**Sibling statistics and probabilities**

**Difference between relationships between parents-progeny and between siblings**

Two siblings share their parents’ genes. With available DNA SSR we can see how they share the alleles at least those of non-functional DNA selected by marker-pairs.

Lets limit ourselves to diploids. Take as an example Emneth Early and Lord Suffield; both are progeny of Keswick Codlin and Hawthornden.

With marker-pair CH04c07, both parents have different alleles, respectively {96,106} and {94,108}. These can be described as having two pairs with unique numbers. Both progeny inherit the same two alleles: {106,108}. That is 106 from Keswick Codlin and 108 from Hawthornden. Neither 94, 96 are passed-on.

That Emneth Early and Lord Suffield end up have the same two alleles might indicate at least for this marker-pair that a parent-progeny relationship could be satisfied.

With marker-pair CH02c11, alleles are again different. Again both progeny inherit one allele from each parent. Emneth Early inherits {217,235}, and Lord Suffield {213,233}. All four alleles are passed to progeny two to one and two to the other.

Now Emneth Early and Lord Suffield have no alleles in common for this marker-pair, showing they do not have a parent-progeny relationship.

**Examples**

Looking at a few examples quickly shows how important it is to ground thinking in reality. sometimes it reveals that siblings were bred with little or no apparent selection (at least among the SSR fragments sampled) and sometimes there does appear a clear bias with siblings showing a non-random inheritance of their parents’ alleles in the marker-pair defined regions of non-coding regions.

**Reinette Franche x Reinette des Carmes**

SSR suggests that the old French variety Reinette Franche may be parent to perhaps 100 varieties, and the old French (Belgian or German) variety Reinette des Carmes to perhaps 20-30. As a pair, these may be parents to 12 varieties, five of which are confirmed from SNP.

**Part of the family tree of two old French varieties, Reinette Franche and Reinette des Carmes**



These are listed in the table below in chronological order of ‘introduction’.

It appears that breeding of this group of siblings began in Europe during C17, and accelerated during C18 and C19 with progeny raised south of London, in Norfolk and followed by several in Herefordshire and Gloucestershire. These siblings form an interesting group because it is quite large (nine diploids and three triploids) and there is no evidence to suggest these weren’t bred (or found) by unconnected individuals and that there was no systematic or consistent selection of attributes beyond the implicit desire to have an appealing fruit.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **first date** | **origin** | **comments from NFC, NAR, Herfs Pom, Hogg** |
| Reinette Franche \* | 1500s | Normandy? | This is a very old French apple, |
| Reinette des Carmes $ | <1667 | France, Belgium or Germany? | first described be =Merlet 1667 |
| Golden Reinette ! | <1650s | Europe | Thought to have originated in Europe. It has been known in England since the mid 1600s. |
| Mannington's Pearmain | 1770 | Uckfield, Sussex | Grown from seed found in cider pomace in the garden of Mr Turley, Uckfield, Sussex, England in about 1770. It was introduced by John Mannington in 1847. |
| Hubbard's Pearmain | <1800 | Norfolk | Originated in Norfolk, England. It was known before 1800. |
| Reinette d'Anjou ! | <1817 | Belgian or German | Thought to be of either Belgian or German origin. First mentioned in 1817. |
| Hunt's Duke of Gloucester | 1820 | Gloucester | Raised by Dr Fry at Gloucester. It was introduced in 1820 by Thomas Hunt of Stratford-on-Avon. |
| Claygate Pearmain ! | 1821 | Claygate, Surrey | Discovered by John Braddick at Claygate, Surrey and exhibited to the Horticultural Society in 1821. |
| Adams's Pearmain | <1826 | Hereford | Brought to notice in 1826 and introduced and exhibited by R. Adams from Herefordshire. |
| Gipsy King | <1876 | UK | first catalogued in River's catalogue |
| Mabbott's Pearmain ! | <1883 | Kent | Raised in Kent, England. It was first described in 1883. Introduced by Lewis Killick of Langley. |
| Pig's Nose Pippin | <1884 | Hereford | Thought to have originated in Hereford. Described in 1884. |
| Hoyaske Guldreinette ! | <1950 | England? Sweden? | As NFC unknown 1974-381, matches trees in Sweden and Denmark; Bernwode & RVRogers claim it is Red Ingestrie, disputed by MAN. |
| A724 | <1950? | Wick Gloucester |  |

\*Reinette Franche has synonym in NFC Herceg Batthyanyi Alma

$ Reinette des Carmes has synonym in Czech collection Karmelitská reneta

! Confirmed parentage with SNP (Muranty et al., Howard et al.)

