final report

Project code: B.BSC.0095 and B.BSC.0097

Prepared by: Dr James Kijas, Dr John Henshall and Dr Jill Maddox

Date published: September 2014

ISBN: 9781740362481

PUBLISHED BY
Meat & Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

SNP selection and pre-commercial development of a high accuracy parentage assay for sheep industry use

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

The ability to correctly assign parentage is important to producers. The advantages of DNA based methods are well known, however industry uptake has been limited due to cost. To reduce cost and increase accuracy, this project sought to develop and evaluate panels of SNP markers. SNP were selected based on assay quality and allele frequency across breeds. Following three rounds of optimization, six marker panels were developed containing 383 SNP. This included markers for horn - poll, muscularity, pigmentation and a small number of inherited diseases. To determine the minimum number of panels required to obtain high accuracy assignment, blood cards were collected from industry flocks that contain differing levels of genetic relatedness between candidate sires. Genotyping was performed using SEQUENOM, before a maximum likelihood based approach was developed and applied to examine parentage. Use of two panels (or 127 SNP) gave high rates of correct paternity and may be sufficient for many flocks. Use of three panels (or 191 SNP) provided higher confidence and is the recommendation for initial commercial application. The current reagent cost associated with genotyping 3 panels, using the US based GeneSeek company as a supplier, is approximately $10 per sample, opening the possibility to offer a DNA based parentage product for under $20.

Executive summary

The ability to correctly assign parentage is important to producers. The advantages of DNA based methods are well known, however industry uptake is conditional on cost effectiveness. The activities within this project sought to develop and evaluate multiplexed sets of SNP capable of delivering high accuracy parentage at low cost. DNA markers were prioritised for inclusion in the project, drawn mainly from SNP collections discovered by the International Sheep Genomics Consortium. Prioritisation considered assay performance across genotyping platforms as well as allele frequency within a broad spectrum of breeds. Selected SNP were entered into the design process for "SEQUENOM" testing, currently the most cost effective genotyping platform available. The design process was successfully completed to produce six multiplex panels. These were subsequently tested using 96 high quality DNA samples, before poor performing SNP were dropped and replacement markers introduced to ensure each multiplex contained a high number (> 60) of working SNP assays. The final result was a total of 383 SNP assays formatted into six panels. Markers for horn - poll, muscularity, pigmentation, a selection of inherited disease and a sex determination marker were included. The optimised molecular reagents required to genotype the 383 SNP were subsequently transferred to GeneSeek, a production genotyping facility.

A critical component of the pre-commercial R&D performed in this project involved pilot testing the SNP panels to evaluate their power to deliver parentage and to assess their behaviour on different sample types including blood cards, nasal swabs and ear clips. The sample type testing demonstrated that whilst ear clip samples generated more genotype calls, blood cards performed sufficiently well, with similar mismatch rates to ear clips, so that the better cost, storage and transport properties of blood cards meant that they would be more suitable for large scale industry collection. To address their parentage power, blood cards were collected from 290 Merinos (sires, dams and progeny) at Karbullah in late 2011. These were selected to represent an industrially relevant breeding scenario where parentage assignment was required. At the same time, the potential benefit to producers of a SNP based parentage tool was recognised within the SheepCRC and Sheep Genetics. This prompted a collaboration to perform parentage testing on these and 1,711 samples from an additional five flocks sourced from Sheep Genetics clients. Blood cards from all 2,001 animals were genotyped for each of the six SNP panels at GeneSeek, before the data was used to evaluate the minimum number of panels required to obtain high accuracy assignment.

Two analytical approaches were tested, exclusion and maximum likelihood. Exclusion provides information on SNP specific error rates and does not depend on flock specific information. Error rates varied between SNP, and a small number of loci performed poorly. If at any time a re-design of the panels is proposed then these would be removed and replaced with better performing SNP. However, they do not present a problem for the immediate implementation of the test, as the maximum likelihood approach accounts for SNP genotyping errors. In addition, flock specific assignment thresholds are defined through simulation, using flock specific observed allele frequencies. When applied to the validation data, use of two panels (or 127 SNP) gave high rates of correct paternity and may be sufficient for many flocks. Use of three panels (or 191 SNP) provided higher confidence and is the recommendation for initial commercial application. The results from commercial use of the three panel test in the Australian sheep industry should be evaluated after sufficient

commercial tests have been performed so as to clarify which flock structures need three panels as opposed to two panels. Cost-benefit analysis should also be performed on two panel versus three panel testing for different flock structures and the industry as a whole. Given the reagent cost required to genotype three panels is around $10 per sample, the results hold a lot of promise for delivering cost effective DNA based parentage in Australian flocks.

The SNP panels identified and tested in this project need to be made available to industry to deliver a benefit to breeders. This will require additional investment, primarily to establish the high volume blood-card handling pipelines that will be needed to accurately process tens of thousands of parentage tests annually. Existing expertise and infrastructure is likely to be adequate to process the associated data, however the ability to receive and dispatch 50,000 blood cards, along with the accurate reporting of results to the correct customer, is a large and complicated undertaking. A number of sample and annotation errors were detected within this project and rigorous procedures need to be used to both minimise such errors and to detect them when they occur. It will also be desirable for Sheep Genetics to develop a database system to store parentage genotypes to minimize regenotyping costs where sheep are part of multiple parentage data sets. The additional work required to make the SNP panels available to producers is to be undertaken within the SheepCRC from July 2012 (J. van der Werf, pers comm.).

Report History:

This combined report describes the outcomes from two aligned projects: B.BSC.0095 and B.BSC.0097. It extends the findings of project B.BSC.0095, published in May 2012, through the description of analysis performed by Dr Maddox concerning SNP error rates and an investigation of exclusion analysis for the assignment of parentage.

Table of Contents

Abstract 2

Executive summary 3

1. Background 6

2. Project objectives 6

3. Methodology 7

Materials 7

DNA quality and sample type testing by SEQUENOM (Australia) 8

Sampling of industry flocks 9

Genotype data from industry flocks 10

Analytical methods 11

4. Results 14

Development and content of six multiplex panels 14

Comparison of SEQUENOM HQ and LQ samples with SNP50 genotypes 16

Sample issues for industry flocks 19

SNP panel performance in Australian flocks 20

5. Discussion, conclusions and recommendations 32

6. Acknowledgements 34

7. Appendices 35

Appendix A 35

Appendix B 36

Appendix C 43

Appendix D 45

8. Bibliography 59

1 Background

The ability to accurately and cost effectively assign parentage would confer a number of benefits to sheep producers. Key advantages include eliminating the need to identify parents of lambs via the ewes they suckle from together with mating records, lowering labour costs and pedigree errors through mismothering. Similarly, the ability to assign paternity would remove the infrastructure required for single sire matings. The ability to perform syndicate matings, and then retrospective assignment of paternity using DNA methods, means increased weaning rates associated with multi-sire programs. Further, access to simple cheap pedigree would simplify adoption of MerinoSelect and allow increased rates of genetic progress, particularly for traits such as number of lambs weaned and worm egg count, both of which are significant for lamb and sheep meat production. Finally, knowledge of parentage allows the incorporation of trait data from relatives when estimating BLUP EBVs. The associated increase in accuracy may be as much as 20% (pers comm., van der Werf).

To date, the adoption of DNA based methods for parentage have been hindered by high assay costs and low transparency in terms of commercial service quality. Two commercial providers offer microsatellite based paternity tests that cost between $16 and $35 per sample. Microsatellites are well suited to paternity in that they are multi-allelic DNA loci that individually have high power to exclude animal pairs as direct relatives. Unfortunately the cost to collect genotypic data from microsatellites is high, and they are not amenable to cost effective automated systems of data collection. Given that we anticipate a commercial DNA based parentage product may be used on tens of thousands of samples per year, microsatellites were not selected as the marker type of choice.

Single nucleotide polymorphisms (SNP) are bi-allelic substitutions that are common throughout the sheep genome. Initial estimates from sequencing indicate the presence of a SNP every 200 bp across the sheep genome (KIJAS *et al*. 2009). In addition to high abundance, SNP are evolutionarily stable (low mutation rate) and importantly they are well suited to high throughput and therefore lower cost genotyping compared with microsatellites. Given cost has been the major barrier to wider uptake of DNA based parentage testing within industry, SNP markers were selected as the marker of choice in this project. Efforts to identify SNP suitable for parentage have been made within the public domain. The International Sheep Genomics Consortium (ISGC) successfully identified tens of thousands of SNP (KIJAS *et al*. 2009), before subsets were genotyped using multiple genotyping platforms across a variety of sheep breeds (KIJAS *et al*. 2012). This information served as the starting point for the activities within this project, the aim being to design and test SNP panels and measure their ability to deliver parentage in Australian sheep populations..

2 Project objectives

The objectives of the project involved the pre-commercial R&D required to deliver a technically robust and widely applicable DNA based tool for sheep parentage. The sub-objectives included:

1. *In silico* selection and priority ranking of approximately 360 well spaced SNPs suitable for parentage determination in sheep.

2. Identify SNP of large effect for priority consideration in SNP panel design. Examples might include SNP diagnostic of poll – horn, pigmentation status or monogenic disease.

3. Collaboration with SEQUENOM during the development of robust SNP multiplexes to formulate the minimum number of plexes with optimal power to resolve parentage/pedigree relationships in sheep with an accuracy >99%.

4. Assist with sample selection and experimental design suitable for pilot testing the developed SNP multiplexes.

5. Development of analytical strategies that utilise SNP data to generate exclusion probabilities and parentage assignment.

3 Methodology

Materials

The SNP markers used in the project were identified by the International Sheep Genomics Consortium (ISGC). Three 'types' of SNP were used, and each is described below:

1) ISGC Parentage SNP

The best SNP for use in parentage testing are those that i) have a high minor allele frequency (MAF) within multiple key breeds, ii) can be genotyped with high accuracy and iii) perform reliably independent of the assay platform used. Activities within ISGC identified a panel of SNP with each of these attributes. In brief, high MAF SNP were identified that displayed reproducible and accurate genotyping results using both the 1.5K Illumina Golden Gate (KIJAS *et al*. 2009) and Infinium SNP50 assays (KIJAS *et al*. 2012). In addition, subsets of the SNP were evaluated using the 'Fluidigm' assay platform, by re-sequencing at the USDA and using the SEQUENOM Iplex platinum design in use at AgResearch. At each step unsuitable SNP were excluded, leaving a core panel of 89 markers. The identifier, chromosomal position and MAF for each SNP was published at the Plant and Animal Genome conference in January 2011 and the poster appears in Appendix A. The marker panel is being promoted to the International Society of Animal Genetics (ISAG) as the standard for testing in sheep. All 89 of the SNP were used in the design of the assays used in this analysis (see below).

2) Performance and Y chromosome SNP

SNP are available that directly cause variation in phenotype, or are linked to the mutations that cause variation in phenotype. A total of 10 SNP were successfully designed into the multiplexes used in this analysis (refer to the Results section). Two SNP were included linked to the *Poll* locus. A further four SNP are associated with hoof pigmentation, one directly underpins muscling in Texel, one is located on the Y chromosome and performs confirmation of sex while the remaining two markers are diagnostic of monogenetic diseases. A third SNP at the *Poll* locus was entered into the design phase but failed repeated testing and was excluded. It is worthwhile noting the monogenic diseases may not be present within Australian flocks.

3) Filler SNP

A third type of SNP was included to ensure sufficient markers were available to achieve high accuracy parentage assignments. These were referred to as ‘filler’ SNP as they were used to fill in around the first two types of SNP during assay design. A set of 4,538 filler SNP were identified as follows. First, 49,034 SNP passing quality control in the sheep HapMap experiment (Kijas *et al.* 2012) were examined to identify 18,565 with MAF > 0.35 in Australia Merino and Poll Merino. This set was evaluated to identify a subset of 4,817 SNP that displayed MAF > 0.25 in Australian flocks of White Faced Suffolk, Poll Dorset, Border Leicester and Texel. A total of 183 SNP were removed that showed segregation ratio irregularities or other typing problems, before another 96 markers were pruned that are inefficient to genotype (as Infinium type I).

Using the three types of markers described above, sets of SNP were designed using SEQUENOM software. These SNP sets are referred to as “multiplexes” and are complimentary combinations of SNP that can be assayed in a single reaction. To reduce the cost of genotyping per SNP, the design phase sought to maximise the number of SNP within each multiplex. After initial design, a total of six multiplexes were tested using a set of 96 genomic DNA samples. The animals were all drawn from Australian industry sires (Merino *n* = 10; Poll Merino *n* = 10; White Suffolk *n* = 20; Border Leicester *n* = 18; Poll Dorset *n* = 20; Texel *n* = 12; Romney *n* = 6). The testing identified underperforming SNP assays that were removed and replaced by other assays. Three rounds of multiplex optimisation were performed.

DNA quality and sample type testing by SEQUENOM (Australia)

A panel of 96 high quality (HQ) DNA samples derived from 92 rams and 4 ewes was made available to the project to test the multiplexes. More details about these samples are given in BSC.0095 Milestone 5. Seventy-five of the 96 samples had previously been genotyped with the SNP50 array under the auspices of the International Sheep Genomics Consortium (ISGC).

MLA arranged for a set of "low quality" (LQ) samples to evaluate genotyping performance when using SEQUENOM. A detailed description of the samples is provided in Table 1. These samples were derived from 53 female sheep that had been SNP50 genotyped as part of the sheepGENOMICS programme and comprised two sets of samples: (1) DNA purified from blood cards, ear clips and nasal swabs (BEN samples) from each of 30 sheep (DNA details in BSC.0095 Milestone 5); and (2) 60 samples of DNA of various quality (DNAQUAL) and concentrations derived from 41 sheep including 18 of the sheep that provided the BEN samples. The BEN samples were all genotyped at least twice whilst the DNAQUAL samples were genotyped at least 3 times. Some of the replicates included incomplete multiplex sets (i.e. missed one or more of Multiplex1, Multiplex 3, Multiplex 5 and Multiplex 6). The rates of "No call" and errors were determined for the sample set. The genotype results for replicate samples were compared and these were also compared to the SNP50 genotypes.

There was no overlap between the HQ and LQ DNA samples so no comparisons between these could be made.

Table 1. Samples genotyped at SEQUENOM (Australia)

Flock	Number Samples	Number of Sheep	
High Quality (HQ)	96	94*	* from 95 sheep; 75 samples SNP50 genotyped
LQ Blood Card#	30	30	all samples SNP50 genotyped
LQ Ear clip#	30	30	all samples SNP50 genotyped
LQ Nasal Swab#	30	30	all samples SNP50 genotyped
LQ DNA	60	41	all samples SNP50 genotyped

The same 30 animals were used for the blood cards, ear clips and nasal swabs

Sampling of industry flocks

In order to pilot test the performance of the SNP panels, samples were collected from Australian industry flocks (Table 2). Contact with producers commenced in September 2011 with Marc Murphy from the Karbullah Merino stud. The objective was to sample animals from a relevant industry flock which may include use of sire syndicates that include half-sibs and/or examples of father – son pairs. A total of 290 samples were collected which included 32 sires, 111 dams, 87 lambs and a further 44 animals within unconfirmed class (ewe or progeny) (Table 2). A total of 16 of the 32 sires were used in a syndicate mating to produce the progeny. The other 16 Karbullah sires were sampled, even though they did not participate in generating any of the lambs. Sheep Genetics supplied cards were used for collection of blood, before the cards were returned to Sheep Genetics. These were subsequently shipped to GeneSeek for SEQUENOM genotyping.

It is important to note that during the last half of 2011, the potential value of SNP based parentage testing for the Australian sheep industry was recognised within the SheepCRC. This prompted the collection of additional flocks to significantly bolster the sample obtained from Karbullah in work coordinated by Sam Gill (Sheep Genetics). Actual sample collection from these flocks and sample plate design was performed external to this project by the SheepCRC. A summary of each flock used in the analysis is provided in Table 2.

Table 2. Summaries of sheep with SNP data

FLOCK	SIRES		DAMS		LAMBS		UNKNOWN	
	Total	Genotyped	Total	Genotyped	Total	Genotyped	Total	Genotyped
Flock 1	11	11	302	302	415	415	0	0
Flock 2	0	0	111	111	122	118	0	0
Flock 3	32	32	111	111	103	103	44	44
Flock 4	7	7 (5)*	21	21 (29)	180	180	3	3
Flock 5	(7)	(7)			(160)	(156)	167	167
Flock 6[#]	24	24 (8)[ª]	238	0	520	331	0	0

* Number of actual sires, dams and progeny identified by genotyping are indicated in brackets

[#] Samples were collected from 782 samples for Flock 6 but only 355 samples were genotyped for this project.

[ª] A sire from flock 4 was found to be the sire of 26 flock 6 progeny and in a parent-child relationship with a flock 6 sire. No progeny were found for the other putative 16 flock 6 sires, and no sire was found for 167 progeny. One flock 6 sire was found to be the sire of another flock 6 sire.

In addition to these animals there were genotyped samples that could not be linked to an animal ID.

Marc Murphy's Karbullah population is Flock 3.

Genotype data from industry flocks

Blood cards from all animals were sent to GeneSeek (Nebraska, USA) for genotyping using the SEQUENOM platform. SNP data was generated from 2,001 sheep from 6 flocks as part of the SheepCRC. For many of the sheep there was no information returned to the project to indicate which individuals should be treated as a sire, a dam or as a lamb. Without this information it is very difficult to draw conclusions regarding pedigree accuracy from the SNP. Consequently, only where the sheep was known to be a prospective sire, dam or lamb was the SNP data used for maximum likelihood analysis in the study. It is important to note this information would be available to a commercial testing laboratory. Details of the number of sheep from each of the 4 flocks with information on animal class is summarised in Table 2. When analysing these data we had in advance no knowledge of the history of the flocks, for example, no knowledge of the closeness of the relationships between sires, or of the long term level of inbreeding. A summary of this information was subsequently provided by Sheep Genetics and is listed in Table 3.

Table 3. Flock characteristics

Flock	Number Genotyped	Number Passed	Breed	Information	Sire Sampled	Dam Sampled	Horn/Poll
1	730	730	Merino	Syndicate joining, use Pedigree Matchmaker	Yes	Yes	Yes
2	235	235	Dohne	Closely related lines (15% inbreeding coefficient)	No	Yes	No data
3	290	290	Merino	Closely related syndicate mated lines	Yes	Yes	Yes
4	212*	212*	Merino	Embryo transfer progeny, closely related sires and dams	Yes	Yes	Yes
5	167	164	Coopworth	Cards from sheepGENOMICS	Yes but not specified	No	No data
6	355	347	Merino	Closely related syndicate lines, poll/horn, dam pedigree	Yes	No	Yes

* 12 additional blood cards all of which were successfully genotyped were subsequently identified as having been derived from this flock.

Analytical Methods

Development of a C^{++} Parentage Program Based on Exclusion Analysis

A command-line C^{++} program termed mla_parent was developed based on exclusion analysis. This program compares genotypes (alleles scored as A or B) for SNPs that are both homozygous for all pairs of animals. The number of mismatches is counted and then used to identify potential parent-child (pc) relationships based on the number of mismatches being below a specified threshold. The program checks samples with more than one parent-child relationship to identify whether any trio (two parents, one child) relationships exist. For a trio relationship to exist the child genotype for each SNP needs to be consistent with it getting one allele from each parent, and the number of mismatches needs to be below a specified threshold. Input to the program consists of a 4 column tab-delimited ascii file containing SNP identifiers, animal identifiers and genotypes (one allele per column scored as A, B or -) together with command line specification of output file prefix, maximum number of double homozygote mismatches for parent-child relationships, and sample and SNP error thresholds (minimum proportion of successful SNP assays needed for a sample to pass, minimum proportion of successful sample assays for a SNP to pass) if these differ from the defaults of 0.95, 0.95. The outputs to the program include all replicate samples found, and all parent-child and trio relationships found within the data set together with the numbers of mismatch errors for SNPs for the replicates and the relationships found within the data set. It should be noted that whilst both parents and children can be identified for trios without other information only the relationship can be identified for a parent-child relationship, i.e. it is not possible to identify which is the parent and which is the child. However, if one animal in the relationship has large numbers of parent-child relationships then it is likely that animal is a parent for most, if not all, of the relationships.

Replicate and relationship data were then compared to the pedigree information provided by MLA using customized awk scripts and a mysql database to generate various relationship and other reports that were provided to MLA.

