ddRADseq -mediated detection of genetic variants in sugarcane

Catalina Molina (molina.catalina@inta.gob.ar)
INTA: Instituto Nacional de Innovacion y Transferencia en Tecnologia Agropecuaria
https://orcid.org/0000-0003-1103-2748

Pablo Alfredo Vera
INTA: Instituto Nacional de Tecnologia Agropecuaria

Natalia Cristina Aguirre
INTA: Instituto Nacional de Tecnologia Agropecuaria

Carla Valeria Filippi
INTA: Instituto Nacional de Tecnologia Agropecuaria

Andrea Fabiana Puebla
INTA: Instituto Nacional de Tecnologia Agropecuaria

Susana Noemí Marcucci Poltri
INTA: Instituto Nacional de Tecnologia Agropecuaria

Norma Beatríz Paniego
INTA: Instituto Nacional de Tecnologia Agropecuaria

Alberto Acevedo
INTA: Instituto Nacional de Tecnologia Agropecuaria https://orcid.org/0000-0002-6415-7504

Research Article

Keywords: Genotyping by sequencing, Single nucleotide polymorphism, Saccharum hybrids, Polyploid genome, Sugarcane sequencing

Posted Date: April 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1546552/v1

Abstract

Sugarcane (*Saccharum sp.*), a world-wide known feedstock for producing sugar, bioethanol, and energy, has an extremely complex genome due to its highly polyploid and aneuploid nature. A double digestion restriction site-associated DNA sequencing protocol (ddRADseq) was tested in four commercial sugarcane hybrids and one high-fibre biotype for the detection and potential application of single nucleotide polymorphisms (SNPs) in genetic breeding. This genotyping approach compared two Illumina sequencing platforms and different filtering schemes, with and without a reference genome. A greater number of reads were obtained with Illumina platform Novaseq6000 than with Nextseq500, being this consistent with a greater number of SNPs determined with high accuracy. Additionally, more SNPs were found using a denovo pipeline compared to de refmap pipeline of the Stacks software. Longer read size, paired-end reads and 4M reads per individual represent the most efficient combination in recovering SNPs. The optimal combination of filter parameters varied depending on the different matrices created based on the different platforms. In both matrices, the highest number of SNPs localized in the longest chromosome 1, whereas the fewest landed in the shortest chromosomes 5 and 8. Multivariate comparisons of the SNPs matrices showed closer relationships among commercial hybrids than with the high-fibre biotype. Functional analysis of the obtained SNPs, performed with the Variant Effect Predictor, demonstrated the appearance of variants throughout the sugarcane genome. The pipeline applied in this study can be exploited to identify useful molecular markers that can be used in sugarcane breeding programs to reduce selection cycles.

Introduction

Sugarcane (*Saccharum sp.*) is a major economic crop in the world, with a global production of about 166.18 million tons in the 2019/2020 agricultural season (Walton 2020). In Argentina, it is the most important industrial crop with 381,138 ha cultivated and the main source of sugar (INDEC 2018). It is grown in the northwestern provinces, mainly in Tucumán (225,480 ha), but also in Jujuy (113,241 ha) and Salta (34,233 ha), according to the records of the 2018 National Agricultural Census (INDEC 2018). There are a few hectares in the east of the country, in the provinces of Santa Fe (2,330 ha) and Misiones (2,270 ha) (INDEC 2018).

The production of first-generation bioethanol (1G) through the fermentation of juices or molasses is becoming increasingly important. In fact, the national production in 2018/2019 was 16.8 million tons of sugar and 4.7 million tons of first-generation bioethanol (1G) (García et al. 2020; Benedetti 2018). Sugar-based bioethanol production as a second-generation (2G) biofuel has attracted interest as a renewable energy and represents a promising alternative to conventional fuels: Twenty-eight percent bagasse is produced per ton of sugar, which can be used to generate heat and electricity (Bottcher et al. 2013). Although there are only 1G biorefineries in Argentina, the prospects for the introduction of 2G industries are promising.

Sugarcane has one of the most complex genomes in crops due to its size (ca. 10 Gb), high degree of ploidy, and allopolyploid nature (Yang et al. 2017). It is also an aneuploid crop, which is one of the reasons why total chromosome number varies between genotypes (Vieira et al. 2018; Thirugnanasambandam et al. 2018). Modern and cultivated sugarcane hybrids in Argentina have five of the six species that integrated the genus Saccharum into their genealogy: *S. barberi* (2n = 116–120), *S. officinarum* (x = 10; 2n = 80–120), *S. robustum* (x = 10; 2n = 60–80), *S. sinense* (2n = 80–124) *and S. spontaneum* (x = 8; 2n = 40–126) (Acevedo et al. 2017). Acevedo et al. (2017) studied the genealogies of 16 cultivars used in Tucumán province (nine Argentine and seven American) and found that all of them differ in the proportion of founder species. They also confirmed that *S. officinarum* is the largest contributor, followed by *S. barberi*, and *S. spontaneum*. In addition, LCP 85–384, the most cultivated genotype (67.7%, according to Ostengo et al. 2021) in Tucumán, has 3 sorghum genotypes in its genealogy *(Sorghum bicolor cv Wiley, TADA and Rex)* (Acevedo et al. 2017).

This complexity directly affects the rearrangement of chromosomes during meiosis and generates additional variation through the gain or loss of genetic material (Glyn et al. 2004; Vieira et al. 2018). This high genetic variability can certainly be used to develop new genotypes with improved industrial traits, such as increasing sugar content and improving digestibility of sugarcane bagasse. However, conventional breeding requires years of testing, both in greenhouses and in experimental fields before a new hybrid can be released. Moreover, breeding programs usually evaluate morphological and agronomic traits, which are prone to error given the environmental influence on vegetative traits (Medeiros et al. 2020). In this context, the use of molecular assisted breeding would benefit crop improvement, not only by shortening the breeding cycle, but also by improving knowledge of accessions used for crosses.

Several years ago, the initial study of plant genomes by low-throughput molecular markers, such as RFLP, RAPD, AFLP, SSR and EST, provided some insights into the organisation and content of genomes, as well as the functionality of genes and noncoding DNA and their precise location (Thirugnanasambandam et al. 2018). These markers, which are characterized by their accuracy, reproducibility, and in some cases neutrality, are also becoming increasingly important in genotype fingerprinting and genetic diversity assessment. However, the use of low-throughput markers in large genomic studies, results in very low coverage, which compromises the scope of studies (Fickett et al. 2018). With the advent of next generation sequencing (NGS) technologies, more efficient molecular markers have been developed, such as single nucleotide polymorphisms (SNPs), which can be easily automated and provide whole genome coverage. SNPs have a significant advantage over other types of markers in polyploids such as sugarcane because they are codominant, and widespread throughout the genome, and can determine the number of copies of each allele (gene dosage). Gene dosage is in turn thought to be involved in variation in gene expression and thus phenotypic variation caused by the relative abundance of each allele (Garcia et al. 2013). Implementation of SNPs and phenotypic data helps identify correlations between these variants and quantitative trait loci, associations that can be potentially useful for future sugarcane breeding programs.

