HelmholtzZentrum münchen
German Research Center for Environmental Health

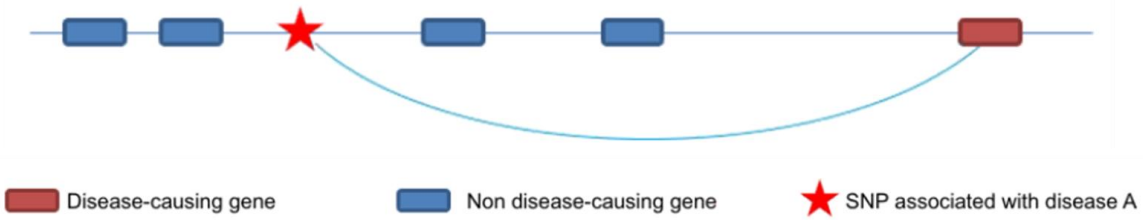
Public bioinformatics resources for GWAS interpretation

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

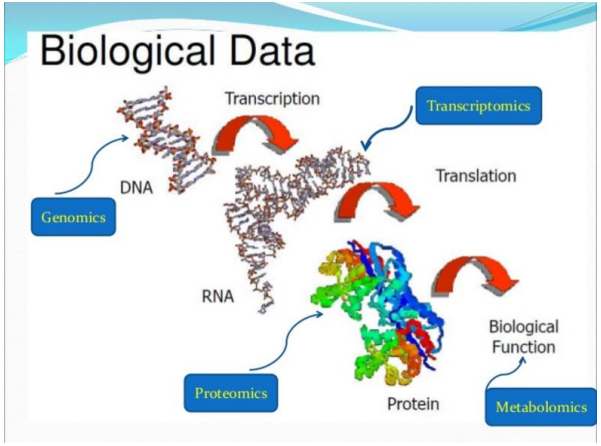
Dr. Konstantinos Hatzikotoulas
Institute of Translational Genomics

Motivation

- Greatest challenge in genetic association studies is to assign functionality (causality) to associated SNPs and genes.
- Predict the most likely genes and variants driving the phenotype associations.
- Need to integrate as much information as possible through genome annotation tools to drive the experimental validation.



What is Bioinformatics?



+

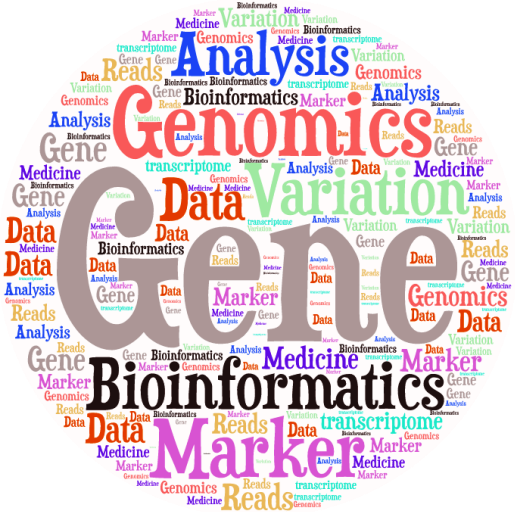
Computers aid to collect, intergrade and analyse biological data ->
Optimal biological interpretation

Goal of Bioinformatics

- Develop **databases** and **computational tools** to generate knowledge to better understanding a living cell and how it functions at molecular level.

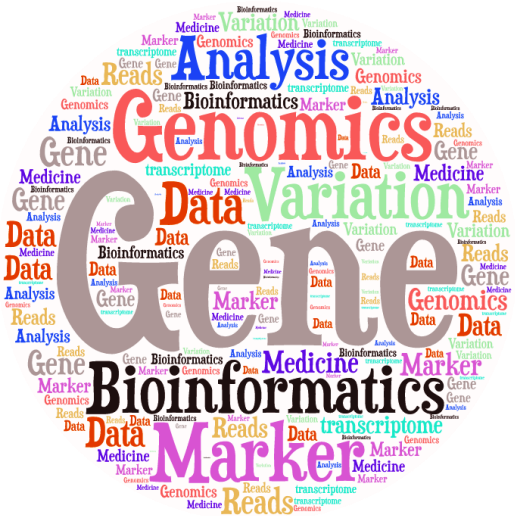
Publicly available bioinformatics resources

- Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database



Publicly available bioinformatics resources

- Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database

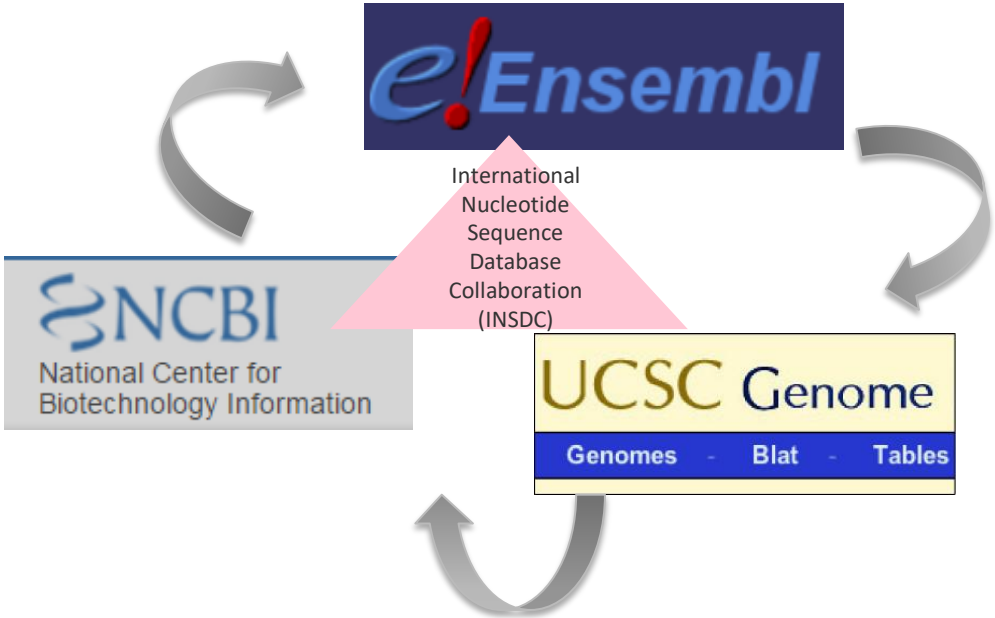


Genome Browsers

- Ensembl : http://www.ensembl.org
- UCSC: http://genome.ucsc.edu/
- NCBI: https://www.ncbi.nlm.nih.gov/

Genome Browsers

- Agreement: To begin bringing all genome assembly data into compliance moving forward



Ensembl … more than a browser!

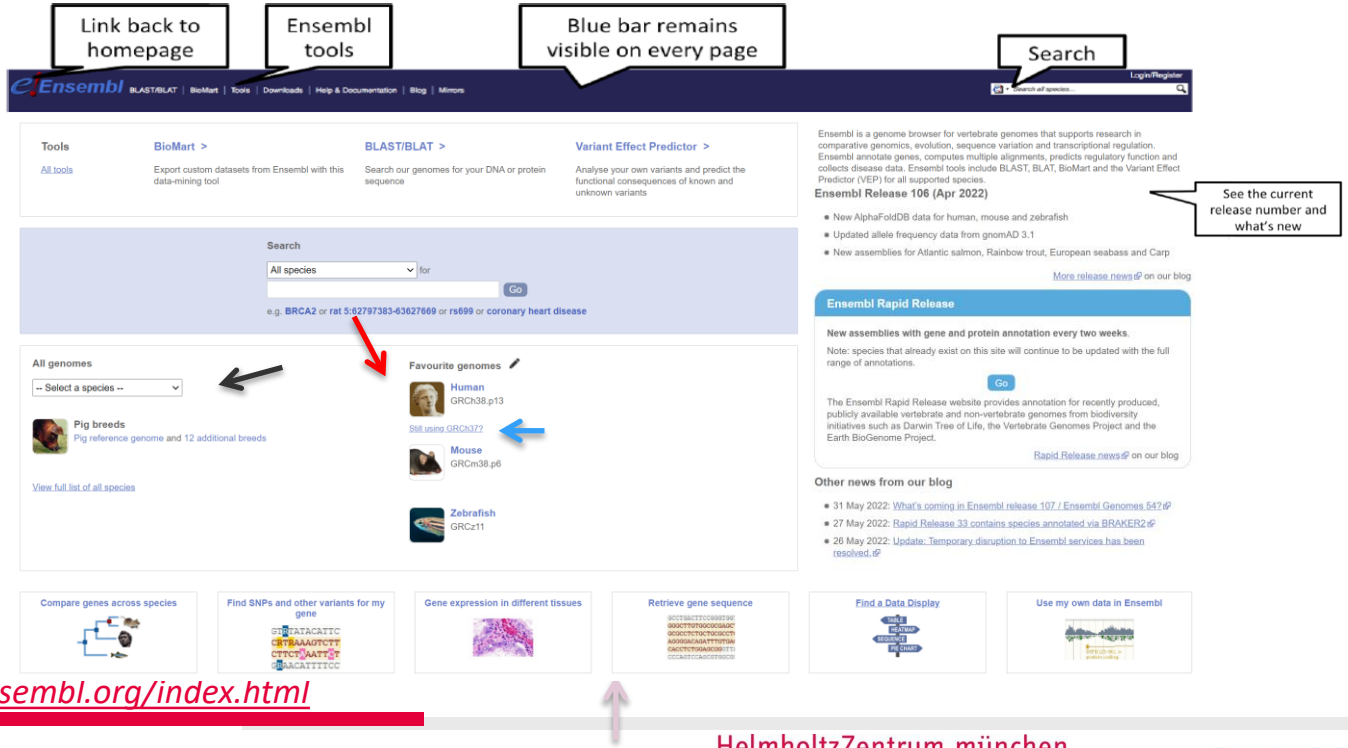
- Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute
- **What can I do with Ensembl?**
  - Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region.
  - View SNPs across strains (rat, mouse), populations (human) or breeds (dog).
  - View genes with annotations along the chromosome.
  - View alternative transcripts (i.e. splice variants) for a given gene.
  - Upload your own data.
  - Export sequence or create a table of gene information with BioMart.
  - Determine how your variants affect genes and transcripts using the Variant Effect Predictor.
  - Share Ensembl views with your colleagues and collaborators.

Ensembl … more than a browser!

- Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute
- **What can I do with Ensembl?**
  - Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region.
  - View SNPs across strains (rat, mouse), populations (human) or breeds (dog).
  - View genes with other annotation along the chromosome.
  - View alternative transcripts (i.e. splice variants) for a given gene.
  - Upload your own data.
  - Export sequence or create a table of gene information with BioMart.
  - Determine how your variants affect genes and transcripts using the Variant Effect Predictor.
  - Share Ensembl views with your colleagues and collaborators.

Ensembl … the front page

- It contains lots of information and links to help you navigate Ensembl.



Hint: https://www.ensembl.org/index.html

Ensembl … the front page … Exercise 1

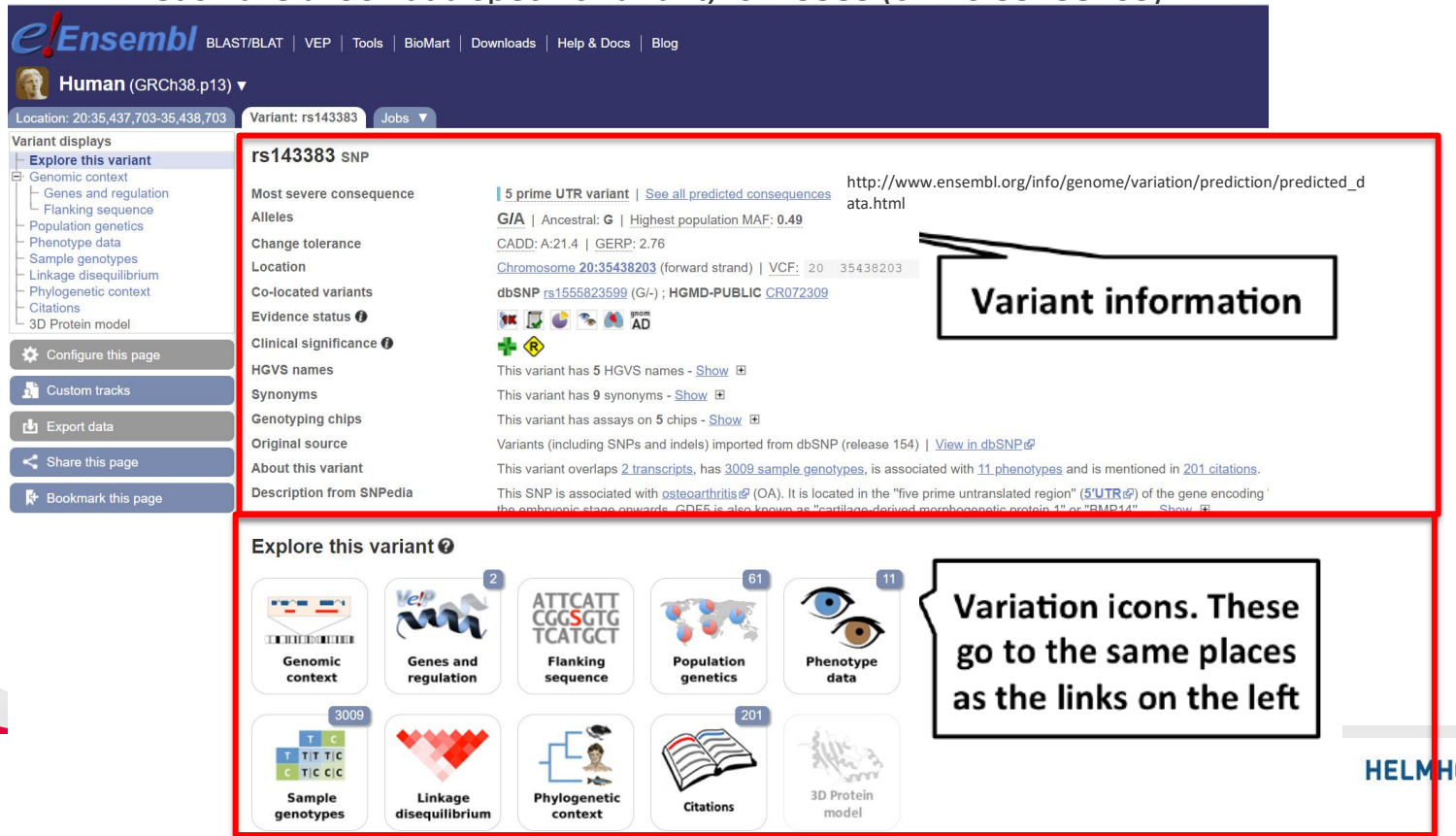
1. Go to the species homepage for **Giant Panda**. What is the name of the genome assembly for Panda?