SNP specific error rates from samples genotyped by SEQUENOM (Australia) and GeneSeek

The HQ and LQ samples (Table 1) were compared with SNP50 genotypes provided by the International Sheep genomics Consortium (ISGC) and SheepCRC using the mla_parent program. The genotypes for replicates within the LQ data set were also compared.

An agreed experimental design was negotiated between MLA and GeneSeek for the GeneSeek genotyping. This included the use of a number of replicated samples to enable assessment of quality control (reproducibility of calls including an assessment of the reliability of the oY1 SNP for sex assignment and an estimate of relative plate performance) and plate identification (based on positioning of replicates on plates). Each plate was meant to have both a male control DNA sample and a female control DNA sample. As stated above, the sample collection was handled external to this project and for some reason the replication of samples did not occur so that the reproducibility of calls could not be properly investigated. This lack of duplicate samples meant that a convoluted procedure had to be used in an attempt to identify error rates for SNPs for the samples genotyped by GeneSeek. Briefly this entailed an iterative approach where putative relationships were identified by exclusion analysis followed by the removal of a set of SNPs with too high error rates (overrepresented in multiple parent-child relationship mismatches) from the data set. The data was reanalysed with the reduced set of SNPs and additional relationships identified. SNPs with too high error rates were again removed from the data set and the data set was reanalysed for possible relationships. Error rates were then determined by determining the number of mismatches for the final set of relationships with all of the multiplex SNPs.

Maximum likelihood pedigree assignment

A maximum likelihood method was used for pedigree assignment (KALINOWSKI *et al.* 2007; KALINOWSKI *et al.* 2010; MARSHALL *et al.* 1998). Given the SNP data for a sire and a lamb, the likelihood that the sire is the parent of the lamb is evaluated, along with the likelihood that the sire is not the parent of the lamb. The estimations make use of an assumed allele frequency for each SNP in the population, and an assumed genotyping error rate. Whilst the exclusion analysis shows that variation exists in the error rate between SNP (Appendix D), as in Marshall *et al.* 1998, a common SNP genotyping error rate was assumed. The method could be improved by incorporating SNP specific error rates, estimated as described above. For consistency with (MARSHALL *et al.* 1998) the log of the ratio (likelihood that the sire is the parent / likelihood that the sire is not the parent) is referred to here as the LOD score (from log odds). LOD scores were also estimated for lamb-dam pairs, and for lamb-sire-dam trios.

To derive an appropriate threshold for parentage assignment it is necessary to know the distribution of the LOD score when the parent being tested is correct, and when the parent being tested is incorrect. The distribution of the LOD score is dependent on the number of SNP markers being used in the likelihood calculations, the allele frequencies of the markers, and when the parent is incorrect, on the relationship between the parent being tested and the true parent. There is no mathematical equation that, given these inputs, produces the threshold, especially as the relationships between the parents being evaluated may be unknown. Accordingly, we used simulation to derive an appropriate LOD threshold for each test type (lamb-sire pair, lamb-dam pair, or lamb-sire-dam trio) for each flock.

1. 1,000 simulated lamb progeny were produced, each with a randomly chosen sire and a randomly chosen dam from the flock. Where one parent type (sire or dam) was not represented in the flock a parent was simulated using the allele frequencies estimated for the flock.
2. For each simulated lamb, LOD scores were estimated for each sire and for each dam. For the most likely 5 sires and the most likely 5 dams the LOD score was estimated for each of the 25 possible parent pairs.
3. For sire parentage, the most likely sire was identified and the LOD score stored (mLOD1), along with the difference between mLOD1 and the LOD score for the second most likely sire. For consistency with (MARSHALL *et al.* 1998) we refer to this difference as Δ1, with the '1' indicating that the difference relates to LOD1. Similarly, the LOD score for the second most likely sire (mLOD2) and associated Δ2 were stored. Whether or not the most likely parent was the parent used in the simulation was also recorded. The same method was used for dam parentage and for sire-dam parentage.
4. For Δ, a threshold (TΔ) was declared as TΔ = 3, and was used in all flocks. This can be interpreted as: parentage was only assigned if the most likely parent was at least 3 times more likely than the second most likely parent. The value of 3 was chosen based on earlier simulation work, and as "three times more likely" was judged to be a sufficiently rigorous threshold. Particularly for trios, a more stringent threshold could be applied with little effect on the assignment rates.
5. Given the threshold TΔ = 3, a threshold for mLOD, (TmLOD) was found that balanced the number of false positives (i.e. mLOD2 > TmLOD) and false negatives (i.e. mLOD1 < TmLOD), subject to the constraint that the percentage of false positives was less than 10%. The objective function also included a term favouring solutions midway between the minimum mLOD1 and the maximum mLOD2.

For the real lambs, the values of mLOD and Δ were compared to the thresholds TmLOD and TΔ, and parentage assigned if mLOD ≥ TmLOD and Δ ≥ TΔ, or not assigned if mLOD < TmLOD or Δ < TΔ. In all simulations and analyses we assumed a genotyping error rate of 1%. Note exclusion analysis error tests demonstrated that 1% was a significant underestimate for some SNPs (Appendix D).

Exclusion analysis pedigree assignment

Exclusion analysis was carried out using the mla_parent program with the genotype data generated by this project together with SNP50 genotype data from the sheepCRC and ISGC. Genotype data for the parentage SNPs was first extracted from the SNP50 data and then combined with the Sequenom data so as to identify matching animal, parent-child and trio relationships. The relationships found were compared with those provided and reports detailing found relationships were provided to MLA. The relationships identified using all six panels were compared with those identified from just using panels 1, 2 and 3.

4 Results

Development and content of six multiplex SNP panels

Three types of SNP were used to design genotyping multiplexes: ISGC parentage SNP, performance SNP and filler SNP. Selection and priority ranking of SNP is described in the methodology section. In order to meet the requirements of the SEQUENOM design process, an excess of SNP were made available to ensure highly multiplexed sets could be developed. After initial design (by SEQUENOM, B.BSC.0098), a total of six multiplexes (or panels) were tested using 96 genomic DNA samples to identify underperforming assays. These were removed and compatible replacement SNP tested. Three rounds of optimisation were performed in collaboration with SEQUENOM Australia (B.BSC.0098), with the result that 383 working assays were formatted into six panels (referred to as W1 – W6, Table 4). The number of markers in each multiplex ranged from 63 (in W2, W3 and W4) to a maximum of 66 SNP (in W5). During design, it was anticipated that two or perhaps three multiplexes would be required to achieve high accuracy results. Six were developed to identify the minimum number required to delivery high accuracy parentage results. The design strategy involved prioritising inclusion of the ISGC parentage SNP and performance SNP into W1 and W2. Filler SNP were subsequently added to maximise the number of markers that could be analysed in parallel.

Table 4. SNP types within multiplexes

			SNP Type	
Panel	**Total_SNP**	**ISGC SNP**	**Performance**	**Fill SNP**
W1	64	38	6	20
W2	63	28	3	32
W3	63	18	1	44
W4	63	2	0	61
W5	66	1	0	66
W6	64	0	0	63
totals	383	87	10	286

SNP identifiers are given in Appendix D

An important objective of the project involved inclusion of performance SNP. This aimed to enrich the value proposition for producers, through reporting on monogenic traits in addition to parentage at little or no additional cost. A total of 10 performance and sex assignment SNP were identified and successfully included into the final multiplex panels. The identification of many of the SNP has been performed outside this project, and the published reference is given in Table 5 along with SNP identifiers. Analysis was performed within the project to identify SNP of large effect for poll – horn, parasite resistance or pigmentation traits that were identified as part of the FMFS study. A detailed explanation of the analysis is contained within milestone report 3 (BSC.0095). Four SNP associated with hoof pigmentation were identified through analysis and were included into the SNP panels. Accurate identification of poll / horn status in animals is likely to be of most interest.

The panel also contained a SNP from the Y chromosome (oY1.1). This SNP should only return a genotype in male sheep, and the genotype should always be homozygous given that there is only a single Y chromosome. Heterozygous genotypes for the Y chromosome SNP result from DNA problems (sample contamination). The proportion of heterozygosity for male samples will underestimate the male sample contamination rate as heterozygosity will only occur when combined male samples come from sires with different oY1.1 genotypes. The oY1.1 SNP can also be used to check the expected sex of an animal and information can be obtained on sex error rates (shows as male but meant to be female, and *vice versa*).

Table 5. Details of performance and sex assignment SNP used in multiplexes

SNP Identifier	Associated Trait	Reference
MSTN	Muscling QTL Myostatin	CLOP *et al*. 2006
OAR10_29389966	Horn Poll	KIJAS *et al*. 2012; JOHNSTON *et al*. 2011
OAR10_29448537	Horn Poll	KIJAS *et al*. 2012; JOHNSTON *et al*. 2011
PITX3	Microthpthalmia	BECKER *et al*. 2010
oY1.1	Sex checking	MEADOWS *et al*. 2004
OAR19_33531772	Hoof Pigmentation	DOMINIK *et al*. (unpublished)
OAR19_33968342	Hoof Pigmentation	DOMINIK *et al*. (unpublished)
s52103	Hoof Pigmentation	DOMINIK *et al*. (unpublished)
OAR19_34354181	Hoof Pigmentation	DOMINIK *et al*. (unpublished)
FM872310_2746delCA	Epidermolysis bullosa	MOMKE *et al*. 2011

Analysis of SEQUENOM HQ and LQ samples and comparison with SNP50 genotypes

SEQUENOM provided both provisional, interim and finalised genotype data for the HQ samples with there being a number of differences between the data sets. In the finalised version SEQUENOM failed nine of the multiplex SNPs tested against the HQ data set with the reasons given being "No amplification" (W1 DU488903_267, W2 DU364754_308, W4 OAR2_137718678), "Poor amplification" (W1 PITX3_MMFwd-12(v2)); "Gene Duplication" (W2 OAR18_10748526, W3 DU213735_493-1(v2)), "Skewed Het (W3 s52103(v2)) and "Lower call rate" (W5 OAR5_31695916, W5 OAR17_6395356).

Eighty-eight of the 96 HQ samples were successfully genotyped for at least 315 (82%) SNPs whilst the remaining eight HQ samples had genotypes for fewer than 80% of the SNPs with the worst sample only having genotypes for 247 (65%) SNPs (Appendix D). The SEQUENOM SNP scoring rate was much lower than that found for the both the Illumina BeadArray 1.5K and SNP50 data sets. Comparison of the SNP50 and SEQUENOM data where genotypes were scored in both data sets revealed that 298 SNPs had no mismatches whilst there were 544 mismatches from 85 loci. The maximum number of mismatches per SNP was 29 for s43800. Whilst the sample set was meant to include only one duplicate (BL

0244112002020008A, BL 0244112002020008) both the SEQUENOM and Illumina genotyping found the expected duplicate and an unexpected duplicate (BL 020041200000A444, PD 1600322001010444). Four parent-child relationships were found within the data set (WS 2300992000000050, 2300012002020070/5 – inconsistent sample identification; BL 210902005050159, 0240752000000112A; PD 1640732003030387(0), 1640732004040464(0); NZRom_M_6, NZRom_M_4). This number of duplicates and parent-child relationships was insufficient to determine error rates from. All genotypes for the oY1.1 SNP were consistent with the sex of the animal (92 males, 4 females).

Comparisons of the LQ parentage genotypes with sheepCRC SNP50 data identified no mismatches for 127 of the SNPs, while 34 of the remaining SNPs had more than 20 mismatches. The most common mismatches were for samples genotyped as heterozygous by SNP50 and homozygous by SEQUENOM genotyping (Table 6). The highest number of mismatches was 34 for s43800. Genotyping variation for duplicate samples included two animals having both oY1.1 positive and oY1.1 negative genotypes when all LQ samples should have genotyped as oY1.1 negative. The average number of discrepant genotypes between the Sequenom and SNP50 genotyping was 2%.

One sample discrepancy was identified within the LQ DNAQUAL data set based purely on comparisons within the SEQUENOM data set, with sample o12678 being shown to be a duplicate of o12633 when it was not meant to be. Comparisons with SNP50 data revealed a further sample error with the wrong RFID being supplied for sample o12641 (951 000000836438 not 951 000006556408).

Table 6. Types of genotypic discrepancies comparing SNP50 and SEQUENOM data

SNP50 genotype	SEQUENOM genotype	Number	% of discrepancies
homozygous	heterozygous	461	14.1
heterozygous	homozygous	2,599	84
opposing homozygotes		35	1.1

Table 7. Performance of multiplex panels for LQ – No calls and SNP50 mismatches

Multiplex Panel	Number of No calls	Number of Mismatches
1	206	147
2	520	149
3	238	219
4	437	261
5	820	387
6	406	233

Multiplex panel 1 performed the best having both the smallest number of "No calls" and the smallest number of SNP50 mismatches (Table 7). The three best panels from a Sequenom-SNP50 mismatch perspective were panels 1, 2 and 3.

Table 8. Results for SEQUENOM (Australia) genotyping of HQ and LQ samples*****

Flock	Average number no calls/sample	Average number mismatches/sample	Comment
High Quality (HQ)			One unexpected sample duplicate also found
LQ Blood Card	15.5	7.4	
LQ Ear clip	6.7	7.3	
LQ Nasal Swab	21.6	8.7	
LQ DNA			

* comparisons only included samples that had been SNP50 genotyped

Table 8 indicates the ear clip samples gave the best results having both the lowest rates of no calls and the lowest rates of mismatches. The blood card samples were almost as good in terms of rate of mismatches but had more than twice as many no calls. Given that blood cards are cheaper, easier to store, transport and extract DNA from the project opted to continue using blood cards for subsequent multiplex testing by GeneSeek.

A number of problems were found with the LQ sample set that had been provided to the project from an experimental design perspective. These included there being no samples from male sheep, an imbalance in the design for the LQ samples such that not all 30 animals used for BEN samples had low quality purified DNA samples, and there was no overlap between the animals used for the HQ and LQ experiments. In terms of the HQ data set it would also have been preferred if all animals had also been SNP50 genotyped.

Some of the discrepancies found were consistent differences between SNP50 BeadArray and SEQUENOM calling which means that one would have to allow for a higher error rate if one was comparing parentage results from sets of genotyping data that have be combined from different genotyping technologies (and possibly service providers). Unfortunately no overlap between SEQUENOM Australia and GeneSeek samples so not able to compare service provider SNP differences.

Sample issues for industry flocks

Of the 2,001 blood cards sent to GeneSeek, 12 lacked sample identification. Parentage exclusion analysis revealed that all of these samples appeared to have come from flock 4. The overall genotyping success rate for samples was over 99%, with only 11 of the 2,001 blood cards not being successfully genotyped (Table 2). The maximum number of SNPs with genotypes for the failed samples was 283 i.e. < 75%. All other samples had genotypes for more than 95% of successfully genotyped SNPs.

The experiment was complicated by the fact that the initial data provided used three identifying systems (DNA ID, electronic ID and 16 Digit ID) however neither the electronic or 16 digit identifier system included all animals. Consequently convoluted procedures were needed to identify expected parent-child and trio relationships to compare with found parent-child and trio relationships. The DNA ID was the most consistent of the identifiers as many sheep in the initial data file lacked either an electronic ID or a 16 digit ID. Whilst, subsequent data provision expanded the number of sheep with electronic and 16 digit Ids, the data analysis would have been significantly easier if this information had been provided at the same time as the genotype data and the initial data analysis.

Some flock samples did not have associated animal class (progeny, sire, dam) information or putative parents, and some animals belonged to multiple classes (i.e. were both a progeny and a parent). This could be checked with exclusion analysis but not with the maximum likelihood method used. No animal class information was received for flock 5.

The animal identifier data revealed that six pairs of duplicate samples were in the 2001 samples sent to GeneSeek. However, comparison of genotype data only revealed four duplicate pairs. Three of these were expected duplicates and one was unexpected (flock 2). The most likely reason for not all of the expected duplicate samples being found was that three of the purported pairs were likely to be artefacts due to sample labelling errors. It is unknown as to what the true level of sample handling errors was. In order to estimate SNP error rates a number of animals would need to be resampled so that the two sets of genotypes could be compared. The level of duplicates in the experiment was insufficient for quality control purposes.

There were a small number of samples with higher heterozygosity than expected. Higher heterozygosity could be an indicator of sample contamination.

Table 9. Details of duplicate samples

Flock	Number of Duplicates Sent	Number of Duplicates Found
1	2	1
2	2	3
4	1	0
5	1	0

SNP panel performance in Australian flocks

1. Maximum likelihood analysis

A critical component of pre-commercial R&D involved pilot testing the SNP panels to evaluate their power to deliver parentage. Of particular interest was investigating the minimum number of panels that could be used to achieve high accuracy results. Based on a preliminary *in silico* study (Appendix C), we knew that one panel would almost certainly be too few, and that 3 panels would likely be sufficient. Panels 1 and 2 contain the performance SNP so these were always included. Panels 3 and 6 contained the largest numbers of polymorphic SNP, and were therefore likely to add the most power if used as a third panel. Consequently we tested panel combinations (1+2), (1+2+3), and (1+2+6). The results of the maximum likelihood based method used for Flocks 1 and 4 are presented in Figures 1a, b, c and 2a, b and c. Results for Flocks 2 and 3 are presented in Appendix B, as similar trends to Flock 1 appear for these flocks. Tabular results for all flocks are presented in Tables 10 - 12, which include a summary of TmLOD results for all flocks, a summary of rates of false negatives and positives for all flocks, and a summary of assignment rates for all flocks respectively.

To define flock specific thresholds for parentage assignment, simulation was used to define LOD values based on the observed allele frequencies. The simulated results are shown in panels on the left hand side, for lamb – sire pairs (top), lamb – dam pairs (middle) and lamb-sire-dam trios (bottom panel). The threshold defined by simulation (the vertical red line), estimated from the simulated data, is the line that best separates the correct assignments (dark blue points) from the assignments made if the true parent was not present in the data being analysed. Likely false positive rates can be estimated by counting the proportion of light blue points to the right of the line and above the Δ threshold of 3, and likely false negative rates can be estimated by counting the proportion of dark blue points to the left of the line or below the Δ threshold of 3. The thresholds are then used with the real progeny (right hand side plots, one point plotted per lamb), with lambs to the right of the vertical line and above the horizontal line declared to have parentage assigned, and lambs either to the left of the vertical line or below the horizontal line failing to achieve the thresholds required to assign parentage.

