Configuration File Parameters for GUIDANCE v0.1.1

Last update: **9** August 2019

- **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 (R^2) 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.

1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using '#' starting a new line.

- **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.

For a complete example of a configuration file, see Figures 3 and 4.

1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using ‘#’ starting a new line.

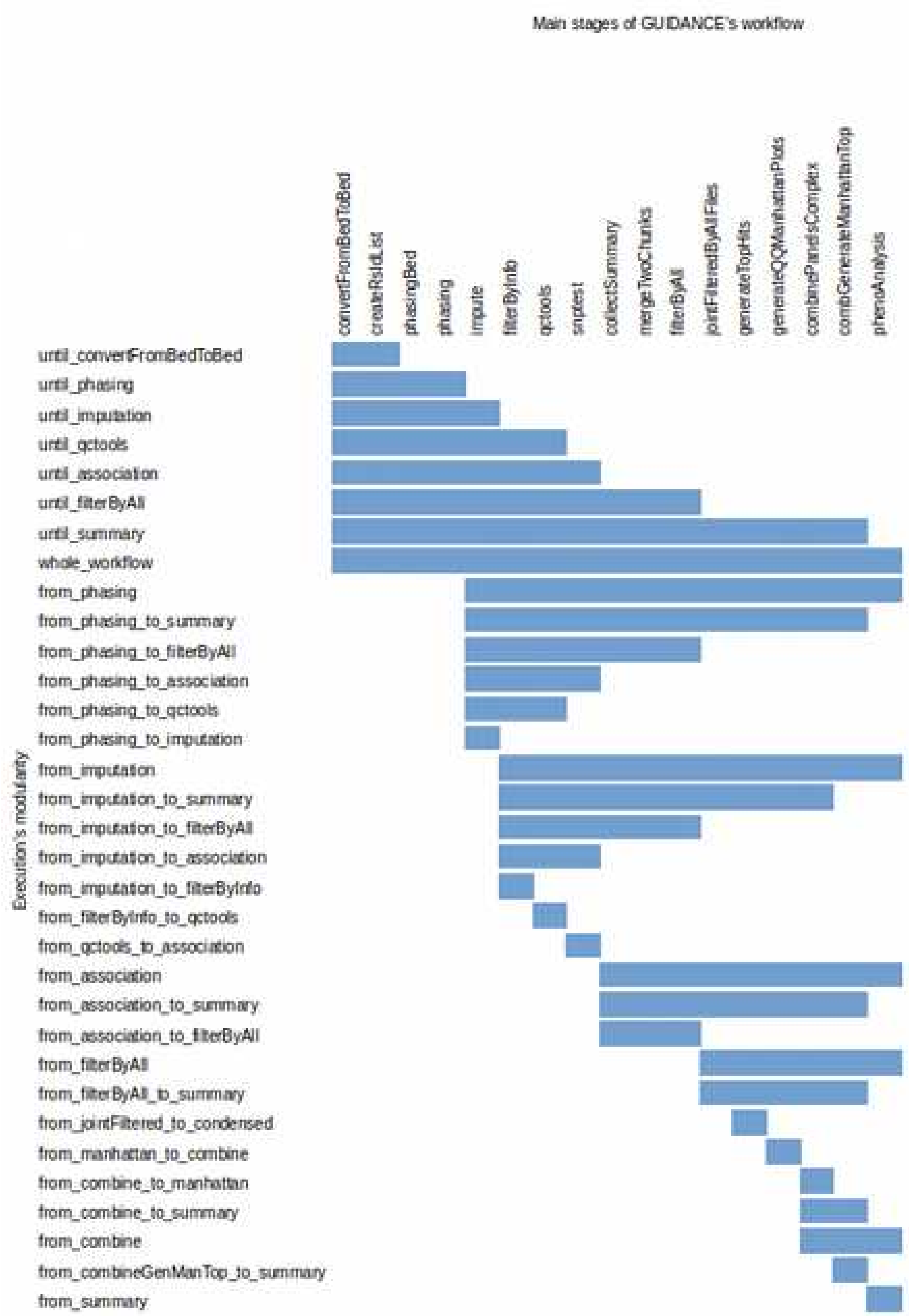


Figure 1. Modularity of the GUIDANCE workflow with IMPUTE2 as imputation tool. The user can choose between using running the whole workflow, or just a subset of stages. The bar represents the number of stages that will be run by each category of modularity.

