

# Ensembl Variation API

Cambridge, December 2013

Sarah Hunt

Ensembl Variation

Materials:

<http://www.ebi.ac.uk/~seh/workshops/>



# Important URLs

- Links for the course materials
  - <http://www.ebi.ac.uk/~seh/workshops/>
- 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

### Key data/objects:

- Variation
- Allele
- Genotypes
- Variation feature (locations)
- Structural variation
- Structural variation feature (locations)

### Additional information

- Variation sets
- Consequences of variations
- Linkage disequilibrium
- Phenotype data
- Citation data

### The Variation Effect Predictor (VEP)

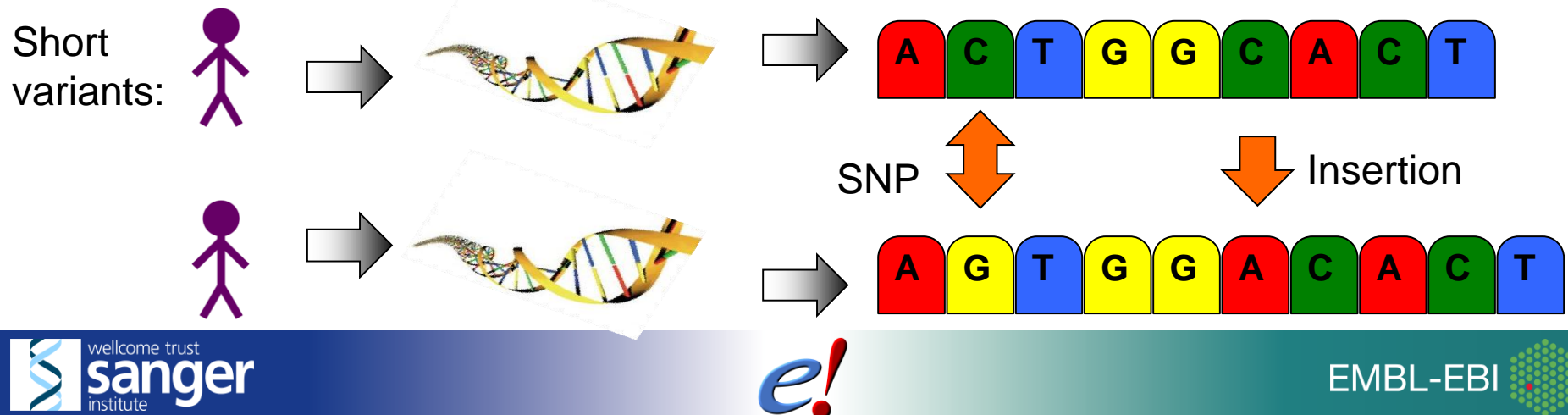
The screenshot displays the Ensembl genome browser interface for the rs1333049 SNP. The top navigation bar includes links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors, along with a search bar and a Login/Register link. The main content area is divided into several sections: 'Variation displays' on the left with a tree view for exploring the variation (Genomic context, Genes and regulation, Flanking sequence, Population genetics, Individual genotypes, Linkage disequilibrium, Phenotype Data, Phylogenetic Context, Citations, External Data, SNPedia, LOVD), 'Configure this page', 'Manage your data', 'Export data', 'Bookmark this page', and 'Share this page'. The central panel shows the 'rs1333049 SNP' details, including the original source, alleles, location (Chromosome 9:22125503), evidence status, synonyms, HGVS name, and genotyping chips. It also lists variants imported from dbSNP and reference/alternative alleles. The 'Explore this variation' section at the bottom features icons for Genomic context, Genes and regulation, Population genetics, Individual genotypes, Linkage disequilibrium, Phenotype data, Citations, Phylogenetic context, and Flanking sequence.

The screenshot displays the Ensembl genome browser interface for the ns916030 structural variation. The top navigation bar is similar to the previous screenshot, with a search bar and a Login/Register link. The main content area shows 'Structural variation displays' on the left, including a tree view for exploring the SV (Genomic context, Genes and regulation, Supporting evidence, Phenotype Data), 'Configure this page', 'Manage your data', 'Export data', 'Bookmark this page', and 'Share this page'. The central panel shows the 'Structural variation: ns916030' details, including the variation class (CNV), allele type(s) (Gain), source (DGVa), study (nstd37), alias (ISCA\_VAR\_v5\_860), location (Chromosome 13:100636950-100638575), and genomic size (1,626 bp). The 'Explore this SV' section at the bottom features icons for Genomic context, Genes and regulation, Supporting evidence, and Phenotype data.

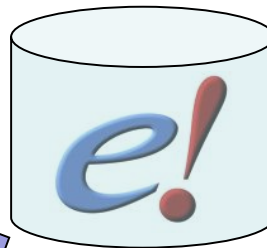
# An introduction to variation

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

- We study 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 62.7 million short human variants in dbSNP (version 138)
- There are over 9 million longer (structural) variations in the Database of Genomic Variants



# Data flow



- HGMD-Public
- COSMIC
- OMIM
- OMIA
- UniProt
- NHGRI
- 1000 Genomes

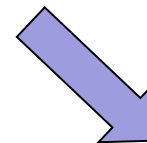
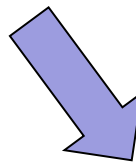
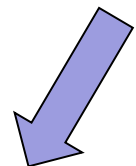
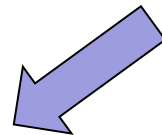
## QC

- Quality filtering
- Evidence summary

## Analysis

- Ancestral allele
- Gene impact

API



## API scripts

```
my $registry_file = "/usr/local/ensembl/registry";
my $reg_load_all($registry_file);
my $reg_load_registry_from_db($host => "ensembl.org", $user => "anonymous");
my $sdb = $reg->get_adapter("human", "core", "slice");
my $slice;
if (scalar @ARGV) {
    $slice = $sdb->fetch_by_region("chromosome", $ARGV[0], $ARGV[1], $ARGV[2]); # strain's long
} else {
    $slice = $sdb->fetch_by_region("chromosome", 21, 3458530, 3458570);
    $slice = $sdb->fetch_by_region("chromosome", 11, 3566110, 3566140);
}
# create a new mapped slice container
my $msc = Bio::EnsEMBL::MappedSliceContainer->new($slice, -EXPANDED => 1);
# create a new strain slice adaptor and attach it to the mapped slice container
my $ssa = Bio::EnsEMBL::DBSQL::StrainSliceAdaptor->new($ssa->db);
$msc->set_strain_slice_adaptor($ssa);
# now attach strains
$msc->attach_strain("Watson");
$msc->attach_strain("Yamanaka");
```









