Explanatory notes regarding 21 sample tracking errors detected in DArT of Synthetic Hexaploids in runs DW07-159_210 and DW07-128

Compiled by Clare Johnson, Wheat CRC, January 2008

180 Primary Synthetic Hexaploids from Synthetic imports from 2001-2004 (AWCC codes ZSE01, ZSF01, ZSF03 and ZSF04, (as listed on http://mendel.lafs.uq.edu.au:8080/ICIS5/GWIS_LBY4.htm) were submitted for DArT analysis in runs numbered DW07-159 and DW07-210. Run DW07-159 was performed on over 25,000 DArT clones in order to identify those that were polymorphic in this germplasm. This information aided selection of clones for use in future analyses of similar material. Thus, in run DW07-210, the analysis was conducted on the prototype of Triticarte's version 2.5 array containing most of the markers polymorphic in wheat. In run DW07-159, scores were available for 1099 markers for 94 samples (including 4 controls) and in run DW07-210, scores were reported for the same set of 1099 markers in 91 samples (including 4 controls, and excluding AUS30673, AUS30639 and AUS30653, which failed to germinate, so wells were left blank), plus an additional 143 markers not scored in DW07-159. The overall call rate for the 1242 markers in run DW07-210 was 94.9%. CloneID is the unambiguous identifier for any markers not yet assigned a name.

DArT markers are biallelic dominant markers. Each marker is scored for each sample: 0, 1, or -. The "-" stands for missing data, in cases where a marker could not be reliably scored for that sample. For each marker, Triticarte provides the Q value, based on ANOVA, as an estimate of marker quality. Q reflects how well the two phases (Present = 1 vs Absent = 0) of the marker are separated in the sample set. Markers with Q value above 80 are considered very reliable. Triticarte advises caution using markers with Q value under 77, at which point the risk of scoring errors becomes significant.

The call rate is the number of effective scores divided by the number of samples. The PIC value measures the distribution of 0 and 1 scores in the sample set. The highest PIC value is 0.5 and means that the marker is distributed 50-50 in the population analysed.

In column C (Chromosome) Triticarte provides a tentative assignment based on the 9 well-curated hexaploid wheat genetic maps, built and analysed by their team, that can be downloaded from www.triticarte.com.au/content/wheat_diversity_analysis.html Triticarte requests that if you find any discrepancy in the chromosome assignment based on your knowledge of the material, please email them.

They noted that control varieties Kennedy and Sunco were consistently scored between service 159 and service 210, as the few scoring differences were within normal limits for scoring errors or genuine heterogeneity between various samples of the same variety. This made sense, as due to a technical error, only a single seed had been used for the preparations in service 159, whereas the recommended 10 seeds had been used for the controls in service 210, and a single seed had been used for the primary synthetic samples throughout. For DW07-159, the controls had come from the University of Sydney Plant Breeding Institute, Cobbitty. Because Triticarte detected a sample tracking problem between replicates of Spica (the genetic distance between them being no different from the average distance among all samples in

service 159), the controls in later service 210 were obtained from Mui-Keng Tan at NSW DPI, who said they had behaved as expected in her marker work.

Variety Spica also scored differently between service 159 and service 210 and, as hoped, was closer to a previous analysis of Spica sourced from NSW DPI than the sample labelled "Spica" from service 159. Unfortunately however, in service 210, the sample labelled "Seri-82" scored nearly identically to Kennedy as if a mistake had been made when extracting the DNA or filling up the plate for this sample, and was clearly different from Seri-82 in service 159. Note that Kennedy and Seri M 82 are closely related by pedigree (Kennedy = Seri M 82/Hartog). An independent data set for Kennedy will be available in the IAT DArT data to be provided by March 2008.

It was therefore decided

a) to retain the Spica result from service 210 and **discard the "Spica" result from service 159.**

b) to retain the Seri-82 result from service 159 and **discard the "Seri-82" result from service 210.**

Because of these sample-tracking issues in the controls, it seemed important to check the data set against an earlier, lower density DArT run (DW07-128) kindly made available by Francis Ogbonnaya, to evaluate the integrity of the data from the primary synthetics.

