Manual Reference Pages - OutcrossSeq

OutcrossSeq is a highly scalable genotyping software for identifying genomic variation,genotyping, haplotype phasing and mapping agronomically important loci in the populations of outcrossing crops, based on whole-genome low-coverage resequencing data.

> *OutcrossSeq works with VCF format files as input. After imputation, the output file is a genotype file.*
> *"Diploid-Outcrossing" and "Double-Cross" imputation results are in bin matrix, take the bin matrix came from OutcrossSeq and plus a formated phenotype table, one can run GACD, MapQTL etc.*
> *“Autopolyploid Plant” imputation result is a SNP matrix, one can run fastGWA, EMMAX etc. using SNP matrix from OutcrossSeq and formated phenotype table.*
> *For more details, please check Demo run below.*

Availability

You can download the source code and the Demo data here:
www.xhhuanglab.cn/tool/OutcrossSeq.html
You can download the Docker container here:
Docker container

Dependencies

We have made OutcrossSeq into a Docker container, you can skip all installation steps (you should having Docker installed at first). In Docker container we have packaged all alignment and mapping softwares.

1. BWA (version 0.7.16)
2. SAMtools (1.5)
3. Picard (1.119)
4. GATK (3.4)
5. Perl （v5.16.3)
6. R (3.4.1)
7. Statistics::Basic perl modular
8. Plink (1.90b5)
9. GCTA (1.92.4)
10. GACD (1.1) :*Win2003,Win2008,WinXP,WinVista,Win7,Win8,Win10*

Used to generate input vcf files: BWA,SAMtools,Picard,GATK.
Used to run OutcrossSeq:Perl, R, Statistics::Basic.
Used for genetic mapping: Plink, GCTA, GACD.

In Docker container GATK should usw GetFullPath :

Software structure

There are three modules in OutcrossSeq.

- "Diploid-Outcrossing" module designed for diploid outcrossing F1 population
- "Double-Cross" module designed for double cross population
- "Autopolyploid Plant" module designed for autopolyploid F1 population.

Data preparation

The input file of "Diploid-Outcrossing" module is variant calling file of two heterozygous parents and all F1 individuals.Normally in VCF format.
See Demo data: `demo_diploid.vcf`

The input file of "Double-Cross" module is variant calling file of four inbred parents and all offspring individuals.

See Demo data:`demo_double.vcf`

The input files of "Autopolyploid Plant" module are two separated variant calling files of parents and offsprings, since for parents we call variants in polyploid mode, meanwhile, for offsprings we call variants in diploid mode.
See Demo data for two hexaploid parents: `demo-Parents.vcf` and their F1 offsprings `demo_autopolyploid.vcf`

- **For variant calling，we recommend to generate VCF files for each chromosome which is more convenient to run the code and to tune parameters in following steps.**

The pipeline for generating vcf is not described here. For details, please refer to the GATK.

Algorithm

Diploid-Outcrossing module

This module is designed for all diploid outcrossing species such as trees,fruits etc.
F1 population with about 200 individauls is good enough for a well study using this module.
It can also be applied to other types of populations when the founders are no more than two heterozygous lines.

1:Filter out low-quality SNP sites

High coverage sequencing of founders: The SNP sites need to be screened by the information of the parents, so the information of the parents must be accurate. Because homozygous SNPs in both parents led to fixed genotypes in the population (for example, all offspring individuals must be AT genotype with AA in the male parent and TT in the female parent, or must be TT genotype with TT in the male parent and TT in the female parent), only the SNP sites that were heterozygous in at least one parental line (e.g., TT and AT for male and female, respectively; or AT for both parents) were left in the following analyses, at the same time filter out the high missing rate and abnormal heterozygosity sites in the offspring (for example,offspring heterozygosity rate should be 0.5 with AT for male and female or with TT in the male parent and AT in the female parent,so we can choose the range between 0.3 and 0.7) (Supplemental note Figure 1).Moreover, based on the raw genotypic data of the population, all SNP sites with high missing data rate (>60%) or of low genotyping quality population < 300) were removed.(QUAL value in VCF of the

```
cd /OutcrossSeq/diploid_outcrossing_species
perl diploid_cross_Genotype_filter.pl 5 parent1 parent2 100 0.6 demo_diploid 0.3
```

Supplemental note Figure 1. Construction of outcrossing F1 population and F2 population. Four color rectangles represent four haplotypes, the genotypes at different positions show different expected genotype frequencies.

