PlotDivergence for plotting the natural variation in sequence identity of the chosen genetic locus encompassing multiple ORFs

Magdalena Julkowska[1], Like Fokkens[2]

[1]King Abdullah University of Science and Technology, [2]University of Amsterdam

Working

Sep 27, 2018

dx.doi.org/10.17504/protocols.io.t2ieqce

Salt Lab KAUST

Magdalena Julkowska

King Abdullah University of Sc...

ABSTRACT

This in silico protocol allows to have an insight into the sequence divergence of selected genomic region, including missing information, gaps and allignment with Open Reading Frames. The protocol is written for model plant Arabidopsis thaliana, but can be possibly adapted to plotting the divergence of any organism, for which multiple accessions were resequenced. We used this script to produce divergence plots in our paper

Magdalena M. Julkowska, Karlijn Klei, Like Fokkens, Michel A. Haring, M. Eric Schranz, Christa Testerink; Natural variation in rosette size under salt stress conditions corresponds to developmental differences between Arabidopsis accessions and allelic variation in the *LRR-KISS* gene, *Journal of Experimental Botany*, Volume 67, Issue 8, 1 April 2016, Pages 2127–2138, https://doi.org/10.1093/jxb/erw015

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Magdalena M. Julkowska, Karlijn Klei, Like Fokkens, Michel A. Haring, M. Eric Schranz, Christa Testerink; Natural variation in rosette size under salt stress conditions corresponds to developmental differences between Arabidopsis accessions and allelic variation in the LRR-KISS gene, Journal of Experimental Botany, Volume 67, Issue 8, 1 April 2016, Pages 2127–2138, https://doi.org/10.1093/jxb/erw015 link

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

BEFORE STARTING

1. Download gnuplot (http://www.gnuplot.info/download.html) and install it by putting the download folder into your "Applications"
2. Download ClustalO (http://www.clustal.org/omega/) and install it by putting the download folder into your "Applications"

