Ensembl Variation API

Hinxton, 26th of January 2016

Laurent Gil
Ensembl Variation

lgil@ebi.ac.uk

Materials:
http://www.ebi.ac.uk/~lgil/workshops/

EMBL-EBI

wellcome trust sanger institute

Schedule

09:30 - 10:45 Variation

10:45 - 11:15 Coffee break

11:15 - 12:30 Variation

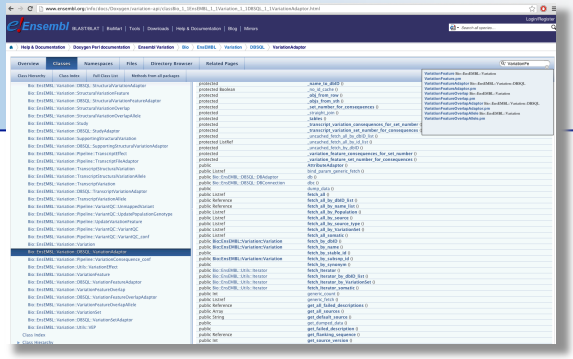
12:30 - 13:30 Lunch break

13:30 - 15:30 Variation

15:30 - 16:00 Tea break

16:00 - 17:00 Variation + VEP

Important URLs



- Links for the course material
  - http://www.ebi.ac.uk/~lgil/workshops/2016-01_Hinxton/
- Ensembl **Variation** API documentation
  - http://www.ensembl.org/info/docs/Doxygen/variation-api/index.html
- Ensembl **Core** API documentation
  - http://www.ensembl.org/info/docs/Doxygen/core-api/index.html
- Variation data documentation:
  - http://www.ensembl.org/info/genome/variation/index.html

Overview

Introduction to the Variation data and API

Key data/objects

- Variation
- Allele
- Genotypes
- Variation feature (location of the variants)
- Structural variation and structural variation feature (location)

Additional information/annotation

- Variation set
- Phenotype data
- Publication data

Analysis

- Variation consequences
- Linkage disequilibrium

Variant Effect Predictor (VEP)

An introduction to genomic variation

Two human genomes differ by one base in every 1000 on average

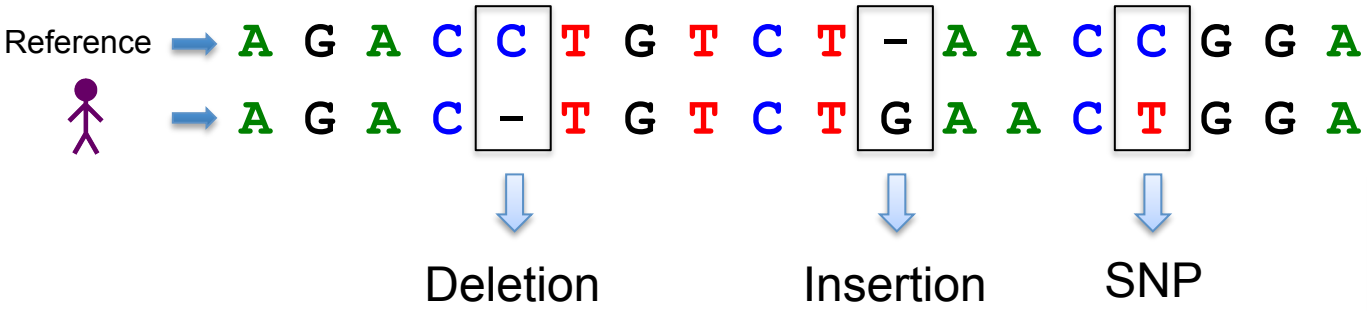
- We study genomic variation to understand:
  - Why some individuals are more susceptible to disease
  - Why some individual have an adverse response to drugs or environmental factors
  - Population history

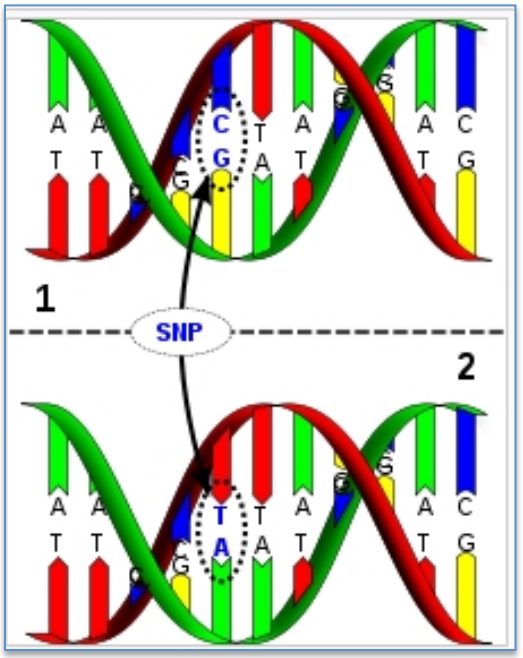
- There are over **150 million** short human variants in dbSNP (v144)
- There are over **3 million** longer (structural) variants in the Database of Genomic Variants archive (DGVa)

Variation types – Short variants

Short scale: one or few nucleotides

- Single nucleotide polymorphism (SNP)
- Small insertions and/or deletions (DIPs or indels)

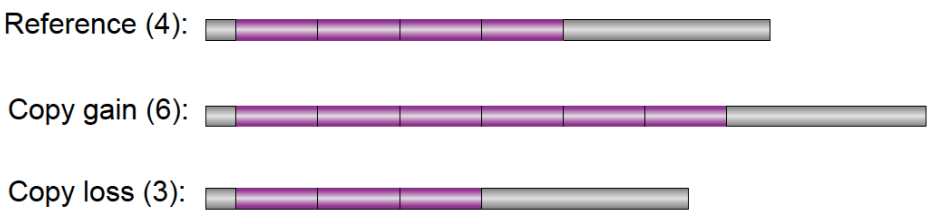




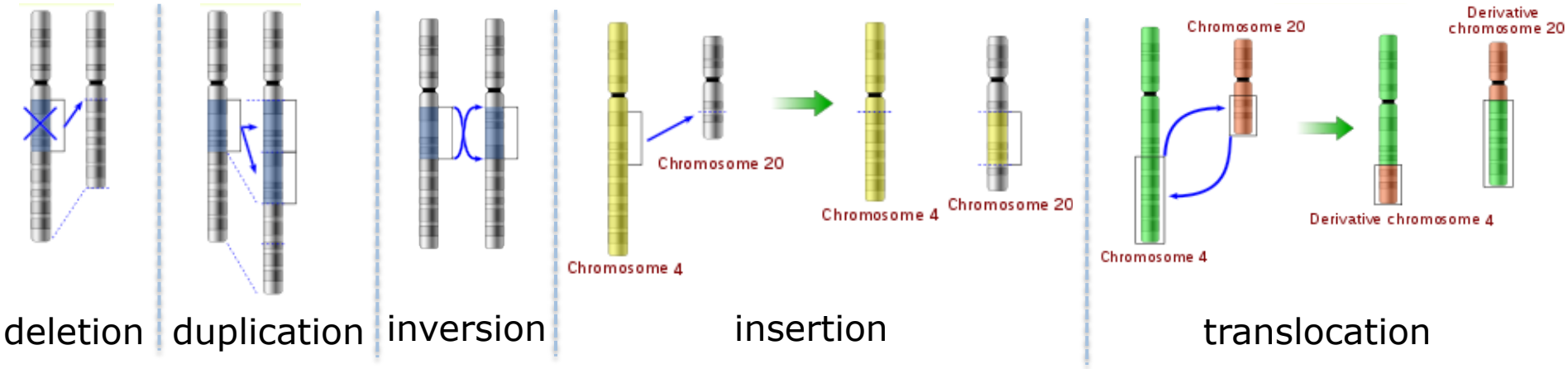
Variation types – Long variants

Large scale: in chromosomal structure (structural variation)

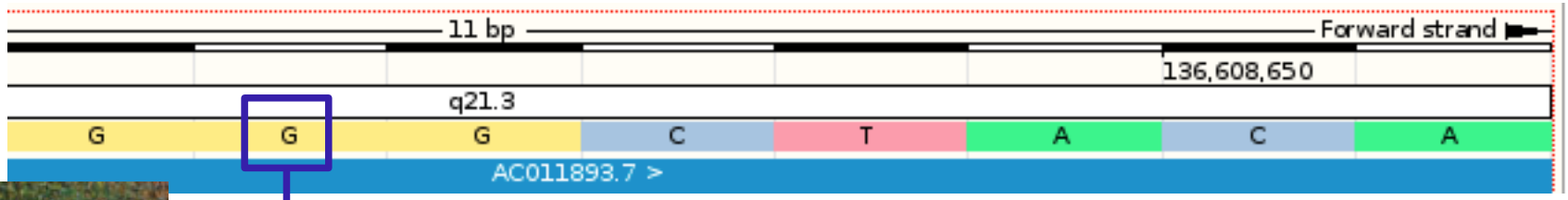
- Copy number variations (CNVs): sequence repeated ‘n’ times in an individual



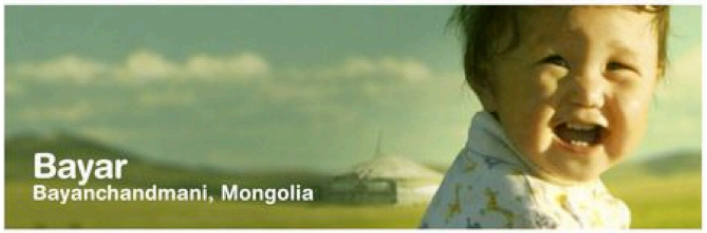
- Large structural variants:

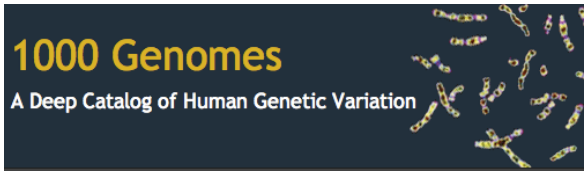


Variation



GAGCTACA
GAGCTACA
GGGCTACA
GGGCTACA





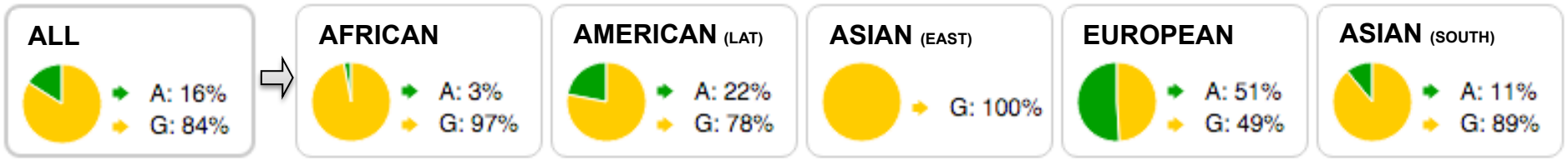
Variation

Lactose Intolerance

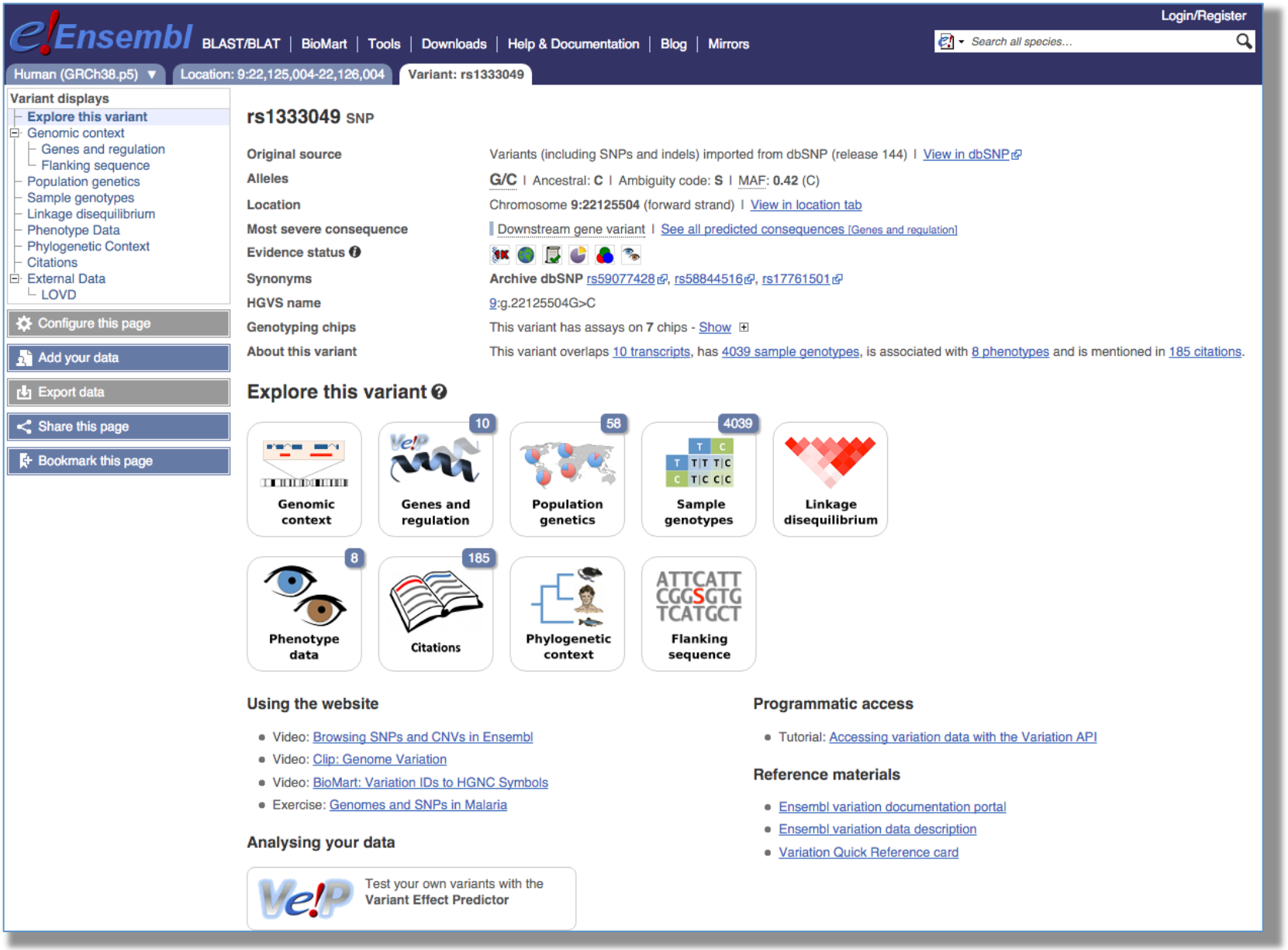
rs4988235: a SNP near the LCT gene controls whether lactase enzyme is turned on or off as a person grows older.

