A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato

Jan G. A. M. L. Uitdewilligen[1,2], Anne-Marie A. Wolters[1,2], Bjorn B. D'hoop[1], Theo J. A. Borm[1,2], Richard G. F. Visser[1,2,3], Herman J. van Eck[1,2,3]*

1 Laboratory of Plant Breeding, Wageningen University, Wageningen, The Netherlands, 2 The Graduate School for Experimental Plant Sciences, Wageningen, The Netherlands, 3 Centre for BioSystems Genomics, Wageningen, The Netherlands

Abstract

Assessment of genomic DNA sequence variation and genotype calling in autotetraploids implies the ability to distinguish among five possible alternative allele copy number states. This study demonstrates the accuracy of genotyping-by-sequencing (GBS) of a large collection of autotetraploid potato cultivars using next-generation sequencing. It is still costly to reach sufficient read depths on a genome wide scale, across the cultivated gene pool. Therefore, we enriched cultivar-specific DNA sequencing libraries using an in-solution hybridisation method (SureSelect). This complexity reduction allowed to confine our study to 807 target genes distributed across the genomes of 83 tetraploid cultivars and one reference (DM 1–3 511). Indexed sequencing libraries were paired-end sequenced in 7 pools of 12 samples using Illumina HiSeq2000. After filtering and processing the raw sequence data, 12.4 Gigabases of high-quality sequence data was obtained, which mapped to 2.1 Mb of the potato reference genome, with a median average read depth of $63\times$ per cultivar. We detected 129,156 sequence variants and genotyped the allele copy number of each variant for every cultivar. In this cultivar panel a variant density of 1 SNP/24 bp in exons and 1 SNP/15 bp in introns was obtained. The average minor allele frequency (MAF) of a variant was 0.14. Potato germplasm displayed a large number of relatively rare variants and/or haplotypes, with 61% of the variants having a MAF below 0.05. A very high average nucleotide diversity ($\pi = 0.0107$) was observed. Nucleotide diversity varied among potato chromosomes. Several genes under selection were identified. Genotyping-by-sequencing results, with allele copy number estimates, were validated with a KASP genotyping assay. This validation showed that read depths of ~60–$80\times$ can be used as a lower boundary for reliable assessment of allele copy number of sequence variants in autotetraploids. Genotypic data were associated with traits, and alleles strongly influencing maturity and flesh colour were identified.

Introduction

DNA variants such as single nucleotide polymorphisms (SNPs), multinucleotide polymorphisms (MNPs), and insertions and deletions (indels) are differences at the nucleotide sequence level among individuals or alleles and represent the basic units of genetic diversity. They can be assayed and exploited as high-throughput molecular markers and are widely used for marker assisted and genomic selection, association analysis and mapping of quantitative trait loci (QTL), and haplotype and pedigree analysis in several crops and model plants [1–7]. Genotyping of DNA sequence variants in highly heterozygous polyploid species, such as potato (*Solanum tuberosum*), is more challenging than in diploid species, because a given gene may be represented by up to four different alleles per locus per genotype. Therefore, genetic analysis in tetraploid potato requires a genotyping system that can distinguish among alleles and quantify the allele copy number. In tetraploid species, the possible allele copy number (zygosity) categories include nulliplex (0), simplex (1), duplex (2), triplex (3), and quadruplex (4).

Several studies have shown that direct resequencing of amplicons by Sanger sequencing is sufficiently quantitative to allow the simultaneous discovery and genotyping of sequence variants in polyploids [8,9]. Amplicon sequencing is a reasonable method to analyse a single or small number of target gene(s), but beyond that scale, more effort is required to design unique primers and optimize PCR parameters to ensure equal amplification of all alleles. Furthermore, in species like potato that exhibit high nucleotide diversity, the many indels usually create uninterpretable sequence reads that further reduces the throughput of Sanger-based amplicon resequencing. With the introduction of the next-generation massively-parallel sequencing (MPS) technologies, several novel approaches have been developed to discover, sequence and genotype not tens, but hundreds to thousands of genes simultaneously [2,10–14]. Similar to Sanger sequencing, MPS can be used directly to genotype sequence variants. Using the

former technology, genotype data are retrieved from chromato-gram peak intensities at variant positions. For MPS, accumulation of sequence reads provides allele copy number estimates for each variant. Accuracy of this allele copy number estimate is dependent on sequencing depth. Genotyping-by-sequencing (GBS) of tetra-ploids will require higher read depth compared to diploids. In humans, for example, a sequence read depth of 30–35× is considered appropriate for accurate genotyping [15]. In order to identify each allele in tetraploids a simulation model suggested a read depth of at least 15× [13]. However, accurate genotyping ($p = 0.95$) of an autotetraploid organism implies also to discrim-inate digenic-simplex (aaab) as opposed to digenic-duplex (aabb) zygosity. Based on a binominal distribution this will require a sequence depth of at least 48×.

To achieve an increase in read depth, the portion of the genome that is sequenced can be reduced by applying, for example, RNAseq [2,16], restriction enzyme based complexity reduction [17,18], or sequence capture methods such as SureSelect, Nimblegen, and Raindance [19,20,21]. RNAseq is less suitable for GBS, since alleles may vary in transcription level across genes, tissues and stages, and thus may generate inaccurate genotyping data. Methods based on restriction enzyme treatments, on the other hand, are more likely to target non-coding parts of the genome, producing less useful data for functional gene analysis. Furthermore, restriction-based methods cannot target specific regions of interest, and nucleotide variants in the restriction site may interfere with digestion and cause null alleles. Sequence capture methods like SureSelect use oligonucleotide baits designed to bind to regions of interest, which can be specifically selected and enriched before sequencing [22]. For example, whole exomes [23] or regions associated with particular traits [19,24] can be targeted. Sequence capture approaches require *a priori* availability of sequence data to design DNA capture probes. The recently sequenced genome of *Solanum tuberosum* group Phureja is appro-priate for this purpose in potato [25].

Here we describe DNA resequencing results obtained from dozens of autotetraploid potato cultivars and one monoploid accession after genome complexity reduction using hybridisation-based in-solution enrichment. We subsequently used this data to identify sequence variants within and across cultivars and to call the genotypes of resequenced individuals. The accuracy of genotyping-by-sequencing was validated using a SNP genotyping assay. The resulting marker dataset is useful for describing allele frequencies, nucleotide diversity, and population structure in potato, and for validating QTLs via association analysis. Our approach is an efficient means of producing data for the design of both high and low-density SNP genotyping assays applicable to a wide range of potato cultivars, and the resulting tools can be used to address questions in population genetics and marker-trait association research.

Materials and Methods

Design of the Capture Library

A custom-designed SureSelect capture library containing 57,054 RNA oligonucleotide baits of 120-bp length each was developed. Genes targeted for enrichment were selected from the PotCyc metabolic pathway database [26], the Potato Maps and More database (PoMaMo) [27], an in-house QualitySNP marker database [28,29], and a subset of single-copy genes homologous to the Conserved Ortholog Set II (COSII) [30]. Functional genes widely used as genetic markers for carbohydrate metabolism [31] and secondary metabolism [32–34], and a number of additional putative candidate genes for potato quality traits, were also

included as targets. In addition to these functional genes, a number of intergenic sequences corresponding to AFLP markers, and a number of chloroplast and mitochondrial genes were included (Table 1). To avoid intron spanning baits initial cDNA targets were aligned to the *S. tuberosum* Group Phureja DM whole genome assembly v.1 and baits were designed based on the genomic reference sequence with highest homology. Genome annotation for the DM sequence was not yet available at the time the baits were designed, so genomic coding regions and intron/exon boundaries were estimated using GeneSeqer [35]. Other reference sequences used for bait design included the chloroplast genome of *Solanum tuberosum* cv. Desiree (NC008096), *Solanum tuberosum* mitochondrial sequences (S66866, X74826, X80386, X83206, X93576) and sequenced BAC clones of genotype RH89-039-16 [25].

During bait design, we aimed to optimize the in-solution hybridisation enrichment by avoiding targets exhibiting repeat elements and paralogous sequences in the reference genome, which can affect target-bait hybridization during enrichment and add difficulty to read mapping after resequencing. Stretches of repetitive sequences within the target regions were excluded from bait design using RepeatMasker (http://www.repeatmasker.org). BLAST homology search against the *S. tuberosum* Group Phureja DM whole genome assembly v.1 [25] was conducted to avoid the use of targets with paralogous and/or duplicated sequences. Except for a small number of target gene families of specific interest (e.g. polyphenol oxidases), target sequences having a secondary hit with E-value $<10^{-10}$ were excluded from bait design. Regions consisting mainly of introns (>200–1000 bp) were avoided as targets. For each gene target sequence, we used an average of 3–4 continuous regions (contigs) for bait design, each region having an average length of 475 bp and, where applicable, including both exons and introns. OligoTiler [36] was used to tile the reference strand of each target region, with baits (of 120 bp in length) starting approximately every 20 bp. This produced a 6× bait tiling coverage and resulted in 57,054 unique baits for the SureSelect capture library (ELID 0274451). In total, the library targets 2,945 contigs (1.44 Mb, GC-content 39%; Table 1). Complete lists of sequencing targets and oligonucleotide bait sequences are available in XLS-file S1 and FASTA-files S1&S2.

Plant Collection

A subset of 83 tetraploid potato cultivars (Table A in Supporting Information S1) was selected from a larger collection [37] using marker-based genetic distance estimates. The panel represents the global gene pool of commercial potato, both heirloom and contemporary, with emphasis on cultivars with high value in breeding and/or use. We also included a monoploid potato clone DM 1–3 511, derived by anther culture of the heterozygous diploid *S. tuberosum* group Phureja, clone BARD 1–3 of accession PI225669 [38]. The DM 1–3 511 clone is highly related to the recently sequenced clone DM 1–3 516R44 (CIP801092) [25], a doubled monoploid derived from the same BARD 1–3 clone [38].

