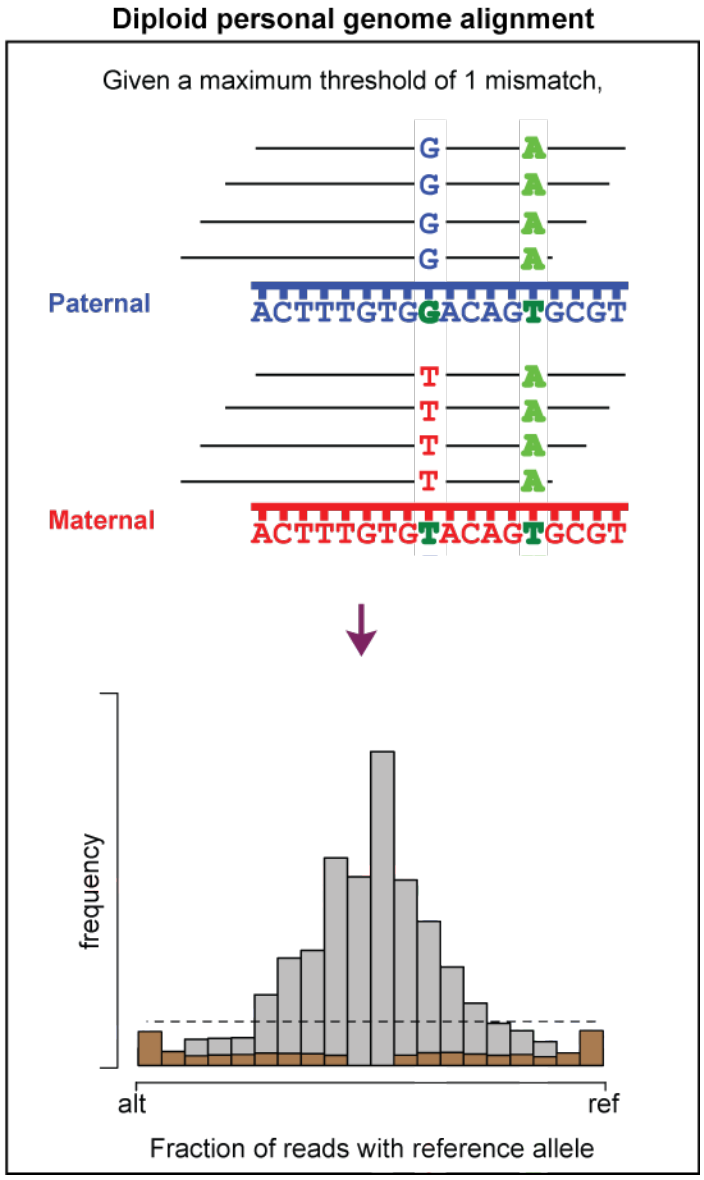
	Ref genome	Pgenome: snvs only	Pgenome: snvs + indels only	Pgenome: snvs + indels + SVs
Reads processed	37,558,398			
# reads uniquely aligned	25,303,498 (67.37%)	25,486,837 (67.86%)	25,538,449 (68.00%)	25,568,042 (68.08%)
# reads that multimap	4,041,495 (10.76%)	4,010,417 (10.68%)	4,012,297 (10.68%)	3,972,990 (10.58%)

PG alleviates reference bias in alignment



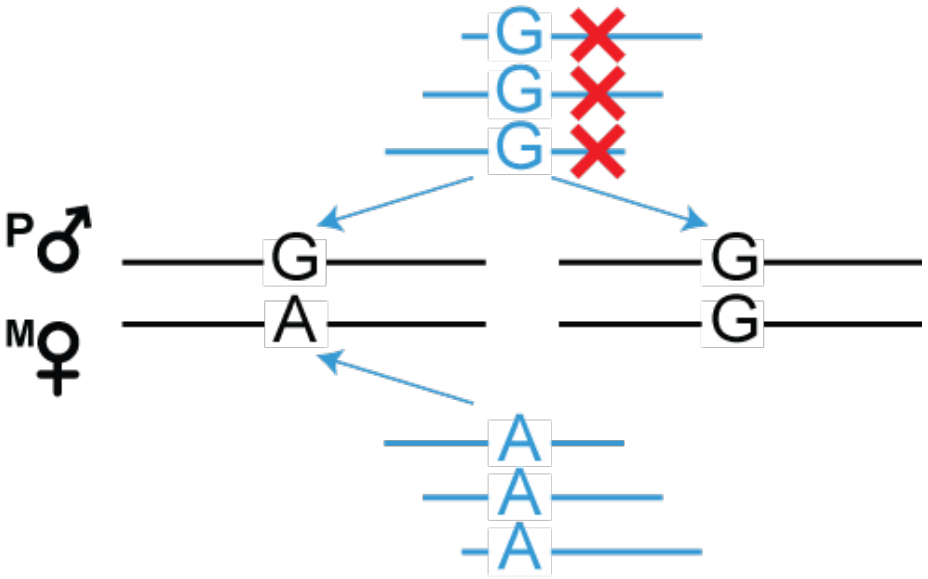
PG alleviates reference bias in alignment





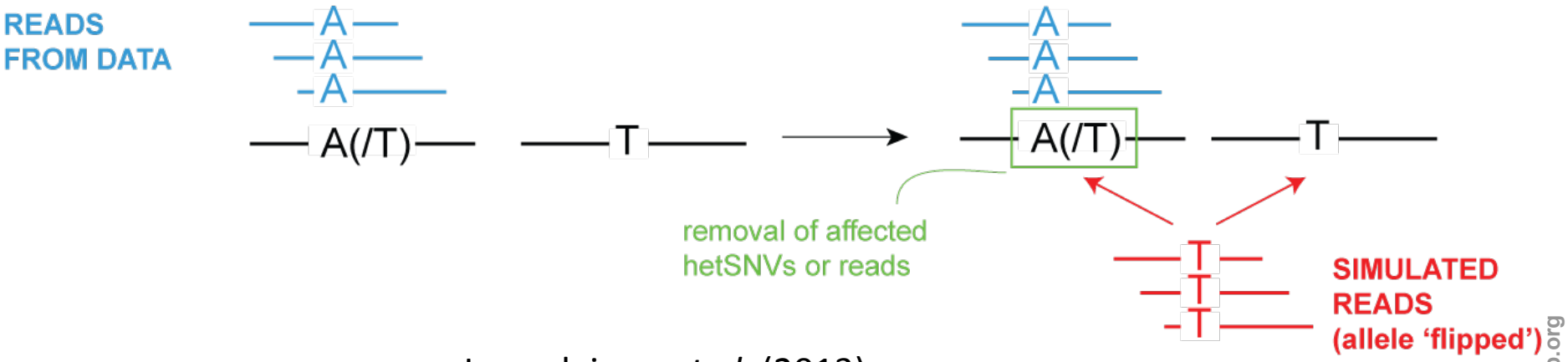
Ambiguous mapping bias due to sequence similarity

- For AS analyses, discard reads that multi-map



Account for ambiguous mapping bias

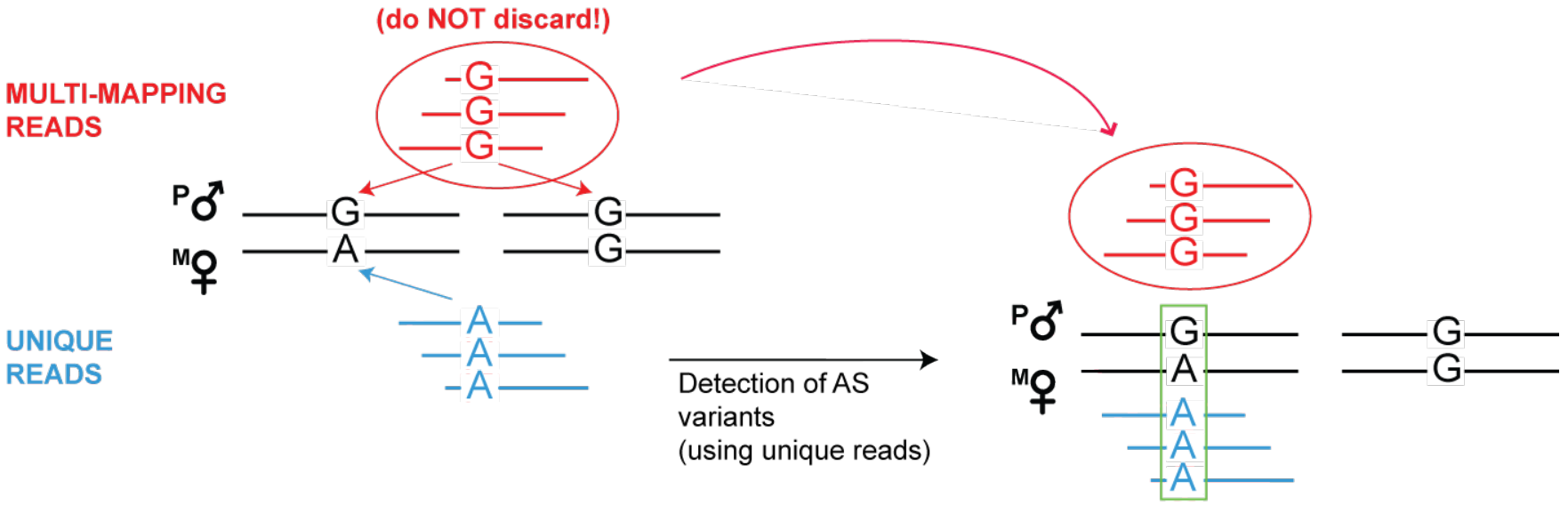
- Using the reference genome, new simulated reads are created where alleles of the original reads are flipped (at het SNV positions)



