The SNPstar manual is NOT completely finished! The complete version is available at paper submission!

Manual SNPstar

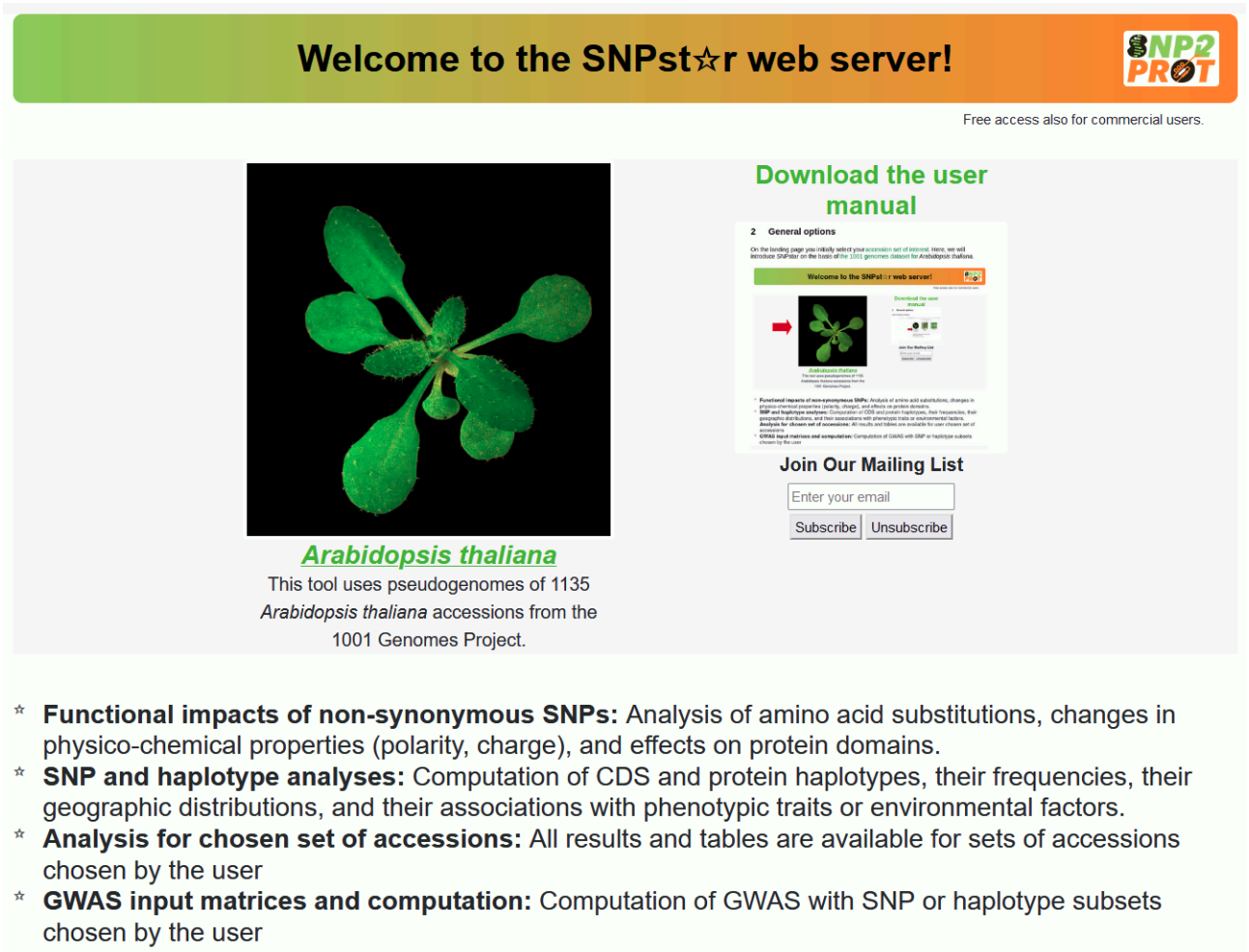
Index

1 Use of SNPstar for scientific research

What can SNPstar do for you? SNPstar is a tool to help researchers in different ways. SNPstar serves as some sort of SNP atlas which collects different kinds of information, characteristics and features for each SNP in genic regions based on genomic information from the 1001 genomes project (https://1001genomes.org/) of *Arabidopsis thaliana*. Currently, SNPstar has a focus on nonsynonymous SNPs, thereby serving as a starting point to gather meaningful information about SNPs that potentially affect protein structure and function. This manual provides a step-by-step guideline for gathering this information for SNPs in (a) gene(s) of interest based on the complete or a subset of the naturally occuring accessions present in the 1001 genomes collection.

2 General options

On the landing page you initially select your accession set of interest. Here, we will introduce SNPstar on the basis of the 1001 genomes dataset for *Arabidopsis thaliana.*

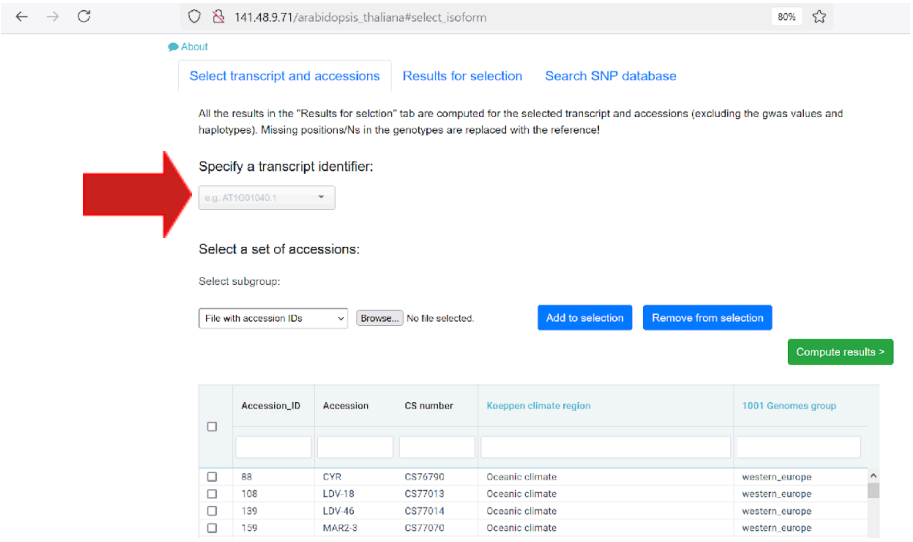


- **Functional impacts of non-synonymous SNPs:** Analysis of amino acid substitutions, changes in physico-chemical properties (polarity, charge), and effects on protein domains.
- **SNP and haplotype analyses:** Computation of CDS and protein haplotypes, their frequencies, their geographic distributions, and their associations with phenotypic traits or environmental factors.
- **Analysis for chosen set of accessions:** All results and tables are available for sets of accessions chosen by the user
- **GWAS input matrices and computation:** Computation of GWAS with SNP or haplotype subsets chosen by the user

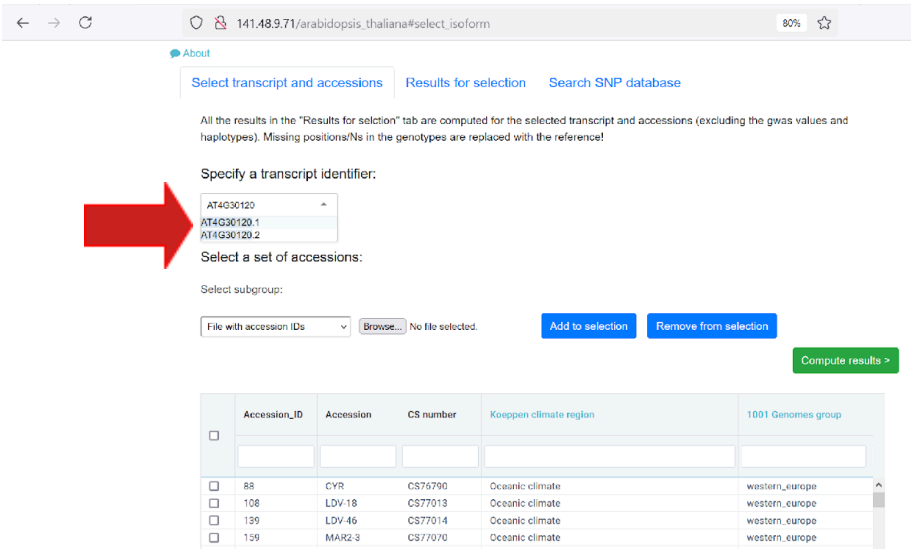
2.1 Gene selection

2.1.1 Select a gene and transcript via AT number

The second step is to select your gene of interest. As an example, we will introduce all the features of SNPstar on the basis of the gene *HMA3* (AT4G30120). You can select any gene of interest either by typing the common gene name (e.g. HMA3) or the AGI identifier in the drop down menu as shown below.



Select your gene model (transcript) of interest. Here, we select AT4G30120.1, but you can select any of the other known splice variants.



2.1.2 Select a group of genes

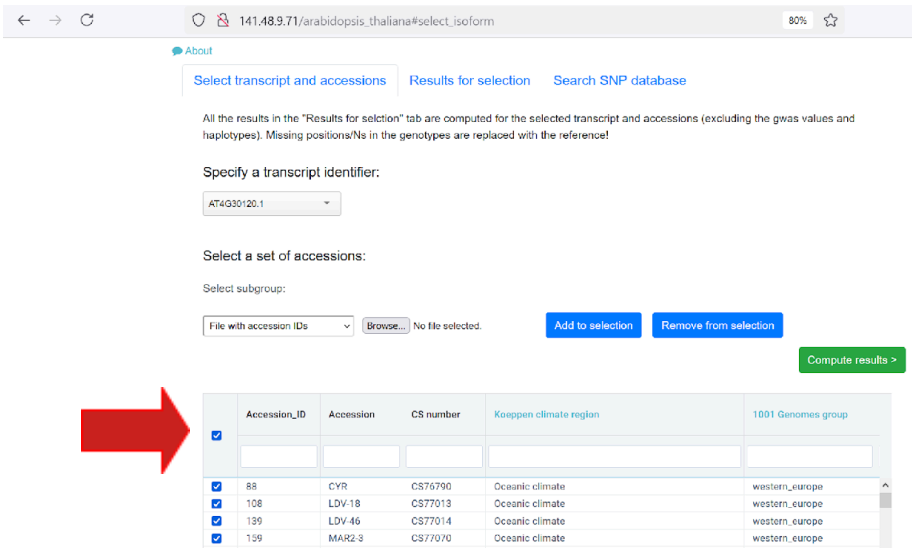
In progress!

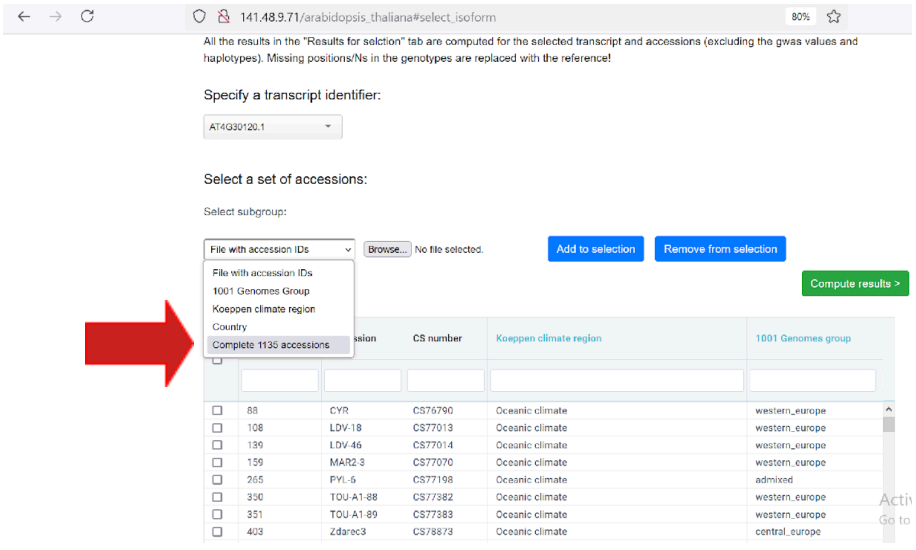
2.2 Accession selection

There are several options to select subsets of accessions for your analysis. We will introduce these in the following subchapters 2.2.1 - 2.2.5.

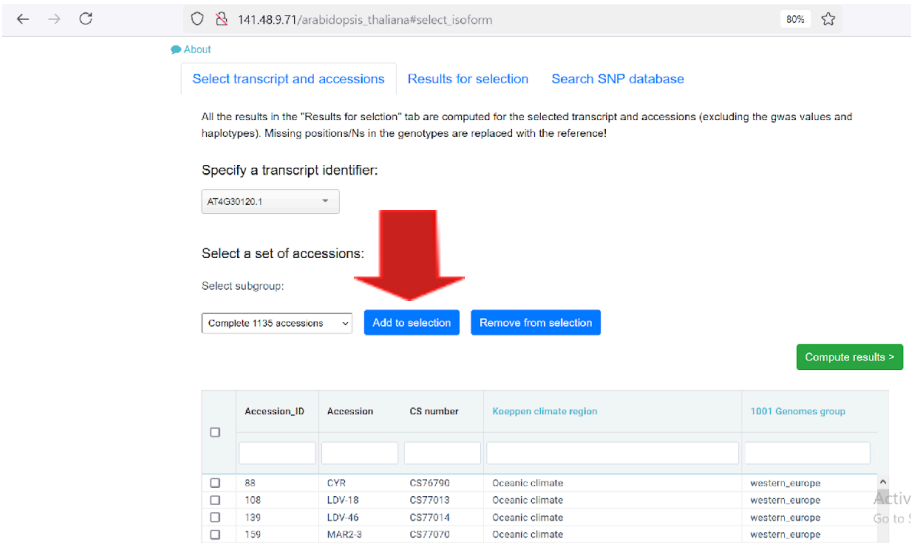
2.2.1. Direct selection all or specified accessions

You can select either the complete set of 1135 accessions or, alternatively, compile a subpopulation by checking the boxes on the left.

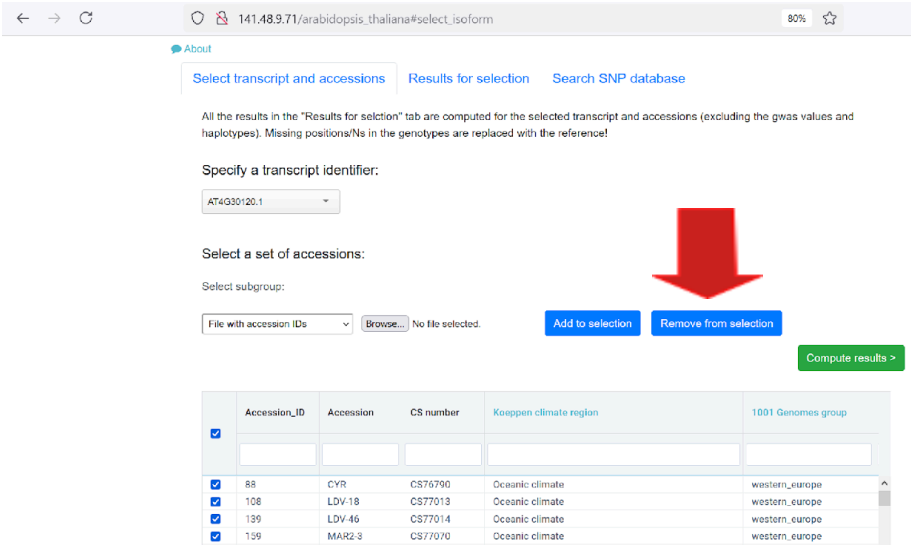




To finish the selection click ‘Add to selection’.

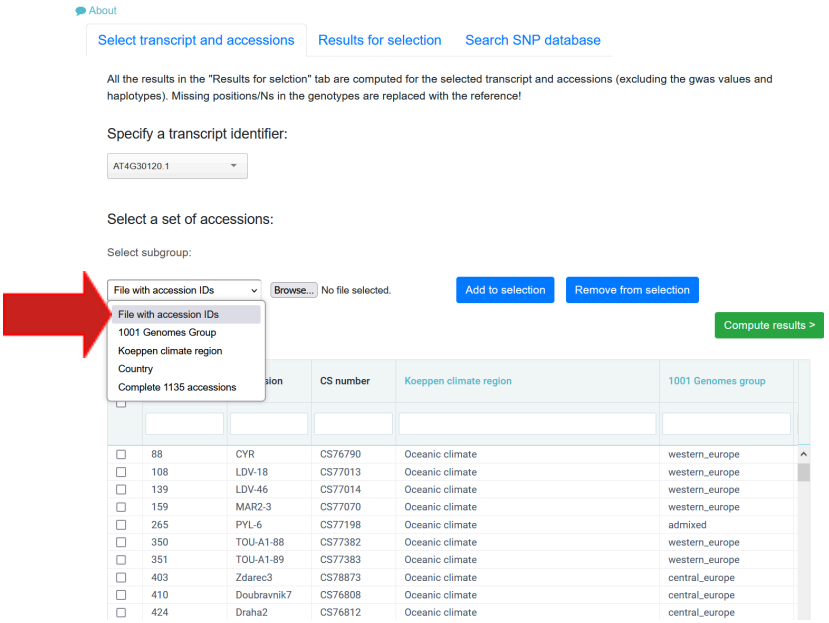


You can remove the selection by clicking ‘Remove from selection’.



2.2.2 Select a subset of accessions via uploading accession IDs

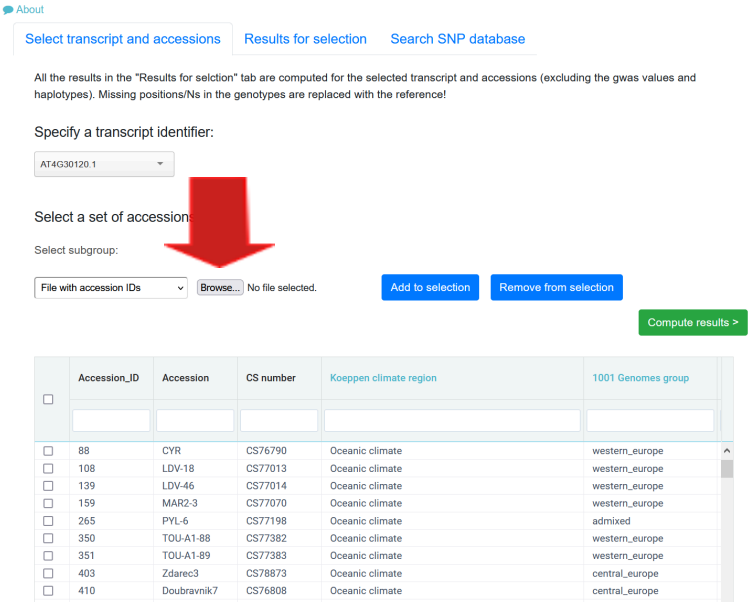
Instead of manually checking all the corresponding boxes, you can upload a file of accessions by marking ‘File with accession IDs’ in the drop down menu.



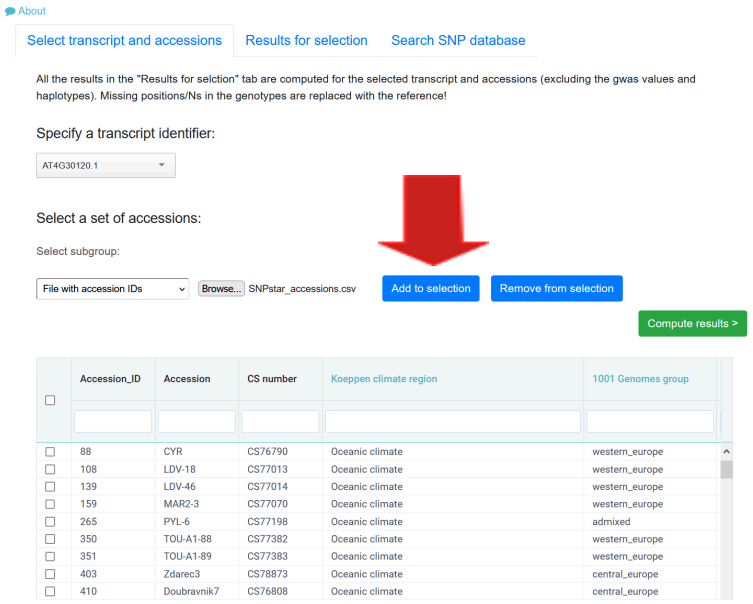
The file with the accession IDs should be in the following format:

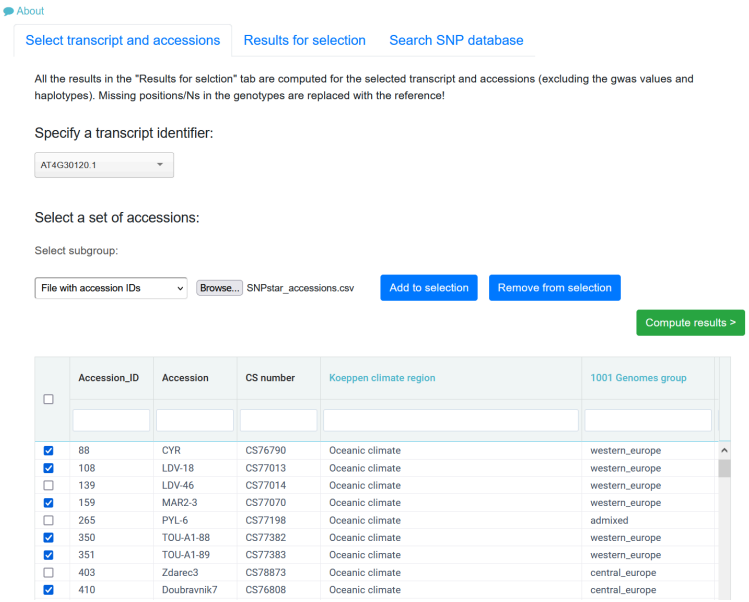
	A	B
1	88	
2	108	
3	159	
4	350	
5	351	
6	410	
7		
8		
9		
10		

Click ‘Browse’ and upload your specified accession list.



Click ‘Add to selection’ for adding.

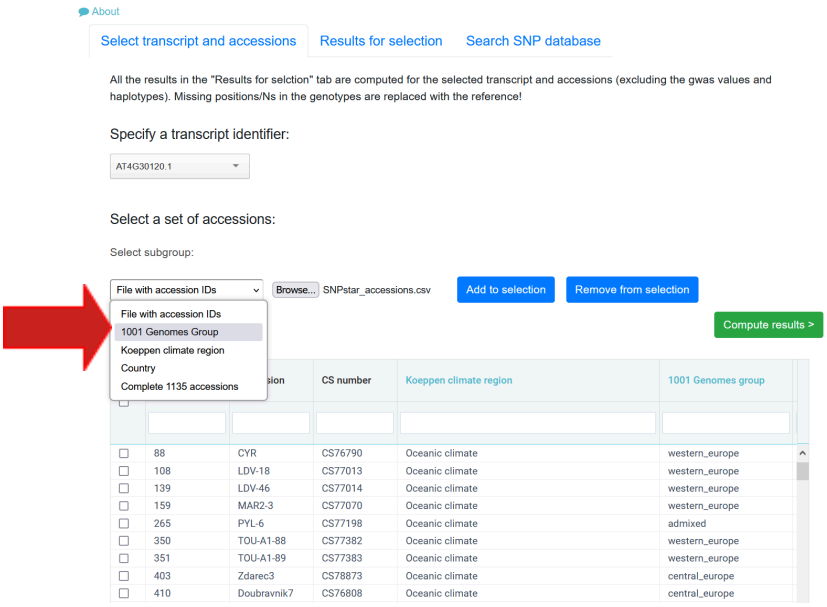




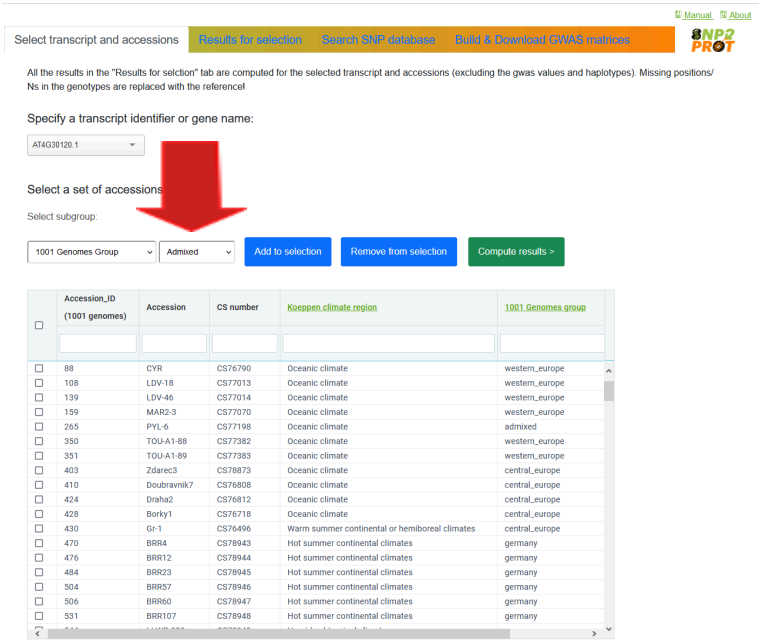
You can always remove the selection via ‘Remove from selection’ (see chapter 2.2.1).

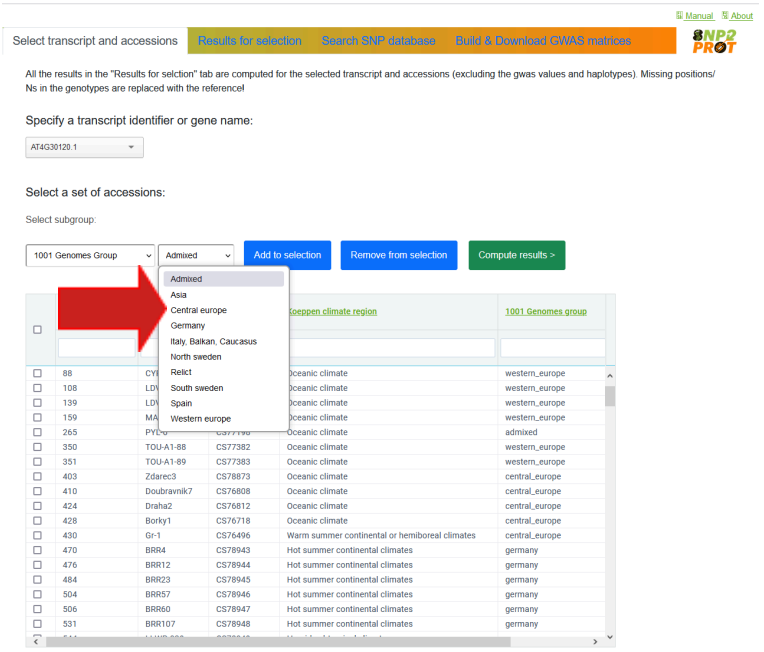
2.2.3 Selecting accession via 1001 genomes groups

Another straight-forward option to select sub-populations is to select genetic diversity groups from the 1001 genomes project. Use the drop-down menu and select ‘1001 Genomes Group’.

