Pedigree Validation and Reconstruction Applied in the Cultivated Potato Using 35 Microsatellites Markers

Martin Spanoghe[1] · **Alexandra Nirsha[2]** · **Thierry Marique[1]** · **Deborah Lanterbecq[1,2,3]**

Abstract

Knowledge of the parental origin of crop species is a prerequisite for formulating effective strategies for genetic conservation, management, and utilisation of genotypes in breeding programmes and genetic studies. In the case of the potato, although pedigree information is well documented, the possible lack of accurate pedigree data for certain cultivars constitutes a loss of opportunity to exploit the selection gain, the precise targeting of a market niche, or other breeding objectives. In this study, the genetic profiles of 1249 varieties were examined using a set of 35 polymorphic simple sequence repeats (SSR) markers with the available pedigree information to achieve pedigree validation. The maximum likelihood (LOD) was used to validate pedigree information of potato accessions for which the genetic data of at least one parent was available in the dataset. Our results showed that among the accessions for which both parents were genotyped, 81.7% presented a validated pedigree, while 18.3% presented a pedigree conflict (for at least one parent). Among the accessions for which only one parent was available in the dataset, 91.3% presented a validated pedigree, while 8.7% presented a pedigree conflict. Then, pedigree reconstruction was investigated to determine the most likely parent candidates for varieties whose pedigree is either questionable or lacking information.

✉ Martin Spanoghe
martin.spanoghe@condorcet.be

[1] Laboratoire de Biotechnologie Et Biologie Appliquée, Haute Ecole Provinciale de Hainaut-CONDORCET, 7000 Mons, Belgium

[2] Centre Pour L'Agronomie Et L'Agro-Industrie de La Province du Hainaut (CARAH), 7800 Ath, Belgium

[3] Hainaut Analyses (HA), 7000 Mons, Belgium

Introduction

Data on the parental origin of crop varieties provide a wealth of information on breeding choices, enabling breeders to make informed decisions regarding existing divergences in progeny, hybrid vigour, and effects of inbreeding depression (Tarn et al. 1992). As demonstrated in the cultivated potato, reliable pedigree information is helpful to estimate breeding values (Slater et al. 2014), to improve genetic gains for low heritability traits (Slater et al. 2016; Sood et al. 2020), and to track the inheritance of disease resistance alleles (Song et al. 2005; Song and Schwarzfischer 2008). This is also relevant for genetic relationship studies (Demeke et al. 1996; Isenegger et al. 2001), given that close clustering of full-sibs and half-sibs is an indication of a strong connection between estimated relationships and known pedigrees (Hosaka et al. 1994). Knowledge about parental information is likewise an essential key to classify recombinant and parental genotypes for linkage analysis (Luo et al. 2000), to achieve appropriate association mapping (Malosetti et al. 2007; Baldwin et al. 2011), to decipher genome-wide conserved patterns in elite parental lines (Li et al. 2018), and to determine the accuracy of genome-wide prediction (Endelman et al. 2018). Information about parents may also be useful to understand the differential interactions between potato genotypes and their pathogens (van Berloo et al. 2007).

A large amount of pedigree information on potato varieties is nowadays available on a web database (van Berloo et al. 2007). Although this is continually updated, pedigree data are missing for some accessions (i.e. orphans or unavailable information). Diverse sources of pedigree data may sometimes cause conflicting parental compositions or refer to identical varieties under different names. However, the orphan varieties could present unsuspected genetic resources that could be exploited in breeding programmes. Likewise, incorrect information about pedigree information can lead to unsuccessful results in field breeding trials. Thus, the lack of accurate pedigree data for many cultivars constitutes a loss of opportunity to exploit the precise targeting of a market niche or other breeding objectives. Given the major socio-economic and sustainability challenges facing agri-food systems, this is an important shortfall.

