NAME

bammds - Creates a multidimensional scaling (MDS) plot of populations for genetic data. The input files are a reference panel with genotypes and one or several bam files.

SYNOPSIS

bammds [--legend *legendfile*] [--dim1 *d1*] [--dim2 *d2*] [--mapquality *qual*] [--ntquality *qual*] [--output *outputfile*] [--pca] [--mds] [*file.bam* ...] [*ref.bed* | *ref.ped* | *ref.tped* | *ref.txt* | *ref.vcf* >] [--tfam *file.tfam*]

bammds [-l *legendfile*] [-x *d1*] [-y *d2*] [-m *qual*] [-n *qual*] [-o *outputfile*] [--pca] [--mds] [*file.bam* ...] [*ref.bed* | *ref.ped* | *ref.tped* | *ref.txt* | *ref.vcf* >] [--tfam *file.tfam*]

DESCRIPTION

bammds will read files with sequencing data (one bam file per individual/ngs library) and will generate an MDS (multidimensional scaling) plot to see how closely related those individuals are to the individuals in the reference panel.

Note that at each position one allele (reference panel) or one read (bam file) is drawn randomly for each individual.

OPTIONS

file.bam

BAM file for individuals to plot. The automatically generated legend and the names of the output files will be based on this file name, so it is recommended to call the files: Population.Individual.whatever.bam

ref.bed

File containing the genotype information of your reference panel in a bed format (see pngu.mgh.harvard.edu/~purcell/plink/ under "binary ped files" for more info). Will be converted in a tped file automatically using plink.

(The information about the grouping of individuals in populations is expected in the tfam file.)

(Please check the notes on reference files below.)

ref.ped

File containing the genotype information of your reference panel in ped format (see pngu.mgh.harvard.edu/~purcell/plink/ under "ped files" for more info). Will be converted in a tped file automatically using plink.

(The information about the grouping of individuals in populations is expected in the tfam file.)

(Please check the notes on reference files below.)

ref.tped

File containing the genotype information of your reference panel in transposed ped format (see pngu.mgh.harvard.edu/~purcell/plink/ under "tped files" for more info).

(The information about the grouping of individuals in populations is expected in the tfam file.)

(Please check the notes on reference files below.)

ref.txt

File containing the genotype information of your reference panel in a custom format (example under XX).

(Please check the notes on reference files below.)

ref.vcf

File containing the genotype information of your reference panel in a vcf format (see

e.g. http://en.wikipedia.org/wiki/Variant_Call_Format). vcftools (http://vcftools.sourceforge.net/) is used to automatically convert this file in a tped file.

(Please check the notes on reference files below.)

--dim1 *d1*

--dim2 *d2*

--xvector *d1*

-x *d1*

--yvector *d2*

-y *d2*

The first and second dimension for the MDS plot. Integer that can go from 1...nindivs-1, where nindivs is the number of individuals in your reference panel plus the number of BAM files. Default: 1 and 2.

--legendfile *legendfile*

-l *legendfile*

The *legendfile* specifies the legend for the individuals, which individuals that should be included and how they should be plotted (shape, size and color of each point).

If left out **bammds** will generate a default legend.csv template (in the tmp directory) and based on this it generates legend_filled.csv. The format is a comma separated value (CSV) file. Change legend.csv using a spreadsheet and look in legend_filled.csv to see examples of values. Also read the section below on the legend file.

--mapquality *qual*

-m *qual*

Skip reads with mapping quality smaller than *qual*. Default: 30.

--ntquality *qual*

-n *qual*

Skip bases with base quality smaller than *qual*. Default: 20.

--outputfile *name*

-o *name*

Name of output file to be generated. Default is to use a name based on the input file names. Valid extensions are: .pdf .png .svg .jpg .jpeg .tif .tiff .csv.

CSV will generate a .csv file with the coordinates of the plots for making your own plots.

--tfam *file.tfam*

If a ref.ped, ref.bed or ref.vcf file is provided, a .tfam file will automatically be generated and used.

If a ref.tped is provided a .tfam file is expected in the same dir.

If you want to use a different .tfam file use **--tfam** to specify that file.

The format is a TAB-separated file with population in column 1 and individual in column 2.

