I. Frustration

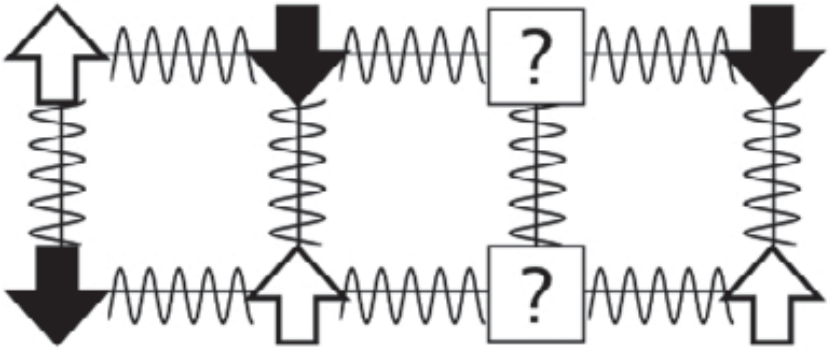
- Background & conceptualization (advantages of secondary calculations)
- Corrected formulation
- Data survey and processing
- MAF analysis (rare alleles associated with extreme delta_F)
- Cancer SNVs & genes (rationalize in TSGs + Oncogenes)
- Thresholding to classify SNVs
- Example case of rescued false negatives: Glucokinase

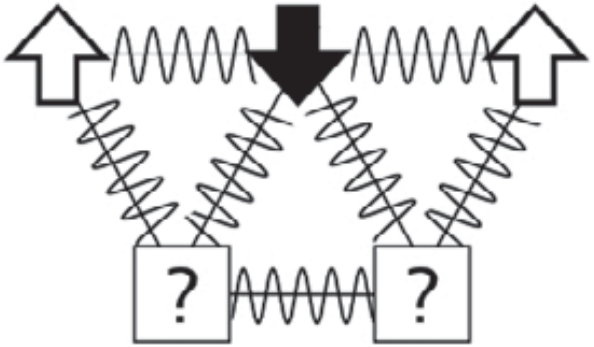
II. eQTLs

- Background
- Reproducibility in Covariates
- Reproducibility in RPKM
- Framingam data (miRNA-eQTLs)
- Current Objectives

III. Supplementary Slides

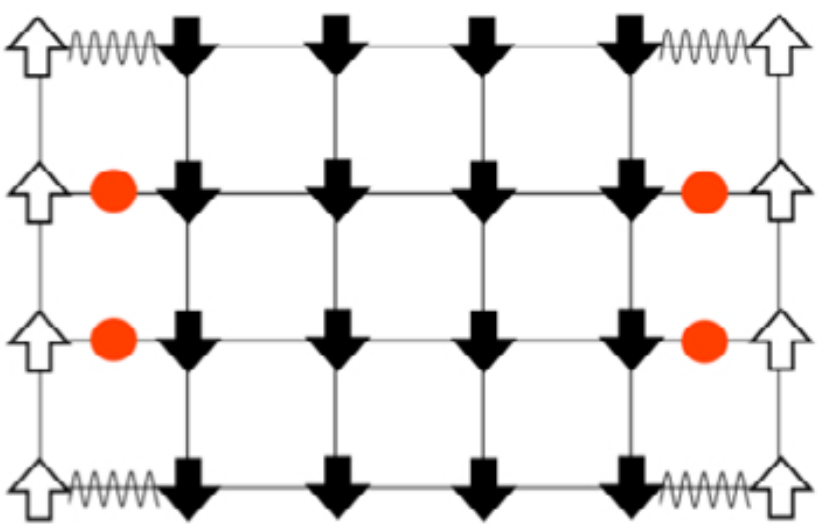
Background & conceptualization

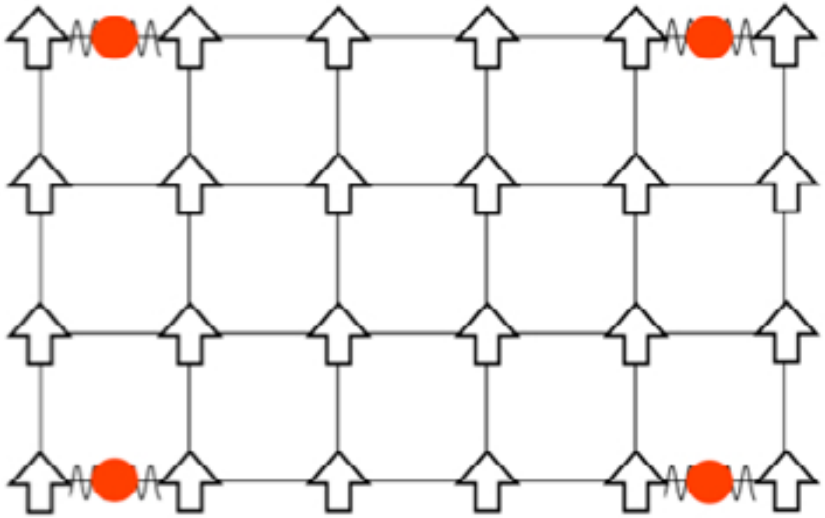




Ferreiro, et al, 2014

Background & conceptualization





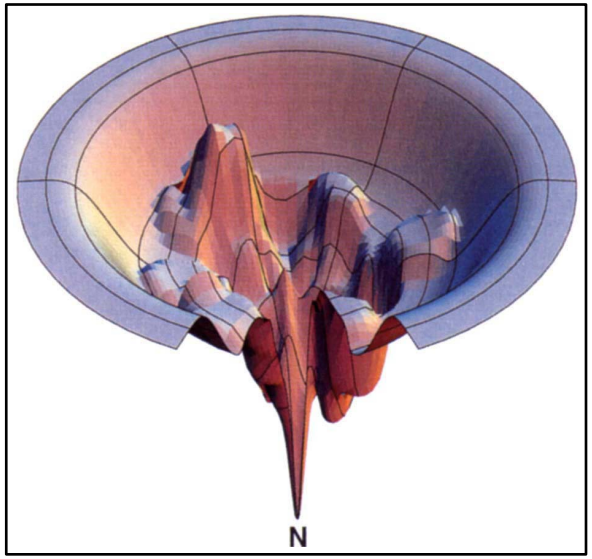
Ferreiro, et al, 2014

straight lines: Favorable ferromagnetic interactions
squiggly lines: Favorable antiferromagnetic interactions

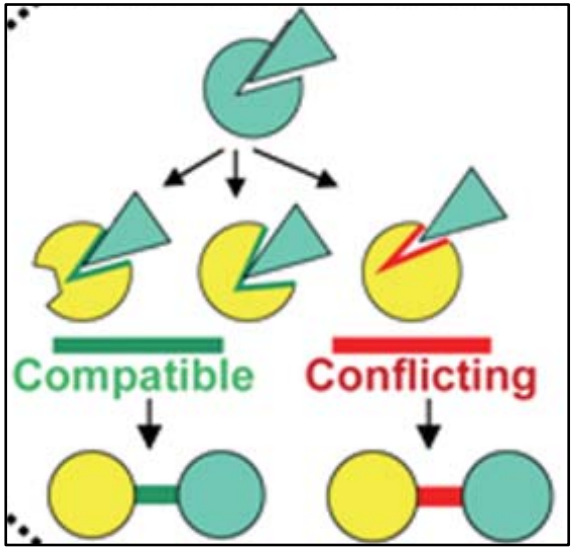
Changes in localized frustration may disrupt essential functionality without introducing global destabilization

Note that frustration is intrinsic to many biological processes!

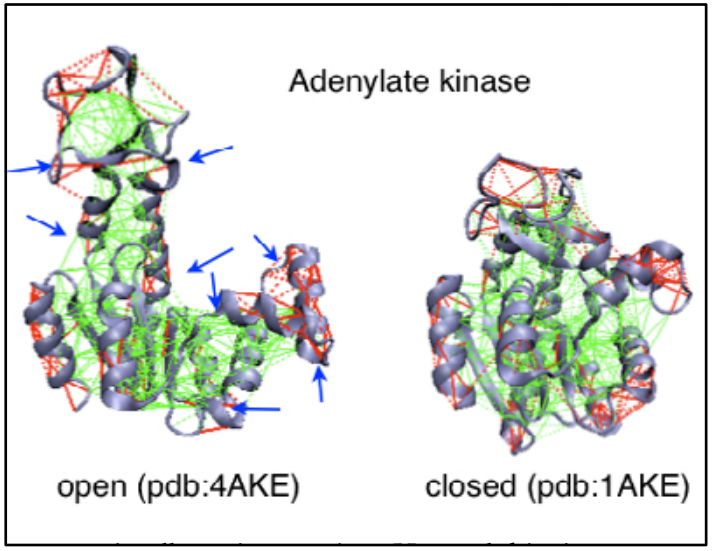
- *Catalytic centers*
- *Allosteric contexts & local conformational switches*
- *Binding sites are often frustrated*
- *Metastable and multi-stable proteins*
- *Protein aggregation*
- *Nucleic acids & protein complexes*



Dill et al, 1997

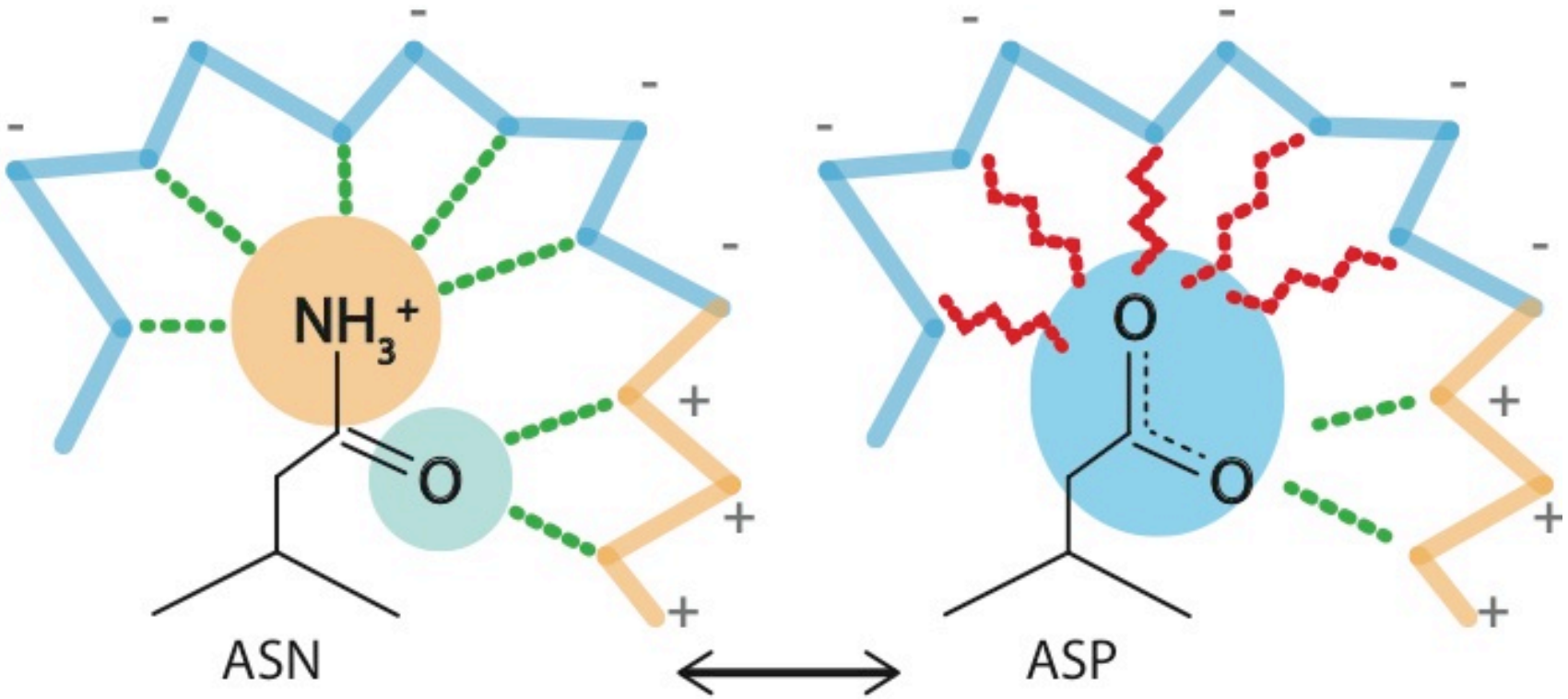


Bhardwaj et al, 2011

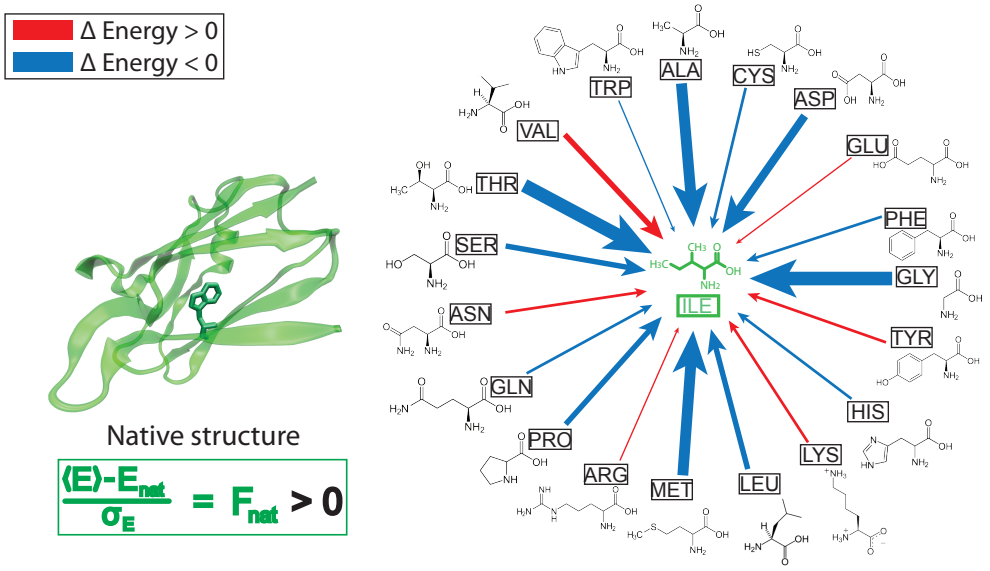


Ferreiro et al, 2014

Background & conceptualization

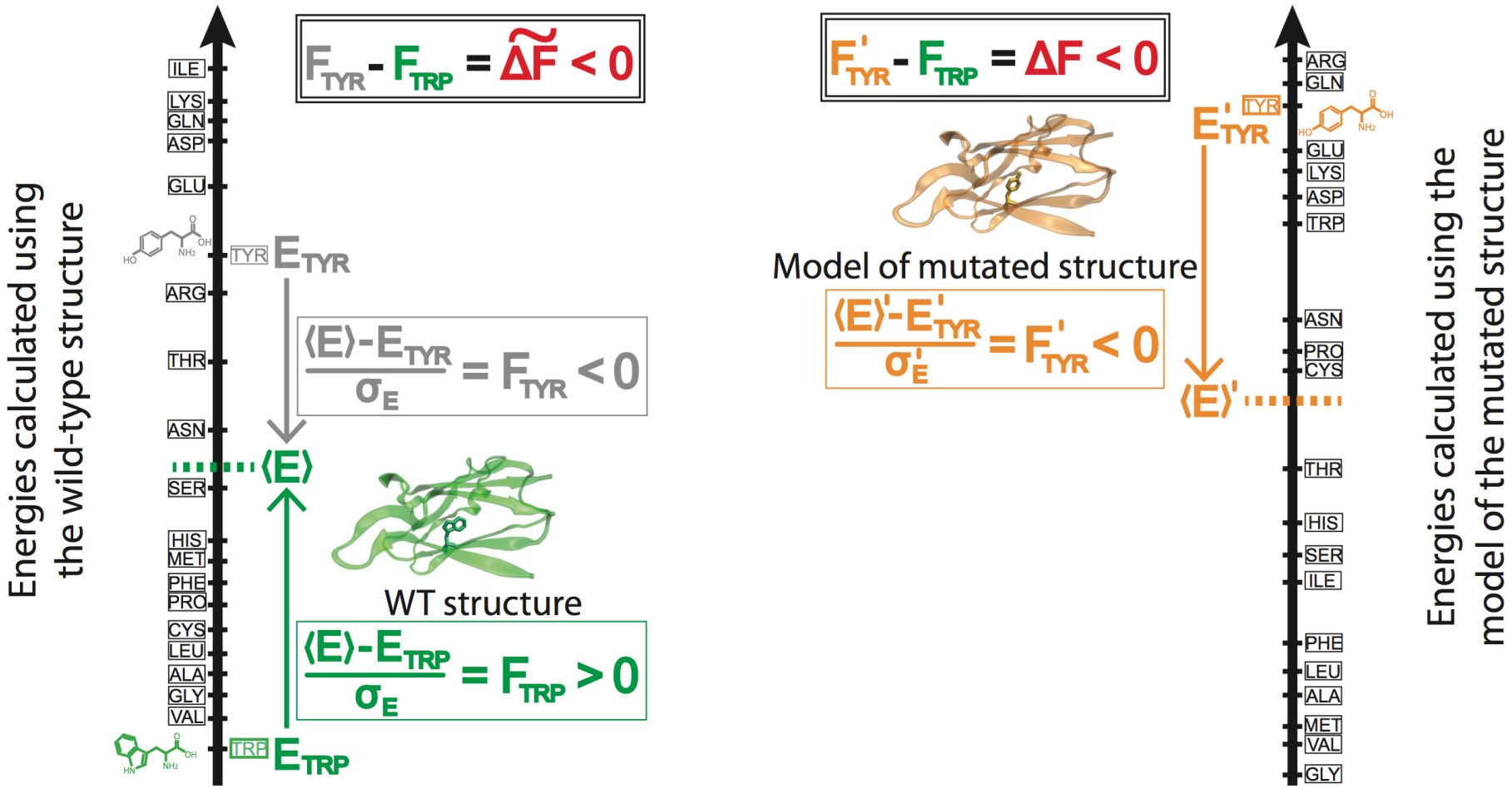


Computational simplicity offers opportunities for application to large SNV datasets



Computational simplicity offers opportunities for application to large SNV datasets

Demonstration of a typical deleterious SNV



Corrected formulation

total energy of the wt protein

Mean energy of all decoys

$$F_i = \frac{E_i^{T,N} - \langle E_{i'}^{T,U} \rangle}{\sqrt{1/N \sum_{k=1}^{n} (E_{i'}^{T,U} - \langle E_{i'}^{T,U} \rangle)^2}}.$$

Ferreiro et al, 2014

Criteria for "minimal frustration" (think of wt structures):

F_i >= + 0.78

(a contact is highly frustrated if Fi < -1)

Corrected formulation

Frustration in biomolecules

Diego U. Ferreiro[1], Elizabeth A. Komives[2]* and Peter G. Wolynes[3]*

[1] Protein Physiology Lab, Dep de Química Biológica, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-IQUIBICEN, Buenos Aires, Argentina
[2] Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093, USA
[3] Department of Physics, Department of Chemistry, and Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA

Abstract. Biomolecules are the prime information processing elements of living matter. Most of these inanimate systems are polymers that compute their own structures and dynamics using as input seemingly random character strings of their sequence, following which they coalesce and perform integrated cellular functions. In large computational systems with finite interaction-codes, the appearance of conflicting goals is inevitable. Simple conflicting forces can

Corrected formulation

Frustration in biomolecules

PNAS

On the role of frustration in the energy landscapes of allosteric proteins

Diego U. Ferreiro[a], Joseph A. Hegler[b,c], Elizabeth A. Komives[b], and Peter G. Wolynes[b,c,1]

[a]Protein Physiology Lab, Department of Biological Chemistry, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina, Buenos Aires, Argentina C1428EGA; [b]Department of Chemistry and Biochemistry; and [c]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92107.

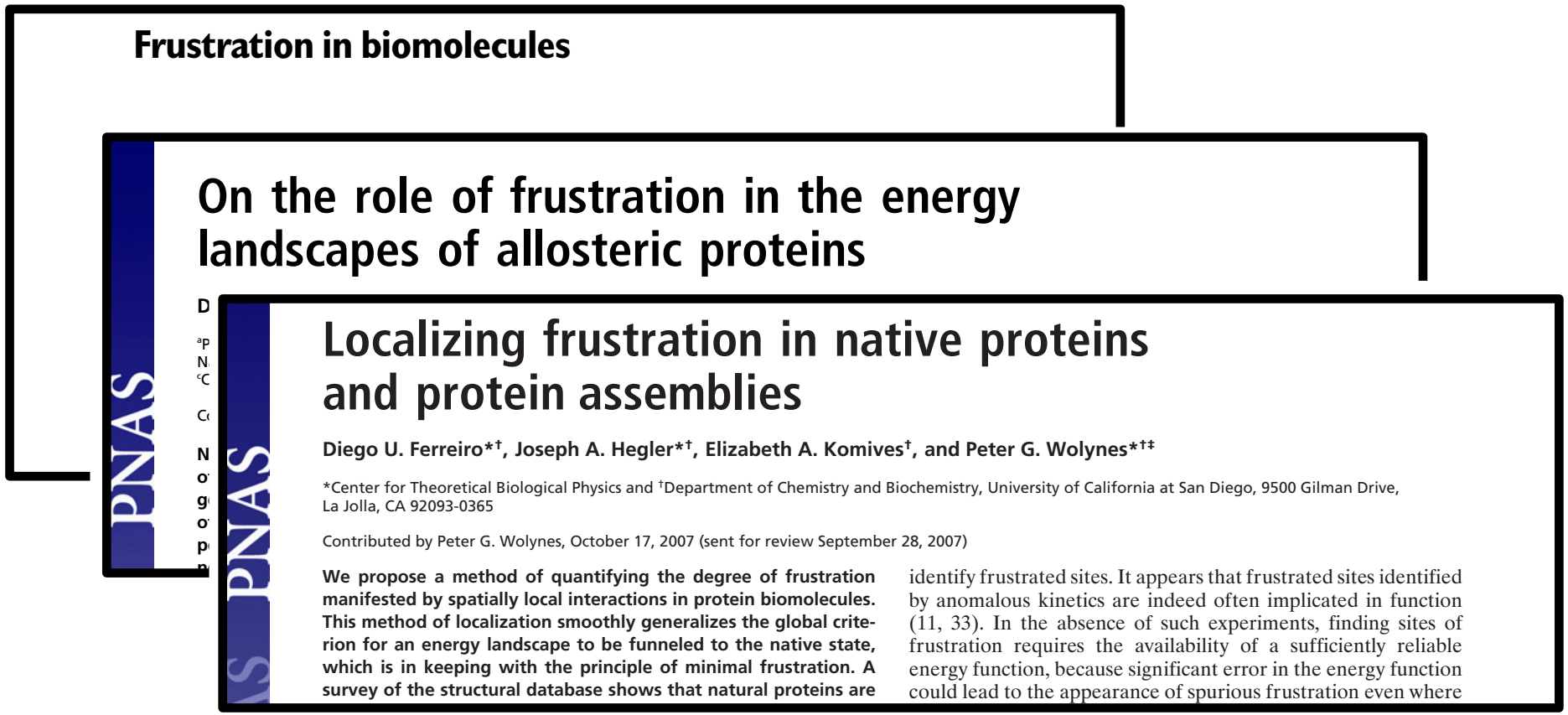
Contributed by Peter G. Wolynes, December 16, 2010 (sent for review November 24, 2010)

Natural protein domains must be sufficiently stable to fold but often need to be locally unstable to function. Overall, strong energetic conflicts are minimized in native states satisfying the principle of minimal frustration. Local violations of this principle open up possibilities to form the complex multifunnel energy landscapes needed for large-scale conformational changes. We survey the

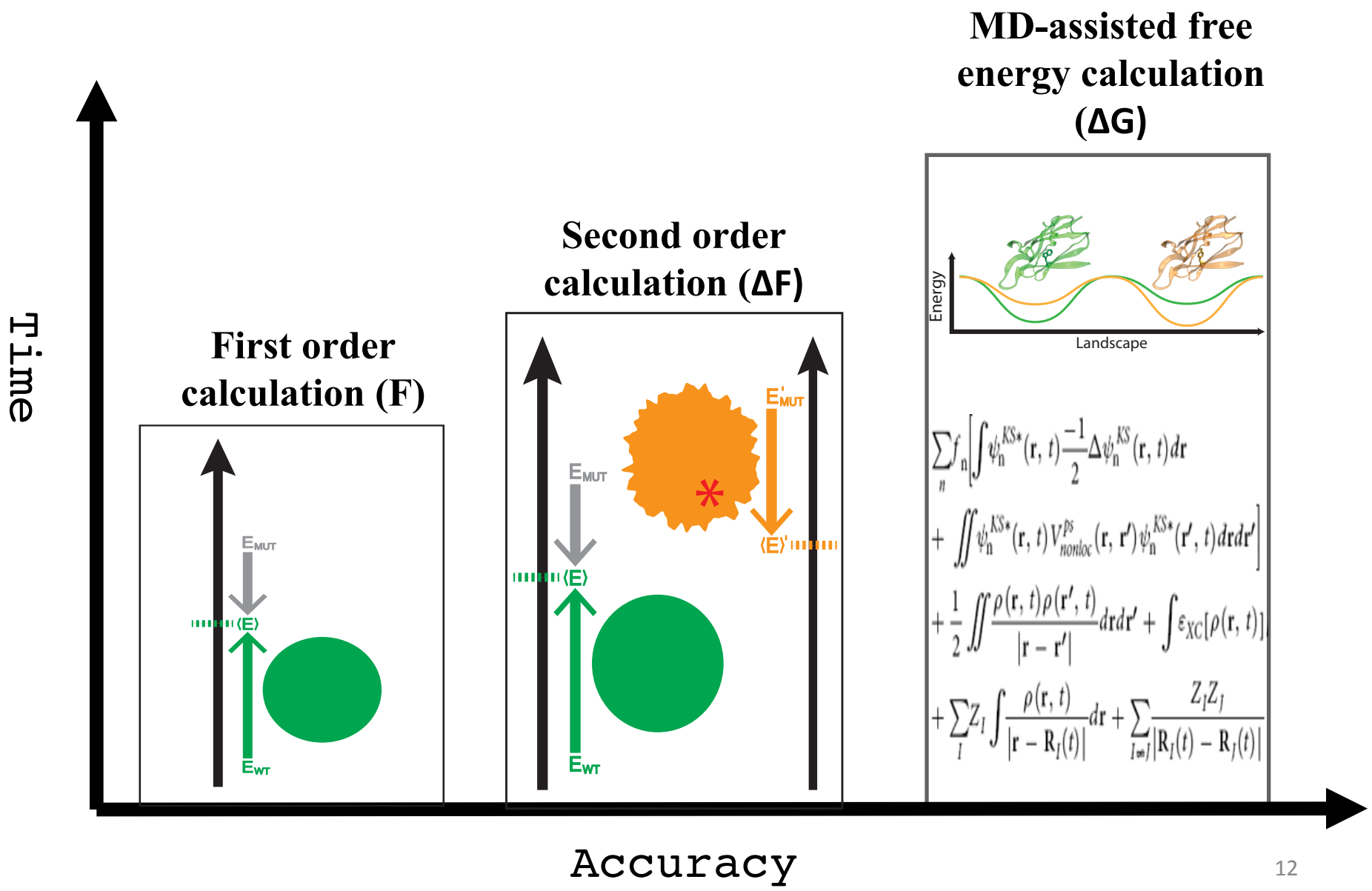
here how local violations of the minimal frustration principle open up possibilities for more complex energy landscapes needed for allostery and large-scale conformational changes (12, 13).

