Title Help notes for scoring SSR alleles in *Phytophthora infestans*

Version *Version 2.0* (29 April 2019)

Author David E L Cooke, The James Hutton Institute, Dundee, david.cooke@huttonl.ac.uk

This guide assumes you are using an ABI capillary sequencer (e.g. 3730) and ABI GeneMapper software 3.7 to 5.0. It also assumes using the 12-plex SSR marker system in Li et al., (2013).

Li Y, Cooke DEL, Jacobsen E, Van Der Lee T, 2013. Efficient multiplex simple sequence repeat genotyping of the oomycete plant pathogen Phytophthora infestans. *Journal of Microbiological Methods* **92**, 316-22.

Based on experience of scoring peaks from thousands of isolates over a 20 year period!

1. General logistical issues

a. Name samples before entry into ABI capillary sequencer so all samples have a logical name. Imagine the confusion if all samples in a large project are exported with the names "1" to "96"! Also name batches of samples/96well plates (Run 1 – etc) in a folder of the same name (see example Figure 1). This allows a simple means of checking back through raw peak data after data is exported to Excel sheets or *poppr* analysis (erroneous data will stand out in PCA for example).

b. Store data in separate projects (e.g. by year). Do not store too many samples (e.g. >1500-2000) in a single project as it becomes slow to load and can crash. We have had problems with memory allocation in GeneMapper.

c. Keep raw txt files when you export data from GeneMapper. Helps check back to raw data.

d. Backup project files of edited data (projectname.ser files) as peak calls are lost when GeneMapper Analysis is re-run.

e. Water controls. We run one water-only PCR sample on each plate to check for contamination. If peaks present in the control sample then one has to take care in interpretation of samples with weak peaks in the same run/plate.

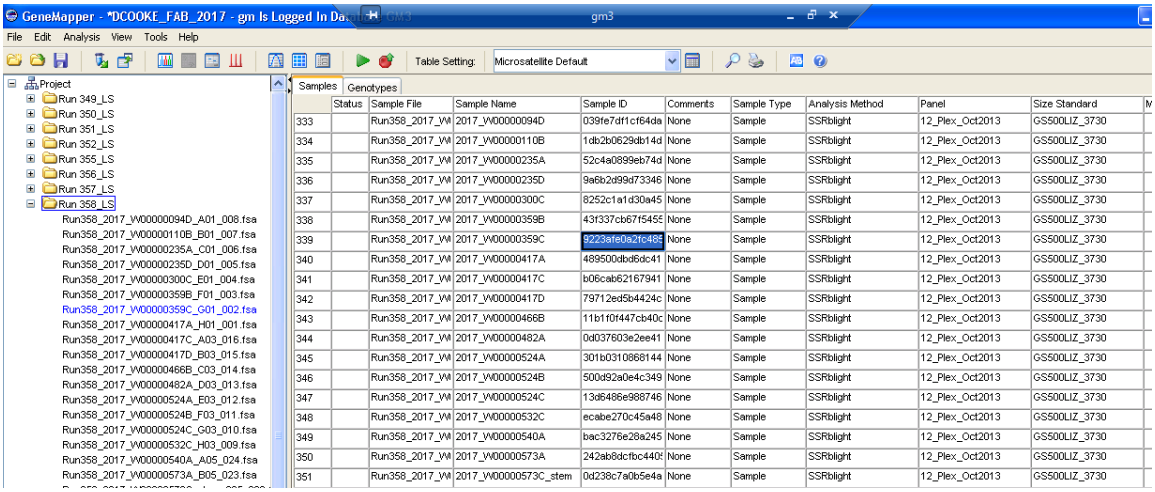


Figure 1. Example of a typical project (DCOOKE_FAB_2017) in which 8 96-well plates have been run (Run349 to Run358) with one folder open to show the sample names and the sample file name which includes the Run number the sample name and the well number.