1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using '#' starting a new line.

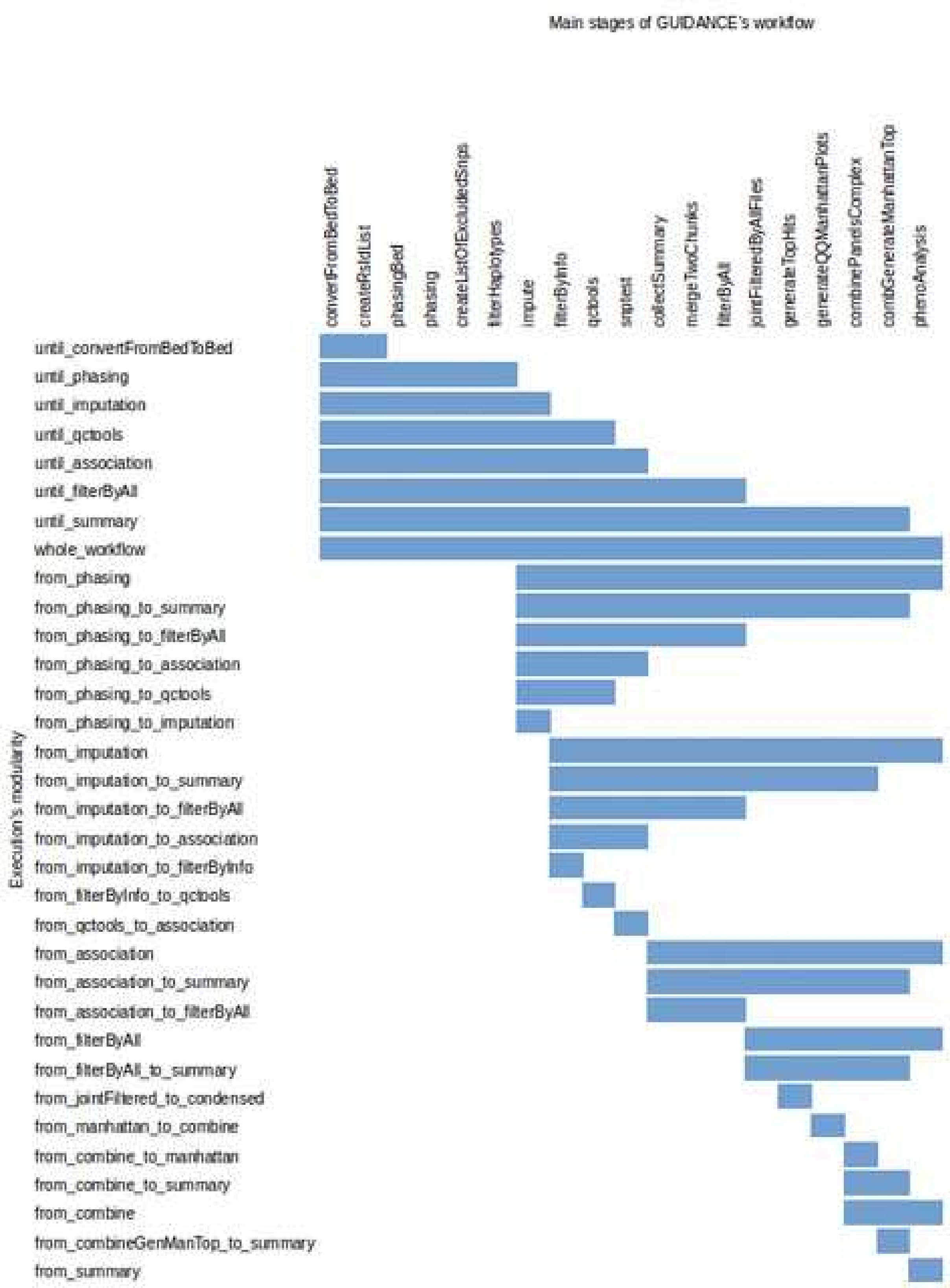
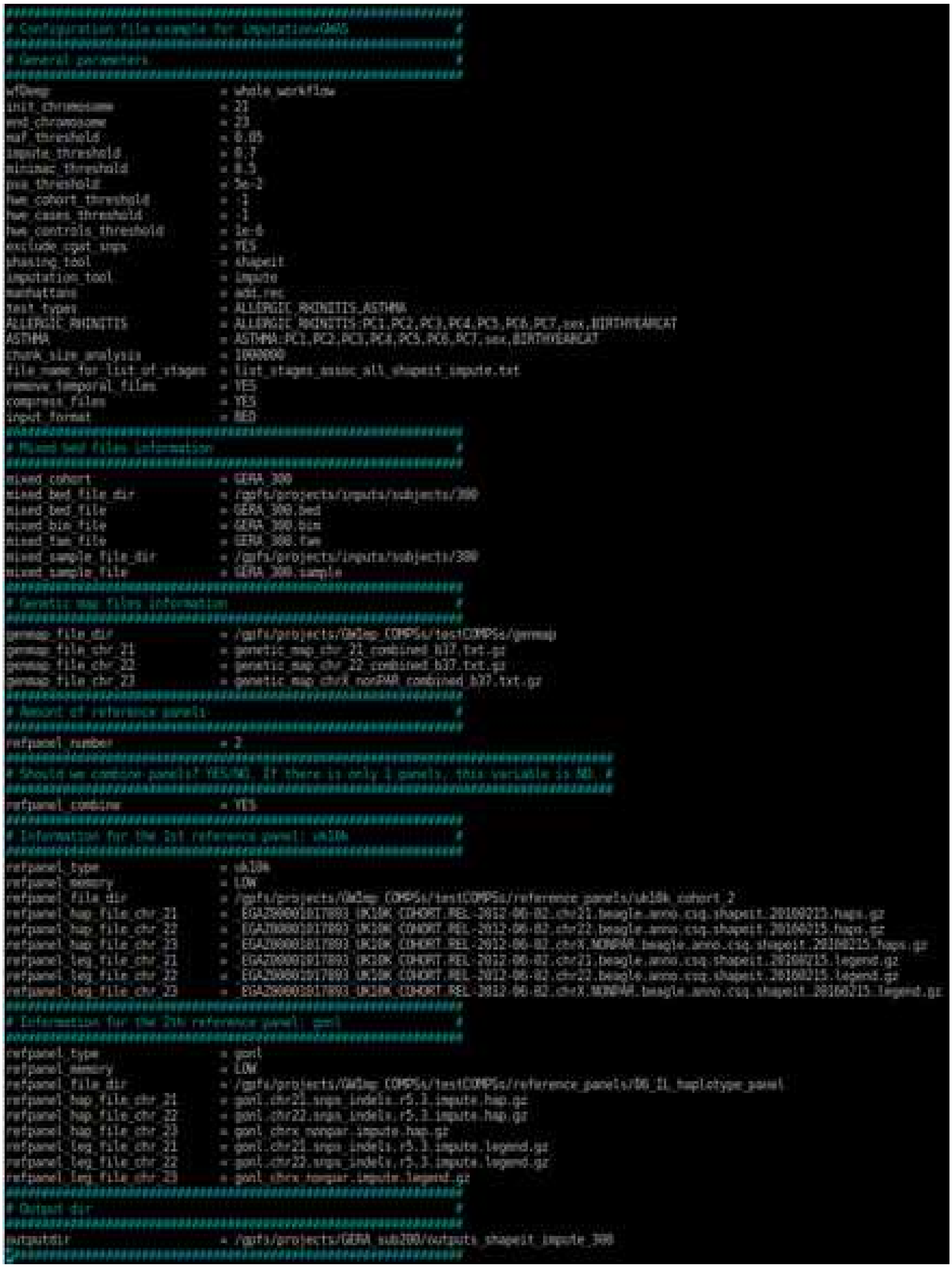


Figure 2. Modularity of the GUIDANCE workflow with Minimac4 as imputation tool. The user can choose between using running the whole workflow, or just a subset of stages. The bar represents the number of stages that will be run by each category of modularity.

