Received: 28 April 2019 | Revised: 5 August 2019 | Accepted: 13 August 2019

DOI: 10.1002/ppp3.10074

RESEARCH ARTICLE

Plants People Planet PPP Open Access

The Long Ashton Legacy: Characterising United Kingdom West Country cider apples using a genotyping by targeted sequencing approach

Helen Harper[1] | Mark O. Winfield[1] | Liz Copas[2] | Sacha A. Przewieslik-Allen[1] | Gary L. A. Barker[1] | Amanda Burridge[1] | Bob R. Hughes[3] | Les Davies[4] | Keith J. Edwards[1]

[1]School of Biological Sciences, University of Bristol, Bristol, UK

[2]Lullingstone, Fore Street, Winsham, Somerset, UK

[3]7 Warren Lane, Long Ashton, Bristol, UK

[4]4 Springfield Terrace, Street, Somerset, UK

Correspondence
Mark O. Winfield, School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK.
Email: mark.winfield@bristol.ac.uk

Funding information
Bristol Centre for Agricultural Innovation (Project 81), Biological Sciences, University of Bristol.

Societal Impact Statement

The English West Country is the home of cider making, providing the region with jobs and industry, as well as cultural reference points such as Laurie Lee's *Cider with Rosie*. Many important cider apple varieties were developed at Long Ashton Research Station (LARS), near Bristol, UK, including 29 varieties known collectively as 'The Girls'. After its closure, some of the knowledge and expertise acquired at Long Ashton was lost, including the pedigree of 'The Girls'. We sampled LARS' derived trees and, using a novel genotyping technique, rediscovered the pedigree of 'The Girls', ensuring that this important cider apple collection will be available for future generations.

Summary

- Our research had two objectives: (a) record the influence of Long Ashton Research Station on the introduction of new cider apple cultivars to the UK; (b) rediscover the parentage of the cider apple cultivars known collectively as 'The Girls'.
- For rapid, cost effective and accurate genotyping, we used the recently developed, medium density, single nucleotide polymorphism-based genotyping procedure, SEQSNP®, to characterize the cultivars.
- We generated a medium density (1,500 markers), whole genome genotype for 245 apple cultivars that allowed us to determine the relationship between cultivars and, in so doing, rediscover the parentage of 'The Girls'.
- We show that SNP genotyping is an efficient tool for the analysis of genetic diversity in cider apples and apples in general, and that the cider apple breeding programme carried out at Long Ashton Research Station made, and continues to make, a unique contribution to UK cider production.

KEYWORDS

apple, cider, genotyping, Long Ashton Research Station, SEQSNP®, SNP

Helen Harper and Mark O. Winfield contributed equally.

1 | INTRODUCTION

In 2019, the UK cider sector was worth £3.1 billion (Westons Cider Report, 2019) with exports, representing 38% of the global cider market, reaching approximately £100 million (National Association of Cider Makers, https://cideruk.com). The success of the sector is reflected in the more than 500 UK cider makers that, collectively, employ over 10,000 people (National Association of cider Makers, https://cideruk.com). These companies range in size from small, artisan cider makers to large producers and they create ciders ranging in quality from farmhouse (sometimes referred to as 'scrumpy') to single variety ciders that many believe rival the best of wines. At the core of UK cider production is the availability of a large collection of apple cultivars exhibiting a rich array of flavors and aromas.

Many regions of the UK have their own, locally adapted apple cultivars and some of these areas produce commercial cider. The English West Country, consisting of the counties of Somerset, Devon, Hereford, Gloucestershire, Dorset, and Cornwall, has a rich history of cider making dating back several centuries. Early experimental work with cider apples and cider making was sponsored by the Bath and West and Southern Counties Society with later funding from the Bath and West of England Society (Barker, 1952). This work, initiated by Robert Neville-Grenville, led to the formation, in 1903, of a fixed institute for research and instruction in cider making and fruit growing, the National Fruit and Cider Institute (NFCI). The institute, which was situated on the estate of Lady Emily Smyth in Long Ashton near Bristol, became the Ministry of Agriculture funded Long Ashton Research Station (LARS) and continued to provide a source of expertise for both apple cultivation and cider production. As part of these activities, LARS undertook an apple breeding programme which led to the introduction of dessert cultivars such as Cheddar Cross (Allington Pippin x Star of Devon) and, later, cider apples such as Ashton Bitter (Stoke Red x Dabinett). In the early 1970s, LARS began research into induced clonal variation using Cobalt 60 irradiation; this work led to the development of the widely grown Bramley Clone 20 and the self-fertile Cox's Orange Pippin (Anderson, Lenton, & Shewry, 2003).

Although the threat of closure was ever present during the 1970s and '80s, work on cider apples continued. This work was partly sponsored by the Bulmers, Taunton, and Showerings cider companies whose factories were experiencing fruit processing problems because most cider apples matured simultaneously in October. To combat this problem, LARS undertook a breeding programme to generate early maturing cider apples that could be harvested prior to the main October to November season. The crosses made were between the cider apples Dabinett (D) and Michelin (M) as female parents and the dessert apples James Grieves (JG) and Worcester Pearmain (WP) as pollen donors. These crosses generated 1,500 seedlings (500 D x JG; 200 D x WP; 650 M x JG and 150 M x WP). In 2007, after extensive trials, 29 of these lines were selected and named. In most cases, the names given to these cultivars were chosen from female workers associated with the breeding project and so collectively the lines became known as 'The Girls' (Copas, 2014; Morris, 2010). These cultivars combined the desired characteristic of early maturity (late September) with regular cropping of good sized, bittersweet fruit and an easily managed tree shape (Anderson et al., 2003). Since their selection, a number of 'The Girls' have proven to be highly popular with cider makers such that, between 2006 and 2017, over one million trees, mainly Amanda, Angela, Debbie, Fiona, Gilly, Hastings, Helen's Apple, Jane, Lizzy, Prince William, Three Counties and Vicky, have been planted for cider apple production (Morris, 2010 and Copas, personal communication). Unfortunately, during the propagation of the original seedlings, records of parentage were lost (Copas, 2014).

