Computational analysis of variants: coding versus non-coding

Mark Gerstein, Yale

Slides freely downloadable from **Lectures.GersteinLab.org** & “tweetable” (via **@MarkGerstein**).

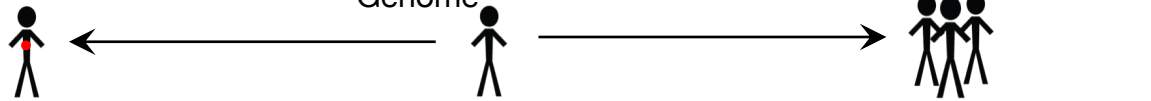
No Conflicts for this Talk. See last slide for more info.

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

A Cancer Genome

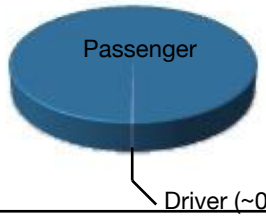
A Typical Genome

Population of 2,504 peoples



Origin of Variants

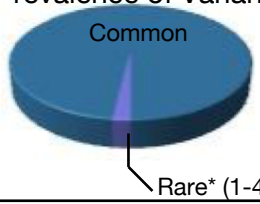
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



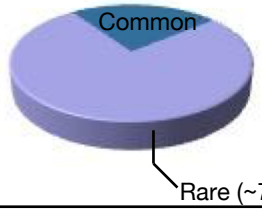
Class of Variants

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

Prevalence of Variants

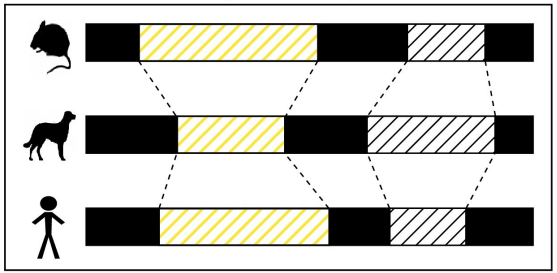


SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M

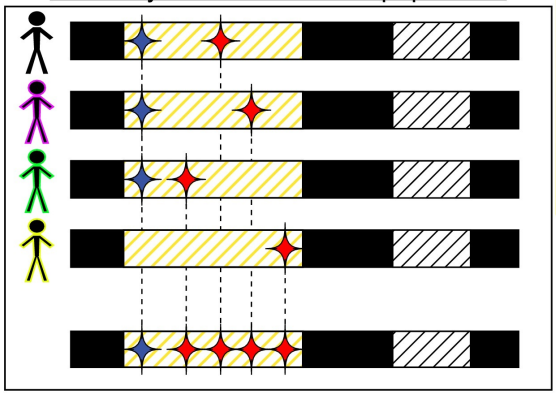


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

Similarity across species



Similarity within the human population



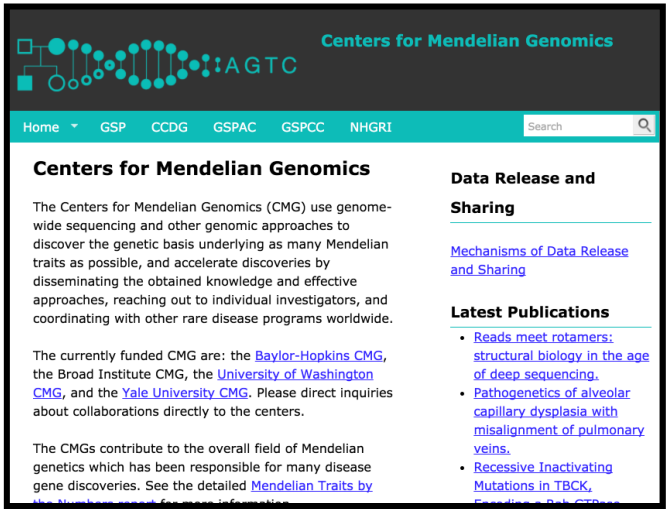
The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74

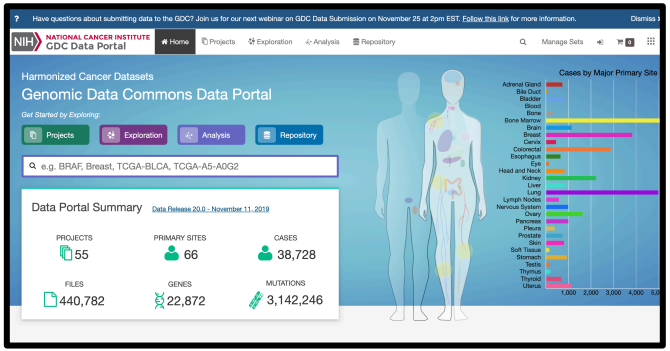
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

[Sethi et al. COSB ('15)]

Rare (as opposed to common) variant analysis particularly applicable at the moment

- CMG rare-disease & TCGA somatic variants
  - Main NIH disease genomics projects
  - Both focus on ”rare” variant for which GWAS is not meaningful
  - For purpose of assessing impact & aggregating “burden” across a gene, somatic & germline rare can be considered similar – no notion of LD
  - Differences:
    drivers under “cell-level” positive selection while
    Mendelian variants are under strong negative selection from an organismal perspective





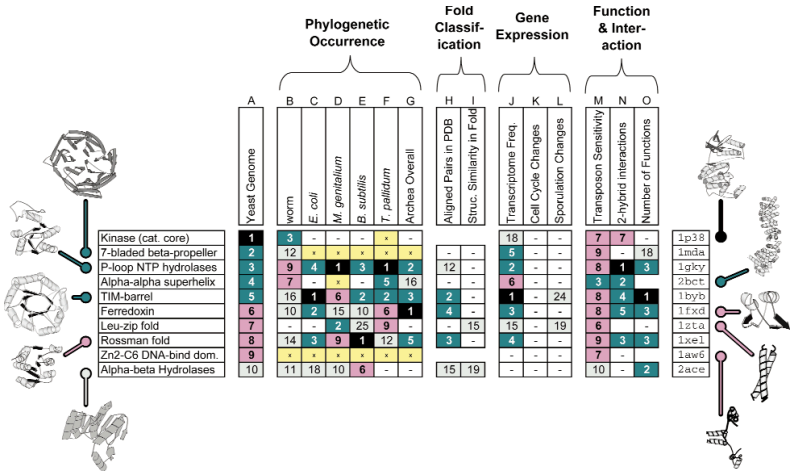
Coding v Non-coding

- Coding
  - Easily interpretable, particularly related to structure
  - Available in large quantities
  - Exomes have the current potential for great scale (Scale of EXAC, >60K exomes [Lek et al. ‘16])
- Non-coding
  - Not as interpretable & hard to connect to genes
- “Near coding”
  - Bits of non-coding, close to genes & readily linked to them
  - EX: Splice sites, promotors, uORF

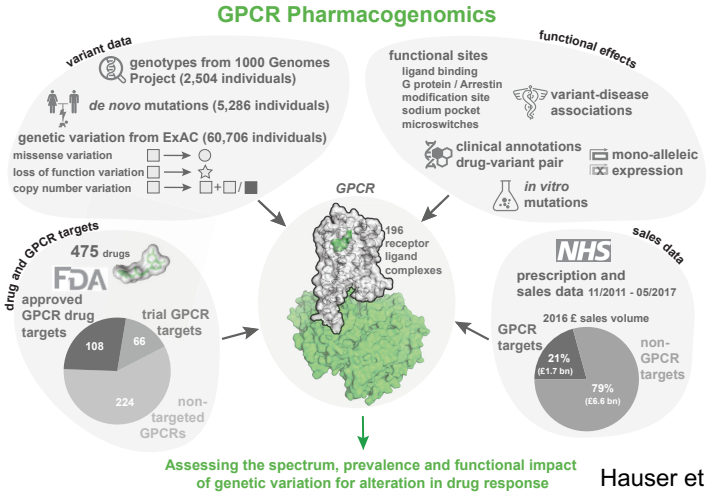
Structure & genomics

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.



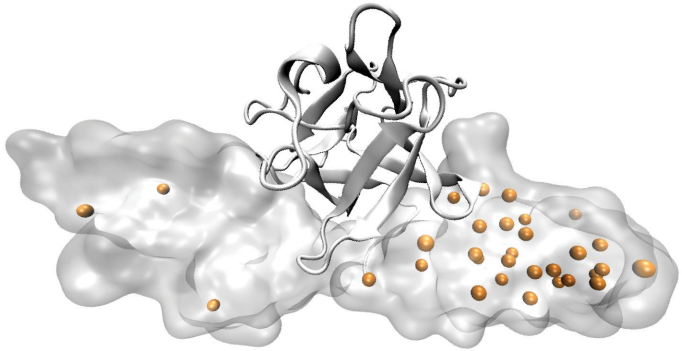
Gerstein et al., Nat. Struct. Bio. 2000



Hauser et al., Cell 2018

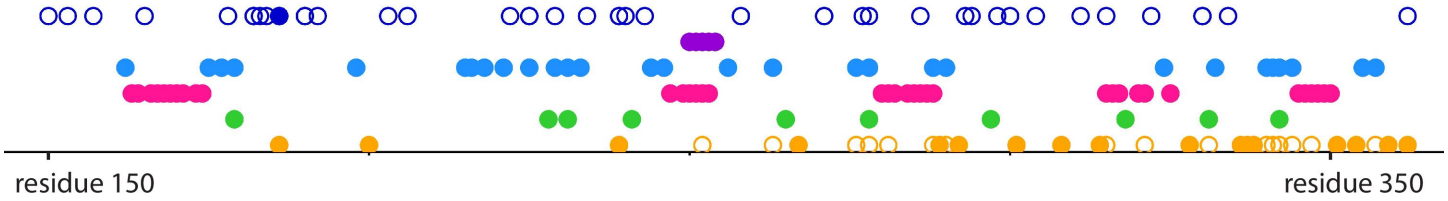
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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



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)]

Computational analysis of variants: coding versus non-coding

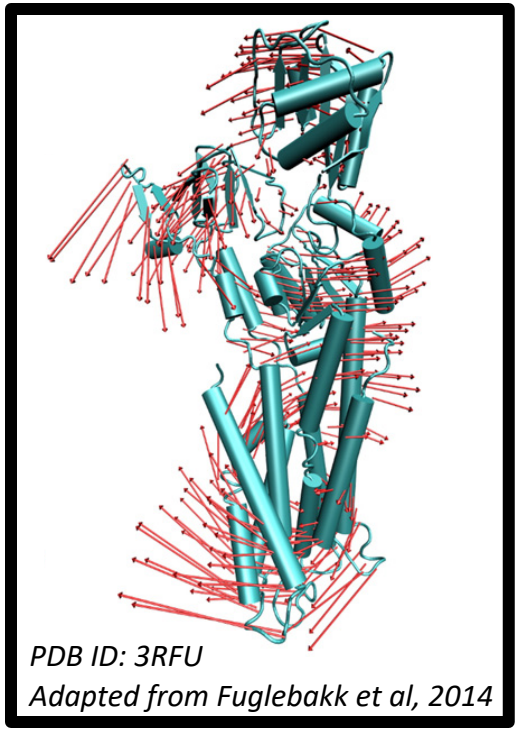
- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **Identifying cryptic allosteric sites with STRESS**
  - On surface & in interior bottlenecks
- **Frustration as a localized metric of SNV impact**
  - Differential profiles for oncogenes v. TSGs
- **ALoFT: Annotation of LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

Computational analysis of variants: coding versus non-coding

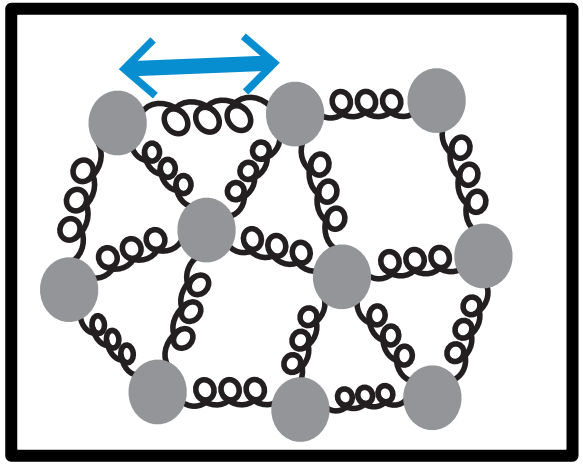
- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **Identifying cryptic allosteric sites with STRESS**
  - On surface & in interior bottlenecks
- **Frustration as a localized metric of SNV impact**
  - Differential profiles for oncogenes v. TSGs
- **ALoFT: Annotation of LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



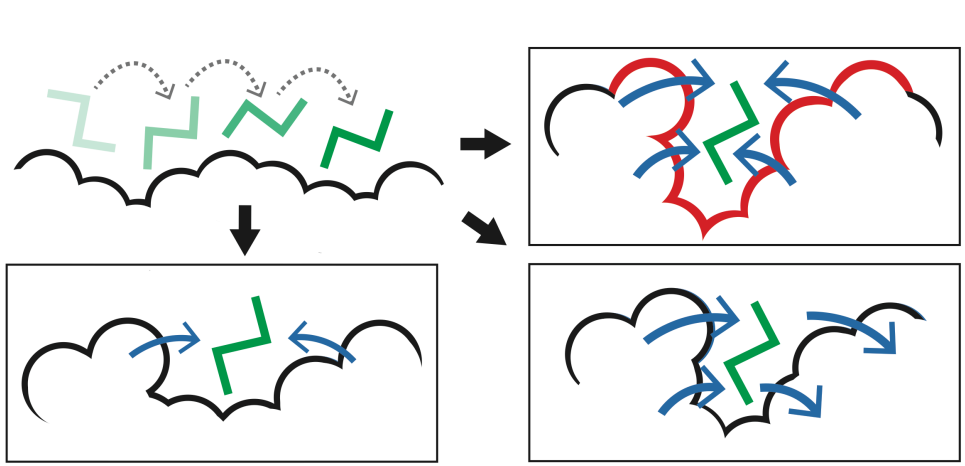
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



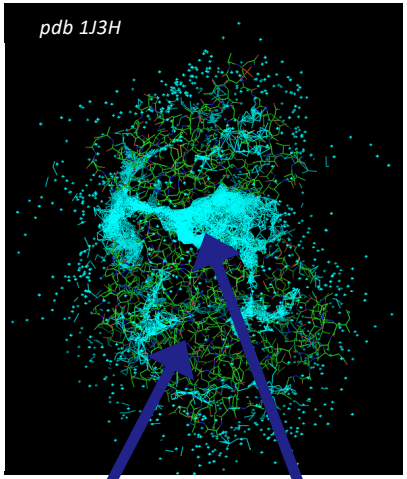
Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion

Predicting Allosterically-Important Residues at the Surface

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

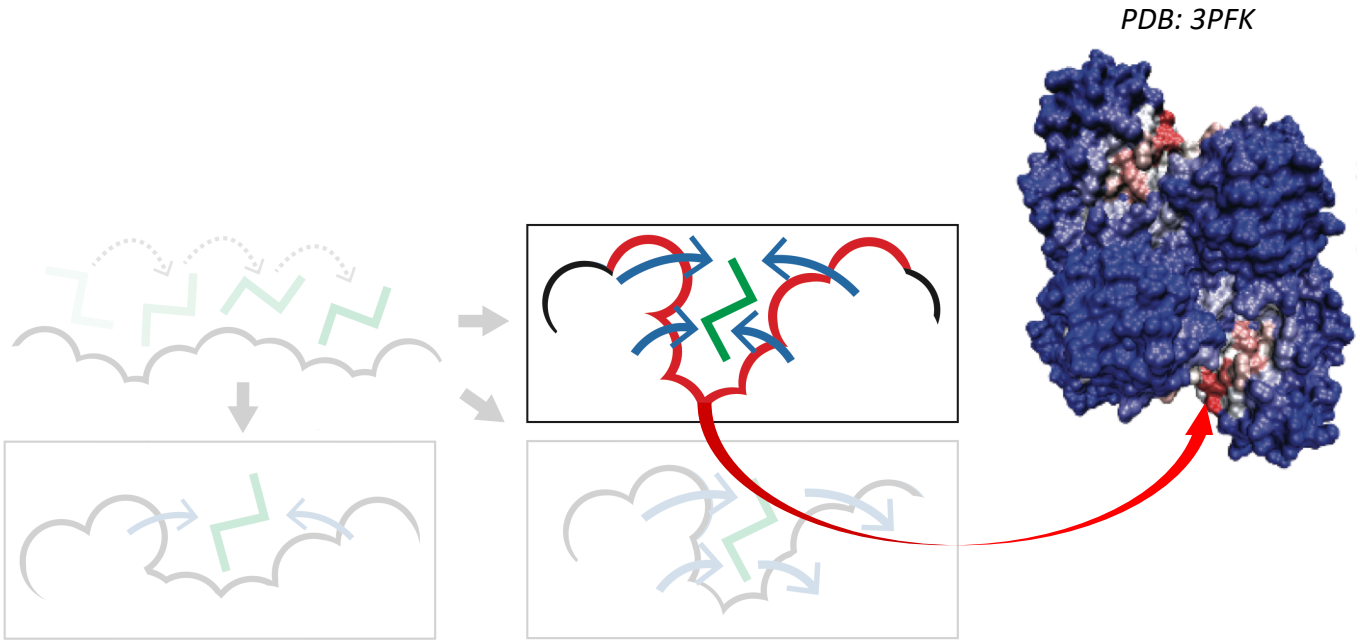


$$\textit{binding leverage} = \sum_{m=1}^{10}\left(\sum_{i}\sum_{j}\Delta d_{ij(m)}^{2}\right)$$



