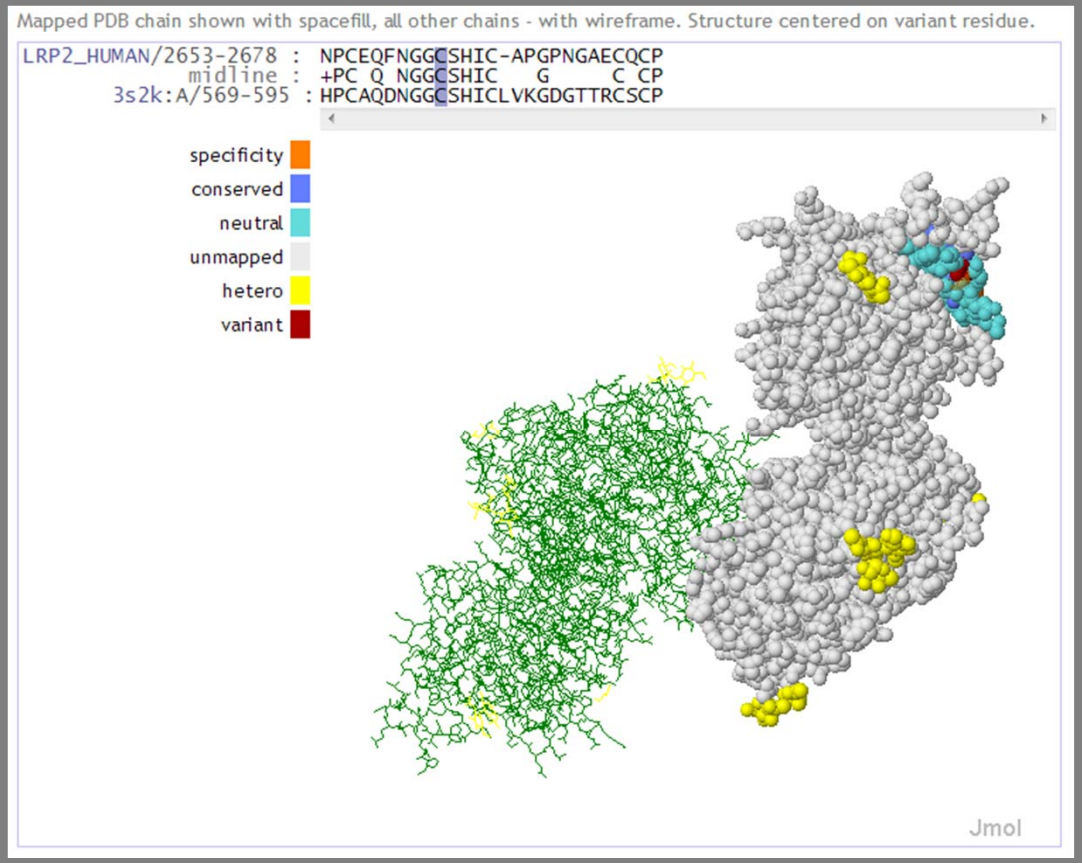


Knowing Your NGS Downstream: Functional Predictions

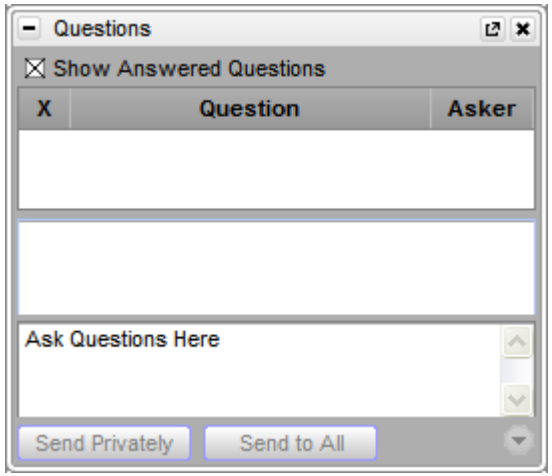
May 15, 2013

Bryce Christensen
Statistical Geneticist / Director of Services

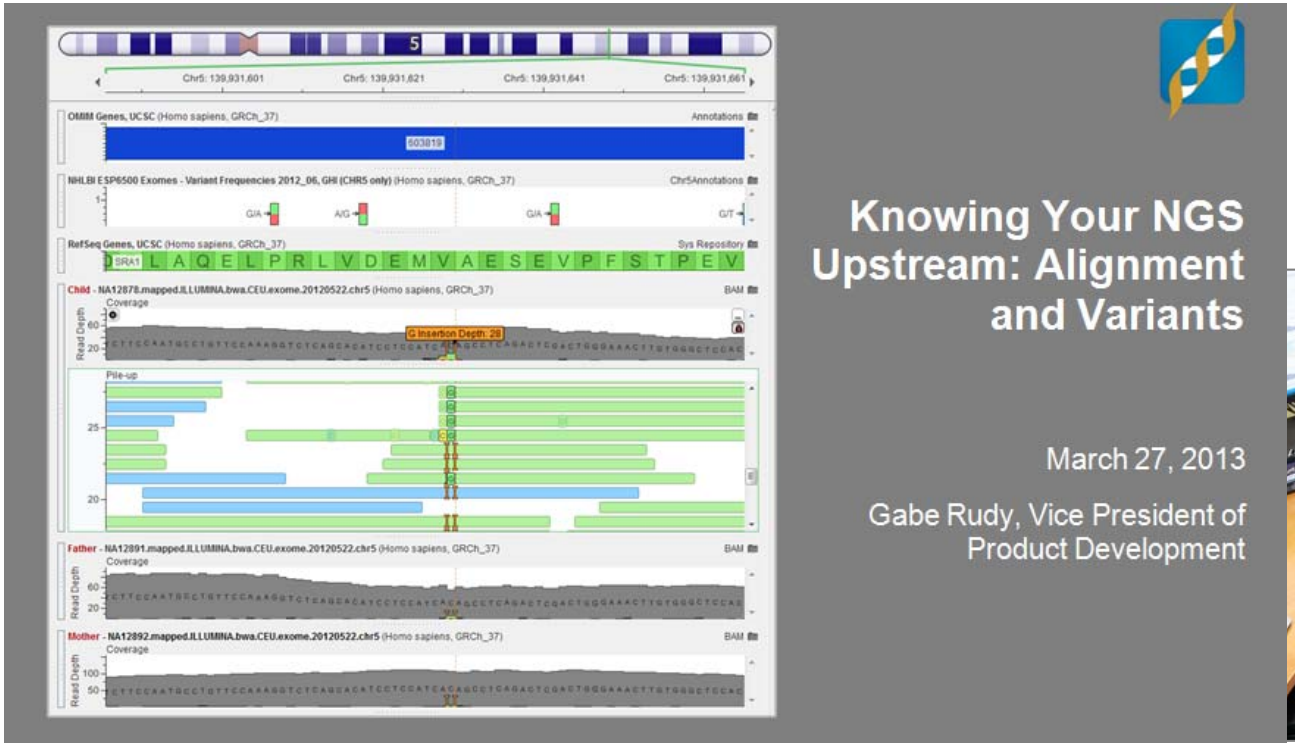
GOLDEN HELIX
Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



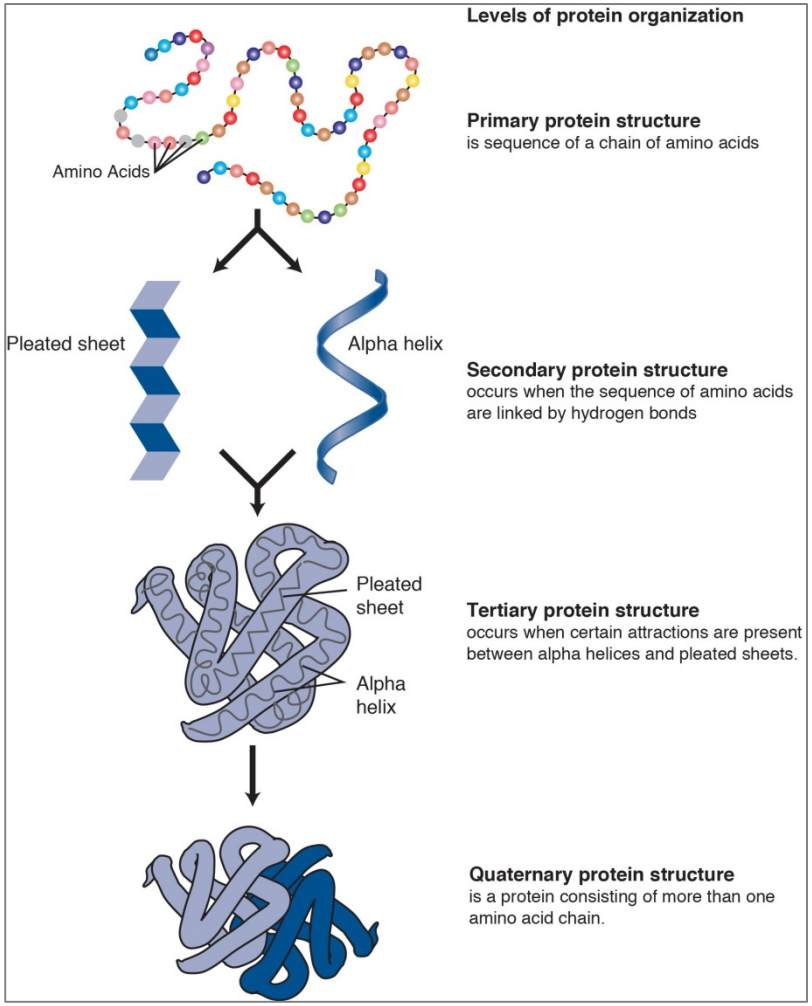
Previous Webinar



- Extremely popular
- Available to view at www.goldenhelix.com
- Feedback inspired today's presentation about downstream analysis

Today's Presentation

- **What I Assume About You**
  - Some experience with NGS technology and downstream analysis of genomic data
  - Not intimidated by the figure at the right→
  - Curious to learn more about the process and practice of predicting functional consequences of genetic variants
- **What You Will Learn**
  - The informatics processes that underlie functional predictions
  - How to apply functional predictions in your own research
- **What You Won't Learn**
  - One true way to make functional predictions



www.genome.gov

NGS Analysis

Primary Analysis

- Analysis of hardware generated data, on-machine real-time stats.
- Production of sequence reads and quality scores

Secondary Analysis

- QA and clipping/filtering reads
- Alignment/Assembly of reads
- Recalibrating, de-duplication, variant calling on aligned reads

Tertiary Analysis

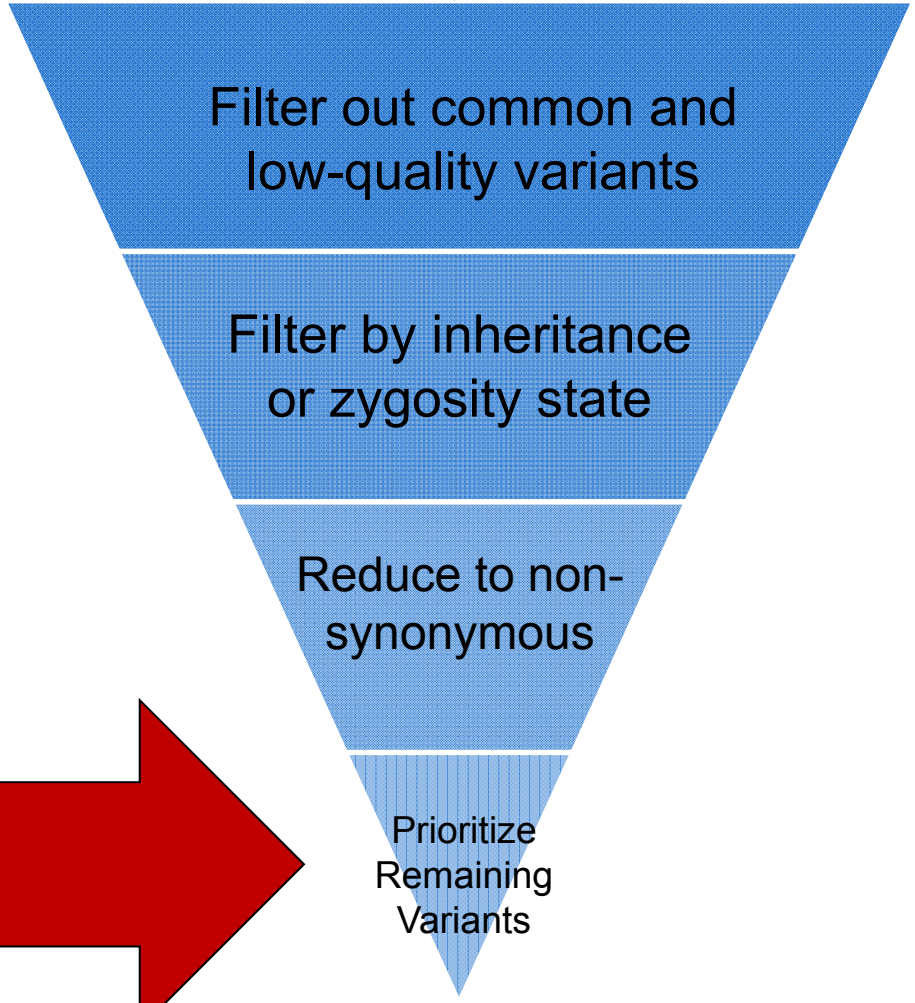
"Sense Making"

- QA and filtering of variant calls
- Annotation and filtering of variants
- Multi-sample integration
- Visualization of variants in genomic context
- Experiment-specific inheritance/population analysis

Sample Variant Analysis Workflow

VCF file goes in

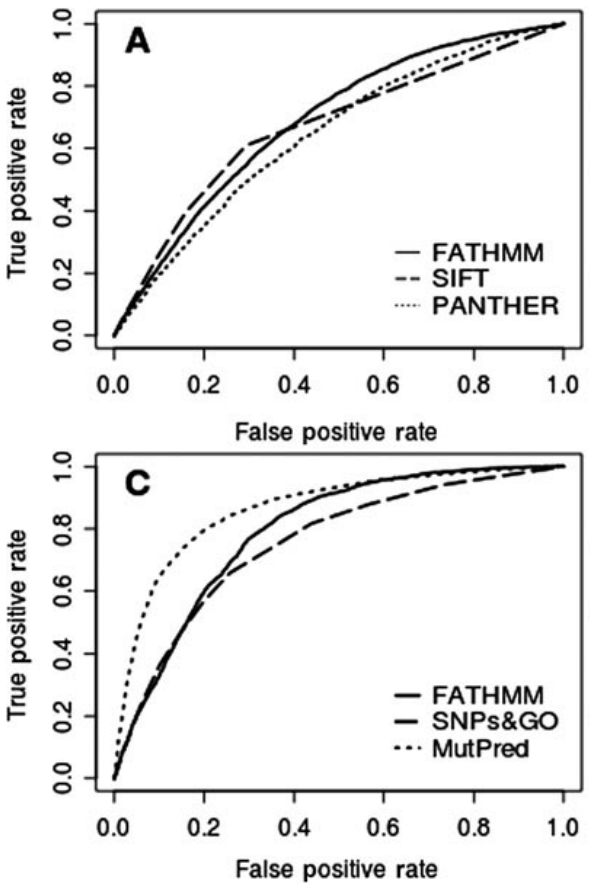
- Many NGS tertiary analysis workflows follow a system of **annotation-based filtering**
- Common to have **a long list** of candidate variants
- Variants need to be prioritized for validation experiments
- **Prioritizing those candidates** is extrememly important, but can be a very difficult process