Thereby, there is a crucial need for reliable genetic tools in the confirmation of existing pedigree data and in the reconstruction of unknown or incorrect pedigree. Although molecular markers have been widely used for the analysis of parentage in natural populations (Jones and Ardren 2003) and to detect pedigree errors in breeding programmes (Visscher et al. 2002; Muñoz et al. 2014), few genetic studies (Douches et al. 1991; Endelman et al. 2017) attempted to elucidate this issue in the cultivated potato. In our previous study (Spanoghe et al. 2015), we investigated several approaches to both validate pedigree information and to infer parent candidates to orphan varieties using SSR data profiles. Our data showed that the simple exclusion method based on the comparison of individual SSR alleles between offspring and the assumed parents has proven to be complicated to implement due to the time required and known limitations. On the other hand, the maximum-likelihood method (LOD) was easier to implement and particularly useful to determine the most likely parent candidate(s) of any variety.

The objective of this study is to apply this inferential approach using a large amount of SSR marker data, with the aim of validating pedigree potato accessions in the dataset and of proposing parent candidate(s) for those whose pedigree information is either unknown, uncertain, or incorrect.

Materials and Methods

Genotype data were consolidated from a previous project (Spanoghe et al. 2022), which included 1249 distinct potato accessions, originating from different spatio-temporal groups, genotyped using 35 highly informative SSR markers. Pedigree information and date of release (between pre-1800 and 2021) were both registered for each accession in accordance with breeder records, the online pedigree database (van Berloo et al. 2007), and variety release publications. Accessions not fulfilling these two pieces of information were classified as orphans. An overview of the accession's background is available in the supplementary materials (Table S1).

The inferential maximum-likelihood method, as introduced by Meagher (1986) and previously investigated in the cultivated potato (Spanoghe et al. 2015), was used on the entire data set ($n=1249$). Formally, the method involved calculating a logarithm of the likelihood ratio (LOD) to compare the likelihood of different relationships for each possible parent, which provided either a null, positive, or negative likelihood parental score (LOD score) (Meagher and Thompson 1986). Due to the polysomic inheritance of the potato and consequently to the constraints of the specific locus dosage during the scoring of polymorphic fragments, we used the likelihood ratio statistics developed to analyse dominant markers (Gerber et al. 2000). For each of the 1249 accessions, parental assignment was performed on a ranking of the most probable varieties according to decreasing LOD scores. The formulas of the likelihood ratio were written in PHP language (see supplementary material file 1 for details) to automate the process for each parent/offspring combination. The pedigree validation procedure was then carried out for varieties whose genetic data of at least one parent was available in the dataset. The pedigree was validated when one or both parent(s) were found in the top 5 of the ranking (out of 1249); otherwise, a "pedigree conflict" was declared and a pedigree reconstruction was investigated. To increase the probability of ranking the true parent(s) in the top position, full sibs were removed from the ranking because of the interfering effects of family structure on the assignment (Jones and Ardren 2003). Moreover, only candidate parents created before the hybrid were displayed in the parental ranking. In addition, additive relationship coefficient, denoted by A, was computed from the pedigree by the tabular method (Henderson 1976) using the nadiv package in R software (Wolak 2012) (The script is provided in a supplementary file 2).

Results

The maximum-likelihood method was conducted on the entire data set to determine the most likely parent candidates for each accession (Table S2). The frequency distribution of all parent/offspring combinations followed a Gaussian curve where the

mean estimated LOD score was close to 0. The LOD scores ranged from a minimum of −44 to a maximum of +75. Parental ranking was first examined among accessions for which both parents were available in the dataset. Of the 1249 assessments, 251 parent–offspring trios (two parents and one offspring) in the dataset answered to this criterion (20.1%). Pedigree records were validated for 205 trios (81.7%), while for the remaining 46 trios (18.3%), a "pedigree conflict" was declared. Among the pedigree conflicts, 78.3% consisted of only one parent absent from the top 5 ranking (pedigree conflict type 1), while 21.7% consisted of the case where neither of the parents appeared in the top 5 ranking (pedigree conflict type 2). All the outcome scenarios were summarised in Fig. 1. Table 1 shows three examples of outcome scenarios. Details of all the results are available in the supplementary material (Table S3). Then, parental ranking was examined among accessions for which only one parent was available in the dataset. Of the 1249 assessments, 530 single-parental records answered this criterion (42.4%). Pedigree record was validated for 486 single-parental records (91.7%), while for the remaining 44 assessments (8.3%), a "pedigree conflict" was declared. All the outcome scenarios were summarised in Fig. 1. Table 2 shows two examples of outcome scenarios. Details of all the results are available in the supplementary material (Table S4). In total, mistakes on the male side are as frequent (50.6%) as mistakes on the female side (49.4%). To reconstruct the pedigree in case of non-validated assessment or lacking information, the highest ranked individuals were proposed as the most plausible parents if the true parents were in the dataset.

