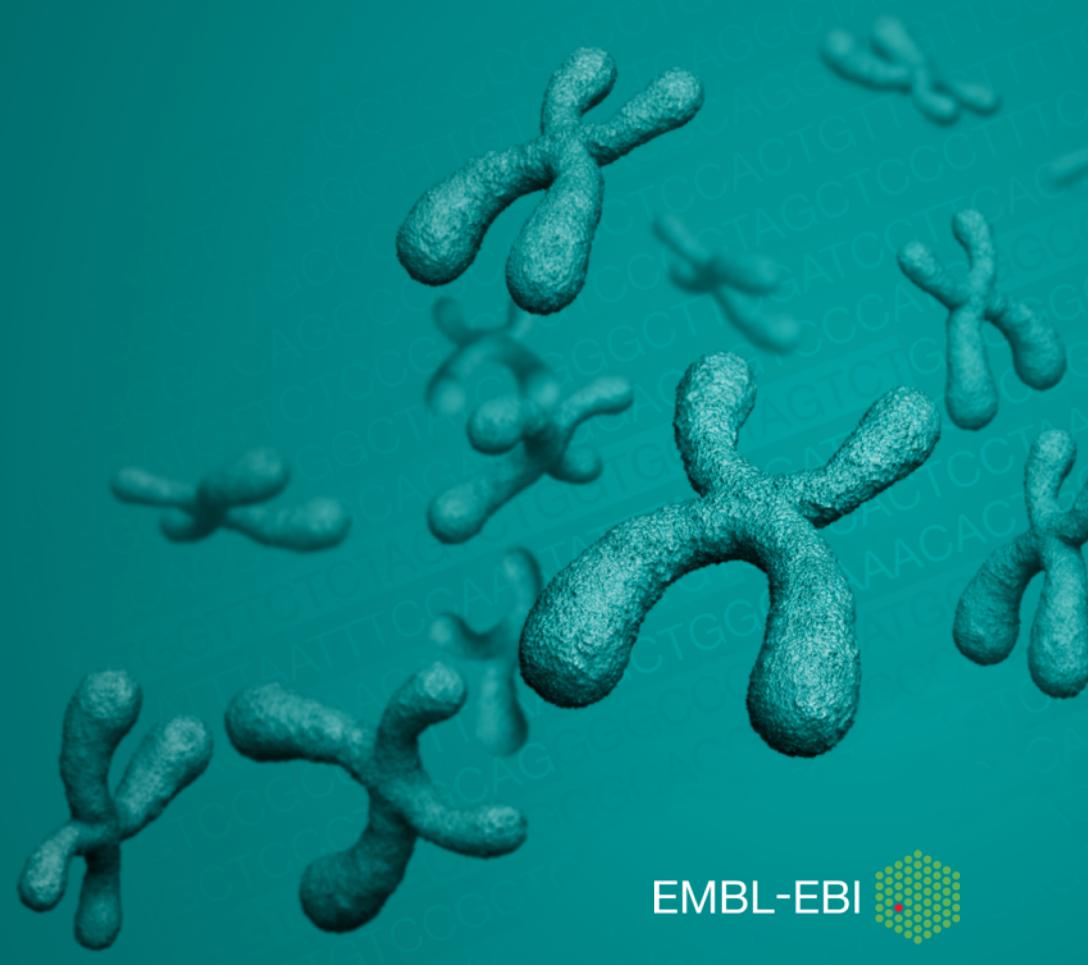


Ensembl Variation API – Part 1

Anja Thormann

Andrew Parton

April 24th 2018 Hinxton



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

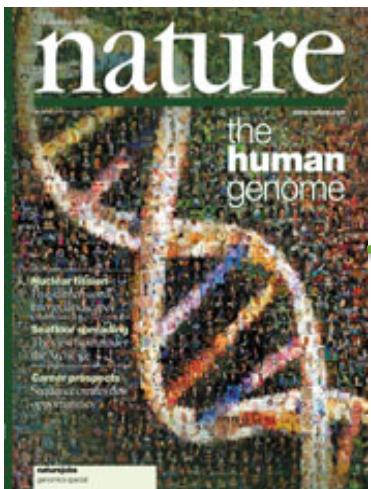


image credit: humanae.tumblr.com

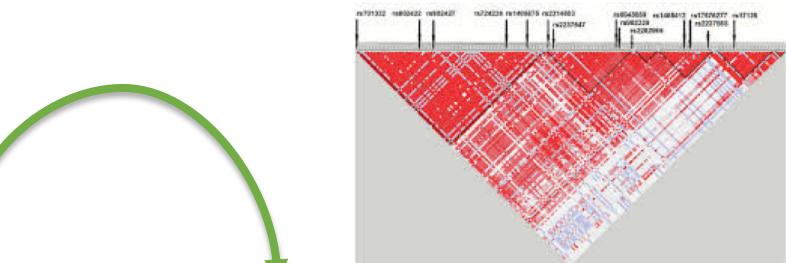
<https://www.yourgenome.org/facts/what-is-genetic-variation>

Sequencing Revolution

2000

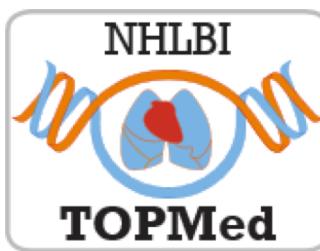


2010



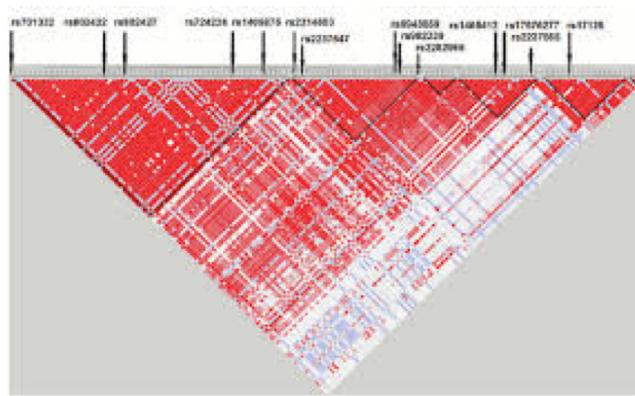
2015

- ExAC
- gnomAD



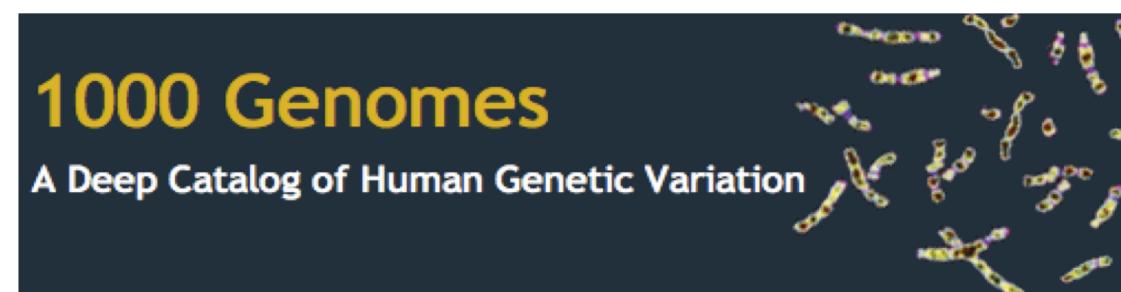
International HapMap consortium

- 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

The diagram illustrates a comparison between two genomes. The top row, labeled "Reference genome", shows the sequence: C G C A G - T G A. The bottom row, labeled "A typical genome", shows the sequence: C G C G G C T G -. Two arrows point from the "C" in the reference genome to the "G" in the typical genome, indicating a SNP. An arrow points from the "T" in the reference genome to the "C" in the typical genome, indicating an insertion. An arrow points from the "A" in the reference genome to the end of the sequence, indicating a deletion.

