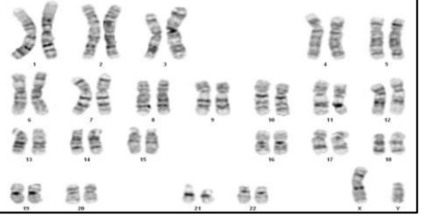
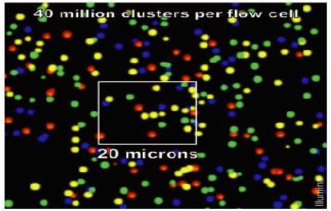
生物信息学：导论与方法
Bioinformatics: Introduction and Methods

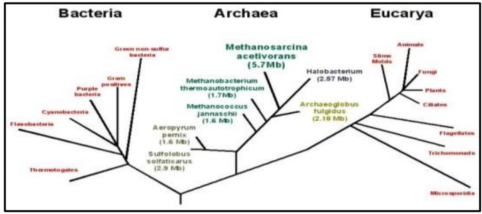
Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University

https://www.coursera.org/course/pkubioinfo



```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



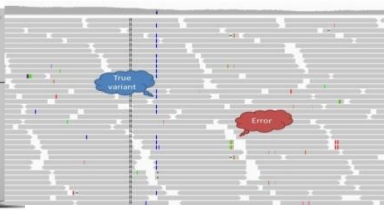

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Slime Molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

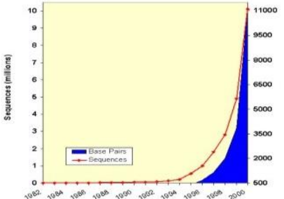
Week 6: Functional prediction of genetic variations

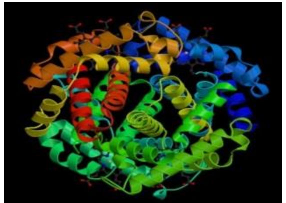
北京大学生物信息学中心 魏丽萍

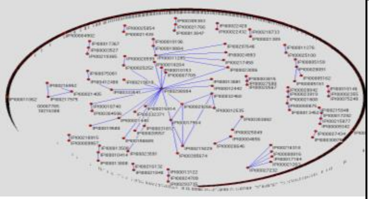
Liping Wei, Ph.D.

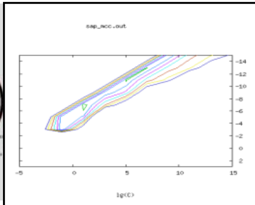
Center for Bioinformatics, Peking University

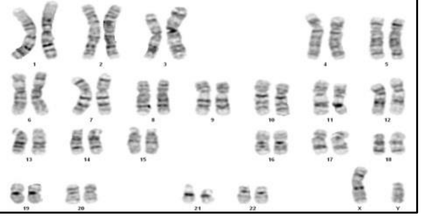




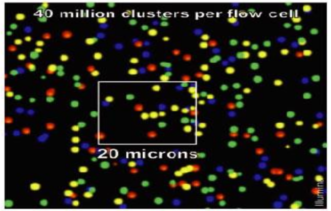


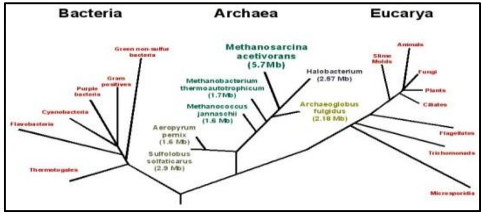






```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



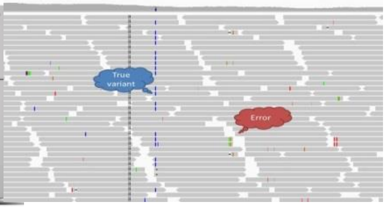

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Slime Molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

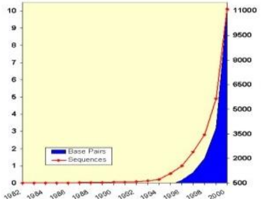
Unit 3: Conservation-based and Rule-based methods: SIFT & PolyPhen

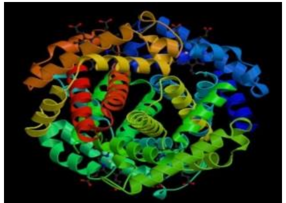
北京大学生物信息学中心 魏丽萍

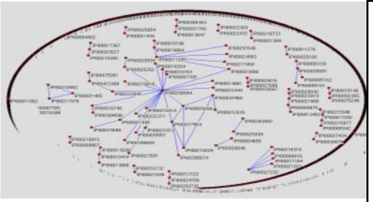
Liping Wei, Ph.D.

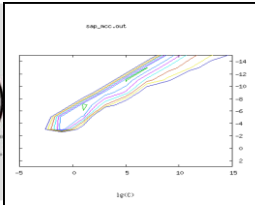
Center for Bioinformatics, Peking University