To investigate this matter further, the LOD score assessed between all other individuals in the dataset was plotted against the additive relationship coefficient (A) calculated from the pedigree, according to an adapted methodology of Endelman et al. (2017). For an individual without pedigree errors, any offspring should be in the top right corner of the figure with an additive relationship of 1 and the parents in the top centre corner with an additive relationship of 0.5 (in the absence of inbreeding) and high LOD score based on the markers. Figure 2 shows such a plot for clone KOMEET, which is a breeding line with recorded parentage EIGENHEIMER x
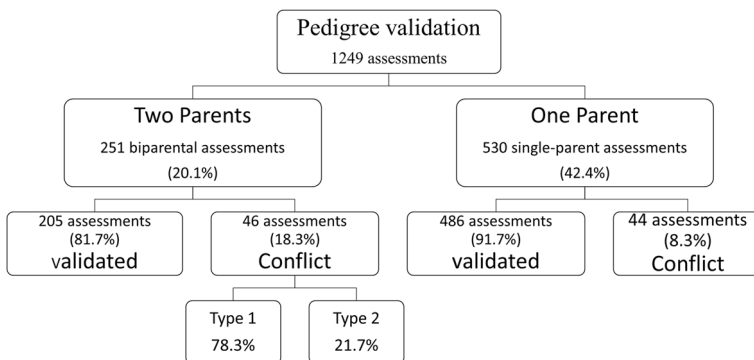

Fig. 1 Schematic view of the outcome scenarios and the results obtained during the pedigree validation procedure

Table 1 Three examples of outcome scenarios encountered during the pedigree validation step carried out among 251 parent–offspring trios for which both recorded parents were genotyped in the dataset. The pedigree was validated when both parents were found in the top 5 of the LOD ranking (Retrieved parent is emphasized in bold, and LOD score noted between brackets); otherwise, a "pedigree conflict" of type 1 or 2 was declared (number of conflict records is noted between brackets) and a parental reconstruction was proposed

Name	Date	Parent 1	Parent 2	Top 5 ranking (LOD score)	Results	Reconstruction
ACKERSEGEN	1929	HINDENBURG	ALLER-FRUHESTE GELBE	**ALLERFRUHESTE GELBE (27)**; **HINDENBURG (22)**; JUBEL (16); ARRAN VICTORY (13); TRIUMF (12)	Validated	n/a
DITTA	1989	BINTJE	QUARTA	NICOLA (26); **QUARTA (22)**; BRITTA (12); VALETTA (12); INDIRA (12)	Conflict (1)	(1) NICOLA
ELECTRE	1955	EIGENHEIMER	AQUILA	**ERDGOLD (24)**; **INDUSTRIE (23)**; JUBEL (22); BELLE DE LOCRONAN (20); JOELLE (19)	Conflict (2)	(1) ERDGOLD; (2) INDUSTRIE

Table 2 Two examples of outcome scenarios encountered during the pedigree validation step carried out among 530 single-parental records for which one recorded parent was genotyped in the dataset. The pedigree was validated when the parent was found in the top 5 of the LOD ranking (Retrieved parent is emphasized in bold, and LOD score noted between brackets); otherwise, a "pedigree conflict" was declared (number of conflict records is noted between brackets) and a parental reconstruction was proposed

Name	Date	Parent 1	Parent 2	Top 5 ranking (LOD score)	Results	Reconstruction
NICOLA	1973	CLIVIA	6430/1011	CLIVIA (29); SPERBER (12); FEUERGOLD (10); FORTUNA (9); RONDA (9)	Validated	n/a
NINETYFOLD	1897	WHITE ELEPHANT	PATERSON'S VICTORIA	UP TO DATE (19); EARLY ROSE (16); MAY QUEEN (15); RURAL NEW YORKER 2 (13); MYATT'S ASHLEAF (12)	Conflict (1)	UP TO DATE

FURORE. The four recorded grandparents for KOMEET are BLAUE RIESSEN, FRANSEN, RODE STAR, and ALPHA, which are plotted at $A = 0.25$ and are more distant than the parents based on LOD score. Varieties with high LOD score but low additive relationship, such as PINK FIR APPLE and ROSA (i.e. GAUMAISE is a self-fertilisation of Rosa), constitute orphan varieties having an unexpected parentage with KOMEET. When there is no pedigree error, as in this case, the overall appearance is a cloud of points with positive slope, stretching from unrelated individuals at $A = 0$ to the parents. ORION, an offspring of KOMEET, is displayed at an additive relationship of 0.5, while half siblings of KOMEET are displayed at an additive relationship of 0.25.

