PattRec

v1.0

Roca I, González-Castro L, Fernández H, Couce ML,
Fernández-Marmiesse A.

CONTENTS

1. Introduccion 3
2. Installation 3
2.1. Ubuntu systems 3
2.2. Windows systems 3
3. Getting ready 4
4. Analysis 4
5. Output 5
6. Acknoledgements 7

1. Introduccion

PattRec is a bioinformatics tool designed to detect rare copy number variants (CNVs) in targeted Next Generation Sequencing (tg-NGS) data. It is presented as a Java-based GUI, with its CNV detection algorithm implemented in R.

This tool was designed for use with target gene panels, sequenced in Illumina platforms.

2. Installation

PattRec was tested for Ubuntu 14.04, 16.04, 18.04 and Windows 10.

Minimum system requirements: 8GB RAM.

2.1. Ubuntu systems

2.1.1. Dependencies

PattRec has the following dependencies:

- Java 8 (https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html)
- MySQL (https://dev.mysql.com/downloads/) or MariaDB (https://mariadb.org/download/)
- R Project (https://www.r-project.org/)
- R package rJava (http://www.rforge.net/rJava/)
- Bedtools (https://bedtools.readthedocs.io/en/latest/)
- SAMtools (https://sourceforge.net/projects/samtools/)

Please make sure to have all the dependencies installed before running PattRec.

2.1.2. Installation

Some configurations have to be performed before running the program:

- If you haven't done it yet, install rJava:
  - From R: *$ install.packages("rJava")*
  - From Terminal: *$ R CMD INSTALL lib/rJava_0.0-11.tar.gz*
- Set configurations (change the path if needed):
  - *$ export R_HOME=/usr/lib/R*
  - *$ export LD_LIBRARY_PATH=/usr/lib/R/site-library/rJava/jri*
- Run the jar file from Terminal:
  - *$ java -jar pattrec.jar*

2.2. Windows systems

2.2.1. Dependencies

PattRec depends on R v4.3 (with rJava package), Java 8, MySQL or MariaDB, Visual Studio, and Perl 5.12. The executable allows the user to install all the dependencies, or just those missing, so there is no need for users to manually install anything.

2.2.2. Installation

If any of the dependencies are already installed on the computer, please uncheck the corresponding box. If the dependencies already installed are R v3.4 or Java 8, the installer will ask you to select their complete path. Please make sure that you have installed the R package rJava; if you haven't, just open a R session and type:

```
$ install.packages("rJava")
```

If you choose to install MySQL via the installer, please make sure to complete the following steps in the MySQL installer dialogue:

- Select 'Typical' installation
- Launch the 'Wizard' in order to configurate the database.
- Select 'Detailed Configuration'
- Add MySQL to the firewall
- Create a password for 'root'

3. Getting ready

In order to perform the analysis, PattRec needs the following files:

- BAM test: the BAM file of the sample user wants to perform the analysis on.
- BAM control/s: one or more BAM files to compare with the BAM file. (For optimal results, it is highly recommended that all the BAM files were sequenced on the same run.)
- BED file: file containing the targeted regions (sequenced regions) in tab separated format. For Windows users, the header of the BED file must be removed, so the first line in the file corresponds to a region; example:
  - *chr1 1000 2000 Gene_name*
- FASTA file: genome reference in fasta format (it must be the same used for the alignment of the BAM files).
  - GRCh37-hg19 can be downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/
  - GRCh38-hg38 can be downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/

4. Analysis

PattRec creates and automatically updates an in-house database with all the CNVs detected in the analysis, to help users filter out false positives and identify polymorphic CNVs in the genome. The first time you run PattRec, the database must be configured: click "Database configuration" and enter the user and password for MySQL or MariaDB server.

Once the input files are selected (test and controls BAM files, BED file and FASTA file), click 'RUN'. This will open a new prompt where several parameters and options are displayed. Once these parameters are configured (if needed), click 'Proceed'. When the analysis finished you will be asked to save the results in the database. If you agree, the results of this analysis will be used in the next execution of the algorithm.

Results are placed in the user home, in the folder 'PattRec'.

