Variant annotation

ANNOVAR and HPG-VARIANT

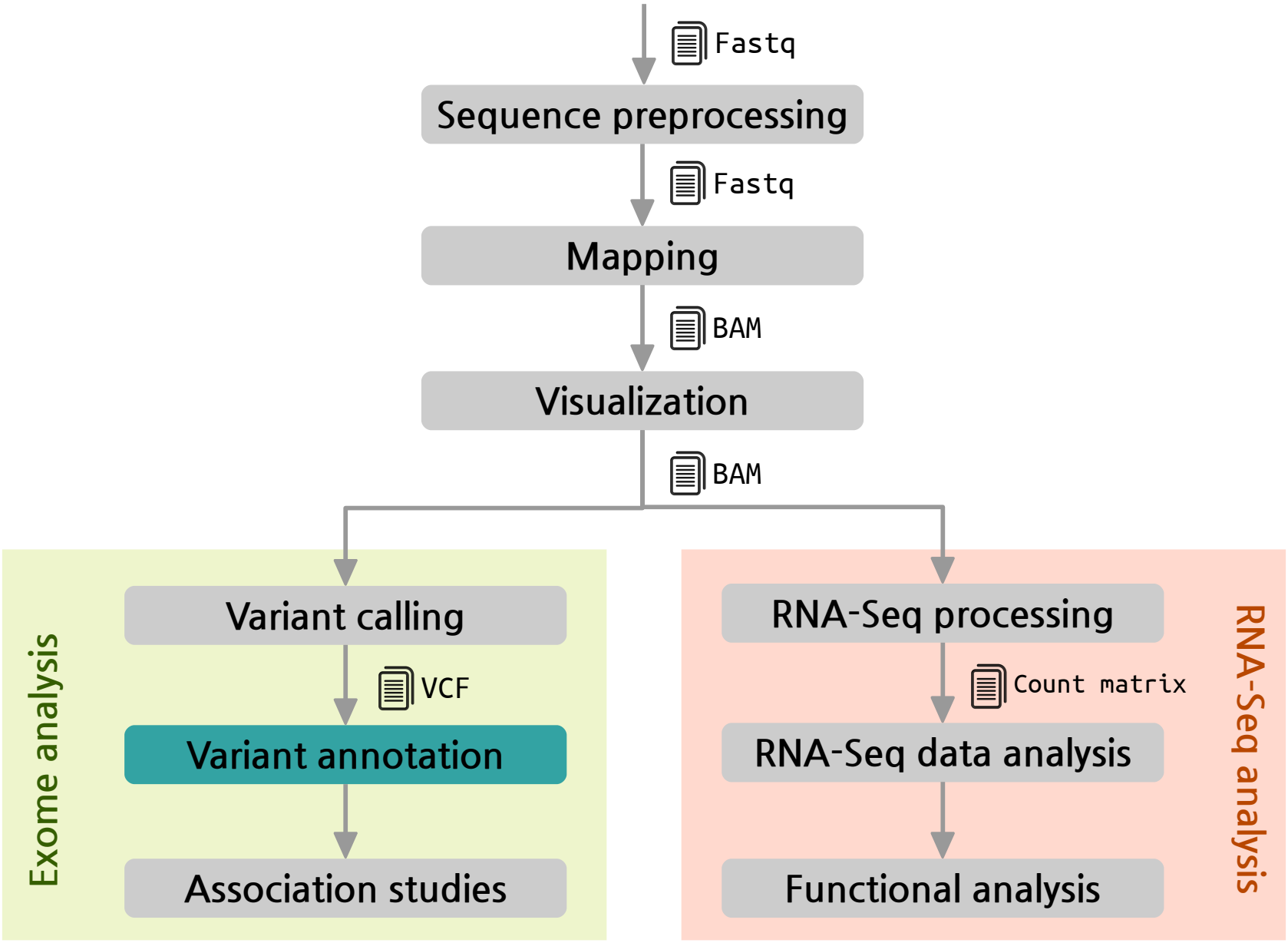
University of Cambridge
Cambridge, UK
30th September 2014