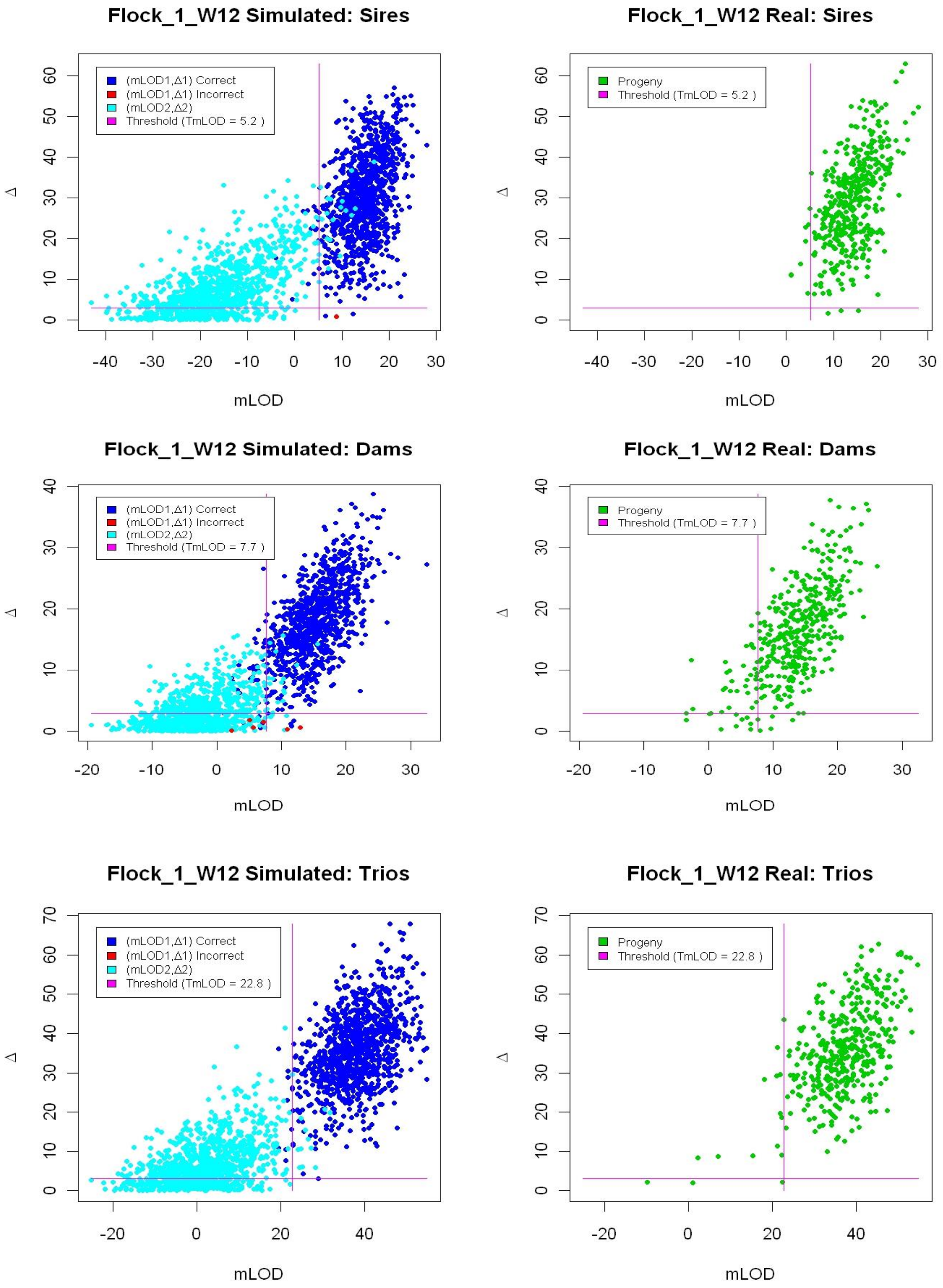


Figure 1a. Plot of Δ against mLOD for Flock 1, Panels W1 and W2

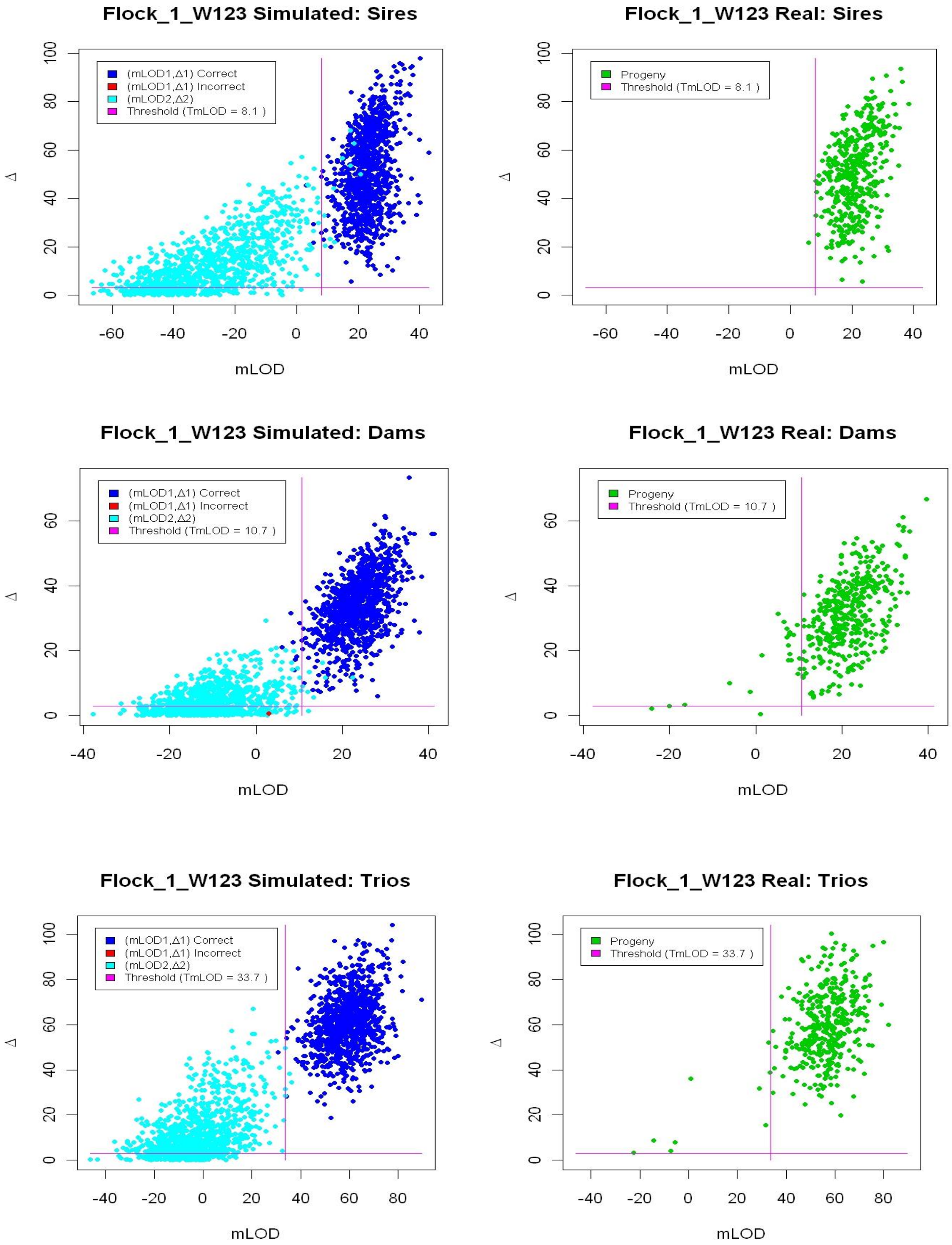


Figure 1b. Plot of Δ against mLOD for Flock 1, Panels W1, W2 and W3

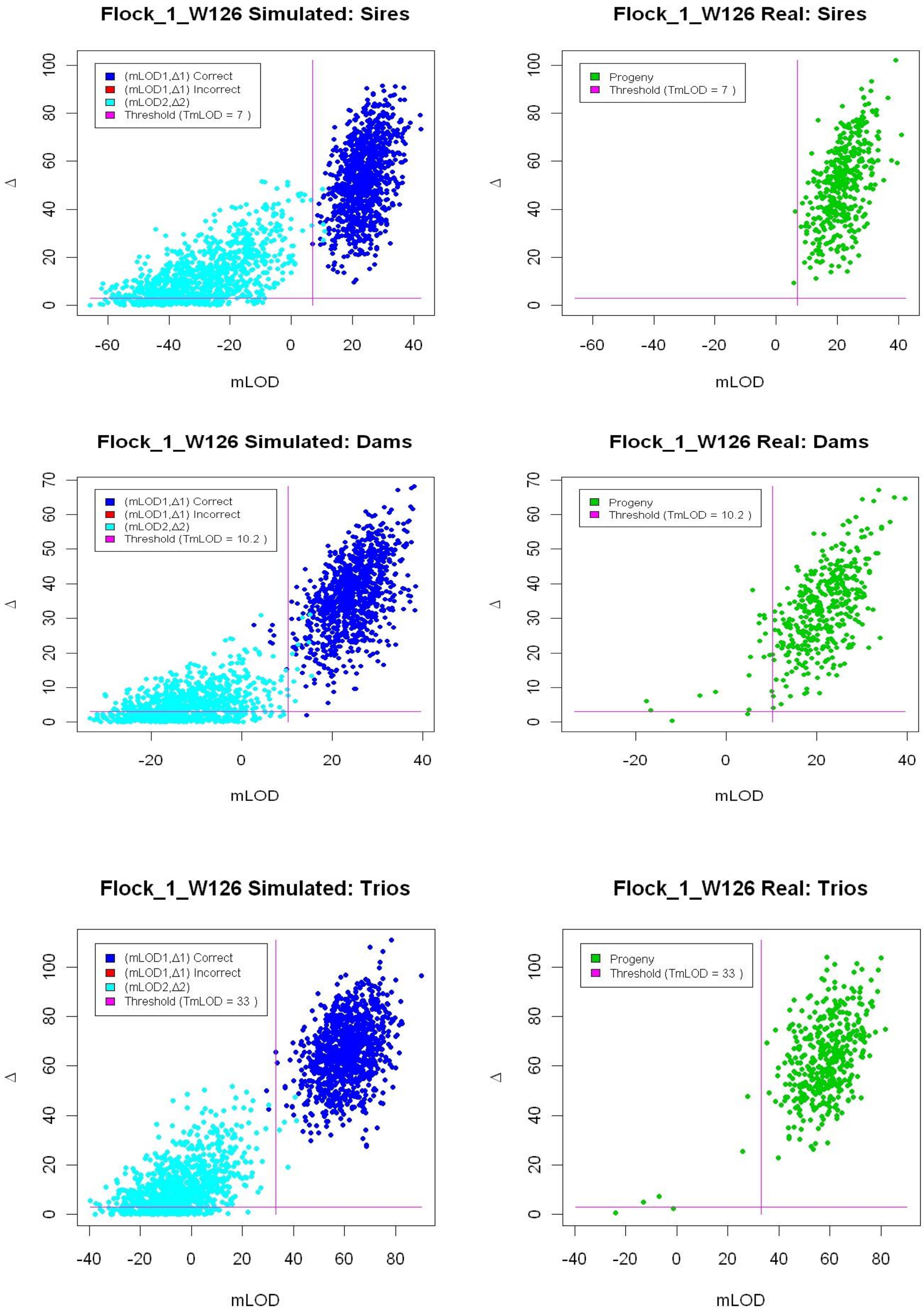


Figure 1c. Plot of Δ against mLOD for Flock 1, Panels W1, W2 and W6

In Figure 2a, results from the analysis of dams as parents, using panels W1 and W2 are plotted. In the simulated data (left), the most likely dam was not the true dam in only a few cases (red spots), and for these the value of Δ is low, below the threshold of 3 in most cases. The optimisation function to estimate the TmLOD threshold appears to produce a meaningful result. The cluster of correctly assigned simulated dams (dark blue dots) has a similar distribution to the cluster of real dams that exceed the threshold (right panel, green dots). For a significant proportion of lambs none of the prospective dams are likely to be the true mother in this flock.

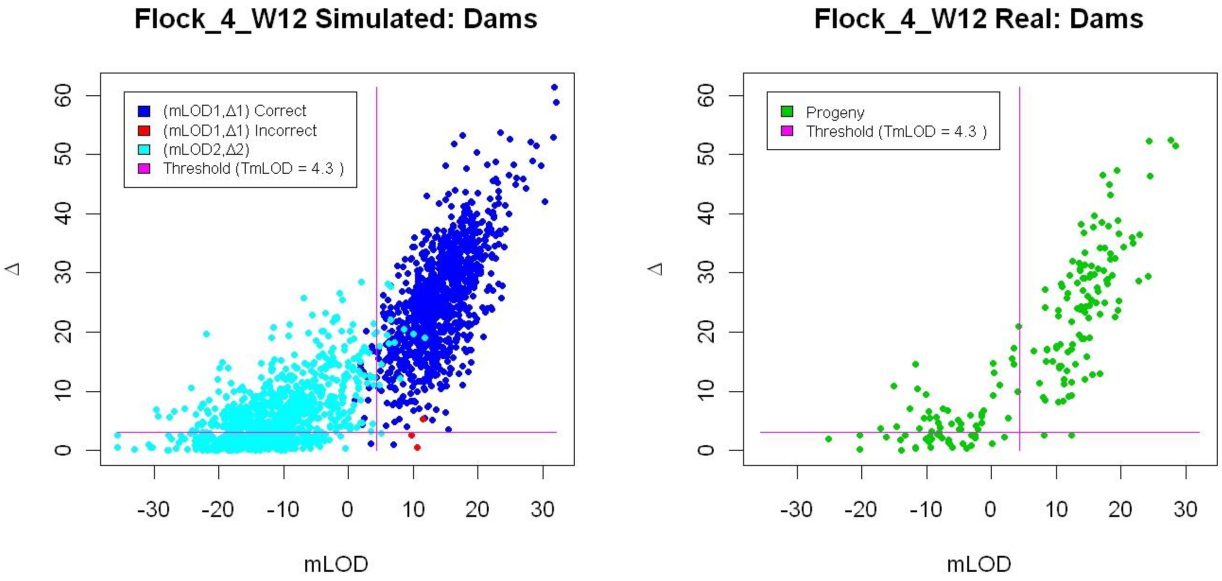


Figure 2a. Plot of Δ against mLOD for Flock 4 dams, panels W1 and W2. The threshold for Δ is 3.0

In Figure 2b the results for a trio analysis of the Flock 4 lambs, using panels W1 and W2 are presented. Compared to the dam analysis it is notable that the TmLOD threshold is much higher (18.4 instead of 4.3). In the simulated data there are fewer false positives and false negatives, and the true parents are always identified (i.e. no red dots). For the real data (right) there is a very clear cluster of trios that exceed the threshold, the rest are clearly below the threshold. There are fewer trios assigned than dams (Figure 1), indicating that for some lambs a dam is present in the data but not the sire.

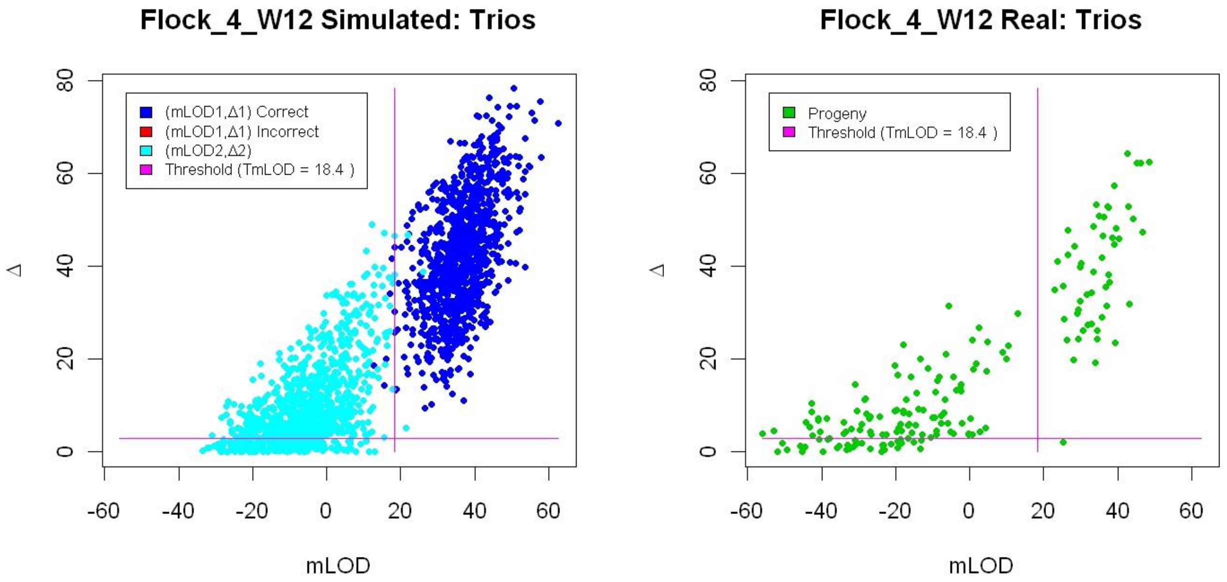


Figure 2b. Plot of Δ against mLOD for Flock 4 trios, panels W1 and W2. The threshold for Δ is 3.0

Figure 2c presents the same animals as Figure 2b, but with panel W3 added. In the simulated data (left) there is even more discrimination between mLOD1 and mLOD2. Adding markers to the test increases the appropriate threshold TmLOD to 32.8. For the real trios, the cluster for which a parent is assigned is even more distinct from those that fail to exceed the threshold. Plots for all of the flocks appear in Appendix B. Similar patterns occur, and in all cases the estimated threshold TmLOD appears to be a reasonable choice.

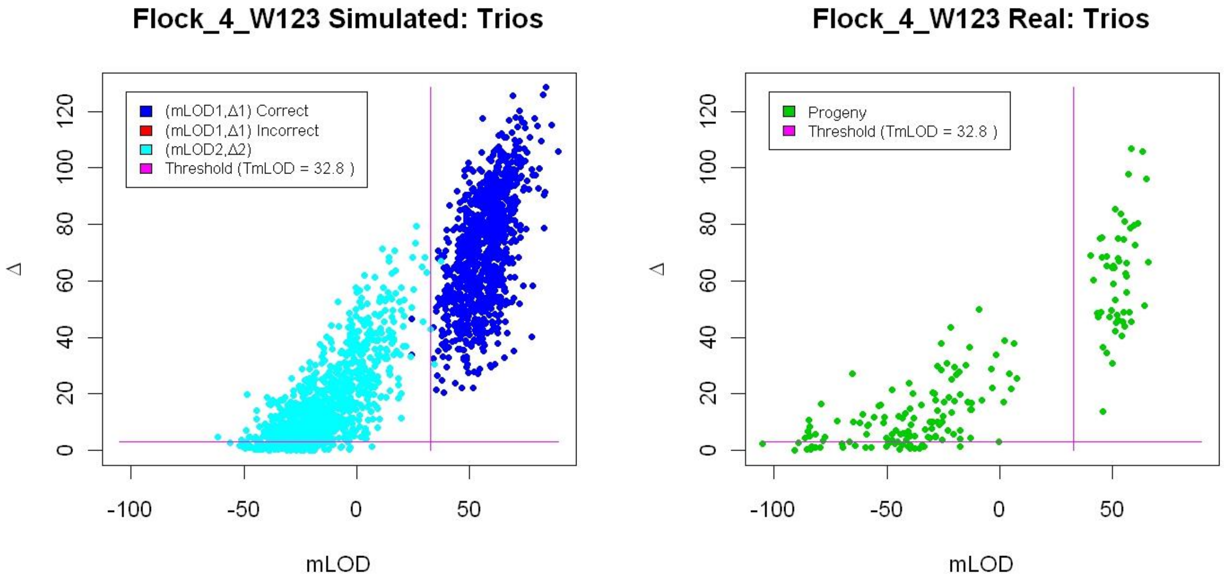


Figure 2c. Plot of Δ against mLOD for Flock 4 trios, panels W1, W2 and W3. The threshold for Δ is 3.0

Comparing the results for all flocks using panels W1 and W2 and either a sire or a dam analysis (Table 10), the values of TmLOD range from 2.1 (Flock 4) to 7.7 (Flock 1). With trios there is less variation (22.8, 19.6 and 18.4 for the three flocks with trio data). The TmLOD range for trios with 3 panels is also close (30.1 to 33.7). False negative (i.e. mLOD1 < TmLOD or Δ1 < TΔ) rates and false positive (i.e. mLOD2 > TmLOD and Δ1 > TΔ) rates fall as the number of panels in the assay goes from 2 to 3. For Flock 1, panels W1, W2 and W6 are perhaps slightly better than panels W1, W2 and W3, as there are fewer false negatives. However, for the other flocks including panel W3 is perhaps better than including panel W6 however there is very little difference. The proportion of simulated parents that were not correctly identified is very low for all of the flocks and assays. In the real data, generally the assignment rate increases as the number of panels goes from 2 to 3. The exception is Flock 4, for which assignment rates are much lower, and even with 2 panels there are two clear clusters in the data (Figure 2c, right). This suggests that the low assignment rate is not an error: it appears that 40% of dams and 50% of sires are not present in the data.

