Project 1.3: Analyzing the Protein Coding Mutations in the Zimmerome

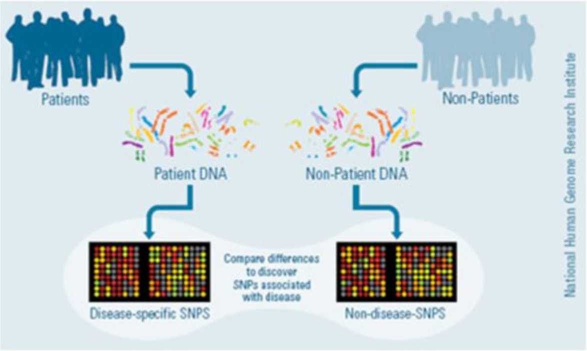
Ramya Prathuri, Megan Brady, and Nir Neumark

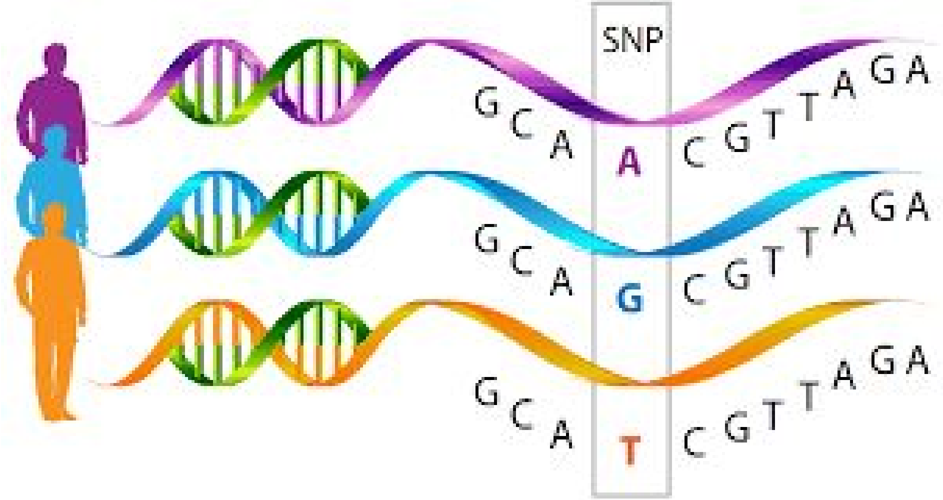
What is Variant Prioritization?

The process of identifying deleterious SNVs

- SNP = seen in > 1% population
- SNV = no limitation on frequency

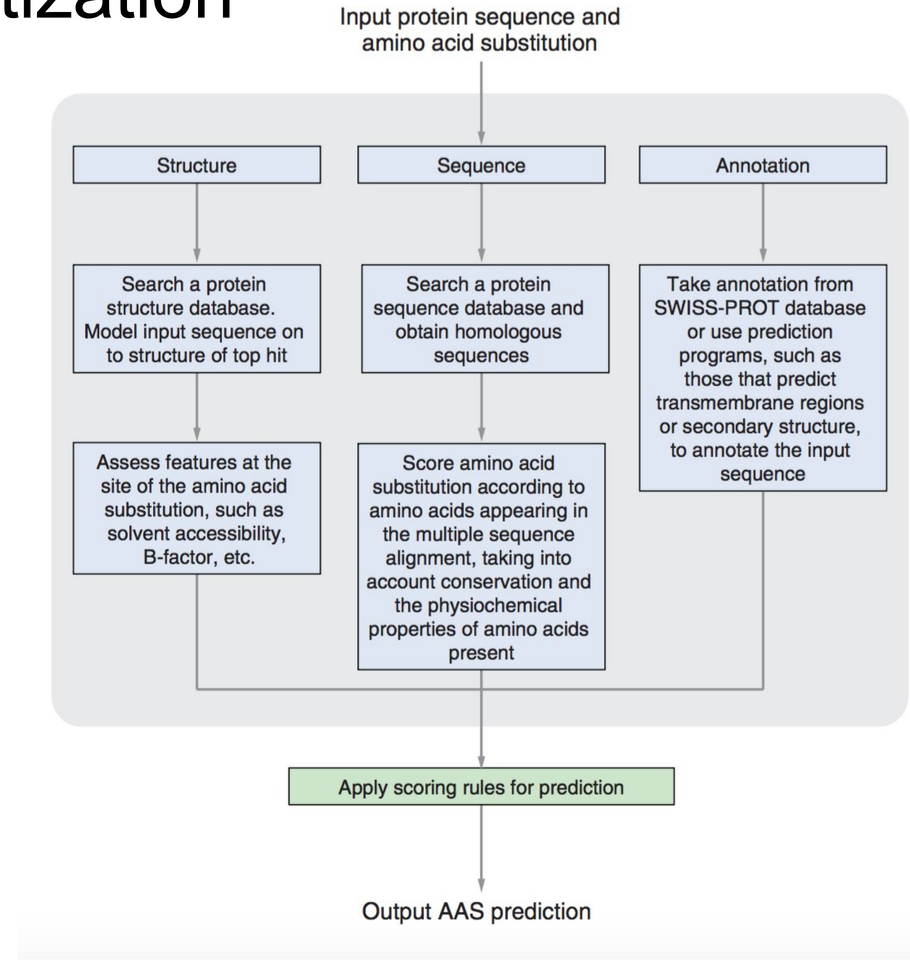
GWAS


Patients
Non-Patients
Patient DNA
Non-Patient DNA
Disease-specific SNPS
Compare differences to discover SNPs associated with disease
Non-disease-SNPS
National Human Genome Research Institute



Principles of Variant Prioritization

- Sequence:
  - synonymous or not?
  - missense or nonsense?
  - aa charge? size?
- Structure:
  - Phyre2, TMHMM
- Annotation:
  - dsSNP? GO term?



Input protein sequence and amino acid substitution

Structure

Search a protein structure database. Model input sequence on to structure of top hit

Assess features at the site of the amino acid substitution, such as solvent accessibility, B-factor, etc.

Sequence

Search a protein sequence database and obtain homologous sequences

Score amino acid substitution according to amino acids appearing in the multiple sequence alignment, taking into account conservation and the physiochemical properties of amino acids present

Annotation

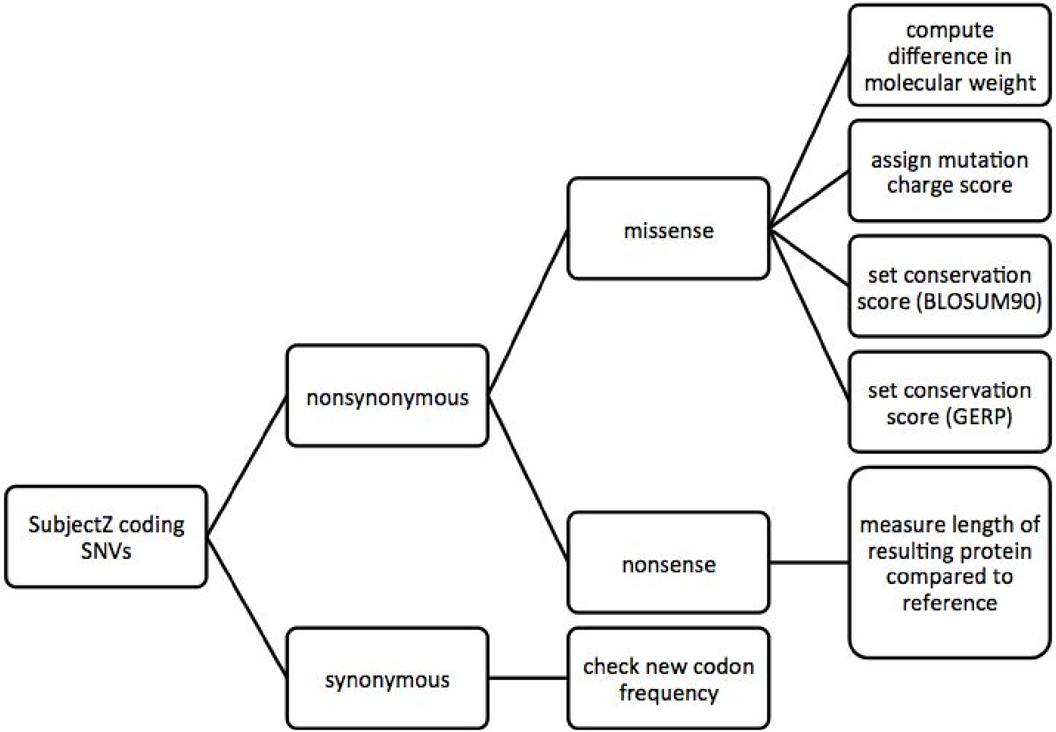
Take annotation from SWISS-PROT database or use prediction programs, such as those that predict transmembrane regions or secondary structure, to annotate the input sequence

Apply scoring rules for prediction

Output AAS prediction

Goal: How to prioritize variants?