Since LARS closed in 2003, its centenary year, many of the cultivars it produced or introduced have been used for commercial cultivation or maintained in local orchards where they have been cared for by passionate individuals such as John Thatcher of Thatcher's Cider. With the passage of time, however, these cultivars may be lost to cultivation or become mislabeled as they pass from one orchard to the next. To ensure that future generations can identify them, we have collected numerous samples and characterized them using a novel, single nucleotide polymorphism (SNP)-based, combined genotyping and sequencing platform, SEQSNP®. While the main aim of our work was to reassign each of the 'Girls' to their correct parents, the sequence data generated will allow identification of LARS trees in future breeding programmes.

2 | MATERIALS AND METHODS

2.1 | Collection of plant material

A total of 245 apple cultivars were collected; principally, these came from locations in Somerset and Bristol although samples were also provided by the John Innes Centre in Norwich (Dataset S1, '*Source of Cultivars*'). Included in the samples were 58 lines derived from the LARS' breeding programme that produced 'The Girls'. Twenty-nine of these were the named 'Girls' and a further 29 were lines that were considered of inferior quality and so not given names (number lines in Dataset S1, '*Source of Cultivars*'). For 169 of the 254 cultivars in the study, only a single tree was sampled. For all others, samples were taken from more than one tree catalogued or labeled as a specific cultivar. From the cultivar Yeovil Sour, 24 replicates were taken from a single tree in order to test reproducibility of SEQSNP® genotyping. Sampling took place in September (2018) so that features of the fruit could be observed.

To test the ability of the SEQSNP® genotyping to aid in the identification of unknown samples, leaves were collected from eight apple trees of unknown identity; no attempt to identify these cultivars was made prior to genotyping. In addition, a small number of samples (10) were collected from local gardens. These samples were of uncertain origin although a provisional name was given to them by the person who provided the sample (Dataset S1, '*Source of Cultivars*'). In total, 380 samples were collected.

In all cases, following LGC 'Plant Sample Collection Kit' instructions, three leaf discs from a single young leaf were sampled and placed in a 96-well plate. These samples were sent to LGC for DNA extraction and SEQSNP® analysis.

2.2 | SEQSNP® design

Validated SNPs from the 480K Axiom Apple Array and their map locations were obtained from Supplementary Table S1 of Bianco et al. (2016) and converted to SEQSNP® markers. The 487,249 SNPs from the array were filtered to include only robust, poly high resolution markers as defined by Bianco et al. (2016). In addition, we only selected markers with concordant chromosome assignments on the Golden Delicious (Valesco et al., 2010), Renetta Grigia di Torriana (Falginella et al., 2015), Fuji (Kunihisa et al., 2016) and Pinova (Di Pierro et al., 2016) maps. For SEQSNP® design, additional flanking sequence of 100 bases either side of each Axiom SNP were obtained by cross referencing the *Malus domestica* sequences (downloaded from https://www.rosaceae.org/species/malus/malus_x_domestica/genome_v3.0.a1). Where this additional sequence was unavailable, SNPs were discarded. Finally, multiple SNPs within a *Malus* v3.0.a1 contig were discarded if they were located less than 100 bases from one already selected. SNPs were chosen to be evenly distributed across the genome by selecting an initial SNP from every integer centimorgan (cM) position of each genetic linkage group (defined here as a locus). This process was repeated to add additional markers evenly to each locus until all SNPs were allocated. After the first iteration, a single marker had been allocated to each locus and a further five iterations allocated up to a total of six markers suitable for SEQSNP® design to each locus (1,700 in total). This process maximized the chances of designing at least one successful assay for every genetic locus with SNP data meeting our design thresholds.

2.3 | Genotyping protocol

DNA was extracted from leaf tissue by LGC using their proprietary extraction method, sbeadex™. Genotyping was performed according to the SEQSNP® protocol by LGC (SEQSNP® guidance notes, LGC web site). The number of reads was calculated for each probe and cultivar after adaptor and quality trimming. Probes with less than 50 reads per cultivar were removed from further analysis.

2.4 | Dimensionality reduction

The relationship between the cultivars was determined from the SNP data. A pair-wise similarity matrix including all 380 samples (all 245 cultivars) was constructed using a custom Python script (available on request): similarity was calculated as the number of calls in common between two cultivars divided by total number of markers scored for them; markers that had missing calls for either of the cultivars being compared were not used to estimate similarity. The resulting matrix was imported into the R statistical software package version 3.3.1 (R Core Team, 2013); multi-dimensional scaling was performed using '*cmdscale*' with K = 5, and the first coordinate plotted against the other four; dendrograms were created using the '*hclust*' function; plotting was performed using the '*as.pyhlo*' function of the ape library.

2.5 | Determining the parentage of 'The Girls'

Parentage of 'The Girls' was inferred from the similarity matrix derived from the genotype data. For each 'Girl', the two parental cultivars with the greatest similarity were assumed to be the parents. Unfortunately, samples of Shamrock, one of 'The Girls', failed genotyping and so are not included in the analysis.

2.6 | Calculation of heterozygosity levels

Levels of heterozygosity were calculated for all samples based on 1,301 markers by dividing the number of heterozygous loci by the total number of genotyped loci. Average heterozygosity was calculated for all 380 samples together. It was also calculated separately for the 27 known triploids, the 27 samples of unknown or provisional identity (Dataset S1, '*Unknown Samples*'), and for the remaining 326 samples which were assumed to be diploid. Since this latter group contained the 24 replicates of Yeovil Sour, in order to eliminate any bias, 22 of these were removed prior to calculating of heterozygosity; thus, only 304 samples were used. An *F* test was performed to compare variance of the diploid and triploids samples and a *t* test was performed to compare the means.

2.7 | Calculating a minimum number of SNPs required to identify a specific cultivar

To identify a minimal set of SNP markers capable of differentiating all cultivars, we first selected the marker with the highest minor allele frequency. Using a Perl script (available on request), we then evaluated all remaining markers to see which one differentiated the highest number of cultivars that were not split by the first marker. The script iterated this process until either adding more SNPs did not provide any further splits or all cultivars were resolved.

3 | RESULTS

3.1 | SNP design

Of the 487,249 available Axiom SNP markers, 54,202 met our design criteria of being robust, poly high resolution and mapping to concordant linkage groups in the four available genetic maps (Di Pierro et al., 2016; Falginella et al., 2015; Kunihisa et al., 2016; Valesco et al., 2010). From these 54,202 SNPs, we selected 1,700 highly polymorphic markers evenly distributed across the genome. These 1,700 SNPs were processed using LGC's SNP pipeline to identify 1,500 suitable for the SEQSNP® genotyping platform (Dataset S2).

