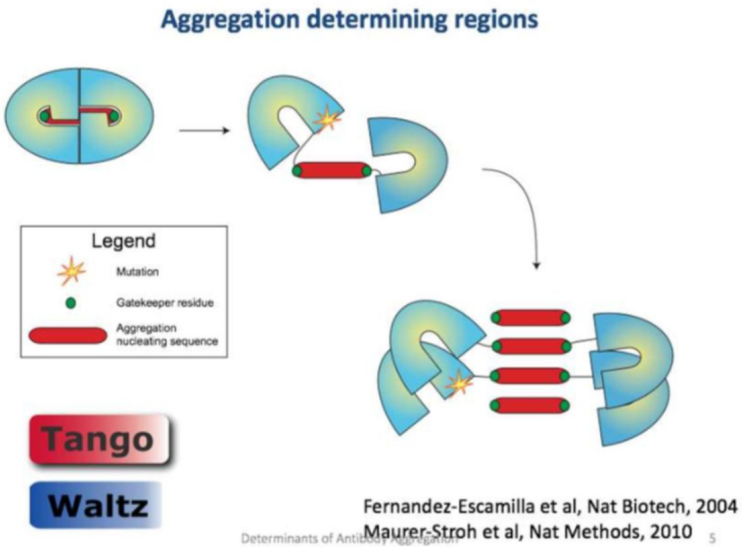
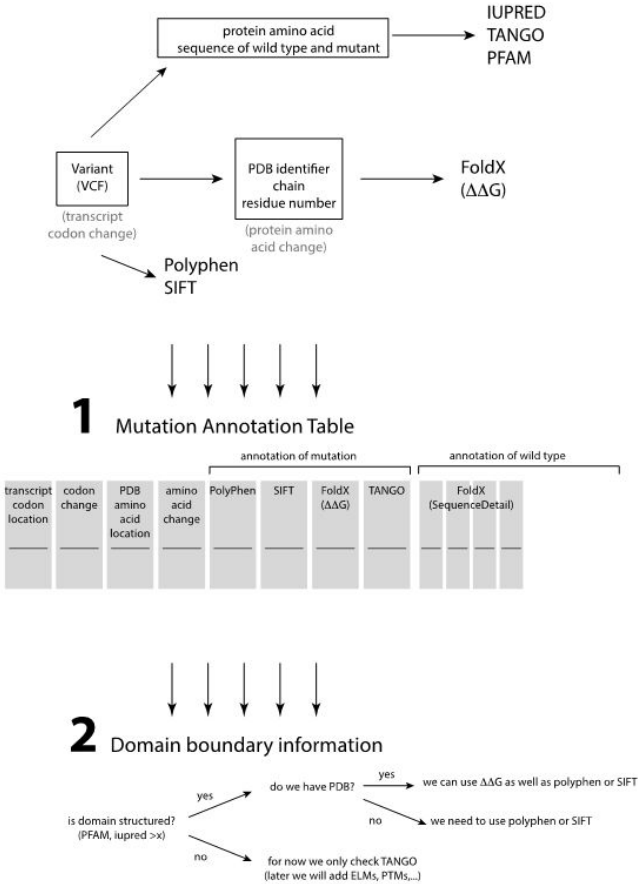
SNP - EFFECT 5

Colton Gowan
Xu Xiao
Qian Yu

Overview





Goal : Create a pipeline to annotate an entire mutation calling set so we can calculate ‘mutational load’ in terms of protein damage, a cell carries. Starting from VCF files for NGS studies.