Table 10. Summary of TmLOD results for all flocks – Simulated data

PANEL	W12	W123	W126
Flock 1 Sire	5.2	9.4	7.1
Flock 1 Dam	7.7	10.7	10.2
Flock 1 Trio	22.8	33.7	33.0
Flock 2 Dam	5.9	9.7	9.1
Flock 3 Sire	4.6	4.7	9.8
Flock 3 Dam	6.1	6.7	8.0
Flock 3 Trio	19.6	33.3	30.1
Flock 4 Sire	2.1	4.3	30.3
Flock 4 Dam	4.3	4.6	7.8
Flock 4 Trio	18.4	32.8	30.3

Table 11 Rates of false positive and negative results for all flocks – Simulated data

	False + %	False - %	False + %	False - %	False + %	False - %
Flock 1 Sire	1.7	1.7	0.8	0.7	0.4	0.0
Flock 1 Dam	1.6	3.7	0.7	0.7	0.9	0.8
Flock 1 Trio	1.2	0.8	0.2	0.1	0.5	0.2
Flock 2 Dam	5.0	5.4	2.1	0.9	2.2	2.5
Flock 3 Sire	0.8	0.9	0.0	0.0	0.2	0.0
Flock 3 Dam	0.8	1.2	0.1	0.1	0.2	0.1
Flock 3 Trio	0.5	0.1	0.0	0.0	0.1	0.0
Flock 4 Sire	1.2	1.1	0.5	0.8	0.2	0.5
Flock 4 Dam	1.3	1.7	0.2	0.2	0.9	0.4
Flock 4 Trio	0.4	0.5	0.2	0.2	0.1	0.1

Table 12. Assignment rates for all flocks – Real data

PANEL	W12	W123	W126
Flock 1 Sire	97.3	98.3	99.5
Flock 1 Dam	88.2	94.7	95.4
Flock 1 Trio	96.4	97.8	98.6
Flock 2 Dam	81.4	81.4	90.7
Flock 3 Sire	80.6	86.4	79.6
Flock 3 Dam	81.6	91.3	88.3
Flock 3 Trio	77.7	78.6	78.6
Flock 4 Sire	49.4	48.3	48.3
Flock 4 Dam	60.0	62.2	60.0
Flock 4 Trio	28.3	28.9	28.9

In these data it was not possible to assign 100% of lambs to parents, and the distributions of the LOD scores indicate this was not due to a lack of power in the panels, but to the absence of the true parent in the data analysed. This could be due to our pre-screening of the data to include only SNP where we knew with certainty that a sample was to be treated as a sire, or treated as a dam, or treated as a lamb in the analysis. For a significant number of samples we did not have sufficient information to include the sample. This highlights an important point, the parentage assignments are conditional on knowing which animals are potential parents, the sex of the potential parents, and which animals are progeny. If this information is not available then erroneous parentage assignments may arise.

2. Exclusion analysis

The iterative approach to error and relationship detection identified subsets of 305 SNPs (including oY1) from the six panel set and 151 SNPs (including oY1) from panels 1, 2 and 3 that gave more accurate parentage detection using exclusion analysis than all SNPs in each set of panels. The 305 set of SNPs performed significantly better in exclusion analysis than did the 151 set of SNPs confirming the findings in Appendix C. Exclusion analysis using the 305 SNP panel identified 1,752 of the genotyped animals as being in relationships (Table 13). More relationships were found by exclusion analysis using 305 SNPs than were found using maximum likelihood analysis. A small number of differences in sire assignment were found between the two analysis methods. These discrepancies were largely caused by the maximum likelihood analysis requiring that samples be pre-classified into parent and progeny groups and the flocks being run separately. Unlike the exclusion analysis, the maximum likelihood approach ignored the possibility of sample class errors between parents and progeny. In a small number of instances this lead to mis-assignments by the maximum likelihood analysis where the true parent was not in the parent class. In addition, a number of animals were found to be represented as both progeny and parents (flocks 4, 5, 6) and these relationships were not found in the maximum likelihood approach.

A more stringent mismatch threshold was needed for flock 4 than for the other flocks when determining parent-child relationships with the 305 SNP panel as some full-siblings generated false parent-child and trio relationships with a parent. The false relationship problem was also found for flocks 1 and 5 when only the 151 set of SNPs was used to determine parentage. In addition a number of relationships were missed by the 151 set due to the number of mismatches exceeding the permitted threshold.

Table 13. Number and type of relationships found using 305 SNPs

Flock	Number of sheep in relationships	Number of Trios	Number of parent-child relationships not in trios	Number of progeny	Number of sires	Number of dams
1	690	419	27 + 18	464	10	253
2	220 + 3*	0 + 25 (from 3 CRC sires)	9 + 131	165	7 + 3*	92
3	274 + 2*	127 + 2 (from 2 CRC sires)	8 + 18	155	15 + 2*	114
4	217 + 5*	92 + 37 (from 4 CRC sires) + 6 (from 1 CRC dam)	4 + 49	188	5 + 4*	29 + 1*
5	163	0	156	156	7	0
6	188 + 1#	0	155 + 27 (from 1 flock 4 sire)	155 + 27#	7 + 1#	0

* parents found within CRC samples

# parent from flock 4

3. Poll/horn and other performance SNPs

A selection sweep peak for horns/poll identified OAR10_29511510 based on global F_{ST} (F_{ST} = 0.682) (OARv1 29.5 Mb near the RXFP2 gene - relaxin/insulin-like family petide receptor2; Kijas *et al,. PLoS Biology* 2012 **10** e1001258). This region was independently identified in a separate set of animals by Johnston *et al.*, (2011) *Mol Ecol* **20** 2555-2566. The OAR10_29511510 SNP was not included in the parentage set, however two nearby SNPs (OAR10_29389966_X.1 (Dominik *et al.,* 2012) and OAR10_29448537.1) were included in the panel to assess the correlation between genotypes and poll or horn phenotypes. Poll and horn phenotype information was provided for sires and progeny of the one flock, facilitating a preliminary investigation of the concordance between SNP genotypes and horn/poll status. The data was incomplete as no dams had phenotypes and not all sires were phenotyped. Where possible, the phenotype was inferred for non-phenotyped sires based on database information. The numbers of phenotypes was small, and possibly subject to error, but the results were not entirely consistent with OAR10_29389966_X.1 and OAR10_29448537.1 being able to correctly predict poll/horn genotype. Further validation on a data set with phenotypes known with certainty should be conducted before using these SNP as a reliable test for poll.

For the other performance SNP in the panels there was no phenotypic variation in the samples, so no validation could be conducted.

4. Sex determination

Gender information was provided for 1,335 (646 females, 689 males) of the 2,001 sheep tested by GeneSeek. Only two animals of the 2,001 genotyped by GeneSeek had oY1.1 heterozygous genotypes. Ten (1.5%) of the putative males genotyped as females (i.e. were oY1.1 –ve) and 66 (10.2%) of the putative females genotyped as males (oY1.1 +ve). This corresponded to an overall sex concordancy rate of 94%. It is likely that some of the sex discrepancies were due to sample handling or annotation errors.

5 Discussion, conclusion and recommendations

The project was commissioned to develop a pre-commercial SNP based tool for high accuracy and cost effective parentage assignment. The results indicate that a small number of SNP panels returned high assignment rates. In some flocks, use of two SNP panels would be sufficient to generate adequate assignment rates and deliver accurate parentage. During the early phase of market uptake it is likely that accuracy and high assignment rate are as important as low cost, meaning use of three panels may be desirable. Comparison of Figure 1a and 1 b clearly showed the improvement in assignment obtained when moving from use of two panels (W1 and 2) to three (W1, 2 and 3). The fact that three panels are sufficient (given the animals evaluated here) gives high confidence that a test can be offered at a price point below the existing microsatellite based tests. It is recommended that at some future date when sufficient industry parentage tests have been conducted from a range of flock structures that the issue of using two versus three panels for all industry flocks or only using two panels for flocks with suitable flock structures be further investigated. This will be of greater importance if the parentage tests continue to be performed in the US and the Australian dollar suffers a big loss in purchasing power resulting in a much more expensive parentage test.

A key outcome from the project is the successful development of SNP panels that are technically robust, and operate on a genotyping platform (SEQUENOM) that is able to genotype 180 SNP for around $10 in reagent costs. This opens the possibility of delivering a genomic product for parentage that will have deep industry uptake.

Another key outcome from the project was the development of an analytical method to assign parentage. The maximum likelihood based approach has at least two key advantages:

1. An assumed error rate.

A genotyping error rate was assumed that accounts for inevitable data errors. The SNP selected for incorporation in the 6 panels have technically robust assays, however the ability to assume a small error rate means the approach is superior to other methods which eliminate the true sire using a single data point (which may be an error). This assisted in achieving high assignment rates.

2. Empirically derived exclusion thresholds.

The approach uses the allele frequency at each SNP within the flock to generate population specific thresholds. The result is a responsive approach that tailors to the structure present within each flock. The observation that the clustering of simulated animals was highly similar to the real animals provided high confidence.

Together, the approach appears suitable to implement for delivering parentage assignment to industry.

The technical development of the SNP sets is complete, and an analytical approach to utilise the data for generating parentage results has been developed. Improvements can be made, for example by incorporation of SNP specific error rates to increase accuracy. It will also be desirable to establish the predictive power of the poll – horn test within Australian flocks. In addition, the successful commercialisation of a parentage test will require development of significant sampling handling and result reporting pipelines. The infrastructure required to receive, process and generate genotypes from 50,000 samples per year is currently resident within only a limited number of service laboratories.

The validation populations were not designed to provide a validation of the performance SNP, but for the polled/horn SNP there was some limited phenotypic data available. We conclude that the polled/horn test needs to be further validated, and possibly refined, prior to using the SNP parentage test also as a test for polled.

Discrepancies between genotyped and recorded sex represent an obvious, but low power, indication of sample handling or recording problems. The panel currently includes only a single Y chromosome SNP that can be used for sex determination. One problem with using a single Y chromosome SNP is that if the SNP fails to amplify then the animal is genotypically female (false –ve). However, more samples that were meant to be female were genotyped as male for the Australian Industry sheep population than the converse. The sex genotyped discrepancies could be due to either sample handling and annotation problems or assay problems (false +ve) or chimaerism. The robustness of the genotypic sex information would be improved if additional sex determining SNPs were added to the panel. The best alternative would be to add SNPs or alternate markers for genes in common with different sequences between the Y chromosome and the nonPAR X chromosome such as the Amelogenin genes, AMELX and AMELY (Pfeiffer & Brenig, 2005; sequencing has revealed that the AMELY gene itself contains a SNP). At least 6 high MAF nonPAR X SNPs would be needed for > 98% confidence that an animal homozygous for all was male. Consideration should be given to replacing some of the poorly performing SNPs (Appendix D) in the panel with additional sex determination SNPs.

A number of recommendations have been provided to assist with commercial implementation of the SNP described in this report:

1. SheepGenetics Australia (SGA) should implement a parentage genotype database system that stores the historic parentage genotypes and facilitates parentage analyses for flocks. SGA should note that parentage assays will likely be based on a mix of Sequenom parentage and SNP50 data and that the SNPs used in the parentage assay may change over time as may the service providers. The database and method used for assigning parentage should be capable of handling data from a combination of SNP50, the proposed HD chip, Sequenom (and other small scale

technology) assays. There will be more parentage discrepancies between Sequenom and SNP50 relationships than for either Sequenom-Sequenom or SNP50-SNP50. Note that whilst there was some overlap in poorly performing SNPs in Sequenom assays between Australian Sequenom and GeneSeek, there were also poorly performing SNPs that were specific to just one of the service providers.

2. Sire samples should be genotyped with each batch of progeny and compared to historic genotyping for the sire if available. If there is a conflict then either the sire will need to be re-genotyped to confirm which data set is correct, or if one set of genotypes is consistent with the putative progeny then this set of genotypes will be assumed to be the correct set. If sires have not previously been genotyped then it would be better to provide two samples per sire the first time they are genotyped so as to maximise the chance that the sire sample works well and there is not a sample error.
3. SGA should perform automated quality control procedures with appropriate alert thresholds on each batch of genotype data to identify any problems with a batch (SNP and sample call rates for different call qualities, SNP performance variation, duplicate identification, samples meant to be duplicates but not found, sex discrepancies, average SNP and sample intensity variation, heterozygosity rate per sample etc). These metrics should be performed per component panel set and for the whole data set. A sufficient number of duplicate samples should be included to check quality control/assay performance as it is possible that assay performance may be poorer for some batches or may degrade over time. The duplicate samples should include a male and a female sample so that sex determination is also checked. SGA may need to get intensity data to redo sex SNP calls if the service provider does not call sex SNPs correctly.
4. A SNP mismatch report should be produced for all the relationships found. This report should include the number of homozygous mismatches and number of homozygous comparisons for each parent as well as the number of trio mismatches and number of trio comparisons for any trios found. This report should be used to assist with QC with regard to SNP performance, sample quality and batch variation.
5. Further consideration should be given to performing the double homozygous parentage exclusion test on all samples in a batch to check that all potential sires and dams (samples with more than two relationships) have been correctly identified as potential parents in the sample list as an adjunct to performing the maximum likelihood parentage analysis. SGA could also consider using a double homozygous parentage exclusion test on historic male data to identify potential missing sires.
6. The producer should provide full sample identification (SG 16 digit identifier) with the samples. Failure to do this will reduce the power of the parentage assay as will not be able to perform likelihood calculations and will only be able to do double homozygote exclusion.
7. There will be a need to check that the allele calling system for GC or AT SNPs matches that recorded in the database if the service provider is changed (these SNPs can generate scoring inconsistencies - depending on how they are called whereas the strand and system is obvious for the AC (TG) or AG (TC) SNPs).

6 Acknowledgements

CSIRO Livestock Industries project staff Russell McCulloch (sampling handling and DNA preparations), Amy Bell (parentage analysis), Sonja Dominik (QTL analysis) and John Henshall (parentage analysis) have worked hard to deliver the results contained within this report.

A number of individuals have held the long term view that a DNA based tool for parentage has the potential to deliver large benefits to producers. Felice Driver had the vision and perseverance required to ensure this project was commissioned. She coordinated the transfer of reagents to GeneSeek, worked to put in place a Material Transfer Agreement that cleared the way for inclusion of the ISGC parentage SNP and played the key role needed to ensure the work was successful. Similarly, the support of Terry Longhurst and Hutton Oddy are acknowledged.

More recently, the work has been assisted by input from the Sheep CRC. Specifically, part of John Henshall's contribution to SheepCRC project 4.1 (Design and Analysis) was redirected into this project to develop the analytical approaches presented in the report. Julius van der Werf is acknowledged for his willingness to have the work adopted into the CRC program from July 2012, and liaising with GeneSeek which is the commercial provider of SNP genotyping. Within GeneSeek, Daniel Pomp has enthusiastically assisted with implementation of the existing SNP multiplexes, and Jeremy Walker is acknowledged for supervising the laboratory based activities that generated the SNP data.

These projects (BSC.0095 and BSC.0097) complemented the work of a commercial service and technology provider (Dr Darryl Irwin, SEQUENOM BSC.0098) for the purpose of developing the SNP multiplex panels, suitable protocols and QC systems. The International Sheep Genomics Consortium is acknowledged for making available the SNP used in this project.

Sam Gill (Sheep Genetics) coordinated collection of the industry flocks 1, 2 and 4. All of the producers are acknowledged for their willingness to participate through the provision of samples, with a special mention made to Marc Murphy at Karbullah.

Dr Jill Maddox performed the comparison of HQ and LQ samples, developed the mla_parent program for exclusion analysis and applied it to generate SNP specific error rates. In September 2014, she augmented the Final Report for project B.BCS.0095 to generate this combined report (projects BSC.0095 and BSC.0097).

7 Appendices

Appendix A: ISGC SNP parentage panel

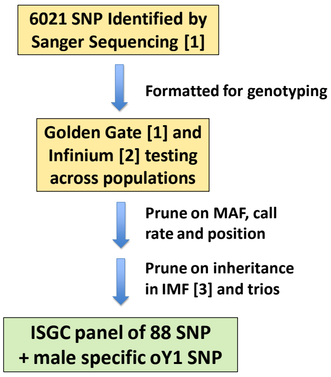
Development of a SNP Panel for Parentage Assignment in Sheep

James Kijas[1], John McEwan[2], Shannon Clarke[2], Hannah Henry[2], Jill Maddox[3], Russell McCulloch[1], Felice Driver[4], Katica Ilic[5], Mike Heaton[6] on behalf of the International Sheep Genomics Consortium[7]

[1] CSIRO Livestock Industries, Australia. [2] AgResearch, New Zealand. [3] University of Melbourne, Australia. [4] Meat and Livestock Australia. [5] Fluidigm Corporation, USA. [6] USDA, USA. [7] www.sheephapmap.org

The use of accurate pedigrees is important for livestock production systems and research projects. We present the development and attributes of a SNP panel for the assignment of parentage in sheep.

Figure 1: Work Flow for SNP Panel Design



6021 SNP Identified by Sanger Sequencing [1]

Formatted for genotyping

Golden Gate [1] and Infinium [2] testing across populations

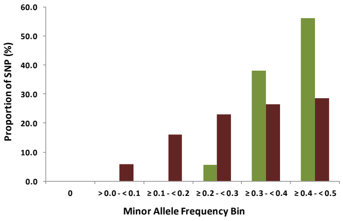
Prune on MAF, call rate and position

Prune on inheritance in IMF [3] and trios

ISGC panel of 88 SNP + male specific oY1 SNP

The genomic position and average minor allele frequency (MAF) of each SNP is given in Table 1. SNP were selected using population allele frequencies obtained from over 70 breeds sampled from 5 continents. The panel is biased towards high MAF markers (Figure 2) to ensure it will be of use across a wide range of breeds.

Figure 2: MAF



MAF was calculated by genotyping 2384 sheep from 74 breeds using the *ovine* SNP50 BeadChip [2]. MAF distributions are shown for all 49034 SNP on the BeadChip (red bars) and 89 SNP in the parentage panel (green bars, Table 1).

Four additional testing procedures (filters) have been applied to some, but not all SNP. Table 1 records the passage of SNP through each of the four additional filters.

