Exercise 7:

Results Interpretation

Aim:

- Read SamBada output
- Apply multiple testing corrections
- Investigate spatial patterns of associations
- Visualize genomic background of associations
- Validate the results via annotation and sequence similarity

INTRODUCTION

Genotype-environment association models can produce a high rate of false positives. For this reason, it is important to learn how to control the false discovery rate and how to manually check the results in order to detect meaningful relationships. In this section we will process the SamBada output in R and then try to validate the best associations by using online genomics databases.

Data Access

The genetic and the environmental data, as well as the SamBada output you produced during the previous exercise will serve as input here. If you missed a step, you can find any data you need in the course website as well. To start this exercise, you will need:

- MOOA_ENV_ps.shp (Morocco environmental data)
- fSNPS.robj (filtered genotype matrix Morocco)
- mol-data-MOOA-Out-1.txt (output of SamBada for univariate models from Morocco)

- StSn_env.shp (Australia environmental data)
- StSn_fSNPS.robj (filtered genotype matrix Australia)
- StSn_PCA.robj (genetic structure data for Australia)
- Mol-data-StSn-Out-3.txt (output of SamBada for trivariate modesl from Australia)
- StSn_SNPseq.txt (table contain the sequence for each SNP-tag).

- plot_map_gradient.Rfun (function to plot a gradient on a map on R)
- qvalue_by_env.Rfun (function to compute qvalues by environmental variables)
- showbest.Rfun (function to show best significant models from SamBada)
- get.gen.position.Rfun (function to retrieve chromosome and position of a SNP).