Lappalainen *et al*. (2013)
Panousis *et al.* (2014)
Van de Geijn *et al.* (2015)

PG facilitates the resolution of ambiguous mapping bias

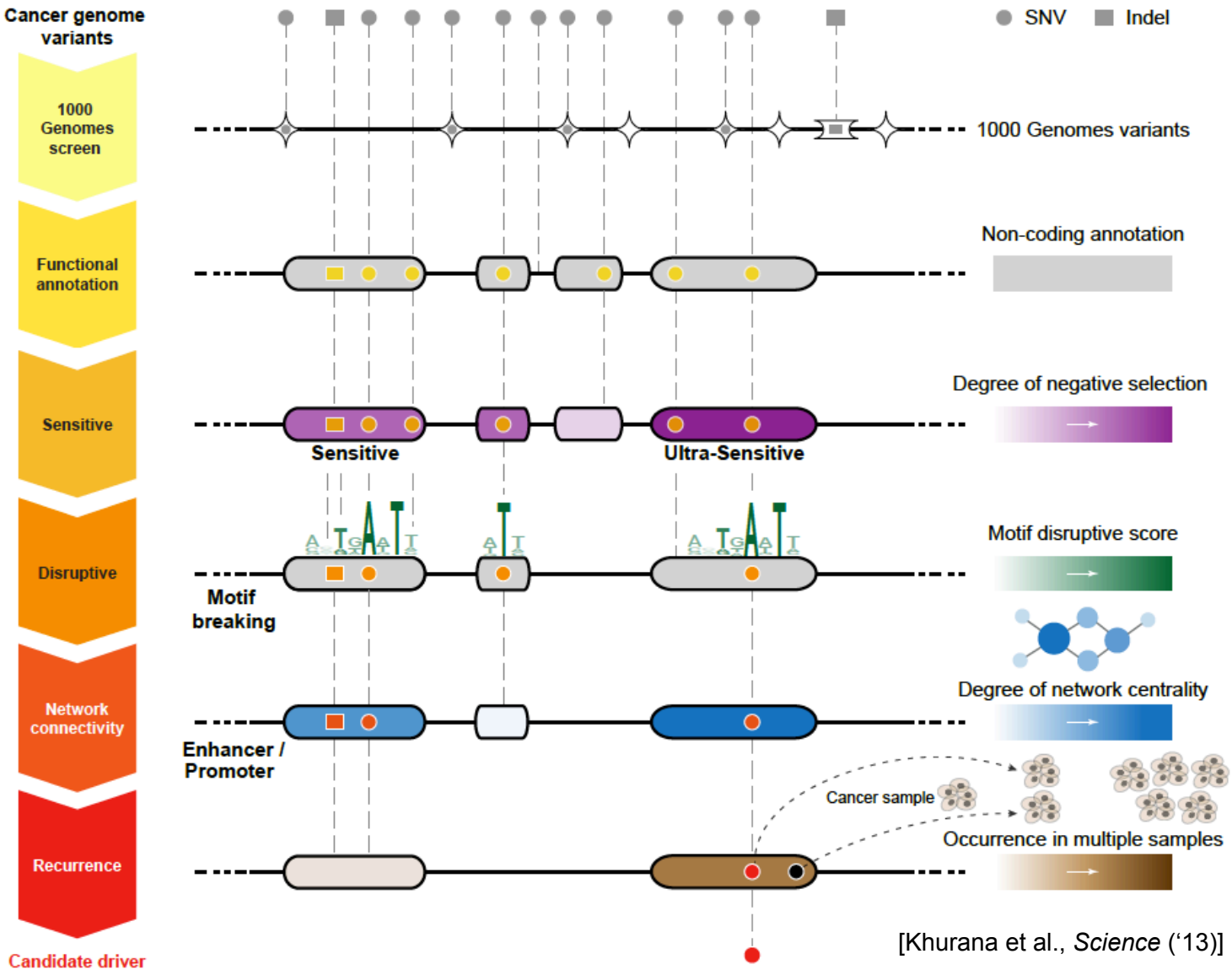
- Using the personal genome, we do not need to simulate reads.
- We can directly test affected sites using multi-mapping read pile



Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

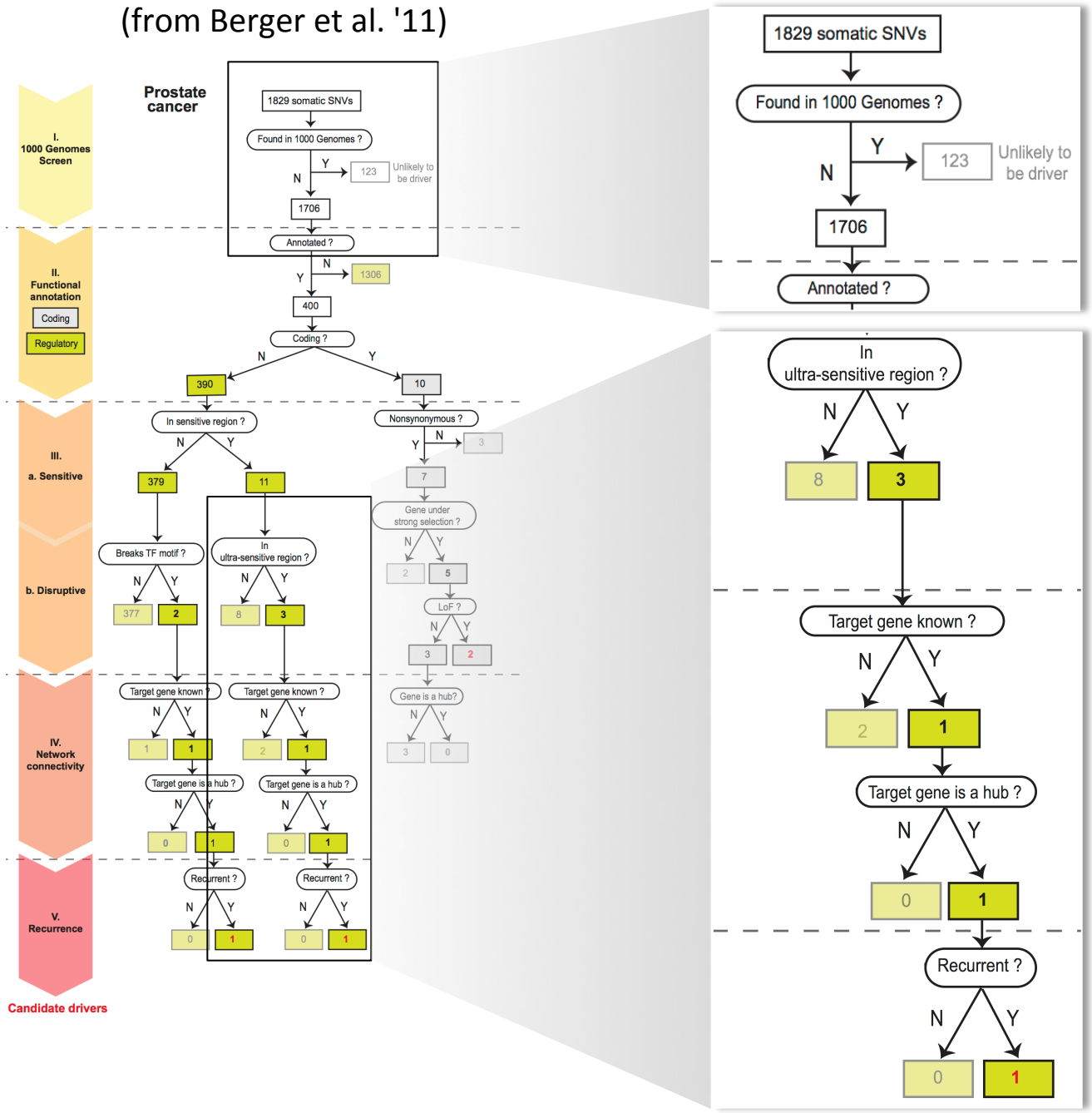
- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

