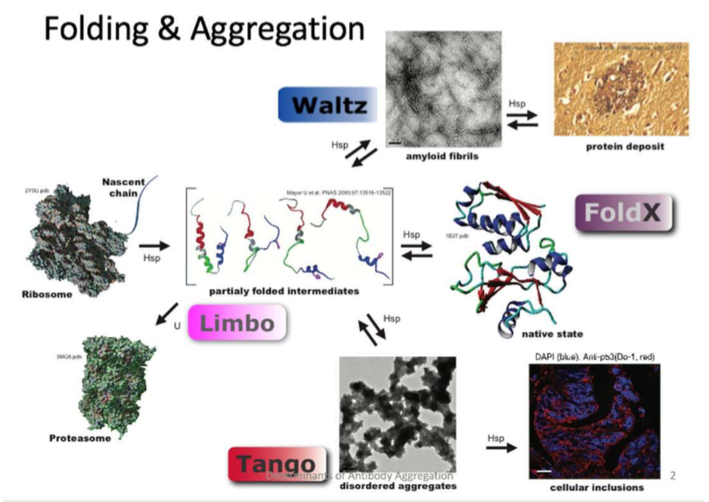
# SNP-EFFECT 4.X

Project by: Colton Gowan, Qian Yu, Xu Xiao
Supervisors: Prof. Joost Schymkowitz, Mark Fiers, Rob van der Kant
Master's of Bioinformatics, Faculty of Bioscience Engineering. Leuven, Belgium

## Introduction




Mutations in amino acid sequences have the potential to cause a range of detrimental effects ranging from truncation and loss-of-function to possibly deadly formations of protein aggregates. Many labs have created tools to annotate mutant proteins in terms of instability. In the past, these inputs were done mutation-by-mutation. Our project enables high-throughput use of these tools by beginning from a VCF file and processing into batch files for use by various algorithms.

The combination of outputs from these tools will enable researchers to simultaneously collect the available prediction information for each SNP and create a database from the results. The utilization of information from various sources will add an emergent functionality by allowing users to subset SNP's within PFAM designated regions and collect their associated algorithm outputs.

To demonstrate our novel database, we will search a cancer dataset for SNP's likely to contribute to protein aggregation as presented in the figure to the left.

## Project Pipeline

### Inputs / Outputs

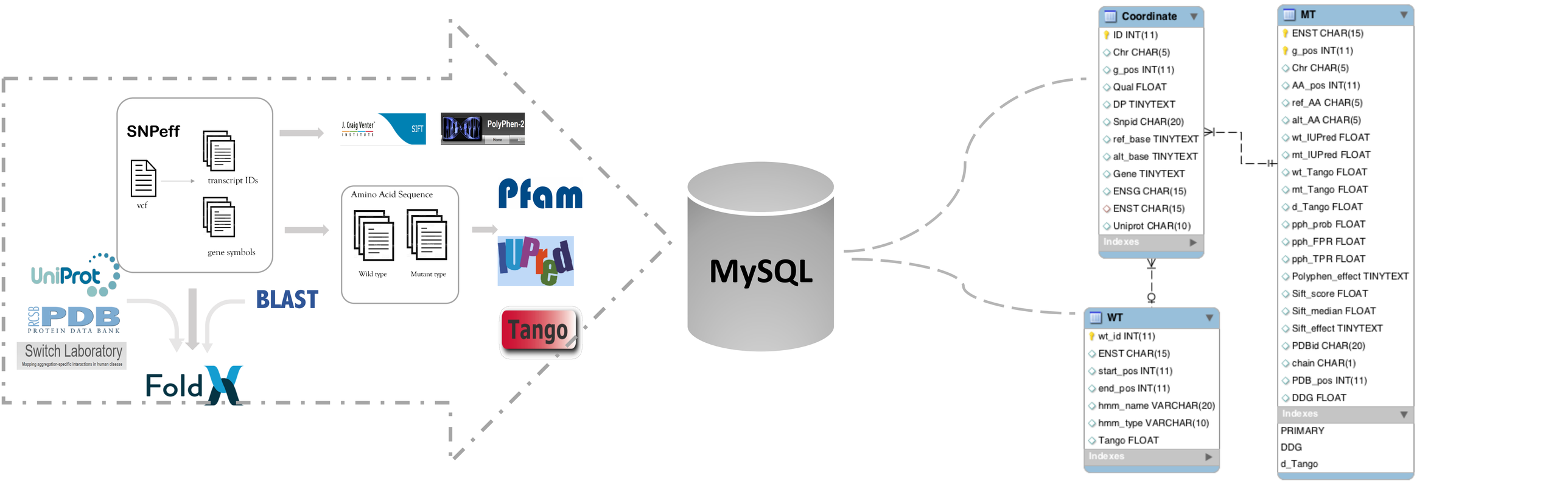
Our pipeline begins with a VCF file which contains thousands of SNP's mapped to a reference genome. The input files for the algorithms require 1 of 3 types of files: VCF, sequence information, or PDB files. Therefore the first step is to parse the VCF file in order to obtain these collections of associated files. Next, the inputs are run through the included algorithms, providing individual outputs. The outputs are finally parsed for relevant information and processed into the database structure.

