PennCNV pipeline
User guide

Aurélien Macé
Statistical Genetics Group (University of Lausanne)
aurelien.mace@unil.ch

March 23, 2016

Contents

1 Overview

This pipeline creates a framework around PennCNV for CNV detection and association with a given phenotype. It works in three parts:

- CNV detection using pennCNV
- Calculation of a Quality Score (QS) for each CNV
- Association between this QS and a given phenotype

It is possible to run this pipeline on multiple cores to reduce processing time. 1150 samples on Illumina 1M platform will take approximately 3h using 16 cores.

2 Requirements

Three software have to be installed and set in the system path:

- *pennCNV*: `http://www.openbioinformatics.org/penncnv/penncnv_download.html`
- *R*: `www.r-project.org/`
- *Perl*: `https://www.perl.org/`

And one package for *R*:

- *stringr*: `http://cran.r-project.org/web/packages/stringr/index.html`

3 Installation

3.1 pennCNV

1. Download:
   In a folder, download from the above link the penncnv.latest.tar.gz file.
2. Uncompress:
   Using the command line go in this folder and uncompress this file: `tar -xvzf penncnv.latest.tar.gz`
3. Check:
   With the command line go in the *pennCNV* folder.
   Check if the installation is ok by just launching `detect_cnv.pl` in the command line.
4. Error:
   If you get an error message concerning `khmm` please inside the *pennCNV* folder go in the `kext` folder and execute the Makefile: `make`
   Depending on the installation of your machine you may get another error concerning -lperl, if so please install the perl library by executing e.g. on Debian the following command: `sudo apt-get install libperl-dev`
   Then execute the `make` and you should be able to run the *pennCNV* functions like `detect_cnv.pl`.
5. Example:
   See the Example section for more information

3.2 R

- Binary file:
  Download the binary files e.g. from `http://cran.at.r-project.org/`
  Choose the one corresponding to your OS and follow the instructions.
- Source code:
  1. Download the source code from the same url address as for the binary files
  2. *R* compilation:
     - On the command line go in the *R* folder
     - Run the configuration `./configure`
     - Execute the makefile `make`
     - For more information read the INSTALL text file
- Packages:
  1. Install the following libraries:
     - *stringr*
     - *parallel*
     - *plyr*
  2. To install a package, open *R* and execute the command: `install.packages("stringr")`

3.3 Perl

Most probably *Perl* is installed by default on your machine.

4 Files

4.1 Input files

4.1.1 Signal intensity files

It can either be the raw report file(s) from Illumina (created with Genomestudio) or file(s) already formatted for pennCNV. The type of file has to be specified in the configuration file.
The signal intensity file has to contain for each probe the **Log R ratio** (called LRR) and the **B allele frequency** (called BAF). The first value is the `log` of the ratio between the total observed intensity and the expected intensity. The second value is the intensity ratio between the B allele and the A allele, e.g. 1 means BB and 0 means AA. For more information, you can refer to the following page: `http://www.openbioinformatics.org/penncnv/penncnv_input.html#_Toc214852003`

4.1.2 HMM files

Some HMM files for Illumina platforms are provided by pennCNV. The default one currently used, either for the classic Illumina platform or the Metabochip platform, is called hhall.hmm and is located in the lib folder of pennCNV. The HMM file is used by pennCNV to get the transition matrices for the Hidden Markov Model to determine a copy number state in function of the LRR, the BAF and the state of the previous probe. For more information, you can refer to the following page: `http://www.openbioinformatics.org/penncnv/penncnv_input.html#_Toc214852003`

4.1.3 Population Frequency of B allele (PFB)

The user can either specify a given PFB file or decide to compile a new one based on the raw signal intensity files. PennCNV provide PFB files for classical Illumina platforms in the lib folder but usually we compile a specific PFB file for each cohort. This can be specified in the configuration file. The PFB file represents the frequency of the B allele for each probe in a given population. For more information, you can refer to the following page: `http://www.openbioinformatics.org/penncnv/penncnv_input.html#_Toc214852003`
To compile a PFB file the user needs to give as input the report files from BeadStudio and not the formatted files for PennCNV. In case you want to compile a new PFB but already have the formatted file for PennCNV, please give the path to the report file in the section *DATA* of the configuration file and the path the formatted files in the section *FormatedPath*.

