Sugar Research Australia™

Development and testing of a SNP marker platform in sugarcane

Final report submitted to Sugar Research Australia

By:

Karen Aitken

Research Organisations:
Sugar Research Australia, CSIRO, Syngenta

May 2016

Researcher Contact Details

Name: Karen Aitken
Address: CSIRO Sustainable Ecosystems
Level 3 Queensland Bioscience Precinct
306 Carmody Road, St Lucia QLD 4067
Phone: 1300 363 400
Fax: N/A
Email: N/A
In submitting this report, the researcher has agreed to RIRDC publishing this material in its edited form.

SRA Contact Details
Sugar Research Australia
50 Meiers Road
Indooroopilly Q 4068 Australia
PO Box 86
Phone: 07 33313333
Fax: 07 38710383
Web: sugarresearch.com.au

This report is an addition to SRA's diverse range of research publications and it forms part of our Sugarcane Breeding program, which aims to increase optimally-adapted varieties.

Most of SRA's publications are available for viewing and free downloading online at www.sugarresearch.com.au.

Keywords: Single Nucleotide Polymorphism

How to cite: Aitken, K, Development and testing of a SNP marker platform in sugarcane, Sugar Research Australia, 2016.

Contents

1. Executive Summary

Marker assisted selection is being used in many industries to develop new and improved crops. This project sought to develop a selection tool for the Australian sugarcane industry by developing a set of DNA markers that could be used to enhance rates of genetic gain in the Sugarcane Breeding Program when compared to phenotypic selection alone. This was a collaborative project between CSIRO, Syngenta and SRA and has been proven successful in combining the skills of each organisation to achieve the best outcome. The extensive experience within Syngenta on other crops was used in combination with the sugarcane genetic knowledge within CSIRO to select the Affymetrix Axiom array system as the platform most likely to generate results on such a complex polyploid as sugarcane. This is the first time that this technology has been utilised on a species that has a complex high ploidy genome.

To date, genetic markers in sugarcane have lagged behind other crops. This is due to both its complex genetic structure and large genome size which makes generation of enough markers to cover the genome very laborious. The array development has removed this obstacle in sugarcane and made available a tool that can generate thousands of polymorphic markers in a single experiment. The marker of choice for all plant and animal selection programs has become the single nucleotide polymorphism (SNP). This is due to their presence throughout the genome in large numbers, robust detection and applicability to high throughput technologies.

To develop SNP markers in sugarcane for the Australian Sugarcane Breeding Program, this project sequenced the gene-rich regions of a selection of basic germplasm and important parents that encompasses the variation that exists in the germplasm. The sequence data was aligned to an existing sugarcane assembly generated in a previous project and SNP markers identified. To maximise the chance of identifying informative markers, the development of the array was carried out in a two-step process. Firstly a 400K array was developed and screened with 480 individuals from the association mapping population. The majority of these markers were non-informative, due to the ploidy in sugarcane where many of the markers are at higher dosage and thus not polymorphic in the breeding population.

The informative markers were used to create a highly informative 40K array, cheaper to screen than the 400K array. This was used to screen 1850 breeding lines for genomic selection analysis to test the gain in prediction accuracy from the addition of SNP marker information to the pedigree information. The results have been highly encouraging and show that SNP information has a much better predictive accuracy for breeding value than using pedigree information for all traits tested. The predictive accuracies of around 0.47 for CCS, and 0.3 for TCH, give high confidence that this method can be used effectively for parental improvement and lead to new varieties with increased CCS and yield.

This project has been very successful with the generation of the highly informative cane array which has already been shown to generate thousands of SNP markers in the Australian Sugarcane Breeding Program population. The project organised two workshops on marker development and implementation in sugarcane which were well attended by both the sugarcane breeders and molecular geneticists.

The workshop was successful in informing the sugarcane breeders about marker implementation and the potential for the use of the sugarcane array. The data from the array is already being used by the sugarcane breeders to select clones with high parent breeding value for inclusion in the breeding program.

This project has laid the foundations for the implementation of markers into sugarcane breeding, for the first time giving SRA an important additional tool to realise increased genetic gain in their breeding program. A great deal of important data has been generated along with new methods for analysis by experts in this field at SRA. The information from the SNP array and the analysis methods developed in this project will guide the implementation of markers in the future of sugarcane breeding. The cane array has other applications and is already being used in a new project to assay the diversity of the Indian breeding program based in Coimbatore, India.

The SNP markers identified in this project, linked to increased CCS and TCH, along with improved smut and pachymetra resistance, will be utilised in other projects. They will be validated and converted to different high throughput low-cost platforms to be implemented into the sugarcane breeding program in the next few years. Through genomic selection using these tools, new high yielding resistant sugarcane varieties will be developed faster.

2. Background

DNA markers can enhance rates of genetic gain in breeding programs and are being applied in many animal and crop species. Based on past research in Australia (particularly in CRC-SIIB), practical evaluation of marker-assisted breeding in sugarcane is now in progress, to measure realised genetic gains for parental improvement compared with (traditional) phenotypic selection.

DNA markers have been used in sugarcane to generate genetic maps, identify quantitative trait loci for a wide variety of traits and fingerprint clones but markers in sugarcane remains very challenging and still lags behind other species. In particular, the very large sugarcane genome means that many current "best bet" markers associated with agronomic traits of interest are unlikely close in linkage with underlying causal genes. This will limit gains from future applications of markers in sugarcane breeding.

In recent years, new DNA marker technologies developed in humans, animals and other crops can overcome these limitations. These technologies are based on single nucleotide polymorphisms (SNPs). SNPs are now the molecular marker of choice in animal breeding programs and important crops because massive numbers per genotype (e.g. >100,000) can be accurately screened for, and a proportion of these markers are usually directly within the genes causing genetic variation in traits of interest.

In preparation for this project proposal, significant resources were deployed in a pilot project in 2010/11 by CSIRO and Syngenta to test SNPs in sugarcane. This pilot study employed "next generation" sequencing technologies to generate large amounts of sequence data from the parents of an Australian genetic mapping population. Analysis of the data using novel bioinformatic tools, developed in collaboration with project partners, identified large numbers of potentially useful SNP markers which could be matured into a high density SNP genotyping platform. This project used this information to expand the SNP identification to encompass the diversity within the Australian Sugarcane Breeding Program and to generate a tool than could be used for a number of applications to assist in generating new sugarcane varieties.

3. Outputs and Achievement of Project Objectives

All the objectives for this project were completed. The goal of the project was to develop and generate a sugarcane SNP chip that could be used for a number of different applications including gene discovery and screening of parental lines for selection in a parental improvement program. This project had four main objectives.

3.1 Generation of sequence data from key ancestors of the Australian sugarcane breeding program to generate the SNP data for a 60K-90K SNP chip

3.1.1 Selection of ancestor clones

Identification of SNP markers relevant to the Australian Sugarcane Breeding Program must be identified within the Australian breeding germplasm pool. Through consultation with scientists from SRA, CSIRO and Syngenta a list of 16 clones were selected for sequencing. Due to their ancestry, these clones will contain the majority of the alleles of genes that are present in the current breeding program. As a result, sequence from these clones will contain SNP markers that are linked to traits of economic importance within the Australian Sugarcane Breeding Program. As the list contains some initial hybrids that were used in sugarcane breeding programs around the world, the SNP chip will also be relevant to other international breeding programs (Table 1).

Table 1. The 16 selected lines selected for sequencing and SNP identification

Clone	Saccharum species	Why it was selected	Correctly identified
Badila	*Saccharum officinarum*	One of the original *S. officinarum* clones used to generate the first hybrids and an ancestor of most sugarcane breeding programs around the world	Yes
CP74-2005	Sugarcane cultivar	A parent in the Australian breeding program and in some important genetic populations under study	Yes
NCO310	Early hybrid	An ancestor of many of the Australian top sugarcane varieties	Yes
Q117	Sugarcane cultivar	A modern parent in many of the Australian sugarcane varieties	Yes
Q208	Sugarcane cultivar	A cultivar grown widely in Australia and a parent in the breeding program as well as one of the parents of a genetic population under study	Yes
QN66-2008	Sugarcane cultivar	A common parent used in the Australian breeding program	Yes
QN80-3425	Sugarcane cultivar	A parent used in the Australian breeding program and a parent of one of the genetic population under study.	Yes
SP83-5073	Sugarcane cultivar	Brazilian clone which has been used as a parent in the Australian breeding program	Yes
POJ2878	Early hybrid	An early hybrid used in most of the breeding programs around the world	Rechecked and the correct line selected
QN58-829	Sugarcane	A parent used in the Australian breeding	Rechecked and the

	cultivar	program	correct line selected
SP70-1143	Sugarcane cultivar	A parent in the Syngenta breeding program	Rechecked and the correct line selected
Trojan	Sugarcane cultivar	An ancestor of the Australian breeding program previously a major cultivar	Rechecked and the correct line selected
Q155	Sugarcane cultivar	An ancestor of the Australian breeding program	Yes
Co475	Sugarcane cultivar	An ancestor of the Australian breeding program	Only one source (therefore n/a)
SP80-3280	Sugarcane cultivar	A cultivar and parent in the Syngenta breeding program	Only one source (therefore n/a)
RB72-454	Sugarcane cultivar	A cultivar and parent in the Syngenta breeding program	Only one source (therefore n/a)

All clones were sourced from Meringa, QLD and nuclear DNA preps carried out for sequencing (Figure 1). To verify these clones they were checked against clones sampled from another source in Macknade, QLD. Two SSR markers were run across the 11 lines that could be obtained from both Meringa and Macknade (Figures 2 and 3).

