

Protein Predictions: From Function to Impact of Genetic Variation

Sarah Pendergrass
Center for Systems Genomics

Outline

- What are homology and non-homology based tools for characterizing new protein sequences?
- What are methods for identifying if my SNPs will have an impact on protein function?
 - Why would I want to use them?

Homology Based Tools

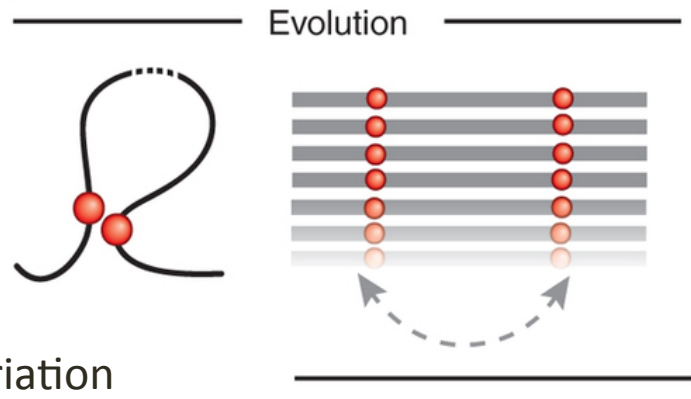
- Identify what a protein will do from a potential protein sequence
- Sequence similarity commonly done with software like BLAST
 - Aligning a pair of sequences
- P-fam can be used
 - Database of conserved protein domain families to annotate and classify proteins
- Multiple sequence alignments and protein 3-D structures can be combined for analyses

Non-Homology Based Tools

- We know many more protein sequences than protein three-dimensional structures
 - Sequencing data identifying coding regions at a fast rate
 - 40% of known human genes don't have functional classification by sequence similarity
- CHALLENGE
 - The homology based approach is not always possible
- Many proteins contain enough information in their amino acid sequence to determine their three-dimensional structure
 - Non-homology based tools

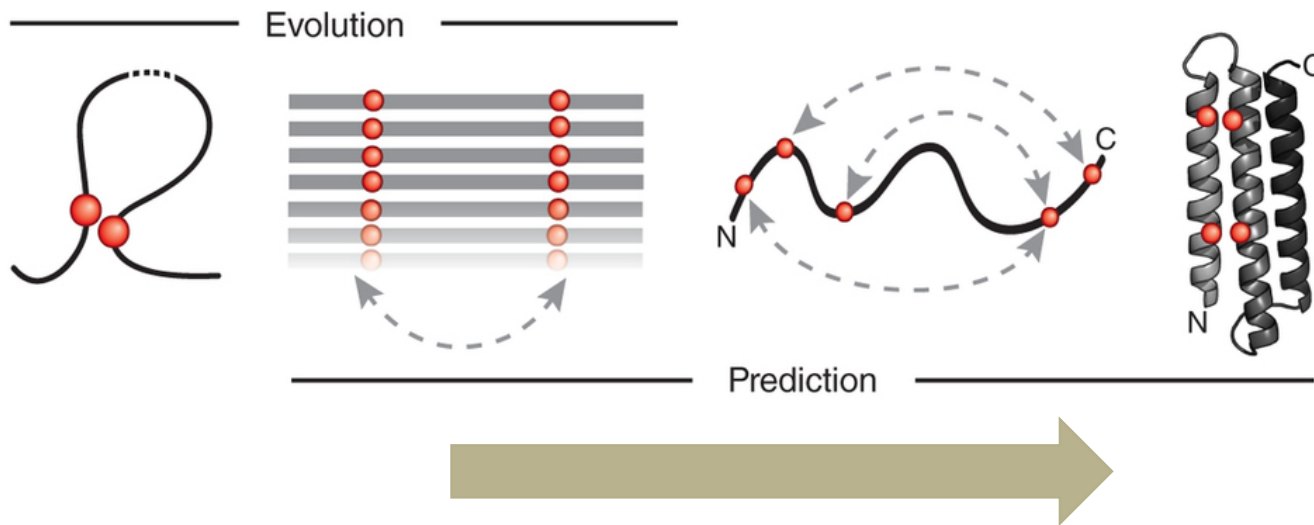
Non-Homology Based Tools

- Evolutionary information found in patterns of correlated protein sequences can be used
 - Evolutionary pressure to maintain favorable interactions between interacting amino acid residues
 - Residue co-variation across related protein sequences
- Evolutionary trajectory of a protein through sequence space is constrained by its function
- Covarying residues be predictive of functional sites, protein interactions, and alternative conformations



