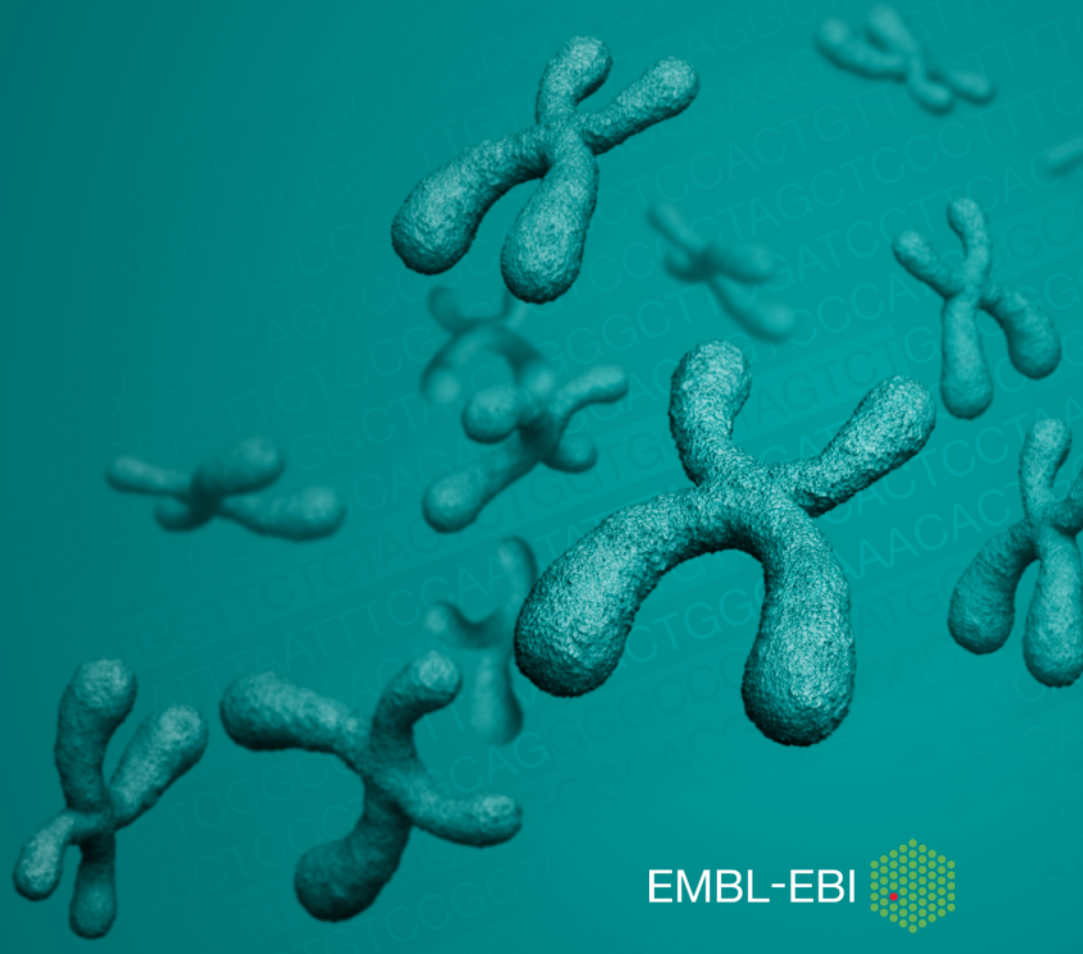
Ensembl Variation API – Part 1

Anja Thormann

Andrew Parton

April 24th 2018 Hinxton



EMBL-EBI

Important URLs

- 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
- Variation data documentation
  - http://www.ensembl.org/info/genome/variation/index.html

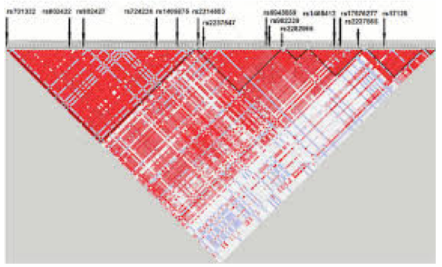
Genomic variation

image credit: humanae.tumblr.com

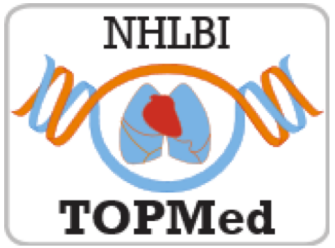
https://www.yourgenome.org/facts/what-is-genetic-variation

