RFAPtools.sh v2.0 Manual

1. Introduction

Following the development of protocols for reduced representation library (RRL) construction that employing RE digestion, only a few software tools have been designed to detect high-quality SNPs. And available software packages could not efficiently discriminate allelic SNPs in the presence of homoeologous or paralogous sequences, especially in natural populations. RFAPtools_v2.0 was designed in particular to detect SNPs in segregating and natural populations of polyploid species, where it would be able to detect allelic SNPs unconstrained by the presence of homoeologous or paralogous sequences. RFAPtools_v2.0 would appear to contribute greatly to SNP discovery and genotyping for linkage map construction and genome association studies. Hence RFAPtools_v2.0 with fewer false-positive SNPs is an effective tool for SNP calling in the absence of a reference genome sequence, and particularly valuable for polyploid crops.

2. Installation

RFAPtools_v2.0 depends on the SOAP2 software, so you need to pre-install it firstly. And we integrated three scripts prf_building.sh, prf_snpcalling.sh and prf_allele.sh/prf_allele_natural.sh into RFAPtools_v2.0.sh as a whole.

RFAPtools_v2.0.sh is one shell script, hence simply unpackage RFAPtools_v2.0.gz, and move the folders of "scripts" and "samples", and RFAPtools_v2.0.sh file to your work folder. After that type the command:

```
$ chmod a+x RFAPtools_v2.0.sh
```

3. Usage

1) Firstly you need to move the sequence data of each sample under the folder (you also could find the test sequencing data "test_data_segregation.tar.gz" or "test_data_natural.tar.gz" from the website). You also need to change the name of the sequencedata of the samples. For

segregation population: You need to name them as p1-F.fastq, p1-R.fastq (if pared-end), p2-F.fastq and p2-R.fastq (if pared-end), for each segregation individual, you need to name them as 1-F.fastq, 1-R.fastq (if pared-end), 2-F.fastq, 2-R.fastq (if pared-end) and et al...; For natural population: You just need to name them as 1-F.fastq, 1-R.fastq (if pared-end), 2-F.fastq, 2-R.fastq (if pared-end) and et al...

2) Then you could justify some parameters use the command of:

```
$ vi RFAPtools_v2.0.sh
```

3) You need to edit these parameters like:

(1) population : segregation or natural population, default: segregation;

(2) first_strand : The first strand of sample used to construct pseudo-reference, default: p1-F.fastq;

(3) second_strand : The second strand of sample used to construct pseudo-reference, default: p1-R.fastq;

(4) output_file : The name of the output file for pseudo-reference, default: p1_tag;

(5) sequence_type : Mean that the type of reads like "single" or "pair", default: pair;

(6) read_count : The tags contained more than that are selected to build pseudo reference sequence, and the selection of this value depend on the volume of sequencing data of parent which you select to build prf like p1, default :2;

(7) sample_number : Please specify the total number of samples of segregation population, default: 40;

(8) output : The name of allele_calling.pl output file, default: SNP.out;

(9) else : The file contained dump genotypes, default: else;

(10) minor_genotype : The minor genotype number of the locus should be higher than this value, or this locus would be dumped, default: 8;

(11) loss_genotype : The loss genotype number of the locus should be less than this value, or this locus would be dumped, default: 8;

(12) hr_relative : Relative heterozygous ratio which express as (heterozygous

genotype/minor genotype), the value of each allelic SNP locus should be lower than it, or this locus would be dumped, default: 0.2;

(13) allele_number : The number of candidate alleles belong to one candidate locus that more than that would be filtered, This parameter is only for natural population, default: 20;

(14) read_length : The length of read sequence, default: 70;

4) After that, you just need to type the command below, and you would get the SNP results:

```
$ ./RFAPtools_v2.0.sh
```

4. Output

The output format of **segregation population** is listed:

1) The start position of each locus on the pseudo-reference sequence;
2) Nucleotide variations position and type;
3) SNP numbers of the tag as the same as p1, and p1 tag is set as “A” genotype;
4) Map number of the locus as the same as p1, when aligning to pseudo-reference sequence;
5) The coverage of “A” genotype of p1, p2 and segregation population;
6) SNP number of the tag as the same as p2, and p2 tag is set as “B” genotype;
7) Map number of the locus as the same as p2, when aligning to pseudo-reference sequence;
8) The coverage of “B” genotype of p1, p2 and segregation population;
9) The number of “A” genotype in segregation population;
10) The number of “B” genotype in segregation population;
11) The number of heterozygous genotypes
12) The number of loss genotypes
13) The value of relative heterozygous rate
14) p1 raw genotype;
15) p2 raw genotype
16) The genotype of sample 1;
17) The genotype of sample 2;

…

n) The genotype of sample n;

The output format of **natural population** is listed:

1) The start position of each sequence tag locus on the pseudo-reference sequence;
2) The number of nucleotide variations on this sequence tag locus;
3) The total number of genotypes per locus;
4) Nucleotide variations position and type;
5) The genotype of sample 1;
6) The genotype of sample 2;

…

7) The genotype of sample n;