

# **Genome bioinformatics of tomato and potato**

**Erwin Datema**

## **Thesis Committee**

### **Thesis Supervisor**

Prof. Dr. W.J. Stiekema  
Professor of Genome Informatics  
Wageningen University, the Netherlands  
Professor of Bioinformatics  
University of Amsterdam, the Netherlands

### **Thesis Co-supervisor**

Dr. R.C.H.J. van Ham  
Vice-President Bioinformatics and Modeling  
Keygene N.V., Wageningen, the Netherlands

### **Other members**

Dr. G. Giuliano, Ente per le Nuove tecnologie, l'Energia e l'Ambiente (ENEA), Rome, Italy  
Prof. Dr. J.H.S.G.M. de Jong, Wageningen University, the Netherlands  
Dr. W.H. Lindhout, Solynta, Wageningen, the Netherlands  
Dr. Ir. D. de Ridder, Delft University of Technology, the Netherlands

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences.

# **Genome bioinformatics of tomato and potato**

**Erwin Datema**

**Thesis**

submitted in fulfillment of the requirements for the degree of doctor  
at Wageningen University  
by the authority of the Rector Magnificus  
Prof. Dr. M.J. Kropff,  
in the presence of the  
Thesis Committee appointed by the Academic Board  
to be defended in public  
on Friday November 11<sup>th</sup> 2011  
at 11.00 a.m. in the Aula.

Erwin Datema  
Genome bioinformatics of tomato and potato  
140 pages

Thesis, Wageningen University, Wageningen, NL (2011)  
With references, with summaries English and Dutch  
ISBN 978-94-6173-047-3

# Contents

<b>Chapter 1</b>	Introduction .....	<b>7</b>
<b>Chapter 2</b>	BlastIf: BLAST analysis of long nucleotide sequences .....	<b>29</b>
<b>Chapter 3</b>	Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato .....	<b>35</b>
<b>Chapter 4</b>	<i>Solanum lycopersicum</i> cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements .....	<b>57</b>
<b>Chapter 5</b>	Genome sequence and analysis of the tuber crop potato .....	<b>81</b>
<b>Chapter 6</b>	Discussion .....	<b>101</b>
<b>Bibliography</b> .....		<b>107</b>
<b>Summary</b> .....		<b>123</b>
<b>Samenvatting</b> .....		<b>127</b>
<b>Curriculum vitae</b> .....		<b>131</b>
<b>Publications</b> .....		<b>133</b>
<b>Acknowledgements</b> .....		<b>135</b>
<b>Education statement of the graduate school Experimental Plant Sciences</b> .....		<b>137</b>



# Chapter 1

## Introduction

*Erwin Datema, Roeland C.H.J. van Ham*

The genome sequence of an organism represents its molecular blueprint that contains the list of parts in that organism and the instructions for how these parts interact with each other. Deciphering a genome sequence, that is, determining the linear order of nucleotides for each chromosome in the genome, allows molecular biologists to understand and manipulate this blueprint. For plants in particular, this in turn enables breeders to more efficiently engineer solutions for crop improvement to respond to the growing demand for food and energy from modern society. In the past two decades genome sequencing has developed from a laborious and costly technology employed only by a small number of large international consortia to a widely used, automated and affordable tool employed worldwide by many research groups. In 2001, the human genome sequence was made publicly available as the result of ten years of work by an international collaboration of several large sequencing “factories” with an associated price tag of three billion US dollars [1]. The original human genome sequence was derived from a mixture of individuals to provide anonymity to the sequence donors, but only seven years later the first genome sequence of an individual was generated by a single research center and a sequencing company [2]. Today, many laboratories around the world are sequencing hundreds of human samples [3] for only tens of thousands of dollars each in order to map the nucleotide and structural variation of the genome in a population. Genome-wide sequencing is currently employed as a tool to link inheritable diseases to their underlying genetic mutations [4] and will likely become a routinely used tool for disease diagnostics in the near future [5].

Not surprisingly, the breakthrough in genome sequencing has also impacted life sciences other than the human field. The genome sequences of many food animals such as chicken [6], cow [7], pig [8] and sheep [9] have been deciphered and are being exploited for fundamental research and applied to improve their breeding programs. A similar trend is visible for crop plant genomes, where the successful determination of the genome sequences of two rice cultivars [10, 11] paved the way for many other crops including cucumber [12], grape [13], maize [14], papaya [15] and soybean [16]. The developments in sequencing technologies have also impacted the associated bioinformatics strategies and tools, both those that are required for primary analyses such as data processing, management, and quality control, and those used for secondary analyses and interpretation of the data (e.g., sequence assembly and gene identification). The current thesis focuses on

the application of genome sequencing, assembly and annotation to potato and tomato, two members of the *Solanaceae* family and both major food crops with a worldwide production of 141 and 329 million tonnes in 2009, respectively [17].

## The *Solanaceae* family

The *Solanaceae* belong to the Asterid clade of Eudicot plants and include many species of high economic importance. Members of the family are employed in agriculture (Figure 1.1) for their edible fruits and tubers, including domesticated and wild tomato (*Solanum lycopersicum* and *Solanum pimpinellifolium*), pepper (*Capsicum* spp), eggplant (*Solanum melongena*), and potato (*Solanum tuberosum*). Other *Solanaceae* are cultivated for their medicinal and drug-related properties, for example tobacco (*Nicotiana tabacum*) and mandrake (*Mandragora officinarum*), or their ornamental flowers such as petunia (*Petunia x hybrida*). Potato is the economically most important species within the *Solanaceae*. It produces underground stems called stolons that under suitable environmental conditions form tubers, which are used by the plant for energy storage and vegetative propagation. As a food crop, the tubers contribute to dietary intake of starch, protein, antioxidants, and vitamins. Many commercial potato cultivars are highly heterozygous autotetraploids that suffer acute inbreeding depression and are susceptible to a plethora of devastating pests and pathogens. Nonetheless diploid accessions such as those from the *S. tuberosum* Phureja group exist and are important breeding stocks for the generation of modern potato varieties. Tomato fruits are the second most consumed vegetable after potatoes, and are a globally important dietary source of lycopene, beta-carotene, vitamin C, and fiber. In addition to its agricultural value and due to its diploid genetics and inbreeding potential, tomato is a widely used model species for fundamental research on subjects including fruit development and pathogen response. Wild *S. pimpinellifolium* is the closest wild progenitor of domesticated *S. lycopersicum* and together they are the only two species with red fruits in the tomato clade of the *Solanaceae* as a result of accumulation of lycopene. *S. pimpinellifolium* is widely used in tomato breeding as a source of disease resistance, stress tolerance and fruit quality.

The nuclear genome of many species of *Solanaceae* consists of twelve chromosomes, and potato and tomato are no exception. Their genomes are expected to measure approximately 840 Mb [18] and 950 Mb [19] in size, respectively, and there is a general genome-wide colinearity between them [20]. Several large-scale rearrangements have been identified between the two genomes, for example on chromosome 6 [21-23]. Potato and tomato chromosomes display a similar morphology, having long continuous stretches of less condensed euchromatin in both chromosome arms flanked by highly condensed heterochromatin at the telomere ends and the centromeres [24]. The pericentromeric heterochromatin spans the majority of most chromosomes and consists primarily of transposable elements, whereas the bulk of the genic sequences are located in the relatively

small euchromatic exterior of the chromosome arms [25]. The significant economic impact of these species combined with their importance in diploid and tetraploid plant genetics make them excellent targets for genome sequencing. The availability of their genome sequences will provide the community with a first glimpse into genome evolution of *Solanaceae* (and Asterids in general) and will impact both fundamental research and breeding strategies in these species for the coming years.



**Figure 1.1:** Agricultural products produced by Solanaceae family members include the fruits of *S. lycopersicum* (a; Penny Greb, Agricultural Research Service, public domain), *Capsicum annuum* (b; Leon Brooks, public domain) and *S. melongena* (c; Leon Brooks, public domain); the flowers of *Petunia x hybrida* (d; Rosendahl, public domain); the leaves of *N. tabacum* (e; net\_efekt, CC-BY-2.0); and the tubers of *S. tuberosum* (f; Scott Bauer, Agricultural Research Service, public domain).

## Strategies for genome sequencing

Genome sequencing is the process of deciphering the linear order of nucleotides in an organism's chromosomes. Two dominant strategies exist for sequencing genomes: Whole Genome Shotgun (WGS) sequencing and clone-based sequencing (also known as hierarchical shotgun sequencing). In the WGS approach (Figure 1.2), multiple copies of the target DNA are sheared into smaller, partially overlapping fragments. The fragments are separated on size by running them over a gel, after which fragments of a size appropriate for the sequencing technology are excised from the gel and purified. These are then processed and amplified either *in vitro* through PCR, or *in vivo* by inserting them into a cloning vector that is propagated within a host organism. Sequencing of the fragments is performed either from one end or from both ends ("double barreled shotgun sequencing").

The nucleotide sequences produced by the sequencing apparatus are referred to as reads, and they represent a measurement -including an associated error rate- of the DNA fragment. The genome sequence can then be reconstructed from these reads in a process called genome assembly, as explained further on in this chapter.

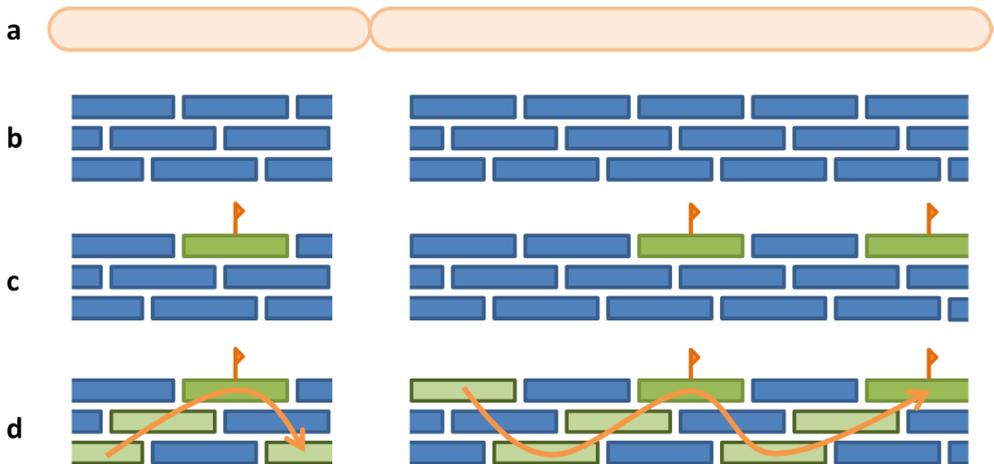
The clone-based sequencing strategy (Figure 1.3) first divides the global genome-wide sequencing problem into several smaller, local problems. Multiple equivalents of the target genome are sheared into fragments of a particular size range, which are inserted into cloning vectors such as fosmids, Bacterial Artificial Chromosomes (BACs) or Yeast Artificial Chromosomes (YACs). The cloning vectors are then propagated in a host organism (*Escherichia coli* for fosmids and BACs; *Saccharomyces cerevisiae* for YACs) such that each host cell contains one unique cloning vector. Together, these host cells constitute a library that contains a redundant representation of the target genome, with many partially overlapping insert sequences (inserts) in the cloning vectors. A physical map of these inserts can be constructed through fingerprinting [26, 27], in which inserts that physically overlap are grouped together into contigs. From each contig a minimal tiling path of inserts that spans the whole length of the contig and has the lowest possible amount of overlapping (redundant) sequence is then selected for sequencing. Preferably, at least one BACs in a contig is anchored to its corresponding physical location on the chromosome through Fluorescence *In Situ* Hybridization (FISH) [28] or its position on a sequence-based genetic map through overgo screening with marker sequences [29]. The selected inserts from each contig are then individually sequenced using the shotgun sequencing strategy, and the sequence of the complete contig is readily derived from the order of, and overlap between, the inserts.

## **Recent developments in sequencing technologies**

In the seventies of the last century, Sanger and co-workers laid the foundations for the chain-terminator sequencing technology [30]. This technique is often referred to as Sanger sequencing, and involves the generation of a set of truncated copies of the DNA fragment through use of labeled dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. The fragments are separated by size using high-resolution gel electrophoresis, after which the DNA sequence is determined from the linear order of labels that pass through the sequencer. Signal detection occurs when the fluorophores attached to the ddNTPs are excited by a laser and emit a fluorescence signal, the color of which determines the base call. Initially this technology was applied to sequence small bacteriophage genomes [31, 32], but automation and improvement in throughput were achieved during the drive to sequence the human genome [33, 34]. Currently, automated Sanger sequencing has achieved a throughput of two to three megabases per machine per day, with individual reads containing up to 900 bases of high-confidence information.



**Figure 1.2:** Whole genome shotgun sequencing approach. Target DNA (a) is randomly sheared into small fragments (b) followed by selection of fragments in a particular size range (c). Selected fragments are then partially sequenced either from one end, resulting in a single sequence (left side of panel d); or from both ends, resulting in two sequences with an known distance between them (right side in panel d).



**Figure 1.3:** Clone-based genome sequencing approach. Target DNA (a) is restriction-digested or sheared into large fragments that are propagated in cloning hosts. Contigs of overlapping clones are subsequently constructed through physical mapping (b). Seed clones from each contig (shown in green in panel c) are then anchored to their physical or genetic location on the genome (indicated by the orange flags). A tiling path of extension clones is selected from each seed clone such that the maximum span of the genome is covered with a minimum of overlap between the clones (panel d).

Sanger sequencing has remained the *de facto* standard for genome sequencing until the start of the 21<sup>st</sup> century. Driven by the quest for the “thousand dollar genome” [35], novel technologies have revolutionized the field of genome sequencing through both an increase in throughput of several orders of magnitude and a corresponding large decrease in cost per sequenced base. Importantly, these technologies require no individual sub-cloning of the fragments to be sequenced. The increased throughput of the next-generation sequencing (NGS) platforms has been achieved at the expense of shorter read lengths and lower accuracy. In contrast to capillary Sanger sequencing, these platforms implement a sequencing-by-synthesis or sequencing-by-ligation approach, in which one or more nucleotides are progressively added to a growing DNA strand. Each cycle of nucleotide addition is followed by image capture, where light signals indicate the successful incorporation of nucleotides into the newly synthesized DNA strand. The next-generation technologies differ from each other, both in implementation (i.e., sequencing methodology and chemistry, image capture, and data processing) and in read length and throughput. The Roche/454, Illumina/Solexa and AB/SOLiD technologies have been widely adapted in many genome sequencing projects since their inception [36, 37] and their current implementations will be discussed here in more detail. An outlook on the ongoing developments in sequencing technology will be presented in Chapter 6 of this thesis.

## **Pyrosequencing on the Roche/454 platform**

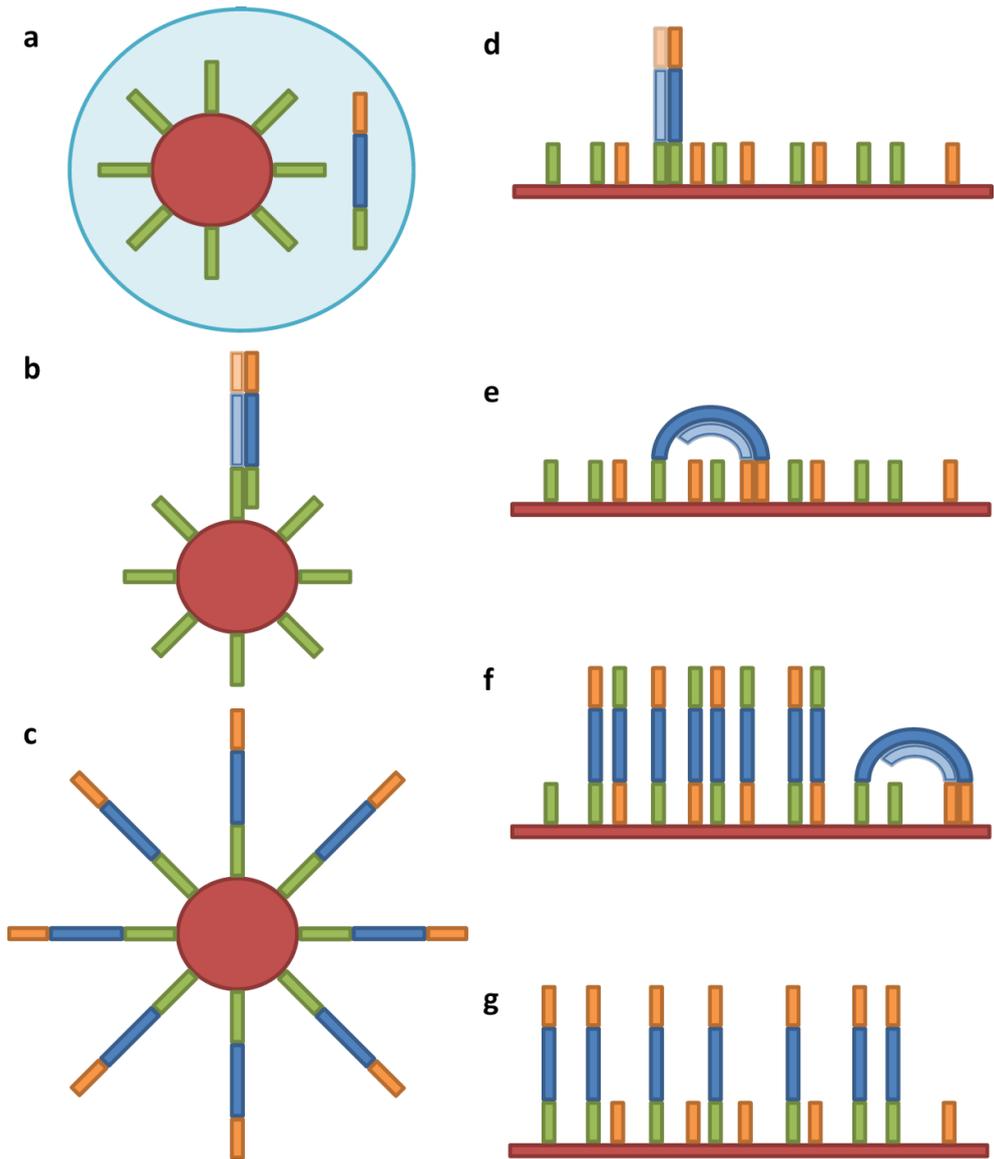
The Genome Sequencer (GS) FLX system developed by Roche/454 [38] is based on pyrosequencing [39]. It produces the longest reads of the currently available next-generation sequencing platforms, having a modal read length of up to 500 bp and a throughput of approximately 400 megabases per instrument run. DNA templates are prepared by shearing the target DNA, followed by fragment size selection and ligation of adapters specific for the 5' and 3' end of DNA, respectively (Figure 1.4a-c). Single-stranded DNA is subsequently amplified and hybridized to beads under conditions that favor one DNA molecule per bead. The bead-bound library is emulsified in a water-in-oil mixture such that each water droplet constitutes a microreactor that ideally contains one bead with one unique DNA fragment. Each bead is populated with several millions of identical copies of its target DNA molecule through emulsion-PCR (emPCR). Sequencing of the DNA fragments takes place by sequentially flowing the four nucleotides (A,C,G and T) in a fixed order over a PicoTiterPlate (PTP) device on which the bead library has been loaded (Figure 1.5a). The wells of PTP device have a diameter that accommodate exactly one bead per well, and as a consequence each well is populated by at most one distinct DNA fragment. If a nucleotide in a flow is complementary to the template strand of a DNA fragment in a well, then the polymerase incorporates the appropriate number of nucleotides, resulting in a light emission that is proportional to the number of nucleotides incorporated. The light emissions are recorded by a Charge Coupled Device (CCD) camera and stored as

flowgrams, which contain the signal intensity as a function of the linear flow order. The nucleotide sequence of each DNA fragment is readily derived from its flowgram.

