ASSIsT: An Automatic SNP ScorIng Tool for in and out-breeding species Reference Manual

Di Guardo M, Micheletti D, Bianco L, Koehorst-van Putten HJJ,
Longhi S, Costa F, Aranzana MJ, Velasco R, Arús P,
Troggio M, van de Weg WE

Version 1.02
April 28, 2017

Contents

ASSIsT is a tool for efficient filtering of Illumina Infinium/BeadExpress based SNP markers. This software can analyse different types of experimental populations: Cross-pollinated (CP – F1), Back Cross (BC), F2 and collections of unrelated individuals (Germplasm). It is possible to export the filtered data in several formats according to the most widely used software for marker-trait association analysis.

1 Getting started

ASSIsT is written in Python; therefore, it can run virtually in any platform with python installed.

1.1 Availability

Source code and Windows executables (built using pyinstaller) are available for download at:

- http://compbiotoolbox.fmach.it/assist
- http://bioinformatics.tecnoparco.org/fruitbreedomics/assist-tool

When using ASSIsT, please cite: Di Guardo and Micheletti et al. 2015, referenced as: Di Guardo M, Micheletti D, Bianco L, Koehorst-van Putten HJJ, Longhi S, Costa F, Aranzana MJ, Velasco R, Arús P, Troggio M, van de Weg EW (XXXX) ASSIsT: An Automatic SNP ScorIng Tool for in- and out-breeding species. Bioinformatics, DOI: 10.1093/bioinformatics/btv446

1.2 What's new in version 1.02

- Fixed the shift of two individuals in the gtypes.csv output file with Germplasm population type.
- Added the ability to deal with crosses derived from self-pollination.

1.3 Installing ASSIsT

The Windows executables is distributed as a zip archive. It not necessary to install ASSIsT, just extract the ASSIsT_Windows.xx.zip archive (xx is the version number).

The source code is a collection of Python scripts. They can be executed from any operating system with Python 2.7 installed. The following additional Python modules are needed to run the software:

- PyQt4 (v.4.8 or higher)
- NumPy (v.1.8 or higher)

- matplotlib (v.1.3 or higher)
- SciPy (v.0.14 or higher)

1.4 Running ASSIsT

In Windows, double click on `ASSIsT.exe` to start ASSIsT. To run ASSIsT from the source code execute `ASSIsT.py` from a command line shell.

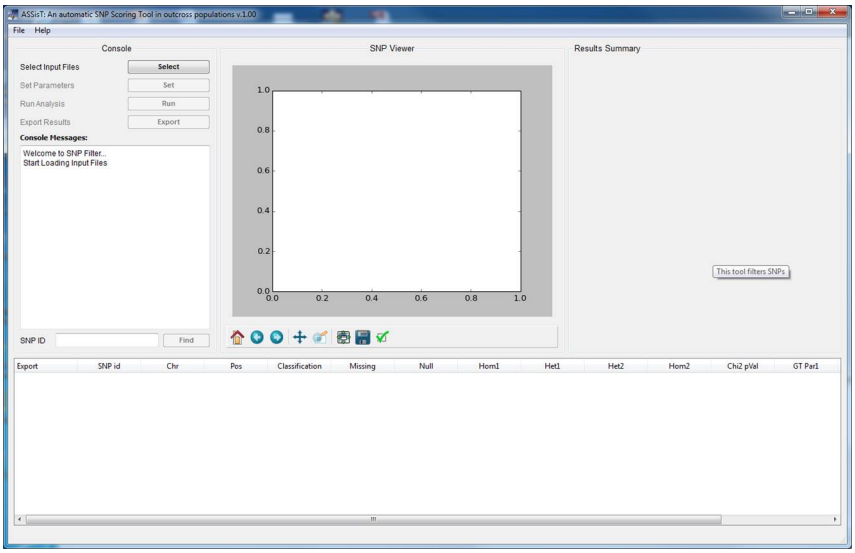


Figure 1: ASSIsT layout.

The analysis start by loading the four input files (GenomeStudio® Final Report, Genome Studio® DNA Report, pedigree, and map file) using the **Select** button. Example data files are available at the previously mentioned web-pages. Example data are provided for Cross Pollinators (CP) and F2-germplasm (F2 for inbreeding crops). Please note that the map file is not mandatory. To load the files click on ... and select the appropriate input file or enter the file name (with the full path) in the text box.

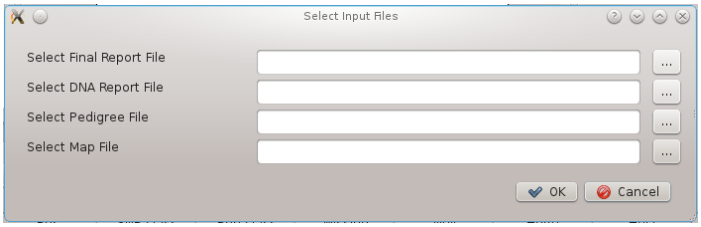


Figure 2: Menu opened by clicking on the Select button

After the Input files are correctly imported, it is necessary to set up the filtering

parameters by clicking on the **Set** button. The first parameter to set is the population type ("CP (F1)", BC, F2 or Germplasm) and then the related statistical and germplasm parameters (see below for details).