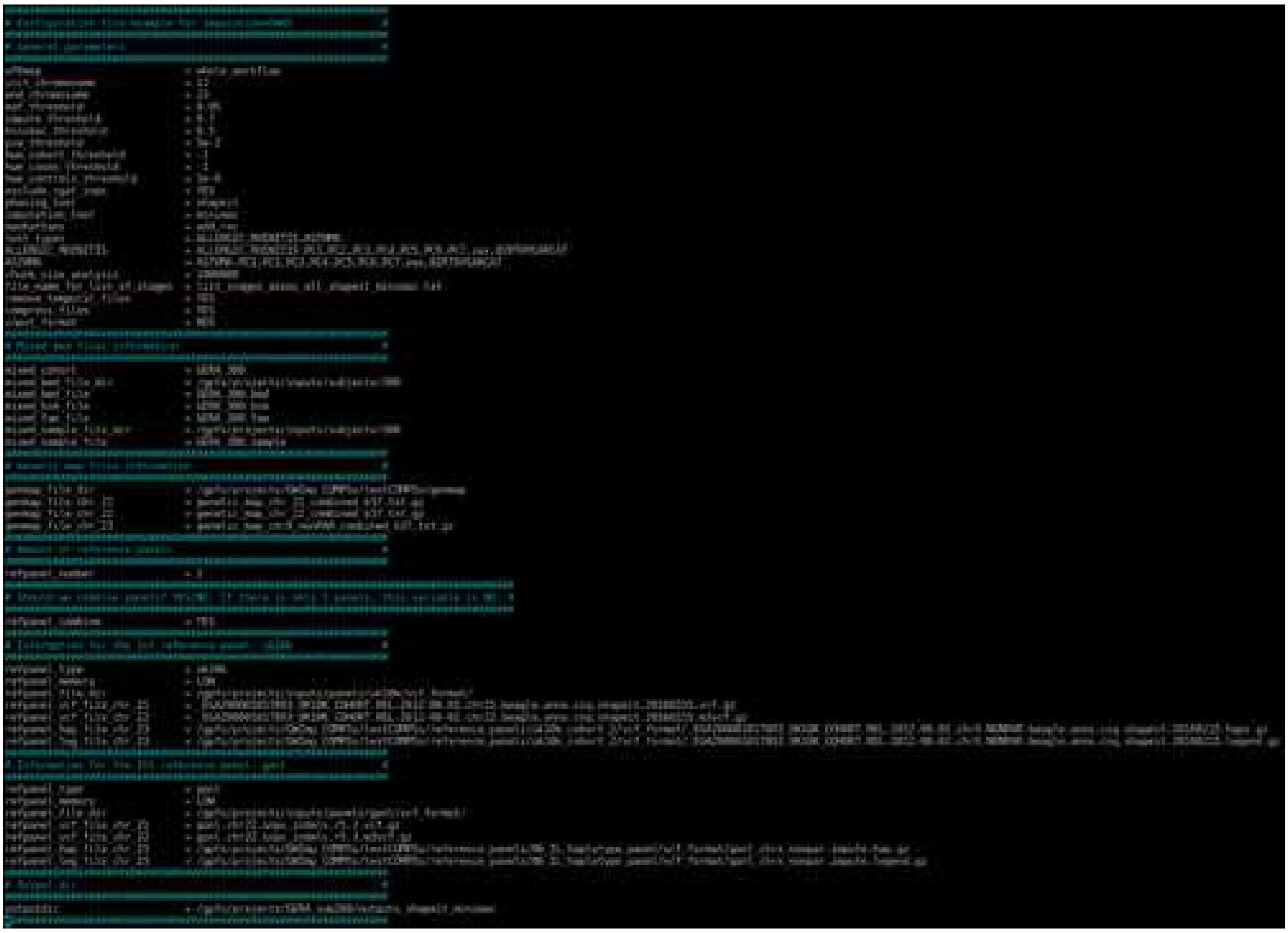
1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using '#' starting a new line.



```
##############################################################
# Configuration file example for imputationGWAS              #
##############################################################
# General parameters                                         #
##############################################################
wfDeep                        = whole_workflow
init_chromosome               = 21
end_chromosome                = 23
maf_threshold                 = 0.05
impute_threshold              = 0.7
minimac_threshold             = 0.5
pva_threshold                 = 5e-2
hwe_cohort_threshold          = -1
hwe_cases_threshold           = -1
hwe_controls_threshold        = 1e-6
exclude_cgat_snps             = YES
phasing_tool                  = shapeit
imputation_tool               = impute
manhattans                    = add,rec
test_types                    = ALLERGIC_RHINITIS,ASTHMA
ALLERGIC_RHINITIS             = ALLERGIC_RHINITIS:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
ASTHMA                        = ASTHMA:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
chunk_size_analysis           = 1000000
file_name_for_list_of_stages  = list_stages_assoc_all_shapeit_impute.txt
remove_temporal_files         = YES
compress_files                = YES
input_format                  = BED
##############################################################
# Mixed bed files information                                #
##############################################################
mixed_cohort                  = GERA_300
mixed_bed_file_dir            = /gpfs/projects/inputs/subjects/300
mixed_bed_file                = GERA_300.bed
mixed_bim_file                = GERA_300.bim
mixed_fam_file                = GERA_300.fam
mixed_sample_file_dir         = /gpfs/projects/inputs/subjects/300
mixed_sample_file             = GERA_300.sample
##############################################################
# Genetic map files information                              #
##############################################################
genmap_file_dir               = /gpfs/projects/GWImp_COMPSs/testCOMPSs/genmap
genmap_file_chr_21            = genetic_map_chr_21_combined_b37.txt.gz
genmap_file_chr_22            = genetic_map_chr_22_combined_b37.txt.gz
genmap_file_chr_23            = genetic_map_chrX_nonPAR_combined_b37.txt.gz
##############################################################
# Amount of reference panels                                 #
##############################################################