4.1.4 Configuration file

This file (tab-delimited) is used to define some specific parameters for the CNVs call and summary statistics calculation. Find below the parameters which can be configured:

pennCNVpath:
Define the path to the folder containing all the pennCNV functions

HMMpath:
Define the path to the HMM file or the path to the folder where to save the created HMM file

HMMcreate:
Define whether a HMM file has to be created. 1 stands for yes and 0 for no

PFB:
Define the path to the PFB files or the path to the folder where to save the created PFB file

CompilePFB:
Define whether a PFB file has to be created. 1 stands for yes and 0 for no.
If you want to compile a new PFB file you need to give as input the report files from Beadstudio and not the formatted ones for PennCNV

GCmod:
Define the path to the GC model files or the path to the folder where to save the created GC model file

UseGCmod:
Define whether to use or not the GC model file. 1 stands for yes and 0 for no

InputData:
Integer defining the type of input data. 0 stands for already formatted, 1 for Illumina format, 2 for Metabochip format

DATA:
Path to the signal intensity files (formatted or not). All the files in the folder will be processed, so **do not put any other files in this folder**.

OUTPUT:
Path where to save the results of the CNVs calls.

FormatedPath:
Path where to save the formatted input files, if the input files have to be formatted

Chromosome:
Chromosomes on which the CNV filtering function can be applied

CNVcall:
Boolean to define whether the CNV call has to be done or not. 1 stands for yes and 0 for no

Cleancall:
Boolean to define whether the CNVs merging has to be done or not. 1 stands for yes and 0 for no

format:
Boolean to define whether the input files have to be formatted or not. 1 stands for yes and 0 for no

CreateRfile:
Create the R files which will be used for the association section.

AssoData:
Boolean to define whether summary statistics have to be calculated or not. 1 stands for yes and 0 for no

NbCores:
Integer defining the number of cores to use for the CNVs detection.

PhenoPath:
Path to the phenotypic data. It has to be a tab-delimited file with column names ID and pheno

Phenotype:
The type of phenotype, e.g. *BMI* or *WT*.

4.1.5 Phenotype file

This tab-delimited file contains the phenotypic data of the samples in the cohort. The first column corresponds to the sample id and the second to the phenotype, they have to be called ID and pheno.
The sample's ids have to be the same as the ones in the raw data files (column Sample ID in the final report files).
The phenotypic values have to be corrected for sex, age and age^2.

4.2 Scripts

4.2.1 Main file

The main script file is a bash file called pennCNV_pipeline.sh which calls:

- the functions to convert the input files if needed
- the pennCNV function for the CNVs detection
- the *R* functions to calculate the Quality Score and the summary statistics

4.2.2 *R* scripts

Four *R* scripts are used to process the results from pennCNV. They :

- convert the pennCNV output in a .rdata format
- calculate for each CNV a specific quality score (QS)
- build SNP by samples matrices for each chromosome with the QS values
- calculate values used for association with the phenotype of interest

5 Usage

5.1 Configuration file

5.1.1 General

Fill the configuration file with the proper paths and define the action to do. The configuration file has to be **tab-delimited** and located in the same folder as the main bash script or in another folder if the entire path is specified. Below an example of a configuration file:

```
pennCNVpath:    /data/sgg/aurelien/software/pennCNV/penncnv_2
HMMpath:        /data/sgg/aurelien/software/pennCNV/penncnv_2/lib/hhall.hmm
HMMcreate:      0
PFB:    example/pfb
CompilePFB:     1
GCmod:  my/GC_model/path
UseGCmod:       0
InputData:      1
DATA:   example/raw
OUTPUT: example/results
FormatedPath:   example/formated
Chromosome:     1-22
CNVcall:        1
Cleancall:      1
format: 1
CreateRfile:    1
AssoData:       1
NbCores:        16
PhenoPath:      example/phenotype/phenotype.txt
Phenotype:      example
```

Figure 1: Configuration example

5.1.2 In the case of the study

For the current study please follow the rules below:

- Set the path to the folder where you store *pennCNV*
- Used the default hmm file given by *pennCNV* and called hhall.hmm located in the *pennCNV* lib folder
- Compile a pfb file: give the folder path at the line PFB in the configuration file and set the value CompilePFB to 1
- Nothing to do with the GC model, no path and UseGCmod set to 0
- For the first time, format the raw data:
  - set InputData to 1
  - set DATA to the path to the folder with the Final Reports from GenomeStudio
  - set the path of the folder where to store the formatted data at the line FormatedPath
- For a second use, take directly the already formatted data:
  - set InputData to 0
  - set DATA to the path to the folder with the formatted data
- at the line OUTPUT, write the path of the folder where to store the results
- set the line Chromosome to 1-22. We are interested in autosomes
- set the value CNVcall to 1 to run the CNV call
- set the value Cleancall to 1 to clean the CNV call
- set the value format to 1 to format the raw data
- set the value CreateRfile to 1 to create the R files used for the association
- set the value AssoData to 1 to run the association
- set NbCores to the right number of cores, the more the faster
- set the path to the phenotypic data on the line PhenoPath
- set Phenotype to the phenotype name e.g. BMI or WT

