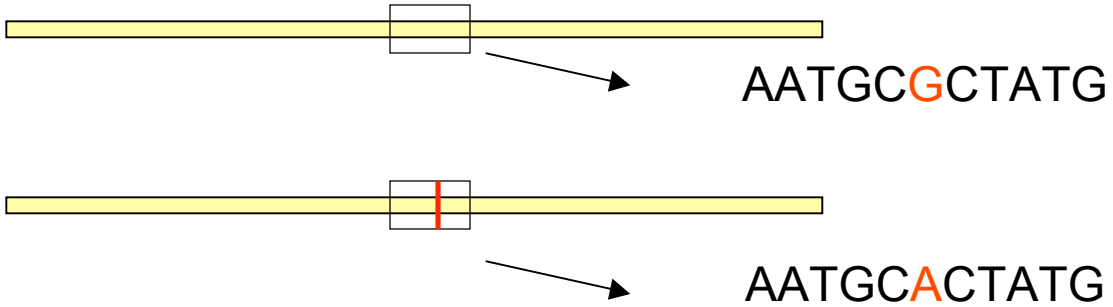
Software Prediction of the Effects of Single Nucleotide Polymorphisms

Angela Collie
CSE 527
December 13, 2004

Objective

Examine the ability of two web-based programs to predict the effect of a single nucleotide polymorphism on a protein.

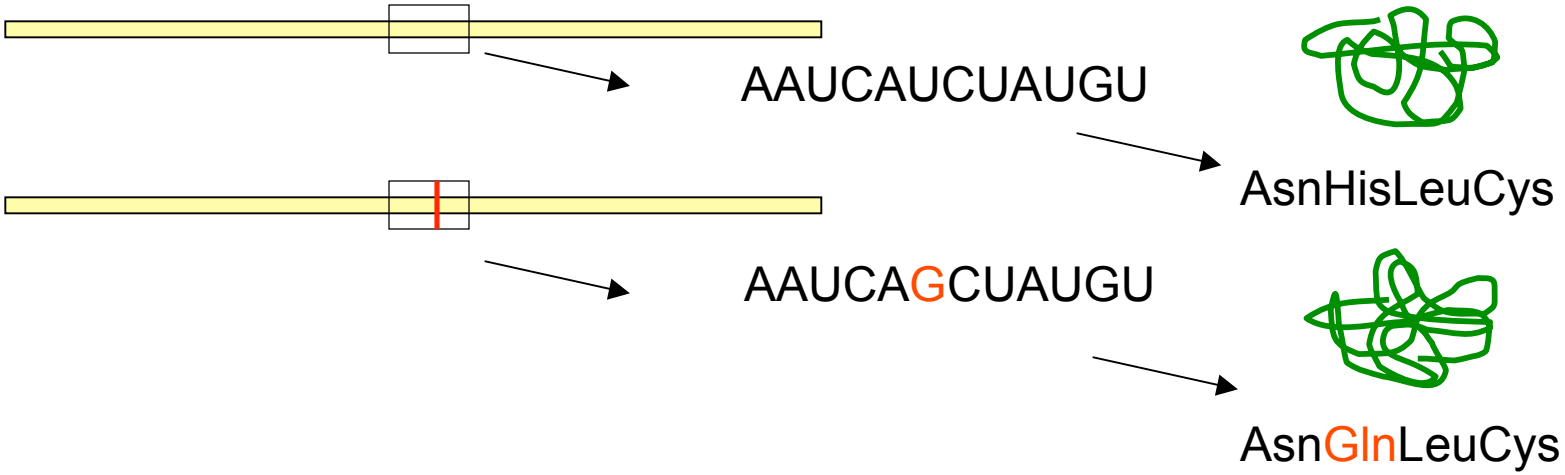
Single Nucleotide Polymorphisms (SNPs)



- 99.9% of the 3.2 billion base pairs in the human genome are the same.
- SNPs are single base pair changes and account for much of the variation.
  - Minor allele is defined as present in >1% of the population.
  - “Common” alleles are present in >10% of the population.
  - There are approximately 11 million SNPs in the genome, corresponding to 1 base pair change every 300 bases.
- Haplotype

Kruglyak L and Nickerson DA. Nat Genet. 2001 Mar;27(3):234-6.

