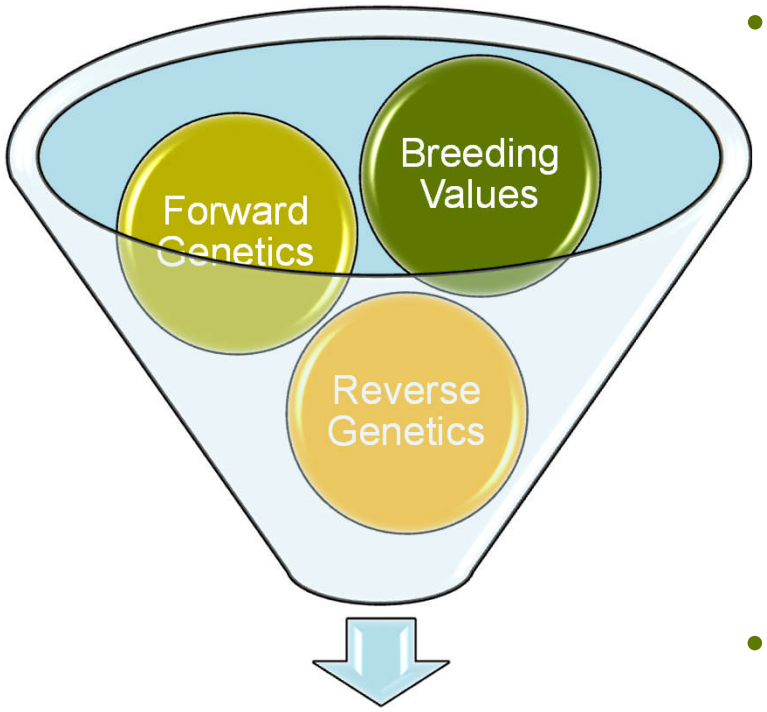
Generation of a 345K Sugar cane SNP chip

Karen S Aitken, Andrew Farmer, Paul Berkman, Cedric Muller, Mike Magwire, Bob Dietrich, Xianming Wei, Emily Deomano and **Raja Kota**

Classification: Public

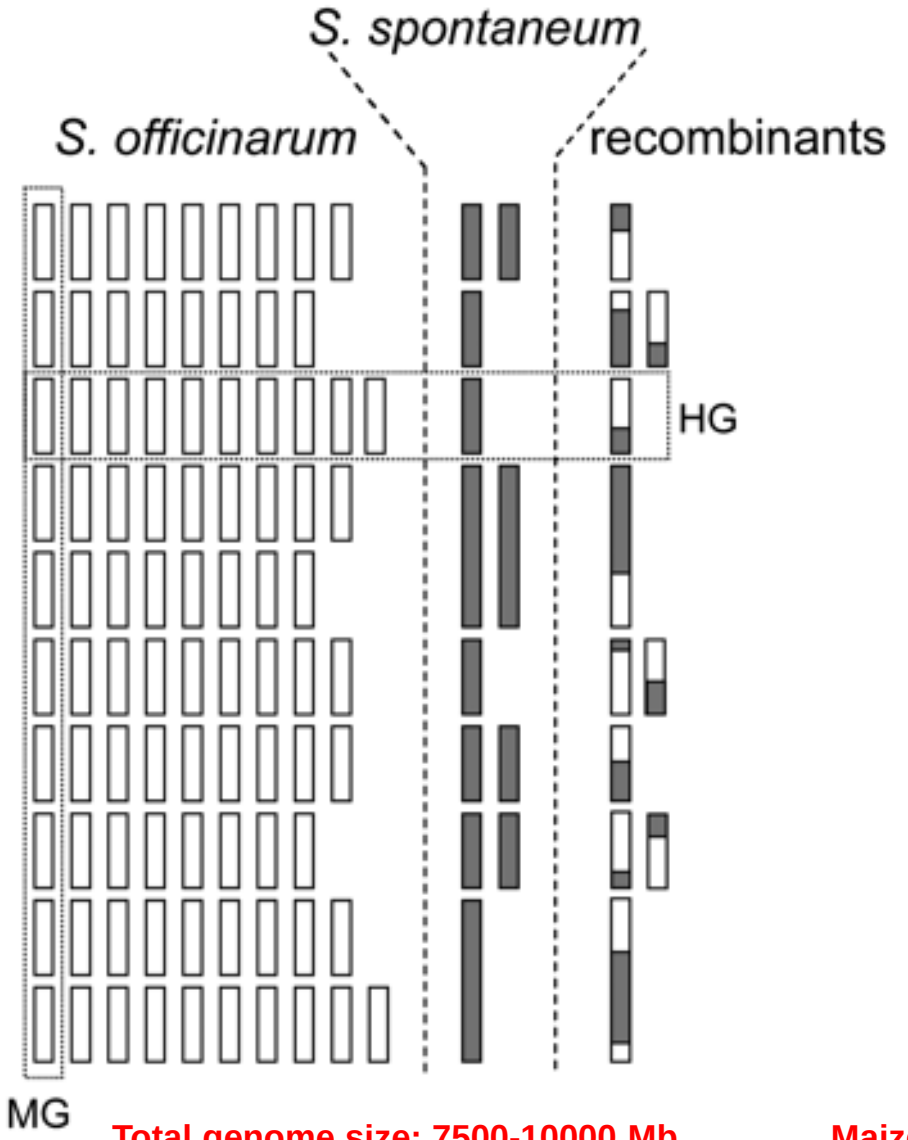
Strategy: Integrating -Omics with Breeding