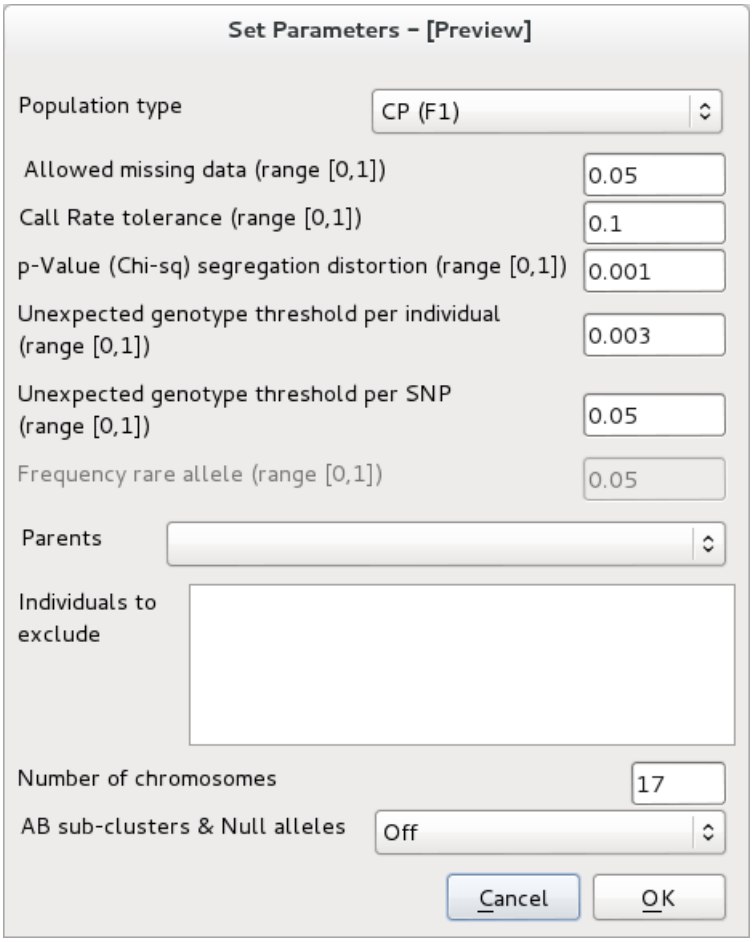


Figure 3: Dialog to select analysis parameter.

When this step is completed, the **Run** button becomes available, and the analysis can be performed by clicking on it.

At the end of the analysis, it is possible to choose the files to export by clicking **Export** Some outputs provide more detailed information on the performance of the filtering analysis itself, e.g., "Summary", "Custom gtypes", "Custom SNP information table" and "Custom Mendel error report".

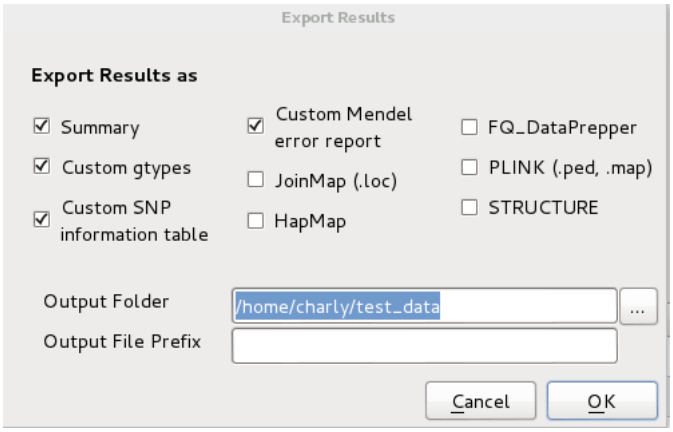


Figure 4: Dialog to select the files to Export.

Additionally, it is possible to export the results in additional formats (JoinMap®, PLINK, HapMap, FlexQTL™DataPrepper and STRUCTURE) that can be used as inputs to third-party programmes.

Note: Through the export section, a customized prefix can be added to the names of the output files.

2 Input files

Sample and marker names have to be consistent in all 4 input files.

Genome Studio Final Report: Using the Report Wizard (Open **GenomeStudio®** → **Analysis** → **Reports** → **Report wizard**), select **Final Report**, and press **Next**. On the following page, use the **"redo with the best 10th Percentile GC Score"** option, and press **Next**. If some samples have been excluded from the GenomeStudio® project you need to remove or zero-out your sample in the report. Make your choice according to how you want to account the excluded data and press **Next** (if no samples have been excluded from GenomeStudio® project this page is not displayed). The format of the Final Report must be set to "Standard" (the other choices are "Matrix" or "3rd Party"). The Final Report should include the following columns (in the specified order):

1. SNP Name
2. Sample ID
3. Allele1 - Top (or Allele1-AB or Allele1-design, depending on the desired type of output)
4. Allele2 - Top (or Allele2-AB or Allele2-design, has to be in line with the choice for Allele1)
5. GC Score

6. GT Score
7. Theta
8. R

Select **Group by SNP** (and not by "sample"). In **General Option**, select "Tab" as the field delimiter. Press **Next**, and select the folder in which the file has to be stored, and enter its name. Press **Finish**.

