

22

**Bioinformatic Tools Practical**

**EL017**

**Vid Prijatelj & Cindy Boer & Linda Broer**

# Foreword

All of the tools and databases in this practical are web-browser based, but some web-browsers work better than others. Please use Chrome or Firefox – sometimes issues with Edge/Internet explorer arise.

# Objective

he objective of this practical is to get acquainted with a number of online bioinformatics tools and databases for the annotation of GWAS.

Feel free to do this with some of your own variants or genes you are interested in, or ask us questions how you could use databases/bioinformatics tools in your own research.

# Introduction

## Results from a genome-wide association study on hand Osteoarthritis

For this first part of the practical we will be looking at the results of a (real) genome-wide association study (GWAS) on osteoarthritis of the hand. Osteoarthritis is a degenerative joint disease characterized by the breakdown of joint cartilage(chondrocytes) and the underlying bone (osteoblasts) in the affected joint. Osteoarthritis is a complex disease meaning that both genetic and environmental factors influence the disease risk, for osteoarthritis of the hand the heritability is estimated to be around ~60%, and well known environmental factors are obesity and mechanical loading.

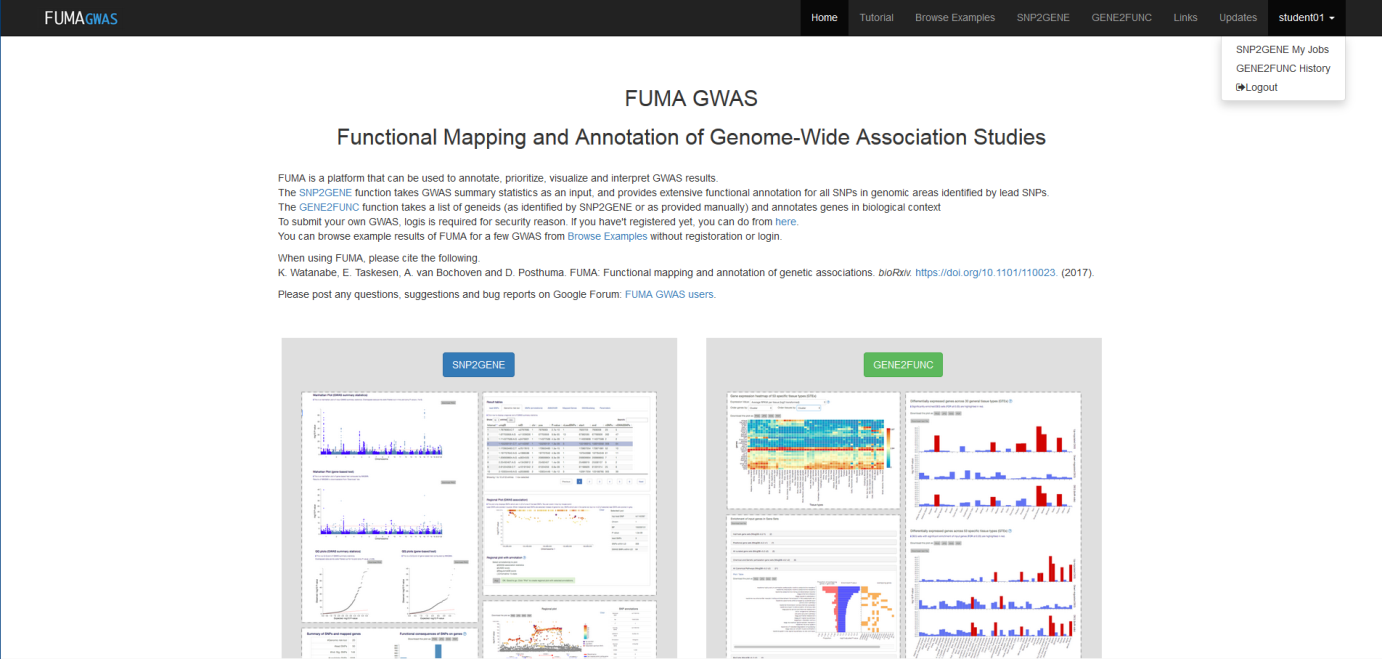
You have performed a GWAS on ~8,700 individuals, to analyse your results you have uploaded them into the FUMA tool (online tool for Functional Mapping and Annotation of GWAS).