2:Calculate kinship of each sample in every windows

After the SNP filtering, consecutive SNPs were examined for each <1 cM window (<1 cM in genetic distances for many plants, freely adjusted by users for different species) across each chromosome. In diploid-Outcrossing F1 population, can be regarded as a recombinant of four haplotypes. The genotypes of per locus are from 1 to 4 (Supplemental note Figure 1).

3:Clustering in each window

The genotypes of per locus are from 1 to 4 (both parents: H1H2 and H3H4; four major groups: H1H3, H1H4, H2H3 and H2H4) were all clear after genotyping (Supplemental note Figure 2). In each window, the gametic homolog pairs (H1H3, H1H4, H2H3 or H2H4) were assigned for each individual in the four major groups. In some windows, a few individuals were out of the four major groups, probably due to their poor coverage at the local regions. Start cuttree threshold can be determine by the evolutionary tree, as example in the upplemental note Figure 2,you can set a threshold > 0.3. You can randomly select several windows on the chromosome to draw the evolutionary tree to determine the threshold.

Supplemental note Figure 2. Clustering results of a window. Four genotypes represented by four red rectangles

4:Haplotying across the chromosome

In each window, the gametic homolog pairs (H1H3, H1H4, H2H3 or H2H4) were assigned for each individual in the four major groups. But each window is clustered independently, the groups of different clusters in different windows will be inconsistent. Each group should be unified .

5:Generate the genotype file

For these individuals, missing data were imputed based on their genotypes in the flanking windows (e.g., H1H3 was called in this window for the individual when H1H3 was called for the individual in the prior window and H1H3 was called in the forward window), or their nearest kinship neighbors at the local region when the genotype in the prior window and that in the forward window were inconsistent. After that, the information of gametic haplotype combinations were collected across the whole genome, which could be directly used in high-resolution genetic mapping.

Double-Cross module

This module is designed for typical Double-Cross population in crop breeding programs, such as rice and mazie etc.
Population size with 500 individauls is sufficient to carry out a nice study.
It can also be applied to other types of populations when the founders are no more than four inbred lines.

1:Filter out low-quality SNP sites

The haplotype information of H1, H2, H3 and H4 were clear after genotype calling from deep sequencing data of four inbred lines (or haplotype phased by other experiments).Among them, type-1 SNPs (segregation ratio in four haplotypes = 1:3, e.g., A, T, A and A for H1, H2, H3 and H4, respectively) and type-2 SNPs (segregation ratio in four haplotypes = 2:2, e.g., A, T, A and T for H1, H2, H3 and H4, respectively) were selected (Supplemental Figure 8).

2:Determine the genotype of the first allele

Consecutive SNPs were examined for each 300 kb window, and the 300-kb genomic regions with <150 SNPs were combined with flanking 300-kb regions into a new window. For each individual in the window, all type-1 SNPs containing the minor allele (e.g., the allele T in A, T, A and A was the minor one, representing H2 type) were used to evaluate the genotype. A type-1 score matrix was generated using the information from all the minor-alleles, and haplotypes were called based on the highest probability among four haplotypes (e.g., H2 type would be called for an individual, in which minor alleles of >95% SNPs pointed to H2 type) (Supplemental Figure 8).

