

Leveraging Protein Structure & Dynamics for Variant Interpretation in Coding Regions



Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via [@MarkGerstein](https://twitter.com/MarkGerstein)).

No Conflicts for this Talk. See last slide for more info.

Human Genetic Variation: the prevalence of rare variants in population studies

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



Driver (~0.1%)

Class of Variants

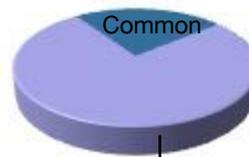
SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants



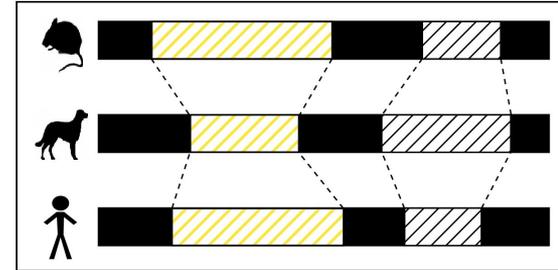
Rare* (1-4%)

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

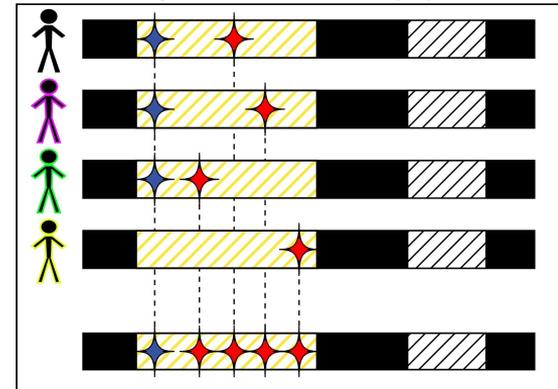


Rare (~75%)

Similarity across species



Similarity within the human population



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

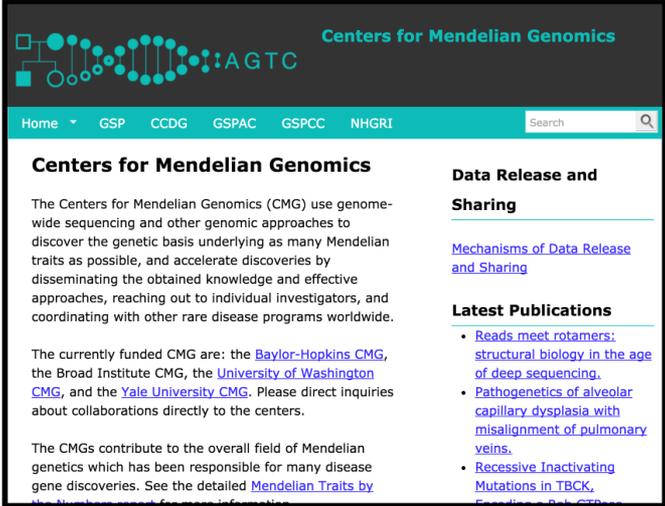
Rare variant analysis particularly applicable at the moment because of the many exomes

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



The screenshot shows the homepage of the Centers for Mendelian Genomics (CMG). The header features the CMG logo and the text "Centers for Mendelian Genomics" and "AGTC". Below the header is a navigation menu with links for "Home", "GSP", "CCDG", "GSPAC", "GSPCC", and "NHGRI", along with a search bar. The main content area is divided into two columns. The left column contains the title "Centers for Mendelian Genomics" followed by a paragraph describing the CMG's mission: "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." Below this is a paragraph listing the currently funded CMG centers: "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 right column contains the title "Data Release and Sharing" followed by a link to "[Mechanisms of Data Release and Sharing](#)". Below this is the title "Latest Publications" followed by a list of three publications: "• [Reads meet rotamers: structural biology in the age of deep sequencing.](#)", "• [Pathogenetics of alveolar capillary dysplasia with misalignment of pulmonary veins.](#)", and "• [Recessive Inactivating Mutations in TBCK.](#)"

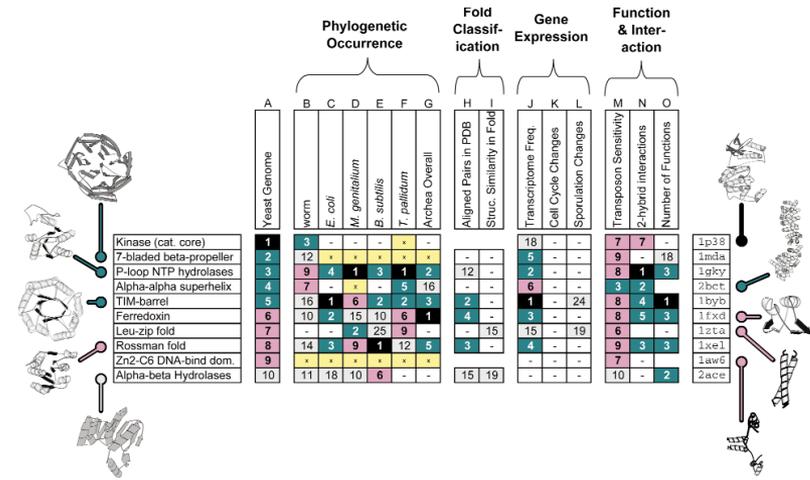
Exomes have the current potential for great scale with the better impact interpretability of coding variants, often in a region of known protein structure

Scale of EXAC, >60K exomes [Lek et al. '16]

Structure & genomics

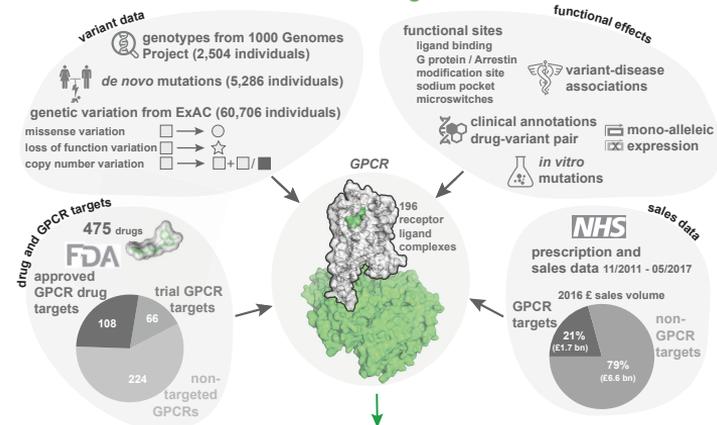
Structure particularly useful for interpreting the impact of the many rare variants whose effect can not be found via GWAS

Also, integration of structure data with genomic variants, EHR & drug data will be key for realizing the goal of precision medicine.



Gerstein et al., Nat. Struct. Bio. 2000

GPCR Pharmacogenomics

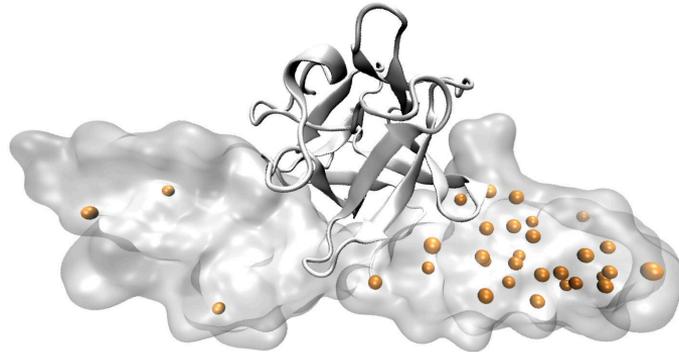


Assessing the spectrum, prevalence and functional impact of genetic variation for alteration in drug response

Hauser et al., Cell 2018

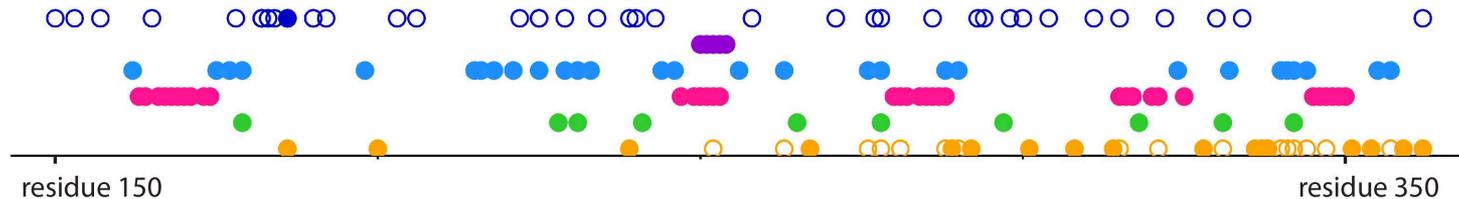
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