Eric Huttner (Triticarte P/L) kindly performed an analysis to compare data sets DW07-128 and DW07-210_159, based on the 491 markers in common to 2 data sets (see Appendix A: scoring).

Of a total 180 samples, three (AUS30639, 30653 and 30673) did not grow, and no comparison was possible for AUS29020 (CROC 1/AE.SQUARROSA (517) from 1993) or the following 35 samples from early nursery ZSE01, as they had not been run in service DW07_128:

AUS29020

AUS29636, AUS29637, AUS29638, AUS29640, AUS29641, AUS29642, AUS29644, AUS29645, AUS29648, AUS29649, AUS29651, AUS29652, AUS29653, AUS29654, AUS29655, AUS29656, AUS29657, AUS29659, AUS29660, AUS29664, AUS29666, AUS29667, AUS29668, AUS29669, AUS29670, AUS29672, AUS29673, AUS29674, AUS29675, AUS29676, AUS29677, AUS29678, AUS29681, AUS29684, AUS29685,

Within the same region of the sample plate (Appendix B), 8 lines (AUS29639, AUS29646, AUS29658, AUS29671, AUS29679, AUS29680, AUS29682 and AUS29683) were confirmed to have tracked correctly. There was some evidence of clustering rather than a systematic error. Therefore data for the above 36 samples has been left in the spreadsheet; however, please notify us if your data suggests that any of these samples have been mislabelled.

In the 141 samples of run DW07-159_210 left available for comparison, the identities of 122 samples were confirmed and 19 suspected tracking errors were identified. (One sample was clearly in common but with a different name: AUS26860 and "equivalent of AUS26860" in the different runs.) In order to check whether any samples in service 159_210 were present in service 128 under another name, Eric Huttner then

calculated a first distance matrix for the ambiguous lines, that is, all lines which were not clearly replicated between the two experiments, based on the 491 markers in common, and looked for near-identical samples.

The resulting table ("Distance1" worksheet) showed that 6 lines of service 210 had a probable counterpart in 128. Then he created a distance matrix on a more comprehensive list of lines from service 128, having removed the duplicate lines and the lines from 210 with a partner (to remain within the Excel limit of 256 columns). The second distance matrix ("Distance2") identified a few more samples. Care was taken in using lower thresholds of distance to assume identity, as the distribution of distances suggested that beyond 0.03 - 0.04, identity could not be assumed as the likelihood of scoring errors increased. This can be checked by looking at the markers separating the putative replicates: if all the differences are in the low quality markers then scoring errors are more likely than identity. If the differences are spread out over the whole range of marker quality, then one cannot exclude a true (small) difference between the samples.

Conclusions

On the basis of the analysis with reference to 491 markers, 8 sample tracking errors were found to have occurred at Horsham within service 128, one involving 4 lines.

The tracking errors in the elite SHW lines were:

AUS30262 = AUS30266 = AUS33378 = AUS33380;
AUS30290 = equivAUS26860;
AUS30296 = AUS30267;
AUS30299 = AUS33423.

Tracking errors in other SHW lines were:

AUS30282 = AUS30283
AUS30628 = AUS30629
AUS33405 = AUS33415
AUS30407 = AUS33413

This would have occurred prior to September 2006, and is of particular concern if it may have affected what has been deposited at AWCC. Potential errors in the accessions could be checked by comparing AWCC and CIMMYT sources of these lines, particularly the elite SHW AUS30262, 30290, 30296 and 30299.

Although in service DW07-159_210 with 1,099 DArT markers there were 6 differences over 1099 markers between AUS30290 and 26860, and very slight differences could be seen between AUS30262, 30266 and 33380 (AUS33378 was not run), between AUS30299 and 33423, note that as samples properly tracked should have $\geq$99% consistency, this difference of 0.54% is not sufficient to discriminate the samples.

