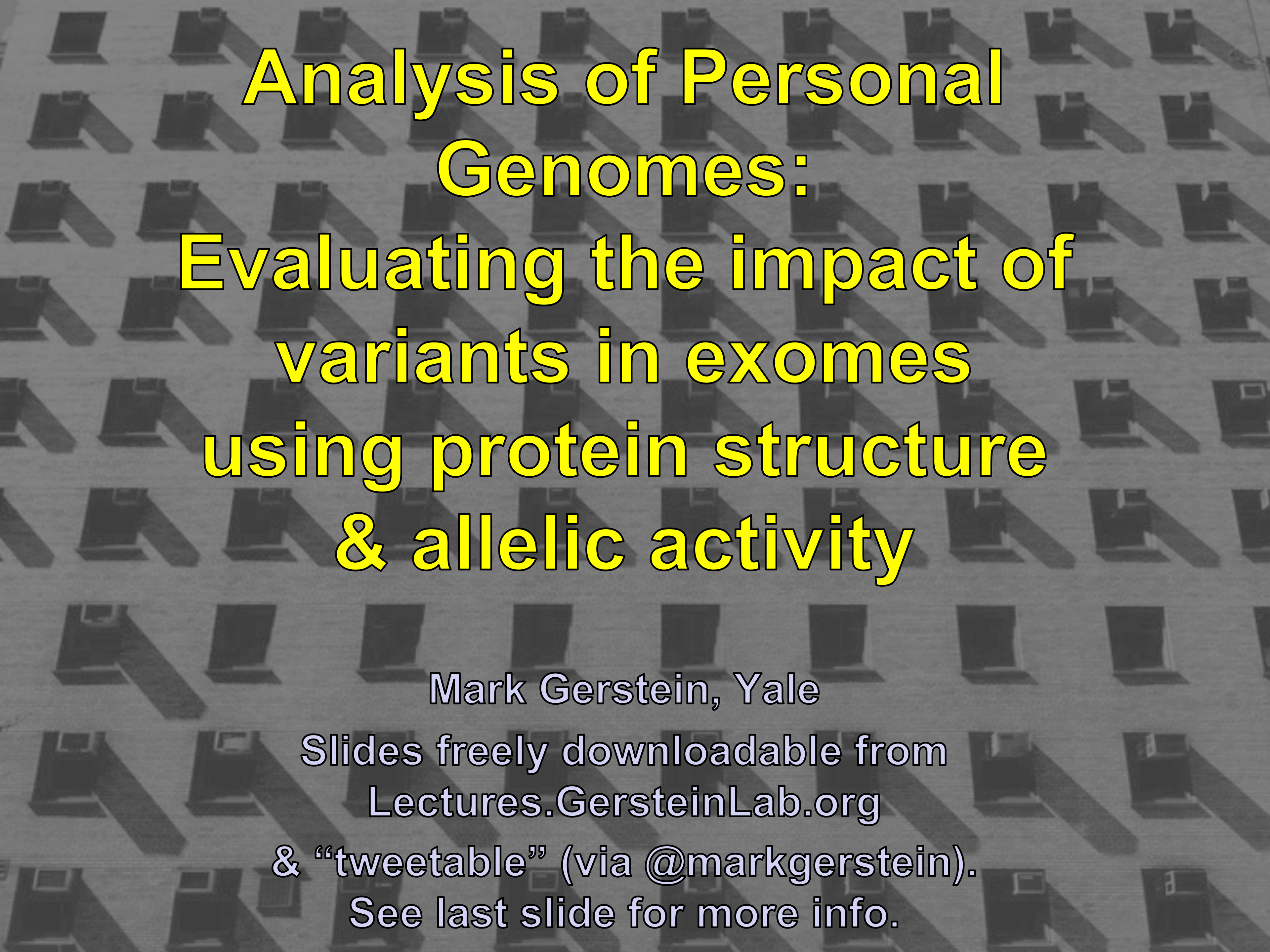


Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

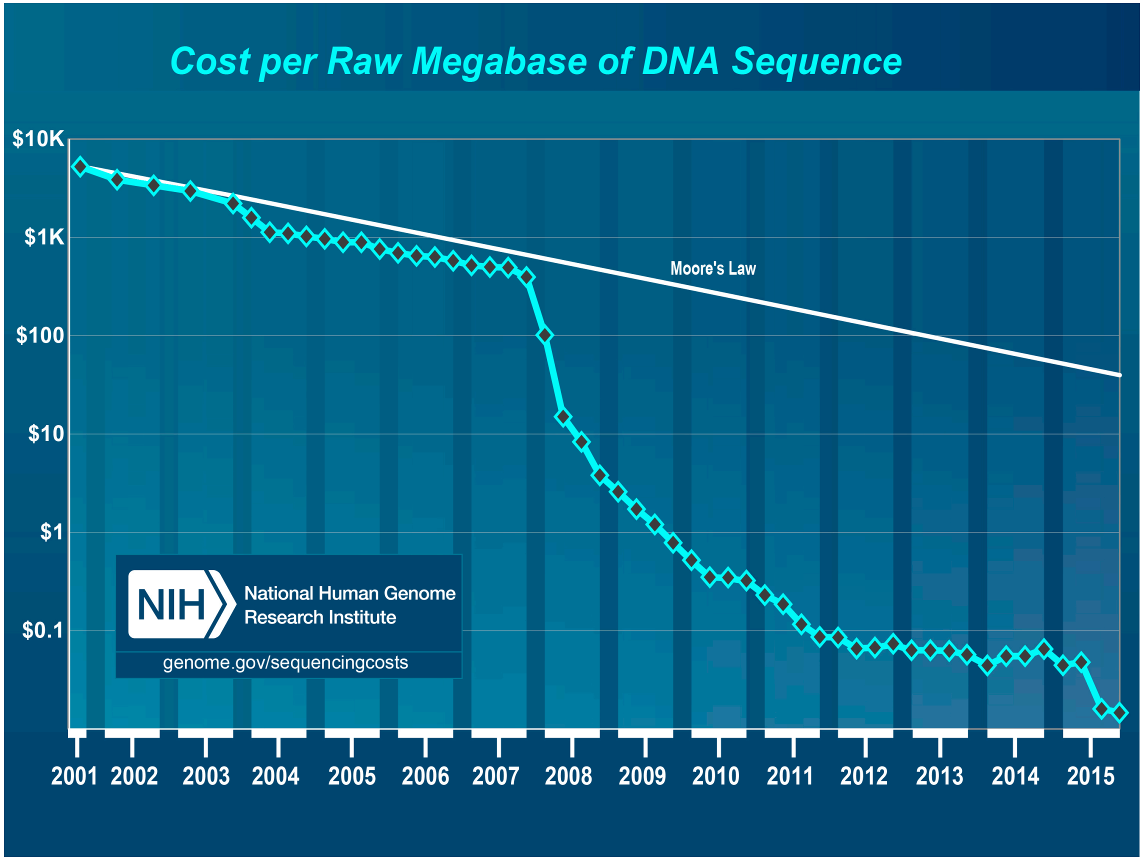
Mark Gerstein, Yale

Slides freely downloadable from Lectures.GersteinLab.org

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

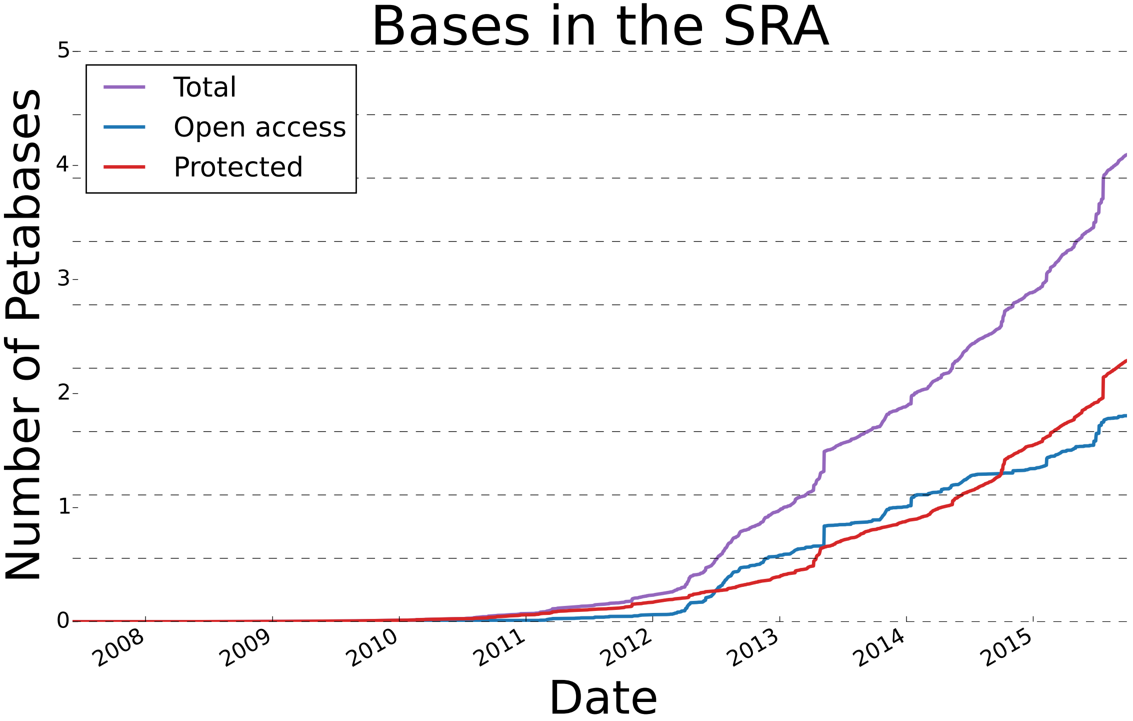
Sequencing Data Explosion: Faster than Moore's Law for a Time (or a S-curve)

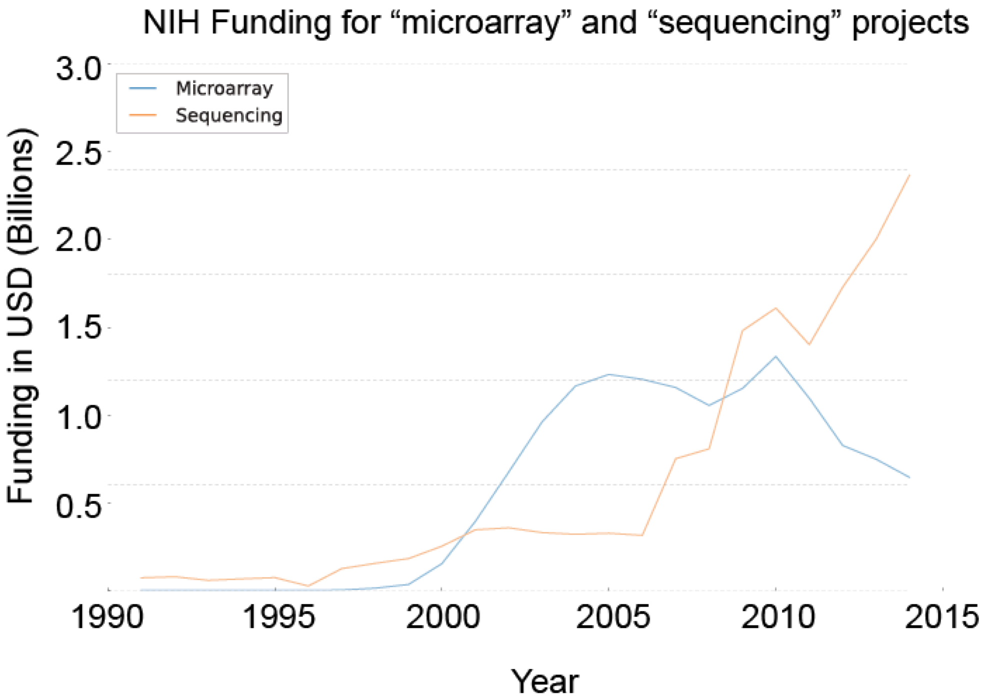
- DNA sequencing has gone through technological S-curves
  - In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
  - The advent of NGS was a shift to a new technology with dramatic decrease in cost).



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.





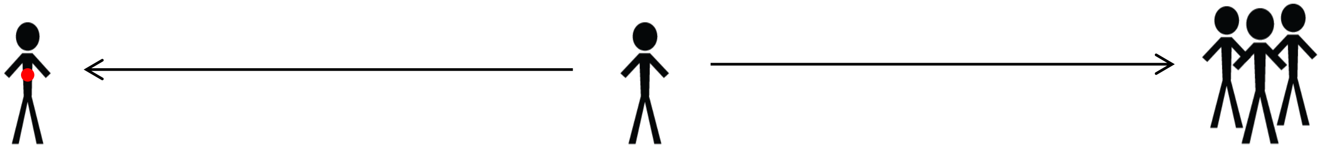
[Muir et al. (‘15) GenomeBiol.]

Human Genetic Variation

A Cancer Genome

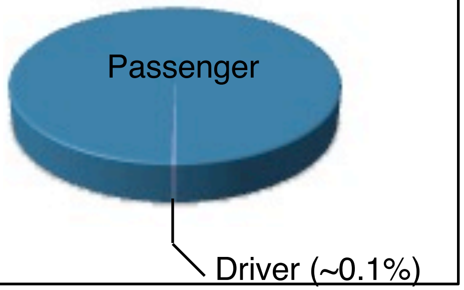
A Typical Genome

Population of
2,504 peoples



Origin of Variants

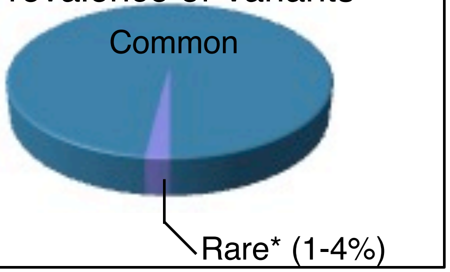
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



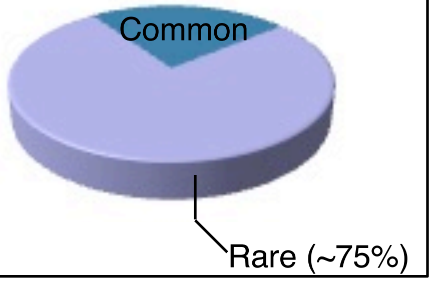
Class of Variants

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

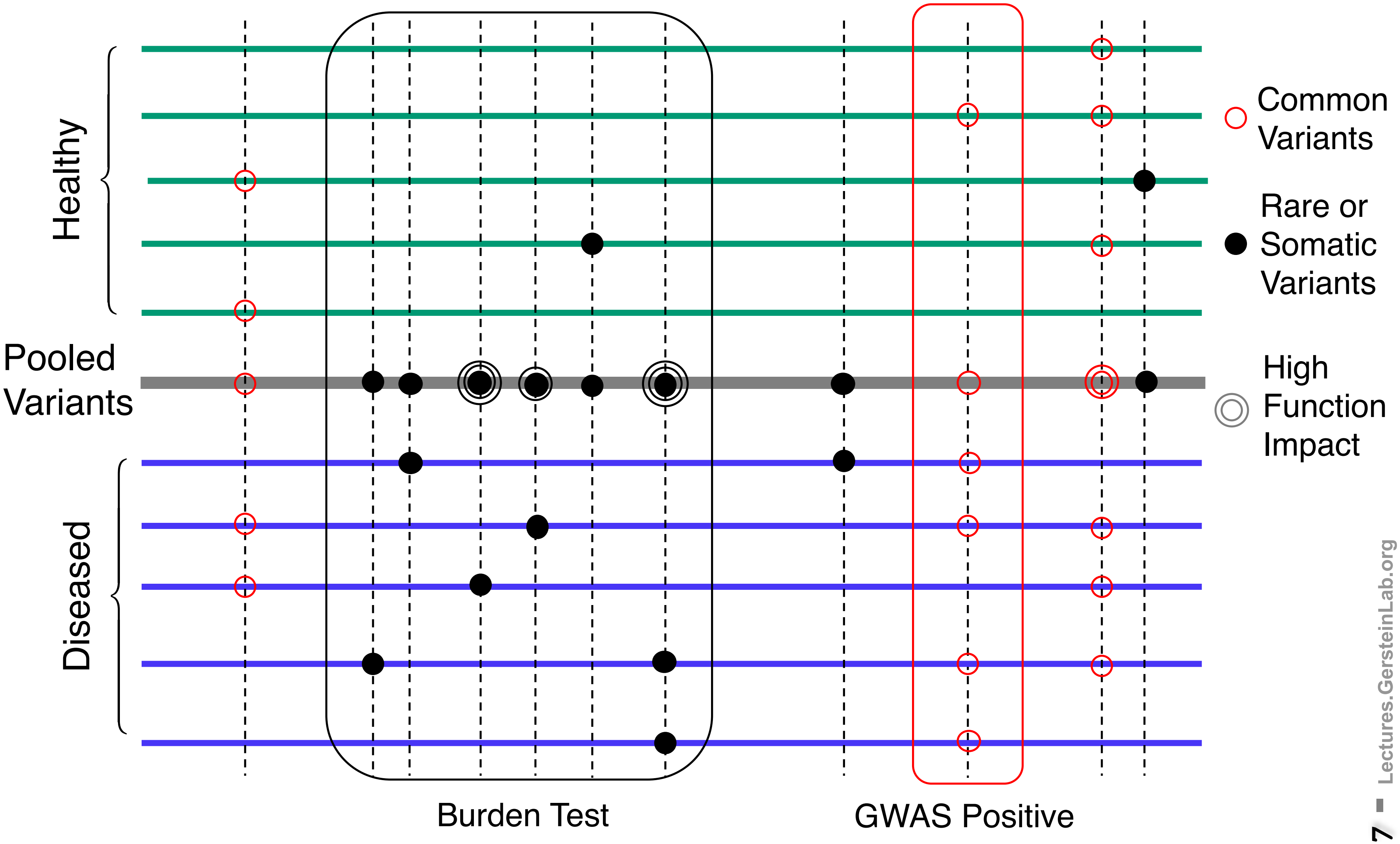
Somatic

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.