Multiple funnels to structurally distinct low-free-energy states can also be achieved by other mechanisms (14), symmetry being the main route to such degeneracy (15). Nearly rigid macromo-

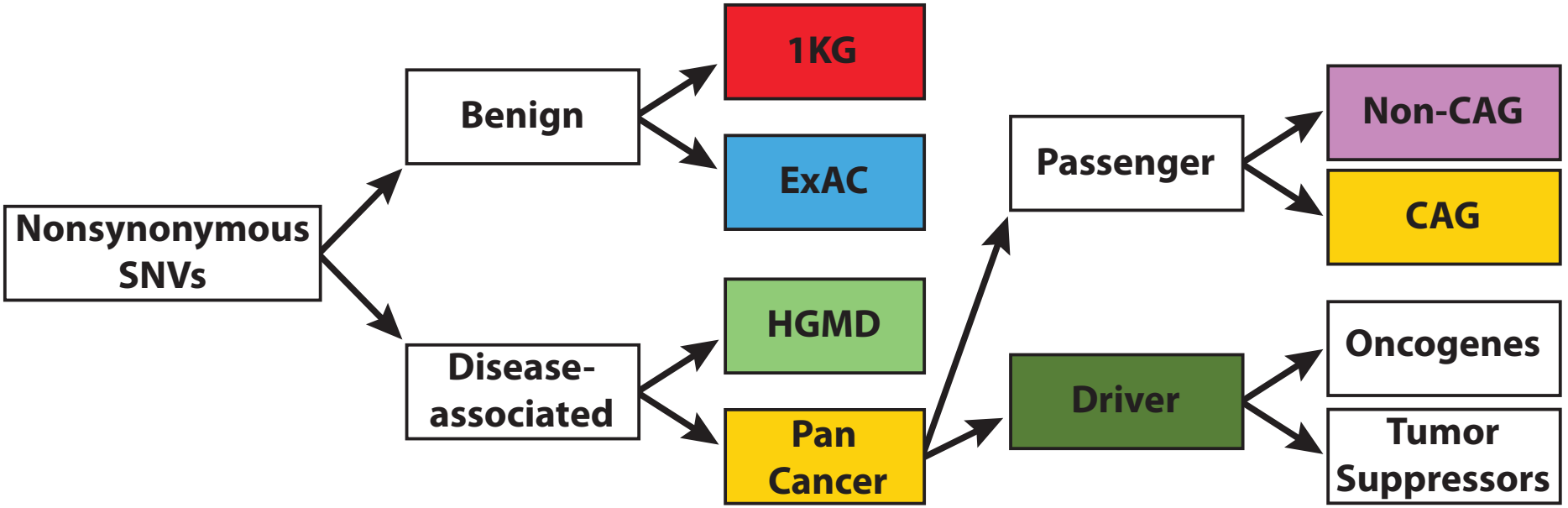
Corrected formulation

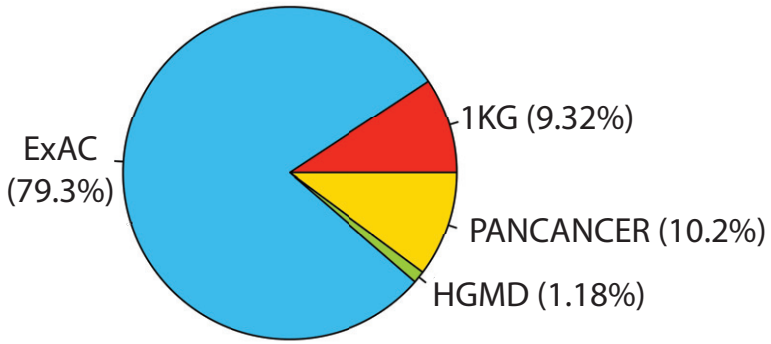


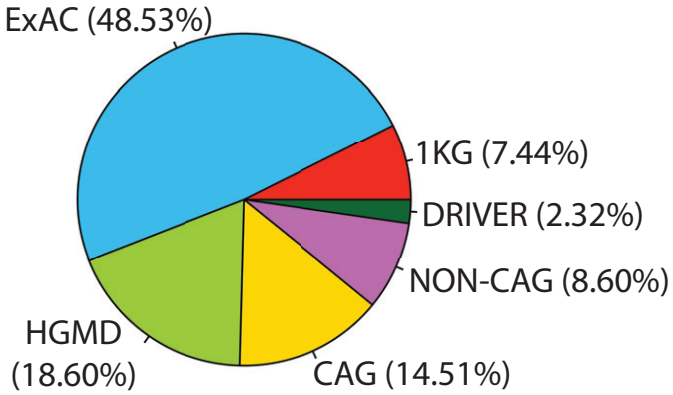
(advantages of secondary calculations)



Data survey and processing







Kumar et al, NAR 2016

Data survey and processing

Table 1. Summary statistics on the number of SNVs used in comparative analyses. Shown are variant counts for non-disease *(top)*, HGMD *(bottom-left)*, and pan-cancer SNVs *(bottom-right)*.

Conservation measure	1000 Genomes core	1000 Genomes surface	ExAC core	ExAC surface
DAF rare (common)	2267 (85)	1570 (106)	17972 (102)	11550 (83)
GERP conserved (variable)	1552 (287)	1132 (212)	12165 (2174)	7637 (1406)

Conservation measure	HGMD core	HGMD surface
GERP conserved (variable)	5158 (961)	1113 (221)

SNV type	PANCAN core	PANCAN surface
non-CAG	2153	1848
CAG	4140	2767
driver	877	486

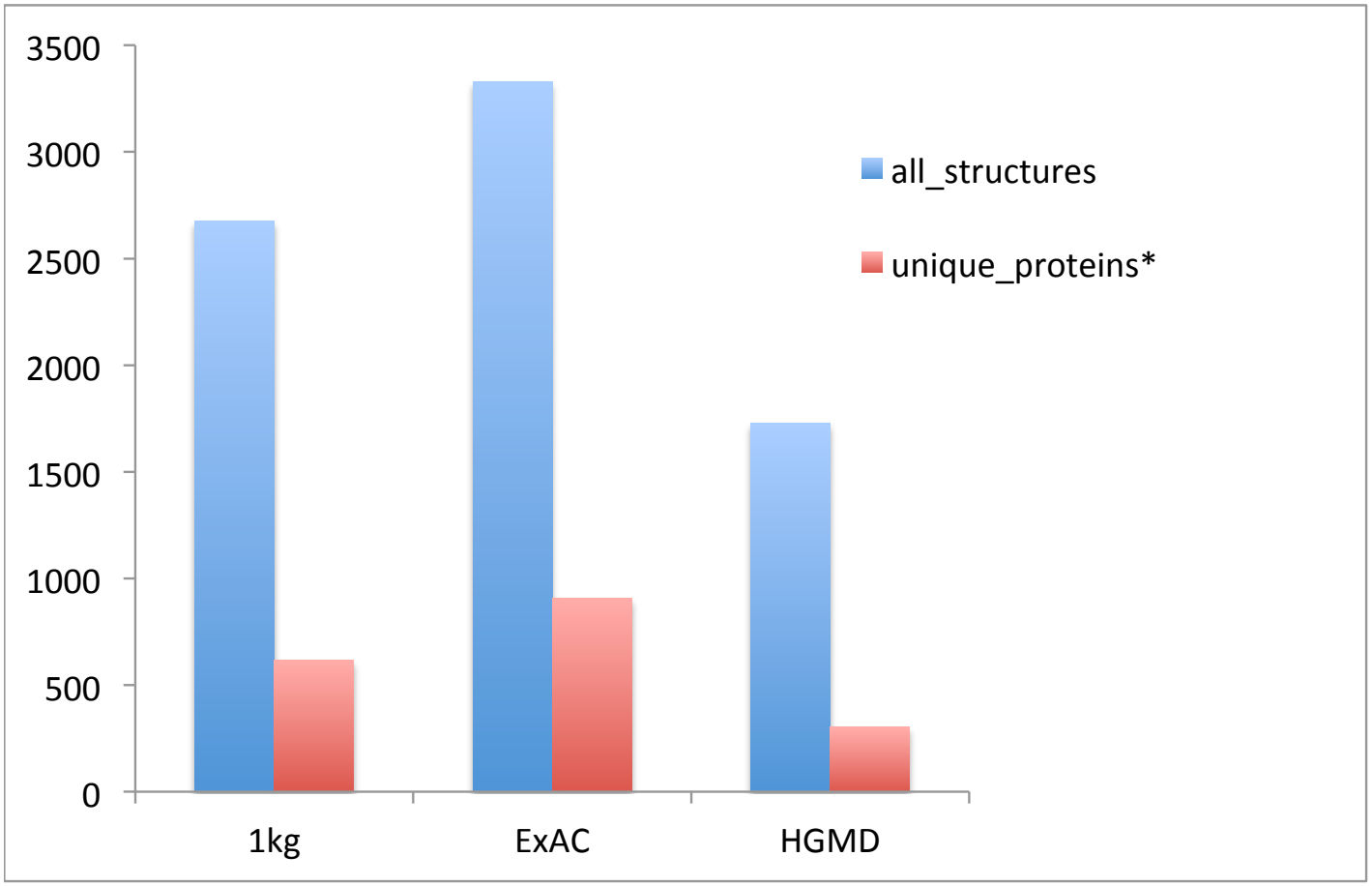
Kumar et al, NAR 2016

Data survey and processing

"Another known issue is strong annotation disparity between known Mendelian disease mutations (e.g. HGMD disease variants) and other variants: most of HGMD mutations are reported in a small subset of proteins, while majority of the proteins only have fewer and mostly benign or unknown significance variants reported for them. This creates bias when performing comparisons between the two functional classes of variants. In case of PDB-mapped variants, such annotation bias might have been alleviated to some extent by the PDB intrinsic bias (mentioned above, skews PDB & HGMD data towards the same proteins) but it requires further investigation. **Authors should present statistics for the number of unique proteins and the distribution of variants in the unique proteins for each of their datasets. They should also attempt to perform their analysis on a (semi-)balanced set(s) of variants, using sets of proteins where both disease and neutral mutations are present**. See Grimm et al. (2015) Human Mut. 36:513-523 for an example of such balanced sets and trends analysis."

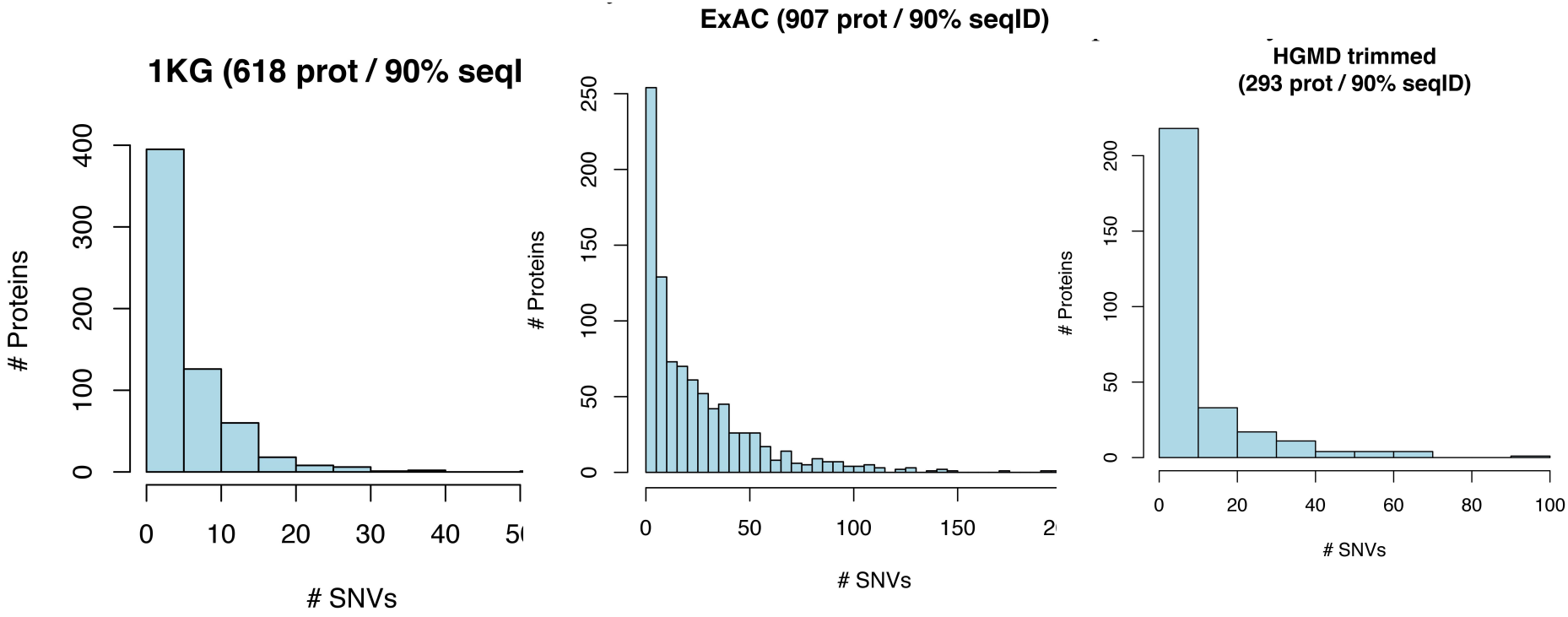
1) Determine the # of unique proteins in each dataset

	1kg	ExAC	HGMD
all_structures	2675	3327	1728
unique_proteins*	618	907	303



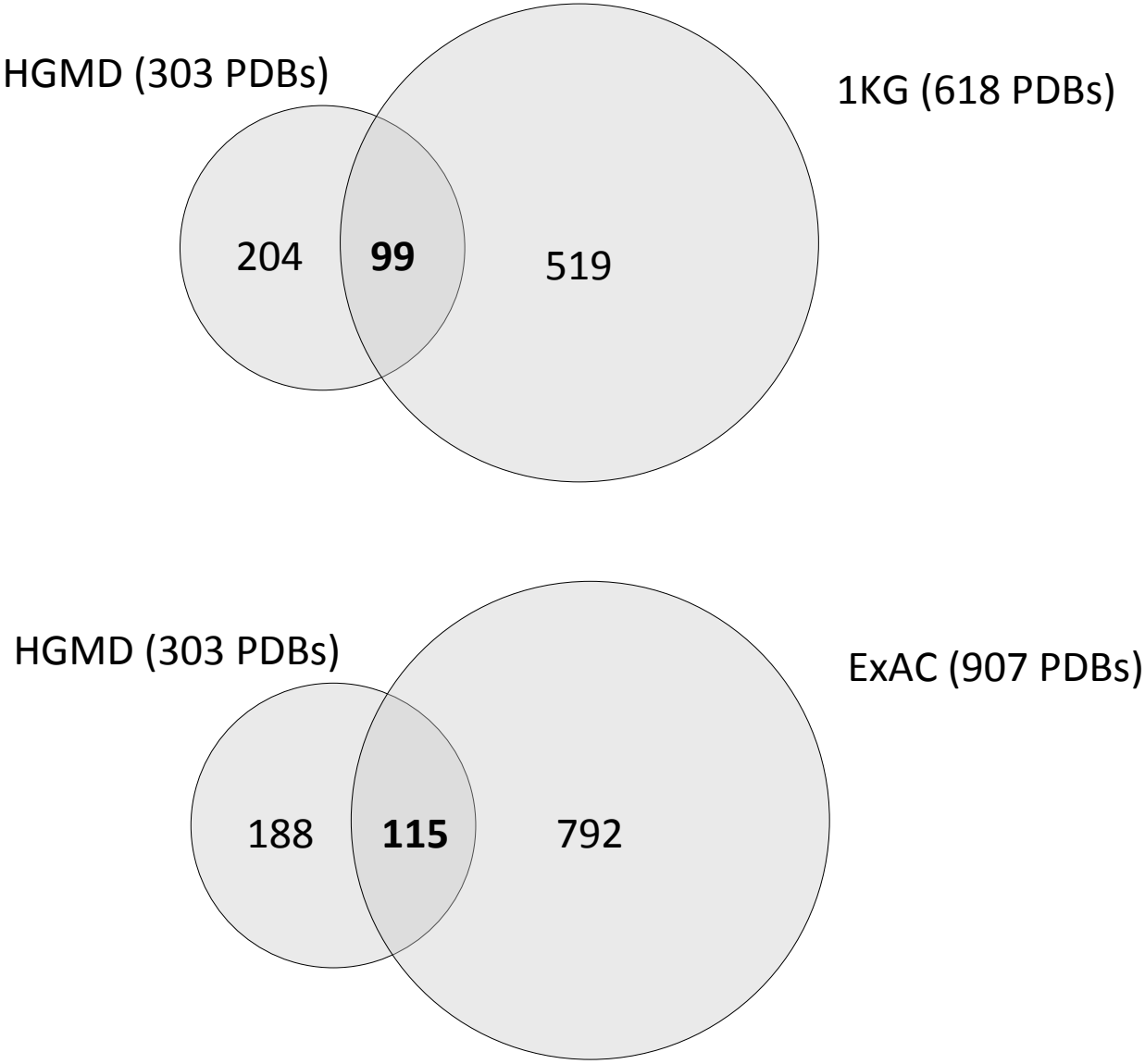
* Defined to be unique if no 2 proteins within the set have chains sharing more than 90% sequence similarity

2) Within the set of non-redundant (i.e., unique) set of proteins: "**present statistics for the number of unique proteins and the distribution of variants in the unique proteins**"

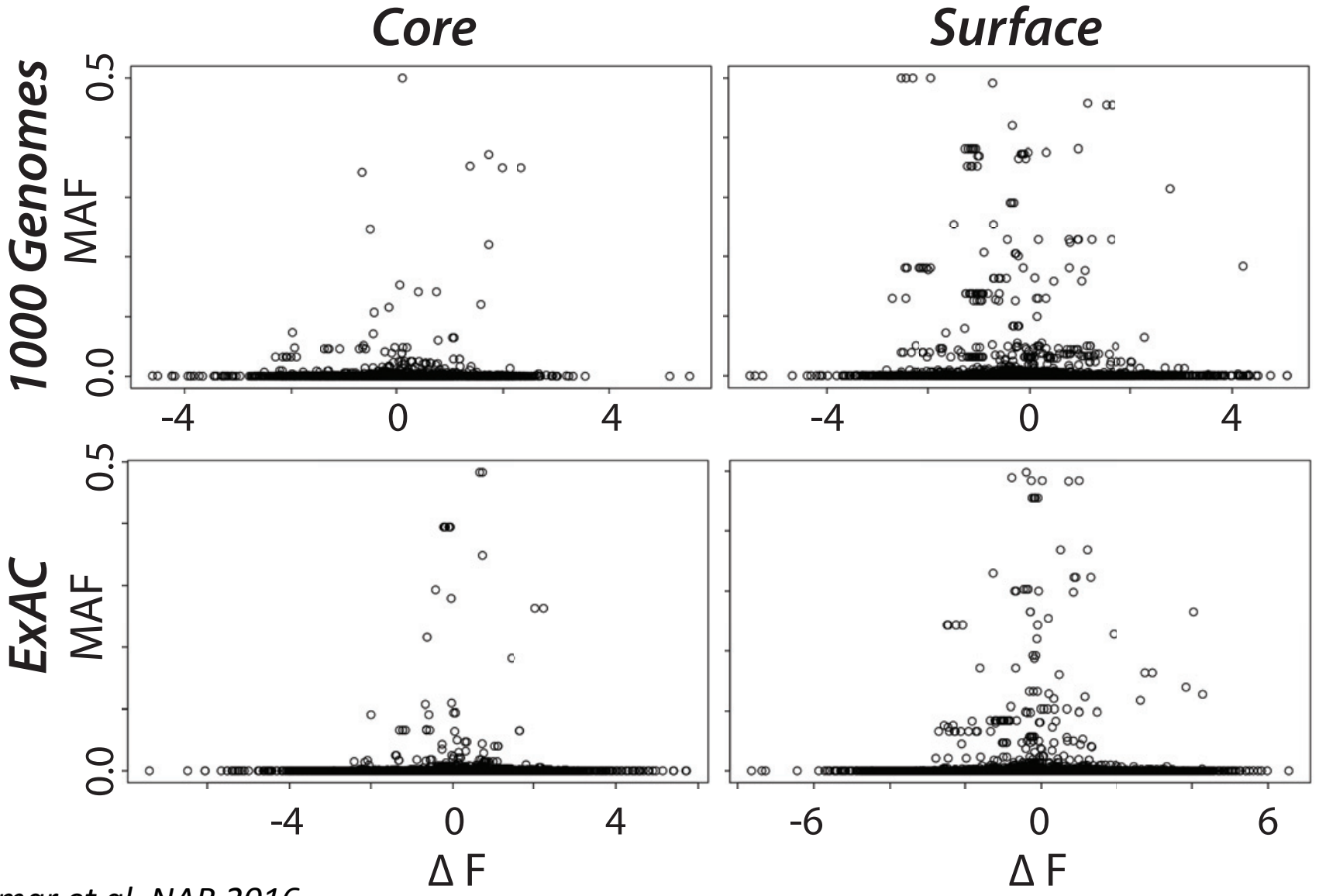


Kumar et al, NAR 2016

2) Within the set of non-redundant (i.e., unique) set of proteins: "**present statistics for the number of unique proteins and the distribution of variants in the unique proteins**"



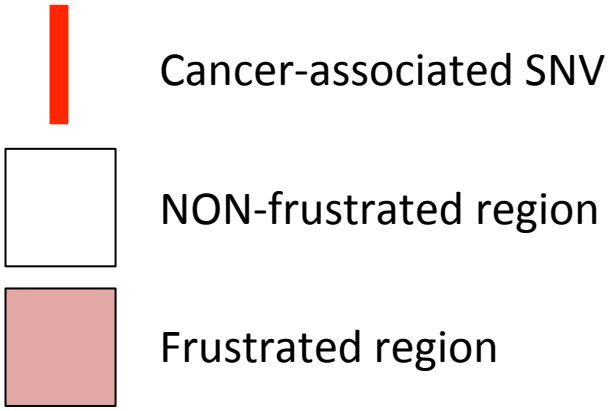
MAF analysis (rare alleles associated with extreme delta_F)



Kumar et al, NAR 2016

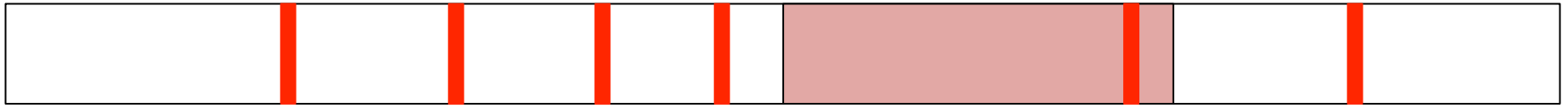
Cancer SNVs & genes (rationalize in TSGs + Oncogenes)

Are Cancer-Associated SNVs enriched in frustrated regions?



Observed: X = # of cancer-associated SNVs that intersect frustrated regions (5 in this case)

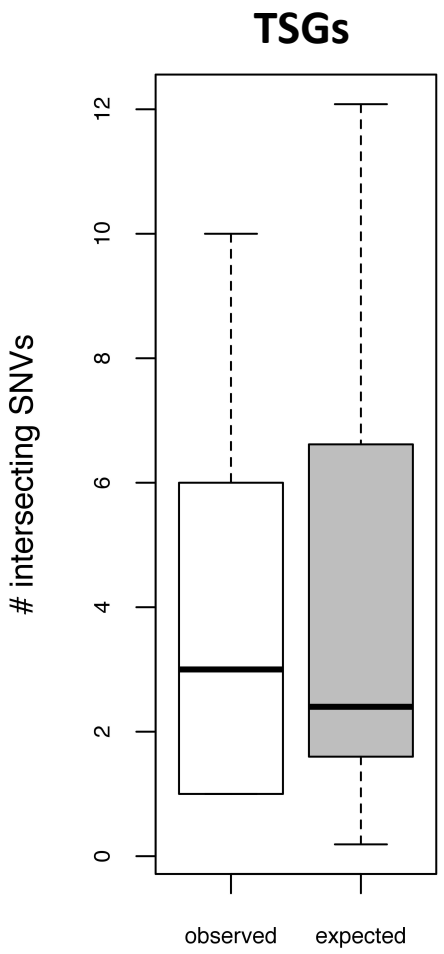
Expected: E[X] = [# frustrated residues / total # residues in protein] * [total # of cancer-associated SNVs]



Cancer SNVs & genes (rationalize in TSGs + Oncogenes)

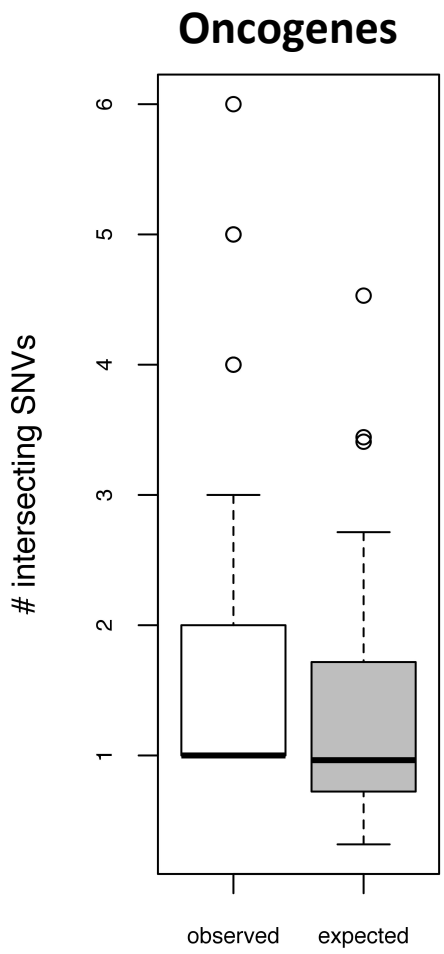
Are Cancer-Associated SNVs enriched in maximally frustrated regions?

