Training materials

- Ensembl training materials are protected by a CC BY license
- http://creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers
- http://www.ensembl.org/info/about/publications.html

http://training.ensembl.org/events

Viewing the data: the Ensembl browser and its possibilities

Benjamin Moore
Ensembl Outreach Officer

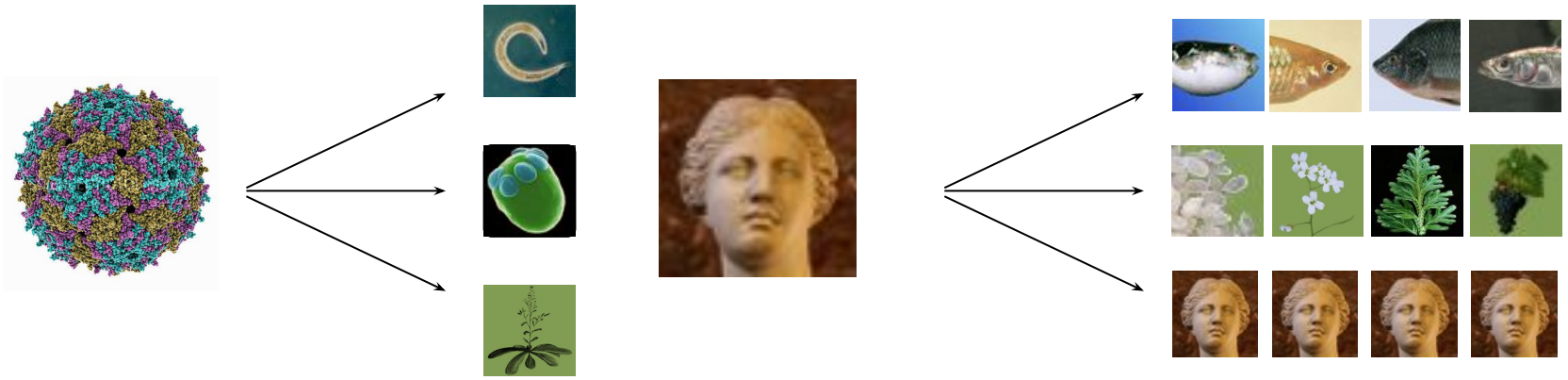
Slides available from
training.ensembl.org/events/

EMBL-EBI

Why do we need genome browsers?

1977: 1st genome to be sequenced (5 kb)

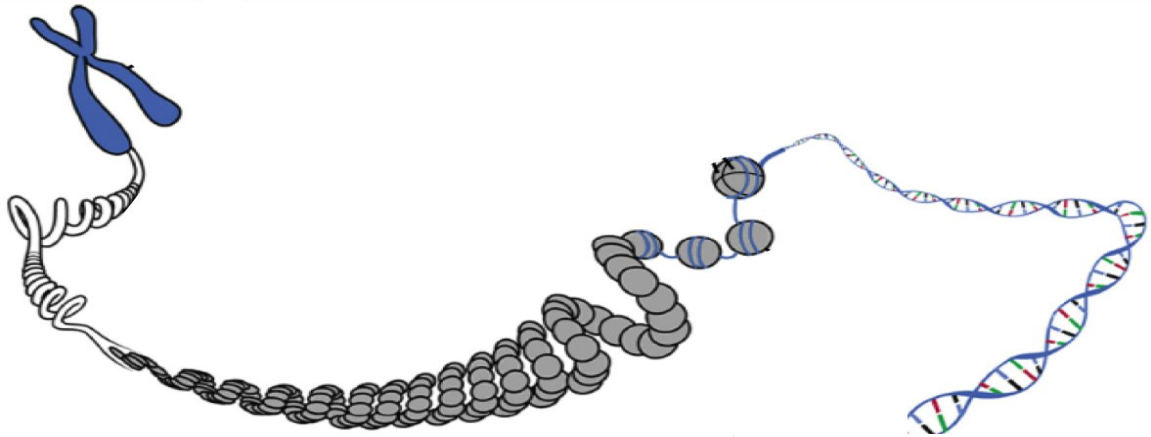
2004: finished human sequence (3 Gb)



Why do we need genome browsers?

http://training.ensembl.org/events

Ensembl- unlocking the code

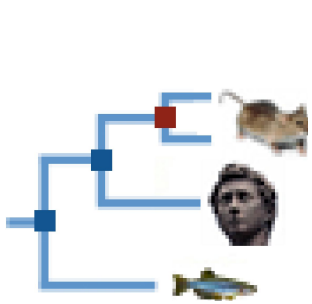


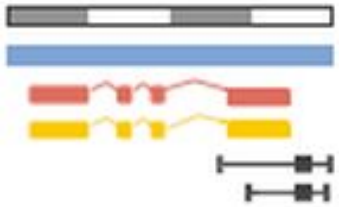
- *Genomic assemblies* - automated gene annotation
- *Variation* - Small and large scale sequence variation with phenotype associations
- *Comparative Genomics* - Whole genome alignments, gene trees
- *Regulation* - Potential promoters and enhancers, DNA methylation

http://training.ensembl.org/events

Ensembl Features

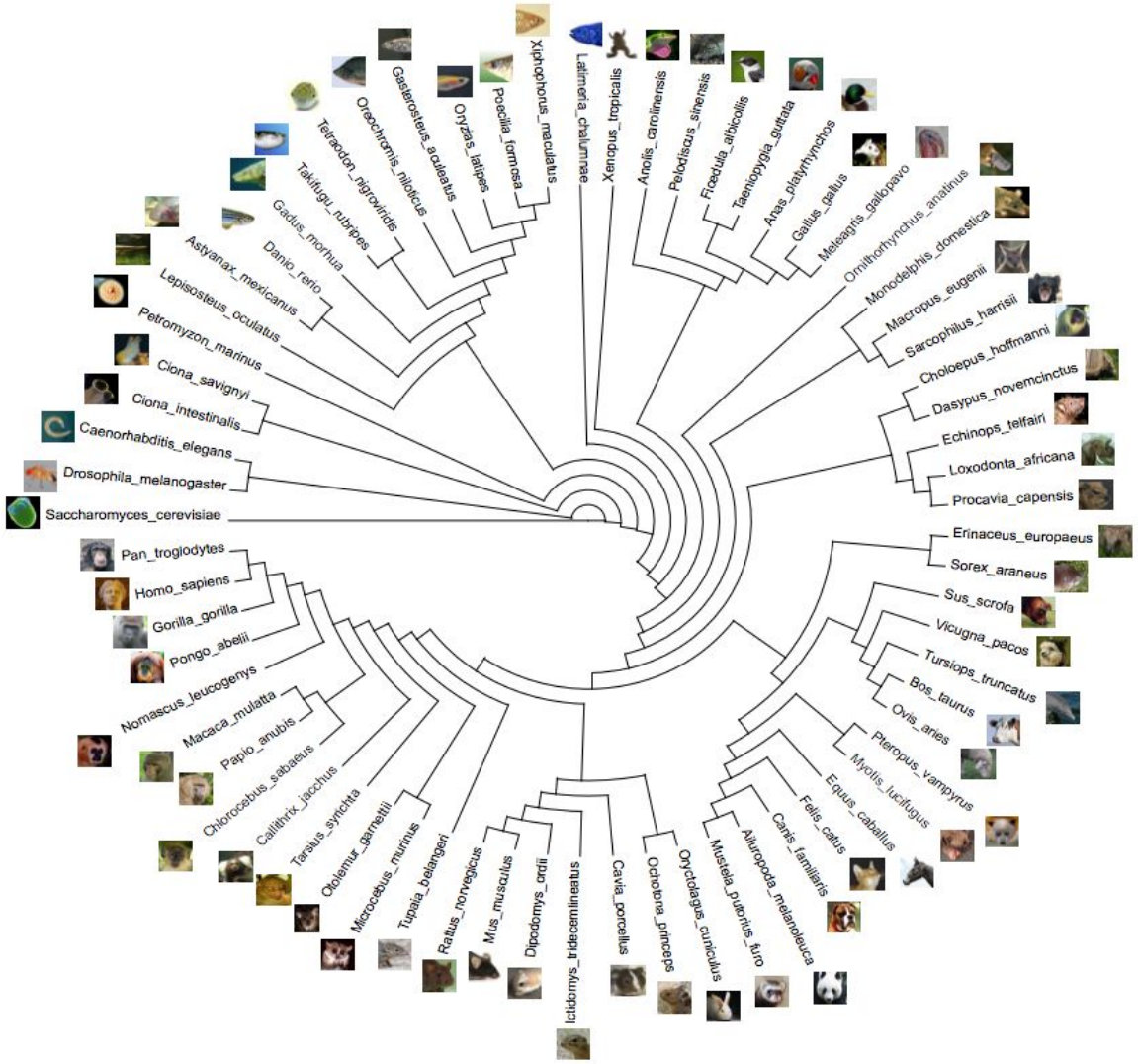
- Gene builds for ~100 species
- Gene trees
- Regulatory build
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source





