REVIEW

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

Richard G. F. Visser · Christian W. B. Bachem · Jan M. de Boer · Glenn J. Bryan ·
Swarup K. Chakrabati · Sergio Feingold · Robert Gromadka ·
Roeland C. H. J. van Ham · Sanwen Huang · Jeanne M. E. Jacobs · Boris Kuznetsov ·
Paulo E. de Melo · Dan Milbourne · Gisella Orjeda · Boris Sagredo · Xiaomin Tang

Abstract Potato is a member of the Solanaceae, a plant family that includes several other economically important species, such as tomato, eggplant, petunia, tobacco and pepper. The Potato Genome Sequencing Consortium (PGSC) aims to elucidate the complete genome sequence of potato, the third most important food crop in the world. The PGSC is a collaboration between 13 research groups from China, India, Poland, Russia, the Netherlands, Ireland,

R. G. F. Visser · C. W. B. Bachem (✉) · J. M. de Boer · X. Tang
Wageningen UR Plant Breeding,
Wageningen University & Research Centre,
P.O. Box 386, 6700 AJ Wageningen, The Netherlands
e-mail: Christian.Bachem@wur.nl

G. J. Bryan
Genetics Programme, SCRI,
Dundee DD2 5DA, UK

S. K. Chakrabati
Central Potato Research Institute,
Shimla 171 001 Himachal Pradesh, India

S. Feingold
Laboratorio de Agro-Biotecnología,
EEA Balcarce - INTA,
cc276 (7620) Balcarce, Argentina

R. Gromadka
Institute of Biochemistry and Biophysics
Polish Academy of Sciences,
ul. Pawinskiego 5a,
02-106 Warsaw, Poland

R. C. H. J. van Ham
Applied Bioinformatics, Plant Research
International/Centre for BioSystems Genomics,
WUR,
Droevendaalsesteeg 1,
6708PB Wageningen, The Netherlands

S. Huang
Institute of Vegetables and Flowers,
Chinese Academy of Agricultural Sciences,
No. 12 Zhong Guan Cun Nan Da Jie,
Beijing 100081, China

J. M. E. Jacobs
The New Zealand Institute for Plant & Food Research Ltd,
Private Bag 4704,
Christchurch, New Zealand

B. Kuznetsov
Centre for Bioengineering,
RAS, Oktyabrya 7-1,
Moscow, Russia

P. E. de Melo
Embrapa, Vegetables,
C. Postal 218, 70.359-970 Brasília – DF, Brasil

D. Milbourne
Teagasc, Crops Research Centre,
Oak Park,
Carlow, Ireland

G. Orjeda
Genomics Research Unit,
Universidad Peruana Cayetano Heredia,
Av Honorio Delgado 430, Urb Ingenieria, SMP,
Lima, Peru

B. Sagredo
INIA-Rayentué, Casilla Nº 13, Rengo.
Región Libertador Bernardo O'Higgins,
Rancagua, Chile

Argentina, Brazil, Chile, Peru, USA, New Zealand and the UK. The potato genome consists of 12 chromosomes and has a (haploid) length of approximately 840 million base pairs, making it a medium-sized plant genome. The sequencing project builds on a diploid potato genomic bacterial artificial chromosome (BAC) clone library of 78000 clones, which has been fingerprinted and aligned into ~7000 physical map contigs. In addition, the BAC-ends have been sequenced and are publicly available. Approximately 30000 BACs are anchored to the Ultra High Density genetic map of potato, composed of 10000 unique AFLP^{TM} markers. From this integrated genetic-physical map, between 50 to 150 seed BACs have currently been identified for every chromosome. Fluorescent *in situ* hybridization experiments on selected BAC clones confirm these anchor points. The seed clones provide the starting point for a BAC-by-BAC sequencing strategy. This strategy is being complemented by whole genome shotgun sequencing approaches using both 454 GS FLX and Illumina GA2 instruments. Assembly and annotation of the sequence data will be performed using publicly available and tailor-made tools. The availability of the annotated data will help to characterize germplasm collections based on allelic variance and to assist potato breeders to more fully exploit the genetic potential of potato.

Resumen La papa es un miembro de las Solanaceae, una familia de plantas que incluye varias otras especies económicamente importantes, tales como tomate, berenjena, petunia, tabaco y ají o chili. El consorcio de secuenciación del genoma de la papa (PGSC) tiene por objeto dilucidar la secuencia completa del genoma de la papa, el tercer cultivo alimentario más importante del mundo. El PGSC es una colaboración entre 13 grupos de investigación procedentes de China, India, Polonia, Rusia, los Países Bajos, Irlanda, Argentina, Brasil, Chile, Perú, EE.UU., Nueva Zelanda y el Reino Unido. El genoma de la papa consiste de 12 cromosomas y tiene una longitud (haploide) de aproximadamente 840 millones de pares de bases, por lo que es una planta con un genoma de tamaño mediano. El proyecto de secuenciación se basa en una biblioteca de 78000 clones de cromosoma artificial bacteriano genomico de papa diploide (BAC), del que se ha obtenido la huella genética y alineado en 7000 ~ contigs de mapa físico. Además, los extremos terminales BAC se han secuenciado y están a disposición del público. Aproximadamente 30000 BACS están anclados al mapa genético de ultra alta densidad de la papa, compuesto de 10000 marcadores AFLP^{TM} únicos. De esta mapa genético-físico integrado, entre 50 a 150 semillas BACs han sido identificadas para cada cromosoma. Experimentos de hibridación *in situ* fluorescente en clones BAC selectos confirman estos puntos de anclaje. La clones semilla proveen el punto de partida para la estrategia de secuenciación de BAC a BAC. Esta estrategia se complementa con los enfoques de secuenciación escopeta del genoma completo usando los instrumentos 454 GS FLX e Illumina GA2. El ensamblaje y anotación de los datos de la secuencia será realizados mediante herramientas publicas disponibles y hechas a la medida. La disponibilidad de los datos anotados ayudarán a caracterizar las colecciones de germoplasma basándose en variación alélica y ayudará a los fitomejoradores de papa a explotar más plenamente el potencial genético de la papa.

Introduction

Potato is a member of the *Solanaceae*, a large plant family with more than 3,000 species. The *Solanaceae* family includes several other economically important species such as tomato, eggplant, petunia, tobacco and pepper. Potato is an important global food source. After wheat and rice, potato is the third most important food crop, with a worldwide production of 325 thousand tons in 2007 (FAO Crops statistics database: http://faostat.fao.org/). Optimization of production levels and resistance to biotic and abiotic stresses are key objectives of global potato breeding programs. Root and tuber crops will play an important role in feeding the developing world in the coming decades. The growth rates in production are particularly strong for potato with an annual average increase of 4.5 million tons per year, exceeding those estimated for rice and wheat. Recent increases in Asia have been particularly striking. By 2020, more than two billion people in Asia, Africa and Latin America will depend on these crops for food, feed, or income (Kuang et al. 2005; Song et al. 1998). Current decisions on research investments for root and tuber crops and the strategy chosen for this research will have profound global implications for decades to come. For the developed world, consumer demands require breeders to produce novel cultivars applicable to specific market segments, such as consumption, processing or varieties compliant with "organic" standards. This diversification is also driven by a "whole chain" approach, for example, demands of potato processors directly affect cultivar selection and quality standards in the agriculture sector. For developing countries, breeding efforts should be focused on high yielding and highly nutritious crops in adverse biotic and abiotic conditions. In order to fulfill the above mentioned demands, it is necessary to develop cultivars combining many high performance characteristics. These include traits

such as high yields for different climates, broad-spectrum disease resistance, high quality storage characteristics and applicability for both processing and consumption markets.

