A pipeline for high throughput detection and mapping of SNPs from EST databases

A. M. Anithakumari · Jifeng Tang · Herman J. van Eck · Richard G. F. Visser · Jack A. M. Leunissen · Ben Vosman · C. Gerard van der Linden

Received: 21 August 2009 / Accepted: 16 December 2009 / Published online: 20 January 2010

Abstract Single nucleotide polymorphisms (SNPs) represent the most abundant type of genetic variation that can be used as molecular markers. The SNPs that are hidden in sequence databases can be unlocked using bioinformatic tools. For efficient application of these SNPs, the sequence set should be error-free as much as possible, targeting single loci and suitable for the SNP scoring platform of choice. We have developed a pipeline to effectively mine SNPs from public EST databases with or without quality information using QualitySNP software, select reliable SNP and prepare the loci for analysis on the Illumina GoldenGate genotyping platform. The applicability of the pipeline was demonstrated using publicly available potato EST data, genotyping individuals from two diploid mapping populations and subsequently mapping the SNP markers (putative genes) in both populations. Over 7000 reliable SNPs were identified that met the criteria for genotyping on the GoldenGate platform. Of the 384 SNPs on the SNP array approximately 12% dropped out. For the two potato mapping populations 165 and 185 SNPs segregating SNP loci could be mapped on the respective genetic maps, illustrating the effectiveness of our pipeline for SNP selection and validation.

Keywords EST database · Illumina GoldenGate assay · QualitySNP · Potato

Introduction

Genetic variation is the basis for the biodiversity of life (Schlotterer 2004). Variations in the DNA sequence of genes and their regulatory regions underlie most of the phenotypic variation that has been exploited in modern crops (Bryan et al. 2000; Masouleh et al. 2009). Breeding strategies aiming to

Electronic supplementary material The online version of this article (doi:10.1007/s11032-009-9377-5) contains supplementary material, which is available to authorized users.

J. Tang: formerly at Wageningen UR Laboratory of Bioinformatics, Wageningen, The Netherlands.

A. M. Anithakumari
Graduate School Experimental Plant Sciences,
Wageningen UR Plant Breeding, PO Box 386, 6700 AJ
Wageningen, The Netherlands

H. J. van Eck · R. G. F. Visser · B. Vosman ·
C. G. van der Linden (✉)
Wageningen UR Plant Breeding, Wageningen University
& Research Centre, PO Box 386, 6700 AJ Wageningen,
The Netherlands
e-mail: gerard.vanderlinden@wur.nl

J. Tang
Keygene N.V, Wageningen, The Netherlands

J. A. M. Leunissen
Wageningen UR Laboratory of Bioinformatics,
Wageningen University & Research Centre, Wageningen,
The Netherlands

improve crop agronomical performance have gained momentum in the last few decades by the use of molecular marker technologies that visualize DNA polymorphisms (Collard et al. 2005). Molecular markers have proven to be extremely useful in breeding, for genome-wide screens for variation, genotype identification and/or fingerprinting, evolutionary and ecological studies.

In breeding programs that are aimed at transferring genes or alleles within or between different species with the aid of molecular markers several steps can be discerned. The first step in this process is the identification of one or more markers closely linked to or within the traits to be introgressed. For this, a high density map of markers on the genome and/or markers in genes that are likely to be involved in the trait of interest can be invaluable tool. SNPs are very well suited for this purpose. Their astonishing abundance has been reported in several discovery projects in many species including humans (Sachidanandam et al. 2001), model species such as *Arabidopsis thaliana* (Jander et al. 2002) and *Drosophila melanogaster* (Hoskins et al. 2001) and in crop plants such as barley (Rostoks et al. 2005), maize (Ching et al. 2002), rice (Shen et al. 2004; McNally et al. 2006), soybean (Zhu et al. 2003) and wheat (Ablett et al. 2006).

Recent technological advancements in discovery and detection platforms have made SNP markers attractive for high-throughput use not only in model species, but also in crop plants (Rafalski 2002). In species for which no genome sequence is available, large scale SNP discovery has generally relied on sequence variation found in libraries of expressed sequence tags (ESTs) (Somers et al. 2003) or on re-sequencing (Choi et al. 2007).

Several software tools are available for SNP discovery from nucleotide databases, including PolyBayes, AutoSNP, and QualitySNP (Marth 1999; Barker et al. 2003; Tang et al. 2006). QualitySNP is especially useful in extracting reliable SNPs from EST sequence databases that lack quality information, and is in many cases capable of distinguishing paralogs from allelic sequences effectively (Tang et al. 2006). Along with the development of tools to mine a large number of SNPs from nucleotide databases, new SNP genotyping platforms were developed that can analyze a large number of SNPs in parallel in a large set of individuals (Syvanen 2005). An increasing number of reports indicate that the GoldenGate system of Illumina is a reliable and cost-effective SNP genotyping platform. It is capable of multiplexing from 96 to 1536 SNPs in a single reaction (Fan et al. 2003).