Alleles: G/**A**

1000 Genomes Project Phase 3 allele frequencies



Genotype	What it means
GG	Likely to be lactose intolerant.
G**A**	Likely to be tolerant due to lactase persistence.
AA	



http://www.ensembl.org/Homo_sapiens/Variation/Explore?v=rs1333049

Ensembl Variation - Roles

Data:

- Build the variation databases: import variations from different sources
- Quality control (quality filtering, evidence summary)
- Data annotation: e.g. phenotype data, ancestral alleles, …
- Calculate consequences of the variants on the transcripts
- SIFT and PolyPhen scores

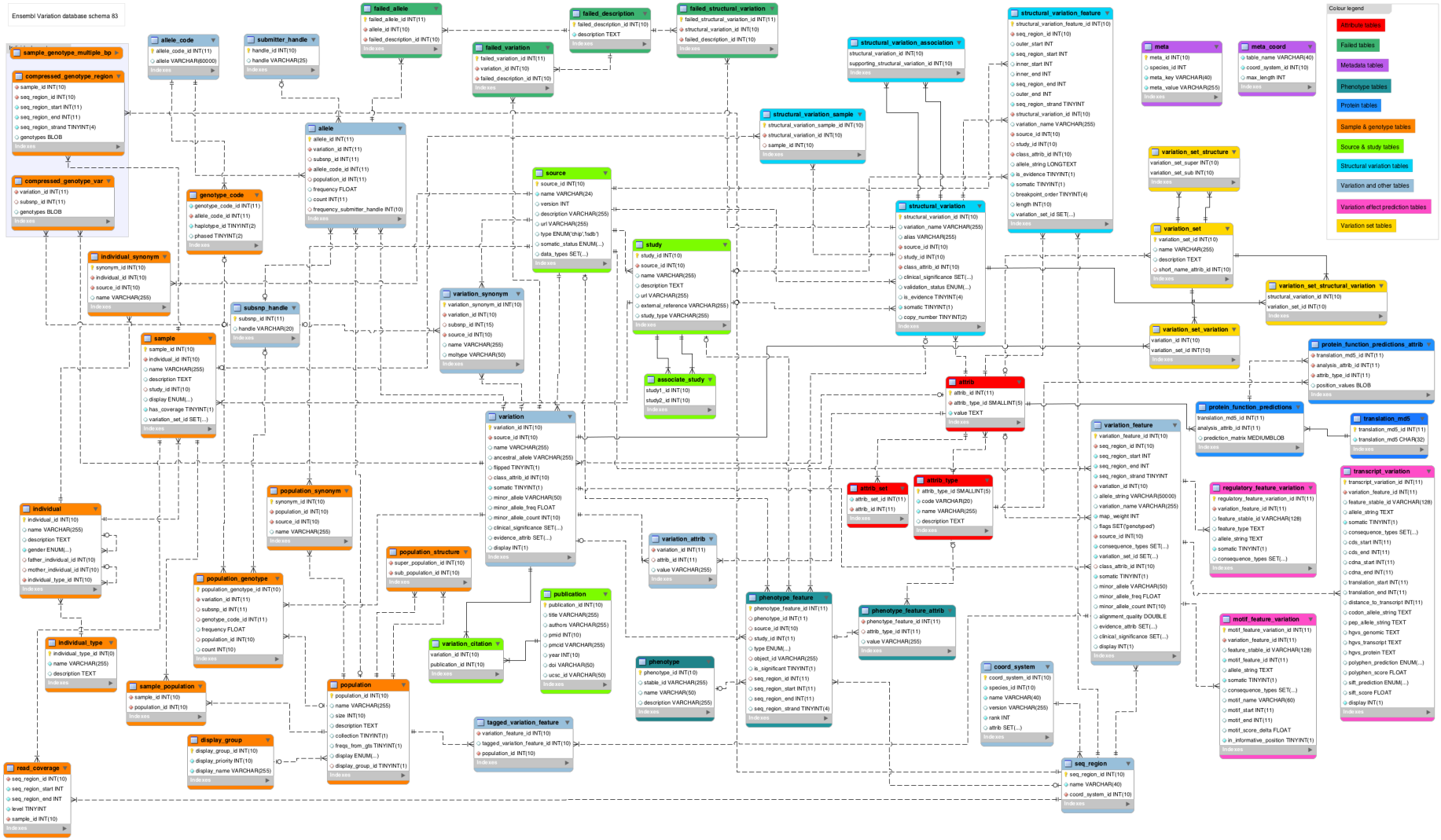
API:

- Build and maintain the Variation API

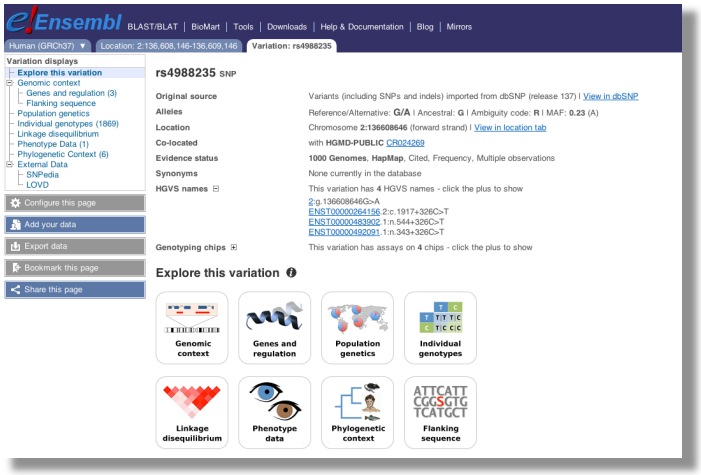
Web:

- Build and maintain part of the Variation Web API

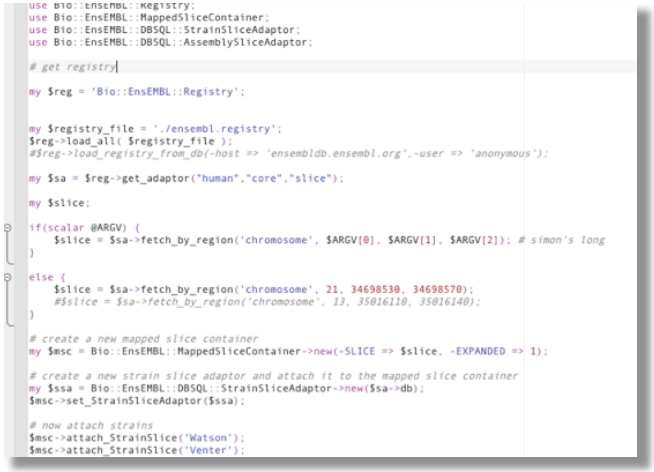
Variation database schema (MySQL)



➡ http://www.ensembl.org/info/docs/variation/variation_schema.html



Website



Scripts

REST
web service

API
perl

Ve!P

Tools

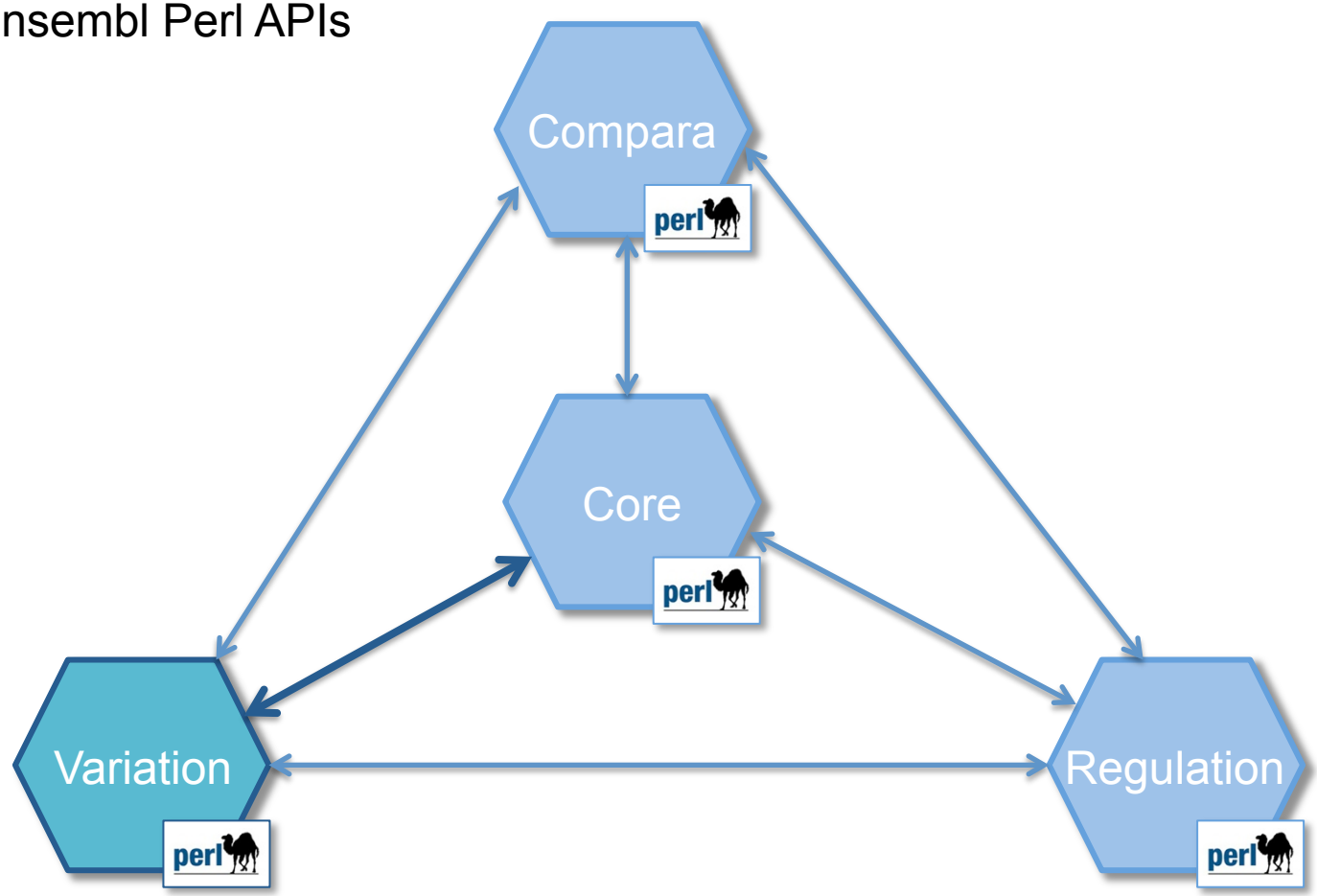
NCBI
dbSNP

Variation
MySQL

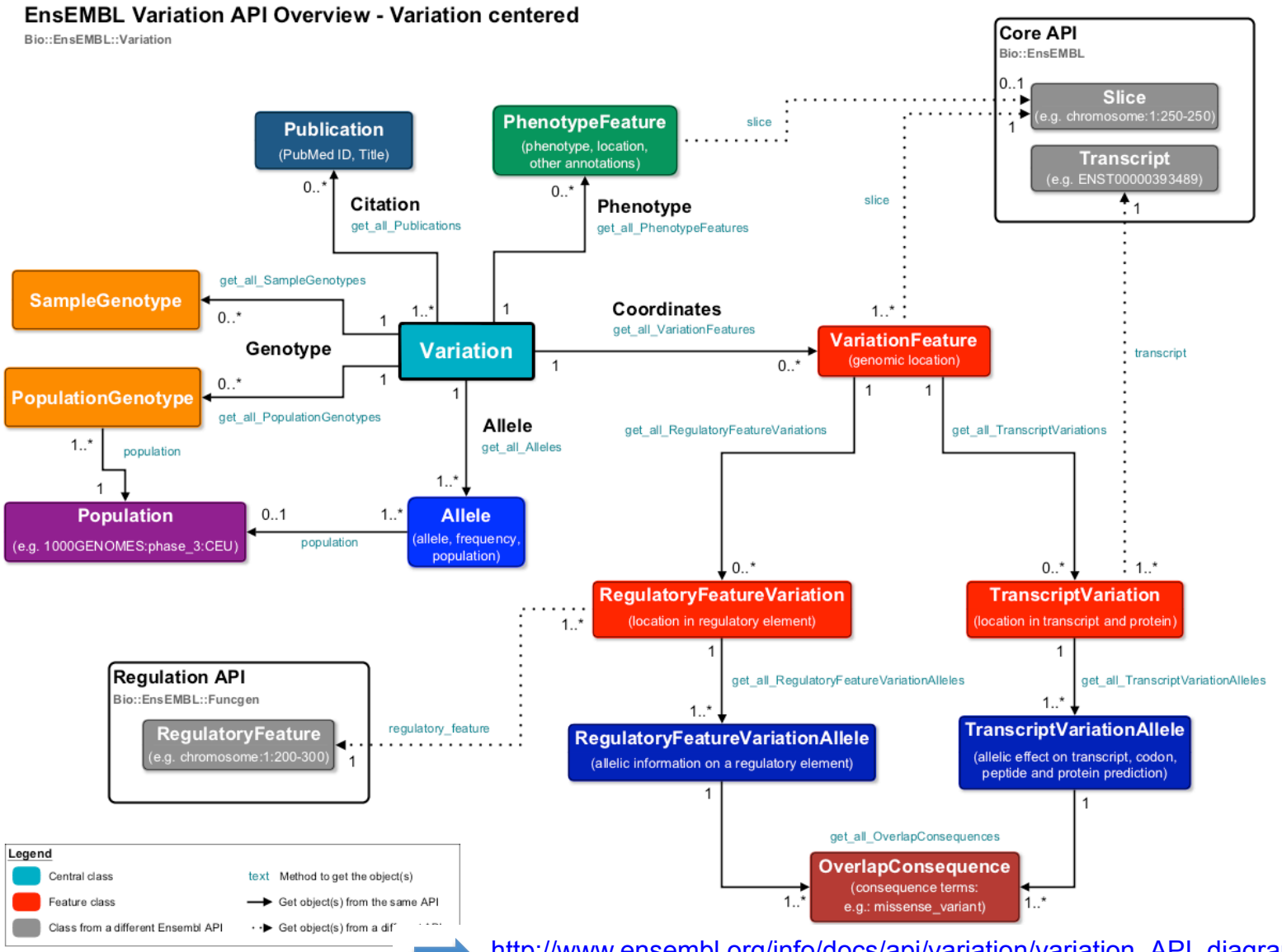
- HGMD-Public
- COSMIC
- OMIM
- UniProt
- NHGRI-EBI
- DGVa
...

About Ensembl API

4 main Ensembl Perl APIs



Simple objects diagram



http://www.ensembl.org/info/docs/api/variation/variation_API_diagram.html

Objects and adaptors

- The API deals with objects representing database entities
- Adaptors are "**factories**" for generating objects
  - Adaptors are retrieved from the Registry

```
use Bio::EnsEMBL::Registry;

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

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

my $va = $reg->get_adaptor('human', 'variation', 'variation');

...

my $variation = $va->fetch_by_name('rs334');
```

Object creation

- Using adaptors
  - **Fetch** object(s) according to some property e.g. name, location
    - "fetch_**all**_..." -> returns a list reference of items
    - "fetch_**by**_..." -> usually returns only 1 item
  - Check documentation which methods the adaptor provides

- Using API objects: e.g. **Variation**
  - **Get** other object(s) from an API object

    e.g. `$variation->get_all_Alleles()`

    returns a list reference of Ensembl **Allele** objects
  - Usually the object is written with a upper case in the method

    e.g. `get_all_Alleles()`

Key data/objects

Variation object

Represents a short difference between at least two genomic sequences

- Basic unit in Ensembl Variation database
- Retrieve using variation adaptor
- May have an evidence class, ancestral allele, clinical significance
- Key attributes:

Attribute	Example value(s)	Method(s)
Variant name	rs1333049, COSM679126	`$v->name()`
Source	dbSNP, COSMIC	`$v->source_name()`
Class	SNP, insertion, deletion	`$v->var_class()`
Minor allele	Allele: C Frequency: 0.43	`$v->minor_allele()` `$v->minor_allele_frequency()`
Supporting evidence	Array: [Frequency, Multiple_observations,etc...]	`$v->get_all_evidence_values()`

Exercise 1

- Retrieve the following information:
  - Variation class
  - Source

- for the following variations in human:
  - rs1333049
  - rs56385407
  - COSM998
  - CI003207

Hint: You will need to use the **VariationAdaptor** and the method "**fetch_by_name**" to retrieve the Variation objects

- Variation API documentation:
  http://www.ensembl.org/info/docs/Doxygen/variation-api/index.html
- Variation tutorial:
  http://www.ensembl.org/info/docs/api/variation/variation_tutorial.html

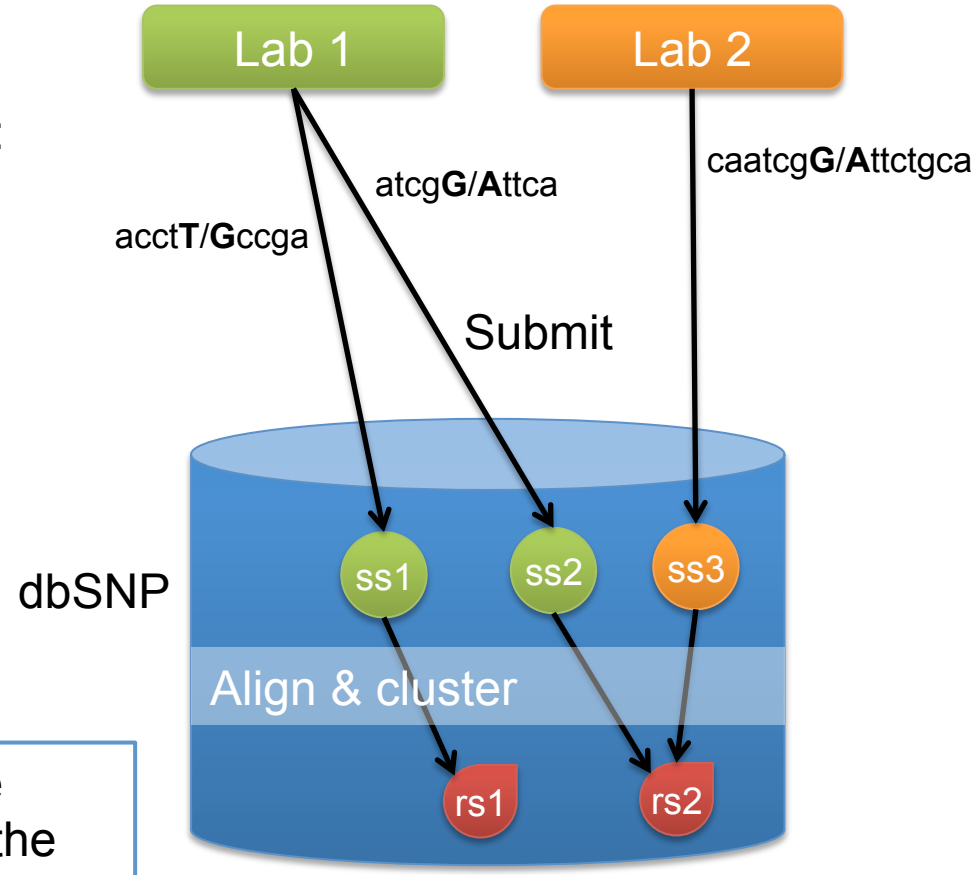
Allele information

Population genetics

A variant at a specific location can have two or more alleles.
Usually a reference allele and at least 1 alternative allele.