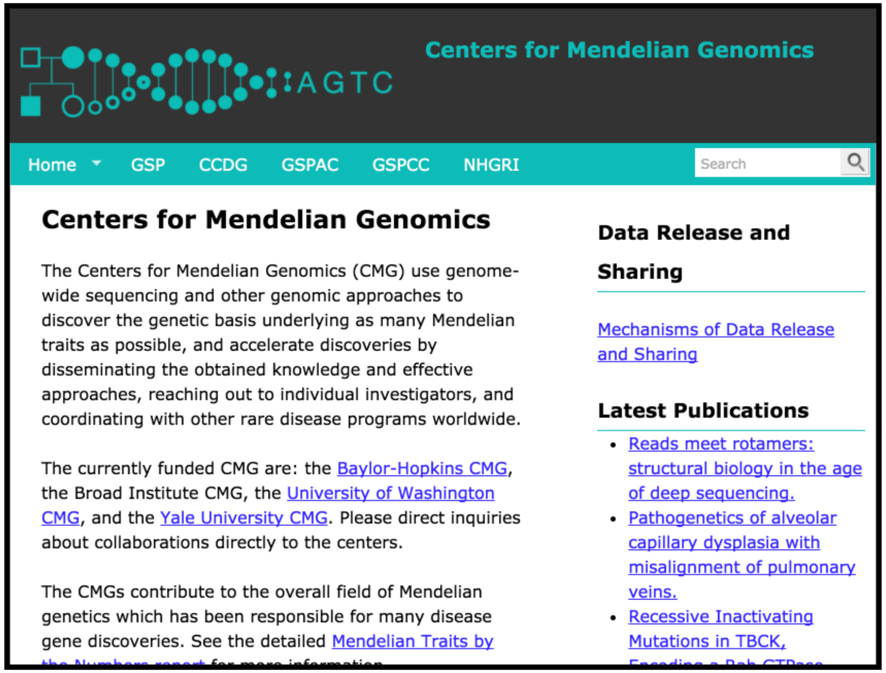
Vogelstein B. Science 2013. 339(6127):1546-1558

Association of Variants with Diseases



Rare variant analysis particularly applicable at the moment to Exomes

- CMG rare disease variants & TCGA somatic variants
  - Main NIH disease genomic project
  - Both of these focus on ”rare” variant for which GWAS is not meaningful
  - Larger numbers of individual exomes more important than WGS



Centers for Mendelian Genomics

The Centers for Mendelian Genomics (CMG) use genome-wide sequencing and other genomic approaches to discover the genetic basis underlying as many Mendelian traits as possible, and accelerate discoveries by disseminating the obtained knowledge and effective approaches, reaching out to individual investigators, and coordinating with other rare disease programs worldwide.

The currently funded CMG are: the Baylor-Hopkins CMG, the Broad Institute CMG, the University of Washington CMG, and the Yale University CMG. Please direct inquiries about collaborations directly to the centers.

The CMGs contribute to the overall field of Mendelian genetics which has been responsible for many disease gene discoveries. See the detailed Mendelian Traits by the Numbers report for more information.

Data Release and Sharing

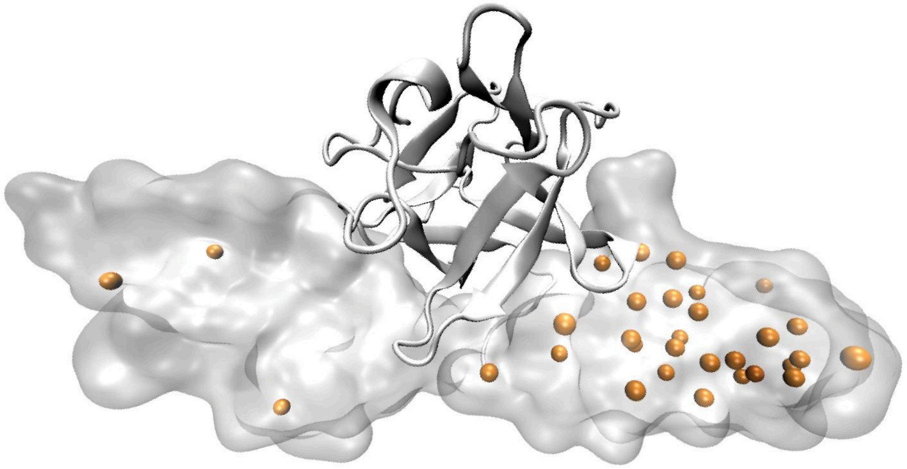
Mechanisms of Data Release and Sharing

Latest Publications

- Reads meet rotamers: structural biology in the age of deep sequencing.
- Pathogenetics of alveolar capillary dysplasia with misalignment of pulmonary veins.
- Recessive Inactivating Mutations in TBCK, Encoding a Rab GTPase

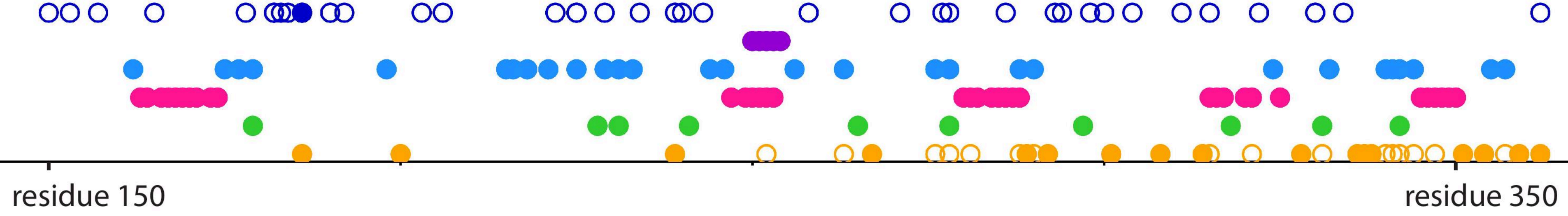
- Exomes have the current potential for great scale with the better impact interpretability of coding variants, often in a region of known protein structure
  - Scale of EXAC, >60K exomes [Lek et al. ‘16]

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



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

Fibroblast growth factor receptor 2 (pdb: 1IIL)



[Sethi et al. COSB ('15)]

Developing Tools for evaluating the impact of rare variants in coding regions

- New tools to wring everything out of protein structure
  - Stress for finding cryptic sites
  - Frustration for rapidly evaluating packing changes
  - (MotifVar) Intensification for using the amplifying power of protein structural motifs (eg TPR)
- Another approach – looking for allelic variants

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

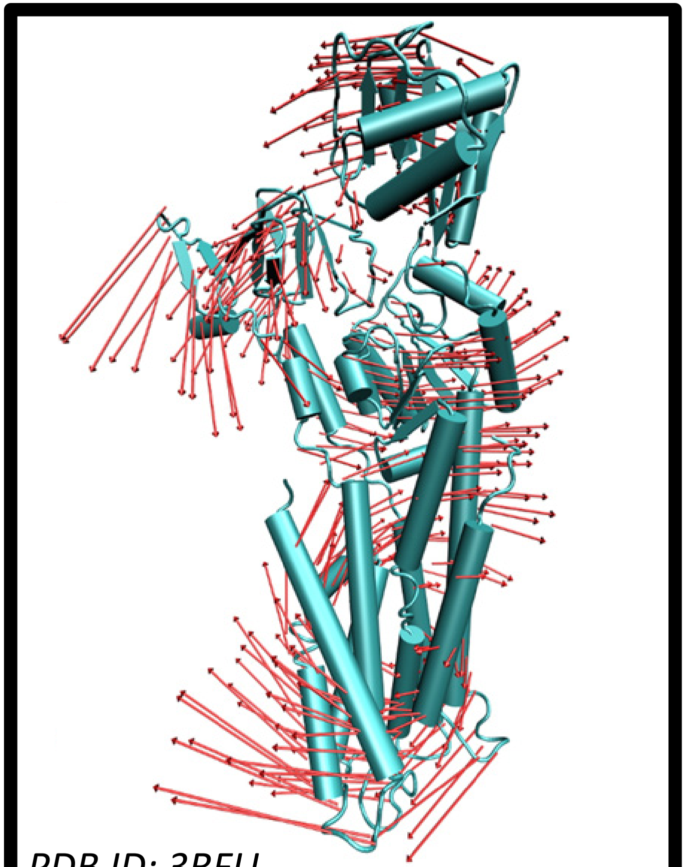
- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

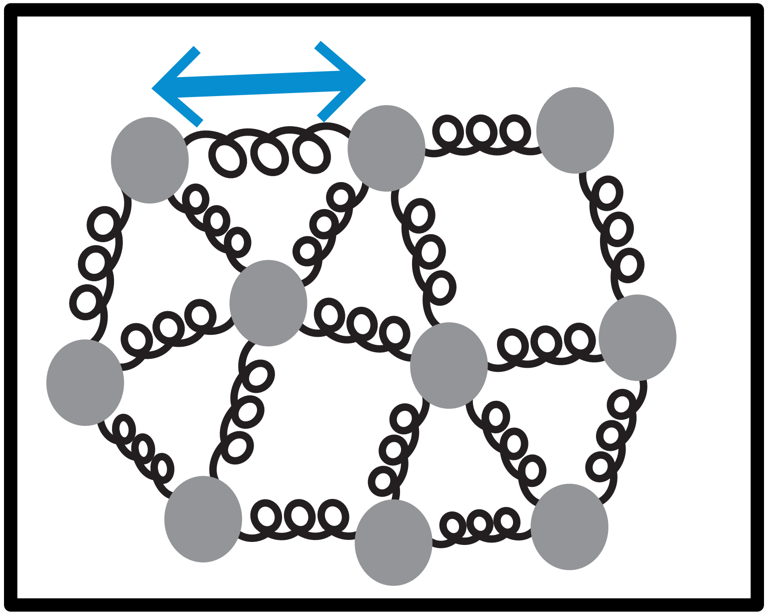
Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



