MISTARtools: a suite of utilities for the management of allele frequency information

Gabriel Renaud
gabriel [dot] reno [at] gmail.com

Abstract

MISTARtools is a set of command-line utilities to store allele count data, merge and split dataset, compute summary statistics and exporting data to various formats.

Contents

1 Introduction

MISTARtools aims at storing allele frequency information for a given individual or population from various input sources (VCF, BAM etc), creating intersections, querying the data and exporting it to various formats used by current software tools.

Program	Use
23andme2mistar	Create a mistartools file from 23andme file
affyVCF2mistar	Create a mistartools file from Affymetrix microarray data
axt2mistar	Create a mistartools file from axt files representing multiple sequence alignement
bam2mistar	Create a mistartools file from a BAM file
bamtable2mistar	Create a mistartools file from a BAM table file
closestMistar	Compute the distance to the nearest neighboring site for each site
epo2mistar	Create a mistartools file from a EPO alignment
generateCoords	Generate coordinate (random or windows) for a given reference genome
mistar2AlleleMatrix	Export a mistartools file to an allele matrix file
mistar2bed	Produce a BED file from the sites in mistartools file
mistar2binary	Export a mistartools file to an allele matrixtools file of 0=ancestral,1=derived
mistar2BinaryPLINK	Export a mistartools file to the PLINK format
mistar2EIGENSTRAT	Export a mistartools file to the EIGENSTRAT format
mistar2fasta	Produce a fasta file from a mistartools file
mistar2nexus	Produce a nexus file from a mistartools file
mistar2segtor	Produce a Segtor file from a mistartools file
mistar2treemix	Export a mistartools file to the treemix format
mistarcat	Concatenate multiple mistartools files
mistarcompress	Compress a mistartools file to a binary one
mistarcompute	Compute summary statistics like pairwise differences and D-statistics
mistardecompress	Decompress a binary mistartools file
mistarfilter	Filter a mistartools file according to certain criterias
mistarfreqSpec	Compute the frequency of alleles
mistarintersect	Intersect multiple mistartools files
mistarmeld	Merge two populations together
mistarRenamePop	Change the name of a population
mistarstats	How many sites are there in that file
mistarunion	Unite multiple mistartools files
mistaruniq	Like mistaruniq but returns only the unique lines
mpileup2mistar	Convert samtools mpileup to mistartools
ms2mistar	Convert ms output to mistartools
ms2nj	Create a neighbor-joining tree from a mistartools file
vcf2mistar	Create a mistartools file from VCF file
vcfcompute	Compute summary statistics directly from VCF files
vcfMulti2mistar	Create a mistartools file from VCF file with multiple individuals

2 File format

A mistartools file is composed of a header and the allele frequency. Each line in the header starts with a # and has the following format:

```
#MISTAR
#PG:[program line]
#GITVERSION: [github revision]
```

```
#DATE: YYYY-MM-DD
#chr    coord   REF,ALT root    anc     pop1     pop2     ...
```

The remaining lines have the following format:

```
chr    coord   REF,ALT [root info]     [anc info]     [pop1 info]     [pop2 info]     ..
```

The ALT allele is set to N if there were no alternative allele found. The info has the following format:

```
[REF allele count],[ALT allele count]:[CPG flag 0=no,1=yes]
```

Here is an example of a line:

```
1 1689574 A,N 1,0:0 1,0:0 1,0:0
1 1689575 T,C 0,1:0 1,0:0 1,0:0
```

2.1 Tabix indexing

If the file was zipped using bgzip, the following command can be used to index it:

```
tabix -s 1 -b 2 -e 2 [mistar file]
```

3 Importing data

This program produces a mistar matrix given a BAM file:

3.1 From BAM

```
bam2mistar <options> [name sample] [fasta file] [bam file]  [EPO alignment file]
```

3.2 From VCF

This program convert VCF files into mistar (prints to the stdout):

```
vcf2mistar <options> [vcf file] [name sample] [EPO alignment file]
```

If it is a vcffile with multiple individuals, use:

```
vcfMulti2mistar <options> [vcf file]
```

3.3 From 23andme

This program convert 23andme files into mistar (prints to the stdout)

```
23andme2mistar <options> [23andme file] [name sample] [EPO alignment file]
```

3.4 From AXT alignment

This program will parse an axt alignment and print a mistar file

```
axt2mistar [chr name] [name sample]  [axt file]
```

3.5 From ms simulations

This program converts ms output into a mistar matrix

```
ms2mistar  [mistar file] [correspondance individuals to pop] [size of chromosome]
```

The correspondence has to have the following format

```
pop1:individual1,individual2-npop2:individual3,individual4
```

The size of the chromsome is the parameter used as -r

4 Transforming data

4.1 cat

This program concatenates many files where the header is found in the first file and does not use the headers from the remaining ones It prints to the /dev/stdout

```
mistarcat [mistar file#1] [mistar file#2]
```

4.2 filtering

This program filters a mistartools matrix given certain criterias.

```
mistarfilter [mode]
```

The mode can be one of the following:

mode	use
noundef	No undefined sites for populations
bedfilter	Filter mistartools file using sorted bedfile
segsite	Just retain segregating sites (or trans./transi)
popsub	Keep a subset of the populations
removepop	Remove a subset of the populations
sharing	Retain sites that share alleles between populations
nosharing	Retain sites that do not share alleles between populations
znosharing	Retain sites that strickly do not share alleles between populations

Here is a description of the different filter modes:

4.2.1 noundef

This will filter out any site where the allele count is nul (0,0) for both reference and alternative

```
mistarfilter noundef [mistar file]
```

4.2.2 bedfilter

This will keep only the positions in the bed file

```
mistarfilter noundef [mistar file] [sorted bed file]
```