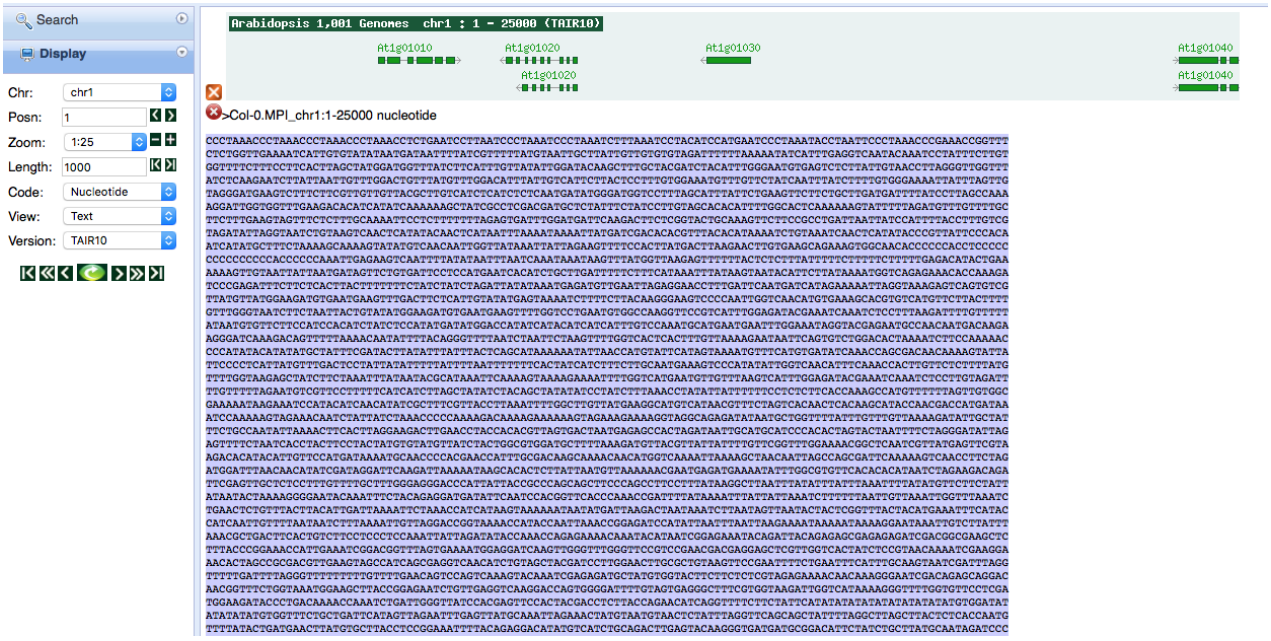
You can best do the installation by using HomeBrew package by typing the following command in your terminal window:

```
usr/bin/ruby -e"$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

Once you have the Homebrew installed, you can install packages by typing in your terminal following command:

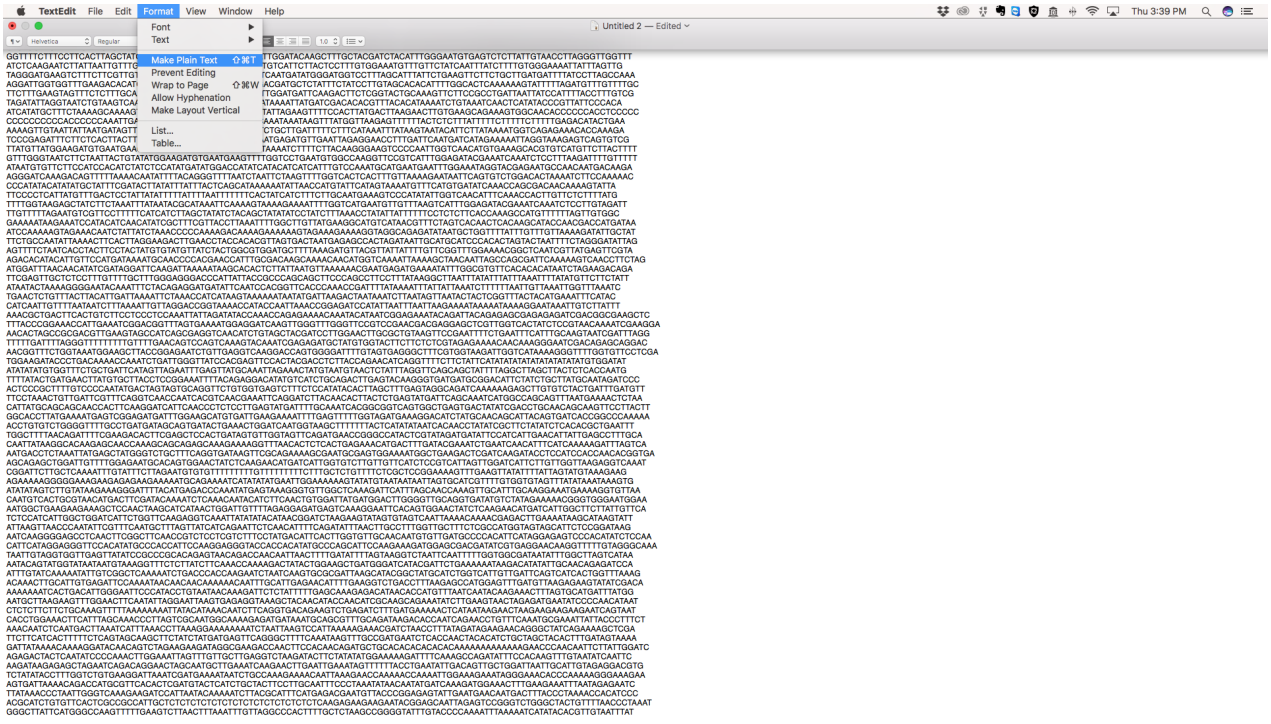
```
brew install gnuplot
```

1 Download the sequences of the group of accessions you want to compare from 1001 genomes SALK Genome Browser DB 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



2 Make sure that Col-0 is the first sequence in the text-edit file with all the sequences from different accessions

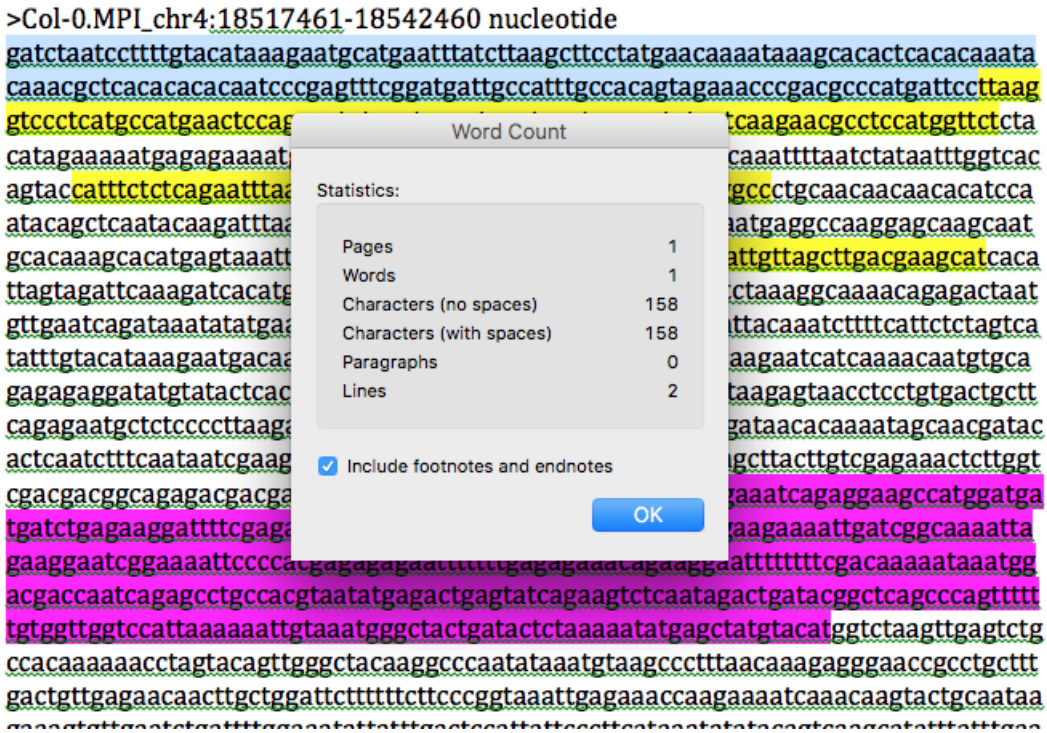
3 Make sure to save the file in the “Plain Text Module”



4 Save the file as .fasta by replacing the .txt suffix