DNA Report: Using the Genome Studio Report Wizard (see the above section on the GenomeStudio® Final Report) select **DNA Report**. Press **Next** once or twice (twice if there are excluded samples in the GenomeStudio® project). Use the **redo with the best 10th Percentile GC Score** option and press **Next** once or twice. Finally select **SampleID**.

Pedigree: The pedigree file is composed of 3 columns: The first contains the list of individuals to analyse, while in the second and third columns, the female and male parents are reported. Be aware that this order (sample, mother, father) is important for some of the output files and that the only compulsory column is the first. In case of populations derived from a self pollination write both in second and third column the name of the selfed parent. Tabulation ("tab") must be used as the field delimiter. The file must also include the following header row:

//SampleID Mother Father

Sample names should not include white spaces (blanks) or special characters (non-ASCII symbols, http://it.wikipedia.org/wiki/ASCII).

Map file (Optional): The map file specifies the physical or genetic coordinates of the Genotyped SNPs. This file has to include three columns: the SNPid, the chromosome, and the position. The position can be expressed in base pairs (bp), Megabase (Mb), or centiMorgan (cM). The following file header line is necessary:

//SNPid Chromosome Position

3 Customizable parameters

Population type: Type of population analysed. The possible choices are "CP (F1)", "BC", "F2" and "Germplasm". The Back-cross (BC) population is analysed as Cross-pollinated (CP) population, as the segregation types are the same (ABxAA or ABxAB and occasionally ABxAC=EFxEG). This tool does not make any assumption on the Grand-parental origin of the alleles. If the populatin type is "CP (F1)"or "BC" and the two parents name are identical ASSIsT consider the population as derived from a sel pollination and the data of the parent are duplicated to simulate two independent individuals.

Allowed missing data: Frequency of allowed missing data (No Call) by SNP and by individuals. The default value is 0.05. Range [0,1].

Call Rate tolerance: Maximum tolerance for the distance between an individual call rate and the analysed population mean. This parameter is used to exclude individuals (rather than SNP markers) for which too many SNPs could not be called. The default value is 0.1. Range [0,1]

p-Value (Chi-sq) segregation distortion: p-value of the Chi-squared test to test the allelic segregation. This check is based on the Hardy-Weinberg equilibrium test for unstructured germplasm or the expected and observed segregation ratio's when bi-parental populations are analysed. The lower the threshold, the more distortion is allowed. The default value is 0.001. Range [0,1].

Unexpected genotype threshold per individual: Proportion of allowed unexpected genotypes for each individual (Trio's Mendel Errors). This parameter is applied at a very final stage of the filtering process, after having accounted for "AB- sub-clusters and Null alleles" (see below) and is used to exclude individuals that have high probability to be not true to type or that have erroneous pedigree records. The default value is 0.003. Range [0,1].

Unexpected genotype threshold per SNP: Proportion of allowed unexpected genotypes for each SNP (Trio's Mendel Errors). Unexpeced calls will be made missing as long as their proportion does not exceed this threshold. When the threshold is exceeded, this SNP will be excluded. Note that this option is only available when Population Type "CP (F1)", "BC" or "F2" is selected. The default value is 0.05. Range [0,1].

Frequency rare allele: Maximum frequency to define an allele as rare. Note that this option is only available when Population Type "Germplasm" is selected. The default value is 0.05. Range [0,1].

Parents: Parents ("CP (F1)", BCx) or grandparents (F2) of the analysed experimental population. This parameter is used to specify which segregating family will be analysed. Note that this option is not available when germplasm is selected as Population type.

Individuals to exclude: It is possible to select the individuals to remove prior to the filtering analysis.

Number of chromosomes: Chromosome number of the species in the analysis. Since the tool was developed for apple, the default value is 17.

AB sub-clusters & Null alleles: Find and score markers with the AB cluster split into two sub-clusters and markers that show a null-allele. Note that this option is only available when Population Type "CP (F1)" or "BC" is selected.

