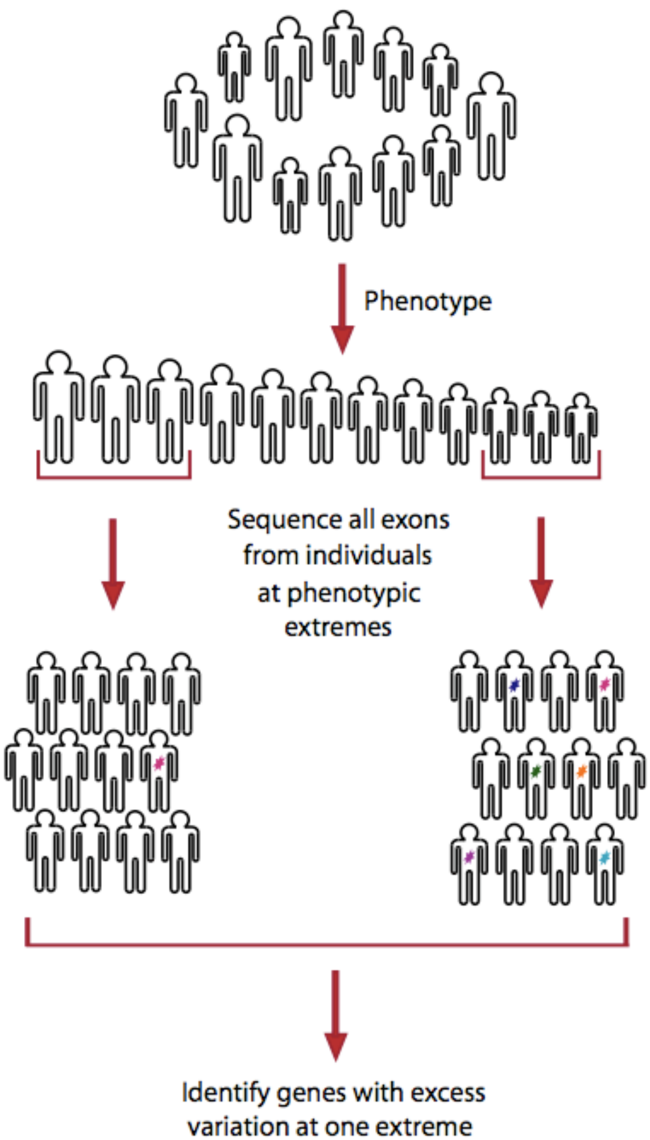
Computational methods for the analysis of rare variants

Shamil Sunyaev

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Harvard-M.I.T. Health Sciences & Technology Division

Combine all non-synonymous variants in a single test



Theory:

1) Most new missense mutations are functional (*mutagenesis, population genetics, comparative genomics*)

2) Most new missense mutations are only weakly deleterious (*population genetics*)

3) Most functional missense mutations are likely to influence phenotype in the same direction (*mutagenesis, medical genetics*)

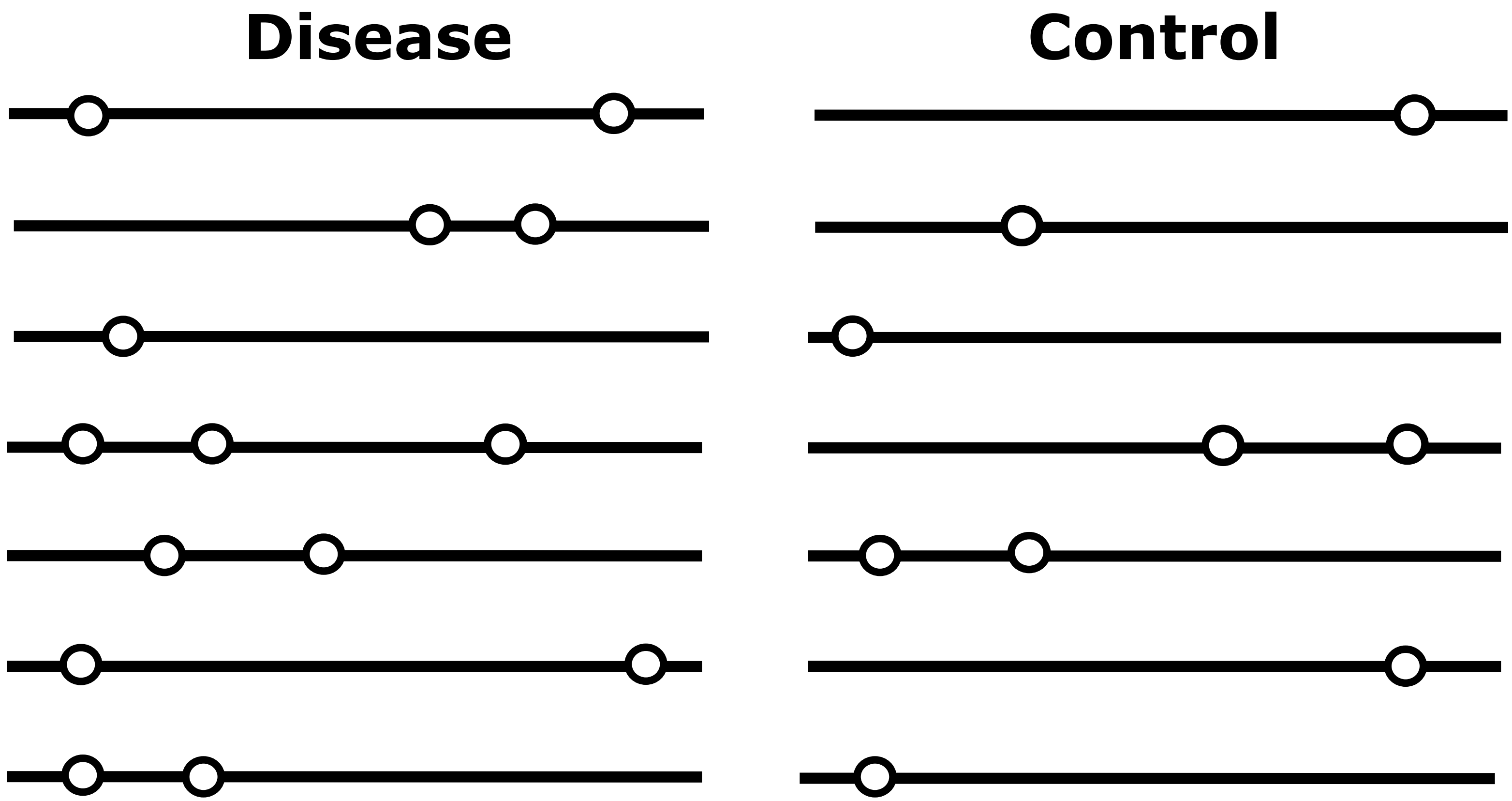
Data:

