Potato Research (2014) 57:327–330
DOI 10.1007/s11540-015-9282-8

Possibilities and Challenges of the Potato Genome Sequence

R. G. F. Visser · C. W. B. Bachem · T. Borm · J. de Boer · H. J. van Eck · R. Finkers · G. van der Linden · C. A. Maliepaard · J. G. A. M. L. Uitdewilligen · R. Voorrips · P. Vos · A. M. A. Wolters

Received: 20 December 2014 / Accepted: 29 December 2014 /
Published online: 30 January 2015

Abstract This paper describes the progress that has been made since the draft genome sequence of potato has been obtained and the analyses that need to be done to make further progress. Although sequencing has become less expensive and read lengths have increased, making optimal use of the information obtained is still difficult, certainly in the tetraploid potato crop. Major challenges in potato genomics are standardized genome assembly and haplotype analysis. Sequencing methods need to be improved further to achieve precision breeding. With the current new generation sequencing technology, the focus in potato breeding will shift from phenotype improvement to genotype improvement. In this respect, it is essential to realize that different alleles of the same gene can lead to different phenotypes depending on the genetic background and that there is significant epistatic interaction between different alleles. Genome-wide association studies will gain statistical power when binary single nucleotide polymorphism (SNP) data can be replaced with multi-allelic haplotype data. Binary SNP can be distributed across the many different alleles per locus or may be haplotype-specific, and potentially tag specific alleles which clearly differ in their contribution to a certain trait value. Assembling reads from the same linkage phase proved to allow constructing sufficiently long haplotype tracts to ensure their uniqueness. Combining large phenotyping data sets with modern approaches to sequencing and haplotype analysis and proper software will allow the efficiency of potato breeding to increase.

Keywords Genome sequencing · Haplotype analysis · Single nucleotide polymorphism · *Solanum tuberosum* · Standardized genome assembly

R. G. F. Visser (✉) · C. W. B. Bachem · T. Borm · J. de Boer · H. J. van Eck · R. Finkers · G. van der Linden · C. A. Maliepaard · J. G. A. M. L. Uitdewilligen · R. Voorrips · P. Vos · A. M. A. Wolters
Wageningen UR Plant Breeding, Wageningen University & Research Centre, PO Box 386, 6700 AJ Wageningen, The Netherlands
e-mail: richard.visser@wur.nl

Introduction

Since obtaining the draft genome sequence of potato, developments in sequencing have been plentiful; costs have dropped and read lengths have increased. Important questions are as follows: What can be done with the available information? and What type of data should be gathered further? Major challenges in potato are standardized genome assembly and haplotype discrimination. With the available sequence, a number of things could be done but one of the most important ones—the classification into haplotypes—is still a difficult task, especially in tetraploid clones. The identification and ability to use single nucleotide polymorphism (SNP) and to assign dosage level to the SNP bring haplotype analysis within reach. In spite of these positive developments, it is also clear that sequencing methods which can deliver longer sequence reads in a high throughput manner than the ones currently available will certainly be necessary to be able to achieve precision breeding in potato. For a long time, breeding was largely done by phenotypic improvement and breeding was considered more of an art than a science. Future breeding will be more directed along the genotypic scale, and the chance of developing successful varieties from a breeding program will have to increase. Opportunities for this are abundant, especially in those crops for which sufficient genomic tools have been developed, like in potato. It is clear, however, that the availability and integration of large amounts of data and their use in well-informed selection of crossing parents will lead to superior varieties for specific targets, but only so when this integrated knowledge can be used to its fullest extent.

Until recently, the big challenge was to identify the gene(s) involved in or responsible for desired traits; however, now it has become an even bigger challenge to identify the most important allele(s) of the gene of interest and at the same time to know how it will express in different genetic backgrounds. Not only are we gaining more and more evidence that not every allele of every gene is giving the same end effect (on phenotype) in different genetic backgrounds, it is also clear that next to this, epistatic interaction between different alleles of genes can lead to a different phenotypic outcome.

Methods

The potato genome consists of 12 chromosomes and has a (haploid) length of about 840 Mbp, which makes it a medium-sized plant genome falling within the reach of full sequencing. The sequencing project builds on a potato genomic DNA library of 78,000 BAC clones from diploid genotype RH, which were fingerprinted and aligned into physical map contigs. These BAC contigs were anchored to the Ultra High Density genetic map of the potato, composed of 10,000 unique AFLP^{TM} markers (Van Os et al. 2006). From this integrated genetic-physical map, between 50 and 150 seed BACs were identified for every chromosome. FISH experiments on selected BAC clones confirmed these anchor points. The seed clones provided the starting point for a BAC-by-BAC sequencing strategy while at the same time the strategy was being complemented by whole genome shotgun sequencing approaches using both 454 GS FLX and Illumina GA2 instruments on the RH and a monoploid genotype (DM). Assembly and annotation of the sequence data was done and published (Visser et al. 2009; PGSC 2011; Sharma

et al. 2013). The BAC-by-BAC sequencing of one chromosome (5) of the diploid RH clone (containing two halpotypes) was completed entirely. Furthermore, the sequencing of 800 genes in over 80 tetraploid varieties was undertaken making use of Sure Select technology (Uitdewilligen et al. 2013). Single nucleotide polymorphisms (SNP) were obtained from these sequencing efforts, and SNP were validated for use in marker analysis studies. The detection of SNP marker-trait associations in genome wide association study (GWAS) panels of tetraploid potato, which is much more challenging compared with association studies in diploid species, was undertaken.