PDB ID: 3RFU

Adapted from Fuglebakk et al, 2014



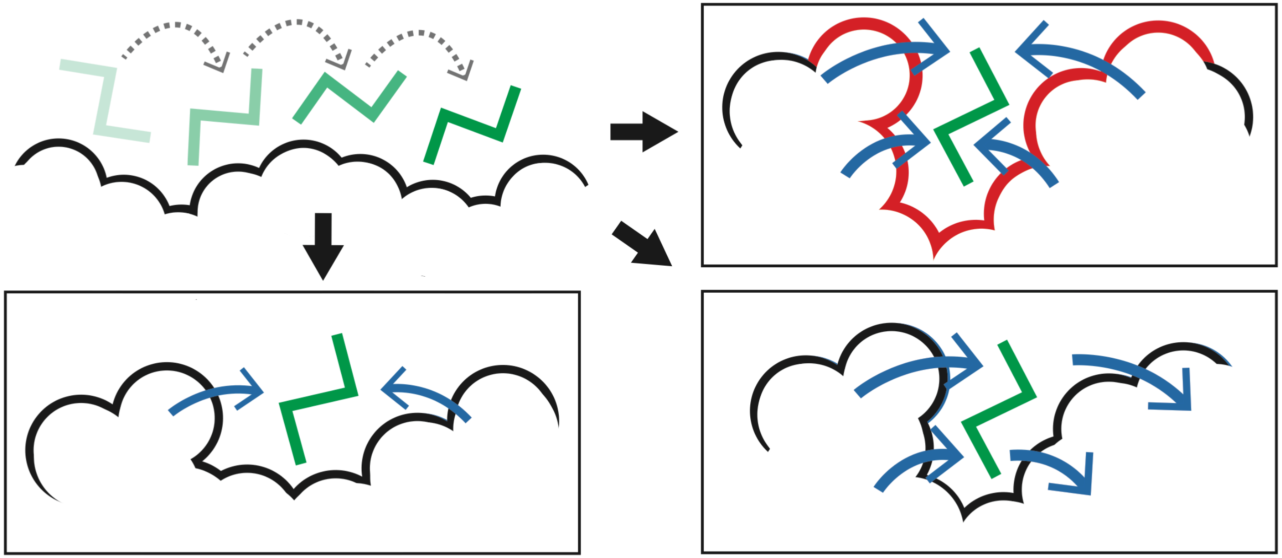
Characterizing uncharacterized variants

<= Finding Allosteric sites

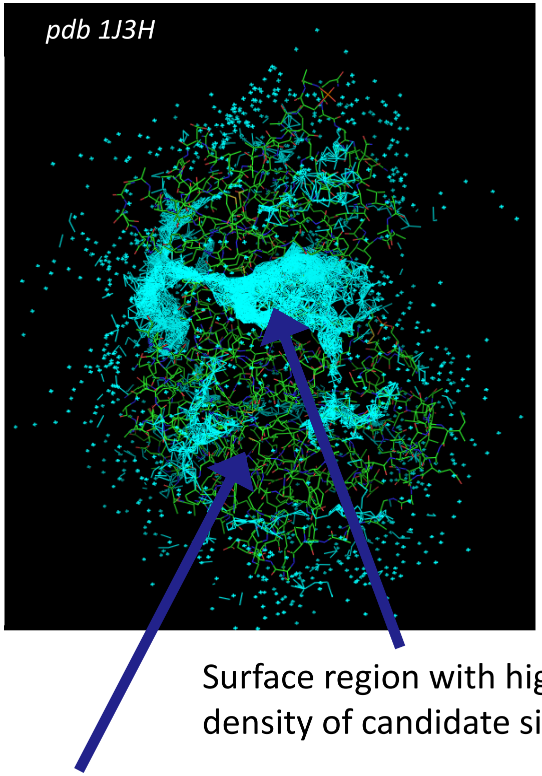
<= Modeling motion

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



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

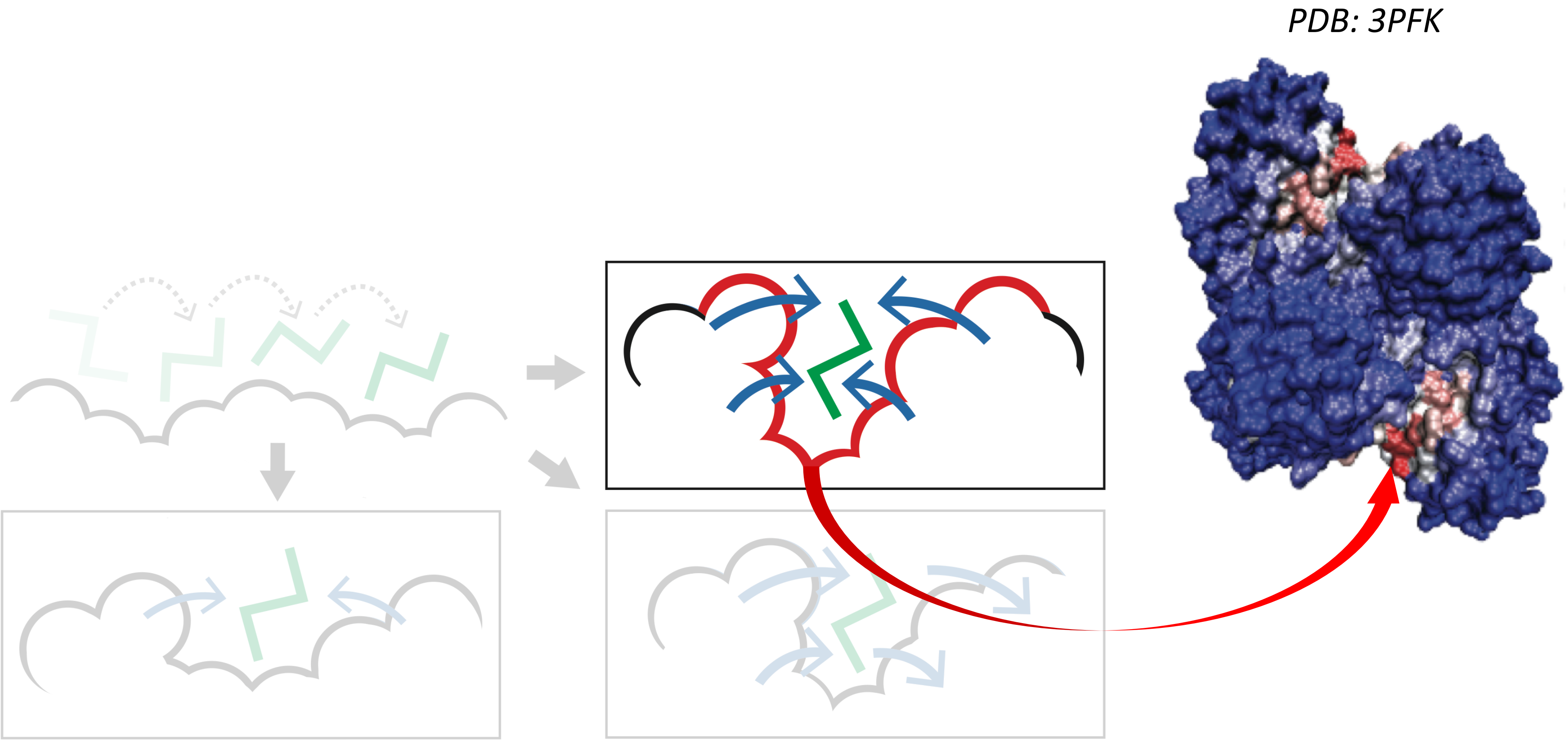


Surface region with high density of candidate sites

Surface region with low density of candidate sites

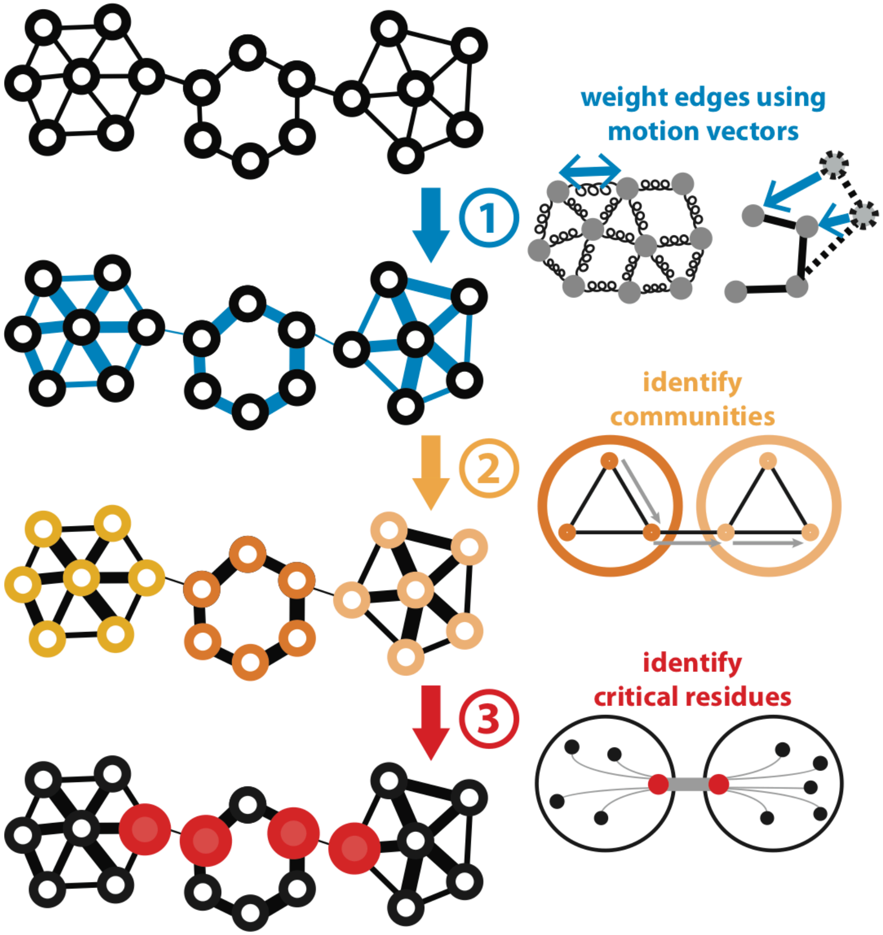
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface



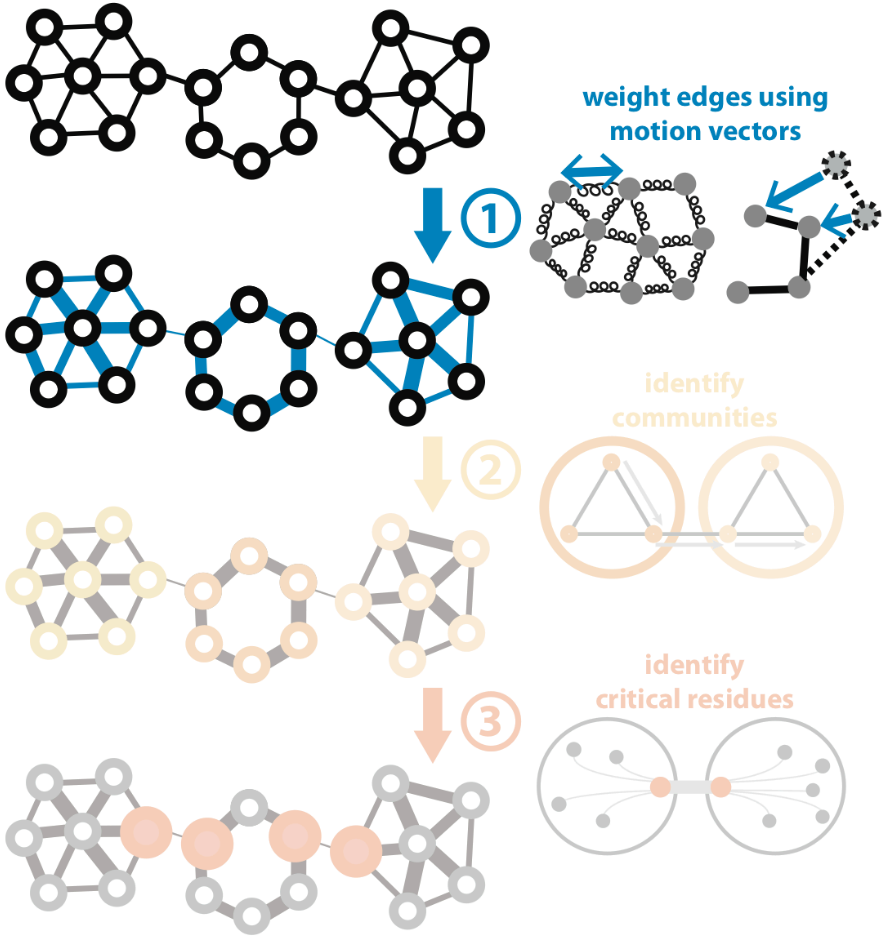
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



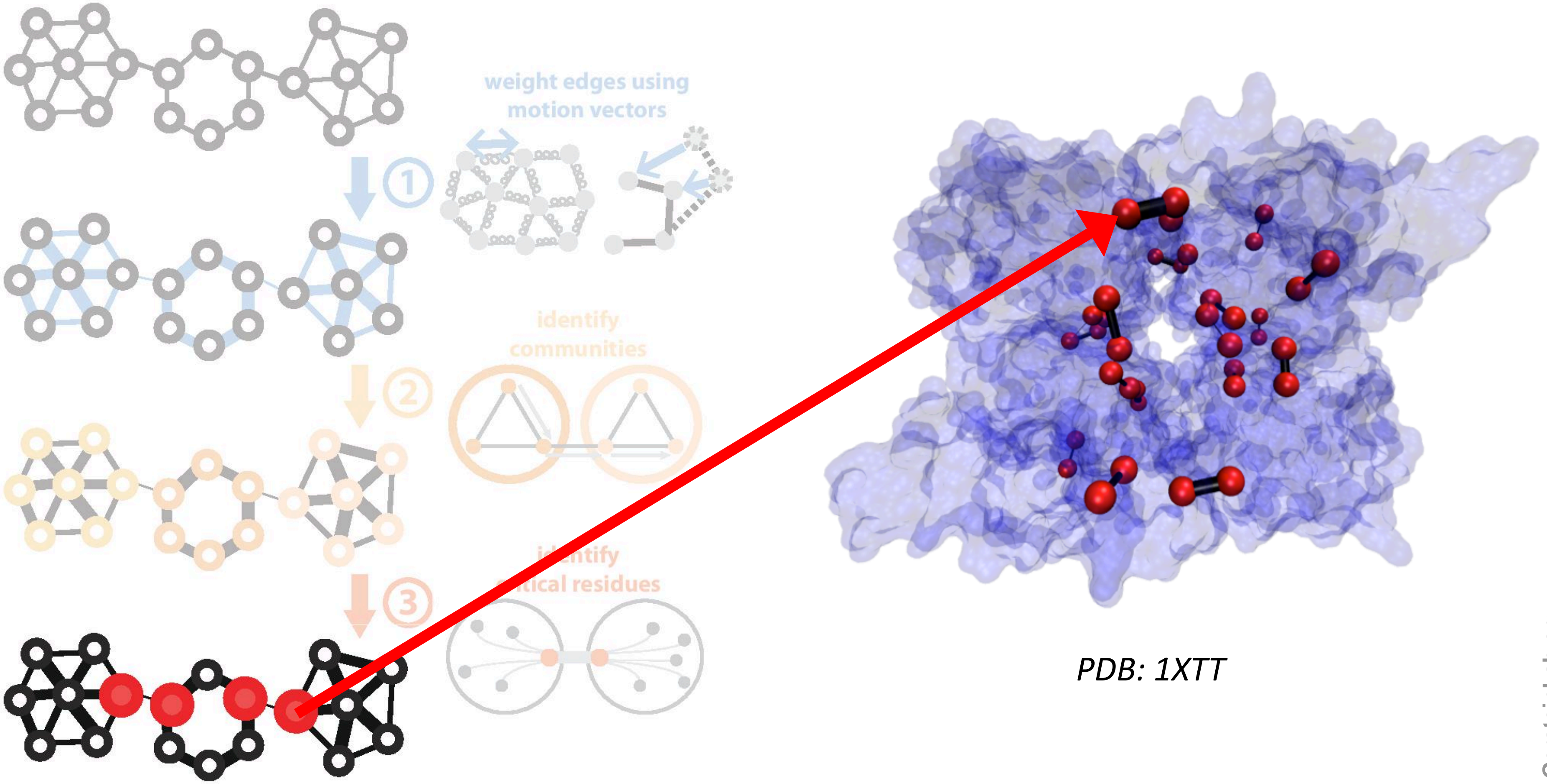
$$Cov_{ij} = \langle \mathbf{r}_{\mathrm{i}} \bullet \mathbf{r}_{\mathrm{j}} \rangle$$

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

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

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

Predicting Allosterically-Important Residues within the Interior

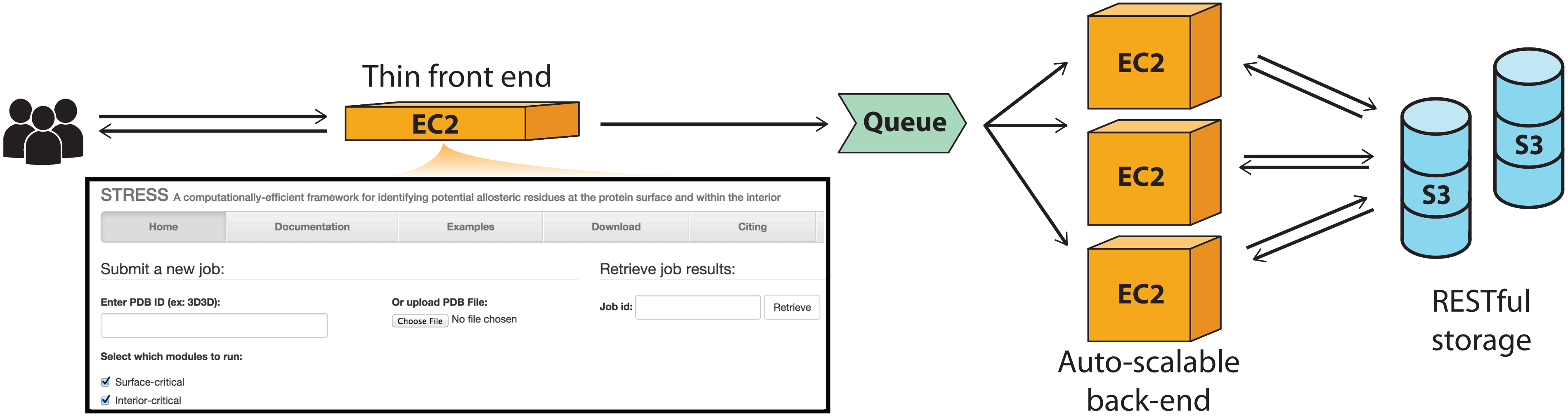


PDB: 1XTT

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: No file chosen

Retrieve job results:

Job id:

Select which modules to run:

☑ Surface-critical

☑ Interior-critical

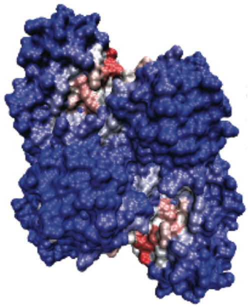
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

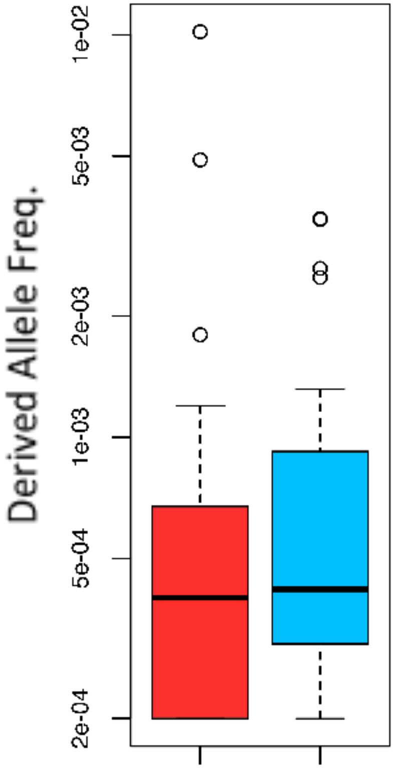
Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