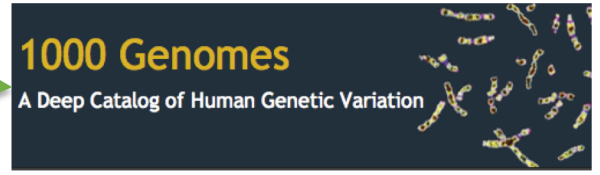
Sequencing Revolution

2000 2010 2015



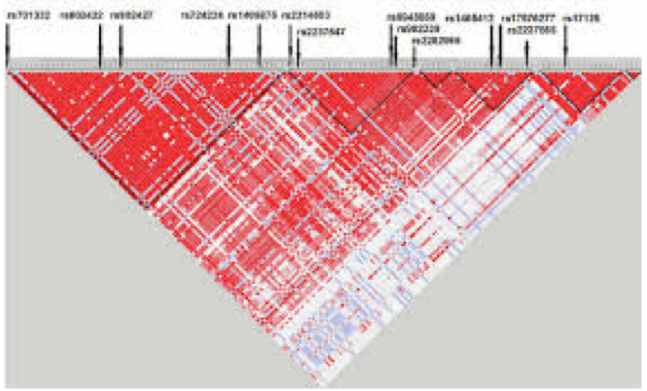
- ExAC
- gnomAD





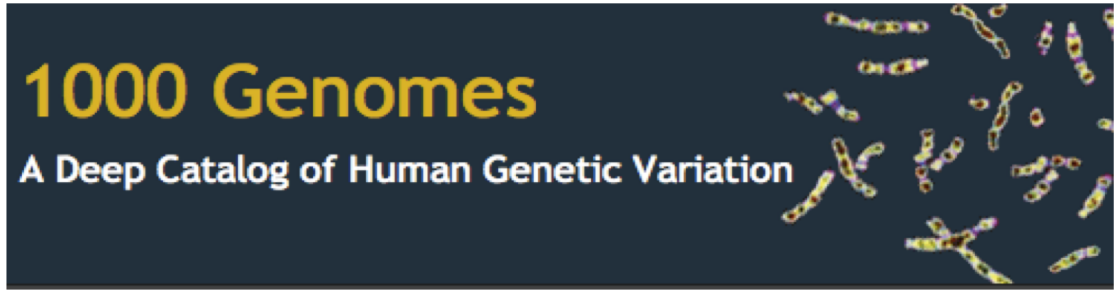
International HapMap consortium

- 1.6 million common single nucleotide polymorphisms (SNPs) from 1,184 individuals from 11 global populations
- Linkage disequilibrium – nonrandom association of alleles at different loci
- Sets of nearby SNPs on the same chromosome are inherited in blocks
- genome-wide association study GWAS: examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait



1000 Genomes Project

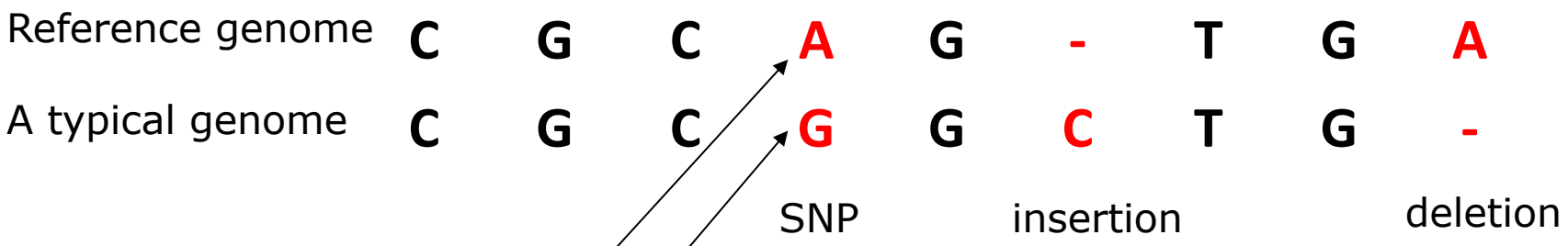
- Properties and distribution of common and rare variation
- Advances in sequence data generation, archiving and analysis
- 2,504 individuals from 26 populations
- Low-coverage whole-genome sequencing
- Deep exome sequencing
- Dense microarray genotyping
- 88 million variants



2,504 individuals from 26 populations

A typical genome

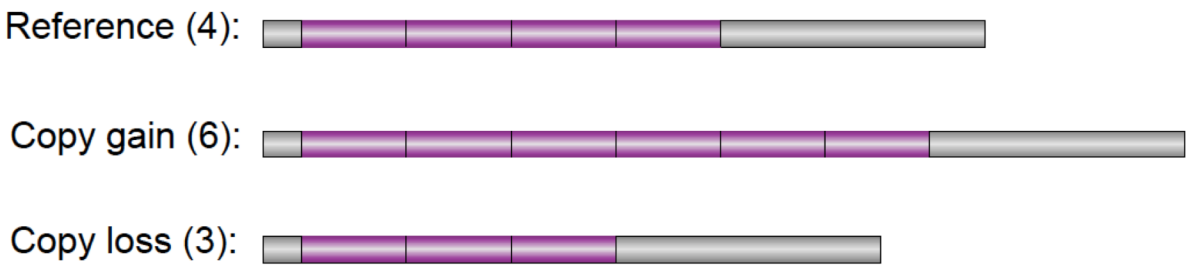
- Differs from the reference human genome at 4.1 million to 5.0 million sites
- >99.9% of variants consist of SNPs and short indels