In this paper we describe a bioinformatics pipeline starting from SNP discovery in ESTs to genotyping using the Illumina GoldenGate assay. Following SNP discovery, the SNP loci are further screened for suitability to be analyzed with the Illumina GoldenGate Genotyping platform. We demonstrate the applicability of this pipeline for potato, which is the third most important food crop in the world. Potato is a heterozygous crop, and commercial varieties are generally tetraploid. For potato, approximately 200,000 ESTs mainly from three cultivars are publicly available. We show here that SNPs identified by QualitySNP from this collection of SNPs can effectively be turned into markers that can be mapped in different diploid potato mapping populations, showing the versatility of the pipeline and the produced SNP markers. Our results indicate that the pipeline produces a large number of SNP markers, and that the selection of SNPs for genotyping on the Illumina GoldenGate genotyping platform yields a high number of reliable functional co-dominant markers that can be easily placed on a genetic map.

Materials and methods

Mapping populations

(a) SH × RH: A cross between two diploid heterozygous potato clones SH83-92-488 and RH89-039-16 (SH × RH) resulted in an F1 mapping population of 135 individuals (van Os et al. 2006). Using a Selective Mapping strategy (Vision et al. 2000) 57 individuals were selected which captured the highest number of recombination events.

(b) C × E: This diploid backcross population consisting of 250 genotypes was obtained from the cross between clones C [USW5337.3; (Hanneman RE 1967)] and E [originally named 77.2102.37; (Jacobsen 1980)]. Clone C is a hybrid between *S. phureja* PI225696.1 and *S. tuberosum* dihaploid USW42. Clone E is the result of a cross between clone C and the *S. vernei*–*S. tuberosum* backcross clone VH^{3-}4211 (Jacobsen 1978). A set of 94 randomly selected individuals was used for this study, along with the parents of the cross.

DNA extraction

Genomic DNA was isolated from 50 to 100 mg of young leaves. After freeze-drying, the leaf material was ground using the MM300 Mixer mill (Retsch Inc., Haan Germany) and DNA extraction was performed using the DNeasy 96 Plant Mini kit (Qiagen, Valencia, California, USA) according to the manufacturer's protocol.

SNP identification pipeline designed for the GoldenGate genotyping platform

For SNP discovery, 219,765 EST reads were downloaded from the EMBL database (http://www.ebi.ac.uk/embl) (version 88). Functional annotation of the ESTs was obtained from the TIGR gene index (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=potato or UniGene (Wheeler et al. 2003) and additional BLASTN and BLASTX analyses (Altschul et al. 1990). The ESTs were aligned into contigs and analyzed for true SNPs using the QualitySNP software (http://www.bioinformatics.nl/tools/snpweb), with D-value set at 0.6 and default values for quality regions and other filters as described by Tang et al. (2006). The resulting data are stored in a 'contig database'. In an additional routine/programme the QualitySNP output was analyzed for SNP loci flanked by 30–50 nt reliable sequences on each side to allow for assay development using the Illumina GoldenGate design tool. The output was formatted to fit the requirements for the assay design tool (http://www.illumina.com/downloads/GoldenGateDesign_TechNote.pdf) and stored in the '100 bp fragment' database. As for potato no reference genome sequence is available we performed an additional BLAST analysis to eliminate fragments that have more than 90% homology with each other to maximize the chances of single locus amplification. Only fragments occurring once in the contig database and with less than 90% similarity to all other contigs in the database were maintained and considered for the GoldenGate assay development.

Selection of SNPs for the Illumina GoldenGate assay

A selection of the SNP loci was made based on putative gene functions in abiotic, biotic stress responses, metabolic and biosynthesis pathways. Functional annotations were taken from the EST annotations in the DFCI potato gene index (hosted at http://compbio.dfci.harvard.edu/tgi/ as part of The Gene Index Project). For some genes several SNPs within the same gene were selected. GoldenGate primers were designed using Illumina's design tool and SNP scoring was performed by Service XS (Leiden, The Netherlands), using Illumina's high-density array technology for standard or custom SNP genotyping of 96 samples. For each sample 250 ng of DNA was used for genotyping with the Illumina standard GoldenGate protocol (Shen et al. 2005). Our experimental setup included two separate genotyping runs; one for the SH × RH population that also included the C and E parents, and one for the C × E population, again including the C and E parents). The data was analyzed using Gencall software (Illumina, San Diego, CA) which is integrated in the Illumina bead station package (http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf) (Shen et al. 2005).

Additional molecular marker development

AFLP markers were generated according to standard protocols with radioactive labels, using 4 Eco-Mse primer combinations (Vos et al. 1995). Bands were scored as present or absent. AFLP markers were encoded by standard AFLP marker coding, with an ID and a chromosomal location; for example E39M60-40c10 is a marker from the Eco39 primer and a Mse60 primer, ID number 40 and mapped on Chromosome 10. The SSR markers used in this study were obtained from different sources (Milbourne et al. 1998; Feingold et al. 2005). The CAPS markers were developed for interesting candidate genes with (putative) functions in amongst others quality traits in the CE population (manuscript in preparation by Werij et al.).

