Identification of alleles involved in *Solanum tuberosum* plant maturity and quantification of alleles effects on phenotypic trait values

PBR-80436 Master Thesis Plant Breeding
Year 2013-2014

Supervisor: Dr. Herman van Eck

May 2014

María del Mar Pérez Nicolás

WAGENINGEN UR
For quality of life

Abstract

In potato (*Solanum tuberosum*) it is known that the main factor controlling plant maturity is the *StCDF1* gene, a member of the DOF (DNA-binding with one finger) family. Two truncated alleles of this gene, *StCDF1.2* and *StCDF1.3*, which contain a 7bp footprint insertion and a 865bp transposon insertion respectively, have been recognized as responsible for early phenotypes regarding plant maturity. A set of 470 potato cultivars, most of them with known phenotype, was screened for the presence or absence of these two alleles, and subsequently, a correlation between our results and phenotype was made. Although our approach was not able to detect *StCDF1.2* allele, a significant difference on phenotype was found between cultivars containing the *StCDF1.3* allele and the cultivars without it.

Moreover, in order to know if plant maturity is a trait determined by the combination of different haplotypes and not only by the presence of the previously mentioned truncated alleles, a subset of 190 cultivars was selected and their PCR products from the first half on the intron were sequenced. Since the intron is a highly polymorphic region, the sequence chromatograms were not clear enough to extract haplotype information from them; thus, PCR products of 15 cultivars were cloned. From the cloned product, 45 haplotypes could be identified, some of which contained haplotype specific DNA sequence variants which allowed us to recognise those haplotypes in 9 of the non-cloned cultivars. A nucleotide diversity value of π=0.115 was found, which correspond to 1 SNP every 8,7bp. The average number of alleles per individual (*Ai*) was 3.4 and three cultivars presented five different haplotypes, a fact that could be due to chimerism. Since only four of the 45 haplotypes were shared by more than one cultivar, association analysis was not successful; hence, further studies are recommended in order to elucidate the role of *StCDF1* alleles on plant maturity.

Table of Contents

List of abbreviations 4
Introduction 5
Materials & Methods 10
Plant material 10
PCR amplification for screening presence/absence of 865bp and 7bp insertions 10
PCR amplification and sequencing 11
Cloning intron of StCDF1 gene 12
Phenotypic data collection 12
Sequence variant detection and analysis 12
Results 13
Screening for presence or absence of the 7bp and 865bp insertions 13
Sequence diversity and haplotype analysis 14
Pedigree analysis 21
Genetic diversity 21
Association of haplotypes with plant maturity trait 21
Discussion 22
Relation of the 7 bp and 865 bp insertion on potato maturity 22
Nucleotide diversity and haplotype analysis 23
Pedigree analysis and genetic diversity 24
Association analysis of StCDF1 with potato maturity 24
Conclusions 25
Recommendations 26
References 27
Appendix 29

List of abbreviations

BAC: Bacterial artificial chromosome

CBSG: Centre for BioSystems Genomics

CDF: Cycling DOF factor

CO: CONSTANS

DOF: DNA-binding with one finger

FKF: Flavin-binding kelch repeat f-box protein 1

FT: Flowering locus T

GBBS: Granule-bound starch synthase

GI: GIGANTEA

GWAS: Genome Wide Association Study

INDEL: insertion-deletion

MNP: Multinucleotide polymorphism

PHYB: Phytochrome B

QTL: Quantitative trait locus

SNP: Single nucleotide polymorphism

Introduction

Potato (*Solanum tuberosum L.*), the third most important non- cereal food crop in the world in 2012 (http://faostat.fao.org), is a valuable source of nutrients supplying vitamins, carbohydrates and important amino acids to the human diet, as well as a treasured product for industries for manufacturing starch and alcoholic beverages (Abelenda, Navarro, & Prat, 2011; Sharma & Nandineni, 2014).

This crop belongs to the *Solanaceae* family, which possesses one of the richest genetic resources of any cultivated plant with approximately 190 wild and primitive species in the section *Petota* of the genus *Solanum* (Jacobs, Smulders, van den Berg, & Vosman, 2011). The current cultivated potato, with an estimated genome size of 850 Mb (de Boer et al., 2011), is an autotetraploid (2n=4x=48) and highly heterozygous species which was first domesticated in the Andes from wild species of *Solanum brevicaule* (Spooner, McLean, Ramsay, Waugh, & Bryan, 2005). In its natural environment, tubers act as vegetative propagation organs to the plant. During unfavourable seasons, these organs stay latent in the soil and sprout next growing season to develop a new plant. Tuber formation, is a complex process in which, under inductive conditions, a stolon stops growing and swells in the subapical region to form a storage organ, the tuber (Abelenda et al., 2011; Morris et al., 2014). Since the control of tuber development, as well as a high and uniform yield, are probably, the most important economic constraint of modern potato production, scientist have been investigating the molecular mechanisms responsible for tuberization during the past few decades (Morris et al., 2014).

It is already known that tuberization rate decreases significantly under high temperatures, mainly if the plant is exposed to warm temperatures during night (Abelenda et al., 2011). Photoperiod is however the main factor controlling tuber formation. In potato's original habitat, the shortening of day length guarantees tubers formation before winter; then, latency of the organs overlaps with the cold months. In wild accessions, photoperiod controls tuber development, making potato to tuberize under short-day conditions and avoiding tubers development under long-day conditions of spring and summer in the northern latitudes (Abelenda et al., 2011). When potato is introduced in these latitudes, wild varieties will develop tubers during autumn, when the day-length starts to become shorter (Kloosterman et al., 2013). Hence, long- day acclimation for tuber formation seems to be one of the first traits selected leading to the varieties nowadays cultivated in Europe (Ames & Spooner, 2008; Kloosterman et al., 2013)

Proper time for tuberization is regulated by environmental signals (Bäurle & Dean, 2006) which are perceived by many genes and signal factors making possible the tuberization process, but it has been shown that a member of the CYCLING DOF FACTOR (StCDF) gene family, mapped on chromosome 5, is the main factor controlling tubers formation (Kloosterman et al., 2013). Flowering in *Arabidopsis thaliana* starts when a 'florigen' called FLOWERING LOCUS T (FT) is expressed (Koornneef, Alonso-Blanco, Peeters, & Soppe, 1998). In potato, there is an homologue of FT (StSP6A), which is a mobile tuberization signal that starts the tuberization process and is independent from the flowering pathway (Kloosterman et al., 2013). StSPS6 expression is controlled by the red light receptor phytochrome B (StPHYB) in combination with the potato homologue for *A. thaliana*

diurnally-regulated transcription factor CONSTANS (StCO). CO acts downstream of a signalling cascade that involves interaction between the clock gene GIGANTEA (GI) and a several blue-light absorbing proteins including FLAVIN-BINDING KELCH REPEAT F-BOX PROTEIN 1 (FKF1), which regulate downstream effectors such as CDF (Imaizumi, Schultz, Harmon, Ho, & Kay, 2005). When FKF1 and GI bind to the CDF protein, they target it for degradation by the proteasome (Sawa, Nusinow, Kay, & Imaizumi, 2007). Apart from the DOF (DNA-binding with one finger), CDF proteins share three conserved basic domains in their C-terminus domain (Imaizumi et al., 2005). In a previous study (Kloosterman et al., 2013) it was shown that loss of the C-terminal domain in truncated alleles makes FKF1 unable to bind these proteins and thus, leading to CO repression and hence in early maturation. In the same study (Kloosterman et al., 2013) was also shown that there is another FT homologue (StSP5G) involved in tuberization. In *A. thaliana* CDF proteins were reported to directly bind to the CO promoter and interfere with StSP5G expression. In potato, two CONSTANS genes (StCO1 and StCO2) were shown to repress tuber development under long-day conditions. Even though the repressive function of these two genes was not well known previously, results of the mentioned study suggested that they are involved in StSP5G activation in long days, thus, StSP5G seems to act as suppressor of StSP6A. Based on these results and in a genome wide-association study using 83 tetraploid potato cultivars, scientists suggested a model for tuberization in which the proteins StGI and StFKF1 regulate the abundance of StCDF1, which downregulates StCO1 and StCO2. These CONSTANS genes supress the transcription of StSP5G, enabling StSP6A expression which results in tuber formation (Figure 1).

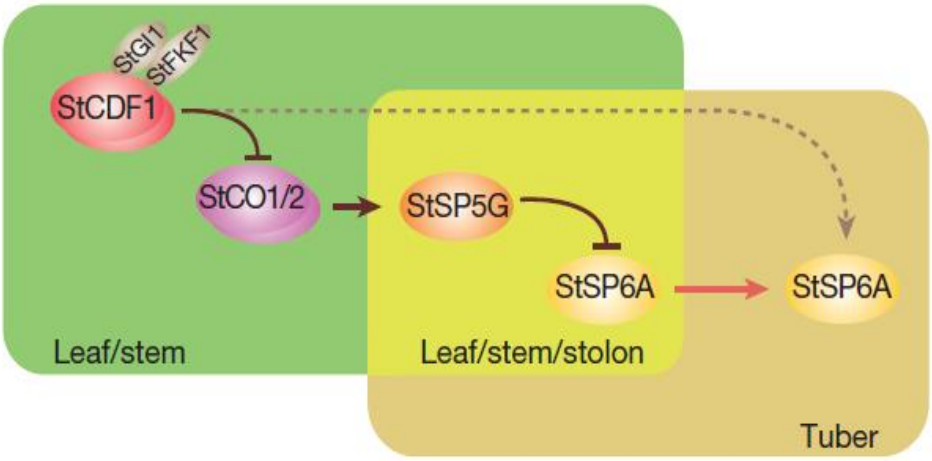


Figure 1. Regulation of tuber formation. Green represents aerial plant organs, the tuber is represented in brown and the overlap represents leaf, stem and stolon. Arrows (→) represent induction and lines (⊣) repression. Transport is represented by the red arrow. StCDF1 acts as an indirect inducer of StSP6A (dotted arrow)(Kloosterman et al., 2013).

To fine map StCDF, the main gene involved in plant maturity, Kloosterman et al. (2013) used two diploid potato populations generated from parental crosses C x E and SH x RH. From the C x E cross, they got a very late-maturing descendant homozygous for an StCDF allele denoted StCDF1.1, while the very early-maturing descendant possessed two different truncated alleles, StCDF1.2 and StCDF1.3. In StCDF1.2 a 7 bp insertion resulted in a frame-shift introducing a premature stop codon

and in StCDF1.3 there was a 864 bp insertion resulting in a fusion protein, replacing the 52 carboxy-terminal residues found in the StCDF1.1 (Figure 2).