SNPeffect 4.0

Phenotypic summary of VAR_064762

Compact overview of VAR_064762

UniProt ID	Gene	Mutation	Mutation type	Disease	OMIM	dbSNP
1433B_HUMAN	YWHAB	V99I	Unclassified	-	No OMIM entry	No dbSNP entry

Short stretch summary of VAR_064762

Predictor	Predicted regions overview	Comparison to WT	Stretches in variant	Stretches in WT
Tango		0 No change	2	2
Waltz		-1 No change	2	2
Limbo		0 No change	4	4

Domain composition of 1433B_HUMAN

Database	Domain composition	Residue details
PFAM		14-3-3 (Residues 5-238)
SMART		14_3_3 (Residues 5-244)

Structure stability summary of VAR_064762

Variant	Stability change	Effect	Structure	Stability frequency histogram
V99I	0.55 kcal/mol	Slightly reduced stability		

TANGO aggregation

WALTZ amylogenicity

LIMBO chaperone binding

single protein variants

IUPred

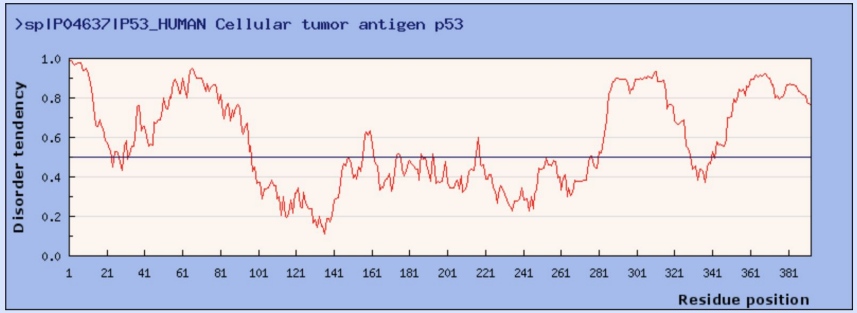
Predict - Intrinsically unstructured/disordered proteins (IUP)

Theory - estimate pairwise interaction energies in the form of quadratic expression, significant separation between the estimated pairwise energies of globular and experimentally verified IUPs. Transform into probabilistic score (0-1). Residue with score above 0.5 can be regarded as disordered.

Some well studied examples of IUPs include p21, the N-terminal domain of p53 or the transactivator domain of CREB. The importance of protein disorder is further underlined by its prevalence in various proteomes. In some eukaryotic genomes more than 20% of the coded residues are predicted as disordered.

IUPred

Prediction of Intrinsically Unstructured Proteins



>sp|P04637|P53_HUMAN Cellular tumor antigen p53

Disorder prediction score

Position	Residue	Disorder Tendency
1	M	0.9854
2	E	0.9883
3	E	0.9711
4	P	0.9677
5	Q	0.9725
6	S	0.9777
7	D	0.9777
8	P	0.9396
9	S	0.9362
10	V	0.9503

p53 - cellular tumor antigen p53

Tango

Predict cross-beta aggregation in peptides and denatured proteins.

Beta-turn, alpha-helix, beta-sheet, alpha-helical aggregation

Method: Calculates the partition function of the phase-space

Application: Tango can correctly predicts pathogenic as well as protective mutations of the Alzheimer beta-peptide, human lysozyme and transthyretin, and discriminates between beta-sheet propensity and aggregation.

Tango - output

res	aa	Beta	Turn	Helix	Aggregation	Conc-Stab_Aggregation
01	D	0.0	0.0	0.000	0.000	0.000
02	N	0.1	0.3	0.176	0.000	0.000
03	E	0.1	0.3	0.176	0.000	0.000
04	W	0.2	0.3	0.176	4.732	4.732
05	G	0.2	0.3	0.715	5.054	5.054
06	Y	1.7	0.0	0.715	9.334	9.334
07	I	3.4	0.0	0.715	9.737	9.737
08	A	3.9	0.0	0.715	9.737	9.737
09	Y	4.8	0.0	0.715	9.233	9.233
10	H	4.6	0.0	0.715	5.334	5.334
11	V	6.0	0.0	0.000	5.015	5.015
12	S	4.8	0.0	0.000	0.173	0.173
13	Q	3.2	0.0	0.000	0.000	0.000
14	D	1.3	0.0	0.000	0.000	0.000
15	P	0.0	0.0	0.000	0.000	0.000

tendency above 5% over 5-6 residues is a potential aggregating segment.

