DNA Fingerprinting update – towards identifying parents

Progress has continued apace. Results from MAN's 2019 campaign were touched upon in the 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). As it isn't involved with inherited characteristics (that bit is in the extrons), it isn't under evolutionary pressure and the intron is thought to mutate less quickly. This is useful as it is makes the method more likely stable over centuries.

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 introns 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 intron 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 intron area 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 the US, 19 marker-pairs are used. 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 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/

	CH04c07_PK1	CH04c07_PK2	CH04c07_PK3	CH04c07_PK4	CH01h10_PK1	CH01h10_PK2	CH01h10_PK3	CH01h10_PK4	CH01h01_PK1	CH01h01_PK2	CH01h01_PK3	CH01h01_PK4	GD07_PK1	GD07_PK2	GD07_PK3	GD07_PK4	CH01f02_PK1	CH01f02_PK2	CH01f02_PK3	CH01f02_PK4	CH01f03b_PK1	CH01f03b_PK2	CH01f03b_PK3	CH01f03b_PK4	GD12_PK1	GD12_PK2	GD12_PK3	GD12_PK4	GD147_PK1	GD147_PK2	GD147_PK3	GD147_PK4	CH04e05_PK1	CH04e05_PK2	CH04e05_PK3	CH04e05_PK4	CH02d08_PK1	CH02d08_PK2	CH02d08_PK3	CH02d08_PK4	CH02c11_PK1	CH02c11_PK2	CH02c11_PK3	CH02c11_PK4	CH02d09_PK1	CH02d09_PK2	CH02d09_PK3	CH02d09_PK4
Cox's Orange Pippin	106	112	0	0	96	0	0	0	117	129	0	0	116	150	0	0	203	205	0	0	158	0	0	0	148	153	0	0	139	150	0	0	173	198	0	0	254	0	0	0	215	0	0	0	232	256	0	0
Braddick Nonpareil	106	0	0	0	96	107	0	0	111	119	129	0	106	118	150	0	168	178	182	0	136	170	178	0	147	148	0	0	135	137	139	0	173	200	0	0	228	254	0	0	215	227	231	0	244	256	0	0
Severn Bank	114	118	120	129	96	101	0	0	111	115	127	0	114	116	0	0	178	180	0	0	158	170	176	0	147	148	0	0	135	139	150	0	173	216	220	0	216	254	0	0	213	229	233	0	232	244	0	0

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 probable or not.

Investigating Parentage

Now let's look at and a MAN example in some detail, Jeffrey Large Red 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.

	CH04c07_PK1	CH04c07_PK2	CH04c07_PK3	CH04c07_PK4	CH01h10_PK1	CH01h10_PK2	CH01h10_PK3	CH01h10_PK4	CH01h01_PK1	CH01h01_PK2	CH01h01_PK3	CH01h01_PK4	Hi02c07_PK1	Hi02c07_PK2	Hi02c07_PK3	Hi02c07_PK4	CH01f02_PK1	CH01f02_PK2	CH01f02_PK3	CH01f02_PK4	CH03d03b_PK1	CH03d03b_PK2	CH03d03b_PK3	CH03d03b_PK4	GD12_PK1	GD12_PK2	GD12_PK3	GD12_PK4	GD37_PK1	GD37_PK2	GD37_PK3	GD37_PK4	CH04e05_PK1	CH04e05_PK2	CH04e05_PK3	CH04e05_PK4	CH02b03_PK1	CH02b03_PK2	CH03d03_PK3	CH03d03_PK4	CH02c11_PK1	CH02c11_PK2	CH02c11_PK3	CH02c11_PK4	Hi03c09_PK1	Hi03c09_PK2	Hi03c09_PK3	Hi03c09_PK4
A421	106	0	0	0	88	96	0	0	111	129	0	0	108	124	0	0	182	193	0	0	136	178	0	0	153	0	0	0	131	137	0	0	173	0	0	0	212	254	0	0	215	227	0	0	232	244	0	0
Newton Wonder	96	106	0	0	88	96	0	0	129	0	0	0	108	116	0	0	180	182	0	0	136	170	0	0	148	153	0	0	131	137	0	0	173	0	0	0	224	254	0	0	227	0	0	0	232	244	0	0
Reinette Rouge Étoilée (syn	106	120	0	0	88	103	0	0	111	129	0	0	118	124	0	0	186	193	0	0	178	182	0	0	148	153	0	0	131	0	0	0	173	206	0	0	210	212	0	0	205	215	0	0	232	244	0	0

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.