3.2 | Sequence coverage and genotype accuracy

In total, 245 cultivars (169 as single samples and 76 with replicates—380 samples in total) were genotyped using the SEQSNP® protocol. This generated 104,207,906 75-base pair sequences resulting in 570,000 genotype calls distributed across the 1,500 probes (Dataset S2). The

number of sequence reads per probe across all the cultivars ranged from 85 to 169,050 with a mean of 60,485. To improve accuracy of allele calling, only cultivars with a sequence read depth of at least 19,386 (average sequence read depth of 50 per probe per cultivar), were taken forward. This resulted in 199 probes being discarded. Of the remaining 1,301 probes, the lowest number of probes (38) mapped on linkage group 16 and the highest number (115) mapped to linkage group 15 (Dataset S2).

To confirm the accuracy of SNP calling, we examined the calls from 24 replicate samples taken from the same tree of the cultivar Yeovil Sour. Across these technical replicates, allele calling was, at worst, 99.5% identical, indicating an error rate of less than 0.5% or six SNP differences across 1,301 SNP markers.

3.3 | Relationship between the apple cultivars genotyped

The genotyping data were used to evaluate the relationship between cultivars. Overall, the cultivars fell into nine broad clusters (Figure 1 and Figure S1): Cluster 1, a small group of 18 samples, contained Cider Lady's Finger and Frederick; Cluster 2, the second largest with 75 samples, contained the 24 replicates of Yeovil Sour and Blenheim Orange; Cluster 3, a small group of 22 samples, contained Michelin (one of the two potential female parents to 'The Girls') and the 'Girl', Early Bird; Cluster 4, the largest group with 104 samples, contained, Dabinett (potential female parent to 'The Girls') and the 'Girls' Angela, Fiona, Gilly, Hastings, Helen's Apple, Jane, Jean, Naomi, Sally, Three Counties, Tina, Tracey, Vicky and Willy; Cluster 5, with 52 samples, was interesting in that it contained most of the known triploid cultivars such as Bramley, Ashmead Kernel, Morgan Sweet and Tom Putt; Cluster 6, with 57 samples, contained James Grieve (one of the two potential male parents to 'The Girls') and the 'Girls' Amanda, Betty, Debbie, Joanna, Lizzy, Margaret and Prince William; Cluster 7, a group of eleven samples, contained Redstreak; Cluster 8, the smallest group with only 6 samples, contained the *Malus* species, *M. niedzwetzkyana*, (Niedzwetzky's apple) and *M. sylvestris*, and the two crab apples Evereste and Red Sentinel; Group 9 contained Worcester Pearmain (potential male parent to 'The Girls') and two cultivars reported to be triploid (Black Vallis and Gennet Moyle) that did not cluster with the Bramley apples (Cluster 5). Interestingly, with the exception of Early Bird, all 'The Girls' are found in Clusters 4 and 6.

We collected supposed replicate samples of all 'The Girls'; for each named cultivar these replicates were collected from different trees in different orchards (Dataset S1). In most cases, replicate samples of 'The Girls' clustered as would be expected. However, in five cases, Amelia, Connie, Debbie, Eleni, and Nicky, the supposed duplicate samples had very different genotypes from each other (Figure S1) The other cultivars with duplicates that did not cluster were Ashton Bitter, Blenheim Orange, Broxwood Foxwhelp, Burrowhill Early, Cap of Liberty, Don's Seedling, Somerset Redstreak, Sweet Alford (however, four of five did), Sweet Coppin, Taylor's Seedling, and White Jersey (Figure S1).

In a small number of cases, SEQSNP® genotyping did not produce distinct genotypes for two different, named cultivars. This was