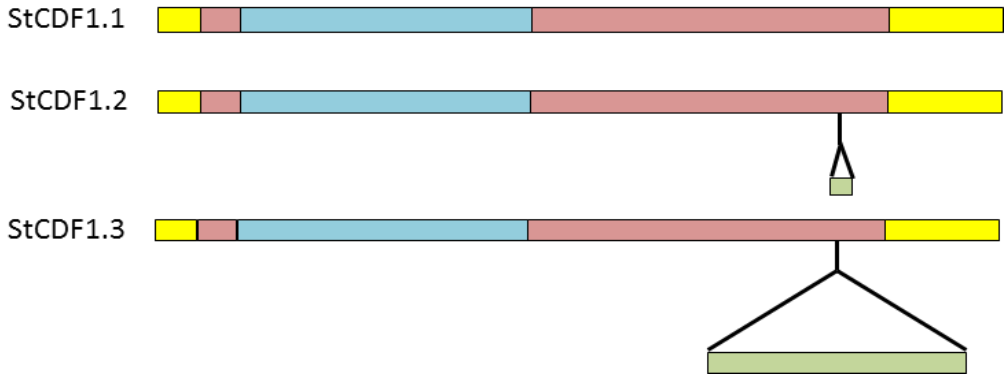


Figure 2. Representation of the three StCDF1 different alleles. StCDF1.1 refers to the wild type, StCDF1.2 refers to the allele with the 7bp insertion and StCDF1.3 refers to the allele with the 865bp transposon insertion.

These two diploid parents, C and E, and their offspring did not represent the whole allelic variation present in commercial cultivars. Thus, in order to know if other alleles or loci are involved in plant maturity trait, a Genome Wide Association Study (GWAS) was also carried out with a panel of 83 tetraploid potato cultivars. Finally, the results of the GWAS confirmed that StCDF1 is the only gene with allelic forms that influences maturation time (Kloosterman et al., 2013). These 83 tetraploid cultivars were used as well to obtain, by targeted re-sequencing of 800 of their genes, a genome-wide dataset of 135000 DNA variants including SNPs, MNPs and small indels. Among these DNA variants, the most significant marker-by-trait association is a SNP situated in the north arm of chromosome 5. On the contrary, none of the 80 sequence polymorphisms found in StCDF1 were related with plant maturity except for an A/G SNP at position 144258 in superscaffold PGSC0003DMB4010000192 (Kloosterman et al., 2013). Based on that and due to the fact that introns are generally less conserved than exons, they show different types of sequence polymorphism (Ahmadvand et al., 2014). Since in Solanaceous plants, the gene structure is conserved enough to use intron sequences as molecular markers (Poczai et al., 2010), during this research we tried to analyse intron sequences from several tetraploid cultivars in order to find DNA variants related with plant maturity trait. However, this analysis of DNA variants such as SNPs, MNPs and indels in a tetraploid specie like potato can be a challenging work or at least, more challenging than in diploids species, being that a specific gene may be represented not only by one to four different alleles but also by different levels of zygosity. In this crop, the level of zygosity includes nulliplex (0), simplex (1), duplex (2), triplex (3) and quadruplex (4). Hence, a genotyping system able to distinguish among alleles and to quantify the allele copy number is required for an correct genetic analysis (Uitdewilligen et al. 2013). Sanger amplicon sequencing could be an adequate method for discovering and genotyping sequence variants in polyploid species. On the other hand, potato shows high level of DNA sequence diversity, thus, due to the many indels present in its genome, sequences of PCR products result from the amplification of a mixture of four alleles, and therefore they are often unreadable with the Sanger sequencing method. For that reason, for this research, apart from direct sequencing of amplicons, also sequencing of cloned products was needed in order to obtain valid and clear

sequence information. Once most haplotypes are known, unique tag SNPs can be assigned to each haplotype. Therefore, when SNPs are used as "haplotype tags", that is numerous SNPs that tag all the known haplotypes in a given locus, they can be as informative as multi-allelic molecular markers. Besides, depending on the resolution, tag SNPs can be defined as either a single tag SNP or a combination of tag SNPs recognizing a single haplotype (J. Uitdewilligen, 2012)

Aside from trying to identify more DNA variants which could be involved in plant maturity, or variants that are in high LD with such functional variants, this research was also aimed to try to quantify alleles effects on the previously mentioned trait. Since the potato cultivars used for the GWAS could not be divided into two clearly different categories, early or late maturing, but showed a continuous distribution of phenotypes, a model of polygenic inheritance could be the most likely reason to explain this phenotypic variation. However, as discussed in previous paragraphs, for potato maturity it has been demonstrated that is only the StCDF1 gene the responsible for the variation. In this context and as cited in the study by Kloosterman et al. (2013), Sirks (Sirks, 1929) proposed a theory in which a single gene model with multiple alleles, each one having its own phenotypic effects, could be an explanation for quantitative inheritance . An example of this phenomenon in potato was observed at the *Ro* locus affecting tuber shape. Different tuber shapes can be observed, round, long or oval, making difficult a classification of phenotypic variation into Mendelian ratios; nevertheless this variability can be explained by the model of a single gene with multiple alleles. Moreover, multiple allelism in heterozygous autopolyploid species like potato not only happens for genes affecting morphological traits but also for genes involved in metabolic pathways. Several alleles encoding granule-bound starch synthase I (GBSSI) which is responsible for amylose biosynthesis were identified in potato (Van de Wal, Jacobsen, & Visser, 2001). Hence, a single gene with multiple alleles could be a plausible explanation for the distribution of phenotypes regarding potato maturity.

As mentioned before, three different alleles were found at the StCDF1 locus in experimental diploids, being StCDF1.1 recessive and StCDF1.2 and StCDF1.3 dominants. According with Robertson's hypothesis (Robertson, 1985) on the relationship of qualitative mutants to quantitative traits in which he proposed that major mutants are null or near-null alleles at a QTL, we can consider StCDF1.1 allele as the null or near-null allele which is recognized qualitatively, whereas variation among dominant alleles StCDF1.2 and StCDF1.3 is quantitative. Although in the case of StCDF1 is the wild type and not the mutated alleles the one which has to be considered as the null or near null-allele, the model suggested by Forkmann and Seyffert, 1977 (Forkmann & Seyffert, 1977), to describe the quantitative effects at a multiallelic locus is in agreement with the view that the recessive allele, in our case StCDF1.1, can be considered as a null or near-null allele. They suggested that the most recessive homozygote is the reference point and contributions of all other alleles are unidirectional and positive, while interactions between alleles are unidirectional and negative (van Eck, 1995). As plant maturity trait is influenced by a single locus, only intralocus interactions can be expected. Among these interactions, just additive ones can be easily extrapolated from diploid to tetraploid populations, whereas dominance and overdominance phenomena at a tetraploid level are still not well known. Throughout this research we will try to study the type of allele interactions.

To summarize and taking into account all the aforesaid, since the gene responsible for plant maturity in potato is already known, more targeting breeding can be done. Now it is possible to adapt the potato to make new varieties adjusted to specific geographical region and make this plant which is already the third non-cereal most important crop even more important. Potato could be now introduced into new areas and the production could also be enhanced in places where potato could not be produced before or where it could only be produced once in a year. Thus, the aim of this research is to try to identify different alleles affecting plant maturity in potato and to study quantitative effects of each allele of the StCDF locus on plant maturity in commercially grown tetraploid cultivars. To reach the objectives of this study, the following questions need to be solved.

- Are the DNA variants present at the StCDF1 C-terminus domain the only sequence variants responsible for the observed variability in plant maturity?
- Can we estimate the contribution of each allele to the phenotypic value?
- Is the contribution of each allele due to an additive effect or due to an interaction between them?

In order to answer these questions, the material and procedures explained in the next section were used.

Materials & Methods

Plant material

To aid screening for the presence/absence of the 865bp transposon insertion and the 7bp footprint insertion, DNA already extracted from six diploid cultivars with known genotypes were used as a reference: C, E, SH, RH, 3027 and 3130-2 (Supplementary table 1). The cultivar collection used contained four diploid and 466 tetraploid cultivars and they were all selected in order to have a range of commercial cultivars (Supplementary table 2). Among these 466 tetraploid cultivars, 184 were selected along with the four diploid cultivars in order to carry out the haplotype identification. For this analysis, the reference genotypes were the potato DM reference genome sequence, as well as BAC clone RH166D06 (StCDF1.1 allele) and BAC clone RH048D11 (StCDF1.3 allele). No DNA isolation was needed since all the DNA samples were already available.

PCR amplification for screening presence/absence of 865bp and 7bp insertions

PCR amplicons were obtained from 30 to 50 ng genomic DNA templates. PCR reactions were performed in a volume of 15 µl, using 0.03 µl of Taq polymerase, 10 x reaction buffer, 200 nM dNTP and 250 nM of each primer. For footprint amplification and detection primer pairs 1 and 2 were used (Figure 3) while for transposon amplification and detection primer pairs 2, 3, 4, 5 and 6 were used (Figure 4) (Supplementary table 3).

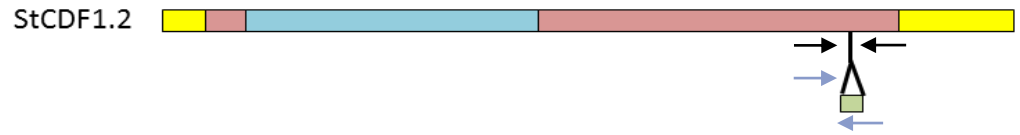


Figure 3. Primers position for the screening of the 7bp insertion in StCDF1.2. Blue arrows represent primer pair 1 and black arrows represent primer pair 2. (Supplementary table 3)

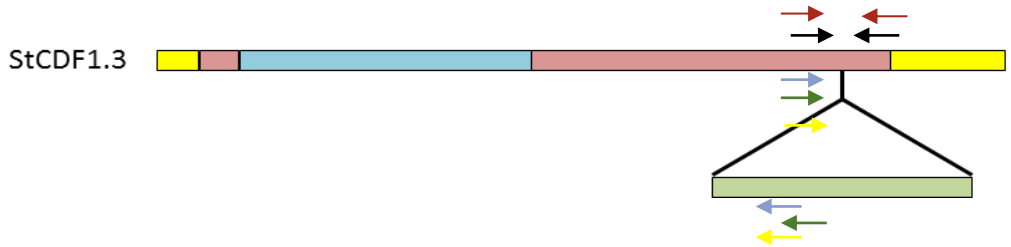


Figure 4. Primers position for the screening of the 865bp insertion in StCDF1.3. Black arrows represent primer pair 2, red arrows represent primer pair 3, green arrows represent primer pair 4, yellow arrows represent primer pair 5 and blue arrows represent primer pair 6. (Supplementary table 3)

The standard PCR cycle conditions were: 3 minutes at 94 °C for initial denaturation, followed by 35 cycles at 94 °C for 30 seconds for denaturation, 30 seconds annealing at 57 °C and 1 minute at 72°C for extension. The final extension step was carried out for 3 minutes at 72 °C and finally incubation was done at 10 °C. For the majority of PCRs SuperTaq Polymerase buffer and enzyme were used (Applied Biosystems). However, when amplification of whole transposon was tried, also Phusion DNA Polymerase and buffer were applied. In that case, concentration of the primers and polymerase were different from the used with Taq polymerase; now 0.5 nM of each primer (primer pairs 2 and 3. Supplementary table 3) and 0.2 µl Phusion DNA Polymerase were used. For Phusion Polymerase, the cycle conditions were: 30 seconds at 98 °C for initial denaturation, followed by 35 cycles at 98 °C during 5-10 seconds for denaturation, 30 seconds annealing at 57 °C and 30 seconds at 72°C for extension. The final extension step was carried out for 5 minutes at 72 °C and finally incubation was done at 10 °C. All PCR products were tested for quality on agarose gel by making use of GelRed.

The experiment was started with optimization of PCR conditions where all primer pairs and the repeatability of amplified fragments were tested by making use of a subset of diploid and tetraploid individuals.

