Find SNPs distinguish two populations

Jinliang Yang

Jan. 28th, 2015

INTRODUCTION

Develope a piece of software to identify a set of SNPs that distinguish two populations.

INSTALL and USAGE

```
git clone https://github.com/yangjl/zmSNPtools.git
cd zmSNPtools
cp packages/fpSNP/fpSNP_v0.2.py bin/fpSNP
chmod +x bin/fpSNP
### make sure the bin/ is in your searching path
export PATH=$PATH:~/bin/
fpSNP -h
```

input data

The input SNP data should be in BED-like format. The first three columns are **snpid**, **chr** and **pos**. The following columns in the first row should be any number of individuals with SNP genotype information. In the second row, the first three columns should be "N" and the followings are two levels of intergers (such as 1 and 2) to specify the populations they belonging to. See the example below.

```
dsf <- read.table("../data/simusnps.txt", nrow = 5,
    header = TRUE)
head(dsf[, 1:7])
```

Run the following command to get the population frequencies:

```
cd largedata/assignprb
fpSNP -i usgbs_tot50k_5619.txt -o usgbs_tot50k_5619.snpfrq
```

test the correctness of the frequency calculations

The following R codes do the same population specific allele frequency calculation. But it is less efficient.

```
### simulate 1000 SNP data
source("../profiling/1.A.1.simulate_SNP_data.R")
### results computed from R were written to
### tests/routput.txt
source("../profiling/1.A.2.R_fpSNP_test.R")
```

After comparisons, only one SNP got different minor allele frequency values in population one and population two. While, this SNP has a overall MAF of 0.5, the minor allele was determined randomly by `Python` and `R`. So it is not a problem for our calculation.

FINDSNPs to distinguish two populations

With the computed allele frequencies in the two populations, we will be able to distinguish the two populations by select the most different SNPs in terms of their allele frequecies. To do this, a statistical test was conducted. The null hypothesis for this test is that for a given set of independent SNPs in two identical populations the allele frequencies are the same. If the possibility that the observed allele frequency differences for a given set of SNPs is smaller than a threshould (say pvalue< 0.05), we will have evidence to reject the null hypothesis and claim the two populations are different.

We used the `paired t-test` to conduct the analysis. To control the independence of the SNPs, we used a selected bin window (for example 1-Mb) to remove SNPs that might be in LD.

```
### simulate 1000 SNP data
source("../lib/getfpsnp.R")
test <- findSNP(frqfile = "../tests/output.txt",
    binsize = 100, pcutoff = 0.001, missingrate = 0.5)
head(test)
```