4 Output files format

summary.txt: Gives an overview of the assay performances both by markers (number of markers for each class) and by individuals (number of samples analysed and list of individuals that did not pass the quality check: outcross or individuals with poor DNA quality). Moreover, it presents the data and parameter settings that were used for the analyses.

gtypes.csv: Custom file reporting the genotypes of all the successfully genotyped SNPs for each individual in the pedigree file. Each row represents an SNP. The individuals of the analysed population are sorted lexicographically, and the identified outcrosses are reported at the end of each row. The file contains the following information: SNP name (SNP id), Chromosome (Chr) position (Pos), Classification of SNP performance (Classification), Number of No Calls (Missing), number of individuals for each genotype (HomozygousNull, Hom1, Het1, Het2, Hom2), Chi-Squared p-value and the genotypes for each individual analysed. Note that Hom stands for homozygous (*AA* or *BB*), Het for heterozygous (*AB*, *Ab*, *aB*, *AO*, *BO*) and HomozygousNull for the contemporary presence of a null allele in both chromosomes (*OO*). *O* is used to indicate a null allele while a lowercase *a* or *b* indicates the presence of an additional SNP at the *A* or *B* probe site, respectively.

mendel_error.tsv: For each unexpected genotype, the individual involved is specified together with the genotypes of the two parents and the marker name (only for "CP (F1)", BCx or F2, and for the SNP that passed filtering).

snp_table.csv: Reports the segregation and classification information for each SNP (excluded and included). Each line reports the information for a single SNP in the following order: SNP id, genetic position (Chr and Pos), whether the marker has been exported (Exported), Classification of SNP performance (Classification), number of missing values (Missing), number of individuals for each allelic class (HomNull, Hom1, Het1, Het2, Hom2), Chi-Squared p-value. The genotype of the parents (GT Par1,GT Par2) is reported only when an experimental population is analysed, while the MAF information is provided only when a Germplasm set is analysed. Het2 represents the second heterozygous state and is present only for the SNPs with *AB* cluster showing a significant split in two sub-clusters or for the *A0 x BO* cross.

joinmap.loc: Input file for JoinMap® . This file is created only for the CP population. More details on the file format are available on the JoinMap® user manual beginning on page 46. This file contains only the "approved" markers while the "discarded" markers are left out.

flexQTL DataPrepper: This output is helpful when preparing an input file for FlexQTL™.

PLINK: Creates the ped and map file that can be used as an input for PLINK (Purcell et al. 2007); this file includes all the SNPs that pass the quality filtering. Details on the file format can be found at pngu.mgh.harvard.edu/ purcell/plink/data.shtml.

HapMap: creates a file including all the SNPs that pass the quality filtering. File format specifications can be found at www.broadinstitute.org.

Structure: Input file for Structure. The file includes two header lines with the marker position and the relative distance between them. Each individual is stored in a single line. The missing data are coded as '-9', while the nucleotides are stored as digits (*1=A, 2=C, 3=G, 4=T*).

5 SNP classification

A pre-screening of the fully genotyped germplasm is performed to identify poorly performing SNPs and individuals with low DNA quality. In this first phase all the SNPs showing more than 75% of NoCall in GenomeStudio® are classified as Failed and excluded from further analysis. Additionally, the accessions showing a CallRate lower than the average CallRate minus the "Call Rate tolerance" are also excluded. The remaining SNPs are further classified based on their performances on the accessions from the pedigree file.

Robust: All the successfully genotyped SNPs in which the segregation follows Mendelian rules and the number of NoCall is lower than the maximum allowed in the dataset. In "CP (F1)" and "BC" populations, the SNPs can be segregated into two or three clusters depending on the parent genotype (*AA*x*AB* and *AB*x*AB*). In "F2" and "Germplasm", populations, the SNPs show three clusters with a not significant p-value for the Chi-squared test.

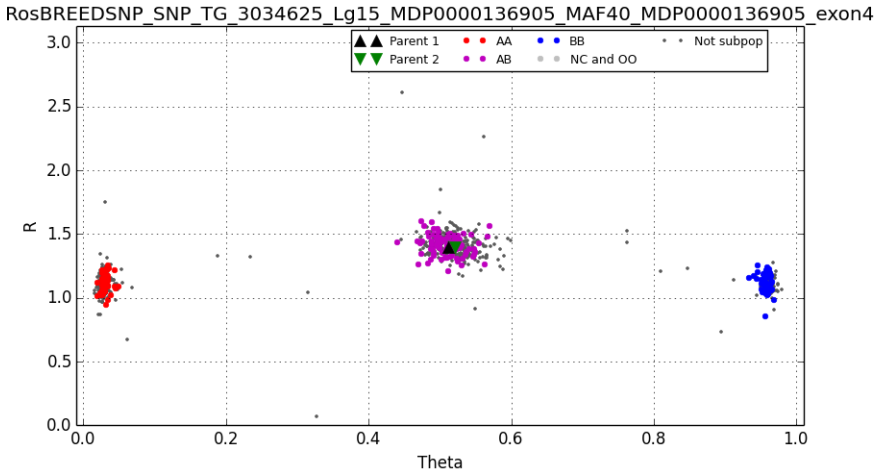


Figure 5: Plot of a Robust SNP.

NullAllele-Failed: This class may appear when "AB sub-clusters & Null alleles" is set to "Off". NullAllele-Failed are the SNPs in which the frequency of the HomozygousNull genotypes (No Call with an intensity of the luminous signal, R, lower than the threshold for null-alleles) is higher than the "Unexpected genotype threshold per SNP" in "CP (F1)", BCx or F2 or higher than the "Frequency rare allele" in "Germplasm". When "AB sub-clusters & Null alleles" is set to "On" the NullAllele-Failed are the SNPs in which the frequency of HomozygousNull exceeds the "Unexpected genotype threshold per SNP", or for null-allele including segregation patterns that ASSIsT does not account for (see last page of the manual), or when the segregation is too skewed (Chi-squared p-value lower than the maximum allowed distortion) to fall in Null_2_Clusters or Null_4_Clusters.

