

# Prediction of Pathogenicity of Sequence Variations

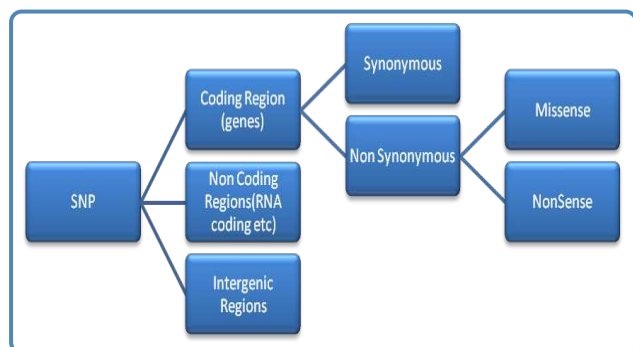
Divya Matta and Ashwin Dalal

*Diagnostics Division, Centre for DNA Fingerprinting and Diagnostics, Hyderabad*

Email: ashwindalal@gmail.com

## Introduction

The Human Genome Project revealed full landscape of the human genome at nucleotide level. This enabled us to identify differences between normal human genome (reference genome) and nucleotide sequence of patients. Sequence variations exist at defined positions within genomes and are responsible for individual phenotypic characteristics, including a person's propensity toward a disorder.



**Figure 1** Classification of SNPs.

A Single Nucleotide Polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals. If more than 1% of a population carries the same nucleotide at a specific position in the DNA sequence, this variation can be classified as an SNP. If an SNP occurs within a gene, the gene is described as having more than one allele. SNPs may lead to variations in the amino acid sequence. SNPs, however, are not always present in genes; they can also occur in non-coding regions of DNA. Types of SNPs are listed in figure 1. The synonymous SNPs are probably responsible for inter-individual phenotypic variation. On the other hand, non-synonymous variants are most likely to

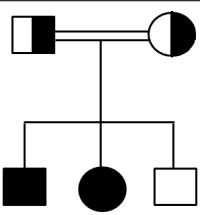
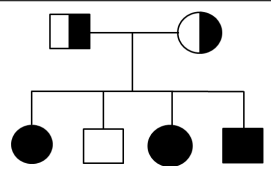
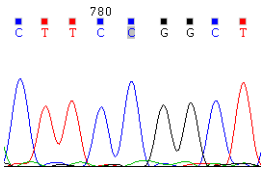
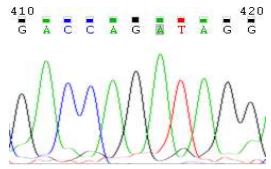
cause human disease and constitute about 90% of the mutations known to be involved in human inherited diseases.

Human genetic diseases can be caused due to de novo or inherited mutations in genes. The mutations can be detected as variants differing from the sequence in the human reference genome. However many variants may not be disease causing. It is very important to confirm whether a variation is a disease causing mutation or simply a polymorphism since many decisions like treatment, genetic counseling and prenatal diagnosis are dependent on this information.

Pathogenic changes in the nucleotide sequence usually lead to changes in the protein sequence. Protein sequences are subject to mutations in natural evolution as well as in somatic development. The direct effect of a mutation on a protein can be an effect on protein function by a number of different mechanisms. These include

- Changes in protein stability (e.g. destabilization leading to higher degradation rates, and, in the steady state, altered protein concentration).
- Change in the interaction of the protein with other biomolecules, such as other proteins, DNA, RNA or lipids, or change in the interaction with ligands, such as enzyme substrates.
- Changes in the molecular function of a protein can affect the phenotype of cells, tissues and the organism.

The importance of amino acid variation and mutations as genetic factors of human diseases has been known for many years. If the identified variation is a known disease-causing mutation then the prediction is straight forward as it is already classified as a disease-causing variant. On the other hand, if the variation is novel, then it needs to be classified as a disease-causing variant or a polymorphic SNP.