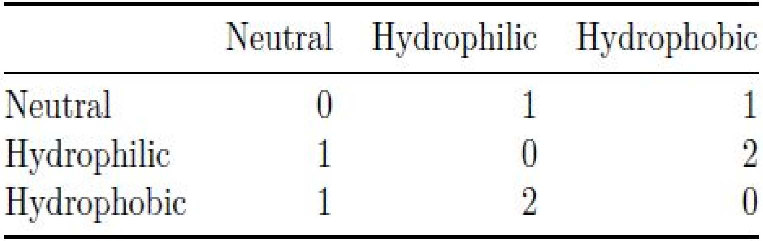
Consider a variety of parameters to assess SNVs


SubjectZ coding SNVs
nonsynonymous
synonymous
missense
nonsense
check new codon frequency
compute difference in molecular weight
assign mutation charge score
set conservation score (BLOSUM90)
set conservation score (GERP)
measure length of resulting protein compared to reference

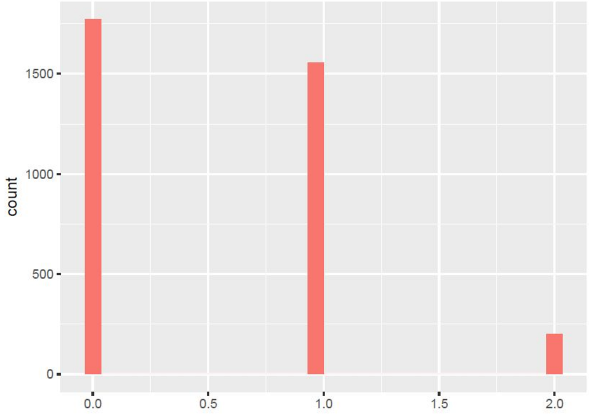
Mutation features

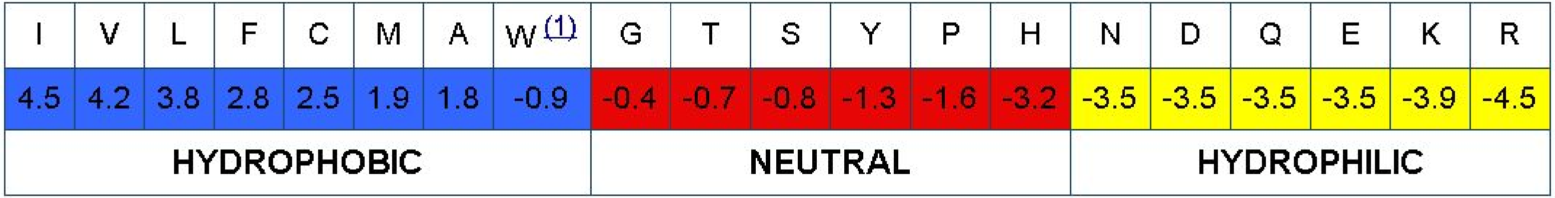
- Location
- Charge Change
- Size Difference
- Substitute
  - Blosum
  - PAM
  - GERD

Charge Change



	Neutral	Hydrophilic	Hydrophobic
Neutral	0	1	1
Hydrophilic	1	0	2
Hydrophobic	1	2	0

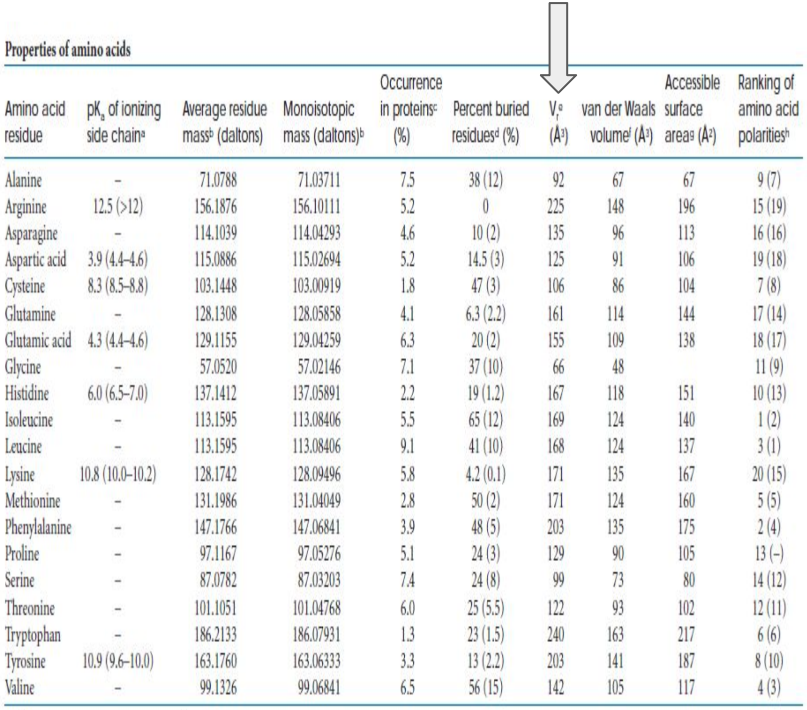




I	V	L	F	C	M	A	W [(1)]	G	T	S	Y	P	H	N	D	Q	E	K	R
4.5	4.2	3.8	2.8	2.5	1.9	1.8	-0.9	-0.4	-0.7	-0.8	-1.3	-1.6	-3.2	-3.5	-3.5	-3.5	-3.5	-3.9	-4.5
HYDROPHOBIC								NEUTRAL						HYDROPHILIC					

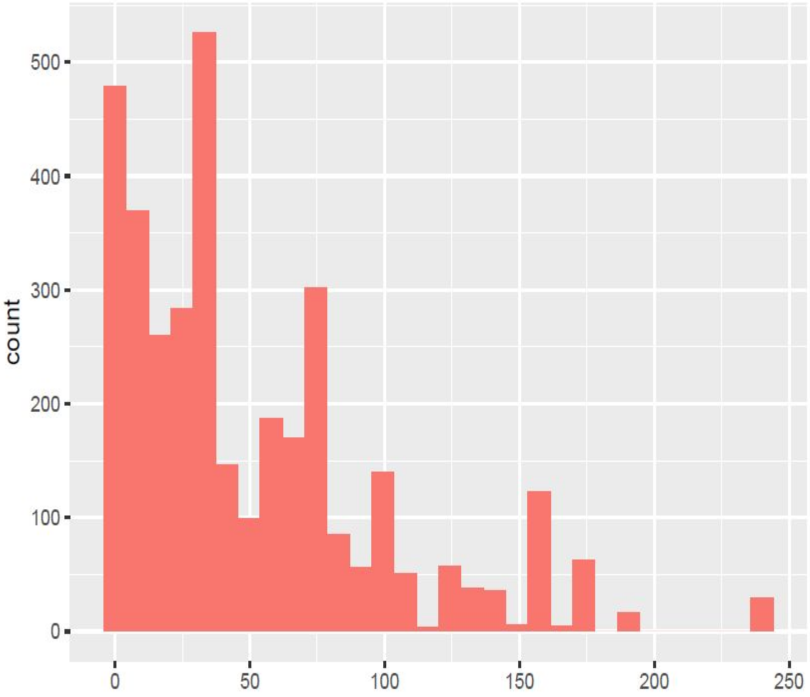
Adapted from: http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html