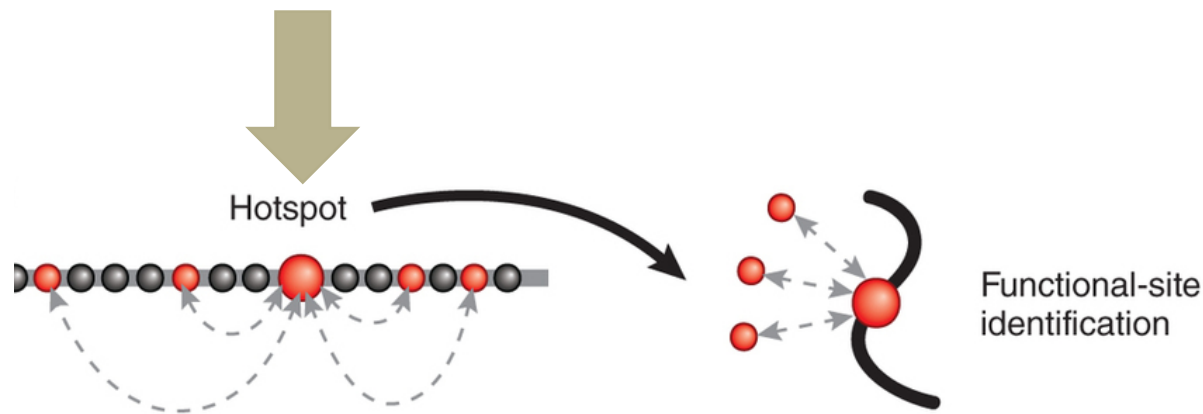
Non-Homology Based Tools

- Once co-evolutionary couplings are identified they can be used to predict unknown three-dimensional structure of a protein



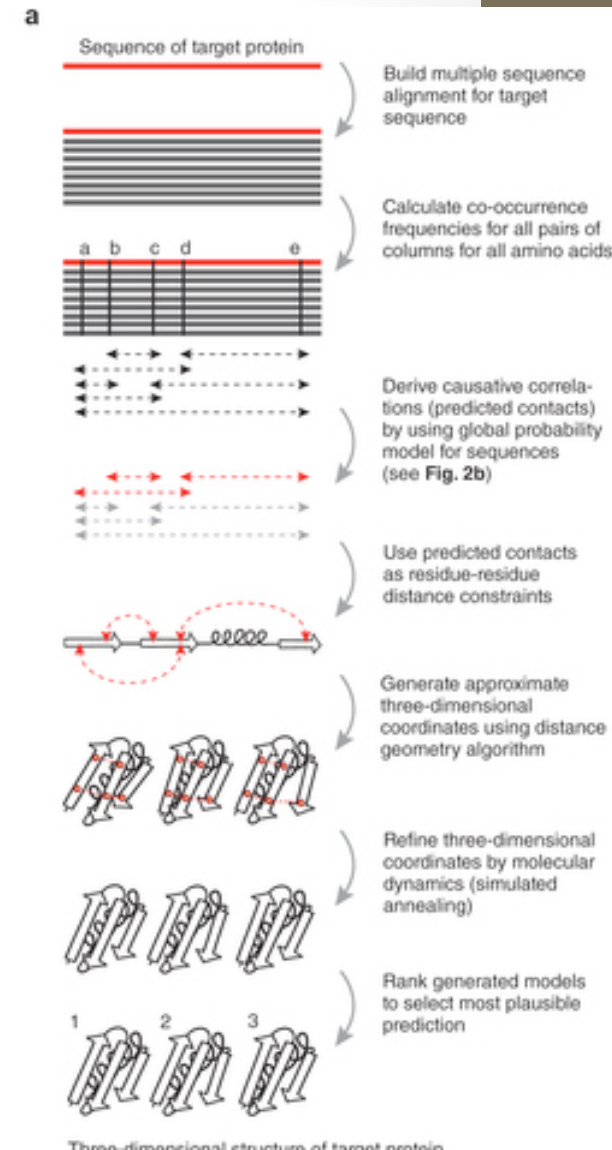
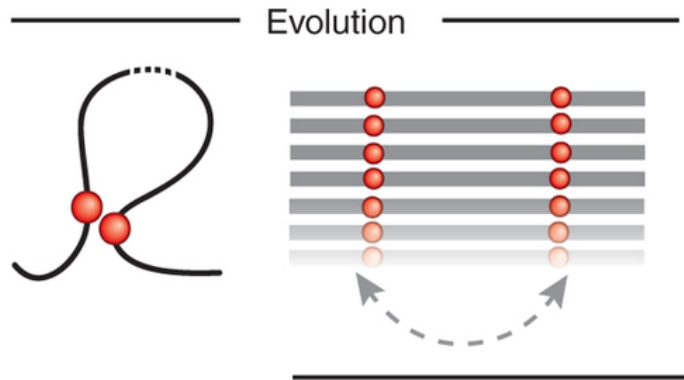
Non-Homology Based Tools