£ Golden Reinette also known as Reinette de Hollande



\* Reinette Franche has synonym in NFC Herceg Batthyanyi Alma

$ Reinette des Carmes has synonym in Czech collection Karmelitská reneta

& triploid with Reinette Franche as diploid gamete donor (may be the mother)

% triploid with Reinette des Carmes as diploid gamete donor (may be the mother)

The tables above show the SSR of the two parents and twelve progeny, with the tree triploid varieties listed at the bottom. Cells with different alleles have been highlighted in colours for showing inheritance of these. For each of the progeny, each of the twelve marker-pairs are derived from inheriting an allele one from each of the parents (with two from one parent for the triploid varieties).

For the marker-pair CH04e05, both parents have the same alleles, 173 and 200. Among the diploid varieties the probability of inheriting them should be 50% if random. Looking at the summary below, the results are 44% for 173 and 56% for 200. Given the relatively small sample size, it appears close to random. Inspection of the other eleven marker-pairs also show a near random distribution of inherited alleles.



While diploid progeny inherit one allele from each of their parents, it is possible that siblings have no alleles in common, at least for one or more market-pairs. For instance, consider the marker-pair CH01h01 with the parents having {115,129,0,0} and {111,119,0,0}. Hubbard’s Pearmain has {119,129,0,0} and Mabbott’s Pearmain {111,115,0,0}. This is quite a common situation; siblings do not have to have all marker-pairs matching. A simple test can be conceived for two varieties being siblings in which the number of alleles in common and the number of marker-pairs with at least an allele in common, are counted. Is it useful?

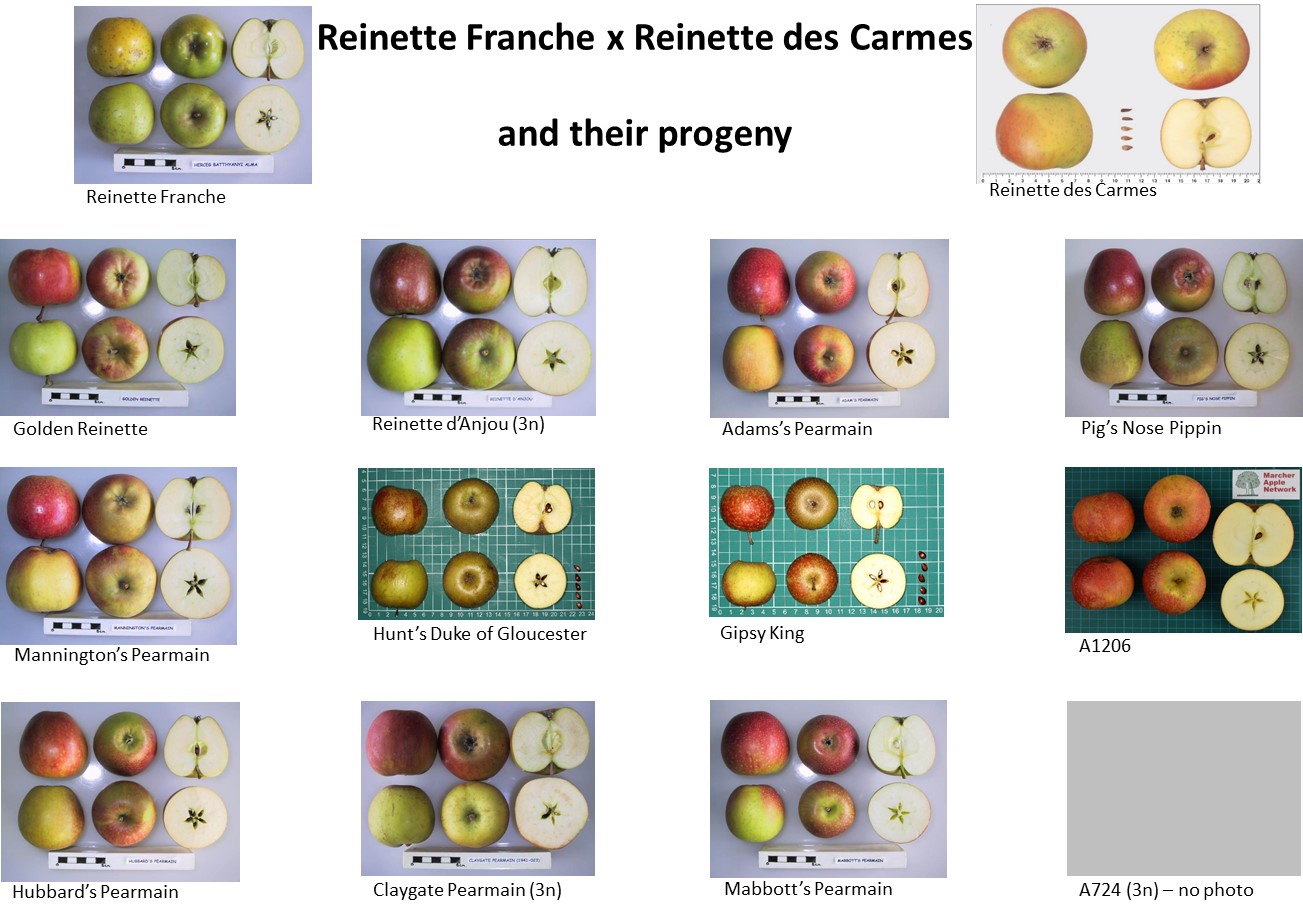
Carrying out this on the 36 pairs of diploids considered above results in the matrix grid as shown below. The numbers in cells refer to the number of alleles in common between varieties in that row and column. The cell colour reflects how many marker-pairs have at least a common allele, green is all twelve, yellow is eleven, beige is ten, pink is nine and red is eight.