The potato has one of the richest genetic resources of any cultivated plant, with about 190 wild tuber-bearing species being recognized in the section *Petota* of the genus *Solanum* (Spooner and Hijmans 2001) as well as in the highly diverse landrace material, for which the taxonomy is currently under revision (Spooner et al. 2007). The tuber-bearing *Solanum* species are very widely distributed in the Americas, from the South Western USA to Southern Chile and Argentina and from sea level to the highlands of the Andes Mountains. Many wild species can be crossed directly with the common potato and moreover, possess a wide range of resistances to pests and diseases, tolerances to frost and drought and many other valuable traits, making them a useful resource for breeding new cultivars.

Despite the importance of the potato, the genetics and inheritance of many important qualitative and quantitative agronomic traits is poorly understood. Likewise, little knowledge is available with respect to compositional and processing traits of the potato tuber. This is mainly due to the tetraploid nature of the genome, the high degree of heterozygosity and the absence of homozygous inbred lines or a collection of genetically well-defined marker stocks. In addition, the frequently observed distorted segregation ratios, probably due to a high genetic load, discourage geneticists to choose potato as a model species for genetic research. Yet, a profound understanding of its genetic composition is a basic requirement for developing more efficient breeding methods. The potato genome sequence will provide a major boost to gaining a better understanding of potato trait biology, underpinning future breeding efforts.

Susceptibility to diseases such as late blight is one of the major causes of loss in production levels. Worldwide, an economic loss on the potato crop of about € 3 billion per year is estimated (Haverkort et al. 2008). Although late blight resistance in temperate conditions and bacterial wilt resistance in the tropics are important traits in potato breeding, these diseases are still largely controlled by either frequent application of fungicides for late blight or practically not controllable in the case of bacterial wilt. It is expected that one of the first benefits of a potato sequence will be a major breakthrough in our ability to isolate, characterize and deploy genes involved in disease resistance. To date, the DNA sequences of only a limited number of disease resistance genes have been isolated and no genes controlling wide-spectrum resistance have yet been definitively identified (Ballvora et al. 2002; Huang et al. 2005; Paal et al. 2004; Song et al. 2003; van der Vossen et al. 2003; van der Vossen et al. 2005; van der Vossen et al. 2000). For bacterial wilt, once temperature and humidity are favorable, there is no practical control available. In addition,

the disease is the major cause of seed tuber losses in the tropics. Different levels of resistance are found in wild relatives, but progress in breeding has been very slow so far (Fock et al. 2000; Kim-Lee et al. 2005; Uhrig et al. 1992).

Regardless of whether marker-assisted breeding or genetic modification approaches are adopted, a fundamental prerequisite for biotechnology-based enhancement of potato varietal improvement is the identification of the genes involved in the target traits and the allelic variation within these genes that results in the phenotypic variation observed for the traits. While there has been some success in achieving this for monogenically inherited traits (primarily the aforementioned disease resistance genes) progress in identifying the genes and alleles underlying traits exhibiting quantitative inheritance has been much slower. Unfortunately, many desirable traits in potato, including almost all tuber quality traits and many desirable forms of horizontal disease resistance, are assumed to be under polygenic control. Genetic mapping in segregating populations and more recently association mapping, have identified potential candidate genes involved in some of these quantitative traits such as disease resistance (reviewed by Gebhardt and Valkonen 2001) and tuber traits (Li et al. 2005; Menendez et al. 2002). While these studies have been made possible by the availability of a large number of Expressed Sequence Tags (ESTs) (Bachem et al. 2000; Rensink et al. 2005; Ronning et al. 2003) and a relatively small number of full length gene sequences available for potato, a major limiting factor to progress has been a lack of a genome sequence resource allowing the positional context of all of the genes in the potato genome to be taken into account. A high quality, well-annotated genome sequence of potato, combined with the mapping techniques described above and the continuing advances in high throughput analyses of the transcriptome, proteome and metabolome promises to radically enhance our ability to identify the desirable allelic variants of genes underlying important quantitative traits in potato. The PGSC seeks to provide such a resource to the potato research and breeding community in the near future, allowing the full potential of biotechnology-based improvement of this important crop plant to be realized.

The Basis for the Potato Genome Sequence Project

The international Potato Genome Sequence Consortium (PGSC) project has its basis in long-standing research on the molecular genetics of potato within the partner organizations, ranging from the construction of genetic linkage maps in diploid and tetraploid potato (Bradshaw et al. 2004; van Eck et al. 1995; van Os et al. 2006) and the use of BAC libraries and map-based gene cloning (Hein et al. 2007; Huang et al. 2005; Song et al. 2003; van der

Vossen et al. 2000), to an integrated physical map currently under construction (Borm 2008).

The framework for assigning sequences to each of the 12 chromosomes of potato is given by the Ultra High Density (UHD) genetic map. This linkage map was constructed in a European Union partnership project and is composed of approximately 10,000 unique AFLP markers. The UHD map was developed using an F_1 mapping population of 130 lines from a cross between the diploid lines SH (SH83-92-488) and RH (RH89-039-16) (van Os et al. 2006). It is by far the most extensive genetic linkage map available in any crop species to date. BAC libraries have been constructed from both parental clones (known as SH and RH) of the UHD map. The RH clone is less heterozygous and the BAC library has a larger average insert size (120 kb) and was therefore chosen for genome-wide physical map construction and genome sequencing. With around 78,000 BACs, the RH BAC library contains approximately 10 genome equivalents of the 840 Mb potato genome (Borm 2008). Sequenced clones from the RH library are publicly available from the company ImaGenes GmbH in Berlin. As an additional resource, the BAC end sequences of the library have been generated by the NSF-funded project on sequencing chromosome 6 (Zhu et al. 2008; http://solanaceae.plantbiology.msu.edu/projects_potato_chr6.php).

Physical Map and Tiling Path Construction

A unique feature of the potato sequencing project is the approach taken in the construction of the physical map (Fig. 1), where AFLP fingerprinting of the RH BAC library has been used to produce a map of contiguous overlapping BAC clones called contigs with the aid of the program FPC (Soderlund et al. 2000; Soderlund et al. 1997). The BAC fingerprint contigs are anchored to the Ultra High Density genetic map using the KeyMaps™ (Jesse et al. 2004) procedure (Fig. 1). In this procedure, DNA pools of the RH BAC library are screened for genetic map markers, this is followed by a identification of the individual BACs containing these markers (Jesse et al. 2004). BAC contigs are thus anchored to the genetic map and provide 'seed' BACs and 'seed' contigs from which to begin sequencing. At present, more than 1600 seed contigs are available across the 12 chromosomes (Fig. 2). On most chromosomes, the seed contigs are well distributed along the euchromatic arms of the genetic map as is visible from the example of a few chromosomes (Fig. 3). In the pericentromeric heterochromatin regions of the genetic map, however, the physical distribution of the anchored contigs remains as yet unresolved.

Fluorescence *in situ* hybridization experiments (FISH, see below in more detail) showed that the chromosome assignments of the seed clones are of high confidence. A minimal tiling path of BAC clones is established from these seeds clones. This is achieved by looking for extension clones, either within the same contig or in a connecting contig that have fingerprint and BAC-end sequence overlaps with the seed clone. The minimal tiling path of the entire potato genome is expected to comprise about 10,000 BAC clones, with an average overlap between the BAC clones of about 10–20%.

FISH Quality Control of the Physical Mapping

Fluorescence *in situ* hybridization (FISH) mapping of potato BAC clones, generates a cytogenetic map that is a valuable complement to the potato genome sequencing project (Iovene et al. 2008). The aim is to determine and verify the positions of BAC clones on the genetic and physical maps and to explore the extent of the euchromatin regions both in potato and the closely related tomato, the genome of which is also being sequenced (http://www.sgn.cornell.edu/about/tomato_sequencing.pl). DAPI staining of pachytene chromosomes shows a clear division between heterochromatic DNA in the pericentromeric region and euchromatic DNA in the distal chromosome arm, as shown for chromosome 1 in Fig. 4a).