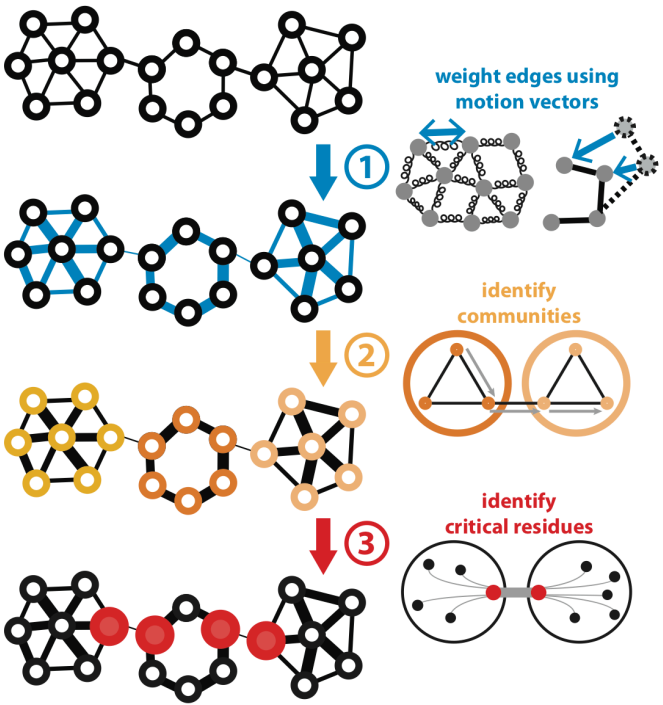
Adapted from Clarke*, Sethi*, et al *('16)*

Predicting Allosterically-Important Residues at the Surface



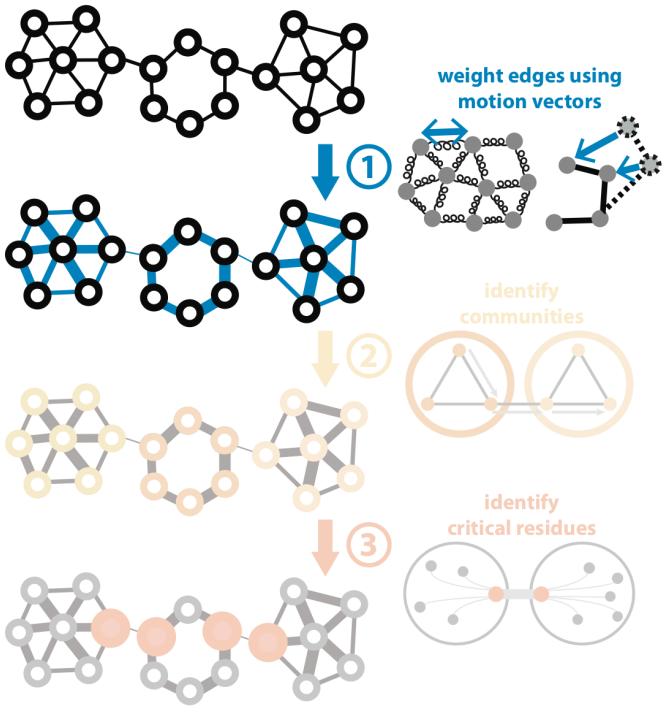
Adapted from Clarke*, Sethi*, et al *('16)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



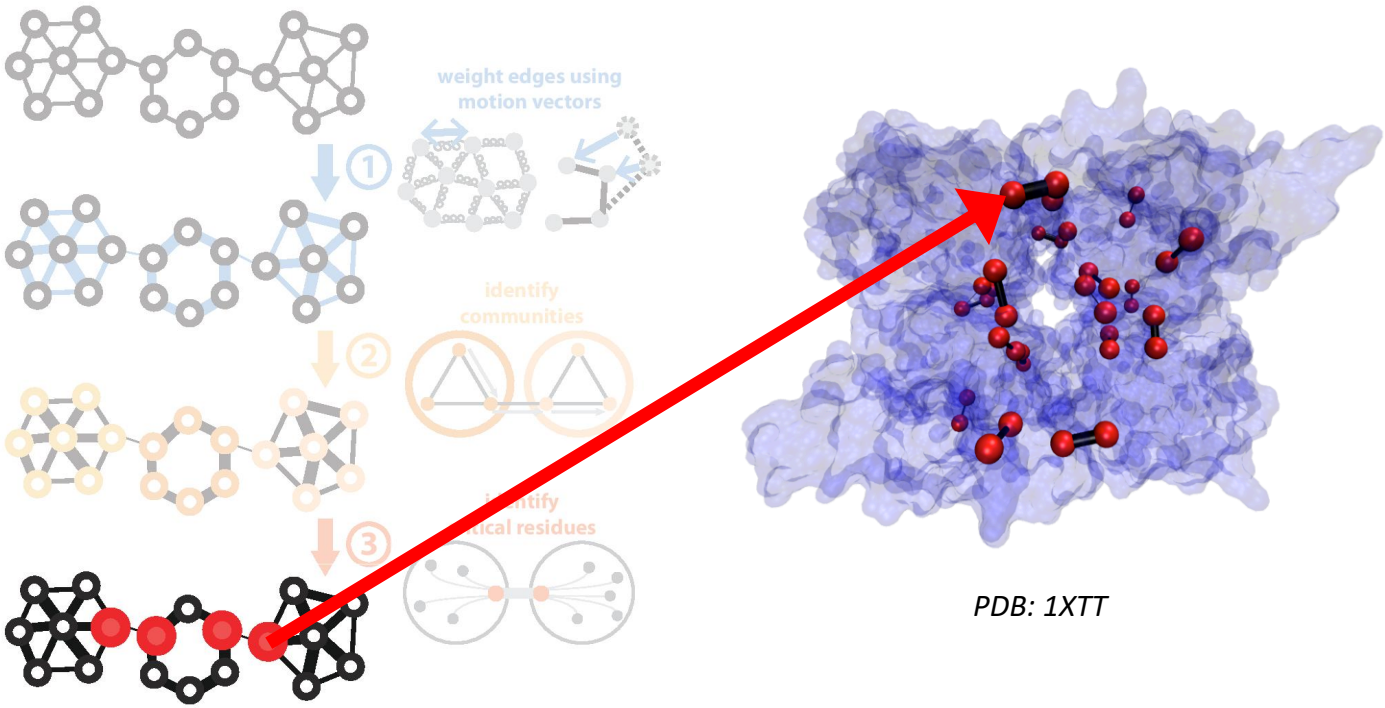
$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

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

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

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

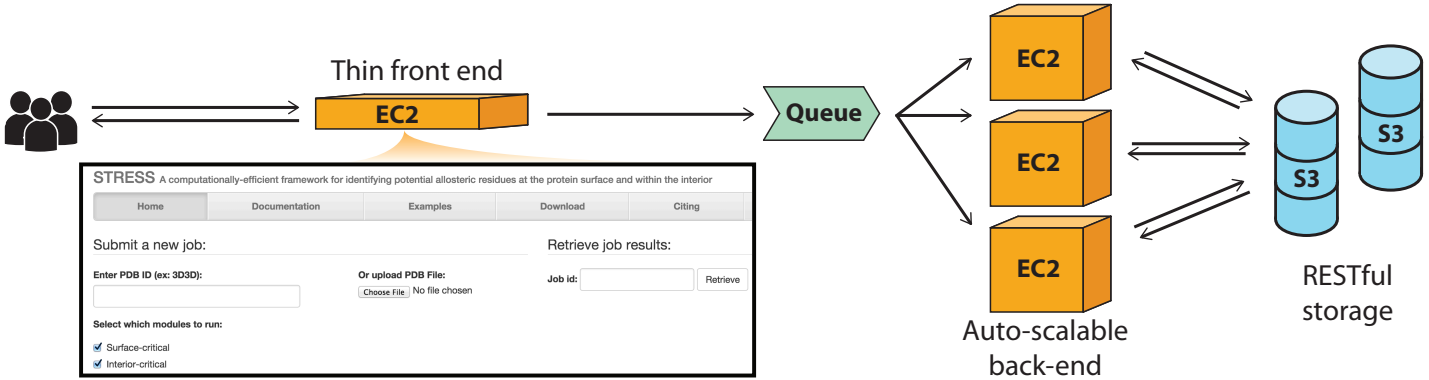
Predicting Allosterically-Important Residues within the Interior



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

STRESS Server Architecture: Highlights

stress.molmovdb.org



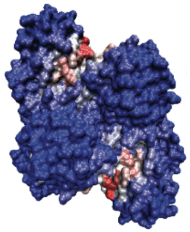
- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are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

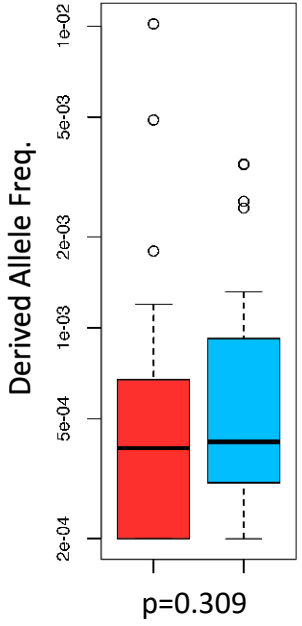
Adapted from Clarke*, Sethi*, et al *('16)*

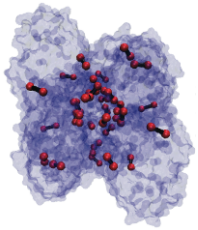
Intra-species conservation of predicted allosteric residues

1000 Genomes

Surface



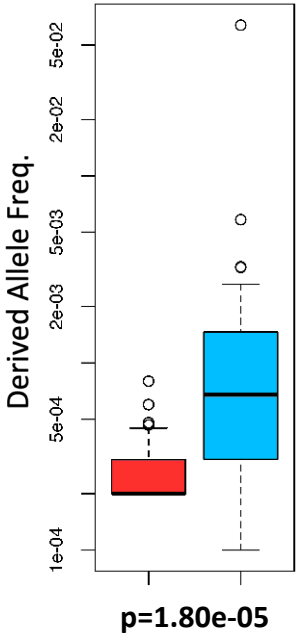




critical

non-critical

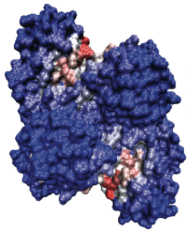
Interior



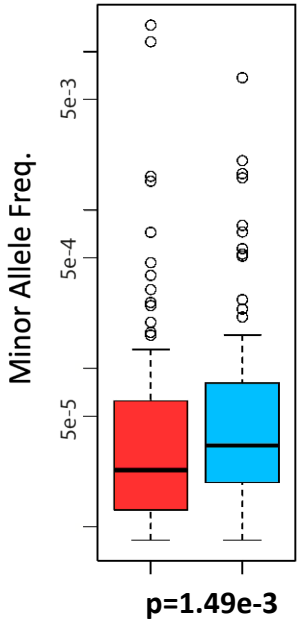
Adapted from Clarke*, Sethi*, et al *('16)*

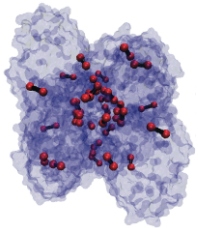
Intra-species conservation of predicted allosteric residues

ExAC



Surface

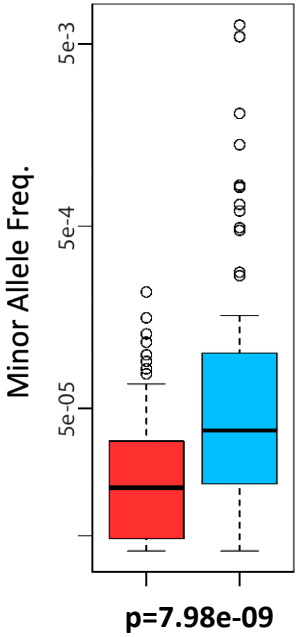




critical

non-critical

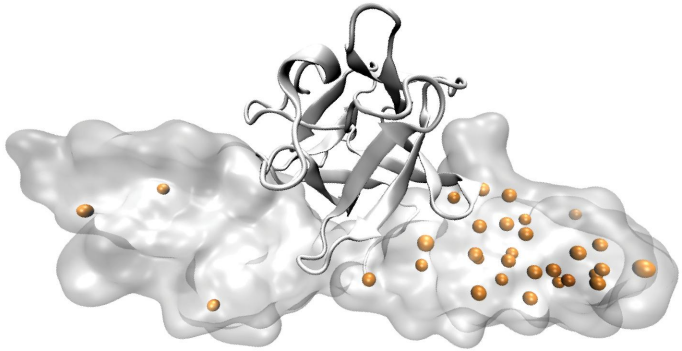
Interior



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

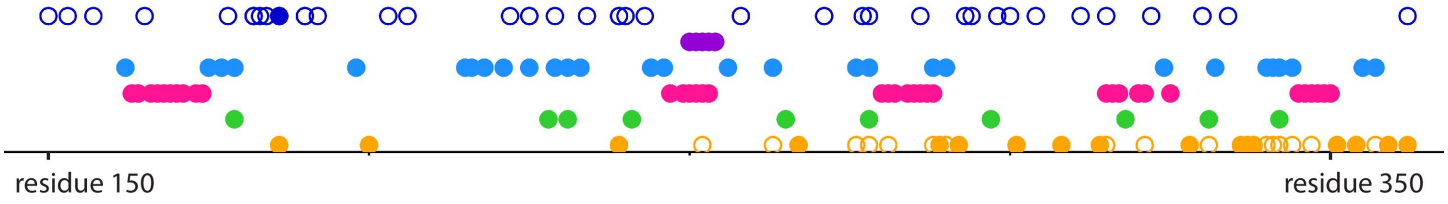
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

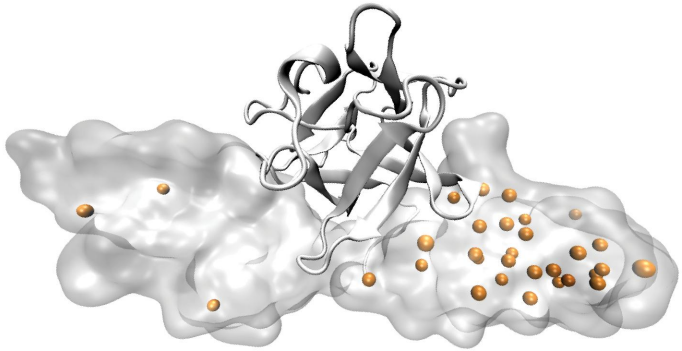
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

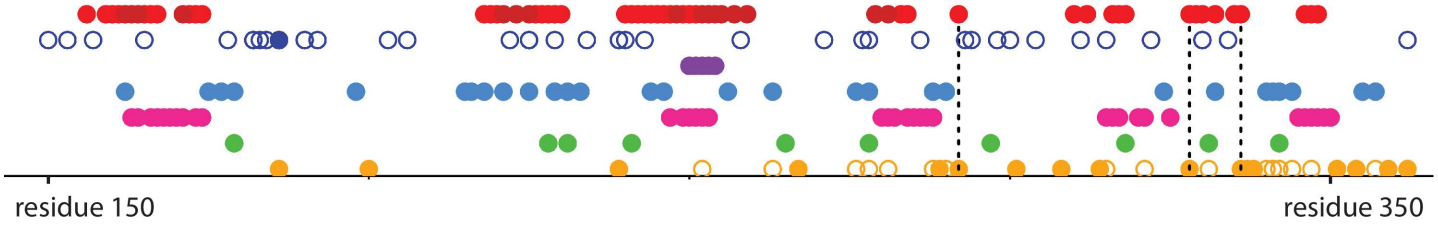
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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



Fibroblast growth factor receptor 2 (pdb: 1IIL)