2. How long is the Panda genome (in bp)? How many coding genes have been annotated?

3. Repeat for **Human**.

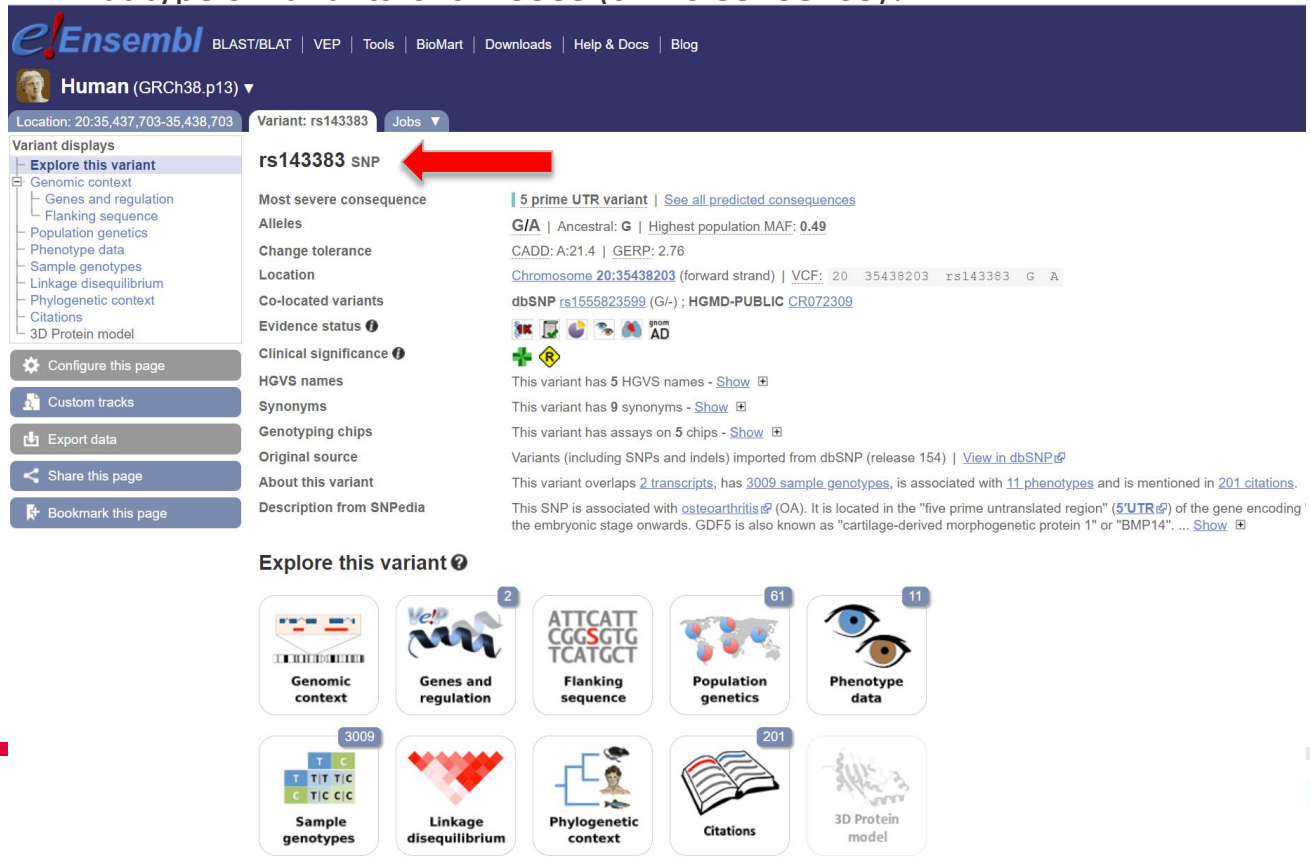
Ensembl … Exploring variants

- Let's have a look at a specific variant, rs143383 (chr20:35438203).



Ensembl … Exploring variants … Variant types

➢ What type of variants is rs143383 (chr20:35438203)?



Ensembl … Exploring variants …
Variant types

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

Type	Description	Example (Reference / Alternative)
SNP	Single Nucleotide Polymorphism	Ref: ...TTG**A**CGTA... Alt: ...TTG**G**CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA... Alt: ...TTGA**TG**CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG**AC**GTA... Alt: ...TTGGTA...
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: ...TTG**A**CGTA... Alt: ...TTG**GCT**CGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG**AC**GTA... Alt: ...TTG**TA**GTA...

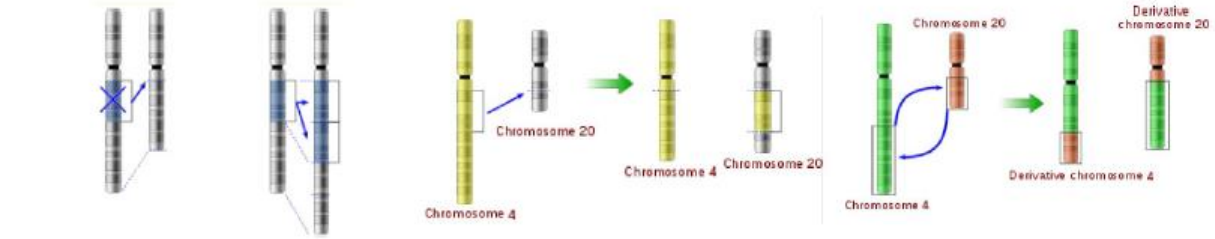
2. Large scale in chromosomal structure

CNV Copy Number Variation: increases or decreases the copy number of a given region

Reference:

"Gain" of one copy:

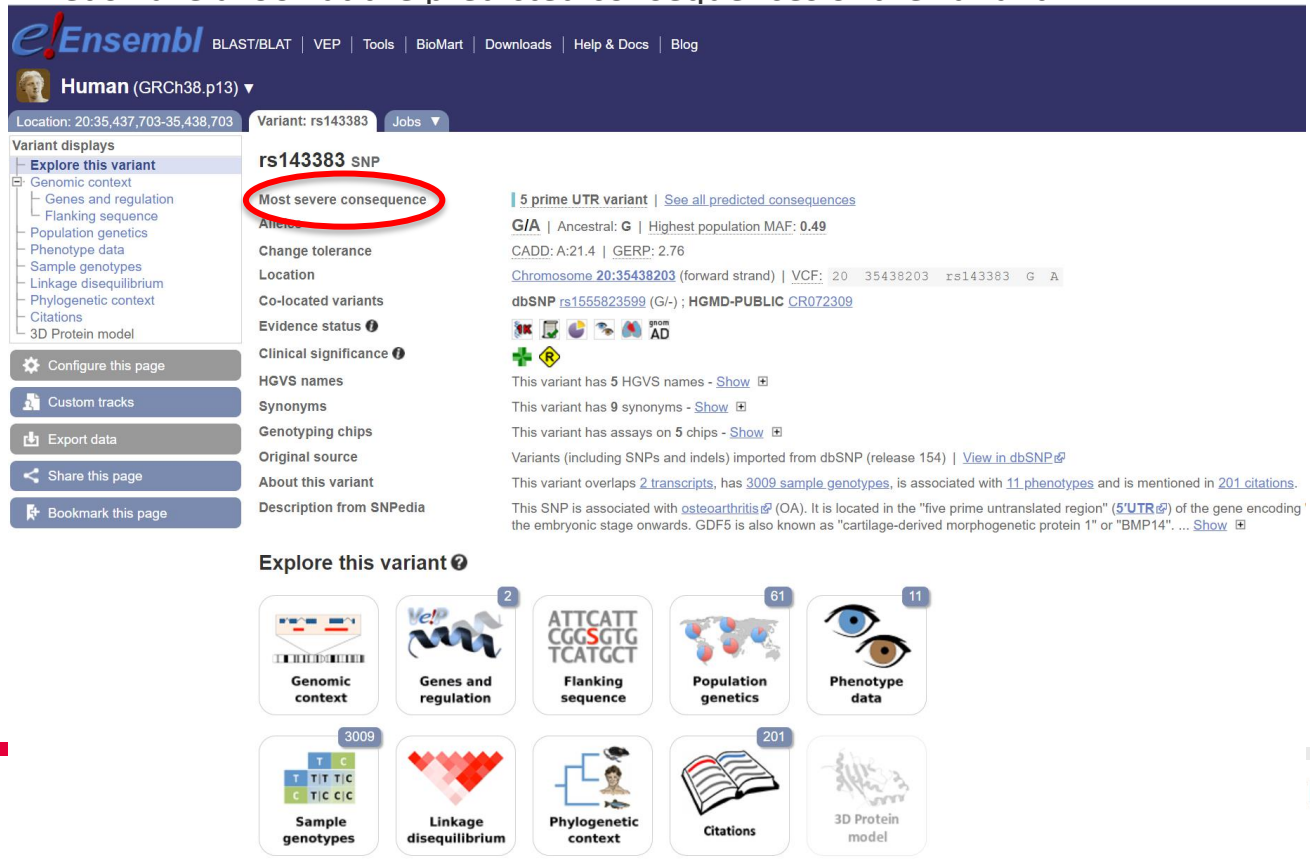
"Loss" of one copy:



deletion duplication insertion translocation

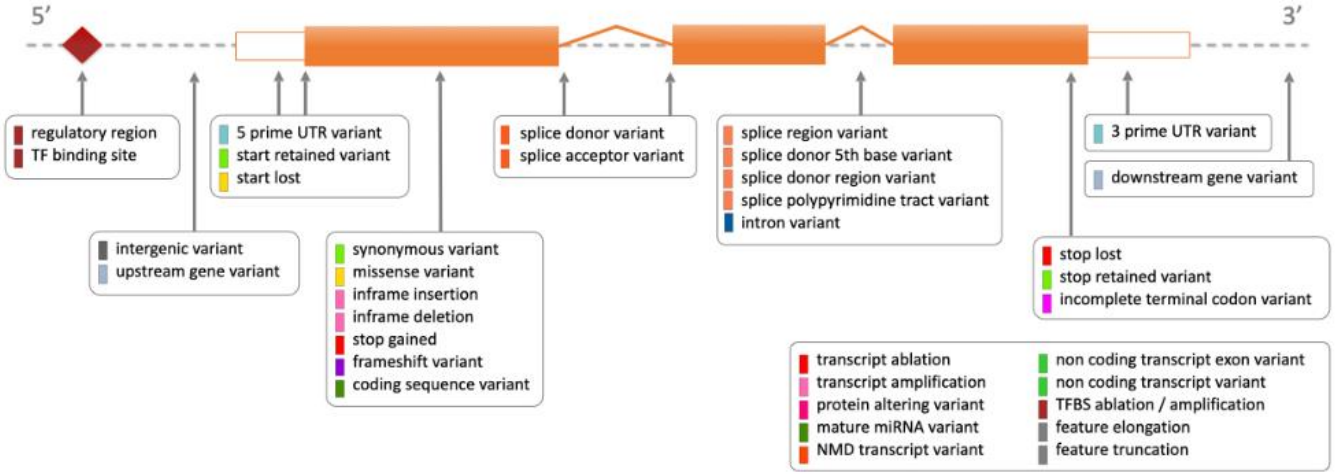
Ensembl … Exploring variants … Calculated variant consequences

➢ Let's have a look at the predicted consequences of the variant.