Recently, a method for using multi-colour FISH for BAC localization on the pachytene phase of meiosis chromosomes using directly labeled BAC probes has been developed (Tang et al. 2008). Hybridization of repetitive DNA sequences from the BACs was effectively suppressed by adding an excess of unlabeled Cot100 genomic DNA to the hybridization mixture. We have used multi-colour staining (Tang et al. unpublished results) with 158 RH BAC clones in FISH localization experiments on all 12 chromosomes. The results of these experiments show that the physical map positions were almost all exactly as predicted from the AFLP generated ultra dense genetic map and marker anchoring procedures (Fig. 4b and c).

Of the 158 clones that were examined, 141 had FISH positions that were as predicted by the genetic map. Three BAC clones hybridized to positions only a few map units away from their expected marker positions. Ten clones, however. clearly hybridized to chromosomal locations that differed from the genetic-physical map. Eight of these discrepancies were errors with AFLP marker anchoring or other errors in the physical map. The two other discrepancies were due to mistakes in clone culturing and tracking. Four BACs bound to multiple locations, including to heterochromatic regions of other chromosomes, and their FISH positions thus could not be verified. These clones presumably harbored repetitive sequences.
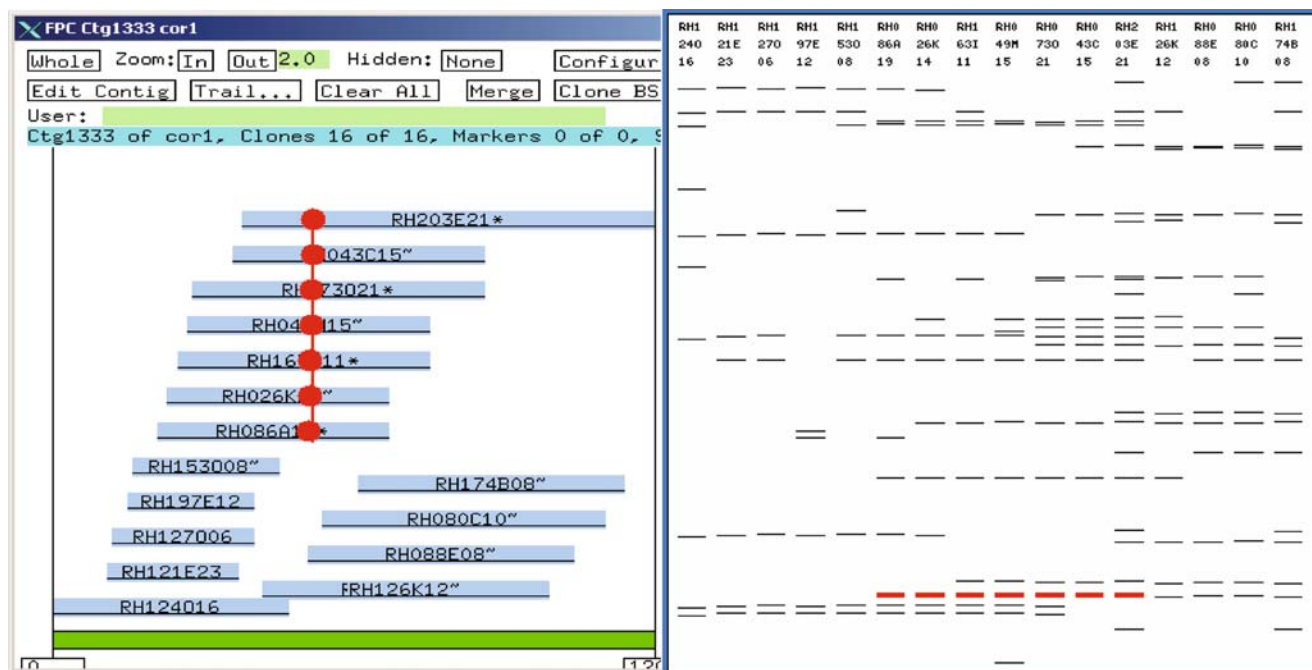
Fig. 1 Physical map construction and anchoring. DNA fingerprint patterns have been generated from the individual BAC clones of the RH BAC (RHPOTKEY) library using the non-selective *Eco*RI/*Mse*I AFLP PCR technique. Based on similarities between these fingerprints (right), the BACs have been stringently aligned into an auto assembly map of 7000 contigs with the program FPC. Each contig (left) represents a set of overlapping clones that originate from the same location of the genome. Genetic map AFLP markers are then identified in the physical map contigs with the KeyMaps procedure (red DNA bands) and anchor them to specific chromosomal positions

From selected anchor points throughout the physical map, a reference set of five landmark FISH BACs has been created for each of the 12 chromosomes, establishing a basic FISH map of the diploid potato (Tang et al. unpublished results). This reference set will be useful for precise chromosomal mapping of unanchored BAC clones, to ensure the resulting genomic maps are highly accurate and integrated with one another.

Fig. 2 Physical map contigs per chromosome. Currently, the AFLP markers from 135 *Eco*RI/*Mse*I primer combinations have been analyzed and 1,600 contigs of the FPC auto-assembly map have been anchored. These anchored contigs are estimated to represent 420 Mb (50%) of the 840 Mb potato genome. Chromosomes 3 and 8 have relatively few seed contigs, because they have fewer genetic markers than the other chromosomes. Chromosome 1 is the largest of the twelve potato chromosomes, which explains its relative high number of anchored contigs

In order to determine the physical size of the euchromatic regions, we are especially interested in locating BACs as close as possible to the euchromatin / heterochromatin borders. As shown in Fig. 4f, the clone RH061A13 is a BAC clone defining the boundary between the euchromatin and pericentromeric heterochromatin of the short arm of chromosome 9.
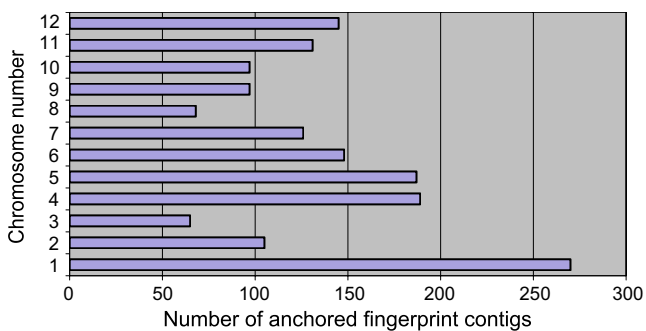
An important test of the quality of a genetic map is to verify that the chromosome ends are fully covered by the markers. We are therefore interested in BAC clones that are anchored to the terminal bins of each of the genetic map linkage groups. For example, the BAC clone RH106H24 contains the AFLP marker EAACMCAA_467. This anchors the BAC to Bin101 at the south end of potato linkage group 1. The FISH signals from RH106H24 partially overlapped with the signals derived from the Arabidopsis telomeric (TTTAGGG) DNA clone pAtT4 on pachytene chromosome 1 (Fig. 4d and e). Thus on chromosome 1 both the physical and genetic maps were shown to extend to the very end of the south arm.