multiple candidate gene studies

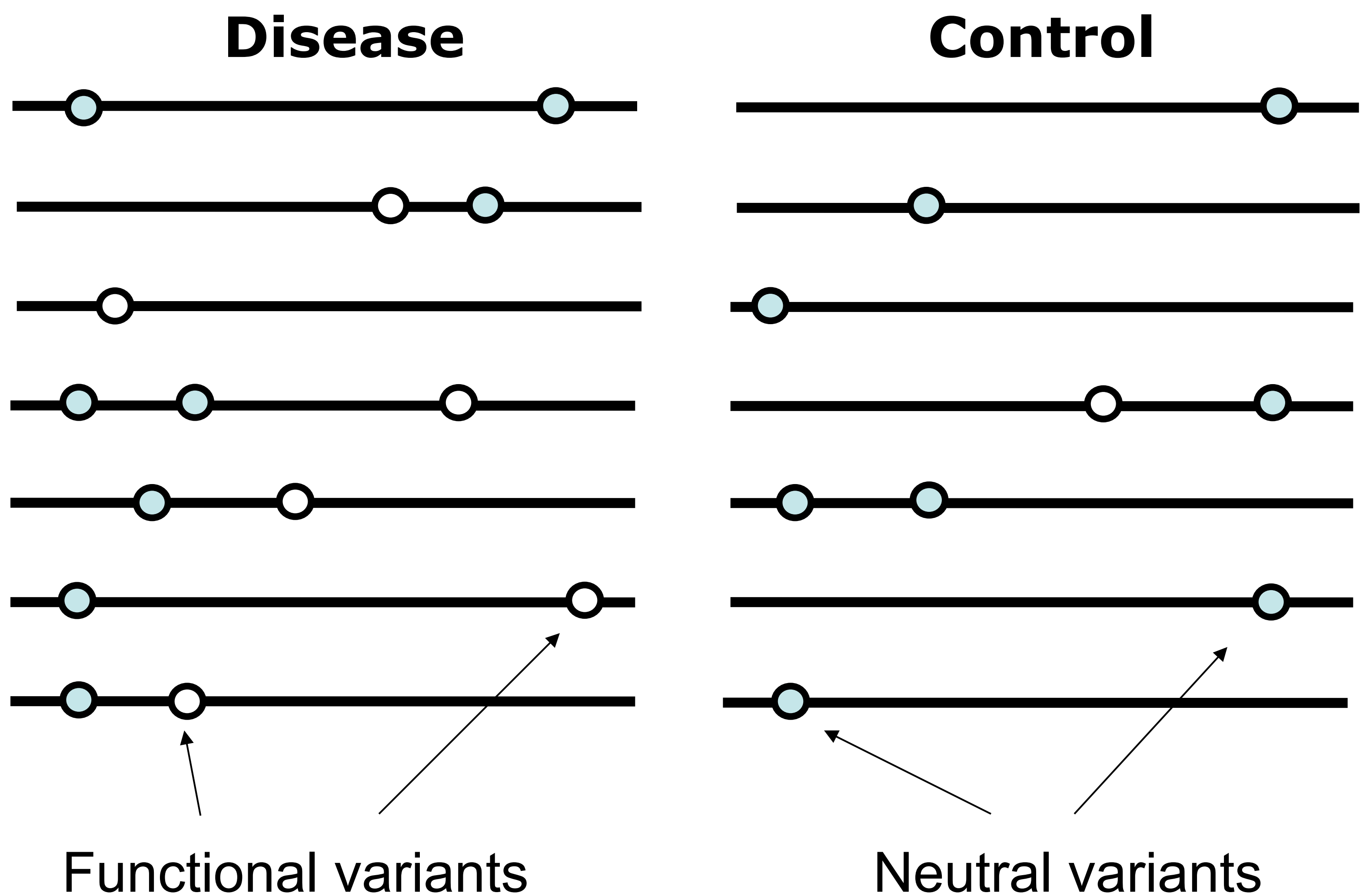
HDL-C, LDL-C, Triglycerides, BMI, Blood pressure, Colorectal adenomas

Kryukov et al., PNAS 2009

Combining variants in a single test

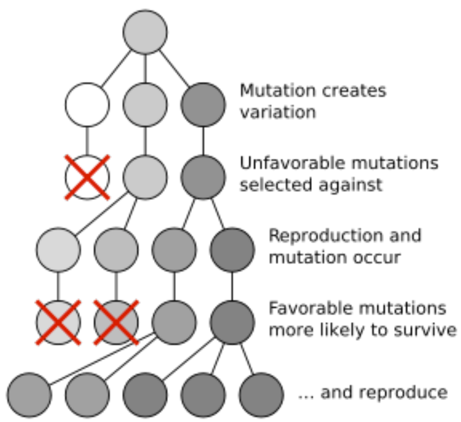


Combining variants in a single test

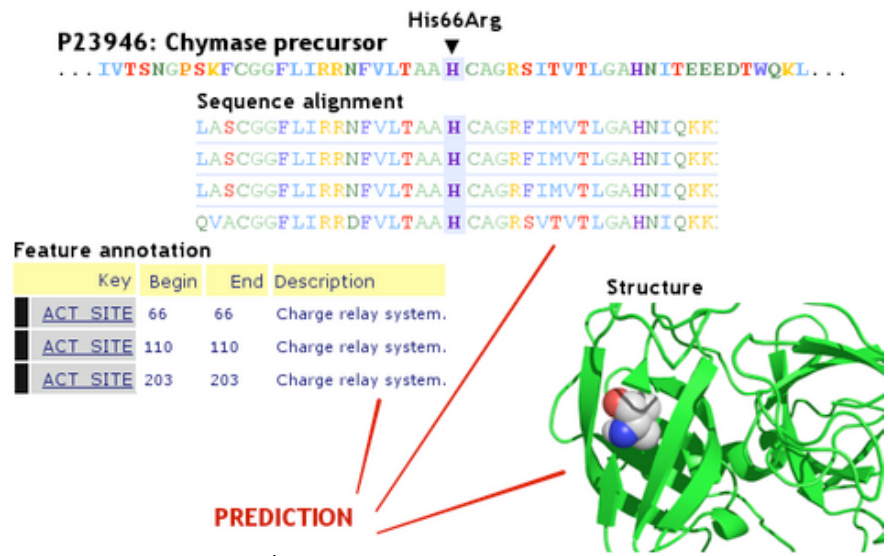


How do we know that the variant is functional?