PCR amplification and sequencing

For amplification and sequencing, primer pair 7 (Supplementary table 3) was designed based on the DM reference genomic sequence. For sequencing the first half of the StCDF1 gene, a fragment with a length of 628 bp (Figure 5), amplification was done using 15 µl reactions with 0.03 µl of Taq polymerase, 10 x reaction buffer, 200 nM dNTP and 250 nM of each primer. Standard cycling conditions were: initial denaturation for 3 minutes at 94 °C, followed by 35 cycles at 94 °C for 30 seconds for denaturation, 30 seconds annealing at 58 °C and 1 minute at 72°C for extension. The final extension step was carried out for 3 minutes at 72 °C and finally incubation was done at 10 °C. SuperTaq Polymerase buffer and enzyme were used (Applied Biosystems). PCR products were tested for quality on agarose gel using GelRed. Before amplicons were sequenced, PCR products were purified with the QIAquick PCR purification kit (QIAGEN). After purification, they were directly sequenced by GATC Biotech AG. The forward primer was used for sequencing in most cases, however, reverse primer was used as well for some tests. Amplicons of fifteen cultivars (Supplementary table 4) were cloned in order to obtain phased haplotypes. On average 16 cloned PCR products were sent for sequencing to obtain consensus sequences.

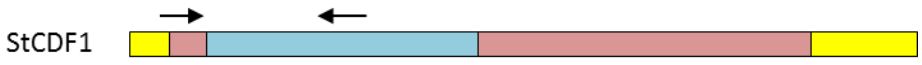


Figure 5. Primer pair position for the amplification of the first half of the intron.

The experiment was started with optimization of PCR conditions where the primer pair and the repeatability of amplified fragments were tested by making use of DNA of a subset of diploid and tetraploid individuals.

Cloning intron of StCDF1 gene

Primer pair 7 (Supplementary table 3) was used to obtain amplicons from the first half of the intron. Subsequently, the amplicons were cloned into pGEM-Teasy Vector (Promega). Plasmid DNA was isolated following plasmid miniprep protocol (Kotchoni, Gachomo, Betiku, & Shonukan, 2003). After plasmid miniprep, a PCR was carried out in order to confirm the presence of the fragment of interest in the plasmid. When the presence was confirmed, purified PCR products or plasmid DNA were sent for sequencing. In case of PCR products, primer StCDF-F 142886 was used for sequencing while in case of using plasmid DNA, the M13 forward primer was used as sequencing primer.

Phenotypic data collection

The phenotypic information that was used in this project for the association between different haplotypes and plant maturity trait, was already available since the studied tetraploid cultivars have been phenotyped before in the framework of the Centre for BioSystems Genomics (CBSG) project (CBSG).

Sequence variant detection and analysis

As stated in the introduction, Kloosterman et al. (2013), by crossing C and E diploid cultivars they got a very late-maturing descendant homozygous for an StCDF1 allele denoted StCDF1.1, and two very early-maturing descendant which possessed two different truncated alleles, StCDF1.2 and StCDF1.3. Apart from the allelic variants StCDF1.1, StCDF1.2 and StCDF1.3, aforementioned, there might be more alleles involved in plant maturity to be discovered. Thus, in this study by using the following techniques and software, we tried to identify them and to analyse the effects that different combination between those alleles could have regarding plant maturity.

As done in a similar research (J. Uitdewilligen, 2012), sequences alignment was performed using the Molecular Evolutionary Genetic Analysis (MEGA) software. Scoring of the allele copy number of SNPs was carry out by manual scoring and by using Data Acquisition and Data Analysis software DAx7.1. Similarity between each pair of sequences was measured on the basis of percentage identity and tree construction was done using the Neighbor-joining method. Phylogenetic tree was built by making use of the software MEGA and Past 3.0. The pedigree information was obtained from the potato pedigree database (van Berloo et al. 2007).

In addition to the sequences and sequence information obtained during this research, sequence information from 83 of the cultivars that were used for this investigation was already available from Jan Uitdewilligen. This information was available in SNPs form and in the sequence reads also phase information was available, but short stretches only. A subset of these SNPs was used to create a SNP genotyping array with 20,000 SNPs. Some of these SNPs were used in order to find a correlation between them and the presence or absence of the truncated alleles.

Results

Screening for presence or absence of the 7bp and 865bp insertions

In order to find DNA variants related with potato maturity, this research was first started with a screening for the presence or absence of the two early alleles, StCDF1.2 and StCDF1.3, in the 470 cultivars set, and for that purpose, different primer pairs were designed. First, with the aim of detecting the 7bp footprint, a reverse primer including the 7bp insertion was designed (Supplementary table 3) (Figure 3). With this primer, our objective was to obtain a band in the agarose gel only when the StCDF1.2 allele was present. However, when tested in cultivars C, E, SH, RH, 3027 and 3130-2, with known genotype (Supplementary table 1) a band demonstrating presence of the footprint was always present (Figure 6). In view of the failure of the first approach, a second primer pair was designed for the amplification of a 77bp fragment. Thus, being the fragment relatively short, we expected to be able to distinguish between the 77bp band when the footprint was absent and the 84bp band when it was present. Nevertheless the two bands mentioned before, were not possible to distinguish and hence, the presence or absence could not be determined in the 470 cultivars set.

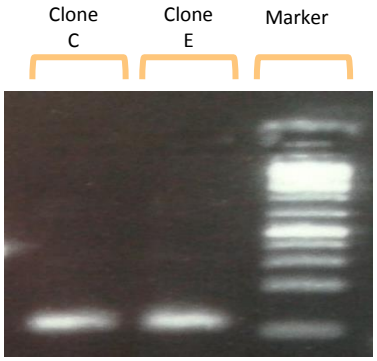


Figure 6. Image from an agarose gel. Primer pair 1 was used on diploid clone C and clone E. The image shows a band present in both cultivars although clone E does not contain StCDF1.2 allele.

Secondly for the transposon screening, the first strategy was to try the amplification of the whole fragment and for that purpose, two primer pairs were designed (Figure 4) (Supplementary table 4Supplementary table 3). However, we were not able to amplify the 865 bp long fragment; therefore, we designed a primer on the transposon in order to see a band on the agarose gel if the transposon was present. Following this strategy, we could distinguish between cultivars carrying the transposon and cultivars without it, but we could not identify the dosage in a cultivar. After all the cultivars were screened for the presence or absence of the StCDF1.3 allele, a t-test was performed in order to know if this allele significantly influences plant maturity trait and, as can be seen in Figure 7, results showed that cultivars with the transposon are significantly earlier that the ones without it.

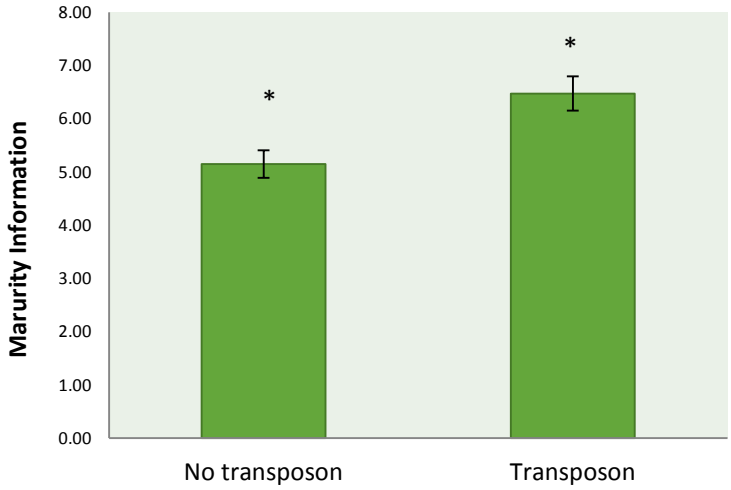


Figure 7. Maturity information on cultivars with and without the 865bp transposon insertion. A significant difference ($P < 0.05$) was detected between the two groups. The phenotypic values of this trait can be found in Supplementary table 2.

Sequence diversity and haplotype analysis

A set of 4 diploid and 184 tetraploid cultivars were selected to be directly sequenced in order to gain insight into StCDF1 gene DNA sequence variants. Amplicons obtained from diploid cultivars showed double chromatogram peaks at some nucleotide sites as expected for heterozygous samples. In the case of tetraploid accessions an average read length of 568bp was obtained and only in a few cases the chromatogram quality was low. From these 188 directly sequenced PCR product, 147 polymorphisms were scored from which seven were multi-allelic (six tri-allelic and one tetra-allelic) and 140 were bi-allelic variants. The average number of DNA sequence variants was 1 every 4 bp.

Since sequence information from homozygous or monoploids was not available, haplotypes from diploids and tetraploids could not be directly extracted from chromatograms due to high rate of polymorphisms and the highly heterozygous condition. Only in four cases, when the chromatograms showed no polymorphism we could recognise homozygous cultivars sharing same haplotype as the DM reference cultivar (Supplementary table 5). The absence of polymorphism in these four cases could also be due to too diverse alleles that were not amplified by the primes, thus resulting in null-alleles. In order to identify individual haplotypes, fifteen cultivars among the 184 tetraploids were selected and cloned from which 65 polymorphisms were scored, giving as a result an average number of 1 DNA sequence variant per 8,7 bp. Information extracted from the cloned PCR products allowed us to identify the surprisingly high number of 45 haplotypes (Table 1). Some of these haplotypes contained tag SNPs which were used to try to extract haplotypes from the initial set of 184 tetraploid cultivars (Table 2).

Table 1. Phased StCDF1 haplotypes. Haplotype defining tag variants are color-coded; Green bases indicate variants tagging a single haplotype, orange bases indicate polymorphism shared by multiple haplotypes. Deletions are shown as asterisk and variants used to identify haplotypes in non-cloned cultivars are shown in bold.