* Go to: <http://fuma.ctglab.nl/>
* And Login using:

Username: prbiostudent01@gmail.com

Password: Prbioinformatics2018

* Go to “Student01”in the top right hand corner and go to the dropdown menu click on “SNP2GENE My Jobs”
* Click on “go to results” of the job with the jobID “HandOA”



You are taken to the results of the SNP2GENE analysis. As the name suggests this section primarily focusses on identifying potential causal genes. Additionally, it gives you some nice graphs which can easily be exported and used in publications (such as the Manhattan plot & QQ plots).

1. What is the difference between the two Manhattan plots provided?

SNP/GENE

1. How many Genome-wide significant signals do you see in the Manhattan plot

872

Scrolling down further on this page you will get the QQ plots for your GWAS and also some functional analyses. In this practical we will focus on the functional analysis and annotation. Let’s go to the ‘Results’ button on the left side. There are many different tables here, but let’s go to the Genomic risk loci table.

1. What is the most significant SNP in this GWAS and where is it located?

rs4764133 12:15064363

Let us see if we can find more information on our Top SNP

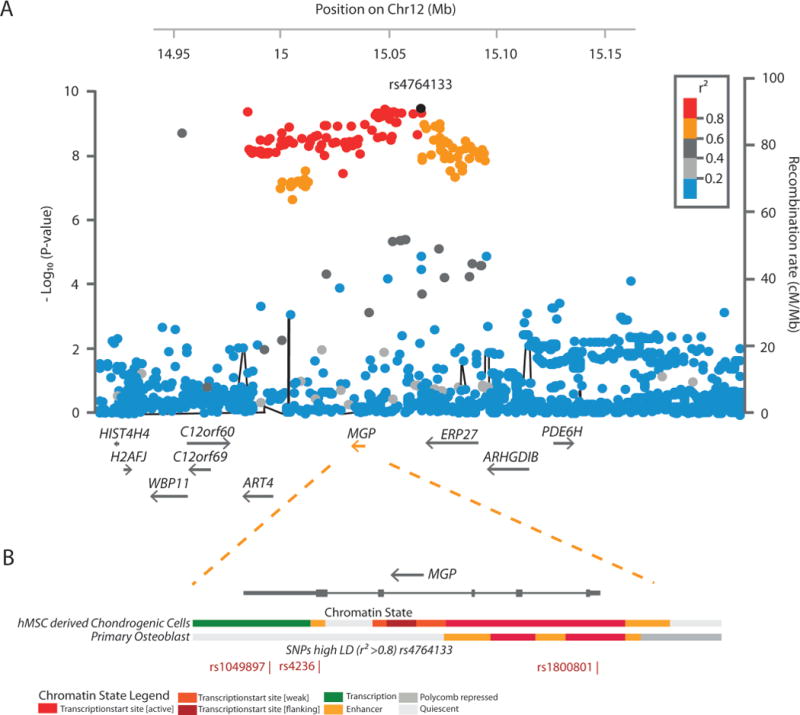
* Select Click on the Top SNP in this table, so that it is highlighted
* Scroll down to the “Regional Plot (GWAS association)
* Select “GWAS association statistics” and “unselect” the rest of the options
* Click on Plot, now a new tab will open

On this newly opened tab, FUMA will plot an interactive LocusZoom plot. This might take a while, if it takes too long, you can move to question 6 (and do the steps described above this question) and then come back to this question later.

* **NOTE the Locus Zoom part of FUMA might (still) bugged and not work – below is an image of the locus zoom as you would have seen in FUMA if the Locus zoom had worked.**