Note: No need to rerun the entire pipeline to calculate association with a new phenotype, it is possible to only run the last part by setting AssoData variable to 1 and fixing all the others to 0 (CompilePFB, CNVcall, Cleancall, format, CreateRfile). Example below:

```
pennCNVpath:    /data/sgg/aurelien/software/pennCNV/penncnv_2
HMMpath:        /data/sgg/aurelien/software/pennCNV/penncnv_2/lib/hhall.hmm
HMMcreate:      0
PFB:    example/pfb
CompilePFB:     0
GCmod:  my/GC_model/path
UseGCmod:       0
InputData:      1
DATA:   example/raw
OUTPUT: example/results
FormatedPath:   example/formated
Chromosome:     1-22
CNVcall:        0
Cleancall:      0
format: 0
CreateRfile:    0
AssoData:       1
NbCores:        16
PhenoPath:      example/phenotype/New_phenotype.txt
Phenotype:      New_example
```

Figure 2: Configuration to only run association with a new phenotype

5.2 Launch the main script

- Fill the configuration file and save it
- Open the command line
- Go in the pipeline script folder
- Launch the CNV_detection.sh file followed by the configuration file name and path (if not in the same folder as the main script), e.g.: `./CNV_detection.sh my_config_file.txt`.

For a basic test please refer to the Example section.
Depending of the number of samples, the need or not to compile a PFB file and the number of cores used for the detection, the global process will take more or less time.

6 Results

Depending of the options in the configuration file, the pennCNV pipeline will produce more or less files and folders

6.1 To upload

The files to upload to our ftp server are stored in the *to_upload_PHENONAME* folder in the output directory:

pheno_info_PHENONAME.txt:
: A text file with the number of samples with genotypic and phenotypic information, the mean value of the phenotype, the sum of squares of the phenotypic values

association_summary_PHENONAME.txt:
: For each probe the mean of the genotypic values, the sum of squares of the genotypic values, the sum of the product genotype by phenotype.

association_summary_burden_PHENONAME.txt:
Summary statistics for burden test: mean of sum for each sample of the absolute Quality Score, sum of square of the absolute QS sum, sum of the product phenotype by absolute QS sum.

log_info_raw.txt:
Gives some information on the total number of CNVs, of samples, deletions, duplications.

log_CNV_summary_dataframe.txt:
Corresponds to the print of the *R* `summary` function applied to the dataframe containing information on all the CNVs detected by *pennCNV*.

log_pheno_histogram_PHENONAME.rdata:
A *R* format file with the histogram information of the phenotype of interest.

log_Quality_Score_histogram.rdata:
A *R* format file with the histogram information of the Quality Scores.

log_pipeline_PHENONAME.log:
Log file of the global pipeline.

6.2 PennCNV outputs

ex1.rawcnv:
The pennCNV raw CNVs without any filtering and merging.

ex1.log:
The log of the pennCNV calls.

clean.rawcnv:
The merged pennCNV raw CNVs.

goodCNV.good.cnv:
The merged and filtered CNVs in the pennCNV format.

QCpass.qcpass:
The list of samples surviving the quality control.

QCsum.qcsum:
Summary value for all the samples produced by pennCNV.

SNPs_list.txt:
The list of SNPs on the platform used for the calls.

Folder *R*:
This folder contains the .rdata files used to calculate the associations data described above.

7 Example

As example a set of raw data (final reports from GenomeStudio) has been provided in the example/raw folder and a configuration file called Config_default_Example.txt, located in the piepline forlder, is almost ready to use.

- Open the Config_default_Example.txt file
- Set the right path to the pennCNV library where the HMM file is stored.
- Open the command line
- In the pipeline folder execute the command: `./CNV_detection.sh Config_default_Example.txt`

It uses by default 16 cores and should take less than a minute to process.
Some warnings may appears especially one concercing a folder that already exist and some in R concerning NAs, these warnings are normal, no need to worry.

8 Questions

Don't hesitate to contact me if you have any question or issue.
Aurélien Macé: *aurelien.mace@unil.ch*