Because of the relatively high degree of DNA sequence similarity among the Solanaceae, the available tomato and potato BACs can be used to study co-linearity between species in the *Solanum* genus. To this end, we have developed a cross-species multi-colour FISH strategy to reveal BAC positions in species related to potato and tomato (Tang et al. 2008).
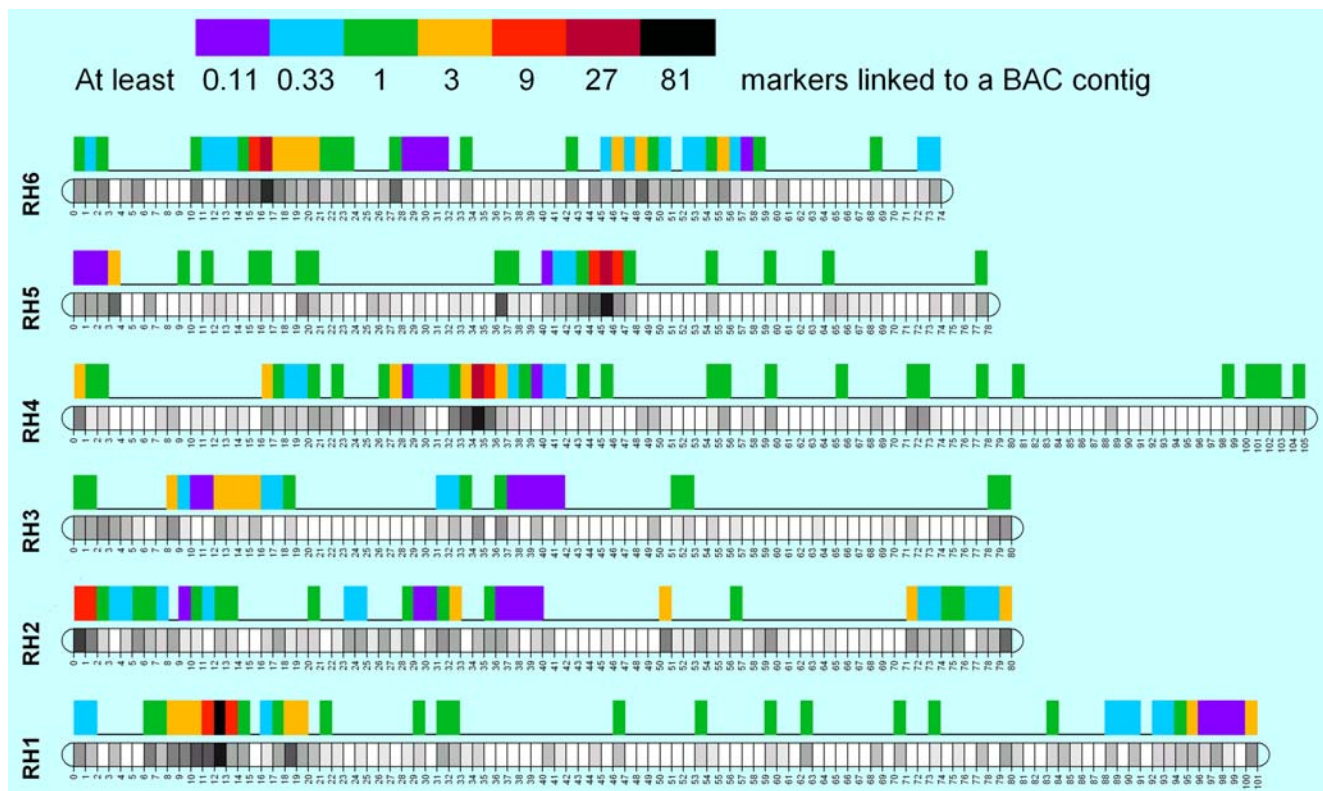
Fig. 3 Distribution of anchored contigs across the RH genetic map (chromosomes RH1 to RH6). In the potato ultra-dense genetic map, each chromosome is divided into an array of bin segments, which are separated from each other by a single crossover event in the F1 mapping population. The genetic map markers were placed in these bins and a grey value indicates the density of (co-segregating) markers per bin (light grey = 1 marker; black = over 100 markers). White bins contain no markers: presumably these represent regions of the genome with a very high recombination frequency or which are homozygous in RH. The density of anchored contigs per bin is indicated with a colour code above the bins and closely follows the marker density in the genetic map. Some of the genetic markers are mapped with less accuracy and are located in an interval of 2 or more bins. Such cases are depicted as bins having 0.5 (or less) anchored contigs, with all occupied bins together adding up to one anchor point. On several chromosomes, the seed contigs are distributed along the entire length of the euchromatic arms of genetic map. The most marker-dense bin of each chromosome contains the centromere and the pericentromeric heterochromatin. Similar to tomato, these pericentromeric bins represent the regions of the genome where genetic recombination is heavily suppressed. Physically, these bins are expected to span the largest portion of the chromosomal sequence

Improvements of the Physical Map

The potato physical map now has around 1,600 seed contigs, which have been anchored with the markers from 135 AFLP primer combinations using the restriction enzyme combination *Eco*RI/*Mse*I. The experience with the current sequencing of Chromosome 5 has been that 187 seed contigs connect to 193 unanchored contigs generated by the FPC program, which brings the total number of anchored BAC clones for Chromosome 5 to 3551. Assuming that contig merging can provide such a doubling of the number of anchored clones also for the other chromosomes, it is anticipated that the current set of AFLP seed contigs will anchor approximately 30,000 BACs. Because of the genome heterozygosity, fingerprint contigs of both haplotypes stay separated and thereby leading to an inflation of the potato fingerprint map. Nevertheless, BAC

sequence information can help to identify pairs of parallel contigs from both haplotypes and thus further improve the quality of the physical map. Still, a substantial fraction of the fingerprint contigs are yet without a chromosome assignment and various strategies are being employed to alleviate this problem. Such an improvement of the physical map is particularly important for chromosomes 3 and 8, where the current number of seed contigs is limited.

As an example, PCR-based molecular markers with known position on the genetic map of potato are being used for marker-assisted selection of BAC clones on chromosome 9, thereby identifying previously unanchored BAC clones and BAC contigs. Marker sequences that give no hits to previously sequenced BACs or to known BAC ends are screened by PCR against pooled DNA of BAC clones. To date eight chromosome 9 specific SSR markers were successfully identified. A further 17 previ-
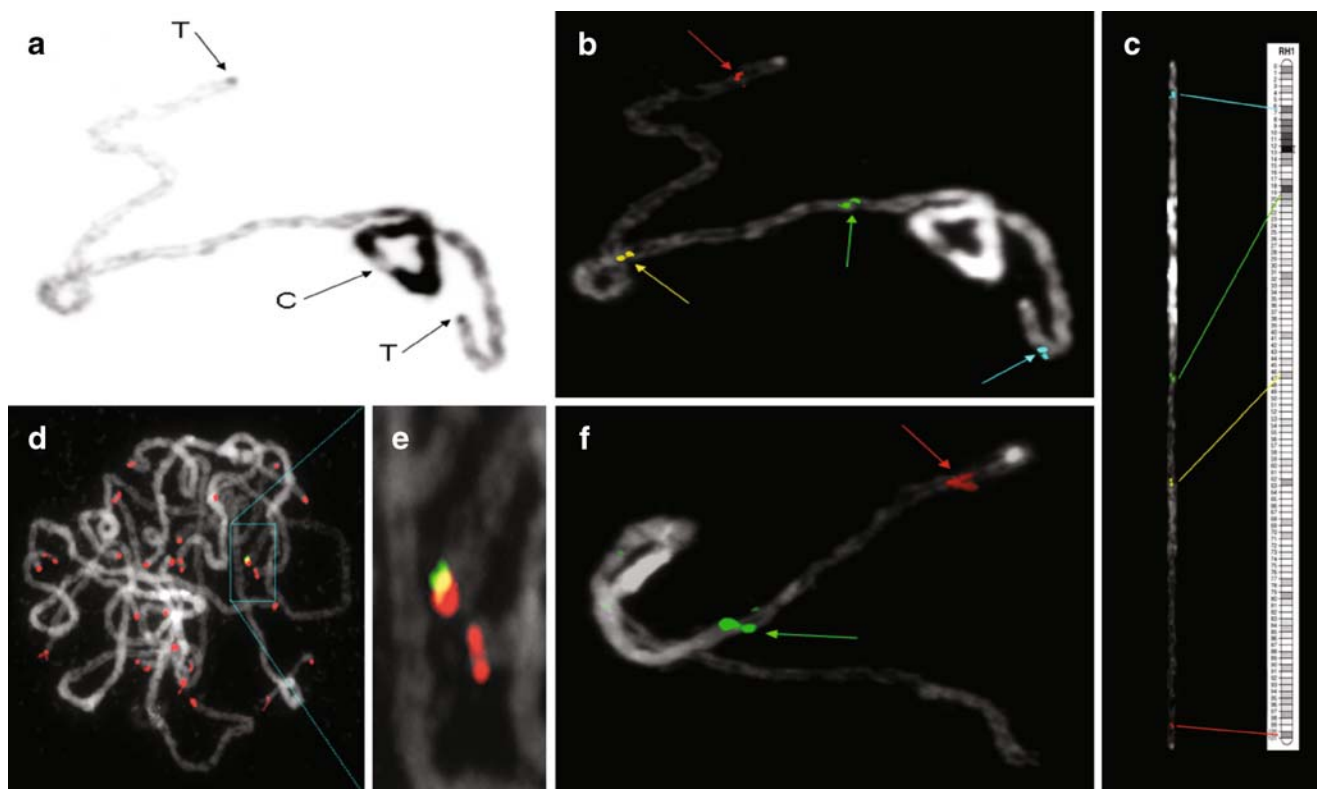
Fig. 4 Diploid potato pachytene chromosomes. **a** Chromosome 1 stained with DAPI. The fluorescent image was reversed so that chromatin appears dark on a light background. The centromere (C) is a constriction within the block of thick, darkly stained pericentromeric heterochromatin. More distally, the arms are thinner and consist of more lightly stained euchromatin. Telomeres (T) are shown as darkly stained spots at the ends of the arms. **b** Multi-colour fluorescence *in situ* hybridization (FISH) of four BAC clones from the euchromatic portion of chromosome 1. **c** Straightened chromosome 1 from B) and the FISH localizations are perfectly correlated with the genetic mapping positions. **d** and **e** One of the most distal BAC clones (RH106H24, green) from chromosome 1 was shown to partially overlap (indicated by yellow colour) with telomeric repeats (pAtT4, red). **f** FISH of two BAC clones from the euchromatic portion of chromosome 9 of which the BAC clone (RH061A13, green signal) borders the pericentromeric heterochromatin

