ASEQ manual v1.1.x

Alessandro Romanel, Francesca Demichelis
romanel@science.unitn.it
demichelis@science.unitn.it

1. PILEUP execution modes

USAGE:
./ASEQ [vcf=string | vcflist=string] [bam=string | bamlist=string] [threads=int] [mbq=int] [mrq=int] [mdc=int] [out=string]

FUNCTION:
*A*llows to execute the pileup for a list of positions using a NGS data file. The application allows also for multiple pileup computations.

INPUT FILES:
A positions list in VCF format and a corresponding NGS data file in BAM format. VCF files should follow VCF format guidelines and have 8 columns with at least reference base provided for each position. For multiple pileup computations the user should provide

- a single VCF file and a file containing a list of paths of multiple BAM files
- a file containing a list of paths (one per line) of multiple VCF files and file containing a list of paths of multiple BAM files; the order in the two files is important because i-th entry in the VCF list corresponds to the i-th entry in the BAM list.

OUTPUT:
For each pileup computation a file *.PILEUP.ASEQ is generated. It is a file in CSV format (we consider CSV format as tab-separated plain text file) and contains the pileup computation.

Download file ASEQ-examples.tar.gz to a concrete example and input/output files structures.

2. GENOTYPE execution mode

USAGE:

./ASEQ vcf=string [bam=string | bamlist=string] mode=GENOTYPE [threads=int] [mbq=int] [mrq=int] [pht=double | htperc=double] [mdc=int] [out=string]

FUNCTION:
*A*llows to call genotype genotype at a position using DNA NGS data.

INPUT FILES:
A positions list in VCF format and a sample NGS data file in BAM format. VCF files should follow VCF format guidelines and have 8 columns. The input position list can be genotyped against different samples NGS data files. In this case the user should provide the file containing a list of paths (one per line) of the multiple BAM files.

OUTPUT:
For each sample three files are generated. The first file *.GENOTYPE.ASEQ is in CSV format (we consider CSV format as tab-separated plain text file) and contains the pileup computation. The second file *.genotype.vcf contains the genotype calls for all input positions. The third file *.heterozygous.vcf contains only the heterozygous calls. Both are files in VCF format. The field INFO of VCF format is updated adding information regarding the genotype test results and the number of read counts for reference and alternative alleles.

Download file ASEQ-examples.tar.gz to a concrete example and input/output file structures.

3. ASE execution mode

USAGE:
./PaPI vcflist=string bamlist=string [genes=string | transcripts=string] mode=ASE [threads=int] [mbq=int] [mrq=int] [pht=double] [pft=double] [mdc=int] [genestat=int] [snpstat=double] [out=string]

FUNCTION:
*A*llows to perform Allele Specific Expression (ASE) analysis.

INPUT FILES:
The user should provide a file containing a list of paths (one per per line) to VCF files containing the heterozygous SNPs, one file per sample, and a file containing a corresponding list of paths (one per line) to RNAseq data BAM files, one file per sample. In addition, the user should provide a file containing a list of genes in BED format or of transcripts in INTERVAL format. We require the BED file to have 4 columns, while INTEVAL format to have 6 columns.

OUTPUT:
For each sample a file *.ASE.ASEQ is in CSV format that contains the pileup computation is generated. In addition the following files are created:

- ASE_genes.csv = Plain text tab-separated file containing information of the genes that show ASE. For each gene we provide basic information of the gene (CHR,FROM,TO,ID), a value representing the result of the ASE calculation and the number of samples that were available for ASE calculation for that gene.
- genes_statistics.csv = Plain text tab-separated file containing detailed ASE information for all the genes with at least one sample available for ASE calculation. For each gene we show basic information (CHR,FROM,TO,ID) a list of values representing the ASE calculation for all the samples (with -1 we indicate a sample not available for ASE calculation).
- snps_per_gene.csv = Plain text tab-separated file containing the number of hitting SNPs for all genes in all the samples.

Download file ASEQ-examples.tar.gz to a concrete example and input/output files structures.

4. Options

mode=string
Execution mode [PILEUP|GENOTYPE|ASE]
(default PILEUP).

vcf=string
List of positions in VCF format.

bed=string
List of positions in BED format.

vcflist=string
File containing a list of positions file names

bam=string
NGS data file in BAM format.

bamlist=string
File containing a list of NGS data files names

genes=string
List of genes coordinates in BED format

transcripts=string
List of transcripts coordinates in INTERVAL format.

threads=int
number of threads used (if available) for the pileup computation
(default 1).

mbq=int
min base quality
(default 1).

mrq=int
min read quality
(default 1).

mdc=int
min depth of coverage that a position should have to be considered in the analysis
(default 1).

pht=double
p-value for heterozigosity test (null hypothesis: alleles are heterozigous)
(default 0.05).

htperc=double
activates heterozigosity test based on percentages. The value specifies the percentage range to be considered for (alt/(ref+alt)) ratio
(default 0.2).

zht
do not automatically consider homozygous a position when reference or alternative allele have zero depth of coverage.

pft=double
p-value for Fisher Exact Test
(default 0.05).

genestat=int
number of samples that have to support ASE in order for a gene/transcript to be flagged as showing ASE
(default 1).

snpstat=double
minimum percentage (the specified value is not included) of ASE positions in a gene/transcript to flag the gene/transcript as ASE in a sample
(value in [0,1] default value 0.5).

out=string
Path of output directory. Keyword STDOUT is used to tell ASEQ to print results on standard output. STDOUT mode is available only with PILEUP mode and only for the pileup of a single file.
(default is the current directory).