Genetic mapping

The potato SNP markers were first mapped in the two mapping populations using JoinMap 4.0 (Van Ooijen 2006) together with AFLP (only 1:1 segregating markers), SSRs and CAPS as backbone markers. SNP markers were also mapped on the existing SH × RH genetic map using a bin mapping approach (van Os et al. 2006).

Results

Potato SNP array construction

In the 219,765 EST sequences 12,184 reliable SNPs were discovered. A set of 7592 SNPs remained after extra filters were set to select for SNP loci with flanking regions suitable for primer design in Illumina's Goldengate assay (at least 30–50 nt flanking sequence on each side, no SNPs detected in the flanking regions and no other sequences that are more than 90% similar present in other contigs/clusters).

The last selection of 384 SNPs for the Illumina array was based on putative functions of the genes containing the SNP loci as deduced from annotations at the DFCI potato gene index website (hosted at http://compbio.dfci.harvard.edu/tgi/). The final selection (hereafter called 384PotSNP array) of SNP markers with their putative functions, locations along with their database ID's (TC numbers) is provided in Supplementary Table.

Evaluation of the 384PotSNP array

The 384PotSNP array was evaluated by genotyping two diploid potato populations and mapping the SNP markers. The quality of each SNP is reflected in the Gencall (GC) score, a value between 0 and 1 (Shen et al. 2005). The Gencall score is a representation of the separation between the heterozygote and homozygote clusters for a particular SNP, and how a SNP score is placed in these clusters. R values below 0.2 generally indicate failed SNP detection, while scores above 0.5 are considered as highly reliable SNP scores. For C × E, 45 SNPs (12%) did not produce a detectable signal or the signal was too low to use it as a reliable marker. Forty-two of these (11%) were also not successful in SH × RH, indicating that these were SNPs for which the assay was not working. Another 7 SNPs did not produce a good result in SH × RH. Of the remaining 339 SNPs in C × E, 173 were not polymorphic between both parents and did not show a segregating polymorphism. For SH × RH, 149 markers were not polymorphic between both parents. Ninety markers were not polymorphic in both the populations.

Six markers in C × E were homozygous in both parents and polymorphic between the parents with a uniform heterozygous offspring (AA × BB → AB). Another set of 6 markers that were homozygous in both parents C and E, polymorphic between parents and segregated according to a 1:2:1 Mendelian ratio (AB × AB → AA, AB and BB). Table 1 summarizes the results of the 384PotSNP array for both populations.

Table 1 Results of 384 PotSNP array performed in two (C × E and SH × RH) independent assays

384 PotSNP array	Mapping
309 out of the 384 are useful markers (80%)	165 markers could be mapped in C × E
42 dropped out in any sample (11%)	186 markers could be mapped in SH × RH
33 were monomorphic in all material[a](9%)	99 markers could be mapped in both populations

[a] Including a set of 220 tetraploid varieties

Genetic mapping of SNP loci

C × E

Out of 165 polymorphic SNP markers, fifty were heterozygous only in parent C; 59 were heterozygous only in parent E and 56 segregating markers were heterozygous in both the parents (AB × AB → AA, AB, BB). These 165 SNP markers were placed on parental genetic linkage maps using Joinmap 4.0 (van Ooijen 2006) together with 93 AFLPs, 45 SSRs, and 24 CAPS markers. Only markers with LOD scores of 3.0 and above were considered. Thirteen and 12 linkage groups were obtained in C and E parental maps, respectively (Fig. 1). Linkage group 6 was divided into two subgroups in the maternal (C parent) map. Nineteen of the 165 SNP markers could not be assigned to a parental linkage group. The C and E genetic parental map span 1012.4 and 774.6 cM, respectively with average distance between adjacent loci 7.2 and 4.5 cM.

SH × RH

A set of 151 AFLP markers from the same four AFLP primer enzyme combinations used in the C × E

Fig. 1 Location of the SNP markers on parental maps C and E. The number on the *left side* is the genetic distance in centiMorgans (cM*) right side* is marker designations. The parental maps were drawn by the MapChart 2.2 program (Voorrips 2002)