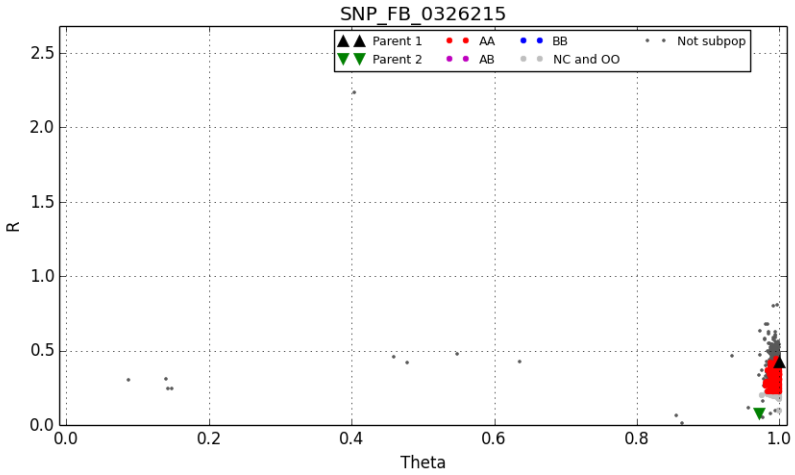


Figure 6: Plot of a NullAllele-Failed SNP.

Null_2_Clusters: SNPs fall in this category if the frequency of homozygous Null genotypes is higher than the "Unexpected genotype threshold per SNP" and if the frequency of one of the homozygous as well as heterozygous classes are lower than the "Unexpected genotype threshold per SNP". The presence of the null allele is coded with "*O*". According to the genotypes of the two parents it is possible to distinguish two different classes of markers: If parents are *AO* x *OO*, half of the offspring will be *AO* and half will be *OO*. If both parents are *AO*, one quarter of the offspring will be *AA*, half will be *AO* and one quarter will be *OO*. **AA** and **AO** clusters are often partially or totally merged so ASSIsT will score the marker according to the presence/absence of the *A* allele in the offspring. This re-calling analysis results in two genotype configuration: *A*- (the second allele is not specified as it is not possible to determine whether the genotype is *AA* or *AO*). This class can be present only in "CP (F1)" and "BC" populations when "AB sub-clusters & Null alleles" is set to "On". Note that ASSIST does not account for crosses of type *AB* x *OO* or *AB* x *AO*.

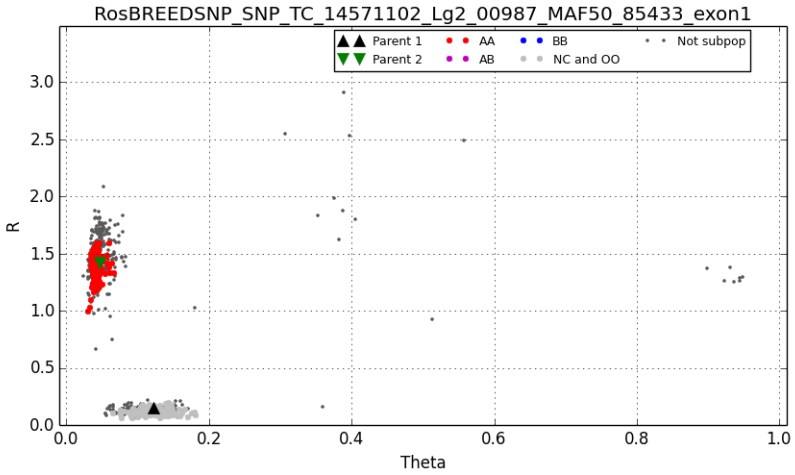


Figure 7: Plot of a Null-one-parent 2 Clusters SNP.

Null_4_Cluster: SNPs fall in this category when the frequencies of *AA*, *AB*, *BB* and *HomNull* (Both chromosomes with null allele) are higher than the "Unexpected genotype threshold per SNP" and one of the parents is initially called AA and the other BB. Based on the observed segregation pattern of the family (1 *AA*: 1 *AB*: 1 *BB*: 1 *OO*), these parents are recoded as *AO* and *BO*, and their progeny is recoded as *AO*, *AB*, *BO* and *OO*, respectively. This class can be present only in a "CP (F1)" and "BC" population when "AB sub-clusters & Null alleles" is set to "On".