Questions revisited

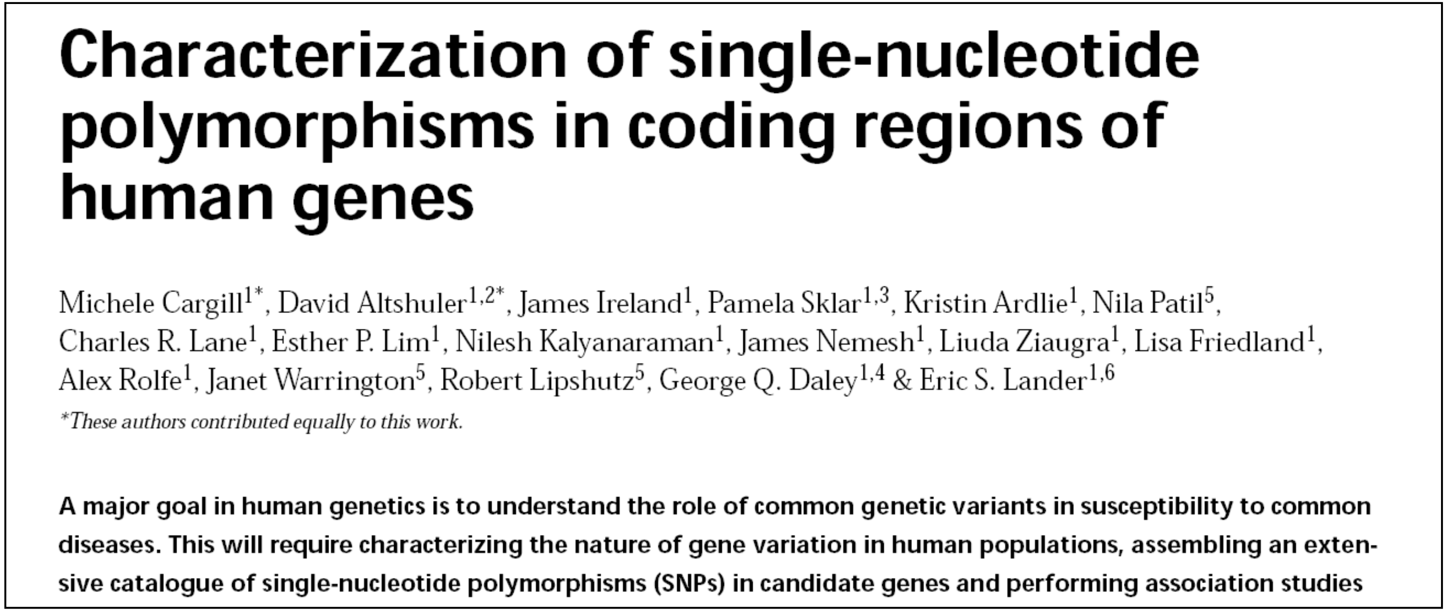
- What features differentiate disease-causing missense SNVs from neutral ones?

- How can we use these features to predict whether a missense SNV is disease-causing?

How would YOU predict the functional and phenotypic effects of an amino acid change?

1999: An early attempt based on BLOSUM substitution matrix

- Assumption: if the substitution score between a variant residue and the wild type residue is positive, then the variant is neutral. If the substitution score is negative, then the variant is deleterious.



Characterization of single-nucleotide polymorphisms in coding regions of human genes

Michele Cargill[1*], David Altshuler[1,2*], James Ireland[1], Pamela Sklar[1,3], Kristin Ardlie[1], Nila Patil[5], Charles R. Lane[1], Esther P. Lim[1], Nilesh Kalyanaraman[1], James Nemesh[1], Liuda Ziaugra[1], Lisa Friedland[1], Alex Rolfe[1], Janet Warrington[5], Robert Lipshutz[5], George Q. Daley[1,4] & Eric S. Lander[1,6]

*These authors contributed equally to this work.

A major goal in human genetics is to understand the role of common genetic variants in susceptibility to common diseases. This will require characterizing the nature of gene variation in human populations, assembling an extensive catalogue of single-nucleotide polymorphisms (SNPs) in candidate genes and performing association studies

More successful methods since 2001

- Conservation-based methods (e.g., SIFT)
- Rule-based methods (e.g., PolyPhen)
- Machine learning classifier-based methods (e.g., PolyPhen2, SAPRED)

Sort Intolerant From Tolerant substitutions (SIFT)

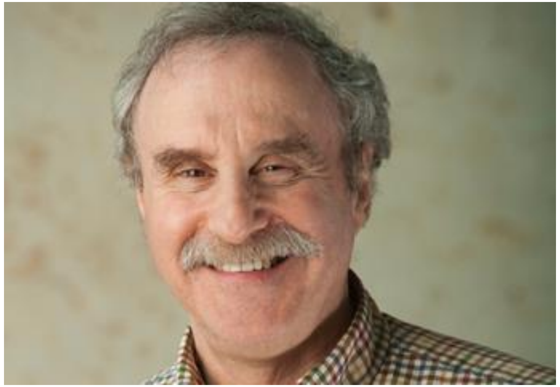
http://sift.jcvi.org/

Important positions (such as active sites) tend to be conserved in the protein family across species.

- Mutations at well-conserved positions tend to be deleterious.

Some positions have a high degree of diversity across species.

- Mutations at these positions tend to be neutral.



www.hhmi.org

Given a protein sequence and an amino acid variation:

Step 1. Search for the most similar sequences and add iteratively

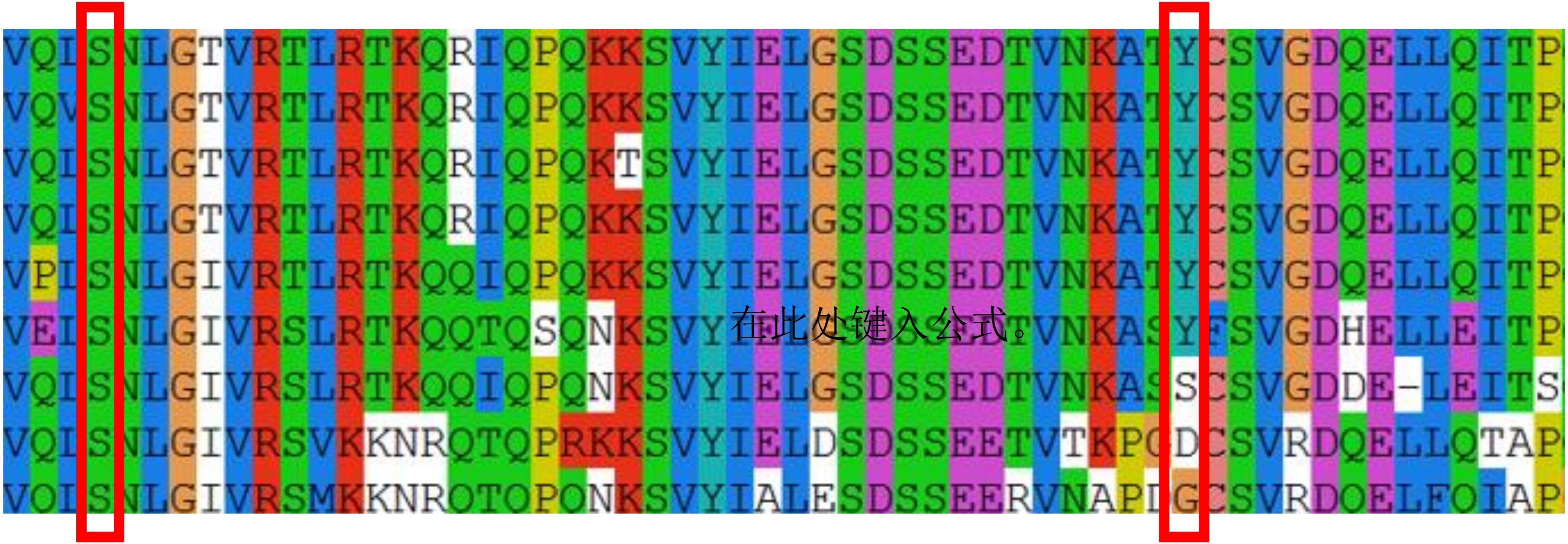
- Sequence search database: SWISS-PROT
- PSI-blast is run for four iterations with parameters –e 0.0001 and –h 0.002.
- PSI-blast results are grouped together if they are >90% identical in the regions aligned, and a consensus sequence is made for each group.
- MOTIF is used to find conserved regions and generate consensus sequences.
- Iteratively add more sequences until conservation in the conserved regions decreases.