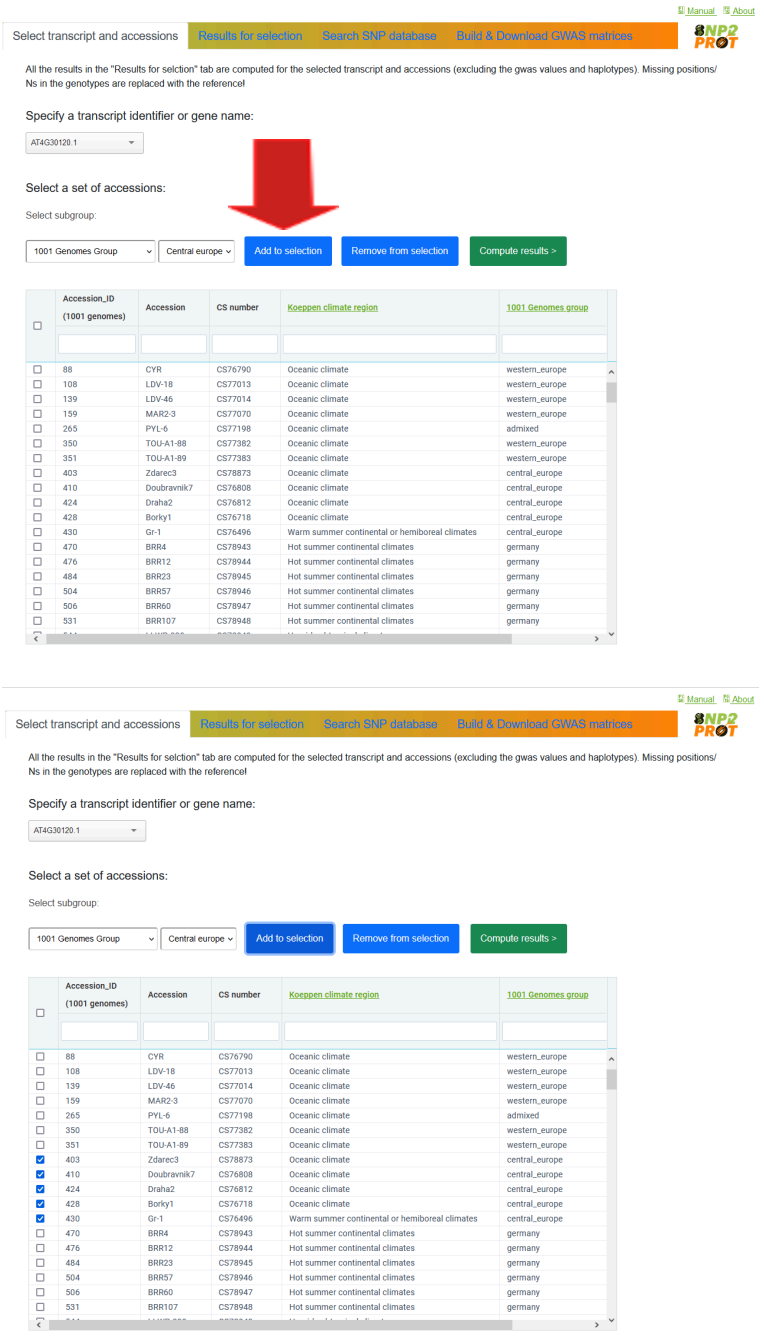


Next, a second drop-down menu appears where you can select the genetic diversity group.





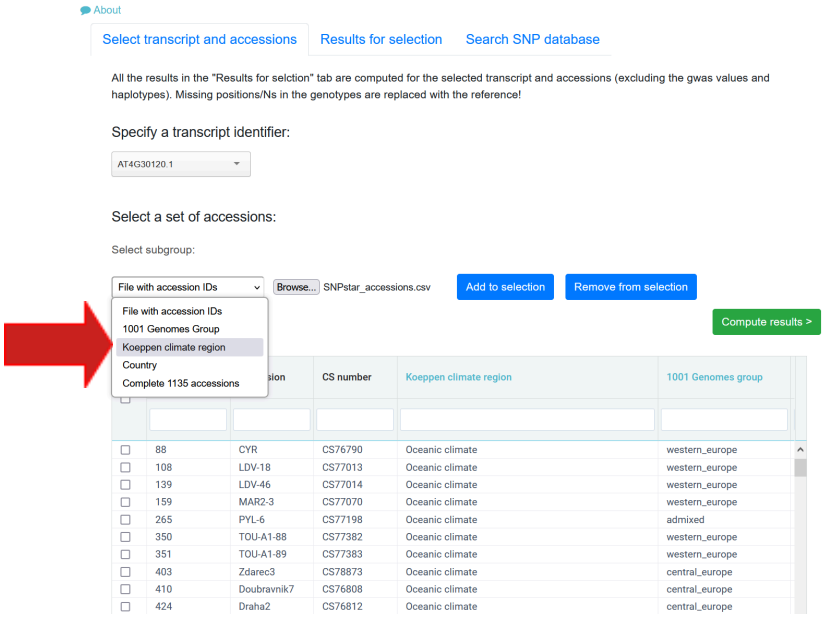
Click ‘Add to selection’.



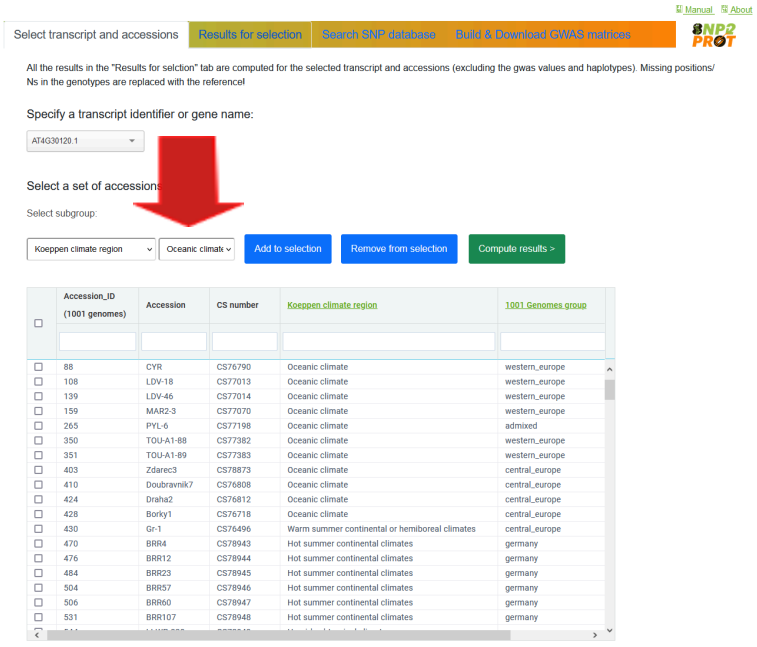
It’s possible to add more 1001 Genome groups if you repeat the procedure with other regions of origin. You can remove specific groups via ‘Remove from selection’ (see chapter 2.2.1).

2.2.4 Select a subset of accessions via Koeppen climate region

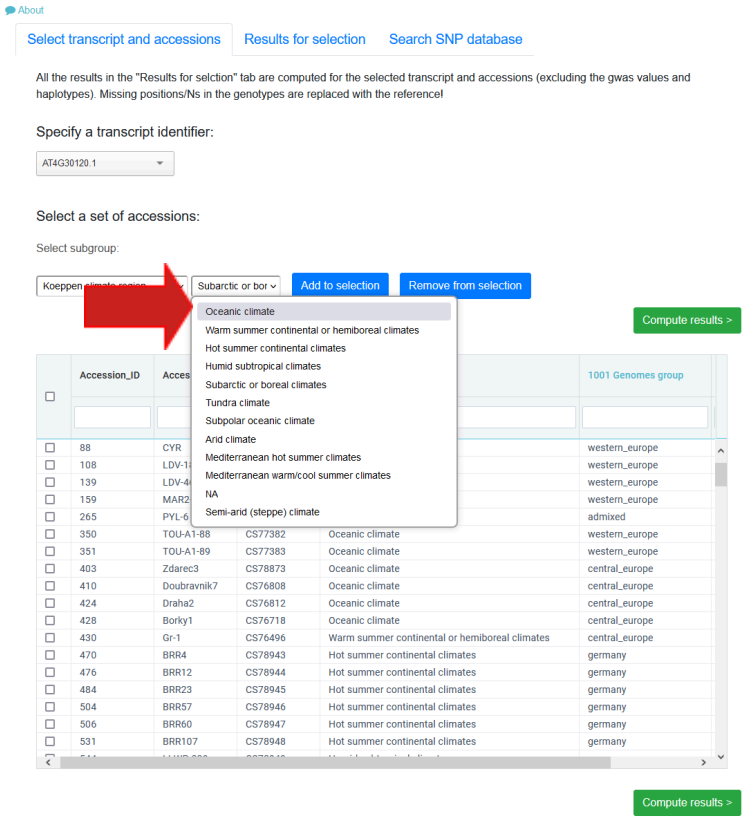
If you would like to select a subset of available accession based on the Koeppen climate regions then use the drop-down menu Select subgroup and select Koeppen climate region. The climate regions according to the Koeppen climate classifications.



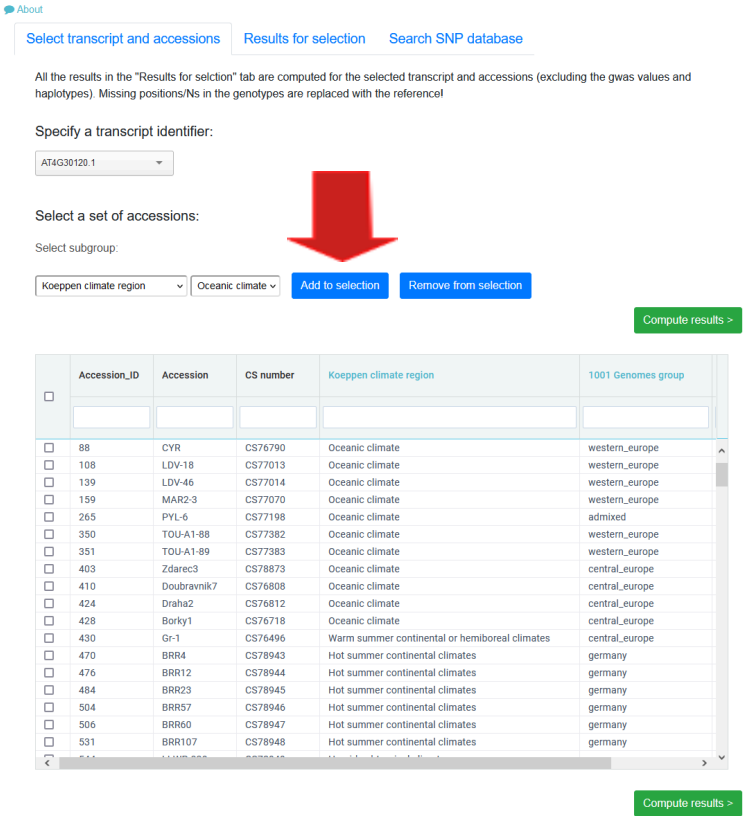
A second drop-down menu appears. In that menu you can select your climate regions of interest.

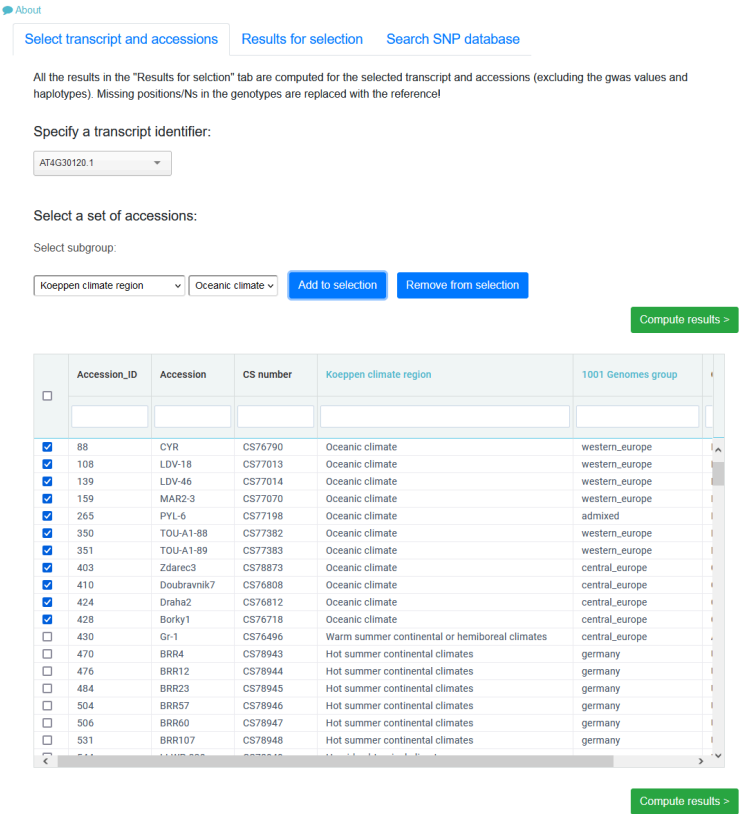


Now you can select your 1001 Genome groups of interest (e.g., Oceanic climate, etc.).



Click ‘Add to selections’.

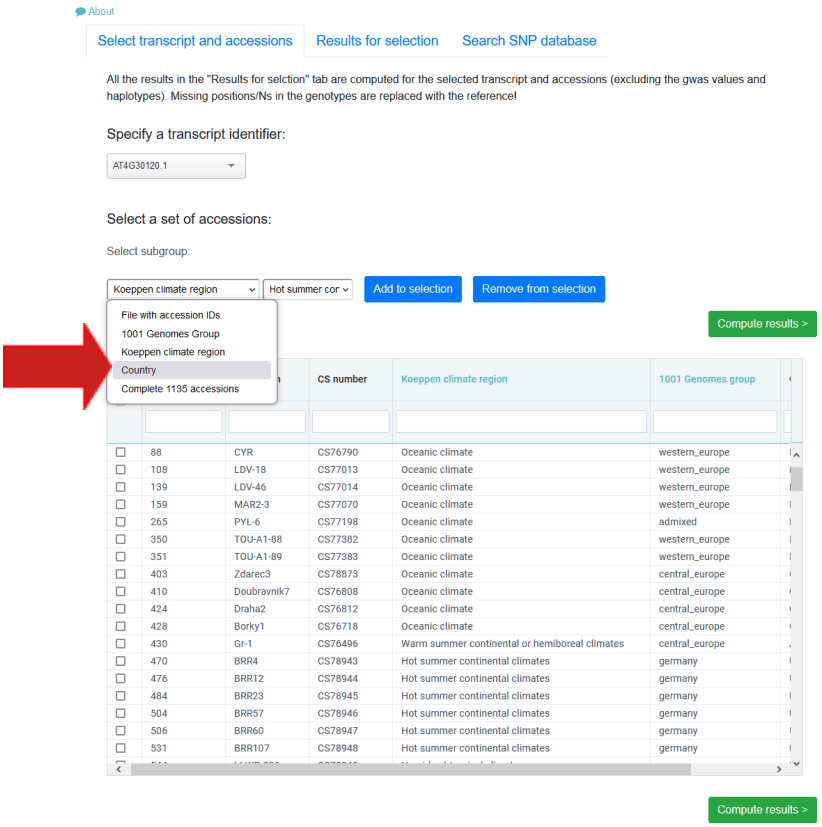




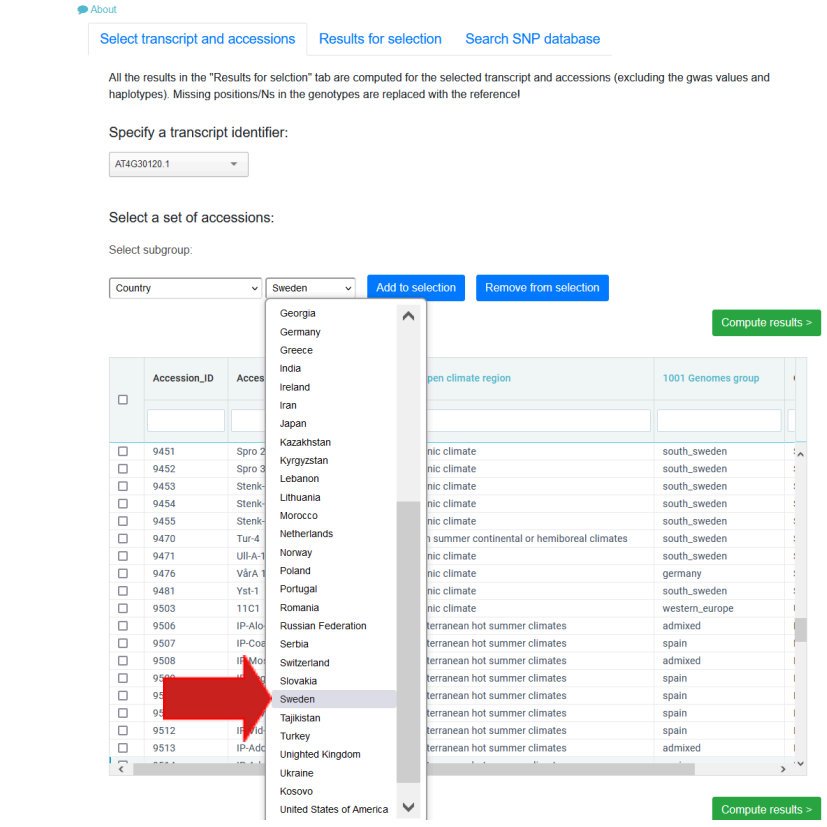
It’s possible to select more than one Koeppen climate region. To do this, repeat the procedure with other climate regions. You can remove specific regions if you select the Koeppen climate region of interest and click ‘Remove the selection’.

2.2.5 Select a subset of accessions via country

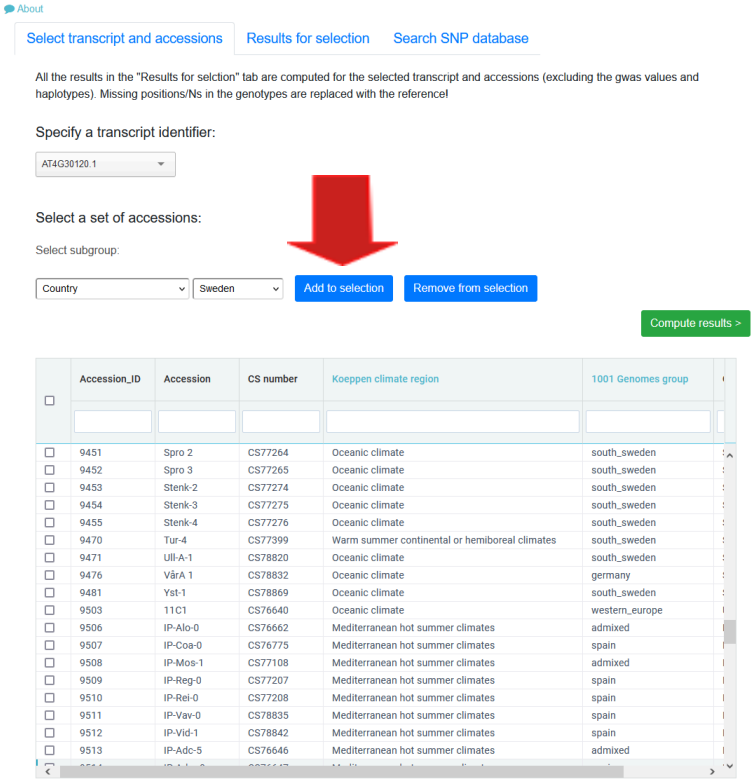
If you would like to select a subset of available accessions based on country of origin, then use the drop-down menu ‘Select subgroup’ and select ‘Country’.

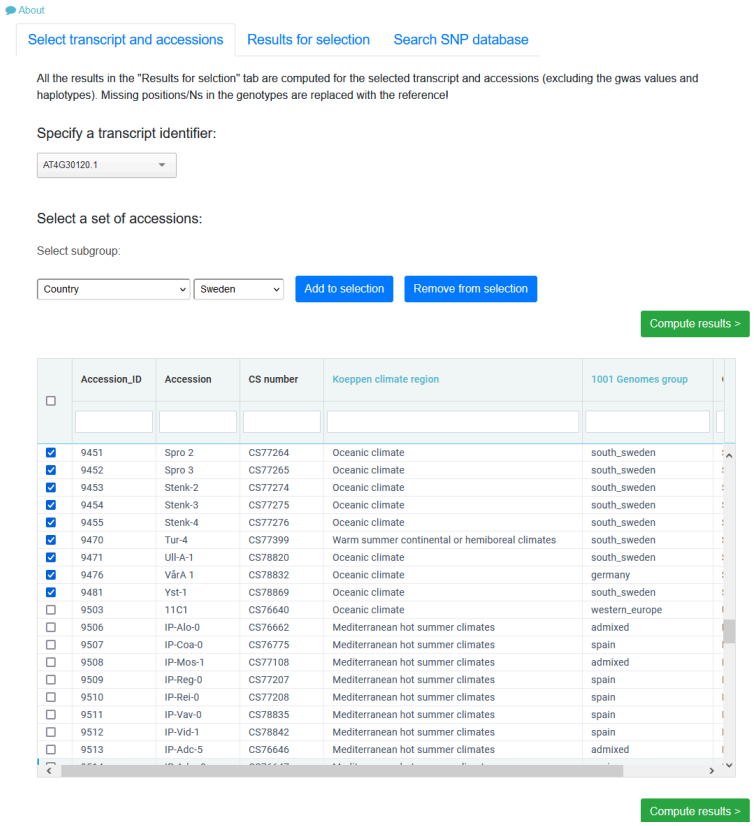


Now you can select your ‘Country’ of interest (e.g., Sweden, etc.).



Click ‘Add to selections’.

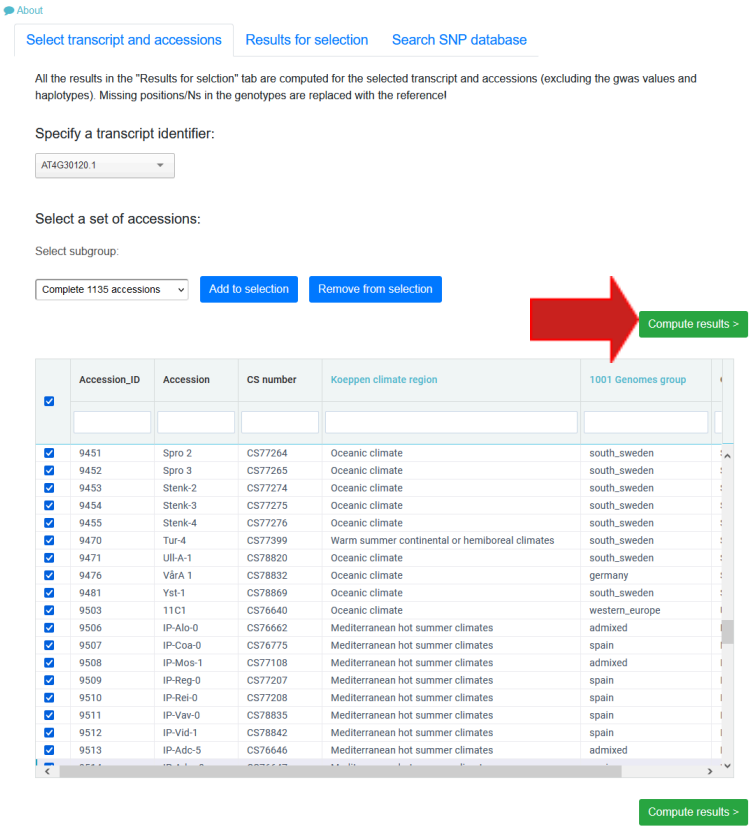

About
Select transcript and accessions
Results for selection
Search SNP database
All the results in the "Results for selction" tab are computed for the selected transcript and accessions (excluding the gwas values and haplotypes). Missing positions/Ns in the genotypes are replaced with the reference!
Specify a transcript identifier:
AT4G30120.1
Select a set of accessions:
Select subgroup:
Country
Sweden
Add to selection
Remove from selection
Compute results >
Accession_ID
Accession
CS number
Koeppen climate region
1001 Genomes group
9451 Spro 2 CS77264 Oceanic climate south_sweden
9452 Spro 3 CS77265 Oceanic climate south_sweden
9453 Stenk-2 CS77274 Oceanic climate south_sweden
9454 Stenk-3 CS77275 Oceanic climate south_sweden
9455 Stenk-4 CS77276 Oceanic climate south_sweden
9470 Tur-4 CS77399 Warm summer continental or hemiboreal climates south_sweden
9471 Ull-A-1 CS78820 Oceanic climate south_sweden
9476 VårA 1 CS78832 Oceanic climate germany
9481 Yst-1 CS78869 Oceanic climate south_sweden
9503 11C1 CS76640 Oceanic climate western_europe
9506 IP-Alo-0 CS76662 Mediterranean hot summer climates admixed
9507 IP-Coa-0 CS76775 Mediterranean hot summer climates spain
9508 IP-Mos-1 CS77108 Mediterranean hot summer climates admixed
9509 IP-Reg-0 CS77207 Mediterranean hot summer climates spain
9510 IP-Rei-0 CS77208 Mediterranean hot summer climates spain
9511 IP-Vav-0 CS78835 Mediterranean hot summer climates spain
9512 IP-Vid-1 CS78842 Mediterranean hot summer climates spain
9513 IP-Adc-5 CS76646 Mediterranean hot summer climates admixed
Compute results >


About
Select transcript and accessions
Results for selection
Search SNP database
All the results in the "Results for selction" tab are computed for the selected transcript and accessions (excluding the gwas values and haplotypes). Missing positions/Ns in the genotypes are replaced with the reference!
Specify a transcript identifier:
AT4G30120.1
Select a set of accessions:
Select subgroup:
Country
Sweden
Add to selection
Remove from selection
Compute results >
Accession_ID
Accession
CS number
Koeppen climate region
1001 Genomes group
9451 Spro 2 CS77264 Oceanic climate south_sweden
9452 Spro 3 CS77265 Oceanic climate south_sweden
9453 Stenk-2 CS77274 Oceanic climate south_sweden
9454 Stenk-3 CS77275 Oceanic climate south_sweden
9455 Stenk-4 CS77276 Oceanic climate south_sweden
9470 Tur-4 CS77399 Warm summer continental or hemiboreal climates south_sweden
9471 Ull-A-1 CS78820 Oceanic climate south_sweden
9476 VårA 1 CS78832 Oceanic climate germany
9481 Yst-1 CS78869 Oceanic climate south_sweden
9503 11C1 CS76640 Oceanic climate western_europe
9506 IP-Alo-0 CS76662 Mediterranean hot summer climates admixed
9507 IP-Coa-0 CS76775 Mediterranean hot summer climates spain
9508 IP-Mos-1 CS77108 Mediterranean hot summer climates admixed
9509 IP-Reg-0 CS77207 Mediterranean hot summer climates spain
9510 IP-Rei-0 CS77208 Mediterranean hot summer climates spain
9511 IP-Vav-0 CS78835 Mediterranean hot summer climates spain
9512 IP-Vid-1 CS78842 Mediterranean hot summer climates spain
9513 IP-Adc-5 CS76646 Mediterranean hot summer climates admixed
Compute results >

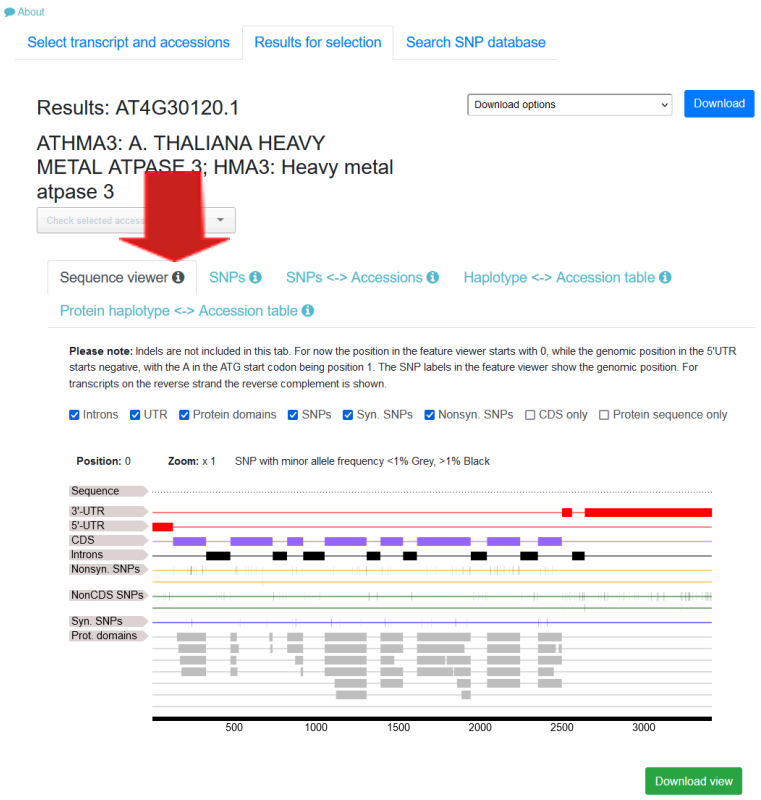
It’s possible to add more ‘Countries’ when you repeat the procedure with other ‘countries’. You can remove specific regions if you select the ‘Country’ of interest and click ‘Remove the selection’ (see chapter 2.2.1).

