Contribution of Natural Genome Diversity to Domestication of Potato

Submitted by Haoyang Duo

Student number: 890225201050

Plant breeding specialization, MSc Plant Science (MPS)

Supervisor: Christian Bachem

Laboratory of Plant Breeding, Wageningen University

September, 2014

Abstract

Plant maturity of potato (*Solanum tuberosum*) is an essential criterion for the selection of potato breeding. It was shown that StCDF1 gene belonging to the DOF (DNA-binding with one finger) family regulates tuberization and plant cycle length. Three alleles of StCDF1 have been identified, namely StCDF1.1 (wild type), StCDF1.2 and StCDF1.3 which contain a 7bp footprint insertion and a 865bp transposon insertion respectively. Genotypes containing the two latter alleles perform an early tuber-bearing behaviour.

In order to explore the allele diversity of StCDF1 in potato genome, a total of 239 accessions representing 183 wild species originating from 14 South American countries as well as 17 commercial cultivars were selected and around 2.4 kb of target sequence of StCDF1 were sequenced by using Pacific Biosciences single-molecule sequencing technology. Homo-polymer compression reduced the sequence error rate significantly from PacBio sequencing reads. A phylogenetic tree was constructed combining various haplotype sequences and geographic information (latitude) by using the homo-polymer-compressed sequences retrieved from PacBio platform. The allele variation was observed between high latitude area (between 11N/S and 30 N/S) and the equatorial region (between 10N and 10S) from the phylogenetic tree. Further analysis is needed to draw solid conclusions in terms of the relation between latitude and allele variation in StCDF1.

Table of Contents

Introduction

Potato crop

Potato (*Solanum tuberosum*) originates from the Andes in South America and is the third important food crop in the global economy after rice and wheat (Kloosterman et al., 2013). It is vegetatively propagated by potato tubers formed from a underground organism called stolon.

Potato is a short- day plant. Tuber formation is strictly induced under short day condition of 12 h or less due to its equatorial origin, although critical night length and the strength of photoperiodical response vary among different genotypes (Jackson, 1999; Morris, et al., 2014; Kloosterman et al., 2013).

Potato genome comprises a complex heterozygosity ranging from homozygous, heterozygous to hemizygous levels. Ploidy levels of potato range from diploid (2n=2x=24), to triploid (2n=3x=36), to tetraploid (2n=4x=48), to pentaploid (2n=5x=60). Wild species even possess hexaploid (2n=6x=72) (Spooner et al., 2005; Cai et al., 2012). Most potato cultivars, however, are tetraploid (2n=4x=48), highly heterozygous and suffer severely from inbreeding depression.

The whole genome sequence of potato released in 2011 is in total 844 mega base (Mb) large (The Potato Genome Sequencing Consortium, 2011). The laboratory of plant breeding of Wageningen University was responsible for part of the Potato Genome Project that was to isolate and analyse the genetic factors that contribute to the Plant Maturity phenotype mapping onto chromosome 5. This research is the continuation of the genome project. The complete sequence of chromosome 5 was determined by Whole Genome Sequencing (WGS) technology and by using a BAC-by-BAC approach in the *S. tuberosum* diploid heterozygous RH genome.

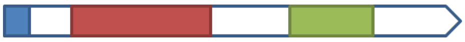
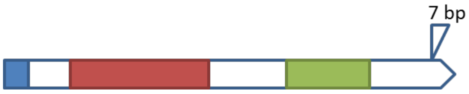
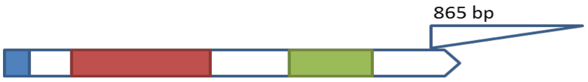
Plant maturity trait includes many pleiotropic sub-traits such as the onset of leaf senescence, life cycle length, time point of tuberization and potato yield. It is one of the first selected traits in potato breeding and cultivation, especially for European potato types that have been long-day acclimated for tuberization. The investigation in allele diversity for tuber initiation and development is of vital importance in potato breeding for local adaption of cultivation at different latitudes across the world. For plant maturity trait, more natural alleles may be present in wild germplasm which is known to contain a tremendous diversity in terms of genetic variation. Identification of new alleles will help the understanding of tuberization to extend growing regions that are hampered by extreme day length variation as well as revealing the complexity of potato genome by developing molecular markers (Kloosterman et al., 2013).

Current study on StCDF1 gene

Plant maturity trait has been mapped under a major-effect quantitative trait locus (QTL) on chromosome 5. It seems to be regulated by a transcription factor called StCDF1 that was fine mapped to this locus. Allelic variation at this locus was shown to be responsible for the major Plant Maturity traits with the help of complementation analysis in late-maturing genotypes (Kloosterman et al., 2013).

StCDF1 (PGSC003DMG400018408) stands for *S. tuberosum* Cycling DOF (DNA-binding with one zinc finger) Factor. It is a transcription factor mediating in the pathway between circadian clock and the StSP6A mobile tuberization signal. Three alleles have been identified in two diploid potato populations, namely, StCDF1.1 coding for late tuberization, StCDF1.2 and StCDF1.3 both coding for early tuber development. Different sizes of insertion and excision events were found in the 3' region of early alleles (a 7bp insertion from transposon excision in StCDF1.2 and an 865bp insertion representing a transposon insertion in StCDF1.3) (Table 1). In order to exploit gene diversity of StCDF1 in potato associated with tuber maturity, this research will mainly focus on exploring wild type resources through sequencing of a large number of South American genotypes. Cultivated materials will also be studied in a parallel project.

Table 1 Cartoon showing the structure of StCDF1 variants. Blue blocks indicate promoter region, red and green parts are exons. 7 bp and 865 bp site is the transposon position.

Name	Structure of variants
StCDF1.1	
StCDF1.2	
StCDF1.3	

The main role of StCDF1 was verified to indirectly promote expression of the tuber initiation signal StSP6A, resulting in promoting tuberization (Kloosterman et al., 2013). The pathway (Figure 1) below shows the relations between tuber development-related genes. StCO1/2 and StSP5G functioning as repressors of tuberization are down-regulated by StCDF1. The sequence of StCDF1 is around 2.8 kb, containing three conserved domains, the DOF domain (DNA-binding) and the StGI1 and StFKF1 binding domains. Ubiquitination of the StCDF1 protein by FKF1 normally results degradation of the protein. The absence of StCDF1 inhibition to StCO1/2 allows these proteins to induce StSP5G which is an inhibitor of the mobile signal StSP6A. The loss of StFKF1 binding domain in the C terminal of StCDF1.2 and StCDF1.3 due to the insertions stabilizes StCDF1 protein resulting in consistent expression in leaf and causes a constitutive repression of StSP5G, allowing StSP6A to express and leading to early tuberization phenotype (Figure 1)(Kloosterman et al., 2013).