Figure 3a shows the plot for DITTA, which presented a pedigree conflict of type 1, with its recorded parentage BINTJE x QUARTA. The recorded female parent QUARTA is in the top centre of the figure with an additive relationship of 0.5 as expected, and recorded progenies of QUARTA with an additive relationship of 0.25. However, the recorded male parent BINTJE is too genetically distant (LOD score $= -4$) to be the true parent. One explanation is that the genotype for BINTJE is incorrect, but this was excluded because the pedigree of BINTJE was indeed validated during the previous step of pedigree validation. NICOLA, presenting the highest LOD score (26), seems to be the true parent (Fig. 3b). MISS ANDES and MISS MIGNONNE are thus full-sibs of DITTA. When using the simple exclusion method to support this statement, six segregation mismatches were found between the alleged parents BINTJE x QUARTA and their offspring DITTA, whereas no segregation mismatches were found between the reconstructed parents NICOLA x QUARTA and DITTA (Table S5).
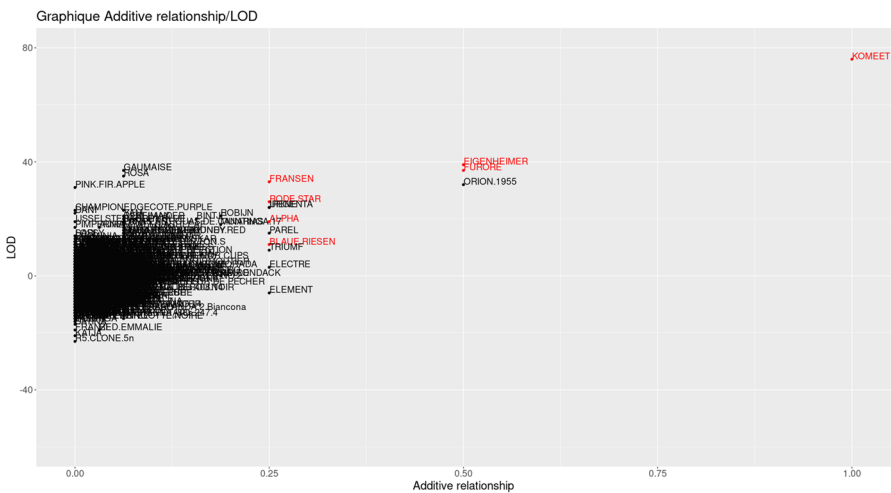

