**DNA Fingerprinting update – towards identifying parents**

Based upon the results DNA 2016-22 campaigns, much guidance from folk in other orchard groups, and recent publications on SNP analysis there is a growing confidence in the ability of DNA SSR to confirm or refute identification by matching to samples previously taken. Some mistakes have been made along the road, particularly failing to recognise that triploid varieties generally appear to have non-viable pollen and seeds, and thus are unlikely to be parents. ‘The jury is still out in this’, hence triploids have now been excluded as parents. Increasingly confidence is growing that it is possible to establish or refute parentage. Interim reports were touched upon in a recent MAN News Sheet and Newsletter. Now we’re beginning to see further benefits as other groups submit more and more samples resulting in matches of our varieties to theirs, and this then demonstrates these aren’t chance seedlings but rather cultivars and sometimes gives suggested identities. Additionally, there has been an effort to see whether parentage can be extracted and what it may reveal. We’ll address examples of these in the following article.

First, though, let’s make a brief recap of how our samples are fingerprinted and what it looks like; it is essentially the same for pears and cherries.

**Fingerprinting**

The DNA fingerprinting we’ve been using is relatively cheap (about £30 a sample) so has to be simple yet able to distinguish between many varieties. A double helix has about 1 billion units, or nucleotides, along its length. The method analyses a contiguous section about one millionth of this, in a part that does not make proteins. This bit, typically 80% of the whole helix, is called the intron (it has been called ‘junk’, because we had no idea what it does); a better name for these sections is non-coding regions. As it isn’t involved with inherited characteristics (that bit is in the genes), it isn’t under evolutionary pressure and these regions are thought to mutate less quickly. This is useful as it is makes the method more likely stable over generations.

DNA is usually extract from leaves (though stipules, buds, stems and fruit can also be used) by grinding and digesting them with mild chemicals. The quantity recovered is tiny, far too little to analyse directly; of this only about a millionth part is of interest. A procedure known as amplification using polymerase chain reaction (PCR) is applied, very similar to that being used for corona virus testing.

DNA comprises four nucleotides known usually by their bases as A,C,G,T; and the two chains of the helix are bound together by A linking to T and C to G. The sequence in which these four types appear determines “the world and everything” of the apple. Within the non-coding regions there are many short sequences of a dozen or more nucleotides common to all apples. The finger printing method involves adding two synthesized chemicals (the marker-pair) that cut out a section of the non-coding region which has characteristic sequences at both ends. This tiny fragment is typically 100-250 nucleotides long, and must then be amplified in quantity by PCR roughly one billion-fold to make it measurable. As such the method would not distinguish between one variety and another but for a fortunate fact which can be exploited. Notwithstanding the comments above, over time when the DNA has been reproduced some errors in copying occur, and those in the non-coding region have no evolutionary pressure to be corrected. The adroitness of this method is that different apple varieties have ‘evolved’ with different numbers of nucleotides present between the two end members defined by our marker-pairs. It may be that during replication 40 or 41 or 42 … repeats of simple sequences such as .TA. or .GCA. etc. occurred. Then length of these fragments or alleles varies from one variety to another and it is length we measure and report as so many base pairs (i.e. nucleotides). We measures the length of these Simple Sequence Repeat (SSR).

No doing this with just one marker-pair would not unambiguously differentiate between the thousands of varieties of apples. Twelve different marker-pairs are used as the standard in the European Cooperative Programme for Plant Genetic Resources. In Europe 16 marker-pairs are used, in US 19. These marker-pairs each select a strand of about 100-250 nucleotides.