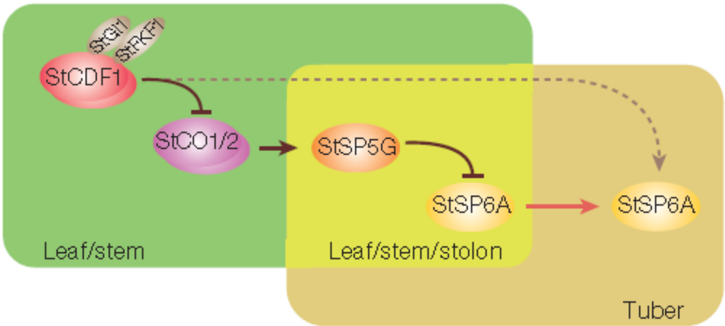


Figure 1 Regulation model of tuber formation. Green represents aerial plant organs, the tuber is represented in brown and the overlap represents leaf, stem and stolon. Arrows (→) represent induction and lines with vertical end (-|) represent repression. Transport is represented by the red arrow. StCDF1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

Pacific Bioscience sequencing technology

Pacific Bioscience (PacBio) technique, known as a third generation sequencing platform (Koren,et al., 2012), uses a single molecule real time sequencing done in wells on a chip, which is called single molecule real time (SMRT) cell, developed by Pacific Biosciences Inc with high consensus accuracy and long read lengths (up to 20 kb) (Pacific biosciences. Inc). DNA polymerase enzyme is immobilized on the “zero-mode waveguide” with a single molecule of DNA as a template. Phosphor-linked nucleotides attached with fluorescent tag on one end are provided in the substrate. When DNA polymerase starts to work, a phosphor-linked nucleotide will be incorporated along the template and will release the fluorescent molecule which will be read and recorded as sequence result at the same time. This technology can sequence DNA fragment up to more than 10 kb. Lone reads of sequencing is an advantage for assembly of large complex genomes. New advanced technology promotes more possibilities to scientific studies and vice versa.

Single-molecule sequencing technology applies a novel strategy by using a template structure called “SMARTbell”. This template structure consists of a double-stranded portion, the insert of interest, a single-stranded hairpin loop on either end which provides a primer binding site (Figure 2A). Once a primer attaches to the hairpin loop, DNA polymerase start to extend using one strand as a template and displace the other. When polymerase returns to the 5’-end of the primer, it starts strand displacement of the primer and continues to synthesize DNA. Therefore, both sense and antisense strand sequences are obtained from single-molecule sequencing technology (Figure 2B) (Travers et al., 2010).

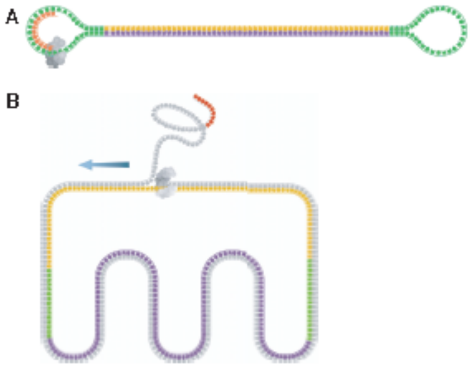


Figure 2 SMARTbell structure. A: the SMARTbell consists of a double-stranded region (purple and yellow strands) flanked by two hairpin loops (green part). The single –stranded hairpin loops provide a primer binding site (orange part). Sequence of interest is inserted between green parts. B: DNA polymerase (grey one) displaces the primer when it returns the 5'end of the primer and continue to synthesize DNA along the other strand (Travers et al., 2010).

Research objectives

The aim of this research is to identify allele diversity at StCDF1 locus in wild potato species originating from South America with the attempt to accumulate knowledge on evolution and domestication of potato by the means of PacBio sequencing technology.

The previous study has identified alleles varying in the 3' region where a transposon insertion or excision occurred in StCDF1 (Kloosterman et al., 2013). Further new allele variations in wild germplasm are highly expected to exist. Such allelic variants may be additional variants in the 3' region or in the promoter region that could be linked to functional variations. SNPs related to early tuberization could also be explored and identified. Variations in sequence may reveal the footprint of potato evolution and domestication along widespread latitudes.

Material and methods

Plant material and genotype selection

In total 239 accessions representing 183 different wild species and 12 hybrids from 14 South American countries were sampled from the online SolRgene dataset (containing totally approx. 5000 species) supported by WUR (Vleeshouwers et al., 2011). In order to discover as many diverse alleles as possible, in principle, at least 1 genotype was selected from each species of one country of origin to cover a wide geographical habitats in latitudes from the southern United States of America to the northern Chile. Seventeen reference genotypes were selected for sequencing, containing early to late commercial European tuber-bearing cultivars with different StCDF1 alleles (Appendix III). The selecting steps are illustrated as below in Figure 3.

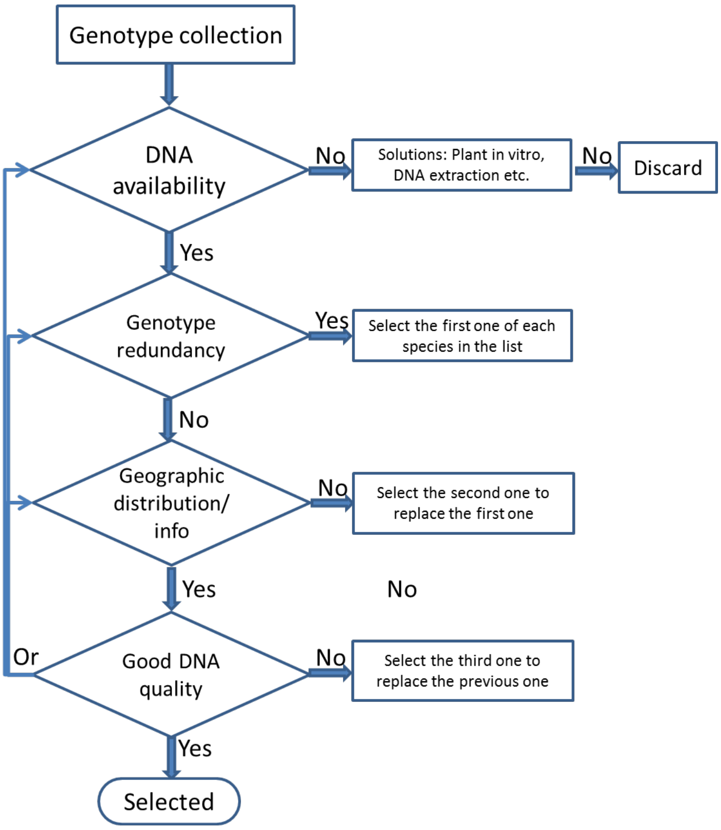

Genotype collection
DNA availability
No
Solutions: Plant in vitro, DNA extraction etc.
No
Discard
Yes
Genotype redundancy
Yes
Select the first one of each species in the list
No
Geographic distribution/ info
No
Select the second one to replace the first one
Yes
No
Or
Good DNA quality
No
Select the third one to replace the previous one
Yes
Selected

Figure 3 Genotype selection procedures.