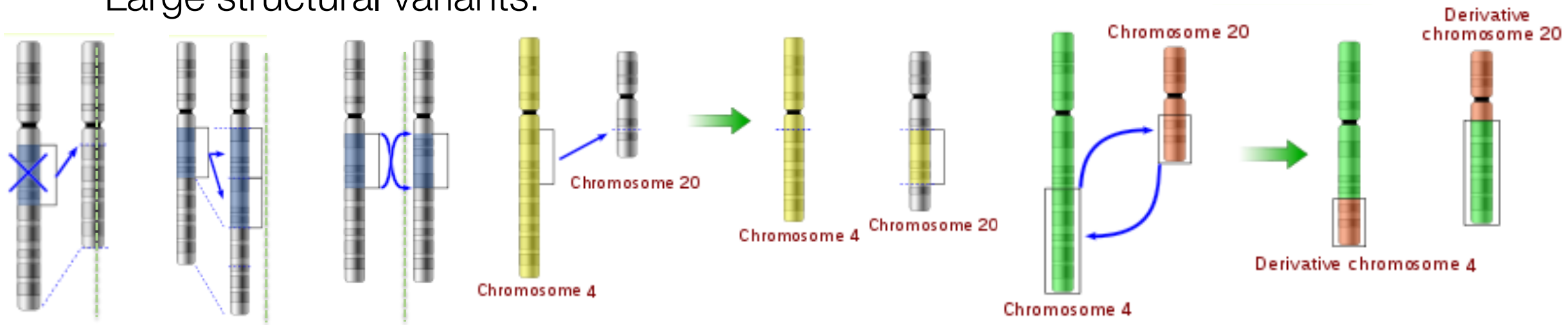
- Reference allele: A
- Alternative allele: G
- Contains 2,100 to 2,500 structural variants

Structural variation

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



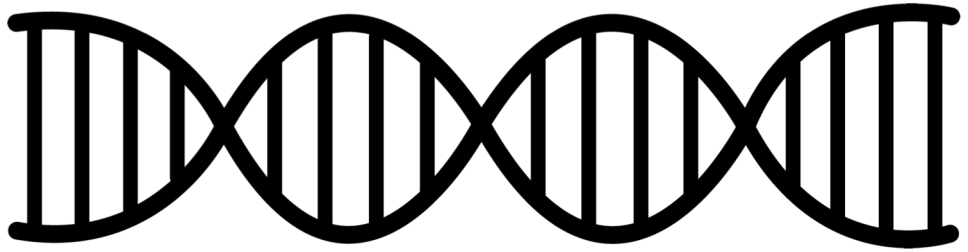
- Large structural variants:



deletion duplication inversion insertion translocation

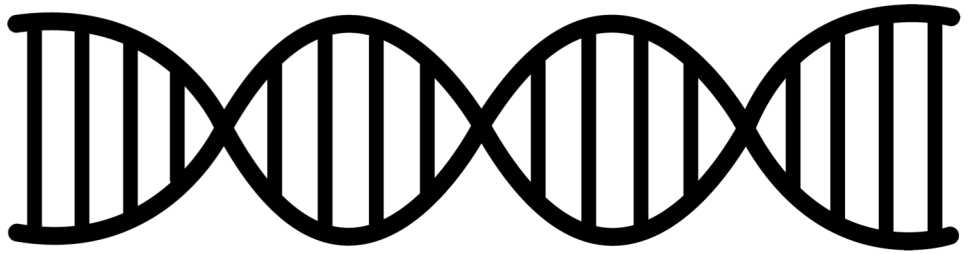
Diploid: 2 versions of each autosomal chromosome

A



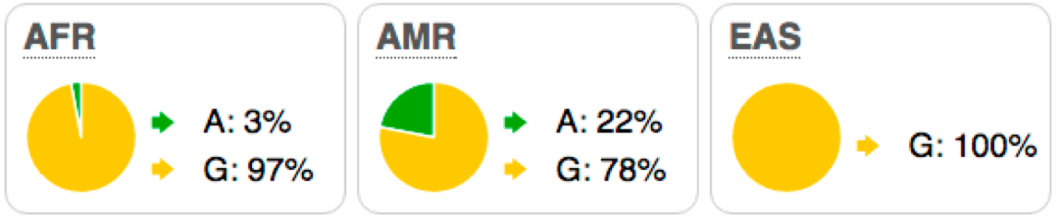
Mother

G



Father

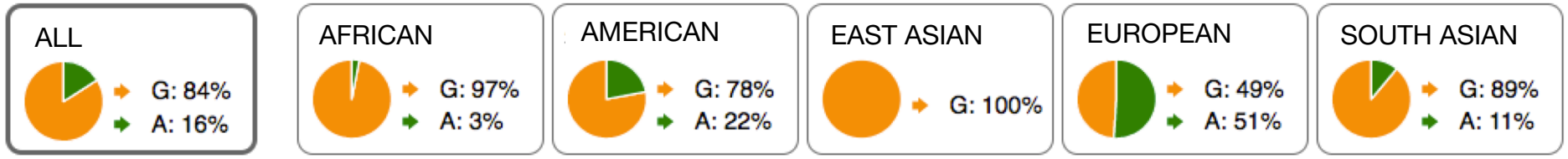
Genotype A|G



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

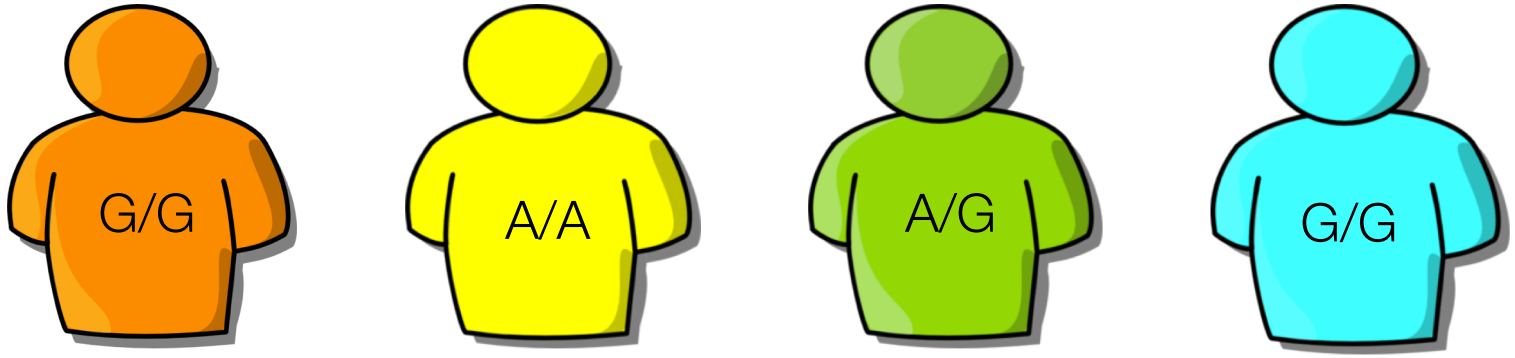
1000 Genomes Project Phase 3 allele frequencies