	CH04c07_PK1	CH04c07_PK2	CH04c07_PK3	CH04c07_PK4
A421	106	0	0	0
Newton Wonder	96	106	0	0
Reinette Rouge Étoilée (syn	106	120	0	0

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 JLR is OK.

	CH01h10_PK1	CH01h10_PK2	CH01h10_PK3	CH01h10_PK4
A421	88	96	0	0
Newton Wonder	88	96	0	0
Reinette Rouge Étoilée (syn	88	103	0	0

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 JLR, then it is NW that passed the 129.

	CH01h01_PK1	CH01h01_PK2	CH01h01_PK3	CH01h01_PK4
A421	111	129	0	0
Newton Wonder	129	0	0	0
Reinette Rouge Étoilée (syn	111	129	0	0

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 JLR.

	Hi02c07_PK1	Hi02c07_PK2	Hi02c07_PK3	Hi02c07_PK4
A421	108	124	0	0
Newton Wonder	108	116	0	0
Reinette Rouge Étoilée (syn	118	124	0	0

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

Using the 4000 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) 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 JLR 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 less likely than their statistical presence may suggest.

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.

News Sheet last spring noted that more information can be squeezed from the DNA than just from matching to identify. It has been realised for some time that the DNA fingerprints can give information about parentage, at least which varieties are not the parents of a given variety! Further work has shown that it can identify plausible pairs of parents, for instance 'Golden Delicious' x 'Kidd's Orange Red' produced 'Gala'. How far can it go and with what certainty?

There are two datasets for testing how useful is the parentage from DNA SSR. A paper by Hélène Muranty et al. with the 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 was published earlier this year. It is a study that revealed with high certainty 295 families of mother, father and their progeny, called 'trios'. Of these there are 115 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 Jeffrey

Large Red 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' and is available to download at https://www.fruitid.com/#help as a 5 MB workbook. 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.

A complementary tool has been developed by MAN. Though the workflow is essentially the same, it is more seamlessly integrated. The process is to enter the named variety of interest and immediately see all plausible combinations of varieties together with the number of marker-pairs that they match. If a match is not revealed then the number of marker-pairs matched by the plausible parents can be reduced to 11 or 10. Additionally fingerprints of other samples that have a lesser confidence can be searched too. As parents are much more likely to be diploid, a filter can be applied to select only diploids as one parent. The tool is loosely called 'P2P', for parent 2 progeny, and requires little Excel skill to use, but at 17 Mb it is much larger than Explorer. It is available via fruitID.com/#help; there is also a user guide. Once plausible combinations of parents are found if required their fingerprints can be relatively simply compared visually.

Here's the output from P2P for the variety 'Grenadier' found within a fraction of a second after typing the variety name. There are 14 plausible parents listed in the second column and repeated in the top row. 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. Rhode Island Greening is tetraploid and has 39 non-zero alleles so has a bigger chance (bite of the ~~cherry~~, oops apple) in matching, and parentage rules with tetraploids are not yet clear to me; again these are blanked out.

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)

Accession/ / DNA sample number	variety of plausible parent 2	ssible triploid or tetraploid?	A1884	A3195	Emneth Early (LA)(syn Early Victoria)	Fiessers Erstling	Grenadier (LA 68A)	Hawthornden (syn Red Hawthornden)	Keswick Codlin	Mrs. Lakeman's Seedling	Norfolk Beauty	Rhode Island Greening (4x)	Ringer	Robert Blatchford	Scotch Dumpling	Sowman's Seedling
A1884	A1884	2n	XX	9	10	8	CC	8	7	10	8	4n!	11	9	5	9
A3195	A3195	2n	9	XX	11	10	CC	7	10	9	9	4n!	11	7	11	9
1975-321	Emneth Early (LA)(syn Early Victoria)	2n	10	11	XX	8	CC	9	9	9	9	4n!	11	8	11	7
1947-469	Fiessers Erstling	2n	8	10	8	XX	CC	11	6	10	11	4n!	11	10	9	8
1974-347	Grenadier (LA 68A)	2n	CC	CC	CC	CC	XX	CC	CC	CC	CC	CC	CC	CC	CC	CC
1999-078	Hawthornden (syn Red Hawthornden)	2n	8	7	9	11	CC	XX	12	7	5	4n!	10	4	9	7
2000-053	Keswick Codlin	2n	7	10	9	6	CC	12	XX	11	11	4n!	11	11	9	9
1930-029	Mrs. Lakeman's Seedling	2n	10	9	9	10	CC	7	11	XX	8	4n!	9	5	10	9
2000-073	Norfolk Beauty	3n	8	9	9	11	CC	5	11	8	XX	4n!	10	5	10	7
1965-044	Rhode Island Greening (4x)	4n	4n!	4n!	4n!	4n!	CC	4n!	4n!	4n!	4n!	XX	8	5	8	8
1924-004	Ringer	2n	11	11	11	11	CC	10	11	9	10	8	XX	9	11	12
1961-047	Robert Blatchford	2n	9	7	8	10	CC	4	11	5	5	5	9	XX	10	7
1949-276	Scotch Dumpling	2n	5	11	11	9	CC	9	9	10	10	8	11	10	XX	10
1927-070	Sowman's Seedling	2n	9	9	7	8	CC	7	9	9	7	8	12	7	10	XX

