Haplotype analysis of the CDS in Arabidopsis thaliana

Magdalena Julkowska

Abstract

The simple pipeline to examine the natural variation in the coding region of your favourite gene, that you either picked by GWAS or just working on for the entire life and would like to examine what else is there to it. The pipeline described in this protocol is based on the SNP data from Arabidopsis, and so you will miss the variation due to non-Col-0 specific insertions or deletions. Nevertheless, this pipeline will give you an insight in the possible alterations of the coding region of your favourite protein

Citation: Magdalena Julkowska Haplotype analysis of the CDS in Arabidopsis thaliana. **protocols.io**
dx.doi.org/10.17504/protocols.io.i2gcgbw
Published: 12 Apr 2018

Protocol

Step 1.
Download the sequence of the Gene of Interest in all HapMap accessions from 1001 genomes SALK Genome Browser DB (http://signal.salk.edu/atg1001/index.php) and save them in text-editor (make sure to remove all "[", "]", and "-". Dots representing the missing data have to stay in the file as it is Alternatively – more genotypes are available from http://tools.1001genomes.org/2.

Step 2.
Make sure that the GENE OF INTEREST is highlighted and the CDS is spelled in capital letters – this is crucial for later!

Step 3.
save the file in the "Plain Text Module" as ".txt".

Step 4.
Start your terminal and type in following, to remove all the bits that are not CDS

Step 5.
Change your directory to where you have your .txt files stored by typing - "cd /Users/julkowmm/Desktop/Haplotype"

Step 6.
Remove all the "small" letters representing non-CDS sequences by typing - "cat AT5G66270_Zink_finger_CDS.txt | perl -e '$count=1; foreach $line () {if ($count==1) {print $line; $count=$count+1;} elsif ($line=/>/) {print "\n$line";} else {chomp $line; print $line; }}' | sed -e 's/^.*rev//g' | sed '/^>/! s/[atgc]//g' | sed -e 's/\.//g'> AT5G66270_Zink_finger_protein.txt ". Make sure to replace "AT5H66270_Zink_finger_CDS.txt" and "AT5G66270_Zink_finger_protein.txt" by the name of your input and output file respectively

Step 7.
Put the sequences in FAbox for collapsing them into individual haplotypes Go to: http://users-birc.au.dk/biopv/php/fabox/click on "DNA to haplotype collapser and converter" insert

your sequences in the window and collapse the sequences

Step 8.
Download the individual Haplotype sequences as well as the file where you have which accessions are represented by which phenotype.

Step 9.
Translate the haplotype sequences into protein using JalView - to download JalView - http://www.jalview.org/

Step 10.
Sort the aligned sequences by pairwise comparisons – eliminate Haplotypes which are exactly the same on protein level

Step 11.
Remove redundant haplotypes from the input file and start from #10 IMPORTANT: make sure that you note which haplotypes were collapsed together – otherwise you will end up with quite a big mess

Step 12.
Match the accessions with the specific haplotype groups as identified with FaBox collapser, and replace the haplotypes that could be further collapsed into one group based on the amino acid sequence

Step 13.
Match the accessions and the haplotype grouping with the phenotypes

Step 14.
Examine the significant differences between your haplotype groups. ADVICE: use only haplogroups that consist of 3 or more genotypes. Otherwise the haplotype data analysis is not very reliable.

Step 15.
Select the haplotypes that are giving you the most interesting / significant differences in your phenotype data and isolate the amino acid sequences of those specific haplotypes in a separate ".txt" file

Step 16.
Import the ".txt" file with the most interesting haplotypes into Jalview and color the amino acids by ClustalX in JalView

Step 17.
Format the final alignment by wrapping the text, adding scale above, below and right, boxes, txt, show gaps. The alignment should be 125 amino acids long (otherwise not readable at print-out) and saved in the PDF format.

Step 18.
Check the location of the known protein domains with Uniprot - http://www.uniprot.org/ - and mark the identified domains in the PDF. Check whether different haplotypes have altered sequence in identified protein domains to further build up your hypothesis.

Step 19.
Pick the most interesting haplotypes for cloning and allele complementation, based on the differences observed in the phenotypes between different "haplogroups", as well as protein sequence.