High-Throughput Sequencing of the Potato Genome

6

Virupaksh U. Patil, Nitya N. Sharma and Swarup Kumar Chakrabarti

Abstract

Potato is globally the most important food crop, with the potential to be an alternate staple food. It is being cultivated in over 125 countries on all five continents. It is clonally propagated, highly heterozygous, auto-tetraploid, and suffers acute inbreeding depression. Potato is the first crop to be sequenced under the asteroid clade of eudicot plants that represent ~25% of flowering plant species. A total of 96.6 Gb raw data was generated using the next generation sequencing (NGS) technologies, Pyrosequencing and Illumina GAII along with the conventional Sanger sequencing, to complete the genome sequencing. The genome sequencing and its successful alignment of the 827 Mb potato genome into un-gapped super-scaffolds covering 93.9% of the total genome size of 840 Mb were completed by the Potato Genome Sequencing Consortium (PGSC), consisting of 26 institutes belonging to 13 countries. Ninety-nine per cent of the aligned sequence fell into 443 super-scaffolds. It took nearly four years for complete sequencing and assembling of the sequence. A brief discussion of the entire potato genome sequencing using the next generation sequencing technologies is covered in this chapter.

6.1 Introduction

Potato (*Solanum tuberosum* L.) is the world's most important non-grain food crop and is central to global food security. After wheat and rice, potato is the third most important in terms of consumption as food, with a world-wide production of 364.8 million tons in 2012 (FAOSTAT 2013). By 2020, it is estimated that more than two billion people worldwide will depend on potato for food, feed, or income. It is clonally

V. U. Patil · N. N. Sharma · S. K. Chakrabarti (✉)
Central Potato Research Institute, Shimla,
Himachal Pradesh, India
e-mail: Chakrabarti.SK@icar.gov.in

V. U. Patil
e-mail: veerubt@gmail.com

S. K. Chakrabarti et al. (eds.), *The Potato Genome*,
Compendium of Plant Genomes, https://doi.org/10.1007/978-3-319-66135-3_6

propagated, highly heterozygous, auto-tetraploid, and suffers acute inbreeding depression. Optimization of production levels and resistance to biotic and abiotic stresses are key objectives of global potato breeding programmes. The need to develop a high quality, well-annotated genome sequence of potato, combined with established mapping techniques to radically enhance our ability to identify the desirable allelic variants of genes underlying important quantitative traits in potato, led to the foundation of The Potato Genome Sequencing Consortium (PGSC)—a collaboration of 13 countries Argentina, Brazil, China, Chile, India, Ireland, The Netherlands, New Zealand, Peru, Poland, Russia, the United Kingdom and the United States. Potato genetics is complicated by its polyploid genome and many important qualitative and quantitative agronomic traits are poorly understood. An understanding of its genetic composition is a basic requirement to develop 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.

Potato belongs to the asterid clade of eudicot plants that represents ~25% of flowering plant species and the complete genome sequence information of any species in this group was not available. Solanaceae, the family of potato, includes several other economically important species, such as tomato, eggplant, petunia, tobacco and pepper. Worldwide, an economic loss on the potato crop of about €3 billion per year is estimated from diseases such as late blight. These diseases are still largely controlled by frequent application of fungicides. It is expected that one of the first benefits of a potato sequence will be a major breakthrough in our ability to characterize and select genes involved in disease resistance.

6.2 Sequencing Technologies

The development of DNA sequencing strategies has been a high priority in genomics research since the unearthing of the structure of DNA and the basic molecular mechanisms of heredity. However, it was not until the works by Maxam and Gilbert (1997) and Sanger et al. (1977), that the first practical sequencing methods were developed and implemented on a large scale. These two landmark researches were responsible for the introduction of first automated DNA sequencers led by Caltech (Smith et al. 1986) and subsequently commercialized by Applied Biosystems (ABI), European Molecular Biology Laboratory (EMBL), Pharmacia-Amersham and General Electric (GE) healthcare. These are now categorized as first-generation sequencing technologies. Despite their popularity as a 'gold standard' among the research community, these suffer certain limitations such as limited length of DNA sequenced, biological biasness, higher amount of sample required, lower number of samples that can be analysed and most important is the higher cost of sequencing. With the advances made in the field of micro-fluidics, imaging power, detection power and computational tools, unconventional sequencing technologies with increased throughput and lower sequencing cost are continuously emerging. The completion of the first human genome drafts (Yamey 2000) was just the start of the modern DNA sequencing era, which resulted in further invention, improved development towards new advanced strategies of high-throughput DNA sequencing, which were collectively called the 'High-Throughput Next Generation Sequencing' (HT-NGS) (Varshney et al. 2009). The first of these NGS technologies was pyrosequencing, which was followed by Illumina and SOLiD (Sequencing by Oligo Ligation and Detection). These HT-NGS technologies have the capacity to produce 100 times more data as compared to the first-generation sequencers and a comparison of features of these three technologies is given in Table 6.1. The horizons and expectations have broadened due to the technological advances in the field of genomics, especially the HT-NGS and its wide range of applications such as: chromatin immune-precipitation coupled to DNA microarray (ChIP-chip) or sequencing (ChIP-seq), RNA sequencing (RNA-seq), whole genome genotyping, de novo assembling and