Results and Discussion

More and more genome sequences of many important crops become available. The promises of using this type of information to improve and speed up breeding processes are numerous. Major challenges in different crop plants, especially the cross-fertilizing polyploid ones, are genome assembly and haplotype discrimination. Having different genomic tools available (like SNP) makes every crop potentially amenable to marker-assisted selection. Cultivated potato germplasm is characterized by a large number of different alleles, often exceeding 10 alleles per locus. Cultivars are highly heterozygous with over three different alleles per locus. GWAS between marker loci and trait phenotypes have limited power, because binary marker data (0/1) are insufficient to unambiguously follow these many alleles. Some SNP markers, however, uniquely tag a single specific allele, and with an allelic series of such TagSNPs, it should be possible to achieve full classification or haplotyping of potato genotypes at any given locus. This approach for genotyping-by-sequencing is a valid and cost effective alternative for high-density SNP arrays to allow GWAS.

We propose that GWAS will gain statistical power when binary SNP data can be replaced with multi-allelic haplotype data. Binary SNP can be distributed across the many different alleles per locus or may be haplotype-specific, and potentially tag specific alleles, which clearly differ in their contribution to a certain trait value. Haplotype reconstruction based on statistical methods to infer the linkage phase of SNP is prohibitively complicated. Therefore, we tested an approach to make use of the original data. The individual sequence reads that were generated to call the sequence variants also display the linkage phase between SNP occurring on the same (paired-end) read. Assembling reads from the same linkage phase indeed allows constructing sufficiently long haplotype tracts to ensure their uniqueness, i.e., haplotypes are identical by descent and indicative for a breeding history as perceived from the pedigree database. At read depths of at least 80× coverage, the short read lengths (2×100 bp paired-ends) are sufficient to construct haplotypes in SNP dense regions of the genome, but in more conserved (i.e., coding) regions of the genome, the current next generation sequencing (NGS) read length is posing limitations to extend the haplotypes; 135,000 unphased SNP were identified from 800 gene loci in a panel of 83 tetraploid potato cultivars. This dataset was used to try and phase SNP into haplotypes. Despite the fact that potato has a very high SNP density (one in every 16 bp), we found the actual SNP counts on the short NGS fragments to be too low. Dedicated alignment software was written to overcome this problem and still have haplotypes of an appreciable size (~2 kbp in length). Comparison of the NGS-based

haplotypes with previous Sanger sequence-based haplotypes confirmed their accuracy for a number of different genes.

Having sequence data as such is not the solution to all problems. Knowing which genes play a role in particular processes but, even more importantly, which alleles are contributing the largest effect to the trait and which combinations of alleles can be best combined to obtain the desired amount of improvement in a trait are key. Knowing where to find and how to combine the different alleles and traits in crossing programs is a challenge but slowly becoming available. For this, good databases with extensive information about many phenotypes and genotypes are important. Likewise, the availability of (software) tools to query all these kinds of databases and be able to extract the essential information is a major challenge. At Wageningen, we have experience with running projects (like, for example, the Virtual Lab of Plant Breeding (VLPB)) which try to deliver tools and concepts to make the best use of all kinds of available omics data sets and increase the efficiency of current breeding programs.

In the VLPB-I project, a total of 14 sub-projects have been defined that cover areas such as smart visualization of single nuclear polymorphisms (SNP) in large collections of sequenced accessions; convenient visualization of the comparison of SNP from parental lines, offspring and a reference; implementation of authentication methodology to safely access private data within BreeDB (http://www.plantbreeding.wur.nl/UK/software_breedb.html); and the implementation of methodology that estimates associations between high-density genome-wide SNP and phenotypic traits. A further VLPB project is planned, to create a professional ICT production environment for all VLPB tools.

References

PGSC (The Potato Genome Sequencing Consortium) (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, Eguiluz M, Guo X, Guzman F, Hackett CA, Hamilton JP, Li G, Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K, Mejía N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ, Torres Y, Waugh R, Zhang Z, Huang S, Visser RGF, Bachem CWB, Sagredo B, Feingold SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JME, Milbourne D, Martin DMA, Bryan GJ (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 3:2031–2047

Uitdewilligen JGAML, Wolters AMA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8(5):e62355

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P, Lefebvre V, Milbourne D, Ritter E, Rouppe van der Voort JNAM, Rousselle-Bourgeois F, van Vliet J, Waugh R, Visser RGF, Bakker J, van Eck HJ (2006) Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173:1075–1089

Visser RGF, Bachem CWB, de Boer JM, Bryan GJ, Chakrabati SK, Feingold S, Gromadka R, van Ham RCHJ, Huang S, Jacobs JME, Kuznetsov B, de Melo PE, Milbourne D, Orjeda G, Sagredo B, Tang X (2009) Sequencing the potato genome: outline and first results to come from the elucidation of the sequence of the world's third most important food crop. Am Potato J 86:417–429