-- YES --



p-value = 0.005519

N = 28



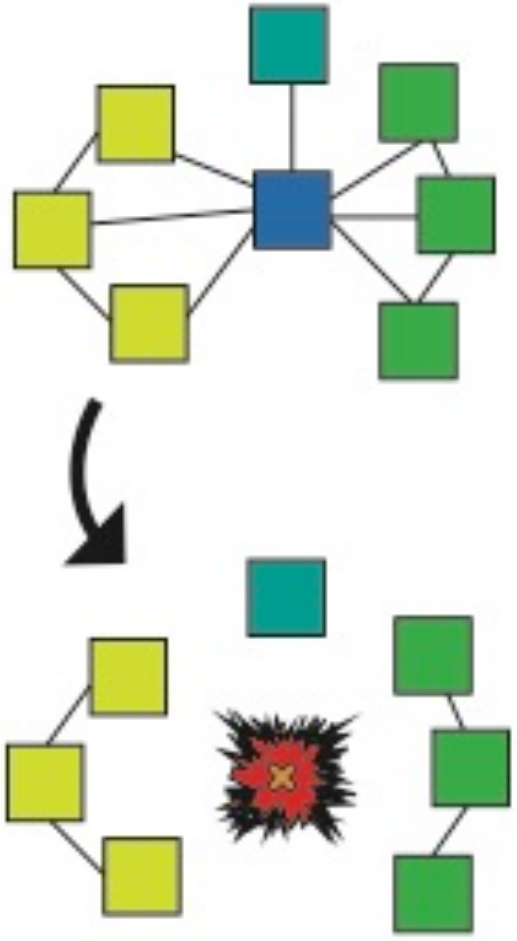
p-value = 2.834e-07

N = 80

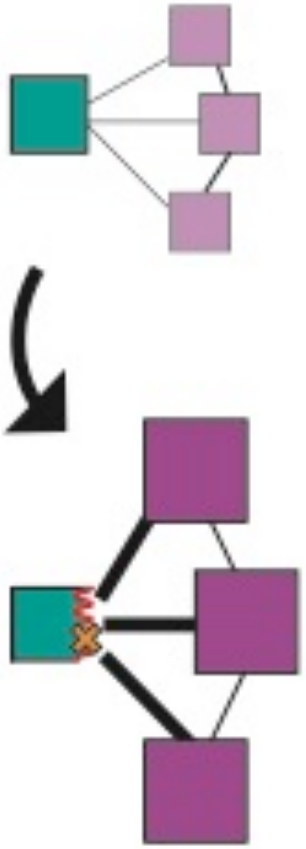
Cancer SNVs & genes (rationalize in TSGs + Oncogenes)

Drilling into potential mechanisms

Naive mechanism for the effects of many **TSG**-associated SNVs **Loss-of-Function Affects**



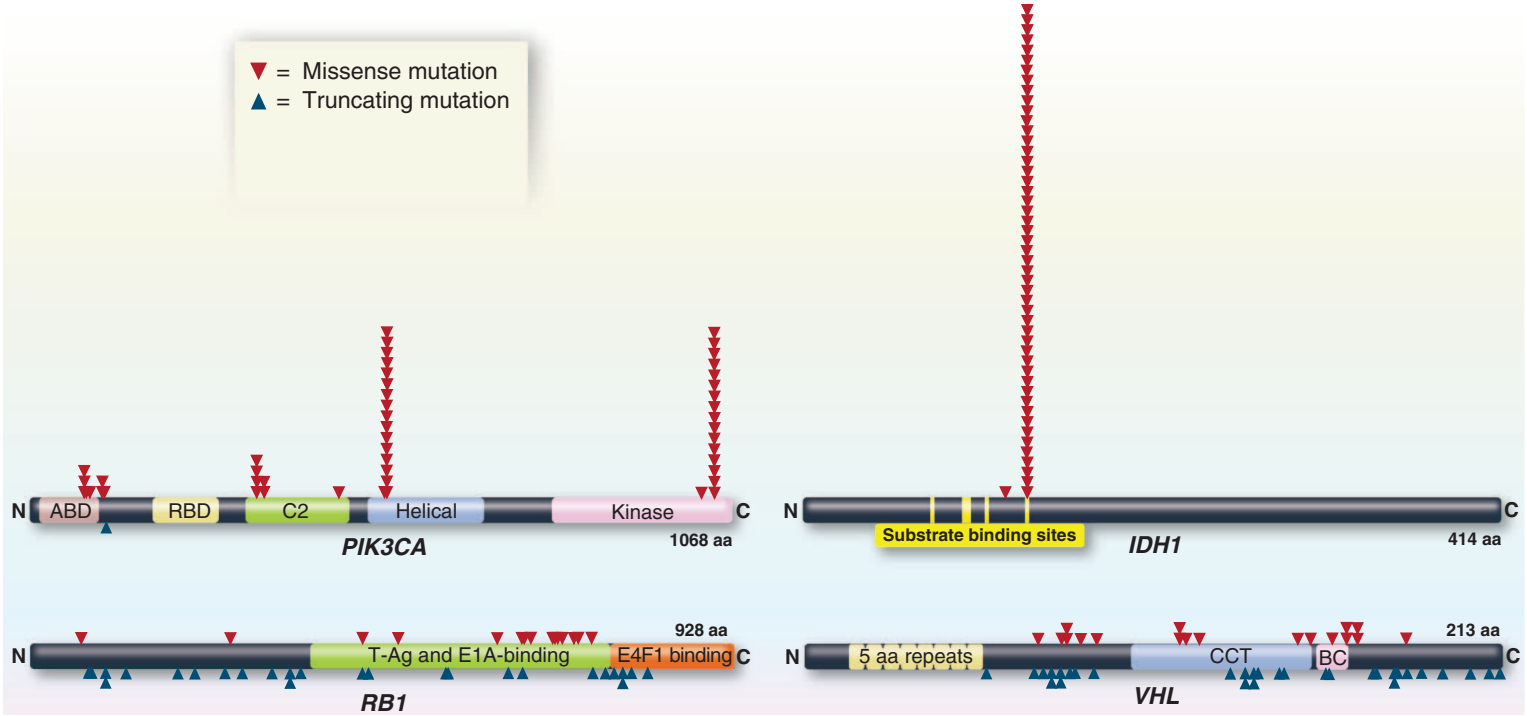
Naive mechanism for the effects of many **oncogene**-associated SNVs **Gain-of-Function Affects**



Cancer SNVs & genes (rationalize in TSGs + Oncogenes)

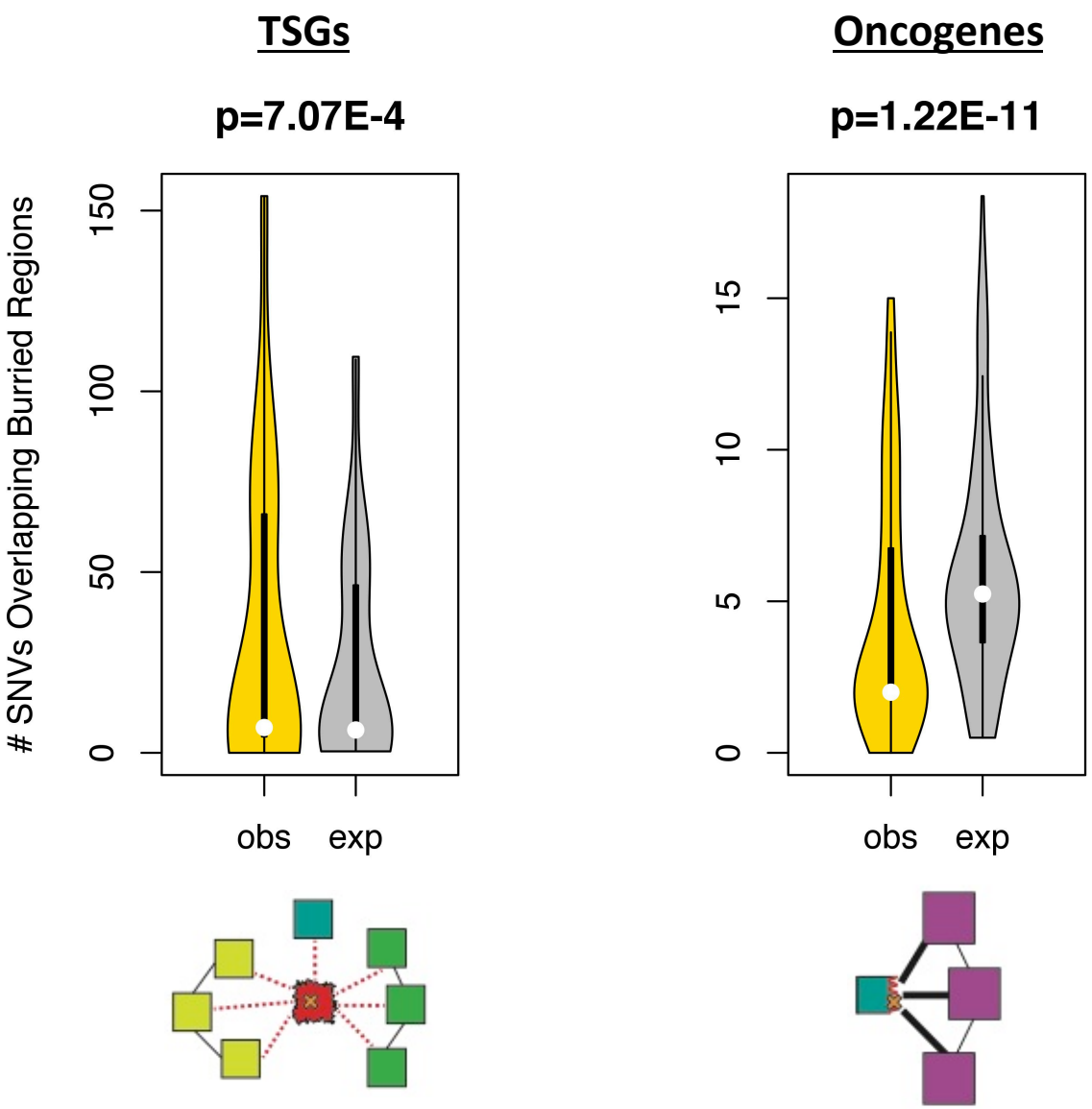
Oncogenes

TSGs

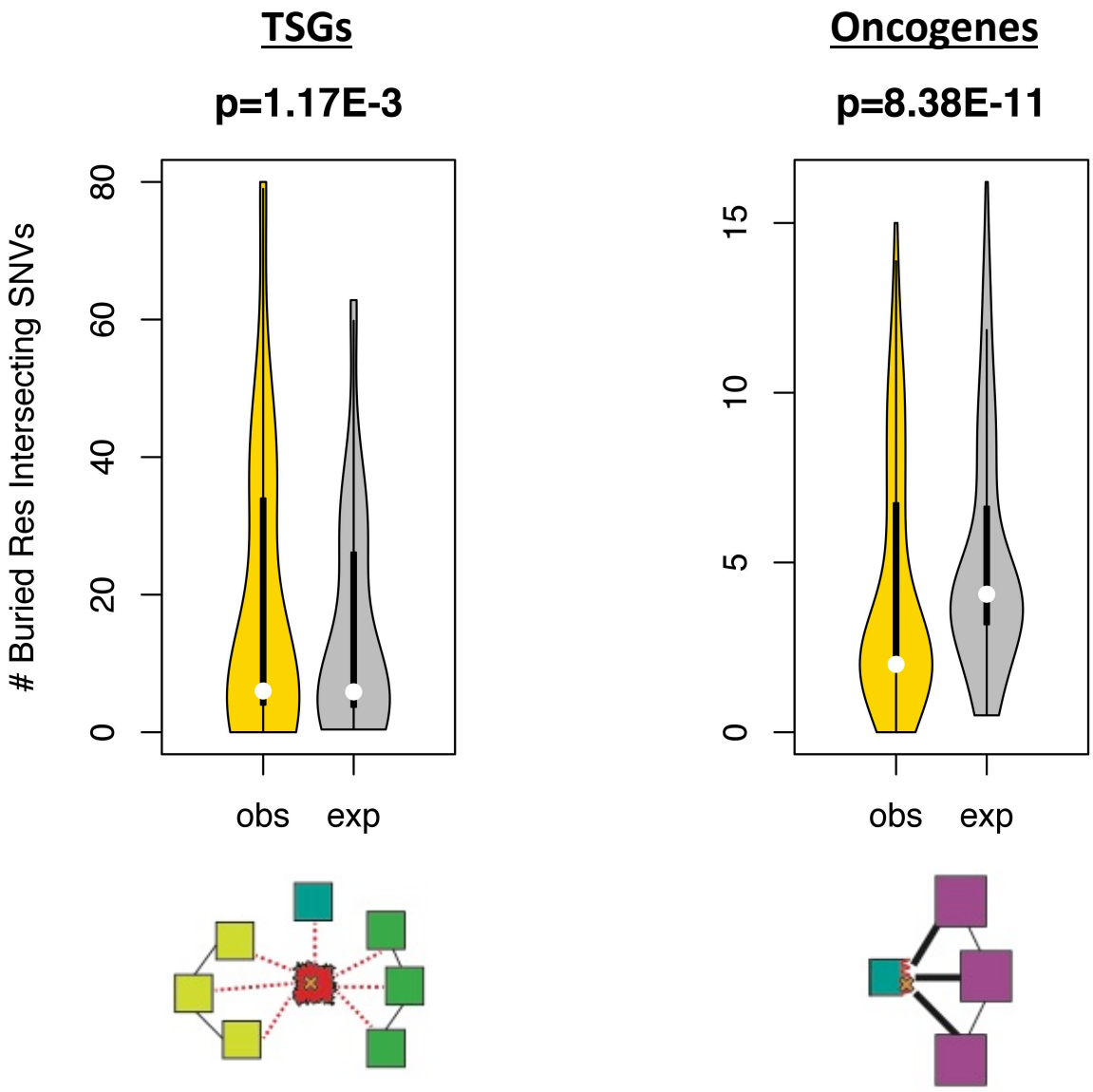


Vogelstein, Bert, et al. "Cancer genome landscapes." Science (2013)

"Redundant" model: Counting the *# of SNVs* that intersect buried regions



"Non-Redundant" model: Counting the *# of buried residues* that intersect cancer-associated SNVs

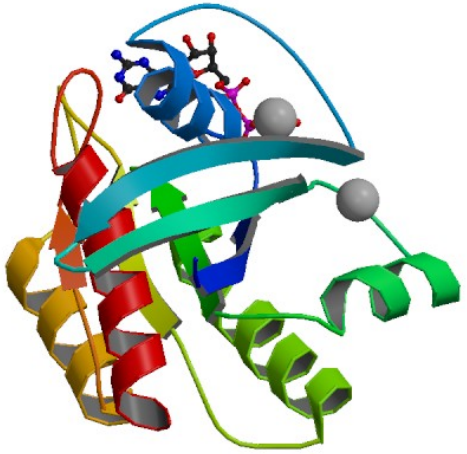


Asymmetric Unit vs. Biological Assembly

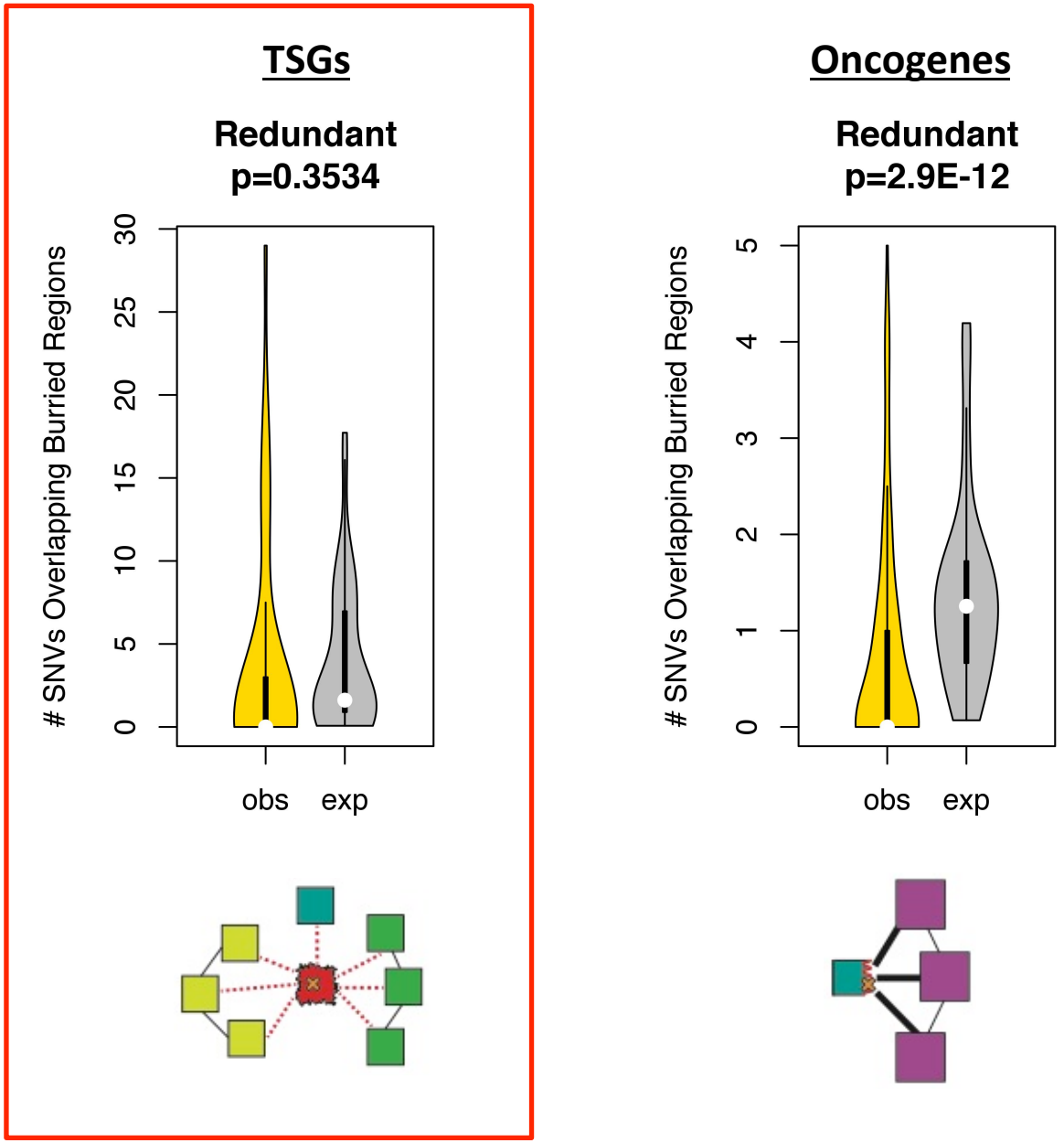
Ex PDB: 3GFT

Asymetric

Bio Assembly



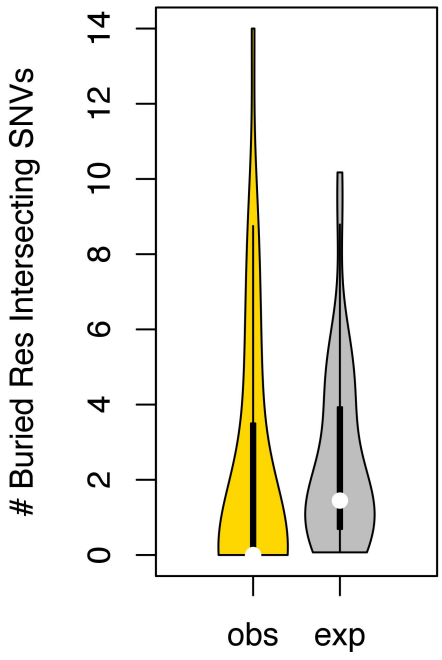
"Redundant" model: Counting (using Bio Assembly Files) the *# of SNVs* that intersect buried regions

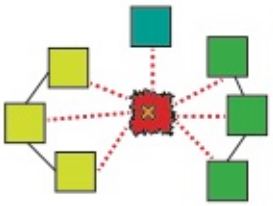


"Non-Redundant" model: Counting (using Bio Assembly Files) the *# of buried residues* that intersect cancer-associated SNVs

TSGs

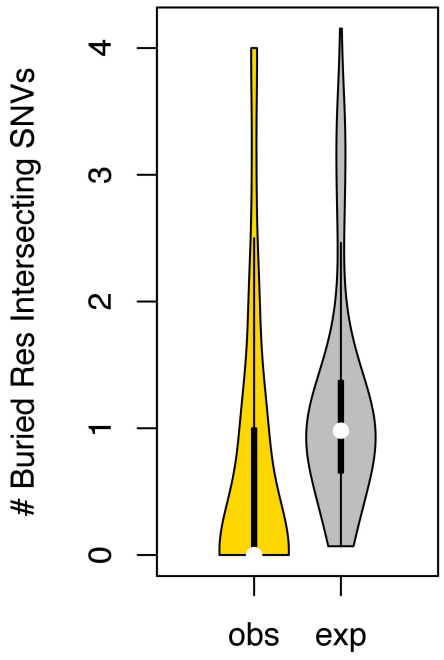
Non-Redundant
p=0.3391

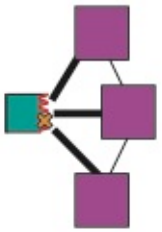




Oncogenes

Non-Redundant
p=3.8E-11

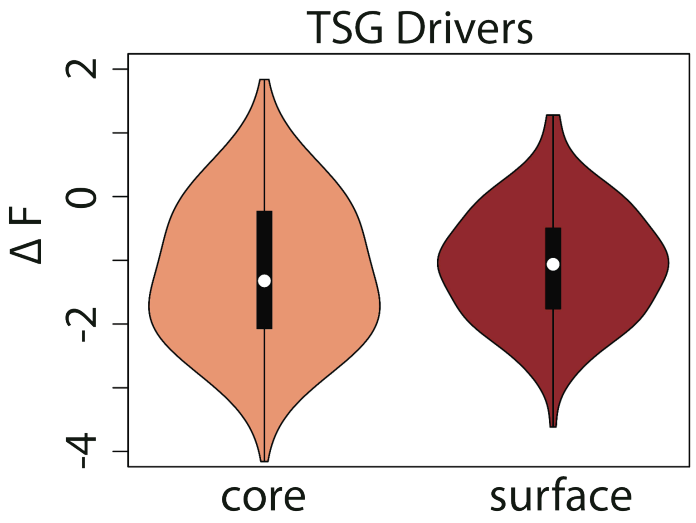


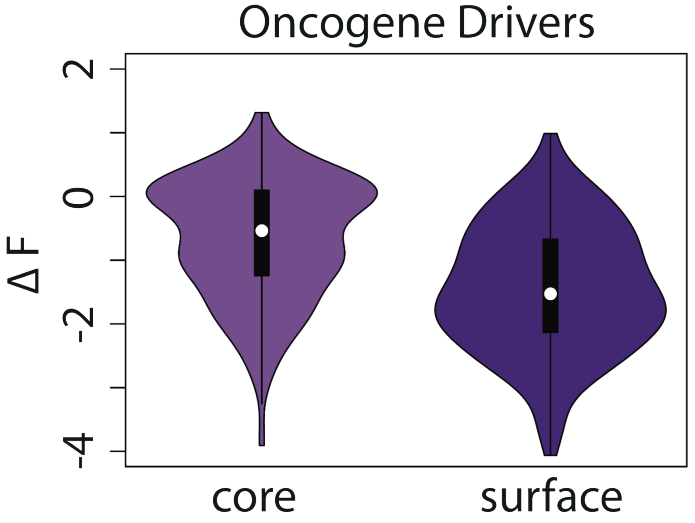


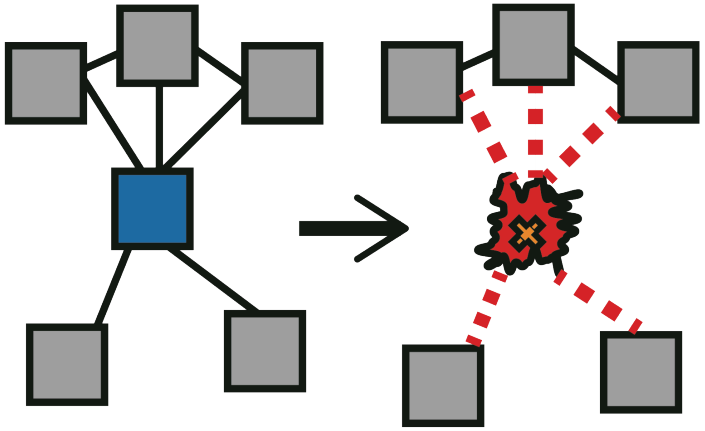
Applications to Cancer-Associated SNVs & Genes

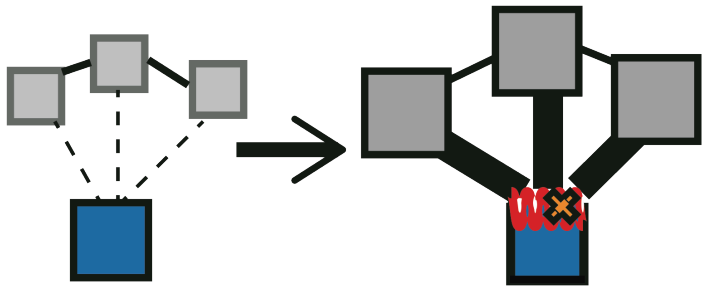
Drilling into potential mechanisms

Frustration is a continuous quantity – go beyond counts & enrichment!









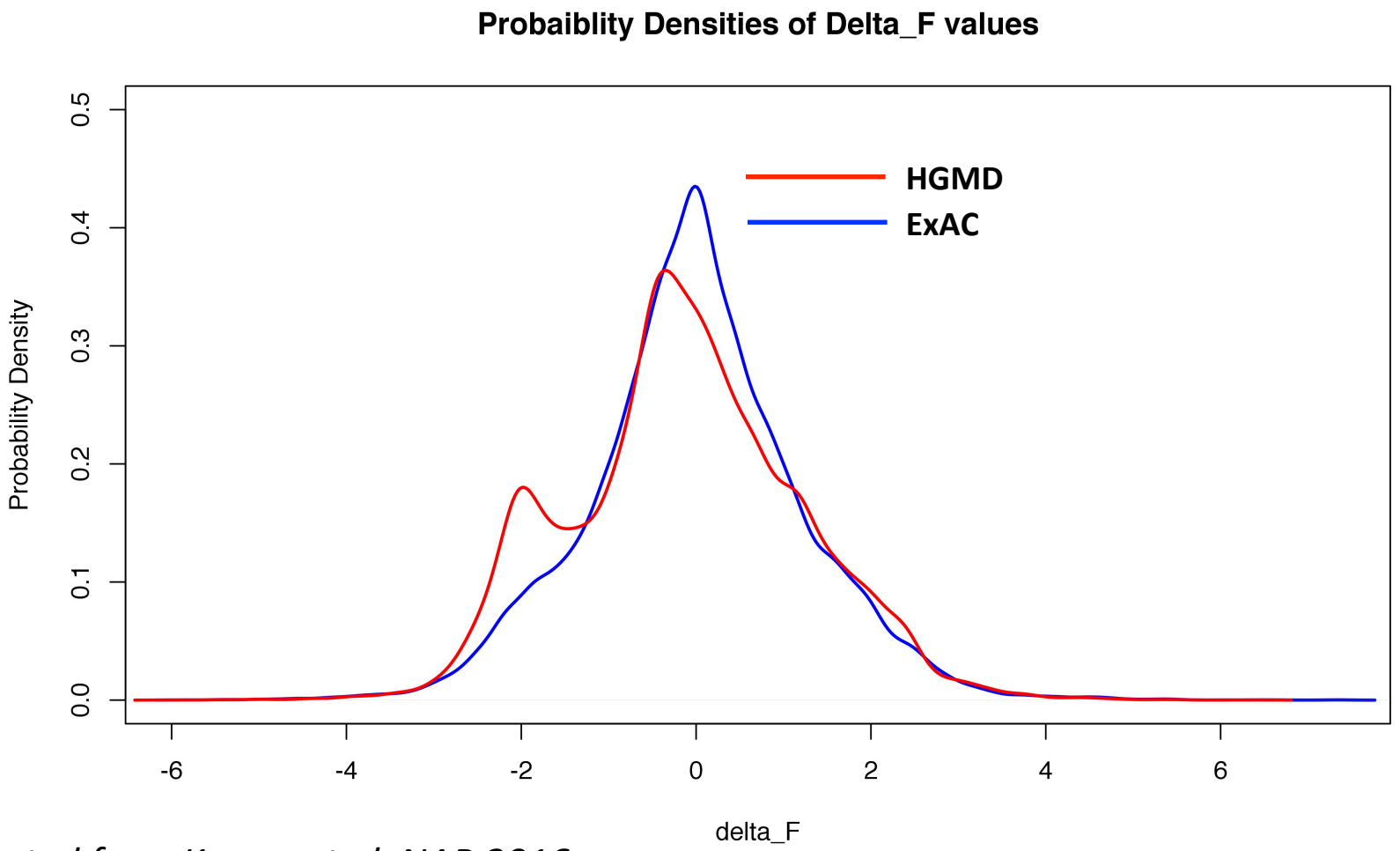
Kumar et al, NAR 2016

Thresholding to classify SNVs

"… how the workflow was applied to variants of unknown significance to help classify/ predict their impact, e.g., **using a certain value of ΔF as a threshold**. This would be extremely valuable and useful for other investigators."

Given an SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious?

HGMD SNVs generally induce more negative ΔF values relative to benign SNVs



Adapted from Kumar et al, NAR 2016

Given an SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious?