Information held against an allele includes:

- Base(s) observed
- Source population
- Submitter information (dbSNP 'handle')
- Frequency in an assayed population



Frequencies for large studies, such as the 1000 Genomes project are calculated on the fly from VCF files

Allele objects

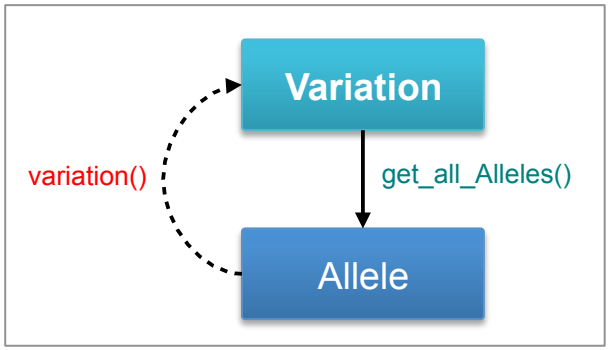
Population genetics

- Represents an allele observed at a variation and usually in a population
- Can be retrieved from **Variation** objects, as well as the **Allele adaptor**

Attribute	Example value(s)	Method(s)	Comment(s)
Allele	G, AC	`$a->allele()`	
Frequency	0.15, 1	`$a->frequency()`	Not always defined
Population	Bio::EnsEMBL::Variation::**Population**	`$a->population()`	Returns an object if defined
Variation	Bio::EnsEMBL::Variation::**Variation**	`$a->variation()`	Returns an object
Submitter id	TSC_CSHL	`$a->subsnp_handle()`	Not always defined

Note: most API objects have a method to retrieve reference to "parent" object
e.g. to retrieve variation object from allele object:

```
$allele->variation()
```



Exercise 2

- For SNP rs1333049 in human, retrieve the following for each of its alleles:
  - Allele
  - Frequency *
  - Population name *
  - Submitter name in dbSNP ("handle") *

* if exist

Note:

Phase 3 data from the 1000 Genomes Project is accessed directly from VCF files.

It is not default API behaviour to read VCF's, so to see frequencies for this project you will need to set the '**use_vcf**' option on the database adaptor to 1.

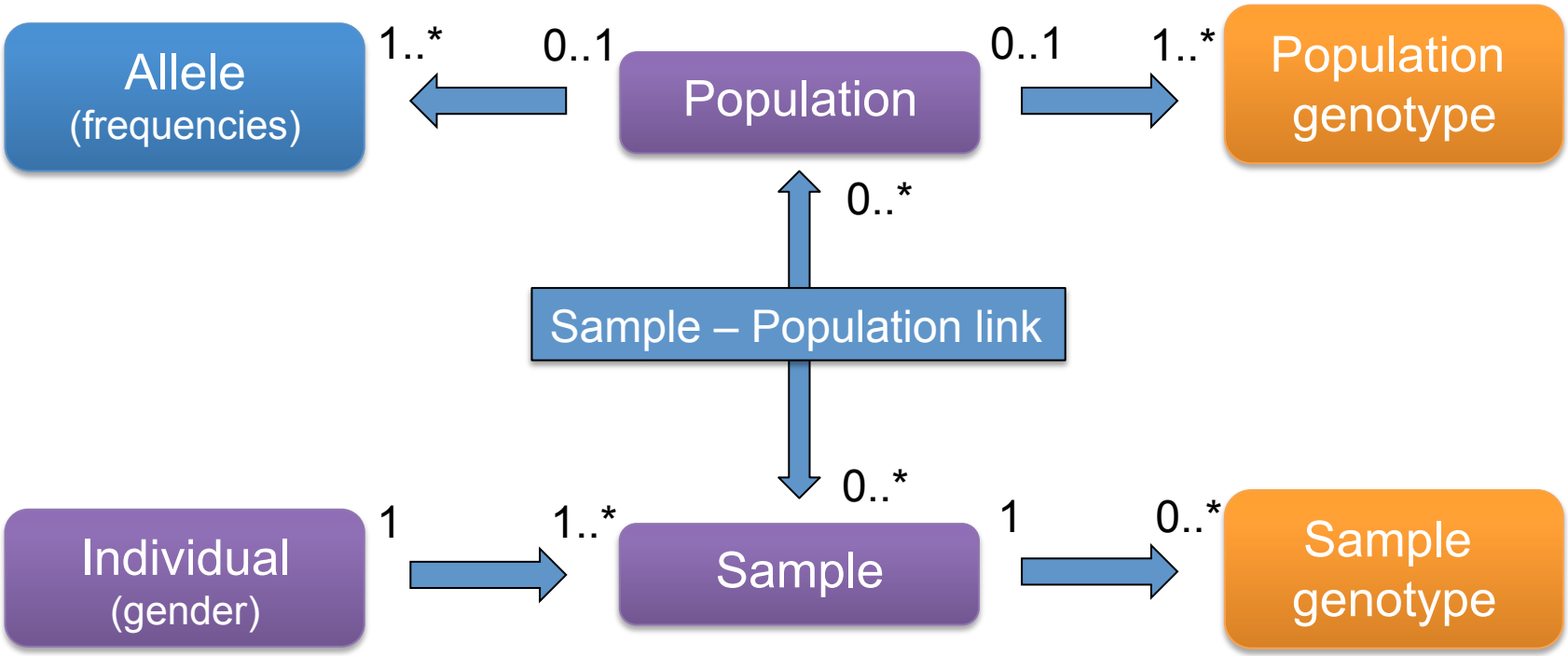
```
$va->db->use_vcf(1);
```

Hint:

You may need to test whether the Population object returned is empty to avoid a script error.

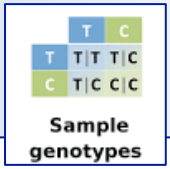
Sample, Individuals & Populations

A population is a group of samples used in a study, rather than a geographical grouping



Experimental results are attached to samples, which are taken from individuals; this supports results from multiple samples from the same individual

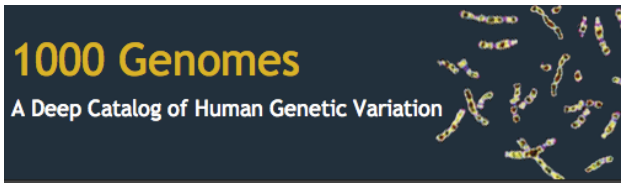
Sample Genotypes



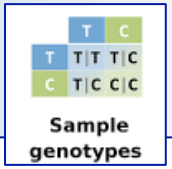
Sample genotypes show which combination of the possible alleles a specific sample from a specific individual has at the site of a specific variant.

We import genotype data from dbSNP and large scale sequencing project such as the 1000 Genomes project, the WTSI Mouse Genomes Project and the NextGen Project.

Genotype data is large, so it is stored either in VCF files or in a compressed format in the Ensembl databases. It cannot be accessed using direct SQL



Sample Genotype object

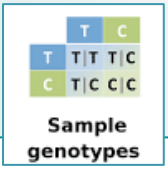


Represents an instance of the genotype of a variant in a specific sample from an individual

Can be retrieved from **Variation** object, as well as the **SampleGenotype adaptor** (Bio::EnsEMBL::Variation::DBSQL::**SampleGenotypeAdaptor**)

Attribute	Example value(s)	Method(s)
Genotype	List reference [A,C]	`$sg->genotype()`
Genotype string	A|C	`$sg->genotype_string()`
Variation object	Bio::EnsEMBL::Variation::**Variation**	`$sg->variation()`
Sample object	Bio::EnsEMBL::Variation::**Sample**	`$sg->sample()`
Individual object	Bio::EnsEMBL::Variation::**Individual**	`$sg->sample->individual()`

Exercise 3



- Fetch all available individual genotypes for the human variation rs1333049.
- Extract:
  - The genotype string
  - The name of the sample
  - The gender of the individual

Note:

Phase 3 Data from the 1000 Genomes Project is accessed directly from VCF files.

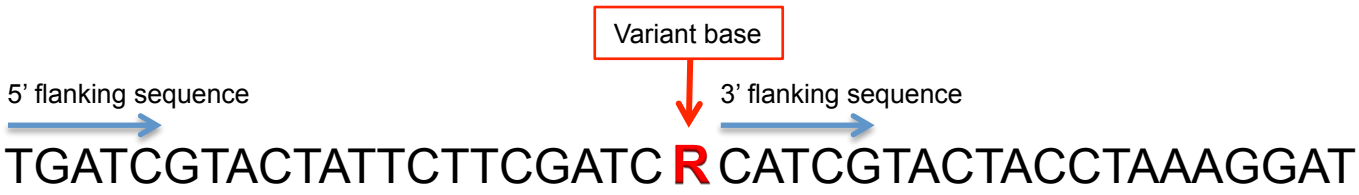
It is not default API behaviour to read VCF's, so to extract genotypes from this project you will need to set the '**use_vcf**' option on the database adaptor to 1:

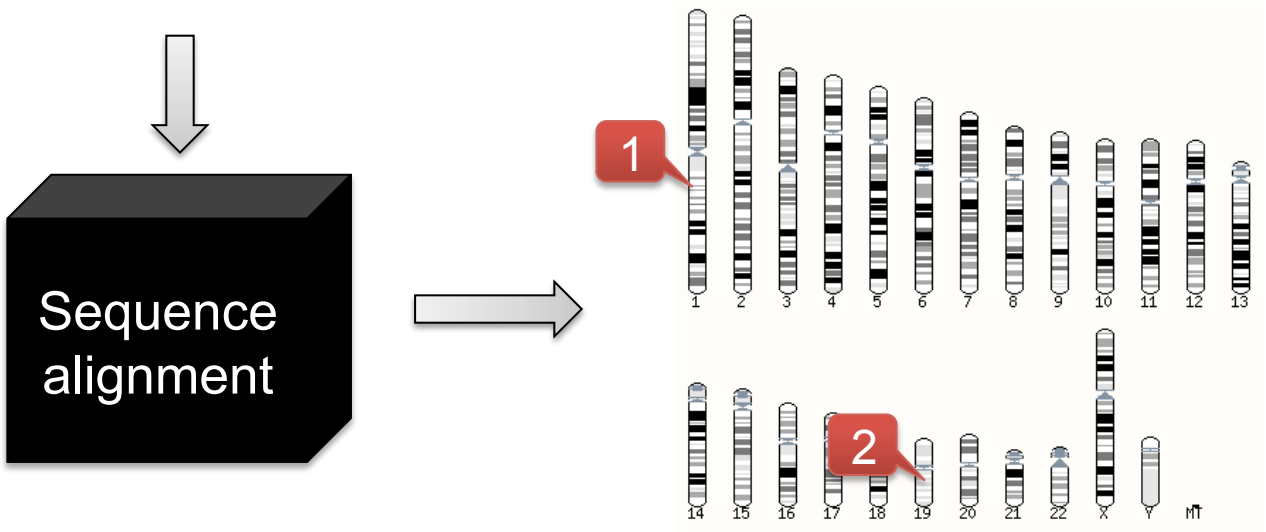
```
$va->db->use_vcf(1);
```

Variation mapping

Variations are mapped to the genomic reference sequence using their flanking sequences

- A variation may have multiple mappings on the current reference
- A variation may not map reliably to the current reference





Variation feature object

- An instance of a variation mapping to the genome
- Can be retrieved from **Slice** (Core) and **Variation** objects, as well as **VariationFeature adaptor**

Attribute	Example value(s)	Method(s)	Comment(s)
Allele string	A/G, -/C	`$vf->allele_string()`	
Chromosome	15, X	`$vf->seq_region_name()`	
Coordinates	103019234	`$vf->start()` `$vf->end()`	slice relative
		`$vf->seq_region_start()` `$vf->seq_region_end()`	chromosome relative
Slice object	Bio::EnsEMBL::**Slice**	`$vf->slice()`	Returns a Core API object

Note: Allele strings are reported as Reference/Alternative where possible:

A/G → "reference/alternative"

Exercise 4

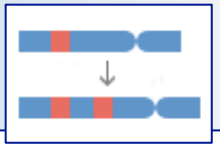
- Retrieve all variations in human located on chromosome 13 from 48987260 to 48990000 and get:
  - Variation name
  - Alleles (“allele_string”)
  - Location (e.g. Chromosome:Start-End)

- Extra:

  Find the the genomic location(s) of the following human variants:
  - rs7107418
  - rs671
  - rs17646946
  - rs4988235

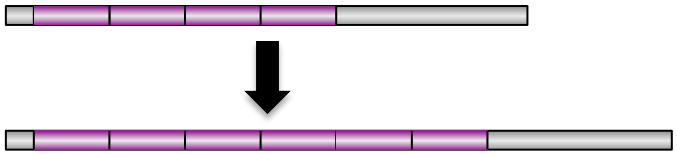
Hint: usually when you have to search something by location, you need to use the **Slice** object (Core API) first

Structural variation



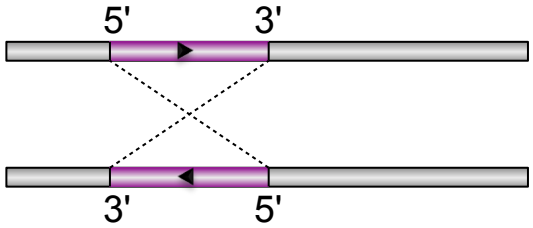
- Usually represents a large segment of variable DNA sequence (> 1kb)
- Different processes give rise to different types of structural variations:

Copy number variation (CNV)

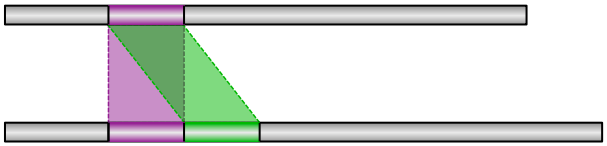


→ Allele type “Gain”