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

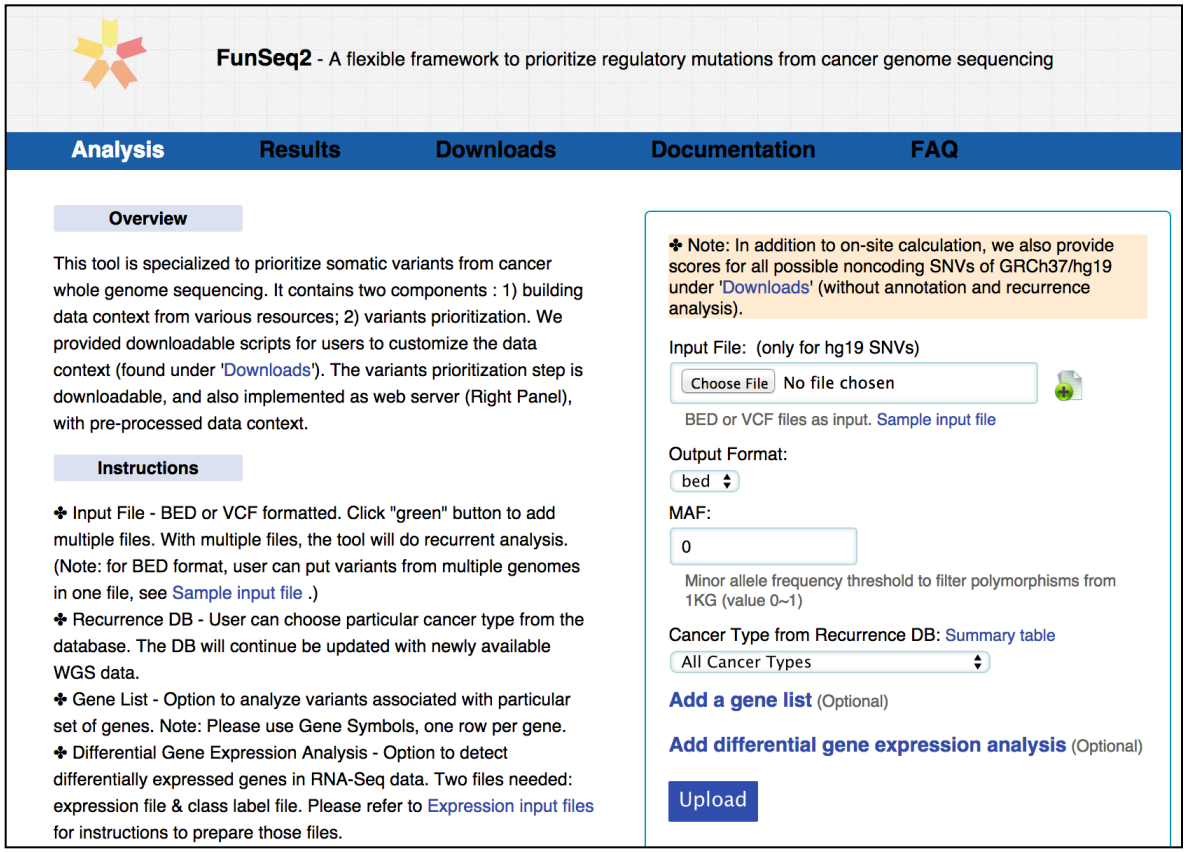


[Khurana et al., *Science* ('13)]

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

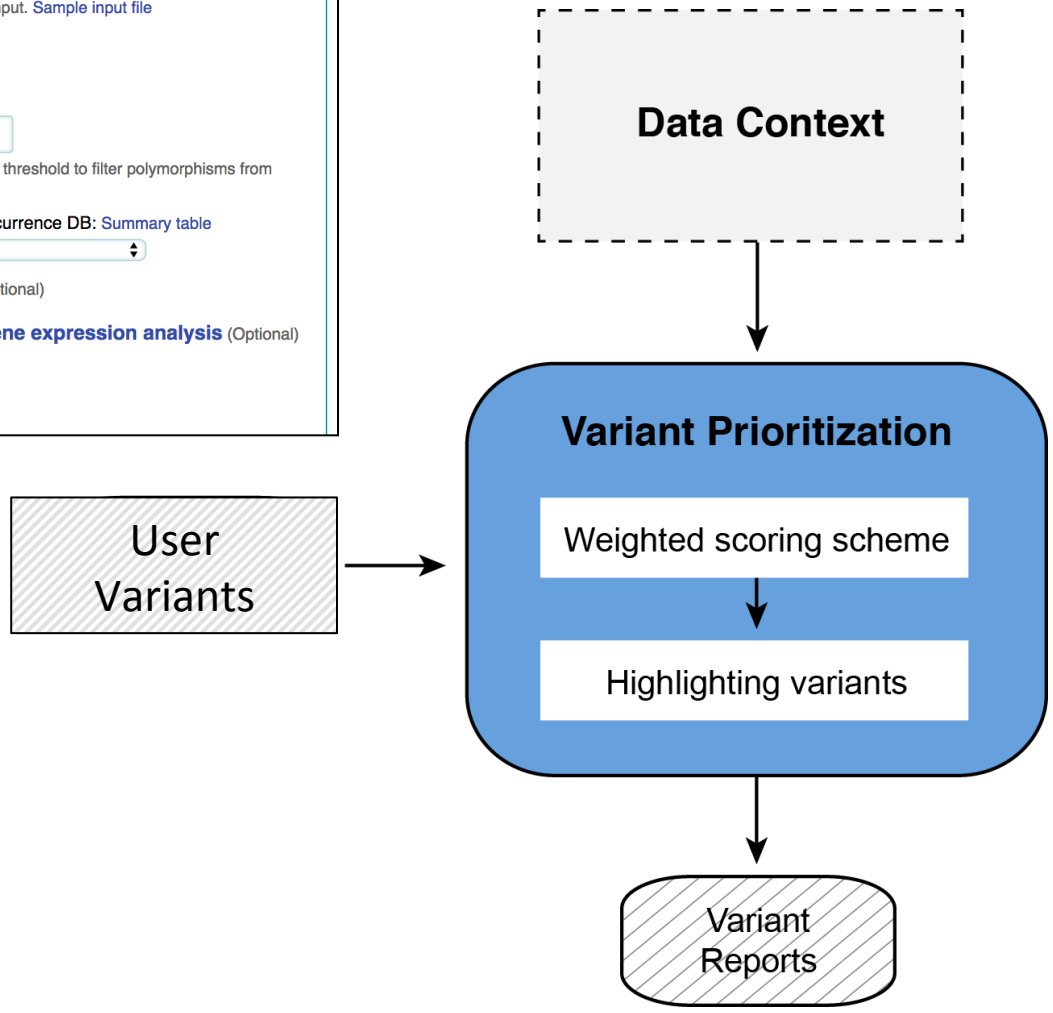


[Khurana et al., *Science* ('13)]



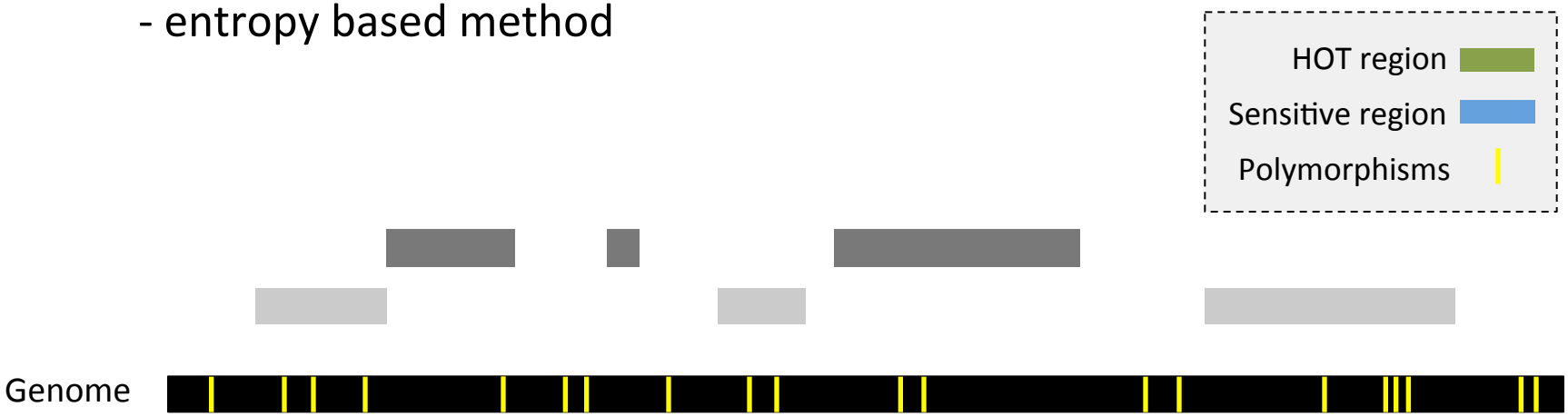
Site integrates user variants with large-scale context

FunSeq.gersteinlab.org



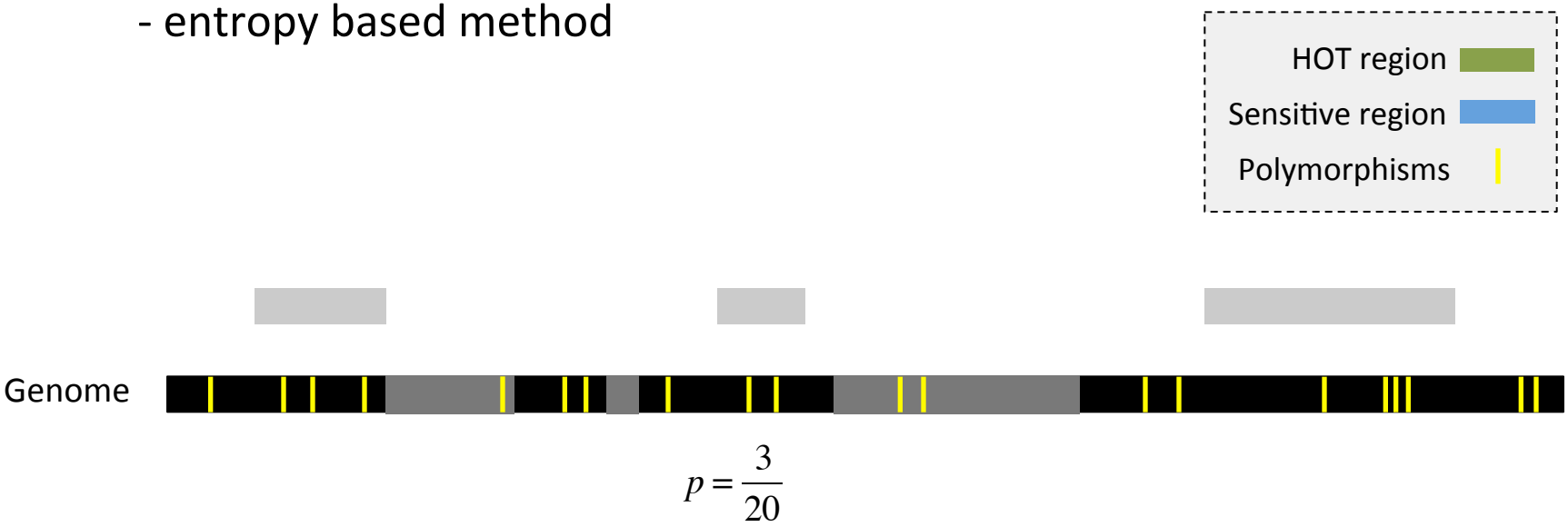
[Fu et al., GenomeBiology ('14)]