2.3 Compute results

To calculate the result of your selections press ‘Compute results’.



If you hover over any ⓘ symbol, additional information (e.g., for tabs or table headers) are shown.



2.4 Results for selection

If you compute the results SNPstar shifts directly to the section ‘Results for selection’.

Here, you have different options to display your results as described in the following chapters.

2.4.1 Sequence viewer

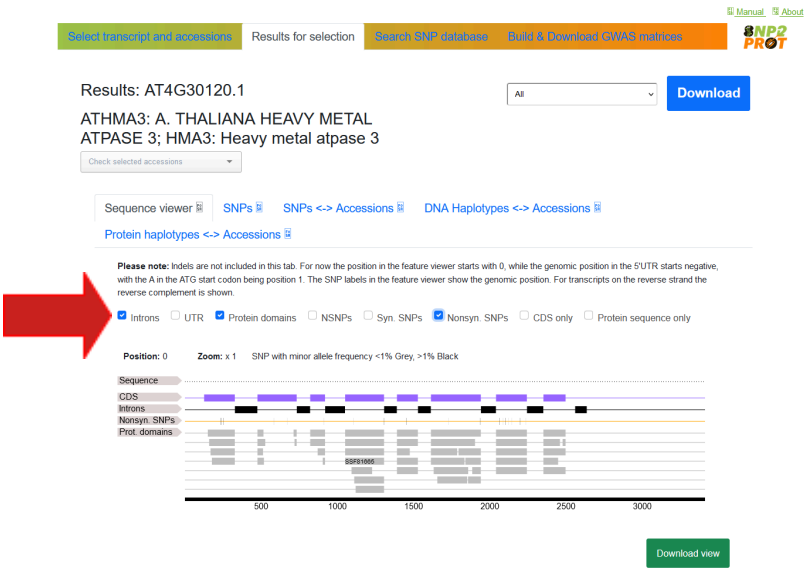
The first tab of results is the ‘Sequence viewer’.

Top: Graphical overview of SNP positions in relation to nucleotide and amino acid sequence features like UTRs, intron, protein domains. Bottom: Plain genomic, coding and amino acid sequence with SNPs color coded by SNP type. Amino acid substitution (not context dependent but for single SNPs).
All options and possibilities are described in the following.

You have different possible selections for ‘Sequence viewer’. You can mark none or all of the available genomic features for display.

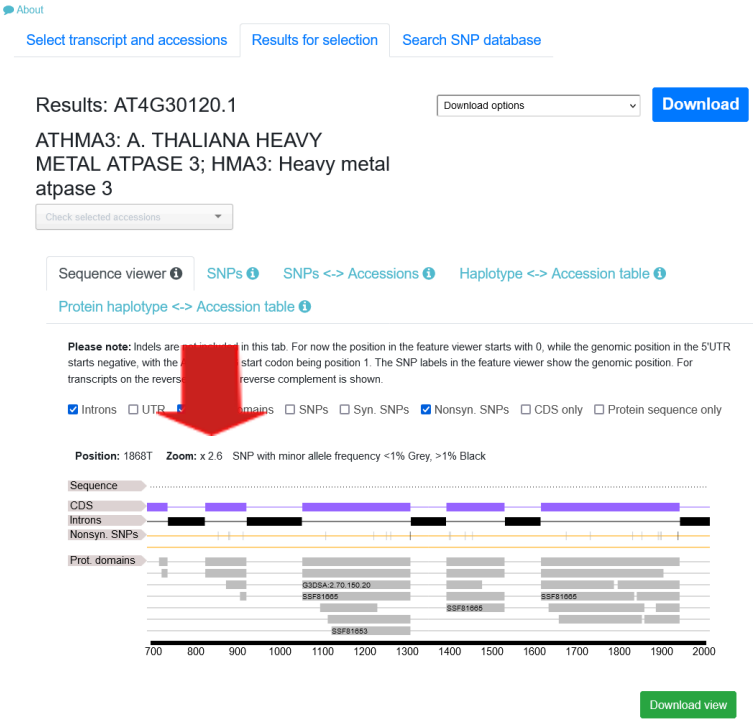
nothing selected: shown is the whole sequence in base pairs (bp) and coding sequence (CDS)
Introns: shown is the whole sequence (in bp), CDS and introns of selected gene
UTR (untranslated region): shown is the whole sequence (in bp), 3'-UTR, 5'-UTR and coding sequence (CDS) of selected gene
Prot. domains (protein domains): shown is the whole sequence (in bp), coding sequence (CDS) and protein domains of selected gene
Noncoding SNPs (Noncoding single nucleotide polymorphisms): shown is the whole sequence (in bp), coding sequence (CDS) and SNPs in non coding regions (NonCDS SNPs)
Syn. SNPs (Synonymous SNPs): shown is the whole sequence (in bp), coding sequence (CDS) and synonymous SNPs in coding regions
Nonsyn. SNPs (Nonsynonymous SNPs): shown is the whole sequence (in bp), coding sequence (CDS) and nonsynonymous SNPs in coding regions
CDS only (coding sequence only): shown is the whole sequence (in bp), nonsynonymous SNPs, synonymous SNPs and protein domains
Protein sequence only: shown is the whole sequence (in amino acids [AA]), nonsynonymous SNPs, synonymous SNPs and protein domains

To show multiple features, mark all elements of interest. As an example shown are introns, protein domains and nonsynonymous SNPs.



It's possible to zoom in. Click and hold the left mouse button to mark a region of interest.

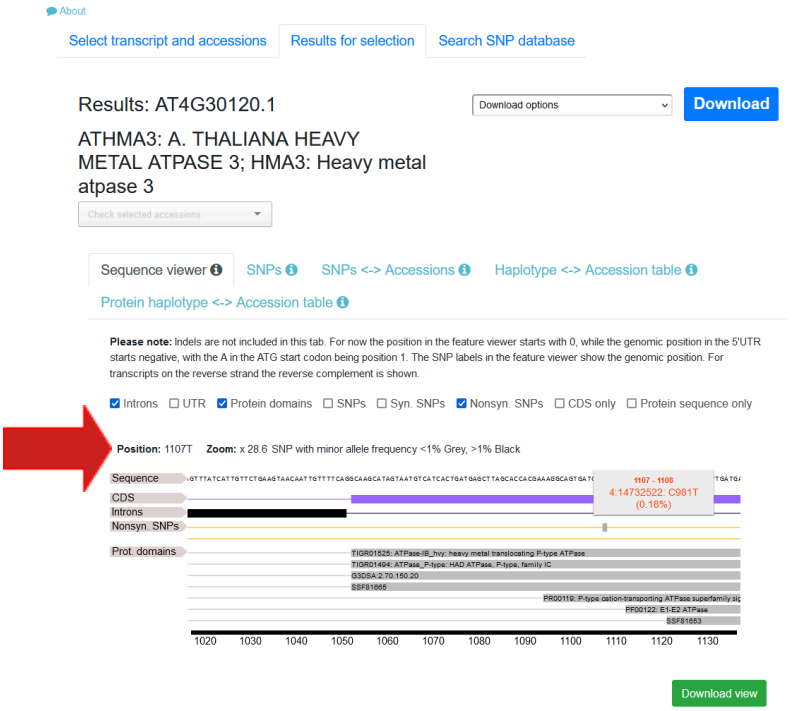
The zoom factor is shown as well.



To zoom out (to zoom factor 1) click with the right mouse button on the figure.

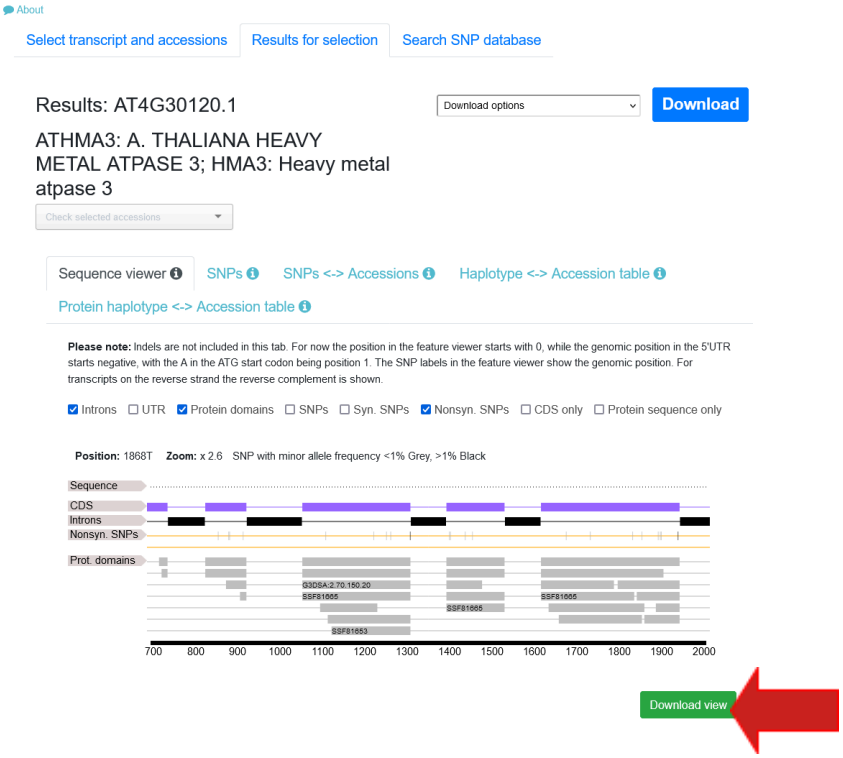
SNPs are displayed as vertical lines. Gray color indicates a SNP with a minor allele frequency <1%, while and black vertical lines indicate SNPs with a minor allele frequency >1%.

If you hover with the mouse cursor over a SNP the sequence position (in bp) and specific base are shown (e.g. 1107T).



With the mouse cursor on a specific SNP the position in bp (1107 - 1108), chromosome:whole chromosome position (4:14732522), polymorphism with CDS position in bp (C981T) and minor allele frequency (0.18%) are shown.

You can download the ‘Sequence viewer’ figure (as svg file) by clicking the ‘Download view’ button.



Beneath of the above described figure, you can find DNA or amino acid sequence views.

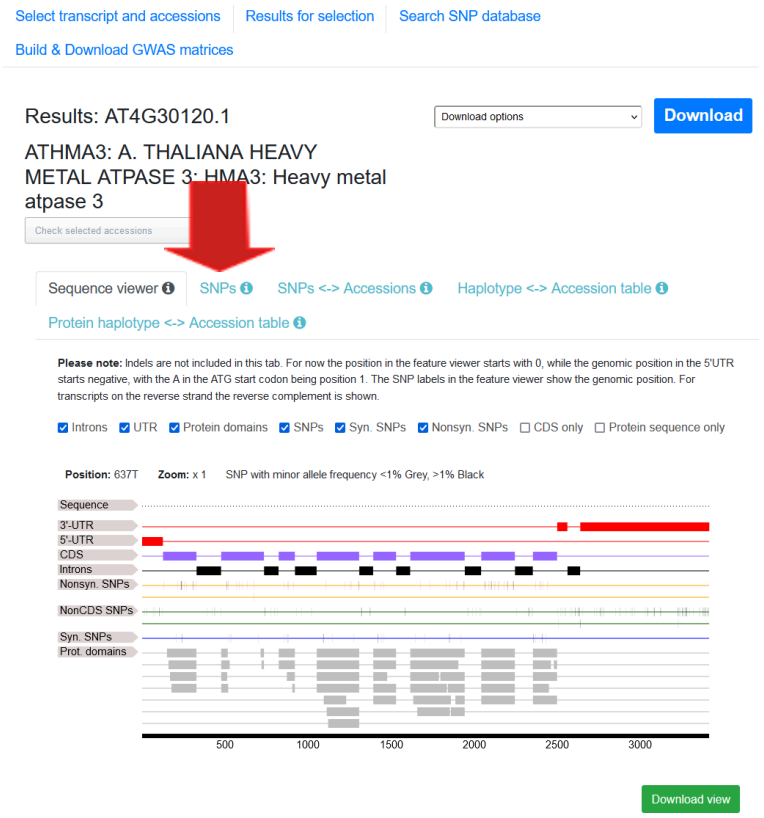
There are three different sequence types available.
genomic sequence: shows the whole sequence region of selected gene
CDS: shows only coding sequence of selected gene
Amino acid sequence: shows AA sequence of selected protein

The different colors indicate (for all three sequence types):
black letters: intron sequence
purple letters: CDS
red letters: 5’ and 3’ UTR sequence
yellow background: Nonsynonymous SNP
blue background: Synonymous SNP
green background: a SNP in a UTR or intron region

2.4.2 SNPs

The next tab of results is ‘SNPs’. This part describes the characteristics of all detected SNPs in detail.

SNPs : Information about each polymorphic position in the selected transcript computed for the selected accessions (GWAS, always for all accessions); e.g. base frequencies, protein description, amino acid frequencies, GWAS results etc.
All options and possibilities are described in the following.



SNP_ID (Genomic position): Chromosome number:genomic position e.g. 4:14730221. SNP identification number. Shown are chromosome:bp position (4:14730221) of each SNP of the selected gene. The empty field below the header can be used for a specific SNP ID search.

Base frequencies: You can sort according to the highest base frequency for each SNP and you can input a cutoff for the lowest allowed minor allele frequency. Shown are polymorphisms (base → base) and the frequencies (number behind) specific SNPs. The polymorphism with the same bases (e.g. C → C) is for the reference genotype (Col-0).

Genic position: Position 1, is the A in the ATG start codon. The 5’-UTR has negative indices. For isoforms on the reverse strand we count the same position, but in the reverse complement. SNP base pair position of selected gene gDNA (genomic DNA).

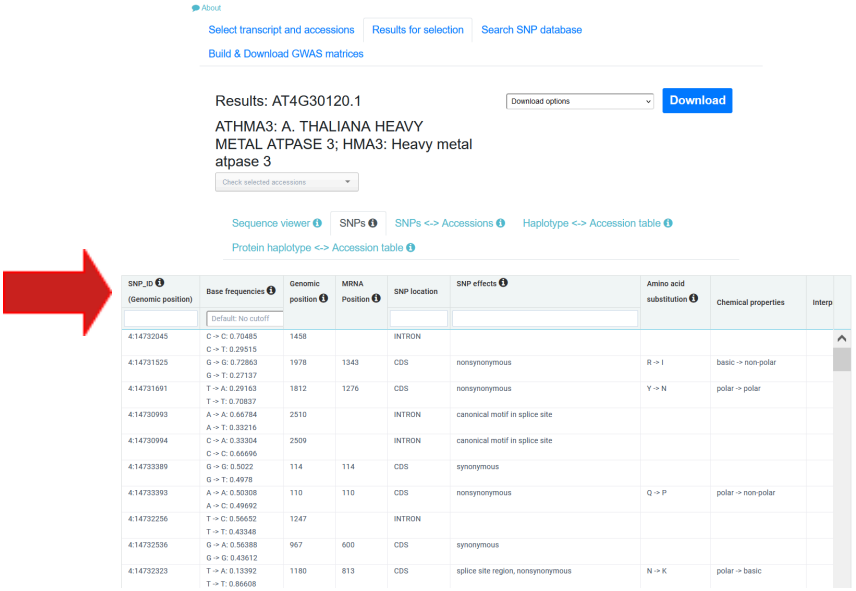
mRNA Position: Position 1 is the A in the ATG start codon. The 5’-UTR has negative indices. Introns are not included. For isoforms on the reverse strand we count the same position, but in the reverse complement. SNP base pair position of selected gene mRNA.

SNP location: Gene structure location (e.g. 3’ UTR (untranslated region), INTRON etc.) of SNP and selected gene.

SNP effects: This includes the effect of the SNP at this position in the reference, ignoring other SNPs in the neighborhood. Canonical motif in splice site means the first dinucleotides in the intron of the splice site (donor GT, acceptor AG). Splice site region describes the region 1 – 3 in the flanking exon and the positions 3 – 8 in the flanking intron of the splice site. Effects are equally defined to SnpEff (A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3., Cigolani P; Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu W, Ruden DM. Fly (Austin): 2012). Effect of SNP on AA sequence of selected gene (e.g. stop codon (premature stop codon), nonsynonymous (AA substitution), canonical motif in splice site etc.)

Amino acid substitution: Amino acid exchange caused by a nonsynonymous SNP in the codon (example: Q -> H means a glutamine in the reference sequence was substituted by a histidine in the SNP harboring allele).

Chemical properties: Physicochemical properties of AA substitution (e.g. non-polar → polar, basic → acidic etc.)

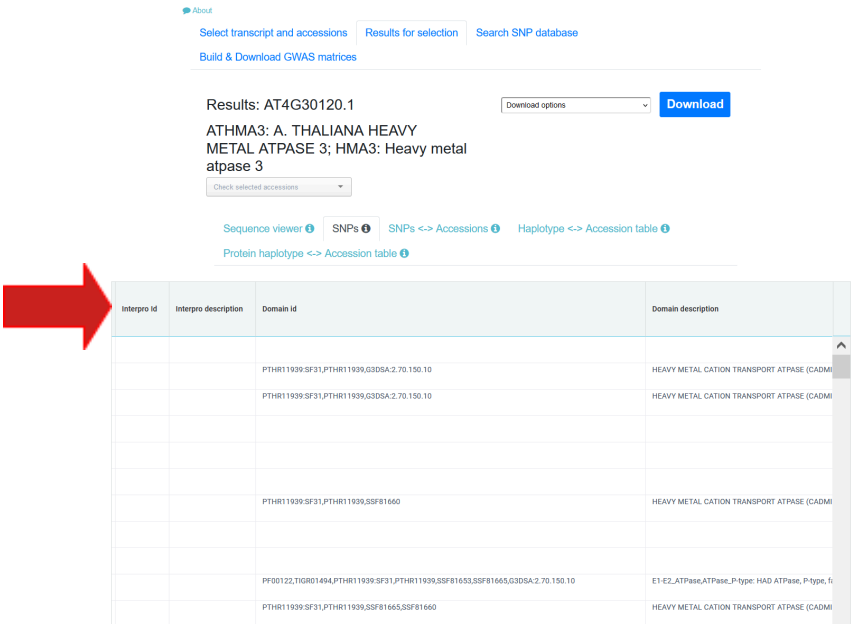


Interpro ID: Protein domain ID in which the AA substitutions are located. This information is based on the InterPro database (functional analysis of proteins by classifying them into families and predicting domains and important sites).

Interpro description: Description of InterPro domain

Domain id: Protein domain ID in which the AA substitutions are located. This information is based on Interproscan.

Domain description: Description of domain.



Strand: Strand direction of sequence (e.g. - and +).

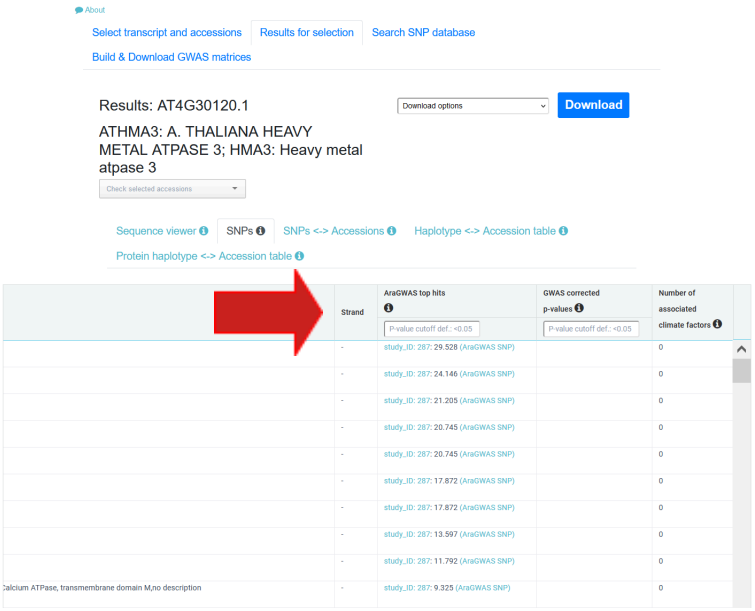
AraGWAS top hits: Top associations imported from the AraGWAS tool as of 10/11/20242024. The score is -log10(p-value) of the hit. The link links to the study on AraGWAS. The table can be sorted according to the lowest score in each cell. e.g. study_ID : 29.528 (AraGWAS SNP)

AraGWAS Catalog link of study information of selected SNP and gene.
AraGWAS score (-$\log_{10}$ (*p*-value)) of selected SNP and gene.
AraGWAS Catalog link of GWAS information of selected SNP and gene.

AraGWAS data based on the AraGWAS Catalog, which is a publicly and manually curated database for standardized GWAS (genome wide association study) results for *Arabidopsis thaliana*. Database is published on the 1001 genome homepage (https://aragwas.1001genomes.org/#/).

GWAS corrected p-values: GWAS analysis was performed with the software GEMMA. Every climate factor was used as a phenotype and for every climate factor, one analysis run was carried out. Only SNPs with a minor allele frequency of 1% or higher were analysed. The obtained p-values were corrected for multiple comparisons using the Benjamini-Yekutiel method. In this column you can see all the correlations between SNPs and the 200 climate factors which have a corrected p-value lower than or equal to the threshold. You can sort the column according to the lowest p-value in each cell.

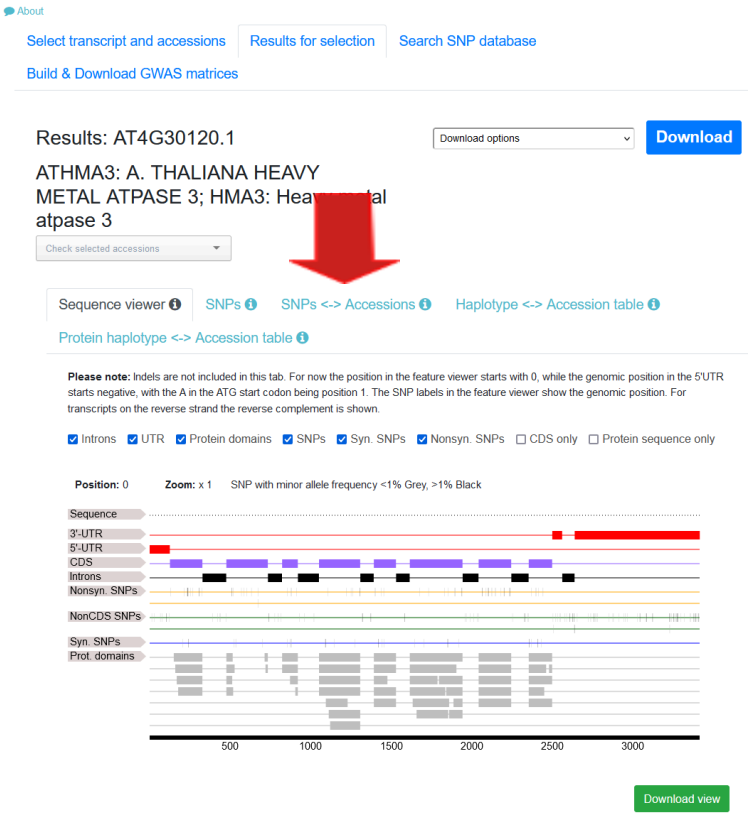
Number of associated climate factors: Number of associated climate factors with a corrected P-value in the GWAS lower than or equal to the threshold.