Table 6.1 Special features of NGS technologies

NGS technologies	Approach	Read length (bp)	Bp per run	Quality	Cost per Mb ($)	Sources of error
Roche 454 Titanium	Pyrosequencing	400–800	800–1000 Mb	10^{-4}–10^{5}	45.00	Amplification, mixed beads, intensity thresholding, homoplymers, neighbour interference
Illumina GAII	Sequencing by synthesis with reversible terminators	100	600 Gb	10^{-2}–10^{-3}	5.97	Amplification, mixed clusters/neighbour, interference, phasing, base labelling
SOLiD	Massively parallel sequencing by ligation	50	1–4 Gb	10^{-2}–10^{-3}	5.81	Amplification, mixed beads, signal decline, neighbour interference

re-assembling of genome, genome-wide structural variation, mutation detection and carrier screening, DNA library preparation, paired ends and genomic captures, and the sequencing of the mitochondrial and chloroplast genome. Besides the advances in sequencing techniques, the past decade will be remembered as the decade of the genome research. Since the publications of the first composite genomes of humans (Venter et al. 2001), many draft genomes from many plants and animal species have been published (www.ensembl.org/info/about/species.html), including the potato genome. For the genome sequencing of potato, three main sequencing technologies were used, namely, Illumina, Pyrosequencing and the Sanger sequencing, which are discussed in detail by Xu et al. (2011).

6.2.1 Roche/454 FLX Pyrosequencer

The 454 sequencing technology (http://www.454.com) was the first of the NGS, derived from a technical combination of pyrosequencing and emulsion PCR. The basis of this technology was sequencing by synthesis (Melamede 1985), a different approach to DNA sequencing by pyrophosphate detection was also reported (Hyman 1988). A team led by Nyren in 1993 came out with a sequencing approach based on chemi-luminescent detection of pyrophosphate released during deoxynucleotide triphosphate (dNTP) incorporation (Nyren et al. 1993). Later, upgrading of a technique by Ronaghi and his co-workers laid the foundation for the commercial development of pyrosequencing at the Royal Institute of Technology, Stockholm in 1996 (Ronaghi et al. 1996). 454 Life Sciences, founded by Jonathan Rothberg in 2000, launched the first commercially available NGS platform called GS 20 in 2005. In the same year, Margulies and colleagues for 454 Life Sciences sequenced the whole genome of *Mycoplasma genitalia* at 96% coverage and 99.96% accuracy in a single run using GS 20 (Margulies et al. 2005). The technology has continuously been upgraded several times into a routine functioning method. The first major technological improvement was the replacement of dATP with that of dATPαS (Ronaghi et al. 1996), followed by the introduction of light phase pyrosequencing and the addition of ssDNA-binding proteins to pyrosequencing (Ronaghi et al. 2000). In 2007, Roche introduced a newer version as GS FLX with the same sequencing chemistry as GS 20 with a unique flowcell referred to as a ‘picotiter plate’ (PTP). An advanced version of the instrument with a PTP plate is GS FLX comprising 3.4 × 10^{6} separate sequencing reaction wells, allowing hundreds of thousands of sequencing reactions to be carried out in parallel and in a massive high-throughput way.

Pyrosequencing is basically a two-stage approach. First, single-stranded DNA is fractionated into smaller fragments (300–1000 bp), polished (made to have a blunt end), and short oligo adapters having a 5′ biotin tag are ligated to the fragments. These adapters provide the priming sequence for the attachment; amplification as well as sequencing the fragment. DNA fragments to be sequenced are then individually immobilized onto streptividin decorated beads which are amplified by the PCR in the water-oil emulsion droplets. These droplets act as individual amplification reactors producing manifold replicas ($\sim 10^7$) of the same DNA sequence on each bead. Template single-stranded DNA is hybridized to a sequencing primer and loaded on to the PTP plate along with DNA polymerase, ATP sulfurylase (a recombinant version from *Saccharomyces cerevisiae*), luciferase (from the American firefly *Photinus pyralis*) (Ronaghi et al. 1998), the nucleotide-degrading enzyme Apyrase (from potato), along with the substrates adenosine 5′ phosphosulfate (APS) and luciferin. One of the four dNTPs is added and, if complementary, DNA polymerase incorporates onto the template accompanied by the release of pyrophosphate (PPi) equal to the molarity of the incorporated nucleotide. This PPi released is quantitatively converted into adenosine tri phosphate (ATP) in the presence of APS. The ATP acts as a fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates light in a comparative amount to the ATP produced. Unincorporated nucleotides and ATPs are continuously washed away by the apyrase and the next reaction starts with another nucleotide addition cycle. One picomole of DNA in a pyrosequencing reaction yields 6×10^9 photons at a wavelength of 560 nm, which is easily detected by a 16 mega pixel CCD camera maintained at −24 °C for its higher resolution and performance. The sequence of DNA is yielded in the form of a 'pyrogram', which corresponds to the order of nucleotides that has been incorporated. The current 454 instrument, the GS FLX + produces an average read length of approximately 1000 bp and throughput of approximately 800 Mb to 1 Gb of high quality sequence data per 7–8 h run (www.454.com).