Extraction and Fragmentation of DNA

DNA was extracted from leaves ground in liquid nitrogen using KingFisher Genomic DNA Purification Kit (Thermo Scientific) and the KingFisher Ml magnetic nucleic acid extraction system (Thermo Scientific) according to the manufacturer's procedures. DNA concentrations were quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) and diluted to 35 ng/µl. Of each DNA sample, 3.5 µg was fragmented by Adaptive Focused Acoustics on a Covaris S2 instrument (Covaris, Inc.), using a 10% duty cycle at intensity 4 for 120 seconds with 200

Table 1. Targets for SureSelect bait library development.

	Source	Genes	Target Contigs	Target Sequences	# of Baits (1 bait/20 bp)	Proportion of baits in library
Nuclear genome	COSII database	248	853	439.6 kb	17,618	30.9%
	PotCyc database	149	467	142.5 kb	4,940	8.7%
	PoMaMo & Candidate genes	116	523	283.1 kb	11,481	20.1%
	In-house database	249	789	424.7 kb	17,174	30.1%
	AFLP sequences	45	202	87.8 kb	3,400	6.0%
Sub-total		807	2,834	1377.7 kb	54,613	95.8%
Organelle genome	Chloroplast	64	99	55.9 kb	2,272	4.0%
	Mitochondrial	4	12	4.7 kb	169	0.3%
	Total	**875**	**2,945**	**1,438.3 kb**	**57,054**	**100%**

For most targets, (cDNA) sequences were aligned to the potato reference genome, and the genomic sequence of each locus was used to design the baits.
doi:10.1371/journal.pone.0062355.t001

cycles per burst. Fragmentation of DNA to an average size of about 300 bp was verified using Bioanalyzer High Sensitivity DNA chips (Agilent).

Indexed Sequence Library Preparation

Custom-ordered, HPLC-purified, indexed adapters consisted of complementary Illumina adapters PE1 and PE2 [39], with PE1 extended by a 4-bp index sequence and an extra terminal T to facilitate sticky-end ligation. The reverse (PE2) adapter was extended by the reverse complement of the PE1 index (Table B in Supporting Information S1). The twelve indices had a balanced base composition and a minimal edit distance (i.e. the number of mutations required to change one index to another) of 2 bp to detect sequencing errors in the index region. To pair the adapter strands, mixtures of the forward and reverse strands (each 50 μM in TE) were incubated at 95°C for 2.5 minutes, followed by a cool down (-1°C/30 sec.) to 25°C and subsequently diluted to a working concentration of 12.5 μM. DNA sequencing libraries were prepared using the NEBNext DNA Sample Master Mix Set 1 (New England Biolabs). Purifications were carried out between end-repair, dA-tailing, and adapter ligation steps using AMPure XP beads (Agencourt Bioscience). To index the 84 potato DNA extracts, each of the twelve unique adapters was ligated to seven different DNA samples by mixing adapter and DNA in a molar ratio of approximately 20:1. After the initial ligation of 15 minutes at 20°C, samples were held at 4°C overnight and purified using AMPure XP beads. Adapter ligation was verified by fragment size analysis using Bioanalyzer High Sensitivity DNA chips.

In-solution Hybridization and Target Enrichment

The 84 indexed paired-end sequencing libraries were hybridized to the SureSelect capture library according to the manufacturer's instructions (Agilent SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library Protocol, Version 1.0 May 2010), with minor modifications. Size selection on gel was omitted, allowing us to skip the standard pre-amplification step prior to in-solution enrichment and reduce the number of clonal reads in the later generated sequences. The use of index-specific blocking oligos to reduce non-specific pull-down due to adapter-adapter hybridisation can also be avoided when pre-amplification is excluded, as non-amplified Y-adapters do not concatenate. We therefor excluded from the hybridization mix Block #1 (Human C0t-1 fraction) [40], which is irrelevant for plants, and excluded

Block #3 (PE adapter block) [40]. We used half of the specified hybridization volumes. Of each indexed DNA sample, 50–400 ng was mixed with 1.25 μg salmon sperm DNA (Block #2), denatured, and hybridized to 100 ng SureSelect biotinylated RNA baits developed from the capture library. The hybridization mix was held at 65°C for 24 h for hybridization, then added to 250 ng (25 μl) T1 streptavidin Dynabeads (Invitrogen), and pulled down. Bait-selected DNA was purified using AMPure beads, and amplified for 14 cycles (Tm 60°C) using 1 U Herculase II Fusion proofreading DNA polymerase (Agilent) and 25 μM each of custom primers PE1.0 and PE2.0 (Table C in Supporting Information S1). The alternative DNA polymerase was chosen after the polymerase included in the NEBNext DNA Sample Master Mix Set 1 was seen to cause a negative shift in average fragment size.

Amplified, enriched DNA libraries were purified using AMPure XP beads and evaluated by agarose gel electrophoresis, Nano-Drop, and Bioanalyzer High Sensitivity DNA chips. Library subsamples were also quantified by qPCR using a two-step amplification protocol [95°C activation for 2 min, 40× (95°C 10 sec., 60°C 30 sec.)] and primers qPCR_1.1 and qPCR_2.1 (Table D in Supporting Information S1).

Hi-Seq2000 Sequencing and Data Preprocessing

Seven distinct resequencing pools were made by combining equimolar amounts of twelve unique, differently-indexed, enriched DNA library samples (Table A in Supporting Information S1). Pooled libraries were sequenced on a Hi-Seq2000 (Illumina) lane using 100-base paired-end sequencing at the Genome Analysis Facility of the University Medical Center Groningen (UMCG). Four sequenced pools that generated low cluster numbers and/or partly failed reverse read sequences were repeated. Initial quality checks (average read quality per cycle, average read quality, base call % per position) were performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequences in each pool were categorized to their biological sources according to the 4-bp index using NovoBarCode (Novocraft) with the average edit distance set to 3. The first 5 bp of each de-multiplexed sequence read (i.e., the 4-bp index plus an extra T used for ligation) were removed.

Sequence read-pairs containing ≥63 nucleotide sequences with 100% match to chloroplast genome sequences of *Solanum tuberosum* cv. Desiree (NC008096) were extracted from the de-multiplexed

sequences and kept for separate analysis. Remaining sequence reads were aligned using the Burrows-Wheeler Alignment tool (BWA) [41] with default alignment parameters, except for the maximum edit distance, which was relaxed to seven due to the expected high sequence divergence between potato alleles. Sequences were aligned to the annotated superscaffolds (DMGv.3.4, comprising 705.5 Mb of sequence in 1,419 super-scaffolds and 43 Mb of non-ACGT bases) of the *S. tuberosum* Group Phureja DM whole genome assembly v.3 reference genome [25]. The 120-bp SureSelect bait sequences of genomic origin were correspondingly mapped to the annotated superscaffolds, and baits with a mapping quality (MQ) $\geq$37 were used to define the genomic target regions.

Alignment data was processed with SAMtools and Picard [42] to mark duplicate reads and estimate the average insert size of the paired-end reads. The Genome Analysis Toolkit (GATK) [43] was used for indel realignment, base-score recalibration, and extraction of read depth information. Read depth and coverage data were processed with custom Perl scripts and BEDTools [44]. The raw sequencing reads and mapping assembly are available on our server (http://datarelease.plantbreeding.nl/).

Sequence Variant Detection and Genotype Calling

For covered regions (see Results for definition), sequence variants were identified simultaneously among the aligned reads from all 83 tetraploids and the single monoploid using the FreeBayes polymorphism discovery algorithm [45]. Sequence variants included binary SNPs, MNPs, and small indels, as well as allelic series of tri-SNPs and tetra-SNPs, multi-allelic MNPs, and indels with a variable number of (repetitive) nucleotides. Reads marked as duplicate, with more than seven base mismatches, more than three separate gaps, or with MQ<30 were excluded for variant calling. The expected mutation rate or pairwise nucleotide diversity was set to 0.01. In order to include an alternate allele as a variant, supporting bases required a minimum base quality (BQ) of at least 13, and at least one supporting alignment was required to have BQ$\geq$20. Furthermore, the alternate allele had to be observed in at least 5 reads and represent at least 12.5% of the observations of reads at that locus within a single potato sample. Mapping quality of alleles was included when calculating posterior probabilities, and variants were only called for sites that had a probability of polymorphism greater than 0.95. Sequence variants adjacent to indels, which were likely due to local misalignment, were filtered using GATK [43].

For genotype calling, zygosities of all sequence variants were resolved by allele-specific read depths for all non-duplicated reads with MQ$\geq$13 using FreeBayes. This resulted in nulliplex (*aaaa*), simplex (*aaab*), duplex (*aabb*), triplex (*abbb*) and quadruplex (*bbbb*) genotype calls relative to the reference sequence for the tetraploid potato samples, while variants in the monoploid sample were genotyped as either absent or present. Hence a variant is called not only if a genotype differs from the reference, but also if a genotype differs from any other genotype. Only variants previously identified by variant calling were used for genotyping, and samples required a minimum read depth of 15$\times$ at the variant position to yield a genotype call. In addition to this 15x coverage threshold we applied a threshold GQ$\geq$26. Genotyping Quality (GQ) was calculated by a likelihood estimate to determine the probability that a genotype call was different from the true genotype using FreeBayes. It was encoded as a phred quality score ($-10\times\log_{10}(p)$) and included read depth at a variant position as a parameter.

Diversity Analysis and Prediction of Functional Consequences of Allelic Variants

To calculate the nucleotide diversity, we first calculated the gene diversity (heterozygosity) of each binary variant; $GD = 1 - \Sigma P_i^2$, where P_i is the frequency of the *i*th allele and GD is summed across *n* alleles. Nucleotide diversity (π) was calculated by averaging GD over all nucleotide sites – or all coding and/or non-coding nucleotides sites – on a contig or gene.