1000 Genomes

Surface



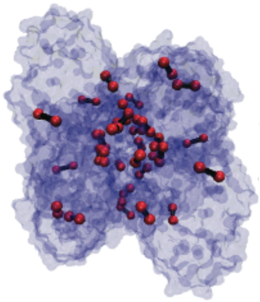


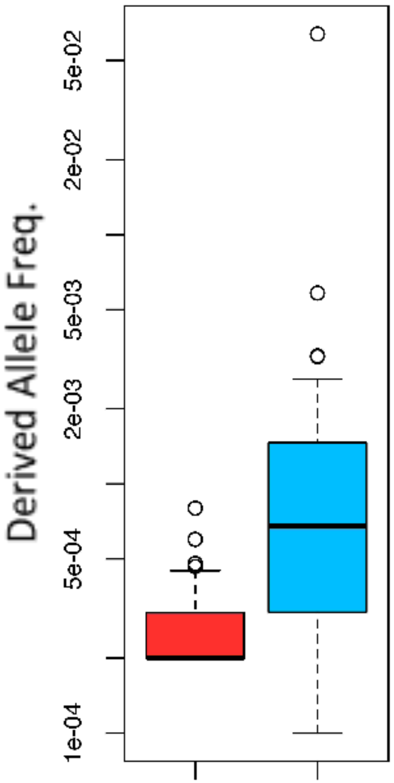
p=0.309

🟥 critical

🟦 non-critical

Interior





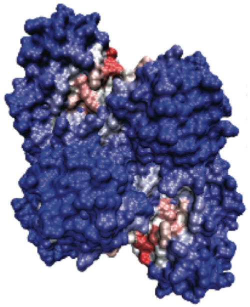
p=1.80e-05

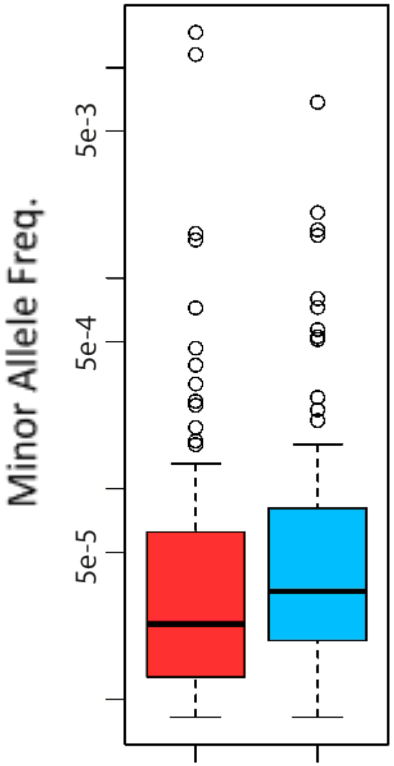
Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

ExAC

Surface



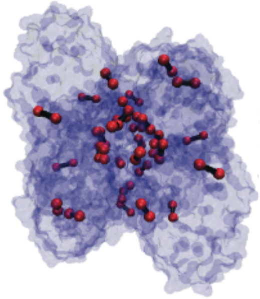


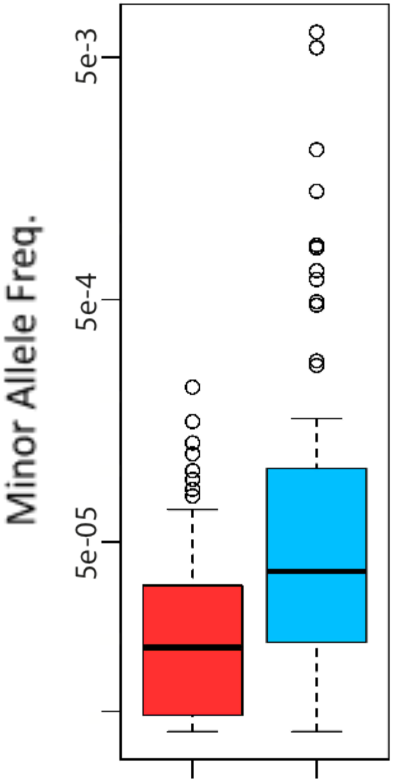
p=1.49e-3

🟥 critical

🟦 non-critical

Interior



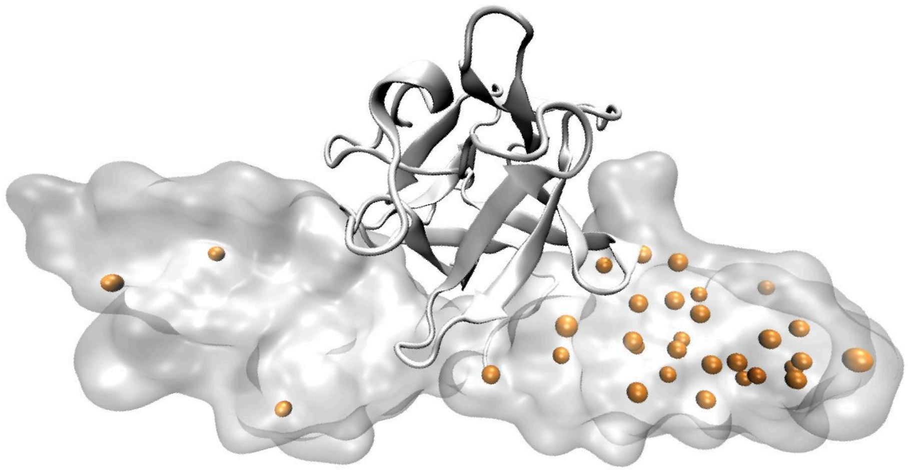


p=7.98e-09

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

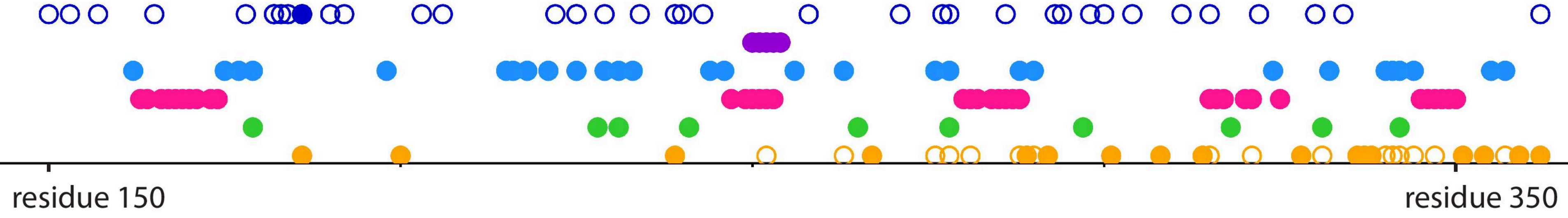
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

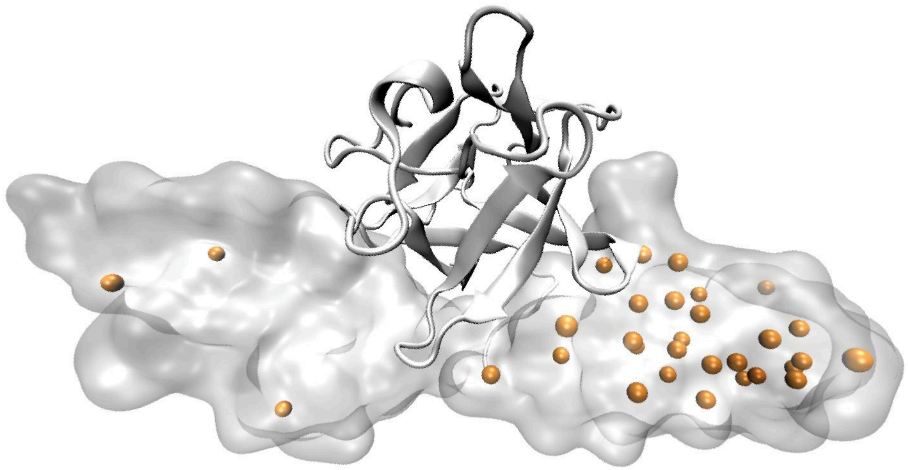
- ● ○ 1000G & ExAC SNVs (common | rare)
- ● Hinge residues
- ● Buried residues
- ● Protein-protein interaction site
- ● Post-translational modifications
- ● HGMD site (w/o annotation overlap)
- ○ HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

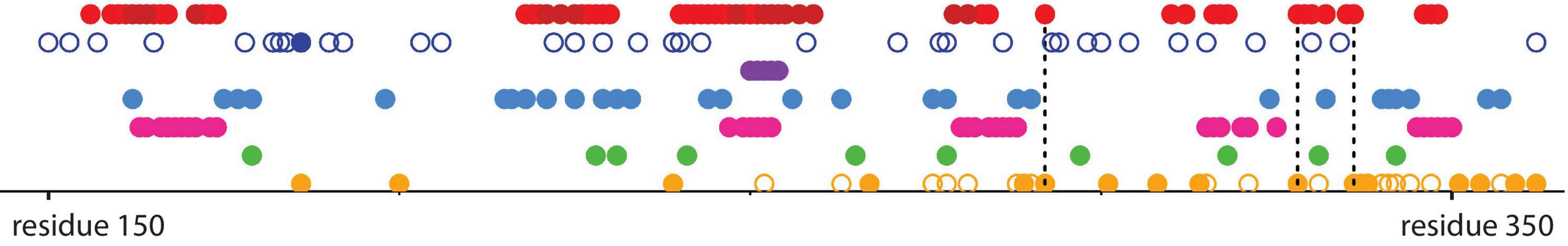
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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)

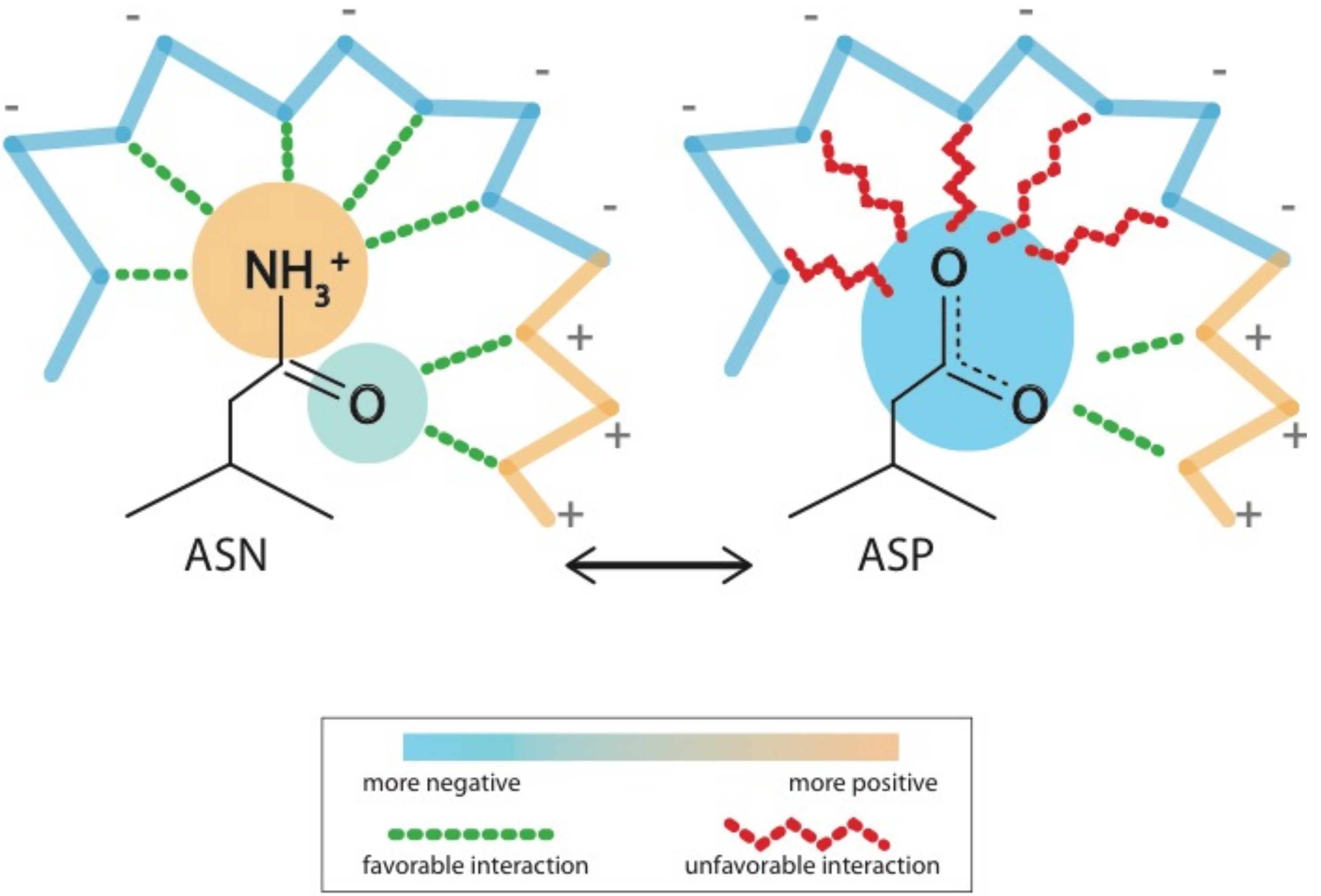


[Sethi et al. COSB ('15)]

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

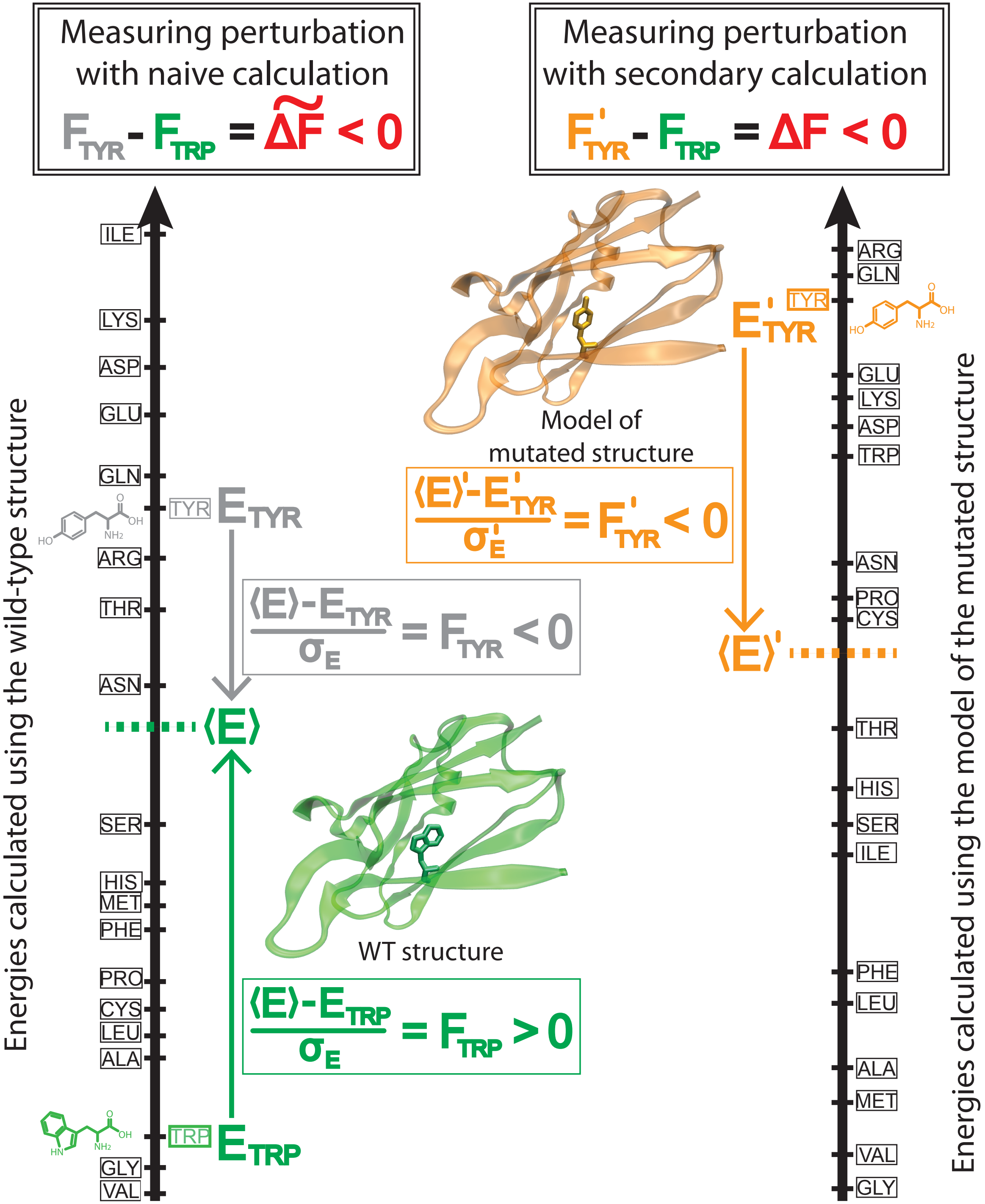
- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Schematic illustration of localized frustration