Population genetics

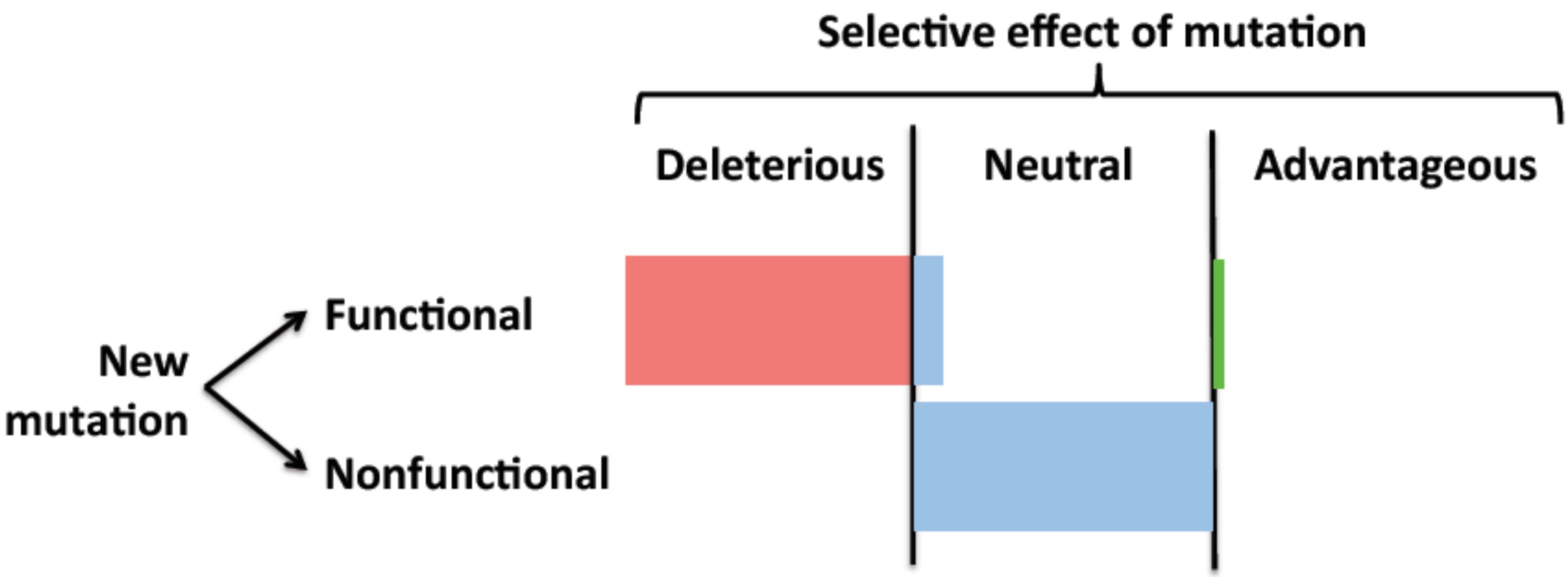


Bioinformatics



Probability that the variant is functional

Most functional mutations are under selective pressure even if the trait is not



Selection indicates functional mutations, whether or not the tested trait is under selection

Allele frequency is informative about selective pressure

- What is the optimal way to incorporate allele frequency information into a burden test?

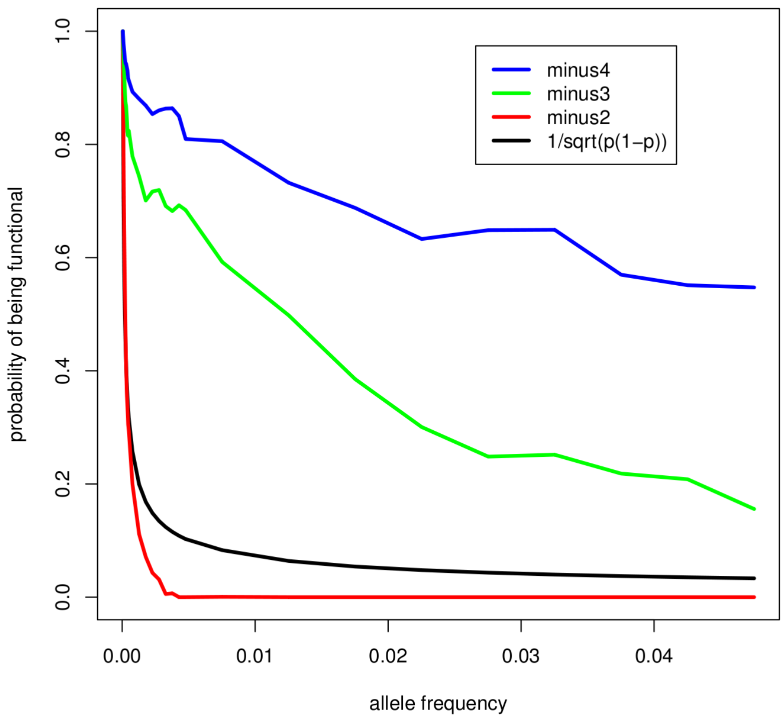
- Is there a natural threshold of allele frequency?

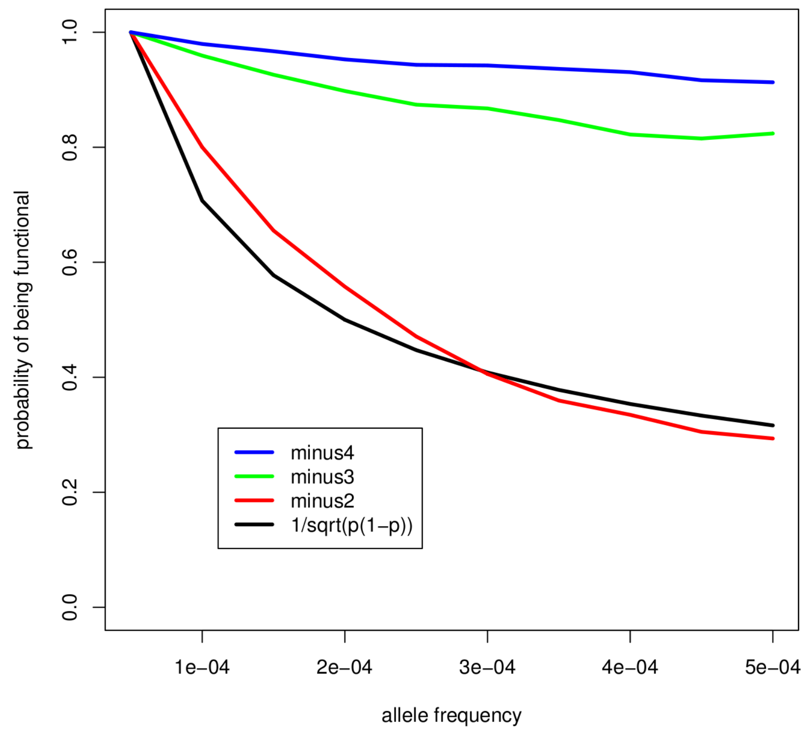
- Is there an optimal way to weight allelic variants with respect to their allele frequency?

Allele frequency has to be taken into account when combining variants

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

Probability that a variant is functionally significant given its allele frequency





However, this dependence is not robust with respect to s_0!

"Goldilocks" alleles

- Special case in terms of study design: alleles of large effect that are frequent enough to be followed up individually in a larger population sample.

- Such "goldilocks" alleles are observed in the simulations.

There is no optimal and robust weighting scheme or optimal threshold!

