Bioinformatics, 31(23), 2015, 3873–3874
doi: 10.1093/bioinformatics/btv446
Advance Access Publication Date: 6 August 2015
Applications Note

Data and text mining

ASSIsT: an automatic SNP scoring tool for in- and outbreeding species

Mario Di Guardo[1,2,3,†], Diego Micheletti[2,4,†], Luca Bianco[2], Herma J. J. Koehorst-van Putten[1], Sara Longhi[1], Fabrizio Costa[2], Maria J. Aranzana[4], Riccardo Velasco[2], Pere Arús[4], Michela Troggio[2] and Eric W. van de Weg[1,*]

[1]Wageningen UR Plant Breeding, 6700 AA Wageningen, The Netherlands, [2]Research and Innovation Centre, Fondazione Edmund Mach, Trento, Italy, [3]Graduate School Experimental Plant Sciences, Wageningen University, 6700 AJ Wageningen, The Netherlands and [4]IRTA, Centre de Recerca en Agrigenómica CSIC-IRTA-UAB, Beellaterra (Cerdanyola del Vallés), 08193 Barcelona, Spain

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on April 22, 2015; revised on July 3, 2015; accepted on July 25, 2015

Abstract

ASSIsT (Automatic SNP ScorIng Tool) is a user-friendly customized pipeline for efficient calling and filtering of SNPs from Illumina Infinium arrays, specifically devised for custom genotyping arrays. Illumina has developed an integrated software for SNP data visualization and inspection called GenomeStudio® (GS). ASSIsT builds on GS-derived data and identifies those markers that follow a bi-allelic genetic model and show reliable genotype calls. Moreover, ASSIsT re-edits SNP calls with null alleles or additional SNPs in the probe annealing site. ASSIsT can be employed in the analysis of different population types such as full-sib families and mating schemes used in the plant kingdom (backcross, F1, F2), and unrelated individuals. The final result can be directly exported in the format required by the most common software for genetic mapping and marker–trait association analysis. ASSIsT is developed in Python and runs in Windows and Linux.

Availability and implementation: The software, example data sets and tutorials are freely available at http://compbiotoolbox.fmach.it/assist/.

Contact: eric.vandeweg@wur.nl

1 Introduction

Advances in whole genome genotyping technologies enabled the investigation of several hundred thousand SNP markers simultaneously on a genome-wide scale. To date, Illumina (GoldenGate® and Infinium®) and Affimetrix (Axiom®) are the most widely used array-based genotyping platforms worldwide. Illumina has developed GenomeStudio®, a proprietary software with a graphical user interface (GUI) for SNP data visualization and filtering that enables the selection of high-quality markers showing robust performance across the examined germplasm. However, the actual filtering of such SNPs requires a deep understanding of the performance of SNP markers, genetic segregation patterns and familiarity with the many tools and parameters in GenomeStudio® (GS). ASSIsT accounts for this by offering a user friendly, automated pipeline that builds on the results of Illumina's GenCall algorithm (Kermani, 2006) as incorporated in GS.

In addition to filtering, ASSIsT also re-edits GS-calls in order to better explore the available information for SNPs showing null alleles or additional SNP clusters® due to additional polymorphisms at the probe annealing site. This re-editing enhances correct SNP calling and reduces unnecessary removal of potentially valuable markers.

2 Methods

The analysis and selection of SNPs performed by ASSisT is based on the calls produced by Illumina's GenCall algorithm (Kermani, 2006). A two tiers approach that employs a bi-allelic genetic model, and then a tri-allelic model is used to classify SNPs on the basis of their real performance on examined germplasm. The tri-allelic model is used to describe more complex segregation patterns due to null-alleles or alleles with variable signal intensity due to additional SNP, as the bi-allelic genetic model used by GS cannot account for such polymorphisms (Bassil *et al.*, 2015; Gardner *et al.*, 2014; Pikunova *et al.*, 2014; Troggio *et al.*, 2013). In this case, ASSIsT may re-edit GS-calls by applying *de novo* filters using the original light intensity data and the segregation patterns in the germplasm.

3 Results

ASSIsT supports the analyses of different population types, such as full-sib families (e.g. human, livestock, cross pollinating plants), mating schemes common in plants (backcross, F1, F2) and individuals with unknown genetic relationships. ASSIsT's GUI allows easy parameter setting and provides a visual output of the SNP clustering analysis. The results produced by ASSIsT can be directly exported to the input format of the most widely used software for genetic and marker-trait association analysis (FlexQTL^{TM}, GAPIT, JoinMap, PLINK, Structure and Tassel). This straightforward integration will improve marker performance in association and QTL mapping studies. ASSIsT is developed in Python (www.python.org). Its source code is released under the GNU General Public Licence (GNU-GPLv3) to allow its integration into bioinformatic pipelines.