Conservation at position *c* is defined as:

$$R_c = log_2\, 20 - \sum_{\{20\ a.a.\}} p_{ca} log\, p_{ca}$$

Ng & Henikoff, Predicting Deleterious Amino Acid Substitutions, Genome Research, 2001

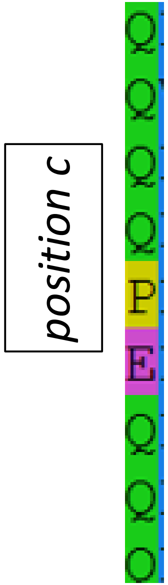
Step 2. Extract the multiple alignment of these chosen sequences



```
VQISNLGTVRTLRTKQRIQPQKKSVYIELGSDSSEDTVNKATYCSVGDQELLQITP
VQVSNLGTVRTLRTKQRIQPQKKSVYIELGSDSSEDTVNKATYCSVGDQELLQITP
VQISNLGTVRTLRTKQRIQPQKTSVYIELGSDSSEDTVNKATYCSVGDQELLQITP
VQISNLGTVRTLRTKQRIQPQKKSVYIELGSDSSEDTVNKATYCSVGDQELLQITP
VPISNLGIVRTLRTKQQIQPQKKSVYIELGSDSSEDTVNKATYCSVGDQELLQITP
VEISNLGIVRSLRTKQQTQSQNKSVYIELGSDSSEDTVNKASYFSVGDHELLEITP
VQISNLGIVRSLRTKQQIQPQNKSVYIELGSDSSEDTVNKASSCSVGDDE-LEITS
VQISNLGIVRSVKKNRQTQPRKKSVYIELDSDSSEETVTKPGDCSVRDQELLQTAP
VQISNLGIVRSMKKNRQTQPQNKSVYIALESDSSEERVNAPIGCSVRDQELFQIAP
```

在此处键入公式。

Step 3. Convert to a Position Specific Scoring Matrix (PSSM) & calculate probability



N_c: total number of sequences aligned at position c

g_{ca}: sequence-weighted frequency that amino acid a appears at position c

(if g_{c-} is the frequency of gaps observed at position c, then for all 20 amino acids, g_{ca} is increased by $g_{c-}/20$.)

f_{ca}: pseudocount of amino acid a at position c, calculated from a 13-component Dirichlet mixture. (see Sjolander, K. et al., 1996, CABIOS, 12:327)

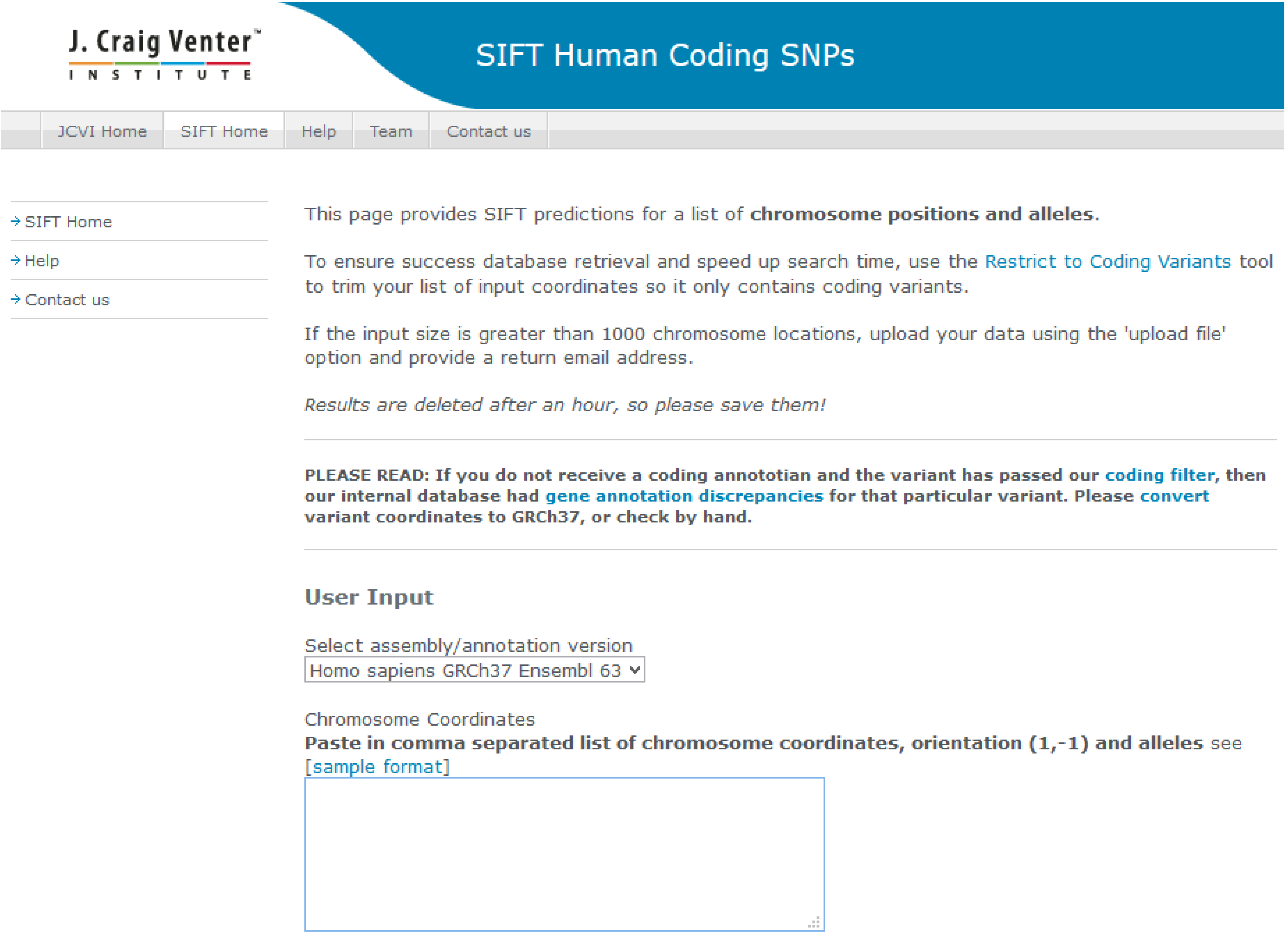
B_c: 0 at an invariant position, or $\exp(\sum_a (g_{ca} * r_a))$ at a variant position