DNA of all genotypes had been extracted from potato leaves before this project and stored at -80°C, previously used for the Potato Genome Sequencing Consortium project. No extra DNA extraction was necessarily needed.

Primer design and PCR amplification of StCDF1

StCDF1 gene specific forward primer, StCDF1-F (5’-CAGCAAACAAACCACACACA-3’) and reverse primer, StCDF1-R (5’-GGAATCAGTACACATCTCTCG-3’) were designed based on the sequence of StCDF1.1, StCDF1. 3 from RH population and StCDF1 genomic sequence (PGSC003DMG400018408) from DM population by using Oligo 6.0 software. Each 20 µl reaction mix included 4 µl Phusion HF buffer, 1 µl of each primer, 0.8 µl10mM dNTPs, 0.2 µl Phusion DNA polymerase 1 µl of genomic DNA of each genotype (15-30 ng/µl) and 12 µl MQ water. The PCR reaction was carried out on a Thermal Cycler (Bio-Rad) as follows: 30 s initial denaturation at 98 °C followed by 30 cycles of (10s at 98 °C, 30 s at 67 °C, and 90 s at 72 °C) and a final extension of 10 min at 72 °C. Amplified products were tested and separated by electrophoresis in a 0.8% agarose gel (1 x TAE buffer at 100 mA for 50 min) using GelRED and visualized under UV light.

Template preparation for PacBio RS sequencing

Gene specific primer pairs combined with 16 sets of barcodes (totally 256 unique primer combinations) were used in PCR for PacBio template preparation (barcoded primer sequence in Appendix I). Products were amplified from all selected genotypes by applying the same PCR protocol as above. Equimolar amplicons of each genotype were pooled and purified by using QIAquick PCR purification kit (QIAGEN, GmbH, Germany). The quality and quantity of purified PCR products was determined through spectrophotometry using the Nanodrop 8000 platform (Thermo Scientific) with a concentration of 45.5 ng/µl, 260/280 at 1.85; and agarose gel electrophoresis (0.8%), respectively. A total of approximately 4 µg of amplified PCR products was delivered to Keygene N.V., Wageningen, the Netherlands for PacBio sequencing.

Data processing and analysis

The barcode-deconvoluted sequence data from PacBio RS platform was received from the provider in which both sense and antisense sequences were present. Consequently data normalization was performed resulting in all the sequences in 5’-3’ order. Since the two main sequence errors from PacBio were nucleotide errors and homo-polymer length errors, the later error was removed by homo-polymer compression in order to acquire high quality data, leading to each homo-polymer run was replaced by a single nucleotide of the same type, for example, a run of “AAAAA” is placed by a single “A” (Au et al., 2012). Next, with the aim to reduce the error rate further, clustering reads per genotype was taken place to discard evident errors by using Dynamic programming and a distance matrix was produced based on sequence differences, followed by sequence assembly executed by MIRA software.
Sequence alignment was accomplished by MEGA 5.2 by Muscle method (Figure 4).

The selected genotypes except reference commercial cultivars were mapped to their geographic locations by Google Map Engine Pro based on their latitude information. A phylogenetic tree was constructed based on the alignment of homo-polymer compressed sequences due to the intrinsic complexity of PacBio sequence data by using MEGA and SeaView 4.5.2 by neighbour-joining.

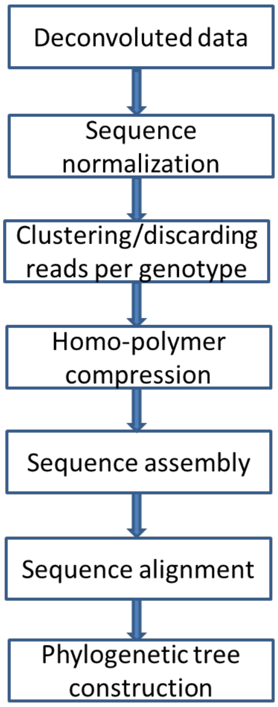


Figure 4 Data processing flowchart

Results and discussion

Intermediate results

Since the current identified alleles vary in C terminal region of StCDF1, more variations could be found in that region very likely. A triple-primer strategy, one forward primer and two reverse primers used in a single PCR reaction, therefore, was designed in the initial experiment plan (Figure 6). The Forward primer was assigned in the minus 200 region based on the StCDF1 5' flanking sequence and one reverse primer (Reverse primer 2) was in the 5' region of the transposon and the other (Reverse primer R1) right after the insertion site, respectively (Figure 6). The amplicons will be span the transposon insertion and excision site and the approximately 2.5 kb upstream of this.

The mixed-primer long-range PCR will give rise to potentially at least 3 different fragments depending on the template (Figure 5). Fragment 1.1 is the wild type allele. Fragment 1.2 is the allele with 7 bp insertion; Fragment 1.3 is the longest fragment with 865 bp transposon.

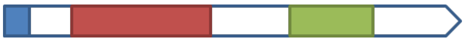
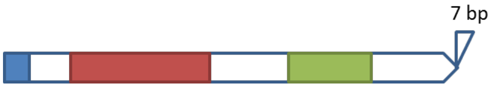
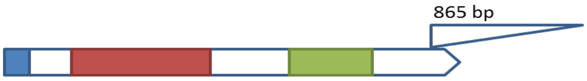
No.	PCR Fragment	Expected length (bp)
1.1		2500
1.2		2507
1.3		3365

Figure 5 Possible fragments from the initially designed primers.

Many different primer combinations were tested in 6 known diploid genotypes, namely, 3130 (StCDF1.2, StCDF1.3), 3027 (StCDF1.1homozygous), RH (StCDF1.1, StCDF1.3), SH (StCDF1.2 homozygous), C (StCDF 1.1, StCDF1.2) and E (StCDF1.2, StCDF1.3). However, no fragment containing the transposon could be amplified by any primer combinations (Figure 7).

The reason of no amplification of the 865 bp transposon might be the potential template competition for primers during PCR procedure. Shorter DNA templates, in this case, StCDF1.1 or/and StCDF1.2, outcompeted (865bp) longer template (StCDF 1.3) to be amplified by the primers in all tested genotypes. Another attempt carried out to verify primer workability was to use a BAC clone of StCDF1.3 allele as a test template where no template competition existed. Only the primer pair with Reverse primer 2 worked successfully indicating the primer pair assigned for the shorter target sequence worked predominately in PCR. Therefore, the triple primer strategy for multiple sequence amplification could not be feasible to realize.