‘Perfect’ matching means all 24 are in common. Both parents have all twelve marker-pairs matched with progeny and are highlighted green. They have between 12 and 17 alleles in common with progeny, though there are only seven in common between Reinette Franche x Reinette des Carmes. Between siblings the number of alleles in common varies from 10 to 18, with an average of 14.83 and standard deviation of 2.02. In principle siblings could have two completely different sets of alleles; in practice this is statistically of essentially zero likelihood.

The seven alleles in common between these diploid parents maybe inherited. Random inheritance of the remaining 17 is to be expected, of which on average half will be inherited by both siblings. Thus the expected number of alleles in common is (7+17/2) = 15.5. Actual number is quite close, well within a standard deviation of the expectation.

These siblings do have quite a number of morphological similarities as may be seen in the montage below.

They are likely to have been bred, and effectively selected, with relatively little bias. What happens when siblings are bred more selectively by the same or similar folk? Furthermore what happens to the statistics, and ability to spot siblings, when their number is fewer than in this rather large set?



**Cox’s Orange Pippin x Jonathan**

There are eight progeny all raised at the Horticultural Laboratory in Wageningen about 1935; details are shown below. Lucullus and Prins Bernhard are DNA matches, though the morphological descriptions in NAR suggest significant difference, but with the same date of introduction, 1935, this is inconsistent with one being a sport.



Monte Carlo modelling of the marker-pair outcomes was made in which it was taken that both alleles of each parent have an equal probability of being passed to progeny, there is no correlation modelled between the outcome for marker-pairs. This may well not be a realistic assumption



Marker Pair CH01f03b is inevitably going to be matched by any model. Four other marker-pairs had a modelled outcome with a probability of occurrence of 10% or more. The remaining seven marker-pairs had a probability of occurrence roughly of 1%, in several cases this can be traced to the sibling set having zero or only one instance of a given allele pair. If marker-pair outcomes were random and uncorrelated with each other, the total probability of finding these eight siblings with these distributions would be the product of all twelve marker-pair probabilities, or about 10-18. It suggests that some non-functional DNA strands are passed on in a concerted manner, with several marker-pairs likely being inherited as a bundle.

Assuming a breeder may carry out perhaps 10000 trials of progeny from a pair of parents, the set of eight siblings will have a probability of occurrence roughly at one part in a thousand. And this is indeed about an order of magnitude less than the lowest marker-pair probability.

**Cox’s Orange Pippin x Worcester Pearmain**

The SSR study of parents and progeny identified seven progeny from these parents; they were raised by several different groups though in the SE of England about 1930. There fingerprints are shown below.



For the most part transfer of alleles in each marker-pair is broadly consistent with statistical expectations.