For most apples, there are two copies of each of the 17 chromosomes. These are the diploid varieties. About 10-20% of apples in curated collections are triploids; they have three copies of the chromosomes, 17 from one parents and 34 from the other. And about 1% of apple are tetraploid with four sets of chromosomes. When the DNA fingerprint is analysed for diploids, as there is a pair of chromosomes which may have a different length for the section between the markers, i.e the allele, then there will be two possible lengths for each marker pair. For triploids there are three possible lengths, and for tetraploids four. To add confusion, for a given marker pair the lengths of alleles may be the same for some or all the chromosomes. Severn Bank, a tetraploid, has a finger print 114,118,120,129 for marker-pair CH04c07. By convention regardless of ploidy, four numbers are carried for each marker pair; with triploids there is no fourth chromosome and the length is zero for all varieties, e.g. Bridstow Wasp has 96,106,120,0. For diploids the third and fourth digits are zero by convention, e.g. Cox’s Orange Pippin 106,112,0,0. And if the alleles from chromosomes are the same, then that length is only listed once, e.g. Braddick Nonpareil is by convention 106,0,0,0, though actually 106,106,106.

Now let’s look at real examples in the form you may already have seen. Three of the varieties above are shown in the figure, for each variety there are 12x4 columns of numbers, the twelve marker-pairs being shown having successively three repeats of the colours blue, green, yellow, red, each divided into four columns with numbers corresponding to the allele lengths or with zero if duplicated. Title in the top row are the name of the marker-pair with and appendage ‘\_PKx’ indicating the four peaks of the alleles. These 48 numbers make up the fingerprint; usually at least half are zeros. Despite those zeros, as each allele has typically at least six common values, there are about 6 raised to the power of 24 possible combinations, that’s more than a million million million. That’s the basis of the methods capacity to distinguish between varieties. Further details are given on our website <http://www.marcherapple.net/research/dna-analysis/>

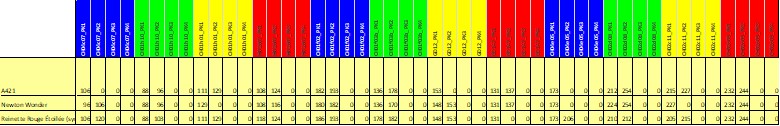


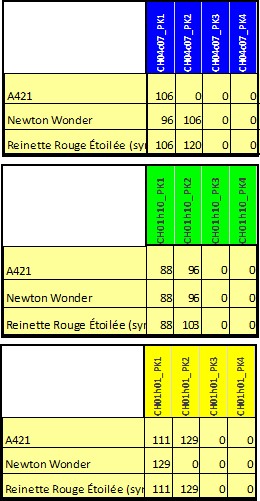
When comparing two fingerprints, we may conclude that they are the same variety if all values (really the non-zero values) are the same. Occasionally small experimental errors/glitches can occur with one or perhaps two alleles having a measured value 1, 2, 3 or 4 base pairs more or less than is correct. And experts are essential for confirming if a difference is experimental or ‘significant’.

**Investigating Parentage**

Now let’s look at and a MAN example in some detail, Afal Cas Gwent (ACG) which has the DNA sample number A421. In 2016, after the first DNA campaign, it was reported that nothing matched this. It was unique. Three years later it is still unique. May be it’s a seedling… of what?

Let’s compare it with the fingerprints of Newton Wonder (NW) and Reinette Rouge d’Etoilee (RRdE), both diploids, which I’m hoping to convince you are its parents.





Take each marker-pair in turn. The first marker-pair CH04c07 shows A421 has two alleles of 106 (recall the convention), and these can come from either of the two other varieties. They also could have passed on 96 or 120, respectively. But in this case they didn’t.

Now the second marker-pair, CH01h10. In this case both the possible parents have 88 in one of their chromosomes, and in the other 96 and 103, respectively. If NW passes the chromosome with 96 and RRdE passes 88, then ACG is OK.

The third marker-pair is CH01h01. A similar situation arises with both possible parents having 129 in a chromosome, but because only RRdE can pass on 111 to ACG, then it is NW that passed the 129.



And finally let’s consider the fourth marker-pair, Hi02c07. It’s clear, NW has to pass on 108 and RRdE 124. Neither the 116 nor 118 are required in ACG.

The same process can then be applied to the other eight marker-pairs. ACG fingerprint could arise then from a mix of NW and RRdE, one or other of their chromosomes being passed to ACG.