2. **Analysis steps**
    a. Run analysis on 1 plate at a time and score these peaks before exporting.
    b. Always chose "*Analysis/Analyse Selected Samples*" only as all your peak edits will be lost each time <u>all</u> data is re-analysed by GeneMapper.
    c. Always check size-standard issues before starting editing. Normal problems are off-scale data due to insufficient dilution in some samples. Off-scale data generates rogue size peaks that 'confuses' the size-calling algorithms
3. **Common peak scoring issues**
    a. *Spectral Overlap* (excessive off-peak signals from one dye causing a spurious peak of the same size in another colour). It is thus best to look at data for all 4 colours and over whole size range first and then focus in on the specifics of uniform peak calling as a second step. Spectral overlap peaks look very different from 'real' SSR peaks as they do not have the classic stutter patterns (Figure 3).
    b. *Dye blobs and artefacts* – some commonly occur in same place independent of the sample and do not have typical stutter patterns.
    c. *Locus-specific amplification patterns* - each locus is different – dinucleotide repeats will look different from trinucleotides and the length of the set of tandem repeats will affect pattern of stutter peaks. All SSR peaks have some sort of stutter that reflects slippage of the DNA strand during amplification across the tandem repeats. Also remember that some isolates of *P. infestans* are diploid and others may be triploid which will affect the pattern of peaks. Some general rules follow:
        i. Stutter patterns depend on the locus. D13 has between 11 copies (band size 106 bp) and 71 copies (band size 218 bp) of the CT repeat. Whereas Pi70 has between 6 (189bp) and 9 (198bp) repeats of the AAG repeat. D13 generally has more stutter (Figure 2 and 3) than Pi07 (Figure 4). The amount of stutter can vary across the size range within a locus in some cases (Figure 5). Dinucleotides generally have more stutter than tri-nucleotide repeats.

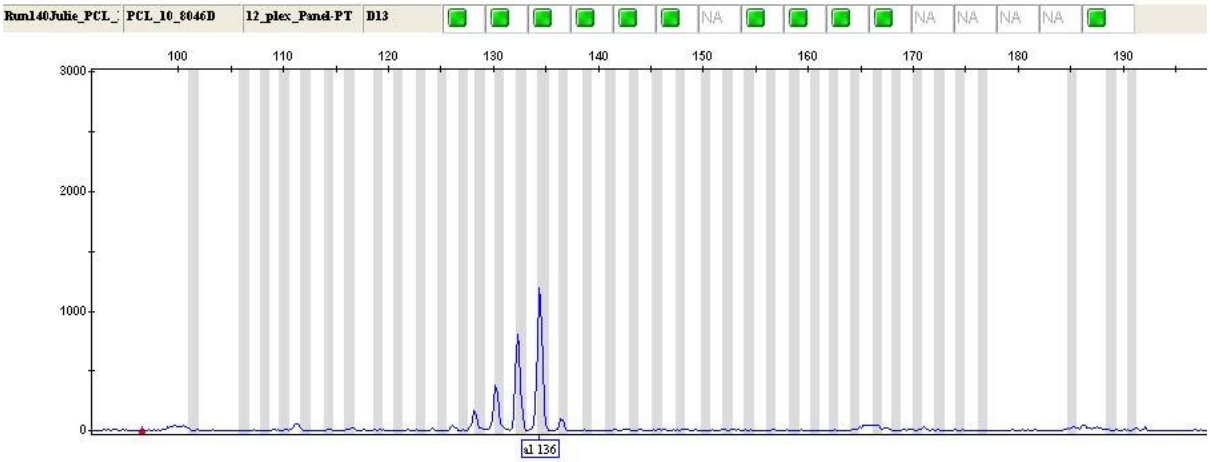


Figure 2. Amplification of 136 bp homozygous allele of locus D13 showing typical pattern of four smaller stutter peaks at 2bp intervals. Note the <u>very</u> slight amplification of a larger 138 bp product.