Marta Bleda Latorre
mb2033@cam.ac.uk
Research Assistant at the Department of Medicine
University of Cambridge
Cambridge, UK

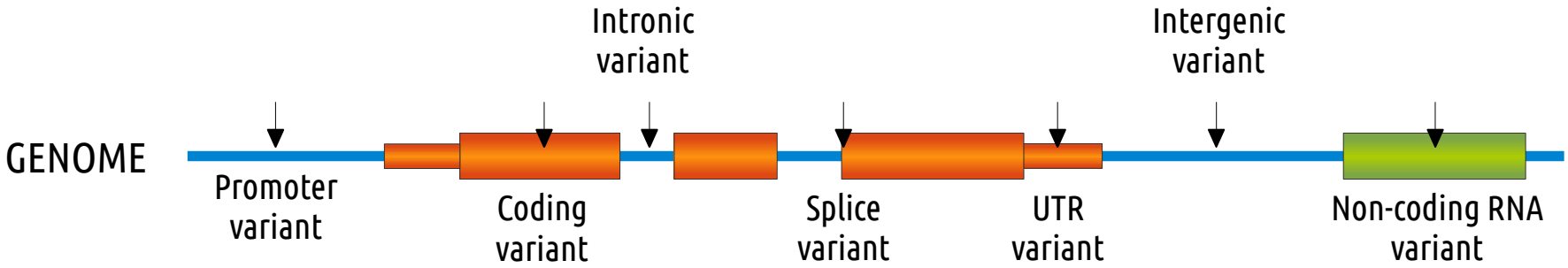
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

UNIVERSITY OF CAMBRIDGE

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 remove 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.

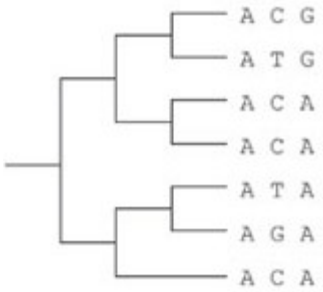
Use **biochemical data** such as amino acid properties, binding information and structural information to estimate the impact.

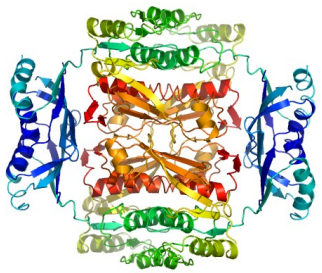
Effects in non-coding variants

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.

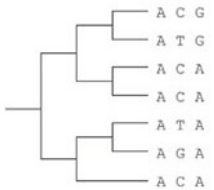




Cooper GM, Shendure J. *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* **Nature Reviews Genetics**. 2011 Aug 18;12(9):628-40. Pubmed PMID: 21850043

Computational methods and tools

Prediction scores for non-synonymous variants



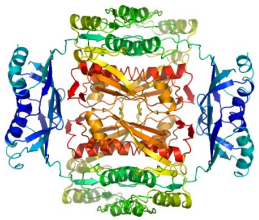


Table 1 | **Tools for protein-sequence-based prediction of deleteriousness**

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.