Size Difference



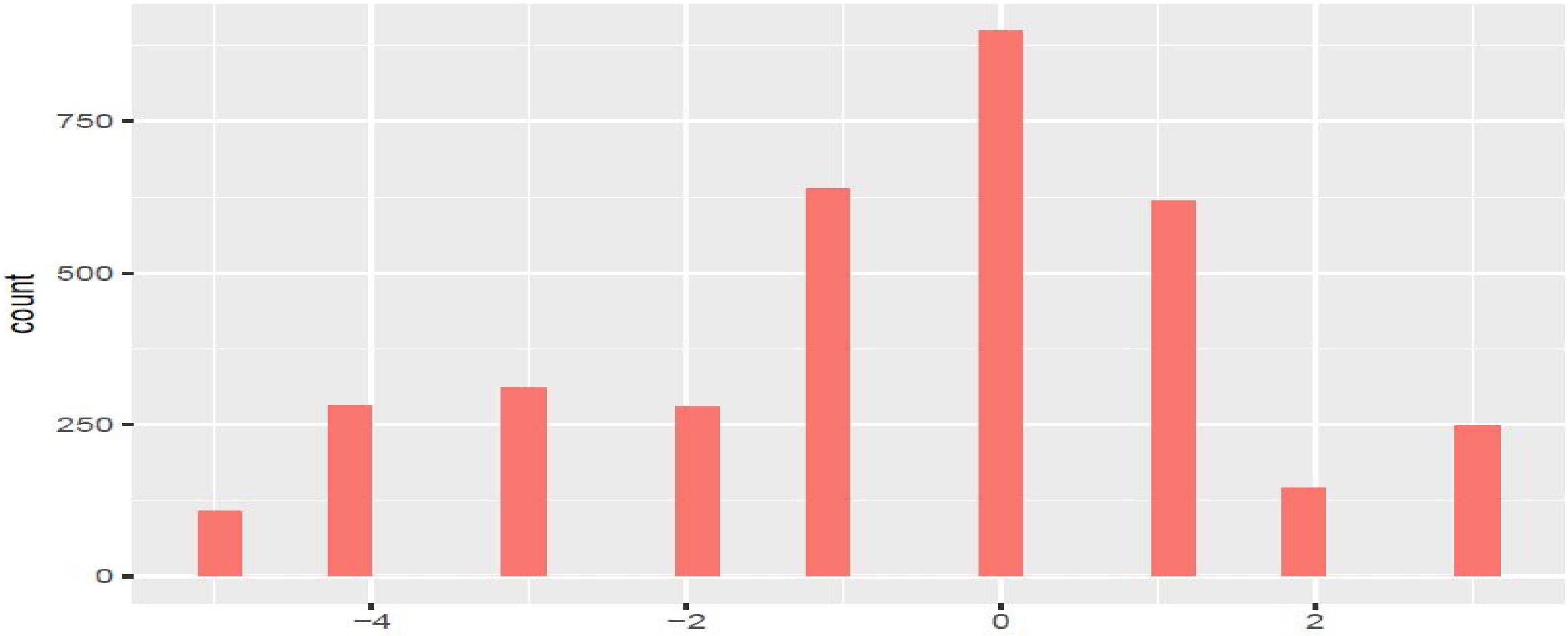
Properties of amino acids

Amino acid residue	pK_a of ionizing side chain[a]	Average residue mass[b] (daltons)	Monoisotopic mass (daltons)[b]	Occurrence in proteins[c] (%)	Percent buried residues[d] (%)	V_r[e] ($Å^3$)	van der Waals volume[f] ($Å^3$)	Accessible surface area[g] ($Å^2$)	Ranking of amino acid polarities[h]
Alanine	–	71.0788	71.03711	7.5	38 (12)	92	67	67	9 (7)
Arginine	12.5 (>12)	156.1876	156.10111	5.2	0	225	148	196	15 (19)
Asparagine	–	114.1039	114.04293	4.6	10 (2)	135	96	113	16 (16)
Aspartic acid	3.9 (4.4–4.6)	115.0886	115.02694	5.2	14.5 (3)	125	91	106	19 (18)
Cysteine	8.3 (8.5–8.8)	103.1448	103.00919	1.8	47 (3)	106	86	104	7 (8)
Glutamine	–	128.1308	128.05858	4.1	6.3 (2.2)	161	114	144	17 (14)
Glutamic acid	4.3 (4.4–4.6)	129.1155	129.04259	6.3	20 (2)	155	109	138	18 (17)
Glycine	–	57.0520	57.02146	7.1	37 (10)	66	48		11 (9)
Histidine	6.0 (6.5–7.0)	137.1412	137.05891	2.2	19 (1.2)	167	118	151	10 (13)
Isoleucine	–	113.1595	113.08406	5.5	65 (12)	169	124	140	1 (2)
Leucine	–	113.1595	113.08406	9.1	41 (10)	168	124	137	3 (1)
Lysine	10.8 (10.0–10.2)	128.1742	128.09496	5.8	4.2 (0.1)	171	135	167	20 (15)
Methionine	–	131.1986	131.04049	2.8	50 (2)	171	124	160	5 (5)
Phenylalanine	–	147.1766	147.06841	3.9	48 (5)	203	135	175	2 (4)
Proline	–	97.1167	97.05276	5.1	24 (3)	129	90	105	13 (–)
Serine	–	87.0782	87.03203	7.4	24 (8)	99	73	80	14 (12)
Threonine	–	101.1051	101.04768	6.0	25 (5.5)	122	93	102	12 (11)
Tryptophan	–	186.2133	186.07931	1.3	23 (1.5)	240	163	217	6 (6)
Tyrosine	10.9 (9.6–10.0)	163.1760	163.06333	3.3	13 (2.2)	203	141	187	8 (10)
Valine	–	99.1326	99.06841	6.5	56 (15)	142	105	117	4 (3)

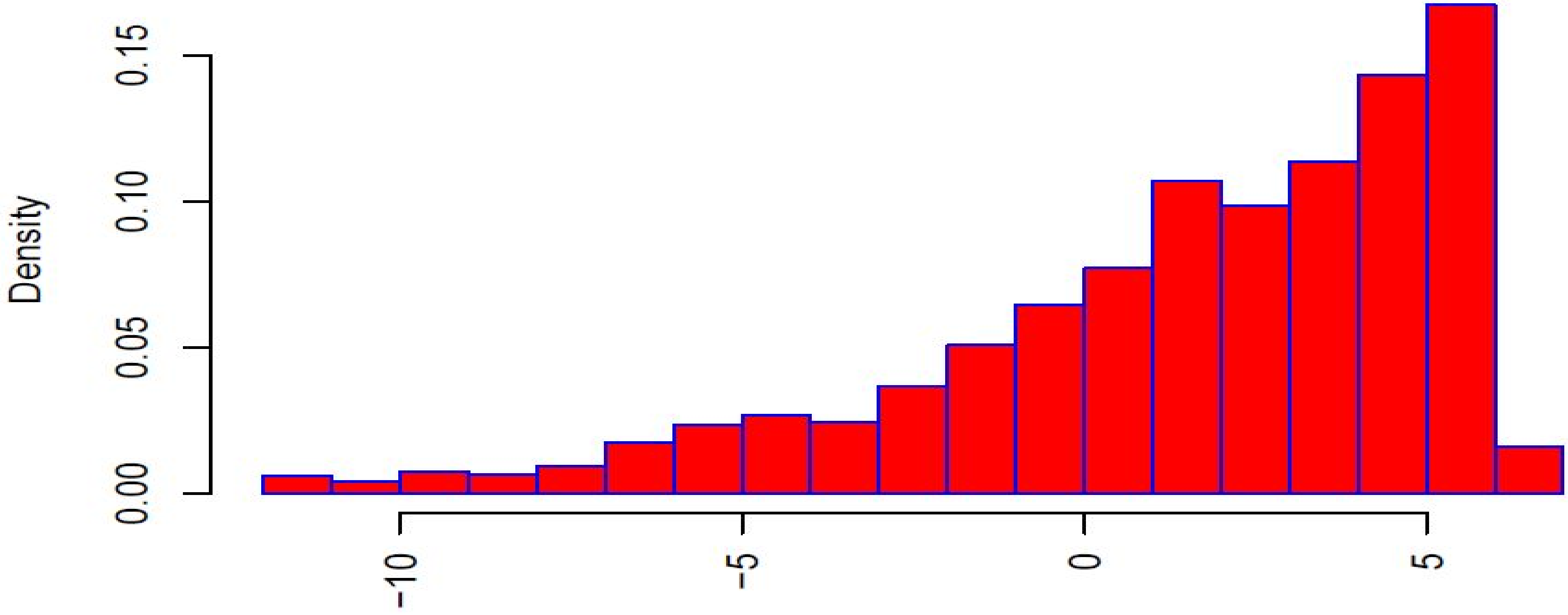


Average volume (Vr) of buried residues, calculated from the surface area of the side chain (Richards 1977; Baumann et al. 1989).