The objective is to maximize *f(x)*

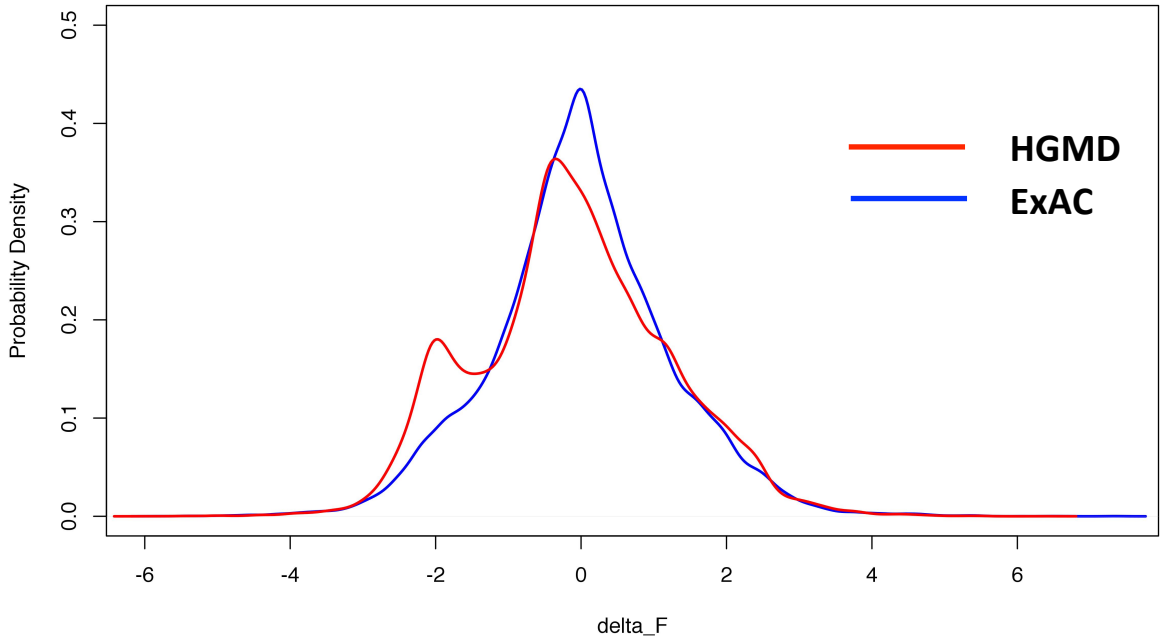
$$f(x) = h(x) + e(x)$$

$$h(x) = fract[\Delta F_{HGMD} < x]) - fract[\Delta F_{HGMD} > x])$$

$$e(x) = fract[\Delta F_{ExAC} > x]) - fract[\Delta F_{ExAC} < x])$$

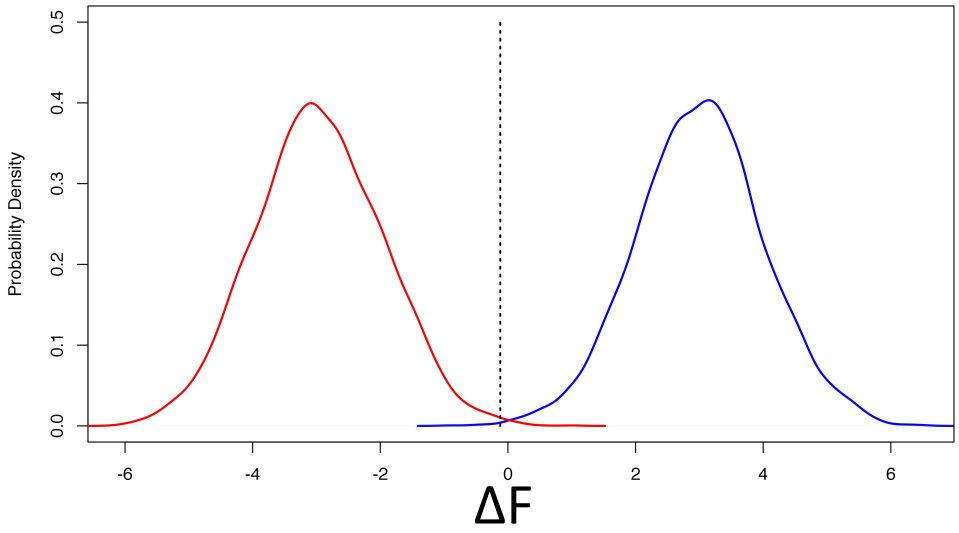
Let ΔF_{HGMD} denote the distribution of ΔF scores induced by HGMD SNVs.

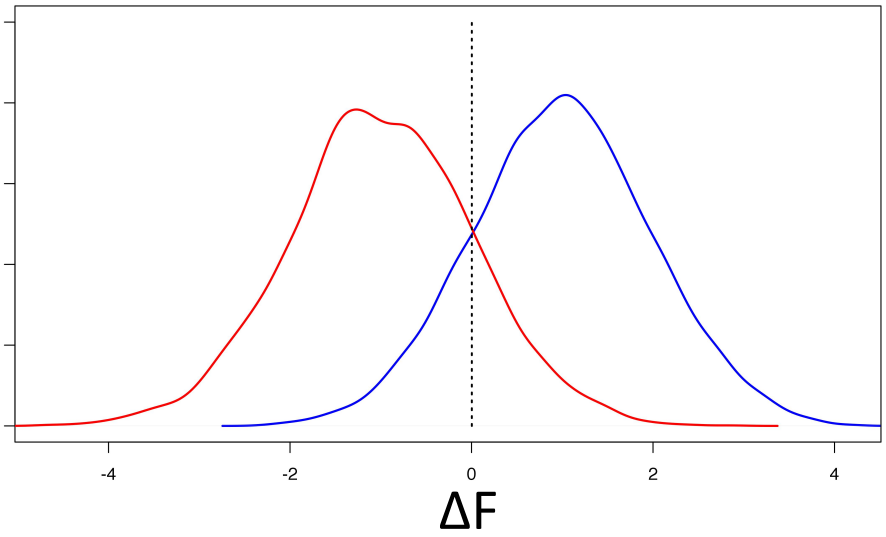
ΔF_{ExAC} is defined for the distribution of ΔF values associated with ExAC SNVs (note the reversed directions relative to the equation above):

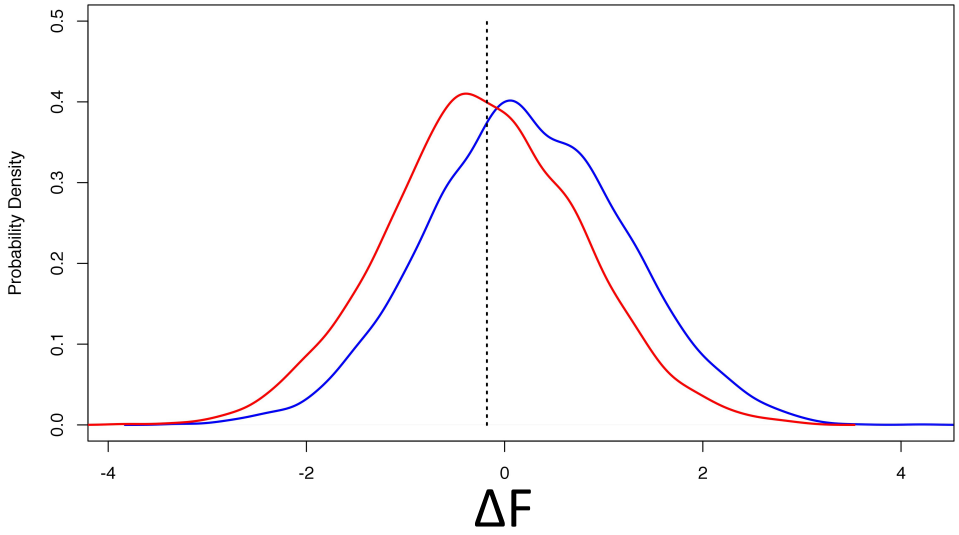


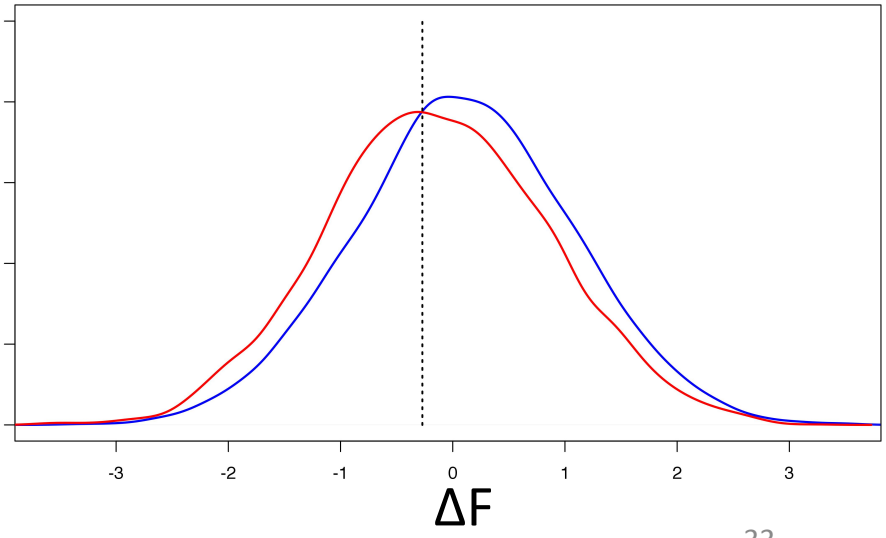
Given an SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious?

Sanity checks on simulated data



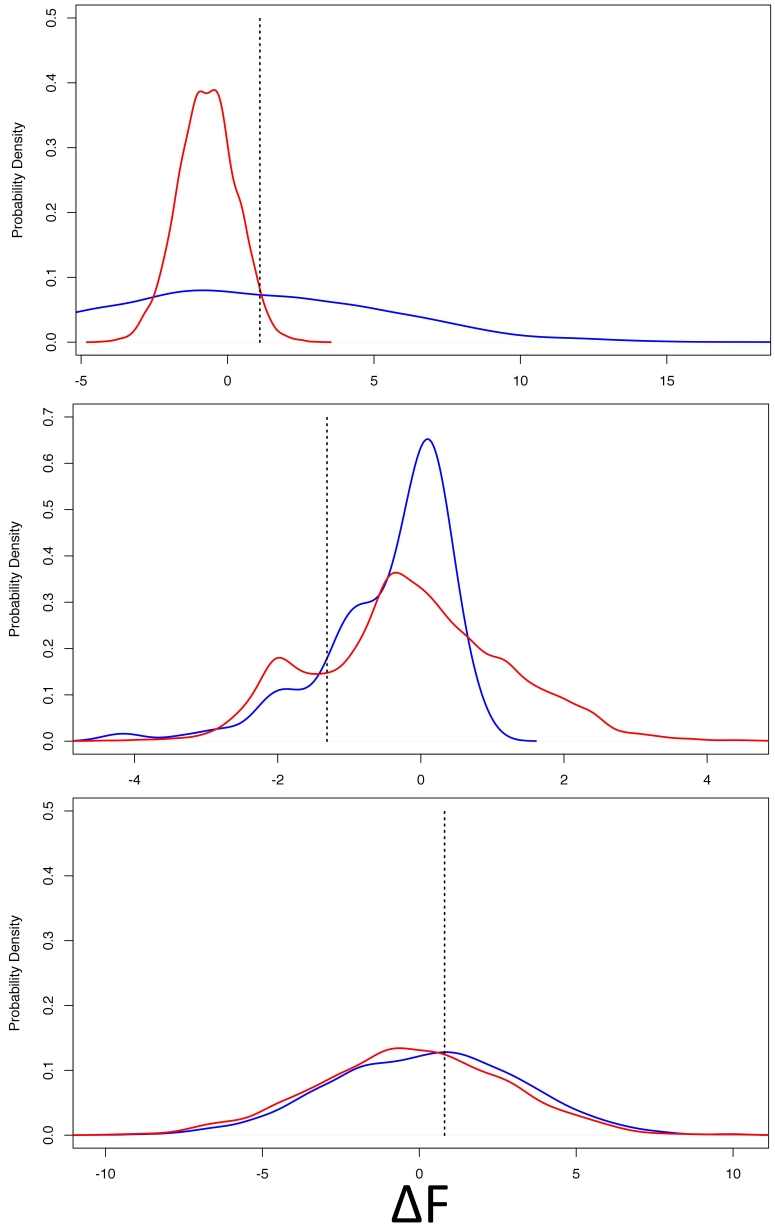




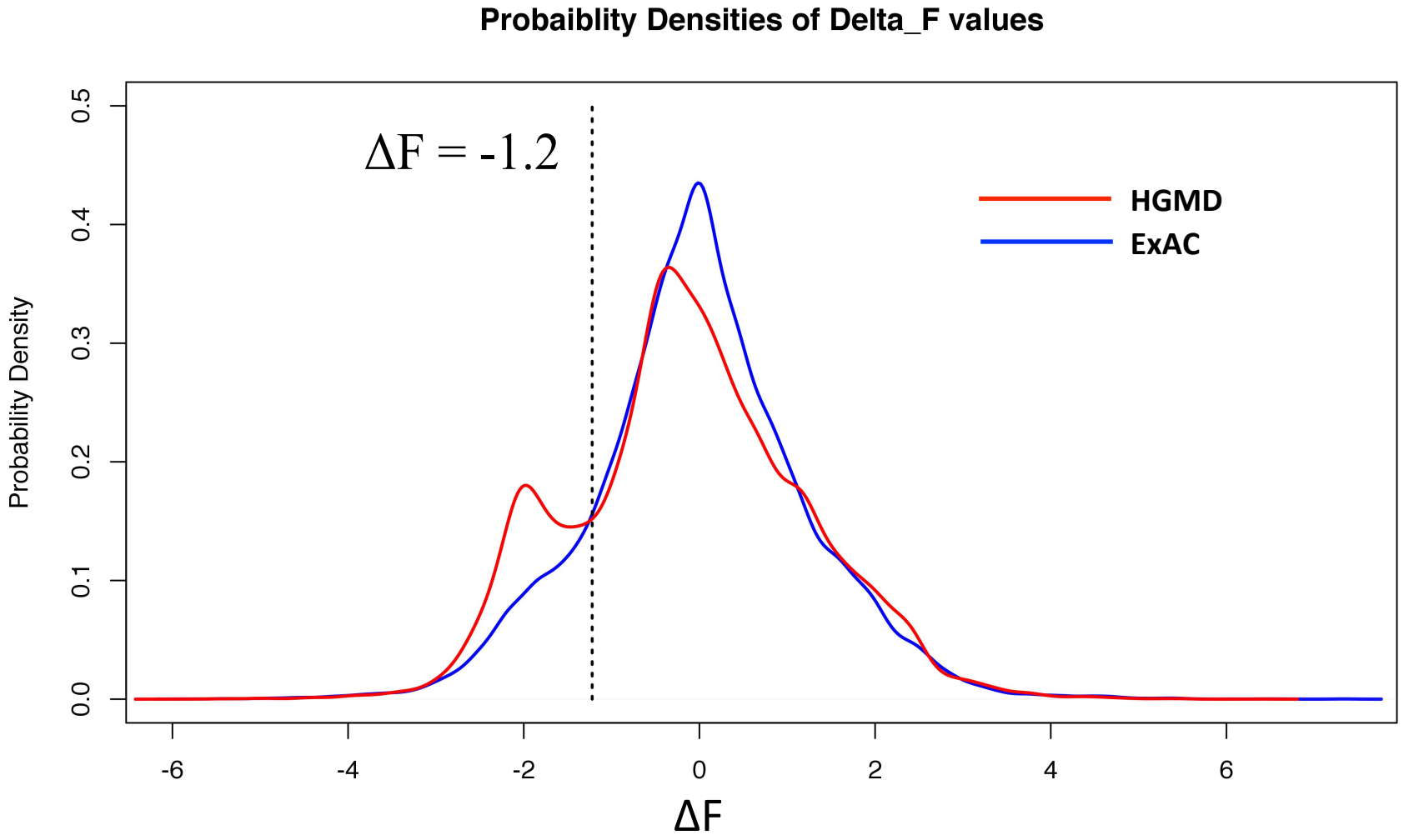


Given an SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious?

Sanity checks on simulated data



Given an SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious?



Kumar et al, NAR 2016

"There are methods existing in order to evaluate potential effects of low-allele-frequency variants in unbiased ways (SIFT, PolyPhen2, MutationTaster, and many others). I would like to see how exactly your method adds up to this ... One could [use] tools to predict the deleteriousness of SNVs (e.g. PolyPhen2 and MutationTaster2) and then check ***if there are disease variants predicted as "harmless" by these tools (i.e. false negative) which are then correctly seen as locally maximal frustrated by your method****..."*

→ Find HGMD SNVs not captured by PolyPhen (yet are captured through frustration)

single chain PDBs

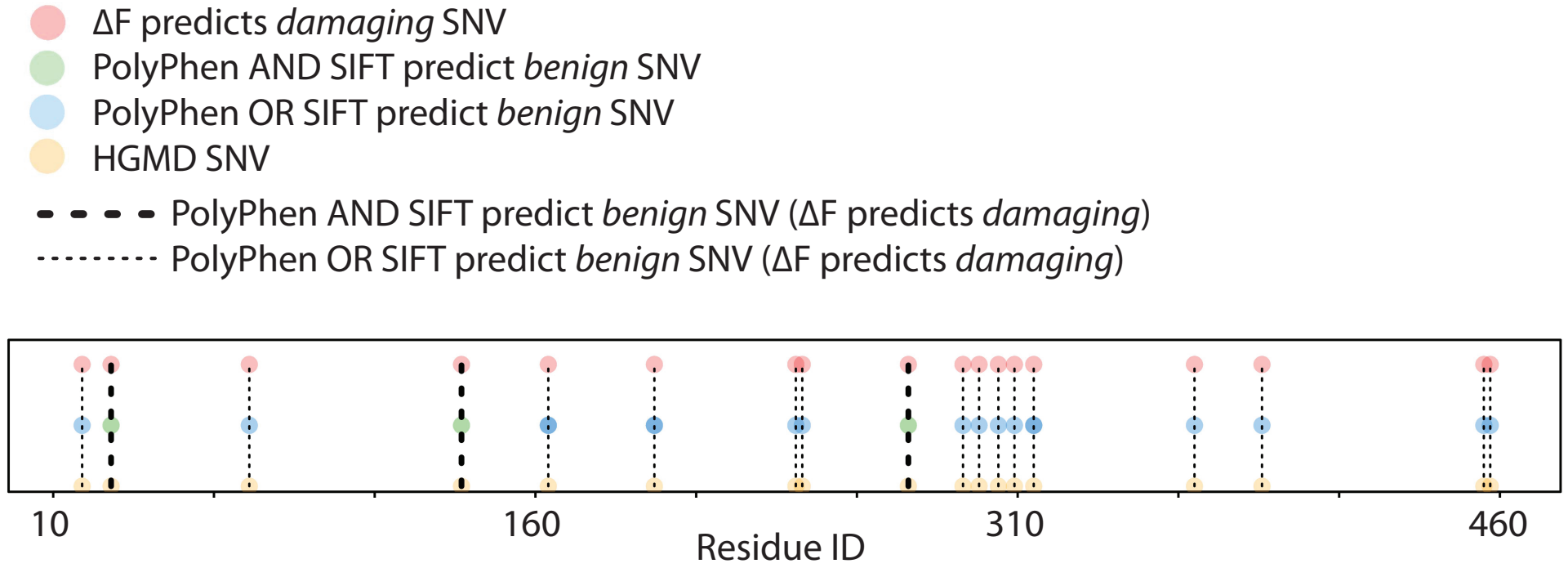
PDB	*# HGMD SNVs*
1T45	2
1V4S	15
1KQ6	1
3PXA	1
1AD6	1
2AMY	1
1OG5	1
2X6U	1

Multi- chain PDBs

PDB	*# HGMD SNVs*
2VGB	2
3GXP	7
1A4I	1
1IIL	1
2O4H	1
3HN3	1

Linearized depiction of HGMD SNVs that constitute ΔF-rescued false negatives

Glucokinase (PDB ID: 1V4S)
SNVs associated with type 2 diabetes



Adapted from Kumar et al, NAR 2016

I. Frustration

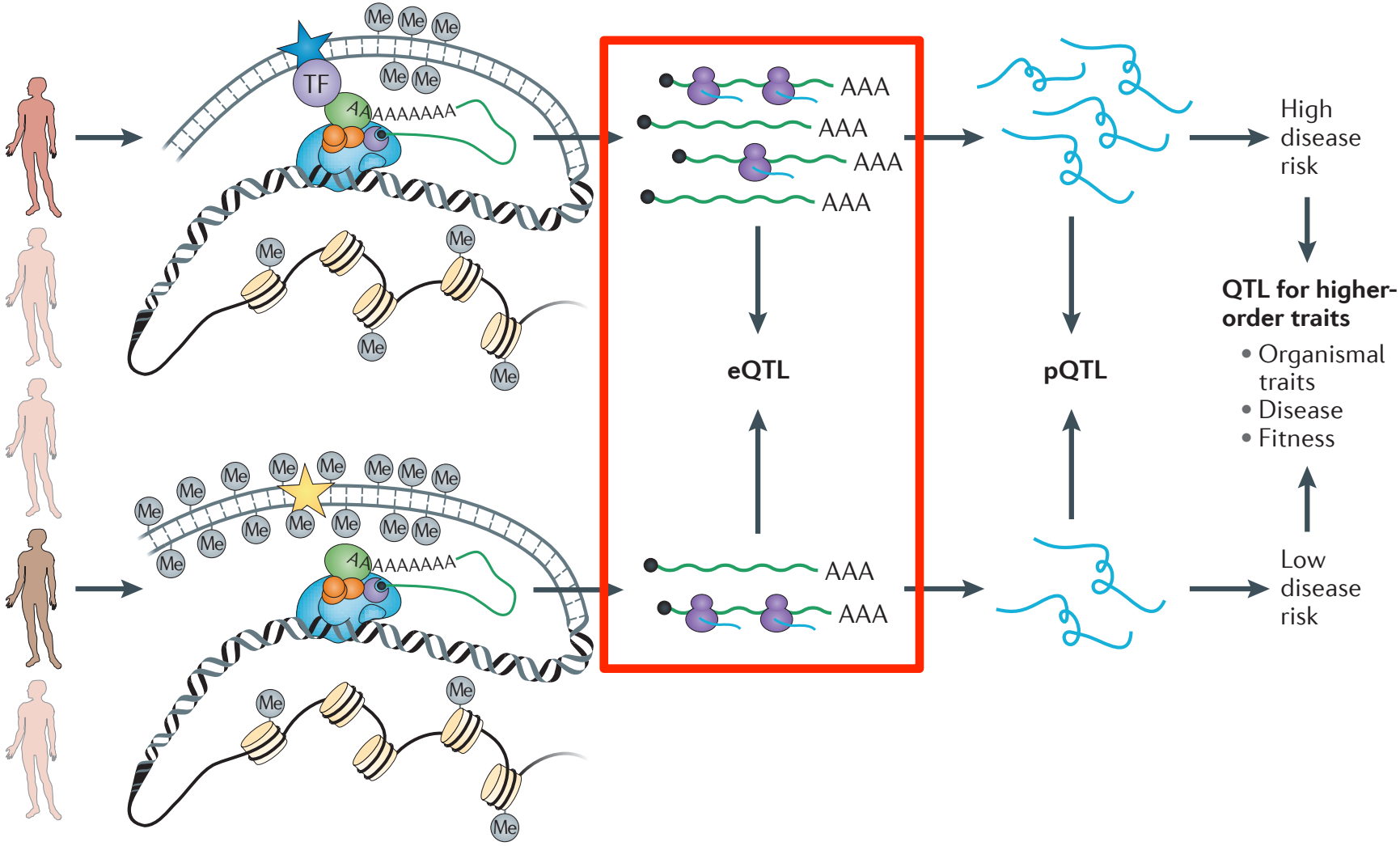
- Background & conceptualization (advantages of secondary calculations)
- Corrected formulation
- Data survey and processing
- MAF analysis (rare alleles associated with extreme delta_F)
- Cancer SNVs & genes (rationalize in TSGs + Oncogenes)
- Thresholding to classify SNVs
- Example case of rescued false negatives: Glucokinase

II. eQTLs

- Background
- Reproducibility in Covariates
- Reproducibility in RPKM
- Framingam data (miRNA-eQTLs)
- Current Objectives

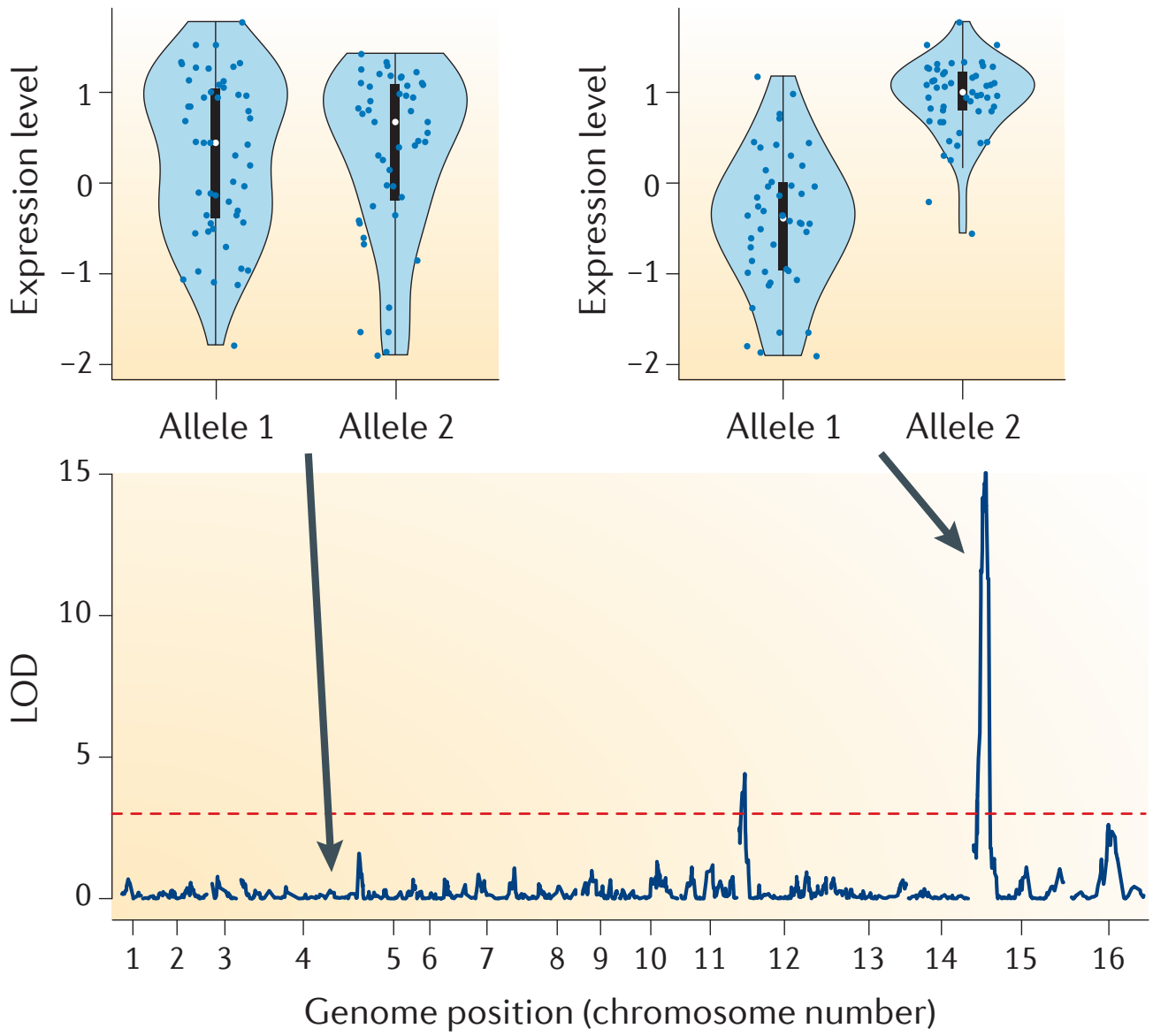
III. Supplementary Slides

Background



Albert and Kruglyak. Nature Reviews Genetics (2015)

Identifying the causal variants in differential gene expression



Albert and Kruglyak. Nature Reviews Genetics (2015)

Reproducibility in Covariates

Gene-level normalized expression matrices (one per tissue)

Reads must:

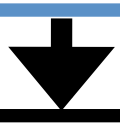
- fall exclusively within exons or span them (i.e. not align into introns)
- contain no more than six non reference bases
- not map equally well to another locus

Genes must:

- have at least 10 samples with
- RPKM > 0.1 and
- raw read counts greater than 6

Covariate correction -- Includes:

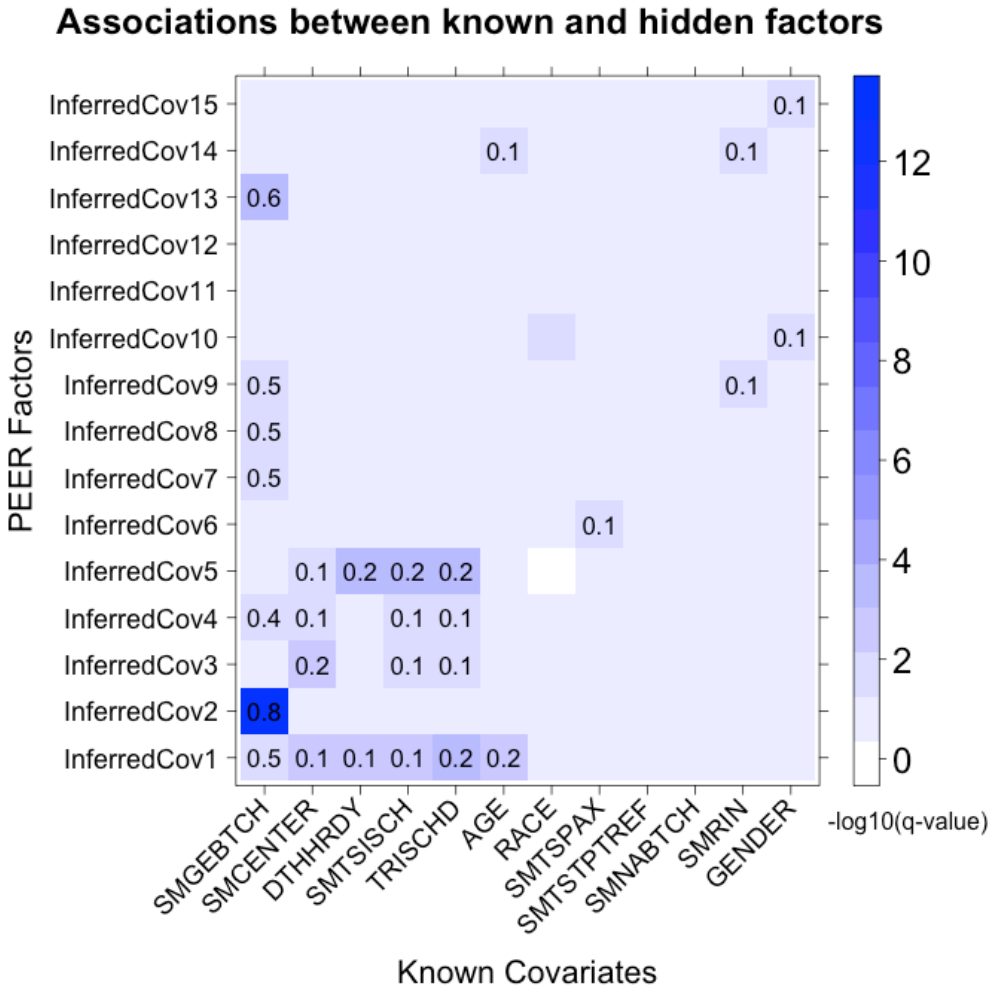
known covariates (ex: gender, genotyping platform)
hidden covariates (PEER factors)



Selecting eQTL (Matri eQTL, FastQTL)

significant gene/snp pairs

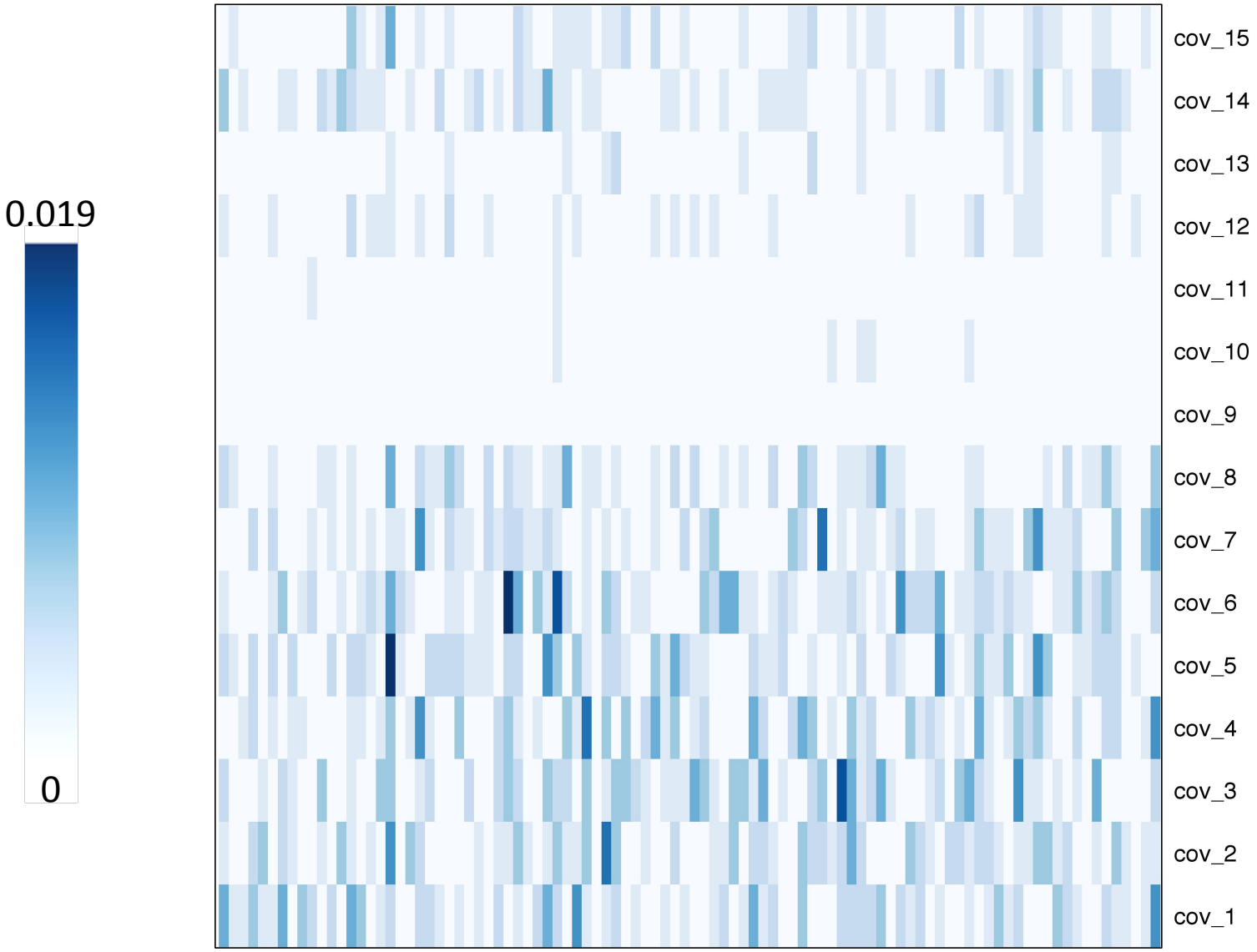
Correlations btwn PEER factors & known covariates (adipose tissue)



Code	Meaning
SMGEBTCH	Expression batch ID
SMCENTER	Collection center
DTHHRDY	Hardy scale
SMTSISCH	Ischemic time for sample
TRISCHD	Ischemic time for individual
AGE	Age of individual
RACE	Self reported race
SMTPAX	Time spent in fixative
SMTSTPTREF	Procurement reference point
SMNABTCH	Nucleic acid isolation batch
SMRIN	RNA quality score (RIN)
GENDER	Gender of individual

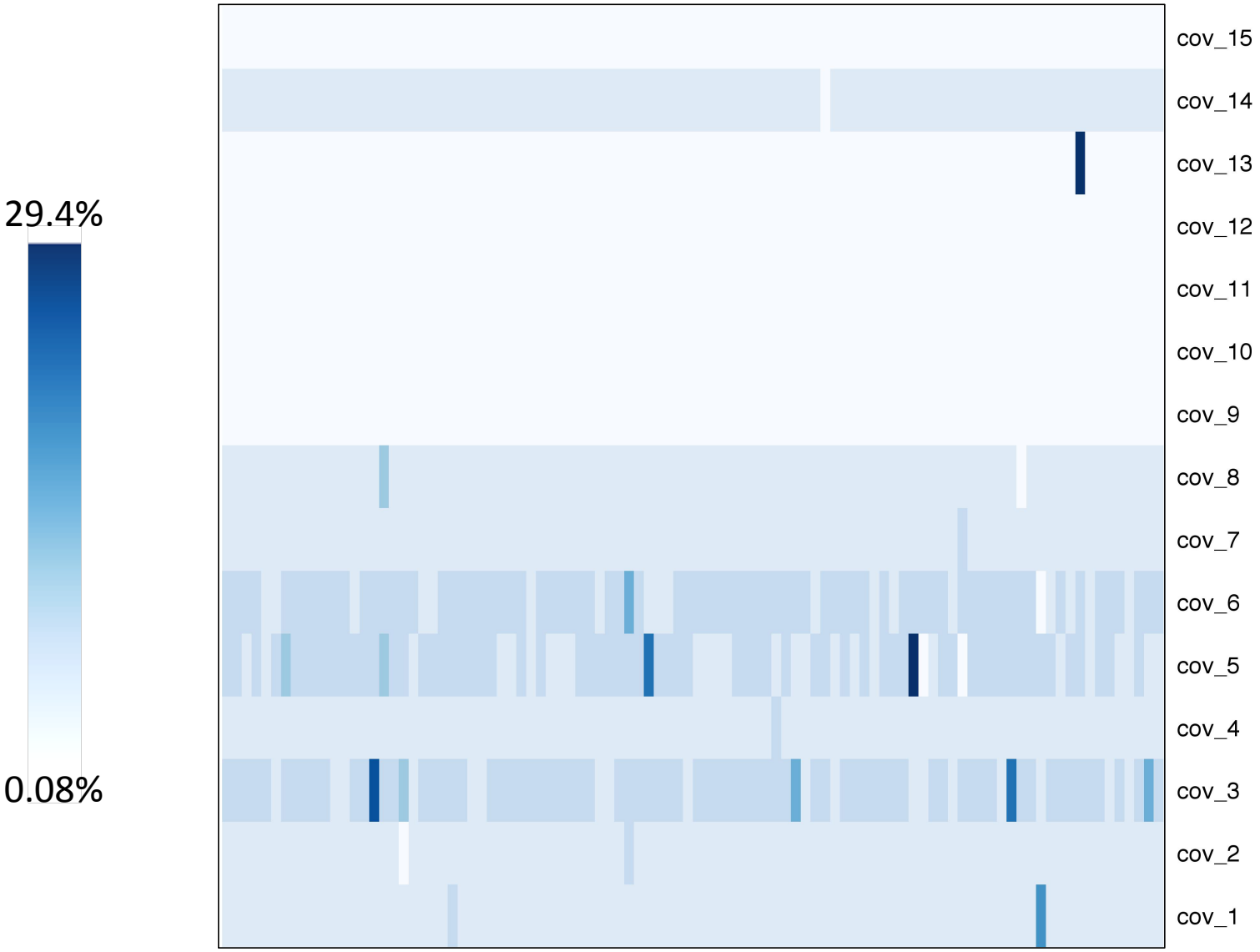
GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." *Science*. (2015)

Absolute deviations: |reported-computed|



96 samples

Percentage deviations



96 samples

Potential Confounding Factors

PEER factors are generated using the top 1000 expressed genes per tissue

PEER version differences (not specified in original paper)

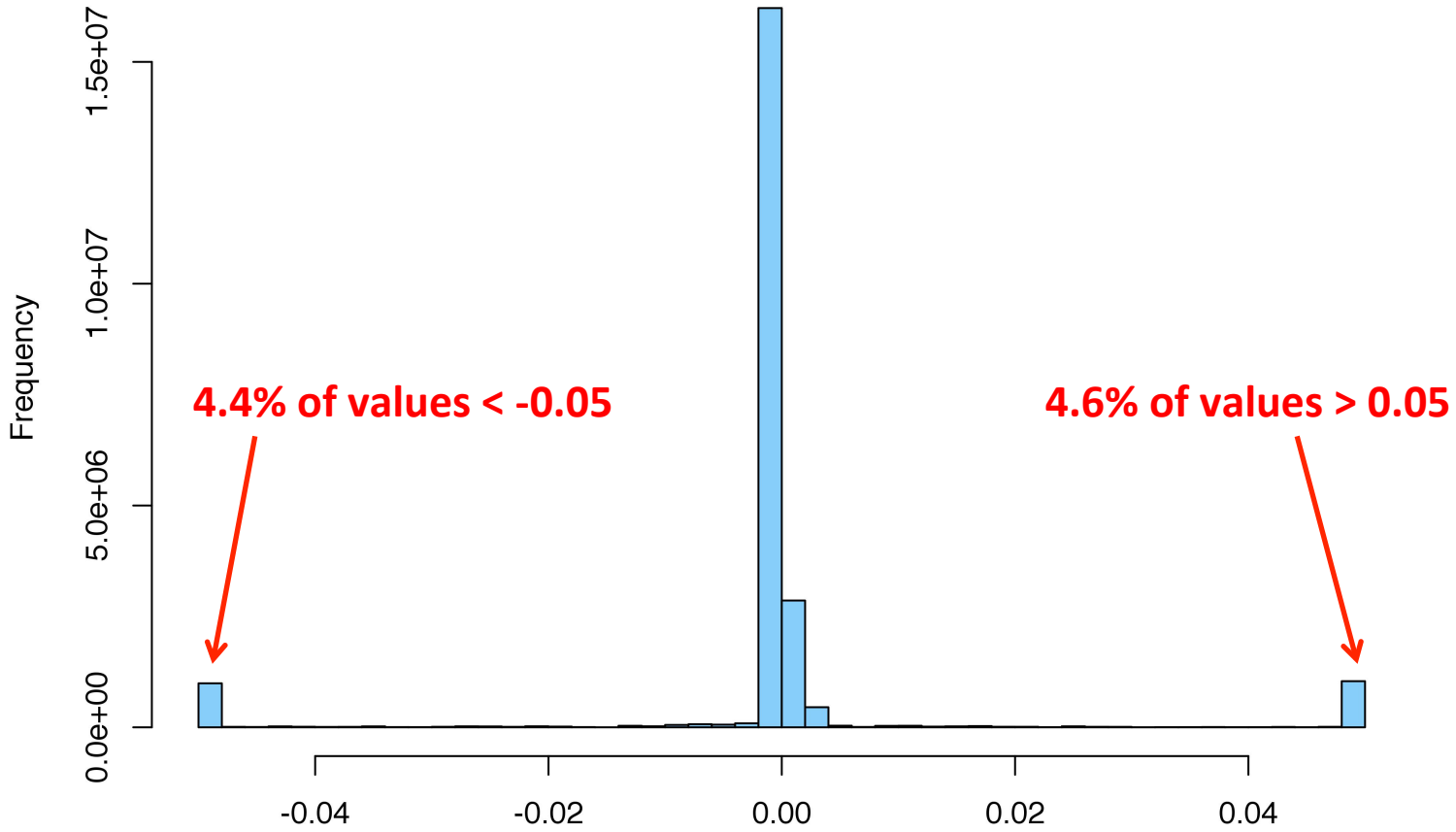
Known covariates somehow included?

Parameters -- gamma distributed for noise & weight factors not reported (a black-box!)

Number of PEERs is determined by N (number of samples per tissue)

Reproducibility in RPKM

Distribution of all fractional errors*



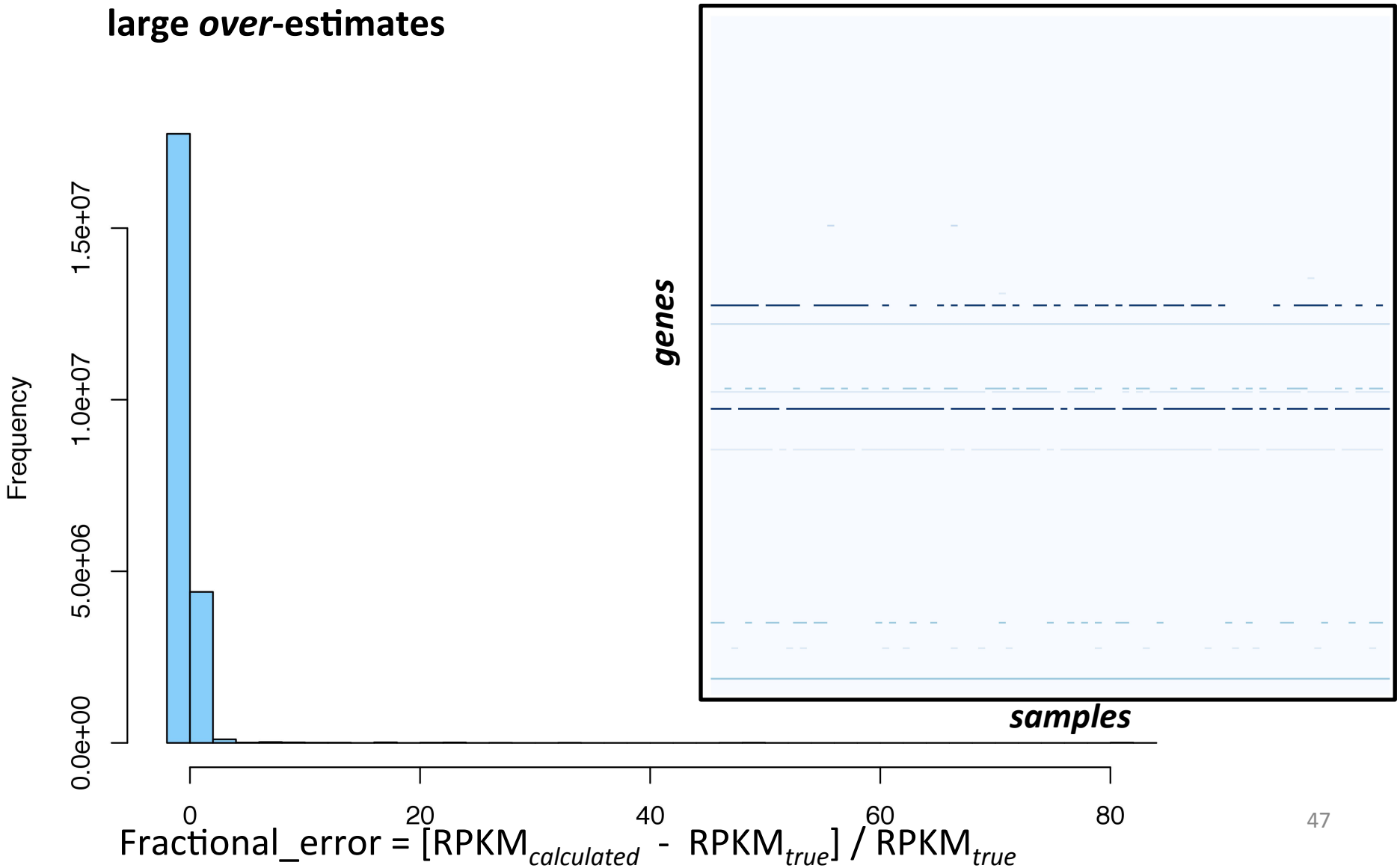
Fractional_error = [$RPKM_{calculated}$ - $RPKM_{true}$] / $RPKM_{true}$

*for single-exon genes

Reproducibility in RPKM

The substantial errors in calculated values are very large *over*-estimates

Specific genes (not samples) account for disparities



Framingham data (miRNA-eQTLs)

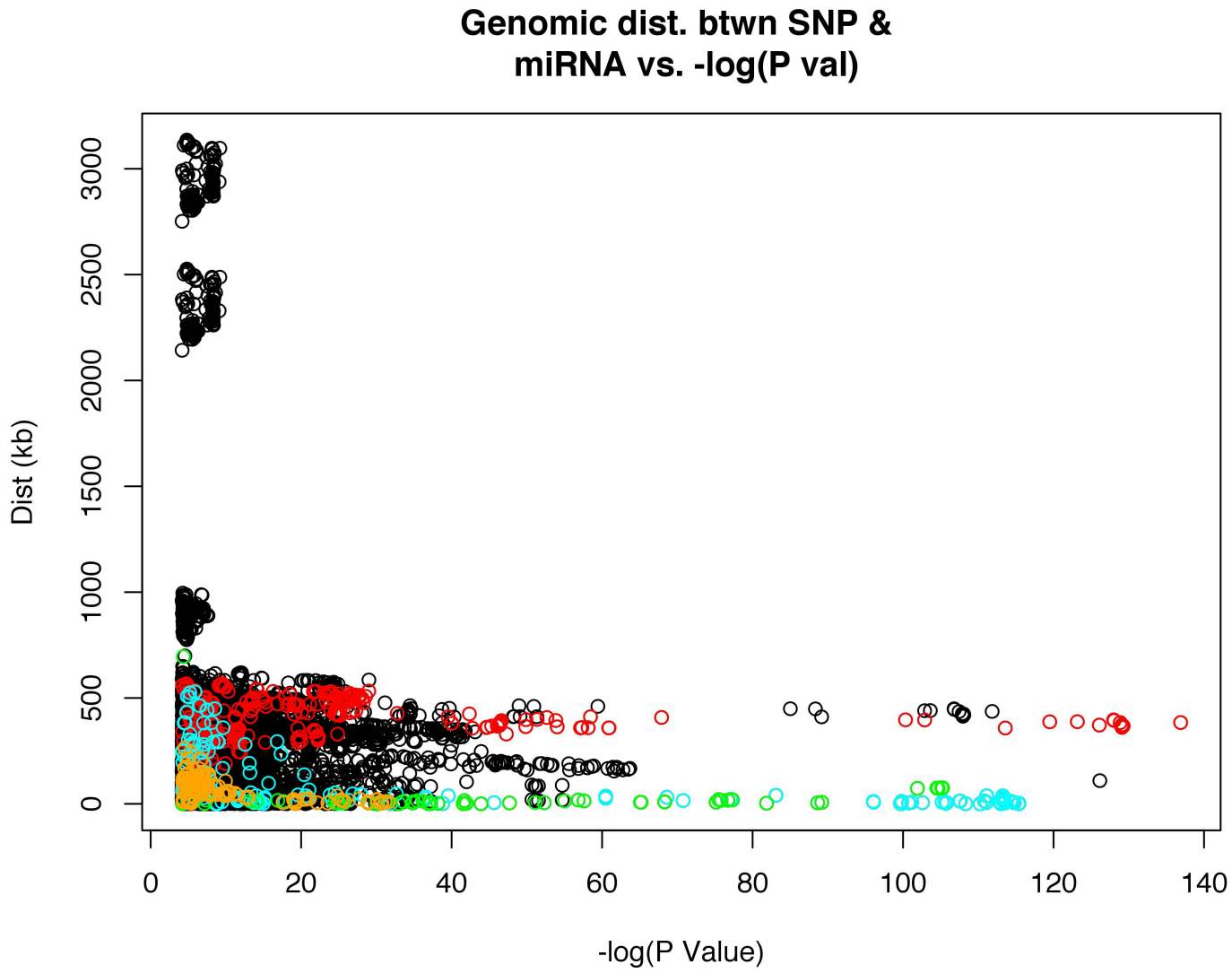
Available attributes for each miR-eQTL (SNV-miRNA pair)

(5,269 cis-miR-eQTLs for 76 mature microRNAs)

- snpID
- miRNA_FHS
- sample size
- **beta**
- MAF
- **Tvalue**
- **Pval**
- **h2q**
- BH_FDR
- chr.SNP
- **SNP.pos**
- SNP.strand
- SNP.func (ex: intron)
- Chr.miR
- **miR.Start**
- miR.End
- miR.strand
- hsa_miR_name
- CisMark (ie: cis or trans)
- miRNA_alter_ID
- miR_Type* (ex: "intron" or "Intergenic")
- mutated base
- wt base
- abs_dist_btwn_SNP_and_miRNA(kb)

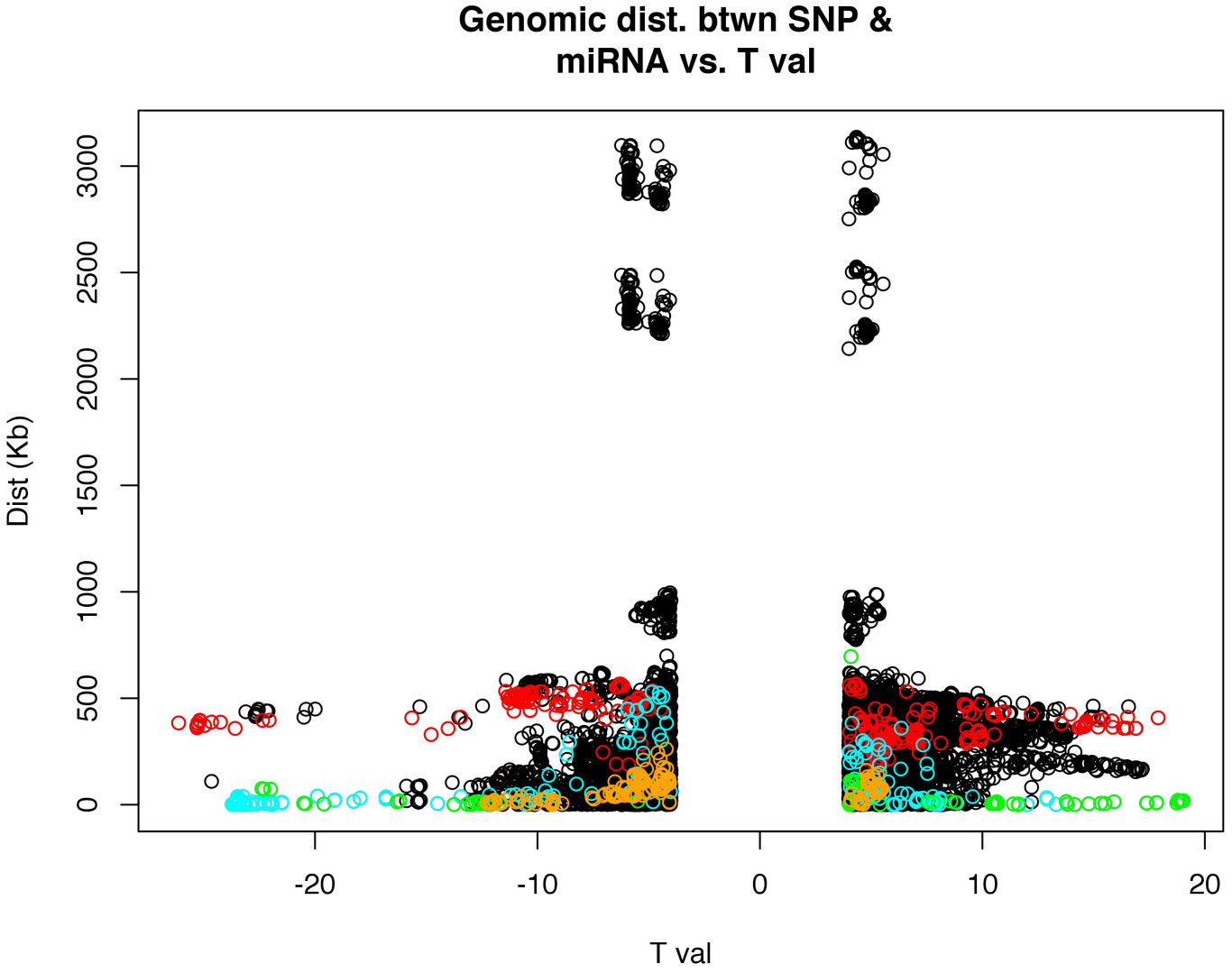
Huan, Tianxiao, et al. "Genome-wide identification of microRNA expression quantitative trait loci." Nature communications 6 (2015).