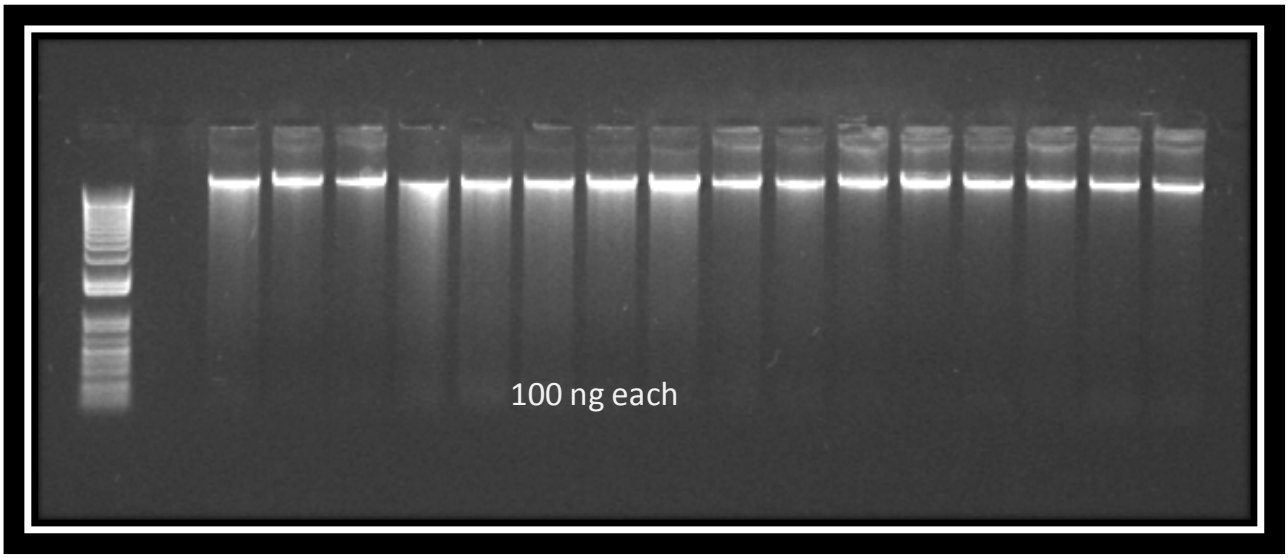


Figure 1. Nuclear DNA from the 16 selected lines run on a 1% agarose gel showing that the DNA is high molecular weight, not degraded and of sufficient quality for sequencing

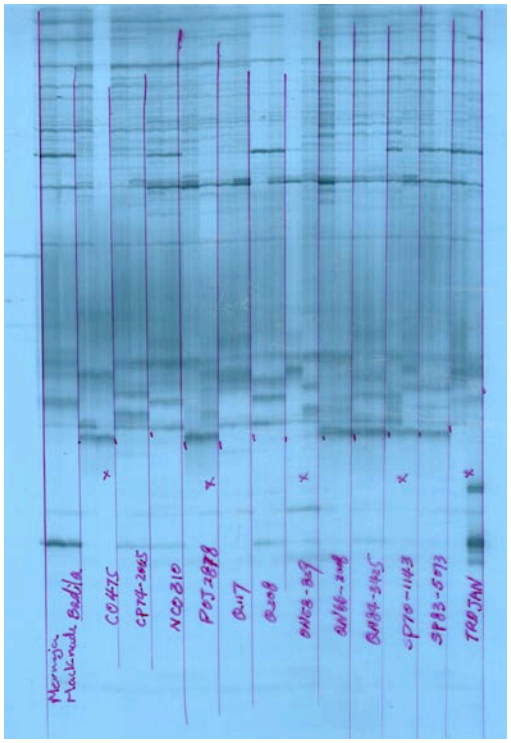


Figure 2. Microsatellite analysis of plant lines to check correct identification.

The gel image shows paired samples from Meringa (left) and Macknade (right) for each plant line. Thirteen of the lines that could be sourced from two sites were checked with SSR SSCIR33 showing that five of the lines do not match.

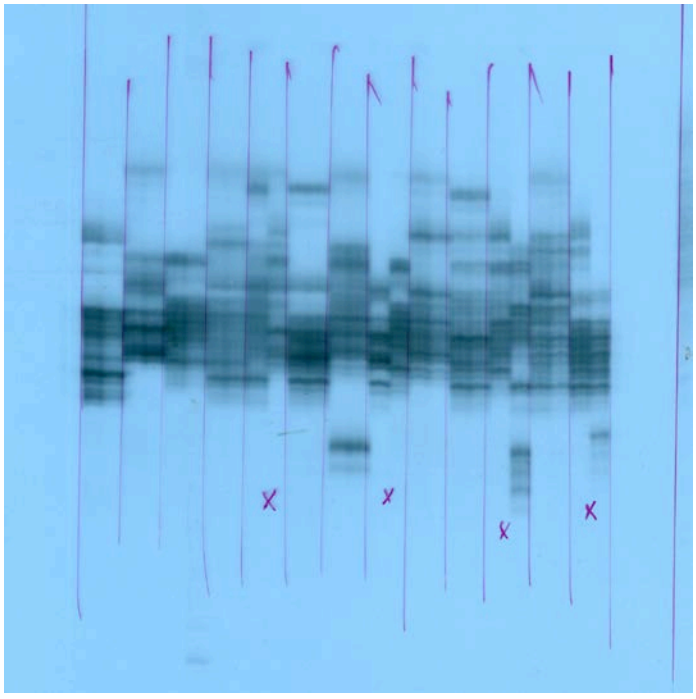


Figure 3. Microsatellite analysis of plant lines to check correct identification

The gel image shows paired samples from Meringa (left) and Macknade (right) for each plant line. The same 13 lines as in Figure 1 screened with SSR SSCIR36 confirming that the same lines do not match.

The lines that were different when checked against the second location were checked against another set of DNA and the correct line was resolved before sending for sequencing.

3.1.2 Sequence Generation from the 16 lines

A reduced representation (RRS) method using *Pst*I was used to generate the library preps for sequencing the 16 core lines. The aim of this sequencing was to target and identify SNPs within genes. Two samples were run per lane of an Illumina Hiseq 2000, with an expected coverage of at least 50x of a given genomic region (Table 2).

Table 2. Number of reads generated for each genotype

Sample	Lane	Reads
Badila	1	171,909,294
Co475	1	178,010,130
CP74-2005	2	154,340,644
Nco310	2	217,972,160
POJ2878	3	119,639,502
Q117	3	159,933,544
Q208	4	141,936,024
QN58-829	4	218,476,606
QN66-2008	5	191,146,084
QN80-3425	5	176,658,030
Trojan	6	71,507,454
Q155	6	342,257,032
SP70-1143	7	104,978,170
RB72454	7	262,617,118
SP80-3280	8	68,080,482
SP83-5073	8	159,297,874

The paired end reads were then assembled using a custom pipeline developed by Syngenta (Fig. 4)

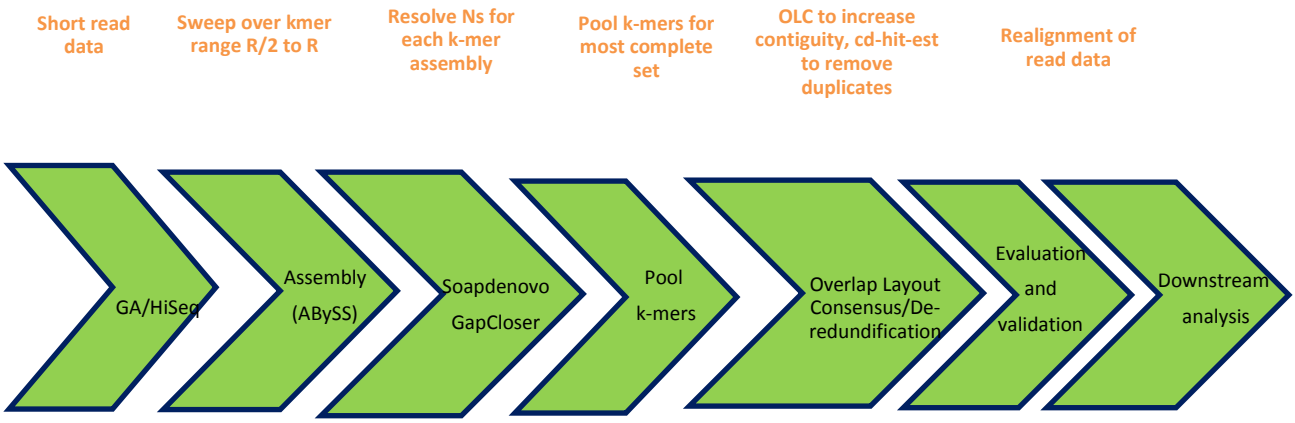


Figure 4. The sequence assembly process that was applied to the reads generated for each of the clones

Once the sequences had been assembled they were aligned to a reference RRS contig de novo assembly from a previous collaborative project (CSIRO, Syngenta) that involved sequencing two Australian lines (Q165 and IJ76-514). Alignment rates were around 50% which was expected based on the similar rates of incorporation of the read data used to produce the assembly (Table 3).

Table 3. Number of reads aligned to the contig de novo assembly

Sample	Total	Total Alignments	Reads Aligned	Percent Reads Aligned	Uniquely aligned	Percent Reads Uniquely Aligned	Unmapped
Badila	176754924	121609941	100718476	57	83565321	47.3	75769582
CP74-2005	162622094	106401332	88173225	54.2	73157434	45	74225635
Co475	181411010	118890694	98727841	54.4	82165076	45.3	82439673
Nco310	225134518	142160579	118927747	52.8	99774836	44.3	105919757
POJ2878	284140504	176731793	146898522	51.7	122345002	43.1	136879730
Q117	245199762	157410196	131154842	53.5	109518446	44.7	113728176
Q155	342257032	234706673	195762880	57.2	163589662	47.8	146042590
Q208	148208146	95520719	79284087	53.5	65953952	44.5	68724599
QN58-829	226299514	148297919	122785285	54.3	101802692	45	103210667
QN66-2008	197351276	124577415	103758204	52.6	86646407	43.9	93341926
QN80-3425	183092040	117294670	97575508	53.3	81373034	44.4	85278676
RB72454	262617118	165844535	137830346	52.5	114863267	43.7	124444444
SP70-1143	270049816	164455984	136690819	50.6	113969221	42.2	133003767
SP80-3280	244655304	156437980	128741395	52.6	106164923	43.4	115591983
SP83-5073	239526332	151479215	125122536	52.2	103490154	43.2	114086160
Trojan	240879830	158105983	132066875	54.8	110629442	45.9	108492741

Objective 1 was completely achieved with all clones generating the 50x coverage required for further analysis.

3.2 Development of the methodology for the analysis of SNP markers and the maximum use of the data including dosage

3.2.1 Initial SNP detection

The initial variant calling was performed using variant filtering criteria developed in the earlier collaborative project. In this earlier project the intent was to capture single dose markers but, as this was only between two individuals for the generation of a genetic map, the process was relatively simple. The initial filters required 4 uniquely aligned reads within the data for a sample to show the variant, with at least 5% of all reads in the same sample supporting the same variant and an average quality of the bases supporting the variant allele calls of at least 10.

Once this first step was completed, sites meeting these criteria were then genotyped across all samples with coverage at the site; assuming a ploidy level of 10 and estimating dosage of the variant according to the frequency of the variant allele within the reads for the sample, using a simple algorithm based on division of the frequency spectrum into a number of ranges based on the estimated ploidy level. In order to determine a genotype for a sample at a given site, it was required to have a minimum unique coverage of 50, so that incorrect dosage estimates for lower coverage samples would not impact subsequent classifications based on variant dosage. Number of SNP markers identified is shown in Table 4.