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

Workflow for evaluating localized frustration changes (ΔF)



Measuring perturbation with naive calculation

$F_{TYR} - F_{TRP} = \widetilde{\Delta F} < 0$

Measuring perturbation with secondary calculation

$F'_{TYR} - F_{TRP} = \Delta F < 0$

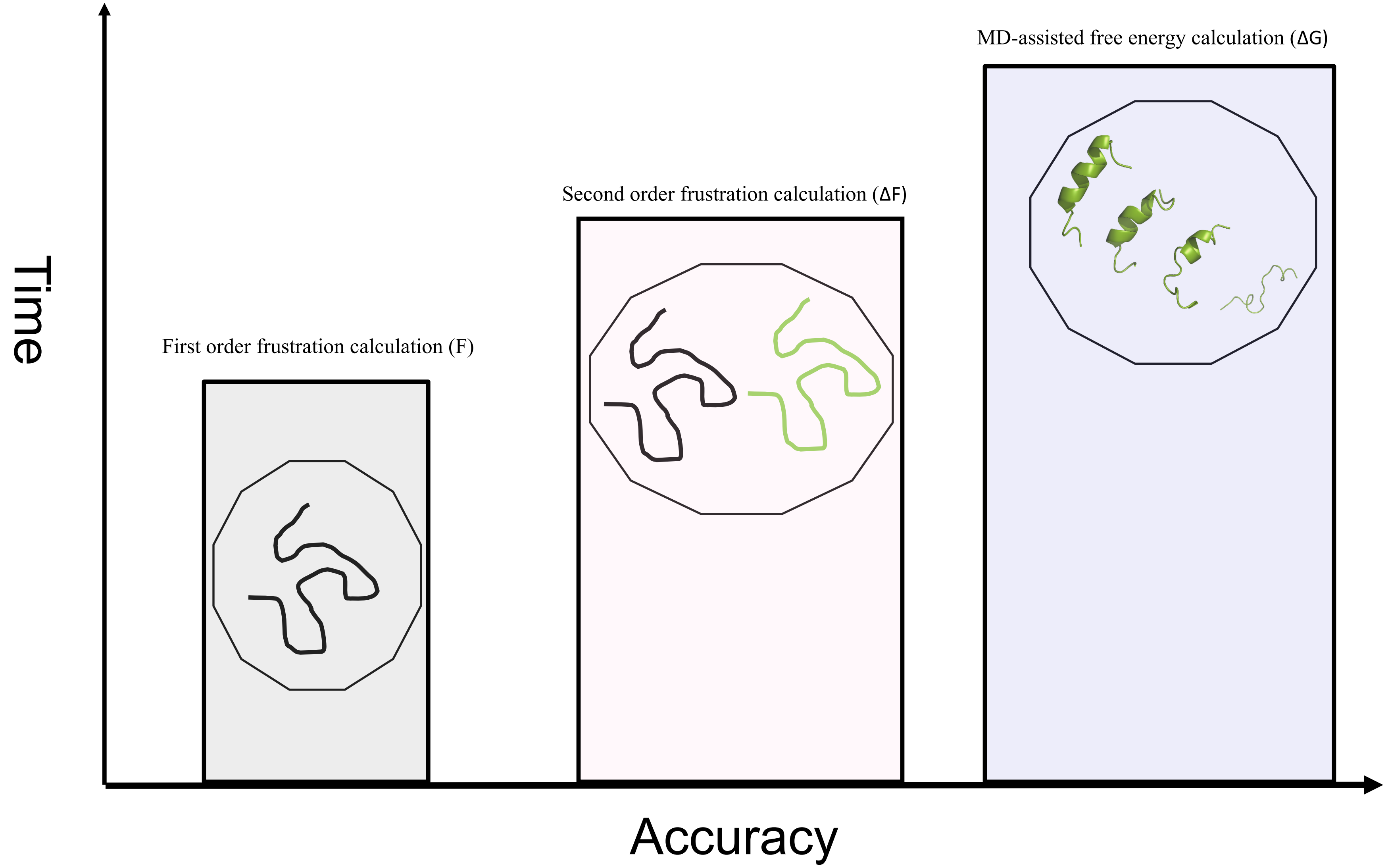
$$\frac{\langle E\rangle' - E'_{TYR}}{\sigma'_E} = F'_{TYR} < 0$$

$$\frac{\langle E\rangle - E_{TYR}}{\sigma_E} = F_{TYR} < 0$$

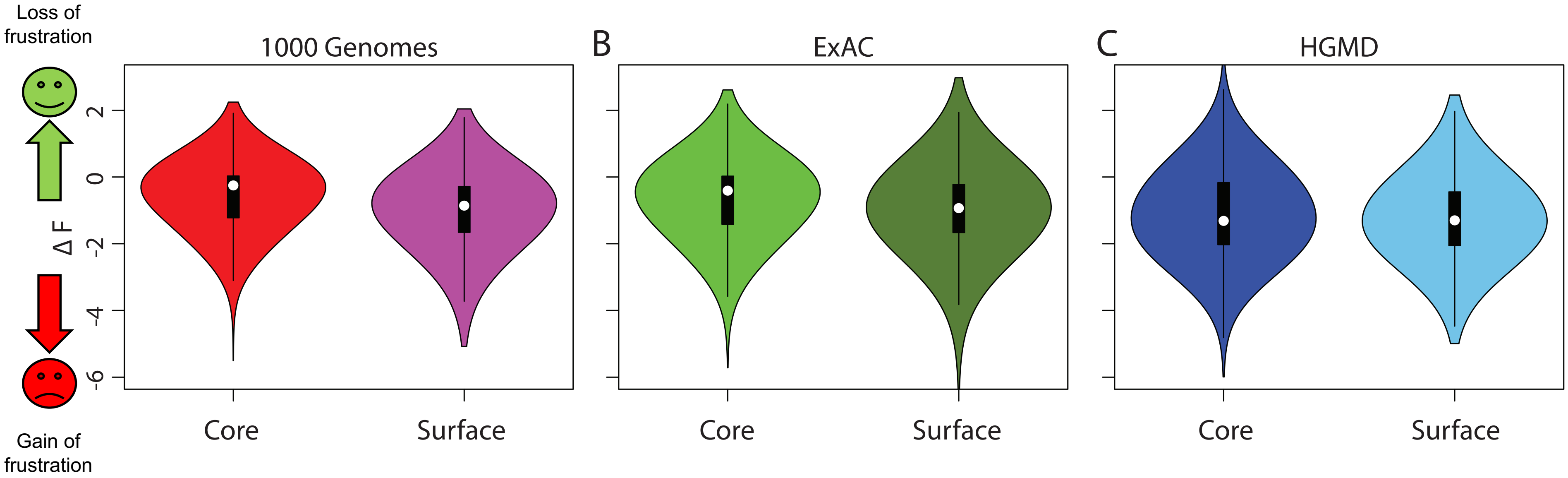
$$\frac{\langle E\rangle - E_{TRP}}{\sigma_E} = F_{TRP} > 0$$

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

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


Time
First order frustration calculation (F)
Second order frustration calculation (ΔF)
MD-assisted free energy calculation (ΔG)
Accuracy

Comparing ΔF values across different SNV categories: Normal v disease

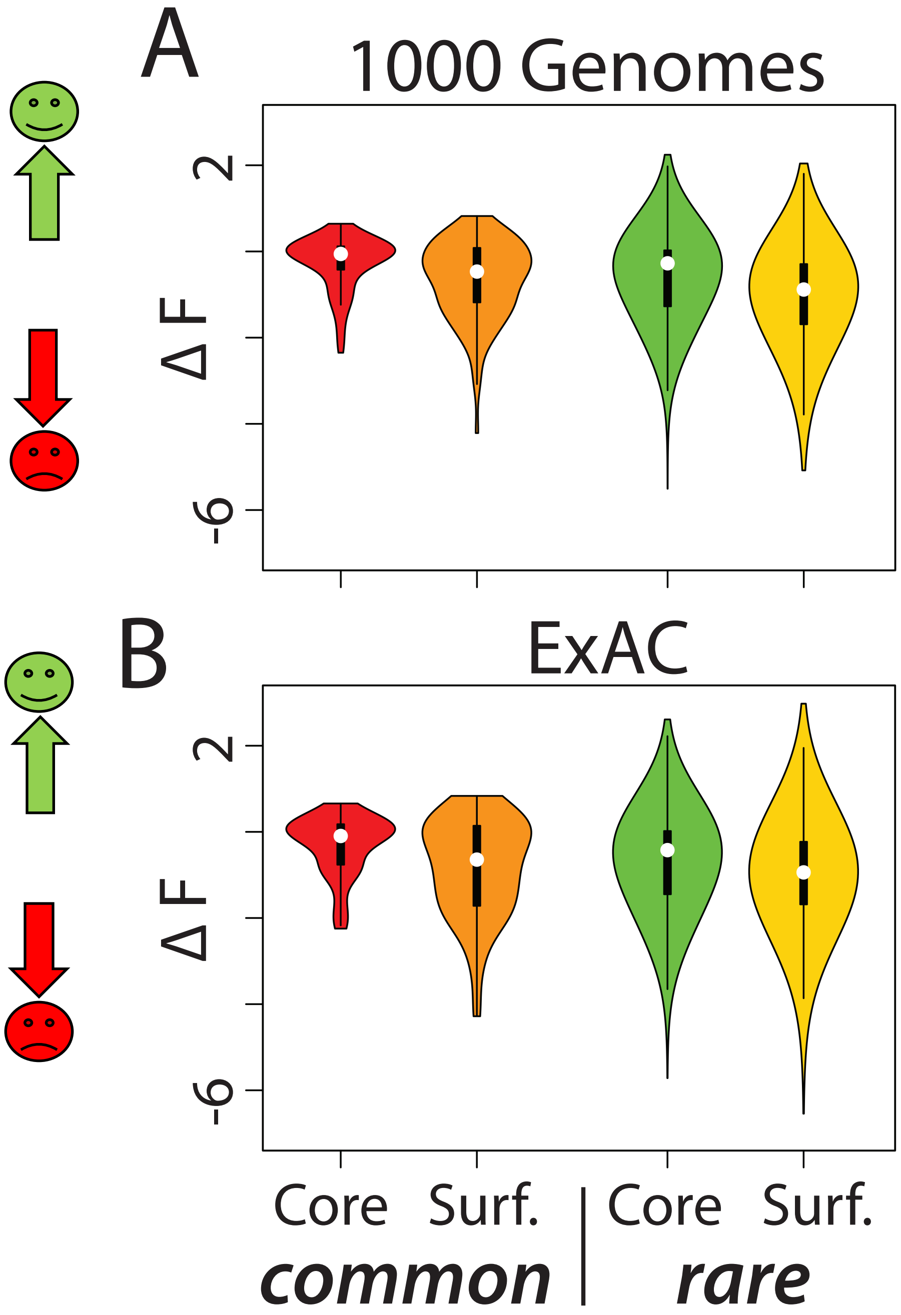


Normal mutations (1000G) tend to unfavorably frustrate (less frustrated) surface more than core,
but for disease mutations (HGMD)
no trend & greater changes

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

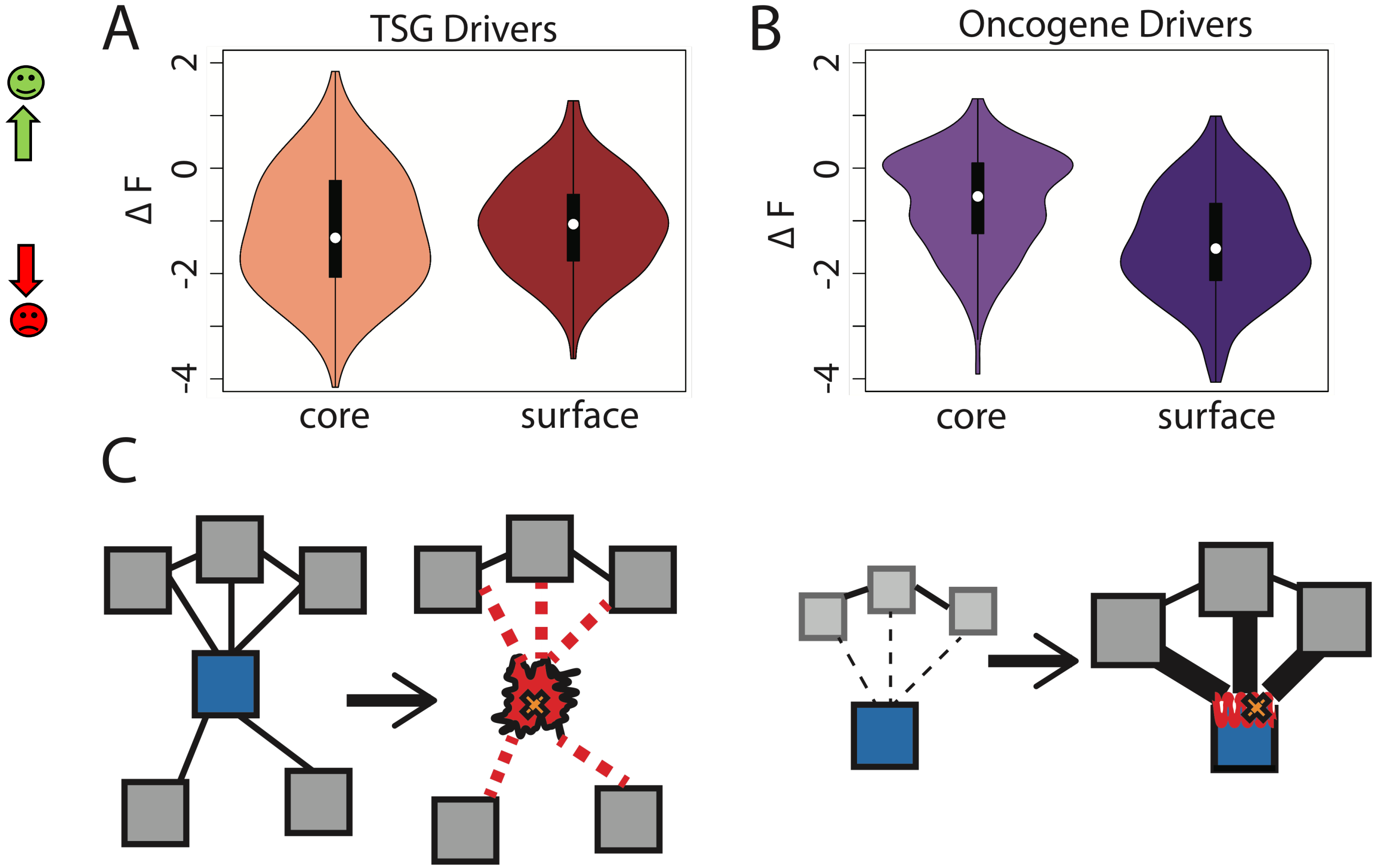
ΔF distributions among rare v. common SNVs

Rare mutations cause more unfavorable frustration change than common ones



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

Comparison between ΔF distributions: TSGs v. oncogenes



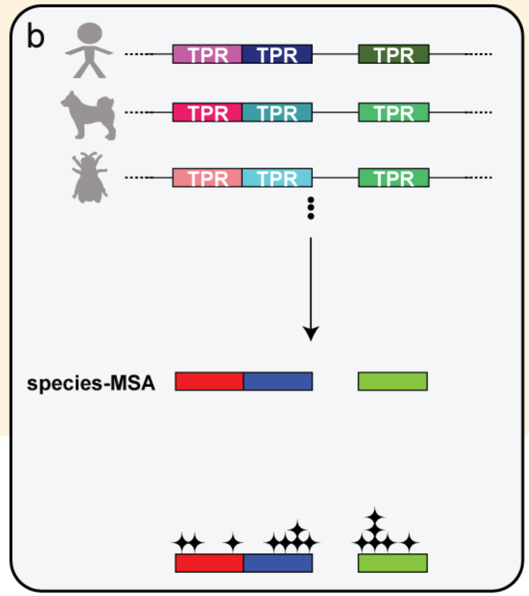
SNVs in TSGs change frustration more in core than the surface, whereas those associated with oncogenes manifest the opposite pattern. This is consistent with differences in LOF v GOF mechanisms.