We do not need to search further in this case. But, 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).

Accession/ / DNA sample number	variety of plausible parent 2	ssible triploid or tetraploid?	Alice	Gyllenkroks Astrakan	Ingrid Marie (LA)
1968-034	Alice	2n	XX	CC	CC
1927-019	Gyllenkroks Astrakan	2n	CC	XX	11
1968-017	Ingrid Marie (LA)	2n	CC	11	XX

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.

Accession/ / DNA sample number	variety of plausible parent 2	ssible triploid or tetraploid?	A2951	A352	A578	A579	A883	Alice	Ansell	Belle de Tours	Belledge Pippin	Bellida	Carswell's Honeydew	Cox's Pomona	Cure	Dunwich Heath	Elan	Gyllenkroks Astrakan	Hubbardston Nonsuch	Ingol	Ingrid Marie (LA)	Merton Worcester (EMLA 1)	Milwa (syn. Junami)	Newland Sack	Norfolk Royal	Odin	Pearson's Plate	Peche Melba	Polly Prosser	Shropshire Hills	Tonino	TVm122	TVm233	Unknown
A2951	A2951	2n	XX	3	4n!	6	4	CC	6	4	4n!	7	6	6	5	5	6	8	8	5	6	6	7	5	7	5	5	6	6	6	6	4	4	4
A352	A352	2n	3	XX	4n!	7	4	CC	5	4	4n!	7	7	7	4	4	7	6	7	5	7	7	7	4	8	6	5	5	5	5	6	4	4	4
A578	A578	4n	4n!	4n!	XX	8	6	CC	6	6	4n!	8	7	8	6	5	7	8	7	6	7	7	6	5	8	7	6	6	5	6	6	6	7	5
A579	A579	3n	6	7	8	XX	7	CC	4	5	4n!	5	5	3	7	6	4	11	7	3	4	4	6	7	4	2	6	7	9	9	6	7	5	6
A883	A883	2n	4	4	6	7	XX	CC	7	4	4n!	5	7	7	5	6	7	6	8	5	7	7	6	6	8	6	6	5	6	6	4	4	4	5
1968-034	Alice	2n	CC	CC	CC	CC	CC	XX	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
A270	Ansell	2n	6	5	6	4	7	CC	XX	5	4n!	4	6	3	4	4	5	8	5	4	5	5	6	5	4	3	5	4	7	6	6	7	5	5
1947-295	Belle de Tours	3n	4	4	6	5	4	CC	5	XX	4n!	5	6	4	4	6	6	7	7	5	6	6	8	6	5	4	5	5	7	5	6	4	4	5
1949-134	Belledge Pippin	4n	4n!	4n!	4n!	4n!	4n!	CC	4n!	4n!	XX	3	5	3	5	4	4	9	6	2	4	4	5	5	4	3	5	5	7	7	4	5	3	4
1994-013	Bellida	2n	7	7	8	5	5	CC	4	5	3	XX	6	4	5	5	5	7	5	4	5	5	4	6	4	4	5	4	7	6	3	6	5	5
1964-035	Carswell's Honeydew	2n	6	7	7	5	7	CC	6	6	5	6	XX	5	8	5	3	9	8	4	4	3	5	6	4	4	6	9	8	8	5	5	5	5
1957-188	Cox's Pomona	2n	6	7	8	3	7	CC	3	4	3	4	5	XX	6	6	4	10	6	3	4	4	6	7	3	2	6	6	9	8	6	7	5	6
1948-300	Cure	3n	5	4	6	7	5	CC	4	4	5	5	8	6	XX	6	8	6	6	6	8	8	7	6	7	6	6	4	6	4	7	6	5	6
A149	Dunwich Heath	2n	5	4	5	6	6	CC	4	6	4	5	5	6	6	XX	4	7	5	4	5	4	4	2	5	5	3	5	5	5	4	4	5	2
1982-230	Elan	2n	6	7	7	4	7	CC	5	6	4	5	3	4	8	4	XX	9	7	3	3	2	4	5	3	3	5	8	8	8	4	5	5	4
1927-019	Gyllenkroks Astrakan	2n	8	6	8	11	6	CC	8	7	9	7	9	10	6	7	9	XX	7	10	11	9	8	6	8	10	7	5	7	4	7	6	8	7
1948-721	Hubbardston Nonsuch	2n	8	7	7	7	8	CC	5	7	6	5	8	6	6	5	7	7	XX	7	8	7	6	5	6	7	5	5	7	5	6	7	8	5
1974-035	Ingol	2n	5	5	6	3	5	CC	4	5	2	4	4	3	6	4	3	10	7	XX	3	3	5	5	4	2	5	6	7	8	4	5	3	4
