

Genotyping-by-sequencing facilitates whole genome visualization of Canadian oat cultivars

Benazir K. Marquez^{1,2}, Charlene P. Wight¹, Jennifer Mitchell Fetch³, Aaron D. Beattie⁴, Douglas A. Johnson², Jesse Poland⁵, Nicholas A. Tinker^{*1}

¹Agriculture and Agri-Food Canada, Bldg. 20, 960 Carling Avenue, Ottawa, ON K1A 0C6, Canada ²Department of Biology, University of Ottawa, 30 Marie Curie Private, Ottawa, ON K1N 6N5, Canada ³Agriculture and Agri-Food Canada, 2701 Grand Valley Road, Brandon, MB R7A 5Y3, Canada ⁴Crop Development Centre, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada ⁵Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

*correspondence: nick.tinker@agr.gc.ca



Introduction

Neutral genetic variation can remain as residual heterogeneity within a cultivar. Multiple marker technologies have detected such variation within different crop species. Variation within a cultivar may also indicate an error in seed handling. Thorough assessment of seed purity requires the testing of large samples with a large number of markers. Genotyping-by-sequencing (GBS) produces an ample amount of data with significant genome coverage at low cost, providing an opportunity for seed purity evaluation. Graphical genotyping of GBS data would allow for efficient data interpretation.

Materials & Methods

Plant material and DNA preparation: Seeds of 'AC Morgan', 'Jordan', 'Leggett', and 'Ronald' were acquired from multiple sources so that the following were represented: breeder seed, certified seed, seed harvested from an internal cultivar increase, and seed harvested from a registration or PBR trial. Ten seeds from each of these sources (together referred to as CDOC-160) were grown for at least ten days. Young leaf tissue was harvested from single plants and dried separately. DNA was extracted from each plant's leaf tissue using a DNeasy Plant Mini kit according to the manufacturer's protocol. Extracts were quantified using Quant-iT Picogreen assays^[1]. Extracts were diluted to 1 ng/μL working volumes with sterile distilled water.

Genotyping-by-sequencing (GBS): was performed on libraries containing CDOC-160 DNA (one full and one partial 95-plex) which were constructed using the P384 adapter set^[2]. Genomic DNA was digested simultaneously with *Pst*I and *Msp*I. Barcoded adapters were ligated to each individual sample following digestion. Samples from each plate were pooled into a single library and PCR-amplified. Each 95-plex library was sequenced on a single lane of Illumina HiSeq 2000 at the National Research Council (Saskatoon, SK) by the DNA Technologies core facility. **The universal network enabled analysis kit (UNEAK) non-reference pipeline^[3]:** was used for SNP discovery and genotyping. The primary GBS analysis was done using the full CDOC-160 set as the reference sample. Reads containing a barcode sequence followed by the expected *Pst*I (CTGCAG)-*Msp*I (CCGG) cut sites were retained and trimmed to 64 bp. Identical reads present in more than five samples were retained and collapsed into tags. Pairwise alignment was implemented with the retained tags to find tag pairs with a 1 bp mismatch, considered afterward as candidate SNPs. In order to reduce false positive SNP calls, a network filter was employed to identify only reciprocal tag pairs. An error tolerance rate of 0.03 was applied to filter out possible paralogs and sequencing errors. Initial quality and genetic filtering in the UNEAK pipeline was performed with default parameters for minor allele frequency (MAF, 0.05-0.5) and call rate (0-1). Additional genetic filtering was performed using CbyT to retain only those SNPs showing maximum 5% heterozygosity, minimum MAF of 10%, and maximum 5% missing scores with CDOC-160 as the reference sample. **Principal component analysis (PCA):** was performed with the *R* statistical package using the *prcomp* function. **Graphical genotypes (GGs):** were generated using GGT 2.0^[4] and the map locations^[5] of 1344 GBS SNPs.