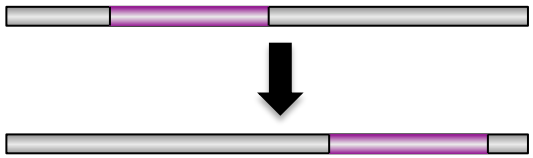
Inversion



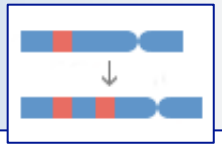
Duplication



Translocation



Example of structural variation



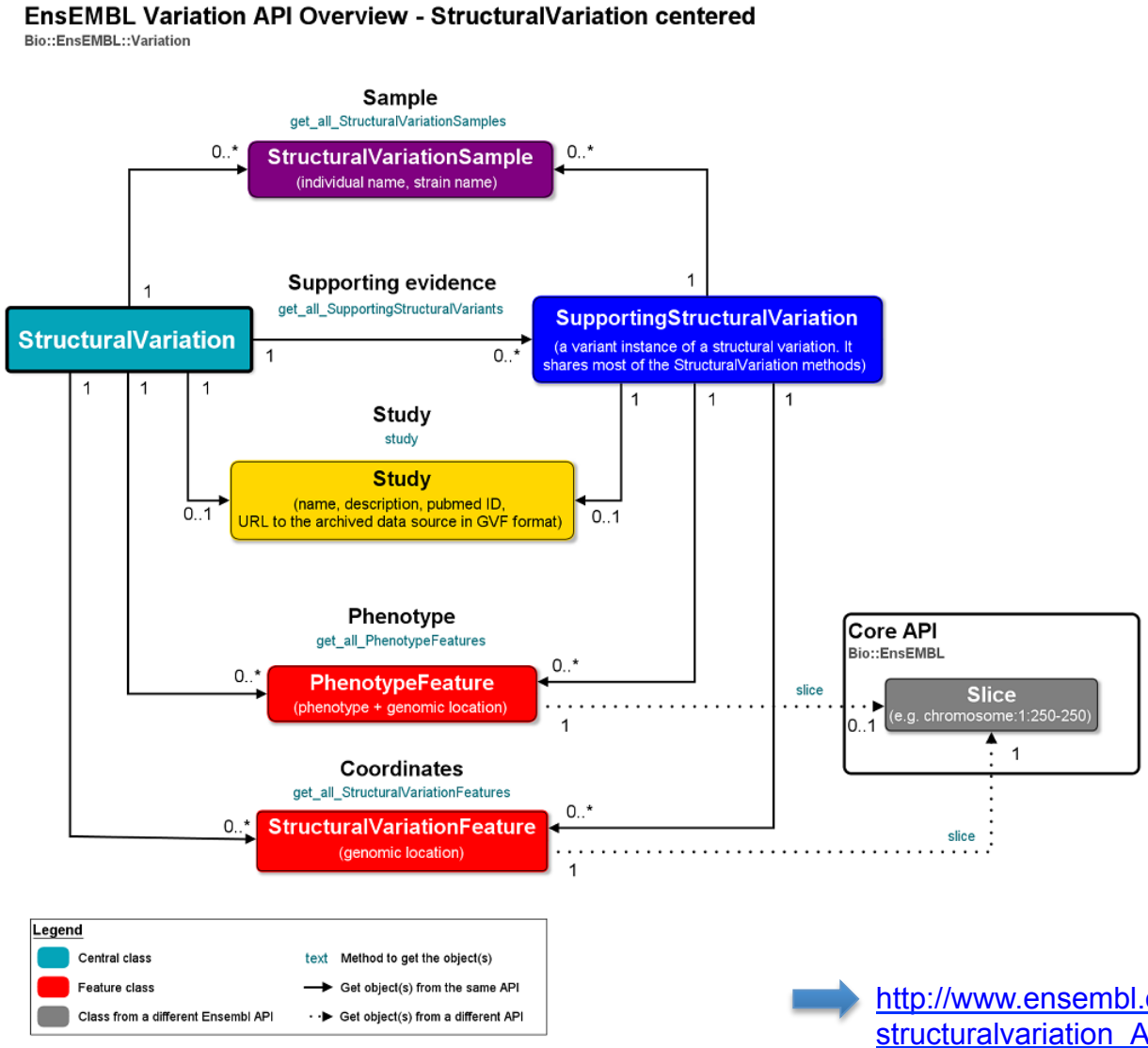
Structural variant: nsv917561

Variation class	CNV (SO:0001019)
Source	DGVa – Database of Genomic Variants Archive
Study	**nstd75** – "International Standards for Cytogenomic Arrays Consortium (prenatal dataset)." PMID:21844811, PMID:20466091
Location	Chromosome **1:146987841-148359881** (forward strand)

Supporting evidence

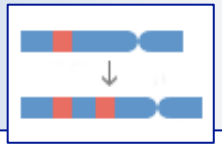
Supporting evidence	Chr:bp (strand)	Allele type	Clinical significance	Individual name
nssv1605931	1:146987841-148359881 (+)	Gain	pathogenic	ISCA_ID_pn_1622
nssv1606072	1:146987841-148359881 (+)	Gain	-	ISCA_ID_pn_1746

Simple objects diagram



http://www.ensembl.org/info/docs/api/variation/structuralvariation_API_diagram.html

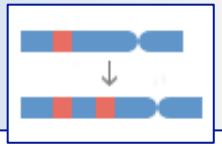
Structural variation object



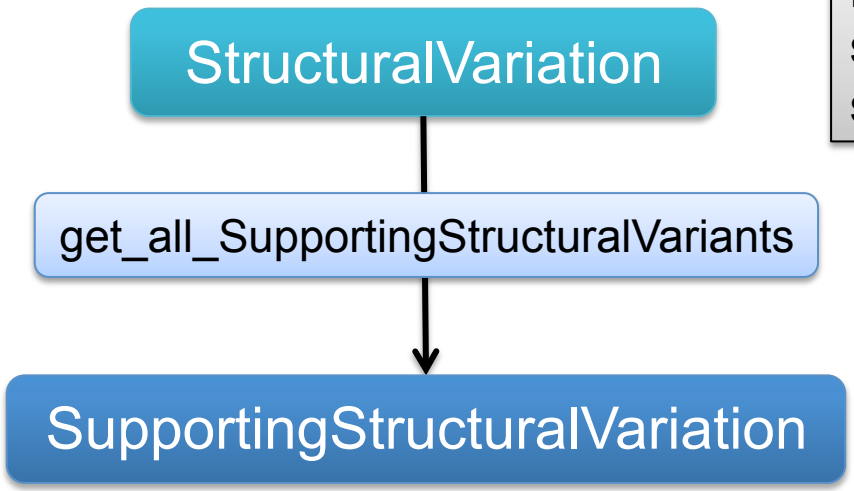
- Similar to the **Variation** object, but represents longer variants
- Data from:
  - DGVa (with study and supporting evidences)
  - CNV probes from Illumina and Affymetrix

Attribute	Example value(s)	Method(s)	Comment(s)
Structural variation name	esv214236	`$sv->variation_name()`	
Class	CNV	`$sv->var_class()`	
Study	Bio::EnsEMBL::Variation::**Study**	`$sv->study()`	Returns an object if defined
Supporting evidence(s)	Reference list of objects Bio::EnsEMBL::Variation:: **SupportingStructuralVariation**	`$sv->get_all_SupportingStructuralVariants()`	

Supporting structural variation



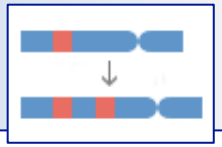
Supporting evidence <=> Allele for a Variation (i.e. experimental data)



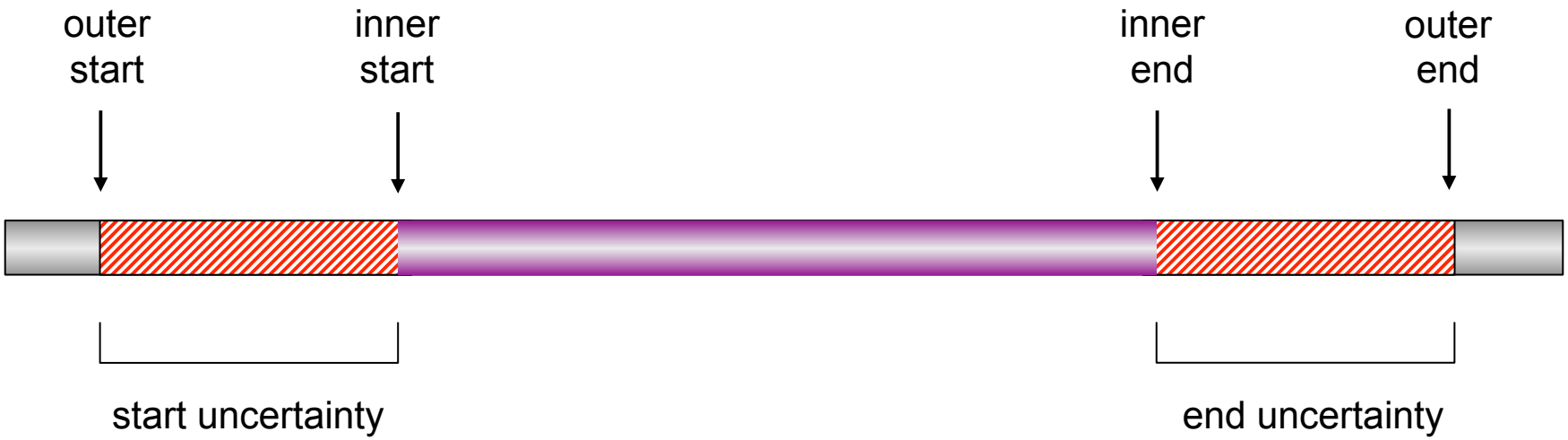
Region of the genome that a submitter has defined as containing structural variations.

Actual **variant calls** that were made within a study.

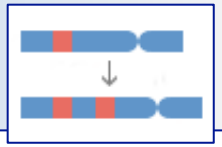
Structural variation coordinates



- Usually a structural variation has only a start and an end
- But sometimes, the breakpoint locations cannot be determined precisely



Structural variation feature object



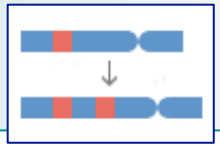
- Similar to **VariationFeature** but represent structural variations
- Can be retrieved from a **Slice** (Core) as well as the **StructuralVariationFeature adaptor** and the **StructuralVariation** object
- No allele string, alleles or associated variation
- Only coordinates, class and the list of the supporting evidences

Attribute	Example value(s)	Method(s)
Structural variation name	esv214236	`$svf->variation_name()`
Chromosome	15, X	`$svf->seq_region_name()`

Coordinates		`$svf->start()`
		`$svf->end()`
		`$svf->seq_region_start()`
		`$svf->seq_region_end()` — "Classic" coordinates
		
		`$svf->outer_start()`
		`$svf->inner_start()`
		`$svf->inner_end()`
		`$svf->outer_end()` — SV specific coordinates

Class	CNV	`$svf->var_class()`

Exercise 5



- Fetch the following information for the structural variation esv275264 in Human :
  - Structural variation class
  - Study name and description
  - Coordinates

- Fetch the names and the classes (sequence ontology term) of its supporting structural variations.

Additional information/annotation

Variation Sets

Variation sets are arbitrary grouping of variations

- Useful for limiting scripts to specific subsets of data
- May group data from multiple sources
- May be arranged in a parent set/subset formation

Parent set	Set
All phenotype-associated variants	HGMD-PUBLIC variants, ClinVar , OMIM, NHGRI-EBI GWAS catalog...
1000 Genomes 3 - All	1000 Genomes 3 - AFR – common, 1000 Genomes 3 – EUR...
Genotyping chip variants	Affy GeneChip 500K, Illumina_HumanOmni2.5 Illumina_Cardio-Metabo_Chip
	ESP_6500

http://www.ensembl.org/info/genome/variation/data_description.html#variation_sets

Variation Sets

- Arbitrary collections of **Variations** or **Structural variations**
- Useful to limit your scripts to important subsets
- Sometimes contain millions of **Variations**, in which case you can fetch an Iterator instead of a list

Attribute	Example value(s)	Method(s)
Description	Variations called by the 1000 Genomes Project…	`$vs->description()`
Members	List of Variations Iterator over Variations	`$vs->get_all_Variations()` `$vs->get_Variation_Iterator()`
Sub/Super sets	Bio::EnsEMBL::Variation::**VariationSet**	`$vs->get_all_sub_VariationSets()` `$vs->get_all_super_VariationSets()`

Phenotype data

ClinVar

OMIM®

DGVarchive

NHGRI-EBI GWAS Catalog

COSMIC
Catalogue of somatic mutations in cancer

europhenome

european genome-phenome archive

HGMD®

DECIPHER
GRCh37

IMPC
INTERNATIONAL MOUSE PHENOTYPING CONSORTIUM

GIANT CONSORTIUM

orphanet

dbGaP
GENOTYPES and PHENOTYPES

ZFIN

RGD

LOVD
Leiden Open Variation Database

AnimalQTLdb

MAGIC

UniProt

OMIA - ONLINE MENDELIAN INHERITANCE IN ANIMALS

GEFOS

Variant

Gene

Structural Variant

QTL

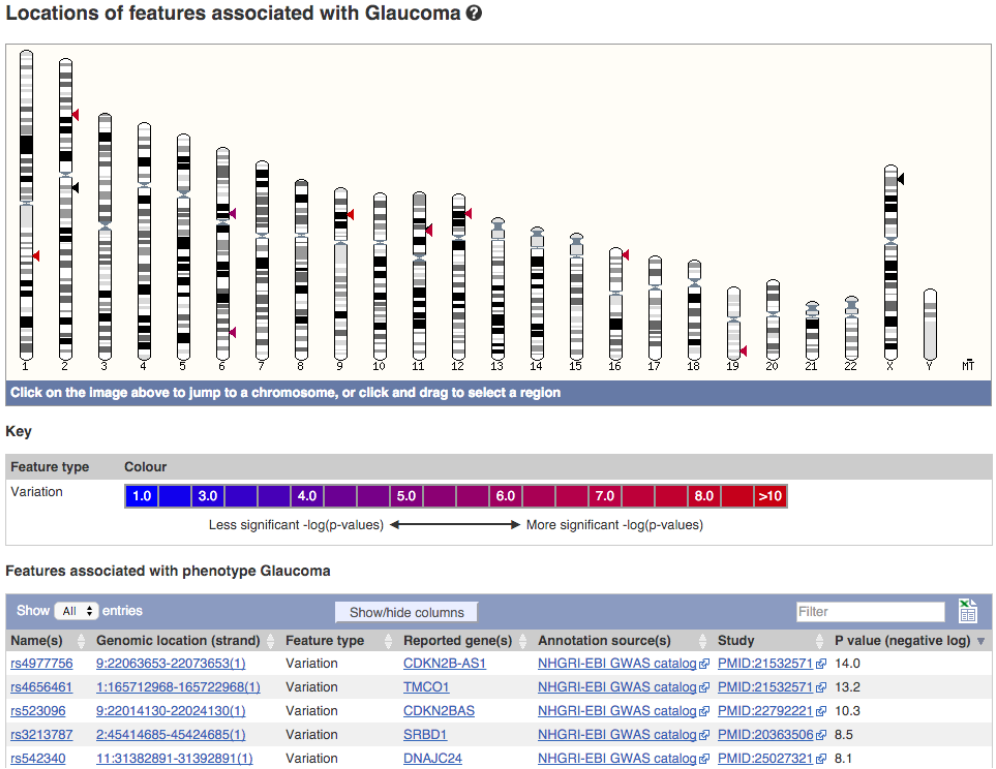
Phenotype data

Phenotypes are mapped to genomic locations via the variations or genes they have reported associations with to create a **PhenotypeFeature**

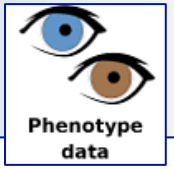
We are collaborating with groups working on phenotype ontologies, but the data current held is as it appears in the source database.