The functional effect (codon mutations, splice site mutations, frame shift mutations) of sequence variants was predicted using snpEff [46] and gene annotation DMGv3.4 of the DM genome.

In order to identify population structure, principal components analysis (PCA) was performed on genotype scores using the FactoMineR library of R [47]. Only genotype scores of variants called in all 84 cultivars were included. Since all genotype scores were measured in units of allele copy number, the data were not scaled. K-mean clustering of the first three principal components was used to identify genotype clusters.

Association Analysis

Adjusted phenotype means for plant maturity and tuber flesh colour in each of the 83 tetraploid cultivars, measured previously over a period of five years [37,48], were used as trait values for conducting association analysis. Additive and dominant genotype models were each tested both with and without correction for population structure. The genotype clusters identified by PCA analysis were used as the adjustment factors for population structure. For dominant association models, linear regression models were used as implemented in PLINK [49], with tetraploid data recoded into diploid homozygous/heterozygous scores, using binary allelic variants only. Adjustment of p-values to correct for multiple testing effects was carried out using step-up FDR control as implemented in PLINK. For the additive tetraploid genotype models, we applied linear regression models implemented in Genstat. For each trait and each marker the model fitted was: response = allele copy number (+ structure)+error.

Validation of Genotype Calling

The accuracy of GBS genotype calls was validated using the Kbioscience Allele-specific Polymorphism Assay (KASP) SNP genotyping platform. Binary SNPs identified in the sequence data that exhibited a minor allele frequency between 0.15 and 0.35 were selected as candidate for assay design. To assure independence among all SNPs in the KASP assay (i.e., use of a unique SNP from each haplotype block), SNP data from GBS were clustered using hierarchical cluster-analysis, and a single SNP from each cluster having a correlation coefficient of $r^2\geq0.16$ was used for the KASP assay, yielding 768 SNPs in the final design. We KASP-assayed DNA from 65 potato cultivars included in GBS, with two of these cultivars measured in duplicate to assess KASP genotyping consistency. A number of additional diploid potato clones were assayed ($\sim$96) and used to examine the signal ratio of the nulliplex, duplex and quadruplex genotype signals. The software package fitTETRA [50] was used for full tetraploid zygosity genotype calling. In total, 270 of the 768 KASP assayed SNPs were selected for validation of the GBS calls. This selection was based on (1) sharp clustering of the signal ratios of discrete genotype classes, (2) clustering of signal ratios of heterozygous diploids with duplex tetraploids, and (3) clustering of signal ratios of homozygous diploids with nulliplex or quadruplex tetraploids.

Results

Sequencing and Mapping of Enriched Libraries of 84 Potato Genomes

We designed an in-solution hybridisation capture library targeting primarily introns and exons of nuclear coding genes, but also including intergenic, chloroplast, and mitochondrial sequences. Baits targeting nuclear genome sequences in the enrichment library covered approximately 1.3 Mb of the potato DM reference genome sequence, scattered across all 12 chromosomes (Figure 1 & BED-file S1).

Genomic DNA sequencing libraries from 83 diverse tetraploid potato cultivars and progenitor lines, and a single monoploid potato clone, were indexed by cultivar with 12 distinct 4-bp index sequences and individually enriched using the capture library. Indexed and enriched samples were multiplexed in pools of 12 unique indices and paired-end sequenced. In total, 592,100,112 read-pairs were obtained, representing approximately 100 Gigabases of sequencing data. The cultivar-specific sequence index could be identified in 96% of the read-pairs (Figure A in Supporting Information S1). For 19% of the read-pairs, cultivar identification was based on the index from only one read of the pair. In most cases, this resulted from technical failure of the reverse-read sequencing.

Within each multiplexed pool, the cultivar-specific read counts rarely reached a twofold difference between cultivars, indicating approximately equivalent amounts of DNA were added to a pool from each cultivar present. Initial inspection of the cultivar-tagged sequences showed that chloroplast-derived sequences were highly abundant (60% of the read-pairs). These sequences were filtered using broad homology criteria and saved for a later analysis of its own. The remaining 227,263,706 read-pairs were aligned to the DM reference genome, with 80% mapping. This resulted in 23.9 Gigabases of high quality (MQ≥13) genome-aligned sequence data. Among mapped reads, 8% were marked as duplicate reads aligning with identical start and end positions on the reference genome.
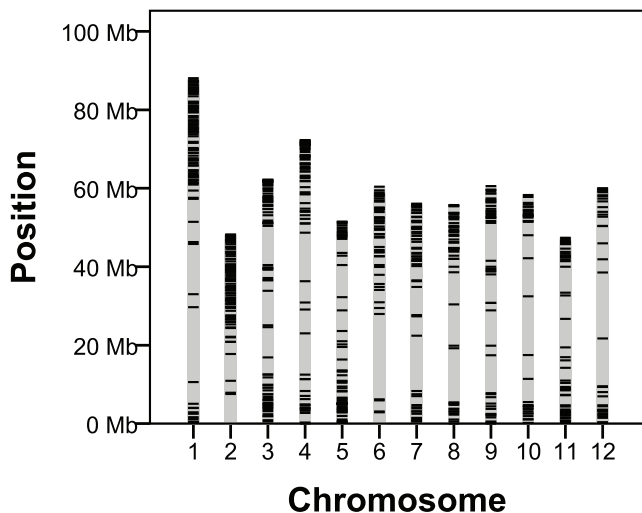
Genome and Target Coverage

As a consequence of the enrichment method sequences aligned not only to target regions but also to flanking and off-target regions. To define the "accessible" part of the genome, i.e., the part with adequate coverage to allow identification of sequence variants, we used the following criteria: (1) at least five cultivars had to be sequenced at the base position, with a threshold criterion of at least 20 component reads with MQ>30 per cultivar; (2) in view of the fragment sizes of mapped reads (261±94 bp, mean±SD), regions defined as covered were at least 261 bp long; (3) regions were discarded when more than 5% of reads within the region aligned to multiple locations with equal probability (MQ = 0); and (4) finally, regions with >80% homology to chloroplast and mitochondrial sequences were removed. This resulted in 2,445 sequenced contigs with a total length of 2.1 Mb (XLS-file S2 and BED-file S2). A total of 12.5 Gigabases of high quality sequence data aligned to these contigs, providing an average read depth of 5,871× when all cultivars were considered together. Median depth per cultivar within this accessible genome was 63×, ranging from 15× to 177× (Table 2 and Table A in Supporting Information S1).

Almost all genomic regions and genes targeted by the enrichment library fell within the accessible regions, with 97% of target sequences and 793 of 807 nucleartarget genes covered. In total, 10.7 Gigabases of high quality sequence data aligned to target regions with a median depth per cultivar of 88×, ranging from 20× to 240×. Accessible flanking regions had an average length of 150 bp and comprised 554 kb of additional sequence. Regions flanking target sequence but interrupted by poor coverage for a small number of nucleotides, and more remote off-target regions, accounted for 288 kb (13.5%) of the accessible genome.

DNA Sequence Variants

A total of 129,156 putative sequence variants (SNPs, MNPs and indels) were identified in the accessible genome (Table 3, CSV-file S1, and VCF-file S1). The density of substitution variants (SNPs and MNPs) was 1.6 times higher in non-coding regions than in coding regions, and the indel density was 12 times higher in non-coding regions. The transition/transversion ratio (T_s/T_v), calcu-

Figure 1. Mapping of resequencing targets onto the potato genome. Physical position of the approximately 800 target genes scattered across the 12 potato chromosomes. Physical positions are ordered according to the latest pseudomolecule order (v.3 beta). The distribution of the ~800 genes reflect the physical proportions of gene rich chromosome arms and gene poor heterochromatic regions.
doi:10.1371/journal.pone.0062355.g001

Table 2. Summary of target enrichment sequence coverage for 84 potato cultivars.

"Accessible" genome parameter	Value
Number of covered regions (contigs)	2,445
Sequence length	2,136,143 bp
Coding sequence (DMGv.3.4)	655,930 bp
Sequencing depth per cultivar at contigs[b,c]	70±36× (Median 63×)
Sequencing depth per cultivar at target sequence[b,c]	95±48× (Median 88×)
Genes covered	977
Target genes covered	793 out of 807
Target sequence covered	1,294,097 bp (97%)[a]
Target+directly flanking sequences	1,848,192 bp
Off-target sequence	287,951 bp (13%)

[a]Percentage of target bait sequence mapped to the DM genome.
[b]Only sequenced nucleotides that aligned with high quality (MQ≥13).
[c]Mean and standard deviation.
doi:10.1371/journal.pone.0062355.t002

Table 3. Overview of DNA variants observed across 84 cultivars in the accessible potato genome.

	Number of sequence variants called		
Variant type	Accessible genome (2136 kb)	Non-coding (1480 kb)	Coding (656 kb)
Di-nucleotide SNPs	105,812	84,454	25,358
Tri-SNPs	5,304	4,097	1,207
Tetra-SNPs	96	66	30
Indels	13,094	12,641	453
MNPs	4,850	4,084	766
Total	129,156	101,342	27,814
Average variant density	1/16.5 bp	1/14.6 bp	1/23.6 bp

doi:10.1371/journal.pone.0062355.t003

lated only for biallelic SNPs, was 1.55, and the ratio of non-synonymous to synonymous SNPs (pN/pS) was 0.64.

Across all cultivars, an average variant density of 1/24 bp in coding regions and 1/15 bp in non-coding regions was observed. Within a single tetraploid cultivar, on average 52,233 sequence variants (1/42.5 bp) were observed. On average each cultivar had 116 cultivar-unique variants, ranging from 0 to 2,688. Cultivars like Vitelotte Noir, the only cultivar with purple flesh colour in our samples, and those with wild species introgression segments contained a relatively high number of cultivar-unique variants (e.g., up to 2.0% of all variants within cv. Vitelotte Noir). Cultivars without unique variants either had ancestors that were widely used in breeding of novel potato cultivars or had themselves been used for this purpose (e.g., cv. Agria and cv. Katahdin respectively).