1968-017	Ingrid Marie (LA)	2n	6	7	7	4	7	CC	5	6	4	5	4	4	8	5	3	11	8	3	XX	3	5	6	4	3	6	8	8	9	5	6	5	5
1979-176	Merton Worcester (EMLA 1)	2n	6	7	7	4	7	CC	5	6	4	5	3	4	8	4	2	9	7	3	3	XX	4	5	3	3	5	8	8	8	4	5	5	4
2002-044	Milwa (syn. Junami)	3n	7	7	6	6	6	CC	6	8	5	4	5	6	7	4	4	8	6	5	5	4	XX	5	5	5	5	6	7	7	3	6	6	4
2002-052	Newland Sack	2n	5	4	5	7	6	CC	5	6	5	6	6	7	6	2	5	6	5	5	6	5	5	XX	6	6	3	5	5	5	4	4	6	2
1933-004	Norfolk Royal	2n	7	8	8	4	8	CC	4	5	4	4	4	3	7	5	3	8	6	4	4	3	5	6	XX	3	5	7	9	7	5	6	6	5
1966-046	Odin	2n	5	6	7	2	6	CC	3	4	3	4	4	2	6	5	3	10	7	2	3	3	5	6	3	XX	6	6	8	8	5	6	4	5
A2119	Pearson's Plate	3n	5	5	6	6	6	CC	5	5	5	5	6	6	6	3	5	7	5	5	6	5	5	3	5	6	XX	5	6	5	4	4	6	2
1931-012	Peche Melba	2n	6	5	6	7	5	CC	4	5	5	4	9	6	4	5	8	5	5	6	8	8	6	5	7	6	5	XX	6	5	5	7	7	5
1961-058	Polly Prosser	3n	6	5	5	9	6	CC	7	7	7	7	8	9	6	5	8	7	7	7	8	8	7	5	9	8	6	6	XX	6	6	6	7	5
A428	Shropshire Hills	2n	6	5	6	9	6	CC	6	5	7	6	8	8	4	5	8	4	5	8	9	8	7	5	7	8	5	5	6	XX	6	5	7	5
2000-109	Tonino	2n	6	6	6	6	4	CC	6	6	4	3	5	6	7	4	4	7	6	4	5	4	3	4	5	5	4	5	6	6	XX	4	5	3
TVm122	TVm122	3n	4	4	6	7	4	CC	7	4	5	6	5	7	6	4	5	6	7	5	6	5	6	4	6	6	4	7	6	5	4	XX	4	3
TVm233	TVm233	2n	4	4	7	5	4	CC	5	4	3	5	5	5	5	5	8	8	3	5	5	6	6	6	4	6	7	7	7	5	4	XX	5
1947-190	Unknown	3n	4	4	5	6	5	CC	5	5	4	5	5	6	6	2	4	7	5	4	5	4	4	2	5	5	2	5	5	5	3	3	5	XX

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

And that's about all there is to it... now, about a thousand investigations later. 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 115 SNP parents-progeny (known as 'trios') that also have complete DNA SSR data were unequivocally and immediately matched in 52 cases. Parents of a further 59 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.
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. In only 9 cases was there significant ambiguity: 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, 45% of parentages were correctly found without ambiguity, and a further 45% were found reasonably correctly with a little careful consideration. Most of the remainder found one or other parent with little ambiguity.

Just four cases gave a wrong parentage: Golden Melon, Laxton's Superb, Norfolk Beauty and Rubens. In all four there was a mismatch between one parent and progeny, for three it was just 2 bp in one marker-pair. For Rubens there were eight mismatched marker-pairs when using the SNP derived parents, why? Matthew Ordidge has confirmed that EMR switched Rubens and Saltcote Russet around; the latter indeed has parents Cox's Orange Pippin and Knobby Russet, it would appear the switch was applied to SSR but not SNP dataset.