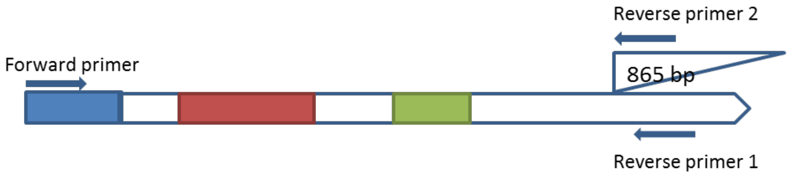


Figure 6 Primer position in StCDF1.

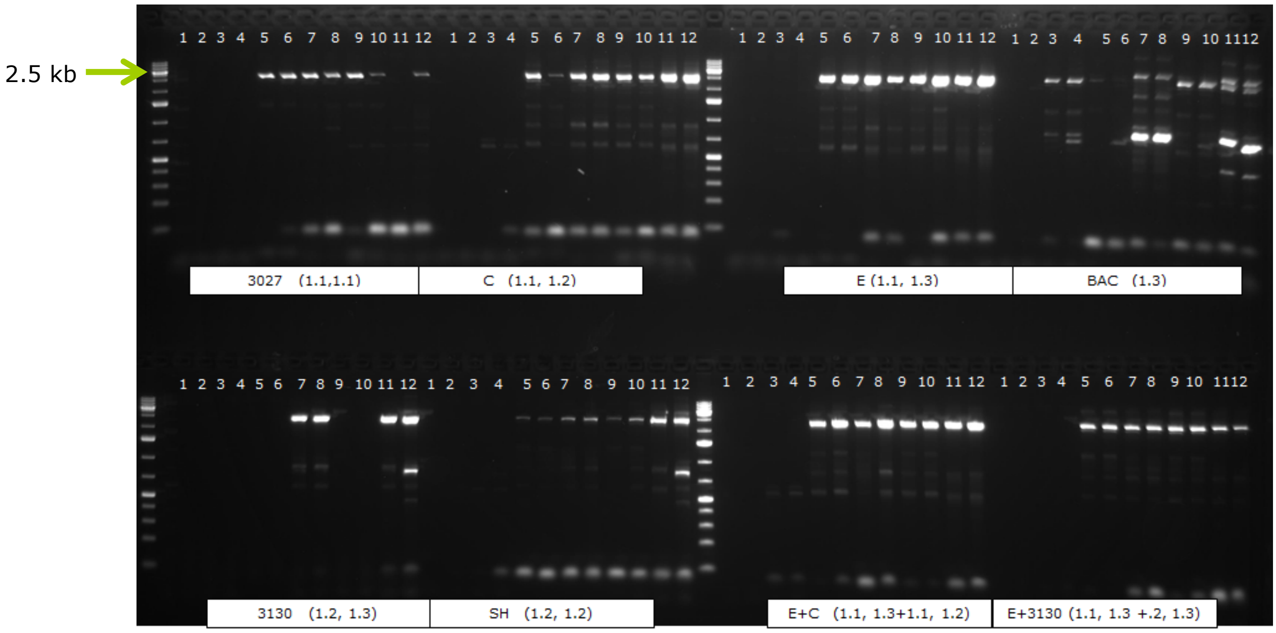


Figure 7 electrophoretic gel image of StDCF1 target region in genotype 3027, C, E, 3130, SH and a BAC clone of StCDF1.3 for primer testing. Lane 1-4 indicate the Forward primer and the Reverse primer 2 were tested, Lane 5-8 is from the combination of Forward primer and the Reverse 1, Lane 9-12 contains all the three primers. Green arrow indicates the position of 2.5 kb from the DNA ladder.

Since the previous strategy attempting for transposon amplification was hindered, it was altered to another design which was described in Material and methods chapter, new reverse primers were designed in the 5' region of StCDF1 before the transposon insertion site. The results in rest of this chapter are reported on the new strategy with the aim to target the sequence region before transposon site of StCDF1.

Data output from PacBio RS platform

The forward primer started in the promoter region and the reverse primer 20 before the insertion site. PCR products amplified from the specific gene primers was expected to be 2349 bp and primer position was shown in Figure 8 based on genomic DM sequence (PGSC003DMG400018408).

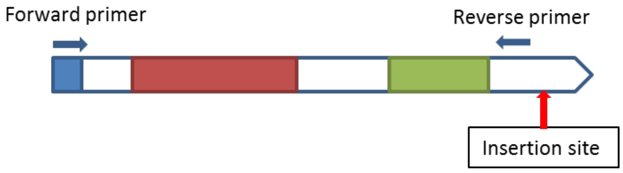


Figure 8 Primer positions in StCDF1.

13,627 reads, 254 primer combinations were retrieved from PacBio RS platform. Sequence read lengths generated ranged from 70 bp to 16,384 bp long with the average read length of 2544 bp and 1607 bp before and after homo-polymer compression respectively. All the amplicons were sequenced 2 or more cycles on SMRTbell. The depth of average sequencing coverage was 54 × to 28× per genotype with a standard deviation of 28. The nucleotide error rate was between 2-4 %. (results received form T. Borm). Due to time limitation the analysis of homo-polymer compressed sequences was the main focus of this report.

As introduced before, the characteristics making PacBio sequencing technology distinguish from the second generation sequencing technologies includes single molecule template, lower cost per base, easy sample preparation, significantly faster run times and long-read lengths to enable de novo sequencing and simplify data analysis.

All sequencing technology manifests bias and errors including PacBio. Bias and errors can be introduced at the library construction, sequencing or computational stages.

During library preparation, all the samples were pooled in an equal-molar way, while the Nanodrop 8000 platform used for DNA quantification also measured the primers left in the PCR, resulting in different amount of available templates of each genotype for sequencing, then further leading to huge deviation in (low) coverage depth.

PacBio RS can generate long length of reads (up to 10 kb). However, the PacBio instrument generates reads that average only 82.1%–84.6% nucleotide accuracy, with uniformly distributed errors dominated by point insertions and deletions (Koren et al., 2012). This high error rate could result in unreliable mismatched alignment between reads, and complicates analysis as the pairwise differences between two reads are approximately twice their individual error rate. Such a high error rate is not acceptable by most genome assemblers which tolerate only 5% -10% error rates. Moreover, it is not computationally feasible by simply increasing the alignment sensitivity of traditional assemblers. This holds true to MIRA assembler used in this study as well. The alignments was performed by default setting, yet it needs to take in to consideration that the alignments were very sensitive to the assembly parameters such as nucleotide substitution and gap penalty (personal communication with T. Borm).

Although the sequencing accuracy can be greatly improved by sequencing both sense and antisense strands as well as by approach called “circular consensus sequencing” which uses the additional information from multiple passes across the insert (Koren et al., 2012). In our case, considerable part of the reads were obtained from 2 or even 1 full pass across the insert which reduced the quantity and the credibility of data tremendously.

Additionally, double stranded DNA can be read by PacBio instrument under continuous high-energy laser excitation, which can result in chimeric reads if the sequencing reaction processes both strands of the DNA and homo-polymer error since the detector cannot accurately recognize constant light excitation of same fluorescent molecule (Koren et al., 2012).