To evaluate the increase in variant density per additional sequenced cultivar, we permutated the order of the cultivars a thousand times and calculated the variant frequency at each incremental step (Figure 2). More than half of all variants were detected by selecting three random cultivars. When 16 random cultivars were selected, 84% of all variants were detected, and the number of novel variants that could be identified by sequencing an additional cultivar dropped below 1% of the variants already discovered. To detect 95% of all the identified variants, 46 random cultivars were required.

Genotype Calling and Allele Frequencies

To be assigned a valid genotype call for a specific variant position, a cultivar required a minimal read depth of 15× at that position. Using this criterion, 86.6% of all possible genotype calls were valid (i.e., of the matrix of 84 cultivars by 129,156 sequence variants, 13.4% had insufficient read depth to make a call), equivalent to a per-locus average of 73 out of 84 cultivars receiving a genotype call. For 42,625 sequence variants (33% of all identified sequence variants), all 84 cultivars were genotyped, and more than 90% of all sequence variants were genotyped in at least half of the cultivars. Population-level allele frequencies were calculated using all valid (≥15×) calls. The distribution of minor allele frequencies (MAF) is shown in Figure 3. The average MAF was 0.14; 17.4% of all sequence variants had MAF<0.01, and 60.9% had MAF>0.05. For 13,458 sequence variants, 10.4% of all identified, the allele in the DM reference genome differed from the major allele in the population.

Nucleotide Diversity

We next calculated the nucleotide diversity (π) for each of the 2,445 sequenced contigs, for each gene, and for each chromosome. Mean π of the covered genome was 1.07×10^{-4} (Table 4). As

expected due to functional constraints on evolutionarily tolerable mutations, π was lower for coding regions than for non-coding regions (Table 4 and Figure 4). The physical position of contigs on the potato pseudomolecules was used to plot π over the twelve potato chromosomes (Figure B in Supporting Information S1). Mean nucleotide diversity for chromosomes 5 and 11 was significantly higher relative to other chromosomes (Figure 5). In contrast, the mean π for chromosome 10 was significantly lower than for other chromosomes. Individual genes with low nucleotide diversity were observed on all chromosomes (XLS-file S3).

Population Structure

Population structure was analysed using the sequence variants that were genotyped in all 84 cultivars. The first three components of a principal component analysis described respectively 6.4%, 4.5% and 3.8% of the variance. In the centre of the PCA plot, a cluster of cultivars of diverse origin was located (Figure 6). Three groups diverging from this set of cultivars were noted, which consist of (a) heirloom British cultivars, (b) a number of typical frying cultivars from continental Europe, and (c) progenitors of potato cyst nematode (PCN) resistance and cultivars bred for the
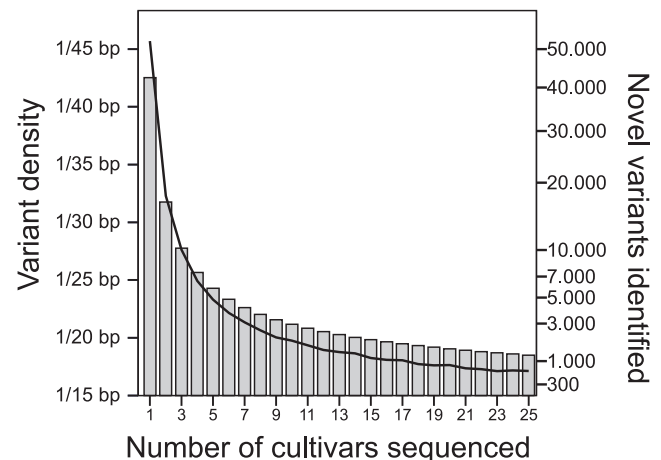

Figure 2. Sequence variant density as the number of randomly-added cultivars increases. The bars show variant density (primary Y-axis), and the black line shows the number of newly-identified variants (secondary Y-axis) as a function of the number of sequenced cultivars. Data is not shown after the 25th cultivar, but continues to drop to a variant density of 1/16.4 bp and an average of 116 novel variants at the 84th cultivar.
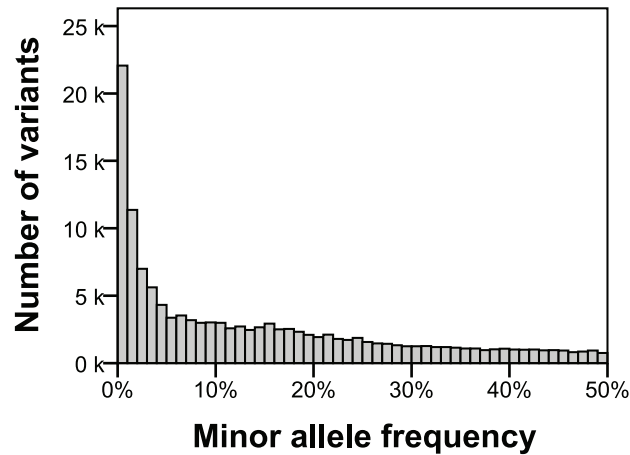
doi:10.1371/journal.pone.0062355.g002

Figure 3. Distribution of minor allele frequencies (MAF) of all 129,156 genotyped sequence variants.
doi:10.1371/journal.pone.0062355.g003

starch industry with resistance to PCN. The heirloom cultivar group is closest to the more distant monoploid *S. tuberosum* Group Phureja clone, which fell into a cluster of its own.

Association Analysis

The genotype dataset was used for a genome wide association study (GWAS) to validate the sufficiency of this data for identifying known QTLs for plant maturity and tuber flesh colour. Association analysis was performed using both additive and dominant genotype models with separate tests with and without correction for population structure. To a large extend, the results of the dominant and additive models overlapped. The results of the dominant allele models are shown in Figure 7. Even with the small population size of 83 phenotyped cultivars, the well-known QTL for early plant maturity on potato chromosome 5 was clearly detected $(-\log_{10}(p) = 6.0)$ and explained 44% of the observed phenotypic variance. Using the current data, this QTL mapped to a region of approximately 371 kb within superscaffold PGSC0003DMB000000192 containing 28 strongly associated variants $(-\log_{10}(p) \geq 5)$. For tuber flesh colour, a major QTL was observed on chromosome 3, mapping to a region of approximately 683 kb containing 27 strongly associated variants $(-\log_{10}(p) \geq 7)$. These sequence variants were located within and near *CHY2* (*β-carotene hydroxylase*), a well-known gene influencing flesh colour via carotenoid synthesis, and explained 61% of the phenotypic variance. Two additional minor-effect QTLs for flesh colour were found on chromosomes 4 and 12. When corrected for population structure, the QTL on chromosome 12 was not significant. The flesh colour QTL on chromosome 4 $(-\log_{10}(p) = 3.7)$ explained 9% of additional phenotypic variance beyond that explained by the major QTL.

Validation of Genotyping-by-sequencing Data

A subset of 270 binary SNPs was selected to validate genotype calls made by GBS using KASP genotyping assays. For one of the two cultivars duplicated in the KASP assay, all genotype calls were identical between replicates, and for the other cultivar, only two genotype calls varied between replicates. The expected KASP genotyping error rate of this subset of SNPs is thus very low (0.4%). For concordance analysis, expected genotype calls were obtained from these 270 KASP assay SNPs and observed genotype calls were obtained by the GBS results. Scored as either
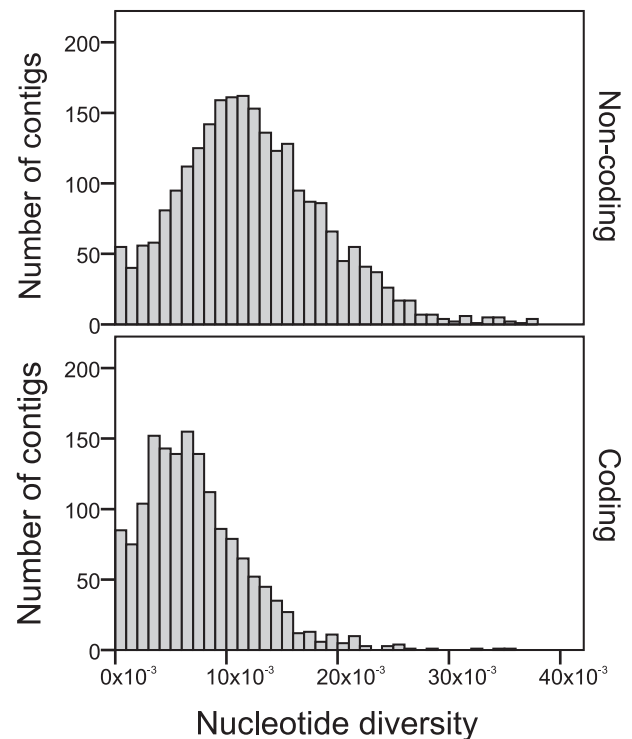

Figure 4. Distribution of nucleotide diversity across sequenced contigs for coding and non-coding regions.
doi:10.1371/journal.pone.0062355.g004

homozygous (i.e., nulliplex or quadruplex) or heterozygous (i.e., simplex, duplex, or triplex), 97.9% of the homozygous and 99.9% of the heterozygous GBS calls were concordant with the results of KASP genotyping. When heterozygous GBS calls were split into simplex, duplex and triplex categories, overall concordance with KASP genotyping dropped to 94.4%. For duplex calls, which were most difficult to classify, only 90.3% of the calls were concordant. We therefore applied a genotype quality (GQ) filter to validate GBS/KASP concordance in the five discrete zygosity classes. A
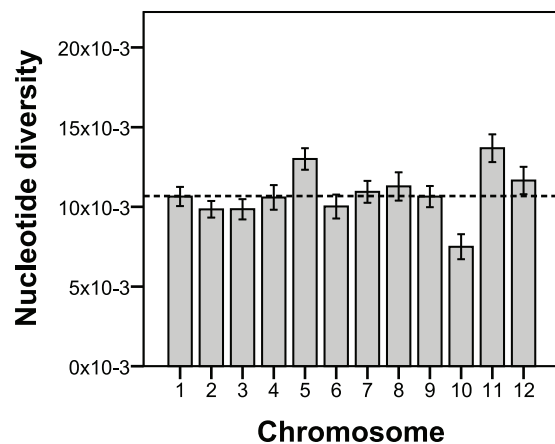

Figure 5. Mean nucleotide diversity across sequenced contigs per chromosome. Dashed line represents genome wide nucleotide diversity (both coding and noncoding sequences), and error bars represent 95% confidence intervals.
doi:10.1371/journal.pone.0062355.g005

Table 4. Nucleotide diversity in potato (mean value and standard deviation across contigs).