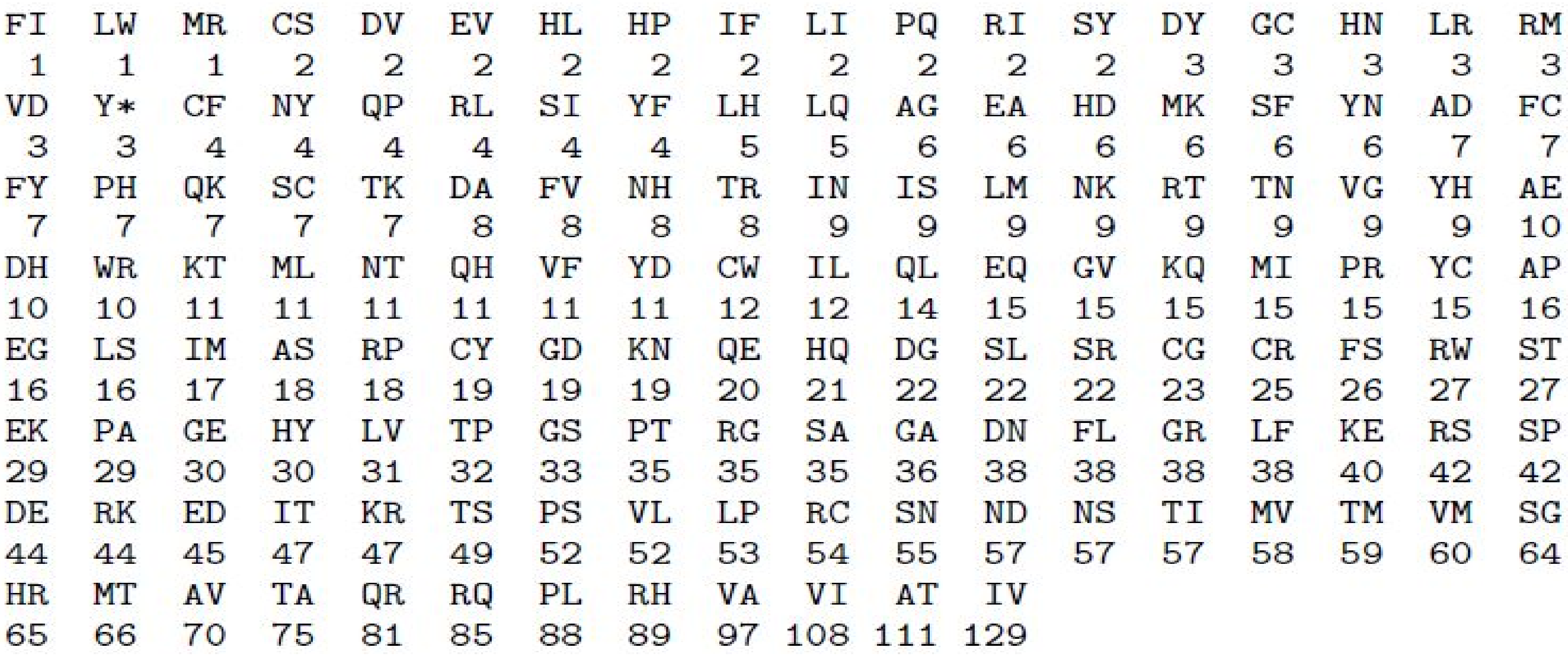
Blosum 90 Histogram



GERP Score Histogram

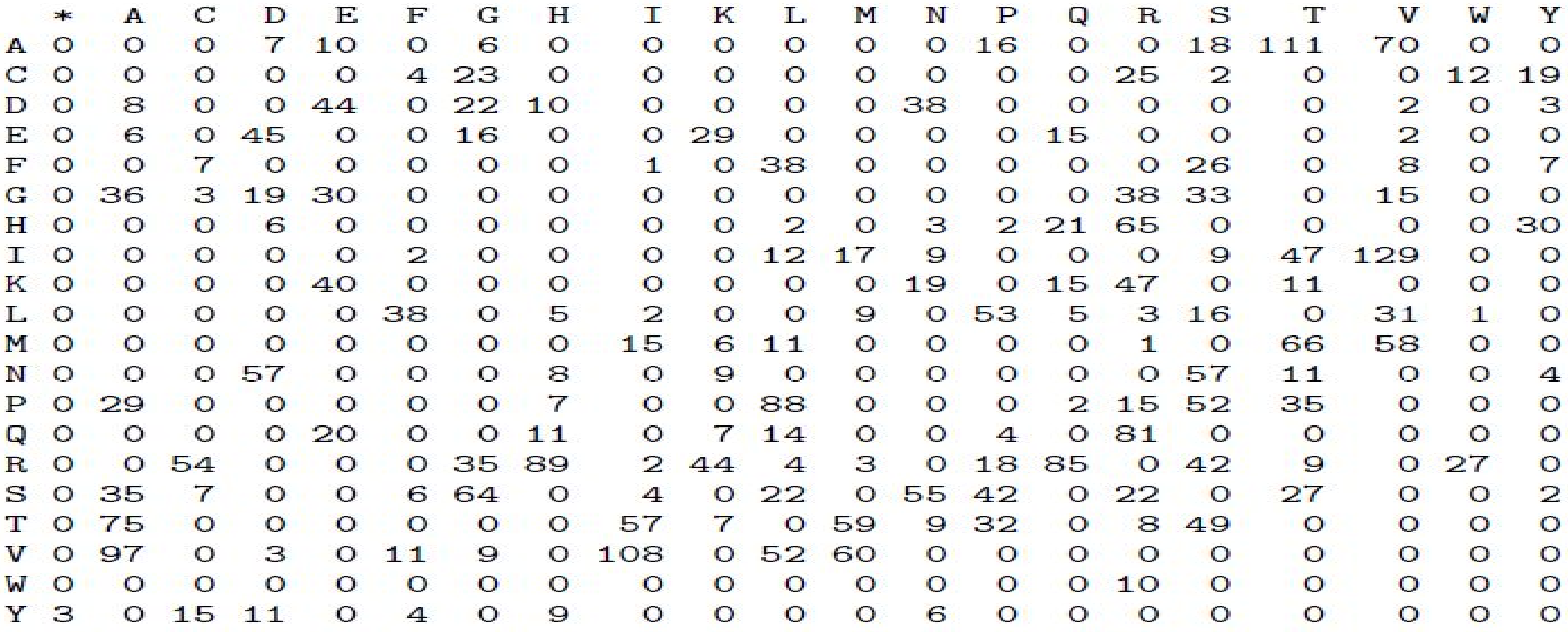


Mutation Frequency



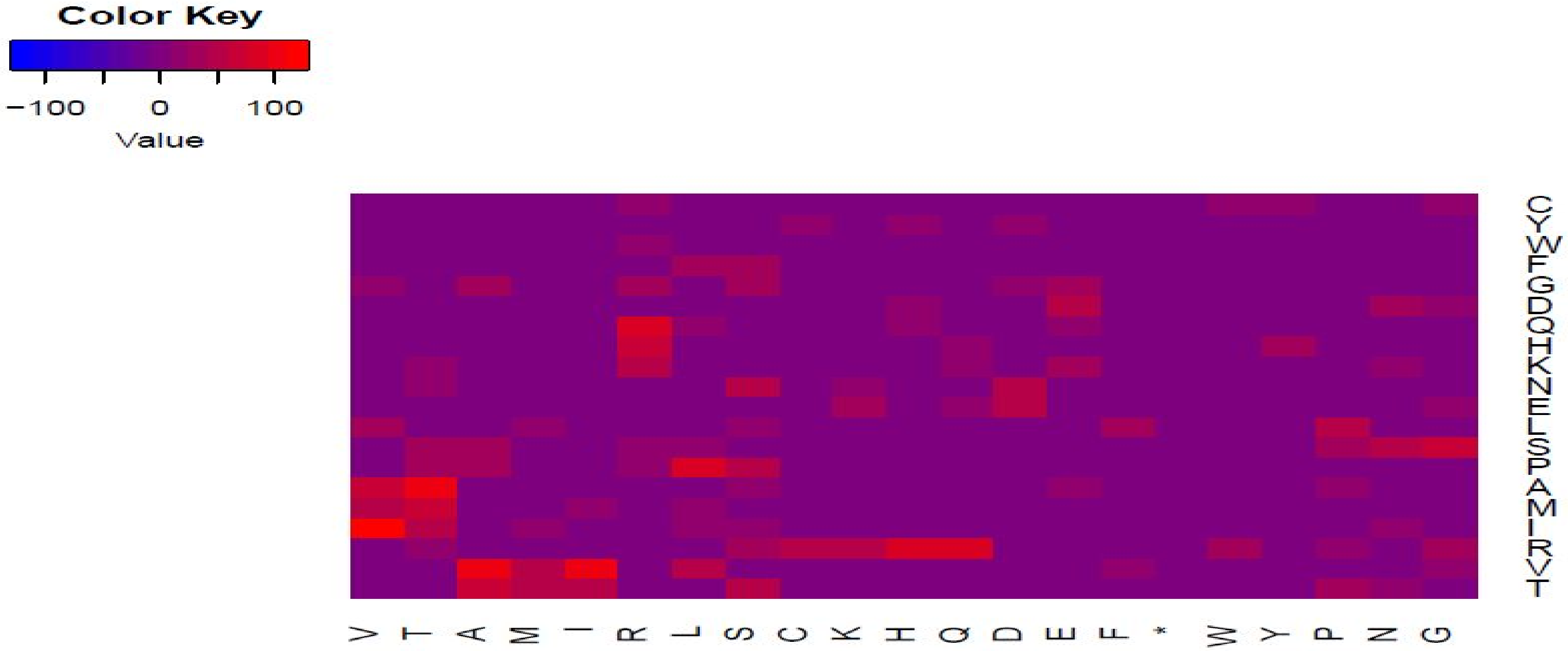
```
 FI  LW  MR  CS  DV  EV  HL  HP  IF  LI  PQ  RI  SY  DY  GC  HN  LR  RM
  1   1   1   2   2   2   2   2   2   2   2   2   2   3   3   3   3   3
 VD  Y*  CF  NY  QP  RL  SI  YF  LH  LQ  AG  EA  HD  MK  SF  YN  AD  FC
  3   3   4   4   4   4   4   4   5   5   6   6   6   6   6   6   7   7
 FY  PH  QK  SC  TK  DA  FV  NH  TR  IN  IS  LM  NK  RT  TN  VG  YH  AE
  7   7   7   7   7   8   8   8   8   9   9   9   9   9   9   9   9  10
 DH  WR  KT  ML  NT  QH  VF  YD  CW  IL  QL  EQ  GV  KQ  MI  PR  YC  AP
 10  10  11  11  11  11  11  11  12  12  14  15  15  15  15  15  15  16
 EG  LS  IM  AS  RP  CY  GD  KN  QE  HQ  DG  SL  SR  CG  CR  FS  RW  ST
 16  16  17  18  18  19  19  19  20  21  22  22  22  23  25  26  27  27
 EK  PA  GE  HY  LV  TP  GS  PT  RG  SA  GA  DN  FL  GR  LF  KE  RS  SP
 29  29  30  30  31  32  33  35  35  35  36  38  38  38  38  40  42  42
 DE  RK  ED  IT  KR  TS  PS  VL  LP  RC  SN  ND  NS  TI  MV  TM  VM  SG
 44  44  45  47  47  49  52  52  53  54  55  57  57  57  58  59  60  64
 HR  MT  AV  TA  QR  RQ  PL  RH  VA  VI  AT  IV