Fig. 2 Population-wide comparison of the LOD calculated from markers with the additive relationship calculated from pedigree records, for KOMEET with the recorded pedigree (EIGENHEIMER x FURORE). KOMEET's offspring (i.e. ORION) is displayed closely at $A = 0.5$. For legibility, affiliation of KOMEET is displayed in red
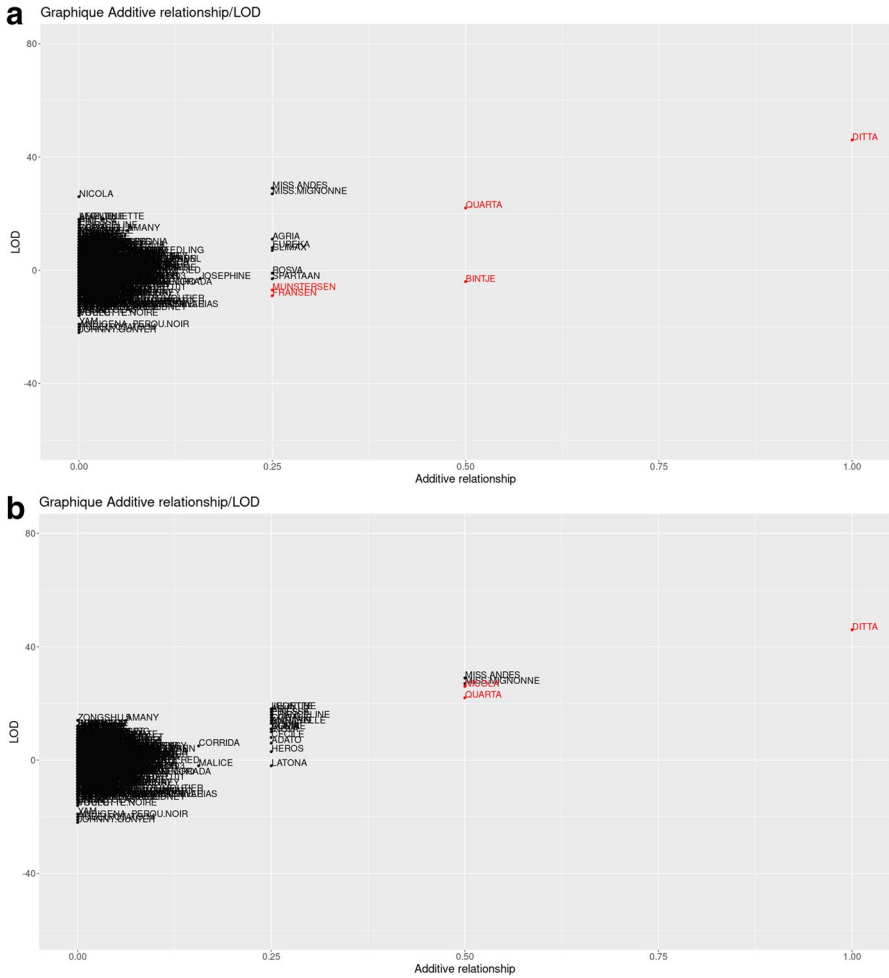
Fig. 3 LOD vs. additive relationship for DITTA with **a** the recorded pedigree (BINTJE x QUARTA), **b** based on the hypothesis that NICOLA is the male parent of DITTA. For legibility, affiliation of DITTA is displayed in red

Discussion

Validation and reconstruction of pedigrees are necessary for enhanced breeding programmes and genetic studies. Such information facilitates the estimation of reliable heritability, the maximisation of genetic gain, and the design of an effective breeding programme. It also ascertains the genetic identity of mislabelled genotypes used in breeding programmes.

In this study, we noted a high proportion of correct pedigree based on both the available information and the presence of hybrid and the parent(s) in the dataset, which means that the tested individual is mostly the expected one. However, a

non-negligible proportion of genotypes did not meet these achievements, questioning the authenticity of the variety (i.e. referring to our DNA isolates). We thus faced genotypes that were not the expected progeny of the alleged parents (and vice versa), resulting in erroneous pedigree information. This concerns the assessments carried out both on biparental individuals with pedigree conflict of type 2 and on single-parent individuals with pedigree conflict of type 1. Unintentional errors in pollination, seed harvest, or labelling as well as an accidental mixing in the breeder's collection can all generate such pedigree conflicts. Despite all the precautions taken, sampling, genotyping, or handling errors could also explain such outcomes. The ten cases where both parents are conflicting can support the assumption that the authenticity progeny is doubtful. It is the same for the forty-four cases where one parent is conflicting. In this case, new DNA isolates from an examination office that performed the DUS test are to be expected. This would allow the error to be attributed to this work or to prior work. Regarding the evaluations carried out on biparental individuals with pedigree conflict of type 1, we cannot establish the same interpretations as below since the individual tested may be the one expected because one of the parents is confirmed. However, the full pedigree should be taken with caution as the other parent has not been confirmed yet. The conflict detected in DITTA's parentage points to the possibility of an erroneous genotype for the conflicting parent, BINTJE, which is ruled out due to the validation of BINTJE's pedigree in a previous step. Indeed, the validation of BINTJE as a parent was successful for CLIMAX, ROSVA, EUREKA, and SPARTAAN. However, a pedigree conflict has been detected for one of BINTJE's two supposed parents, MUNSTERSEN, which has a LOD score of -12. A new DNA isolate from this conflicting parent should be retested.

