

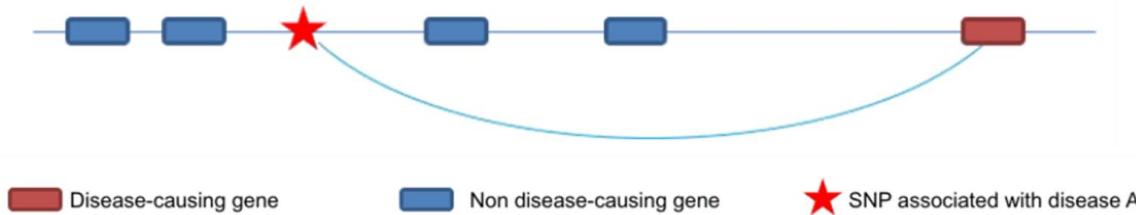
# Public bioinformatics resources for GWAS interpretation

**HELMHOLTZ** RESEARCH FOR  
GRAND CHALLENGES

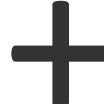
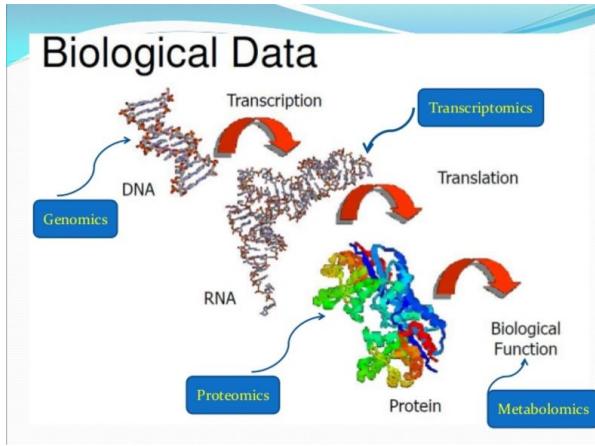
Dr. Konstantinos Hatzikotoulas  
Institute of Translational Genomics

# 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

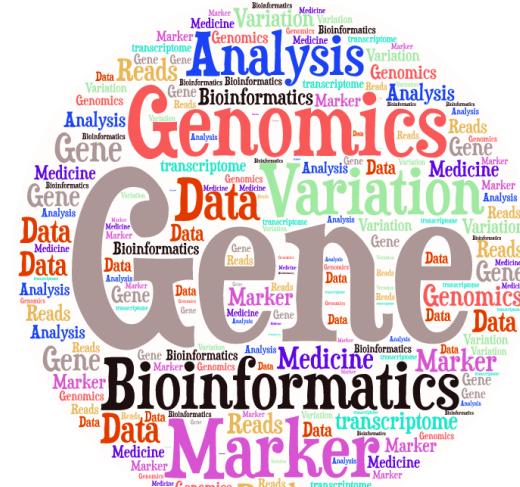
# Goal of Bioinformatics

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



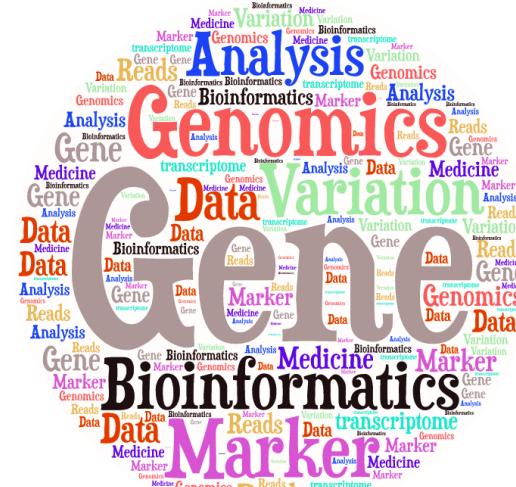
# Publicly available bioinformatics resources

- 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



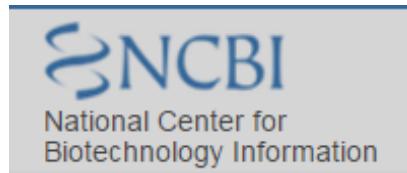
# Publicly available bioinformatics resources

- 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



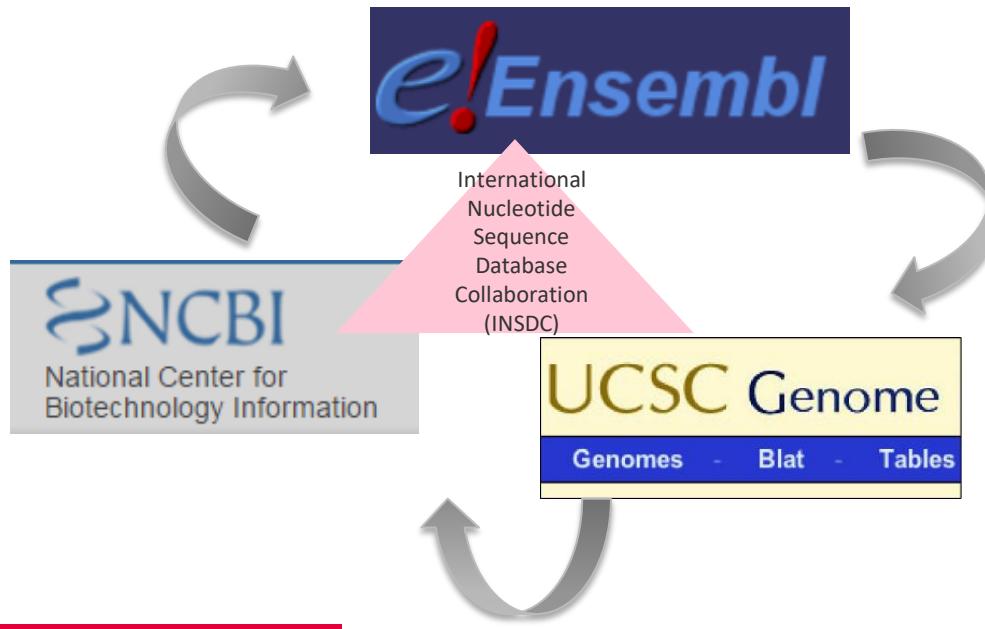
## Genome Browsers

- 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.

The screenshot shows the Ensembl homepage with several key features highlighted:

- Link back to homepage**: A button in the top left corner.
- Ensembl tools**: A button in the top center.
- Blue bar remains visible on every page**: A speech bubble pointing to the top navigation bar.
- Search**: A search bar at the top right with a "Search" button and a "Login/Register" link.
- See the current release number and what's new**: A button on the right side.
- Ensembl Release 106 (Apr 2022)**: Information about the latest release.
- Search bar**: A search bar with dropdown options for "All species" and a "Go" button.
- Variant Effect Predictor >**: A section for analyzing variants.
- BioMart >**: A section for exporting custom datasets.
- BLAST/BLAT >**: A section for searching genomes.
- Tools**: A sidebar with links to "All tools" and specific sections like "All genomes" (with a dropdown for "Select a species"), "Favourite genomes" (listing Human, Mouse, Zebrafish, and Pig breeds), and "View full list of all species".
- Ensembl Rapid Release**: A section for new assemblies.
- Other news from our blog**: A section with a list of recent posts.
- Footer links**: A row of links including "Compare genes across species", "Find SNPs and other variants for my gene", "Gene expression in different tissues", "Retrieve gene sequence", "Find a Data Display", and "Use my own data in 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).

Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

Variant displays

Explore this variant

- Genomic context
  - Genes and regulation
  - Flanking sequence
  - Population genetics
  - Phenotype data
  - Sample genotypes
  - Linkage disequilibrium
  - Phylogenetic context
  - Citations
  - 3D Protein model
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

rs143383 SNP

Most severe consequence | See all predicted consequences

GIA | Ancestral: G | Highest population MAF: 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203

dbSNP rs1555823599 (G-) ; HGMD-PUBLIC CR072309

AD

This variant has 5 HGVS names - Show

This variant has 9 synonyms - Show

This variant has assays on 5 chips - Show

Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis (OA). It is located in the "five prime untranslated region" (5'UTR) of the gene encoding the embryonic stage onwards. GDE5 is also known as "cartilage-derived morphogenic protein 1" or "BMP14". Show

http://www.ensembl.org/info/genome/variation/prediction/predicted\_d ata.html

### Variant information

Explore this variant

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

3009 201

Variation icons. These go to the same places as the links on the left

# Ensembl ... Exploring variants ...

## Variant types

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

 Ensembl

Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

Variant displays

- Explore this variant **rs143383 SNP** (highlighted with a red arrow)
- Genomic context
  - Genes and regulation
  - Flanking sequence
  - Population genetics
  - Phenotype data
  - Sample genotypes
  - Linkage disequilibrium
  - Phylogenetic context
  - Citations
  - 3D Protein model
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

Most severe consequence

5 prime UTR variant | See all predicted consequences

G/A | Ancestral: G | Highest population MAF: 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

dbSNP rs1555823599 (G-) ; HGMD-PUBLIC CR072309



This variant has 5 HGVS names - [Show](#)

This variant has 9 synonyms - [Show](#)

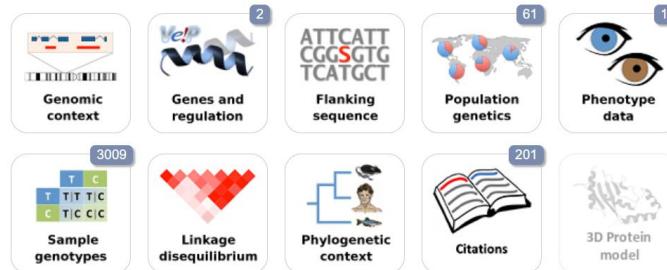
This variant has assays on 5 chips - [Show](#)

Variants (including SNPs and indels) imported from dbSNP (release 154) | [View in dbSNP](#)

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis (OA). It is located in the "five prime untranslated region" (5'UTR) of the gene encoding GDF5. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14". ... [Show](#)

### Explore this variant



# Ensembl ... Exploring variants ...

## Variant types

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

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

### 2. Large scale in chromosomal structure

**CNV**

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

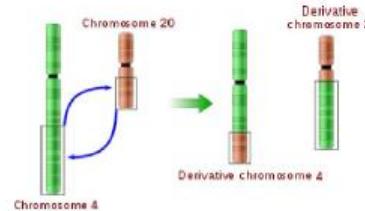
Reference:



"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.

 Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

**Variant displays**

- Explore this variant
- Genomic context
  - Genes and regulation
  - Flanking sequence
  - Population genetics
  - Phenotype data
  - Sample genotypes
  - Linkage disequilibrium
  - Phylogenetic context
  - Citations
  - 3D Protein model
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

**rs143383 SNP**

Most severe consequence (circled in red)

5 prime UTR variant | See all predicted consequences

G/A | Ancestral: G | Highest population MAF: 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

dbSNP rs1555823599 (G-) ; HGMD-PUBLIC CR072309

This variant has 5 HGVS names - [Show](#)

This variant has 9 synonyms - [Show](#)

This variant has assays on 5 chips - [Show](#)

Variants (including SNPs and indels) imported from dbSNP (release 154) | [View in dbSNP](#)

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis (OA). It is located in the "five prime untranslated region" (5'UTR) of the gene encoding the embryonic stage onwards. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14" ... [Show](#)

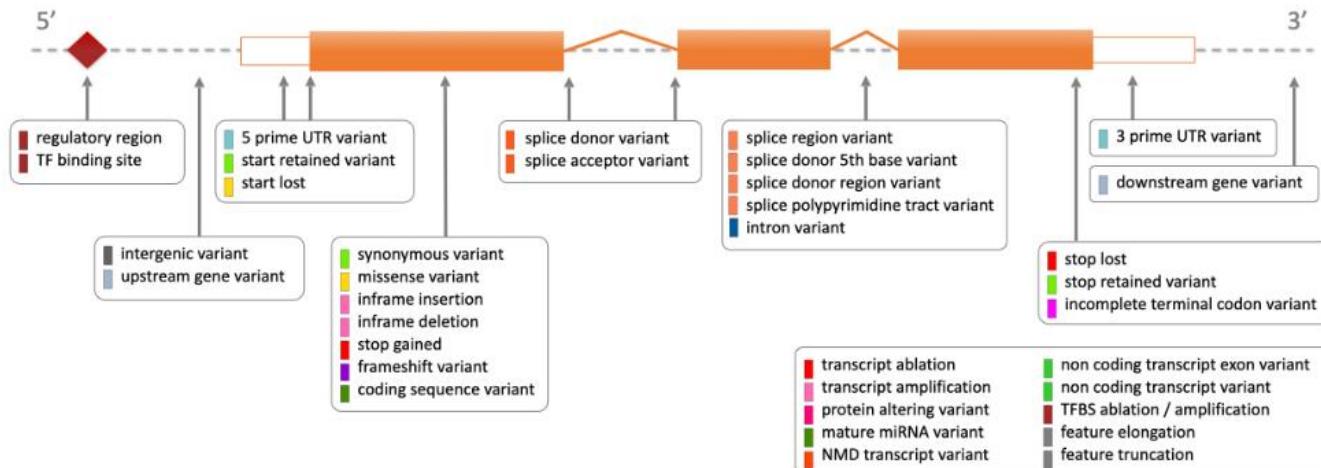
**Explore this variant**

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

3009 2 61 11 201

# 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](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#F	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO_0001574#F	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#F	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#F	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#F	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#F	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO_0002012#F	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO_0001889#F	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into the coding sequence	SO_0001821#F	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO_0001822#F	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#F	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO_0001818#F	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#F	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#F	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#F	Splice donor region variant	LOW
splice_polyypyrimidine_tract_variant	A sequence variant that falls in the polyypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO_0002169#F	Splice polyypyrimidine 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#F	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#F	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#F	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO_0001819#F	Synonymous variant	LOW
coding_sequence_variant	A sequence variant that changes the coding sequence	SO_0001580#F	Coding sequence variant	MODIFIER
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO_0001620#F	Mature miRNA variant	MODIFIER
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO_0001623#F	5 prime UTR variant	MODIFIER
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO_0001624#F	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#F	Non coding transcript exon variant	MODIFIER
intron_variant	A transcript variant occurring within an intron	SO_0001627#F	Intron variant	MODIFIER
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO_0001621#F	NMD transcript variant	MODIFIER
non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO_0001619#F	Non coding transcript variant	MODIFIER
upstream_gene_variant	A sequence variant located 5' of a gene	SO_0001631#F	Upstream gene variant	MODIFIER
downstream_gene_variant	A sequence variant located 3' of a gene	SO_0001632#F	Downstream gene variant	MODIFIER
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO_0001895#F	TFBS ablation	MODIFIER
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	SO_0001892#F	TFBS amplification	MODIFIER
TF_binding_site_variant	A sequence variant located within a transcription factor binding site	SO_0001782#F	TF binding site variant	MODIFIER
regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	SO_0001894#F	Regulatory region ablation	MODERATE
regulatory_region_amplification	A feature amplification of a region containing a regulatory region	SO_0001891#F	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#F	Feature elongation	MODIFIER
regulatory_region_variant	A sequence variant located within a regulatory region	SO_0001566#F	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#F	Feature truncation	MODIFIER
intergenic_variant	A sequence variant located in the intergenic region, between genes	SO_0001628#F	Intergenic variant	MODIFIER

# Ensembl ... Exploring variants ...

## Calculated variant consequences

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

**Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

**Variant displays**

- Explore this variant
  - Genomic context
    - Genes and regulation
    - Flanking sequence
    - Population genetics
    - Phenotype data
    - Sample genotypes
    - Linkage disequilibrium
    - Phylogenetic context
    - Citations
    - 3D Protein model
  - Most severe consequence
  - Alleles
  - Change tolerance
  - Location
  - Co-located variants
  - Evidence status ⓘ
  - Clinical significance ⓘ
  - HGVIS names
  - Synonyms
  - Genotyping chips
  - Original source
  - About this variant
  - Description from SNPedia
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

**rs143383 SNP**

5 prime UTR variant [See all predicted consequences](#) (circled)

G/A | Ancestral: G | Highest population freq = 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

dbSNP rs1555823599 (G-) ; HGMD-PUBLIC CR072309

This variant has 5 HGVIS names - [Show](#) ⓘ

This variant has 9 synonyms - [Show](#) ⓘ

This variant has assays on 5 chips - [Show](#) ⓘ

Variants (including SNPs and indels) imported from dbSNP (release 154) | [View in dbSNP](#) ⓘ

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis ⓘ (OA). It is located in the "five prime untranslated region" (5'UTR ⓘ) of the gene encoding the embryonic stage onwards. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14" ... [Show](#) ⓘ

**Explore this variant ⓘ**

- Genomic context
- Genes and regulation (circled)
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes (3009)
- Linkage disequilibrium
- Phylogenetic context
- Citations (201)
- 3D Protein model

# Ensembl ... Exploring variants ... Calculated variant consequences

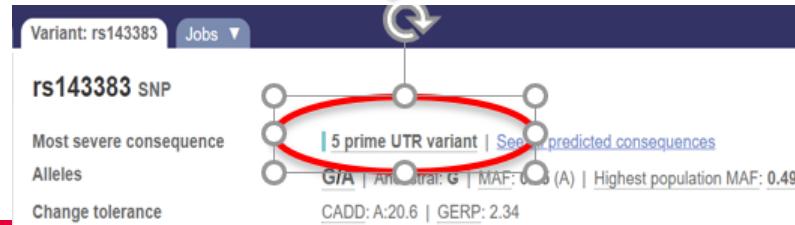
➤ This variant is found in 2 transcripts of the *GDF5* gene.