Double-digest restriction site associated with DNA sequencing (ddRADseq), is one of the many technologies included in the so-called Genotyping by sequencing (GBS) (Peterson et al. 2012). This protocol involves the use of two different restriction enzymes and the selection of a specific fragment size. This method can identify and simultaneously genotype thousands of loci distributed throughout the genome, providing a genomic perspective to the genotypes analysed. It also provides high reproducibility, large loci depth coverage, and selection of effective single nucleotide polymorphisms (SNP) distributed across the genome (Aguirre et al. 2019). For polyploid crops, detecting true SNPs is more challenging than for diploid crops due to subgenomes within genotypes. However, many tools developed for diploids can also be applied to allopolyploids because they behave similarly to diploid species (Bourke et al. 2018). Nevertheless, quality control and filtering are important steps in bioinformatic analysis to detect false positive SNPs.

Herein, we modified the ddRADseq protocol, originally developed by Peterson et al. (2012) and optimized for non-model diploid species by Aguirre et al. (2019), to apply it to a highly polyploid species to identify genomic variants for breeding applications in modern and cultivated Argentine sugarcane hybrids. Our working hypothesis is that for ddRADseq protocols applied to complex genomes, SNPs recovery is affected by read size, number of reads per individual, and single-reads versus paired-end reads. In this article, we tested sequences from two Illumina sequencing platforms, with a reduced number of samples to evaluate the above conditions. In addition, various bioinformatic filtering schemes were discussed to determine the application of parameters that balance the rate of missing data and the accuracy of detected SNP.

Materials And Methods

Plant material

Five sugarcane (*Saccharum* spp.) accessions from the germplasm collection of the National Institute of Agricultural Technology at the Famaillá Research Experimental Station, Famaillá (27°03’S, 65°25’W, 363 masl), Tucumán, Argentina, were selected for this study. Of the selected accessions, one was the high-fibre biotype INTA 05-3116, and four were the commercial hybrids NA 78–724, NA 56 – 30 (Argentine origin), LCP 85–384 and HOCP 92–665 (American origin), actively used as progenitors in the sugarcane breeding program due to their industrial and agronomic characteristics (Acevedo et al. 2017; García et al. 2021).

ddRADseq libraries

Leaves from three plants in each accession were used for isolation and purification of high-quality DNA using an EasyPure Plant Genomic DNA Kit (TransGen Biotech). DNA quality was checked using Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) and agarose gel electrophoresis analysis. DNA was quantified using a Qubit 2.0 fluorometer (Thermo Fisher Scientific). The ddRADseq protocol 1 optimised for *Eucalyptus dunnii* (Aguirre et al. 2019) was modified for sugarcane by using the *PstI/MboI* combination in the digestion step. A Fragment Analyzer was used for validating the DNA fragment size (a

range between 400 and 650 bp, representing 260–510 + 140 bp of barcodes) (Agilent Technologies, Santa Clara, USA).

The library was sequenced using both the Illumina Nextseq500 platform, here referred to as Ne-70 (75 bp single-end read trimmed at 70 bp; 4M reads/individual), and the Illumina Novaseq6000 platform, here referred to as No-150 (150 bp paired-end reads; 10M reads/individual). To evaluate the efficiency of the sequencing platforms in terms of read lengths, depth of coverage and single-reads vs. paired-end sequencing: (a) reads obtained in No-150 were shortened to 70 bp (while maintaining 10M reads/individual and paired-end, hereafter referred to as No-70-10M) and compared to No-150; (b) 4M reads/individual (No-70-4M) were randomly selected among the already shortened No-70-10M reads, in order to analyse the effect of the number of reads per individual; (c) finally, the selected No-70-4M were compared to the original Ne-70 to analyse paired-end versus single-end reads.

Data analyses

The bioinformatics workflow used to analyse the raw data is shown in Fig. 1. Quality was checked visually using FastQC (Andrews 2010) and summarised with MultiQC (Ewels et al. 2016). The program “process radtags” from the Stacks v 1.42 package (Catchen et al. 2011 2013 and 2016) was used for quality cleaning. Those reads whose average quality within a window fell below a Phred score of 10 (Ewing and Green 1998), or have un-called nucleotides, adapter sequences or deficient restriction enzyme cut sites were discarded (Catchen et al. 2011). Two strategies were followed for SNP calling: (1) reference guided and (2) de novo (Fig. 1).

For the first strategy, the cleaned reads were mapped with Bowtie2 under default parameters (Langmead and Salzberg 2012) against the genome of monoploid sugarcane cultivar R570 (Garsmeur et al. 2018) which was used as a reference, SNP identification was performed using the “ref_map.pl” pipeline (default parameters). In the second strategy, the “denovo_map.pl” pipeline of Stacks was used for SNP discovery, due to the absence of a well-annotated reference genome. Following both steps, the corresponding raw SNP matrix was obtained under Variant Call Format (VCF) using the “Population” program implemented in Stacks. The raw matrix statistics were calculated to determine subsequent filtering thresholds.

Filtering of markers was performed using VCFtools (Danecek et al. 2011). An initial and global filter of 0% missing data was applied to all data sets to ensure good and reliable data. Minor allele frequency (MAF) > 0.1, commonly applied to diploids, was used as a conservative threshold, while a minimum MAF of 0.3 was used as recommended for allopolyploids (Clevenger and Ozias-Akins 2015). Three different depth coverage ranges with minimum and maximum thresholds were investigated:

- 10x-36x: being 10x the usual recommendation for diploids (high confidence base-call) and 36x as maximum threshold (Ravinet and Meier 2020).
- 30x-60x: being 30x the usual recommendation for allopolyploids (Clevenger and Ozias-Akins 2015) and 60x as maximum threshold.

- 56x-200x: being 56x the threshold described by Yang et al (2017) for sugarcane and 200x as maximum threshold.

Six filtering schemes (FS) resulted from the combination of different parameters of MAF and read depth (Fig. 1).

Figure 1 Filtering schemes (FS) combining minor allele frequency (MAF) and read depth (RD)

SNP density per chromosome for Ne-70 with FS1 and No-150 with FS5 was plotted using R package CMplot (Yin et al. 2021). Correlation analysis of chromosome length and number of SNPs per chromosome was performed using “cor” and “cor.mtest” function (corrplot package) (Wei and Simko 2021). Correlation coefficients were considered significant for p-values < 0.05 (*).

A distance matrix between the samples of Ne-70 with FS1 and No-150 with FS5 was calculated using the R package vcfR, pegas, stats (Knaus BJ and Grünwald 2017; Paradis 2010; R Core Team 2019). According to the cophenetic correlation, the highest cophenetic correlation was obtained with the “average” method (best clustering model for both datasets) (Borcard et al. 2018).