	Parentage from DNA SSR				
	unequivocal agreement	best easily selected	wrong	with minor mismatches	major mismatches
yes	52	50	4	12	3
maybe		9			

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 271 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 183 varieties with two parents, for only 87 do both parents appeared plausible, perhaps another 16 trios might be plausible if there were experimental 'glitches' in DNA fingerprints; 80 were clearly implausible. The detailed assessment is a substantial table and available on our website.

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 13. 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.

While checking the parentages, other possibilities arose from the SSR fingerprints (including 15 with SNP based parents) for the 96 trios with wrong or questionable parentage. There were 73 trios for which alternative pairs of parents seemed more probable. Together with the 87 with 'correct' parents, that is 160 plausible trios from the 183 varieties investigated. Overall 87% of the trios end up with plausible parents. That there are about 10% of the total where there may be experimental glitches is similar in number to those found above when comparing DNA SSR and SNP derived parents.

Summary of varieties investigated

selected from NAR with one or two parents	271
which have two parents listed	183
for which both parents are plausible	87
for which at least one is improbable	80
for which further consideration may be warranted	16
of those quoted parents found improbable, SSR fingerprint gives parents of	73
total plausible trios	160
number where marker-pair mismatches are small and may be glitches	19

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 794 DNA samples submitted by MAN, members and WPCS. For various reasons a number of varieties have been sampled and analysed several times. There are in total 423 different varieties.

MAN DNA samples submitted	738
WPCS samples submitted	56
Total varieties	423

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	varieties
DNA samples with two plausible parents	137	91
having high confidence	39	19
that are probable	17	12
that are possible	49	35
that are unlikely	19	15
no confidence at all	13	10

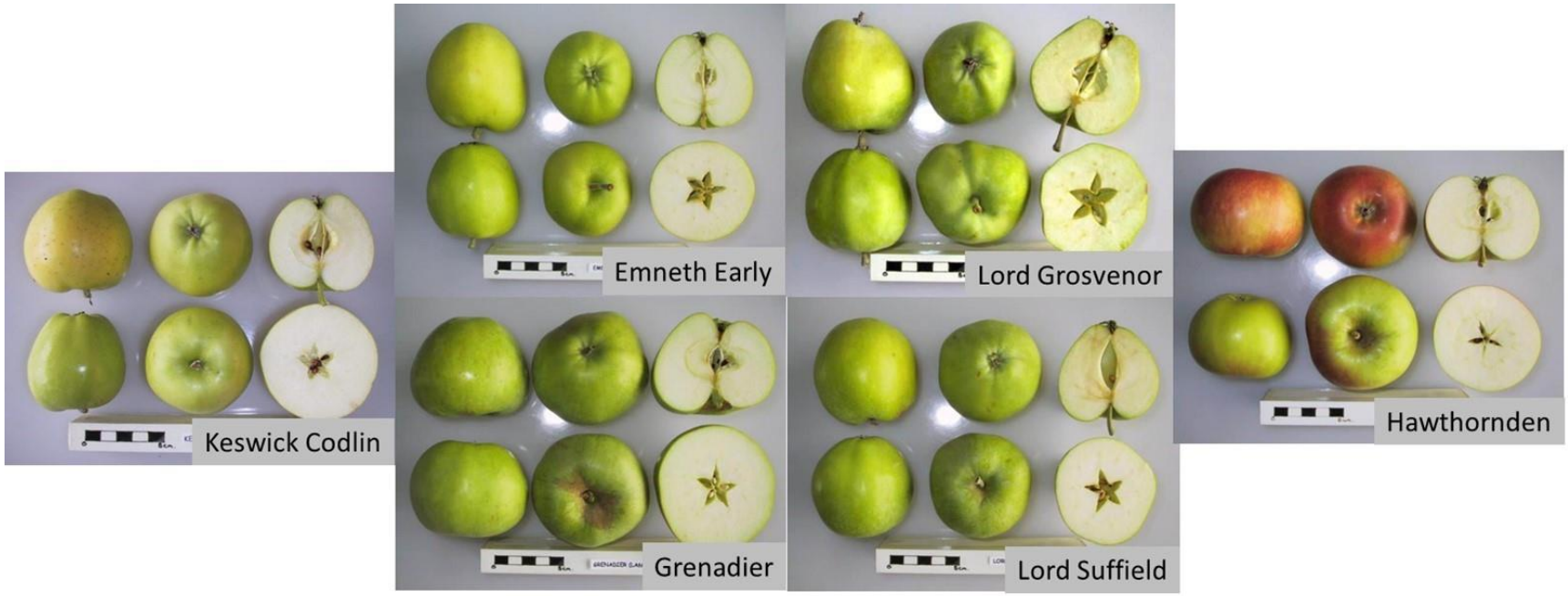
There are 68 varieties that have two parents identified with some degree of confidence; high (SNP), probable (SSR unambiguous), possible (SSR with some caveats). That is only about 15% of the 423 varieties being considered. It's much less than encountered with the plant breeder's records. Is it surprising or shocking?