 65  66  70  75  81  85  88  89  97 108 111 129
```

Mutation Frequency (matrix presentation)



	*	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0	0	7	10	0	6	0	0	0	0	0	0	16	0	0	18	111	70	0	0
C	0	0	0	0	0	4	23	0	0	0	0	0	0	0	0	25	2	0	0	12	19
D	0	8	0	0	44	0	22	10	0	0	0	0	38	0	0	0	0	0	2	0	3
E	0	6	0	45	0	0	16	0	0	29	0	0	0	0	15	0	0	0	2	0	0
F	0	0	7	0	0	0	0	0	1	0	38	0	0	0	0	0	26	0	8	0	7
G	0	36	3	19	30	0	0	0	0	0	0	0	0	0	0	38	33	0	15	0	0
H	0	0	0	6	0	0	0	0	0	0	2	0	3	2	21	65	0	0	0	0	30
I	0	0	0	0	0	2	0	0	0	0	12	17	9	0	0	0	9	47	129	0	0
K	0	0	0	0	40	0	0	0	0	0	0	0	19	0	15	47	0	11	0	0	0
L	0	0	0	0	0	38	0	5	2	0	0	9	0	53	5	3	16	0	31	1	0
M	0	0	0	0	0	0	0	0	15	6	11	0	0	0	0	1	0	66	58	0	0
N	0	0	0	57	0	0	0	8	0	9	0	0	0	0	0	0	57	11	0	0	4
P	0	29	0	0	0	0	0	7	0	0	88	0	0	0	2	15	52	35	0	0	0
Q	0	0	0	0	20	0	0	11	0	7	14	0	0	4	0	81	0	0	0	0	0
R	0	0	54	0	0	0	35	89	2	44	4	3	0	18	85	0	42	9	0	27	0
S	0	35	7	0	0	6	64	0	4	0	22	0	55	42	0	22	0	27	0	0	2
T	0	75	0	0	0	0	0	0	57	7	0	59	9	32	0	8	49	0	0	0	0
V	0	97	0	3	0	11	9	0	108	0	52	60	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
Y	3	0	15	11	0	4	0	9	0	0	0	0	6	0	0	0	0	0	0	0	0

Mutation Frequency (heatmap)

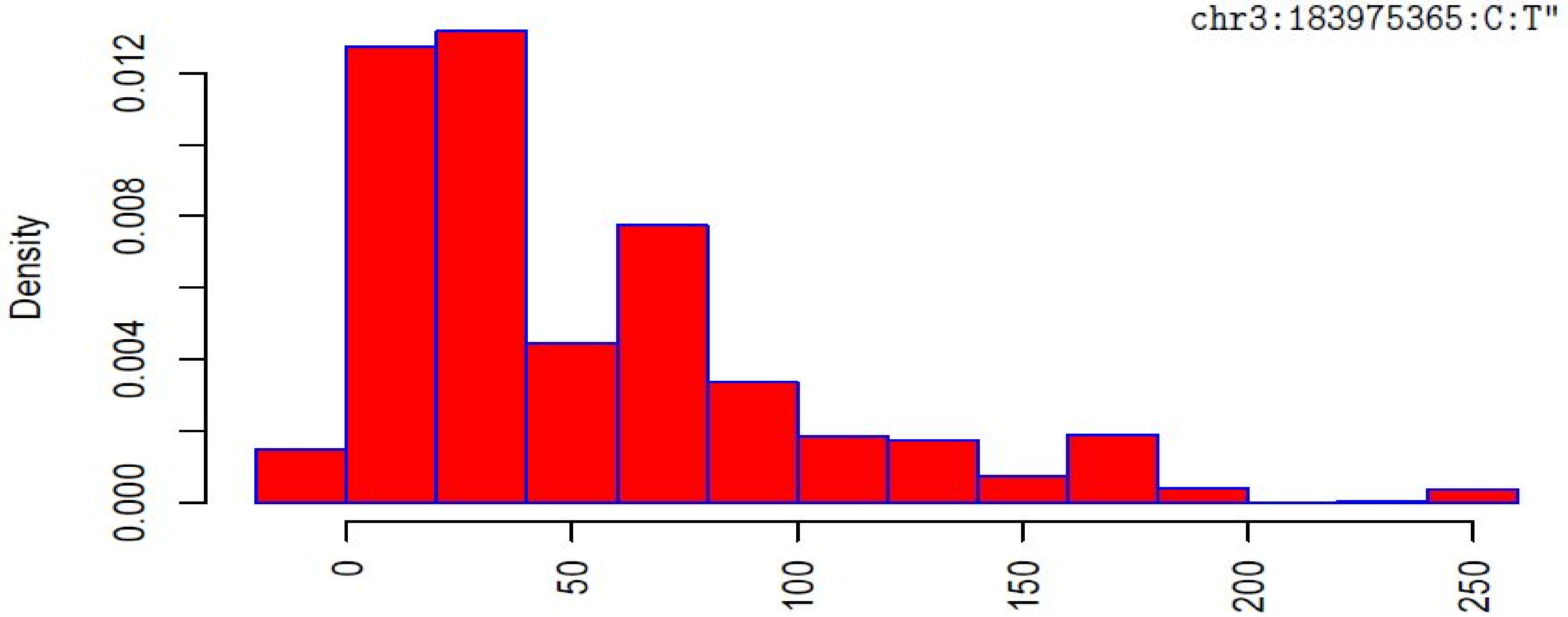


Deleterious Score (DScore)