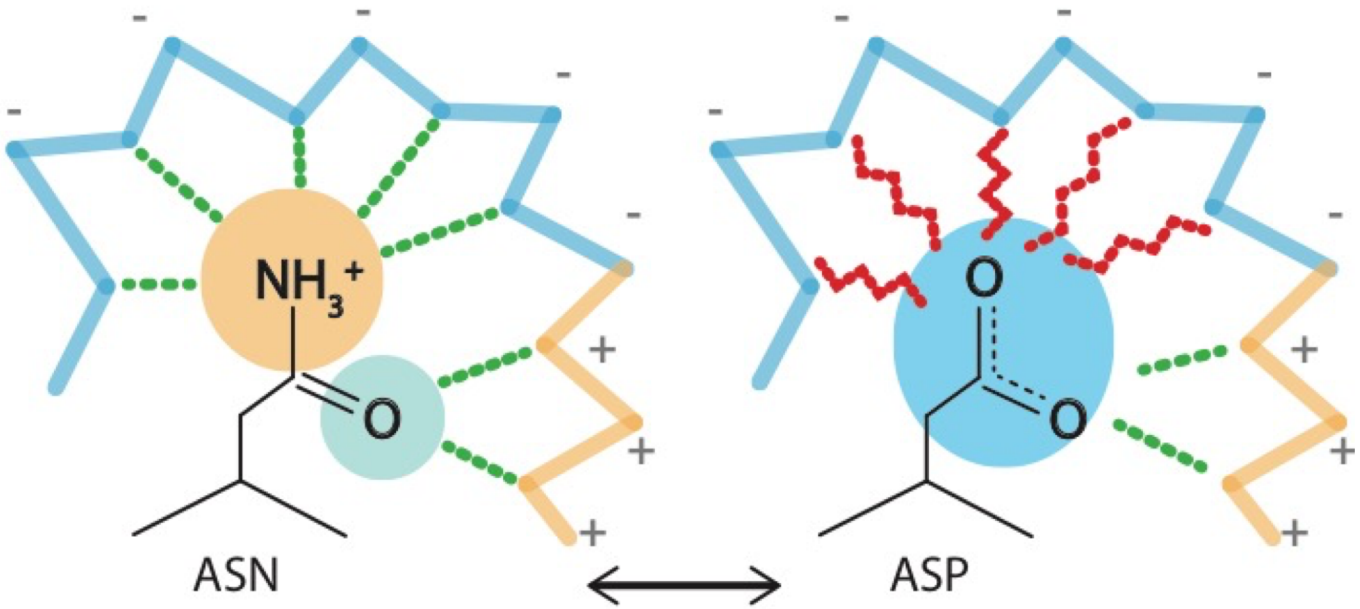
- Predicted allosteric (surface | interior)
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



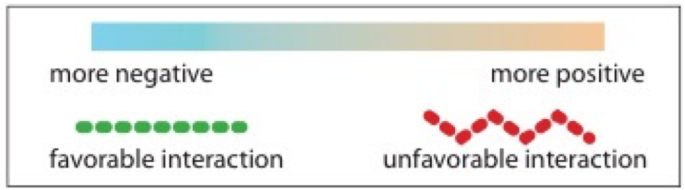
[Sethi et al. COSB ('15)]

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

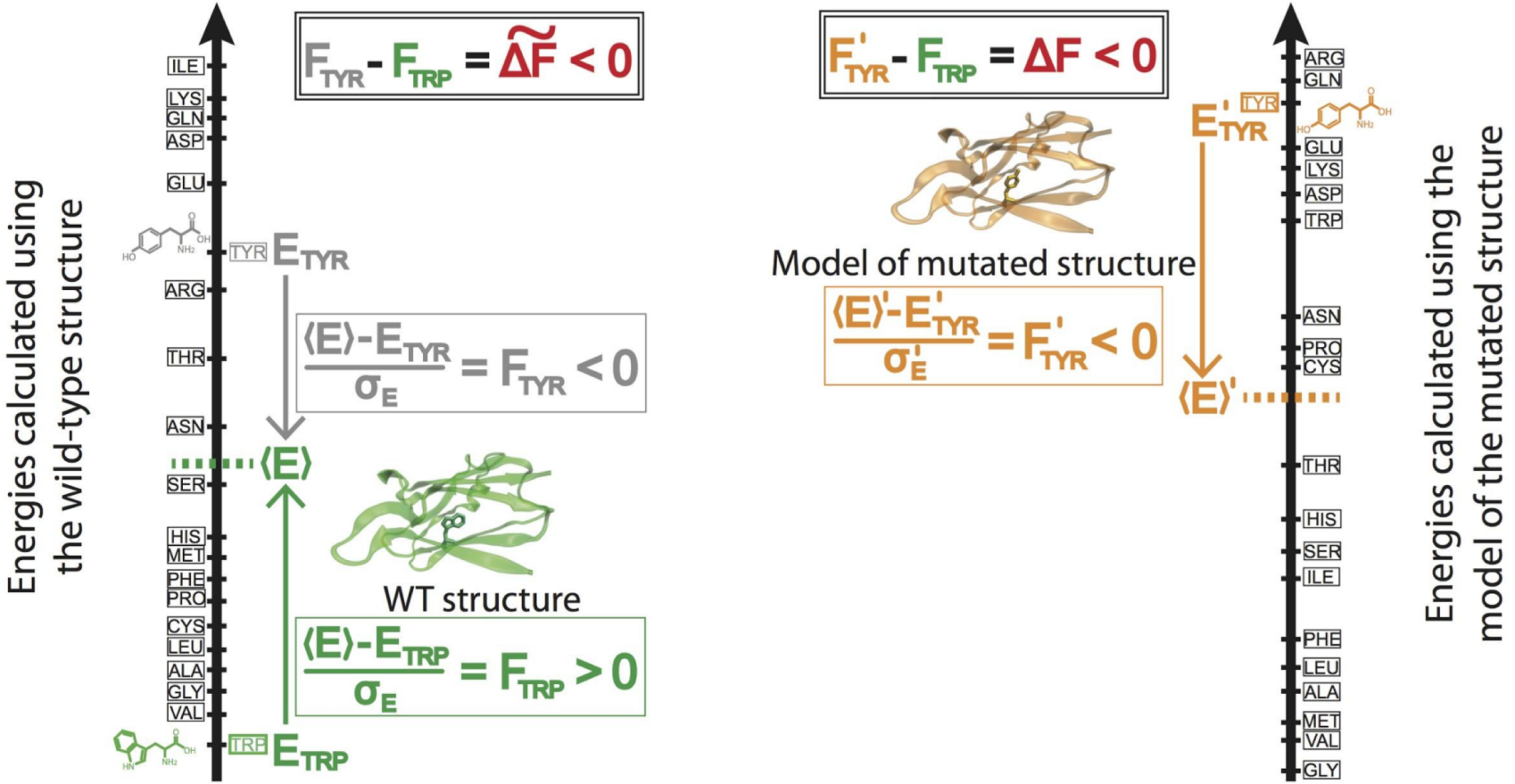


What is localized frustration ?



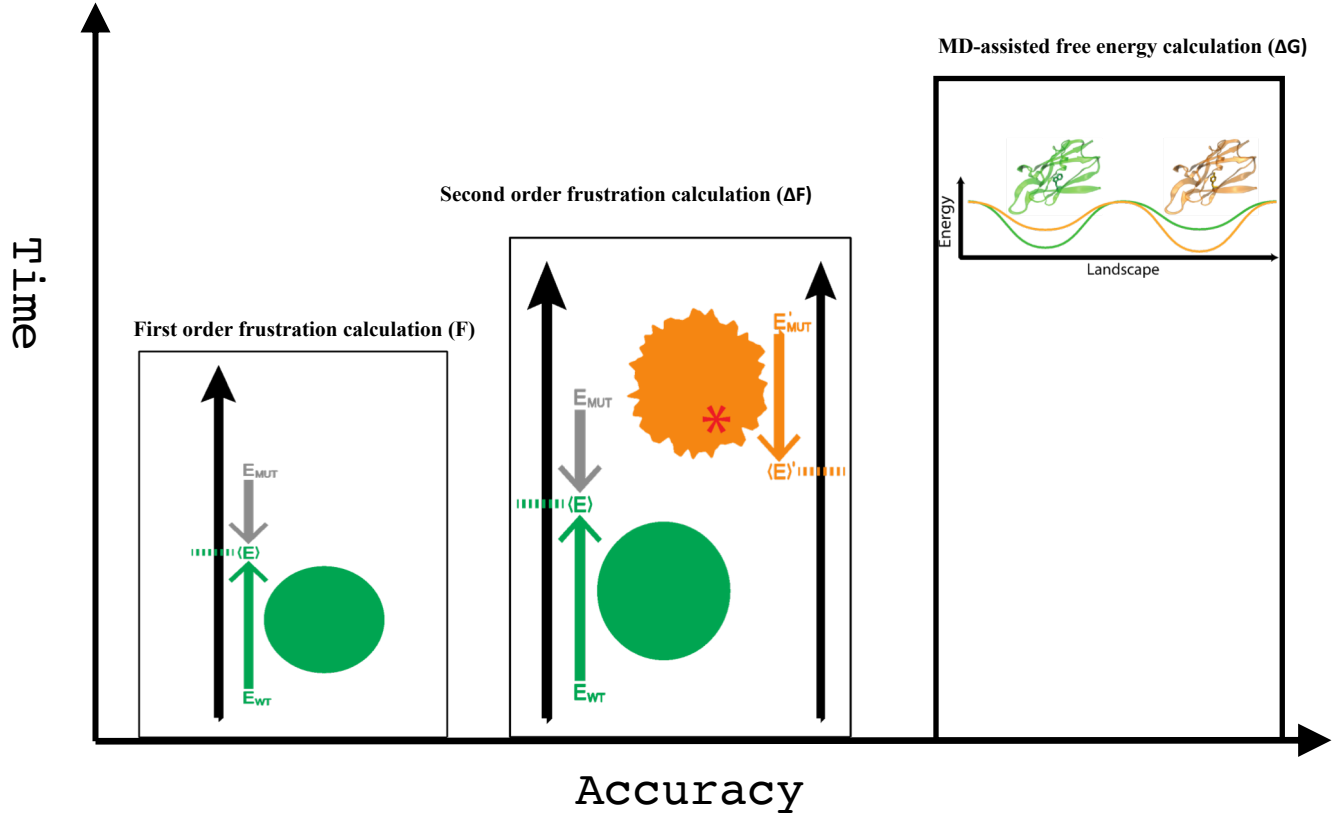
[Ferreiro et al., *PNAS* ('07)]

Workflow for evaluating localized frustration changes (ΔF)

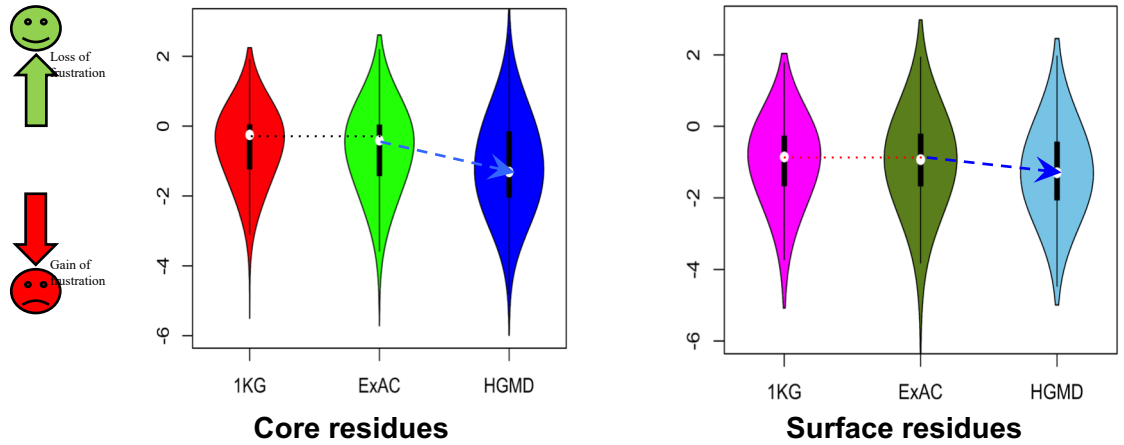


[Kumar et al. *NAR* (2016)]

Complexity of the second order frustration calculation



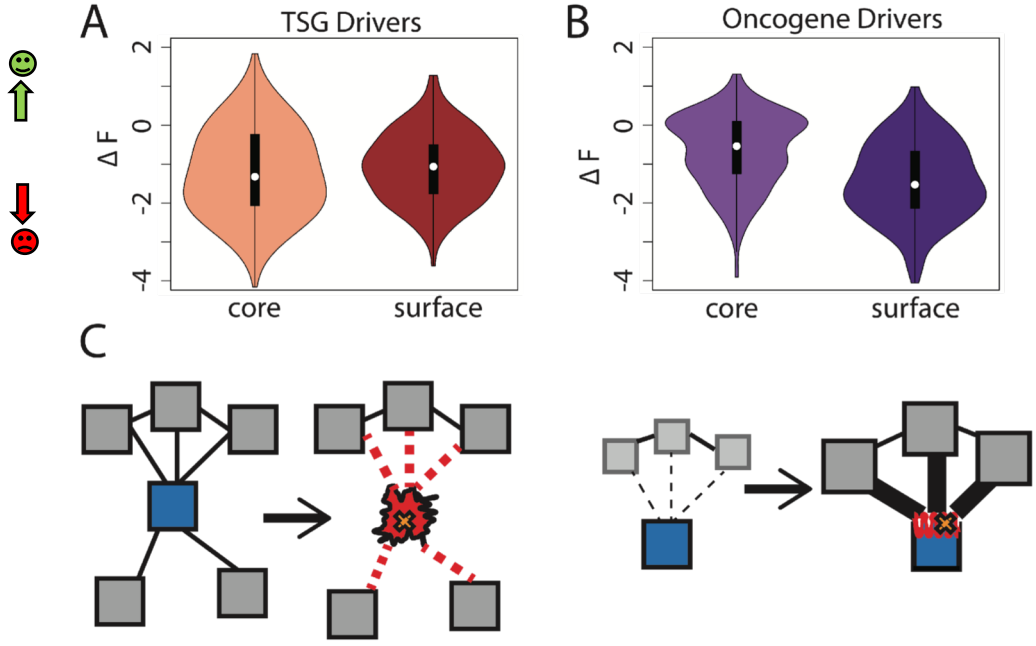
Comparing ΔF values across different SNV categories: disease v normal



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

[Kumar et al, *NAR* (2016)]

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* (2016)]

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

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

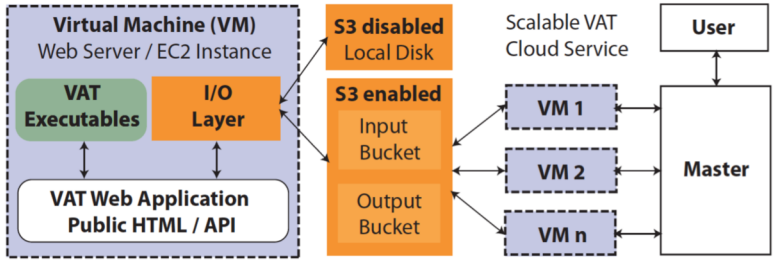
VCF Input

Output:

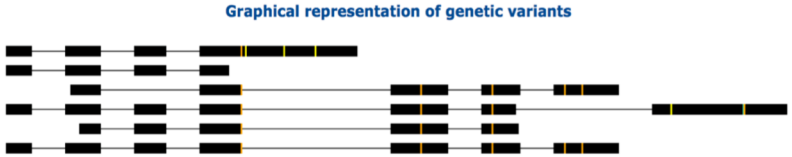
- Annotated VCFs
- Graphical representations of functional impact on transcripts

Access:

- Webserver
- AWS cloud instance
- Source freely available



CLOUD APPLICATION



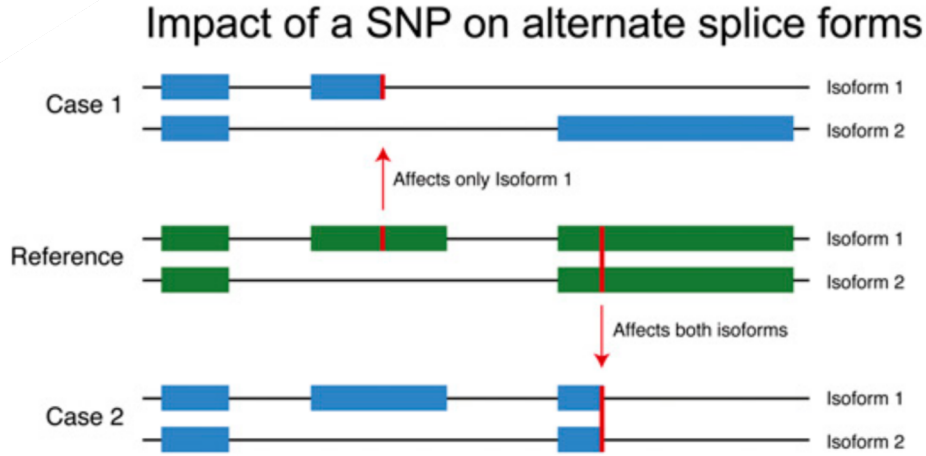
vat.gersteinlab.org

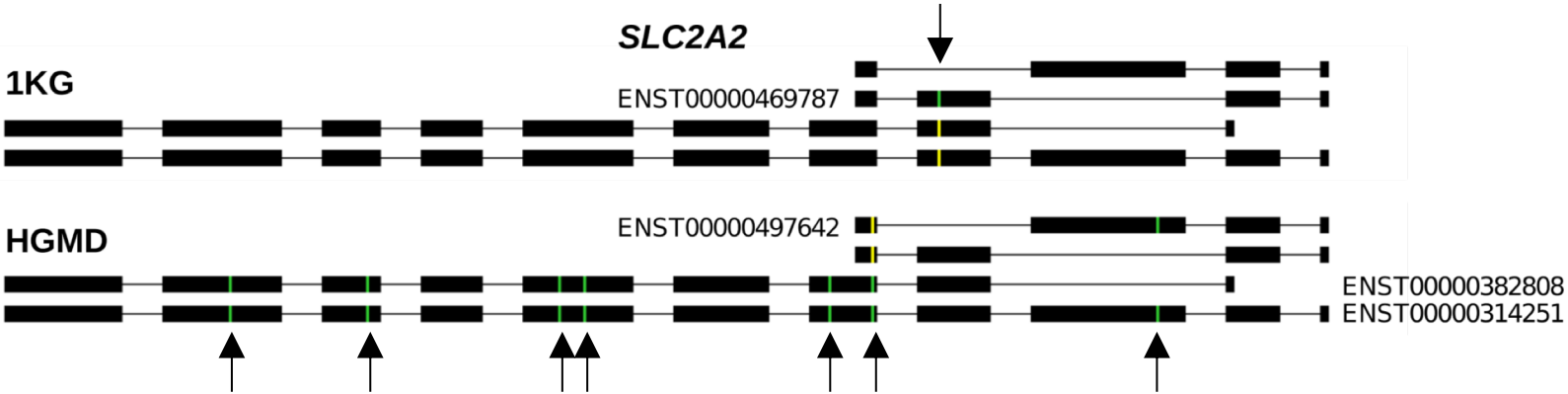
Habegger L., Balasubramanian S.*, et al. Bioinformatics, 2012*

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



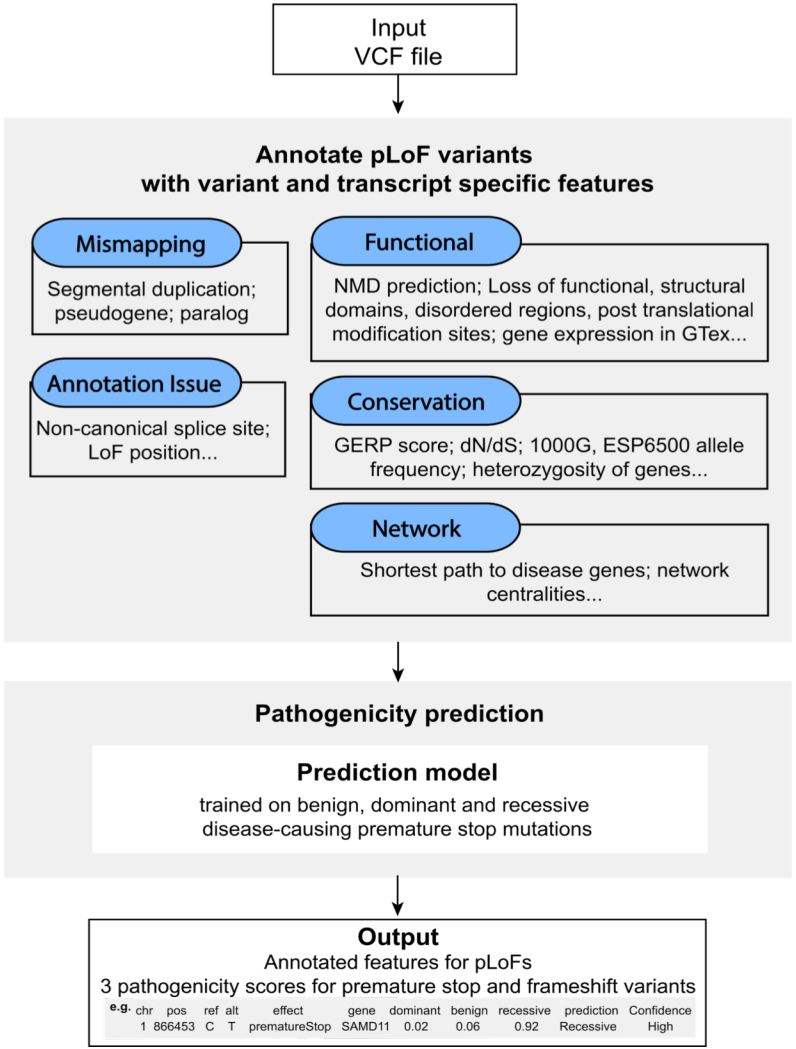


Annotation of Loss-of-Function Transcripts (ALoFT)

Runs on top of VAT

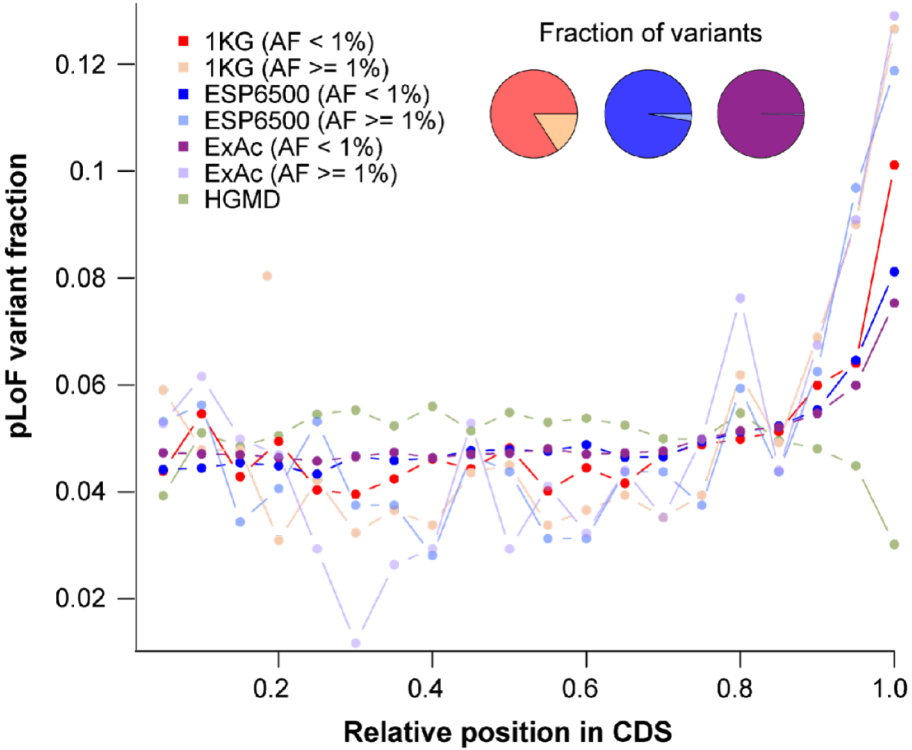
Output:

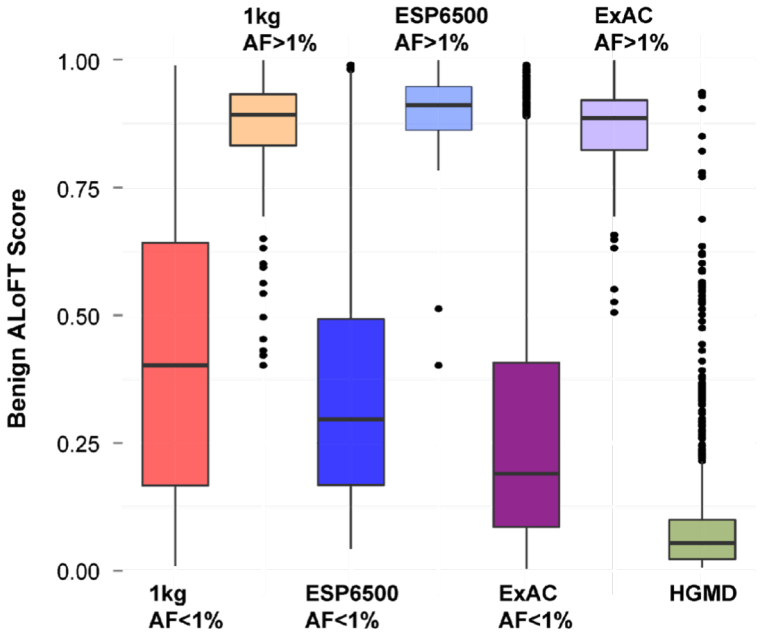
- Impact score: benign or deleterious.
- Decorated VCF.



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

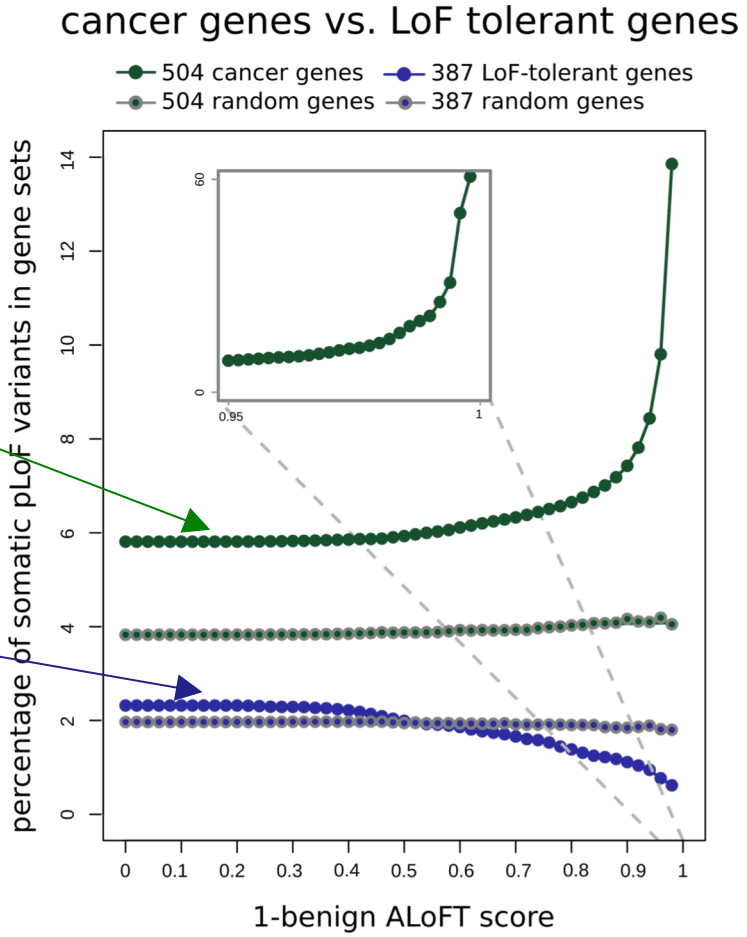
ALoFT identifies deleterious somatic LoF variants

Cancer genes:

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

LoF tolerant genes:

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

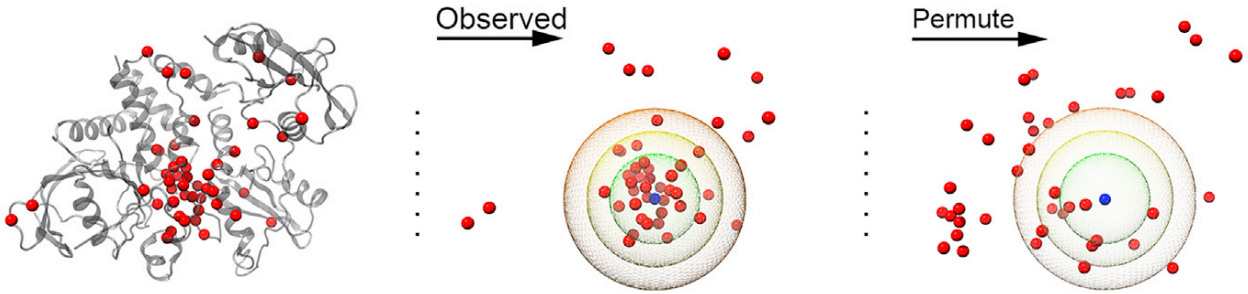


Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

Structures have been used successfully to “aggregate” the burden of mutations



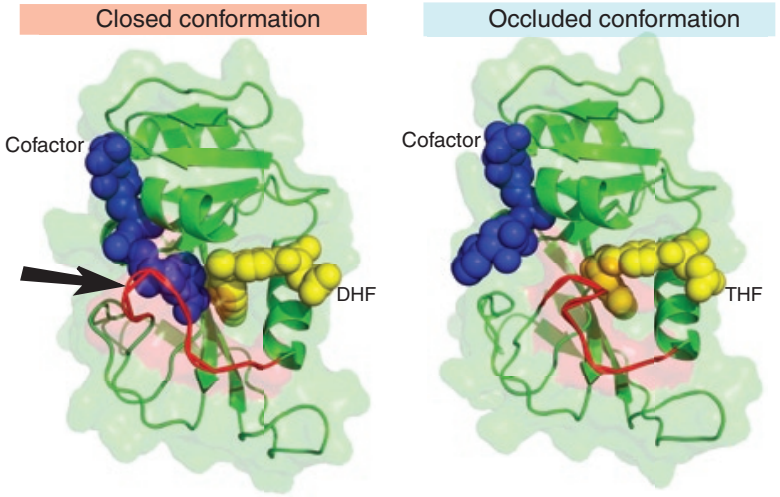
Sivley , et al., AJHG. (2018)

These approaches search for mutational clusters on protein structure using distance cutoff.

Permutation is performed to identify statistically significant mutational clusters on static protein structure.

Both rare germline & somatic

Protein dynamics is important for protein function



Proteins are inherently dynamic bio-molecules and sample large ensembles of conformations.

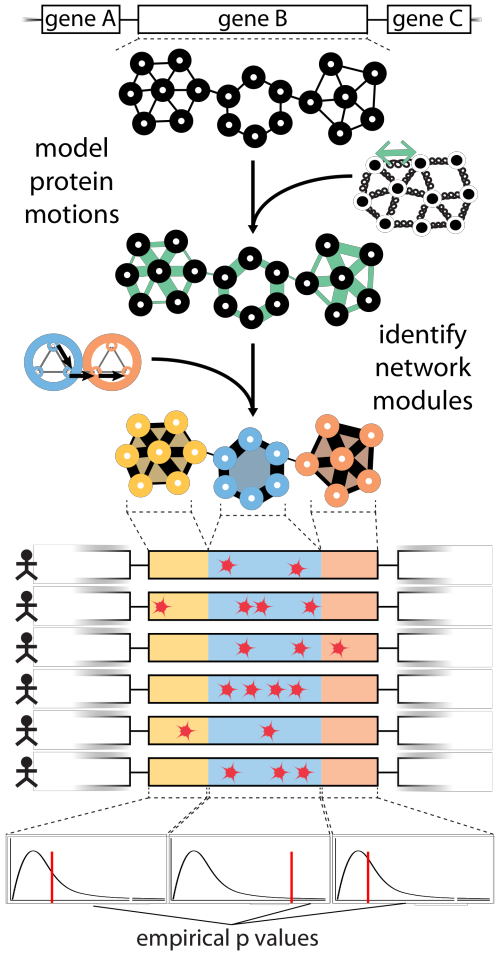
prior structure-based methods are potentially less sensitive to identify functional residues through the mutation clustering approach.