$$Probability\ of\ amino\ acid\ a\ at\ position\ c = \frac{N_c}{(N_c + B_c)} * g_{ca} + \frac{B_c}{(N_c + B_c)} * f_{ca}$$

Divided by the $probability$ of the consensus amino acid to get the $normalized\ probability$

$Normalized\ probability\ of\ the\ amino\ acid\ variant \leq$ 0.05: damaging; >0.05: tolerated

Ng & Henikoff, Predicting Deleterious Amino Acid Substitutions, Genome Research, 2001



SIFT Human Coding SNPs

This page provides SIFT predictions for a list of **chromosome positions and alleles**.

To ensure success database retrieval and speed up search time, use the Restrict to Coding Variants tool to trim your list of input coordinates so it only contains coding variants.

If the input size is greater than 1000 chromosome locations, upload your data using the 'upload file' option and provide a return email address.

Results are deleted after an hour, so please save them!

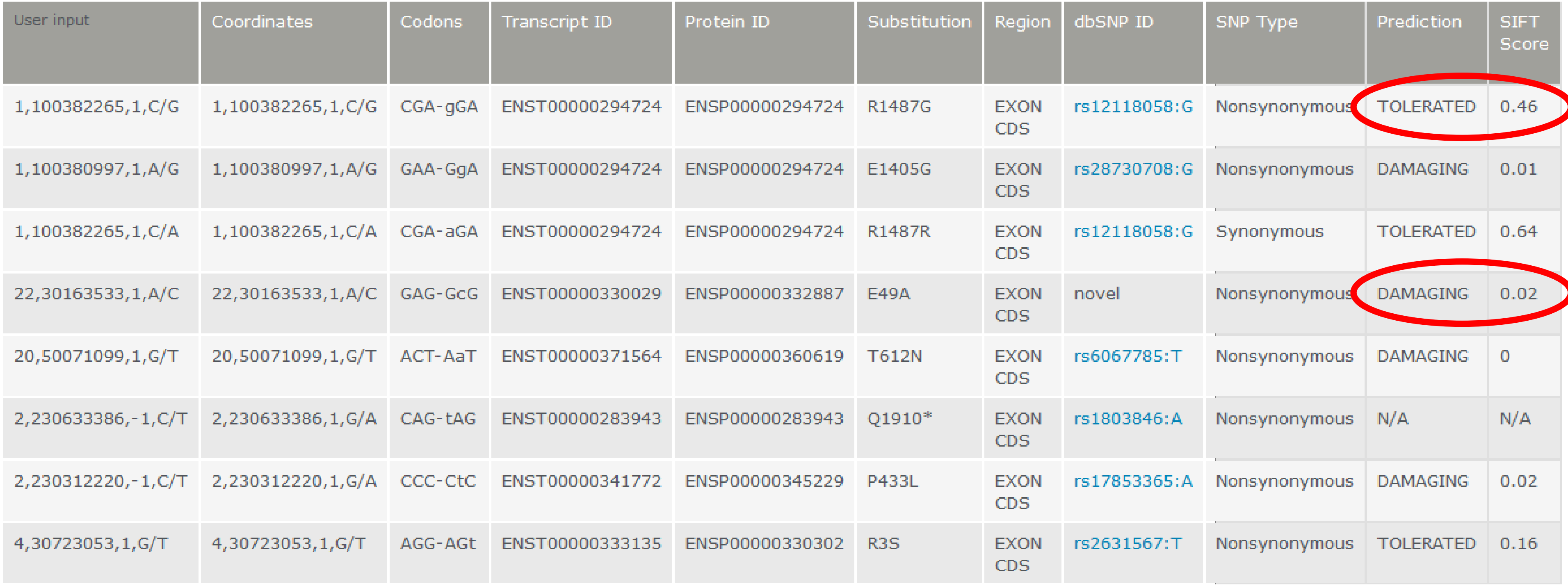
PLEASE READ: If you do not receive a coding annototian and the variant has passed our coding filter, then our internal database had gene annotation discrepancies for that particular variant. Please convert variant coordinates to GRCh37, or check by hand.