1. What do the “dots” in the locusZoom plot mean? What do the colours of these dots mean?

Dots are snv, color is red



1. Which genes are covered by the LD region of the top SNP (LD>0.8)?

MGP erp27 art4

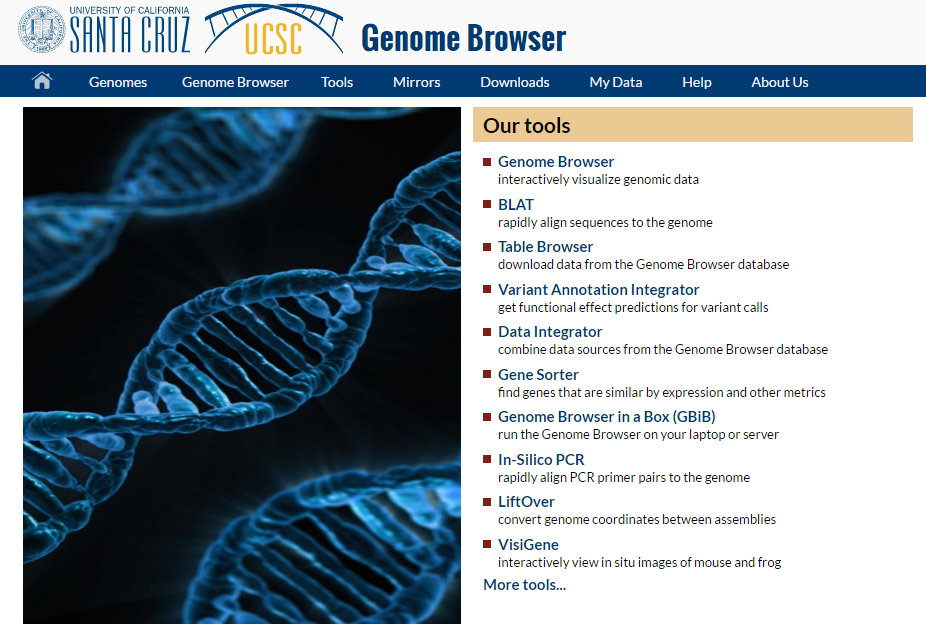
* Now Go back to the Manhattan plots

10. are these genes (from the LocusZoom plot) also in the Gene based Manahattan plot, and are they significant associated?

yes

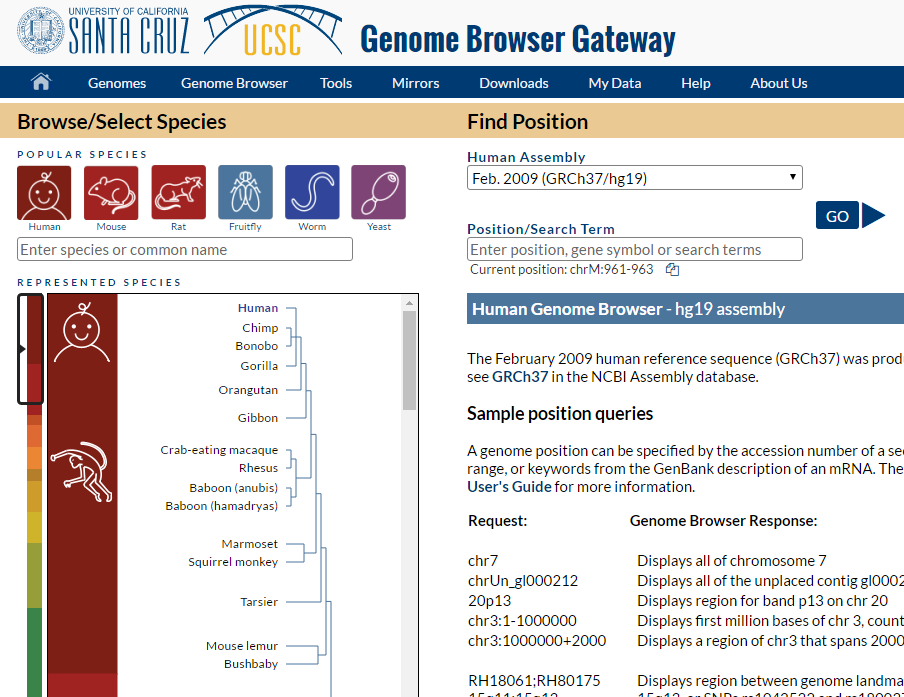
That does not really tell us anything on what could possibly be our Causal gene. Let’s see if we can find some more information about the genes surrounding your SNP. We will do this using the UCSC genome browser. To get a more visual image of where your SNP is located were are now going to the UCSC genome browser in a new tab or window:

* Go (**in a new window or tab**) to the UCSC genome browser <https://genome-euro.ucsc.edu>



* Click on Genome Browser (first option under ‘Our tools’). Allow yourself to be redirected to a closer mirror.

Below you see the Genome Browser search window, which should also be on your screen now.