4.2.3 segsite

This will retain sites where the allele count is greater than 0 for either the reference or alternative for at least one individual. It has options to retain only transitions or transversions.

```
mistarfilter  segsite [options] [mistar file]
```

4.2.4 popsub

This will keep only the population specified in the list. Please note that it will set the alternative allele to 'N' if no population has the alternative allele

```
mistarfilter   popsub [mistar file] [comma separated group to keep]
```

4.2.5 removepop

This will remove the population specified in the list.Please note that it will set the alternative allele to 'N' if no population has the alternative allele

```
mistarfilter removepop [mistar file] [comma separated group to remove]
```

4.2.6 sharing

This will only retain sites where every individuals in population group 1 share the same allele(s) as every individual in population group 2. It requires that the allele count for every individual for both groups be non-zero. A random allele is picked (biased for allele count) for heterozygous position so do not be surprised if you get different outputs every time.

```
mistarfilter  sharing [mistar file] [comma separated group 1] [comma separated group 2]
```

4.2.7 nosharing

This will filter sites where individuals in population group 1 do not share at least one allele with individual in population group 2. It requires that the allele count for every individual for both groups be non-zero. A random allele is picked (biased for allele count) for heterozygous position so do not be surprised if you get different outputs every time. In other words, the individuals in the first group have to be all reference and the second all alternative or vice-versa.

```
mistarfilter nosharing [mistar file] [comma separated group 1] [comma separated group 2]
```

4.2.8 znosharing

This will filter sites where individuals in population group 1 strickly do not share any allele with individual in population group 2. It requires that the allele count for every individual for both groups be non-zero. Please remember that this will exclude any hetezygous sites

```
mistarfilter  [mistar file] [comma separated group 1] [comma separated group 2]
```

4.3 intersect

This program will print the intersection of the mistar files to stdout, it will skip triallelic sites.

```
mistarintersect [mistar file 1] [mistar file 2] ...
```

4.4 Meld two populations

This program will merge different specified populations into a single one. You

```
mistarmeld  <options> [mistar file zipped] "popToMerge1,popToMerge2,.." "newid"
```

Example of usage:

```
mistarmeld data.mst.gz "Papuan,Austalian" "oceanians"
```

4.5 Rename populations

This program will rename different specified populations.

```
mistarRenamePop <options> [mistar file] "popOldName1,popOldName2,..." "popNewName1,popNewName2,..."
```

Example of usage:

```
mistarRenamePop data.mst "Papuan,Austalian" "Oceanians1,Oceanians2"
```

4.6 union

This program will print the union of the mistar files to stdout, it will skip triallelic sites.

```
mistarunion <options> [mistar file 1] [mistar file 2] ...
```

4.7 Unique of union

This program will print the unique union of the mistar files to stdout. Used to merge the same files from different filters.

```
mistaruniq [mistar file 1] [mistar file 2] ...
```

4.8 replace ancestor

This program will print the first mistar file but with the ancestral information from the second one to stdout.

```
replaceAncestor [mistar file 1] [mistar file 2]
```

4.9 use population as root and ancestor

This program will use specified populations as root and ancestor and produce lines with only those two populations.

```
usePopAsRootAnc <options> [mistar file] "poproot" "popanc"
```

5 Statistics

5.1 Closest sites

This program will print to stdout the distance to the closest site for each record.

```
closestMistar <options> [mistar file]
```

5.2 Site frequency spectrum

This program will print the number of observed alleles for the reference and alternative alleles.

```
mistarfreqSpec <options> [mistar file]
```

5.3 Summary statistics

This program takes a mistar matrix and prints some how many lines it has.

```
mistarstats  <options> [mistar file]
```

5.4 Relationships between individuals

5.4.1 Neighbor joining tree

Compute a neighbor-joining tree using the mistar file.

```
mistarcompute nj  <options> [mistar file]
```

5.4.2 Pairwise average coalescence

To compute pairwise average coalescence.

```
mistarcompute paircoacompute  <options> [mistar file]
```

5.4.3 Pairwise nucleotide differences

To compute pairwise nucleotide differences.

```
mistarcompute pairdiff <options> [mistar file]
```

5.4.4 D-statistics

To compute triple-wise D-statistics

```
mistarcompute dstat <options> [mistar file]
```

6 Exporting data

6.1 To treemix

To print treemix input.

```
mistar2treemix <options> [mistar file]
```

6.2 To allele matrix

This program produces a matrix where each record for population becomes a single allele A,C,G,T.

```
mistar2AlleleMatrix  [mistar file]
```

6.3 To binary allele matrix

This program takes a mistar matrix and prints the alleles as a binary matrix (0=ancestral,1=derived)

```
mistar2binary <options> [mistar file] [comma separated group]
```

6.4 To binary PLINK

This program takes a mistar matrix and prints the genotype and SNP file in PLINK format

```
mistar2BinaryPLINK <options> [mistar file] [out (.bed)] [out (.bim)] [out SNP file (.fam)]
```

6.5 To EIGENSTRAT

This program takes a mistar matrix and exports the data in EIGENSTRAT.

```
mistar2EIGENSTRAT <options> [mistar file] [out genotype file (.geno)] [out SNP file (.snp)] [out SNP file (.ind)]
```

6.6 To fasta

This program takes a mistar matrix and exports the data in EIGENSTRAT.

```
mistar2EIGENSTRAT <options> [mistar file] [out genotype file (.geno)] [out SNP file (.snp)] [out SNP file (.ind)]
```

6.7 To fasta

This program takes a mistar matrix and prints a FASTA file using the allele information with one record per population. Each site generates one base pair.

```
mistar2fasta <options> [mistar file]
```

6.8 To NEXUS

This program takes a mistar matrix and prints the alleles in nexus format.

```
mistar2nexus <options> [mistar file]
```