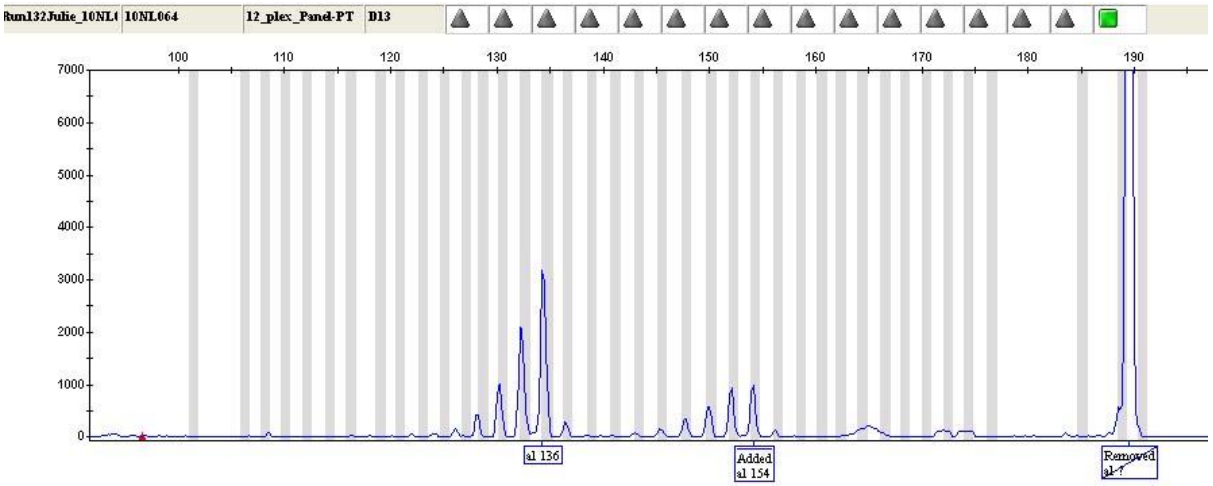


Figure 3. Amplification of 136 and 154 bp heterozygous alleles of locus D13 showing 2bp stutter pattern plus the less efficient amplification of the larger 154 bp product. Also note the spectral overlap peak at around 190bp that has no stutter peaks.

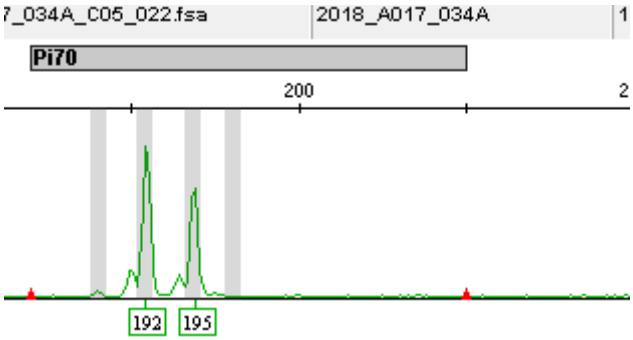


Figure 4. Amplification of 195 and 195 bp heterozygous alleles of locus Pi70 showing almost no stutter peaks. A slight 'shoulder' peak 1bp smaller than each allele is present and a very weak stutter 3bp lower than the 192bp peak is evident.

