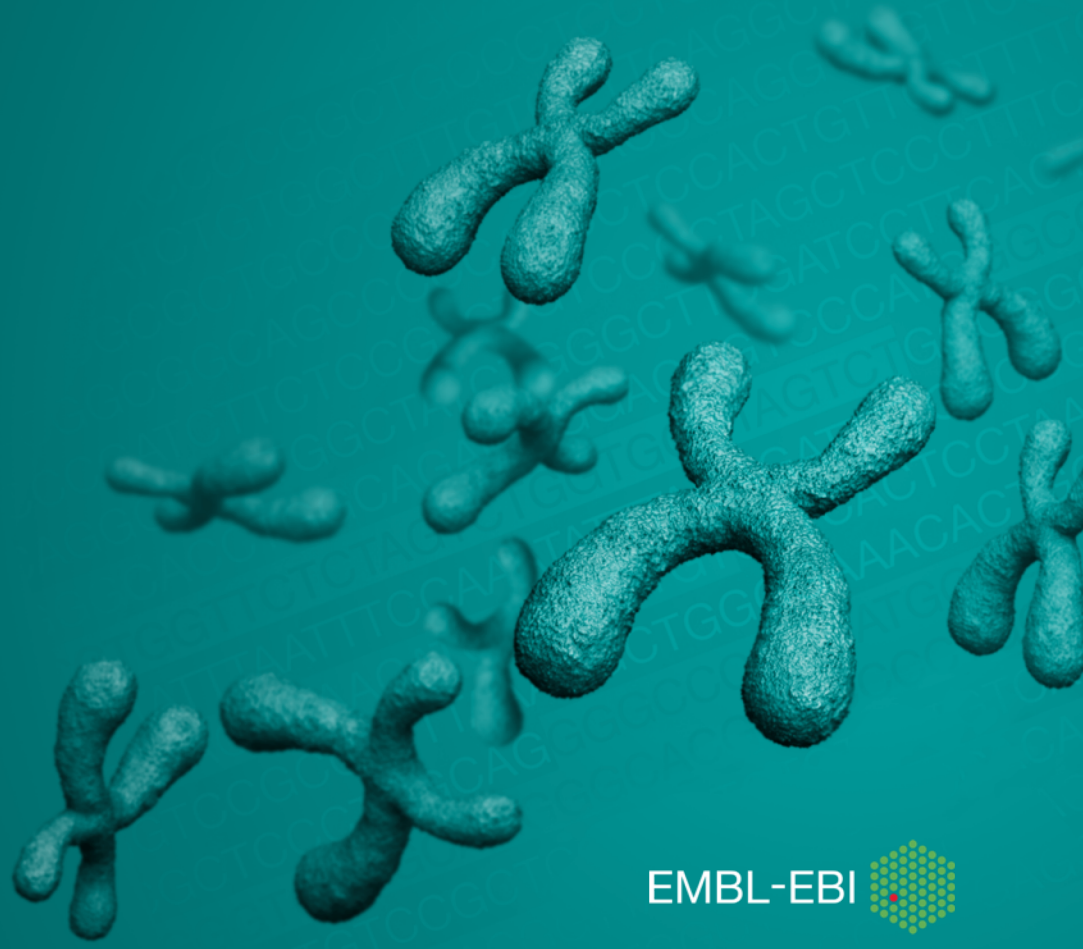
Ensembl Variation API – Part 2

Andrew Parton

April 24th 2018



EMBL-EBI

Useful Links

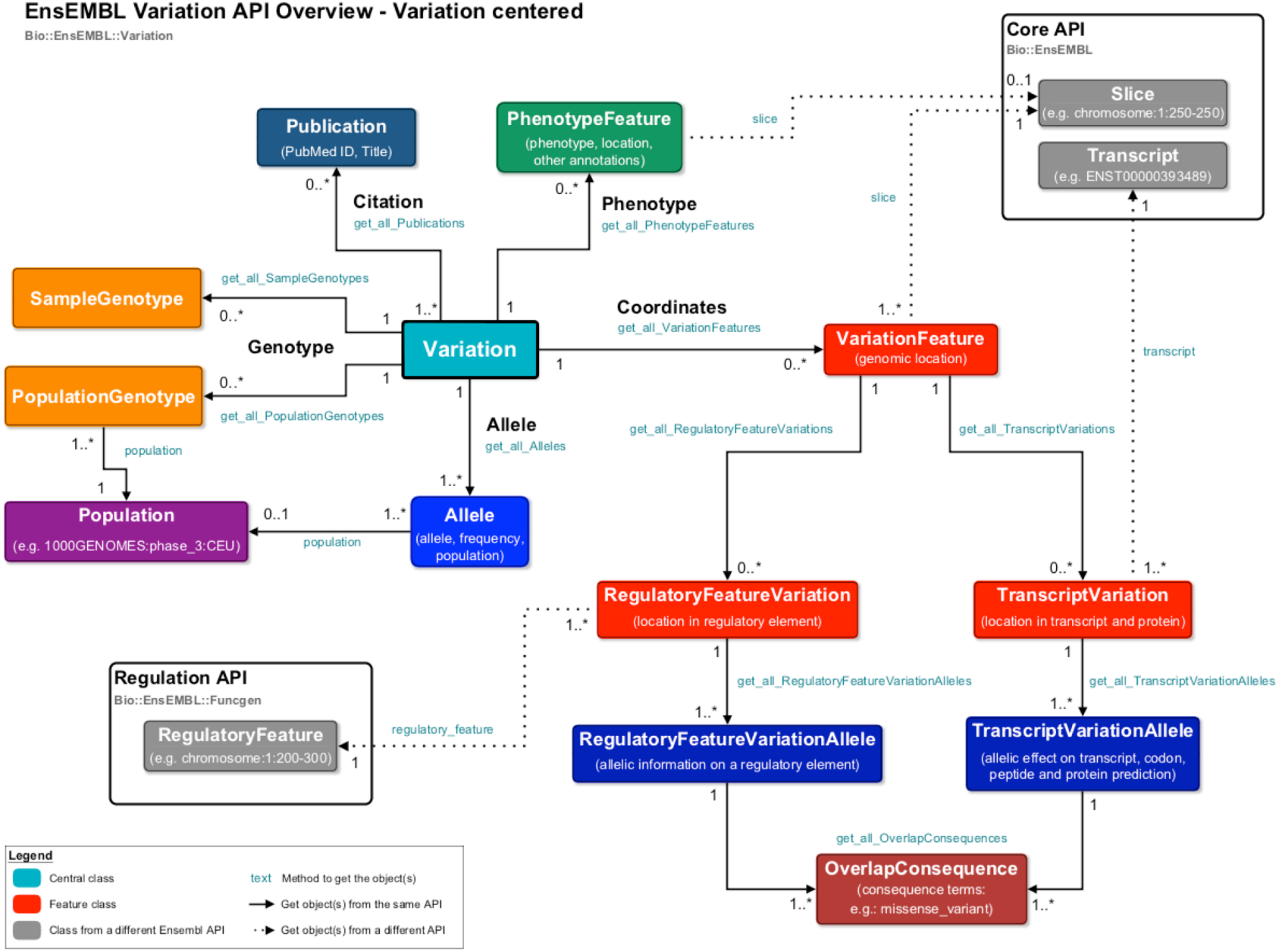
- Course Material
  - https://github.com/Ensembl/ensembl-presentation/tree/master/API/Variation/Hinxton1804

