vcf2gwas

Contents

About The Project

Performing a genome-wide association study (GWAS) on a dataset can be a laborious task, especially when analysing multiple phenotypes. VCF and input files have to be processed and prepared in the right way depending on the way the analysis is performed and afterwards various operations need to be carried out.

vcf2gwas is a Python-built API for GEMMA, PLINK and bcftools performing GWAS directly from a VCF file as well as multiple post-analysis operations.
Some of the benefits of this pipeline include:

- VCF file does not need to be converted or edited by the user
- Input files will be adjusted, filtered and formatted for GEMMA
- GEMMA analysis will be carried out automatically (both GEMMA's linear (mixed) models and bayesian sparse linear mixed model available)
- Dimensionality reduction via PCA or UMAP can be performed on phenotypes / genotypes and used for analysis.
- Once the analysis has been executed, the results will be analyzed:
  - Manhattan plots, Q-Q plots and diagnostic plots (dependent on GEMMA's model)
  - Summaries of the SNPs
  - Comparison to genes
- vcf2gwas is able to analyze several input files with different sets of individuals and multiple phenotypes in a efficient manner due to parallelization, saving the user a lot of time compared to standard GWAS procedure
- Results are reproducible on any compatible machine
- Figures are publication-ready

If you use vcf2gwas in your research please cite us: vcf2gwas - Python API for comprehensive GWAS analysis using GEMMA

Built with

vcf2gwas was built using Python, bcftools, PLINK and GEMMA.

GEMMA is the software implementing the Genome-wide Efficient Mixed Model Association algorithm for a standard linear mixed model and some of its close relatives for GWAS. It fits either a univariate linear mixed model, a multivariate linear mixed model or a Bayesian sparse linear mixed model.

The exact versions of Python, bcftools, PLINK and GEMMA used to build the pipeline are available in the environment file.

Getting Started

These instructions will provide an easy way to get vcf2gwas running on your local machine.
vcf2gwas works on macOS and Linux systems.

Prerequisites

The only requirement is an up to date version of conda installed on your machine.

Installation

It is a good practice to install the package in a clean environment.
So first create a new environment (you can name it as you like), here with the exemplary name 'myenv':

```
conda create -n myenv
```

Next, activate the environment by typing:

```
conda activate myenv
```

Now, the vcf2gwas package can be installed:

```
conda install vcf2gwas -c conda-forge -c bioconda -c fvogt257
```

Everything is ready for analysis now.
Optionally, to test the installation and copy the example files to your current working directory, run:

```
vcf2gwas -v test
```

Once the analysis is completed, the environment can be deactivated:

```
conda deactivate
```

Usage

The items below will explain the required format of the input files, the basic usage and available options as well as the structure of the output files.

Input Files

There are multiple files that can be provided as input for vcf2gwas, below you can find an overview over these files.
For more information about the example files provided with vcf2gwas, please refer to the manual.

VCF file:

A VCF file containing the SNP data of the individuals to be examined is required to run vcf2gwas. This file does not need to be altered in any way and can be in either `.vcf` or `.vcf.gz` format.

Phenotype file(s):

One or multiple phenotype files can be used to provide the phenotype data for GEMMA. These files need to be in the comma separated `.csv` format.
In the first column one has to put the IDs of the individuals. These IDs must match the individuals' IDs of the VCF file, since mismatched IDs will be removed from analysis. The remaining columns resemble the phenotypes with the phenotype description as the column name. vcf2gwas will recognize either "-9" or "NA" as missing values and the phenotypes can be either continuous or binary.
Example files to run GEMMA can be found in the input folder (VCF file + corresponding phenotype file with one phenotype). Below is an excerpt of the exemplary phenotype file `example.csv`:

	avrRpm
5837	1
6009	1
6898	1
6900	0
6901	0

Covariate file:

Note: A covariate file can only be used to provide covariates for the GEMMA analysis when running the linear model or the linear mixed model.
The covariate file has to be formatted in the same way as the phenotype file, with individual IDs in the first column and the covariates in the remaining columns with their respective names as column names.