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



Structural variation

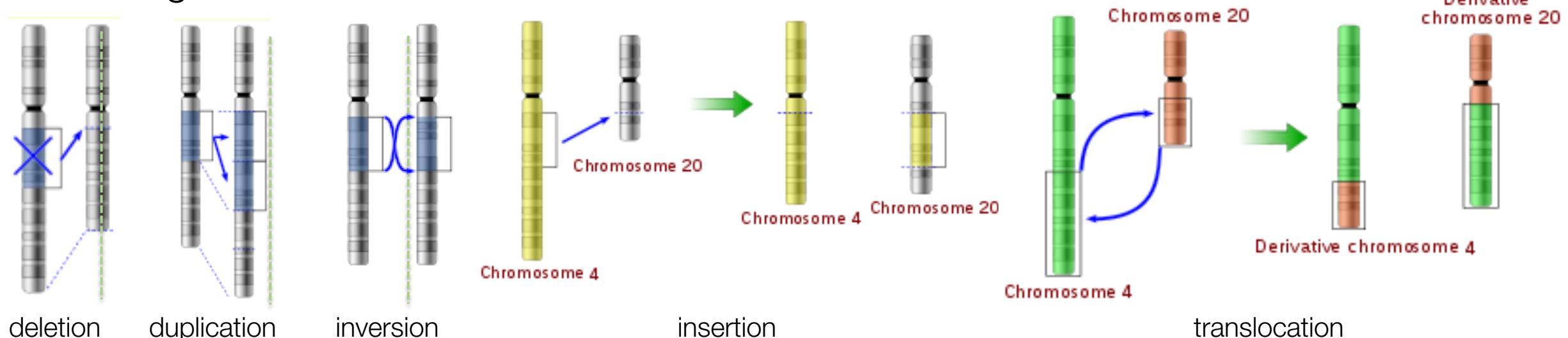
- Copy number variations (CNVs): sequence repeated ‘n’ times in an individual

Reference (4): 

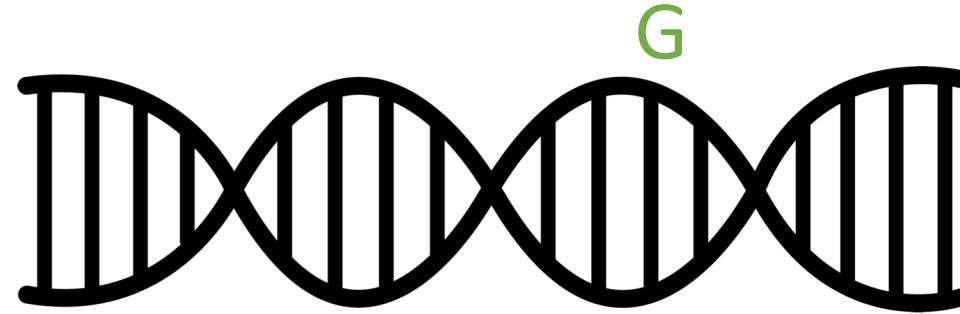
Copy gain (6): 

Copy loss (3): 

- Large structural variants:



Diploid: 2 versions of each autosomal chromosome



Mother

Father

Genotype A | G

AFR



AMR



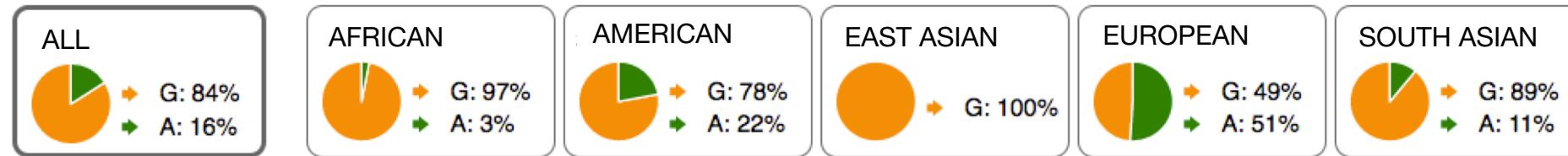
EAS



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

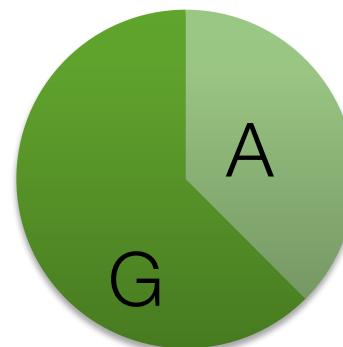
Reference Allele: G



Count alleles: 3 A, 5 G

Frequency for A: $3/8 = 0.375$

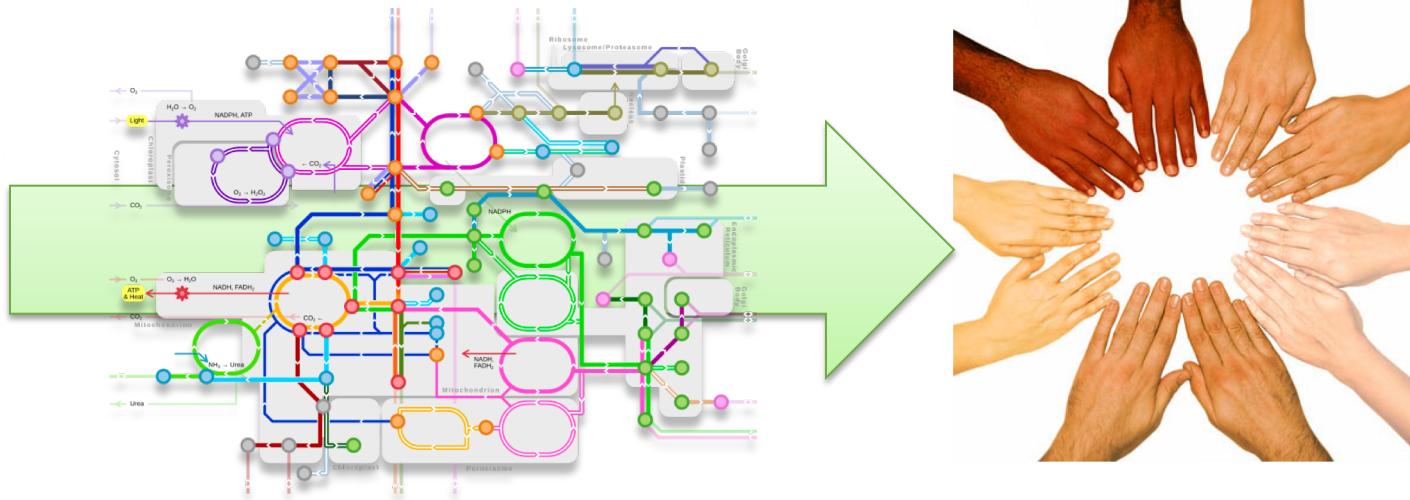
Frequency for G: $5/8 = 0.625$



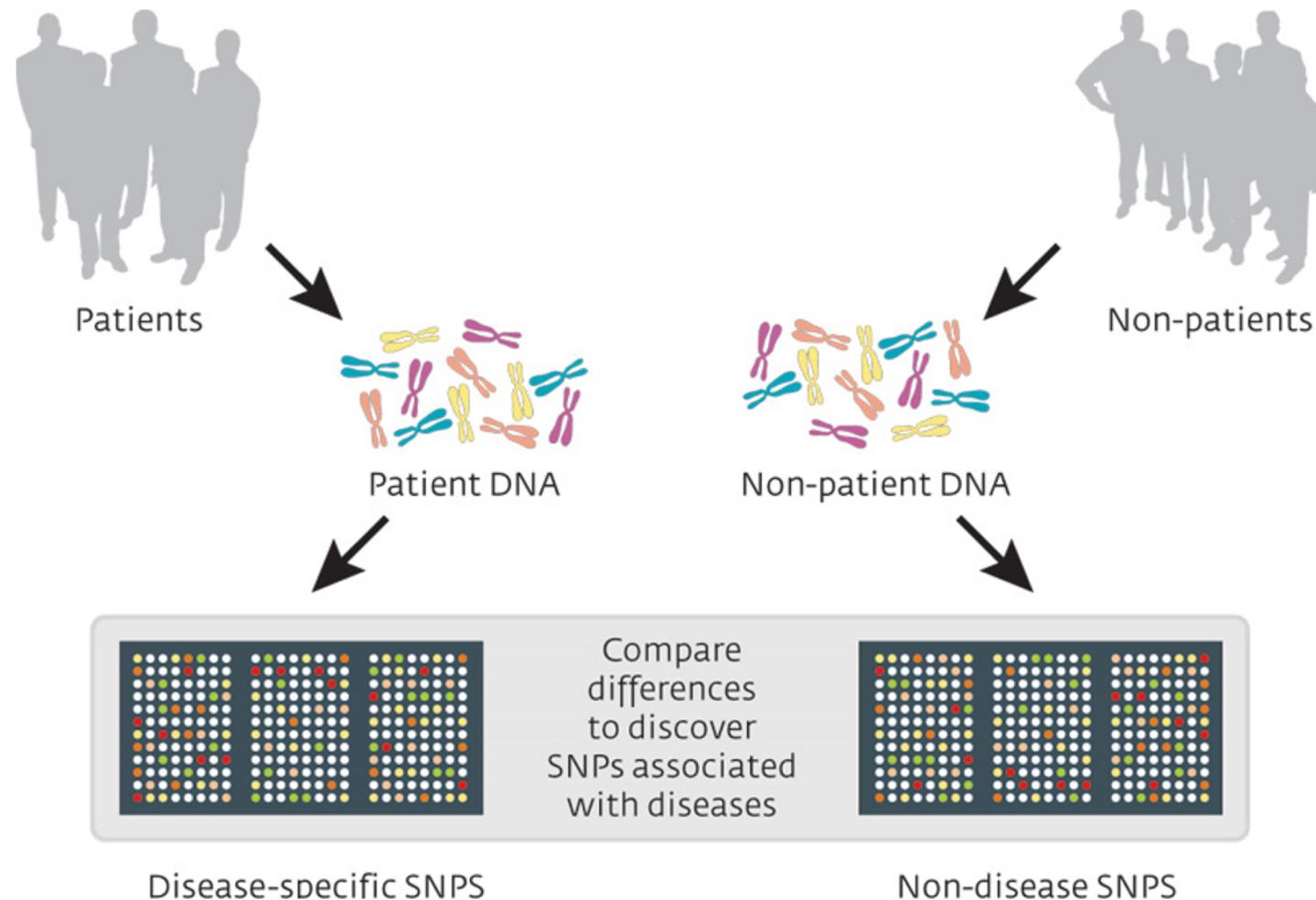
A 0.375
G 0.625

Genotype to Phenotype

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

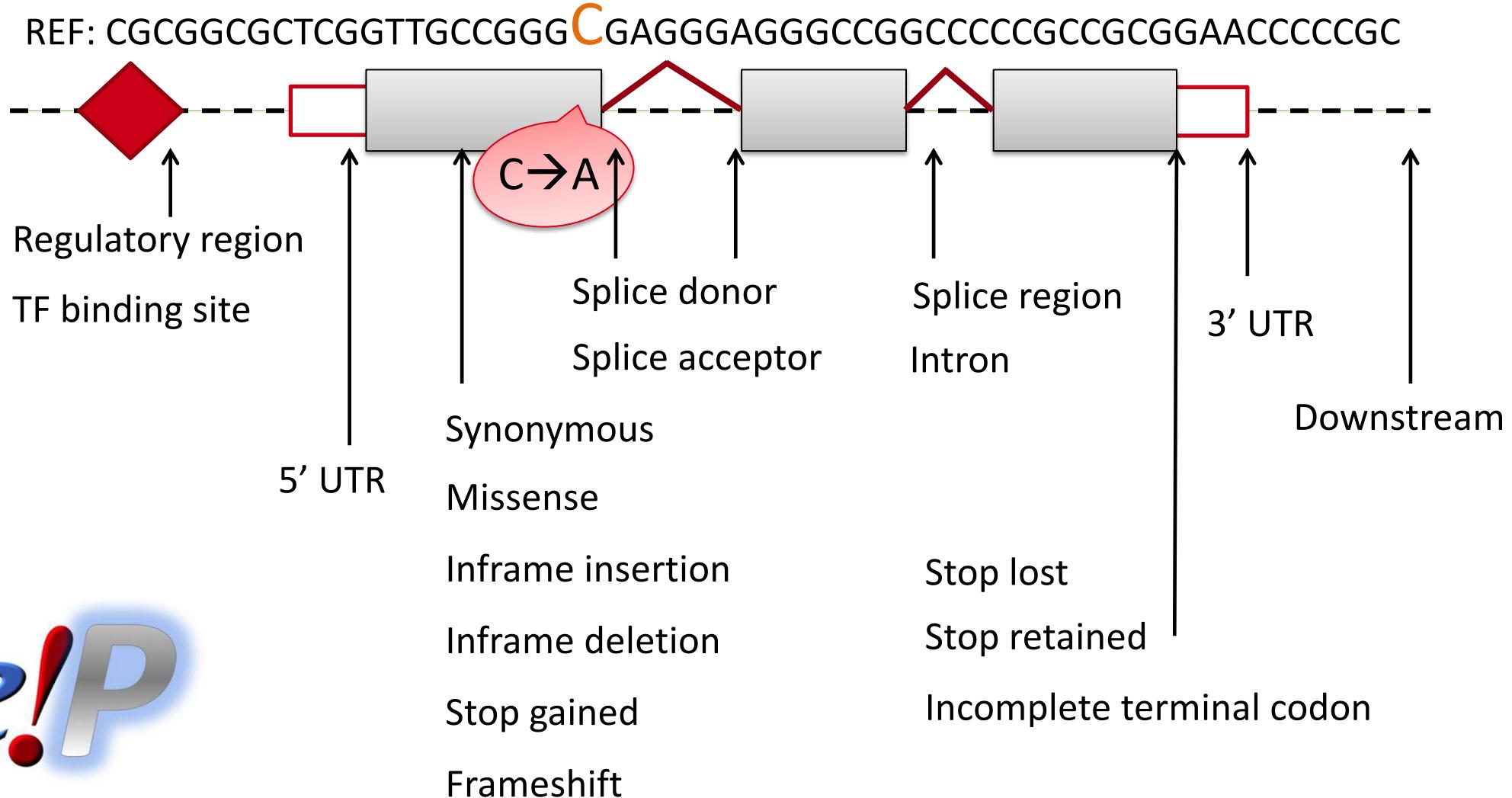


Genome-Wide Association Studies (GWAS)



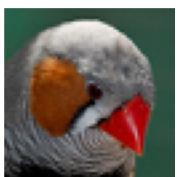
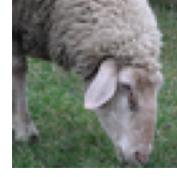
© Pasieka, Science Photo Library

Variation consequence



