

# Variant annotation

Irene Keller

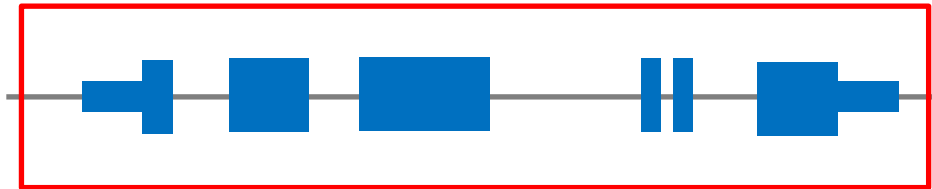
Department for Biomedical Research, University of Bern

Total number of variants ( = differences to reference genome) detected in two example datasets:

	Whole exome (WES)	Whole genome (WGS)
Total nbr variants	106'693	4'175'605

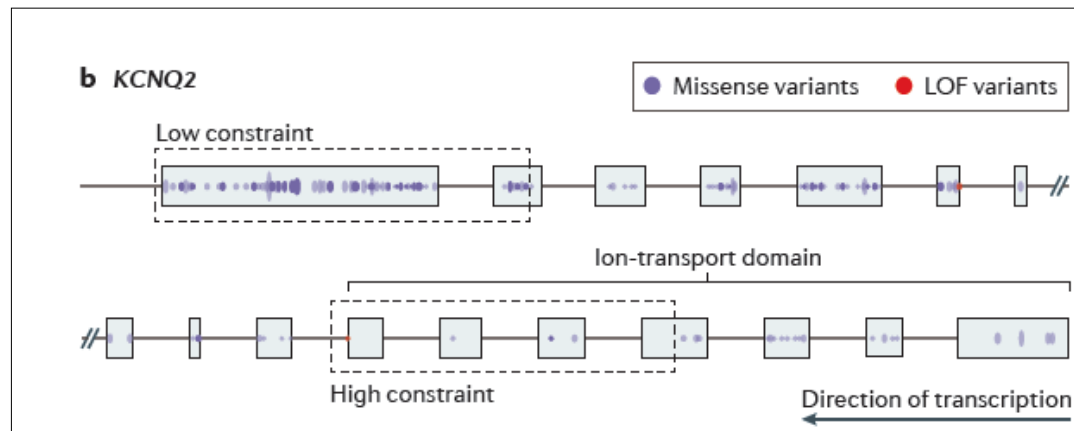


We need to prioritise these variants to be able to identify possible disease causing mutations



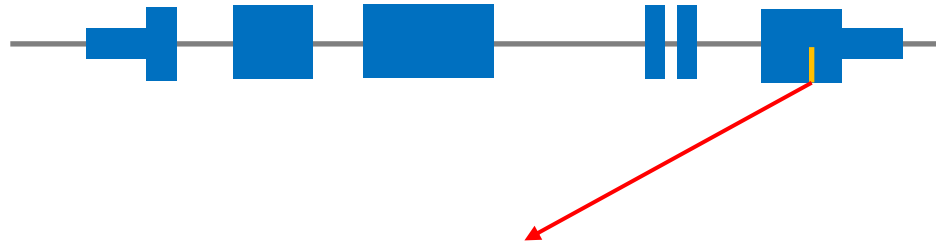
Annotation for the **entire gene**. For example:

- Selective constraint: How often is the gene mutated?