http://training.ensembl.org/events

Ensembl- access to 100+ genomes



http://training.ensembl.org/events

Ensembl Genomes- expanding Ensembl

www.ensembl.org

- Vertebrates



- Other representative species

http://training.ensembl.org/events

Ensembl Genomes- expanding Ensembl

www.ensembl.org

www.ensemblgenomes.org

- Vertebrates

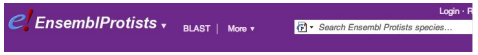


- Other representative species

- Bacteria

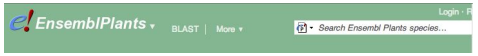

- Fungi


- Protists

- Metazoa


- Plants

http://training.ensembl.org/events

Variation types

1) Small scale in one or few nucleotides of a gene

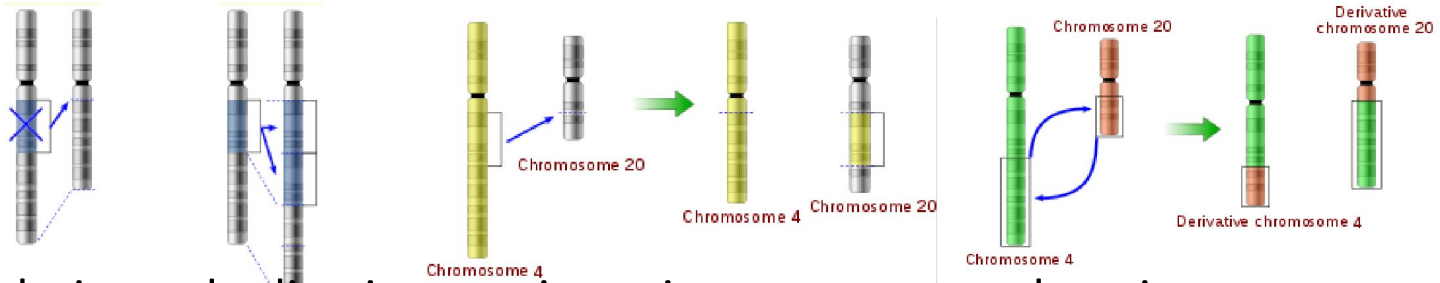
- Small insertions and deletions (DIPs or indels)
- Single nucleotide polymorphism (SNP)

A G A C T T G A C C T G T C T - A A C T G G A

T G A C T T G A C - T G T C T G A A C G G G A

2) Large scale in chromosomal structure (structural variation)

- Copy number variations (CNV)
- Large deletions/duplications, insertions, translocations



deletion duplication insertion translocation

22 species with variation data

	**Cat** *Felis catus*	3.6 million+	
	**Chicken** *Gallus gallus*	21 million+	
	**Chimpanzee** *Pan troglodytes*	1.6 million+	
	**Cow** *Bos taurus*	100 million+	
	**Dog** *Canis familiaris*	5.9 million+	
	**Fruitfly** *Drosophila melanogaster*	6.7 million+	
	**Gibbon** *Nomascus leucogenys*	1.1 million+	
	**Horse** *Equus caballus*	5 million+	
	**Human** *Homo sapiens*	329 million+	
	**Macaque** *Macaca mulatta*	3 million+	
	**Mouse** *Mus musculus*	80 million+	
	**Opossum** *Monodelphis domestica*	1.1 million+	
	**Orangutan** *Pongo abelii*	10 million+	
	**Pig** *Sus scrofa*	60 million+	
	**Platypus** *Ornithorhynchus anatinus*	1.3 million+	
	**Rat** *Rattus norvegicus*	5 million+	
	**S. cerevisiae** *Saccharomyces cerevisiae*	263,000+	
	**Sheep** *Ovis aries*	51 million+	
	**Tetraodon** *Tetraodon nigroviridis*	902,000+	
	**Turkey** *Meleagris gallopavo*	9,000+	
	**Zebra Finch** *Taeniopygia guttata*	1.7 million+	
	**Zebrafish** *Danio rerio*	17 million+	

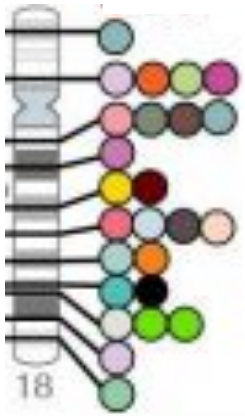
http://training.ensembl.org/events

22 species with variation data

	**Cat** *Felis catus*	3.6 million+		**Opossum** *Monodelphis domestica*	1.1 million+
	**Chicken** *Gallus gallus*	21 million+		**Orangutan** *Pongo abelii*	10 million+
	**Chimpanzee** *Pan troglodytes*	1.6 million+		**Pig** *Sus scrofa*	60 million+
	**Cow** *Bos taurus*	100 million+		**Platypus** *Ornithorhynchus anatinus*	1.3 million+
	**Dog** *Canis familiaris*	5.9 million+		**Rat** *Rattus norvegicus*	5 million+
	**Fruitfly** *Drosophila melanogaster*	6.7 million+		**S. cerevisiae** *Saccharomyces cerevisiae*	263,000+
	**Gibbon** *Nomascus leucogenys*	1.1 million+		**Sheep** *Ovis aries*	51 million+
	**Horse** *Equus caballus*	5 million+		**Tetraodon** *Tetraodon nigroviridis*	902,000+
	**Human** *Homo sapiens*	650 million+ ~~329 million+~~		**Turkey** *Meleagris gallopavo*	9,000+
	**Macaque** *Macaca mulatta*	3 million+		**Zebra Finch** *Taeniopygia guttata*	1.7 million+
	**Mouse** *Mus musculus*	80 million+		**Zebrafish** *Danio rerio*	17 million+

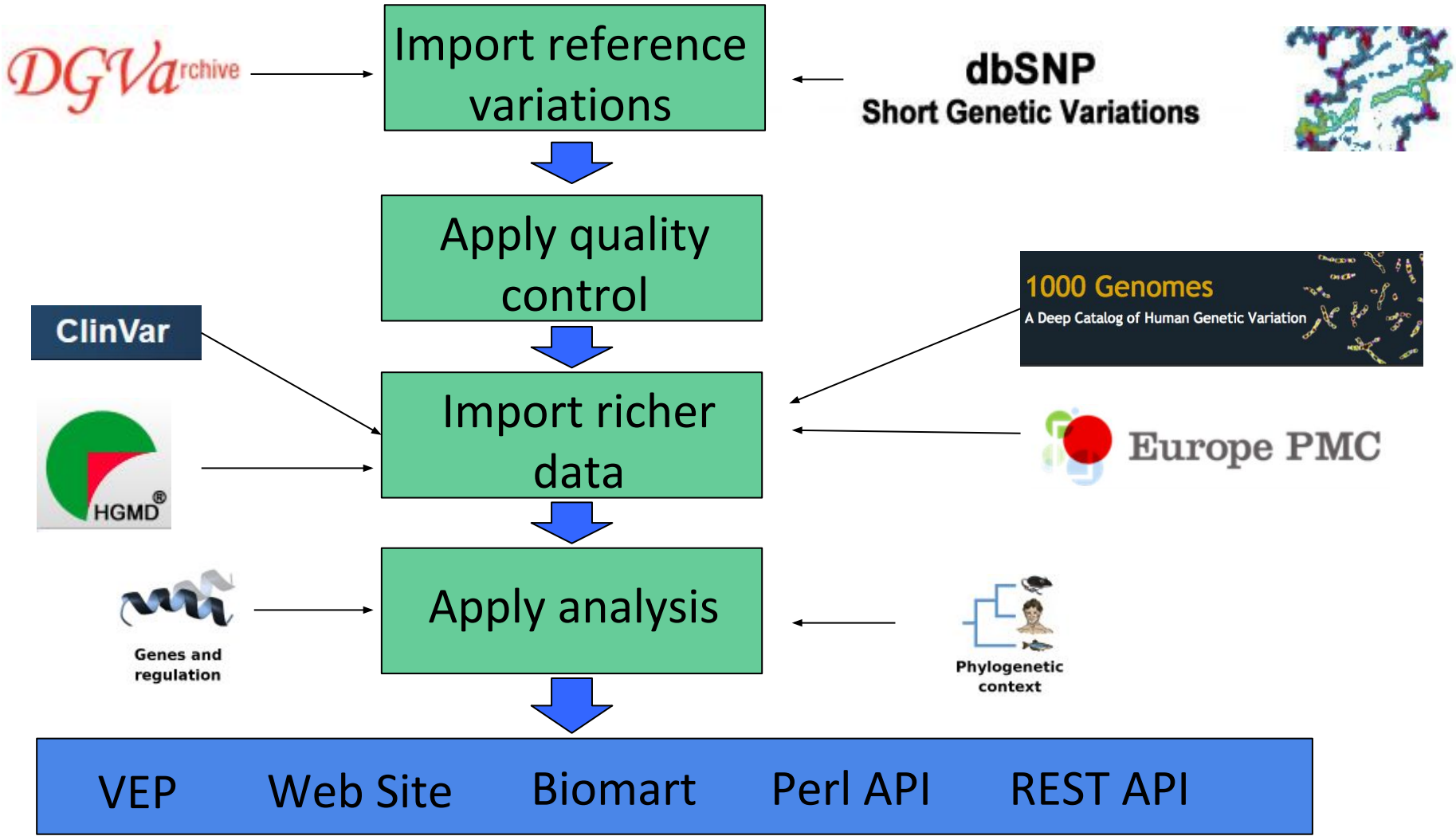
http://training.ensembl.org/events

Variation sources



http://www.ensembl.org/info/docs/variation/sources_documentation.html

The Ensembl variation process



http://training.ensembl.org/events

Quality Control

Icon	Name	Description
	Multiple observations	The variant has multiple independent dbSNP submissions, i.e. submissions with a different su different discovery samples
	Frequency	The variant is reported to be polymorphic in at least one sample
	Cited	The variant is cited in a PubMed article.
	Phenotype or Disease	The variant is associated with at least one phenotype or disease.
	1000 Genomes	The variant was discovered in the 1000 Genomes Project (human only)
	ExAC	The variant was discovered in the Exome Aggregation Consortium (human only).
	HapMap	The variant is polymorphic in at least one HapMap panel (human only)
	ESP	The variant was discovered in the Exome Sequencing Project (human only).