Attributes types held for phenotype features include:

- Clinical significance from ClinVar
- Reported genes
- Most associated risk allele
- P-value
- Inheritance type
- Odds ratio
- Beta coefficient



Phenotype Feature object



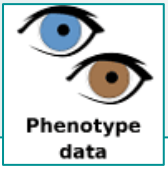
Represents an association between a phenotype an a genomic feature

- Can be retrieved from a **Slice** (Core) or a **Variation** object as well as its **PhenotypeFeature adaptor**
- Not all the data has the same attributes populated
- Phenotype + genomic location

Attribute	Example value(s)	Method(s)
Phenotype object	Bio::EnsEMBL::Variation::**Phenotype**	`$pf->phenotype()` *returns object
Reported gene(s)	CDKN2BAS, NOTCH2	`$pf->associated_gene()`
p-value	6e-07	`$pf->p_value()`
Risk allele	C	`$pf->risk_allele()`
Chromosome	15, X	`$vf->seq_region_name()`
Coordinates	103019234	`$vf->start()` `$vf->end()` } slice relative `$vf->seq_region_start()` `$vf->seq_region_end()` } Chromosome relative

To get the phenotype name: `$pf->phenotype->description()`

Exercise 6



Find all phenotypes associated with the human variants **rs6495122** and **rs2470893** and extract:

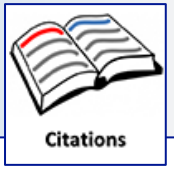
- The phenotype description
- The source name
- The p-value for the association
- The reported risk allele
- Any reported gene(s)

Extra

For each variant find:

- The minor allele and its frequency (this is calculated from the full set of 1000 Genomes samples)
- The ancestral allele (this is calculated by aligning 8 primate assemblies)

Variation citation data



We import variation citation data from:

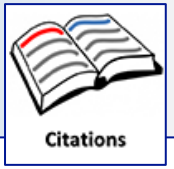
- **dbSNP**
  - Publication information is submitted with the variant submission.
- **EuropePMC**
  - Publications for which the full text is freely available in PubMed Central are mined for refSNP identifiers.
- **UCSC Genocoding Project**
  - Publications for which the text is freely available of those from publishers who have agreed to data mining are mined for refSNP identifiers.

In Ensembl 83, **9** species have citation data.

For human we have:

- 156,000 distinct variants associated with at least 1 publication
- 53,000 distinct publications

Publication object



- Represents an article in a journal
- Can be retrieved from a **Variation** object as well as the **Publication adaptor**
- The type of information available will depend on publisher and data source

Attribute	Example value(s)	Method(s)
Cited variants	List reference of Bio::EnsEMBL::Variation::**Variation** objects	`$p->variations()` *returns listref of objects
Title	Mitochondrial acetylation and diseases of aging.	`$p->title()`
Authors	Wagner GR, Payne RM.	`$p->authors()`
Year	2011	`$p->year()`
PubMed Identifier	21437190	`$p->pmid()`

Exercise 7

Obtain Variation objects for the human variations:

- rs3730070
- rs11765954
- rs671

Find all articles we hold which cite these variants and report :

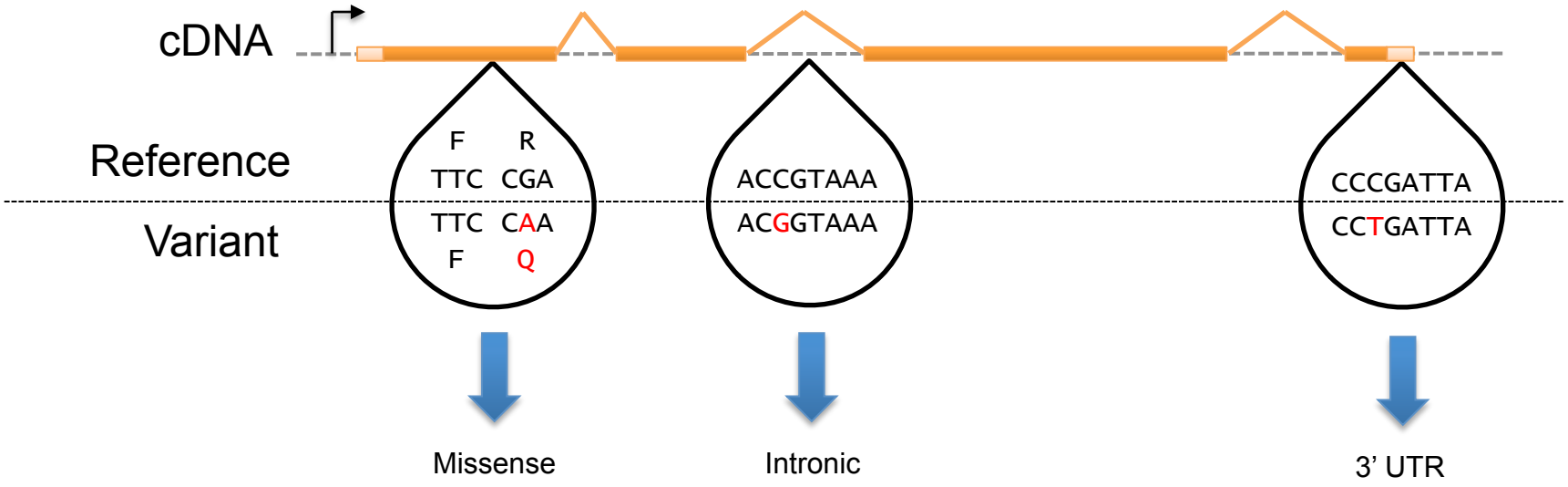
- The Pubmed ID
- The year of publication
- The title

Extra:
Find all variants cited in the publication with the PMID:18187665

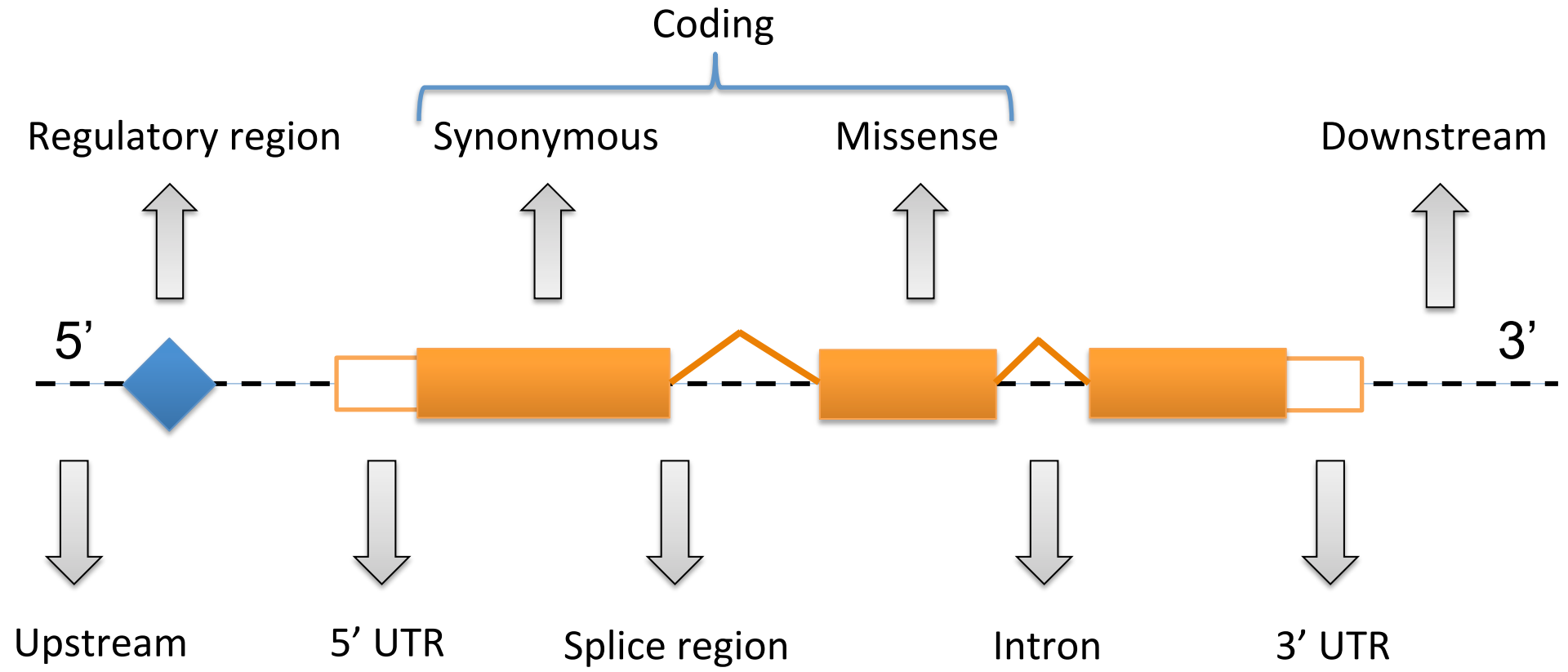
Analysis

Variation consequences

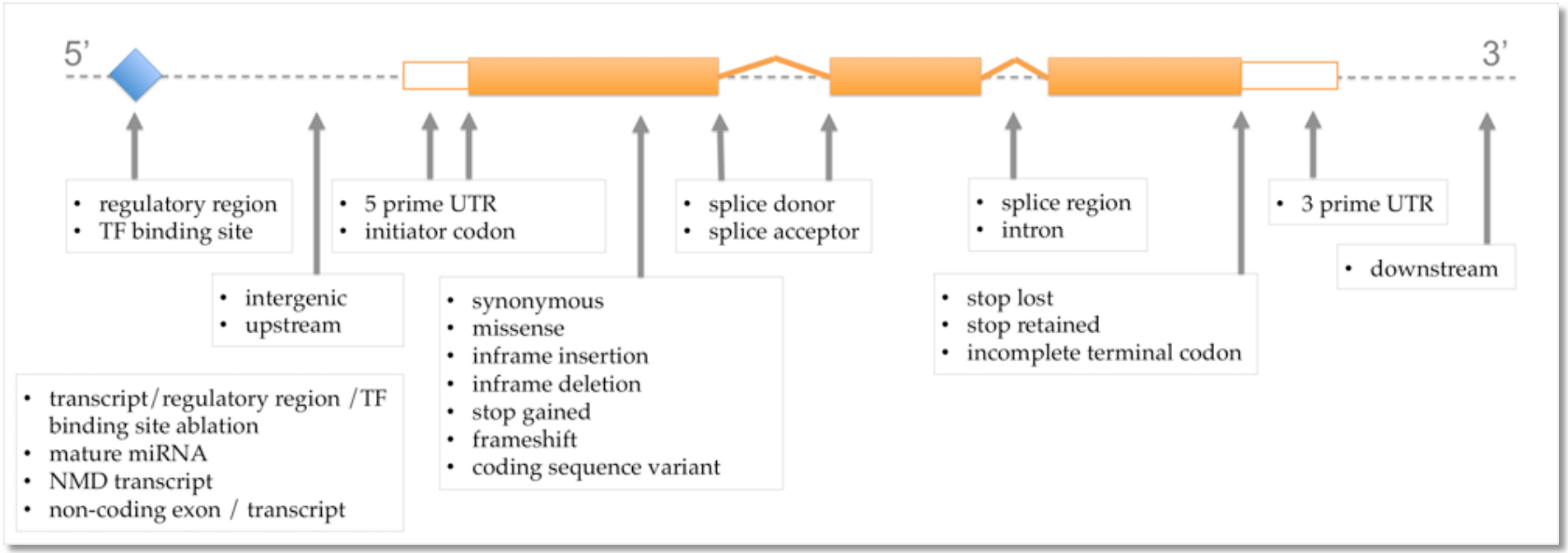
Defines consequence of a variation in relation to the transcript structure it overlaps



Variation consequences



Variation consequences



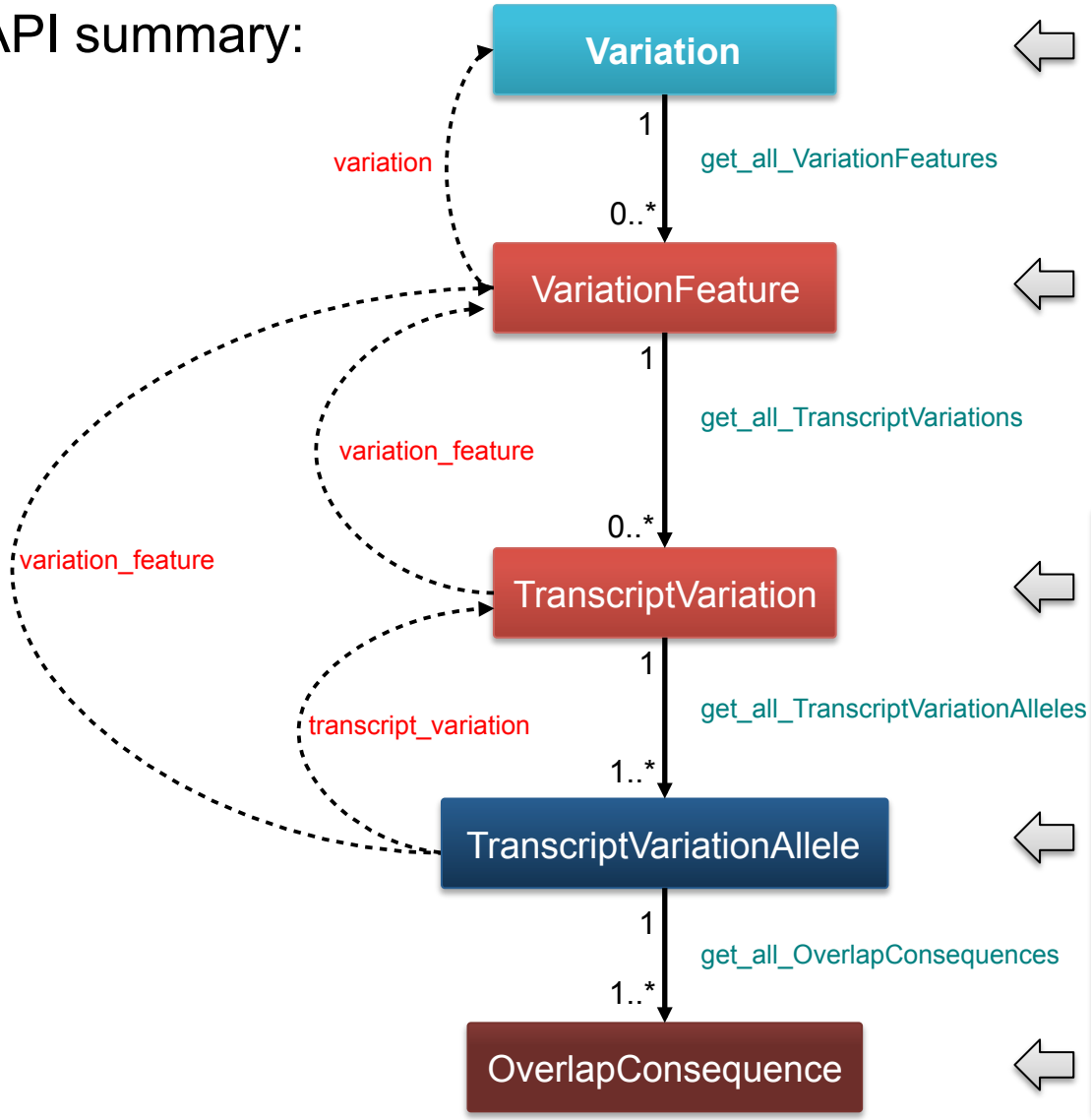
Sequence ontology provides controlled vocabulary to describe consequences

SO The Sequence Ontology Project

http://www.ensembl.org/info/genome/variation/predicted_data.html#consequence_type_table

Transcript variation

API summary:



Variant information

Variant information at the genomic level

VariationFeature information at the transcript level

Specific allele in the transcript and its consequence(s)

Consequence(s) terms

Transcript variation object

- Transcript variation = an instance of a variation feature inside or near a transcript (i.e. can be more than one per variation feature → e.g. alternative splicing)
- Most "severe" consequence string is stored in the VariationFeature object:

```
$vf->display_consequence()
```

- Can be retrieved from the **VariationFeature** objects as well as its object adaptor

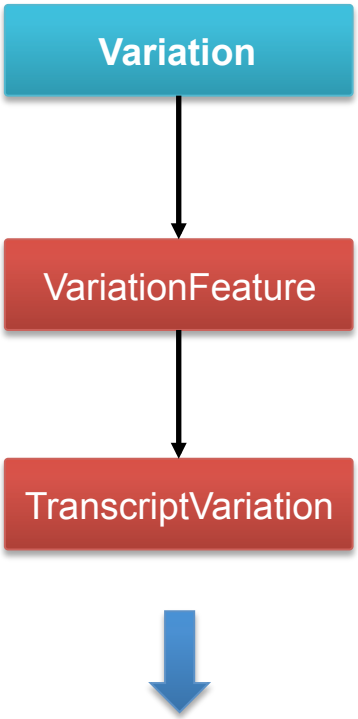
Attribute	Example value(s)	Method(s)	Comment(s)
Transcript	Bio::EnsEMBL::**Transcript**	`$tv->transcript()`	Returns a Core API object
Consequence type	intron_variant, stop_lost	`$tv->display_consequence()` `$tv->consequence_type()`	Only the most severe type All types in an array