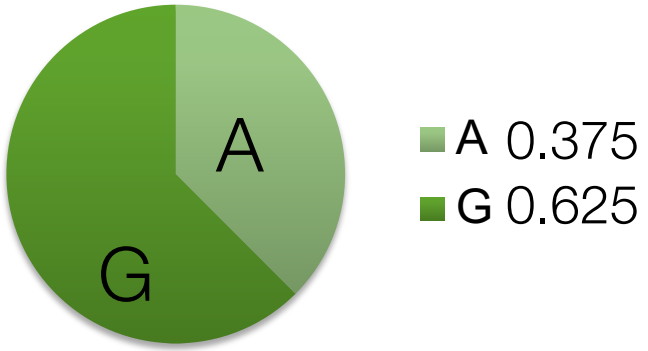
Genotype	What does it mean
G/G	Likely to be lactose intolerant.
G/A	Likely to be tolerant due to lactase persistence.
A/A	Lactose tolerant

Allele Frequency in a population

Chromosome 7: 5678987-5678987
Reference Allele: G

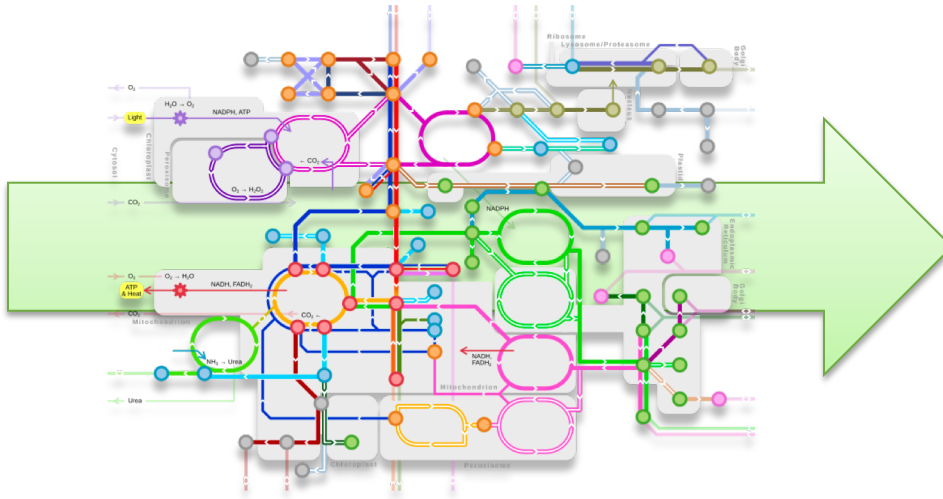


Count alleles: 3 A, 5 G
Frequency for A: 3/8 = 0.375
Frequency for G: 5/8 = 0.625