- Residues subject to a high number of evolutionary pair constraints are likely functional hotspots
 - Such as interaction with external ligands
- These hotspots may not be detectable by just analyzing single-residue conservation



Non-Homology Based Tools

- Evolutionary Couplings
- Evfold.org
- Amino acid sequence of the target protein is used to perform a database search for sequence similarity
- Then correlated residue co-variation
- Then protein folding prediction



Non-Homology Based Tools

- Evolutionary Couplings
- Evfold.org
- Hundreds of sequences are needed to derive plausible causative evolutionary couplings
 - Primary limitation

EVcouplings Which residues are the most evolutionarily constrained?

Calculate ECs between residues, explore these for functional relevance and map them onto known structures.

EVfold Predict Unknown 3D Structure for individual protein domains

Protein structure prediction from sequence variation
Nature Biotechnology 30, 1072–1080 (2012)

Non-Homology Based Tools

- Evolutionary Couplings
- EvFold.org
- Predicted co-evolved contacts between 50 E. coli complexes of unknown structure
 - Also had results consistent with detailed experimental data

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A Hopf, Charlotta P.I Schärfe, João P.G.L.M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M.J.J. Bonvin, Debora S Marks ✉

DOI: <http://dx.doi.org/10.7554/eLife.03430>

Published September 25, 2014

Cite as eLife 2014;10.7554/eLife.03430

Download PDF

Non-Homology Based Tools

- Evolutionary Couplings
- EvFold.org
- Also used this to predict previously unknown 3D structures for 11 transmembrane proteins

Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

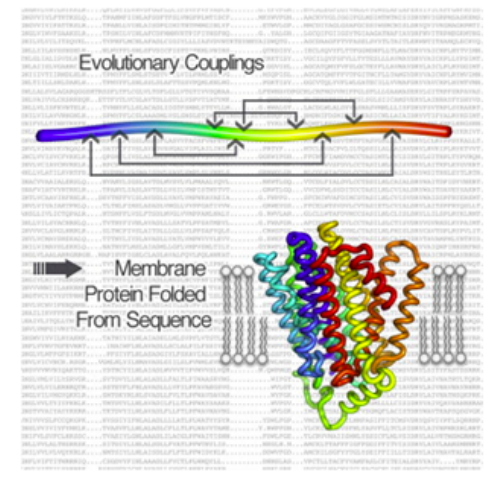
Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, Debora S. Marks

Cell, Volume 149, Issue 7, 1607-1621, 10 May 2012

[PubMed PMID: 22579045] >> [View on Journal Website](#)

Abstract

We show that amino acid covariation in proteins, extracted from the evolutionary sequence record, can be used to fold transmembrane proteins. We use this technique to predict previously unknown 3D structures for 11 transmembrane proteins (with up to 14 helices) from their sequences alone. The prediction method (EVfold_membrane) applies a maximum entropy approach to infer evolutionary covariation in pairs of sequence positions within a protein family and then generates all-atom models with the derived pairwise distance constraints. We benchmark the approach with blinded de novo computation of known transmembrane protein structures from 23 families, demonstrating unprecedented accuracy of the method for large transmembrane proteins. We show how the method can predict oligomerization, functional sites, and conformational changes in transmembrane proteins. With the rapid rise in large-scale sequencing, more accurate and more comprehensive information on evolutionary constraints can be decoded from genetic variation, greatly expanding the repertoire of transmembrane proteins amenable to modeling by this method.



