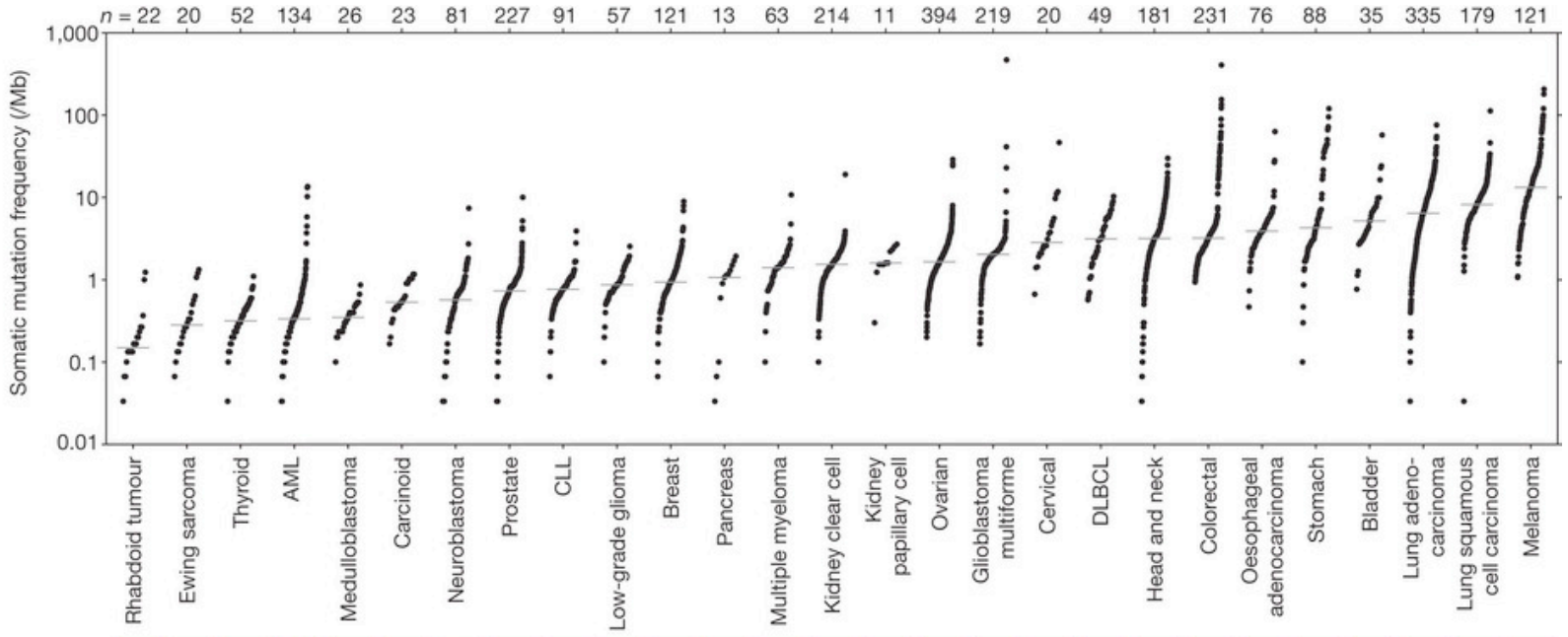
Annotating and prioritizing SNVs

Introduction

What to do with all the SNVs?

- So far: making sure the SNVs are biologically relevant difference and not technical artefacts



Lawrence et al. Nature 2013

Prioritizing

Does the SNV affect a gene, a transcription factor or miRNA binding site?

How likely is the SNV to affect protein function?

Does the SNV fall within a region that is highly duplicated in the genome?

Does the gene affect a known regulatory element, e.g. enhance or promoter?

How does the SNV affect a gene/ transcript?

Is the region containing the SNV evolutionary conserved?

Is the SNV known in other cancers?

Is the SNV known to be prevalent in a healthy population?

Tools for annotating SNVs/SNPs

- There are many…
- …we will use ANNOVAR:

ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data

Kai Wang ✉, Mingyao Li, Hakon Hakonarson

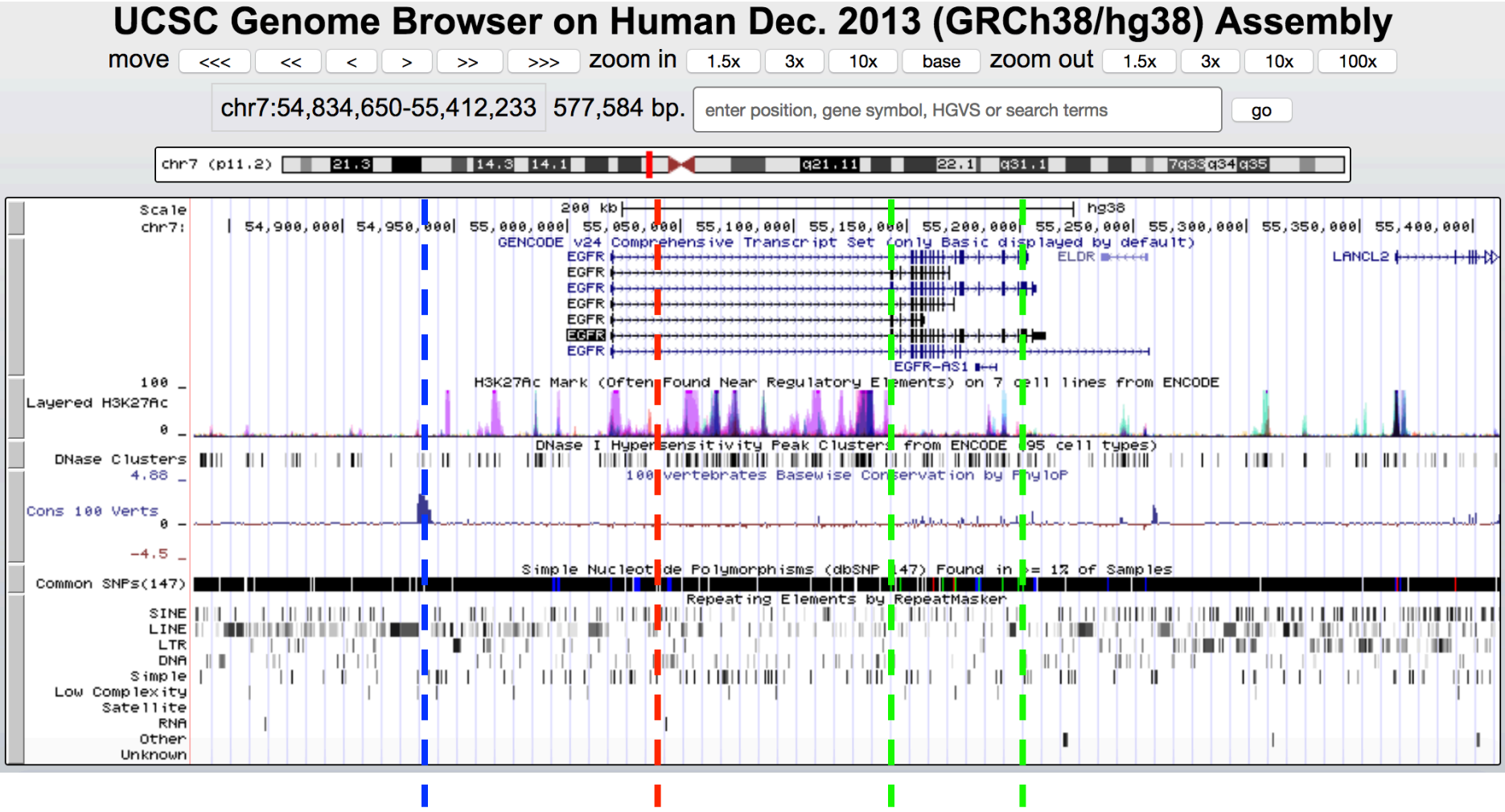
Nucleic Acids Res (2010) 38 (16): e164. **DOI:** https://doi.org/10.1093/nar/gkq603

Published: 03 July 2010 **Article history** ▾

Region-based annotation

- Annotations of variants based on overlap with specific genomic elements, e.g.:
  - Gene regions
  - conserved genomic regions,
  - (predicted) transcription factor binding sites,
  - (predicted) microRNA target sites
- Especially important for whole-genome sequencing data

Does the SNV affect a gene, a known enhancer, etc.?



Gene-based annotation – distance to gene

- Annotate variants to distinguish between intergenic, exonic, intronic, etc variants

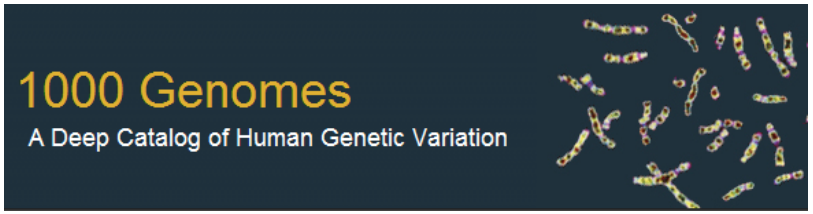
Value	Default precedence	Explanation
exonic	1	variant overlaps a coding
splicing	1	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)
ncRNA	2	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)
UTR5	3	variant overlaps a 5' untranslated region
UTR3	3	variant overlaps a 3' untranslated region
intronic	4	variant overlaps an intron
upstream	5	variant overlaps 1-kb region upstream of transcription start site
downstream	5	variant overlaps 1-kb region downtream of transcription end site (use -neargene to change this)
intergenic	6	variant is in intergenic region

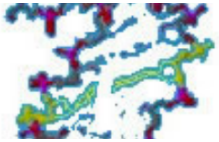
Gene-based annotation – effect on coding region

Annotation	Precedence	Explanation
frameshift insertion	1	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift deletion	2	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift block substitution	3	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
stopgain	4	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site. For frameshift mutations, the creation of stop codon downstream of the variant will not be counted as "stopgain"!
stoploss	5	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift deletion	7	a deletion of 3 or mutliples of 3 nucleotides that do not cause frameshift changes in protein coding sequence