To enable to check the error presence of chimeric reads, technical control should be concerned in later experimental design which was missing in this research. Only the biological control, the cultivars with known allele sequence of StCDF 1 were included. A spare primer combination could be used to quantify this type of error if it is detected.

Long homo-polymer runs during sequencing process are a main source of errors from PacBio sequencing (Quail et al., 2012). Since both strands of DNA template were sequencing at the same time, the function speed of the polymerase in the sequencing reaction for the two strands could be different, making difficulty in detecting the light excitation for each strand, especially for the detection of homo-polymers. Hence homo-polymer length error is the major error source (Koren et al., 2012; personal communication with C. Bachem and T. Borm)

To reduce the homo-polymer error, the approach of homo-polymer compression (HC) was used to allow preliminary alignment study. HC transformation has been proven to be useful in seeking possible alignment matches for pyrosequencing reads (454 platform) (Au et al., 2012; Gilles et al., 2011). By replacing the homo-polymers with a single nucleotide, the error rate is significantly reduced at the cost of no longer being able to easily identify corresponding amino acid residues (personal communication with T. Borm).

The assemblies were performed before and after homo-polymer compression.
From the uncompressed sequences, 3 “allele” (assemblies) were identified per genotype on average, while only 1 “allele” (assembly) of each genotype was identified in the homo-polymer compression reads. This odd results proposed the hypothesis

that most of the sequence differences between alleles were within the homo-polymers, which directed us to cluster all the reads per genotype and re-assemble them to verify the hypothesis.

All the reads without homo-polymer compression per genotype were aligned and grouped based on the distance score by using Dynamic programming. 1063 dubious reads (7.8%) were removed by the automatic check of barcode assignment. The groups containing less than 2 reads (3541 reads) were removed. Furthermore, 370 reads with extreme lengths were removed as well. Then the remaining reads from each genotype which were homo-polymer compressed were clustered and aligned by neighbour joining, resulting in 3 clusters per genotype on average. The number of clusters for each genotype gave an indication of possible allele number per genotype.

Another approach to analyse the data was suggested by combining and comparing sequence data from 454 and Illumina to correct the innate error in long, single molecule sequencing. The data accuracy was improved from ~85% to over 99.9% (Koren et al., 2012).

Besides utilizing data from the second generation sequencing, new technology also intrigues us to explore DNA sequencing possibilities. Oxford Nanopore technology, rather than using a sequencing-by-synthesis approach, it employs an exonuclease-based ‘sequencing by deconstruction’ approach. A strand of DNA is fed through a biological pore containing an exonuclease and the various bases are identified by measuring the difference in their electrical conductivity as they pass through the pore. Distinct advantages of this system include a low instrument fabrication and operation cost due to no need for labelled nucleotides and optical detection systems (like the laser in PacBio). A clear disadvantage, however, is that redundant sequencing (and the associated high accuracy) is not feasible because the template is digested by exonuclease during sequencing. Yet, this weakness could be limited by replacing the exonuclease coupled to the nanopore with a DNA polymerase. However, nanopore-based sequencing is still under development (Munroe & Harris, 2010).

From the homo-polymer compressed reads, some accessions contain more haplotypes than its ploidy lever, for instance, *S. infundibuliforme* is diploid (Spooner and Hijmans, 2001) while 5 different alleles was recognized from the reads; the same situation with *S. coelestipetalum*. Ploidy level is also a good reference to verify the analysis results.

Phylogenetic analysis

In order to aid the phylogenetic analysis, geographic information (country of origin, latitude and longitude) was linked to the wild genotypes resulting in a unique distribution map of the wild potato species sampled in this study (Figure 9). In total of 134 accessions representing 113 wild species and 4 interspecific hybrids from 11 countries of which latitude information were available were mapped on the map. The species from URY, BRA, and USA were not included. All the accessions were categorized into three groups according to the latitudes, Group N includes accessions originating 40N to 12N, Group E from 11N to 10S and Group S from 11S to 30S), with an equal latitude spread range approximately.

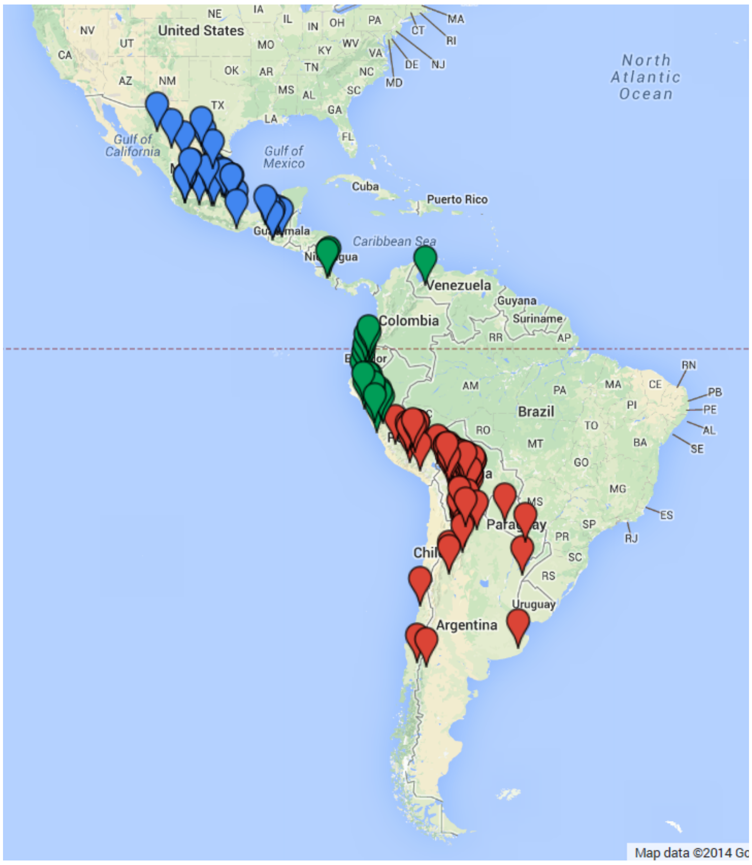


Figure 9 Geographic distribution of selected wild genotypes.

Blue: Group N, green: Group E, red: Group S. Dash line indicates the equator.