Results & Discussion

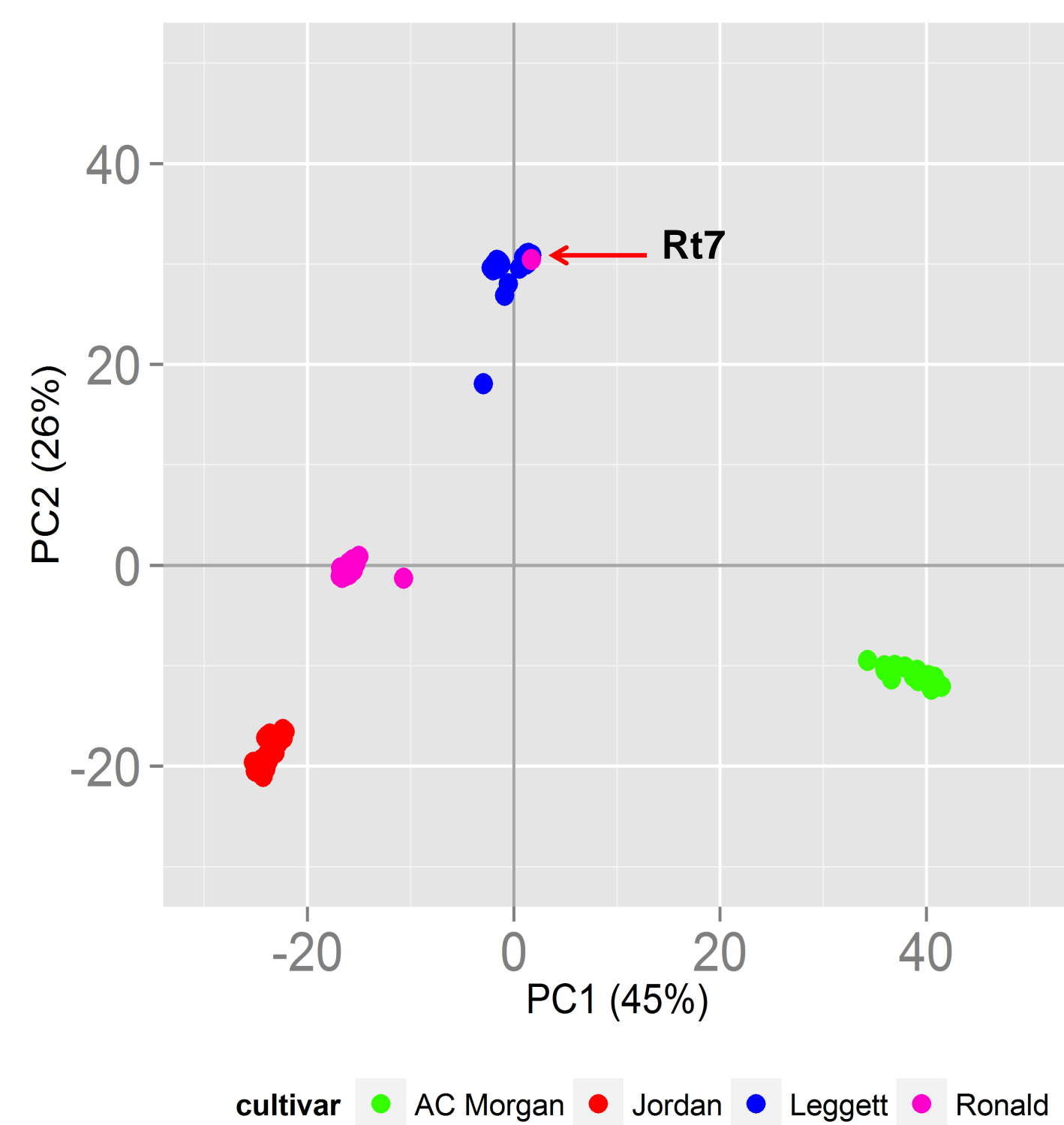


Figure 1. Principal component analysis (PCA) of GBS SNPs in oat cultivars AC Morgan, Jordan, Leggett, and Ronald. Each circle represents one seed sample.

A total of 2212 SNPs were called using UNEAK. Of these, 1344 (60.8%) were mapped on an oat consensus framework^[5]. Only 21.8% of all GBS loci (20.5% of mapped loci) showed substantial variation (defined as PIC>0.30) within one or more oat cultivars. PCA (Fig.1) revealed that all seeds clustered by cultivar, but that all cultivars contained some allelic variation. Jordan, AC Morgan, and Ronald formed tight clusters, while Leggett formed two noticeably separate clusters. One Ronald trial seed (Rt7) clustered tightly with seeds of Leggett, suggesting it is a contaminant.