	Family I	Family II		
<b>Pedigree</b>				
<b>Sequence Chromatogram</b>				
<b>Prediction</b>	<b>Mutation</b>	<b>W393R (c.1177T&gt;C)</b>	<b>Mutation</b>	<b>459+1 G&gt;A</b>
	<i>MutationTaster</i>	Disease Causing	<i>MutationTaster</i>	Damaging
	<i>SIFT</i>	Score (0) Damaging	<i>Human Splicing Finder</i>	Broken Wild type donor. Most Probably affecting the splicing.
	<i>PolyPhen</i>	Score (1) Probably disease causing		
<b>Interpretation</b>	Variant is likely to be disease causing mutation based on these evidences: 1. Variant is homozygous in patient 2. Variant is heterozygous in both parents and normal sibling 3. Variant is not present in SNP databases and in 100 normal individuals 4. Variant is predicted to be disease causing by prediction software	Variant is likely to be disease causing mutation based on these evidences: 1. Variant is homozygous in patient 2. Variant is heterozygous in both parents and normal sibling 3. Variant is not present in SNP databases and in 100 normal individuals 4. Variant is predicted to be disease causing by prediction software		
<b>Disorder</b>	<b>Niemann-Pick Disease</b>	<b>Metachromatic Leukodystrophy</b>		
<b>Gene</b>	<b>SMPD1</b> • Located on Chromosome 11p15 • Has 6 exons • So far 116 mutations have been identified	<b>ARSA</b> • Located on Chromosome 22q13 • Has 8 exons • So far 189 mutations have been identified		

**Table 1** Illustration of the utility of various methods for assessing the pathogenic potential of a genetic variant.

The gold standard for classifying a variant as disease-causing or a polymorphism, is to conduct functional analysis. This is done by recreating the mutation in vitro and studying its effect on the function of that particular protein. Functional analysis can be done by employing cells in culture or in transgenic animals modified with specific variant genes or sequence polymorphisms of interest. The major drawback of these methods is that these procedures are laborious, expensive and time consuming and hence not feasible in routine clinical diagnostics.

In the absence of readily available functional validation methods, different approaches are used to gather evidence "for or against" the likelihood that the particular variant is disease-causing or not. In order to classify the obtained variation as mutation or polymorphism, the following strategies can be followed:

- The identified variation is screened against

Tool	Input	Output	Interpretation
<b>MutationTaster</b> <a href="http://www.mutationtaster.org/">http://www.mutationtaster.org/</a>	Sequence with specific mutation or mutation position and mutated nucleotide	Effect of Mutation	-
<b>SIFT</b> <a href="http://sift.jcvi.org/">http://sift.jcvi.org/</a>	Ensemble Protein ID and Mutation position and Mutated amino acid	Score (0-1)	Score <0.05 damaging Score > or = 0.05 tolerated
<b>PolyPhen-2</b> <a href="http://genetics.bwh.harvard.edu/pph2/">http://genetics.bwh.harvard.edu/pph2/</a>	Amino acid sequence, Wild type amino acid and mutated amino acid along with position of mutation	HumVar score (0-1) HumDiv score (0-1)	Higher the score higher the probability to cause disease
<b>HANSA</b> <a href="http://www.cdfd.org.in/HANSA/">www.cdfd.org.in/HANSA/</a>	Amino acid sequence, Mutation position, wild type and mutated amino acids	Difference of wild type and mutated amino acid	Higher the difference higher the disease causing ability

**Table 2** List of Mutation prediction software and their score interpretation.

SNP databases like dbSNP, 1000 Genome Project, Exome Variant Server etc. dbSNP is a database of known polymorphisms seen in humans. Hence if a variant is found to be described in dbSNP as a polymorphism then the variant can be classified as a non-disease causing polymorphism. Further the variant also needs to be looked for in 100 normal individuals from same ethnic population to identify whether it is a polymorphism or not.

