Variant calling quality enhancement within a large breeding population

Authors: Jan van Oeveren[1] , Antoine Janssen[1], Bas Tolhuis[2] and Roeland van Ham[1]
Affiliations: [1] Keygene N.V., Agro Business Park 90, 6708 PW Wageningen, The Netherlands
[2] Genalice Core B.V., Paasbosweg 18-20, 3862 ZS Nijkerk, The Netherlands

For more information contact
roeland.van-ham@keygene.com or
bas.tolhuis@genalice.com

Abstract

Whole genome sequencing applications are rapidly being adopted in agrigenomics to improve molecular breeding strategies. This massively increases data production and creates growing computational and storage challenges in a time and budget constraint environment. As a solution to this bottleneck in NGS analysis, KeyGene has implemented GENALICE MAP as secondary analysis software for ultra-fast mapping and variant calling of complex crop genomes. Processing times have been reduced over a 100 fold, whereas former hardware requirements of a cluster >1000 cores decreased to a single machine with 12 cores. The acceleration thus achieved enables large cohort, multi sample variant calling for population genomics applications. In this study the use, accuracy and efficiency of the GENALICE Population Caller is characterized for SNP variation of a large breeding population of highly homozygous individuals, derived from complex crosses between multiple parental lines. This population set-up allows for the assessment of true genotype composition from parental haplotype block information. The data set is assessed with individual line and population based variant scores. We demonstrate the quality enhancement capabilities.

1. Multi-parent Advanced Generation Inter-Cross (MAGIC) population

In this study, we use the MAGIC population to assess the quality of genotype calls for a large set of individuals. This method circumvents lack of golden standard variants and large scale wet lab validation.

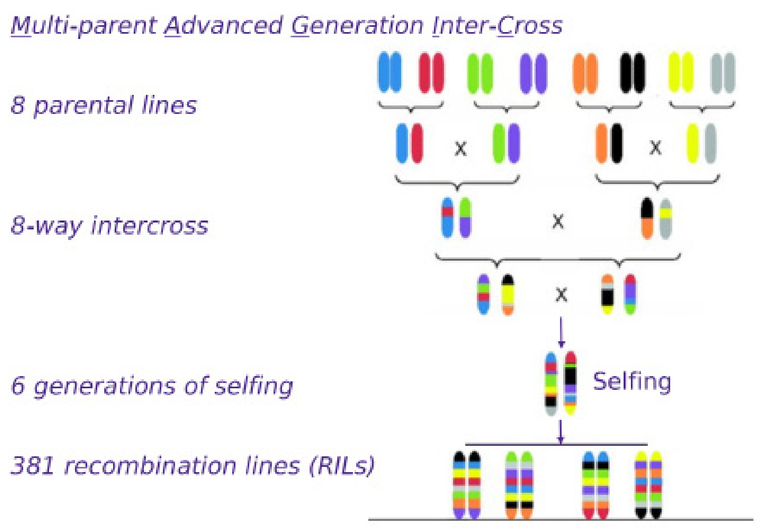

Figure 1 Breeding scheme of the MAGIC population. The foundation of the population are eight homozygous parental lines. The parents are crossed using an eight-way intercross. The offspring is bred to homozygosity using six generations of self-crosses. This resulted in 381 recombination inbred lines (RILs). Variants present in the parents can be used to determine haplotype blocks in the RIL offspring. Moreover, parental variants can be used as true variants in accuracy assessment.

2. Whole genome re-sequencing and data analysis

We used Illumina short-read technology to sequence the parent and RIL individuals. We sequenced the parents at high depth (~60x) to get high quality variant calls. We used shallow (~7x) sequencing depths for the 381 RILs. We analyzed the sequence reads with GENALICE MAP read mapping and variant calling. This is a fast and accurate tool to analyze short-read sequencing data [1]. The tool can analyze individual samples and the population as a whole. Moreover, it can use genetic profiles or a population context to enhance the calling of variants.