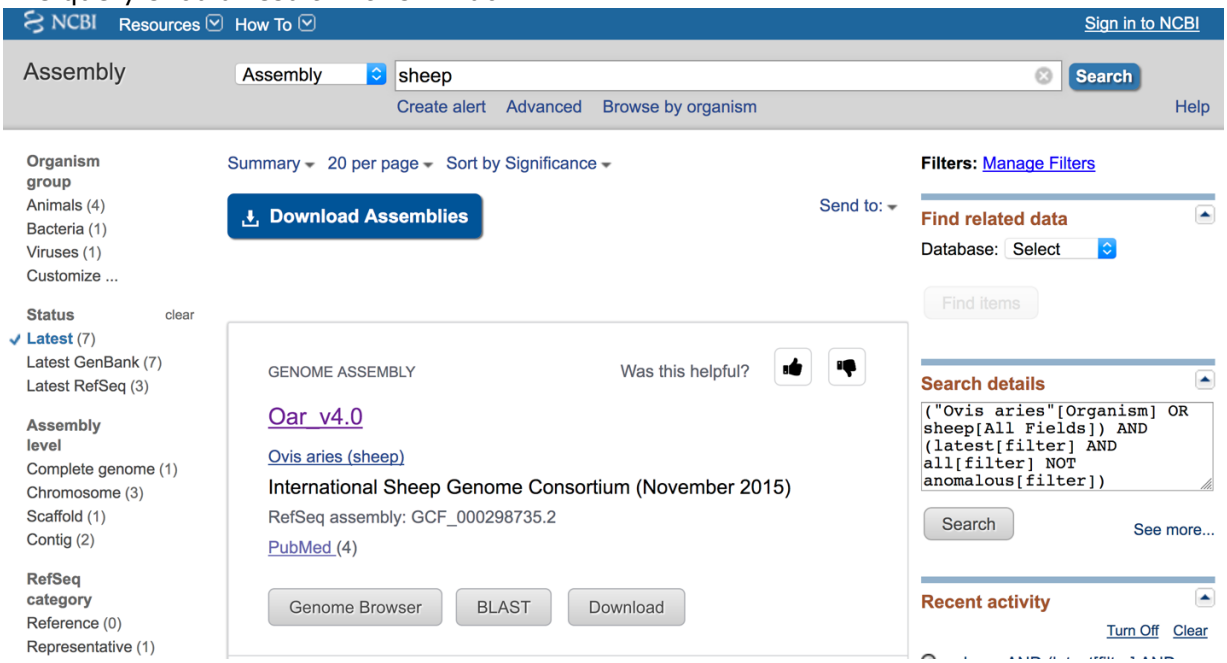
EXERCISE

Read SamBada output

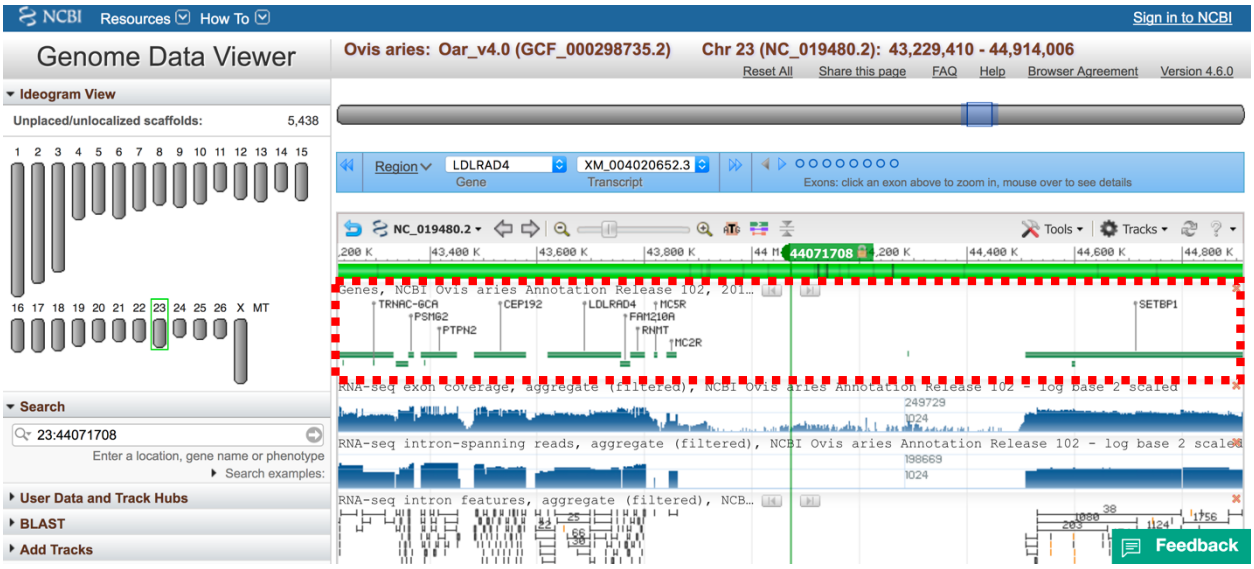
We will start working on the dataset of sheep from Morocco. We ran SamBada setting the flag of outputting all the models. This means that all models are present and they are ranked by significance. Since for Morocco we ran only univariate models, we'll have to read the results on the file: *Mol-data-MOOA-Out-1*. Usually SamBada filters the results using a corrected p-value using the Bonferroni method. In our case, we will use a less conservative approach, the q-value method correction. The aim of this approach is reduce false positives due to the fact that a huge number of tests has been performed (number of markers x number of environmental variables).
The q-value correction we are using is implemented by the qvalue R package. This package is accessible from an online repository of R packages dedicated to biology (bioconductoR).
Create a Rstudio project in the working directory, open the *read_output_Morocco.R* script and follow the instructions.

Validation via annotation

Reference genomes are nowadays available for a wide range of species and can be greatly beneficial for the interpretation of the results of a landscape genomics study. More specifically, annotations (i.e. the lists of genetic features associated to reference genomes) can be used to understand the functional role of a genetic region.
During the previous section, we identified a genetic marker on position 44071708 of chromosome 23 that is associated to the bioclim19 variable (precipitation in the coldest quarter). Go to the NCBI website: https://www.ncbi.nlm.nih.gov/. NCBI is a repository of genetic sequences of various type and represents one of the most employed databases to retrieve genomic data. In the search box, choose Assembly and look for the word "Sheep". The query should result in this window:



Now click on genome browser. You will access to the genome browser interface.



In the left panel, you can perform a search for a specific genome location by typing chromosome#:position. In the case here above we looked for the marker we identified with SamBada. Use the icons on the upper panel to move across the chromosome and to zoom-out to a larger extent. The dotted red rectangle indicates where the annotations are listed. You can see that on the range we are observing there are several genes (the green boxes).

Q1) Which genes are the closest to our SNP of interest? What is their role? Does it make sense with the environmental gradient it is associated to?

This approach allows to filter the findings of a landscape genomics study and fits well with the exploratory nature of experiments of this kind. Nonetheless, it does not allow a full validation since it is usually hard to draw an unquestionable link between the molecular function of a gene and the selective pressure that might cause adaptation.