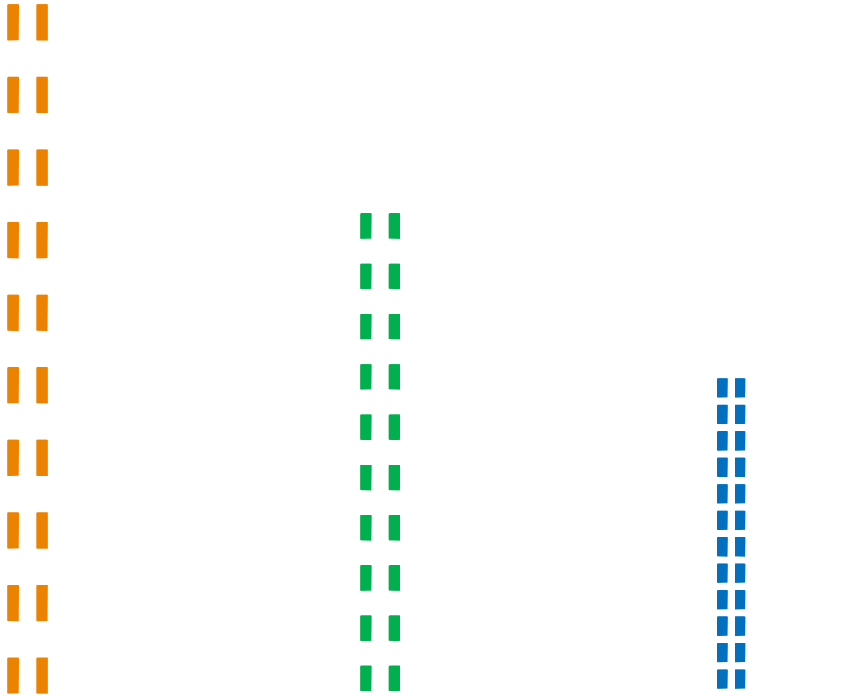
- Using marker-trait associations to assemble an ideal genotype (ideotype)
  - ♦ Forward Genetics
    - QTL mapping (bi-parental and multi-parental populations)
    - Genome Wide Association mapping
  - ♦ Reverse Genetics
    - Candidate gene association mapping
- Using multiple trait selection indices to develop lines with superior whole genome breeding values using molecular markers (genomic/genome wide selection - GWS)

syngenta

Modern Sugar Cane Genome Size



- Each individual in a cross yields a unique arrangement of chromosomes due to random pairing during meiosis



Total genome size: 7500-10000 Mb

Maize: 2500 Mb

Sorghum: 800 Mb

Rice: 400 Mb

Classification: Public

syngenta

Developing a SNP chip for GWAS/GS

- Sequence 16 core lines representing Australian and Brazilian breeding programs

- SNPs called using parameters (yet to be finalized) to be used for developing a SNP chip using Affy's "Axiom" Technology
  - ~345K SNP screening array for 480 AU clones and will include mapped SNPs from Illumina (positive controls)
  - Running a screening array ensures that the SNPs that are used on the smaller array will perform.
  - Smaller chip will be in a 96 array format
  - Maximum number of SNPs on the 96 array format would be 50K