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



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

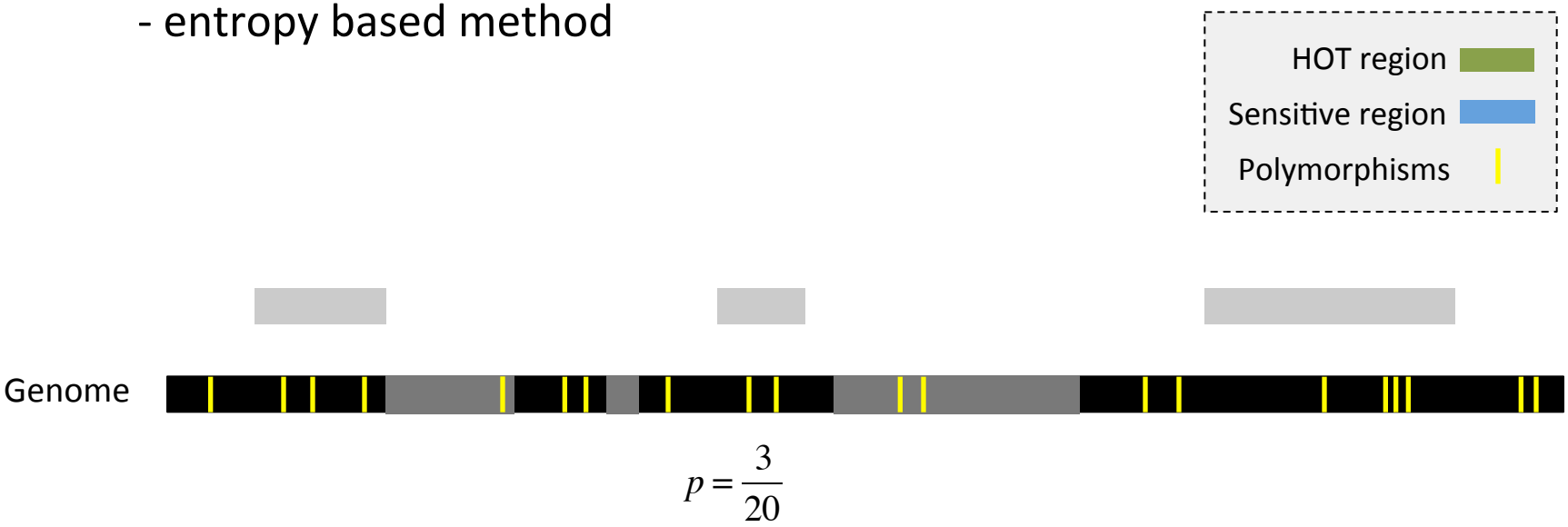
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



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

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

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



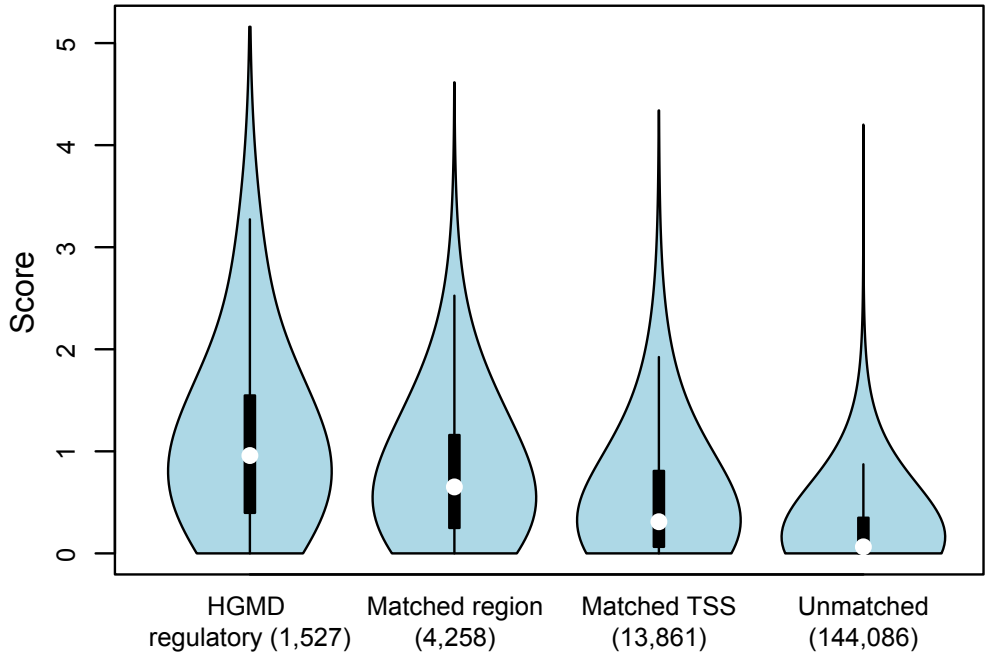
Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

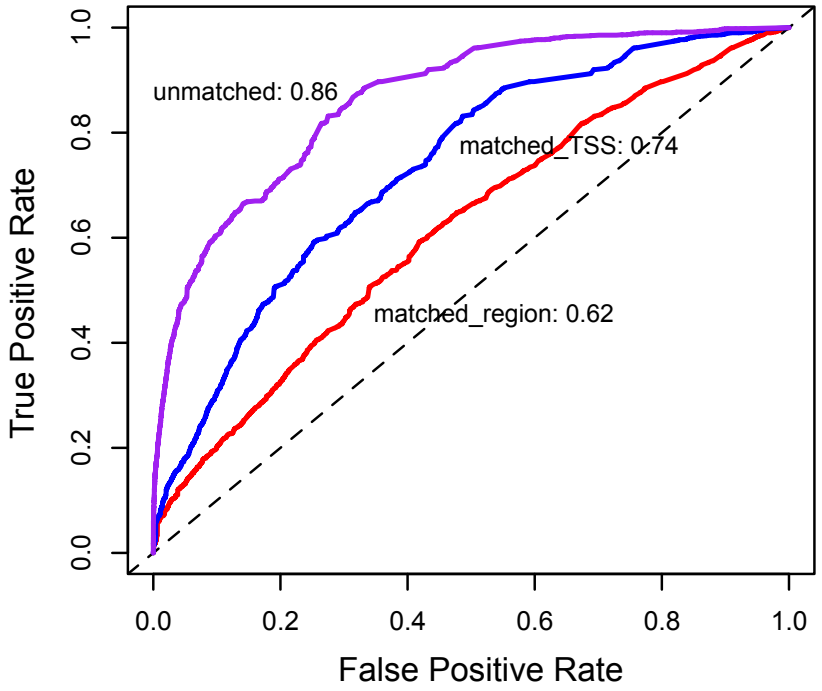
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