Both the study with varieties in the DNA SNP dataset and those of the plant breeders have explicitly selected cultivars that are likely to have parent-progeny relationships. If anything the MAN/WPCS dataset is biased away from 'normal' cultivars to rarer varieties and ones that may be seedlings. There are about 300 varieties for which no suggestion has been made for either parent, of which 54 have at least 20 plausible parents identified, and ten have more than a 100. These include King of Tomkins County, Byford Wonder, Yorkshire Greening, Norfolk Beefing, Cockpit and Webb's Kitchen Russet. Perhaps the question is more how the DNA of these varieties became so well 'connected' in the dataset when the varieties aren't widespread in our area?

A total of 1048 DNA samples submitted by MAN, GOT, The Pippin Trust (TPT) and Welsh Perry and Cider Society (WPCS) have been investigated. Suggested parents are given with an impression of confidence in that assessment. Please remember, these are just suggestions, not definite evaluations. Results are given in a table.

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:

NFC accession number	Cultivar known as (leave blank if unknown)	Ploidy	CH04c07				CH01h10				CH01h01				Hi02c07				CH01f02				CH01f03b			
1999-078	Hawthornden (syn Red Hawthornden)	2n	94	108	0	0	96	96	0	0	119	121	0	0	116	120	0	0	180	182	0	0	136	158	0	0
2000-053	Keswick Codlin	2n	96	106	0	0	88	113	0	0	119	119	0	0	116	116	0	0	170	191	0	0	162	176	0	0
1975-321	Emneth Early (LA)(syn Early Victoria)	2n	106	108	0	0	96	113	0	0	119	119	0	0	116	116	0	0	180	191	0	0	136	176	0	0
1974-347	Grenadier (LA 68A)	2n	94	106	0	0	88	96	0	0	119	121	0	0	116	116	0	0	180	191	0	0	136	162	0	0
2000-062	Lord Grosvenor	2n	94	106	0	0	96	113	0	0	119	121	0	0	116	120	0	0	180	191	0	0	136	162	0	0
2000-063	Lord Suffield	2n	106	108	0	0	88	96	0	0	119	119	0	0	116	120	0	0	180	191	0	0	158	176	0	0

NFC accession number	Cultivar known as (leave blank if unknown)	Ploidy	GD12				GD147				CH04e05				CH02d08				CH02c11				CH02c09			
1999-078	Hawthornden (syn Red Hawthornden)	2n	148	182	0	0	131	139	0	0	173	173	0	0	254	254	0	0	213	235	0	0	232	232	0	0
2000-053	Keswick Codlin	2n	153	160	0	0	131	154	0	0	173	196	0	0	212	254	0	0	217	233	0	0	232	254	0	0
1975-321	Emneth Early (LA)(syn Early Victoria)	2n	148	160	0	0	131	154	0	0	173	196	0	0	212	254	0	0	217	235	0	0	232	232	0	0
1974-347	Grenadier (LA 68A)	2n	160	182	0	0	131	131	0	0	173	196	0	0	212	254	0	0	217	235	0	0	232	254	0	0
2000-062	Lord Grosvenor	2n	148	160	0	0	139	154	0	0	173	173	0	0	254	254	0	0	217	235	0	0	232	254	0	0
2000-063	Lord Suffield	2n	160	182	0	0	131	131	0	0	173	173	0	0	212	254	0	0	213	233	0	0	232	232	0	0