3:Determine the genotype of the second allele

Similarly, all type-2 SNPs were used to infer the other haplotype in the individual, based on the highest probability of the type-2 score matrix as well. For example, when most type-2 SNPs point to H3 type for the individual mentioned above, the gametic haplotype combination would be clear (that is, H2H3) (Supplemental Figure 8).

4:Generate the genotype file

In some windows, a few individuals contained too many missing data, leading to the failure in the two genotype calling steps (type-1 and type-2). For these individuals, missing data were imputed based on flanking windows (e.g., H2H3 was called when H2H3 was called for all flanking windows of the individual). After the processes were repeated for the whole genome, a map of gametic haplotype combinations was generated for the population (Supplemental Figure 8).

Supplemental Figure 8. The schematic diagram of OutcrossSeq for double cross population or outcrossing species with well-phased haplotype information. The haplotypes (H1, H2, H3 and H4) are indicated by different colors. Type-1 loci are the SNPs with segregation ratio in four haplotypes of 1:3 while type-2 loci are those with segregation ratio in four haplotypes of 2:2. Type-1 loci and type-2 loci are used in two consecutive steps for genotyping, respectively.

Autopolyploid Plant module

This module is designed dedicated for F1 population of autopolyploid or near-autopolyploid plants such as potato and sweet potato etc.
A F1 Population with about 1000 or more individauls is recommended, for a reasonable practice.
It might be applied to other types of large populations but which is not typical in autopolyploid plants.

1:Selecet simplex or double-simplex SNPs

Filtered for simplex (single-dose variant present in only one parent, e.g., ATTTTT × TTTTTT) and double-simplex SNPs (single-dose variant present in both parents, e.g., ATTTTT × ATTTTT).

2:Transform .vcf to .geno file

It's just a format conversion, converting the genotypes determined in vcf into a format that is easier to understand.

3:Filter low-quality and error ratio sites

This step can be skipped
The proportion of different genotypes in the population was calculated for each SNP. According to expected ratios for simplex (e.g., “A”:“T”=1:11 in ATTTTT × TTTTTT) and double-simplex SNPs (e.g., “A”:“T”=2:10 in ATTTTT × ATTTTT) along with the bias due to skim sequencing, SNPs with the proportion of heterozygous genotypes (0/1) in the population between 0.04 to 0.28 were left for following analyses. Moreover, SNP sites with >60% missing calls were excluded for imputation.

4:Select relevant sites

Consecutive SNPs (filtered after the above steps) were examined in a window of 500 SNPs in size across each chromosome in the reference genome of the sextuploid. In each window, the Pearson correlation coefficients between any two SNPs were calculated based on the unimputed genotypes using Statistics::Basic in Perl, which led to a 500×500 relationship matrix to infer the linkage disequilibrium patterns in polyploidy genome. The matrix was further used in clustering analysis using the method of unweighted pair-group method with arithmetic means in Statistics::Basic of Perl. With the cutoff of 0.9, the clustering analysis enabled the partition of 500 SNPs into multiple groups, and the groups with the SNP number of <10 were removed.

5:Select mark readsNum

Our data is in low coverage, so there will be many missing and wrong genotype loci, extract the known sequencing data to prepare for imputation

6:Imputation missing SNPs

Each of the resulting SNP groups was actually the collection of highly linked SNP markers with their minor alleles located in the same homologous chromosome, keeping nearly the same paces in recombination. Missing data were imputed by taking three variables (the variables were diverse among different SNP groups) into consideration: the SNP number of the group, the missing data rate and the genotypic proportion in the population, based on which a total of 40 circumstances were divided for imputations. Taking one circumstance as an example, when the SNP number was >80, the missing data rate was <80% for the group and the heterozygous genotype proportion = 0%, all the genotypes would be imputed to be homozygous (e.g., AAAAAA). After the processes were completed and the window slided across the whole genome, a fine-scale genotype map was generated to enable high-resolution genetic mapping.

7:Merge genotype files

The above steps are performed by chromosomes, so the last step needs to merge all the genotype data together