Genes and regulation ?

Gene and Transcript consequences

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

	<a href="#">coding_sequence_variant</a>	A sequence variant that changes the coding sequence	<a href="#">SO_0001580</a>	Coding sequence variant	MODIFIER
	<a href="#">mature_miRNA_variant</a>	A transcript variant located with the sequence of the mature miRNA	<a href="#">SO_0001620</a>	Mature miRNA variant	MODIFIER
	<a href="#">5_prime_UTR_variant</a>	A UTR variant of the 5' UTR	<a href="#">SO_0001623</a>	5 prime UTR variant	MODIFIER
	<a href="#">3_prime_UTR_variant</a>	A UTR variant of the 3' UTR	<a href="#">SO_0001624</a>	3 prime UTR variant	MODIFIER
	<a href="#">non_coding_transcript_exon_variant</a>	A sequence variant that changes non-coding exon sequence in a non-coding transcript	<a href="#">SO_0001792</a>	Non coding transcript exon variant	MODIFIER
	<a href="#">intron_variant</a>	A transcript variant occurring within an intron	<a href="#">SO_0001627</a>	Intron variant	MODIFIER
	<a href="#">NMD_transcript_variant</a>	A variant in a transcript that is the target of NMD	<a href="#">SO_0001621</a>	NMD transcript variant	MODIFIER
	<a href="#">non_coding_transcript_variant</a>	A transcript variant of a non coding RNA gene	<a href="#">SO_0001619</a>	Non coding transcript variant	MODIFIER



# Ensembl ... Exploring variants... Population genetics

➤ Let's have a look at population genetics.

**Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

**Variant displays**

- Explore this variant
- Genomic context
  - Genes and regulation
  - Flanking sequence
  - Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

**rs143383 SNP**

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

Alleles: G/A | Ancestral: G | Highest population MAF: 0.49

Change tolerance: CADD: A:21.4 | GERP: 2.76

Location: Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

SNP rs155582350 (rs143383): HGMD-PUBLIC CR072309

This variant has 5 HGVS names - Show

This variant has 9 synonyms - Show

This variant has assays on 5 chips - Show

Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis (OA). It is located in the "five prime untranslated region" (5'UTR) of the gene encoding the embryonic stage onwards. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14". ... Show

**Explore this variant**

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

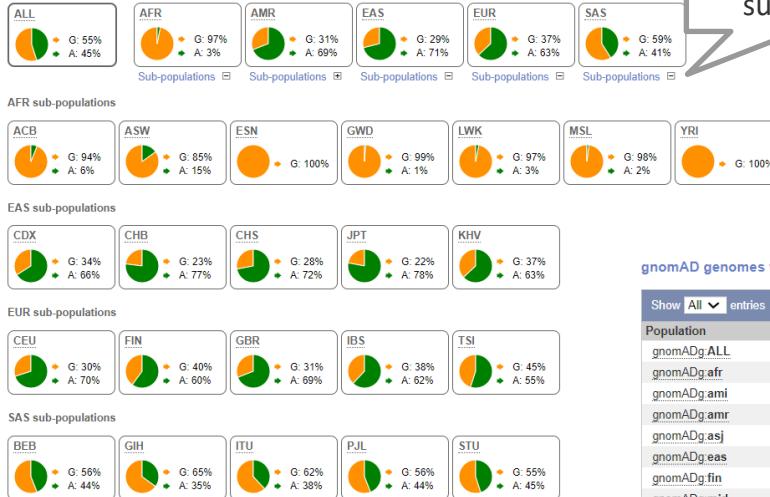
3009 201 61 11

# 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.

## Population genetics ↗

### 1000 Genomes Project Phase 3 allele frequencies



Expand  
subpopulations

Pie charts of  
allele frequencies

Tables of more  
detailed data

		G: 0.085 (43)	A: 0.512 (26)	A/G: 0.403 (203)
EAS	G	0.287 (289)	0.713 (119)	
CDX	G	0.344 (64)	0.656 (122)	A/G: 0.409 (38)
CHB	G	0.233 (48)	0.767 (158)	
CHS	G	0.276 (58)	0.724 (152)	A/G: 0.369 (38)
JPT	G	0.221 (46)	0.779 (162)	A/G: 0.367 (7)
KHV	G	0.369 (73)	0.631 (125)	A/G: 0.374 (37)
				A/G: 0.515 (51)

### gnomAD genomes v3.1.2 (11) ↗

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.

The screenshot shows the Ensembl Variant displays page for SNP rs143383. The URL is [https://www.ensembl.org/Homo\\_sapiens/Variation/Variant-display?var=rs143383](https://www.ensembl.org/Homo_sapiens/Variation/Variant-display?var=rs143383). The variant is located at GRCh38.p13, position 20:35,437,703-35,438,703. The main content area shows the variant details: rs143383 SNP, 5' prime UTR variant, G/A allele, Ancestral G, Highest population MAF: 0.49, CADD score: 21.4, GERP score: 2.76, Chromosome 20:35438203 (forward strand), VCF ID: 20 35438203 rs143383 G A, dbSNP ID: rs155823599 (G-), HGMD-PUBLIC ID: CR072309. Below this, there are sections for HGVS names (circled in red), Synonyms, Genotyping chips, Original source, About this variant, and Description from SNPedia. The "Explore this variant" sidebar on the left has a section for "Phenotype data" which is also circled in red. At the bottom, there is a grid of icons for "Explore this variant" including Genomic context, Genes and regulation, Flanking sequence, Population genetics, Phenotype data (circled in red), Sample genotypes, Linkage disequilibrium, Phylogenetic context, Citations, and 3D Protein model. The "Phenotype data" icon features an eye and a brain.

Ensembl Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

Variant displays

Explore this variant

- Genomic context
  - Genes and regulation
  - Flanking sequence
  - Population genetics
  - Phenotype data
  - Sample genotypes
  - Linkage disequilibrium
  - Phylogenetic context
  - Citations
  - 3D Protein model
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

rs143383 SNP

Most severe consequence

Alleles

Change tolerance

Location

Co-located variants

Evidence status ⓘ

Clinical significance ⓘ

HGVS names

Synonyms

Genotyping chips

Original source

About this variant

Description from SNPedia

5 prime UTR variant | See all predicted consequences

G/A | Ancestral: G | Highest population MAF: 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

dbSNP rs155823599 (G-) ; HGMD-PUBLIC CR072309

AD

This variant has 5 HGVS names - Show ⓘ

This variant has 9 synonyms - Show ⓘ

This variant has assays on 5 chips - Show ⓘ

Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP ⓘ

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis (OA). It is located in the "five prime untranslated region" (5'UTR) of the gene encoding GDF5. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14". ... Show ⓘ

Explore this variant ⓘ

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

# Ensembl ... Exploring variants... Phenotype/Disease data

- This variant is associated with 11 phenotypes/diseases.

