**Key terminology**

domain (of a protein)

family (of proteins)

Hidden Markov Model (HMM)

homologs

orthologs

paralogs

**branch (of a phylogenetic tree) [includes “branch length”]**

Lineages of taxonomic units that link nodes within a phylogenetic tree.

In a rooted tree, branches indicate direct transmission of genetic information from the taxonomic unit located at one end (parent) of the branch to the other (child). To determine the direction of transmission, consider removing the branch of interest from the tree to yield two unconnected subtrees, each of which contains only one of the nodes directly linked by the branch of interest. The subtree containing the root node contains the more ancestral of the two nodes linked by the branch of interest; hence, the direction of transfer of genetic information is from this more ancestral node to the other node linked by the branch of interest. For an unrooted tree, it is unknown which of the subtrees contains the root node, hence in unrooted trees the direction of transmission of genetic information is not specified.

If the branch is part of a scaled phylogenetic tree, then a value is associated with the branch that indicates some measure of the difference between the two taxonomic units directly linked by the branch; this value is often referred to as the "branch length". If the tree is "unscaled", no such value is associated with the branch, that is, no branch length is specified or defined.

A branch that links two internal nodes is known as an internal, inner, or interior branch. Branches linking an internal and an external node are referred to as external branches (also terminal branches).

**branch length**

See “branch”.

**node (of a phylogenetic tree)**

In phylogenetic trees, nodes represent taxonomic units. Nodes between which there is direct transfer of genetic information are linked to each other by branches. Nodes in a phylogenetic tree that are attached to only a single terminal branch are referred to as terminal nodes (also external nodes, leaves, or tips), and represent operational taxonomic units. Nodes attached to more than one branch are referred to as internal (also interior) nodes, and represent hypothetical taxonomic units.

**phylogenetic tree [includes “root”]**

A description of a path of transmission of genetic information between a set of operational (and usually also hypothetical, if the tree contains any internal nodes) taxonomic units (see Operational Taxonomic Units (OTUs) and Hypothetical Taxonomic Units (HTUs)). Tree structures, as understood in graph theory, can be used to represent phylogenetic trees. In graph theory, trees are defined as undirected graphs for which exactly one path connects any two nodes (or "vertices"), i.e., a tree is any connected graph that does not contain any cycles. Phylogenetic trees consist of nodes (operational or hypothetical taxonomic units) that are connected via branches.

Phylogenetic trees are described as either "rooted" or "unrooted". In a rooted tree, there is one node (the "root node") that represents the most recent common ancestor of all other taxa in the tree. While it is assumed that such an ancestor also exists for an unrooted tree, i.e., that all the OTUs share a most recent common ancestor, in an unrooted tree no inference is made about where on the tree this HTU might be.

**polytomy**

An internal node of a phylogenetic tree is described as a multifurcation if (i) it is in a rooted tree and is linked to three or more daughter subtrees or (ii) it is in an unrooted tree and is attached to four or more branches. A tree that contains any multifurcations can be described as a multifurcating tree.

The rooted tree, A, in the figure above, contains five bifurcations (thick horizontal lines) and one multifurcation (dashed horizontal line). The unrooted tree, B, in the figure above, contains 3 bifurcations (solid dots) and one

**root**

See “phylogenetic tree”

**sister group**

In a rooted bifurcating tree, any internal node represents an ancestor of two subtrees. These two subtrees are sometimes referred to as "sister groups" of each other, i.e., subtree A is the sister group of subtree B (and accordingly subtree B is the sister group of subtree A). As this definition depends on knowing the direction of genetic transmission along the branches of the tree (i.e., which nodes/branches are ancestral and which are descendant), it is only possible to identify sister groups in the context of a rooted phylogenetic tree; in unrooted trees the direction of transmission of genetic information is not specified, so it is not possible to identify which of the subtrees linked to an internal node are ancestral and which are descendant. To overcome this problem, the concept of adjacent groups was developed for referring to subtrees linked to the same internal node in unrooted trees.

**subtree**

A tree obtained by detaching a branch from a larger phylogenetic tree.

**taxonomic unit**

Phylogenetic trees can be used to describe patterns of genetic transmission between different kinds of entities, for example: different species, different individuals within a population of the same species, different genes within a gene family. The term "taxonomic unit" is used to refer to the entities between which patterns and paths of genetic transfer are described. Thus, for some trees, the taxonomic units will be individuals within a population, in other trees they will be different species.

**topology (of a phylogenetic tree)**

The pattern of linkage between the nodes in a tree.

A given tree topology can be represented in many different ways. Consider, for example, the tree topology in which the following list describes all pairs of nodes that are directly linked to each other: AE, ED, DB, EF, and DC. Three different representations of this topology are shown in the figure below.

Note that the pattern of links between nodes, and hence the tree topology, is independent of the lengths of the branches.

Definitions taken from entries written by Aidan Budd and Alexandros Stamantakis in the 2nd edition of the “Dictionary of Bioinformatics and Computational Biology” (eds John M. Hancock and Marketa J. Zvelebil)

**Virtual Machine**

Password: tsl2017

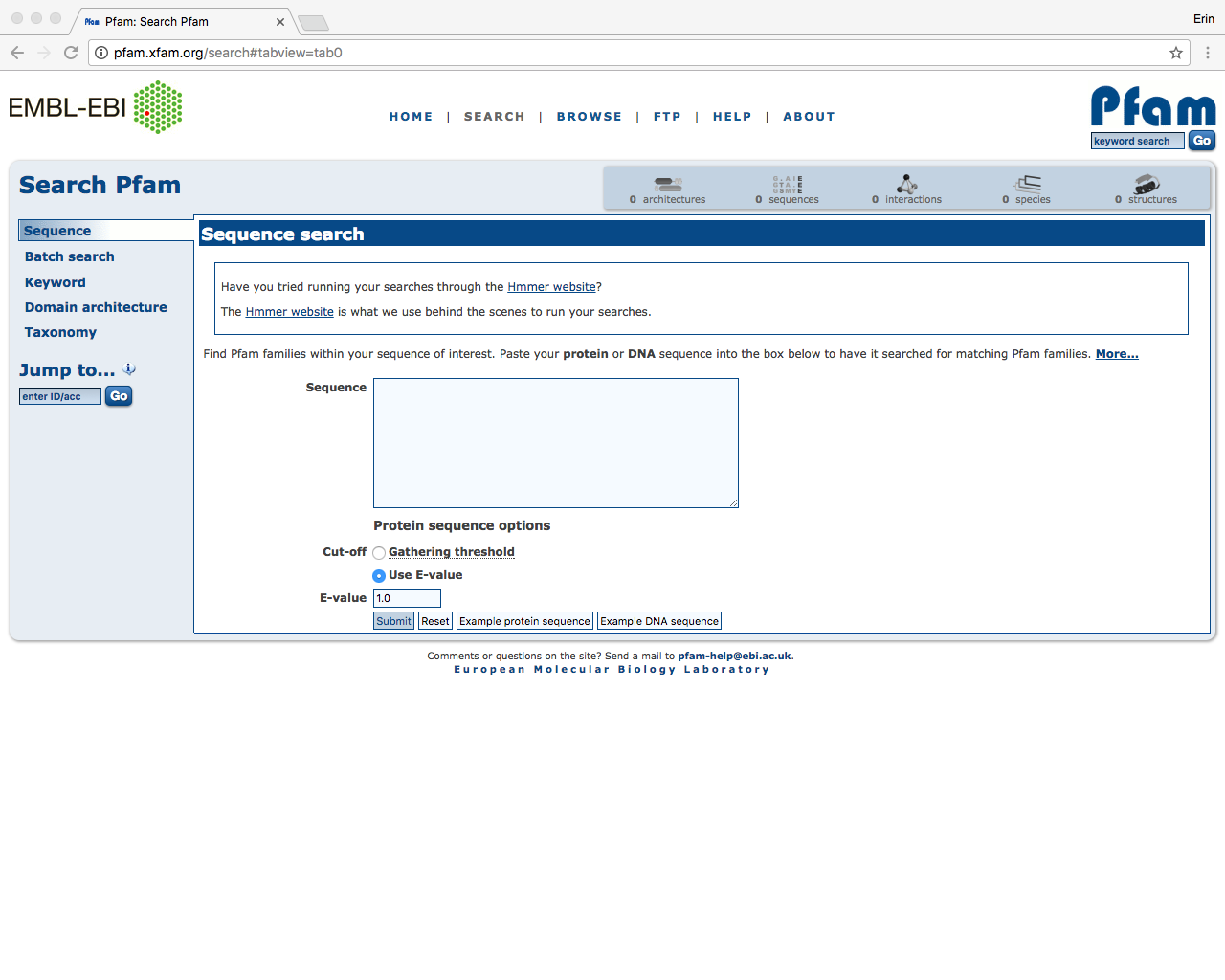
**Project 1 Planting tree sequences (Level 1 skills - web servers)**

This project will allow you to gain the skills necessary to identify the origin of an integrated domain and analyze its homologs. We will annotate the ID and identify the closest Arabidopsis protein that they are related to. Since we know so much about Arabidopsis, you can think why a homolog of a particular Arabidopsis protein is integrated into an NLR.