Hapl.	142351	142360	142397	142399	142405	142410	142411	142411	142413	142415	142419	142434	142435	142436	142439	142450	142468	142468	142493	142497	142497	142499	142506	142513	142514	142520	142520	142524	142525	142529	142554	142559	142563	142585	142585	142587	142595
e	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	A	T	C	A
g	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	T	T	C	A
1	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	T-	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
2	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
3	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	T	G	C	A
4	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
5	T	A	C	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	T	A	T	C	A
6	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
7	T	A	T	T	T	C	T	T-	T-	A	AA	A	A	G	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	T	T	C	A
8	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	G	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	T	T	C	A
9	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	T	T	T	C	A	T	C	A
10	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
11	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
12	T	A	T	T	T	C	T	TT	T-	A	AA	A	A	A	T	TT	T	T	A	A	AAA	A	C	G	T	C	C	A	A	C	T	T	C	A	T	C	A
13	T	A	T	T	G	C	T	TT	T-	A	A	A	A	A	T	TT	T	*	A	A	AAA	G	T	G	T	C	C	A	A	C	T	T	C	A	T	C	A
14	T	A	T	T	G	C	T	TT	T-	A	A	A	A	A	T	TT	T	T	A	A	AAA	G	T	G	T	C	C	A	A	C	T	T	C	A	T	C	A
15	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	G	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	A	T	C	A
16	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	A	T	C	A
17	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	A	T	C	A
18	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	T	TT	T	T	G	A	AAA	A	T	G	C	C	C	A	A	C	G	T	C	A	T	C	A
19	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	A	C	G	T	C	A	T	C	A
20	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	A	C	G	T	C	T	T	C	A
21	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	A	A	AAA	A	T	G	T	C	C	A	A	C	T	T	C	T	T	C	A
22	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	A	A	AAA	A	T	A	T	C	C	A	A	C	G	T	C	T	T	C	A
23	C	A	T	T	T	C	T	TT	T-	A	A	A	**G**	A	T	TT	T	T	A	A	A--	A	T	G	T	T	C	A	C	C	T	T	C	A	T	C	A
24	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	A	A	AAA	A	T	G	T	C	C	A	C	C	T	T	C	A	T	C	A

Hapl.	142351	142360	142397	142399	142405	142410	142411	142411	142413	142415	142419	142434	142435	142436	142439	142450	142468	142468	142493	142497	142497	142499	142506	142513	142514	142520	142520	142524	142525	142529	142554	142559	142563	142585	142585	142587	142595
25	T	A	T	T	T	C	T	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	A--	A	T	G	T	C	C	A	C	C	G	T	C	A	T	C	A
26	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	A	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	T	C	T	T	C	A
27	T	A	T	T	T	C	T	TT	TT	A	A	A	A	A	T	TT	A	T	G	A	AAA	A	T	G	T	C	A	A	C	C	G	T	C	T	T	C	A
28	T	A	T	T	T	C	T	TT	T-	A	A	A	A	A	T	TT	T	T	A	A	AAA	G	T	G	T	C	A	A	C	C	T	T	C	A	T	C	A
29	T	A	T	T	T	C	T	TT	T-	A	A	A	A	A	T	TT	A	T	G	A	AAA	A	T	G	T	C	A	A	C	C	G	T	C	T	T	C	A
30	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	A	T	C	A
31	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	A	T	C	A
32	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	A
33	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	G	C	C	G	T	C	T	T	C	A
34	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	G	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	G
35	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	C	C	T	T	C	A
36	T	A	T	GAATTATAAAGTTT	T	C	T	TT	T-	A	A	C	A	A	G	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	A	T	C	A
37	T	A	T	T	T	A	A	TT	T-	T	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	T	A
38	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	A
39	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	A
40	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	A
41	T	G	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	G	T	C	C	A	C	C	G	T	C	T	T	C	A
42	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	T	C	T	T	C	A
43	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	T	C	T	T	C	A
44	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	T	C	T	T	C	A
45	T	A	T	T	T	A	A	TT	T-	A	A	A	A	A	T	TT	T	T	G	A	AAA	A	T	A	T	C	C	A	C	C	G	T	C	T	T	C	A

Hapl.	142616	142627	142628	142623	142628	142628	142628	142628	142640	142647	142676	142699	142700	142704	142734	142735	142736	142740	142748	142749	142787	142791	142795	142801	142802	142803	142805	142805	142807	142810	142830	142831	142837	142839	142840	142840	142849	142869
e	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	A
g	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
1	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
2	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
3	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	A	A	T	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	C	A	T	T	T	T	*	T	G
4	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
5	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
6	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
7	A	AA	A	A	A	A	A	AAAA	A	T	T	A	G	T	G	T	T	A	G	A	A	A	A	*	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	T	G
8	A	AA	A	A	A	A	A	AAAA	A	T	T	A	G	T	G	T	T	A	G	A	A	A	A	*	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	T	G
9	A	AA	A	A	A	AA	A	A	A	T	T	A	A	T	G	C	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
10	A	AA	A	A	A	AA	A	A	A	T	T	A	A	T	G	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
11	A	AA	A	A	A	AA	A	A	A	T	T	G	G	T	G	T	T	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
12	A	AA	A	A	A	A	AAA	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
13	A	AA	C	T	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
14	A	AA	C	T	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
15	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	*	T	*	T	A
16	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	A
17	G	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	A
18	A	C-	A	A	A	A	A	A	A	C	T	A	G	C	A	T	C	A	G	G	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	A
19	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	A
20	A	C-	A	A	A	A	A	A	G	T	T	A	G	T	A	T	C	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	G
21	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
22	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	T--------	T----	A	A	T	T	T	T	T	T	G
23	A	AA	C	A	A	A	A	A	A	T	T	A	G	T	G	T	T	A	A	A	T	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	C	A	T	T	T	T	*	T	G
24	A	AA	C	A	A	A	A	A	A	T	T	A	G	T	G	T	T	A	A	A	T	T	A	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G

Hapl.	142616	142627	142628	142623	142628	142628	142628	142628	142640	142647	142676	142699	142700	142704	142734	142735	142736	142740	142748	142749	142787	142791	142795	142801	142802	142803	142805	142805	142807	142810	142830	142831	142837	142839	142840	142840	142849	142869
25	A	AA	C	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	*	T	G
26	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	T	*	A	T----	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
27	A	AA	C	A	A	A	A	A	A	T	C	A	G	T	A	T	T	A	G	A	A	A	T	*	A	TATAT	TAT	TATGGATAT	T----	A	A	T	C	T	T	T	T	G
28	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	A	T	*	A	T----	TAT	TATGGATAT	T----	A	A	G	T	T	T	*	T	A
29	A	AA	C	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	A	*	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	T	G
30	A	AA	A	A	A	A	AAA	A	A	T	T	G	G	T	G	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	C	G
31	A	AA	C	T	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
32	A	AA	C	T	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
33	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
34	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	G	T	T	A	G	A	A	A	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
35	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	G	T	T	A	G	A	A	T	A	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	T	G
36	A	C-	A	A	A	A	A	A	A	T	T	A	G	T	A	T	C	A	G	A	A	A	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	T	T	T	T	T	C	G
37	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	TATAT	A	A	T	T	T	A	T	T	G
38	A	AA	A	A	*	A	A	A	A	T	T	A	G	T	A	T	T	G	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	C	A	G	T	T	T	T	T	G
39	A	AA	A	A	A	AA	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	G	TATAT	T--	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
40	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
41	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
42	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	*	A	TATAT	TAT	TATGGATAT	T----	A	G	G	T	T	T	T	C	G
43	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	TATAT	TAT	TATGGATAT	T----	A	G	G	T	T	T	T	C	G
44	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G
45	A	AA	A	A	A	A	A	A	A	T	T	A	G	T	A	T	T	A	G	A	A	T	T	TATATAT	A	T----	TAT	TATGGATAT	T----	A	A	G	T	T	T	T	C	G

From the 35 tag sequence variants found in the cloned sequences, only three could be used to identify haplotypes in the set of the non-cloned cultivars (Table 2). By making use of these tag variant, was possible to extract haplotypes from 9 cultivars which represents only 5.26% of the total.

Table 2. Haplotype tag variants used to extract haplotypes from the non-cloned cultivars.

Reference Haplotype	Variant position	Cultivars with same haplotype
17	142616	IVP4X-119-37 IVP4X-217-21 Ar 96-0746 Stefano Hzpc06-36 Mpi 19268
20	142640	IVP4X-121-40
23	142435	Kantara Vr-94-268

In order to investigate the degree of sequence similarity between the discovered haplotypes, the e, f and g alleles and the DM reference genome, a Neighbor-joined dendogram was constructed, from which it is necessary to highlight that three cultivars presented five different haplotypes instead of a maximum of four (Figure 8). These results will be discussed in detail in the section ¨Discussion¨.

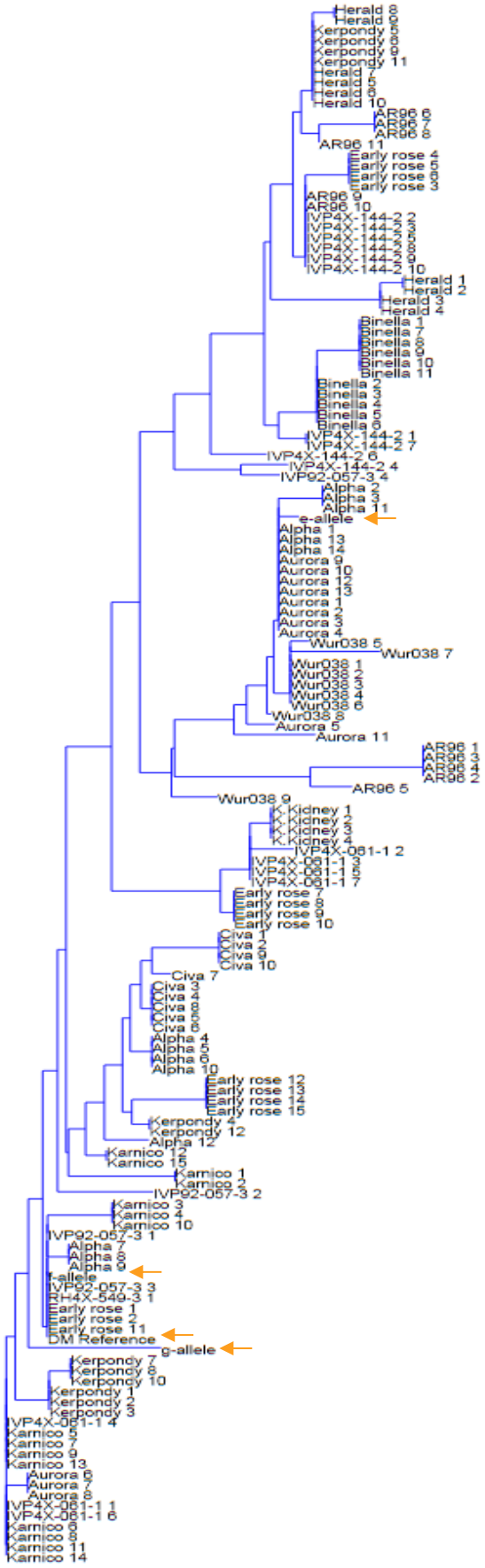


Figure 8. Dendrogram of the of the 45 StCDF1 haplotypes observed across 15 tetraploids, including the e, f and g alleles and DM reference genome. The distances were computed using the Jukes-Cantor method and the tree inferred using the Neighbor-joining method.

Pedigree analysis

From the 45 haplotypes identified during this research, 5 were shared by more than one cultivar. Thus, in order to verify that the haplotypes were identical-by-descendent and to identify possible sources of haplotypes, a pedigree analysis was carried out by making use of the potato pedigree database (van Berloo et al. 2007). Since most of the cultivars sharing haplotypes are not present in the database or have unknown parents, only haplotypes 16 and 40 could be analysed. Haplotype 16 is common for Alpha and Aurora. The pedigree database shows that Alpha is an ancestor of cultivar Aurora and the fragment they share remained without variation for four generations (Supplementary figure 1). In contrast, Herald, which is one of the parental of Saskia, does not have any common haplotype with it, which means that in only one generation it suffered allele variation (Supplementary figure 1). The same situation can be seen in the case of Civa and Binella, both descending from Herald and in case of Early Rose which is one of the parental of Alpha. The differences between related cultivars could be due to accumulation of mutation events but also due to wrong identity of the sample, for example, in case of Alpha there are four different entries in the pedigree database and only one of them descends directly from Early Rose (Supplementary figure 3).