1. Which options are available in the Assembly dropdown box?

The different assembly releases basically have a different genome build. For this practical, you will be **working with the Feb 2009 assembly, as the newer assembly is on a different build than you used for your GWAS.** Now type your lead SNP from the Hand Osteoarthritis GWAS (**rs4764133**) and hit the GO button.

1. How many search results do you obtain? Can you find out what the different results are?  
    37 origine

* Go to the Simple Nucleotide Polymorphisms (dbSNP 155)

A figure of your SNP and surrounding DNA with all kinds of information should now show up. Your SNP is located in the middle of the figure, and at the top of the figure are options to move your screen and/or zoom in/out of your picture, in the bottom track named: “Simple Nucleotide Polymorphisms (dbSNP 155)

The UCSC Genome Browser is build-up of several tracks (Genes, mRNA, ESTs, Conservation, SNPs, etc.). These tracks can be added or deleted using the drop down controls below the figure.To modify the tracks that are displayed you can also use the grey bars on the left of the figure. By right-clicking on one of the bars you get options to change that specific track. By left-clicking on the bar, you get more information on the track, descriptions of the colours used, and more information on the database where the information in the track is coming from (and you can continue to the database itself.

Before you do anything else, you will want to clean your screen. There is a lot of information shown you don’t want.

* **Below your figure is the option ‘Hide all’. Click on it**

Your figure should be empty now. Now we can build up the figure to our liking. Below the ‘Hide all’ option is a long list of options under blue bars(tracks).

* Select the following tracks in the blue bar – genes and predictions - ‘UCSC Genes’ (pack) (press the refresh button on the tracks to load these in the figure)
* Select the following tracks in the blue bar – variation - ‘dbSNP (153)’ (pack) (press the refresh button on the tracks to load these in the figure)

Let’s look at where our lead SNP is located

* Examine the figure (zoom out to see more of the genomic region 100x)

1. What kind of SNP is rs4764133 and which gene(s) are surrounding your SNP?, Is this the same as you saw in the LocusZoom plot in FUMA?

Non-coding

ERP27

1. Look at the genes in the figure. What is the difference between the thin horizontal lines and the much thinker horizontal lines in the gene tracks of the picture?

Extron intron

Let’s see if we can find some more information about the genes surrounding your SNP. In the UCSC genome-browser look at the Genes surrounding your SNP. Let’s look at the closest gene first ERP27

* Go to the description of the ERP27 gene in the UCSC track. Do this by clicking on the gene name of the **upper transcript** in UCSC genes track.

1. What is the full name of your gene, EPR27?

Homo sapiens endoplasmic reticulum protein 27 (ERP27)

On this page, in the table under ‘sequence and links to tools and databases’ you can find links to other databases with more information on this gene. Due to time restraints we can’t possible investigate all of them. The rest of the page gives a summary of information of these databases on your gene

* On this page, in the table under ‘sequence and links to tools and databases’ select the Entrez Gene database.

1. How Many Exons does ERP27 have?

8

1. Can you find what is known about the function of ERP27 here?

This gene encodes a noncatalytic member of the protein disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins. The canonical protein has an N-terminal signal sequence, two thioredoxin (TRX)-like domains and a C-terminal ER-retention sequence. Alternative splicing results in multiple transcript variants encoding distinct isoforms; some of which lack domains present in the canonical protein. [provided by RefSeq, Dec 2016]

1. Does this function suggest a functional role of this gene in Osteoarthritis?

no

Go Back to the UCSC genome-browser and now examine the Other genes nearby the lead SNP

1. What are the known functions for ART4, C12orf60 and MGP?

C12orf60: Homo sapiens chromosome 12 open reading frame 60

ART4: This gene encodes a protein that contains a mono-ADP-ribosylation (ART) motif. It is a member of the ADP-ribosyltransferase gene family but enzymatic activity has not been demonstrated experimentally. Antigens of the Dombrock blood group system are located on the gene product, which is glycosylphosphatidylinosotol-anchored to the erythrocyte membrane. Allelic variants, some of which lead to adverse transfusion reactions, are known. [provided by RefSeq, Jul 2008]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data to make the sequence consistent with the reference genome assembly. The genomic coordinates used for the transcript record were based on transcript alignments.Transcript (Including UTRs)

Mgp: this gene encodes a member of the osteocalcin/matrix Gla family of proteins. The encoded vitamin K-dependent protein is secreted by chondrocytes and vascular smooth muscle cells, and functions as a physiological inhibitor of ectopic tissue calcification. Carboxylation status of the encoded protein is associated with calcification of the vasculature in human patients with cardiovascular disease and calcification of the synovial membranes in osteoarthritis patients. Mutations in this gene cause Keutel syndrome in human patients, which is characterized by abnormal cartilage calcification, peripheral pulmonary stenosis and facial hypoplasia. [provided by RefSeq, Sep 2016].