Ensembl … Exploring variants … Calculated variant consequences

- A rule-based approach to predict the effects that each allele of the variant may have on the transcript.



Ensembl … Exploring variants …
Calculated variant consequences

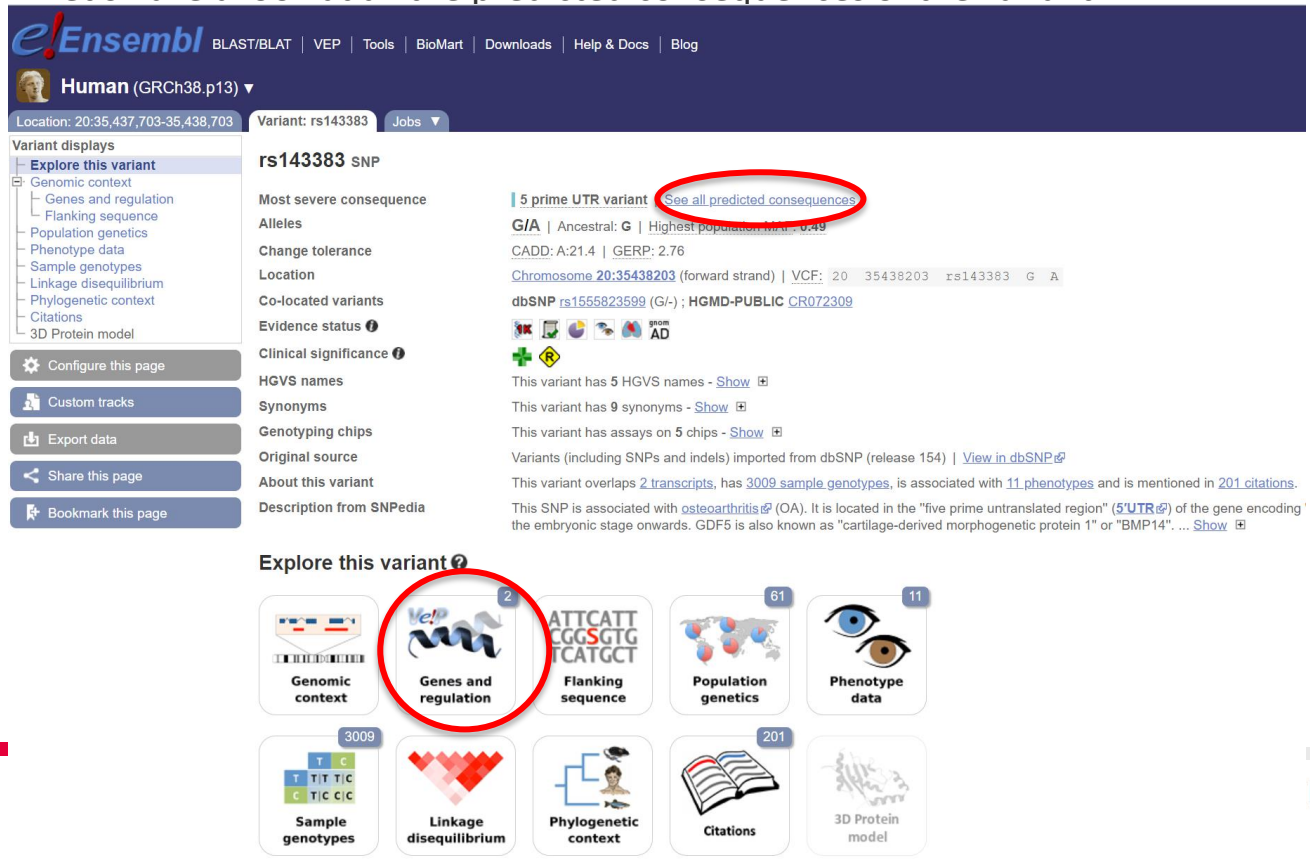
➢ Sequence Ontology (SO) consequence term :

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

SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
splice_donor_5th_base_variant	A sequence variant that causes a change at the 5th base pair after the start of the intron in the orientation of the transcript	SO:0001787	Splice donor 5th base variant	LOW
splice_donor_region_variant	A sequence variant that falls in the region between the 3rd and 6th base after splice junction (5' end of intron)	SO:0002170	Splice donor region variant	LOW
splice_polypyrimidine_tract_variant	A sequence variant that falls in the polypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO:0002169	Splice polypyrimidine tract variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW
coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding sequence variant	MODIFIER
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Mature miRNA variant	MODIFIER
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5 prime UTR variant	MODIFIER
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3 prime UTR variant	MODIFIER
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	SO:0001792	Non coding transcript exon variant	MODIFIER
intron_variant	A transcript variant occurring within an intron	SO:0001627	Intron variant	MODIFIER
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript variant	MODIFIER
non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO:0001619	Non coding transcript variant	MODIFIER
upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream gene variant	MODIFIER
downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream gene variant	MODIFIER
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO:0001895	TFBS ablation	MODIFIER
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	SO:0001892	TFBS amplification	MODIFIER
TF_binding_site_variant	A sequence variant located within a transcription factor binding site	SO:0001782	TF binding site variant	MODIFIER
regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	SO:0001894	Regulatory region ablation	MODERATE
regulatory_region_amplification	A feature amplification of a region containing a regulatory region	SO:0001891	Regulatory region amplification	MODIFIER
feature_elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence	SO:0001907	Feature elongation	MODIFIER
regulatory_region_variant	A sequence variant located within a regulatory region	SO:0001566	Regulatory region variant	MODIFIER
feature_truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	SO:0001906	Feature truncation	MODIFIER
intergenic_variant	A sequence variant located in the intergenic region, between genes	SO:0001628	Intergenic variant	MODIFIER

Ensembl … Exploring variants … Calculated variant consequences

➢ Let's have a look at all the predicted consequences of the variant.



Ensembl … Exploring variants … Calculated variant consequences

➢ This variant is found in 2 transcripts of the *GDF5* gene.

Genes and regulation

Gene and Transcript consequences

Show/hide columns

Filter

Gene	Transcript (strand)	Allele (Tr. allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	AA	Codons	Detail
ENSG00000125965 HGNC: GDF5	ENST00000374369.8 (-) biotype: protein_coding	A (T)	5 prime UTR variant	26 (out of 2368)	-	-	-	-	Show
ENSG00000125965 HGNC: GDF5	ENST00000374372.1 (-) biotype: protein_coding	A (T)	intron variant	-	-	-	-	-	Show

coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding sequence variant	MODIFIER
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Mature miRNA variant	MODIFIER
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5 prime UTR variant	MODIFIER
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3 prime UTR variant	MODIFIER
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	SO:0001792	Non coding transcript exon variant	MODIFIER
intron_variant	A transcript variant occurring within an intron	SO:0001627	Intron variant	MODIFIER
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript variant	MODIFIER
non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO:0001619	Non coding transcript variant	MODIFIER

Variant: rs143383 | Jobs

rs143383 SNP

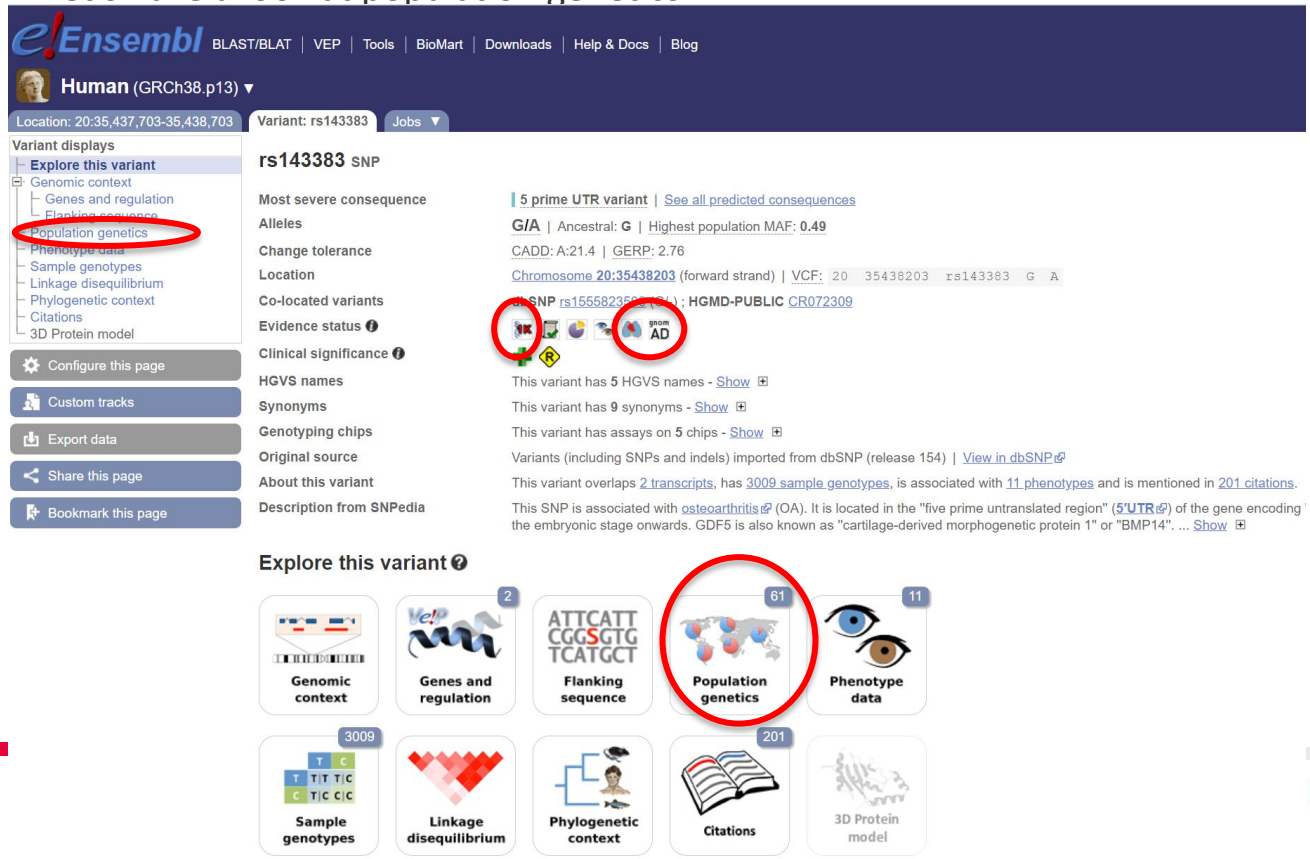
Most severe consequence: 5 prime UTR variant | See all predicted consequences

Alleles: G/A | Ancestral: G | MAF: … (A) | Highest population MAF: 0.49

Change tolerance: CADD: A:20.6 | GERP: 2.34

Ensembl … Exploring variants… Population genetics

- Let's have a look at population genetics.



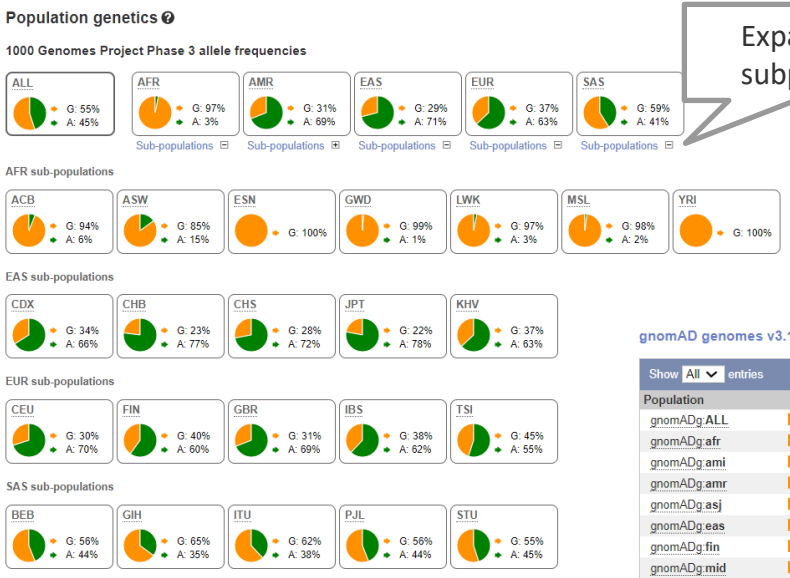
Ensembl … Exploring variants … Population genetics

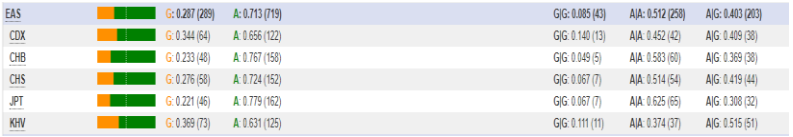
- Frequencies are grouped by project for studies with multiple subpopulations.
- Pie charts can be displayed only for 1000 Genomes allele frequencies.