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

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Intensification amplifies signals from motif-based MSAs



[Chen et al., JMB ('17, in revision)]

Intensification amplifies signals from motif-based MSAs

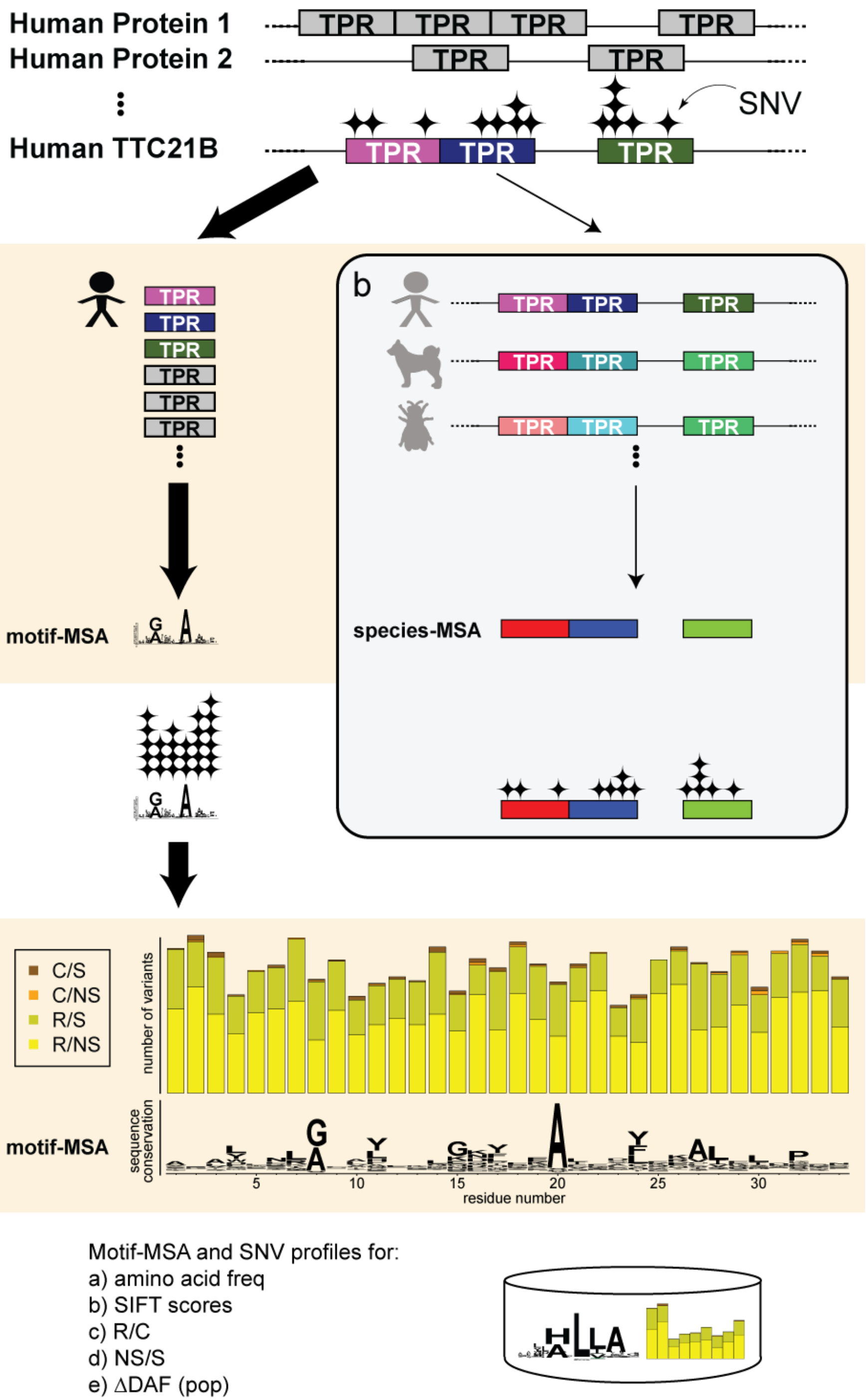
1. Find motifs

1. Generate motif-MSA

1. Map SNVs to motif-MSA

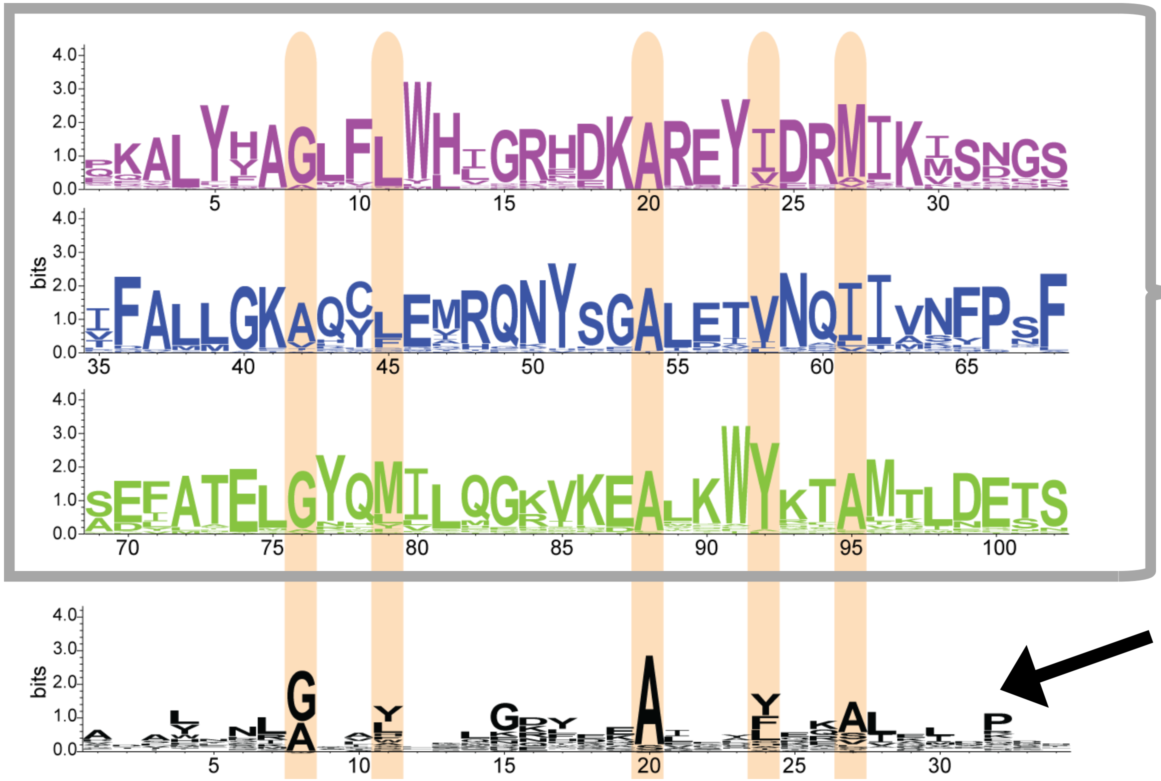
1. Evaluate SNV profiles

1. Store in database



Motif-MSA and SNV profiles for:
a) amino acid freq
b) SIFT scores
c) R/C
d) NS/S
e) ΔDAF (pop)

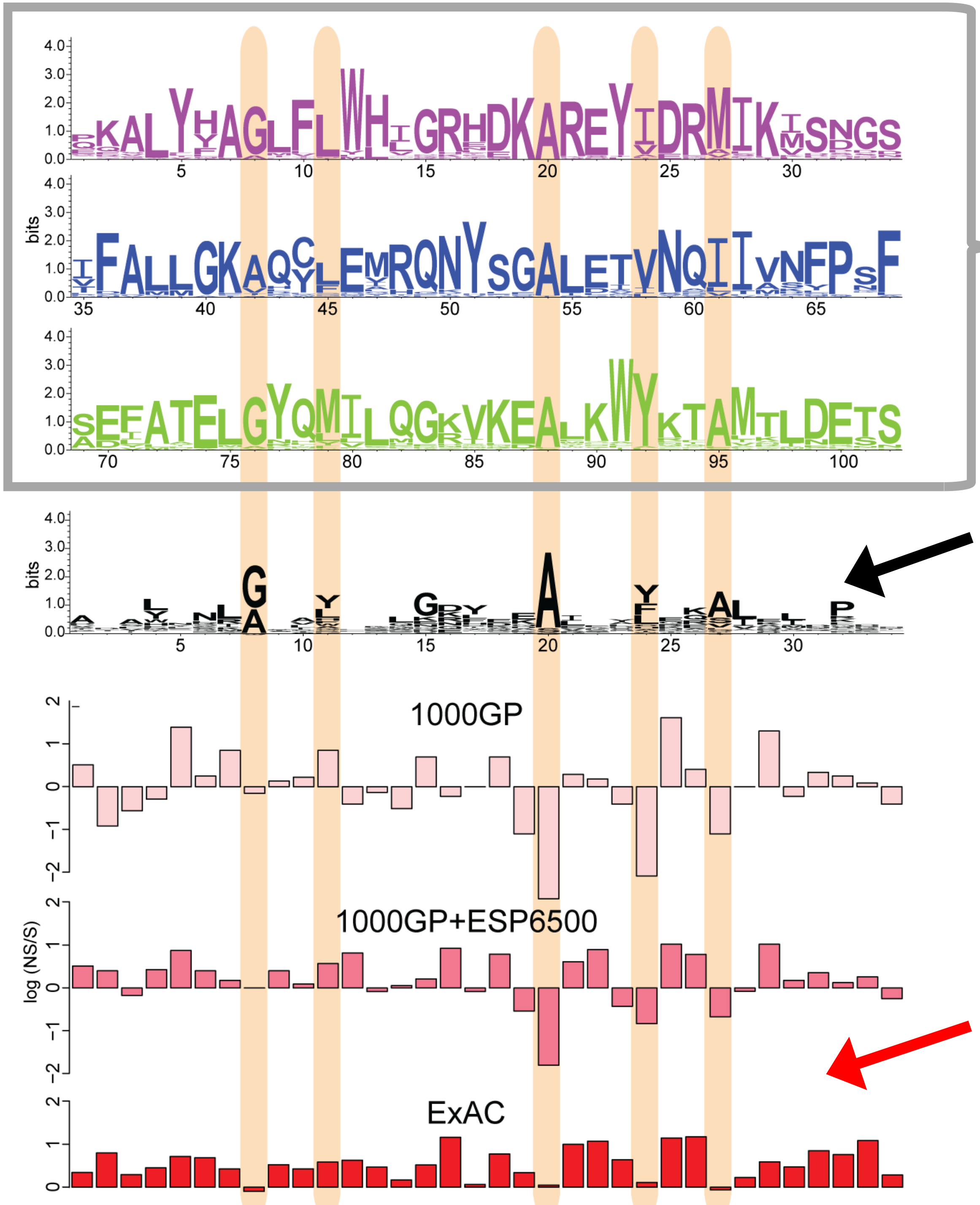
[Chen et al., JMB ('17, in revision)]



Species MSAs

Motif-MSA uncovers important positions missed by species-MSA

[Chen et al., JMB ('17, in revision)]

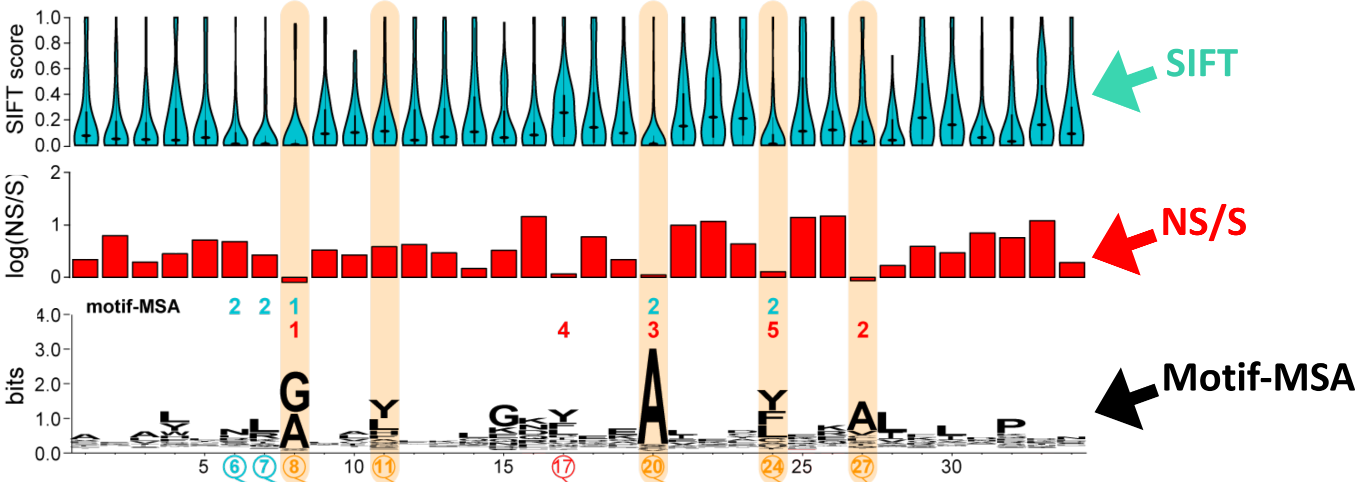


Species MSAs

Motif-MSA uncovers important positions missed by species-MSA

Signal-to-noise is the best in ExAC

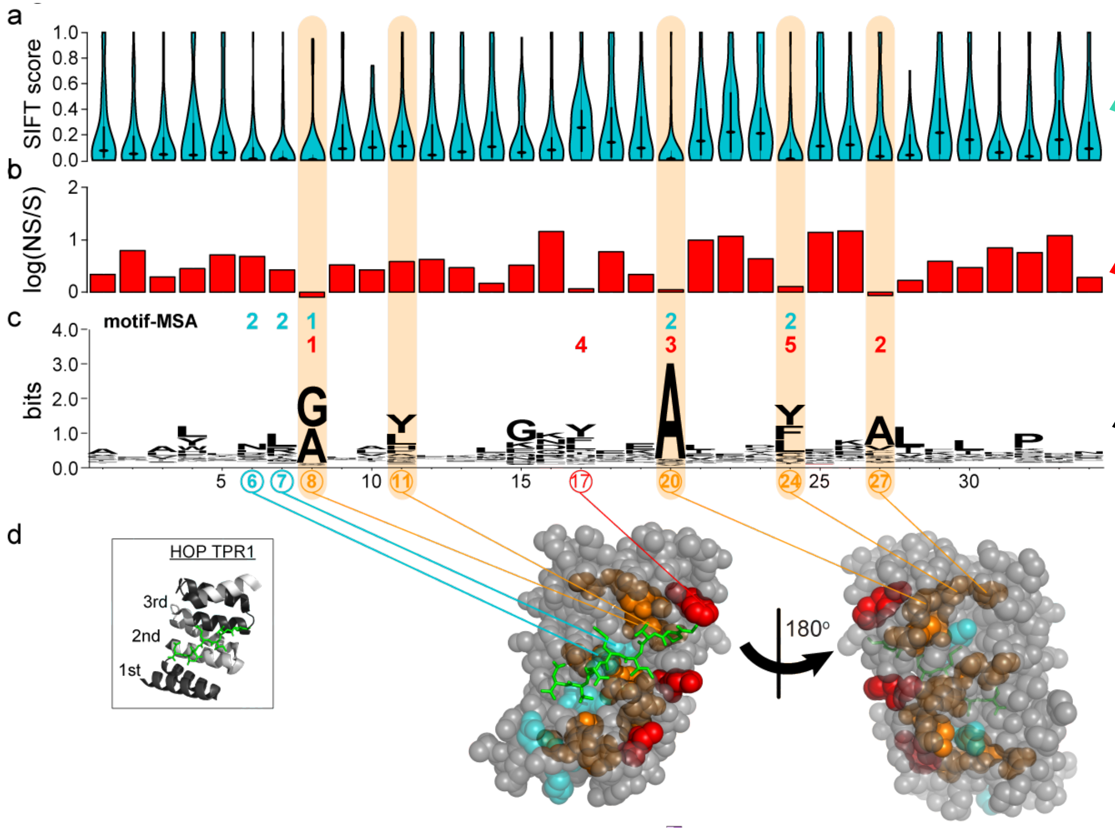
[Chen et al., JMB ('17, in revision)]