For each market pair, there are 4 parental alleles, so these should be present on average 3.5 times in the seven progeny, if they are inherited randomly; often four are present and usually there is at least 2 of any unique value. Two exceptions are striking. No progeny has Worcester Pearmain allele 110 in marker-pair CH04c07, and in GD147 neither Cox’s Orange Pippin 139 nor Worcester Pearmain 137 alleles are passed to any of the progeny. Both these observations amplify concerns raised above and add further thoughts. Here, for some marker-pairs, the alleles do not appear in progeny on anything close to a random statistical basis.

**Hawthornden x Keswick Codlin**

A third example of a smaller set of four progeny derived fromHawthornden x Keswick Codlin show a similar bias in a couple of marker-pairs even though these were raised by four separate groups with an elapsed time of about 70 years. For marker-pair CH04c07, the allele 96 in Keswick Codlin is not passed to any of the four progeny. Similarly in marker-pair CH01f02 neither Hawthornden 182 nor Keswick Codlin 170 are passed on. For GD12, Keswick Codlin 153 is not passed on.





There is a low probability of these marker-pair combinations arising as a result of random combinations of uncorrelated scrambling of parental DNA. Most striking is in CH01f02. It again suggests that other factors are affecting inheritance of specific parts of the non-coding genome.

Gold Medal maybe a fifth sibling. It hasn’t been included here as there is a (low) possibility that its parents are Keswick Codlin x Queen.

**Beauty of Bath x Worcester Pearmain**

A similar picture emerges for this family with siblings raised by by unconnected folk around the SE of England from 1906 to about 1964.

Three marker pairs illustrate cases where one or more alleles are dominant and some suppressed, though not as strikingly as in the previous examples. CH01f03b has {136,176} dominant with just one occurrence of 178. Similarly in CH02d08 {228,250} is dominant with only two occurrences of 210 and 224 is suppressed. Finally in CH02c09, 256 dominates and 242 suppressed.



**Suggested learning from Examples**

Comparing the occurrences found in modelling these four examples, there is not obvious trend, generally each marker-pair has instances of high (>15%) and low (<1.5%) occurrences. There is tentative evidence that occurrences are high or low in the four marker-pairs CH01h01, Hi02c07, CH04e05 and CH02c09, while both CH04c07 and GD147 are either high or low. The most striking result is with the Cox’s Orange Pippin x Worcester sibling set where three marker-pairs seemingly are much more prevalent than random statistics would suggest they should be. The sibling sets are small and it is prudent not to over-interpret these differences. Yet the distinct impression remains that Breeder choice selects not just genes but also markers pairs, which means there is a linkage on some chromosomes between a gene and its associated marker pair (situated in a nearby non-functional region). With twelve marker-pairs, i.e. 24 (or 36 or 48) alleles, selecting some non-functional DNA disfavours other parts. I am not aware of an investigation of this effect[[1]](#footnote-1). However, it is clear that using frequency of marker-pair occurrences as criteria for testing whether two or more varieties are siblings is unlikely to be reliable.

**Counting alleles in common**

An alternative approach for assessing whether varieties are actually siblings is to count the number of alleles that they have in common. It isn’t potentially as clear as marker-pair statistics could have been, but it holds some promise, even if some selective breeding has occurred.

**Cox’s Orange Pippin x Jonathan**

There are eight siblings, though there is a matching pair, Lucullus and Prins Bernhard.

The cross correlation matrix is revealing of the extremes that may be encountered. In this the number of alleles in common can be counted by comparing two fingerprints as shown in the extracts given above (note a zero in the PK2 position means that the first value in PK1 is to be duplicated). Parents Cox’s Orange Pippin and Jonathan have 10, meaning that there are only 14 independent alleles that may be passed on. The average number between siblings given in the cross-correlation table is 16.96, close to the expected mean of 10+14/2 = 17; standard deviation of (different) sibling alleles in common is 2.67. However, the extreme cases stands out as a warning: Prinses Irene and Prinses Margriet have just 12 alleles in common while Lucullus and Prins Bernhard have double that number, both of these are at roughly at or below the 5% occurrence level. In the event that in the latter pair one is sport of the other, the mean and standard deviation are 16.57 and 2.21.

**Cox’s Orange Pippin x Worcester Pearmain**

There are seven siblings. In the cross-correlation table for Cox’s Orange Pippin x Worcester Pearmain have 8 alleles in common, so 16 in both parents can be passed to progeny. Numbers between siblings is in the range 13-20, with a mean of 16.62, close to the random statistical number of 8+(24-8)/2=16. For the 21 sibling alleles in common, the standard deviation is 1.91.