You can now try to validate other associations found via SamBada. Remember that it is rare that the SNP found as associated is the one actually under selection. Most of the times, it is useful to check in the region around a SNP. Keep in mind that a mutation can modify the action of a gene in several ways and that generally speaking the genes closer to the mutation are more likely to be affected. As a rule of thumb, consider that one mutation on a regulatory region can impact the activity of a gene located up to 1 MB away, so any gene outside this window is very unlikely to be affected.

Can you find other promising associations in the SamBada output?

Australia Case Study

The case study of the Australian Stripey Snapper brings some extra complications to the table. The first one consists in the fact that this population shows a stressed genetic structure

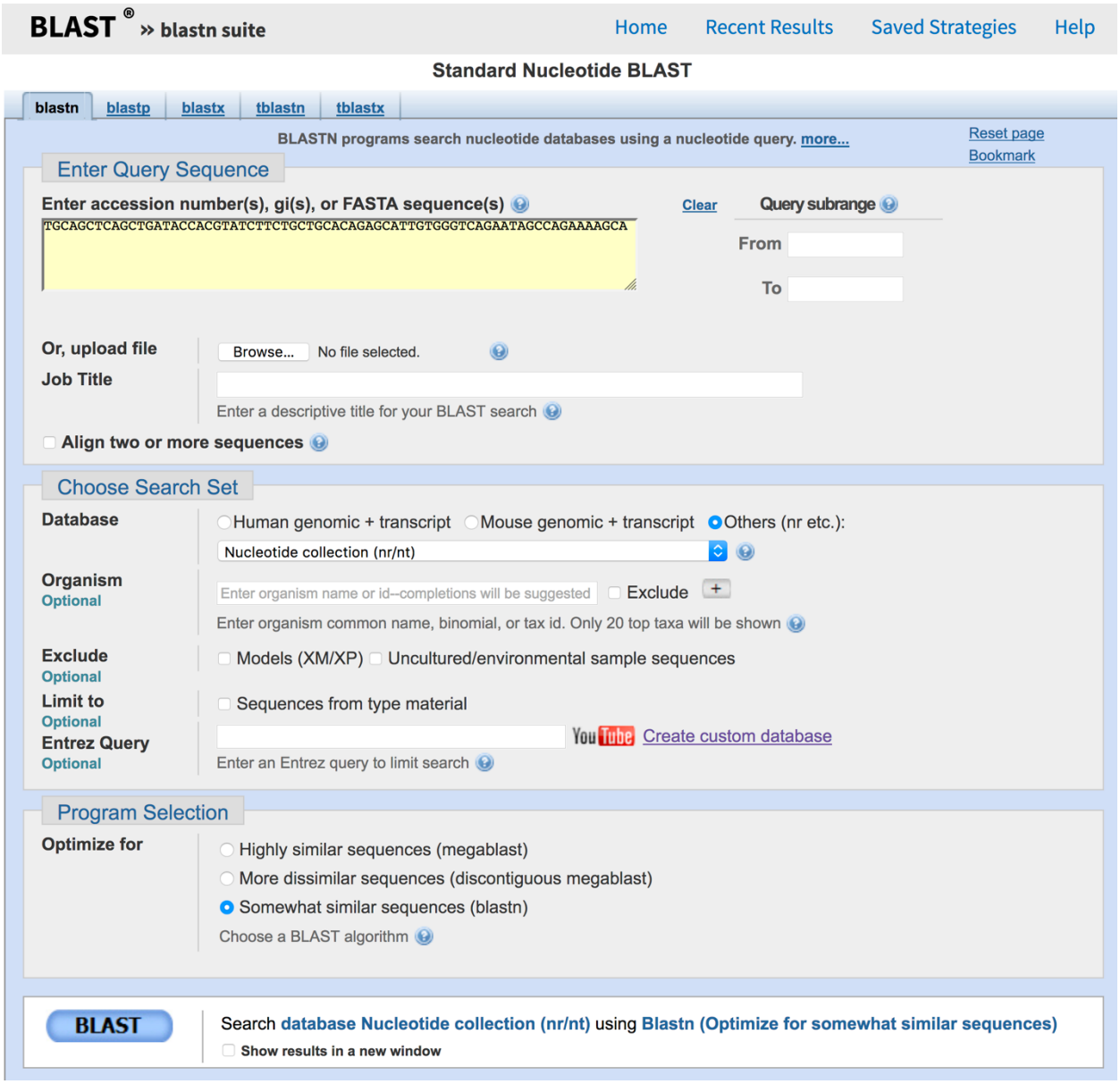
along a north-south axis. Since there are many environmental variables that aligns similarly (for instance temperature), it is necessary to correct the association models for population structure, which is what we did by running SamBada with tri-variate models. Another issue concerns the fact that we are working with a non-model species that does not have a reference genome available and for which little is known about its genes. Here we won't be able to navigate across a genome browser but we will have to rely on another approach.
You can start exploring the results of SamBada by opening the *read_output_Australia.R* script and following the instructions.