Selection in PPI motifs

[Chen et al., JMB ('17, in revision)]

Selection in PPI motifs



SIFT

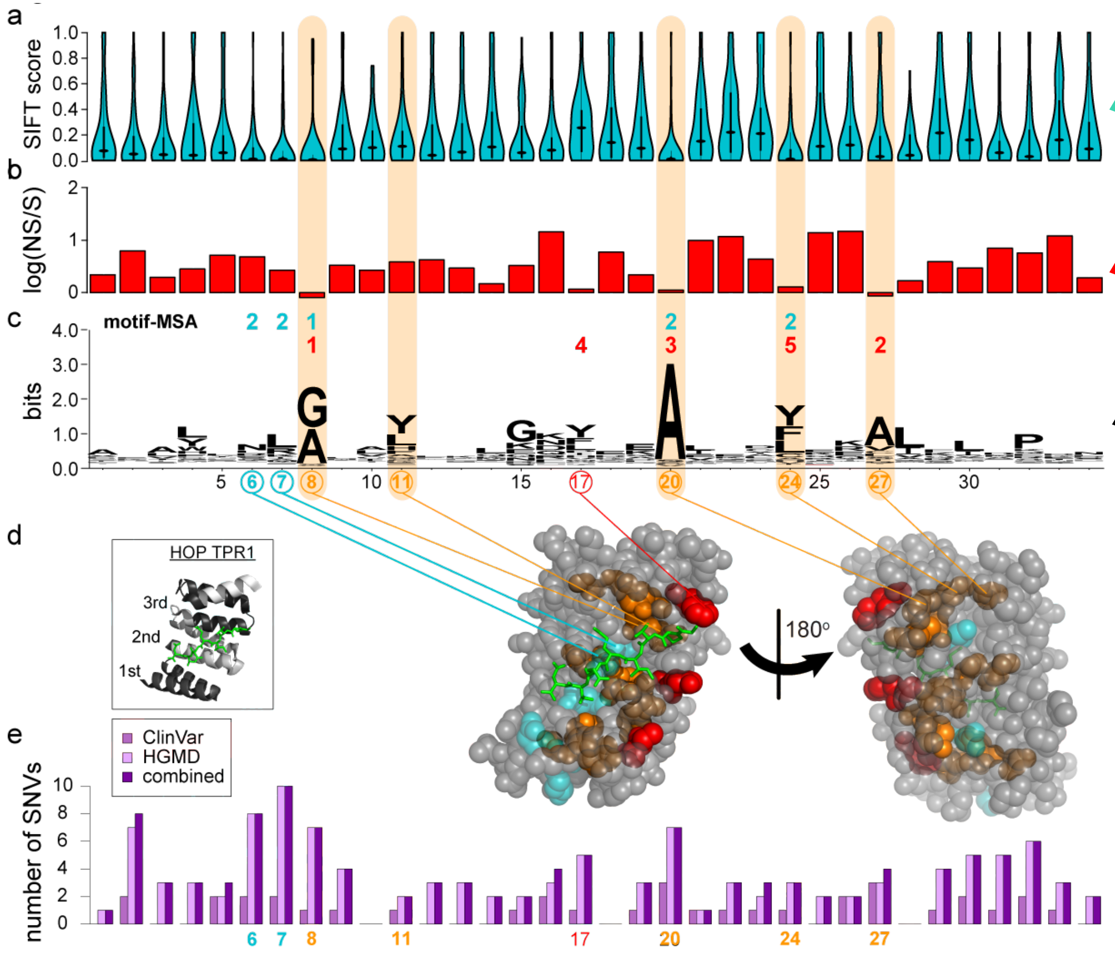
NS/S

Motif-MSA

How to check possible significance:
Burial within structure

[Chen et al., JMB ('17, in revision)]

Selection in PPI motifs



SIFT

NS/S

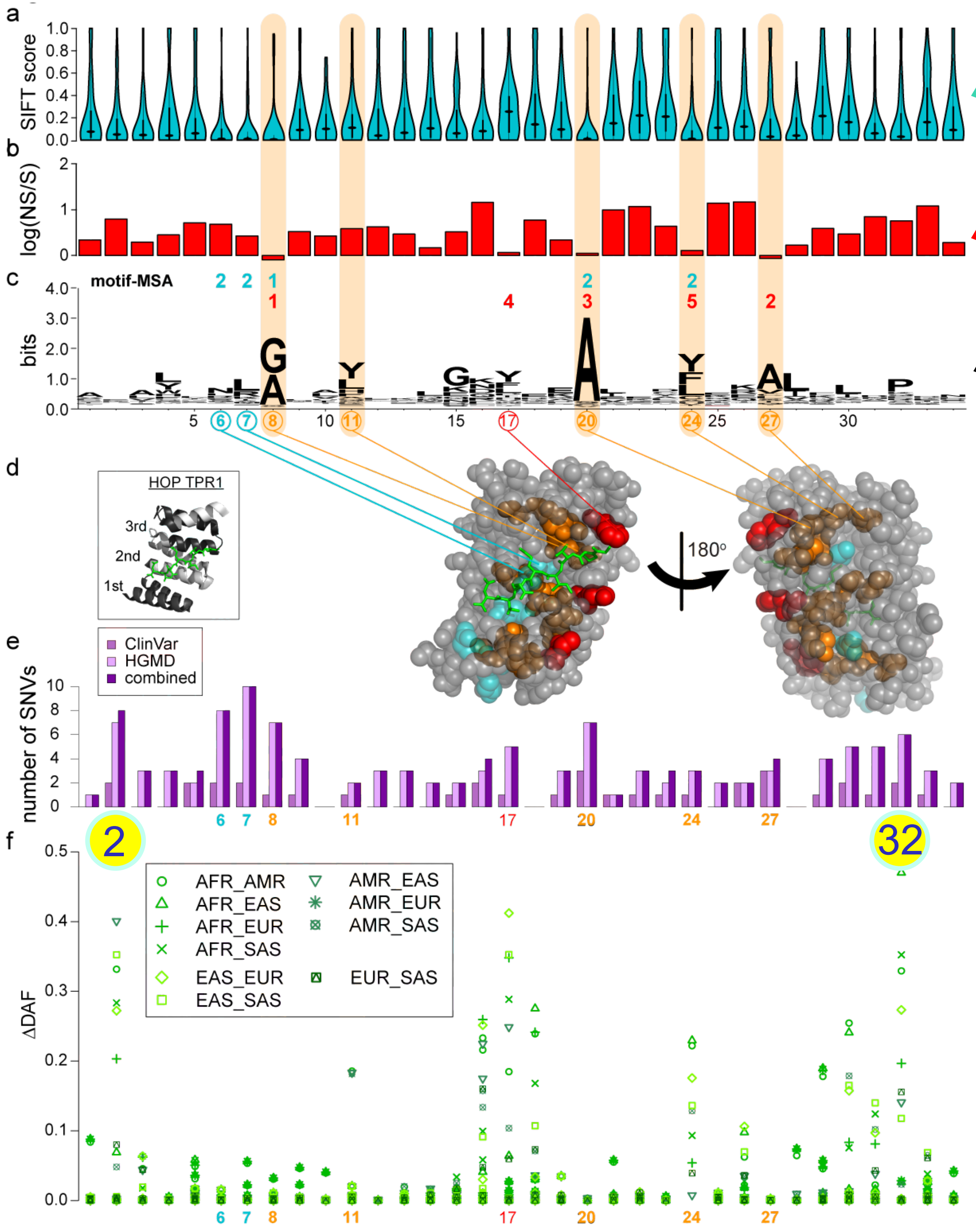
Motif-MSA

How to check possible significance:

-> Burial within structure

-> more SNVs implicated in diseases in ClinVar and HGMD

Selection in PPI motifs



SIFT

NS/S

Motif-MSA

How to check possible significance:

-> burial within structure

-> more SNVs implicated in diseases in ClinVar and HGMD

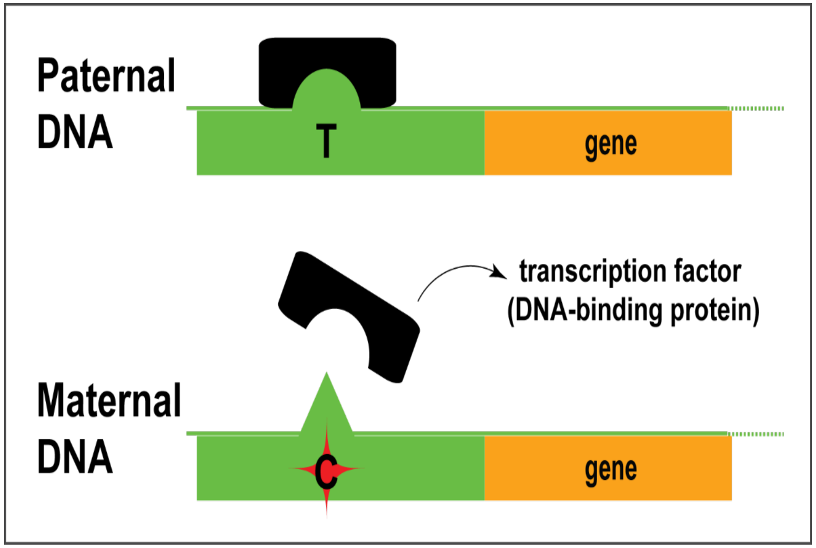
-> sites with increased human pop. differentiation might indicate important position

[Chen et al., JMB ('17, in revision)]

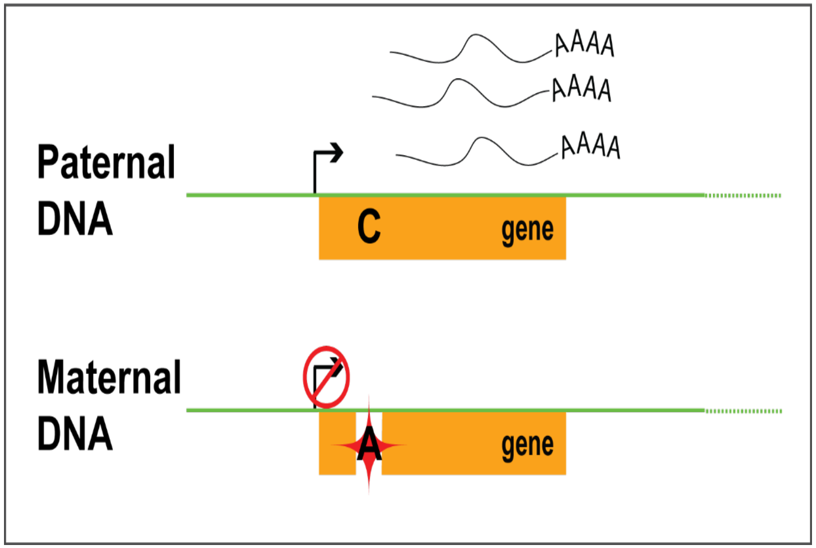
Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Allele-specific binding and expression



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

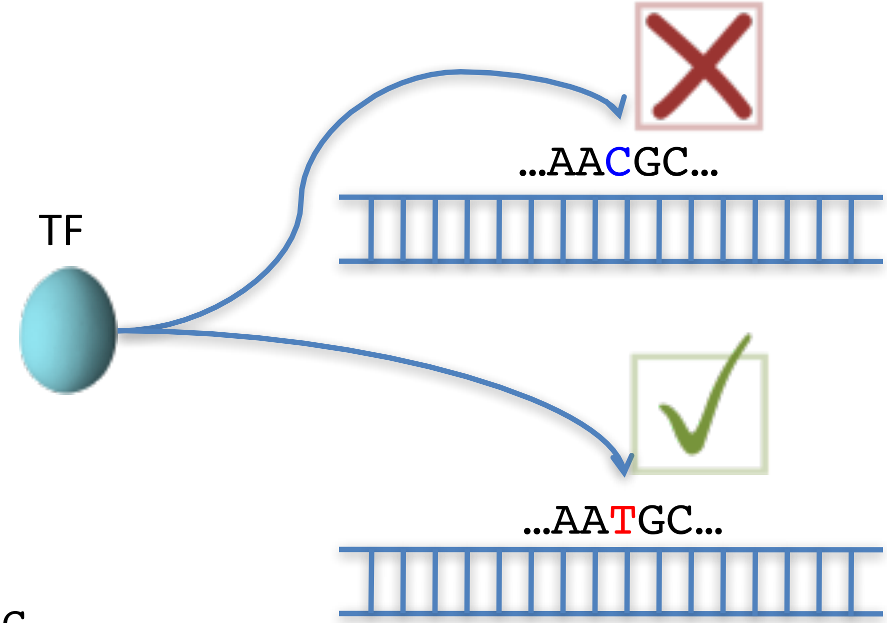


e.g. allele-specific expression
(ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

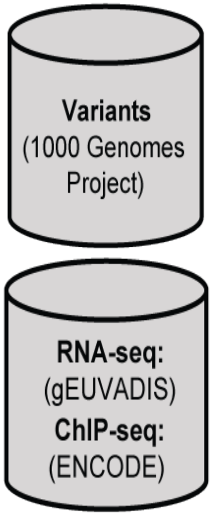
```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                TGCACGTTGGGAGTTGGC

           10 x T
            2 x C
```



Haplotypes with a Heterozygous Polymorphism

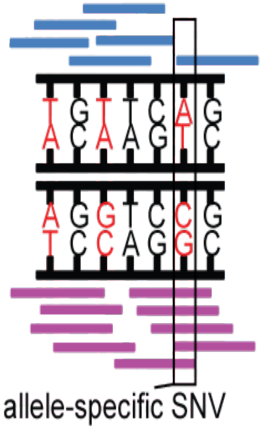
AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes



2. Align ChIP-seq & RNA-seq reads

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

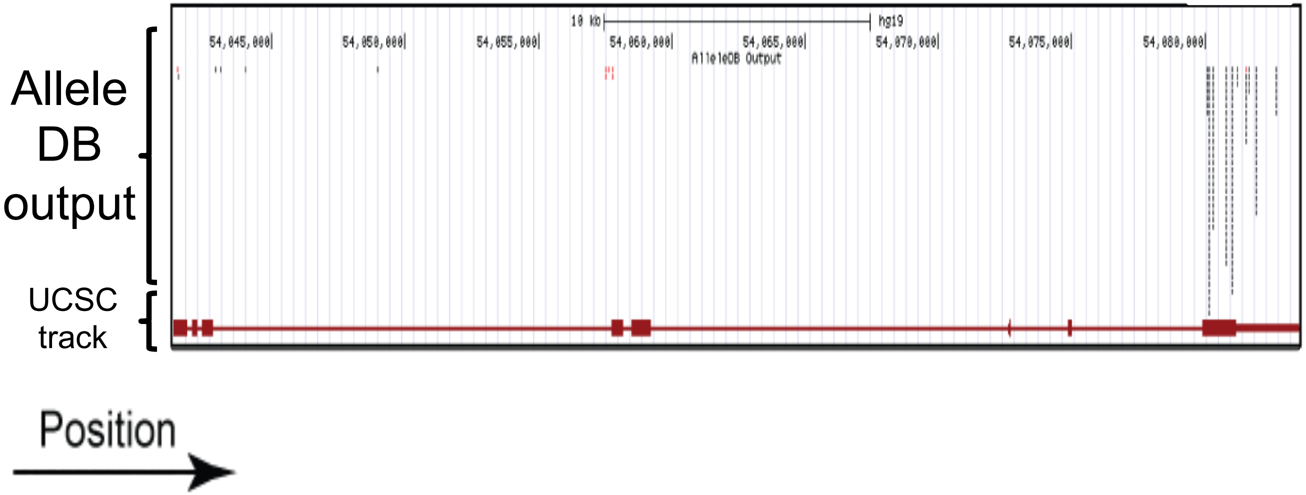


Many Technical Issues: Reference bias, Ambiguous mapping bias, Over-dispersed (non binomial null)

alleledb.gersteinlab.org

[Chen *et al.* ('16) *Nat. Comm.*]