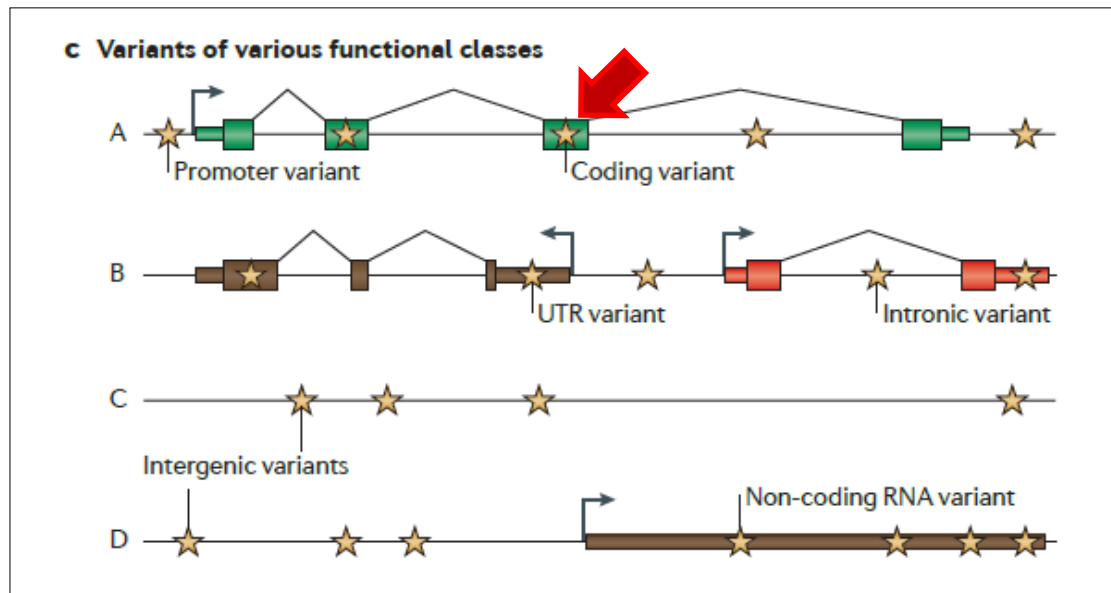


From Eilbeck et al. 2017 Nature Methods

- Is the gene known to be associated with a particular (disease) phenotype?

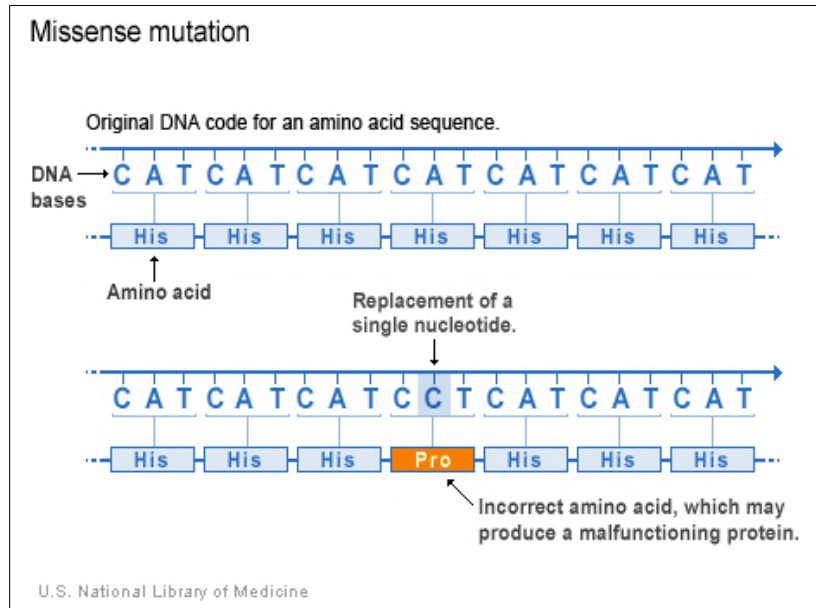


Annotation for individual variant, e.g. SNP



From Cooper & Shendure 2011 Nature Reviews Genetics

# Identify coding variants



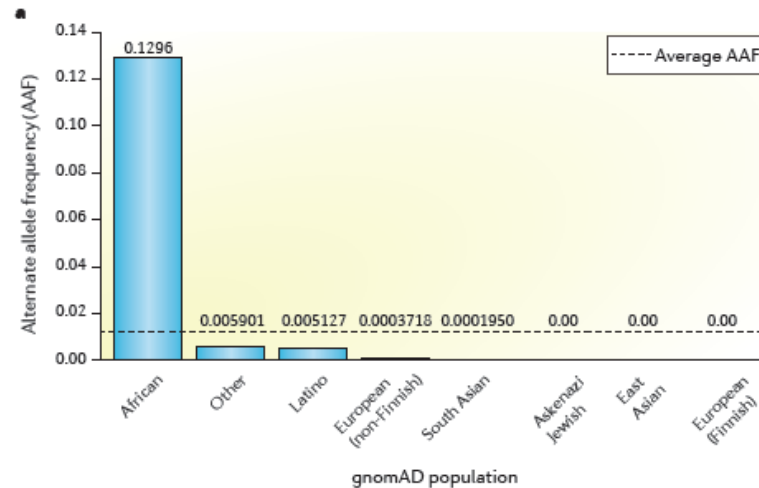
	WES	WGS
Total nbr variants	106'693	4'175'605
Coding variants	10'494	10'263

Of course, non-coding variants can also have functional effects but these are, in general, more poorly understood

## Variant ranking

Based on **allele frequency in the general (healthy) population**. Important to consider the correct reference population!

