

Variant annotation

ANNOVAR and HPG-VARIANT

University of Cambridge

Cambridge, UK

17th June 2015

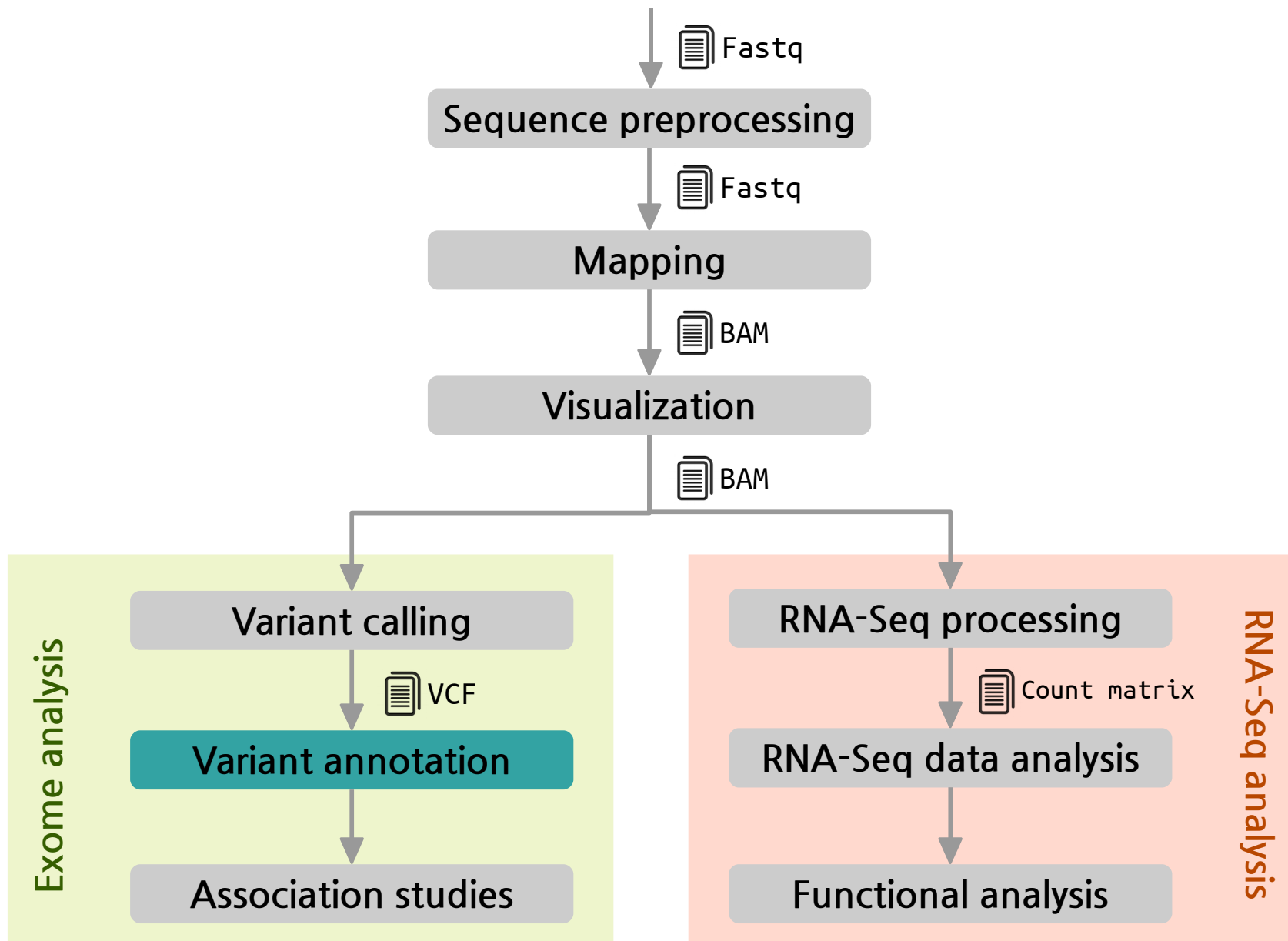
Javier López

fjlopez@ebi.ac.uk

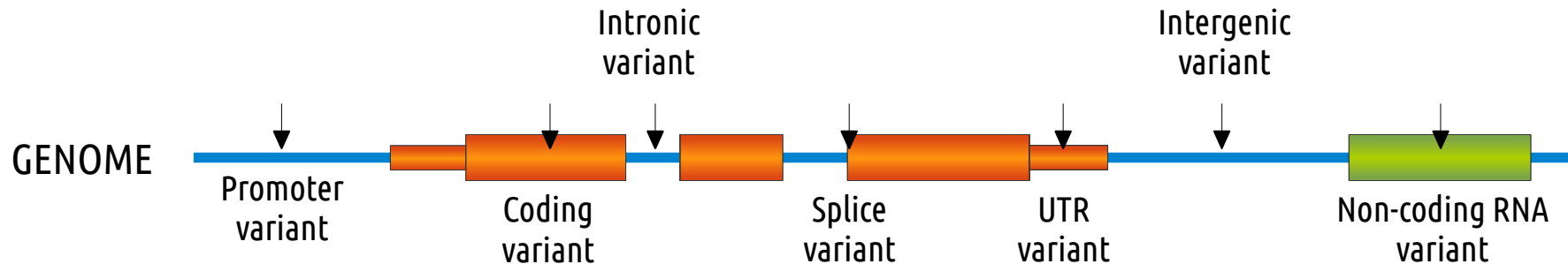
EBML-European Bioinformatics Institute

Acknowledgements: Marta Bleda Latorre

The pipeline



What is functional annotation?



Why we do that?

- ▶ Each individual exome carries ~25,000 variants → **PRIORITIZATION!**
- ▶ We want to identify a **small subset** of functionally important variants to pinpoint the putative disease causal variants
- ▶ We need strategies to **estimate the deleteriousness** of our variants to better identify disease-causal 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

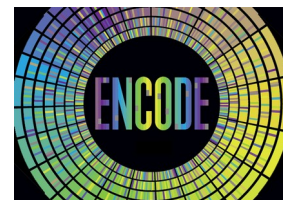
50 to 100 variants previously implicated in inherited disorders

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



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

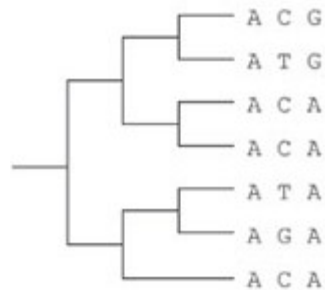
- ▶ **Annotated information** is sometimes **limited**, particularly for rare and complex traits
- ▶ Computational methods can measure deleteriousness by using **comparative genomics** and knowledge of **protein biochemistry and structure**

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**.

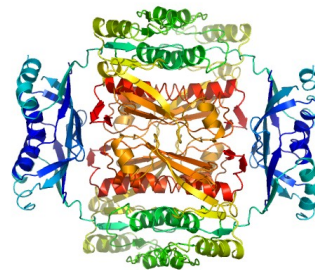


Effects in protein-coding variants

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

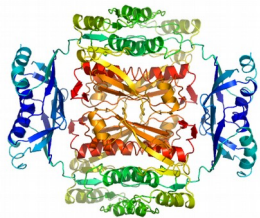
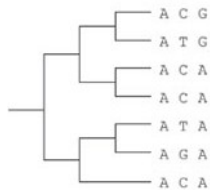


Table 1 | Tools for protein-sequence-based prediction of deleteriousness

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://sift.bii.a-star.edu.sg/	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://www.pantherdb.org/	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	http://www.mutationtaster.org/	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	http://snpanalyzer.uthsc.edu/	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	http://mmb2.pcb.ub.es:8080/PMut/	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	http://genetics.bwh.harvard.edu/pph2/	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	http://sapred.cbi.pku.edu.cn/	