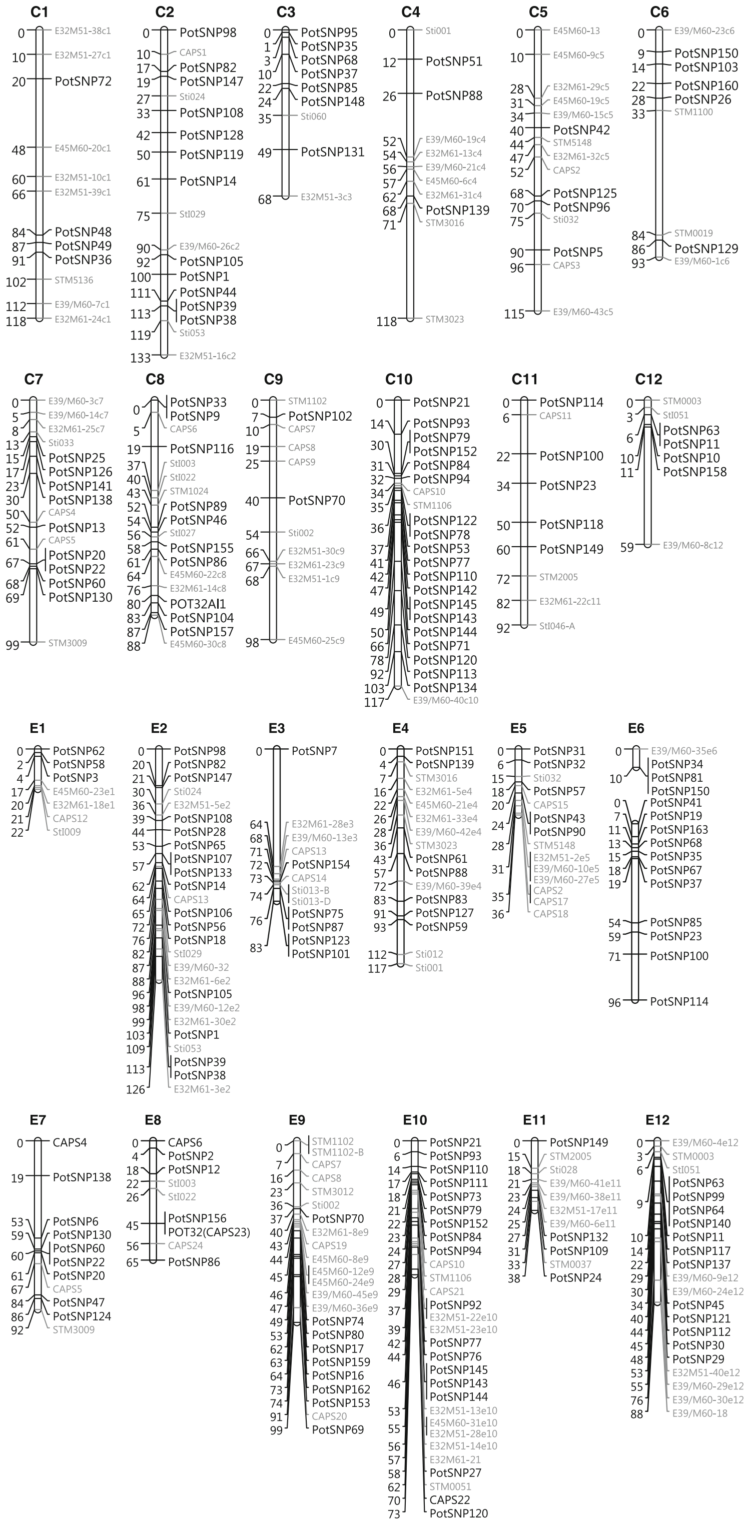


Table 2 Number of markers used for construction of parental maps (C, E and SH, RH) according to marker type

Marker type	Total markers used in construction of parental maps		Markers on the map	
	C and E	SH and RH	C and E	SH and RH
SNP markers	165	186	146	168
AFLP markers	93	151	82	131
SSR markers	45	16	33	16
CAPS markers	24	21	22	21

population were selected from the ~10,000 available AFLP markers in SH × RH. Parental maps of SH and RH were constructed with 15 SSR, 24 CAPS, 151 AFLP and 186 SNP markers using Joinmap 4.0 (van Ooijen 2006). Out of 186 polymorphic SNP markers, 71 were heterozygous in parent SH; 69 were heterozygous in parent RH and 46 segregating markers were heterozygous in both parents (AB × AB → AA, AB, BB). Table 2 lists the markers used for mapping in both populations.

Twelve parent specific linkage groups were obtained for both SH and RH (Fig. 2) The Linkage group RH01 was divided into two subgroups. In SH the length of the linkage groups ranged from 52.6 cM to 115.9 with the average distance between the loci of 4.05 cM. The RH parental map spans 686.7 cM and the average distance between loci is 3.8 cM.

To confirm the SH × RH SNP markers with their bin signatures to calculate error frequency of our mapping results, we compared the marker segregation pattern with the map segregation patterns (bin signatures) and placed these 186 SNP markers in the ultra dense potato map (Van Os et al. 2006). All of the markers were anchored to the bins of the highly saturated parental reference maps and distributed over all linkage groups. Marker order was identical to the map positions in the parental maps constructed in this study. Bin mapping procedure not only gives the bin position but also the goodness of fit to that position. Distance to the bin is a measure for the number of singletons or error in the data. Out of 186 markers, 183 showed error scores below 0.1%, the remaining 3 markers had error scores more than 0.1 and showed LOD scores less than 3.