The interspecific hybrids and the commercial cultivars were removed, while the corresponding latitude information was added to the names of wild species with the aim to find the relation between allele diversity and latitude. A total of 102 out of 187 species were mapped in the tree with latitude information attached, using muscle alignment by Neighbour joining (Figure 10). (For the clear tree information, see Supplementary file 1)

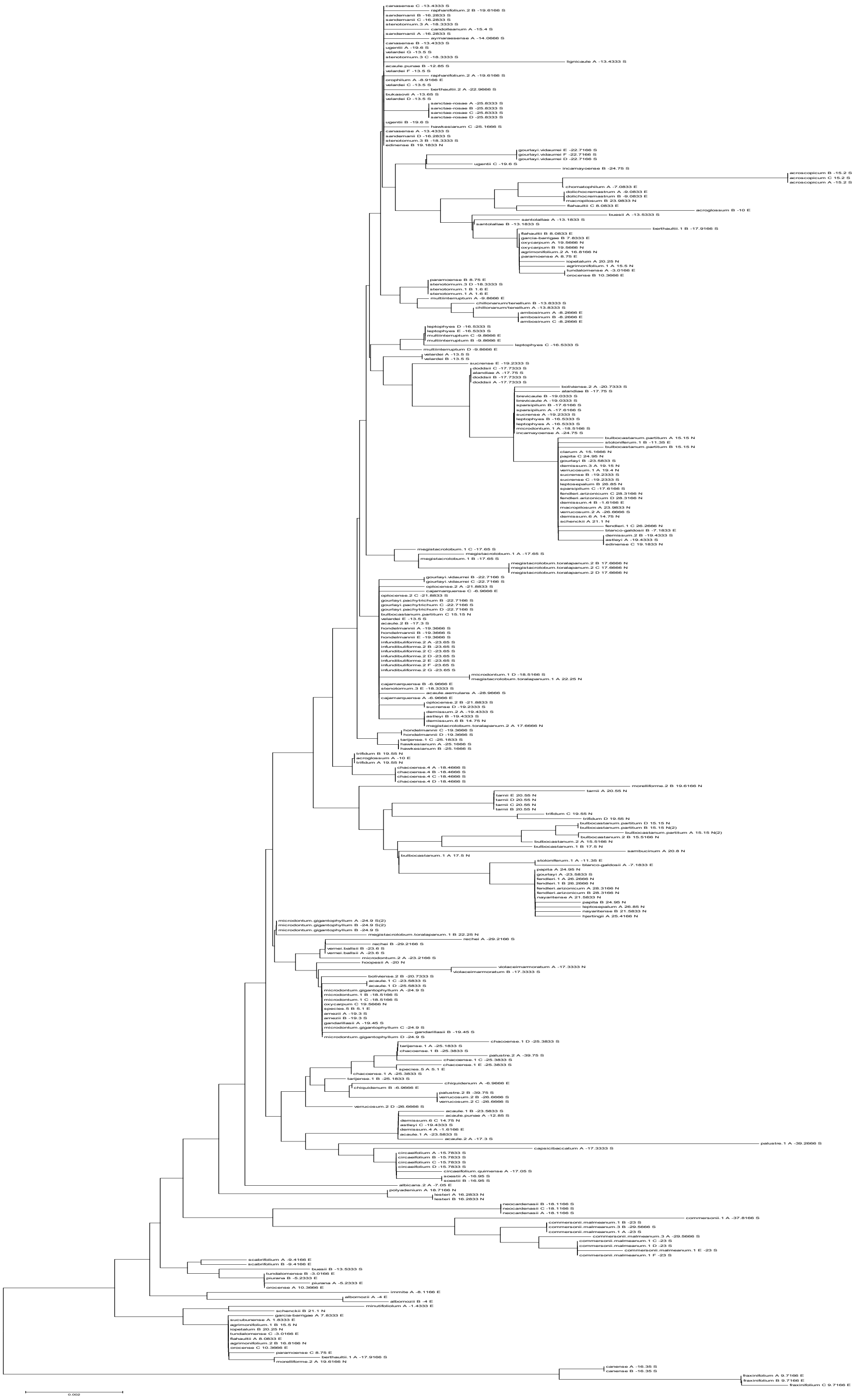


Figure 10 Phylogenetic tree of the sampled wild species.

Header structure: accession name _allele name_latitude_group name (N, E, S)

Based on the categorization of latitudes, the accessions from the same geographic group tend to fall into one branch indicating various alleles does present in this genetic pool. This sheds light on possible allele variance does occur between high latitude area (between 11N/S and 30 N/S) and the equatorial region (between 10N and 10S) from the phylogenetic tree. Some haplotypes are spread along the latitudes showing same alleles may exist in diverse regions (Supplementary file 1).

In this research, only the latitude information of each accession was used for the analysis, while the altitude could also be taken into consideration when analysing. Further analysis is needed to draw any solid conclusions.

Conclusions

Homo-polymer compression reduced the sequence error rate significantly from PacBio sequencing reads. The selected wild specie collection was representative in terms of a wide distribution along latitudes. The allele variation was observed between high latitude area (between 11N/S and 30 N/S) and the equatorial region (between 10N and 10S) from the phylogenetic tree. Further analysis is needed to draw solid conclusions in terms of the relation between latitude and allele variation in StCDF1.

References

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. PLoS One, 7(10), e46679.

Cai, D., Rodríguez, F., Teng, Y., Ané, C., Bonierbale, M., Mueller, L.A., Spooner, D.M.(2012) Single copy nuclear gene analysis of polyploidy in wild potatoes (Solanum section Petota). BMC evolutionary biology 2012, 12(1).

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., & Martin, J. F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC genomics, 12(1), 245.

Jackson, S. D.(1999) "Multiple signalling pathways control tuber induction in potato." Plant Physiology 119.1 (1999): 1-8.

Kloosterman, B., Abelenda, J.A., Gomez Mdel, M., Oortwijn, M., de Boer ,J.M., Kowitwanich, K., Horvath, B.M., van Eck, H.J., Smaczniak, C., Prat, S. (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 2013, 495(7440):246-250.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., ... & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology, 30(7), 693-700.

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

Munroe, D. J., & Harris, T. J. (2010). Third-generation sequencing fireworks at Marco Island. Nature biotechnology, 28(5), 426-428.

Pacific biosciences. Inc, assessced on August 31st, 2014 at <http://www.pacificbiosciences.com/products/>

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics, 13(1), 341.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., ... & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. Genome Biol, 14(5), R51.

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 2005, 102(41):14694-14699.