Using the 5000 unique fingerprints now in the fruitID dataset, which includes all from National Fruit Collection (NFC), Irish Seed Saver Association (ISSA), Tamar Valley (TVm), MAN, Gloucestershire Orchard trust (GOT), and many continental European ones via INRAE etc., there is no other combination of two varieties that can do this. Final questions to ask:

* Morphology: Does the progeny look a bit like both possible parents? When I suggested this relationship to Mike Porter he immediately said “Yes”.
* Provenance: Were both the parents in existence (well) and with a reasonably high probability of co-location somewhere before the progeny was introduced? NW is from 1887 and RRdE 1830, both are popular cultivars. If ACG is a seedling this is all consistent (or at very least not inconsistent); if we start to find it occurring as very old trees in many locations, we’ll have to think again.
* Ploidy: it is well known that triploids produce little viable pollen, and seeds from the apples of triploids are often weak. Both parents being triploid is improbable, even one triploid parent is very much less likely than their statistical presence may suggest. A recent study using SNP by Nick Howard et al. (2022) of hundreds of triploid varieties found no evidence that any were parents of other varieties.

Now you’re well ready to look at a few examples of identification arising from the DNA 2019 campaign and at parentage, including comparison with those from a recent whole genome study and the National Apple Register.

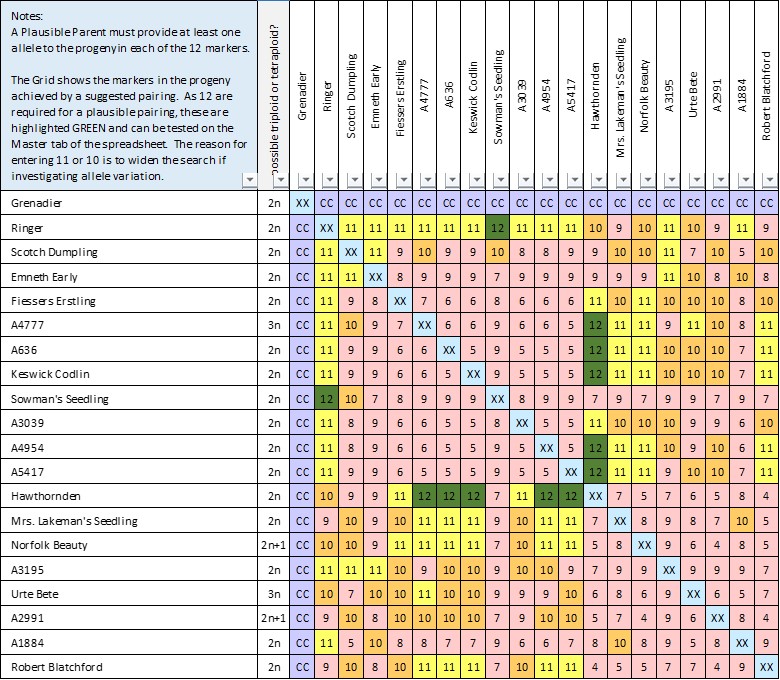
There are two datasets for testing how useful is the parentage from DNA SSR. A paper by Hélène Muranty et al. has a punchy title ‘Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm’, BMC Plant Biology (2020) 20:2; https://doi.org/10.1186/s12870-019-2171-6. It is a study that revealed with high certainty 295 families of mother, father and their progeny, called ‘trios’. Of these there are 116 with DNA SSR data available from the NFC for comparison, and many show at least one parent being Cox’s Orange Pippin, Jonathan or McIntosh.

A second source is the historic record of breeders. Muriel Smith in the National Apple Register (1971) lists both parents for 181 varieties of which we also have DNA data and a further 90 with one (European) variety as parent. These include lots with one of the parents being a Cox’s Orange Pippin, a Jonathan, or a McIntosh.

In the following we shall test the DNA SSR method discussed above against parentages given in these two sources. Then we’ll see what the parentages might be of MAN’s accessions.