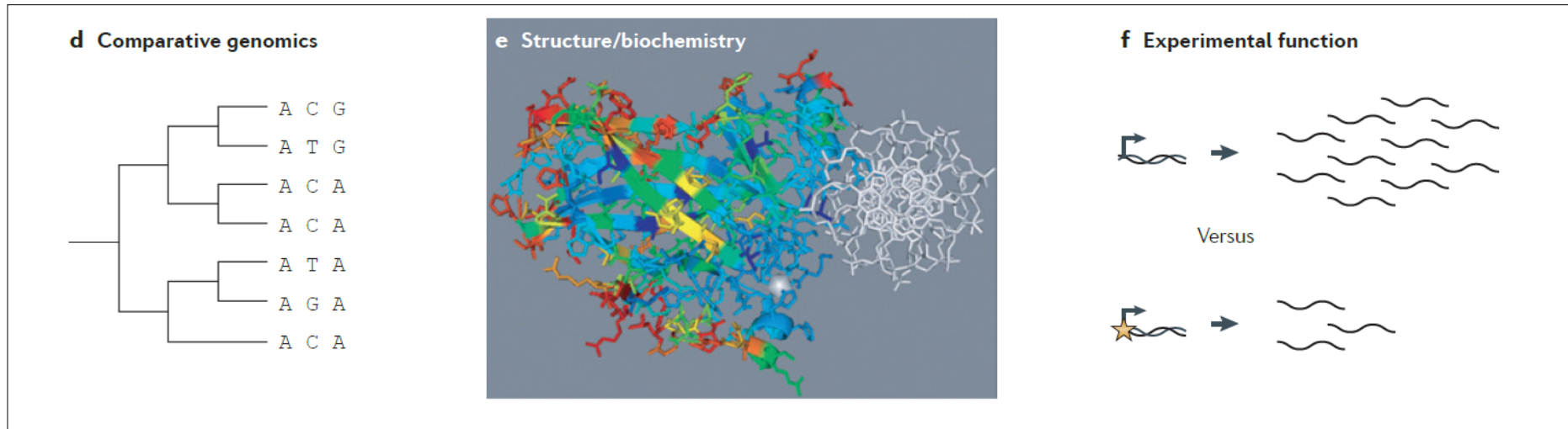
Allele frequencies for rs79444516



Eilbeck et al. 2017 Nature Methods

	WES	WGS
Total nbr variants	106'693	4'175'605
Coding variants	10'494	10'263
Coding, rare (<1%) variants	411	

Based on **effect prediction scores** (see Ritchie & Flicek 2014 Genome Medicine for a very good overview)

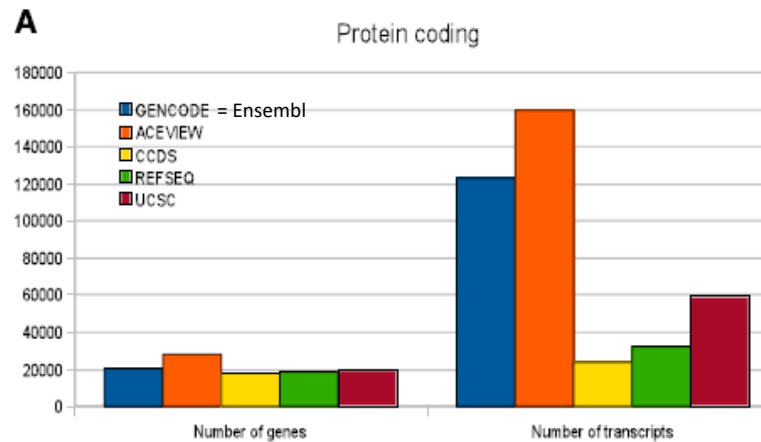


Cooper & Shendure 2011 Nature Reviews Genetics

- For the human genome, many of these scores have been precomputed:  
<https://sites.google.com/site/jpopgen/dbNSFP>
- Important: For many of these predictors, it has been shown that the average scores differ between known deleterious and putatively neutral variants but, often, there is quite some overlap and individual variants may be misassigned

