# Making Daniele-style fastPHASE haplotype plots

### Things you need to start:

1. The [fastPHASE program](https://www.sciencedirect.com/science/article/pii/S000292970763701X?via%3Dihub). It’s easiest if this is in your $PATH, of course. To get a license and download, go [here.](http://scheet.org/software.html) Change the name of the binary to “fastphase” (or change the relevant line in the runFastphase function to reflect the name of your fastphase binary).

2. R and a basic knowledge of how to manage your own directories in R. Work each fastPHASE analysis that you want to do in its own unique directory. The analysis will generate a lot of intermediate files! If you run fastPHASE twice in the same directory, it WILL overwrite the previous analysis (at least as it’s set up now - this can be changed. See me...).

3. Function file from Daniele: fastPHASE.functions.R, which includes the R functions csvToFastphase, runFastphase, parseFastphase, plotFastphase. You will need to source these functions: see “running.fastPHASE.R” for an example.

4. The R file running.fastPHASE.R which shows an example of how to use the functions in the function file.

5. SNP data (as a .csv)

A. Use unimputed data ONLY. Call method 74 (unimputed 250K SNP data, 1000+ worldwide accessions) or resequencing data (unimputed SNPs, 200+ Swedish accessions, or Fernando’s unimputed 1001g SNP calls) are good choices.

B. From this data, pull out a subset of SNPs that corresponds to the desired window around your gene (or area) of interest. I recommend starting with a window size of at least +/- 50kb (if not 100kb). This window may seem large, but it will give you a good feeling for what’s going on in the region. It is much easier to start with a large window and then zoom in if needed!

C. Make sure accession names are ecotype\_id. The raw call\_method52 files are indexed using array\_id rather than ecotype\_id. If using call\_method 52 or 74, make sure to remove any accessions that are NOT also included in call\_method75; these were filtered out by Yu for quality reasons and shouldn’t be included in analysis.

D. Format should be as follows: One SNP per row. First two columns are the chromosome and position of the SNPs and the rest of the columns should be the base calls in each of the 1000+ accessions (if using worldwide data). ***Missing data should be called NA***.

E. One also might want to filter to remove SNPs with low minor allele frequencies??

E. Ask if you have questions or need help. This can be fiddly!

6. Phenotype data (as a .csv). First column is ecotype\_id, second is phenotypic value. File should have column names. The name of the first column is unimportant, but the name of the second column should be the phenotype name - it will appear on the plot. As is generally the case with R, don’t use any spaces in these names!

7. Latitude and longitude data (call\_method\_75\_info.csv).

### Here’s how it works!

Before you start running these scripts, be sure to double check that you have put SNPs into the correct format (see #4 above). This is very important and I haven’t written any “checks” for this into the code. It’s your responsibility to format the SNP file correctly!

1. Open up running.fastPHASE.R to see the general flow of things.

2. Run fastPHASE. This will take a bit of time (hours) and will generate a lot of output files! All the output files will be written to the current working directory unless you go in and edit the functions. Any old fastPHASE output files in your folder could be overwritten. (functions used are *runFastphase* and *csvToFastphase*)

2. Parse the fastPHASE output files. *(*function is *parseFastphase)*

3. Make the pretty plot. (function is *plotFastphase*)

4. Enjoy staring at the plot, hopefully while drinking the beverage of your choice.

### Resources:

FastPHASE paper:

<https://www.sciencedirect.com/science/article/pii/S000292970763701X?via%3Dihub>

Download link for fastPHASE:

<http://scheet.org/software.html>

FLC paper where this method was used and published:

doi: [10.1101/gad.245993.114](https://dx.doi.org/10.1101%2Fgad.245993.114)

<http://genesdev.cshlp.org/content/28/15/1635>