Table 4. Number of variants identified using initial parameters

Parameter	Variants identified
10_05_4uniq ploidy=8	~5.9 million
10_05_4uniq ploidy=12	~6.2 million
10_05_4uniq ploidy=10	~6.2 million
min coverage 50	~4.6 million
one or more samples with single/double dosage	~4.1 million

From this initial genotype-assigned variant report, the first round of selected SNPs delivered to Affymetrix for their assessment of designability was filtered to obtain ~400K SNPs, based on a categorisation of the genotypes for each site into dosage based classes (zero/low/medium/high). Then the variants themselves were categorized on the basis of the sample dosage assignments according to the following logic:

a) Class 1: Low dose (single/double) in at least 4 lines, 0 dose in at least 4 lines and high dose (5 or more copies) in at least 1 line (~50K SNPs submitted)
b) Class 2: Low dose (single/double) in at least 4 lines, 0 dose in at least 4 lines and rest can be medium dose (3-4 copies) but cannot be high dose (~200K SNPs submitted)
c) Class 3: Low dose (single/double) in at least 2 lines, 0 dose in at least 2 lines and rest cannot be either HD or MD (~150K SNPs submitted)

Note that the basic definition of Classes 2 and 3 was such that some SNPs could qualify for inclusion in both classes. In the numbers of SNPs given above for each of the classes, the Class 2 SNPs refer to SNPs belonging exclusively to Class 2 (meaning that they had at least 1 sample with a medium dosage call). In order to maximise the utility of the Class 3 SNPs (regardless of whether they also belonged to Class 2), these were filtered for having at least 60 possible informative crosses (i.e. the number of zero-dosage samples * the number of low-dosage samples); in practice, this means that all of the Class 3 SNPs submitted also qualified for inclusion in Class 2, but had no medium dosage calls, and the Class 3 number listed above corresponds to this qualified set.

In addition to the SNPs from these classes that were considered of interest as targets for the chip design, the Affymetrix design protocol requires that off-target variation also be represented as a separate file of variants in the submission (similar in principle to the way that variation in flanking sequences of the targeted SNPs is represented with IUPAC ambiguity codes in other genotype design protocols).

For this reason, it was decided not to compromise the scoring of the on-target SNPs by including minimally supported variants in their flanks, so variants not falling into the dosage categorisation classes, a further requirement that the variant allele had been seen in at least two samples was implemented. This minimised the chances of library-specific artifacts) as well as requiring an allele frequency across all samples genotyped at the site of >= 5% (assuming a ploidy of 10). This resulted in ~3.3 million variants included in this Class. Note that (as described below) some of these variants were included in the ultimate design set, due to the relatively high attrition rate of designable SNPs in the initially defined dosage classes.

Invariant regions in the RRS contigs were extracted to serve as controls per the Affymetrix chip design instructions. For this purpose, the initial variant report was examined. It identified both contigs with no variants reported as well as segments between reported variants of at least 50 bp in those contigs present. These regions were then compared to the unique coverage reports to determine the sub-regions of at least 32bp with a minimum coverage of 50 across each of a minimum of 16 of the lines assayed. Regions containing Ns in the reference were excluded and the result was delivered to Affymetrix for inclusion in the chip design.

In order to get a sense for the expected genome-wide distribution of our submitted SNPs, the sorghum genome was used as a reference to BLAST the sugarcane contigs from which the submitted SNPs were selected. The distributions for contigs containing SNPs from each of the submission classes (contigs containing SNPs from multiple classes included in each distribution separately) were all quite similar. They appeared to reflect the expected bias away from centromeric regions as expected from the gene-enrichment of the RRS protocol used for the sequencing (Figure 5).

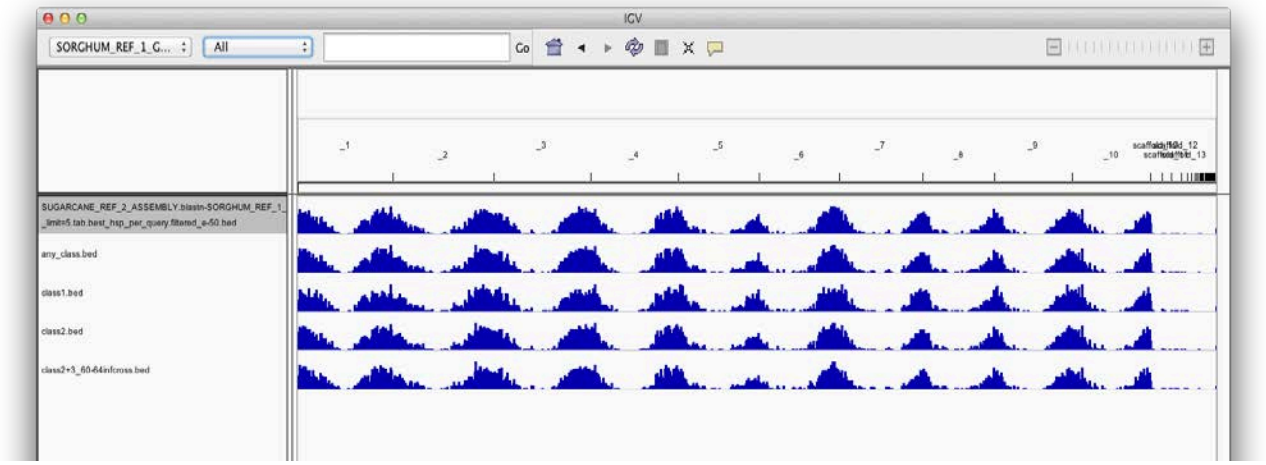


Figure 5. SNPs aligned to the sorghum genome showing low coverage in the centromeric region as expected

The top track is all the SNPs identified, the second track is SNPs identified in Class 1, the next track is Class 2 and the third track is Class 2 and 3 SNPs.

A more detailed look at chromosome 1 of sorghum reveals that the SNP markers of all classes are distributed across the chromosomes with no SNPs again present in the centromeric region as expected (Figure 6).

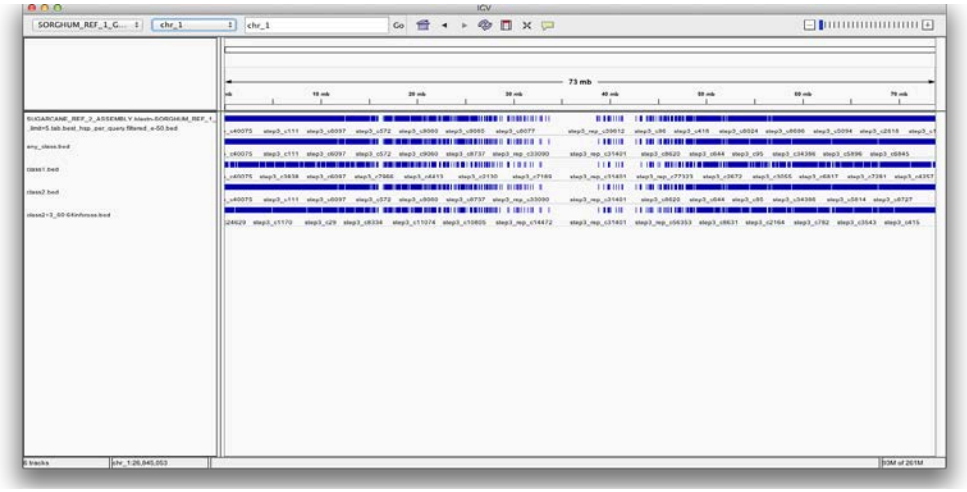


Figure 6. Sugarcane sequence aligned against sorghum chromosome 1

The top track is the alignment of the sugarcane contigs and the other tracks are the different classes of SNP markers.

Assessment of the distribution of SNPs was undertaken. This was based on their dosage on all 16 lines to ensure that there was definitively no bias based on origin of clones. In Figure 6 the number of variants that would be informative (i.e. low dosage vs zero dosage) for a cross between the samples in row and column are shown.

	Badila	CP74-2005	Co475	Nco310	POJ2878	Q117	Q155	Q208	QN58-829	QN66-2008	QN80-3425	RB72454	SP70-1143	SP80-3280	SP83-5073	Trojan
Badila	0	57809	68725	67939	68003	71083	68193	63255	72348	60021	69026	75198	70469	65377	69747	57627
CP74-2005	57809	0	67594	61640	74142	75413	67546	61848	75921	64096	64335	64348	63148	61980	62871	72466
Co475	68725	67594	0	69771	72346	66712	68172	62897	56287	68677	67577	74007	72786	74905	72603	80194
Nco310	67939	61640	69771	0	79327	81211	66684	58469	79316	70453	58316	71142	70214	73170	69013	82221
POJ2878	68003	74142	72346	79327	0	79829	78773	73258	87514	62111	74568	82184	76341	76356	81052	79681
Q117	71083	75413	66712	81211	79829	0	78377	72121	56434	73364	75246	80550	80697	77777	81723	72178
Q155	68193	67546	68172	66684	78773	78377	0	63623	72234	70643	65278	78327	75296	74903	74298	73613
Q208	63255	61848	62897	58469	73258	72121	63623	0	68217	67429	60065	65924	67192	68447	65793	70377
QN58-829	72348	75921	56287	79316	87514	56434	72234	68217	0	76548	74827	83127	82898	85210	82743	62221
QN66-2008	60021	64096	68677	70453	62111	73364	70643	67429	76548	0	68060	71336	69473	69485	69410	69401
QN80-3425	69026	64335	67577	58316	74568	75246	65278	60065	74827	68060	0	70821	70609	71642	72883	77517
RB72454	75198	64348	74007	71142	82184	80550	78327	65924	83127	71336	70821	0	72462	68853	70743	85522
SP70-1143	70469	63148	72786	70214	76341	80697	75296	67192	82898	69473	70609	72462	0	73390	67761	80951
SP80-3280	65377	61980	74905	73170	76356	77777	74903	68447	85210	69485	71642	68853	73390	0	64920	81937
SP83-5073	69747	62871	72603	69013	81052	81723	74298	65793	82743	69410	72883	70743	67761	64920	0	80751
Trojan	57627	72466	80194	82221	79681	72178	73613	70377	62221	69401	77517	85522	80951	81937	80751	0

