offPVP Diallel: using GERP scores in IBD region for phenotypic prediction

Jinliang Yang

Jan. 9th, 2015

SUMMARY

In this study, we used summary statistics of GERP scores in the IBD regions to fit the genomic selection model for phenotypic prediction. Four summary statistics were used for each IBD block, they are 1) number of conserved SNPs in additive mode, 2) sum of the conservation scores in additive mode, 3) number of conserved SNPs in dominant mode and 4) sum of the conservation scores (GERP) in dominant mode.

First, a python script `gerpIBD.py` was developed to compute the conservation statistics of each IBD block across a diallel population. The scores were normalized across samples. The script was uploaded to the repo of `zmSNPtools`. Use `-help` for more detail of the program.

Second, gerp scores were circular shuffled and were used to generate N set of random assigned conservation statistics for each IBD block.

Third, seven phenotypic traits were trained with the conservation statistics in IBD block as genotype. The results were compared with circular shuffled genotypes. The results showed the real data fit always better than the shuffled genotypes.

Finally, we trained the GS model with subset of the phenotypic data and predict the validation set using a 5-fold cross-validation strategy.

---

*Data preparation for **gerpIBD***

SNPs were mapped to the IBD block using **bedtools** as below.

```
### load the GERP element data
source("../profiling/4.IBD/5.A.1_snp2IBD_block.R")
```

Two other files for **gerpIBD** contain information of GERP scores and SNP genotype for founder lines

*Run **gerpIBD** for real and circular shuffled data*

A **gerpIBD** wrapper was created in order to simplify the procedure to run the python code in **farm**. Basically, we just run the below line:

```
  gerpIBD -d largedata/IBD/allsnps_11m_IBD.bed -s largedata/SNP/allsnps_11m.dsf5
\\ -g largedata/SNP/allsnps_11m_gerpv2_cs1.csv -o largedata/SNP/gerpIBD_output"

### run gerpIBD for real data
source("../profiling/4.IBD/5.A.2_run_gerpIBD.R")
### run gerpIBD for circular shuffled data:
source("../profiling/4.IBD/5.A.3_cirshuffling.R")
source("../profiling/4.IBD/5.A.4_run_gerpIBD_cs.R")
```

A total of `29,869,451` conserved elements were identified, which accounted for `1.4%` of the total genomic space.

From these conserved regions, we randomly selected 1,000 SNPs with GERP score >2 for genomic prediction. A random set of 1,000 SNPs were selected from unconserved genomic regions.