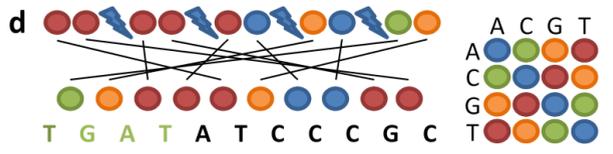
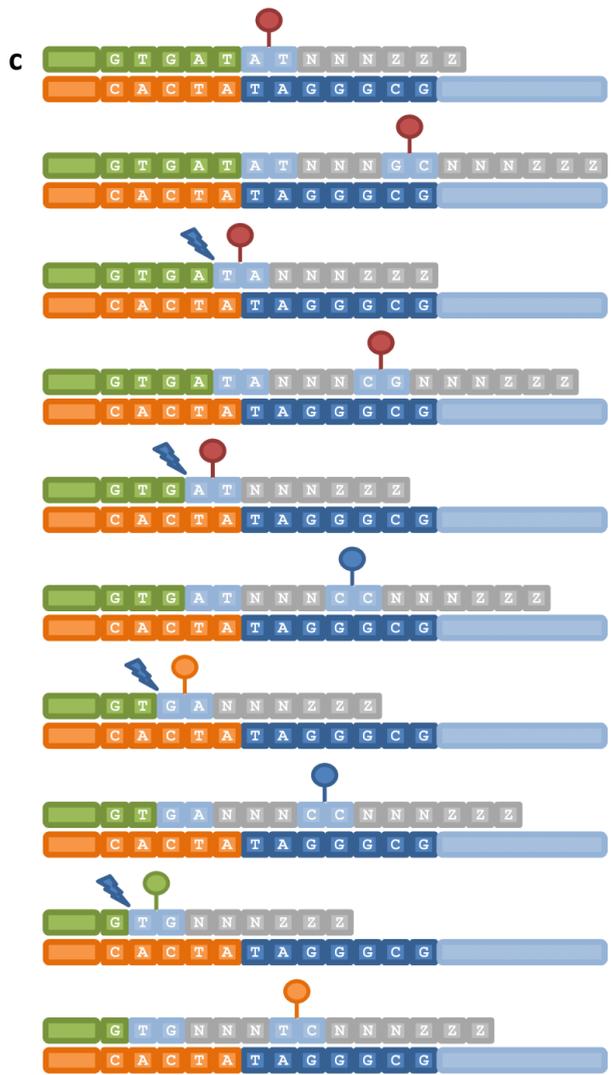
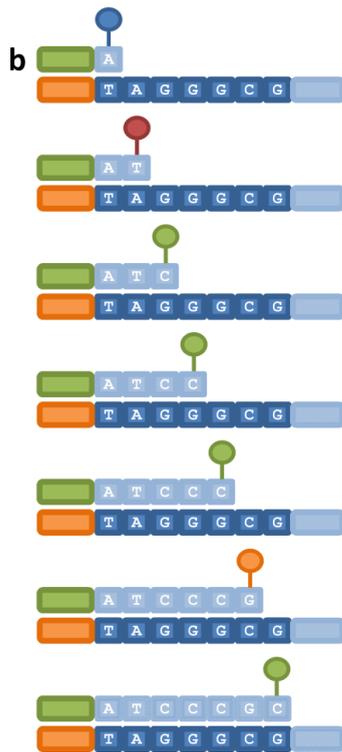
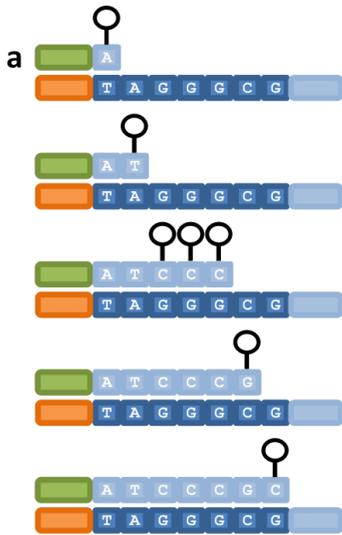
While the conditions during the hybridization of the DNA fragments to the beads have been optimized to yield beads with a single DNA molecule attached, two suboptimal situations can occur. Some microreactors can contain a single bead with multiple distinct DNA fragments. A bead in such a microreactor will be populated with a mixture of DNA fragments and will result in nonsense flowgram that represents a mix of fragment sequences. These “mixed reads” are readily detected and removed by the basecalling software that processes the flowgram images. The other suboptimal type of microreactor contains a single DNA fragment and multiple beads, resulting in two or more beads with identical DNA fragments. These beads will be loaded in distinct wells of the PTP device and will yield multiple reads of the same clonally duplicated fragment. Such reads can be distinguished from true repetitive sequences in the target genome since the former all start from the same position in the sequence. The primary source of read errors in 454 sequencing is found in the inaccurate determination of homopolymer tract length from the flowgrams. The intensity of the light signal that results from the incorporation of nucleotides within a given flow is linear only up to eight nucleotides, and there is a substantial error rate in the length of longer homopolymer tracts [39]. In contrast, substitution errors are more rare in 454 sequencing.

### **Cyclic reversible termination sequencing on the Illumina/Solexa platform**

The HiSeq 2000 system introduced by Illumina/Solexa [40] is based on the Cyclic Reversible Termination (CRT) sequencing technology [41]. With a current maximum of 100 bp, it has a shorter read length than the Roche/454 technology, but it comes with a considerably larger throughput of up to 200 Gb per instrument run. DNA template preparation for sequencing is similar to that employed for the Roche/454 system (Figure 1.4d-g). Single-stranded DNA templates are chemically cross-linked to a glass slide on which forward and reverse primers are randomly distributed. Each DNA fragment is copied into cluster of identical, single-strand sequences through solid-phase bridge amplification. The amount of DNA that is deposited on the glass slide is chosen such that the population density of the glass slide is maximized while clusters remain spatially separated from each other. Sequencing is performed by flowing a mixture of all four nucleotides, each covalently bound to a different fluorophore, over the glass slide (Figure 1.5b). The nucleotides are modified with reversible blocking groups, meaning that DNA polymerase incorporates exactly one nucleotide into the growing DNA fragment. The dyes are cleaved off the nucleotides and detected by Total Internal Reflection Fluorescence (TIRF) using two lasers, resulting in four intensity values (one for each nucleotide) for each cluster after every flow. The most probable nucleotide per flow is then derived from these four intensities, resulting in a sequence read with a length proportional to the number of flows.



**Figure 1.4:** Temple immobilization and amplification. In emulsion PCR as employed in Roche/454 and AB/SOLiD sequencing, water droplets containing a single bead and DNA fragment (a) are emulsified in oil. The target DNA hybridizes to the bead (b) and the bead is populated with thousands of identical DNA fragments through PCR amplification (c). In solid-phase bridge amplification as used in Illumina/Solexa sequencing, target DNA hybridizes to adapter sequences chemically attached to a glass slide (d). The slide-bound adapter is extended into a single-stranded template that is subsequently amplified through bridge amplification (e) into a cluster of identical, glass-bound DNA fragments in both orientations (f). Prior to sequencing, one of the two adapters is cleaved from the glass slide to retain a cluster of identically oriented fragments (g).



**Figure 1.5:** Imaging and base-calling in next generation sequencing. In Roche/454 sequencing (a), nucleotides are sequentially flown over a PicoTiterPlate device containing DNA-bound beads. After each flow in which nucleotides were successfully incorporated, a light signal is emitted with an intensity corresponding to the number of nucleotides. The resulting nucleotide sequence can be readily derived from the order and intensity of the signals. The Illumina/Solexa sequencing protocol (b) flows all four nucleotides over the glass slide at once. A single nucleotide is incorporated during every flow, which is detected through the color of its fluorescent dye. The order of the colors corresponds directly to the nucleotide sequence. On the AB/SOLiD platform (c), a library of all possible 1,2-probes is flown over the glass side to which the DNA-bound beads are attached and a matching probe is ligated to the growing DNA strand. Each probe contains two specific nucleotides, three degenerate nucleotides and three universal nucleotides to which a fluorescent dye is coupled, the color of which corresponds to the two specific nucleotides. The three universal nucleotides are cleaved off during signal detection followed by ligation of a new matching probe in the next flow. After a predetermined number of ligation cycles, the newly synthesized strand is disassociated and a novel sequencing primer, one base shorter than the previous one, is hybridized. In total, five rounds of ligation cycles and primer resets are performed. The two-base encoded signals in their original order (top of panel d) can be converted to the linear colorspace sequence (bottom of panel d) based on order of the primers. The nucleotide sequence can then be derived from the colorspace sequence through the two-base encoding scheme (right side of panel d) given that the first interrogated base of the primer (in this example “T”) is known.

As a result of overlapping emission spectra of the four fluorescent dyes, there exists a strong correlation between both the A and T intensities, and between the C and G intensities. While filters are present in the sequencing apparatus to distinguish between the nucleotides, this separation is somewhat limited and occasionally results in incorrect basecalls. As a result, substitutions are the most common type of error in Illumina sequencing [42]. Moreover, in CRT sequencing the signal for a flow is dependent on both the preceding and the following flow due to processes known as pre-phasing and phasing, respectively. These phenomena concern the incomplete removal of the reversible terminator and/or dye moiety, or the addition of more than one nucleotide in a given flow as a result of nucleotides that lack an effective terminator, resulting in asynchronous extension of the DNA molecules in subsequent flows. Due to (pre-)phasing, the signal intensities for each of the nucleotides in a flow consist of a mixture of signals from the current flow as well as noise from the previous and next flows. This effect increases with the number of flows that have been performed, causing a reduction in sequence fidelity towards the 3' end of the reads [43].

### **Sequencing by ligation on the AB/SOLiD platform**

The 5500xl SOLiD System developed by Applied Biosystems [44] employs sequencing-by-ligation [45], with read lengths of up to 75 bp and a total throughput of up to 300 Gb per

instrument run. Similar to the Roche/454 platform, sequencing is performed on single-stranded DNA fragments that have been hybridized to beads and amplified with emPCR (Figure 1.4a-c). The DNA-containing beads are chemically cross-linked to an amino-coated glass surface and sequencing of fragments is performed by hybridizing a mixture of fluorescently labeled probes to the primed target DNA (Figure 1.5c). These probes consist of two specific nucleotides that together determine the color of the fluorescent label (two-base encoding), three degenerate nucleotides and three universal nucleotides that together allow the probe to bind to any DNA molecule that starts with the complement of the two specific nucleotides. After ligation of the hybridized probe to the target strand, the universal nucleotides including the fluorescent label are cleaved off, and the color of the label is detected. Multiple successive cycles of probe hybridization, ligation and cleavage result in a pattern of color calls spaced in five-base intervals, with each color representing a combination of two bases. The pairs of nucleotides have been coupled to four colors such that each color represents four pairs of nucleotides. After a set number of cycles, the extension product is removed from the template DNA and a new primer, one nucleotide shorter than the previous, is used to start the next round of hybridization, ligation and cleavage cycles. In total, five rounds of such cycles are performed, after which all the combined color calls are ordered into a linear sequence. Each color in this sequence now represents two nucleotides, with the second nucleotide of any given pair being identical to the first nucleotide of the next pair. The nucleotide sequence can then be determined from the colorspace sequence (Figure 1.5d).

Basecalling errors in SOLiD sequencing have a larger impact than in the other technologies. Since each base is encoded in two consecutive color calls and each color is used to represent four distinct pairs of nucleotides, the determination of a given nucleotide from the colorspace sequence is dependent on both the current and the previous color call. Thus, a single incorrect color call results in all nucleotides after that position being incorrect when the nucleotide sequence is translated from the colorspace sequence. While this is a disadvantage in *de novo* assembly it is a strong advantage when mapping reads to a reference genome, as sequencing errors can be distinguished from true sequence polymorphisms with high fidelity.

## **Construction of sequencing templates**

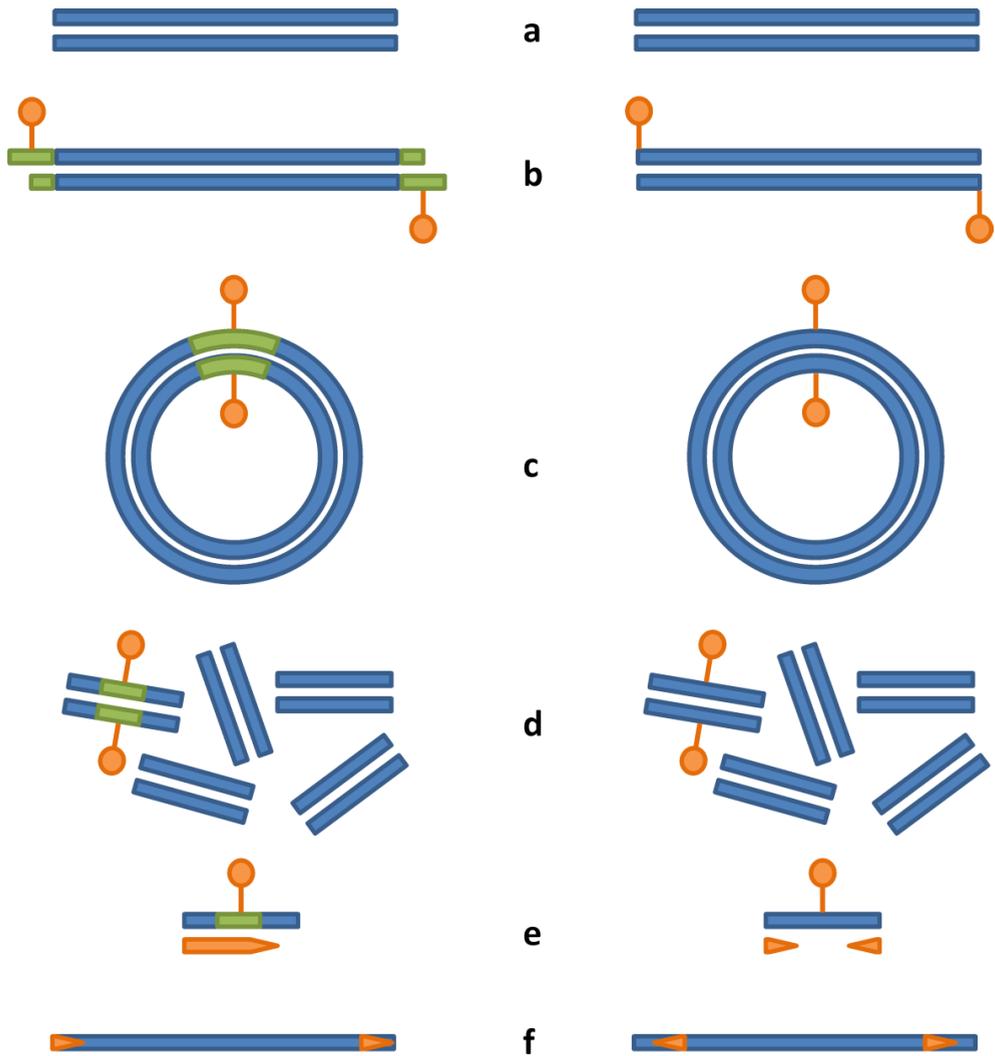
All three NGS platforms discussed here allow for the generation of fragment libraries from randomly sheared DNA. For such libraries, fragments are selected that are several hundred nucleotides in length. This size selection is not arbitrary, but instead is based on restrictions within the emPCR and bridge amplification protocols. While Roche/454 sequencing can only produce a single read per fragment, both the Illumina/Solexa and the AB/SOLiD platforms can produce two reads from a single DNA fragment, one from both ends. This technology is commonly referred to as paired-end sequencing and produces two (short)

sequence tags that are oriented with respect to each other, and have a known approximate distance between them that corresponds to the average fragment size. While both ends of the fragment can directly be sequenced from the beads on the AB/SOLiD platform, paired-end sequencing on the Illumina/Solexa platform involves “flipping” the template through a single bridge-amplification round in order to make the other end of the fragments accessible for the DNA polymerase.

The three technologies also allow for the construction of matepair libraries of various fragment lengths, typically between one and twenty kilobases. Matepair libraries are created by circularization of DNA fragments of the desired length through intramolecular ligation. The circularization protocol differs between the platforms; a biotinylated linker sequence is ligated between the two fragment ends in the Roche/454 and AB/SOLiD protocols, whereas the ends of the fragment are biotinylated and ligated without addition of a linker molecule in the Illumina/Solexa protocol (Figure 1.6a-c). The circularized fragments are sheared to produce fragments of a length similar to that used in the fragment sequencing protocols (Figure 1.6d). After enrichment for (or ideally, purification of) fragments that contain the biotin tag, the library is prepared further according to the regular fragment library protocol and sequenced. On the Illumina/Solexa and AB/SOLiD platforms, the resulting fragments are sequenced using the paired-end protocol to produce two sequence tags that are spatially separated by the originally selected matepair fragment length. In contrast, a single sequence that reads through the linker sequence is produced on the Roche/454 platform, from which the two tags from either end of the original fragment can be extracted (Figure 1.6f).

## **Assembling genome sequences from millions of short reads**

Sequence assembly is the reconstruction of the complete sequence of a DNA molecule from the reads that have been generated from it. The problem of sequence assembly stems from the inability to sequence long DNA molecules such as chromosomes in a single read. Two principal methods, overlap-layout-consensus and De Bruijn graph, have been developed to assemble the relatively short DNA sequence reads that are produced by current sequencing hardware into longer, more complete sequences [46, 47]. Both methods are equally suited for the WGS and the clone-based sequencing strategy, and the choice of method depends primarily on the read length and the error model underlying the sequencing technology.