REST  
API

## Website



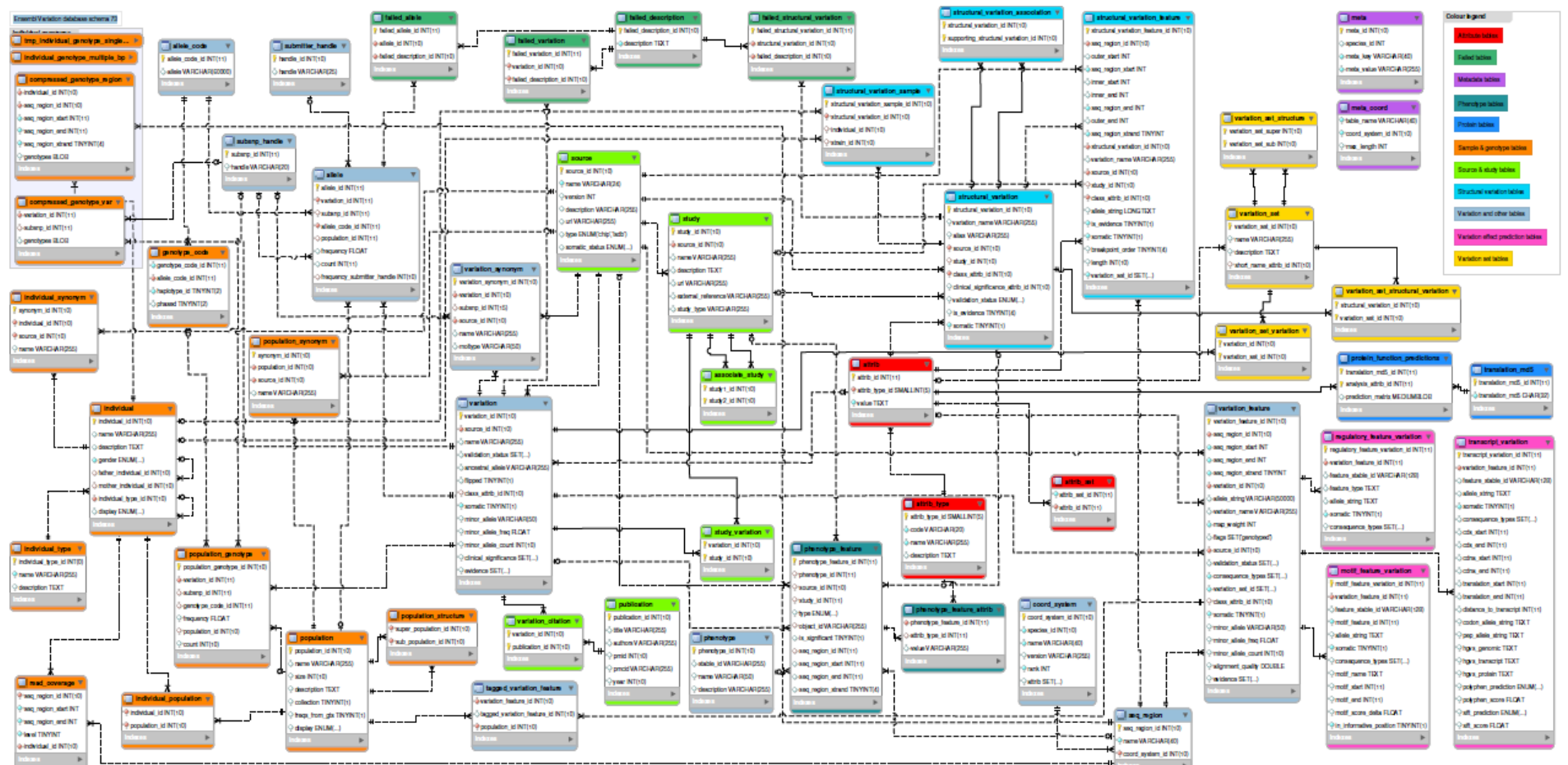
# Quality control

Icon	Name	Description	QC Type	Reported failure reason
	Multiple_observations	The variant has multiple independent dbSNP submissions, i.e. submissions with a different submitter handles or different discovery samples	Mapping checks	Variation does not map to the genome
	Frequency	The variant is reported to be polymorphic in at least one sample		Variation maps to more than 1 location
	HapMap	The variant is polymorphic in at least one HapMap panel (human only)		Mapped position is not compatible with reported alleles
	1000 Genomes	The variant was discovered in the 1000 genomes project (human only)		None of the variant alleles match the reference allele
	Cited	The variant is cited in a PubMed article.	Checks on the alleles of refSNPs	Loci with no observed variant alleles in dbSNP
	ESP	The variant was discovered in the Exome Sequencing Project (human only).		Variation has more than 3 different alleles
				Alleles contain ambiguity codes
			Checks on the alleles in dbSNP submissions	Alleles contain non-nucleotide characters
				Additional submitted allele data from dbSNP does not agree with the dbSNP refSNP alleles
			External failure classification	Flagged as suspect by dbSNP

We run basic checks and flag suspicious variants. By default such variants are not returned by the API; to retrieve them set the variable `DBAdaptor::include_failed_variations()`

We summarise the information supporting each dbSNP variant to give a guide to its reliability  
Variants which are cited are returned by the API even if they fail one of the QC filters

# Variation database schema



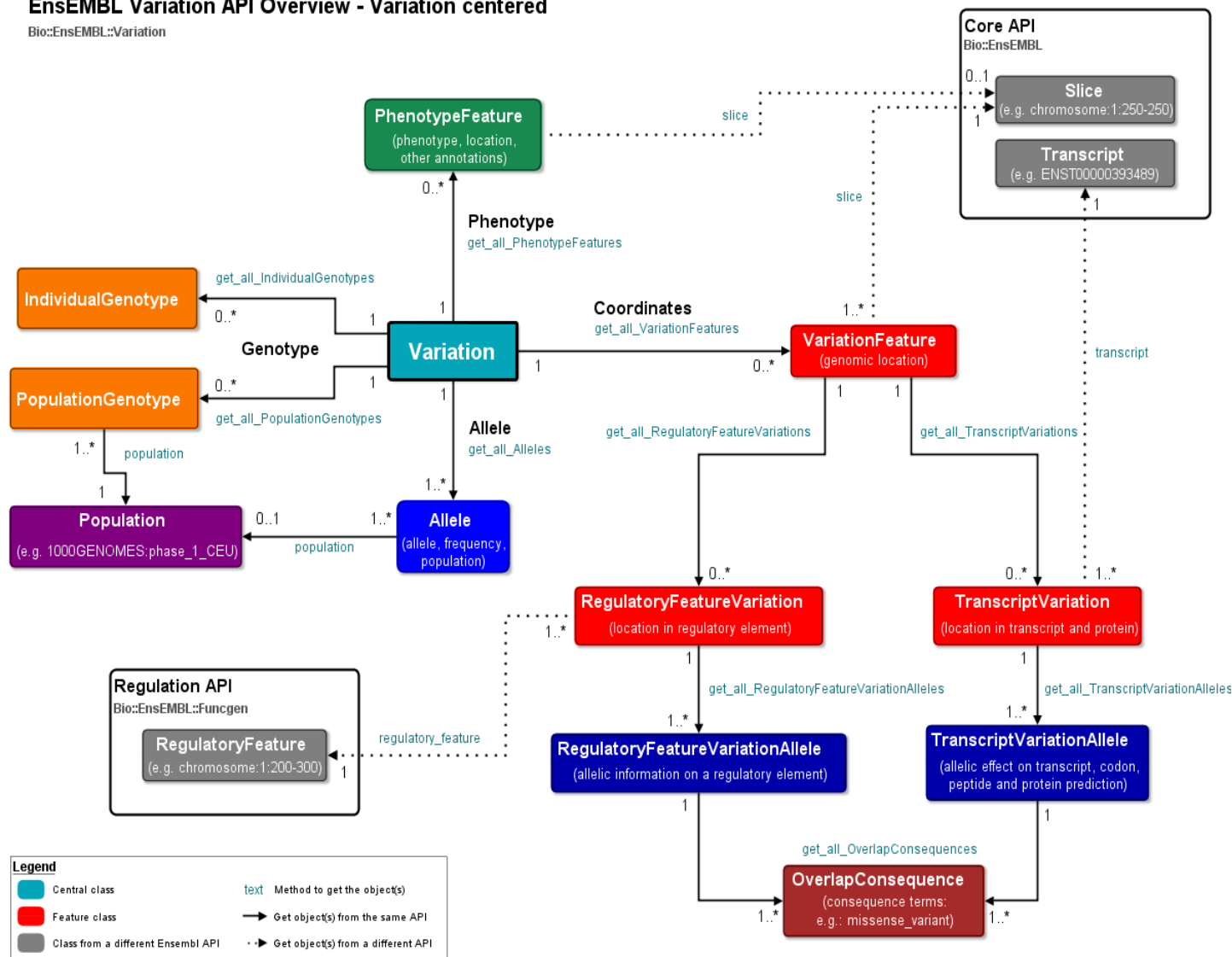
[http://www.ensembl.org/info/docs/variation/variation\\_schema.html](http://www.ensembl.org/info/docs/variation/variation_schema.html)