Table 1: SNP ID, Genomic Location and Filter Testing

ISGC Parentage SNP	Chr	Mb Pos	Allele	MAF	Filter 1 Re-seq	Filter 2 Fluidigm	Filter 3 SQ_AgR	Filter 4 SQ_CLI	5k Chip	Filter problem
DU290101_408.1	1	7.8	A	0.337					🟩	3
DU518561_359.1	1	14.2	G	0.381		🟩	🟩	🟩	🟩	2
DU351298_316.1	1	69.6	A	0.445	🟩	🟩	🟩	🟩	🟩	
DU232924_365.1	1	95.8	G	0.250		🟩	🟩	🟩	🟩	
DU271929_382.1	1	97.5	A	0.483	🟩	🟩	🟩		🟩	4
DU502334_443.1	2	19.1	A	0.437		🟩	🟩	🟩	🟩	
DU469454_586.1	2	26.2	G	0.394	🟩		🟩	🟩	🟩	1
DU425907_184.1	2	50.1	G	0.358		🟩	🟩	🟩	🟩	
DU501115_497.1	2	62.8	A	0.239		🟩	🟩		🟩	
DU492516_411.1	2	63.4	T	0.478	🟩	🟩	🟩			
DU470875_383.1	2	91.5	G	0.357		🟩	🟩			
250506CS_*1	2	100.9	G	0.345			🟩	🟩	🟩	
DU191879_495.1	2	157.6	A	0.335		🟩	🟩	🟩	🟩	
DU480434_533.1	2	192.2	A	0.480		🟩	🟩	🟩	🟩	
DU260201_585.1	2	226.7	A	0.422		🟩	🟩	🟩	🟩	
DU503161_123.1	2	237.2	A	0.352		🟩	🟩	🟩		
DU425259_620.1	3	21.4	A	0.461		🟩	🟩			4
DU231007_156.1	3	59.0	G	0.463	🟩	🟩	🟩	🟩	🟩	2
DU225323_218.1	3	91.0	A	0.467		🟩	🟩	🟩	🟩	
DU260081_579.1	3	108.8	A	0.383			🟩	🟩	🟩	
DU394537_289.1	3	181.6	G	0.371		🟩	🟩	🟩	🟩	
CL635241_413.1	3	181.9	A	0.455	🟩	🟩	🟩	🟩	🟩	
DU408817_431.1	3	205.0	A	0.343		🟩	🟩	🟩	🟩	
DU202116_405.1	4	58.2	A	0.444	🟩	🟩	🟩		🟩	
DU460511_423.1	4	61.1	G	0.443			🟩	🟩	🟩	
DU305004_417.1	4	70.1	A	0.270		🟩	🟩	🟩	🟩	
DU369175_467.1	4	73.0	G	0.375		🟩	🟩	🟩	🟩	
DU446213_412.1	5	12.5	A	0.394		🟩	🟩	🟩	🟩	
DU444709_372.1	5	56.0	A	0.489		🟩	🟩	🟩	🟩	
DU453259_440.1	5	64.8	G	0.346		🟩	🟩	🟩	🟩	
DU194639_560.1	6	56.7	G	0.442		🟩	🟩		🟩	
CZ925803_293.1	6	100.8	A	0.443		🟩	🟩	🟩	🟩	
DU337465_337.1	6	106.0	A	0.338		🟩	🟩	🟩	🟩	
CL635944_160.1	6	115.0	A	0.490			🟩	🟩	🟩	4
DU467751_524.1	7	10.6	A	0.429		🟩	🟩	🟩		
DU499587_509.1	7	74.0	A	0.325		🟩	🟩		🟩	
CZ920950_468.1	7	74.8	A	0.456	🟩	🟩	🟩	🟩	🟩	
DU530067_219.1	7	100.0	G	0.327		🟩	🟩	🟩	🟩	
DU213735_493.1	8	6.6	A	0.437	🟩	🟩	🟩		🟩	
DU411204_551.1	8	13.8	A	0.361		🟩	🟩	🟩	🟩	
DU189970_325.1	9	86.6	C	0.374		🟩	🟩	🟩	🟩	
DU471913_499.1	9	91.1	G	0.490		🟩	🟩	🟩	🟩	
DU364754_308.1	9	93.9	A	0.397			🟩	🟩	🟩	
DU372582_268.1	9	94.4	G	0.247		🟩	🟩		🟩	2
DU468275_284.1	10	33.1	A	0.352		🟩	🟩	🟩		
DU310747_445.1	10	38.2	G	0.470		🟩	🟩		🟩	
DU269694_582.1	11	1.9	A	0.473		🟩	🟩	🟩	🟩	
DU433863_261.1	11	15.5	A	0.419		🟩	🟩	🟩	🟩	
DU417675_79.1	11	19.6	A	0.344			🟩	🟩	🟩	
DU508448_227.1	11	25.3	A	0.485		🟩	🟩	🟩	🟩	4
DU326572_241.1	11	59.5	A	0.446	🟩	🟩	🟩		🟩	
DU314655_578.1	12	26.7	A	0.365		🟩	🟩	🟩	🟩	
DU310703_497.1	12	75.3	A	0.492		🟩	🟩		🟩	1
DU275428_276.1	13	10.9	A	0.460		🟩	🟩	🟩	🟩	
DU435573_466.1	13	30.1	A	0.449		🟩	🟩	🟩	🟩	
DU411403_398.1	13	41.3	G	0.427		🟩	🟩	🟩	🟩	
DU462008_263.1	14	44.6	A	0.330		🟩	🟩	🟩	🟩	
DU223894_556.1	14	57.5	G	0.449		🟩	🟩	🟩	🟩	1
DU381045_479.1	14	60.7	A	0.403		🟩	🟩	🟩	🟩	
DU464373_638.1	15	2.3	A	0.467	🟩	🟩	🟩	🟩	🟩	
DU426312_454.1	15	44.4	G	0.375			🟩	🟩	🟩	
DU301502_402.1	15	73.7	G	0.441		🟩	🟩	🟩	🟩	
DU241306_191.1	15	78.6	G	0.279			🟩	🟩	🟩	
DU324670_456.1	17	10.2	A	0.400		🟩	🟩	🟩	🟩	
DU206327_107.1	17	14.4	A	0.499		🟩	🟩	🟩	🟩	
DU378819_632.1	17	22.3	A	0.475		🟩			🟩	3
DU511222_139.1	17	27.4	A	0.351		🟩	🟩	🟩	🟩	
DU300156_445.1	17	38.0	G	0.456		🟩	🟩	🟩	🟩	
DU463532_137.1	17	56.0	A	0.443		🟩	🟩	🟩	🟩	
DU492379_209.1	18	3.9	A	0.385			🟩	🟩	🟩	
DU488903_267.1	18	21.4	G	0.334		🟩	🟩			
DU325612_517.1	18	25.4	A	0.433	🟩	🟩	🟩	🟩	🟩	1
DU440765_491.1	18	60.5	A	0.474					🟩	3
DU345394_399.1	18	61.1	A	0.450		🟩	🟩	🟩	🟩	4
DU264531_279.1	19	0.6	A	0.388		🟩	🟩	🟩	🟩	
DU258053_237.1	19	57.1	A	0.400			🟩	🟩	🟩	
DU411432_523.1	19	57.2	C	0.406		🟩	🟩	🟩	🟩	
DU183112_480.1	20	31.1	A	0.453	🟩	🟩	🟩	🟩	🟩	
DU442373_141.1	20	48.4	A	0.342			🟩	🟩	🟩	
DU380983_440.1	21	28.3	G	0.451		🟩	🟩	🟩	🟩	
DU383863_376.1	21	38.2	G	0.443		🟩	🟩		🟩	1
DU196132_525.1	21	42.7	G	0.388		🟩	🟩	🟩	🟩	
DU413316_575.1	22	13.1	A	0.419		🟩	🟩	🟩	🟩	
DU302760_528.1	23	11.6	G	0.494		🟩	🟩	🟩	🟩	
DU313102_671.1	23	17.3	G	0.484			🟩	🟩	🟩	
CZ920359_258.1	24	3.2	G	0.382	🟩	🟩	🟩	🟩		
DU455254_479.1	25	0.1	G	0.453		🟩	🟩	🟩	🟩	
DU512685_259.1	25	1.2	G	0.495		🟩	🟩		🟩	
oY1	Y	0.0	G	0.320				🟩	🟩	

Table 1
The genomic location of each SNP (**Chr / Mb Pos**) is taken from the genome assembly version OAR2.0 available at [4]. SNP identifiers can be used to obtain additional information about each SNP [4]. The minor allele is given along with its frequency (**MAF**) in 2384 animals [2]. Four filters are described below, and irregularities arising from these additional tests are shown at right.

Filter1: Re-Sequencing
SNPs were re-sequenced from two nested PCR fragments produced from 96 diverse sires from 10 breeds. Passing SNPs could be reliably amplified and sequenced from genomic DNA without interference from nearby SNPs or other sequence features.

Filter2: Fluidigm Testing
SNP assays were designed using Early Access SNPType Assay Design Service and were tested on GT.96.96 microfluidic chip with SNPtype Assay Reagents. A panel of 95 animals was genotyped using genomic DNA. SNP assays exhibiting robust performance and high concordance rates against available SNP50 genotypes are shown.

Filter3: Sequenom Testing at AgResearch
Sequenom multiplexes were designed containing both parentage SNP (Table 1) and trait performance SNP unrelated to ISGC activities (not shown). Multiplexes were used to genotype pedigree material to prune SNP based on incorrect inheritance and call rate. SNP are shown that passed the filter.

Filter4: Sequenom Testing at MLA / CSIRO
Independent Sequenom multiplexes were designed by a second team. Multiplexes were used to genotype both high and low quality DNA samples. SNP are shown that passed both a call rate and concordance against SNP50 QC filter.

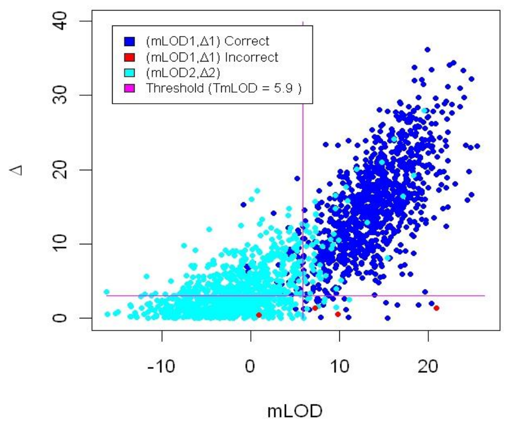
Key Points

- ✷ We have identified a technically robust set of SNP suitable for parentage analysis in a wide variety of sheep.
- ✷ Disclosure of the SNP and their attributes is intended to promote uptake by commercial partners.
- ✷ The panel will be promoted to the International Society of Animal Genetics (ISAG) as the standard for testing in sheep.
- ✷ The SNP presented form the backbone of panels in beta-testing by GeneSeek [5] and Pfizer Animal Genetics [6].
- ✷ Wide applicability of the SNP panel opens the way for international level trace-back and product of origin testing.

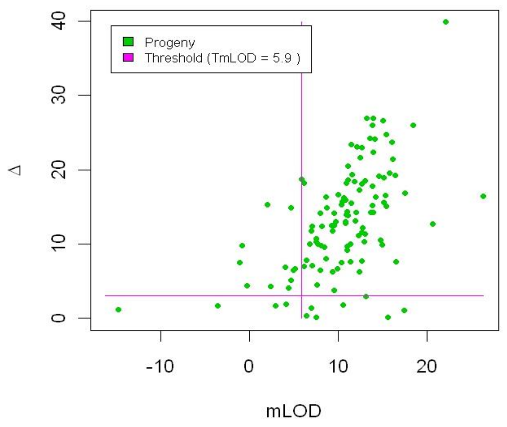
[1] Kijas et al. (2009) *PLoS ONE* 4:e4668
[2] Kijas et al. (2012) *PLoS Biology* (accepted pending minor revision)
[3] Crawford et al. (1995) *Genetics* 140:703-724.
[4] http://www.livestockgenomics.csiro.au/sheep/oar2.0.php
[5] http://www.neogen.com/GeneSeek/index.html
[6] http://www.pfizeranimalgenetics.co.nz/sites/PAG/aus/Pages/sheep.aspx