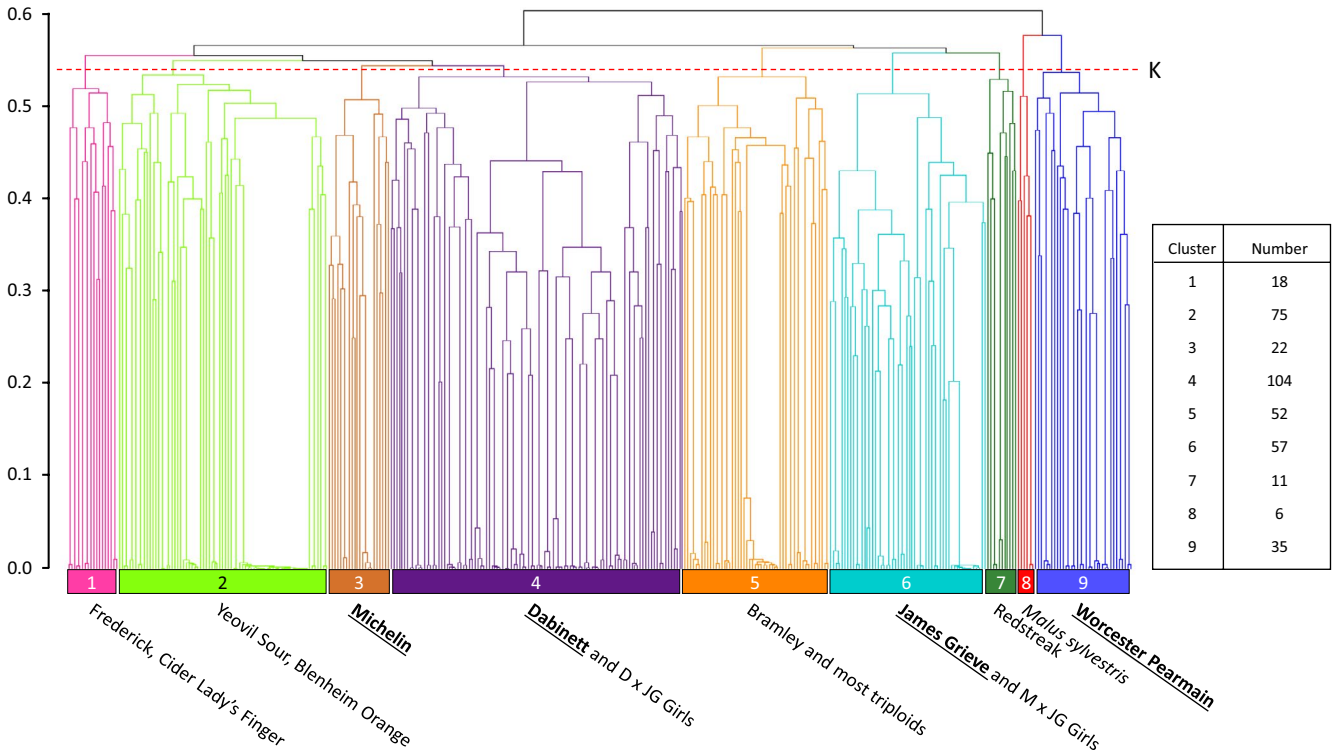

FIGURE 1 Dendrogram showing the relationship of all apple lines used in the study. The dendrogram is split into nine clusters by the line K. Each cluster has been highlighted in a different color and has been numbered. Selected cultivars belonging to each group are listed below the dendrogram; highlighted in bold are the parental lines that LARS used in the crosses that gave rise to 'The Girls' (Dabinett (♀), James Grieve (♂), Michelin (♀) and Worcester Pearmain (♂))

the case for Cox's Orange Pippin and its sport, Queens Cox; only one SNP difference between them. Similarly, Tom Putt and the sport derived from it, Red Tom Putt which has heavily red flushed fruit, were 99.6% identical; there were only 5 SNP differences (0.39% difference) between them. These differences are less than the difference (error) between the technical replicates of Yeovil Sour. What is more, most of the duplicates of other cultivars showed at least this level of difference so we cannot say that these cultivars are clearly different.

Conversely, Loders M and Loders P, two cultivars found during a hunt for old Dorset cider apple trees and thought to be the same type because they had similar growth habit and shared the same juicing characteristics (Copas, 2014), had only 50% of SNP markers in common.

3.4 | Parents of the 'Girls'

The genotyping data allowed us to infer which two of the four cultivars used in the LARS' breeding programme were the parents to each of the 'Girls'; unfortunately, samples of one of 'The Girls', Shamrock, failed genotyping and so inferences could be made about only 28 rather than all 29 of 'The Girls'. On a PCO plot, coordinate one separates lines from the LARS' breeding programme into two groups. One of these groups is positioned between the female parents Michelin and Dabinett and the male parent, James Grieve, the other between the females and Worcester Pearmain (Figure 2a). The cluster most similar to James Grieve contains cultivars that are clearly distinct from all other cultivars in the study (this group contains 74 lines: 59 of the named 'Girls' and 15 of the inferior lines that were not given a name). The group closer to Worcester Pearmain, which contains 14 named 'Girls' and 8 unnamed lines, is much less distinct from the other cultivars in the study. Each of these two groups splits into two further groups when coordinate 4 is plotted (Figure 2b). The most distinct groups are those which lie between James Grieve and Michelin, and James Grieve and Dabinett. These are clearly distinct from all other cultivars studied indicating that they are probably the offspring of the parental cultivars that flank them. Although less distinct from the other cultivars studied, there are also lines from the LARS' breeding programme that lie between Dabinett and Worcester Pearmain, and Michelin and Worcester Pearmain.

It is assumed that each of the 'Girls' will have more SNP markers in common with its two parents than to pretenders for that role. Using a subset of the data from the similarity matrix (Dataset S1, '*Similarity Matrix*'), we constructed a table showing the similarity of each of the 'Girls' with each of the four possible parents (Figure 3a). The cultivars Angela, Fiona, Gilly, Hastings, Helen's Apple, Jane, Jean, Naomi, Sally, Three Counties, Tina, Tracey, Vicky, and Willy, all of which fell into Cluster 4 of the dendrogram (Figure 1), are clearly more similar to Dabinett and James Grieve than they are to the other two "parents" (Figure 3a,b). Amanda, Betty, Debbie, Joanna, Lizzy, Margaret, and Prince William (found in Cluster 6), on the other hand, are more similar to Michelin and James Grieve. Hannah and Jenny are most similar to Dabinett and Worcester Pearmain. Strangely, Early Bird is most similar to the two female parents, Dabinett and Michelin, which were not reported to have been crossed. The genotypes of