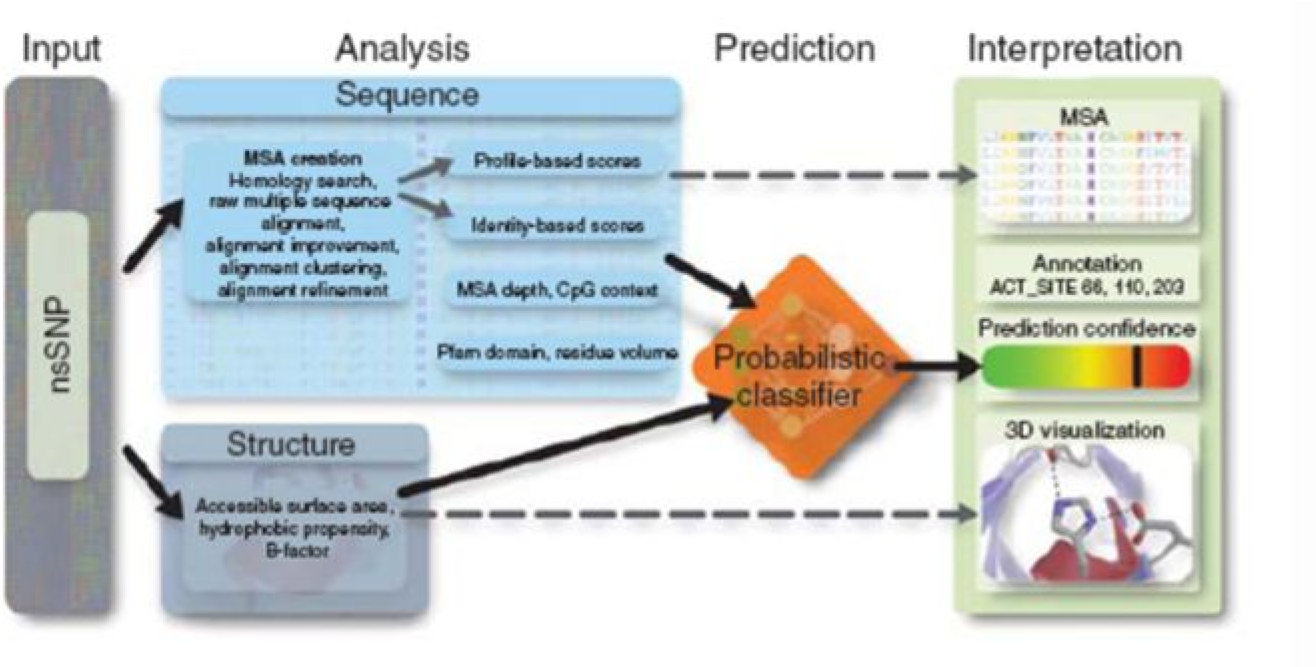
$$DScore = \alpha Charge + \beta Size + \gamma GERP + \delta Blosum$$

$$\alpha = \beta = \gamma = \delta = 1$$

Deleterious Score Histogram

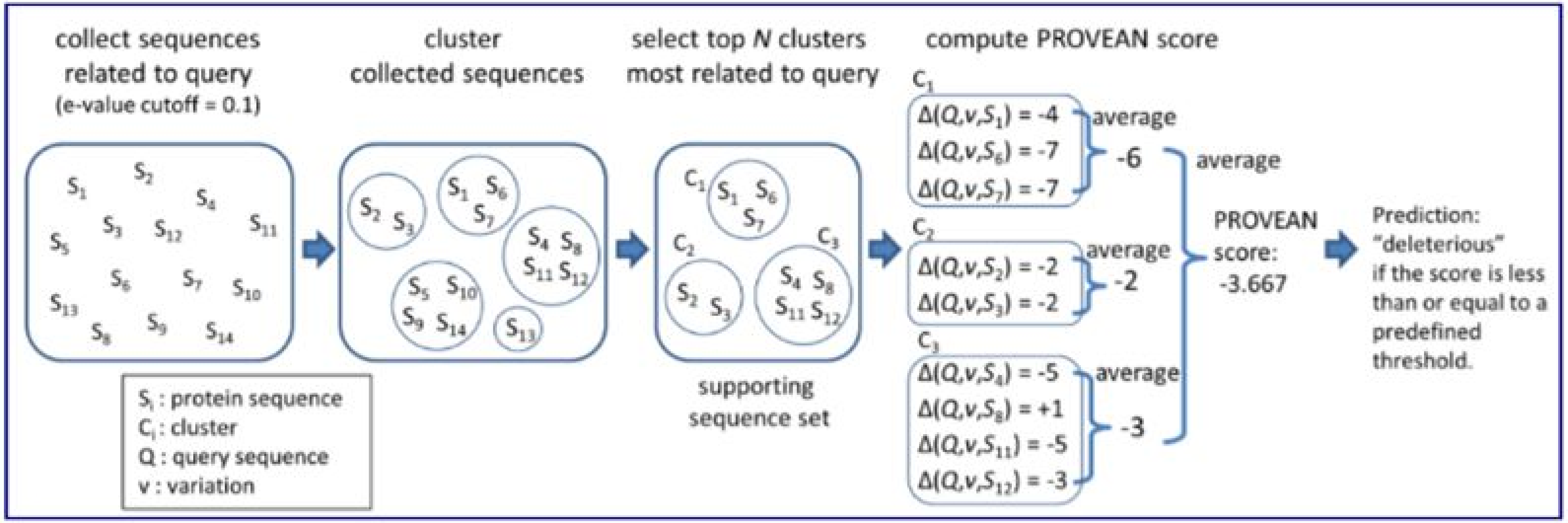


PolyPhen



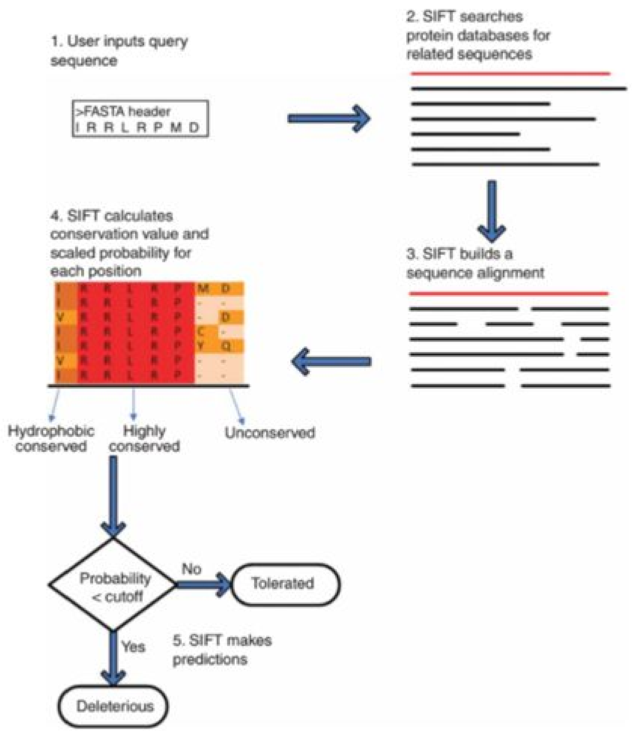
Adzhubei et al. 2010

PROVEAN



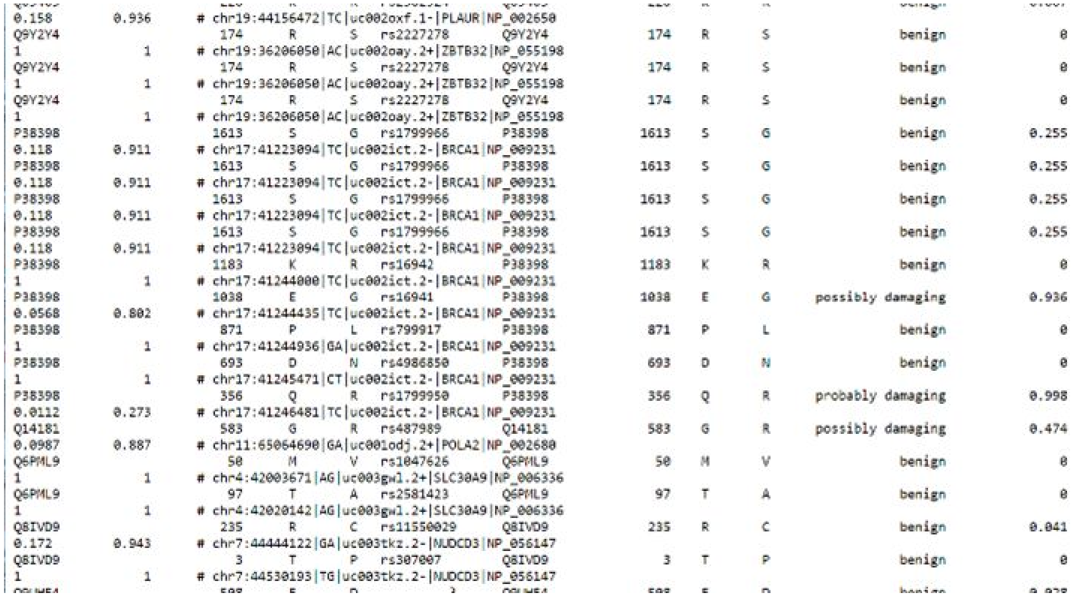
http://provean.jcvi.org/about.php

SIFT


1. User inputs query sequence
>FASTA header
I R R L R P M D
2. SIFT searches protein databases for related sequences
3. SIFT builds a sequence alignment
4. SIFT calculates conservation value and scaled probability for each position
Hydrophobic conserved
Highly conserved
Unconserved
Probability < cutoff
No
Tolerated
Yes
Deleterious
5. SIFT makes predictions

