

# Variant annotation

## Variant Effect Predictor (VEP)

Edinburgh Genomics

Edinburgh, UK

23<sup>rd</sup> October 2015

Marta Bleda Latorre

[mb2033@cam.ac.uk](mailto:mb2033@cam.ac.uk)

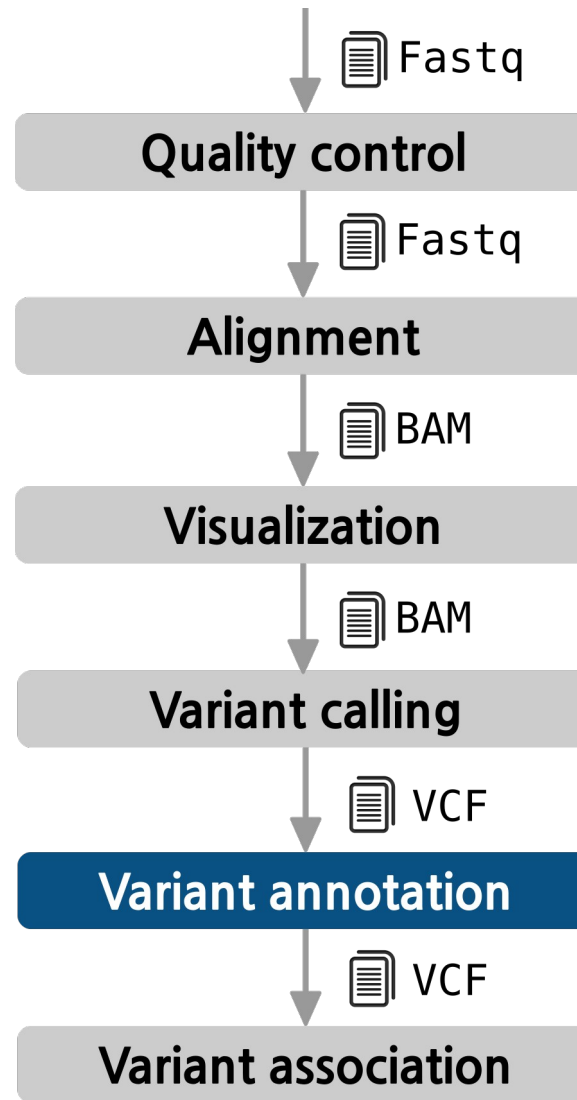
Research Assistant at the Department of Medicine

University of Cambridge

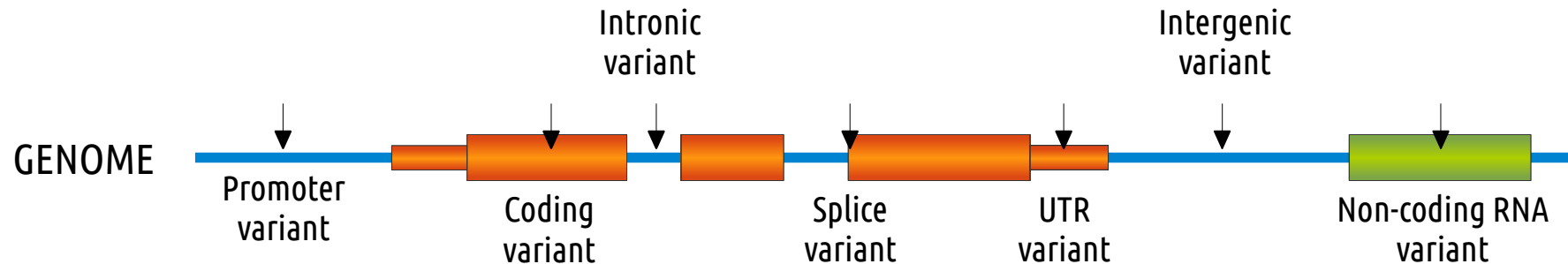
Cambridge, UK

# The pipeline

---



# What is functional annotation?



## Why we do that? PRIORITIZATION!

- ▶ Each individual exome carries between **25,000 and 50,000 variants**
- ▶ A **whole genome** can carry 3.5 million variants on average
- ▶ After annotating there will be **hundreds of deleterious variants**

### CAUTION!

On average, each *normal* person is found to carry:

~11,000 **synonymous** variants

~11,000 **non-synonymous** variants

**250 to 300 los-of-function** variants in annotated genes

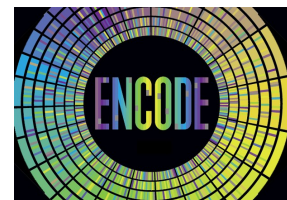
**50 to 100** variants previously implicated in inherited disorders