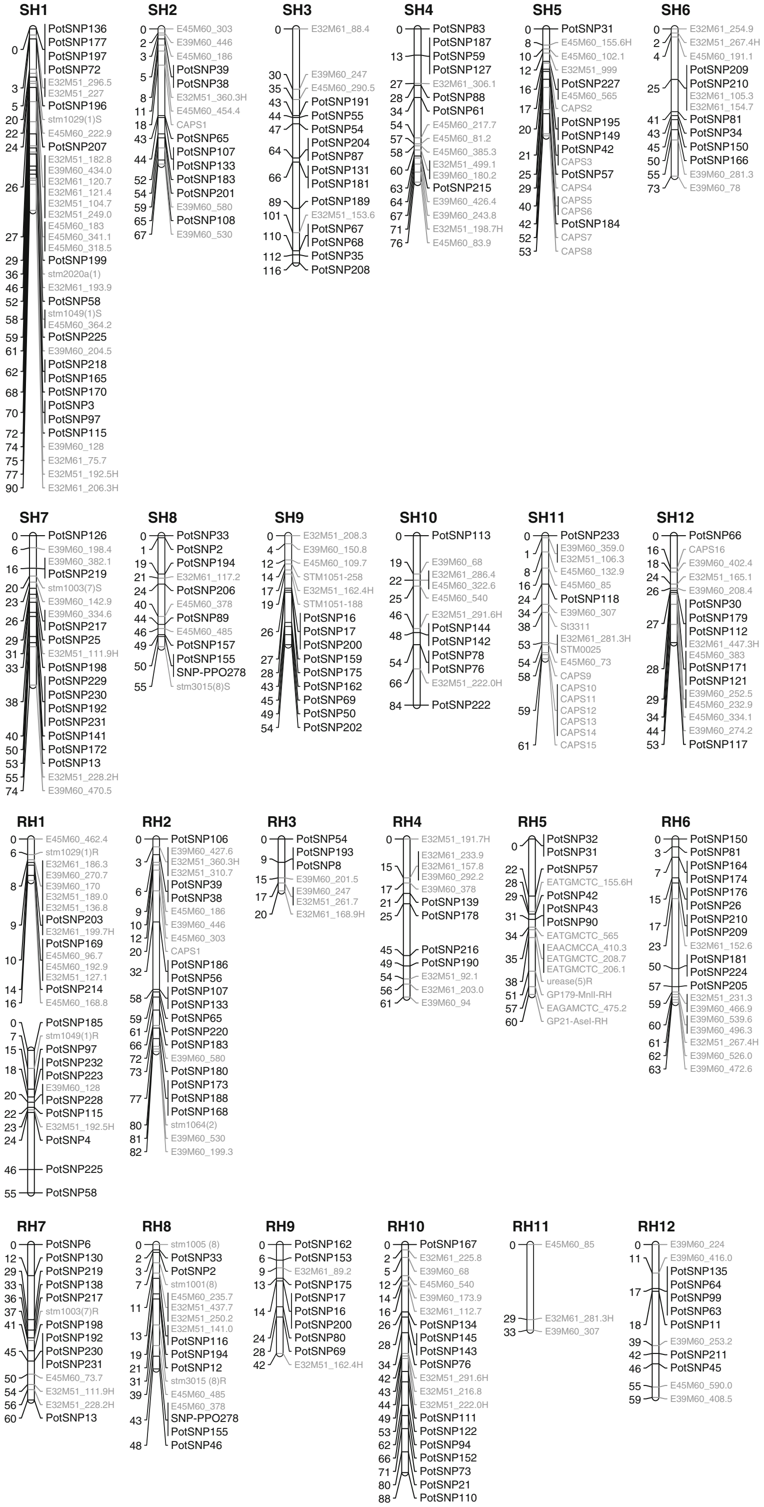
There were 99 markers segregating in both the potato populations. These markers were used to align the C × E with SH × RH maps (Fig. 3), linking the genetic loci of the C × E population are aligned to the ultra dense map and the increasing amount of genomic sequence information of clone RH generated by the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net/).

Discussion

This paper describes the successful development and implementation of a bioinformatics pipeline for the identification of putative SNPs in public EST databases, to convert these SNPs in assays compatible to the Illumina GoldenGate SNP platform, and to map the SNP markers using this genotyping platform. The identification and selection of potato SNPs for the GoldenGate assay results in a score of 89% of working GoldenGate assays, and at least 77% of the full electronic SNP dataset are true SNPs amenable to the GoldenGate genotyping platform.

