CKDGen Round 5 Script

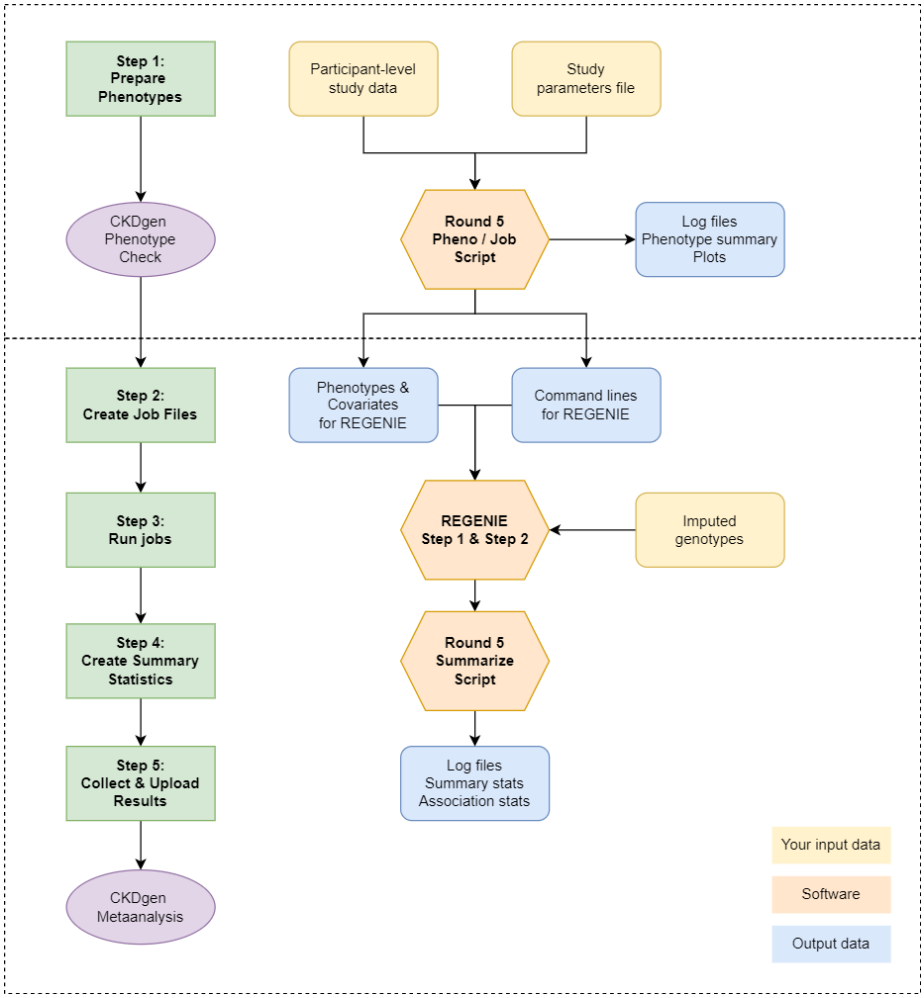

Figure 1: Script Workflow

Usage

1. **Download CKDGen R5 script**

   ```
   cd PATH/FOR/ANALYSIS
   git clone https://github.com/genepi-freiburg/ckdgen-r5.git
   cd ckdgen-r5
   ```

2. **Generate input.txt** for your study

Tab-delimited text file, including a header; rows: participants; columns: variables; please find details in the analysis plan;
R example:

```
write.table(file = "input.txt", x = df, col.names = TRUE, row.names = FALSE, sep = "\t
```

3. **Configure parameters.txt**

   The provided file contains documentation and example values and can be edited. Rename the parameters file to include your study name.

   ```
   cp parameters.txt parameters-<your study name>.txt
   nano parameters-<your study name>.txt
   ```

4. **Run phenotype preparation script**

   ```
   bash 01-ckdgen-r5-phenotypes.sh parameters-<your study name>.txt
   ```

   Check output and log files (folders "return_pheno" and "output_pheno") for any errors.

5. **Send** everything in the "return_pheno" folder to the CKDGen R5 team

   Please send via email to:

   - ckdgenconsortium@gmail.com

   Please **wait for us** to check the files before proceeding with the association analysis. We would like to avoid unnecessary work for you.
   (Now would be a good time to prepare your genotypes - see analysis plan.)