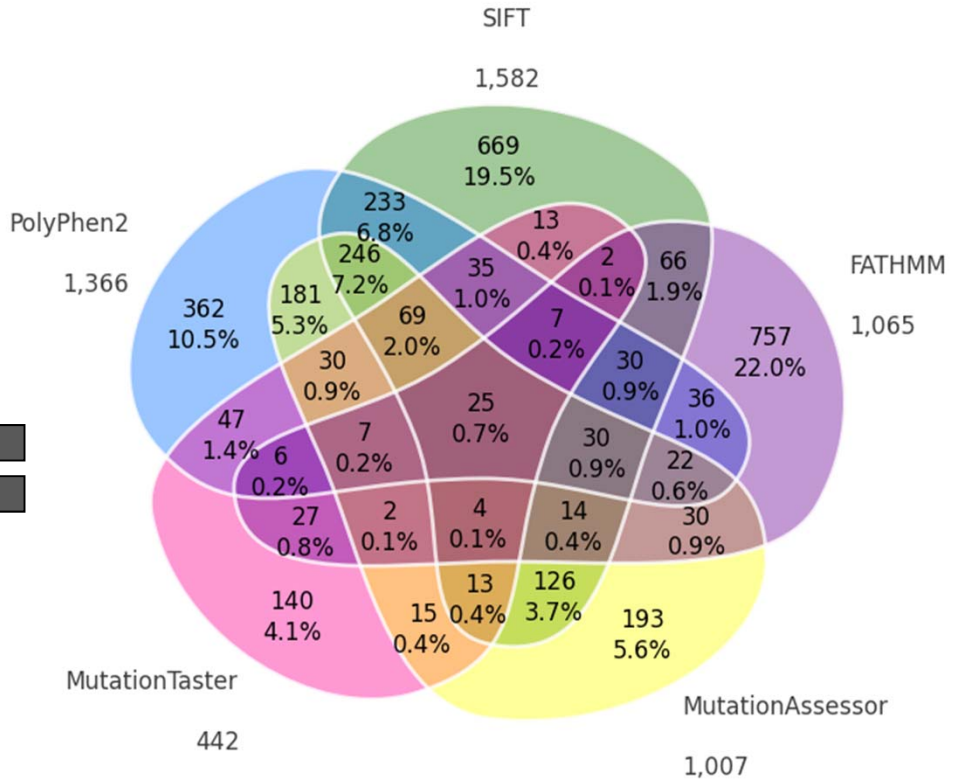
Functional Prediction Algorithms

- SIFT
- PolyPhen
- PolyPhen-2
- MutationTaster
- MutationAssessor
- FATHMM
- PANTHER PSEC
- SNPs&GO
- MutPred
- SNAP
- PMut
- TopoSNP
- SNPs3D
- VEST
- PhD-SNP
- X-Var
- Align-GVGD
- PROVEAN
- nsSNPAnalyzer
- LRT

Motivation



≠



- Published comparisons indicate that **most prediction algorithms are similar** in their ability to detect true functional variants
- But in practice, **they rarely agree about much of anything**

Agenda

1. The Basics of Molecular Biology & Functional Predictions
2. Overview of Commonly Used Algorithms
3. Comparisons
4. Applying Functional Predictions

The Central Dogma of Molecular Biology

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."
-- Francis Crick, 1958

- **In other words:**
  - DNA is transcribed to RNA
  - RNA is translated to create proteins
  - Unidirectional process
- **Protein is where damaging effects of a DNA mutation will be observed**
- **Functional prediction algorithms are based almost entirely on protein sequences**

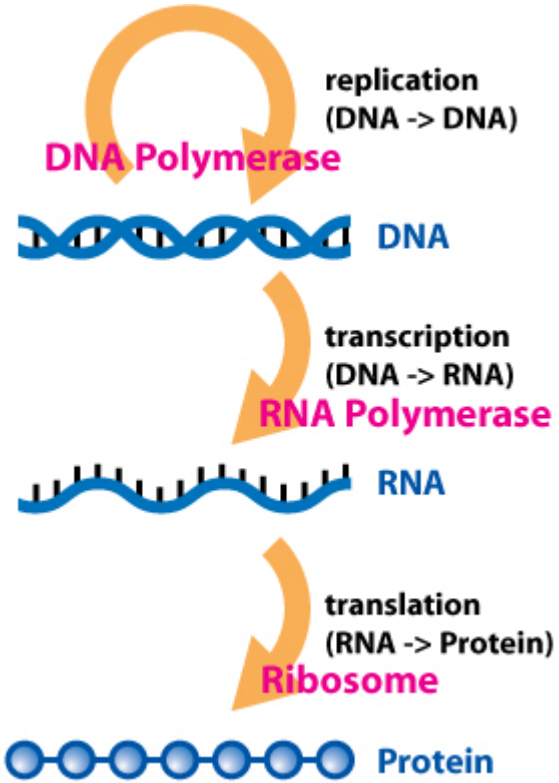


Image from Wikimedia Commons, Dhorspool

Transcription

- Transcription is the process by which an **RNA transcript is created from DNA within the cell nucleus** before moving to the cytoplasm

- Includes splicing exons together to create **meaningful transcripts**

- The complete collection of mRNA transcripts in a given cell or tissue is often called the **“transcriptome”**

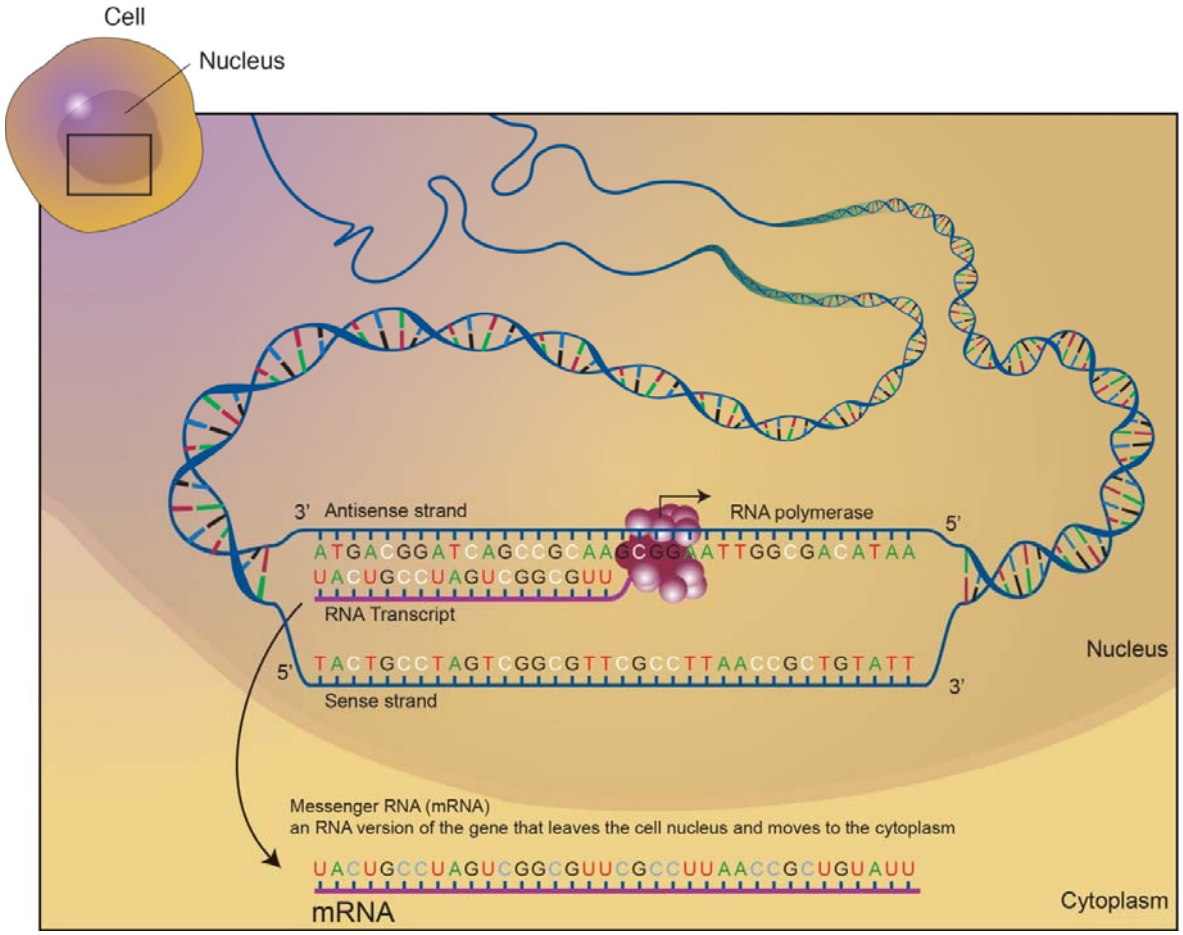


Image from genome.gov

Translation

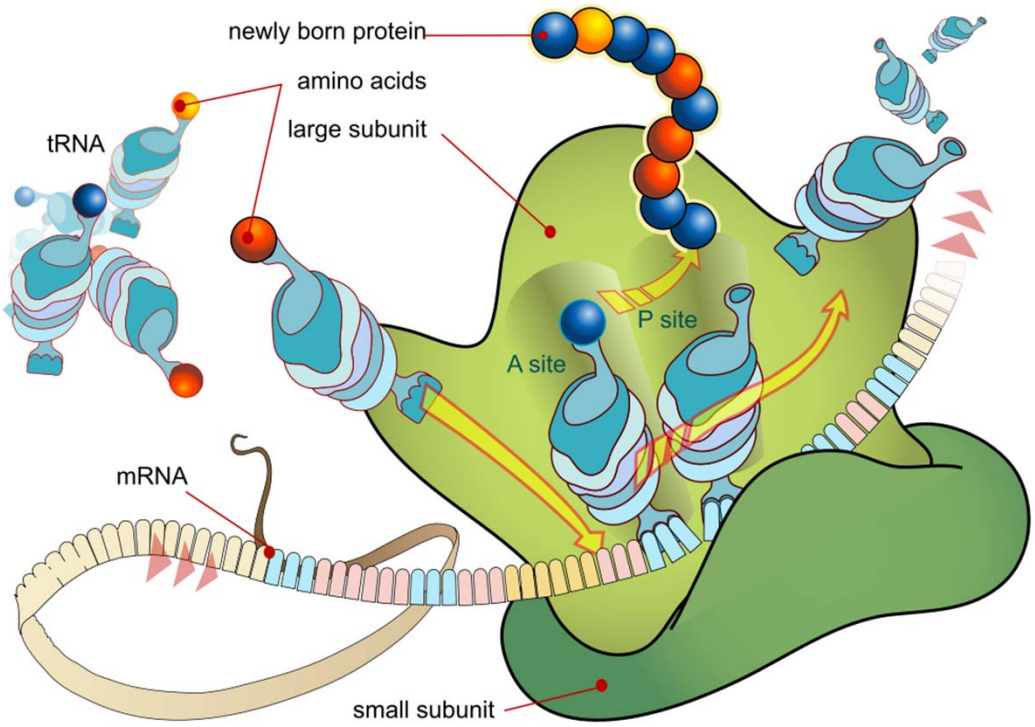
RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G
	Amino Acids				

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

GOLDEN HELIX
Accelerating the Quest for Significance™

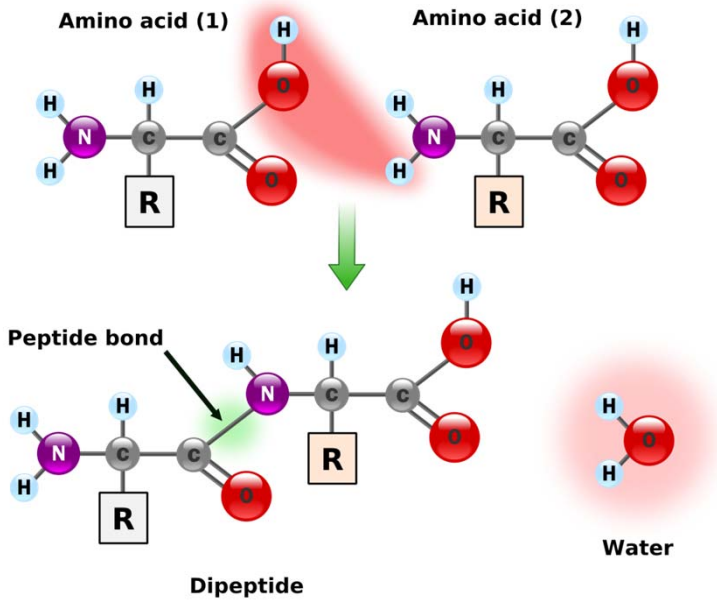
- **mRNA transcripts are converted to amino acid sequences** via the translation process
- Think of it as a **different language**; nucleic acids versus amino acids



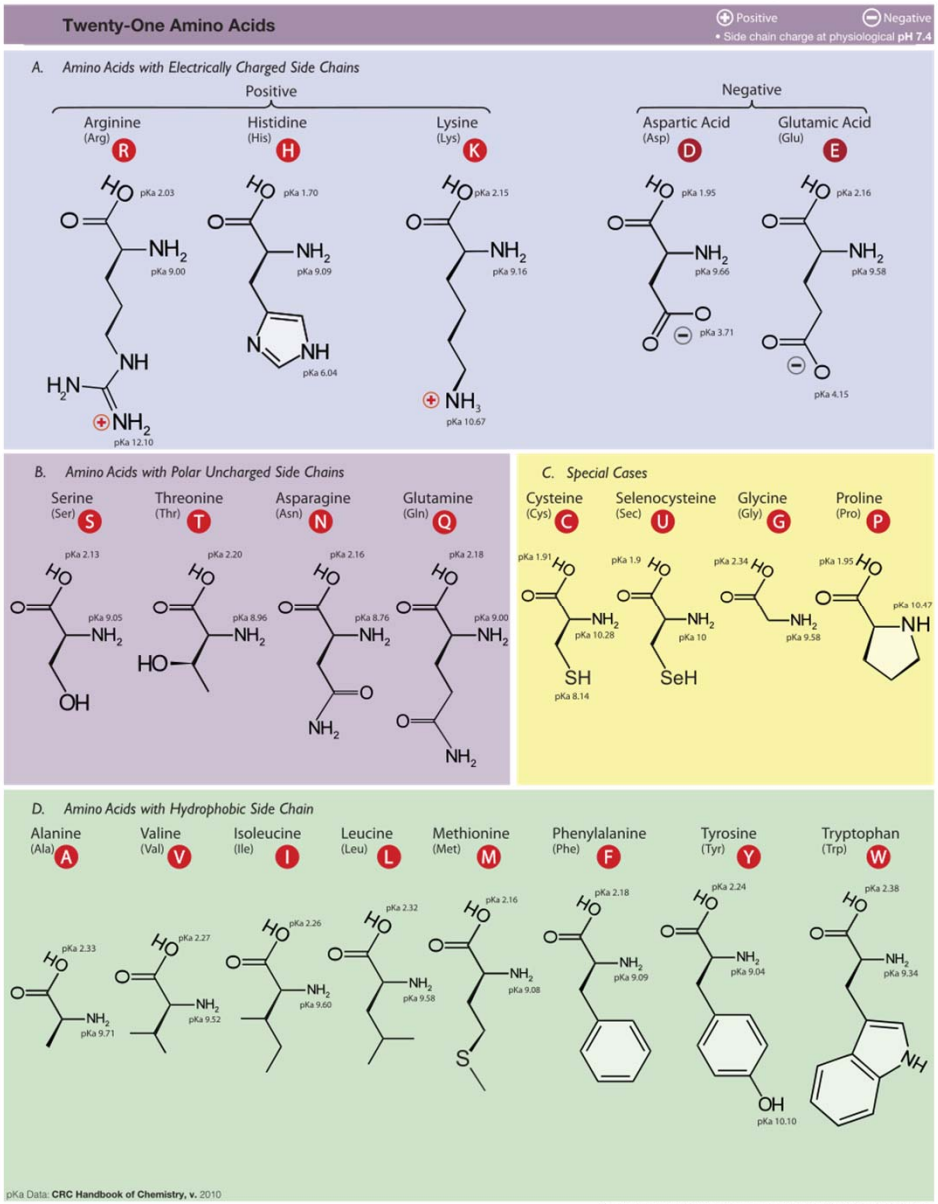
Images from genome.gov and WikiMedia Commons, by ladyofhats