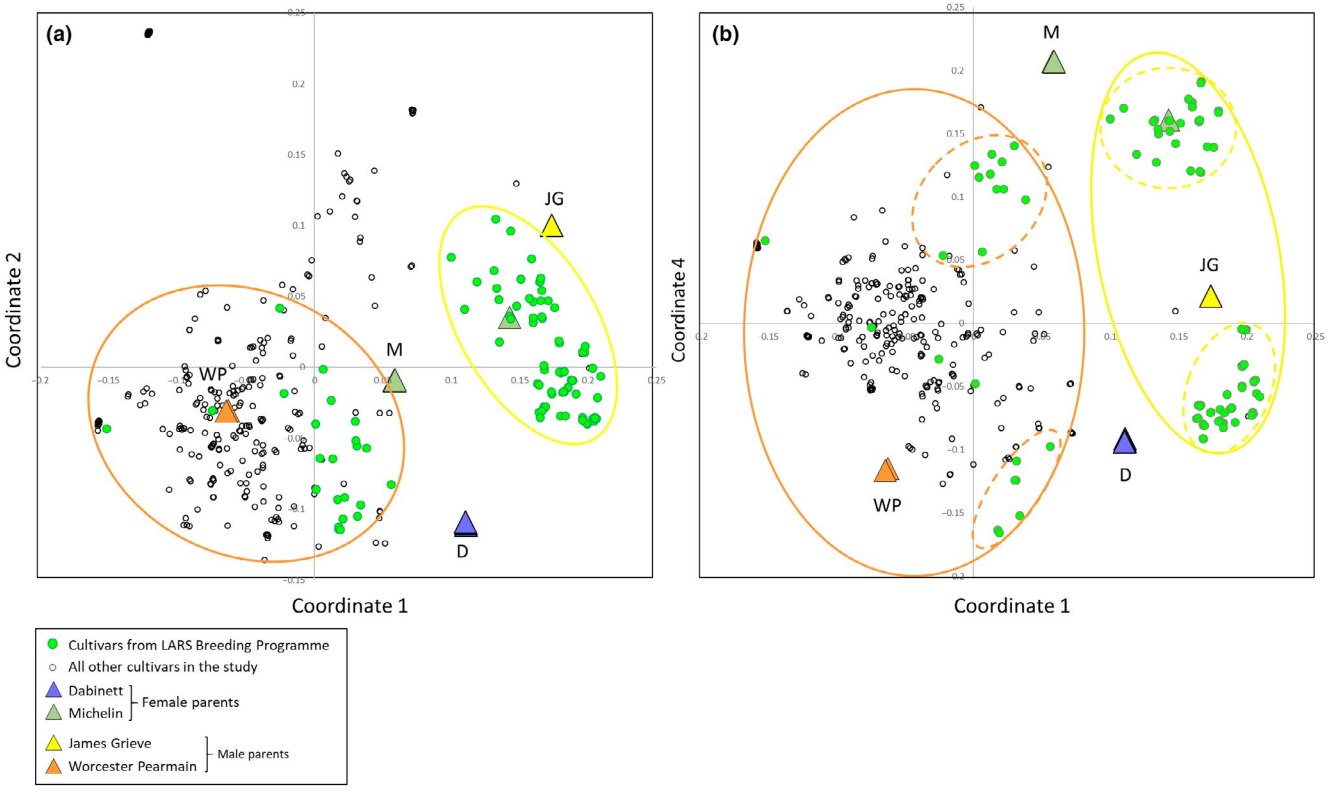

FIGURE 2 Principal coordinate plot of all samples. (a) Plot showing coordinate 1 versus coordinate 2. (b) Principle coordinate 1 versus coordinate 4; 'The Girls' essentially fall into four clear groups. In each plot, the lines from the LARS' breeding programme that produced 'The 'Girls' are highlighted in green; the four possible parents are highlighted by large triangles of different colors

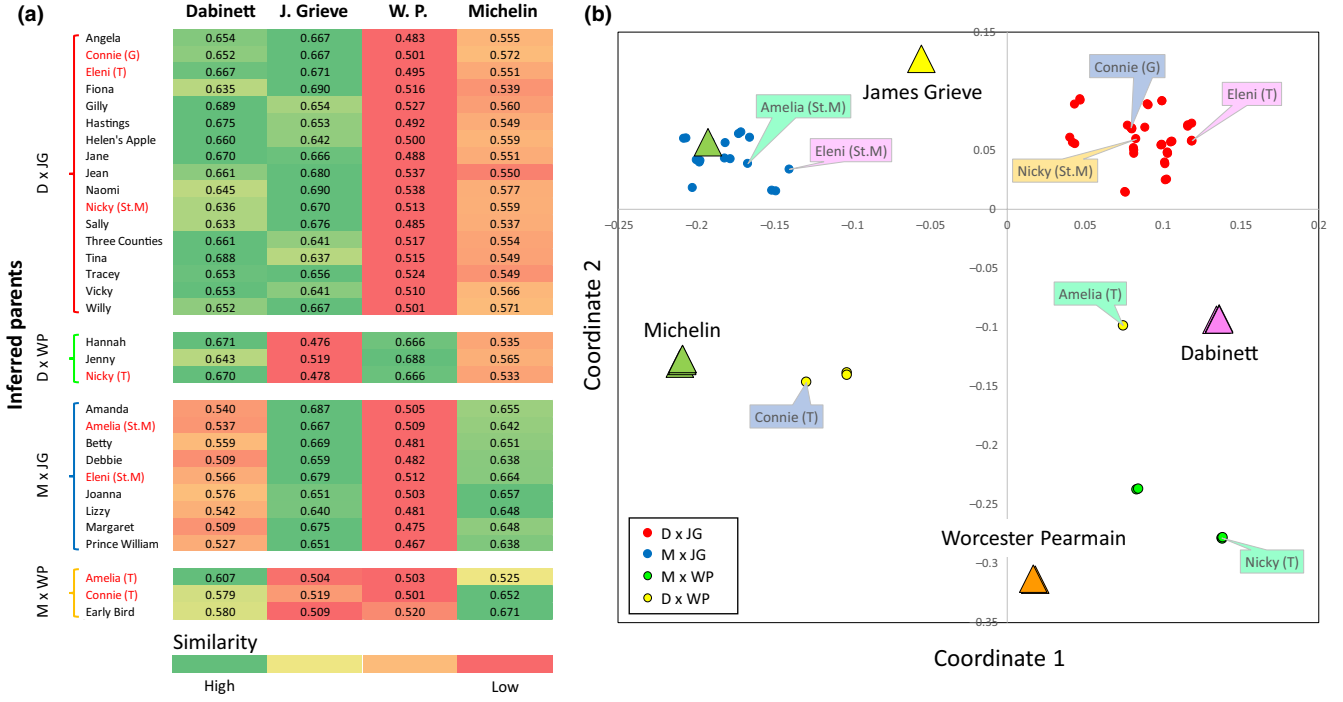

FIGURE 3 Inferred parentage of 'The Girls'. (a) Heat map of similarity between 'The Girls' and their potential parent lines (Dabinett (♀), James Grieve (♂), Michelin (♀) and Worcester Pearmain (♂); dark green is highly similar; dark red is highly dissimilar; percentage similarity is written on the heat map. Red text highlights the duplicates that appear in to two different groups. (b) PCO plot of 'The Girls' (colored circles) and their potential parent lines (colored triangles); duplicates of a cultivar that does not lie close together are named

the supposed replicate samples of the four cultivars Amelie, Connie, Eleni and Nicky (highlighted in red text in Figure 3a and labeled in Figure 3b) were very different from each other and no conclusions about their relationship to the 'parents' could be drawn.

3.5 | Using genotyping to identify unknown cultivars

As part of our study, we included eighteen samples of unknown or unconfirmed identity (Dataset S1, '*Unknown Samples*'). Of these, based on their position in the dendrogram, we were able to suggest an identity for eleven; the other seven could not be identified as they did not cluster with any of the cultivars examined in this study.