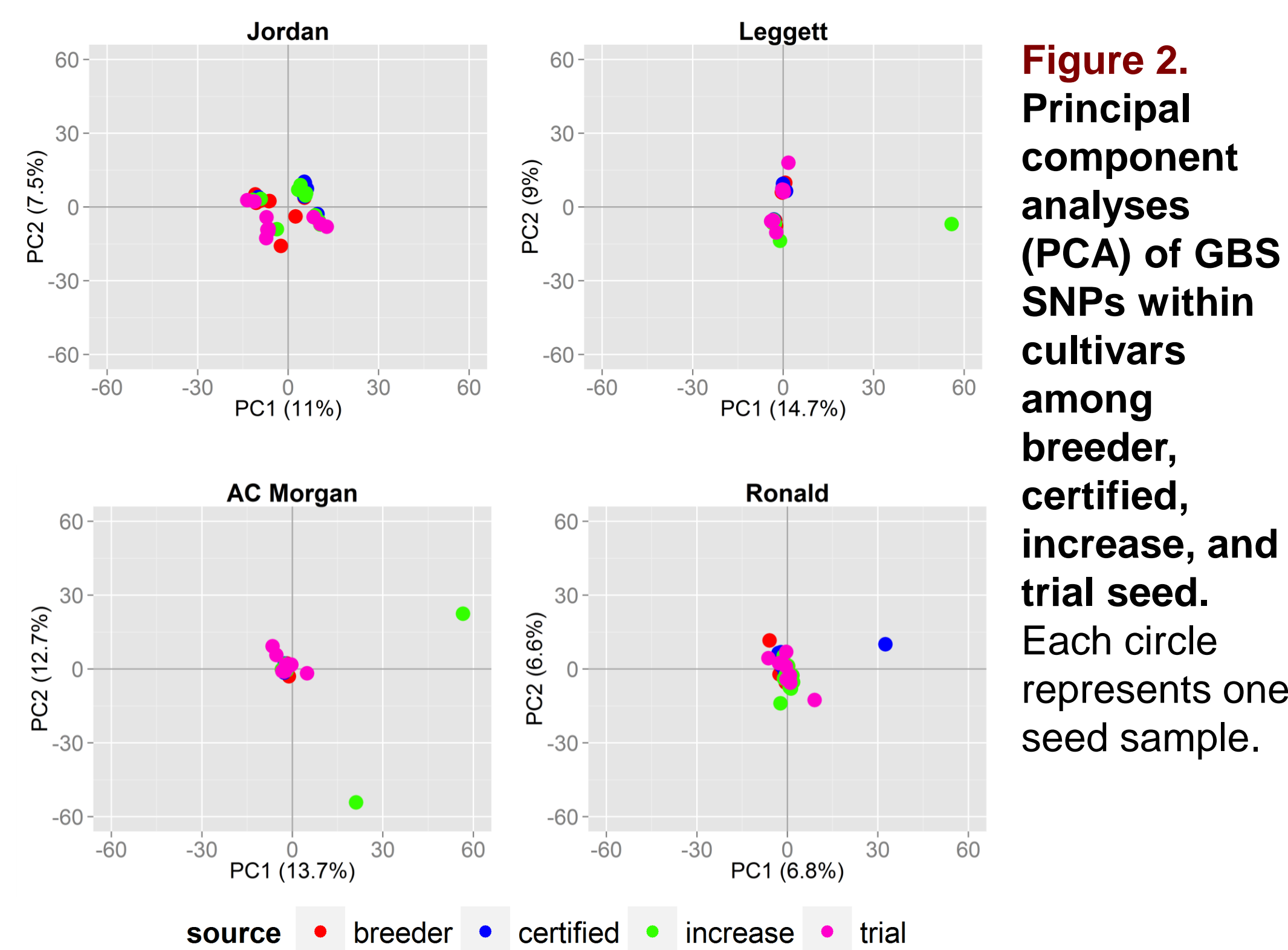


Figure 2. Principal component analyses (PCA) of GBS SNPs within cultivars among breeder, certified, increase, and trial seed. Each circle represents one seed sample.

To further investigate the genetic heterogeneity of each cultivar, PCAs were performed for each cultivar using only polymorphic SNPs (Fig.2). Jordan seeds appear to be the most heterogeneous, followed by Ronald, Leggett, and AC Morgan, based on cluster tightness observed in individual PCA plots. Jordan separated into four clusters, none of which reflected seed source. Leggett seeds segregated into two clusters, mirroring the results of the first PCA. Ronald and AC Morgan both appeared as tight clusters with a few seeds distinctly different. Cluster patterns suggest that variation did not appear to be limited to a specific seed source.

Whole genome GGs allowed for localization of this variation. Most of the observed variation fell within distinct map regions and showed allele frequencies between 40% and 60% in the cultivar in