Potentially miss many critical mutational clusters

We leverage protein dynamics to identify mutation clusters

Focus on data from from TCGA pancan atlas data

Workflow to aggregate mutations taking into account dynamics



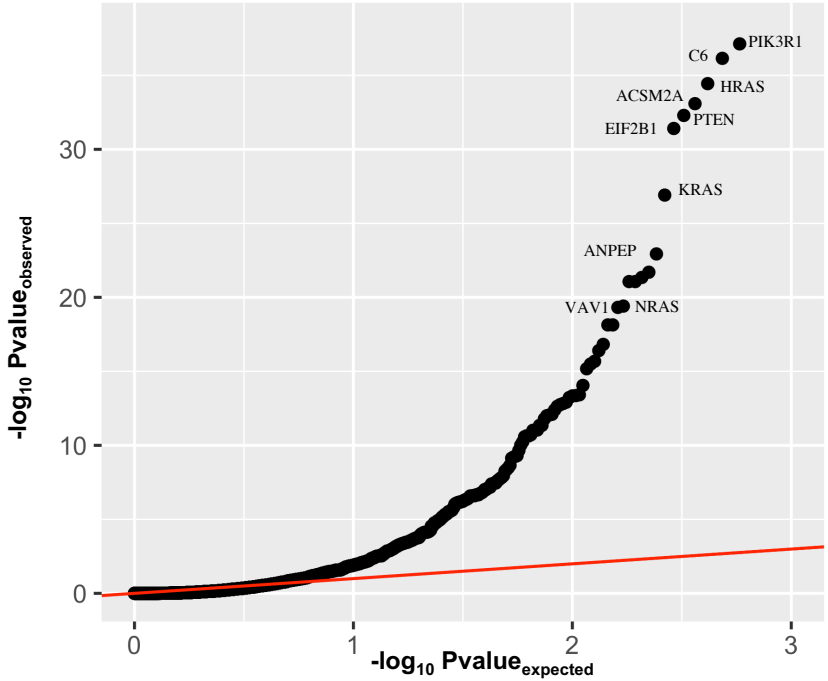
Our framework leverages large-scale conformational changes of a protein to identify dynamic sub-regions of proteins (or “communities”).

we mapped missense mutations
onto three-dimensional protein structures.

For each community with mapped mutations, we performed a Fisher exact test to determine whether variants fall within a given community is more frequently observed than what would be expected by chance.

[Kumar et al. (‘19). PNAS]

Pancancer Q-Q plot for genes with hotspot communities

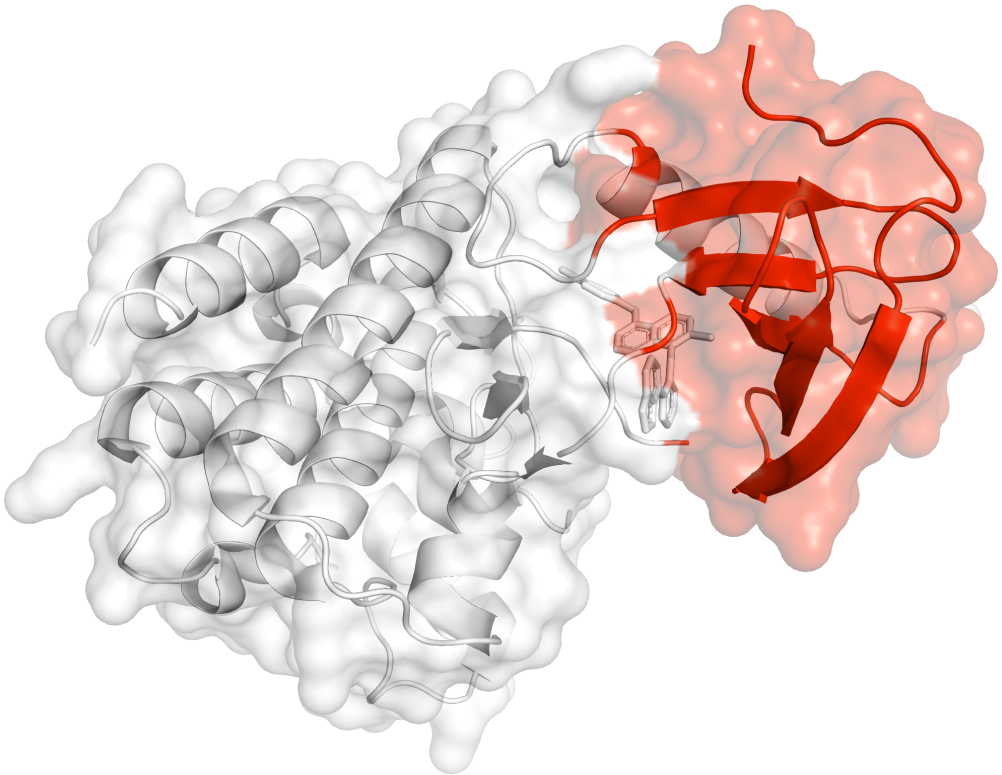


Our pan-cancer analysis identifies hotspot communities present on protein structures of 434 putative driver genes.

Our workflow identifies well known driver genes as well as novel putative driver genes.

[Kumar et al. (‘19). PNAS]

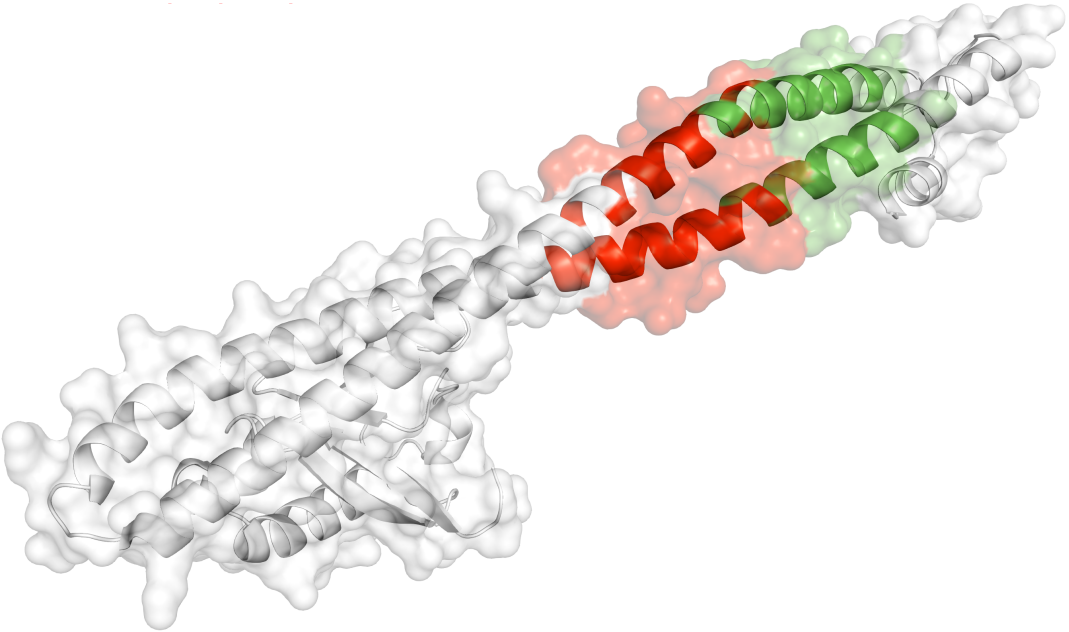
Example of oncogene with hotspot communities



BRAF: We identify one hotspot community comprising of 52 residues on the co-crystal structure of the BRafV600E kinase domain .

[Kumar et al. (‘19). PNAS]

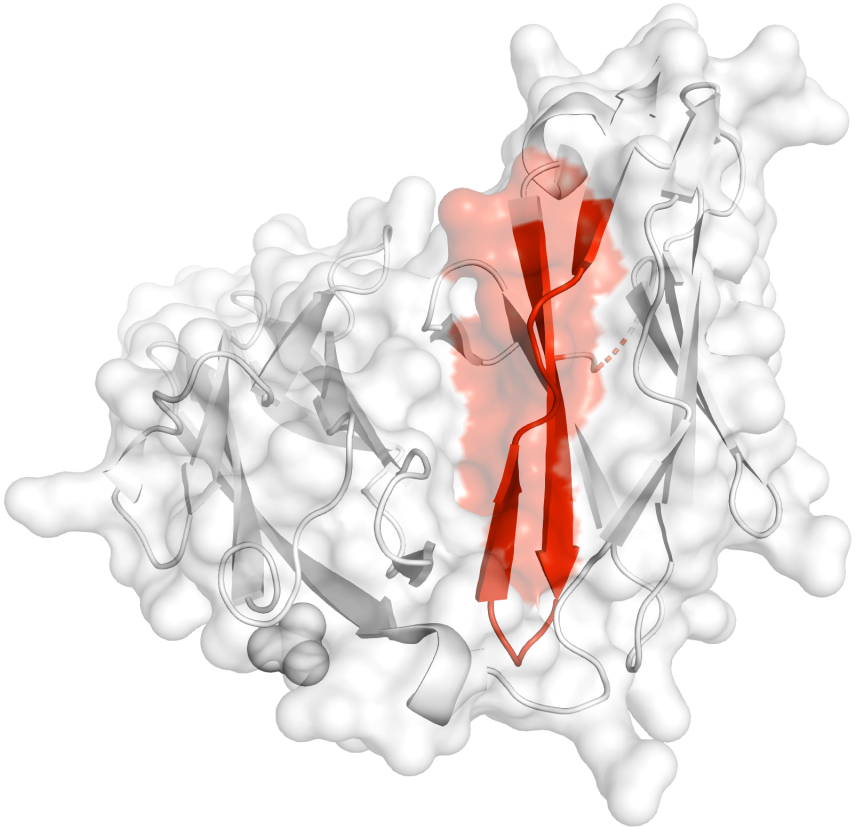
Example of TSG with hotspot communities



We identify two hotspot communities adjacent to each other on the co-crystal structure of the PIK3R1 gene.

[Kumar et al. (‘19). PNAS]

Example of novel drivers with hotspot communities



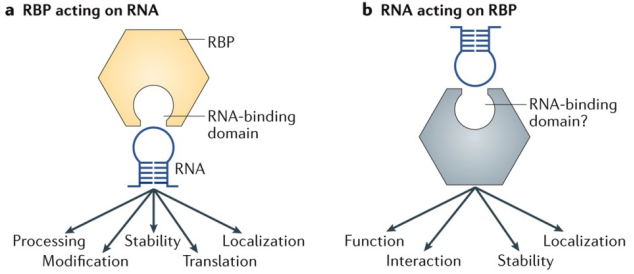
Our workflow predicts one hotspot community that comprise of 47 residues in the crystal structure of PTPRD gene.

[Kumar et al. (‘19). PNAS]

Computational analysis of variants: coding versus non-coding

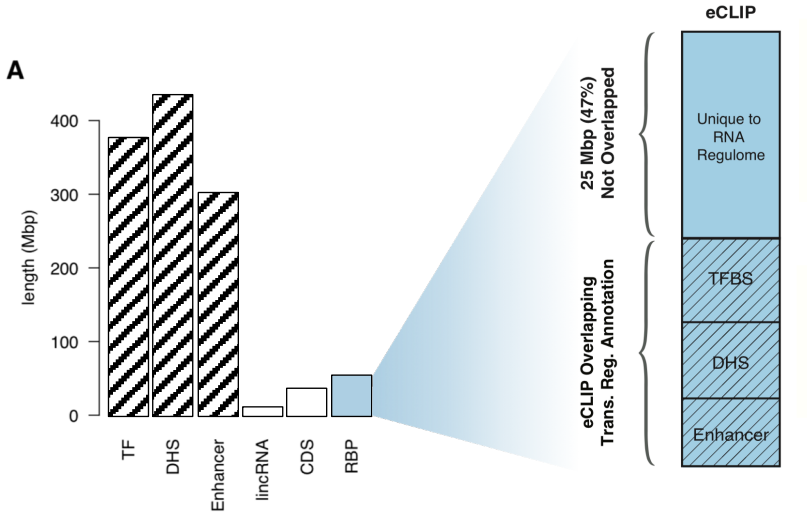
- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **Identifying cryptic allosteric sites with STRESS**
  - On surface & in interior bottlenecks
- **Frustration as a localized metric of SNV impact**
  - Differential profiles for oncogenes v. TSGs
- **ALoFT: Annotation of LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

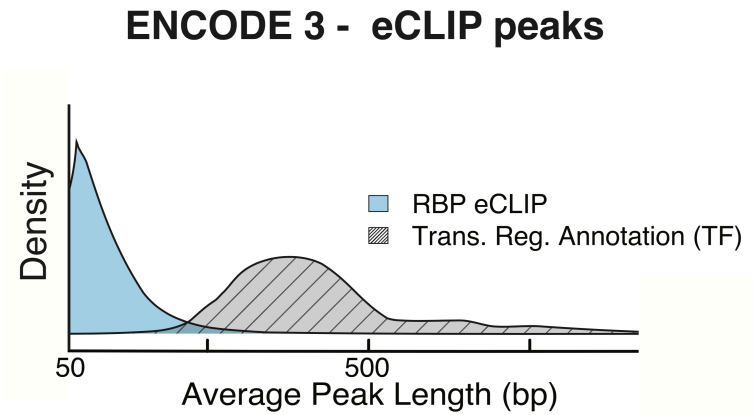
RNA Binding Proteins (RBPs)



Nat Rev Mol Cell Biol. 2018 May;19(5):327-341. doi: 10.1038/nrm.2017.130. Epub 2018 Jan 17.

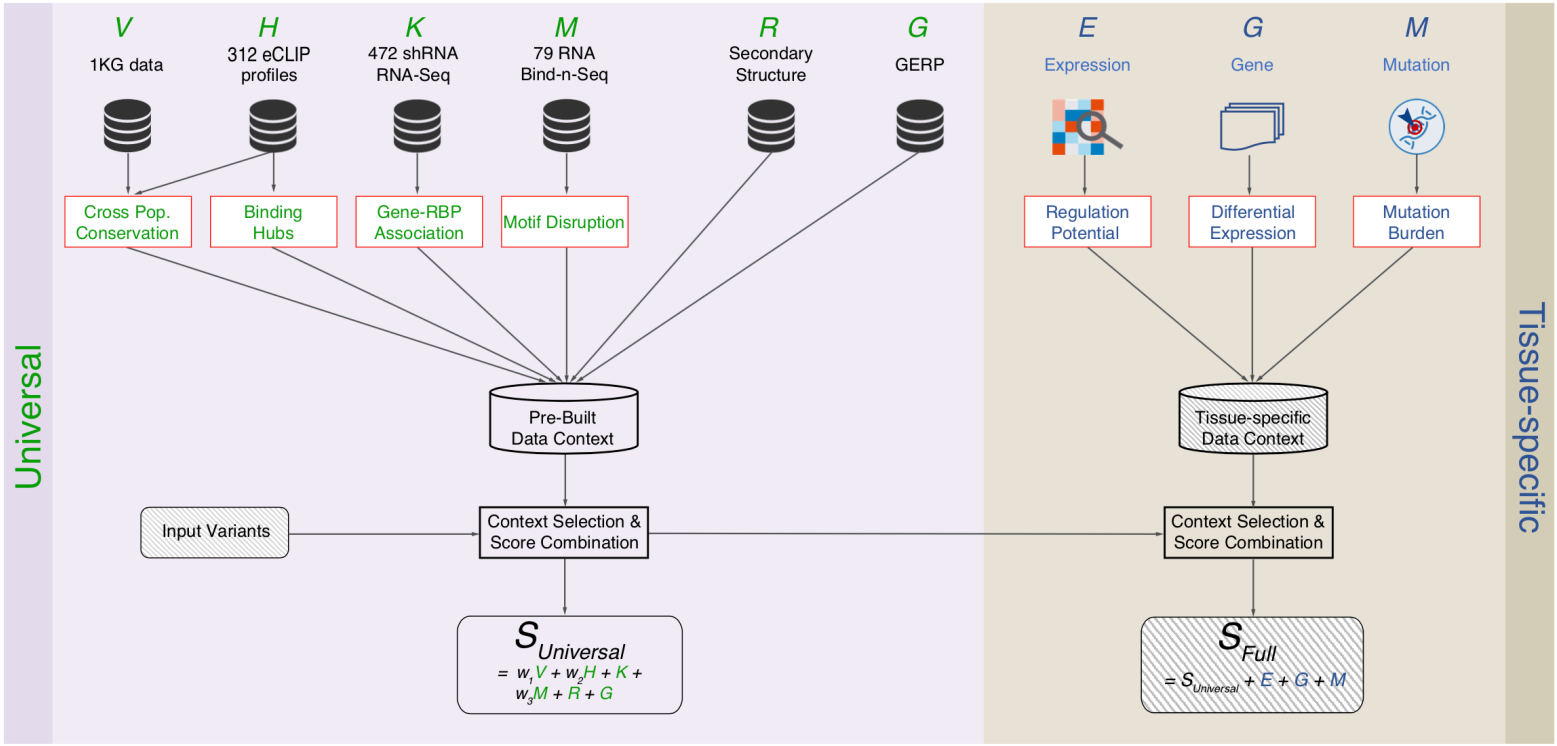
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