We summarise the information supporting each dbSNP variant to give a guide to its reliability

We run basic checks and flag suspicious variants

rs14390 SNP

This variation has been flagged

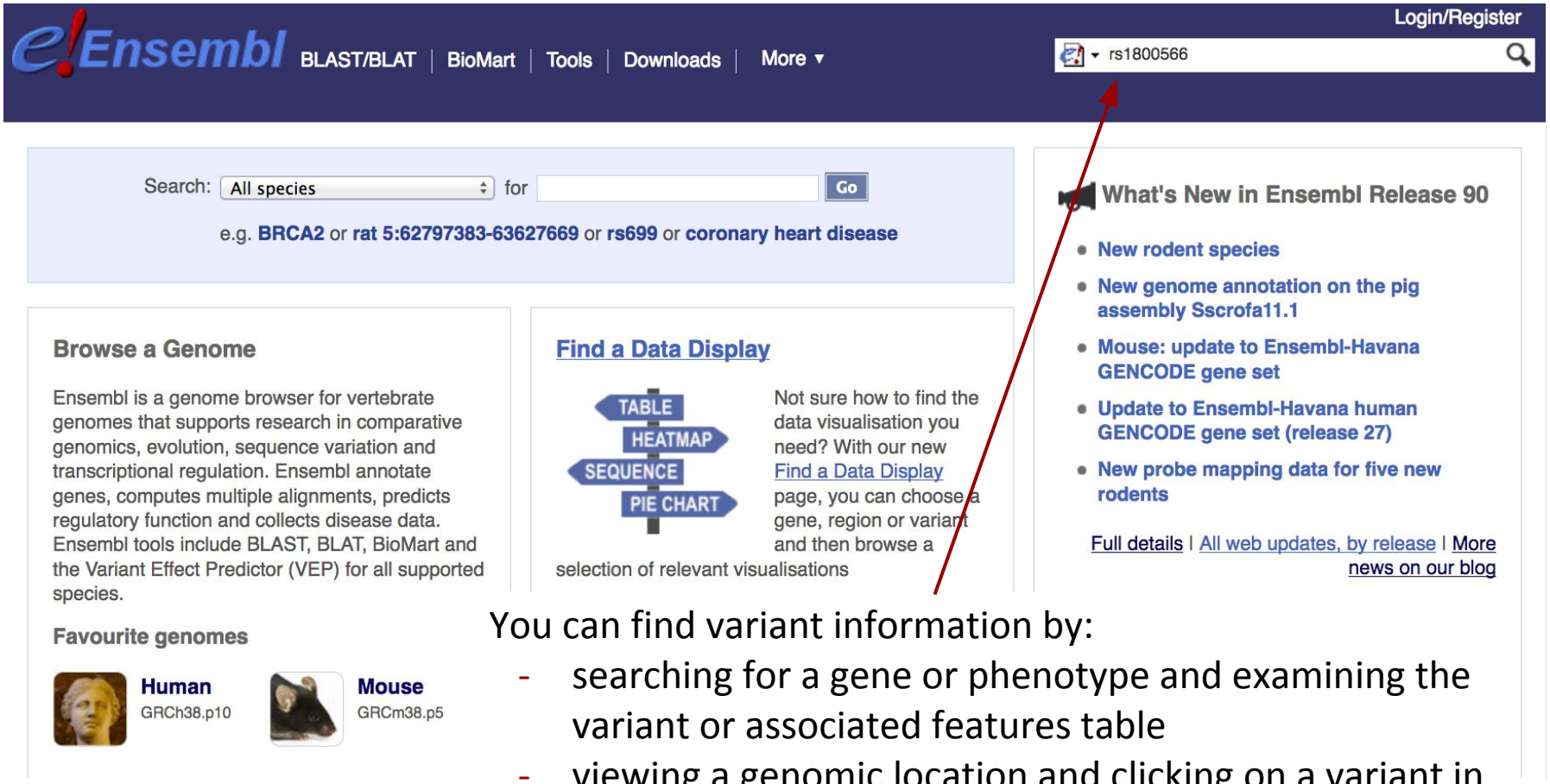
- Flagged as suspect by dbSNP
- Variation maps to 2 genomic locations

Select a location: None selected Go

Original source	Variants (including SNPs and indels) imported from dbSNP (release 138)
Alleles	C/T | Ancestral: T | Ambiguity code: Y
Location	This variation maps to 2 genomic locations; **None selected**
Evidence status	

QC Type	Reported failure reason
Mapping checks	Variant does not map to the genome
	Variant maps to more than 1 location
	Mapped position is not compatible with reported alleles
	None of the variant alleles match the reference allele
Checks on the alleles of refSNPs	Loci with no observed variant alleles in dbSNP
	Alleles contain ambiguity codes
	Alleles contain non-nucleotide characters
Checks on the alleles in dbSNP submissions	Additional submitted allele data from dbSNP does not agree with the dbSNP refSNP alleles
External failure classification	Flagged as suspect by dbSNP

Finding your variant

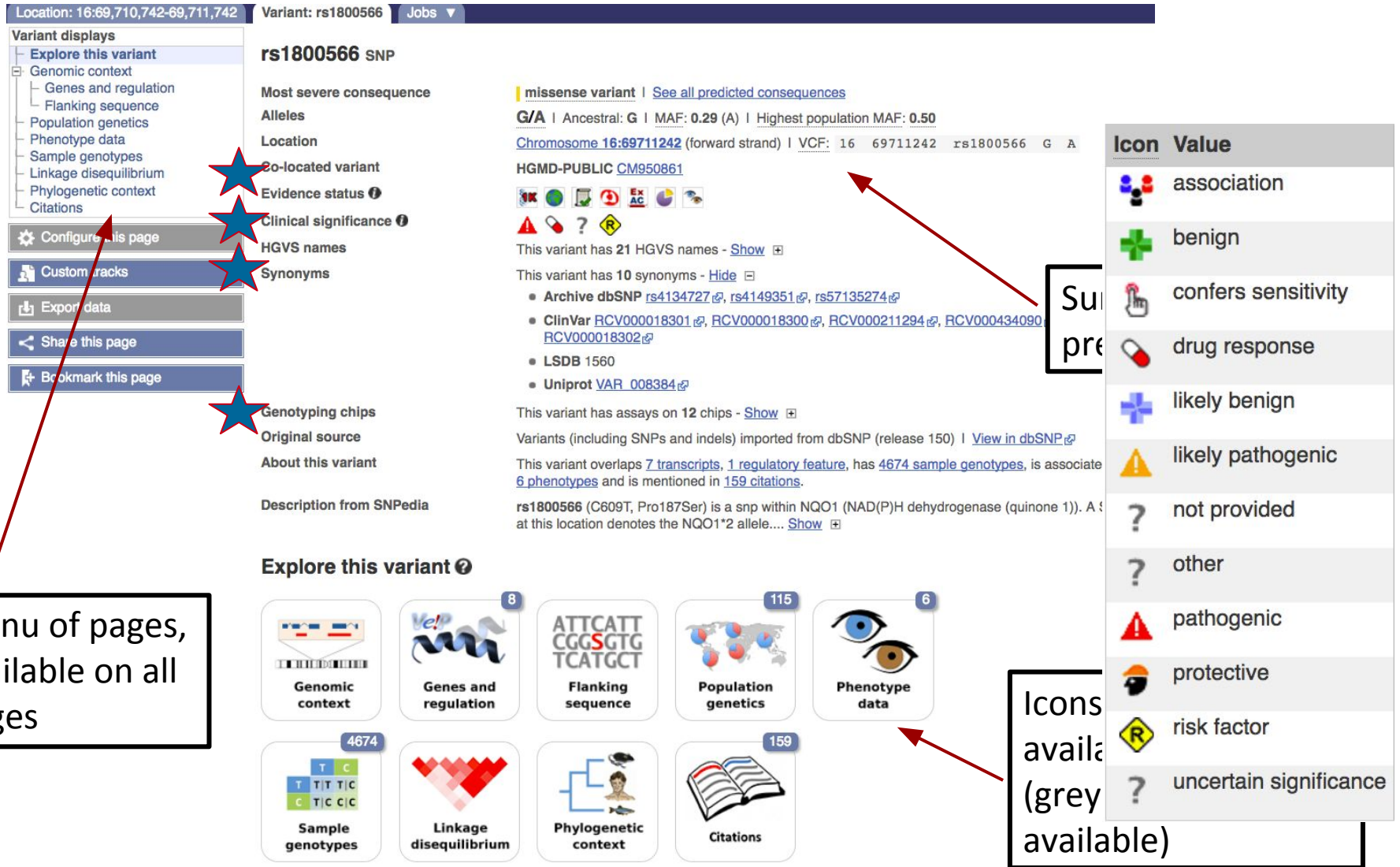


You can find variant information by:

- searching for a gene or phenotype and examining the variant or associated features table
- viewing a genomic location and clicking on a variant in one of the variant tracks
- searching directly by variant name

http://training.ensembl.org/events

The Variant Tab



Menu of pages, available on all pages

Icons available (grey available)

http://training.ensembl.org/events

Allele Frequencies- the 1000 Genomes Project

Name	Size	Description
1000GENOMES:phase_3:ALL	**2504**	**All phase 3 individuals**
1000GENOMES:phase_3:AFR	**661**	**African**
• 1000GENOMES:phase_3:**ACB**	96	African Caribbean in Barbados
• 1000GENOMES:phase_3:**ASW**	61	African Ancestry in Southwest US
• 1000GENOMES:phase_3:**ESN**	99	Esan in Nigeria
• 1000GENOMES:phase_3:**GWD**	113	Gambian in Western Division, The Gambia
• 1000GENOMES:phase_3:**LWK**	99	Luhya in Webuye, Kenya
• 1000GENOMES:phase_3:**MSL**	85	Mende in Sierra Leone
• 1000GENOMES:phase_3:**YRI**	108	Yoruba in Ibadan, Nigeria
1000GENOMES:phase_3:AMR	**347**	**American**
• 1000GENOMES:phase_3:**CLM**	94	Colombian in Medellin, Colombia
• 1000GENOMES:phase_3:**MXL**	64	Mexican Ancestry in Los Angeles, California
• 1000GENOMES:phase_3:**PEL**	85	Peruvian in Lima, Peru
• 1000GENOMES:phase_3:**PUR**	104	Puerto Rican in Puerto Rico
1000GENOMES:phase_3:EAS	**504**	**East Asian**
• 1000GENOMES:phase_3:**CDX**	93	Chinese Dai in Xishuangbanna, China
• 1000GENOMES:phase_3:**CHB**	103	Han Chinese in Bejing, China
• 1000GENOMES:phase_3:**CHS**	105	Southern Han Chinese, China
• 1000GENOMES:phase_3:**JPT**	104	Japanese in Tokyo, Japan
• 1000GENOMES:phase_3:**KHV**	99	Kinh in Ho Chi Minh City, Vietnam
1000GENOMES:phase_3:EUR	**503**	**European**
• 1000GENOMES:phase_3:**CEU**	99	Utah residents with Northern and Western European ancestry
• 1000GENOMES:phase_3:**FIN**	99	Finnish in Finland
• 1000GENOMES:phase_3:**GBR**	91	British in England and Scotland
• 1000GENOMES:phase_3:**IBS**	107	Iberian populations in Spain
• 1000GENOMES:phase_3:**TSI**	107	Toscani in Italy
1000GENOMES:phase_3:SAS	**489**	**South Asian**
• 1000GENOMES:phase_3:**BEB**	86	Bengali in Bangladesh
• 1000GENOMES:phase_3:**GIH**	103	Gujarati Indian in Houston, TX
• 1000GENOMES:phase_3:**ITU**	102	Indian Telugu in the UK
• 1000GENOMES:phase_3:**PJL**	96	Punjabi in Lahore, Pakistan
• 1000GENOMES:phase_3:**STU**	102	Sri Lankan Tamil in the UK

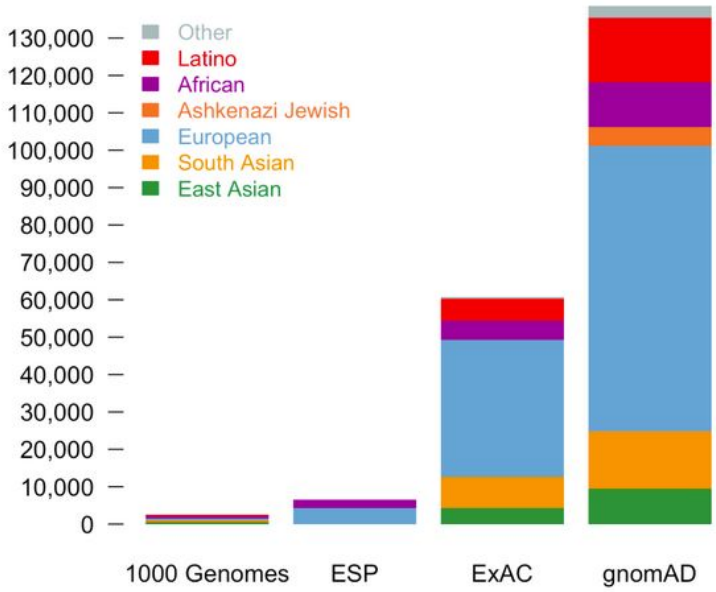
The 1000 Genomes Phase 3 data provides genotype data for 2504 individuals from 26 different narrowly defined populations in 5 groupings.

Allele Frequencies- GnomAD

The Genome Aggregation Database provides allele frequency data for over 130,000 samples from 7 different populations

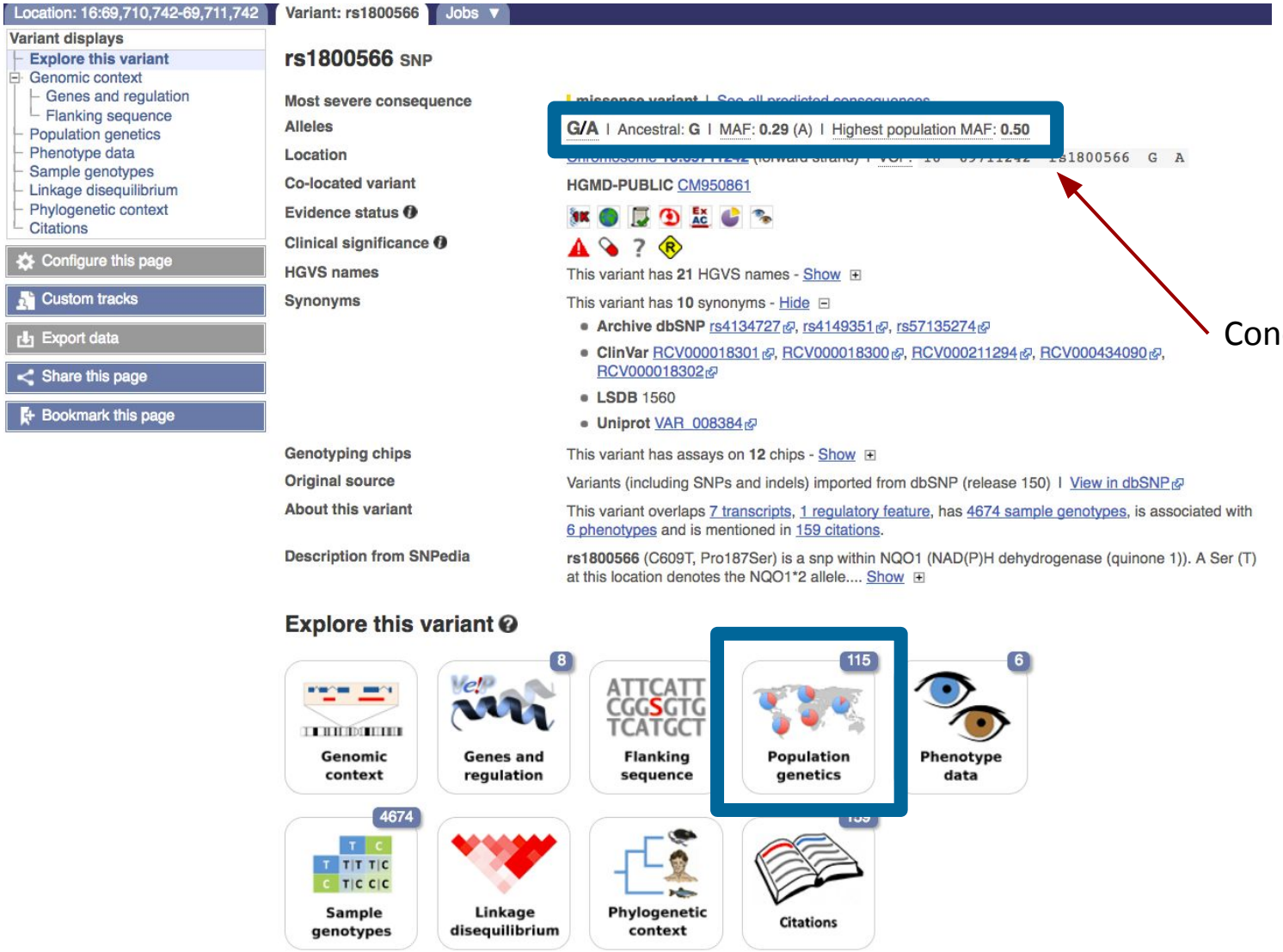
POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632