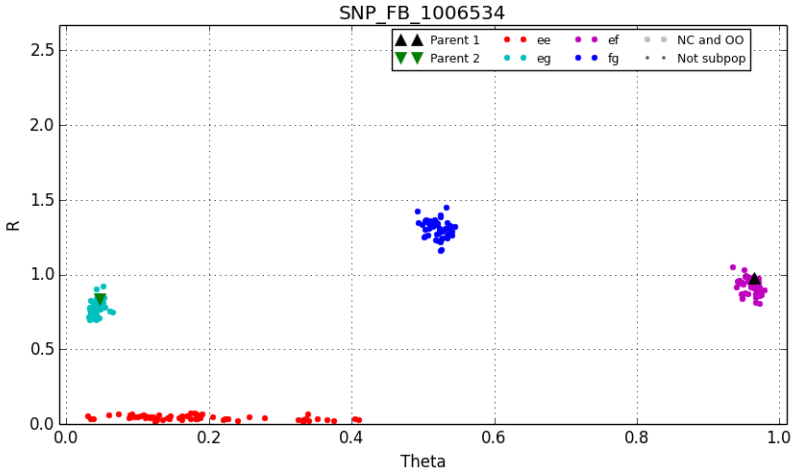


Figure 8: Plot of a Null-two-parents 4 Cluster SNP.

AB_2_sub-clusters: The separation of the *AB* cluster into two distinct sub-clusters is tested when the "AB sub-cluster & NullAllele" option is activated. The

presence of 2 sub-clusters within the *AB* genotypes is assessed by looking at the presence of one gap in the derivatives of the distances between the Theta of a contiguous data point. To exclude spurious separation at the lower or higher bound of the *AB* cluster, please be aware that the derivate is computed after a 10% trim of the extreme values of Theta. The separation is accepted when less than three consecutive values are over 2 * 95th percentile of the derivative distribution. This class can be present only in the "CP (F1)" and "BC" populations.

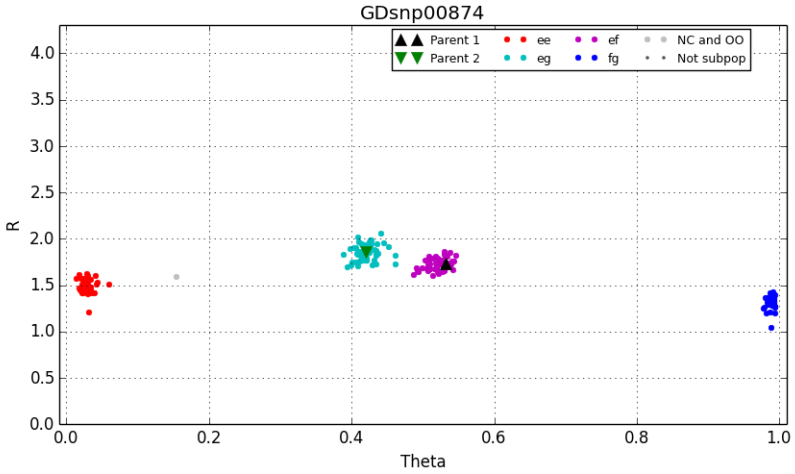


Figure 9: Plot of a AB 2 sub-clusters SNP

OneHomozygRare_HWE: The SNPs are classified as "OneHomozygRare_HWE" when the frequency of one homozygote cluster is lower than the threshold for the "Frequency rare allele" but the proportions of the three genotype classes respect the Hardy-Weinberg equilibrium. This class can be present only in the "Germplasm" population. In this case the "Frequency rare allele" is actually used as genotype frequency and not as allelic frequency to warn the users about the presence of clusters comprising few individuals. This situation in some cases can hide the presence of an unspecific annealing that causes a shift of part of the homozygote cluster at higher or lower Theta values in correspondence to the heterozygote cluster.

OneHomozygRare_NotHWE: The SNPs are classified as "OneHomozygRare_ NotHWE" when the frequency of one homozygote cluster is lower than the threshold for the "Frequency rare allele" and the proportions of the three genotype classes does not follow the Hardy-Weinberg equilibrium. This class can be present only in the "Germplasm" population. In this case the "Frequency rare allele" is actually used as genotype frequency and not as allelic frequency to warn the users about the presence of clusters comprising

few individuals. This situation in some cases can hide the presence of an unspecific annealing that causes a shift of part of the homozygote cluster at higher or lower Theta values in correspondence to the heterozygote cluster.

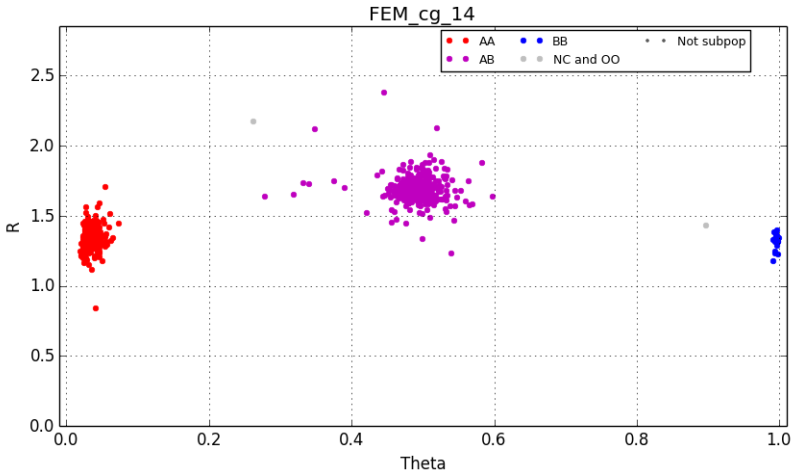


Figure 10: Plot of a OneHomozygRare_(Not)HWE SNP

Monomorphic: The SNPs are classified as False-SNP when a single Genotype class is present and its frequency is higher than the rare allele frequency threshold.