	Badila	CP74-2005	Co475	Nco310	POJ2878	Q117	Q155	Q208	QN58-829	QN66-2008	QN80-3425	RB72454	SP70-1143	SP80-3280	SP83-5073	Trojan
Badila	0	67454	77544	74672	71510	74964	73942	75032	75303	72089	78950	79128	75015	70937	74145	54560
CP74-2005	67454	0	83012	73968	81204	84782	79186	78902	86985	76093	81144	72938	71069	69689	70999	76998
Co475	77544	83012	0	82308	76204	68846	73374	75878	57351	81979	75376	83056	83111	82835	81861	87752
Nco310	74672	73968	82308	0	85338	84370	71872	65984	87185	81113	66178	78516	75881	83669	72281	86446
POJ2878	71510	81204	76204	85338	0	74822	80072	84380	83861	64395	79772	80896	74601	75025	80921	74830
Q117	74964	84782	68846	84370	74822	0	82008	83798	55765	81721	80534	79928	82735	79247	85133	68994
Q155	73942	79186	73374	71872	80072	82008	0	78450	75121	83093	74884	80840	82781	79657	81377	74398
Q208	75032	78902	75878	65984	84380	83798	78450	0	80127	82315	71446	78798	81045	82767	75887	76876
QN58-829	75303	86985	57351	87185	83861	55765	75121	80127	0	82950	81761	82945	88366	88156	88562	58405
QN66-2008	72089	76093	81979	81113	64395	81721	83093	82315	82950	0	82383	78029	74974	78594	77224	71185
QN80-3425	78950	81144	75376	66178	79772	80534	74884	71446	81761	82383	0	81604	81115	78183	82241	81314
RB72454	79128	72938	83056	78516	80896	79928	80840	78798	82945	78029	81604	0	74659	71161	74103	81986
SP70-1143	75015	71069	83111	75881	74601	82735	82781	81045	88366	74974	81115	74659	0	75170	73210	82449
SP80-3280	70937	69689	82835	83669	75025	79247	79657	82767	88156	78594	78183	71161	75170	0	67492	80617
SP83-5073	74145	70999	81861	72281	80921	85133	81377	75887	88562	77224	82241	74103	73210	67492	0	81111
Trojan	54560	76998	87752	86446	74830	68994	74398	76876	58405	71185	81314	81986	82449	80617	81111	0

Figure 6. The number of informative variants for a cross between any two of the samples for Class 1, Class 2 and Class 2 +3 SNPs

The numbers represent the number of SNP that are different between the samples. The green is high numbers of SNPs and the orange low numbers. The distribution shows that there is no large bias towards any one clone.

3.2.2 Final selection of SNP markers

The files returned from the Affymetrix designability assessment protocol contained information for both the “on target” and “off target” SNPs, and include a number of considerations as to the likelihood of success for each. In particular, they give a “recommendation” classification based on the following criteria derived from other scores provided in the file:

1. A marker/strand is recommended if: pconvert > .6, and there are no wobbles, and poly count = 0.
2. A marker/strand is not_recommended if: duplicate count > 0, or poly count > 0, or pconvert < .4, or wobble distance < 21, or wobble count >= 3
3. A marker is not possible on a given strand if we cannot build a probe to interrogate the SNP in that direction. This is mainly due to ambiguous bases (W, Y, N, S etc.) we found in the sequence. At least one unambiguous base on the far side of the polymorphism is required in order to build a probe from the opposite direction.
4. All other markers are considered neutral.

Only ~100K of the SNPs considered “on-target” actually met all these designability criteria and were included in the “recommended” category. In order to adjust for this high level of attrition without requiring another time-consuming round of designability assessment and in consultation with Affymetrix, two new categories were made, from which the final selection algorithm proceeded.

In the first of these, all SNPs from our original dosage-based categories were placed as long as they had a pconvert score of >= 0.65; this score represents the probability of conversion of the SNP based on its flanking sequence and absent any consideration of the likelihood of the duplication of the assayed region or variability in the flanking sequence. Given the nature of the reference and the relaxed allele-frequency requirement on the off-target SNPs, ignoring the possible effects of these other dimensions of the “recommendation” categorisation seemed to be justified. This included ~220K SNPs for consideration.

To ensure that SNPs from as many of the contigs as possible were included to maximise the genomic distribution, the original dosage-classification logic was rerun with reduced requirements on the number of lines with zero/low dosage (at least 2 in each of these categories, and no longer restricting the Class 3 SNPs to have more informative crosses than implied by the basic definition). SNPs from the set previously considered off-target, that met these adjusted dosage classification criteria and were also “recommended” according to Affymetrix’s full criteria, were considered as possible for inclusion when a given contig did not have enough SNPs from the Class 1 to meet the targeted average number of SNPs per contig. Finally, a set of ~2300 SNPs that had been validated in previous work with Infinium genotyping chips were assessed for pconvert scores by Affymetrix, and 2073 of these were included with a pconvert threshold of 0.55.

The algorithm used for final SNP selection tried to maximise the evenness of the genomic distribution by:

1. Starting with the contigs with the fewest number of candidate SNPs
2. Keeping a running average of the number of SNPs that would be preferred per contig
3. Selecting up to that many SNPs if possible (selecting as many as possible otherwise)
4. Prioritizing SNPs selected based on their meeting the original stringent dosage criteria first, and secondarily by pconvert score (with the 0.65 minimum imposed)

Following the submission of the 200K set chosen using this strategy and the receipt of a proposed design for the chip in October 2014, Affymetrix advised that their array format would accommodate ~420K markers, giving the option of tiling the submitted markers on both strands or submitting additional SNPs for tiling new markers. After some discussion of the benefits of double-strand tiling (only useful for high-importance markers to increase the chance of successful assay conversion) and consideration of project timelines, it was decided to tile the Infinium-derived SNPs and the Class1-stringent and Class2-stringent SNPs on both strands when possible. It was decided to choose additional markers for the remainder, using the same strategy as previously, but increasing the target number of SNPs.

After a submission of 420K (Table 5) markers (representing a somewhat smaller number of SNPs, due to double-stranded tiling as well as certain classes of SNPs that require each allele to have a separately designed probeset), a small proportion of submitted markers were rejected by the design protocol. Among these were a number of the Infinium SNPs, but it was determined that in this

special case it was due to their having been duplicated among the SNPs derived from the new sequencing data. Having re-added the Infinium SNPs (except for one that had no flanking sequence on one side, leaving 2072 total), and accepting some additional double-tiled markers nominated by Affymetrix to fill in the remaining deficiency in total capacity ("21,508 probesets for the opposite strand of the rank 1 markers that have the highest pconvert values"), the final numbers of SNPs were selected from the revised categorisation as:

Table 5. Number of SNPs selected for each SNP class (where "stringent" denotes the original dosage classification and "non-stringent" the relaxed version)

The total number of tiled elements on the chip is 424,048 markers, corresponding to 345,704 unique SNPs.

Class	Number of SNPs	Both strands tiled	One strand tiled
Infinium	2072	665	1407
Class1-stringent	26355	11467	14888
Class2-stringent	108563	48353	60210
Class2+3-stringent	76461	6377	70084
Class1-nonstringent	22156	1866	20290
Class2-nonstringent	34932	3009	31923
Class2+3-nonstringent	75165	6607	68558

According to the contig-sequential selection algorithm described above, the number of contigs allowing selection of a given number of SNPs is as follows (Table 6).

Table 6. Number of SNP markers per contig selected

Contig_count	SNP_count
9581	1
7089	2
5462	3
4266	4
3406	5
20941	>=6

When the original samples were analysed for the selected SNPs, the number of variants was similar across all lines apart from Badila. Badila contained the lowest number of SNP markers due to it having only 80 chromosomes, compared to around 110-120 for the other lines and to it being a pure *S. officinarum* rather than hybrids as all the other lines are (Table 7).

Table 7. Number of SNPs per plant line selected for the array

Sample	No. of SNPs
Badila	74843
CP74-2005	132444
Co475	146555
Nco310	160732
POJ2878	175135

Q117	179788
Q155	161121
Q208	146194
QN58-829	171205
QN66-2008	146649
QN80-3425	155654
RB72454	185154
SP70-1143	169815
SP80-3280	164689
SP83-5073	165287
Trojan	168824

This objective was achieved with the methods developed for selection of low dosage SNP markers across all the germplasm.

3.3 Analysis of an association mapping and biparental mapping population to identify markers linked to traits of economic importance

3.3.1 Screening the association mapping population across the 400K SNP array (Canechip)

The Affymetrix Axiom Technology has been selected for the sugarcane SNP array design as the most appropriate method for screening SNP markers in sugarcane. It has had successful outcomes in a pilot project on wheat, another polyploid crop, with great improvement for SNP calling over other technology tested. They have also developed an Axiom array for strawberry which is an octoploid. In addition to this the Axiom technology used a two-step approach to the development of the SNP array with the first array developed containing 345K SNP markers. This array was screened with the SRA/CSIRO association mapping population containing 476 individuals. The results from this array meant that low dose markers can be selected for the smaller second array of a 40K SNP, which will be used to screen a number of populations in a more cost-effective way. Axiom Biofx Services has been contracted as the service provider for producing the SNP chip. This will maximise the number of low dose usable SNP markers on the second smaller array. The design of the Axiom chip is an iterative process between Axiom Biofx Services, Syngenta, CSIRO and SRA (as in Figure 7).