The first step of this pipeline is the identification of putative SNPs, for which we used QualitySNP. For many SNP assays, including Illumina's GoldenGate assays, the SNP locus needs to be amplified with locus-specific primers that do not amplify any other locus. The paralogous sequences that are placed in separate clusters by QualitySNP may be putative binding targets of the SNP amplification primers designed for a SNP detected in the allelic clusters. The Illumina design tool can eliminate paralogous sequences only when a fully sequenced reference genome is available. However, this is currently not the case with many crop species like potato. To circumvent this problem of paralogous sequences, our pipeline includes a similarity search using the flanking sequences of the SNP to identify clusters with high similarity to the SNP locus. In this study we eliminate SNPs for which the similarity search found other clusters with more than 90% similarity. This implies that SNPs are eliminated that might be suitable for a SNP assay if the primer binding sites

Fig. 2 Location of the SNP markers on parental maps SH and RH. The number on the *left side* is the genetic distance in centiMorgans (cM) *right side* is marker designations. The parental maps were drawn by the MapChart 2.2 program (Voorrips 2002)



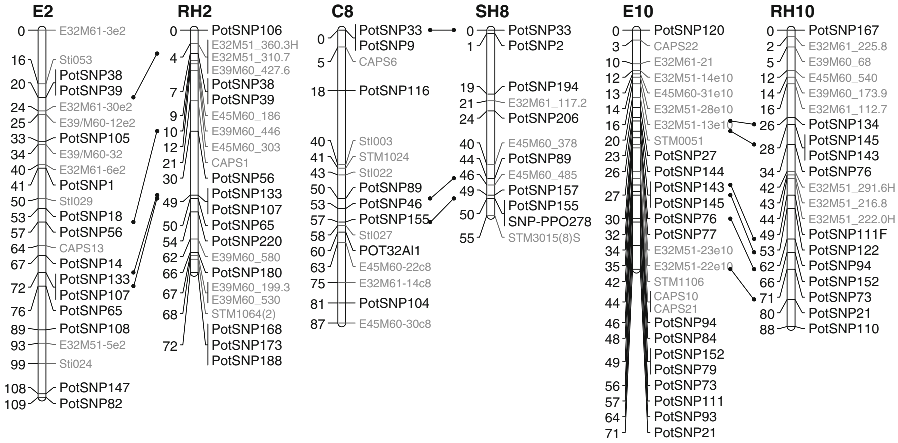


Fig. 3 Alignment of C8 linkage group with the SH8, and E2, E10 with the RH2, RH10 linkage groups, respectively using markers common to both populations. *Left side* number indicates genetic distances in centiMorgan (cM), *right side* marker designations

are carefully chosen. If a SNP in a specific gene is required, or only a limited number of SNPs have been identified, it may be worthwhile to look into the SNP loci for which a similarity conflict has been identified, and design primers for these SNPs. However, we intend this pipeline to be used for high through-put analysis of SNPs from databases to produce a genome-wide SNP array. For efficiency purposes, the SNP loci that might be problematic for GoldenGate assays are therefore eliminated from the list that is used for SNP selection for the SNP array rather than evaluated manually.

Performance of the 384PotSNP array

Of the 384 SNPs that we assembled on a GoldenGate SNP genotyping array and used for genotyping two diploid potato mapping populations only 42 SNPs (11%) failed to produce an interpretable output in two separately performed assays. There are several possible explanations for these SNPs to fail. Firstly, failure may be due to incorrect primer synthesis. In other studies it was observed that 10% of validated SNP loci do not give a result in standard GoldenGate assays, pointing to failure as a result from the assay design (Rostoks et al. 2006, Hyten et al. 2008). Secondly, the SNP frequency in potato is exceptionally high, and therefore SNPs in the primer target sequences could be common. This will likely affect primer annealing as well as signal interpretation. Thirdly, this might have been caused by large introns within the amplicon or introns at the primer sites; the amplification primers were designed on EST sequence information, whereas the SNP assays were performed on genomic DNA. Presence of introns in the SNP amplified region can be detected if genomic sequence information of the SNP locus is available from potato or related species. This was the case for only four of the SNPs with failing assays. Nevertheless, for two out of those four indeed an intron was present within the SNP amplified locus, indicating that in approximately 50% of the failing SNP assays intron presence may be the cause of failure. With the advent of an available genome sequence for potato, the detection of intron–exon boundaries in the vicinity of a SNP will be possible for most if not all SNP loci selected for the array. This information can then be used for primer design, either by filtering out the SNP loci with introns near the SNP site, or by designing primers based on genomic sequence. However, assays with primers based on intron sequences may be more prone to failure than assays with primers in exon sequences, as intron sequences are more variable than exon sequences.

We have shown that the SNP GoldenGate assay linked to the pipeline is a proficient strategy for SNP genotyping in potato, with SNP markers successfully mapped in two potato populations. In total, 342 out of 384 SNP account for the 89% success rate of the combination of QualitySNP with the GoldenGate assay which is comparable to the 90% success rate previously reported in barley (Rostoks et al. 2006) and 89% in soybean (Hyten et al. 2008). However, the barley SNP array from Rostoks et al. (2006) is based on resequencing of selected genes with the parents of a mapping population included, whereas the 384PotSNP array contains SNPs from whatever information available in the EST databases, demonstrating the effectiveness of our pipeline in selecting SNPs that are likely to produce reliable data on the GoldenGate genotyping platform.