Amino Acid Properties

- Amino Acids are distinguished by their **respective residues** (aka side-chains or R-groups)
- Residues are classified by polarity, volume, hydrophobic and other physicochemical properties



Images from WikiMedia Commons, by YassineMrabet and DanCojocari



Levels of Protein Structure

- **Primary Structure**
  - Linear sequence of amino acids
- **Secondary Structure**
  - Interaction between amino acids via hydrogen bonding results in regular substructures called alpha helices and beta sheets
- **Tertiary Structure**
  - The final three-dimensional form of an amino acid chain
  - Is influenced by attractions between secondary structures
- **Quaternary Structure**
  - Several tertiary structures may interact to form quaternary structures

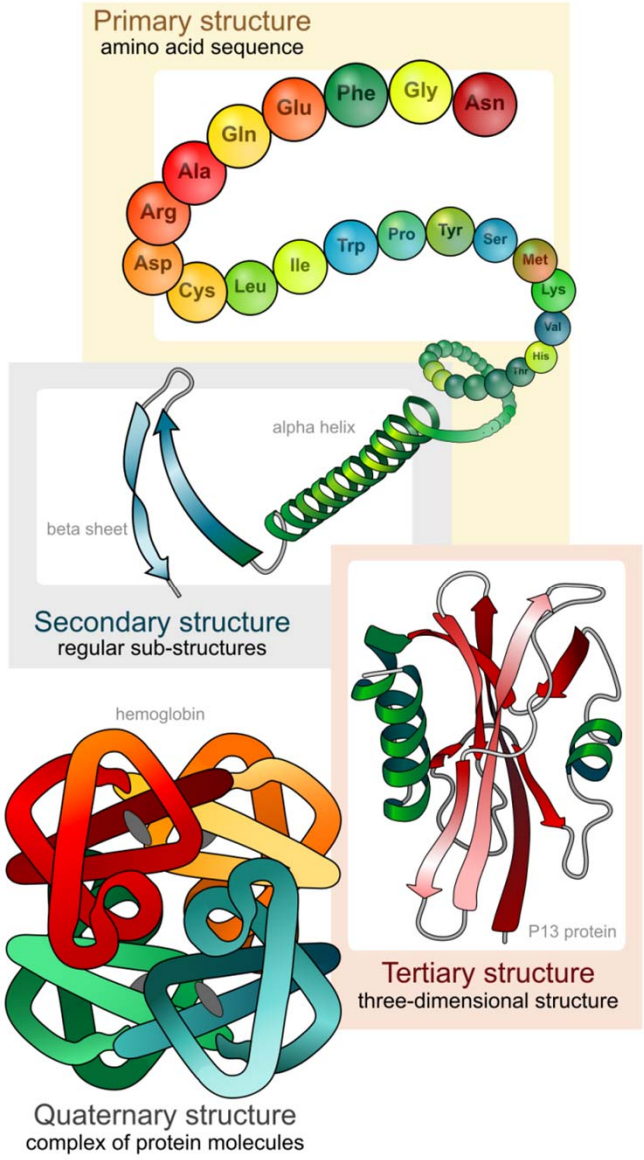
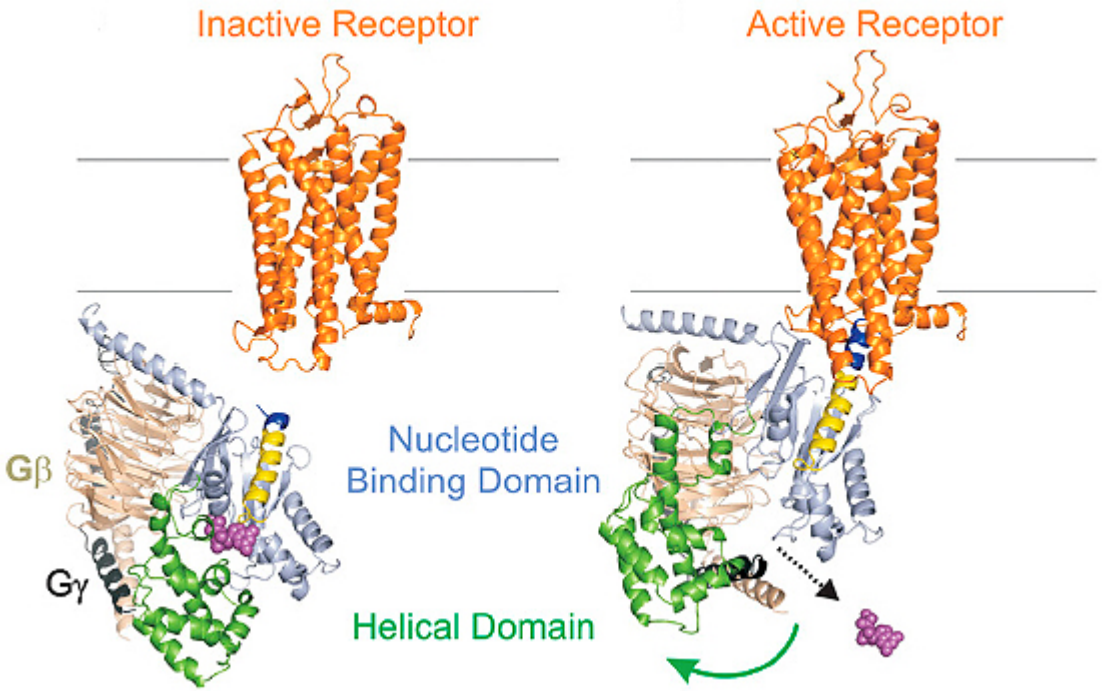


Image from WikiMedia Commons, ladyofhats

From Structure to Function

- Proteins include various types of functional domains, binding sites and other surface features
  - This determines how the protein interacts with other molecules
- Replacing certain amino acids may have **drastic effects** on the protein structure
  - Thereby affecting the protein function

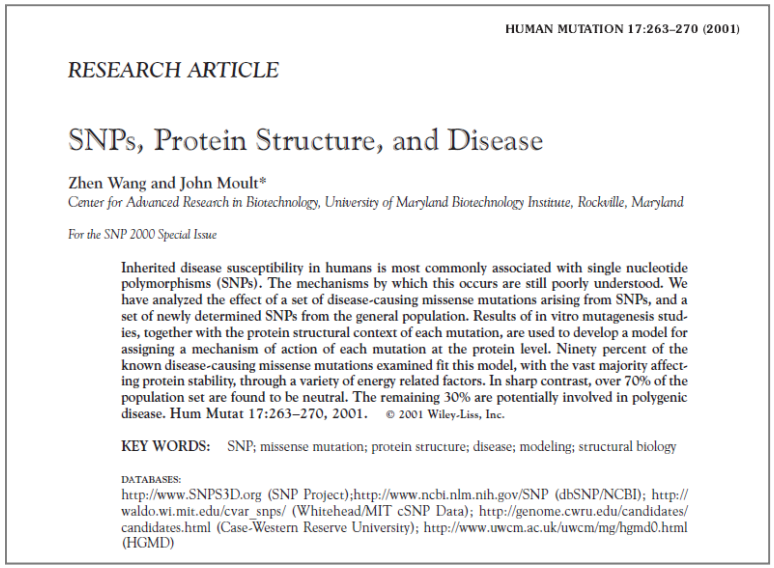
- If we know how the protein structure is affected by an amino acid substitution, we can make a **good guess** about functional consequences.
- **The problem is that we don't know the wild-type 3D strucuture of most proteins.**



http://www.vanderbilt.edu/vicb/DiscoveriesArchives/g_protein_receptor.html

Using Primary Structure as Proxy for Tertiary

- 83% of disease-causing mutations affect **stability** of proteins (Wang and Moult, 2001)
- 90% of disease-causing mutations can be detected using **structure and stability**
- Many human proteins have numerous homologs:
  - **Paralogs:** Separated by a gene duplication event
  - **Orthologs:** Separated by speciation
- Don't know the exact structure of most proteins, but we can **compare amino acid sequences** to identify domains and motifs conserved by evolution
- **Disease causing mutations are overrepresented at conserved sites in the primary structure** (Miller and Kumar, 2001)



HUMAN MUTATION 17:263–270 (2001)

RESEARCH ARTICLE

SNPs, Protein Structure, and Disease

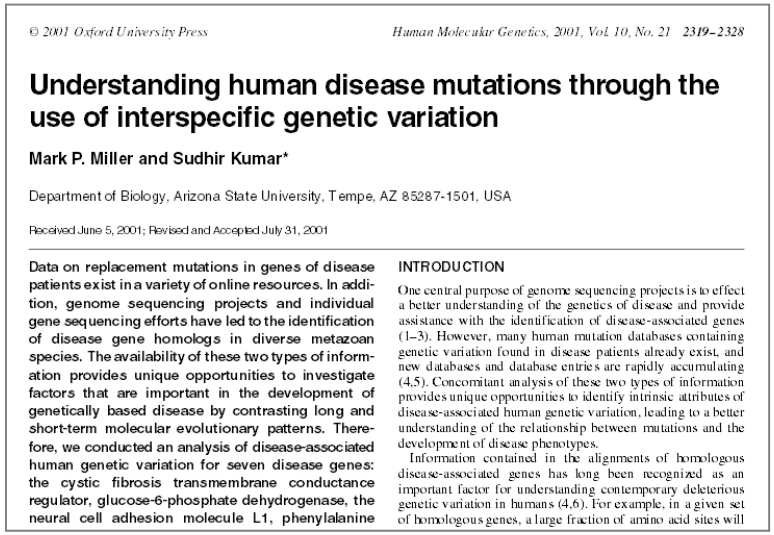
Zhen Wang and John Moult*
Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland

For the SNP 2000 Special Issue

Inherited disease susceptibility in humans is most commonly associated with single nucleotide polymorphisms (SNPs). The mechanisms by which this occurs are still poorly understood. We have analyzed the effect of a set of disease-causing missense mutations arising from SNPs, and a set of newly determined SNPs from the general population. Results of in vitro mutagenesis studies, together with the protein structural context of each mutation, are used to develop a model for assigning a mechanism of action of each mutation at the protein level. Ninety percent of the known disease-causing missense mutations examined fit this model, with the vast majority affecting protein stability, through a variety of energy related factors. In sharp contrast, over 70% of the population set are found to be neutral. The remaining 30% are potentially involved in polygenic disease. Hum Mutat 17:263–270, 2001.

KEY WORDS: SNP; missense mutation; protein structure; disease; modeling; structural biology

DATABASES:
http://www.SNPS3D.org (SNP Project);http://www.ncbi.nlm.nih.gov/SNP (dbSNP/NCBI); http://waldo.wi.mit.edu/cvar_snps/ (Whitehead/MIT cSNP Data); http://genome.cwru.edu/candidates/candidates.html (Case-Western Reserve University); http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html (HGMD)



 Human Molecular Genetics, 2001, Vol. 10, No. 21 2319–2328

Understanding human disease mutations through the use of interspecific genetic variation

Mark P. Miller and Sudhir Kumar*

Department of Biology, Arizona State University, Tempe, AZ 85287-1501, USA

Received June 5, 2001; Revised and Accepted July 31, 2001

Data on replacement mutations in genes of disease patients exist in a variety of online resources. In addition, genome sequencing projects and individual gene sequencing efforts have led to the identification of disease gene homologs in diverse metazoan species. The availability of these two types of information provides unique opportunities to investigate factors that are important in the development of genetically based disease by contrasting long and short-term molecular evolutionary patterns. Therefore, we conducted an analysis of disease-associated human genetic variation for seven disease genes: the cystic fibrosis transmembrane conductance regulator, glucose-6-phosphate dehydrogenase, the neural cell adhesion molecule L1, phenylalanine

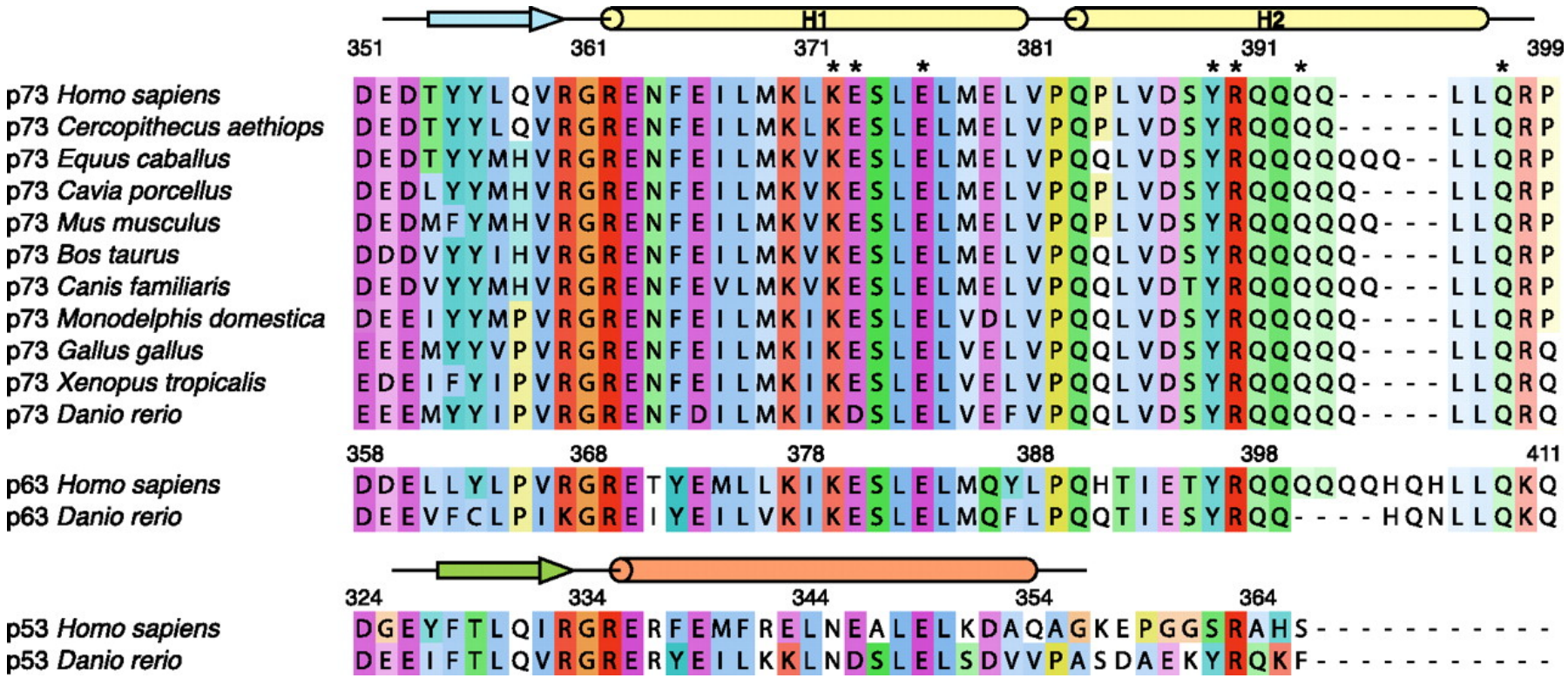
INTRODUCTION

One central purpose of genome sequencing projects is to effect a better understanding of the genetics of disease and provide assistance with the identification of disease-associated genes (1–3). However, many human mutation databases containing genetic variation found in disease patients already exist, and new databases and database entries are rapidly accumulating (4,5). Concomitant analysis of these two types of information provides unique opportunities to identify intrinsic attributes of disease-associated human genetic variation, leading to a better understanding of the relationship between mutations and the development of disease phenotypes.