Kumar et al. 2009

PolyPhen2

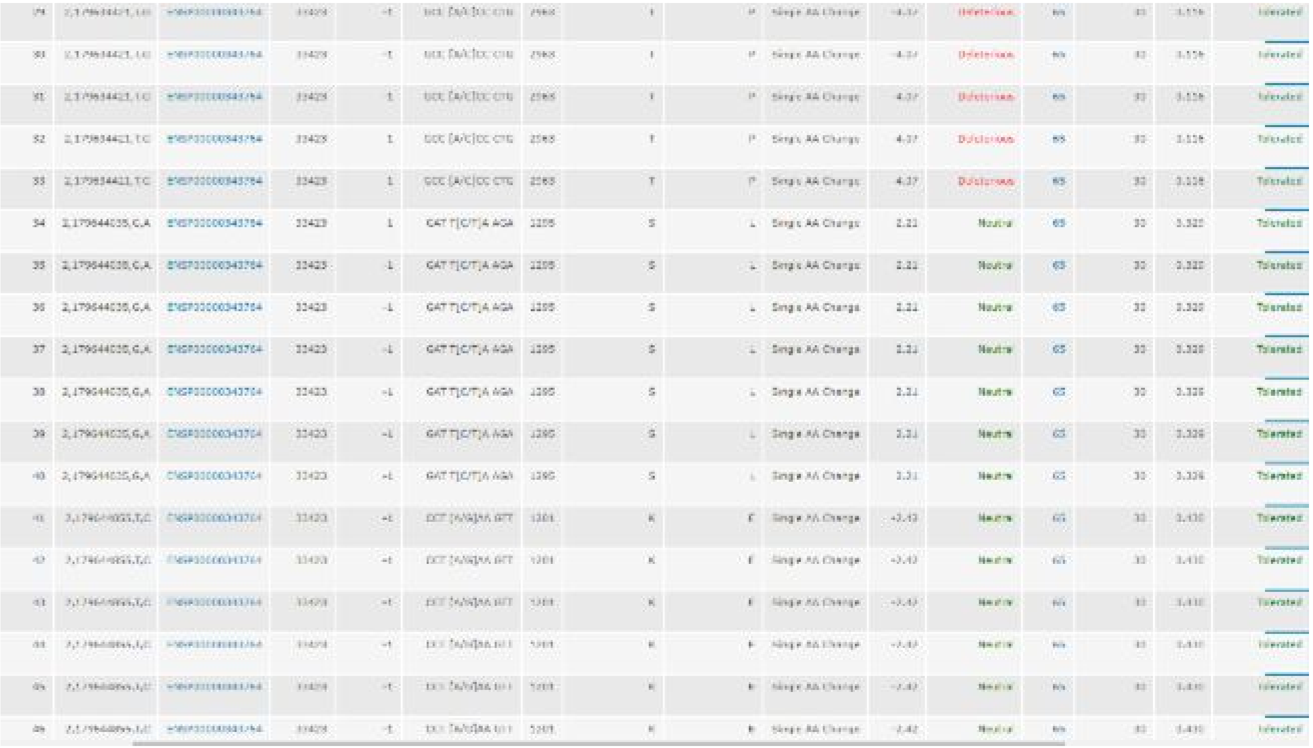


```
0.158       0.936     # chr19:44156472|TC|uc002oxf.1-|PLAUR|NP_002650
Q9Y2Y4                   174     R      S   rs2227278        Q9Y2Y4         174   R    S            benign        0
1               1     # chr19:36206050|AC|uc002oay.2+|ZBTB32|NP_055198
Q9Y2Y4                   174     R      S   rs2227278        Q9Y2Y4         174   R    S            benign        0
1               1     # chr19:36206050|AC|uc002oay.2+|ZBTB32|NP_055198
Q9Y2Y4                   174     R      S   rs2227278        Q9Y2Y4         174   R    S            benign        0
1               1     # chr19:36206050|AC|uc002oay.2+|ZBTB32|NP_055198
P38398                  1613     S      G   rs1799966        P38398        1613   S    G            benign    0.255
0.118       0.911     # chr17:41223094|TC|uc002ict.2-|BRCA1|NP_009231
P38398                  1613     S      G   rs1799966        P38398        1613   S    G            benign    0.255
0.118       0.911     # chr17:41223094|TC|uc002ict.2-|BRCA1|NP_009231
P38398                  1613     S      G   rs1799966        P38398        1613   S    G            benign    0.255
0.118       0.911     # chr17:41223094|TC|uc002ict.2-|BRCA1|NP_009231
P38398                  1613     S      G   rs1799966        P38398        1613   S    G            benign    0.255
0.118       0.911     # chr17:41223094|TC|uc002ict.2-|BRCA1|NP_009231
P38398                  1183     K      R   rs16942          P38398        1183   K    R            benign        0
1               1     # chr17:41244000|TC|uc002ict.2-|BRCA1|NP_009231
P38398                  1038     E      G   rs16941          P38398        1038   E    G  possibly damaging    0.936
0.0568      0.802     # chr17:41244435|TC|uc002ict.2-|BRCA1|NP_009231
P38398                   871     P      L   rs799917         P38398         871   P    L            benign        0
1               1     # chr17:41244936|GA|uc002ict.2-|BRCA1|NP_009231
P38398                   693     D      N   rs4986850        P38398         693   D    N            benign        0
1               1     # chr17:41245471|CT|uc002ict.2-|BRCA1|NP_009231
P38398                   356     Q      R   rs1799950        P38398         356   Q    R  probably damaging    0.998
0.0112      0.273     # chr17:41246481|TC|uc002ict.2-|BRCA1|NP_009231
Q14181                   583     G      R   rs487989         Q14181         583   G    R  possibly damaging    0.474
0.0987      0.887     # chr11:65064690|GA|uc001odj.2+|POLA2|NP_002680
Q6PML9                    50     M      V   rs1047626        Q6PML9          50   M    V            benign        0
1               1     # chr4:42003671|AG|uc003gwl.2+|SLC30A9|NP_006336
Q6PML9                    97     T      A   rs2581423        Q6PML9          97   T    A            benign        0
1               1     # chr4:42020142|AG|uc003gwl.2+|SLC30A9|NP_006336
Q8IVD9                   235     R      C   rs11550029       Q8IVD9         235   R    C            benign    0.041
0.172       0.943     # chr7:44444122|GA|uc003tkz.2-|NUDCD3|NP_056147
Q8IVD9                     3     T      P   rs307007         Q8IVD9           3   T    P            benign        0
1               1     # chr7:44530193|TG|uc003tkz.2-|NUDCD3|NP_056147
```

- Probability calculated through Bayes classifier
- Benign (0-.45), possibly damaging (.45-.95), probably damaging (.95-1)