Level of polymorphism

The PotSNP array permits the rapid generation of a high number of polymorphic markers. Out of 339 SNPs in CxE (342 in SH × RH), 164 could be mapped (186 in SH × RH). In CxE 161 SNP (155 SH × RH) were monomorphic. The high number of monomorphic SNPs is not surprising; the SNPs were discovered in ESTs from only three varieties namely Shepody, Kennebec and Bintje and the parents in the mapping population are not directly related to any of these varieties. Preliminary data obtained using the potato genotyping array with potato cultivars indicated that 60% of these non-segregating SNP loci were in fact polymorphic in a large cultivar set (data not shown) indicating that these are true SNPs. Part of the remaining monomorphic SNP may also be examples of failure of primer design of the allele-specific primer in the Goldengate assay. This would result in the detection of only a single allele. Six markers in C × E were homozygous in both parents and polymorphic between the parents, with heterozygous offspring. For the population C × E this is a highly unexpected result, as C × E is a backcross population (C is a parent of E). At least one allele of parent E should have been inherited from C, so an AA × BB → AB genetic model for these six loci should not be possible. Neither could this be caused by primer annealing polymorphisms creating a null allele (A0 × B0) as this would result in a segregating rather than a uniform offspring. Similarly, for nine SSRs four alleles were detected in the CE population. These markers could be mapped consistent with an AB × CD → AC, AD, BC, BD genetic model for these 9 loci so this is not an artifact. We currently do not have a satisfying explanation for these observations.

Another set of 6–8 markers that were homozygous in both parents and polymorphic between parents segregated in the population consistent with e.g. an A0 × B0 → A0, AB, B0 genetic model. These were found in both populations, and may represent markers with null alleles in one or both parents. These null alleles may be caused for instance by SNPs or other polymorphisms in one of the alleles that interfere with primer binding and/or subsequent amplification. The results from both runs with the 384PotSNP array were highly comparable, indicating that the reproducibility of the GoldenGate assay is high.

Mapping

The SNP markers in both the potato populations are well distributed over the chromosomes, with minimal clustering. In the parental map of RH linkage group RH01 was divided into two subgroups most likely because the number of markers was not high enough. For some genes more than one SNP within the same open reading frame was selected, for instance SNP38, SNP39 (from ESTs identical to *S. tuberosum* clone transcription factor APFI-like mRNA, TC1649610) and SNP143, 144, 145 (from ESTs identical to *S. tuberosum* StPDC mRNA for pyruvate decarboxylase, TC167230). The SNP markers originating from a single gene all mapped at the same positions. PotSNP156 is located in the coding region of the POT32 gene, and maps at the same position as the CAPS marker POT32A developed for the same gene on chromosome 8 by Werij et al. (2007). For some of the SNP marker loci (genes), the chromosomal location was already known either in potato or in the related species tomato. For each of those markers, the mapping positions agreed with the published mapping positions of the genes. For example PotSNP002 mapped on chromosome 8 in our two populations, and is nearly identical to tomato clone 132639F which also maps on chromosome 8 of tomato. PotSNP009 on chromosome 8 showed a high homology with tomato BiP/grp78 gene, also located on tomato chromosome 8.

In the SH × RH population the positions of 186 markers were confirmed by placing them on the ultra dense bin map (van Os et al. 2006). Most (98.4%) of the polymorphic markers showed error scores below 0.1%. Hence, the Illumina GoldenGate assay is capable of producing high number of error free markers. These SNP markers can be used not only to align CxE map with SH × RH but also as anchors in the potato physical map (van Os et al. 2006).

Perspectives

Our bioinformatics pipeline produced over 7500 SNPs using the EST dataset that are amenable to be assayed on the GoldenGate genotyping platform. Therefore, it is reasonable to expect that more than 7000 remaining SNPs will produce a similar percentage of true and technically scorable SNPs as obtained from the current pilot of 384 SNPs, and are a valuable source for SNP markers in potato populations and cultivars. The EST dataset that was used to mine the SNPs contains sequences from four cultivars: Bintje, Kennebec, Shepody and Kuras. Bintje is an ancient cultivar; Kennebec is a variety from the USA with a pedigree that differs significantly from Bintje. Shepody and Kennebec have pedigree that is partly overlapping. Especially the parents of the C × E population are only distantly related to these cultivars, but still half of the SNP markers generated by the pipeline are polymorphic in the population. This illustrates the wide usability for mapping, association, marker assisted breeding and biodiversity studies of SNP marker assays such as offered by the GoldenGate platform.