Filter-based annotation

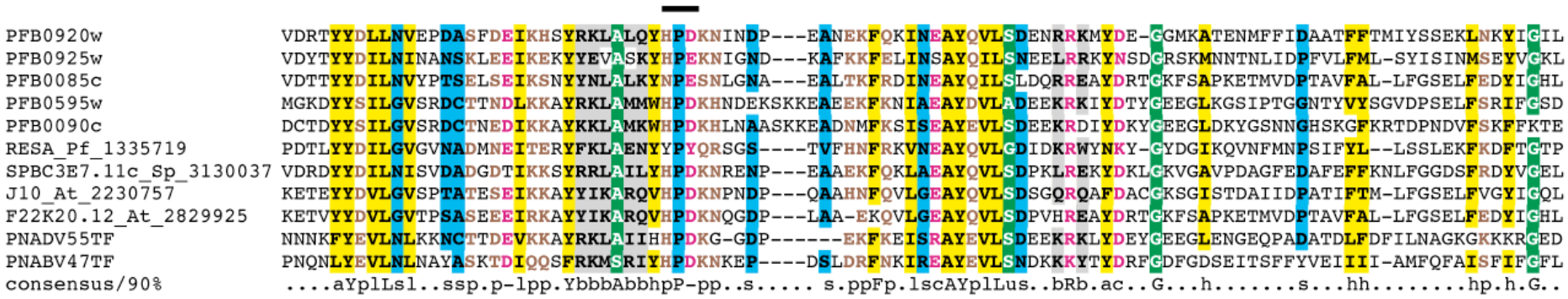
- Identify exact variant (with same start position, end position and alleles) in databases of …
  - Known (disease-related or common) variants
  - Predicted functional effect





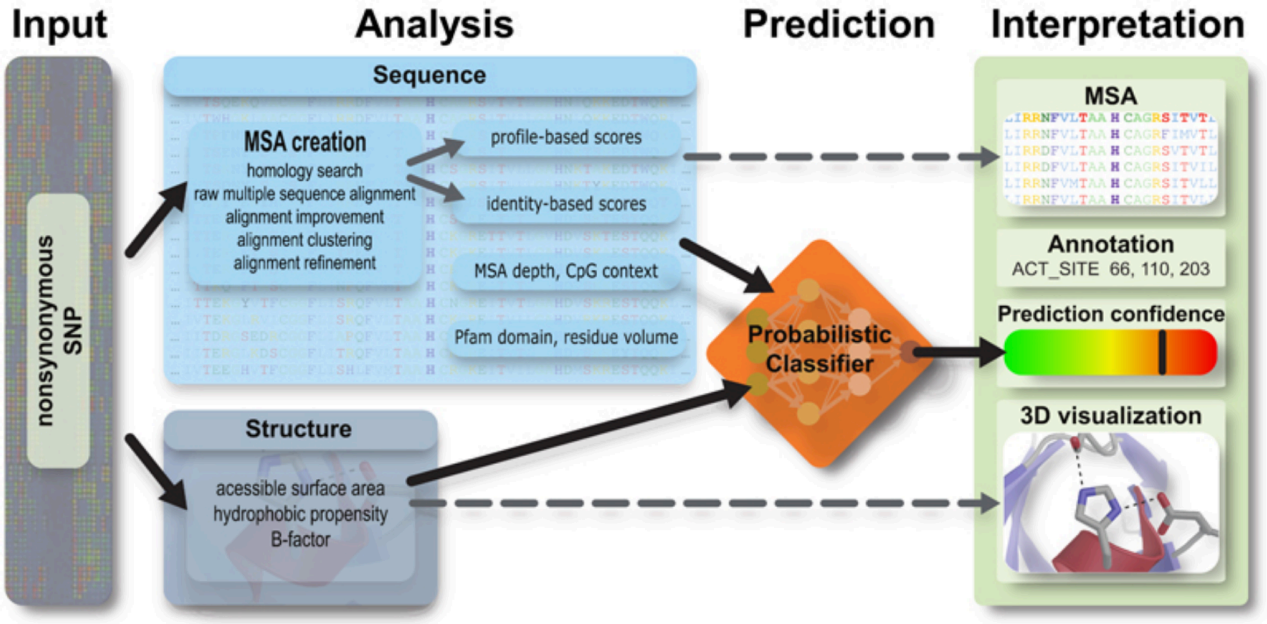
Predicted functional effect – SIFT score

- Rational:
  - Residues that are conserved completely in the protein family are expected to be important for function
- Give a protein sequence SIFT
  - …searches for closely related protein sequences
  - …performs multiple alignment
  - …calculates normalized probabilities for all possible substitution at each position (SIFT score)



Predicted functional effect – PolyPhen

- Prediction using sequence- and structure-based features
- Returns probability of a mutation to be deleterious
- Uses training set to learn important characteristics of a mutation, e.g.:
  - 13,032 human disease-causing mutations from UniProt
  - 8,946 human nsSNPs without annotated involvement in disease



Cohort-Based Prioritization