Haplotype 40 descends from Early Rose to another 7 samples including Urgenta, Saskia and VTN 62-33-3 that are confirmed by pedigree analysis. This haplotype seems the most common version with a much higher allele frequency (although still small). Remarkably this most common haplotype is also the version observed in earlier studies (Potato Genome Sequence Consorium, 2011; Kloosterman et al., 2013).

Genetic diversity

Potato and all others tetraploid genotypes can be defined as monogenic (aaaa), digenic-simplex (aaab), digenic-duplex (aabb), trigenic (aabc) and tetragenic (abcd) (Gallais 2003). For this study the StCDF1 alleles from 15 cultivars were cloned in order to get the four possible haplotypes present in each cultivar. For the total of 45 haplotypes found among these 15 cultivars, the average number of alleles per individual (*Ai*) was 3.4. Only two of the 15 cultivars were observed as monogenic, however, since not enough colonies to ensure statistical reliability were sequenced, this cannot be confirmed. From the 13 remaining heterozygous cultivars, four were observed as tetragenic, four as trigenic, two as digenic-duplex and finally three cultivars, Alpha, AR96 and Wur038 were found to have five different haplotypes (Supplementary table 5), which for a tetraploid cultivar is an unexpected result which will be studied in detail in the discussion part.

Association of haplotypes with plant maturity trait

Maturity trait value of the cultivars used in our research was measured in the framework of the Centre for BioSystems Genomics (CBSG) project. Maturity is expressed in a scale ranging from 1 to 10, 1 standing for very late cultivar and 10 standing for very early cultivar. Three of the fifteen cultivars with known haplotypes lack of maturity information. Among the thirteen cultivars with known phenotype, Herald showed the highest value. Three Herald`s haplotypes were unique for this cultivar, thus no conclusions about whether only one of this haplotypes or the combination between them is the responsible for the early phenotype, however the fourth haplotype, which is shared with Kerpondy can be discarded as responsible for early maturity since Kerpondy is a late cultivar. Four other haplotypes are shared by more than one cultivar; nevertheless in all the case except for one,

the cultivars sharing haplotype have significantly different phenotypic values. On the other hand, haplotype 44, which is common for Karnico and IVP4x-061-1, could be a candidate responsible for their late phenotype. Regarding the 40 haplotypes left, no conclusion can be drawn since they are all unique.

Discussion

Relation of the 7 bp and 865 bp insertion on potato maturity

A set of 470 cultivars, including 6 diploid accessions were screened for the presence or absence of the StCDF1.2 and StCDF1.3 alleles, the two alleles responsible for early maturation in potato (Kloosterman et al., 2013). The approaches used to detect the StCDF1.2 allele failed. Using a reverse primer containing the footprint sequence, a band was always visible independently of the presence of the 7bp insertion. A possible explanation for this phenomenon could be based on the fact that the primer was designed on the footprint, which means that it contained twice the same sequence. Thus, it is possible that when the footprint was present the primer annealed correctly but when the footprint was not in a cultivar, the primer formed a loop skipping one of the repeated sequenced and hence, making amplification possible even when the target fragment was absent. This problem could be solved by making use of different techniques such as RT-PCR, Taqman assay and Beacon probes.

Regarding now the 865bp transposon insertion, we could detect when this allele was present. With that information a t-test was performed in order to know if there was a significant difference between cultivars with the transposon and cultivars without it. The results showed that cultivars with the StCDF1.3 allele were significantly earlier than cultivars in which the allele was absent. However, since with our approach was not possible to detect the dosage of the transposon in a cultivar and we failed as well in the detection of the StCDF1.2 allele, we could not draw other conclusions from the results, regarding additivity, dominance or allele interactions on the quantitative effects of StCDF1 alleles on plant maturity. Only with information about presence/absence it is not possible to know whether the early phenotype is due to the transposon, the combination of several StCDF1.3 alleles in the same cultivar, the presence of both, StCDF1.2 and StCDF1.3, just because of the presence of StCDF1.2 or due to other still unknown DNA sequence variants. This can be illustrated with the example of cultivars Astarte and KA02-2220 among others (Supplementary table 2). Cultivar Astarte is a late cultivar with a phenotypic value of 4 but in our screening it showed a band, which means that the transposon was present at least once. On the contrary, cultivar KA02-2220, which is an early cultivar with a phenotypic value of 8.5 did not show any band. In consequence, we cannot conclude that StCDF1.3 alone is the responsible for the early maturation and further research is needed to find a solution to this enigma.

Nucleotide diversity and haplotype analysis

The first half of the StCDF1 intron, a 628 bp long fragment, was analysed by direct sequencing of PCR products from diploid and tetraploid cultivars. By analysing the sequence chromatograms, haplotype information could not be extracted, hence, to clone PCR products was needed. Cloned PCR products allowed us to identify 45 haplotypes, of which not all have tag variations, among the 45 haplotypes only in 19 of them tag variations could be identified. The 35 identified tag sequence variations were used to try to identify haplotypes presents in the other 173 cultivars that were sequenced in the framework of this research and in that way to study their effect on the StCDF1 gene.

Potato is a crop with remarkably abundant DNA sequence variation. Moreover, this study is focussed on the intron and in general, intronic regions are less conserved and hence, more polymorphic then exonic ones (Ahmadvand et al., 2014). In the directly sequenced PCR products an overall frequency of polymorphic sites of 1 every 4 bp was found. However, since the sequence chromatograms were not clear enough to identify real polymorphism, this information is not reliable. On the contrary, chromatograms derived from cloned PCR were clear enough to trust them. In this case the frequency of sequence variants was 1 polymorphism every 8.7bp, which exceeds the level of 1 polymorphism per 12bp found in a previous study (J. Uitdewilligen, 2012). Nevertheless, the number of sequence variants can vary widely, depending on the number of clones, the regions and the genotypes analysed. In our case, working with the intronic region of the StCDF1, which is not under selection pressure, could be the reason of this high rate of DNA sequence variation.

It also has to be noticed that sequence variation could be interpreted in different ways. In table 1, it seems that certain elements can be interpreted as insertions/deletions, but they can also be interpreted as SSR. That is the case of polymorphism between position 142801 and 142807 which can be view as $(TA)^n$ dinucleotide repeat or SSR and also the case of polymorphisms between position 142628 and 142640 which can be regarded as poly A repeat or SSR. This different interpretation would result in a frequency of sequence variants of 1 polymorphism every 9,8bp.

Regarding haplotypes, an average number of alleles per cultivar (*Ai*) of 3.4 were found. It has to be emphasized that in the case of cultivars Alpha, AR96 and Wur038, five alleles were detected instead of a maximum of four, which is expected in a tetraploid cultivar. It is logical to think that this incongruence could be due to a mixture of DNA samples or due to a PCR or sequencing error; however another less frequent factor, called chimerism, could be the responsible. In a previous study (Dorst, 1952) about skin colour, it was found that cultivars Groene Rode Star and Bonte Rode Star, which are bud-sports of the variety Rode Star, were periclinal chimera. Groene Rode Star has a sub-epidermal layer genetically different from the others and Bonte Rode Star had a genetically different epidermis while the other layers were identical to the ones of Rode Star (Figure 9). Therefore, if cells of more than one genotype (genetic makeup) are found growing adjacent in the tissues of that plant, it could explain why five alleles were found in Alpha and Wur038.

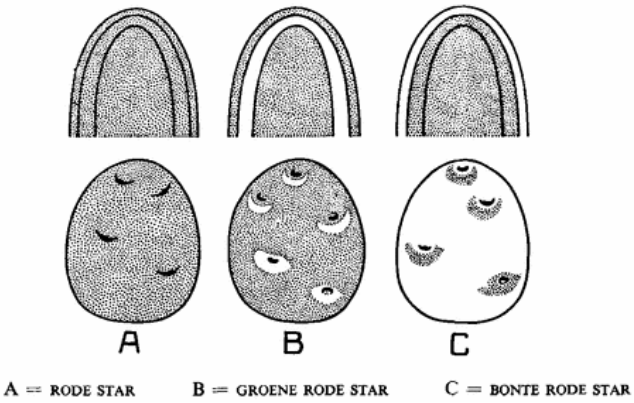


Figure 9. Illustration showing chimerism in potato. (Dorst, 1952)

Pedigree analysis and genetic diversity

With identifyed haplotypes, we expected to be able to study the haplotype transmission from parents to offspring. However from the 45 haplotypes found, only four were share by more than one cultivar and among this four, just one of them was shared by cultivars with complete pedigree information, thus, our study was limited. Haplotype 16 was common for cultivar Alpha and Aurora; looking at their pedigree (Supplementary figure 1) we can see that Alpha is one of Aurora's ancestors. These two cultivars are four generation distant; as a result, the fact that they share the same haplotype could suggest that the segment under study is quite stable and remained unchanging through several generations. However in the same figure it can also be seen that Alpha is the offspring coming from the cross Early Rose x Sebec and it does not share any haplotype with Early Rose. The same situation happens with Herald which is also a descendant of Early Rose, and Karnico, Civa and Binella which all descent from Herald. This is in concordance with the fact that the intron is not under selection pressure, therefore a higher rate of mutations can occur without effect on the phenotype, which explains why intronic region are more polymorphic than exonic ones. However allele divergence between related cultivars could also be due to accumulated mutation since the parental was released, for example Early Rose was released in 1867. Also regarding Early Rose, its differences with Alpha could be due to the sequencing of a Dutch Alpha cultivar and not the American Alpha cultivar which is the offspring resulting from the cross Early Rose x Sebec (Supplementary figure 3).

Association analysis of StCDF1 with potato maturity

Up to the moment, only StCDF1.2 and StCDF1.3 alleles are regarded as responsible for potato early maturation. In a previous study (Kloosterman et al., 2013), an A/G SNP at position 144258 in superscaffold PGSC0003DMB4010000192 was considered as a marker related with maturity trait. Hence, in order to find more markers related with this trait and also to find haplotypes and study their effects on maturation PCR products from fifteen cultivars were cloned and sequenced.

Among all the DNA sequence variants detected after sequencing, none of them showed a correlation with plant maturity trait, however, by sequencing, we could identify 45 different haplotypes, among which, just four were shared by multiple cultivars. Haplotype 6 was shared by

AR96 and IVP4X-061-1, but since the phenotype information of the second cultivar is not available, the influence of this haplotype on the phenotype cannot be studied. The same happens with haplotype 41, common for Early Rose IVP92-057-3 and RH4X-549-3. In case of haplotype 16, which is shared by Alpha and Aurora the phenotypes are different, Alpha is considered as a late cultivar and Aurora as an early one, therefore, this haplotype does not seem to have influence on the phenotype. Finally in case of haplotype 44 common for Karnico and IVP4X-061-1, it could have the effect of delaying maturation, since both cultivars containing it are regarded as late cultivars. Since the number of cultivars with full haplotype identification in this study was limited, further research with more cultivars is needed in order to understand the quantitative effects of different haplotypes combination on plant maturity trait. Only when analysing more cultivars that have certain haplotypes will allow to estimate the effect on maturity.

Conclusions

1. A significant difference on plant maturity was found between cultivars containing the 865 bp transposon insertion and cultivars without it. However no other conclusions regarding additivity, dominance or allele interactions on the quantitative effects of StCDF1 alleles on plant maturity can be drawn from it since the presence of the 7 bp footprint insertion, the other allele responsible for early maturation, could not be detected.