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

Fibroblast growth factor receptor 2 (pdb: 1IIL)



[Sethi et al. COSB ('15)]

Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

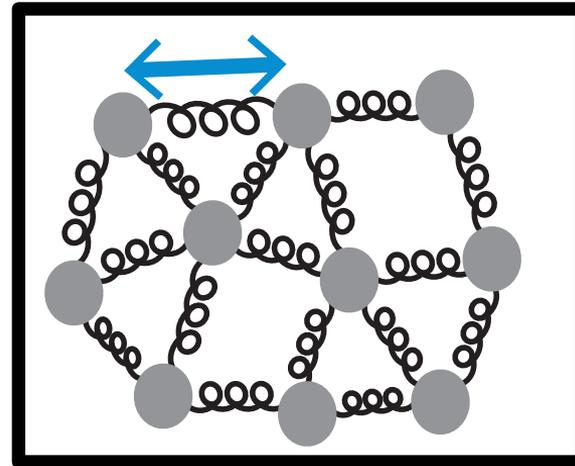
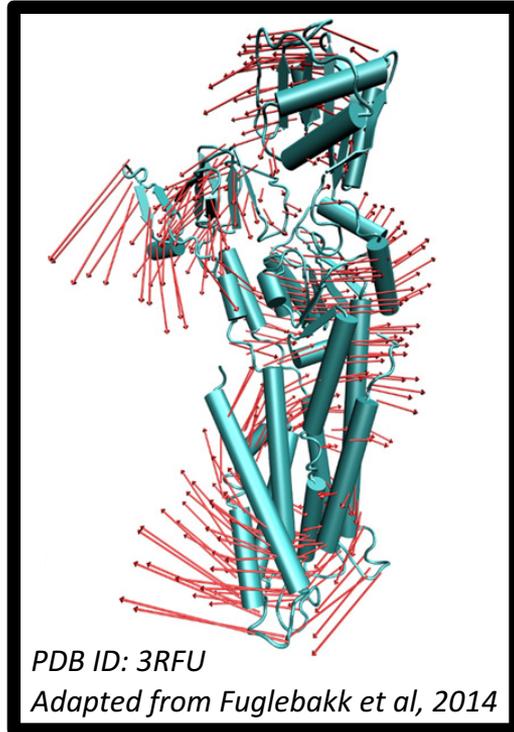
- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

Models of Protein Conformational Change

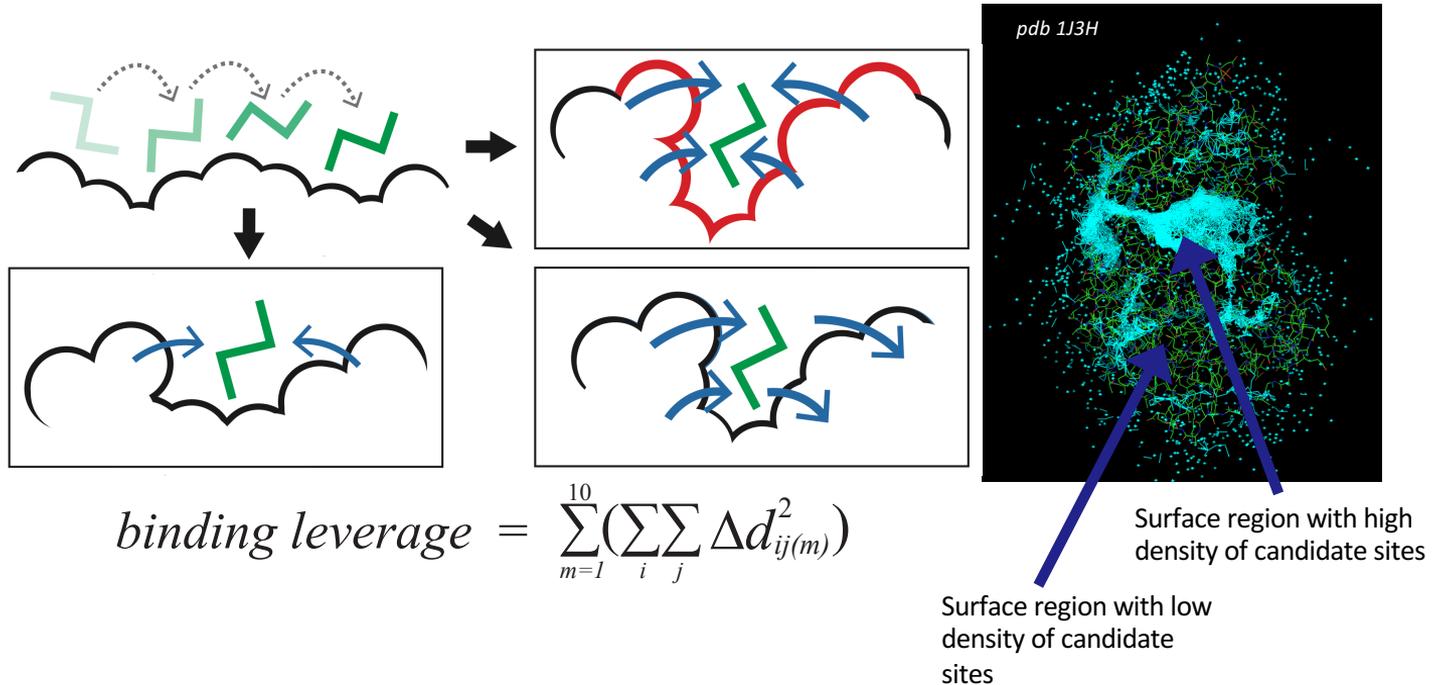
Motion Vectors from Normal Modes (ANMs)



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

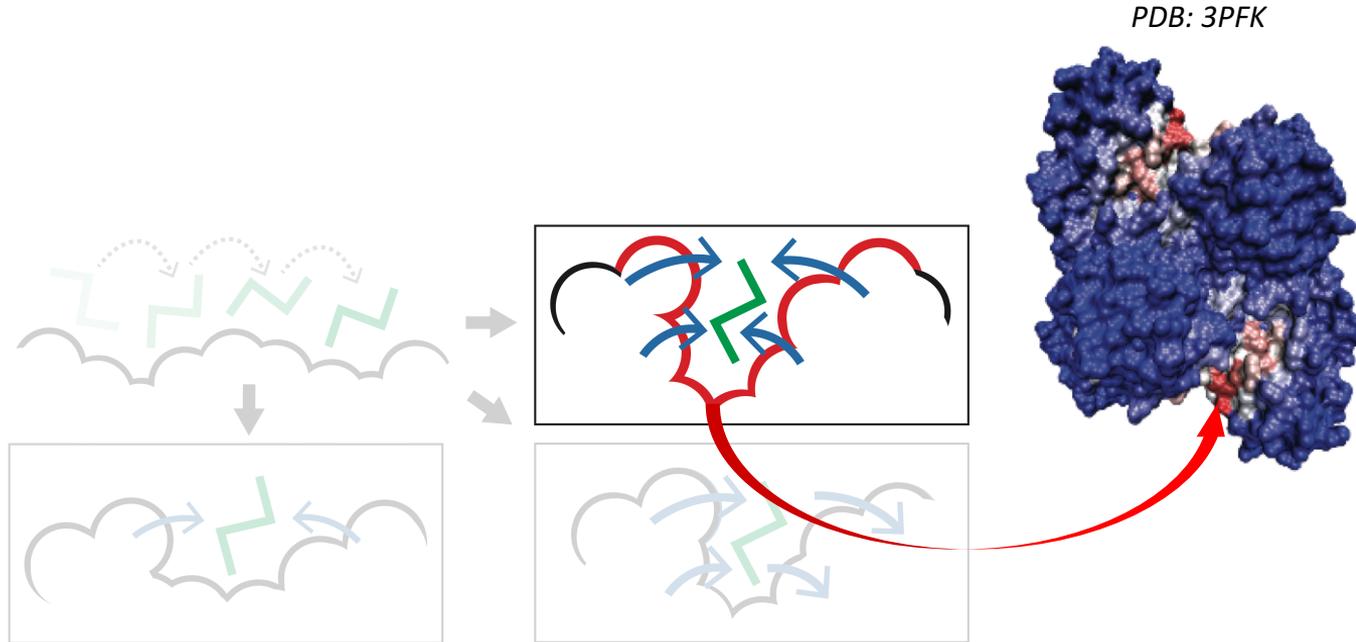
Predicting Allosterically-Important Residues at the Surface