Validation by Sequence Similarity

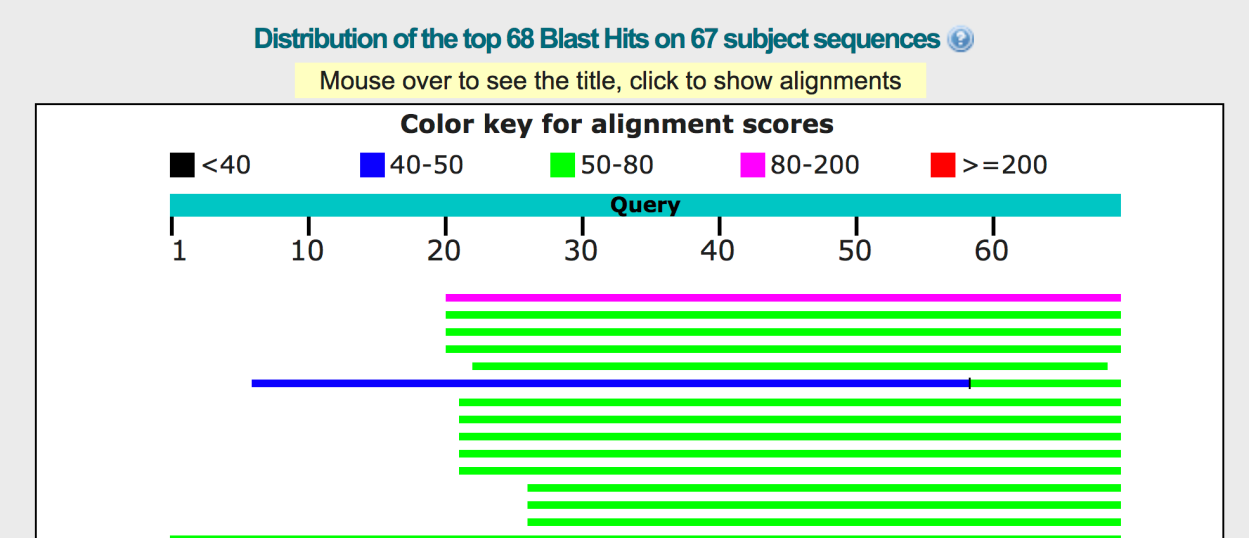
When there is no reference genome available the quest for the putative function of a SNP becomes more challenging. In the worst case, there is only a nucleotide sequence associated to the SNP of interest and it is necessary to get the most out of it. A common strategy is to employ a sequence similarity search across repositories of genomic data. The idea is that genes and regulatory regions are usually well conserved among relative species and we could therefore discover the putative role of our SNP by finding similar sequences.
We try this approach for the SNP we identified with SamBada:

>snp009567
TGCAGCTCAGCTGATACCACGTATCTTCTGCTGCACAGAGCATTGTGGGTCAGAATAGCCAGAAAA
GCA

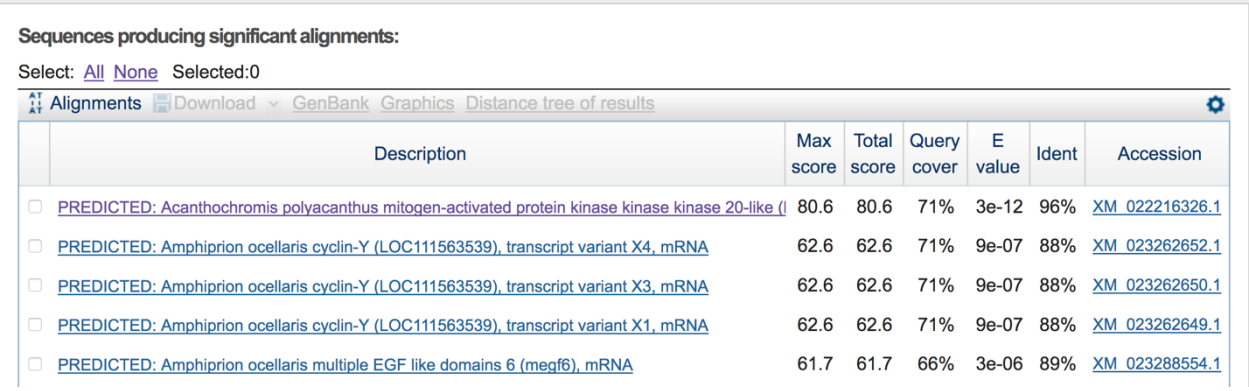
As a reminder, this SNP was found associated to Tm006 (i.e. the average sea surface temperature in June).
NCBI provides an online tool to query for sequence similarity across its databases. This tool is called BLAST (basic local alignment search tool) and comes in different flavors, depending on the type (nucleotide or protein) of the query (our sequence) and subject (the one from the database). Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi and choose Nucleotide Blast. You will get to this page:



Copy paste the nucleotide sequence in the "Enter accession number…" box. The central box of this page allows to customize your search. For instance, you can look within a specific database (e.g. whole genome sequences or transcripts only) and filter the search by species. For our case, we interrogate the *Nucleotide collection (nr/nt)* database which merges non-redundant data from various sources (click on the question mark for more information). The box on the bottom of the page allows to choose a specific algorithm to run the search. In our case, we use the "blastn" method because it is more suitable for short input sequences (as ours). After you have set these parameters, click on the BLAST button. The job will take some minutes to run and should bring you to this result:

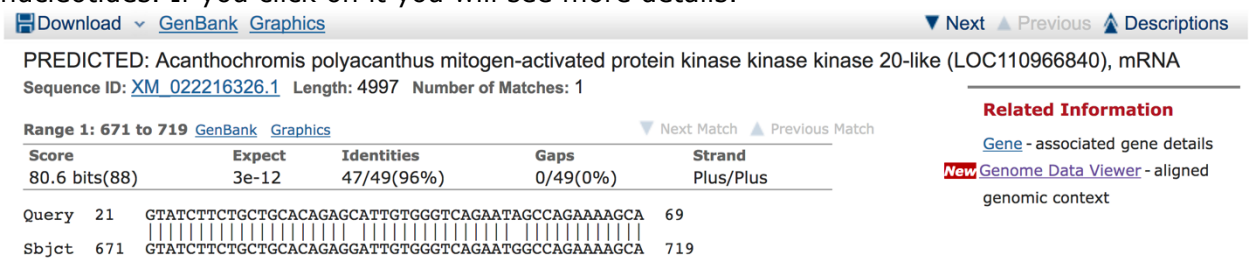


This graphic summary shows the matches found between our sequence (the query) and the sequences from the database. For instance, the purple box represents the best match: a sequence that aligns with more than half of our query with a high score. If we scroll down we can see the same alignment written as text.



Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Acanthochromis polyacanthus mitogen-activated protein kinase kinase kinase 20-like (I	80.6	80.6	71%	3e-12	96%	XM_022216326.1
PREDICTED: Amphiprion ocellaris cyclin-Y (LOC111563539), transcript variant X4, mRNA	62.6	62.6	71%	9e-07	88%	XM_023262652.1
PREDICTED: Amphiprion ocellaris cyclin-Y (LOC111563539), transcript variant X3, mRNA	62.6	62.6	71%	9e-07	88%	XM_023262650.1
PREDICTED: Amphiprion ocellaris cyclin-Y (LOC111563539), transcript variant X1, mRNA	62.6	62.6	71%	9e-07	88%	XM_023262649.1
PREDICTED: Amphiprion ocellaris multiple EGF like domains 6 (megf6), mRNA	61.7	61.7	66%	3e-06	89%	XM_023288554.1

The best alignment covers 71% of the query and this match resulted in 96% of identical nucleotides. If you click on it you will see more details.



Q2) To which species does the Subject sequence belongs to? What kind of sequence is it? What is its putative function? Could this be linked to the environmental variable associated to the SNP?

You can now go back to the SamBada results and try to validate other significant associations.

Can you find other promising associations in the SamBada output?