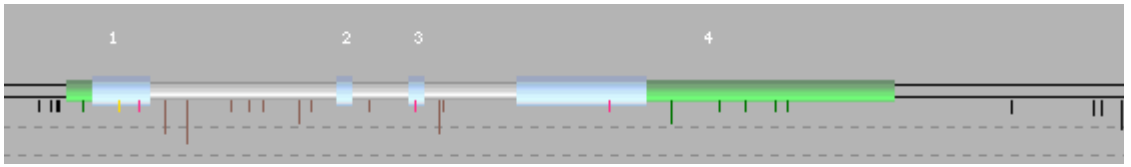
Nonsynonymous SNPs



- nsSNPs are SNPs that are present in the coding region of a gene and result in an amino acid change in the resulting protein.
  - This can affect the 3D structure or interactions with other proteins.
- SNPs in the promoter and exons of a gene are thought to be the most harmful to a protein.

Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

dbSNP

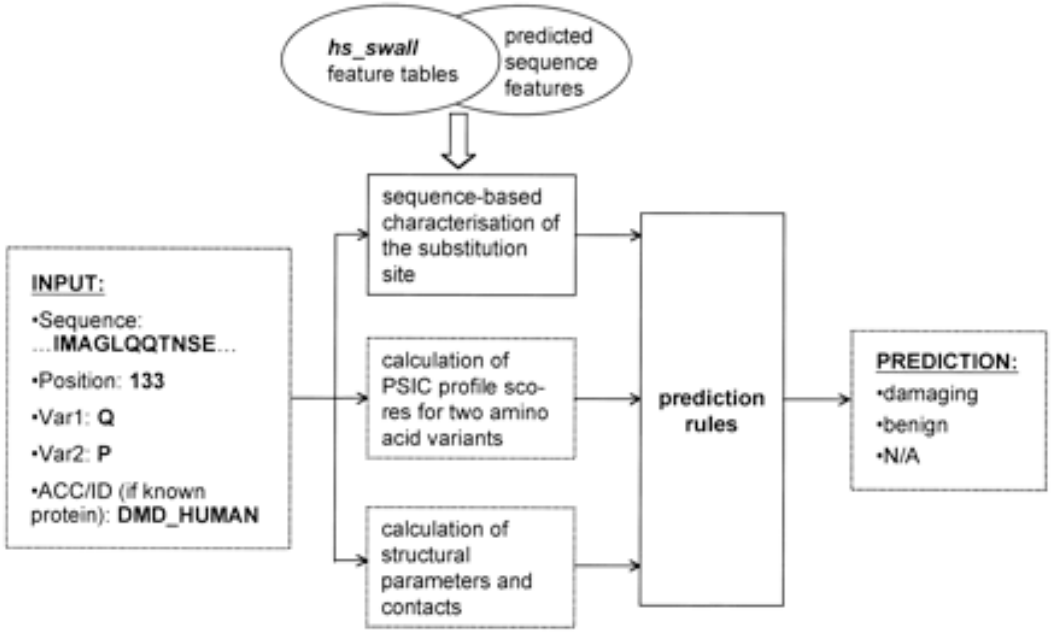


TNF gene: dbSNP provides similar pictorial representation of SNPs

- Public database maintained by the NCBI.
- Must recent build had over 10 million SNPs. 5 million have been validated.
- Data is linked to gene and other NCBI databases, including 3D structure representations for some SNPs.

http://www.genome.utah.edu/genesnps/cgi-bin/frame.cgi?gene_id=440
http://www.ncbi.nlm.nih.gov/SNP/