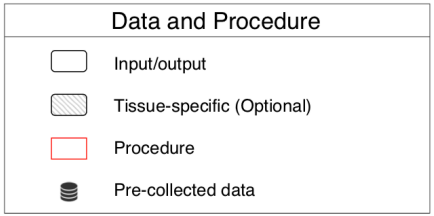


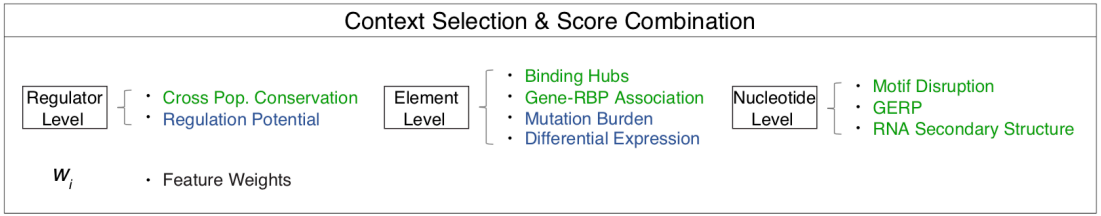


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

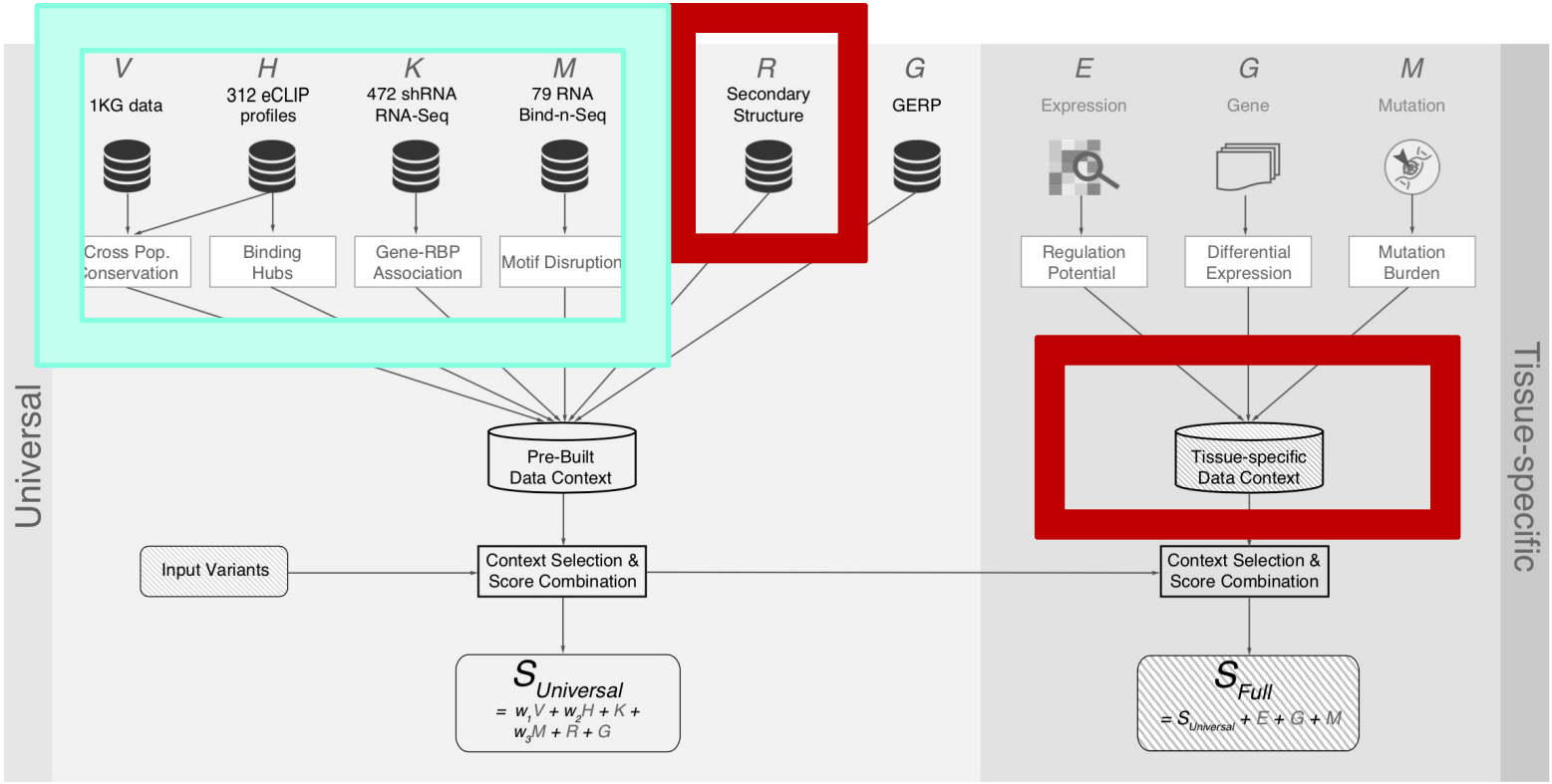
Schematic of RADAR Scoring

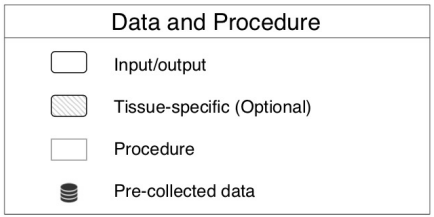


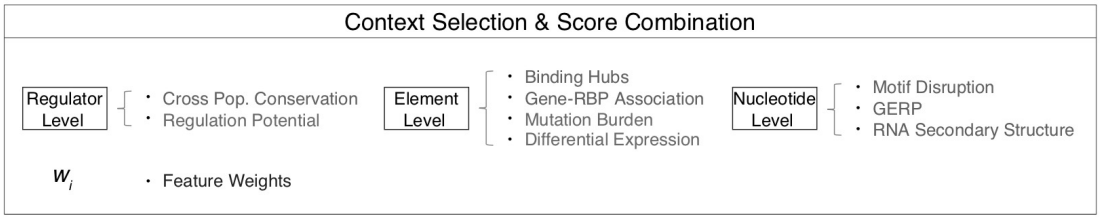




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

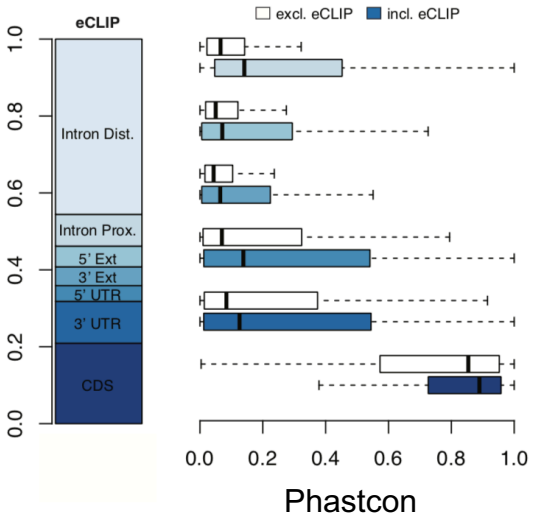




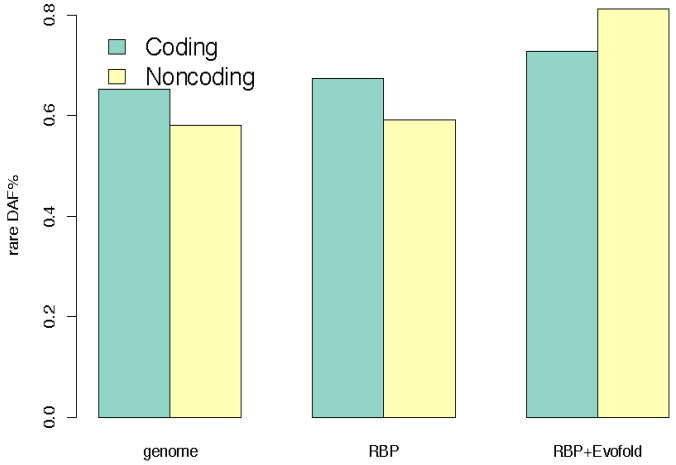


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

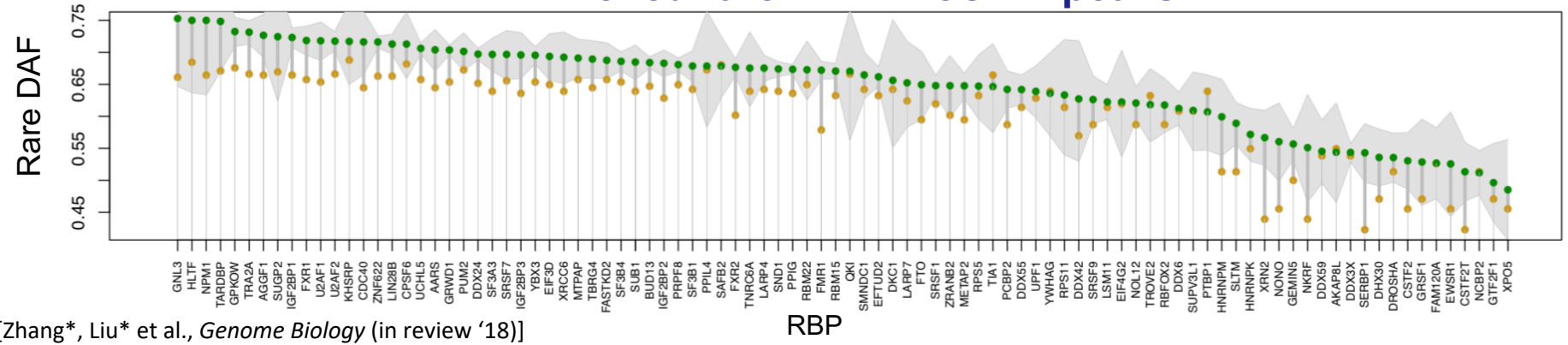
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



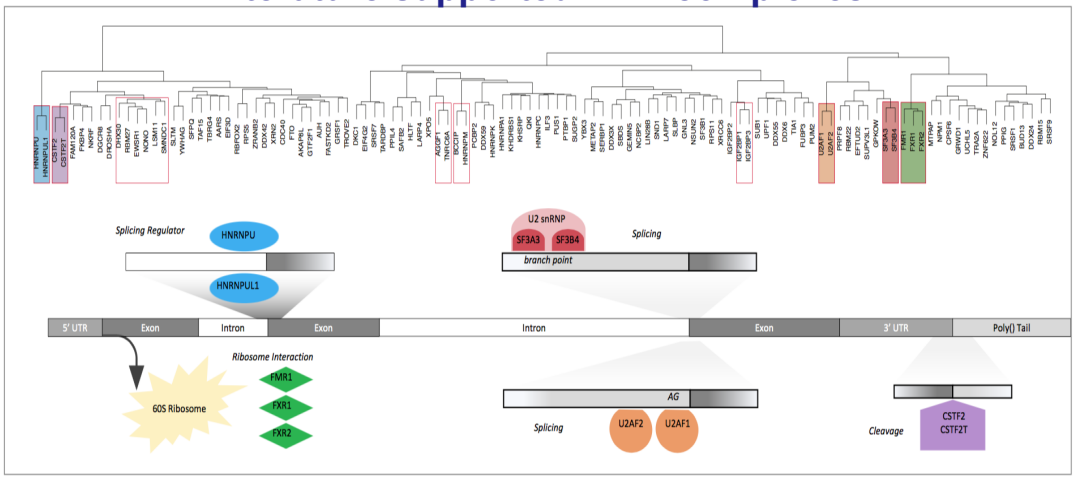
Enriched rare DAF in eCLIP peaks



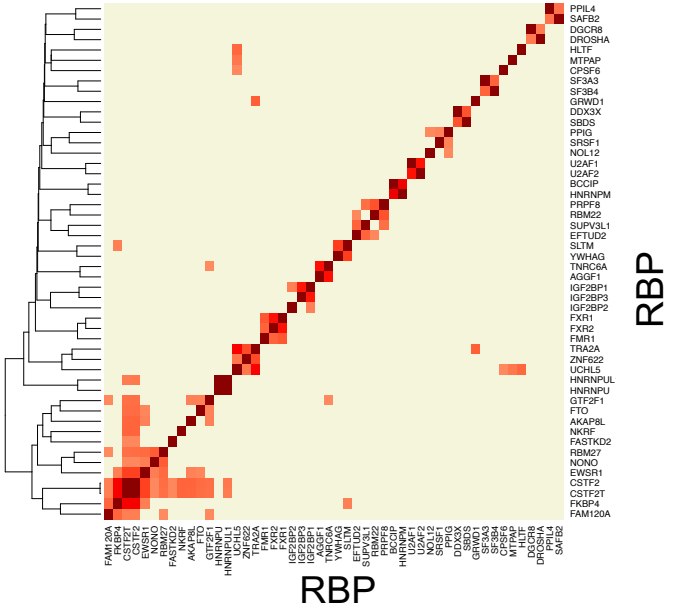
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