3.6 | Heterozygosity and ploidy

Heterozygosity scores for the 380 samples studied ranged in value from 0.19 to 0.53, with an average of 0.36. Of these samples, however, 27 were from cultivars known to be triploid, such as Bramley and Ashmead's Kernel (Dataset S1, '*Heterozygosity*'). These lines had an average heterozygosity of 0.48 (max = 0.52, min = 0.34), whereas the average for the assumed diploid lines (unknown lines and 22 of the Yeovil Sour lines excluded) was 0.35 (max = 0.53, min = 0.19). In a box and whiskers plot of these data, diploids and triploids are clearly distinct (Welch Two Sample *t* test *p*- value of 8.0e-14). However, it was apparent that, among the supposedly diploid samples, there was a small number of samples (13) with levels of heterozygosity similar to those of the known triploids (Figure 4). Similarly, among the known triploids, there were three samples, one Blenheim Orange sample and the two samples of Genet Moyle, that had levels of heterozygosity comparable to diploids (Dataset S1, "*Heterozygosity*").

The 27 samples of unknown or provisional identity (Dataset S1, '*Unknown Samples*') had average heterozygosity of 0.39 (max = 0.51, min = 0.30). However, this group of samples clearly contained individuals falling into two distinct groups (Figure 4): seventeen samples had low heterozygosity (mean of 0.33; max = 0.36, min = 0.30), and 10 had high heterozygosity (mean = 0.50; max = 0.51, min = 0.48). The provisional names of the former group were of diploid cultivars and the provisional names of the latter were of reported triploids such as 'Bramley' and 'Cooker', a common coinage for Bramley-like apples.

3.7 | Developing a minimum set of SNP markers

We were able to identify a set of just 25 SEQSNP® markers capable of discriminating all cultivars genotyped in this study. With the addition of six markers, this set also covers all seventeen linkage groups (Dataset S1, '*Minimum Marker Set*').

4 | DISCUSSION

The main objectives of our study were to provide a permanent genetic record of apple cultivars developed or introduced to the UK by LARS and, more particularly, determine the parentage of the cultivars

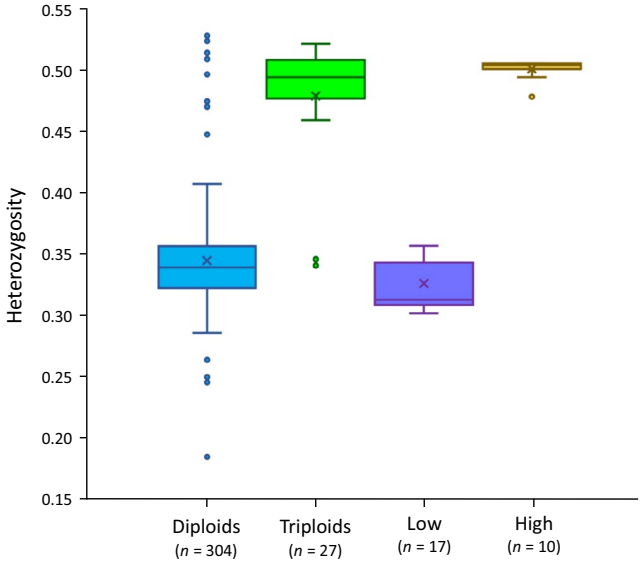

FIGURE 4 Box and whisker plot of heterozygosity of apple cultivars based on genotypes determined by SEQSNP®. The diploid box contains all samples that had not previously been reported to be triploid. The triploid box contains cultivars reported to be triploid (DEFRA report, 2010). The outliers to the triploids are the two Gennet Moyle samples and one of the Blenheim Orange samples. The "Low" and "High" high boxes are of the unknown and provisional samples

collectively known as 'The Girls'. We chose to use SNP markers rather than the more commonly used 12 microsatellites for two main reasons: (a) although the apple microsatellites are highly polymorphic and able to distinguish between the majority of, although not all, apple cultivars (DEFRA, 2010), they do not cover all 17 chromosomes (Coart, 2003); (b) microsatellite analysis is difficult to automate, especially with regard to data capture and scoring. Indeed, many breeders and scientists working on agronomically important crops have moved away from using microsatellites and are now using SNPs. In addition, as one of the main purposes of our study was to generate sufficient sequence and genotyping information for future generations to identify LARS-derived material, we took the view that SNP markers would provide our dataset with a degree of future proofing. We chose to use SEQSNP®, a novel SNP-based genotyping technique, because it is relatively inexpensive (£20/sample) compared to SNP-array-based genotyping and it generates both genotype and sequence information which is relatively easy to present in a spreadsheet and that can be replicated in-house or via a commercial service provider. This is in contrast to Genotyping by Sequencing (GBS) protocols which result in sets of markers specific for each cross or for each collection and so are difficult to compare across studies. Finally, the use of whole genome, SNP-based genotyping allows one to easily convert informative SNPs into individual Kompetitive Allele Specific Primer (KASP) markers (LGC Genomics). For instance, examination of the SNPs used in this study suggested that, when used together, just thirty-one SNPs converted to KASP markers would be capable of distinguishing all the cultivars examined here (Dataset S1, '*Minimum Marker Set*'). The requirement for such a small number of SNP-based KASP markers would make it relatively inexpensive (~£4 per sample) to genotype and re-catalogue the entire UK apple collection, especially if many of the DNA samples were already available due to their previous genotyping with microsatellites.