Polymorphism Phenotyping (PolyPhen)

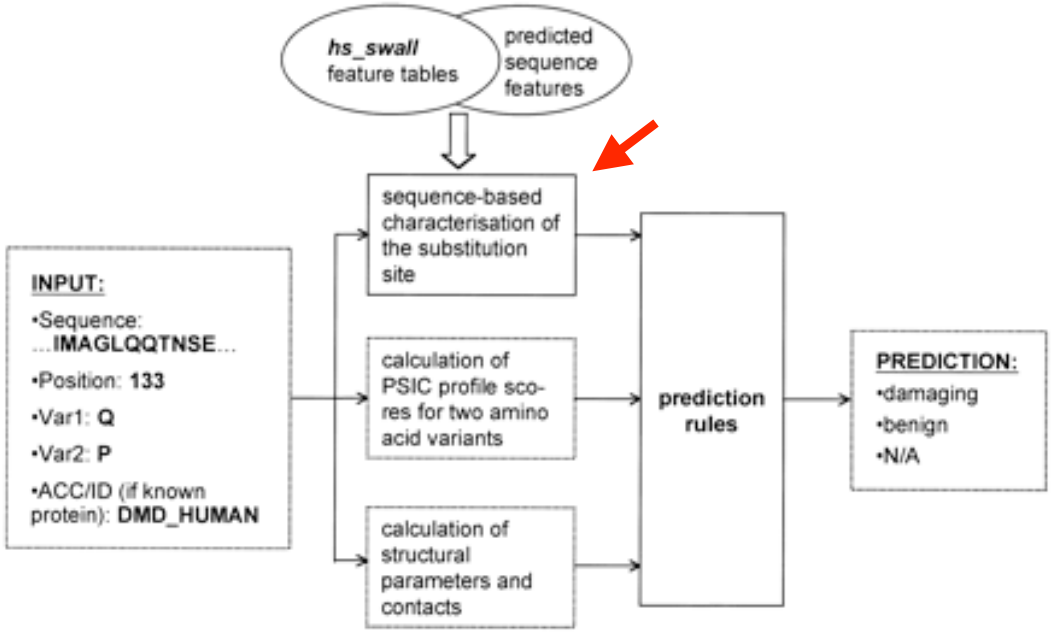


- Web-based tool for predicting the effect of a nsSNP on a protein.
- Utilizes a combination of 3D structural parameters and sequence homology to make prediction based on rules.
- Input is protein sequence (or ID #) and position of amino acid substitution and amino acid variants.
- Returns predictions of "probably damaging," possibly damaging," "benign," and "unknown."