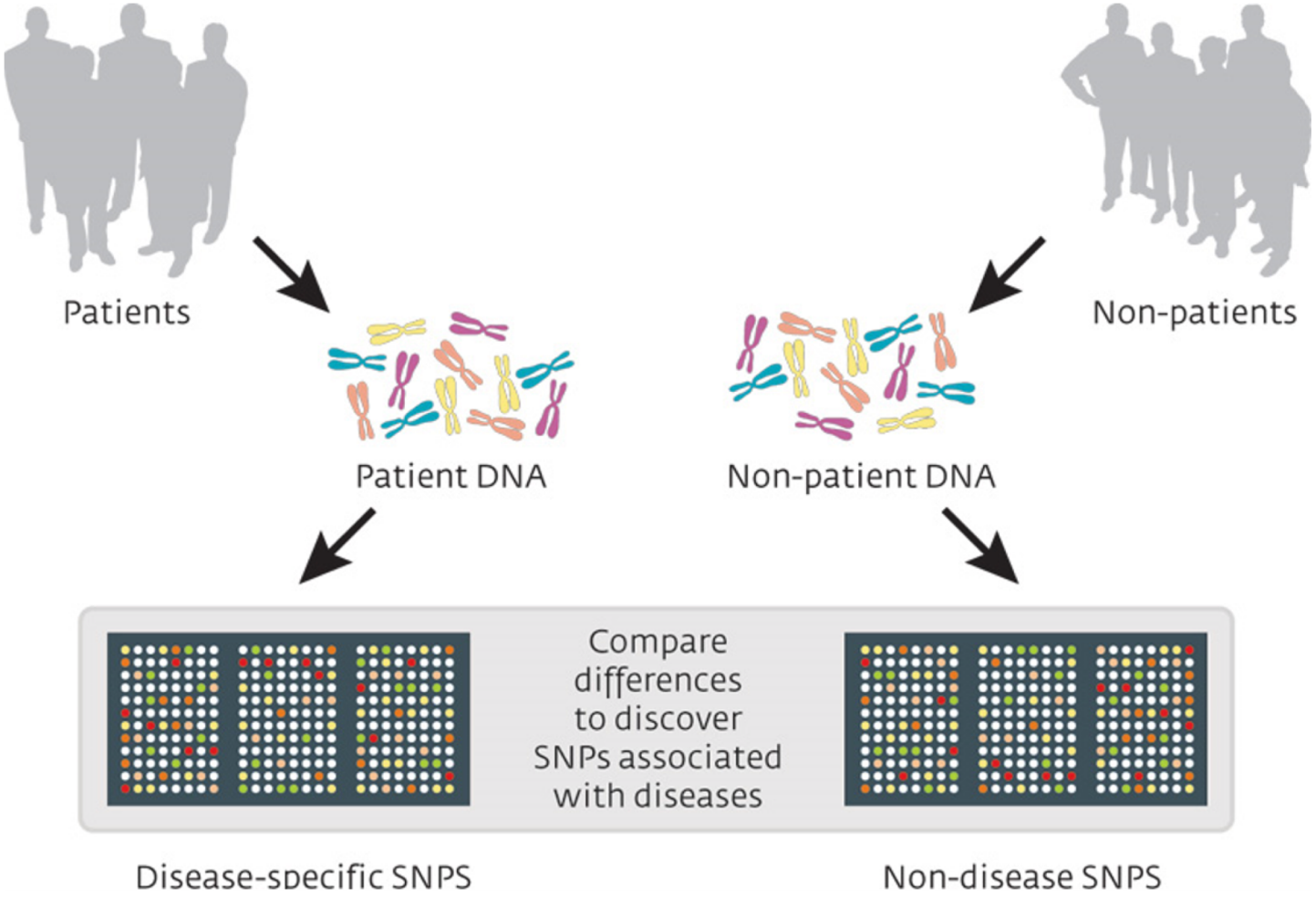


Genotype to Phenotype

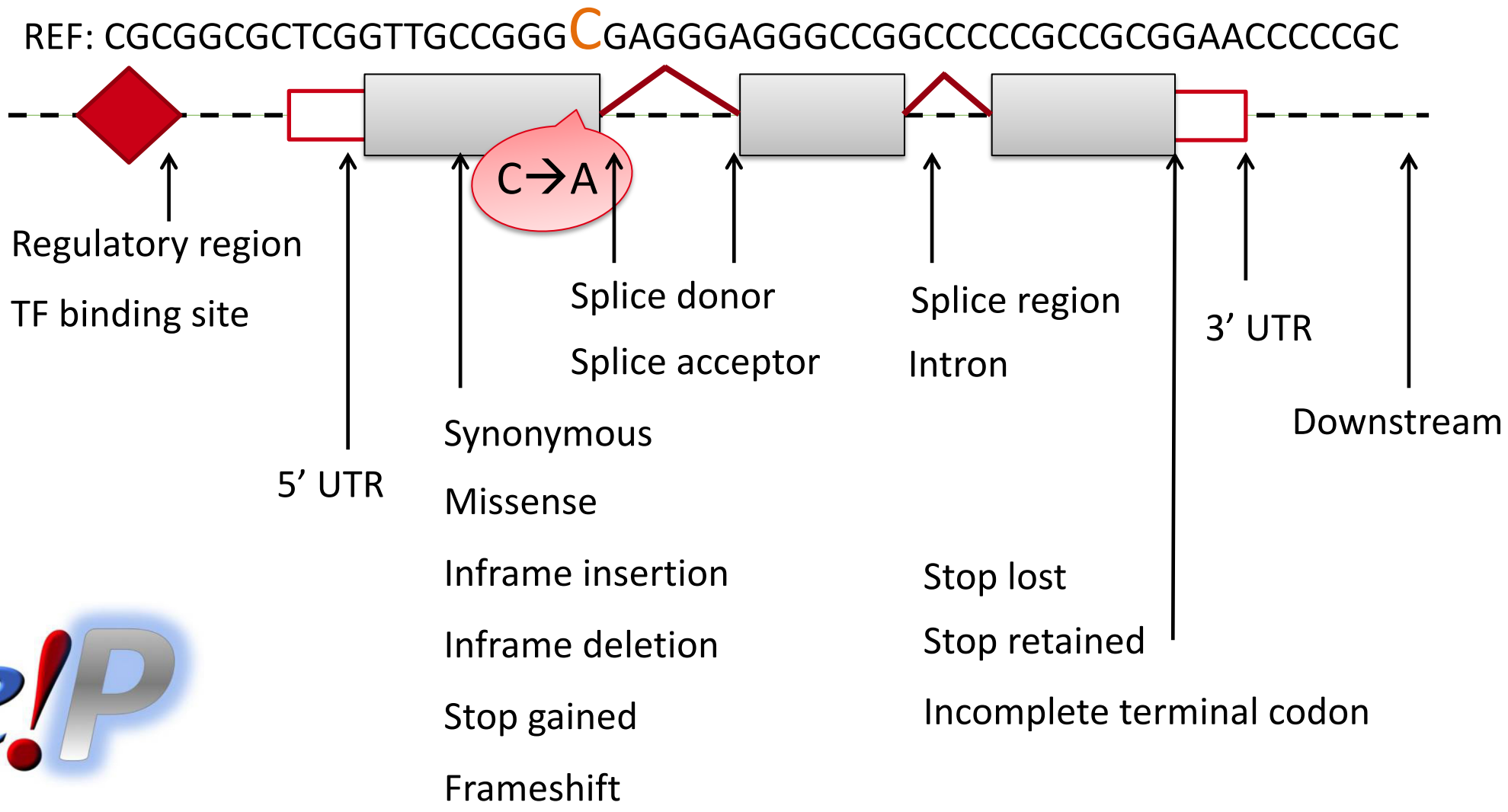
GIG: 0.318 (34)
AIG: 0.449 (48)
AIA: 0.234 (25)