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

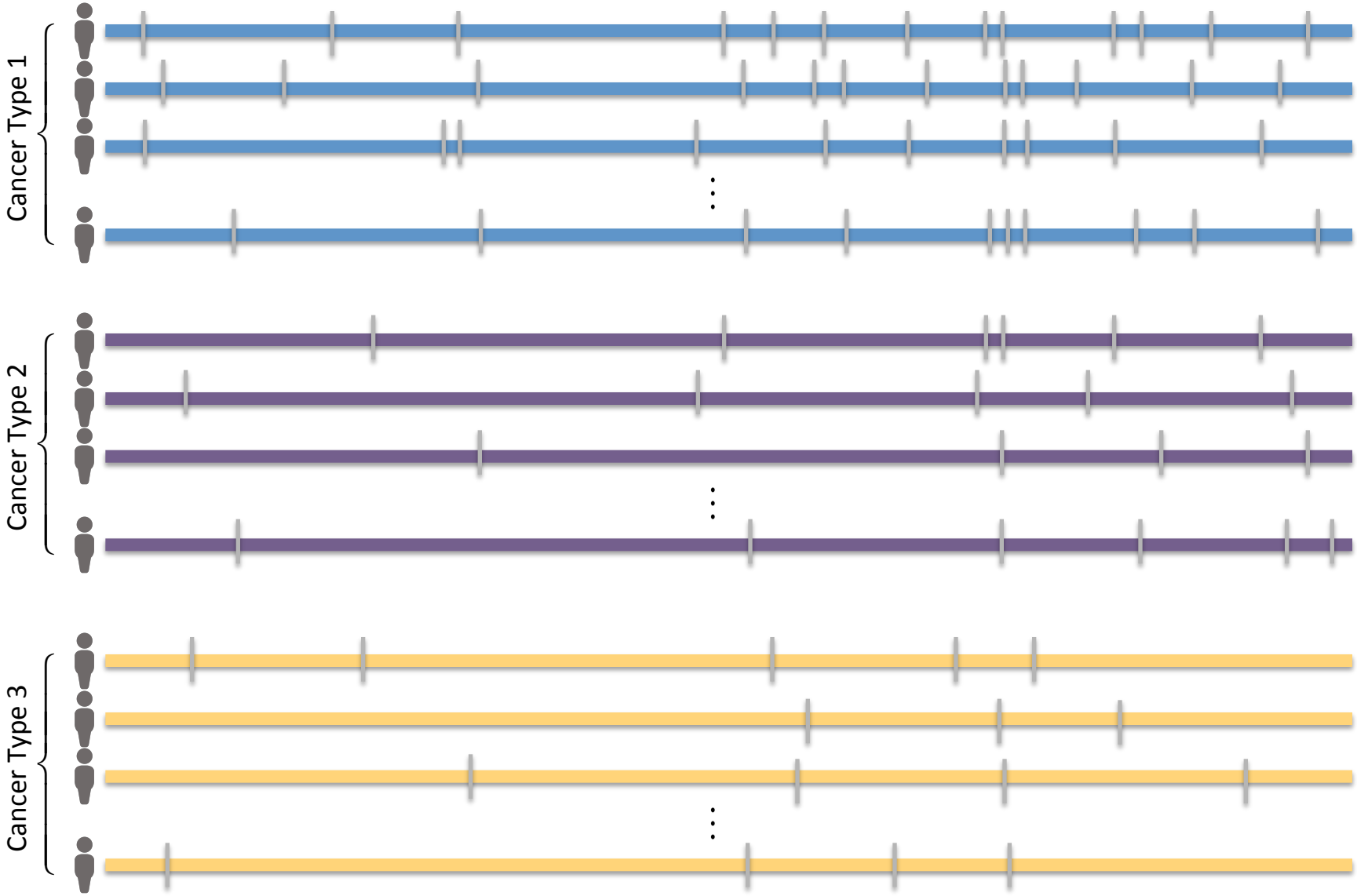
Ritchie et al., Nature Methods, 2014

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

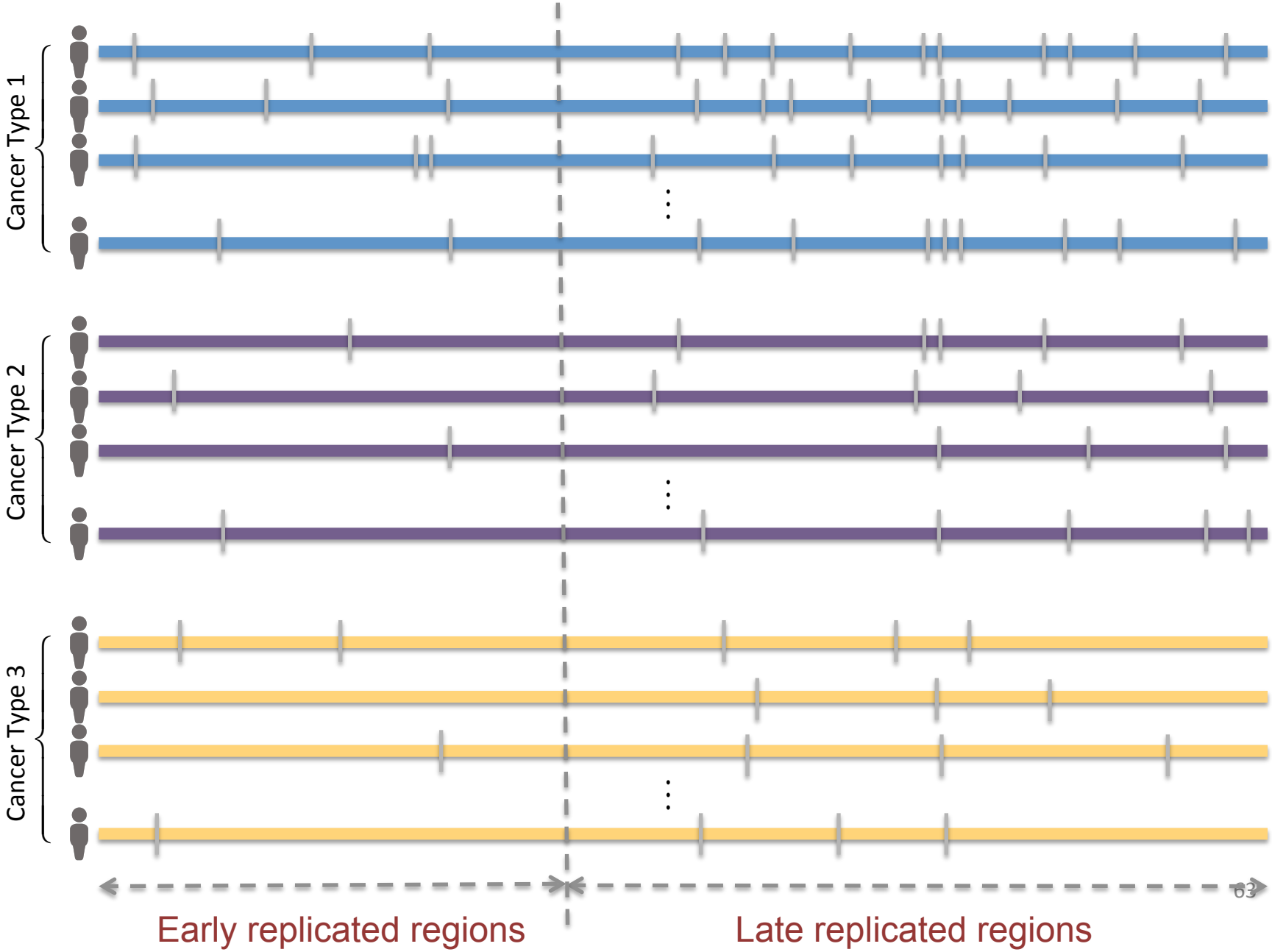
Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

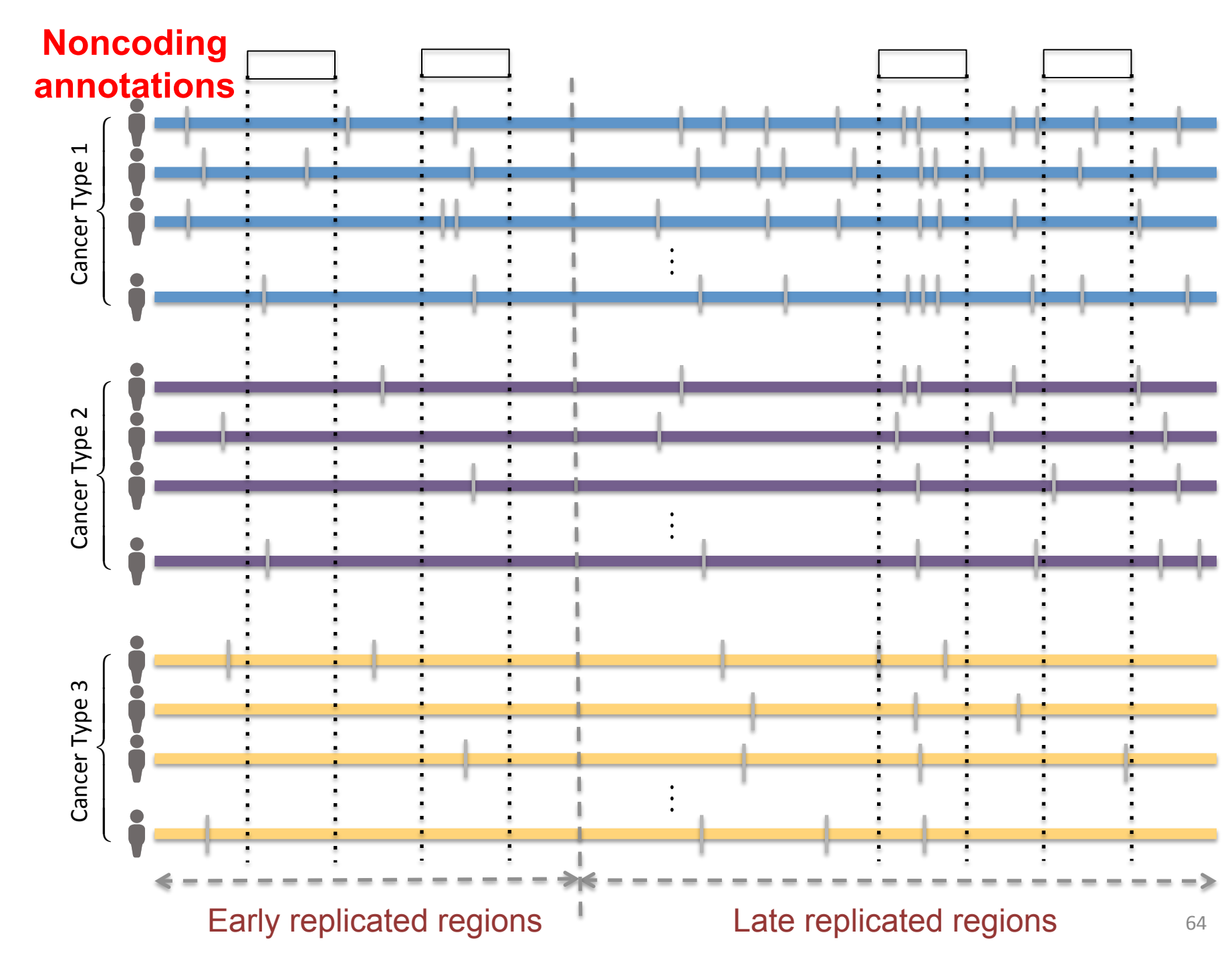
- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