Framingham data (miRNA-eQTLs)



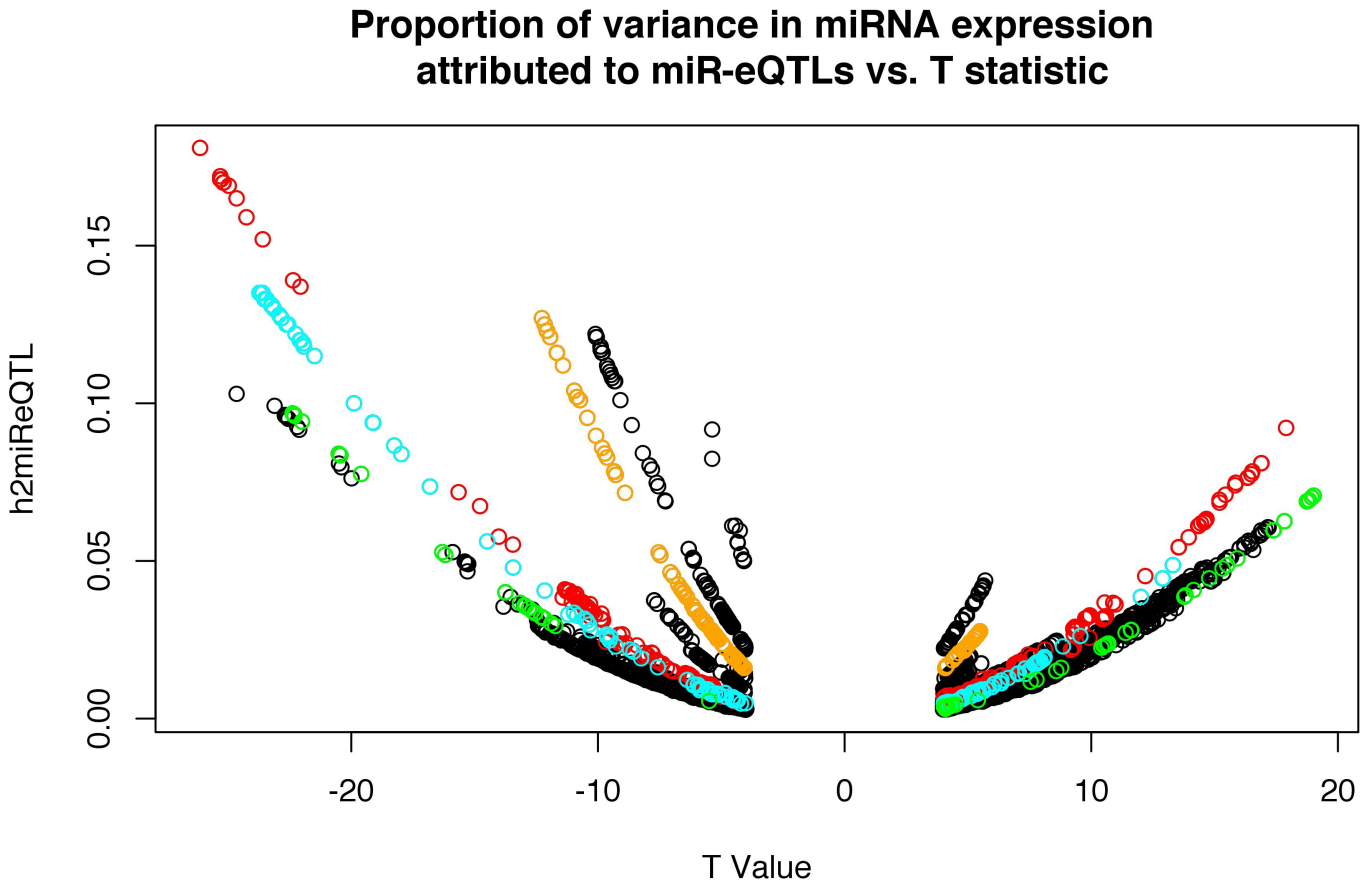
miR_100_5p miR_1303 miR_133a miR_30a_3p

Framingham data (miRNA-eQTLs)



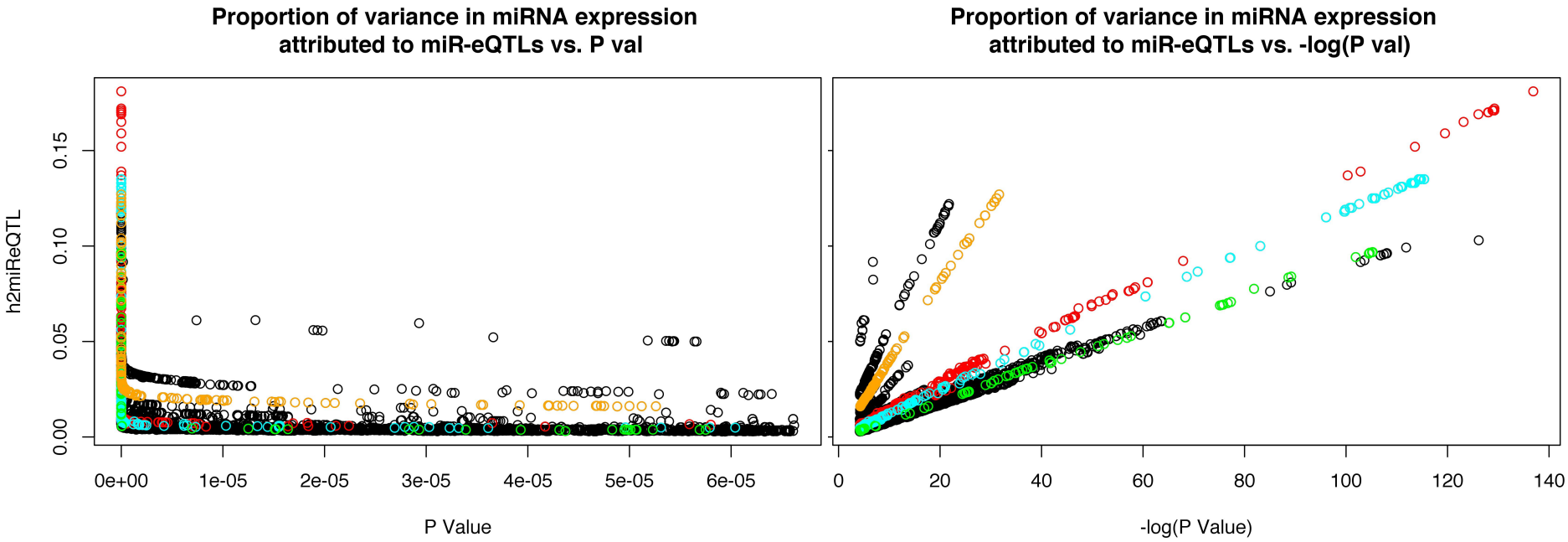
miR_100_5p miR_1303 miR_133a miR_30a_3p

Framingham data (miRNA-eQTLs)



miR_100_5p miR_1303 miR_133a miR_30a_3p

Framingham data (miRNA-eQTLs)

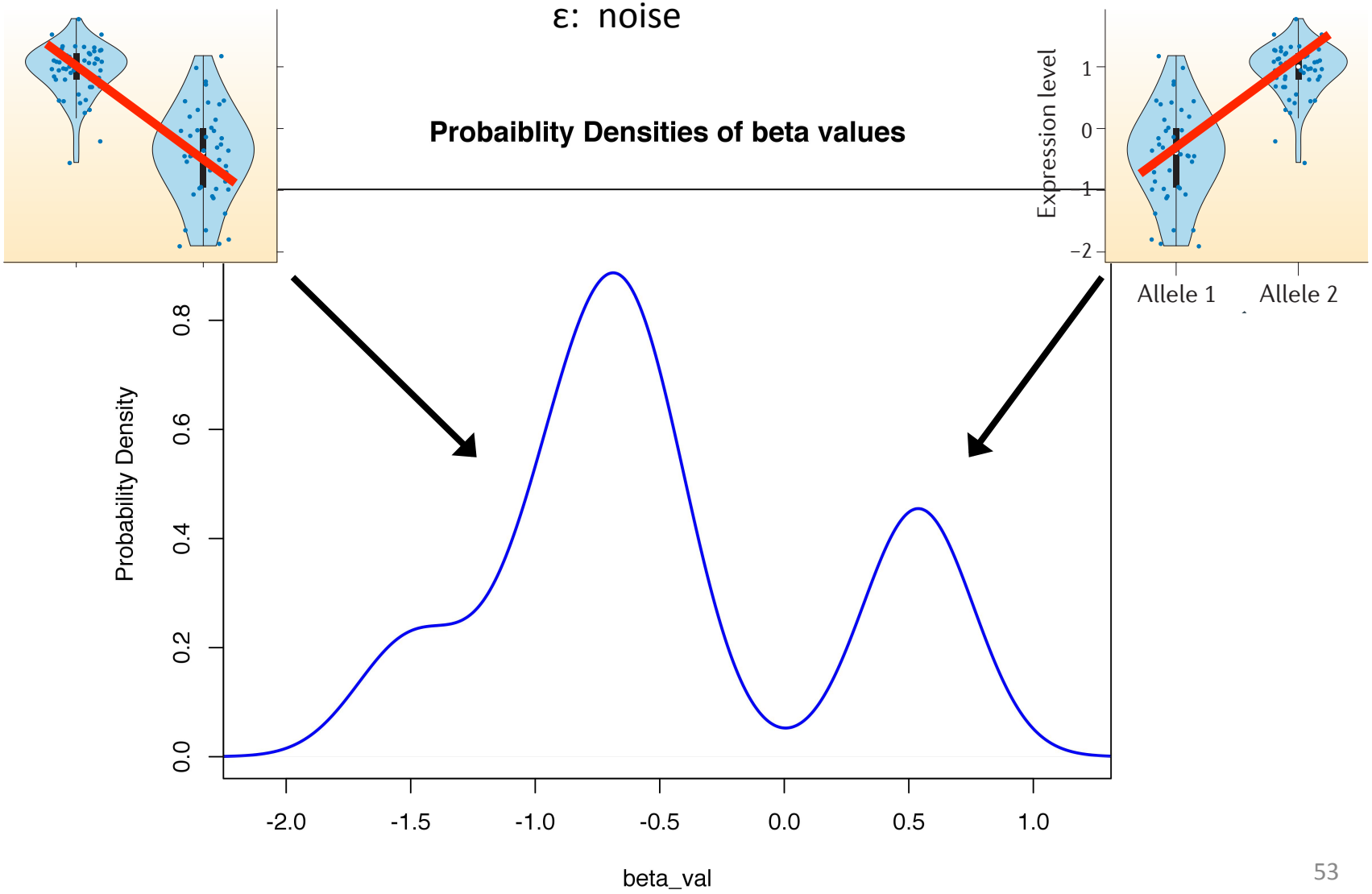


miR_100_5p **miR_1303** **miR_133a** **miR_30a_3p**

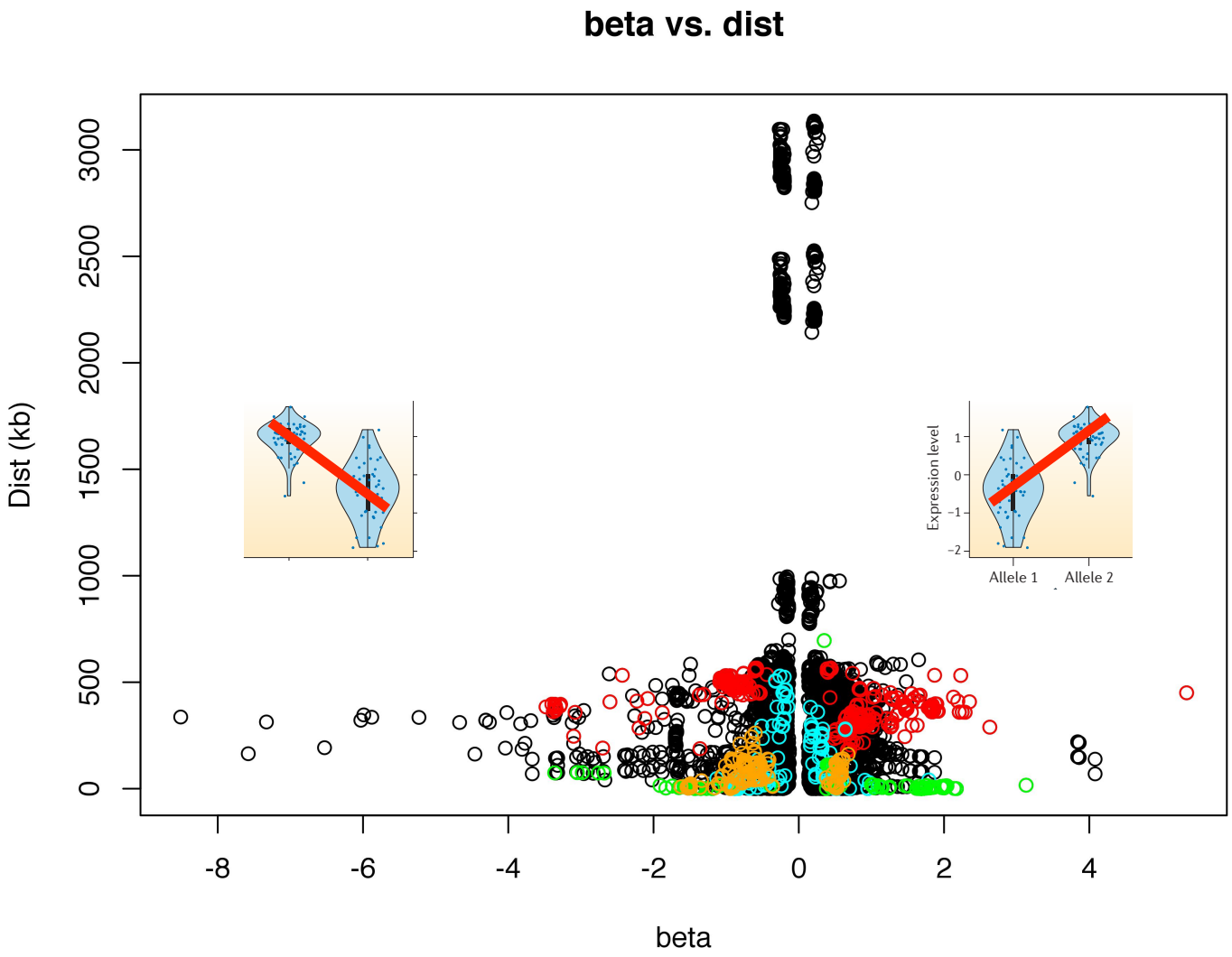
Model w/Simple Linear regression

$$g = \alpha + \beta s + \varepsilon$$

g: gene expression
s: genotype
ε: noise

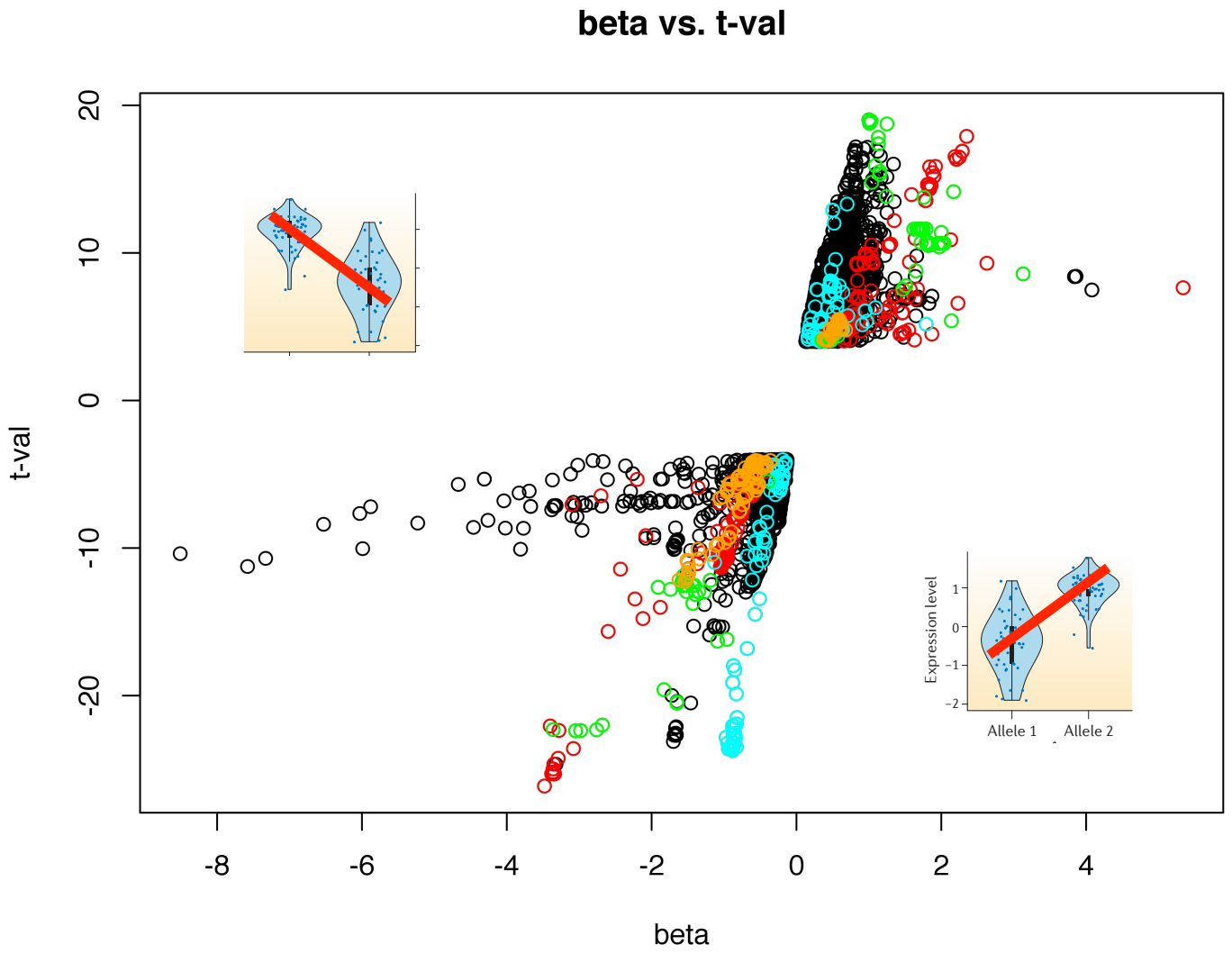


Framingham data (miRNA-eQTLs)



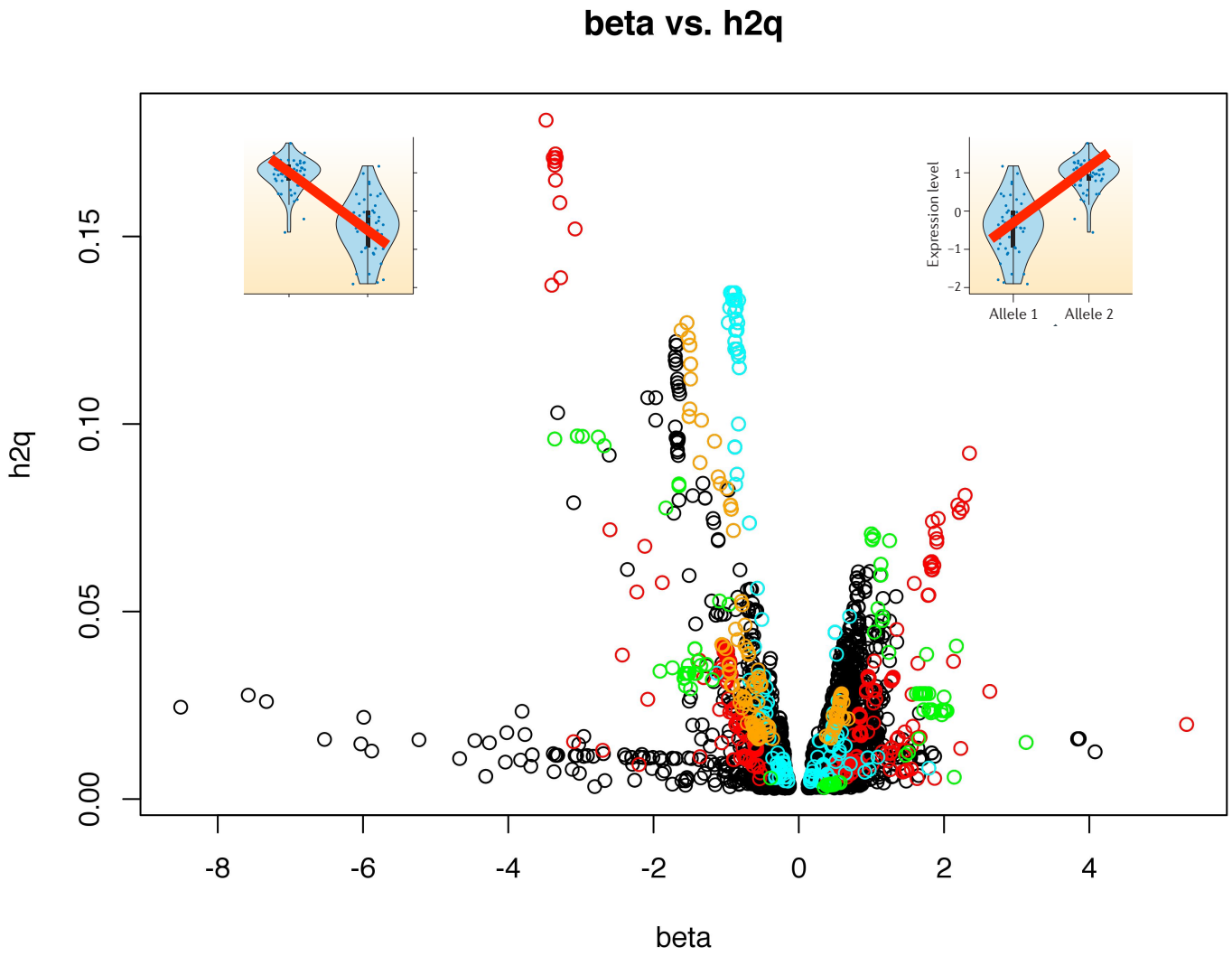
miR_100_5p miR_1303 miR_133a miR_30a_3p

Framingham data (miRNA-eQTLs)



miR_100_5p miR_1303 miR_133a miR_30a_3p

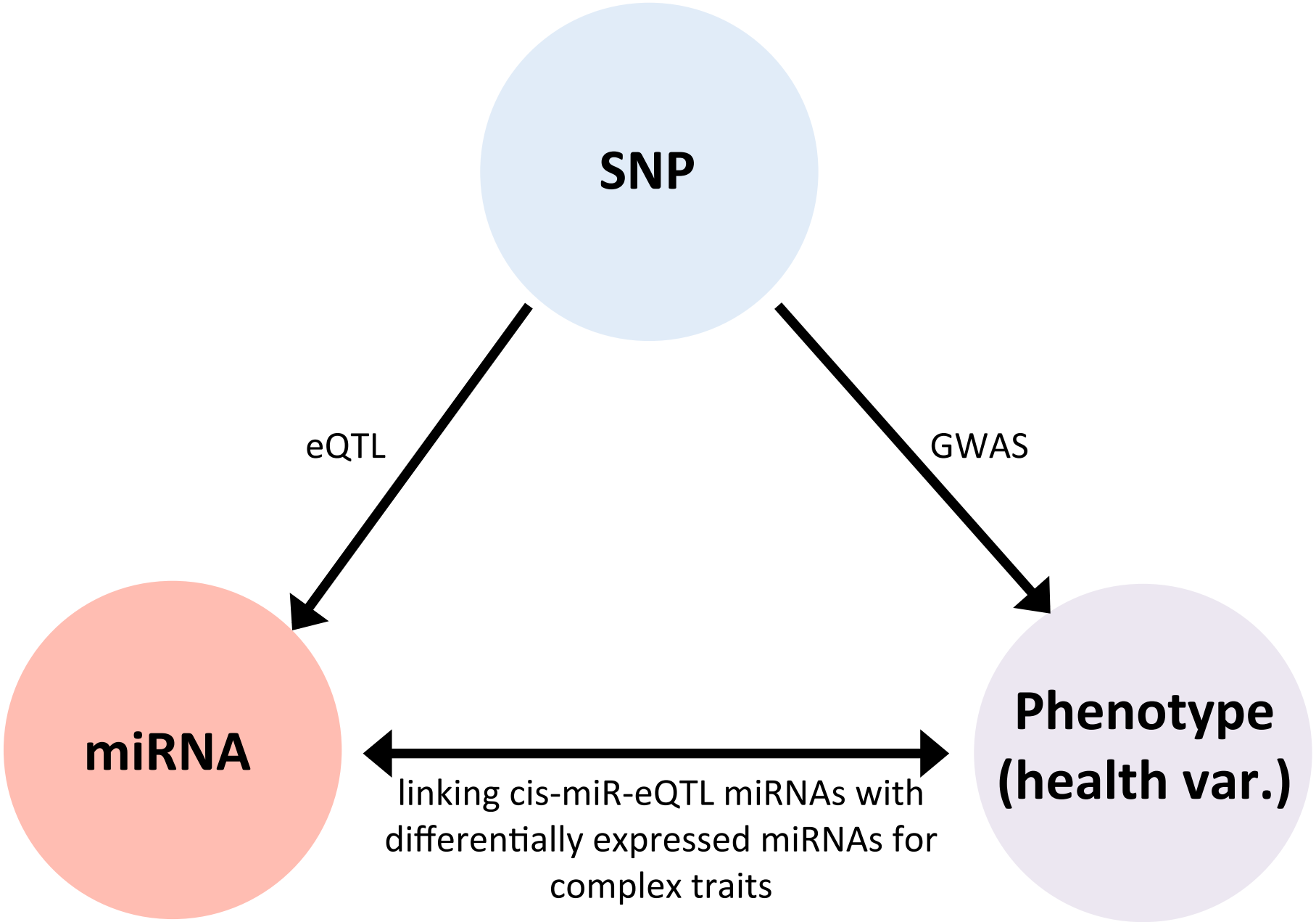
Framingham data (miRNA-eQTLs)



miR_100_5p miR_1303 miR_133a miR_30a_3p

Current Objectives (re eQTLs items)

> in brain group: commonMind bam files ---[htseq]---> raw reads
> HZ currently implementing k-means on Framingham data
> hear back from J. Freedman re. IRB
> hear back from RM
> networks analysis (eqtls constitute edges between miRNA & SNV)
> GTEx
    + consistent w/gtex (cis-only, and under different models)?
    + GO enrichment of affected genes
    + narrow in on specific cases (bio annotations – HZ)


SNP
eQTL
GWAS
miRNA
Phenotype
(health var.)
linking cis-miR-eQTL miRNAs with
differentially expressed miRNAs for
complex traits