- 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

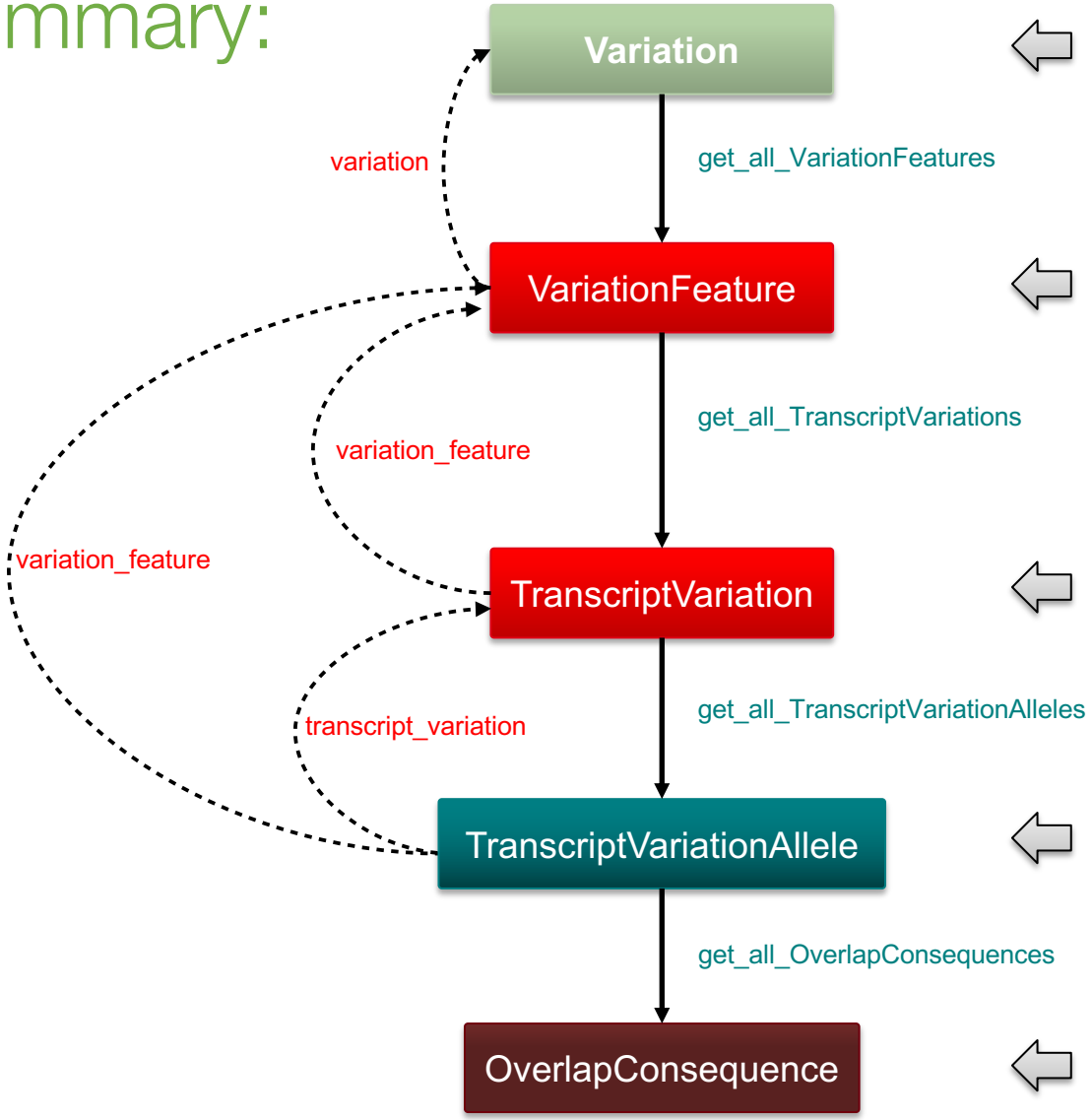
EnsEMBL Variation API Overview - Variation centered

Bio::EnsEMBL::Variation



EMBL-EBI

API Summary:



Variant information

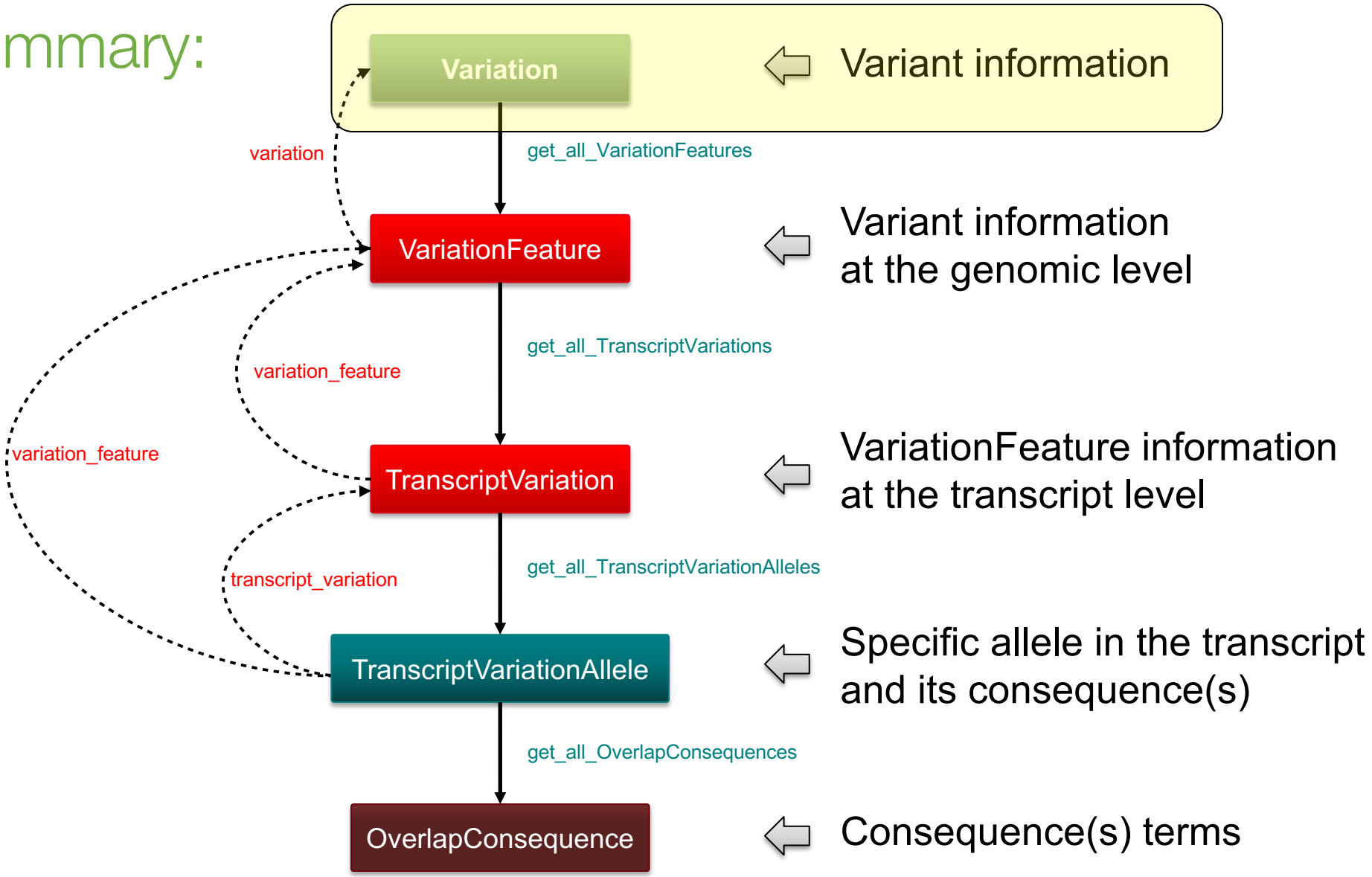
Variant information
at the genomic level

VariationFeature information
at the transcript level

Specific allele in the transcript
and its consequence(s)

Consequence(s) terms

API Summary:


Variation
Variant information
variation
get_all_VariationFeatures
VariationFeature
Variant information
at the genomic level
get_all_TranscriptVariations
variation_feature
variation_feature
TranscriptVariation
VariationFeature information
at the transcript level
get_all_TranscriptVariationAlleles
transcript_variation
TranscriptVariationAllele
Specific allele in the transcript
and its consequence(s)
get_all_OverlapConsequences
OverlapConsequence
Consequence(s) terms