2.4.3 SNPs ↔ Accessions

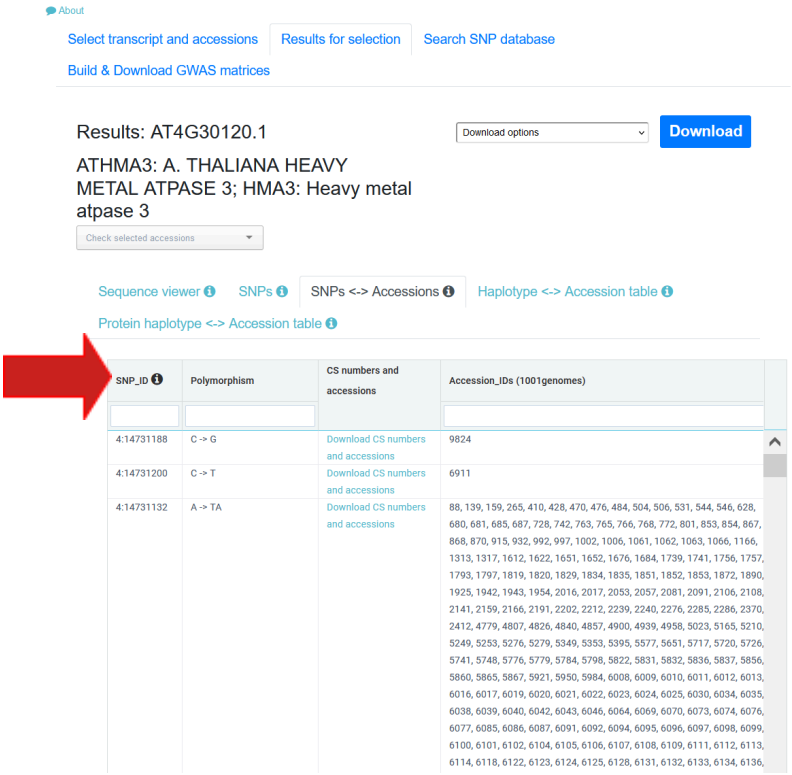
The tab ‘SNPs ↔ Accessions’ shows a table for SNP/accession associations.

SNPs ← Accessions : Information about each polymorphism in the selected transcript computed for the selected accession. List of accessions in which the polymorphism exists, effect of the polymorphism, amino acid change and SNP quality.
All options and possibilities are described in the following.



All empty fields beneath column titles can be used for specific searching (it is important to use the correct format!)

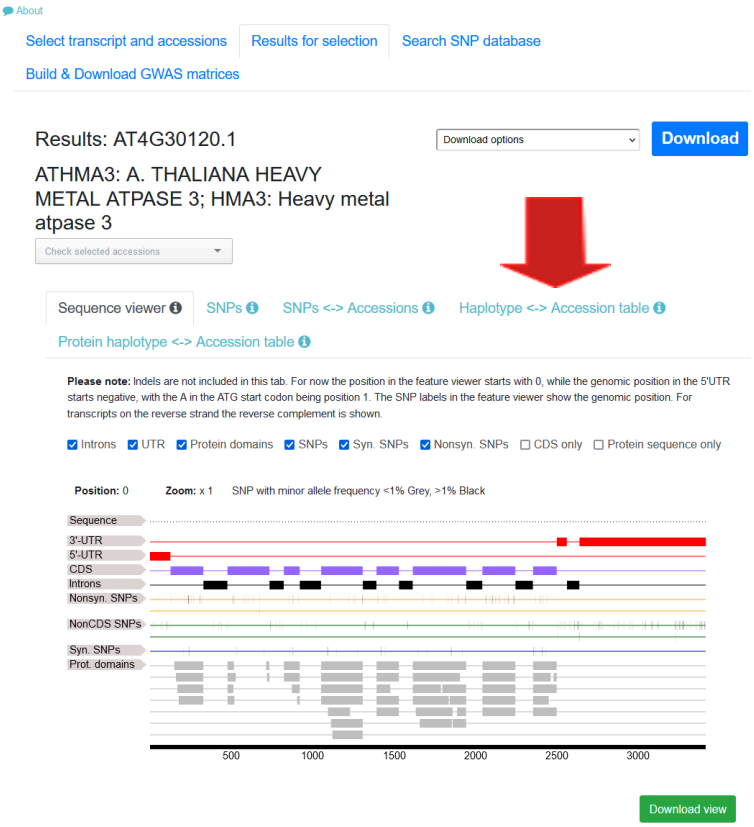
SNP Id: Chr.Pos e.g. 4:14731188. SNP identification number. Shown are chromosome:bp positions (4:14731188) of each SNP of the selected gene.
Polymorphism: List of polymorphisms (SNPs or InDels). For example: C→ G, always the first base (C) is from Col-0 (reference genotype) → the second base (G) from other accessions.
CS numbers and accessions: Download link for a list of genotypes different to Col-0. Shown are accessions with 1001 genome project Id, accession name and CS numbers.
Accession_IDs (1001 genomes): List of genotypes different to Col-0. Shown are accession IDs from 1001 Genome project.



2.4.4 DNA Haplotype ↔ Accessions

The tab ‘DNA Haplotype ↔ Accessions’ shows a table for DNA haplotype/accession associations.

DNA Haplotype ↔ Accessions : Accessions with the same polymorphisms in the selected CDS of the selected transcript have the same transcript specific haplotype. All options and possibilities are described in the following.



All empty fields beneath column titles can be used for specific searching (use the correct format!)

Haplotype color: Color of each DNA haplotype which is shown in the map of all haplotypes (map is beneath the table!)

Haplotype_ID: Transcript Id.Haplotype number_dna (e.g. AT4G30120.1_dna).
AT4G30120.1 1 dna

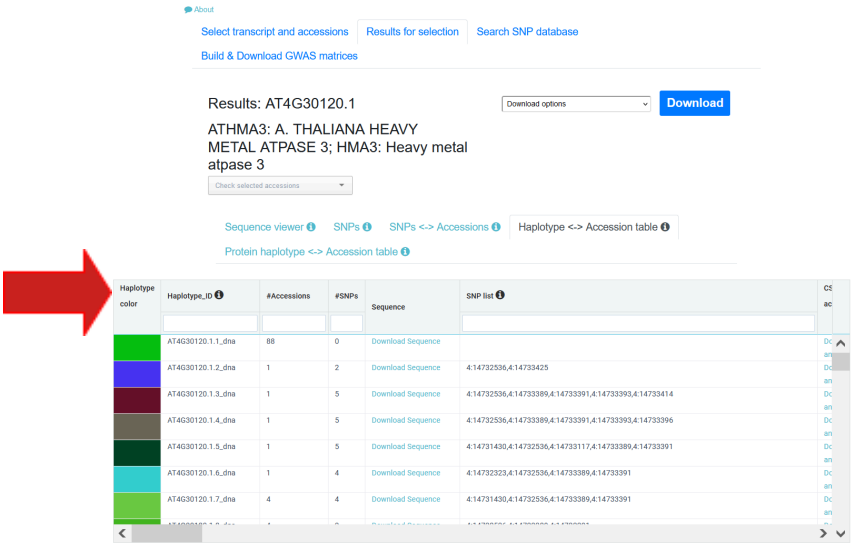
Gene ID haplotype number type of sequence

 # Accessions: Count of accessions of each haplotype.

 # SNPs: Count of SNPs of each haplotype compared to reference (Col-0). The haplotype group with a SNP number of 0 is the reference haplotype group (accession haplotype group with Col-0).

Sequence: Download link of CDS haplotype sequence of each haplotype (FASTA file).

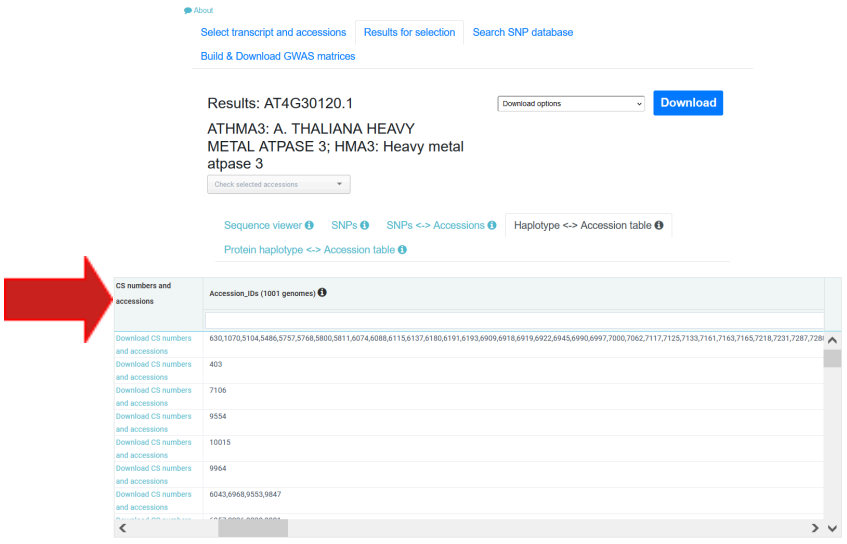
SNP list: A list of the SNPs/InDels that differentiate this haplotype from the haplotype of the reference accession (Col-0). List of each SNP per each haplotype. Shown are SNP IDs in chromosome:bp position (4:14732536) of each SNP.



Map of all haplotypes

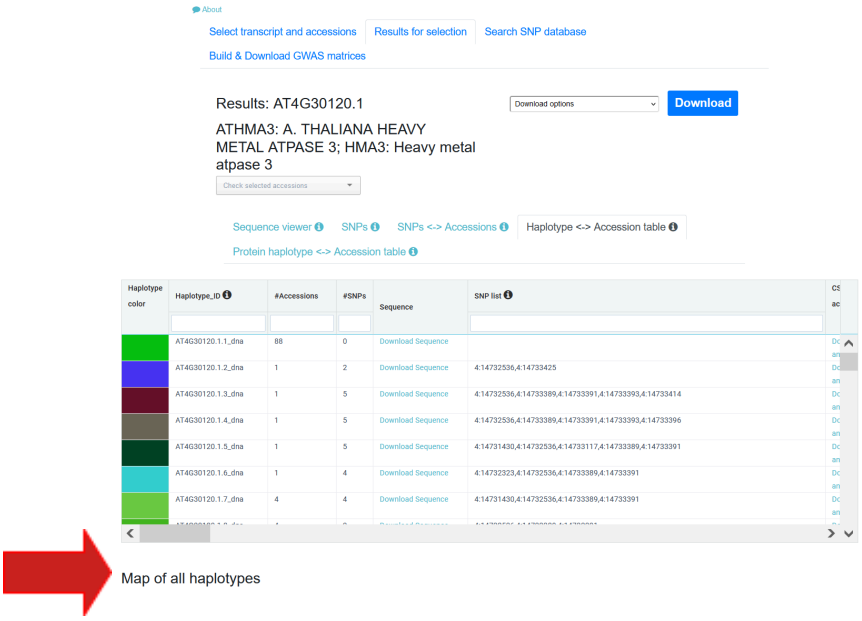
CS numbers and accessions: Download link for all accessions of specific haplotype. In the table 1001 genome ID, genotype name and CS number are shown.

Accession_IDs (1001 genomes): The haplotype of one transcript might be the same for several accessions. So for each haplotype there is an accession list with the different accession IDs. The Accession_IDs originate from the 1001 genomes project. Genotype list of all haplotype groups of selected gene. Shown are accession IDs from the 1001 Genomes project.

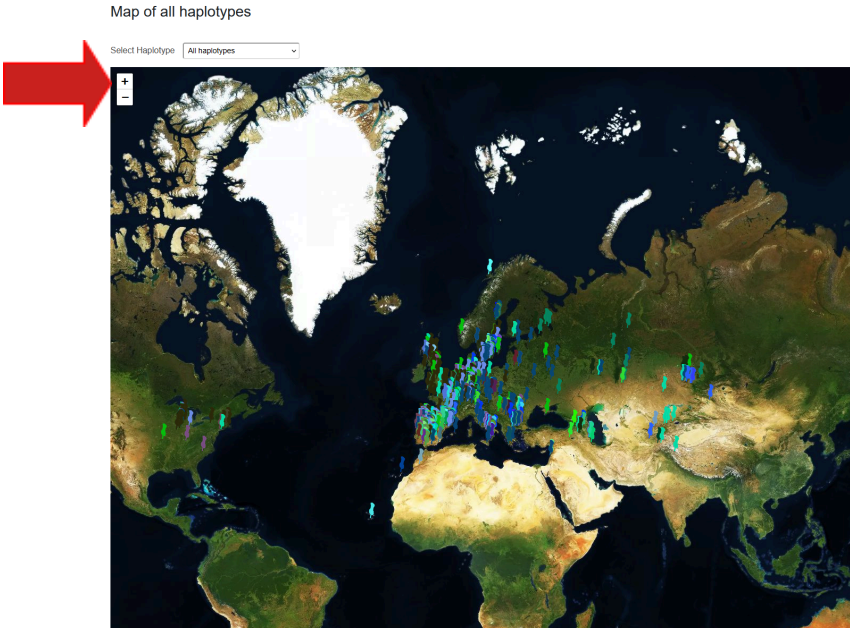


Map of all haplotypes

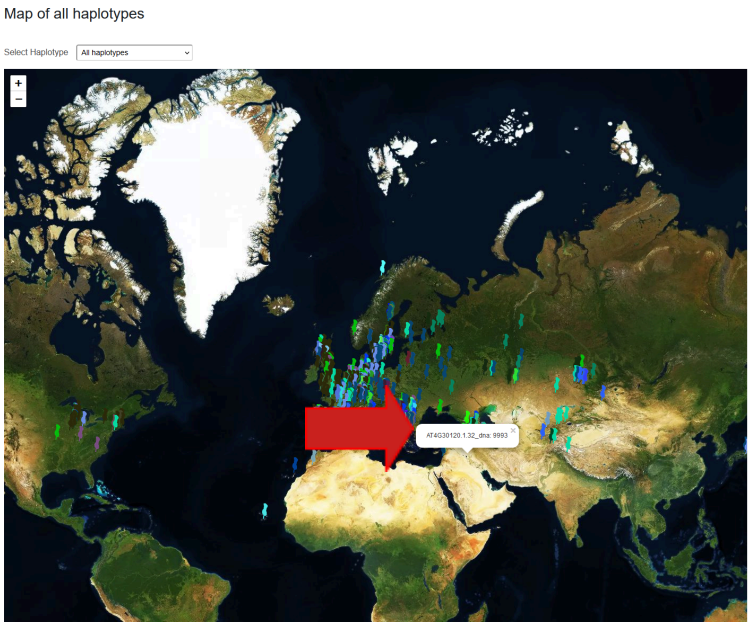
Beneath of the ‘DNA Haplotype ↔ Accessions’ table a ‘Map of all haplotypes’ is shown.



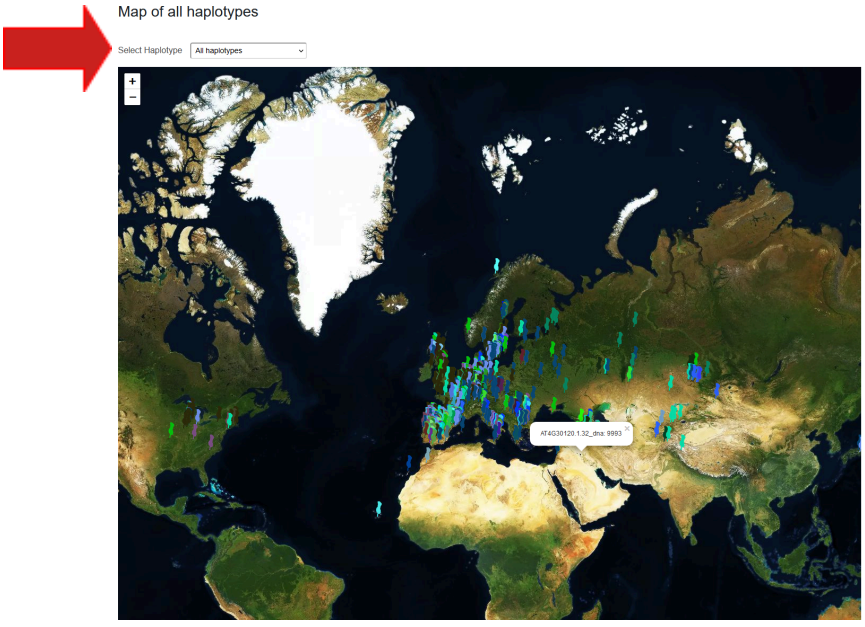
The ‘Map of all haplotypes’ shows the geographic origin of all selected accessions. Each color is specific for a group of haplotypes. The zoom works via the buttons in the left upper corner or the mouse wheel. With holding the left mouse button and moving the mouse cursor the map can be shifted.



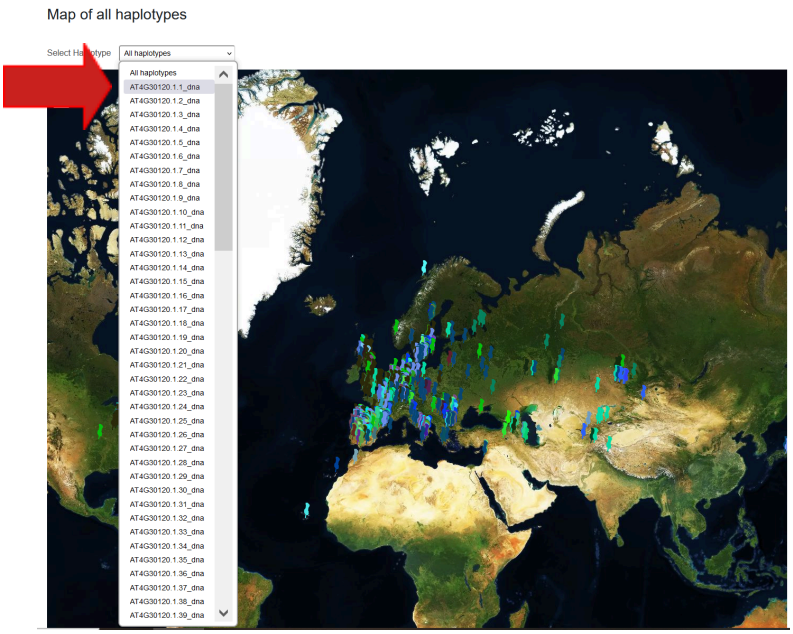
Left mouse click on a mark of the map shows haplotype and accession IDs (1001 genome ID). Both parameters are separated via a double colon (:).



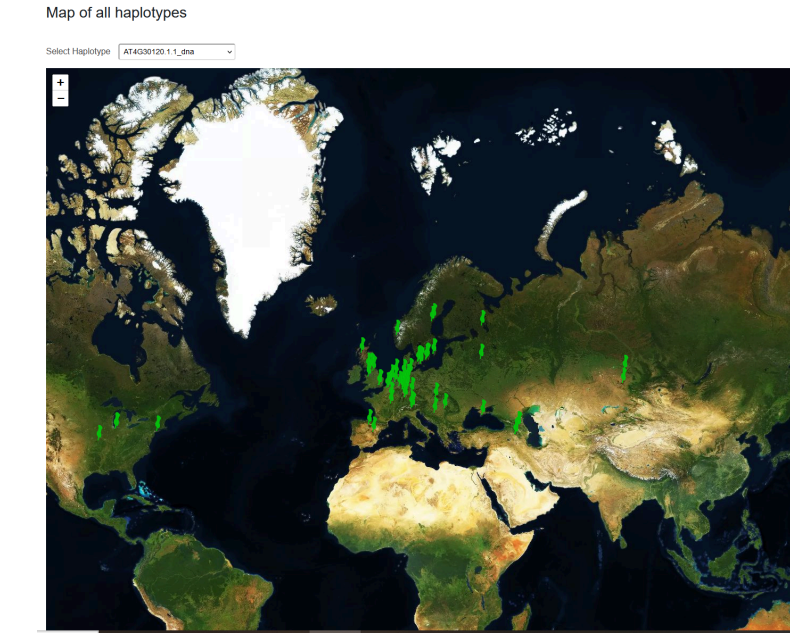
Select haplotype groups (accessions with the same haplotype) via the ‘Select Haplotype’ option.



Selection of a specific haplotype (haplotype ID).



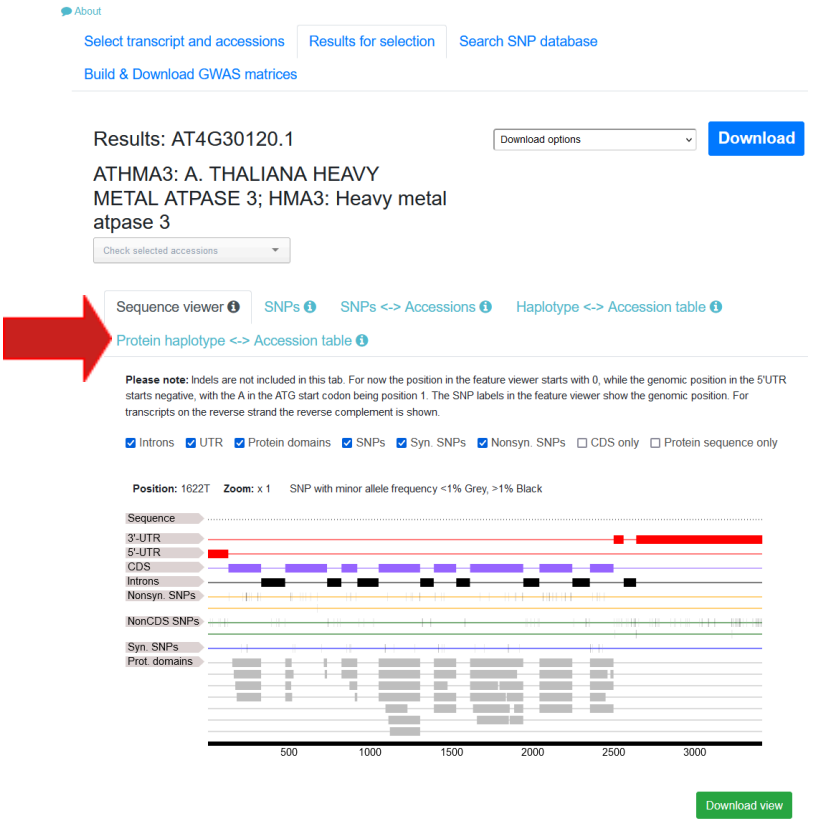
The geographic origins of all accessions within the haplotype AT4G30120.1.1_dna (haplotype ID) are shown as an example.



2.4.5 Proteotype ↔ Accession

The tab ‘Proteotype ↔ Accessions’ shows a table for protein haplotype/accession associations. The proteotype describes the protein haplotype.

Proteotype ← Accessions ℹ : Accessions with polymorphisms in the selected CDS of the selected transcript that result in the same protein sequence have the same haplotype.
All options and possibilities are described in the following.



All empty fields beneath column titles can be used for specific searching (use the correct format!)

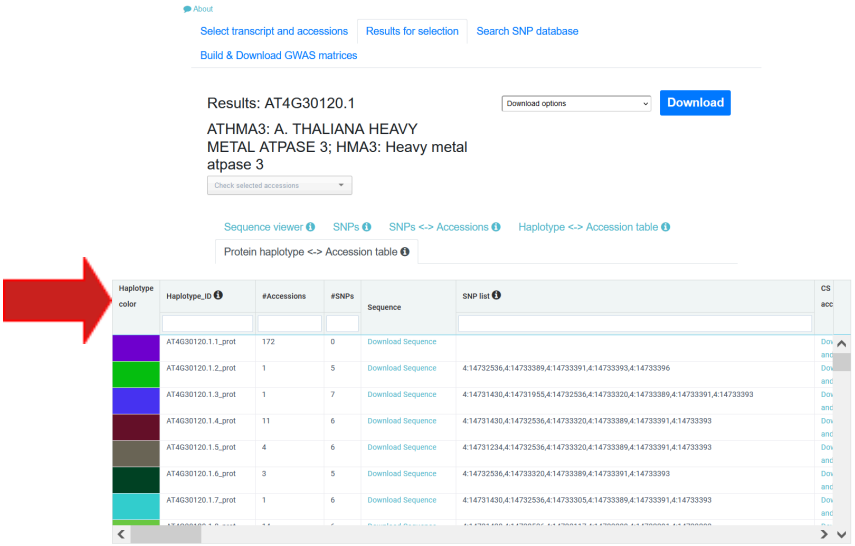
Proteotype color: Color of each AA haplotype which is shown in the map of all haplotypes (map is beneath the table!)

Proteotype_ID: Transcript Id.Haplotype number_prot (e.g. AT4G30120.1.1_prot).
AT4G30120.1.1_prot

Gene ID.haplotype number_type of sequence

 # Accessions: Count of accessions of each haplotype.
 # SNPs: Count of SNPs of each proteotype compared to reference (Col-0). The proteotype group with a SNP number of 0 is the reference proteotype group (accession proteotype group with Col-0).
Sequence: Download link of AA sequence of each proteotype (FASTA file).

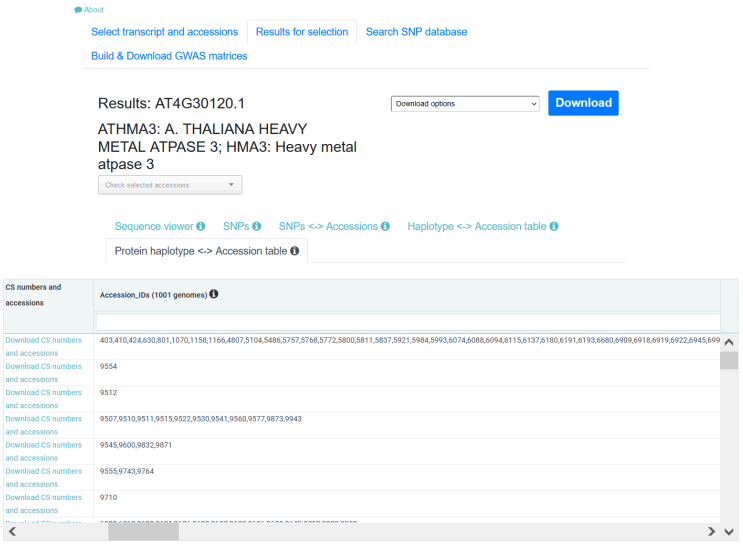
SNP list: A list of the SNPs/InDels that differentiate this proteotype from the proteotype of the reference accession (Col-0). List of SNPs causing AA substitutions per each proteotype. Shown are chromosome:bp positions (4:14732536) of each SNP.