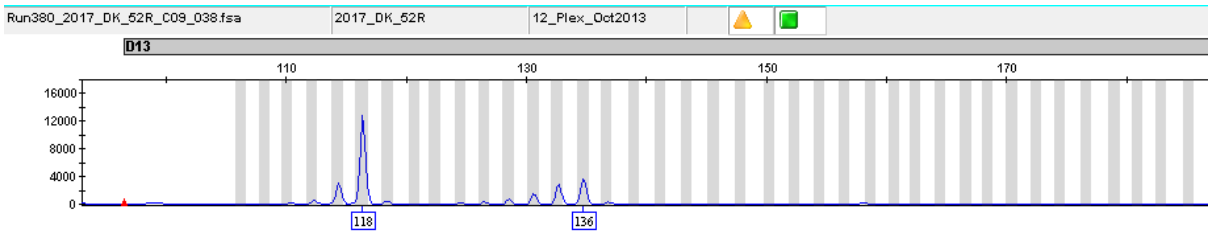
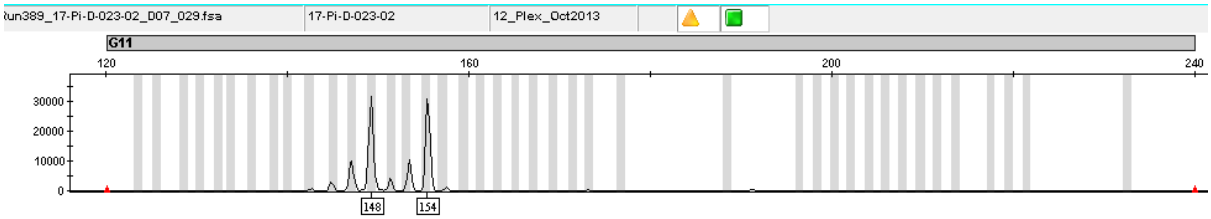


Figure 5. Amplification of 118 and 136 bp heterozygous alleles of locus D13 showing less stutter associated with the fewer repeats amplified in the 118 allele (17) than the 136 bp allele with 26 repeats.

ii. In <u>almost all</u> cases the largest peak in a stutter series is the 'real' allele (Figure 6a). However, there are exceptions such as the additive effects of a 'real' peak combined with the stutter peak of another peak that that is only one dinucleotide repeat different in size (Figure 6b).

a



b

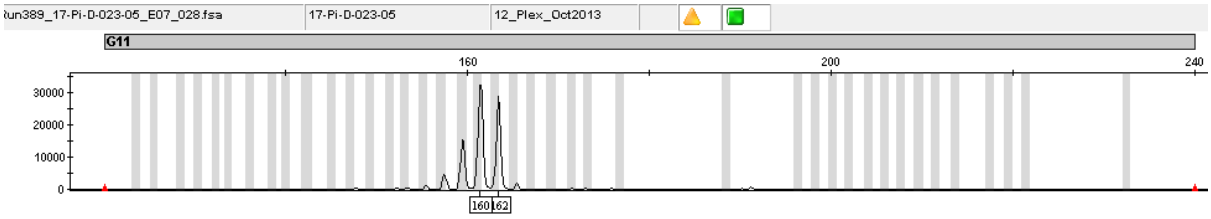
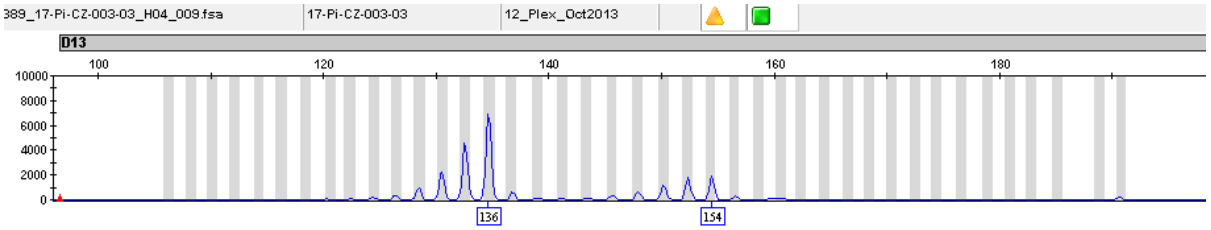


Figure 6. Cumulative stutter peaks. a) typical amplification of 148 and 154 bp heterozygous alleles at locus G11 with their stutter peaks independent of one another. b) amplification of 160 and 162 bp heterozygous alleles at locus G11 with the 160 bp allele being taller than the 162 bp allele because of an accumulation of the 'real' 160 bp peak plus the first stutter peak of the 162 bp allele.