Pfam

Pfam: A large protein families database, represented by multiple sequence alignments and hidden Markov model (HMMs) [pfam]

Aim: Analyze query protein sequence to obtain domain information.

Combine with IUPred score to make more sense on identified disordered regions in Pfam.

Example(one mutation): 1A01_HUMAN

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. **More...**

Download the data used to generate the domain graphic in JSON format.

Source	Domain	Start	End	Gathering threshold (bits)		Score (bits)		E-value	
				Sequence	Domain	Sequence	Domain	Sequence	Domain
sig_p	n/a	1	24	n/a	n/a	n/a	n/a	n/a	n/a
low_complexity	n/a	8	18	n/a	n/a	n/a	n/a	n/a	n/a
Pfam	MHC_I	25	203	28.30	28.30	310.00	309.40	5.7e-90	8.5e-90
disorder	n/a	69	72	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	75	81	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	86	97	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	104	107	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	120	121	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	199	217	n/a	n/a	n/a	n/a	n/a	n/a
Pfam	C1-set	210	290	21.00	21.00	64.60	63.50	1.2e-14	2.7e-14
disorder	n/a	244	259	n/a	n/a	n/a	n/a	n/a	n/a
disorder	n/a	282	284	n/a	n/a	n/a	n/a	n/a	n/a
transmembrane	n/a	308	332	n/a	n/a	n/a	n/a	n/a	n/a
low_complexity	n/a	312	329	n/a	n/a	n/a	n/a	n/a	n/a
Pfam	MHC_I_C	337	364	20.40	20.40	60.80	59.90	1.3e-13	2.4e-13
disorder	n/a	341	348	n/a	n/a	n/a	n/a	n/a	n/a
low_complexity	n/a	345	358	n/a	n/a	n/a	n/a	n/a	n/a

Pfam:
MHC_I (25-203),
C1-set (210-290),
MHC_I_C (337-364).

Disorder:
Many regions not covered by Pfam-A are predicted to be intrinsically disordered, which doesn't mean they are lack of function.
Incorporating IUPred predictions to provide more explanation on the disordered regions. [pfam]

FoldX

FoldX: An empirical force field for the effect of mutations on stability, folding and dynamics of proteins. It calculates the free energy of a molecule based on its 3D structure. [foldx]

Aim: Predict the effect of mutation on stability based on free energy difference(DDG) between WT and MT.

Example: 1433S_HUMAN

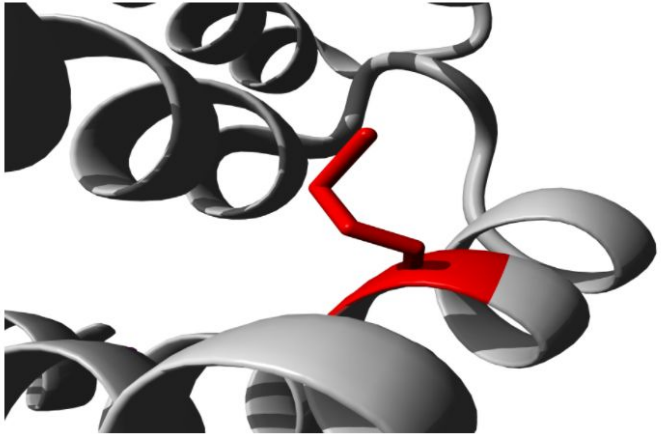
FOLDX structural profile

PDB ID	Chain	Homology	Position in PDB	Free energy change	Standard deviation	Mainchain burial	Sidechain burial
3iqu	A	100	155	0.52 kcal/mol	0.01 kcal/mol	1	0.86