## Phenotype Data

### Significant association(s)

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	Supporting evidence	External reference	Clinical significance	Reported gene(s)	Associated allele	Statistics
ACROMESOMELIC DYPLASIA, HUNTER-THOMPSON TYPE	ClinVar  [Illumina Clinical Se...] -	-	Orphanet:93437  , Orphanet:968 	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
Brachydactyly	ClinVar  [Illumina Clinical Se...] Brachydactyly	-	HP:0001156  , Orphanet:294937 	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
ClinVar: phenotype not specified	ClinVar  [Invitae]	-	-	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
FIBULAR HYPOPLASIA AND COMPLEX BRACHYDACTYLY	ClinVar  [Illumina Clinical Se...] -	-	Orphanet:2639 	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
Grebe syndrome	ClinVar  [Illumina Clinical Se...] -	-	Orphanet:2098 	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
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 	-	I	-	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 SYNOSTOSIS SYNDROME 2	ClinVar  [Illumina Clinical Se...] -	-	Orphanet:3237 	-	-	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-
Osteoarthritis of hip	ClinVar  [OMIM]	Hip osteoarthritis	HP:0008843 	PMID:17384641 	MIM:601146 	 ★★★★	LOC109461476, <a href="#">GDF5</a>	-	-

# Ensembl ... Exploring variants... Citations

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

The screenshot shows the Ensembl Variant displays page for SNP rs143383. The URL is [https://www.ensembl.org/Homo\\_sapiens/Variants/Variant-display?var=rs143383&db=core](https://www.ensembl.org/Homo_sapiens/Variants/Variant-display?var=rs143383&db=core). The variant is located at 20:35,437,703-35,438,703. The sidebar on the left has a red circle around the "Explore this variant" section, specifically the "Citations" link. The main content area shows the SNP details, including its consequence as a 5' prime UTR variant, its G/A allele, and its association with dbSNP and HGMD-PUBLIC. Below this, there are links to HGVS names, synonyms, genotyping assays, and clinical significance. At the bottom, there is a "Explore this variant" section with a red circle around the "Citations" icon, which shows 201 citations.

Ensembl Human (GRCh38.p13) ▾

Location: 20:35,437,703-35,438,703 Variant: rs143383 Jobs ▾

Variant displays

Explore this variant

- Genomic context
- Flanking sequence
- Population genetics
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

**rs143383 SNP**

Most severe consequence

Alleles

Change tolerance

Location

Co-located variants

Evidence status ⓘ

Clinical significance ⓘ

HGVS names

Synonyms

Genotyping chips

Original source

About this variant

Description from SNPedia

5 prime UTR variant | See all predicted consequences

G/A | Ancestral: G | Highest population MAF: 0.49

CADD: A:21.4 | GERP: 2.76

Chromosome 20:35438203 (forward strand) | VCF: 20 35438203 rs143383 G A

dbSNP rs1555823599 (G-) ; HGMD-PUBLIC CR072309

AD

This variant has 5 HGVS names - Show ⓘ

This variant has 9 synonyms - Show ⓘ

This variant has assays on 5 chips - Show ⓘ

Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP ⓘ

This variant overlaps 2 transcripts, has 3009 sample genotypes, is associated with 11 phenotypes and is mentioned in 201 citations.

This SNP is associated with osteoarthritis ⓘ (OA). It is located in the "five prime untranslated region" (5'UTR ⓘ) of the gene encoding the embryonic stage onwards. GDF5 is also known as "cartilage-derived morphogenetic protein 1" or "BMP14".... Show ⓘ

Explore this variant ⓘ

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

3009 201 61 11 2 11 201

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

# Ensembl ... Exploring variants ...

## Citations

- This variant is mentioned in 201 citations.

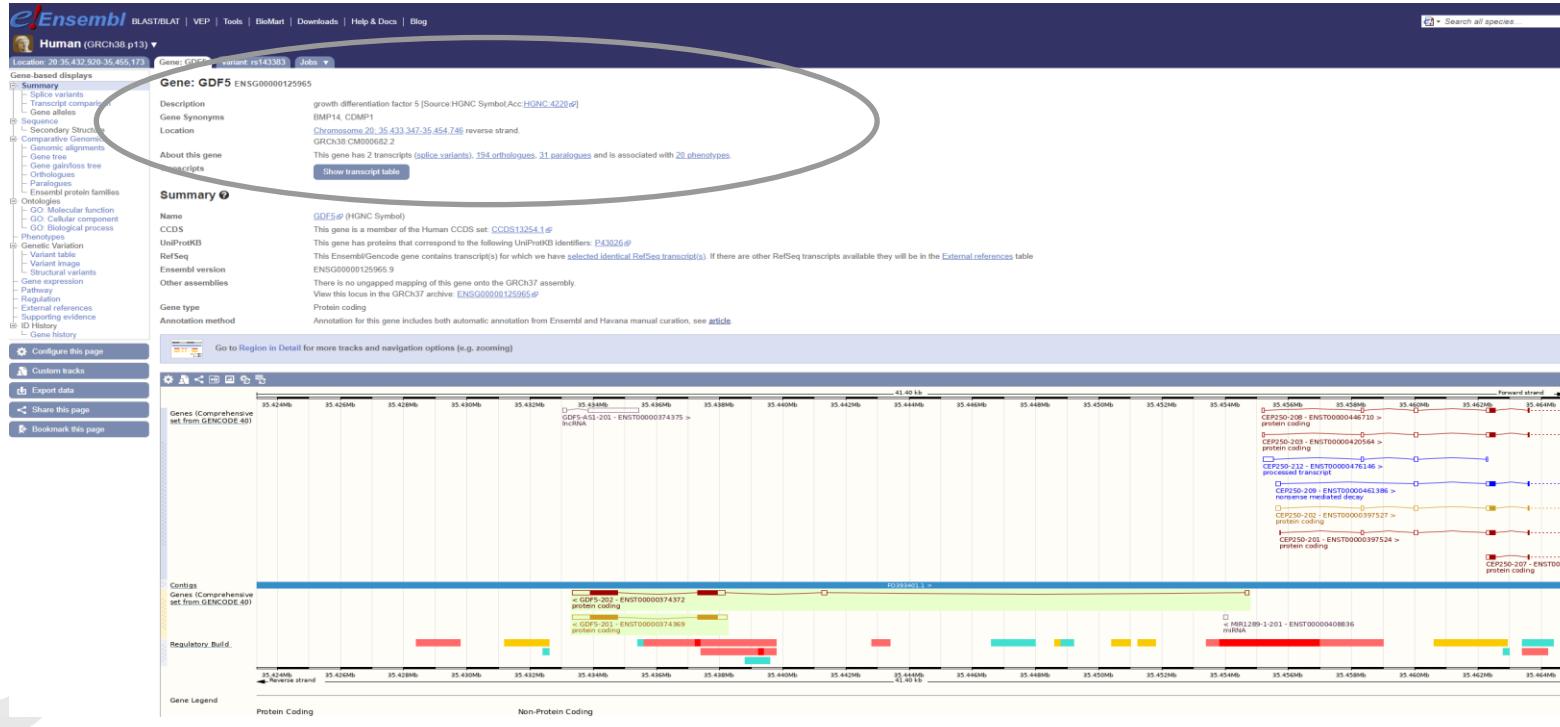
### Citations

rs143383 is mentioned in the following publications

Show	All	entries	Phenotype	Title	Author(s)	Full text	Citation source
Year	PMID						
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	<a href="#">PMC678638</a>	EPMC	
2021	34206824	Current Evidence about Developmental Dysplasia of the Hip in Pregnancy.		Simionescu AA, Cirstoiu MM, Cirstoiu C, Stanescu AMA, et al	<a href="#">PMC8305660</a>	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	<a href="#">PMC847225</a>	EPMC	
2021	33817978	The Mechanisms and Functions of GDF-5 in Intervertebral Disc Degeneration		Guo S, Cui L, Xiao C, Wang C, et al	<a href="#">PMC8126946</a>	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	<a href="#">PMC7941218</a>	EPMC	
2021	34203285	Genetic Study of IL6, GDF5 and PAPPA2 in Association with Developmental Dysplasia of the Hip.		Harsanyi S, Zamborsky R, Krajcová L, Kokavec M, et al	<a href="#">PMC8303839</a>	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.		Vatlinchinov 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	<a href="#">PMC6957535</a>	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	<a href="#">PMC7488690</a>	EPMC	
2020	32244273	Developmental Dysplasia of the Hip: A Review of Etiopathogenesis, Risk Factors, and Genetic Aspects.		Harsanyi S, Zamborsky R, Krajcová L, Kokavec M, et al	<a href="#">PMC7230892</a>	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	<a href="#">PMC7558486</a>	EPMC	
2020	32144293	Generation and characterization of human induced pluripotent stem cells (iPSCs) from hand osteoarthritis patient-derived fibroblasts.		Castro-Vilà E, Gómez R, Sanjurjo-Rodríguez C, Piñeiro-Ramírez M, et al	<a href="#">PMC7060311</a>	EPMC	