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**

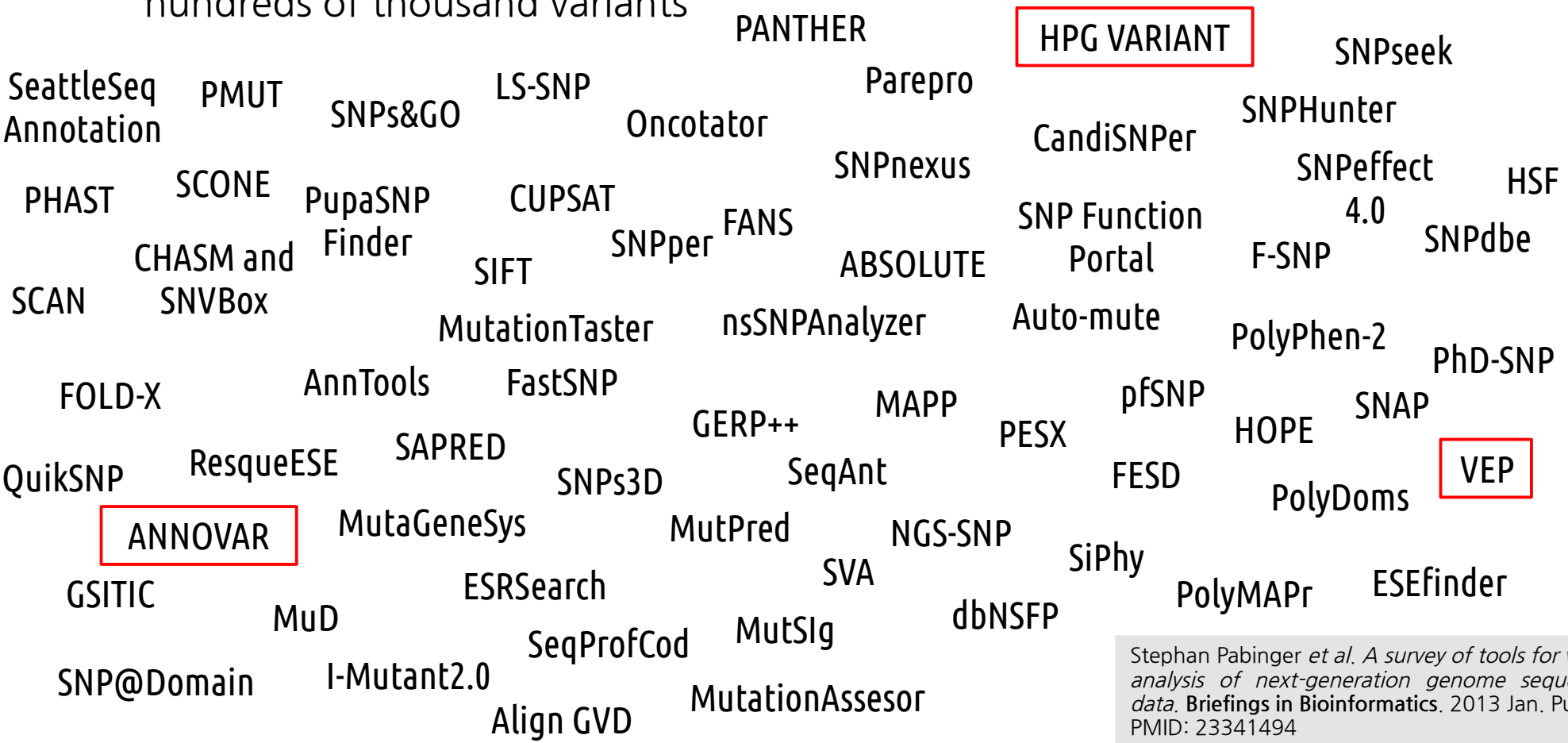
Name	Type	Information	URL	Refs
phastCons	Phylogenetic HMM	Evolutionary	http://compgen.bscb.cornell.edu/phast/	60
GERP	Single-site scoring	Evolutionary	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	67
Gumby	Single-site scoring	Evolutionary	http://pga.jgi-psf.org/gumby/	21
phyloP	Single-site 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

Tools for functional annotation

- We need to measure the **impact** of each variant in the genome
- We **cannot** annotate 25,000 variants **manually** checking more than 20 databases
- **Tools integrate** biological information and **ease** the functional annotation of hundreds of thousand variants



Stephan Pabinger *et al. A survey of tools for variant analysis of next-generation genome sequencing data.* **Briefings in Bioinformatics**. 2013 Jan. Pubmed PMID: 23341494