Nucleotide diversity (×10-3)	10.7±5.4
Nucleotide diversity (×10-3), coding sequence	7.3±4.8
Nucleotide diversity (×10-3), non-coding sequence	12.4±6.7

doi:10.1371/journal.pone.0062355.t004

threshold of GQ≥26 was required to reduce the number of discordant duplex calls below 5%. Applying the GQ26 filter to the complete set of 129,156 sequence variants yielded 74.8% of variants with an assigned genotype call, with an average of 63 out of 84 cultivars genotyped per variant position. This is equivalent to 25.2% of unassigned genotype calls, yielding approximately two-fold more unassigned genotype calls compared to the set subject only to the 15× read-depth threshold. Median read depth of all GBS genotype calls meeting the GQ26 threshold was 61×, and for duplex calls this was 81×. Overall concordance between the GQ26 filtered set and KASP genotyping was 98.4%, with 96.2% concordant duplex calls (Table 5).

Analysis of Chloroplast Reads

As an initial analysis of chloroplast reads, 100,000 paired-end reads per cultivar, with index sequences on both sides, were mapped to the chloroplast reference genome. A total of 241 sequence variants, covering the chloroplast genome, were identified (VCF-file S2). Since chloroplast sequences are monomorphic, the chloroplast haplotypes could be directly inferred. Four main chloroplast types, with a number of sub-types, were found using a phylogenetic approach (Figure 8 and Table A in Supporting Information S1). A number of cultivars contained distinct chloroplast genomes resembling those of *S. demissum*, *S. vernei* and *S. tuberosum* Group Phureja, but most resembled those commonly found in *S. tuberosum* Group Tuberosum cultivars.

Discussion

Target-enriched Genome Sequencing

This paper presents evidence that GBS in complex polyploid genomes can be performed accurately, especially with regard to the specific requirement of accurate allele dosage detection. Below, we will argue that our method of target enrichment was of great importance for genotyping accuracy. Target enrichment allowed the achievement of sufficient read depth. Furthermore, bait-hybridisation enrichment allowed exclusion of complex multi-copy or repetitive regions of the genome, and avoidance of potential difficulties in distinguishing allelic variants from paralogous sequences [51]. We focussed on a single copy genes, e.g. from the set of conserved orthologous sequence genes (COSII) [30] and used uniquely mapped genes to define the genomic target regions. It is not excluded that individual potato genotypes may be polymorphic for specific gene duplications (Copy Number Variants). Indeed, for some genes read depths were observed twice or more than average in specific cultivars, suggesting gene duplications in these cultivars (data not shown). In these rare cases the duplication maps to the same location on the reference genome and could affect genotype calling. However, we feel confident that this is a relatively rare phenomenon, because the validation of genotype calling with KASP assays showed high concordance.

The SureSelect enrichment method yielded sequences of which 45% aligned to the target sequences. Other in-solution DNA
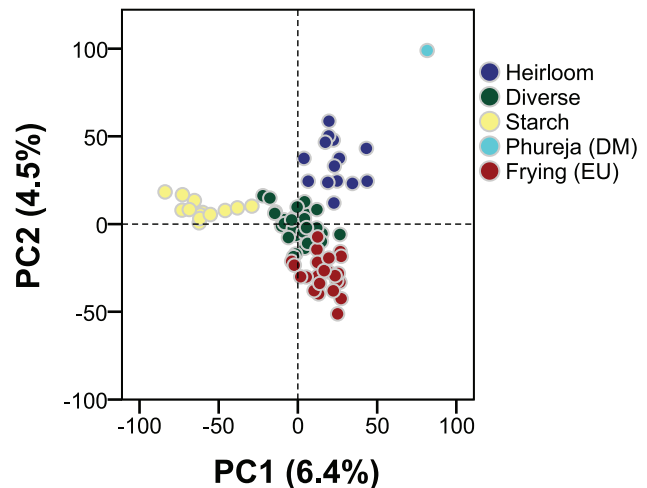

Figure 6. First and second components from principal component analysis of potato sequence variant genotypes. Population structure was analysed using ~43 K sequence variants genotyped in all 84 cultivars. The first three components describe 14.7% of the variance. Based on these three components, the cultivars were clustered into five groups. The most distant cultivar is the monoploid *S. tuberosum* Group Phureja clone. In the centre of the PCA plot cultivars of diverse, worldwide origins are observed. Three additional divergent groups can be observed, consisting of heirloom cultivars, frying cultivars from continental Europe and cultivars and germplasm used in starch industry.
doi:10.1371/journal.pone.0062355.g006

enrichment studies in plant species have not yet been published, but this on target percentage is consistent with the 40–50% reached in human and animal studies that have used the SureSelect system for target enrichment [19,40].

To reduce sequencing costs we multiplexed twelve cultivars in a sequencing pool. The custom index adapters have proven valuable as multiplex adapters; no index-specific bias in read counts was observed and over 96% of the generated reads were assigned to cultivars. The inclusion of these indices permits tracking of the alleles' cultivar source so that zygosity can be determined in individual genotypes, allowing GBS of many individuals. Larger numbers of indices can be easily created [17] and allow multiplexing of tens to hundreds of individuals. We found consistent enrichment across all indexed samples, and virtually all target sequences were covered covered at a sufficient depth, with a median average read depth of 88× per cultivar for nuclear target sequences (63× for the extended "accessible" genome). The read depth of nuclear sequences was reduced by the large number of cpDNA reads (60% of total reads) generated in this study. This large share of cpDNA reads was unforeseen in view of the 4% share of cpDNA baits used for target enrichment. In retrospect, this large proportion is almost certainly due to the copy number variation between nuclear DNA and cpDNA. The latter is present in about 100 copies per plastid and about 100 plastids per leaf cell [52]. Hence, bait library design should avoid sequences from, or homologous to extra-nuclear DNA.

In the estimate of allele copy number we assume no ascertainment bias; i.e., the relative number of allele-specific sequencing reads is proportional to the allele copy number. However, in-solution hybridisation based target enrichment may bias the pool of captured DNA targets that are sequenced towards those variants which preferentially hybridize and have higher sequence similarity to the reference bait sequence [19]. Anticipating this bias, we used a high probe tiling redundancy (6×, one
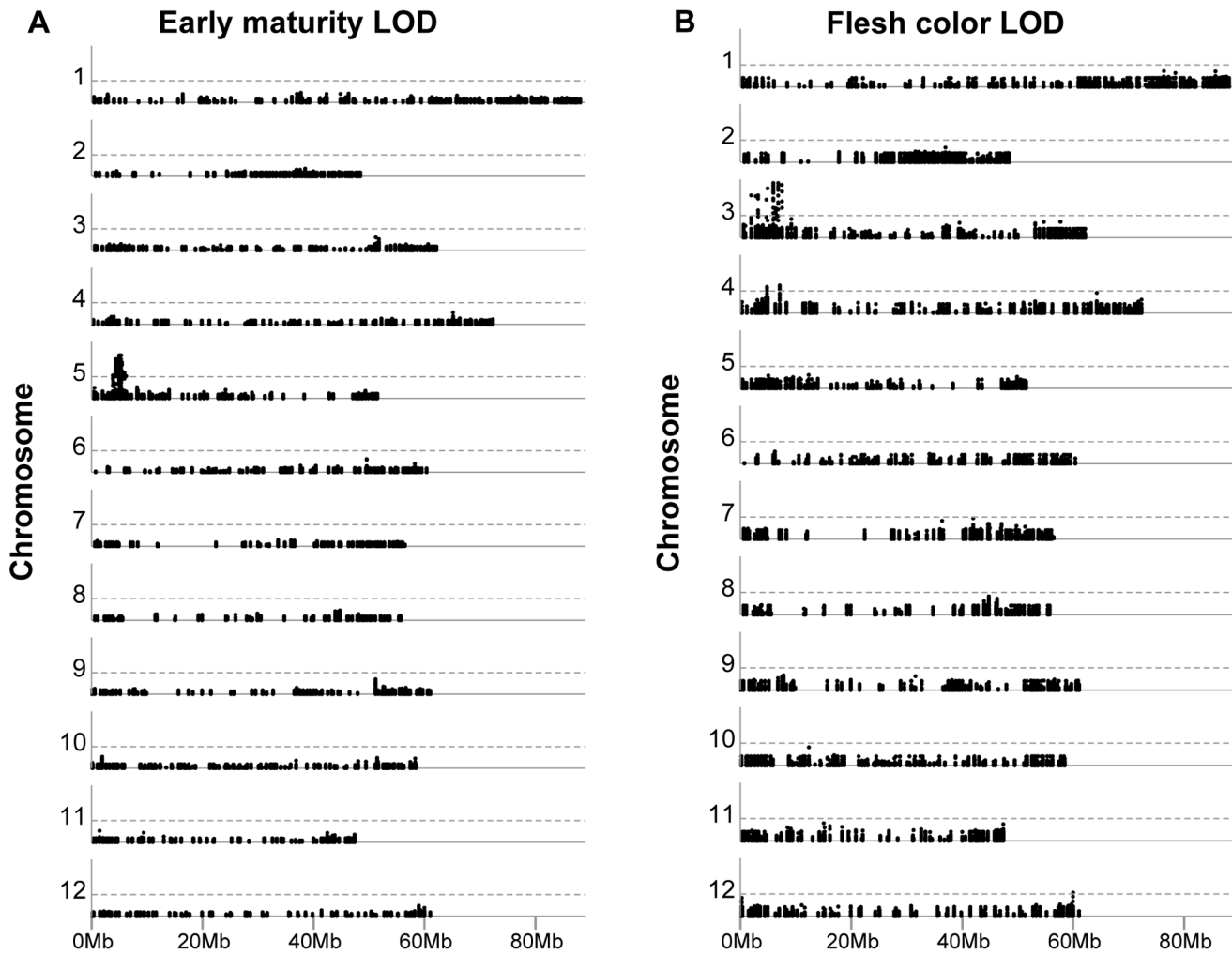
Figure 7. Manhattan plots of *p* values for associations between DNA sequence variants and two phenotypic traits in potato: (A) plant maturity and (B) tuber flesh colour. The FDR corrected $-\log_{10}(p)$ values from GWAS analysis are plotted relative to the physical position on each of 12 potato chromosomes. The horizontal dashed line is plotted at the FDR-corrected significance threshold of $\alpha = 0.001$ ($-\log_{10}(p) = 3$).
doi:10.1371/journal.pone.0062355.g007

