This text describes how to use the different scripts to produce the figures and statistics present in the article "***Tempo and mode of genome evolution in a 50,000-generation experiment***" by Olivier Tenaillon, Jeffrey E. Barrick, Noah Ribeck, Daniel E. Deatherage, Jeffrey L. Blanchard, Aurko Dasgupta, Gabriel C. Wu, Sébastien Wielgoss, Stéphane Cruveiller, Claudine Médigue, Dominique Schneider & Richard E. Lenski

To compute the figures and tables several scripts have to be executed.

1) First, **GenomeCompositionComputer.pl** can be used with command

perl GenomeCompositionComputer.pl REL606.gff3 REL606.L20.G15.P0.M35.mask.gd

The first argument**, REL606.gff3** is the gff3 file for the genome under study, the second, **REL606.L20.G15.P0.M35.mask.gd,** a file giving the regions of the genome excluded due to the sequencing technology used (limitations due to read length and repeated regions). This program produces an output file: **GenomeComposition.txt**, which gives the composition of the genome in terms of types (synonymous, non-synonymous and intergenic) of mutations for each possible category of mutations (AT->CG, AT->GC, AT->TA, CG->GC, CG->TA, GC->TA). It will be used to compute expectation of non-synonymous mutations based on neutrals.

2) The second program, **ComputeMutationThroughTimeDryad.perl**, uses a file with the position of the mutations to compute their type (synonymous, non-synonymous, intergenic, in non-coding genes), to filter them according to their appearance in a mutator background or not and finally to produce a file that will be used to compute the phylogeny. The program can be launched with the following command:

perl ComputeMutationThroughTimeDryad.pl Cmdfile.txt

Cmdfile.txt is a command file that provides the following information:

genome: REL606.gff3 # the gff3 of the genome
genomedata: oli.LTEE.final_masked.no_IS_adjacent.tab #the tabulated input data
genomeinfo: LTEE_Mutator_info.txt #a file with information relative to the sampling time of the clones and their mutator status
mutator: 0 #0 for all mutations 1 for the ones that appeared only in mutator and 2 for the ones that occurred in non mutators
Outputfiles_Suffix: none # a suffix to add at the end of the files, "none" can be used to avoid adding a suffix.

LTEE_Mutator_info.txt is of the form:

Population	*Strain_ID*	*Generation*	*CloneName*	*PointMutatorStatus*
Ara+1	*REL768A*	*500*	*Ara+1_500gen_768A*	*0*
Ara+1	*REL768B*	*500*	*Ara+1_500gen_768B*	*0*
Ara+1	*REL958A*	*1000*	*Ara+1_1000gen_958A*	*0*

The *CloneName* has to match the names of **oli.LTEE.final_masked.no_IS_adjacent.tab** file. The *PointMutatorStatus* is 0 for non mutator, 1 if the clone is a point mutator clone.

This program is run on the LTEE data and MAE (for the figures) and for the LTEE data with the mutator option set to 1 or 2 to compute the G scores on the mutator only mutations and on non-mutator mutations.

3) The R script **ScriptFiguresDryad.r** creates all the figures. This script can be run with the following command line:

Rscript ScriptFiguresDryad.r inputfolder ouputfolder

Inputfolder is the path to the folder in which the following files should be present:

count.LTEE.final_masked.csv
GenomeComposition.txt
MAE_fitness.csv
MutationTypesThroughTime.txt
MutationTypesThroughTimeMAE.txt
MutRArray.txt
Spectrum_counts.csv

Outputfolder is the folder in which the pdf of the figures and postscripts of the extended figures will be produced. The code of each figure can also be run independently on the R shell.

4) **ConvergenceGstatDryad.pl** can be then used to produce the G score tables, and to compute the Zscore with simulations the following code has to be run:

perl ConvergenceGstatDryad.pl REL606.gff3 REL606.L20.G15.P0.M35.mask.gd EventByStrainsSorted.txt 1000 outputfolder

The first argument is the gff3 file of the reference genome, the second, the masked file, the third, the output file of **ComputeMutationThroughTimeDryad.pl** of the form **EventByStrainsSorted.txt** that has only the subset of mutations of interest (all non mutator clones or only mutator clones). The fourth is the number of replicates for the simulations and the fifth the outputfolder where the output files will be created.
The output file of interest is **GenesGstat.txt** which is the base of supplementary tables 2 and 3. **GenesGstat_SimulsNS.txt** stores in a single line per simulation the G score for each

gene. The first line corresponds to the real dataset. The simple following code can be used in R to compute the Zscore :

```
Tab<-read.table("GenesGstat_SimulsNS.txt")
h<-rowSums(Tab)
nbrep<-length(h)
m<-mean(h[2:nbrep])
s<-sd(h[2:nbrep])
Zscore<-(h[1]-m)/s
```

5) Overall to produce all the tables and figures, put the following files in the working folder as well as the 4 previous source codes

count.LTEE.final_masked.csv

CmdfileLTEE.txt

CmdfileLTEE_nonMutator.txt

CmdfileLTEE_Mutator.txt

CmdfileMAE.txt

MAE_fitness.csv

MAE_Mutator_info.txt

LTEE_Mutator_info.txt

oli.LTEE.final_masked.no_IS_adjacent.tab

oli.MAE.final_masked.no_IS_adjacent.tab

REL606.gff3

REL606.L20.G15.P0.M35.mask.gd

Spectrum_counts.csv

Then, from that directory, run:

```
perl GenomeCompositionComputer.pl REL606.gff3 REL606.L20.G15.P0.M35.mask.gd
perl ComputeMutationThroughTimeDryad.pl CmdfileLTEE.txt
perl ComputeMutationThroughTimeDryad.pl CmdfileLTEE_nonMutator.txt
perl ComputeMutationThroughTimeDryad.pl CmdfileLTEE_Mutator.txt
perl ComputeMutationThroughTimeDryad.pl CmdfileMAE.txt
mkdir Figurefolder
Rscript ScriptFiguresDryad.r ./ Figurefolder/
mkdir GscoreNonMut
mkdir GscoreMutator
perl ConvergenceGstatDryad.pl REL606.gff3 REL606.L20.G15.P0.M35.mask.gd
EventByStrainsSortedNoMut.txt 1000 GscoreNonMut
perl ConvergenceGstatDryad.pl REL606.gff3 REL606.L20.G15.P0.M35.mask.gd
EventByStrainsSortedMutator.txt 1000 GscoreMutator
```

While the first programs take a few minutes, the randomisation to compute the Gscore can last several hours.