Another source of information on the function of genes is Gene Ontology (GO). In the Page Index table on UCSC you can click on ‘GO Annotations’ to find out what the GO terms associated with your gene are. GO terms can be subdivided in three groups, namely molecular function, biological process and cellular component.

1. What are the GO terms for biological process of MGP? Does this match with what you found earlier?

GO:0001502 cartilage condensation

GO:0001503 ossification

GO:0007275 multicellular organism development

GO:0030154 cell differentiation

GO:0030500 regulation of bone mineralization

GO:0051216 cartilage development

yes

.

Now that we know more about the genes surrounding your lead SNP, let’s see if we can find a potential functional link between our GWAS results and some of the genes nearby.

* **In FUMA,** go to the “Results” section, there in the results tables go to the tab SNPs(annotation).

1. What does the header IndSigSNP mean? And why are there many more SNPs here than in your lead SNP list? (hit = LD/r2)

In LD

1. Are there any SNPs which are located in an exonic (or coding) regions which are in high LD with any of our lead SNP (**rs4764133)**? And what is their GWAS P-value?

Yes

rs1861698 ART4 synonymous

rs11276 ART4 nonsynonymous SNV

rs3088189 ART4 synonymous

rs4236 MGP nonsynonymous

rs11056243 ERP27 nonsynonymous

1. Look up the SNPs from answer 17 under the “Annovar” tab, what kind of exonic functions do these SNPs have?
2. What is the difference between Synonymous and non-synonymous?

If it makes a difference in transcripyion

* In FUMA go to the “GWAScatalog” tab in the Results section

1. Are there any known GWAS associations with our Lead SNP? (**rs4764133**) and what about the other top Signal: rs7139060?

no

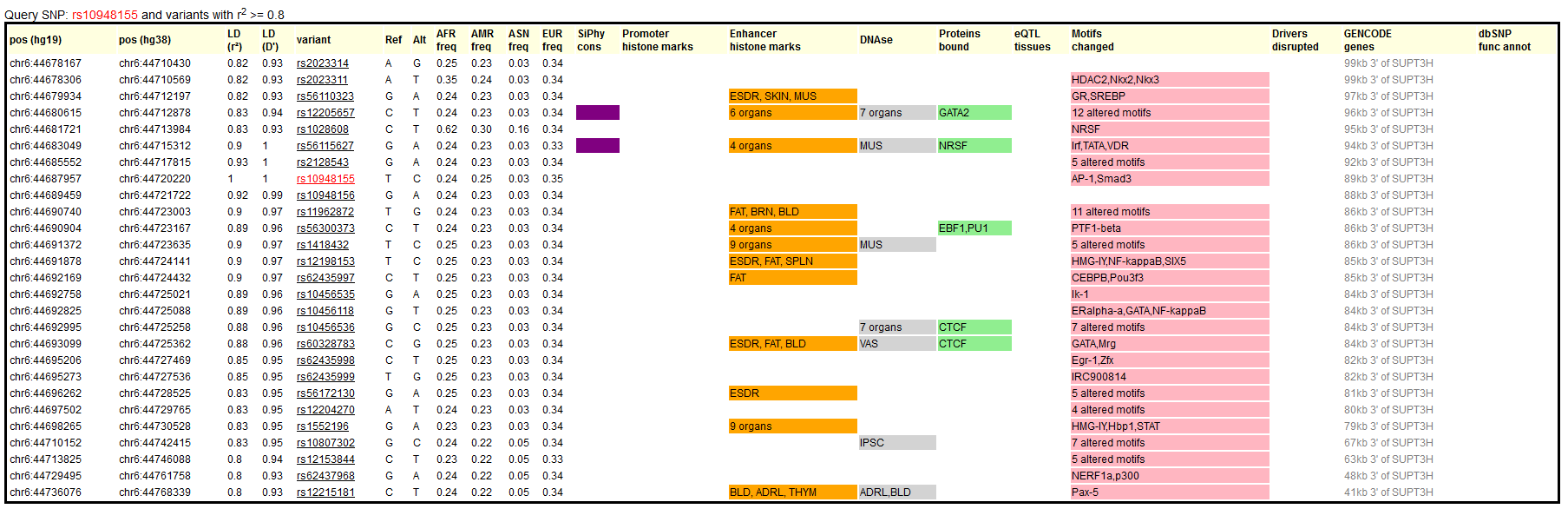
YES HEIGHT

1. Based on the already gathered information which SNPs from the hit **rs4764133** and genes do you think might be interesting to examine further?