Appendix B: Plots of ∆ against mLOD for all flocks

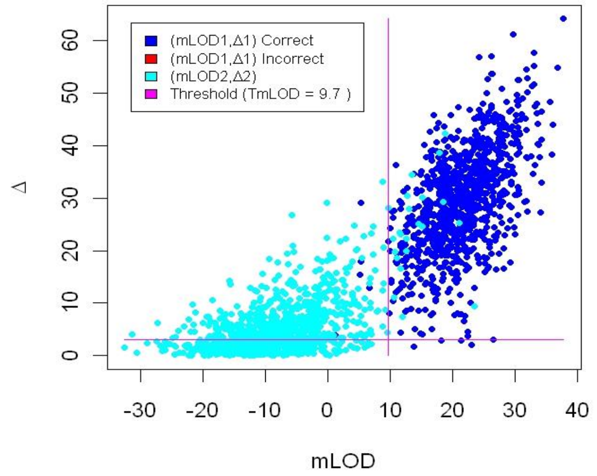
Flock_2_W12 Simulated: Dams



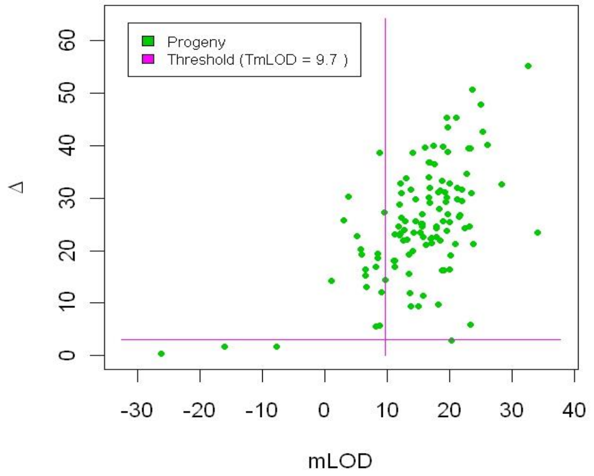
Flock_2_W12 Real: Dams



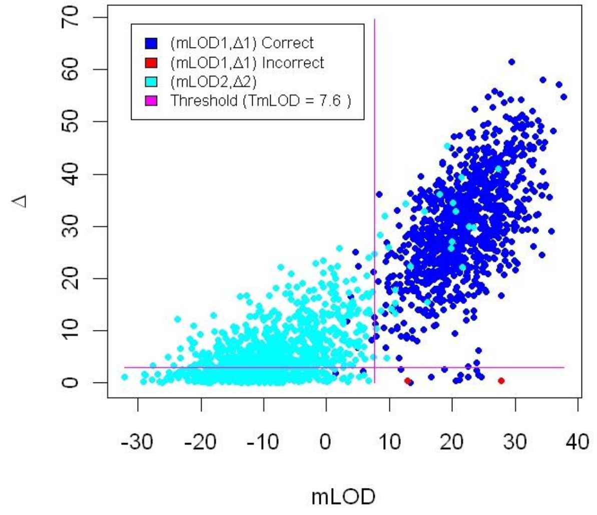
Flock_2_W123 Simulated: Dams



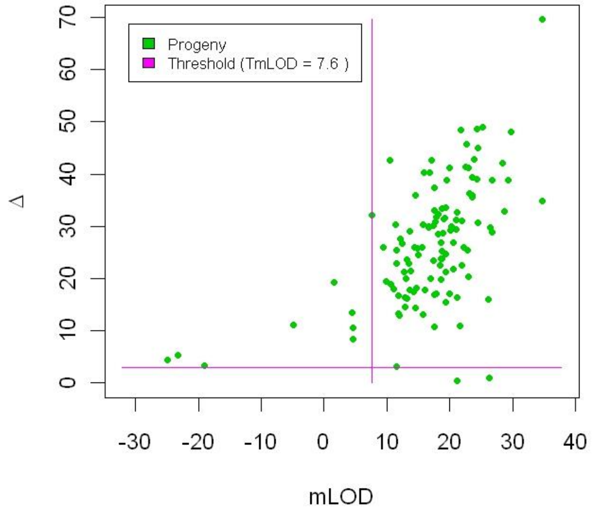
Flock_2_W123 Real: Dams



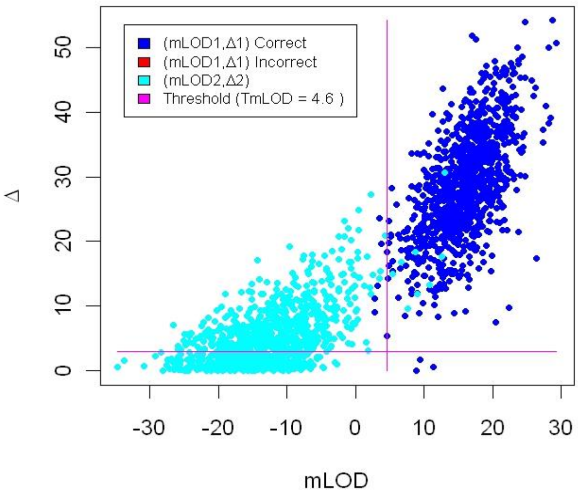
Flock_2_W126 Simulated: Dams



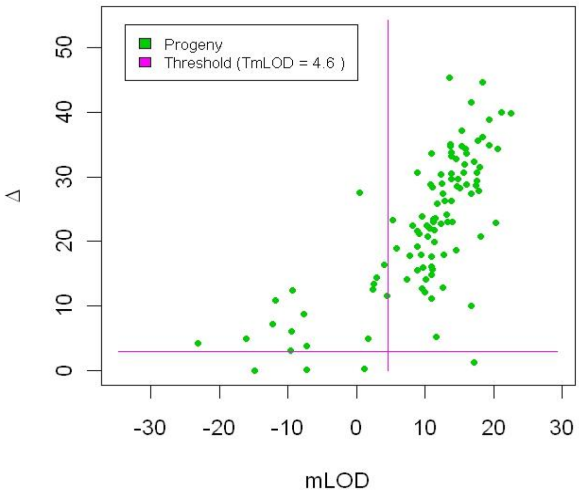
Flock_2_W126 Real: Dams



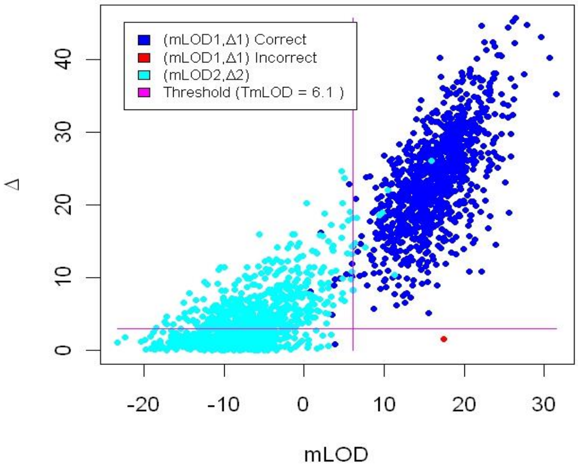
Flock_3_W12 Simulated: Sires



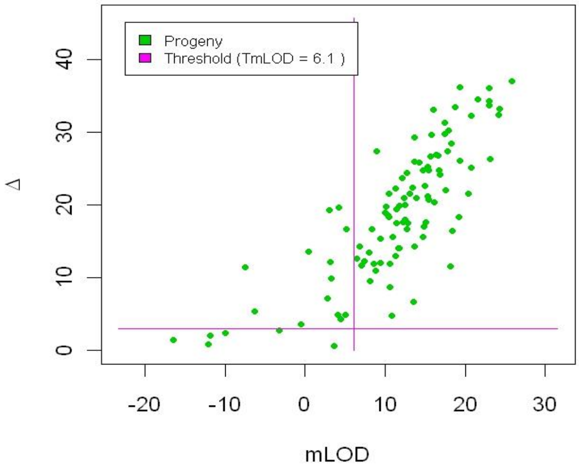
Flock_3_W12 Real: Sires



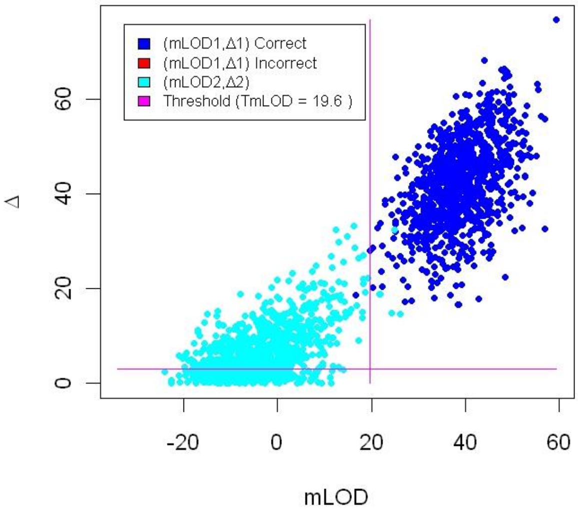
Flock_3_W12 Simulated: Dams



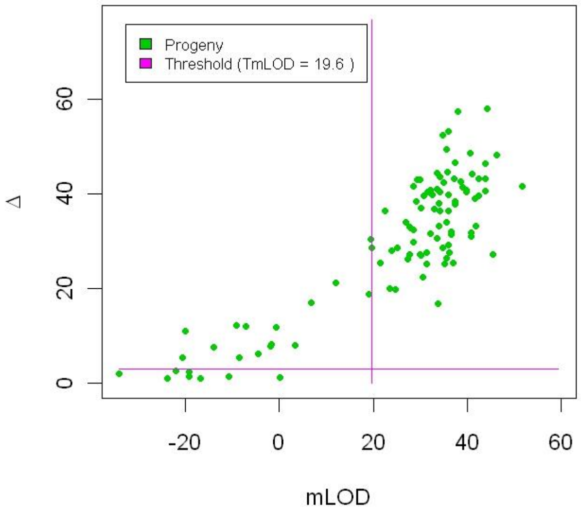
Flock_3_W12 Real: Dams



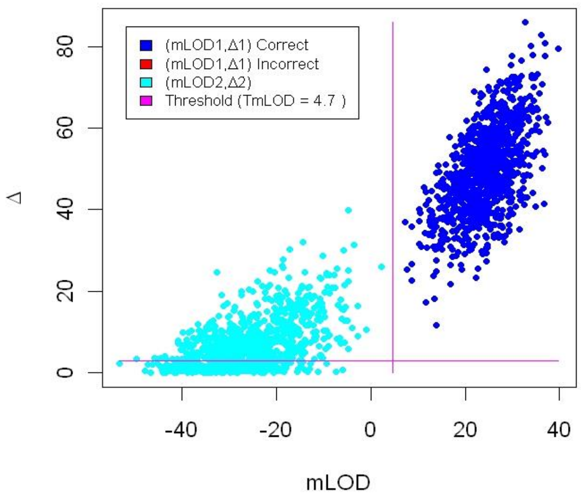
Flock_3_W12 Simulated: Trios



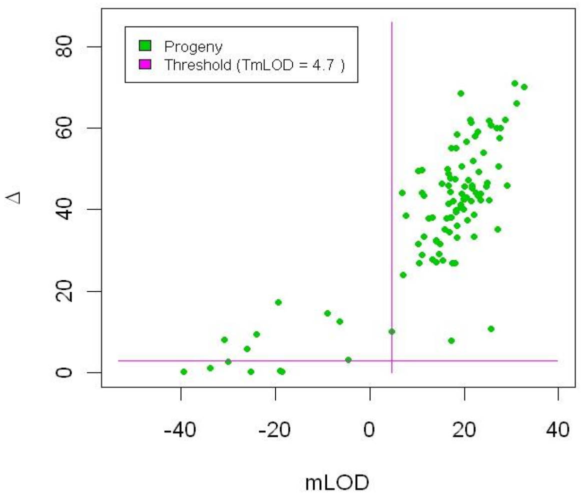
Flock_3_W12 Real: Trios



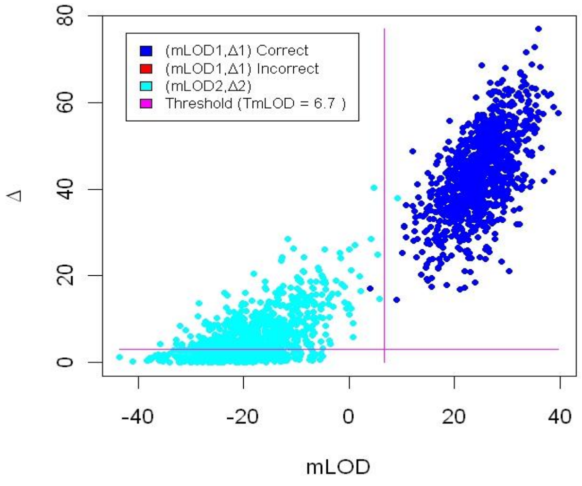
Flock_3_W123 Simulated: Sires



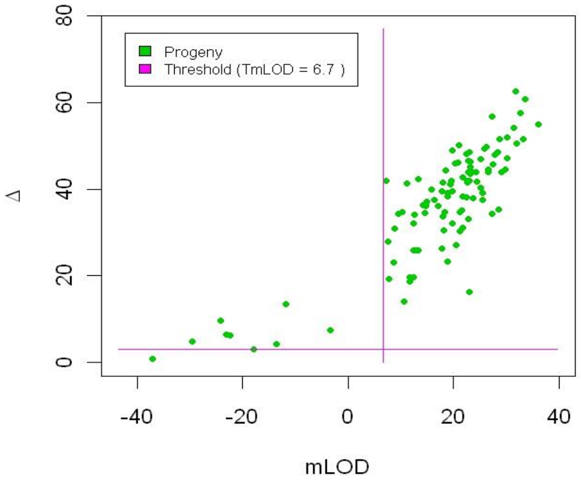
Flock_3_W123 Real: Sires



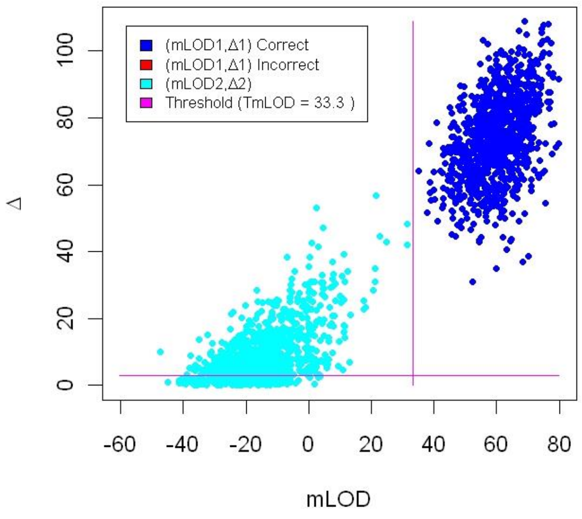
Flock_3_W123 Simulated: Dams



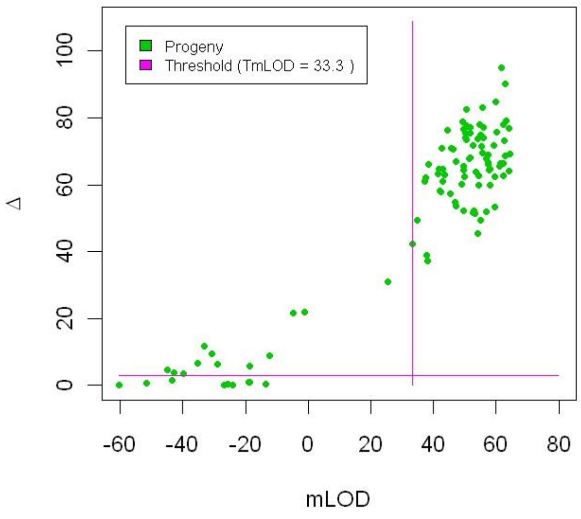
Flock_3_W123 Real: Dams

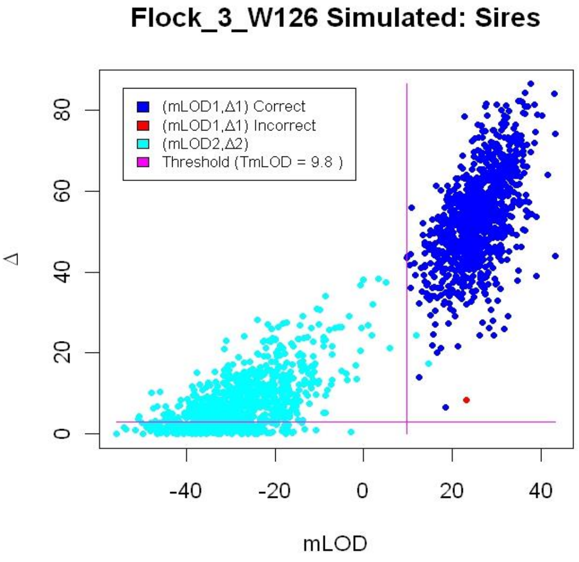


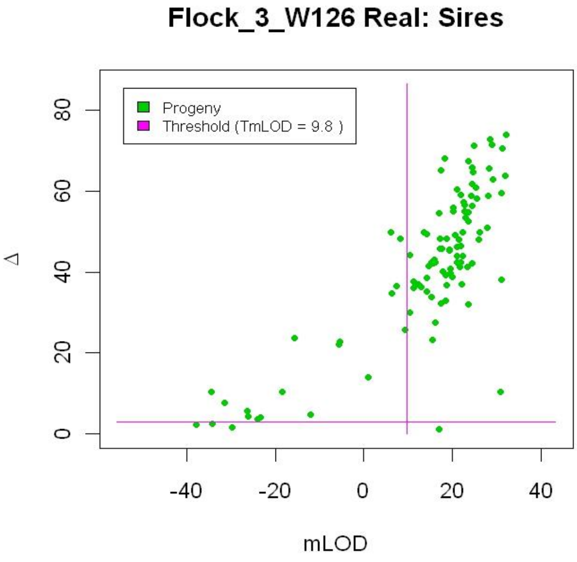
Flock_3_W123 Simulated: Trios

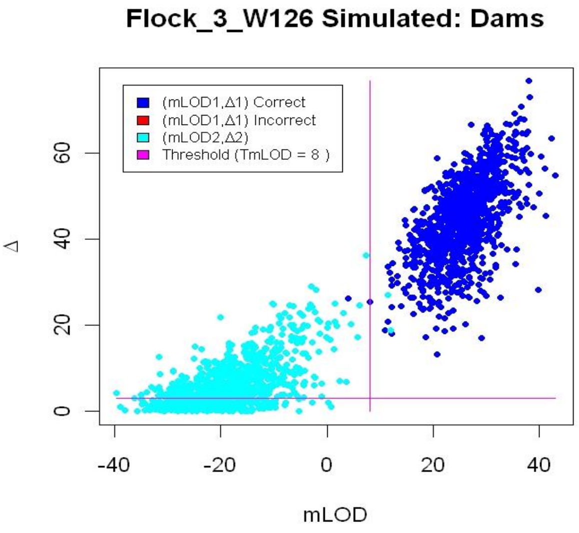


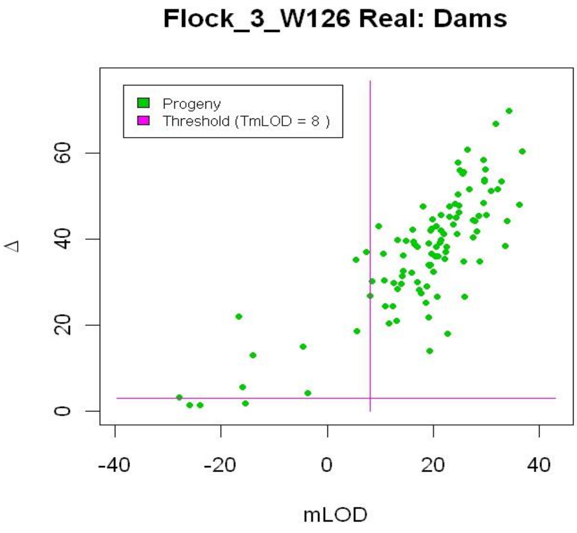
Flock_3_W123 Real: Trios

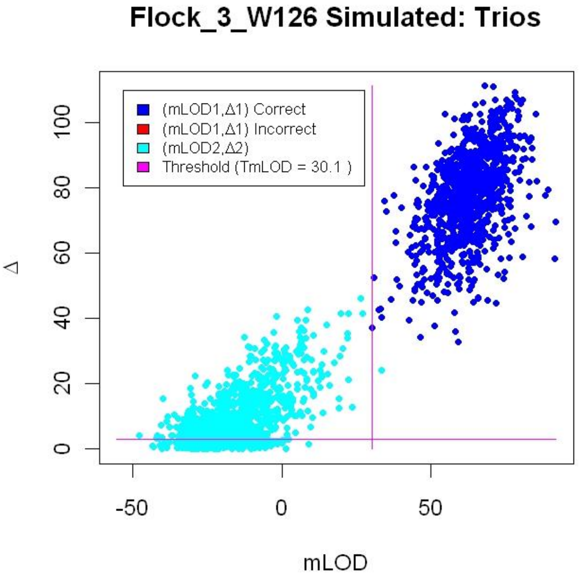


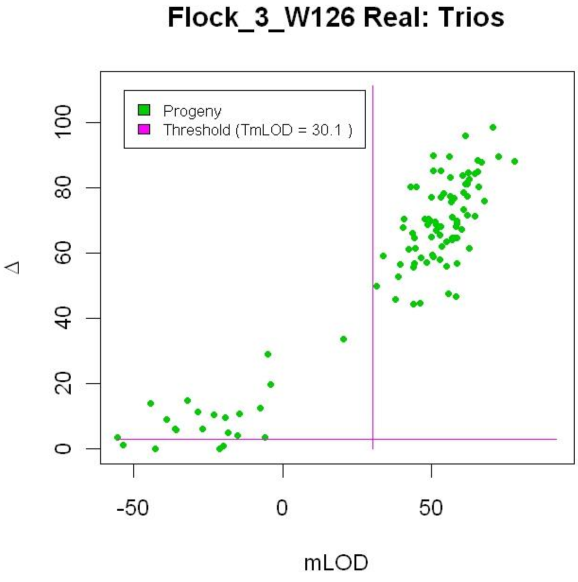

Flock_3_W126 Simulated: Sires
(mLOD1,Δ1) Correct
(mLOD1,Δ1) Incorrect
(mLOD2,Δ2)
Threshold (TmLOD = 9.8)
Δ
0
20
40
60
80
-40
-20
0
20
40
mLOD


Flock_3_W126 Real: Sires
Progeny
Threshold (TmLOD = 9.8)
Δ
0
20
40
60
80
-40
-20
0
20
40
mLOD

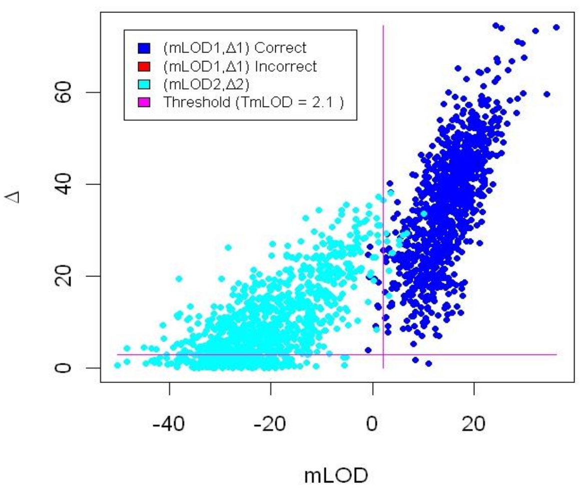

Flock_3_W126 Simulated: Dams
(mLOD1,Δ1) Correct
(mLOD1,Δ1) Incorrect
(mLOD2,Δ2)
Threshold (TmLOD = 8)
Δ
0
20
40
60
-40
-20
0
20
40
mLOD


Flock_3_W126 Real: Dams
Progeny
Threshold (TmLOD = 8)
Δ
0
20
40
60
-40
-20
0
20
40
mLOD

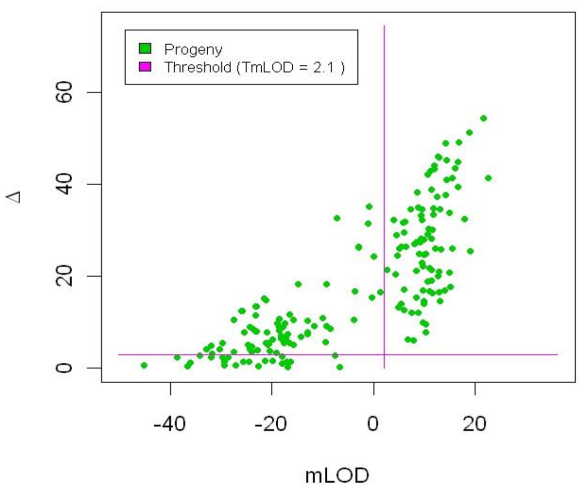

Flock_3_W126 Simulated: Trios
(mLOD1,Δ1) Correct
(mLOD1,Δ1) Incorrect
(mLOD2,Δ2)
Threshold (TmLOD = 30.1)
Δ
0
20
40
60
80
100
-50
0
50
mLOD


Flock_3_W126 Real: Trios
Progeny
Threshold (TmLOD = 30.1)
Δ
0
20
40
60
80
100
-50
0
50
mLOD