refpanel_number               = 2
###################################################################################
# Should we combine panels? YES/NO. If there is only 1 panels, this variable is NO. #
###################################################################################
refpanel_combine              = YES
##############################################################
# Information for the 1st reference panel: uk10k             #
##############################################################
refpanel_type                 = uk10k
refpanel_memory               = LOW
refpanel_file_dir             = /gpfs/projects/GWImp_COMPSs/testCOMPSs/reference_panels/uk10k_cohort_2
refpanel_hap_file_chr_21      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr21.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_hap_file_chr_22      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr22.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_hap_file_chr_23      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chrX.NONPAR.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_leg_file_chr_21      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr21.beagle.anno.csq.shapeit.20160215.legend.gz
refpanel_leg_file_chr_22      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr22.beagle.anno.csq.shapeit.20160215.legend.gz
refpanel_leg_file_chr_23      = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chrX.NONPAR.beagle.anno.csq.shapeit.20160215.legend.gz
##############################################################
# Information for the 2th reference panel: gonl              #
##############################################################
refpanel_type                 = gonl
refpanel_memory               = LOW
refpanel_file_dir             = /gpfs/projects/GWImp_COMPSs/testCOMPSs/reference_panels/06_1L_haplotype_panel
refpanel_hap_file_chr_21      = gonl.chr21.snps_indels.r5.3.impute.hap.gz
refpanel_hap_file_chr_22      = gonl.chr22.snps_indels.r5.3.impute.hap.gz
refpanel_hap_file_chr_23      = gonl_chrx_nonpar.impute.hap.gz
refpanel_leg_file_chr_21      = gonl.chr21.snps_indels.r5.3.impute.legend.gz
refpanel_leg_file_chr_22      = gonl.chr22.snps_indels.r5.3.impute.legend.gz
refpanel_leg_file_chr_23      = gonl_chrx_nonpar.impute.legend.gz
##############################################################
# Output dir                                                 #
##############################################################
outputdir                     = /gpfs/projects/GERA_sub200/outputs_shapeit_impute_300
##############################################################
```

