Canola SNPro™

JUNE 2022

Sharon Reikhav, Nir Kfir, Gil Ben Zvi, Kobi Baruch, Evan Staton, Masood Rizvi, Edial Risseeuw, Daphna Tako and Arie Zackay

ABSTRACT

Canola SNPro™ is a cost-effective solution for high-density genotyping of Brassica napus to support Genomic Selection and other molecular breeding methods. It includes a designed set of 500 SNPs and an imputation pipeline customized to produce highly accurate and reliable genotype data that conform to the current standard 60K Canola panel. Canola SNPro™ was designed by filtering the 60K SNPs for conservation and robustness using a Canola pan-genome of 12 de novo assemblies, adjusted for the diversity of two different publicly available data sets, and benchmarked by comparison with full data of 470 NAM samples. The reported results were empirically validated, with a call rate of 98% and median sample imputation accuracy of >95%.

INTRODUCTION

Genotyping is widely used in crop breeding to identify and select lines with favorable genetic content. The routine use of Genomic Selection (GS), a method that requires genotype information across the entire genome at high density, is growing and reaching ever more breeding programs and crops. Optimizing a genotyping strategy for GS and implementing the strategy successfully in a crop-breeding program requires extensive resources.

Rapeseed (Brassica napus subsp. napus), also known as rape, or oilseed rape, and commonly referred to as Canola plant when grown in Canada or Australia, is presently the third-largest oil crop and accounting for the production of over 72.3 million tons in 2020[1].

Rapeseed has an allotetraploid genome (2n = 4× = 38) with an estimated haploid size of 1.1Gbp[2] . The complex polyploid structure has resulted in abundant structural and copy number variations[3] , which has posed a challenge for accurate genotyping of genome-wide markers for association studies and genomic predictions. Many breeding programs utilize an existing microarray for B. napus consisting of ~60K SNPs[4]. However, the use of this array for commercial purposes is limited for several reasons:

1. A large proportion of the SNPs produce erroneous calls due to Presence Absence Variation (PAV), mispriming of the probes on the array, or picking up additional loci in the genome. The recent ability to produce multiple reference sequences (commonly referred to as pan genomes) has shed light on the extent of this phenomenon while also suggesting the tools to resolve it.
2. Many of the SNPs will be monomorphic and thus uninformative in a given breeding population.
3. The high-throughput genotyping of a dense marker set culminates in a high cost per sample. As a result, smaller SNP subsets have been generated from the 60K set tailored to the needs of the different rapeseed breeding programs. However, these smaller arrays are often still cost-prohibitive and cannot easily be adapted to account for novel diversity.

Extensive application of data imputation has been increasingly used to account for these issues[5].

SNPer™ is NRGene's genotyping solution, which combines the genotyping of a small set of highly informative markers in the progeny samples, with the imputation to a high-density marker set sampled in the parental lines. SNPer™ is customized to the genetic diversity within an individual breeding program of any crop to maximize the genetic data gained from a minimum number of SNPs genotyped. Therefore the informativeness and fidelity of the genetic data is maximized, which translates into higher prediction accuracy while minimizing the cost.

NRGene recently launched Canola SNPro™, which features a generic version of the SNPer™ solution. The minimal panel is customized for Spring Canola and, similarly to SNPer™, combines low-density genotyping with high-density imputation to an industry-standard panel such as the 60K set or derivatives thereof. This report presents the development and the validation of Canola SNPro™.

---

[1] https://www.fao.org/faostat/en/#data/QCL [2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950721/ [3] https://www.nature.com/articles/s41477-019-0577-7
[4] https://pubmed.ncbi.nlm.nih.gov/28220206/ [5] https://pubmed.ncbi.nlm.nih.gov/32331884/

METHODS AND RESULTS

Reliable SNP selection

The 60K Canola SNPs were mapped to 12 B. napus reference genomes, previously assembled de novo by NRGene[6]. The 12 genomes were contributed by commercial and academic participants of the B. napus pan-genome consortium and consisted of 4 Winter and 8 Spring lines (Table 1).