1. MC simulations generate a large number of candidate sites
2. Score each candidate site by the degree to which it perturbs large-scale motions
3. Prioritize & threshold the list to identify the set of high confidence-sites



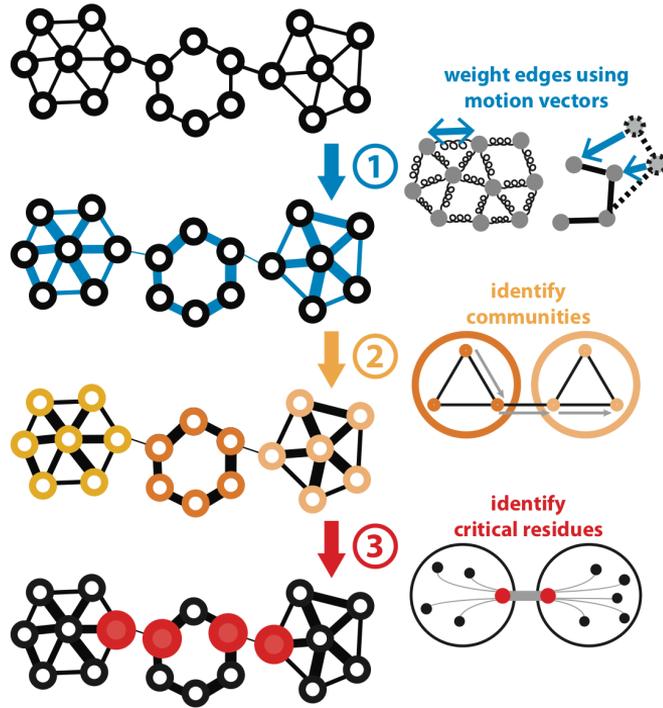
Adapted from Clarke*, Sethi*, et al ('16)

Predicting Allosterically-Important Residues at the Surface



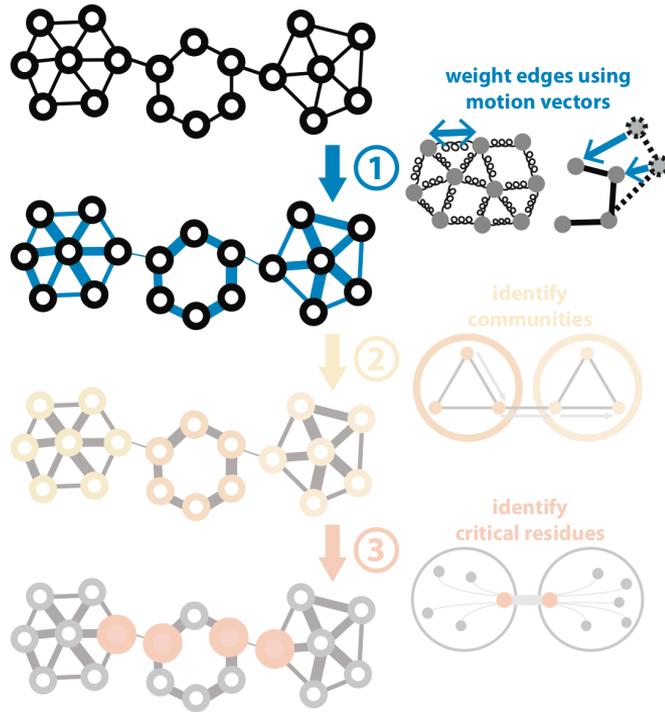
Adapted from Clarke*, Sethi*, et al ('16)

Predicting Allosterically-Important Residues within the Interior



Adapted from Clarke*, Sethi*, et al ('16)

Predicting Allosterically-Important Residues within the Interior

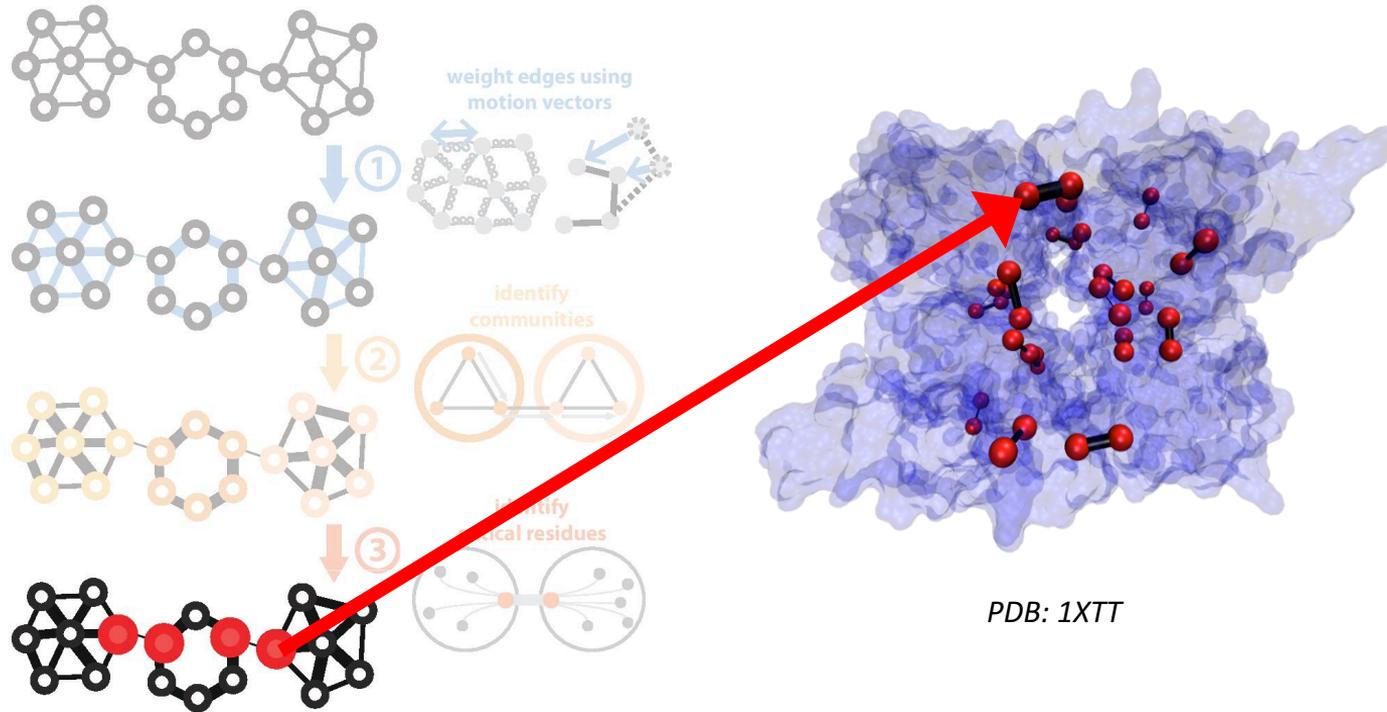


$$Cov_{ij} = \langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle$$

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

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

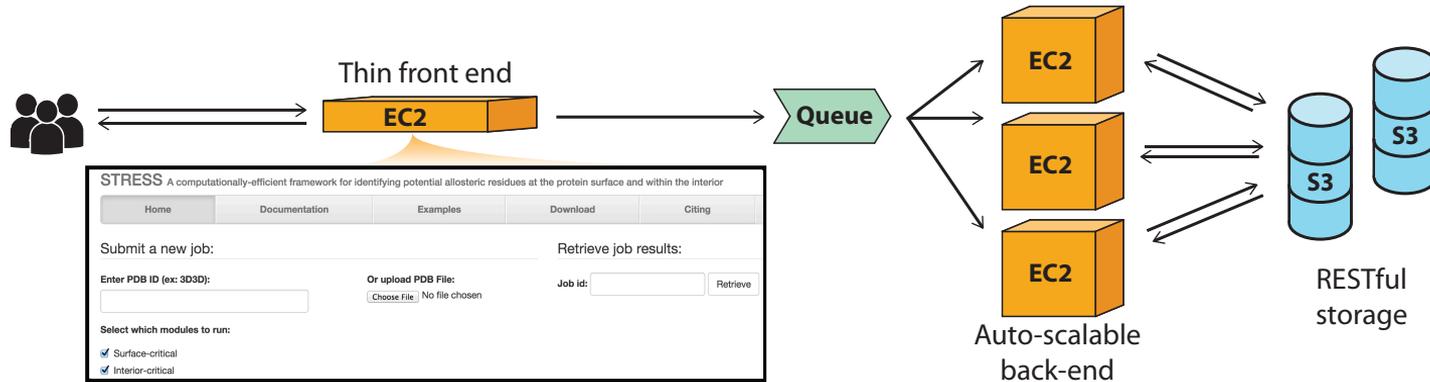
Predicting Allosterically-Important Residues within the Interior



