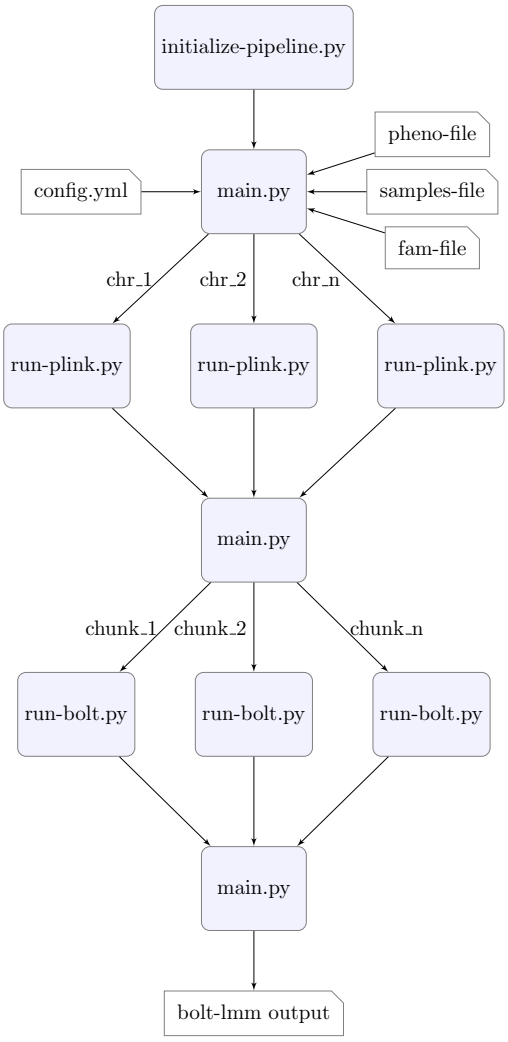
The bolt-lmm pipeline on the high-performance computer cluster

Introduction

This pipeline runs bolt-lmm (Loh et al, Nat Genet 2015; Loh et al. Nat Genet 2018) with UK biobank data on the Imperial hpc cluster. It performs steps to format data, divide them into chunks and run the chunks through bolt-lmm in



parallel (see)

Environment

The pipeline needs software and python packages installed and in the environment path. On the Imperial hpc cluster, this is achieved by two methods:

1. Modules to load installed software into the search path. This is carried out by the pipeline itself. Required modules need to be listed in the configuration file.
2. The Conda enironment, providing python, python packages and other software defined by the user. For instructions on how to use a conda environment see https://www.imperial.ac.uk/admin-services/ict/self-service/research-support/rcs/support/applications/python/ Before using the conda environment for this pipeline for the first time, you need to set it up and install some python modules.

```
module load anaconda3/personal
anaconda-setup
conda install pandas
conda install pyyaml
```

Configuration

A pipeline run is configured by the yaml-format file config.yml. An example configuration file is located in /rds/general/project/uk-biobank-2020/live/software/bolt-lmm-pipeline/config/config.yml. Copy this file to a convenient location and edit the configuration to your needs. For pipeline tests, the phenotype file in /rds/general/project/uk-biobank-2020/live/software/bolt-lmm-pipeline/data/sample.phenotype.txt can be used. At least the output directory needs to be adjusted.

Temporary files directory The program produces temporary files and directories, the location of which can be set with the 'tempdir' variable. These files take a lot of space, therefore it is recommended to choose a location on the ephemeral directory (default is /rds/general/user/$USER/ephemeral/). The variable temp-delete (True/False) determines if the temporary directory gets deleted at the end of the pipeline run.

Covariate syntax To use columns in the phenotype file in the model, the config file uses the following syntax:

```
cov-1: "cat_cov1,...,cat_covn;quant_cov1,...,quant_covn"
```

A comma-separated list of categorial covariates, followed by a semicolon, followed by a comma-separated list of quantitative covariates. For example:

1. Categorial and quantitative covariates:

```
cov-1: "Sex,Center;Age,PC1,PC2,PC3,PC4"
```

2. Quantitative covariates only:

```
cov-1: ";age,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10"
```

How to run the pipeline

1. Loading the conda environment.

```
module load anaconda3/personal
```

2. Starting the pipeline

```
python /rds/general/project/uk-biobank-2020/live/software/bolt-lmm-pipeline/bin/initialise-p
```

3. Pipeline help message

```
python /rds/general/project/uk-biobank-2020/live/software/bolt-lmm-pipeline/bin/initialise-p
```

Data files

1. Phenotype file, containing phenotypes and covariates, with the first line containing column headers and subsequent lines containing records, one per individual. bolt-lmm requires this to be a whitespace-delimited file, i.e. tab-delimited will do. The first two columns must be FID and IID (the PLINK identifiers of an individual). Any number of columns may follow. Values of -9 and NA are interpreted as missing data. All other values in the column should be numeric.
2. fam file (plink –bfile argument).
3. Sample file (bolt –sampleFile argument).
4. Genotype file(s) in .bgen format (bolt –bgenFile argument). Currently these are the ukb_imp_chr*.bgen files in /rds/general/project/uk-biobank-2017/live/reference/sdata_latest/ by default.
5. File listing missing samples to remove (bolt –remove argument), e.g. if samples in fam-file are missing in sample-file. Tab-delimited text file, no header, FID IID must be first two columns. If this is the case and no remove-file is provided, bolt-lmm produces a file listing the samples to remove and exits with an error. The generated file can be used as remove-samples-list in a new run.

Version history

- 0.01 (2022-07-28) First version running on hpc cluster

TODO

- variant annotation
- multiple models in parallel
- bolt core SNPs concatenation?
- mail upon job completion
- resource allocation, multi-threading
- check queues (medbio?)
- check warning: Overlap of sample file and fam file $< 50\%$
- dedicated conda environment?
- accessible location