Source	Type	Assembly size (Gbp)	N50 (Mbp)	N50 #sequences	N90 (Mbp)	N90 #sequences
GIFS/AAFC	Spring	1	24.08	14	1.78	68
GIFS/AAFC	Spring	0.98	14.75	18	1.05	97
Commercial	Winter	0.99	20.34	14	2.11	56
Commercial	Winter	0.97	22.9	13	1.98	55
Nuseed	Spring	0.98	10.66	27	0.71	148
Nuseed	Spring	0.99	22.87	15	3.58	57
Nutrien	Spring	1	24.37	15	1.64	68
Nutrien	Spring	1	18.75	17	2.42	59
Corteva	Winter	1	34.07	10	3.37	42
Corteva	Spring	1	22.71	13	2.31	59
Bayer	Winter	0.98	19.69	16	1.58	78
Bayer	Spring	0.99	21.12	16	1.7	65
	Average	**0.99**	**21.36**	**15.7**	**2.02**	**71**

Table 1. Assembly statistics of NRGene's B. napus pan-genome

Of the 60K SNP sequences, 39.5K were aligned to a single conserved genomic position in all 8 Spring Canola assemblies, while 33K passed the single locus criterion in the combined 12 Spring and Winter lines. This means that the selected SNPs exist (not absent) in all the reference sequences and are also specific for the A and C sub-genomes that constitute the B. napus tetraploid genome. The removal of the polygenic and absent SNPs from further analysis is important for reliable genotyping, accurate imputation, and ultimately genomic prediction.

Target and Minimal SNP sets simulations and design

Target Set definition

The allele frequency of the selected SNPs were determined in two datasets that were based on the 60K array: a Chinese semi-winter rapeseed dataset of 203 lines that were genotyped with a 24K subset[7], and an Australian/Canadian dataset of 61 lines genotyped with 36K SNPs[8]: 19.7K SNPs complied with a call rate threshold of >0.1 and a Minor Allele Frequency (MAF) cutoff of 0.05. This Target SNP Set was used as the target panel for the imputation and benchmarking processes described below. In the Canola SNPro™ process, the parental lines of a typical breeding program will be genotyped for these markers by means of an array or by whole genome sequencing. The high density 19.7K panel can also be used in any other project with the objective of high-resolution genotyping.

In silico validation using a semi-winter lines' data

The 19.7K Target Set was the input for the selection of a Minimal Set in the range of 100 to 1000 SNPs. The Minimal Set was selected using a proprietary algorithm that optimizes for marker informativeness and maximal Linkage Equilibrium between the SNPs to enable maximal coverage and accuracy of the imputation.

30 lines out of the 203 Chinese semi-winter rapeseed lines[7] were crossed in silico to produce 20,000 synthetic progeny lines (~400 unique in-silico populations with 50 progenies each). Recombination frequencies for the simulated crosses were based on a genetic map (courtesy of Dr. Isobel Parkin, AAFC). Two population types were generated to account for different breeding applications: Doubled Haploid (DH) for fully fixed material in a single generation, and F4 to simulate mostly inbred material in a Single Seed Descent or a pedigree breeding program.

The inputs for imputation were the Target Set genotypic data of the parents, the Minimal Set genotypic data of the progeny, and the pedigrees that link each progeny to the parental lines. The imputation accuracy was calculated by comparing the Target Set genotypes to the imputed genotypes from the Minimal Set for the simulated DH and F4 progenies.

[6] https://www.nrgene.com/press-release/pan-genome-consortium/ [7] https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6711-0

Figure 1 shows that 96% of the Target Set SNP genotypes were accurately imputed on the simulated F4 lines when the Minimal Set contained 500 SNPs. As expected, the accuracy was higher in fully homozygous DH progenies where a median value of 97% was observed for the same 500 SNPs in the Minimal Set.