Look into the Project 1 and find a folder with a number you have drawn. You will find a mystery NLR-ID sequence that we previously identified in our genomic screens. Now, you will identify this IDs and analyze it through genomics and phylogenetics. In this exercise, no command line skills are required, however you are welcome to use them if you prefer.

**Domain prediction**

**Tool:** Pfam search - <http://pfam.xfam.org/browse>





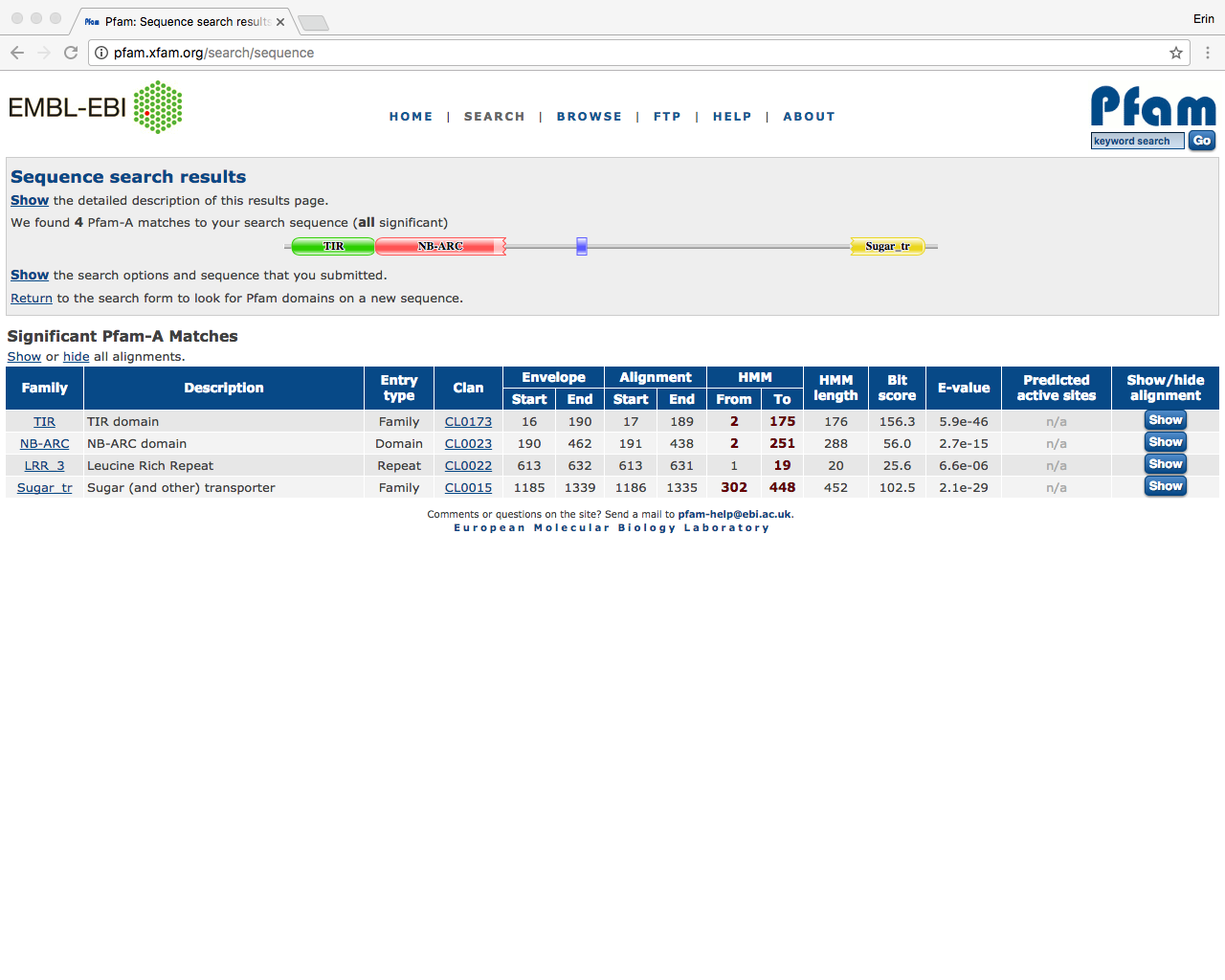
**Method:**

Follow link

Navigate to **search** at top of web page

Click on **sequence** then **paste** your sequence in the box and press submit.

**Results**



1. Shows you the predicted domain composition for your protein sequence.
2. The conserved type of protein sequence predicted
3. Where in your sequence the region is predicted. If overlapping domain predictions treat with caution
4. A description of confidence in the hit accounting for likelihood of random matches due to multiple searching. The lower the e-value the greater confidence in the hit.
5. Clicking show alignment will give you a view of the match between the HMM and your query (#SEQ) if your query is mainly highlighted in green this means it is high posterior probability match whilst regions in red have low posterior probability.

Normal domains you would expect are NB-ARC, TIR, CC and LRR. You can follow the link next to non-canonical domain name (1) to find out more about your identified domain including its Pfam ID which you need for later.

If your protein has more than one non-canonical domains then note both down in the group spreadsheet and then paste the protein sequence into CDD search then press live search. If a domain appears in both searches with a low e-value you can be more confident it exists and it will be more likely to work with the remaining steps.

Tool: NCBI CDD https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?

Link to group spreadsheet: <https://docs.google.com/spreadsheets/d/10CpRp6iwr1mYy3XlvwqkBOK-ML0Odj3EP7gtZLghVzI/edit?usp=sharing>

Pfam ID of domain:

**Positives:**

Easy to use

Clean visualisation

Can search multiple protein sequences at once

**Limitation:**

Limit to bulk sequence upload and sequences must be protein

Data not available for bulk downstream

Cannot search many sequences for just one specific domain of interest

**NB command line users**: you can use HMMER (hmmscan) and Pfam-A installed on this machine in place of web-servers.

**Retrieving proteins with domain of interest**

**Tool: Biomart**

**Link:**

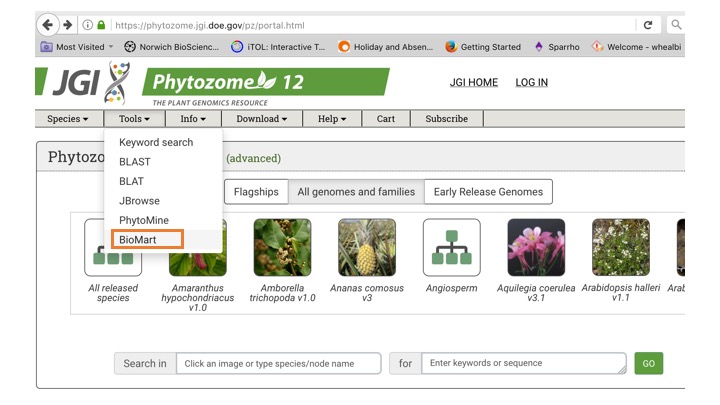
**Phytozome:**

[**https://phytozome.jgi.doe.gov/biomart/martview/fa2c76032665114d5ef856f01ad6df97**](https://phytozome.jgi.doe.gov/biomart/martview/fa2c76032665114d5ef856f01ad6df97)