- Association mapping populations from Australia and Brazil will be genotyped using the smaller array

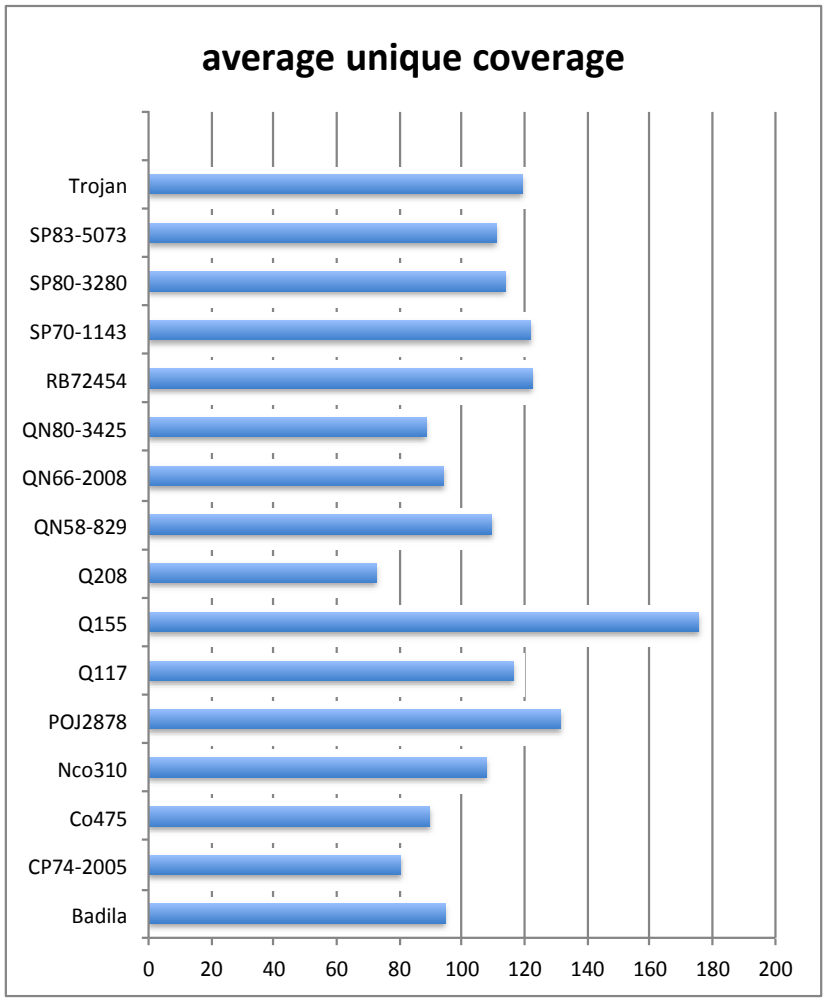
Sequencing Results

- 2 samples/lane with an expected coverage of at least 50X of a given genomic region
- Sequences were assembled using previous data from a previous project that involved sequencing two Australian lines Q165 and IJ76

Sample	Reads (in bp)
Badila	171909294
Co475	178010130
CP74-2005	154340644
Nco310	217972160
POJ2878	**119639502**
Q117	159933544
Q208	141936024
QN58-829	218476606
QN66-2008	191146084
QN80-3425	176658030
Trojan	**71507454**
Q155	342257032
SP70-1143	**104978170**
RB72454	262617118
SP80-3280	**68080482**
SP83-5073	159297874

