SNP effect 5

Introduction:

The SNPEffect platform seeks to aggregate information from several software packages which predict the effects of protein folding on experimentally derived amino-acid substitutions. Ultimately, the next iteration of SNPeffect will be able to take in large datasets and annotate these substitutions to find candidate targets and better understand the role of mutational load in cancer datasets. For example, with a time series of cancer lines, potentially identify how the mutational load changes over time, and which proteins (and the locations within them) are more likely to gain mutations earlier cancer stages.

Given a VCF file from an RNA-Seq data experimental group, (eg cancer, treatment, etc…) we wish to decipher the effects of SNP’s that are present within proteins. This procedure will be done through mapping, aggregating scores, and finally querying in a new and growing database.

Mapping is performed to link the SNP’s in the VCF file to their relevant information: genomic location, transcript ID’s, novel amino acid substitution/sequence, and ultimately the PDB structure files containing this SNP. This is performed through the utilization of several databases including UniprotKB, RCSB, SNPeff, SIFTS, and more.

Scoring software include SIFT, Polyphen, and FoldX. These may require the Wild Type amino acid sequence or the PDB files with the associated versions with the mutant amino acid. These scores are calculated through various algorithms and give us the power to discern between innocuous and destructive amino acid substitutions for normal protein behavior.

Querying from the final output of this platform will enable researchers to ask complex questions about the information aggregated into the database created from their VCF file. These queries may include specific parameters for scores from specific software and be as specific as choosing mutations within specific protein family domains.

Ultimately, the goal is to create an ever-expanding database which will be continuously updated as users upload new data so the number of potential queries can become more exact and have more support by public data.

Methods

All of the following was performed with the Human Genome 19 since our VCF cancer dataset was created with this version. The scripts could be updated to accommodate other versions.

We used the UniProtKB database to find which proteins already had PDB files. Conveniently, a webpage on UniProt contained a complete list of UniprotKB ID’s with the PDB Id’s that contain these proteins. This webpage was parsed for human proteins. The output was a list of many-to-many association between a UniprotKB ID and PDB ID. This is due to the fact that a single protein could have many PDB files and a single PDB file may be a small region of interest which is shared between many proteins.

To find which proteins were affected by each SNP, we parsed the VCF after it had been processed by SNPeff. SNPeff added an additional column of information to the standard VCF which included key information: The ENSG and the associated ENST. With this information, we were able to cross-reference the ENST to the human genome annotation file to find out the exact genomic location of the transcripts. The human genome annotation file was parsed to find the lines with CDS which contained the beginning and end of each CDS region as well as the ENST Id’s in which they were located. Each ENST was parsed providing the mutated amino acid sequence of each transcript.