**Plant Ensembl**

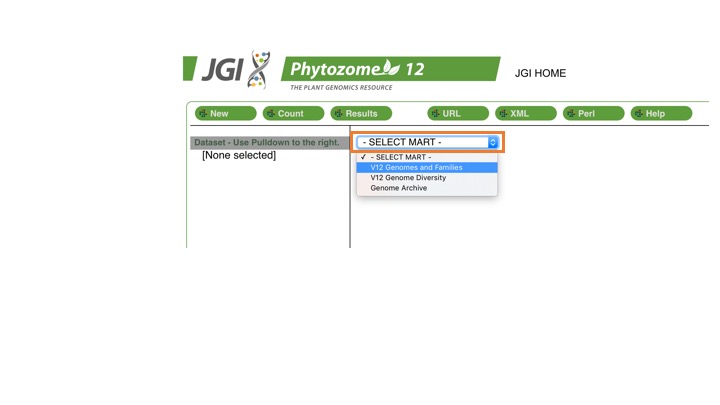
[**https://plants.ensembl.org/biomart/martview/4dd2ed2030be39022f6b96b974e838c7**](https://plants.ensembl.org/biomart/martview/4dd2ed2030be39022f6b96b974e838c7)

Biomart allows you to extract sequences and many other features from genomes and proteomes by specifying your requirements. Biomart is available under Tools on the Phytozome web-server



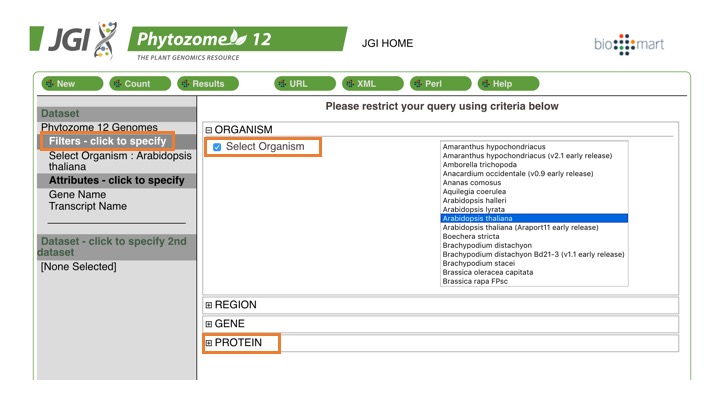
**Method:**

1. After navigating to the Biomart tool, choose ‘V12 Genomes and Families’ from the drop down menu on the right:

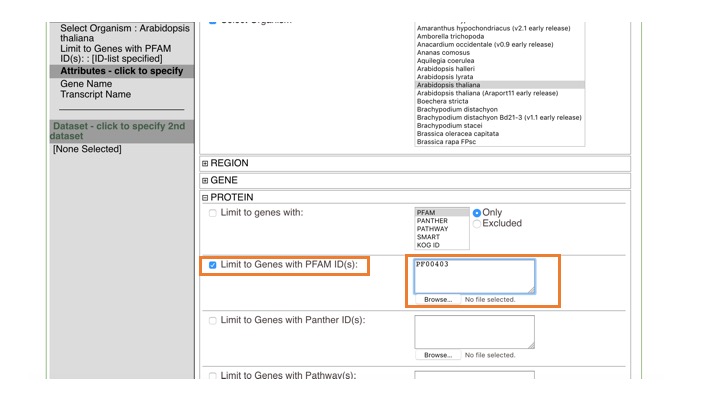


In the next drop down menu, choose ‘Phytozome 12 Genomes’

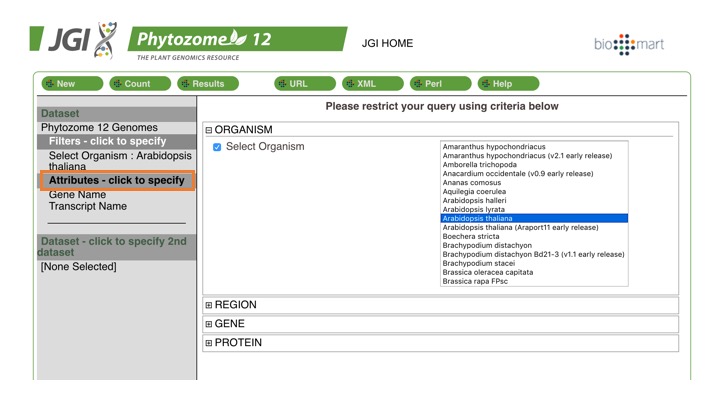
2. Then on the left, click on ‘Filters’ and check ‘Select organisms’, then choose ‘Arabidopsis thaliana’ (or another organism if you want to try it). Next, click on ‘Protein’ to expand this section.



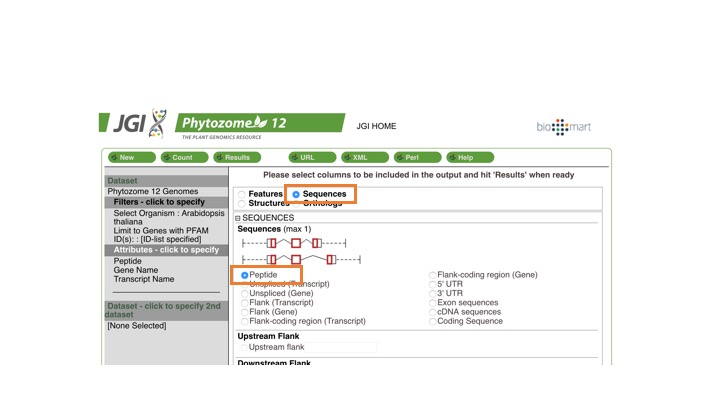
3. Click ‘Limit to genes with PFAM IDs’ and type the Pfam domain identification number that you got for your integrated domain from the Pfam webserver.

****

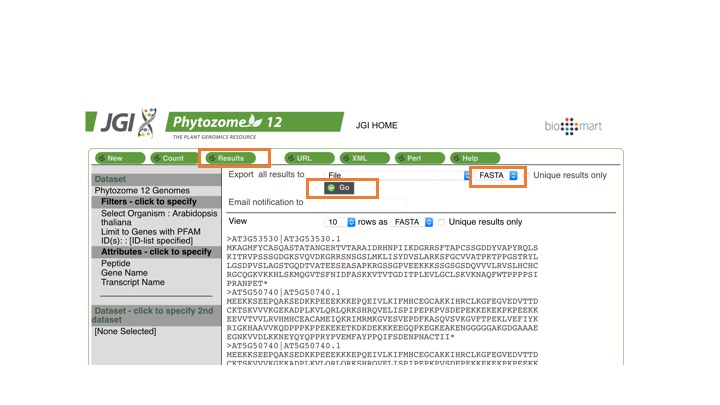
4. After you are done with filters, you need to go back to the panel on the left and click on

****

Here, you specify that in your results, you want to download sequences, and peptides

****

5. Finally, you can click on ‘Results’, check that you have results downloaded in ‘FASTA’ format and hit ‘Go’ to download a file with all sequences in Arabidopsis that have the domain you specified.

****

6. The results should be in your downloads folder. Open the file and you should see fasta with all full length sequences containing your domain of interest.

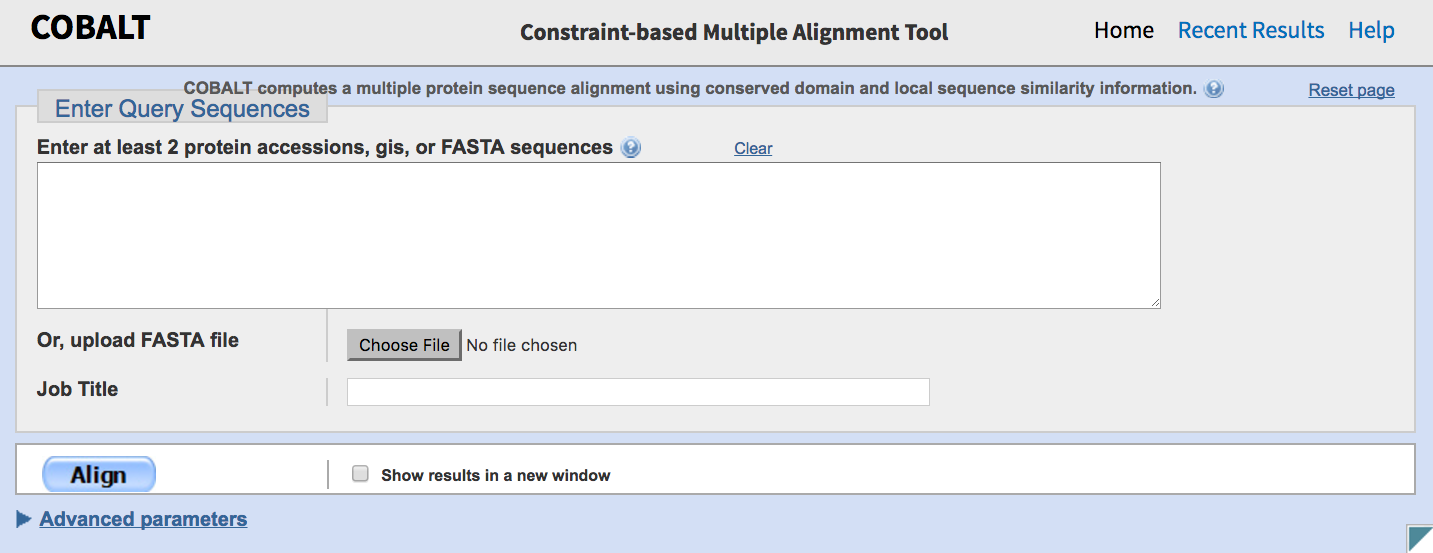
Add your NLR-ID sequence to this file for next steps of the analyses and save file in your Project 1 folder.