Map of all haplotypes

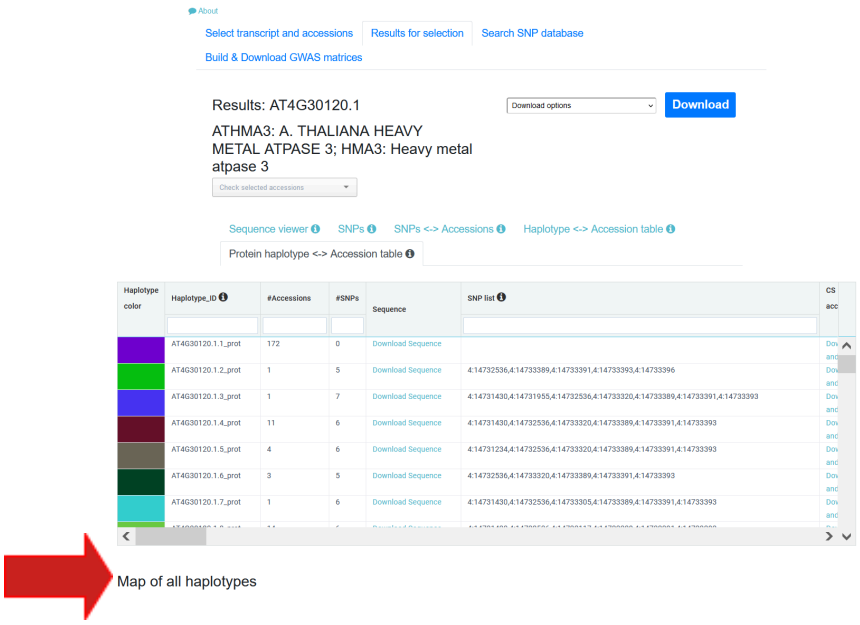
CS numbers and accessions: Download link for all accessions of specific haplotype. In the table 1001 genome Id, genotype name and CS number are shown.

Accession_IDs (1001 genomes): ℹ The proteotype of one transcript might be the same for several accessions. So far for each proteotype there is an accession list with the different accession IDs. The Accession_IDs originate from the 1001 genomes project. Genotype list of all proteotype groups of selected gene. Shown are accession IDs from the 1001 Genomes project.



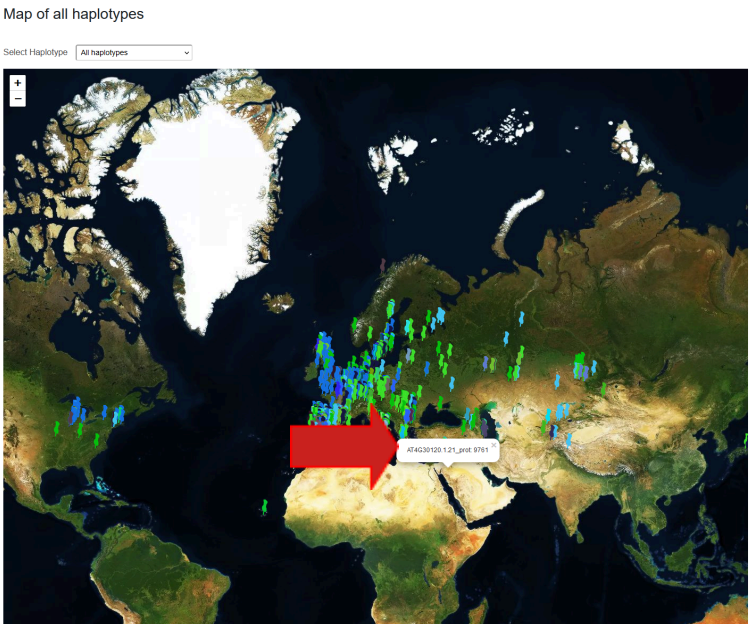
Map of all haplotypes

Map of all proteotypes is shown beneath the table.

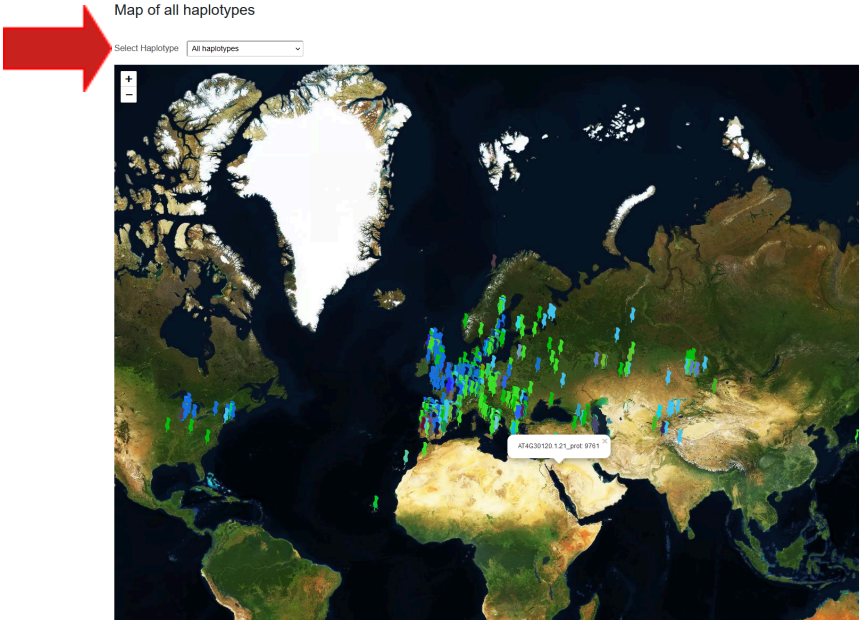


The ‘Map of all proteotypes’ shows the geographical origin of all selected accessions. Each color is specific for a group of proteotypes. The zoom works via the buttons in the left upper corner or the mouse wheel. With holding the left mouse button and moving the mouse cursor the map can be shifted.

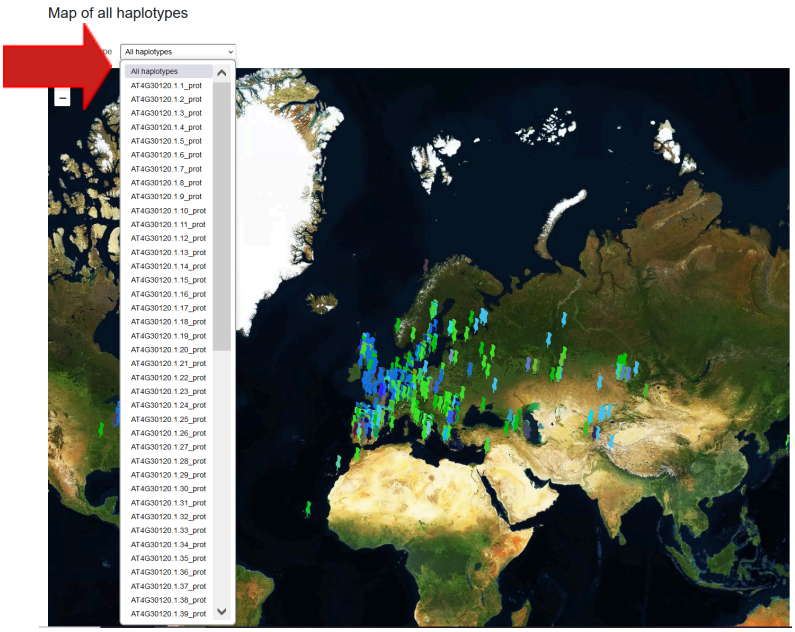
Left mouse click on a mark of the map shows protein haplotype and accession IDs (1001 Genomes ID). Both parameters are separated via a space.



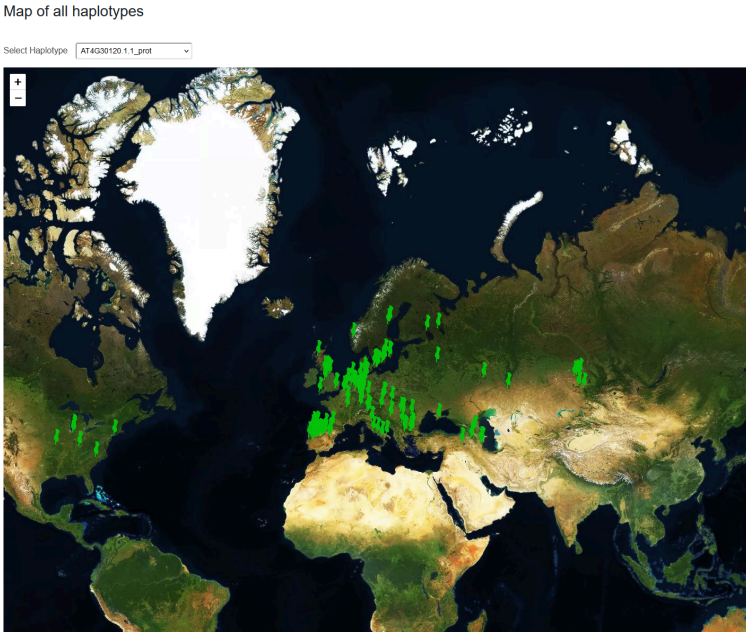
Select protein haplotype groups (accessions with the same protein haplotype) via ‘Select Proteotype’ option.



Selection of a specific proteotype (protetype ID).



The origin of all accessions from the protein haplotype AT4G30120.1.1_prot group (proteotype ID) are shown as an example.

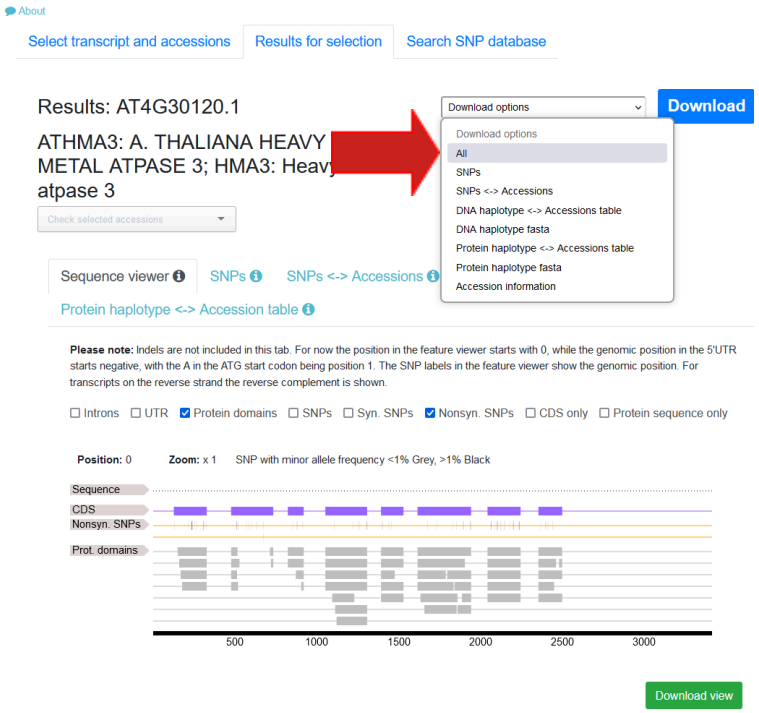


2.4.6 Download options

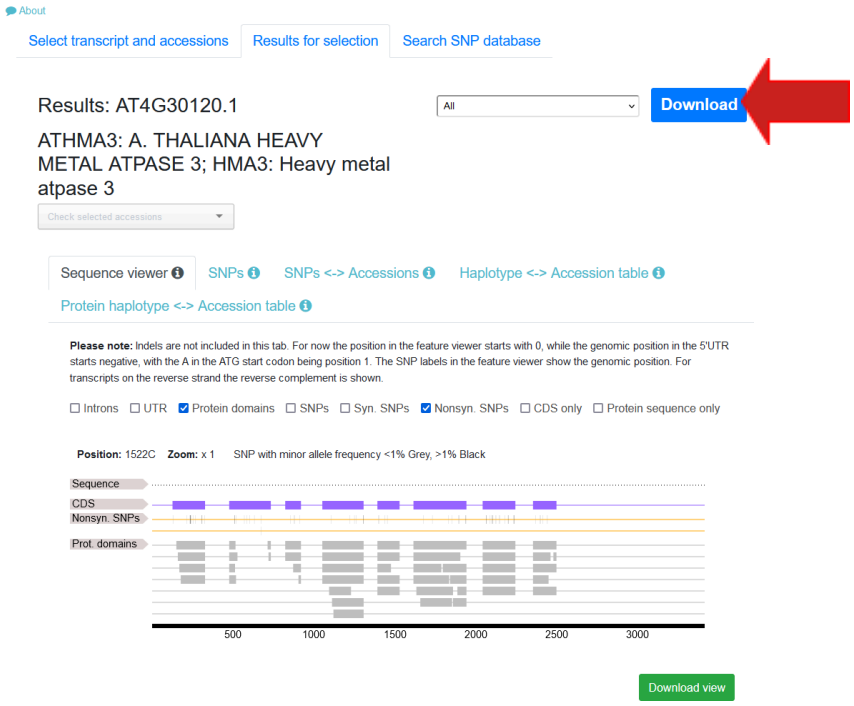
In this chapter we describe download options for the data from the result tables.

2.4.6.1 Download all

To download all available result tables, open the drop-down menu on the upper right corner of the ‘Results for selection’ tab. Select the option ‘All’.



Press the ‘Download’ button.



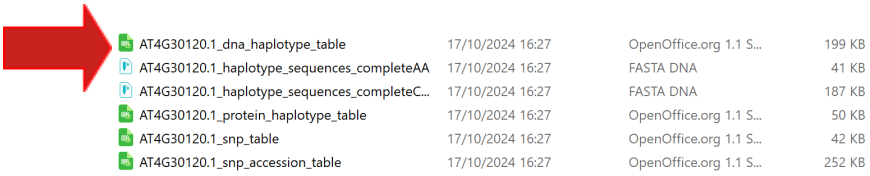
For all other download options select your file of interest and download it as described above.
For the description of the download options and the corresponding output files, see the following manual sections:

- For ‘DNA haplotype ↔ Accession table’ download see chapter 2.4.6.2.
- For ‘Protein haplotype fasta’ download see chapter 2.4.6.3.

- For ‘DNA haplotype fasta’ download see chapter 2.4.6.4.
- For ‘Protein haplotype ↔ Accession table’ download see chapter 2.4.6.5.
- For ‘SNPs’ download see chapter 2.4.6.6.
- For ‘SNPs ↔ Accessions’ download see chapter 2.4.6.7.
- For ‘Accession information’ download see chapter 2.4.6.8.

2.4.6.2 DNA haplotype table

The ‘dna_haplotype_table’ is dumped as a csv file.



DNA haplotype table shows:

haplotype _ID: AT4G30120.1_1_dna

Gene ID_haplotype number_type of sequence

number_of_accessions: Count of accessions of each haplotype.
list_of_snps: List of each SNP per each haplotype. Shown are chromosome:bp positions (4:14732536) of each SNP.

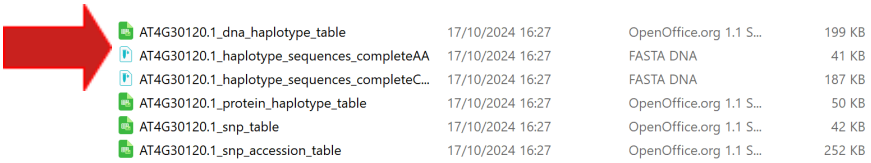
	A	B	C
1	#all	accessions	selected
2	haplotype_ID	number_of_accessions	list_of_snps
3	AT4G30120.1.1_dna	88	
4	AT4G30120.1.2_dna	1	4:14732536,4:14733425
5	AT4G30120.1.3_dna	1	4:14732536,4:14733389,4:14733391,4:14733393,4:14733414
6	AT4G30120.1.4_dna	1	4:14732536,4:14733389,4:14733391,4:14733393,4:14733396
7	AT4G30120.1.5_dna	1	4:14731430,4:14732536,4:14733117,4:14733389,4:14733391
8	AT4G30120.1.6_dna	1	4:14732323,4:14732536,4:14733389,4:14733391
9	AT4G30120.1.7_dna	4	4:14731430,4:14732536,4:14733389,4:14733391
10	AT4G30120.1.8_dna	4	4:14732536,4:14733389,4:14733391
11	AT4G30120.1.9_dna	1	4:14732536,4:14733110,4:14733389,4:14733391
12	AT4G30120.1.10_dna	1	4:14731525,4:14731691,4:14732176,4:14733044,4:14733376

list_accession_IDs: All accessions belonging to the specified haplotype.
haplotype_sequence_complete_CDS: Coding sequence of haplotypes.
length_haplotype_sequence: Sequence length of haplotypes in bp.

	D	E	F
1			
2	list_accession_IDs	haplotype_sequence_completeCDS	length_haplotype_sequence
3	630,1070,5104,5486,5757,5768,5800,5813	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629
4	403	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629
5	7106	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629
6	9554	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629
7	10015	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629
8	9964	ATGGCGGAAGGTGAAGAGTCAAAGAA	1629

2.4.6.3 Proteotype sequence complete AA

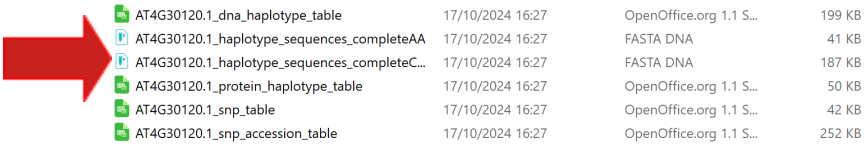
The ‘proteotype_sequence_completeAA’ is dumped as a FASTA file.



This file contains all AA sequences of each selected haplotype.

2.4.6.4 Haplotype sequences complete CDS

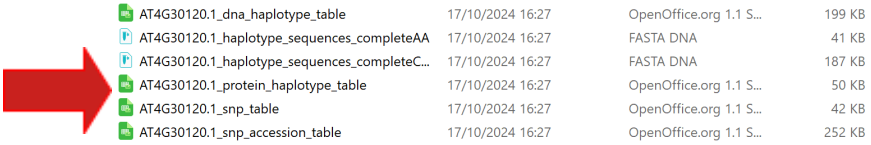
The ‘haplotype_sequence_completeCDS’ is dumped as a FASTA file.



This file contains all CDS sequences of each selected haplotype.

2.4.6.5 Proteotype table

The ‘proteotype_table’ is dumped as a csv file.



The proteotype table shows:

proteotype _ID: AT4G30120.1_1_prot

Gene ID_proteotype number_type of sequence

number_of_accessions: count of accessions of each haplotype
list_of_snps: list of each SNP per each proteotype per gene. Shown are chromosome:bp positions (4:14732536) of each SNP

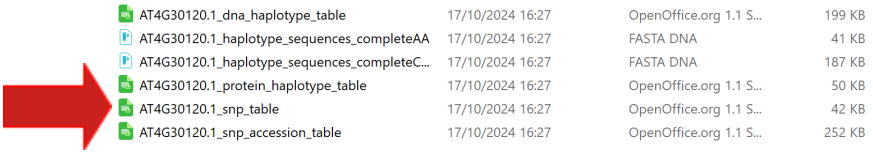
	A	B	C
1	#all	accessions	selected
2	haplotype_ID	number_of_accessions	list_of_snps
3	AT4G30120.1.1_prot	172	
4	AT4G30120.1.2_prot	1	4:14732536,4:14733389,4:14733391,4:14733393,4:14733396
5	AT4G30120.1.3_prot	1	4:14731430,4:14731955,4:14732536,4:14733320,4:14733389,4:14733391,4:14733393
6	AT4G30120.1.4_prot	11	4:14731430,4:14732536,4:14733320,4:14733389,4:14733391,4:14733393
7	AT4G30120.1.5_prot	4	4:14731234,4:14732536,4:14733320,4:14733389,4:14733391,4:14733393
8	AT4G30120.1.6_prot	3	4:14732536,4:14733320,4:14733389,4:14733391,4:14733393

list_accession_IDs: all accession belonging to the specified haplotype
proteotype_sequence_complete_AA: AA sequence of proteotype
length_proteotype_sequence: AA sequence length of proteotype

	D	E	F
1			
2	list_accession_IDs	haplotype_sequence_completeAA	length_haplotype_sequence
3	403,410,424,630,801,1070,1158,1166,4807,5104,5486,5757,576	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLRQVDGVKP	543
4	9554	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLHPLDGVKE	543
5	9512	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLRPLDGVKE	543
6	9507,9510,9511,9515,9522,9530,9541,9560,9577,9873,9943	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLRPLDGVKE	543
7	9545,9600,9832,9871	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLRPLDGVKE	543
8	9555,9743,9764	MAEGEESKKMNLQTSYFDVVGICCSSEVSIVGNVLRPLDGVKE	543

2.4.6.6 SNP table

The ‘SNP_table’ is dumped as a csv file.



The SNP table shows:

SNP ID: SNP identification number. Shown are chromosome:bp position (4:14730221) of each SNP of selected gene.
Variant_0: Base pair variant on a specific position with the highest frequency of selected gene.
Frequency_0: Base frequency of variant 0 from selected gene.
Variant_1: Base pair variant on a specific position with the second highest frequency of selected gene.
Frequency_1: Base frequency of variant 1 from selected gene.
Variant_2: Base pair variant on a specific position with the third highest frequency of selected gene.
Frequency_2: Base frequency of variant 2 from selected gene.
SNP relative position mRNA: SNP base pair position of selected gene mRNA.
SNP relative position genome: SNP base pair position of selected gene gDNA (genomic DNA).

	A	B	C	D	E	F	G	H	I
1	#all accessions selected								
2	SNP Id	Variant_0	Frequency_0	Variant_1	Frequency_1	Variant_2	Frequency_2	SNP relative position mrna	SNP relative position genome
3	4:14730221	C -> C	0.99824	C -> A	0.00176	noval	noval	2463	3282
4	4:14730223	T -> T	0.9859	T -> C	0.0141	noval	noval	2461	3280
5	4:14730234	C -> C	0.99912	C -> A	0.00088	noval	noval	2450	3269
6	4:14730241	G -> T	0.70749	G -> G	0.29251	noval	noval	2443	3262
7	4:14730247	A -> A	0.99031	A -> T	0.00969	noval	noval	2437	3256
8	4:14730252	T -> G	0.67489	T -> T	0.32511	noval	noval	2432	3251
9	4:14730270	C -> C	0.99912	C -> T	0.00088	noval	noval	2414	3233
10	4:14730277	C -> C	0.99912	C -> A	0.00088	noval	noval	2407	3226
11	4:14730312	G -> G	0.99824	G -> A	0.00176	noval	noval	2372	3191
12	4:14730332	A -> A	0.99912	A -> G	0.00088	noval	noval	2352	3171
13	4:14730350	C -> C	0.9815	C -> T	0.0185	noval	noval	2334	3153
14	4:14730354	A -> A	0.95419	A -> G	0.04581	noval	noval	2330	3149
15	4:14730358	T -> T	0.99648	T -> A	0.00352	noval	noval	2326	3145
16	4:14730360	A -> A	0.99648	A -> T	0.00352	noval	noval	2324	3143

SNP location: Gene structure location (e.g. 3' UTR, INTRON etc.) of SNP and selected gene.
SNP effects: Effect of SNP on AA sequence of selected gene (e.g. stop codon (premature stop codon), nonsynonymous (AA substitution), canonical motif in splice site etc.)
Substitution_0: AA substitution on a specific position with the highest frequency of selected gene.
Substitution_frequency_0: AA substitution of substitution 0 from selected gene.
Substitution_1: AA substitution on a specific position with the second highest frequency of selected gene.
Substitution_frequency_1: AA substitution of substitution 1 from selected gene.
Substitution_properties: Physico-chemical properties of AA substitution (e.g. non-polar → polar, basic → acidic etc.)
Domain Id: Protein domain ID in which the AA substitutions are located.

	J	K	L	M	N	O	P	Q
1								
2	SNP location	SNP effects	Substitution_0	Substitution_frequency_0	Substitution_1	Substitution_frequency_1	Substitution properties	Domain Id
51	INTRON	noval	noval	noval	noval	noval	noval	noval
52	INTRON	noval	noval	noval	noval	noval	noval	noval
53	INTRON	noval	noval	noval	noval	noval	noval	noval
54	3UTR	noval	noval	noval	noval	noval	noval	noval
55	3UTR	noval	noval	noval	noval	noval	noval	noval
56	3UTR	noval	noval	noval	noval	noval	noval	noval
57	3UTR	noval	noval	noval	noval	noval	noval	noval
58	3UTR	noval	noval	noval	noval	noval	noval	noval
59	3UTR	noval	noval	noval	noval	noval	noval	noval
60	CDS	stop codon	stop codon	frame shift	synonymous	noval	noval	noval
61	CDS	nonsynonymous	N -> K	0.00088	noval	noval	polar -> basic	PTHR11939:SF31
62	CDS	synonymous	noval	noval	noval	noval	noval	noval
63	CDS	nonsynonymous	M -> T	0.00088	noval	noval	non-polar -> polar	PTHR11939:SF31
64	CDS	synonymous	noval	noval	noval	noval	noval	noval
65	CDS	nonsynonymous	G -> D	0.00088	noval	noval	non-polar -> acidic	PTHR11939:SF31
66	CDS	nonsynonymous	T -> N	0.00617	noval	noval	polar -> polar	PTHR11939:SF31
67	CDS	synonymous	noval	noval	noval	noval	noval	noval
68	CDS	nonsynonymous	I -> T	0.00088	noval	noval	non-polar -> polar	PTHR11939:SF31
69	CDS	splice site region	noval	noval	noval	noval	noval	noval
70	INTRON	canonical motif in splice site	noval	noval	noval	noval	noval	noval
71	INTRON	splice site region	noval	noval	noval	noval	noval	noval

Domain description: Description of domain.

	R
1	
2	Domain description
51	noval
52	noval
53	noval
54	noval
55	noval
56	noval
57	noval
58	noval
59	noval
60	noval
61	HEAVY METAL CATION TRANSPORT ATPASE (CADMIUM/ZINC TRANSPORTING ATPASE),CATION-TRANSPORTING ATPASE,HMA_2,HMA, heavy metal-associated domain
62	noval
63	HEAVY METAL CATION TRANSPORT ATPASE (CADMIUM/ZINC TRANSPORTING ATPASE),CATION-TRANSPORTING ATPASE,HMA_2,HMA, heavy metal-associated domain
64	noval
65	HEAVY METAL CATION TRANSPORT ATPASE (CADMIUM/ZINC TRANSPORTING ATPASE),CATION-TRANSPORTING ATPASE,HMA_2,HMA, heavy metal-associated domain
66	HEAVY METAL CATION TRANSPORT ATPASE (CADMIUM/ZINC TRANSPORTING ATPASE),CATION-TRANSPORTING ATPASE,HMA_2,HMA, heavy metal-associated domain
67	noval
68	HEAVY METAL CATION TRANSPORT ATPASE (CADMIUM/ZINC TRANSPORTING ATPASE),CATION-TRANSPORTING ATPASE,HMA_2,HMA, heavy metal-associated domain
69	noval
70	noval
71	noval

Strand: Strand direction (e.g. - and +) of sequence.
AraGWAS_score: AraGWAS score (-$\log_{10}$ (p-value)) of selected SNP and gene.
AraGWAS_links: AraGWAS Catalog links of GWAS and study information of selected SNP and gene.