Information contained in the alignments of homologous disease-associated genes has long been recognized as an important factor for understanding contemporary deleterious genetic variation in humans (4,6). For example, in a given set of homologous genes, a large fraction of amino acid sites will

Multiple Sequence Alignment

- A **multiple sequence alignment** (MSA) comparing the amino acid (AA) sequence of protein homologs can be generated by BLAST or similar algorithms
- Almost all contemporary functional prediction algorithms incorporate MSAs in some manner



Joerger A C et al. PNAS 2009;106:17705-17710

More on Multiple Sequence Alignments (MSAs)

- MSAs may include **700 or more homologous sequences** in some methods
- Prediction algorithms may incorporate orthologs and/or paralogs in the MSA
- Distantly related orthologs are frequently cited (especially SIFT authors) as giving optimum prediction performance
  - **Be cautious**—phylogenetic relationship doesn't always mean that the protein has the same function or is similarly important in both species
  - Some authors (especially PolyPhen2) argue that **a combination** of paralogs and orthologs is best
- While most functional prediction algorithms incorporate MSAs, they differ in how the MSA is interpreted and how AA substitutions are scored

Trained and Untrained Algorithms

Trained/Weighted Algorithms

- **Machine learning** methods
- Classify the functional consequence of a given mutation based on characteristics observed in a **selected set of mutations** known to be either damaging or benign
- May include known disease sequences in the MSA
- Selection of training data is important factor in algorithm performance and appropriateness for any given analysis project
- *Examples: PolyPhen-2, MutationTaster*

Untrained Algorithms

- Do not incorporate machine learning techniques.
- A given mutation is classified based on a **theoretical model** incorporating important prior knowledge about the types of mutations that are expected to cause disease
- **May not carry some of the biases** present in a trained algorithm and may have more general applicability for various analysis projects
- *Examples: SIFT, MutationAssessor, FATHMM-unweighted*

Agenda

1. The Basics of Molecular Biology & Functional Predictions
2. **Overview of Commonly Used Algorithms**
3. Comparisons
4. Applying Functional Predictions

Five Algorithms to Review

Algorithm	Pub Year	Citations (G.Schol.)	Host Inst.	Category	Distinguishing Characteristic
SIFT	2003	>1200	JCVI (UW)	Untrained	Popular, broadly applicable and intuitive method to identify functional mutations.
PolyPhen2	2010	>1000	Harvard/ BWH	Trained	Provides 2 scores (HumDiv and HumVar) for applications to complex and Mendelian disease, respectively.
Mutation Assessor	2011	57	MSKCC	Untrained	Considers AA conservation in protein subfamilies to refine important functional regions. Interactive user interface.
Mutation Taster	2010	199	Charité - Berlin	Trained	Native support for DNA (rather than AA) variant analysis. Allows online submission of VCF files.
FATHMM	2013	NA	U Bristol	Trained (weighted)	Uses HMM method (rather than BLAST) to create MSA. Weighted extensions for human disease and cancer analysis.

These five methods were selected due their inclusion in the Database for NonSynonynous Functional Predictions (*dbNSFP: Liu et al., 2011*) which can be accessed within Golden Helix SNP & Variation Suite (SVS)

dbNSFP

- **The Database for NonSynonymous Functional Predictions (dbNSFP)** is a **free tool** developed by Dr. Xiaoming Liu**.** [Hum Mutat 32(8):894, 2011]
- Catalogs several pre-computed conservation and functional prediction scores for all possible nsSNPs in the human genome
- Downloadable database and Java program for annotating variants in VCF
  - 75 variables returned for each queried variant
- **Conservation scores:**
  - PhyloP, GERP++, SiPhy
- **Functional Predictions:**
  - SIFT, PolyPhen-2, LRT, MutationAssessor, MutationTaster, FATHMM
- **Other Annotations:**
  - Variant frequencies, disease associations, transcript data, haploinsufficiency
- Available at (*https://sites.google.com/site/jpopgen/dbNSFP*)

UTHealth
The University of Texas
Health Science Center at Houston

SIFT

- "**Sorting Intolerant From Tolerant**" (*sift.jcvi.org*)

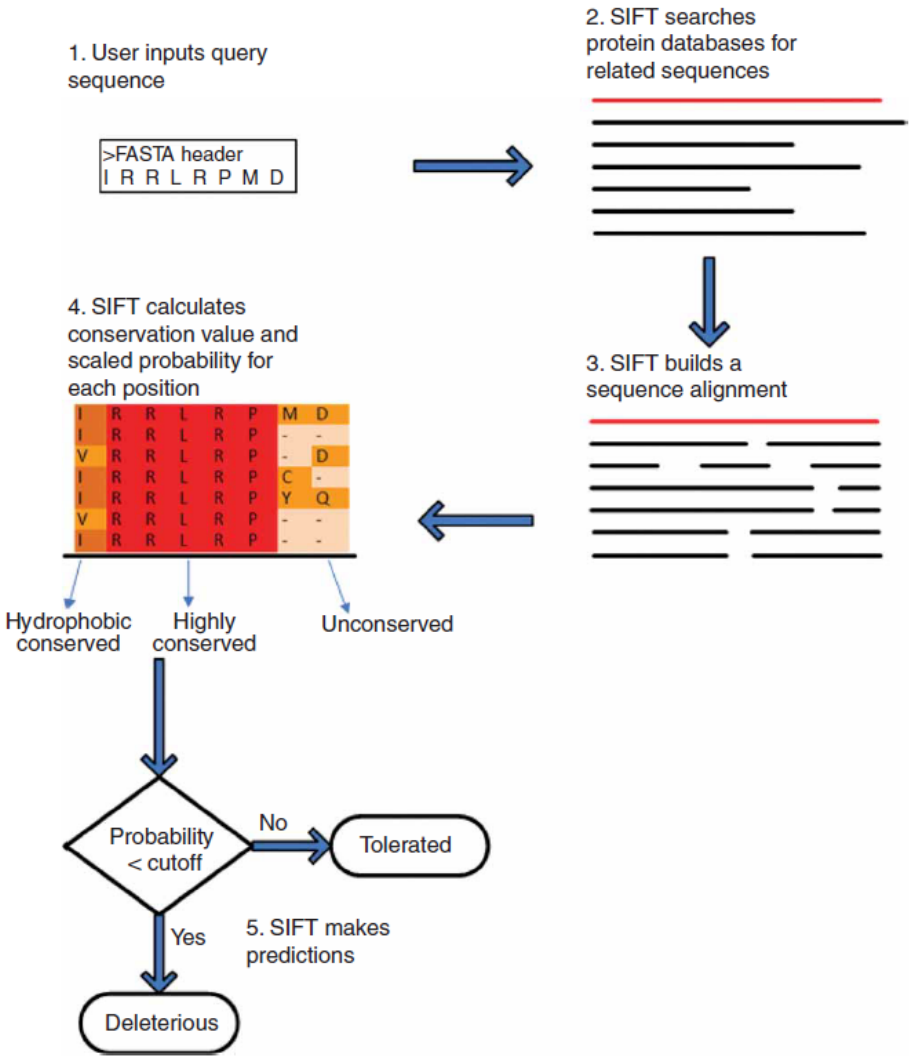
J. Craig Venter™ INSTITUTE

- "SIFT predicts whether an amino acid substitution affects protein function. 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."

- **Publications:**
  - Predicting Deleterious Amino Acid Substitutions. *Genome Res.* 2001 May;11(5):863-74.
    - Cited by 844 (per Google Scholar)
  - SIFT: predicting amino acid changes that affect protein function. Nucl. Acids Res. (2003) 31 (13): 3812-3814
    - Cited by 1,248
  - Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc. 2009*;4(7):1073-81
    - Cited by 564

SIFT: How It Works

- Relies **entirely on sequence and does not include structural features**
- Builds an MSA based on PSI-BLAST and considers several features in **scoring a variant AA**:
  - Is the position highly conserved for a single amino acid?
  - Is the position highly conserved for amino acids with a particular polarity, charge, or other chemical property?
  - How different is the mutant AA from the most common AA in the MSA?



Nat Protoc. 2009;4(7):1073-81

SIFT Scores and Predictions

Predict not tolerated	Position	Seq Rep	Predict tolerated
d c g w h n e s p r k q y t	3M	0.60	a f v i ML
w	4A	0.60	c f m y i h v l p r q t n k s e GDA
w m h f	5C	0.80	y i q r e l k d p n v g t a SC
w c m f	6R	0.80	y i h v p l d g n q e t a k SR
w d h q p n c e r g k s	7V	0.80	y t a mF l I V
w d h g	8I	0.80	n c r q p e y k s f m t AvLI
	9N	0.80	w c h p f y Mi q r v g e d t a k s l N
c w m f i d	10R	0.80	p v y g s l t n a e q Hk R
c w f m y i v d h p g l	11R	0.80	n s t a e k Q R
c w f m y i v d h p g l	12R	0.80	n s t a e k Q R
y w v t s r q p n m l k i g f e d c a	13H	0.80	H

AAs in capital letters appeared at least once in the MSA

Scores in black are predicted to be tolerated

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1M 0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2E 0.25	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3M 0.50	0.07	0.02	0.02	0.03	0.12	0.02	0.02	0.24	0.03	1.00	0.63	0.02	0.03	0.03	0.03	0.03	0.05	0.17	0.02	0.04
4A 0.50	1.00	0.05	0.13	0.17	0.04	0.68	0.04	0.05	0.16	0.09	0.04	0.12	0.14	0.10	0.10	0.30	0.15	0.11	0.01	0.05
5C 0.75	0.59	1.00	0.17	0.15	0.09	0.33	0.08	0.11	0.17	0.15	0.06	0.19	0.18	0.12	0.12	0.93	0.44	0.21	0.03	0.11
6R 0.75	0.37	0.04	0.24	0.36	0.06	0.23	0.11	0.10	0.58	0.17	0.06	0.26	0.15	0.29	1.00	0.63	0.33	0.15	0.02	0.09
7V 0.75	0.10	0.03	0.02	0.03	0.22	0.04	0.03	0.99	0.04	0.48	0.10	0.03	0.03	0.03	0.04	0.05	0.09	1.00	0.02	0.09
8I 0.75	0.44	0.07	0.07	0.13	0.21	0.07	0.07	0.86	0.13	1.00	0.21	0.08	0.09	0.10	0.10	0.13	0.24	0.85	0.04	0.12
9N 0.75	0.90	0.15	0.73	0.85	0.41	0.60	0.37	0.54	0.94	1.00	0.36	0.95	0.36	0.59	0.66	0.93	0.84	0.64	0.10	0.46
10R 0.75	0.14	0.02	0.08	0.16	0.07	0.10	0.38	0.06	0.46	0.12	0.04	0.13	0.08	0.19	1.00	0.13	0.13	0.09	0.02	0.11
11R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
12R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
13H 0.75	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

GOLDEN HELIX
Accelerating the Quest for Significance™

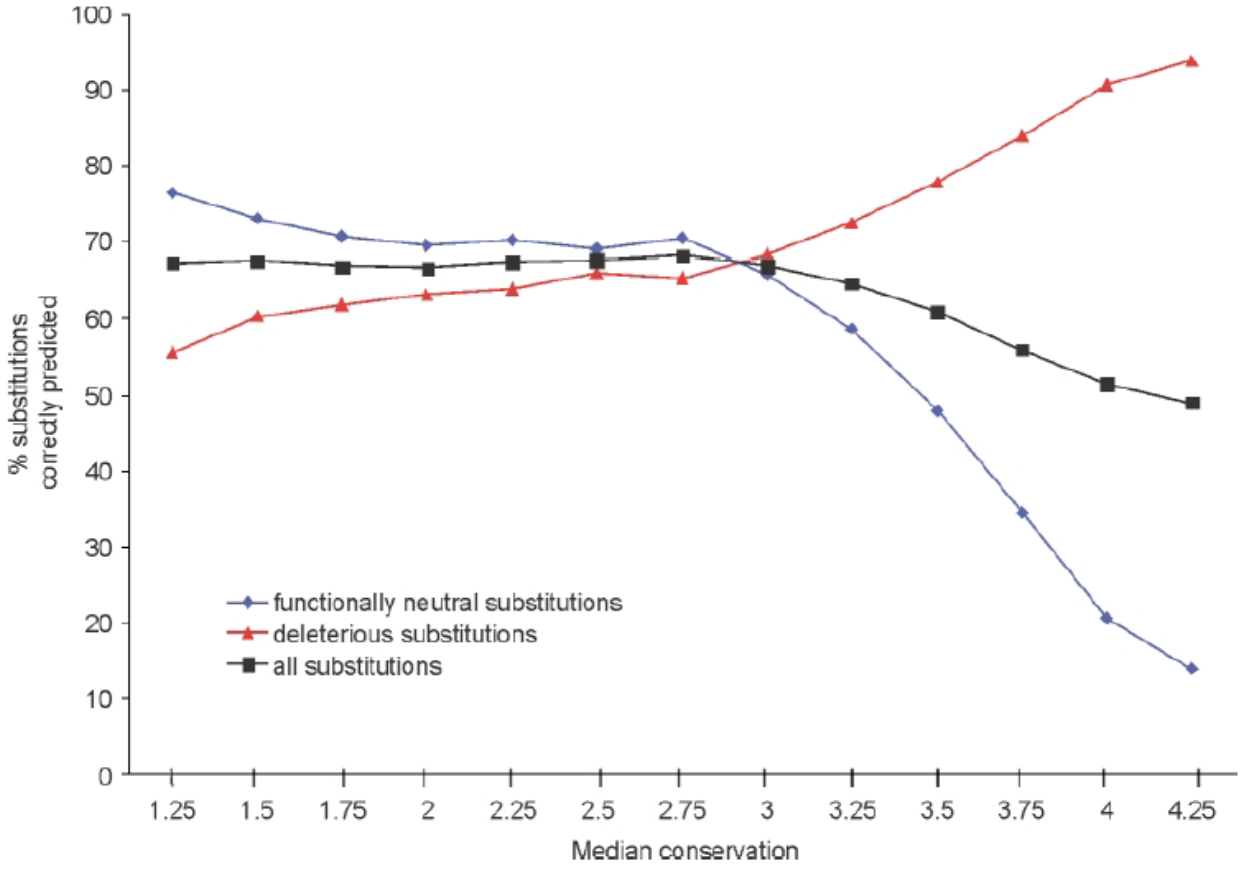
Nat Protoc. 2009;4(7):1073-81

SIFT MSA Diversity vs. Prediction Accuracy

- "Confidence in a substitution **predicted to be deleterious depends on the diversity of the sequences in the alignment**. If the sequences used for prediction are closely related, then many positions will [wrongly] appear conserved… This leads to a high false positive error..."