2. The average sequence variant was 1 polymorphism every 8,7bp, which is higher compared with other studies. This result suggests that intronic regions are less conserved than exonic ones.

3. No DNA sequence variants found in the intron showed significant association with plant maturity trait.

4. In cultivars Alpha, AR96 and Wur 038 five haplotypes were found, which suggests that these cultivars might be chimeras or unknown artifacts.

5. A higher number of fully classified cultivars is needed to study the possible quantitative effects of haplotype combinations on plant maturity trait. However, haplotype 44 could be a candidate responsible for late phenotype.

6. A follow up project is needed to further confirm or reject the aforementioned conclusions.

Recommendations

The original plan was to identify the presence or absence of the 865 bp and 7 bp insertions, both responsible for early maturation in potato (Kloosterman et al., 2013), in a set of 470 potato cultivars, and to quantify the correlation between the presence or absence of these two insertions with potato maturity. As part of the plan, we also wanted to try to identify, by sequencing the intron, a marker related with plant maturity trait different from these two insertions, as well as haplotypes present in potato in order to understand the effects that all these haplotypes could have on phenotype. Since the identification of the 7bp was not possible and our approach was not able to count how many alleles with the transposon were present in a cultivar, we cannot extract a solid conclusion from the results presented in this project.

Furthermore, by sequencing PCR products without cloning is not possible to identify haplotypes, consequently the effects of different haplotypes combination on the phenotype could not be studied. Thus, in order to have a better understanding of the correlation of the phenotype and the 7bp and 865bp insertions and to investigate if the phenotypic value is the result of additive effects of different haplotypes, a follow up project is needed

For a following up thesis project, a different technique to screen for the presence or absence of the 7bp footprint insertion is needed. It has been proven that by using the Taqman assay technique, in which a more specific probe than a primer is used, the identification is correct. Moreover, by using primers, the presence of the transposon could be sensed, however, was not possible to identify in how many alleles the transposon was present. Thus, to allow a proper correlation between plant maturity and presence or absence of the two insertions, not only a technique able to identify presence or absence is needed, but also a technique which can detect copy numbers.

With this research it has also been shown that in potato, a tetraploid cultivar, and more specifically while working with the intron, which is a region less conserved and thus with higher lever of polymorphism, the direct sequencing of PCR products is not the most suitable approach. Due to all the polymorphism present and to the tetraploid condition of potato, sequence chromatograms were not clear enough to identify haplotypes. Hence, to sequence cloned PCR product or to use a more advance technique, such as PacBio, is recommended in order to fine recognize different haplotypes present in a cultivar.

It is also recommended to assign year of market release to cultivar name in order to avoid mistakes due to synonymous cultivar names. That could have been the case of cultivar Alpha which has 4 different entries in the pedigree database (Supplementary figure 3). Adding information about year of release can avoid wrong conclusions regarding pedigree analysis.

Finally, in order to understand the effects of different haplotypes and their combination on plant maturity, a higher number of haplotypes is needed; therefore, more cultivars have to be selected.

References

Abelenda, J. A., Navarro, C., & Prat, S. (2011). From the model to the crop: Genes controlling tuber formation in potato. *Current Opinion in Biotechnology, 22*(2), 287-292.

Ahmadvand, R., Poczai, P., Hajianfar, R., Kolics, B., Gorji, A. M., Polgár, Z., & Taller, J. (2014). Next generation sequencing based development of intron-targeting markers in tetraploid potato and their transferability to other Solanum species. *Gene, 540*(1), 117-121.

Ames, M., & Spooner, D. M. (2008). DNA from herbarium specimens settles a controversy about origins of the European potato. *American Journal of Botany, 95*(2), 252-257.

Bäurle, I., & Dean, C. (2006). The Timing of Developmental Transitions in Plants. *Cell, 125*(4), 655-664.

CBSG. from www.cbsg.nl

de Boer, J. M., Borm, T. J., Jesse, T., Brugmans, B., Tang, X., Bryan, G. J., . . . Visser, R. G. (2011). A hybrid BAC physical map of potato: a framework for sequencing a heterozygous genome. *BMC genomics, 12*(1), 594.

Dorst, J. C. (1952). Two remarkable bud-sports in the potato variety Rode Star.*Euphytica*, 1(3), 184-186.

FAO. from http://faostat.fao.org

Forkmann, G., & Seyffert, W. (1977). Simulation of quantitative characters by genes with biochemically definable action. VI. Modifications of a simple model. *Genetics, 85*(3), 557-572.

Gallais, A. (2003). Génétique quantitative et méthodes de sélection chez les espèces autopolyploïdes. Editions Quae

Imaizumi, T., Schultz, T. F., Harmon, F. G., Ho, L. A., & Kay, S. A. (2005). Plant science: FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in Arabidopsis. *Science, 309*(5732), 293-297.

Jacobs, M. M., Smulders, M. J., van den Berg, R. G., & Vosman, B. (2011). What's in a name; Genetic structure in Solanum section Petota studied using population-genetic tools. *BMC evolutionary biology, 11*(1), 42.

Kloosterman, B., Abelenda, J. A., Gomez, M. d. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., . . . Prat, S. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*.

Koornneef, M., Alonso-Blanco, C., Peeters, A. J. M., & Soppe, W. (1998). Genetic control of flowering time in arabidopsis *Annual Review of Plant Biology* (Vol. 49, pp. 345-370).

Kotchoni, S. O., Gachomo, E. W., Betiku, E., & Shonukan, O. O. (2003). A home made kit for plasmid DNA mini-preparation. *African Journal of Biotechnology, 2*(4).

Morris, W. L., Hancock, R. D., DUCREUX, L. J. M., Morris, J. A., Usman, M., Verrall, S. R., . . . HEDLEY, P. E. (2014). Day length dependent restructuring of the leaf transcriptome and metabolome in potato genotypes with contrasting tuberization phenotypes. *Plant, cell & environment*.

Poczai, P., Cernák, I., Gorji, A. M., Nagy, S., Taller, J., & Polgár, Z. (2010). Development of intron targeting (IT) markers for potato and cross-species amplification in Solanum nigrum (Solanaceae). *American Journal of Botany, 97*(12), e142-e145.

Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355), 189-195.

Robertson, D. S. (1985). A possible technique for isolating genic DNA for quantitative traits in plants. *Journal of Theoretical Biology, 117*(1), 1-10.

Sawa, M., Nusinow, D. A., Kay, S. A., & Imaizumi, T. (2007). FKF1 and GIGANTEA complex formation is required for day-length measurement in Arabidopsis. *Science, 318*(5848), 261-265.

Sharma, V., & Nandineni, M. R. (2014). Assessment of genetic diversity among Indian potato (Solanum tuberosum L.) collection using microsatellite and retrotransposon based marker systems. *Molecular Phylogenetics and Evolution, 73*(1), 10-17.

Sirks, M. (1929). The interrelations of some anthocyane-factors in the potato. *Genetica, 11*(3), 293-328.

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., & Bryan, G. J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America, 102*(41), 14694-14699.

Uitdewilligen, J. (2012). *Discovery and genotyping of existing and induced DNA sequence variation in potato*: Wageningen University.

Uitdewilligen, J. G. A. M. L., Wolters, A. M. A., D'Hoop, B. B., Borm, T. J. A., Visser, R. G. F., & van Eck, H. J. A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS ONE, 8*(5).

van Berloo, R., R. C. B. Hutten, H. J. van Eck and R. G. F. Visser (2007) An online potato pedigree database resource. *Potato Research 50*: 45-57.

Van de Wal, M., Jacobsen, E., & Visser, R. (2001). Multiple allelism as a control mechanism in metabolic pathways: GBSSI allelic composition affects the activity of granule-bound starch synthase I and starch composition in potato. *Molecular Genetics and Genomics, 265*(6), 1011-1021.

van Eck, H. J. (1995). *Localisation of morphological traits on the genetic map of potato using RFLP and isozyme markers*: Landbouwuniversiteit te Wageningen.

Appendix

Supplementary table 1 Genotype of reference cultivars. e refers to StCDF1.1, f refers to StCDF1.2 and g refers to StCDF1.3.

Cultivar	Genotype
C	ef
E	eg
RH	eg
SH	ff
3027	ee
3130	fg

Supplementary table 2. Results derived from the screening for the presence/absence of the transposon. 1 refers to presence and 0 refers to absence. Phenotypic information is also shown. Regarding phenotype information, 1 refers to very late cultivars and 10 refers to very early cultivars.

Cultivar	Transposon	Phenotype	Cultivar	Transposon	Phenotype
707TG-11-1	1	5.7	IVP4X-218-13	0	5.7
Abundance (Sutton'S)	0		IVP4X-218-14	1	
Accord	0	7.4	IVP4X-218-15	1	
Ackersegen	0	3.8	IVP4X-223-13	0	6.5
Adirondack	1	5.2	IVP4X-225-15	0	5.5
Adora	1		IVP92-057-3	0	
Adretta	1		IVP97-079-9	0	
Agata	1	8.2	IVP98-082-14	0	
Agria	1	5.4	Jaerla	1	7.4
Ajiba	0	6.0	Jaune D'Or	1	5.1
Aladin	1	5.5	KA 07-4188	1	5.9
Albion	0	6.0	KA 07-4190	1	6.9
Allure	0	4.5	KA 07-4192	1	5.9
Almera	1	6.7	Ka 77-0133	0	
Alpha	0	4.5	Ka 96-0202	0	2.7
Altus	0	3.9	Ka 97-0743	0	3.8
Am 66-42	1		KA02-2220	0	8.5
Am 78-3704	1	7.6	Kantara	0	4.1
Ambition	0	5.5	Karakter	0	3.7
Amorosa	1	6.6	Kardal	0	
Ampera	1	6.6	Kardent	0	4.6
Amyla	1	4.1	Karnico	1	3.5
Annabelle	1	7.6	Kartel	0	3.7
Anosta	1	7.7	KASTELLI	1	6.7
ANUSCHKA	1	7.9	Katahdin	1	5.6
Ar 76-168-1	1	8.3	Katinka	0	3.2
Ar 80-127-5	1	5.3	Kennebec	1	6.4