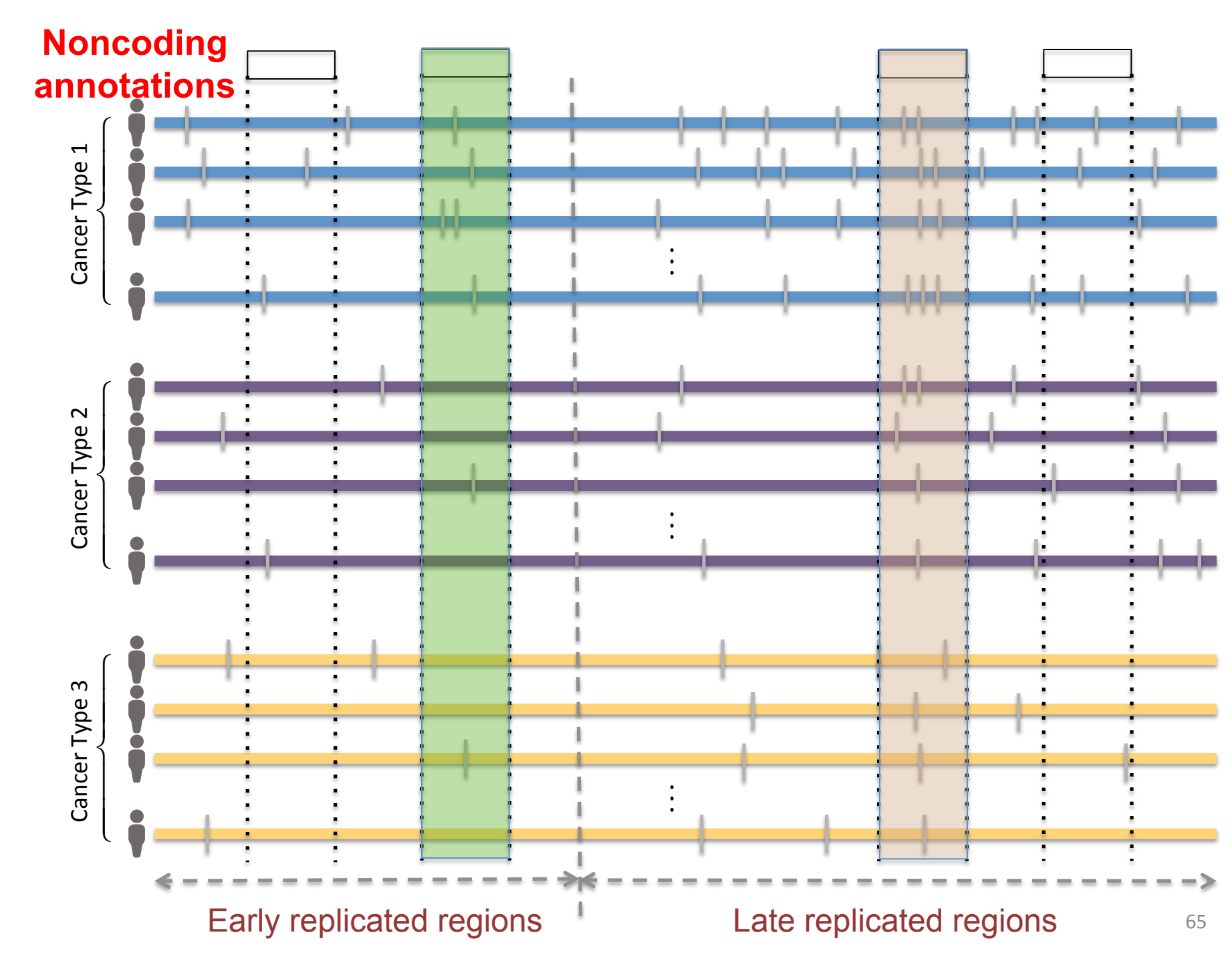
Mutation recurrence



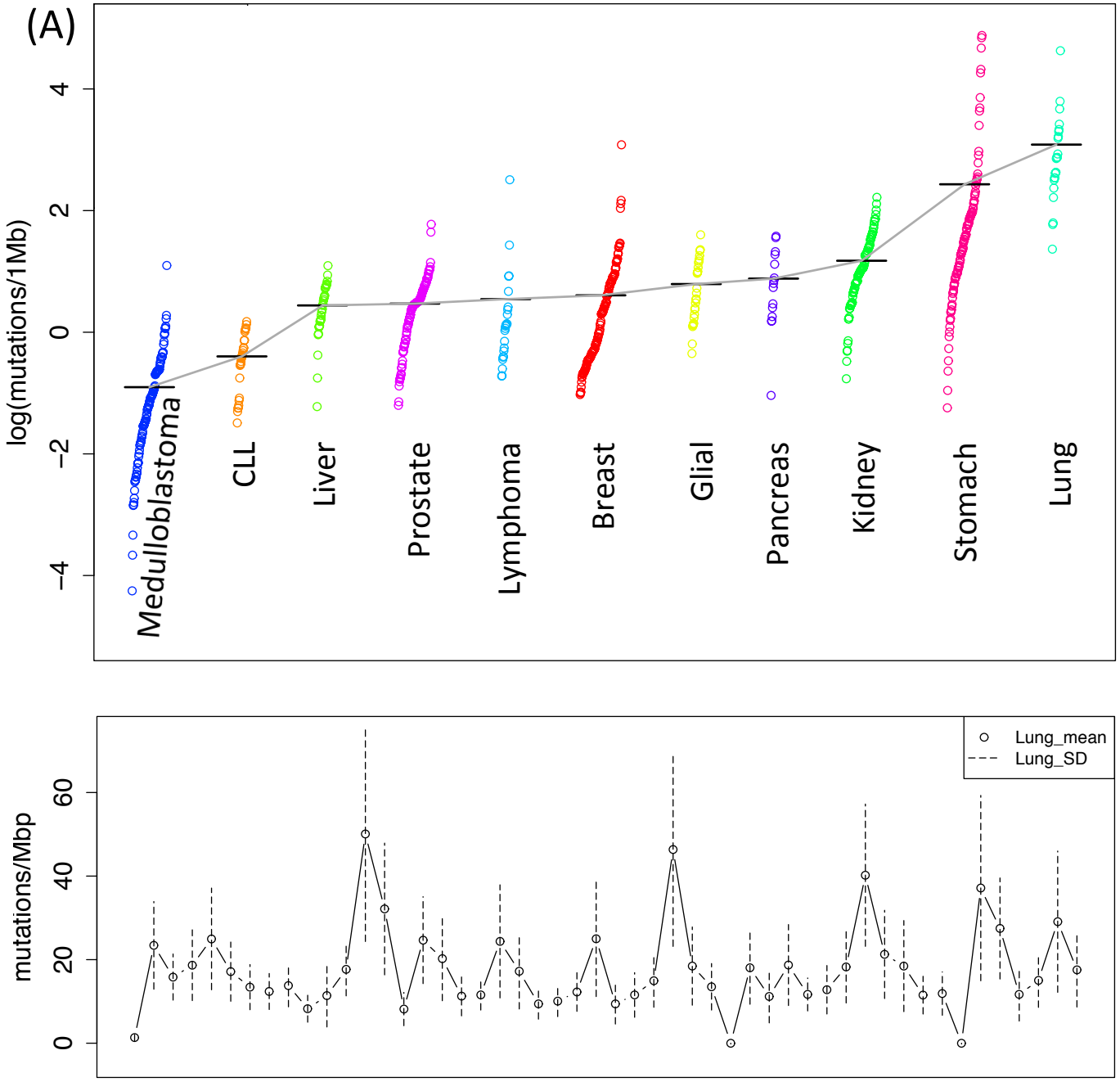
Mutation recurrence







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

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

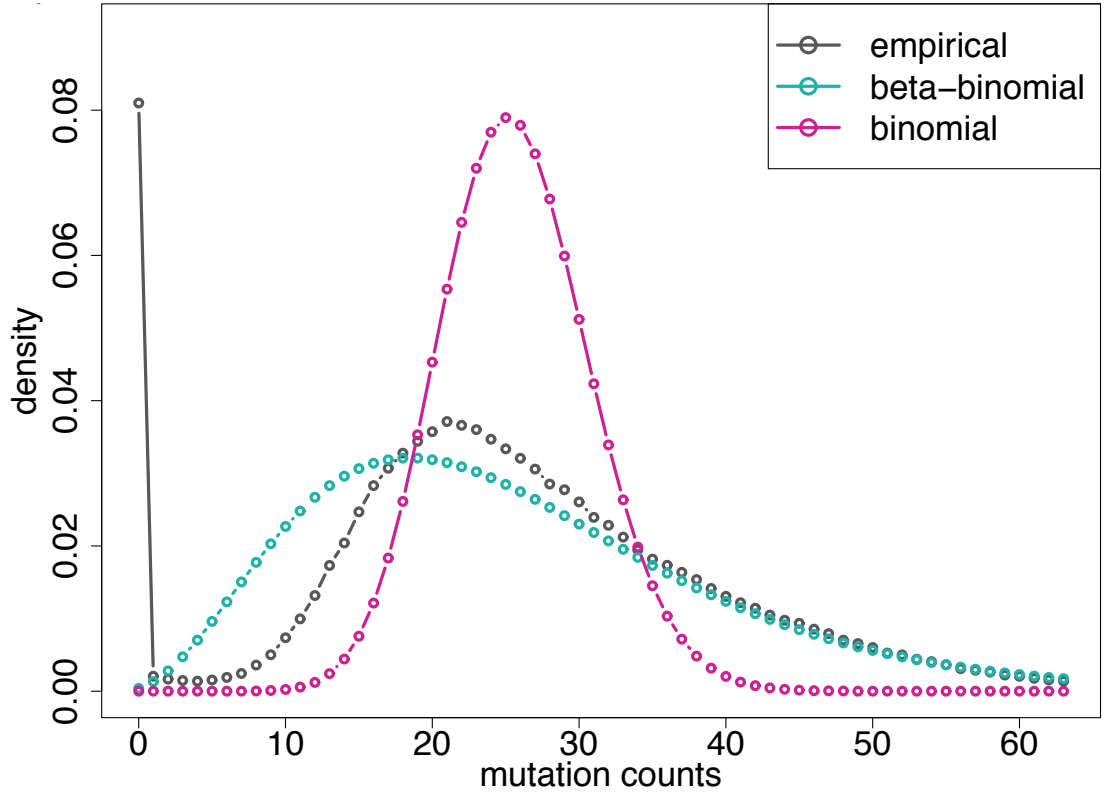
$$\mathbf{p_i} : Beta(\mu|\mathbf{R}, \sigma|\mathbf{R})$$