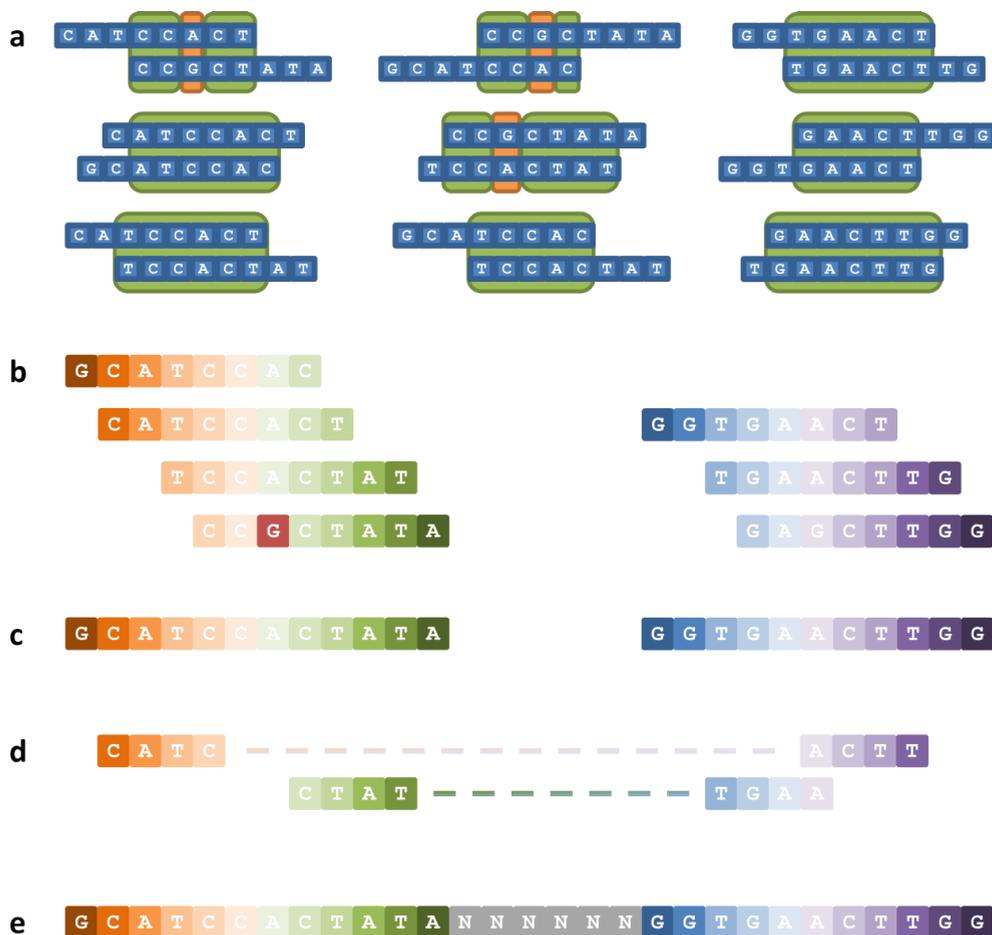


**Figure 1.6:** Construction of matepair libraries. A similar protocol is followed for Roche/454 sequencing on the one hand, (left) and Illumina/Solexa and AB/SOLiD sequencing on the other hand (right). Double-stranded DNA fragments of the desired size are isolated (a) and ligated to biotinylated adaptors, or directly biotinylated (b). The fragments are circularized (c) and sheared into smaller fragments of a size appropriate for the sequencing platform (d). Fragments containing the biotin tag purified and single-stranded DNA is subsequently sequenced using the shotgun of paired-end protocol (e). The matepair sequencing protocol results in two oriented sequence tags that are physically separated from each other by the selected fragment length (f).

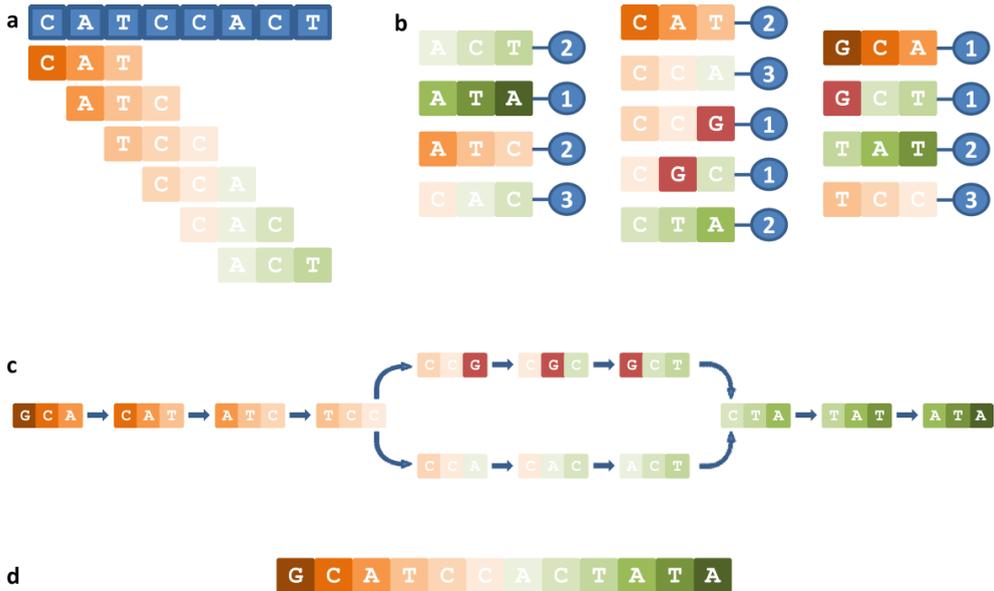
The overlap-layout-consensus assembly strategy (Figure 1.7) was pioneered in 1995 with the assembly of the *Haemophilus influenzae* genome sequence [48]. As the name implies, the algorithm consists of three phases. In the overlap phase, potential overlaps between all pairs of reads are identified, taking into account potential sequence polymorphisms and errors. Reads with a high degree of similarity are then aligned to each other in the layout phase. Finally, a consensus sequence is derived from the aligned reads by taking the most probable nucleotide at every column in the alignment. This strategy is often employed on Sanger and Roche/454 reads as these are relatively long, thereby reducing the chance on spurious overlaps due to repetitive sequences. Examples of overlap-layout-consensus assemblers are ARACHNE [49], CABOG [50] and Newbler [39]. All-versus-all sequence similarity searches are computationally infeasible for reads produced from Illumina/Solexa and AB/SOLiD sequencing due to the short read length, relatively high error rate and massive data volume of these platforms. Novel assembly tools such as ABySS [51], SOAPdenovo [52] and Velvet [53] have been developed to reconstruct genome sequences from these type of data using the De Bruijn graph algorithm first implemented in Euler [54] (Figure 1.8). In this approach, all reads are partitioned into  $k$ -mers (substrings of length  $k$ ) that become the nodes in the graph. The directed edges between the nodes represent the exact overlap of  $k - 1$  nucleotides between two  $k$ -mers. Sequencing errors manifest themselves in the graph as short “tips” that branch off the linear path representing the consensus sequence, whereas true sequence polymorphisms between alleles or multiple copies of a related sequence element form “bubbles”. The errors are readily ignored by the assemblers by cutting the “tips” off the graph, whereas a consensus sequence is generally determined for the true polymorphisms in a process called “bubble pinching”.

While the aim of genome assembly is to produce a single consensus sequence for each chromosome, insufficient read depth and repetitive sequences often result in a fragmented assembly. Ideally, all reads from a single chromosome form a single group of overlaps in the overlap-layout-consensus approach, or a single linear path in a De Bruijn graph; however, (local) undersampling of the genome results in multiple distinct groups of overlaps, or multiple disjoint paths, respectively. Repetitive sequences can lead to misassembled, chimeric sequences as well as assembly fragmentation. In the overlap-layout-consensus approach, reads from distinct repeat copies on the genome can be combined into a single group of overlaps, whereas in the De Bruijn graph such repeats can result in branches and loops in the graph that cannot be resolved into a single linear sequence. Both types of algorithm often prefer to break the assembly when such repeats are encountered rather than maintaining incorrect connections, resulting in multiple contiguous sequences (“contigs”) for each chromosome. Sequence contigs can be ordered and oriented into larger, gapped sequences called scaffolds using paired-end and matepair sequences (Figure 1.7c-d). The gaps in these scaffolds represent either one or more repetitive sequences that were not assembled (or assembled incorrectly), or a section of the genome that was not captured by reads. Based on the original fragment lengths of the paired sequences, the gaps within the scaffolds have an approximately known size. Algorithms to close these inter-contig gaps have been developed both within genome assembly software

[55] and as separate tools [56]. Clone-based sequencing reduces the repeat problem and the risk of long-range misassembly by partitioning the genome assembly problem into multiple smaller, local assembly problems. Sequences that are repetitive in a genome-wide context need not be repetitive within a single clone sequence.



**Figure 1.7:** Overlap-layout-consensus sequence assembly. Sequence reads are aligned pairwise to determine all sequence overlaps (a). Groups of overlapping reads are then laid out in a multiple sequence alignment (b) that tolerates some amount of sequence variation (for example, the bright red “G” in the lower row of the leftmost multiple alignment). The consensus sequence for each contig is derived from the corresponding reads by taking the most probable letter in each column of the multiple alignment (c). Multiple contig sequences can be oriented and ordered together into a scaffold through alignment of matepair sequences (d). The resulting gap size can be estimated from the fragment size distribution of the matepair libraries and is represented as N’s in the scaffold sequence (e).



**Figure 1.8:** De Bruijn-graph sequence assembly. Sequence reads are partitioned into  $k$ -mers (a;  $k = 3$ ) and the frequency of each  $k$ -mer in all reads is recorded (b). A directed graph is constructed in which the nodes are the  $k$ -mers and the edges are overlaps between two  $k$ -mers that occur consecutively in at least one read (c). Sequence variation (the bright red “G”) is visible in the graph as a “bubble” of length  $k$ . The consensus sequence is derived through traversal of the linear path in the graph in which the  $k$ -mers are supported by the largest number of reads (d).

The accuracy of an assembly is dependent not only on the decisions of the assembly algorithm, but also on the quality of the underlying read data. While the quality values produced by the sequencing apparatus provide a good estimate of the accuracy of a nucleotide within a read [57], several tools have been developed to exploit multiplicity of read coverage to correct potential sequencing errors. Prior to the general availability of NGS technologies, base error correction software employed multiple sequence alignment to identify erroneous basecalls [58, 59]. However, alignment-based methodologies are impractical for the large data volumes commonly produced by Illumina/Solexa and AB/SOLiD sequencing, and in response the spectral correction methodology was developed by Chaisson *et al.* [60]. In this method, each sequence substring of length  $k$  (commonly referred to as a  $k$ -mer) in a read is classified as either solid or insolid, based on whether its multiplicity in the complete read data is above or below a threshold, respectively. Reads containing insolid  $k$ -mers can then be trimmed [61] or corrected [62-64]. Despite stringent quality control of the read data, the assembled sequences may still contain errors in both content and structure. Base errors in the consensus sequence can result from insufficient read depth, misplaced reads or even the underlying error model of the sequencing technology such as homopolymer errors in Roche/454 sequencing.

Structural errors occur when reads from disjoint regions in the genome are assembled together into a chimeric contig, or when paired sequences inadvertently connect physically separated contigs. Tools have been developed to identify structural errors, but they are limited to medium sized assemblies that have been generated by overlap-layout-consensus assemblers [65, 66]. More recently there has been a focus on the integration of multiple genome assemblies from various assemblers and/or read datasets in order to produce a single assembly of superior quality [67, 68]. Ideally, a whole-genome assembly is validated through a reference data set such as a manually curated correction of BAC clone assemblies, but in the absence of such a costly resource a high-quality genome sequence of a related organism can also be used as a reference [69].

As a consequence of the limitations outlined above, an assembled genome sequence often consists of hundreds to thousands of disjoint scaffolds. If a clone-based sequencing strategy is followed, a physical map of the clones can be utilized to group scaffolds from a single clone contig together. Scaffolds from different clone contigs may be connected to each other through end sequencing of the clone library, after which the resulting gaps may be closed by sequencing the gap-spanning clones. The clone library can additionally be exploited to anchor the sequence scaffolds to their physical location on the genome by “painting” one or more clones from a scaffold on the chromosomes through use of fluorescence *in situ* hybridization. Irrespective of the sequencing approach the scaffolds can also be anchored and oriented on their probable chromosomal location if a sequence-based genetic map of the species is available. The success of this approach depends strongly on the ratio between the marker density and distribution in the genetic map on the one hand, and the number and length of the sequence scaffolds on the other hand.

## **Interpretation of genome sequences through annotation**

The linear strings of nucleotides that together comprise a genome sequence are of limited interest by themselves. Genome annotation encompasses the process of assigning a plausible biological interpretation to a genome sequence through identification and characterization of the elements contained therein. Ideally, each element in a genome such as a gene or a transposon is annotated through experimentally obtained evidence. The disparity in throughput with which genomes can be sequenced and assembled compared to the laboriousness of experiments to determine the precise structure and function of a single element has resulted in the development of algorithms that can predict these features in a genome. Prediction algorithms can be divided into *ab initio* algorithms that exploit only the properties of the genome sequence itself, and evidence-based algorithms that incorporate experimentally obtained evidence (often produced in high-throughput experiments) into their predictions. The quality of the predictions from both types of algorithms is strongly dependent on the knowledge that is implicitly encoded in the algorithm, and the quality of the data available to the algorithm to make the predictions. There are two components to

genome annotation: structure annotation and function annotation. Structural genome annotation involves the determination of the location, boundaries and composition of distinct sequence elements, whereas function annotation attempts to assign a probable biological function to each of these sequence elements.

## Structural genome annotation

A logical first step after the assembly of a new genome is the identification of the precise locations and boundaries of the functional elements in the sequence. Of all the features in a genome, protein-coding genes are the most extensively studied. It is therefore not surprising that many tools have been developed to identify the gene complement in a genome sequence [70, 71]. Protein-coding genes have a number of features in common through which it is computationally feasible to predict them *ab initio* in a genome sequence. For example, the protein-coding region of a gene is delimited by a start and stop codon, has no internal stop codons and has a length that is modulo three. Moreover the distribution of nucleotide hexamers differs between coding and non-coding sequence [72]. In eukaryotes, the coding region may be interrupted by introns, which can be identified computationally through conserved signals on the border between the coding exons and non-coding introns (splice sites). Tools to predict gene sequences in a genome range from naïve Open Reading Frame (ORF) predictors like *getorf* from the EMBOSS suite [73] to complex eukaryotic gene finders such as *geneid* [74], *genscan* [75] and *GlimmerHMM* [76] that incorporate Hidden Markov Models (HMMs) and Weight Matrix Models (WMMs) to identify conserved gene properties. More recently, Support Vector Machines (SVMs) have also been applied successfully to the gene prediction problem [77]. Whereas some sequence properties are conserved over a wide evolutionary range, other properties including the sequence surrounding the splice site and the distribution of intron lengths are unique for a particular genus or species. To overcome this variation, gene finders can be trained on the specific properties of the genome under investigation through a set of known gene structures from the corresponding species, or a closely related species [78]. In practice such data are often not available, and an alternative strategy involves the fine-tuning of gene finders for a genome through iterative prediction and self-training on the predicted gene set [79, 80].

In contrast to pure *ab initio* predictors, evidence-based gene finders exploit experimental evidence to enhance their prediction accuracy. *Augustus* [81] complements the *ab initio* gene finding from HMMs with aligned transcript sequences, whereas *GenomeThreader* [82] predicts gene structures from aligned transcripts or proteins alone. The predictions made by these tools however depend strongly on the availability of good sequence coverage of the transcriptome or proteome. More complex tools such as *EuGene* [83] and *JIGSAW* [84] take advantage of the benefits of multiple *ab initio* and evidence-based predictors by integrating the results from different tools, either through linearly combining

and voting schemes or through integration of these results into their own trained prediction algorithms. These tools often show improved prediction accuracy compared to the individual *ab initio* and evidence-based predictors that they integrate [85].

While protein-coding genes have been the focus of gene discovery methods for many years, non-coding RNA (ncRNA) genes have more recently attracted the focus of the scientific community [86]. This class of genes produces functional RNA molecules that often perform a regulatory role, and in recent years many software tools have been developed for the prediction of ncRNA genes. Examples of *ab initio* prediction tools include MiRPara [87] for miRNA genes, RNAmmer [88] for rRNA genes, SnoScan [89] for snoRNA genes and tRNAscan-SE [90] for tRNA genes. While their short lengths prohibit effective sequenced similarity searches, small RNA genes are often conserved in structure between related species. The INFERNAL software suite [91] exploits this information by searching a genome sequence for small non-coding RNAs using a database built from consensus RNA secondary structure profiles.

A striking characteristic of plant genomes, and large eukaryotic genomes in general, is their high repeat content [92]. Retrotransposable elements occupy a substantial fraction of many plant genomes [93] and interfere with accurate gene prediction as these elements contain ORFs required for their multiplication in the genome. Prior to gene finding, a genome sequence is often “repeat masked” to eliminate the prediction of false-positive and chimeric gene structures that may result from complete and fragmented retrotransposons. Identification and masking of repetitive elements in a genome sequence is generally performed through alignment against a database of known repeat sequences, for example using the RepeatMasker software [94] with the RepBase database [95]. For newly completed genomes such databases do not exist; however, due to their conserved structure some repeat types including retroelements can also be identified *ab initio* using tools such as LTRharvest [96] and MGEScan-LTR [97]. The RECON software [98] identifies repeats through their multiplicity in the genome sequence, without taking into account *a priori* knowledge about the structure of the elements. The false discovery rate of *ab initio* repeat prediction tools is high, as tandemly duplicated genes and ubiquitous gene families often meet the criteria these tools impose [99]. In contrast, there exist several classes of repeats that can be identified with high fidelity by their structure alone. Tandem repeats, inverted repeats and Simple Sequence Repeats (SSRs) can readily be identified by tools like tandem repeats finder [100], inverted repeats finder [101] and SciRoKo [102], respectively. While they do not often interfere with gene prediction, SSRs are of particular interest to geneticists and breeders as they are often highly polymorphic within populations, making them an excellent tool for marker development.

## Function annotation

Once the elements in a genome sequence have been identified, the next step is to assign to them a plausible biological function. Computational inference of the function of a particular sequence can be achieved either directly through sequence similarity searches, or indirectly through the identification of common motifs or domains between a group of functionally related sequences. Both methodologies exploit the wealth of sequence annotations that have been generated and deposited in public databases in the past decades to annotate a newly generated sequence. The accuracy and reliability of annotations derived from database searches depend strongly on both the availability of evolutionarily related sequences in the databases, and the quality of their annotations.

A similarity search involves the comparison of a novel sequence of interest to a database of annotated sequences, the function of which have preferably been experimentally validated. The function annotation of the most similar sequence within a defined similarity threshold can then be transferred to the query sequence under the assumption that primary sequence similarity implies functional homology. BLAST [103] is the best-known tool for sequence similarity searches and can be used on many public as well as private sequence databases. The method does not take into account the conceptual difference between orthologous sequences, which likely have the same function between different species, and paralogous sequences that have arisen from intraspecies duplication and likely fulfill a different biological role [104]. Public databases such as GenBank [105] contain millions of sequences spanning billions of basepairs, but many of the sequences lack proper experimental annotation of function or have themselves been annotated through similarity searches. Using the public databases to annotate new sequences can therefore lead to propagation of inaccurate and erroneous annotations. Curated databases like Swiss-Prot [106] have been developed to limit the propagation of annotation errors, but contain only a fraction of the number of sequences in GenBank and are therefore less likely to result in a sequence match with high similarity.

Motif and domain searches provide a more coarse-grained alternative to sequence similarity searches. While the latter focus on the similarity between two sequences over their whole length, the former rely on the conservation of small subsequences within a group of functionally related sequences. Prime examples of such conserved subsequences are protein domains, the modular functional sub-parts of proteins. Domains can be identified and extracted from a multiple sequence alignment of functionally related proteins and represented as HMMs or WMMs, which in turn can be used to query novel protein sequences. InterProScan [107] employs a large collection of protein domain databases [108] to identify conserved protein signatures in sequences of interest. In addition to the free-text annotation that is also present in the public sequence databases, the annotations in the domain databases have been enriched with Gene Ontology (GO) [109] terms that describe the biological process, molecular function and/or cellular component that are associated with the protein domain. The GO consists of a fixed set of clearly defined terms

and relationships that result in a high-level annotation of function and location and allow for systematic comparisons between annotated protein datasets.