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');
```

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

For the human variants rs1333049, rs56385407, COSM998 and CI003207, retrieve the following information:

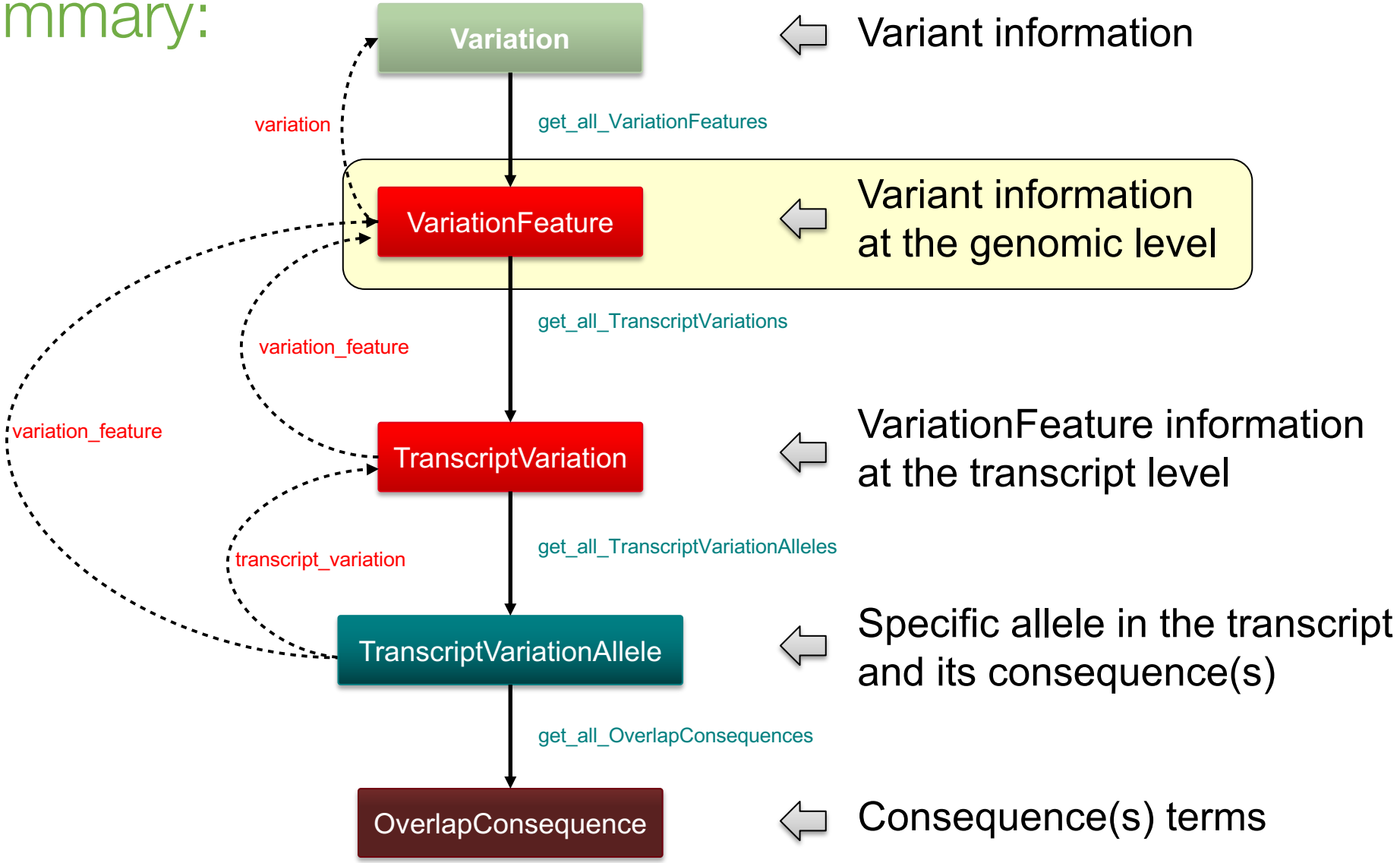
- Variation class
- Source

Find the same information for the platypus (Ornithorhynchus anatinus) variant rs69772326.

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

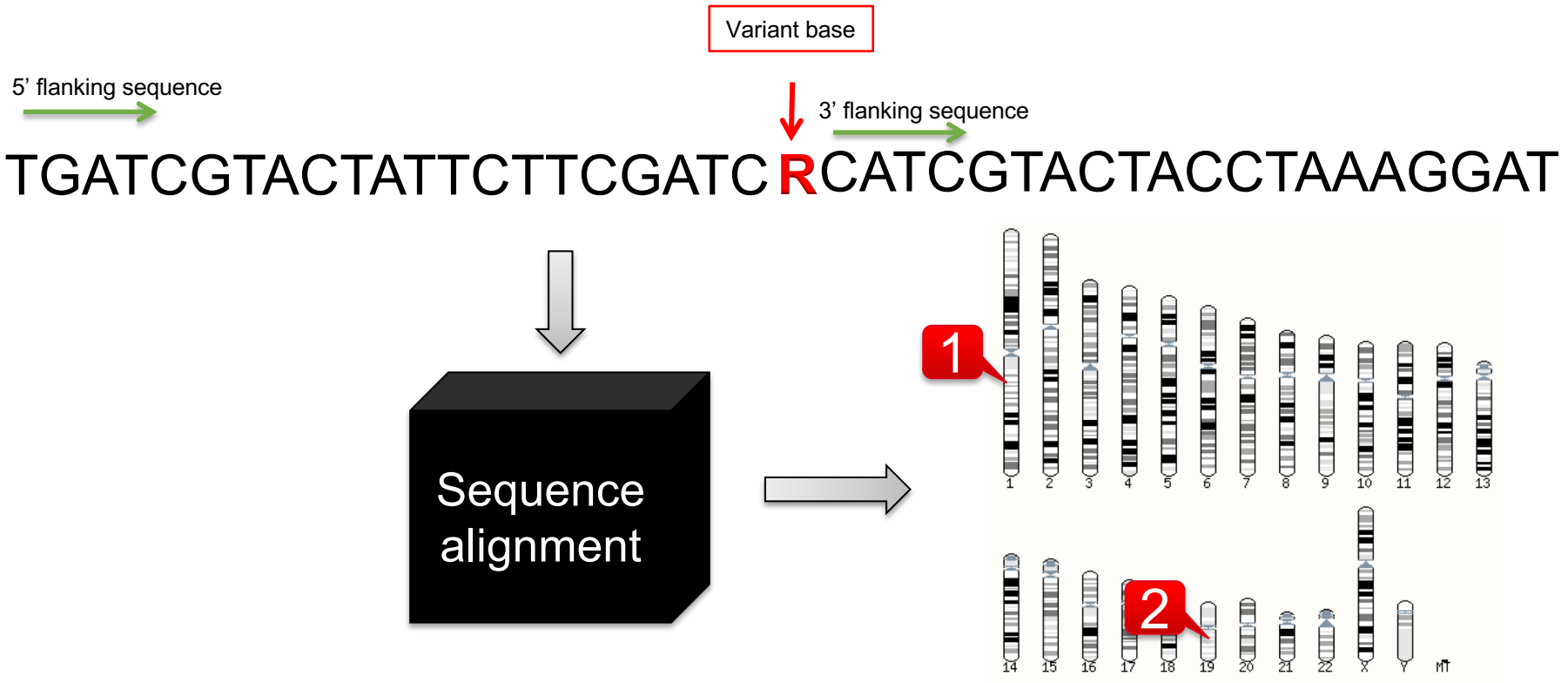
API Summary:


Variation
Variant information
variation
get_all_VariationFeatures
VariationFeature
Variant information
at the genomic level
get_all_TranscriptVariations
variation_feature
variation_feature
TranscriptVariation
VariationFeature information
at the transcript level
get_all_TranscriptVariationAlleles
transcript_variation
TranscriptVariationAllele
Specific allele in the transcript
and its consequence(s)
get_all_OverlapConsequences
OverlapConsequence
Consequence(s) terms

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()`
		
		`$vf->seq_region_start()`