## Issues/difficulties with variant annotation

- **Different annotations** for the same species, for example Ensembl, RefSeq etc



From Harrow et al. 2012

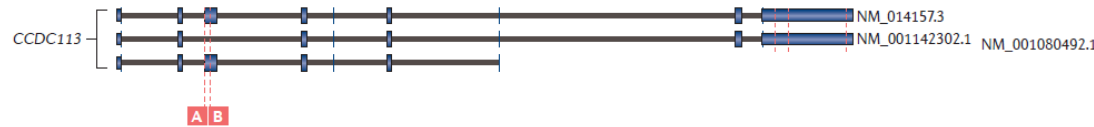
- Annotations are regularly updated

→ It is critical to keep track of exact annotation + version used in each project!



## Issues/difficulties with variant annotation

- Many genes have **multiple isoforms**
- Annotation may differ between isoforms



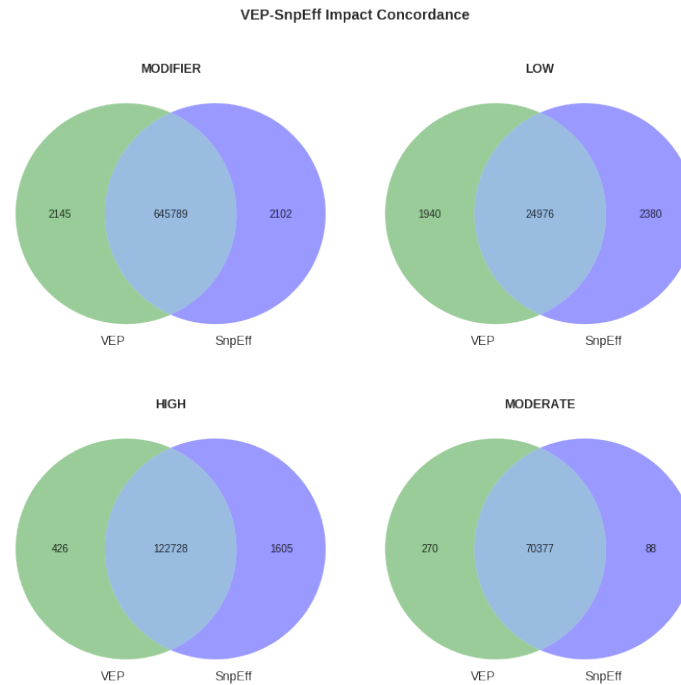
	Variant allele	Gene	Transcript change	RefSeq	Protein change	Molecular consequence	
A	rs765957496	G	CCDC113	c.228+1143A>G	NM_001142302.1	—	Intron variant
		G	CCDC113	c.229+2A>G	NM_014157.3	—	Splice acceptor variant
B	rs775877153	A	CCDC113	c.228+1182T>A	NM_001142302.1	—	Intron variant
		A	CCDC113	c.266T>A	NM_014157.3	Met89Lys	Missense variant

Eilbeck et al. 2017 Nature Methods

- We may either output all possible annotations or select a single annotation per variant, e.g.
  - Most serious effect
  - Canonical transcript (= longest transcript)
  - Biologically most relevant transcript  
(e.g. [https://m.ensembl.org/info/genome/genebuild/transcript\\_quality\\_tags.html](https://m.ensembl.org/info/genome/genebuild/transcript_quality_tags.html))

# Issues/difficulties with variant annotation

- Annotation may depend on **tool** that is used → But big efforts to improve consistency




<http://andrewjesaitis.com/2017/03/the-state-of-variant-annotation-in-2017/>


# Widely used tools for vcf annotation


## Ensembl VEP

### Web interface




- Point-and-click interface
- Suits smaller volumes of data


 [Documentation](#)





### Command line tool




- More options and flexibility
- For large volumes of data

 [Documentation](#)

 [Clone from GitHub](#)

 [Download \(zip\)](#)

 [Pull Docker image from DockerHub](#)

<https://www.ensembl.org/info/docs/tools/vep/index.html>

## Annovar

### Command line:

<https://doc-openbio.readthedocs.io/projects/annovar/en/latest/#annovar-documentation>

### Web-based: wAnnovar

<http://wannovar.wglab.org/>

## SnEff

Genetic variant annotation and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).  
Features:

- Supports over **38,000 genomes**.

### Command-line only (and Galaxy)

<https://pcingola.github.io/SnpEff/>