d. Triploid samples. Many clonal lineages are triploid and the presence of three alleles at a single locus can generate distinctive patterns. Firstly, peak height ratios are skewed in favour of the allele with 2 copies (see Figure 7a). Secondly, mutations at heterozygote loci

a



b

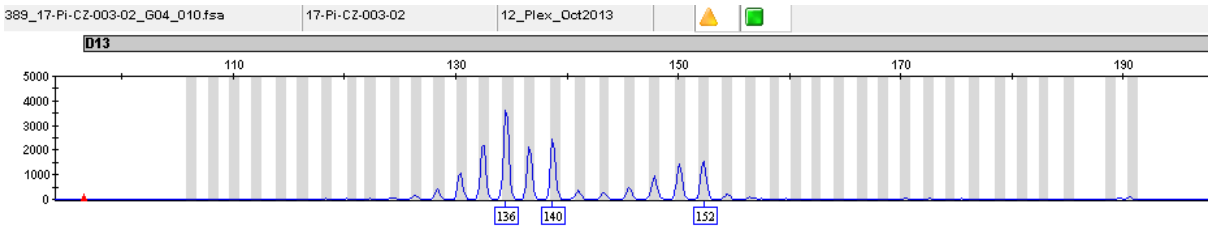


Figure 7. Triploid patterns. a) amplification of 136 and 154 bp heterozygous alleles at locus D13 in a triploid isolate of genotype EU_13_A2. Two copies of the 136 bp loci accounts for its taller peak. b) amplification of a different isolate of the same genotype in which 2 stepwise mutations in one copy of the 136 allele has generated the 140 bp allele. In addition, the former 154 bp allele has contracted by one dinucleotide repeat to generate a 152 bp allele.

can generate additional alleles so isolates commonly have three alleles (Figure 7b and 8b). All three alleles should be scored as new population genetic analysis software such as *poppr* can accommodate this data.

a

b

Figure 8. Triploid patterns. a) amplification of 154 and 160 bp heterozygous alleles at locus G11 in a triploid isolate of genotype EU_13_A2 with two 160 bp loci which accounts for the taller 160 bp peak. b) amplification of a different isolate of the same genotype in which mutation has generated an additional 150 bp allele.

e. Anomalous peak sizes. These can form when a repeat changes in size due to a partial repeat sequence or due to an DNA base insertion or deletion in the DNA sequence flanking the tandem repeats (Figure 9). Subsequent population genetic analysis that accounts for step-wise mutation in the number of repeats will require peaks that conform to the expected repeat pattern (e.g. dinucleotide alleles called 166 and 170 bp with a matching repeat series such as 4 and 6 repeats in Pi04). See appendix of repeat patterns and names at the end of this document. It is important to NOT to record peaks out of the expected repeat patterns. In clonal genotype EC-1 for example a 1bp deletion in P4B 205 allele generates a consistent 204 bp peak but recording this 204 peak would prevent analysis so this is recorded as 205 bp.

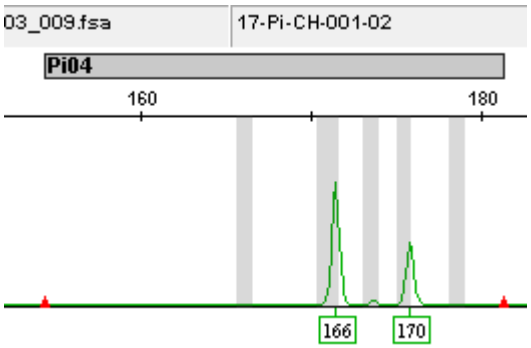
f. Variation in peak height between loci. This depends on the fluorescent dye and PCR efficiency. Oligonucleotide primers can lose fluorescent dyes if exposed to light so always keep primers in dark (e.g. foil lid in ice box) when out on the bench. Make small aliquots (we use enough to run 2 x 96 samples – 2 plates) to prevent exposure to light and to freeze-thaw cycles.