4.1 | Accuracy of SNP-based genotyping compared to the written record

Our initial examination of the large number of replicate samples taken from the same tree, suggested there were some inaccuracies in the genotyping (data not presented). Further examination of these inaccuracies showed that they were due to a small number of probes for which sequence coverage was low. The removal of probes that, on average, had less than 50 sequences per cultivar eliminated these inconstancies and led to a >99% similarity between technical replicates. In addition to these replicates from a single tree, we also collected replicates for 76 other cultivars; that is, we collected samples from the same named cultivar, but from different trees in different orchards (Dataset S1). Of these 76 cultivars, in 59 cases, the putative replicates clustered in accordance with the labeling provided by the various orchards. In 17 cases, however, putative replicates failed to cluster (Figure S1), indicating that many cultivars are labeled incorrectly. For example, of the five biological replicates of Michelin, four clustered while one was very different and was obviously not a true Michelin. Similarly, of the five replicates of Sweet Alford, only four clustered indicating that one of them had been labeled incorrectly. Unfortunately, for those cultivars for which we had only two samples and these had a different genotype, it was not possible to decide which, if either, was the correct genotype for the named cultivar. This was the case for the four 'Girls', Amelia, Connie, Eleni and Nicky (Figure 3 and Figure S1) and so we were unable to draw any conclusions about their parentage. Outside of 'The Girls', some of the other cultivars appeared to have been named erroneously. One of the Don's seedling samples, for example, clustered with Ashton Bitter while two others clustered with Tremletts Bitter (from which the records suggest it was partially derived). In these cases, there are two possibilities, either the samples were mislabeled when supplied to the orchard/person concerned or the samples were mixed during the genotyping procedure. We can discount the latter as leaf samples were collected by two people directly into a 96-well microtiter plate. Following this, samples remained in a 96-well plate until they were barcoded and processed automatically with results being directly fed into an SNP database. As an example of the errors in naming that can be made, two Cox's Orange Pippin trees that had been purchased from a reputable wholesale nursery, Keith 1 and 2, proved not to be identical to each other genotypically; Keith 1 was clearly labeled correctly as it formed part of a tight group with other samples of Cox's Orange Pippin; Keith 2, was distinctly different from any other cultivar in the study. Other cultivars, for which we did not have replicates, also hinted at mislabeling. For example, the cultivars Cadbury and Reinette d'Obry lay together on the dendrogram as did Langworthy and Reine des Pommes, both in Cluster 1 (Figure 1; Figure S1). Our results suggest that mislabeling of nursery

trees in both retail and wholesale nurseries can be significant, presumably occurring during replication, transplanting or shipping. Indeed, to some extent this problem is understandable since cultivars that are very similar phenotypically may be quite diverse genotypically, as would appear to be the case for the two samples of the cultivar Loders, which only shared 50% of SNP markers. One must assume that mislabeling will continue to be a problem, highlighting the need for this study and for centralized and well-characterized collections such as the National Fruit Collection, Brogdale (https://www.brogdalecollections.org/).

4.2 | The ability of SEQSNP® to discriminate between triploid and diploid cultivars

As both diploid and triploid cultivars are commonly grown, and both have been associated with LARS, we were interested in determining how well the SEQSNP® genotyping platform would perform with cultivars of differing ploidy. Larsen et al. (2018) reported that triploid cultivars have a higher level of heterozygosity than diploids. Our SNP data clearly support this finding. That is, mean heterozygosity for the known triploids such as Bramley, Ashmead Kernel, and Bulmer's Norman was higher than that for diploids (Figure 4; Dataset S1). Indeed, in our study, of the 17 named cultivars (represented by 27 samples) reported to be triploid in the DEFRA report GC0140 (2010), 16 (24 samples) had high heterozygosity (range 0.46–0.52). The exceptions to this were the two samples of Gennet Moyle and one of the three supposed replicates of Blenheim Orange. The replicates of Gennet Moyle, a cultivar reported to be triploid, had low heterozygosity (0.3421 and 0.3472). In addition, whereas most of the known triploids fell within Cluster 5 on the dendrogram (Figure 1; Figure S1), the two samples of Gennet Moyle did not; they fell into Cluster 9. Taken together, this might suggest that either the two samples are not Gennet Moyle or that Gennet Moyle is not a triploid. Of the three replicates of Blenheim Orange, only two had heterozygosity within the range of the other triploid cultivars. The outlying sample of 'Blenheim Orange' had heterozygosity similar to that seen for diploid cultivars and, therefore, probably represents a further mislabeled sample. A potential fourth replicate of Blenheim Orange, collected from a private garden (Daniel's garden) was almost certainly a Cox's Orange Pippin as it clustered with several other Cox samples in Cluster 6 of the dendrogram (Figure S1) and had low heterozygosity (0.30).

Contrary to this, samples from 13 cultivars not previously reported to be triploids had high heterozygosity and were outliers on the box and whiskers plot (Figure 4). These samples also fell into Cluster 5 on the dendrogram (Dataset S1). With further study, some or all of these might prove to be triploid. As a case in point, the cultivar that we called Stubbard had high heterozygosity (0.510), but is not recorded as a triploid. However, we found that the name Stubbard is a synonym of Stibbert that is recorded as a triploid. Tom Putt, which is recorded as a triploid, appears to be identical to Red Tom Putt (not reported to be triploid) which is a sport of it. These lie together on the dendrogram and have high heterozygosity (0.469 and 0.472, respectively) so both are probably triploids.

The lines with the lowest heterozygosity were the two *Malus* species, *M. niedzwetzkyana* (0.19) and *M. sylvestris* (0.247 and 0.252). These fell into Cluster 8 with the crab apples Red Sentinel, and Evereste, and the Dorset cider apple, Marnhull Mill, all of which had low levels of heterozygosity. This observation might reflect the self-fertile nature of these two species and three cultivars, suggesting that a degree of inbreeding has occurred to reduce heterozygosity. An acquisition bias, however, cannot be excluded as the SNPs included in the study were based on a *Malus domestica* reference genome. This contrasts with the third *Malus* species, *M. sieversii* (Cluster 4), which, reflecting its self-incompatibility and relatedness to domestic apple cultivars, had a heterozygosity level in the mid-range for the diploid cultivars.

4.3 | What does genotyping tell us about Long Ashton Cider apples and 'The Girls'?

The 380 trees genotyped using SEQSNP® fell into nine clusters. Of these, only clusters 4 and 6 included cultivars belonging to 'The Girls'. Cluster 4 contained all 'Girls' derived from the Dabinett x James Grieve cross, whereas Cluster 6 contained all those 'Girls' thought to be derived from the Michelin x James Grieve cross (Figure 2 and Dataset S1). In addition to the above, Cluster 4 also contains the two 'Girls', Hannah and Jenny, inferred to be derived from the Dabinett x Worcester Pearmain cross.

Clearly, the most successful crosses were those involving Dabinett x James Grieve and Michelin x James Grieve since all but three of 'The Girls' (Hannah, Jenny and Early Bird) appear to be derived from these two crosses. Hannah and Jenny are most probably derived from the cross between Dabinett and Worcester Pearmain. Early Bird, the only 'Girl' that does not appear to be derived from any of the three crosses mentioned so far, does not really appear to be derived from a Michelin x Worcester Pearmain cross either (Figure 2). It is highly similar to Michelin (0.671 similarity), and phenotypically it shares similarity too (Copas, 2014), but the second most similar parental cultivar is Dabinett rather than Worcester Pearmain. It may well be that Early Bird is derived from an open cross between Dabinett and Michelin.