*The tools*

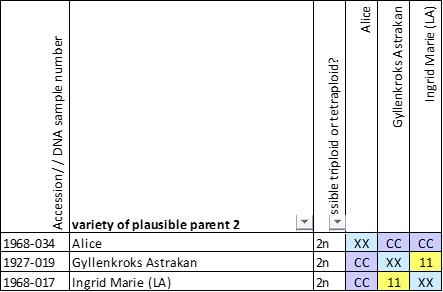
Peter Laws has pioneered the use of a workbook for searching for plausible parental combinations that could match the DNA fingerprint of a given variety. The example of Afal Cas Gwent showed the details of what is required of matching. Peter’s triumph was to find a way to implement this in a workbook with a relatively simple manual procedure. He employs a two-step approach, first identify those varieties that have at least one matching allele for each of the twelve marker-pairs, then second investigate whether in pairs they have all alleles of the progeny. This has been implemented as ‘Explorer-P2P’ and is available to download at <https://www.fruitid.com/#help>. Peter has provided regular upgrades. It really nice to use for those with some familiarity with Excel, it keeps close to the actual data, though its procedure does require some manual intervention which means workflow is a little slow.

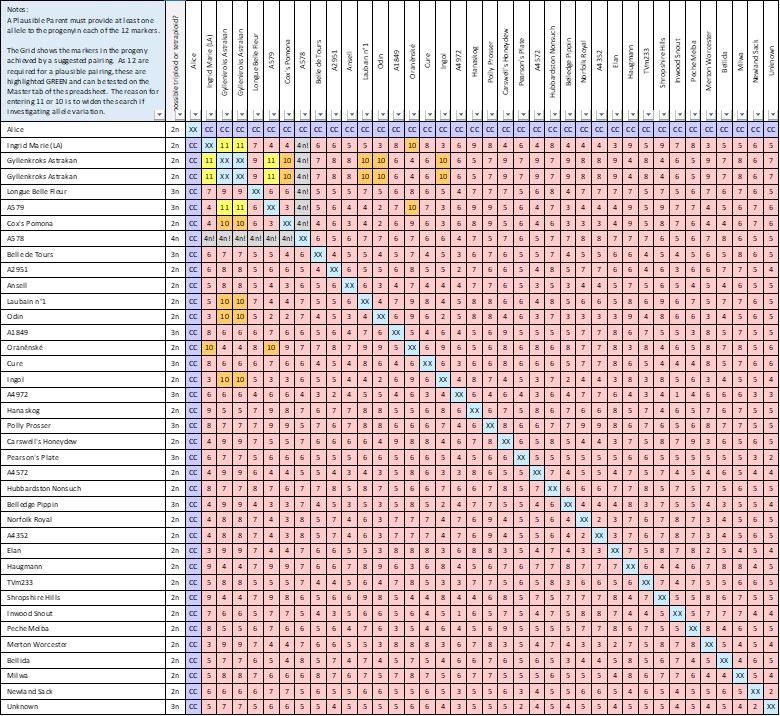
Here’s the output from Explorer-P2P for the variety ‘Grenadier’ found within a few seconds of selecting it. There are 19 plausible parents listed in the second column and repeated in the top row (some are very similar to one another). The named variety itself is included, though most varieties are self-sterile, and it is unlikely an exact copy would result even if fertile. The row and column opposite Grenadier has been blanked out with ‘CC’ (though handy to have a reminder of the variety in question). The diagonal from top left to bottom right has XX as again these self-pollinating combinations are unlikely. The table is symmetric about this diagonal, as DNA SSR does not distinguish between male and female; a combination Keswick Codlin x Hawthornden is indistinguishable from Hawthornden x Keswick Codlin.