**NB command line users:** We annotated all domains in Arabidopsis proteins with pfamscan, so you can use grep and/or scripts outlined in project 2 to search for and extract proteins with a given domain.

**Alignment**

**Tool:** Cobalt aligner

**Link:** https://www.ncbi.nlm.nih.gov/tools/cobalt/re\_cobalt.cgi

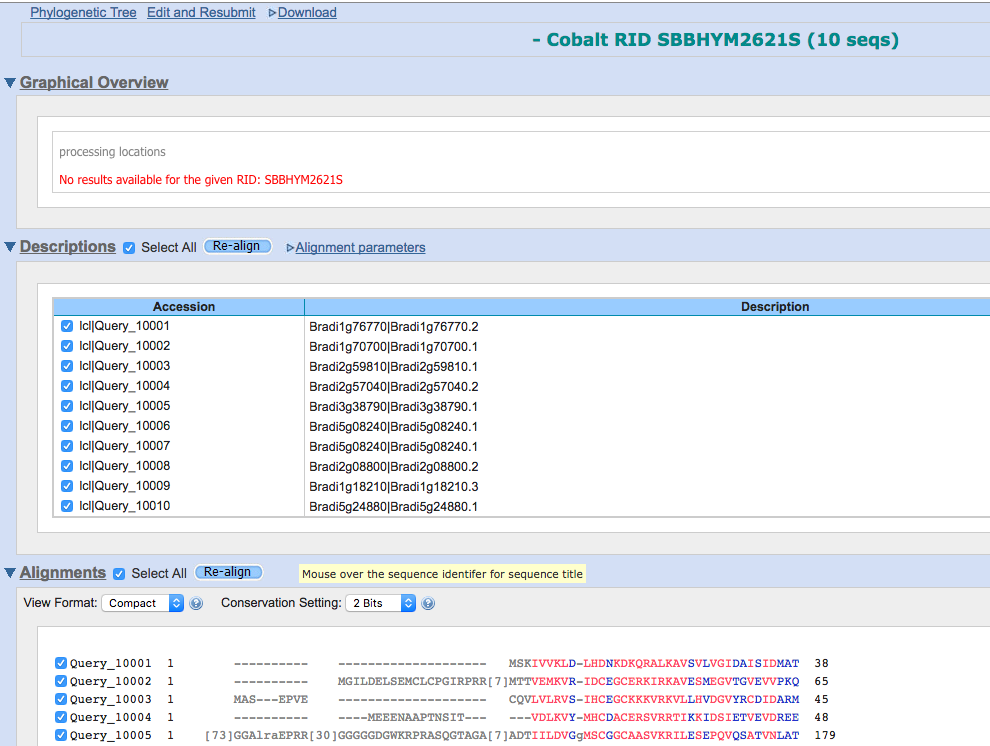
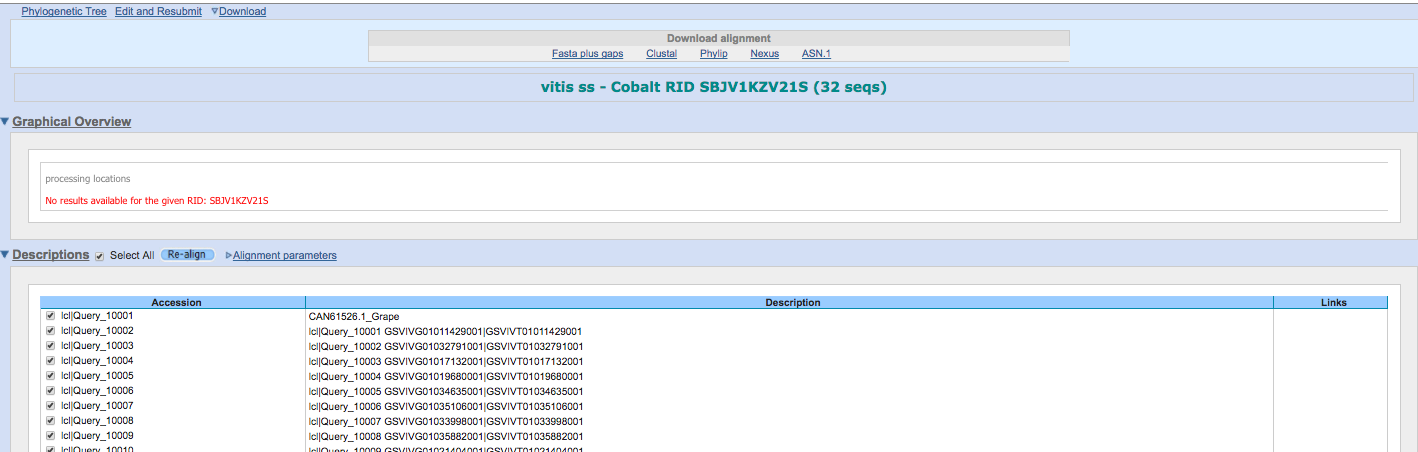


**Method**

1. Choose file upload (if you have used biomart this should be in your downloads folder).
2. Look at advanced options and consider if these are necessary
3. **Align**

**Results**

If a pop-up message “No results available for the given RID: SBBHYM2621S” appears on your screen close the message.



The dashes represent and insertion/deletion events in this sequence compared to others with which it is aligned. The regions which align well will be coloured. This is a good point to check if the NLR-ID is aligning to your other proteins with the same domain.

1. Click **download** then **gapped fasta**

**Are various options available which you can look at in more detail if these are of interest to you later.**

**Advantages**

Easy to use

Takes multiple formats

Lots of output information which is downloadable

Multiple output formats

**Limitations**

Limit on number of sequences

Less control over parameters

Can take longer or fail due to internet connection and weight of traffic on web page