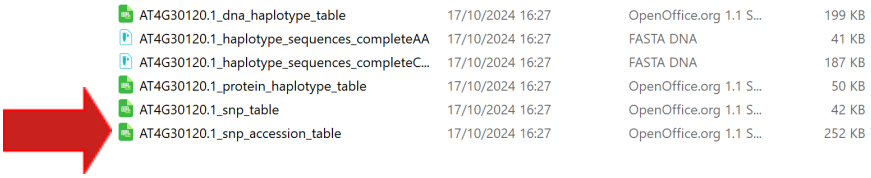
	S	T	U
1			
2	Strand	AraGWAS_score	AraGWAS_links
81	-	noval	noval
82	-	noval	noval
83	-	noval	noval
84	-	noval	noval
85	-	24.1462113508663	SNP_link: https://aragwas.1001genomes.org/#/study/287/associations/4_14731525, study_link: https://aragwas.1001genomes.org/#/study/287, score: 24.1
86	-	noval	noval
87	-	noval	noval
88	-	noval	noval
89	-	noval	noval
90	-	noval	noval
91	-	noval	noval
92	-	noval	noval
93	-	noval	noval
94	-	noval	noval
95	-	21.204987069462	SNP_link: https://aragwas.1001genomes.org/#/study/287/associations/4_14731691, study_link: https://aragwas.1001genomes.org/#/study/287, score: 21.2

GWAS corrected: GWAS analysis was performed with the software GEMMA. Every climate factor was used as a phenotype and for every climate factor, one analysis run was carried out. Only SNPs with a minor allele frequency of 1% or higher were analysed. The obtained p-values were corrected for multiple comparisons using the Benjamini-Yekutiel method. In this column you can see all the correlations between SNPs and the 200 climate factors which have a corrected p-value lower than or equal to the threshold. You can sort the column according to the lowest p-value in each cell.
Number of associated climate factors: Number of associated climate factors, with a corrected P-value in the GWAS lower than or equal to the threshold.

	V	W
1		
2	GWAS corrected	Number of associated climate factors
81	noval	0
82	noval	0
83	noval	0
84	noval	0
85	noval	0
86	noval	0
87	noval	0
88	noval	0
89	noval	0
90	noval	0
91	noval	0
92	noval	0
93	noval	0
94	noval	0
95	noval	0
96	noval	0

2.4.6.7 SNP accession table

The ‘SNP_accession_table’ is dumped as a csv file.



The SNP accession table shows:

SNP ID: List of SNP IDs of selected gene. Shown are chromosome:bp position (4:14731188) of each SNP
Polymorphism: List of polymorphisms (SNPs or InDels). For example: C → G, always the first base (C) is from Col-0 (reference genotype) and his group → the second base (G) from other accessions.
Genotypes: List of genotypes different to Col-0. Shown are accession IDs from 1001 Genome project.
CS number: List of genotypes different to Col-0. Shown are the CS accession IDs.
Accession: List of genotypes different to Col-0. Shown are accession names.

	A	B	C	D	E
1	#all accessions selected				
2	SNP id	Polymorphism	Genotypes	CS number	Accession
3	4:14731188	C -> G	9824	CS76702	Bes-5
4	4:14731200	C -> T	6911	CS76789	Cvi-0
5	4:14731132	A -> TA	88,139,159,265,410,428,	CS76790,CS77014,CS7	CYR,LDV-46,MAR2-3,PY
6	4:14731216	T -> C	6911	CS76789	Cvi-0
7	4:14731221	A -> T	9508,9518,9524,9528,95	CS77108,CS76660,CS7	IP-Mos-1,IP-Alm-0,IP-Ber
8	4:14731228	G -> A	9828	CS76721	Bra-0
9	4:14731234	C -> A	9533,9542,9545,9574,96	CS76763,CS76872,CS7	IP-Cem-0,IP-Fun-0,IP-He
10	4:14731260	T -> A	9543	CS76886	IP-Gra-0
11	4:14731261	T -> C	9879	CS77169	Per-0
12	4:14731275	T -> C	1158,1166,7169,8337,90	CS76656,CS77651,CS7	Aledal-6-49,Aledal-14-73,
13	4:14731393	G -> A	6090,7305,8343,9544,95	CS77288,CS78915,CS7	T1000,Pt-0,Na-1,IP-Gua-
14	4:14731395	GA -> A	9777	CS76880	Gn-1
15	4:14731396	G -> A	6897,9534,9841,9910	CS76430,CS76774,CS7	Ag-0,IP-Cmo-3,Ees-0,BR
16	4:14731430	T -> G	628,763,765,766,768,772	CS78951,CS76522,CS7	LI-OF-061,Kar-1,Sus-1,D
17	4:14731450	G -> A	9759,10013	CS76439,CS76388	Anz-0,Lerik1-3
18	4:14731471	AGAACTTTCC -> C	9099	CS76953	Istisu-9
19	4:14731480	A -> G	9991	CS76391	Vash-1

2.4.6.8 Accession information

The ‘accession_information’ is dumped as a csv file.

The accession information shows:

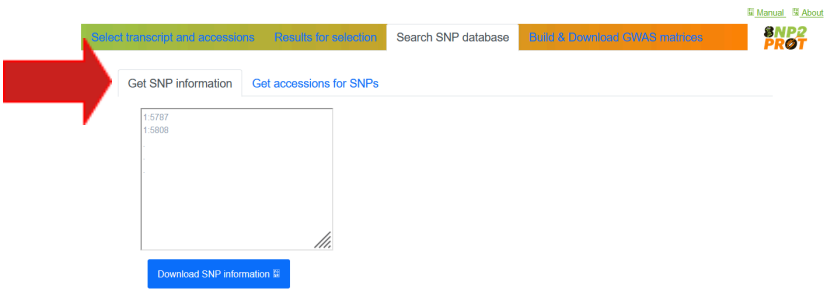
accession_id: Accession identification number based on the 1001 Genomes project.
accession: Name of accession.
cs_number: CS accession Ids.
country: Country of origin.
group: Population structure computed in 1001 genomes paper.

	A	B	C	D	E
1	accession_id	accession	cs_number	country	group
2	88	CYR	CS76790	FRA	western_europe
3	108	LDV-18	CS77013	FRA	western_europe
4	139	LDV-46	CS77014	FRA	western_europe
5	159	MAR2-3	CS77070	FRA	western_europe
6	265	PYL-6	CS77198	FRA	admixed
7	350	TOU-A1-88	CS77382	FRA	western_europe
8	351	TOU-A1-89	CS77383	FRA	western_europe
9	403	Zdarec3	CS78873	CZE	central_europe
10	410	Doubravnik7	CS76808	CZE	central_europe
11	424	Draha2	CS76812	CZE	central_europe
12	428	Borky1	CS76718	CZE	central_europe
13	430	Gr-1	CS76496	AUT	central_europe
14	470	BRR4	CS78943	USA	germany
15	476	BRR12	CS78944	USA	germany
16	484	BRR23	CS78945	USA	germany
17	504	BRR57	CS78946	USA	germany
18	506	BRR60	CS78947	USA	germany
19	531	BRR107	CS78948	USA	germany
20	544	LI-WP-039	CS78949	USA	germany
21	546	LI-WP-041	CS78950	USA	germany
22	628	LI-OF-061	CS78951	USA	western_europe
23	630	LI-OF-065	CS77036	USA	western_europe
24	680	LI-RR-096	CS78952	USA	germany
25	681	LI-RR-097	CS78953	USA	germany

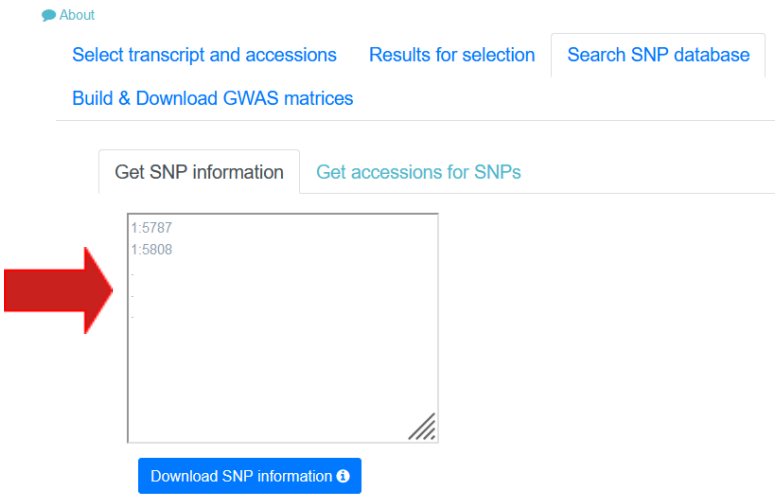
2.5 Search SNP database

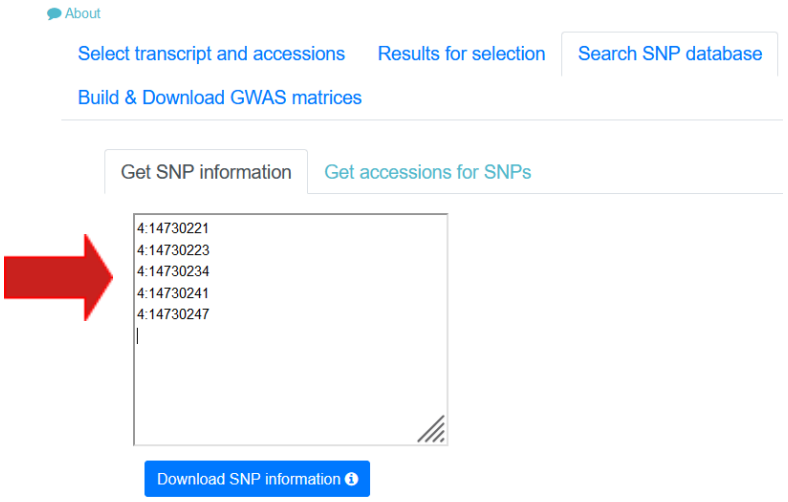
This database can be used to retrieve SNP information and accessions for SNPs. To get that information you need to identify SNP IDs of interest.

2.5.1 Get SNP information



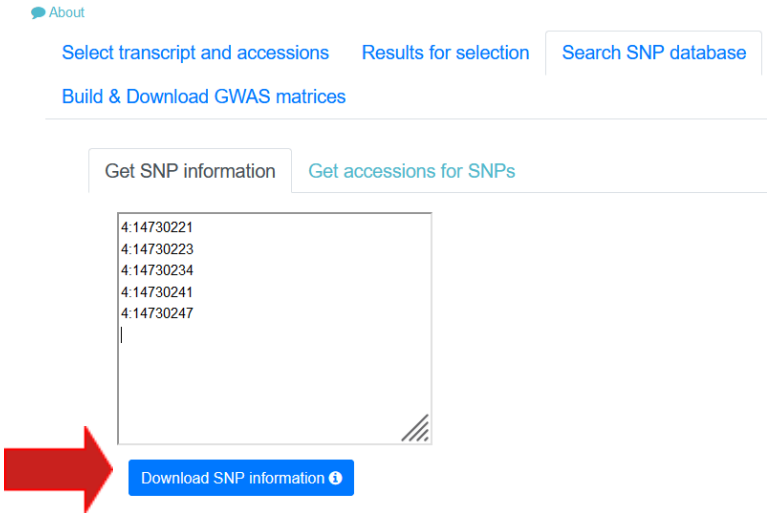
Copy SNP ID or IDs of interest into the empty field.





Click ‘Download SNP information’.

Download SNP information : For each SNP ID in the list all the transcripts covering the SNP are listed. Additional information like SNP effect, SNP region and base frequency are added.



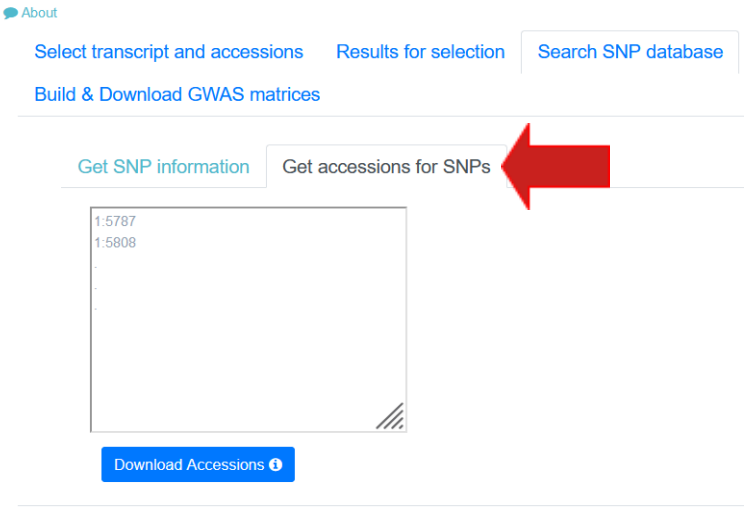
Transcript ID: Shown are *A. thaliana* gene Ids and splicing variant (e.g. AT4G30120.1) in which the selected SNPs are located
SNP ID: SNP identification number. Shown are chromosome:bp positions (4:14730221) of each SNP of selected gene.
Variant_0: Base pair variant on a specific position with the highest frequency of selected gene.
Frequency_0: Base frequency of variant 0 from selected gene.
Variant_1: Base pair variant on a specific position with the second highest frequency of selected gene.
Frequency_1: Base frequency of variant 1 from selected gene.
SNP relative position mRNA: SNP base pair position of selected gene mRNA.
SNP relative position genome: SNP base pair position of selected gene gDNA (genomic DNA).
SNP location: Gene structure location (e.g. 3' UTR (untranslated region), INTRON etc.) of SNP and selected gene.
SNP effects: Effect of SNP on AA sequence of selected gene (e.g. stop codon (premature stop codon), nonsynonymous (AA substitution), canonical motif in splice site etc.)

	A	B	C	D	E	F	G	H	I	J
1	Transcript id	SNP Id	Variant_0	Frequency_0	Variant_1	Frequency_1	SNP relative position mrna	SNP relative position genome	SNP location	SNP effects
2	AT4G30120.2	4:14730221	C -> C	0.99824	C -> A	0.00176	2296	2967	3UTR	noval
3	AT4G30120.2	4:14730223	T -> T	0.9859	T -> C	0.0141	2294	2965	3UTR	noval
4	AT4G30120.2	4:14730234	C -> C	0.99912	C -> A	0.00088	2283	2954	3UTR	noval
5	AT4G30120.2	4:14730241	G -> T	0.70749	G -> G	0.29251	2276	2947	3UTR	noval
6	AT4G30120.2	4:14730247	A -> A	0.99031	A -> T	0.00969	2270	2941	3UTR	noval
7	AT4G30120.1	4:14730221	C -> C	0.99824	C -> A	0.00176	2589	3408	3UTR	noval
8	AT4G30120.1	4:14730223	T -> T	0.9859	T -> C	0.0141	2587	3406	3UTR	noval
9	AT4G30120.1	4:14730234	C -> C	0.99912	C -> A	0.00088	2576	3395	3UTR	noval
10	AT4G30120.1	4:14730241	G -> T	0.70749	G -> G	0.29251	2569	3388	3UTR	noval
11	AT4G30120.1	4:14730247	A -> A	0.99031	A -> T	0.00969	2563	3382	3UTR	noval

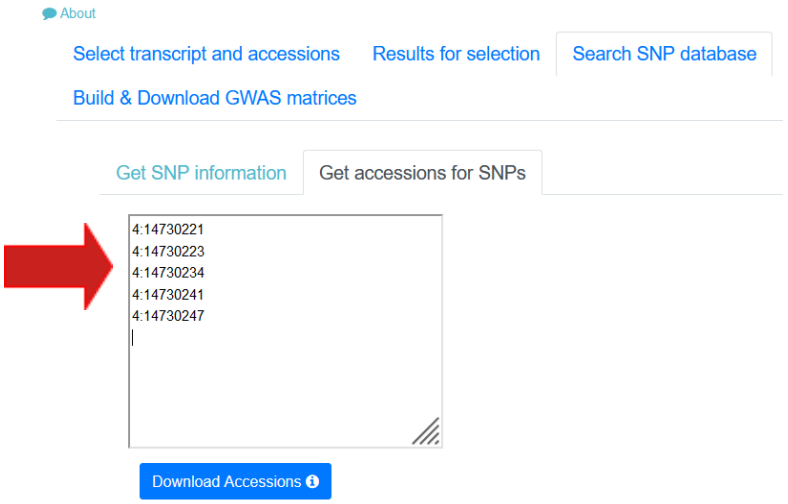
Substitution_properties: Physico-chemical properties of AA substitution (e.g. non-polar → polar, basic → acidic etc.)
Domain ID: Protein domain ID in which the AA substitutions are located.
Domain description: Description of domain.
Strand: Strand direction (e.g. - and +) of sequence.
GWAS corrected: GWAS analysis was performed with the software GEMMA. Every climate factor was used as a phenotype and for every climate factor, one analysis run was carried out. Only SNPs with a minor allele frequency of 1% or higher were analysed. The obtained p-values were corrected for multiple comparisons using the Benjamini-Yekutiel method. In this column you can see all the correlations between SNPs and the 200 climate factors which have a corrected p-value lower than or equal to the threshold. You can sort the column according to the lowest p-value in each cell.
Number of associated climate factors: Number of associated climate factors, with a corrected P-value in the GWAS lower than or equal to the threshold.

	K	L	M	N	O	P	Q	R
1	Substitution properties	Domain Id	Domain description	Strand	AraGWAS_score	AraGWAS_links	GWAS corrected	Number of associated climate factors
2	noval	noval	noval	-	noval	noval	noval	0
3	noval	noval	noval	-	noval	noval	noval	0
4	noval	noval	noval	-	noval	noval	noval	0
5	noval	noval	noval	-	noval	noval	noval	0
6	noval	noval	noval	-	noval	noval	noval	0
7	noval	noval	noval	-	noval	noval	noval	0
8	noval	noval	noval	-	noval	noval	noval	0
9	noval	noval	noval	-	noval	noval	noval	0
10	noval	noval	noval	-	noval	noval	noval	0
11	noval	noval	noval	-	noval	noval	noval	0

2.5.2 Get accessions for SNPs

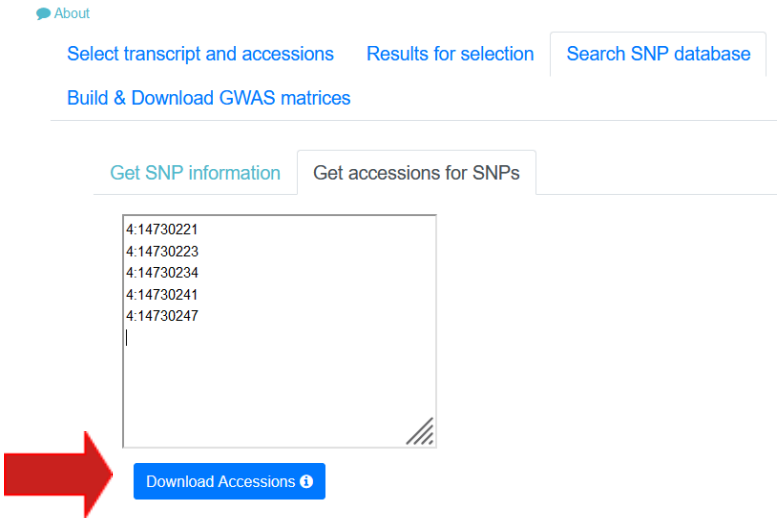


Copy and paste SNP ID or IDs of interest into the empty field.

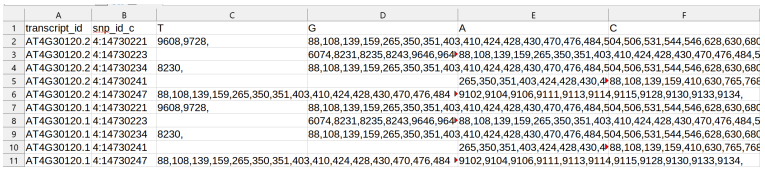


Retrieve accessions for SNPs of interest by clicking ‘Download Accessions’.

Download SNP information ⓘ : For each SNP ID in the list all the transcripts covering the SNP are listed. For each combination of transcript, SNP ID and variant all the genotypes with this specific variation are listed.



Transcript ID: Shown are *A. thaliana* gene IDs and splice variants (e.g. AT4G30120.2) in which the selected SNPs are located
SNP_ID: SNP identification number. Shown are chromosome:bp positions (4:14730221) of each SNP of selected gene.
T: Accessions with the base T on this position.
G: Accessions with the base G on this position.
A: Accessions with the base A on this position.
C: Accessions with the base C on this position.

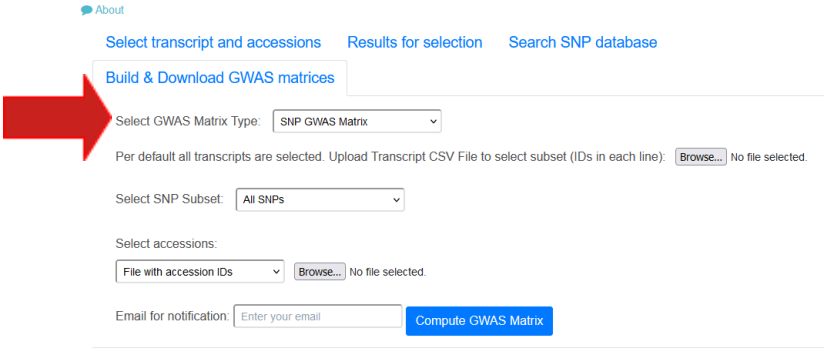


	A	B	C	D	E	F
1	transcript_id	snp_id_c	T	G	A	C
2	AT4G30120.2	4:14730221	9608,9728,	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,504,506,531,544,546,628,630,680		
3	AT4G30120.2	4:14730223		6074,8231,8235,8243,9646,964	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,5	
4	AT4G30120.2	4:14730234	8230,	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,504,506,531,544,546,628,630,680		
5	AT4G30120.2	4:14730241			265,350,351,403,424,428,430,	88,108,139,159,410,630,765,768
6	AT4G30120.2	4:14730247	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484		9102,9104,9106,9111,9113,9114,9115,9128,9130,9133,9134,	
7	AT4G30120.1	4:14730221	9608,9728,	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,504,506,531,544,546,628,630,680		
8	AT4G30120.1	4:14730223		6074,8231,8235,8243,9646,964	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,5	
9	AT4G30120.1	4:14730234	8230,	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484,504,506,531,544,546,628,630,680		
10	AT4G30120.1	4:14730241			265,350,351,403,424,428,430,	88,108,139,159,410,630,765,768
11	AT4G30120.1	4:14730247	88,108,139,159,265,350,351,403,410,424,428,430,470,476,484		9102,9104,9106,9111,9113,9114,9115,9128,9130,9133,9134,	

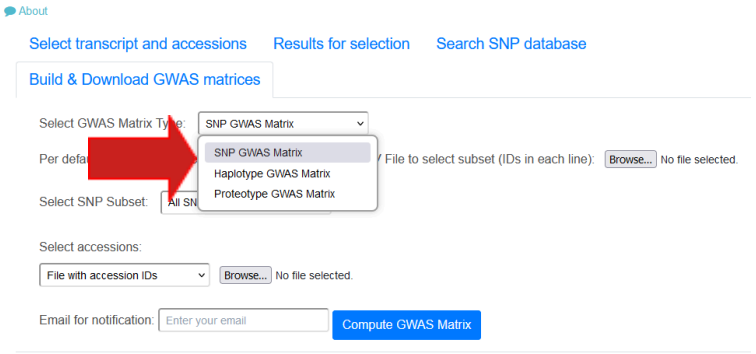
2.6 Build & Download GWAS matrices

SNPstar enables the generation of GWAS matrices of interest. The computed matrix contains accessions and SNPs, DNA haplotypes or protein haplotypes.

Select three different types of polymorphisms from the drop-down menu ‘Select GWAS Matrix Type’.



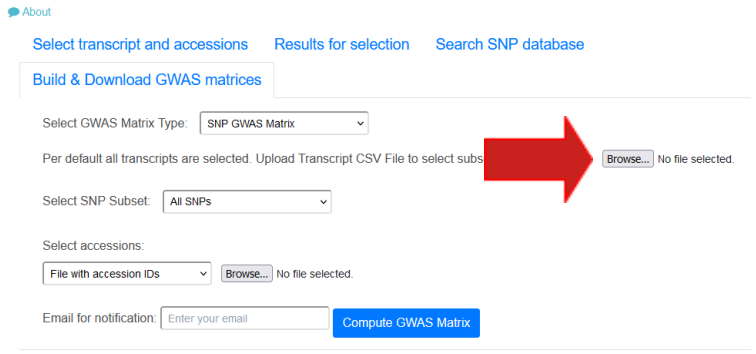
‘Select GWAS Matrix Type’ options are ‘SNP GWAS Matrix’, ‘Haplotype GWAS Matrix’ and ‘Proteotype GWAS Matrix’. If the ‘SNP GWAS Matrix’ is applied, a matrix for all available SNPs of selected genes is generated. For ‘Haplotype GWAS Matrix’ only DNA haplotypes (based on bases) and for ‘Proteotype GWAS Matrix’ only protein haplotypes (or proteotypes based on AAs).



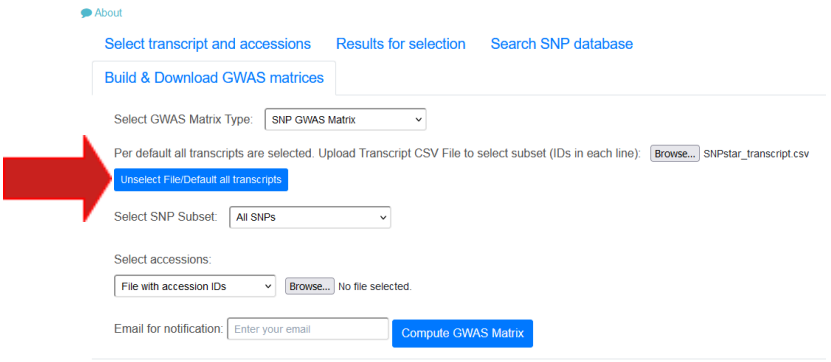
You can upload a transcript table (csv file). If the default settings are used, a matrix for all transcripts (genes) is computed. If a download transcript file is used, a matrix for transcripts (genes) of interest is computed. The uploaded file needs to contain gene and transcript numbers (e.g. AT4G30120.1).