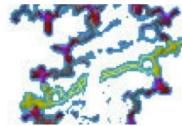
Ve!P

23 species with a variation database



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

dbSNP
Short Genetic Variations



Import variants

DGVA^{rchive}

Phenotype data

Population genetics

Citations

Quality control

Import richer data

Apply analysis

VelP

Genes and regulation

Variation database



Variation details

- Explore this variation
- Genomic context
- Flanking sequence
- Population genetics
- Individual genotypes (1869)
- Linkage disequilibrium
- Phylogenetic Data (1)
- External Data
- SNPedia
- LOVD

Original source Variants (including SNPs and indels) imported from dbSNP (release 137) | [View in dbSNP](#)

Reference/Alternative: G/A | Ancestral: G | Ambiguity code: R | MAF: 0.23 (A)

Alleles Chromosome 2:136608646 (forward strand) | [View in location tab](#)

Location with HGMD-PUBLIC [CR024269](#)

Co-located 1000 Genomes, HapMap, Cited, Frequency, Multiple observations

Evidence status None currently in the database

Synonyms HGVS names

This variation has 4 HGVS names - click the plus to show

ENST00000294156.2:c.1917<326>C>T
 ENST00000485902.1:n.544<326>C>T
 ENST00000450991.1:n.343<326>C>T

Genotyping chips This variation has assays on 4 chips - click the plus to show

Explore this variation

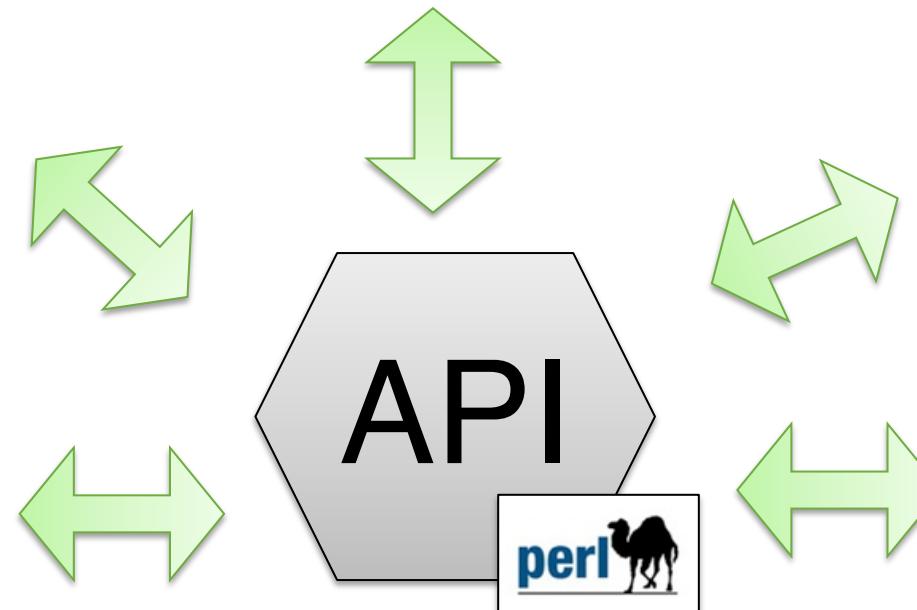
- Genomic context
- Genes and regulation
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phenotype data
- Phylogenetic context
- Flanking sequence