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	http://www.rostlab.org/services/SNAP/	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	http://www.snps3d.org/	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	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	http://compgen.bscb.cornell.edu/phast/	60
GERP	Single-site scoring	Evolutionary	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	67
Gumby	Single-site scoring	Evolutionary	http://pga.jgi-psf.org/gumby/	21
phyloP	Single-site scoring	Evolutionary	http://compgen.bscb.cornell.edu/phast/	66
SCONE	Single-site scoring	Evolutionary	http://genetics.bwh.harvard.edu/scone/	68
binCons	Sliding-window scoring	Evolutionary	http://zoo.nhgri.nih.gov/binCons/index.cgi	69
Chai Cons	Sliding-window scoring	Evolutionary and structural	http://research.nhgri.nih.gov/software/chai	71
VISTA	Visualization tool (various scores)	Evolutionary	http://genome.lbl.gov/vista/index.shtml	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

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/
```

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 Ensembl's Perl 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
- Perl **API** requires:
 - **Installation**
 - **Downloads** and **update**
 - API documentation → **Hard to understand**

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

Input

Species:	 Human (Homo sapiens) 
Assembly: GRCh37	
Name for this data (optional):	<input type="text"/>
Input file format (details) :	Ensembl default 
Either paste data:	<div>1 909238 909238 G/C + 3 361464 361464 A/- + 5 121187650 121188519 DUP</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

Identifiers and frequency data 	Additional identifiers for genes, transcripts and variants; frequency data
Extra options 	e.g. SIFT, PolyPhen and regulatory data
Filtering options 	Pre-filter results by frequency or consequence type
<input type="button" value="Run >"/> Reset	

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)

EXAMPLE of API usage: Getting all variations in a particular human transcript and see what is the effect of that variation in the transcript