		`$vf->seq_region_end()`
			} slice relative
			
			} 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 2

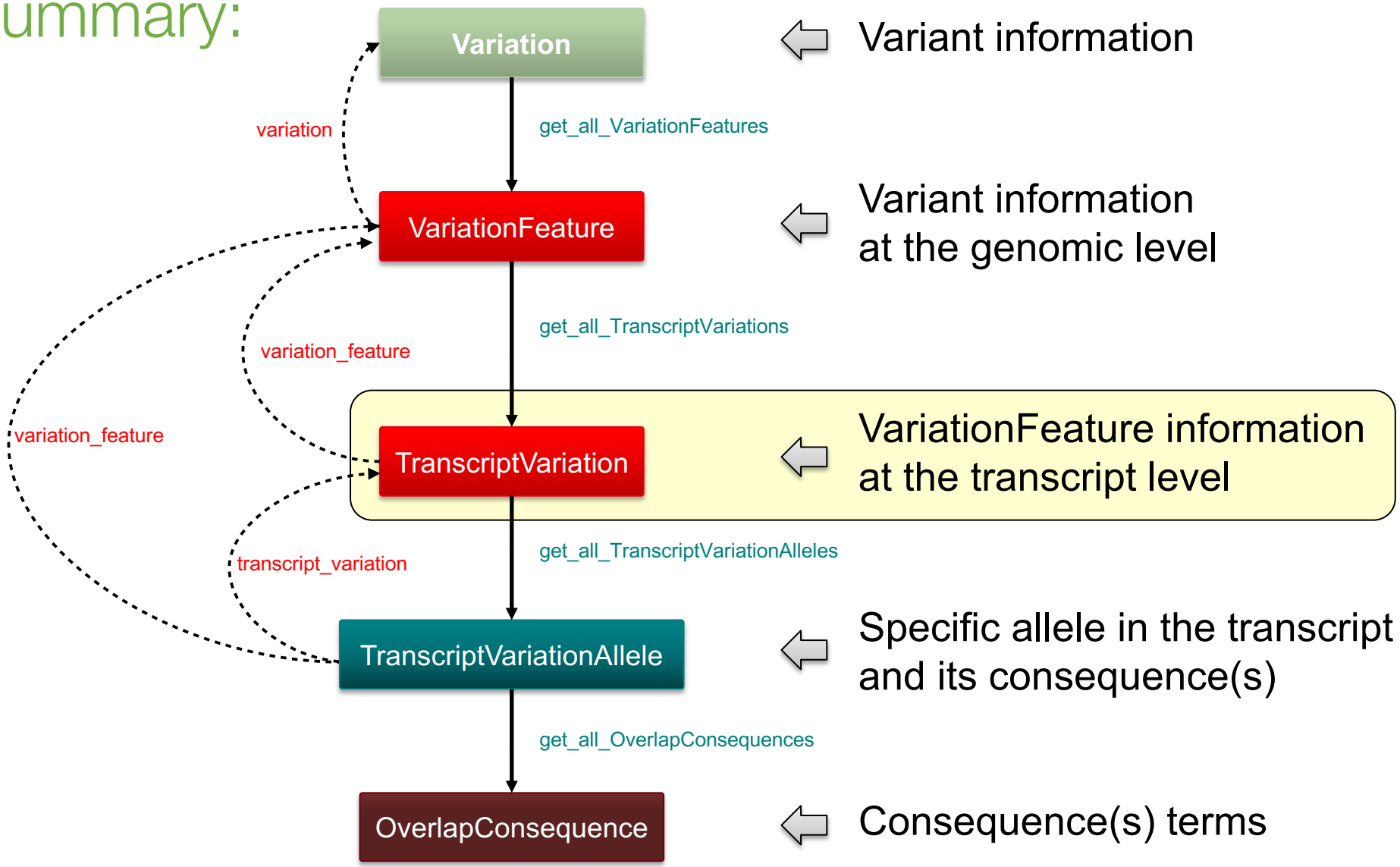
- Retrieve all variations in human located on chromosome 13 from 48987260 to 48990000 and get:
  - Variation name
  - Alleles (“allele_string”)
  - Location (e.g. Chromosome:Start-End)

- Extra:

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

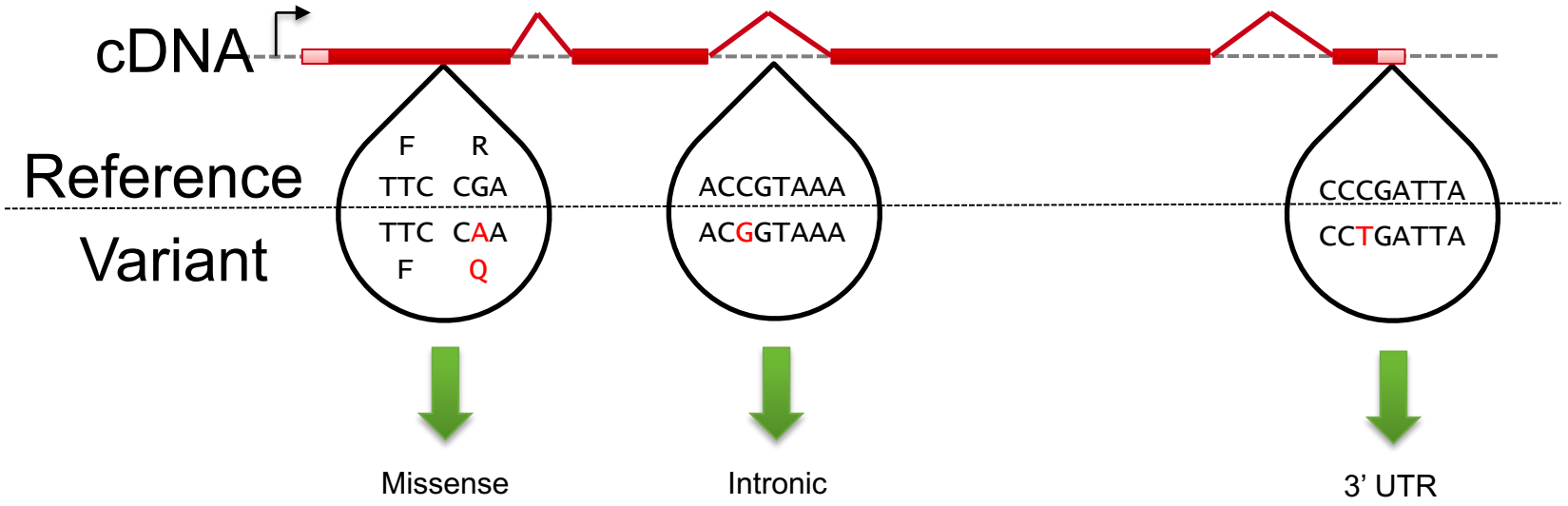
Hint: usually when you have to search something by location, you need to use the **Slice** object (Core API) first

API Summary:

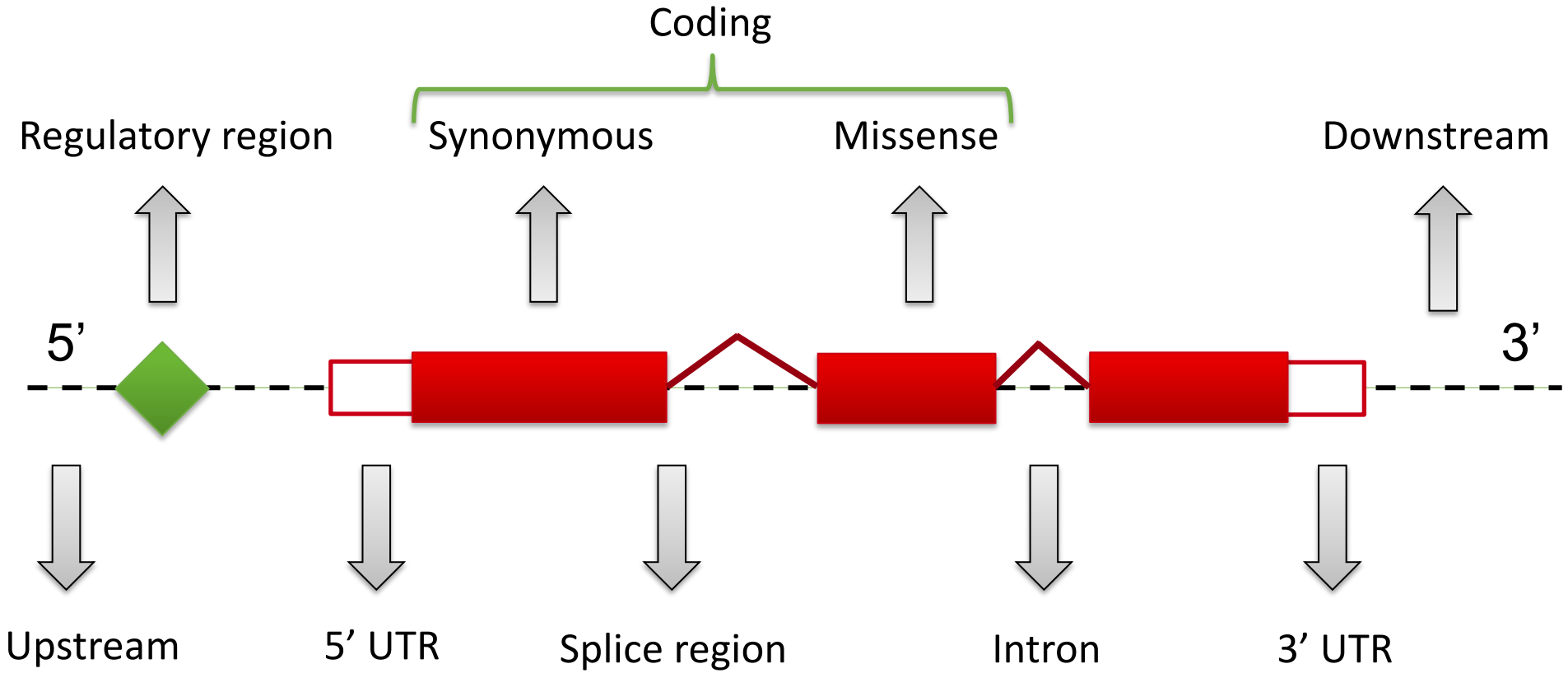

Variation
Variant information
variation
get_all_VariationFeatures
VariationFeature
Variant information
at the genomic level
get_all_TranscriptVariations
variation_feature
variation_feature
TranscriptVariation
VariationFeature information
at the transcript level
get_all_TranscriptVariationAlleles
transcript_variation
TranscriptVariationAllele
Specific allele in the transcript
and its consequence(s)
get_all_OverlapConsequences