Pie charts of allele frequencies

Expand subpopulations

Tables of more detailed data





gnomAD genomes v3.1.2 (11)

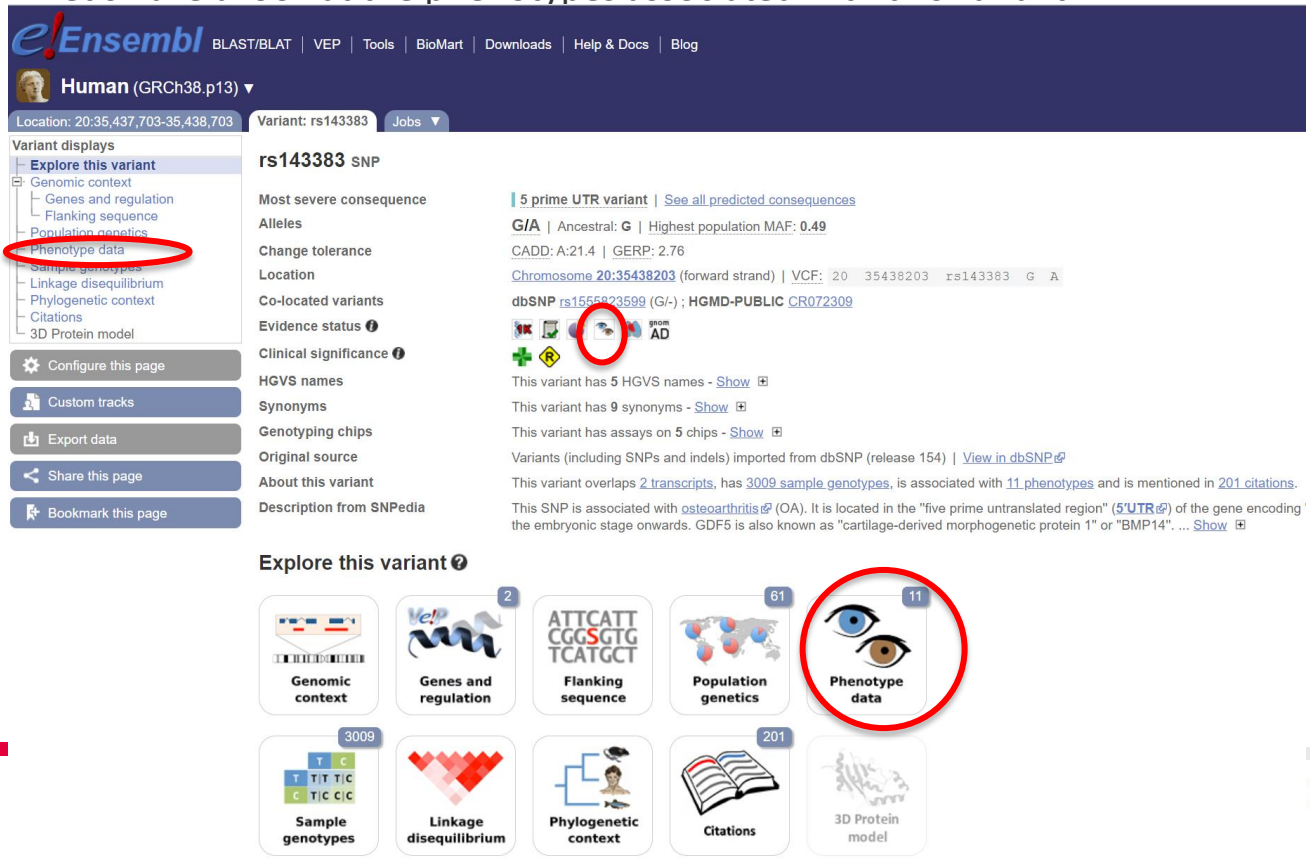
Show All entries — Show/hide columns

Population	Allele: frequency (count)	
gnomADg:ALL	G: 0.522 (79301)	A: 0.477 (72469)
gnomADg:afr	G: 0.879 (36292)	A: 0.121 (5006)
gnomADg:ami	G: 0.401 (365)	A: 0.599 (545)
gnomADg:amr	G: 0.376 (5729)	A: 0.624 (9491)
gnomADg:asj	G: 0.455 (1576)	A: 0.545 (1890)
gnomADg:eas	G: 0.283 (1462)	A: 0.717 (3710)
gnomADg:fin	G: 0.415 (4371)	A: 0.585 (6163)
gnomADg:mid	G: 0.494 (156)	A: 0.506 (160)
gnomADg:nfe	G: 0.377 (25640)	A: 0.623 (42306)
gnomADg:oth	G: 0.485 (1010)	A: 0.515 (1072)
gnomADg:sas	G: 0.559 (2700)	A: 0.441 (2126)

Jump to: 1000 Genomes Project Phase 3 (32) | gnomAD genomes v3.1.2 (11) | NCBI ALFA (12) | TOPMed (1) | Gambian Genome Variation Project (5)

Ensembl … Exploring variants… Phenotype/Disease data

- Let's have a look at the phenotypes associated with this variant.



Ensembl … Exploring variants… Phenotype/Disease data

- This variant is associated with 11 phenotypes/diseases.

Phenotype Data

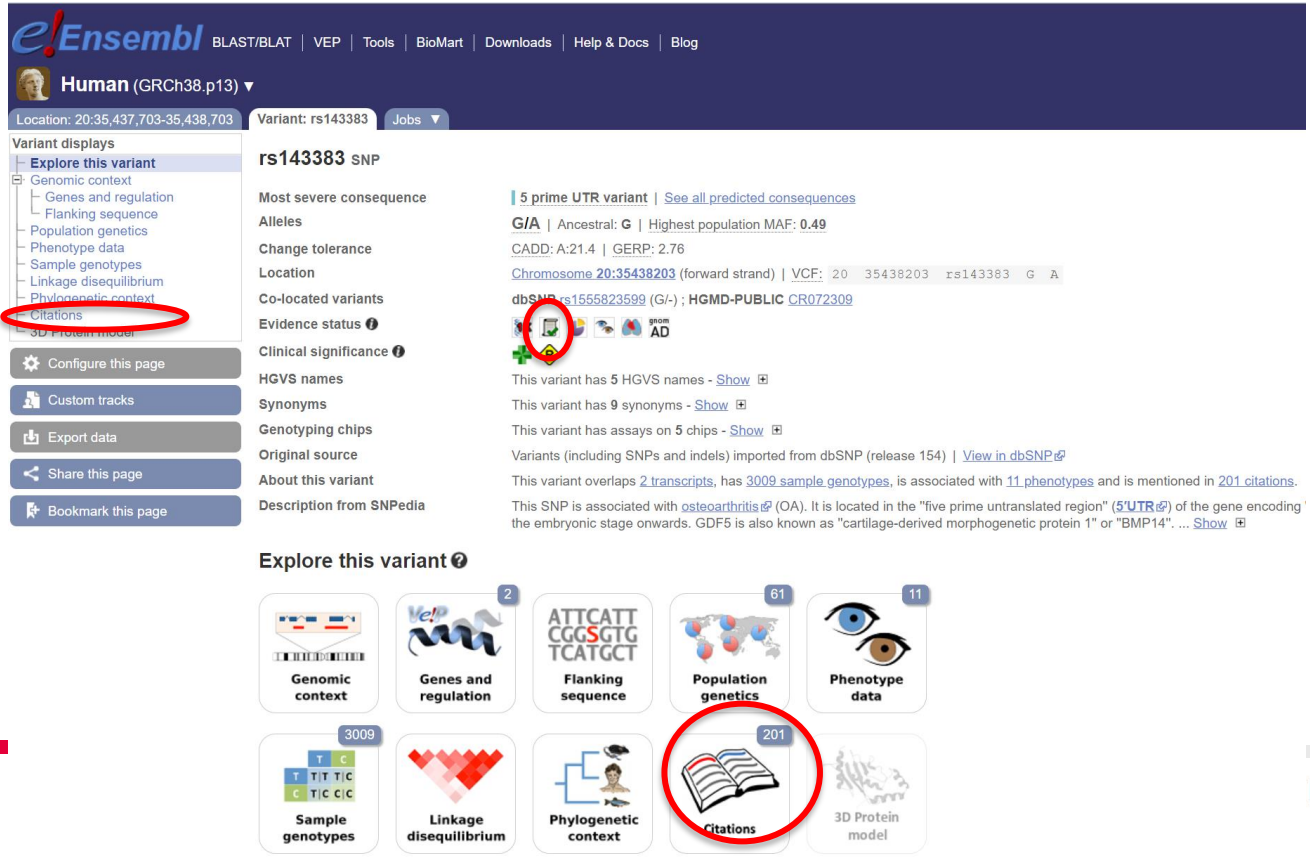
Significant association(s)

Show All entries | Show/hide columns | Filter

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	Supporting evidence	External reference	Clinical significance	Reported gene(s)	Associated allele	Statistics
ACROMESOMELIC DYSPLASIA, HUNTER-THOMPSON TYPE	ClinVar [Illumina Clinical Se...]	-	Orphanet:93437, Orphanet:968	-	-		LOC109461476, GDF5	-	-
Brachydactyly	ClinVar [Illumina Clinical Se...]	Brachydactyly	HP:0001156, Orphanet:294937	-	-		LOC109461476, GDF5	-	-
ClinVar: phenotype not specified	ClinVar [Invitae]	-	-	-	-		LOC109461476, GDF5	-	-
FIBULAR HYPOPLASIA AND COMPLEX BRACHYDACTYLY	ClinVar [Illumina Clinical Se...]	-	Orphanet:2639	-	-		LOC109461476, GDF5	-	-
Grebe syndrome	ClinVar [Illumina Clinical Se...]	-	Orphanet:2098	-	-		LOC109461476, GDF5	-	-
Height	NHGRI-EBI GWAS catalog	body height	EFO:0004339	-	PMID:31562340	-	GDF5	-	**p-value:** $9.00e^{-78}$ **beta coefficient:** 0.06367727 unit increase
Height	GIANT	body height	EFO:0004339	-	-	-	-	G	**p-value:** $2.81e^{-35}$
Knee osteoarthritis	NHGRI-EBI GWAS catalog	Knee osteoarthritis, osteoarthritis, knee	EFO:0004616, HP:0005086	-	PMID:30374069	-	-	T	**p-value:** $8.00e^{-18}$ **odds ratio:** 1.09287
Knee osteoarthritis	NHGRI-EBI GWAS catalog	Knee osteoarthritis, osteoarthritis, knee	EFO:0004616, HP:0005086	-	PMID:30664745	-	GDF5	-	**p-value:** $5.00e^{-7}$ **odds ratio:** 1.0495905
MULTIPLE SYNOSTOSES SYNDROME 2	ClinVar [Illumina Clinical Se...]	-	Orphanet:3237	-	-		LOC109461476, GDF5	-	-
Osteoarthritis of hip	ClinVar [OMIM]	Hip osteoarthritis	HP:0008843	PMID:17384641	MIM:601146		LOC109461476, GDF5	-	-

Ensembl … Exploring variants… Citations

➢ Let's have a look at the citations of the variant.



Ensembl … Exploring variants … Citations

- This variant is mentioned in 201 citations.

Citations

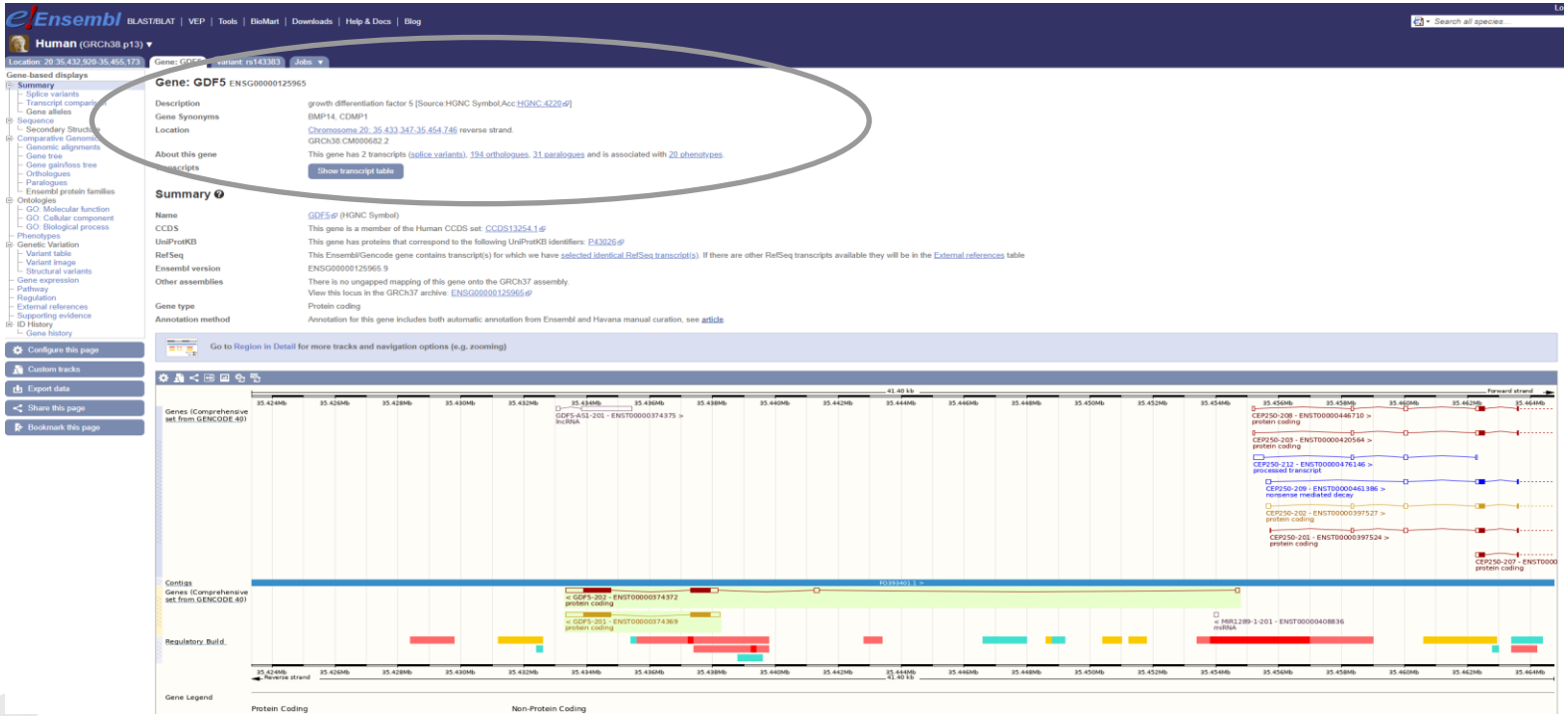
rs143383 is mentioned in the following publications