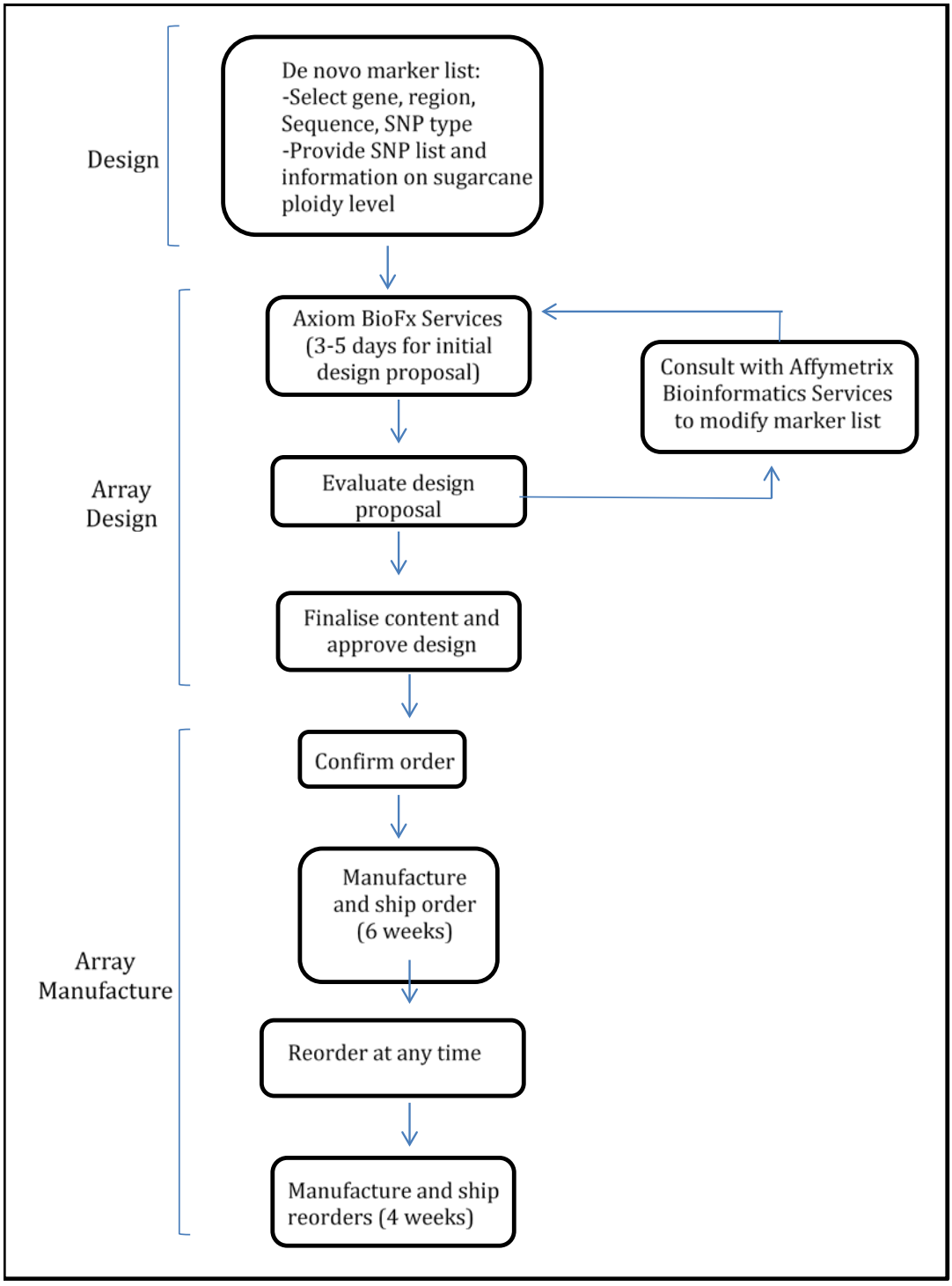


Figure 7. Axiom custom array design process

The list of samples in the association mapping population is given in Supplementary Data 1. This population consists of parental clones from the Australian Sugarcane Breeding Program plus commercial cultivars and the other half are clones randomly taken from 30 unselected families (eight clones per family). This allows the structure in the analysis of the data to be implemented. In addition 20 lines from the mapping population (Q165 x IJ76-514) were included to help determine dosage of the SNP markers from the segregation frequency in this set of DNA. The DNA was purified and run

out on agarose gels to check the level of degradation of the DNA (Figure 8). The DNA was then placed into the 96 well plates and sent to Affymetrix for screening across the arrays.

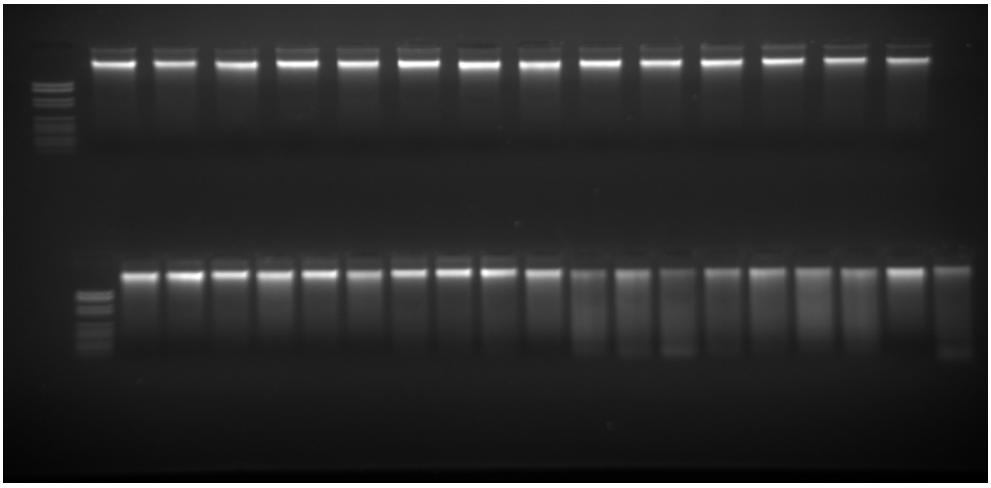


Figure 8. A subset of purified DNA from the association mapping population, showing that the DNA was high molecular weight and under-graded

3.3.2 Analysis of the association mapping population on the 400K sugarcane axiom array

The Axiom GTv1 Genotype Calling Algorithm was used to identify genotypes. This uses a Bayesian procedure in which a prior (calculated from expected genotypic frequencies) for each SNP was combined with the observed data to obtain a posterior estimate of cluster centers. The posterior estimate was then used to call genotypes. The prior used may be a generic prior common to all SNPs or a specific prior computed for that SNP from a set of training data. In the case of sugarcane no training data was available so this was initial analysis.

The array was designed to include some SNPs that had two probe sets to interrogate the SNP with both forward and reverse strand probes. This increases the chances of obtaining a result but reduces the number of SNPs that can be placed on the array. The array was designed to include more single probe SNPs to increase the number of SNPs on the array. The SNP calling is automated and examples of the other types of polyploid species that have used this system were given in the report generated by Affymetrix for the array (Supplementary Data 2). As seen from the report, the higher the ploidy of the plant, the more complex the SNP clustering becomes as the SNP clusters become less distinct and blend together making it more difficult to assign genotypes to individuals.

The association mapping population was screened across the array and included part of a mapping population and some more diverse germplasm. The greater the genetic diversity the more likely it is to observe off-target caused by too much variation in the screened population. This is because the SNP sequence varies from individual to individual and may, in more diverse germplasm, have double deletions or sequence non-homology. As association mapping population contains germplasm from the breeding program, the DArT data has been used to assign groupings to the population to reduce this problem.

The data from the array was automatically clustered and the genotype called. For a diploid population only three clusters are expected: AA, AB, and BB. In a polyploid for example, an autotetraploid with 4 copies of every chromosomes, five clusters would be expected. In sugarcane which has at least 10 copies of every chromosome, many more clusters would be expected. This results in a much more complex output from the array and makes clustering the data and assigning the genotype far more difficult. This is partly due to the fact that as the SNP increases in dosage in the population, the clusters become less distinct (see Supplementary Data 2).

SNP Polisher was used to classify the output from the array. The SNPs from the array were classified into genotypic classes (Table 8).

Table 8. SNP classification for all the SNPs on the Canechip

SNP Category	Count	Percentage
PolyHighResolution	11440	3.3%
No Minor Hom	20248	5.9%
MonoHighResolution	149160	43.2%
OffTargetVariant	225	0.1%
ABvarianceY	129	0.0%
AAvarianceY	162	0.1%
BBVarianceY	105	0.0%
ABVarianceX	146	0.0%
AAvarianceX	57	0.0%
BBVarianceX	97	0.0%
UnexpectedHeterozygosity	456	0.1%
HomHomResolution	162	0.1%
CallRateBelowThreshold	27905	8.1%
Other	135412	39.2%
Total	345704	100%

The cluster output also generates graphs. Figure 9 shows the results for the most important categories from the array. Interestingly 3.3% of the SNP markers were classified into the polymorphic high resolution group. This would seem to indicate that sugarcane does have a small amount of preferential pairing which has been indicated in the genetic studies carried out in CSIRO (Aitken et al 2005). The most interesting group is the 'No minor homozygous' cluster which are the genotypes that are mapped in sugarcane using dominant marker methods and will be used for the QTL analysis.

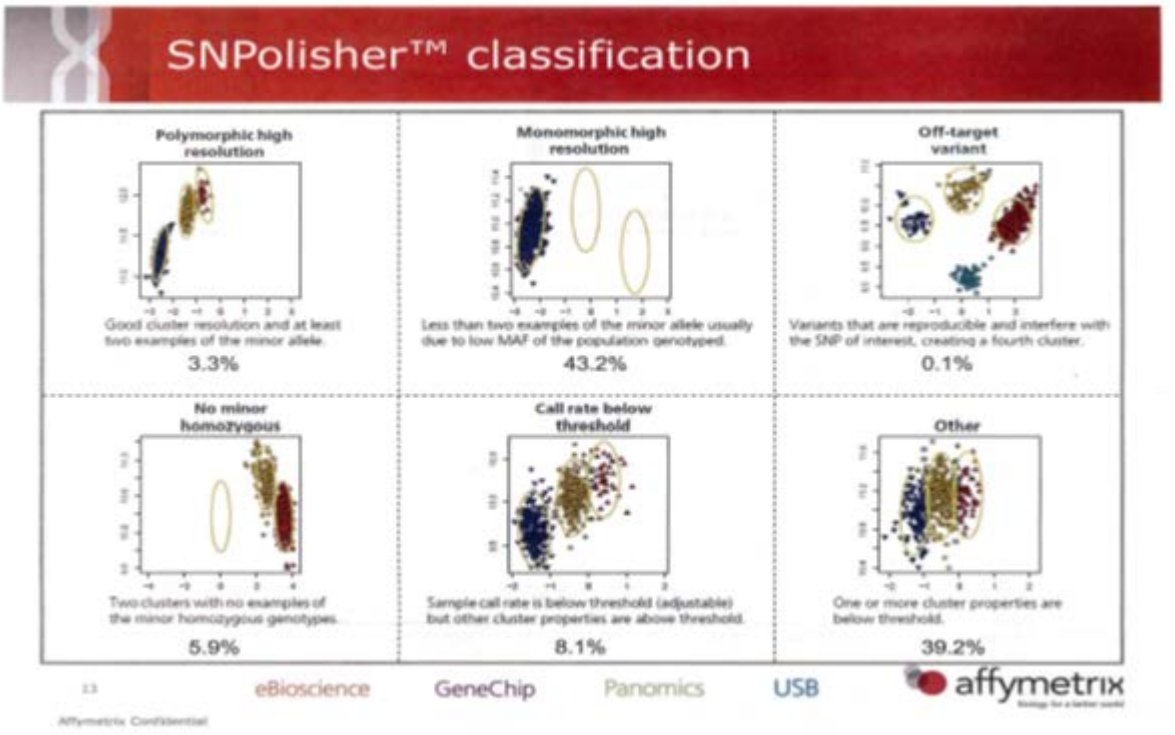


Figure 9. Cluster diagrams for the most important SNP markers on the array

There are 31688 SNP markers that can be used for QTL analysis within the association mapping population. These include the PolyHighResolution SNPs and the No Minor Hom SNPs (Table 8). Further work will be carried out using cluster analysis of the DArT data to determine if with this additional information more of the SNP markers can be transferred, either from the 'Call Rate Below Threshold' category or the 'Other' section (Table 8) to usable and repeatable SNP markers. Sugarcane is the most complex polyploid that has been screened for a SNP array of any kind and the data presented here gives confidence that this technology can be used in sugarcane breeding.