## **Comparative genome analysis**

An annotated genome sequence, while a powerful tool on its own, contributes to the understanding of the genetic blueprint of a single (individual of a) species only. Alignments between genome sequences of multiple accessions or varieties of a single species allow for the study of genome diversity and evolution through the identification of Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion polymorphisms (InDels). Moreover, alignments between the genomes of related species, for example from the same genus, can be generated to identify structural variation such as translocations, inversions, Copy Number Variation (CNV) and Presence/Absence Variation (PAV). Software like MUMmer [110] and BLASTZ/LASTZ [111] has been developed to align complete genome sequences and extract the variation between them. Another method to identify sequence variation between related genomes involves mapping the unassembled short reads of a newly sequenced genome to the high-quality reference assembly of an existing, related genome [112]. Tools such as BWA [113] and Bowtie [114] excel in revealing SNP and InDel variations between two related genomes through read mapping, whereas other software like Pindel [115] focuses on the detection of structural variation through readpair alignment. The identified sequence variation from both approaches can be utilized to study the evolution of genomes, and to generate molecular markers that can be exploited to screen large populations. In such studies, variation in protein-coding sequences is of particular interest. Comparing the gene complement of related species can distinguish the species-specific genes or pathways that may perform specialized roles in these species from the common set of genes that are shared between them. Moreover, studying the variation on the nucleotide level between different genotypes within a species can aid in the identification of new, superior alleles of genes or regulatory sequences that underlie a trait of interest.

## **Automation of genome annotation and analysis**

Annotation of a genome sequence involves the execution of a number of different tools in a particular combination and order on a collection of sequences, ranging from a small number of complete chromosomes to hundreds or thousands of sequence contigs and scaffolds. Genome sequences from evolutionary related species can readily be annotated using the same software tools albeit sometimes with modified parameters and training data. These properties make genome annotation a repetitive, modular task with many inter-task dependencies that can best be described as a pipeline. Efficient execution of complex pipelines requires a flexible workflow management environment in which the pipelines can

be managed and monitored, as well as adjusted for genome-specific datasets and parameters. With this in mind, the Cyrille2 workflow management system [116] was developed to automate the annotation of partial and complete genome sequences. It provides a graphical user interface to create, modify, execute and manage genome annotation pipelines that can be composed from more than twenty distinct third-party software tools as well as a similar number of customized scripts. Tools are included for *ab initio* and evidence-based gene structure prediction, integration of gene prediction, repeat identification and masking, gene function annotation and comparative genome analysis, among others. Given a user-defined pipeline, the system automatically determines which analyses are eligible to be executed at a given time and uses the Sun Grid Engine (SGE) to schedule these on a computer cluster without user intervention. Once completed, the results of an analysis are available in a relational database that can be accessed by the Generic Genome Browser software [117] to visualize the annotations on the genome sequence. This combination of a robust automated pipeline system with an accessible and intuitive interface to browse the annotations accelerates the discovery of novel information from a genome sequence.

## Scope of this thesis

The aim of this thesis is to uncover the genomes of potato and tomato and the information contained within their sequences in order to both provide a framework for future fundamental research into solanaceous genomes (and Asterid genomes in general) and to accelerate the breeding programs of these crops. In Chapter 2 the BlastIf algorithm is introduced, which extends the commonly used BLAST program to efficiently annotate long nucleotide sequences. This tool was employed within the Cyrille2 workflow management system to annotate tomato and potato BAC sequences. A preliminary sequence comparison of the tomato and potato genomes based on a large collection of BAC End Sequences (BESs) is presented in Chapter 3. This study discusses the gene and repeat content of these genomes and provides a first glimpse into the evolution of *Solanaceous* genomes. Chapter 4 provides a more detailed look into the structural organization of protein-coding and repetitive elements on tomato chromosome 6. In this chapter, the first draft sequence of tomato chromosome 6 is explored through a combination of genomics, genetics and cytogenetics. Many of the results from this study are complemented by the findings in Chapter 5, where the whole genome sequence of a homozygous diploid potato is generated and compared to a collection of BAC and WGS sequences from a heterozygous diploid potato. Finally, Chapter 6 discusses the future developments in genome sequencing and the impact this is expected to have on plant breeding, as well as the lessons learned from sequencing two Solanaceous genomes.

## Chapter 2

# BlastIf: BLAST analysis of long nucleotide sequences

*Erwin Datema, Mark W.E.J. Fiers, Roeland C.H.J. van Ham*

### Summary

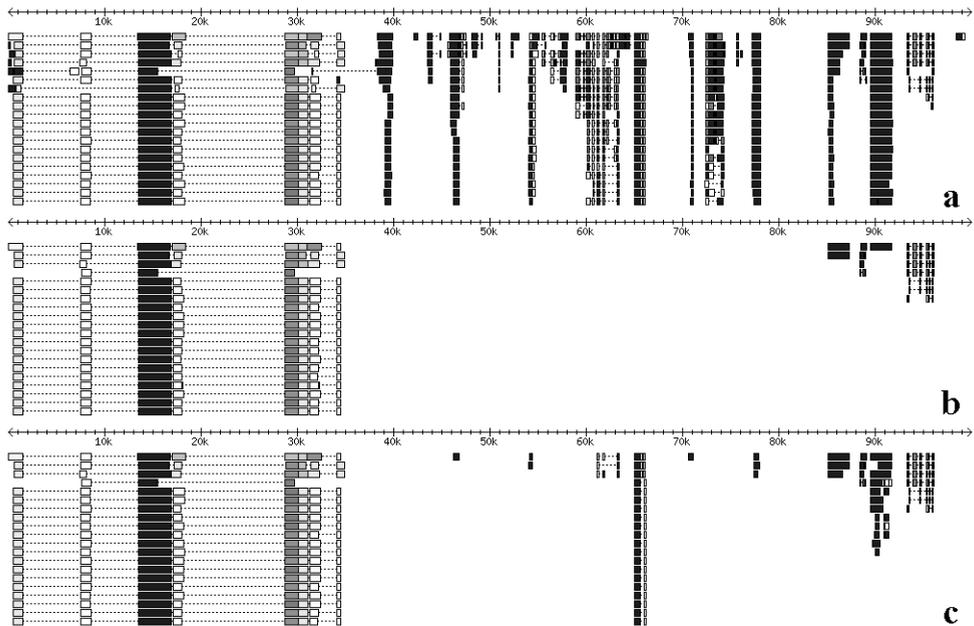
BlastIf generates a comprehensible BLAST output for long nucleotide sequences through intelligent filtering of the results produced by the blastall program. The filtering results in a strong reduction of similar BLAST hits while revealing most of the variation present among hits, irrespective of the length of the query sequence and the composition of the elements therein.

### Introduction

The BLAST algorithm [118, 103] implemented in the blastall software is an excellent tool for assigning a putative function to a novel gene, transcript or protein sequence. It can also be applied to gain insight in the functional and structural organization of long nucleotide sequences such as genomic BAC clones. In contrast to sequence alignment tools that focus solely on the discovery of gene structures, for example Genewise [119], BLAST can identify many additional genetic elements such as retro-elements, transposons and other repeat sequences. Two problems are associated with running blastall on a long nucleotide sequence: (i) the large number of similar hits that can be generated due to database redundancy and (ii) the relationships implied between High-scoring Segment Pairs (HSPs) within a hit that in fact correspond to distinct elements on the query sequence.

The first problem occurs, for example, when a BLASTX analysis is performed on a 100 kb genomic sequence that contains a genetic element that is overrepresented in the sequence database, for example a retrotransposon or a cytochrome P450 gene. The blastall software will return thousands of hits to this element alone, while in practice, a user will only evaluate the highest scoring hits. A high redundancy will thus imply that many lower scoring hits will be missed by the user, even though these might reveal new elements or hitherto unobserved variation on already identified elements.

Several strategies have previously been proposed to address this problem. Firstly, blastall includes a `-K` parameter [120] that is described to limit the number of hits per region on a query sequence to a fixed number. Closer evaluation of this parameter however reveals that that the intended functionality is not implemented in the current version (2.2.13) of the program, and that using this parameter can inadvertently result in a dramatic loss of identified regions on a genomic sequence (Figure 2.1).



**Figure 2.1:** Evaluation of the blastall `-K` parameter on the output of blastx. A 100 kb of genomic sequence from *Arabidopsis thaliana* chromosome 2 (GenBank accession NC\_003071.3, bases 30.000 through 130.000) was blasted three times against the nr database using settings only different for the `-K` parameter: (a) `-K` disabled; (b) `-K` 10; (c) `-K` 100. In each run, parameters for the number of descriptions (`-b`) and hits (`-v`) was set to 10.000.000 in order to retrieve all hits to the database. Without the `-K` parameter (a), the blast hits to the nr database are distributed over most of the length of the sequence. With the `-K` parameter set to 10 (b), only a fraction of the regions identified in (a) are found. The number of hits to each region varies, and differs significantly from the requested number of hits per region (i.e., 10). In the output of `-K` 100 (c), hits to additional regions in sequence are identified when compared to `-K` 10 (b). Furthermore, highly significant hits that are not present in the `-K` 10 output appear in the `-K` 100 output. The number of hits identified for each region varies between the regions, and exceeds the requested number of 100 more than 3-fold in the region 0 – 35 kb. Note that each of the figures has been truncated to show only the first 20 rows of alignments. Similar results were obtained with the use of `-K` on other genomic sequences (not shown). Images were created using the Graphics module from Bioperl [121].

Power-BLAST [122] addresses this problem by masking repetitive elements, but this requires *ab initio* knowledge of the repetitive elements (e.g. interspersed repeats) that may be present in the sequence of interest. Another approach has recently been outlined by Cantalloube *et al.* [123], in which redundancy between hits is reduced by suppressing hits that produce identical alignments, such as a hit to a well-conserved protein domain. As demonstrated by the authors, this method is highly practical for reducing the redundancy in an output without loss of information. However, the presence of a single residue difference between the alignments contained in two hits will result in the retention of both these hits in the final output. In practice, sequence databases contain many sequences that differ only in a few residues, such as gene paralogs and orthologs, and in many cases this stringent criterion for filtering will still result in a large number of hits.

The second problem stems from the fact that the HSPs within a single hit can span a large fraction of the query sequence, even though they are unlikely to represent only a single genetic element. This greatly hinders the separation of multiple similar elements such as gene family members on the same query sequence from each other.

To increase the usefulness of using BLAST in comprehensively annotating long sequences, these problems should be overcome by, firstly, reducing the redundancy of the output while retaining those alignments that display variation, and secondly, by separating HSPs when they appear to belong to different elements such as tandemly repeated genes. The redundancy among database hits can most easily be reduced by limiting the number of similar, best scoring hits to each element in a sequence to a specified number. However, because this is likely to remove potentially relevant variation present among the lower scoring hits, all hits should be evaluated at the HSP level. Furthermore, hits that span a large part of the query sequence need to be broken up into smaller groups of HSPs to separate biologically distinct elements such as duplicated genes. Ideally, the structural features of these elements should be taken into account to separate a hit into groups of HSPs. Since the structure of a biological element on a novel sequence cannot be known beforehand, automated separation of these elements on the basis of their structure is difficult to implement. A pragmatic approach is to separate HSPs within a hit on the basis of their physical distance on the query sequence.

Here we describe BlastIf, an application which retrieves all possible database hits from the blastall program, parses these data, and presents the user with a comprehensive output that has a highly reduced redundancy.

## Methods

BlastIf is written in Python [124] and uses the Numeric package [125]. BioPython [126] is used to parse the initial results of blastall before filtering. The application runs on Linux and requires the blastall software to perform the BLAST searches. BlastIf supports blastn,

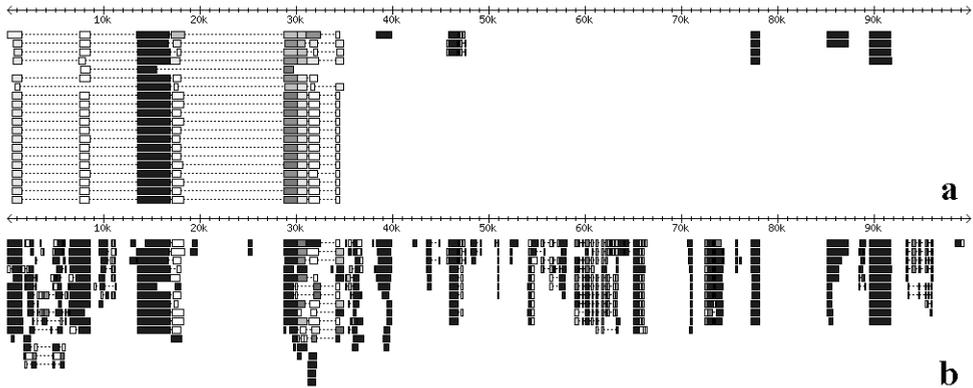
blastp, blastx, tblastn and tblastx, and its output conforms to these formats. It is released under the GNU General Public License version 2 [127].

Upon execution of BlastIf, blastall is run with the parameters supplied to BlastIf, while `-b` and `-v` are set to 10,000,000 (user configurable). In very rare cases where this number is not sufficient to retrieve all available hits from the database, for example when using a very large query sequence (several Mbs) or database, the following approach is taken. Regions of the query sequence that produced a large number of hits are masked, and blastall is iteratively run on the masked input sequence until no new hits can be retrieved. A drawback of this method is that part of the diversity between hits to the same element can be lost, however this only involves the lowest scoring hits.

The alignments produced by blastall are evaluated in order of their statistical significance. To prevent alignments from spanning large parts of the query sequence they are separated into HSP groups, based on the distances between individual HSPs on the query strand. In agreement with the method used by blastall, the expectation value of an HSP group is defined as the lowest expect value of any single HSP within that group. If any HSP group created in this manner has an expectation value larger than the threshold, it is removed from further consideration. A configurable rule system has been implemented that allows the user to control the level of redundancy and variation in the output. Multiple rules can be combined using logical ORs to express different conditions for accepting an HSP group in the output. The rule system depends on a vector that stores the coverage for each nucleotide in the query sequence. Here, the coverage of a nucleotide is defined as the number of HSP groups that have previously been accepted by the rule system at this position in the query sequence. Individual HSPs are accepted or rejected by a rule on the basis of their coverage, and an HSP group is accepted by that rule and included in the output if a user defined number of its HSPs pass the rule.

## **Application**

Figure 2.2 illustrates the power of BlastIf in generating an orderly output for a moderately long genomic sequence. In comparison to the same number of best hits from a regular blastall analysis, BlastIf reveals many additional elements and variation between hits. Repeated elements are separated from each other and the redundancy between hits is strongly reduced. We expect BlastIf to be a valuable tool for molecular biologists who wish to get a quick overview of the genetic elements present in a newly sequenced segment of genomic DNA, prior to more elaborate efforts of gene structure prediction and annotation.



**Figure 2.2:** Comparison between blastall (a) and BlastIf (b) in a blastx analysis of 100 kb of genomic sequence from *A. thaliana* chromosome 2 (GenBank accession NC\_003071.3, bases 30.000 through 130.000) versus the nr database. (a) The blastall output of the 500 best hits contains a large stack of highly similar hits in the region between 0 and 35 kb. Because this region displays a high similarity to a large number of retro-element and polyprotein sequences in the nr database, the rest of the sequence is largely devoid of hits. (b) In contrast, the BlastIf output contains only 281 hits, and it is compacted and comprehensive. The large stack of alignments in the region between 0 and 35 kb has been greatly reduced, and the alignments at this position have been split into distinct smaller ones, allowing for the identification of individual elements. Hits to numerous additional elements in the query sequence can be observed, for example in the region between 59 and 65 kb, which shows significant similarity (1e-99) to several helicase-domain containing proteins. Note that figure (a) has been truncated to show only the first 20 rows of alignments. Images were created using the Graphics module from Bioperl [121].



## Chapter 3

# Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato

*Erwin Datema, Lukas A. Mueller, Robert Buels, James J. Giovannoni, Richard G.F. Visser, Willem J. Stiekema, Roeland C.H.J. van Ham*

A modified version is published in *BMC Plant Biology* (2008), 8:34

### Summary

Tomato (*S. lycopersicum*) and potato (*S. tuberosum*) are two economically important crop species, the genomes of which are currently being sequenced. This study presents a first genome-wide analysis of these two species, based on two large collections of BAC end sequences representing approximately 19% of the tomato genome and 10% of the potato genome.

The tomato genome has a higher repeat content than the potato genome, primarily due to a higher number of retrotransposon insertions in the tomato genome. On the other hand, simple sequence repeats are more abundant in potato than in tomato. The two genomes also differ in the frequency distribution of SSR motifs. Based on EST and protein alignments, potato appears to contain up to 6,400 more putative coding regions than tomato. Major gene families such as cytochrome P450 mono-oxygenases and serine-threonine protein kinases are significantly overrepresented in potato, compared to tomato. Moreover, the P450 superfamily appears to have expanded spectacularly in both species compared to *A. thaliana*, suggesting an expanded network of secondary metabolic pathways in the *Solanaceae*. Both tomato and potato appear to have a low level of microsynteny with *A. thaliana*. A higher degree of synteny was observed with *Populus trichocarpa*, specifically in the region between 15.2 and 19.4 Mb on *P. trichocarpa* chromosome 10.

The findings in this chapter present a first glimpse into the evolution of Solanaceous genomes, both within the family and relative to other plant species. When the complete genome sequences of these species become available, whole-genome comparisons and protein- or repeat-family specific studies may shed more light on the observations made here.

## Introduction

The *Solanaceae*, or Nightshade family, is a dicot plant family that includes many economically important genera that are used in agriculture, horticulture, and other industries. Family members include the tuber bearing potato (*S. tuberosum*); a large number of fruit-bearing vegetables, such as peppers (*Capsicum spp*), tomatoes (*S. lycopersicum*), and eggplant (*S. melongena*); leafy tobacco (*N. tabacum*); and ornamental flowers from the *Petunia* and *Solanum* genera.

Tomato is generally considered to be a model crop plant species, for which many high-quality genetic and genomic resources are available, such as high-density molecular maps [128], many well-characterized near-isogenic lines (NILs), and rich collections of ESTs and full-length cDNAs [129, 130]. Potato is the most important crop within the *Solanaceae*, ranking fourth as a world food crop following wheat, maize and rice. Similar resources are available for potato, including an ultra-high density linkage map [18], a collection of phenotype data [131], and a large transcript database [132]. Like most other nightshades, tomato and potato both have a basic chromosome number of twelve, and there is genome-wide colinearity between their genomes [20].

Much effort is currently being invested to sequence the nuclear and organellar genomes of these organisms. The International Tomato Genome Sequencing Project [133] is sequencing the tomato (*S. lycopersicum* cv. Heinz 1706) genome in the context of the family-wide Solanaceae Project (SOL). Rather than sequencing the complete genome, which is approximately 950 Mb [19], only the gene-rich euchromatic regions (estimated at 240 Mb) are being sequenced using a BAC-by-BAC walking approach [134]. The Potato Genome Sequencing Consortium (PGSC) [135] aims to sequence the complete potato (*S. tuberosum*, genotype RH89-039-16) genome of approximately 840 Mb [18] using a similar marker-anchored BAC-by-BAC sequencing strategy.