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

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

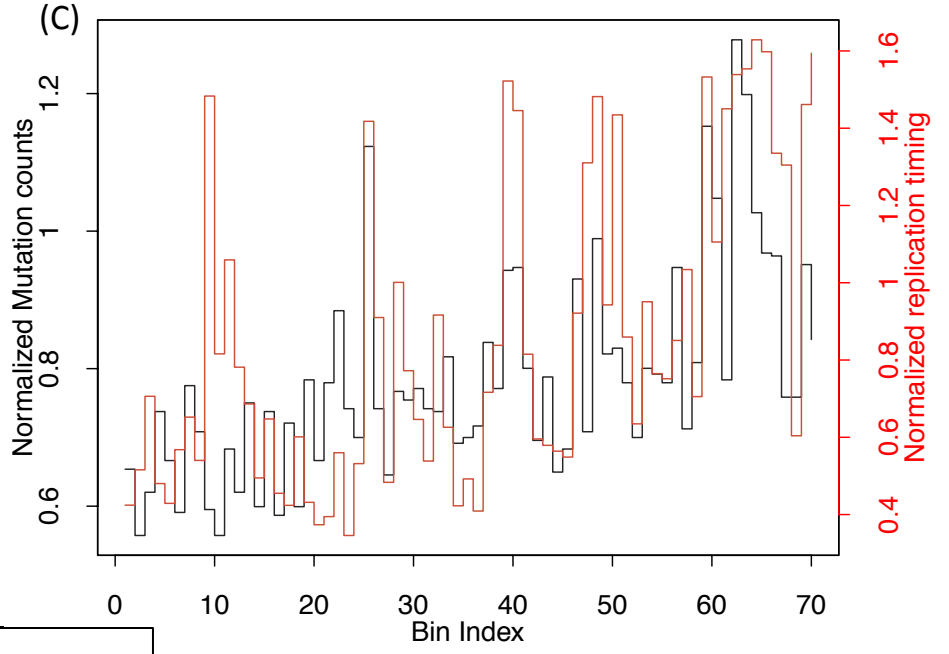
LARVA Model Comparison

- 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

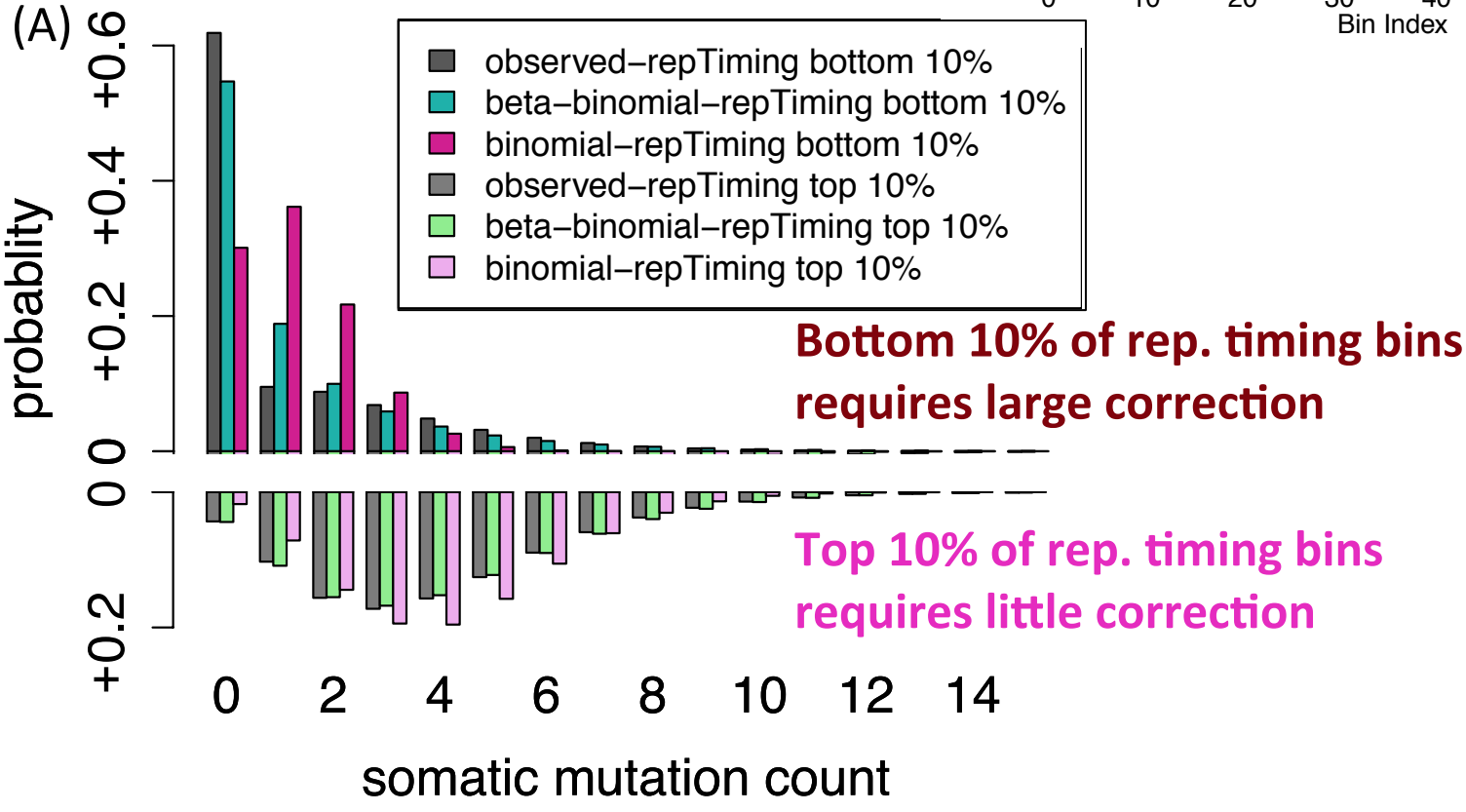


[Lochovsky et al. *NAR* ('15)]