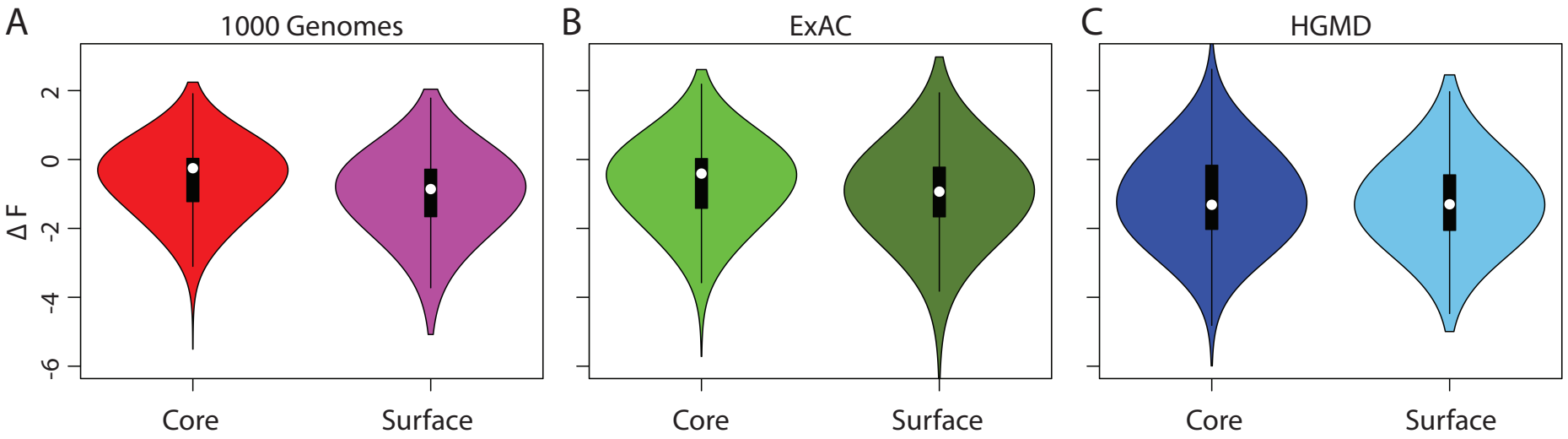
Acknowledgements

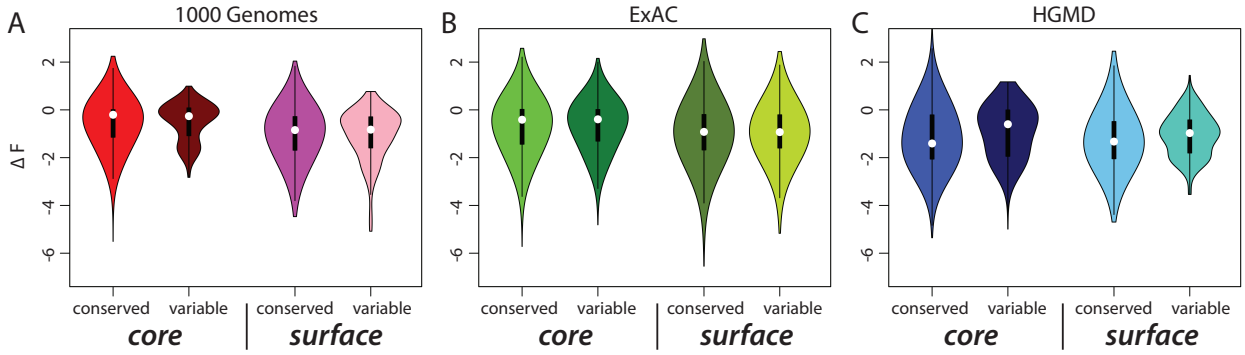
Mark

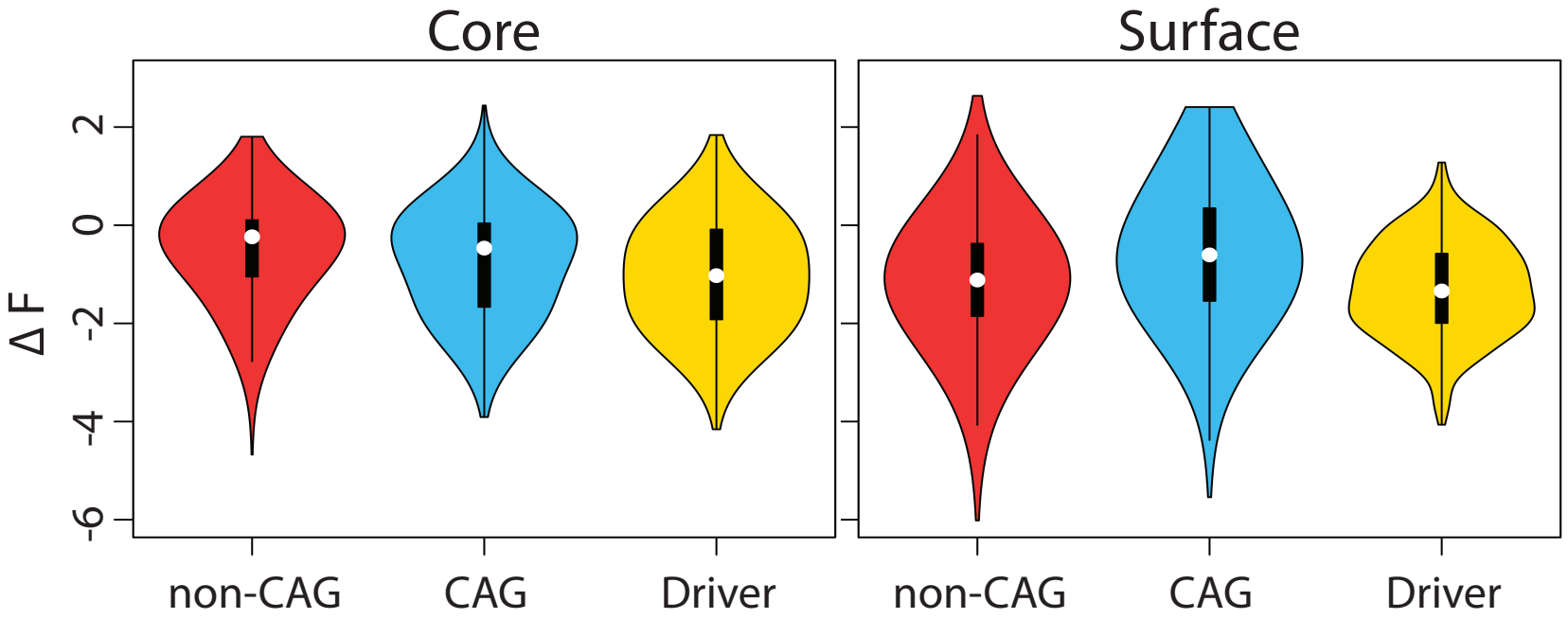
Frustration:
- Sushant

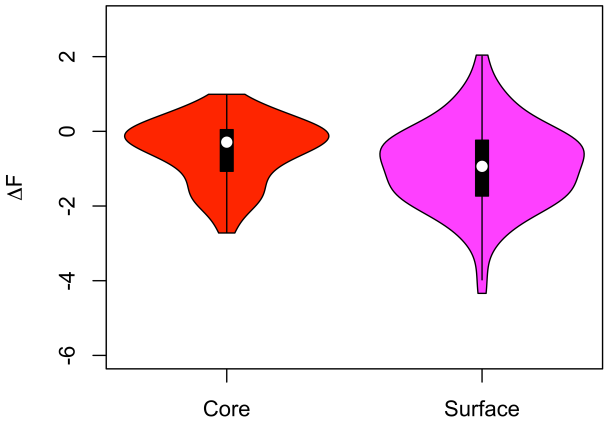
eQTLs:
- Joel
- Holly
- Shuang
- Fabio

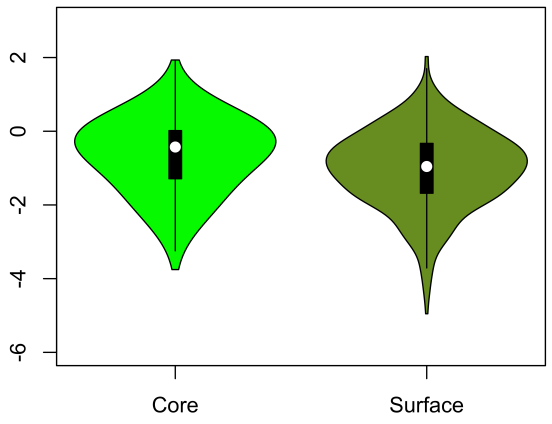
SUPPLEMENTARY SLIDES

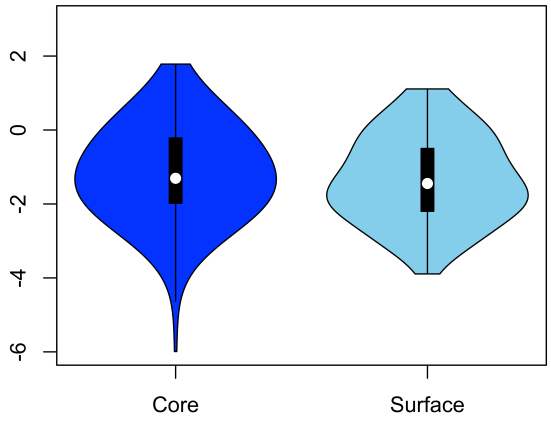

A
1000 Genomes
B
ExAC
C
HGMD
Δ F
2
0
-2
-4
-6
Core
Surface
Core
Surface
Core
Surface


A
1000 Genomes
B
ExAC
C
HGMD
ΔF
2
0
-2
-4
-6
conserved
variable
conserved
variable
core
surface

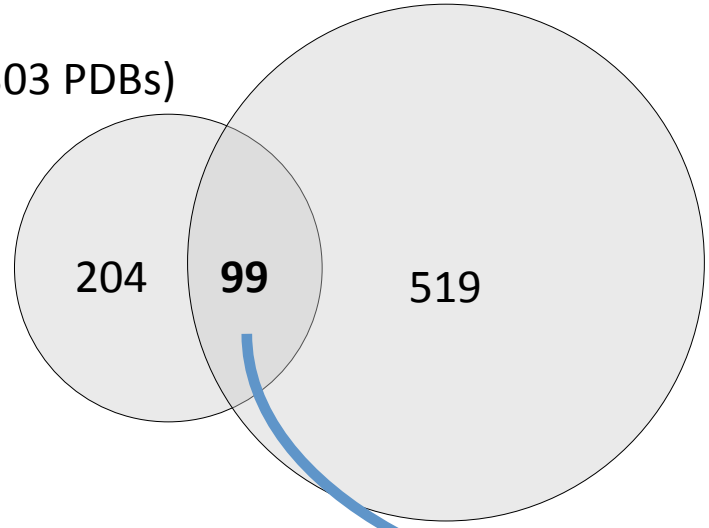

Core
Surface
ΔF
2
0
-2
-4
-6
non-CAG
CAG
Driver
non-CAG
CAG
Driver


ΔF
2
0
-2
-4
-6
Core
Surface

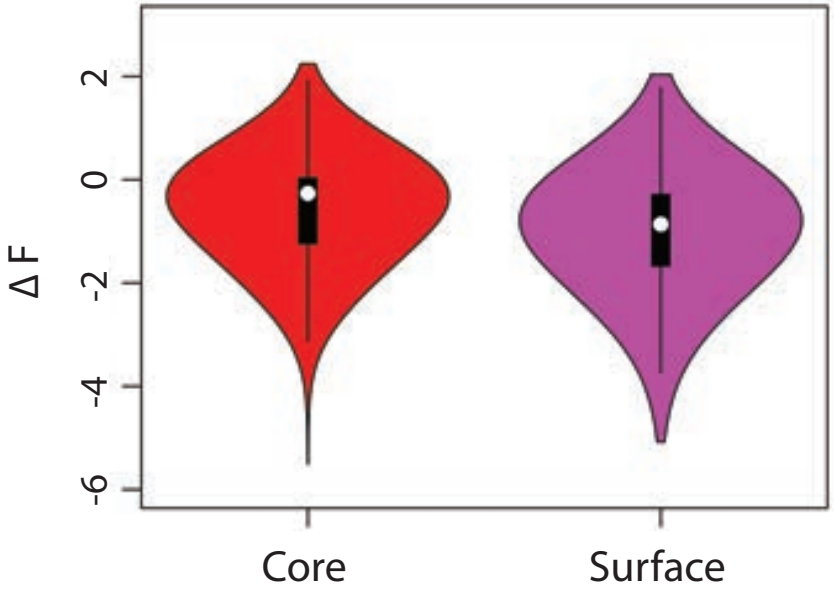

2
0
-2
-4
-6
Core
Surface


2
0
-2
-4
-6
Core
Surface

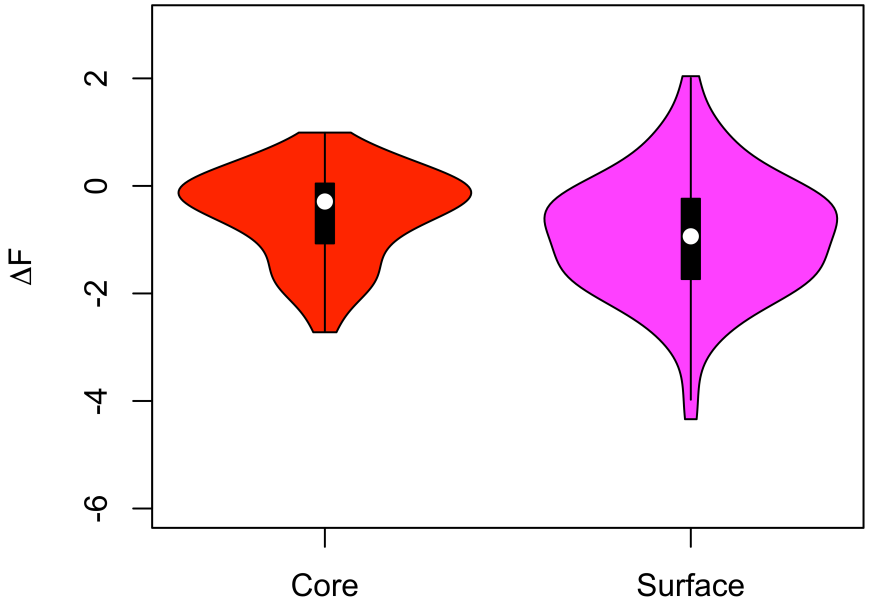
“They should also attempt to perform their analysis on a (semi-)balanced set(s) of variants, using sets of proteins where both disease and neutral mutations are present.”



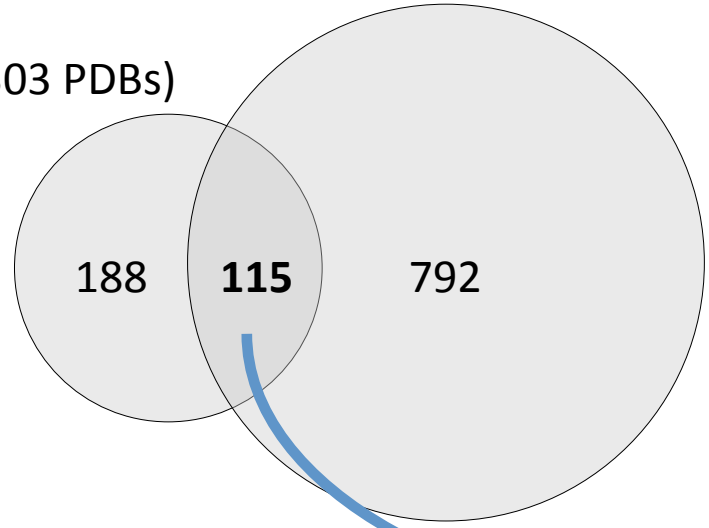
Previous



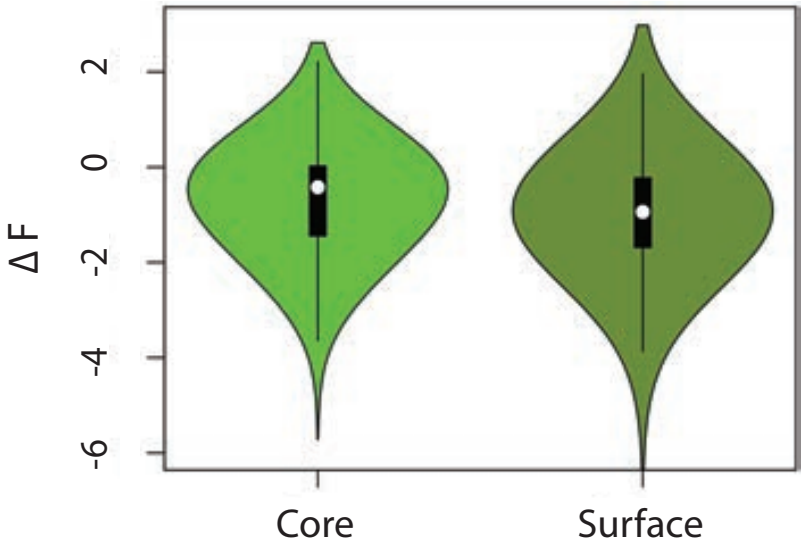
Revision



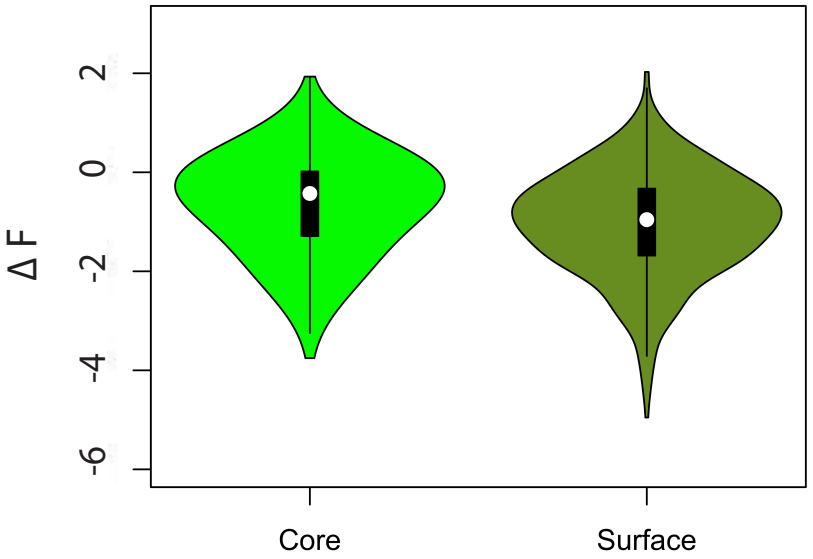
"They should also attempt to perform their analysis on a (semi-)balanced set(s) of variants, using sets of proteins where both disease and neutral mutations are present."

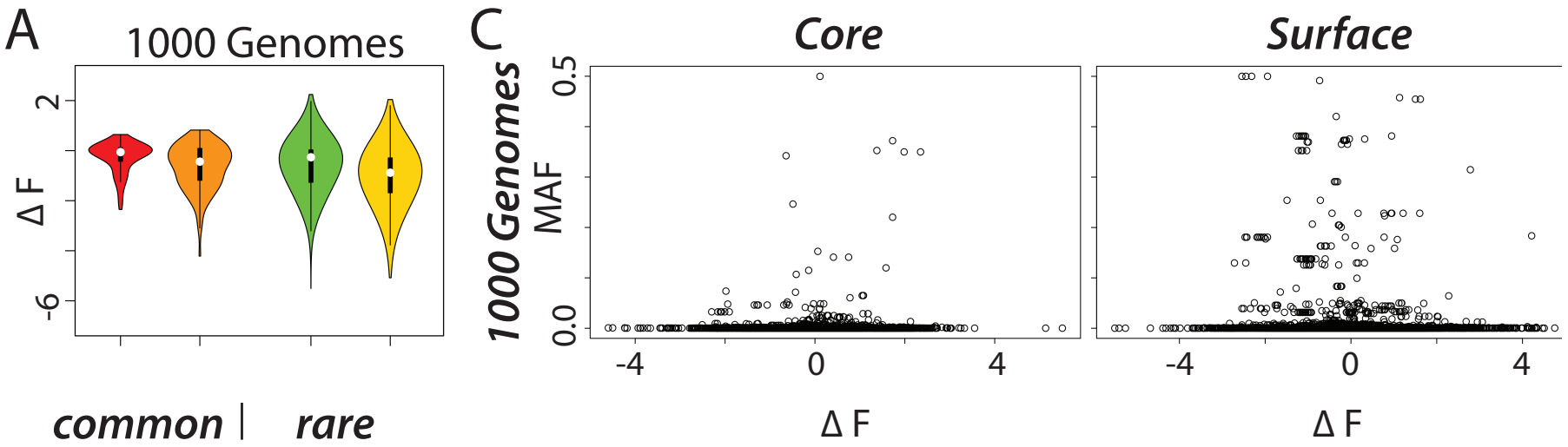


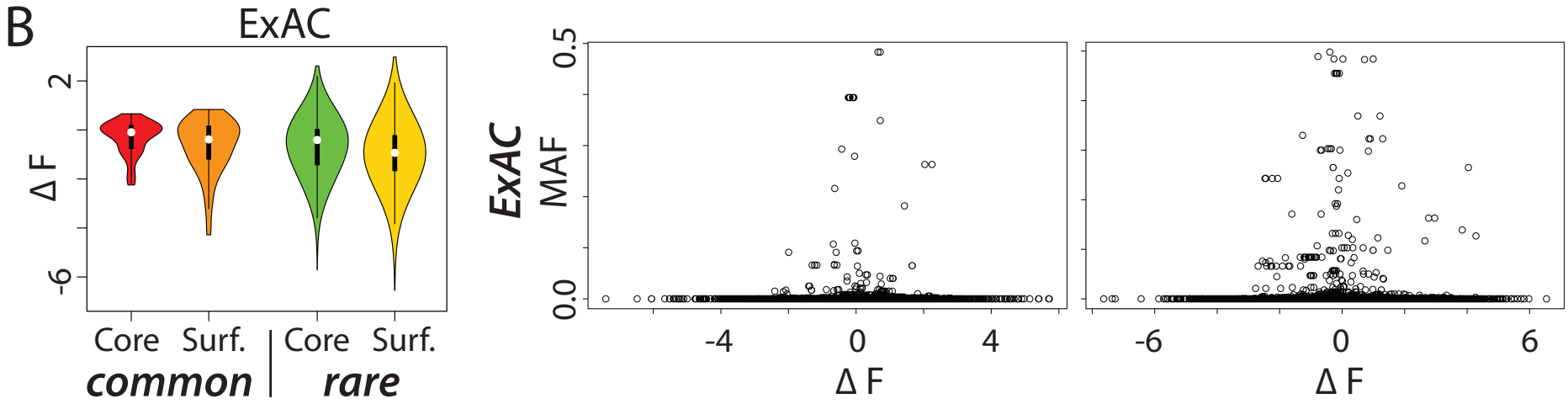
Previous



Revision




A
1000 Genomes
Δ F
2
-6
common
rare
C
Core
Surface
1000 Genomes
MAF
0.5
0.0
-4
0
4
Δ F
Δ F


B
ExAC
Δ F
2
-6
Core
Surf.
Core
Surf.
common
rare
ExAC
MAF
0.5
0.0
-4
0
4
Δ F
-6
0
6
Δ F

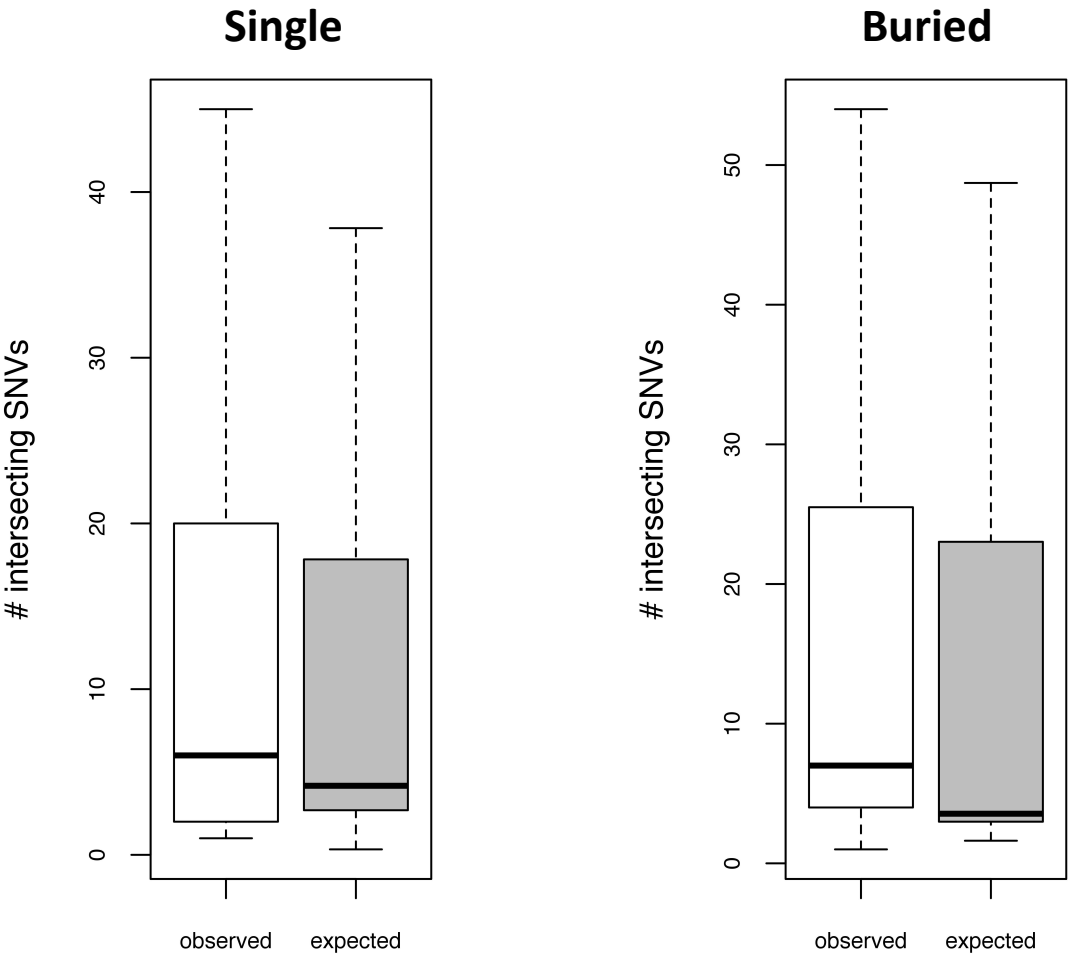
Additional scatterplots:

.//plot_delta_frustr_vs_MAF/frustr_vs_MAF_scatter_plots/

.//plot_delta_frustr_vs_MAF/

Tumor Suppressor Genes

Minimally Frustr. Residues

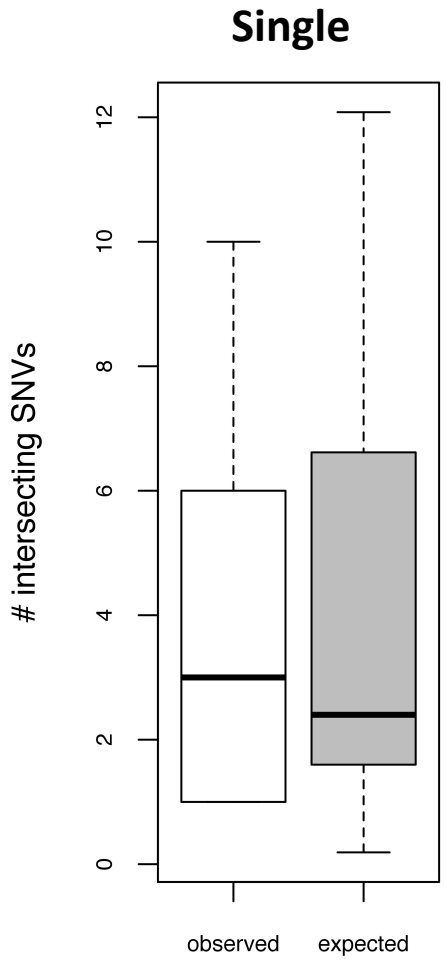


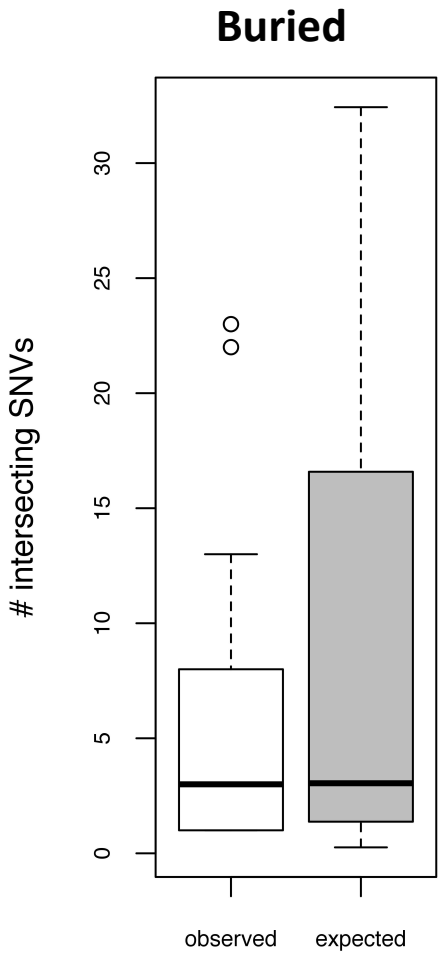
p-value = 0.001912
N = 39

p-value = 3.163e-06
N = 36

Tumor Suppressor Genes

Maximally Frustr. Residues





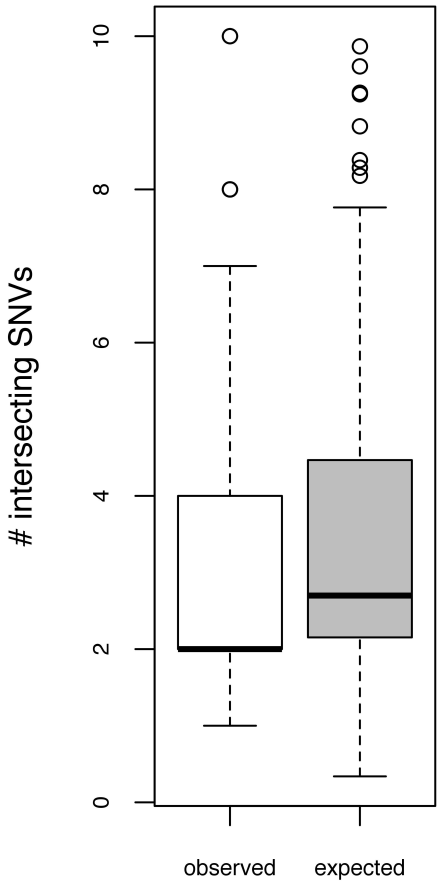
p-value = 0.005519
N = 28

p-value = 0.001818
N = 34

Oncogenes

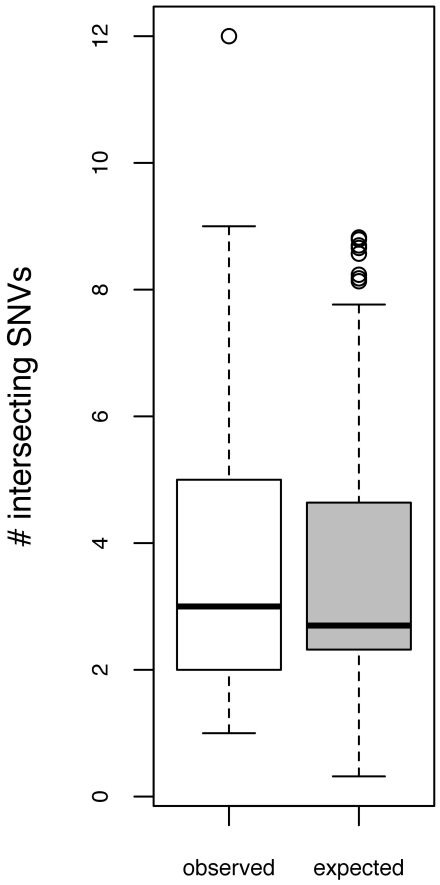
Minimally Frustr. Residues

Single



p-value = 0.01498
N = 118

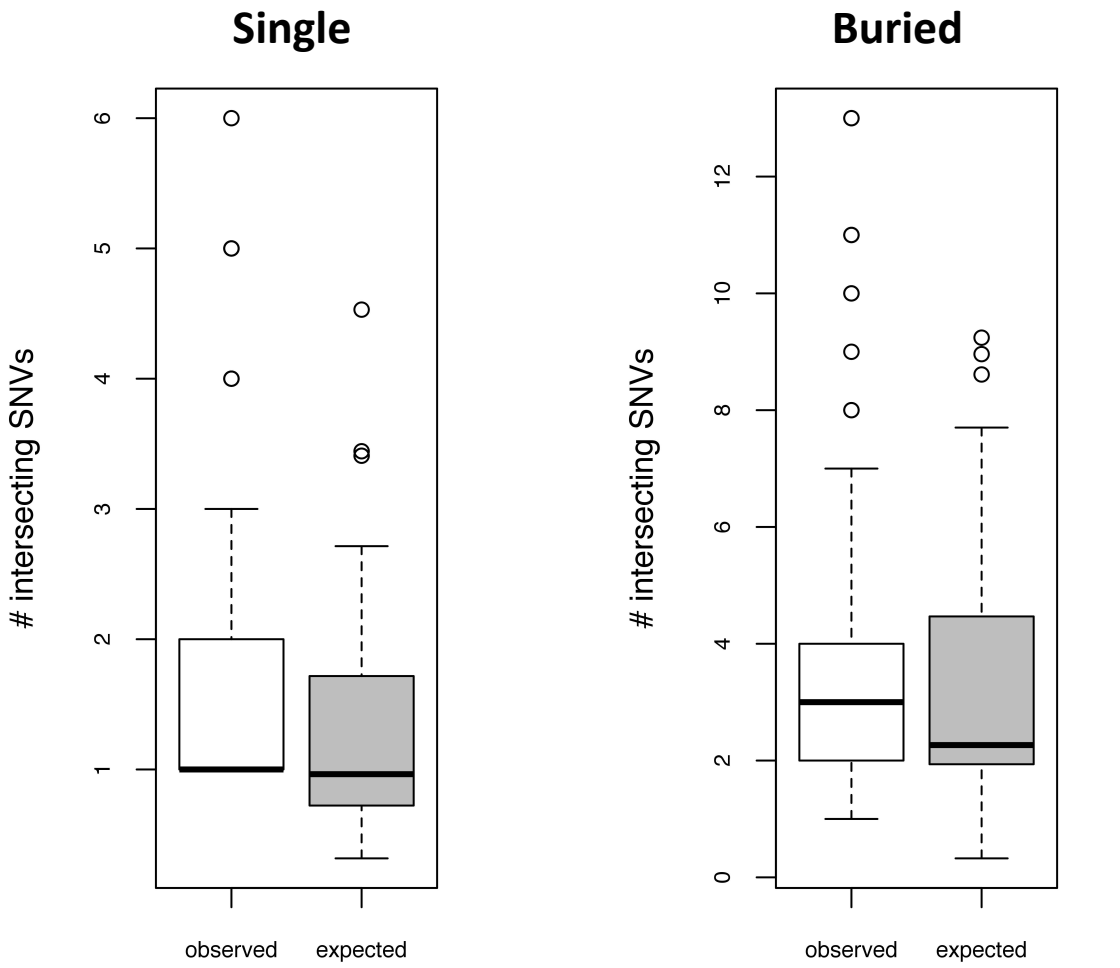
Buried



p-value = 0.7594
N = 96

Oncogenes

Maximally Frustr. Residues



p-value = 2.834e-07
N = 80

p-value = 4.159e-06
N = 112

**jpg and *pdf*

/Users/admin/Desktop/rsch/frustration/surf_and_core_enrichment/

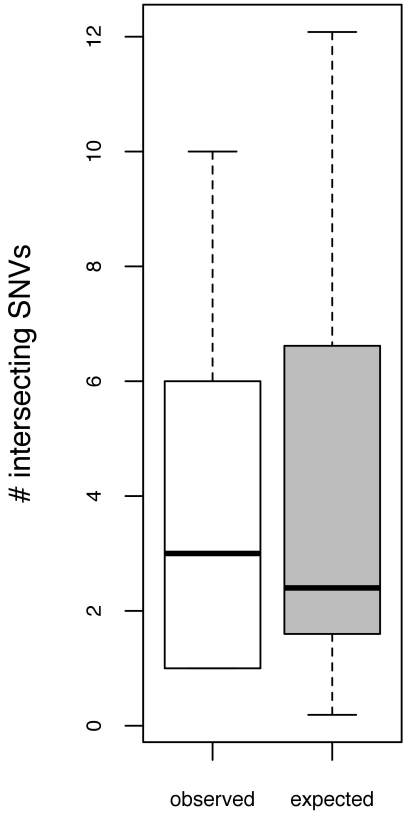
.//feb4_prs.pdf

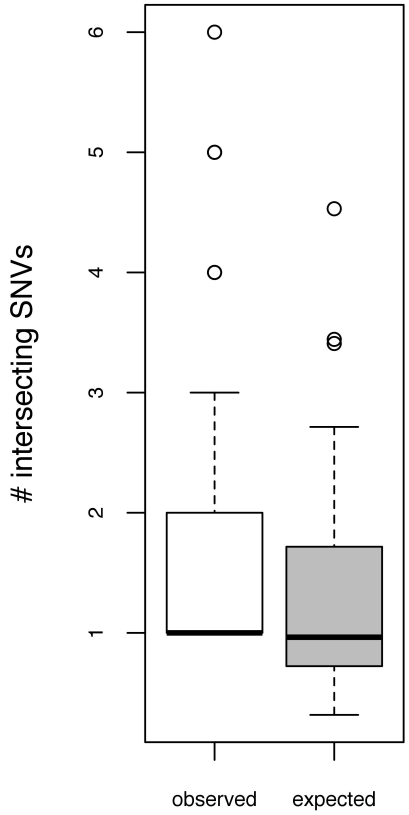
.//surf_and_core_enrichment/frustr_Mar16_mtg.pdf

Maximally Frustr. Residues Using the Single-Residue Index

TSGs

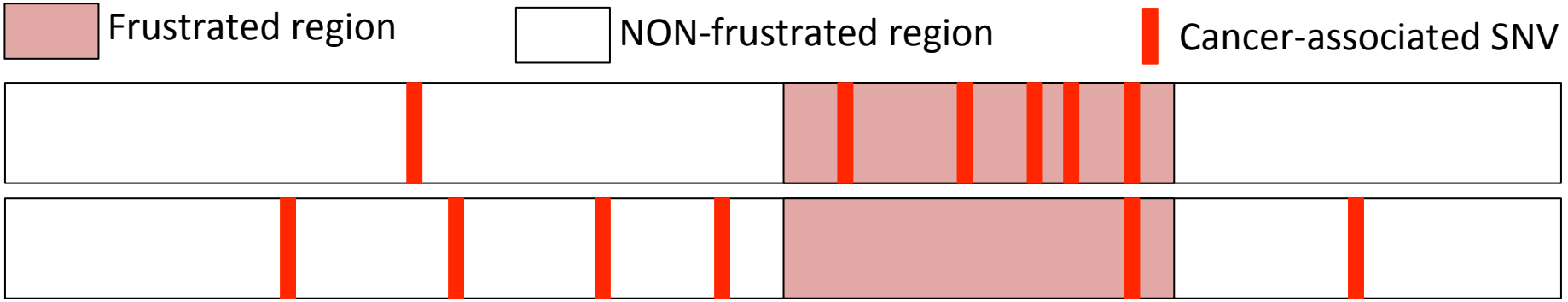
p-value = 5.5E-3
N = 28





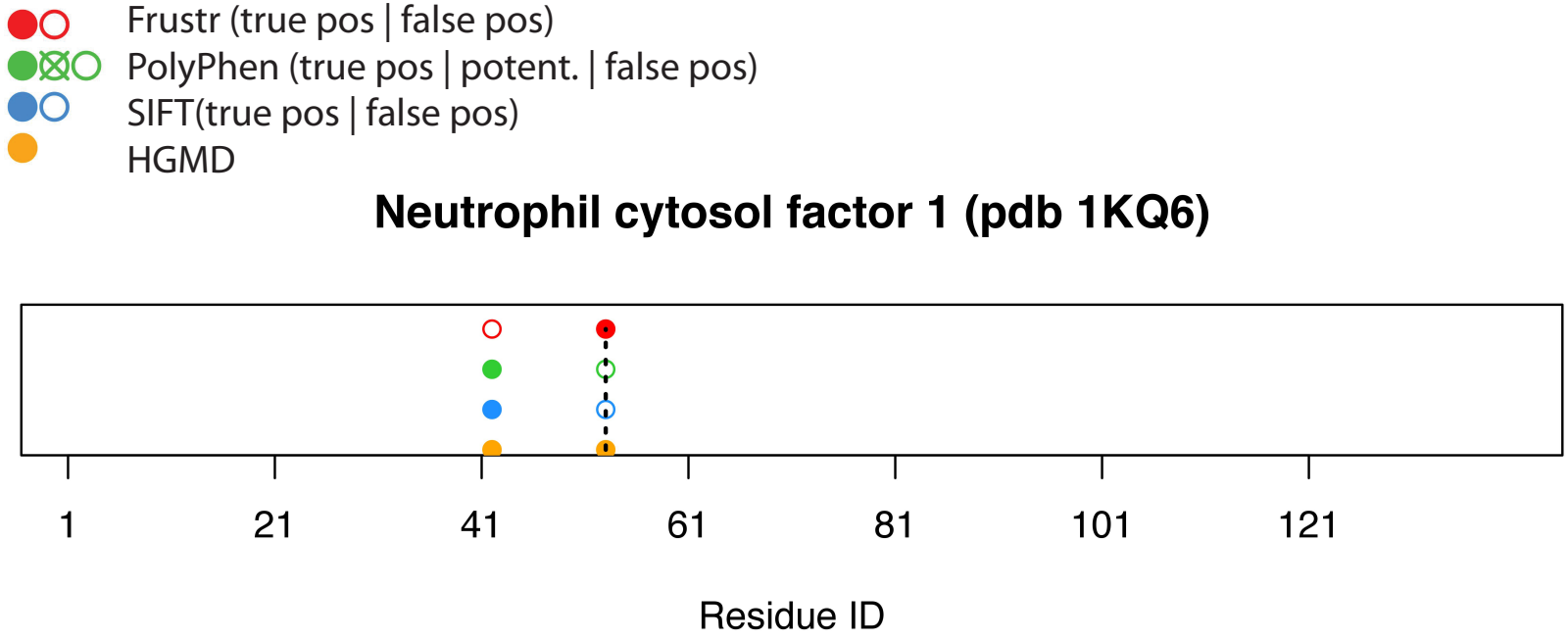
Oncogenes

p-value = 2.83E-7
N = 80



Observed: X = # of cancer-associated SNVs that intersect frustrated regions (5 in this case)
Expected: E[X] = [# frustrated residues / total # residues in protein] * [total # of cancer-associated SNVs]

To underline the usefulness of your method, which is, ... to meet a "growing and urgent need to evaluate the potential effects of low-allele-frequency variants in unbiased ways using high-throughput methodologies", I miss some extra calculations / benchmarking. There are methods existing in order to evaluate potential effects of low-allele-frequency variants in unbiased ways (SIFT, PolyPhen2, MutationTaster, and many others). I would like to see how exactly your method adds up to this. Is the additional information gained from structural analysis really an advantage over existing methods? If you could show this, this would surely be an argument for people to use and cite your method ... One could for ... create a small set of variants and analyse these with one or two of the "common" tools to predict the deleteriousness of SNVs (e.g. PolyPhen2 and MutationTaster2, since these are generally considered the most accurate ones) and then check if there are disease variants predicted as "harmless" by these tools (i.e. false negative) which are then correctly seen as locally maximal frustrated by your method. Or any other way how it can be shown that the method is indeed useful for the analysis of high-throughput data (e.g. compare with other existing "structural prediction" tools, if those exist).



+ (reminder on nets disc)

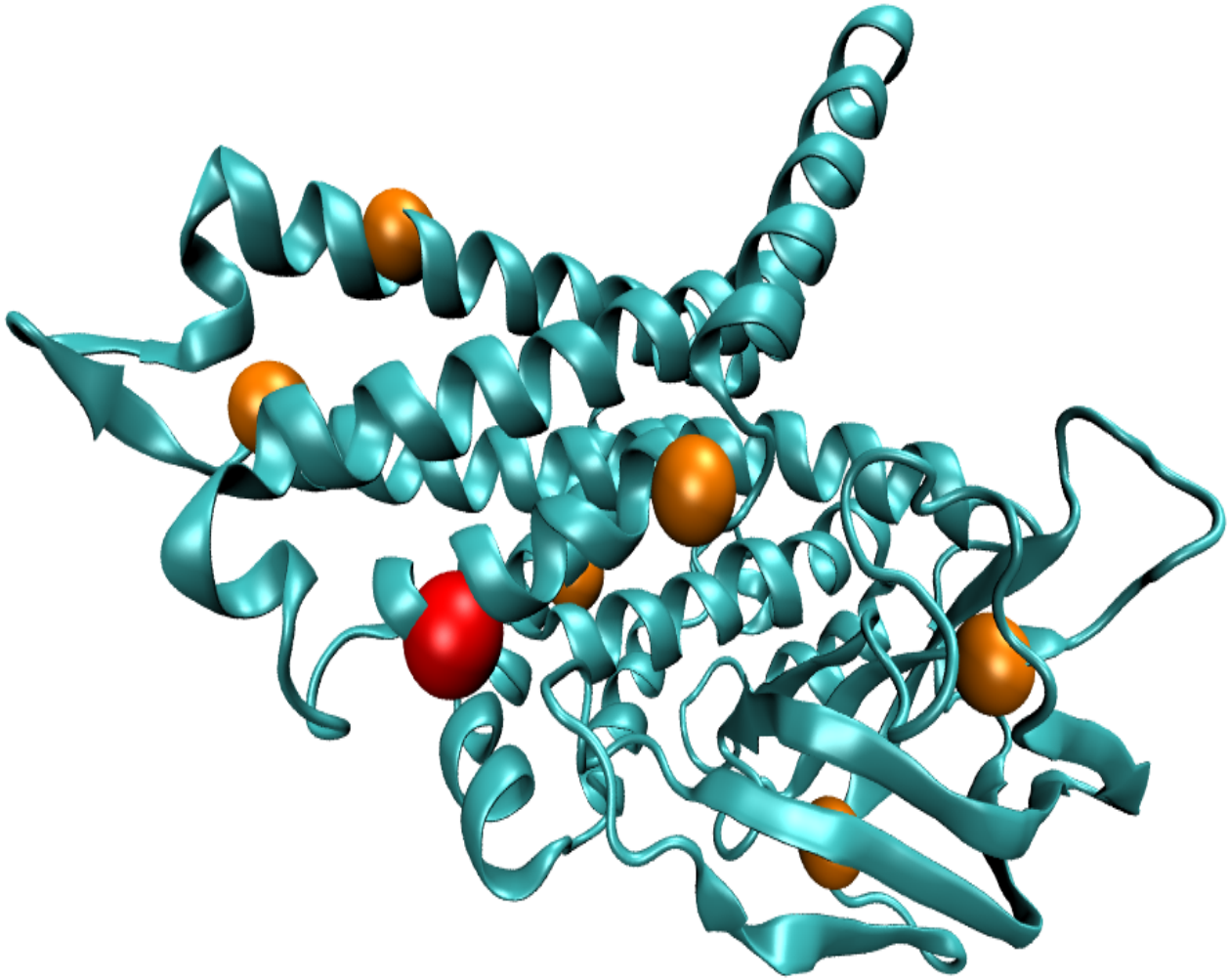
Acyl-CoA-dehydrogenase deficiency

PDB	SNPs	ResPos	origRes	mutRes	
2VIG	chr12:121176108:T:G	217	MET	ARG	-1.802
2VIG	chr12:121174892:T:A	105	ILE	ASN	-2.855
2VIG	chr12:121176421:C:A	294	ALA	ASP	-1.728
2VIG	chr12:121177182:C:G	390	ILE	MET	-2.909
2VIG	**chr12:121177150:C:T**	**380**	**ARG**	**TRP**	**-5.352**
2VIG	chr12:121175763:C:T	199	ALA	VAL	0.693
2VIG	chr12:121176633:C:T	315	ALA	VAL	0.297

HGMD SNP disrupting core residues to different extent in a particular disease

8

Acyl-CoA-dehydrogenase deficiencv



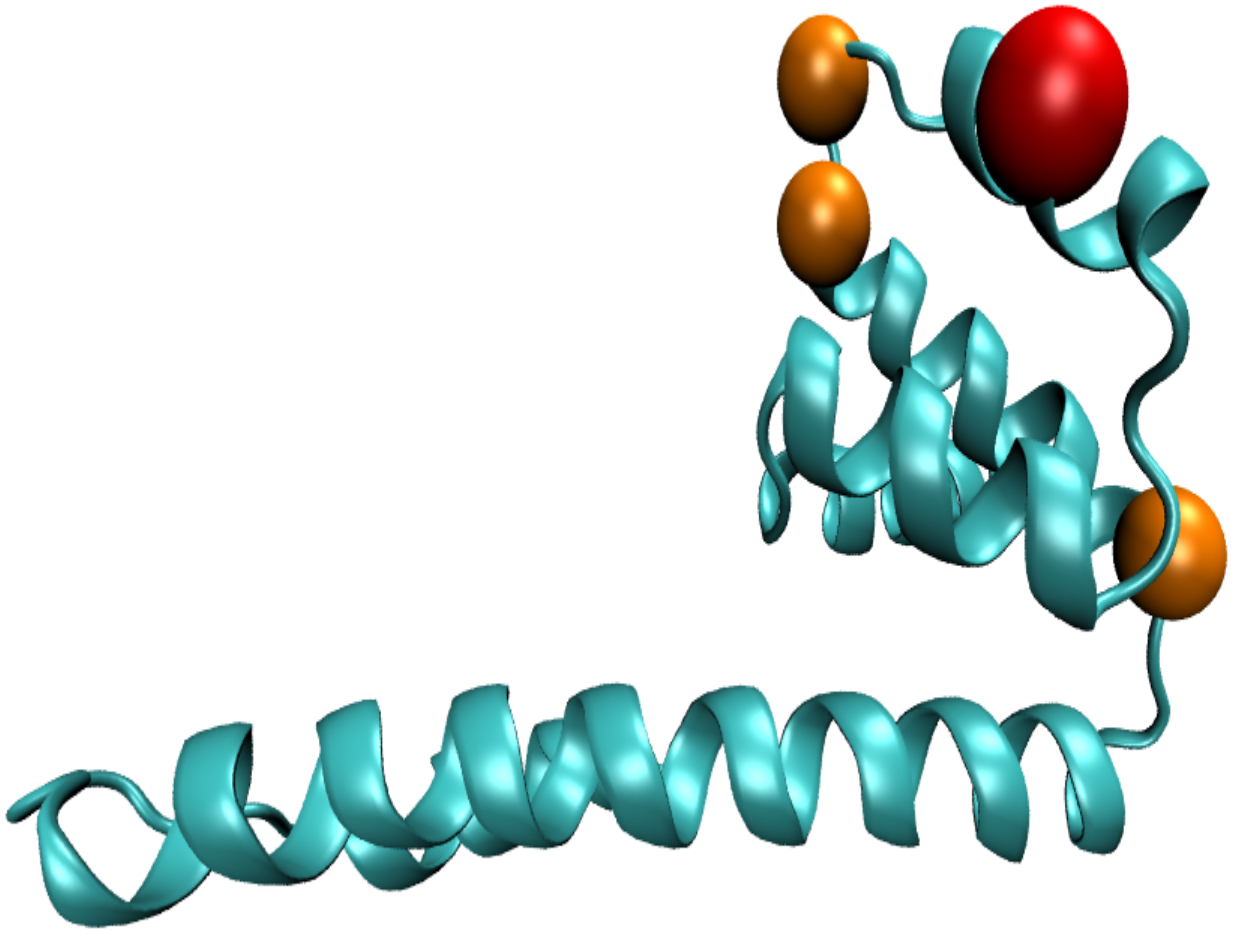
Autoimmune Lymphoproliferative Syndrome

PDB	SNPs	ResPos	origRes	mutRes	
3EZQ	chr10:90773977:G:A	260	ASP	ASN	-0.615
3EZQ	chr10:90773978:A:G	260	ASP	GLY	-2.129
3EZQ	chr10:90773977:G:C	260	ASP	HIS	-1.355
3EZQ	chr10:90774008:C:T	270	THR	ILE	0.396
3EZQ	chr10:90774008:C:A	270	THR	LYS	-0.044
3EZQ	chr10:90774002:A:C	268	GLN	PRO	-0.334
3EZQ	chr10:90774050:T:C	284	LEU	PRO	0.015

HGMD SNPs disrupting surface residues to different extent.

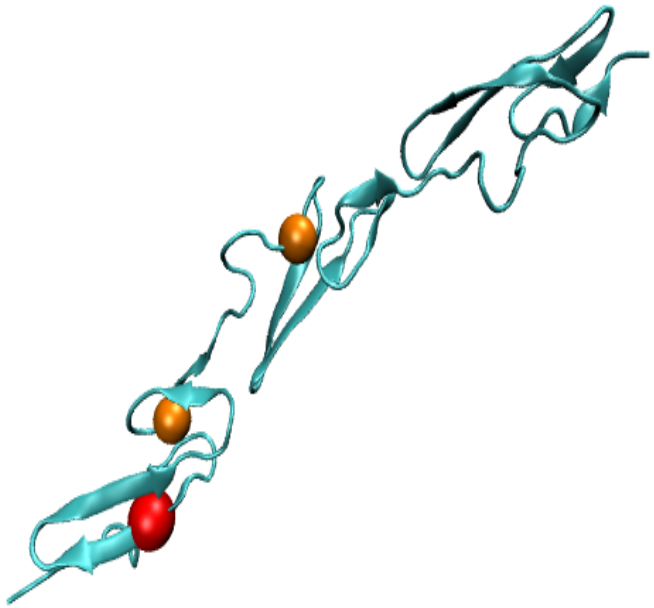
9

Autoimmune Lymphoproliferative Syndrome



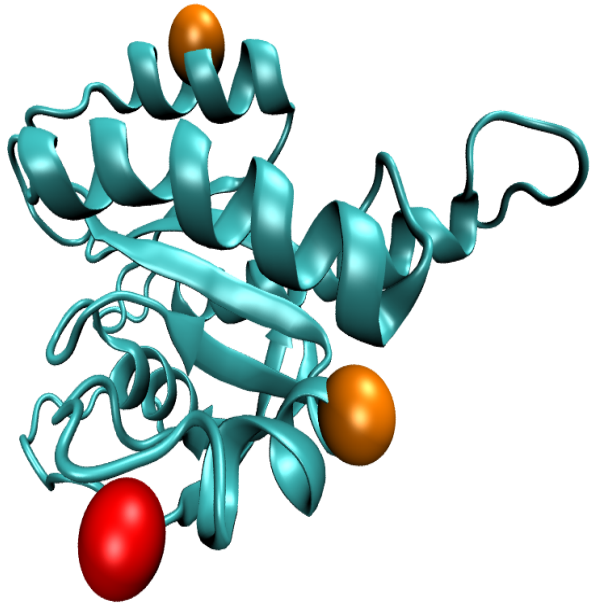
TSG Driver disrupting core residues : NOTCH1 gene

PDB	SNPs	ResPos	origRes	mutRes		cancerType
2VJ3	chr9:139412263:C:T	461	CYS	TYR	-2.813	Head & Neck
2VJ3	**chr9:139412359:C:T**	**429**	**CYS**	**TYR**	**-3.477**	**Head & Neck**
2VJ3	chr9:139412360:A:T	429	CYS	SER	-1.572	Lung
2VJ3	chr9:139412299:C:T	449	CYS	SER	-1.085	Head & Neck



Oncogene Driver disrupting surface residues : KRAS gene

PDB	SNPs	ResPos	origRes	mutRes		cancerType
4DSO	chr12:25398213:T:A	36	ILE	LEU	-0.928	Esophageal
4DSO	chr12:25380240:C:A	73	ARG	MET	-2.495	Astrocytoma
4DSO	**chr12:25398211:T:C**	**36**	**ILE**	**MET**	**-3.631**	**AML**
4DSO	chr12:25378603:T:C	132	ASP	GLY	-1.408	Stomach



Potential causes of outlier error rates

- single-exon genes (often associated w/pgenes) → duplications → low mappability scores
- low mappability scores? → check using intersect bed w/encode in UCSC Genome Browser. Genome browser has a track for mappability → first download and check w/intersect bed?
- exon lengths from same genome build? In any case GTEx is reporting in GENE read counts
- With BAM file as input, GTEx uses RNA-SeQC: "Expression levels were produced at the gene and exon level in RPKM units using RNA-SeQC"
  → Black box & confounding factors (GC bias, mapability, uniqueness, etc)

Misc Notes

- Strange that *processed* read counts data are not available at the GTEx Portal
- BAM files not available to re-compute RPKM from RNA-SeqC
- GTEx: tophat/bowtie, though will be STAR 2.4.2a in v7 (CommonMind= STAR)
- PsychENCODE currently processing all to be uniform?

Framingham data (miRNA-eQTLs)