MGP

There is also another, way to see if your GWAS top SNP and SNPs in LD are in coding regions of a gene an in regulatory features/regions )See part II/PartIII of the preactical): HaploReg. HaploReg uses data from the UCSC, ENCODE and the Roadmap Epigenetics database and cross correlates this with the location of your SNP.

* Go to HaploReg (<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>)
* (V4.1 is the default)
* Enter **rs4764133**  in the search bar (the webpage is slow)
* Below you see the typical Haploreg output: you will see your top SNP and all SNPs in high LD (r2>0.8) with you top SNP



1. According to HaploReg in which genes does rs4764133 have coding variants?

C12orf60

MGP

1. Click on rs11276 to get more information about the SNP, to which genes is rs11267 annotated?

ART4

C12orf60

* Go back to the UCSC genome-browser
* In the search bar at the top of the screen above your genome-browser figure type in The two RSID’s separated by the ;

**rs11276; rs4236**

You will now see that both these SNPs are highlighted in the UCSC genome browser (Sometimes this feature does not work, if so just type in one variant and do the questions and later/other tab the other variant). By left-clicking on the bars near the tracks, you get more information on the track, descriptions of the colours used, and more information on the database where the information in the track is coming from (and you can continue to the database itself).

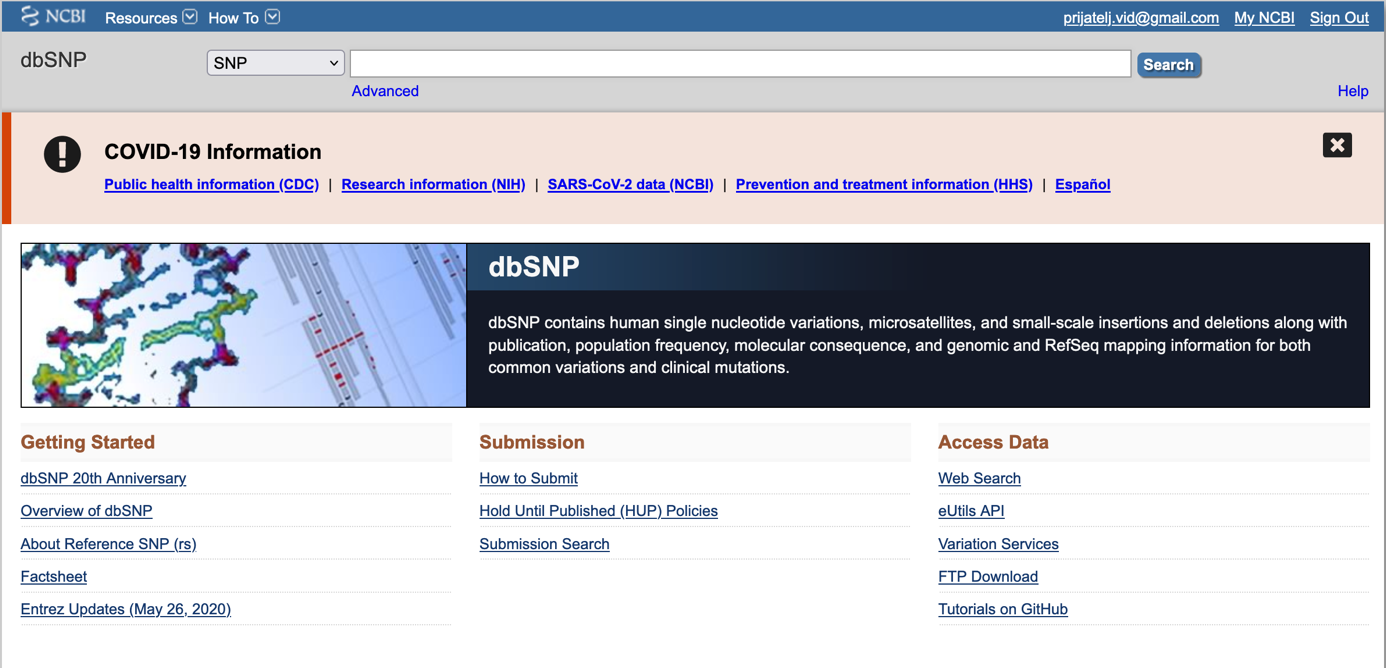
1. Where is rs11276 located according to UCSC? In which gene is the SNP located?