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. **Nature**. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

# Sources of functional information

**Table 1 Publicly available tools and databases for various tasks of genetic variant annotation and prioritization**

Category	Database/tool/project	Description	URL
Genetic variant data sources	dbSNP <sup>68</sup>	Comprehensive, curated SNP and short indel database	<a href="http://www.ncbi.nlm.nih.gov/projects/SNP">http://www.ncbi.nlm.nih.gov/projects/SNP</a>
	DbVar <sup>69</sup>	Comprehensive, curated database for structural variants	<a href="http://www.ncbi.nlm.nih.gov/dbvar">http://www.ncbi.nlm.nih.gov/dbvar</a>
	DGV <sup>70</sup>	Human structural variants from samples with no phenotype	<a href="http://projects.tcag.ca/variation">http://projects.tcag.ca/variation</a>
Functional characterization of genomic elements	ENCODE <sup>71</sup>	High-throughput functional characterization of DNA elements, including noncoding regions	<a href="http://www.genome.gov/10005107">http://www.genome.gov/10005107</a>
	SIFT <sup>72</sup> , PolyPhen <sup>73</sup>	Prioritization of nonsynonymous SNPs	<a href="http://sift.jcvi.org">http://sift.jcvi.org</a> , <a href="http://genetics.bwh.harvard.edu/pph2">http://genetics.bwh.harvard.edu/pph2</a>
Public gene–trait associations	dbGaP <sup>34</sup>	Comprehensive listing of genotype-to-phenotype mappings	<a href="http://www.ncbi.nlm.nih.gov/gap">http://www.ncbi.nlm.nih.gov/gap</a>
	EGA <sup>74</sup>	Genotype–phenotype experiment archive	<a href="http://www.ebi.ac.uk/ega">http://www.ebi.ac.uk/ega</a>
Disease-associated mutations	HGMD <sup>35</sup>	Database for human disease mutations	<a href="http://www.hgmd.org">http://www.hgmd.org</a>
	OMIM <sup>36</sup>	Mendelian disease gene associations	<a href="http://www.ncbi.nlm.nih.gov/omim">http://www.ncbi.nlm.nih.gov/omim</a>
	SwissVar <sup>76</sup>	Variant catalog of the UniProt knowledge bases	<a href="http://swissvar.expasy.org">http://swissvar.expasy.org</a>
	GAD <sup>77</sup>	NCBI source for genotype–disease associations	<a href="http://geneticassociationdb.nih.gov">http://geneticassociationdb.nih.gov</a>
	GWAS catalog from NHGRI <sup>78</sup>	SNP-phenotype associations found by GWAS	<a href="http://www.genome.gov/gwastudies">http://www.genome.gov/gwastudies</a>
Whole-genome repositories	Complete genomics public genomes <sup>79</sup>	Complete genomics for 69 genomes from multiple ancestries (includes samples from the NHGRI and NIGMS repositories)	<a href="http://www.completegenomics.com/sequence-data/download-data">http://www.completegenomics.com/sequence-data/download-data</a>
	1,000 Genomes <sup>80</sup>	Expanding resource currently housing three low-coverage whole genomes of multiple ancestries	<a href="http://www.1000genomes.org">http://www.1000genomes.org</a>
Ancestry-focused variant data sources	HapMap <sup>26</sup>	Haplo-block mapping for diverse populations	<a href="http://www.hapmap.org">http://www.hapmap.org</a>
	HGDP <sup>27</sup>	SNP profiles of samples from several endogenous populations	<a href="http://hagsc.org/hgdp">http://hagsc.org/hgdp</a>
Pharmacogenomic associations and data sources	PharmGKB <sup>56</sup>	Variant–pharmacokinetic/pharmacodynamic trait associations and gene–drug interactions	<a href="http://www.pharmgkb.org">http://www.pharmgkb.org</a>
	DrugBank <sup>81</sup>	Drug-target database with biochemical properties	<a href="http://drugbank.ca">http://drugbank.ca</a>



Cordero P, Ashley EA. Whole-genome sequencing in personalized therapeutics. *Clin Pharmacol Ther.* 2012 Jun ;91(6):1001-9. PubMed PMID: 22549284

# Computational method and tools

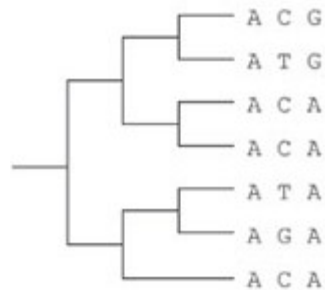
- ▶ **Annotated information** is sometimes **limited**, particularly for rare and complex traits
- ▶ Computational methods can measure deleteriousness by using **comparative genomics** and knowledge of **protein biochemistry and structure**

## Comparative Genomics

Focus on sequences that have not been removed by **natural selection**.