Further interrogation of the data has increased the initial number of polymorphic makers from 31,688 to 47,803 SNPs. This was done by selecting markers from other categories that did not initially meet the high cut-off rates, but on further analysis demonstrated good clustering of the data; for example, the 'Call Rate Below Threshold' group of markers (see milestone 5). The data was also run through fitTetra, an analysis package designed for tetraploid plants. This package assigns the clones to one of five clusters that correspond to the five possible genotypes generated in a tetraploid. This allows the data to be clustered into the 5 genotype clusters rather than the maximum of 3 clusters used for the initial analysis. It is highly probable that parts of the sugarcane genome will act as a tetraploid due to its hybrid genome structure and this was verified as 1386 SNP markers fitted the 5 clusters of a tetraploid (Figure 10).

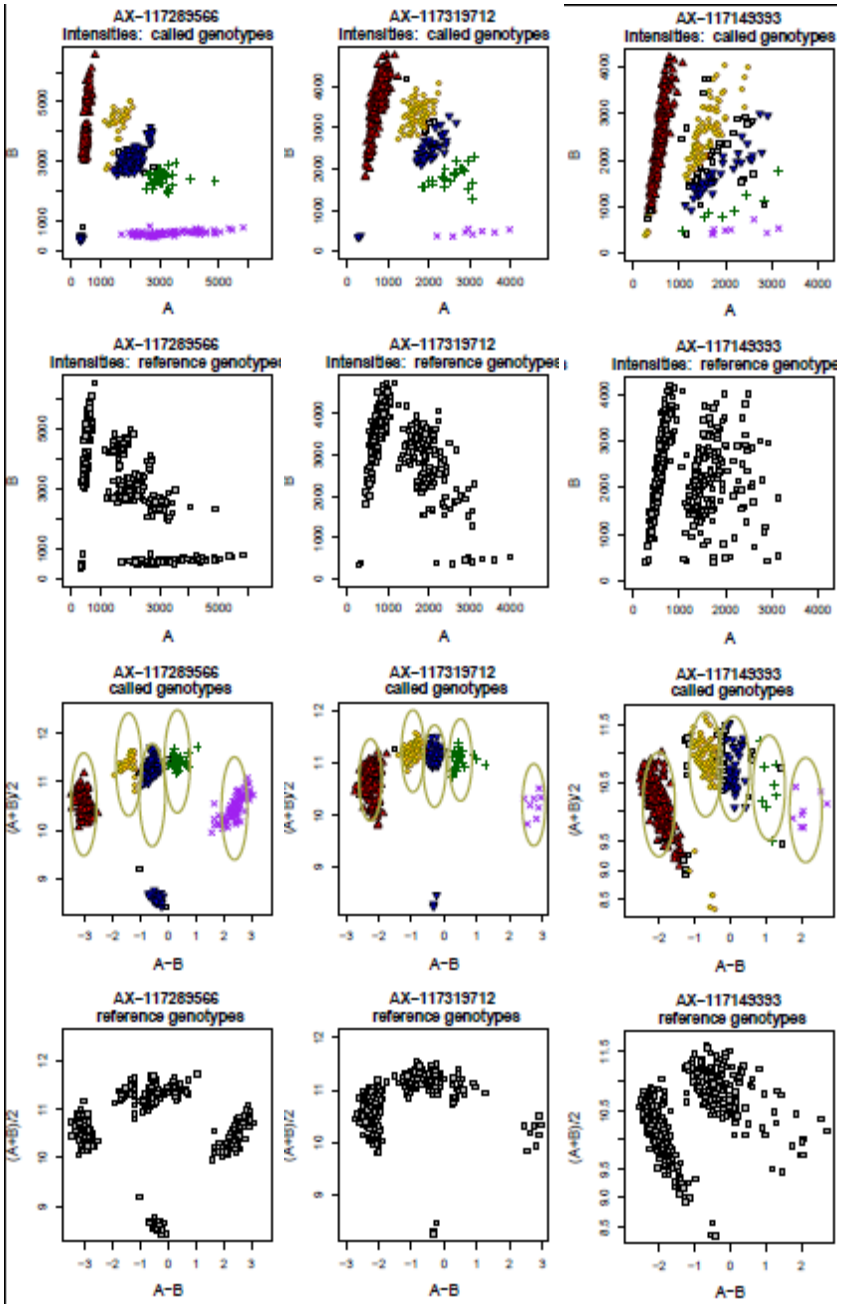


Figure 10. Examples from fitTetra showing the clustering patterns for clones for a selection of SNP markers

The clones have been assigned to one of the five genotype clusters that are predicted for a tetraploid plant (genotypes are in different colours). The graphs show the progressive analysis to classify the plants lines into genotypic clusters.

As the SNP markers have an associated sequence, we can align the SNPs to sorghum, the most closely related plant that has a genome sequence, to determine the genome coverage of the polymorphic markers. The polymorphic SNP markers were located on all of the sorghum chromosomes (Table 9). Interestingly only 18,856 SNP markers aligned to the sorghum genome, the rest did not at $\geq e^{-51}$ and therefore may not be within a gene. This suggests that at least 18,856 genes are tagged with a SNP using this method and probably more align but at a lower significance level.

Table 9. Number of SNP markers aligned to the sorghum genome at greater than e^{-51}

Sugarcane homology group	Sorghum chromosome	Number of SNP markers
HG1	Sb4	2092
HG2	Sb6 and Sb5	2842
HG3	Sb3	2696
HG4	Sb1	3216
HG5	Sb7	1280
HG6	Sb9	1543
HG7	Sb10	1640
HG8	Sb8 and Sb2	3401
Scaffolds		146
Total		18856

The final report from Affymetrix on the analysis of the SNP data and selection of the polymorphic high quality markers is included in Supplementary Data 3.

3.3.3 Analysis of the association mapping population

Two methods were used to identify QTL for important traits. The first was a mixed model which was applied in analysis of the data to simultaneously account for the impact of population structure, genotype by environment interaction and spatial variation within a trial. This analysis has been carried out for CCS and TCH and indicates that substantially more markers were identified linked to both traits than were identified with the DArT markers, even given the much larger number of SNP markers (Table 10).

Table 10. Number of SNP and DArT markers identified that were associated with cane yield and sugar content at different p values

Significant level	Number of SNPs expected by Random chance	Number of DArTs expected by Random chance	TCH		CCS	
			DArT	SNP	DArT	SNP
0.05	2295.75	768	1228	15177	1380	10033
0.01	459.15	154	352	5373	377	2775
0.001	45.9	15	64	1212	55	495
0.0001	4.59	2	8	284	8	93

Similar results were obtained for the disease resistance data with a large number of markers being identified at a high significance level (Table 11).

Table 11. The number of SNP markers associated with disease resistance at difference levels of significance

Significant level	Smut	Pachymetra	Leaf scald	Fiji leaf gall
0.05	3588	3842	2728	3278
0.01	1050	1031	799	844
0.001	350	168	148	137
0.0001	216	30	39	26

Significantly more markers were identified using the SNP chip, indicating that SNP markers are more robust and not subjected to variation in hybridization intensity to the same degree as DArT markers. The results indicate there are a large number of markers linked to TCH and CCS which could be used to improve the genetic gain within the SRA Sugarcane Breeding Program.

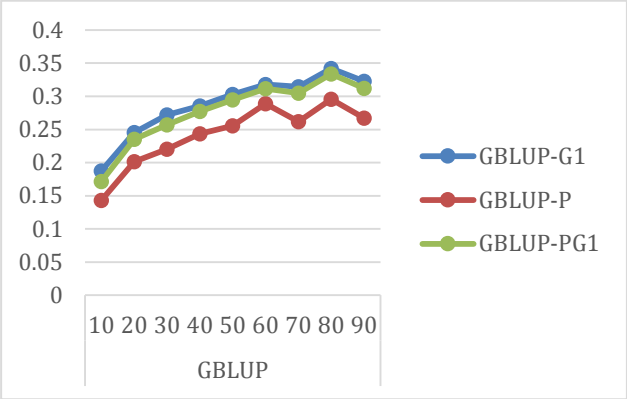
The second method used for analysis of the association mapping data was again using mixed models but in this case all the markers are fitted simultaneously and the data used to predict TCH and CCS. Firstly, the SNPs that have extreme frequencies were removed from the data and for each SNP the missing values were replaced by the most frequent allele. This was because the methods used cannot tolerate missing data. The mean for each SNP is computed and was used to center the data values. Standard deviation for each SNP was then computed on the centered data. The centered values were then standardised to a unit variance by its standard deviation.
The clones without phenotypic data were dropped from the analysis. CCS and TCH were both centered by subtracting the corresponding phenotypic mean.

Two types of mixed model were fitted, either GBLUP model or RKHS model. For these models μ is the only fixed effect which is the overall mean of the phenotypic data. All other factors including the error are random effects. For each type of model, three models fit that differed in the information used. Two of these used only either the pedigree data or SNP data. The third model used both pedigree and SNP data.

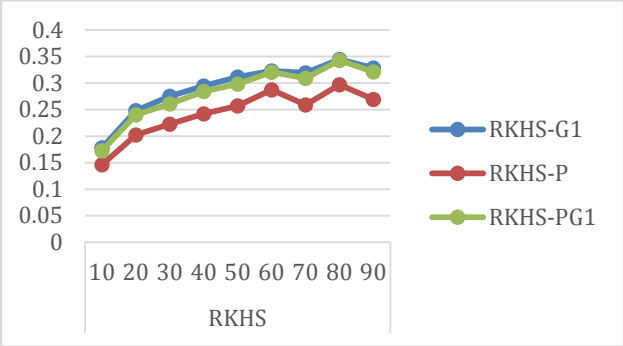
Variance parameters for these models were treated as unknown. A scaled inverse chi-square prior distribution with 5 degrees of freedom and scale parameter of 1 was assigned to these variances. For each model, samples from posterior distributions were obtained with a Gibbs sampler. Inferences were based on 12000 samples with 200 samples burn-in. The BGLR R- package (de los Campos and Perez-Rodriguez, 2013) was used to fit the models.

The models described above were evaluated for their ability to predict TCH and CCS. To do this, 50 sets of training:test (TRN:TST) datasets were used. For each set, ninety percent of the 456 clones were randomly selected and assigned to the TRN dataset. The remaining ten percent was tagged as the TST dataset which gave the 90 TRN:10 TST ratio. Similarly, clones were randomly selected and assigned to the TRN and TST dataset respectively for 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90 TRN:TST ratios to test the impact of training size on the accuracy. Phenotypic data for clones in the TST dataset were set to missing (Figure 11).