PattRec has several parameters and options for a better adjustment to each analysis:

- ***Parameters***:
  - **MIN DUP (0-1)**: minimum percentage of increase (compared to 'normal' samples) to call a duplication (range between 0-1). Default parameter: 0.3 (30%)

  - **MIN DEL (0-1)**: minimum percentage of decrease (compared to ‘normal’ samples) to call a deletion (range between 0-1). Default parameter: 0.35 (35%)
  - **MIN CONT COV (>0)**: minimum depth of coverage mean in control samples to call a CNV. Default parameter: 50X.
  - **MIN GENE DUP (0-1)**: minimum percentage of increase (compared to ‘normal’ samples) to call a duplication in whole gene comparison (range between 0-1). Default parameter: 0.3 (30%)
  - **MIN GENE DEL (0-1)**: minimum percentage of decrease (compared to ‘normal’ samples) to call a deletion in whole gene comparison (range between 0-1). Default parameter: 0.35 (35%)
- ***Options***:
  - **USE VCF**: whether to use SNVs filtering for the analysis or not. With this option, regions annotated as a deletion with heterozygous SNVs are discarded from the final output. If selected, Windows users must insert a vcf file; Ubuntu users, on the other hand, can choose to either insert one, or to create a new one.
  - **IGNORE POLYMORPHIC**: when chosen, regions contained in the ‘polymorhic.bed’ file are filtered out from the analysis. A different bedfile than the one provided can be chosen.
  - **XLSX OUTPUT**: whether to write the output in xlsx format or not (ie, output in plain text). Default: TRUE.
  - **GENE ANALYSIS**: whether to perform whole gene analysis or not. Default: TRUE.
  - **RESTRICTIVE**: whether to allow a more restrictive filtering or not. If selected, and if the final output has more than 10 regions, then the regions with a p-value higher than the median of all of them are removed from the final output. Default: TRUE.
  - **FIXED CONTROL SAMPLES**: whether to use all the control samples provided by the user or not. If this option is selected, all the control samples are used; if it is not selected, then the program chooses the control samples more similar to sample test. Default: TRUE.
  - **DOWNSAMPLING**: if this option is selected, then the program performs the same analysis on 5 copies of the sample test (these copies are obtained by ‘downsampling’ the original BAM file a 20% of its original depth of coverage); if a region is not present in the final output of 80% of the copies, it is remove from the final output of the original comparison. Default: FALSE.
  - **PLOT**: if this option is selected, the program will plot the depth of coverage of the genes with regions considered CNVs. Default: FALSE.

5. Output

Either by selecting the option to write the output in xlsx format or in plain text, the user would have two or three sheets or txt files: one corresponding to the exons calls (“regions”), one corresponding to the genes calls (“genes”; only if the “GENE ANALYSIS” option is chosen), and one containing global information about the samples.

Each sheet or file has the following information:

- ***Regions***:
  - **Chr**, **Start**, **End**: chromosomal position of each region.
  - **Gene**: name of the gene containing each region.
  - **Type**: type of the CNV (deletion/duplication).