Sample numbers



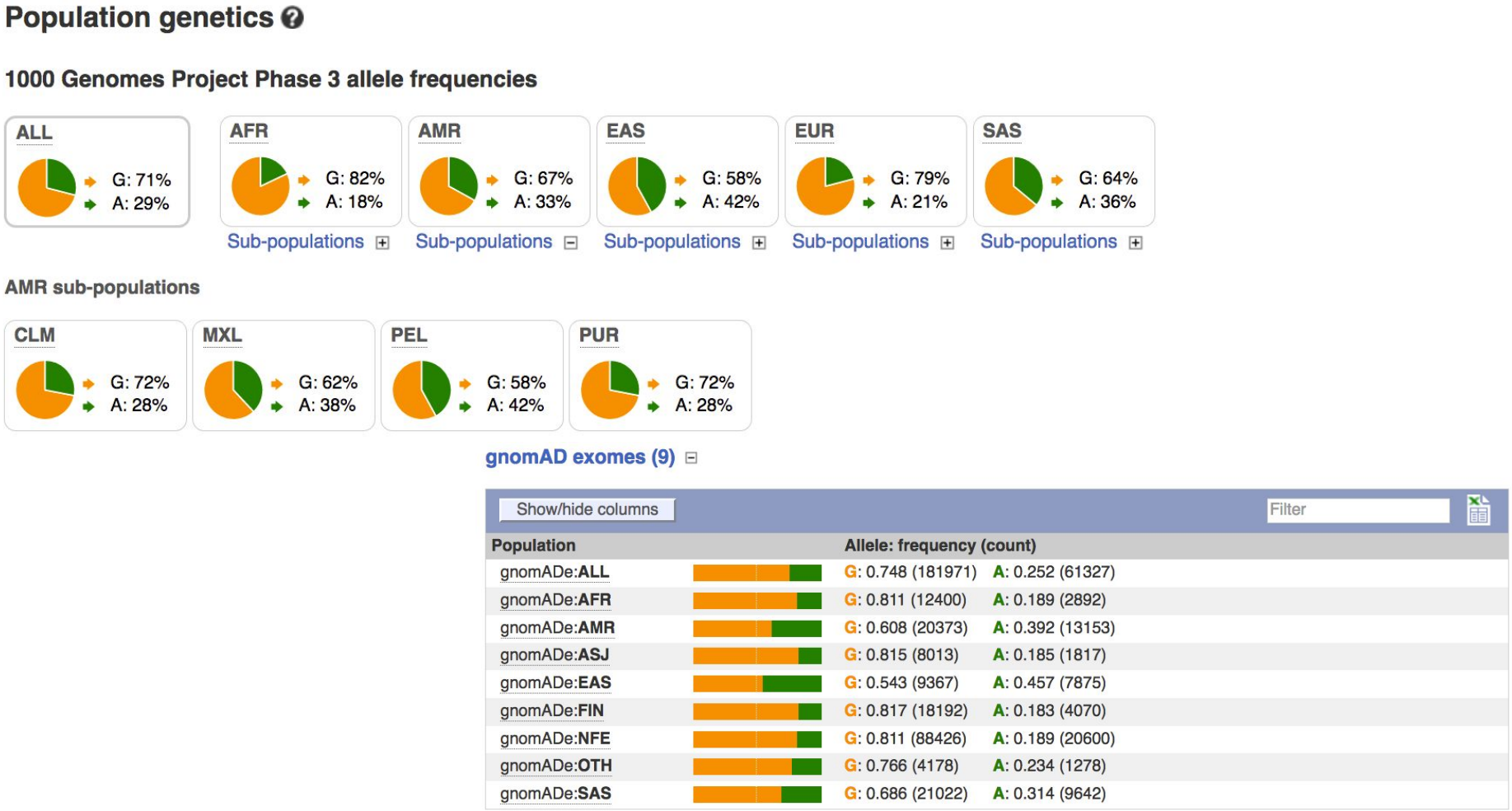
From: https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

Allele Frequencies



http://training.ensembl.org/events

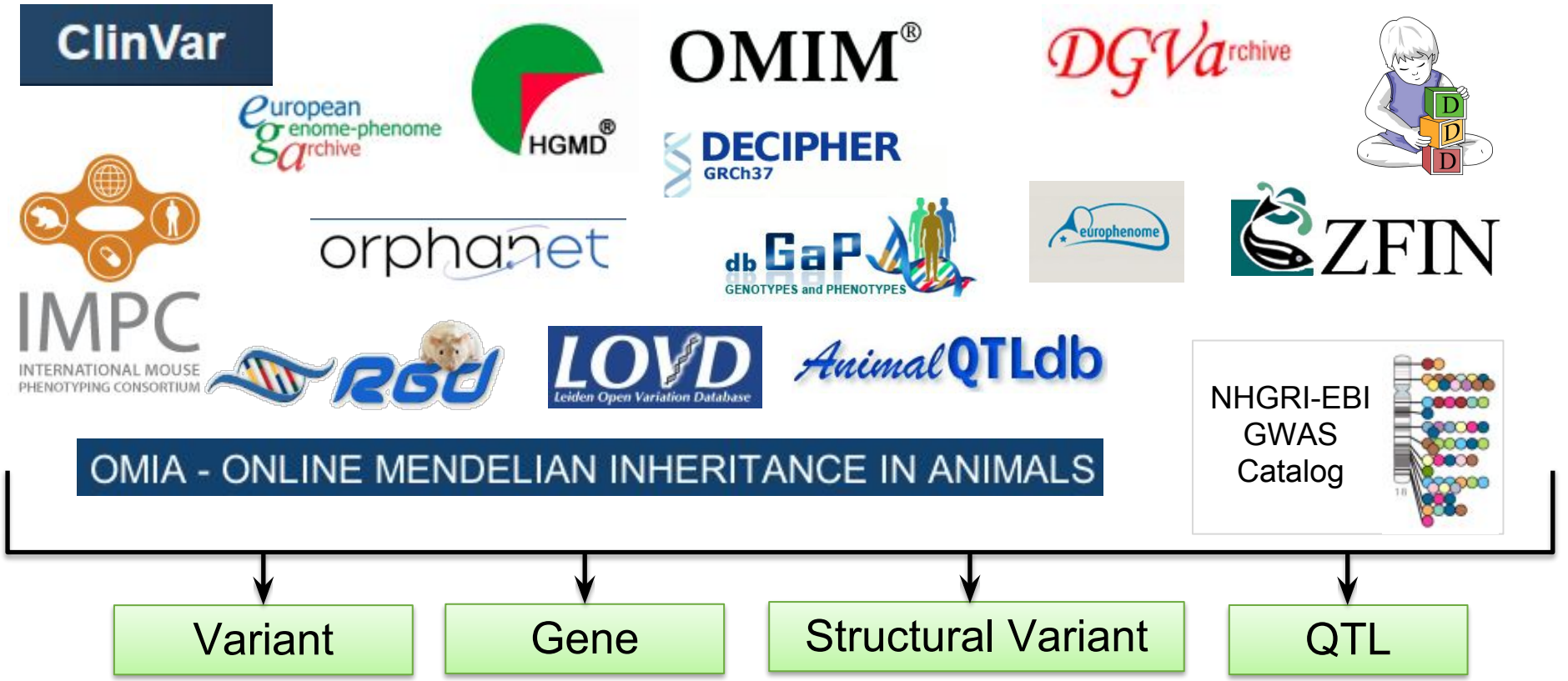
Allele Frequency Distributions



Pie charts and tables display frequency data for different studies. These include TOPMed and UK10K (ALSPAC and TWINS UK) for human, WTSI Mouse Genomes Project and NextGen

http://training.ensembl.org/events

Phenotype and Disease Data

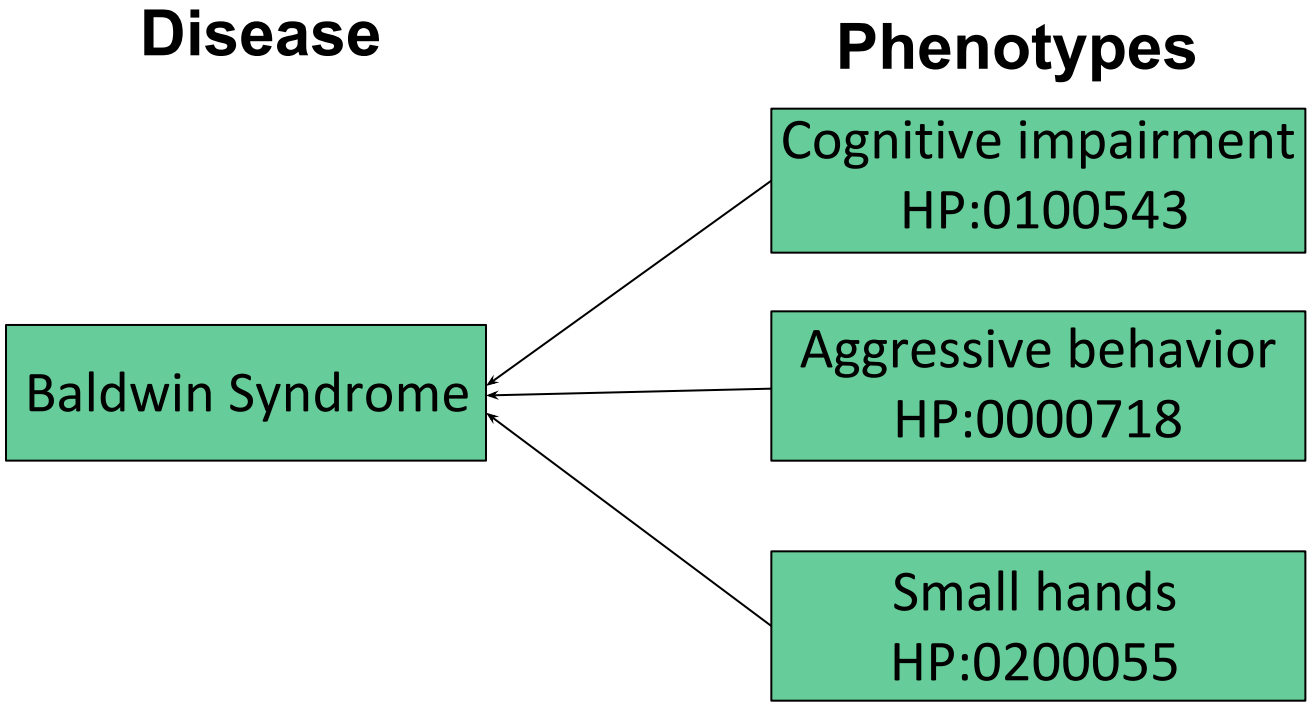


Attributes types held for phenotype features include:

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

Terminology

A disease can be considered as a bundle of phenotypes observed in the subjects, often named after the person who studied and characterised it.