**Quantify evolutionary changes** in genes or genomes and define conserved and neutral regions.

Variants observed in conserved sites are highly likely to be **deleterious**.

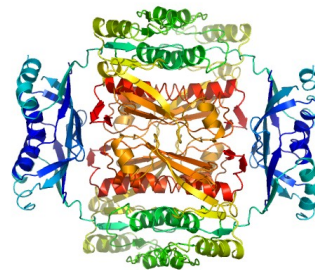


## Effects in protein-coding variants

Can combine **evolutionary** and **biochemical** information.

Use **alignments of homologous proteins** to estimate mutational deleteriousness.

Use **biochemical data** such as amino acid properties, binding information and structural information to estimate the impact.



## Effects in non-coding variants

The majority of the human genetic variation is in non-coding regions.

**No detectable conservation outside vertebrates.**

Main strategy for estimation is testing the **mammalian conservation** of the non-coding variants.

Cooper GM, Shendure J. *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* **Nature Reviews Genetics.** 2011 Aug 18;12(9):628-40. Pubmed PMID: 21850043

# Computational methods and tools

## Prediction scores for non-synonymous variants

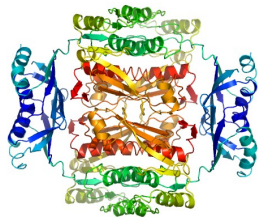
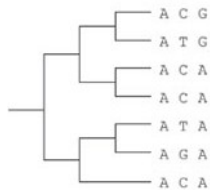


Table 1 | Tools for protein-sequence-based prediction of deleteriousness

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	<a href="http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html">http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html</a>	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	<a href="http://sift.bii.a-star.edu.sg/">http://sift.bii.a-star.edu.sg/</a>	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	<a href="http://www.pantherdb.org/">http://www.pantherdb.org/</a>	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	<a href="http://www.mutationtaster.org/">http://www.mutationtaster.org/</a>	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	<a href="http://snpanalyzer.uthsc.edu/">http://snpanalyzer.uthsc.edu/</a>	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	<a href="http://mmb2.pcb.ub.es:8080/PMut/">http://mmb2.pcb.ub.es:8080/PMut/</a>	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	<a href="http://genetics.bwh.harvard.edu/pph2/">http://genetics.bwh.harvard.edu/pph2/</a>	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	<a href="http://sapred.cbi.pku.edu.cn/">http://sapred.cbi.pku.edu.cn/</a>	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	<a href="http://www.rostlab.org/services/SNAP/">http://www.rostlab.org/services/SNAP/</a>	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	<a href="http://www.snps3d.org/">http://www.snps3d.org/</a>	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	<a href="http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html">http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html</a>	37

\*Also makes predictions for synonymous and non-coding variant effects: for example, splicing. MAPP, Multivariate Analysis of Protein Polymorphism; polyPhen, polymorphism phenotyping.

Kircher M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. **Nature Genetics**. 2014; Pubmed PMID: 24487276.

Shihab H. A.. et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. **Bioinformatics**. 2014; Pubmed PMID: 25583119.

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. **Nature Reviews Genetics**. 2011 Aug 18;12(9):628-40. Pubmed PMID: 21850043



# Computational methods and tools

## Prediction scores for non-coding variation

Table 2 | **Tools for nucleotide-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
phastCons	Phylogenetic HMM	Evolutionary	<a href="http://compgen.bscb.cornell.edu/phast/">http://compgen.bscb.cornell.edu/phast/</a>	60
GERP	Single-site scoring	Evolutionary	<a href="http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html">http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html</a>	67
Gumby	Single-site scoring	Evolutionary	<a href="http://pga.jgi-psf.org/gumby/">http://pga.jgi-psf.org/gumby/</a>	21
phyloP	Single-site scoring	Evolutionary	<a href="http://compgen.bscb.cornell.edu/phast/">http://compgen.bscb.cornell.edu/phast/</a>	66
SCONE	Single-site scoring	Evolutionary	<a href="http://genetics.bwh.harvard.edu/scone/">http://genetics.bwh.harvard.edu/scone/</a>	68
binCons	Sliding-window scoring	Evolutionary	<a href="http://zoo.nhgri.nih.gov/binCons/index.cgi">http://zoo.nhgri.nih.gov/binCons/index.cgi</a>	69
Chai Cons	Sliding-window scoring	Evolutionary and structural	<a href="http://research.nhgri.nih.gov/software/chai">http://research.nhgri.nih.gov/software/chai</a>	71
VISTA	Visualization tool (various scores)	Evolutionary	<a href="http://genome.lbl.gov/vista/index.shtml">http://genome.lbl.gov/vista/index.shtml</a>	70