6.2.2 The Illumina Genome Analyzer

In 1997, British chemists Shankar Balasubramanian and David Klenerman conceptualized an approach for sequencing single DNA molecules attached to microspheres. They funded Solexa in 1998; however, their goal of sequencing single DNA molecules was not fulfilled. The idea was then shifted towards sequencing clonally amplified templates. The year 2006 marks the commercial launch of the first 'short read' sequencing platform *Solexa Genome Analyzer*. The idea was based on *sequencing by synthesis*, one of the high throughput DNA sequencing in NGS. The templet DNA sample is fractionated to the average size ~800 bp. The fragmented DNA ends are repaired; 5′ end phosphorylated while a 3′ poly A tail is added. Repair of DNA is carried out using T_4 DNA Polymerase (digests 3′ protruding ends), Klenow DNA polymerase (extension of 3′ recessive ends) and T_4 PNK (phosphorylates 5′ ends and dephosphorylates 3′ ends). Like 454/Roche, Illumina sequencing also requires the template sequence to be converted to a special sequencing library which ensures the immobilization and amplification for sequencing (Fedurco et al. 2006). Therefore, two unique folked adaptors (adaptor oligonucleotides are complementary to flow cell anchors) are added at the 5′ and 3′ ends of the DNA fragment. The prepared samples are immobilized on an 8-channelled flowcell surface, allowing bridge amplification. Hybridization of the library fragments and the adapter with that of flow cell occurs by active heating and cooling stages. Subsequently, reactants and an isothermal polymerase are incubated to amplify the fragment in a discrete area 'cluster' on a flow cell surface (for animation: http://www.illumina.com/) to form small clusters of single-stranded fragments called 'bridge amplification'. Clusters are formed spontaneously due to the fact that the newly produced copies of the fragment get attached in close proximity to the original fragment. After the bridge amplification is complete, densely packed clusters of fragments have formed, and each cluster consists of many copies of the same fragment, which begins the sequencing by synthesis step. For single-strand sequencing of forward strands, clusters are denatured, chemically cleaved and washed. Sequencing of a forward strand starts with the hybridization of sequencing primer complementary to the adapter sequence followed by the addition of DNA polymerase and a mixture of four differently coloured fluorescent dye terminator nucleotides. All four nucleotides are modified with a distinct fluorochrome, and the reversible terminator group attached at its 3′ hydroxyl group is chemically blocked, so that when one nucleotide is incorporated, replication stops. This ensures the uniqueness of each event. DNA polymerase incorporates the appropriate nucleotide and unused nucleotides are washed away. After every incorporation cycle, the imaging step occurs to determine each incorporated nucleotide followed by the chemical cleavage step which removes the fluorescent nucleotide and unblocks the 3′ end with the help of reducing agent tris (2-carboxymethyl phosphine) for the next sequencing cycle. The process of adding nucleotides, imaging and removing the terminator is called a cycle. The sequencing run requires 2–8 days with 50×10^6 clusters per flow cell to generate read lengths of 35–75 bases. The system generates overall sequencing output of 2–15 Gb per run. The latest technology 'Hi-seq 2500' produces around 600 Gb throughput per 11-day run with dual flow cell and another higher version of the same MiSeq® system with much higher throughput and quality is almost ready to be released (http://www.illumina.com/).

Recently single molecule-based sequencing technologies have hit the market, which are collectively called the Next-Next Generation Sequencing (NNGS) or Third Generation Sequencing (TGS) technologies. Pacific Biosciences Inc. was the first to introduce the NNGS in the global market. The NNGS technologies are said to be more efficient compared to NGS in terms of throughput, cost of sequencing and time consumption, but their utility and performance are yet to be proved as a real advance of technologies over NGS. These never-ending advances in sequencing technologies provide opportunities to target not only the model plant species with small genome sizes, but many cultivated and other economically important plant species like potato for sequencing, identifying millions of novel markers, agronomically important genes, knowledge of which can be directly translated into crop improvement.

6.3 Sequencing the Potato Genome

The potato has one of the richest genetic resources of any cultivated plant (Spooner and Hijmans 2001). 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. Many wild species can be crossed directly with the common potato and, moreover, possess a wide range of resistances to pests and diseases, tolerance to frost and drought and many other valuable traits, making them a useful resource for breeding new cultivars. Outside of its natural range in South America, the cultivated potato is considered to have a narrow genetic base, resulting originally from limited germplasm introductions to Europe. Most potato cultivars are auto-tetraploid ($2n = 4x = 48$), highly heterozygous, suffer acute inbreeding depression, and are susceptible to many devastating pests and pathogens, as exemplified by the Irish potato famine in the mid-nineteenth century. Together, these attributes present a significant barrier to potato improvement using classical breeding approaches. The knowledge of the genome sequence in potato facilitates advanced breeding targeting many important agronomic traits.