On top of the Horsham tracking errors, there were a further 13 sample tracking errors at Cobbitty. These are shown in the "List of tracking errors" worksheet in the Excel file. The identity of AUS30265_210 was confirmed. The lines with sample tracking errors either within service 159-210, or in 159-210 as compared to service 128, were thus excluded from the data in the main worksheet, but were labelled and copied into

a separate worksheet, along with correct runs (from service 159-210) of lines they match in service 128. They have been colour coded for clarity. Those samples that tracked correctly in service 159-210 but had mis-been tracked in run 128 were colour coded for ease of identification. This was done to enable users of the data to check the current data against any anomalous data they may have encountered in their own work if samples were mislabelled prior to shipping.

Acknowledgements: Eric Huttner, Triticarte P/L, Francis Ogbonnaya, DPI Vic/ICARDA, Jayne Wilson, DPI Vic, Dinesh Khatcar and Kate Vincent, USyd PBIC, Mui-Keng Tan, NSW DPI.

Appendix A:

Scoring the consensus between the 2 sets:

1 vs 1 or 1 vs X was scored as 1;

0 vs 0 or 0 vs X was scored as 0;

X vs X was scored as - ;

1 vs 0 was scored as X.

That is, the X's represent the errors and counting them provides an indication of the identity of the 2 extracts. Samples not tracked properly will have about 50% of their scores consistent by chance. Samples tracked properly should have $\geq$99% consistency. (Note that the call rate of the 2 data sets was not taken into account: when a score (1 or 0 in one set) is facing a missing data point (X in the other set), there is no effective comparison, and the consensus becomes overestimated.)

Appendix B: Microtitre sample plate layout and sample tracking errors detected

	A	B	C	D	E	F	G	H		DW07-210
1		B1 AUS30305								
2				D2 AUS29678						Samples confirmed vs
3	Blank (AUS30639, no seed)									F.Ogbonnaya DW07-128
4						F4 AUS30652	Blank (AUS30653, no seed)			
5				D5 AUS30658						Tracking errors in
6							G6 AUS30670			DW07-210
7	A7 AUS30672	Blank (AUS30673, no seed)								
8		B8 AUS33686	C8 AUS33387	D8 AUS33388			G8 AUS33392			Samples not run by
9		B9 AUS33395	C9 AUS33396					H9 AUS33402		Francis so can't check
10										for tracking errors
11					E11 AUS33417					
12			Spica (10 seed, MKTan)	Kennedy (10 seed, MKTan)	??Seri-82 -looks like Kennedy (10 seed, MKTan)	Sunco (10 seed, MKTan)	no sample	no sample		no DNA

	A	B	C	D	E	F	G	H	DW07-	159, plate 1, Sept 2006
1	A1 AUS29636	B1 AUS29637	C1 AUS29638		E1 AUS29640	F1 AUS29641	G1 AUS29642	H1 AUS29644		Samples confirmed vs
2	A2 AUS29645		C2 AUS29648	D2 AUS2969	E2 AUS29651	F2 AUS29652	G2 AUS29653	H2 AUS29654		F.Ogbonnaya DW07-128
3	A3 AUS29655	B3 AUS29656	C3 AUS29657		E3 AUS29659	F3 AUS29660	G3 AUS29664	H3 AUS29666		
4	A4 AUS29667	B4 AUS29668	C4 AUS29669	D4 AUS29670		F4 AUS29672	G4 AUS29673	H4 AUS29674		Tracking errors in
5	A5 AUS29675	B5 AUS29676	C5 AUS29677			F5 AUS29681				DW07-159_plate1,
6	A6 AUS29684	B6 AUS29685		D6 AUS29020						Sep-06
7										
8						F8 AUS32072				Samples not run by
9	A9 AUS30275			D9 AUS30278		F9 AUS30280				Francis so can't check
10										for tracking errors
11		B11 AUS30292		D11 AUS30295						
12			??**not**Spica (1 seed, PBIC)	Kennedy (1 seed, PBIC)	Seri-82 - poor but looks right (1 seed, PBIC)	Sunco (1 seed, PBIC)	no sample	no sample		no DNA