- SIFT therefore returns a **conservation score** to indicate the diversity of sequences used in the alignment

- Using predictions with median conservation >3.25 is discouraged

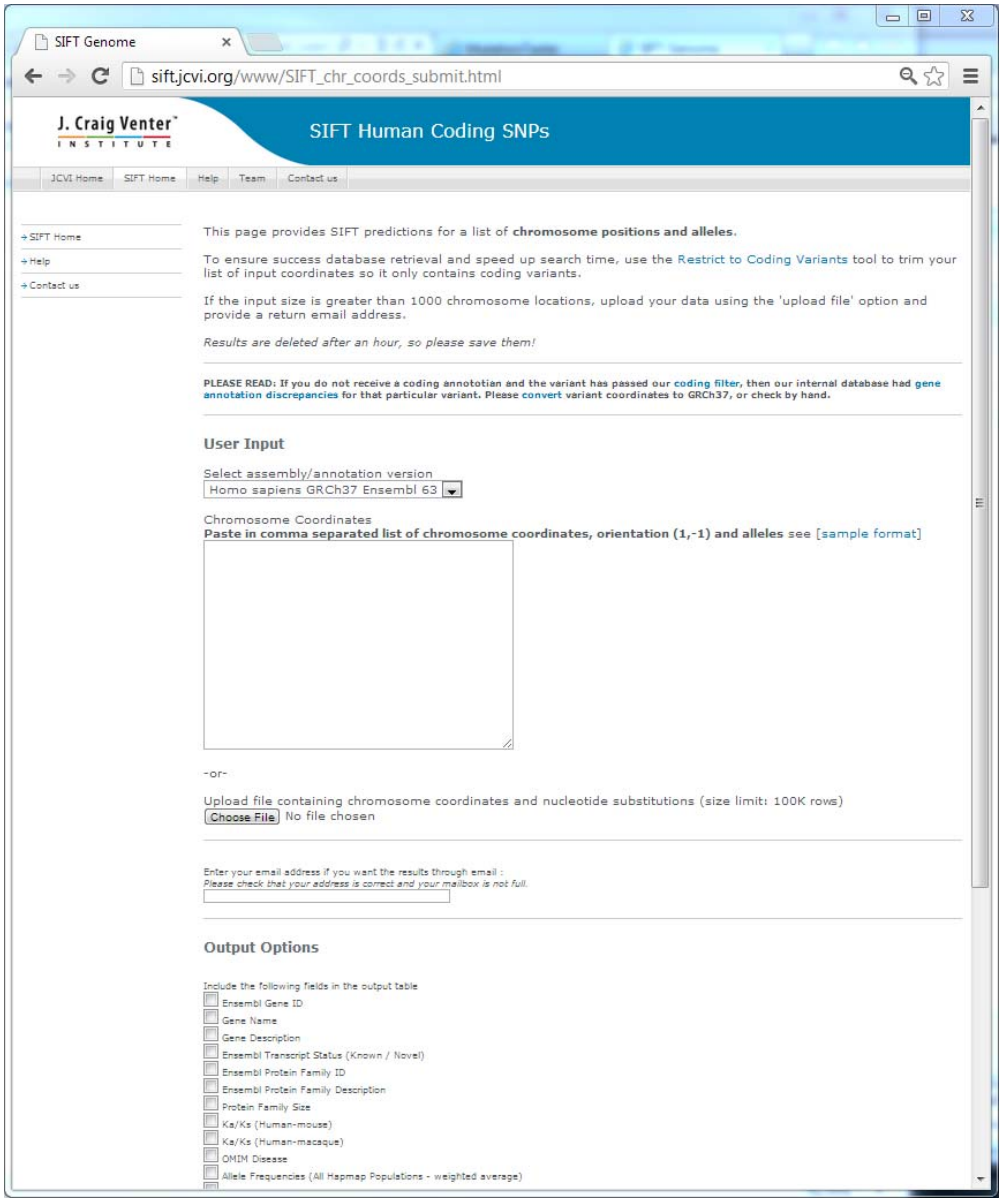


GOLDEN HELIX
Accelerating the Quest for Significance™

Nucl. Acids Res. (2003) 31 (13): 3812-3814

Using SIFT

- Web interface for making queries at *sift.jcvi.org*
- Classify amino acid substitutions, SNPs, or indels
- Can run **interactively or via batch** upload (maximum 100k variants)
- Requires simple text format for describing variants
- **Extensive annotations** provided with output
- Output returned in html or downloadable text table



PolyPhen-2

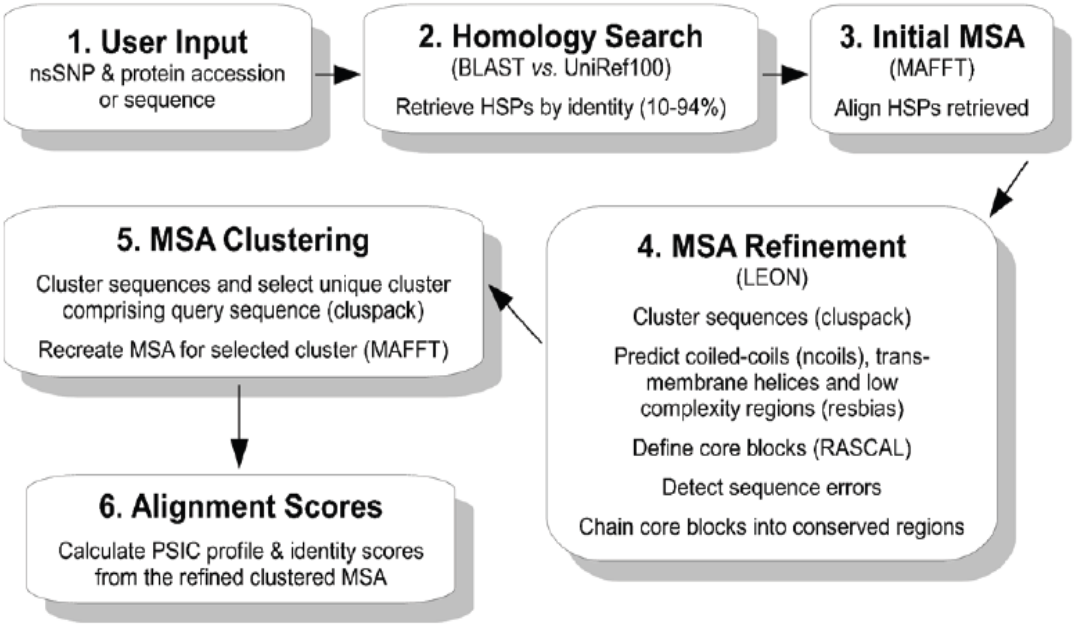
- **Polymorphism Phenotyping v2**
  (*genetics.bwh.harvard.edu/pph2*)

- "PolyPhen-2 is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations."

- **Publication:**
  - A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248 - 249 (2010)
    - Cited by 1058 on Google Scholar

PolyPhen-2: How It Works

- PolyPhen-2 is a trained algorithm that uses a naive **Bayes classifier to score variants** based on 11 predictive features.
- The most informative predictive features characterize:
  - How likely the two human alleles (WT/alt) are to **occupy the site given the pattern of AA replacements** in the MSA (aka PSIC score [Sunyaev et al, 1999])
  - How **distant** the protein harboring the first deviation from the human wild-type allele is from the human protein
  - Whether the mutant allele originated at a **hypermutable** site



Nature Methods 7, 248 – 249 (2010) [Suppl]

Features in the PolyPhen-2 Prediction Model

"We have found that including both orthologs and paralogs of the analyzed sequence in MSA leads to more accurate predictions, perhaps because a majority of disease-causing replacements affect protein structure, rather than specific aspects of function"

Eight Sequence Features:

- PSIC score of the wild-type AA
- Difference in PSIC score between wild-type and alternate AA
- Sequence identity to the closest homolog carrying any mutant AA
- Congruency of the mutant allele to the multiple alignment
- CpG context of transition mutations
- Alignment depth at mutation site
- Change in amino acid volume
- Whether mutation site is in an annotated Pfam domain

Three Structural Features (for proteins with known 3D structures):

- Accessible surface area of the wild-type residue
- Change in hydrophobic propensity
- Crystallographic β-factor reflecting conformational mobility of wild-type residue

Two PolyPhen-2: Two Prediction Models

PolyPhen-2 calculates two unique predictions. Both use the same basic methods, but the predictions are trained with different training datasets.

HumVar

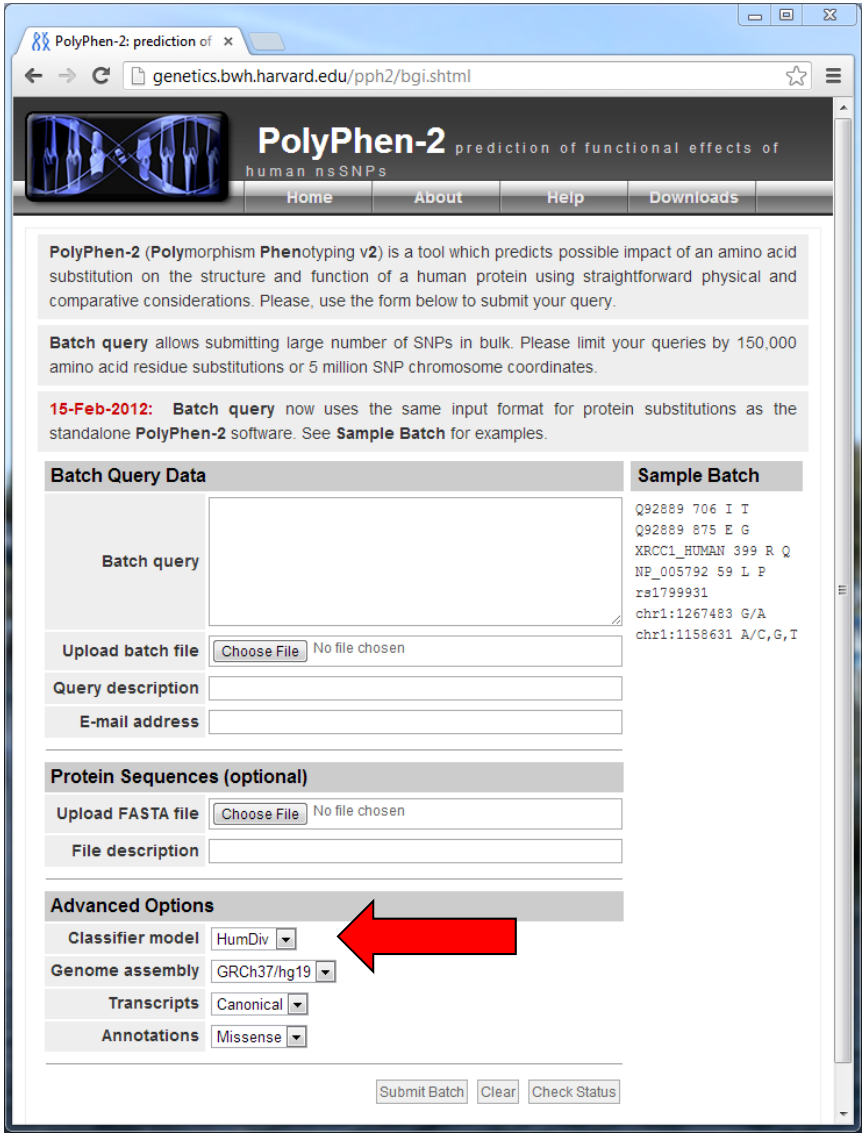
- Trained on all 13,032 human disease-causing mutations from UniProt and 8,946 human nsSNPs without annotated involvement in disease
- "Non-damaging" set includes a sizable fraction of mildly deleterious alleles. HumVar is tuned to detect drastic effects and is best used in analysis of **Mendelian** traits

HumDiv

- Trained on all 3,155 damaging alleles annotated in UniProt as causing human Mendelian diseases and affecting protein stability or function, together with 6,321 differences between human proteins and closely related mammalian homologs, assumed to be nondamaging and close to selective neutrality
- Should be used to evaluate rare alleles at loci potentially involved in **complex disease**. HumDiv is likely to classify even mildly deleterious alleles as damaging

Using PolyPhen-2

- Web interface for making queries at *genetics.bwh.harvard.edu/pph2*
- Classify amino acid substitutions or SNPs
- Requires simple text format for describing variants
- Can run interactively or via batch upload
- Standalone software may be downloaded and installed locally
- **Watch Out:** Documentation and user guides for both the web app and standalone program are incomplete.