User Input

Select assembly/annotation version
Homo sapiens GRCh37 Ensembl 63

Chromosome Coordinates
Paste in comma separated list of chromosome coordinates, orientation (1,-1) and alleles see [sample format]

Prediction results



User input	Coordinates	Codons	Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction	SIFT Score
1,100382265,1,C/G	1,100382265,1,C/G	CGA-gGA	ENST00000294724	ENSP00000294724	R1487G	EXON CDS	rs12118058:G	Nonsynonymous	TOLERATED	0.46
1,100380997,1,A/G	1,100380997,1,A/G	GAA-GgA	ENST00000294724	ENSP00000294724	E1405G	EXON CDS	rs28730708:G	Nonsynonymous	DAMAGING	0.01
1,100382265,1,C/A	1,100382265,1,C/A	CGA-aGA	ENST00000294724	ENSP00000294724	R1487R	EXON CDS	rs12118058:G	Synonymous	TOLERATED	0.64
22,30163533,1,A/C	22,30163533,1,A/C	GAG-GcG	ENST00000330029	ENSP00000332887	E49A	EXON CDS	novel	Nonsynonymous	DAMAGING	0.02
20,50071099,1,G/T	20,50071099,1,G/T	ACT-AaT	ENST00000371564	ENSP00000360619	T612N	EXON CDS	rs6067785:T	Nonsynonymous	DAMAGING	0
2,230633386,-1,C/T	2,230633386,1,G/A	CAG-tAG	ENST00000283943	ENSP00000283943	Q1910*	EXON CDS	rs1803846:A	Nonsynonymous	N/A	N/A
2,230312220,-1,C/T	2,230312220,1,G/A	CCC-CtC	ENST00000341772	ENSP00000345229	P433L	EXON CDS	rs17853365:A	Nonsynonymous	DAMAGING	0.02
4,30723053,1,G/T	4,30723053,1,G/T	AGG-AGt	ENST00000333135	ENSP00000330302	R3S	EXON CDS	rs2631567:T	Nonsynonymous	TOLERATED	0.16

- Score cutoff: 0.05

Accuracy of SIFT

False Negative rate:	31%
False Positive rate:	20%
Coverage:	60%

Ng, P. C. Henikoff, S. Annu Rev Genomics Hum Genet. 2006;7:61-80.

Definitions of accuracy

		Test Outcome			
		Test Positive	Test Negative		
Truth ("Gold standard")	Positive	**True Positive** (hit)	**False Negative** (miss)	**True Positive Rate** (TPR) = **Sensitivity** = **Recall** = TP / (TP+FN)	**False negative rate** (FNR) = **Type II error (β)** = 1- sensitivity = FN / (TP+FN)
	Negative	**False Positive** (false alarm)	**True Negative** (correct rejection)	**True Negative Rate** (TNR) = **Specificity** = TN / (TN+FP)	**False positive rate** (FPR) = **Type I error (α)** = 1- specificity = FP / (TN+FP)
		Positive predictive value (PPV) = **Precision** = TP / (TP+FP)	**Negative predictive value** (NPV) = TN / (TN+FN)	**Accuracy** = (TP+TN) / total	
		False discovery rate (**FDR**) = 1 - precision = FP / (TP+FP)			

Polymorphism Phenotyping (PolyPhen): a rule-based method

- PolyPhen predicts impact of amino acid variations based on both multi-sequence alignment AND protein 3D structure features

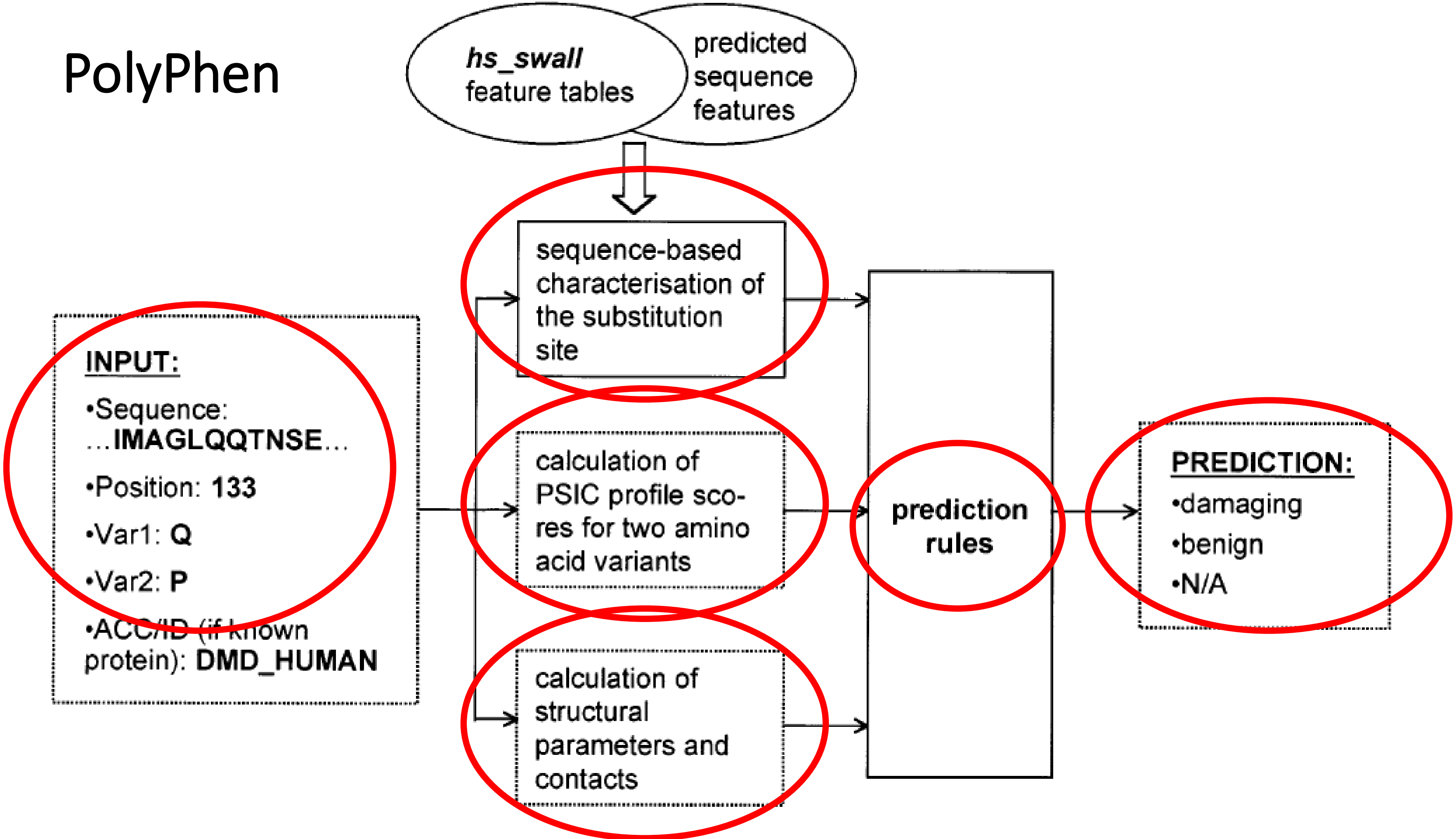
Peer Bork

- Presumptions:
  1. Amino acid variations at conserved positions are more likely to cause functional changes.
  2. Amino acid variations that affect the active sites, interaction sites, solubility, or stability of a protein are likely to affect protein structure.
  3. Changes in protein structure are likely to cause changes in protein function, which are likely to cause changes in phenotype.