# Ensembl ... Exploring genes

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



# Ensembl ... Exploring variants ... Exercise 2

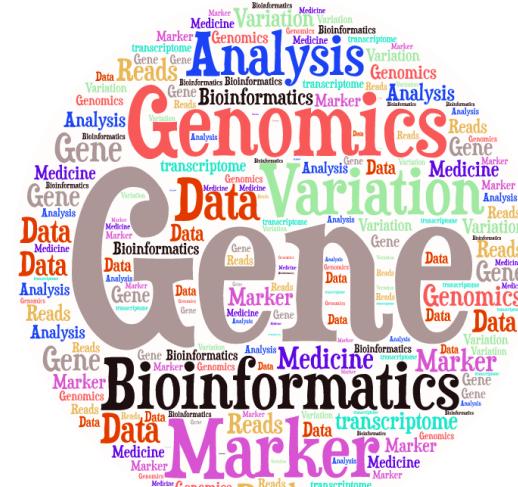
Explore rs2075650 (build 37).

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?



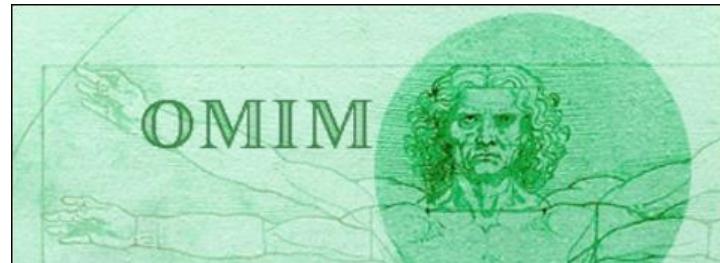
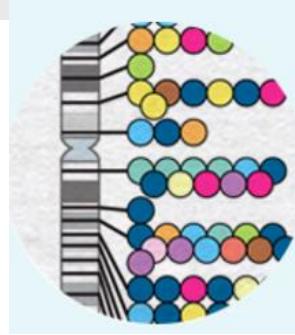
# Publicly available bioinformatics resources

- 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

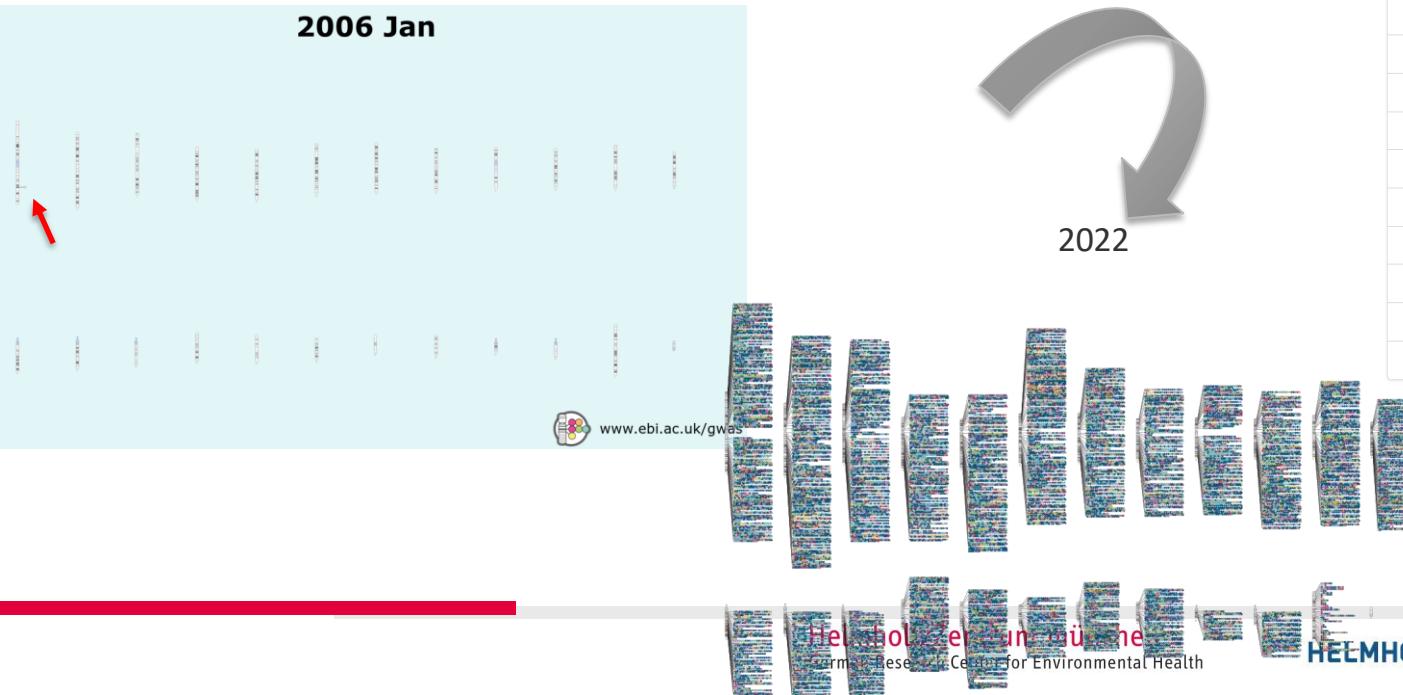
- GWAS catalog (<https://www.ebi.ac.uk/gwas/> )
- OMIM (<http://omim.org/> )
- Orphanet (<http://www.orpha.net> )



# Human Traits & Diseases Databases

## GWAS catalog

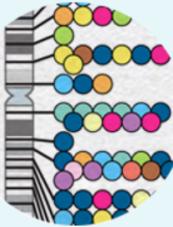
- “GWAS catalog” is a catalog of published/unpublished genome-wide association studies.



Show SNPs for	
Digestive system disease	640
Cardiovascular disease	1607
Metabolic disease	1096
Immune system disease	2380
Nervous system disease	3120
Liver enzyme measurement	1363
Lipid or lipoprotein measurement	4772
Inflammatory marker measurement	469
Hematological measurement	11555
Body weights and measures	3984
Cardiovascular measurement	1875
Other measurement	2764
Response to drug	446
Biological process	2523
Cancer	2713
Other disease	3005
Other trait	2822

# 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:** 20:3543347-35454746 **Cytogenetic region:** 20q11.22 **Biotype:** protein coding

Associations 92 Studies 67

# Human Traits & Diseases Databases

## GWAS catalog

V 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

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 x 10 <sup>-16</sup>		0.37	-	8.25 z score increase	-	GDF5	Vertex-wise cortical surface area	cortical surface area measurement	-	GCST90095130	20:35438203
rs143383-T	9 x 10 <sup>-78</sup>		NR	-	0.06367727 unit increase	[0.057-0.07]	GDF5	Height	body height	-	GCST008839	20:35438203
rs143383-T	1 x 10 <sup>-9</sup>		NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-G	1 x 10 <sup>-17</sup>		0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-T	8 x 10 <sup>-16</sup>		0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-T	5 x 10 <sup>-7</sup>		NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203

Studies 6

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
Aoyama 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

### G GDF5

**Description:** growth differentiation factor 5

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

Available data:

Associations 112

Studies 78

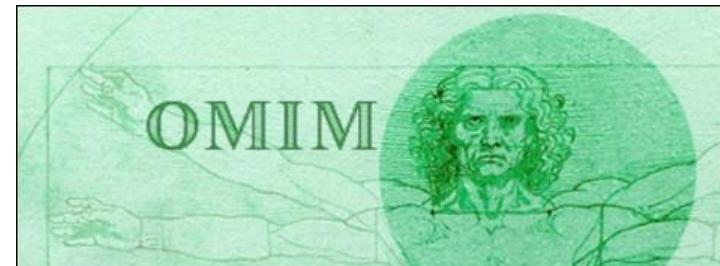
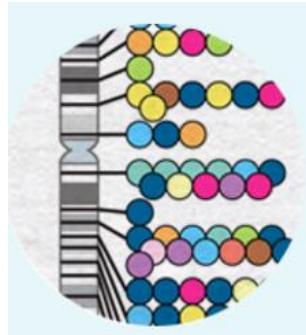
Traits 40