SNP Effect Prediction

The Variant Effect Predictor (VEP) from the Ensembl Plants (version release 51) was used to predict the functional effect of SNP, following the instructions described in Ensembl/plant-scripts: Scripting analyses of genomes in Ensembl Plants, with the reference genome and its annotation, in FASTA and gff3 formats.

Results

Even though sugarcane is a relevant crop worldwide, little genomic information is available about its complex genome. The ddRADseq protocol implemented in this work allowed the generation of a library with pooled samples with good quality and quantity. The restriction enzymes chosen, *PstI/MboI*, were a successful combination for the digestion step and the manual size selection was able to recover the size of the DNA fragments of interest (Fig. S1). To investigate the performance of the ddRADseq protocol implemented for allopolyploid sugarcane genome we compared the results obtained using two sequencing platforms, considering the information content obtained, the best cost-benefit ratio, and different filtering schemes. Furthermore, the library accomplished the performance expectancy of the Illumina platforms (Fig. S2).

The number of reads and their alignment to the reference genome for each sample, in Ne-70 and No-150 platforms are shown in Table 1. The number of reads with the No-150 was on average more than seven times-fold the number detected with the Ne-70 (Table 1, Table S1), exceeding the expected 2.5:1 ratio. Because the same library was used in both assays, differential ratios could probably be due to sample nature, sequencer loading and clustering in cell flow. Samples NA 78–724 and NA 56 – 30 accounted for the highest and the lowest number of reads, respectively, in both protocols (Table 1). The overall alignment rate of the sequences against the monoploid sugarcane genome was higher in Ne-70 than in

No-150. These rates, which were similar across samples in the former platform, averaged 54%, and were in contrast with the rates of the latter platform (~ 42%), with the exception of sample NA 78–724, which reported 50% (Table 1).

Table 1
Number of sequenced reads (Nreads) obtained across samples with two platforms. Overall alignment rate of reads against the monoploid genome of sugarcane cultivar R570 (Garsmeur et al. 2018), number of reads with unique alignment, and number of reads with multiple alignments.

Sequencing platform	Accesion	Nreads	unique alignment reads	aligned more than once	overall alignment rate/Nreads	unique alignment reads/ Nreads
Ne-70	INTA 05-3116	3,055,298	1,182,532	430,432	0.528	0.387
	NA 78–724	3,759,023	1,499,488	569,465	0.550	0.399
	LCP 85–384	3,307,885	1,315,870	462,541	0.538	0.398
	HOCP 92–665	2,826,753	1,098,629	435,747	0.543	0.389
	NA 56 – 30	2,535,384	1,006,667	368,096	0.542	0.397
No-150	INTA 05-3116	10,504,705	4,162,735	2,208,809	0.415	0.396
	NA 78–724	13,056,991	5,696,204	2,484,897	0.504	0.436
	LCP 85–384	10,810,451	4,375,030	2,314,485	0.427	0.405
	HOCP 92–665	9,302,776	3,714,998	2,100,697	0.433	0.399
	NA 56 – 30	8,269,031	3,356,088	1,794,476	0.435	0.406

Sugarcane SNP matrices obtained with Stacks program

The raw SNP data matrices determined using the “Population” program from Stacks varied in Ne-70 and No-150 platforms. In the former, 188,199 SNPs for the *denovo_map.pl* and 86,051 SNPs in the *ref_map.pl* were recovered, while 2,831,922 SNPs for the *denovo_map.pl* and 460,890 SNPs in the *ref_map.pl* were identified in the latter (Table 2). The *denovo_map.pl* strategy allowed identifying an amount of SNPs 2.1 and 6.1 times larger than *ref_map.pl* in Ne-70 and No-150, respectively. Additionally, 15.0 times more SNPs with *denovo_map.pl* were recovered in No-150 compared with Ne-70, whereas 5.4 times more SNPs were recovered with *ref_map.pl* strategy.

Table 2
Number of SNPs per matrix obtained with the Stacks program. Matrices for de novo analysis or reference strategy with raw data and after applying six different filtering schemes (FS). MAF = minor allele frequency; RD = read depth.

Sequencing platforms and simulations	Strategy	Raw data	FS 1	FS 2	FS 3	FS 4	FS 5	FS 6
			MAF > 0.1	**MAF > 0.3**	**MAF > 0.1**	**MAF > 0.3**	**MAF > 0.1**	**MAF > 0.3**
			RD = 10x-36x		**RD = 30x-60x**		**RD = 56x-200x**	
Ne-70 4M/ind	reference	86,051	9,094	6,237	4,246	3,111	2,122	1,619
	de novo	188,199	14,360	9,206	3,222	2,193	723	555
No-150 10M/ind	reference	460,890	23,174	13,403	28,081	18,290	47,964	33,400
	de novo	2,831,922	30,167	16,925	15,568	9,102	5,430	3,392
No-70 4M/ind	reference	174,585						
	de novo	363,919						
No-70 10M/ind	reference	351,864						
	de novo	767,203						

SNP matrices obtained from trimmed No-150 reads, No-70-10M were compared with matrices obtained from No-150 reads to examine the effect of read length on the number of SNPs identified. The number of SNPs identified in No-150 with *denovo_map.pl* and *ref_map.pl* was 3.7 and 1.3 times the number of SNPs identified in No-70-10M, respectively. The minority of SNPs in *denovo_map.pl* were detected on the first 70 bp (767,203 SNPs in No-70-10M from 2,831,922 SNPs in No-150), whereas in *ref_map.pl* the majority of SNPs were located on the first 70 base pairs (351,864 SNPs in No-70-10M from 460,890 SNPs in No-150) (Table 2).

To examine the effect of the number of reads per individual, the No-70-10M matrices were compared with a random subsample of 4M reads per individual, No-70-4M. SNPs identified in No-70-10M doubled the amount of the ones identified in No-70-4M, and this was true for both SNP calling strategies (Table 2). This further suggests that the number of SNPs is not directly proportional to the number of reads per individual, because SNPs detected in No-70-4M represented almost 50% of the ones detected in No-70-10M instead of the 40% expected if it was.

Finally, to examine the effect of paired-end versus single reads on the number of SNPs identified, the original Ne-70 was compared to the No-70-4M subsample. As expected, the number of SNPs identified in No-70-4M were almost twice the number of SNPs identified in Ne-70, and this was true for both SNP calling strategies (Table 2).

Clearing SNP matrices

The raw SNP matrices generated from Ne-70 and No-150 datasets were filtered for quality, missing data, read depth, and MAF, as shown in Fig. 1. Twenty-four SNP panels were generated from six filtering schemes (FS). Ne-70 showed a decreasing trend of SNPs from FS1 to FS6 in both strategies (*denovo_map.pl* and *ref_map.pl*), indicating that the SNPs retained decreased with read depth as fewer SNPs fulfilled the restrictive conditions imposed by read depths 56x-200x (Table 2). No-150 showed a similar SNP recovery behaviour as Ne-70 in *denovo_map.pl*, but No-150 had the highest SNP recovery values in *ref_map.pl* with read depth 56x-200x (FS5 and FS6) (Table 2). At this depth, single-dose alleles could be recovered at a heterozygous locus, while at a shallower depth they could be lost. SNPs retained within each read depth were higher with MAF > 0.1 for both platforms. Since the application of FS1 in Ne-70 and FS5 in No-150 retained the highest number of SNPs, both filtering schemes were selected for further comparisons.