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



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

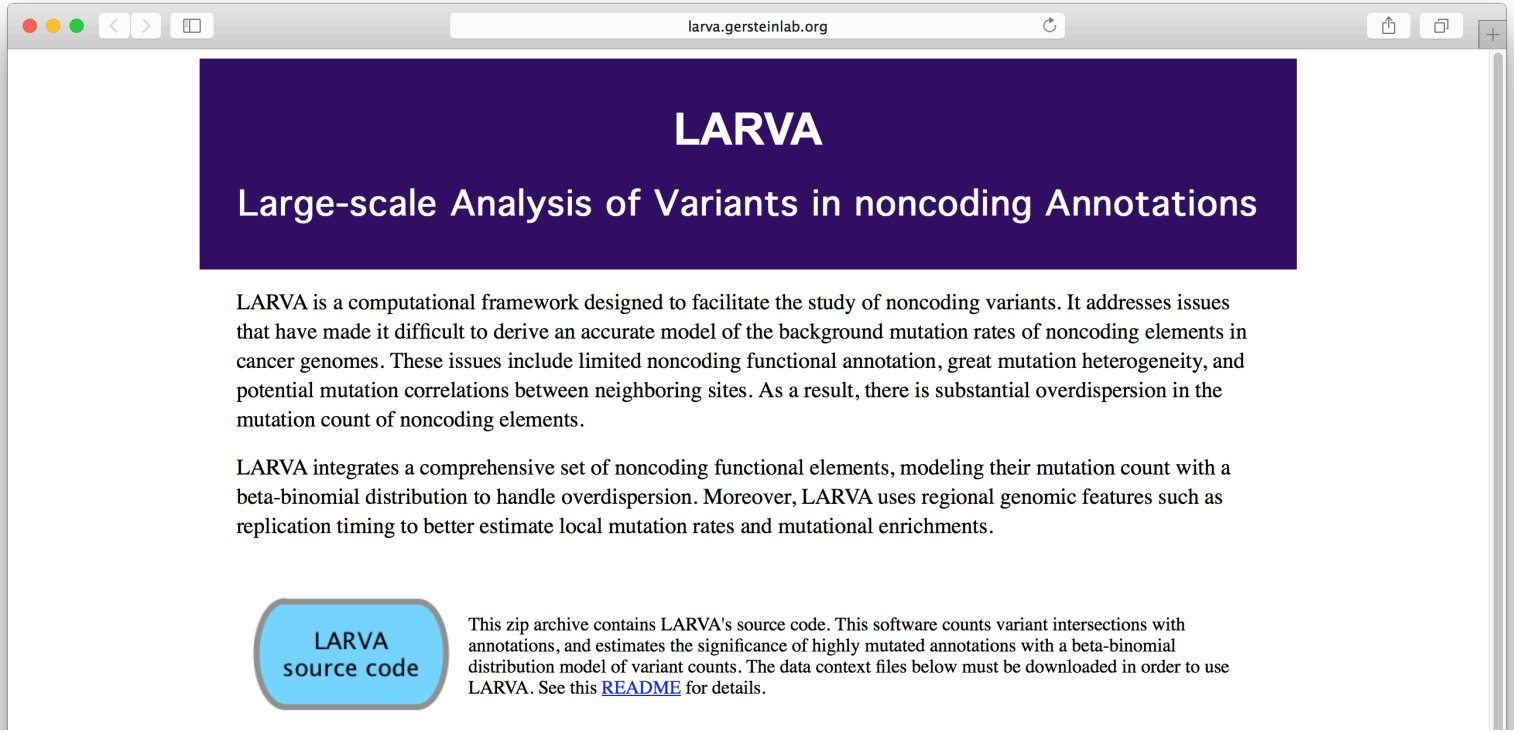


Bottom 10% of rep. timing bins requires large correction

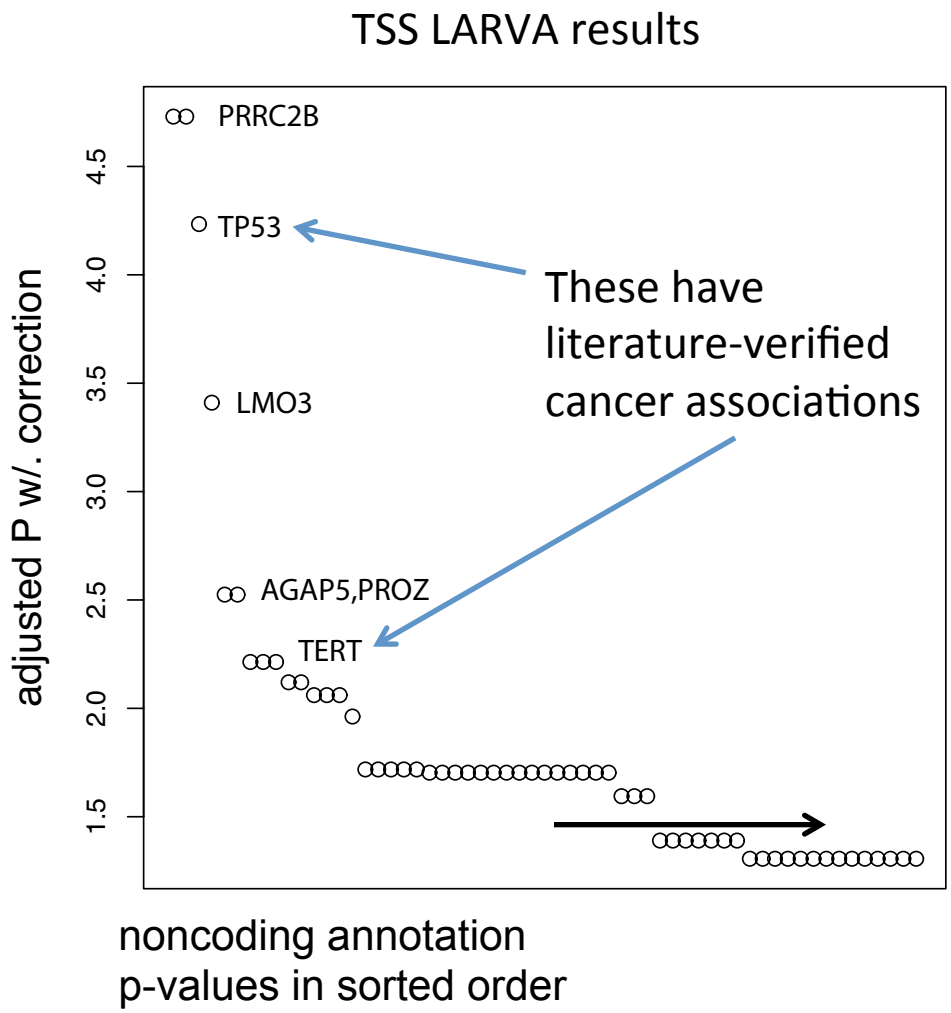
Top 10% of rep. timing bins requires little correction

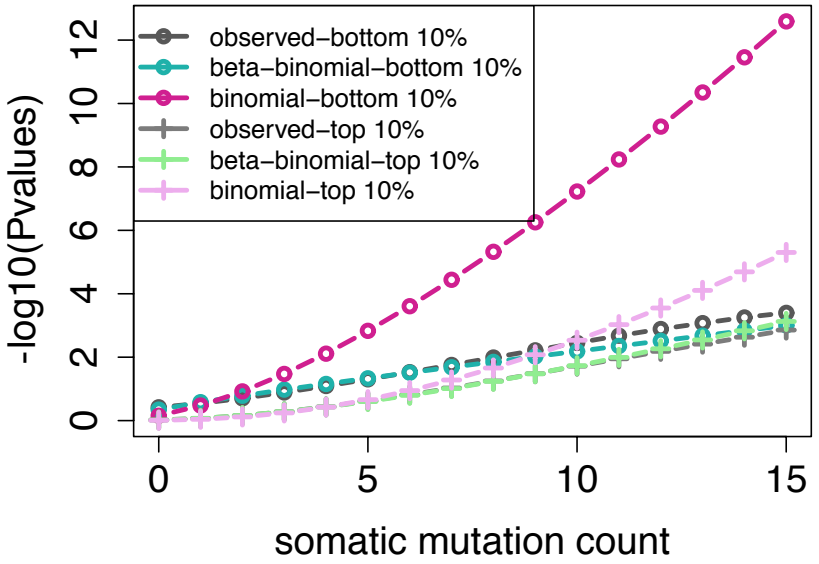
LARVA Implementation

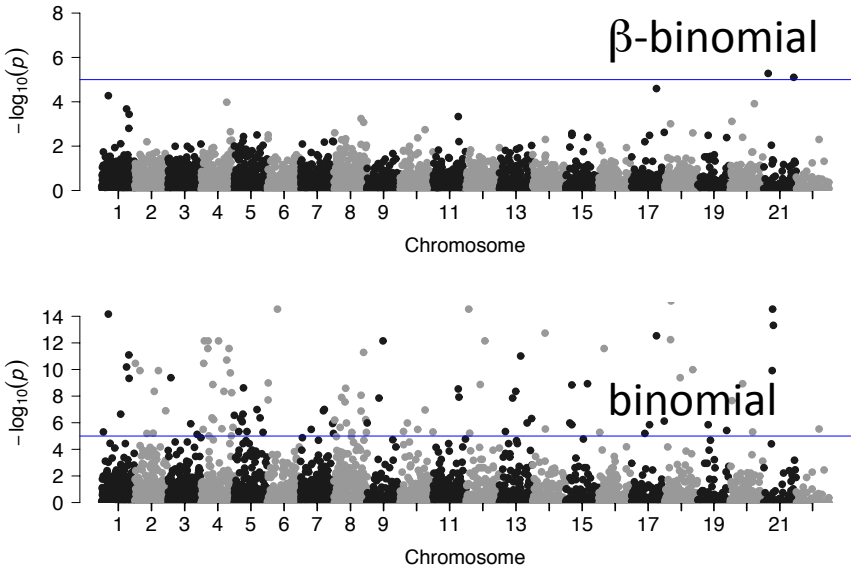
- http://larva.gersteinlab.org/
- Freely downloadable C++ program
  - Verified compilation and correct execution on Linux
- A Docker image is also available to download
  - Runs on any operating system supported by Docker
- Running time on transcription factor binding sites (a worst case input size)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



LARVA Results







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

Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

- Introduction: the landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
- Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)
- Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
  - Key technical Issue: Need to build personal genomes
    - Assessing their quality via read mapping
- Putting it together in workflows
  - Integrating evidence on non-coding variants with FunSeq
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - Using Larva to do burden testing on non-coding annotation
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

Personal Genomics: Prioritizing High-impact Rare & Somatic Variants

- **Introduction: the landscape of variants in personal genomes**
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
    - On surface & in interior bottlenecks
- **Non-coding Variants #1: Prioritizing rare variants with "sensitive sites" (human-conserved)**
- **Non-coding Variants #2: Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
  - **Key technical Issue: Need to build personal genomes**
    - Assessing their quality via read mapping
- **Putting it together in workflows**
  - **Integrating evidence on non-coding variants with FunSeq**
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritzing rare germline variants
  - **Using Larva to do burden testing on non-coding annotation**
    - Need to correct for over-dispersion in bionomial
    - Parameterized according to replication timing

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

CostSeq2
P **Muir**, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, F Isaacs, J Rozowsky

FunSeq.gersteinlab.org
- & -
FunSeq2.gersteinlab.org
Y **Fu**, E **Khurana**, Z Liu, S Lou, J Bedford, XJ Mu, KY Yip, V Colonna, XJ Mu, … ,
1000 Genomes Project Consortium, et al

LARVA.gersteinlab.org
L **Lochovsky**, J **Zhang**, Y Fu, E Khurana

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

Acknowledgments

Hiring Postdocs. See gersteinlab.org/jobs

Extra