To overcome the key issue of heterozygosity to generate a high-quality draft potato genome sequence, a unique homozygous form of potato called a doubled monoploid is derived using classical tissue culture techniques (Paz and Veilleux 1999). The draft genome sequence from this genotype, *S. tuberosum* group *Phureja* DM1-3 516 R44 (hereafter referred to as DM), was used to integrate sequence data from a heterozygous diploid breeding line, *S. tuberosum* group *Tuberosum* RH89-039-16 (hereafter referred to as RH). These two genotypes represent a sample of potato genomic diversity; DM with its fingerling (elongated) tubers was derived from a primitive South American cultivar whereas RH more closely resembles commercially cultivated tetraploid potato. The combined data resources, allied to deep transcriptome sequence from genotypes, explored the potato genome structure and organization, as well as key aspects of the biology and evolution of this important crop.

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 PGSC originally started out with sequencing RH. This part of the project builds on a diploid potato genomic bacterial artificial chromosome (BAC) clone library of 78,000 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 30,000 BACs are anchored to the Ultra High Density genetic map of potato, composed of 10,000 unique AFLP markers. From this integrated genetic-physical map, between 50 and 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 have been carried out by the researchers 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. Sequencing of DM was also started because the overall progress in RH was slow. The heterozygosity of RH has limited the progress of physical mapping and will complicate the assembly of the genome sequence. Whole-genome shotgun sequencing of DM1-3 516R44 (CIP801092), a doubled monoploid potato clone, is expected to eliminate the complexity in assembly. In 2011, the findings of genome and transcriptome sequencing of DM and RH was published in *Nature* in 'Genome sequence and analysis of the tuber crop potato' (Xu et al. 2011). In the later sections we will discuss the strategies and major conclusions of this project.

6.3.1 DM Whole-Genome Shotgun Sequencing and Assembly

The nuclear and organellar genomes of DM were sequenced using a whole-genome shotgun sequencing (WGS) approach. In total, 96.6 Gb of raw sequence data was generated from two next-generation sequencing (NGS) platforms, Illumina Genome Analyser and Roche Pyrosequencing, as well as the conventional Sanger sequencing technologies. The genome was assembled using SOAPdenovo, resulting in a final assembly of 727 Mb, of which 93.9% is non-gapped sequence. Ninety per cent of the assembly falls into 443 superscaffolds larger than 349 kb. The 17-nucleotide depth distribution suggested a genome size of 844 Mb, consistent with estimates from flow cytometry. Analysis of the DM scaffolds indicated 62.2% repetitive content in the assembled section of the DM genome, less than the 74.8% estimated from bacterial artificial chromosome (BAC) and fosmid end sequences, indicating that much of the unassembled genome is composed of repetitive sequences.

Libraries from DM genomic DNA were constructed for Illumina Genome Analyser II (paired end 200 and 500 bp; mate pair 2, 5 and 10 kb insert size) and Roche 454 platforms (8 and 20 kb) for sequencing using standard protocols. A BAC library and three fosmid libraries were end-sequenced using the Sanger platform. For the Illumina GAII platform, 70.6 Gb of 37–73 bp paired-end reads from 16 libraries with insert lengths of 200–811 bp was generated. The mate-pair libraries (2, 5 and 10 kb insert size) were used to generate 18.7 Gb. In total, 7.2 Gb of 454 single-end data were generated and applied to gap filling to improve the assembly, of which 4.7 Gb (12,594,513 reads) were incorporated into the final assembly. For the 8 and 20 kb 454 paired-end reads, representing 0.7 and 1.0 Gb of raw data respectively, 90.7 Mb (511,254 reads) and 211 Mb (1,525,992 reads), respectively, were incorporated into the final assembly.

A high-quality potato genome was constructed using the short read assembly software SOAPdenovo (Version 1.014). The 69.4 Gb data of GAII paired-end short reads were first assembled into contigs, which were sequence assemblies without gaps composed of overlapping reads. To increase the assembly accuracy, only 78.3% of the reads with high quality were considered. Then contigs were further linked into scaffolds by paired-end relationships (~300 to ~550 bp insert size), mate-pair reads (2 to approx. 10 kb), fosmid ends (~40 kb, 90,407 pairs of end sequences) and BAC ends (~100 kb, 71,375 pairs of end sequences). Then filled gaps with the entire short-read data were generated using Illumina GAII reads. The primary contig N_{50} size (the contig length such that using equal or longer contigs produces half of the bases of the assembled genome) was 697 bp and increased to 1318 kb after gap-filling. When only the paired-end relationships were used in the assembly process, the N_{50} scaffold size was 22.4 kb. Adding mate-pair reads with 2, 5 and 10 kb insert sizes, the N_{50} scaffold size increased to 67, 173 and 389 kb, respectively. When integrated with additional libraries of larger insert size, such as fosmid and BAC end sequences, the N_{50} reached 1318 kb. The final assembly size was 727 Mb, 93.87% of which is non-gapped sequence. The gap filling was further carried out using 6.74 fold coverage of 454 data, which increased the N_{50} contig size to 31,429 bp with 15.4% of the gaps filled.