Both sequencing projects rely heavily on BAC libraries, of which three exist for tomato (*HindIII* [136], *MboI*, and *EcoRI*) and two exist for potato (*HindIII* and *EcoRI*). The tomato libraries are available through the SOL Genomics Network (SGN) [137] and the potato libraries will soon be available at through the PGSC [135]. All of these libraries have been end-sequenced to support BAC-by-BAC sequencing and extension, and to provide a base of genome-wide survey sequences to support studies such as the one presented here.

This chapter describes a detailed sequence analysis of 310,580 tomato BAC End Sequences (BESs), representing 181.1 Mb (~19%) of the tomato genome, and 128,819 potato BESs, corresponding to 87.0 Mb (~10%) of the potato genome (for an overview of the tomato and potato BES data, see Table 3.1). This comparative genomics study aims to gain insight into the similarity between the tomato and potato genomes, both on the structural level through repeat and gene content analyses and on the functional level through gene function analyses. Furthermore, we investigate micro-syntenic relationships between these two

Solanaceous genomes, and several other sequenced plant genomes. The sequence content of BESs from a particular library is biased by which restriction enzyme was used to make the library. To avoid comparing sequence sets with different biases, tomato-potato comparisons are made only between BESs from libraries made with the same enzyme.

**Table 3.1:** Overview of tomato and potato BES data. The sequences are subdivided into libraries, which are labeled with a three-letter code, with the corresponding restriction enzyme listed between brackets.

	No. sequences	Total length	Average length	GC content
<b>Tomato</b>	<b>310,580</b>	<b>181,076,819</b>	<b>583</b>	<b>36.10%</b>
HBa ( <i>HindIII</i> )	144,307	89,649,564	621	35.50%
Eco ( <i>EcoRI</i> )	77,141	46,398,406	601	35.20%
Mbo ( <i>MboI</i> )	89,132	45,028,849	505	38.30%
<b>Potato</b>	<b>128,819</b>	<b>86,972,687</b>	<b>675</b>	<b>35.60%</b>
POT ( <i>HindIII</i> )	76,930	52,695,698	685	36.00%
PPT ( <i>EcoRI</i> )	51,889	34,276,989	661	35.00%

## Results

### Repeat density and categorization

Based on similarity searches of the repeat database, between 13.0% and 22.9% of the nucleotides in the tomato BESs were identified as belonging to a repeat (see Table 3.2, second through fourth columns). The most common repeat families in the tomato libraries were the *Gypsy* (5.0 – 11.6%) and *Copia* (4.2 – 5.3%) classes of retrotransposons. Another prominent class of repeats comprised the ribosomal RNA genes (<0.1 – 8.6%). The tomato Eco (*EcoRI*) library had the lowest repeat density at 13.0%, which can be attributed to a lower amount of *Gypsy* retrotransposons (5.0%). The highest repeat content was found in the tomato Mbo (*MboI*) library (22.9%), more than a third of which (8.6%) consisted of ribosomal RNA genes. Note that, since the repeat detection was based on sequence similarity, different segments in a BES could be assigned to more than one repeat family. As a result, the sum of the repeat content per repeat type can be slightly larger than the total repeat content.

In contrast to the tomato BESs, only between 10.0% and 12.5% of the nucleotides in the potato BESs showed similarity to known *Magnoliophytæ* repeats (see Table 3.2, fifth and sixth columns). As in tomato, the majority of the repeats were found in the *Gypsy* (5.4 – 8.6%) and *Copia* (2.5 – 2.6%) retrotransposon families, whereas the fraction of ribosomal RNA genes was small (<0.1 – 0.5%). Potato appeared to contain approximately two times as many LINE and SINE elements as tomato (see Table 3.2), although the absolute percentages were low. Furthermore, a higher percentage of class II DNA transposons was

observed in potato (1.0 – 1.2%, versus 0.5 – 0.7% in tomato), the majority of which could not be classified. In agreement with the differences observed between the tomato HBa (*HindIII*) and Eco libraries, the potato PPT (*EcoRI*) library had an overall lower repeat content than the POT (*HindIII*) library, and more specifically, a lower amount of *Gypsy* retrotransposons (5.4% versus 8.6% in the POT library). The PPT library was also enriched in ribosomal RNA genes in comparison to the POT library (0.5% versus less than 0.1%), just as was found comparing the Eco library to the HBa library in tomato.

Since similarity-based repeat detection can be limited by the size and diversity of the repeat database, a self-comparison of the BESs was performed in order to estimate the redundancy within the BESs. Even with the stringent requirement that at least 50% of a given query sequence match another BES with at least 90% identity, 52.0% of the nucleotides in the tomato BESs had a match to one or more other tomato BESs, and 19.0% matched five or more other BESs. The redundancy in the potato BESs was lower than in tomato; 39.0% of the nucleotides in the potato BESs had a hit to at least one other potato BESs, and 12.9% had a hit to five or more BESs. This difference could not be attributed solely to the larger number of tomato BESs, compared to the number of potato BESs; a self-comparison of the tomato HBa library, which is of approximately the same size as the potato POT and PPT libraries combined, showed that 50.7% of the nucleotides in this library matched at least one other HBa BES, and 16.8% matched five or more other HBa BESs. The percentage of nucleotides in both species that matched five or more other BESs was only slightly higher than the findings from the RepeatMasker analysis (see Table 3.2), suggesting that the repeat database used in this study was sufficient to detect the majority of highly abundant repeats in these species. These findings also confirm the observation from the similarity-based repeat detection that the tomato BESs are more repetitive than the potato BESs.

### Simple sequence repeats

A total of 28,423 SSRs with a motif length between one and five nt, and a total length of at least 15 nt were detected in the tomato BESs, representing one SSR per 6.4 kb of genomic sequence. The term ‘motif length’ is used here to describe the length of the motif that is repeated in the SSR; for example, an ATATAT repeat has a motif length of two (with AT being the motif). The most abundant motif length was five nucleotides (11,177 SSRs), followed by motif lengths of two (6,588 SSRs), four (4,596 SSRs), three (4,135 SSRs), and lastly one (1,927 SSRs).

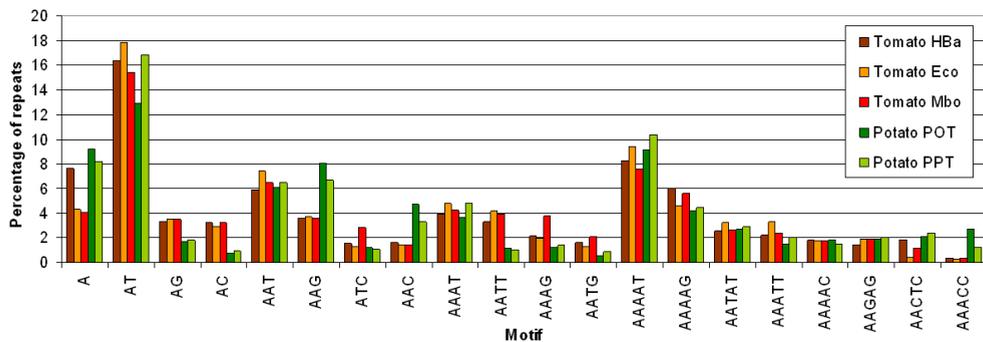
In potato, 19,019 SSRs were found, out of which 3,964 (21%) belonged to class I (i.e., SSRs containing more than 10 motif repeats). Thus, the potato BESs had one SSR per 4.6 kb of genomic sequence, which is higher than that in tomato (one SSR per 6.4 kb). As in tomato, the most abundant motif length in the potato SSRs was five nucleotides (7,922 SSRs). However, the next most abundant length was three (3,941 SSRs), followed by motif lengths of two (3,270 SSRs), four (1,980 SSRs) and one (1,906 SSRs).

**Table 3.2:** Classification and distribution of known plant repeats in the BAC end sequences. Numbers represent percentages of nucleotides that show similarity to a repeat of the indicated category. An 'x' represents the absence of a repeat family; '0.00' indicates that the repeat is present, but at a frequency lower than 0.005 % of the nucleotides in the BESs. Tom.: tomato; Pot.: potato.

	Tom. HBa	Tom. Eco	Tom. Mbo	Pot. POT	Pot. PPT
<b>Class I retrotransposons</b>	<b>16.95</b>	<b>9.30</b>	<b>13.81</b>	<b>11.42</b>	<b>8.19</b>
<b>LTR retrotransposons</b>	<b>16.81</b>	<b>9.19</b>	<b>13.72</b>	<b>11.16</b>	<b>7.92</b>
Ty1/Copia	5.25	4.17	4.39	2.55	2.48
Ty3/Gypsy	11.56	5.02	9.33	8.60	5.43
Unclassified	0.00	0.00	0.00	0.01	0.01
<b>Non-LTR retrotransposons</b>	<b>0.14</b>	<b>0.11</b>	<b>0.09</b>	<b>0.26</b>	<b>0.27</b>
LINE	0.09	0.06	0.05	0.15	0.13
SINE	0.05	0.05	0.04	0.11	0.14
<b>Class II DNA transposons</b>	<b>0.64</b>	<b>0.66</b>	<b>0.49</b>	<b>1.03</b>	<b>1.23</b>
En-Spm	0.26	0.26	0.21	0.27	0.27
Harbinger	0.00	0.00	0.00	0.00	0.00
Mariner	0.00	0.00	0.00	0.00	0.00
MuDR	0.07	0.09	0.05	0.10	0.11
Pogo	0.02	0.03	0.02	0.03	0.08
Stowaway	0.02	0.02	0.02	0.01	0.02
TcMar-Stowaway	x	x	x	0.00	0.00
Tourist	x	x	0.00	0.00	x
hAT	0.02	0.04	0.02	0.05	0.19
hAT-Ac	0.01	0.00	0.01	0.01	0.01
hAT-Tip100	0.02	0.02	0.02	0.11	0.10
Unclassified	0.22	0.20	0.14	0.45	0.45
<b>Satellites</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.04</b>	<b>0.03</b>
Centromeric	0.00	x	0.00	0.00	0.00
Subtelomeric	x	x	x	0.00	0.00
Unclassified	0.00	0.00	0.00	0.04	0.03
<b>Ribosomal genes</b>	<b>0.04</b>	<b>2.98</b>	<b>8.58</b>	<b>0.03</b>	<b>0.53</b>
rRNA	0.04	2.98	8.58	0.03	0.53
<b>Unclassified</b>	<b>0.08</b>	<b>0.11</b>	<b>0.07</b>	<b>0.07</b>	<b>0.11</b>
Centromeric	x	x	x	0.00	x
Composite	x	x	x	x	0.00
RC/Helitron	0.08	0.11	0.07	0.06	0.11
Unknown	0.00	0.00	0.00	0.01	0.00
<b>Total</b>	<b>17.66</b>	<b>13.01</b>	<b>22.91</b>	<b>12.54</b>	<b>10.02</b>

Figure 3.1 shows the distribution of the primary SSR motifs in the tomato and potato BESs, ordered by motif length and relative frequency within the motifs of the same length. The most abundant SSR motifs in both datasets were AT-rich, with the di-nucleotide repeat AT/TA being the most abundant (16.6% of all tomato and 14.7% of all potato SSRs, respectively). Several motifs, such as AG/CT, AC/GT, AATT/AATT and AAAG/CTTT were more frequent in tomato than in potato, whereas other motifs, such as AAG/CTT, AAC/GTT, AACTC/GAGTT and AAACC/GGTTT were found predominantly in potato.

Considering only the class I SSRs, the most abundant SSR motifs in tomato and potato were AT/TA (50.8 and 39.1% of all class I SSRs, respectively) and A/T (25.8 and 42.1%). In tomato, the di-nucleotide motifs AC/GT (6.3%) and AG/CT (5.7%) were the most abundant after these two, whereas in potato the mononucleotide C/G (6.0%) and trinucleotide AAT/ATT (4.5%) and AAG/CTT (3.7%) occurred at the second, third and fourth highest frequency, respectively. This suggests that the differences in primary motif frequencies between tomato and potato also hold when considering only class I SSRs



**Figure 3.1:** Distribution of the most abundant SSR motifs in the tomato and potato BESs. The values on the Y axis represent the fraction of SSRs for each dataset that consist of the motifs listed on the X axis.

### Gene content

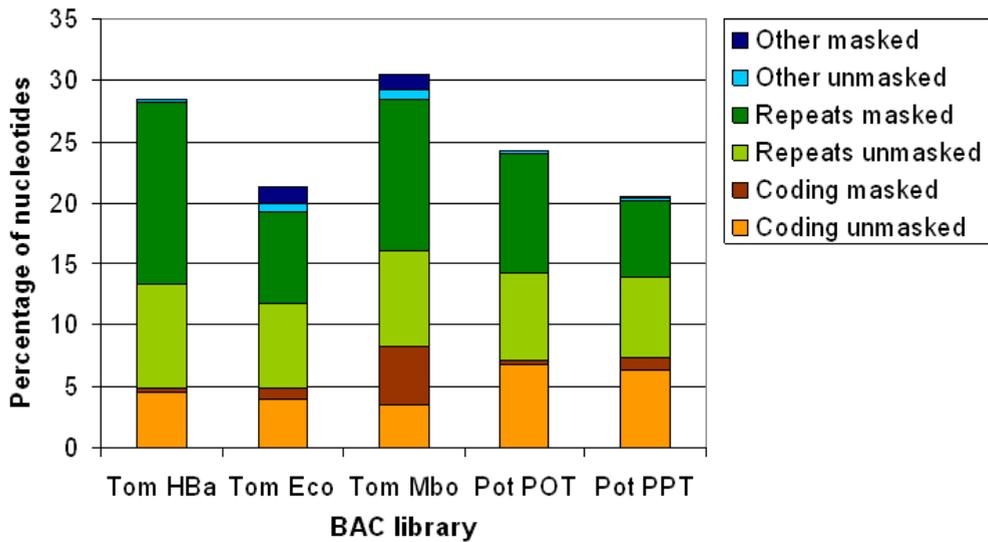
In the tomato BESs, the percentage of nucleotides that matched by at least one database sequence ranged from 21.3% for the Eco library, to 30.5% for the Mbo library. Figure 3.2 presents a breakdown of these BLAST hits into three main categories ('coding', 'repeats', and 'other'), based on the keyword filtering described in Materials and Methods. Each category was then subdivided into 'masked' and 'unmasked' subcategories, with 'masked' indicating an overlap with repetitive sequences identified by RepeatMasker, and 'unmasked' indicating a lack of such overlap. In this way, the BLAST and RepeatMasker results were combined in order to generate the best possible estimation of the percentage of

putative protein-coding nucleotides in the BESs. The 'coding' category represents the percentage of nucleotides that matched one or more database sequences, and were not identified as repetitive by the keyword filtering. After removing the overlap with repeats identified by RepeatMasker, the percentage of coding nucleotides in the three libraries ranged from 3.5% for the Mbo library to 4.6% for the HBa library (the 'coding unmasked' category in Figure 3.2). The Mbo library had the highest percentage of the three libraries in the 'coding masked' category, which is likely the result of the high number of ribosomal repeat sequences in this library that have escaped the keyword filtering. The 'repeats' category contains the BLAST matches to transposon and other repeat related sequences. In all three libraries, there was a considerable fraction of nucleotides that the keyword filtering assigned to the 'repeats' category but that did not overlap with the repeats identified by RepeatMasker (i.e. the 'repeats unmasked' category). This fraction ranged from 6.9% in the Eco library to 8.4% in the HBa library and may represent a combination of repeats that were missed by RepeatMasker and true protein-coding genes that were miss-classified by the keyword filtering. The final category in Figure 3.2, 'other', represents all non-transposon-related repetitive sequences that were identified by the keyword filtering.

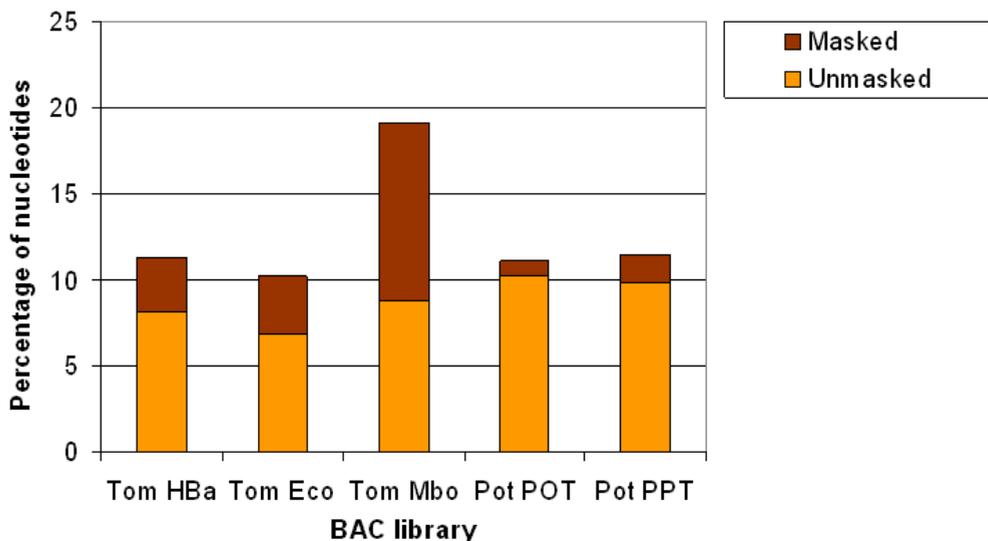
In the potato POT and PPT libraries, 24.3 and 20.5% of the nucleotides matched the protein database, respectively. While these numbers were slightly lower than those for the tomato HBa and Eco libraries (28.5 and 21.3%, respectively), the percentage of nucleotides assigned to the 'coding' category (6.8 and 6.3%) was larger than those of the corresponding tomato libraries (4.6 and 3.9%), suggesting that potato may have a larger gene repertoire than tomato. Furthermore, the number of transposon regions and other repeat-related regions that was found in this comparison to the protein database was more than 1.5-fold higher for tomato than for potato. This is consistent with the difference in transposon content that was found in the repeat analysis.

Figure 3.3 shows the results of the BLASTN comparison of the BESs to species-specific EST databases. The matches were divided into two categories, 'masked' and 'unmasked'. The 'masked' category contains the nucleotides that had a match in the EST database, but were found to be repetitive in the RepeatMasker analysis; the 'unmasked' category contains the nucleotides that did not overlap with repeats. In the tomato libraries, between 10.2 and 19.1% of the nucleotides matched one or more tomato EST sequences. The Mbo library had the highest EST coverage (19.1%), but more than half of these matches (10.3%) were 'masked'. The percentage of nucleotides in the 'unmasked' category ranged from 6.8% in the Eco library to 8.8% in the Mbo library.

For the potato BESs, 11.1% (POT) and 11.5% (PPT) of the nucleotides had match in the potato EST database, which is in fairly good agreement with the tomato HBa and Eco comparisons versus the tomato database (11.3 and 10.2%, respectively; see also Figure 3.3). Fewer matches in the potato BESs were 'masked' than in tomato, confirming the observation from the BLASTX comparison to the protein database that the potato BESs have more protein coding nucleotides and lower repeat content.



**Figure 3.2:** Percentage of nucleotides in the BESs covered by BLASTX hits to the non-redundant protein database. The BLAST hits have been divided into three categories ('coding', 'repeats', 'other') based on keyword filtering. Each category has subsequently been divided into 'masked' (i.e., overlapping with repeats identified by RepeatMasker) and 'unmasked' (i.e., no overlap with repeats identified by RepeatMasker) subcategories. Tom.: tomato; Pot.: potato.