Coordinates		`$tv->cdna_start()`
		`$tv->cdna_end()`
		`$tv->cds_start()`
		`$tv->translation_start()`	

Amino acid	F/L, */W	`$tv->pep_allele_string()`	

Transcript variation

Example:



Name: rs201036189

Location: 1:206342907
Allele string: C/T

Transcript: ENST00000573034
Codon: Cgg/Tgg
Amino acid string: R/W

Transcript (strand)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons
ENST00000573034 (+) biotype: protein_coding	Missense variant	384	322	108	R/W	CGG/TGG

Transcript variation allele object

- Consequences are actually established at the allele level
- These represent the most specific information available

```
$transcript_variation->get_all_TranscriptVariationAlleles()
```

- All the consequences of a **TranscriptVariationAllele** are represented as a list of **OverlapConsequence** objects

Attribute	Example value(s)	Method(s)
TranscriptVariation	Bio::EnsEMBL::Variation::**TranscriptVariation**	`$tva->transcript_variation()` *NB returns an object
OverlapConsequences	Reference list of Bio::EnsEMBL::Variation::**OverlapConsequence**	`$tva->get_all_OverlapConsequences()` *NB returns a list of objects

HGVS notation at various levels	1:g.206516261C>T
	ENST00000295713.5:c.62C>T
	ENSP00000295713.5:p.Arg22Trp
		`$tva->hgvs_genomic()`
		`$tva->hgvs_coding()`
		`$tva->hgvs_protein()`

Affected codons	Cgg/Tgg	`$tva->display_codon_allele_string()`
SIFT and PolyPhen predictions	string	`$tva->sift_prediction()` `$tva->polyphen_prediction()`

Protein function predictions

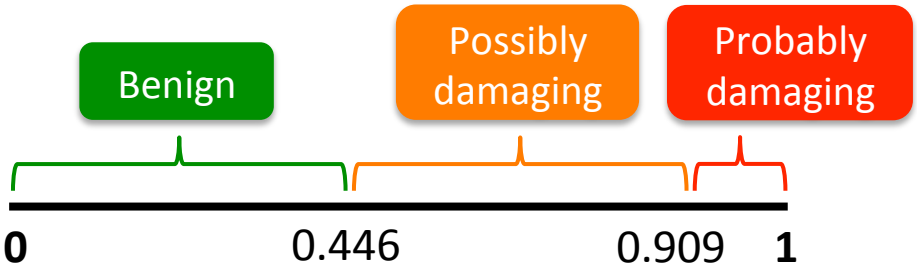
We run tools to identify non-synonymous mutations that are likely to affect protein function

SIFT

- Uses sequence homology and amino acid similarity to calculate if the substitution is: *tolerated* (>0.05) or *deleterious* (<=0.05)
- Supported species: Human, Mouse, Chicken, Cow, Dog, Horse, Pig, Rat, Sheep, Zebrafish

PolyPhen

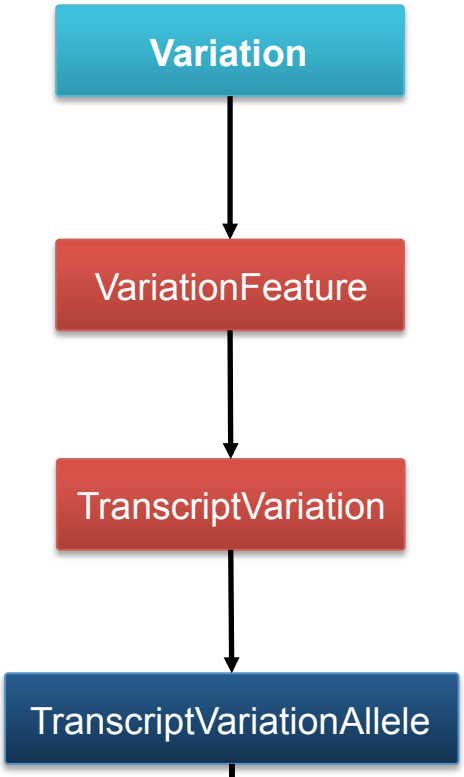
- Uses sequence homology, PDB 3D structures, Pfam annotation, etc. to predict if a substitution is: *benign*, *possibly damaging*, *probably damaging* or *unknown*
- Supported species: human only



- Methods on a *TranscriptVariationAllele*:
  - `$tva->sift_prediction()`
  - `$tva->polyphen_prediction()`
  - `$tva->sift_score()`
  - `$tva->polyphen_score()`

Transcript variation allele

Example 1:
with 2 alleles



Name: rs201036189

Location: 1:206342907
Allele string: C/T

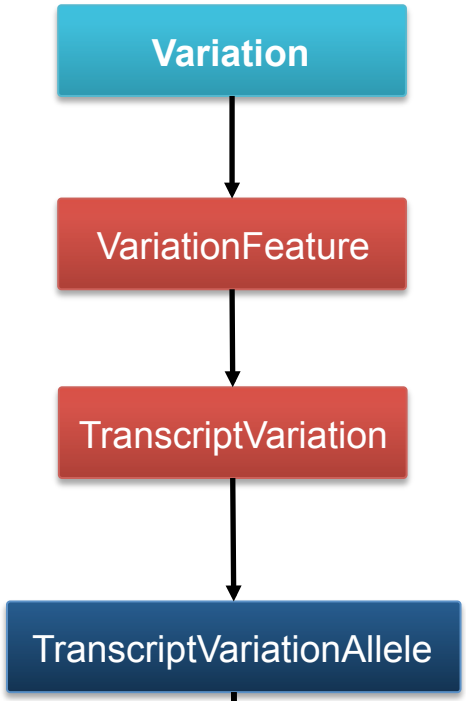
Transcript: ENST00000573034
Codon: Cgg/Tgg
Amino acid string: R/W

Allele: T (alternative allele)

Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen
ENST00000573034 (+) biotype: protein_coding	T (T)	Missense variant	384	322	108	R/W	CGG/TGG	0	0.998

Transcript variation allele (2)

Example 2:
with 3 alleles



Name: rs191084422

Location: 6:56553483
Allele string: C/G/T

Transcript: ENST00000293925
Codons: gCa/gGa/gTa
Amino acid string: A/G/V

Allele: G (alternative allele 1)
Allele: T (alternative allele 2)

Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen
ENST00000293925 (+) biotype: protein_coding	G (G)	Missense variant	1898	1898	633	A/G	GCA/GGA	0.01	0.117
ENST00000293925 (+) biotype: protein_coding	T (T)	Missense variant	1898	1898	633	A/V	GCA/GTA	0.21	0.006

Overlap consequences

- Contain all the information we have about a specific consequence
- Also contains a ‘predicate’ which tests if this consequence applies to a **TranscriptVariationAllele**

Attribute	Example value(s)	Method(s)
Sequence Ontology term	missense_variant	`$oc->SO_term()`
Label term	Missense variant	`$oc->label()`
Ensembl term	NON_SYNONYMOUS_CODING	`$oc->display_term()`
NCBI term	missense	`$oc->NCBI_term()`
Predicate	True/False	`$oc->predicate($tva)`

Exercise 8a

- Fetch all transcript variations in transcript ENST00000001008 in human and retrieve the following:
  - variation name
  - consequence type (most severe)
  - amino acids *
  - position in cDNA *
  - position in translation *

* if exist

- In a second attempt, filter for the transcript variations with the consequence type “missense_variant”

Hint: `fetch_all_by_Transcripts()` method requires a list reference of objects.
You have only one so use as parameter e.g.:

```
fetch_all_by_Transcripts([$transcript]) instead of
fetch_all_by_Transcripts($transcript)
```

Exercise 8b

- Fetch all the coding transcript variation alleles in the transcript ENST00000001008 in human and retrieve the following:
  - Allele string
  - Codon change (with the allele position displayed, e.g aAa)
  - Amino acid change
  - SIFT and PolyPhen predictions *

* if exist

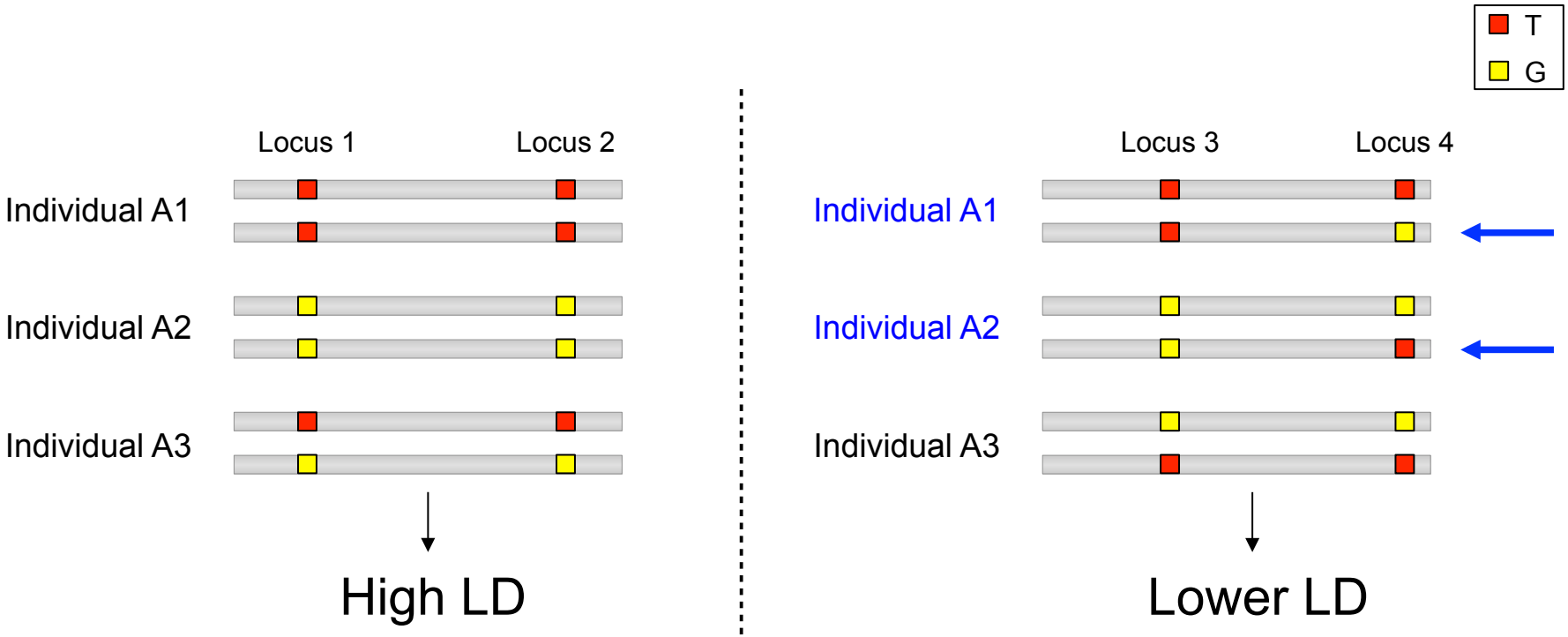
In this exercise, we only want information about the alternate allele.

Hint: You need to fetch the **TranscriptVariation** objects first, as you did in the exercise 8a

Linkage disequilibrium

Linkage disequilibrium (LD) is a measure of how frequently alleles at two separate loci are inherited together on the same haplotype in a specific population

- We provide LD statistics for the 1000 Genome Project phase 3 populations
- Two common measures:
  - r^2, D' ($r^2 = 1$ → perfect LD)



Linkage disequilibrium container object

- Represents an instance of a container of pair-wise LD values in a region (2 by 2)
- Calculated on the fly (using a script written in C)
- Can contain values for multiple populations
- Most methods return hash references
  e.g.: To retrieve the r^2 value you need to use: `$ldc->{r2}`

Method
`$ldc->get_all_ld_values()`
`$ldc->get_all_r_square_values()`
`$ldc->get_all_d_prime_values()`

Key	Value
variation1	*VariationFeature* object of one of the two variations used for the calculation
variation2	*VariationFeature* object of the second variation used for the calculation
r2	0 to 1
d_prime	0 to 1

Exercise 9

Create an LD feature container for region chromosome **9:22124000-22126000** in human and find the names of all pairs of variants in high LD with each other in the 1000 Genomes 'Toscani in Italy' population, which is named "**1000GENOMES:phase_3:TSI**"

Hints:

- Try r^2 >= 0.95
- The `fetch_by_Slice()` method can take an optional argument "Population object"

Note:

Phase 3 Data from the 1000 Genomes Project is accessed directly from VCF files.

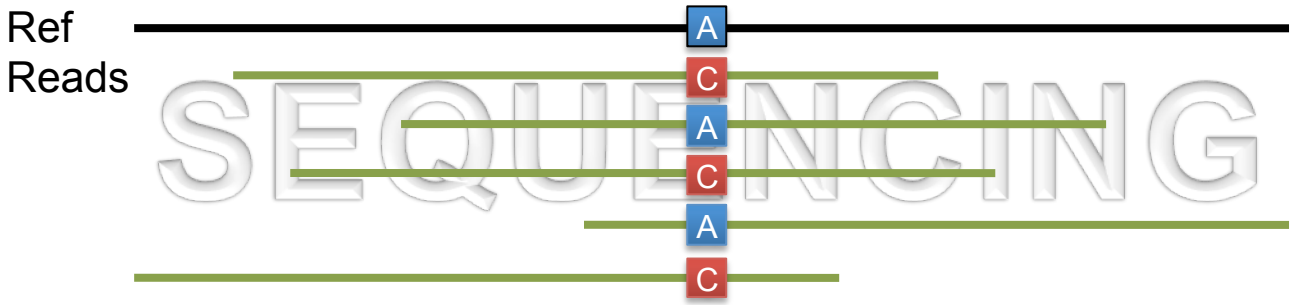
It is not default API behaviour to read VCF's, so to extract LD statistics for this population you will need to set the '**use_vcf**' option on the database adaptor to 1.

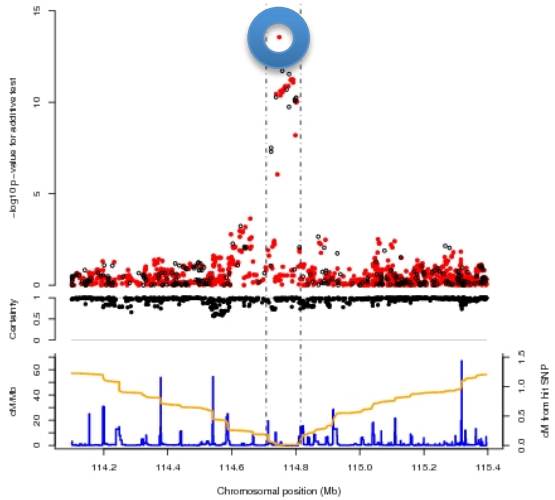
```
$lda->db->use_vcf(1);
```

Variant Effect Predictor

http://www.ensembl.org/info/docs/tools/vep/index.html

λBackground



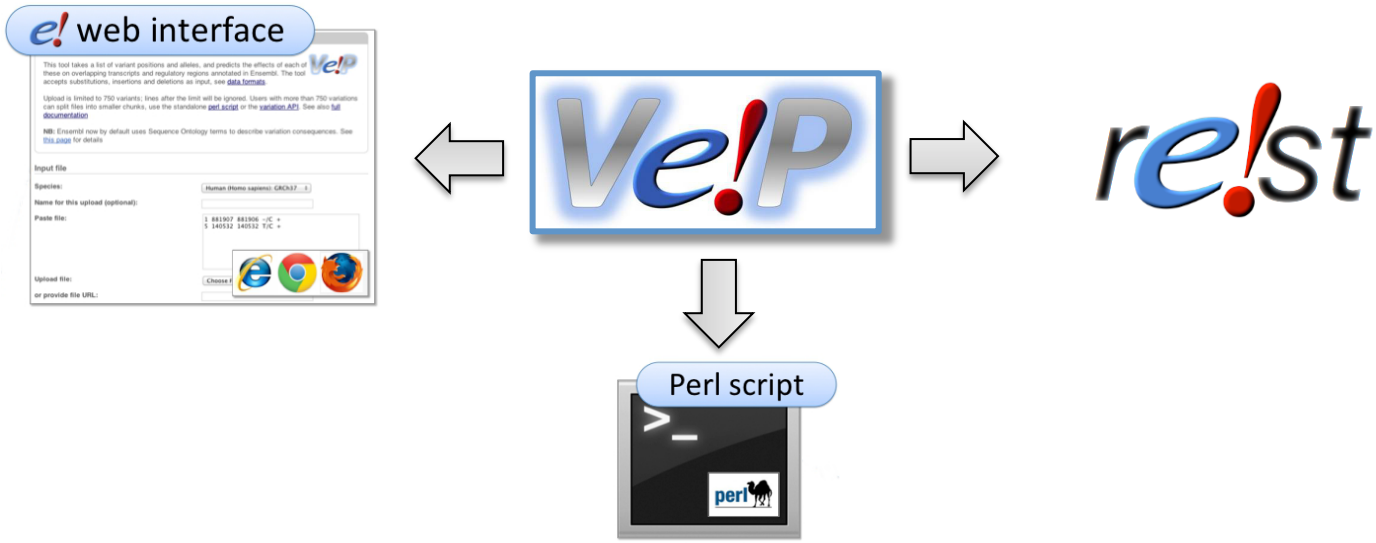


?