Adapted from Clarke*, Sethi*, et al ('16)

STRESS Server Architecture: Highlights

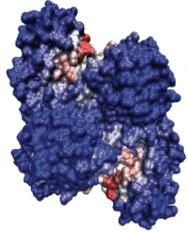
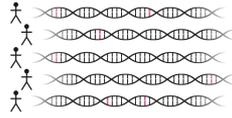
stress.molmovdb.org



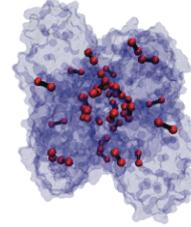
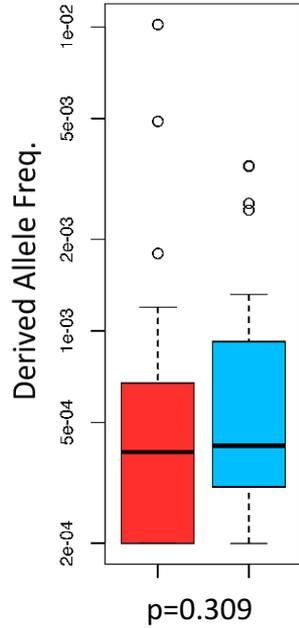
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

Intra-species conservation of predicted allosteric residues

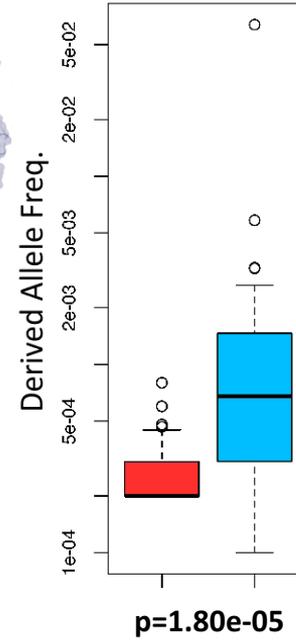
1000 Genomes



Surface



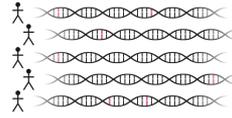
Interior



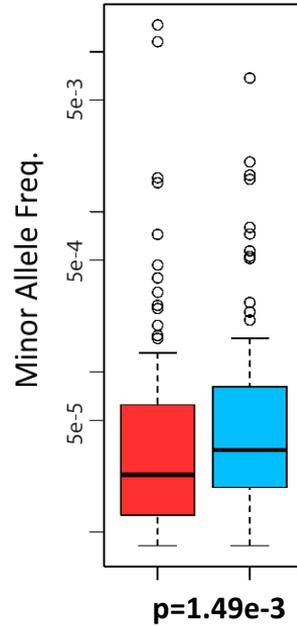
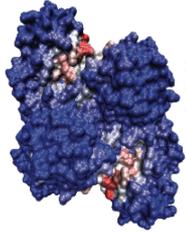
critical
non-critical

Intra-species conservation of predicted allosteric residues

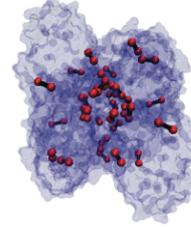
ExAC



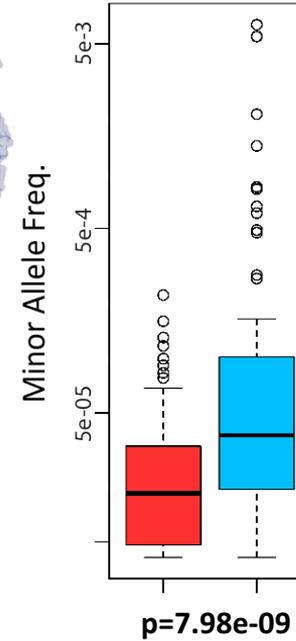
Surface



Interior

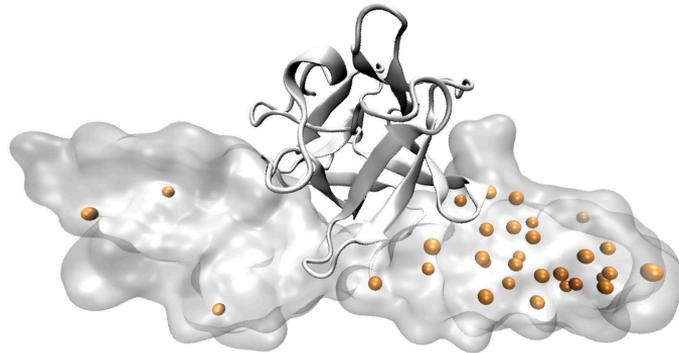


critical
non-critical



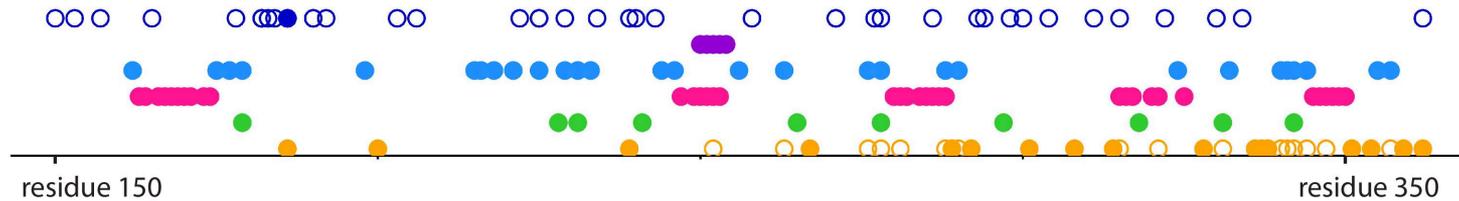
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

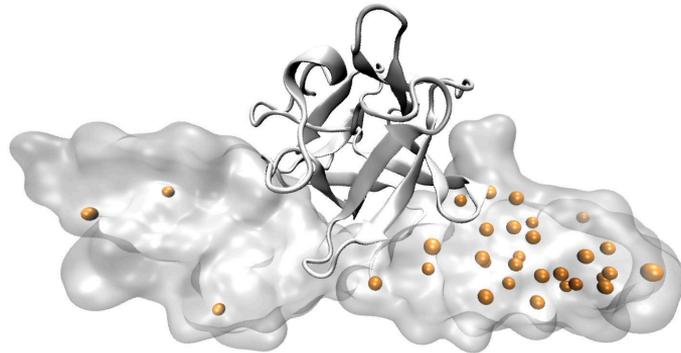
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

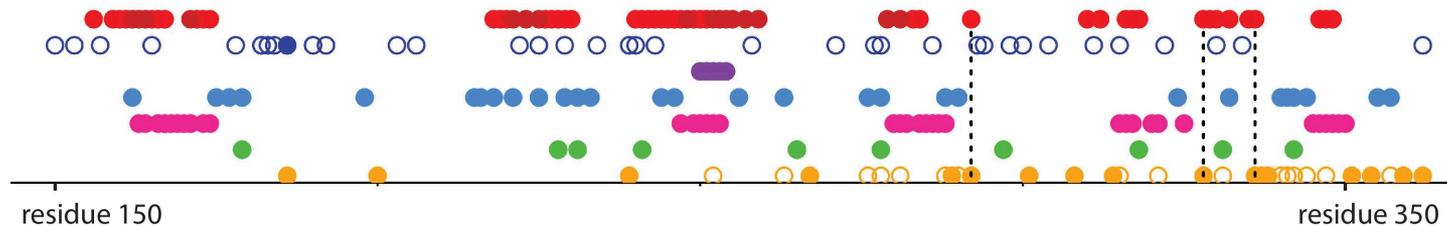
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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