Not the best aligner

**NB command line users:** You can use HMMER (hmmalign) and hmm for the domain of interest instead of webservers.

**Alignment curation**

**Tool:** Jalview

**Location:** Start >

**Method**

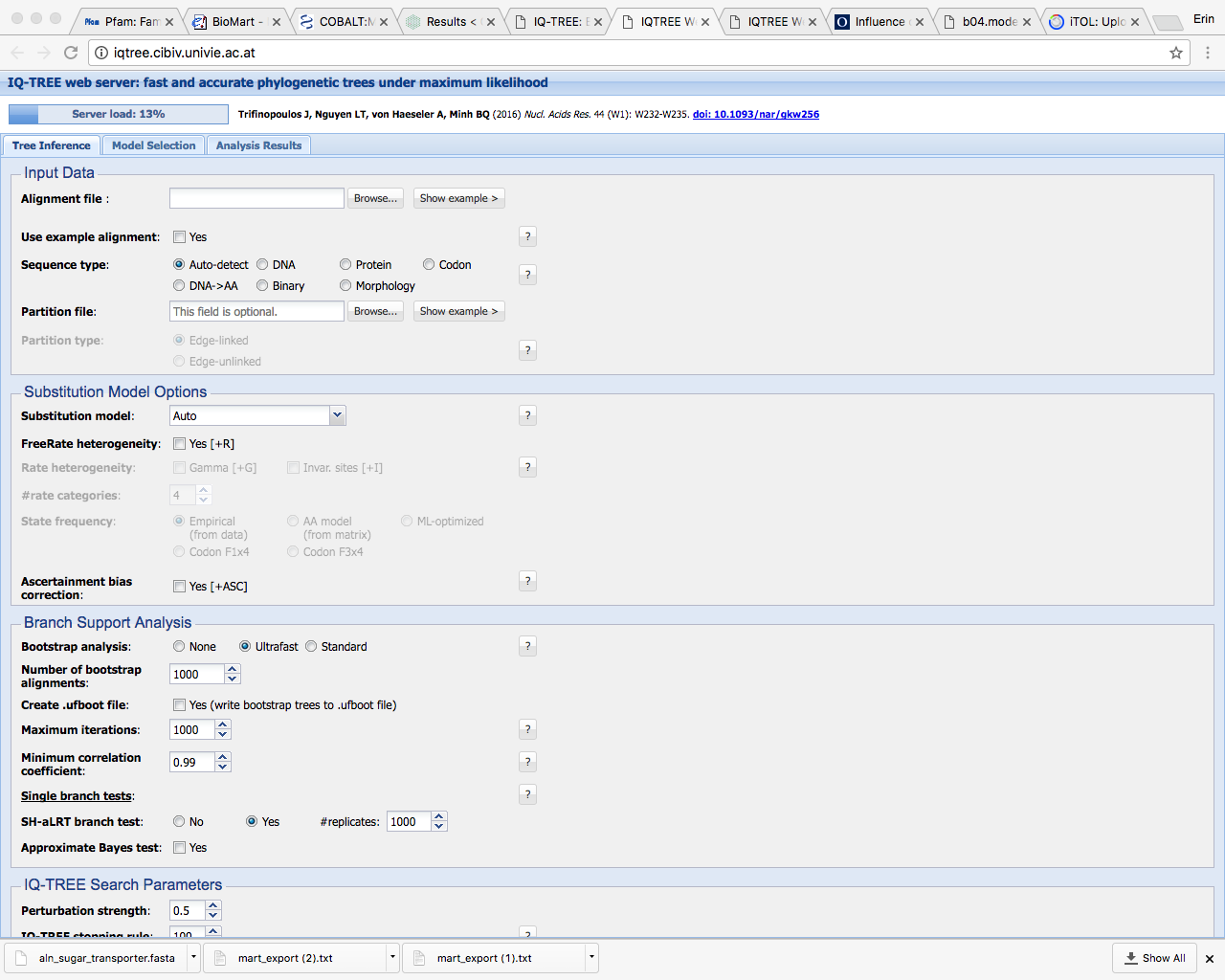
1. Visualise conservation of amino-acids, this can be done through clicking on colour then ClustalX, then By conservation .
2. Drag the scroll bar to look the whole way across the alignment, this will allow you to identify regions of high conservation which is also indicated by the bar chart at the bottom. This is the domain region which you want to keep.
3. Remove region composed of many gaps before where nearly all sequences clearly align. Do this by clicking on the first ‘-’ or letter in the first sequence and dragging across and down to cover the whole region to remove. Then navigate to edit at the top of the alignment window and click cut. Repeat this for the region after which the sequences do not align well. Jalview often requires you to drag from left to right when trying to select a region.
4. Highlight sequences, other than your NLR-ID, which do not align well using a mouse click on the sequence name and holding down Ctrl to highlight multiple at once. Remove these sequences using Edit - Cut.
5. Remove columns of all gaps by Edit > Remove empty columns
6. Save alignment File > Save as > Save in Project 1 in folder of NLR-ID you are working on

**NB command line users:** we have installed belvu to enable alignment viewing and editing

**Phylogeny construction**

**Tool:** IQ-Tree

**Link:** http://www.iqtree.org/



**Method**

1. Navigate to **Web Service**
2. Input data
   1. Alignment file : click option **Browse** then select alignment from clustal omega
   2. Sequence type: **Protein**
3. Substitution modal
   1. **Auto**
   2. **Free rate model** tick(This allows variation in substitution rate beyond gamma distribution bounds? I think.. )
4. **Submit** job

**Results**

IQtree will re-direct you to a results page and give you a link which you can return to later to find your results.

**Link to tree :**

Whilst IQ-Tree is running you can navigate to the run log tab to see the tree builders progress. The sidebar and email will inform you of when the run is complete.

Once finished you can view the log to see if there were any problems such as sequences removed as two where identical or sequences with few aligned columns being present in the alignment.

Click the **download** selected jobs button on the bottom left of the page.

Then rename the .contree file with your name and the domain of interest as this will make it easier to find your own tree later on iTOL.

**Advantages**

Allows use of complex models and options

Easy to use

Visually appealing

**Limitations**

Memory and run time are limited

Model selection is not tested after tree construction by IQ-Tree and this can result in over-fitting a model to your data (only a concern for those very interested in phylogenetics).

Phylogeny run time for 373 seq of length 100+ proteins - A whole day

<http://iqtree.cibiv.univie.ac.at/?user=guest&jobid=170805160731> (NLR-ID-5 example )