The single-base accuracy of the assembly was estimated by the depth and proportion of discordant reads. For the DM v3.0 assembly, 95.45% of 880 million usable reads could be mapped back to the assembled genome by SOAP 2.20 using optimal parameters. The read depth was calculated for each genomic location and the peak depths for the whole genome and the CDS regions are 100 and 105, respectively. Approximately 96% of the assembled sequences had more than 20-fold coverage. The overall GC content of the potato genome is about 34.8% with a positive correlation between GC content and sequencing depth. The DM potato should have few heterozygous sites and 93.04% of the sites can be supported by at least 90% reads, suggesting high base quality and accuracy.

6.3.2 RH Genome Sequencing and Assembly

Whole-genome sequencing of genotype RH was performed on the Illumina GAII platform using a variety of fragment sizes and reads lengths, resulting in a total of 144 Gb of raw data. These data were filtered using a custom C program and assembled using SOAPdenovo 1.03 (Li et al. 2009). Additionally, four 20-kb mate-pair libraries were sequenced on a Roche/454 Titanium sequencer, amounting to 581 Mb of raw data (Tables 6.1 and 6.2). The resulting sequences were filtered for duplicates using custom Python scripts.

The RH BACs were sequenced using a combination of Sanger and 454 sequencing at various levels of coverage. Consensus base calling errors in the BAC sequences were corrected using custom Python and C scripts, using a similar approach to that described previously (Chaisson et al. 2004). Sequence overlaps between BACs within the same physical tiling path were identified using megablast from BLAST 2.2.21 (Altschul et al. 1997) and merged with megamerger from the EMBOSS 6.1.0 package (Rice et al. 2000). Using the same pipeline, several kilobase-sized gaps were closed through

Table 6.2 RH whole-genome data from Illumina and 454 as per insert size

Library type	Insert size	Read length	Total reads	Total bp
A. Illumina				
Single-end	500 bp	75	107,492,068	8,061,905,100
Paired-end	200 bp	75	174,979,788	13,123,484,100
Paired-end	200 bp	125	326,814,010	40,851,751,250
Paired-end	300 bp	100	82,324,780	8,232,478,000
Paired-end	500 bp	75	528,763,314	39,657,248,550
Paired-end	500 bp	125	200,147,478	25,018,434,750
Matepair	2 kb	35	72,401,032	2,534,036,120
Matepair	5 kb	35	167,488,622	5,862,101,770
Matepair	10 kb	35	41,822,754	1,463,796,390
Total			1,702,233,846	144,805,236,030
B. 454				
Matepair	20 kb	259	686,844	178,053,005
Matepair	20 kb	255	653,410	166,359,937
Matepair	20 kb	310	765,621	237,055,547
Matepair	20 kb	293	643,577	188,612,498
Total			2,105,875	581,468,489

alignment of a preliminary RH whole-genome assembly. The resulting non-redundant contigs were scaffolded by mapping the RH whole-genome Illumina and 454 mated sequences against these contigs using SOAPalign 2.20 (Li et al. 2009) and subsequently processing these mapping results with a custom Python script. The scaffolds were then ordered into superscaffolds based on the BAC order in the tiling paths of the FPC map. This procedure removed 25 Mb of redundant sequence, reduced the number of sequence fragments from 17,228 to 3768, and increased the N50 sequence length from 24 to 144 kb.

6.3.3 Construction of the DM Genetic Map and Anchoring of the Genome

To anchor and fully orientate physical contigs along the chromosome, a genetic map was developed de novo using sequence-tagged-site (STS) markers comprising simple sequence repeats (SSR), SNPs, and diversity array technology (DArT). SSR and SNP markers were designed directly from assembled sequence scaffolds, whereas polymorphic DArT marker sequences were searched against the scaffolds for high-quality unique matches. A total of 4836 STS markers including 2174 DArTs, 2304 SNPs and 358 SSRs were analysed on 180 progeny clones from a backcross population ((DM × DI) × DI) developed at CIP between DM and DI (CIP no. 703825), a heterozygous diploid *S. tuberosum* group Stenotomum (formerly *S. stenotomum* ssp. *goniocalyx*) landrace clone. The data from 2603 polymorphic STS markers comprising 1881 DArTs, 393 SNPs and 329 SSR alleles were analysed using JoinMap 4 and yielded the expected 12 potato linkage groups. Anchoring the DM genome was accomplished using direct and indirect approaches. The direct approach employed the ((DM × DI) × DI) linkage map whereby 2037 of the 2603 STS markers comprised of 1402 DArTs, 376 SNPs and 259 SSRs could be uniquely anchored on the DM superscaffolds. This approach anchored ~52% (394 Mb) of the assembly arranged into 334 superscaffolds.