Show All entries | Show/hide columns | Filter

Year	PMID	Phenotype	Title	Author(s)	Full text	Citation source
2021	33235627		Identification of key genes in osteoarthritis using bioinformatics, principal component analysis and meta-analysis.	Sun X, Duan H, Xiao L, Yao S, et al	PMC7678638	EPMC
2021	34206824		Current Evidence about Developmental Dysplasia of the Hip in Pregnancy.	Simionescu AA, Cirstoiu MM, Cirstoiu C, Stanescu AMA, et al	PMC8305660	EPMC
2021	34210342		The effect of common variants in GDF5 gene on the susceptibility to chronic postsurgical pain.	Yan S, Nie H, Bu G, Yuan W, et al	PMC8247225	EPMC
2021	33817978		The Mechanisms and Functions of GDF-5 in Intervertebral Disc Degeneration.	Guo S, Cui L, Xiao C, Wang C, et al	PMC8126946	EPMC
2021	33522652		Growth differentiation factor 5 in cartilage and osteoarthritis: A possible therapeutic candidate.	Sun K, Guo J, Yao X, Guo Z, et al	PMC7941218	EPMC
2021	34203285		Genetic Study of IL6, GDF5 and PAPPA2 in Association with Developmental Dysplasia of the Hip.	Harsanyi S, Zamborsky R, Krajciova L, Kokavec M, et al	PMC8303839	EPMC
2021	33973562		A Case-Control Study of Major genetic pre-disposition risk alleles in developing DDD in the Northeast US population: Effects of Gene-Gene Interactions.	Valtchinov VI, Zhai BK, Hida T, Lacson R, et al		EPMC
2021	33861510		Frequency of Growth Differentiation Factor 5 rs143383 and asporin D-repeat polymorphisms in patients with hand and knee osteoarthritis in Kurdistan province, Iran.	Moghimi N, Nasseri S, Ghafouri F, Jalili A.		EPMC
2020	31923126		A Meta-analysis Assessing the Association Between COL11A1 and GDF5 Genetic Variants and Intervertebral Disc Degeneration Susceptibility.	Wu F, Huang X, Zhang Z, Shao Z.		dbSNP
2020	32103374		A novel variant near LSP1P3 is associated with knee osteoarthritis in the Chinese population.	Li Y, Liu F, Xu X, Zhang H, et al		EPMC,dbSNP
2020	31932746		Regulation of Gdf5 expression in joint remodelling, repair and osteoarthritis.	Kania K, Colella F, Riemen AHK, Wang H, et al	PMC6957535	EPMC
2020	32928309		Association between growth differentiation factor 5 rs143383 genetic polymorphism and the risk of knee osteoarthritis among Caucasian but not Asian: a meta-analysis.	Peng L, Jin S, Lu J, Ouyang C, et al	PMC7488690	EPMC
2020	32244273		Developmental Dysplasia of the Hip: A Review of Etiopathogenesis, Risk Factors, and Genetic Aspects.	Harsanyi S, Zamborsky R, Krajciova L, Kokavec M, et al	PMC7230892	EPMC
2020	32967120		Can Implementation of Genetics and Pharmacogenomics Improve Treatment of Chronic Low Back Pain?	Suntsov V, Jovanovic F, Knezevic E, Candido KD, et al	PMC7558486	EPMC
2020	32144293		Generation and characterization of human induced pluripotent stem cells (iPSCs) from hand osteoarthritis patient-derived fibroblasts.	Castro-ViÃ±uelas R, Sanjurjo-RodrÃguez C, PiÃ±eiro-Ramil M,	PMC7060311	EPMC

Ensembl … Exploring genes

- Let's have a look at *GDF5* (growth differentiation factor 5).



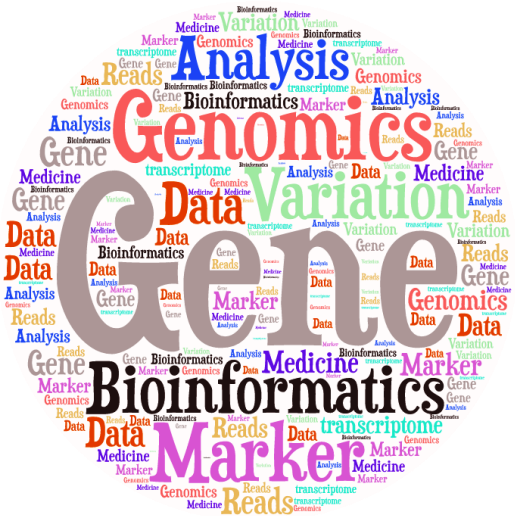
Ensembl … Exploring variants … Exercise 2

Explore rs2075650 (build 37).

1. What type of variants is?
2. Where is this variant located (chr - position - strand)?
3. Are there big differences in allele frequencies between populations?
4. What is the ancestral allele frequency in the Japanese in Tokyo (JPT) population from the 1000G set?
5. What is the least frequent genotype for this variant in the KHV population from the 1000G set?
6. In how many transcripts is this variant found?
7. Are all transcripts from the same gene?
8. How many and which consequence type are predicted for this variant?
9. How many phenotypes are associated with this variant?
10. How many phenotypes are associated with the gene that this variant is located in?
11. How many publications refer to rs2075650?

Publicly available bioinformatics resources

- Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database

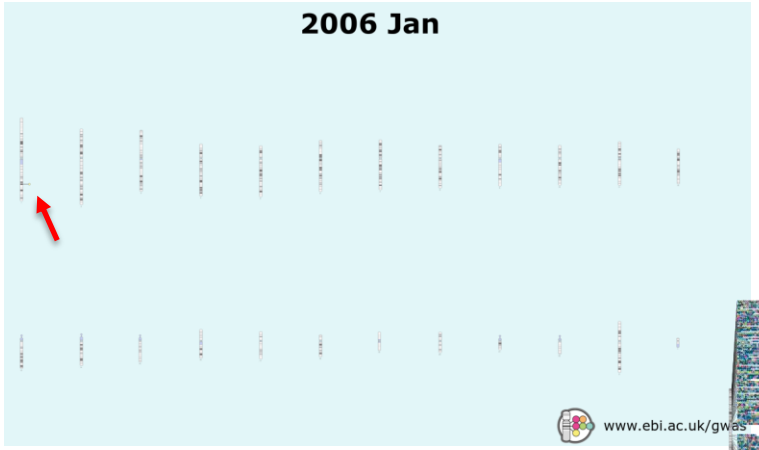


Human Traits & Diseases Databases

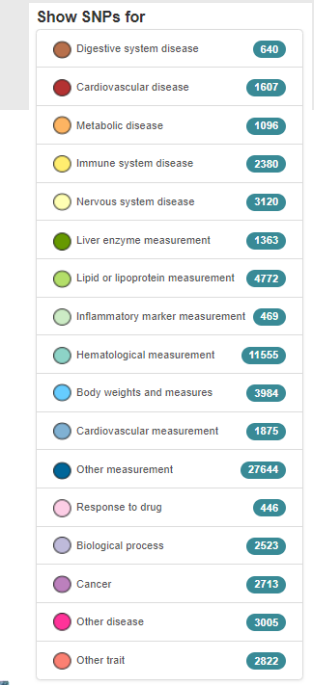
- GWAS catalog (https://www.ebi.ac.uk/gwas/)
- OMIM (http://omim.org/)
- Orphanet (http://www.orpha.net)

Human Traits & Diseases Databases
GWAS catalog

- "GWAS catalog" is a catalog of published/unpublished genome-wide association studies.



2022



Human Traits & Diseases Databases
GWAS catalog

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

- Let's have a look at a specific variant, rs143383.

Search results for *rs143383*

V rs143383

Location: 20:35438203 **Cytogenetic region:**20q11.22 **Most severe consequence:** 5 prime utr variant **Mapped gene(s):** GDF5

Associations 6 Studies 6

G GDF5

Description: growth differentiation factor 5
Location: 20:35433347-35454746 **Cytogenetic region:** 20q11.22 **Biotype:** protein coding

Associations 92 Studies 67

Human Traits & Diseases Databases
GWAS catalog

rs143383

Location: 20:35438203 **Cytogenetic region:**20q11.22 **Most severe consequence:** 5 prime utr variant **Mapped gene(s):** GDF5

Associations 6 Studies 6

Associations 6

Search

Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background trait(s)	Study accession	Location
rs143383-G	2×10^{-16}		0.37	-	8.25 z score increase	-	GDF5	Vertex-wise cortical surface area	cortical surface area measurement	-	GCST90095130	20:35438203
rs143383-?	9×10^{-78}		NR	-	0.06367727 unit increase	[0.057-0.07]	GDF5	Height	body height	-	GCST008839	20:35438203
rs143383-?	1×10^{-9}		NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-G	1×10^{-17}		0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-T	8×10^{-18}		0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-?	5×10^{-7}		NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203

Studies 6

Search

First author	Study accession	Publication date	Journal	Title	Reported trait	Trait(s)	Background trait(s)	Discovery sample number and ancestry	Replication sample number and ancestry	Association count	Summary statistics
Styrkarsdottir U	GCST006925	2018-10-29	Nat Genet	Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis.	Knee osteoarthritis	osteoarthritis, knee	-	• 586030 European	-	7	NA
Tachmazidou I	GCST007090	2019-01-21	Nat Genet	Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data.	Knee osteoarthritis	osteoarthritis, knee	-	• 403124 European	-	13	FTP Download or API access
Lin E	GCST90090967	2021-07-16	Hum Mol Genet	Genome-wide association study in the Taiwan biobank identifies four novel genes for human height: NABP2, RASA2, RNF41 and SLC39A5.	Height	body height	-	• 14571 East Asian	• 20506 East Asian	31	NA
Akiyama M	GCST008839	2019-09-27	Nat Commun	Characterizing rare and low-frequency height-associated variants in the Japanese population.	Height	body height	-	• 159095 East Asian	• 32692 East Asian	608	NA
van der Meer D	GCST90095130	2021-12-15	Sci Adv	The genetic architecture of human cortical folding.	Vertex-wise cortical surface area	cortical surface area measurement	-	• 33748 European	-	659	FTP Download

Human Traits & Diseases Databases
GWAS catalog

GDF5

Description: growth differentiation factor 5

Location: 20:35433347-35454746 **Cytogenetic region:** 20q11.22 **Biotype:** protein coding

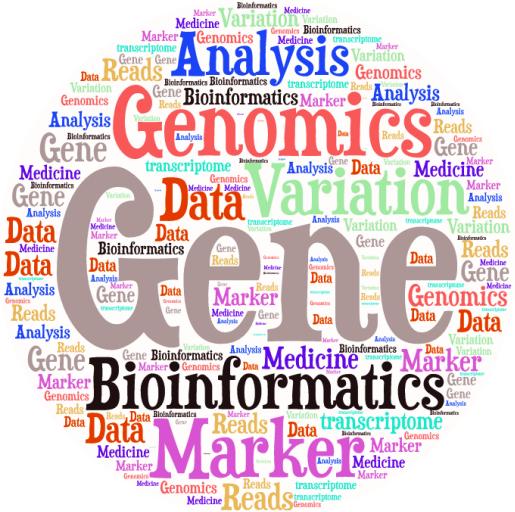
Available data: Associations 112 | Studies 78 | Traits 40