```
1 use strict;
2 use warnings;
3 use Bio::Ensembl::Registry;
4
5 my $registry = 'Bio::Ensembl::Registry';
6
7 $registry->load_registry_from_db(
8     -host => 'ensembl.ensembl.org',
9     -user => 'anonymous'
10 );
11
12 my $stable_id = 'ENST00000393489'; #this is the stable_id of a human transcript
13 my $transcript_adaptor = $registry->get_adaptor('homo_sapiens', 'core', 'transcript'); #get the adaptor to get the Transcript from the database
14 my $transcript = $transcript_adaptor->fetch_by_stable_id($stable_id); #get the Transcript object
15
16 my $trv_adaptor = $registry->get_adaptor('homo_sapiens', 'variation', 'transcriptvariation'); #get the adaptor to get TranscriptVariation objects
17 my $trvs = $trv_adaptor->fetch_all_by_Transcripts([$transcript]); #get ALL effects of Variations in the Transcript
18
19 foreach my $tv (@{$trvs}) {
20     my $tvas = $tv->get_all_alternate_TranscriptVariationAlleles();
21
22     foreach my $tva(@{$tvas}) {
23         my @ensembl_consequences;
24         my @so_consequences;
25
26         my $ocs = $tva->get_all_OverlapConsequences();
27
28         foreach my $oc(@{$ocs}) {
29             push @ensembl_consequences, $oc->display_term;
30             push @so_consequences, $oc->SO_term;
31         }
32
33         my $sift = $tva->sift_prediction;
34         my $polyphen = $tva->polyphen_prediction;
35
36         print
37             "Variation ", $tv->variation_feature->variation_name,
38             " allele ", $tva->variation_feature_seq,
39             " has consequence ", join(", ", @ensembl_consequences),
40             " (SO ", join(", ", @so_consequences), ").";
41
42         if(defined($sift)) {
43             print " SIFT=$sift";
44         }
45         if(defined($polyphen)) {
46             print " PolyPhen=$polyphen";
47         }
48
49         print "\n";
50     }
51 }
```

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 v3.1.0

<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 20 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 v3.1.0

- New CLI that integrates CellBase building and query commands
- 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 v3.1.0 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, GoNL, Spanish Variation, EVS and uk10k
- **Gene/Transcript Expression**: Gene Expression Atlas
- Clinical phenotype information: ClinVar, COSMIC, GWAS catalog
- **13 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, DGldb, DisGeNET, GERP++ and many others
 - Many more species (~25 new species)
 - 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/releases>

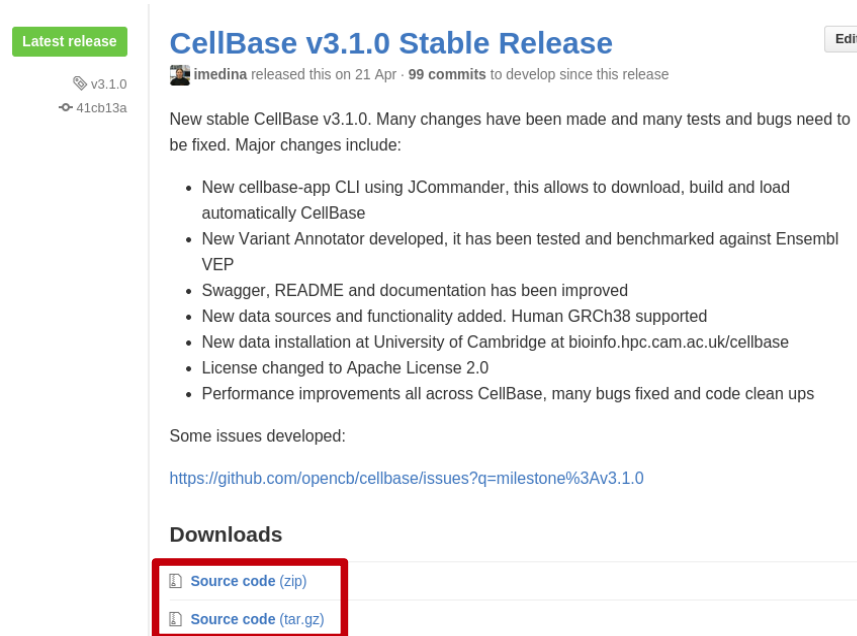
- Extract the contents

- Usage:

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

- REST API:

```
http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v3/hsapiens/genomic/variant/19:45411941:T:C/full_annotation
```



The screenshot shows the GitHub release page for CellBase v3.1.0. The page title is "CellBase v3.1.0 Stable Release" with an "Edit" button. It states that the release was made on 21 Apr with 99 commits. A list of major changes includes: a new CLI using JCommander, a new Variant Annotator, improved documentation, new data sources (Human GRCh38), a new data installation at the University of Cambridge, a license change to Apache 2.0, and various performance improvements. A link to issues for this milestone is provided. Under the "Downloads" section, two links are listed: "Source code (zip)" and "Source code (tar.gz)", both of which are highlighted with a red rectangular box.

THANK YOU.