Hand size can also be considered a **trait**

The Naming Problem

Different sources describe the phenotypes/ diseases in different ways

- We need to normalise the descriptions for easy querying
- Also need to respect the original data

Phenotype Data

Significant association(s)

rs2470893 SNP

Show All entries | Show/hide columns (3 hidden) | Filter

Phenotype, disease and trait	Source(s)	Study	Reported gene(s)	Associated allele	Statistics
Caffeine consumption	NHGRI-EBI GWAS catalog	PMID:21490707	EDC3, CYP1A1, CYP1A2, LMAN1L, CSK	T	p-value: $5.00e^{-14}$ beta coefficient: 0.12 mg/day increase
Coffee consumption	NHGRI-EBI GWAS catalog	PMID:21876539	CYP1A1, CYP1A2	T	p-value: $2.00e^{-11}$ beta coefficient: 0.0675 unit increase
Coffee consumption	NHGRI-EBI GWAS catalog	PMID:25288136	CYP1A1, CYP1A2	T	p-value: $5.00e^{-19}$ beta coefficient: 0.2 unit increase

Features are often reported as associated with a disease sub-type making it complex to extract all loci of interest

The Solution

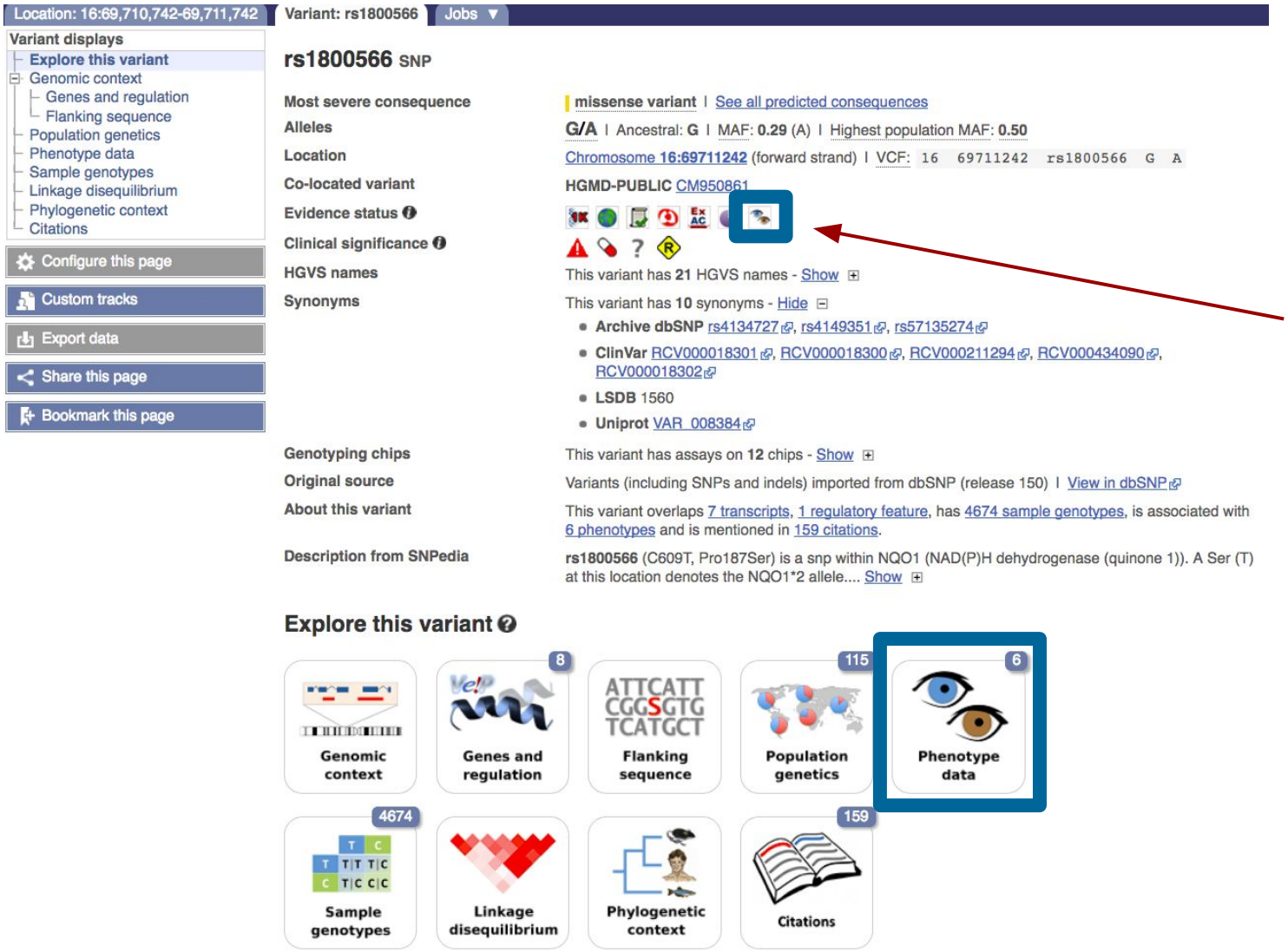
Use phenotype and disease ontologies to organise descriptions and allow query and display by synonym and child/parent term.

Viewing Ontology Mappings by Variant

Additional columns in the table of diseases linked to a variant show the Ontology terms we have mapped to the description

It is now clear these descriptions refer to the same disease

Phenotype and Disease Data



Very concise summary

http://training.ensembl.org/events

Phenotype and Disease Data

Phenotype Data

Significant association(s)

Show/hide columns | Filter

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	Study	Clinical significance	Reported gene(s)	Associated allele
BENZENE TOXICITY, SUSCEPTIBILITY TO	OMIM	-	-	MIM:125860	-	NQO1	0001
Alkylating Agents, anthracyclines and related substances, fluorouracil, and Platinum compounds response - Efficacy	ClinVar	-	-	-	★★★☆	NQO1	A
Lung Cancer	ClinVar	lung adenocarcinoma, lung carcinoma, squamous cell carcinoma	EFO:0000571, EFO:0000707, EFO:0001071	-	? ☆☆☆☆	NQO1	A
BENZENE TOXICITY, SUSCEPTIBILITY TO	ClinVar	-	-	-	R ☆☆☆☆	NQO1	A
Leukemia, post-chemotherapy, susceptibility to	ClinVar	leukemia	EFO:0000565	-	R ☆☆☆☆	NQO1	A
Breast cancer, post-chemotherapy poor survival in	ClinVar	breast carcinoma	EFO:0000305	-	! ☆☆☆☆	NQO1	A

Citation Data - Sources

dbSNP

- Publication information is submitted with the variant submission.

EuropePMC

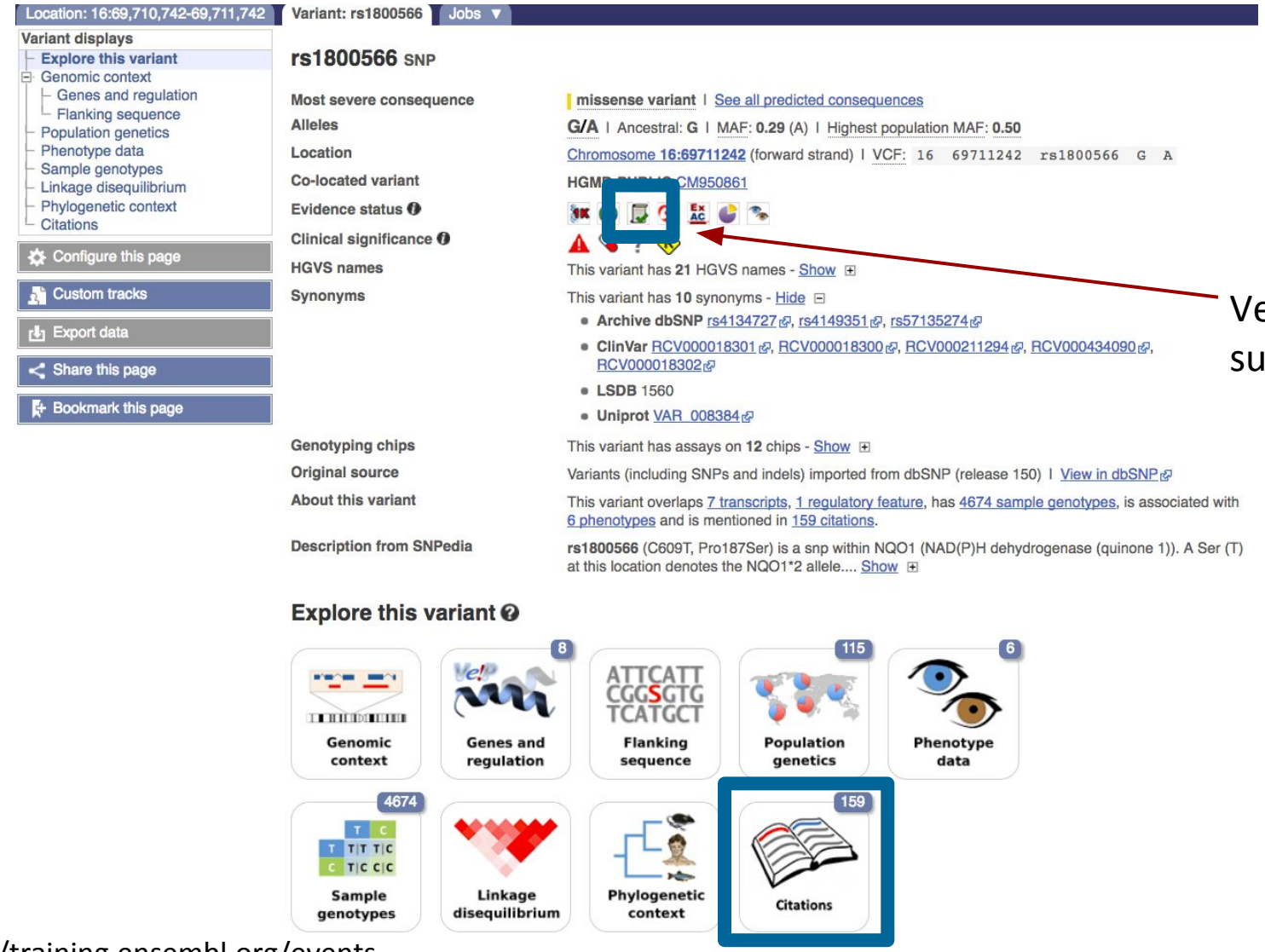
- Publications for which the full text is freely available in PubMed Central are mined for refSNP identifiers.