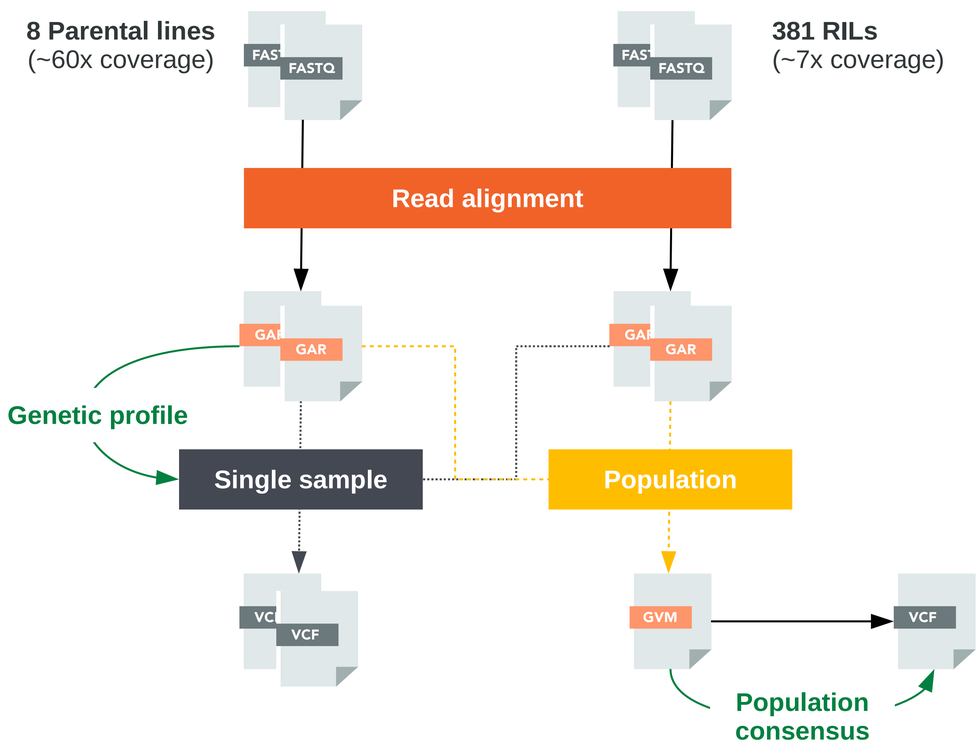

Figure 2 Illumina re-sequencing and GENALICE MAP data analysis methods. We analyzed 8 parental and 381 RIL lines with GENALICE MAP. First we mapped short sequence reads to the reference genome for all individual samples, resulting in GENALICE Aligned Reads (GAR) files holding read mapping results. Next, we either used GENALICE MAP single sample (gray) or population (yellow) variant calling. Single sample calling results in one VCF file per sample, while population calling stores variants in an extensible container called GENALICE Variant Map (GVM). The collected variants can be exported to one multi-sample VCF file. A parental genetic profile can enhance the quality of single sample variant calls. Consensus in the population can enhance the quality of population variant calls.

3. Genetic profile and population consensus increase variant yield

Single sample and population calling yield a similar number of variants in the MAGIC population. The yield can be increased using either a parental genetic profile (single sample) or Consensus Based Call Enhancement [CBCE] (population). The increase in variant calls might indicate more value retrieved from the data.

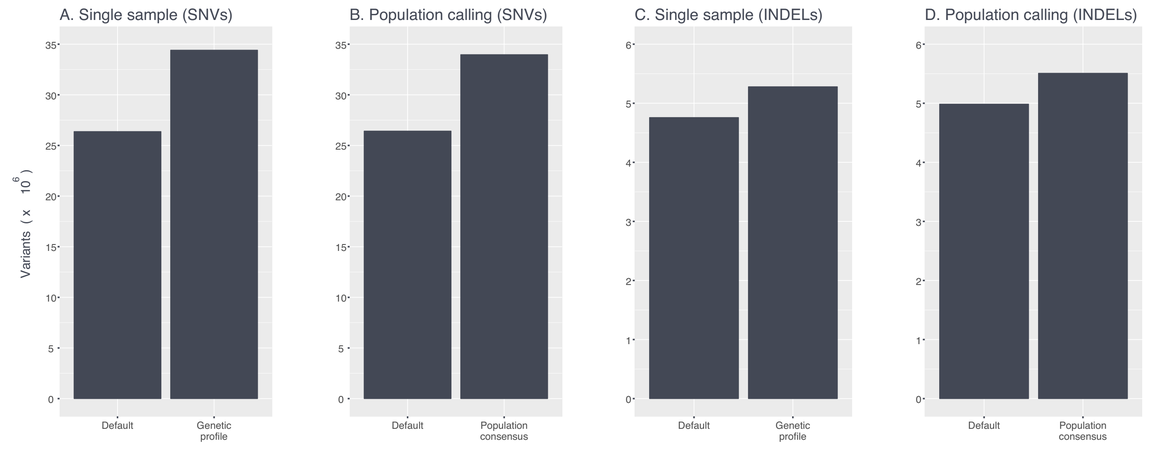

Figure 3 Number of variants in the MAGIC population. Number of variants (SNVs and INDELs) called with either GENALICE MAP Population calling or Single sample calling using the standard method (default) or quality enhancement through Consensus Based Call Enhancement (population calling) / parental genetic profile (single sample).