  - **pvalue**: p-value of the comparison. If the name is “pvalue_bonf”, then the Bonferroni correction was used; if the name is “pvalue_benj”, then the Benjamini-Hochberg correction was applied.
  - **n_exons**: number of exons within the region.
  - **Test_mean**: mean depth of coverage of the test sample in the region.
  - **Test_gene_max**: maximum depth of coverage of the test sample in the gene containing the region.
  - **Cont_mean**: mean depth of coverage of the control samples in the region. In xlsx output, a color code is applied: orange for regions with values under 100X, blue for regions with values over 100X.
  - **Cont_gene_max**: maximum depth of coverage of the control samples in the gene containing the region.
  - **Cont_sd**: standard deviation of the control samples’ depth of coverage in the region.
  - **CV**: coefficient of variation of the control samples’ depth of coverage in the region.
  - **CV_norm**: coefficient of variation of the control samples’ depth of coverage in the region, normalized by the maximum coverage of the gene.
  - **GC**: GC-content of the region. In xlsx output, a color code is applied: regions with GC content over 0.6 or under 0.35 are colored in red.
  - **%decrease/increase**: Percent decrease or increase of the test sample’s depth of coverage compared to the control samples. In xlsx output, a color code is applied: regions with a percentage over 40% are colored in blue.
  - **#DUP_SNPs**: if the vcf filtering option is chosen, number of SNVs within each duplicated region with a ratio wild-type reads to rare-allele reads compatible with duplication (ie, ratio <0.75 or >1.75).
  - **#Normal_SNPs**: if the vcf filtering option is chosen, number of SNVs within each duplicated region with a ratio wild-type reads to rare-allele reads not compatible with duplication (ie, ratio ≥0.75 and ≤1.75).
  - **Freq**: number of samples in the in-house database with a CNV intersecting the region. In xlsx output, a color code is applied: regions with values over 10 are colored in red, regions with values under 5 are colored in blue.
  - **Samples**: name of the in-house database samples with a CNV intersecting the region, only reported if the number in “Freq” is >0 and <10.
  - **Percentage_Samples**: % of decrease/increase of the in-house database samples, only reported if the number in “Freq” is >0 and <10.
  - **DCVar**: *docvar* of the control samples within the region; this parameter measures the variability of the depth of coverage between samples. In xlsx output, a color code is applied: regions with values over the global DCVar (ie, the mean *docvar* of all exons sequenced; reported in the “information” sheet/file) plus its standard deviation are colored in red, regions with values under the global DVar minus its standard deviation are colored in blue.
  - **Region_info**: exon number (according to each RefSeq accession number of the gene) of all the exons contained within the region.
- ***Genes*****:**
  - **Chr**, **Start**, **End**: chromosomal position of each gene.
  - **Gene**: name of the gene.
  - **Type**: type of the CNV (deletion/duplication).
  - **pvalue**: p-value of the comparison. If the name is “pvalue_bonf”, then the Bonferroni correction was used; if the name is “pvalue_benj”, then the Benjamini-Hochberg correction was applied.

    - **Test_mean**: mean depth of coverage of the test sample in the gene.
    - **Cont_mean**: mean depth of coverage of the control samples in the gene. In xlsx output, a color code is applied: orange for cells with values under 100X, blue for cells with values over 100X.
    - **Cont_sd**: standard deviation of the control samples' depth of coverage in the gene.
    - **CV**: coefficient of variation of the control samples' depth of coverage in the gene.
    - **GC**: GC-content of the gene. In xlsx output, a color code is applied: genes with GC content over 0.6 or under 0.35 are colored in red.
    - **%decrease/increase**: Percent decrease or increase of the test sample's depth of coverage compared to the control samples. In xlsx output, a color code is applied: regions with a percentage over 40% are colored in blue.
    - **#DUP_SNPs**: if the vcf filtering option is chosen, number of SNVs within each duplicated gene with a ratio wild-type reads to rare-allele reads compatible with duplication (ie, ratio <0.75 or >1.75).
    - **#Normal_SNPs**: if the vcf filtering option is chosen, number of SNVs within each duplicated gene with a ratio wild-type reads to rare-allele reads not compatible with duplication (ie, ratio ≥0.75 and ≤1.75).
- ***Information***:
    - **Sample**: samples' (test and controls) names.
    - **Sex**: sex of each sample (only reported if chrX or chrY is sequenced).
    - **Cov_mean**: global depth of coverage's mean of each sample.
    - **Cov_SD**: global depth of coverage's standard deviation of each sample.
    - **CV**: global depth of coverage's coefficient of variation of each sample.
    - **Correlation**: Pearson's correlation between the test and each control sample's depth of coverage normalized by the total number of base pairs sequenced.
    - **DCVar_global(mean)**: *docvar* mean of all the exons sequenced.
    - **DCVar_global(SD)**: *docvar* standard deviation of all the exons sequenced.

6. Acknoledgements

We thank Rosanna Abal García for her invaluable work, and Ignacio Aneiros and Manoel Xosé Polo for their help with the Windows version of PattRec.

This work was supported by the Instituto de Salud Carlos III [grant numbers: PI13/02177, JR13/00019, and PI13/02451].