Add your data Export data Bookmark this page Share this page

Variation database



Website



Scripts

```

use Bio::ENSEMBL::Registry;
use Bio::EnsEMBL::MappedSliceContainer;
use Bio::EnsEMBL::DBSQL::StrainSliceAdaptor;
use Bio::EnsEMBL::DBSQL::AssemblySliceAdaptor;

# get registry
my $reg = 'Bio::EnsEMBL::Registry';

my $registry_file = '/ensembl.registry';
$reg->load_all($registry_file);
#$reg->load_registry_from_db(-host => 'ensembl.ensembl.org', -user => 'anonymous');

my $sa = $reg->get_adaptor("human", "core", "slice");

my $slice;

if(scalar @ARGV) {
    $slice = $sa->fetch_by_region('chromosome', $ARGV[0], $ARGV[1], $ARGV[2]); # simon's long
}

else {
    $slice = $sa->fetch_by_region('chromosome', 21, 34698530, 34698570);
    #$slice = $sa->fetch_by_region('chromosome', 13, 35016110, 35016140);
}

# create a new mapped slice container
my $msc = Bio::ENSEMBL::MappedSliceContainer->new(-SLICE => $slice, -EXPANDED => 1);

# create a new strain slice adaptor and attach it to the mapped slice container
my $ssa = Bio::EnsEMBL::DBSQL::StrainSliceAdaptor->new($sa->db);
$msc->set_StrainSliceAdaptor($ssa);

# now attach strains
$msc->attach_StrainSlice('Watson');
$msc->attach_StrainSlice('Venter');
  
```

Tools



rs4988235 SNP

Most severe consequence

intron variant | See all predicted consequences

Alleles

G/A/C | Ancestral: G | MAF: 0.16 (A) | Highest population MAF: 0.49

Location

Chromosome 2:135851076 (forward strand) | 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](#) ⏺

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

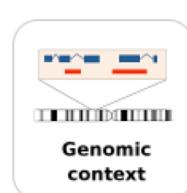
Original source

Variants (including SNPs and indels) imported from dbSNP (release 150) | [View in dbSNP](#) ⓘ

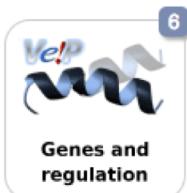
About this variant

This variant overlaps 3 transcripts, has 3271 sample genotypes, is associated with 5 phenotypes and is mentioned in 117 citations.