RH is the male parent of the mapping population of the ultra-high-density (UHD) linkage map used for construction and genetic anchoring of the physical map using the RHPOTKEY BAC library. The indirect mapping approach exploited in silico anchoring using the RH genetic and physical map (Van Os et al. 2006; Visser et al. 2009), as well as the tomato genetic map data from SGN (http://solgenomics.net/). Amplified fragment length polymorphism markers from the RH genetic map were linked to DM sequence scaffolds via BLAST alignment (Altschul et al. 1997) of whole-genome-profiling sequence tags (Van der Vossen et al. 2010) obtained from anchored seed BACs in the RH physical map, or by direct alignment of fully sequenced RH seed BACs to the DM sequence. The combined marker alignments were processed into robust anchor points. The tomato sequence markers from the genetic maps were aligned to the DM assembly using SSAHA2. Positions of ambiguously anchored superscaffolds were manually checked and corrected. This approach anchored an additional ~32% of the assembly (229 Mb). In 294 cases, the two independent approaches provided direct support for each other, anchoring the same scaffold to the same position on the two maps. Overall, the two strategies anchored 649 superscaffolds to approximate positions on the genetic map of potato covering a length of 623 Mb. The 623 Mb (~86%) anchored genome includes ~90% of the 39,031 predicted genes. Of the unanchored superscaffolds, 84 were found in the N90 (622 scaffolds greater than 0.25 Mb), constituting 17 Mb of the overall assembly or 2% of the assembled genome. The longest anchored superscaffold is 7 Mb (from chromosome 1) and the longest unanchored superscaffold is 2.5 Mb.

6.3.4 Identification of Repetitive Sequences

Transposable elements (TEs) in the potato genome assembly were identified at the DNA and protein level. RepeatMasker (Chen 2004) was applied using Repbase (Jurka et al. 2005) for TE identification at the DNA level. At the protein level, RepeatProteinMask (Chen 2004; Jiang et al. 2008) was used in a WuBlastX 9 (Altschul et al. 1997) search against the TE protein database to further identify TEs. Overlapping TEs belonging to the same repeat class were collated, and sequences were removed if they overlapped >80% and belonged to different repeat classes. Repetitive sequences account for at least 62.2% of the assembled genome (452.5 Mb) with long terminal repeat retrotransposons comprising the majority of the transposable element classes, representing 29.4% of the genome. In addition, subtelomeric repeats were identified at or near chromosomal ends. Using a newly constructed genetic map based on 2603 polymorphic markers in conjunction with other available genetic and physical maps, we genetically anchored 623 Mb (86%) of the assembled genome, and constructed pseudomolecules for each of the 12 chromosomes, which harbour 90.3% of the predicted genes.

6.3.5 Gene Prediction

To predict genes, ab initio predictions on the repeat-masked genome was carried out and then results were integrated with the spliced alignments of proteins and transcripts to genome sequences using GLEAN (Elsik et al. 2007). The potato genome was masked by identified repeat sequences longer than 500 bp, except for miniature inverted repeat transposable elements which are usually found near genes or inside introns (Kuang et al. 2009). The software Augustus (Stanke et al. 2004) and Genscan (Burge and Karlin 1997) was used for ab initio predictions with parameters trained for *A. thaliana*. For similarity-based gene prediction, the protein sequences of four sequenced plants (*A. thaliana*, *Carica papaya*, *V. vinifera* and *Oryza sativa*) were aligned onto the potato genome using TBLASTN with an *E*-value cut-off of 1×10^{-5}, and then similar genome sequences were aligned against the matching proteins using Genewise (Birney et al. 2004) for accurately spliced alignments. In EST-based predictions, EST sequences of 11 *Solanum* species were aligned against the potato genome using BLAT (identity ≥ 0.95, coverage ≥ 0.90) to generate spliced alignments. All these resources and prediction approaches were combined by GLEAN (Elsik et al. 2007) to build the consensus gene set. To finalize the gene set, the RNA-Seq from 32 libraries, of which eight were sequenced with both single- and paired-end reads, were aligned to the genome using Tophat (Trapnell et al. 2009) and the alignments were then used as input for Cufflinks (Trapnell et al. 2010) using the default parameters. Gene, transcript and peptide sets were filtered to remove small genes, genes modelled across sequencing gaps, TE-encoding genes, and other incorrect annotations. The final gene set contains 39,031 genes with 56,218 protein-coding transcripts, of which 52,925 non-identical proteins were retained for analysis.

6.3.6 Transcriptome Sequencing

RNA was isolated from many tissues of DM and RH that represent developmental, abiotic stress and biotic stress conditions (Xu et al. 2011). cDNA libraries were constructed (Illumina) and sequenced on an Illumina GA2 in the single- and/or paired-end mode. To represent the expression of each gene, we selected a representative transcript from each gene model by selecting the longest CDS from each gene. The aligned read data were generated by Tophat (Trapnell et al. 2009) and the selected transcripts used as input into Cufflinks (Trapnell et al. 2010), a short-read transcript assembler that calculates the fragments per kb per million mapped reads (FPKM) as expression values for each transcript. Cufflinks was run with default settings, with a maximum intron length of 15,000.