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

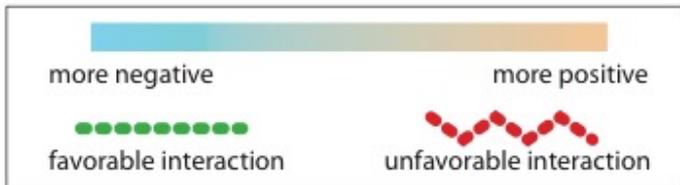
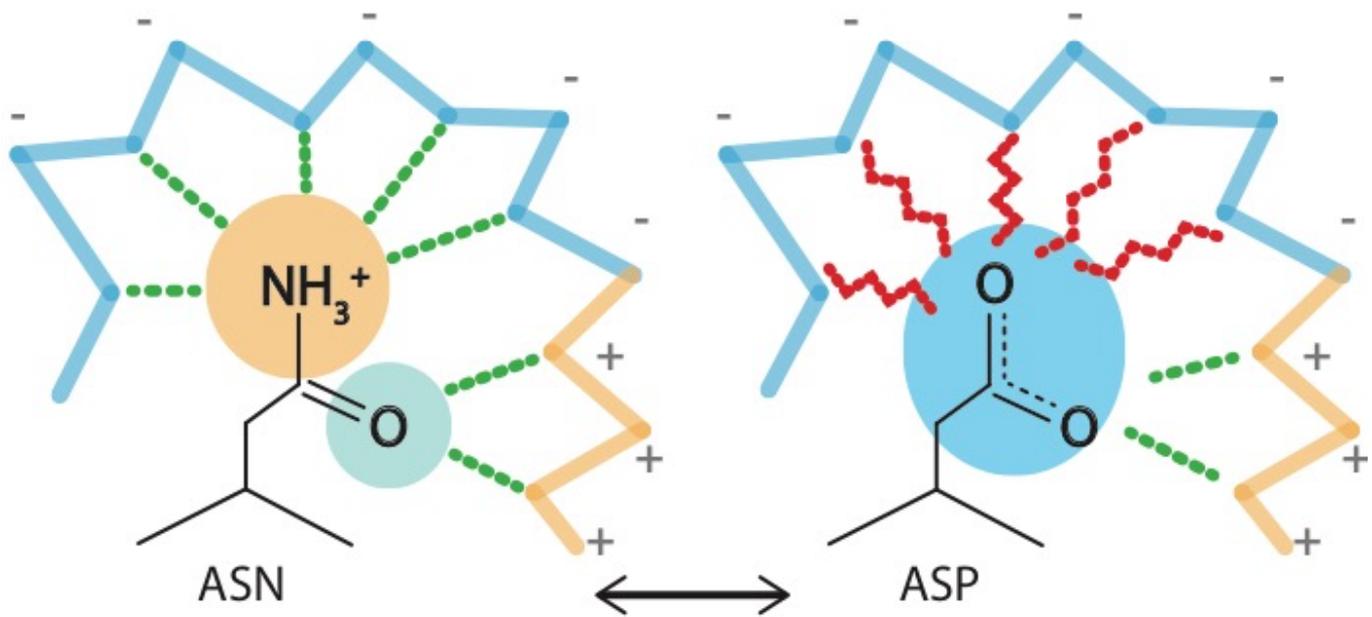
Fibroblast growth factor receptor 2 (pdb: 1IIL)



[Sethi et al. COSB ('15)]

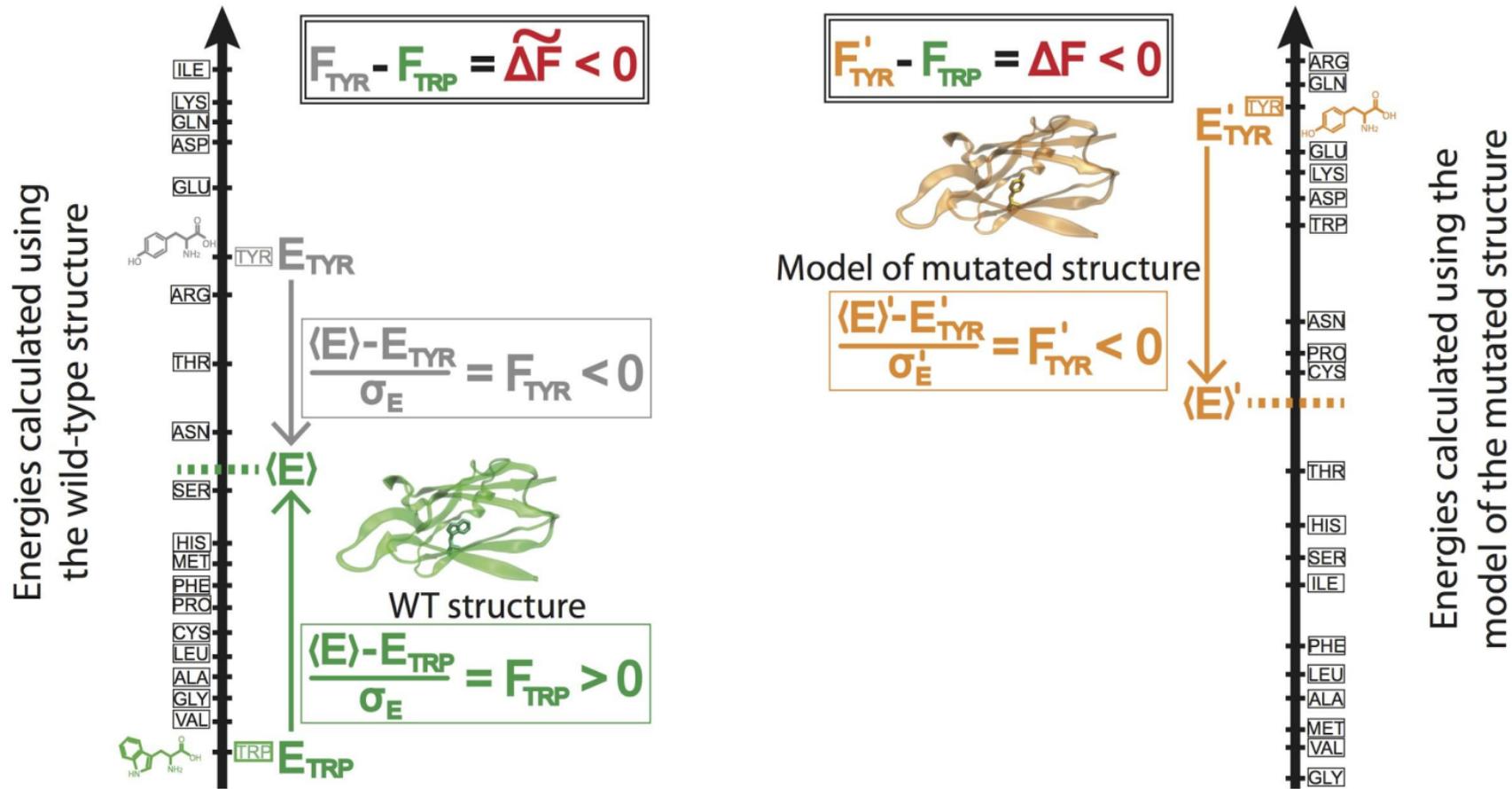
Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