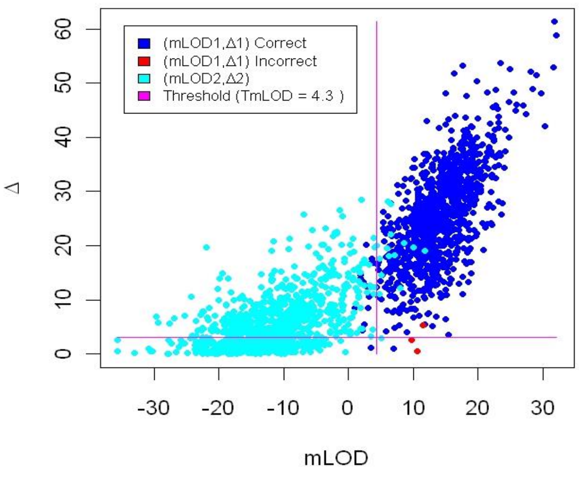
Flock_4_W12 Simulated: Sires



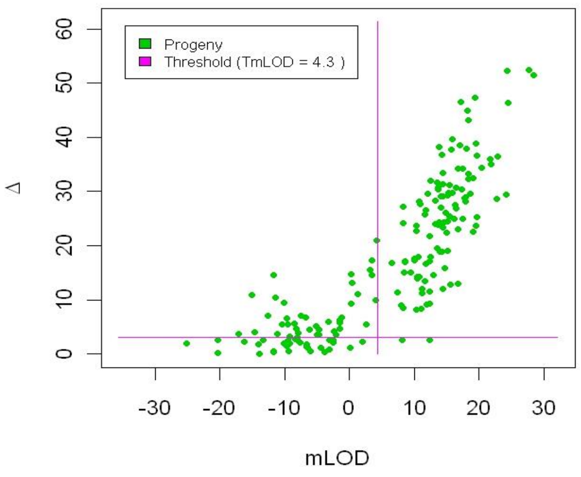
Flock_4_W12 Real: Sires



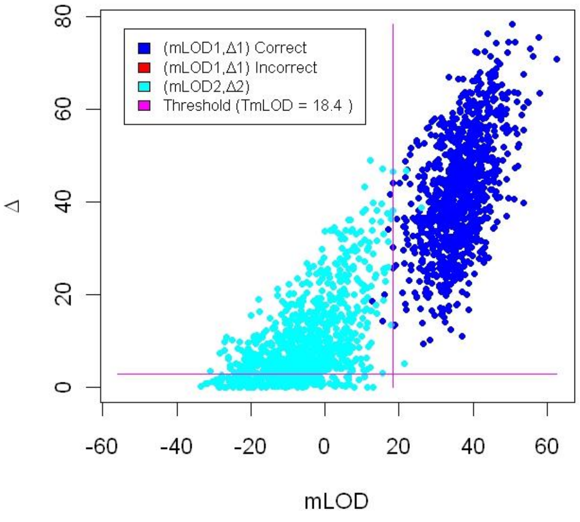
Flock_4_W12 Simulated: Dams



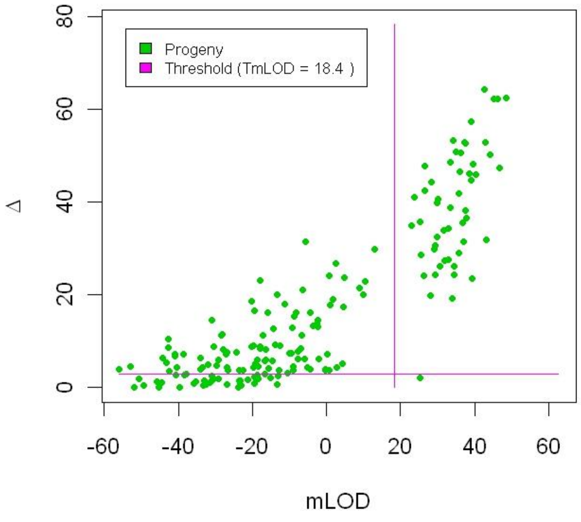
Flock_4_W12 Real: Dams



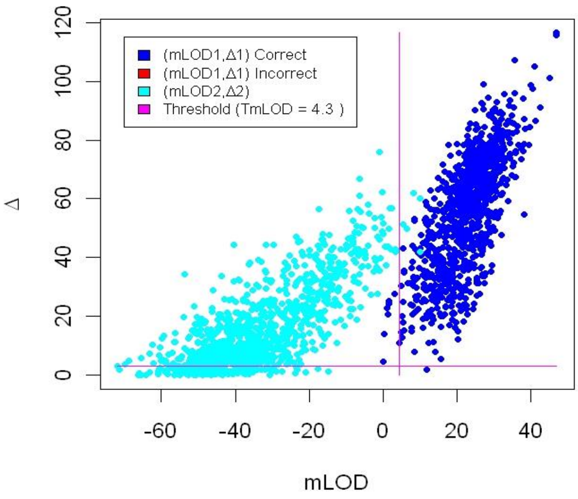
Flock_4_W12 Simulated: Trios



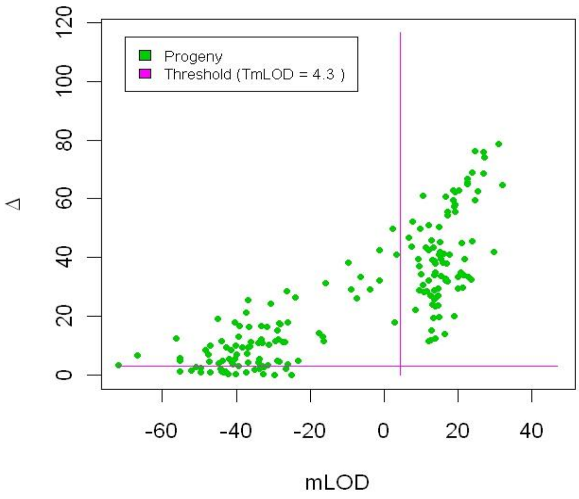
Flock_4_W12 Real: Trios



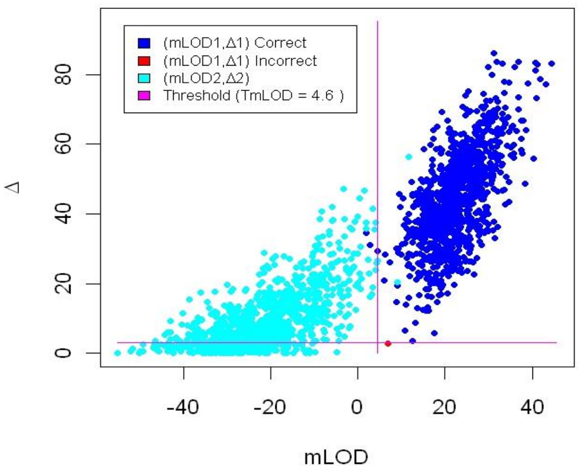
Flock_4_W123 Simulated: Sires



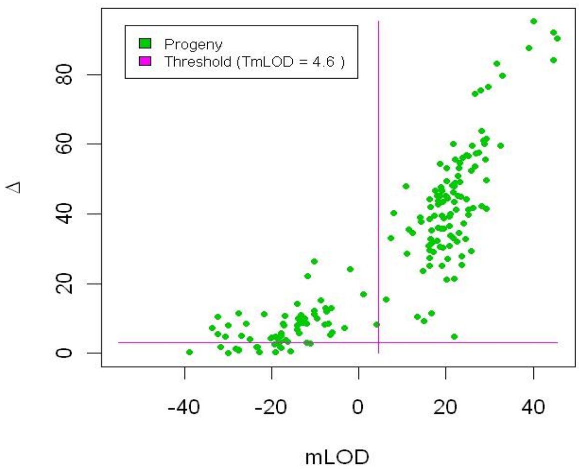
Flock_4_W123 Real: Sires



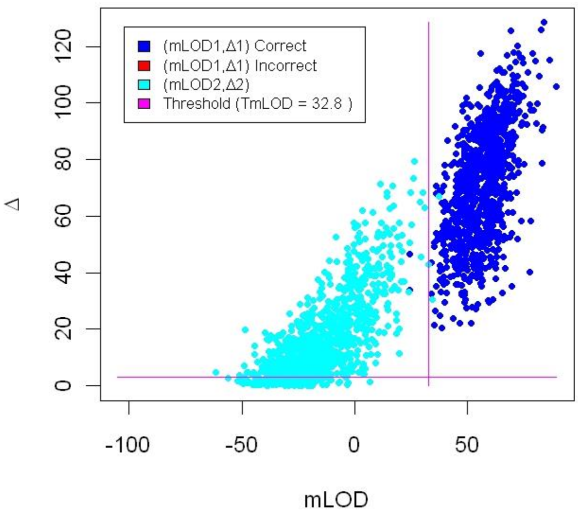
Flock_4_W123 Simulated: Dams



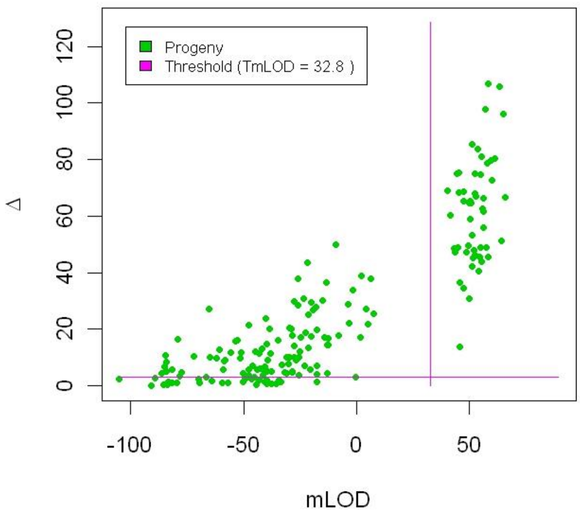
Flock_4_W123 Real: Dams



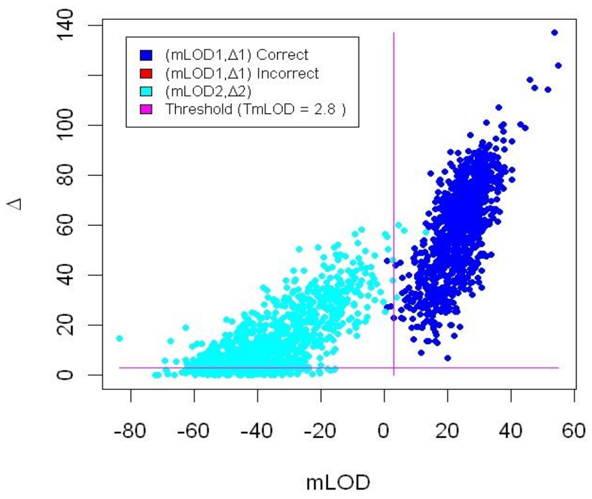
Flock_4_W123 Simulated: Trios



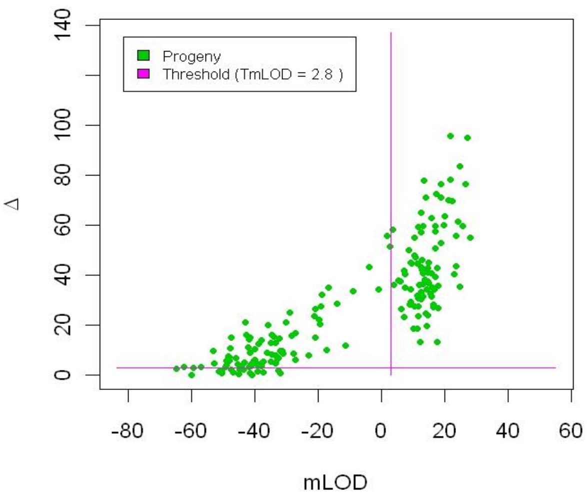
Flock_4_W123 Real: Trios



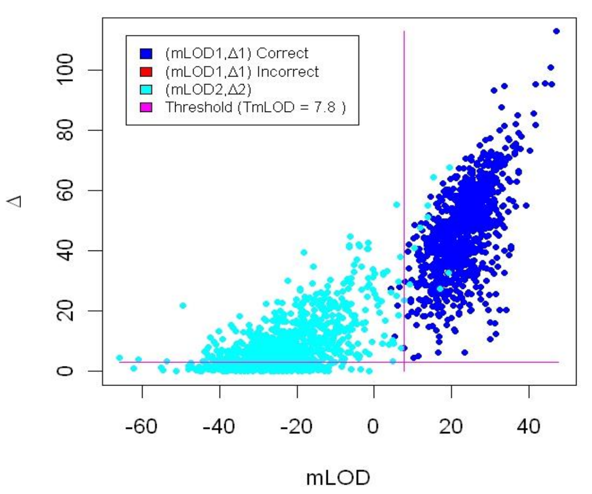
Flock_4_W126 Simulated: Sires



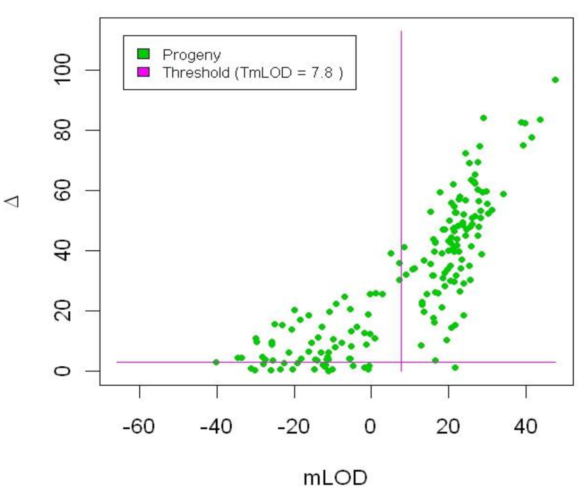
Flock_4_W126 Real: Sires



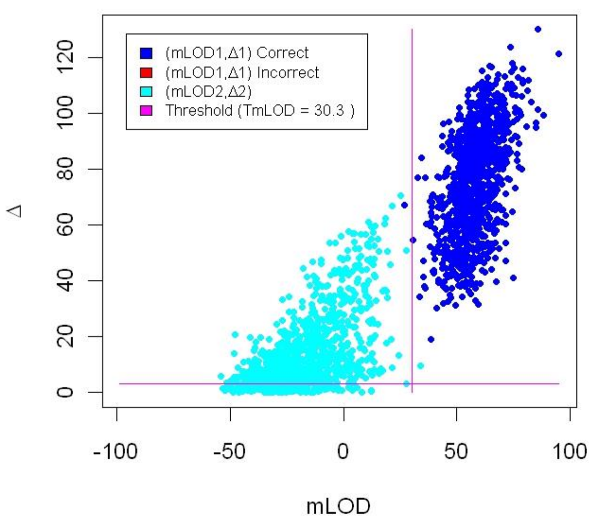
Flock_4_W126 Simulated: Dams



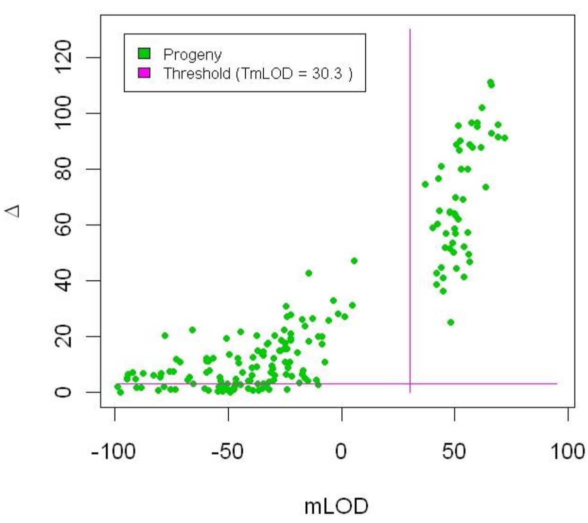
Flock_4_W126 Real: Dams



Flock_4_W126 Simulated: Trios



Flock_4_W126 Real: Trios



Appendix C: Using exclusion analysis to test different SNP panel sizes in a range of sheep populations

The mla_parent program was used to assess the usefulness of different sized SNP panels for determining parentage. Actual parentage assignment was first determined using a panel of 4,646 SNPs that was selected as part of the sheepGENOMICS SG.542 project. The 4,646 set was selected on the basis of high MAF, high GC scores, lack of heterozygote excess, presence on autosomes or X chromosome pseudo autosomal region and lack of other reasons for SNP failure. This set had a MAF of at least 0.4 for the first batch of sheepGENOMICS FMFS genotyping.

Prior to the design of the SEQUENOM sets the program was used to test three small parentage SNP panels on populations of sheep genotyped with either the 1.5K or SNP50 arrays as part of other projects. The test SNP panels comprised: (1) the set of 98 parentage SNPs provided by AgResearch plus oY1.1; (2) a set of 149 SNPs that includes the 98 AgResearch SNPs together with an additional 51 SNPs from the 4,538 filler SNP list provided by James Kijas plus oY1.1; and (3) a set of 239 SNPs plus oY1.1 which comprised the 149 set plus an additional 90 SNPs from the filler SNP list. The additional filler SNPs were selected based on their OarV1 genome sequence positions.

Only the 98 SNP panel was able to be tested on the AWI pedigree test population (AWI Project WP89) as it was not genotyped with the SNP50 chip. Tested population structures were either 2-generation, half-sibling or multi-generation full and half-sibling. The male oY1.1 SNP was also included in the panels and all animals that were heterozygous for oY1.1 were failed. It should be noted that the genotype data revealed that some service providers had problems with oY1.1 scoring as they failed to realise that it was not an autosomal SNP when performing genotype calls.

SheepGENOMICS data set – number of pcs (duplicates removed from data set) when 4,222 true pcs in data set

	98 SNPs + oY1	149 SNPs + oY1	239 SNPs + oY1
pcs 0 mismatches true (false)	25458 4203 (3640)	6272 4203 (257)	4363 4201 (1)
1 mismatch true (false)	1 (17614)	1 (1795)	40 (3)
2 mismatches true (false)		0 (16)	118 (0)
number of missed pcs*	18	18	18

* pcs missed relate to sample failures due to insufficient SNPs genotyping successfully

Results for the sheep CRC data set*

number of	98 SNPs + oY1	149 SNPs + oY1	239 SNPs + oY1	4646 + oY1
samples passed	5038	5042	5035	5008
samples failed	408	404	411	438
pc pairs	8300	1199	671	587
trios	12	10	0	0
duplicate pairs	1659	1659	1659	1659
failed snps#	0	0	0	3

* includes sample duplicates

SNPs failed when

It is clear from the two tables that the panel of 239 SNPs gave the best results for the three small panels that were tested with there being only a very small number of false assignments even when 2 mismatches were allowed. It is also clear from the above that a panel of 98 SNPs is definitely not enough for parentage assignment as there are lots of false positive possibilities re parent-child combinations. A panel of 149 SNPs performs a lot better than does the 98 SNP panel but there are still a lot of false findings.

Appendix D: SNP composition of multiplexes and error data

A description of each quality control test 1 - 8 and error rates are provided below the table.

		Quality Control Test								Error Rate	
Multiplex	**SNP Identifier**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**
W1	250506CS3901012300001_913.1	7		1						0.004	0.010
W1	CL635241_413.1	1								0.000	0.010
W1	CL635944_160.1	33		16						0.059	0.059
W1	DU183112_480.1	1								0.000	0.010
W1	DU191879_495.1	5	3							0.016	0.016
W1	DU194639_560.1	8		3						0.011	0.011
W1	DU202116_405.1	12	1	8						0.035	0.035
W1	DU232924_365.1									0.000	0.010
W1	DU258053_237.1	1								0.000	0.010
W1	DU259120_464.1	15				Y	Y	Y	Y	0.000	0.010
W1	DU269694_582.1	2		3						0.011	0.011
W1	DU302760_528.1	1								0.000	0.010
W1	DU310703_497.1	1								0.000	0.010
W1	DU325267_788.1	27	2	23		Y	Y			0.096	0.096
W1	DU325612_517.1	1	1							0.005	0.010
W1	DU329154_467.1	25	2	37			Y			0.148	0.148
W1	DU337465_337.1	1	1							0.005	0.010
W1	DU342117_350.1	40		28		Y	Y			0.104	0.104
W1	DU345394_399.1	13		7						0.026	0.026
W1	DU348827_210.1	8	1	5						0.024	0.024
W1	DU369175_467.1	1								0.000	0.010
W1	DU408817_431.1	2								0.000	0.010
W1	DU411432_523.1	7	1							0.005	0.010
W1	DU413316_575.1	4					Y			0.000	0.010
W1	DU417675_79.1	1								0.000	0.010
W1	DU425907_184.1			1						0.004	0.010
W1	DU426312_454.1	71	4	65						0.262	0.262

W1	DU434120_194.1	6	2							0.010	0.010
W1	DU442373_141.1									0.000	0.010
W1	DU444709_372.1	2								0.000	0.010
W1	DU453259_440.1	7	1							0.005	0.010
W1	DU455254_479.1	11	3	3						0.027	0.027
W1	DU460511_423.1	9	5							0.026	0.026
W1	DU464373_638.1	1	1							0.005	0.010
W1	DU467751_524.1	1								0.000	0.010
W1	DU468275_284.1	2	2							0.010	0.010
W1	DU469454_586.1	4	2							0.010	0.010
W1	DU470875_383.1	13	4	1						0.025	0.025
W1	DU488903_267.1	8	3		Y					0.016	0.016
W1	DU492516_411.1	5		1		Y				0.004	0.010
W1	DU508448_227.1	14		14						0.052	0.052
W1	DU512685_259.1									0.000	0.010
W1	DU518561_359.1	18	2							0.010	0.010
W1	DU530067_219.1									0.000	0.010
W1	FM872310_2746delCA.1									0.000	0.010
W1	MSTN.1									0.000	0.010
W1	OAR10_29389966_X.1	1								0.000	0.010
W1	OAR13_15714447.1	3	2	3						0.022	0.022
W1	OAR13_19104238.1									0.000	0.010
W1	OAR17_2380024.1	5	2							0.010	0.010
W1	OAR19_33531772.1					Y				0.000	0.010
W1	OAR19_37492499.1	2								0.000	0.010
W1	OAR19_38337561.1			3			Y			0.011	0.011
W1	OAR1_163269760.1	7	3							0.016	0.016
W1	OAR20_2774290.1	3	1	1						0.009	0.010
W1	OAR25_25184395.1	3	1							0.005	0.010
W1	OAR2_33972713.1	23	7	7						0.063	0.063
W1	OAR3_72722977.1	8	1	1						0.009	0.010
W1	OAR6_21138069.1	19		9		Y	Y			0.033	0.033
W1	oY1.1	24	9							0.047	0.047
W1	PITX3_MMFwd.12.1				Y			Y	Y	0.000	0.010
W1	s22341.1	8	1							0.005	0.010

W1	s31833.1	40	14	6						0.096	0.096
W1	s68324.1	49	1	29		Y				0.113	0.113
W2	CZ920950_468.1									0.000	0.010
W2	DU196132_525.1	1	1							0.005	0.010
W2	DU206327_107.1	1								0.000	0.010
W2	DU223894_556.1	10	2	1		Y				0.014	0.014
W2	DU225323_218.1	1								0.000	0.010
W2	DU241306_191.1									0.000	0.010
W2	DU264531_279.1	3								0.000	0.010
W2	DU300156_445.1									0.000	0.010
W2	DU305004_417.1	2	1							0.005	0.010
W2	DU310747_445.1	3								0.000	0.010
W2	DU314655_578.1	7								0.000	0.010
W2	DU324670_456.1									0.000	0.010
W2	DU326572_241.1	1	1							0.005	0.010
W2	DU351298_316.1									0.000	0.010
W2	DU364754_308.1				Y	Y		Y	Y	0.000	0.010
W2	DU378819_632.1	3	1							0.005	0.010
W2	DU380983_440.1	5	2	2						0.018	0.018
W2	DU381045_479.1	1								0.000	0.010
W2	DU394537_289.1									0.000	0.010
W2	DU405213_575.1	17		12		Y				0.044	0.044
W2	DU433863_261.1	1	1	1		Y				0.009	0.010
W2	DU435573_466.1	2		2						0.007	0.010
W2	DU438000_178.1	27		27		Y	Y			0.100	0.100
W2	DU446213_412.1	1								0.000	0.010
W2	DU471913_499.1	4	3		Y	Y			Y	0.016	0.016
W2	DU480434_533.1	1								0.000	0.010
W2	DU492379_209.1	2								0.000	0.010
W2	DU499587_509.1	1								0.000	0.010
W2	DU502334_443.1	2		3	Y	Y	Y		Y	0.011	0.011
W2	DU503161_123.1	1								0.000	0.010
W2	OAR10_29448537.1					Y				0.000	0.010
W2	OAR10_51775366.1	1	1							0.005	0.010
W2	OAR10_92199067.1									0.000	0.010

W2	OAR12_2436404.1	3	2							0.010	0.010
W2	OAR12_45884879.1	3								0.000	0.010
W2	OAR14_26710874.1	70	2	48		Y	Y			0.188	0.188
W2	OAR17_31146799.1	5	2							0.010	0.010
W2	OAR17_57234757.1	2								0.000	0.010
W2	OAR18_10748526.1	6	2							0.010	0.010
W2	OAR19_33968342.1					Y				0.000	0.010
W2	OAR19_34354181.1	1				Y				0.000	0.010
W2	OAR1_155904685.1	4							Y	0.000	0.010
W2	OAR1_185320639.1									0.000	0.010
W2	OAR1_62521467.1	1								0.000	0.010
W2	OAR21_20371526.1	1								0.000	0.010
W2	OAR22_3499143.1									0.000	0.010
W2	OAR24_17892863.1									0.000	0.010
W2	OAR3_202992242.1									0.000	0.010
W2	OAR3_238210924.1	51	7	22			Y	Y		0.118	0.118
W2	OAR3_57525359.1	18	2	11						0.051	0.051
W2	OAR3_95548607.1	4		1						0.004	0.010
W2	OAR5_89502724.1									0.000	0.010
W2	s03538.1	21	1	11		Y				0.046	0.046
W2	s06750.1	4								0.000	0.010
W2	s13861.1	3		2						0.007	0.010
W2	s44590.1	7								0.000	0.010
W2	s45944.1	2			Y					0.000	0.010
W2	s50176.1									0.000	0.010
W2	s62387.1			3		Y				0.011	0.011
W2	s62974.1			1						0.004	0.010
W2	s68067.1	36	2	24			Y			0.099	0.099
W2	s69652.1	13					Y		Y	0.000	0.010
W2	s73968.1	3	1							0.005	0.010
W3	CZ920359_258.1	6		7						0.026	0.026
W3	CZ925803_293.1	1	1							0.005	0.010
W3	DU189970_325.1	3							Y	0.000	0.010
W3	DU213735_493.1	15	12		Y	Y			Y	0.063	0.063
W3	DU231007_156.1	1								0.000	0.010