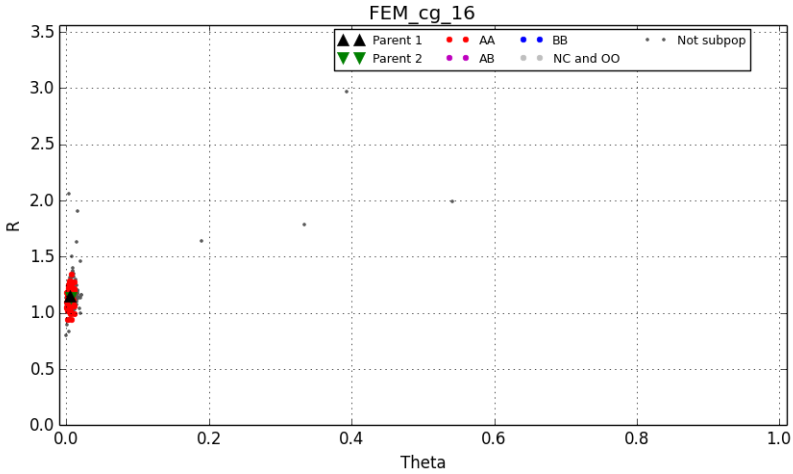


Figure 11: Plot of a Monomorphic SNP.

DistortedAndUnexpSegreg: The segregation in the full-sib families shows a severe skewedness (Chi-squared p-value higher than the value set in the pa-

rameters), or one of the genotype classes is missing in a "Germplasm" population, or a genotype class occurs which is not supported by the parental genotypes. This could be due to for instance a AB x AO marker.

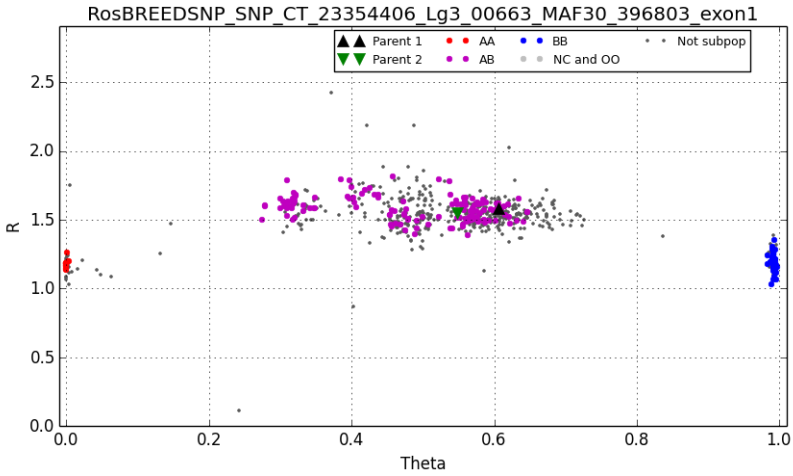


Figure 12: Plot of a Distorted SNP.

OneClassMissing: In an "F2" population, SNPs fall into this class when one of the three genotypes has a frequency lower than the rare allele frequency threshold.

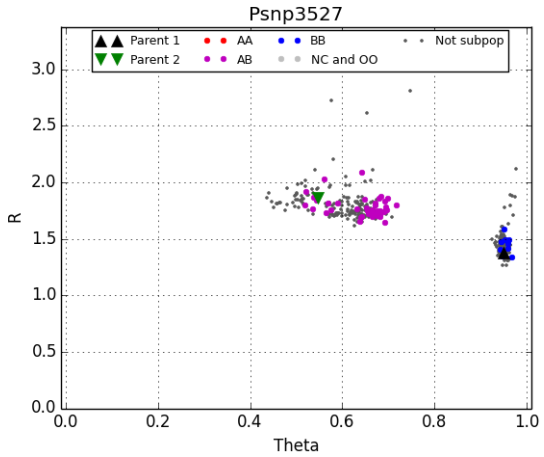


Figure 13: Plot of a oneClassMissing SNP.

ShiftedHomo: In a "Germplasm" population, SNPs are classified as ShiftedHomo when one of the two homozygous classes is absent. This is normally due to an unspecific annealing that causes a shift of the cluster at higher or lower

Theta values, depending on the allele that is the concern of the paralogy.