GERP, Genomic Evolutionary Rate Profiling; HMM, hidden Markov model; SCONE, Sequence Conservation Evaluation.

Cooper GM, Shendure J. *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* **Nature Reviews Genetics.** 2011 Aug 18;12(9):628-40. Pubmed PMID: 21850043





# Variant Effect Predictor (VEP)

---

VEP documentation site: <http://www.ensembl.org/info/docs/variation/vep/index.html>




- Backed by **Ensembl**
- Free and open source
- **3 ways of functionality**: web interface, standalone Perl script and REST API
- **Input formats**: CSV, VCF, Pileup and HGVS
- **Regulatory information**: TFBSs
- **Filtering** by coding regions and MAF
- **Predictions**: SIFT, PolyPhen
- 1000 genomes and dbSNP information
- Uses **Sequence Ontology**
- **Many species**
- Regulatory information does **not include miRNA targets**
- The **standalone Perl script** needs:
  - **Perl** and **MySQL** support
  - **Download, install and update** every ~ 2 months

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor*. **BMC Bioinformatics** 26(16):2069-70(2010) Pubmed PMID: 20562413

# Variant Effect Predictor (VEP)

VEP web interface: [http://www.ensembl.org/Homo\\_sapiens/Tools/VEP](http://www.ensembl.org/Homo_sapiens/Tools/VEP)

**Input**

Species:	 Human (Homo sapiens) 
Assembly: GRCh37	
Name for this data (optional):	<input type="text"/>
Input file format ( <a href="#">details</a> ):	Ensembl default 
Either paste data:	<div><pre>1 909238 909238 G/C + 3 361464 361464 A/- + 5 121187650 121188519 DUP</pre></div>
Or upload file:	<input type="button" value="Choose File"/> No file chosen
Or provide file URL:	<input type="text"/>
Transcript database to use:	<p><input checked="" type="radio"/> Ensembl transcripts</p> <p><input type="radio"/> Gencode basic transcripts</p> <p><input type="radio"/> RefSeq transcripts</p> <p><input type="radio"/> Ensembl and RefSeq transcripts</p>

**Output options**

Additional identifiers for genes, transcripts and variants; frequency data

e.g. SIFT, PolyPhen and regulatory data

Pre-filter results by frequency or consequence type

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor*. **BMC Bioinformatics** 26(16):2069-70(2010) Pubmed PMID: 20562413

# CellBase v3.2.0

<https://github.com/opencb/cellbase/>

---

- NoSQL database that integrates the most relevant biological information:
  - Core features: genes, transcripts, exons, proteins, genome sequence, etc.
  - Regulatory: Ensembl regulatory, TFBS, miRNA targets, CTCF, Open chromatin, etc.
  - Functional annotation: OBO ontologies (Gene Ontology, Human Disease Ontology), etc.
  - Genomic variation: Ensembl Variation, ClinVar, COSMIC, etc.
  - Systems biology: IntAct, Biogrid, Reactome, gene co-expression, etc.
  - More than 30 species available
- An exhaustive RESTful Web service API has been implemented:  
<http://wwwdev.ebi.ac.uk/cellbase/webservices/rest/v3/hsapiens/genomic/region/22:17449263:17449264/gene>
- New features included in v3.2.0
  - New Variant Annotation functionality
  - New data:
    - RefSeq annotation
    - DisGeNET
    - Gene Expression Atlas

Bleda M, et al. *CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources*. **Nucleic Acids Research**. 2012 Pubmed PMID: 22693220