syngenta

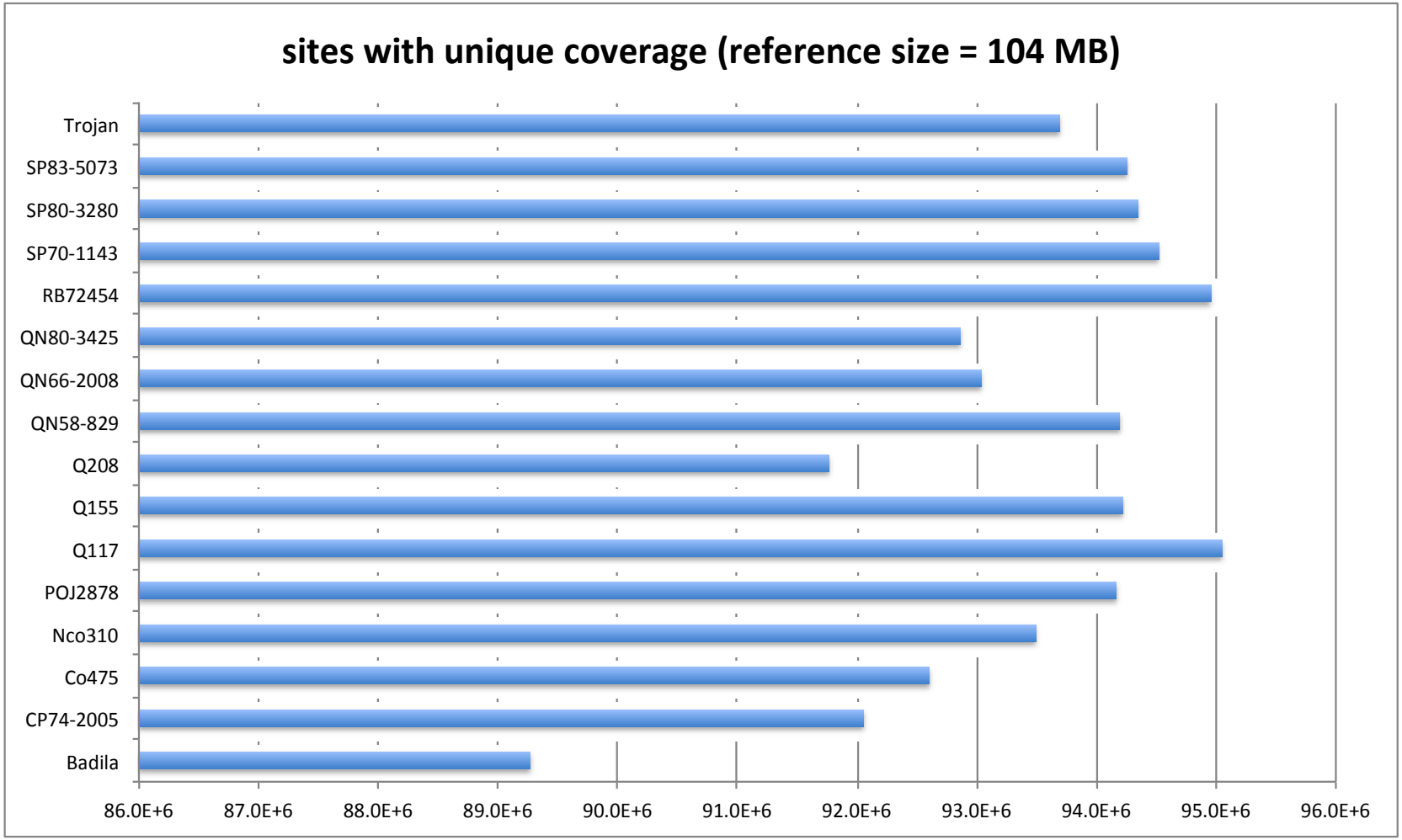
Average coverage across all Samples



Classification: Public

syngenta

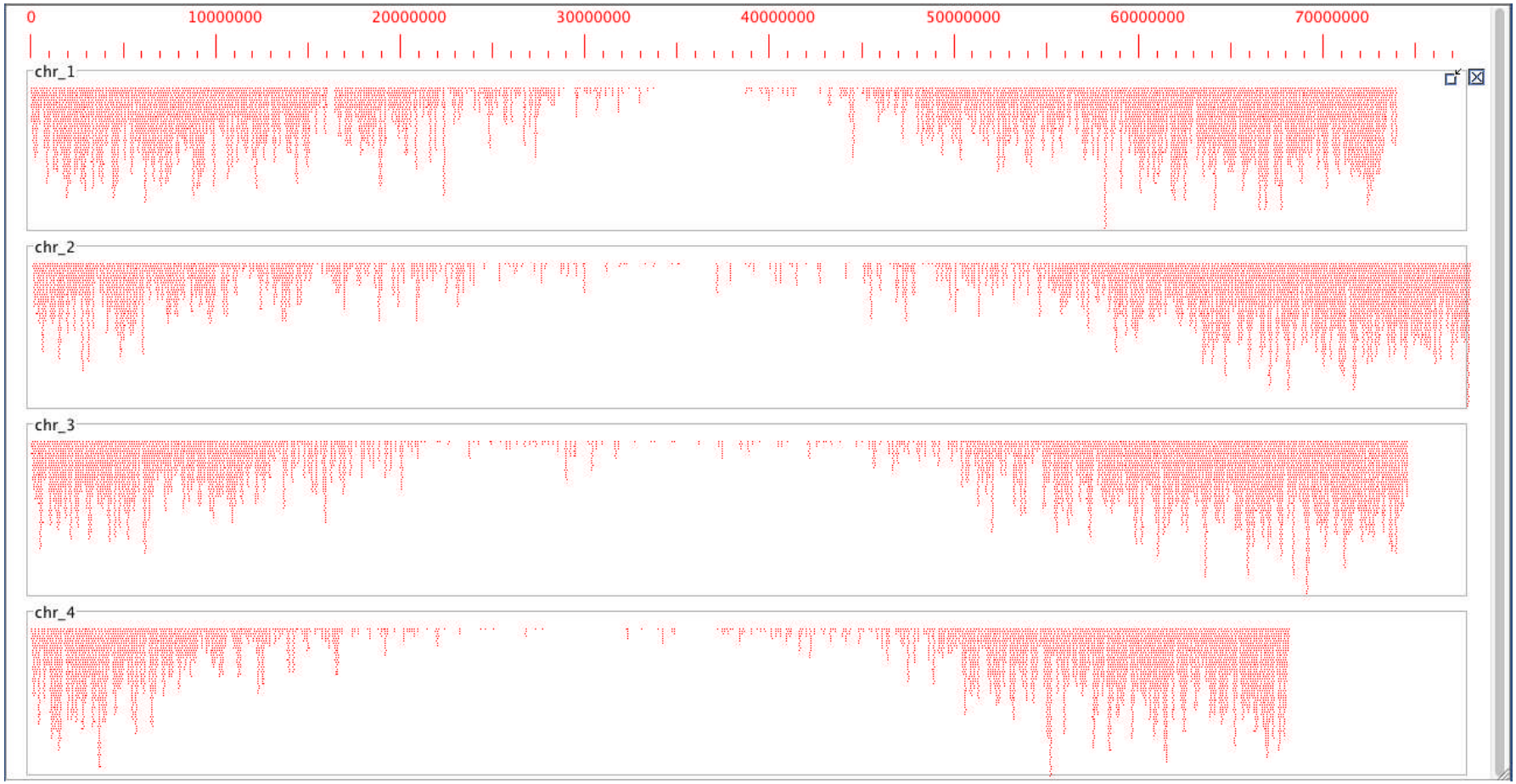
Variation in number of sites with uniquely aligning read coverage across lines



Classification: Public

syngenta

Distribution of sugarcane contigs in sorghum genome



SNP calling Criteria

For SNP calling, following parameters were selected:

1. Addressing dosage:

   Class 1: Low dose (single/double) in at least 4 lines, 0 dose in at least 4 lines and high dose in at least 1 line

   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

   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
2. Compare the total number of low dose SNPs selected based on lines derived from Australia or Brazil. Data from this analysis will be used to remove any bias in the SNP selection process
3. If using the above filters does not result in sufficient number of SNPs, reduce coverage from 50X to a lower number with the limit being 20X
4. Once SNP calling is done, align the number of SNPs that fit the above criteria to a given line (i.e. total number of SNPs selected from line 1, line 2 and so on)
5. Ensure preselected regions (DArT mini array (~400), and successfully mapped Infinium SNPs (~2400) are enriched (2-3 SNPs per marker sequence) is selected
6. Map the selected SNPs to the Sorghum reference

Classification: Public

Results from applying Dosage selection criteria

- Using 50x minimum coverage:

%	count	class code
3	50843	100 (class 1)
11	206346	010 (class 2)
37	682885	011 (class 2 and class 3)
49	892155	001 (class 3)