# Human Traits & Diseases Databases ...

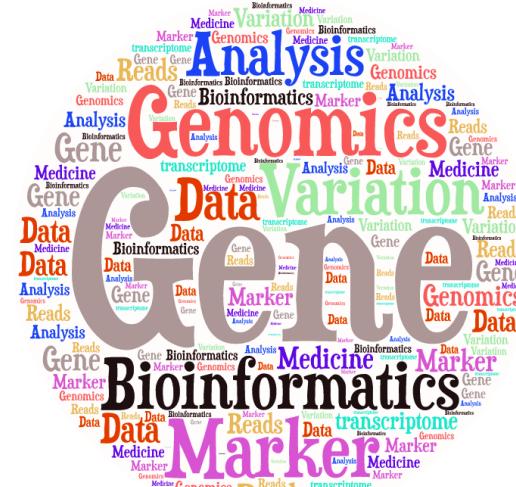
## Exercise 3

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



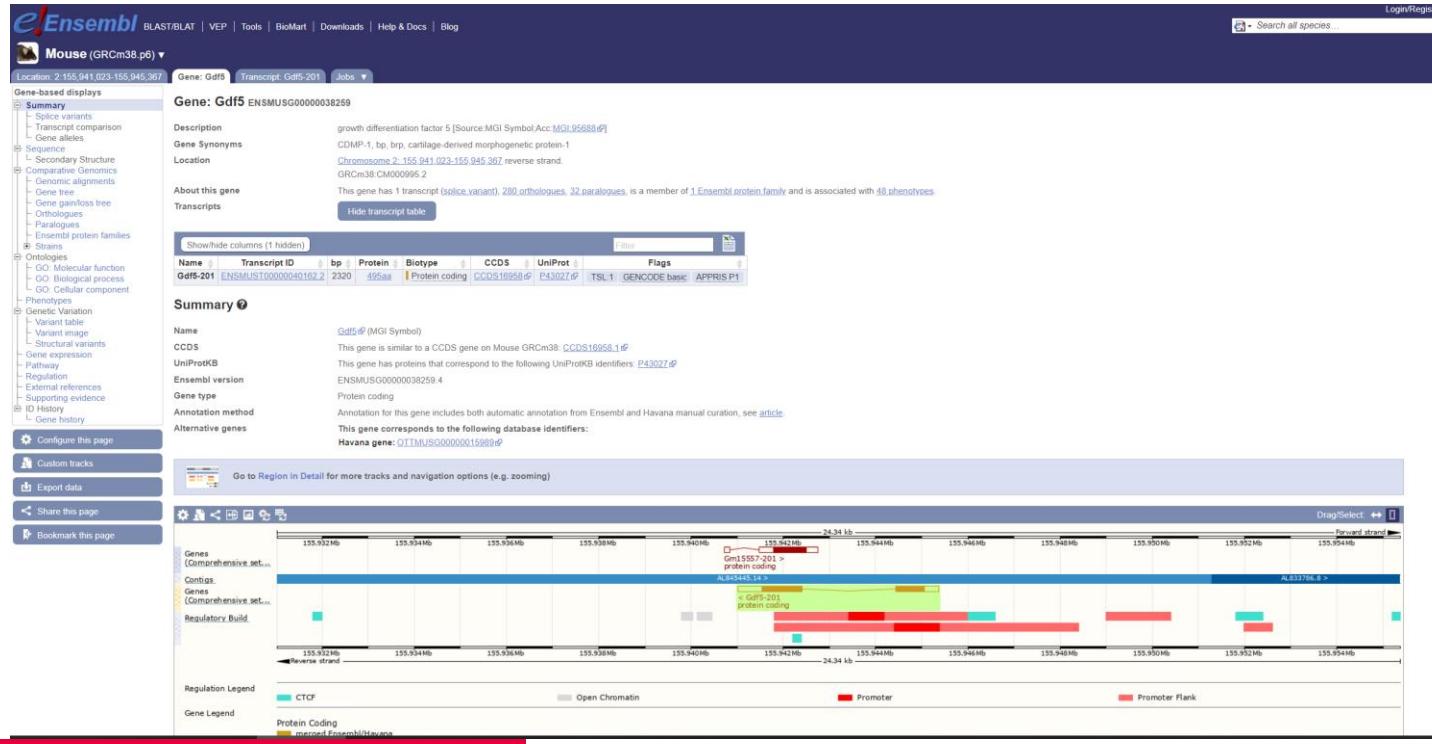
# Publicly available bioinformatics resources

- 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](http://www.ensembl.org/Mus_musculus/Info/Index)



The screenshot displays the Ensembl gene detail page for *Gdf5* (ENSMUSG00000038259). The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar at the top right allows searching across all species. The main content area shows the gene's description, synonyms, and location on Chromosome 2. It also provides information about orthologues, paralogues, and phenotypes. The central part of the page features a table of transcripts for *Gdf5*, with columns for Name, Transcript ID, bp, Protein, Biotype, CCDS, UniProt, and Flags. Below this is a summary section with details like MGI Symbol (Gdf5), UniProtKB ID (P43027), and Ensembl version (ENSMUS00000038259). The bottom half of the page contains a genomic track viewer showing the gene's position from 155.932 Mb to 155.954 Mb. The tracks include Genes (Comprehensive set), Contigs, Regulatory Build, Regulation Legend (CTCF, Open Chromatin, Promoter, Promoter Flank), and Protein Coding (CCDS16958.1, P43027.6, TSL1, GENCODE basic, APPRIS P1).

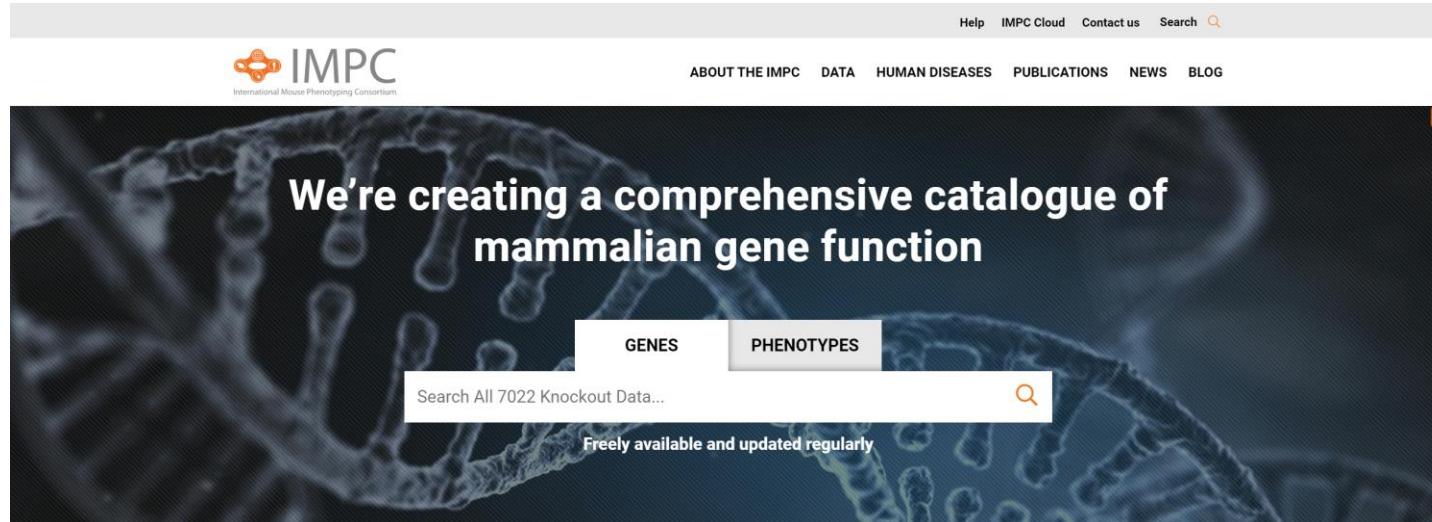
# Databases of other Organisms

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

The screenshot shows the homepage of the Mouse Genome Informatics (MGI) database. At the top, there's a navigation bar with links for Search, Download, More Resources, Submit Data, Find Mice (IMSR), Analysis Tools, Contact Us, and Browsers. The MGI logo is on the left, and the Alliance of Genome Resources logo is on the right. A search bar at the top left says "Type your search here". Below it is a sidebar with various search and analysis tools: Genes, Phenotypes & Mutant Alleles, Human-Mouse: Disease Connection, Gene Expression Database (GXD), Recombinase (cre), Function, Strains, SNPs & Polymorphisms, Vertebrate Homology, Mouse Models of Human Cancer, Batch Data and Analysis Tools, and Nomenclature. A "Quick Search" button is also in this sidebar. To the right of the sidebar is a large banner titled "Mouse Models for Coronavirus Research" which highlights MGI's role in providing resources for COVID-19 research. Below the banner, there's a "What's new at MGI" section with a list of updates and a "Getting Started" section with links to introductory materials. The footer contains social media links (Facebook, Twitter) and a copyright notice.