Non-Homology Based Tools

- Maybe also try MAKER
 - <http://www.yandell-lab.org/software/maker.html>
- Purpose is to allow smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases
- Identifies repeats, aligns contigs and proteins to a genome, produces gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values

Why Important

- So protein predicting methods are important for identifying more information about protein coding regions
- Leveraging this information for understanding more about the effect of individual SNPs on proteins
 - Databases and tools that combine protein information with the potential impact of genetic variation on proteins

Novel Missense Variants

- I have SNPs or SNVS
 - Perhaps sequencing data or whole-exome, or genotype array data
- Non-synonymous protein coding SNVs may have the greatest impact on phenotypic variation
 - Excess of rare alleles among those predicted to be functional for proteins
 - *But don't forget all the other ways genetic variation can affect gene transcription*
- So for my SNPs or SNVs
 - How can I identify if they will have an impact on a protein?

Novel Missense Variants

- Protein information can be used to prioritize study of specific SNPs/SNVs
- Can help identify key SNPs or SNVs for further study after identifying them in association testing
 - How to sort through hundreds of results
- Identification of genes that can then be migrated to gene based testing
- Easier translation to molecular/cellular assay

PolyPhen

- Polymorphism Phenotyping (PolyPhen)
- Predicts possible impact of an amino acid substitution on the structure and function of a human protein
- <http://genetics.bwh.harvard.edu/pph2/>

PolyPhen

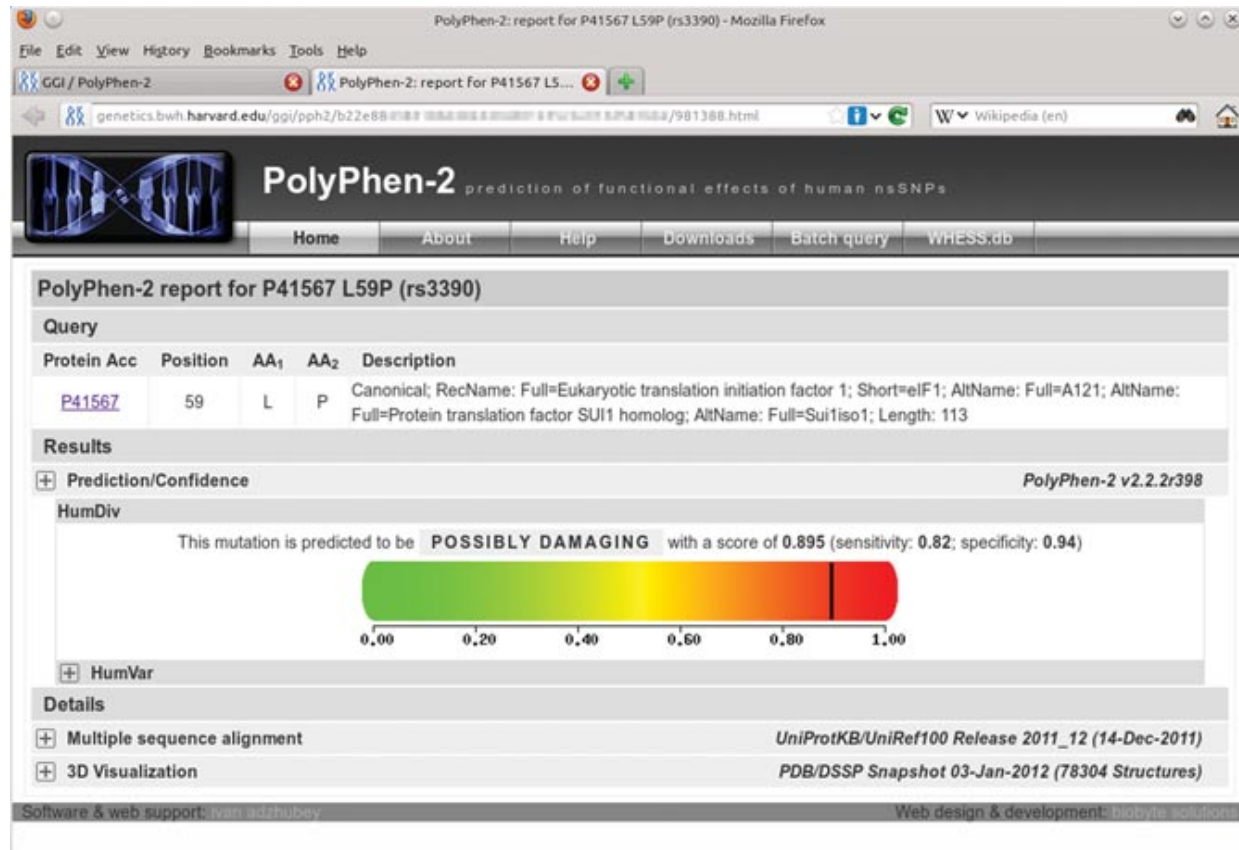
- Probabalistic classifier
- Prediction is based on a number of sequence, phylogenetic, and structural features characterizing the substitution
 - Maps coding SNPs to gene transcripts
 - Extracts protein sequence annotations and structural attributes
 - Builds conservation profiles
 - Then estimates the probability of the missense mutation being damaging based on a combination of all these properties