Appendices

Appendix I. 16 pairs of barcoded primers

Bar-coded primer	Sequence
F1	TCATGAGTCGACACTACAGCAAACAAACCACACACA
R1	CGTGTGCATAGATCGCGGAATCAGTACACATCTCTCG
F2	TATCTATCGTATACGCCAGCAAACAAACCACACACA
R2	ATGTATCTCGACTGCAGGAATCAGTACACATCTCTCG
F3	ACGTACGCTCGTCATACAGCAAACAAACCACACACA
R3	CGATGACGTCGCTGTAGGAATCAGTACACATCTCTCG
F4	TGTGAGTCAGTACGCGCAGCAAACAAACCACACACA
R4	CACACGTAGTCTGCGCGGAATCAGTACACATCTCTCG
F5	CTGCTAGAGTCTACAGCAGCAAACAAACCACACACA
R5	CGAGCTATCTCATACTGGAATCAGTACACATCTCTCG
F6	TCATGCACGTCTCGCTCAGCAAACAAACCACACACA
R6	CAGCGACTGTGATACTGGAATCAGTACACATCTCTCG
F7	AGAGCATCTCTGTACTCAGCAAACAAACCACACACA
R7	TGTCGCATCATATGATGGAATCAGTACACATCTCTCG
F8	CGCATCGACTACGCTACAGCAAACAAACCACACACA
R8	GCTGTGATCTACGTCTGGAATCAGTACACATCTCTCG
F9	CGTAGCGTGCTATCACCAGCAAACAAACCACACACA
R9	TGAGTAGCATGACACGGGAATCAGTACACATCTCTCG
F10	CGATCATCTATAGACACAGCAAACAAACCACACACA
R10	CTGCGTGCGCGATAGTGGAATCAGTACACATCTCTCG
F11	CGACGTATCTGACAGTCAGCAAACAAACCACACACA
R11	CGCGTGCAGAGTGTCAGGAATCAGTACACATCTCTCG
F12	CACGTCACTAGAGCGACAGCAAACAAACCACACACA
R12	ATATCAGTCACGTCTGGGAATCAGTACACATCTCTCG
F13	TGTCGCAGCTACTAGTCAGCAAACAAACCACACACA
R13	ACGATCACTACAGTGCGGAATCAGTACACATCTCTCG
F14	AGTCGCATGACTGTGTCAGCAAACAAACCACACACA
R14	GTCGAGTAGCACTACTGGAATCAGTACACATCTCTCG
F15	CAGTACTGCACGATCGCAGCAAACAAACCACACACA
R15	TGAGCAGATCTCGCATGGAATCAGTACACATCTCTCG
F16	CACTGATCGATATGCACAGCAAACAAACCACACACA
R16	GTACACTAGTGCACATGGAATCAGTACACATCTCTCG

Appendix II. Complete wild genotype list

Nr.	**species P4**	**Country**	**Lattitude**	**Longitude**
1	chacoense	ARG		
2	microdontum gigantophyllum	ARG	-24.9	-65.65
3	tarijense	ARG	-25.1833	-65.8
4	acaule	ARG	-23.5833	-65.2
5	vernei	ARG		
6	berthaultii	ARG	-22.9666	-65.45
7	boliviense	ARG	-25.3333	-65.8333
8	gourlayi	ARG	-23.5833	-65.3333
9	gourlayi vidaurrei	ARG	-22.7166	-65.2
10	hannemanii	ARG		
11	hawkesianum	ARG	-25.1666	-65.8666
12	incamayoense	ARG	-24.75	-65.7333
13	infundibuliforme	ARG	-23.65	-66.3
14	megistacrolobum toralapanum	ARG	-22.25	-65.05
15	neorossii	ARG		
16	ruiz-lealii	ARG		
17	sanctae-rosae	ARG	-25.8333	-65.5833
18	setulosistylum	ARG		
19	palustre	ARG	-39.75	-71.3333
20	tuberosum andigena	ARG		
21	microdontum	ARG	-23.2166	-64.9166
22	maglia	ARG		
23	venturii	ARG		
24	commersonii	ARG	-37.8166	-58.2166
25	species	ARG		
26	acaule aemulans	ARG	-28.9666	-67.7
27	commersonii malmeanum	ARG	-29.5666	-57.5333
28	oplocense	ARG	-21.8833	-66.1833
29	verrucosum	ARG	-26.6666	-65.8166
30	vernei ballsii	ARG	-23.6	-65.1333
31	spegazzinii	ARG		
32	kurtzianum	ARG		
33	megistacrolobum	ARG		
34	rechei	ARG	-29.2166	-67.65
35	canense	BOL	-16.35	-67.5833
36	capsicibaccatum	BOL	-17.3333	-66.35
37	chaparense	BOL		
38	circaeifolium quimense	BOL	-17.05	-67.2833
39	gandarillasii	BOL	-18.3333	-65.1666

40	species	BOL	-19.2166	-65.8333
41	soestii	BOL	-16.95	-67.1833
42	acaule	BOL	-17.3	-66.1666
43	arnezii	BOL	-19.3	-64.45
44	avilesii	BOL	-18.6333	-64.15
45	boliviense	BOL	-20.7333	-64.85
46	circaeifolium	BOL	-15.7833	-68.6666
47	gourlayi pachytrichum	BOL	-19.45	-65.2166
48	hondelmannii	BOL	-19.3666	-64.7833
49	megistacrolobum	BOL	-17.65	-66.9833
50	megistacrolobum toralapanum	BOL	-17.6666	-66.5
51	oplocense	BOL		
52	sucrense	BOL	-19.2333	-65.85
53	astleyi	BOL	-19.4333	-65.4
54	subandigena	BOL		
55	hannemanii	BOL		
56	berthaultii	BOL	-17.9166	-65.9166
57	microdontum gigantophyllum	BOL		
58	hoopesii	BOL	-20	-64.4166
59	infundibuliforme	BOL		
60	microdontum	BOL	-18.5166	-64.1166
61	achacachense	BOL		
62	virgultorum	BOL		
63	curtilobum	BOL		
64	doddsii	BOL	-17.7333	-65.1
65	palustre	BOL		
66	alandiae	BOL	-17.75	-65.2
67	candolleanum	BOL	-15.4	-69.0666
68	demissum	BOL	-19.4333	-65.2166
69	okadae	BOL	-17	-67.2333
70	raphanifolium	BOL	-19.6166	-65.75
71	stenotomum goniocalyx	BOL		
72	leptophyes	BOL	-16.5333	-68.1
73	sparsipilum	BOL	-17.6166	-66.3166
74	chacoense	BOL	-18.4666	-64.0666
75	tuberosum andigena	BOL		
76	tarijense	BOL		
77	ajanhuiri	BOL		
78	brevicaule	BOL	-19.0333	-65.2833
79	violaceimarmoratum	BOL	-17.3333	-65.7666
80	neocardenasii	BOL	-18.1166	-64.2
81	stenotomum	BOL	-18.3333	-67.6
82	ugentii	BOL	-19.6	-64.6166