Phylogeny run time for 10 seq - A few minutes

<http://iqtree.cibiv.univie.ac.at/?user=guest&jobid=170603160036>

**NB command line users:** you can run your phylogeny with RaxML as outlined in project 2.

**Phylogeny visualisation**

**Tool:** iTOL

**Link:** <http://itol.embl.de/login.cgi>

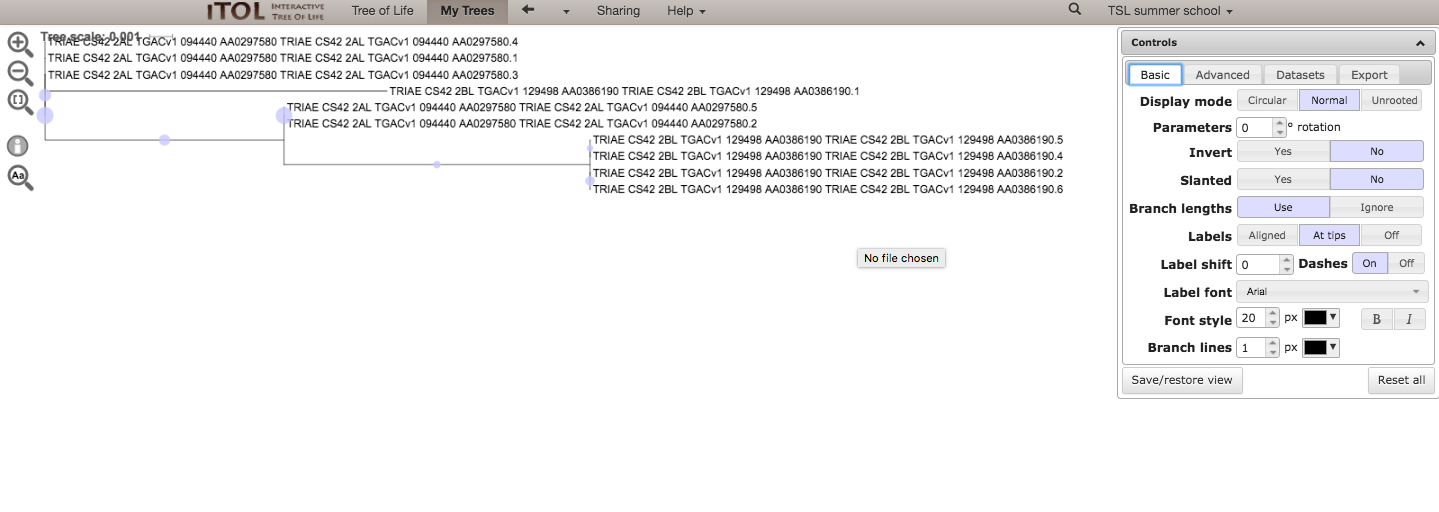
**Method**

1. Navigate to **login**

Username: TSLSummer2017

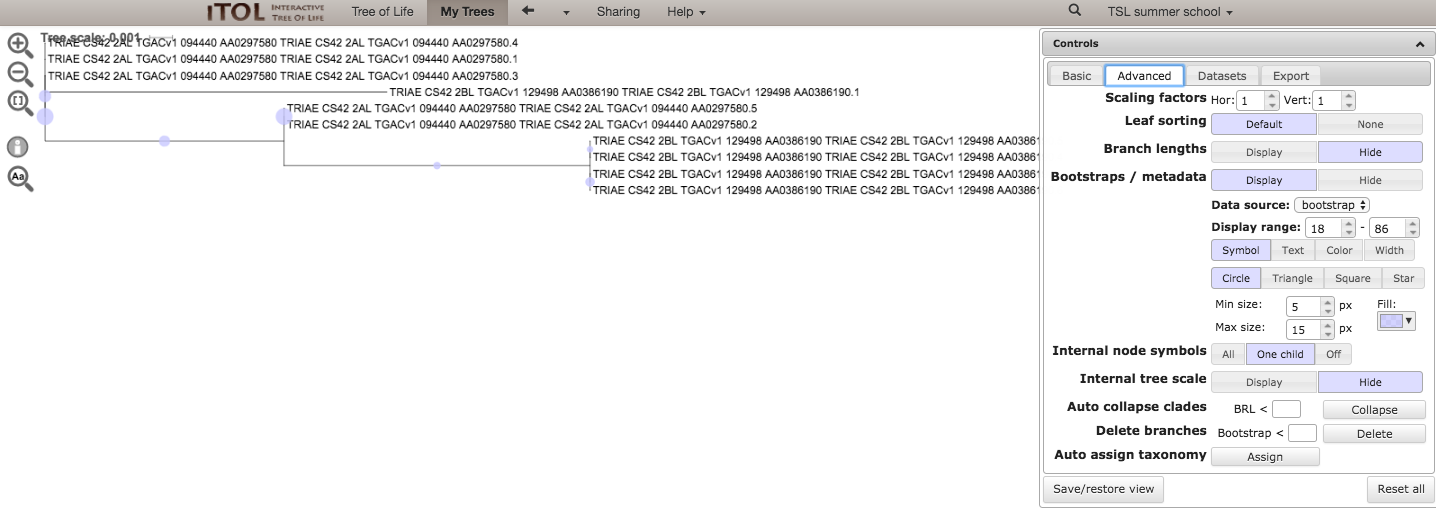
Password: path0gen

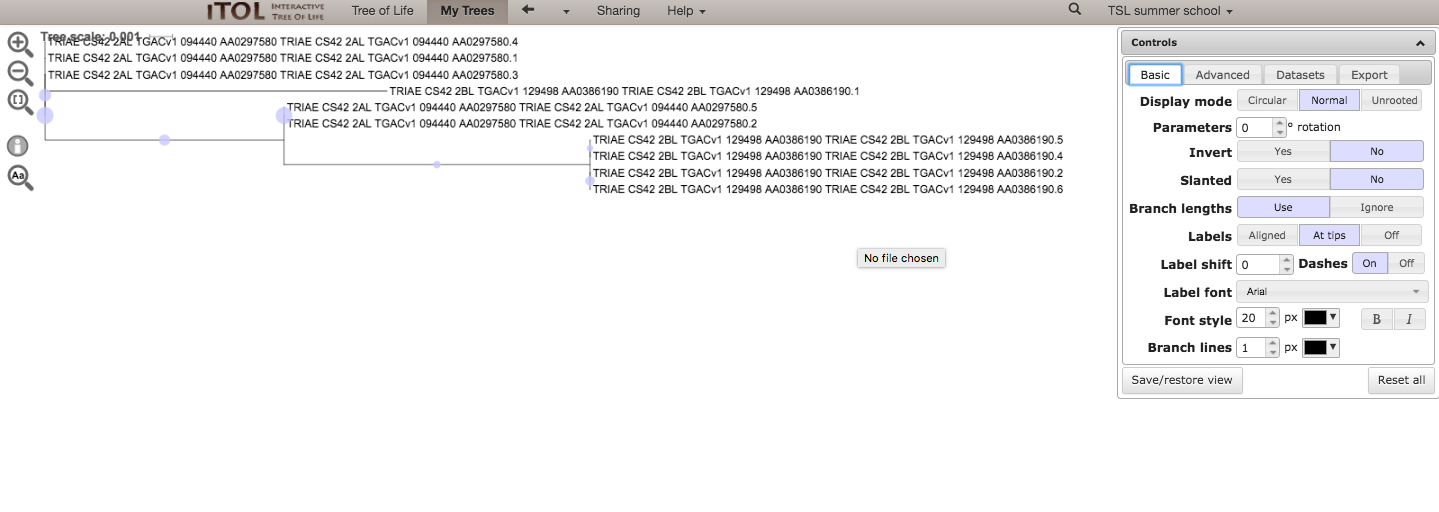
1. Navigate to **upload tree files** under TSL summer school project heading and upload the tree saved from IQ-Tree <your\_file\_name>.contree
2. Click on the blue hyperlink with the name of the file you uploaded
3. Visualising tree, alter the view of the tree using the different control options.









1. **Save/restore view**, recommended to do regularly to prevent losing work. Can save multiple views of the same tree. 
2. Exporting tree, click on the export tab of the control panel then you can use the format drop down menu to select more conventional formats such as PDF and PNG.

NB. Command line experts can try formatting data to from tree to be compatible with dataset display files. For more detail see iTol help pages. <http://itol.embl.de/help.cgi#annot>

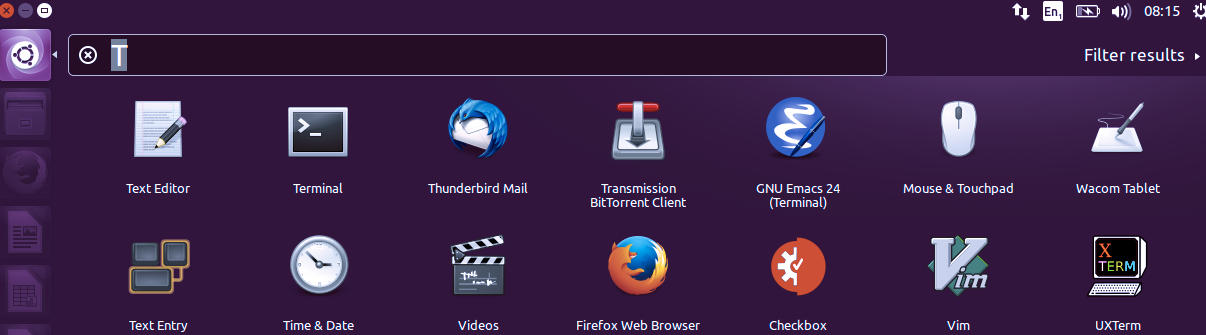
**KEGG analyses**

**Project 2 Annotation of NLRs, NLR-IDs and constructing NLR Phylogeny (Level 2 - command line skills)**

This command-line oriented project will show you how to extract all NB-ARC containing genes and construct their phylogeny with help of our plant\_rgenes pipeline (https://github.com/krasileva-group/plant\_rgenes) (Sarris et al 2016). If you are command-line novice - give it a try, you can alway go back and do most of this project with skill set you acquired in project 1. If you are command-line expert, take it to the end and annotate NLR-IDs and label where they occur on NLR tree.

**Cheatsheet for first time command line users**

1. Open Terminal



1. Navigate to our project by typing

cd ~/Documents/TSL\_SummerSchool/Project2/

cd change directory (navigate between folders and subfolders)

ls list all files and folders in current directory

head <filename> see first 10 lines of a file

less <filename> see content of a file (type ‘q’ to quit)

command -h or --help usually will display how to use it

Hitting TAB key helps you auto-complete file path

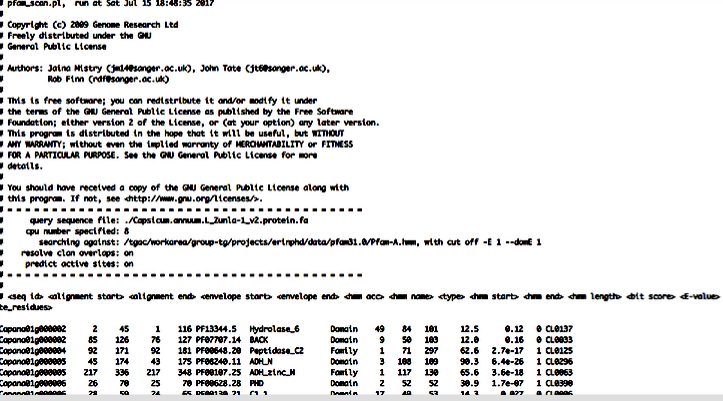
Ctrl A takes you to the beginning of command line

Ctrl E takes you to the end

Ctrl C exit a process or writing of a command

Hitting up arrow allows you to scroll through previous commands that you typed

**Pfam scan**



This step takes all the fasta sequences of proteins and scans the sequence for regions of similarity to the defined Pfam domains. If a match is found an e-value is assigned and the co-ordinates of where in the HMM and in the protein sequence the domain aligns to are given.