Human Traits & Diseases Databases …
Exercise 3

1. Try to find more information about rs2075650 and the gene that is located in GWAS catalogue and OMIM databases.

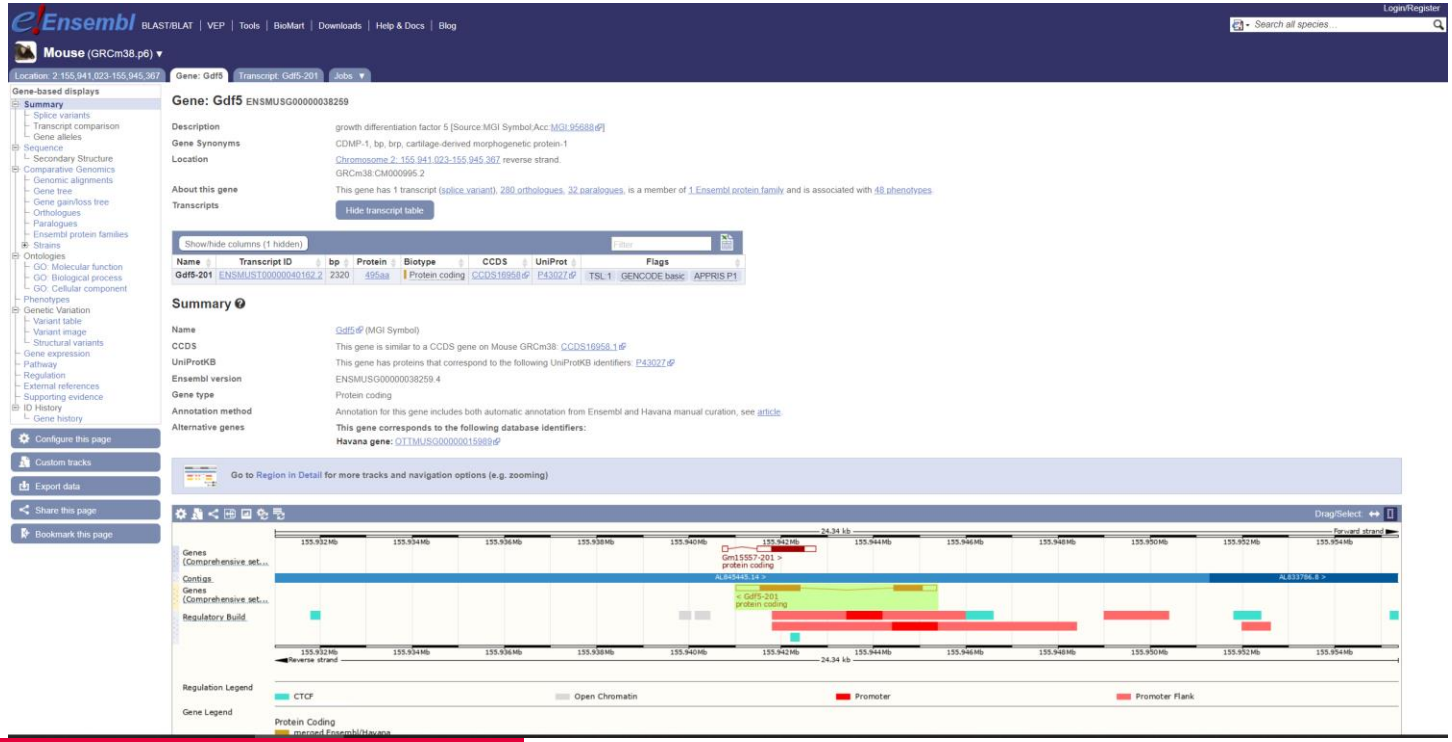
Publicly available bioinformatics resources