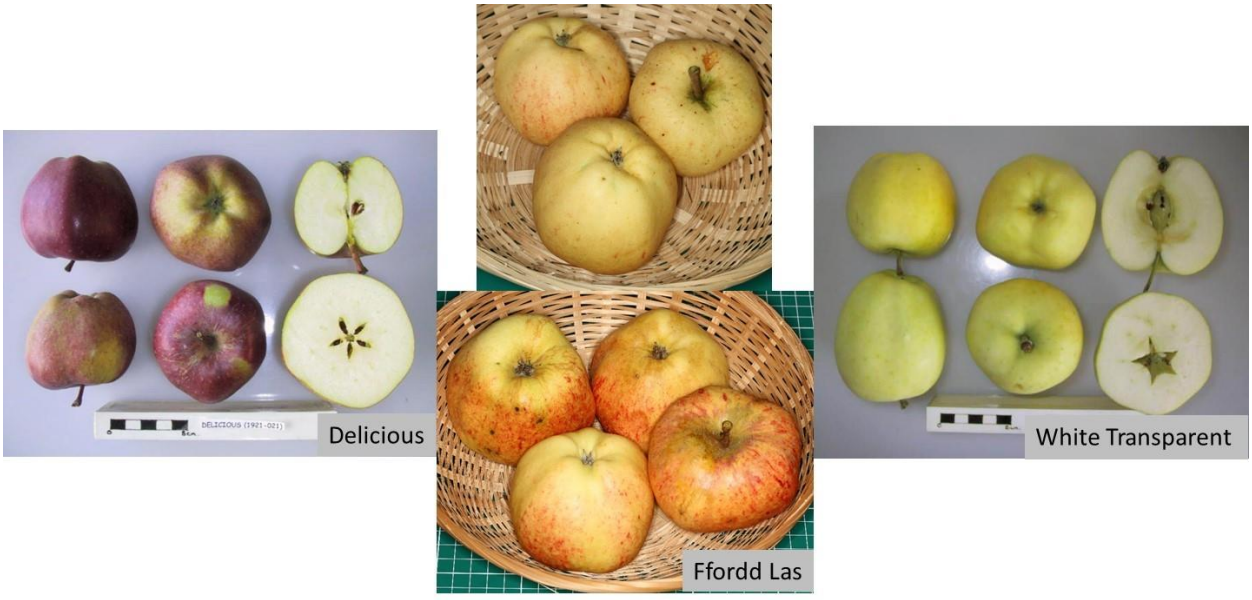
	Hawthornden	Keswick Codlin	Emneth Early	Grenadier	Lord Grosvenor	Lord Suffield
Hawthornden						
Keswick Codlin	7					
Emneth Early	13	17				
Grenadier	14	16	17			
Lord Grosvenor	16	13	16	17		
Lord Suffield	15	14	16	15	12	

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.



Gipsy King (diploid provenance 1872) from Barcelona Peamain (diploid,1831) x Reinette d'Anjou (triploid,1817); there is a mismatch on marker-pair CH02c11 of 10bp according parentage is assessed as unlikely

Grosmont Glory (diploid) from Marged Nicolas (diploid) x (Claygate Pearmain (triploid); parentage assessed as possible.

Hambling's Seedling (triploid,1893); Hawthornden (diploid,1780) x London Pearmain (diploid,1842); there is a mismatch on CH04c07 of 2bp; parentage assessed as unlikely

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

King Coffee (diploid,1934) from Bess Pool (diploid,1824) x Carter's Pearmain (triploid,1934); parentage assessed as possible.

King's Acre Pippin (triploid,1897) from Nonpareil (diploid,1500s) x A1142 (triploid); King's Acre Pippin has entire Nonpareil finger print into triploid; parentage assessed as possible. (Issue of triploid inheritance will be considered below.)

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.


Keswick Codlin


Martin Nonpareil (MAN)


Chatley Kernel

Onibury Pippin (diploid,<1883) from Golden Harvey (diploid,C16) x Keswick Codlin (diploid, <1793); mismatches on CH01h01 by 4 bp; parentage assessed as possible.