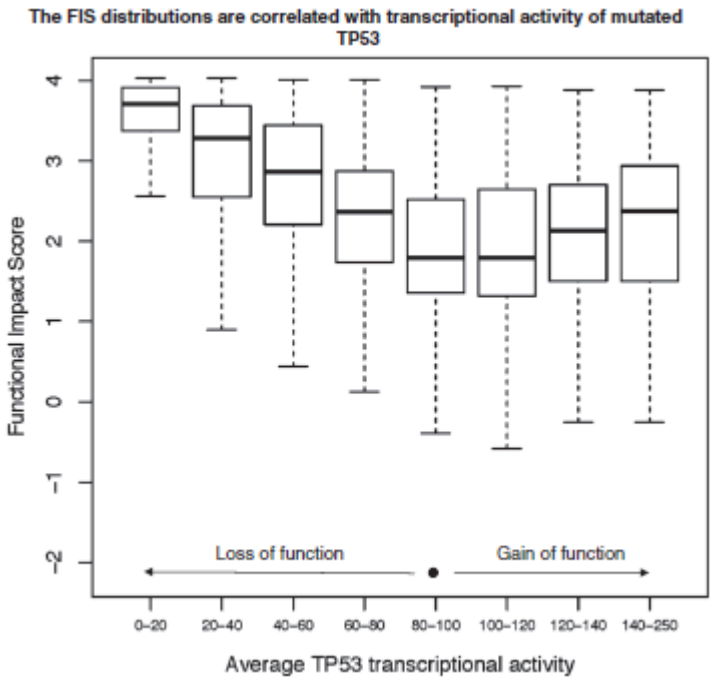


MutationAssessor

- MutationAssessor (*mutationassessor.org*)
- "The server predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The functional impact is assessed based on evolutionary conservation of the affected amino acid in protein homologs."
- "We use this rich evolutionary information for the prediction of the functional impact of mutations in general and in cancer in particular."
- **Publications:**
  - Method and server white paper:
    - **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucl. Acids Res*. 39. (2011)
    - 57 citations
  - Original method paper:
    - **Determinants of protein function revealed by combinatorial entropy optimization.** *Genome Biology* 8, R232. (2007)
      - 62 citations

About MutationAssessor

- Unique in that it was designed with special consideration for evaluating somatic variants in cancer
- Authors are careful in selection of terminology: **refer to variants as "functional" rather than "damaging" or "disease causing"**
- MutationAssessor concept is to capture variants with various consequences:
  - Loss of function
  - gain of function
  - drug resistance
  - switch-of-function



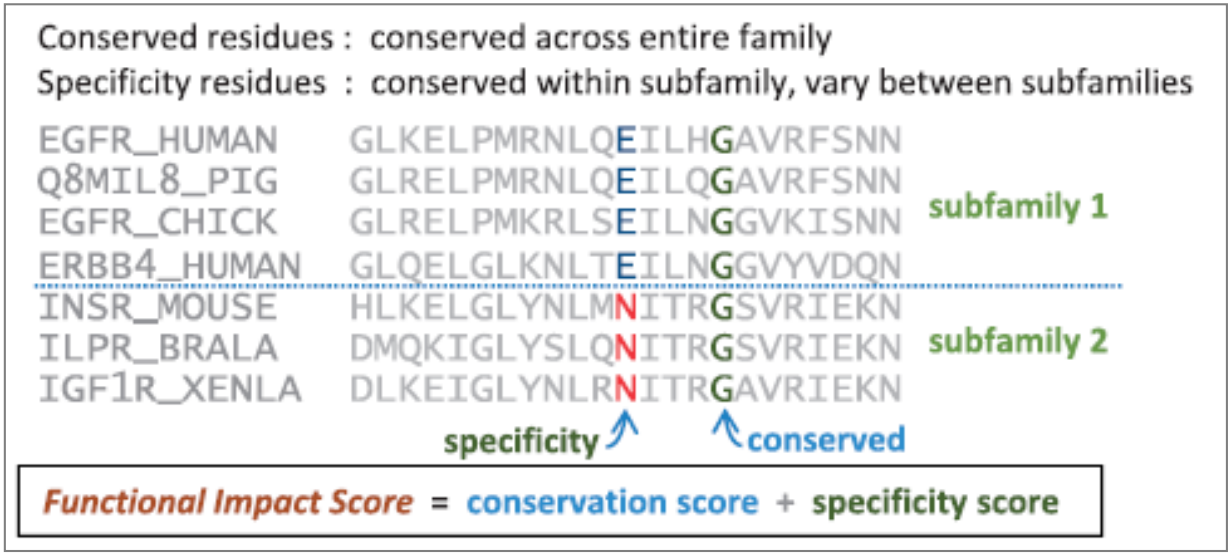
Nucl. Acids Res. 39 (2011)

MutationAssessor: How It Works

- **Uses multiple sequence alignments together with known 3D structures of sequence homologs**
  - 3D structures are annotated in output, but aren't part of the functional impact score.
- Stands out from other methods in the use of **protein subfamilies**

- Calculates two scores for each AA substitution:

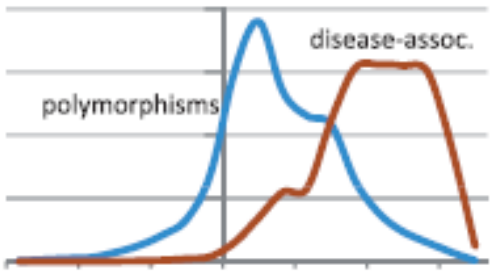
1. **Conservation** (across entire protein family)
2. **Specificity** (conserved within subfamily, but not conserved in entire family)



Nucl. Acids Res. 39 (2011)

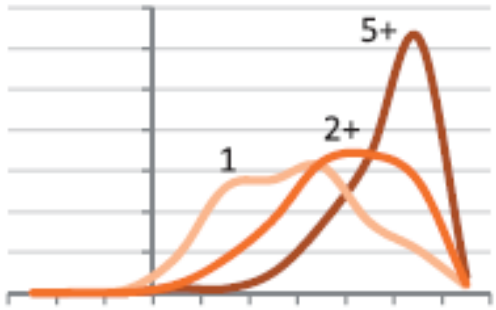
MutationAssessor: Schematic

- **Functional Impact Score** is the sum of the conservation and specificity scores
- "The specificity residues are predominantly located on protein surfaces in known or predicted binding interfaces and often directly linked to protein functional interactions."



Functional impact: disease or neutral?

80% classification accuracy in separation of
36K common polymorphisms (assumed neutral) from
19K disease-associated variants (assumed functional)
AUC = 0.86



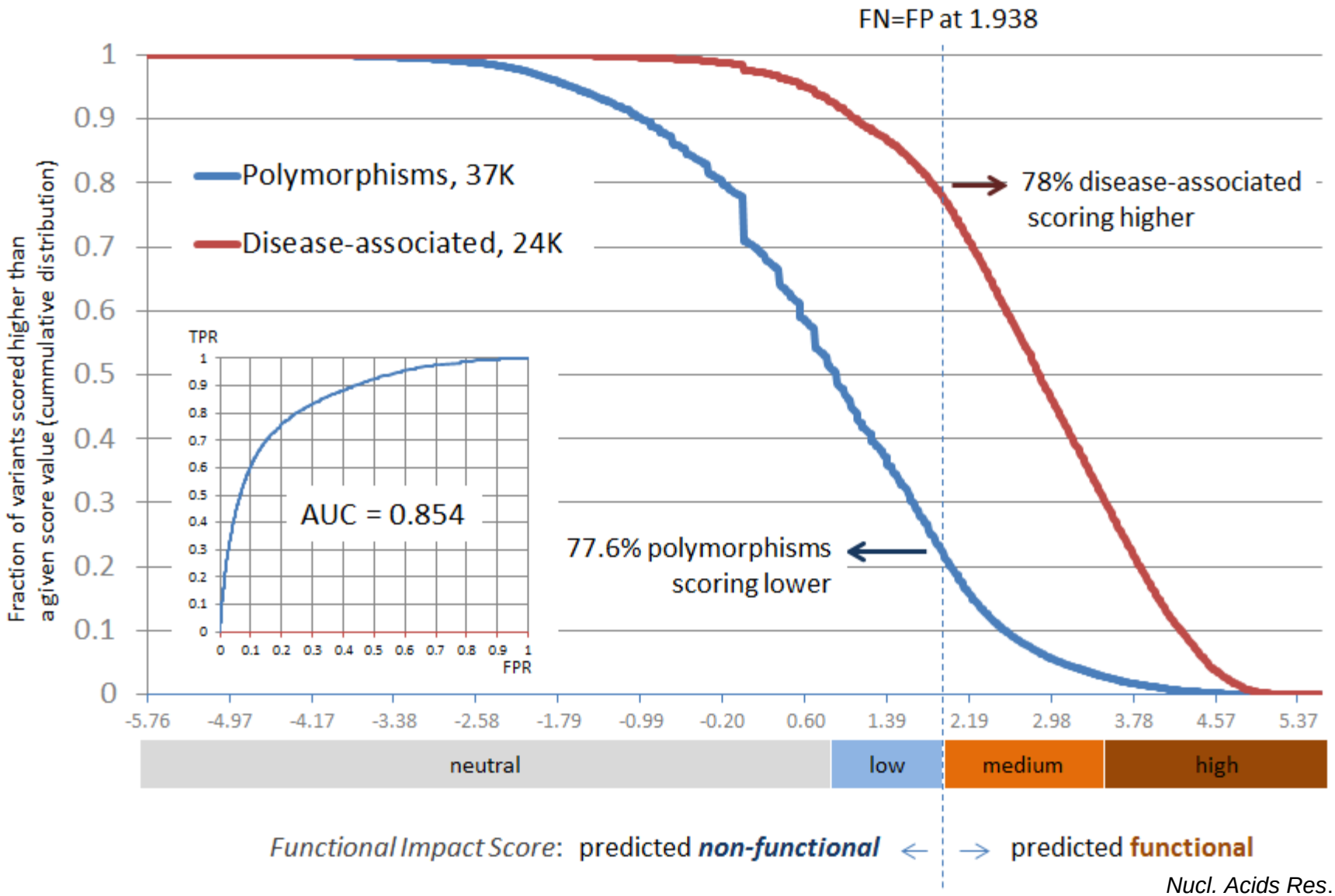
Functional impact in cancer : stronger or weaker?

10K point non-synonymous mutations in COSMIC.v49 :

- non-recurrent (observed in only one sample) vs recurrent (observed in 2 or more samples) : 69% classification accuracy, AUC=0.75
- non-recurrent vs highly recurrent (observed in 5 or more samples) : 78% classification accuracy, AUC=0.84

Nucl. Acids Res. 39 (2011)

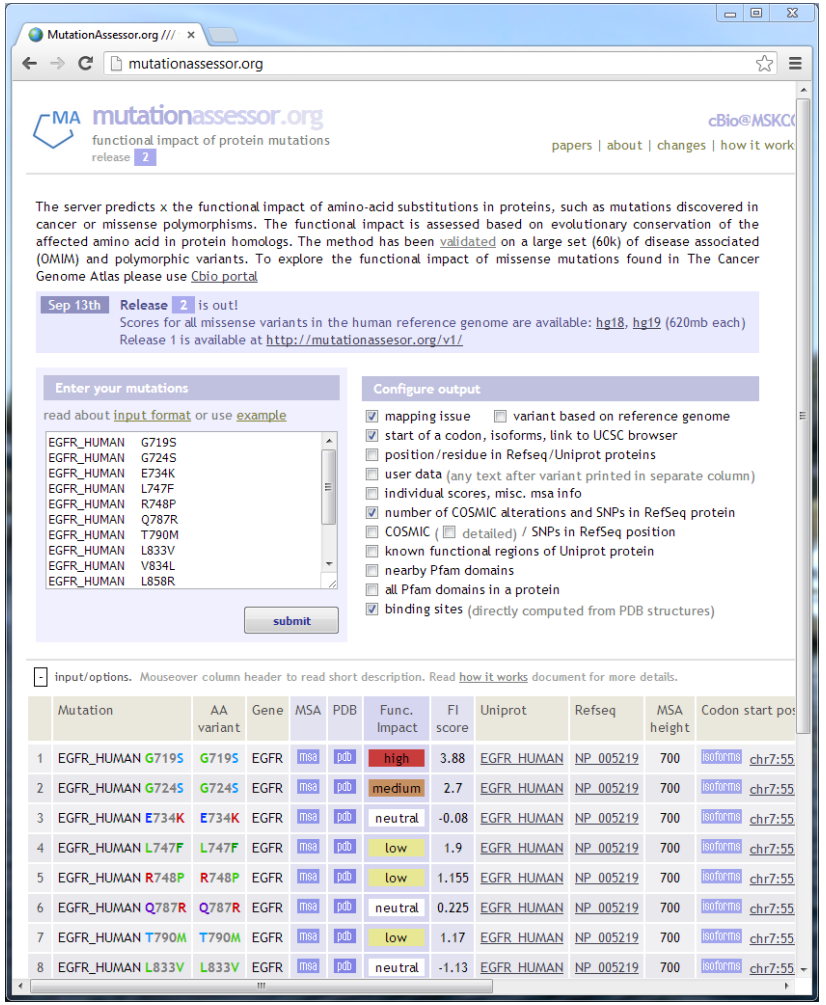
MutationAssessor: Validation



Nucl. Acids Res. 39 (2011)

Using MutationAssessor

- Web query interface at *mutationassessor.org*
- **Classifies amino acid substitutions only**
  - Also allows submission of variants using DNA coordinates and base changes
  - Mutations classified as neutral, low, medium, or high functionality.
- Best run interactively, has option for batch upload via WEBAPI
- **Extensive output** including domain annotations and options to display MSA and 3D structure of most similar protein with known tertiary structure
- **Simple by powerful user interface.** Let's take a look at it…



MutationTaster

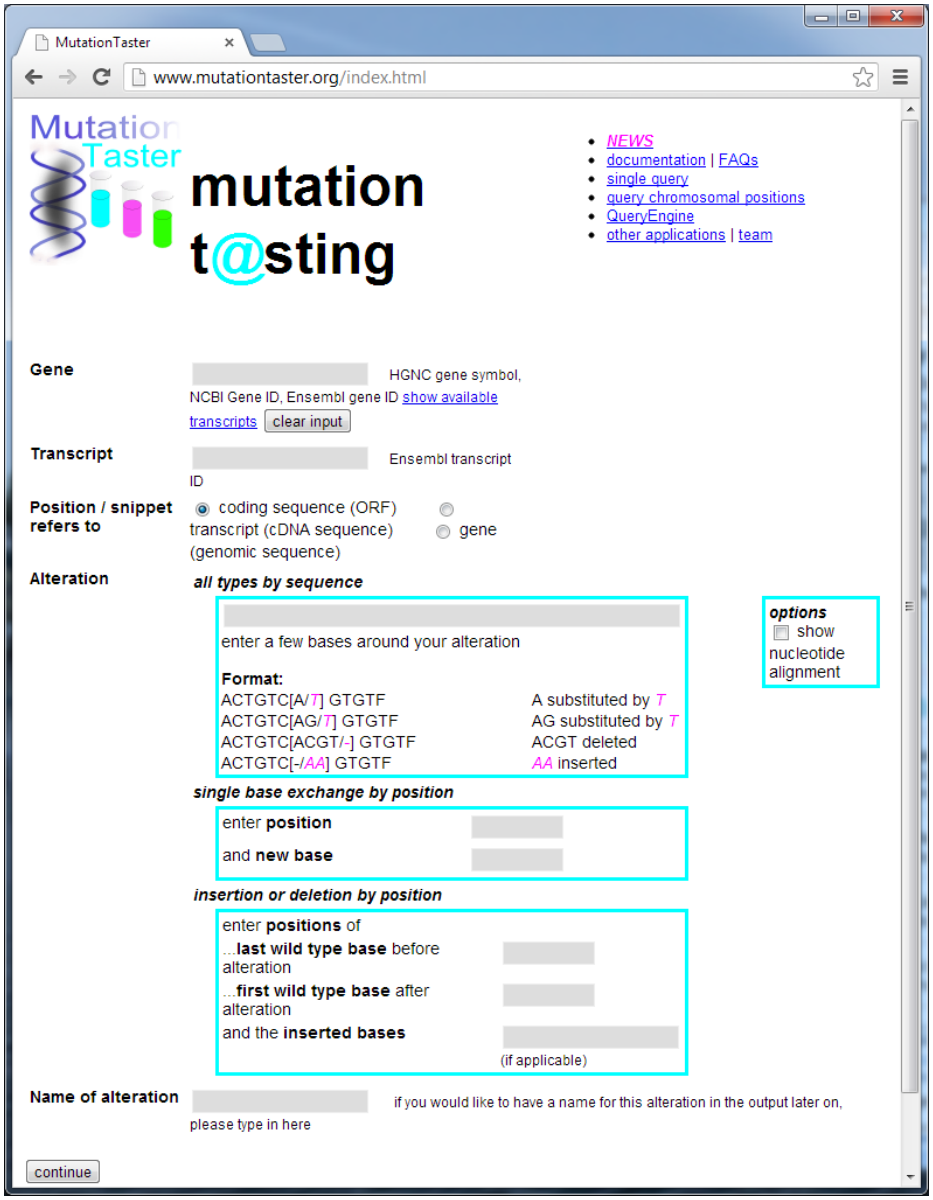
- MutationTaster (*mutationtaster.org*)

- "MutationTaster integrates information from different biomedical databases and uses established analysis tools. Analyses comprise evolutionary conservation, splice-site changes, loss of protein features, and changes that might affect the amount of mRNA. Test results are then evaluated by a naïve Bayes classifier, which predicts the disease potential."