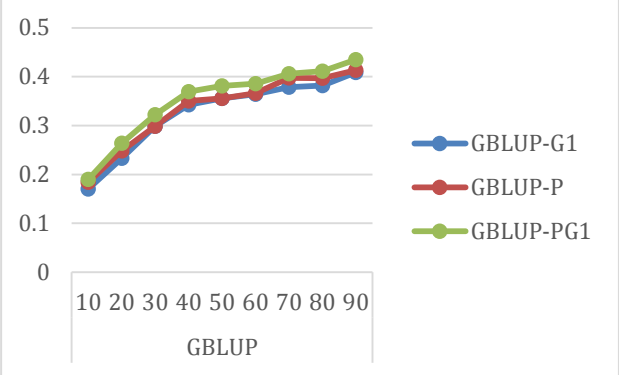
a)


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10 20 30 40 50 60 70 80 90
GBLUP
GBLUP-G1
GBLUP-P
GBLUP-PG1

b)


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10 20 30 40 50 60 70 80 90
RKHS
RKHS-G1
RKHS-P
RKHS-PG1

c)


0.5
0.4
0.3
0.2
0.1
0
10 20 30 40 50 60 70 80 90
GBLUP
GBLUP-G1
GBLUP-P
GBLUP-PG1

d)

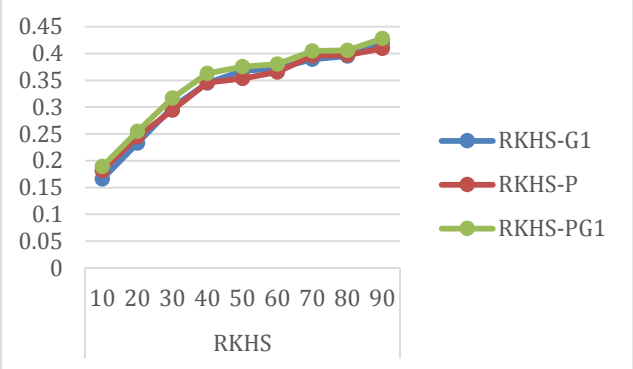
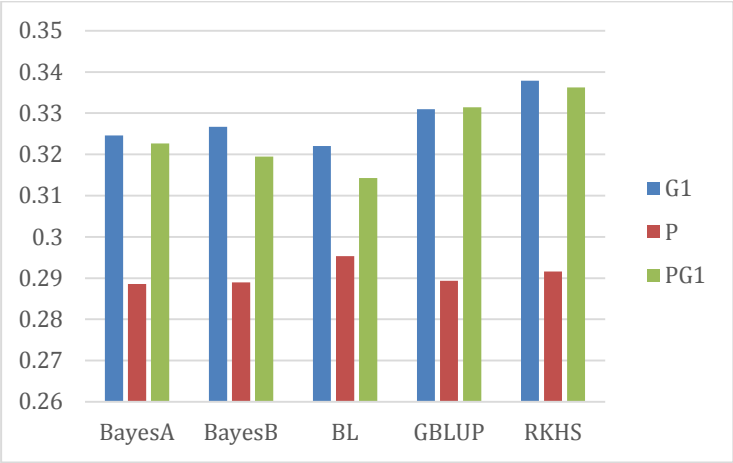

0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10 20 30 40 50 60 70 80 90
RKHS
RKHS-G1
RKHS-P
RKHS-PG1

Figure 11. Graphs of prediction accuracy for each of the (TRN:TST) for 47,803 SNP markers for a) CCS using the GBLUP model, b) CCS using the RKHS model, c) TCH using the GBLUP model and d) TCH for the RKHS model. The blue line is using the SNP data only, the red line is the pedigree data only, and the green line is using both pedigree and SNP data

The graphs all show that as the training populations increase in size the prediction accuracies increase, indicating that a population size of 371 clones is not enough to achieve an accurate prediction value. All graphs show that using either the SNP data only, or combining the SNP data with the pedigree data increases the prediction value over using pedigree data alone. The size of training populations in other species using these analysis methods are much greater. Work by Syngenta on soybean used 1000 lines and achieved a prediction accuracy of 0.38 with a genetic gain of 3%. In corn the predictive values were slightly higher at 0.45 but with a similar number of lines. This indicates that sugarcane should achieve equivalent or high prediction values when screened with 1000 lines. Higher prediction values were achieved using the GBLUP models than with the RKHS. The graphs also indicate that the SNPs give higher prediction values for TCH than for CCS, possibly due to less variation for CCS in the training population than for TCH.

This same method was used for smut and pachymetra in the association mapping population.

a)



b)

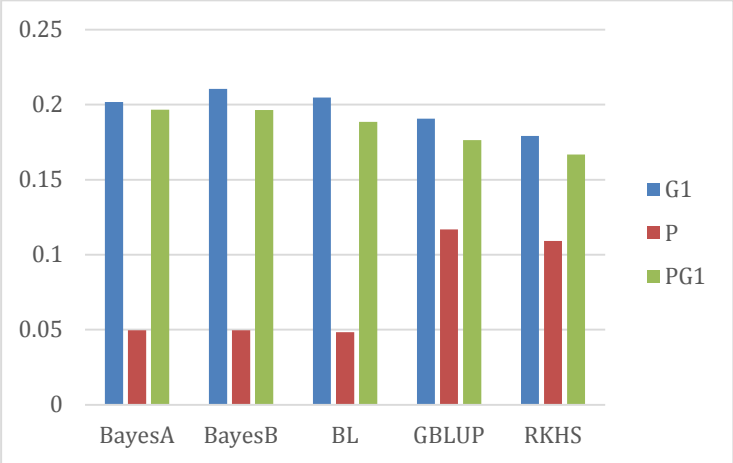


Figure 12. Graphs of the prediction accuracies of 47,803 SNP markers for a) smut using the BayesA, BayesB, Bayesian LASSO, GBLUP and RKHS models, b) Pachymetra using the BayesA, BayesB, Bayesian LASSO, GBLUP and RKHS models

G1 is using only SNP markers, P is pedigree data and PG1 is both SNP and pedigree data.

The results indicate low predictive accuracy for both smut and pachymetra probably due to the small size of the data set.

3.3.4 Selection of population for screening

The polymorphic markers identified in the 400K sugarcane were re-arrayed to generate the final Canearray. The two populations selected for screening across the smaller polymorphic Canechip were the FAT trials and the SmutBuster population. The reasons for their selection are listed in Table 12.

Table 12. Populations selected for screening across the 40K Canechip

Option	Number of clones	Possible advantages	Possible disadvantages
FAT trials	2011 planted series: x clones 2012 planted series: y 2013 planted series: z Total 764	High quality data, especially cane yield Ratooning data Smut data & some other disease data (better than other options except parents)	Selected material, so will not have extremely low performing clones, especially for CCS Disconnected sets of clones Not as good as parents for disease data
SmutBuster series	1086	Relatively unselected genotypes and high variation (especially for CCS) – similar to what end application may be OK phenotypic data (except single row cane yield) – across 4 sites	Cane yield only single row data Some concern about "representativeness" since does not include smut resistant parents

Tonnes of Cane per Hectare (TCH) and Commercial Cane Sugar (CCS) were collected from two trials setup in 2013. These trials were the SmutBuster CAT trial in 4 regions at North, Burdekin, Central and South with 7, 220 clones and the FAT trial where the clones were repeated in 4 trials per region but not across regions with 764 clones. Disease rating data on the FAT clones was also collected from several pathology trials. A total of 58, 363 SNP marker data was collected on these clones. After data cleaning and merging, there were 1, 850 SmutBuster CAT and FAT clones with TCH and CCS data, 790 clones with SMUT ratings, 351 clones with Pachymetra ratings and 23, 667 SNP markers left for further analysis.

Both analysis methods were used (as before) for the association mapping population. The first was a mixed model which was applied in analysis of the data to simultaneously account for the impact of population structure, genotype by environment interaction and spatial variation within a trial. This analysis was only carried out for TCH and CCS and the results indicate that many more markers are significantly associated with this trait than would be expected by random chance (Table 13).

Table 13. Number of markers identified that were associated with TCH and CCS at different significance levels

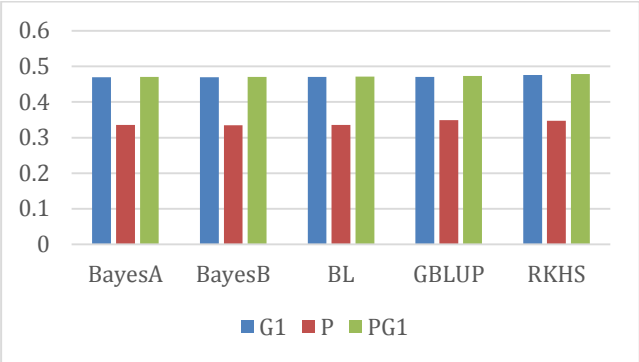
Significant level	Number of SNPs expected by Random chance	TCH	CCS
0.05	1183.3	3164	3276
0.01	236.7	1270	1573
0.001	23.7	253	380
0.0001	2.37	52	110

The second method used was to assess genomic selection on these clones. The objective was to determine the gain in prediction accuracy from SNP marker information. Three sets of models were used according to the information in the model. The first set was with all available SNP marker data included in the model. One of the remaining two sets used only pedigree information, the other both pedigree information and SNP marker data. For each of these sets, several Bayesian models were considered, namely, BayesA, BayesB, Bayesian LASSO, GBLUP, and RKHS.

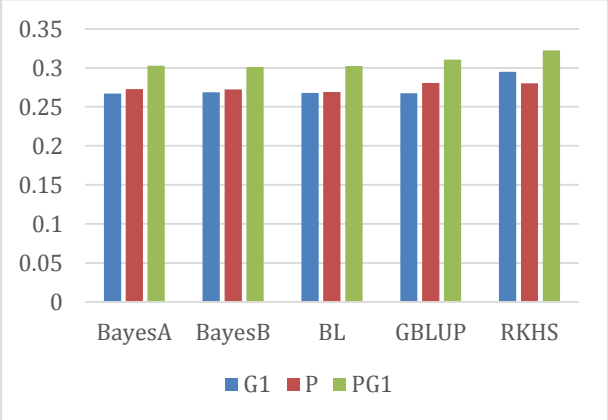
A cross validation on each model per set was carried out. The data was partitioned into two sets where 90% of the clones was used as the training dataset and the remaining 10% as the test dataset. Clones were randomly assigned to the training and test datasets. Fifty (50) data partitions were prepared for each trait. A model was fitted on the training dataset and evaluated for prediction accuracy on the test dataset.

The following graphs show the predictive accuracy for CCS, TCH, smut and Pachymetra (Figure 13).

