# Guidance

GUIDANCE is an integrated framework that is able to perform haplotype phasing, genome-wide imputation, association testing and PheWAS analysis of large GWAS datasets. Moreover, this application allows performing all these steps in a single execution, as well as in a modular way with optional user intervention.

The GUIDANCE strength is based on the possibility of increasing the potential of GWAS by means of the analysis of the X chromosome, as well as the autosomes analysis, using one or multiple reference panels, do the assosiation testing using several models (additive, dominant, recessive, heterodominant and genotypic inheritance models) and realizing the cross-phenotype association analysis when more than one disease is available in the cohort under study.

# Table of Contents

1. [Getting started](#getstarted)

\* [Installing on a Singularity image](#singularity)

\* [Installing on bare metal](#bare)

2. [Configuration file description](#config)

\* [Input description](#input)

\* [Configuration file examples](#examples\_conf)

3. [GUIDANCE execution](#execution)

\* [Running on a Singularity image](#singularityexec)

\* [Running on bare metal](#bareexec)

4. [Authors](#authors)

5. [License](#license)

6. [Acknowledgments](#acknowledgments)

## Getting Started <a name="getstarted"></a>

GUIDANCE was developed to mainly run on HPC infrastructures. The instructions given here consider this execution environment. In order to run it in the cloud, specific instructions are given in this [project](https://github.com/ramonamela/guidance\_cloud).

These instructions will get you a copy of the project up and running on your machine. There are two main options to run GUIDANCE: directly on top of your OS and inside a [Singularity image](https://github.com/sylabs/singularity). We strongly recommend to use Singularity as it makes all the process much more easier.

### Installing on a Singularity image <a name="singularity"></a>

<details><summary>Show instructions</summary>

<p>

#### Prerequisites

In order to run GUIDANCE in this mode, you only need to install the guidance [COMPSs’](https://github.com/bsc-wdc/compss/tree/guidance) branch and generate the Singularity image containing all the depedencies. In addition, if you want to use the framework with [SHAPEIT](https://mathgen.stats.ox.ac.uk/genetics\_software/shapeit/shapeit.html#download), you need to obtain the binary v2r727 and copy it in the path “$PROJECT\_PATH/singularity/docker/TOOLS/shapeit.v2.r727.linux.x64”.

#### Installation

It is possible to find the instructions to install COMPSs in their github repository.

Regarding the Singularity image generation, it is enough to run [this script](https://github.com/ramonamela/guidance/blob/master/singularity/build\_container.sh). It will automatically create a local docker registry and upload to it the base docker image necessary to create the Singularity image. Nevertheless, you should modify the file placed in [$PROJECT\_PATH/singularity/singularity/base.def](https://github.com/ramonamela/guidance/blob/master/singularity/singularity/base.def) in order to include the folder in your local file system that should be visible into the image. More precisely, you should only touch this lines:

```

%setup

mkdir -p $SINGULARITY\_ROOTFS/gpfs/home/

mkdir -p $SINGULARITY\_ROOTFS/gpfs/scratch/

mkdir -p $SINGULARITY\_ROOTFS/gpfs/apps/MN4

mkdir -p $SINGULARITY\_ROOTFS/gpfs/projects/

mkdir -p /opt/intel

mkdir -p /scratch

```

In addition, a GUIDANCE binary should be generated from this repository. In order to do so, you should run the following command in the root of the repository:

```

mvn clean install

```

This will generate a binary guidance.jar.

</p>

</details>

### Installing on bare metal <a name="bare"></a>

<details><summary>Show instructions</summary>

<p>

#### Prerequisites

First of all, you need to install the guidance [COMPSs’](https://github.com/bsc-wdc/compss/tree/guidance) branch in your system.

Afterwards, all the dependencies should be installed in the machine in which GUIDANCE will run. More precisely, the following binaries should be installed and accessible from all the worker nodes:

\* Minimac4

\* Minimac3

\* Plink 1.9

\* Eagle 2.4

\* Impute 2.3.2

\* Qctool 1.4

\* Snptest 2.5

\* Tabix

\* Bgzip

\* Shapeit v2 r727

\* Samtools

\* Bcftools

\* R (tested version 3.5.1)

Next, the following environment variables should be set in order to tell GUIDANCE where the binaries are located in each machine:

```

export BCFTOOLSBINARY=/PATH\_TO/BCFTOOLS/1.8/INTEL/bin/bcftools

export SAMTOOLSBINARY=/PATH\_TO/SAMTOOLS/1.5-DNANEXUS/INTEL/IMPI/bin

export PLINKBINARY=/PATH\_TO/plink\_1.9/plink

export EAGLEBINARY=/PATH\_TO/Eagle\_v2.4.1/eagle

export RSCRIPTDIR=/PATH\_TO/R\_SCRIPTS/

export QCTOOLBINARY=/PATH\_TO/qctool\_v1.4-linux-x86\_64/qctool

export SHAPEITBINARY=/PATH\_TO/shapeit.v2.r727.linux.x64

export IMPUTE2BINARY=/PATH\_TO/impute\_v2.3.2\_x86\_64\_static/impute2

export SNPTESTBINARY=/PATH\_TO/snptest\_v2.5

export MINIMAC3BINARY=/PATH\_TO/Minimac3/bin/Minimac3

export MINIMAC4BINARY=/PATH\_TO/Minimac4/release-build/minimac4

export TABIXBINARY=/PATH\_TO/SAMTOOLS/1.5-DNANEXUS/INTEL/IMPI/bin/tabix

export BGZIPBINARY=/PATH\_TO/SAMTOOLS/1.5-DNANEXUS/INTEL/IMPI/bin/bgzip

export RSCRIPTBINDIR=/PATH\_TO/R/3.5.1/INTEL/bin/

export R\_LIBS\_USER=/gpfs/home/bsc05/bsc05997/TOOLS/R\_libs/