Conclusion

The combined use of Quality SNP and Illumina GoldenGate assay in a pipeline has proven to be an efficient tool for the construction of a genetic linkage map. The pipeline produces a large number of co-dominant, polymorphic loci rapidly with a good distribution of markers over the chromosomes. The SNP markers have been selected from EST sequences which were annotated based on sequence similarity to genes with a known function, or in an isolated case based on gene function in potato. The SNP based genetic map therefore allows a candidate gene-based QTL mapping approach. This SNP array offers markers in genes with a variety of putative functions, including biotic and abiotic stress tolerance. Marker assisted breeding with such SNP markers can accelerate the improvement of potato for important traits.

Acknowledgments We would like to thank Dr. Eric van de Weg for helpful discussions, Dr. Jan de Boer for bin mapping and we are grateful to Service XS Leiden, The Netherlands for performing the Illumina GoldenGate assay.

References

Ablett G, Hill H, Henry R (2006) Sequence polymorphism discovery in wheat microsatellite flanking regions using pyrophosphate sequencing. Mol Breeding 17:281–289

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422

Bryan GT, Wu KS, Farrall L, Jia Y, Hershey HP, McAdams SA, Faulk KN, Donaldson GK, Tarchini R, Valent B (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene Pi-ta. Plant Cell 12:2033–2045

Ching A, Caldwell K, Jung M, Dolan M, Smith O, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3:19

Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, Hwang EY, Yi SI, Young ND, Shoemaker RC, Van Tassell CP, Specht JE, Cregan PB (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196

Fan JBOA, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS (2003) Highly parallelSNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78

Feingold S, Lloyd J, Norero N, Bonierbale M, Lorenzen J (2005) Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L.). Theor and Appl Gen 111:456–466

Hanneman RE PS (1967) Crossability of 24-chromosome potato hybrids with 48-chromosome cultivars. Potato Research 10:62–73

Hoskins RA, Phan AC, Naeemuddin M, Mapa FA, Ruddy DA, Ryan JJ, Young LM, Wells T, Kopczynski C, Ellis MC (2001) Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. Genome Res 11:1100–1113

Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor and Appl Gen 116:945–952

Jacobsen E (1978) Die chromosomen verdopplung in der zuchtung dihaploider kartoffeln. PhD thesis, Rheinischen Friedrich Wilhelm Universitat, Bonn. 159 pp

Jacobsen E (1980) Increase of diplandroid formation and seed set in 4× × 2× crosses in potatoes by genetical manipulation of dihaploid and some theoretical consequences. Zeitschrift für Pflanzenzuchtung 85:110–121

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiol 129:440–450

Marth GT (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genet 23:452–456

Masouleh AK, Waters DLE, Reinke RF, Henry RJ (2009) A high-throughput assay for rapid and simultaneous analysis of perfect markers for important quality and agronomic traits in rice using multiplexed MALDI-TOF mass spectrometry. Plant Biotechnol J 7:355–363

McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol 141:26–31

Milbourne D, Meyer RC, Collins AJ, Ramsay LD, Gebhardt C (1998) Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol Gen Genet 259:233–245

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biol 6:R54

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varietes. Proc Natl Acad Sci USA 103:18656–18661

Sachidanandam R, David Weissman SCS, Kakol JerzyM, Lincoln D, The International SNP Map Working Group et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Schlotterer C (2004) The evolution of molecular markers—just a matter of fashion? Nature Reviews Genetics 5:63–69

Shen Y-J, Jiang H, Jin J-P, Zhang Z-B, Xi B, He Y-Y, Wang G, Wang C, Qian L, Li X, Yu Q-B, Liu H-J, Chen D-H, Gao J-H, Huang H, Shi T-L, Yang Z-N (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol 135:1198–1205

Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A (2005) High-throughput SNP genotyping on universal bead arrays. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 573:70–82

Somers DJ, Kirkpatrick R, Moniwa M, Walsh A (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 46:431–437

Syvanen A-C (2005) Toward genome-wide SNP genotyping. Nature Genet 37:S5–S10

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

Van Ooijen JW (2006) JoinMap 4.0, software for the calculation of genetic linkage maps in experimental populations. Plant Research International, Wageningen

van Os HV, Andrzejewski S, Bakker E, Barrena I, Bryan G, Caromel B, Ghareeb B, Isidore E, Jong WD, Koert Pv, Lefebvre V, Milbourne D, Ritter E, Rouppe van der Voort JNAM, Rousselle-Bourgeois F, Vliet Jv, Waugh R, Visser RG, Bakker J, Eck HJV (2006) Construction of a 10,000 marker ultra-dense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genome-wide physical map. Genetics 173:1075–1087

Vision TJ, Brown DG, Shmoys DB, Durrett RT, Tanksley SD (2000) Selective mapping: a strategy for optimizing the construction of high-density linkage maps. Genetics 155:407–420

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucl Acids Res 23:4407–4414

Werij JS, Kloosterman B, Celis-Gamboa C, De Vos CHR, America T, Visser RGF, Bachem CWB (2007) Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. Theor Appl Gen 115:245–252

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L (2003) Database resources of the National Center for Biotechnology. Nucl Acids Res 31:28–33

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134