FACS

1. What organisms are covered?

Currently the Simple tool can map genes for the following species: Arabidopsis\_thaliana, Oryza\_sativa\_Japonica, Zea\_mays, Solanum\_lycopersicum, Drosophila\_melanogaster. However, it is very easy to enable Simple to map mutants of other species; see question 2.

1. What if my organism is not included?

You can simply add it; open the file data\_base.txt with Excel. You want to add another new row with information about your organism in a way simple can read it (each row in data\_base.txt stores information for a single organism). In the first column write your organism name (no special characters besides \_ (underscore), in the second column paste the link to the gz (compressed) genome fasta file. Many are hosted on Ensembl. You can look at the other species rows to figure out how to find this link. In the third column you will need the compressed, gz link to the known variants gz file. You can find it the same way you found the link for the compressed fasta file. If your species doesn’t have this file, copy the following string to your knownsnps column: <https://raw.githubusercontent.com/wacguy/test/master/empty.vcf.gz>. The forth column information should be taken from the snpEff.database.xlsx file inside the folder SnpEff which is located in the folder programs. Find your organism name (column #2 in the snpEff.database.xlsx file) and copy the respective text of column #1 (Genome); if your genome has more than one entry we recommend to use the highest number (genome annotation). Paste it to the forth column of your newly added species in the data\_base.txt file. Save the file and close it.

1. How much time will it take the program to run?

We experienced time ranging from a few hours to ~48 hours depending on your machine, the size of your fastq files (which actually reflects the size of your genome).

1. How do I know that the program finished running?

The terminal will show a beautiful “Simple is done” red-yellow-blue text.

1. What if I have no genes in my EMS.cands4 list?

Don’t Panic! We have mapped several genes even though the EMS.cands4.txt file was empty. There can be three possible reasons for that:

* It is possible that as a result of sampling error, i.e., collecting a few wt individuals into your mutant bulk your causal mutation was not recognized as homozygous in the mutant bulk. Open the file EMS.plot4.txt. Now you are looking for an entry (row/SNP) that has a low number in the mut.ref column #10 a much higher number in mut.alt column #11 and wt.ref/wt.alt ratio should be ~2/1. Additionally, you might prefer SNPs that have a significant impact on the coding region (column #9) and ones with C>T or G>A changes (column #8) which are the majority of changes induced by EMS. Obviously you can use some advanced functionality of your spreadsheet editor like sorting.
* Simple uses reference FASTA and VCF files downloaded from the internet but we noticed that some releases have mistakes in them. For example, in the FASTA file of Arabidopsis thaliana, release 31, the chromosomes are organized incorrectly; the VCF file of release 32 of the same species is missing most of the known SNPs in chromosome 1. These errors lead to incorrect execution of commands in the pipeline. We have already checked that the listed species links to the FASTA are correct but if you are working with a different species and had a problem with the pipeline, you might want to check these files in the refs folder (you will most likely have to open them with the Terminal since they are too large for applications such as Word or Excel).
* Simple will only report indels up to a certain size; large genomic deletions that are rare but possible in EMS like screens are not likely to be detected.
* There is a very low chance that your mutation originated from a non-reference nucleotide. These SNPs are reported in the EMS.cands\_alt4.txt (since this is a very unlikely event, this file will usually be empty).

1. How much memory do I need?

Quite a lot, the output folder will fill up with ~100 gb, depending on the size of the input files (determined by the number of reads) and species (large genome size requires more reads) but essentially, only the three Rplot.pdf, EMS.cands4.txt, EMS.plot4.txt and EMS.cands\_alt4.txt are necessary; the rest of the files are not necessary for final analysis. You might want to keep them for further analysis e.g., visualizing the data with tools like the IGV browser.

1. I have data from two different runs, can I use SIMPLE?

Most likely, yes. If the two bulks are from two different runs, it is possible that the fastq files have different quality scores encoding; the pipeline should be able to deal with this issue. However, if a single bulk was generated by concatenating fasq files with different encodings, the program might spit an error message. In such case you should try and fix this. Here are some tools that can help you deal with this kind of problem: <https://en.wikipedia.org/wiki/FASTQ_format#cite_note-7>