UCSC Genocoding Project

- Publications for which the text is freely available or those from publishers who have agreed to data mining are mined for refSNP identifiers.

Species	Citations
sheep	31
horse	58
pig	73
rat	90
dog	200
chicken	414
cow	1,625
mouse	8,309
human	594,364

Citation Data



Very concise summary

http://training.ensembl.org/events

Citation Data

Citations

rs1800566 is mentioned in the following publications

Show All entries | Show/hide columns | Filter

Year	PMID	Title	Author(s)	Full text
2017	28291250	Site-to-site interdomain communication may mediate different loss-of-function mechanisms in a cancer-associated NQO1 polymorphism.	Medina-Carmona E, Neira JL, Salido E, Fuchs JE, Palomino-Morales R, Timson DJ, Pey AL.	PMC5349528
2017	28529598	Biological interaction of cigarette smoking on the association between genetic polymorphisms involved in inflammation and the risk of lung cancer: A case-control study in Japan.	Yamamoto Y, Kiyohara C, Suetsugu-Ogata S, Hamada N, Nakanishi Y.	PMC5431513
2017	28282928	Oxidative Stress: A New Target for Pancreatic Cancer Prognosis and Treatment.	Martinez-Useros J, Li W, Cabeza-Morales M, Garcia-Foncillas J.	PMC5372998
2016	26801900	Pharmacogenetics driving personalized medicine: analysis of genetic polymorphisms related to breast cancer medications in Italian isolated populations.	Cocca M, Bedognetti D, La Bianca M, Gasparini P, Girotto G.	PMC4722680
2016	27015811	Associations of air pollution exposure with blood pressure and heart rate variability are modified by oxidative stress genes: A repeated-measures panel among elderly urban residents.	Kim KN, Kim JH, Jung K, Hong YC.	PMC4807581
2016	27019599	Association between L55M polymorphism in Paraoxonase 1 and cancer risk: a meta-analysis based on 21 studies.	Chen L, Lu W, Fang L, Xiong H, Wu X, Zhang M, Wu S, Yu D.	PMC4786067
2016	26873362	Gene Polymorphism Association with Type 2 Diabetes and Related Gene-Gene and Gene-Environment Interactions in a Uyghur Population.	Xiao S, Zeng X, Fan Y, Su Y, Ma Q, Zhu J, Yao H.	PMC4755665
2016	26426434	Necrotizing Enterocolitis Is Not Associated With Sequence Variants in Antioxidant Response Genes in Premature Infants.	Sampath V, Helbling D, Menden H, Dimmock D, Mulrooney NP, Murray JC, Dagle JM, Garland JS.	PMC5055643
2016	26445852	Genetic Biomarkers of Barrett's Esophagus Susceptibility and Progression to Dysplasia and Cancer: A Systematic Review and Meta-Analysis.	Findlay JM, Middleton MR, Tomlinson I.	PMC4700058
2016	26868429	Heme Oxygenase-1 and 2 Common Genetic Variants and Risk for Multiple Sclerosis.	Agúndez JA, Garcí…, Benito-León J, Millán-Pascual J, D… a P, Pisa D, Turpín-Fenoll L, Al… r P, Ortega-Cubero S, Ayuso-P… García-Albea E, Plaza-Nieto JF, Jiménez-Jiménez FJ.	PMC4751624

Link to full text

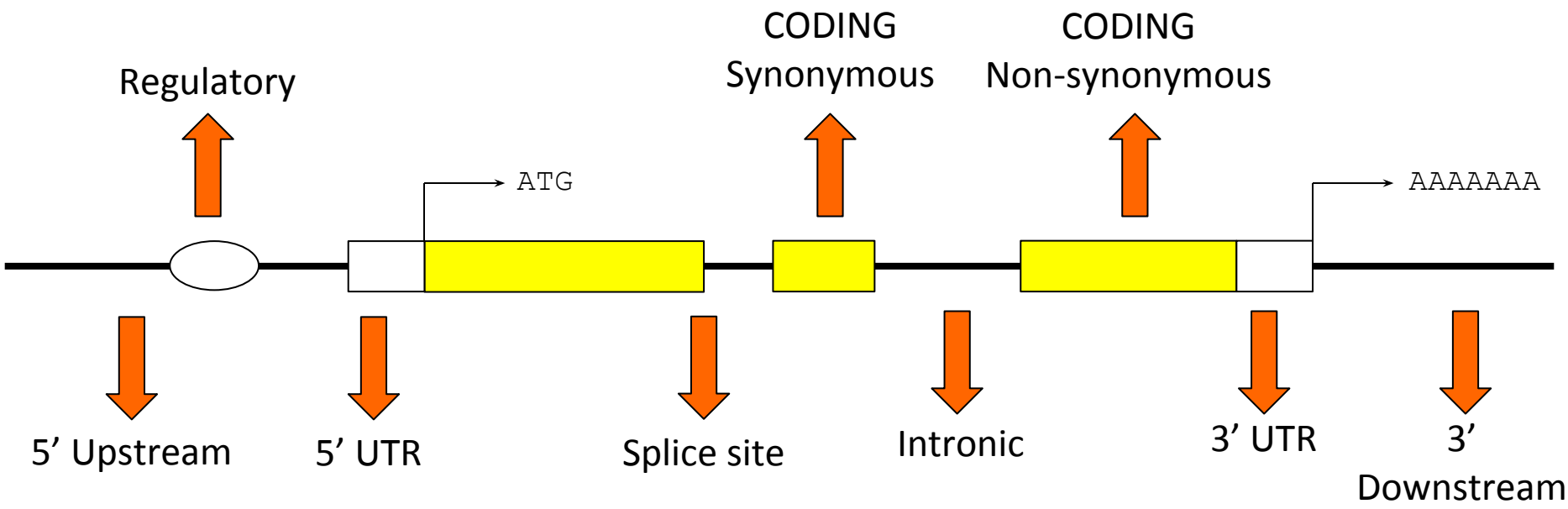
http://training.ensembl.org/events

Variant Annotation

We analyse the imported variation data utilising other Ensembl resources. Annotation provided includes:

- Comparison to Ensembl gene structures to predict transcript/ protein consequences
- Comparison to Ensembl regulatory features to find co-location
- Comparison to other species to predict ancestral allele, show phylogenetic context
- Linkage disequilibrium calculations for the 1000 Genomes populations

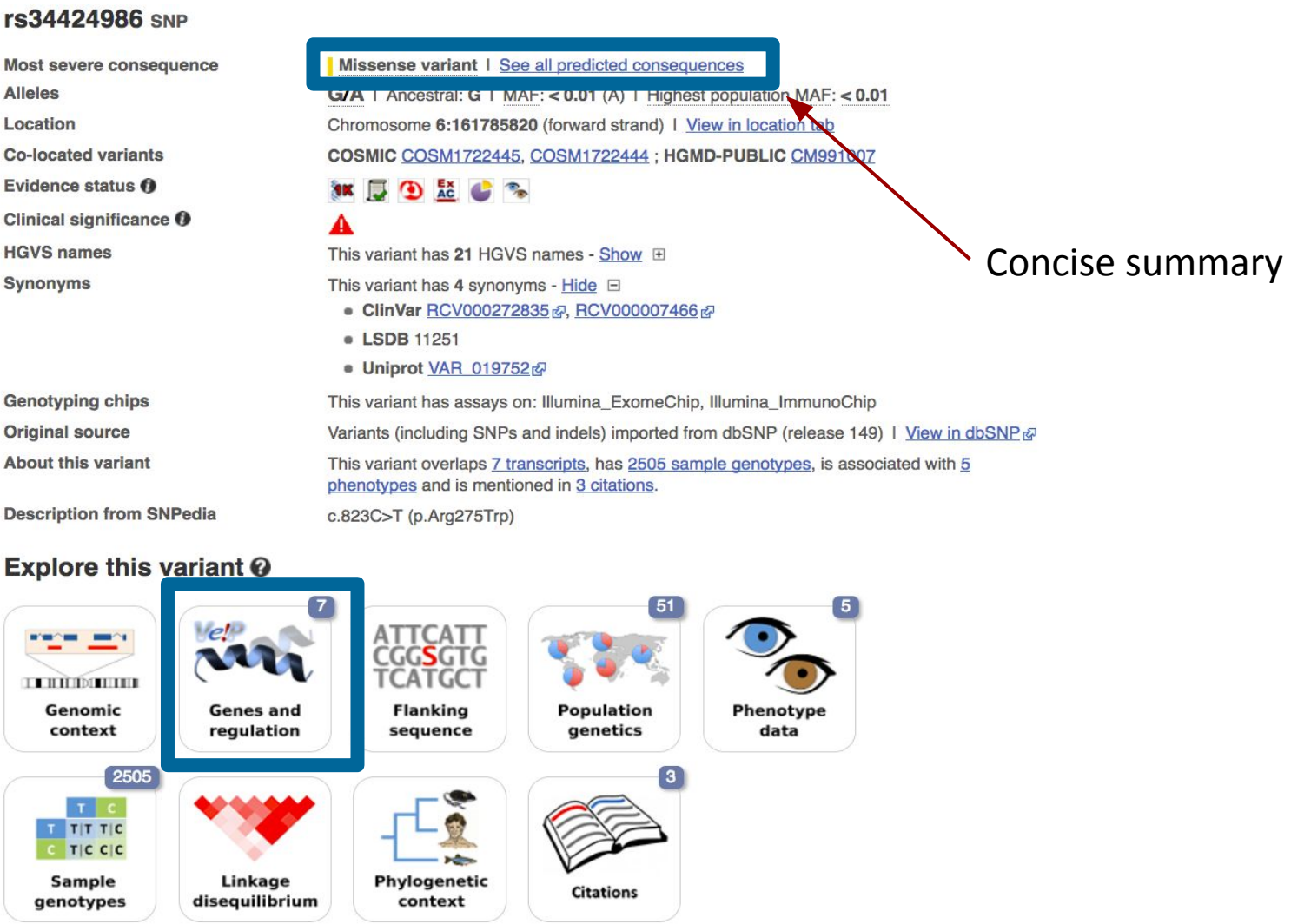
Variant consequences



We calculated the predicted consequences of variants on the Ensembl transcript set and assign Sequence Ontology terms