In general, you can search any HMM database using HMMER’s hmmscan, such as:

hmmscan -o <output.out> <db.hmm> <sequence.fa>

NB here and below: In examples, we sometimes enclose <filename> in <> brackets. This is to indicate that you should type in your specific filename here. The <> should be omitted when you put the specific name of your file.

Try it with your sequence from project 1. Hmmscan can output the data in different format, oriented towards the sequence, or each domain. Try putting options --tblout, --domtblout and --pfamtblout and see the difference.

Example

cd ~/Documents/TSL\_SummerSchool/Project\_1/nlr-id/Example\_NLR\_ID

hmmscan -o example.hmmscan\_db.out db.hmm example.fa

hmmscan -o NLR\_ID\_<no.>/<nlr\_id\_name>.hmmscan\_db.out ~/Documents/TSL\_SummerSchool/reference/pfamA\_db/Pfam-A.hmm ~/Documents/TSL\_SummerSchool/Project\_1/nlr-id/NLR\_ID\_<no.>/<nlr\_id\_name>

hmmscan --tblout example.hmmscan\_db.tblout db.hmm example.fa

hmmscan --tblout <NLR\_ID\_<no.>/<nlr\_id\_name>>.hmmscan\_db.tblout ~/Documents/TSL\_SummerSchool/reference/pfamA\_db/Pfam-A.hmm ~/Documents/TSL\_SummerSchool/Project\_1/nlr-id/NLR\_ID\_<no.>/<nlr\_id\_name>

**Processing**

For downstream processing, you need a very consistent format, ideally sorted by the order of appearance of domains in the sequence. Therefore, we like to use pfam\_scan.pl script distributed by the developers of Pfam database. Here is how you would use it for an example protein.

perl ~/Documents/TSL\_SummerSchool/scripts/pfam\_scan.pl -e\_seq 1 -e\_dom 1 -as -outfile example\_pfamscan.out -cpu 8 -fasta example.fa -dir ~/Documents/TSL\_SummerSchool/reference/pfamA\_db

However, when you analyze a whole proteome, this annotation process can run hours (or even days if you are using your old laptop). Therefore, we have run the following method for you on our clusters

**Method**

Prerequisites

hmmer-3.1b2

Pfamscan

Pfam-A.hmm

Pfam-A.dat

active\_site.dat

perl ~/Documents/TSL\_SummerSchool/scripts/pfam\_scan.pl -e\_seq 1 -e\_dom 1 -as -outfile example\_pfamscan.out -cpu 8 -fasta example.fa -dir ~/Documents/TSL\_SummerSchool/reference/pfamA\_db

The results are available in

~/Documents/TSL\_SummerSchool/reference/

**Pfamscan Reformat**

The output of a command line pfam scan is a large table where protein domains even with relatively high e-values (likely matched by chance) are retained, there is also no criteria for non-overlapping domains. We therefore need to reformat the output to an easier to interpret format and to retain only non-overlapping domains and later those of interest or with a significant evalue.

**Method**

Prerequisites

bioperl

plant\_rgenes

1. cd ~/Documents/TSL\_SummerSchool/Project\_2
2. perl ~/Documents/TSL\_SummerSchool/scripts/K-parse\_Pfam\_domains\_v3.1.pl -p <pfamscan-of-interest.out> -e 0.001 -o <Pfamscan\_name>.parsed.verbose -v T

Example

perl ~/Documents/TSL\_SummerSchool/scripts/K-parse\_Pfam\_domains\_v3.1.pl -p ~/Documents/TSL\_SummerSchool/Project\_2/pfamscan\_out/Athaliana\_167\_TAIR10\_pfamscan-12-13-2016.out -e 0.001 -o athaliana\_pfamscan-07-08-2-17.parsed.verbose -v T

-p|--pfam <pfamscan.out>

-e|--evalue <evalue cutoff>

-o|--output

-v|--verbose <T/F> default F. Display more information about each domain (start, stop, evalue)



1. python ~/Documents/TSL\_SummerSchool/scripts/filterDomainv2.py <\*.parsed.verbose> <domain of interest> <e-value> > <input file name>.<evalue.domain of interest>.txt

Example

python ~/Documents/TSL\_SummerSchool/scripts/filterDomainv2.py athaliana.parsed.verbose NB-ARC 0.001 > athaliana .0.001.NB-ARC.txt



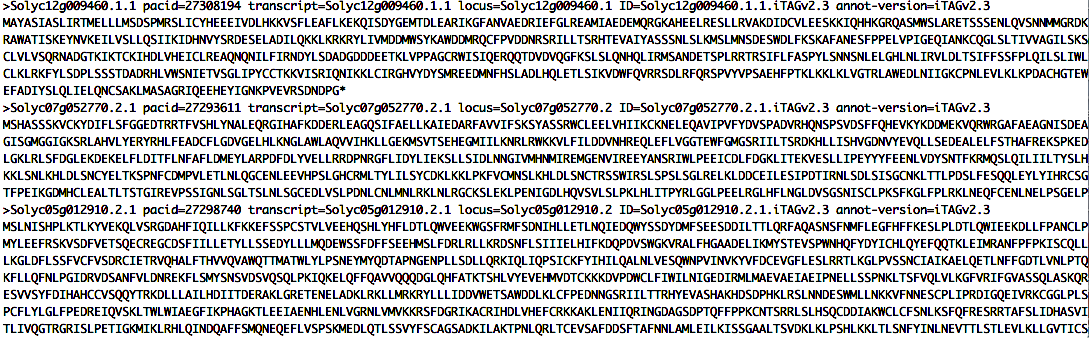
1. perl ~/Documents/TSL\_SummerSchool/scripts/K-get\_fasta\_from\_ids.pl -i <domain name .txt> -f ~/fasta/<species working on>.fa -o <input file name>.fa

Example

perl ~/Documents/TSL\_SummerSchool/scripts/K-get\_fasta\_from\_ids.pl -iathaliana.0.001.NB-ARC.txt -f ~/Documents/TSL\_SummerSchool/reference/proteomes/Athaliana\_167\_TAIR10.protein\_primaryTranscriptOnly.fa -o athaliana.0.001.NB-ARC.fa

-i |--ids location of the file with sequence ids

-f |--fasta location of the fasta file

-o|--output location of the output file

4. Command line

cat ~/Documents/TSL\_SummerSchool/NLR-ID/<\*.fa> >> <\*NB-ARC.fa>

Example

cat ~/Documents/TSL\_SummerSchool/NLR-ID\_<\*>/<\*>.fa> >> athaliana.0.001.NB-ARC.fa

**Alignment**

Now we have all the amino-acid sequences for the proteins with our domain of interest we can align these proteins to the hmm which represents that domain. From the alignment we should start to be able to see patterns which represent similarity of sequences to one another.

Since all NLRs have a conserved NB-ARC domain, you can align them using this domain as a common template. Go to <http://pfam.xfam.org/family/NB-ARC>

and download <http://pfam.xfam.org/family/PF00931/hmm>

If you are curious - look inside the .hmm file to see how the domain is described as a statistical model of aa probabilities at each position of the domain

Place this hmm file into you project 2 folder.

As a bonus, you can download any of the curated functionally annotated NLRs from <http://prgdb.crg.eu/wiki/Category:Reference_R-Genes,_manually_curated> and include them in alignment and phylogeny. Place them in project 2 folder as well.

**Method**

Prerequisites:

hmmer-3.1b2

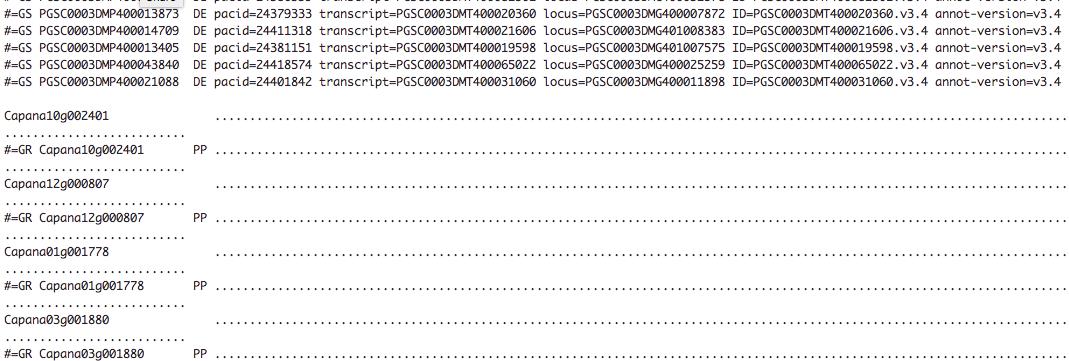
python\_anaconda-2.5.0

bioscripts.convert

1. hmmalign -o <domains of interest.fa>\_hmmalign.sth <Domain of interest>.hmm <domains of interest.fa>

Example

hmmalign -o athaliana.0.001.NB-ARC\_hmmalign.sth ~/Downloads/NB-ARC.hmm athaliana.0.001.NB-ARC.fa



Hmmalign produces an alignment in Stockholm format however for visualisation purposes we would like the alignment in fasta format.