	A	B
1	AT4G30110.1	
2	AT4G30110.2	
3	AT4G30120.1	
4	AT4G30120.2	
5		

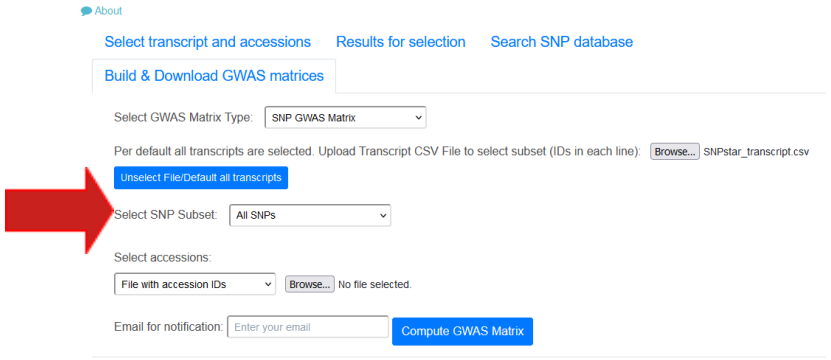
Click ‘Browse’ and select your transcript file of interest for upload.



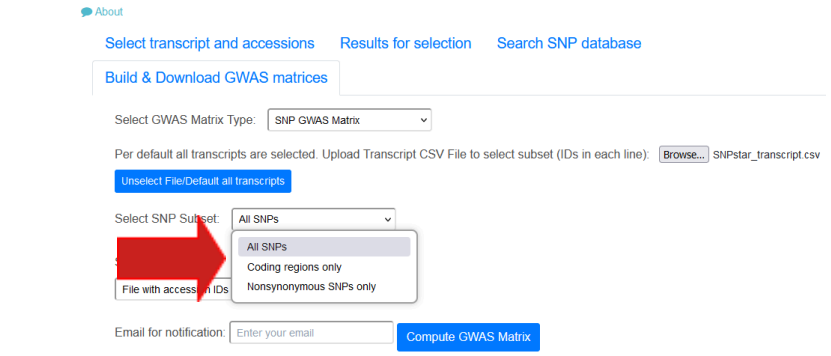
To remove the selected transcript (Gene) file, click ‘Unselect File/Default all transcripts’.



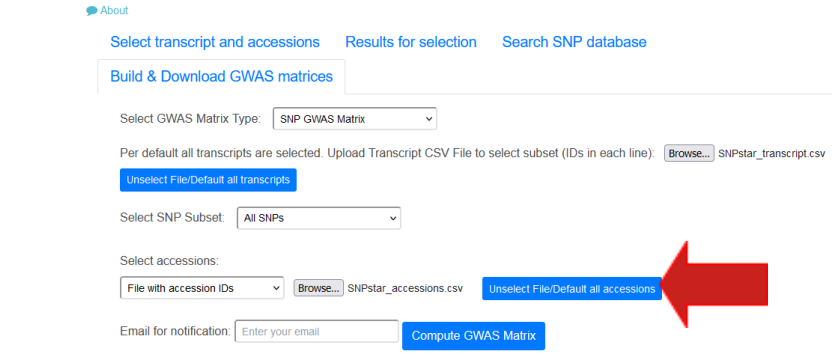
Next, select a SNP subset.



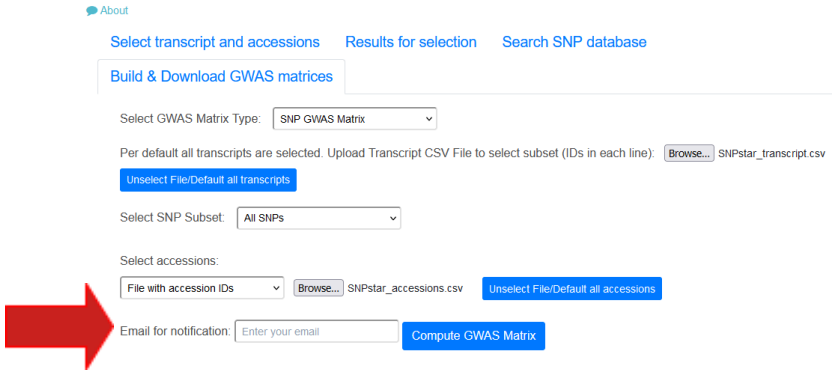
'Select SNP Subsets' options are 'All SNPs', 'Coding region only' and 'Nonsynonymous SNPs only'. If 'All SNPs' are selected all SNPs are used. For 'Coding region only' SNPs in coding regions and for 'Nonsynonymous SNPs only' SNPs causing AA substitutions are considered.



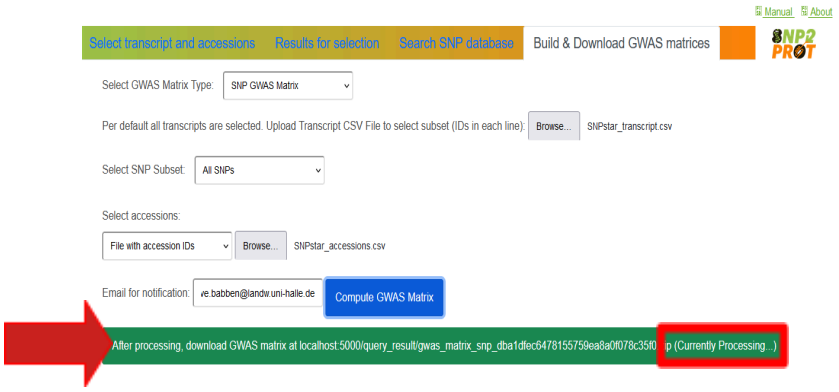
If you would like to select accessions of interest you can upload an accession file. For the structure of that file see chapter 2.2.2. If the default settings are used a matrix for all accessions is computed. The upload works like for the transcript table (refer to 2.6).
To remove the selected accession file use the 'Unselect File/Default all accession' button.



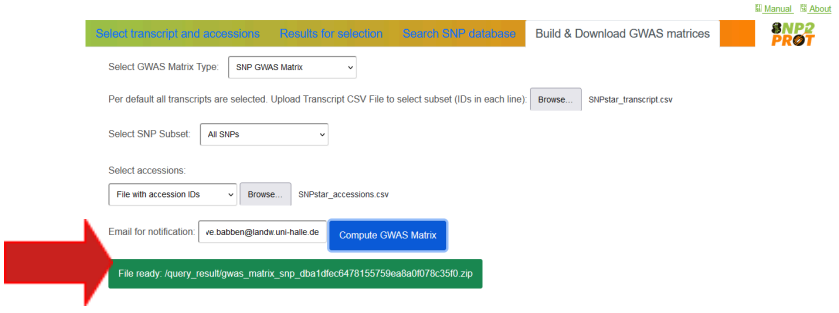
Provide an email address for notification and press ‘Compute GWAS Matrix’.



If the calculation is in progress, an information window is displayed.

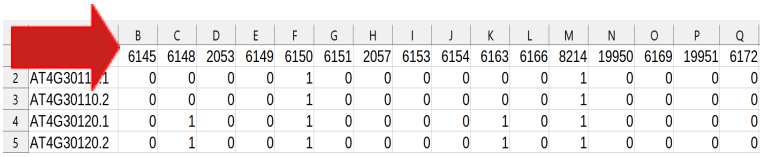


To retrieve the GWAS matrix click on the green window (only when processing is finished).



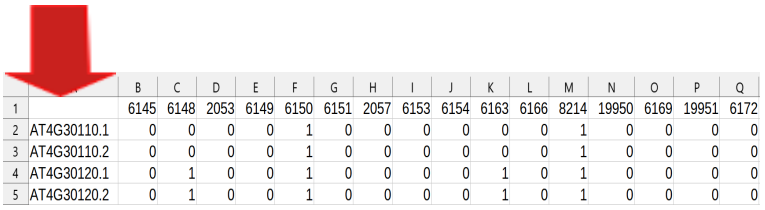
2.6.1 Layout of SNP GWAS Matrix

In the first horizontal row all selected accessions (IDs from the 1001 Genomes project) are shown.



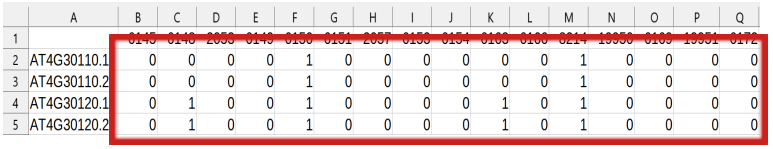
		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
3	AT4G30110.2	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
4	AT4G30120.1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0
5	AT4G30120.2	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0

In the first vertical column all selected transcripts are shown.



		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
3	AT4G30110.2	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
4	AT4G30120.1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0
5	AT4G30120.2	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0

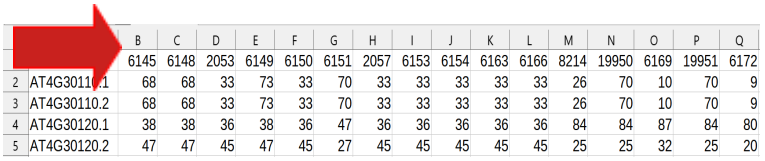
The matrix consists of numbers. The number zero (0) describes the status of Col-0. One (1) describes an allele different from Col-0. This matrix can be used for GWAS and/or kinship calculations.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
3	AT4G30110.2	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
4	AT4G30120.1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0
5	AT4G30120.2	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0

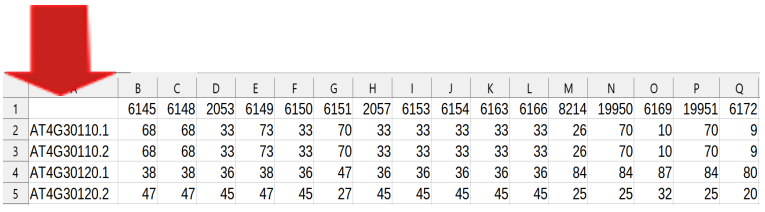
2.6.2 Layout of Haplotype GWAS Matrix

In the first horizontal row all selected accessions (IDs from the 1001 Genomes project) are shown.



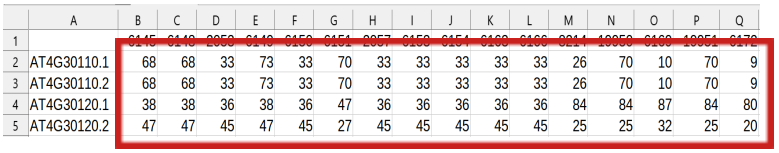
		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
3	AT4G30110.2	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
4	AT4G30120.1	38	38	36	38	36	47	36	36	36	36	36	84	84	87	84	80
5	AT4G30120.2	47	47	45	47	45	27	45	45	45	45	45	25	25	32	25	20

In the first vertical column all selected transcripts are shown.



		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
3	AT4G30110.2	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
4	AT4G30120.1	38	38	36	38	36	47	36	36	36	36	36	84	84	87	84	80
5	AT4G30120.2	47	47	45	47	45	27	45	45	45	45	45	25	25	32	25	20

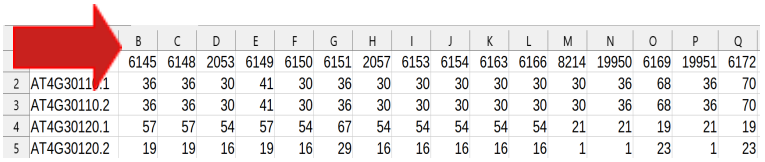
The matrix consists of numbers which describe a specific haplotype of specific accession and transcript. This matrix can be used for GWAS calculation.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
3	AT4G30110.2	68	68	33	73	33	70	33	33	33	33	33	26	70	10	70	9
4	AT4G30120.1	38	38	36	38	36	47	36	36	36	36	36	84	84	87	84	80
5	AT4G30120.2	47	47	45	47	45	27	45	45	45	45	45	25	25	32	25	20

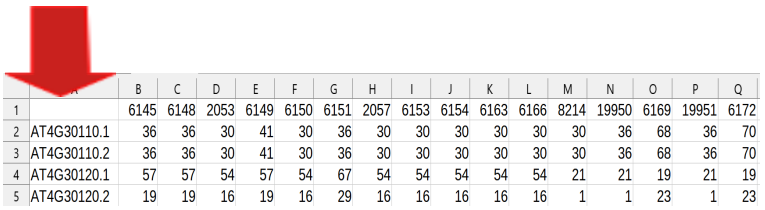
2.6.3 Layout of Proteotype GWAS Matrix

In the first horizontal row all selected accessions (IDs from the 1001 Genomes project) are shown.



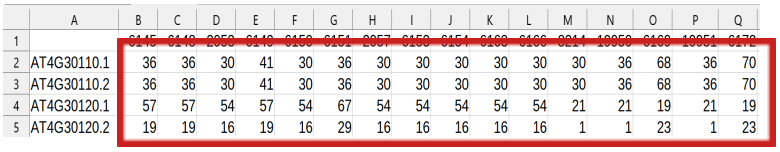
		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
3	AT4G30110.2	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
4	AT4G30120.1	57	57	54	57	54	67	54	54	54	54	54	21	21	19	21	19
5	AT4G30120.2	19	19	16	19	16	29	16	16	16	16	16	1	1	23	1	23

In the first vertical column all selected transcripts are shown.



		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
3	AT4G30110.2	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
4	AT4G30120.1	57	57	54	57	54	67	54	54	54	54	54	21	21	19	21	19
5	AT4G30120.2	19	19	16	19	16	29	16	16	16	16	16	1	1	23	1	23

The matrix consists of numbers which describe a specific protein haplotype/proteotype of a specific accession and transcript. This matrix can be used for GWAS calculation.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		6145	6148	2053	6149	6150	6151	2057	6153	6154	6163	6166	8214	19950	6169	19951	6172
2	AT4G30110.1	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
3	AT4G30110.2	36	36	30	41	30	36	30	30	30	30	30	30	36	68	36	70
4	AT4G30120.1	57	57	54	57	54	67	54	54	54	54	54	21	21	19	21	19
5	AT4G30120.2	19	19	16	19	16	29	16	16	16	16	16	1	1	23	1	23

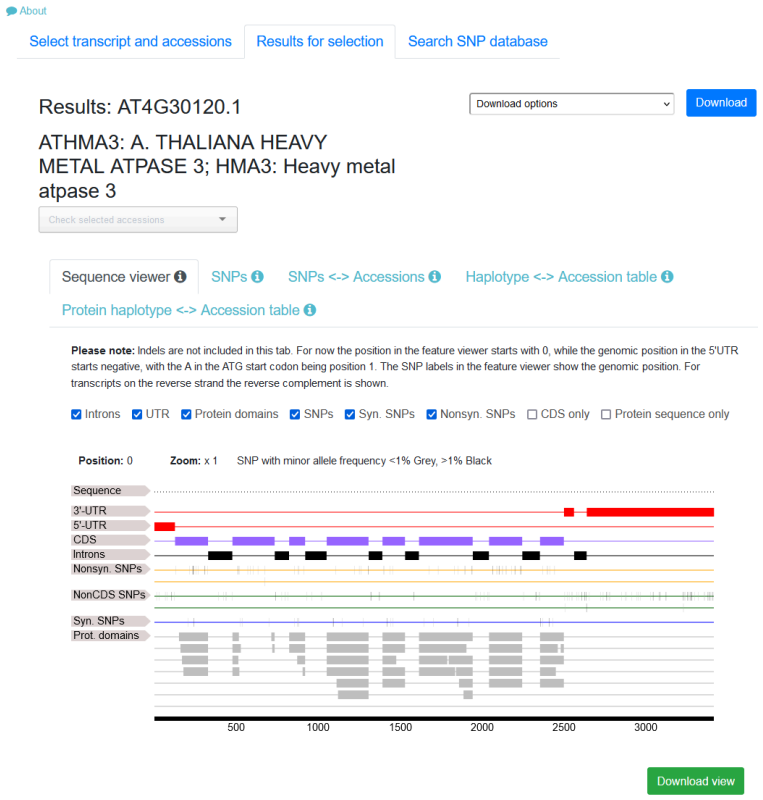
3 Use cases

SNPstar covers a number of different use cases. In this chapter, we will explain the most prominent use cases, which can be roughly grouped into single target gene, multiple target genes and GWAS.

3.1 Single target gene

In this first use case, we will consider a single target gene and all SNPs in this gene. We will again use *A. THALIANA HEAVY METAL ATPASE 3* (*HMA3*; *AT4G30120*) as an example.

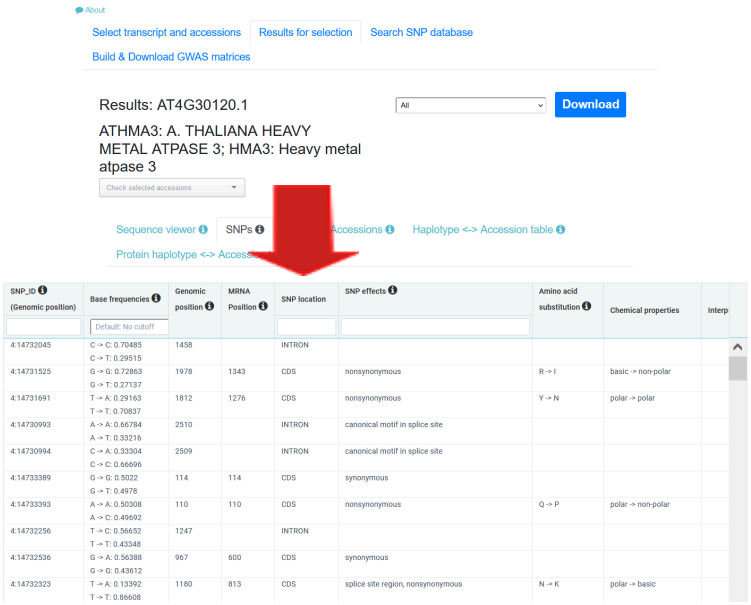
First, select gene and gene model (AT4G30120.1; see chapter 2.1.1). Then we select all available accessions (see chapter 2.2.1) and click ‘Compute results’ (see chapter 2.3) to get to the ‘Sequence viewer’ tab as displayed below.



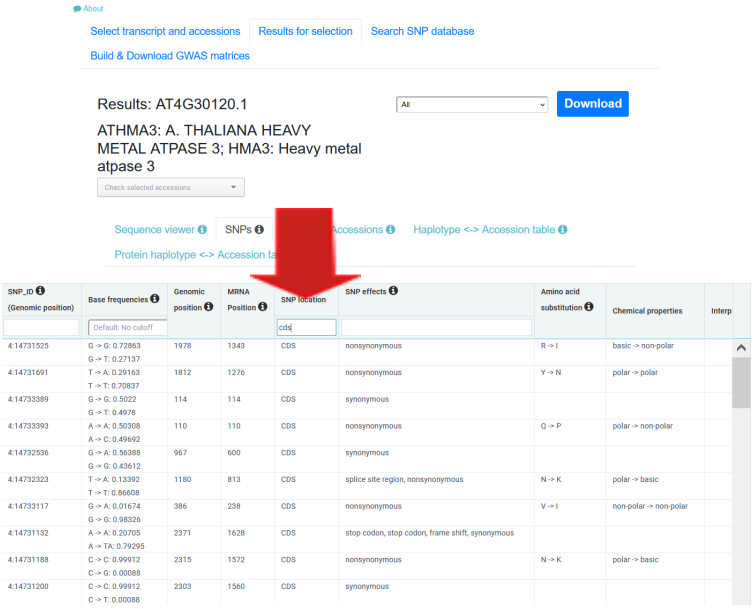
3.1.1 Identification of SNPs in a single gene of interest

3.1.1.1 General selection and SNP identification

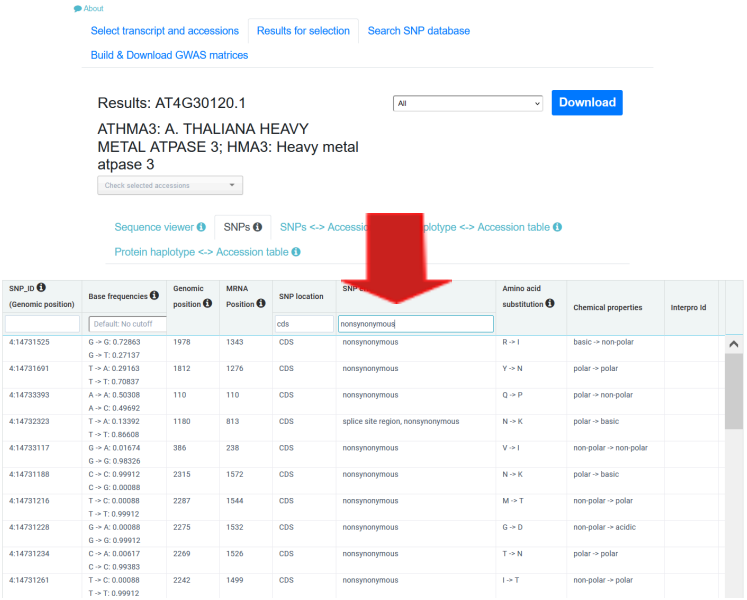
The Sequence viewer tab (see chapter 2.4.1) shows an overview of all SNPs represented in all accessions. Additional information are gene structure (e.g. CDS, Intron, etc.), detected protein domain locations and effect of SNPs (nonsynonymous, synonymous and non-CDS). Beneath the overview figure the sequence is shown.
For a closer look of detected SNPs use the ‘SNPs’ tab (see chapter 2.4.2). Different selection options can be used depending on your research interest. If you are interested in a specific SNP location use the column ‘SNP location’ for selection. As an example we use CDS for selection.



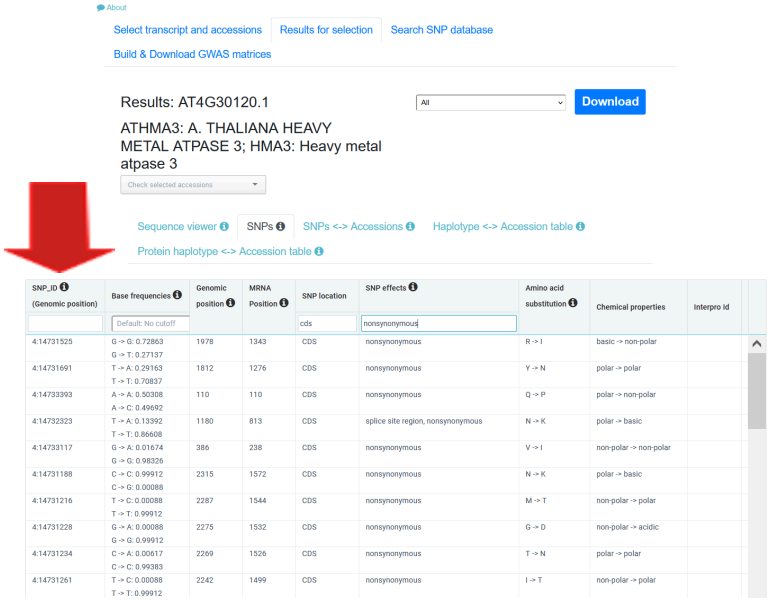
Type ‘CDS’ into the empty field below the column header. The selection works automatically (no need to press enter). Shown are all SNPs which are located in the CDS of the selected target gene.



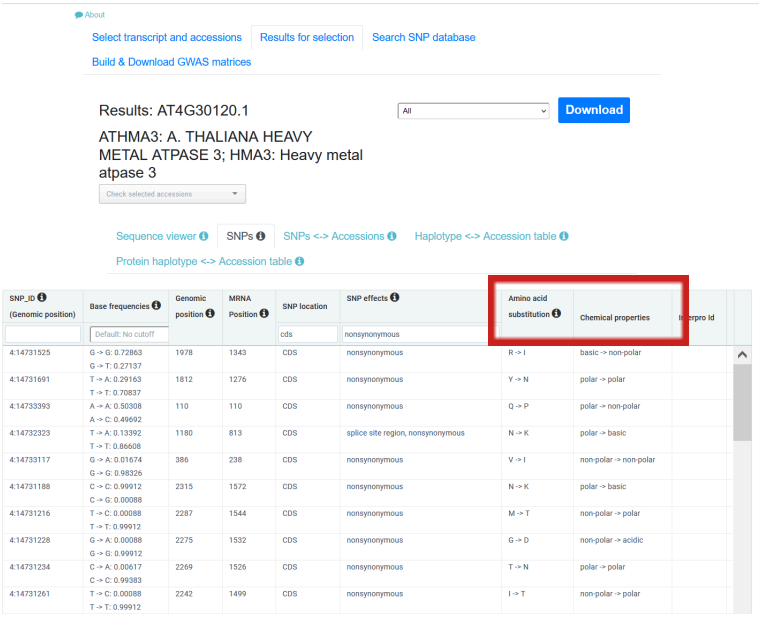
If you are interested in further filtering the SNPs in the CDS for, for example, those causing missense mutations, type ‘nonsynonymous’ into the empty field below the column header ‘SNP effects’. Shown are all SNPs which are located in the CDS with an AA substitution effect.



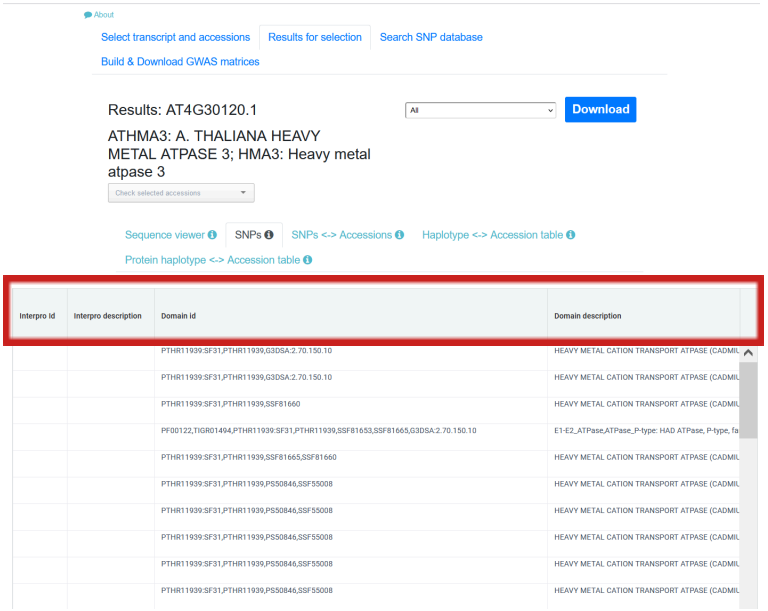
The first column ‘SNP ID’ shows all SNPs which are selected with the implemented filter steps by now. The SNP IDs are important to identify accessions which contain SNPs of interest (see chapter 3.1.1.2).



Additional interesting information are the nature of AA substitutions and whether the mutated residue may have different physico-chemical properties to specifically focus on such SNPs with a potential consequence on protein structure and/or function.