RNA bait every ~20 bp) to reduce allele-specific bias during hybridization. Furthermore, we relaxed mapping quality settings for counting the reads mapped to the reference genome. In a

preliminary genotype analysis performed after sequence variant calling, only reads with a very high mapping quality ($\geq$MQ30) were used to determine the relative number of allele-specific

Table 5. Concordance between genotyping-by-sequencing and KASP genotyping calls.

Genotyping-by-sequencing	Expected genotype call (KASP)					
	Nulliplex	*Simplex*	*Duplex*	*Triplex*	*Quadruplex*	*Total*
Concordant calls	4,975	4,861	2,204	1,373	1,272	14,685
Discordant calls	67	25	86	34	25	237
Percentage of concordant calls	98.7%	99.5%	96.2%	97.6%	98.1%	98.4%
Median read-depth of concordant calls	85×	91×	101×	92×	75×	89×
Median read-depth of discordant calls	52×	56×	73×	83×	47×	61×
Percentage of alternative reads for concordant calls	0%	23%	49%	76%	100%	–
Percentage of alternative reads for discordant calls	4%	36%	34%	56%	97%	–

The genotype calls derived from each method for 270 binary SNPs were compared. Sequencing calls were filtered by a minimum read depth of 15× and a minimum genotype quality score of GQ26.
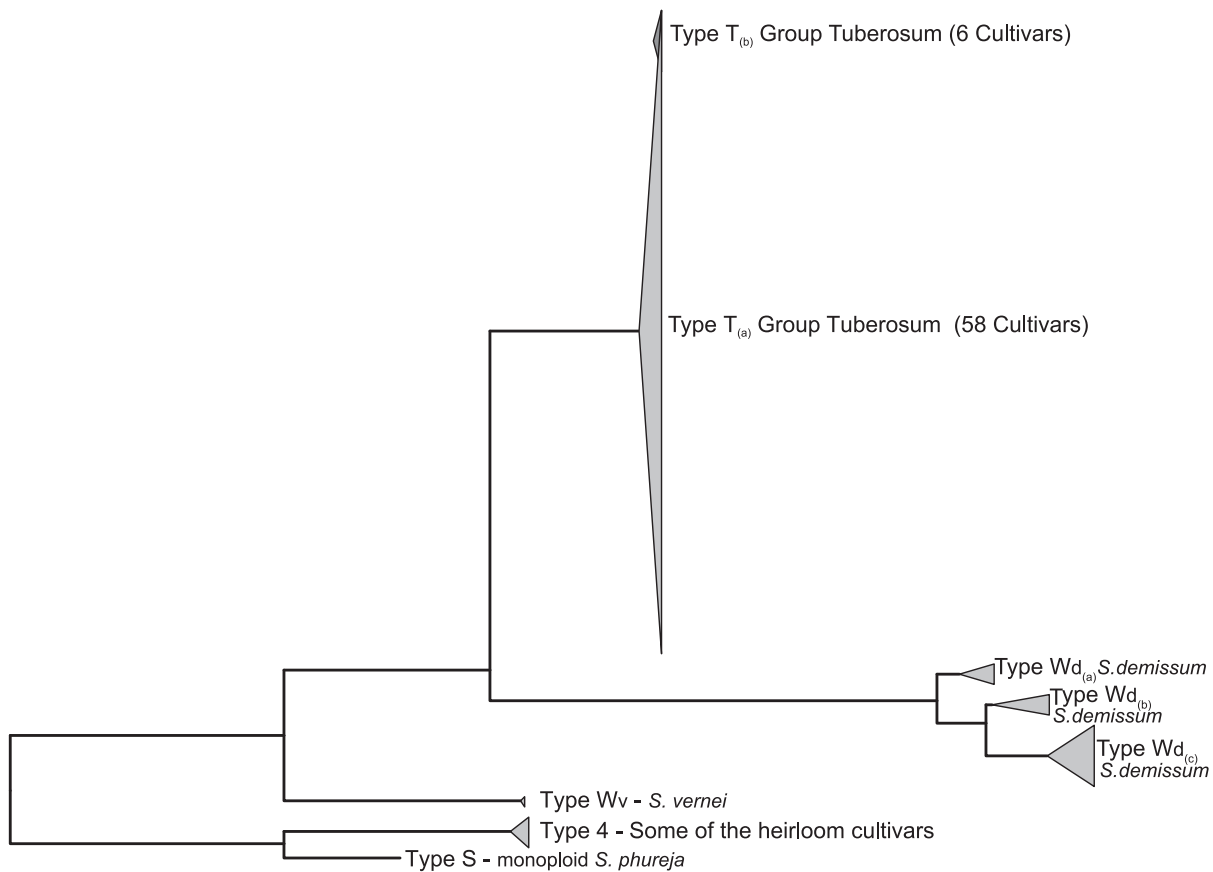doi:10.1371/journal.pone.0062355.t005

Figure 8. Neighbour-Joining tree of chloroplast haplotypes. The distances of 241 sequence variants for the 84 cultivars were computed using the Jukes-Cantor method and the tree inferred using the Neighbor-joining method.
doi:10.1371/journal.pone.0062355.g008

sequencing reads. This caused a large ascertainment bias due to severe underrepresentation of non-reference alleles. Inclusion of reads with a lower mapping quality (MQ≥13) strongly reduced this ascertainment bias. When the GQ-filter was applied, 95% of the sequence variants were concordant, as shown by the KASP validation (Table 5). The average read ratio for these calls shows no bias. We conclude from this study that hybridisation based capture bias hardly affected genotype calling, even in potato germplasm where reference and non-reference reads often have only about 95% sequence homology. We have not examined the genotyping accuracy of MNPs or indels, but larger indels are likely to reduce capture and mapping efficiency.

We used a mapping approach to align the sequencing data to the potato genome sequence [25]. An advantage of alignment to an annotated reference genome is that it allows prediction of whether a sequence variant falls within or near a gene of interest, and whether it is expected to cause a functional change in the protein product (e.g. synonymous vs. non-synonymous) that might alter the enzyme activity of the protein. This can be very useful in determining whether a particular sequence variant is likely to be responsible for a phenotype of interest. We observed approximately twenty-five thousand non-synonymous variants, with 50 variants that cause premature stop codons in 43 out of 807 genes. Hence, natural diversity of potato offers a substantial resource for the identification of knockouts for gene-function studies.

A disadvantage of mapping sequence reads toward a reference sequence is that structural variation like chromosome rearrangements, inversions and large (transposon) insertions are likely to be

missed. Structural variation between reads and reference genome complicates the mapping of reads. This disadvantage can be partly avoided by *de novo* assembly, but computational difficulties associated with the assembly of highly diverse polyploid species like potato makes mapping sequence reads to a reference sequence a more straightforward approach. In this study we did not systematically look for structural variants, nor did we consider gene copy number variants (CNVs). Methods for detecting both structural variants and CNVs using a mapping approach have however been developed [53,54] and could be applied. Fortunately, structural variants can frequently be detected using sequence variants in linkage disequilibrium (LD). For example, orange potato flesh colour due to zeaxanthin accumulation, caused by a transposon insertion in the zeaxanthin epoxidase gene was initially detected using SNPs [34].

Sequence Diversity Analysis

This study confirms the high sequence diversity of potato, albeit using a much higher number of potato cultivars and on a more genome-wide scale than previous studies [55,56]. We found an overall frequency of one variant every 17 bp in the cultivar population, 1/15 bp in non-coding and 1/24 bp in coding regions. This is somewhat higher than the one variant per 21–23 bp reported in previous studies on potato [55,56], reflecting the differences in the number of potato cultivars used. Compared to other crops where both coding and noncoding regions were analyzed for nucleotide diversity, total nucleotide diversity in potato cultivars ($\pi = 10.7 \times 10^{-3}$) is larger than that in sugar beet

(7.6×10^{-3}) [57], maize elite lines (6.3×10^{-3}) [58], and soybean (1.25×10^{-3}) [59]. Simko et al. (2006) presented a somewhat higher nucleotide diversity for potato of $\pi = 14.6 \times 10^{-3}$, but their sample included a number of distantly related wild potato species [56].