SNP

Variant Effect Predictor

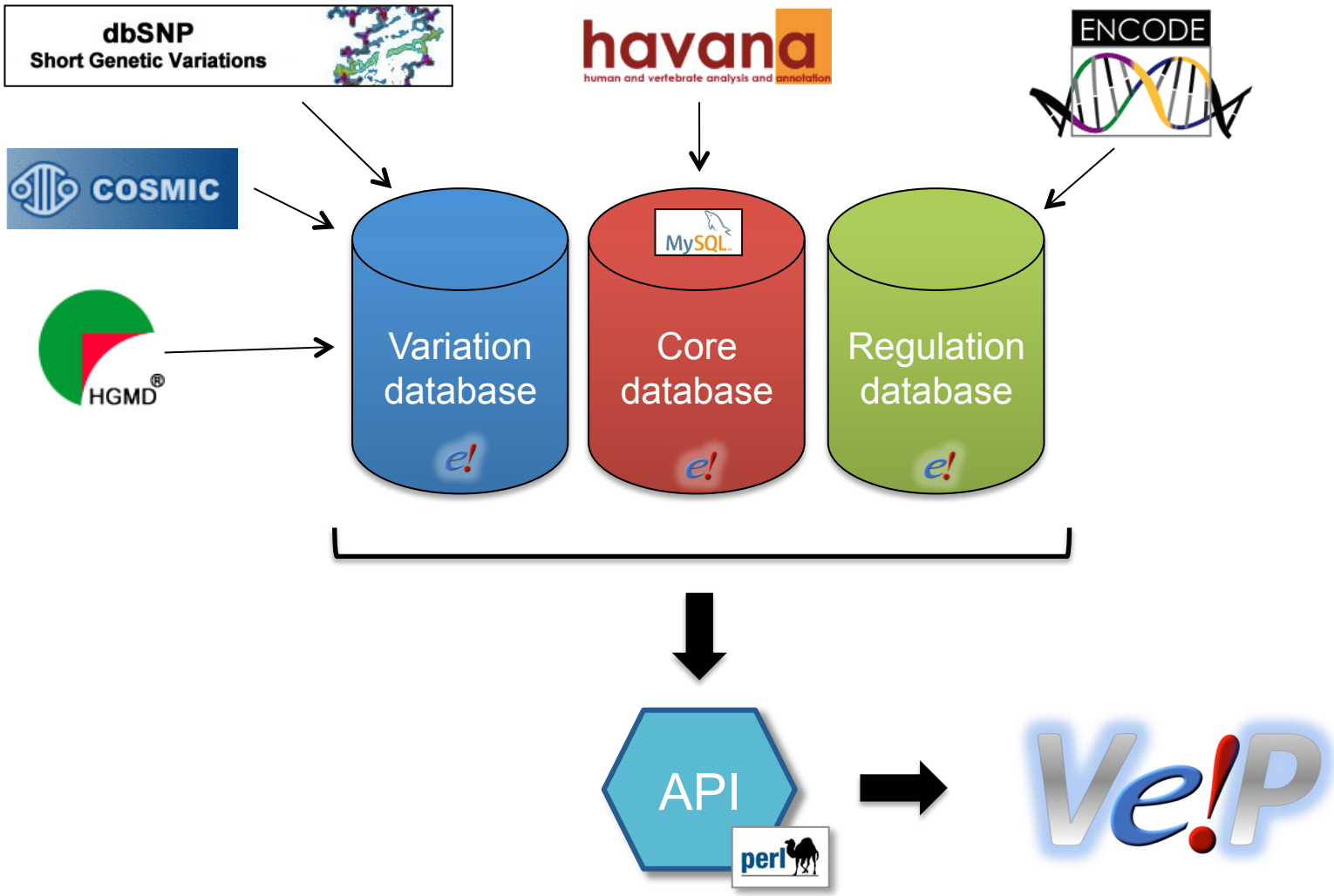
A tool to predict the functional consequences of variants, using the Ensembl API



- Use through simple web interface
- Use locally and completely privately by downloading the script and reference data

- Multiple input formats supported including VCF
- Many different annotations available
- Many filtering options available

Variant Effect Predictor



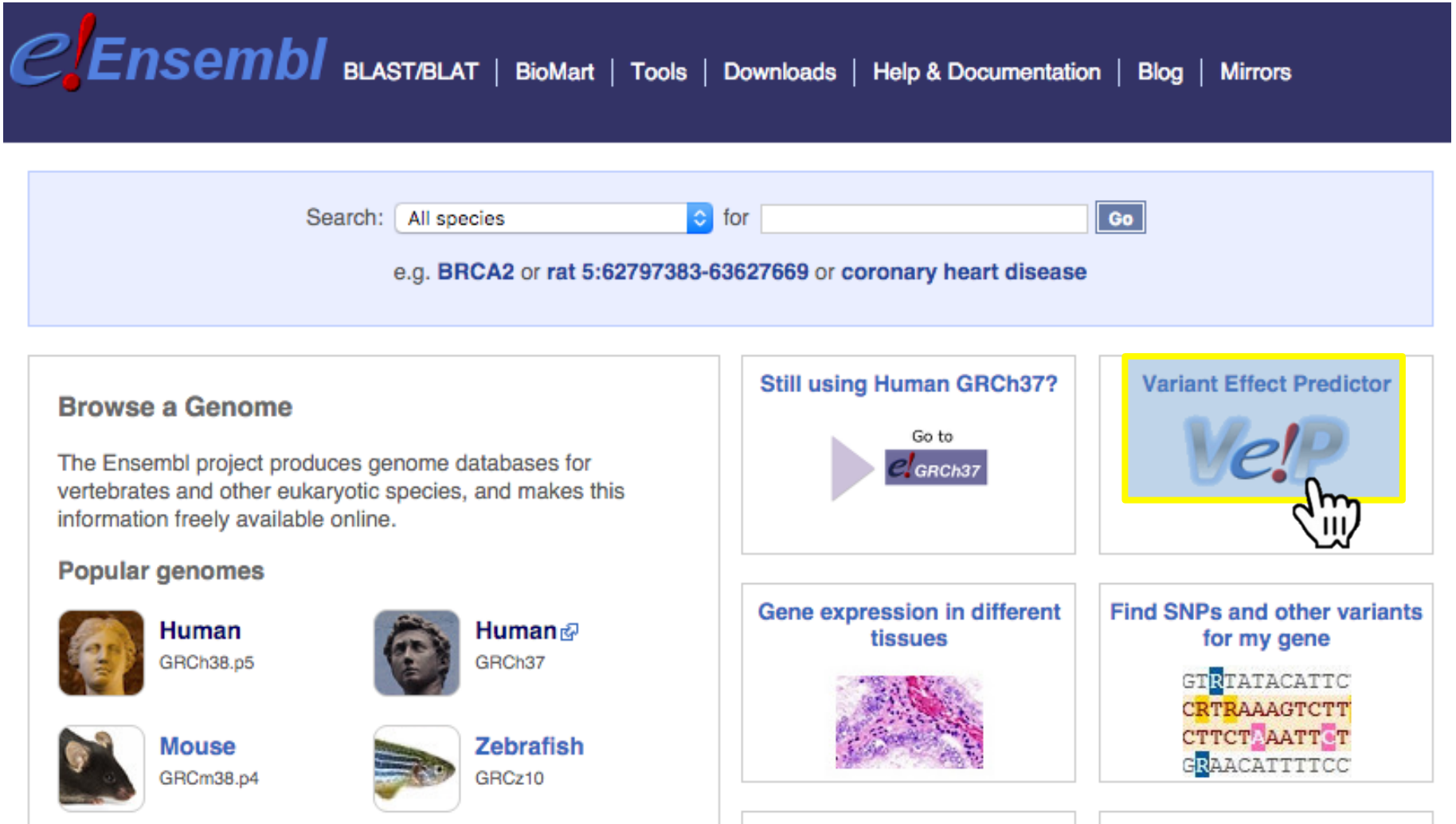
Variant Effect Predictor

- Uses all the consequences types from Ensembl Variation
- Can use aligned RefSeq transcripts (uses the Ensembl gene set by default)
- Provides variant descriptions using HGVS nomenclature
- Can use SIFT and PolyPhen2 predictions for popular species
- Provides regulatory region consequences
- Supports some types of Structural Variants encoded in VCF format
- Provides information on known variants at the same location:
    - Allele frequencies in reference populations
    - Pubmed identifiers if the variant is cited

The web interface

Variant Effect Predictor: Web

ensembl.org



Variant Effect Predictor: Web

ensembl.org/vep

Help & Documentation > API & Software > Ensembl Tools > Variant Effect Predictor

Variant Effect Predictor

The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **genes** and **transcripts** affected by the variants
- **location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen** scores for changes to protein sequence
- ... And more!

Web interface

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

Documentation
Launch the web interface

Standalone perl script

- More options, more flexibility
- For large volumes of data

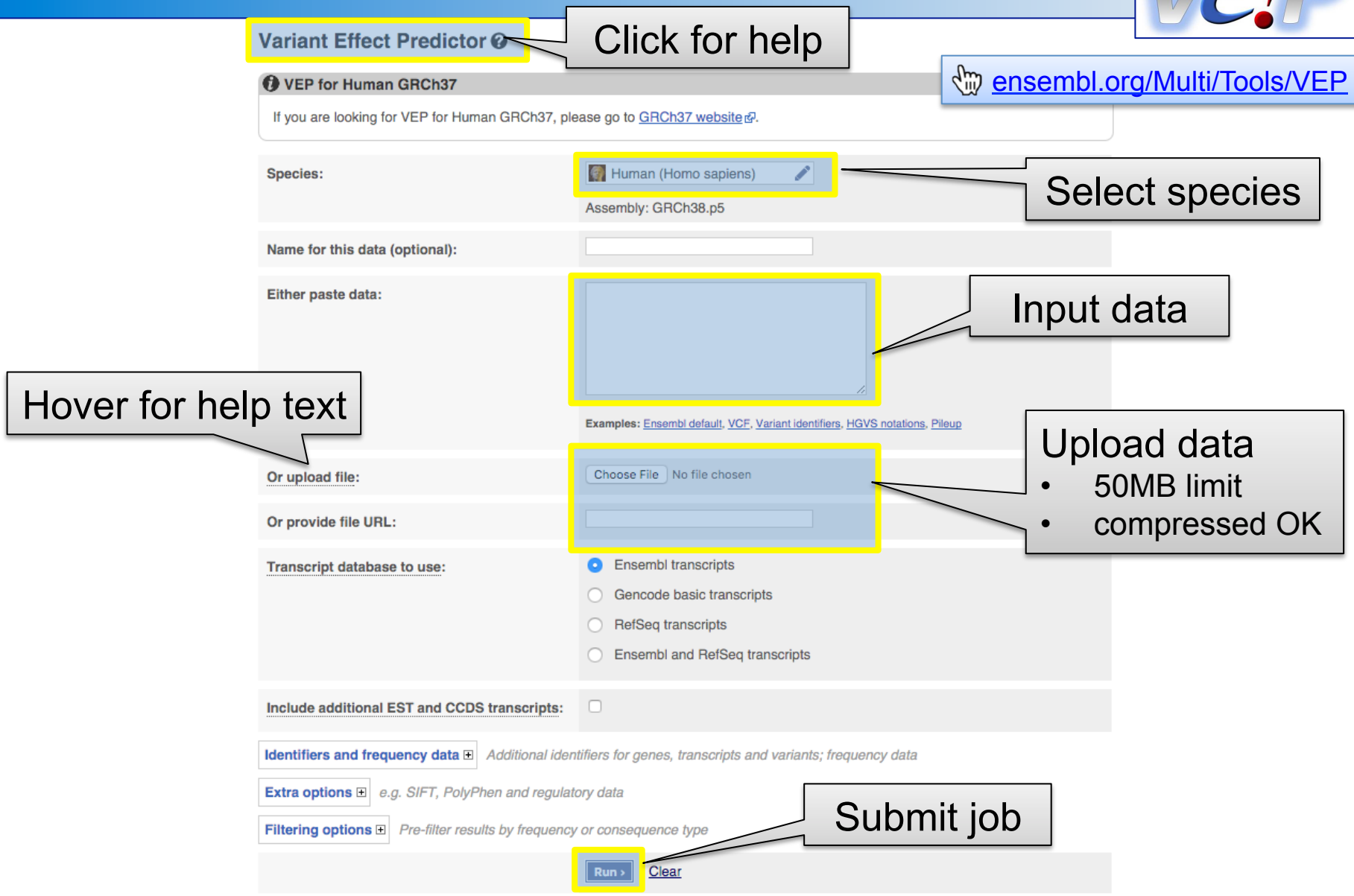
Documentation
Download latest version

REST API

- Language-independent API
- Simple URL-based queries
- GET single variants, POST many