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

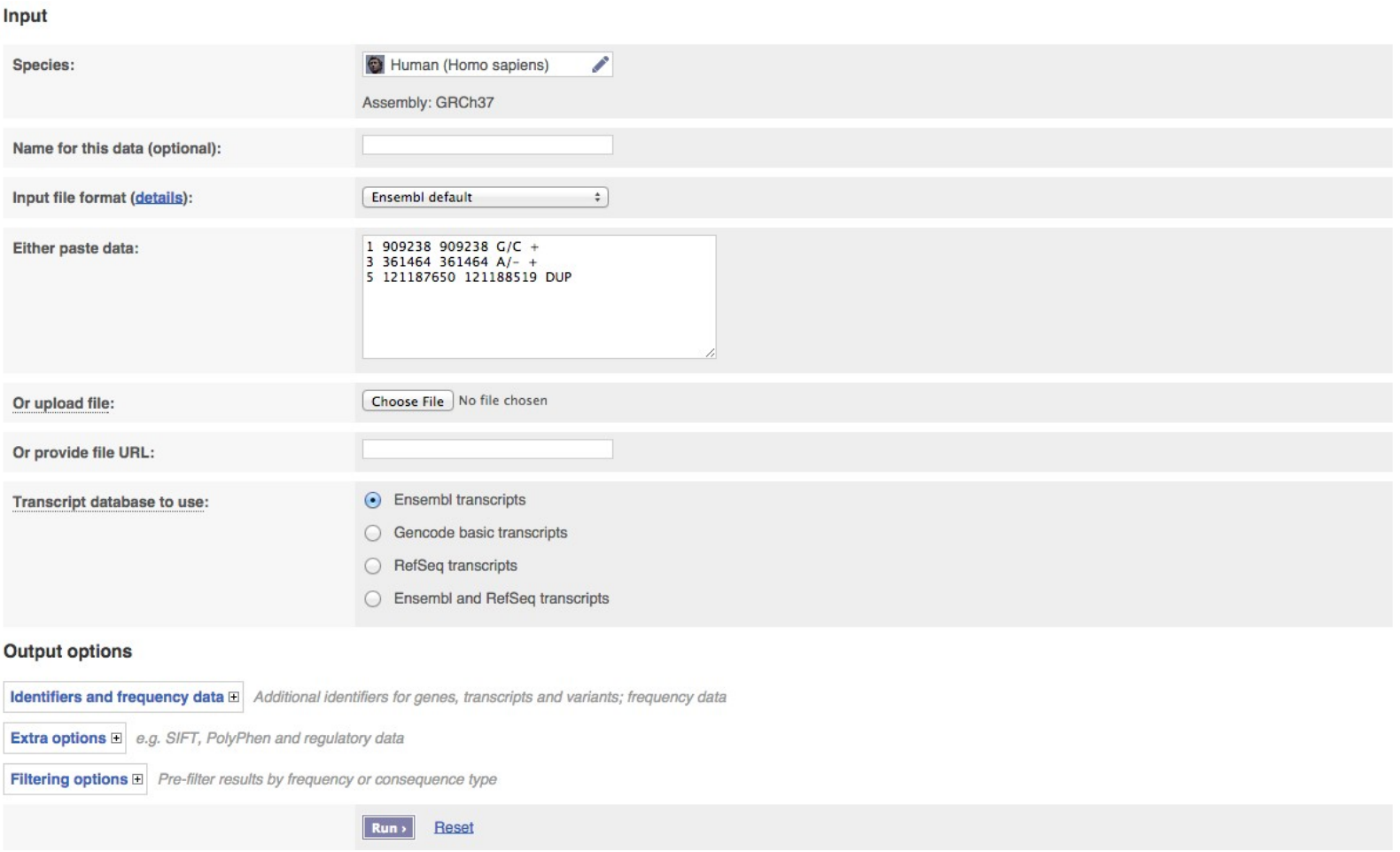
- Backed by **Ensembl**
- Free and open source
- **3 ways of functionality**: web interface, standalone Perl script and Ensembl's Perl API
- **Input** formats: CSV, VCF, Pileup and HGVS
- **Regulatory information**: TFBSs
- **Filtering** by coding regions and MAF
- **Predictions**: SIPF, PolyPhen
- 1000 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