In developing DM and RH tubers, 15,235 genes were expressed in the transition from stolons to tubers, with 1217 transcripts exhibiting >5-fold expression in stolons versus five RH tuber tissues (young tuber, mature tuber, tuber peel, cortex and pith). Of these, 333 transcripts were upregulated during the transition from stolon to tuber, with the most highly upregulated transcripts encoding storage proteins. Foremost among these were the genes encoding proteinase inhibitors and patatin (15 genes), in which the phospholipase A function has been largely replaced by a protein storage function in the tuber. In particular, a large family of 28 Kunitz protease inhibitor genes (KTIs) was identified with twice the number of genes in potato compared to tomato in the tuber (Gore et al. 2009).

The stolon to tuber transition also coincides with strong up-regulation of genes associated with starch biosynthesis. It was observed that several starch biosynthetic genes were 3–8-fold more highly expressed in tuber tissues of RH compared to DM. Together this suggests a stronger shift from the relatively low sink strength of the ATP-generating general carbon metabolism reactions towards the plastidic starch synthesis pathway in tubers of RH, thereby causing a flux of carbon into the amyloplast.

6.3.7 Identification of Disease-Resistant Genes

Predicted open reading frames (ORFs) from the annotation of *S. tuberosum* group Phureja assembly V3 were screened using HMMER V3 (http://hmmer.janelia.org/software) against the raw hidden Markov model (HMM) corresponding to the Pfam NBS (NB-ARC) family (PF00931). The HMM was downloaded from the Pfam home page (http://pfam.sanger.ac.uk/). The analysis using the raw HMM of the NBS domain resulted in 351 candidates. From these, a high quality protein set ($<1 \times 10^{-60}$) was aligned and used to construct a potato-specific NBS HMM using the module ‘hmmbuild’. Using this new potato-specific model, we identified 500 NBS-candidate proteins that were individually analysed. To detect TIR and LRR domains, Pfam HMM searches were used. The raw TIR HMM (PF01582) and LRR 1 HMM (PF00560) were downloaded and compared against the two sets of NBS-encoding amino acid sequences using HMMER V3. Both TIR and LRR domains were validated using NCBI conserved domains and multiple expectation maximization for motif elicitation (MEME) (Bailey and Elkan 1995). In the case of LRRs, MEME was also useful to detect the number of repeats of this particular domain in the protein. As previously reported (Mun et al. 2009), Pfam analysis could not identify the CC motif in the N-terminal region. CC domains were thus analysed using the MARCOIL (Delorenzi and Speed 2002) program with a threshold probability of 90 (Mun et al. 2009) and double-checked using paircoil with a *P*-score cut-off of 0.025 (Porter et al. 2009). Selected genes (±1.5 kb) were searched using BLASTX against a reference *R*-gene set (Sanseverino et al. 2010) to find a well-characterized homologue. The reference set was used to select and annotate as pseudogenes those peptides that had large deletions, insertions, frameshift mutations, or premature stop codons. DNA and protein comparisons were used.

6.3.8 Haplotype Diversity Analysis

RH reads generated by the Illumina GA2 were mapped on to the DM genome assembly using SOAP2.20 (Li et al. 2009) allowing at most four mismatches, and SNPs were called using SOAPsnp. Q20 was used to filter the SNPs owing to sequencing errors. To exclude SNP calling errors caused by incorrect alignments, we excluded adjacent SNPs separated by <5 bp. SOAPindel was used to detect the indels between DM and RH. Only indels supported by more than three uniquely mapped reads were retained. Owing to the heterozygosity of RH, the SNPs and indels were classified into heterozygous and homozygous SNPs or indels.

On the basis of the annotated genes in the DM genome assembly, we extracted the SNPs located at coding regions and stop codons. If a homozygous SNP in RH within a coding region induced a premature stop codon, we defined the gene harbouring this SNP as a homozygous premature stop gene in RH. If the SNP inducing a premature stop codon was heterozygous, the gene harbouring this SNP was considered a heterozygous premature stop codon gene in RH. In addition, both categories can be further divided into premature stop codons shared with DM or not shared with DM. As a result, the numbers of premature stop codons are 606 homozygous PS genes in RH, 1760 heterozygous PS genes in RH but not shared with DM, 288 PS in DM only, and 652 heterozygous premature stop codons in RH and shared by DM. To identify genes with frame-shift mutations in RH, we identified all the genes containing indels of which the length could not be divided by 3. We found 80 genes with frame-shift mutations, of which 31 were heterozygous and 49 were homozygous.

