Genome analysis
gsrc: an R package for genome structure rearrangement calling

Fabian Grandke, Rod Snowdon and Birgit Samans

Department of Plant Breeding, IFZ Research Centre for BioSystems, Land Use and Nutrition, Justus Liebig University, 35392 Giessen, Germany
*To whom correspondence should be addressed.

Abstract
Summary: Genome structure rearrangements are a common phenomenon in allopolyploid species. Deletions, duplications and homeologous non-reciprocal translocations (HNRT) between the highly similar subgenomes can be observed, which are known to have a large impact on phenotypic traits. Current research is limited because these rearrangements can be located genome wide only by cost intensive sequencing approaches and not reliably in high-density array genotyping data. We developed gsrc, an R-package to detect genome structure rearrangements from genotyping data in allopolyploid species including exchanges between subgenomes. We exemplarily apply gsrc to a publicly available Brassica napus dataset.

Availability and Implementation: The compiled R-package and source code are available at http://cran.r-project.org/web/packages/gsrc/.

Contact: birgit.samans@uni-giessen.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction
Polyploidization is a common event in flowering plants and has a high impact on agronomically important plants. Some of the economically most relevant crops are allopolyploid (e.g. wheat, rape seed), where the subgenomes merged through hybridization. Synthesis of two genomes often leads to structural rearrangements including deletions, duplications and homoeologous exchanges where homoeolog areas between the subgenomes are translocated non-reciprocally, leading to a deletion in one subgenome and a duplication in the corresponding area of the other subgenome (Cai et al., 2014). Genomic rearrangements have been shown to affect phenotypic traits and therefore it is of interest to monitor them on a genome wide scale. High-density single nucleotide polymorphism (SNP) arrays are well established in plant sciences for monitoring genetic and genomic variation. As they are more affordable than alternative technologies (e.g. sequencing), large cohorts of samples can be measured. Numerous tools for calling genotypes or copy-number variations (CNVs) from high-density array genotyping data are available. However, all of them are limited to specific array designs or species (e.g. celmm) (Carvalho et al., 2010). The proprietary tools to generate these data formats, usually provided by the microarray manufacturers, are closed source and closed access (Ritchie et al., 2011). Additionally, none of them includes the ability to detect homoeologous exchanges in allopolyploid species. Here we present gsrc, an R package (R Core Team, 2015) for genome structure rearrangement calling in allopolyploids. It enables scientist to detect and visualize CNVs and reciprocal translocations in species with homoeologous regions.

2 Methods and algorithms
A conceptual overview of gsrc is provided in Figure 1 A. It starts with reading in binary raw data from two-signal SNP arrays, quality control and normalization. Polar coordinates transformations (PCT) and k-means clustering based genotype calling are applied to calculate B-allele frequencies (BAF) and the corresponding Log R ratios (LRR) for each marker (Peiffer et al., 2006). Circular binary segmentation (CBS) combines LRR values with SNP positions and groups consecutive markers into regions with similar intensities indicating the same copy number level (Olshen et al., 2004). User defined
thresholds categorize these regions to estimate the level of different copy number variations (CNVs) for each sample. The detection of HNRTs requires information about corresponding synteny blocks between the subgenomes. Synteny calculation is based on markers which can be mapped to both subgenomes. gsrc employs density-based spatial clustering of applications with noise (DBSCAN) to remove outliers and reveal global synteny regions (Ester et al., 1996). These are refined into smaller blocks to increase the resolution of the HNRT detection. The blocks are scanned for opposing CNVs with deletions in one subgenome and duplications in the other subgenome. The final step of our pipeline visualizes LRR, CNVs, translocations and synteny blocks (Fig. 1 B). gsrc can be tailored to different datasets and combined with other R packages and algorithms. For instance, synteny block calculation can be based on genome sequences instead of unigene mapping.

3 Example

The vignette (Supplementary data) shows an example workflow of gsrc with 140 samples of a dataset for the allotetraploid crop Brassica napus (Körber et al., 2012). Genotyping was performed with the Brassica 60k Illumina\textsuperscript{®} Infinium consortium SNP array (Edwards et al., 2013). The genome of Brassica napus consists of two subgenomes A and C with ten and nine chromosomes, respectively. Unigene mapping and SNP positions are publicly available (Bancroft et al., 2015). An example of the result is shown in Figure 1 B.

Funding

This work was supported by the DFG (234585441, PREDICT) and the BMBF (0315964, PRE-BREED YIELD).

Conflict of Interest: none declared.

References


Richie, M.E. et al. (2011) Comparing genotyping algorithms for Illumina’s Infinium whole-genome SNP BeadChips. BMC Bioinformatics, 12, 68.