Ar 85-96-13	0	4.3	Kepplestone Kidney	0	6.2
Ar 96-0746	1	6.3	Kerpondy	0	4.4
Ar 96-1120	1	5.3	Kerr'S Pink	0	
Ar 97-0576	0		King George	1	6.0
Ar 97-3342	1	6.3	Kondor	1	5.8
Ar 98-0313 (arizona)	1	6.7	Kuras	0	3.7
Ar 98-0342	1		Kuroda	1	5.7
Ar 98-1196	1		Labadia	1	6.7
Ar 98-1428	1	6.3	Lady Amarilla	1	6.9
Ar 99-1184	1	6.4	Lady Blanca	1	7.5
Ar 99-1333	1	5.4	Lady Christl	1	7.9
Ar 99-1343	1	5.3	Lady Felicia	1	7.5
AR95	0		Lady Jo	1	7.0
AR96	1		Lady Olympia	1	6.2
Arcade	1	6.6	Lady Rosetta	1	6.3
Arinda	0	6.7	Lady Valora	1	5.6
Arlanda	1	6.7	LANORMA	0	5.6
Armundo	1	4.3	Laura	1	5.9
Arnova	0	6.9	Laurene	1	6.0
Arran Chief	0	4.8	Lenape	1	6.8
Arran Pilot	1	6.8	Leyla	1	7.9
Arran Victory	0	4.1	Liseta	1	7.7
Arrow	1	8.2	Ma 96-0419	1	
Arsenal	1	5.8	Magnum Bonum	1	8.5
Artemis	1	7.4	Majestic	1	5.1
Astarte	1	4.0	Mantra	0	3.2
Asterix	0	5.4	Mara	0	4.6
ATLANTIC	0	6.5	Marabel	1	7.9
Aurora	1	7.9	Marfona	1	6.8
Ausonia	1	7.0	Marilyn	1	7.1
Avano	1	4.5	Maris Piper	1	5.7
Avarna	0	3.5	Maritiema	0	
Avaya	1	5.9	Maritta	0	
Aveka	0	5.2	Markies	0	4.0
Avenance	0	4.3	May Queen	1	7.7
Aventra	0	5.0	Melody	0	5.5
Averia	0	4.3	Menco	0	4.7
Aviala	0	4.2	Merano	0	3.1
Axion	0	4.3	Mercator	0	3.6
Ballydoon	1	6.3	Monaco	0	
Bartina	1	5.3	Monalisa	1	7.5
BELLANOVA	1	4.9	Mondial	1	5.0
Belle De Fontenay	1	8.1	MONTE CARLO	1	7.1
Bellini	1	6.1	Morene	1	4.6

Berber	1	7.8	Mpi 19268	1	
Bildtstar	1	6.1	Musica	1	6.6
Binella	1	7.4	Mustang	1	6.2
Bintje	1	6.4	Myatt'S Ashleaf	1	6.6
Bionica	1	7.0	Nadine	1	6.8
British Queen	1	6.3	Nicola	1	6.0
Brodick	1	6.7	Niska	1	6.0
Bronko	1	6.1	Noisette	1	7.1
Caesar	0	5.5	Nomade	0	5.9
Canberra	1	7.0	Noordeling	0	
Cara	1	4.7	Obelix	1	6.4
Casteline	1	7.3	Opera	1	6.6
Challenger	1	5.8	Orchestra	1	7.2
Champion	0	3.4	Oriana	1	8.4
Charlotte	1	7.7	OSIRA	1	6.9
Cherie	1	7.4	Pallas_Duplo	1	6.3
Chopin	1		Pallina	0	
Cilena	1	7.8	Panda	0	4.8
Citadel	1	6.4	PANTHER	1	7.7
Civa	1	8.3	Paterson'S Victoria	1	4.9
Clivia	1	5.1	Peerless	1	2.7
Cmk1998-035-053	1	7.3	Pekaro	1	6.3
Cmk1998-601-035	0	7.4	Pentland Dell	0	5.6
Cmk1999-018-011	1		Pepo (1919)	1	4.4
Cmk1999-038-039	1	6.9	Picasso	0	5.7
Cmk1999-079-034	1	7.0	Premiere	1	8.0
CMK200	1		Prevalent	0	3.7
Cmk2000-032-009	1	6.4	Primura	1	8.2
Cmk2001-003-001	1	7.1	Princess	1	7.3
Cmk2001-023-039	1	7.0	Profijt	1	5.1
Cmk2001-034-056	1	6.8	Quarta	1	6.8
Cmk2001-049-007	0	4.9	Raja	1	6.2
Cmk2001-050-016	1	7.1	Ramos	1	6.0
Cmk2001-053-011	1	6.2	Ranger Russet	1	6.2
Cmk2001-067-018	1	7.1	Ratte	0	6.1
Cmk2001-069-050	0	4.9	Record	1	5.8
Cmk2002-004-004	1	6.7	Red Baron	1	5.9
Cmk2002-010-008	1	6.5	Red One	1	6.2
Cmk2002-023-018	1	6.3	Red Scarlett	1	7.2
Cmk2002-204-002	0	6.2	Redstar	1	6.0
Cmk2002-210-007	1		Remarka	1	5.2
Cmk2002-210-026	1	6.2	RH06-973-8	0	
Cmk2003-607-004	0	5.3	RH07-027-7	0	
CMK201	1		RH07-066-14	0	
CMK202	0		RH4X-131-15	1	7.2

CMK203	1		RH4X-238-9	0	
CMK204	1		RH4X-353-16	1	6.0
CMK205	0		RH4X-361-2	1	8.1
Courage	1	6.7	RH4X-362-4	1	7.5
Craigs Alliance	1	6.6	RH4X-362-5	1	6.9
Craigs Bounty	1	5.8	RH4X-367-29	1	6.6
Craigs Defiance	1	6.0	RH4X-532-40	1	7.8
Cupido	1	7.4	RH4X-533-12	0	5.8
Daisy	0	5.9	RH4X-540-23	1	7.8
Defender	0	5.4	RH4X-544-15	0	5.2
Deodara	1		RH4X-549-3	1	7.2
Desiree	1	5.8	RH4X-598-7	1	5.7
Di Vernon	1		RH4X-614-10	0	5.7
Diamant	1	6.0	RH4X-638-2	1	6.2
Ditta	1	6.2	RH4X-638-20	1	7.7
Dob2000-534-036	1	7.2	RH4X-638-21	1	7.2
Dolce Vita	0	5.2	RH4X-648-17	0	4.7
Donald	0	7.3	Rijk98-049-008	1	6.7
Doon Star	1	6.6	Robeny	1	6.0
Dorado	1	7.0	Rode Eersteling	1	
Dore	1		Romano	1	7.1
Dr Mcintosh	0	5.3	Rosella	1	5.5
Draga (1970)	1		Russet Burbank	1	5.7
Dynamica	1	5.9	Sagitta	0	6.4
Early Rose	1	6.3	Samba	1	6.7
Eden (2000)	0	5.9	Santana	0	6.2
Eersteling	1		Sante	1	6.3
Ehud	1	7.4	Saphire	0	4.9
Eigenheimer	1	6.2	Saskia	1	7.8
El Paso	0	5.9	Saturna	1	6.4
Elisabeth	1	7.1	Seresta	1	5.2
Elkana	1	5.2	Shamrock	0	3.5
ENDEAVOUR	1	5.6	Shepody	1	6.6
Eos	1	4.4	Sifra	1	5.6
Epicure	1	8.1	Signum	1	
Escort	1	6.1	Silvester	1	4.8
Estima	1	7.2	Simply Red	0	4.8
Everest	1	7.3	Sinora	1	6.9
Exquisa	0	5.8	Sirtema	1	
Fabula	1	5.8	Smart	1	7.3
Felsina	1	6.6	Soprano	1	7.2
Festien	0	3.4	Spunta	1	6.3
Fianna	0	4.5	Starga	0	4.6
Fichtelgold	0	4.3	Stefano	1	3.4
Florida	1	5.6	Stirling	1	4.7
Flourball (Sutton's)	1	5.0	Sunrise	1	
Fontane	1	5.8	Susanne	1	7.0

Fresco	1	7.8	Sw 98-1680	1	
Frieslander	1	8.4	Sylvana	1	6.5
Furore	0		Tahi	0	5.0
Gladstone	1	5.0	Tasso	0	4.0
Gloria	1	7.6	Taurus	1	6.5
Golden Wonder	0	3.4	Terra Gold	0	5.0
Goya (2000)	1	4.2	The Alness	1	7.2
Granola	1	6.3	Timate	1	6.3
Great Scot	1	5.5	Tinwald'S Perfection	1	4.8
Hansa	1	6.0	Toyoshiro	1	6.7
Herald	1	8.5	Triplo	1	6.9
Hermes	1	6.3	Ulme	1	4.5
Home Guard	1	7.5	Ulster Glade	1	4.9
Hzpc06- 10	0		Ulster Knight	1	4.9
Hzpc06- 11	1	6.5	Ulster Sceptre	1	8.3
Hzpc06- 13	1	6.0	Ultimus	0	5.3
Hzpc06- 16	1		Umatilla Russet	1	6.2
Hzpc06- 18	1		Up To Date	1	5.1
Hzpc06- 19	1	6.8	Urgenta	1	5.5
Hzpc06- 21	0	6.0	Valiant	0	3.6
Hzpc06- 23	1		Valor	1	
Hzpc06- 29	1	8.1	Ve 66-295	1	5.8
Hzpc06- 3	1	6.6	Ve 74-45	0	5.2
Hzpc06- 33	1		Vectra	1	4.3
Hzpc06- 34	1	7.1	Verba	1	3.8
Hzpc06- 36	0		Veronie	0	7.4
Hzpc06- 5	0		Victoria	1	6.3
Hzpc06- 8	0	6.5	Virgo	1	6.7
Hzpc06- 9	1	6.2	Vision	1	5.2
Impala	0		Vivaldi	1	7.5
Industrie	0	4.0	VK 69-491	1	
Innovator	1	7.0	VK 69-491_AR	1	
Inova	1	8.0	Voran	0	4.4
Irish Queen	0	4.3	Vos2001-017-001	1	6.4
IVP-3916-22	1		Voyager	1	5.9
IVP-3916-71	1		Vr-00-1071	0	
IVP-3916-87	1		Vr-00-1077	0	5.9
IVP4X-041-1	1	4.5	Vr-00-1078	0	6.5
IVP4X-046-33	1	7.5	Vr-90-44	1	7.9
IVP4X-061-1	0	4.4	Vr-92-216	1	
IVP4X-066-16	0	7.6	Vr-92-501	0	5.0
IVP4X-072-17	0	5.8	Vr-93-146	1	5.6
IVP4X-075-50	1	7.4	Vr-94-268	1	
IVP4X-076-24	1	6.8	Vr-95-313	0	5.6
IVP4X-081-10	1	5.8	Vr-95-562	1	8.1

IVP4X-088-13	0	6.1	Vr-96-325	1	6.6
IVP4X-104-26	0	7.6	Vr-97-1127	0	6.1
IVP4X-116-14	0	5.9	Vr-97-670	1	7.0
IVP4X-119-37	1	5.2	Vr-98-1077	1	
IVP4X-121-40	1	7.6	Vr-98-1111	1	
IVP4X-124-3	1	5.9	Vr-98-265	1	
IVP4X-132-6	0	5.4	Vr-98-321	1	6.8
IVP4X-134-3	0	4.9	Vr-98-377	1	6.6
IVP4X-142-8	1	5.9	Vr-98-413	1	5.9
IVP4X-144-2	1	7.8	Vr-98-72	1	7.9
IVP4X-150-10	1	6.2	Vr-98-808	1	
IVP4X-156-12	1	7.8	Vr-98-967	1	8.4
IVP4X-163-10	1	8.3	Vr-99-406	1	7.4
IVP4X-184-1	1	7.2	Vr-99-579	1	6.9
IVP4X-184-42	1	6.8	Vr-99-680	1	6.6
IVP4X-185-31	1	6.8	Vr-99-715	1	8.5
IVP4X-185-8	1	5.8	Vr-99-906	1	6.3
IVP4X-203-27	1	6.7	Vtn 62-33-3	1	5.5
IVP4X-203-28	1	6.2	W 72-22-496	1	7.3
IVP4X-214-10	0	5.5	Wauseon	1	
IVP4X-215-13	1	5.2	Wilja	1	7.2
IVP4X-217-21	1	6.2	Winston	1	6.9
IVP4X-217-22	1	6.7	Wisent	0	4.0
IVP4X-217-23	1	6.7	Wur038	1	
IVP4X-217-26	1	7.7	Yam	1	4.0
IVP4X-217-4	1	6.2	Yukon Gold	1	7.7
IVP4X-218-10	0	6.2	Zafira	1	6.3

Supplementary table 3. Primer sets used. Hash symbol (#) represents amplification length when the 7bp and the 865bp insertions are not present Asterisk (*) represents amplification when the insertions are present, if they are not, amplification will not happen.