```

In this case, R\_LIBS\_USER specify the path into which are installed the R dependencies in case they are not installed in the default path. More precisely, the following packages should be available:

\* data.table

\* plyr

\* dplyr

\* reshape

\* gap

\* sfsmisc

\* BiocManager

\* IRanges

In addition, in the folder “RSCRIPTDIR” there shuould be all the scripts available [here](https://github.com/ramonamela/guidance/tree/master/src/main/R).

Finally, maven should be installed in the system in order to compile the GUIDANCE binary.

#### Installing

Once all the dependencies are in the system, you should checkout the GUIDANCE repository and generate the guidance binary. In order to do so, you should run the following command in the root of the repository:

```

mvn clean install

```

This will generate a binary guidance.jar.

</p>

</details>

## Configuration file <a name="config"></a>

In order to run GUIDANCE the user has to edit a configuration file, where the basic input and output characteristics have to be specified. This file also allows the tuning of multiple parameters related to, among others, covariates, chunk size for genotype imputation, info scores, minor allele frequency and Hardy-Weinberg thresholds. The user can also decide to run several phenotypes and several combinations of phenotypes/covariates in a single run.

A detail description of each of the configurable parameters (configuration\_file\_parameters.pdf) is provided in this [file](https://github.com/ramonamela/guidance/tree/master/webpage/configurationfileparameters.pdf) or in the collapsable list placed at the end of this paragraph.

<details><summary>Show the configuration file parameters</summary>

<p>

• wfDeep: Name that defines the number of stages to be executed. These stages are defined in Figures 1 and 2.

• init\_chromosome: First chromosome to analyse.

• end\_chromosome: Last chromosome to analyse.

• maf\_threshold: Minor allele frequency cut-off used to filter final results.

• impute\_threshold: IMPUTE2 info score cut-off used to filter final results.

• minimac\_threshold: MINIMAC Estimated imputation accuracy (R²) cut-off used to filter final results.

• hwe\_cohort\_threshold: Hardy-Weinberg equilibrium p.value threshold for cohort.

• hwe\_cases\_threshold: Hardy-Weinberg equilibrium p.value threshold for cases.

• hwe\_controls\_threshold: Hardy-Weinberg equilibrium p.value threshold for controls.

• exclude\_cgat\_snps: Logical. Whether or not G>C or A>T SNPs should be excluded. We strongly recommend activating this flag as to avoid strand orientation issues. Most of the genotyping arrays have a very small number of such SNPs, and their exclusion should not result in any noticeable loss of imputation performance.

• imputation\_tool: The name of the imputation tool to impute genotypes. To date, only “impute” to select IMPUTE2 and “minimac” to select MINIMAC4 are accepted.

• test\_types: Names for the different analysis to be carried out by GUIDANCE, separated by commas. The association results for each “test\_type” will be created in a directory with the same name inside the “associations” directory. Below this flag, different “test\_types” have to be listed with the phenotype name and the covariates names to take into account in the association analysis (for instance, to analyse “test\_types = DIA2,CARD” users should add: “DIA2 = DIA2:sex,BMI” and “CARD = CARD:sex,BMI” below, where sex and BMI are covariates).

• chunk\_size\_analysis: Size of the chunks considered to partition the data.

• file\_name\_for\_list\_of\_stages: File into which all the commands launched in the workflow are stored.

• input\_format: (I think that now we only support BED input since we have not tried with the other formats since I am working in the project…).

• mixed\_cohort: Name of the cohort.

• mixed\_bed\_file\_dir: The path to the directory with genotype files.

• mixed\_bed/bim/fam/\_file: Name of the file containing genotypes.

• mixed\_sample\_file\_dir: Path to the directory where the sample file is located.

• mixed\_sample\_file:. Name of the sample file.

• genmap\_file\_dir: Path where genetic map files are located.

• genmap\_file\_chr\_n: Name of the genetic map file for each chromosome in every new line.

• refpanel\_number: Number of reference panels.

• refpanel\_combine: 'NO' if there is only one panel or imputed results from different reference panels should not be integrated; 'YES' when different reference panels are expected to be used in the analysis and also the integration of all the results is required.

• refpanel\_type: Name of the reference panel.

• refpanel\_memory: Required amount of memory demanded by each particular panel. Currently, “HIGH”, “MEDIUM” and “LOW” are supported.