Variable threshold (VT) approach

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

Variable threshold (VT) approach

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

Variable threshold (VT) approach

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

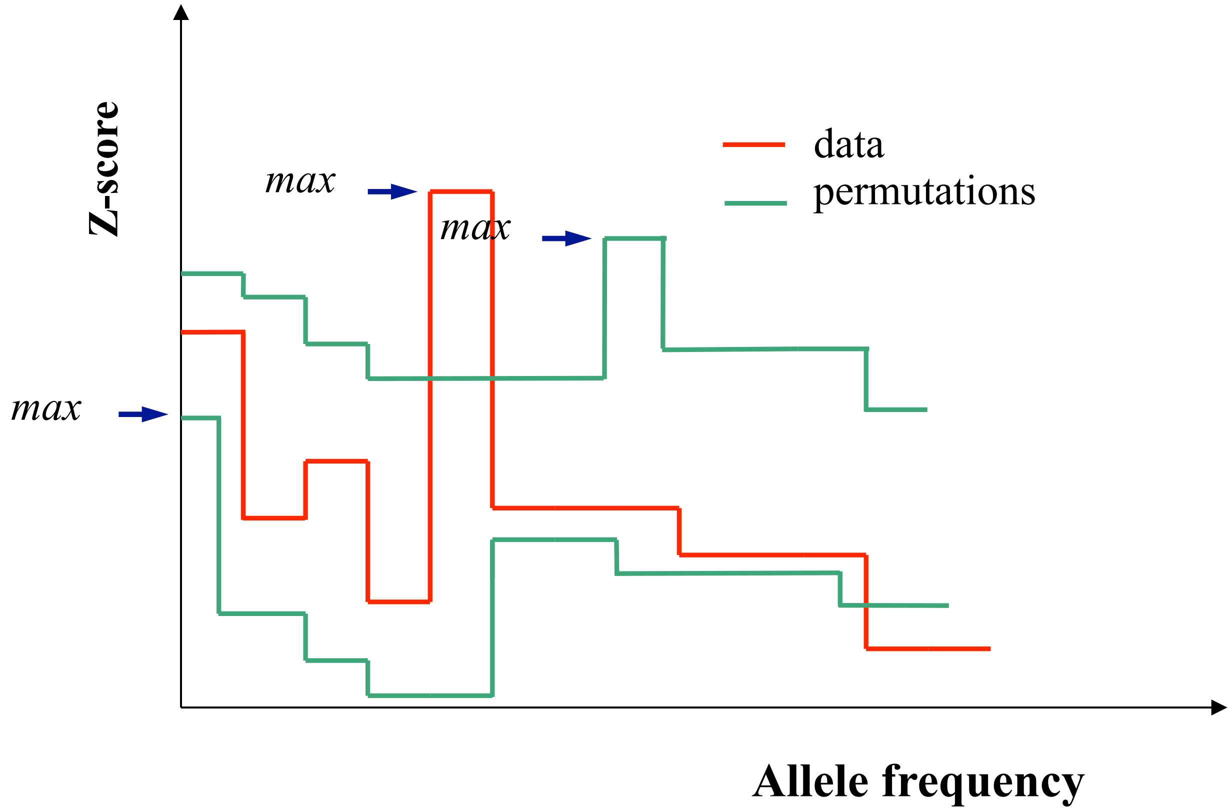
Variable threshold (VT) approach

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

Variable threshold (VT) approach

	Cases	Controls
SNP1	1	0
SNP4	1	0
SNP7	1	0
SNP8	0	1
SNP2	2	2
SNP6	4	2
SNP3	10	1
SNP5	195	210

Variable Threshold (VT) approach



$z(T)$ is the z-score of a regression across samples of phenotypes vs. counts of alleles with frequency below threshold T.

We maximize $z(T)$ over T. Type I error is controlled by permutations.

Price, Kryukov et al., AJHG 2010

Beyond allele frequency

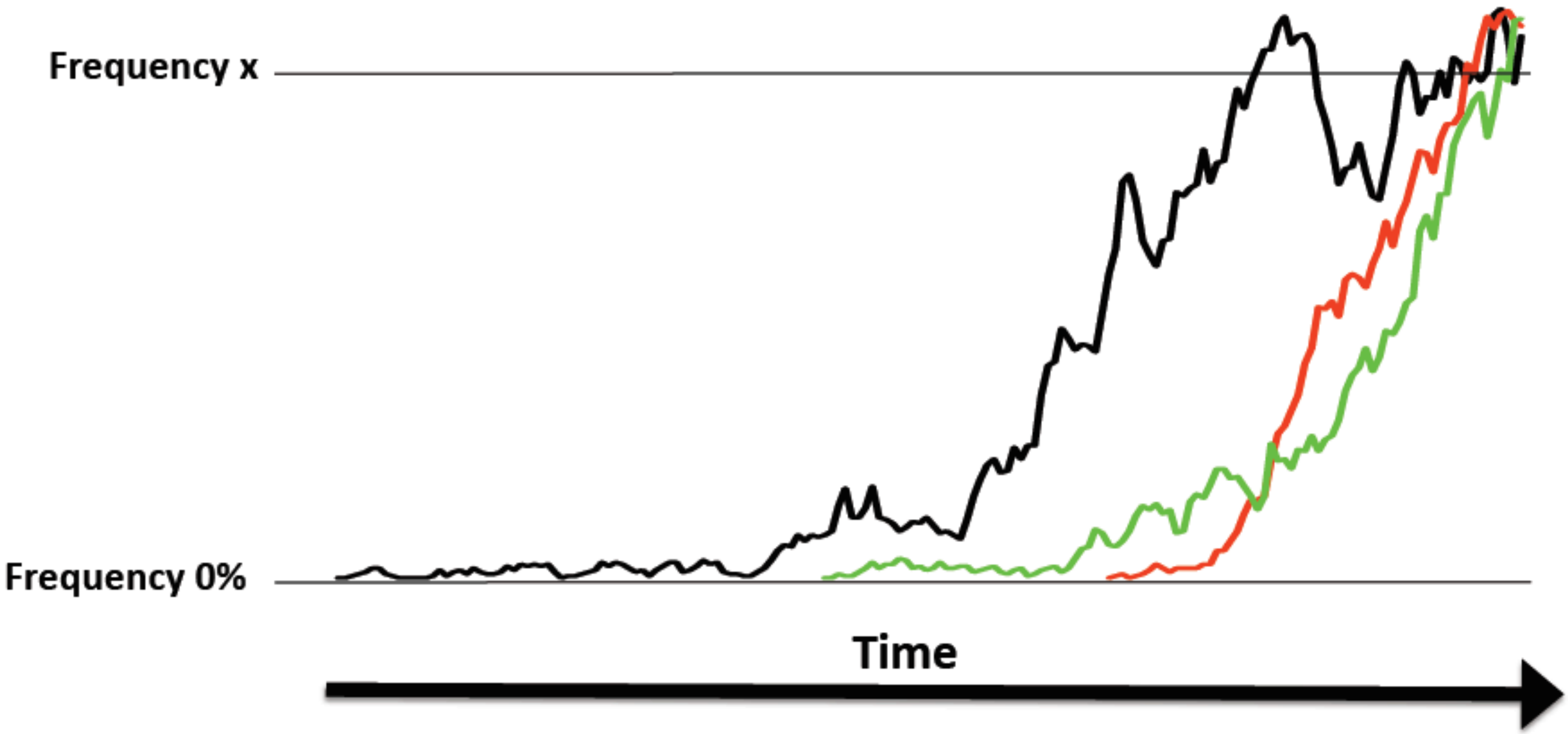
- Allele frequency does not capture all information about selective pressure.

- It is possible to further stratify allelic variants of the same frequency.