The Sequence Ontology Project

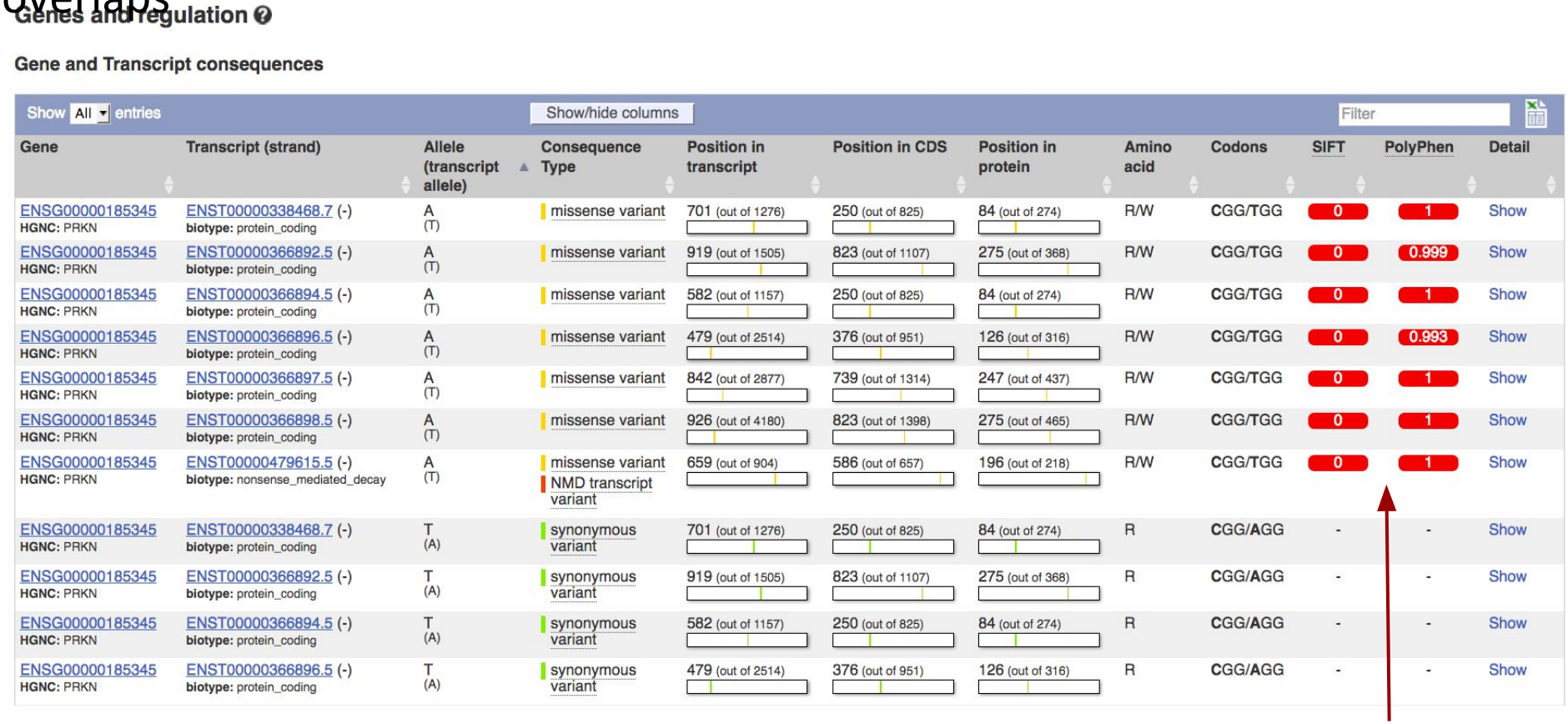
Genes and Regulation



http://training.ensembl.org/events

Genes and Regulation

Tables show the predicted impact of a variant on all of the transcripts it overlaps



Genes and regulation

Gene and Transcript consequences

Show All entries | Show/hide columns | Filter

Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen	Detail
ENSG00000185345 HGNC: PRKN	ENST00000338468.7 (-) biotype: protein_coding	A (T)	missense variant	701 (out of 1276)	250 (out of 825)	84 (out of 274)	R/W	CGG/TGG	0	1	Show
ENSG00000185345 HGNC: PRKN	ENST00000366892.5 (-) biotype: protein_coding	A (T)	missense variant	919 (out of 1505)	823 (out of 1107)	275 (out of 368)	R/W	CGG/TGG	0	0.999	Show
ENSG00000185345 HGNC: PRKN	ENST00000366894.5 (-) biotype: protein_coding	A (T)	missense variant	582 (out of 1157)	250 (out of 825)	84 (out of 274)	R/W	CGG/TGG	0	1	Show
ENSG00000185345 HGNC: PRKN	ENST00000366896.5 (-) biotype: protein_coding	A (T)	missense variant	479 (out of 2514)	376 (out of 951)	126 (out of 316)	R/W	CGG/TGG	0	0.993	Show
ENSG00000185345 HGNC: PRKN	ENST00000366897.5 (-) biotype: protein_coding	A (T)	missense variant	842 (out of 2877)	739 (out of 1314)	247 (out of 437)	R/W	CGG/TGG	0	1	Show
ENSG00000185345 HGNC: PRKN	ENST00000366898.5 (-) biotype: protein_coding	A (T)	missense variant	926 (out of 4180)	823 (out of 1398)	275 (out of 465)	R/W	CGG/TGG	0	1	Show
ENSG00000185345 HGNC: PRKN	ENST00000479615.5 (-) biotype: nonsense_mediated_decay	A (T)	missense variant, NMD transcript variant	659 (out of 904)	586 (out of 657)	196 (out of 218)	R/W	CGG/TGG	0	1	Show
ENSG00000185345 HGNC: PRKN	ENST00000338468.7 (-) biotype: protein_coding	T (A)	synonymous variant	701 (out of 1276)	250 (out of 825)	84 (out of 274)	R	CGG/AGG	-	-	Show
ENSG00000185345 HGNC: PRKN	ENST00000366892.5 (-) biotype: protein_coding	T (A)	synonymous variant	919 (out of 1505)	823 (out of 1107)	275 (out of 368)	R	CGG/AGG	-	-	Show
ENSG00000185345 HGNC: PRKN	ENST00000366894.5 (-) biotype: protein_coding	T (A)	synonymous variant	582 (out of 1157)	250 (out of 825)	84 (out of 274)	R	CGG/AGG	-	-	Show
ENSG00000185345 HGNC: PRKN	ENST00000366896.5 (-) biotype: protein_coding	T (A)	synonymous variant	479 (out of 2514)	376 (out of 951)	126 (out of 316)	R	CGG/AGG	-	-	Show

The SIFT and PolyPhen2 packages predict how likely a variant is to be damaging to a protein. Red indicates damaging; green indicates tolerated; blue indicates a low quality prediction

Summary

The Ensembl genome browser displays:

- imported variation data, including phenotype associations, allele frequencies and literature citations from a variety of different sources

- additional variant annotation, including functional consequence predictions

- this data can be accessed through the web-interface, BioMart, Perl API, REST API and the FTP site

Ensembl Acknowledgements

The Entire Ensembl Team

Daniel R. Zerbino[1], Premanand Achuthan[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Jyothish Bhai[1], Konstantinos Billis[1], Carla Cummins[1], Astrid Gall[1], Carlos García Giroń[1], Laurent Gil[1], Leo Gordon[1], Leanne Haggerty[1], Erin Haskell[1], Thibaut Hourlier[1], Osagie G. Izuogu[1], Sophie H. Janacek[1], Thomas Juettemann[1], Jimmy Kiang To[1], Matthew R. Laird[1], Ilias Lavidas[1], Zhicheng Liu[1], Jane E. Loveland[1], Thomas Maurel[1], William McLaren[1], Benjamin Moore[1], Jonathan Mudge[1], Daniel N. Murphy[1], Victoria Newman[1], Michael Nuhn[1], Denye Ogeh[1], Chuang Kee Ong[1], Anne Parker[1], Mateus Patricio[1], Harpreet Singh Riat[1], Helen Schuilenburg[1], Dan Sheppard[1], Helen Sparrow[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Brandon Walts[1], Amonida Zadissa[1], Adam Frankish[1], Sarah E. Hunt[1], Myrto Kostadima[1], Nicholas Langridge[1], Fergal J. Martin[1], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Dan M. Staines[1], Stephen J. Trevanion[1], Bronwen L. Aken[1], Fiona Cunningham[1], Andrew Yates[1] and Paul Flicek[1,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Eagle Genomics Ltd., Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK and [3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Funding

EMBL

wellcome

Open Targets

BBSRC bioscience for the future

National Human Genome Research Institute

TRANSFORMING GENETIC MEDICINE INITIATIVE

SAVE THE TASMANIAN DEVIL PROGRAM APPEAL

Co-funded by the European Union

Training materials

- Ensembl training materials are protected by a CC BY license
- http://creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers
- http://www.ensembl.org/info/about/publications.html

http://training.ensembl.org/events