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

PolyPhen Algorithm: Step 1

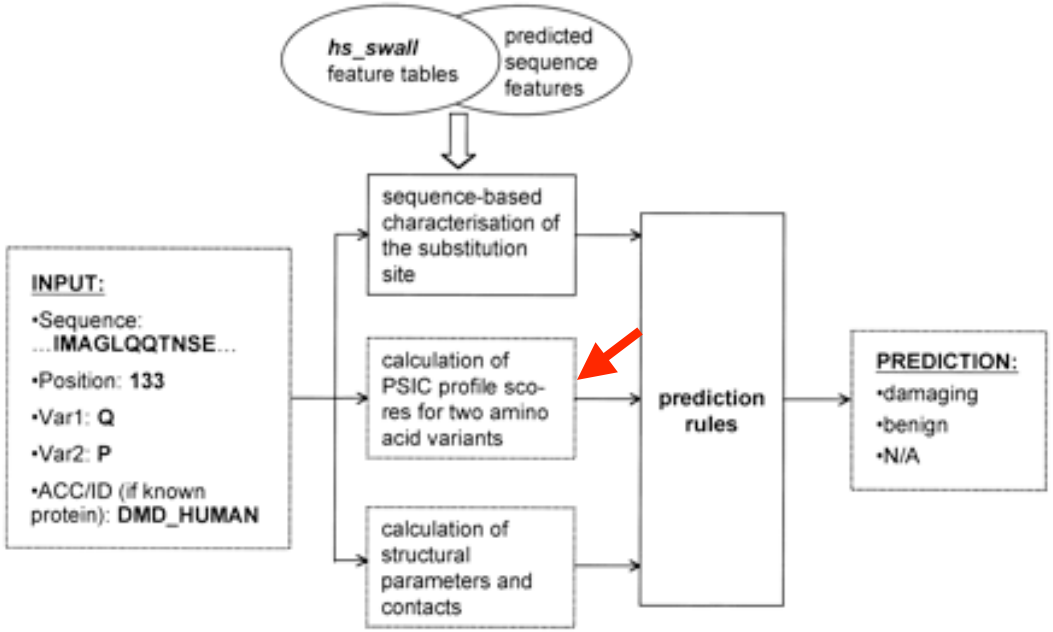


- Characterization of substitution site.
- Checks protein database for protein features.
- Uses several program add-ins to identify transmembrane, coil and signal peptide regions.
- If substitution is in a transmembrane region, a score is calculated to determine effect.

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

PolyPhen Algorithm: Step 2



$$W(a,i) = \ln\left[\frac{p(a,i)}{q_a}\right]$$

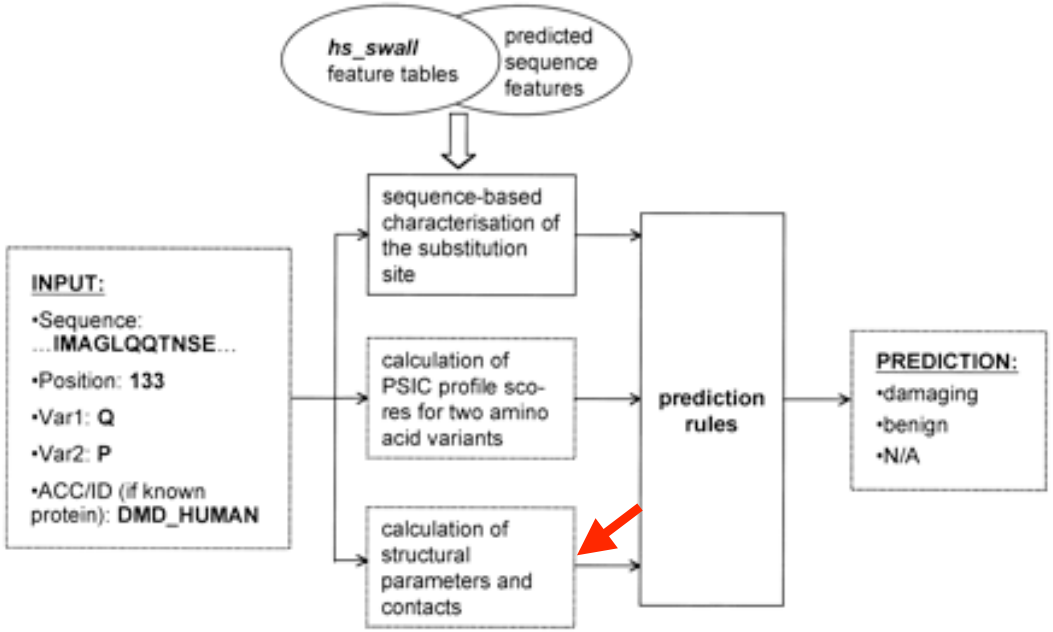
- Sequence homology scores are based on this log likelihood ratio.

- PolyPhen uses BLAST against a protein database to identify sequences with 30-94% homology to input sequence.
- Position-Specific Independent Counts (PSIC) is run. This returns a score that is based on the log likelihood ratio of the amino acid *a* occurring at position *i* compared to the background frequency of amino acid *a*.
- Ratio is corrected to account for the limited number of sequences available and the interdependence of sequences.