Now that we have identified these two SNPs and examined where they are located, Let’s see if we can find more information about them:

1. In the UCSC SNP track: What is the meaning of the colours of SNPs in the SNP-track? Hint: Use the bar on the left of the track and check under colouring options)
2. What kind of SNP are **rs11276; rs4236** according to their colour?

Try to get more information on your top SNP from dbSNP. To get to the dbSNP record for specific SNPs, click on the rs-number in the figure itself. Then click on the rsid next to dbSNP.

* Go to dbSNP for rs4236



1. What is the amino-acid change caused by this variant?

Now scroll down the page to the Population diversity section. Here you see the allele frequencies a bunch of different submitters (Study information) have posted for this variant. Some of the submitters have provided allele frequencies in only Europeans or in multiple ethnicities grouped together. Be careful – do not look under the ALFA Allele Frequency table. You are interested in the frequencies in European, East-Asian and African populations separately.

1. Given the above, which result for the allele frequencies would you trust most?
2. What is the allele frequency of T and C in Caucasian, East-Asian and African populations?

SNP information can are also aggregated in other databases such as Ensemble:

* Go to <http://www.ensembl.org>
* In the search bar in the centre or on top of the screen search for **rs11276**
* Click on the search result for rs11276 (human variant)

1. Wat is the evidence status for rs11276? Is the variant validated?
2. What is the consequence of this variant (what is the amino acid change caused by this variant and the predicted effect, benign or deleterious)?

Let’s recap what you have found out so far. We have found two variants in high LD with our lead SNP which are located in the coding regions of 2 genes and cause an amino acid change in these genes. You have checked the frequency and the validity of these variants. You also know the location of these genes and what their functions are. This has given you a lot of useful information, including an important clue, but you will have to link this gene to Osteoarthritis.

To know more about diseases or traits your gene is involved in, you can also consult the OMIM (Online Mendelian Inheritance in Man) database. You can enter the OMIM database via the ‘Sequence and Links to tools and databases’ table at the UCSC genome browser. Or in the Ensemble Browser under “phenotype data” and in the table under Source you will have a link to OMIM

OMIM contains an historical overview of disease histories and information genes with links to literature.

1. If you go to the OMIM database for ART4 and MGP, what is the accession number of the gene?
2. Is there anything known about diseases or traits your gene is involved in?

You are also interested in what can be learned from model organisms, in particular mice. Go back to the UCSC Gene information page, and in the ‘Sequence and Links to Tools and Databases’ table, you select ‘MGI’. This link will take you to the MGI database. This is the international database resource for the laboratory mouse, providing integrated genomic and biological data to facilitate the study of human health and disease.

1. What does the abbreviation MGI stand for?

Search for your gene. Under Genome Features click on your gene symbol (first option) and read through the page you get. Under Mutations, alleles and phenotypes you get information on existing mouse models for your gene.

1. Is there a knock-out mouse available for ART4 and MGP? What phenotypes are seen?
2. Another mouse phenotype website is the International Mouse Phenotyping Consortium (IMPC – <https://www.mousephenotype.org>). Search for both MGP and ART4 knockouts. What phenotype summary data can you provide? What phenotypes would be of interest (but they aren’t)? Is in this case the significance cut-off really so important?

Let’s examine the gene expression of our genes and if our lead SNP have an eQTL with one of our genes”. For this we will use the GTeX portal.

* Go to GTeX - <https://gtexportal.org/home/>
* Seach for our genes: *MGP, ERP27 and ART4*

1. In which tissues are the genes expressed?

* Seach for **rs4764133** (enter this into the search bar)

1. What kind of eQTL can be found for our lead SNP?
2. Are these eQTL(s) important four our GWAS results? why or why not?