5 | CONCLUSIONS

We have shown that the SNP-based genotyping platform SEQSNP® is a useful tool to identify apple cultivars. Due to the high level of heterozygosity in apple, we were able to discriminate between all cultivars sampled except those derived as sports from other named cultivars. What is more, due to the accuracy of the procedure, replicate samples clustered making it possible to identify cases where mislabeling has probably occurred. We believe it is significant that we have identified just 31 individual SNP probes, distributed across all 17 linkage groups, that would be sufficient to discriminate all the cultivars examined here; if this set of probes were to be used on the UK wide collection, it would provide a cheaper, more cost-effective and readily automated tool for the fingerprinting of apple cultivars than that presently

available. Finally, we are particularly pleased that the parentage of 'The Girls' could be inferred from our data as this part of the Long Aston Legacy has made a major contribution to the production of West Country cider with over one million 'Girls' sold since 2009 (pers. communication John Worle Nursery).

ACKNOWLEDGMENTS

This work could not have been carried out without funding from the Bristol Centre of Agricultural Innovation (BCAI). BCAI was set up as a Trust to manage the funds generated via the sale of some of the Long Ashton Research Station site. In addition, we would like to thank John Thatcher of Thatcher's Cider for his knowledge and enthusiasm and for maintaining the Thatcher's Heritage Orchard. We thank the following nurseries for allowing us to access and sample cultivars from their orchards: Worle Nursery, Field Farm Nursery, Shepton Mallet Nursery and Linden Lea Nursery. We also thank Miranda Krestovnikoff and the University of Bristol for allowing access to their orchards, which contain transplanted trees from the original LARS orchards. Furthermore, we thank the following for collecting samples of unknown identity for the project: Diane Hird, Daniel Robert, and Wendy Gibson. Finally, we acknowledge all of those who worked at Long Ashton Research Station, especially Peter Shewry and those who were employed during in its final years. Through its people and its apple cultivars, the Long Ashton legacy will continue to have a positive effect on UK agriculture well into the future.

AUTHOR CONTRIBUTIONS

M.O.W. carried out data analysis, produced the figures and supplementary data and made a major contribution to the writing of the manuscript. H.H. and K.J.E. secured the funding required to do the work, helped with sample collection and contributed to the writing of the manuscript. L.C. provided expertise in the identification and location of the various LARS derived cultivars. A.B. and S.A.P-A. analyzed the data for heterozygosity and contributed to the writing of the manuscript. B.R.H. assisted during sampling and provided valuable historic information on the breeding of 'The Girls'. G.L.A.B. generated many of the computer scripts required to harvest and analysis the SNP data. L.D. was involved in the original discussion, identification of the source material and in the collection of various samples from the Hereford orchards.

ORCID

Mark O. Winfield https://orcid.org/0000-0002-5078-4806

REFERENCES

Anderson, H. E., Lenton, & J. R., Shewry, P. R. (2003). *Long Ashton Research Station: One hundred years of science in support of agriculture*. Bristol, UK: University of Bristol. H.E. Iles (Central Press) LTD.

Barker, B. T. P. (1952). Long Ashton Research Station, 1903–1953. *Journal of Horticultural Science*, *28*(3), 149–151. https://doi.org/10.1080/00221589.1953.11513779

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denance, C., Theron, A., … Troggio, M. (2016). Development and validation of the Axiom(®) Apple480K SNP genotyping array. *The Plant Journal*, *86*(1), 62–74. https://doi.org/10.1111/tpj.13145

Coart, E., Vekemans, X., Smulders, M. M., Wagner, I., Van Huylenbroeck, J., Van Bockstaele, E., Roldan-Ruiz, I. (2003). Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Molecular Ecology*, *12*(4), 845–857. https://doi.org/10.1046/j.1365-294x.2003.01778.x

Copas, L. (2014). Cider apples The New Pomona. *Liz Copas*. ISBN 978-0-9568994-2-2.

DEFRA. (2010). *Fingerprinting the national apple and pear collections - GC0140*. Retrieved from http://sciencesearch.defra.gov.uk/Default.aspx?Menu=Menu&Module=More&Location=None&Completed=0&ProjectID=15150.

Di Pierro, E. A., Gianfranceschi, L., Di Guardo, M., Koehorst-van Putten, H. J. J., Kruisselbrink, J. W., Longhi, S., … van de Weg, W. E. (2016). A high density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. *Horticultural Research*, *3*, 16057. https://doi.org/10.1038/hortres.2016.57

Falginella, L., Cipriani, G., Monte, C., Gregori, R., Testolin, R., Velasco, R., … Tartarini, S. (2015). A major QTL controlling apple skin russeting maps on the linkage group 12 of 'Renetta Grigia di Torriana'. *BMC Plant Biology*, *15*, 150. https://doi.org/10.1186/s12870-015-0507-4

Kunihisa, M., Moriya, S., Abe, K., Okada, K., Haji, T., Hayashi, T., … Yamamoto, T. (2016). Genomic dissection of a 'Fuji' apple cultivar: Re-sequencing, SNP marker development, definition of haplotypes, and QTL detection. *Breeding Science*, *66*, 499–515. https://doi.org/10.1270/jsbbs.16018

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S., & Toldam-Andersen, T. B. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PLoS ONE*, *13*(8), e0201889. https://doi.org/10.1371/journal.pone.0201889

Morris, S. (2010). *'Harvest time for 29 new varieties of English cider apples'* The Guardian. Retrieved from https://www.theguardian.com/lifeandstyle/2010/sep/28/cider-apples-new-uk-varieties.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/. ISBN 3-900051-07-0.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., … Viola, R. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). *Nature Genetics*, *42*, 833. https://doi.org/10.1038/ng.654

Westons Cider Report (2019) Retrieved from www.ashdale-consulting.com/wp-content/uploads/2019/04/Weston-Cider-Report-2019.pdf.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Harper H, Winfield MO, Copas L, et al. The Long Ashton Legacy: Characterising United Kingdom West Country cider apples using a genotyping by targeted sequencing approach. *Plants, People, Planet*. 2019;00:1–9. https://doi.org/10.1002/ppp3.10074