g. Null alleles. Failure to amplify can be due to a loss of the locus or a mutation in the primer binding region and is relatively common in locus D13 and occurs rarely in G11. You can tell if a locus is null in a particular isolate if it consistently fails to amplify when other loci for the same sample amplify well (Figure 10). The sample should be repeated to confirm this.

h. Mixed DNA samples – peaks of >1 isolate present. This is a lot more obvious when one is working with clonal populations but can be hard to detect in sexual populations. Look for anomalous peak numbers – for example I have never observed 4 peaks at a single locus any uncontaminated sample.

i. Consistent peak calling. Edit peaks first - then change to "Genotypes" tab page and sort data by allele size. Then view peak data with all isolates with the same recorded alleles side-by-side one locus at a time. This allows you to check for consistency in peak calling over longer time periods or between different staff members.

a



b

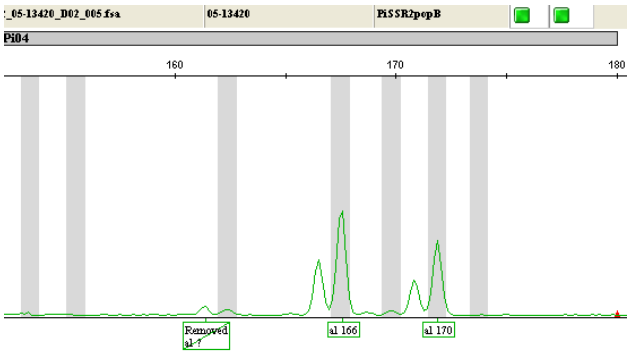


Figure 9. Anomalous peak sizes. a) typical amplification of 166 and 170 bp heterozygous alleles at locus Pi04 with no stutter peaks. b) artefact peaks also amplified 1bp smaller than the real 166 and 170 bp peaks NB b) is an image of the same alleles as in b) but from an older set of results from before the re-design of the primers for the 12-plex assay. The allele sizes are smaller than those in a). This highlights that it is the allele name that is important and not the peak size. See allele sizes and names for Pi04 in Appendix 1 (below).