**Figure 3.3:** Percentage of nucleotides in the BESs covered by BLASTN hits to the species-specific transcript databases. The BLAST hits have been divided into 'masked' (i.e., overlapping with repeats identified by RepeatMasker) and 'unmasked' (i.e., no overlap with repeats identified by RepeatMasker) categories. Tom.: tomato; Pot.: potato.

## Functional annotation

A total of 30,335 GO terms, out of which 585 unique terms, were assigned to the tomato HBa BESs based matches in the Pfam database. Although there were more than half as many Eco BESs as HBa BESs, only 7,647 GO terms (403 unique terms) were assigned to them. In potato, 17,060 terms (544 unique terms) were assigned to the POT library, whereas only 9,312 terms (419 unique terms) were assigned to the PPT library. Comparing the GO annotations of tomato to those of potato (for libraries generated with the same restriction enzyme) resulted in 18 significantly overrepresented terms between the *HindIII* digested libraries (seven in tomato HBa, and eleven in potato POT) and nine significantly overrepresented terms between the *EcoRI* digested libraries (seven in tomato Eco, and two in potato PPT).

In both species, many of the terms that were overrepresented in the *HindIII* libraries compared to their *EcoRI* counterparts were related to retrotransposon activity, such as DNA binding (GO:0003677), DNA integration (GO:0015074), RNA-directed DNA polymerase activity (GO:0005634), and chromatin-related terms (GO:0000785, GO:0003682, GO:0006333). Furthermore, many of these transposon-related terms were significantly overrepresented in tomato, compared to potato (P value < 10<sup>-4</sup>). This is consistent with the findings from the RepeatMasker and BLAST analyses discussed above. Surprisingly, some terms that were overrepresented in both the *EcoRI* digested libraries could be linked to transcription factor genes. In tomato, zinc ion binding (GO:0008270), DNA-dependent regulation of transcription (GO:0006355), and transcription factor activity (GO:0003700) were overrepresented in the Eco library. The potato PPT library was enriched for zinc ion binding (GO:0008270), nucleic acid binding (GO:0003676), and transcription factor activity (GO:0003700).

Analysis of the protein families identified by PANTHER revealed similar trends for the number of matches, both within and between the tomato and potato libraries. In tomato, 1,064 distinct families were found in the HBa BESs for a total of 28,984 hits, and 8,226 hits representing 654 families were found in the Eco BESs. Analysis of the potato POT library revealed 951 distinct PANTHER families for a total of 13,821 hits; however, only 6,926 hits to 716 families were found in the PPT BESs. Two and three PANTHER families were found to be overrepresented in the tomato HBa and Eco libraries, compared to eleven and five overrepresented families in the potato POT and PPT libraries, respectively.

Consistent with the greater abundance of *Gypsy* retrotransposons in the *HindIII* libraries of both tomato and potato, the GAG/POL/ENV polyprotein (PTHR10178) PANTHER family was found to be overrepresented in both *HindIII* libraries, compared to the corresponding *EcoRI* libraries. Furthermore, the GAG-POL-related retrotransposon (PTHR11439) PANTHER family was relatively more abundant in the *EcoRI* libraries, which also agrees with the difference in the *Gypsy:Copia* ratio between the *HindIII* and *EcoRI* libraries (see also Table 3.2). Both of these retrotransposon-related terms were found to be significantly

( $P$  value  $< 10^{-4}$ ) overrepresented in tomato when compared to potato. In the tomato Eco library, transcription-factor related terms such as zinc finger CCHC domain contain protein (PTHR23002), zinc finger protein (PTHR11389) and MADS box protein (PTHR11945) were significantly overrepresented ( $P$  values  $4.0 \times 10^{-13}$ ,  $7.8 \times 10^{-7}$ , and  $1.5 \times 10^{-6}$ , respectively), confirming the results from the GO analysis. No transcription-factor related PANTHER families were significantly overrepresented in the potato PPT library.

Between tomato and potato, the majority of the overrepresented terms in potato corresponded to important biological and biochemical processes. For example, zinc finger CCHC domain containing proteins (PTHR23002) and general transcription factor 2-related zinc finger proteins (PTHR11697) occurred with a significantly ( $P$  value  $2.2 \times 10^{-16}$  for both) higher frequency in potato POT than in tomato HBa; the latter was also overrepresented in the potato PPT library. This was also reflected in the GO annotation through terms such as nucleic acid binding (GO:0003676) and zinc ion binding (GO:0008270). The overrepresentation of these terms relative to tomato suggests an expansion of transcription factors or other genes for DNA binding proteins in the potato genome.

Another example is the cytochrome P450 superfamily (PTHR19383), which was also found in the GO analysis through terms such as iron ion binding (GO:0005506) and mono-oxygenase activity (GO:0004497). Cytochrome P450 proteins play important roles in the biosynthesis of secondary metabolites, and the overrepresentation of these proteins in potato could indicate an expanded network of pathways that synthesize secondary metabolites in potato.

A final example involves the large family of plant-type serine-threonine protein kinases (PTHR23258), which are known to play important roles in disease resistance in various plant species (for example, the Pto gene in tomato [138]). In the PANTHER database, this family consists of 104 different subfamilies, 71 of which were found in the tomato and potato BESs. Out of these 71 subfamilies, 15 were found only in tomato, and five were unique to potato. Most of the subfamilies that were found in both species were overrepresented in potato, such as LRR receptor-like kinases (PTHR23258:SF462) and LRR transmembrane kinases (PTHR23258:SF474). Several subfamilies occurred at a higher frequency in tomato, including serine/threonine specific receptor-like protein kinases (PTHR23258:SF416) and Pto-like kinases (PTHR23258:SF418). Thus, while the complement of serine-threonine protein kinases in potato exceeds that of tomato, several of the subfamilies have expanded specifically in tomato. This may reflect an adaptation for resistance to different pathogens, or a difference in the dominant mechanism of pathogen resistance between these species.

### Comparative genome mapping

Out of the 135,842 pairs of tomato BESs that were compared to the *A. thaliana* genome, 15,283 pairs had one or more matches. These matches were divided into five categories, as is shown in the last five columns of Table 3.3. The 'single end' category represents the

BAC end pairs from which only one of the two sequences had a match to the *A. thaliana* genome, and contained the majority of the matches (10,191). Paired end matches, in which the BESs from the same BAC each had a match to a different chromosome, were assigned to the ‘non-linear’ category. The ‘gapped’ category contained 4,836 BAC end pairs that matched to the same *A. thaliana* chromosome with a distance between the paired matches that was either smaller than 50 kb or larger than 500 kb. The final two categories represented the BACs from which both end sequences were matched to the genome within a distance of 50 to 500 kb of each other, either in the correct orientation with respect to each other (‘colinear’), or rearranged with respect to each other (‘rearranged’). Out of the 4,840 tomato BES pairs that hit to the same *A. thaliana* chromosome, three pairs fell into the ‘colinear’ category, and one pair fell into the ‘rearranged’ category, suggesting the presence of four putative micro-syntenic regions between tomato and *A. thaliana*.

**Table 3.3:** BLASTN hits between the tomato and potato BESs, and the *A. thaliana* genome.

	No hit	Single end	Non-linear	Gapped	Colinear	Rearranged
<b>Tomato</b>	<b>120,559</b>	<b>10,191</b>	<b>252</b>	<b>4,836</b>	<b>3</b>	<b>1</b>
HBa	57,489	5,469	159	50	1	1
Eco	30,529	1,655	33	1,279	2	0
Mbo	32,541	3,067	60	3,507	0	0
<b>Potato</b>	<b>51,361</b>	<b>4,102</b>	<b>82</b>	<b>115</b>	<b>1</b>	<b>1</b>
POT	31,568	2,718	57	18	1	0
PPT	19,793	1,384	25	97	0	1

**Table 3.4:** BLASTN hits between the tomato and potato BESs, and the *P. trichocarpa* genome.

	No hit	Single end	Non-linear	Gapped	Colinear	Rearranged
<b>Tomato</b>	<b>110,633</b>	<b>18,904</b>	<b>5,597</b>	<b>635</b>	<b>51</b>	<b>22</b>
HBa	52,083	10,297	666	68	38	17
Eco	28,630	3,341	1,174	344	6	3
Mbo	29,920	5,266	3,757	223	7	2
<b>Potato</b>	<b>46,189</b>	<b>8,844</b>	<b>554</b>	<b>34</b>	<b>24</b>	<b>17</b>
POT	28,116	5,899	300	19	17	11
PPT	18,073	2,945	254	15	7	6

Potato had 55,662 pairs of BESs, out of which 117 pairs were mapped to the *A. thaliana* genome, with both BESs of the pair matching the same chromosome. Two potato BACs displayed putative microsynteny based on the end sequence matches, one of which was colinear, whereas the other represented a possible rearrangement. In comparison to tomato,

potato had very few BACs that fell into the ‘gapped’ category, although the smaller PPT library had more than five times as many sequences in this category as the POT library. Interestingly, the large majority of the tomato BACs that fell into this category was from the Eco and Mbo libraries (1,279 and 3,507, respectively). The *EcoRI* and *MboI* digested libraries were found to contain a high fraction of ribosomal RNA genes in the RepeatMasker analysis, and indeed more than 80% of the sequences from these libraries that fell into the ‘gapped’ category contained ribosomal RNA genes.

Repeating the same analysis against the *P. trichocarpa* genome, only 708 of the tomato BES pairs matched with both ends to the same chromosome (the sum of the last three columns in Table 3.4). It should be noted here that *P. trichocarpa* has both a larger number of chromosomes than *A. thaliana* (19 versus 5) and approximately twenty-two thousand additional contig sequences that have not yet been integrated into the chromosome pseudomolecules. Based on these numbers alone, one would expect a smaller number of paired BESs to map to the same chromosome or contig sequence. Even so, *P. trichocarpa* displayed more regions of micro-synteny with tomato than *A. thaliana*: 73 pairs of BESs mapped within a distance between 50 and 500 kb of the other BES in the pair. More than two-thirds of these matches (51, the ‘colinear’ category in Table 3.4) showed colinearity between tomato and *P. trichocarpa*, whereas the remaining 22 hits represented rearrangements in their respective regions of micro-synteny.

Consistent with the difference between the tomato – *A. thaliana* and tomato – *P. trichocarpa* mappings, a smaller number of potato BES pairs (75) could be mapped with both ends to the same chromosome in *P. trichocarpa*, than in *A. thaliana*. Of these, there were 41 regions of potential microsynteny, out of which 24 were colinear. Compared to tomato, the ‘non-linear’ and to a lesser extent the ‘gapped’ categories were underrepresented in potato. Again these differences seem to originate from the fact that many of the BESs in the Eco and Mbo libraries contain ribosomal RNA genes. The majority of these sequences fell into the ‘non-linear’ category in the *P. trichocarpa* comparison, rather than the ‘gapped’ category as was the case with *A. thaliana*, due to the ribosomal RNA genes being contained in some of the unassembled contig sequences rather than in the chromosomal pseudomolecules.

## Discussion

### Sequence properties

Based on the differences between the libraries in both tomato and potato, it seems unlikely that any of these partial digestion-based libraries represents an unbiased cross section of the genome. For example, in tomato the Mbo library has a higher GC percentage than the HBA and Eco libraries. This difference is likely caused by the length and GC content of the restriction sites that were targeted in the digestion of the genome: both the *HindIII* and

*EcoRI* sites (AAGCTT and GAATTC, respectively) have a length of six nucleotides and a GC content of 33.3%, whereas the *MboI* site (GATC) has a length of four nucleotides and a GC content of 50%. The consequences of this are clearly visible in the results of the gene and repeat content analyses presented in this chapter: results differ markedly among libraries made with different enzymes. However, we think it reasonable to assume that tomato and potato libraries derived from digestion with the same restriction enzyme would have similar sequence bias. Using this assumption, we strive to minimize any effect of sequence bias on our results by maintaining logical separation of BESs from different libraries, and only directly comparing data for BESs from libraries constructed with the same restriction enzymes.

The tomato BESs (and specifically the *MboI* BESs) are shorter than the potato BESs on average. The difference in average sequence length between the tomato *HindIII* and *EcoRI* libraries and their potato counterparts is approximately 60 nt for both libraries and is most likely the result of a difference in sequencing quality and equipment. However, we think it reasonable to assume that a difference in sequence length on this scale would not influence the results of the similarity-based analyses that have been performed in this study.

#### Repeat density and categorization

Both the tomato and potato libraries vary in total repeat content and in ratios between repeat types. For example, ribosomal DNA sequences are overrepresented in the tomato *MboI* and *EcoRI*, and the potato PPT libraries, relative to the tomato *HBa* and potato *POT* library, respectively. This phenomenon was also observed in a study of *Zea mays* BESs [139], where it was attributed to the presence of many *MboI* sites in the *Z. mays* ribosomal DNA cluster, compared to one *EcoRI* site, and no *HindIII* sites. By similar reasoning, the under-representation of *Gypsy* retrotransposons in the *EcoRI* and PPT libraries might result from a lower frequency of *EcoRI* sites in this element compared to *HindIII* and *MboI* sites.

The discrepancy between the repeats identified by RepeatMasker (Table 3.2) and BLASTX (Figure 3.2) indicates the need for tomato- and potato-specific repeat databases. A repeat database had previously been generated from the tomato BESs (L. Mueller, unpublished data), however comparing the tomato BESs to this database using RepeatMasker resulted in approximately 60% of the tomato BESs being annotated as repetitive (data not shown). The majority of these repeats could however not be assigned to a known repeat family. Thus, while the findings in this chapter may present an underestimation of the actual repeat content of the tomato and potato BESs, the findings from the RepeatMasker and BLASTX analyses both clearly suggest a higher repeat content in the tomato BESs than in the potato BESs.

A correlation between genome size and retrotransposon content has previously been identified in the *Brassicaceae* [140]. There, it was found that the retrotransposon content increases with genome size, from approximately 7 to 10% in *A. thaliana* (genome size 125 Mb), to 14% in *Brassica rapa* (genome size 530 Mb), to 20% in *B. oleracea* (genome size

700 Mb). Comparing this to cereal crops such as *Oryza sativa* (genome size 430 Mb, 35% retrotransposons [141]) and *Z. mays* (genome size 2,365 Mb, 56% retrotransposons [139]) suggests that while the actual retrotransposon content in cereals is higher than in *Brassicaceae*, the correlation with genome size may be universally present in plants. The data presented in this research indicate that genome expansion in the *Solanaceae* is also associated with retrotransposon amplification; potato (genome size 840 Mb) has an estimated retrotransposon content between 8.2 (PPT) and 11.4% (POT), whereas that of tomato (genome size 950 Mb) is notably higher (9.3% for the Eco library, and 17.0% for the HBa library).

The ratio between *Gypsy* and *Copia* retrotransposon sequences in the tomato BESs is between 1:1 and 2:1, whereas this ratio in the potato BESs is between 2:1 and 3:1. While this ratio clearly differs within each species between libraries generated with a different restriction enzyme, the difference in ratios between tomato and potato is observed in both the *HindIII* and the *EcoRI* digested libraries (see Table 3.2). In *A. thaliana* [142], *B. rapa* [140], *Carica papaya* [143] and *Z. mays* [139], this ratio is approximately 1:1. The tomato and potato genomes appear more similar to the *O. sativa* genome in this respect, where the *Gypsy* to *Copia* ratio was found to be around 2:1 [141]. The difference in the *Gypsy:Copia* ratio between tomato and potato suggests that the retrotransposon amplification associated with the genome expansion in tomato is predominantly the result of additional *Copia* copies.

### Simple sequence repeats

The most abundant SSRs in all size categories for both tomato and potato were AT-rich. This is consistent with findings in other plant species, such as *A. thaliana* [144], *B. rapa* [140], *C. papaya* [143], *Glycine max* [145], and *Musa acuminata* [146]. In both potato and tomato, penta-nucleotide repeats are the most common form of SSRs, and AAAAT is the predominant repeat motif. This is in sharp contrast to previously studied plant species, in which di- and penta-nucleotide repeats generally occur least frequently [147]. In many plant species, such as *A. thaliana*, *B. rapa* [140], and *O. sativa* [10, 11], tri-nucleotide repeats are the most abundant microsatellites. However, BES analysis of *C. papaya* [143], *G. max* [145] and *M. acuminata* [146] suggests that di-nucleotide repeats are more common in these plant species. Thus, both tomato and potato display a unique distribution of microsatellite frequencies compared to other studied plant species.

The tomato BESs have a higher fraction of di- and tetra-nucleotide repeats compared to the potato BESs. This may be because one or more of the tomato BAC end libraries are enriched for BACs that are derived from centromeric regions in the tomato genome, as these regions have previously been found to be enriched for long, class I di- and tetra-nucleotide repeats [148]. However, the relative enrichment for di- and tetra-nucleotide repeats in tomato compared to potato is observed in all three tomato libraries; this would

only be compatible with the hypothesis of enrichment for centromeric regions if these regions contain more *HindIII*, *EcoRI* and *MboI* sites than average for the tomato genome.

### Gene content

After repeat masking and keyword filtering, the percentage of nucleotides in the potato POT and PPT BESs that have a match in the non-redundant protein database is 1.5- to 1.6-fold that of the tomato HBa and Eco BESs, respectively. Both the percentage of nucleotides and the number of BESs having a hit to the protein database after repeat masking and keyword filtering are higher in potato (13.8% in the POT library; 12.9% in the PPT library) than in tomato (8.7% in the HBa library; 7.9% in the Eco library), supporting the hypothesis that potato has more putative protein-coding regions than tomato. In the BLASTN comparison of the BESs to the ESTs, a similar discrepancy between potato and tomato was observed, with potato having a 1.3- to 1.4-fold higher EST coverage than tomato. Furthermore, cross-comparisons of the tomato BESs to the potato ESTs and vice versa confirmed that the difference in EST coverage of the BESs was not caused by a difference in number of unique transcripts between the tomato and potato EST collections (data not shown). The difference between the BLAST comparisons to the protein and transcript databases may be attributed to the presence of full-length cDNA sequences in the tomato transcript data, whereas these are not present in the potato data, resulting in an overrepresentation in the tomato BESs for the interior regions of coding sequences. Even if one assumes that this more conservative lower bound is correct, the results still suggest that potato has a larger gene repertoire than tomato since the tomato genome is only approximately 1.1 times larger than the potato genome.

In both tomato and potato, a smaller percentage of nucleotides show similarity to the EST database than to the protein database, while the percentage of non-repetitive coding sequence in the EST database comparison (the 'unmasked' category in Figure 3.3) is higher than that in the protein database comparison (the 'coding unmasked' category in Figure 3.2). Surprisingly, the majority of the matches to the protein and transcript databases do not overlap. For example, in the tomato HBa library, 8.1% and 4.6% of the nucleotides have a match in the EST and protein databases, respectively, while only 1.6% have a match in both. Similarly, for the potato POT library, only 2.5% of the nucleotides have a match in both the transcript and protein sequences, whereas the individual percentages of nucleotides that have a match in these databases are 10.2% and 6.8%, respectively. On one hand, the matches to the EST databases that do not overlap with matches to the protein database may represent unique, taxon- or species-specific protein-coding genes that are not represented in the non-redundant protein database, or transcribed but untranslated regions in these genomes. On the other hand, matches to the protein database that do not overlap with matches in the EST database may indicate either the presence of genes that were not sufficiently expressed in the tissues under the conditions that were sampled during EST library construction, or mis-annotated or otherwise incorrect sequences in the protein database.