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

PolyPhen Algorithm: Step 3



- PolyPhen BLASTs the sequence against the user-chosen PDB or PQS databases to find proteins of sequence identity of at least 50%.
- Several structural parameters are then calculated using a protein database and another add-in.
- Polyphen then checks contacts of the variant amino acid with ligands, interactions between parts of the protein, and critical residues.

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

PolyPhen Algorithm: Rules

Rules (connected with logical AND)			Prediction
PSIC score difference	**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
Less than 0.5	Arbitrary	Arbitrary	Benign
Greater than 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
Between 0.5 and 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
Between 0.5 and 1.5	Normed accessibility ACC 5%	Absolute change of accessible surface propensity is 1.0 or absolute change of side chain volume is 80	Probably damaging
Between 1.5 and 2.0	Arbitrary	Arbitrary	Possibly damaging
Greater than 2.0	Arbitrary	Arbitrary	Probably damaging

Table legend: One row corresponds to one rule, which may consist of several parts connected by logical AND. If no evidence for a damaging effect is seen, substitution is considered benign.

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

Sample PolyPhen Output

Query: TNF P84L

Prediction This variant is predicted to be possibly damaging

Prediction	Available data	Prediction basis	Substitution effect	Prediction data
possibly damaging	FT alignment structure	Structure	1.1.1: structural effect, buried site, hydrophobicity **Hydrophobicity change at buried site**	PSIC score difference: 0.839 normed accessibility: 0.14 hydrophobicity change: 1.07

Remarks

Closest contact with other chains: GLN 125C, distance 2.800 Å

Details

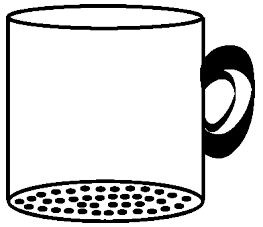
PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS

Score 1	Score 2	|Score1-Score2|	Observations	Diagnostics	Multiple alignment around substitution position
+0.945	+0.106	**0.839**	90	cached	Sequences: Flanks:

http://www.bork.embl-heidelberg.de/PolyPhen/

Sunyaev SR. Protein Eng. 1999 May;12(5):387-94.
Ramensky V. Nucleic Acids Res. 2002 Sep 1;30(17):3894-900.

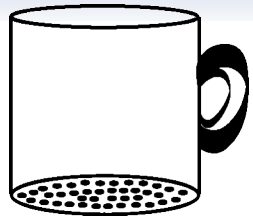
Sorting Intolerant From Tolerant (SIFT)



- Web-based tool for predicting the effect of a nsSNP on a protein.
- Utilizes sequence homology to predict effect.
- Aligned protein sequences are from BLink.
- Input is GI# (unique for protein) or protein sequence. SNP amino acid substitutions and position can also be submitted.
- Returns predictions of "affect protein function" and "tolerated" for each SNP. Also returns normalized score and median sequence information.

Ng PC, Henikoff S. Genome Res. 2001 May;11(5):863-74.
http://blocks.fhcrc.org/sift/SIFT.html

SIFT Algorithm

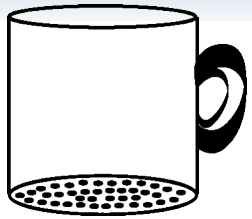


$$p_{ca} = \frac{N_c}{(N_c + B_c)} * g_{ca} + \frac{B_c}{(N_c + B_c)} * f_{ca}$$

- "$p_{ca,}$ the probability of amino acid *a* at position *c*, is a weighted average of the observed amino acid frequencies in the alignment and the estimated unobserved frequencies."
  - N_c is the number of sequences at position *c*.
  - B_c is an exponential function that returns the number of pseudocounts based on amino acid frequencies in a predetermined matrix.
  - g_{ca} is a sequence weighted frequency that *a* appears at *c* in the alignment.
  - f_{ca} is a frequency of pseudocounts.
- Normalization.