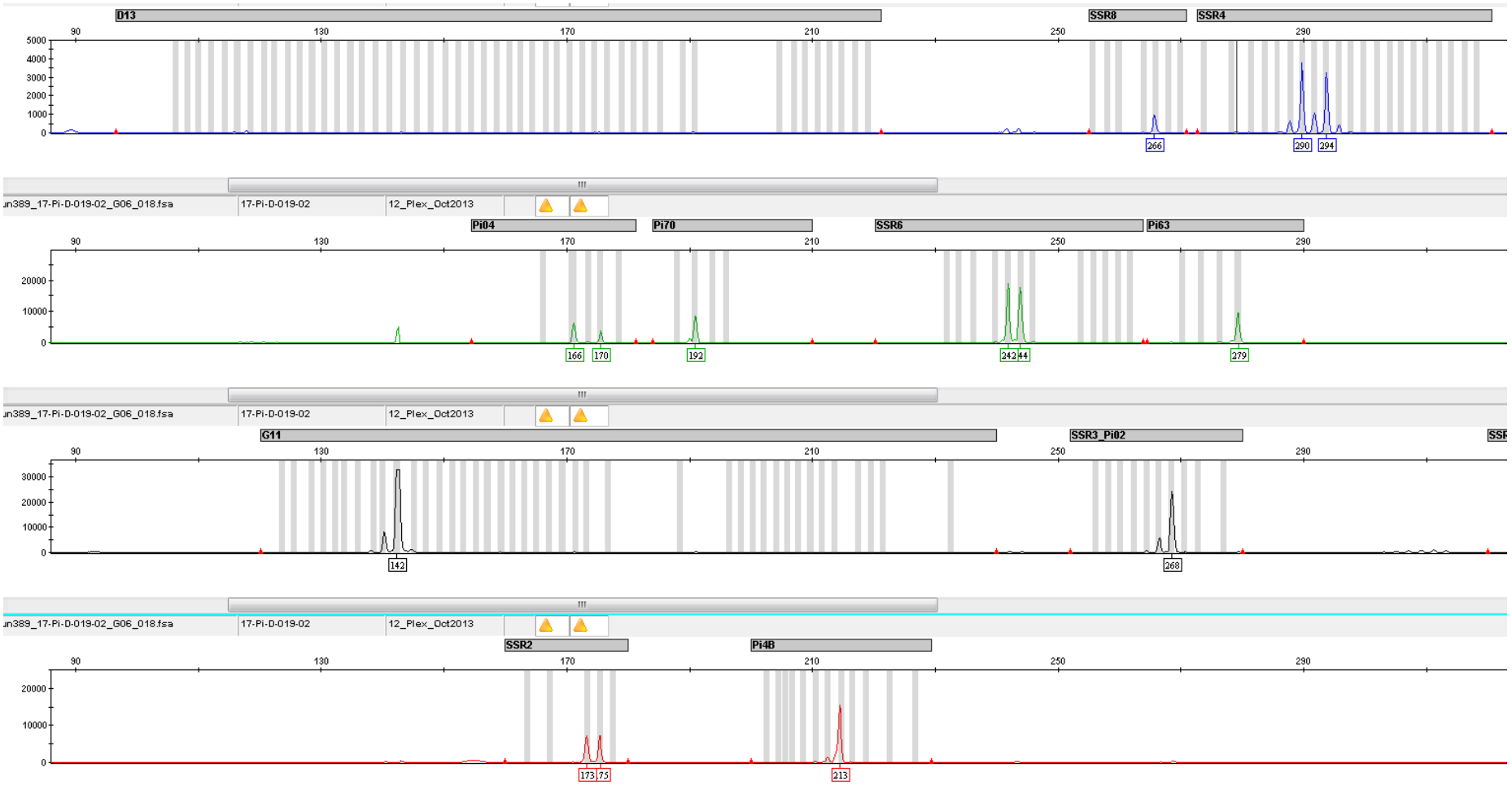


Figure 9. Null alleles. Alleles are amplified clearly in 11 loci but absent in locus D13.

Appendix 1 Table of the 12 Markers (Loci) showing the fluorescent dye colour and the type of repeat (dinucleotide or trinucleotide). The allele names for each, their corresponding size in base pairs (bp) and the number of repeats each represents are also shown. Note that the name and the peak size do not always match exactly.

Version	GM v 4.1		
Chemistry Kit	12_plexPRI_SCRI Kit		
BinSet Name	Blight SSR binset		
Panel Name	12_Plex_Oct2013		
Allele name	**Allele size**	**Number of repeats**	
Marker Name	***SSR8***	*Blue*	*Dinucleotide*
256	255.7	4	
258	258.1	5	
260	259.8	6	
264	263.9	8	
266	265.8	9	
270	269.9	11	
Marker Name	***SSR6***	*Green*	*Dinucleotide*
232	231.8	0	
234	233.8	1	
236	236.1	2	
238	238.1	3	
240	239.6	4	
242	241.7	5	
244	243.8	6	
246	245.8	8	
254	253.7	12	
256	255.8	13	
258	257.6	14	
260	259.8	15	
262	261.7	16	
Marker Name	***SSR4***	*Blue*	*Dinucleotide*
274	273.7	2	
278	278.2	4	
282	281.4	7	
284	283.7	8	
286	285.7	9	
288	287.8	10	
290	289.7	11	
292	291.7	12	