Keswick Codlin


Onibury Pippin


Golden Harvey

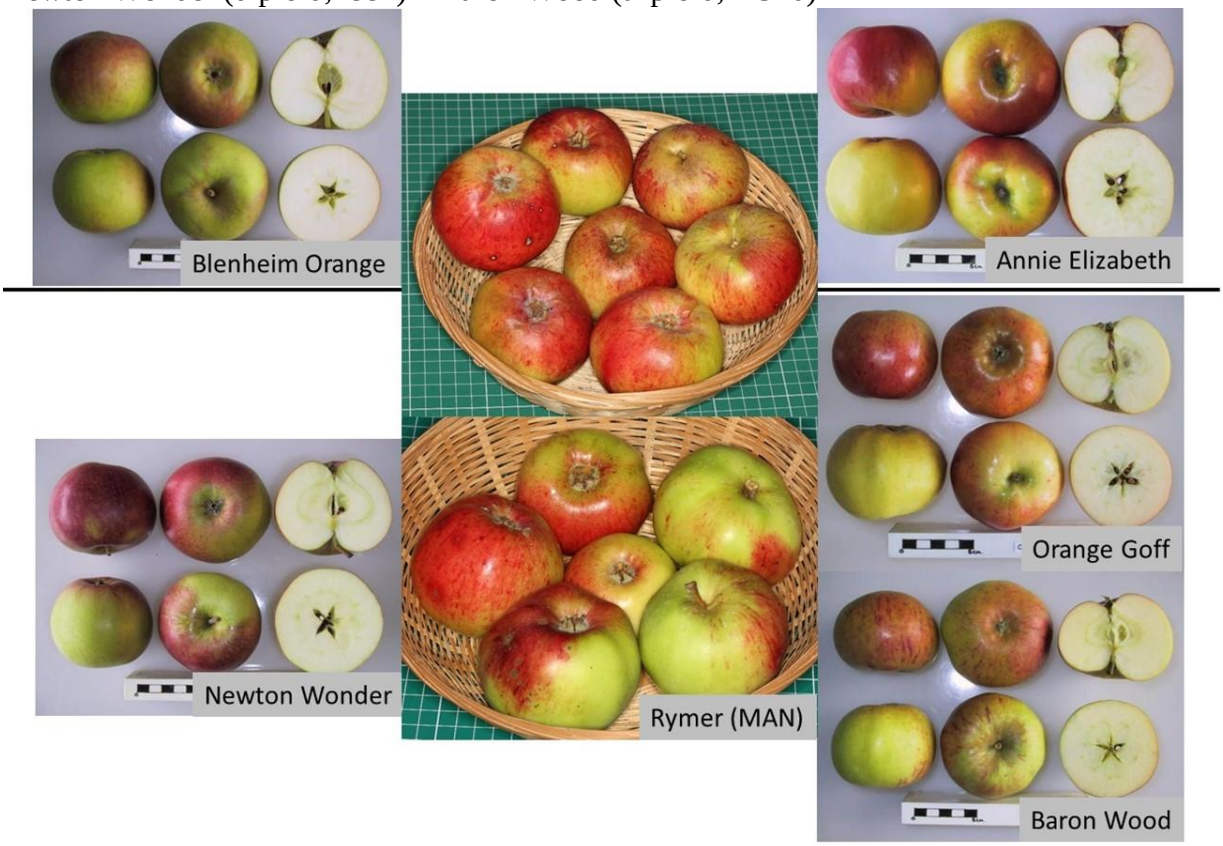
Pearson's Plate (triploid,1831) from Nonpareil (diploid,C16) x Reinette d'Anjou (tripoid,1817); parentage assessed as possible. Pairs of alleles of Reinette d'Anjou of each marker-pair carry across to Pearson's Plate, that makes confidence in parentage greater.


Nonpareil


Pearson's Plate


Reinette d'Anjou

Rymer (MAN) (diploid, likely ca 1900) from three possible pairs:
Annie Elizabeth (diploid,1857) x Bramley's Seedling (triploid,1813);
Newton Wonder (diploid,1887) x Orange Goff (diploid,1842);
Newton Wonder (diploid,1887) x Baron Wood (triploid,<1940)



Resolving which of these three will require some detailed morphological work, though it is interesting to note that Baron Wood is a triploid cultivar from Orange Goff.

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

Tewkesbury Baron (diploid,1883) from Devonshire Quarrenden (diploid,1678) x Rushock Pearmain (diploid,1821); mismatch in CH01h01 of 2 bp; parentage assessed as unlikely, alternative to Rushock Pearmain is A1271 (same as A1430).

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

There appear to be a cluster of varieties with similar morphology and very similar DNA that includes Cornish Pine, Sugar Quinlan (A484), White Castle Quoining (A510), Winter Quoining (Bunn) (A514) and Winter Quoining (Ted) (A1137) which has the same DNA as Walters 10 (A1142,A3401). Interestingly too Scilly Pearl and Vernade also appear to be associated with these. It points to a common family, perhaps related through Cornish Pine's parents? Next project….

Triploid inheritance from diploid parents

Detailed review of the DNA SSR dataset has shown a number of instances where the entire DNA fingerprint of a diploid variety is present in part within a triploid variety. It is not a co-incidence,

it is clear evidence of a parent progeny relationship. A few examples have been mentioned above and those noted are listed in the table.

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 3000 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.