* Go Back to FUMA to the “Results” tab

After looking into these different databases it is time to answer the most important question yet:

1. Which of the researched genes a plausible gene for affecting Hand Osteoarthritis and why?

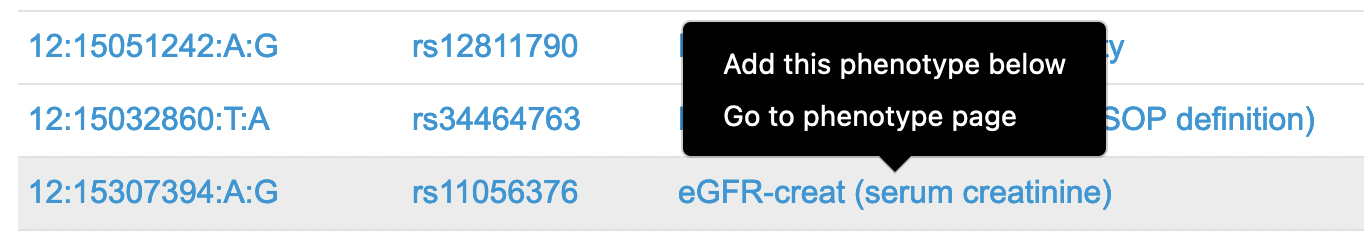
This is the end of Part I of the practical. The above example is real, and published (<https://ard.bmj.com/content/76/12/2046>) but do note that the chance of finding a hit in a coded region with such an obvious effect on the gene is highly unusual. Most variants are not located in coding sequence. In Part II of this practical you will learn how you can still get more information on this locus by using a different real-life example.

The field is focusing more and more into single large databases, where not only is data aggregated, but also additionally tested, thus streamlining *in silico* methods to be tested *in vitro/in vivo*.

One of such portals is the “Musculoskeletal Knowledge Portal”, aggregating data on musculoskeletal genomics.

1. Go to <https://msk.hugeamp.org> and type in MGP. Which traits had a variant within this gene significantly associated.

1. Given information you found top, do they make sense? How can you explain serum creatinine levels?
2. It will be easier to explore serum creatinine levels directly at the publication. Highlight the “eGFR-creat (serum creatinine)” and click on the “Go to phenotype page”



Scroll down the page and find the dataset with the largest sample size. What is the study name, what is the sample size and what ancestry was included in the study?

1. Find the CKDGen Study. Can you find more about the study? What is the article name, first author and which phenotypes were measured?
2. Is MGP mentioned anywhere in the main text? Which variant has the lowest p-value in this study? To which gene is it annotated and based on what? What was the focus of this gene, with lowest p-value?
3. Can you MGP absolutely anywhere? If no, why do you think so?
4. There are also a couple of other portals that offer interesting insights, including the PheWAS. Go to GWAS Atlas (<https://atlas.ctglab.nl/>) and under PheWAS type in the MGP. What traits are genome-wide significant? What exactly is the significance in this PheWAS? Which traits are no longer there?
5. How are SNPs tied to genes in the GWAS Atlas? Which studies? RTFM!
6. So why exactly did we lose a couple of traits?
7. What does that tell us in terms of online databases?
8. We are interested in genetic correlation of the osteoarthritis to other traits. There is a website you can find online called “LD Hub”. You can find the link here <https://www.broadinstitute.org/publications/broad11906>

Navigate then to the LD Hub website.

1. Where can you find the option to perform genetic correlation?
2. This is now a great teaching moment. That website was a monumentally important pillar of genetic epidemiology, that stopped working sometime in July 2021. Why – we don’t know. It just did. When that happens, you are pretty much out of luck.
3. Soooo, we want to perform the genetic correlation analysis. Can we do it some other way? How? What do you think is the biggest hurdle?
4. Did the authors of that paper already perform the genetic correlation?
5. Find the 2021 cell paper by same author (CG Boer). Try finding the genetic correlation information. Which trait shows the greatest genetic correlation to the ALLOA?
6. What do you think is the link between the two?

Given the length of the both practicals, this should provide a nice introduction to the GWAS and related database mining.

As there is always a probability of these online databases vanishing, which also includes tool integration, it is very much relevant you, as participants and scientists, get to know the ins-and-outs of different functional tools so that you can use them locally/by yourselves as well.