Figure 3. Configuration file example for a GUIDANCE execution with IMPUTE2 as imputation tool.

1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using ‘#’ starting a new line.



```
#################################################
# Configuration file example for imputation=[illegible]          #
#################################################
# General parameters                            #
#################################################
wfDeep                       = whole_workflow
init_chromosome              = 22
end_chromosome               = 23
maf_threshold                = 0.05
impute_threshold             = 0.7
minimac_threshold            = 0.3
pva_threshold                = 5e-2
hwe_cohort_threshold         = -1
hwe_cases_threshold          = -1
hwe_controls_threshold       = 1e-6
exclude_cgat_snps            = YES
phasing_tool                 = shapeit
imputation_tool              = minimac
manhattans                   = add,rec
test_types                   = [illegible]
[illegible]                  = [illegible]
[illegible]                  = [illegible]
chunk_size_analysis          = 1000000
file_name_for_list_of_stages = list_stages_assoc_all_shapeit_minimac.txt
remove_temporal_files        = YES
compress_files               = YES
input_format                 = BED
#################################################
# Mixed bed files information                   #
#################################################
mixed_cohort                 = GERA_300
mixed_bed_file_dir           = /gpfs/projects/[illegible]/inputs/subjects/300
mixed_bed_file               = GERA_300.bed
mixed_bim_file               = GERA_300.bim
mixed_fam_file               = GERA_300.fam
mixed_sample_file_dir        = /gpfs/projects/[illegible]/inputs/subjects/300
mixed_sample_file            = GERA_300.sample
#################################################
# Genetic map files information                 #
#################################################
genmap_file_dir              = /gpfs/projects/[illegible]/genmap
genmap_file_chr_21           = genetic_map_chr_21_combined_b37.txt.gz
genmap_file_chr_22           = genetic_map_chr_22_combined_b37.txt.gz
genmap_file_chr_23           = genetic_map_chrX_nonPAR_combined_b37.txt.gz
#################################################
# Amount of reference panels                    #
#################################################
refpanel_number              = 2
################################################################
# Should we combine panels? YES/NO. If there is only 1 panel, this variable is NO. #
################################################################
refpanel_combine             = YES
#################################################
# Information for the 1st reference panel: uk10k #
#################################################
refpanel_type                = uk10k
refpanel_memory              = LOW
refpanel_file_dir            = /gpfs/projects/[illegible]/inputs/panels/uk10k/vcf_format/
refpanel_vcf_file_chr_21     = [illegible]
refpanel_vcf_file_chr_22     = [illegible]
refpanel_hap_file_chr_23     = [illegible]
refpanel_leg_file_chr_23     = [illegible]
#################################################
# Information for the 2nd reference panel: gonl  #
#################################################
refpanel_type                = gonl
refpanel_memory              = LOW
refpanel_file_dir            = /gpfs/projects/[illegible]/inputs/panels/gonl/vcf_format/
refpanel_vcf_file_chr_21     = gonl.chr21.snps_indels.r5.3.vcf.gz
refpanel_vcf_file_chr_22     = gonl.chr22.snps_indels.r5.3.vcf.gz
refpanel_hap_file_chr_23     = [illegible]
refpanel_leg_file_chr_23     = [illegible]
#################################################
# Output dir                                    #
#################################################
outputdir                    = /gpfs/projects/[illegible]/outputs_shapeit_minimac
#################################################
```

Figure 4. Configuration file example for a GUIDANCE execution with Minimac4 as imputation tool.

1. NOTES: The configuration file does not yet accept neither blank lines nor tabs.
2. It is possible to include comments on the configuration file by using '#' starting a new line.