Total = 1832229

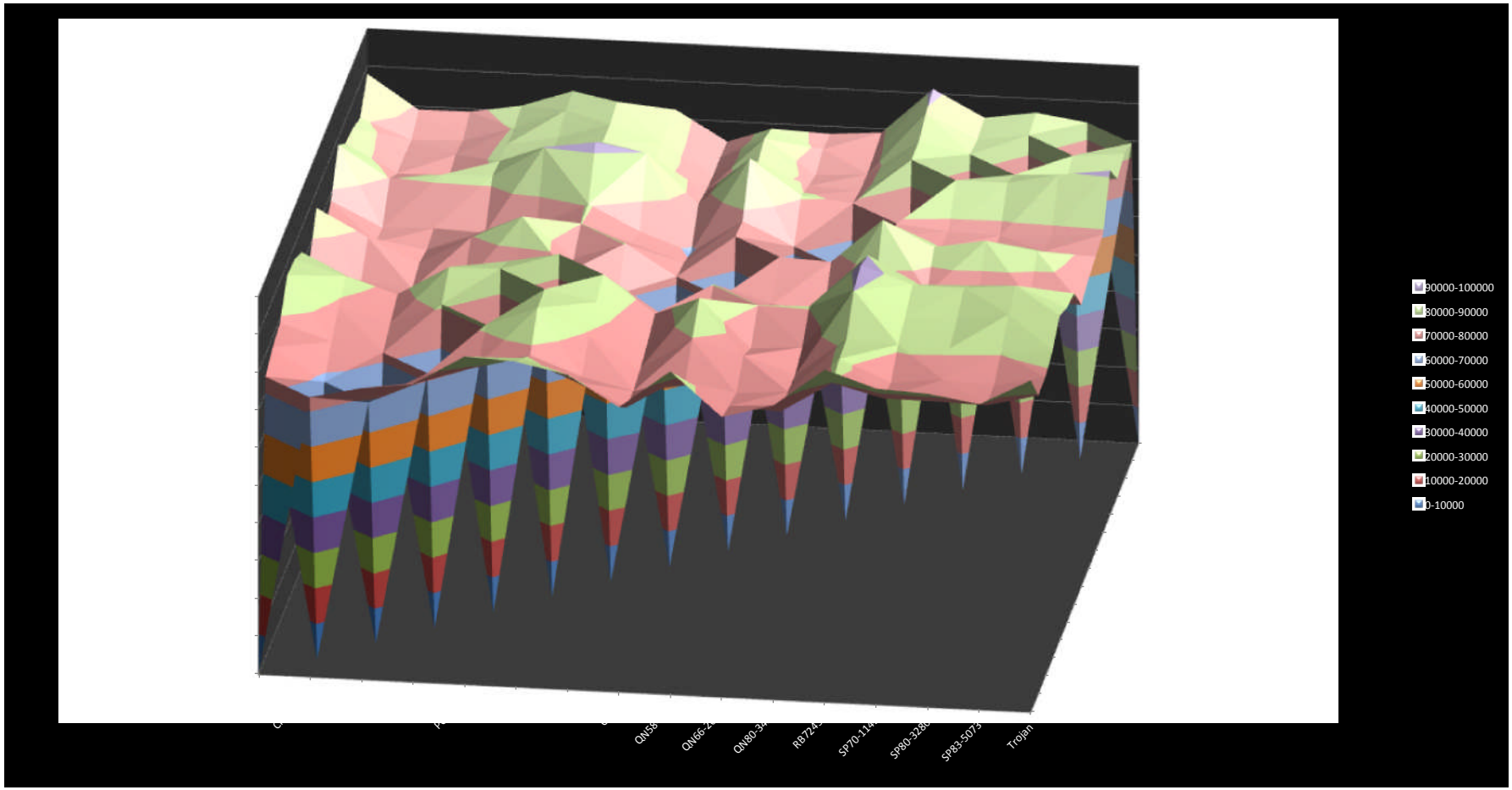
- Using 20x minimum coverage:

%	count	class code
5	131831	100 (class 1)
14	384630	010 (class 2)
38	1015206	011 (class 2 and class 3)
42	1121023	001 (class 3)