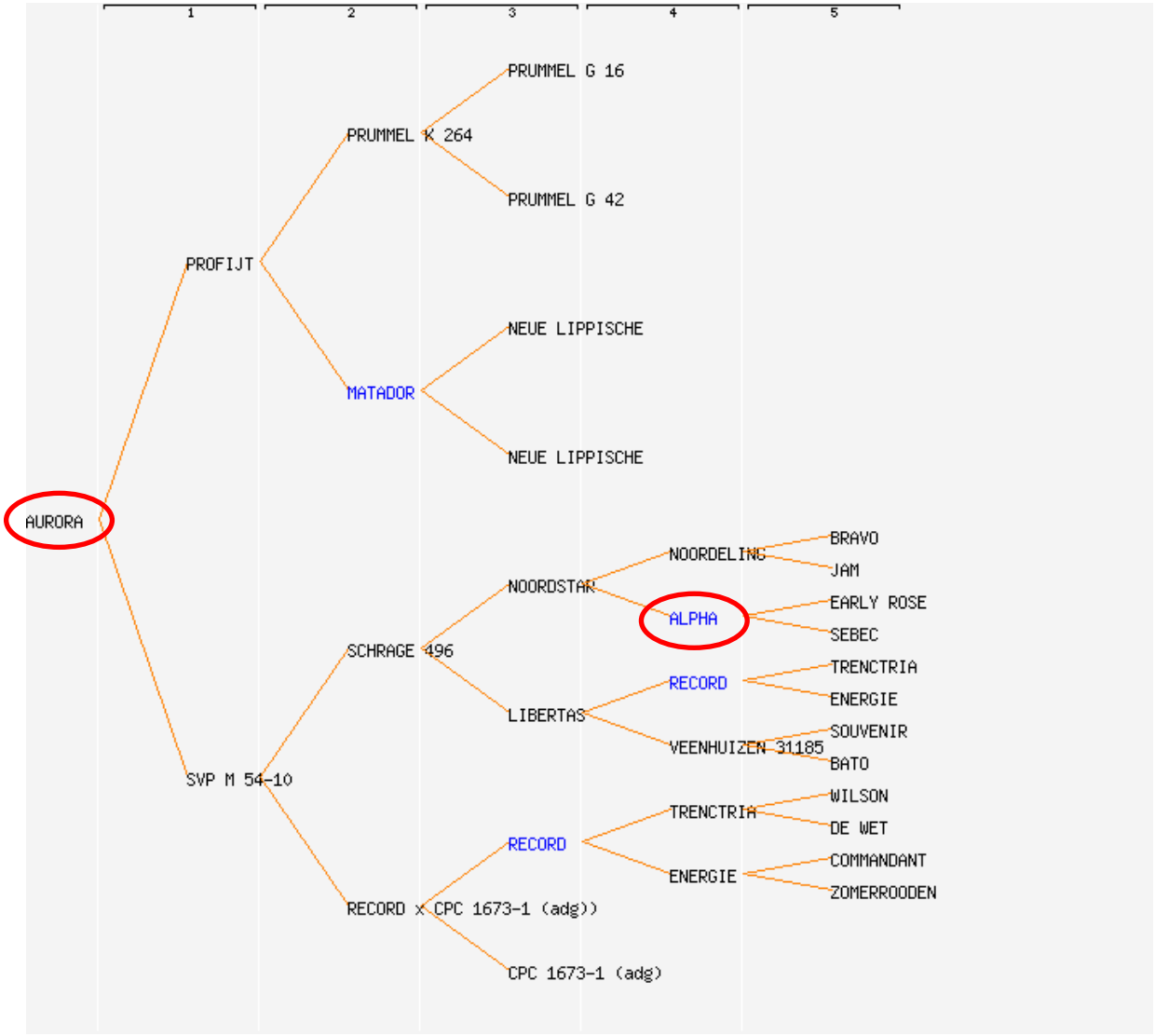
	FORWARD		REVERSE		
Primer Pair	Name	Sequence	Name	Sequence	Amplification length (bp)
1	CDF_F1	AATGAAGCGGCTAAAAGCTC	StCDF-R footprint	TCTTCTCATTCCTGATACCTAGTGA	62*
2	CDF_F1	AATGAAGCGGCTAAAAGCTC	StCDF-R 144340	ATTGTACCACGACCGAATC	77#
3	StCDF-F 144189	TCCATCACAATCACGGAGAA	StCDF-R 144430	GACAAGGCTGCTGGATTAGC	261#
4	StCDF-F 144189	TCCATCACAATCACGGAGAA	StCDF-R transp	CTATATGTGCCGTCCCGTTT	362*
5	CDF_F1	AATGAAGCGGCTAAAAGCTC	RH Phase 0 R3	GGTTCCGGTTCGTTCC	102*
6	StCDF-F 144189	TCCATCACAATCACGGAGAA	RH Phase 0 R3	GGTTCCGGTTCGTTCC	196*
7	StCDF-F 142886	ACAACCTCATCCCCTCTTTC	StCDF-R 142886	CAGGGCATATAATCCTTGTGC	628
8	M13	GTTTTCCCAGTCACGAC	M13	CAGGAAACAGCTATGAC	-

Supplementary table 4. Cultivars cloned in order to find phased haplotypes.

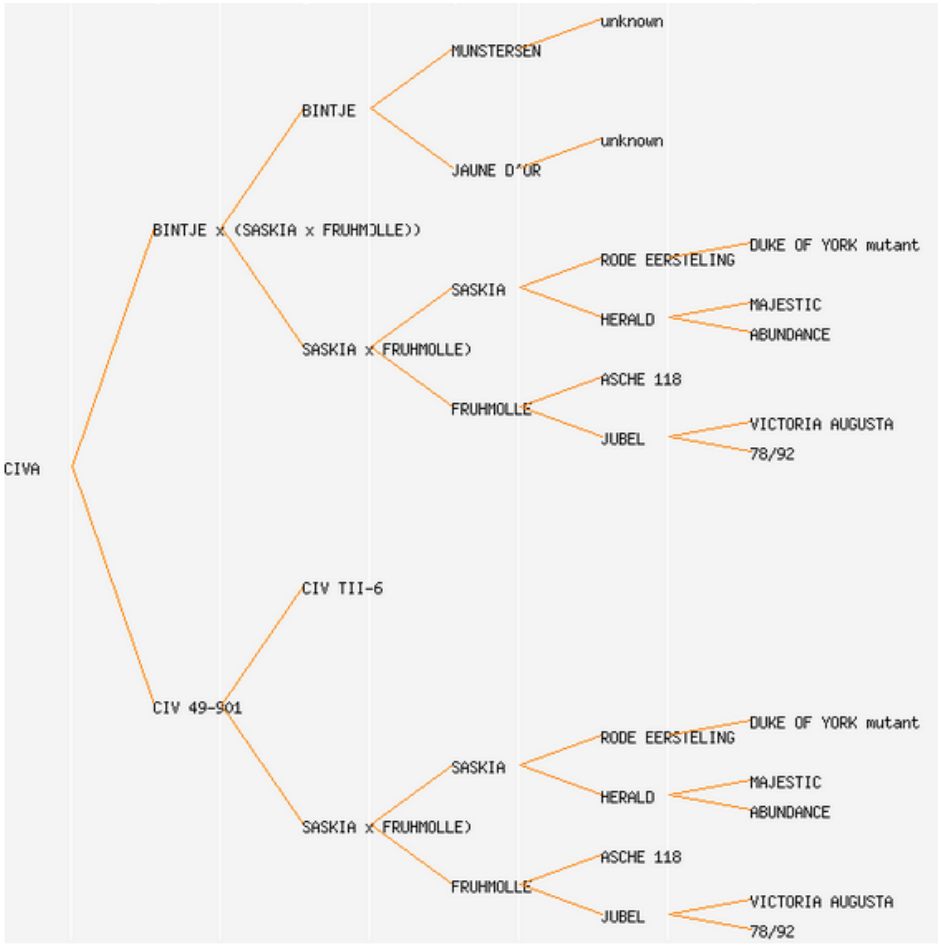
Cultivar	Phenotype
AR96	-
Alpha	4.5
Aurora	7.9
Binella	7.4
Civa	8.3
Early Rose	6.3
Herald	8.5
IVP4X-061-1	4.4
IVP4X-144-2	7.8
IVP92-057-3	-
Karnico	3.5
Kerpondy	4.4
Kepplestone Kidney	6.2
RH4X-549-3	7.2
Wur038	-

Supplementary table 5. Information about haplotypes present in different cultivars. Cultivars presented in bolt represent the cultivars sharing haplotype with the DM reference genome and which were directly extracted from the set of non-cloned cultivars.

Haplotype	Cultivars	Haplotype	Cultivars
1	Herald	**24**	Ar96
2	Herald Kerpondy	**25**	Wur038
3	Ar96	**26**	Karnico
4	Ar96	**27**	IVP4X-061-1
5	Early Rose	**28**	Aurora
6	Ar96 IVP4X-144-2	**29**	Early Rose
7	Herald	**30**	Civa
8	Herald	**31**	CIva
9	Binella	**32**	CIva
10	Binella	**33**	Alpha
11	IVP4X-144-2	**34**	Early Rose
12	IVP4X-144-2	**35**	Kerpondy
13	IVP4X-144-2	**36**	Alpha
14	IVP92-057-3	**37**	Karnico
15	Alpha	**38**	IVP92-057-3
16	Alpha Aurora	**39**	Karnico
17	Wur038	**40**	DM reference StCDF1.2 Early Rose IVP92-057-3 RH4X-549-3 **Urgenta** **Saskia** **Vr-98-377** **Vtn 62-33-3**
18	Wur038	**41**	Alpha
19	Wur038	**42**	Kerpondy
20	Kepplestone kidney	**43**	Kerpondy
21	Wur038	**44**	Karnico IVP4X-061-1
22	Aurora	**45**	Aurora
23	Ar96		



Supplementary figure 1. Pedigree information of cultivar Aurora, showing that cultivar Alpha is one of his ancestors.



Supplementary figure 2. Pedigree information of cultivar Civa, showing that cultivar Herald is one of his ancestors. This figure also shows that Herald is an ancestor of cultivar Saskia.

ID	NAME	crosscomb	orig	jaar	breeder_code	breeder	reference
744	Offspring ALPHA	EARLY ROSE x SEBEC	USA	1874			6
745	Offspring ALPHA		HOL	1895		G. VEENHUIZEN	
746	Offspring ALPHA		HOL	1904		K.R. VELTHUIS	
743	Offspring ALPHA	PAUL KRUGER x PREFERENT	HOL	1925	DORST A 10	FRIESE MIJ. VAN LANDBOUW	2/54-91 4 12/25-07 13/91-93 15

Supplementary figure 3. Figure showing cultivar with same name but different pedigree information.