ously unassigned contigs have been anchored. These results indicate that this method can be very useful for chromosome specific BAC selection, allowing the number of seed BACs to be greatly increased and potentially filling gaps between anchored contigs. Due to the high level of synteny between tomato and potato, BLAST searches using tomato marker DNA sequences on the potato BAC end sequences or PCR amplification of tomato molecular markers using potato library BAC pools have also successfully identified additional BAC contigs.

The Execution of the Potato Genome Sequence Project

In order to determine the sequence of potato in a manageable time frame, in 2005 researchers at Wageningen University initiated the establishment of an international consortium capable of sharing the required tasks. The PGSC has brought together a global community to complete the project. Within the PGSC, individual partners concentrate on different chromosomes. Currently the PGSC comprises 13 partners. Two of these are working on two chromosomes each (The Netherlands working on Chromosomes 1 and 5 and China working on chromosomes 10 and 11). India (chromosome 2), USA (chromosome 6), Poland (chromosome 7), New Zealand (chromosome 9) and Russia (chromosome 12) have all taken on a single chromosome. Chromosomes 3 and 4 are being sequenced in small partnerships. The South American nations Argentina, Brazil, Chile and Peru are sequencing chromosome 3 and the UK and Ireland are sequencing chromosome 4. Until recently chromosome 8 was unaccounted for but the Netherlands has now begun to select seed clones for sequencing. The PGSC partners have access to all data on the genetic and physical map of the potato genome and can use it to facilitate their own sequencing efforts as well as to develop tools which may benefit other PGSC members. A web-portal is available giving access to the genetic and physical mapping data (www.potatogenome.net). Furthermore, tools for sequence submission annotation and

genome browsing have been set up. Sequence data are made available in the public databases after a 6 month grace/quality control period. Currently approximately 1600 BACs have been sequenced by the consortium or are in the sequencing pipeline. Of these about 600 BACs are publicly available. The first stage of the BAC-by-BAC strategy adopted by the consortium comprises a six-times coverage sequencing effort of the 10000 BAC clones (120 kb each), which span the potato genome (as described above). This includes a basic annotation of the sequence data, including identification of open reading frames and initial gene assignment by sequence comparison.

Close interaction with other Solanaceae genome projects, such as the tomato genome sequencing project is being maintained throughout the project, as information from each of these projects can be used in a mutually beneficial manner due to the high levels of conserved synteny between the two genomes. The tomato genome sequencing effort is also organized in a consortium with various laboratories from countries around the globe. It originally set out to sequence only the euchromatic regions of the genome (http://www.sgn.cornell.edu/about/tomato_sequencing.pl). Many of the PGSC partners are already actively collaborating with their counterpart groups sequencing the equivalent tomato chromosome and in some case (UK, China) PGSC members are directly involved in the projects to sequence the equivalent chromosomes. The benefits of collaboration between the two projects extend to aspects such as the ordering of Phase 1 sequence contigs of potato BACs by comparison to tomato BACs completed to Phase 3 (Fig. 5a and b) and the use of sequence information from one species to extend BAC contigs and span sequence gaps in the other species (Fig. 5c).

Implementation of Next Generation Sequencing Technologies in the Potato Genome Sequencing Project

With the rapid development of next generation sequencing technologies (NGS) several laboratories involved in the PGSC are implementing Roche 454 GS FLX sequencing platforms for BAC-by-BAC sequencing. This allows the parallel sequencing of several BACs in one sequencing run using BACs tagged with Multiplex Identifiers (Roche Applied Science) increasing the speed and reducing the cost of the sequencing activities. In a few pilot experiments, we were able to sequence 56 BACs of which 24 were previously sequenced using traditional Sanger sequencing. These BACs varied in repeat content and number of contiguous DNA stretches (DNA-contigs) after initial assembly of the Sanger sequences and some exhibited discrepancies between total (Sanger-sequencing-based) DNA-contig sizes and sizes predicted by pulsed field gel electrophoresis (PFGE). Eight BACs were sequenced individually in a first run using an 8 reaction chambers on the GS-FLX sequencer (8-lane gasket) followed by 48 BACs in two consecutive runs, using two reaction chambers (two lane gasket) and 12 sample IG tags. Because of the fact that some of the BACs were sequenced with Sanger-based technology, we were able to identify BACs of which the sequence results were better or worse than with the traditional Sanger method. In a particular example, BAC clone RH047N06 gave nine DNA-contigs with Sanger sequencing and only a small difference between the visualized size from PFGE and total contig length. On the other hand the massively parallel sequencing (Roche 454) gave 35 contigs and a large difference between the predicted size and total contig length. The reverse was also observed, for instance with BAC clone RH047D21 where the Sanger sequencing resulted in 17 contigs with relatively large difference between PFGE size and total contig length. On the other hand the massively parallel sequencing gave only 14 contigs with relatively small difference between PFGE size and total contig length. In Table 1 the results of 48 BACs are compiled and Fig. 6 gives an example of the comparison of 454 versus Sanger assemblies. Overall, it is likely that next generation sequencing technologies (NGS) will increasingly be used for BAC-based sequencing in this and other genome projects, particularly in the light of advances in both read length and the ability to perform paired end sequencing of longer fragments.

In parallel, the PGSC is launching several pilot projects for whole genome shot-gun sequencing (WGS) of the potato genome. The strategy for this is to combine the considerable volume of chromosomally anchored BAC-by-BAC based sequence data with the random short read sequence data that can be generated by both the Illumina GA2 and the Roche GS FLX platforms. Initially, the aim will be to assemble individual chromosome sequences where a, yet to be defined, critical sequence volume has been achieved. The first target for this approach will be Chromosome 5, where 106% of the chromosomal complement of sequence has been generated. When validated the WGS approaches will be extended to include a community sequencing effort by laboratories with appropriate capacity to increase the sequencing depth. It is envisaged that this combined approach will increase the coverage of sequencing, close gaps in genomic regions not well covered by the BAC library and also help in the ordering of fragmented BAC sequences.