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

```perl
use strict;
use warnings;
use Bio::EnsEMBL::Registry;

my $registry = 'Bio::EnsEMBL::Registry';

$registry->load_registry_from_db(
    -host => 'ensembldb.ensembl.org',
    -user => 'anonymous'
);

my $stable_id = 'ENST00000393489'; #this is the stable_id of a human transcript
my $transcript_adaptor = $registry->get_adaptor('homo_sapiens', 'core', 'transcript'); #get the adaptor to get the Transcript from the database
my $transcript = $transcript_adaptor->fetch_by_stable_id($stable_id); #get the Transcript object

my $trv_adaptor = $registry->get_adaptor('homo_sapiens', 'variation', 'transcriptvariation'); #get the adaptor to get TranscriptVariation objects
my $trvs = $trv_adaptor->fetch_all_by_Transcripts([$transcript]); #get ALL effects of Variations in the Transcript

foreach my $tv (@{$trvs}) {
        my $tvas = $tv->get_all_alternate_TranscriptVariationAlleles();

        foreach my $tva(@{$tvas}) {
                my @ensembl_consequences;
                my @so_consequences;

                my $ocs = $tva->get_all_OverlapConsequences();

                foreach my $oc(@{$ocs}) {
                        push @ensembl_consequences, $oc->display_term;
                        push @so_consequences, $oc->SO_term;
                }

                my $sift = $tva->sift_prediction;
                my $polyphen = $tva->polyphen_prediction;

                print
                        "Variation ", $tv->variation_feature->variation_name,
                         " allele ", $tva->variation_feature_seq,
                         " has consequence ", join(",", @ensembl_consequences),
                         " (SO ", join(",", @so_consequences), ").";

                if(defined($sift)) {
                        print " SIFT=$sift";
                }
                if(defined($polyphen)) {
                        print " PolyPhen=$polyphen";
                }

                print "\n";
        }
}
```

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

HPG-VARIANT (aka VARIANT)

HPG-VARIANT web site: http://www.opencb.org/projects/hpg/doku.php?id=variant:overview

- Free and open source.
- **3 ways of functionality**: C CLI program, Web application and Java RESTful WS API
- **Cloud** variant annotator. Requires **no installation or updates**
- **Regulatory information**: Conserved genomic regions, TFBSs and miRNA targets. ENCODE DNAse I hypersensitive sites and Histone methylations
- dbSNP and 1000genomes information
- **Phenotypic information**: HGMD, COSMIC and OMIM
- **Cross-link** with many other DDBB (Ensembl, UniProt, PDB, etc)
- Polyphen and SIFT
- **Input**: VCF, GFF and BED. Accepts compressed files in ***tar.gz***
- **11 species** (human, mouse, rat, zebra fish, worm, fly, yeast, dog, pig, mosquito and plasmodium)
- HPG-VARIANT-GWAS to test for **association**

- **Young** program, many **new features coming**
  - PhastCons, GERP
  - Many more species (~25 new species)
  - Large structural variants annotation

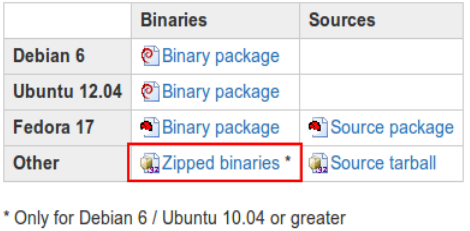
Medina I, De Maria A, Bleda M, Salavert F, Alonso R, Gonzalez CY, Dopazo J. *VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing.* **Nucleic Acids Research**. 2012 Jul;40(Web Server issue):W54-8 Pubmed PMID: 22693211

HPG-VARIANT (CLI Program)

- **Download** the program and save it into your `course/variant_annotation` folder :

  http://wiki.opencb.org/projects/hpg/doku.php?id=variant:downloads

- Extract the contents



	Binaries	Sources
Debian 6	Binary package	
Ubuntu 12.04	Binary package	
Fedora 17	Binary package	Source package
Other	Zipped binaries *	Source tarball

* Only for Debian 6 / Ubuntu 10.04 or greater

- Add the folder to your PATH

```
echo "export PATH=$PATH:/home/Desktop/hpg-variant-1.0" >> ~/.bashrc
source ~/.bashrc
```