Shamil Sunyaev

http://archives.focus.hms.harvard.edu/2005/Feb11_2005/index.html http://www.embl.de/~bork/

PolyPhen



hs_swall feature tables

predicted sequence features

INPUT:

- Sequence: ...**IMAGLQQTNSE**...
- Position: **133**
- Var1: **Q**
- Var2: **P**
- ACC/ID (if known protein): **DMD_HUMAN**

sequence-based characterisation of the substitution site

calculation of PSIC profile scores for two amino acid variants

calculation of structural parameters and contacts

prediction rules

PREDICTION:

- damaging
- benign
- N/A

PolyPhen

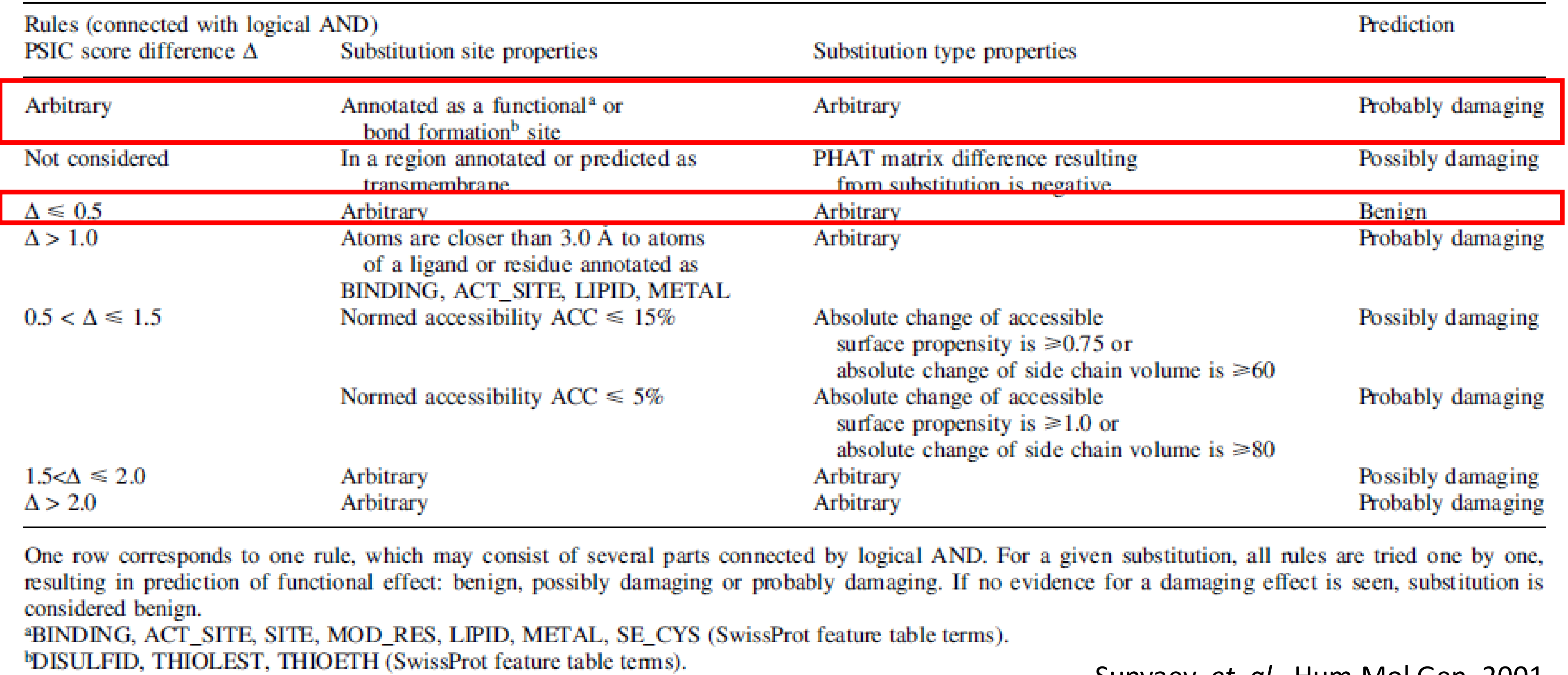
1. Generate multi-sequence alignment of homologous sequences and calculate sequence features and PSIC.

2. Get the protein 3D structure or using homolog modeling to predict its structure

3. Calculate structure-based features of the substitution site

- □disulfide, thiolest or thioeath bond, binding site, active site etc.
- □Secondary structure
- □Solvent accessible surface area
- □$\Phi - \Psi$ dihedral angles
- □Normalized B-factor for the residue
- □Loss of hydrogen bond
- □transmembrane regions
- □coiled coil regions
- □signal peptide regions

4. empirically derived rules to predict damaging or benign



Rules (connected with logical AND)			Prediction
PSIC score difference Δ	Substitution site properties	Substitution type properties	
Arbitrary	Annotated as a functional[a] or bond formation[b] site	Arbitrary	Probably damaging
Not considered	In a region annotated or predicted as transmembrane	PHAT matrix difference resulting from substitution is negative	Possibly damaging
$\Delta \leqslant 0.5$	Arbitrary	Arbitrary	Benign
$\Delta > 1.0$	Atoms are closer than 3.0 Å to atoms of a ligand or residue annotated as BINDING, ACT_SITE, LIPID, METAL	Arbitrary	Probably damaging
$0.5 < \Delta \leqslant 1.5$	Normed accessibility ACC ⩽ 15%	Absolute change of accessible surface propensity is ⩾0.75 or absolute change of side chain volume is ⩾60	Possibly damaging
	Normed accessibility ACC ⩽ 5%	Absolute change of accessible surface propensity is ⩾1.0 or absolute change of side chain volume is ⩾80	Probably damaging
$1.5<\Delta \leqslant 2.0$	Arbitrary	Arbitrary	Possibly damaging
$\Delta > 2.0$	Arbitrary	Arbitrary	Probably damaging

One row corresponds to one rule, which may consist of several parts connected by logical AND. For a given substitution, all rules are tried one by one, resulting in prediction of functional effect: benign, possibly damaging or probably damaging. If no evidence for a damaging effect is seen, substitution is considered benign.