Although the LOD method gave high results in the ranking criterion, the chance of finding the putative parent in the first position is recurrent but not always achieved. If the position of a parent candidate is not in the top 5 of the ranking, it is not considered as the true parent. It is assumed that the ranking of the top 5 is an arbitrary threshold above which all the parents were found for the exact pedigrees and below which the parents are too far apart in the ranking to be exact. This principle therefore makes the method very robust. Nevertheless, the presence of siblings that may be either genetically close to the analysed variety or to the inferred parent explains the poorer performance in the ranking. They were thus considered as interfering candidates. Indeed, the presence of other family members in the pool of parent candidates poses a serious challenge to parentage assessment since it may disrupt the parental ranking (Jones and Ardren 2003). Another aspect of the ranking examination revealed a less efficient ranking for many breeds that have a parent widely used in breeding such as Agria, Desiree, Early Rose, Nicola, and Katahdin. Consequently, their breeds tended to cluster together, also disrupting the ranking. By removing from the ranking the siblings and the varieties created after the one under study, the ranking score was greatly improved so that the probability of finding this putative parent in the first five of the ranking was systematic, in the event of a correct pedigree.

Some erroneous tests were found from the analyses of pedigree validation, and new parent(s) from our dataset was proposed. This parent reconstruction step can be especially useful when the real parent is in the dataset. However, reconstructed

parent(s) may not necessarily be suitable candidates because the dataset may not contain the true parent(s). Moreover, the plausibility of the proposed parents must satisfy several conditions, such as (1) the parent could have been available to the breeder of the descendant variety; (2) the parents are known to be fertile. Otherwise, these proposals should be used with caution, although the proposed candidate is certainly more plausible than the rejected one. However, we noted a logical tendency for closely related candidates (e.g. siblings) to cluster in the proposed parent(s), which complicates the identification of the true parent(s). If the proposed parent is not in the dataset, we can hypothesise that it remains close to the real parent. For these cases, a re-examination is to be proposed to validate or invalidate the proposed candidates, for example by means of more sophisticated technology (Endelman et al. 2017).

In conclusion, the LOD method constitutes a powerful tool for assigning parents, especially if known sibling examination is performed in complement. This step allowed increasing the probability to rank the true parent in the first position by rejecting interfering candidates. The few pedigree conflicts observed were met when the genetic fingerprint of the tested individuals was different from that which was expected. This leads to an error in the publication of the new variety's pedigree, to erroneous origins, and failures reflected in the scheme of varietal selection. The implementation of such a practice could have considerable impacts on the field and on the market, in terms of precision, savings, and time.

Declarations

References

Baldwin SJ, Dodds KG, Auvray B, Genet RA, Macknight RC, Jacobs JME (2011) Association mapping of cold-induced sweetening in potato using historical phenotypic data. Ann Appl Biol 158:248–256. https://doi.org/10.1111/j.1744-7348.2011.00459.x

Demeke T, Lynch DR, Kawchuk LM, Kozub GC, Armstrong JD (1996) Genetic diversity of potato determined by random amplified polymorphic DNA analysis. Plant Cell Rep 15:662–667. https://doi.org/10.1007/BF00231920

Douches DS, Ludlam K, Freyre R (1991) Isozyme and plastid DNA assessment of pedigrees of nineteenth century potato cultivars. Theor Appl Genet 82:195–200. https://doi.org/10.1007/BF00226213

Endelman JB, Schmitz Carley CA, Douches DS et al (2017) Pedigree reconstruction with genome-wide markers in potato. Am J Potato Res 94:184–190. https://doi.org/10.1007/s12230-016-9556-y

Endelman JB, Carley CAS, Bethke PC et al (2018) Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. Genetics 209:77–87. https://doi.org/10.1534/genetics.118.300685

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol Ecol 9:1037–1048. https://doi.org/10.1046/j.1365-294x.2000.00961.x

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83. https://doi.org/10.2307/2529339

Hosaka K, Mori M, Ogawa K (1994) Genetic relationships of Japanese potato cultivars assessed by RAPD analysis. Am Pot J 71:535–546. https://doi.org/10.1007/BF02851325