- **Publication:**
  - MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* 7, 575–576 (2010)
    - Cited by 199

About MutationTaster

- **Trained classifier**
  - Trained on over 50,000 disease mutations and 520,000 common polymorphisms gathered from various sources
- **One of the few prediction tools with native support for DNA alterations rather than AA substitutions**
  - Annotates indels and non-coding regions in addition to protein-coding SNPs
  - Has option to combine adjacent mutations into a complex substitution polymorphism to determine the true amino acid change
- **Web interface allows for upload and annotation of VCF files**
  - Limited to single-sample VCFs
  - Seems very popular. Over 6,200 “very large jobs” were in queue on April 29



More about MutationTaster

- **Uses three different annotation methods** depending on the type of mutation:
  - Alterations that don't affect AA sequence (intronic and intergenic SNVs, indels and substitutions.
  - Alterations that affect a single AA position (SNVs or substitutions)
  - Alterations that affect multiple AA positions (Frameshifts)
- The classifier is trained on a different set of predictors for each type
- Output includes **extensive annotations for coding and non-coding regions**
  - Alterations of Kozak consensus sequence
  - Propensity to affect splice sites (based on 3rd party program "NNSplice")
  - dbSNP, 1kG, ClinVar, HGMD annotations
  - Various regulatory features; both AA and DNA conservation values
- **Caution: Has some quirks.** But ease of use and breadth of application for DNA are attractive

FATHMM

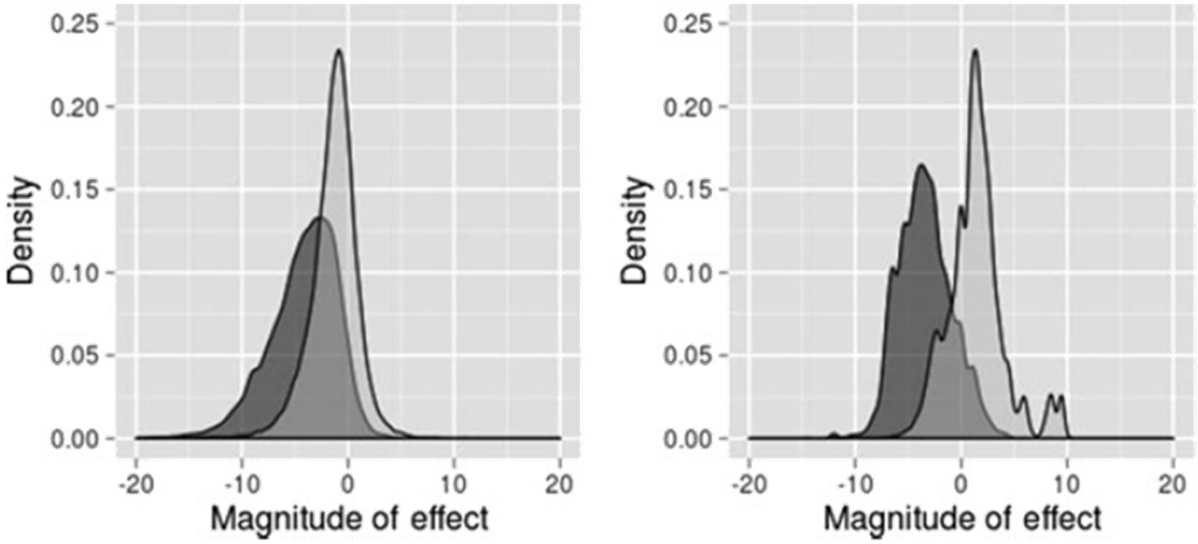
- Functional Analysis through Hidden Markov Models (*fathmm.biocompute.org.uk*)

- "A high-throughput web-server capable of predicting the functional, molecular and phenotypic consequences of protein missense variants using hidden Markov models (HMMs) representing the alignment of homologous sequences and conserved protein domains."

- **Publications:**
  - Predicting the Functional, Molecular and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.*, 34, 57-65 (2013)
  - Predicting the Functional Consequences of Cancer-Associated Amino Acid Substitutions. (Submitted)

About FATHMM

- The authors argue that **MSAs based on hidden Markov models** (HMMs) are **inherently superior** to alignments from BLAST and related methods
- The standard untrained version of FATHMM uses HMM methodology to construct the MSA that is used to assess conservation of AA residues
- FATHMM also queries manually curated HMMs representing the alignment of conserved protein domain families (SUPERFAMILY and Pfam)
- A **species-specific version** incorporates "pathogenicity weights"
  - Derived from the relative frequency of disease associated and functionally neutral sequences mapping onto conserved protein domains

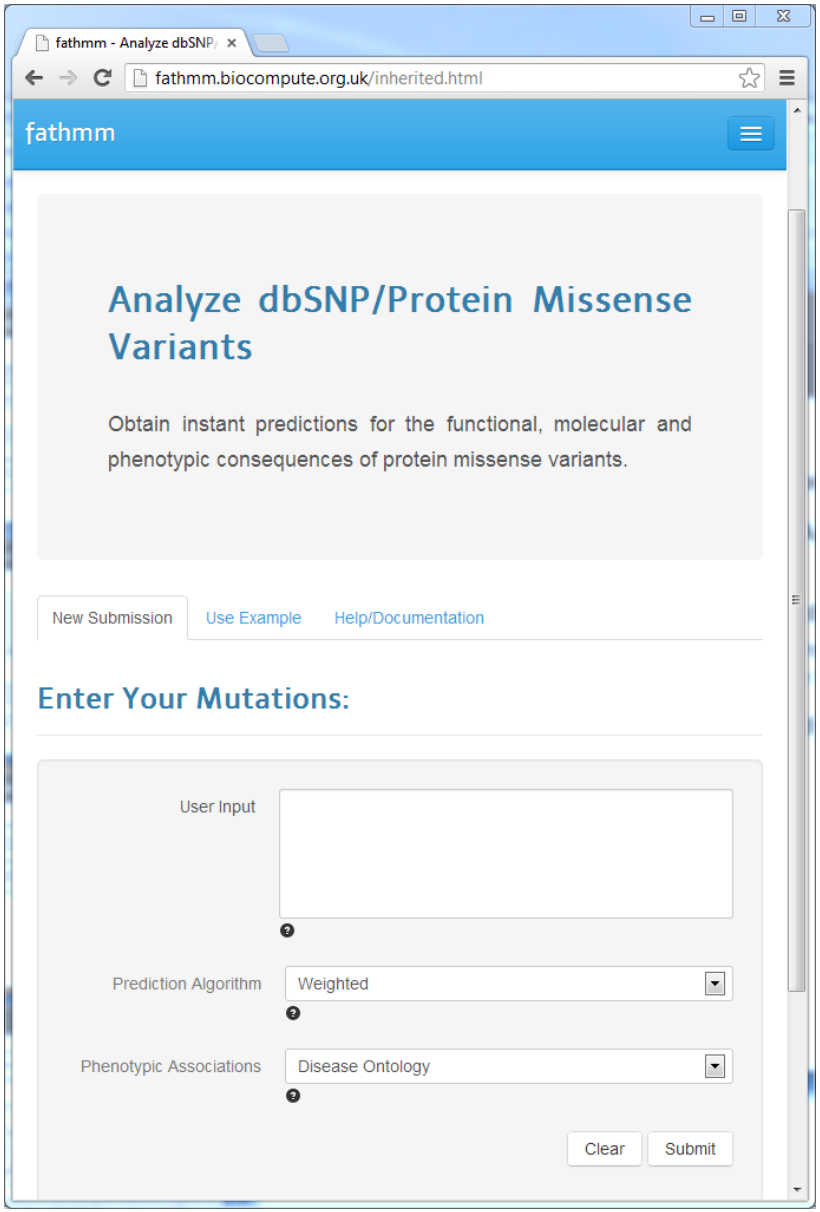


GOLDEN HELIX
Accelerating the Quest for Significance™

Hum. Mutat. **34**, 57-65 (2013)

Using FATHMM

- Web portal at *fathmm.biocompute.org.uk*
- Submit variants **based on AA substitution or by rsID**. No support for other DNA-based formats
- Output returned in html or downloadable text table
  - Output may include optional annotations from Human Phenotype Ontology, Gene Ontology, Disease Ontology or other sources
- **Application can be installed and run locally**
- Cancer-specific version also available, but still unpublished

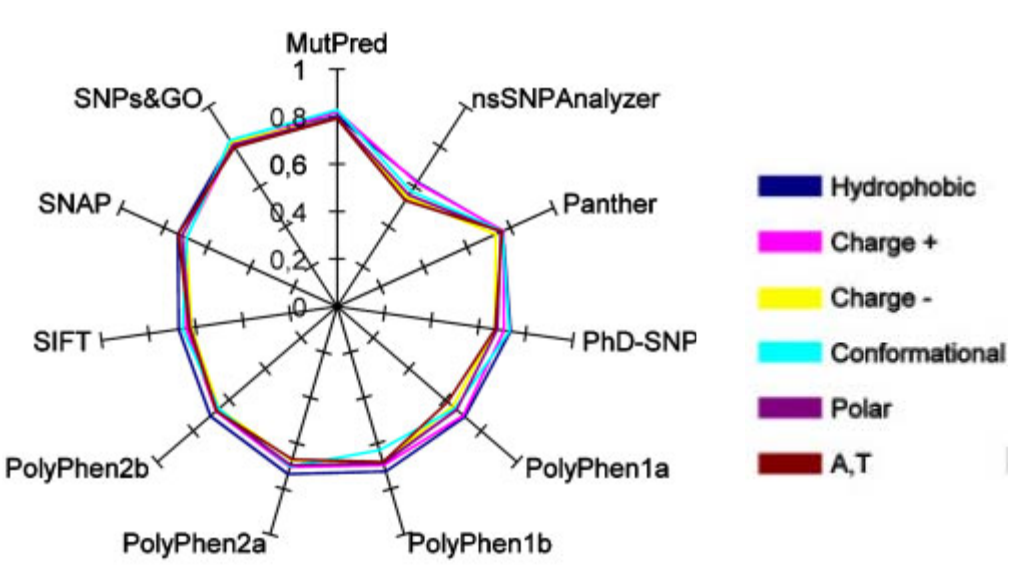


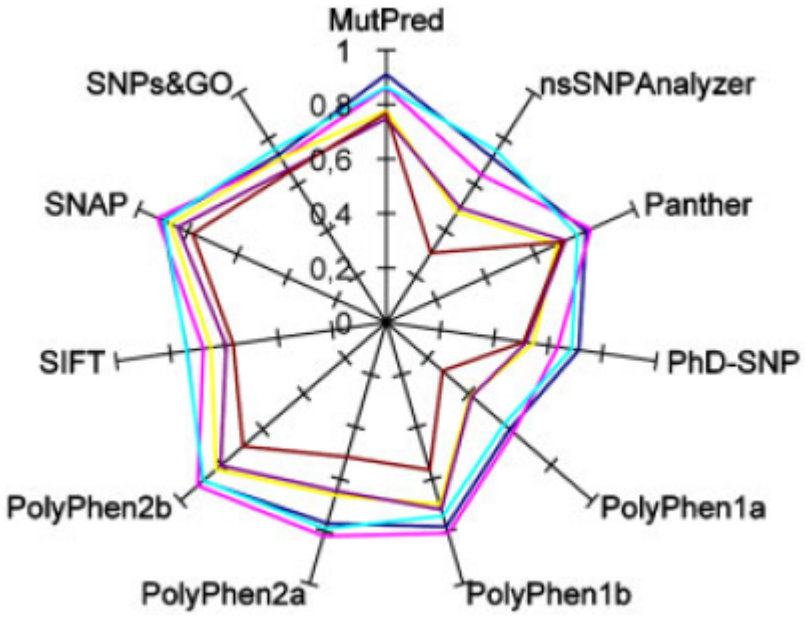
Agenda

1. The Basics of Molecular Biology & Functional Predictions
2. Overview of Commonly Used Algorithms
3. **Comparisons**
4. Applying Functional Predictions

What now?