# Databases of other Organisms

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



# Databases of other Organisms

- Zebrafish: <http://zfin.org/>

Search expert curated zebrafish data

Any ▾ heart contraction abnormal

Search

Fig. 1 of Monroe et al., 2016

**Genes**  
Search for genes, transcripts, clones, and other markers

**Expression**  
Search for gene expression data, and annotated images

**Mutants/Tg**  
Search for mutants, knockdowns, transgenics, and affected phenotypes

**Antibodies**  
Search for antibodies by gene, labeled anatomy, and other attributes

**BLAST**  
Align nucleotide and protein sequences with zebrafish datasets

**Publications**  
Search for zebrafish research publications and scientific literature

**About ZFIN**

The Zebrafish Information Network (ZFIN) is the database of genetic and genomic data for the zebrafish (*Danio rerio*) as a model organism. ZFIN provides a wide array of expertly curated, organized and cross-referenced zebrafish research data.

[Learn More](#)

**Additional Resources**

Data Mining  
ZebrafishMine BioMart

# Databases of other Organisms

- Drosophila: <http://flybase.org/>

The screenshot shows the FlyBase website homepage. At the top, there is a navigation bar with links for Home, Tools, Downloads, Links, Community, Species, About, Help, and Archives. Below the navigation bar, there are several search and browse options: BLAST, GBrowse, JBrowse, Resources, RNA-Seq, Vocabularies, ImageBrowse, and Batch Download. A banner at the top right indicates the version is FB2020\_05, released on Oct 14, 2020. The main content area features a "QuickSearch" bar with various filters like Human Disease, Protein Domains, Gene Groups, Pathways, GO, Data Class, and specific search fields for FlyBase, Homologs, GAL4 etc., Expression, Phenotype, and References. Below the search bar, there is a "Everything" search input field and a "Search" button. To the left of the search bar, there is a video thumbnail titled "Using JBrowse on FlyBase". On the right side of the search bar, there is a note about using wild cards (\*). The bottom half of the page displays a grid of 12 images representing different fly tissues and organs: Adipose, Circulatory, Excretory, Muscle, Imaginal Precursor, Integumentary, Tracheal, Digestive, Nervous, and Reproductive.

# 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/>)

The screenshot shows the homepage of Ensembl Genomes. At the top, there's a grid of small images representing different organisms. Below this is the main header: "e! EnsemblGenomes Providing genome data for non-vertebrate species, with tools for the manipulation, analysis and visualisation of that data". To the right is a "Contact us" link. The page features several sections:

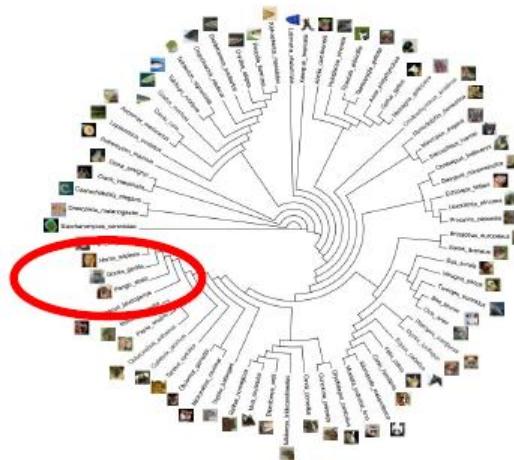
- Ensembl**: COVID-19, SARS-CoV-2 Genome sequence & annotation data.
- Ensembl**: Rapid Release, 2 weekly releases of new assemblies with gene & protein feature annotation.
- EnsemblPlants**: Includes *Triticum aestivum* IWGSC, *Oryza sativa* Japonica Group IRGSP1.0, *Arabidopsis thaliana* TAIR10, and a "Go to Ensembl Plants" button.
- EnsemblMetazoa**: Includes *Cenorhabditis elegans* WBcel235, *Drosophila melanogaster* BDGP6.2B, and *Bambusos merrillii* ASMT16Gv1, with a "Go to Ensembl Metazoa" button.
- EnsemblProtists**: Includes *Plasmodium falciparum* 3D7 ASM270v2, *Dichotomium discoideum* dicty\_2.7, and *Phytophthora infestans* ASM1429v1, with a "Go to Ensembl Protists" button.
- EnsemblFungi**: Includes *Magnaporthe oryzae* MG8, *Saccharomyces cerevisiae* R64-1-1, and *Aspergillus nidulans* ASM1142v1, with a "Go to Ensembl Fungi" button.
- EnsemblBacteria**: Includes *Streptococcus pneumoniae* Spn42298\_v1.0, *Escherichia coli* AL505-MS, and *Bacillus subtilis* ASM73511v1, with a "Go to Ensembl Bacteria" button.

At the bottom, it says "EMBL-EBI Ensembl Genomes is developed by EMBL-EBI". On the far right is the EMBL-EBI logo.

# Databases of other Organisms ...

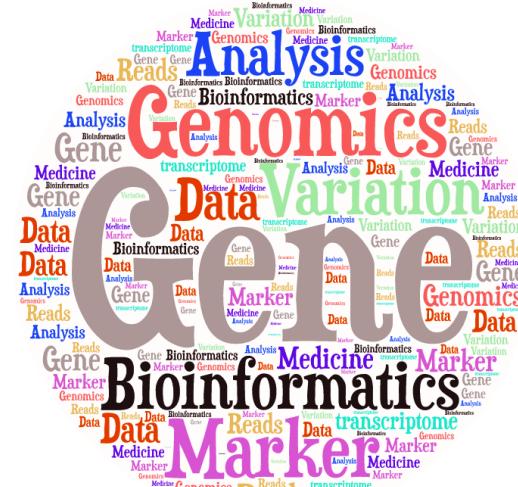
## Exercise 4

- 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/>).

The screenshot shows the GTEx Portal homepage. At the top, there's a navigation bar with links for Home, Downloads, Expression, Single Cell, QTL, IGV Browser, Tissues & Histology, Documentation, About GTEx, Publications, Access Biospecimens, FAQs, and Contact. A search bar and a Google Sign In button are also present.

A prominent banner at the top left announces the "snRNA-Seq Data released" on 2021-07-20, stating that 8 tissues and 16 GTEx donors (25 frozen tissue samples total) have been released. Below the banner, there are two main sections: "Resource Overview" and "Explore GTEx".

The "Resource Overview" section includes links to Current Release (V8), Tissue & Sample Statistics, Tissue Sampling Info (Anatomogram), Access & Download Data, Release History, and How to cite GTEx. It also contains a detailed paragraph about the GTEx project's goal of building a comprehensive public resource for studying tissue-specific gene expression and regulation.

The "Explore GTEx" section is divided into several categories:

- Browse:** By gene ID, By variant or rs ID, By Tissue, Histology Viewer.
- Single Cell:** Data Overview, Multi-Gene Single Cell Query.
- Expression:** Multi-Gene Query, Transcript Browser.
- QTL:** Locus Browser (Gene-centric), Locus Browser (Variant-centric).

At the bottom left, there's a section for "Developmental GTEx" with a brief description of its purpose and goals. On the far left, there's a large decorative graphic element.