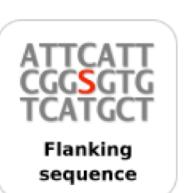
Explore this variant ⓘ



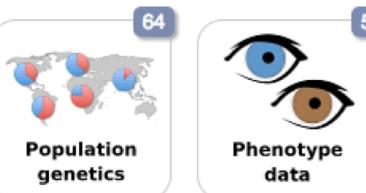
Genomic context



Genes and regulation



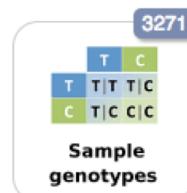
Flanking sequence



Population genetics



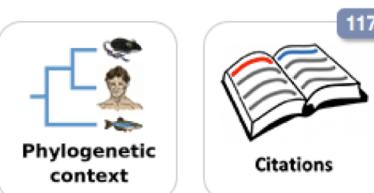
Phenotype data



Sample genotypes



Linkage disequilibrium



Phylogenetic context



Citations

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');
```



The diagram consists of three green rectangular boxes with white text. The first box contains the word 'species' with a green arrow pointing upwards from it to the word 'species' in the code. The second box contains the word 'group' with a green arrow pointing upwards from it to the word 'group' in the code. The third box contains the words 'object name' with a green arrow pointing upwards from it to the word 'variation' in the code.

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

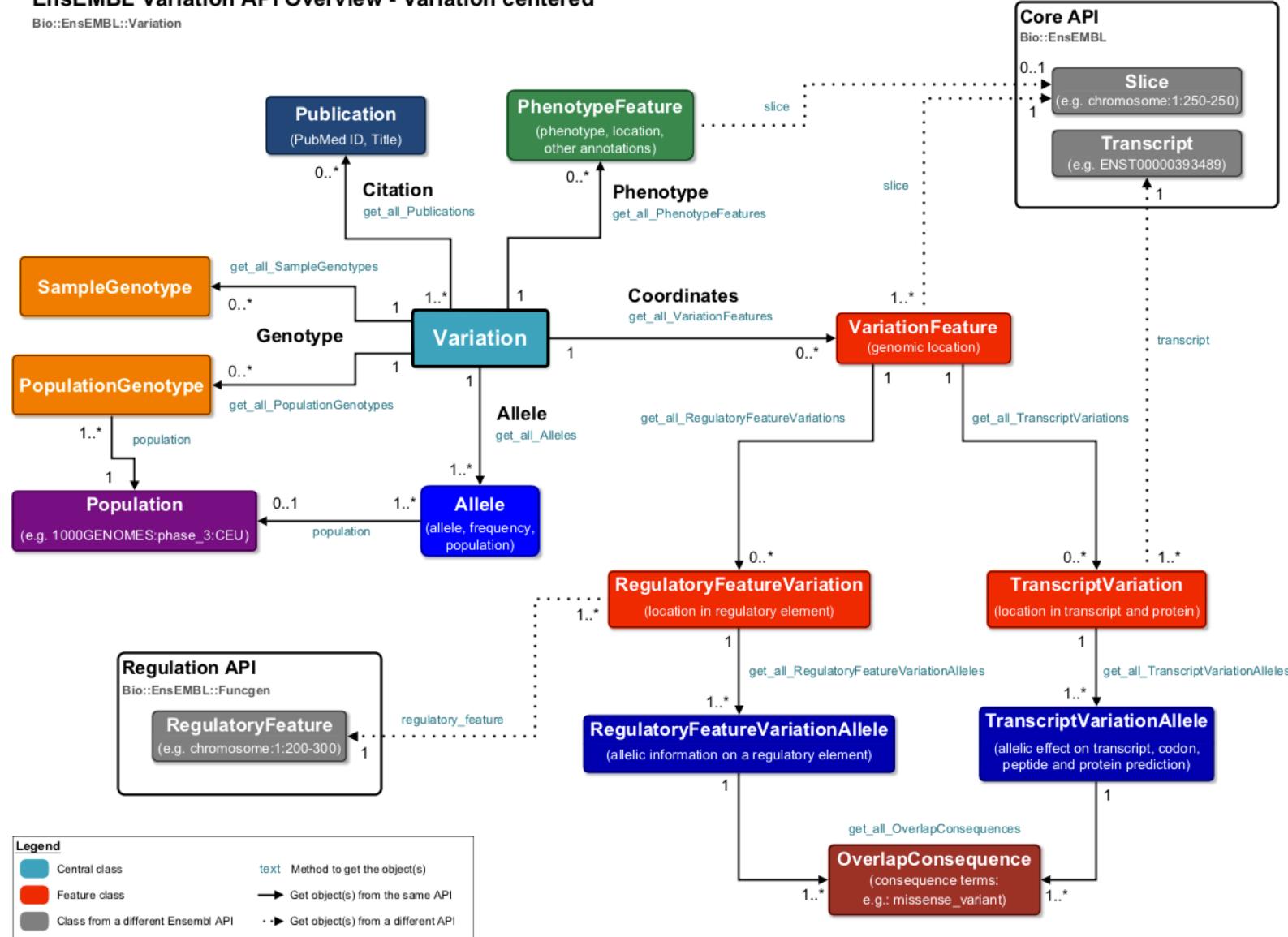
- Get properties
 - `$variation->name()`
- Get other object(s) from an API object
 - `$variation->get_all_Alleles()`

Outline for today

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

EnsEMBL Variation API Overview - Variation centered

Bio::EnsEMBL::Variation



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