4. Genetic profile and population consensus result in better concordance

Enabling either population consensus or a genetic profile results in more variants that are concordant with parent specific variants. In addition, enhancement results in more discordant calls. The single sample approach yields most concordant INDELs.

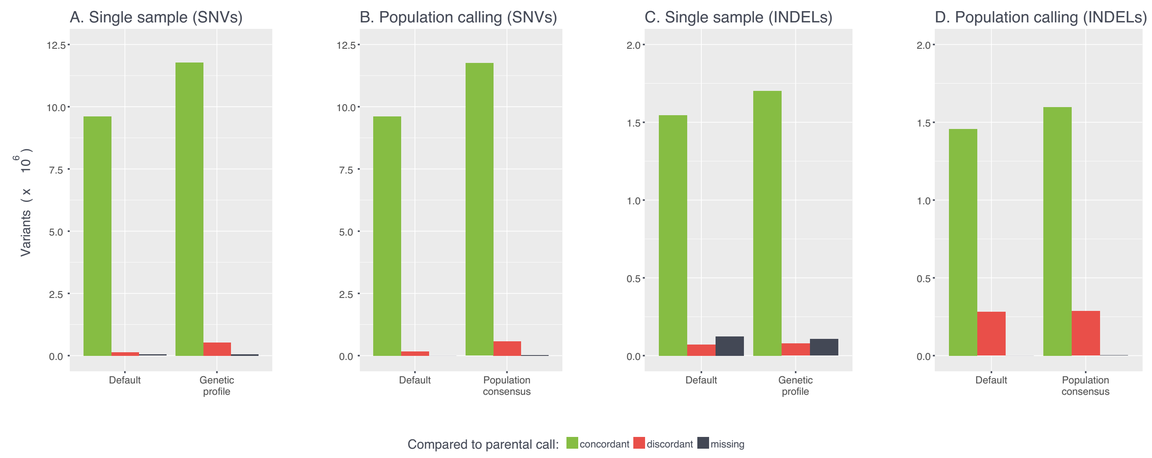

Figure 4 Number of concordant, discordant and not called parental variants in RILs. Number of parental variants (SNVs and INDELs) across the RILs that are concordant (green), discordant (red) or missing (gray). Variants were called with either Population or Single sample calling using standard method (default) or quality enhancement (CBCE/Profile).

5. Haplotype block assessment

Using the constitution of the MAGIC inbred lines and the occurrences of variants unique to a single parent, we defined for each inbred line regions of parental origin, so-called haplotype blocks. The process for deriving this haplotype block information was easier for the Population Calling setup, compared to the single sample calling procedure, as a single vcf file is required as input, with reference allele scores included. This is exactly the output of the Population Caller, whereas the single sample calling requires three steps. First step is deriving all variants, second step is scoring all variant loci for all individuals, including reference alleles, and third a merge of all individual vcf files is performed. The population analysis leads to an increase in mean haplotype block length of 6.2% and improves genome coverage with 5.5%. Where the single sample calling could identify 64.9% of the genome into high confident haplotypes, the population caller resulted in 67.5%, with larger blocks (mean size 1.7 Mbp).

	Single sample	Population calling	Relative change
Total size (bases)	225,216,428	234,264,090	5.5%
Proportion of genome	64.9%	67.5%	5.5%
Mean size (bases)	4,305,962	4,570,215	6.2%
Stddev size (bases)	5,976,076	6,218,160	4.0%
Number of blocks	20,559	20,120	–2.3%

Conclusions

Quality enhancement with GENALICE MAP results in higher variant yield and more concordant calls, suggesting increased value of the data results. Quality enhancement can be done directly on the MAGIC population or on single samples by using a parental genetic profile. The GENALICE MAP Population Callers makes haplotype block detection easy and provides increased yield when calling high quality haplotype blocks.

References
[1] Pluss et al 2017 PNAS, DOI: 10.1073/pnas.1713830114