SNP density was plotted to visualise the distribution of SNPs along each chromosome (Fig. 2). Raw data from Ne-70 showed regions with more than 289 SNPs spread throughout the whole set of chromosomes, with distal regions showing relatively high SNP density, i.e., more than 181 SNPs (Fig. 2a). FS1 application upon the Ne-70 platform diminished dramatically SNP density in the entire set of chromosomes (Fig. 2b), being this reduction consistent with the differential number of SNPs retained in FS1 compared to raw data (Table 2). Raw data from No-150 displayed regions with more than 1621 SNPs spread throughout the whole set of chromosomes (Fig. 2c). FS5 application upon the No-150 platform revealed a lower distribution of SNPs in pericentromeric regions compared to distal ones (Fig. 2d).

Figure 2 SNPs distribution through 10 chromosomes of sugarcane cv R570 a. Ne-70 raw data; b. Ne-70 with FS1; c. No-150 raw data; d. No-150 with FS5 (one Mb window). The colour legend scale indicates the number of markers within a one Mbp window

The chromosome length and the number of SNPs per chromosome were positively correlated in both matrices, Ne-70 with FS1 (rho = 0.94; p-value < 2.2e-16) and No-150 with FS5 (rho = 0.96; p-value < 2.2e-16), (Fig. S3). Thus, chromosome 1, the longest chromosome, accounted for 1545 SNPs in the former matrix (Table S2) while 8850 in the latter one (Table S2). Conversely, chromosomes 8 and 5, the shortest ones, showed 526 and 573 SNPs respectively in Ne-70 with FS1, whereas 2702 and 2534 SNPs respectively in No-150 with FS5 (Table S2).

Hierarchical Cluster Analysis and Principal Component Analysis

Dataset from *ref_map.pl* filtered matrices responsible for the highest numbers of SNPs (Ne-70 FS1 and No-150 FS5, Fig. 2b, d) were analysed according to two different multivariate methods, one based on clustering whereas the other on ordination on reduced space. Cluster Analysis of the *ref_map.pl* filtered matrices (Table S3) resulted in two dendrograms that showed distinct clustering arrangements (Fig. 3a, b). A single-member-cluster was composed of INTA 05-3116 (high fibre biotype) in both dendrograms (Fig. 3a, b). A second cluster, formed by four commercial hybrids: NA 78–724, LCP 85–384, HOCP 92–

665, and NA 56 − 30, was also detected in both dendrograms, although different distances were determined among the hybrids in each dendrogram (Fig. 3a, b).

Figure 3 Relationship among sugarcane cultivars based on SNPs markers. Average agglomerative clustering distance in (a) Ne-70 with FS1 and (b) No-150 with FS5. Plot of principal components in (c) Ne-70 with FS1 and (d) No-150 with FS5. PC, principal component

Principal Component Analysis (PCA) allowed the graphical representations of the relationships among the 5 sugarcane genotypes (Fig. 3c, d). In the PCA plot, a summary of total variation of the SNPs presented by the first (PC1) and second (PC2) principal components explained 56% of total variance in both plots. PC1 contributed to 32% of total variance and distinguished two groups of genotypes (encircled in solid line). Group 1 was composed of INTA 05-3116, whereas group 2 was composed of the four commercial hybrids. PC2 accounted for 24% of total variance separating LCP 85–384, NA 56 − 30, and HOCP 92–665 (upper left quadrant) from NA 78–724 (lower left quadrant) for Ne-70 FS1 (Fig. 3c) whereas NA 56 − 30 and HOCP 92–665 (upper left quadrant) from NA 78–724 and LCP 85–384 (lower left quadrant) for No-150 FS5 (Fig. 3d). Taken together, hierarchical cluster analysis and PCA revealed similar patterns of genotype clustering and ordination on reduced space.

SNP Effect Prediction

To evaluate the potential implications of SNPs on the five sugarcane genotypes assayed the Ensembl Plants Variant Effect Predictor (VEP) was used (Fig. 4). Out of the total variants processed in Ne-70 FS1, downstream plus upstream gene variants hit 60.6%, standing for more than half the variants detected (Fig. 4a). Intergenic and intronic variants reached 2,070 (14.9%) and 1,314 SNPs (9.5%), respectively (Fig. 4a, Table S4). Almost 13% of the variants (1,776 SNPs) corresponded to coding regions (Fig. 4a, Table S4). From the latter number of SNPs, 894 caused changes in amino acids (missense variant, 50.3%), 868 caused no resulting change to the encoded amino acid (synonymous variant, 48.9%), 5 avoided the onset of transcription (start lost, 0.3%), and 5 caused stop codons to be won (0.3%) and 4 to be lost (0.2%) (Fig. 4b, Table S4).

Figure 4 Percentage of SNPs according to position relative to genes and biological effect. SNP variants are categorised and percentages are given for all consequences in **a.** Ne-70 with FS1; or **c.** No-150 with FS5, and consequences in coding regions according to genome site location in **b.** Ne-70 with FS1 or **d.** No-150 with FS5

Downstream plus upstream gene variants in No-150 FS5 accomplished a similar percentage to Ne-70 FS1 (61.4%) (Fig. 4c). Intergenic and intronic variants accomplished 9,834 (13%) and 9,662 SNPs (12.7%), respectively (Fig. 4c, Table S4). Around 10% SNPs were localised in coding regions representing 7,756 variants. From this amount: 1) 4,071 SNPs were missense variants (52.5%), with possible implications in protein functionality; 2) 3,586 SNPs were synonymous variants (46.2%); 3) 19 SNPs caused loss of start codons; and 4) 80 SNPs modified stop codons (55 were gained, 14 were lost, and in 11 codons at least one base was changed, but the terminator codon remained) (Fig. 4d, Table S4).

Discussion

Sugarcane genomics analysis faces major hassles due to the gain and loss of genetic material tightly associated with the highly polyploid and aneuploid nature of the crop. Genotyping sugarcane is not only complex because of the multiple number of allele copies, but also because of the aneuploidy and unknown ploidy of some hybrids (Serang et al. 2012). Furthermore, in polyploids the false-positive rate of SNPs detection rises due to ambiguous mapping of highly similar homeologous loci, becoming challenging to the precise SNPs detection (Clevenger et al. 2015). GBS has been proved to be a successful way to analyse sugarcane (Balsalobre et al. 2017; Yang et al. 2017), however, to our knowledge ddRADseq approach has never been applied to this crop before. Our simulations and comparisons showed that, regardless of the complexity of genotyping sugarcane hybrids, there is consistency in our data results obtained by the application of the ddRADseq protocol.