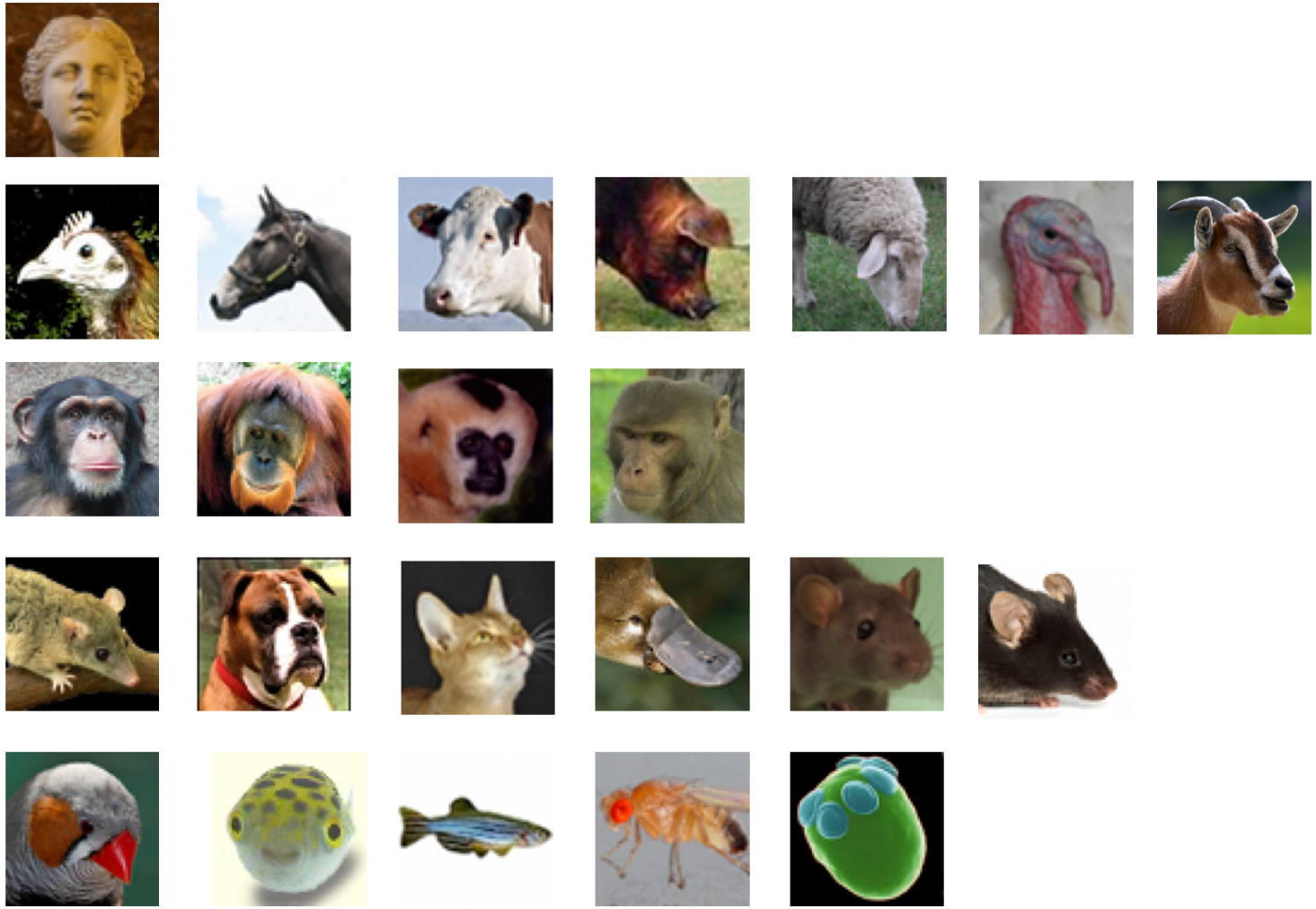
Genome-Wide Association Studies (GWAS)