To identify DM-specific genes, all the RH Illumina GA2 reads were mapped to the DM genome assembly. If the gene was not mapped to any RH read, it was considered a DM-specific gene and in total 35 DM-specific genes, 11 of which are supported by similarity to entries in the KEGG database, were identified (Kanehisa et al. 2004). To identify RH-specific genes, the RH Illumina GA2 reads not mapping to the DM genome were assembled into RH-specific scaffolds. Then, these scaffolds were annotated using the same strategy as for DM. To exclude contamination, the CDS sequences against the protein set of bacteria with the *E*-value cut-off of 1×10^{-5} using Blastx were aligned. CDS sequences with >90% identity and >90% coverage were considered contaminants and were excluded. In addition, all DM RNA-seq reads were mapped onto the CDS sequences, and CDS sequences with homologous reads were excluded because these genes may be due to incorrect assembly. In total, we predicted 246 RH specific genes, 34 of which are supported by Gene Ontology annotation (Shannon et al. 1996).

6.4 Conclusion

The sequencing of a unique doubled-monoploid potato clone to overcome the problems associated with genome assembly due to high levels of heterozygosity can help to generate a high-quality draft potato genome sequence that provides new insights into eudicot genome evolution. A combination of data from the vigorous, heterozygous diploid RH and relatively weak, doubled-monoploid DM, was used to directly address the form and extent of heterozygosity in potato and provide the first view into the complexities that underlie inbreeding depression. Combined with other recent studies, the potato genome sequence may elucidate the evolution of tuberization. This evolutionary innovation evolved exclusively in the Solanum section Petota that encompasses ~200 species distributed from the south–western United States to central Argentina and Chile. Neighbouring *Solanum* species, including the *Lycopersicon* section, which comprises wild and cultivated tomatoes, did not acquire this trait. Both gene family expansion and recruitment of existing genes for new pathways have contributed to the evolution of tuber development in potato.

Given the pivotal role of potato in world food production and security, the potato genome provides a new resource for use in breeding. Many traits of interest to plant breeders are quantitative in nature and the genome sequence will simplify both their characterization and deployment in cultivars. Whereas much genetic research is conducted at the diploid level in potato, almost all potato cultivars are tetraploid and most breeding is conducted in tetraploid material. Hence, the development of experimental and computational methods for routine and informative high-resolution genetic characterization of polyploids remains an important goal for the realization of many of the potential benefits of the potato genome sequence.

References

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3:21–29

Birney E, Clamp M, Durbin R (2004) Genewise and Genomewise. Genome Res 14:988–995

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. Bioinformatics 20:2067–2074

Chen N (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform 25:4.10.1–4.10.14

Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18:617–625

Elsik CG et al (2007) Creating a honey bee consensus gene set. Genome Biol 8:R13

FAOSTAT (2013) www.faostat.fao.org

Fedurco M, Romieu A, Williams S et al (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22

Gore MA et al (2009) A first-generation haplotype map of maize. Science 326:1115–1117

Hyman ED (1988) A new method of sequencing DNA. Anal Biochem 174:423–436

Jiang Z, Hubley R, Smit A, Eichler EE (2008) DupMasker: a tool for annotating primate segmental duplications. Genome Res 18:1362–1368

Jurka J et al (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280

Kuang H et al (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. Genome Res 19:42–56

Li R et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al (2005) Genome sequencing in micro-fabricated high-density picolitre reactors. Nature 437:376–380

Maxam AM, Gilbert W (1997) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Melamede RJ (1985) Automatable process for sequencing nucleotide. US Patent no. US4863849

Mun JH, Yu HJ, Park S, Park BS (2009) Genome-wide identification of NBS-encoding resistance genes in Brassica rapa. Mol Genet Genomics 282:617–631

Nyren P, Pettersson B, Uhlen M (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem 208:171–175

Paz MM, Veilleux RE (1999) Influence of culture medium and in vitro conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. Plant Breed 118:53–57

Porter BW et al (2009) Genome-wide analysis of *Carica papaya* reveals a small *NBS* resistance gene family. Mol Genet Genomics 281:609–626

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84–89

Ronaghi M, Pourmand N, Jain M, Willis T, Davis R (2000) Pyrosequencing for genome resequencing. In: 12th International genome sequencing and analysis conference, Miami, FL

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Sanseverino W et al (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res 38:D814–D821

Shannon JC, Pien FM, Liu KC (1996) Nucleotides and nucleotide sugars in developing maize endosperms: synthesis of ADP-glucose in *brittle-1*. Plant Physiol 110:835–843

Smith LM et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Spooner DM, Hijmans RJ (2001) Potato systematics and germplasm collecting, 1989–2000. Am J Potato Res 78:237–268

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32:W309–W312

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Trapnell C et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol 28:511–515

Van Os H et al (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 173:1075–1087

Van der Vossen E et al (2010) Whole Genome Profiling of the Diploid Potato CloneRH89-039-16 (Plant & Animal Genomes XVIII Conference, 2010)

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ (2001) The sequence of the human genome. Science 291:1304–1351

Visser RGF et al (2009) Sequencing the potato genome: outline and first results to come from the elucidation of the sequence of the world's third most important crop. Am J Potato Res 86:417–429

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

Yamey G (2000) Scientists unveil first draft of human genome. BMJ 321:7