W3	DU260081_579.1	1								0.000	0.010
W3	DU260201_585.1									0.000	0.010
W3	DU271929_382.1	41		32		Y	Y		Y	0.119	0.119
W3	DU275428_276.1									0.000	0.010
W3	DU290101_408.1	5								0.000	0.010
W3	DU301502_402.1									0.000	0.010
W3	DU383863_376.1									0.000	0.010
W3	DU411204_551.1	1								0.000	0.010
W3	DU411403_398.1									0.000	0.010
W3	DU425259_620.1	18		7		Y				0.026	0.026
W3	DU462008_263.1	1								0.000	0.010
W3	DU463532_137.1									0.000	0.010
W3	DU511222_139.1					Y				0.000	0.010
W3	OAR10_62894887.1									0.000	0.010
W3	OAR11_62887032.1									0.000	0.010
W3	OAR12_11657392.1	4								0.000	0.010
W3	OAR12_67278197.1									0.000	0.010
W3	OAR15_12666876.1	64	3	37	Y	Y				0.153	0.153
W3	OAR17_6382522.1	1								0.000	0.010
W3	OAR18_44175536.1	2								0.000	0.010
W3	OAR1_125382442.1	8			Y				Y	0.000	0.010
W3	OAR1_218727571.1	9	3					Y		0.016	0.016
W3	OAR1_226894517.1	33	1	28			Y			0.109	0.109
W3	OAR1_227032731.1									0.000	0.010
W3	OAR1_80034351.1	1								0.000	0.010
W3	OAR1_99762260.1	4	1	1						0.009	0.010
W3	OAR20_14039387.1	2								0.000	0.010
W3	OAR26_10633898.1	4		8						0.030	0.030
W3	OAR26_34360075.1	22		22		Y	Y			0.081	0.081
W3	OAR2_155018930.1	2		1						0.004	0.010
W3	OAR2_47914972.1	10		2		Y				0.007	0.010
W3	OAR3_145344922.1	1								0.000	0.010
W3	OAR3_177781806.1									0.000	0.010
W3	OAR3_226847164.1									0.000	0.010
W3	OAR3_40698853.1									0.000	0.010

W3	OAR4_6967496.1								0.000	0.010
W3	OAR5_101392888.1	123	3	93	Y	Y			0.360	0.360
W3	OAR5_110500655.1								0.000	0.010
W3	OAR6_122433556.1								0.000	0.010
W3	OAR8_36117636.1	49	2	56	Y				0.218	0.218
W3	OAR8_38564574.1	123	1	114	Y	Y			0.427	0.427
W3	OAR9_46531990.1								0.000	0.010
W3	OAR9_88113366.1				Y				0.000	0.010
W3	s03976.1								0.000	0.010
W3	s12930.1								0.000	0.010
W3	s18401.1	1							0.000	0.010
W3	s23738.1								0.000	0.010
W3	s26070.1	2		2					0.007	0.010
W3	s26699.1	1	1						0.005	0.010
W3	s36879.1								0.000	0.010
W3	s37564.1								0.000	0.010
W3	s39944.1	2							0.000	0.010
W3	s42102.1	8	3						0.016	0.016
W3	s44648.1								0.000	0.010
W3	s52103.1	195	8	164	Y	Y			0.649	0.500
W3	s55239.1	4	1						0.005	0.010
W3	s65026.1								0.000	0.010
W3	s65582.1	7	1		Y				0.005	0.010
W4	DU313102_671.1	3	2						0.010	0.010
W4	DU372582_268.1	1							0.000	0.010
W4	OAR10_20810050.1	1							0.000	0.010
W4	OAR10_49494142.1					Y			0.000	0.010
W4	OAR11_3851247.1	64	15	47	Y	Y			0.253	0.253
W4	OAR13_26732874.1	2	1						0.005	0.010
W4	OAR15_54550843.1	2							0.000	0.010
W4	OAR15_82382470.1	34	3	10					0.053	0.053
W4	OAR17_22571786.1	10	1						0.005	0.010
W4	OAR18_10126276.1	17	4			Y		Y	0.021	0.021
W4	OAR18_16861687.1	4		2					0.007	0.010
W4	OAR18_26482558.1	21	4	3				Y	0.032	0.032

W4	OAR19_37223995.1	1	1							0.005	0.010
W4	OAR1_169681073.1	1								0.000	0.010
W4	OAR1_172538566.1	24	6	12						0.076	0.076
W4	OAR1_207602058.1	8								0.000	0.010
W4	OAR1_227870704.1	99	1	98		Y	Y			0.368	0.368
W4	OAR1_260750419.1	66	3	56		Y	Y			0.223	0.223
W4	OAR1_47348433.1	2	1							0.005	0.010
W4	OAR1_79404769.1	12								0.000	0.010
W4	OAR1_86712802.1	3		3						0.011	0.011
W4	OAR20_1216753.1	3	1							0.005	0.010
W4	OAR20_33384221.1									0.000	0.010
W4	OAR20_43815997_X.1	4		3						0.011	0.011
W4	OAR21_5011592.1	2								0.000	0.010
W4	OAR22_1023592.1	1								0.000	0.010
W4	OAR23_8935783.1	5	2	1						0.014	0.014
W4	OAR24_19994470.1									0.000	0.010
W4	OAR24_32232601.1	4	2							0.010	0.010
W4	OAR25_8324086.1									0.000	0.010
W4	OAR26_6517460.1	5	1							0.005	0.010
W4	OAR2_137718678.1				Y	Y		Y	Y	0.000	0.010
W4	OAR3_117626019.1	4	1							0.005	0.010
W4	OAR3_177903329.1									0.000	0.010
W4	OAR3_183012984.1	2	1							0.005	0.010
W4	OAR3_191490835.1									0.000	0.010
W4	OAR3_34209284.1	1	1							0.005	0.010
W4	OAR3_71384072.1	2		1						0.004	0.010
W4	OAR3_83593412.1									0.000	0.010
W4	OAR4_57594871.1	9	5		Y		Y		Y	0.026	0.026
W4	OAR6_101640082.1									0.000	0.010
W4	OAR6_74265158.1	9		3						0.011	0.011
W4	OAR6_94177376_X.1	11	2	2	Y		Y		Y	0.018	0.018
W4	OAR6_98890184.1	2	1							0.005	0.010
W4	OAR8_14718655.1	5		1						0.004	0.010
W4	OAR9_16213053.1	9	4							0.021	0.021
W4	OAR9_25781979.1									0.000	0.010

W4	OAR9_83679492.1	96	3	48		Y	Y			0.194	0.194
W4	s06272.1	59	2	67			Y		Y	0.259	0.259
W4	s11273.1	2								0.000	0.010
W4	s13172.1	5	2							0.010	0.010
W4	s15406.1	1								0.000	0.010
W4	s19793.1	58	1	40		Y	Y			0.153	0.153
W4	s19983.1									0.000	0.010
W4	s20231.1	2	1							0.005	0.010
W4	s21986.1	2		5						0.019	0.019
W4	s28866.1	1								0.000	0.010
W4	s45473.1	2								0.000	0.010
W4	s58112.1	85	4	63		Y	Y			0.254	0.254
W4	s60073.1	33	2	22					Y	0.092	0.092
W4	s62286.1	11	1						Y	0.005	0.010
W4	s62749.1	1								0.000	0.010
W4	s62927.1									0.000	0.010
W5	DU440765_491.1	1								0.000	0.010
W5	OAR10_16479268.1	3	1							0.005	0.010
W5	OAR10_68517121.1	11	1	5			Y		Y	0.024	0.024
W5	OAR10_85903105.1									0.000	0.010
W5	OAR11_56075682.1	1								0.000	0.010
W5	OAR12_28453815.1	5						Y		0.000	0.010
W5	OAR12_4247318.1	3		1						0.004	0.010
W5	OAR13_25099166.1	5	1	1	Y	Y	Y		Y	0.009	0.010
W5	OAR15_32857143.1	6	1	1		Y				0.009	0.010
W5	OAR15_56018209.1	1	1					Y		0.005	0.010
W5	OAR16_36737603.1	33		20					Y	0.074	0.074
W5	OAR17_23630662.1	28	1	30		Y	Y		Y	0.116	0.116
W5	OAR17_34122964.1	8	3		Y	Y			Y	0.016	0.016
W5	OAR17_6395356.1	8	4	4		Y		Y	Y	0.036	0.036
W5	OAR17_66393977.1	4	1							0.005	0.010
W5	OAR19_39818570.1									0.000	0.010
W5	OAR1_145170621.1	2			Y	Y			Y	0.000	0.010
W5	OAR1_152799556.1	1								0.000	0.010
W5	OAR1_288212148.1	12	1	5						0.024	0.024

W5	OAR1_56099906.1	2	1			Y				0.005	0.010
W5	OAR1_77436037.1	9	1	3						0.016	0.016
W5	OAR23_29207406.1	3	1							0.005	0.010
W5	OAR24_14004421.1	1		1						0.004	0.010
W5	OAR24_15364787.1	2				Y				0.000	0.010
W5	OAR25_14645453.1	1								0.000	0.010
W5	OAR2_10464259.1	4								0.000	0.010
W5	OAR2_141253696.1	4	1							0.005	0.010
W5	OAR2_143337545.1	3	1					Y		0.005	0.010
W5	OAR2_156639058.1	13	2	5						0.029	0.029
W5	OAR2_198392100.1	1								0.000	0.010
W5	OAR3_136187521.1									0.000	0.010
W5	OAR3_34567488.1	6								0.000	0.010
W5	OAR3_35053594.1									0.000	0.010
W5	OAR3_79590268.1	43	3	25		Y	Y		Y	0.108	0.108
W5	OAR4_56051945.1	1		1						0.004	0.010
W5	OAR5_25319096.1	66	9	73	Y	Y	Y		Y	0.318	0.318
W5	OAR5_31695916.1	18				Y	Y		Y	0.000	0.010
W5	OAR5_67883800_X.1	5	1							0.005	0.010
W5	OAR6_103085812.1	2		6		Y	Y			0.022	0.022
W5	OAR6_12170461.1	3	1					Y		0.005	0.010
W5	OAR6_58799540.1	2								0.000	0.010
W5	OAR8_5468228.1	2	1	1						0.009	0.010
W5	s03406.1	28		22		Y				0.081	0.081
W5	s03883.1	3	2							0.010	0.010
W5	s06849.1	160	3	107		Y	Y			0.412	0.412
W5	s07581.1	7							Y	0.000	0.010
W5	s18545.1	3								0.000	0.010
W5	s22150.1	10								0.000	0.010
W5	s25829.1									0.000	0.010
W5	s31573.1	77	5	86		Y	Y		Y	0.345	0.345
W5	s32325.1	1	1							0.005	0.010
W5	s36632.1	1								0.000	0.010
W5	s36667.1	34	4	13	Y	Y	Y			0.069	0.069
W5	s40679.1	3			Y	Y			Y	0.000	0.010

W5	s43103.1	2	1						Y	0.005	0.010
W5	s50474.1	5								0.000	0.010
W5	s53904.1	2								0.000	0.010
W5	s55911.1	5	1	2						0.013	0.013
W5	s56367.1	3							Y	0.000	0.010
W5	s57057.1	125	2	87		Y	Y		Y	0.333	0.333
W5	s66040.1	3								0.000	0.010
W5	s70893.1	6	3			Y				0.016	0.016
W5	s70905.1	1		2						0.007	0.010
W5	s72240.1	1	1							0.005	0.010
W5	s72564.1	2								0.000	0.010
W5	s75165.1	1		1						0.004	0.010
W5	s75550.1	6	1	2		Y				0.013	0.013
W6	OAR10_50587660.1	4	1							0.005	0.010
W6	OAR10_91392930.1	2	1							0.005	0.010
W6	OAR12_64163725.1	6	1	2	Y			Y	Y	0.013	0.013
W6	OAR13_15031682.1	2								0.000	0.010
W6	OAR13_34580622.1	49	1	39			Y			0.150	0.150
W6	OAR13_86825729.1									0.000	0.010
W6	OAR14_16510791.1	1								0.000	0.010
W6	OAR15_26462379.1	1	1							0.005	0.010
W6	OAR15_48956698.1									0.000	0.010
W6	OAR15_6580863.1									0.000	0.010
W6	OAR16_24428159.1	6								0.000	0.010
W6	OAR16_59388986.1									0.000	0.010
W6	OAR18_69288226.1	1								0.000	0.010
W6	OAR1_161217407.1	4	1							0.005	0.010
W6	OAR1_260339881.1	1								0.000	0.010
W6	OAR1_48901255.1	8		7						0.026	0.026
W6	OAR21_26255799.1	13								0.000	0.010
W6	OAR22_6200554.1	12	2	12						0.055	0.055
W6	OAR26_13494685.1	5	1	2					Y	0.013	0.013
W6	OAR2_103451201.1	109	2	61			Y			0.236	0.236
W6	OAR2_236658998.1	2								0.000	0.010
W6	OAR2_37366448.1									0.000	0.010

W6	OAR2_49526327.1	39	2	30		Y	Y		0.122	0.122
W6	OAR2_85440747.1	3							0.000	0.010
W6	OAR3_166113095.1	5	1						0.005	0.010
W6	OAR3_31172991.1	2							0.000	0.010
W6	OAR3_32039232.1	5		1	Y		Y	Y	0.004	0.010
W6	OAR3_58317158.1	2	1						0.005	0.010
W6	OAR3_78678231.1	25	1	9	Y		Y	Y	0.039	0.039
W6	OAR3_80078321.1	3	2						0.010	0.010
W6	OAR4_82342246.1	1	1						0.005	0.010
W6	OAR4_85288470.1	3							0.000	0.010
W6	OAR5_109326228.1	9		5					0.019	0.019
W6	OAR5_109867526.1	15	3	4					0.031	0.031
W6	OAR5_81209579.1	1							0.000	0.010
W6	OAR5_97644679.1	4		2					0.007	0.010
W6	OAR6_29796975.1	1							0.000	0.010
W6	OAR6_35078443.1								0.000	0.010
W6	OAR6_50590550.1	2	1						0.005	0.010
W6	OAR6_61182493.1								0.000	0.010
W6	OAR7_101493967.1								0.000	0.010
W6	OAR7_51517611.1			1					0.004	0.010
W6	OAR8_55769107.1	4		1					0.004	0.010
W6	OAR9_90078978.1	1	1						0.005	0.010
W6	s07823.1	15		14					0.052	0.052
W6	s11161.1	1	1						0.005	0.010
W6	s15703.1								0.000	0.010
W6	s15823.1	9	3	1					0.019	0.019
W6	s25104.1	1							0.000	0.010
W6	s26771.1	3		1					0.004	0.010
W6	s30254.1	1							0.000	0.010
W6	s30348.1								0.000	0.010
W6	s33920.1	1	1						0.005	0.010
W6	s37320.1	25	2	3				Y	0.022	0.022
W6	s42171.1	2	1						0.005	0.010
W6	s43800.1	152	2	129		Y	Y	Y	0.488	0.488
W6	s44992.1								0.000	0.010

W6	s47975.1	18	3	10		Y	Y			0.053	0.053
W6	s52578.1	3								0.000	0.010
W6	s54511.1	1								0.000	0.010
W6	s56532.1	4	1							0.005	0.010
W6	s60447.1	1								0.000	0.010
W6	s66880.1	1								0.000	0.010

Quality Control Test 1: Number of mismatches for trios.

Quality Control Test 2: The number of trio mismatches where the mismatch is not due to opposing double-heterozygosity parent-offspring genotypes.

Quality Control Test 3: The number of mismatches for double homozygote parent-progeny.

Quality Control Test 4: GeneSeek Genotyping: low call rate (too few genotypes passed QC)

Quality Control Test 5: High Quality DNA - Fail due to too many mismatches.

Quality Control Test 6: Low Quality DNA - Fail due to too many mismatches.

Quality Control Test 7: High Quality DNA - Too Few Genotypes Passed.

Quality Control Test 8: Low Quality DNA - Too Few Genotypes Passed.

Error Rate 1: Trio parent-progeny error rate

Error Rate 2: Adjusted trio parent-progeny error rate

Notes:

SNPs with > 14 errors for Quality Control Test 1 (Number of mismatches for trios) are considered to exhibit too many errors to be reliable.

Quality control tests 1 and 3 are based on using a set of parentage and trio relationships to look at SNP mismatches. Number of mismatches for trios is adjusted for duplicate parents and progeny genotype (i.e. if parents are represented multiple times for more than 1 progeny with the same progeny genotype then they are only counted once - note this slightly overcounts the trio mismatches due to sires).

Number of mismatches for double homozygote parent-progeny (NUM_HOMO_PC_MM) relates to the number of double homozygote mismatches (e.g. parent AA, progeny BB) adjusted for duplicate parents (parents only counted once if represented more than once).

Number of trio mismatches where mismatch is not due to opposing double-heterozygote parent-progeny (NUM_TRIO_MM_PROG_HET) is the number of trio mismatches that are not due to parent-child double homozygote mismatches adjusted so that each pair of parents is represented only once

HQ_MM_FAIL Australian Sequenom high quality data set - at least 7 mismatches vs SNP50 data (given there were not many samples to compare this is a very high threshold).

LQ_MM_FAIL Australian Sequenom low quality data set - versus duplicate samples

HQ_TOO_FEW_PASS Australian Sequenom high quality data set

LQ_TOO_FEW_PASS Australian Sequenom low quality data set

The formula used for Trio parent-progeny error rate (TRIO_PC_ERROR_RATE) was 2*((NUM_HOMO_PC_MM/540) + (NUM_TRIO_MM_PROG_HET/381))

where NUM_HOMO_PC_MM is the number of double homozygote mismatches adjusted so that each parent is only counted once; 540 is the number of unique parents; NUM_TRIO_MM_PROG_HET is the number of trio mismatches that are not due to parent-child double homozygote mismatches adjusted so that each pair of parents is represented only once; 383 is the number of unique parent-pairs. The rationale for the 2 is that only a subset of the errors is able to be detected.

The formula used for Adjusted trio parent-progeny error rate (ADJ_ERROR_RATE) is

if (TRIO_PC_ERROR_RATE< .01) then ADJ_ERROR_RATE = .01

else if (TRIO_PC_ERROR_RATE > 0.5) then ADJ_ERROR_RATE =.5

else ADJ_ERROR_RATE = TRIO_PC_ERROR_RATE)

The rationale for the .01 is that this is a reasonable minimum rate for Sequenom. Would need to do hundreds more duplicate tests to get a more accurate error estimate. There is much more duplicate information for the Australian Sequenom samples than for the GeneSeek samples - but the SNP the sets of SNPs for which there are high levels of Australian Sequenom and GeneSeek errors are not the same although there is some overlap. The reason for capping the error rate at .5 is that it is meaningless to have a rate greater than .5 which is equivalent to tossing a coin.

Dr Jill Maddox performed this analysis, and can be contacted for additional explanation if required**.**

8 Bibliography

Becker D, Tetens J, Brunner A, Burstel D, Ganter M, Kijas J for the ISGC and Drogemueller, C. (2010) Microphthalmia in Texel sheep is Associated with a Missense Mutation in the Paired-Like Homeodomain 3 (*PITX3*) Gene. *PLoS One* **5**: e8689

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C, Georges M. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics* **38(7):**813-8.

Kalinowski ST, Taper ML, Marshall TC. (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16(5):**1099-106.

Kalinowski ST, Taper ML, Marshall TC. (2010) Erratum: Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **19(7):**1512.

Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, Wilson P, Ingersoll RG, McCulloch R, McWilliam S, Tang D, McEwan J, Cockett N, Oddy VH, Nicholas FW, Raadsma H; International Sheep Genomics Consortium. (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One* **4(3):**e4668.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, Paiva S, Barendse W, Ciani E, Raadsma H, McEwan J, Dalrymple B; International Sheep Genomics Consortium Members. (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology* **10(2):**e1001258.

Marshall TC, Slate J, Kruuk LE, Pemberton JM. (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7(5):**639-55.

Meadows J, Hawken R and Kijas J (2004) Nucleotide diversity on the Ovine Y chromosome. *Animal Genetics* **35(5):** 379-385.

Johnston SE, McEwan JC, Pickering N, Kijas JW, Beraldi D, Pilkington J, Pemberton J, Slate J. (2011) GWAS identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Molecular Ecology* **(12):**2555-66.

Mömke S, Kerkmann A, Wöhlke A, Ostmeier M, Hewicker-Trautwein, Ganter M, Kijas J for the International Sheep Consortium, Distl O. (2011) A frameshift mutation within LAMC2 is responsible for Herlitz type junctional epidermolysis bullosa (HJEB) in Black Headed Mutton sheep. *PLoS ONE* **6(5):**e18943.

Pfeiffer I & Brenig B. (2005) X- and Y-chromosome specific variants of the amelogenin gene allow sex determination in sheep (*Ovis aries*) and European red deer (*Cervus elaphus*). *BMC Genetics* **6**:16