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



Allele DB output

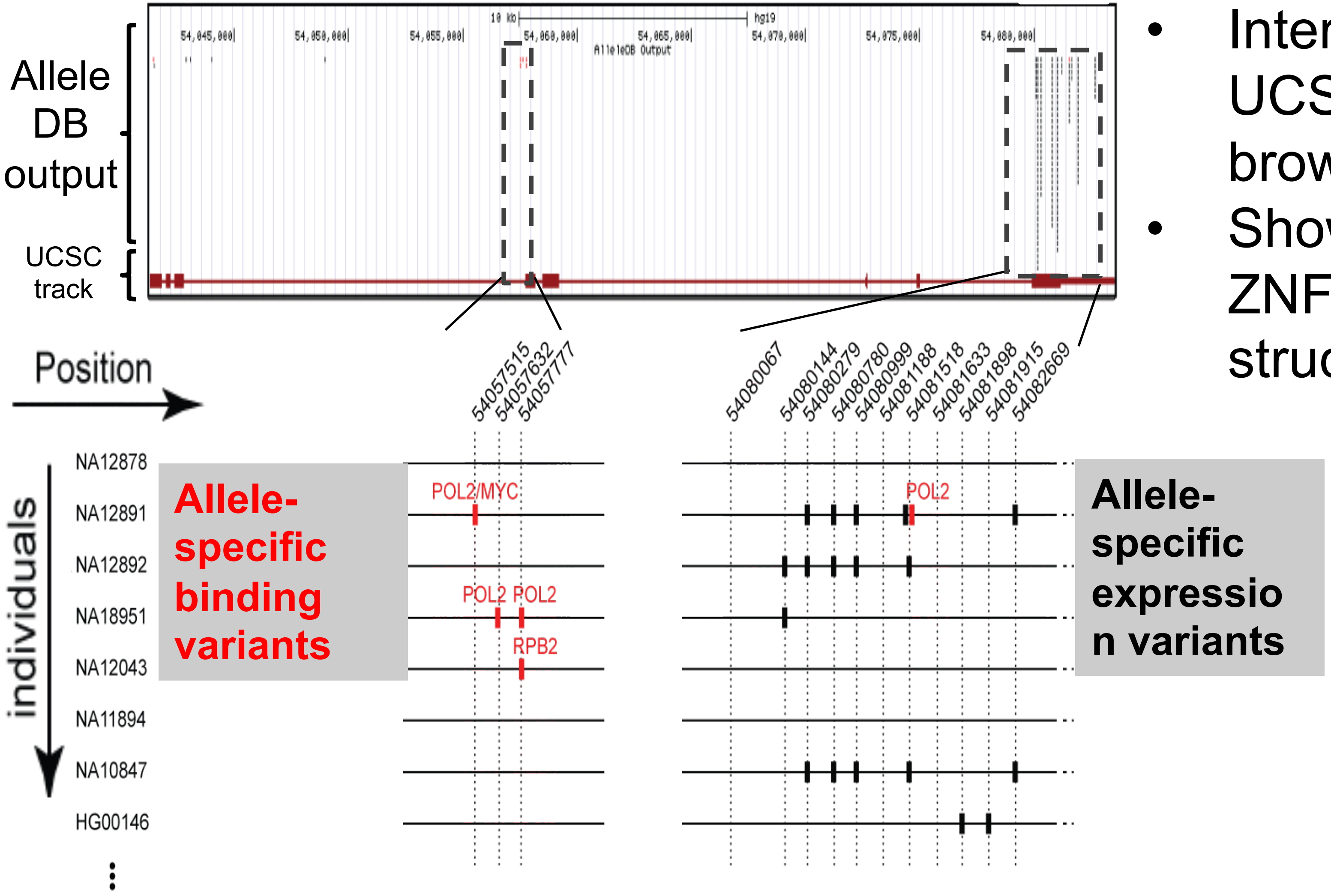
UCSC track

Position

- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

[Chen *et al.* ('16) *Nat. Comm.*]

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



Allele-specific binding variants

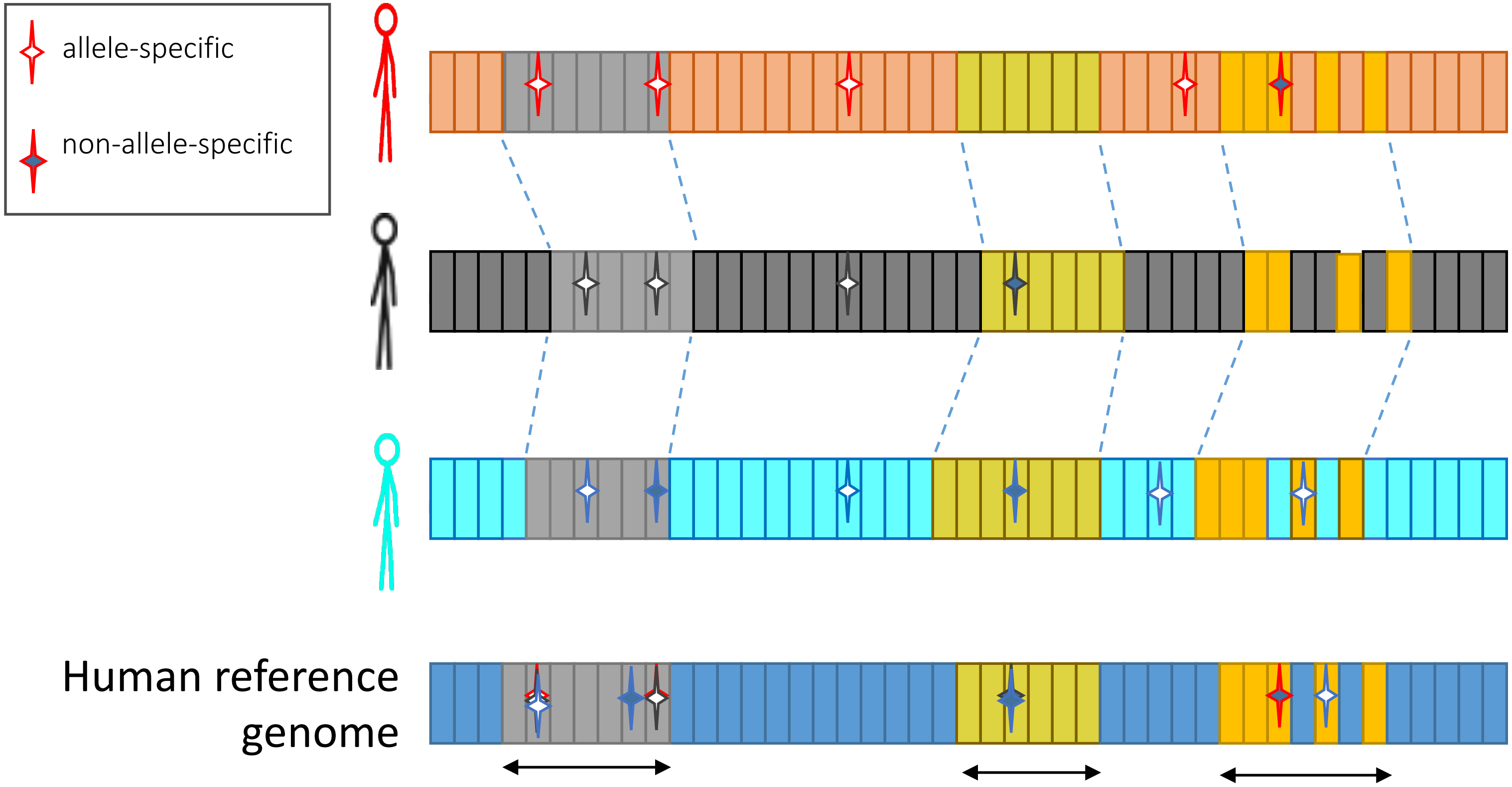
Allele-specific expression variants

- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

[Chen *et al.* ('16) *Nat. Comm.*]

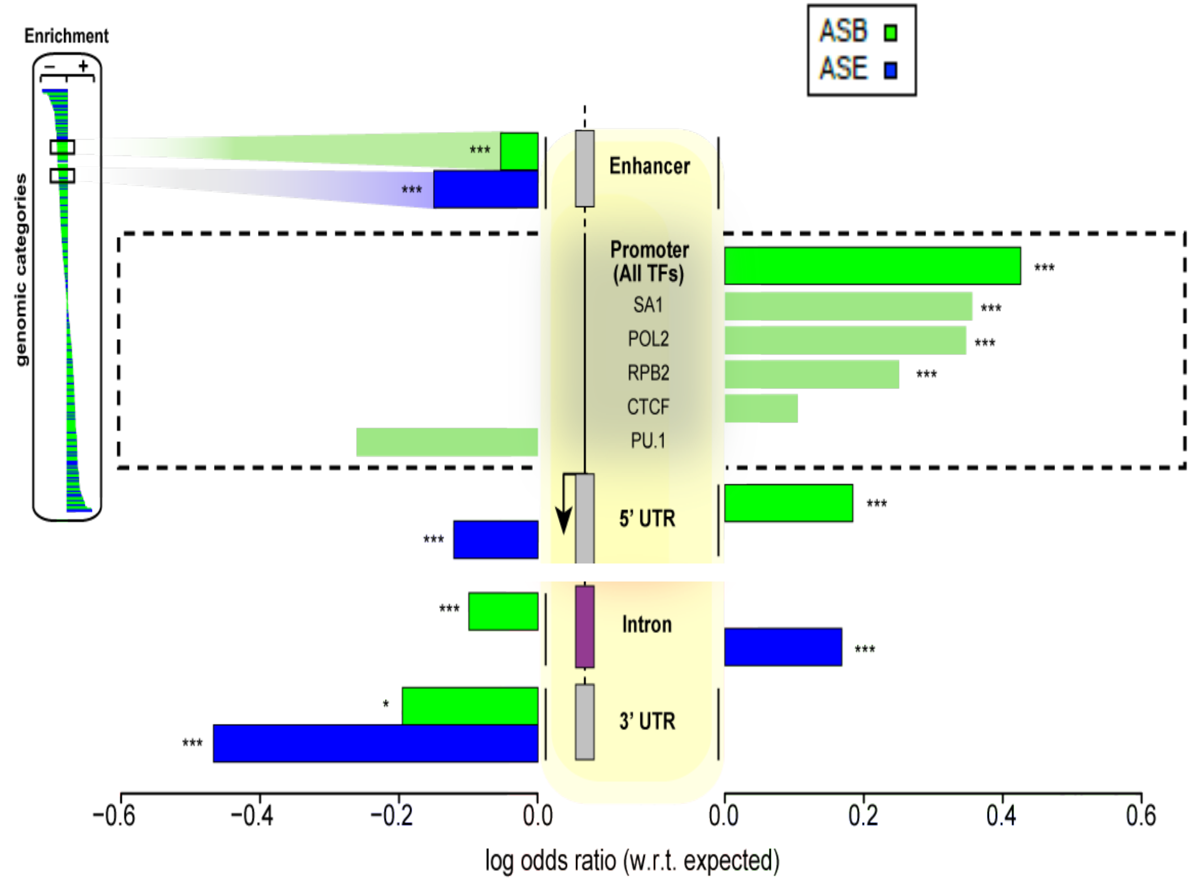
Collecting ASE/ASB variants into allele-specific genomic regions

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)



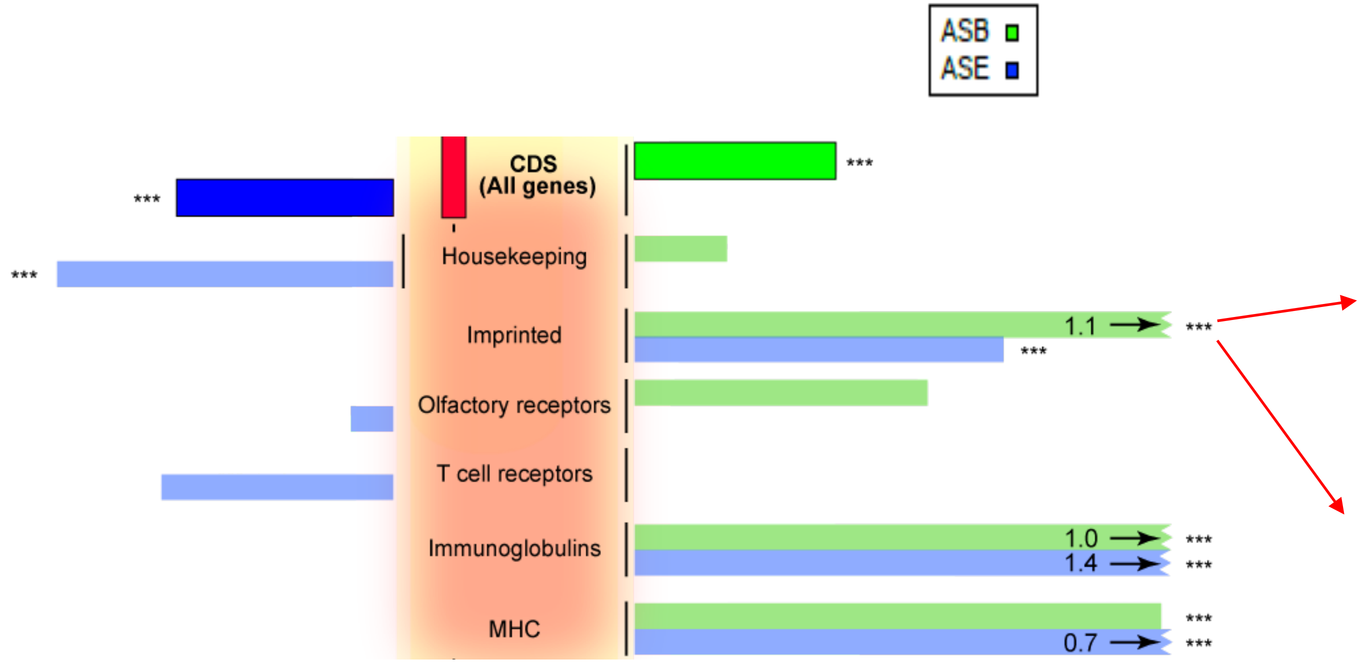
[Chen *et al.* ('16) *Nat. Comm.*]

Groups of elements that are enriched or depleted in allelic activity



[Chen *et al.* ('16) *Nat. Comm.*]

Groups of elements that are enriched or depleted in allelic activity



- SNURF imprinted and implicated in Prader-Willi/Angelman Syndrome
- KCNQ1 is an imprinted gene

[Chen *et al.* ('16) *Nat. Comm.*]

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Analysis of Personal Genomes: Evaluating the impact of variants in exomes using protein structure & allelic activity

- Introduction
  - Rare v common variants
  - The importance of interpreting rare coding variants in the context of disease genomics (CMG,TCGA)
- Identifying cryptic allosteric sites with STRESS
  - On surface & in interior bottlenecks
- Using changes in localized frustration to find further sites sensitive to mutations
  - Difference betw. TSGs & oncogenes
- Using structural motifs (eg TPR) for intensification of weak population genetic signals
  - For both negative and positive selection
- Prioritizing allelic genes using AlleleDB
  - Having observed difference in molecular activity in many contexts

Acknowledgments

github.com/gersteinlab/

Frustration

S Kumar,

D Clarke

(Intensification)

MotifVar.gersteinlab.org

J Chen,

B Wang, L Regan

AlleleDB.gersteinlab.org

J Chen,

J Rozowsky,

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

STRESS.molmovdb.org

D Clarke, A Sethi,

S Li, S Kumar,
R Chang, J Chen

Hiring Postdocs. See gersteinlab.org/jobs

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt