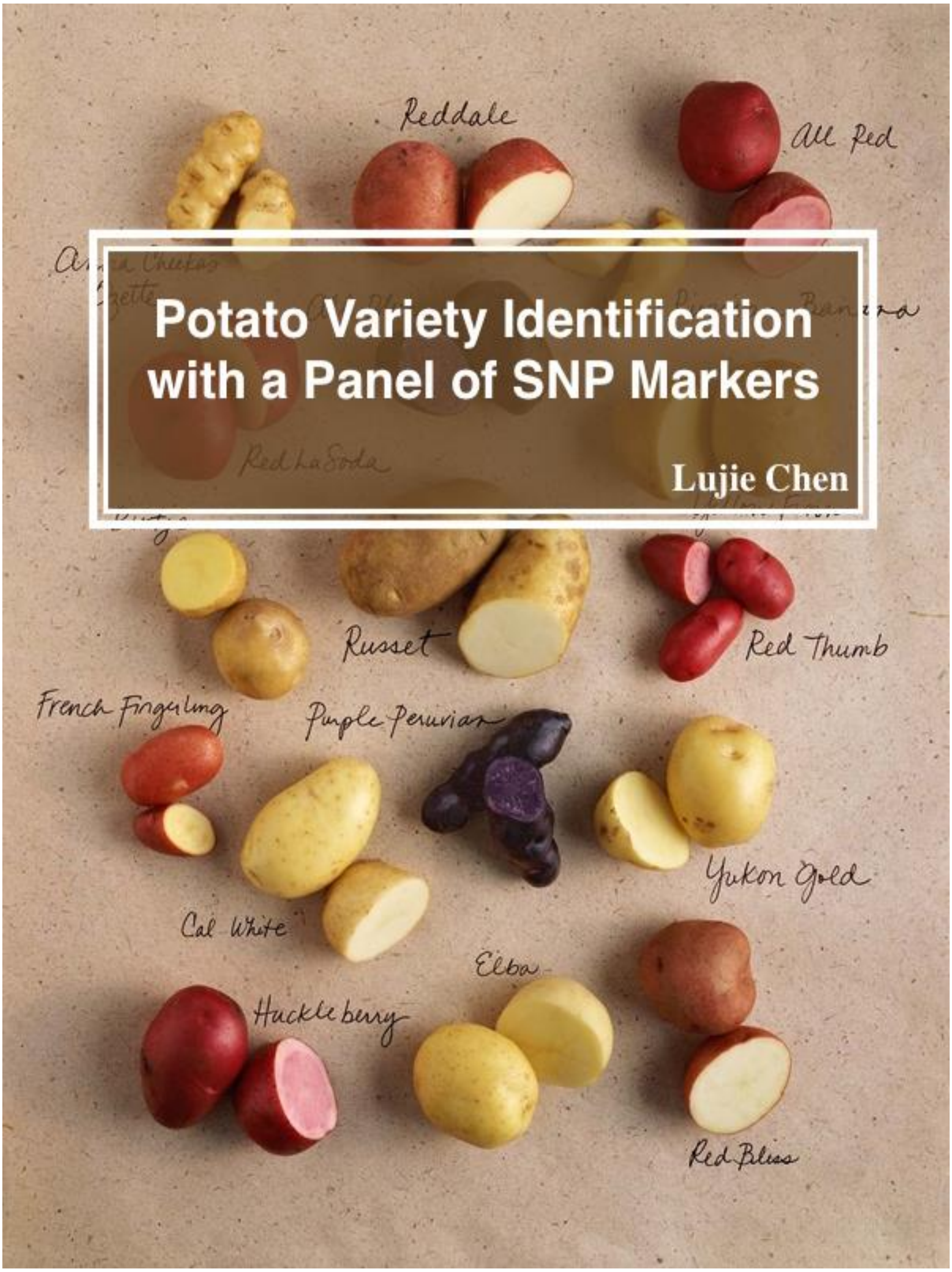
Potato Variety Identification with a Panel of SNP Markers

Lujie Chen



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Potato Variety Identification with a Panel of SNP Markers

Lujie Chen

911226157010

Supervisor: Dr.Ir. Herman J. van Eck

Research group: Biodiversity and Genetic Variation

Wageningen University,

Wageningen, the Netherlands

PBR-80436 • MSc Thesis • Plant Breeding

1st March, 2016

Abstract

From previous study, nine simple sequence repeats (SSR) markers (STM0019, STM2005, STM2028, STM3009, STM3012, STM3023, STM5136, STM5148 and SSR1) identified around 1000 potato varieties successfully. Compared to indirect PCR fragment comparison, using single nucleotide polymorphisms (SNP) markers with dosage estimation, differences between nucleotide and copy number can be revealed. The discriminatory power of SNP markers is influenced by population allele frequency and number of markers. In this study, we analysed 537 tetraploid potato varieties by SNP markers with dosage data. Those SNPs were filtered to five levels of reference allele frequency to investigate the effect of allele frequency for identification. We generated 10 different panel sizes of SNPs with the allele frequency from 45 to 55% to narrow down the effective number of SNP panel. Due to different data collection of SNPs and SSRs, we used two kinds of similarity coefficient (Jaccard's coefficient and Kosman's coefficient) to calculate similarity. Our results showed that a balanced allele frequency contributes to the highest discriminatory power. Moreover, when the panel size of SNP increased to larger than 50, mean value and variance of overall pairwise comparison between 537 varieties did not display a significant difference. The frequency distribution of pairwise comparison resembled a normal distribution. Lastly, compared SSR alleles with adjacent SNP alleles in 91 varieties, 4 SSR loci can find 5 to 24 SNPs with high similarity.

Key words: variety identification, single nucleotide polymorphisms (SNP), simple sequence repeat markers (SSR), similarity coefficient

Acknowledgement

At the very moment, I would like to show my deepest gratitude to my supervisor, Dr.Ir. Herman J. van Eck for his critical and patient supervisions on my thesis project. His skilful guidance and valuable suggestions helped me further improve on different aspects like critical thinking, experiment designing, data analysis and scientific writing. Great thanks to Johan Willemsen, for the help of making convenient software to analysis Kosman's coefficient. In addition, I also want to show my thanks to Dr. MJM (Rene) Smulders who offered the information of this project, which is a suitable project for me and I learnt a lot from it. Last but not least, I want to thank my family and friends for their continuous encouragement and support during my thesis project.

Contents:

1. Introduction

Potato (*Solanum tuberosum* L.) is an important food crop in Europe, originated from South America. According to current statistics, more than 376 million tonnes of potato were produced over the world in 2013, ranking to top 10 productive commodities (FAOSTAT, 2015). Besides, potato variety in the world is diverse. The overall amount of extant potato varieties is over 4500 worldwide (Pieterse and Judd, 2014). Meanwhile, 8537 named varieties can be found in Potato Pedigree Database 2013, including released cultivars and unreleased breeding genotypes (Van Berloo *et al*, 2007).

1.1 Distinctness test

Potato breeding requires plant breeders' right for protection to duplicate and essential derivation. To obtain plant breeder's right, the candidate varieties would be assessed the criteria of Distinctness, Uniformity and Stability (DUS), based on the guidelines set by The International Union for the Protection of New Varieties of Plants (UPOV). In addition, passing DUS test and the Value of Cultivation and Use (VCU) test are requirements for registration in the National List of varieties, which is a precondition for market release (Osman *et al*, 2015). VCU test declares that one or more important characteristics should be clearly distinguishable from other known varieties in the community (Council, 2002). Similarly, in the UPOV guideline of potato, the characteristics like the size of light sprout, the growth habit of plant and green colour of leaf, would be assessed to test distinctness (Anon, 2004). The assessments focus on morphological or physiological characteristics are difficult to measure and influenced easily by environmental effects (Moisan-Thiery *et al*, 2005). Due to environmental effects, the same variety growing in different locations showed considerable inconsistencies of some morphological descriptions (Reid *et al*, 2011). In contrast to morphological and physiological assessments, molecular markers are reliable and rapid tools for genotype identification via DNA fingerprinting. Therefore, molecular methods may assist the morphological approach of UPOV for new potato varieties distinctness test.

1.2 Molecular techniques for identification

In recent decades, an increasing number of studies have used molecular techniques to distinguish potato varieties. For instance, restriction fragment length polymorphism markers (RFLP) (Gebhardt *et al*, 1989; Görg *et al*, 1992), random amplified polymorphic DNAs (RAPDs) (Demeke *et al*, 1993; Hosaka *et al*, 1994) and inter simple sequence repeats polymerase chain reaction markers (ISSR) (Prevost and Wilkinson, 1999) were extensively applied in the end of 20th century. In comparison to the molecular markers mentioned above, simple sequence repeat markers (SSR) or microsatellites occupied a leading position for more than a decade in genetics research and breeding application (Mir *et al*, 2013). Moreover, SSRs have already effectively and efficiently identified hundreds of potato varieties (Ghislain *et al*, 2004; Moisan-Thiery *et al*, 2005; Reid and Kerr, 2007; Spanoghe *et al*, 2014). From the previous study of Reid et al. 2011, nine SSR markers (STM0019, STM2005, STM2028, STM3009, STM3012, STM3023, STM5136, STM5148 and SSR1) identified around 900 potato varieties successfully with a strong discriminatory power. The results of nine SSR markers identification show that 99.5% of the varieties have unique molecular profiles, while the rest 0.5% varieties have common ancestors in their pedigree and include some known mutants.

1.3 Comparison between SSRs and SNPs

SSRs have become the method of choice in identification studies due to their high allelic diversity and reproducibility. However, some drawbacks of SSRs are described below. Firstly, the data collection of SSR alleles does not contain information on allele dosage. Since the majority of cultivated potatoes is highly heterozygous auto-tetraploid (2n=4x=48), both the number of different allele and allele copy number should be considered for the genotyping analysis and variety distinction. Secondly, the variety discrimination by SSR markers lacks of reliability due to its indirect comparison of PCR fragment. Comparing PCR fragments by SSR markers cannot identify the size homoplasy, which is defined as copies of locus are identical in state but not identical by descent (Estoup *et al*, 2002). Lastly, drawbacks like low throughput due to the limitations in automation and data management, and requiring polyacrylamide gel electrophoresis for genotyping limit efficiency and cost effectiveness of SSR markers (Gonzaga *et al*, 2015; Guichoux *et al*, 2011).

Considering the constraints of SSR markers, single nucleotide polymorphisms (SNP) analysis, which has equal reproducibility, may be a substitution. One drawback of SNPs is lower allelic diversity, which means more SNP markers are required for the substitution of SSRs. SNPs have lower mutation rate, higher throughput and can be significantly less expensive than SSRs (Guichoux *et al*, 2011; Maughan *et al*, 2011; Smith and Maughan, 2015). A panel of SNP markers can be applied to validate genetic identity of working collection because of the low assay cost. Hence, in terms of economic feasibility and technical veracity, sooner or later, SNPs will substitute SSR marker analysis in variety identification.

1.4 Varity identification by SNP markers

SNPs are DNA sequence variants due to a nucleotide difference at a specific locus. They are the most abundant and smallest unit of genetic variation in eukaryotic genomes (Dou *et al*, 2012; Edward *et al*, 2008). SNPs have been discovered in many crops, such as soybean, maize and cassava, playing an undisputable role in the analysis of population structure, genetic diversity and association studies (Ching *et al*, 2002; Ferguson *et al*, 2012; Wu *et al*, 2010). Since the method of data capture of SNP genotyping platforms provides allele intensity, genotype calling can be scored into five different dosages (0, 1, 2, 3 and 4) by fitTetra (Voorrips *et al*, 2011) and Illumina GenomeStudio software (version 2010.3, Illumina, San Diego, CA). With the help of dosage estimation, dissimilarity comparison between potato varieties can be determined by one dosage difference. Therefore, the difference of present allele can be detected by comparing dosage 1, 2, 3 and 4.

To our knowledge, no relative study on potato variety identification by SNPs has been published yet. In a relevant study of barley, the ascertainment bias and optimal number of SNPs were mentioned as important issues for developing the panel of SNP markers (Moragues *et al*, 2010). Therefore, the ascertainment bias should be noticed when using SNP markers and the size of SNP panel should be narrowed down in economical aspect.

For the purpose of variety identification, population allele frequency of SNPs is a crucial factor because it affects the probability of each dosage. As bi-allelic SNP markers, the population allele frequency can be presented in two ways, (1) the major and minor allele frequency (MAJ/MAF) or (2) the alternative and reference (ALT/REF) allele frequency related to the reference genome. According to the studies done by Vos et al. (2015) the average MAF of SNP observed specifically in recent varieties is about 10 times smaller than the average MAF of SNP in varieties released before 1945 (Vos *et al*, 2015). The MAF value may also indicate the age of allele. However, using ALT/REF to represent the population allele frequency is more stable because the reference genome from potato clone DM is a stable dataset. In regard to variety identification, the reference allele frequency would be chosen to illustrate similarity based on same reference genome.

1.5 Suitable similarity coefficient for dosage data

Three coefficients (Dice, Jaccard and simple match coefficient) are commonly applied to measure similarity with binary data from molecular markers. In the study of SSRs, Jaccard's coefficient was chosen for the calculation of pairwise similarity matrix. These coefficients cannot handle dosage data from polyploids. Therefore we used the coefficient of Kosman &Leonard (2005), which is suitable for polyploid organisms. It is capable of comparing each allele from different genotypes (AAAA, AAAB, AABB, ABBB and BBBB). Through this method, every allele dosage would be compared rather than the absence or presence of an allele. If two compared allele dosages are the same, it would be counted as 1, otherwise it would be counted as 0. Therefore, the four alleles for each genotype would be compared and the sum of it would be divided by the ploidy level which is 4 in our case. Subsequently, the same comparison can be done among all loci. The sum of them would be divided by number of loci to get average similarity coefficient. In Table 1, all possible results of comparison between five genotypes are shown, while the genotypes were replaced by the dosage of reference gene.

Table 1 Genetic similarity between five dosages with Kosman's coefficient

Dosage	0	1	2	3	4
0	1	3/4	2/4	1/4	0
1	3/4	1	3/4	2/4	1/4
2	2/4	3/4	1	3/4	2/4
3	1/4	2/4	3/4	1	3/4
4	0	1/4	2/4	3/4	1

1.6 Research goals

In this study, we investigated the power of SNP markers with dosage estimation to identify potato varieties. The research objectives include (1) discovering how population allele frequency and size of SNP panel affect discriminatory power, (2) analysing similarity with known genetic relationship to set criteria for estimation of different variety or duplicate, (3) correlating nine SSR markers from the previous study with adjacent SNP markers to substitute SSRs for the distinctness test and (4) using the selected panel to test samples for validation.

2. Materials and Method

2.1 Plant materials

In this study, we used 537 tetraploid potato varieties described before (Vos *et al*, 2015). 48 varieties of them were released before 1945 and the oldest variety (Yam) was released in 1787. Among the 537 varieties, 192 of them were derived from the study of D'hoop et al. (2008), which could represent the commercial potato germplasm around world. Another set of 173 advanced breeding lines were added for subsequent studies (D'hoop *et al*, 2014; D'hoop *et al*, 2011). The rest varieties were provided by Dr. Ronald Hutten (Wageningen UR Plant Breeding) and the five companies Meijer B.V., HZPC, KWS, Averis and Agrico, including 51 cultivars and 120 advanced breeding lines. An additional set of 234 full sibs was analysed as well. This subpopulation was created from a cross between two tetraploid potato varieties, '*Altus*' and '*Colomba*' (Bourke *et al*, 2015).

2.2 SNPs data

In total, 14530 bi-allelic SNP markers were available, with five kinds of dosage (0, 1, 2, 3, and 4). Those SNP markers were successful SNP assays for the 20K SNP array, derived from the study of Vos et al. (2015). Within 14530 SNP markers, most (10707) of them were obtained from Uitdewilligen et al. (2013). In addition, this data also encompassed 3561 SNPs selected from the 8303 SolCAP array (Hamilton *et al*, 2011) and 202 SNPs from SolCAp 69K detection in Hamiliton et al. (2011). The rest of 32 SNPs were manually developed from candidate genes, related to morphological and disease resistance traits.

2.3 SSRs data

A set of 219 varieties with nine SSR markers (STM0019, STM2005, STM2028, STM3009, STM3012, STM3023, STM5136, STM5148 and SSR1) was collected from five sources. Three companies HZPC, Meijer B.V. and Averis provided 160, 12 and 12 varieties respectively. 32 varieties of them were derived from the study of Reid et al. (2011) and the rest 7 varieties were derived from UPOV (Reid, 2014). Within 219 varieties, 91 of them overlap with the in 537 varieties with SNPs data.

2.4 Experimental design and data analysis

In order to achieve research goals, we designed several experiments to test effects of allele frequency and panel size related to identification. Two kinds of similarity coefficient (Jaccard's and Kosman's) were used with two kinds of data (binary data and dosage data). For the effect of allele frequency, we analysed the relation between Polymorphism Information Content (PIC) and reference allele frequency for selection of SNP markers. The formula of PIC is $PIC = 1\text{-}\sum(P_i)^2$, where P_i is the frequency of the ith allele (Nei, 1973). For bi-allelic SNPs data, specifically, the formula can be rewritten as $PIC= 1\text{-} REF^2 \text{-} (1\text{-}REF)^{\ 2}$, here the *REF* represents reference allele frequency. When *REF* is closing to 50%, the PIC is reaching the highest value (0.5).

2.4.1 Experiment 1: allele frequency

In an attempt to investigate how reference allele frequency affected similarity coefficient, random samplings of 94 SNPs were selected by 5 levels of reference allele frequency (40-60%, 30-40%/60-70%, 20-30%/70-80%, 10-20%/80-90% and 0-10%/90-100%), which relate to the largest to smallest PIC value. 94 SNPs is a basic number of one standard micro-titre plate with two wells for no template controls. Then the pairwise comparison among 537 potato varieties (537*536/2=143916 pairs) was done, using coefficient from Kosman & Leonard 2005 by a customer design software. For each level of reference allele frequency, 5 replicates of 94 SNPs were generated to gain understanding of the stochasticity of SNP marker choice. Besides, the mean and variance were calculated to describe the influence caused by reference allele frequency.

2.4.2 Experiment 2: size of SNP panel

For discovering the optimal size of SNP panel, random samplings of 10 different sizes (10, 20, 30, 40, 50, 60, 70, 80, 90 and 100) was generated automatically with five replicates of each size by labelling every SNP marker with automatic numbers and sorting by those numbers in Excel. The SNPs were selected from around 500 SNPs with a reference allele frequency of 45-55%. After the calculation of pairwise similarity of different size of SNPs subsets was done, the mean and variance were calculated.

To describe the frequency distribution of similarity among 537 varieties, one sample of 100 SNPs was used to make a histogram and a normal curve was added as well. Meanwhile, the quantile-quantile plot (Q-Q plot) was plotted to confirm the assumption of normal distribution. Then 95% confidence interval of right side was calculated.

2.4.3 Experiment 3: effect of similarity index

The data selected from experiment 1 were converted from dosage to binary data for similarity comparison with Jaccard's coefficient. For the data conversion, firstly the raw dosage data of REF/ALT allele was converted to minor or major allele. It was done by symmetrically rearrange dosage data when reference allele frequency was larger than 50%. Then the minor allele (1, 2, 3 and 4) and major allele (0) was converted presence and absence (1 and 0) respectively.

The pairwise comparison within 537 varieties was done by DARwin 6.0 (http://darwin.cirad.fr), using Jaccard's coefficient. After comparisons between 5 different levels of reference allele frequency, the mean and variance of them were compared with dosage data, by using the same panel of SNPs to reveal the difference caused by coefficient.

2.4.4 Experiment 4: effect of panel size and similarity index

The data selected by 10 different sizes of random marker subsets in experiment 2 was converted to binary data. Then the similarity Jaccard's coefficient was calculated by DARwin. Mean and variance were calculated as well.

2.4.5 Control experiment: full sibs

The similarity between full sibs presented the worst case scenario for a SNP panel to make accurate differentiation. 234 offspring was analysed by 1144 SNP markers with reference allele frequency from 40% to 60%. The performance of it was compared to the similarity between two parents and three grandparents and one grand-grandparent. Besides, the ten random SNP panels with 100 and 50 SNPs from the experiment 2 and 566 SNPs with reference allele frequency from 45% to 55% were used to calculated similarity as well. The comparison between different panel sizes provided an understanding how panel size influence the similarity coefficient in a highly similar population. Besides, the theoretical similarity of full sibs was estimated from dosage combination of their parents. To evaluate the difference of similarity between full sibs with 537 varieties, 232 out of 537 varieties were randomly selected. The similarity between 232 random varieties together with two parents was analysed, using same 1144 SNPs. The outcome of varieties and outcome of full sibs from 1144 SNPs were used to make a scatter plot to visualize the similarity level between different varieties with full sibs.

2.5 Substitution of SSR alleles with SNP alleles

To replace alleles from known SSRs with SNPs, the equivalent SNPs were identified with the Kosman's coefficient. However, absence or presence of SSR alleles cannot be interpreted as dosage data directly. This bias, we calculated the frequency of each SSR allele (ploidy level/number of alleles per locus per variety), used as allele copy number for the comparison with SNPs. Meanwhile, the SNPs locating within 5Mb of SSR in the same chromosome were filtered from the entire 14530 SNP dataset. The Kosman similarity between SSR and SNP marker pairs was calculated, using the 91 common varieties. For each SSR allele, the SNPs with the highest similarity were selected and sorted by coordinate from smallest to largest. Furthermore, dosage data among 91 varieties of each marker were compared.

2.6 Real test with KASP assay

The data from Infinium assays were validated by real test with KASP assay. To select a subset of SNPs for KASP, several criteria were used to optimize this subset of SNP for validation. Firstly, 384 SNPs were filtered by the following standards: (1) reference allele frequency is between 40% to 60%, (2) allow primer design, (3) no high similarity among markers, (4) no adjacent markers within 0.5 Mb distance, (5) no more than 6 flanking SNPs and (6) no more than 2 SNPs per DMB super scaffold. Afterwards, 224 SNPs with 45-55% allele frequency and 160 SNPs with 40-45% and 55-60% allele frequency were retained. Then this subset was narrowed down to 100 SNPs by checking the even distribution in all chromosomes and distance between SNPs is far enough therefore one SNP selected per scaffold. In the end, 100 SNPs were select as a subset and sent to LGC Company for the KASP assay.

For the plant samples, 73 potato varieties were selected from 537 varieties. For 63 varieties used freeze dried samples were used due to the limitation of available fresh plant materials. Between the fresh and dry samples, 4 of dried samples had 2 replicates of fresh samples and 1 of them had 4 replicates (4*3 and 1*5). Within dry samples, two of them had a replicate of dried sample (2*2). Fortunately, 15 varieties had fresh samples. Among those samples, 7 of them were diploid to confirm whether the tetraploid varieties with dosage of 2 (duplex AAaa) coincide with diploid heterozygous genotypes (Aa). Within the fresh samples, two of them had four replicates (2*5), two of them had three replicates (2*4) and four of them had one replicates (4*2). Those replicates used same plant variety but with different quantity of leaf disk (1 leaf disk or 4 leaf disk) and different quality of leaf sample (old tissue or old and fresh tissue together). In total, 94 leaf samples and two empty wells for no template control was sent to the LGC Company.

3. Results

3.1 The effect of the SNP allele frequency

The outcomes of one replicate with Kosman's coefficient and Jaccard's coefficient were used to obtain an impression of the performance (Figure 1) of each of the estimators. In Figure 1, a scatterplot is shown of the sorted similarity values between 143916 pairs (537*536/2) of varieties as calculated by Jaccard's and Kosman's coefficient. The Jaccard and Kosman similarity value exhibit totally different performance. Firstly, the effect of allele frequency between each "curve" formed by points seems balanced in Figure 1B. However, in Figure 1A, the differences in similarity are increasingly upward biased when the allele frequency is more unbalanced (40-60%, 30-40%/60-70% and 20-30%/70-80%).

Then the ranking of five standards of allele frequency is opposite in two Figures 1A and 1B, and likewise the ranking of average similarity (Figure 2). When using Kosman's coefficient, the reference allele frequency is father away from 50% (e.g. 0-10%/90-100%), the average of similarity value is upwardly biased (close to 1). Therefore, markers with lower PIC (unbalanced allele frequency) can hardly identify the difference between varieties. This confirms the initial assumption that a threshold should be set on the allele frequency of the SNPs. However, when the same SNP dosage data were converted into binary data to allow Jaccard estimates, the result is different. The mean of all pairwise comparison decreases dramatically, when the reference allele frequency closes to unbalanced value like 0% and 100% (Figure 2).

Remarkably, the similarity estimates by Jaccard's coefficient show a very side range from 1 to 0 when using SNPs with an unbalanced allele frequency (0-10%/90-100%). Moreover, around 20000 variety pairs are completely dissimilar (Jaccard's coefficient is 0) for all 94 SNPs, as shown by the beginning part of trend (Figure 1B). After those 20000 pairs, the Jaccard's coefficient increased from 0 to 0.27, resulting in a noticeable break in the trend (Figure 1B). The percentage of present alleles is from 2.12% to 25.53% in the five replicates, when the allele frequency was set to 0-10%/90-100%. Therefore, a small section of alleles were comparable for the Jaccard coefficient, which calculated the difference between present alleles. One dosage difference among a small set of comparable markers resulted in a big increase immediately (Figure 1B).

In conclusion, the SNPs with balanced allele frequency (40-60%) provide the lowest average similarity, using the Kosman's coefficient. Moreover, the results of similarity achieved by Jaccard's and Kosman's coefficient are different. Compared to the Kosman's coefficient, Jaccard's coefficient with rare SNPs (0-10%/90-100%) showed a large variation of similarity value.

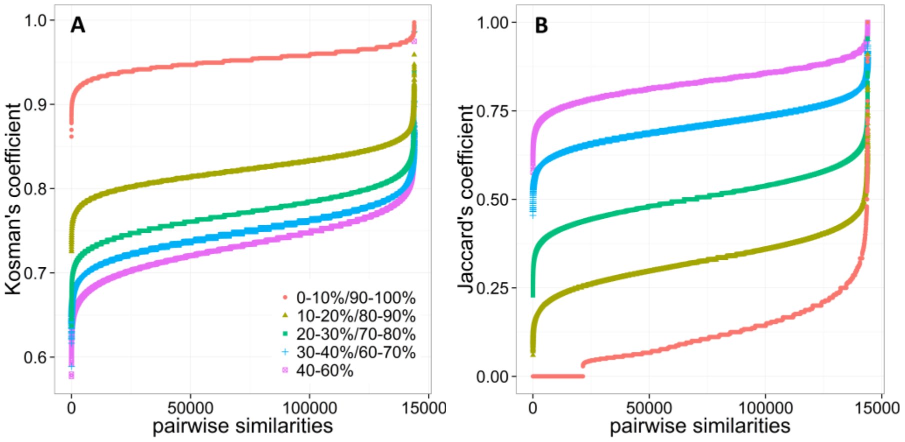


Figure 1 Scatter plots for two similarity coefficients: (A) Kosman & Leonard's coefficient and (B) Jaccard's coefficient. The estimates of the similarity coefficient were sorted from smallest to largest. 94 SNPs from five levels of reference allele frequency (40-60%, 30-40%/60-70%, 20-30%/70-80%, 10-20%/80-90% and 0-10%/90-100%) was used for pairwise comparisons among 537 potato varieties.

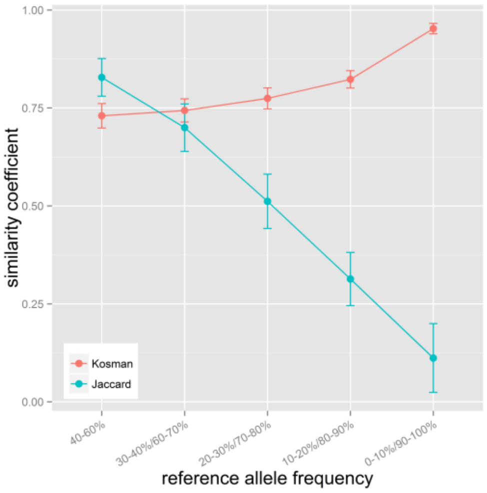


Figure 2 Dot plot for average similarity calculated by Kosman's and Jaccard's coefficient. Each dot represents the average similarity of five replications and error bar shows standard deviation within five replicates.

3.2 The effect of panel size

Using 10 different sizes of SNP panels with a reference allele frequency between 45% and 55%, both Jaccard's and Kosman's coefficient exhibited some similar features. First of all, the more SNP markers are applied, the smoother line is formed by points and the slope of those lines is closer to zero (Figure 3A and 3B). Secondly, many short horizontal lines emerge in the Figure 3A and 3B when the panel size is small, like 10 or 20. Those horizontal lines indicate that many pairs of varieties have an equal level of similarity, given by such a small subset of SNP markers. Subsequently, the average similarity can be estimated without bias when using panel sizes ranging from 40 to 100 SNPs, because the values hardly change (Figure 4A). Last but not least, the standard deviation decreases when the size of SNP panel increases, but it does not decreases anymore when the panel size is larger than 50 (Figure 4B). Thus a panel size exceeding 50 SNPs has not much added value anymore on the accuracy of the estimated similarity.

Focusing to the upper part of Figure 3A, only a few points have a Kosman similarity larger than 0.9 among 10 SNP panels of different sizes. Among all pairs, one variety pair shows 1.0 similarity, which is *Cardinal* with its mutant *Diamant*. Noticeably, the smallest SNP panel (10 SNPs) performed similarly in upper tail as 100 SNPs, using the Kosman's coefficient (Figure 3A). In the Figure 3B, more variety pairs reach a larger than 0.9 similarity, using the Jaccard coefficient. Meanwhile, a large amount of variety pairs showed 1.0 Jaccard similarity with 10 SNPs (Figure 3B). Based on the difference in the upper part of two figures (Figure 3A and 3B), it is concluded that the Kosman's coefficient is more sensitive to detect dissimilarity with a small size of SNP panel. This is easily explained, because the Kosman coefficient does take dosage differences into account as well, but the Jaccard coefficient ignores these subtle differences.

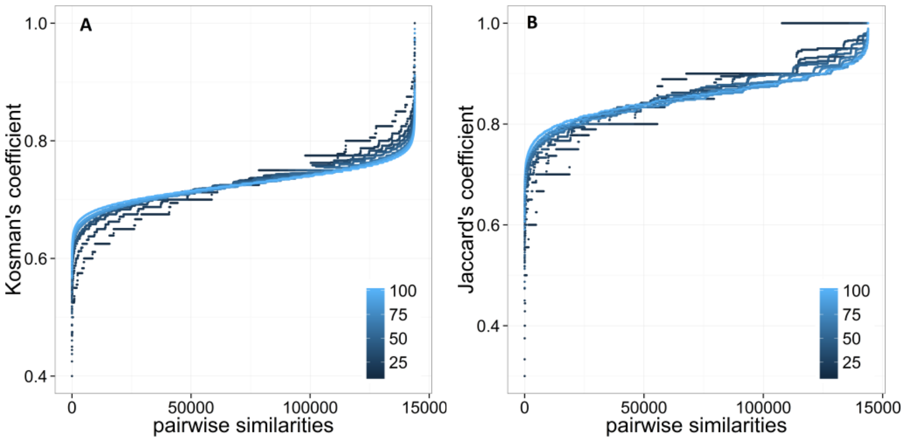


Figure 3 Scatter plots of two similarity coefficient: (A) Kosman & Leonard's coefficient and (B) Jaccard's coefficient. The similarity coefficient was sorted from smallest to largest. 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 SNPs (represented by dark to light blue colour) with reference allele frequency from 45% to 55% was used to do overall pairwise comparison.

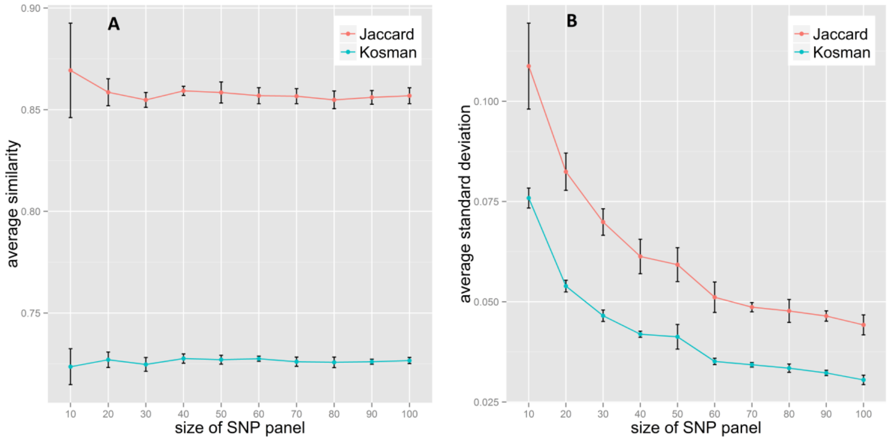


Figure 4 Dot plots of (A) average similarity and (B) standard deviation. Each dot was the mean value (A) and standard deviation (B) calculated from five replications of 10 different sizes of SNP panel (10, 20, 30, 40, 50, 60, 70, 80, 90 and 100), with the error bar of standard deviation of them.

3.3 Theoretical panel size offering sufficient resolution.

When absence/presence data are compared of an allele between two varieties, there are two kinds of results, same (0, 0 or 1, 1) or different (0, 1 or 1, 0). The frequency that two varieties have a same outcome (0, 0 or 1, 1) is 0.5 (0.5^2+0.5^2), when the frequency of presence is 0.5. Then the probability that 10 SNP markers are same within two varieties is 0.00098 (0.5^10), under the assumption that these 10 SNPs are independent.

With the help of dosage information, one bi-allelic SNP marker can identify five different kinds of allele dosages (0, 1, 2, 3 and 4). If the allele frequency is 0.5, the probabilities of variety having 0, 1, 2, 3 or 4 allele dosages are 0.0625, 0.25, 0.375, 0.25 and 0.0625 respectively. Then the probability of all possible comparison can be calculated (Table 2). The probability that a SNP will display a difference of 0, 1, 2, 3 or 4 dosages between two varieties are as follows: 0.27, 0.44, 0.22, 0.06 and 0.01, respectively. In analogy to power calculations using binary data, the frequency of 10 identical dosages is 2.06E-06 (0.27^10) which is 457 times lower than binary data. And if 10 SNPs are independent, the expected number of variety pairs that is indistinguishable by chance alone, is one pair from a panel of 986 varieties (986*985/2=4.86E+05=1/ (0.27^10)).

Table 2 All the possible results of similarity comparisons between five dosages with the probability of each result.

Dosage	**0**	**1**	**2**	**3**	**4**
Frequency	(0.0625)	(0.25)	(0.375)	(0.25)	(0.0625)
0	0	1	2	3	4
(0.0625)	(0.0625*0.0625)	(0.0625*0.25)	(0.0625*0.375)	(0.0625*0.25)	(0.0625*0.0625)
1	1	0	1	2	3
(0.25)	(0.25*0.0625)	(0.25*0.25)	(0.25*0.375)	(0.25*0.25)	(0.25*0.0625)
2	2	1	0	1	2
(0.375)	(0.375*0.0625)	(0.375*0.25)	(0.375*0.375)	(0.375*0.25)	(0.375*0.0625)
3	3	2	1	0	1
(0.25)	(0.25*0.0625)	(0.25*0.25)	(0.25*0.375)	(0.25*0.25)	(0.25*0.0625)
4	4	3	2	1	0
(0.0625)	(0.0625*0.0625)	(0.0625*0.25)	(0.0625*0.375)	(0.0625*0.25)	(0.0625*0.0625)

3.4 Frequency distribution of pairwise similarities

Based on the results regarding the effects of SNP allele frequency and SNP panel size, the balanced allele frequency and a panel size exceeding around 50 SNPs results in a stable mean similarity with a low standard deviation. Therefore, a robust variety identification can be achieved with around 50 SNPs. One replicate of similarity comparison based on 100 SNPs with a reference allele frequency from 45% to 55% was selected to analyse the frequency distribution of similarity values across pairs between 537 varieties. The shape of frequency distribution seems like a normal distribution, thus, a curve of a normal distribution was fitted and one threshold (coefficient is 0.77) separating the 5% highest similarity was added (Figure 5A). Besides, a Q-Q plot was made to confirm it (Figure 5C). As it shown in the Figure 5C, the majority of similarity values followed the normal distribution except the two tails of lowest and highest value. More outliers emerge in the end of range which is caused by pairs of related varieties in the dataset. By magnifying the figure, it is clear that few pairs get lower than 0.6 Kosman's similarity while three pairs are ranging from 0.9 to 1.0 (Figure 5B). Therefore, most of similarity value follows normal distribution, except some exceptional variety pairs.

Nevertheless, the shape of frequency distribution also depends on the diversity between materials. With the exception of Diamant-Cardinal, none of pairs have a similarity coefficient of 1, indicating that no other duplicate exists in the materials. Only 28 pairs (0.019%) of all comparisons generated a Kosman’s similarity from 0.85 to 0.97. The rest of comparisons (99.98%) have a similarity lower than 0.85 which is at least 60 dosages difference across 100 SNPs between each pair. The 60 dosage differences can be based on 15 SNPs which differ four dosages up to 60 SNPs with a single dosage difference.

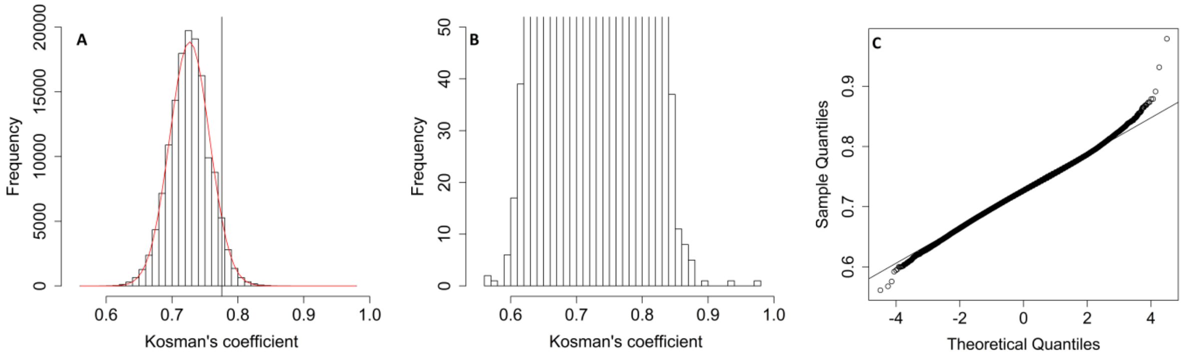


Figure 5 A: Frequency distribution of pairwise comparisons of 537 potato varieties (Kosman’s coefficient) **B:** Two tails of frequency distribution of pairwise comparisons (Kosman’s coefficient) from Figure 5A. **C:** Q-Q plot of similarity coefficient (Kosman’s coefficient) of overall pairwise comparisons.

By analysing the variety pairs with an exceptionally high similarity coefficient (in the right tail), more pairs were identified which have a related pedigree (siblings or parent-child). Across five replicates, the highest Kosman’s coefficient has been found for the same pair: *Diamant* and *Cardinal*. By searching the Potato Pedigree Database 2013, it was found that *Cardinal* is a mutant of *Diamant*. Besides, pairs *Armundo* and *Zafira*, *Early Rose* and *Russet Burbank*, *Spunta* and *Everest* and so on, have a high similarity around 0.85 to 0.89. Within those variety pairs, the former one is a parent of the latter one. *Liseta* and *Mondial* showed a high similarity in two replications (0.85 and 0.86), and from the pedigree database it turns out that they have same parents *Spunta* and *VE 66-295*. Unfortunately, there is no information in pedigree database for the four closest pairs which are RH4X-353-16 and RH4X-638-21 (from 0.86 to 0.90), RH4X-353-16 and RH4X-638-20 (from 0.87 to 0.89), RH4X-353-16 and RH4X-638-2 (from 0.86 to 0.87) and *Sutton’s Flourball* and *Shamcock* (from 0.87 to 0.91). Considering the high similarity between them, they might have some common ancestry in their pedigree. In the left tail, the lowest similarity, varieties *Bellel De Fontenay*, *Vitelotte Noire*, *Kepplestone Kidney*, *Shamrock* and *Ratte* are presented frequently. Those varieties show on average a lower similarity to the remainder of the varieties. Unfortunately, the parents of these heirloom varieites is unavailable from the pedigree database. *Vitelotte Noire* is the only variety with purple flesh and skin colour and thus represents a unique section of the gene pool.

3.5 Control experiments

Using all 1144 SNPs having a 40-60% reference allele frequency across 537 varieties, the pairwise similarity was calculated between 234 full sibs of the *Altus* x *Colomba* population. This control experiment therefore investigates material with an exceptionally high degree of similarity. The coefficient calculated by Kosman's and Jaccard's method was almost same, with the average of 0.81 and 0.82 respectively. This value is much higher than the mean value derived from 537 varieties (0.72). The observed high similarity follows our expectation. The distribution of similarity estimates of full sibs nicely followed normal distribution (Figure 6), and matched nicer as compared to distribution based on 537 varieties. The normal curve and Q-Q plot support the assumption of normality (Figure 6A and 6C). 80% of pairs (21821) have a similarity value in a range from 0.8 to 0.85. Only 218 (1.0%) out of 27261 pairs (234*233/2) reach a similarity above 0.85. Besides, the pairs with similarity over 0.87 occupied 0.01% of all pairs. However, when applying 100 and 50 SNPs selected from the previous experiments, 1250 pairs (4.6%) and 1918 pairs (7%) in average get a similarity above 0.85 respectively. More outlier values in two tails emerged when using a lower number of SNPs, leading to a decrease of amount of pairs which have average similarity (Figure 7). As executed, when the panel size expanded from 50 to 1176 SNPs, the mean similarity value increased from 0.8072 to 0.8133 and the standard deviation decreased from 0.0296 to 0.0154.

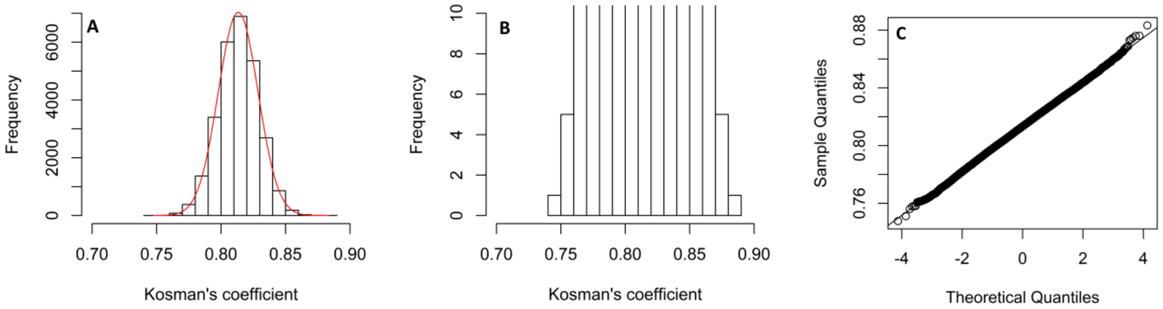


Figure 6 A: Frequency distribution of pairwise comparisons of 234 full sibs (Kosman's coefficient) **B:** Two tails of frequency distribution of pairwise comparisons (Kosman's coefficient) from Figure 6A. **C:** Q-Q plot of similarity coefficient of overall pairwise comparisons.

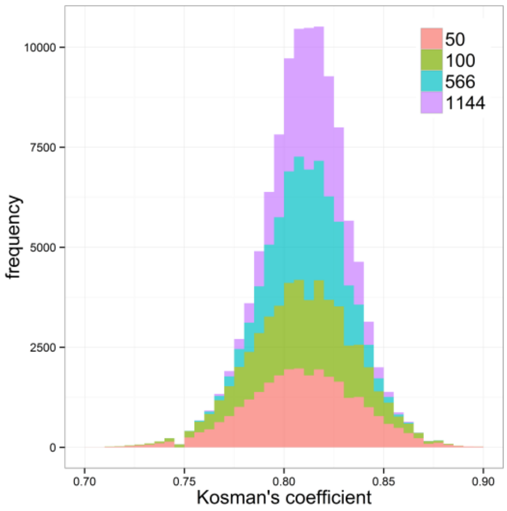


Figure 7 Frequency distribution of pairwise comparison of full sibs used 50, 100, 566 and 1144 SNPs.

The similarity between two parents ‘*Altus*’ and ‘*Colomba*’ (0.7261) is close to mean value of 537 varieties (0.7258) with 100 SNPs. Nevertheless, the average similarity of full sibs increases by 0.1 in comparison to variety pairs (Figure 8). The similarity between parents (*‘Altus’* and *‘Colomba’*) with offspring is around 0.81 which is close to the similarity among full sibs. When comparing grandparents with full sibs, two generation distance, the similarity decreased to 0.75. Likewise, the distance between grand-grandparent with parent is around 0.77. However, the similarity decreased to 0.7 for the pair that contains two varieties which have no close relation, for example in the pair of the parent of *‘Altus’* with variety ‘*Colomba’* (individuals from two lineages).

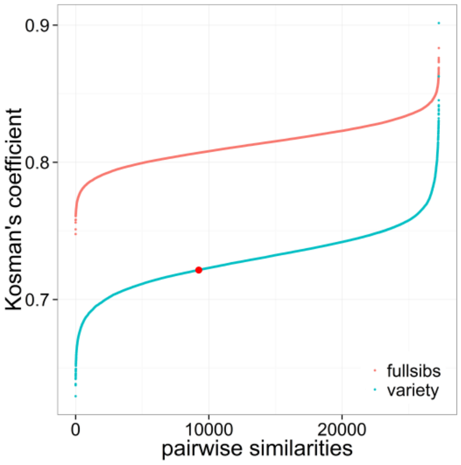


Figure 8 Scatter plot of Kosman’s coefficient. The similarity coefficient was sorted from smallest to largest of comparison between 234 full sibs with 234 random varieties. And the red dot represents the similarity between two parents of full sibs.

By analysing the genotype of the two parents at each of the 1144 SNP loci (1096 SNPs in reality, excluded the missing values) the Mendelian probabilities of five dosages of offspring were calculated (Table 3). For 215 SNPs (19.62%), ‘Colomba’ and ‘Altus’ showed a duplex x triplex genotype, followed by 210 SNPs (19.16%) with a simplex x duplex genotype (Table 4). According to Mendelian expectations, the probabilities of dosages 0 to 4 among offspring were 0.06, 0.24, 0.38, 0.27 and 0.06 respectively among 1096 SNPs. Then the proportions of dosage similarity could be calculated as well, which is 0.25, 0.42, 0.24, 0.08 and 0.01 corresponding to full, 3/4, 2/4, 1/4 and 0 dosage similarity. Therefore, the theoretical similarity of offspring was 0.705 (0.25+0.42*3/4+0.24*2/4+0.08/4), which is closer to similarity of their parents (0.7261) than the value calculated by 234 samples (0.8133).

Table 3 The allele dosage and frequency in full sibs result from a Mendelian prediction based on the allele dosage observed in the parents.

Dosage	0	1	2	3	4
0	0	0, 1	0, 1, 2	1, 2	2
	(1)	(1:1)	(1:4:1)	(1:1)	(1)
1	0, 1	0, 1, 2	0, 1, 2, 3	1, 2, 3	2,3
	(1:1)	(1:2:1)	(1:5:5:1)	(1:2:1)	(1:1)
2	0, 1, 2	0, 1, 2, 3	0, 1, 2, 3, 4	0, 1, 2, 3, 4	2, 3, 4
	(1:4:1)	(1:5:5:1)	(1:8:18:8:1)	(1:5:5:1)	(1:4:1)
3	1, 2	1, 2, 3	1, 2, 3, 4	2, 3, 4	3, 4
	(1:1)	(1:2:1)	(1:5:5:1)	(1:2:1)	(1:1)
4	2	2,3	2, 3, 4	3, 4	4
	(1)	(1:1)	(1:4:1)	(1:1)	(1)

Table 4 The observed frequencies in 1096 SNPs of two parents *'Altus'* and *'Colomba'*.

Parental Dosage	0	1	2	3	4
0	4	30	45	25	9
	0.36%	2.74%	4.11%	2.28%	0.82%
1		63	210	160	43
		5.75%	19.16%	14.60%	3.92%
2			144	215	45
			13.14%	19.62%	4.11%
3				72	26
				6.57%	2.37%
4					5
					0.46%

3.6 Threshold for the recognition of similar or different varieties

To distinguish two varieties, a threshold of similarity coefficient was proposed based on the comparison within varieties and full sibs. From the similarity analysis of full sibs, 99% similarity coefficient within full sibs is lower than 0.85. Therefore, in a highly similar population, the chance to get a 0.85 similarity coefficient is less than 1% (see Results 3.5). As to a mixture of different varieties, the chance to get a 0.85 similarity coefficient is less than 0.02% (see Results 3.4). When the similarity coefficient is larger than 0.85, the chance they are similar varieties is high and the pedigree of them should be similar. In the opposite side, if the similarity coefficient is lower than 0.85, they may not be same variety. Hence, the similarity coefficient of 0.85 was suggested as a threshold for the recognition of similar or different varieties.

3.7 Identification of SNPs in high LD with SSR alleles

Calculation of correlation of nine SSRs with all 14530 SNPs is computationally intensive. Therefore subsets of SNP data were retrieved from the full data using a window of 5 Mb surrounding the physical position of the SSR locus. As it shown in Figure 8, the quantity of SNPs locating within 5Mb of nine SSRs (STM5136, STM3023, STM5148, STM0019, STM3009, SSR1, STM3012, STM2005 and STM2028) and their reference allele frequency varied from chromosome to chromosome. In chromosome 1 the SSR is located in a SNP dense region and the amount of adjacent SNPs is the largest (410 SNPs, Table 4). The allele frequency of those SNPs is shown on the y-axis (Figure 8). For SSR (STM3023) on chromosome 4 only a small number of SNPs were found in the flanking 5Mb region, and most SNPs show a rare allele frequency. It appeared that not the absolute number of SNPs in the 5Mb interval determined the success to identify SNPs in high linkage disequilibrium (LD) with a SSR allele. More important was the physical proximity, because no association between SNP and SSR-allele was found in cases where the most nearby SNPs were more distant than 50kb. This suggests that on chromosome 4, 6, 7, 9 and 12 decay of LD already took place within 50kb to such a level that no strong associations between SNPs and SSR alleles can be expected. Very short physical distance between SNPs and SSR alleles, were found for chromosome 1 and 8, which contain SNPs within 100 bp (Table 4). In Table 4, 1 to 3 SSR alleles of STM5136, STM5148, SSR1 and STM2005 can find some similar SNPs. Those SSRs have closer SNPs compared to other SSRs.

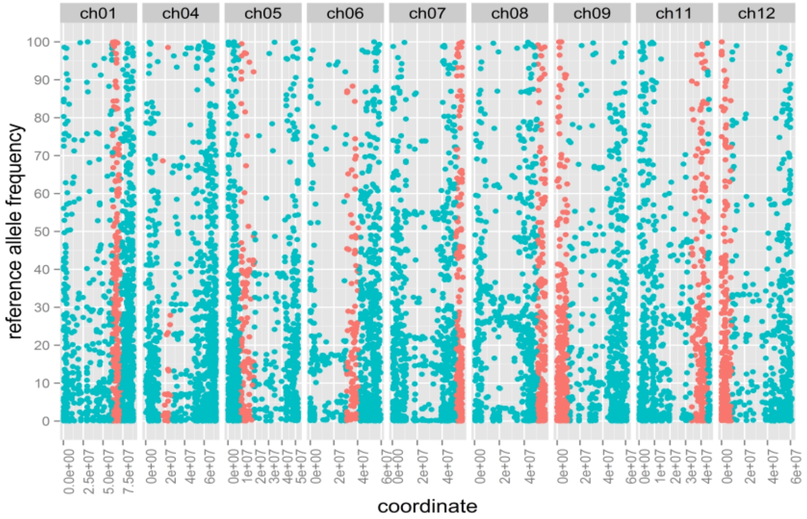


Figure 8 Scatter plot of reference allele frequency of all the SNPs in the nine chromosomes (chromosome 1, 4, 5, 6, 7, 8, 9, 11 and 12). The red dots represent the SNPs within 5Mb of nine SSRs (STM5136, STM3023, STM5148, STM0019, STM3009, SSR1, STM3012, STM2005 and STM2028).

Table 4 Results of comparison between SNPs and SSRs in 91 common varieties.

STM marker	chromosome	No. of adjacent SNPs (5Mb)	The closest distance of SNP marker	No. of SSR alleles	No. of SSR alleles found similar SNPs	No. of SNPs with high similarity*
STM5136	1	410	19	8	2	7
STM3023	4	36	245661	4	0	0
STM5148	5	179	11773	15	3	15
STM0019	6	163	131922	8	0	0
STM3009	7	251	130360	6	0	0
SSR1	8	206	99	10	3	24
STM3012	9	335	154180	6	0	0
STM2005	11	315	25829	5	1	5
STM2028	12	340	55165	9	0	0

* The Kosman's coefficient in the pairs of SNPs and SSR alleles is larger than 0.95, except the rare SSR alleles which get 0.95 similarity with absent SNPs.

The limited success to find SNPs associated with SSR is also due to the limited variety panel, where 29 out of 100 SSR alleles did not exist in these 91 varieties. For the remaining 71 alleles, the SSR allele dosage data was estimated phenotypically and thus with bias. Therefore, the allele frequency was biased downward when the number of different alleles was 3 or less. For the reason that the average allele dosage of 3 phenotypic alleles is 1.33 (4/3), thus, only 3 dosages can be assigned in this case (1:1:1) but the total dosage is 4. When the allele number was 2, the dosage assignment based on average copy number was 2:2. In conclusion, no dosage of 3 was in SSR data due to the bias.

Comparing 9 SSRs with adjacent SNPs respectively, the similarity between SSRs with their adjacent SNPs was influenced by the allele frequency of SSRs, which was calculated by the biased dosage estimation. When the allele frequency of SSR was higher than 30%, the Kosman's coefficient tended to be lower than 0.8. Meanwhile, when the allele frequency was lower than 10%, the Kosman's coefficient was almost larger than 0.9 or even equal to 1. The inaccuracy of Kosman coefficient with increasing allele frequency is explained as the result of greater bias in allele dosage estimates.

As to the SSR marker (STM5136) in chromosome 1, eight phenotypic alleles (A, B, C, D, E, F, G and H) existed in 91 varieties with allele frequencies from 0.3% to 23.4%. SNPs similar to the common alleles E and F (21% and 23% allele frequency) can be found in the closest position with similarity around 0.89. While the less common alleles C, H and D, the most similar SNPs (PotVar0098421, PotVar0098524 and PotVar0098499 respectively) are in the closest coordinates and allelic dosage are same in many varieties. However, there are no identical SNPs in this locus. The highest Kosman's coefficient among pairs with this SSR is 0.98, which was calculated by a pair of allele D with a SNP (PotVar0098499) in a short region (2004 bp). However, in 18.7% variety the allele D presented as a missing value, thus it is incomparable in those varieties. In terms of rare alleles (A, G and B), even the absent SNPs reached a high similarity in the pair with those alleles. Therefore, those rare alleles contain similar information with absent SNPs in 91 varieties.

In chromosome 5, 15 SSR alleles (A, B, C, and D ... P) of STM5148 are available in 91 varieties. One SNP (PotVar0104899) is highly similar (0.967) to allele I and "PotVar0104886" is highly similar to (0.972) allele P. Remarkably, the distances between them are larger than 1 Mb, suggesting that no decay of long range LD took place at this locus. Both the I and P allele have a relatively high allele frequency, which excludes to explain high LD values due to recent introduction of new alleles on a wild species introgression segment. And indeed "PotVar0104899 and PotVar0104886" belong to the group of pre-1945 SNPs (Vos et al., 2015). Another remarkable result is the observation of 13 SNPs with high similarity to allele O, but the closest distance between those SNPs and STM5148 is 11773 bp. Again, allele O is relatively common and the SNPs are not post-1945.

For SSR1, three SNPs (solcap_snp_c2_28480, solcap_snp_c2_28476 and PotVar0023147) locating within 4500 bp are highly similar to allele F and two SNPs (PotVar0023097 and solcap_snp_c1_8759) locating within 500 bp are highly similar to allele A. In addition, there are 13 SNPs have high similarity above 0.95 with allele J, while the most similar one (PotVar0023391) locating within 200,000 bp has 0.99 Kosman value with this allele. Similarly, the 5 SNPs (PotVar0113061, PotVar0112982, PotVar0112966, PotVar0112965 and PotVar0112780) which found highly similar to one STM2005 allele (F) are far away from STM2005, above 700,000bp.

In summary, with our current dataset and methods we successfully identified SNPs that could substitute SSR alleles, but it will not be easy to achieve this situation for many more SSR alleles. Resequencing of PCR amplicons flanking the SSRs may provide new data, but this is a laborious experiment.

3.8 Validation of samples with KASP assay

One hundred SNPs with allele frequency from 45% to 55% were converted into a KASP assay to analyse 94 DNA samples. Only 42 KASP assays provided a grouping that could be clustered to capture allele dosage. For these SNPs the dosage data was obtained with fitTetra software. It should be noted that the quality of 42 SNPs assay is not consistent and high in average. Only 14 SNPs showed five sharp and balanced clusters like PotVar0097245 in Figure 9. However, the rest 28 SNPs displayed mainly three types of less than ideal patterns.

1) Some SNPs show a skewed X/Y ratio, where X and Y is the intensity of fluorescence of either allele. The image of solcap_snp_c1_6036 is shown as an example (Figure 9B). Three clusters of heterozygous (ABBB, AABB and AAAB) diverge from the centre and are closer to one homozygous cluster (AAAA) than the other cluster (BBBB). Besides, the signal intensity of one nucleotide (A) is universally higher than the other. It happened mostly in solcap SNPs, due to the fact that 10 out of 13 SNPs formed this type are from solcap SNPs. 2) For 10 markers, the clusters are grouped but dispersed highly (Figure 9C). Hence, the dosage assigned for the variety between two clusters is not trustworthy. 3) The rest five markers gave a clear separation of five cluster and well distribution in the plot, but the signal intensity varied largely in each cluster (Figure 9D).

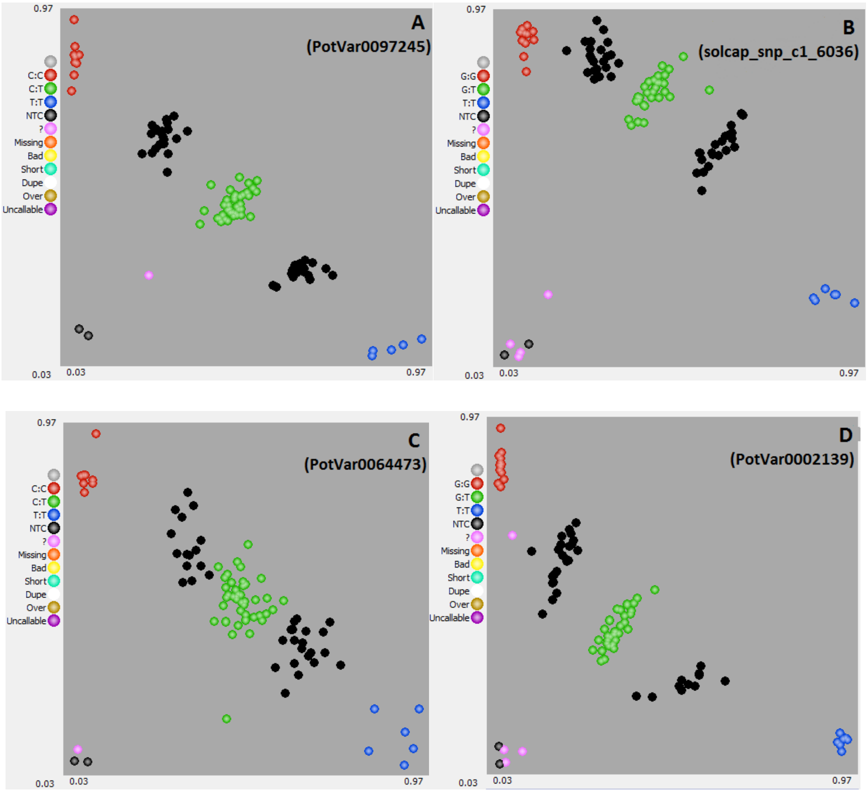


Figure 9: Genotype plot using KASP assay of four types of cluster assignment. Five clusters represent dosage 0, 1, 2, 3 and 4.

The quality of cluster grouping in the genotype plots vary from marker to marker. Nevertheless, compared dosage data from KASP assay with infinium data from Vos et al. (2015), more than 96% dosage data of KASP assay is consistent with infinium data excluding the missing values. From the result of clusters assignment, freeze dried samples generated high quality DNA as did the fresh samples. For the replicates within fresh samples, 9 differences were found. In addition, five varieties were replicated to compare between freeze dried and fresh tissue samples. For 4 pairs of the replicates, the level of inconsistency is only 2%. On the other hand, for the two replicates within freeze dried fresh (*Agria* and *Bintje*), 35% (59/ (42*4)) KASP assay differed from infinium data. Among the 73 potato samples, 7 of them are diploid and expected to get 0 and 4 for homozygous and 2 for heterozygous. Three data points showed unexpected dosage of 3.

In order to figure out the reason of enormous inconsistency between replication of dry samples and to validate the rest dry sample, we calculated the Kosman's coefficient between each sample with 537 varieties. Within 73 samples, three kinds of result were obtained, which are (1) largest similarity is lower than 0.85, (2) largest similarity is close to 1 in the pair with supposed variety and (3) largest similarity is close to 1 but in the pair with another unexpected variety.

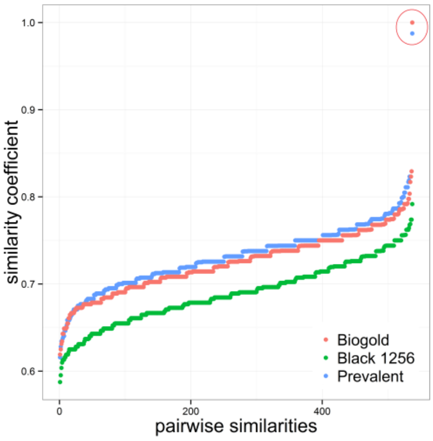


Figure 10: Scatter plot of similarity of three samples compared with 537 varieties. The red circle points out the two largest Kosman's coefficient.

As it is shown in the Figure 10, the 84 samples like *Biogold* are able to find expected variety with a Kosman's coefficient = 1.0, which indicates that there was not a single error in dosage estimates between Infinium and KASP assays, and both samples were correctly labelled as *Biogold*. The second best match between *Biogold* and another variety is reaching a Kosman value of 0.829.

Unexpectedly, for five varieties (*Prevalent*, *Gloria*, *Bartina*, *Triplo* and *Bintje*) their Infinium and KASP data matched poorly. For these pairs, a relative low similarity (0.69 to 0.75) was observed, but a high similarity with another variety (0.98 to 1.0). For example, the sample supposed to be *Prevalent*, reached 0.7439 similarity coefficient with *Prevalent*, which is close to the average similarity (0.7315) among 537 pairs. But the highest value is 0.98 with *Nomade*. Besides, two diploid samples, regarded as C and E, showed high similarity with varieties E and C respectively. These observations suggest that variety names were not correctly assigned to samples. Lastly, five samples (*Black 1256*, *Jaerla*, *Monalisa*, *Toyoshiro* and *Agria*) failed to find a similar variety, like *Black 1256* in the Figure 10. The highest similarity of those samples is lower than 0.85.

4. Discussions

4.1 Comparison of the Jaccard and Kosman similarity coefficient

In this study, two kinds of similarity coefficient were applied, which are Jaccard and Kosman similarity coefficient. Jaccard's coefficient was extensively used in pairwise comparison of SSRs (Reid *et al*, 2011; Spanoghe *et al*, 2014). Kosman's coefficient was adjusted from study of Kosman &Leonard (2005), which is specifically for the comparison of dosage data. As it shown in results 3.1, there are two obvious difference between the similarity comparisons which using Jaccard's and Kosman's coefficient.

Firstly, when the allele frequency was 0-10%/90-100%, the similarity of overall pairs was low and ranging for 0 to 1 using the Jaccard's coefficient. For the Kosman's coefficient, the comparison with rare SNP alleles got the highest average similarity.

The reason of this phenomenon is that the cases of two absent alleles in pairwise comparison were eliminated by Jaccard's coefficient. When using rare SNP alleles (0-10%/90-100%), a big section of SNPs (75% to 95%) was absent among 94 SNPs. Therefore, when using Jaccard's coefficient large amount of absent-absent cases are incomparable. If we used SSR markers, it is a good method. Because each SSR locus may have around 10 different SSR alleles. The absent SSR allele is unpredictable and incomparable. However, for bi-allelic SNP markers, there is only one possibility for absent allele, the alternative allele. If both of varieties show absence, representing both of them had a same alternative. Therefore, the cases excluded by Jaccard's coefficient are the same SNP alleles in two compared varieties. On the contrary, Kosman's coefficient regarded 0 to 0 cases as same SNP alleles and showed high similarity in the rare SNP alleles. In this aspect, Kosman's coefficient provided a more suitable comparison for SNP markers.

Except the comparison between rare alleles, the effect of reference allele frequency differed from Jaccard and Kosman similarity coefficient. From the results 3.1, the ranking of five levels of allele frequency is opposite in two coefficient.

It was caused by the data converting. The raw data was dosage data which used directly for the comparison by Kosman's coefficient. To calculate Jaccard's coefficient, the raw data was converted to binary data. After data converting, dosage 1, 2, 3 and 4 (reference allele is minor allele) or 0, 1, 2 and 3 (reference allele is major allele) was replaced by 1. Then the dissimilarity between presence alleles disappeared after data converting. Moreover, when the reference allele frequency was set close to 50%, in binary data around 0.9375 (1-0.5^4) of the allele is present. The high percentage of presence increases the similarity among varieties. Whereas, when the allele frequency is further away from the 0.5, the ratio of presence and absence is closer to 0.5. Then more dissimilarity between presence and absence results in a lower mean of similarity coefficient. Therefore, the data conversion from dosage data to binary data decreased the dosage dissimilarity in the present alleles.

In conclusion, Kosman's coefficient is more suitable than Jaccard's coefficient, for SNP markers with dosage data.

4.2 SNP ascertainment bias

The selection of SNPs with a particular purpose influences its comprehensive application in different target groups, and a careful selection should maximize the information content of different groups (Thomson *et al*, 2012). For this study, the SNP data are derived from 20k SNP array (Vos *et al*, 2015), which included varieties from different studies with different purposes. Therefore it decreases the bias of selection. Besides of biased selection, choosing SNP alleles with high frequency in the discovery process introduced ascertainment bias (Nielsen, 2000). Moreover, the high frequency SNPs are more polymorphic, which leading to an underestimation of genetic diversity (Moragues *et al*, 2010). However, nearly half of SNP markers in our data have a low allele frequency from 0 to 10% or from 90 to 100% among 537 potato varieties.

In this study, we selected a subset of SNP markers to identify potato variety based on a limited amount of individuals and a certain region of population allele frequency predominately. Hence, questions might come out: (1) whether the SNP panel selected by 537 varieties can be applied for other varieties identification or a more distantly related gene pool such as landraces from Latin America, (2) whether the common alleles derived from founding fathers tell the difference between recent released varieties. According to the results of Vos et al. (2015), the SNPs discovered only in recent varieties do not show a high allele frequency. The allele frequency of pre-1945 SNPs (Vos et al. 2015) is relatively stable and hardly influenced by selective breeding after a century. Hence, the subset we selected by PIC value represents SNPs in old founding fathers. Besides, its allele frequency may not differ too much amongst different potato samples.

4.3 Optimal size of SNP panel

In the study of barley, a 384 SNP subset was proposed for a good combination of power and economy for the germplasm identification (Moragues *et al*, 2010). However, from the theoretical estimation, we concluded that 10 independent SNP markers are already able to distinguish between approximate 1000 varieties. In reality, the results from experiments with 10 SNPs lack stability compared to experiments with higher number of SNPs (20, 30, 40, 50, 60, 70, 80, 90 and 100). It also has the highest standard deviation within 5 replicates and between 5 replicates. Under this condition, the similarity comparison with 10 SNPs is not reliable and easily influenced by the choice of markers easily. Therefore, the number should increase for a stable comparison. Based on the results of increasingly larger SNP panels, the average similarity is stabilized at panel sizes of 40 to 100 SNPs. Besides, the standard deviation decreases slowly since 60 SNPs had been applied. In a word, to identify potato variety, the number of SNPs in the panel can be narrowed down to approximate 48 SNPs and reduce the cost for application afterwards.

The 10 different sizes of SNP panel worked well in 537 varieties, which can be explained by the discriminatory power of panel or the diversity of materials is huge. Only one pair of variety found has high similarity. Hence, the control experiment of full sibs indicates the worst utilization of the panel we selected. In case of unidentifiable cases, 1144 and 566 available SNPs in full sibs with allele frequency from 40% to 60% and 45% to 55% respectively applied to estimate similarity between them accurately. Whilst, the panel of 50 SNPs and 100 SNPs selected from experiments were utilized to test discriminatory power in closely related materials. The average similarity among overall pairwise comparison with 50, 100, 566 and 1144 SNPs is lower than 0.90, revealing the SNPs we selected worked efficiently to distinguish between full sibs. Nevertheless, based on the standard deviation and the number of similarity coefficient in the upper tail of frequency distribution, the panel sizes we selected (≤100 SNPs) increase the percentage of high similarity (>0.85) between varieties compared to 566 and 1144 SNPs. Therefore, larger panel size contributes to less outliers of similarity value, representing more stable test can be acquired from it. However, it also presents the higher cost for the variety identification, which is not feasible and worth to do this kind of test generally.

4.4 Using SNPs to replace SSR

Given the advantages of SNP markers as mentioned in the introduction (low assay cost, high throughput and low mutation rate), and given the work done in the past with SSRs, we analysed the possibility to replace SSRs by SNPs, in order to save the information already obtained with SSRs. To replace all the 95 SSRs alleles currently used, an equally large number of SNPs is required in the studies of diversity and relatedness (Hamblin *et al*, 2007). In a study of maize, 7 to 11 times more SNPs than SSRs were suggested for analysing population structure and genetic diversity (Van Inghelandt *et al*, 2010). It can be argued that all comparisons between the number of SSR and SNPs loci are flawed. Counting the total of SSR alleles across loci is a more fair comparison with the total of SNP loci. When regarding the lack of control of the allele frequencies of all SSR alleles (typically an L-shaped distribution), then a selection of SNPs with balanced allele frequency is expected to outperform SSR-alleles in resolution power.

The replacement between SSRs and SNPs is feasible but not easy in our case for the reasons that (1) the limited number of common potato varieties in SSRs and SNPs leading to nearly 30% SSR alleles is absent, (2) the density of adjacent SNPs are insufficient to compare (e.g. 36 SNPs around STM3023) and (3) the deviation of dosage estimation of SSR alleles, due to phenotypic observations, which biased the allele frequency estimates. From the result we can see that the SNPs which are closer to SSR locus are more likely to get highly similar genotype with SSR alleles. Moreover, the allele frequency of markers impacts on the chance of discovering similar profiles between SSRs and SNPs. Common allele is unlikely to get high similarity value with other alleles. And rare allele can be highly similar to absent allele. From our results, it seemed that the higher allele frequency, the lower similarity value. However, the amount of rare allele (reference allele frequency smaller than 10%) is 7322, taking almost half of makers. Therefore, it has more candidate SNPs to replace for rare SSR alleles.

4.5 Utilization of SNP panel

Based on the results from the effect of allele frequency and SNP panel size, we propose 40 to 50 SNPs with allele frequency from 40% to 60% to create a robust SNP panel for potato variety identification. In order to test discriminatory power of SNP panel, 94 samples from 73 varieties were used. For the selection of SNP platform to analysis samples, multiplexed chip-based technology has highest throughput, but for the applications requiring small number of SNPs with large number of samples it is less suitable (Semagn *et al*, 2014). Besides, when compared results from KASP assay and genotyping by sequencing, both methods generated similar conclusions in the end (Ertiro *et al*, 2015). Therefore, KASP genotyping platform was chosen to test SNP panel, for the reasons that it is cost effective with low scale and precise genotyping analysis. Then we analysed results from KASP assay, which can be used as a tool for variety identification in the future.

The results from KASP assay are mostly in agreement with the data from infinium assay. Then we used the dosage data of 94 samples including the replicates to compare with 537 varieties. With 42 SNPs, the majority samples got high similarity (close to 1) in the pair with the expected varieties. In addition, the second highest similarity coefficient for those samples with non-identical varieties is lower than 0.85. More than 0.15 difference (=1-0.85) between top 1 and 2 similarity coefficient, which is equal to about 25 dosages differences (6 to 25 loci out of 42 SNPs loci). Therefore, the possibility for misidentification is low and our SNP panel is powerful to detect samples with same varieties. However, some samples performed inversely. Those samples reached low similarity with expected variety but high similarity (>0.85) with one unexpected variety. This phenomenon was also detected in two diploid of fresh sample. We assumed that those samples are labelled with incorrect denomination from previous processes. It also proves that some human error during the experiments can be discovered easily by our SNP panel.

5. Conclusion

Two kinds of similarity coefficient (Kosman's and Jaccard's coefficient) were used in all experiments. Kosman's coefficient detected dissimilarity precisely and performed better with common alleles than Jaccard's coefficient. Furthermore, Kosman coefficient makes use of the extra information of dosage data variation in tetraploids. For the discriminatory power of panel we selected, reference allele frequency affects it mostly. Hence, we concluded SNP markers with balanced allele frequency (highest PIC value) are easier to identify different variety than unbalanced allele frequency, according to the average similarity among 537 varieties.

For the size of panel, it affects the stabilization of testing similarity more than average similarity. The more markers applied for identification, the lower standard deviation of similarity value we got. In addition, a panel with 10 SNPs is already able to identify around 1000 varieties theoretically. However, in reality, we proposed that a panel with 40 to 50 SNPs should be sufficient. Therefore, 42 SNPs with 94 samples was used to see how it works in the real test. Then the similarity comparison showed 10 samples got a low similarity with expected variety, indicating wrong variety names were put on it by mistake. It proved the discriminatory power of 42 SNPs is strong enough to detect human error.

We used the Kosman's coefficient from 100 SNPs with reference allele frequency from 45% to 55% to make frequency distribution. It obeys normal distribution largely, except some extreme values of similarity. Besides, the results from 234 full sibs showed that Kosman's coefficient higher than 0.85 represents a close relationship between two varieties in pair. We suggested 0.85 Kosman's coefficient can be used as a threshold for the recognition of similar or different varieties.

Using SNPs to replace SSRs is feasible, but it is not easy in our case. More adjacent SNPs are required and precise dosage data of SSRs would be helpful to analysis.

6. Future research

In this study, we attempted to find highly similar SNP markers to replace nine SSR markers (STM5136, STM3023, STM5148, STM0019, STM3009, SSR1, STM3012, STM2005 and STM2028), which identified around 1000 varieties in study of Reid et al. (2011). However, we did not access the whole dataset from previous study. The SSR data which we used in this study was offered by several companies and collected from relevant studies. Only 91 varieties overlap with SNPs data. Therefore, 30% of SSR alleles are absent among those varieties. Maybe in the following study, using the original SSR data from Reid et al. (2011) can improve the replacement between SSRs and SNPs.

For the control experiments, we used 234 full sibs from the cross of *'Altus'* and *'Colomba'*. However, the similarity between *Altus* and *Colomba* is not very high. It is close to the average value of overall pairwise comparison among 537 varieties. The similarity level of parents may also influence the degree of similarity within their offspring. If it is possible, the full sibs of other two more similar varieties can be used to investigate the influence of parents and how robust of SNP panel.

In the end, for the validation of samples, the panel of 42 SNPs worked nicely. Nevertheless, it would be interesting to test it with more samples or even more similar samples like full sibs. Besides, the panel size can also be tested. Maybe 42 SNPs are not the optimal size. Unfortunately, we cannot compare it with other panel sizes due to the time limitation. Therefore, in the future experiments, how panel size affect the validation can be analysed as well.

7. References

Anon. (2004). *UPOV guideline TG/23/6, Vol. 2015*.

Bourke PM, Voorrips RE, Visser RG, Maliepaard C (2015). The Double-Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map. *Genetics* **201**(3)**:** 853-863.

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S *et al* (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC genetics* **3**(1)**:** 19.

Council (2002). Council Directive 2002/53/EC of 13 June 2002 on the common catalogue of varieties of agricultural plant species. *Off J Eur Communities* **L193:** 11.

D'hoop BB, Keizer PLC, Paulo MJ, Visser RGF, Eeuwijk FA, Eck HJ (2014). Identification of agronomically important QTL in tetraploid potato cultivars using a marker–trait association analysis. *Theoretical and Applied Genetics* **127:** 731-748.

D'hoop BB, Paulo MJ, Visser RGF, Eck HJ, Eeuwijk FA (2011). Phenotypic Analyses of Multi-Environment Data for Two Diverse Tetraploid Potato Collections: Comparing an Academic Panel with an Industrial Panel. *Potato Research* **54:** 157-181.

Demeke T, Kawchuk LM, Lynch DR (1993). Identification of potato cultivars and clonal variants by random amplified polymorphic DNA analysis. *American Potato Journal* **70**(8)**:** 561-570

Dou J, Zhao X, Fu X, Jiao W, Wang N, Zhang L *et al* (2012). Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology direct* **7**(1)**:** 17

Edward KJ, Poole RL, Barker GL (2008). 1 SNP Discovery in Plants. *Plant Genotyping II: SNP Technology***:** 1

Ertiro BT, Ogugo V, Worku M, Das B, Olsen M, Labuschagne M *et al* (2015). Comparison of Kompetitive Allele Specific PCR (KASP) and genotyping by sequencing (GBS) for quality control analysis in maize. *BMC genomics* **16**(1)**:** 908

Estoup A, Jarne P, Cornuet JM (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular ecology* **11**(9)**:** 1591-1604

FAOSTAT (2015). Food and Agriculture Organisation of the United Nations Statistics Division *http://faostat3faoorg/browse/Q/QC/E*.

Ferguson ME, Hearne SJ, Close TJ, Wanamaker S, Moskal WA, Town CD *et al* (2012). Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. *Theoretical and Applied Genetics* **124**(4)**:** 685-695

Gebhardt C, Blomendahl C, Schachtschabel U, Debener T, Salamini F, Ritter E (1989). Identification of 2n breeding lines and 4n varieties of potato (Solanum tuberosum, ssp. tuberosum) with RFLP-fingerprints. *Theoretical and applied genetics* **78**(1)**:** 16-22

Ghislain M, Spooner DM, Rodriguez F, Villamón F, Nunez J, Vásquez C *et al* (2004). Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. *Theoretical and Applied Genetics* **108**(5)**:** 881-890

Gonzaga ZJ, Aslam K, Septiningsih EM, Collard BCY (2015). Evaluation of SSR and SNP markers for molecular breeding in rice. *Plant Breeding and Biotechnology* **3**(2)**:** 139-152

Görg R, Schachtschabel U, Ritter E, Salamini F, Gebhardt C (1992). Discrimination among 136 tetraploid potato varieties by fingerprints using highly polymorphic DNA markers. *Crop Science* **32**(3)**:** 815-819

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O *et al* (2011). Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**(4)**:** 591-611

Hamblin MT, Warburton ML, Buckler ES (2007). Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PloS one* **2**(12)**:** e1367.

Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A *et al* (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC genomics* **12**(1)**:** 302

Hosaka K, Mori M, Ogawa K (1994). Genetic relationships of Japanese potato cultivars assessed by RAPD analysis. *American Journal of Potato Research* **71**(8)**:** 535-546.

Maughan PJ, Smith SM, Fairbanks DJ, Jellen EN (2011). Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (Amaranthus sp.). *The Plant Genome* **4**(1)**:** 92-101

Mir RR, Hiremath PJ, Riera-Lizarazu O, Varshney RK (2013). Evolving molecular marker technologies in plants: From RFLPs to GBS *Diagnostics in Plant Breeding*. Springer, pp 229-247.

Moisan-Thiery M, Marhadour S, Kerlan MC, Dessenne N, Perramant M, Gokelaere T *et al* (2005). Potato cultivar identification using simple sequence repeats markers (SSR). *Potato Research* **48**(3-4)**:** 191-200

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and applied genetics* **120**(8)**:** 1525-1534

Nei M (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**(12)**:** 3321-3323

Nielsen R (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**(2)**:** 931-942

Osman AM, Bonthuis H, Van den Brink L, Struik PC, Almekinders CJM, van Bueren ETL (2015). Adapting Value for Cultivation and Use testing to stimulate the release of improved varieties for the organic sector. The case of spring wheat in The Netherlands. *Organic Agriculture* **5**(2)**:** 101-111.

Pieterse L, Judd J (2014). World Catalogue of potato varieties 6th edition. *Agrimedia, Clenze*.

Prevost A, Wilkinson MJ (1999). A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. *Theoretical and Applied Genetics* **98**(1)**:** 107-112

Reid A. (2014). *Vol. 2015*. UPOV (ed.).

Reid A, Hof L, Felix G, Rücker B, Tams S, Milczynska E *et al* (2011). Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue. *Euphytica* **182**(2)**:** 239-249

Reid A, Kerr EM (2007). A rapid simple sequence repeat (SSR)-based identification method for potato cultivars. *Plant Genetic Resources: Characterization and Utilization* **5**(01)**:** 7-13

Semagn K, Babu R, Hearne S, Olsen M (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular breeding* **33**(1)**:** 1-14

Smith SM, Maughan PJ (2015). SNP genotyping using KASPar assays *Plant Genotyping*. Springer, pp 243-256

Spanoghe M, Marique T, Rivière J, Lanterbecq D, Gadenne M (2014). Investigation and Development of Potato Parentage Analysis Methods Using Multiplexed SSR Fingerprinting. *Potato Research* **58**(1)**:** 43-65

Thomson MJ, Zhao K, Wright M, McNally KL, Rey J, Tung C-W *et al* (2012). High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. *Molecular Breeding* **29**(4)**:** 875-886

Van Berloo R, Hutten RCB, Van Eck HJ, Visser RGF (2007). An online potato pedigree database resource. *Potato research* **50**(1)**:** 45-57

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics* **120**(7)**:** 1289-1299

Voorrips RE, Gort G, Vosman B (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics* **12**(1)**:** 172

Vos PG, Uitdewilligen JG, Voorrips RE, Visser RG, van Eck HJ (2015). Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* **128**(12)**:** 2387-2401.

Wu X, Ren C, Joshi T, Vuong T, Xu D, Nguyen HT (2010). SNP discovery by high-throughput sequencing in soybean. *BMC genomics* **11**(1)**:** 469