- Human Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database



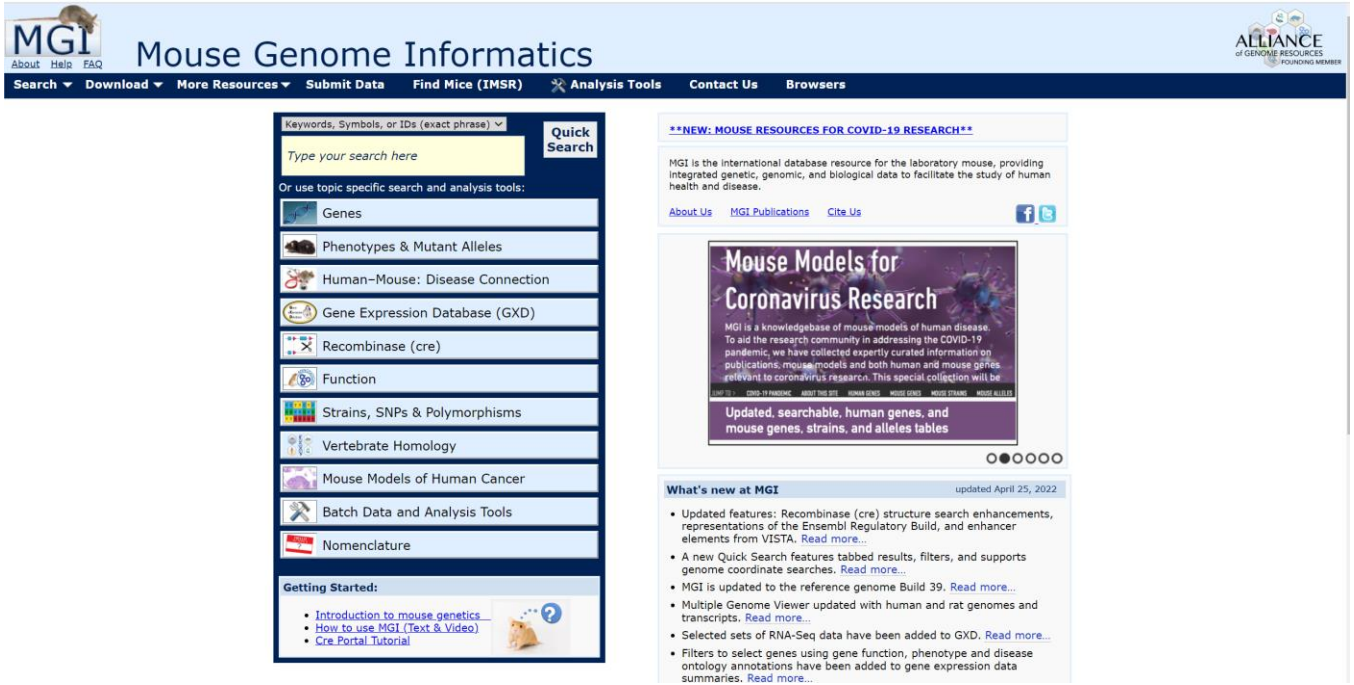
Databases of other Organisms

- Mouse: http://www.ensembl.org/Mus_musculus/Info/Index



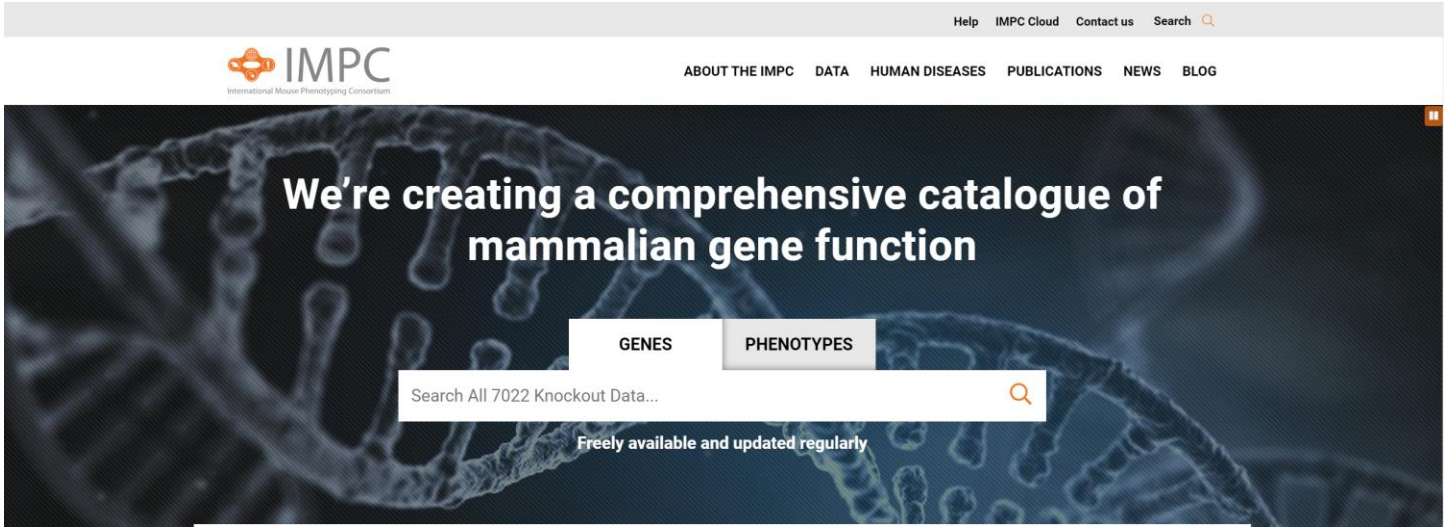
Databases of other Organisms

- Mouse: http://www.informatics.jax.org/



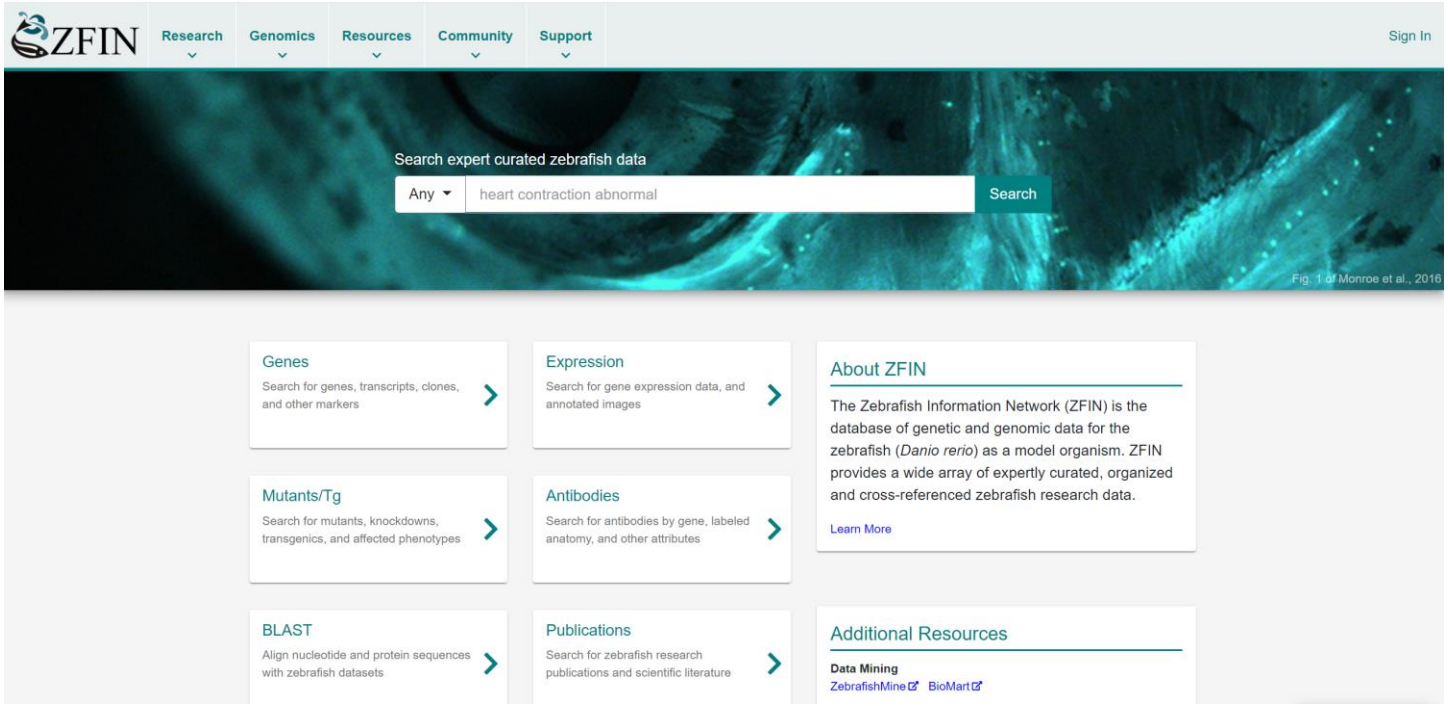
Databases of other Organisms

- Mouse: https://www.mousephenotype.org/



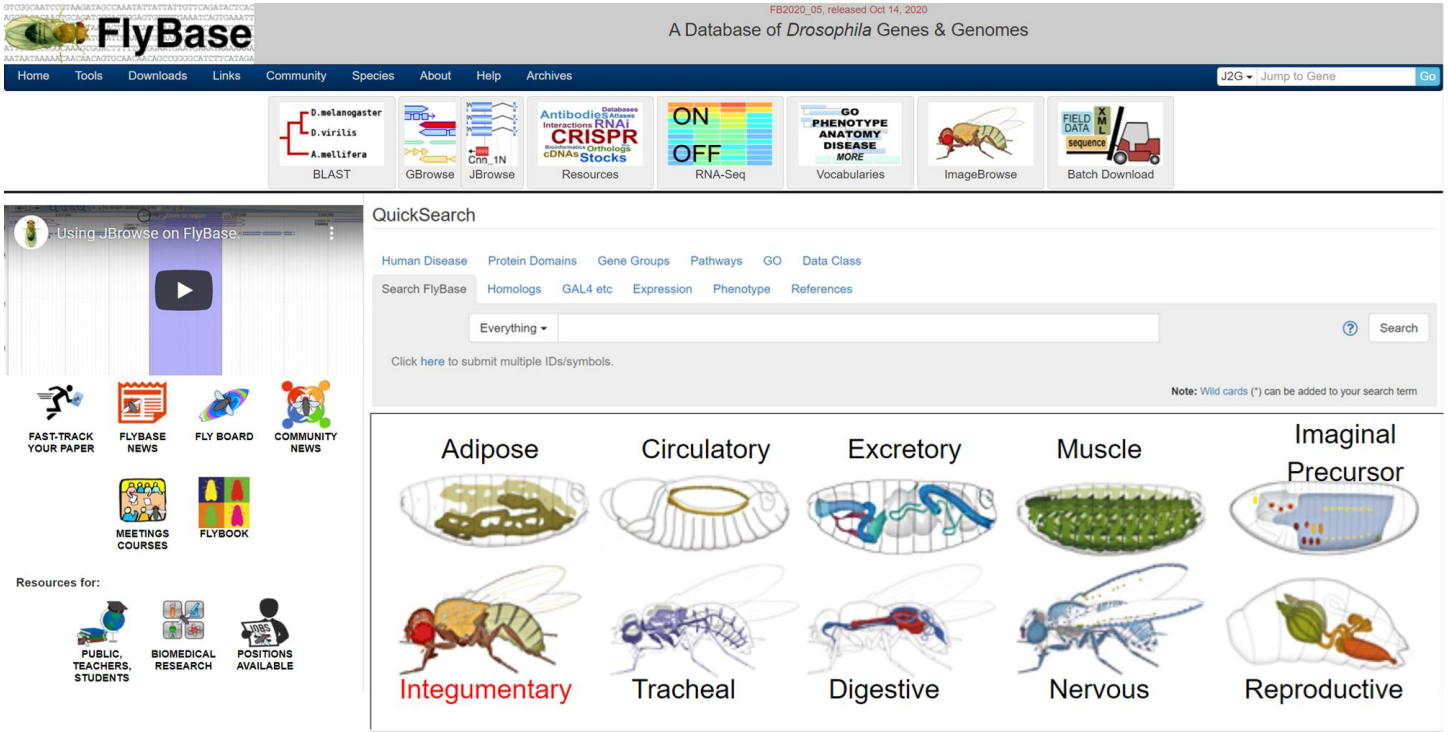
Databases of other Organisms

- Zebrafish: http://zfin.org/



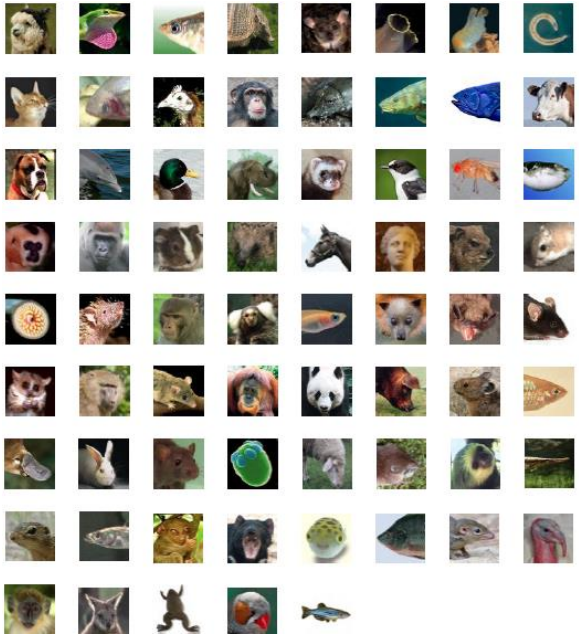
Databases of other Organisms

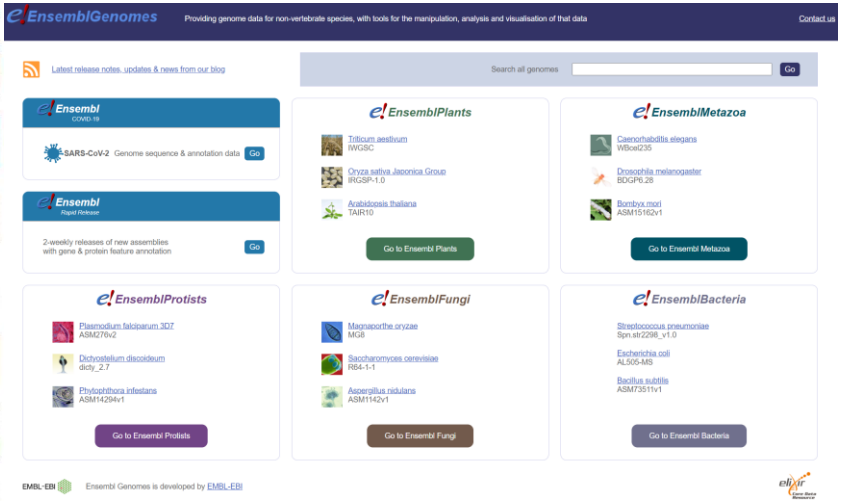
- Drosophila: http://flybase.org/



Databases of other Organisms

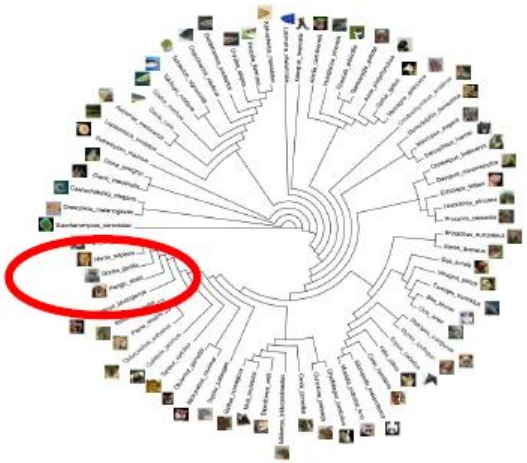
- Ensembl : >300 vertebrate species (https://www.ensembl.org/info/about/species.html) & >45000 genomes from non-vertebrate species (http://ensemblgenomes.org/)





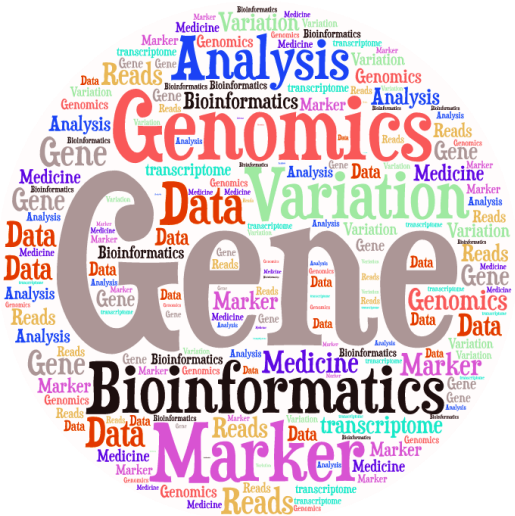
Databases of other Organisms …
Exercise 4

- Explore rs2075650 in the databases of other organisms.
  - *Hint: you can search by rsID or (mapped) gene id or phenotype.*



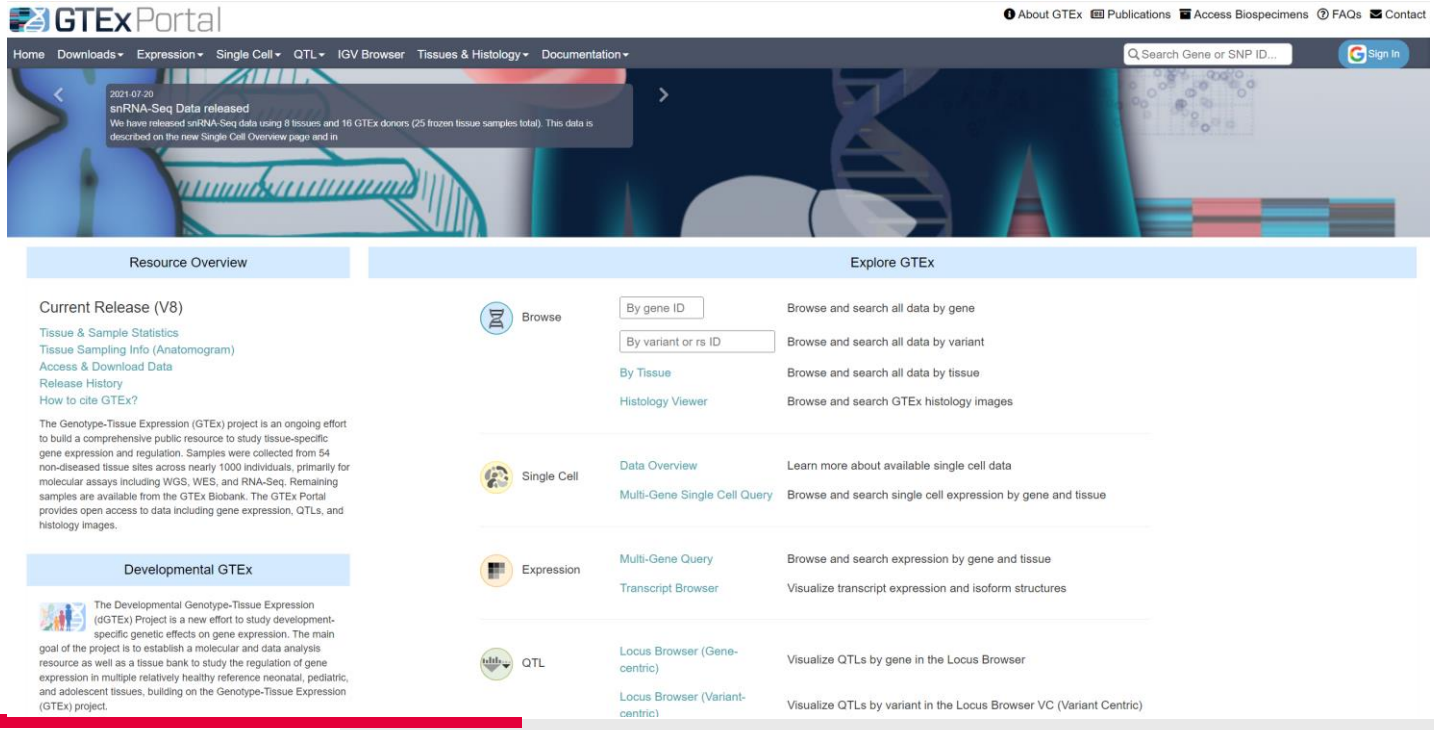
Publicly available bioinformatics resources

- Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database



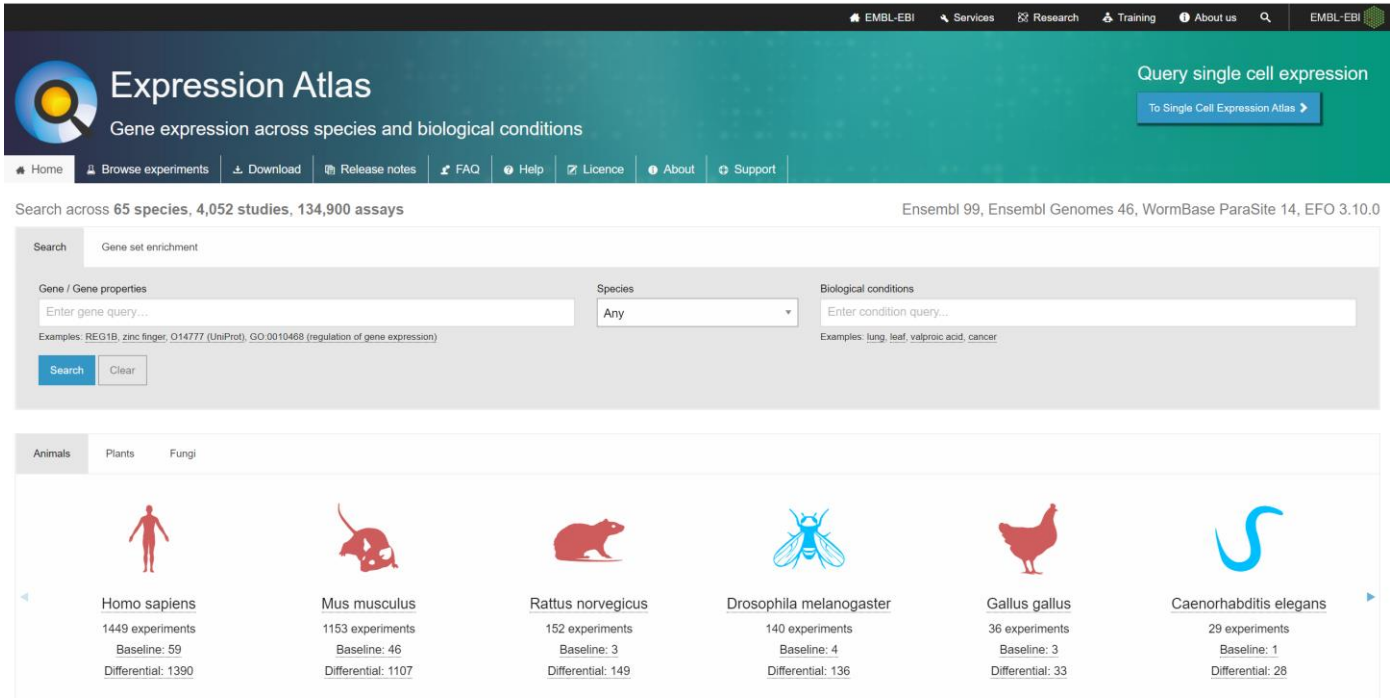
Gene Expression Databases

- Genotype-Tissue Expression (GTEx) : Correlations between genotype and tissue-specific gene expression (https://www.gtexportal.org/home/).