Isenegger DA, Taylor PWJ, Ford R, Franz P, McGregor GR, Hutchinson JF (2001) DNA fingerprinting and genetic relationships of potato cultivars (Solanum tuberosum L.) commercially grown in Australia. Aust J Agric Res 52:911–918. https://doi.org/10.1071/AR00161

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. Mol Ecol 12:2511–2523. https://doi.org/10.1046/j.1365-294X.2003.01928.x

Li X, Xu J, Duan S, Bian C, Hu J, Shen H, Li G, Jin L (2018) Pedigree-Based Deciphering of Genome-Wide Conserved Patterns in an Elite Potato Parental Line. Front Plant Sci 9:690. https://doi.org/10.3389/fpls.2018.00690

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D (2000) Predicting parental genotypes and gene segregation for tetrasomic inheritance. Theor Appl Genet 100:1067–1073. https://doi.org/10.1007/s001220051388

Meagher TR (1986) Analysis of paternity within a natural population of Chamaelirium luteum. 1. Identification of most-likely male parents. Am Nat 128:199–215. https://doi.org/10.1086/284554

Meagher T, Thompson E (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theor Popul Biol 29:87–106. https://doi.org/10.1016/0040-5809(86)90006-7

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879–889. https://doi.org/10.1534/genetics.105.054932

Muñoz PR, Resende MFR Jr, Huber DA, Quesada T, Resende MDV, Neale DB, Wegrzyn JL, Kirst M, Peter GF (2014) Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. Crop Sci 54:1115–1123. https://doi.org/10.2135/cropsci2012.12.0673

Slater AT, Cogan NO, Forster JW, Hayes BJ, Daetwyler HD (2016) Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. Plant Genome 9(3). https://doi.org/10.3835/plantgenome2016.02.0021

Slater AT, Wilson GM, Cogan NOI, Forster JW, Hayes BJ (2014) Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theor Appl Genet 127:809–820. https://doi.org/10.1007/s00122-013-2258-7

Song YS, Hepting L, Schweizer G, Hartl L, Wenzel G, Schwarzfischer A (2005) Mapping of extreme resistance to PVY (Ry (sto)) on chromosome XII using anther-culture-derived primary dihaploid potato lines. Theor Appl Genet 111:879–887. https://doi.org/10.1007/s00122-005-0010-7

Song Y, Schwarzfischer A (2008) Development of STS markers for selection of extreme resistance (Rysto) to PVY and maternal pedigree analysis of extremely resistant cultivars. Am J Potato Res 85:159–170. https://doi.org/10.1007/s12230-008-9012-8

Sood S, Bhardwaj V, Kaushik SK, Sharma S (2020) Prediction based on estimated breeding values using genealogy for tuber yield and late blight resistance in auto-tetraploid potato (Solanum tuberosum L.). Heliyon 6(11):e05624. https://doi.org/10.1016/j.heliyon.2020.e05624

Spanoghe M, Marique T, Rivière J, Lanterbecq D, Gadenne M (2015) Investigation and development of potato parentage analysis methods using multiplexed SSR fingerprinting. Potato Res 58:43–65. https://doi.org/10.1007/s11540-014-9271-3

Spanoghe M, Marique T, Nirsha A, Esnault F, Lanterbecq D (2022) Genetic diversity trends in the cultivated potato: a spatiotemporal overview. Biology 11:604. https://doi.org/10.3390/biology11040604

Tarn TR, Tai GCC, De Long H, Murphy AM, Seabrook JEA (1992) Breeding potatoes for long-day, temperate climates. In: Janick J (ed) Plant breeding reviews. Wiley, New York, pp 217–332. https://doi.org/10.1002/9780470650363.ch8

Van Berloo R, Hutten RCB, Eck HJ, Visser RGF (2007) An online potato pedigree database resource. Potato Res 50:45–57. https://doi.org/10.1007/s11540-007-9028-3

Visscher PM, Woolliams JA, Smith D, Williams JL (2002) Estimation of pedigree errors in the UK dairy population using mi-crosatellite markers and the impact on selection. J Dairy Sci 85:2368–2375. https://doi.org/10.3168/jds.S0022-0302(02)74317-8

Wolak ME (2012) nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. Methods Ecol Evol 3:792–796. https://doi.org/10.1111/j.2041-210X.2012.00213.x