294	293.6	13	
296	295.7	14	
298	297.6	15	
300	299.7	16	
302	301.9	17	
304	304.0	18	
306	305.7	19	
308	307.9	20	
310	309.9	21	
312	312.1	22	
314	314.0	23	
316	316.3	24	
318	318.1	25	
322	323.2	27	
326	327.4	29	
328	329.3	30	
330	331.4	31	
336	337.4	34	
340	341.3	36	
344	345.3	38	
346	347.2	39	
348	349.1	40	
Marker Name	***SSR2***	*Red*	*Dinucleotide*
163	163.5	1	
167	167.3	3	
173	173.3	6	
175	175.4	7	
177	177.4	8	
Marker Name	***SSR11***	*Yellow*	*Dinucleotide*
329	328.4	1	
331	330.7	2	
339	337.9	6	
341	340.9	7	
343	342.6	8	
345	346.3	9	
355	356.0	14	
Marker Name	***SSR3_Pi02***	*Yellow*	*Dinucleotide*
256	256.1	9	
258	258.1	10	
260	260.1	11	
262	262.3	12	
264	264.3	13	
266	266.3	14	
268	268.5	15	
270	270.5	16	
272	272.8	17	
276	276.9	19	

Marker Name	***D13***	*Blue*	*Dinucleotide*
108	106.2	12	
110	108.2	13	
112	110.0	14	
114	112.1	15	
116	114.2	16	
118	116.3	17	
120	118.4	18	
122	120.7	19	
124	122.4	20	
126	124.5	21	
128	126.5	22	
130	128.5	23	
132	130.6	24	
134	132.5	25	
136	134.8	26	
138	136.6	27	
140	138.9	28	
142	141.1	29	
144	143.3	30	
146	145.6	31	
148	147.8	32	
150	150.2	33	
152	152.3	34	
154	154.3	35	
156	156.4	36	
158	158.6	37	
160	160.6	38	
162	162.4	39	
164	164.4	40	
166	166.6	41	
168	168.4	42	
170	170.4	43	
172	172.5	44	
174	174.4	45	
176	176.6	46	
178	179.0	48	
180	180.9	49	
182	182.8	52	
184	185.1	54	
188	188.9	56	
190	190.8	57	
204	204.6	58	
206	206.9	65	
208	208.7	66	
210	210.8	67	
212	212.7	68	
214	214.7	69	
216	217.0	70	

218	219.0	71	
Marker Name	***Pi04***	*Green*	*Dinucleotide*
160	166.1	1	
166	171.2	4	
168	173.5	5	
170	175.4	6	
172	178.5	7	
Marker Name	***Pi63***	*Green*	*Trinucleotide*
270	270.2	5	
273	273.2	6	
276	276.2	7	
279	279.2	8	
Marker Name	***Pi70***	*Green*	*Trinucleotide*
189	188.0	6	
192	190.8	7	
195	193.7	8	
198	196.0	9	
Marker Name	***G11***	*Yellow*	*Dinucleotide*
124	123.5	11	
126	125.5	12	
128	128.5	13	
130	130.3	14	
132	132.3	15	
134	133.9	16	
136	136.0	17	
138	138.5	18	
140	140.0	19	
142	142.2	20	
144	145.0	21	
146	147.1	22	
148	149.3	23	
150	151.4	24	
152	153.1	25	
154	155.2	26	
156	157.0	27	
158	159.4	28	
160	161.1	29	
162	163.1	30	
164	165.1	31	
166	167.0	32	
168	169.2	33	
170	171.4	34	
172	173.2	35	
176	176.7	37	
188	188.5	43	
196	196.4	47	

198	198.3	48	
200	200.2	49	
202	202.0	50	
204	204.1	51	
206	205.8	52	
208	207.7	53	
210	209.7	54	
212	211.5	55	
214	213.6	56	
218	217.5	58	
220	219.5	59	
222	221.4	60	
232	232.5	65	
Marker Name	***Pi4B***	*Red*	*Dinucleotide*
201	202.6	20	
203	204.5	21	
205	206.6	22	
207	208.5	23	
209	210.5	24	
211	212.5	25	
213	214.6	26	
215	216.4	27	
217	218.6	28	
221	222.6	30	
225	226.7	32	