Variation consequence

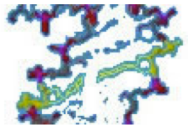


23 species with a variation database



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

dbSNP
Short Genetic Variations



Import variants

DGVarchive

Quality control

Phenotype data

Population genetics

Citations

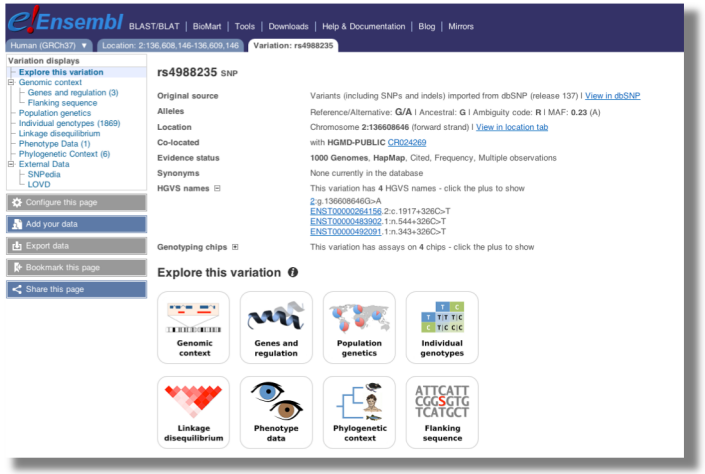
Import richer data

Apply analysis

Genes and regulation

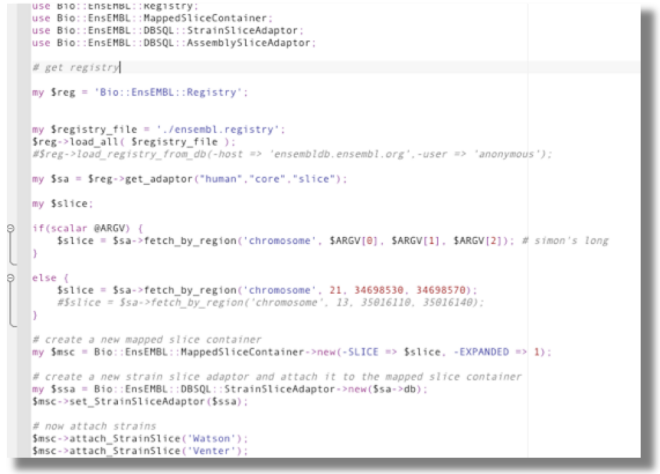
Variation database

MySQL

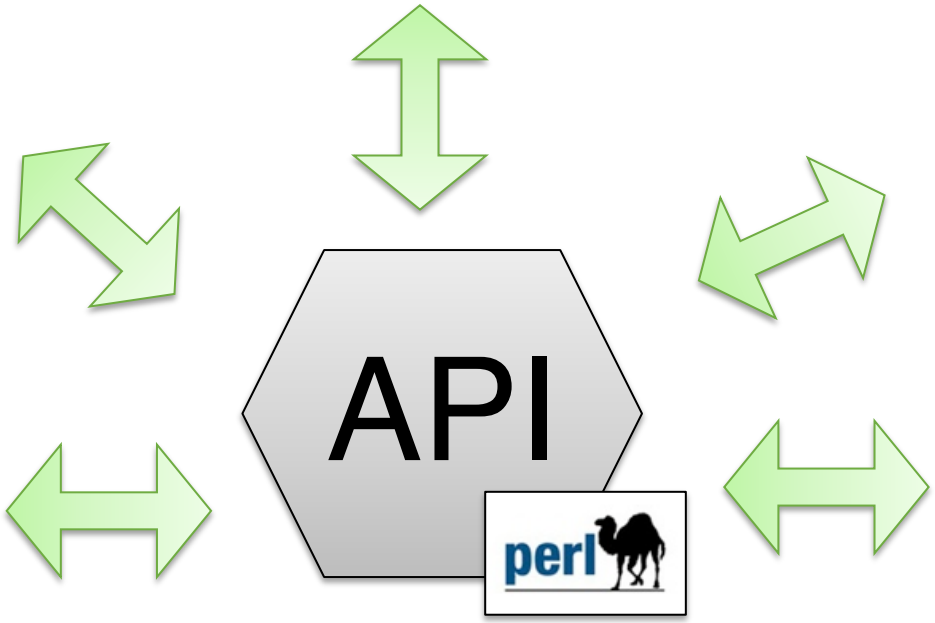


Website

Variation database



Scripts



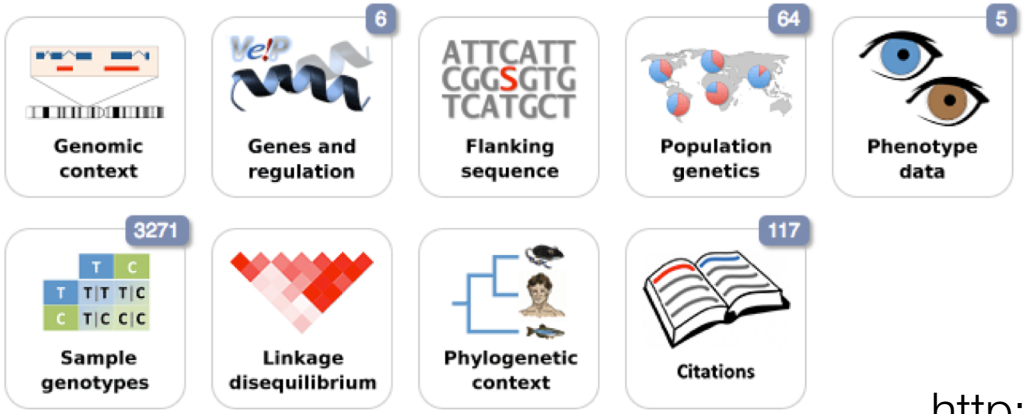
Tools

rs4988235 SNP

Most severe consequence	intron variant I See all predicted consequences
Alleles	G/A/C I Ancestral: G I MAF: 0.16 (A) I Highest population MAF: 0.49
Location	Chromosome 2:135851076 (forward strand) I VCF: `2 135851076 rs4988235 G A,C`
Co-located variant	HGMD-PUBLIC CR024269
Evidence status	
Clinical significance	
HGVS names	This variant has 9 HGVS names - Show
Synonyms	This variant has 2 synonyms - Show
Genotyping chips	This variant has assays on 7 chips - Hide
Original source	Variants (including SNPs and indels) imported from dbSNP (release 150) I View in dbSNP
About this variant	This variant overlaps 3 transcripts, has 3271 sample genotypes, is associated with 5 phenotypes and is mentioned in 117 citations.