[a] BINDING, ACT_SITE, SITE, MOD_RES, LIPID, METAL, SE_CYS (SwissProt feature table terms).

[b] DISULFID, THIOLEST, THIOETH (SwissProt feature table terms).

Sunyaev, *et. al.*, Hum Mol Gen, 2001

PolyPhen

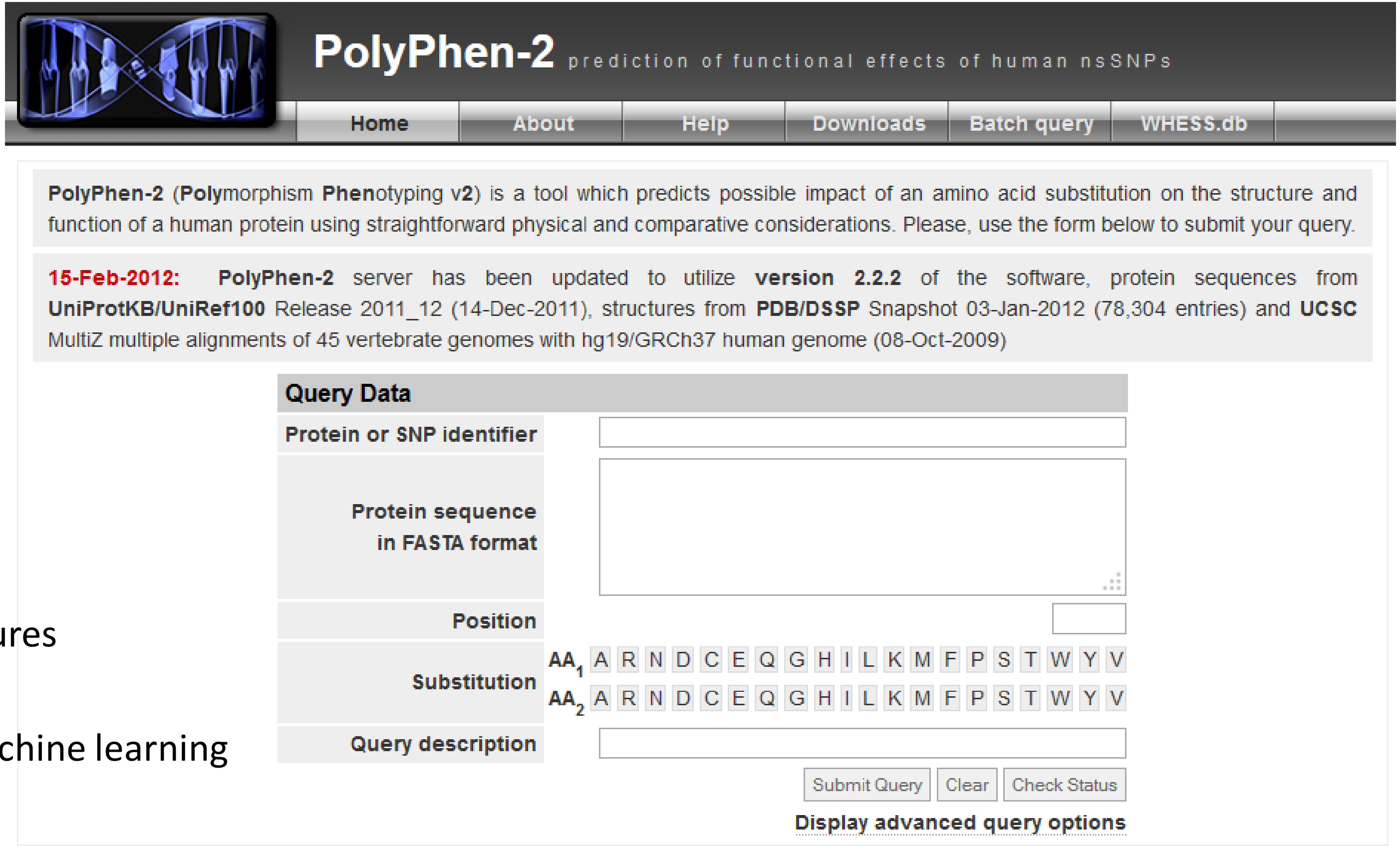
Pros

- improved prediction accuracy when protein 3D structure is available

Cons

- If 3D structure is not available, it can only depend on MSA.
- The rules are empirical.

PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/)



PolyPhen-2 (**Poly**morphism **Phen**otyping **v2**) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.