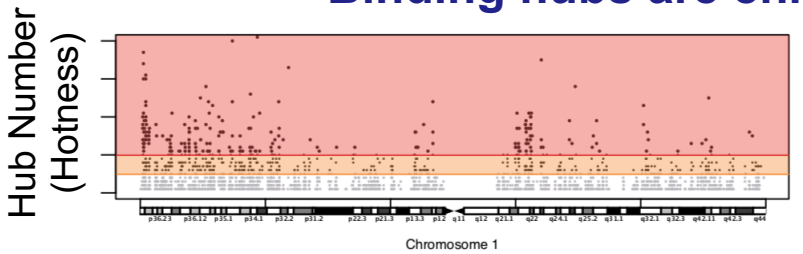
Literature supported RBP complexes

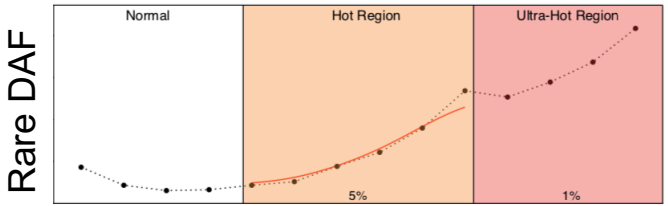


Unique co-binding patterns of RBPs

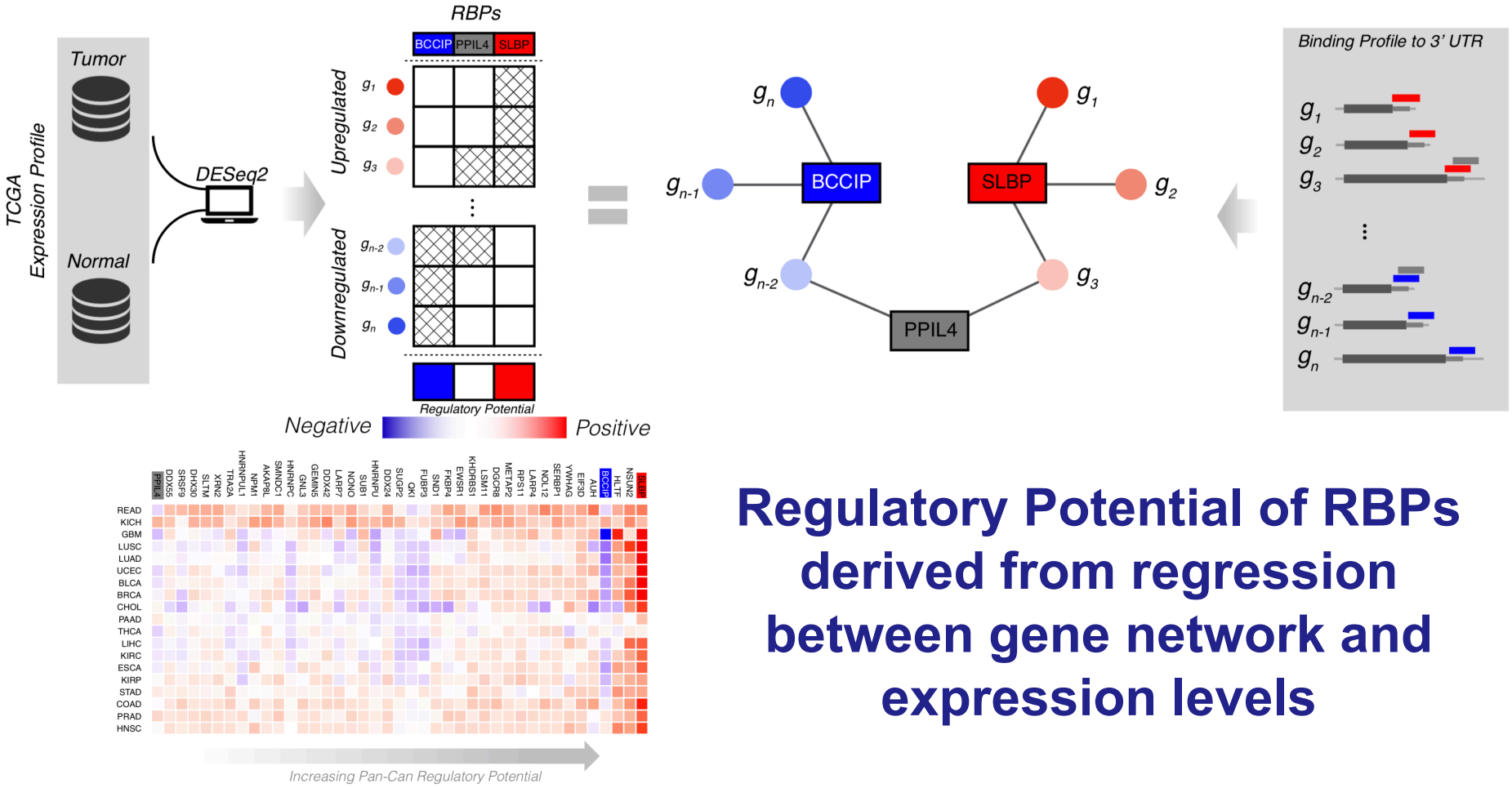


Binding hubs are enriched for rare variants





[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

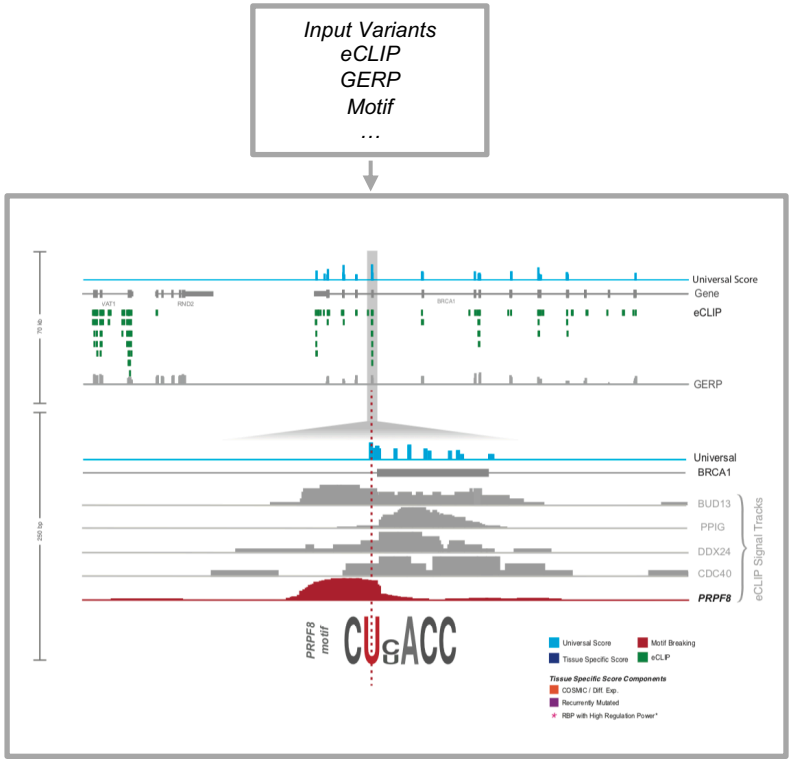


Regulatory Potential of RBPs derived from regression between gene network and expression levels

[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Visualization of RADAR Features and Scoring

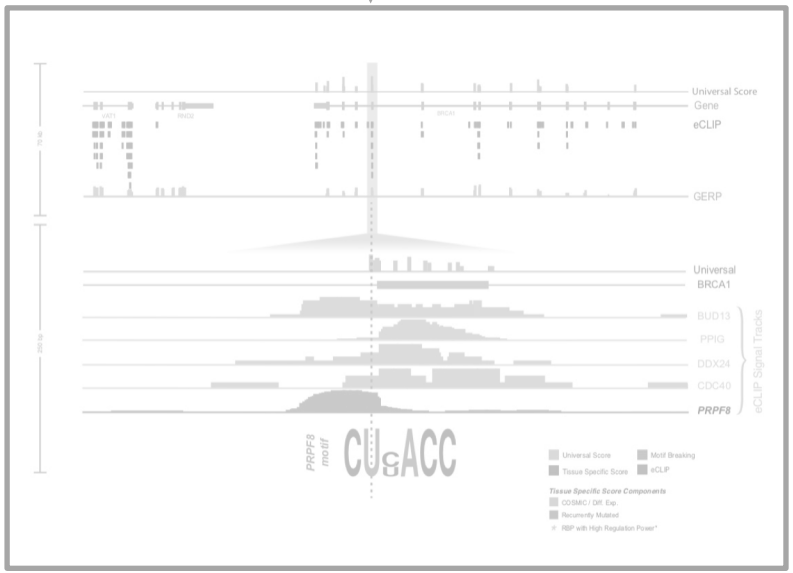
Germline Variants are Score Using a Universal Scoring Scheme



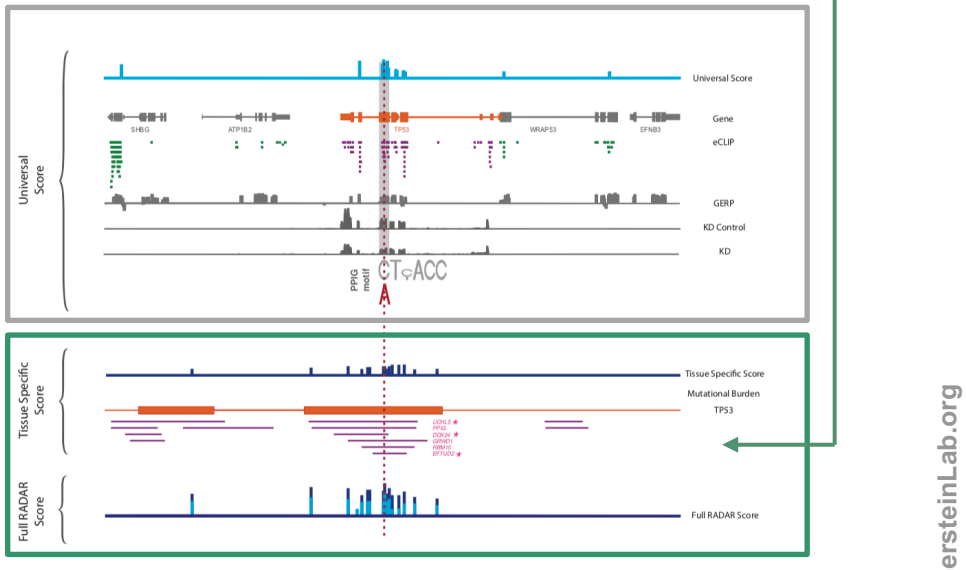
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Visualization of RADAR Features and Scoring

Input Variants
eCLIP
GERP
Motif
…



Tissue Specific:
Variants
Expression
Regulatory Potential



Somatic Variant Scored with Universal + Tissue specific context score

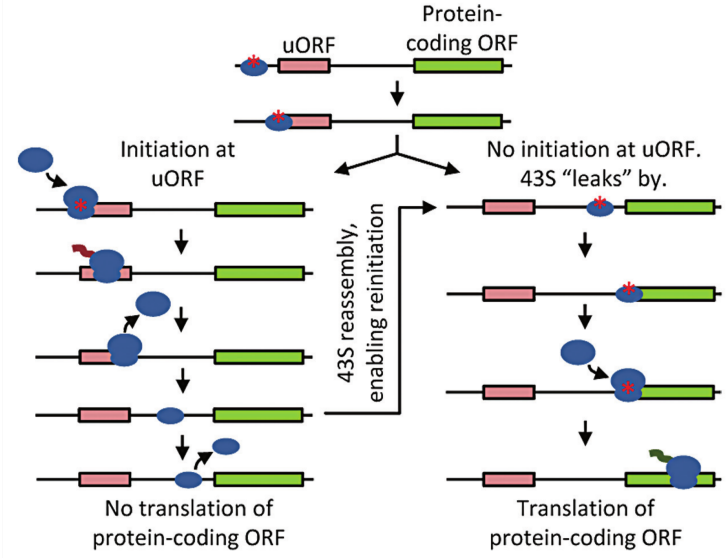
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

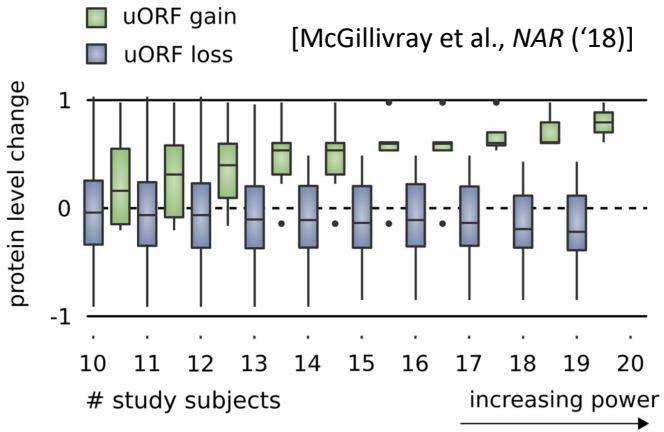
Upstream open reading frames (uORFs) regulate translation are affected by mutation

[Calvo et al., *PNAS* (‘09)]

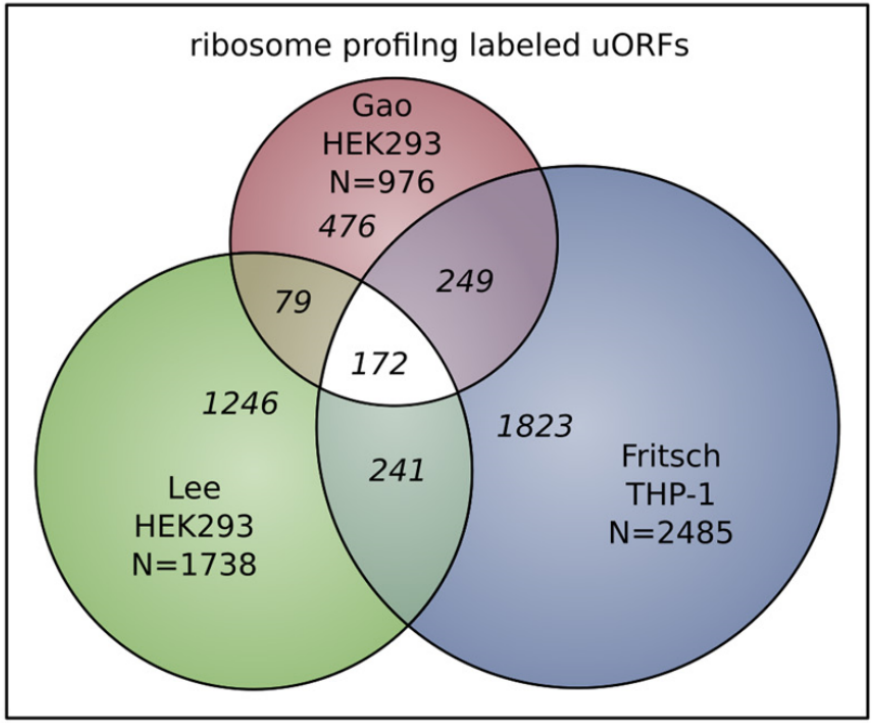


[Ferreira et al., *Bioengineered* (‘14)]

- uORFs regulate the translation of downstream coding regions.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.

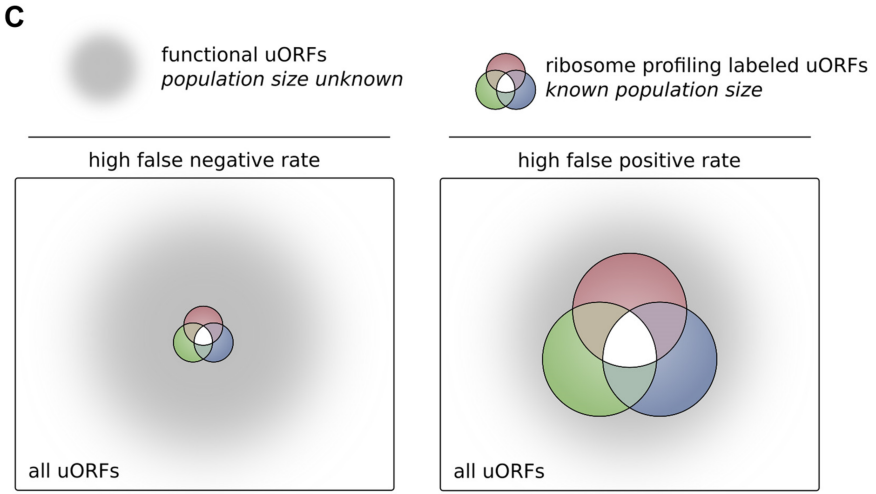


[McGillivray et al., *NAR* (‘18)]



From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

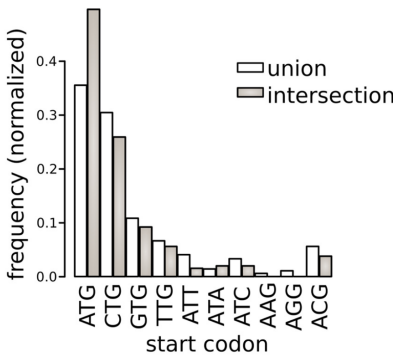


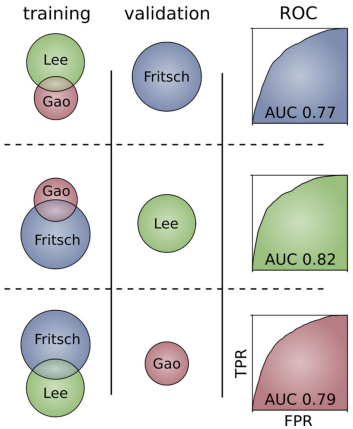
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

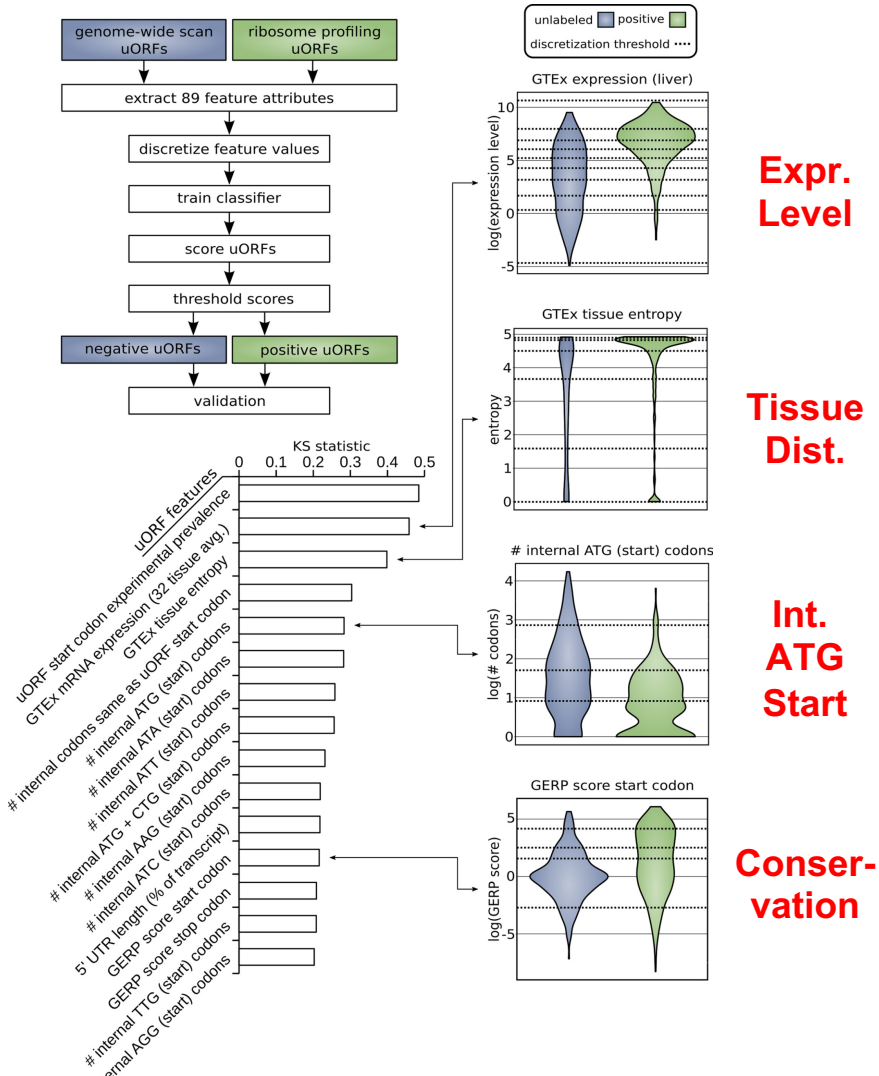
Prediction & validation of functional uORFs using 89 features

- All near-cognate start codons predicted.
- Cross-validation on independent ribosome profiling datasets and validation using in vivo protein levels and ribosome occupancy in humans (Battle et al. 2014).