OverlapConsequence
Consequence(s) terms

Variation Consequences

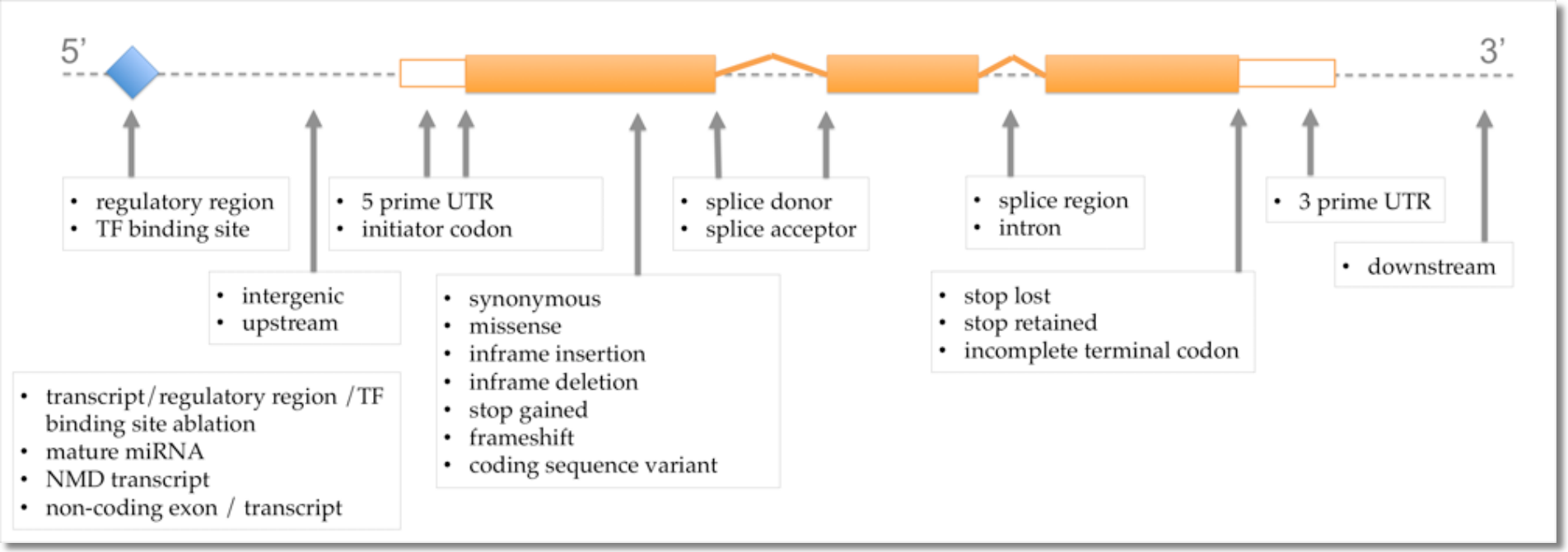
Defines consequence of a variation in relation to the transcript structure it overlaps



Variation Consequences



Variation Consequences



Sequence ontology provides controlled vocabulary to describe consequences

The Sequence Ontology

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

Transcript Variation

- 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:

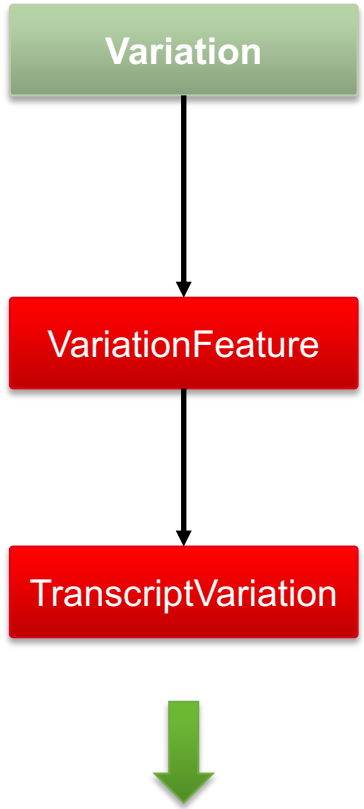
```
$vf->display_consequence()
```

- 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()`	



Name: rs201036189

Location: 1:206342907
Allele string: C/T

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

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

Exercise 3

- 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 *

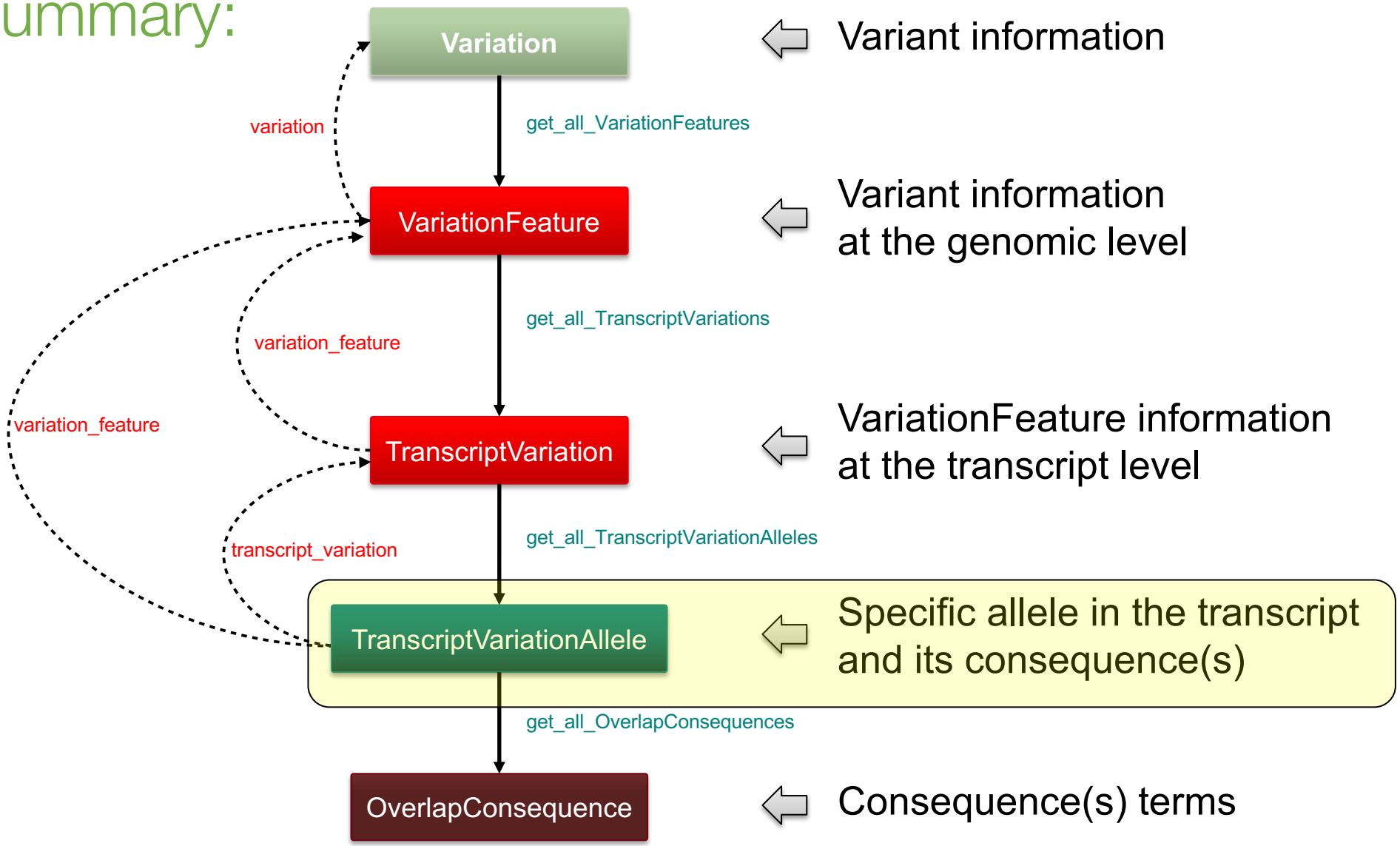
* If data exists

- Filter these transcript variation objects to find variants 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)
```