The PGSC is also employing a hybrid whole genome shotgun sequencing approach to sequence the *S. tuberosum* group *phureja* doubled monoploid clone, DM1-3 516R44 (CIP801092) as a complement to the *S. tuberosum* RH effort. This line, developed by Richard Veilleux of Virginia Tech (Veilleux et al. 1995), was selected as it provides a
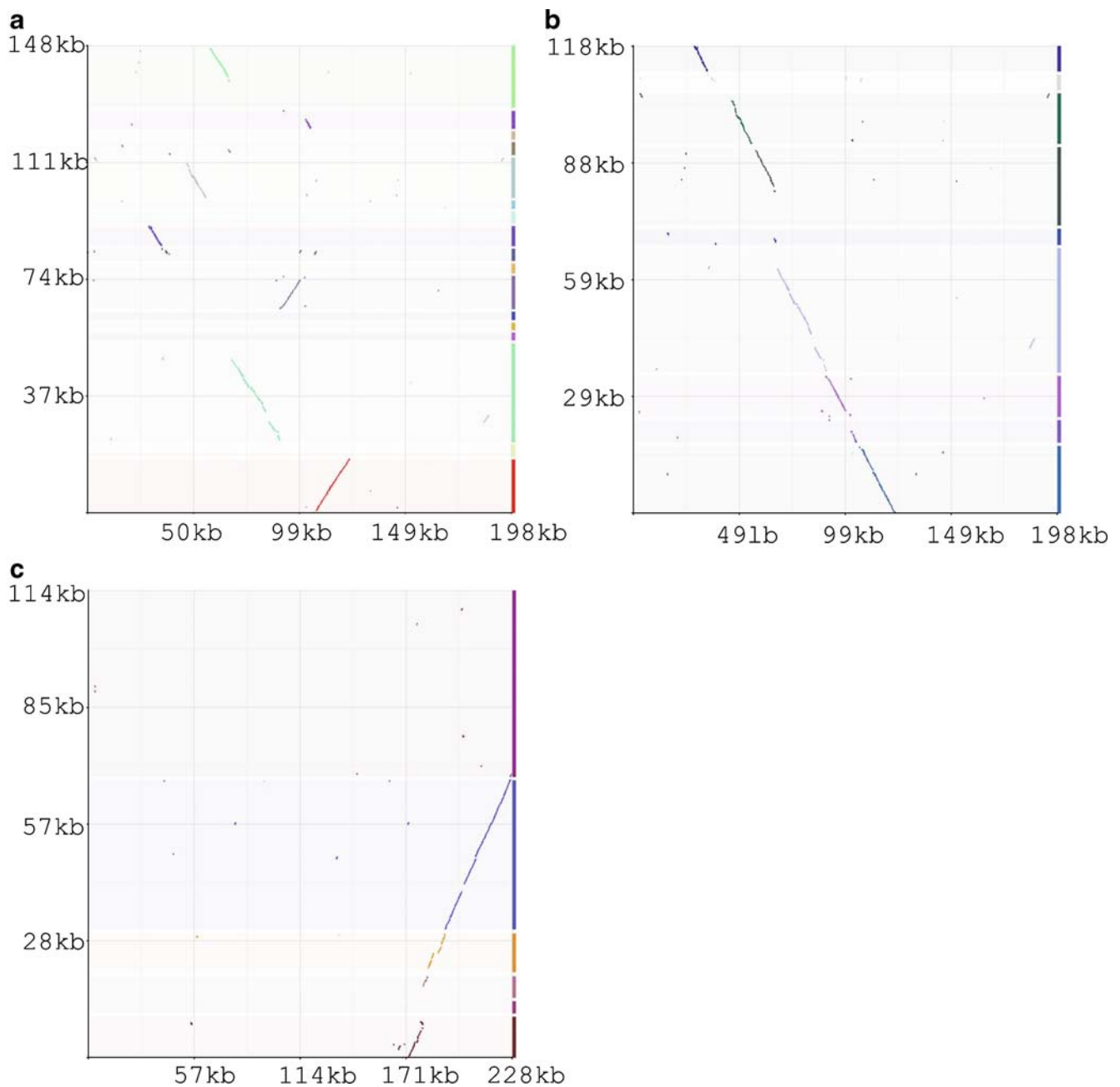
Fig. 5 An illustration of the utility of tomato genome sequence information in potato and vice versa. Potato BAC RH033F05 (AC233612) has been submitted to GenBank as Phase 1 sequence with 17 unordered contigs. A pair wise alignment of the potato BAC **a** to its homoeologous tomato contig, ctg5745, allows 9 of the contigs, representing 80% of the available BAC sequence to be ordered relative to each other **b**. Pair wise alignment of RH91C23(AC233623) with its tomato homeolog ctg916 **c** simultaneously allows the ordering of 6 contigs from this potato BAC relative to each other, while yielding the potential to extend the tomato contig by using the non-overlapping region of the potato BAC to identify overlapping tomato BACs using homology to fosmid and BAC-end sequences. Alignments were performed using the MULAN tool (http://mulan.dcode.org/). Potato contig numbers (Y-axis) represent the order they are found in the submitted sequence, the suffix "R" indicates that the submitted contig has been reverse complemented

completely homozygous line that eliminates the complexity in genome assembly caused by heterozygosity. Three distinct technology platforms (Illumina, Roche and Sanger), will be used to generate a deep whole genome shotgun assembly of this line. A component of this effort will involve anchoring of the scaffolds to the genetic map to ensure the sequence is of high value to breeders along with annotation of the sequence and comparison with the sequenced *S. tuberosum* genome. All data will be made available immediately to the public following quality control.

Table 1 Comparison of assembly statistics of 48 BACs sequenced with both 454 and Sanger technology

Average BAC insert size 137kb	454	Sanger
No. of reads	21871	1337
Number of contigs	16.57	12.44
Contig length	19563	20283
N50 Contig size	39808	70923
Depth (sequence)	35.19	10.25
Read Length	238	750
BAC Coverage Fraction (PFGE)	101.20%	83.78%
BACs with > 20 contigs	28%	16%

Current Status of the Sequencing Project

Our current estimate based on the 840 Mb genome size of completed sequence is about 30%. This data comes from the BAC end sequences (Zhu et al. 2008) and from the approximately 1700 BACs that have been, or are currently being sequenced by the partnership. As mentioned above chromosomes 1 and 5 currently have the highest sequence coverage with 40% and 106%, respectively. The different starting times of the various groups participating in the PGSC have resulted in a large variation of the sequence volume for each chromosome (http://bacregistry.potatogenome.net/pgscreg/overview_chrom_public.py). However, we are confident that progress in 2009 including our WGS will be sufficient to achieve the stated goal of completion of a draft of the complete genome by the end of 2009.

Release and Availability of the Potato Genome Sequence and Other Resources Connected to it