Now we can focus on the rest of the cells with their colours and numbers. Green cells have 12, yellow cells are 11, orange are 10 and the rest are pink. It is the number of marker-pairs that are fully matched by the combination of varieties taking one from the second column and one from the top row. The combination Keswick Codlin x Hawthornden can match all twelve marker pairs of Grenadier. Both are diploid and introduction dates look reasonable Grenadier (1862), Keswick Codlin (1793), Hawthornden (1780). Another possibility is Ringer x Sowman’s Seedling; which can be rejected because of dates: Ringer (1864) and Sowman’s Seedling (1914)

Grenadier might also have resulted from a cross of Hawthornden with any of A636, A4777, A4954 and A5417; however their SSR are so similar to Keswick Codlin that they may be mutations or the results of small experimental glitches. What if there is a data glitch with a mismatch between alleles?

The next example is for ‘Alice’ where there is a mismatch of 2 base pairs (or nucleotides), abbreviated to bp, in marker-pair CH04e05. Parents are Ingrid Marie x Gyllenkroks Astrakan, but we see that only 11 of the marker pairs can be matched with this combination. There are no other varieties that are plausible parents (matching all twelve marker-pairs).

Suppose we believe an experimental glitch or a mutation has occurred; look at all those varieties that only match eleven, not twelve, marker-pairs. Now 32 plausible parents are revealed (again including Alice). However, only one more possible combination of parents is suggested with 11 matches: Gyllenkroks Astrakan x A579. It isn’t very likely as Alice was raised in Sweden in 1943 while A579 is a variety submitted by the Agri- Food & Biosciences Institute and remains unmatched. Parental combinations that have fewer matches to Alice’s DNA are much less probable.



Unless there is reason to include them, I’ve found it better to be cautious before invoking unmatched varieties, asking is it likely that a parent could itself be a seedling? There are just a few such cases that seem reasonable, more later on that. Oh yes, plausible parents could also be plausible progeny, it’s only when pair combinations are made that they become parental.

Both Peter Laws and I record our fulsome thanks to Dr Matthew Ordidge for kindly technical advice and encouragement along our journeys.

*Comparison of parentage derived from DNA SNP and SSR*

Comparison of parentage derived from DNA SSR fingerprints with a definitive set from DNA SNP (that’s Single Nucleotide Polymorphism) gives a good indication of how reliable the method is. It also show where difficulties arise. Both then give an impression about the confidence with which parents are identified.

Of the 116 SNP parents-progeny (known as ‘trios’) that also have complete DNA SSR data were unequivocally and immediately matched in 54 cases. Parents of a further 46 varieties were identified quite easily, though this will often have required using additional information, such as ploidy and provenance. Together that’s 90% of all tested. Details of all these matches are given in a table (link).

The most commonly encountered issue has been the prevalence in our dataset (i.e..the NFC plus other collections) of progeny, and of multiple ‘incestuous’ parentages from a few well-loved varieties including Cox’s Orange Pippin, Jonathan, or McIntosh. The simple approach adopted has been of the type if Cox’ Orange appears, for instance in the case of Ellison’s Orange, along with some or all of Carswell's Orange, Holstein, Honey Pippin, Ivette, Jupiter, Karmijn de Sonnaville, Lynn's Pippin, Merton Pippin, Oranje de Sonnaville, Polly Prosser, Primus, Winter Gem, then the ‘top’ parent will, or may, be presumed. With this simple screening, it might take a minute to ten to scan through 2-100 parental combinations, remember triploids can pretty certainly be discounted. In only 15 cases was there significant ambiguity including Gascoyne's Scarlet, Geheimrat Doktor Oldenburg, Herefordshire Russet, King David, Kyokko, Rival, Sowman's Seedling, Thurso, and Upton Pyne. Ambiguity can arise because of data glitches, when there are many plausible parental combinations and when several parents have similar fingerprints probably because they are closely related.

Overall, 47% of parentages were correctly found without ambiguity, and a further 40% were found reasonably correctly with a little careful consideration. Most of the remainder found one or other parent with little ambiguity.

The table show gives details of ones that gave a wrong parentage. Two other major mismatches were noted: Brighton in which the marker-pair CH01f03b was adrift by 12 or even 22 bp, and Kyokko in CH01f02 by 12bp. No idea why! There are minor dataset mismatches in 12 trios (10%). Sufficiently often that we should be vigilant, even cautious, but not often enough to undermine its utility.