1. convalign fasta <domains of interest\_hmmalign.sth >

Example

convalign fasta athaliana.0.001.NB-ARC\_hmmalign.sth

**Belvu alignment view**

We can use visualisation of the alignment to crop the alignment to just the region which aligns to our domain of interest. This means we remove the rest of the sequence which in some proteins may encode the NB-ARC or LRRs for instance. Where you start to see that most sequences align with less dashes and more colour this is where the aligner is predicting sequences represent the same position in the protein.

The alignment can be viewed using the programme belvu.

**Method**

1. belvu <\*hmmalign.fasta> 2> <\*hmmalign.fasta.log>
2. Right click on the dash before the first region of high conservation, in the example I have marked the position with an arrow.
3. Edit -> Remove columns left of selection....
4. Then right click on the residue flanking the last region, then use Edit -> Remove columns right of selection

Now we have our region of interest we can clean up the alignment

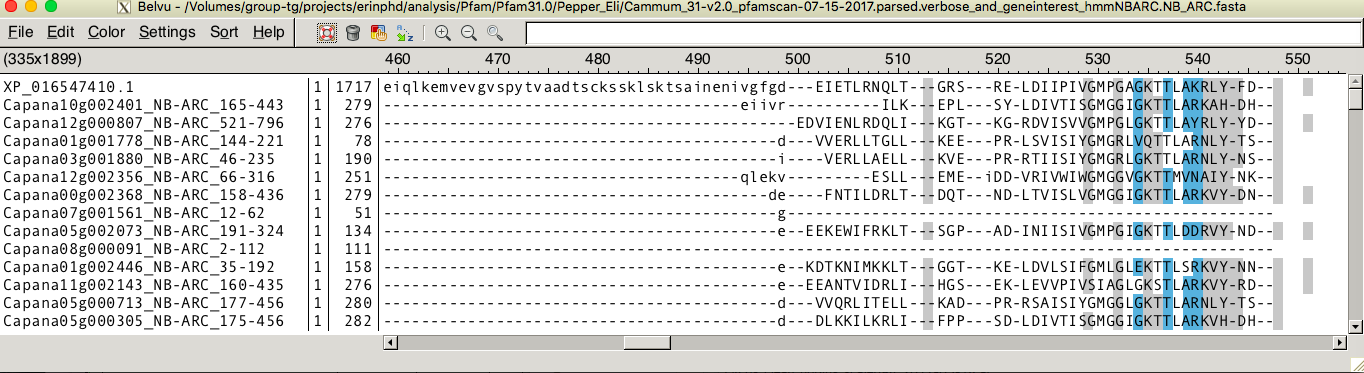
1. Edit remove gappy columns
   1. 90% gaps

I have chosen 90 here as an arbitrary cut off but you can use more stringent cut-offs however as there is no easy undo in belvu if you remove too much of the variation in your sequences you may have to start refinement again.

1. Repeat the above but this time removing gappy sequences.

Once you are happy with your curation then save.

1. File -> Save as <input>.belvu.fa



**Phylogeny with RAXML**

We are now ready to build a tree of the protein domains to visualise how they maybe related evolutionarily. For this we are going to use the RAXML programme to build a bootstrapped maximum likelihood tree.

**Method**

Prerequisites

raxml

1. raxmlHPC-SSE3 -f a -x 1123 -p 2341 -# 100 -m PROTCATJTT -s <input alignment> -n <input alignment>.raxml

Example

~/biotools/standard-RAxML-master/raxmlHPC-SSE3 -f a -x 1123 -p 2341 -#100 -m PROTCATJTT -s cannum.0.001.NB-ARC\_hmmalign.belvu.fa -n cannum.0.001.NB-ARC\_hmmalign.raxml

-f | rapid Bootstrap analysis and search for best-scoring ML tree in one program run

-x | Random starting seed for bootstrapping

-# | Number of bootstraps

-p | Random starting seed for maximum likelihood tree

-m | Model used to explain likelihood of amino-acid change

-s | Multiple sequence alignment

-n | Output file prefix

-U | Skip gappy columns NB only use if have relatively few gaps in alignment

**Itol visualisation**

As in level 1 guide but choose file RAxML\_bipartitionsBranchLabels\*.raxml

**Annotating NLR-IDs with plant\_rgenes**

Through using a final script we can create a table to clearly show us the numbers of different subtypes of NLR in the species of interest and the domains found to be part of the NLR-IDs.

In order to create this table we first need a metadata table which the script calls to identify the species and family to which it belongs. The metadata table is also a useful way to store information about the source of your raw data.

**Method**

Prerequisites

bioperl

plant\_rgenes

metadata table

The metadata table is a tab separated file. Open the file using nano or your preferred text editor.Do not open in excel as this often distorts the tab separation. Be careful not to add in extra spaces and that the species\_ID column entry exactly matches the prefix of your .parsed.verbose file. If you wanted to include multiple species proteomes in an analysis then you just include information on them in the metadata table and locate the respective \*.parsed.verbose file in the same directory. If you are not using Arabidopsis then replace the details for those of the species you are studying if you are not sure on what to fill in a field ask a demonstrator or use NA.

1. nano ~/Documents/TSL\_SummerSchool/Project\_2/test\_dataset/db\_descriptions.txt
2. Ctrl + O Enter (To save changes)
3. Ctrl + X (To exit)

Whilst in the directory of the parsed.verbose file you can then run the perl script.

1. cd ..
2. git clone https://github.com/krasileva/plant\_rgenes.git
3. perl plant\_rgenes/processing\_scripts/K -parse\_Pfam\_domains\_NLR-fusions-v2.2.pl -i <folder containing parsed.verbose> -o <folder for output> -d <path to metadata table> -e <e-value for calling an ID domain>

Example

cd ..

perl ~/Documents/TSL\_SummerSchool/plant\_rgenes/processing\_scripts/K -parse\_Pfam\_domains\_NLR-fusions-v2.2.pl

-i ~/Documents/TSL\_SummerSchool/Project\_2/

-o ~/Documents/TSL\_SummerSchool/Project\_2/

-d ~/Documents/TSL\_SummerSchool/Project\_2/test\_dataset/db\_descriptions.txt

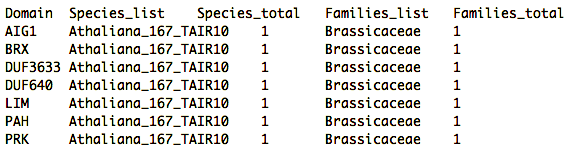
Output file 1 - nlrsd-summary\*

This file provides a summary of the total number of NLR, TNLs, fusions and fused domains find in each proteome.

Screen Shot 2017-08-07 at 10.13.55.png

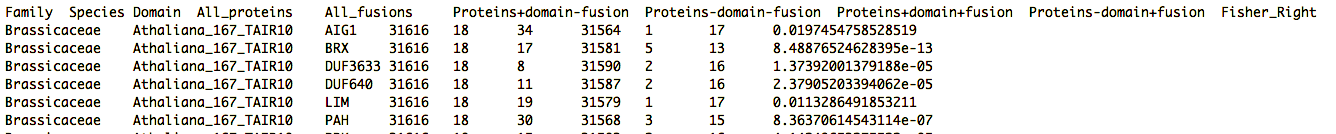
Output file 2 - nlrsd-prevelance

A record of how many times a fused domain is found at a species and family level.



Output file 3 - nlrsd-domains

Provides statistics to describe the likelihood the domain was not found by chance.



Output file 4 - nlrsd\_by\_prevalence\*wordcloud

An ordered list of domains which can be used to generate wordcloud image

Downstream example analysis: integrated domains and the KEGG database

Step 1) identify your integrated domain on the KEGG database (sucrose synthase)

<http://www.genome.jp/kegg/kegg1a.html>

Step 2) identify related NLR-IDs

BLAST