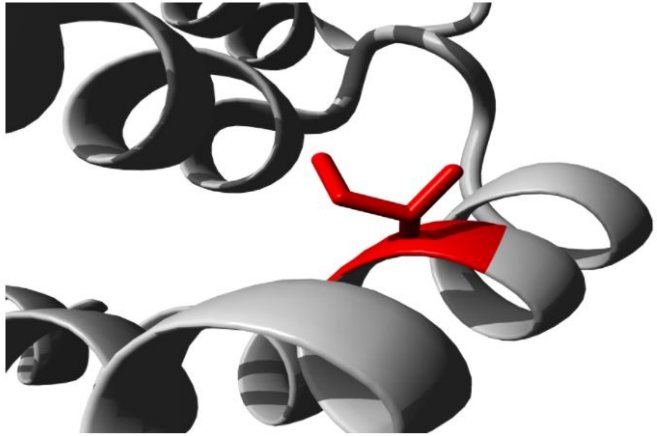
Implications on protein stability

The free energy change of this mutation is 0.52 kcal/mol. The mutation is predicted to slightly reduce protein stability.

Molecular visualization of the variant VAR_048095



Wild type 1433S_HUMAN
Download WT PDB file



Variant VAR_048095
Download Variant PDB file

These two molecular images show the structural environment of the wild type and variant amino acid.
The left image represents the wild type residue, the right represents the variant residue. The residues are colored in red and is depicted in stick representation.
Click on the images to get a larger view and to download the original.

FoldX example:

WT and MT PDB ⇒ FoldX --command=Stability --pdb=WT.pdb ⇒ ddG

```
Output File: MT_VAR_048095_3iqu_0_ST.fxout
Configuration File: config_MT_VAR_048095_3iqu_0_ST.cfg

BackHbond        =                -238.97
SideHbond        =                -65.40
Energy_VdW       =                -274.02
Electro          =                -16.33
Energy_SolvP     =                377.56
Energy_SolvH     =                -353.69
Energy_vdwclash  =                4.18
energy_torsion   =                8.31
backbone_vdwclash=                250.09
Entropy_sidec    =                149.43
Entropy_mainc    =                338.88
water bonds      =                0.00
helix dipole     =                -8.24
loop_entropy     =                0.00
cis_bond         =                0.55
disulfide        =                0.00
kn electrostatic=                 0.00
partial covalent interactions = -7.92
Energy_Ionisation =               0.22
Entropy Complex =                 0.00
-----------------------------------------------------------
Total            =                                  -85.43
```

```
1 models read: WT_VAR_048095_3iqu.pdb

BackHbond        =                -238.88
SideHbond        =                -65.40
Energy_VdW       =                -274.41
Electro          =                -16.30
Energy_SolvP     =                377.37
Energy_SolvH     =                -354.03
Energy_vdwclash  =                4.15
energy_torsion   =                8.10
backbone_vdwclash=                250.07
Entropy_sidec    =                149.82
Entropy_mainc    =                338.98
water bonds      =                0.00
helix dipole     =                -8.21
loop_entropy     =                0.00
cis_bond         =                0.55
disulfide        =                0.00
kn electrostatic=                 0.00
partial covalent interactions = -7.92
Energy_Ionisation =               0.22
Entropy Complex =                 0.00
-----------------------------------------------------------
Total            =                                  -85.95
```

E.g. ddG=0.52 kcal/mol (same with SNPeffect website result)

PolyPhen-2 and SIFT(Sort Intolerant From Tolerant)

- Two tools which calculate a score reflecting a prediction of pathogenicity of missense variants
- Non-Switch Lab
- Output : Value which suggests neutrality or loss/gain of function
- Moderate Specificity / Low Sensitivity
- May provide additional evidence for or against a mutation of interest

PolyPhen-2

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	

Submit Query | Clear | Check Status

Display advanced query options

```
#!/bin/sh
curl \
  -F _ggi_project=PPHWeb2 \
  -F _ggi_origin=query \
  -F _ggi_target_pipeline=1 \
  -F MODELNAME=HumDiv \
  -F UCSCDB=hg19 \
  -F SNPFUNC=m \
  -F NOTIFYME=myemail@myisp.com \
  -F _ggi_batch_file=@example_batch.txt \
  -D - http://genetics.bwh.harvard.edu/cgi-bin/ggi/ggi2.cgi
```