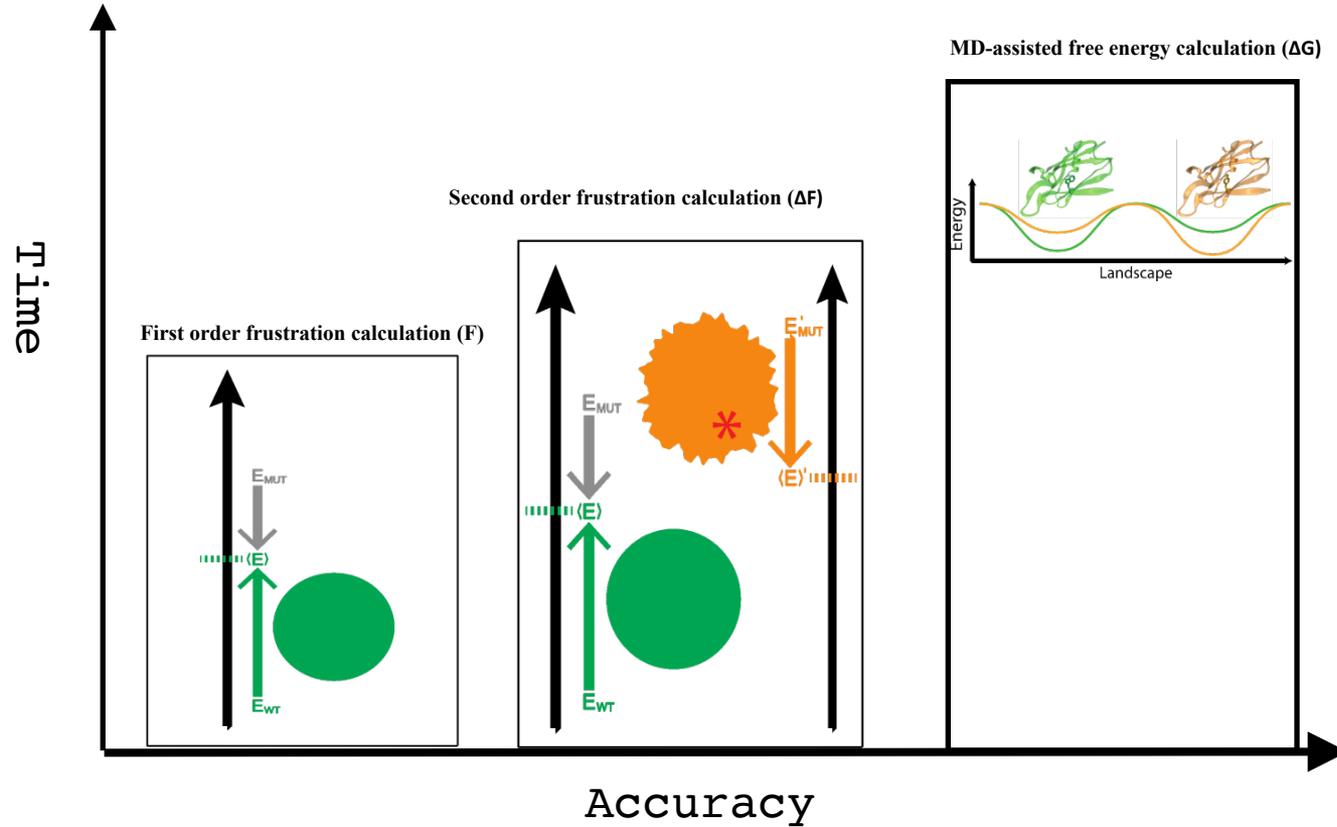


What is localized frustration ?

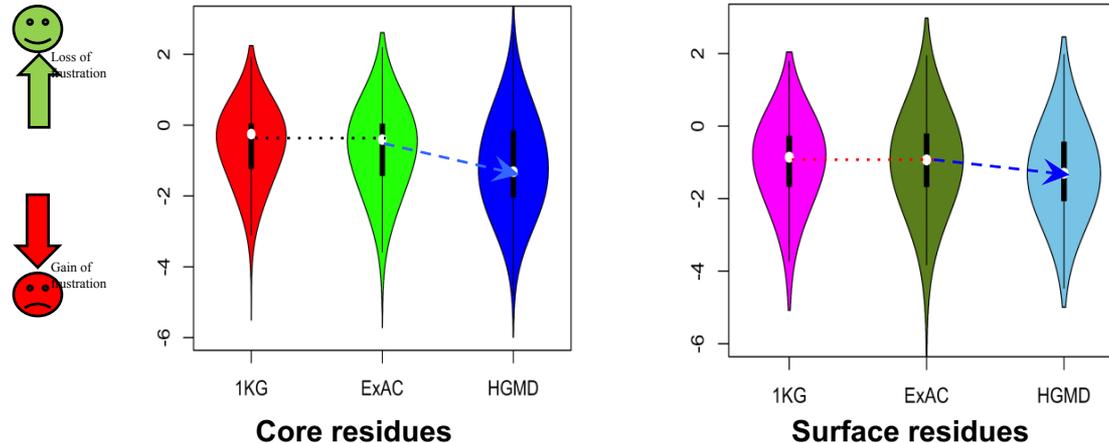
Workflow for evaluating localized frustration changes (ΔF)



Complexity of the second order frustration calculation

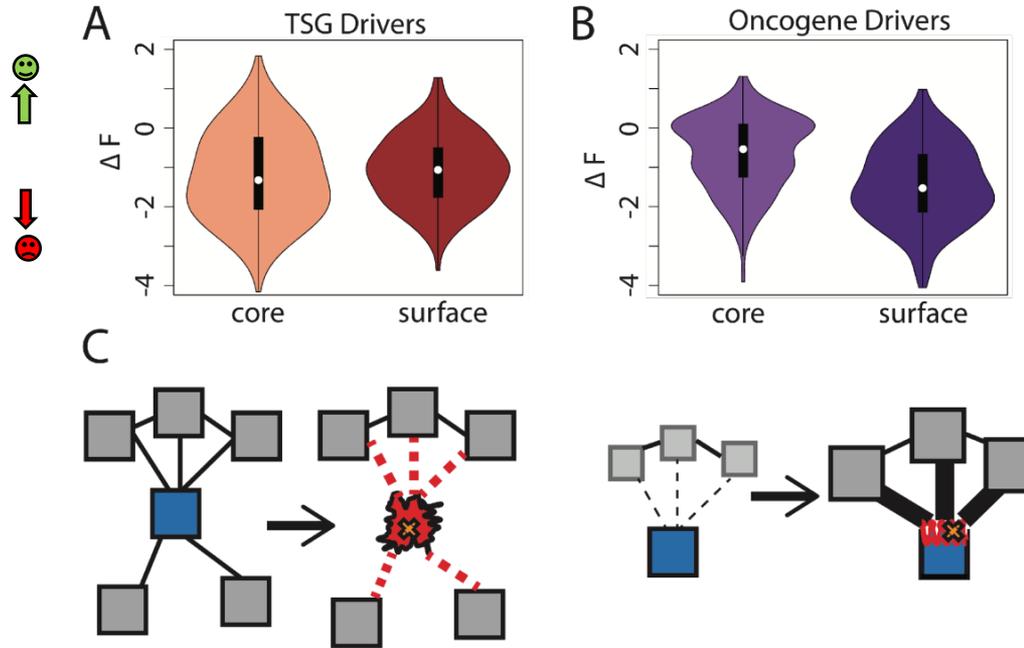


Comparing ΔF values across different SNV categories: disease v normal



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

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.

Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

Variant Annotation Tool (VAT), developed for 1000G FIG

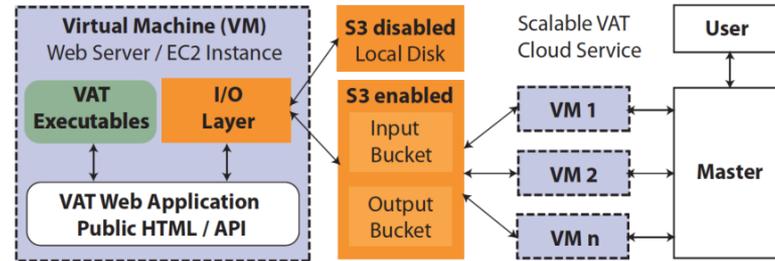
VCF Input

Output:

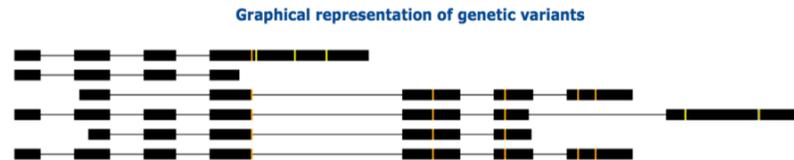
- Annotated VCFs
- Graphical representations of functional impact on transcripts

Access:

- Webserver
- AWS cloud instance
- Source freely available



CLOUD APPLICATION



vat.gersteinlab.org

Habegger L. *, Balasubramanian S. *, et al. *Bioinformatics*, 2012

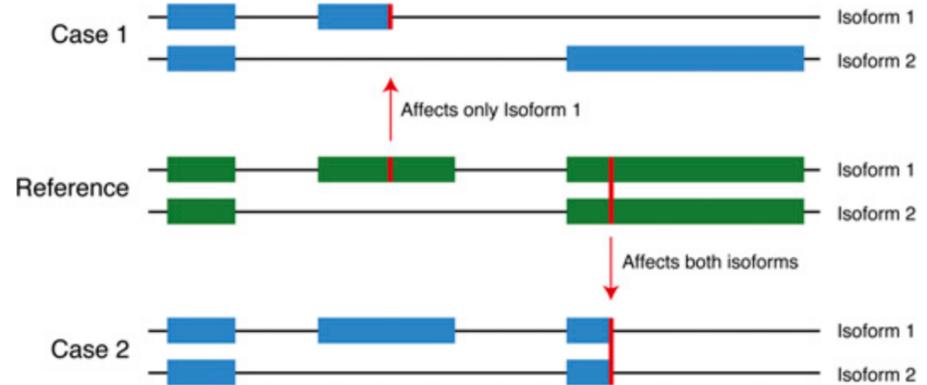
Complexities in LOF annotation

Transcript isoforms,
distance to stop,
functional domains,
protein folding,
etc.