For discovery of common sequencing variants (i.e. variants with a high MAF), only a limited number of potato cultivars have to be sequenced, as more than half of all variants were already detected by sequencing three random cultivars. In a previous study in potato, SNPs have been identified in cDNA of a set of three to six – mainly North-American – potato cultivars, using Sanger EST-sequences available in GeneBank from three potato cultivars, and using high-throughput transcriptome sequencing on three additional cultivars [2]. The SNPs identified in this relatively small sample, are expected to mainly represent the common SNPs from all coding sequences, whereas this study included introns and sampled rare alleles. Based on exons sequenced in both studies an overlap of only 2,572 variants was expected between our 129,156 variants and the 69,586 mapped SNPs identified by Hamilton et al. (2011). We detected 2,362 (92%) of these SNPs. As only one cultivar was included in both studies (cv. Bintje), the 8% of variants undetected in our study might represent rare variants more specific to the North American cultivars predominantly sampled by Hamilon et al. (2011), or may be false positive/negatives in either study. As a result of the at least 14-fold larger sample size and larger geographic diversity targeted by our study, we find approximately 15,000 extra variants in the exon sequences covered by both studies. These include multi-allelic sequence variants at 6% of the base positions where SNPs were called by Hamilon et al. (2011). When these SNPs are assumed bi-allelic, they may complicate genotype calling in a SNP genotyping array.

An important feature affecting the application of sequence variants as molecular markers is their MAF, which influences the type of information provided by the marker in different populations. Moderate-frequency alleles are valuable for mapping studies, where it is desirable to maximize the number of polymorphic markers between two parental lines. Rare variants can, if not assayed, cause ascertainment bias in phylogenetic reconstructions, overestimations of mean diversity [60], and spurious correlations in association mapping [61]. This paper presents an L-shaped distribution of the MAF (Figure 5) which points to an abundance of rare alleles in the cultivated gene pool. In comparison with another vegetatively propagated outbreeder, *Vitis vinifera*, the average MAF of potato (0.14) is lower. *Vitis vinifera* has an average MAF of 0.24 [62]. In a population of 80 grapevine cultivars, over 80% of SNP variants had a MAF above 0.10, while in potato we found only 48% of the variants had a MAF above 0.10. We expect that wild species introgressions with low MAF contribute significantly to the high variant density in potato.

To investigate whether we could identify potential signatures of selection in the potato genome, we examined the nucleotide diversity of loci along the physical position of each chromosome. Nucleotide diversity values were very variable both between and within chromosomes. Chromosomes 5 and 11 exhibited highest nucleotide diversity. Introgression of resistance genes from wild species is a likely explanation, as chromosome 5 and 11 contain the largest clusters of resistance genes, conferring resistance to a wide variety of pathogens [63,64]. In contrast, overall nucleotide diversity was reduced on chromosome 10. Some of the most conserved genes in the dataset are located near the position of the skin colour, tuber shape and eye depth QTL on chromosome 10 [65–67]. An eye-catching difference between commercial western germplasm and Latin American land races are the highly variable tuber shapes and anthocyanin pigmentation patterns in tuber skin and flesh. Hence this reduced nucleotide diversity could well reflect a selective sweep.

Within all chromosomes, individual genes were found with reduced nucleotide diversity as well. It is not possible at this stage to say whether these genes themselves, rather than closely linked loci are under selection, but some of these genes are good candidate genes for phenotypic traits under strong selection like day-length dependent tuberisation and some resistance traits. An example of a good direct candidate gene is the *CONSTANS* gene that has been shown to affect the day-length regulation of tuber induction in potato [68] and is the second most conserved gene we sampled on chromosome 2.

Chloroplast Types and Population Structure

The presence of an extreme cytoplasmic bottleneck in cultivated potato has been described before [69]. Five main cpDNA types (A, S, C, W and T) and a number of sub-types have been described [70]. The A haplotype is most frequent in Group Andigena and the T haplotype in Chilean *S. tuberosum* and modern cultivars. Diploid *S. tuberosum* Group Phureja is assigned to the S-type and wild material like *S. vernei* and *S. demissum* to the W-type. Our phylogeny of chloroplast haplotype data supports previous work suggesting that most modern cultivars have chloroplasts resembling those of the *S. tuberosum* Group Tuberosum genepool (T-Type) [71,72]. Two cultivars have a W-type haplotype originating from *S.vernei* (VE71–105 and VTN62-33-3) and thirteen cultivars have a W-type haplotype originating from *S. demissum* [73]. These chlorotypes have been introduced during introgression of resistance traits. More remarkably, four heirloom cultivars in our sample (cv. Belle De Fontenay, cv. Kepplestone Kidney, cv. Home Guard, and cv. Shamrock) have chloroplast haplotypes phylogenetically close to that from the *S. tuberosum* Group Phureja (S-Type). This chlorotype might represent the *S. tuberosum* Group Andigena type (A-Type) that was more common in cultivars from before the 1840s late blight (*Phytophtra infestance*) epidemic [74].

The genome-wide population structure of this study largely corresponds with the structure observed using AFLP markers [48]. While the AFLP study distinguished five divergent cultivar groups, we identified four, with one outlier forming its own group. Both analyses support genetic similarity within heirloom cultivars and within starch cultivars, but the high-throughput approach used here did not support separate groups of fresh consumption, processing (crisps or fries), and additional miscellaneous cultivars. Instead, PCA analysis suggested the existence of a cluster of cultivars originating from continental Europe, consisting mainly of processing cultivars, along with a cluster containing cultivars of mixed, world-wide origin. Given that the first three components of the principal component analysis account for less than 15% of the total variation, it seems that there is little population structure within the cultivated potato gene pool, as was also observed by D'hoop et al. (2010).

Concordance between GBS and KASP Genotyping in an Autotetraploid

In diploids it is known that the accuracy of variant detection and genotyping depends on sequence depth and the mode of sexual reproduction (homozygous selfers vs. heterozygous outbreeders) [75,76]. Genotyping variants in polyploid species, such as potato, is more challenging than in diploids, because a given gene may be represented not only by a number of different alleles, but also by different zygosity levels. We therefore tested the accuracy of genotype calls made by GBS by testing a small subset of binary SNPs using an independent KASP genotyping platform. The genotype calls of GBS were found to be over 99% consistent with

KASP when scored as homozygous or heterozygous. We applied a genotype quality (GQ) parameter to account for differences in read depth requirement across all zygosity classes (nulliplex, simplex, duplex, triplex, and quadruplex). Over 95% of duplex calls (the zygosity class with lowest consistency between GBS and KASP) were consistent between the two genotyping methods at a threshold of GQ26. Overall concordance of the five zygosity classes was 98.4% at this threshold. In conclusion, for genotyping autotetraploids a GQ26 is recommended to achieve reliable genotype calls using our pipeline. In a more general sense, our results indicate that read depths of 60–80x are sufficient for reliable, dosage diagnostic genotype calls in GBS in autotetraploid potato.

Association Analysis using GBS

The GBS dataset was tested for its performance in GWAS. Even though the data do not represent a complete genome, but a focussed survey at 807 loci randomly distributed across the genome, and although the number of 83 potato cultivars severely limits statistical power, several interesting traits could be mapped. It appeared that the 129,156 sequence variants provide a robust dataset to detect alleles influencing monogenic traits such as plant maturity and tuber flesh colour (this paper) and various other monogenic traits such as tuber shape and pathogen resistance (unpublished data). The sequence variants associated with tuber flesh colour and plant maturity are indeed located in and near the respective candidate genes *CHY2* [34] and *StCDF1*. The latter was recently shown to be the causal gene involved in potato day-length signalling [77]. We observed that a simplistic dominant model (0/1 data reflecting absence/presence of an allele) gave similar GWAS results as the more detailed additive model (0/1/2/3/4 copies of the allele). This makes sense, since 61% of the sequence variants have a MAF below 0.05. If such variants follow Hardy-Weinberg expectations, in practice the minor allele will be absent in ~81% of the cultivars and simplex in ~17% of the cultivars, while a duplex cultivar is hardly expected (1.4%). This notion, that many sequence variants have a low allele frequency, points to the advantage that the input GBS data does not require the more stringent genotype quality filtering necessary for accurate estimation of allele copy number.

GWAS requires marker densities surpassing the decay in LD. In this study only a fraction (2.1 Mb, 0.25%) of the 840 Mb potato genome was re-sequenced, but in most cases the genetic distance between our 807 candidate genes will be less than the estimated LD decay of 4–10 cM [48,56]. For the data generated here, both short- and long-range LD still needs to be analysed. Although short-range LD could have been analysed, long-range LD requires the robust ordering of superscaffolds of the DM reference genome into physical chromosomes, preferentially in combination with an aligned high-resolution genetic linkage map. These maps and pseudo-molecules are currently being developed to estimate LD decay on both genetic (cM) and physical (bp) scale in the near future. We expect that the identified sequence variants and regions covered in this study are in LD with most functional alleles. We also expect that identification of haplotypes and phasing of the sequence variants can be achieved by using a 'read-backed phasing' approach (unpublished data) for use in a multiallelic GWAS analysis. Since a large number of genes in our study were either known candidate genes, or primary- and secondary metabolism genes, the annotated biochemical pathway may suggest a biological link between the gene underlying the sequence variants, and the QTL. Alternatively, the potato genome browser (http://solanaceae.plantbiology.msu.edu) will support the identification of candidate genes in proximity to QTL.

Supporting Information

Supporting Information S1 Supporting Tables and Figures.
(DOC)

BED-file S1 SureSelect baits mapped (MQ≥37) to the superscaffolds of the DM reference genome.
(BED)

BED-file S2 Accessible genome regions of the DM reference genome.
(BED)

CSV-file S1 Annotations for sequence variants identified in the accessible genome, including allele copy numbers of each of the 84 samples.
(GZ)

FASTA-file S1 SureSelect RNA bait sequences.
(FASTA)

FASTA-file S2 SureSelect target sequences.
(FASTA)

VCF-file S1 Sequence variants and genotypes identified in the accessible potato DM genome of 84 samples.
(GZ)

VCF-file S2 Sequence variants identified in the chloroplast genome.
(VCF)

XLS-file S1 Annotations for genomic SureSelect targets, including observed coverage in resequencing data.
(XLS)

XLS-file S2 Annotations for accessible genome regions, based on sequence data collected in this study.
(XLS)

XLS-file S3 Annotations of nucleotide diversity for accessible genes.
(XLS)

Author Contributions

Conceived and designed the experiments: JU AW BD HvE. Performed the experiments: JU. Analyzed the data: JU TB. Contributed reagents/materials/analysis tools: TB. Wrote the paper: JU RV HvE.