PolyPhen-2 report for Q9NYV9 N149S

Query

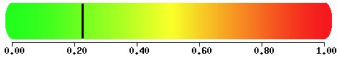
Protein Acc	Position	AA_1	AA_2	Description
Q9NYV9	149	N	S	Canonical; RecName: Full=Taste receptor type 2 member 13; Short=T2R13; AltName: Full=Taste receptor family B member 3; Short=TRB3; Length: 303

Results

Prediction/Confidence

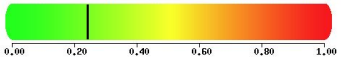
HumDiv

This mutation is predicted to be **BENIGN** with a score of **0.226** (sensitivity: **0.91**; specificity: **0.88**)



HumVar

This mutation is predicted to be **BENIGN** with a score of **0.241** (sensitivity: **0.87**; specificity: **0.75**)



Details

Multiple sequence alignment

SIFT

csh ./SIFT_for_submitting_fasta_seq.csh <seq file> <protein_database> <file of substitutions>

User Input

Enter your email address if you want the results through email :
Please check that your address is correct and your mailbox is not full.

Protein Ensembl ENSP IDs
One ENSP number (starting with ENSP) per line (Limit 1000 proteins, please!).
[Sample format]

Paste in ENSP numbers, and any substitions

-or-

Upload file containing ENSP numbers and substitutions
Choose File No file chosen

Submit Reset

SIFT: PREDICTIONS

User Input	ENSP	Pos	Ref	Subst	Prediction	SIFT Score	Median Information Content	# Seqs
ENSP00000224605,D55	ENSP00000224605	55	D	A	TOLERATED	0.22	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	C	DAMAGING	0.01	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	D	TOLERATED	1	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	E	TOLERATED	0.35	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	F	DAMAGING	0.01	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	G	TOLERATED	0.37	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	H	DAMAGING	0.03	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	I	DAMAGING	0.05	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	K	TOLERATED	0.23	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	L	TOLERATED	0.22	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	M	DAMAGING	0.03	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	N	TOLERATED	0.28	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	P	TOLERATED	0.07	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	Q	TOLERATED	0.21	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	R	TOLERATED	0.13	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	S	TOLERATED	0.38	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	T	TOLERATED	0.17	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	V	TOLERATED	0.07	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	W	DAMAGING	0	2.23	49
ENSP00000224605,D55	ENSP00000224605	55	D	Y	DAMAGING	0.01	2.23	49

Project planning and solution design

1. Softwares practices (09/10 - 16/10)

   IUPred & TANGO ⇒ Qian; Pfam & FoldX ⇒ Xu; PolyPhen & SIFT ⇒ Colton

   Understand required input format, output result and basic algorithm of tools

2. NGS data to first demo (16/10 - 30/10)
   .vcf file ⇒ protein AA sequence & pdb file
   Get results from different tools (shell script, python)
3. Pipeline formation (30/10 - 13/11)
4. Optimization (13/11 - 27/11)

   Database management (Sql? Query, Websites?)

5. Optimization (27/11 - 11/12)

   Poster, reports.

Reference

[SNPeffect]G. De Baets *et al.*, SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* **40**, D935-939 (2012).

[SNPeffect]J. Reumers *et al.*, Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucleic Acids Res* **36**, D825-829 (2008).

[IUPred]Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-3434 (2005).

[Tango]A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**, 1302-1306 (2004).

[pfam]A. Bateman *et al.*, The Pfam protein families database. *Nucleic Acids Research* **32**, D138-D141 (2004).

[foldx] J. Schymkowitz *et al.*, The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-388 (2005)

[sift] P. C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812-3814 (2003).

[polyphen] Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, *0 7*, Unit7.20. http://doi.org/10.1002/0471142905.hg0720s76