Balasubramanian S. et al., *Genes Dev.*, '11

Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

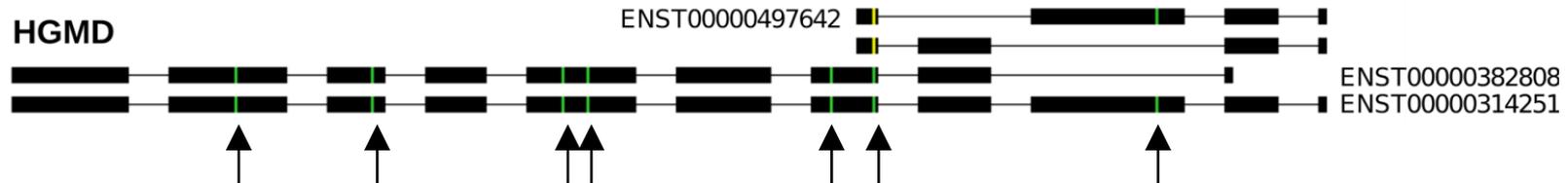
Impact of a SNP on alternate splice forms



1KG



HGMD



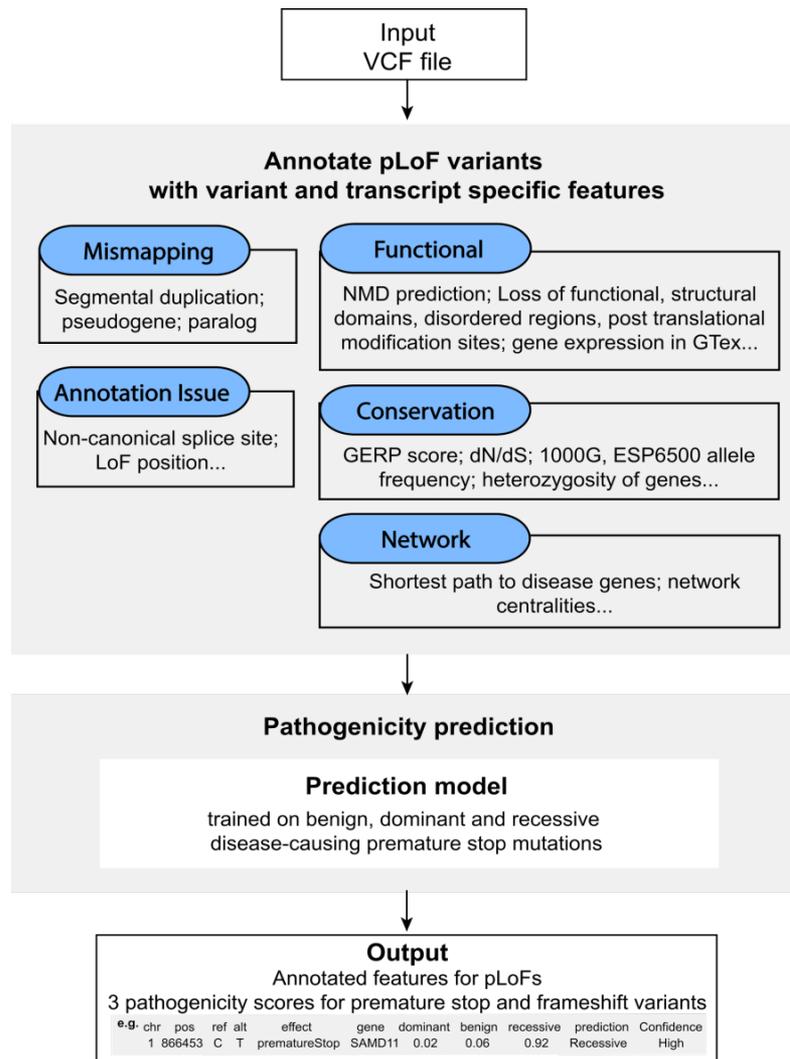
Annotation of Loss-of-Function Transcripts (ALoFT)

Runs on top of VAT

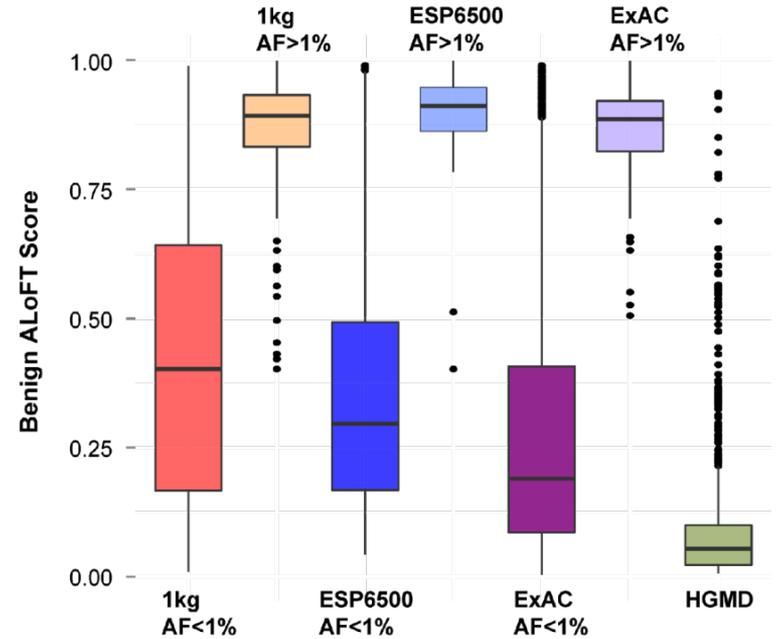
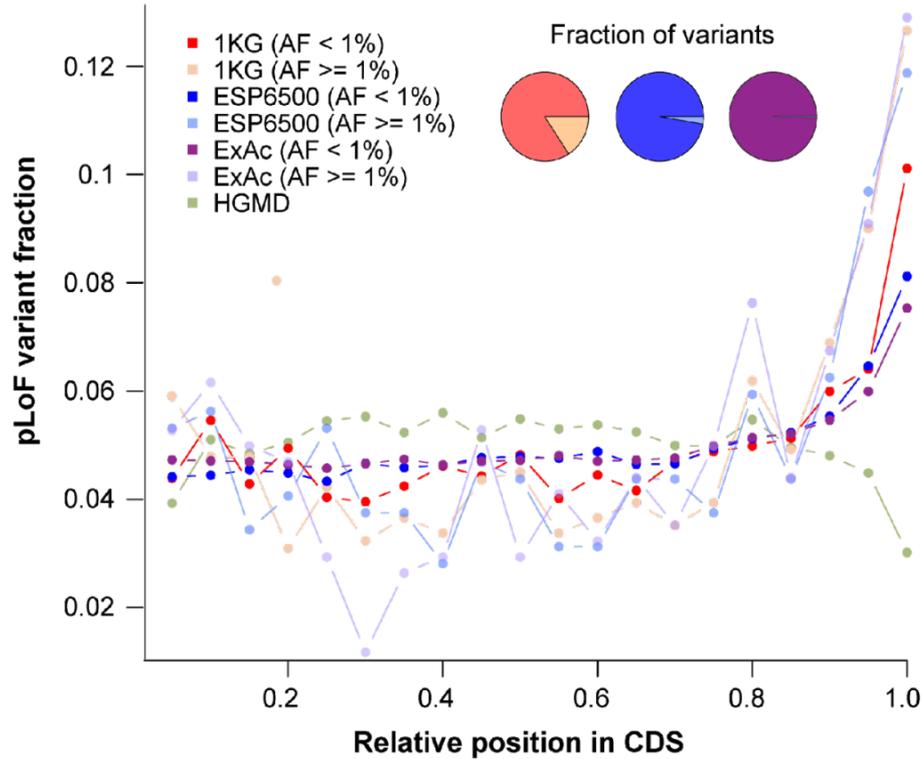
Output:

- Impact score: benign or deleterious.
- Decorated VCF.