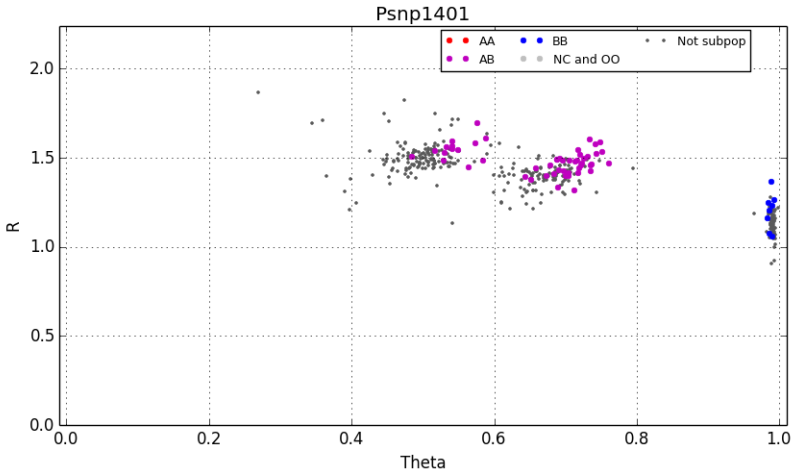


Figure 14: Plot of a shiftedHomo SNP.

Failed: All the SNPs that show a high rate of no-call, that have a mean signal intensity <0.4 or that do not fall in any other class are classified as Failed.

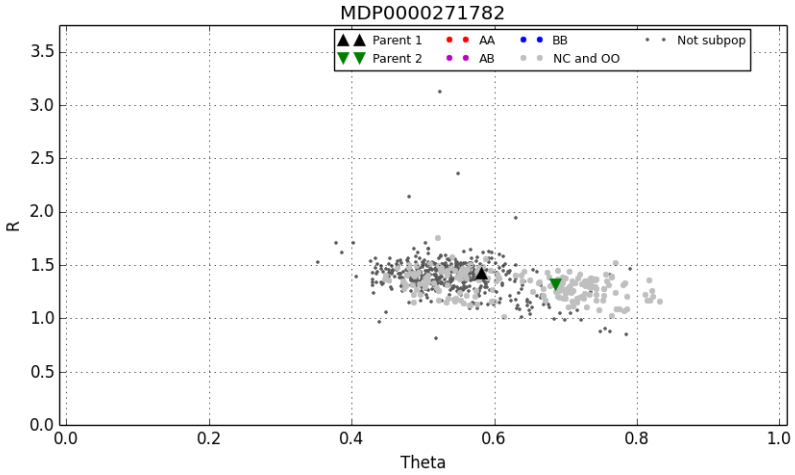


Figure 15: Plot of a Failed SNP

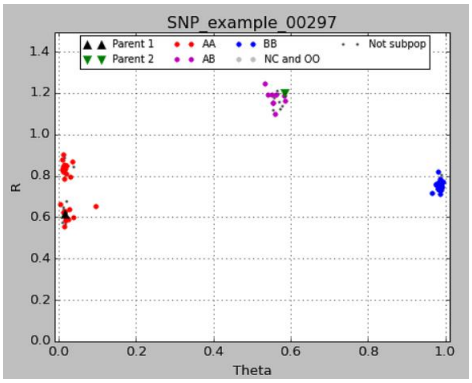
6 Prospects for further development

To our knowledge ASSIsT is the first software that identifies and calls null-alleles from SNP markers. The parental SNP-genotype combinations considered are *AO* x *AO*, *AO* x *OO* and *AO* x *BO*. The combinations *AB* x *AO* (**case (a)**) and *AB* x *OO*

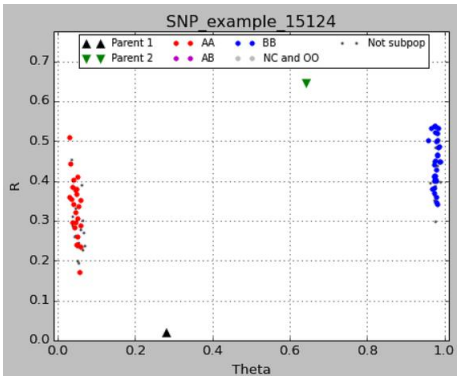
(**case (b)**) do show equally good prospects based on our results on SNP that were filtered and called using Excel based procedures developed in-house. These were not incorporated into ASSIsT due to time constraints.

Currently, neither GenomeStudio® nor ASSIsT supports automated calling of SNP for which one of the clusters for homozygous individuals (*AA* or *BB*) is in between x=0.4 and x=0.6, which is true for part of the paralogous SNP. Part of these SNP markers do show three well separated clusters (**case (c)**), and thus have good prospects for calling through alternative procedures. Another useful extension could be the further classification of excluded markers. Currently, markers with non-allowed genotypes and markers with segregation distortion are both assigned to the class "Distorted and unexpected segregation".

(a) AO x $AB \rightarrow \frac{1}{4}\,AO + \frac{1}{4}\,AA + \frac{1}{4}\,AB + \frac{1}{4}\,B0$



(b) OO x $AB \rightarrow \frac{1}{2}\,AO + \frac{1}{2}\,BO$



(c) AB x $AB \rightarrow \frac{1}{4}\,AA + \frac{1}{2}\,AB + \frac{1}{4}\,BB$