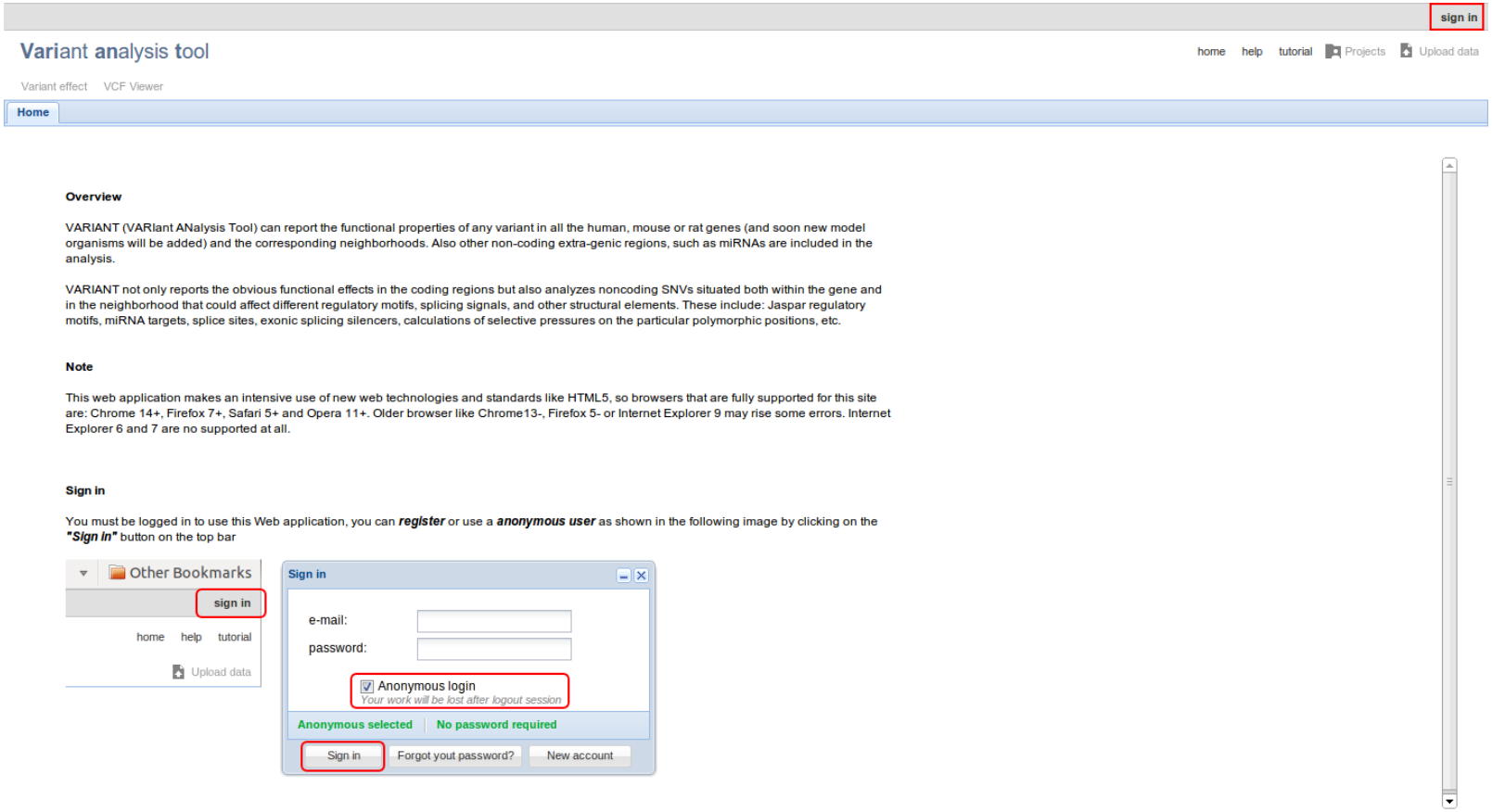
- Usage:

```
hpg-var-effect -v CHB.exon.2010_03.sites.vcf --outdir effect_output/
```

HPG-VARIANT
Web application

- Web application

  http://variant.bioinfo.cipf.es/



THANK YOU.

BierApp

Bierapp.babelomics.org