The outputs of the algorithms range from a single to multiple values. For example, the SIFT output includes a SIFT score, SIFT median conservation value, and a qualitative value which summarizes the prediction as "TOLERATED" or "DELETERIOUS" [2]. Researchers using our tool can choose which of the parameters they wish to query based on their background knowledge.

### Parsing VCF

In order to obtain the proper formats for the varying inputs, the tool SNPeff was utilized to obtain Ensembl gene and transcript IDs and the position of the nucleotide change [1]. From this information we were able to obtain the wild-type and mutant amino acid sequences and positions of amino acid substitution. Initially, we attempted to take advantage of UniprotKB human ID's with PDB files and the SIFTS database. However, an issue arose over locating correct PDB files from RCSB database, due to many-to-many relationships between a protein and PDB file. Additionally, PDB files may contain only sub-regions of the protein, missing the SNP. In lieu of this we utilized a SWITCH Lab-database including 30,000 homology based models. This allowed us to collect information for FoldX to measure predicted protein atomic force fields [3].

### Database




## Results

A researcher studying cancer datasets may ask the question, "which SNP's contribute to protein stability (FoldX) and aggregation (TANGO) within protein domains (PFAM)?". Since our database includes the outputs for these algorithms, we can now query our database while considering all these options. Our goal will be to subset these high probability proteins based on these parameters.

```
select N.snpid,N.ref_AA,N.alt_AA,N.ENST,N.gene,N.uniprot,N.Tango,N.DDG,N.hmm_name
from (select N.*,C.Snpid,C.Gene,C.Uniprot
from (select W.*,E.AA_pos,E.DDG,E.g_pos,E.ref_AA,E.alt_AA
from (select AA_pos,ENST,DDG,g_pos,ref_AA,alt_AA from MT where DDG>5) as E
inner join wt as W on W.ENST=E.ENST) as N inner join coordinate as C on C.ENST= N.ENST and C.g_pos=N.g_pos) as N where N.AA_pos between N.start_pos and
N.end_pos and N.Tango > 0;
```

| Snpid | ref_AA | alt_AA | ENST | Gene | Uniprot | Tango | DDG | hmm_name |
|---|---|---|---|---|---|---|---|---|
| rs1071816 | Y | F | ENST00000412585 | HLA-B | P01889, | 263.905 | 5.33411 | MHC_I |
| NULL | V | A | ENST00000377698 | ALDH1B1 | P30837 | 2042.36 | 5.09428 | Aldedh |
| rs11890 | G | V | ENST00000219302 | NME3 | Q13232 | 903.451 | 5.68082 | NDK |

## Discussion

As more tools are introduced to accomplish similar goals, researchers may opt to use only the latest or popular tools and thereby lose information and perspective from others. On the other hand, as our project illustrates, we can create pipelines and databases to aggregate this information. This enables users to query complete available information for SNP's. Using multiple sources has the potential to create more complex queries which may be more successful in large datasets.

A secondary goal for this project was to create the database structure in a way that new information from other tools can easily be added. Extending the aggregation and data should allow for greater specificity based on new tools. This approach is heavily dependent on the quality of homology models for some tools, like FoldX, otherwise many of the tools depend purely on sequence and SNP data alone. Finally, in the future, we would like this database to be associated with expression data to better understand how mutational load in the proteome level evolves as cancer lines mature.

VCF : 'Variant Calling Format' file contains list of SNP information mapped onto a reference genome
FoldX : Software suite of tools to predict protein stability based on PDB inputs [http://foldxsuite.crg.eu/]
TANGO : Algorithm which predict protein aggregation contribution for each residue in unfolded proteins [http://tango.crg.es/]
SIFT : predicts whether an amino acid substitution affects protein function. Based on conservation [http://sift.jcvi.org/]
PolyPhen-2 : Predicts impact of an amino acid substitution on the structure via physical/comparative considerations [genetics.bwh.harvard.edu/pph2/]
PFAM : database of protein families that includes their annotations and multiple sequence alignments [http://pfam.xfam.org/]
IUPRED : Intrinsically Unstructured Predictions of protein disorder [http://iupred.enzim.hu/]
SNPeff : Open source tool that annotates variants + predicts their effects on genes using an interval forest approach. [http://snpeff.sourceforge.net/]

**GitHub**: https://github.com/yuq1993/IBP-SNPeffect

**References**:
[1] : "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672
[2] : P.C. Ng, S. Heikof, SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*.**31**, 3812-3814 (2003).
[3] : J. Schymkowitz et al., The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-388 (2005).