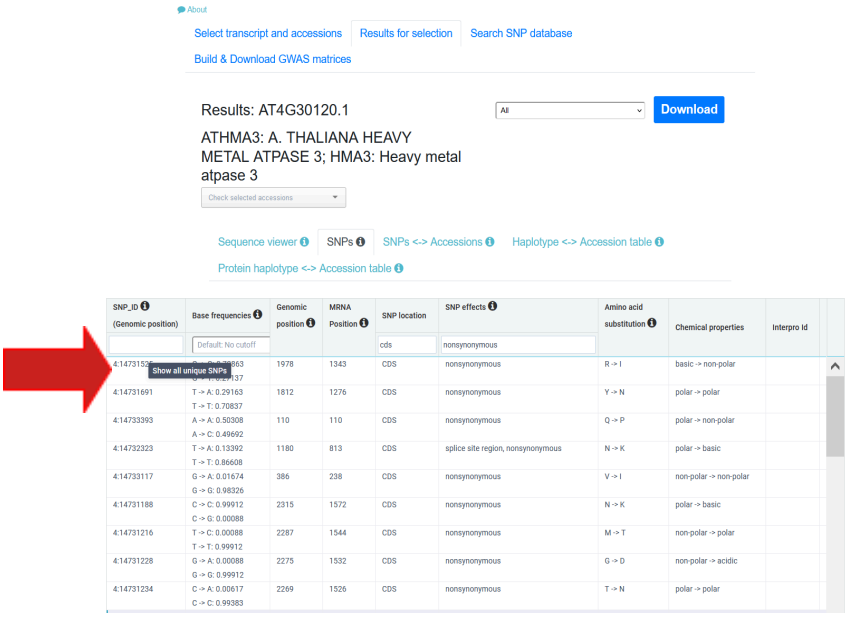
If you are interested in nonsynonymous SNPs which are located in a protein domain or in combination with already shown information (e.g. physico-chemical property, etc.), scroll to the right to the ‘Interpro ID’, ‘Interpro description’, ‘Domain ID’ and ‘Domain description’ columns to retrieve this type of information.



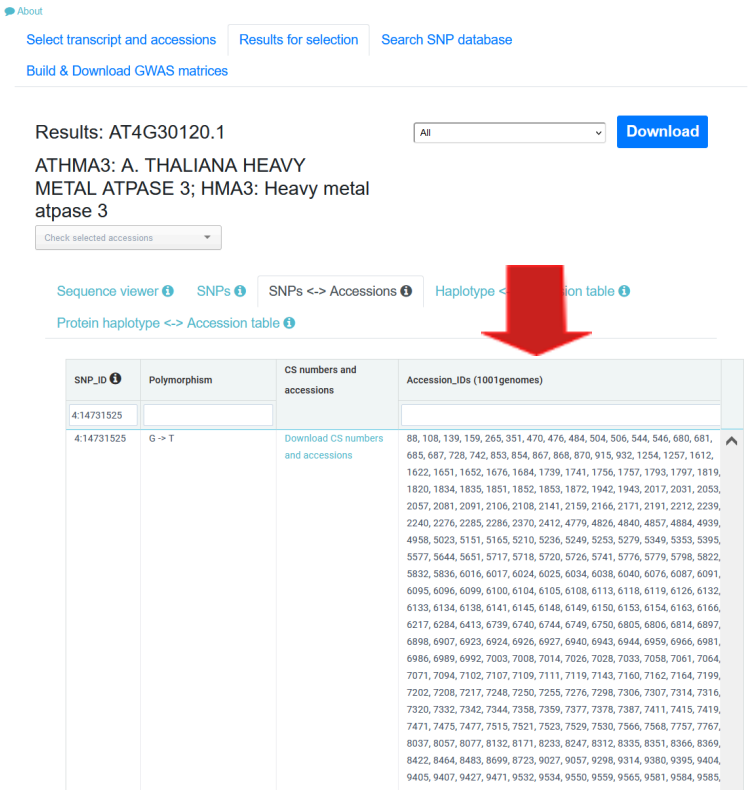
To download information for the SNPs of interest, generate a table with a format as previously described (see 2.5.1) to identify and download accessions carrying your SNP(s) of interest (see 2.5.2).
Another way to identify SNP IDs for SNPs of interest is to download the complete SNP table (see 2.4.6.6). Open the downloaded file in your data processing software of choice (e.g. Excel, etc.) and filter the columns of interest (SNP location, SNP effects, etc.). The resulting lists can also be used for accession identification (chapter 3.1.1.2).
Here we would like to explain filtering for SNP location and SNP effects. Filter CDS in your calculator program of choice (e.g. Excel) in the column SNP location.

3.1.1.2 Identification of accessions with SNPs of interest

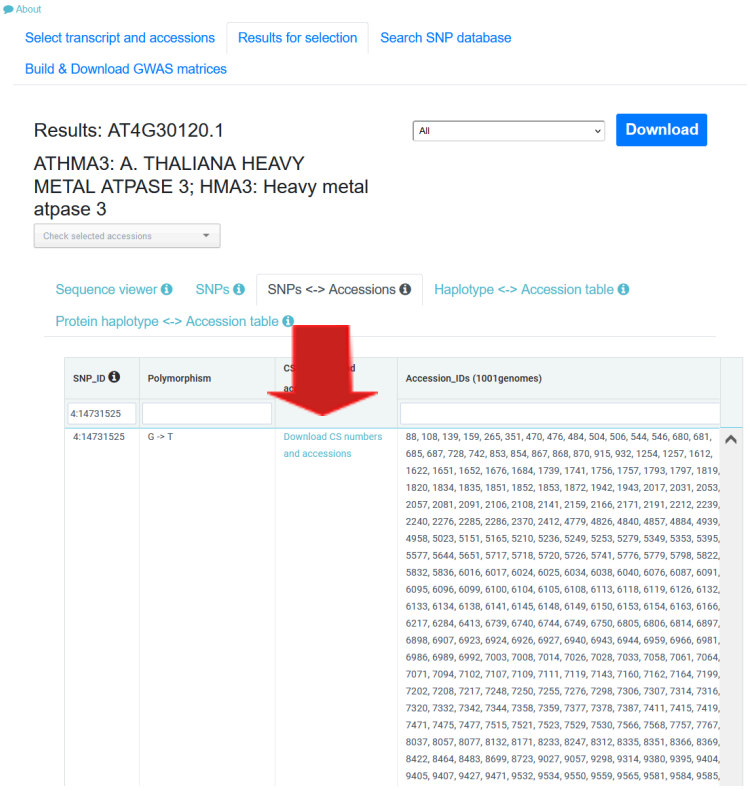
If you have identified a single SNP of interest, you might want to do experiments with accessions carrying different alleles at this locus, or simply clone different alleles from original genetic material. You can click on the ‘SNP ID’ (e.g. 4:14731525) to move directly to the tab ‘SNPs ↔ Accessions’ (see chapter 2.4.3).



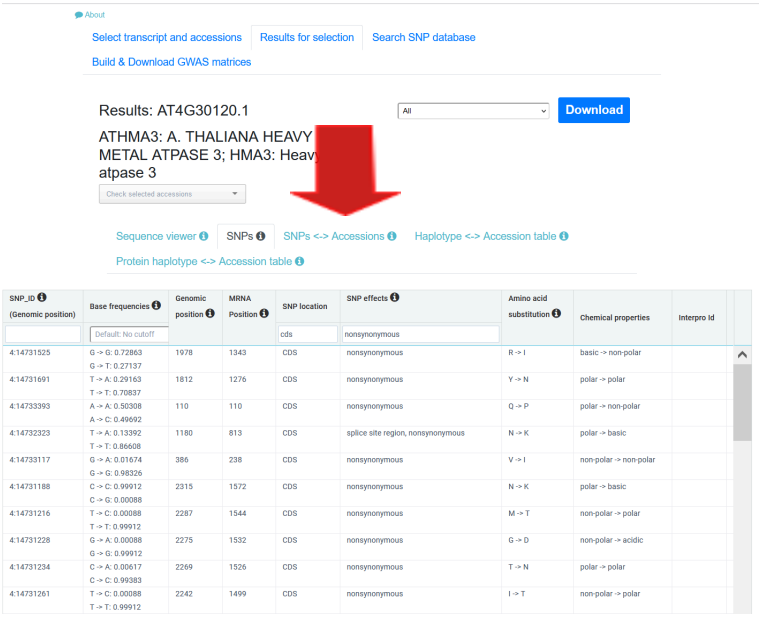
For the selected SNP of interest all accessions carrying the allele different from the Col-0 reference are now shown in the column ‘Accession_IDs (1001 genomes)’. All shown accessions are different to the reference (Col-0). Now you can search for accessions which are in your lab if you write down the accession ID of interest in the empty field below the column header Accession_IDs (1001 genome).



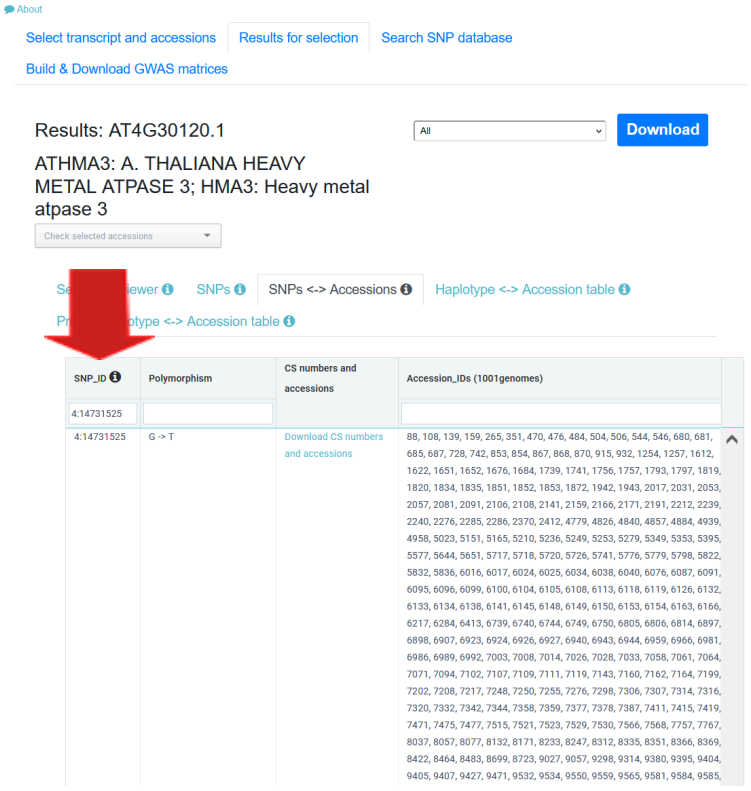
You can download the accession table by clicking on the link in the column ‘CS numbers and accessions’ (see in chapter 2.4.3).



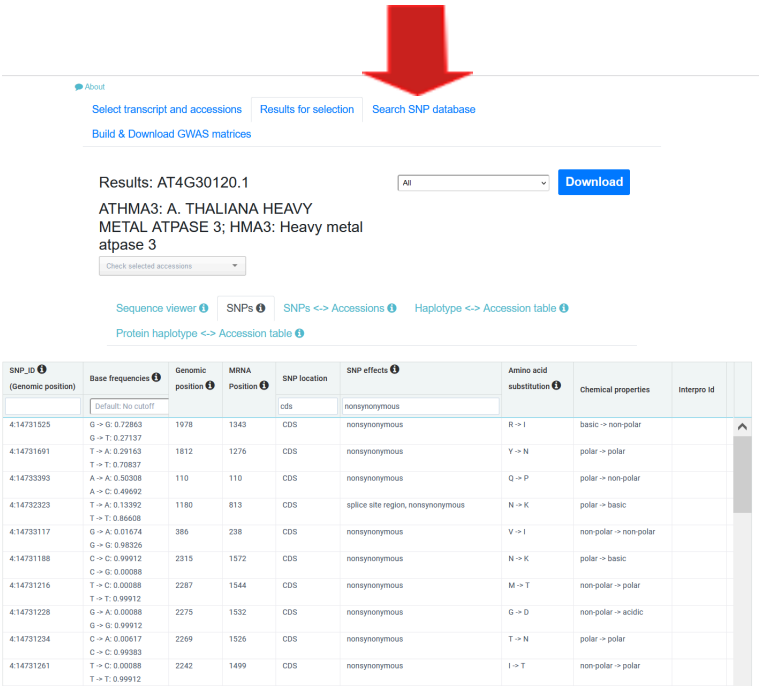
Another way is to move to to the ‘SNPs ↔ Accessions’ tab (see chapter 2.4.3) ...

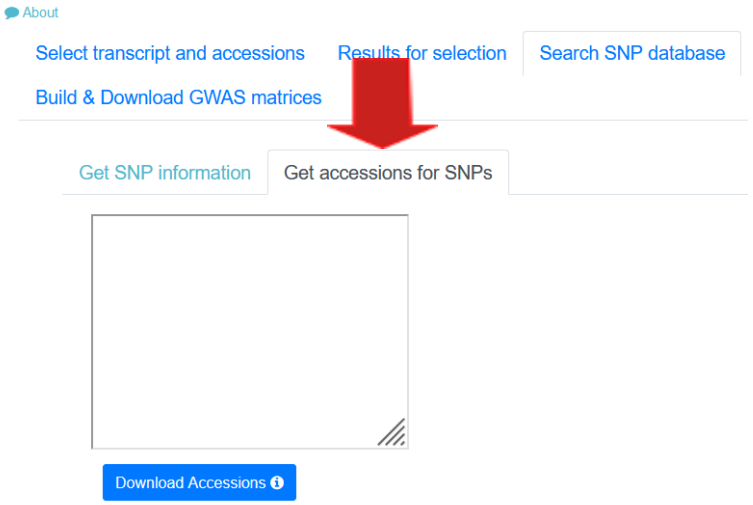


… and type a single SNP ID of interest (e.g. 4:14731525) into the empty field below the column header ‘SNP_ID’.



If you are interested in more than a single SNP, generate an input list with SNP IDs in your data processing software of choice (e.g. Excel) or, alternatively, use a filtered SNP table (see chapter 3.1.1.1). Next, move to the ‘Search SNP database’ tab followed by ‘Get accessions for SNPs’ tab.

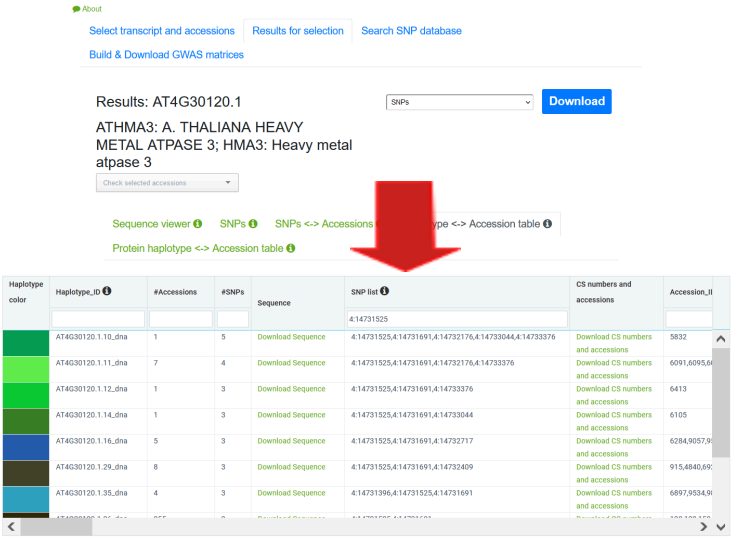




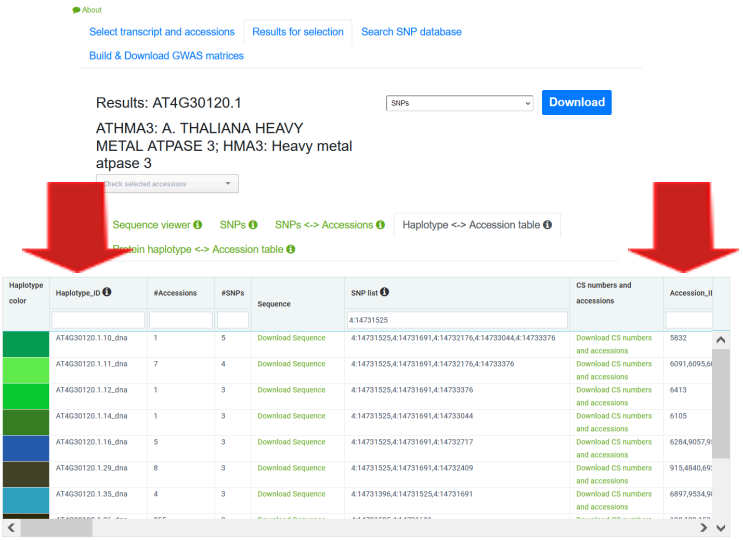
Paste the SNP IDs of interest into the empty field and click ‘Download Accessions’ to retrieve accession information (see chapter 2.5.2).

3.1.1.3 Identification of accessions with haplotypes of interest

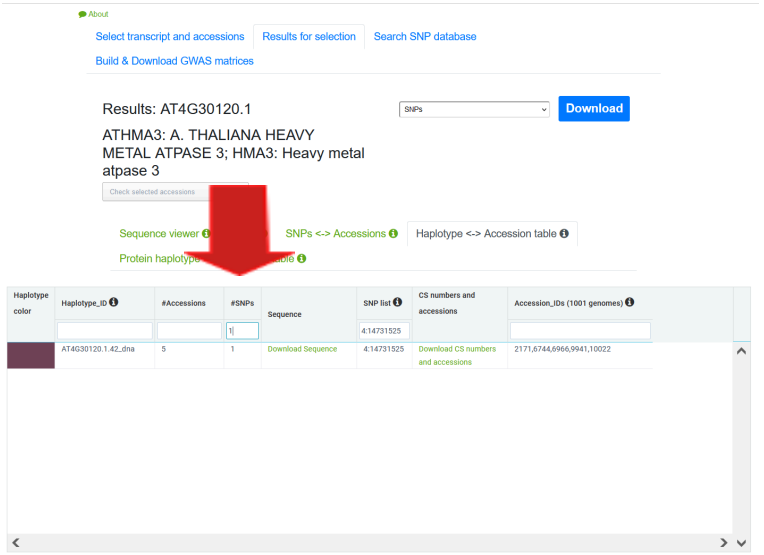
Since SNPs rarely occur in isolation due to phenomena like linkage disequilibrium, you might be interested in identifying haplotypes containing an identical sequence of SNPs for your gene of interest; for example for subsequent cloning projects. We will therefore demonstrate the identification of haplotypes and accessions sharing the same haplotype for a gene of interest.
First, identify SNPs of interest as previously shown (see 3.1.1.2). If you have a list of SNP IDs you can search for a single SNP ID in the tab ‘Haplotype ↔ Accession table’. Type your SNP ID of interest (e.g. 4:14731525) into the column ‘SNP list’ (see chapter 2.4.4).



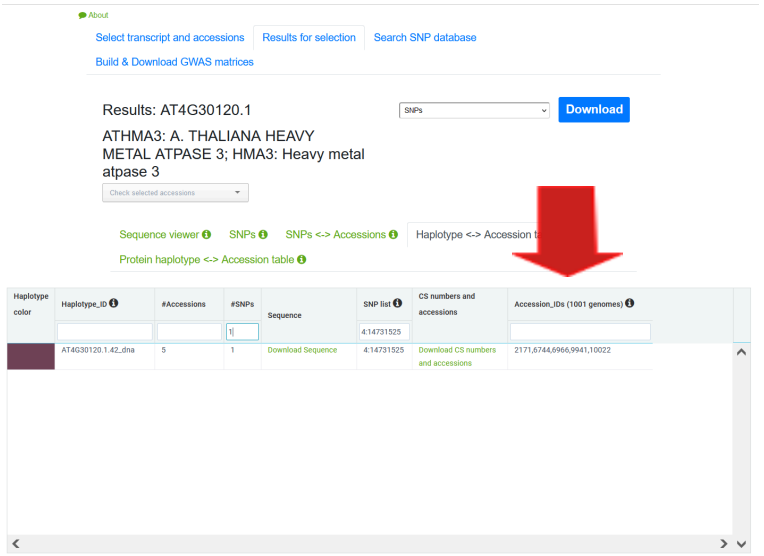
Shown are all haplotypes and the accessions with your SNP of interest.



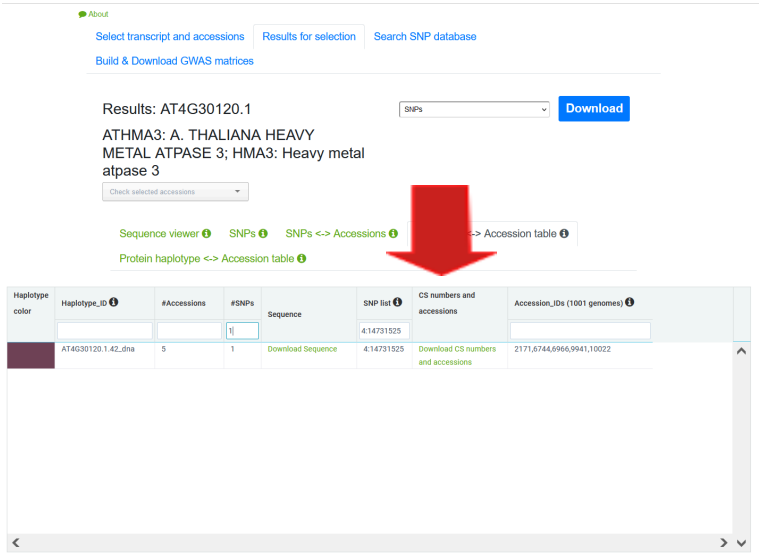
If you would like to identify a haplotype which contains only your SNP of interest, write 1 in the search field of column ‘#SNPs’. Now all haplotypes and accessions which contain only your SNP of interest are shown.



For our example we identify the haplotype AT4G30120.1.42 with the SNP of interest (4:14731525). Now we can identify accessions which have the haplotype of interest. All accessions are listed in the column ‘Accession_IDs’.

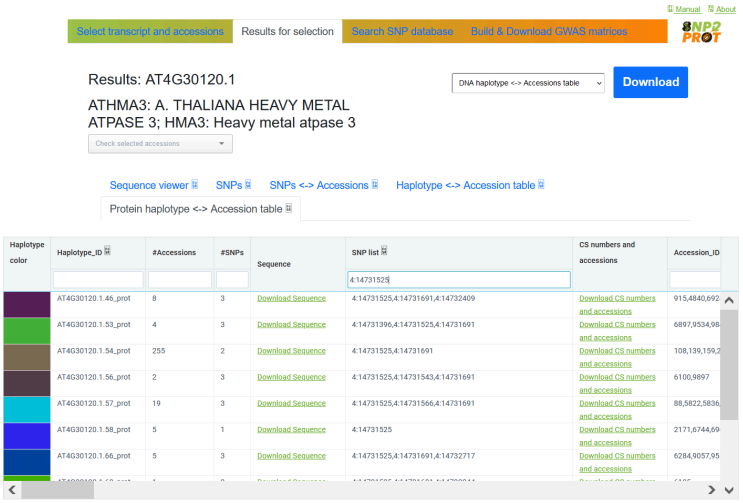


You can also download a list of the selected accessions.



3.1.1.4 Identification of accessions with protein haplotypes of interest

If you are interested in SNPs which show only AA substitutions then use protein haplotypes. In our example we use the same SNP ID as before (4:14731525) and then move to the tab ‘Protein haplotype ↔ Accession table’ (see chapter 2.4.5). Filter for SNP ID 4:14731525.



To identify accessions carrying the SNP alleles of interest, follow the same procedure as for haplotype identification described above (see 3.1.1.3).

3.1.2 Identification of known associated SNPs in a gene of interest

In its SNP table SNPstar also lists published associations of SNPs retrieved from the AraGWAS (https://aragwas.1001genomes.org/#/) database and, in addition, associations of SNPs with environmental/climatic factors (eGWAS). You can exploit this information to find out whether SNP(s) in your gene of interest are associated with such phenotypes or environmental/climatic data. Select your gene of interest and all accessions (see 2.1).

Move to tab ‘SNPs’ and check the columns ‘AraGWAS top hits’, ‘GWAS corrected p-values’ and ‘Number of associated climate factors’ (see 2.4.2) to retrieve this information.

3.2 GWAS

3.2.1 GWAS with all transcripts and a subset of accession.

The initial situation here is the availability of phenotypic data for a subset of accessions or the plan to collect phenotypic data for a subset of accessions that are available in your lab. The widest approach to perform a genome wide association study would include all available SNPs. In this case, you can upload a csv file with the subset of accessions, choose "SNP GWAS Matrix" as matrix type and "All SNPs" as SNP subset and compute the GWAS matrix (see chapter 2.6). This will result in a large dataset. For a more reduced dataset, you could choose a subset of SNPs like "coding SNPs" or "nonsynonymous SNPs" which would focus the subsequent analysis on functional SNPs.
As the conventional GWAS exhibits difficulties in detecting rare alleles and is furthermore limited by the mostly biallelic information of the SNPs, the use of "Haplotype GWAS Matrix" or "Proteotype GWAS Matrix" as matrix type is an effective approach to overcome the limitations of biallelic SNPs and increases the allelic resolution of the candidate genomic region (Qian et al. 2017). The proteotype dataset contains fewer variants per transcript compared to the haplotype dataset as only nonsynonymous SNPs that cause an amino acid exchange are considered (see chapter 2.6).
The computed GWAS matrices can be used directly or with some modification as input for different GWAS tools, e.g. TASSEL etc.

3.2.2 GWAS with a subset of transcripts and a subset of accessions

For this use case, the initial situation is again the availability of phenotypic data for a subset of accessions or the plan to collect phenotypic data for a subset of accessions that are available in you lab. Furthermore, you have already focused your interest on a group of transcripts, may it be the large group of transcription factors, auxin signaling genes or a gene family like ARFs (AUXIN RESPONSE FACTORs). After uploading the csv files of your subset of accession and your transcripts of interest, you could compute the "SNP GWAS Matrix" type with "All SNPs" or a "Haplotype GWAS Matrix", if your analysis should contain all SNPs. If you want to focus more on SNPs that cause an amino acid exchange, you can select the "SNP GWAS Matrix" with "nonsynonymous SNPs" or the "Proteotype GWAS Matrix" (see chapter 2.6). The computed GWAS matrices can be used directly or with some modification as input for different GWAS tools, e.g. TASSEL etc.

3.2.3 GWAS with all or a subset of transcripts and all accessions

This would correspond to the rare occasion that phenotypic data is available for all 1135 accessions. If the GWAS matrix should also contain all transcripts you do not need to upload any file, only choose the SNP matrix type and SNP type and compute the GWAS matrix (see also 3.5.1). If you want to focus on a specific subset of transcripts, you upload only the transcript csv file and then compute the GWAS matrix (see also 3.5.2).