# Gene Expression Databases

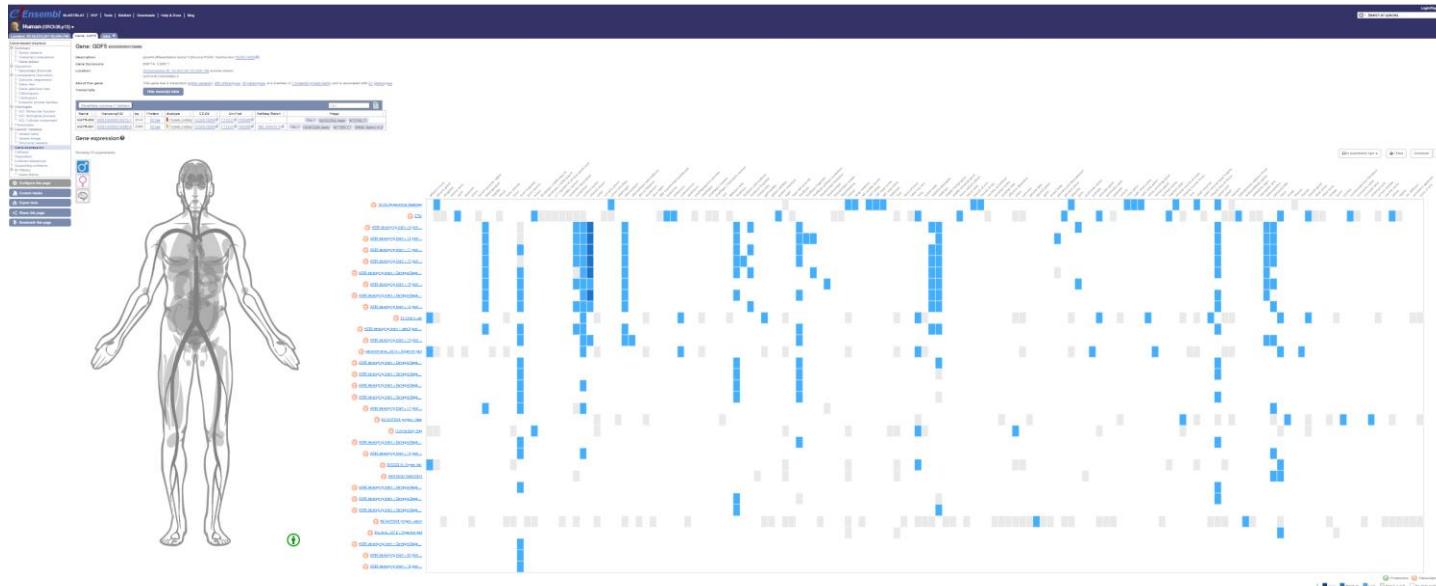
- Expression Atlas: Differential and Baseline Expression (<http://www.ebi.ac.uk/gxa/home> or through Ensembl).

The screenshot shows the Expression Atlas homepage. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search icon. A green banner on the right says "Query single cell expression" and has a link to "To Single Cell Expression Atlas". Below the banner, the main title "Expression Atlas" is displayed next to a logo of a magnifying glass over a blue circle. A sub-header reads "Gene expression across species and biological conditions". A horizontal menu bar includes Home, Browse experiments, Download, Release notes, FAQ, Help, Licence, About, and Support. A search bar at the top says "Search across 65 species, 4,052 studies, 134,900 assays". To the right, it says "Ensembl 99, Ensembl Genomes 46, WormBase ParaSite 14, EFO 3.10.0". The main content area has sections for "Gene / Gene properties" (with a search bar for "Enter gene query...") and "Species" (set to "Any"). It also has a "Biological conditions" section with a search bar for "Enter condition query...". Below these, there's a "Search" button and a "Clear" button. Further down, there's a navigation bar with tabs for "Animals", "Plants", and "Fungi", with "Animals" currently selected. A row of icons and links for various model organisms follows:

Species	Experiments	Baseline	Differential
Homo sapiens	1449	59	1390
Mus musculus	1153	46	1107
Rattus norvegicus	152	3	149
Drosophila melanogaster	140	4	136
Gallus gallus	36	3	33
Caenorhabditis elegans	29	1	28

# Gene Expression Databases

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



# Gene Expression Databases ...

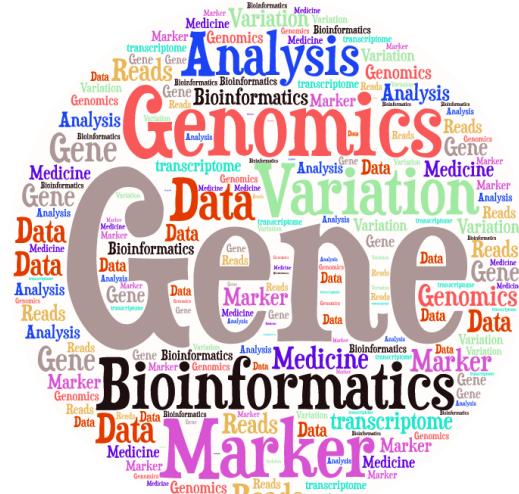
## Exercise 5

- 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:

- [Policy Issues in the Development and Adoption of Biomarkers for Molecularly](#)

- [1. Targeted Cancer Therapies: Workshop Summary.](#)

National Cancer Policy Forum, Board on Health Care Services, Institute of Medicine.  
Washington (DC): National Academies Press (US); 2015.

PMID: 25855848   [Free Books & Documents](#)

[Similar articles](#)

[journal title](#)  
[abbreviation](#)

[authors](#)

- [Four-wave mixing experiments with extreme ultraviolet transient gratings.](#)

- Bencivenga F, Cucini R, Capotondi F, Battistoni A, Mincigrucci R, Giangrisostomi E, Gessini A, Manfredda M, Nikolov IP, Pedersoli E, Principi E, Svetina C, Parisse P, Casolari F, Danailov MB, Kiskinova M, Masciovecchio C. *Nature*. 2015 Apr 9;520(7546):205-8. doi: 10.1038/nature14341.

PMID: 25855456

[Similar articles](#)

[volume & issue](#)

[e-pagination](#)

- [Molecular imaging of angiogenesis after myocardial infarction by \(111\)In-](#)

- [3. DTPA-cNGR and \(99m\)Tc-sestamibi dual-isotope myocardial SPECT.](#)

Hendrikx G, De Saint-Hubert M, Dijkgraaf I, Bauwens M, Douma K, Wiertz R, Pooters I, Van den Akker NM, Hackeng TM, Post MJ, Mottaghy FM. *EJNMMI Res*. 2015 Jan 28;5:2. doi: 10.1186/s13550-015-0081-7. eCollection 2015.

PMID: 25853008   [Free PMC Article](#)

[Similar articles](#)

[publicaton date](#)

# 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 Sugees  
Senior Bioinformatician

# Team144 pipeline

Variant details	
General information:	
rsID:	<a href="#">rs143384</a>
Chromosome:start-end:	<a href="#">20:34025756-34025756</a>
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
Synonyms:	rs431839, rs61433024, rs765106, rs3748435, rs17422934
Phenotypes (risk allele):	<a href="#">Height(A),</a> <a href="#">Infant length(G),</a> <a href="#">Height(G)</a>

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



Dr Daniel Sugees  
Senior Bioinformatician

HelmholtzZentrum münchen  
German Research Center for Environmental Health

Thank you!

**HELMHOLTZ** RESEARCH FOR  
GRAND CHALLENGES

# What is a genome assembly?

## Sequence reads

CGGCCTTGGGCTCCGCCTTCAGCTCAAGA		
CAGCTGTCCCAGATGAC	ACTTAACCTCCCTCCCAGCTGTCC	
GGGCTCCGCCTTCAGCTC		TCCCAGCTGTCCCAGATGACGCCATC
	AACTTCCCTCCCAGCT	
CGGCCTTGGGCTCC		TCCGCCTTCAGCTCAAGACTTAACCTC
	CAGATGACGCC	

## Match up overlaps

CGGCCTTGGGCTCCGCCTTCAGCTCAAGA	AACTTCCCTCCCAGCT	CAGATGACGCC
TCCGCCTTCAGCTCAAGACTTAACCTC	TCCCAGCTGTCCCAGATGACGCCATC	
GGGCTCCGCCTTCAGCTC	ACTTAACCTCCCTCCCAGCTGTCC	
CGGCCTTGGGCTCC		CAGCTGTCCCAGATGAC

## Genome assembly

CGGCCTTGGGCTCCGCCTTCAGCTCAAGACTTAACCTCCCTCCCAGCTGTCCCAGATGACGCCAT

# 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**