The PGSC comprises a mix of partners from universities and research institutes. At the outset we have set up a general data release policy that requires partners to release sequence data six months after generation. Partners are however at liberty to release their data anytime prior to this date. Accordingly, partners such as the USA and UK who are obliged to submit sequence data as it is generated by their funding authority indeed do so. The system for data submission is that phase 1 sequence data is entered into the PGSC database in Wageningen and is simultaneously submitted to GenBank with (or without) a publication moratorium for a maximum of 6 months. As described below, the sequence data is then annotated and made available to the partnership in a generic genome browser (GGB) and in a sequence registry database. A public version of the GGB is also accessible from the PGSC website (www.potatogenome.net).
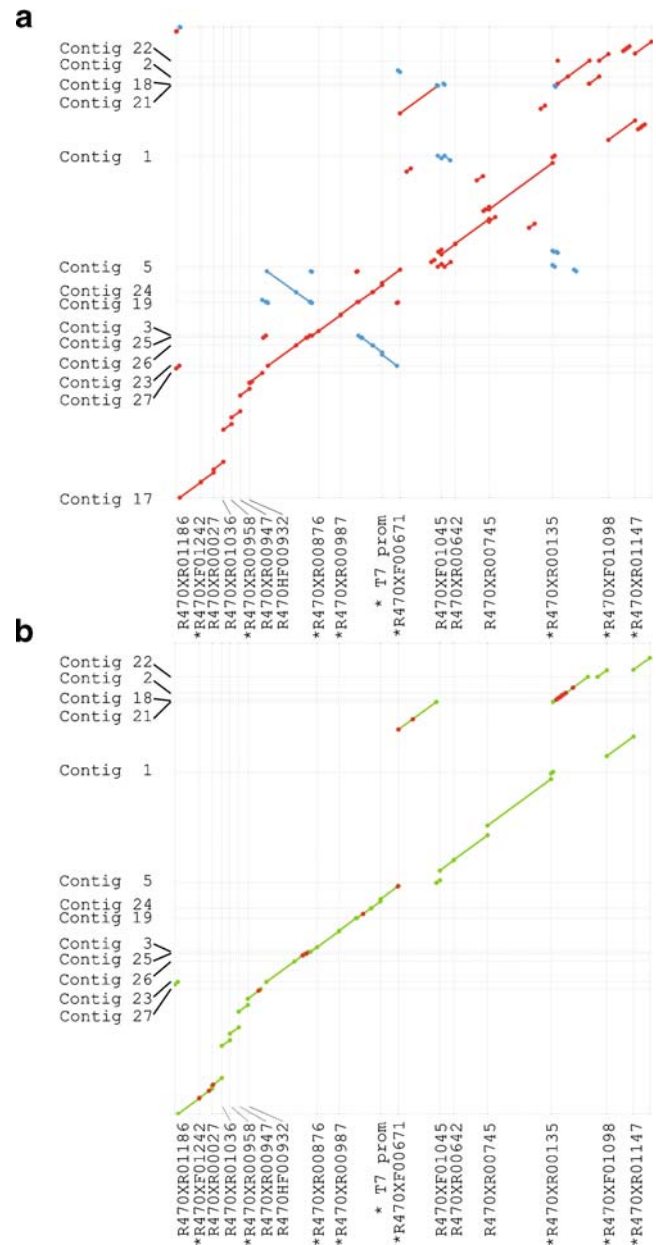

Fig. 6 NUCmer alignment between a Sanger BAC sequence assembly (x-axis) and the corresponding 454 BAC sequence assembly. Aligned segments are represented as lines delimited by dots. **a** Red lines represent matches in the positive strand; blue lines represent matches in the negative strand. Multiple matches on the same x or y position indicate repeats. **b** Identity-filtered NUCmer alignment of the same BAC sequences. The green lines (delimited by green dots) represent matches; red dots represent base differences between the Sanger and 454 assemblies

Potato Genome Sequence Database, Annotation and Assembly

The Potato Genome Sequence Database has been set up and will be maintained by Wageingen University and Research Center. The database contains all raw trace files of each of

the sequenced BAC clones from PGSC-NL. These raw trace files are used for assembly of the BAC sequence into contigs using automated assembly tools such as TOPAAS (Tomato and Potato Assembly Assistance System; Peters et al 2006). TOPAAS is a software package that automates the assembly and scaffolding of contig sequences for low coverage sequencing projects. It uses read pair information, alignments between genomic, EST and BAC end sequences and annotated genes. The application also assists the selection of large genomic insert clones from BAC libraries for walking. TOPAAS is particularly applicable where related or syntenic genomes are sequenced.

The WUR is also annotating all BAC contigs as made available by the partners. The raw data for the BAC sequences from the partners are being submitted to the NCBI's trace file repositories. Annotation of BACs is currently being done using the software package Cyrille2 (Fiers et al. 2008) that has recently been developed. Cyrille2 is an advanced workflow management system geared towards automated annotation and visualization using the Generic Genome Browser GBrowse web interface and database structure. It features a flexible interface to create user defined annotation pipelines. As part of the effort in the USA, all publicly available BAC sequences are annotated for genes, related sequences in other Solanaceae species and similarity to other completed dicotyledonous genomes (Arabidopsis, grapevine and poplar). The genes, their annotation and a GBrowse view of the BACs can be seen at http://solanaceae.plantbiology.msu.edu.

Anticipated Benefits

The members of the PGSC and in due course the entire research community will have access to annotated, genome-anchored sequence data from all participants. Knowledge of the complete genome sequence will provide an invaluable resource for the identification of genes and variant/novel alleles of genes for every trait of interest to potato breeders. This knowledge will revolutionize the way the potato crop can be improved and greatly enhance the development of advanced breeding material and novel cultivars containing important traits. Furthermore, the possibilities to conduct detailed functional genomics and comparative genomics with related Solanaceae, in particular tomato, will open up the possibilities to investigate important traits that differentiate these species and deepen our understanding regarding the evolution of plant species.

An important aspect of the project, in addition to its primary goal, is to foster the development of the capacity of research groups worldwide to exploit the genome sequence of potato. The establishment of a global network of laboratories focusing on potato genomics will help to consolidate the efforts of the individual labs. Academic exchange programs and seminars and workshops, particularly in the area of bioinformatics, have been established to support those laboratories with more restricted experience or limited facilities in the field of genomics research.

The PGSC is conceived as a network whose lifespan is set to extend well beyond the timeframe of the actual sequencing work. The greatest benefits are expected from the post genomic research that will follow from and build upon the sequence data.

mits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Bachem, C., R. van der Hoeven, J. Lucker, R. Oomen, E. Casarini, E. Jacobsen, and R. Visser. 2000. Functional genomic analysis of potato tuber life-cycle. *Potato Research* 43: 297–312.

Ballvora, A., M.R. Ercolano, J. Weiss, K. Meksem, C.A. Bormann, P. Oberhagemann, F. Salamini, and C. Gebhardt. 2002. The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant Journal* 30: 361–371.

Borm, T.J. 2008. *Construction and use of a physical map of potato*, 139. Wageningen, The Netherlands: Wageningen University PhD thesis.

Bradshaw, J.E., B. Pande, G.J. Bryan, C.A. Hackett, K. McLean, H.E. Stewart, and R. Waugh. 2004. Interval mapping of quantitative trait loci for resistance to late blight [Phytophthora infestans (Mont.) de Bary], height and maturity in a tetraploid population of potato (Solanum tuberosum subsp. tuberosum). *Genetics* 168: 983–995.

Fiers, M.W., A. van der Burgt, E. Datema, J.C. de Groot, and R.C. van Ham. 2008. High-throughput bioinformatics with the Cyrille2 pipeline system. *BMC Bioinformatics* 9: 96.

Fock, I.I., C. Collonnier, A. Purwito, J. Luisetti, V.V. Souvannavong, F. Vedel, A. Servaes, A. Ambroise, H. Kodja, G. Ducreux, and D. Sihachakr. 2000. Resistance to bacterial wilt in somatic hybrids between Solanum tuberosum and Solanum phureja. *Plant Science* 160: 165–176.

Gebhardt, C. and J.P. Valkonen. 2001. Organization of genes controlling disease resistance in the potato genome. *Annual review of Phytopathology* 39: 79–102.

Haverkort, A.J., P.M. Boonekamp, R. Hutten, E. Jacobsen, L.A.P. Lotz, G.J.T. Kessel, R.G.F. Visser, and E.A.G. van der Vossen. 2008. Societal costs of late blight in potato and prospects of durable resistance through cisgenic modification. *Potato Research* 51: 47–57.

Hein, I., K. McLean, B. Chalhoub, and G.J. Bryan. 2007. Generation and Screening of a BAC Library from a Diploid Potato Clone to Unravel Durable Late Blight Resistance on Linkage Group IV. *International Journal Plant Genomics* 2007: 51421.

Huang, S., E.A. van der Vossen, H. Kuang, V.G. Vleeshouwers, N. Zhang, T.J. Borm, H.J. van Eck, B. Baker, E. Jacobsen, and R.G. Visser. 2005. Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato. *Plant Journal* 42: 251–261.