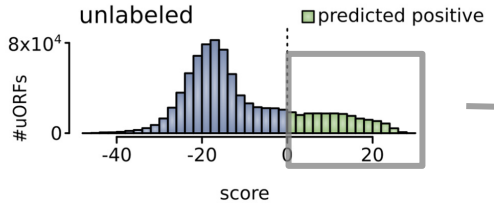


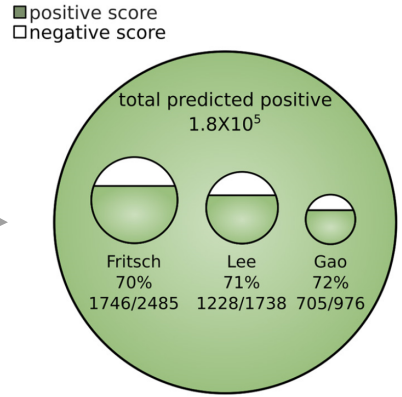
[McGillivray et al., *NAR* ('18)]

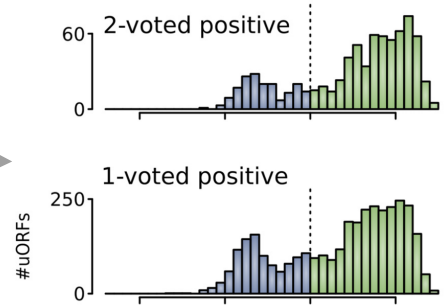


A comprehensive catalog of functional uORFs

Universe of **1.3M** uORFs scored via Simple Bayes algo.







- Predicted functional uORFs may be intersected with disease associated variants.

- **180K**: Large predicted positive set likely to affect translation
- Calibration on gold standards, suggests getting **~70%** of known

[McGillivray et al., *NAR* ('18)]

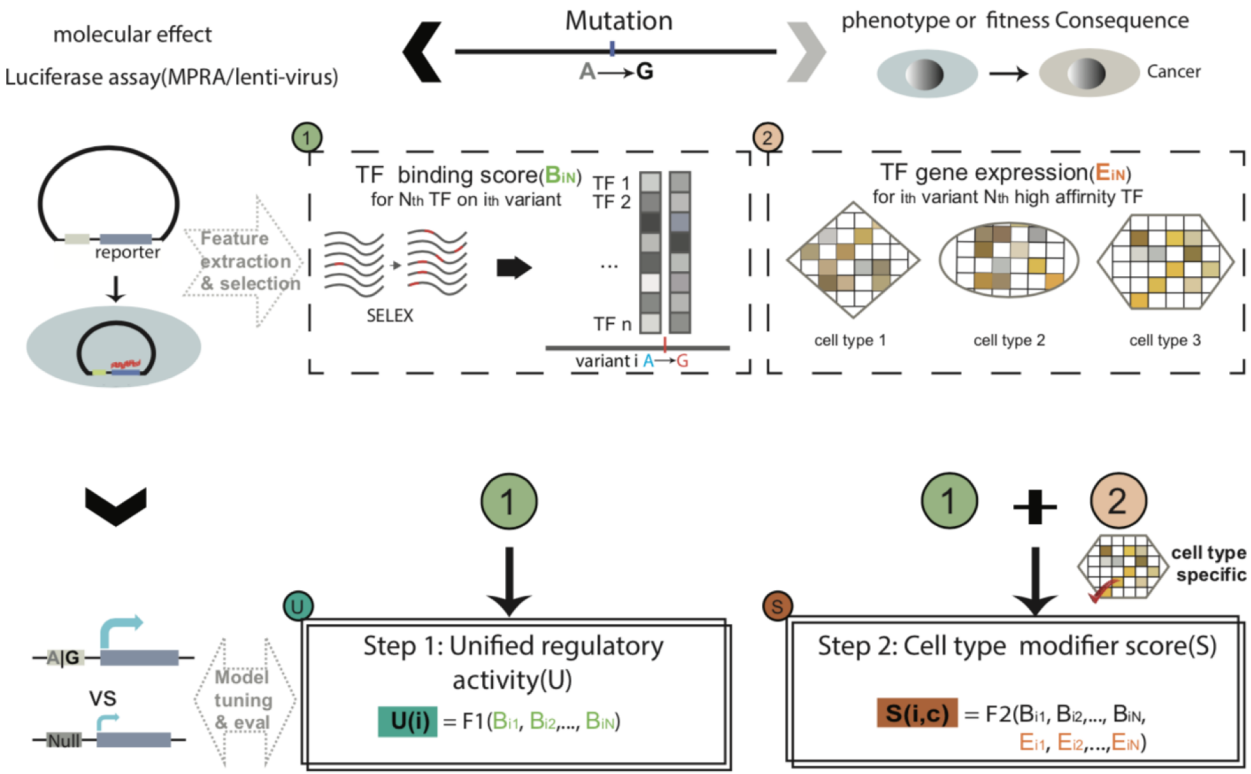
Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

Promotor Mutations

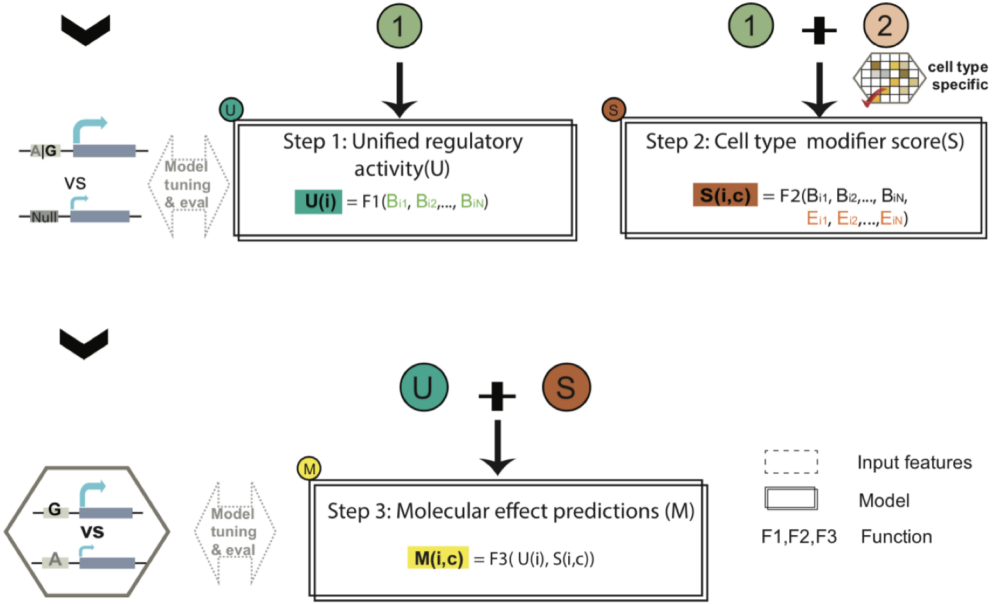
- How do we assess their effect?
  - What's the readout? Expression?
- Molecular v Organismic phenotype
- Importance of specific cellular context

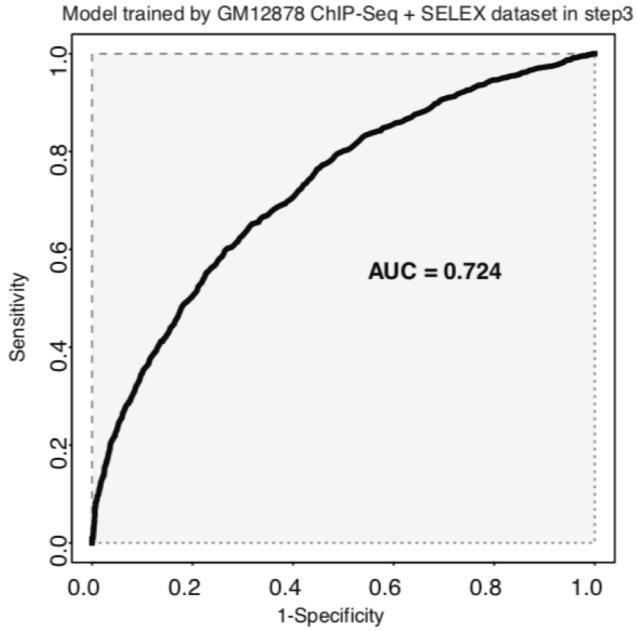
GRAM approach for assessing molecular impact

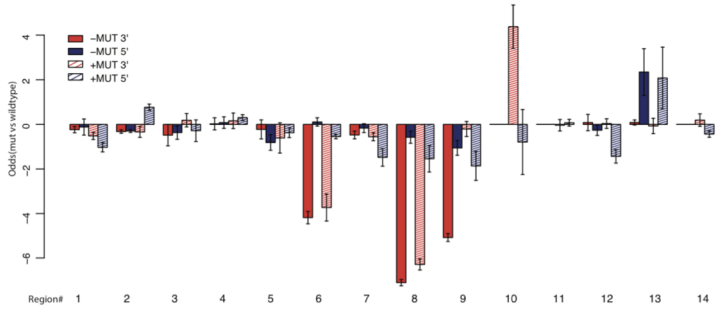


[Lou et al. ('19). PLOS Genetics]

GRAM performance

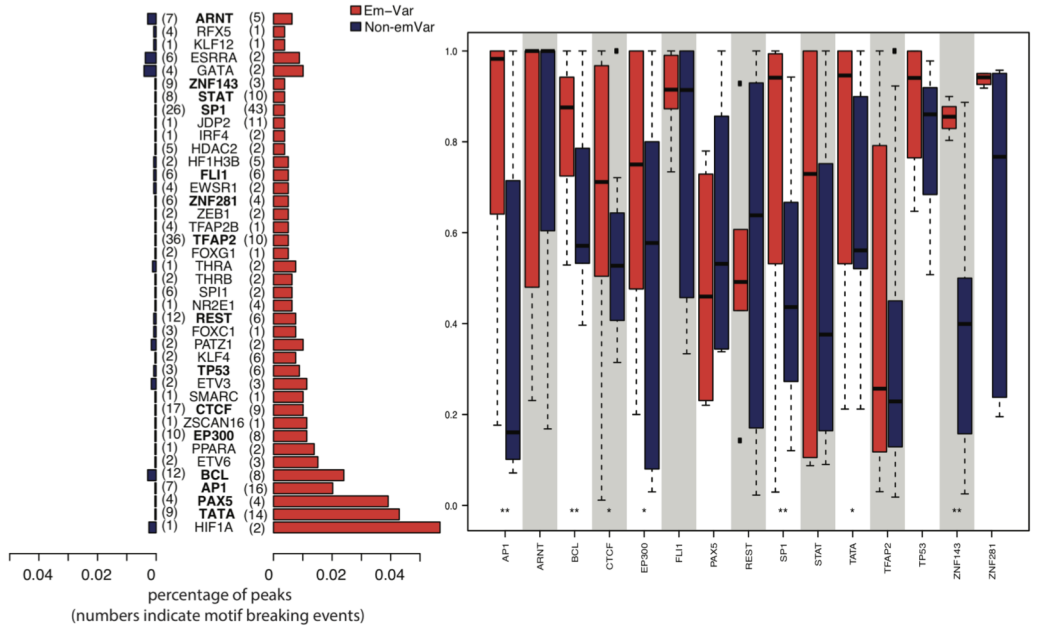


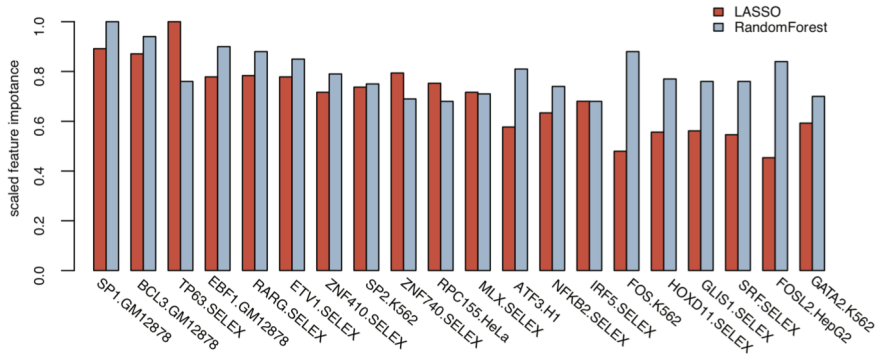




[Lou et al. ('19). PLOS Genetics]

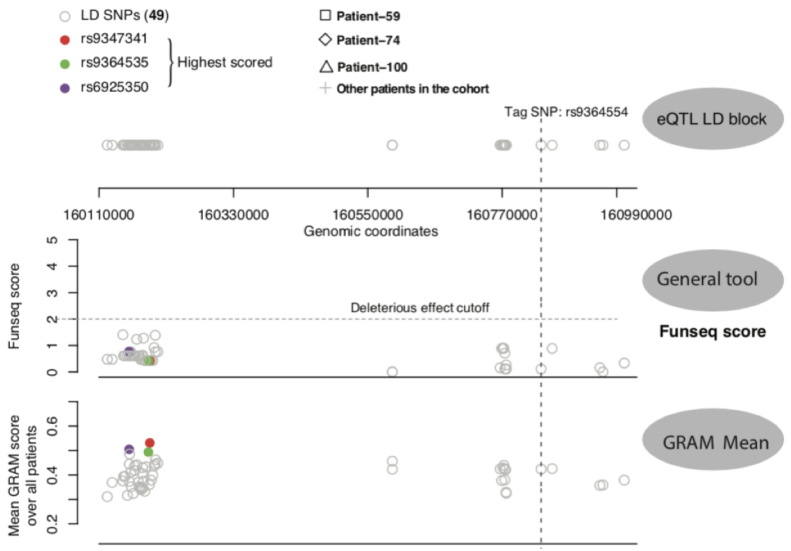
GRAM Important Features

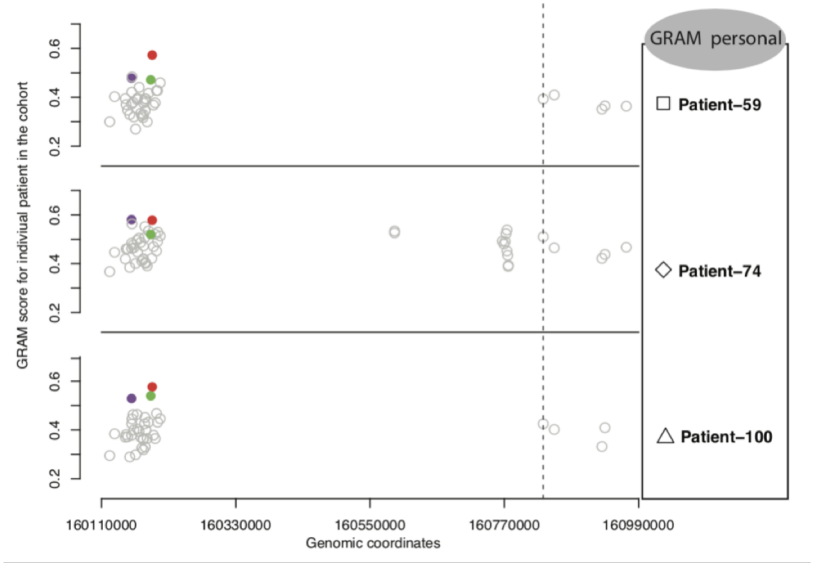




[Lou et al. ('19). PLOS Genetics]

GRAM application to find molecular cause within group of eQTL variants





[Lou et al. ('19). PLOS Genetics]

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

Computational analysis of variants: coding versus non-coding

- **Intro: types of variants**
  - Rare v common,
    somatic v germline, coding v noncoding
- **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 LoF Transcripts**
- **Using dynamics to help identify mutation clusters (Hotcommics)**
  - Find dynamic sub-communities & determine aggregated mutational burden within these
- **RADAR Prioritization for RBP sites**
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **GRAM to assess the molecular effect of (promotor) mutations**
  - Universal score + cell type specific score

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

RADAR.gersteinlab.org
J **Zhang**, J **Liu**, D Lee, J-J Feng,
L Lochovsky, S Lou, M Rutenberg-Schoenberg

github.gersteinlab.org/**uORFs**
P **McGillivray**, R Ault, M Pawashe,
R Kitchen, S Balasubramanian

github.com/gersteinlab/**Frustration**
S **Kumar**, D Clarke

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

github.com/gersteinlab/**gram**
S **Lou**, KA Cotter, T Li, J Liang, H Mohsen, J Liu,
J Zhang, S Cohen, J Xu, H Yu, MA Rubin

github.com/gersteinlab/**hotcommics**
S **Kumar**, D Clarke

ALoFT.gersteinlab.org
S **Balasubramanian**, Y **Fu**,
M Pawashe, P McGillivray, M Jin, J Liu,
K Karczewski, D MacArthur


STOP

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, 2019.
- 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