PolyPhen

- Can use the web interface
 - Provide protein identifier (UniProtKB accession number or entry name)
 - Enter position of the substitution in the protein sequence into the position text box
 - Indicate reference amino acid residue, and the substitution residue
 - Can also analyze large datasets of single nucleotide changes using batch mode
 - Get a web page with download options after the batch runs
- Also command line capability

PolyPhen

- Web interface
 - Heatmap color bar with the black indicator illustrating the strength of the putative damaging effect for the variant



PolyPhen

- Web interface
 - Multiple sequence alignment and black box around variant
 - 3D protein structure with variant location marked in red
 - Interactive

The screenshot displays the PolyPhen web interface. The top section, titled "Multiple sequence alignment", shows a UniProtKB/UniRef100 alignment (Release 2011_12, 14-Dec-2011) for the query sequence (sp|C1C511#1) and several other sequences. The alignment is color-coded by amino acid type, and a black box highlights the mutation position (Leu72). Below the alignment, a text box states: "Shown are 75 amino acids surrounding the mutation position (marked with a black box). An interactive version of the complete alignment is [also available](#)." The bottom section, titled "3D Visualization", shows a PDB/DSSP Snapshot (03-Jan-2012, 78304 Structures) of the protein structure. The structure is a ribbon diagram, and the mutation (Leu72) is highlighted in red. To the right of the structure, the following information is displayed: EntryID: 2IF1, ChainID: A, Residue: Leu72, Identity: 100.0%, and Overlap: 100.0% (113 aa). At the bottom of the 3D visualization, there are buttons for "Zoom into mutation", "Reset view", and a "View size" slider.

VEP

- Variant Effect Predictor (VEP)
- Determines effect of SNPs, insertions, deletions, CNVs, or structural variants on
 - Genes, transcripts, protein sequences, and regulatory regions
- Input coordinates of variants and nucleotide changes to find
 - Genes affected by the variants
 - Location of the variants
 - Upstream of the transcript
 - In a coding sequence
 - In non-coding RNA
 - In regulatory regions
 - Exons, introns
 - **PolyPhen predictions**
- Consequence of your variants
 - Stop gained, missense, stop lost, frame shift

Sometimes Analysis
Sources and Databases
are Like This:



VEP

- Find co-located known variants
 - Report known variants from the Ensembl Variation database
- Report of minor allele frequency data from the 1000 Genomes data and NHLBI-ESP

VEP

- Web interface that suits small volumes of data
- User-friendly command line for Unix, Linux



Variant Effect Predictor ⓘ

New VEP job:

ⓘ VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Input

Species:  Human (Homo sapiens) 
Assembly: GRCh38

Name for this data (optional):

Input file format ([details](#)):

Ensembl default ▾

Either paste data:

```
1 909238 909238 G/C +  
3 361464 361464 A/- +  
5 121187650 121188519 DUP
```

Or upload file:

No file selected.

Or provide file URL:

Transcript database to use:

- ☒ Ensembl transcripts
☐ Gencode basic transcripts
☐ RefSeq transcripts
☐ Ensembl and RefSeq transcripts

Output options

[Identifiers and frequency data](#) ⓘ Additional identifiers for genes, transcripts and variants; frequency data

[Extra options](#) ⓘ e.g. SIFT, PolyPhen and regulatory data

[Filtering options](#) ⓘ Pre-filter results by frequency or consequence type

SIFT

- Sorting Intolerant from Tolerant (SIFT)
- <http://sift.jcvi.org/>
- SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST

SnEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
 - Annotates and predicts the effects of variants on genes (such as amino acid changes)
 - SnpEff is really fast: calculated predictions for all the SNPs in the 1000 Genomes project in less than 15 minutes
- <http://snpeff.sourceforge.net/>
- Also has SnpSift
 - Once you annotated your files using SnpEff, you can use SnpSift to help you filter large genomic datasets in various ways