Current NGS platforms generate data for analysing genes and genomes, through sequencing by synthesis chemistry (López de Heredia 2016). Differences between platforms rely on the output range, reads per run, maximum read length and costs (Illumina 2021). For this reason, the comparison between sequencing platforms is crucial to decide how to invest available resources. Our results showed that the number of reads assigned to each genotype followed the same order in Ne-70 and No-150 platforms, although the final number of reads in the latter one was on average several times the number detected in the former one (Table 1). It is worth noting that NA 78–724 and NA 56 – 30, two Argentinean hybrids coming from the same sugarcane breeding program, stood for extreme numbers of reads in both sequencers, suggesting a consistency in sequencer performance among samples.

The overall alignment rate of both platforms averaged ~ 48%. Several non-mutually exclusive phenomena might be responsible for this value: First, the hybrid R570 reference genome assembly originated from synteny with sorghum, using BACs based on their global alignments with this related species and representing the gene-rich part of its genome (Garsmeur et al. 2018). This indicates that fewer conservative regions are unlikely to align with other sugarcane hybrids. Second, R570 is a cultivar from Reunion Island, located in the Indian Ocean, suggesting that it might be genetically distant from the Argentine and American genomes assayed in this investigation (Acevedo et al. 2017; Garsmeur et al. 2018); then the alignment software was developed for diploid species, representing a technical drawback for a highly polyploid and aneuploid species such as sugarcane along with computational problems. These observations increase the difficulty in locating the reads in the exact position of the reference genome even though Yang et al. (2017) reported that Bowtie 2, was more effective than other aligners in sugarcane. Interestingly, the high fibre biotype and the sugarcane cultivars showed similar alignment rates, independently of the platform used (Table 1).

Stacks software provides two types of analyses for SNP recovery: *de novo* and reference based (Rochette and Catchen 2017). The *de novo* analysis allowed a higher rate of SNPs detection than the reference-based analysis in both platforms, although the latter pipeline let us know the SNP position in the

chromosome and thus further information could be acquired. In addition, more raw data SNPs were recovered using *de novo* strategy compared to reference-based strategy.

Simulations applied to further unravel the role of the read length, number of reads and paired-end versus single-reads impact on the number of called SNPs have demonstrated that the SNPs are not evenly distributed along the read length. This is supported by the fact that the minority of the SNPs in *denovo_map.pl* (767,203 out of 2,831,922), but the majority in *ref_map.pl* (351,864 out of 460,890) landed on the first 70 base pairs. Thus, even though more SNPs were recovered in the longest reads (No-150) than in the shortest one (No-70-10M) for both mapping strategies, the *de novo* strategy was responsible for the greatest difference in SNPs recover (3.7 vs. 1.3 in *ref_map.pl*). Additionally, the number of SNPs is not directly proportional to the number of reads per individual. In relative terms, the amount of SNPs recovered in No-70-4M exceeded half the amount recovered in No-70-10M, suggesting that assigning 4M reads/individual is more efficient than 10 M reads/individual.

The lack of information on the application of the ddRADseq protocol in sugarcane moved us to test six filtering schemes (FS) composed of a combination of different MAF and read depths in order to select the most suitable filtering parameters values for improving SNP calling. The inclusion of minimum and maximum depth filters was recommended for depth coverage (Ravinet and Meier 2020). While a minimum depth cut-off removes false positive calls and ensures higher quality calls, a maximum cut-off tends to avoid paralogous or repetitive regions (Ravinet and Meier 2020) that have been recently reported in the sugarcane genome (Trujillo-Montenegro et al. 2021). The combination of the same MAF with increasing minimum and maximum depths cut-offs diminished the SNPs recovered in both Ne-70 mapping strategies and in No-150 *denovo_map.pl*. Nevertheless, in No-150 *ref_map.pl* the combination of the same MAF with increasing read depths augmented the SNP call. This could be related to paralog regions coming from the duplication events or polyploid nature where we can find several copies of some regions (or genes) within the same genome that belong to different ancestors (subgenomes). In addition, in highly polyploid genomes, such as sugarcane which can achieve a ploidy level of twelve, 56x was the estimated sequence depth for single dose markers to be called, while in lower depth sequencing, a single dose SNP of heterozygous genotype could be called as homozygous genotype (Yang et al. 2017). A series of best practices for SNP calling in polyploids would imply the combination of the highest number of reads, the highest MAF, and the highest depth (Clevenger et al. 2015; Yang et al. 2017).

Distal regions along each sugarcane chromosome exhibited relatively high SNP density both in raw data and, to a minor extent, in FS1 and FS5 (Fig. 2) due to the differential number of SNPs retained in these FS compared to raw data (Table 2). These observations are consistent with the pattern observed in the sorghum reference genome assembly, where distal regions of chromosomes exhibited high gene density and pericentromeric regions displayed low gene density and low rates of recombination (McCormick et al. 2018).

Findings of hierarchical clustering and principal component analyses were in accordance both within and between matrices, highlighting the isolation of the high-fibre biotype from the commercial hybrids (Fig. 3),

likely due to their differential genetic backgrounds (García et al. 2021; Kane et al. 2021). Interestingly, NA 78–724 and NA 56 − 30, the two Argentinean hybrids genetically developed under the same sugarcane breeding program did not share a same cluster in any of the matrices analysed. Furthermore, NA 56 − 30 and HOCP 95–665 showed the nearest genetic distance in No-150 (Fig. 3b) among the samples assayed, in agreement with genetic similarities based on coefficient of parentage estimations (Acevedo et al. 2017).

To shed more light on the effect of the SNPs retained on genes, transcripts, protein sequences, and regulatory regions on the five sugarcane genotypes assayed, the variant effect predictor (VEP) was used. Overall, similar proportions of SNPs were assigned to the same functional consequences in both matrices (Fig. 4a, c). This further shows that independently of the differential number of SNPs detected in each platform, similar variant proportions were allocated to the very same consequence terms. Among the SNPs discovered, missense variants (6.4–5.4%, Fig. 4b, d) responsible for the change of one base resulting in a different amino acid sequence where the protein length is preserved, emerged as interesting variants to consider for further analysis. Some SNPs identified in the coding region, such as splice region variants (0.5% − 0.7%), refer to a change either within 1–3 bases of the exon or 3–8 bases of the intron, indicating that they have low impact for they are unlikely to change protein behaviour, even though they still represent a variation in DNA sequences. Similar considerations may be considered for synonymous variants (6.3% − 4.7%) as they are supposed to have no change to the encoded amino acid. More than half of the variants detected landed within regulatory regions, suggesting that they might alter the biological functionality of the gene (Fig. 4a, c). Finally, in non-coding variants or variants affecting non-coding genes, impact is difficult to predict or there is no evidence of impact at all.