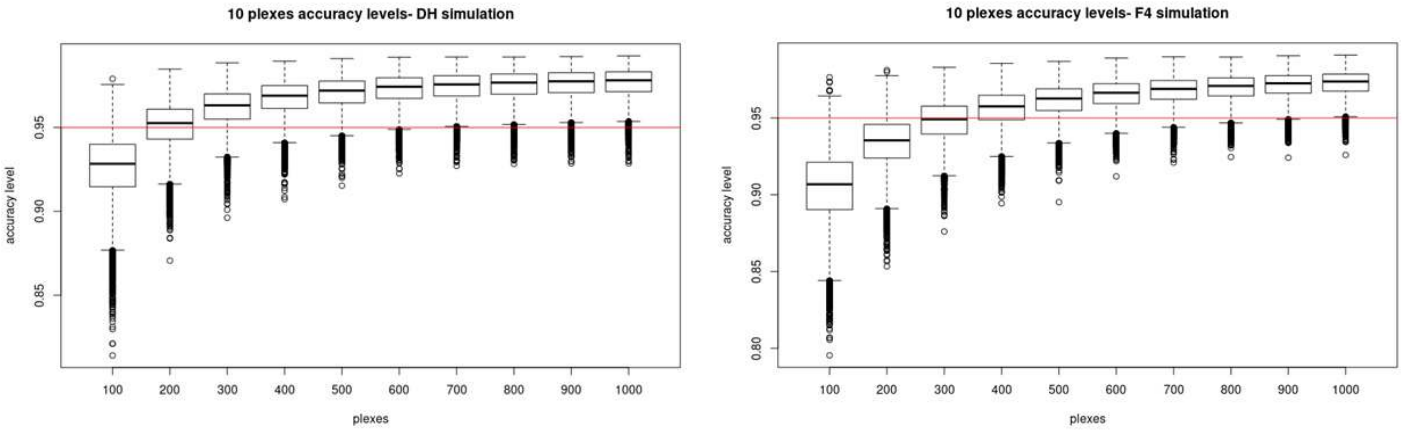

Figure 1: Imputation accuracy simulations with different Minimal Set sizes

In silico validation using spring canola lines' data

A minimal SNP set of 500 SNPs was selected to optimize cost and imputation accuracy. This Minimal Set was used for all further results. 48 parental samples including 18 Winter and 27 Spring lines[8] were crossed in silico to produce 2,256 progeny from all possible 1,128 combinations. Simulations were done as described for the Chinese populations. The median imputation accuracy is 98.2%, and 99.9% of the samples displayed an imputation accuracy level greater than 95%.

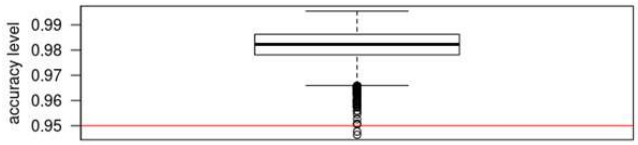

Figure 2. Imputation accuracy simulations of Canola lines (DH)

Minimal panel development

The 500 selected SNPs were used to develop an amplicon sequencing plex (PlexSeq™). Figure 3 displays the distribution of the selected SNPs along the B. napus reference sequence. If needed, a larger ~1000 set can be used to increase the robustness and accuracy of genotype calls.

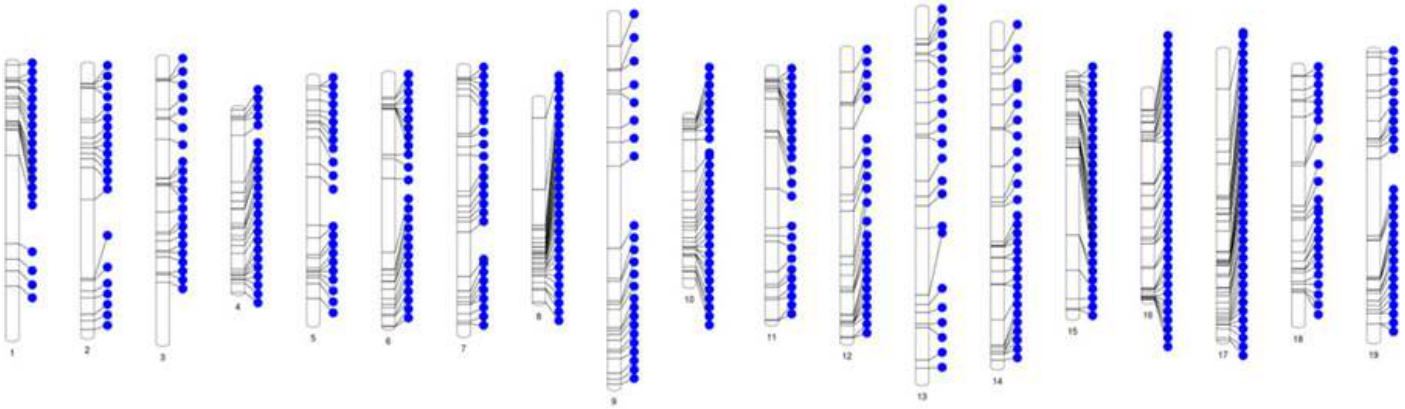

Figure 3. SNP distribution along the B. napus physical reference

[8] https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.14194

Empirical benchmark study

To validate the design and approach empirically, a benchmark set of 470 Recombinant Inbred Lines (RILs) from 20 out of the 50 Spring B. Napus Nested Association Mapping (NAM) populations[10] was used. The samples and their full genotype data were generously contributed by Dr. Isobel Parkin (AAFC).

