Do genotype imputation using LinkImputeR

Last year, I wrote a blog about constructing genetic linkage map with SNPs called from GBS data of the Rabiosa x Sikem F1 population, Constructing genetic linkage map. Before the map construction, in order to recover some missing SNPs for some samples, I did imputation using a tool called LinkImputeR, https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3873-5

You could find the user guide here, UserGuide.pdf

Installing the software is quite straightforward, just follow the tutorial in the user guide, download the package and decompress it, then it's done.

From here you could download the package, http://www.cultivatingdiversity.org/software.html

Then decompress the package by this command:

tar -xvzf LinkImputeR.tar.gz

The input file for LinkImputeR is a standard vcf file containing SNPs, however, you must be aware that some fileds in the vcf file are required: AD, DP and GT. If you use BCFtools for variant calling, GT will be shown in the vcf file by default, but AD and DP are not included, therefore, you have to give explicit option to the command to have AD and DP in your vcf file. For example:

bcftools mpileup -a AD,DP -I -A -f {input.ref} {input.bam} | bcftools call -mv > {output}

In the above command, -a AD,DP will make sure to have AD and DP in the vcf file, -I means skip indel calls and -A means call SNPs from orphan read pairs, which I understand as reads don't have proper mate mapped. -I is not necessary if you want to include indels, but I think -A is very helpful for us as our GBS reads are only single end. Some explanations about -A can be found here, https://www.biostars.org/p/91183/

So, once you have your vcf file ready, you could start imputation.

First, make a new working directory, then copy your vcf file and the control file provided by LinkImputeR to the new folder. Edit the control file and then run the command. The control file is called accuracy.ini and you could find an original one in the folder where you have your LinkImputeR installed. Here's what inside the control file:

```
[Input]
filename = ./rabiosa_05_final.vcf
save = ./filtered.vcf

[InputFilters]


[Global]
depth = 2,4,6,8

[CaseFilters]
mincalled = 1
```

```
[Stats]
root = ./
level = table
eachmasked = yes

[Output]
control = ./impute.xml

[Log]
file = log.txt
level = debug

[Accuracy]
numbermasked = 5000
```

Change parameters accordingly. For more details of each parameter, please refer to the manual and paper. You could notice that it is blank under InputFilters, this is because I deleted options for prefiltering as I have filtered my vcf file after SNP calling. The depth option under Global header could allow you to impute based at different read depth and options under CaseFilters could allow you to test different minor allele frequency and missing value.

Now, after editing the control file, we could run the first step, which would give you an impression which combination of parameters could output the best accuracy for imputation, just run this:

mypath=/home/yutachen/public/Yutangchen/impurte

java -jar $mypath/LinkImputeR.jar -s accuracy.ini

After some while, you will have a table called sum.dat showing some statistics, for example:

Name	Samples	Positions	Accuracy	Correlation	Filters	Additional
Case 1	286	49101	0.8442	0.7773	PositionMinCall(1),SampleMinCall(1)	Depth(2)
Case 2	286	49101	0.8870	0.8340	PositionMinCall(1),SampleMinCall(1)	Depth(4)
Case 3	286	49101	0.9040	0.8568	PositionMinCall(1),SampleMinCall(1)	Depth(6)
Case 4	286	49101	0.9232	0.8705	PositionMinCall(1),SampleMinCall(1)	Depth(8)

You could see that Case 4 with read depth 8 has the best accuracy, so then continue with Case 4's setting for imputation, run this:

java -jar $mypath/LinkImputeR.jar impute.xml 'Case 4' output_05_8.vcf

impute.xml is the product of the first step and the imputed vcf will be output_05_8.vcf (you can give any names you like). So, here we have finished the whole imputation process with LinkImputeR and now you could use the imputed vcf file for linkage map construction.