• refpanel\_file\_dir: Path where the reference panel for each chromosome is located.

• refpanel\_hap\_file\_chr\_n: Haplotypes files per chromosome of the reference panel provided in case IMPUTE2 is chosen as imputation tool and for the chrX in case Minimac4 is used.

• refpanel\_leg\_file\_chr\_n: Legend files per chromosome of the reference panel provided in case IMPUTE2 is chosen as imputation tool and for the chrX in case Minimac4 is used.

• refpanel\_vcf\_file\_chr\_n: VCF files per chromosome of the reference panel provided in case Minimac4 is used.

• outputdir: The path of the directory where the results will be written.

</p>

</details>

### Input description <a name="input"></a>

GUIDANCE accepts PLINK format files, e.g. gwas.bed, gwas.bim, gwas.fam), as input. It also accepts covariates, such as principal compontents, gender or any covariate defined in the sample file. As part of the input, GUIDANCE also requires the inclusion of one, or several reference panels for imputation accepting public (1KG phase 1 or 3, HapMap, DCEG, UK10K, GoNL) or private reference panels. Also a genetic map is needed.

#### Sample File

The user needs to also provide an adequate sample file. Please, check SNPTESTv2 webpage for information on how to prepare a suitable sample file.

#### Genetic Map File

All of IMPUTE2 reference panel download packages come with appropriate recombination map files. Check IMPUTE2 webpage for more information. Hence, this genetic map files can be used for phasing with SHAPEIT. A file per chromosome must be given.

On the other hand, when phasing with EAGLE, a single file must be given. Several compatible options are abailable on the [Broad Website](<https://data.broadinstitute.org/alkesgroup/Eagle/downloads/tables/>).

#### Reference panels

It must be noted that, when using IMPUTE2, the reference panels must be given with format .haps and .leg. On the other hand, Minimac4 only accepts M3VCF. [Regular VCF can be converted to M3VCF using Minimac3](<https://genome.sph.umich.edu/wiki/Minimac4>).

In addition, we have encountered some problems when imputing the ChrX with Minimac4. This is why the imputation of this chromosome is always done with IMPUTE2, so even when using Minimac4 both .haps and .legend must be given for ChrX if this chromosome is included in the study.

### Configuration file examples <a name="examples\_conf"></a>

We give some templates of configuration files for a study from Chr21 to ChrX, using HRC, 1kphase3, uk10k and gonl as reference panels and [Eagle-IMPUTE2](<https://github.com/ramonamela/guidance/tree/master/utils/conf_examples>/[config\_eagle\_impute.file](https://github.com/ramonamela/guidance/blob/master/utils/conf_examples/config_eagle_impute.file)), [SHAPEIT-IMPUTE2](<https://github.com/ramonamela/guidance/tree/master/utils/conf_examples>/config\_eagle\_minimac.file), [Eagle-Minimac4](<https://github.com/ramonamela/guidance/tree/master/utils/conf_examples>/[config\_shapeit\_impute.file](https://github.com/ramonamela/guidance/blob/master/utils/conf_examples/config_shapeit_impute.file)) and [SHAPEIT-Minimac4](<https://github.com/ramonamela/guidance/tree/master/utils/conf_examples>/[config\_shapeit\_minimac.file](https://github.com/ramonamela/guidance/blob/master/utils/conf_examples/config_shapeit_minimac.file)) as phasing and imputation tools.

## GUIDANCE execution <a name="execution"></a>

Explain how to run GUIDANCE

### Running on a Singularity image <a name="singularityexec"></a>

<details><summary>Show instructions</summary>

<p>

</p>

</details>

### Running on bare metal <a name="bareexec"></a>

<details><summary>Show instructions</summary>

<p>

</p>

</details>

## Authors <a name="authors"></a>

\* \*\*Marta Guindo Martinez\*\* - \*R scripts, Java code and Scientific Methodology\* - [ORCID](https://orcid.org/0000-0002-8099-9170)

\* \*\*Ramon Amela Milian\*\* - \*R scripts and Java code\* - [ORCID](https://orcid.org/0000-0001-5943-5824) [github](https://github.com/ramonamela)

\* \*\*Cristian Ramón-Cortés Vilarrodona\*\* - \*Java code\* - [ORCID](https://orcid.org/0000-0003-4170-818X) [github](https://github.com/cristianrcv)

\* \*\*Montserrat Puiggròs\*\* - \*R scripts and Java code\* - [ORCID](https://orcid.org/0000-0001-5034-7924)

\* \*\*Josep Maria Mercader\*\* - \*Scientific Methodology\* - [ORCID](https://orcid.org/0000-0001-8494-3660)

\* \*\*David Torrents\*\* - \*Scientific Methodology\* - [ORCID](https://orcid.org/0000-0002-6086-9037)

<!-- See also the list of [contributors](https://github.com/your/project/contributors) who participated in this project. -->

## License <a name="license"></a>

This project is licensed under the MIT License - see the [LICENSE.md](LICENSE.md) file for details

## Acknowledgments <a name="acknowledgments"></a>

\* Hat tip to anyone whose code was used

\* Inspiration

\* etc