Quality Check of pedigree data

The first step of SNPro™ analysis is to make sure pedigree and genotype data are consistent. As a result of this check, four of the samples showed significant inconsistencies between the pedigree and genotype data. Two of the four samples were matched with different parents and included in the analysis, while the two other samples were removed from further analysis leaving 468 NAM samples.

Call rates and agreement of datasets

A potential pool of 9,234,576 data points represents the complete set of genotypes that can be captured representing the 19.7k SNPs used as the Target Set. The call rate of the array data and the imputed set was 97% and 98%, respectively. This result is typical of the ability of a conservative imputation process to salvage No calls that are a technical artifact of the array. Table 2 shows the agreement between the different calls of the two genotype calls.

		Array				
		Total	NC	AA	BB	AB
SNPro	Total	9,234,576	261,743	4,355,077	4,348,269	269,487
	NC	190,010	1.4%	0.3%	0.3%	0.1%
	AA	4,379,529	0.6%	44.9%	1.3%	0.5%
	BB	4,374,567	0.7%	1.3%	44.9%	0.5%
	AB	290,470	0.2%	0.6%	0.6%	1.8%

Table 2. 60k and SNPro™ agreement. Rows represent the number of resulting calls made by Canola SNPro™ and columns represent the array calls. NC= No Calls, AA BB AB represent homozygous ref, homozygous alt, and heterozygous calls. The data points are distributed on the -2dimensional matrix such that the sum of data points for which both sources made the same call would be on the diagonal of the matrix.

As shown in Table 2, a total of 93% of data points were consistent. The imputation pipeline could also rescue 51% of the data points that were not called by the array. Only 32% of the No Call data obtained by SNPro™ had a genotype call in the 60k array. Imputation accuracy is higher in homozygous loci than in heterozygous loci. The overall imputation accuracy for the entire set was 94.2% (excluding the SNPs that had a No Call in the array data).

Per sample imputation accuracy

Genomic prediction models are calculated to assign a predicted phenotype per sample. It is therefore of vital importance that the genotype data best represents the actual genotype of each sample and that as few samples as possible are excluded from downstream analysis. Assessing the imputation accuracy per sample is therefore the most useful metric to quantify the applicability of the genotype data. Figure 4 represents the imputation accuracy of 468 NAM samples. The median accuracy was 95.6%, and only 26 samples fell below 90% accuracy.

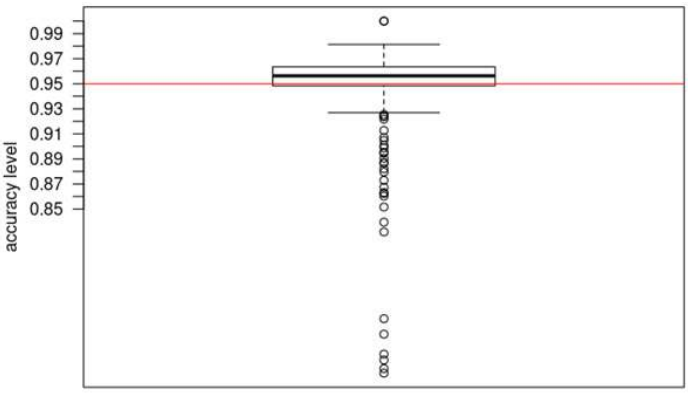

Figure 4. Imputation accuracy of 468 NAM samples

SUMMARY

This white paper details the process of designing and validating a SNPro™ genotyping solution for B. napus based on filtering and optimization of the Canola 60k panel. The SNP set was filtered by aligning to 12 reference genome assemblies and adjusted to the call rates and allele frequency of 2 different B. napus populations to generate a robust target set of SNPs that can be utilized for genomic predictions. A minimal SNP set of 500 SNPs was validated through simulations, to produce an imputation accuracy of 95-98% for different population types. Empirical validation using 468 NAM samples yielded a 1% increase in call rate and 95.6% per sample consistency (accuracy) compared with the full 60K array data of the same samples. The Canola SNPro™ Minimal Set and imputation can be utilized to fit any B. napus population (including winter types) by adjusting the Target Set as shown for the two populations in this study.

[9] https://agriplexgenomics.com/Portals/0/APG%20process%204%20web%20103020.pdf
[10] https://www.frontiersin.org/articles/10.3389/fpls.2021.780250/full