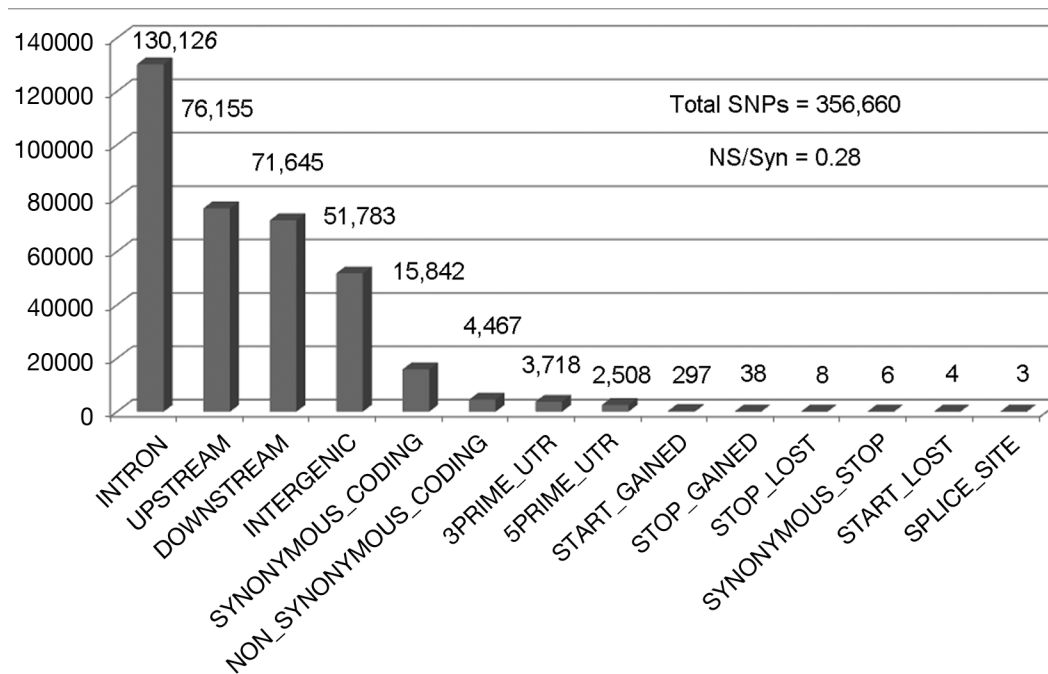
SnpEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
 - Speed—the ability to make thousands of predictions per second
 - Flexibility—the ability to add custom genomes and annotation
 - Can integrate with Galaxy, an open access and web-based platform for computational bioinformatic research
 - Compatibility with multiple species and multiple codon usage tables (including mitochondrial genomes)
 - Integration with the Genome Analysis Toolkit (GATK)
 - Ability to perform non-coding annotations

Sub-field	Notes
Effect	Effect of this variant. See details below
Codon_Change	Codon change: old_codon/new_codon
Amino_Acid_change	Amino acid change: old_AA/new_AA
Warnings	Any warnings or errors
Gene_name	Gene name
Gene_BioType	BioType, as reported by ENSEMBL
Coding	[CODING NON_CODING]. If information reported by ENSEMBL (e.g., has 'protein_id' information in GTF file)
Transcript	Transcript ID (usually ENSEMBL)
Exon	Exon ID (usually ENSEMBL)
Warnings	Any warnings or errors (not shown if empty)

SnPEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
 - More genome versions
 - Open source for any user
 - Supports VCF files



SnPEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
 - Similar in many ways to ANNOVAR and VAAST
 - <http://www.openbioinformatics.org/annovar/>
 - <http://www.yandell-lab.org/software/vaast.html>

VAT

- <http://varianttools.sourceforge.net/Association/HomePage>
- Variant Association Tools
- Large collection of utilities devoted to data exploration, quality control and association analysis of rare/common single nucleotide variants and indels

STRING

- <http://string-db.org/>
- A database of known and predicted protein interactions
- The interactions include direct (physical) and indirect (functional) associations
- Can search by protein name, or protein sequence

Which One Do I Choose?

- It is a challenge
 - Variety of similar and also different information
 - Different approaches for estimating effect on proteins
 - It does become a matter of opinion and specific needs of a given project
 - The good news is that work is in progress to unify multiple sources so you can gather information from multiple sources

Questions?