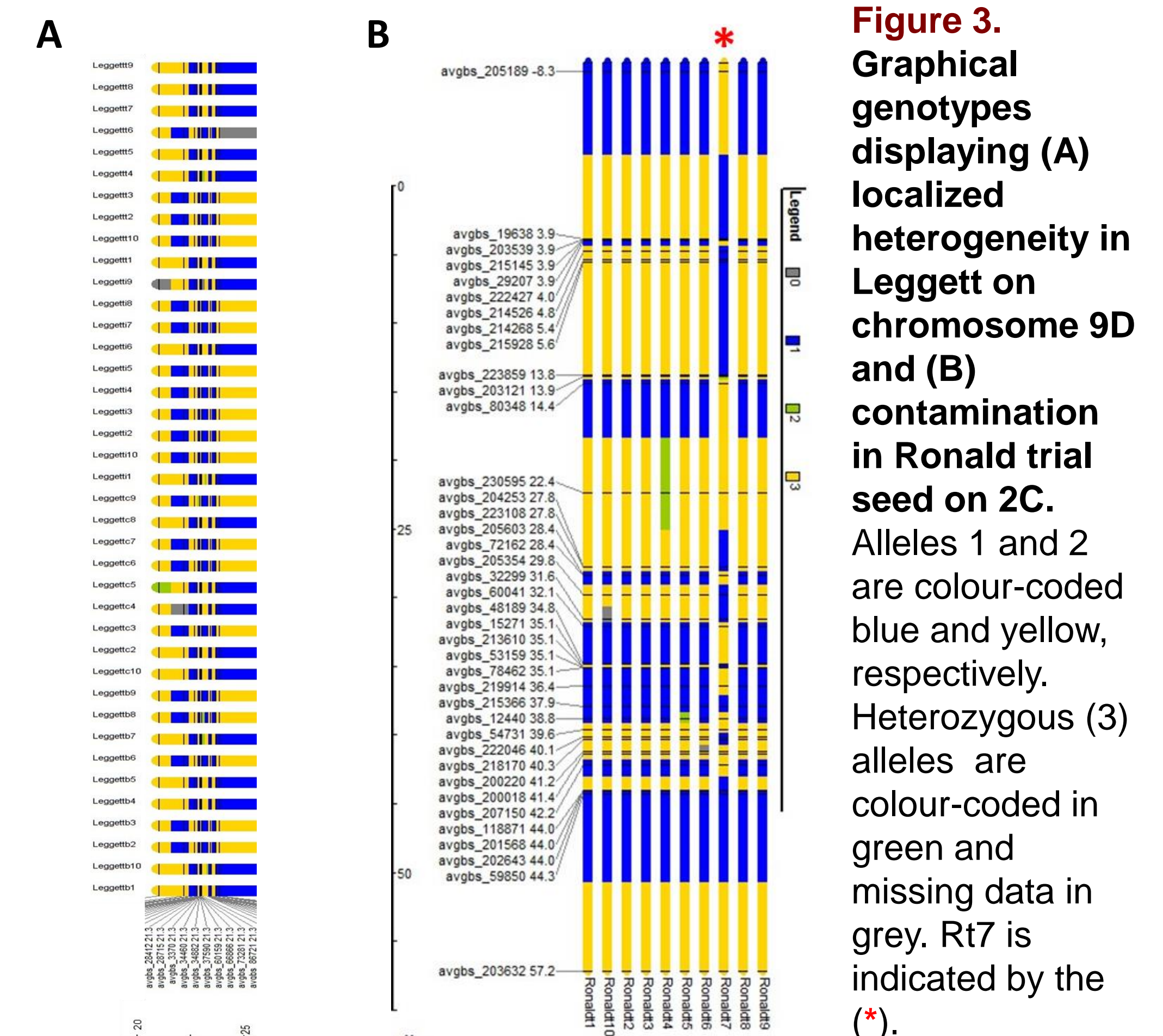


Figure 3. Graphical genotypes displaying (A) localized heterogeneity in Leggett on chromosome 9D and (B) contamination in Ronald trial seed on 2C. Alleles 1 and 2 are colour-coded blue and yellow, respectively. Heterozygous (3) alleles are colour-coded in green and missing data in grey. Rt7 is indicated by the (*).

which polymorphism was observed (*data not shown*). For example, the 20-23 cM region on chromosome 9D was observed to be extremely variable in the 40 Leggett seed samples (Fig. 3A). This region divided the Leggett seeds into two groups that reflected the clusters observed in the PCA score plot (Fig.1). This result is consistent with the expectation that most cultivars contain residual heterogeneity in distinct chromosomal regions that were heterozygous in the last generation from which seed was bulked from a single plant.

GGs also allowed for physical contamination to be distinguished from this genetic heterogeneity. Contamination was characterized by consistent, non-localized variation. Rt7 differed from Ronald seeds at every chromosome in this manner, as shown in Fig. 3B. There is no conclusive way of determining if Rt7 was introduced into the Ronald seed stock by seed handling errors in the field or during sampling for this study.

Conclusion

GBS offers breeders and researchers a whole genome perspective of current cultivars. This method of data interpretation would be suitable for breeders as an efficient tool for selection, and for inspectors as a method of discovering possible seed contaminants.

References

- [1] Blotta I *et al.* (2005) *Ann Ist Super Sanita* 41:119-123.
- [2] Poland J *et al.* (2012) *PLoS One* 7:e32253.
- [3] Lu F *et al.* (2013) *PLoS Genet* 9:e1003215.
- [4] van Berloo R (2008) *J Hered* 99:232-236.
- [5] Huang YF *et al.* (2013) *PLoS One* (*in press*).

Acknowledgements

Special thanks to Dr. Weikai Yan for providing his expert opinion on PCA and in regard to cultivar choice for this study.