- **How do we know if they really work? What should I use?**
- There are several published comparisons based on various standards
- These comparisons serve as a starting point to understand the differences in methods





Accuracy and Sensitivity for different types of AA substitutions. (Thusberg et al., *Human Mutation*, 2010)

Published Comparisons

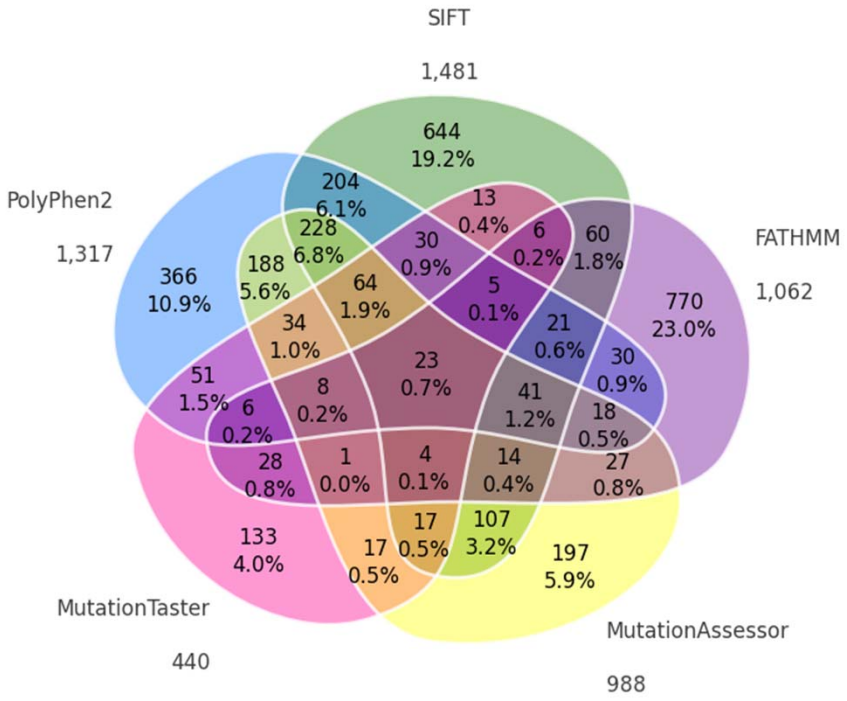
Table 2. Performance of Computational Prediction Methods using the VariBench Benchmarking Dataset

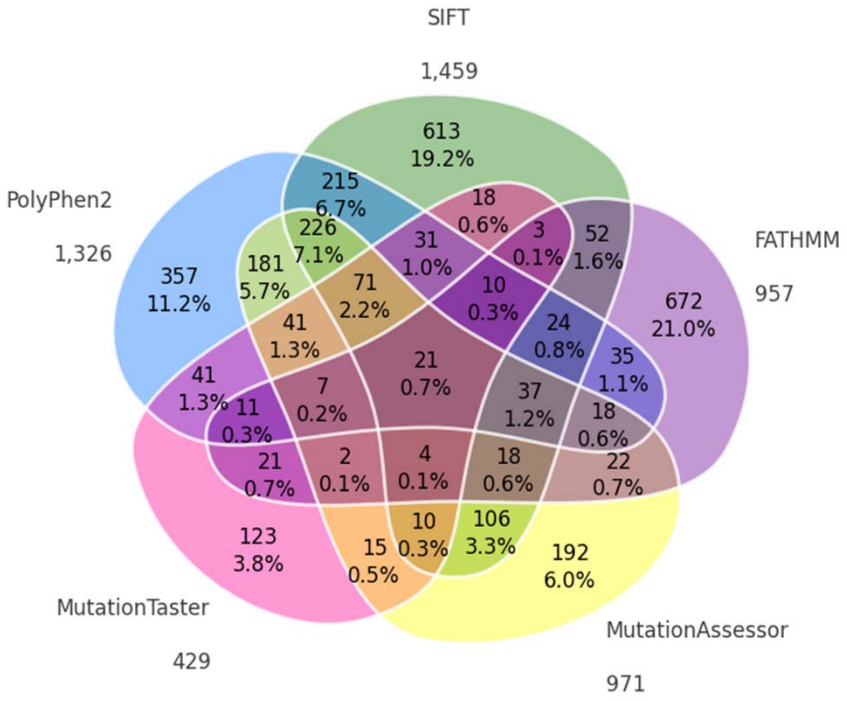
	tp	*fp*	*tn*	*fn*	Accuracy[a]	Precision[a]	Specificity[a]	Sensitivity[a]	NVP[a]	MCC[a]
Theoretical/unweighted computational prediction methods										
SIFT	10,464	4,856	12,188	7,433	0.65	0.64	0.62	0.68	0.66	0.30
PolyPhen 1[b]	10,093	9,185	17,669	3,199	0.69	**0.77**	**0.85**	0.52	0.64	0.39
PolyPhen 1[c]	14,285	4,993	13,671	7,197	0.70	0.68	0.66	0.74	0.72	0.40
PANTHER	9,689	2,859	8,676	2,797	**0.76**	0.76	0.76	**0.77**	**0.77**	**0.53**
FATHMM (unweighted)	11,561	4,839	16,257	7,707	0.69	0.72	0.77	0.60	0.66	0.38
Trained/weighted computational prediction methods										
PolyPhen 2[b]	13,807	5,102	13,863	6,010	0.71	0.71	0.70	0.73	0.72	0.43
PolyPhen 2[c]	16,206	2,703	10,199	9,674	0.69	0.64	0.51	0.86	0.78	0.39
PhD-SNP	11,900	6,896	16,788	4,377	0.71	0.75	0.79	0.63	0.68	0.43
SNPs&GO	13,736	5,487	17,028	1,382	0.82	**0.90**	**0.92**	0.71	0.76	0.65
nsSNPAnalyzer	4,360	2,778	1,319	943	0.60	0.59	0.58	0.61	0.60	0.19
SNAP	16,000	2,146	8,190	6,387	0.72	0.67	0.56	**0.88**	0.83	0.47
MutPred	13,829	2,507	15,891	4,557	0.81	0.79	0.78	0.85	0.84	0.63
FATHMM (weighted)	14,231	1,633	10,146	2,336	**0.86**	0.86	0.86	0.86	**0.86**	**0.72**

- Published comparisons have generally similar findings:
  - **Most algorithms are 65% - 80% accurate** when comparing known disease mutations to neutral mutations, with reasonable ROC curves
- The problem is that in practice, there are many variants with uncertain consequences, and this gray area is where interpretation is especially difficult
- **Most algorithms will predict 10%-20%** of all nsSNPs to be damaging

Hum. Mutat., **34**, 57-65 (2013)

Classifying all nsSNPs in a Sample

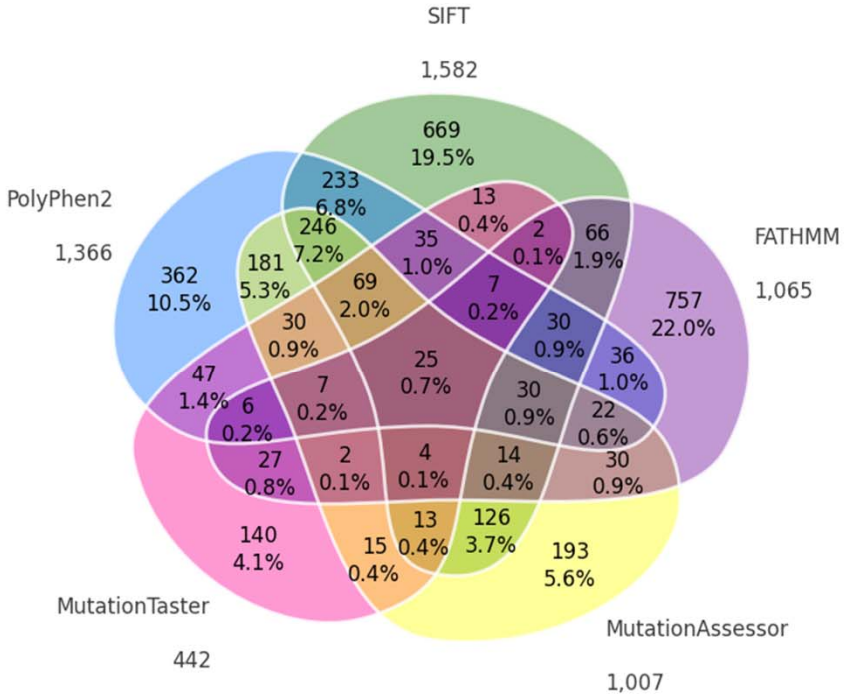


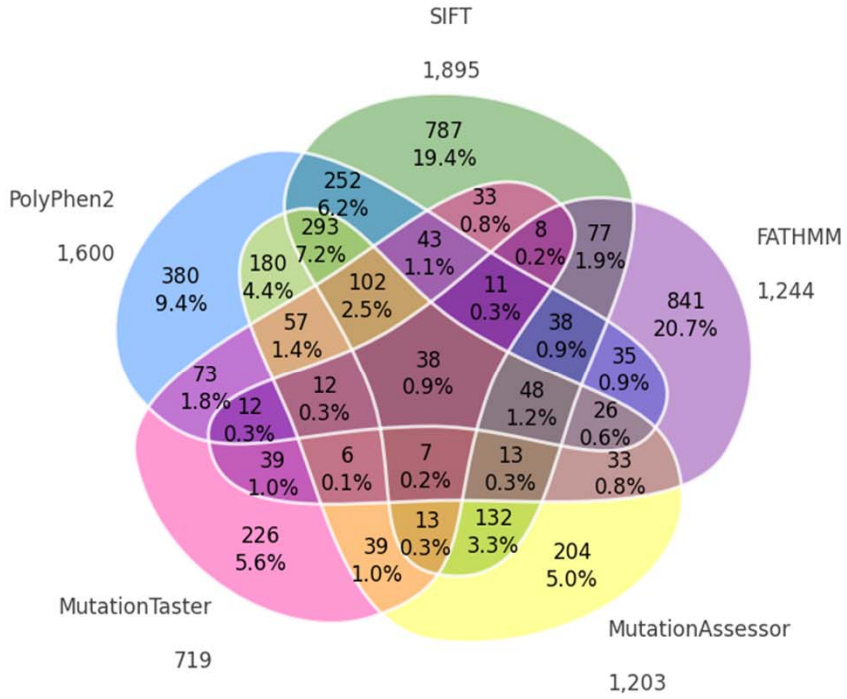


- **NA12878 – CEU**
  - 10,366 total nsSNPs
  - 3355 (32%) called damaging by at least one method
  - 23 (0.22%) called damaging by all 5

- **HG00733 – PUR**
  - 9566 total nsSNPs
  - 3197 (33%) called damaging by at least one method
  - 21 (0.22%) called damaging by all 5

Classifying all nsSNPs in 2 more Samples





- **NA18526 – CHB**
  - 10,407 total nsSNPs
  - 3437 (33%) called damaging by at least one method
  - 25 (0.24%) called damaging by all 5

- **NA19240 – YRI**
  - 11,661 total nsSNPs
  - 4058 (35%) called damaging by at least one method
  - 38 (0.33%) called damaging by all 5

How Many Damaging SNVs per Sample?

		Number of algorithms calling each SNP Damaging/Functional					
Sample	**# nsSNPs**	**0**	**1**	**2**	**3**	**4**	**5**
NA12878 (CEU)	10,366	7011	2110	725	375	122	23
HG00733 (PUR)	9566	6369	1957	706	384	129	21
NA18526 (CHB)	10,407	6970	2121	774	400	117	25
NA19240 (YRI)	11,661	7603	2438	893	509	180	38

- **Of the 23 SNPs that are universally predicted damaging in NA12878:**
  - 13 are in 1000 Genomes Project, 11 have allele frequencies ≥1% in Europeans
  - 15 are in the NHLBI ESP data, 8 have allele frequencies ≥1% in Europeans
- **The YRI sample has 15% more nsSNPs, but 65% more called damaging by all 5 methods**
  - African genomes are very diverse
    - Human reference genome is biased toward European alleles, & protein sequences used in MSAs for prediction are likely to be similarly biased

Which One Should I Use?

- **Common belief is that variants called damaging by multiple algorithms are most likely to have true disease causing potential**
- **Published comparisons aren't exhaustive**, and usually focus on prediction performance for detecting a particular category of mutations
- Each prediction tool has its own strengths and weaknesses, and may carry **certain biases** based on the authors' own research interests
- All of the algorithms generally **perform well** for distinguishing between known damaging variants and known neutral variants
- **False positive rate can be high** when the methods are applied to a broad range of variants of unknown significance.
  - Difficult to quantify this
  - Numerous (most?) nsSNPs have functional consequences, but may not cause disease

Anecdotal Experience

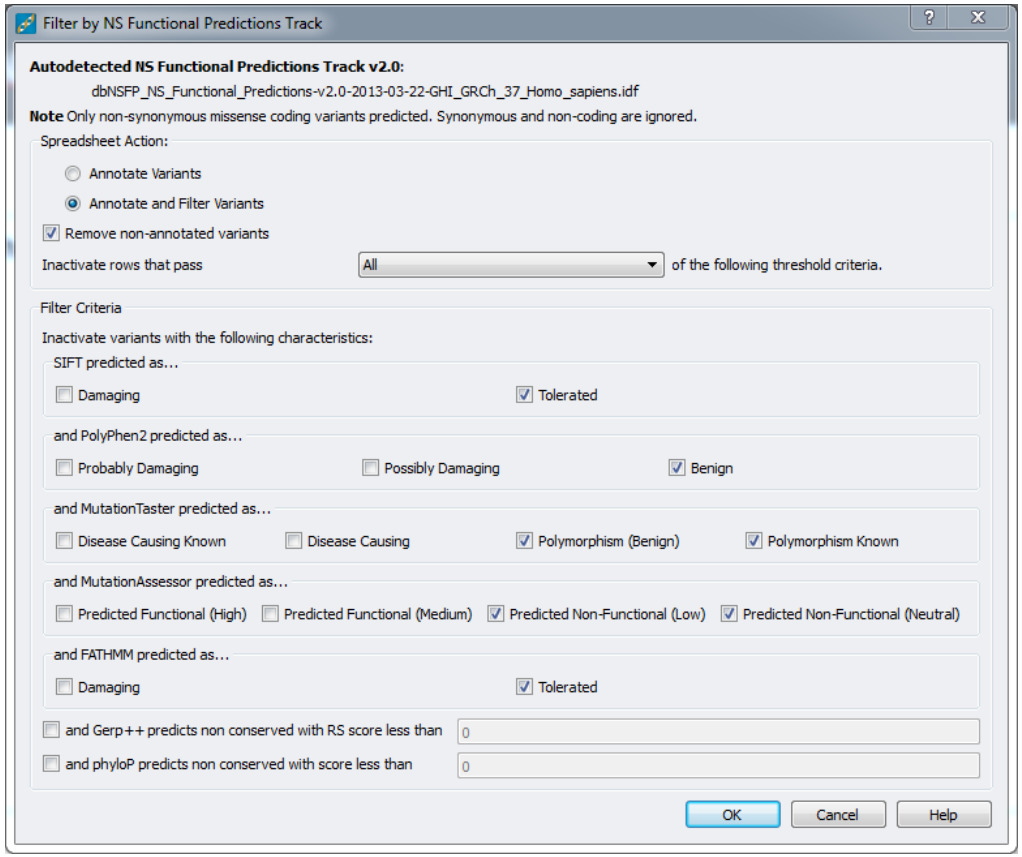
- **Algorithms consider many factors, and it's difficult to identify an obvious reason for most discrepancies.**
- I reviewed several variants called **damaging by SIFT and PolyPhen2**, but called **neutral by MutationAssessor**
  - When submitted to the MutationAssessor website, many of these variants had very low depth in the MSA (1-7 sequences)
  - It seems that MutationAssessor errs toward neutral when there is little data.
- Similarly reviewed several variants called **damaging by PolyPhen2 and MutationAssessor**, but called **tolerated by SIFT**.
  - Sites were generally highly conserved, and SIFT scores trended low (0.08-0.2)
  - Reference and alternate AA usually had similar chemical properties.
  - SIFT may be more sensitive to chemical similarity than the others.

Agenda

1. The Basics of Molecular Biology & Functional Predictions
2. Overview of Commonly Used Algorithms
3. Comparisons
4. **Applying Functional Predictions**

dbNSFP SVS Integration

- Golden Helix **SNP & Variation Suite** allows users to annotate and filter nsSNPs based on functional predictions from dbNSFP
- Users can filter SNVs based on **any or all of the algorithms** described today
- dbNSFP prediction data can also be **viewed interactively in GenomeBrowse**



GOLDEN HELIX
SNP & VARIATION SUITE 7

[Using dbNSFP in SVS]

Further Reading

The following papers were very helpful in preparing this presentation:

- **“Predicting the Effects of Amino Acid Substitutions on Protein Function” by Ng and Henikoff**
  - Annu. Rev. Genomics Hum. Genet. 2006. 7:61-80

- **“Performance of Mutation Pathogenicity Prediction Methods on Missense Variants,” by Thusberg, Olatubosun, and Vihinen**
  - Hum. Mut. 2011. 32(4):358-68

What topics would you be interested to learn about next?

Questions?

Use the Questions pane in your GoToWebinar window