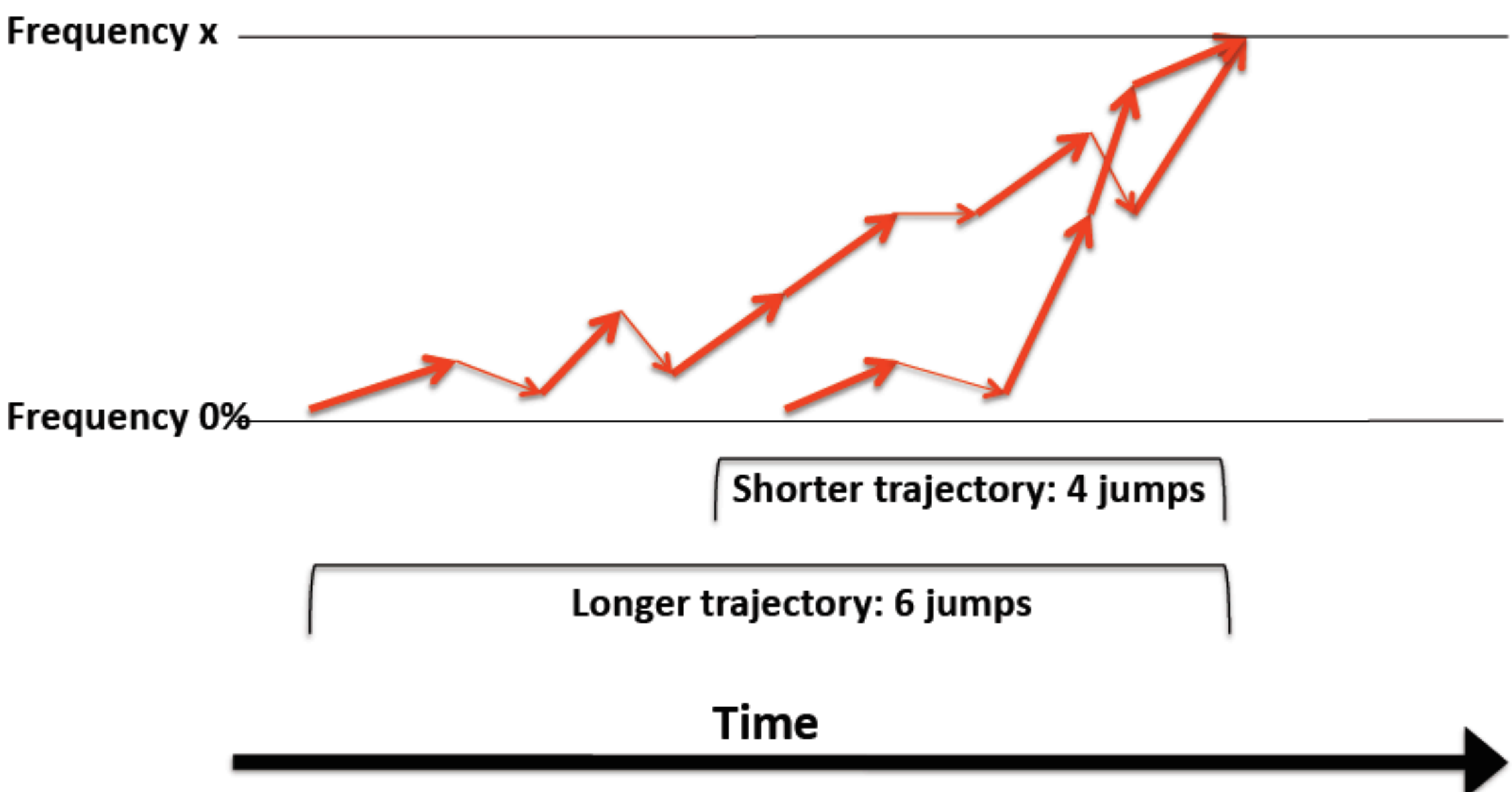
- Alleles under selective pressure are, on average, younger than neutral alleles. This is true even if they are at the same frequency.

Allelic age is informative even conditionally on frequency

Maruyama effect (1974): at any frequency **advantageous** , or **deleterious** alleles are younger than **neutral** alleles