API Summary:


Variation
Variant information
variation
get_all_VariationFeatures
VariationFeature
Variant information
at the genomic level
get_all_TranscriptVariations
variation_feature
variation_feature
TranscriptVariation
VariationFeature information
at the transcript level
get_all_TranscriptVariationAlleles
transcript_variation
TranscriptVariationAllele
Specific allele in the transcript
and its consequence(s)
get_all_OverlapConsequences
OverlapConsequence
Consequence(s) terms

Transcript Variation Allele

- 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	`$tva->display_codon_allele_string()`
SIFT and PolyPhen predictions	0.01, 0.44, 1	`$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)

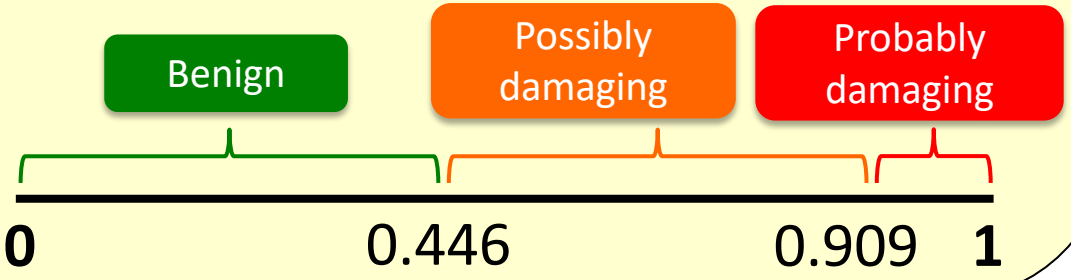
Supported species: Human, Mouse, Chicken, Cow, Dog, Horse, Pig, Rat, Sheep, Zebrafish