References

1. Grattapaglia D, Junior O, Kirst M, de Lima B, Faria D, et al. (2011) High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biology 11: 65.

2. Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, et al. (2011) Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC genomics 12.

3. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631.

4. Close T, Bhat P, Lonardi S, Wu Y, Rostoks N, et al. (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC genomics 10: 582.

5. Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. 42: 961–967.

6. Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166–177.

7. Moose SP, Mumm RH (2008) Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. Plant physiology 147: 969–977.

8. Rickert AM, Premstaller A, Gebhardt C, Oefner PJ (2002) Genotyping of SNPs in a polyploid genome by pyrosequencing (TM). Biotechniques 32: 592−+.

9. Sattarzadeh A, Achenbach U, Lübeck J, Strahwald J, Tacke E, et al. (2006) Single nucleotide polymorphism (SNP) genotyping as basis for developing a PCR-based marker highly diagnostic for potato varieties with high resistance to Globodera pallida pathotype Pa2/3. Molecular Breeding 18: 301–312.

10. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5: 887–893.

11. Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Research 15: 1767–1776.

12. Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, et al. (2009) Large-Scale Discovery of Gene-Enriched SNPs. Plant Gen 2: 121–133.

13. Griffin P, Robin C, Hoffmann A (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC Biology 9: 19.

14. Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, et al. (2010) Rapid Genomic Characterization of the Genus Vitis. PLoS one 5: e8219.

15. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, et al. (2010) The Characterization of Twenty Sequenced Human Genomes. PLoS genetics 6: e1001111.

16. Gore M, Bradbury P, Hogers R, Kirst M, Verstege E, et al. (2007) Evaluation of Target Preparation Methods for Single-Feature Polymorphism Detection in Large Complex Plant Genomes. Crop Sci 47: S-135–S-148.

17. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PloS one 6: e19379.

18. van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, et al. (2007) Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PloS one 2: e1172.

19. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182–189.

20. Kiss MM, Ortoleva-Donnelly L, Reginald Beer N, Warner J, Bailey CG, et al. (2008) High-throughput quantitative polymerase chain reaction in picoliter droplets. Analytical Chemistry 80: 8975–8981.

21. Nijman IJ, Mokry M, van Boxtel R, Toonen P, de Bruijn E, et al. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. 7: 913–915.

22. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. 12: 499–510.

23. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. 7: 111–118.

24. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, et al. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. 4: 960–974.

25. Xu X, Pan S, Cheng S, Zhang B, Mu D, et al. (2011) Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.

26. Menda N, Buels RM, Tecle I, Mueller LA (2008) A community-based annotation framework for linking solanaceae genomes with phenomes. Plant physiology 147: 1788–1799.

27. Meyer S, Nagel A, Gebhardt C (2005) PoMaMo–a comprehensive database for potato genome data. Nucleic Acids Research 33: D666–670.

28. Anithakumari AM, Tang J, van Eck HJ, Visser RG, Leunissen JA, et al. (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Molecular breeding : new strategies in plant improvement 26: 65–75.

29. Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC bioinformatics 7: 438.

30. Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174: 1407–1420.

31. Gebhardt C, Chen X, Salamini F (2001) A potato molecular-function map for carbohydrate metabolism and transport. Theoretical and Applied Genetics 102: 284–295.

32. Martin C, Niggeweg R, Michael AJ (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. Nature Biotechnology 22: 746–754.

33. Mccue KF, Allen PV, Shepherd LVT, Blake A, Rockhold DR, et al. (2007) Manipulation and compensation of steroidal glycoalkaloid biosynthesis in potatoes. Proceedings of the VIth International Solanaceae Conference, Solanaceae VI: Genomics Meets Biodiversity: 343–349.

34. Wolters AMA, Uitdewilligen JGAML, Kloosterman BA, Hutten RCB, Visser RGF, et al. (2010) Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. Plant Molecular Biology 73: 659–671.

35. Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203–211.

36. Bertone P, Trifonov V, Rozowsky JS, Schubert F, Emanuelsson O, et al. (2006) Design optimization methods for genomic DNA tiling arrays. Genome Research 16: 271–281.

37. D'hoop BB, Paulo MJ, Mank RA, van Eck HJ, van Eeuwijk FA (2008) Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica 161: 47–60.

38. Veilleux RE, Lightbourn GJ (2007) Production and evaluation of somatic hybrids derived from monoploid potato. American Journal of Potato Research 84: 425–435.

39. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, et al. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36: e122.

40. Harismendy O, Tewhey R, Nakano M, Wang XY, Pabon-Pena C, et al. (2009) Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology 10.

41. Durbin R, Li H (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

42. Durbin R, Li H, Handsaker B, Wysoker A, Fennell T, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

43. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303.

44. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

45. Garrison EaM G (2012) Haplotype-based variant detection from short-read sequencing. ArXiv e-prints: 1207.3907.

46. Cingolani P, Platts A, Wang L, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6.

47. Lê S, Josse J, Husson F (2008) FactoMineR: An R package for multivariate analysis. Journal of Statistical Software 25: 1–18.

48. D'hoop BB, Paulo MJ, Kowitwanich K, Sengers M, Visser RGF, et al. (2010) Population structure and linkage disequilibrium unravelled in tetraploid potato. Theoretical and Applied Genetics 121: 1151–1170.

49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559–575.

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

51. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. 461: 272–276.

52. Daniell H (2004) Chloroplast genetic engineering for improved agronomic traits and molecular farming, using various selection systems. In Vitro Cellular & Developmental Biology-Animal 40: 18a–18a.

53. Lam HYK, Mu XJ, Stutz AM, Tanzer A, Cayting PD, et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. 28: 47–55.

54. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al. (2012) Copy number variation detection and genotyping from exome sequence data. Genome Research.

55. Rickert AM, Kim JH, Meyer S, Nagel A, Ballvora A, et al. (2003) First-generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. Plant Biotechnology Journal 1: 399–410.

56. Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173: 2237–2245.

57. McGrath J, Shaw R, de los Reyes B, Weiland J (2004) Construction of a sugar beet BAC library from a hybrid with diverse traits. Plant Molecular Biology Reporter 22: 23–28.

58. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, et al. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. Bmc Genetics 3.

59. Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, et al. (2003) Single-Nucleotide Polymorphisms in Soybean. Genetics 163: 1123–1134.

60. Schlötterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in Drosophila melanogaster. Molecular ecology 11: 947–950.

61. Pritchard JK (2001) Deconstructing maize population structure. Nature Genetics 28: 203–204.

62. Lijavetzky D, Cabezas J, Ibanez A, Rodriguez V, Martinez-Zapater J (2007) High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC genomics 8: 424.

63. Bakker E, Borm T, Prins P, Vossen E, Uenk G, et al. (2011) A genome-wide genetic map of NB-LRR disease resistance loci in potato. 123: 493–508.

64. Jupe F, Pritchard L, Etherington G, MacKenzie K, Cock P, et al. (2012) Identification and localisation of the NB-LRR gene family within the potato genome. BMC genomics 13: 75.

65. Li X-Q, De Jong H, De Jong D, De Jong W (2005) Inheritance and genetic mapping of tuber eye depth in cultivated diploid potatoes. 110: 1068–1073.

66. Van Eck HJ, Jacobs J, Stam P, Ton J, Stiekema WJ, et al. (1994) Multiple Alleles for Tuber Shape in Diploid Potato Detected by Qualitative and Quantitative Genetic Analysis Using RFLPs. Genetics 137: 303–309.

67. Van Eck HJ, Van Der Voort JR, Draaistra J, Van Zandvoort P, Van Enckevort E, et al. (1995) The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Molecular Breeding 1: 397–410.

68. González-Schain ND, Díaz-Mendoza M, Żurczak M, Suárez-López P (2012) Potato CONSTANS is involved in photoperiodic tuberization in a graft-transmissible manner. The Plant Journal 70: 678–690.

69. Provan J, Powell W, Dewar H, Bryan G, Machray GC, et al. (1999) An extreme cytoplasmic bottleneck in the modern European cultivated potato (Solanum tuberosum) is not reflected in decreased levels of nuclear diversity. Proceedings of the Royal Society of London Series B: Biological Sciences 266: 633–639.

70. Hosaka K, de Zoeten GA, Hanneman Jr RE (1988) Cultivated potato chloroplast DNA differs from the wild type by one deletion - evidence and implications. Theoretical and Applied Genetics 75: 741–745.

71. Hosaka K (2004) Evolutionary pathway of T-type Chloroplast DNA in potato. American Journal of Potato Research 81: 153–158.

72. Spooner DM, Nuñez J, Rodríguez F, Naik PS, Ghislain M (2005) Nuclear and chloroplast DNA reassessment of the origin of Indian potato varieties and its implications for the origin of the early European potato. Theoretical and Applied Genetics 110: 1020–1026.

73. Van Berloo R, Hutten RCB, Van Eck HJ, Visser RGF (2007) An online potato pedigree database resource. Potato Research 50: 45–57.

74. Powell W, Baird E, Duncan N, Waugh R (1993) Chloroplast DNA variability in old and recently introduced potato cultivars. Annals of Applied Biology 123: 403–410.

75. Kenny EM, Cormican P, Gilks WP, Gates AS, O'Dushlaine CT, et al. (2011) Multiplex Target Enrichment Using DNA Indexing for Ultra-High Throughput SNP Detection. DNA Research 18: 31–38.

76. Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, et al. (2009) Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology 10.

77. Kloosterman B, Abelenda JA, Gomez MdMC, Oortwijn M, de Boer JM, et al. (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495: 246–250.