The EST data likely provides the most reliable sampling of the true protein coding regions in these genomes, since it is based on experimental data that contain species-specific sequences not available in the protein database. Due to the selection for poly-A tails that is normally used in the construction of EST libraries, the number of non-protein coding transcripts will be relatively small. Taking the nucleotides from the HBa and Eco libraries that match ESTs and do not overlap with repeats as a measure of coding sequences, the tomato genome (950 Mb) is estimated to contain between 64.8 and 77.1 Mb of coding regions. Similarly, assuming a genome size of 840 Mb, the total coding region length for potato would be between 82.5 and 85.4 Mb. These numbers set lower bounds on the estimated coding content of these genomes, as the EST data is unlikely to represent the full complement of full-length protein-coding sequences in these genomes.

Previous estimates put the gene content of tomato at 35,000 genes, based on an analysis of 27,274 UniGenes and 6 BAC sequences [149]. If these 35,000 genes are indeed represented by 71.0 Mb of coding sequence (the average of the estimations for the HBa and Eco libraries), then the average transcript length of tomato would be approximately 2.0 kb. This is longer than the average transcript length in *A. thaliana*, which is 1.2 kb according to the TAIR7 genome annotation [150]. Assuming the same average transcript length, potato (84.0 Mb of coding sequence, averaged over the two libraries) would contain approximately 41,400 genes, or 6,400 more genes than tomato. Since the data presented here are based on similarity-searches on short genomic sequences only, this difference does not necessarily represent a difference in functional genes, but may also reflect a larger proportion of pseudogenes or otherwise non-functional alleles in potato.

### Functional annotation

The results from the GO and PANTHER analysis generally show a similar trend. The tomato BESs have more GO terms and PANTHER families associated to them than the potato BESs do. However, the potato BESs have a larger number of unique terms associated to them. This agrees with the results of the BLASTX comparison to the non-redundant protein database, where it was found that the tomato BESs have a higher overall coverage of BLAST hits, but a lower percentage of putative protein coding regions (see also Figure 3.2).

In both the GO term and PANTHER family analyses, the majority of the terms occur at a relatively low frequency. For example, in the tomato HBa versus potato POT comparison, only 131 out of the 730 distinct GO terms that were assigned to the BESs occurred ten or more times in at least one of the species. This group of 131 GO terms contained all 18 of the terms that were significantly ( $P$  values  $< 10^{-4}$ ) overrepresented in one of the species in this comparison. Moreover, 39 out of these 131 terms were found at least 50 times in at least one species, and this subgroup contained 16 out of the 18 significantly overrepresented terms. Similarly, in the PANTHER family analysis, 119 out of the 1,352 distinct families that were found in the BESs occurred at least ten times in at least one

species, out of which 12 families were found at least 50 times. The 119 families that were found at least ten times contained every one of the 13 families that were significantly overrepresented in one of the species; ten of these were counted more than 50 times in at least one species. While only the tomato HBa versus potato POT comparison is shown here, the other comparisons show a similar pattern, indicating that many of the highly abundant GO terms and PANTHER families are significantly overrepresented in either tomato or potato. The majority of these overrepresented terms and families are most abundant in potato, and represent biologically important functions and processes. In tomato, a smaller number of terms and families is overrepresented; these are primarily connected to structural genomic features such as retrotransposons.

The overrepresentation of transposon-related GO terms and PANTHER families in tomato was consistent with the results of the repeat analysis, confirming the observations that tomato is richer in retrotransposons than potato. However, in the PANTHER analysis, reverse transcriptases (PTHR19446) were significantly overrepresented in potato. At first glance, this does not correspond well with the overrepresentation of RNA-directed DNA polymerase activity (GO:0003964) and RNA-dependent DNA replication (GO:0006278) in tomato. However, in both tomato and potato, the large majority of the reverse transcriptases originated from non-LTR retroelements (PTHR19446:SF34), which in fact is consistent with the higher frequency of non-LTR retrotransposons in potato found in the RepeatMasker analysis (see also Table 3.2).

The cytochrome P450 mono-oxygenases represent a large gene superfamily in plants that are commonly associated with the biosynthesis of secondary metabolites. In *A. thaliana*, at least 272 P450 genes have been found, representing approximately one percent of the gene complement of this species. In *O. sativa*, this family is even larger, with 458 P450 genes identified so far [151]. Not all the P450s in these genomes represent true protein-coding sequences; in *A. thaliana*, 90% of the genes are truly protein coding, compared to 72% in *O. sativa*. In total, 66 distinct families of P450 genes were identified in *A. thaliana* and *O. sativa*, several of which were found to be overrepresented in either of these species. Moreover, some families were present in one, but completely absent in the other species [152]. In the *Hind*III and *Eco*RI libraries, 186 and 209 BESs that could be associated with the cytochrome P450 PANTHER family (PTHR19383) were found in tomato and potato, respectively. Since these BAC end sequences represent approximately 14% and 10% of their respective genomes, these data suggest an enormous expansion of P450 genes in the *Solanaceae*. This could be the result of an expansion of specific P450 families, but also of the evolution of species- or family-specific P450s. For example, the allene oxide synthase has currently only been found in Solanaceous species, including tomato and *Petunia inflata* [153]. The overrepresentation of P450s in potato compared to tomato may be another result of species-specific P450 families, but may also indicate expansion of families that are shared between these species.

## Comparative genome mapping

In this study, paired BAC ends have been exploited to detect regions of microsynteny between the Solanaceous species tomato and potato, and the model plant organisms *A. thaliana* and *P. trichocarpa*. Using similar approaches, microsynteny has been observed between *A. thaliana* and *B. rapa* [140]; *C. papaya* and *P. trichocarpa* [143]; and *M. acuminata* and *O. sativa* [146].

A higher number of tomato and potato BACs display microsynteny to *P. trichocarpa*, than to *A. thaliana*. The reduced level of microsynteny between tomato/potato and *A. thaliana* is not likely a difference in evolutionary distances between these species. Both *A. thaliana* and *P. trichocarpa* are part of the rosids clade, whereas tomato and potato belong to the asterids clade. It may be the result of a recent duplication of the *A. thaliana* genome, followed by the loss of roughly 70% of the duplicated genes [154]. Assuming that this loss occurred randomly, the large majority of possible microsyntenic regions that existed before the duplication will have disappeared due to the major local expansions and contractions that would be associated with such a duplication and subsequent loss. This hypothesis is strengthened by the observation that only approximately 1% of 12,000 *A. thaliana* BES pairs could previously be mapped within 300 kb to the *P. trichocarpa* genome, indicating that the organization of these genomes is indeed vastly different [143].

Regions of microsynteny have previously been detected between tomato/potato and *A. thaliana*. A 57 kb region of tomato chromosome 7 containing five genes was shown to be syntenic with a 30 kb region on *A. thaliana* chromosome 1, although the order and orientation of the genes suggested two inversion events [155]. In another study, a 105 kb BAC sequence matched to four different segments on *A. thaliana* chromosomes 2, 3, 4, and 5; however, each of the four regions in *A. thaliana* were shorter than their tomato counterpart [156]. Recently, five microsyntenic blocks were detected between a region on potato chromosome 5 harbouring a QTL for resistance to late blight and root cyst nematodes, and *A. thaliana* chromosomes 1, 3 and 5 [157]. These syntenic blocks spanned between three and seven ORFs, and were interrupted by non-syntenic blocks. In each of these examples, the microsynteny between tomato/potato and *A. thaliana* involves shorter regions on the *A. thaliana* genome than the average tomato and potato BAC sequence length. Furthermore, regions of (micro-)synteny are often detected between coding sequences, whereas the fraction of coding sequences in the tomato and potato BESs is relatively low (< 10%), providing a good explanation for the reduced amount of microsynteny between these species observed here.

Synteny between potato and *A. thaliana* has also been identified on a genome-wide level using a comparative mapping approach. This revealed 90 putative syntenic blocks between potato and *A. thaliana* that cover 41% of the potato genetic map, and 50% of the *A. thaliana* physical map [158]. These syntenic blocks were unevenly distributed over the potato genetic map, and redundant in respect to the number of areas on the *A. thaliana*

genome that displayed synteny to most areas on the potato map. The regions of microsynteny between tomato/potato and *A. thaliana* that were found with the BES-based approach described in this study could not be used to confirm or renounce any putative higher-order syntenic regions, due to the relatively short distances between the BAC ends.

Six paired tomato BAC end matches cluster in the 16.0 – 20.2 Mb interval of *P. trichocarpa* chromosome 10. Furthermore, seven pairs of potato BESs map to the partially overlapping interval between 15.2 – 19.4 Mb, indicating the presence of either a number of distinct microsyntenic regions, or possibly a single region of macrosynteny, between the tomato/potato and *P. trichocarpa* genomes. These findings provide an interesting starting point for a detailed comparison between these species in this region, once more tomato and potato genomic sequences become available.

## Conclusion

The large scale analysis of tomato and potato BESs presented in this chapter revealed many interesting structural and functional differences between the genomes of these closely related species. We have shown that the tomato genome is not only more repetitive than the potato genome, but that these genomes also differ in their repeat composition, most importantly in the distribution of *Gypsy* and *Copia* retrotransposons. In contrast to other studied plant genomes, we have shown that the tomato and potato genomes contain a large number of SSRs with a motif length of five, which may be a unique feature of Solanaceous genomes.

Comparative analysis of the putative protein coding regions in these BESs revealed an enrichment of these regions in the potato genome. Moreover, several protein families were found to be overrepresented in potato compared to tomato, such as cytochrome P450 monooxygenases and serine-threonine protein kinases. The P450 superfamily appears to have expanded dramatically in both species compared to *A. thaliana*, suggesting an expanded network of secondary metabolic pathways in the *Solanaceae*.

Both tomato and potato appear to have low microsynteny with *A. thaliana*, which is likely a result of this species' relatively recent genomic rearrangement. A higher degree of synteny was observed with *P. trichocarpa*. Difference in evolutionary distances is not likely to be the reason for this increased microsynteny, since both *A. thaliana* and *P. trichocarpa* are part of the rosids clade, whereas tomato and potato belong to the asterids clade.

Taken together, these findings present a first glimpse into the evolution of Solanaceous genomes, both within the family and relative to other plant species. When the complete genomic sequences of these species become available, whole-genome comparisons and protein- or repeat-family specific studies may shed more light on the intriguing observations made in this chapter.

## Methods

### BAC end sequences

Tomato BESs from the HBa (*Hind*III), Eco (*Eco*RI) and Mbo (*Mbo*I) libraries were obtained from SGN FTP site [137]. For all analyses, the 'screened\_and\_trimmed' sets (bacends\_combined\_screened\_and\_trimmed.v4.seq) were used, in which low-quality regions and vector sequences have been trimmed, and sequences shorter than 150 nt have been removed. Additionally, this file excludes BESs with hits to the mitochondrial genome of *A. thaliana* [150] and the chloroplast genome of *N. tabacum* (NCBI GenBank accession NC\_001879.2), based on a BLASTN search with an E-value cutoff of  $10^{-10}$ . Potato BESs, which have undergone quality and vector clipping, were downloaded from the GSS section of NCBI GenBank [159] using the query "RHPOKEY". All sequences shorter than 150 nt and sequences with BLASTN (blastall 2.2.15) [103] hits to the *A. thaliana* mitochondrial or *N. tabacum* chloroplast genomes with a E-value lower than  $10^{-10}$  were removed in order to be consistent with the tomato data. Recently, the tomato and potato chloroplast genomes have become available; however, it can be assumed that the *A. thaliana* mitochondrial genome is sufficiently similar to these genomes, and as such additional filtering was not deemed necessary [103, 160].

### Repeat density and categorization

Repeats were identified in the tomato and potato BESs through similarity searches to the *Magnoliophyta* section of the RepBase repeat database (release 2006-10-06) [95], using RepeatMasker 3.1.5 [94] and cross\_match 0.990319 [161]. The repeat density was then calculated as the percentage of nucleotides in the BESs that had one or more hits to the repeat database. Classification of repeat families was derived from the annotation in the RepBase database. Redundancy in the BESs was detected with BLASTN (blastall 2.2.15), by comparing the tomato and potato BES data to itself and removing all matches of a sequence to itself. The E-value cutoff was set to  $10^{-5}$  and BLAST hits were removed if they did not have a minimum coverage of 50% of the query sequence with 90% identity.

### Simple sequence repeats

Microsatellites were detected using a modified version of the Sputnik software [162]. Running parameters were set to return all SSRs spanning at least 15 nucleotides, with a motif length between 1 and 5 (i.e., mono-, di-, tri-, tetra-, and penta-nucleotide repeats), and a minimum score of 8. Microsatellites identified in this manner were divided into two classes; class I, which has 10 or more motif repeats; and class II, which has fewer than 10 motif repeats [145].

## Gene content

The gene content of the BESs was estimated through BLAST searches using blastall 2.2.15. The BESs were compared to the NCBI GenBank non-redundant protein database (release 2007-02-16) [163] using BLASTX, and to the Kazusa KTU2 tomato EST database [164] and the CAB PotatEST potato EST database (January 2007 release) [132] using BLASTN. For all BLAST searches an E-value cutoff of  $10^{-5}$  was used, and the best five hits were evaluated. Additionally, a 90% identity cutoff was used for the BLASTN searches to the transcript databases.

In order to distinguish between true, putative protein-coding regions, and transposon- or contamination-related regions, the BLAST matches to the non-redundant protein database were filtered based on keyword matches in the BLAST hit descriptions. In general, these keywords described sequences that show similarity to bacterial contamination, transposon-related, chloroplast, mitochondrial and ribosomal protein sequences. Any BLAST match that was not filtered out by any of the keywords was considered to represent a non-repetitive, protein-coding region.

## Functional annotation

Tomato HBa and Eco, and potato POT and PPT BESs were functionally annotated through comparisons against the Pfam (version 21.0) [165] and PANTHER (version 6.1) [166] protein family databases, using InterProScan 4.3.1 [167]. GO terms from the Pfam annotations, and PANTHER family (but not subfamily) identifiers from the PANTHER annotations, were extracted from the merged output file of InterProScan. For each GO term and PANTHER family, the number of matching tomato and potato BESs was counted; if a single GO term or PANTHER family was assigned to the same sequence multiple times, for example due to multiple open reading frames in the same sequence, it was only counted once.

Subsequently, the counts were compared pairwise using a two-sided Fisher's exact test from the R software suite [168]. Note that GO term annotations are not always assigned independently of each other (as is required by Fisher's exact test), meaning that some terms often or exclusively occur together as they both describe different aspects of a single biological process or function. However, for simplicity, these higher order dependencies between GO terms are ignored, which may lead to an overestimation of the number of distinct overrepresented terms. Additionally, to mitigate error caused by differences in bias between libraries made with different restriction enzymes, direct inter-species comparisons are made only between BESs from libraries made with the same restriction enzyme. Lastly, the null hypothesis here is that there is no difference in abundance for a GO term or PANTHER family between the tomato and potato BESs, whereas the alternative hypothesis indicates a difference. A conservative P value cut-off of  $10^{-4}$  was selected to distinguish significant differences from non-significant differences.

### Comparative genome mapping

To determine potential areas of microsynteny between the Solanaceous species studied here and dicot model plants, paired BESs were selected and mapped to the *A. thaliana* and *P. trichocarpa* genome sequences with BLASTN alignments. Paired end sequences were available for 135,842 tomato BACs (63,169 HBa, 33,498 Eco and 39,175 Mbo) and 55,662 potato (34,362 POT and 21,300 PPT) BACs. Whole genome sequences of *A. thaliana* and *P. trichocarpa* were downloaded from TAIR [150] and JGI [169], respectively. The *P. trichocarpa* genome sequence used in this study was not finished, but rather consisted of a pseudomolecule sequence for each of the 19 chromosomes plus an additional 177,7 Mb in 21,993 contig sequences.

For each BES, only the best match to the respective genome sequence with an E-value lower than  $10^{-5}$  was evaluated, and the hit was rejected if the distance between subsequent HSPs was larger than 2,000 nt. A BAC was considered to have microsynteny to the target genome if both ends mapped within a distance of between 50 and 500 kb of one another. When both ends were oriented properly with respect to each other, the region was considered colinear; otherwise, the region was considered to be rearranged between the two species.

### **List of abbreviations**

Eco = Tomato *EcoRI* digested BAC library; EST = Expressed Sequence Tag; HBa = Tomato *HindIII* digested BAC library; Mbo = Tomato *MboI* digested BAC library; nt = nucleotides; POT = Potato *HindIII* digested BAC library; PPT = Potato *EcoRI* digested BAC library.

## Chapter 4

# *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements

Sander A. Peters\*, Erwin Datema\*, Dóra Szinay, Marjo J. van Staveren, Elio G.W.M. Schijlen, Jan C. van Haarst, Thamara Hesselink, Marleen H.C. Abma-Henkens, Yuling Bai, Hans de Jong, Willem J. Stiekema, René M. Klein Lankhorst, Roeland C.H.J. van Ham

\* These authors contributed equally

A modified version is published in *The Plant Journal* (2009) 58, 857–869

### Summary

We studied the physical and genetical organization of chromosome 6 of tomato (*S. lycopersicum*) cv. Heinz 1706 by combining BAC sequence analysis, High Information Content Fingerprinting, genetic analysis, and BAC-FISH mapping data. The chromosome positions of 81 anchored seed and extension BACs corresponded in most cases with the linear marker order on the high density EXPEN 2000 linkage map. We assembled twenty-five BAC contigs and eight singleton BACs spanning 2.0 Mb of the short arm euchromatin, 1.8 Mb of the pericentromeric heterochromatin and 6.9 Mb of the long arm euchromatin. Sequence data were combined with their corresponding genetic and pachytene chromosome positions into an integrated map that covers approximately one third of the chromosome 6 euchromatin and a small part of the pericentromeric heterochromatin. We then compared physical length (MB), genetic (cM) and chromosome distances ( $\mu\text{m}$ ) for determining gap sizes between contigs and revealing relative hot and cold spots of recombination. Through sequence annotation we identified several clusters of functionally related genes and an uneven distribution of both gene and repeat sequences between heterochromatin and euchromatin domains. Although a greater part of the non-transposon genes was located in the euchromatin, the highly repetitive (22.4%) pericentromeric heterochromatin displayed an unexpectedly high gene content of one gene per 36.7 kb. Surprisingly, the short arm euchromatin was relatively rich in repeats as well with a repeat content of 13.4%, yet the ratio of *Ty3/Gypsy* and *Ty1/Copia* retrotransposable elements across the chromosome clearly distinguished euchromatin (2:3) from heterochromatin (3:2).

## Introduction

The SOL Genomics Network (SGN) is an international consortium of groups that aims to develop the family *Solanaceae* as a model for a systems approach for understanding plant adaptation and diversification. A cornerstone in SGN is the International Solanaceae Genome Project (SOL) which aims to sequence the genome of tomato (*S. lycopersicum*) cv. Heinz 1706 [170]. Within the SOL project the Centre for BioSystems Genomics (CBSG) in the Netherlands takes responsibility for sequencing chromosome 6.