5 Copy the Col-0 sequence into the word document and make sure that no “ENTERS” are present and that all ORFs from all the genes in the region are highlighted (I usually use different colors for different genes)

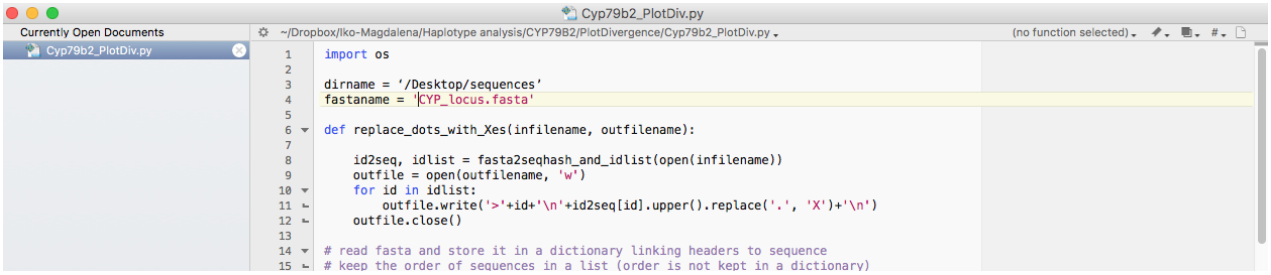
6 Note the position of the ORFs in the Col-0 sequence in format: At4g10310 – (2,200),(400,500) - meaning that the gene At4G10310 is having first exon in the copied sequence starting from the second character untill the 200th character, and the second exon is positioned between 400th and 500th chatacter. You can do it in word by clicking at the lower pane saying Words:... (next to the pane indicating the page number you are currently on).



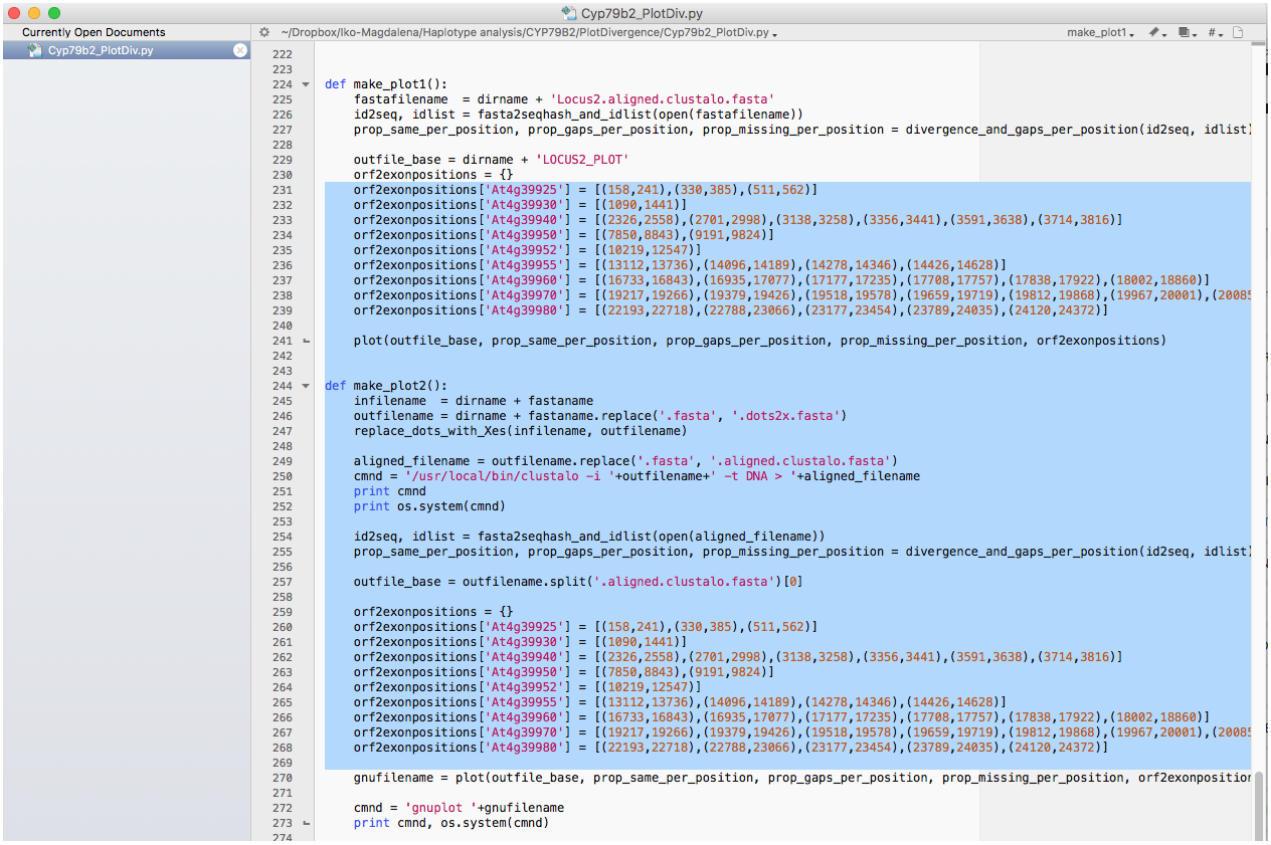