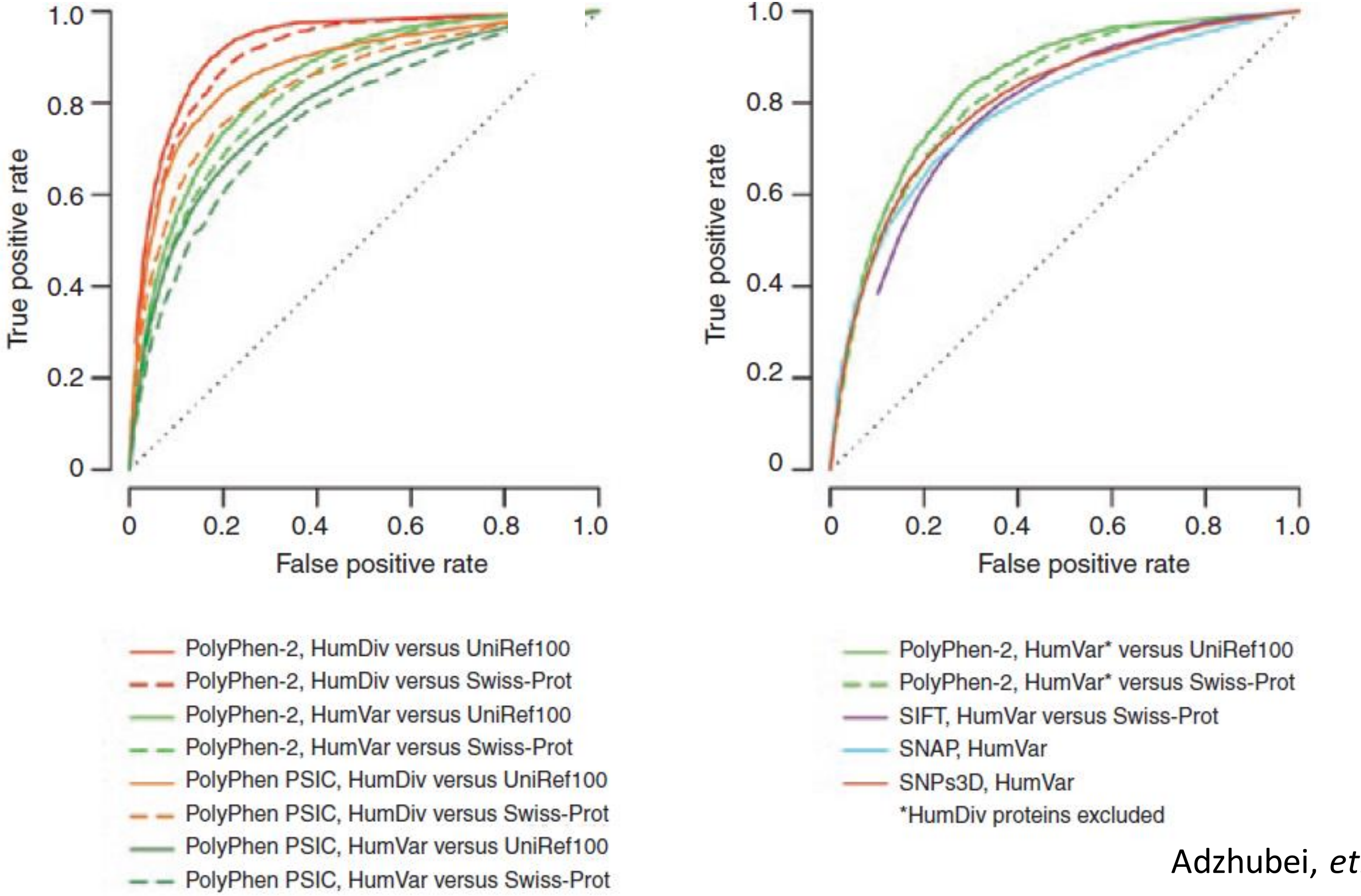
15-Feb-2012: **PolyPhen-2** server has been updated to utilize **version 2.2.2** of the software, protein sequences from **UniProtKB/UniRef100** Release 2011_12 (14-Dec-2011), structures from **PDB/DSSP** Snapshot 03-Jan-2012 (78,304 entries) and **UCSC** MultiZ multiple alignments of 45 vertebrate genomes with hg19/GRCh37 human genome (08-Oct-2009)

Query Data

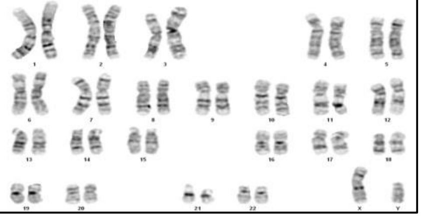
Protein or SNP identifier	
Protein sequence in FASTA format	
Position	
Substitution	**AA_1** A R N D C E Q G H I L K M F P S T W Y V **AA_2** A R N D C E Q G H I L K M F P S T W Y V
Query description	

- □ Use more predictive features
- □ Based on Naïve Bayes machine learning

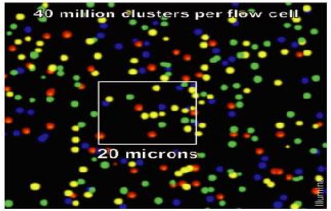
Improved performance compared with PolyPhen

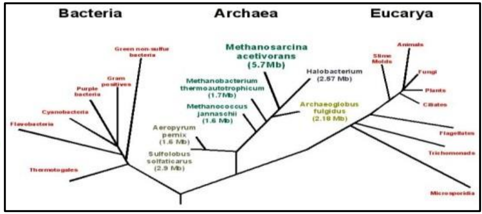


Adzhubei, *et. al.,* Nat Methods, 2010



```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



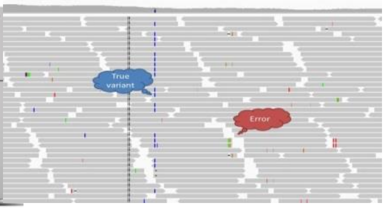

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Slime Molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

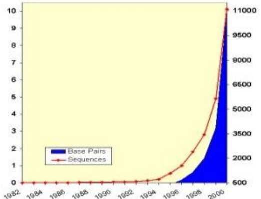
Week 6 Unit 4: Classifier-based methods: SAPRED

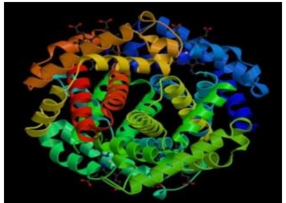
北京大学生物信息学中心 魏丽萍

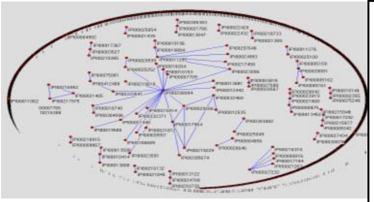
Liping Wei, Ph.D.

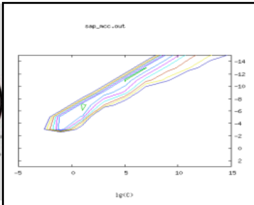
Center for Bioinformatics, Peking University











生物信息学：导论与方法

Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍

Center for Bioinformatics, Peking University

https://www.coursera.org/course/pkubioinfo