Ng PC, Henikoff S. Genome Res. 2001 May;11(5):863-74.
http://blocks.fhcrc.org/sift/SIFT.html

Sample SIFT Output



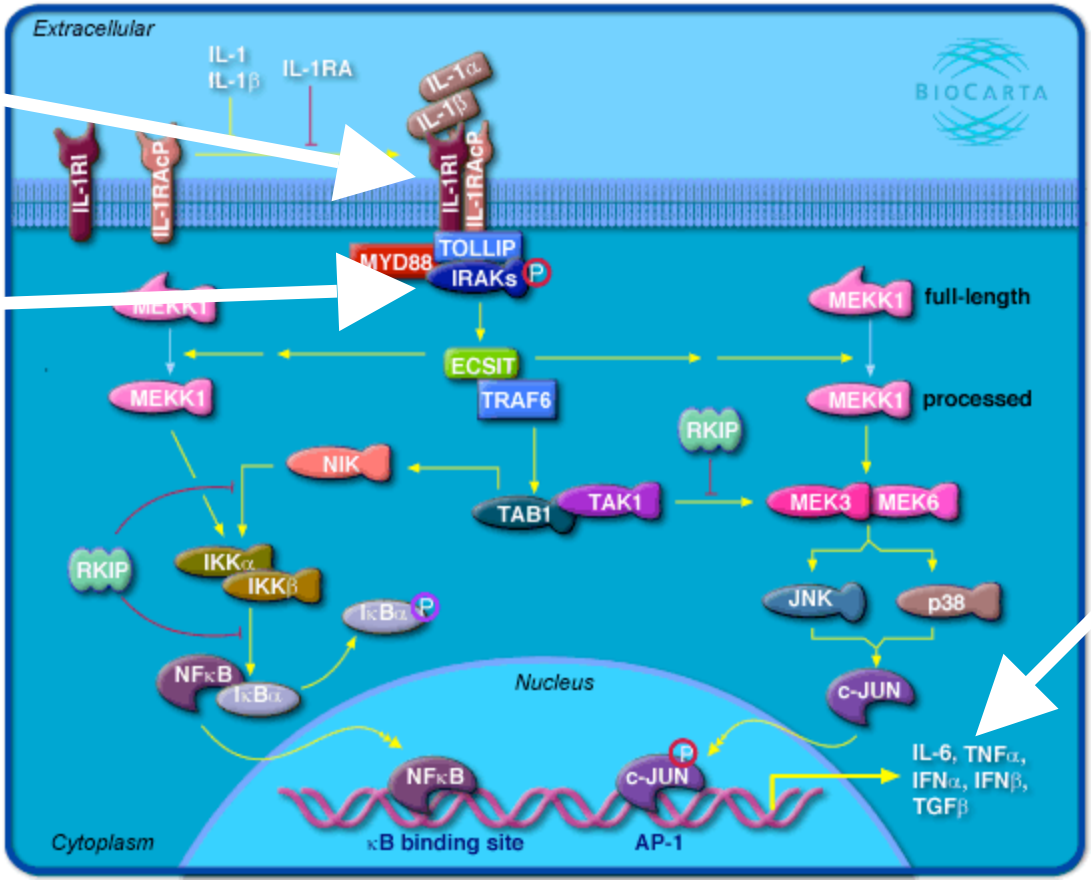
- Threshold for tolerance is 0.5.
- For position 1M
  - Predict not tolerated:ywvtsrqpnlkihgfedca
  - Predict tolerated: M
  - Normalized probabilities for each amino acid can also be obtained
- For substitution: A94T
  - Substitution at pos 94 from A to T is predicted to AFFECT PROTEIN FUNCTION with a score of 0.02.
  - Median sequence conservation: 2.64
  - Sequences represented at this position:48

Ng PC, Henikoff S. Genome Res. 2001 May;11(5):863-74.
http://blocks.fhcrc.org/sift/SIFT.html