# CellBase v3.2.0 annotator

<https://github.com/opencb/cellbase/>

CellBase documentation site: <https://github.com/opencb/cellbase/wiki>

- Free and open source.
- Institutions using it: EBI - EVA, GEL
- 2 ways of functionality: CLI program, Java RESTful WS API
- Cloud variant annotator. Requires no installation or updates
- Consequence type in Ensembl and RefSeq genes (Sequence Ontology)
- Regulatory regions: TFBS, miRNA target, ENCODE
- Conservation scores: phastCons and phyloP.
- Protein substitution effect scores: PolyPhen-2, SIFT
- Control studies frequencies: 1000 genomes, EVS
- Gene/Transcript Expression: Gene Expression Atlas
- Clinical phenotype information: ClinVar, COSMIC, GWAS catalog
- More than 30 species (human, mouse, rat, zebra fish, worm, fly, yeast, dog, pig, mosquito, etc.)
- Young program, many new features coming
  - UniProt, InterPro, Intact, Emory clinical data, DGIdb, GERP++ and many others
  - Large structural variants annotation

Bleda M, et al. *CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources*. **Nucleic Acids Research**. 2012 Pubmed PMID: 22693220

# CellBase v3.1.0 annotator

<https://github.com/opencb/cellbase/>

- **Download** CellBase code and save it into your course/variant\_annotation folder:

<https://github.com/opencb/cellbase/wiki/Download-and-Installation>

- Extract the contents

## Download and Installation

Nacho edited this page 14 hours ago · 15 revisions

This tutorial is intended to explain how to download and install CellBase and RESTful web services. To learn how to clone and build CellBase please visit [README](#)

## Getting CellBase

You can get stable CellBase software in **two** different ways:

1. You can download a built version packaged in a **tar.gz** from [GitHub releases](#). This is the recommended way for using stable releases.
2. or, you can download source code by cloning the repository and following the [README](#) documentation to build CellBase, this is useful for testing or trying the last features from the development branch of CellBase, notice you should **not** use this for production.

To download most featured Cellbase version click here [CellBase v3.2.0](#)

- Usage:

```
cellbase.sh variant-annotation --input-file file.vcf --output-file file.vep
```

- REST API:

[http://bioinfodev.hpc.cam.ac.uk/cellbase/webservices/rest/v3/hsapiens/genomic/variant/19:45411941:T:C/full\\_annotation](http://bioinfodev.hpc.cam.ac.uk/cellbase/webservices/rest/v3/hsapiens/genomic/variant/19:45411941:T:C/full_annotation)



# AnnoVar

ANNOVAR web site: <http://www.openbioinformatics.org/annovar/>

- Free and open source
- Can annotate SNV, insertions and deletions
- **Regulatory information:** Conserved genomic regions, TFBSs, miRNA targets and predicted miRNA secondary structures. ENCODE DNase I hypersensitive sites, Histone methylations, ChIP and RNA-Seq peaks
- DbSNP, 1000 genomes, SIFT and GERP filtering
- **Predictions:** Polyphen, LRT, MutationTaster, PhyloP
- Can handle **custom annotations** in GFF3
- Can handle 1 o 0-based coordinates
- **5 Species** (human, mouse, worm, fly, yeast)

- Accepts VCF4, GFF3-SOLiD and CSV BUT after conversion to their **particular input file:**

Chr	Start	End	Ref	Obs	Comments
1	161003	161003	C	T	comments: rs1000050

- **Perl** written program
- **Installation** required
- Users need to **download** every annotation database and save them locally (~35GB per assembly)
- Need to be **run several times**
- **Output:** several files depending on the query

Wang K, Li M, Hakonarson H. *ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data*. **Nucleic Acids Research**. Sep;38(16):e164 Pubmed PMID: 20601685

# AnnoVar

---

## EXAMPLE of ANNOVAR usage

### DOWNLOADING BIOLOGICAL DATA:

```
user@computer:~$ annotate_variation.pl -buildver hg19 -downdb refgene humandb/
user@computer:~$ annotate_variation.pl -buildver hg19 -downdb snp135 -webfrom annovar humandb/
user@computer:~$ annotate_variation.pl -buildver hg19 -downdb phastConsElements46way humandb/
user@computer:~$ annotate_variation.pl -buildver hg19 -downdb 1000g2012apr -webfrom annovar humandb/
user@computer:~$ annotate_variation.pl -buildver hg19 -downdb cytoBand humandb/
```

### EXTRACTING THE EFFECT:

```
user@computer:~$ annotate_variation.pl -geneanno example/ex1.human humandb/
user@computer:~$ annotate_variation.pl -regionanno -dbtype band example/ex1.human humandb/
user@computer:~$ annotate_variation.pl -filter -dbtype 1000g2012apr_eur example/ex1.human humandb/
```

**THANK YOU.**