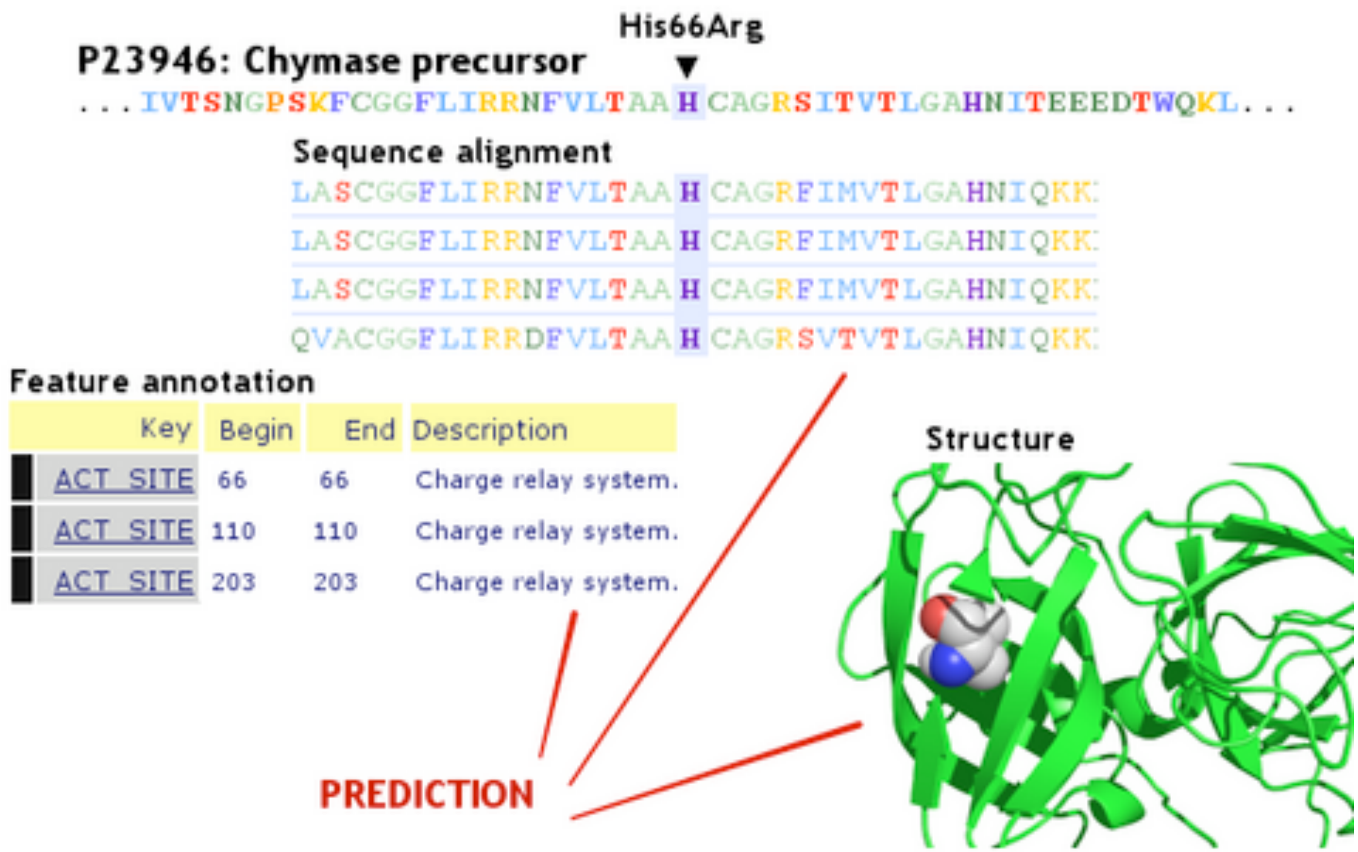


Intuition behind the effect

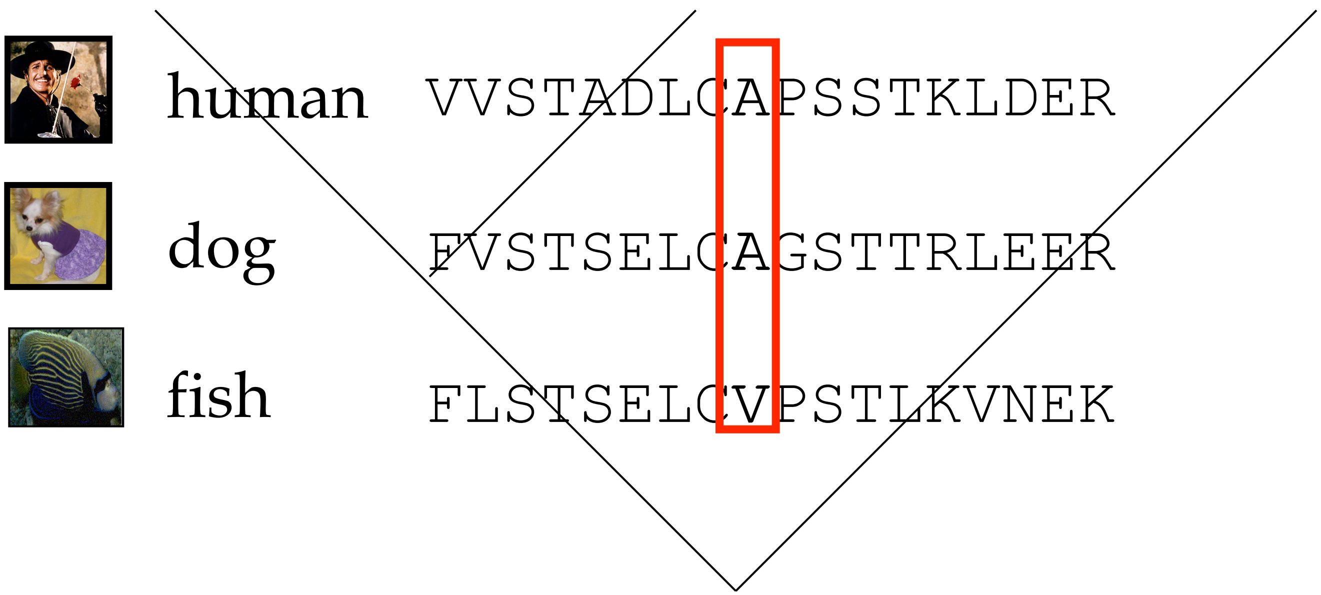


Allelic age can be measured by density of younger mutations and by LD decay

Bioinformatics predictions



Does the mutation fit the pattern of past evolution?

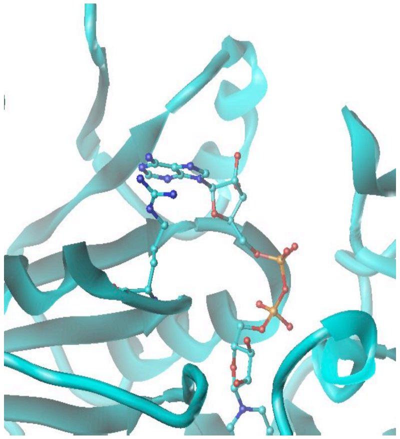
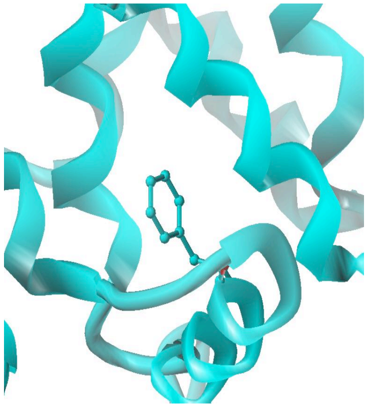
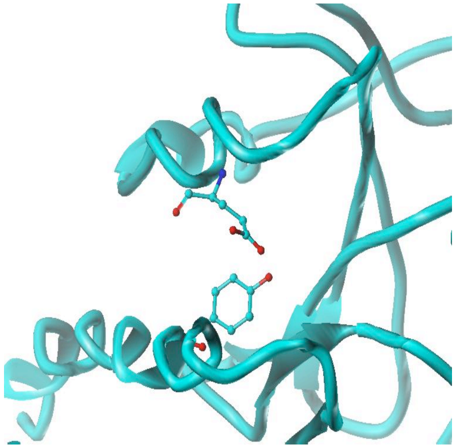


Statistical issues:

-sequences are related by phylogeny

-generally, we have too few sequences

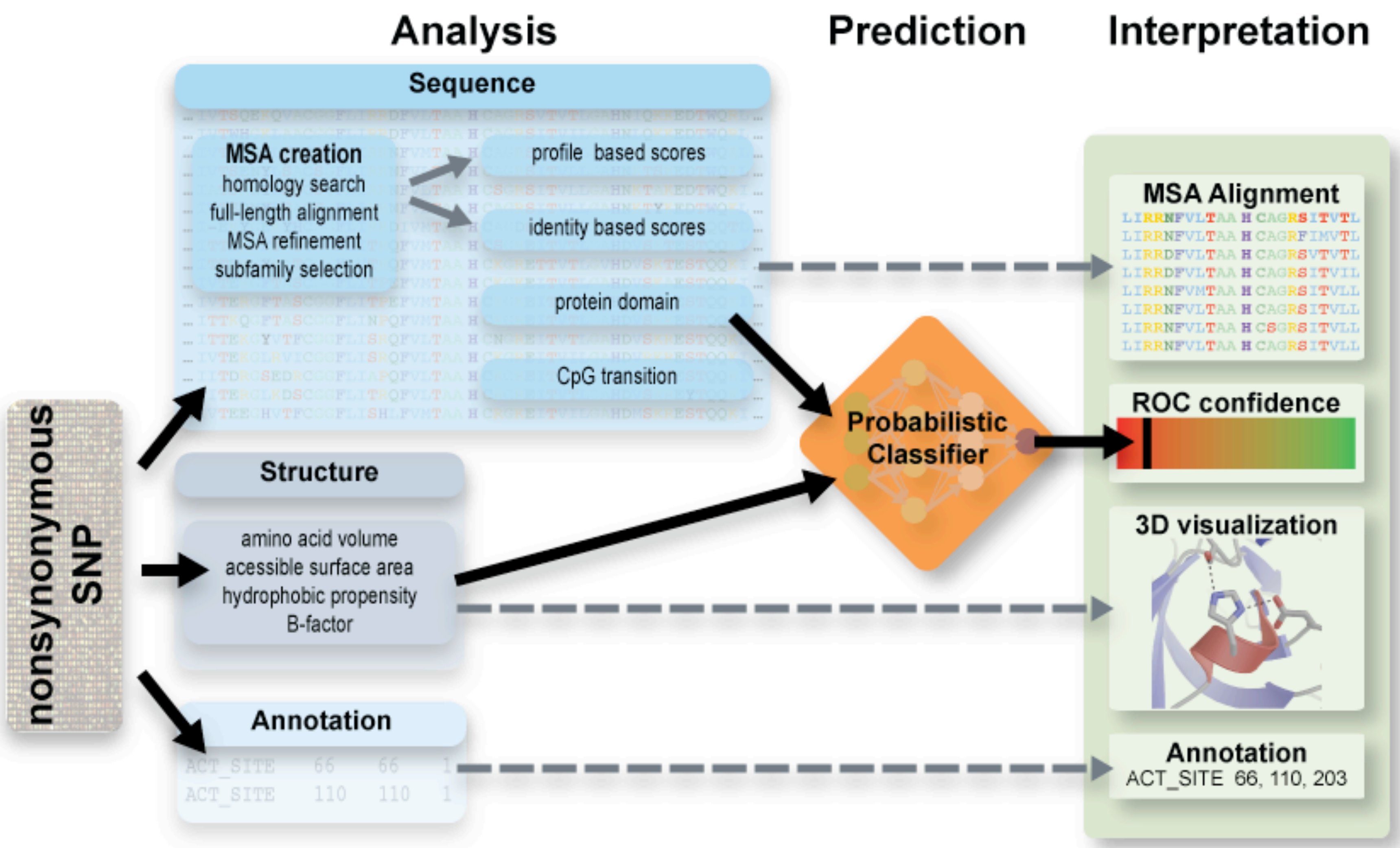
Predictions based on protein structure

- **Most of pathogenic mutations are important for stability.**

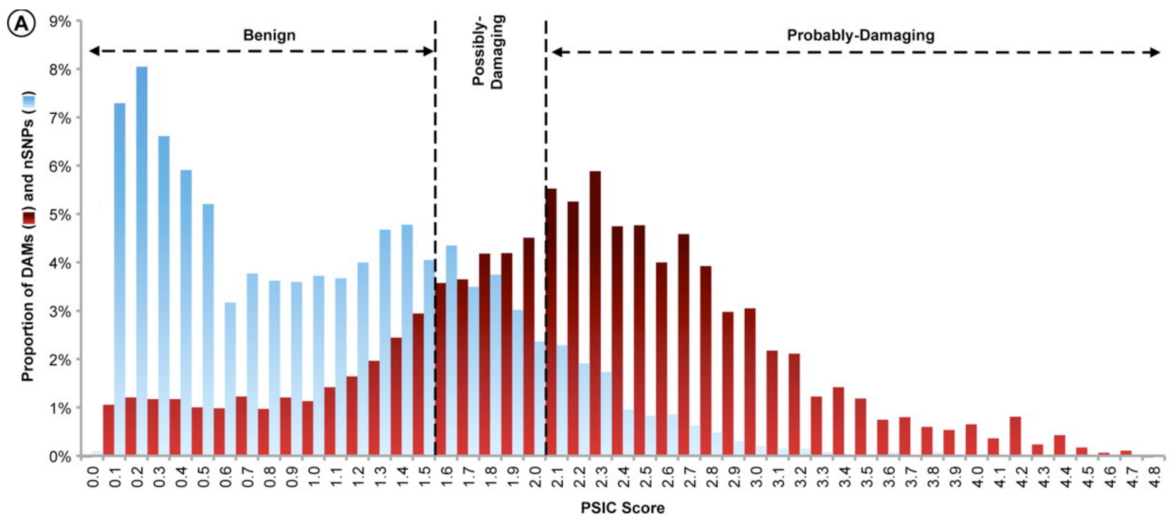
- **Heuristic structural parameters help with predictions (albeit less than comparative genomics)**

PolyPhen-2



www.genetics.bwh.harvard.edu/pph2 Adzhubei, et al. Nature Methods 2010

Incorporation of PolyPhen-2 scores into VT-test



Kumar S et al.

Genome Research 2009

We incorporated weights approximating these distributions into the test for alleles with frequency below 1%

$$z(T) = \sum_{i=1}^{m}\sum_{j=1}^{n} \xi_i^T C_{ij}(\pi_j - \bar{\pi}) \Big/ \left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\xi_i^T C_{ij}\right)^2\right]^{\frac{1}{2}}$$

Price, Kryukov et al., AJHG 2010

Power

Power of Various Approaches Using Simulated Phenotypes

	T1	T5	WE	VT	VTP
$\alpha = 0.001$	0.135	0.180	0.097	0.205	0.257
$\alpha = 0.05$	0.547	0.502	0.545	0.598	0.686

Results for Three Empirical Data Sets

	T1	T5	WE	VT	VTP
TG	0.013	0.00009	0.0024	0.00036	0.00006
T1D	0.001	0.0000006	0.0000009	0.0000012	0.0000001
BMI	0.041	0.064	0.014	0.013	0.0027

This is a general approach

- Prediction scores can be easily incorporated into other tests such as WSS, CMC, RVE, C-alpha etc.

- Other available prediction methods include SIFT, Pmut, SNAP, SNPs3D, GERP etc.

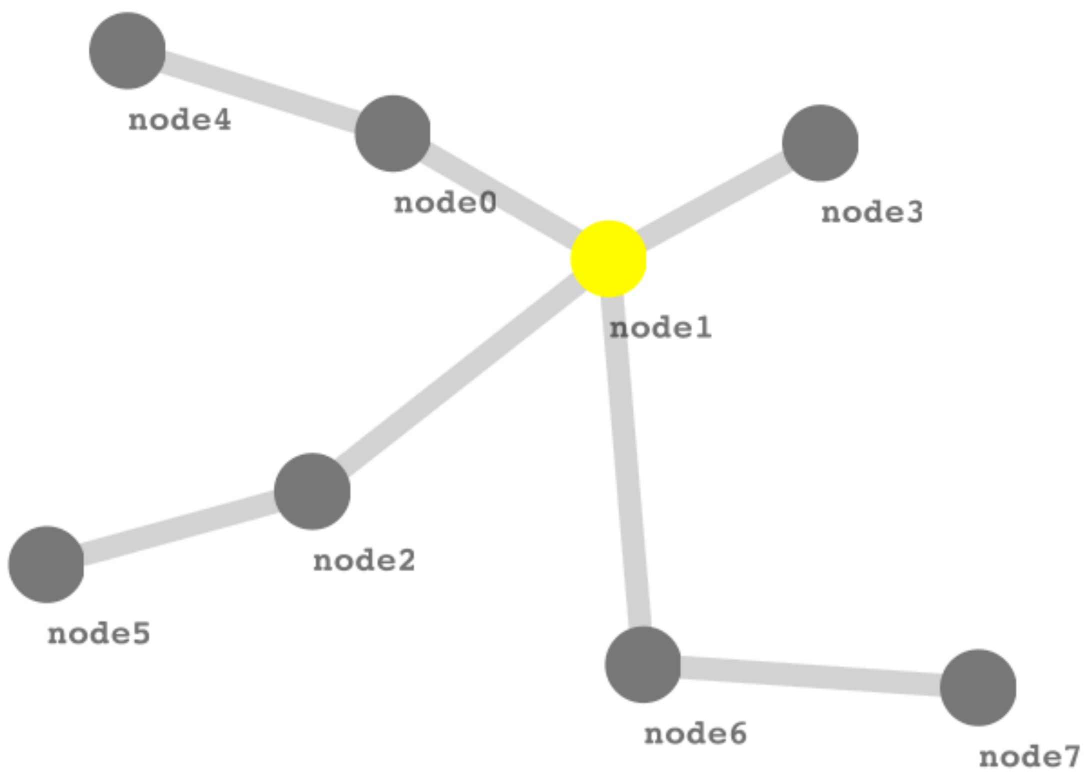
- The same approach cab take into account sequencing errors.