Genes of Interest

IL-1R1

IRAK1
IRAK4



TNF
IL-6

http://www.biocarta.com/pathfiles/h_il1rPathway.asp

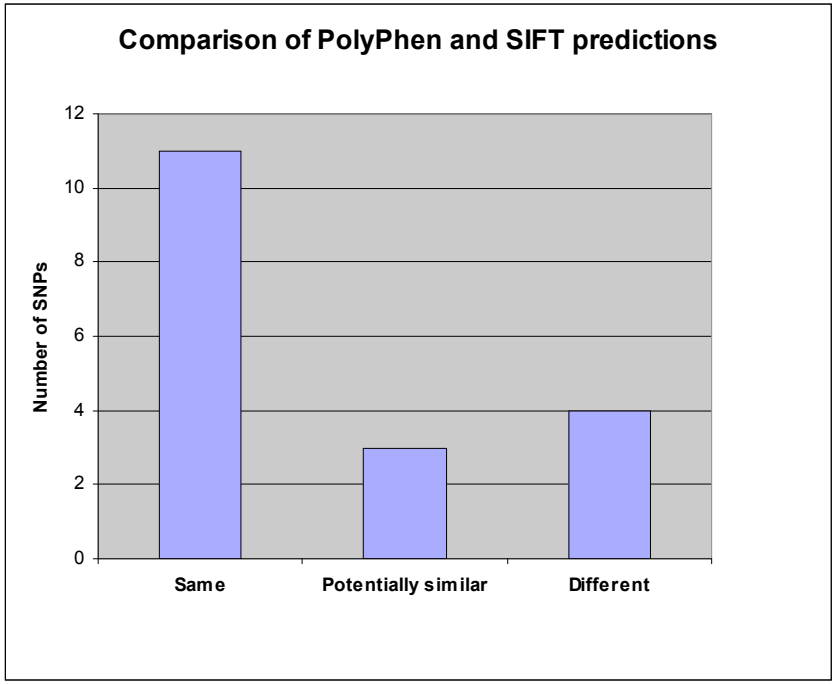
PolyPhen AND SIFT Results

Gene	Reference aa residue / SNP aa residue	PolyPhen prediction	SIFT prediction
IL1R1	Ala [A]/Gly [G]	Benign	Tolerated
IRAK1	Ser [S]/Leu [L]	Possibly damaging	Tolerated
	Arg [R]/Gly [G]	Possibly damaging	Affect Protein Function
	Cys [C]/Ser [S]	Probably damaging	Tolerated
	Phe [F]/Ser [S]	Benign	Affect Protein Function
	Arg [R]/His [H]	Benign	Tolerated
	Thr[T]/Ile [I]	Benign	Tolerated
IRAK4	Ser [S]/Arg [R]	Benign	Affect Protein Function
	His [H]/Arg [R]	Benign	Tolerated
	Ala [A]/Thr [T]	Benign	Tolerated
TNF	His[H]/Asn [N]	Benign	Tolerated
	Pro [P]/Leu [L]	Possibly damaging	Tolerated
	Ala [A]/Thr [T]	Benign	Affect Protein Function
	Ile [I]/Asn [N]	Probably damaging	Affect Protein Function
IL6	Pro [P]/Ser [S]	Benign	Tolerated
	Leu [L]/Pro [P]	Probably damaging	Affect Protein Function
	Asp [D]/Val [V]	Benign	Tolerated
	Asp [D]/Glu [E]	Benign	Tolerated

IRAK4 is recently described so there may not have been enough sequences for prediction.

Software Comparison

- 18 SNPs were examined in 5 genes.
- PolyPhen and SIFT have different predictions for several SNPs.
- None of these SNPs have been described in OMIM or in the literature.



Conclusion

- Two web-based software programs were used to predict the effect of 18 SNPs on 5 genes in the IL-1B signaling pathway.
- Software predictions must be verified with experimental data.
- Predictions will improve with additional homologous sequences and three-dimensional structure.