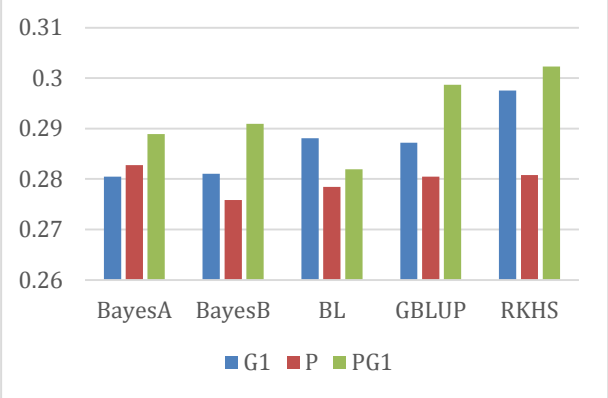
a)



b)



c)



d)

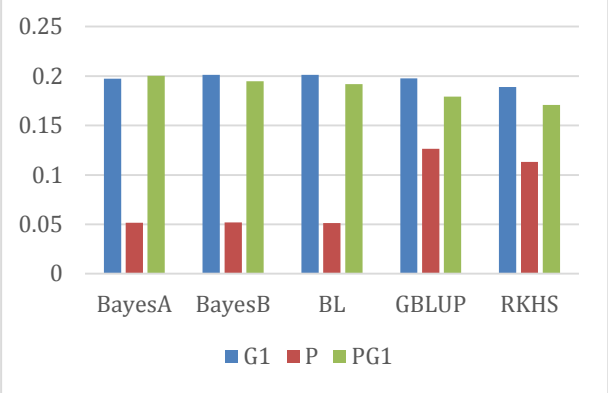


Figure 13. Graphs of the prediction accuracies of 23,667 SNP markers for BayesA, BayesB, Bayesian LASSO, GBLUP and RKHS models, for a) CCH b) TCH, c) SMUT for the FAT trials only and d) Pachymetra for the FAT trials only

G1 is using only SNP markers, P is pedigree data and PG1 is both SNP and pedigree data

The data indicated that predictive accuracies for CCS were around 0.47, a good predictive value. For TCH, the value was lower at 0.3 but this was expected as TCH is controlled by many more genes than CCS. The predictive values of both smut and pachymetra were very low but this is likely due to a small population size and that most of the FAT clones are smut and pachymetra resistant.

This objective was achieved with all the analysis completed. The results are very positive for CCS and TCH and there is strong evidence that the SNPs identified in this project can be used to select for plant lines with increased breeding value in the Australian sugarcane breeding program.

3.4 Determination of the cheapest and most effective strategy to validate the selected SNPs into a breeding program.

The cost of marker implementation depends on a number of factors including number of individuals to screen and number of SNP markers that those individuals will be screened with. So along with determining costs of screening SNP markers there has to be a determination of where in the breeding program those SNPs will be implemented.

The analysis of the data so far indicates that:

1. A more detailed and objective simulation of predicted gain and cost-effectiveness of a rapid recurrent marker assisted breeding program based on a subset of the SNP markers will be carried out. This would provide a clear cost/benefit analysis of implementation and the case for industry investment in a marker assisted breeding program.
2. A subset of the most important markers would be chosen based on examining models of genomic prediction analysis. Methodologies to screen for these sets of markers at low cost per genotype would be identified, tested and costs accurately determined. Some of this information would be available from other SRA funded projects.
3. At the present time the most likely scenario for implementation based on analysis of the data so far is screening the clones at the CAT stage, and selection based on an index using markers and phenotype data. This would give an improved selection of parents which would generate improved varieties faster (Figure 14).

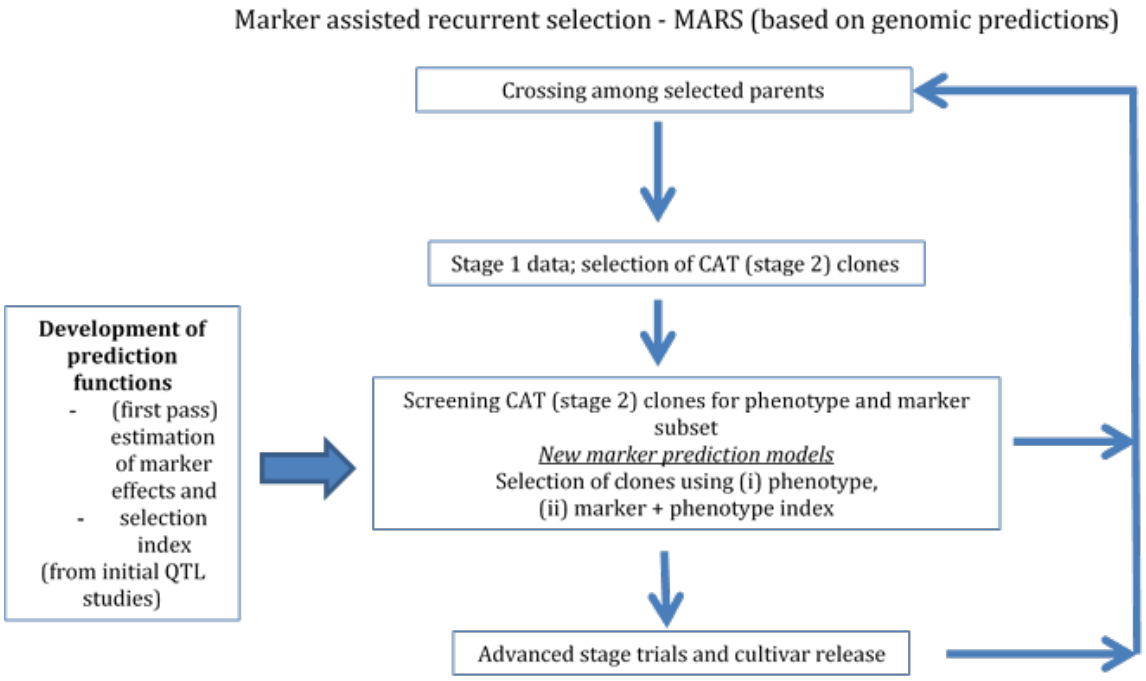


Figure 14. A diagram of one proposed method of using SNP markers in the breeding program

Initial work on testing methods has been carried out in project SRA 2015/025. This tested a number of SNP platforms to determine which works most accurately for sugarcane and is cost effective for implementation of smut and pachymetra markers. The technology for screening SNP markers is in constant development, so the tools selected at this point in time may not be the tools used for implementation in a few years' time.

Implementation depends fundamentally on demonstrating the technical case outlined above in point 1, and obtaining the funds to carry out the program. With the appointment of Bert Collard to SRA, a program is being developed that will use the data and tools that have been developed in this project.

4. Outputs and Outcomes

4.1 Outputs

1. A sugarcane SNP chip with 40K SNP markers that are segregating in the Australia breeding program.
2. 50x coverage of the gene rich regions of Illumina paired end reads for 16 important ancestors of the SRA Sugarcane Breeding Program.
3. A set of parameters to define SNP markers for using in a breeding program to maximise the identification of low dose polymorphic SNPs.
4. Methodologies developed for the genomic prediction of traits using SNP data.
5. SNP markers linked to CCS, TCH and disease resistance to smut and pachymetra.
6. Two workshops were organised for the discussion of marker implementation in sugarcane. They were used to inform the breeders of the project progress and present the results of the work. Workshops were held in March 2014 and 2016 and were well attended by all sugarcane breeders and molecular geneticists.

4.2 Outcomes

1. Increase in capability; there is now a selection tool available for genomic selection in the sugarcane breeding program.
2. Ability to select parents for the breeding program with better breeding values and increased rate of variety development.
3. Sugarcane breeders with a better understanding of SNP markers and how they can be utilised.
4. SNP markers for validation and use in any marker system within the breeding program.
5. An implementation plan for marker assisted recurrent selection for parental improvement.
6. Better communication between sugarcane breeders and molecular marker scientists.

5. Intellectual Property (IP) and Confidentiality

There is no direct IP generated from this project which should be legally protected or treated confidentially.

6. Industry Communication

The key message is there is now a 40K Canechip available that can used for genomic prediction in any sugarcane breeding program. This array has been used for the first genomic prediction analysis in sugarcane. Results indicate that this tool can be used to increase genetic gain in selection of parents in the breeding program and will increase the speed that new better varieties are being developed.

The Canechip can now be used and data from the first screening of this array is being used to select clones from the SmutBuster population to generate new high yielding resistant varieties faster. The Canechip is also being used in an Australian Indian project lead by Prakash Lakshmanan from SRA.

The information created by this project will be further published in international journals but has not been widely communicated to the public, however it work been communicated to several scientific audiences in the ISSCT workshop and the Tropical Crops Congress.

7. Environmental Impact

There has been no environmental impact in conducting this project.

8. Recommendations and Future Industry Needs

This project forms the basis for the future implementation of marker assisted selection in sugarcane. The recommendations for implementation are:

1. The Canechip is a valuable tool that can be used for discovery in any sugarcane breeding program around the world. The array will be used to assay the diversity in the Indian breeding program in Coimbatore. This data can then be used to determine the level of variation between Indian and Australian germplasm. This information will provide the breeders a method to identify which new germplasm they would like to integrate into their breeding program to incorporate new traits.

2. The Canechip has been the first tool to assay genomic selection in sugarcane and determine the predictive power for traits such as CCS, TCH and disease resistance to smut and pachymetra. Further work is needed to develop the statistical methods to increase the power of this analysis.

3. The array allows the analysis of over 40K SNP markers but selection for some traits will rely on much smaller numbers of SNP markers. Work is needed to convert the array SNP markers to high throughput cost effective detection methods. Some of this will be done in other SRA projects such as SRA 2015/025 and CPI030 but more work is needed to complete this process and provide a range of methods for SNP detection.

4. A statistical method for determining the amount of variation that the markers explain would be an invaluable tool for determining the most valuable markers.

9. Publications

A list of publications arising from the project is given below. The full text of the papers is given in Supplementary 4.

1. Aitken KS, Farmer A, Berkman P, Muller C, Wei X, Demano E, Jackson P, Magwire M, Dietrich B, Kota R (2016) Generation of a 345K sugarcane SNP chip. International Society of Sugarcane Technologists, Chiang Mai, Thailand.

2. Demano E, Wei X, Aitken K, Kota A, Jackson P (journal paper in preparation). Genomic-enabled prediction in sugarcane using DArT and SNP markers.

Conference abstracts:

1. Aitken KS, Farmer A, Berkman P, Jackson P, Wei X, Muller C, Magwire M, Dietrich B, Kota R (2015) 11th Germplasm and Breeding , 8th Molecular Biology ISSCT workshop, Reunion.

2. Aitken KS, Farmer A, Berkman P, Muller C, Wei X, Demano E, Jackson P, Magwire M, Dietrich B, Kota R (2015). Generation of a sugarcane SNP array for genomic selection. Tropical Crops Congress, Brisbane.