Recently, a new genome assembly from the sugarcane hybrid CC 01-1940 has been published using NGS technologies (PacBio, Illumina Short Reads and HI-C Reads) (Trujillo-Montenegro et al. 2021). Our preliminary analysis using this brand-new reference genome showed that the overall alignment increased the mapping rate from ~ 54% to ~ 85% in Ne-70, demonstrating that the hybrid CC 01-1940 is genetically closer to our hybrids than the cultivar R570 (Garsmeur et al. 2018). Furthermore, the largest (chromosome 1) and smallest (chromosomes 5 and 8) sizes of chromosomes coincide in Trujillo-Montenegro et al. (2021) and in this investigation. Analyses of commercial and high-fibre sugarcane hybrids using the genome assembly of the hybrid CC 01-1940 as reference, are in progress (Manuscript in preparation). This underscores the fact that the pipeline considered in this study can be used regardless of any newer genome sequence version that might be released in the future. Furthermore, it is a good tool for detecting novel SNPs and, if compared to SNP array technology, it can be applied without previous knowledge on polymorphisms. Even more, this protocol can be applied specifically to the sugarcane population of the breeding program being assessed to discover molecular markers associated with agricultural features.

Declarations

Declarations

Conflict of interest

All authors declare that they have no conflict of interest.

Acknowledgments

We specially thank Sergio Gonzalez (Instituto de Agrobiotecnología y Biología Molecular, INTA-CONICET, Argentina), for the assistance and advice on bioinformatics and Giusi Zaina (Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Italy) for improving the quality of the manuscript through her critical reading. We are also grateful to Consejo Nacional de Investigaciones Científicas y Técnicas for supporting a Doctoral Fellowship to Catalina Molina, and the Unidad de Genómica, Instituto Nacional de Tecnología Agropecuaria, INTA, Argentina for laboratory collaboration. We thank the field team of INTA´s sugarcane breeding program for its technical assistance.

References

1. Acevedo A, Tejedor MT, Erazzú LE, Cabada S, Sopena R (2017) Pedigree comparison highlights genetic similarities and potential industrial values of sugarcane cultivars. Euphytica 213. https://doi.org/10.1007/s10681-017-1908-2
2. Aguirre NC, Filippi CV, Zaina G, Rivas JG, Acuña CV, Villalba PV, García MN, González S, Rivarola M, Martínez MC, Puebla AF, Morgante M, Hopp HE, Paniego NB, Poltri SNM (2019) Optimizing DDRADseq in non-model species: A Case Study in Eucalyptus dunnii Maiden. Agronomy 9 https://doi.org/10.3390/agronomy9090484
3. Andrews S (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Balsalobre TWA, da Silva Pereira G, Margarido GRA, Gazaffi R, Barreto FZ, Anoni CO, Cardoso-Silva CB, Costa EA, Mancini MC, Hoffmann HP, de Souza AP, Garcia AAF, Carneiro MS (2017) GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18:1–19. https://doi.org/10.1186/s12864-016-3383-x
5. Benedetti P (2018) Primer relevamiento del cultivo de caña de azúcar de la República Argentina a partir de imágenes satelitales para la campaña 2018. Inta 2–6
6. Bottcher A, Cesarino I, dos Santos AB, Vicentini R, Sampaio Mayer JL, Vanholme R, Morreel K, Goeminne G, Magalhães Silva Moura JC, Nobile PM, Carmello-Guerreiro SM, dos Anjos IA, Creste S, Boerjan W, de Andrade Landell MG, Mazzafera P (2013) Lignification in sugarcane: Biochemical characterization, gene discovery, and expression analysis in two genotypes contrasting for lignin content. Plant Physiol 163:1539–1557. https://doi.org/10.1104/pp.113.225250
7. Bourke PM, Voorrips RE, Visser RGF, Maliepaard C (2018) Tools for genetic studies in experimental populations of polyploids. Front Plant Sci 9:1–17. https://doi.org/10.3389/fpls.2018.00513

8. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3 Genes|Genomes|Genetics. 1:171–182. https://doi.org/10.1534/G3.111.000240
9. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. https://doi.org/10.1111/mec.12354
10. Catchen J, Cresko WA, Hohenlohe PA, Amores A, Bassham S (2016) Stacks Manual. Most, pp. 1–37. Retrieved from http://catchenlab.life.illinois.edu/stacks/
11. Clevenger JP, Ozias-Akins P (2015) SWEEP: A tool for filtering high-quality SNPs in polyploid crops. G3 Genes, Genomes. Genet 5:1797–1803. https://doi.org/10.1534/g3.115.019703
12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330
13. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354
14. Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 8(3):186–194
15. Fickett N, Gutierrez A, Verma M, Pontif M, Hale A, Kimbeng C, Baisakh N (2018) Genome-wide association mapping identifies markers associated with cane yield components and sucrose traits in the Louisiana sugarcane core collection. Genomics 0–1. https://doi.org/10.1016/j.ygeno.2018.12.002
16. Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry RJ, Souza GM, Van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep 3:1–10. https://doi.org/10.1038/srep03399
17. García JM, Molina C, Acevedo A, Silva M, Gómez LD, Erazzú LE (2020) Caña de azúcar en Argentina: situación actual y uso para fines bioenergéticos. Optimización de los procesos de extracción de biomasa sólida para uso energético. UP Valencia, España, pp 145–158
18. García JM, Silva MP, Simister R, McQueen-Mason SJ, ERAZZÚ LE, GÓMEZ LD, Acevedo A (2021) High variability for cell-wall and yield components in commercial sugarcane (Saccharum spp.) progeny contrasts with parental lines and energy cane. J Crop Improv. https://doi.org/10.1080/15427528.2021.2011521
19. Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, Jenkins J, Martin G, Charron C, Hervouet C, Costet L, Yahiaoui N, Healey A, Sims D, Cherukuri Y, Sreedasyam A, Kilian A, Chan A, Van Sluys MA, Swaminathan K, Town C, Bergès H, Simmons B, Glaszmann JC, Van Der Vossen E, Henry R, Schmutz J, D'Hont A (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat Commun 9. https://doi.org/10.1038/s41467-018-05051-5

