Introgression Analysis

Jinliang Yang

Jan. 5th, 2015

SUMMARY

Test whether Totontepec maize lines have admixture with *mexicana* using SNP50k data.

Get genetic and physical map of the SNP50k

I compared the SNP50k map from Matt's archive and the map downloaded from Illumina website. Matt's map is more inforamtive. It contains less SNPs (N=37,568) but more of them have physical information. However, the top/bot allele issue need to be first addressed.

```
### output top/bot allele with the chr and pos
### info pulled out from Matt's data.
source("../profiling/1.Introgression/1.A.1_snps_topbot.R")
```

Transform top/bot to forward strand

Use the perl code **top2ref** from **zmSNPtools** to do the trick. Here is how I did it:

```
### install the software
export PERL5LIB=$PERL5LIB:~/Documents/Github/zmSNPtools/modules
cp packages/top2ref/top2ref.pl bin/top2ref
chmod +x bin/top2ref

### use samtools to index reference genome in fasta format
sed -i -- 's/chromosome:AGPv2:.*chromosome /chr/g' ZmB73_RefGen_v2.fasta
samtools faidx ZmB73_RefGen_v2.fasta

### run the program to determine the strand first
cd /home/jolyang/Documents/Github/N2/largedata
cat snp50k_topbot.txt | top2ref -r ~/dbcenter/AGP/AGPv2/ZmB73_RefGen_v2.fasta > snp50k_topbot.map

### remove SNPs that could not be determined
source("../profiling/1.Introgression/1.A.2_snp_topbot_remove0.R")
# only 11 SNPs could not be determined, that is acceptable
```

```
### run the program again to flip the 37k SNPs
cat snp50k_topbot.txt | top2ref -m snp50k_topbot.map > snp50k_topbot.fwd
```

Translate top/bot to ref/alt

Assemble all the information together with teh **"top/bot"**, **"fwd-top/fwdbot"** and **"ref/alt"** alleles information. These could be used as the translation table, where top=>fwdtop, bot=>fwdbot.

```
### generate the translation table
source("../profiling/1.Introgression/1.A.3_snp_translation_table.R")
```

*Phasing parental reference populations using **fastPHASE***

Run the following `R` script to format the SNP50k data and the formatted data were used as the `fastPHASE`[1] input. The formatted data included:

1. Mexicana (N=120)
2. Parviglumis (N=130)
3. Landraces (N=94)

[1] Scheet, P., & Stephens, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data : Applications to Inferring Missing Genotypes and Haplotypic Phase, 78(April), 629–644.

```
source("profiling/1.Introgression/1.A.1_fastPHASE_input.R")
```

And then run `fastPHASE` to do the haplotype phasing with HMM. Basically, I ran the code, for example

`fastPHASE -oMex -S1234 mex_120.fp`,

to do the hap phasing. To run on `farm` with `slurm`, I prepared the slurm codes with an `R` script.

```
source("../profiling/1.Introgression/1.A.2_run_fastPHASE.R")
```

```
# run the following code on farm

cd /home/jolyang/Documents/Github/Introgression
sbatch -p bigmeml largedata/fphase/mex_fp_slurm.sh
sbatch -p bigmeml largedata/fphase/parv_fp_slurm.sh
sbatch -p bigmeml largedata/fphase/land_fp_slurm.sh
```