Balasubramanian S.* , Fu Y.* et al., *NComms.*, '17



LoF distribution varies as expected by mutation set (from healthy people v from disease)



Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

ALoFT identifies deleterious somatic LoF variants

Cancer genes:

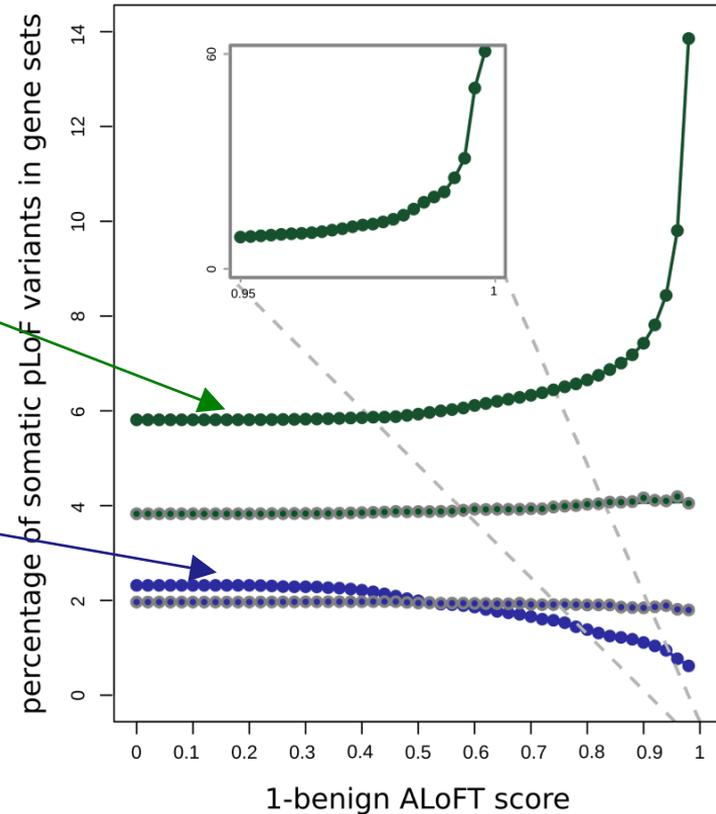
- COSMIC consensus.
- *Enriched in deleterious LoFs.*

LoF tolerant genes:

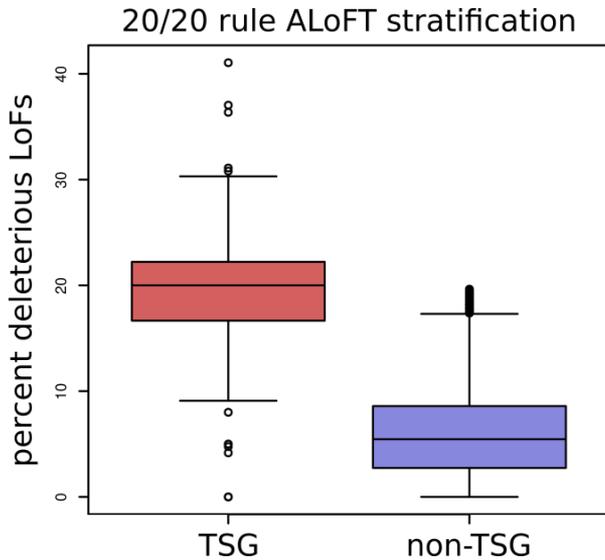
- LoF in the 1KG cohort.
- *Depleted in deleterious LoFs.*

cancer genes vs. LoF tolerant genes

- 504 cancer genes
- 387 LoF-tolerant genes
- 504 random genes
- 387 random genes



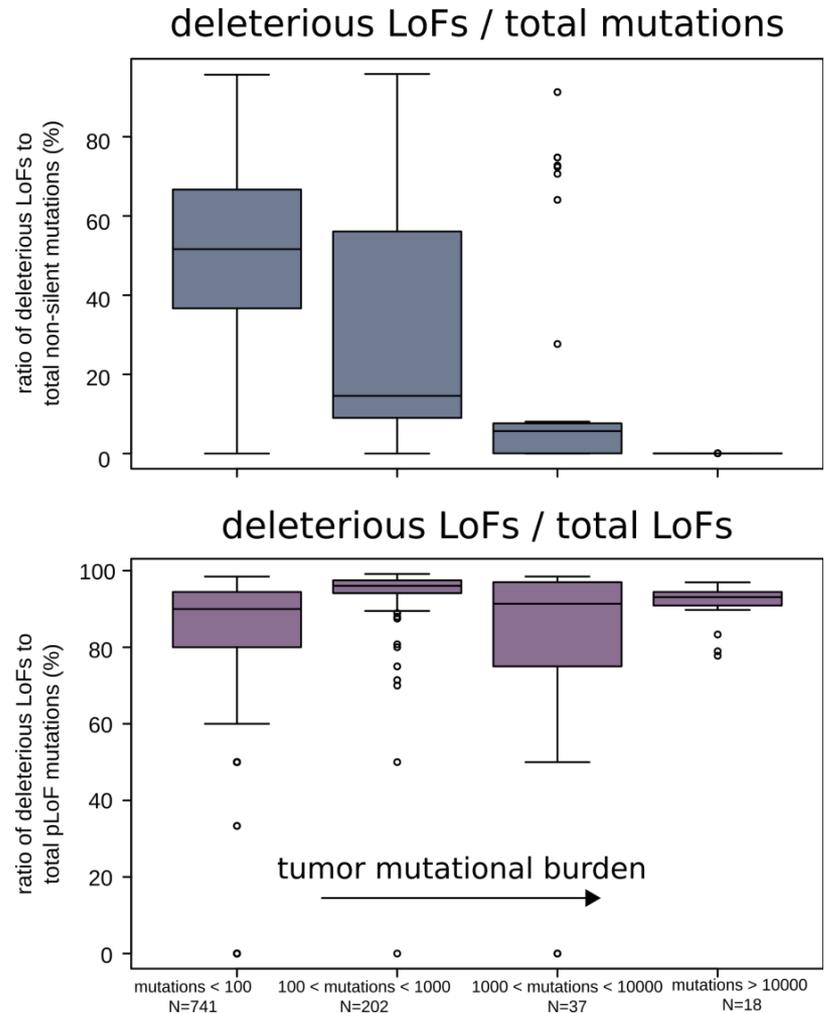
ALoFT refines cancer mutation characterization



Vogelstein *et al.* '13: if >20% of mutations in gene inactivating → tumor suppressor gene (TSG).

ALoFT further refines 20/20 rule predictions.

Balasubramanian S.*, Fu Y.* *et al.*, *NComms.*, '17



Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

- Background on rare & common variants
- Identifying cryptic allosteric sites with **STRESS**
 - On surface & in interior bottlenecks
- **Frustration** as a localized metric of SNV impact
 - Differential profiles for oncogenes v. TSGs
- **ALoFT**: Annotation of Loss-of-Function Transcripts

Acknowledgments! Also, Hiring: See **Jobs**.gersteinlab.org

STRESS.molmovdb.org

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

S Li, S Kumar,
R Chang, J Chen

ALoFT.gersteinlab.org

S **Balasubramanian**,

Y **Fu**, M Pawashe, P McGillivray,
M Jin, J Liu, K Karczewski, D MacArthur

github.com/gersteinlab/**Frustration**

S **Kumar**, D Clarke



Info about this talk

No Conflicts

Unless explicitly listed here. There are no conflicts of interest relevant to the material in this talk

General PERMISSIONS

- This Presentation is copyright Mark Gerstein, Yale University, 2017.
- Please read permissions statement at
sites.gersteinlab.org/Permissions
- Basically, feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or website link). 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 streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as `kwpotppt` , that can be easily queried from flickr, viz:
[flickr.com/photos/mbgmbg/tags/kwpotppt](https://www.flickr.com/photos/mbgmbg/tags/kwpotppt)