20. Glyn J, Rod E, Merry B, Yates D, Yates W, Yates B, Digges P, Forber G, Todd M, Berding N, Cox M, Hogarth M, Bailey R, Leslie G, Irvin J (2004) Sugarcane, Second edi. Blackwel Publishing Company
21. Illumina (2021) International Narcotics Control Board 1999, United Nations, accessed 1 October 1999,
22. INDEC: Instituto Nacional de Estadísticas y Censos- Resultados definitivos - Censo Nacional Agropecuario (2018) Buenos Aires. Retrieved December 21, 2021, from https://www.indec.gob.ar/ftp/cuadros/economia/cna2018_resultados_definitivos.pdf
23. Kane AO, Pellergini VOA, Espirito Santo MC, Ngom BD, García JM, Acevedo A, Erazzú LE, Polikarpov I (2021) Evaluating the Potential of Culms from Sugarcane and Energy Cane Varieties Grown in Argentina for Second-Generation Ethanol Production. Waste Biomass Valorization 2021 1–15. https://doi.org/10.1007/S12649-021-01528-5
24. Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH (2016) Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci 242:14–22. https://doi.org/10.1016/j.plantsci.2015.04.016
25. Knaus BJ, Grünwald NJ (2017) Mol Ecol Resour 17(1):44–53 ISSN 757. http://dx.doi.org/10.1111/1755-0998.12549. “VCFR: a package to manipulate and visualize variant call format data in R.”
26. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
27. López de Heredia U (2016) Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. Munibe Ciencias Nat 64. https://doi.org/10.21630/mcn.2016.64.07
28. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, Mattison A, Morishige DT, Grimwood J, Schmutz J, Mullet JE (2018) The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J 93:338–354. https://doi.org/10.1111/tpj.13781
29. Medeiros C, Almeida Balsalobre TW, Carneiro MS (2020) Molecular diversity and genetic structure of Saccharum complex accessions. PLoS ONE 15:e0233211. https://doi.org/10.1371/JOURNAL.PONE.0233211
30. Ostengo S, Serino G, Perera MF et al (2021) Sugarcane Breeding, Germplasm Development and Supporting Genetic Research in Argentina. https://doi.org/10.1007/s12355-021-00999-z. Sugar Tech
31. Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26:419–420
32. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0037135
33. Ravinet M, Meier J (2020) Speciation & Population Genomics: a how-to-guide. Physalia Courses. Available: https://speciationgenomics.github.io/filtering_vcfs/

34. Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc 12:2640–2659. https://doi.org/10.1038/nprot.2017.123
35. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
36. Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE 7:1–13. https://doi.org/10.1371/journal.pone.0030906
37. Thirugnanasambandam PP, Hoang NV, Henry RJ (2018) The challenge of analyzing the sugarcane genome. Front Plant Sci 9:616. https://doi.org/10.3389/FPLS.2018.00616/BIBTEX
38. Trujillo-Montenegro JH, Rodríguez Cubillos MJ, Loaiza CD, Quintero M, Espitia-Navarro HF, Salazar Villareal FA, Viveros Valens CA, González Barrios AF, De Vega J, Duitama J, Riascos JJ (2021) Unraveling the Genome of a High Yielding Colombian Sugarcane Hybrid. Front Plant Sci 12. https://doi.org/10.3389/FPLS.2021.694859
39. Vieira MLC, Almeida CB, Oliveira CA, Tacuatiá LO, Munhoz CF, Cauz-Santos LA, Pinto LR, Monteiro-Vitorello CB, Xavier MA, Forni-Martins ER (2018) Revisiting meiosis in sugarcane: Chromosomal irregularities and the prevalence of bivalent configurations. Front Genet 9:1–12. https://doi.org/10.3389/fgene.2018.00213
40. Walton J (2020) The 5 Countries That Produce the Most Sugar. In: 5 Ctries. that Prod. Most sugar. https://www.investopedia.com/articles/investing/101615/5-countries-produce-most-sugar.asp. Accessed 22 Jan 2021
41. Wei T, Simko V (2021) R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92)
42. Yang X, Song J, You Q, Paudel DR, Zhang J, Wang J (2017) Mining sequence variations in representative polyploid sugarcane germplasm accessions. BMC Genomics 18:1–16. https://doi.org/10.1186/s12864-017-3980-3
43. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X (2021) rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-wide Association Study. Genomics Proteomics Bioinformatics. Mar 2:S1672-0229(21)00050 – 4. doi: 10.1016/j.gpb.2020.10.007. Epub ahead of print. PMID: 33662620

Figures

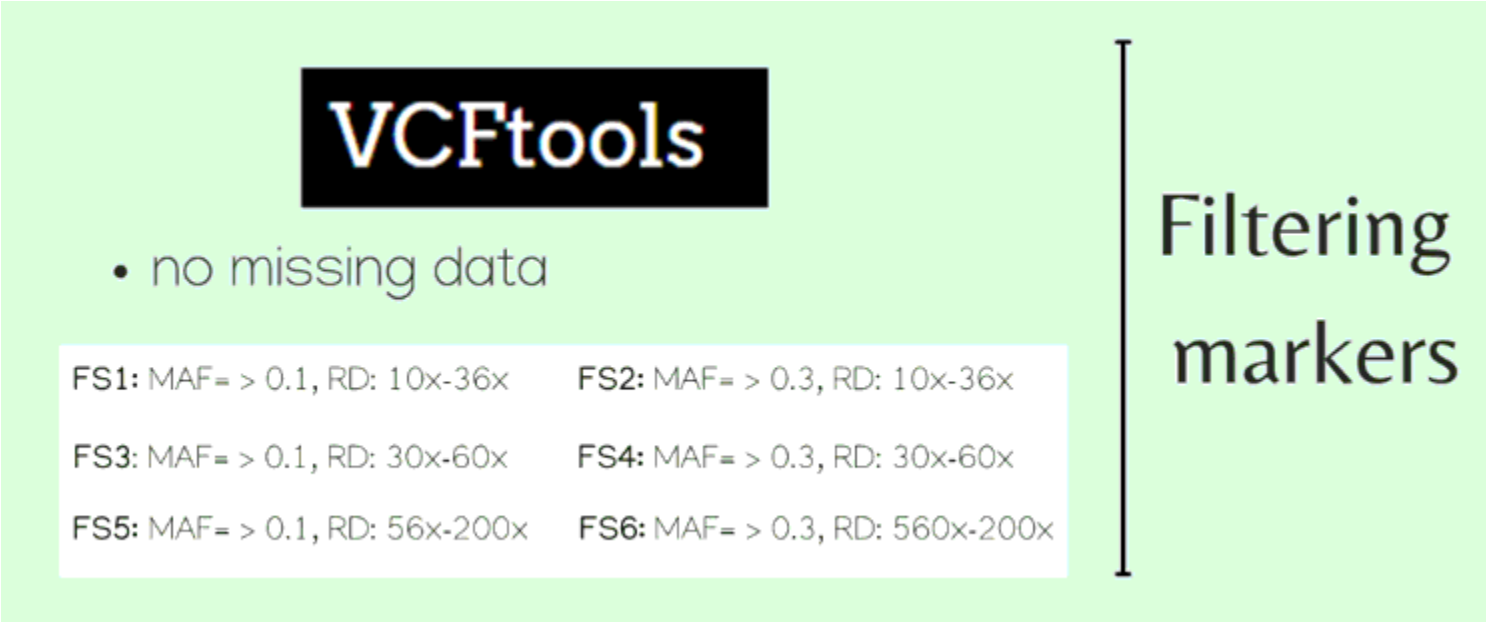

Figure 1

Filtering schemes (FS) combining minor allele frequency (MAF) and read depth (RD)