Gene file:

vcf2gwas has GFF files for the most common species built-in. To compare the results, use the species abbreviation with the `-gf` / `--genefile` option (see File affiliated options). For more information about the available species, their abbreviations and the reference file used, please refer to the manual.
To compare the results of the GWAS analysis with specific genes, a gene file can be provided as input. The gene file has to be either a GFF3 formatted `.gff` file or a comma separated `.csv` file.
If in the `.csv` format, the file needs at least three columns containing information about chromosome, gene start position and gene stop position. These columns have to be named 'chr', 'start' and 'stop'.
Furthermore it is necessary that the chromosome information is in the same format as the chromosome information in the VCF file, otherwise vcf2gwas won't recognize the information correctly.
Optional columns providing additional information have to be called 'ID', 'name' and 'comment'. Below is an excerpt of an exemplary gene file in the `.csv` format:

chr	start	stop	ID	name	comment
1	3644420	3647768	AT1G10920.1	LOV1	
1	4144935	4147817	AT1G12220.1	RPS5	Disease resistance protein family
5	17462611	17467448	AT5G43470.1	RPP8	Disease resistance protein family

Relatedness matrix:

To perform GWAS, GEMMA needs a relatedness matrix, which vcf2gwas will calculate by default. Nonetheless one can provide a relatedness matrix manually.

Running vcf2gwas

Once the virtual environment is activated, vcf2gwas can be run on the command-line by specifying the input files and the statistical model chosen for GEMMA. Below is an exemplary command for running a linear mixed model analysis on all phenotypes in `example.csv` using genotype information from `example.vcf.gz`, both in the input directory.

```
vcf2gwas -v example.vcf.gz -pf example.csv -ap -lmm
```

The available options will be elucidated in the next section.
In the manual, detailed instructions on how to run vcf2gwas and its available options can be viewed.

Available Options:

File affiliated options:

- `-v` / `--vcf`
  Specify genotype `.vcf` or `.vcf.gz` file (required).
- `-pf` / `--pfile`
  Specify phenotype file.

- `-p` / `--pheno`
  Specify phenotypes used for analysis:
  Type the phenotype name
  OR
  '1' selects first phenotype from phenotype file (second column), '2' the second phenotype (third column) and so on.

- `-ap` / `--allphentypes`
  All phenotypes in the phenotype file will be used.

- `-cf` / `--cfile`
  Type 'PCA' to extract principal components from the `VCF` file
  OR
  Specify covariate file.

- `-c` / `--covar`
  If 'PCA' selected for the `-cf` / `--cfile` option, set the amount of PCs used for the analysis
  Else:
  Specify covariates used for analysis:
  Type the covariate name
  OR
  '1' selects first covariate from covariate file (second column), '2' the second covariate (third column) and so on.

- `-ac` / `--allcovariates`
  All covariates in the covariate file will be used.

- `-chr`/ `--chromosome`
  Specify chromosomes for analysis.
  By default, all chromosomes will be analyzed.
  Input value has to be in the same format as the CHROM value in the VCF file

- `-gf` / `--genefile`
  Specify gene file.

- `-gt` / `--genethresh`
  Set a gene distance threshold (in bp) when comparing genes to SNPs from GEMMA results.
  Only SNPs with distances below threshold will be considered for comparison of each gene.

- `-k` / `--relmatrix`
  Specify relatedness matrix file.

- `-o`/ `--output`
  Change the output directory.
  Default is the current working directory.

GEMMA affiliated options:

- `-lm` {1,2,3,4}
  Association Tests with a Linear Model.

optional: specify which frequentist test to use (default: 1)
1: performs Wald test
2: performs likelihood ratio test
3: performs score test
4: performs all three tests

- `-gk` {1,2}
  Estimate Relatedness Matrix from genotypes.
  optional: specify which relatedness matrix to estimate (default: 1)
  1: calculates the centered relatedness matrix
  2: calculates the standardized relatedness matrix

- `-eigen`
  Perform Eigen-Decomposition of the Relatedness Matrix.

- `-lmm` {1,2,3,4}
  Association Tests with Univariate Linear Mixed Models.
  optional: specify which frequentist test to use (default: 1)
  1: performs Wald test
  2: performs likelihood ratio test
  3: performs score test
  4: performs all three tests
  To perform Association Tests with Multivariate Linear Mixed Models, set '-multi' option

- `-bslmm` {1,2,3}
  Fit a Bayesian Sparse Linear Mixed Model
  optional: specify which model to fit (default: 1)
  1: fits a standard linear BSLMM
  2: fits a ridge regression/GBLUP
  3: fits a probit BSLMM

- `-m` / `--multi`
  performs multivariate linear mixed model analysis with specified phenotypes
  only active in combination with '-lmm' option