# Simplified object diagram

## Ensembl Variation API Overview - Variation centered

Bio::Ensembl::Variation





# 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 => 'ensembl.ensembl.org',
  -user => 'anonymous'
);

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



species



group



object name

# Object creation

Using adaptors:

- Example: VariationAdaptor
- **Fetch** object(s) according to some property e.g. name, location
  - “fetch\_all\_...” => returns a listref of items
  - “fetch\_by\_...” => usually returns only 1 item
- Check documentation which methods the adaptor provides

Using API objects:

- Example: Variation
- **Get** other object(s) from an API object
  - Eg: \$variation->**get\_all\_Alleles()** => returns a listref of Allele objects
- Usually the object is written with a upper case in the method (e.g. get\_all\_**Alleles()**)

# Variation object

Represents a short sequence variation

- Retrieved using the variation adaptor
- Has a name, variation class, data source, flanking sequence
- May have an evidence class, ancestral allele, clinical significance

Attribute	Example value(s)	Method(s)
Variation name	rs1333049, COSM29450	<code>\$v-&gt;name()</code>
Source	dbSNP, COSMIC	<code>\$v-&gt;source()</code>
Class	SNP, insertion, deletion	<code>\$v-&gt;var_class()</code>
Flanking sequence	TAGCTAGCTATTAC....	<code>\$v-&gt;five_prime_flanking_seq()</code> <code>\$v-&gt;three_prime_flanking_seq()</code>
Supporting evidence	Array: [Frequency, Multiple_observations,etc]	<code>\$v-&gt;get_all_evidence_values()</code>

# Exercise 1

- Retrieve the following information:
  - Variation name
  - Variation class
  - Data source
- for the following human variations:
  - rs55710239
  - rs56385407
  - COSM998
  - CI003207

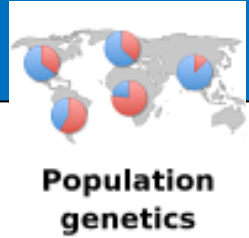
Variation API documentation:

<http://www.ensembl.org/info/docs/Doxygen/variation-api/index.html>

Variation tutorial:

[http://www.ensembl.org/info/docs/api/variation/variation\\_tutorial.html](http://www.ensembl.org/info/docs/api/variation/variation_tutorial.html)

# Allele object



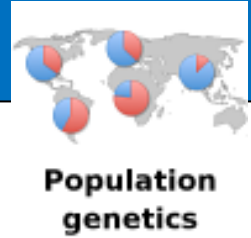
Represents an allele observed for a variation .

- Can be retrieved from variation objects, as well as the allele adaptor
- May have assayed population, frequency and submitter information

Attribute	Example value(s)	Method(s)	Comment
Allele	G, AC	<code>\$a-&gt;allele()</code>	
Frequency	0.15, 1	<code>\$a-&gt;frequency()</code>	not always defined
Population	Population object	<code>\$a-&gt;population()</code>	returns object
Variation	Variation object	<code>\$a-&gt;variation()</code>	returns 'parent' object *

\* NB most API objects have a method to retrieve reference to “parent” object

# Exercise 2



- For human variation rs1333049 retrieve the following for each of its alleles:
  - Allele
  - Frequency\*
  - Population name\*
  - Submitter name in dbSNP (“handle”)\*

**HINT:** You may need to test whether the population object returned is empty to avoid a script error.

\* if exists

# Variant Genotypes

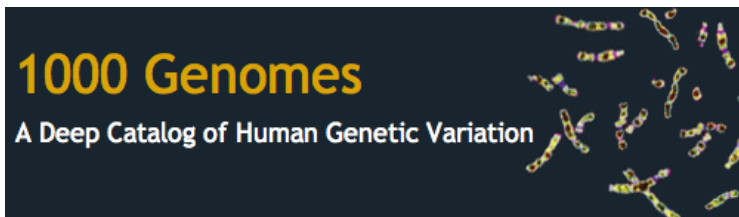
	T	C
T	T	T
C	T	C

Individual  
genotypes

Variant genotypes show which combination of the possible alleles a specific individual has at the site of a specific variant

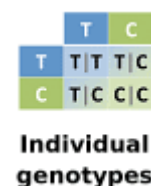
We import genotype data from dbSNP and large scale sequencing project such as the 1000 Genomes project and the WTSI Mouse Genomes Project.

Genotype data is large, so it is stored in a compressed format in the Ensembl databases and cannot be accessed using direct SQL





# Individual Genotype object



Represents an instance of the genotype of a variation in a specific individual

- Can be retrieved from variation objects, as well as the IndividualGenotype adaptor

Attribute	Example value(s)	Method(s)
Genotype	Genotype string (format A T)	<code>\$ig-&gt;genotype_string()</code>
Variation	Variation object	<code>\$ig-&gt;variation()</code>
Individual	Individual object	<code>\$ig-&gt;individual()</code>