- `$tva->sift_score()`
- `$tva->polyphen_score()`

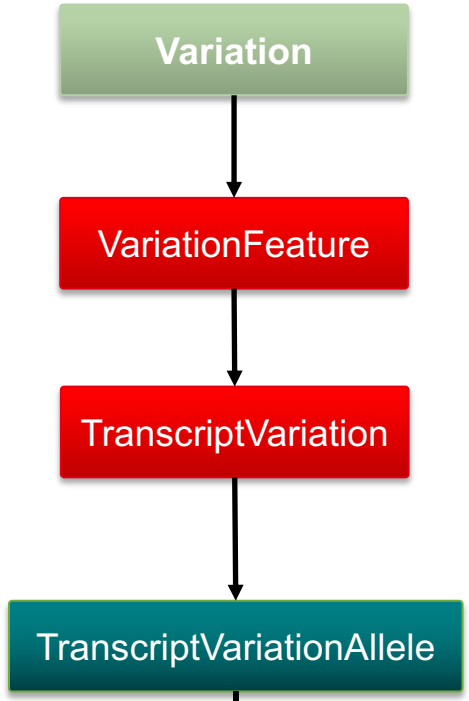
PolyPhen

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



Example



Name: rs191084422

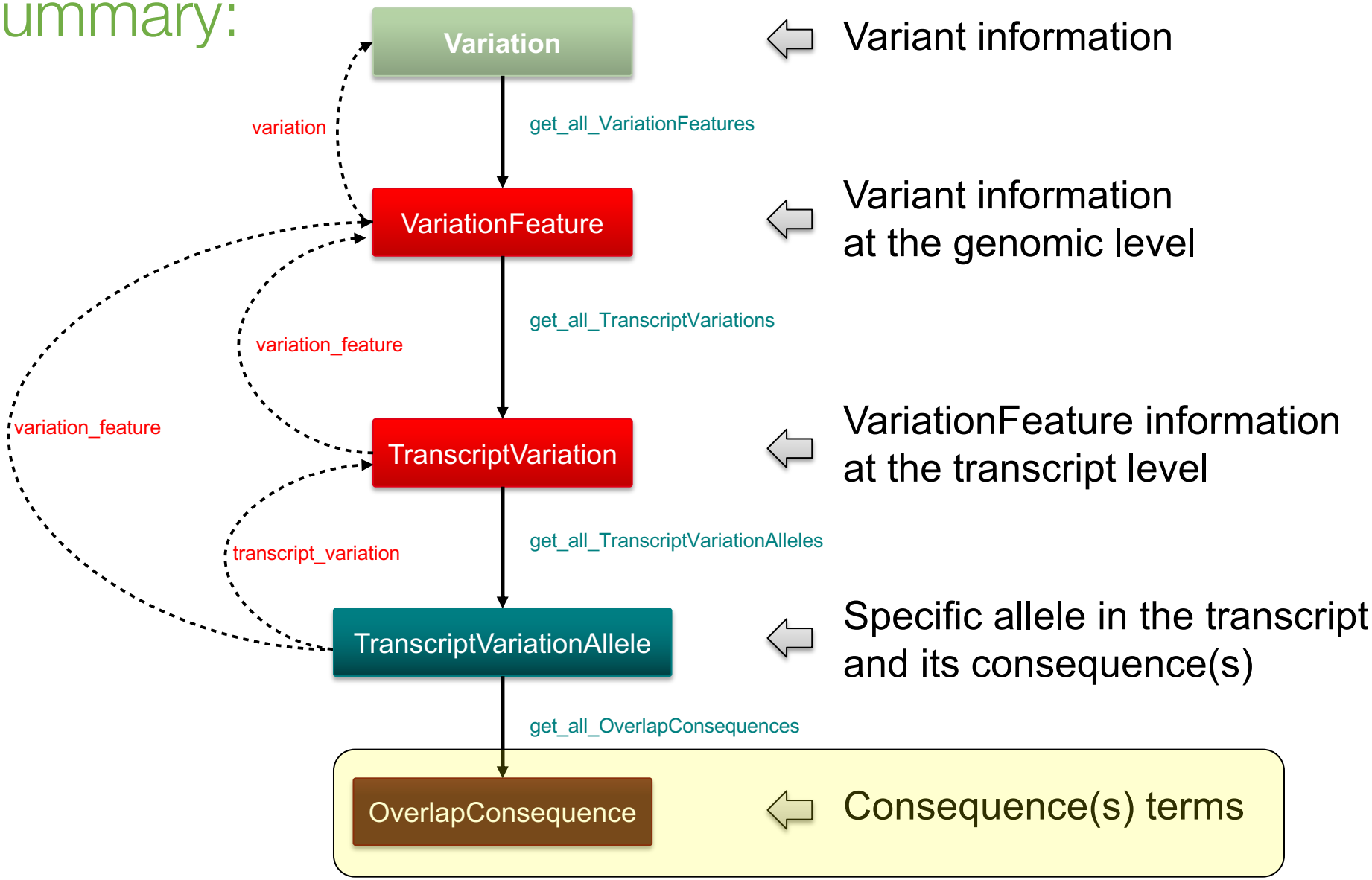
Location: 6:56553483
Allele string: C/G/T

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

API Summary:


Variation
Variant information
variation
get_all_VariationFeatures
VariationFeature
Variant information
at the genomic level
get_all_TranscriptVariations
variation_feature
variation_feature
TranscriptVariation
VariationFeature information
at the transcript level
get_all_TranscriptVariationAlleles
transcript_variation
TranscriptVariationAllele
Specific allele in the transcript
and its consequence(s)
get_all_OverlapConsequences
OverlapConsequence
Consequence(s) terms

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 4

- 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 data exists

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

Hint: You need to fetch the **TranscriptVariation** objects first,
as you did in Exercise 3