6. **Prepare the REGENIE jobs**

   Please also have a look at the following files to use the provided PLINK/REGENIE scripts/pipeline. Please adjust paths as necessary for your environment.

   - plink_qc.sh
   - vcf_to_bgen.sh
   - make-regenie-step1-job-scripts.sh
   - make-regenie-step2-job-scripts.sh

   You will need to adjust at least the following scripts: "make-regenie-step1-job-scripts.sh" and "make-regenie-step2-job-scripts.sh" to suit your environment (paths, job scheduler, etc.). The scripts contain comments which guide you.

   Before you run REGENIE, you might need to use PLINK to perform a QC on the chip genotypes of your study. An example is provided in the "plink_qc.sh" script.

   To generate the REGENIE step 1 and step 2 job files, please run:

   ```
   bash 02-ckdgen-r5-make-regenie-jobs.sh parameters-<your study name>.txt
   ```

In case you need to make any adjustments, you can re-run this step to generate the job files again.

7. **Run REGENIE** to produce association statistics

   Phenotype/covariable input and REGENIE command line scripts are provided in the "output_pheno" folder. Output will go to "logs", "output_regenie_step1" and "output_regenie_step2".

   Job submission files are generated by the phenotype script and stored under "jobs". Please investigate these scripts and see if everything looks right. We suggest to first run a pilot analysis before submitting all jobs.

   Make sure you complete all "step 1" jobs prior to "step 2" jobs. We provide an example file to submit all jobs including the dependency of step 2 on step 1 with the `03-submit-all-jobs.sh` script. This script runs with Slurm, but you may be able to adapt it to be used with other scheduling systems as well.

   ```
   bash 03-submit-all-jobs.sh
   ```

   Furthermore, we provide a script to check the result log files of step 1 for errors ("check_step1_logs.sh"). We suggest you use this before proceeding to step 2. (Both "check_step1_logs.sh" and "check_step2_logs.sh" will also be run as part of the next collect/summarize step.)

   If you need to regenerate the job files (e.g., if you change parameters or paths in the "make-regenie-step*-job-scripts.sh" scripts), you can re-run the `02-ckdgen-r5-make-regenie-jobs.sh` script, or just invoke the `output_pheno/<study>.regenie.jobs.sh` file.

8. **Summarize** your results

   This step will summarize your phenotypes, taking into account actual missingness after merging genotypes and phenotypes. During this step, we also invoke the "check_step1_logs.sh" and "check_step2_logs.sh" scripts to check the REGENIE runs finished successfully. Please watch out for "ERROR" messages.

   ```
   bash 04-postprocess-results.sh parameters-<your study name>.txt
   ```

9. **Upload** GWAS summary statistics and log files to the CKDGen R5 site

   Automatically collect all necessary files for the upload (log files from "output_regenie_step1" and everything in the following folders "return_pheno", "output_regenie_step2", and parameter files/logs):

   ```
   bash 05-collect-files-for-upload.sh <YOURSTUDYNAME>
   ```

   Go to https://ckdgen.eurac.edu/upload/ and upload `ckdgen-r5-upload-<YOURSTUDYNAME>-<date>.tgz`

   User name: `ckdgenR5`
   Password: `ExcitingScience!`

Because of the large file size, you can also upload your results using SFTP:

User name: `sftp01`
Password: `eeb5iesheeBee6raesua`
Host name: `ckdgen.eurac.edu`

In any case, please send us a short mail to ckdgenconsortium@gmail.com

10. **FAQ and bugs**

- In previous versions of the phenodata preprocessing a warning about low cases or low controls swapped cases / controls (e.g. WARNING: Less than 500 cases for gout). Computations were not affected.
- A previous version of the vcf_to_bgen.sh helper script submitted all plink jobs in parallel, which can lead to a memory overload in case of insufficient memory.
- Some studies observed: Step 1 error in stratified analyses for X chr only, there were a few SNPs that had low variance in women (i.e. invariable genotype) and stopped regenie from running. These variants are returned in a file. They needed to be excluded first (did this manually), then rerun step 1.
- Some studies observed: Regenie step 1 stopped running because >10^6 variants were included, but step 1 can be forced. Pruning of variants during QC prior to step 1 should take care of this by reducing the number of SNPs.