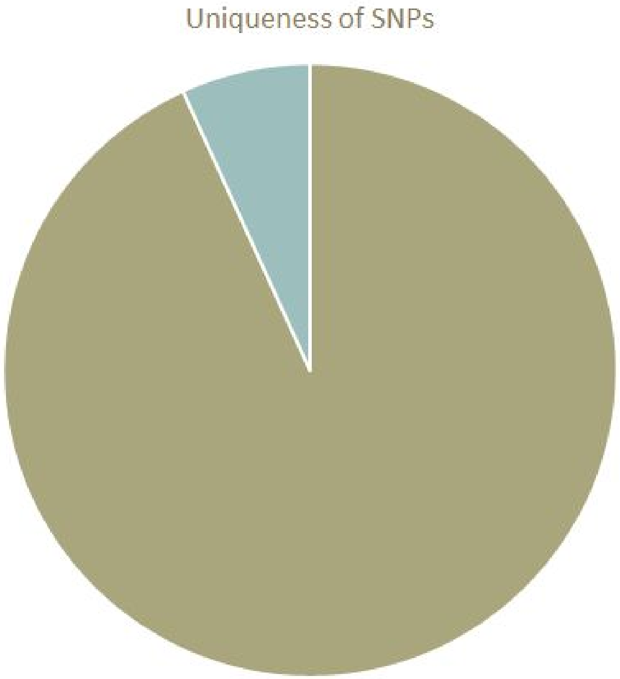
Adzhubei et al. 2015

PROVEAN & SIFT



29	2,179634423,T,C	ENSP00000343764	33423	-1	GCC [A/C]CC CTG	2583	T	P	Single AA Change	-4.37	Deleterious	65	32	3.128	Tolerated
30	2,179634423,T,C	ENSP00000343764	33423	-1	GCC [A/C]CC CTG	2583	T	P	Single AA Change	-4.37	Deleterious	65	32	3.128	Tolerated
31	2,179634423,T,C	ENSP00000343764	33423	1	GCC [A/C]CC CTG	2583	T	P	Single AA Change	4.37	Deleterious	65	32	3.128	Tolerated
32	2,179634423,T,C	ENSP00000343764	33423	1	GCC [A/C]CC CTG	2583	T	P	Single AA Change	4.37	Deleterious	65	32	3.128	Tolerated
33	2,179634423,T,C	ENSP00000343764	33423	1	GCC [A/C]CC CTG	2583	T	P	Single AA Change	4.37	Deleterious	65	32	3.128	Tolerated
34	2,179644035,C,A	ENSP00000343764	33423	1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.322	Tolerated
35	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.322	Tolerated
36	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.322	Tolerated
37	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.326	Tolerated
38	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.326	Tolerated
39	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.326	Tolerated
40	2,179644035,G,A	ENSP00000343764	33423	-1	GAT T[C/T]A AGA	1285	S	L	Single AA Change	2.21	Neutral	65	32	3.326	Tolerated
41	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.43	Neutral	65	32	3.430	Tolerated
42	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.43	Neutral	65	32	3.430	Tolerated
43	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.42	Neutral	65	32	3.430	Tolerated
44	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.42	Neutral	65	32	3.430	Tolerated
45	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.42	Neutral	65	32	3.430	Tolerated
46	2,179644055,T,C	ENSP00000343764	33423	-1	CCT [A/G]AA GTT	1281	K	E	Single AA Change	-2.42	Neutral	65	32	3.430	Tolerated

Analyzing the Zimmer Genome



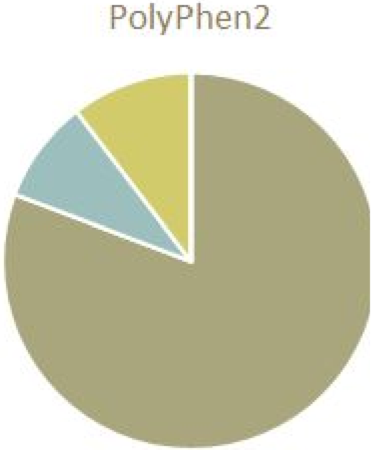
Brown = Present in dbSNP
Blue = Unique

Comparison of Various Programs

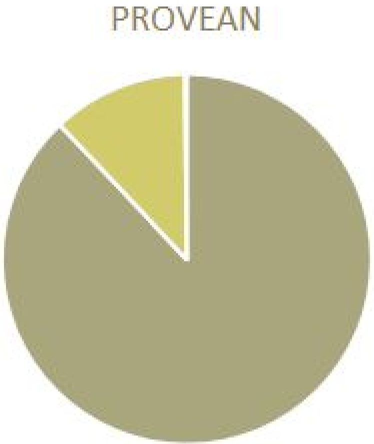
	PolyPhen2	PROVEAN	SIFT
Benign	2669	3104	2851
Possibly Damaging	293	0	0
Damaging	344	420	658
Undetermined	0	6	21

	PolyPhen2	PROVEAN	SIFT
Single Amino Acid Change	3281	3514	3514
Synonymous	15	10	10
Nonsense	10	6	6

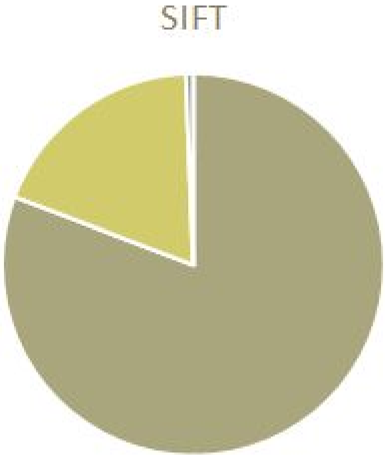
Comparison of Various Programs



PolyPhen2



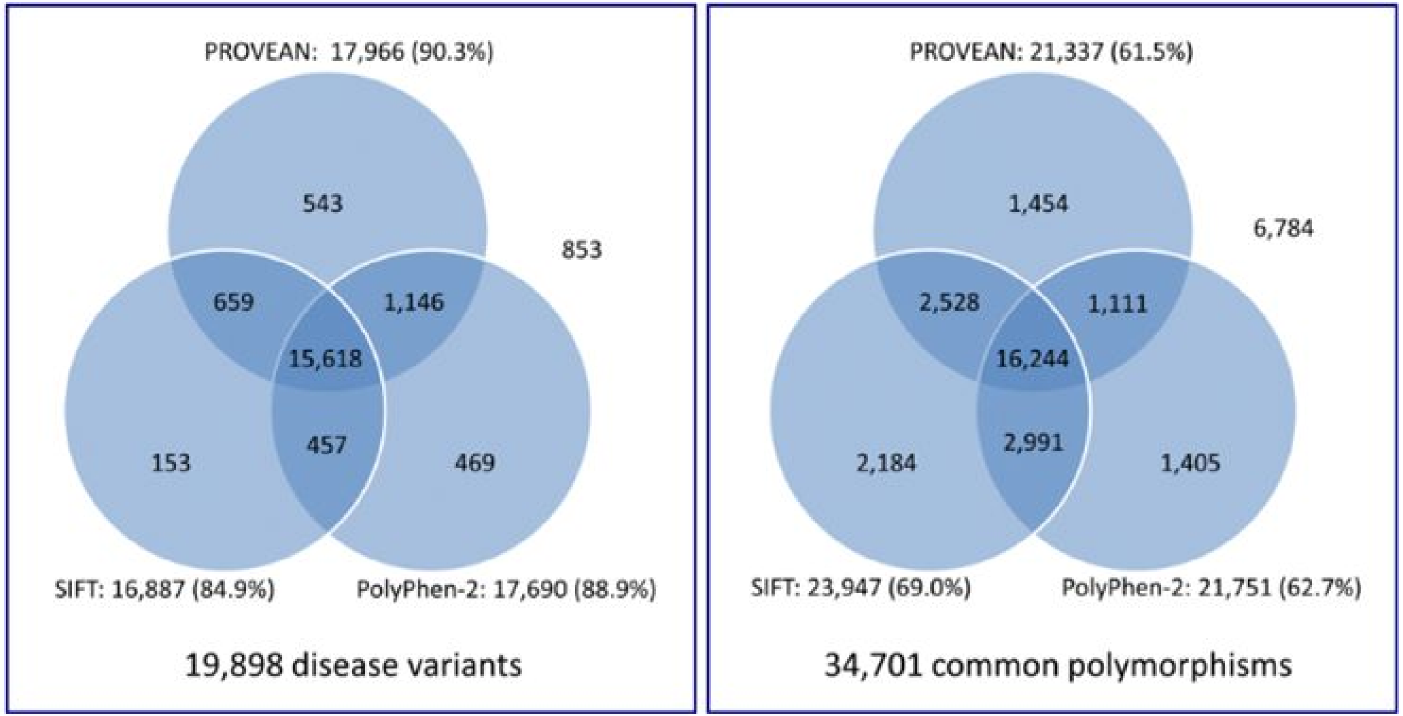
PROVEAN



SIFT

Brown = Neutral
Yellow = Damaging
Blue = Possibly damaging
Gray = Undetermined

Comparison of Various Programs



http://provean.jcvi.org/about.php