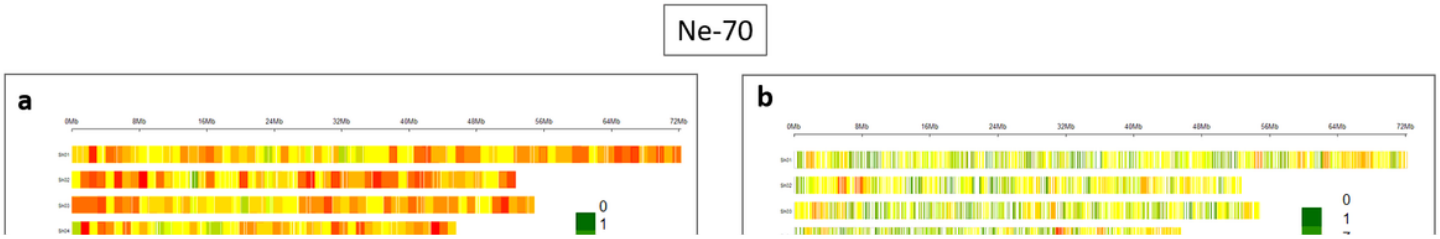

Figure 2

SNPs distribution through 10 chromosomes of sugarcane cv R570 a. Ne-70 raw data; b. Ne-70 with FS1; c. No-150 raw data; d. No-150 with FS5 (one Mb window). The colour legend scale indicates the number of markers within a one Mbp window

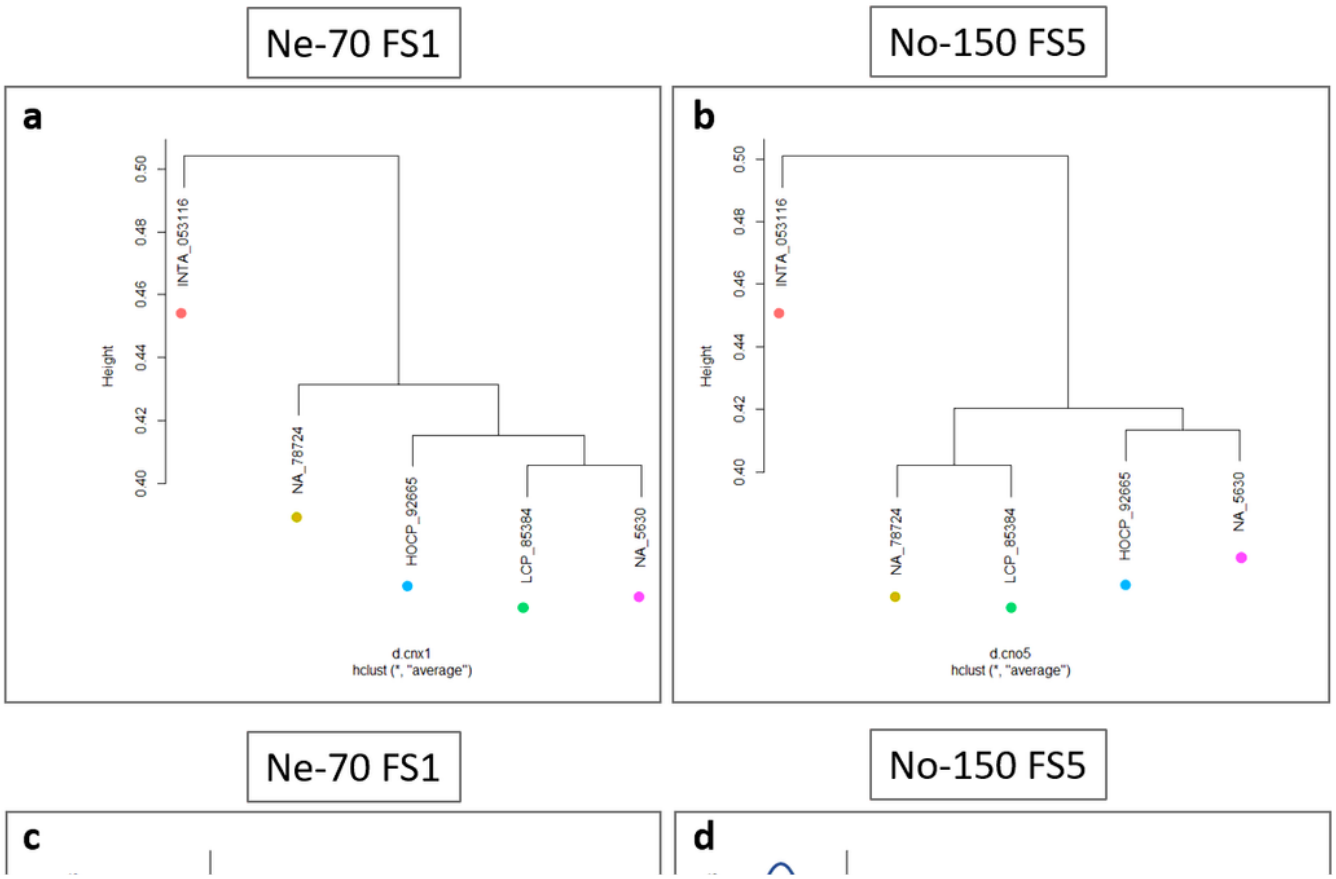

Figure 3

Relationship among sugarcane cultivars based on SNPs markers. Average agglomerative clustering distance in (a) Ne-70 with FS1 and (b) No-150 with FS5. Plot of principal components in (c) Ne-70 with FS1 and (d) No-150 with FS5. PC, principal component

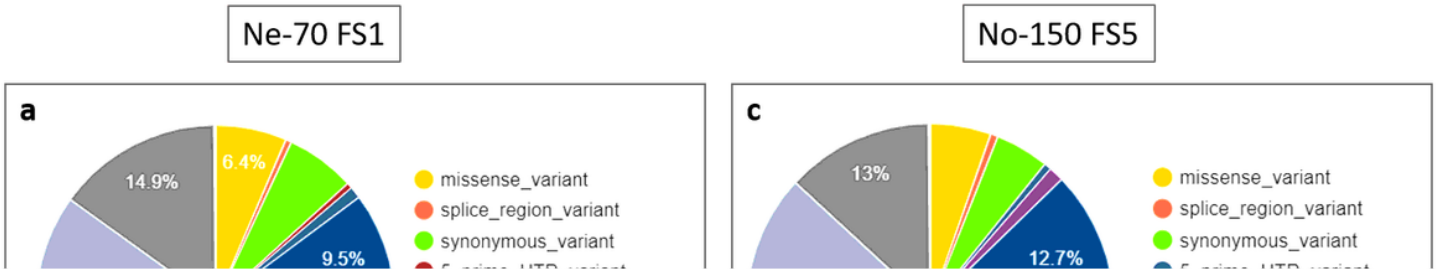

Figure 4

Percentage of SNPs according to position relative to genes and biological effect. SNP variants are categorised and percentages are given for all consequences in **a.** Ne-70 with FS1; or **c.** No-150 with FS5, and consequences in coding regions according to genome site location in **b.** Ne-70 with FS1 or **d.** No-150 with FS5

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryfileMolinaetal.pdf