Total = 2652690

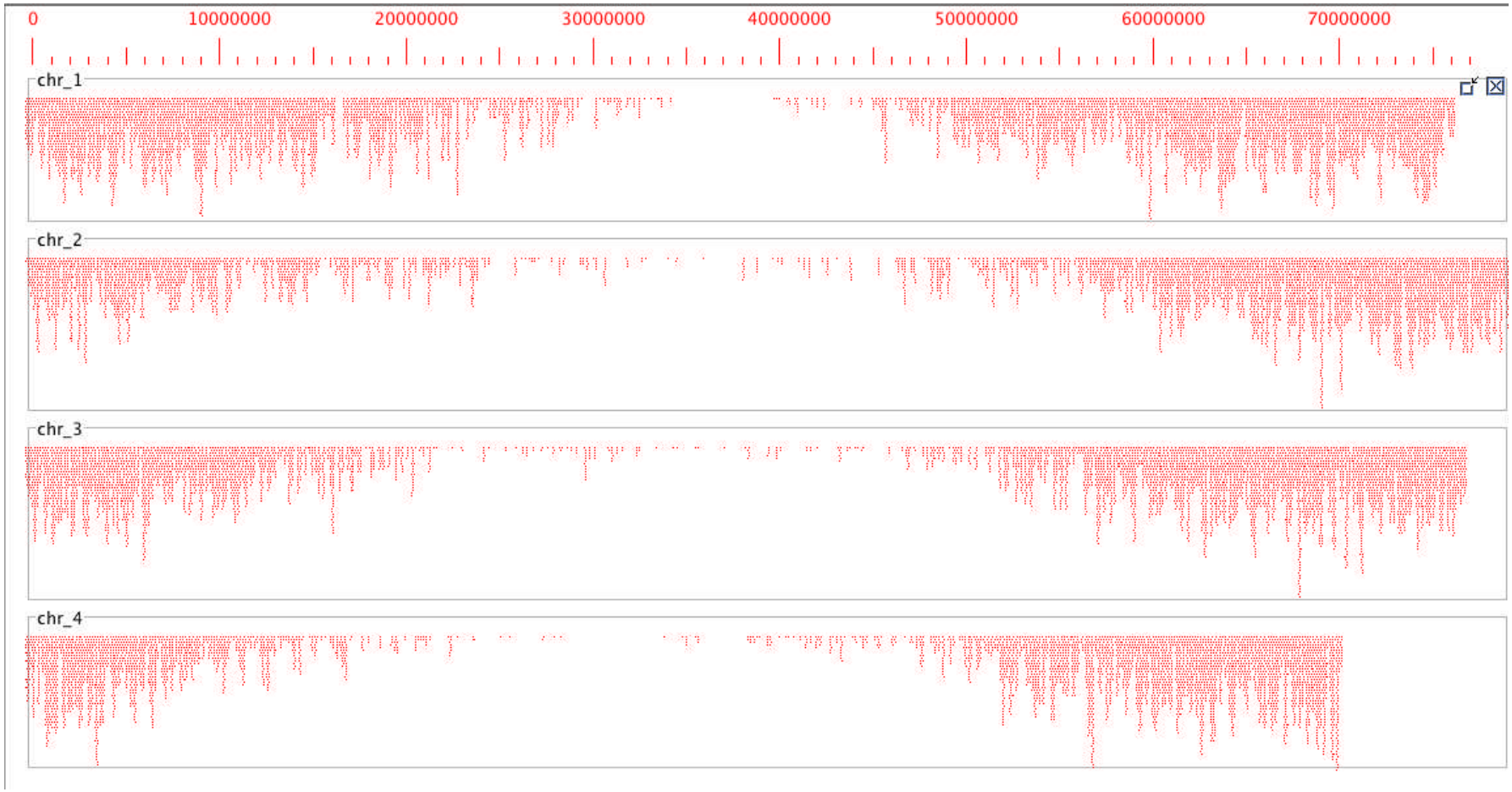
Classification: Public

Pairwise-mappable (LD/0D) variant counts



syngenta

Distribution of sugarcane contigs with mappable LD/0D variants



syngenta

Affymetrix Customized Workflow for data analysis

Generate genotypes following best practice workflow. Execute SNPolisher using Polyploid setting and apply Supplemental Variance filters (Z-score > 3) and HetvMAF = 1.9

Execute Ps_CallAdjust with threshold set to 0.1. Perform reproducibility and MI accuracy analysis and remove probesets with > 1 error in either category

Execute Ps_Metrics using adjusted call table and remove CallRateBelowThreshold SNPs with <2 clusters, <2 observations in the least populated cluster and CallRate <80% and NoMinorHom SNPs with <10 observations of the minor allele

Identify any probesets with high Variance.X.Z scores (> 3) and PHR probesets with >1 MI error for fitTetra analysis

Conversion Type	Count	Percentage
PolyHighResolution	11695	2.76
AAvarianceX	64	0.02
AAvarianceY	162	0.04
ABvarianceX	152	0.04
ABvarianceY	134	0.03
BBvarianceX	109	0.03
BBvarianceY	112	0.03

Conversion Type	Count	Percentage
MonoHighResolution	169446	39.96
NoMinorHomozygote	21366	5.04
OffTargetVariant	251	0.06
CallRateBelowThreshold	33652	7.94
Other	186263	43.92
Unexpected Het	480	0.11
Hom-Hom Resolution	162	0.04
Total number of probesets	424048	100