7 Open the plot_divergence.py in text editor specific for scripts - I strongly encourage you to use “text wrangler” or "atom" - txt editors designed for scripts will give you less errors when editing since they are highlighting different functional modes of the script.

plot_divergence.py

8 Insert the name of the folder / location where your fasta file is located and tell the software the name of your file



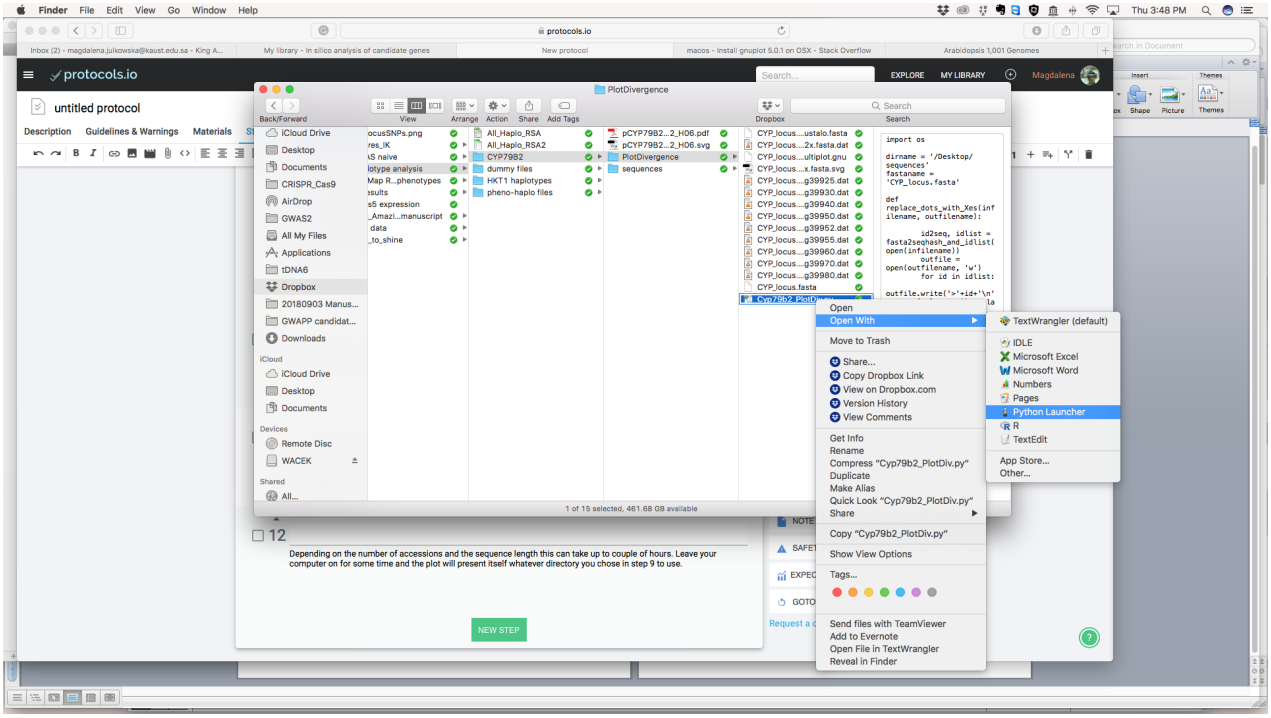
9 At the bottom of the script you have to give the coordinates of the ORFs of your different gene – change accordingly to step 6. WATCH OUT – you have to enter the ORFs in DUPLO!!! If there are too many ATG, please remove whatever you dont need to be plotted as ORF.



10 Save file and open “terminal” (Utilities => terminal)

11 Enter “python Desktop/plot_divergence.py” <= the location of your script (for me it is on Desktop) and click enter.

alternatively you can also make the file run by right-clicking on the python code (plot_divergence.py) and running it with python launcher



12 Depending on the number of accessions and the sequence length this can take up to couple of hours. Leave your computer on for some time and the plot will present itself whatever directory you chose in step 8 to use.

13 You will end up with three plots. The upper graph represent the % of missing data, the middle plot represent the insertion/deletions and the lower plot represent the % identity of the sequence among the chosen accessions. All of this is of course relative only to Col-0. Below the plot you will find the position of individual ORFs. The plot is saved as .svg file which can be opened in Adobe Illustrator and modified for colour and the direction of the ORF can be added too.