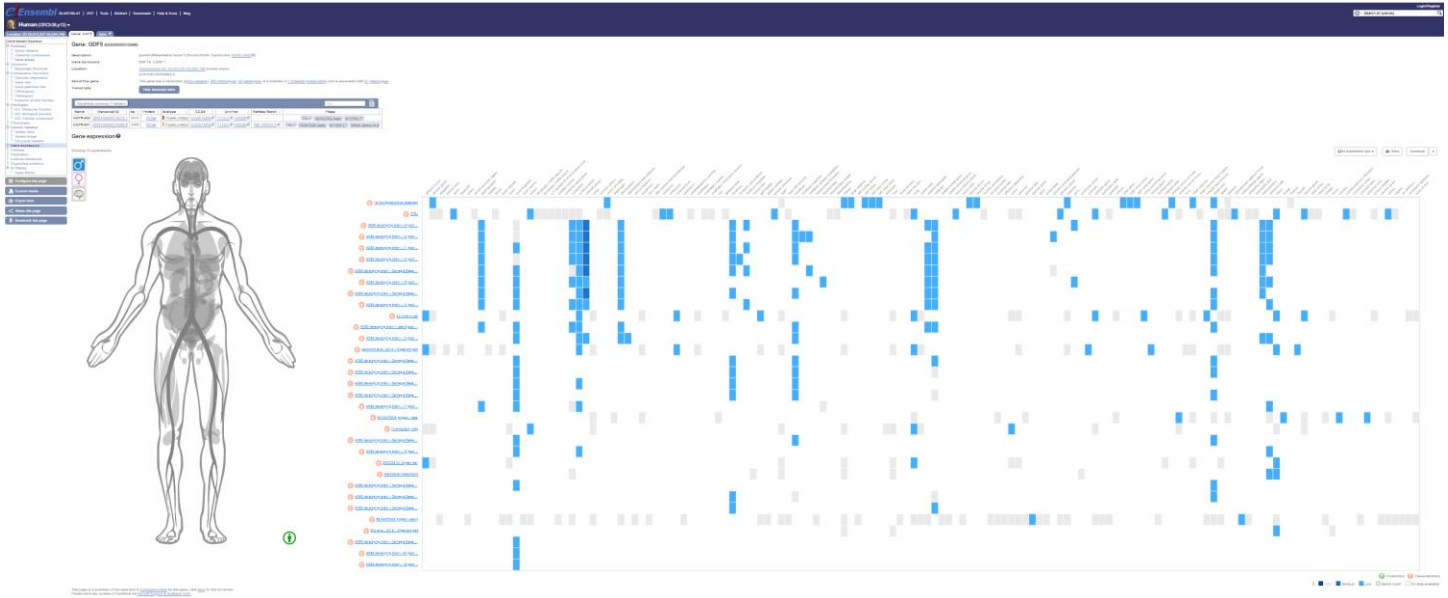
Gene Expression Databases

- Expression Atlas: Differential and Baseline Expression (http://www.ebi.ac.uk/gxa/home or through Ensembl).



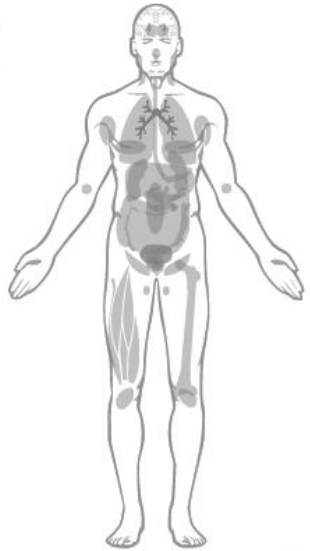
Gene Expression Databases

- Expression Atlas: Differential and Baseline Expression (http://www.ebi.ac.uk/gxa/home or through Ensembl).



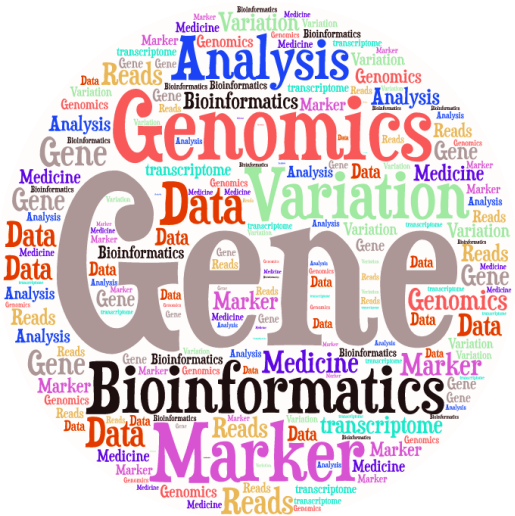
Gene Expression Databases …
Exercise 5

- Does rs2075650 affect the expression of the gene?
  - Hint: Use GTEx portal
- In which tissue *TOMM40* is expressed? Any connections with Alzheimer disease?
  - Hint: Use Expression Atlas



Publications Database

- Genome Browsers
- Nucleotide Sequence Databases
- Protein Sequence Databases
- Database Searching by Sequence Similarity
- Protein Domains: Databases and Search Tools
- Human Traits & Diseases Databases
- Phylogeny & Taxonomy
- Databases of other Organisms
- Gene Prediction
- Gene Expression Databases
- Gene Regulation
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database



Publications Database

- PubMed (https://www.ncbi.nlm.nih.gov/pubmed) comprises over 34 million citations.
- How do I search PubMed?
  - Identify the key concepts for your search.
  - Enter the terms* in the search box.
    - Keywords
    - Author name
    - Journal name
    - …..
  - Suggestions will display as you type your search terms.
  - Click Search.

Publications Database

- Anatomy of the summary results:

Publications Database … Exercise 6

- Try to find more information about rs2075650 and the gene that is located.
- Can you find a paper for which:
  - "Lopes" is one of the authors
  - Dealing with "osteoarthritis"
  - The name of the journal is "Lancet"

Pipeline

Variant:

Ensembl:

- ✓ position,
- ✓ consequences (transcript and regulatory region),
- ✓ 1kG frequencies,
- ✓ GWAS signals around the variant,
- ✓ regulatory marks around the variant
- ✓ nearby genes.

ExAC: allele frequencies

GTEx: associated gene.

Pubmed: papers where the variant is cited.

Gene:

Ensembl:

- ✓ Position
- ✓ Associated phenotypes
- ✓ Gene ontology annotations

Uniprot:

- ✓ Function, Tissue, Localization, Disease

OMIM:

- ✓ gene annotation
- ✓ associated diseases

GWAS catalog:

- ✓ Associated variants and association details.

Gene Expression Atlas:

- ✓ Tissues in which expression of the gene was observed

GTEx:

- ✓ List of variants that affects the expression of the gene.

Mouse Genomics Informatics:

- ✓ Mouse phenotype of the gene.

Dr Daniel Suveges
Senior Bioinformatician

Team144 pipeline

Variant details

General information:

rsID:	rs143384
Chromosome:start-end:	20:34025756-34025756
Allele string:	A/G
MAF:	0.438898
Consequence:	5_prime_UTR_variant
Variation type:	SNP
gerp (average gerp):	5.04 (3.497)
GWAVA score:	0.79
Ancestral allele:	G
Minor allele:	A
Synoyms:	rs431839, rs61433024, rs765106, rs3748435, rs17422934
Phenotypes (risk allele):	Height(A), Infant length(G), Height(G)

GWAS signals

GWAS signals within 1Mbp of the variant:

rsID	SNPID	Distance	Trait	p-value	PMID
rs17310467	chr20:33545616	-481kbp	Hemostatic factors and hematological phenotypes	4E-34	22443383
rs11906160	chr20:33565755	-461kbp	Anticoagulant levels	1E-6	22216198
rs1535466	chr20:33718706	-308kbp	Height	1.8E-29	25282103
rs2295888	chr20:33722863	-303kbp	Prothrombin time	5E-13	22703881
rs6120849	chr20:33730387	-296kbp	Protein C levels	7E-37	20802025
rs6088735	chr20:33745676	-281kbp	Hemostatic factors and hematological phenotypes	4E-34	22443383
rs6060278	chr20:33753262	-273kbp	Hemostatic factors and hematological phenotypes	4E-34	22443383
rs867186	chr20:33764554	-262kbp	Anticoagulant levels	4E-9	22216198
rs867186	chr20:33764554	-262kbp	Coagulation factor levels	6E-37	20231535
rs867186	chr20:33764554	-262kbp	D-dimer levels	4E-6	21502573
rs867186	chr20:33764554	-262kbp	Hemostatic factors and hematological phenotypes	2E-6	22443383
rs867186	chr20:33764554	-262kbp	Protein C levels	1E-64	25376901
rs11167260	chr20:33775200	-251kbp	Amyotrophic lateral sclerosis	4E-6	24529757

Dr Daniel Suveges
Senior Bioinformatician

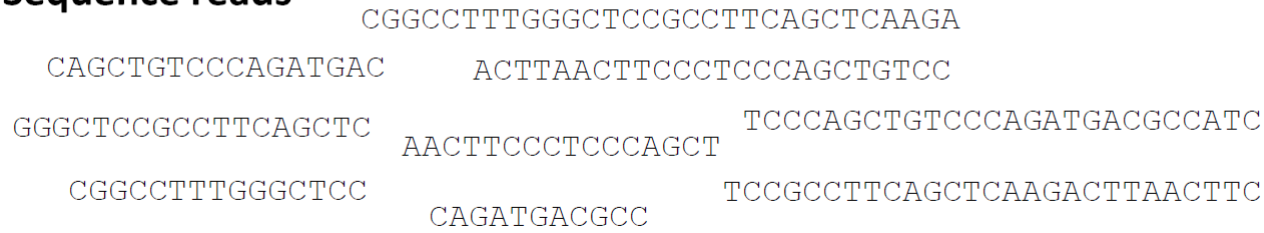
HelmholtzZentrum münchen
German Research Center for Environmental Health

Thank you!

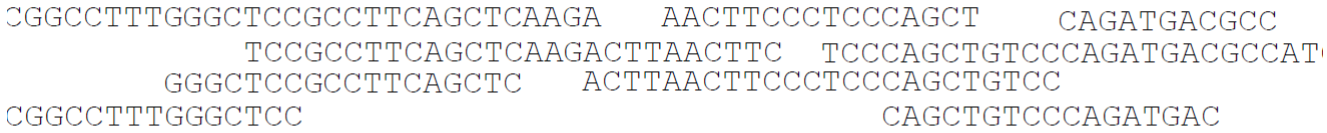
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

What is a genome assembly?

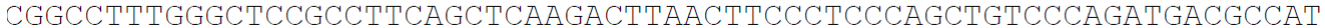
Sequence reads



Match up overlaps



Genome assembly



Ensembl … the front page … Exercise 1

1. Go to the species homepage for **Giant Panda**. What is the name of the genome assembly for Panda?
   - ASM200744v2, INSDC Assembly GCA_002007445.2, Apr 2020
2. How long is the Panda genome (in bp)? How many coding genes have been annotated?
   - 2,444,060,653 & 20,857
3. Repeat for **Human**.

- GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28, Dec 2013
- 3,096,649,726
- 20,471 (incl 662 readthrough)

Ensembl … Exploring variants … Exercise 2

Explore rs2075650 (build 37).

1. What type of variants is? SNP
2. Where is this variant located (chr - position - strand)? Chromosome 19:44892362 (forward strand)
3. Are there big differences in allele frequencies between populations? Not really. Max 8% in some sub-African pop and 10% in som AMR sub-pop too.
4. What is the ancestral allele frequency in the Japanese in Tokyo (JPT) population from the 1000G set? 12%
5. What is the least frequent genotype for this variant in the KHV population from the 1000G set? G|G
6. In how many transcripts is this variant found? 5
7. Are all transcripts from the same gene? Yes, TOMM40
8. How many and which consequence type are predicted for this variant? 1, Intron
9. How many phenotypes are associated with this variant? 67
10. How many phenotypes are associated with the gene that this variant is located in? 1
11. How many publications refer to rs2075650? 313