This simple randomness test does not differentiate between which of the two alleles from a given parent are passed to progeny, and we have seen how for two marker-pairs some alleles are not passed to any of the seven progeny.

**Hawthornden x Keswick Codlin**

There are four siblings in the Hawthornden x Keswick Codlin family.

Parents have seven alleles in common. On that basis, it might be expected siblings will have 15.5 alleles in common, and that is the exact observation from the cross correlation table. Standard deviation is 1.87. One pair, Lord Grosvenor and Lord Suffield have only 12 common alleles at roughly the 5% confidence level.



**Beauty of Bath x Worcester Pearmain**

These two have only six alleles in common.

The six cross-correlations between siblings have an average of 16.5 alleles in common, whereas the expectation is 15. Standard deviation is 1.8. Laxton’s Herald and Folkestone only share twelve alleles.

**Summary from Examples**

A summary of the alleles in common between siblings is given below, rounded to one decimal place.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | alleles in common | | | | |
| Parents | between parents | expectation between siblings | observed between siblings | observed standard deviation | lower bound at 98% confidence |
| Reinette Franche x  Reinette des Carmes | 7 | 15.5 | 14.8 | 2.0 | 10.8 |
| Cox’s Orange Pippin x Jonathan  or excluding Lucullus | 10 | 17 | 17.0  16.6 | 2.7  2.2 | 11.6  12.2 |
| Cox’s Orange Pippin x  Worcester Pearmain | 8 | 16 | 16.6 | 1.9 | 12.8 |
| Hawthornden x Keswick Codlin | 7 | 15.5 | 15.5 | 1.8 | 11.9 |
| Beauty of Bath x  Worcester Pearmain | 6 | 15 | 16.5 | 1.8 | 12.7 |

**Inbreeding – a caution**

Popular varieties such as Reinette Franche, Golden Reinette, Cox’s Orange Pippin, Jonathan, Golden Delicious, McIntosh, and Alexander, have been widely used for breeding, with scores of cultivars produced from each, and sometimes crossed repeatedly (Bannier, 2010). Muranty et al. noted several examples of varieties bred from parents that themselves shared a parent. It results in the progeny have 50% of the genome of its grand-parent, the same fraction expected of a full-sibling. Hood’s Supreme is progeny of James Grieve x Charles Ross, both of which have Cox’s Orange Pippin as a parent. Hood’s Supreme has 14 alleles in common and nine marker pairs with a matching allele, just about sufficient to suspect a sibling relationship.



A more extreme example is Fairie Queen. It is progeny of Cox’s Orange Pippin and James Grieve. James Grieve is progeny of Cox’s Orange Pippin and Pott’s Seedling. Consequently, the genome of Fairie Queen contains 75% of that from Cox’s Orange Pippin. All twelve marker-pairs have a matching allele and there are 17 alleles in common. Sufficient to suspect a full-sibling relationship. It isn’t that, but they are actually closer genetically.



**Rough rules for spotting possible sibling pairs from DNA SSR**

In summary, counting numbers of alleles in common between samples or sets of samples does enable connections to be identified with some degree of confidence. Inheritance of which of the two alleles from each parent appears to be random. But, as the SSR fragment are fairly close to genes that may be under breeder selection, some bias can occur.

Bannier (2011) highlighted some effects from successive breeding from a few popular varieties. During this process certain alleles will become (a little) more prevalent. When seeking siblings, if they are connected with these popular varieties, a slightly higher number of alleles in common should be anticipated.

As noted above the most likely (mode) number that siblings will have in common is the sum of the number common between parents and one-half the remainder. With, say, eight shared and 16 unshared alleles between parents, the mode in distribution of alleles common between siblings is 16. With a standard deviation of typically 2, only about 1 in 20 of samples will depart from the mean by more than 4, giving a probable range from 12 to 20 alleles in common. Only 1 in a thousand occurrences are likely outside the range 10 to 22. The table below extends this to other cases.

|  |  |  |  |
| --- | --- | --- | --- |
| **numbers of alleles in common** | | | |
| shared = numbers common between parents | unshared =  not shared between parents | mode of distribution in numbers between siblings | approx. lower limit on likely number between siblings |
| 4 | 20 | 14 | 10 |
| 8 | 16 | 16 | 12 |
| 12 | 12 | 18 | 14 |
| 16 | 8 | 20 | 16 |
| 20 | 4 | 22 | 18 |

This leads three rough rules:

1. Overall, unless parents have few alleles in common, siblings are likely to have half, or more, of their alleles in common.
2. Siblings do not need to have at least one allele in common for all twelve marker-pairs. On the contrary, one or two marker-pairs without alleles in common may be expected and supports that they are more likely siblings than parent-progeny. Nine, exceptionally eight, is likely a minimum.
3. If a popular variety is thought to be involved in parentage of the suspect siblings, inferring a sibling relationship via SSR data should be treated with greater caution, with parentages checked.