Don’t expect it to be infallible but a potentially useful tool. Cheap and cheerful…. But that’s what DNA SSR is compared with whole genome studies.

*Comparison between parentage from Plant Breeder records and that derived from DNA SSR*

We have established that (well-defined) DNA SSR fingerprints can be about 90% reliable in the identified parents. The alternative test, whether any two parents are **not** likely to be parents of a given variety, has a near 100% chance of being correct. We can now assess whether reported parentages by Plant Breeders and others are likely correct, or wrong and in that case which are more likely parents.

A convenient summary listing of parentages is given in the National Apple Register. A subset of 274 varieties was taken which comprised all the two-parent varieties (i.e. trios) and single-parent varieties of a European origin. Fortunately, almost all are in the NFC, and DNA SSR fingerprints are available.

Of these 187 varieties with two parents and having SSR for progeny and parents, only 96 appear to have both parents plausible, perhaps another 5 trios might be plausible if there were experimental ‘glitches’ in DNA fingerprints; 87 were clearly implausible. The detailed assessment is available on our website (link).

It suggests that either frequent mistakes were made in the original ID of trees, or record keeping was poor, or parentages were retro-actively assessed from morphology. Among other breeders it is notable that Wastie failed to get one right out of 11. By contrast, Research Stations such as East Malling, Long Ashton, Merton, Wageningen, and Ottawa are fairly reliable. It can be done. Parentage records should be viewed with considerable caution.

*Suggested parentage of varieties in MAN’s collection from fingerprint results*

Having established that fingerprints can be used to tease out parentage sometimes, the next investigation has been reviewing all 864 DNA samples planned by MAN and its members, 26 of which were not submitted or had no clear result; there are 838 samples with results. For various reasons a number of varieties have been sampled and analysed several times. There are in total 480 different varieties.

Of all 864 samples 126 also have had SNP analysis made with about 100 having high confidence in parentage found. There are 712 samples without SNP analysis for consideration here.

We expect far fewer instances with parents revealed because we can expect that many varieties (both progeny and parents) haven’t been found, or are seedlings, or haven’t had DNA SSR measured. Indeed it is feared that the parents of many of the varieties we now treasure may be extinct.

|  |  |
| --- | --- |
|  | number |
| DNA samples with results but no SNP also made | 712 |
| DNA samples having at least one plausible parents |  |
| that are probable | 89 |
| that are possible | 92 |
| that are unlikely | 31 |
| no confidence at all | 499 |

There are 67 varieties that have two parents probably identified. About 20% of the samples have at least one parent identified with at least a moderate degree of confidence (‘probables’ and ‘possibles’).

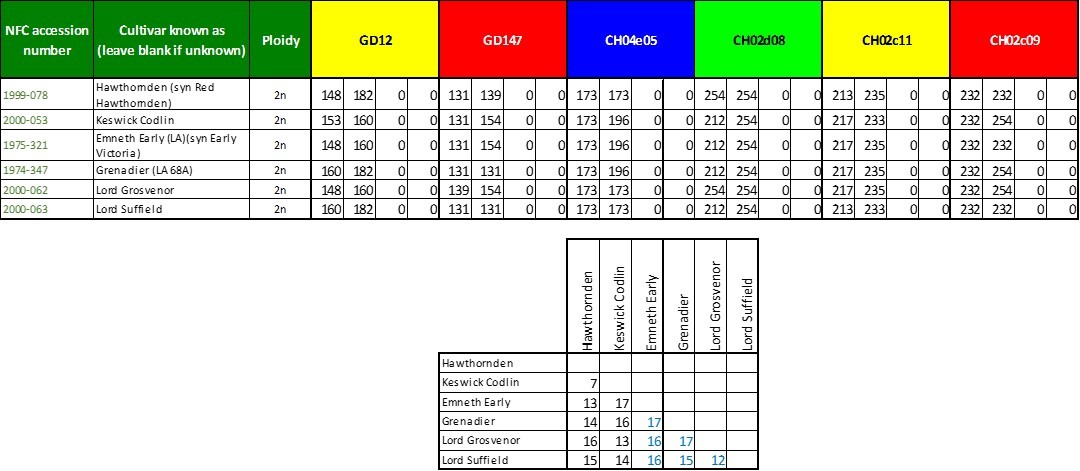
Suggested parents are given **here** (with link) with an impression of confidence in that assessment. Please remember, these are just suggestions, not definite evaluations.