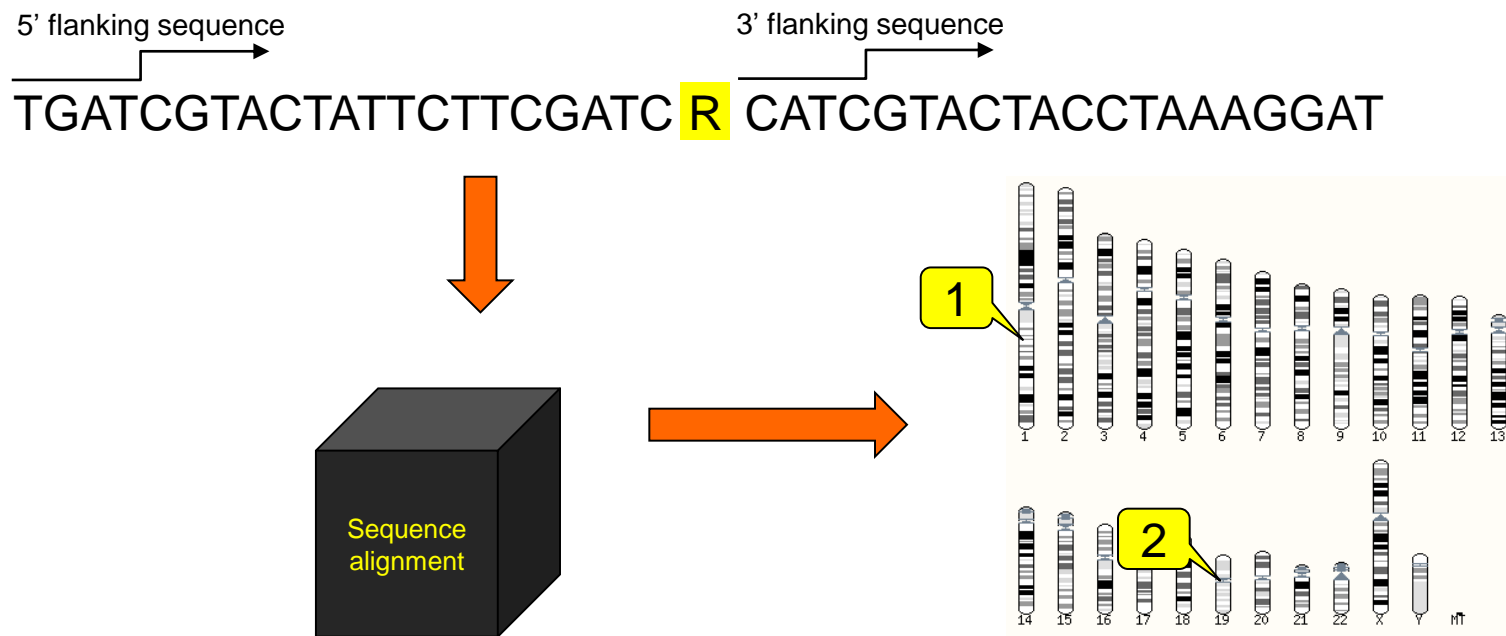
# Exercise 3



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

# 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

Represents an instance of a variation mapping to the genome

- Can be retrieved from slice and variation objects, as well as variation feature adaptor

Attribute	Example value(s)	Method(s)
Allele string	A/G, -/C	<code>\$vf-&gt;allele_string()</code>
Chromosome	15, X	<code>\$vf-&gt;seq_region_name()</code>
Coordinates	103019234	<code>\$vf-&gt;start()</code> } slice relative
		<code>\$vf-&gt;end()</code>
		<code>\$vf-&gt;seq_region_start()</code> } chromosome relative
		<code>\$vf-&gt;seq_region_end()</code>
Slice object	Slice object	<code>\$vf-&gt;slice()</code> [returns core object]

# Exercise 4

- Retrieve all the variation features in the region of **rat** chromosome 20 from 23000000 to 23150000 and extract following information:
  - Variation name
  - Allele string
  - Location

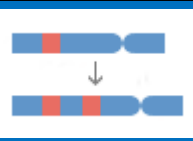
Extra:

- Find the genomic locations of the human variants: rs7107418, rs671, rs17646946 and rs4988235

**HINT:** usually when you have to search something by location or in need to use the Slice object from the Core API

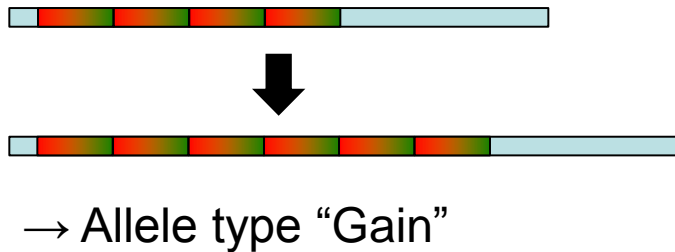
Core API: <http://www.ensembl.org/info/docs/Doxygen/core-api/index.html>

# Structural variation

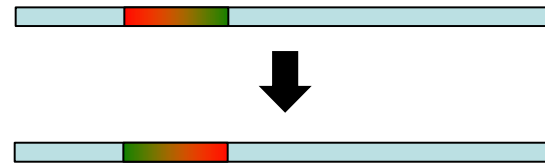


- A structural variation is a long (> 1kb) section of variable DNA sequence
- Different processes give rise to different types of structural variations:

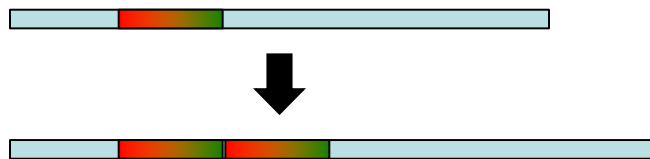
## Copy number variation (CNV)



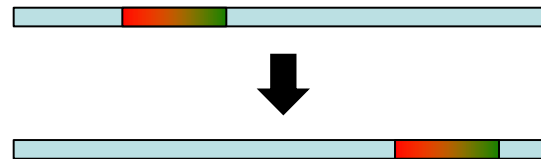
## Inversion



## Duplication



## Translocation



# Structural variation object



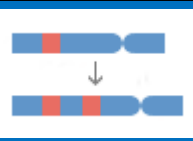
Represent a structural (long) sequence variation

- Data is imported from:
  - DGVa (with study information and supporting evidence)
  - CNV probes from Illumina and Affymetrix genotyping chips

Attribute	Example value(s)	Method(s)
Structural variation name	esv214236	<code>\$sv-&gt;variation_name()</code>
Class	CNV	<code>\$sv-&gt;var_class()</code>
Study	estd59	<code>\$sv-&gt;study-&gt;name()</code>
Supporting evidence	Reference list of Supporting Structural variation objects	<code>\$sv-&gt;get_all_SupportingStructuralVariants()</code>



# Structural variation coordinates



- Usually a structural variation has only a start and stop
- If the breakpoint locations cannot be determined precisely they are defined by a region



# Structural variation feature object



Represents an instance of a structural variation mapping to the genome

- Can be retrieved from a slice as well as its adaptor and SV object
- Only coordinates, class and the list of the supporting evidences are stored (there are no allele strings, alleles, etc)

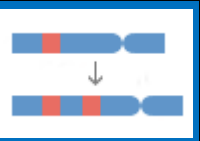
Attribute	Example value(s)	Method(s)
Variation name	esv214236	<code>\$svf-&gt;variation_name()</code>
Coordinates		<code>\$svf-&gt;seq_region_name()</code>
		<code>\$svf-&gt;start()</code>
		<code>\$svf-&gt;end()</code>
		<code>\$svf-&gt;seq_region_start()</code>
		<code>\$svf-&gt;seq_region_end()</code>
		<code>\$svf-&gt;outer_start()</code>
		<code>\$svf-&gt;inner_start()</code>
		<code>\$svf-&gt;inner_end()</code>
		<code>\$svf-&gt;outer_end()</code>
Class	CNV	<code>\$svf-&gt;var_class()</code>

slice relative

chromosome relative

SV specific chromosome relative coordinates

# Exercise 5



- Fetch the following information for the structural variation esv234231 in Human :
  - Structural variation class
  - Study name and description
  - Coordinates
- Fetch the names and the classes (sequence ontology term) of its supporting structural variations