Once groups of possible siblings have been screened against the rough rules, other information can be invaluable. If one or both of the parents are known or suspected, checking the consistency of allele inheritance greatly increases confidence in assigning the varieties as siblings. Similarity of morphology and provenance may also lend some assurance.

**References**

Muriel Smith, ‘National Apple Register’ (1971).

Considine MJ, Wan Y, D’Antuono MF, Zhou Q, Han M, et al. (2012) Molecular Genetic Features of Polyploidization and Aneuploidization Reveal Unique

Patterns for Genome Duplication in Diploid Malus. PLoS ONE 7(1): e29449. <https://doi:10.1371/journal.pone.0029449>

Hélène Muranty, Caroline Denancé, Laurence Feugey, Jean-Luc Crépin, Yves Barbier, Stefano Tartarini, Matthew Ordidge, Michela Troggio, Marc Lateur, Hilde Nybom, Frantisek Paprstein, François Laurens and Charles-Eric Durel, Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm, BMC Plant Biology (2020) 20:2

<https://doi.org/10.1186/s12870-019-2171-6>

Ordidge M, Kirdwichai P, Baksh MF, Venison EP, Gibbings JG, Dunwell JM.

Genetic analysis of a major international collection of cultivated apple

varieties reveals previously unknown historic heteroploid and inbred

relationships. PLoS One. 2018;13(9):e0202405.

Howard, N. P., Micheletti, D., Luby, J. J., Durel, C.-E., Denancé, C., Muranty, H., Ordidge, M., & Albach, D. C. (2022). Pedigree reconstruction for triploid apple cultivars using single nucleotide polymorphism array data. Plants, People, Planet, 1–14. https://doi.org/10.1002/ppp3.

10313

H-J Bannier, Erwerbs-Obstbau Volume 52, 2011, Numbers 3-4, 85-110, DOI: 10.1007/s10341-010-0113-4

1. Dr Matthew Ordidge has commented on an draft of this note (7Sep20):

   “I have to admit that I get a little lost in the discussion of potential sibling similarities (given the potential uncertainties) and, to a large extent, this was why we pulled short of getting involved in our PLOS paper. But I think your observation that a number of the alleles appear to be under selection is interesting.

   ….

   I presume that you are correct in that this deviation from random potentially suggests an element of advantage correlated with some of the marker alleles and this would presumably indicate that there is some functional genetic element that is linked to them (in a form that can be selected, wittingly or unwittingly). I think the apple markers were generally selected to be neutral (as far as anybody knew at the time) in order to allow them to more randomly distribute – but we have recently included two markers with known linkage to fruit size and flesh colour in the cherry marker set. This is different to the absolute location of the marker itself – which is normally in a ‘non-functional’ region (such that the mutations that create the size range itself are not under selection). By linkage – I mean they are physically close on their chromosome to a gene for the trait under selection such that when crossing over occurs (which it does – allowing further reassortment of the contents of each chromosome) they are most likely to remain together in the combination of marker size and allele variant of the linked gene that was inherited from the parent (i.e. they are close enough that the chance of a crossover between them is low).

   But, the accuracy of this using such a small number of SSR markers to investigate this will always remain relatively low. And I have a feeling that using it for the identification of siblings will retain a reasonable amount of uncertainty.

   By comparison – I think Nick Howard (who was carrying out a postdoc project in Germany and has just taken a post as a breeder in Wageningen) has been doing some very powerful analysis using the 20K SNP chips – which has involved trying to trace specific blocks of chromosome through generations and cross over events to follow the segregation between parents and offspring. He has three papers in progress (I haven’t seen the first two – but I’ve been in discussion with him on the third – because it relates to the parentage of triploids that we looked at in our PLOS publication). I’m sure there are others doing similar.” [↑](#footnote-ref-1)