ASSIsT requires three input files: a pedigree file in which the parents of each sample are reported and two standard report files from GS (Final Report and DNA Report). The two GS reports are standard output of commercial service companies; therefore, ASSIsT does not necessarily require access to GS. A map file with the genetic or physical position of the markers may also be included. This information is mandatory for exporting results in Structure or PLINK formats.

ASSIsT allows pre-selection of the stringency of the filtering procedure by customizing the following parameters: (1) Proportion of missing data, (2) Call Rate threshold, (3) Segregation distortion (χ^2 *P*-value), (4) Frequency of not allowed genotypes (structured germplasm) and (5) Minor Allele Frequency.

The first step of the filtering analysis is a quality check of the individuals; samples with a high proportion of unexpected marker genotypes due to outcrossing, different ploidy levels and DNA admixture, among other causes, are considered deviating germplasm and further excluded from the analysis. Samples with poor DNA quality (call rate significantly lower than the average of the dataset) will not be considered in the analysis either. All discarded samples are listed in the 'summary' output file.

Only 'robust' markers (i.e. those showing a clear cluster separation and few No Calls) are allowed through the initial filtering. These markers can show two (one homozygous and one heterozygous) or three clusters (two homozygous and one heterozygous). For some markers, the AB cluster might result in two distinct subclusters, due to additional SNPs at the probe site, which may lead to differential hybridization efficiency and to distinct classes of signal intensity within a marker allele. The variation in signal intensity, generally ignored by GS, is considered by ASSisT instead. For instance, a cross between two heterozygous parents generates three genotype clusters at a single locus (e.g. $CT \times CT$ produces $1/4CC + 1/2CT + 1/4TT$). When one allele (let us say *T*) shows two distinct intensity classes, it may be interpreted as $\underline{CT} \times Ct$, which gives $1/4CC + 1/4CT + 1/4Ct + 1/4Tt$. The discernment between the two heterozygous classes (*CT* and *Ct*) makes this marker fully informative in inheritance studies, where as 'classical' heterozygotes are not informative in the generation of genetic linkage maps as it is not possible to determine the parental origin of the alleles. Additional SNPs in the probe, as well as INDELS (Pikunova *et al.*, 2014), may also give rise to null alleles, due to the lack of signal in one of the DNA templates, which results in additional clusters. GS cannot currently account for this scenario; thus, informative markers are lost. Conversely, ASSIsT succeeds in the analysis of the majority of such markers (A0 × A0, A0 × 00 and A0 × B0), allowing a more efficient marker calling.

All the above-mentioned SNP classes are suitable for the generation of genetic linkage maps or for marker–trait association studies. Discarded markers are grouped according to their performance considering absence of or severe distortion in segregation, presence of not allowed genotypes in segregating families and number of No Calls.

ASSIsT has been used to analyze SNP markers of several bi-parental full-sib families and germplasm of apple (Bianco *et al.*, 2014), peach, melon and grape. For each family, ~99% of the 'approved' (those that passed the filtering procedure) SNPs showed to have high-quality data as they integrated smoothly in the generation of high-quality genetic linkage maps. The remaining 1% presented several types of issues, largely related to the presence of paralog loci where the AB cluster was too close or even merged to one of the two homozygous clusters.

ASSIsT thus proved to be an effective tool for genotyping studies as it allows to easily filter informative and well-performing SNP and to recover potentially useful SNPs from indels or regions of high-sequence divergence, feeding them directly to the most common downstream analysis tools through its easy interface.

Funding

This work was co-funded by the EU seventh Framework Programme by the FruitBreedomics Project No. 265582: Integrated Approach for increasing breeding efficiency in fruit tree crops (www.FruitBreedomics.com). The views expressed in this work are the sole responsibility of the authors and do not necessary reflect the views of the European Commission.

Conflict of Interest: none declared

References

Bassil,N.V. *et al.* (2015). Development and preliminary evaluation of a 90K Axiom® SNP array in the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics*, **16**, 155.

Bianco,L. *et al.* (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* × *domestica* Borkh). *PLoS One*, **9**, e110377.

Gardner,K.M. *et al.* (2014). Fast and cost-effective genetic mapping in apple using next-generation sequencing. *G3*, **4**, 1681–1687.

Kermani,B.G. (2006). Artificial intelligence and global normalization methods for genotyping. US Patent 20 060 22 529.

Pikunova,A. *et al.* (2014). 'Schmidt's Antonovka' is identical to 'Common Antonovka', an apple cultivar widely used in Russia in the breeding for biotic and abiotic stresses. *Tree Genet. Genom.*, **10**, 261–271.

Troggio,M. *et al.* (2013). Evaluation of SNP data from the Malus infinium array identifies challenges for genetic analysis of complex genomes of polyploid origin. *PLoS One*, **8**, e67407.