Documentation

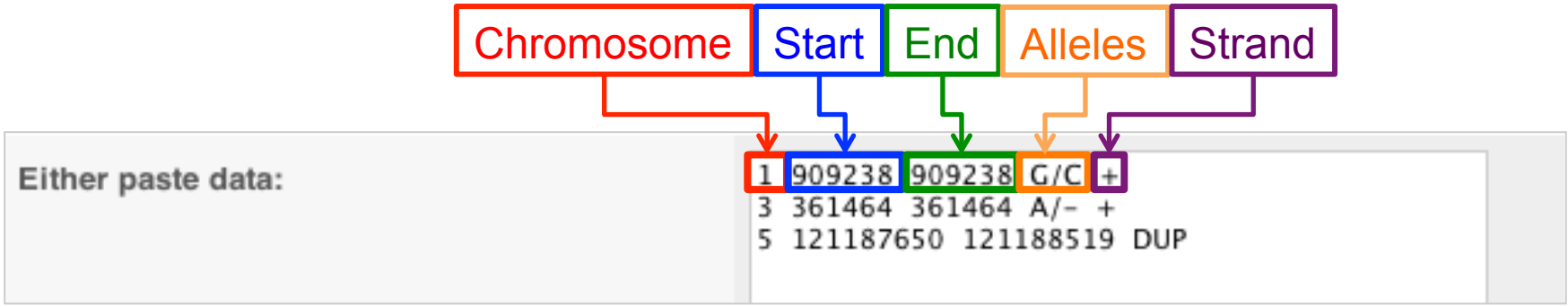
Variant Effect Predictor: Web



Variant Effect Predictor: Web

Input format

ensembl.org/info/docs/tools/vep/vep_formats.html#input

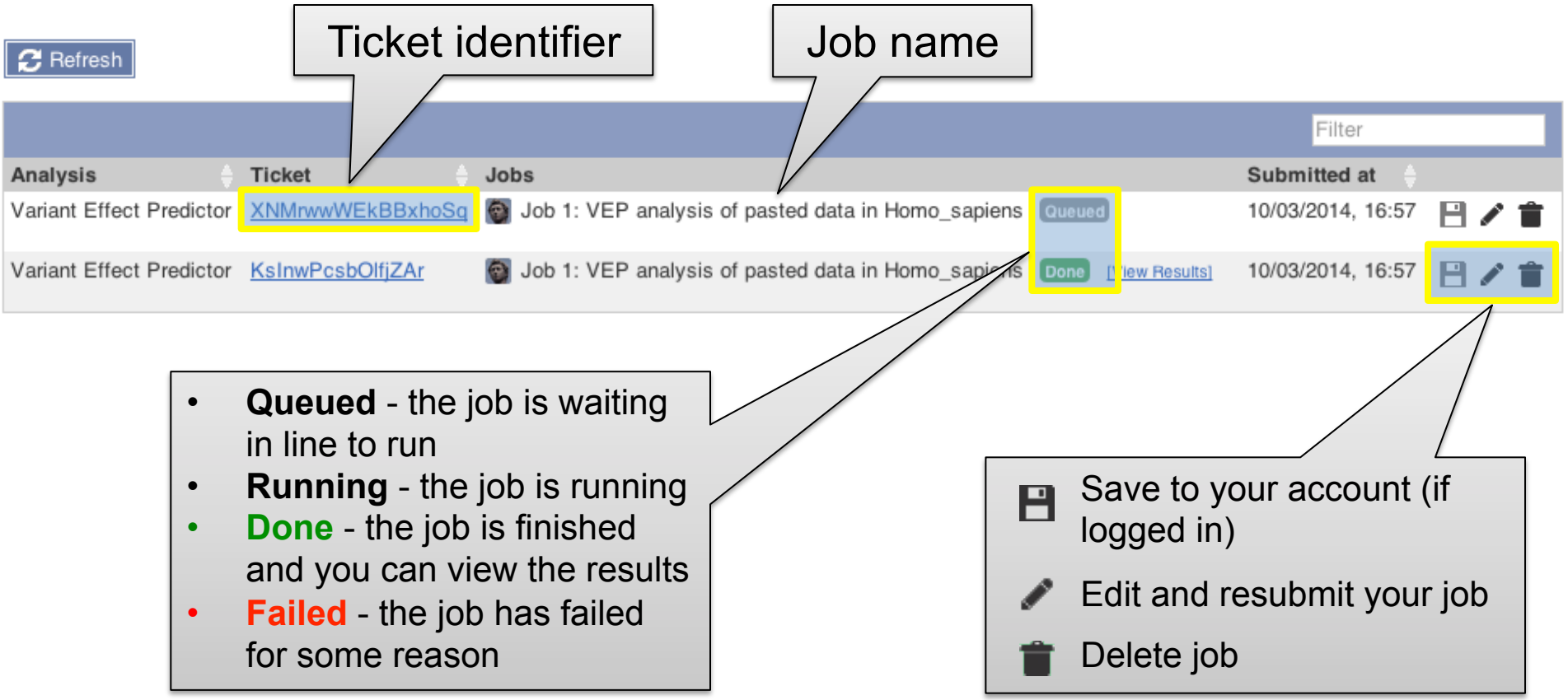


- Coordinates
  - 1-based
  - for a SNP start = end
- Alleles
  - "/" separated list (may be more than two), reference allele first!
  - "-" represents no sequence (insertion, deletion)
- Other formats supported:
  - VCF
  - HGVS nomenclature
  - Variant identifiers (e.g. rsIDs from dbSNP)

Variant Effect Predictor: Web

Jobs

ensembl.org/info/docs/tools/vep/online/input.html#jobs



Variant Effect Predictor: Web

Return to jobs

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors | Search all speci

Human (GRCh37) ▼ | Location: 6:133,002,729-133,035,188 ▼ | Gene: VNN1 ▼ | Transcript: VNN1-001 ▼ | Variation: rs143624951 ▼ | Jobs

Variation displays

- Explore this variation
- Genomic context
  - **Genes and regulation (1)**
  - Flanking sequence
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phenotype Data
- Phylogenetic Context (6)
- Citations
- External Data
  - SNPedia
  - LOVD

Configure this page

Manage your data

Export data

Bookmark this page

Share this page

Download view as CSV

rs143624951 SNP

Original source Variants (including SNPs and indels) imported from dbSNP (release 138) | View in dbSNP

Alleles C/T | Ancestral: **C** | Ambiguity code: **Y**

Location Chromosome **6:133004370** (forward strand) | View in location tab

Most severe consequence Stop gained | See all predicted consequences [Genes and regulation]

Evidence status

HGVS names This variation has **3** HGVS names - click the plus to show

Genes and regulation

Gene and Transcript consequences

Show/hide columns

Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codo
ENSG00000112299 HGNC: VNN1	ENST00000367928 (-) biotype: protein_coding	T (A)	Stop gained	1465	1451	484	W/*	TGG

wellcome trust sanger institute

EMBL-EBI

Variant Effect Predictor: Web

Results summary

ensembl.org/info/docs/tools/vep/online/results.html#summary

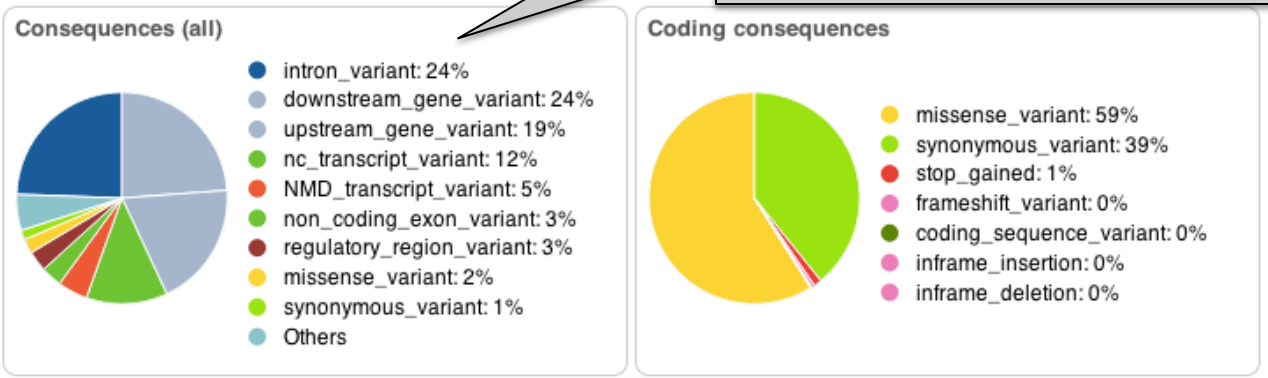
Variant Effect Predictor results

Summary statistics for ticket uTPZdaCscZ7TeOiX:

Category	Count
Variants processed	19976
Variants remaining after filtering	19976
Novel / existing variants	914 (4.6%) / 19062 (95.4%)
Overlapped genes	116
Overlapped transcripts	428
Overlapped regulatory features	373

Variant counts
- before/after filtering
- novel / known variants

Consequence type charts
- total observed - may be more than one per variant
- Separate chart for coding consequences



Variant Effect Predictor: Web

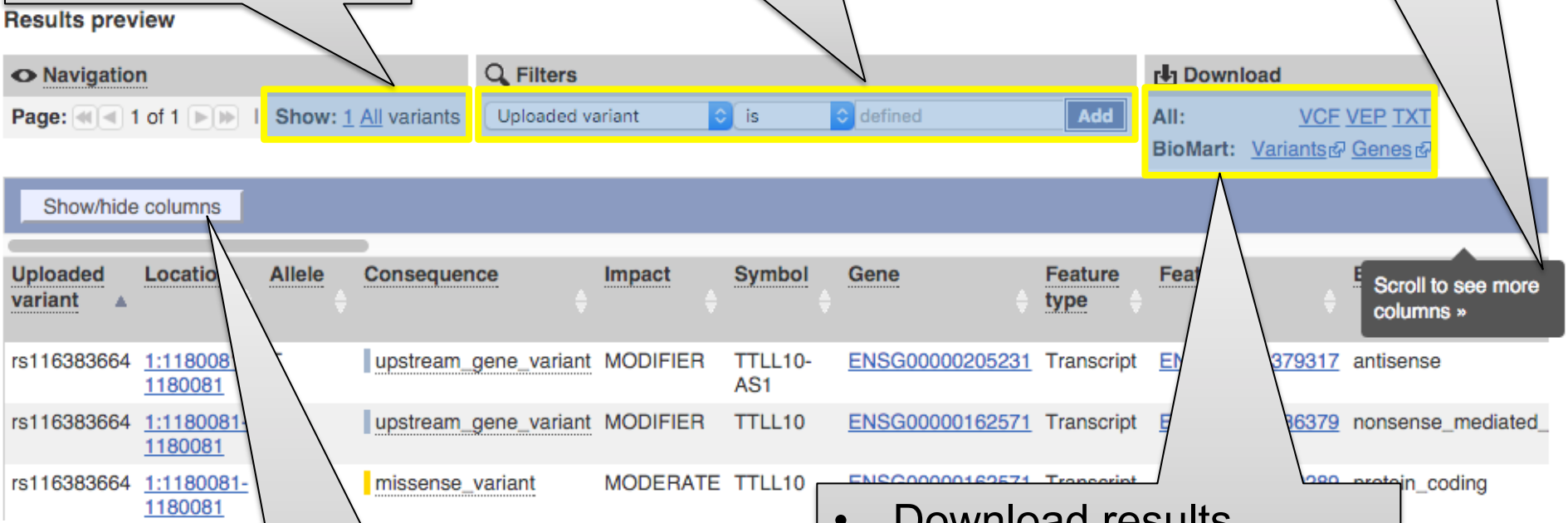
Results preview

ensembl.org/info/docs/tools/vep/online/results.html#table

Navigate results
NB one row per variant/ transcript overlap!

Create and edit filters

NB more columns: scroll right!



Show/hide columns in results table

- Download results
- Send results to BioMart

VEP - Exercise

- Using a nucleotide change of G → A at position 25587758 on chromosome 21:

  a) Does this variant cause an amino acid change?

  b) Is there an existing variant here?

  c) What is its global minor allele frequency (GMAF)?

  d) Which do you think is the “worst” consequence listed?

  e) What is the HGNC symbol of the gene affected?

The script version

Variant Effect Predictor: Script

Ressources

- Documentation
  http://www.ensembl.org/info/docs/tools/vep/script/index.html
- Download
  http://www.ensembl.org/info/docs/tools/vep/script/vep_download.html
- Help
  dev@ensembl.org (public - mailing list)
  helpdesk@ensembl.org (private - ticketed helpdesk)

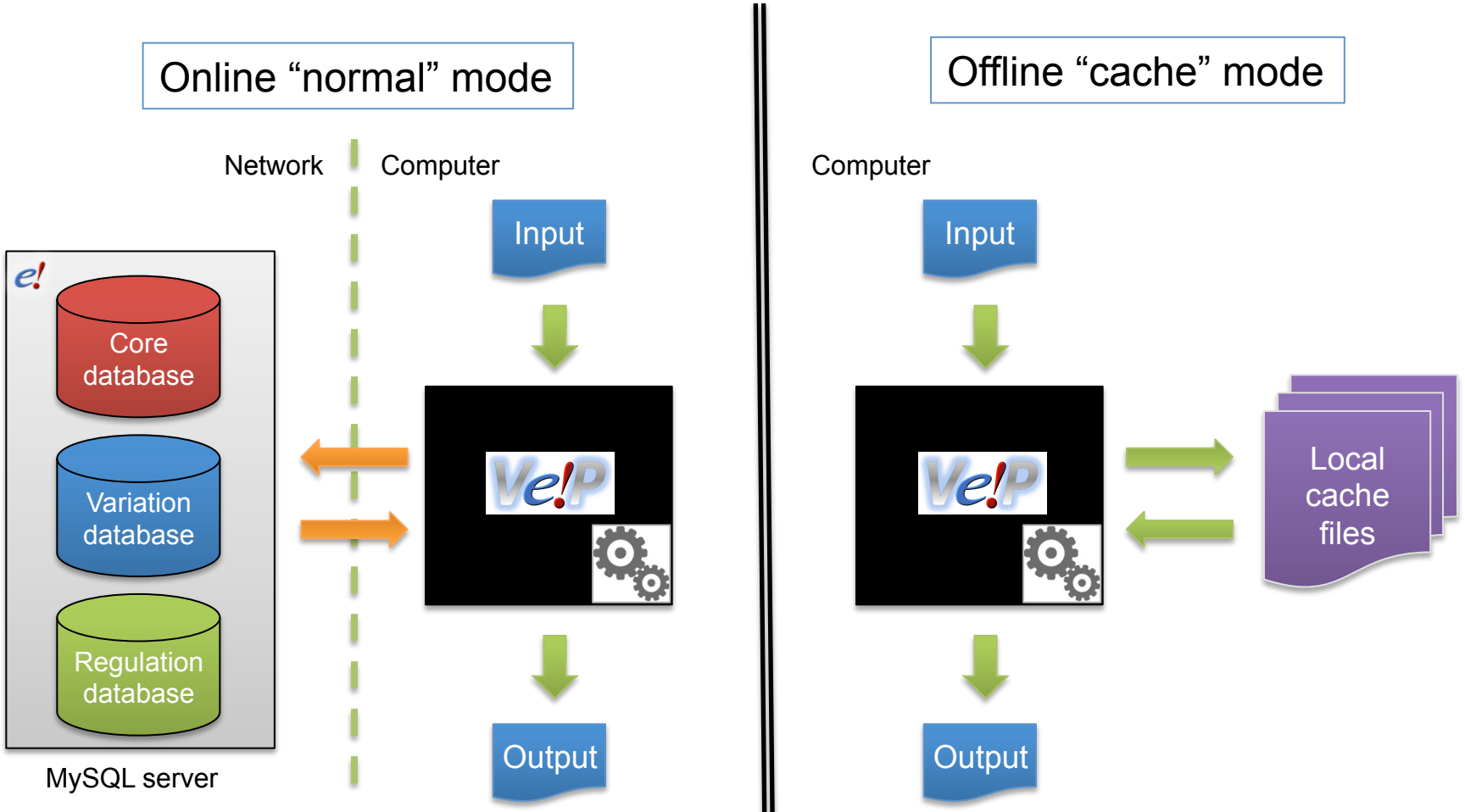
Variant Effect Predictor: Script

Advantages compared to the web version

- Can be included in a pipeline
- No input file limit
- More advanced use (i.e. more filters and options)
- The possibility to use existing plugins or to write your own plugins
  You can see the list of available plugins here:
  https://github.com/Ensembl/VEP_plugins
- The possibility to run the script on stand alone, without internet connection, using the VEP cache files:
  http://www.ensembl.org/info/docs/tools/vep/script/vep_cache.html

Variant Effect Predictor: Script

Online vs offline



Getting more information

- Variation documentation
  http://www.ensembl.org/info/genome/variation/index.html
  http://www.ensembl.org/info/docs/api/variation/index.html
- Online Perl API documentation
  http://www.ensembl.org/info/docs/Doxygen/variation-api/index.html
- Variation API tutorial
  http://www.ensembl.org/info/docs/api/variation/variation_tutorial.html
- Ensembl developers mailing list
  dev@ensembl.org

Acknowledgements

The Entire Ensembl Team

Andrew Yates[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Konstantinos Billis[1], Denise Carvalho-Silva[1], Carla Cummins[1], Peter Clapham[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[1], Leo Gordon[1], Thibaut Hourlier[1], Sarah E. Hunt[1], Sophie H. Janacek[1], Nathan Johnson[1], Thomas Juettemann[1], Stephen Keenan[1], Ilias Lavidas[1], Fergal J. Martin[1], Thomas Maurel[1], William McLaren[1], Daniel N. Murphy[1], Rishi Nag[1], Michael Nuhn[1], Anne Parker[1], Mateus Patricio[1], Miguel Pignatelli[1], Matthew Rahtz[2], Harpreet Singh Riat[1], Daniel Sheppard[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Steven P. Wilder[1], Amonida Zadissa[1], Ewan Birney[1], Jennifer Harrow[2], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Giulietta Spudich[1], Stephen J. Trevanion[1], Fiona Cunningham[1], Bronwen L. Aken[1], Daniel R. Zerbino[1] and Paul Flicek[1,2,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Funding

wellcome trust

National Human Genome Research Institute

CTTV Centre for Therapeutic Target Validation

BLUEPRINT epigenome

EMBL

BBSRC bioscience for the future

BioMedBridges

MedBioinformatics

Co-funded by the European Union

Feedback

https://www.surveymonkey.co.uk/r/API_Jan16

Figures

- Babies: http://www.impawards.com/2010/posters/babies.jpg (slide #2)