THE LEGEND FILE

The legend.csv-file determines:

- The labels of the populations and the individuals.
- Which populations and which individuals to include/exclude.

- Which individuals should be considered samples and thus be plotted special.
- Color, size and point symbols for each point.

First run of **bammds** will create a template legend that you can change if you do not like the defaults.

The structure of a legend.csv file

Edit the legend.csv file using a spreadsheet. To make it less work to fill out the legend file there are a few shortcuts:

- An empty field gets the value of the field above.
- A field with * will get the default value for this field unless this field has no default in which case the * means match anything. See the specifics for each field below.

Population

This is the population as specified in the TPED/VCF/BED-file. For BAM files the population is the string up to the first '.', so it is recommended to name the BAM files: Population.Individual.whatever.bam

A * in Population will match all populations and can thus be used to set the default value for populations with no specified value.

pop_label

If the Population field is not the label you want on the plot, you can override the label by putting it here.

A * in pop_label means the value in Population.

Individual

This is the individual as specified in the TPED/VCF/BED-file. For BAM files the individual is the string between the first and second '.', so it is recommended to name the BAM files: Population.Individual.whatever.bam

A * in Individual will match all individual and can thus be used to set the default value for individuals with no specified value.

indv_label

If the Individual field is not the label you want on the plot, you can override the label by putting it here.

A * in indv_label means the value in Individual.

sample

Sample can take 3 values:

-1
: Remove this population or individual from the plot.

0 This population or individual is a reference and should be plotted normally.

1 This population or individual is a sample and should be plotted special.

A * in sample means 0.

order

For internal reasons the order must sort the same way as the input files. This is only relevant if you change the order of the lines in the legend.csv-file.

A * will make sure order is in increasing order.

Color

The color is the hex color, so #ff0000 is red. The color can be chosen on http://www.w3schools.com/tags/ref_colorpicker.asp

A * will auto select a light color for reference and black for samples.

pch

The point type is a letter.

If the letter is in [] the letter will be plotted in a circle.

For population a * will chose the first letter of the population.

For individual a * will chose the first letter of a hashed value (i.e. it looks random but will remain the same).

For samples the letter will be in [] and thus plotted in a circle.

cex

The size of the point.

Increase this to make the point in the plot bigger.

A * in cex corresponds to 0.8.

Difference between PCA and MDS

Here should be a description of the difference between PCA and MDS.

Notes on reference files

NOTE: The markers in the reference file should be sorted in the same way as the bamfile (same as reference genome). If they are not sorted the same way, the mpileup part will die. The same applies for all the ref formats.

NOTE2: If a marker in the reference file is not in the same strand as the reference genome to which the bam files are mapped, this will turn it into a triallelic site. Triallelic sites are discarded by default, therefore information from the sites that are expressed on the opposite strand will be lost. It is recommended that the reference dataset is expressed in terms of the reference genome.

EXAMPLE: Plot two BAM files against a VCF file.

bammds Polynesian.Pl24.cleaned.bam Polynesian.Pl25.cleaned.bam HGDP_hg19.vcf

ENVIRONMENT VARIABLES

$TMPDIR

Directory for temporary files. **bammds** will store temporary files while running in this directory. These files are only used internally. It defaults to **/tmp** but if **/tmp** is space constrained, it can be changed like this:

```
export TMPDIR=/scratch
```

EXIT STATUS

bammds will exit with non-zero status if it discovers anything is wrong.

REPORTING BUGS

Report bugs to <XX@gmail.com>.

AUTHOR

When using **bammds** for a publication please cite:

<<INSERT ARTICLE HERE>>

LICENSE

Documentation license I

Documentation license II

- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Notice

For any reuse or distribution, you must make clear to others the license terms of this work.

A copy of the full license is included in the file as cc-by-sa.txt.

DEPENDENCIES

bammds uses Perl, R (the modules data.table and gridExtra will automatically be downloaded), GNU Parallel (20130722 or later which is included), samtools (for processing bam files), vcftools (for processing vcf files), p-link (for processing bed and ped files).

Compilation dependencies: C++ compiler, pod2man, fatpack + installation of Perl-modules with .packlist files.

SEE ALSO

samtools(1), **vcftools**(1), **R**(1), **p-link**(1)