Several examples are shown below for which we have photographs available, with thanks to John Savidge, Charles Martell and the NFC.

*Emneth Early, Grenadier, Lord Grosvenor and Lord Suffield* The SNP study showed that Emneth Early (1899) and Grenadier (1862) was progeny of Hawthornden (1780) x Keswick Codlin (1793), now it is seen that so too are Lord Grosvenor (1872) and Lord Suffield (1836). All are diploids. This is a rather neat conclusion and looks consistent with provenance and morphology.



Now let’s look in a little more detail at the DNA of parents and progeny. As there are some marker pairs that have just one allele reported, we recognize that actually both alleles have the same value. Expanding this duplication gives a modified fingerprint:



Count the instances where any two of these cultivars have the same numerical values for a pair of alleles of each marker pairs. Hawthornden and Keswick Codlin share just seven of the same alleles, which is about 30% of the 24 alleles. The other 70% of the non-common alleles gets shared out to progeny like cards from a well shuffled pack. Hawthornden shares respectively 13, 14, 16 and 15 of its alleles with the four progeny. We would expect this to be on average 7 + (24-7)/2 or 15.5. It is actually 14.5 and for Keswick Codlin as parent 15. Given the small number of market pairs, this is well within statistical variations. In all cases there is at least one matching pair of alleles for each marker pair.

Consider the instances of matching pairs of alleles between progeny. It ranges from 12 to 17 with an average of 15.5 for the six combinations. In some cases there is no matching of alleles for a given maker pair, for instance between Lord Grosvenor and Lord Suffield for both marker-pairs CH01f03b and GD147; this is as expected, since in principle two progeny can receive a completely different set of genes to each other. This understanding may be useful in assessing other family groups.

*Dolafallen* (diploid) from Keswick Codlin (diploid,1793) x Yellow Ingestrie (diploid,1800); parentage assessed as possible.



*Ffordd Las* (diploid) from White Transparent (diploid,mid-C19) x Stark's Late Delicious (diploids,1912); parentage assessed as possible.



*Jolly Miller* (diploid,1883), Murfitt's Seedling (diploid, ca 1883) x Lord Lennox (Finzi) (diploid, <1829); parentage assessed as possible



*Martin Nonpareil (MAN)* (diploid, 1795) from Keswick Codlin (diploid,1793) x Chatley Kernel (diploid,1894); parentage assessed as probable. Note for the historic Martin Nonpareil to have been progeny of Chatley Kernel, the latter must have been in existence for at least 100 years earlier; we believe this is not improbable.



*Saint Cecilia* (diploid,1900) from Cox's Orange Pippin (diploid,1825) x Lane's Prince Albert (diploid,1840); parentage assessed as probable.

*A503 Ty Du 2* (diploid) from Annie Elizabeth (dip,1857) x (A306, A1383, Reynold's Crab or Kernel) (dip); parentage assessed as possible.



**In summary**

Assessing parentage from DNA fingerprints was previously thought a step too far; I hope you are minded to believe it maybe just about possible for some.

No worry, there are still another 4000 varieties in the NFC/fruitID inventory for you to work on.

**Parentage of pear varieties**

Just the same principals as apples. A listing is given in ’The Book of Pears’ by Joan Morgan of plant breeder records for 105 varieties. Whether the DNA supports these, or there are alternative parental combinations is being assessed in a similar manner as above.