# Variation sets

We create variation sets which are arbitrary grouping of variations

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

Parent set	Set
All HapMap	HapMap – CEU, HapMap – HCB, HapMap – JPT, HapMap - YRI
All phenotype-associated variants	HGMD-PUBLIC variants, ClinVar , OMIM, NHGRI GWAS catalog...
1000 Genomes - All	1000 Genomes - ASN – common, 1000 Genomes – EUR...
	Affy GeneChip 500K, Illumina_HumanOmni2.5 Illumina_Cardio-Metabo_Chip
	ESP_6500

See: [http://www.ensembl.org/info/docs/variation/data\\_description.html#variation\\_sets](http://www.ensembl.org/info/docs/variation/data_description.html#variation_sets)

# Variation set object

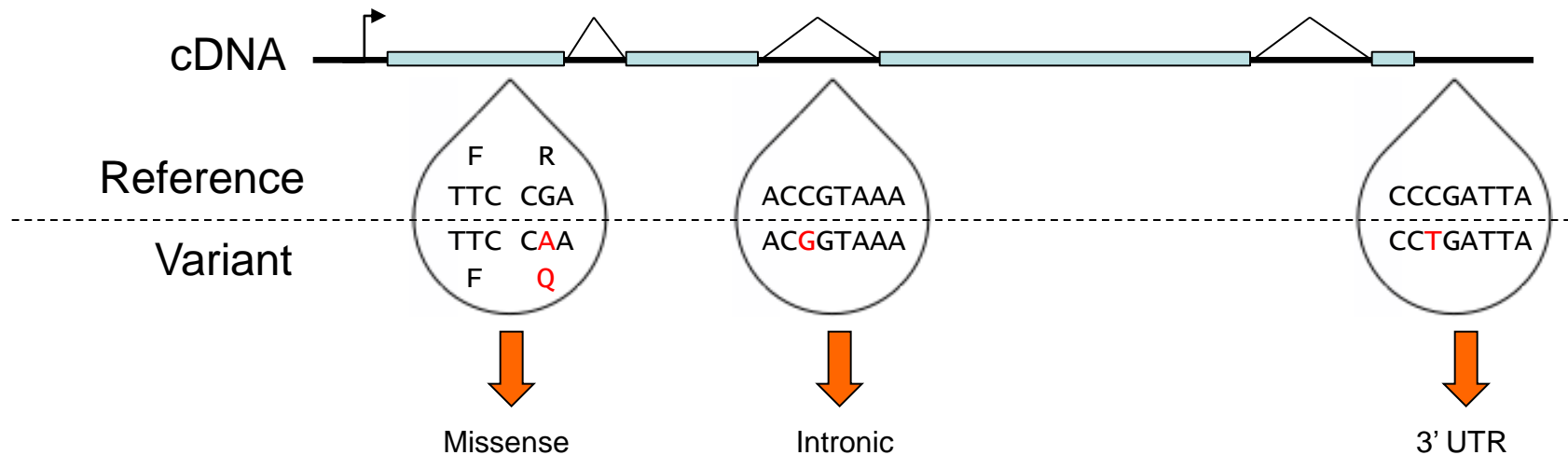
- Represents an arbitrary collections of variations
- Some sets contain millions of variations, in which case you can fetch an Iterator instead of a list

Attribute	Example value(s)	Method(s)
Description	'1000 genomes – All – Common' , 'Affy GenomeWideSNP_6.0 '	<code>\$vs-&gt;description()</code>
Members	List of Variation objects Iterator over Variations	<code>\$vs-&gt;get_all_Variations()</code> <code>\$vs-&gt;get_Variation_Iterator()</code>
Sub/Super sets	VariationSet object	<code>\$vs-&gt;get_all_sub_VariationSets()</code> <code>\$vs-&gt;get_all_super_VariationSets()</code>

# Consequences of variations



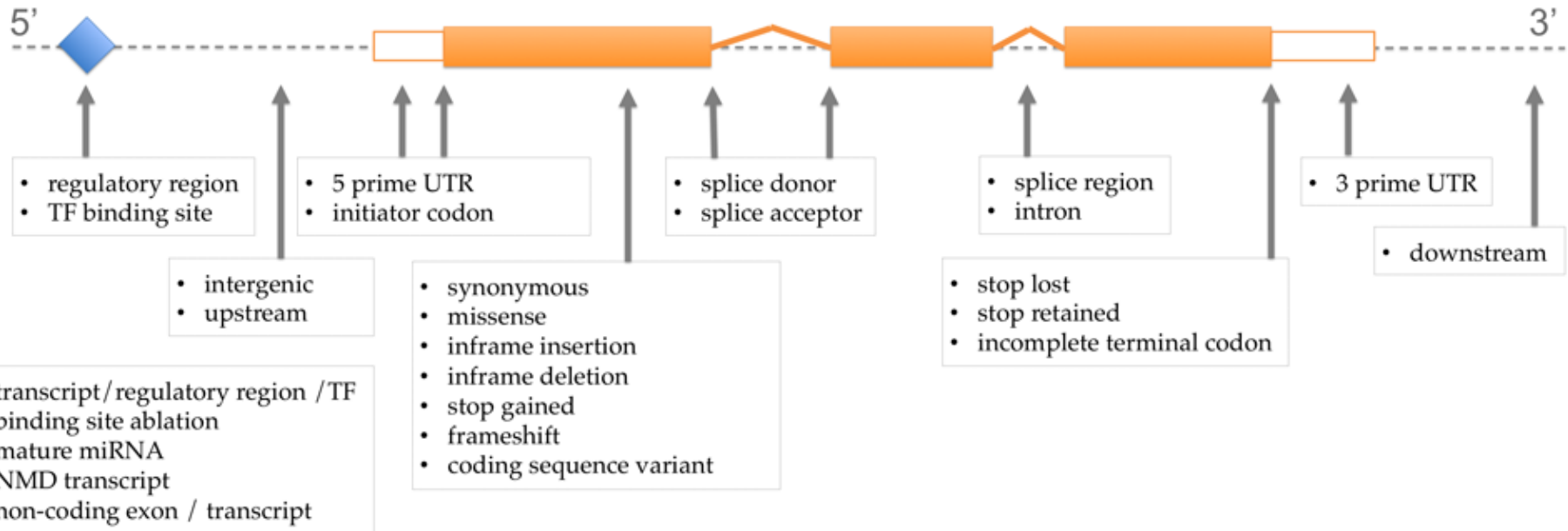
Genes and  
regulation



# Consequence types



Genes and  
regulation



Ensembl uses consequence terms defined by the Sequence Ontology project

See:

[http://www.ensembl.org/info/genome/variation/predicted\\_data.html#consequences](http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences)