- `-w` / `--burn`
  specify burn-in steps when using BSLMM model.
  Default value: 100,000

- `-s` / `--sampling`
  specify sampling steps when using BSLMM model.
  Default value: 1,000,000

- `-smax` / `--snpmax`
  specify maximum value for 'gamma' when using BSLMM model.
  Default value: 300

Miscellaneous options:

- `-M` / `--memory`
  set memory usage (in MB)
  if not specified, half of total memory will be used

- `-T` / `--threads`
  set core usage
  if not specified, all available logical cores minus 1 will be used

- `-q` / `--minaf`
  minimum allele frequency of sites to be used (default: 0.01)
  input value needs to be a value between 0.0 and 1.0

- `-t` / `--topsnp`
  number of top SNPs of each phenotype to be summarized (default: 15)
  after analysis the specified amount of top SNPs from each phenotype will be considered

- `-P` / `--PCA`
  perform PCA on phenotypes and use resulting PCs as phenotypes for GEMMA analysis
  optional: set amount of PCs to be calculated (default: 2)
  recommended amount of PCs: 2 - 10

- `-U` / `--UMAP`
  perform UMAP on phenotypes and use resulting embeddings as phenotypes for GEMMA analysis
  optional: set amount of embeddings to be calculated (default: 2)
  recommended amount of embeddings: 1 - 5

- `-KC` / `--kcpca` Kinship calculation via principal component analysis instead of GEMMA's internal method
  optional: r-squared threshold for LD pruning (default: 0.5)

- `-sv` / `--sigval`
  set value where to draw significant line in manhattan plot
  represents -log10(1e-).
  Default: Bonferroni corrected with total amount of SNPs used for analysis.
  set to '0' to disable line

- `-nl` / `--nolabel`
  remove the SNP labels in the manhattan plot
  reduces runtime if analysis results in many significant SNPs

- `-nq` / `--noqc`
  deactivate Quality Control plots
  reduces runtime

- `-fs`/ `--fontsize`
  set the fontsize of plots.
  Default value: 26

- `-sd` / `--seed`
  perform UMAP with random seed

reduces reproducibility

- `-r` / `--retain`
  keep all temporary intermediate files
  e.g. subsetted and filtered VCF and .csv files

Output

vcf2gwas will create an output folder with a hierarchical structure consisting of multiple folders containing plots, summaries, GEMMA output files, log files and so on, depending on the selected options.
Below are the QQ-plot and manhattan-plot that are produced when running the test command mentioned in Installation:
Manhattan-plot, QQ-plot

The exemplary directory and file structure of the output folder after running a linear mixed model analysis on a single phenotype is shown below:

```
output/
└── 'model'
    ├── 'phenotype'
    │   ├── QQ
    │   │   └── QQ plot figure (.png)
    │   ├── GEMMA output file (.txt)
    │   ├── GEMMA log file (.txt)
    │   ├── best_p-values
    │   │   ├── top 1% variants (.csv)
    │   │   ├── top 0.1% variants (.csv)
    │   │   └── top 0.01 variants (.csv)
    │   └── manhattan
    │       └── manhattan plot figure (.png)
    ├── files
    │   └── files_'file'
    │       ├── PLINK BED files (.bed, .bim, .fam, .nosex)
    │       ├── PLINK log file (.log)
    │       ├── GEMMA relatedness matrix (.txt)
    │       └── GEMMA log file (.log.txt)
    ├── logs
    │   └── analysis log file (.txt)
    ├── QC
    │   ├── phenotype QC plot (.png)
    │   └── genotype QC plots (.png)
    ├── vcf2gwas log file (.txt)
    └── summary
        ├── summarized top SNPs (.csv)
        └── top_SNPs
            └── phenotype top SNPs (.csv)
```

The names of the folders in quotes as well as the file names will vary based on the selected options and the file and phenotype names.

License

Distributed under the terms of the GNU General Public License. See LICENSE for more information.

Contact

If you run into any troubles, please raise an issue on the github page.

Acknowledgements

The GEMMA software was developed by:

Xiang Zhou
Dept. of Biostatistics
University of Michigan

Peter Carbonetto, Tim Flutre, Matthew Stephens, Pjotr Prins and others have also contributed to the development of the GEMMA software.

- Genome-wide efficient mixed-model analysis for association studies
- Efficient multivariate linear mixed model algorithms for genome-wide association studies
- Polygenic Modeling with Bayesian Sparse Linear Mixed Models