After applying advanced filters

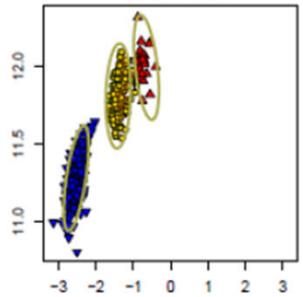
Conversion Type	Count	Percentage	Delta from initial
PolyHighResolution	10474	2.47	1221
NoMinorHom	12106	2.85	9260
CallRateBelowThreshold	25222	5.95	8430
Combined (PHR+NMH+CRB)	**48802**	11.47	---
Total number of probesets	424048	100	---

Number of probesets identified by Variance X.Z > 3 used for fitTetra analysis:	541
Number of PHR probesets identified with >1 MI error for fitTetra analysis:	902

Classification: Public

Classification of SNP calls using fit-Tetra algorithms

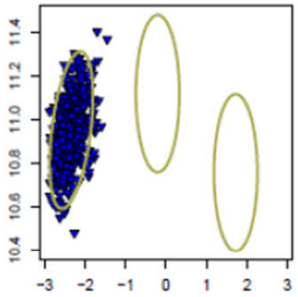
Polymorphic high resolution



Good cluster resolution and at least two examples of the minor allele.

3.3%

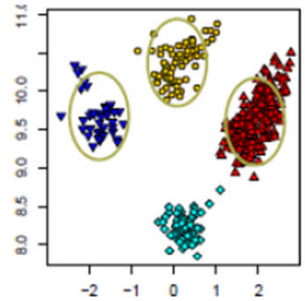
Monomorphic high resolution



Less than two examples of the minor allele usually due to low MAF of the population genotyped.

43.2%

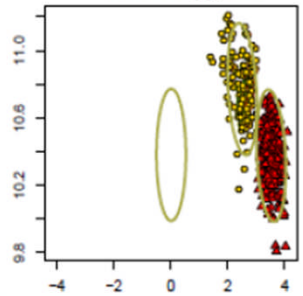
Off-target variant



Variants that are reproducible and interfere with the SNP of interest, creating a fourth cluster.

0.1%

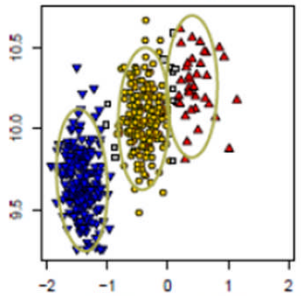
No minor homozygous



Two clusters with no examples of the minor homozygous genotypes.

5.9%

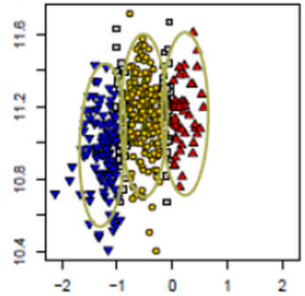
Call rate below threshold



Sample call rate is below threshold (adjustable) but other cluster properties are above threshold.

8.1%

Other



One or more cluster properties are below threshold.

39.2%

syngenta

BLAST results of ~49K SNP markers aligned to the sorghum genome (>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

Classification: Public

Australian Core program – Association mapping Panel

- Population 1 **("association mapping population")**
  - 480 clones from core program
  - Cane yield, sugar content, measured at 3 sites x 2 years
  - Disease (smut) resistance, other diseases on subset

- Subset of lines from the bi-parental mapping population

syngenta

Number of SNP and DArT markers identified associated with TCH and CCS at different p values using mixed model analysis

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

syngenta

Number of SNPs identified associated disease traits at different p values using mixed model analysis

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

Classification: Public

Acknowledgements

- Members of the Analysis, Bioinformatics and Genetics group - Syngenta
- Andrew Farmer – NCGR, Paul Berkman - CSIRO
- Sugar cane team – USA/Brazil
  - Dirk Benson
  - Ian Jepson
  - Yan Zhang
  - Michel Moraes
  - Jair Durate
  - Stacy Miles
- CSIRO/SRA - Australia
  - Karen Aitken
  - Xianming Wei
  - Phil Jackson
  - Emily Deomano
- Funding support from Syngenta and SRA (SRDC)

syngenta



Bringing plant potential to life