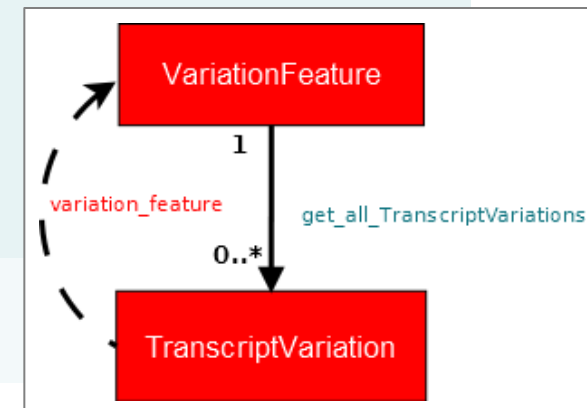
# Transcript variation object



Genes and  
regulation

- Represents an instance of a variation feature in or near a transcript
  - There can be multiple transcriptvariation objects per variation feature object
  - The most “severe” consequence is stored in the variation feature object
  - Can be retrieved from variation feature objects as well as object adaptor

Attribute	Example value(s)	Method(s)
Transcript	Transcript object	<code>\$tv-&gt;transcript()</code> <b>*NB returns core object</b>
Consequence type	intron_variant, stop_lost	<code>\$tv-&gt;display_consequence()</code> <b>*single, most severe</b> <code>\$tv-&gt;consequence_type()</code> <b>*all in an array</b>
Coordinates		<code>\$tv-&gt;cdna_start()</code> <code>\$tv-&gt;cdna_end()</code> <code>\$tv-&gt;cds_start()</code> <code>\$tv-&gt;translation_start()</code>
Amino acids	F/L, */W	<code>\$tv-&gt;pep_allele_string()</code>



# Transcript variation allele object



Genes and  
regulation

- Represents an instance of an allele of a variation feature in or near a transcript
  - Consequences are evaluated at the allele level
  - All the consequences of a *TranscriptVariationAllele* are represented as a list of *OverlapConsequence* objects

Attribute	Example value(s)	Method(s)
TranscriptVariation	TV object	<code>\$tva-&gt;transcript_variation()</code>
OverlapConsequences	Reference list of objects	<code>\$tva-&gt;get_all_OverlapConsequences()</code>
HGVS notation at various levels	string	<code>\$tva-&gt;hgvs_genomic()</code> <code>\$tva-&gt;hgvs_transcript()</code> <code>\$tva-&gt;hgvs_protein()</code>
Affected codons	aAt/aCt	<code>\$tva-&gt;display_codon_allele_string()</code>

# Overlap consequences

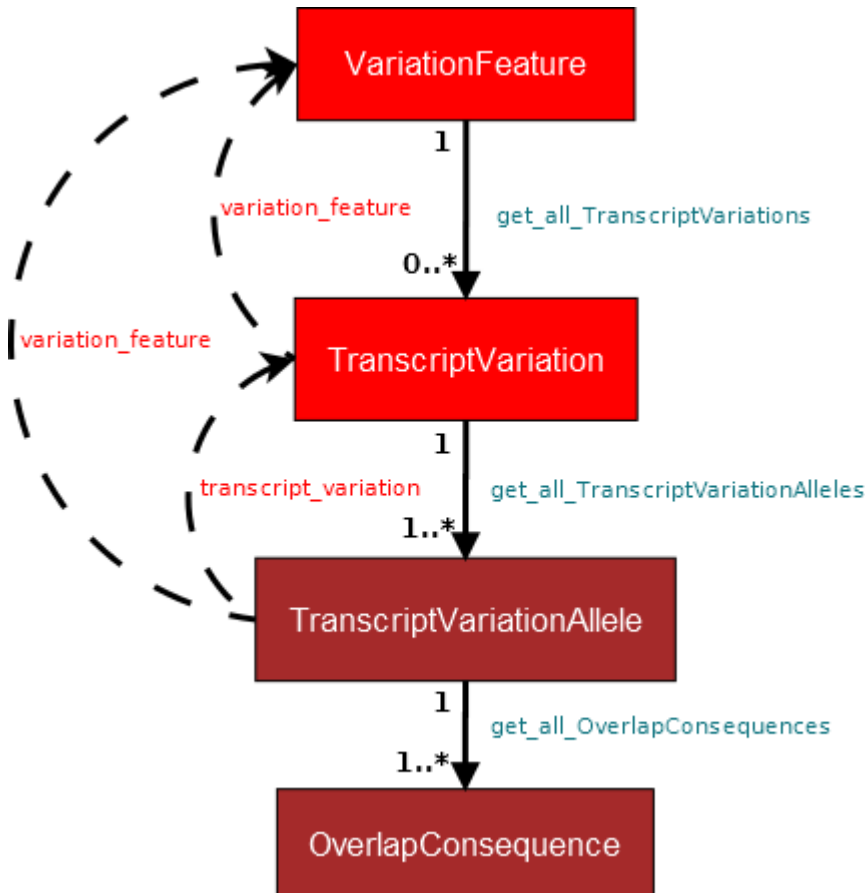


Genes and  
regulation

- 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 (default)	missense_variant	<code>\$oc-&gt;SO_term()</code>
Ensembl term	NON_SYNONYMOUS_CODING	<code>\$oc-&gt;display_term()</code>
NCBI term	missense	<code>\$oc-&gt;NCBI_term()</code>
Predicate	True/False	<code>\$oc-&gt;predicate(\$tva)</code>

# Transcript variation summary



← Variation information  
at the genomic level

← VariationFeature information  
at the transcript level

← Specific allele in the transcript  
and its consequence(s)

← Consequence(s) terms

# Protein function predictions



- We run two tools to predict how missense mutations are likely to affect protein function
- SIFT:
  - Uses sequence homology and amino acid similarity to calculate if the substitution is *tolerated* or *deleterious*
  - Supported species: human, chicken, cow, mouse, pig, rat and zebrafish
- PolyPhen:
  - Uses sequence homology, PDB 3D structures, Pfam annotation etc. to predict if a substitution is *probably damaging*, *possibly damaging* or *benign*
  - Supported species: human only
- Access methods on a *TranscriptVariationAllele*:
  - `$tva->sift_prediction()`
  - `$tva->polyphen_score()`

# Exercise 6a

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

**HINT:** `fetch_all_by_Transcripts` method requires a listref of objects; you have only one so use e.g.  
`[$transcript]` instead of `$transcript`

\* if exist

# Exercise 6b

- Fetch all the coding allele transcript variations in transcript ENST00000001008 in human and retrieve the following:
  - Allele string
  - Codon (with the allele position displayed)
  - Amino acids
  - Sift and PolyPhen predictions\*

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

**HINT:** first you will need to fetch the transcript variation objects like in the exercise 6a

\* if exist



# Predicting consequences

- Consequences can also be predicted for novel variant loci
- Can use the Variant Effect Predictor tool
  - Website version in the Ensembl website
  - Stand-alone Perl script also available for large volumes of data
- Also simple through API:

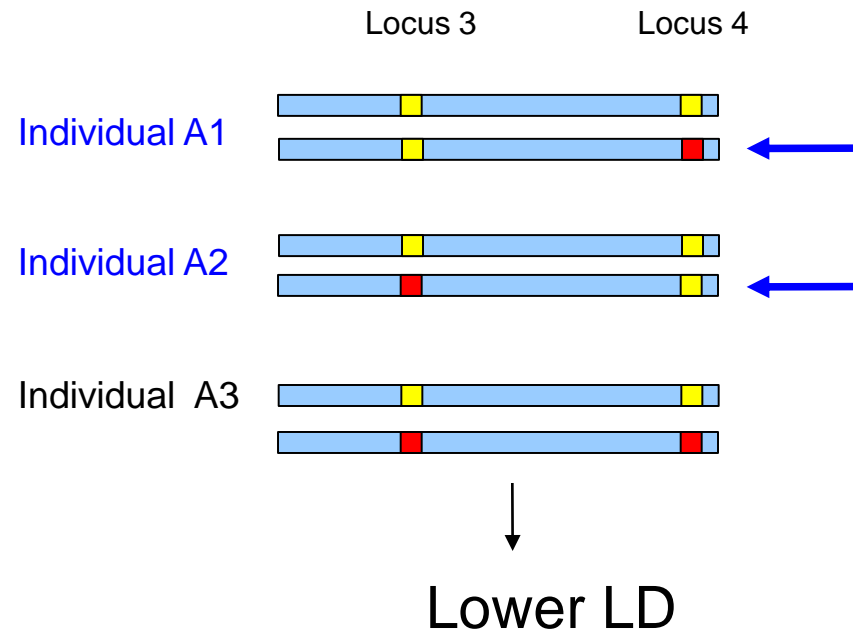
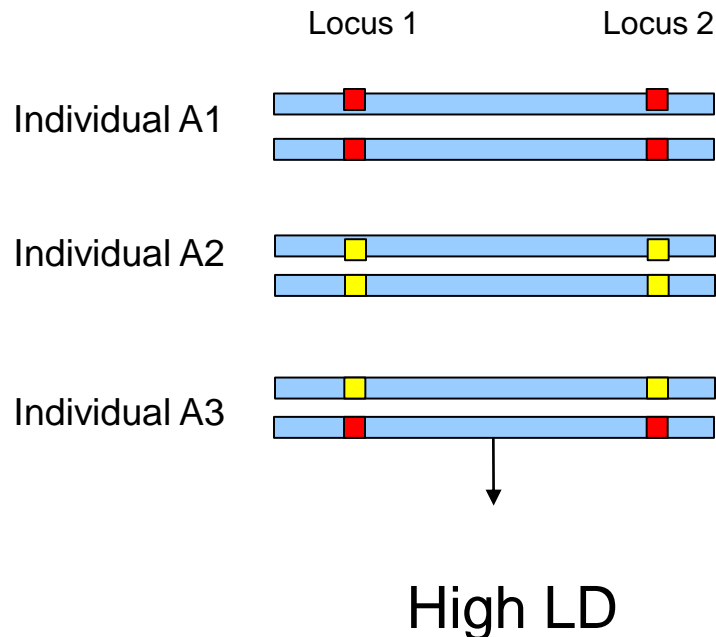
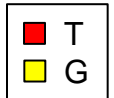
```
my $vf = Bio::EnsEMBL::Variation::VariationFeature->new(  
-start      => $start,  
-end        => $end,  
-slice      => $slice,      # slice object (location)  
-allele_string => "C/T",  
-strand     => 1,  
-adaptor    => $vfa,      # variation feature adaptor  
);  
  
my $tvs = $vf->get_all_TranscriptVariations();
```

[http://www.ensembl.org/info/docs/api/variation/variation\\_tutorial.html](http://www.ensembl.org/info/docs/api/variation/variation_tutorial.html)

# 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 data for the HapMap and 1000 genomes populations
- Two common measures
  - $r^2$ ,  $D'$  ( $r^2 = 1 \rightarrow$  perfect LD)



# LD feature container objects

- Represents an instance of a container of pair-wise LD values in a region (slice)
  - calculated on the fly
  - can contain values for a single or multiple populations
- Most methods return hash references
  - To retrieve e.g.  $r^2$  value use `$ld_hash_ref->{r2}`
  - Some of these hash elements contain objects

Keys	Value
variation1	variation feature object of one of the two variations used for the calculation
variation2	variation feature object of one of the two variations used for the calculation
r2	0 to 1

# Exercise 7

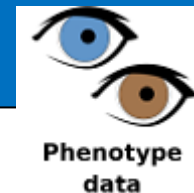


- Create an LD feature container for the region from 22124000 to 22126000 on human chromosome 9 and find the names of all pairs of SNPs in perfect LD in the population named “1000GENOMES:phase\_1\_GBR”

## HINTS:

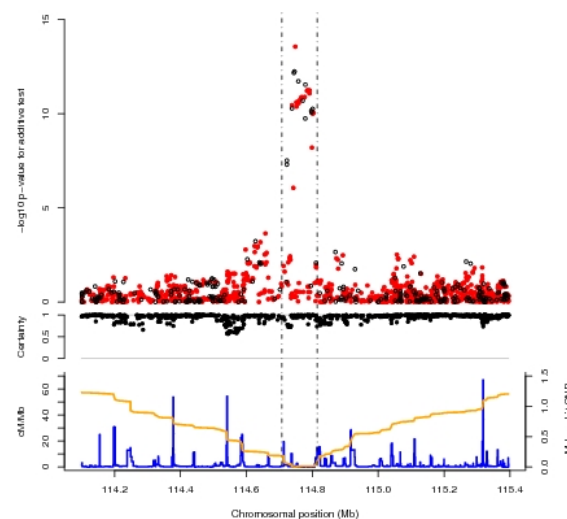
- perfect LD means  $r^2 = 1$
- the `fetch_by_Slice()` method can take an optional population argument.

# Phenotype information



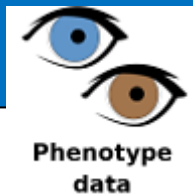
We import data linking phenotypes to short variations, structural variations, genes and QTLs from curated and experimental sources including:

- NHGRI GWAS catalog
- OMIM
- OMIA
- UniProt
- HGMD
- COSMIC
- ClinVar
- EGA
- Association data from large consortia such as MAGIC and GIANT \*



\* Only data reaching genome-wide significance according to the study's criteria is reported by default

# Phenotype information

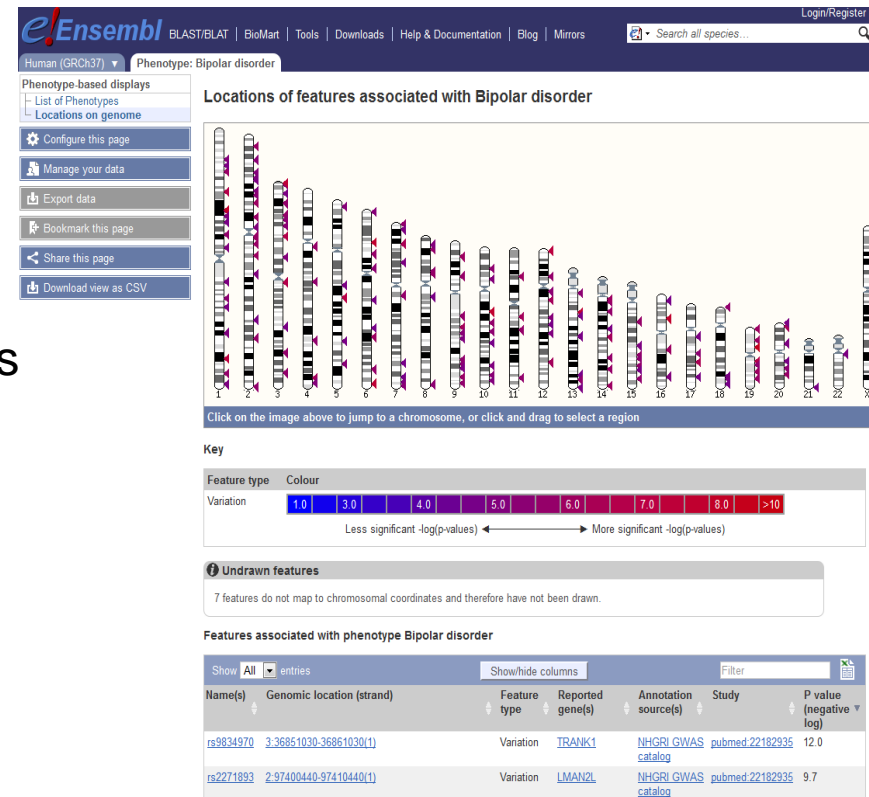


Phenotypes are mapped to genomic locations via the variations or genes they have reported associations with to create a PhenotypeFeature