We are likely to be underpowered to detect the effect of individual genes on traits

- Combining signal from multiple genes can dramatically increase power

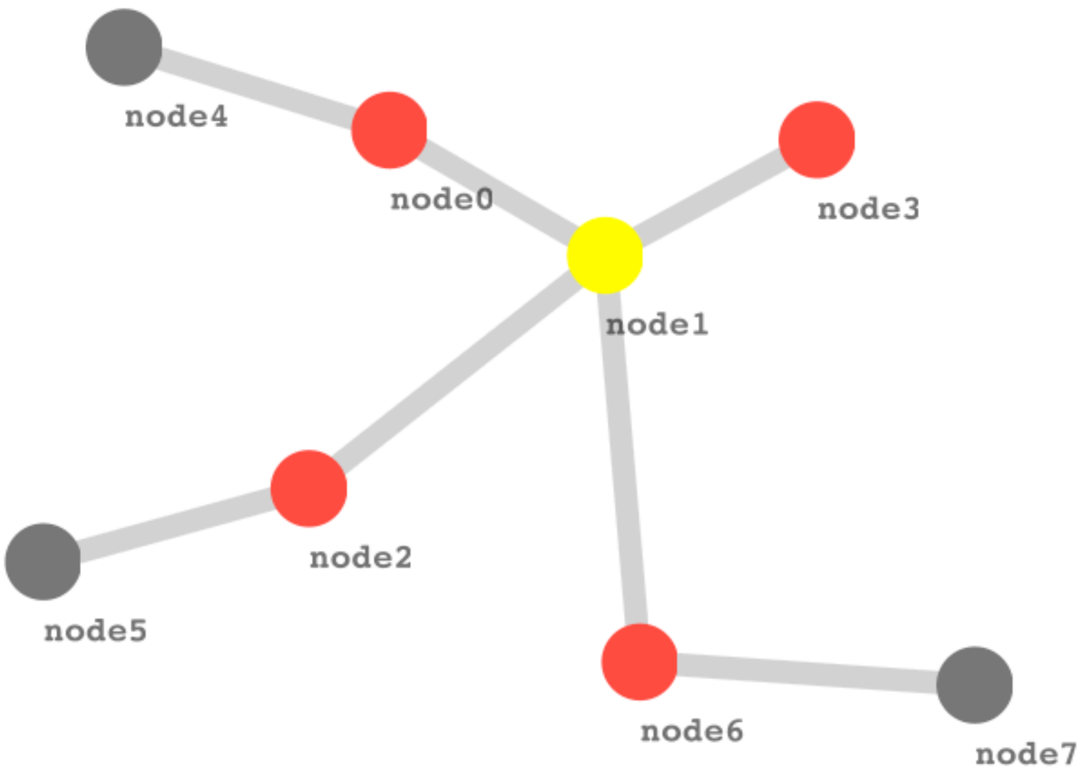
- Although we do not know the right pathways, we can attempt constructing them automatically

SNIPE method



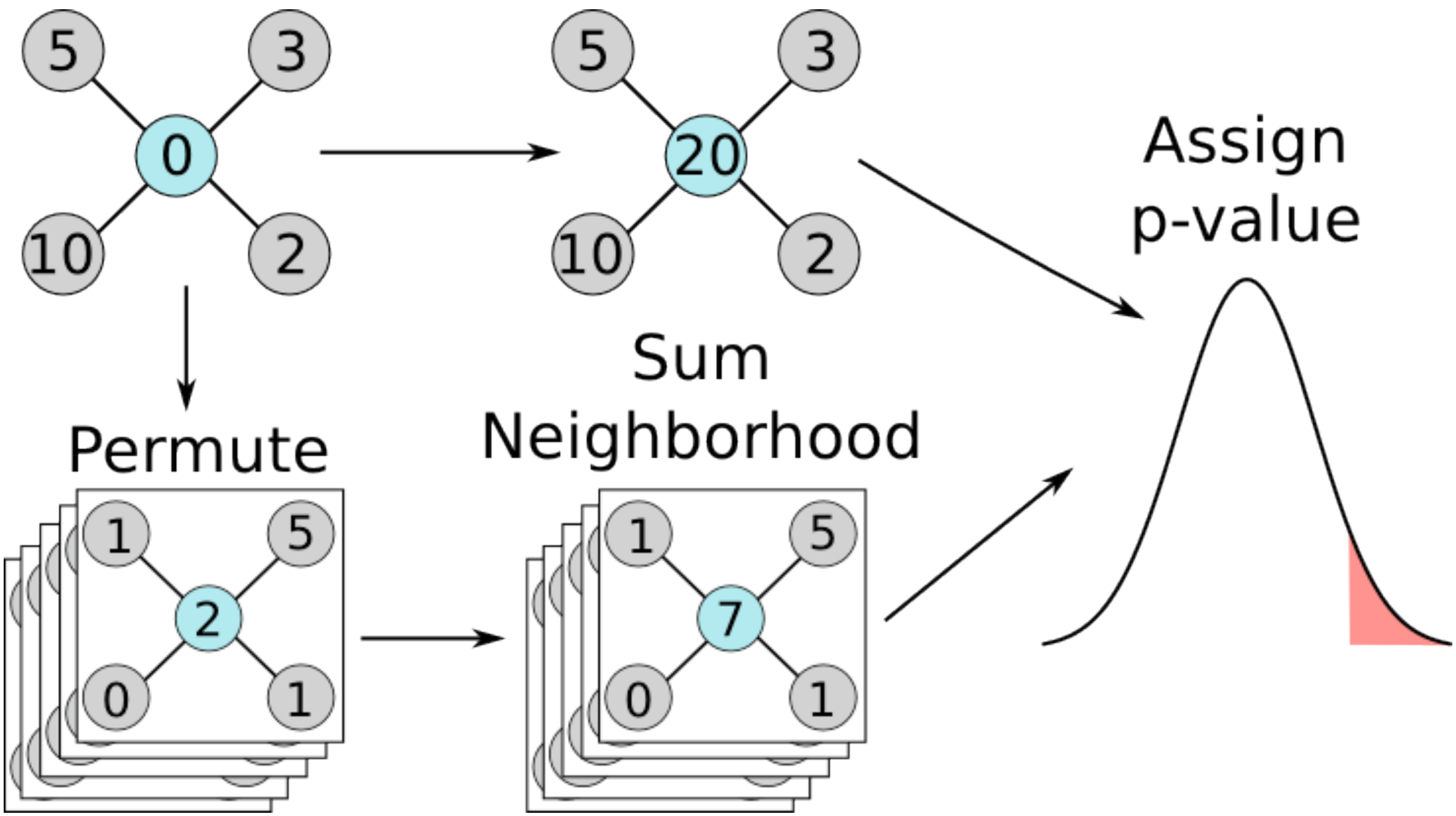
http://string.embl.de/

SNIPE method



http://string.embl.de/

SNIPE method



Acknowledgments

The lab: Gregory Kryukov, Alex Shpunt, Adam Kiezun, Ivan Adzhubei, Victor Spirin, Steffen Schmidt, David Nusinow, Daniel Jordan

HSPH, BWH, MGH
Alkes Price, Lee-Jen Wei, Paul de Bakker, Shaun Purcell

Literature and software

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. "Pooled Association Tests for Rare Variants in Exon-Resequencing Studies." Am J Hum Genet. 2010

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. "A method and server for predicting damaging missense mutations." Nature Methods 2010

Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. "Power of deep, all-exon resequencing for discovery of human trait genes." PNAS 2009

Kryukov GV, Pennacchio LA, Sunyaev SR. "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." Am J Hum Genet. 2007

genetics.bwh.harvard.edu/pph2

genetics.bwh.harvard.edu/rare_variants