- Illumina_ImmunoChip
- Illumina_Human1M-duo
- Illumina_Cardio-Metabo_Chip
- Illumina_HumanOmni5
- HumanOmniExpress
- Illumina_HumanOmni1-Quad
- Illumina_HumanOmni2.5

Explore this variant



http://www.ensembl.org/Homo_sapiens/Variation/Explore?v=rs4988235

Adaptors

- Adaptors are **factories** for generating object
- Adaptors are retrieved from the Registry

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


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


$registry->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');
```

Data retrieval

Adaptors

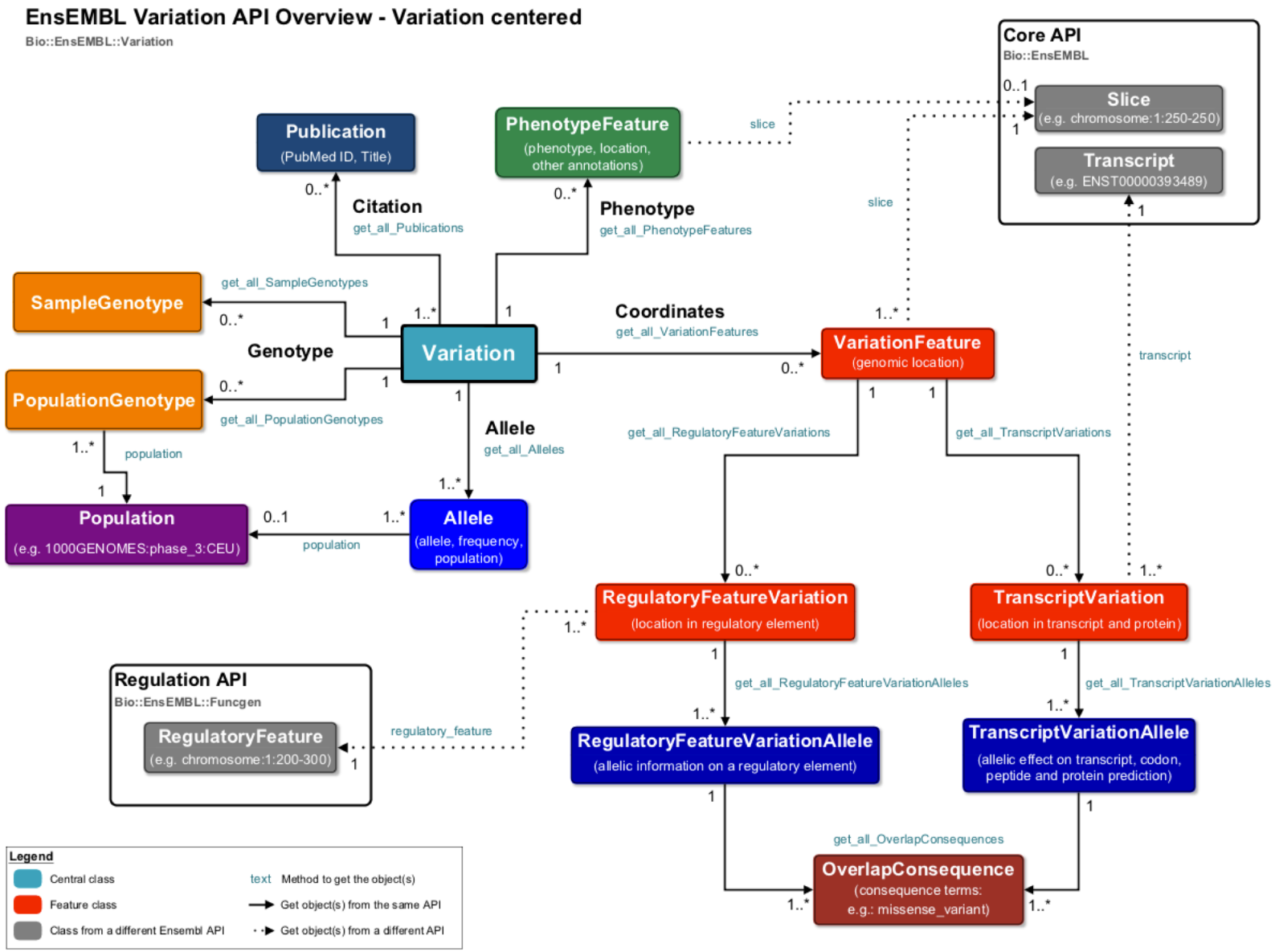
- Fetch object(s) according to some property e.g. name, location
  - `$adaptor->fetch_all_...` returns a list reference of items
  - `$adaptor->fetch_by_...` returns only 1 item
- Check documentation which methods the adaptor provides

API objects: e.g. Variation

- Get properties
  - `$variation->name()`
- Get other object(s) from an API object
  - `$variation->get_all_Alleles()`

Outline for today

- Variation attributes and location
- Variation consequence
- Allele and Genotype Frequencies
- Phenotype
- Linkage Disequilibrium
- Structural Variation



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