- The identified variation is screened in mutation databases like the Human Genome Mutation Database (HGMD) etc. These databases contain a list of known mutations in a particular gene and if the variant is found in any of these databases then it can be classified as a disease-causing mutation. It is always important to refer to the original paper to see if functional analysis was done for the mutation, specifically in cases of rarely described mutations.
- Segregation analysis: Single gene diseases follow typical patterns of inheritance of variants in families. In case of an autosomal recessive disease, the parents and unaffected siblings of the affected child will be heterozygous for the mutation. This information regarding status of the same variant in family members can help us to postulate regarding disease-causing mutations. In the same

way, a severe phenotype of an autosomal dominant condition is more likely to be due to a de novo mutation (i.e. present in the proband but not the parents).

- Pathogenicity potential prediction software: Several pathogenicity prediction software have been developed to predict the likelihood of a particular variant to be disease-causing or not, i.e. MutationTaster, SIFT, PolyPhen, Human Splicing Finder (in case of splice site variations) and HANSA.<sup>1-5</sup> However it is important to note that the results of these software are only predictions and have to be interpreted in association with other information regarding the variant. The obtained variation is analyzed with prediction software. An illustration of the use of these methods is shown in Table 1.

Computer based approaches play an important role in providing reliable results in a shorter time and are very easy to handle. Several methods (Table 2 & 3) for assessing the effects of mutation on protein function and abnormal mRNA splicing patterns (resulting in exon skipping, cryptic splice site use, high levels of intron inclusion) have been developed over the years. To assess a mutational effect, such methods typically use the physicochemical properties of amino acids, as well as information about the role of amino acid side chains in protein structure. These methods

Tool	Input	Output	Interpretation
<b>NNSplice</b>	Single/multiple sequences	Score (0-1)	Higher score implies greater potential for splice site
<b>GENSCAN</b>	Single sequence $\leq$ 1 million bp	Probability score (0-1)	Higher score implies a higher probability of correct exon
<b>MaxEntScan</b>	Single/multiple sequences (5': 9 bp (-3 to +6); 3': 23 bp (-20 to +3))	Maximum entropy score (log odds ratio)	Higher score implies a higher probability of the sequence being a true splice site
<b>Human Splicing Finder</b>	Single sequence $\leq$ 5,000 bp	S & S score (0-100)	Higher score implies greater potential for splice site
<b>SROOGLE</b>	Target exon along with two flanking introns	Different scores with their percentile scores (0-1)	Higher percentile score implies a higher ranking of the splice site within pre calculated distributions

**Table 3** Summary of input, output, and interpretation of prediction scores for selected currently available *in silico* tools for 5' and 3' splice site prediction with user-friendly web interface.

combine all essential properties of both the original and substituted residues (e.g. size, polarity), structural information (e.g. surface accessibility, hydrogen bonding) and evolutionary conservation, and then are trained to distinguish between known functionally deleterious variants and presumably neutral variants. These methods assess effect of a mutation by a score computed based on a particular theoretical model. Most of these computational approaches are validated on variants with pronounced phenotypic effects, e.g. functionally deleterious and disease related variants. Such variants usually involve loss of function of a mutated gene.

Functional studies are the most accurate and reliable method for characterizing the effect of an SNP on the structure and function of the protein. However the limitation lies in the laborious, costly and time consuming procedures needed for these studies. Computer-assisted (*in silico*) technologies are considered to be efficient alternatives to *in vitro* experiments and are thought to have the poten-

tial to speed up the interpretation of pathogenic potential of variants pending functional validation.

## References

1. Acharya, V. and Nagarajaram, H.A. Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Hum Mutat* **33**, 332-7 (2012).
2. Adzhubei, I. *et al.* Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7, **20** (2013).
3. Jian, X. *et al.* In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med* **16**, 497-503 (2014).
4. Kumar, P. *et al.* Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
5. Schwarz, J.M. *et al.* MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-6 (2010).