83	commersonii	BRA		
84	fernandezianum	CHL		
85	palustre	CHL	-39.2666	-71.9666
86	sitiens	CHL		
87	tuberosum	CHL		
88	species	CHL		
89	maglia	CHL	-32.9666	-71.5333
90	brachycarpum	COL	1.6	-77.15
91	stenotomum	COL		
92	curtilobum	COL		
93	phureja	COL	1.1666	-77.1666
94	colombianum	COL		
95	moscopanum	COL	2.4	-76.45
96	tuquerrense	COL		
97	species	COL		
98	flahaultii	COL	5.1	-74.0333
99	garcia-barrigae	COL	8.0833	-73.2166
100	orocense	COL	7.8333	-73.2333
101	otites	COL	10.3666	-72.8
102	sucubunense	COL	1.8333	-76.5166
103	fraxinifolium	CRI	9.7166	-84.1
104	longiconicum	CRI		
105	nigrum	CRI	9.9333	-83.8833
106	phureja	ECU		
107	albornozii	ECU	-4	-79.2833
108	curtilobum	ECU		
109	solisii	ECU		
110	tundalomense	ECU	-3.0166	-79.0333
111	moscopanum	ECU		
112	chacoense	ECU		
113	tuberosum andigena	ECU		
114	demissum	ECU	-1.6166	-79
115	albicans	ECU		
116	minutifoliolum	ECU	-1.4333	-78.4166
117	paucijugum	ECU	-0.65	-78.5
118	morelliforme	GTM	14.75	-91.55
119	demissum	GTM	15.5166	-91.4833
120	agrimonifolium	GTM	15.5	-90.95
121	bulbocastanum	GTM	15.15	-90.3
122	clarum	GTM		
123	bulbocastanum partitum	GTM	15.1666	-91.5166
124	demissum	MEX	19.15	-99.9333
125	hjertingii	MEX	25.4166	-100.85

126	papita	MEX	24.95	-103.9
127	polytrichon	MEX		
128	verrucosum	MEX	19.4	-101.6
129	tuberosum andigena	MEX		
130	agrimonifolium	MEX	16.8166	-92.5833
131	cardiophyllum	MEX		
132	ehrenbergii	MEX	21.4166	-102.65
133	fendleri	MEX	26.2666	-105.45
134	species	MEX		
135	polyadenium	MEX	18.7166	-97.3166
136	schenckii	MEX	21.1	-99.7
137	stoloniferum	MEX		
138	tarnii	MEX	20.55	-98.4833
139	bulbocastanum	MEX	17.5	-96.45
140	lesteri	MEX	16.2833	-96.55
141	guerreroense	MEX	19.65	-103.6333
142	trifidum	MEX	19.55	-103.6333
143	brachistotrichum	MEX		
144	bulbocastanum partitum	MEX	16.1666	-92.2
145	edinense	MEX	19.1833	-99.65
146	michoacanum	MEX		
147	matehulae	MEX		
148	pinnatisectum	MEX		
149	brachycarpum	MEX		
150	hougasii	MEX		
151	iopetalum	MEX	20.25	-98.2166
152	morelliforme	MEX	19.6166	-97.05
153	sambucinum	MEX	20.8	-100.4333
154	nayaritense	MEX	21.5833	-102.85
155	oxycarpum	MEX	19.5666	-97.2333
156	macropilosum	MEX	23.9833	-99.7333
157	fendleri arizonicum	MEX	28.3166	-107.35
158	leptosepalum	MEX	26.85	-101.2666
159	demissum	PER		
160	acaule	PER	-15.65	-71.4333
161	chomatophilum	PER	-7.0833	-78.5833
162	stoloniferum	PER	-11.35	-77.3833
163	raphanifolium	PER		
164	soukupii	PER		
165	acroscopicum	PER	-15.2	-72.9333
166	ambosinum	PER	-8.2666	-77.85
167	cajamarquense	PER	-6.9666	-79.1833
168	coelestipetalum	PER		

169	curtilobum	PER		
170	dolichocremastrum	PER	-9.0833	-77.0166
171	juzepczukii	PER		
172	laxissimum	PER		
173	medians	PER		
174	mochiquense	PER		
175	sogarandinum	PER		
176	stenotomum goniocalyx	PER		
177	tuberosum	PER		
178	aracc-papa	PER		
179	chaucha	PER		
180	humectophilum	PER		
181	multiinterruptum	PER	-9.8666	-77.7333
182	piurana	PER	-5.2333	-79.4666
183	stenotomum	PER		
184	marinasense	PER		
185	canasense	PER	-13.4333	-71.85
186	acaule punae	PER	-12.85	-74.85
187	albicans	PER	-7.05	-78.6
188	bukasovii	PER	-13.65	-73.3833
189	limbaniense	PER		
190	acroglossum	PER	-10	-76.5
191	paucissectum	PER		
192	abancayense	PER		
193	lignicaule	PER	-13.4333	-71.85
194	buesii	PER	-13.5333	-71.9333
195	species	PER		
196	tuberosum andigena	PER		
197	pampasense	PER		
198	chancayense	PER		
199	immite	PER	-8.1166	-79.0333
200	blanco-galdosii	PER	-7.1833	-78.2166
201	chiquidenum	PER	-6.9666	-78.5666
202	sandemanii	PER	-16.2833	-71.5166
203	velardei	PER	-13.5	-72.8
204	huancabambense	PER		
205	amabile	PER		
206	amayanum	PER		
207	ancophilum	PER		
208	augustii	PER		
209	aymaraesense	PER	-14.0666	-73.25
210	chillonanum (=tenellum)	PER	-13.8333	-72.25
211	huarochiriense	PER		

212	hypacrarthrum	PER		
213	irosinum	PER		
214	lycopersicoides	PER		
215	santolallae	PER	-13.1833	-72.55
216	scabrifolium	PER	-9.4166	-76.7833
217	orophilum	PER	-8.9166	-77.8333
218	chacoense	PRY	-25.3833	-57.15
219	commersonii malmeanum	PRY	-23	-60
220	demissum	URY		
221	commersonii malmeanum	URY		
222	jamesii	USA		
223	fendleri	USA		
224	subpanduratum	VEN		
225	paramoense	VEN	8.75	-70.8666
226	phureja	VEN		
227	jamesii	MEX		
228	brachycarpum x sparsipilum	BOL	-17.4833	-65.2666
229	sparsipilum x leptophyes	BOL	-16.55	-68.1
230	sparsipilum x sucrense	BOL		
231	sucrense x oplocense	BOL		
232	tarajense x arnezii	BOL		
233	tarajense x microdontum	BOL		
234	tuberosum andigena x curtilobum	BOL	-17.65	-67.5166
235	tuberosum andigena x sucrense	BOL		
236	violaceimarmoratum X yungasense	BOL	-16.3166	-67.85
237	maglia x microdontum	ARG		
238	rechei x microdontum	ARG		
239	raphanifolium x sparsipilum	PER		

Appendix III. Commercial cultivar list

Table 2 commercial cultivar list, 0 indicates the absence of transposon, 1 for presence

Nr.	Cultivar	Phenotype	Transposon
1	Shamrock	Very late	0
2	Gladstone	Intermediate	1
3	Estima	Early	1
4	Casteline	Early	1
5	Civa	Early	1
6	Karnico	Very late	1
7	Herald	Early	1
8	Daisy	Intermediate	0
9	Katahdin	Intermediate	1
10	ATLANTIC	Intermediate	0
11	OSIRA	Early	1
12	C	intermediate	1
13	E		1
14	3027	late	0
15	3130	early	1
16	SH	early	1
17	RH		1