Iovene, M., S.M. Wielgus, P.W. Simon, C.R. Buell, and J. Jiang. 2008. Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato. *Genetics* 180: 1307–1317.

Jesse, T.P., L. Wiggers-Perbolte, K. Jansen, J. Buntjer, M. van der Meulen, and R. Sommer. 2004. *Keymaps applications in the construction of high resolution integrated genetic and physical maps*. San Diego, USA: In Plant and Animal Genomes XII Conference.

Kim-Lee, H.Y., J.S. Moon, Y.J. Hong, M.S. Kim, and H.M. Cho. 2005. Bacterial wilt resistance in the progenies of the fusion hybrids between haploid of potato and Solanum commersonii. *American Journal of Potato Research* 82: 129–137.

Kuang, H., F. Wei, M.R. Marano, U. Wirtz, X. Wang, J. Liu, W.P. Shum, J. Zaborsky, L.J. Tallon, W. Rensink, S. Lobst, P. Zhang, C.E. Tornqvist, A. Tek, J. Bamberg, J. Helgeson, W. Fry, F. You, M.C. Luo, J. Jiang, C. Robin Buell, and B. Baker. 2005. The R1 resistance gene cluster contains three groups of independently evolving, type I R1 homologues and shows substantial structural

variation among haplotypes of Solanum demissum. *Plant Journal* 44: 37–51.

Li, L., J. Strahwald, H.R. Hofferbert, J. Lubeck, E. Tacke, H. Junghans, J. Wunder, and C. Gebhardt. 2005. DNA Variation at the Invertase Locus invGE/GF Is Associated With Tuber Quality Traits in Populations of Potato Breeding Clones. *Genetics* 170: 813–821.

Menendez, C.M., E. Ritter, R. Schafer-Pregl, B. Walkemeier, A. Kalde, F. Salamini, and C. Gebhardt. 2002. Cold sweetening in diploid potato: mapping quantitative trait loci and candidate genes. *Genetics* 162: 1423–1434.

Paal, J., H. Henselewski, J. Muth, K. Meksem, C.M. Menendez, F. Salamini, A. Ballvora, and C. Gebhardt. 2004. Molecular cloning of the potato Gro1–4 gene conferring resistance to pathotype Ro1 of the root cyst nematode Globodera rostochiensis, based on a candidate gene approach. *Plant Journal* 38: 285–297.

Peters, S.A., J.C. van Haarst, T.P. Jesse, D. Woltinge, K. Jansen, T. Hesselink, M.J. van Staveren, M.H.C. Abma-Henkens, and R.M. Klein Lankhorst. 2006. TOPAAS, a Tomato and Potato Assembly Assistance System for Selection and Finishing of Bacterial Artificial Chromosomes. *Plant Physiology* 140: 805–817.

Rensink, W.A., S. Iobst, A. Hart, S. Stegalkina, J. Liu, and C.R. Buell. 2005. Gene expression profiling of potato responses to cold, heat, and salt stress. *Functional Integrative Genomics* 5: 201–207.

Ronning, C.M., S.S. Stegalkina, R.A. Ascenzi, O. Bougri, A.L. Hart, T.R. Utterbach, S.E. Vanaken, S.B. Riedmuller, J.A. White, J. Cho, G.M. Pertea, Y. Lee, S. Karamycheva, R. Sultana, J. Tsai, J. Quackenbush, H.M. Griffiths, S. Restrepo, C.D. Smart, W.E. Fry, R. Van Der Hoeven, S. Tanksley, P. Zhang, H. Jin, M.L. Yamamoto, B.J. Baker, and C.R. Buell. 2003. Comparative analyses of potato expressed sequence tag libraries. *Plant Physiology* 131: 419–429.

Soderlund, C., I. Longden, and R. Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences* 13: 523–535.

Soderlund, C., S. Humphray, A. Dunham, and L. French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Research* 10: 1772–1787.

Song, J.Y., D.W. Choi, J.S. Lee, Y.M. Kwon, and S.G. Kim. 1998. Cortical tissue-specific accumulation of the root-specific ns-LTP transcripts in the bean (Phaseolus vulgaris) seedlings. *Plant Molecular Biology* 38: 735–742.

Song, J., J.M. Bradeen, S.K. Naess, J.A. Raasch, S.M. Wielgus, G.T. Haberlach, J. Liu, H. Kuang, S. Austin-Phillips, C.R. Buell, J.P. Helgeson, and J. Jiang. 2003. Gene RB cloned from Solanum bulbocastanum confers broad spectrum resistance to potato late blight. *Proceedings National Academy Sciences United States America* 100: 9128–9133.

Spooner, D.M. and R.J. Hijmans. 2001. Potato systematics and germplasm collecting, 1989–2000. *American Journal of Potato Research* 78: 237–268.

Spooner, D.M., J. Nunez, G. Trujillo, R. Herrera Mdel, F. Guzman, and M. Ghislain. 2007. Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. *Proceedings National Academy Sciences United States America* 104: 19398–19403.

Tang, X., D. Szinay, C. Lang, M.S. Ramanna, E.A. van der Vossen, E. Datema, R.K. Lankhorst, J. de Boer, S.A. Peters, C. Bachem, W. Stiekema, R.G. Visser, H. de Jong, and Y. Bai. 2008. Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics* 180: 1319–1328.

Uhrig, H., C. Gebhardt, E. Tacke, W. Rohde, and F. Salamini. 1992. Recent advances in breeding potatoes for disease resistance. *Netherlands Journal of Plant Pathology* 98: 193–210.

van der Vossen, E.A., J.N. Rouppe van der Voort, K. Kanyuka, A. Bendahmane, H. Sandbrink, D.C. Baulcombe, J. Bakker, W.J. Stiekema, and R.M. Klein-Lankhorst. 2000. Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. *Plant Journal* 23: 567–576.

van der Vossen, E., A. Sikkema, B.L. Hekkert, J. Gros, P. Stevens, M. Muskens, D. Wouters, A. Pereira, W. Stiekema, and S. Allefs. 2003. An ancient R gene from the wild potato species Solanum bulbocastanum confers broad-spectrum resistance to Phytophthora infestans in cultivated potato and tomato. *Plant Journal* 36: 867–882.

van der Vossen, E.A., J. Gros, A. Sikkema, M. Muskens, D. Wouters, P. Wolters, A. Pereira, and S. Allefs. 2005. The Rpi-blb2 gene from Solanum bulbocastanum is an Mi-1 gene homolog conferring broad-spectrum late blight resistance in potato. *Plant Journal* 44: 208–222.

van Eck, H.J., J. Rouppe van der Voort, J. Draaistra, P. van Zandvoort, E. van Enckervort, B. Segers, J. Peleman, E. Jacobsen, J. Helder, and J. Bakker. 1995. The Inheritance and chromosomal localiza-tion of AFLP markers in a non-inbred potato offspring. *Molecular Breeding* 1: 397–410.

van Os, H., S. Andrzejewski, E. Bakker, I. Barrena, G.J. Bryan, B. Caromel, B. Ghareeb, E. Isidore, W. de Jong, P. van Koert, V. Lefebvre, D. Milbourne, E. Ritter, J.N. van der Voort, F. Rousselle-Bourgeois, J. van Vliet, R. Waugh, R.G. Visser, J. Bakker, and H.J. van Eck. 2006. Construction of a 10, 000-marker ultradense genetic recombination map of potato: provid-ing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* 173: 1075–1087.

Veilleux, R.E., L.Y. Shen, and M.M. Paz. 1995. Analysis of the genetic composition of anther-derived potato by randomly amplified polymorphic DNA and simple sequence repeats. *Genome* 38: 1153–1162.

Zhu, W., S. Ouyang, M. Iovene, K. O'Brien, H. Vuong, J. Jiang, and C.R. Buell. 2008. Analysis of 90 Mb of the potato genome reveals conservation of gene structures and order with tomato but divergence in repetitive sequence composition. *BMC Genomics* 9: 286.