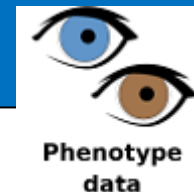
We are collaborating with groups working on phenotype ontologies, but the data current held is as it appears in the source database.

Attributes types held for phenotype features include:

- clinical significance reported by ClinVar
- reported genes
- risk allele
- p-value
- inheritance type
- odds ratio
- beta coefficient



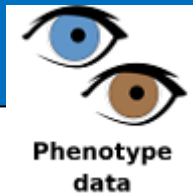
# Phenotype Feature object



- Represents an association between a phenotype and a genomic feature
  - Can be retrieved from a variation or slice object as well as the adaptor
  - A range of object types are supported
  - A range of possible attributes are supported
  - Not all features have the same attributes available

Attribute	Example value(s)	Method(s)
Feature object id	Variation name, gene name	<code>\$pf-&gt;object_id()</code> <b>*returns name</b>
Phenotype	Phenotype object	<code>\$pf-&gt;phenotype()</code> <b>*returns object - for description use:</b> <code>\$pf-&gt;phenotype-&gt;description()</code>
Reported gene(s)	CDKN2BAS, NOTCH2	<code>\$pf-&gt;associated_gene()</code>
p-value	6e-07	<code>\$pf-&gt;p_value()</code>

# Exercise 8



Find all phenotypes associated with the human variation named rs11765954 and find

- The phenotype description
- The associated p-value
- For any phenotypes with an associated risk allele, find the frequency of this allele in the 1000 genomes project populations
- Extra:
  - Try extracting all phenotype features for human variants rs10209020 and rs1052373 after setting the DBAdaptor variation `$adaptotr>db->include_non_significant_phenotype_associations(1)`

Hints and tips:

- The 1000 genomes populations' names begin "1000GENOMES"





We import variation citation data from:

- dbSNP
  - Publication information is submitted with the variant submission. The publication may describe a variant discovery project or data may have been extracted from a pre-existing report and formatted for database submission by curators.
- 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.

# Publication object



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

Attribute	Example value(s)	Method(s)
Cited variants	Array of variation objects	<code>\$p-&gt;variations()</code>
Title	Mitochondrial acetylation and diseases of aging.	<code>\$p-&gt;title()</code>
Authors	Wagner GR, Payne RM.	<code>\$p-&gt;authors()</code>
Year	2011	<code>\$p-&gt;year()</code>
PubMed ID	21437190	<code>\$p-&gt;pmid()</code>

# Exercise 9



Obtain variation objects for the human variations:  
rs2234693 and rs3730070

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

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

The VEP is a tool to predict the functional consequences of variants, using the Ensembl API and data

- A web interface is available
  - No set up required
- A standalone Perl script is available
  - Rapid large scale analysis using local data caches
  - Secure local analysis of private data
- Several input formats are supported: VEP tabulated format, VCF, Pileup, HGVS (e.g. 5:g.140532T>C), Variants identifiers (e.g. rs699)
- Multiple analysis and data filtering options are available
- Multiple output formats are supported
- The stand alone method can be customised using the plugin system

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

Example of simple input:

Chr	Start	End	Allele	Strand	[Identifier]
-----	-------	-----	--------	--------	--------------

1	881907	881906	-/C	+	
<b>5</b>	<b>140532</b>	<b>140532</b>	<b>T/C</b>	<b>+</b>	
12	1017956	1017956	T/A	+	
2	946507	946507	G/C	+	
14	19584687	19584687	C/T	-	
19	66520	66520	G/A	+	
8	150029	150029	A/T	+	
1	230845794	230845795	A/-		+

Reporting options include:

- transcript consequence
- protein impact
- co-location with known variants
- global minor allele frequency
- HGVS notation
- overlaps with regulatory regions
- overlaps with structural variants
- co-location with failed variants
- PubMed ids for citations

Example command line:

```
perl variant_effect_predictor.pl -i input.txt -o output.txt
```

Example output from web interface:

Variation					type		in cDNA	in CDS	in protein	acid change	change	Variation	
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000535572</a>	Transcript	stop_lost	7376	6403	2135	*K	Tag/Aag	<a href="#">rs55650617</a>	GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000468231</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=3699; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000535376</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=4444; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000430095</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=3755; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000315939</a>	Transcript	stop_lost	7790	7147	2383	*K	Tag/Aag	<a href="#">rs55650617</a>	GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000488642</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=3968; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000340908</a>	Transcript	stop_lost	5926	5926	1976	*K	Tag/Aag	<a href="#">rs55650617</a>	GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000228345</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=3365; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000530271</a>	Transcript	stop_lost	8641	8641	2881	*K	Tag/Aag	<a href="#">rs55650617</a>	GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000002016</a>	<a href="#">ENST00000481052</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=3705; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000537603</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=589; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000543065</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=4303; GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000537687</a>	Transcript	stop_lost	8570	7927	2643	*K	Tag/Aag	<a href="#">rs55650617</a>	GMAF=A:0.0005
12_1017956_T/A	<a href="#">12:1017956</a>	A	<a href="#">ENSG000000060237</a>	<a href="#">ENST00000544559</a>	Transcript	downstream_gene_variant	-	-	-	-	-	<a href="#">rs55650617</a>	DISTANCE=834; GMAF=A:0.0005

# Getting more information

- Variation database schema
  - ➔ [http://www.ensembl.org/info/docs/variation/variation\\_schema.html](http://www.ensembl.org/info/docs/variation/variation_schema.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](http://www.ensembl.org/info/docs/api/variation/variation_tutorial.html)
- Ensembl developers mailing list
  - ➔ [dev@ensembl.org](mailto:dev@ensembl.org)

# Ensembl Acknowledgements

## Ensembl Team

### Ensembl Variation

Fiona Cunningham,

Laurent Gil

Will McLaren

Anja Thormann

Paul Flicek, Steve Searle and  
the entire Ensembl Team

## Funding

**wellcome**trust

EMBL



National  
Human Genome  
Research Institute



**BBSRC**  
bioscience for the future

European Commission  
Framework Programme 7



**Quantomics**

From Sequence to Consequence :  
Tools for the Exploitation of Livestock Genomes