Tomato is a diploid species with twelve chromosomes and a genome size of approximately 950 Mb [19]. Its chromosome morphology has been well studied and cytogenetic analysis of pachytene chromosomes displays long continuous stretches of less condensed euchromatin in both chromosome arms flanked by highly condensed heterochromatin at the telomere ends and the centromeres [171-173]. Based on deletion studies, the pericentromeric heterochromatin contains approximately 75% of the nuclear DNA [174]. This was subsequently confirmed by recent heterochromatin estimates [175, 176, 173]. However, approximately 90% of all non-transposon genes were thought to reside in the euchromatin [149, 25] and this led to the initial efforts of SOL partners to concentrate on the euchromatic gene-rich space of the genome [149]. Initial sequencing efforts revealed that the euchromatin was largely devoid of repetitive sequences and had a gene density of 6.7 kb/gene, similar to *Arabidopsis* and rice. In contrast, the pericentromeric heterochromatin displayed a 10 to 100 fold lower gene density and was found to be densely packed with transposable elements, of which the *Ty3/Gypsy* class was the most abundant [25]. The euchromatin / heterochromatin proportion of the tomato genome has been subject of several cytogenetic studies and recent measurements yielded estimates of 31 Mb of euchromatin and 28 Mb of heterochromatin for chromosome 6 of which short and long arm euchromatin measure 4.1 Mb and 26.9 Mb, respectively [173, 28].

To reconstruct the euchromatic part of the tomato genome, SOL follows a BAC-by-BAC sequencing approach. Physical mapping is an integrated part of the reconstruction process as it provides a framework for ordering and joining sequence data, genetically mapped markers and BAC scaffolds. A classical global mapping approach known as ‘map-first-sequence-second’ was used in which a physical map is constructed from a fingerprinted BAC library, from which in turn a Minimal Tiling Path (MTP) of clones is selected for shotgun sequencing. Deep-coverage tomato *HindIII*, *EcoRI*, and *MboI* BAC libraries have been constructed [177, 136] and the *HindIII* library has been fingerprinted at the University of Arizona [178]. However, this physical map was built with FingerPrinted Contigs (FPC) and was based on low resolution and low information content fingerprinting, a technique known to introduce gaps and false overlaps in the MTP [179]. BAC contigs have then been linked to genetically mapped markers and have provided a framework of clones available for sequencing, positional cloning, and comparative analysis. To this end, approximately 500 overgo probes were designed from sequenced markers mapped on the EXPEN 2000

map, which is a high density genetic map of tomato constructed from an F2 population (F2-2000) of 83 individuals derived from the cross *S. lycopersicum* LA925 x *S. pennellii* LA716 [180]. These probes have been hybridized to BAC filter arrays in order to link overgo sequences to specific BAC clones. Although overgo screening is simple and efficient, spurious hybridization may cause the occurrence of both false positive and false negative BACs, as was described for maize [181]. While two-thirds of the probes could be unequivocally assigned to single BACs, the initial physical map was relatively unsaturated with anchor points. The low level of anchoring and the limited resolution of fingerprinted contigs prompted us to follow a local rather than global physical mapping strategy. In this so called Sequenced Tagged Connector (STC) approach [182, 183] large scale BAC end sequencing followed by similarity searches between seed BACs and BAC End Sequences (BES) were used to select favorable BACs for extension walking. After sequencing the extension BAC, the process is reiterated resulting in contigs which are built stepwise as sequencing progresses. This ‘map-as-you-go’ procedure has already been validated for tomato [184]. In the course of the project new BAC extension sequence overlaps and fluorescence *in situ* hybridization data progressively became available and were used to update continuously the contig construction and improve the scaffolding. Following contig building, identification of marker sequences on BAC inserts and cytogenetic mapping information provided us with the opportunity to tie physical map data to genetic map data. Here we present the integrated mapping results and the physical and genetical organization of 139 BACs on tomato chromosome 6 as revealed by combining High Information Content Fingerprinting, genetic mapping data, FISH and sequence annotation.

## Results

### BACs linked to genetic markers and cytogenetic mapping

To obtain seed BACs with mapped anchor positions on chromosome 6, *Hind*III and *Mbo*I tomato BAC libraries have previously been screened using overgo hybridization [185]. In addition, we selected a small amount of BACs using AFLP analysis (data not shown). We verified the cytogenetic position of 113 candidate BACs on pachytene chromosomes using BAC-FISH to construct a backbone for BAC walking (see Figure 4.1 and [28]). Fifty-one seed BACs and 30 extension BACs were confirmed to be located on chromosome 6. An additional three BACs landed on both chromosome 6 and on other chromosomes. The other twenty-nine BACs did not hybridize to chromosome 6 of which 24 showed a single focus on one of the other chromosomes and four had multiple foci onto multiple chromosomes. For one BAC we could not detect a clear signal.

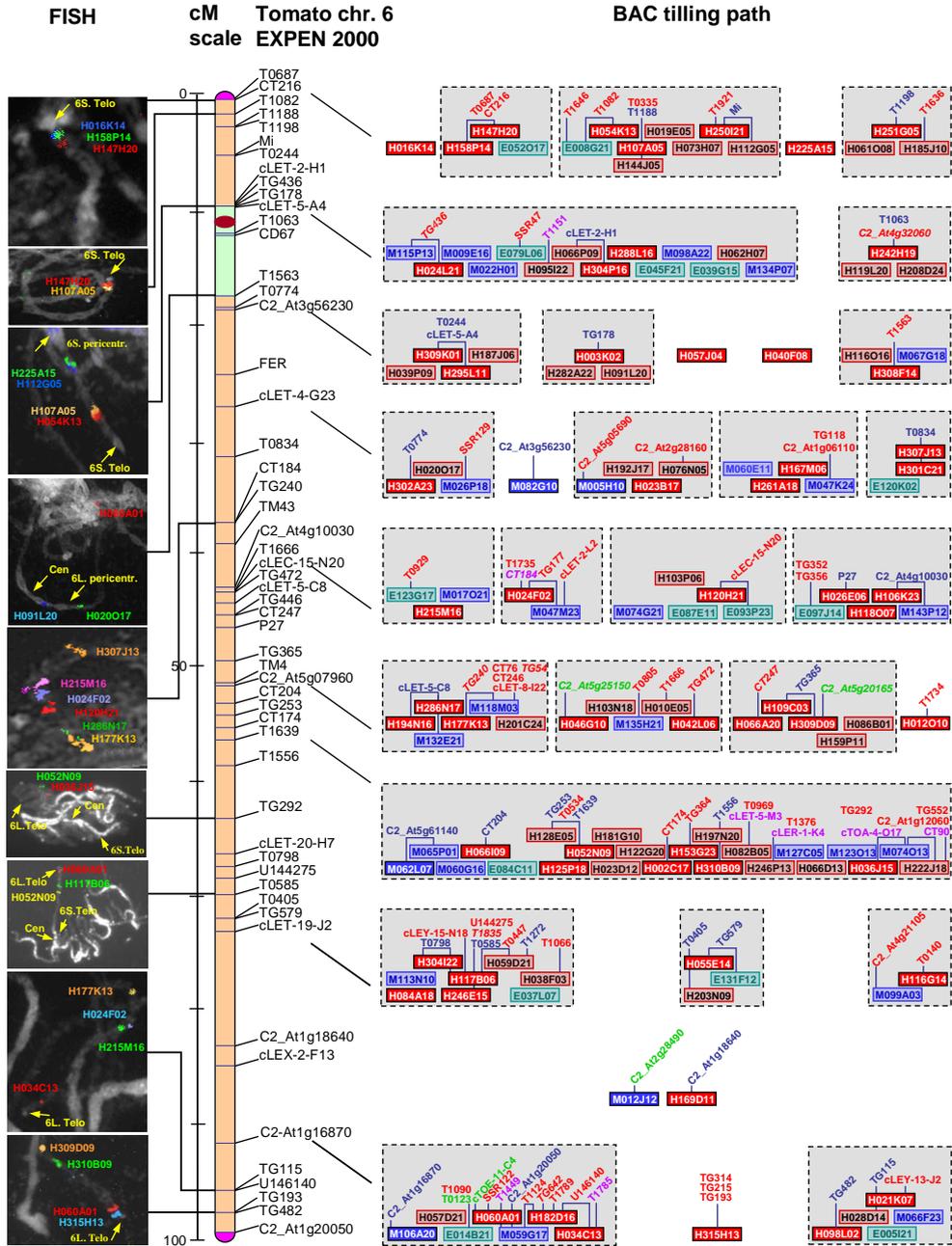
Aiming to link BAC sequences to the tomato EXPEN 2000 genetic map, the BAC sequences were searched against the tomato marker database from SGN [186]. In total, 154 markers were identified, of which eighty-eight have been mapped on chromosome 6 and

twelve have multiple genetic map locations (Figure 4.1 and Supplementary Table 4.1). The remaining fifty-four markers had not previously been mapped on tomato chromosome 6. Surprisingly, we found three markers in chromosome 6 BAC sequences that were genetically mapped on chromosomes 2, 4, and 11, respectively. One marker (cLET-5-M3) mapped on both chromosome 7 and 12, but not on chromosome 6. In addition, six markers were mapped onto both chromosome 6 and on another chromosome, whereas another four markers were genetically mapped at multiple positions on chromosome 6. Nevertheless, for all of these cases the corresponding BACs displayed a single clear fluorescent signal on chromosome 6. We then selected sets of seed BACs for multicolor FISH analysis such that each set shared at least one seed BAC included in another set as a reference. From the cytogenetic positions a linear order of seed BACs was determined. Overall, the cytogenetic mapping order of seed BACs was in agreement with the linear mapping order of anchored BACs to the Tomato EXPEN 2000 genetic map, although some striking discrepancies were observed. For example, the region between 0 and 5 cM on the short arm contains markers for which genetic positions are clearly inverted compared to the relative physical positions. Likewise, markers and corresponding BACs which have been genetically mapped to the long arm euchromatin regions around 47 cM and 97 cM have discrepant genetic and cytogenetic map orders (Supplementary Table 4.1).

#### BAC-by-BAC walking, physical mapping and chromosome coverage

A bidirectional BAC walk from 64 seed BACs was initiated. For the short arm euchromatin we sequenced 29 BACs which comprised 2.0 Mb of non-redundant sequence, covering 49% of the short arm. Approximately 6.9 Mb (26%) of non-redundant sequence was recovered for the long arm euchromatin from 90 BACs and an additional 1.8 Mb (6%) of pericentromeric heterochromatin was sequenced from 20 BACs (Figure 4.1).

We assessed the accuracy of ‘global’ physical mapping for placing the sequenced BACs in contigs on the physical map. To this end, we used 131 SNaPshot fingerprinted BACs, consisting of 12 seed BACs which were cytogenetically mapped to the short arm euchromatin of chromosome 6 and 119 candidate extension BACs. The fingerprints of these BACs were assembled in a single round with FPC resulting in seven contigs. After sequencing and assembly, the BAC order of these contigs was compared with the order of the corresponding BES as they were aligned to the 2 Mb of assembled sequence of the short arm. Within six contigs, of which the largest started at SL\_Mbo\_115P13 and ended with SL\_Mbo\_134P07, the linear BAC order was identical to the order derived from BES. However in the seventh contig seed BACs clustered into a stacked assembly along with other extension candidates. These BACs localized to non-adjacent positions on the genetic map as was confirmed by cytogenetic mapping (Figure 4.1). Assembly analysis indeed indicated that these BACs did not share sufficient sequence overlap to properly assemble and sequence annotation revealed that these BACs had a high repeat content. Only after exclusion of these repetitive BACs a mapping order could be produced that was in agreement with the contig order as determined from the assembled BES.



**Figure 4.1:** Physical coverage and integrated map for tomato chromosome 6. Cytogenetic mapping positions of seed and extension BACs are displayed in the left panel. Chromosome 6 markers and their corresponding genetic positions on the cM scale are displayed in the middle panel. The right panel displays the reconstructed BAC minimal tiling path of 25 supercontigs and 8 singleton BACs in accordance with the order of cytogenetically mapped seed BACs. Each BAC is shown with its identifier displayed in a colored box: *Hind*III BACs in red, *Mbo*I BACs in blue, and *Eco*RI BACs in green. Solid colors represent seed BACs, and transparent colors display extension BACs. Seed BACs identified by overgo hybridization have the corresponding marker identifier depicted in blue. Markers found by BLASTN analysis have red colored identifiers. BACs with a FISH confirmed chromosome 6 position containing a non chromosome 6 genetic marker, have the corresponding marker identifier depicted in green. Markers located on both chromosome 6 and other chromosomes are depicted in pink. Markers from a genetic map other than the EXPEN 2000 map are in italics.

In order to avoid the repeat resolution problem associated with ‘global’ physical mapping we instead focused our efforts on ‘local’ physical mapping. Using the Sequence Tagged Connector approach, BES were assembled on seed BAC insert sequences and analyzed with TOPAAS as described previously [184]. Candidate extension BACs that passed the TOPAAS quality control were subsequently fingerprinted with a High Information Content Fingerprinting (HCIF) technique using SNaPshot [27]. Each set of fingerprinted BACs, consisting of a single seed and corresponding candidate extension BACs was then assembled with FPC into a single contig in multiple rounds.

In total we placed 142 BACs on the physical map, 139 of which were sequenced. The average overlap between BACs in a supercontig was 13.3 kb. The relative order in which supercontigs were placed on the physical map was primarily determined by the FISH map position of the seed BACs (Figure 4.1 and [28]).

During the construction of an MTP for the euchromatic regions of chromosome 6 we identified several domains that were poorly covered by seed BACs, and these areas reflected the physical gaps that were not yet bridged in the BAC walking process. To estimate a global basepair / centimorgan relationship, chromosome distances on a micrometer scale were determined for the euchromatic and heterochromatic portion (Table 1). In addition, we estimated gap sizes as a fraction of the total euchromatin size by measuring physical distances between adjacent BAC FISH positions in pachytene complements flanking supercontigs (Table 4.1 and Supplementary Table 4.1). As an example, we observed a large gap towards the bottom of the long arm which was flanked by BACs LE\_HBa\_055E14 and SL\_Mbo\_106A20 on the physical map. On the genetic map this gap was flanked by marker T0405 (73 cM) and marker C2\_At1g16870 (92.5 cM) (Figure 4.2). The basepair / centimorgan ratio for the corresponding interval was 0.35 Mb/cM and this was comparable to what we observed for the long arm euchromatin portion of the chromosome, but almost two times more than the 0.2 Mb/cM ratio which was observed for most of chromosome 2 [187]. This distance makes up approximately 20% of the total linkage group of chromosome 6 and approximately 7 Mb on the physical map.

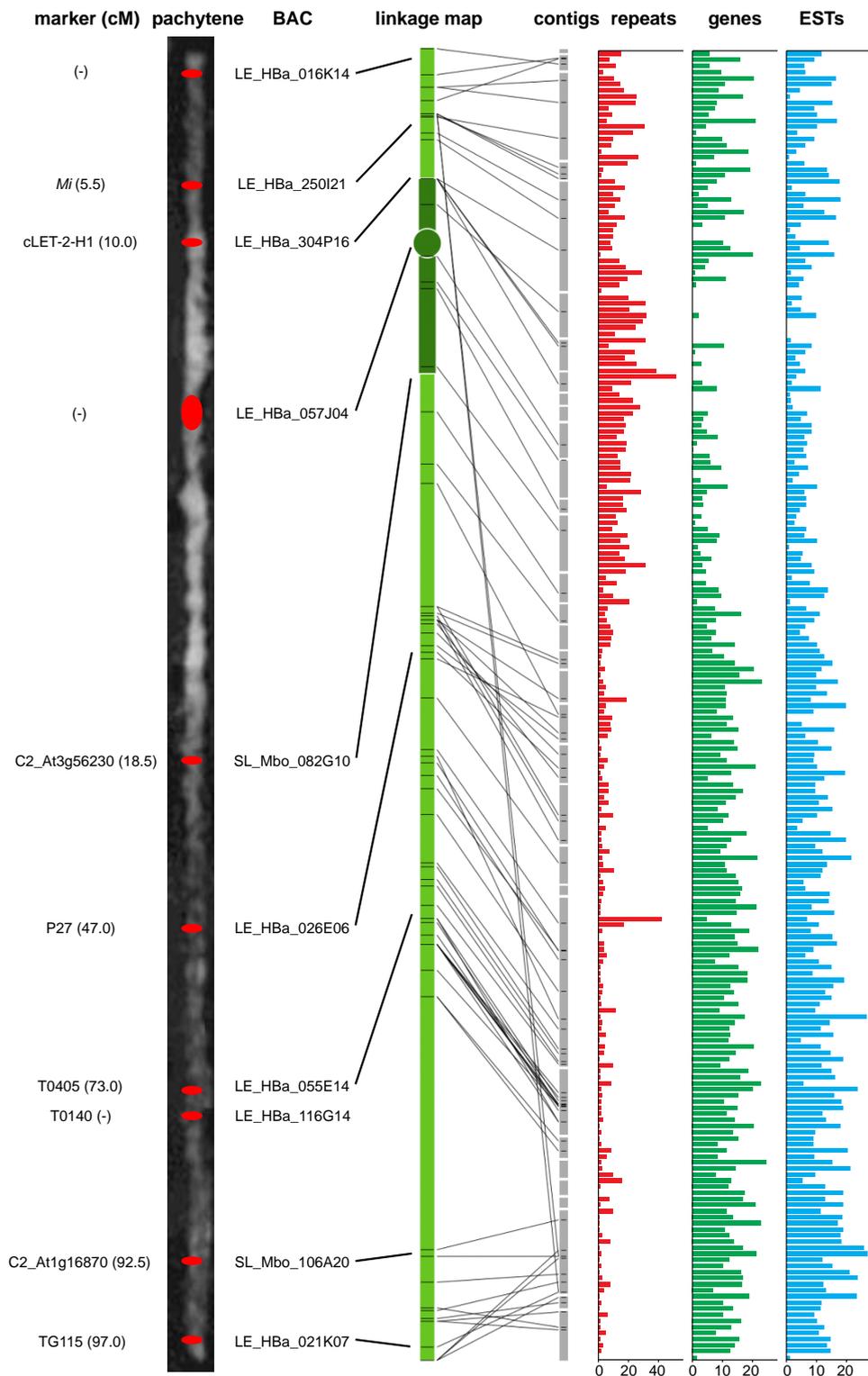
While we could place two seed BACs in this gap, extension of these BACs as well as extension of the BACs bordering the gap was unsuccessful. Proximal to the long arm telomere, in the 97 to 98 cM interval, we observed a gap between LE\_HBa\_034C13 and LE\_HBa\_098L02. This region has a relatively high basepair / centimorgan ratio of 1.2 Mb/cM and thus this small genetic gap corresponds to a considerable physical gap of 0.96 Mb.

**Table 4.1:** Physical and genetic distances of the short arm (6SE) and long arm (6LE) euchromatin region and the heterochromatin region (6SH+6LH) of chromosome 6.

Domain	Chromosome dist ( $\mu\text{m}$ )	Genetic dist. (cM)	Size (Mb)	Mb/ $\mu\text{m}$	cM/ $\mu\text{m}$	Mb/cM
6SE	1.8	10	4.1	2.27	5.55	0.41
6SH+6LH	8	8.5	50	6.25	1.06	5.94
6LE	29.5	82.5	26.9	0.91	2.8	0.33
telomere-H147H20	0.7	n.d.	1.59	2.27	n.d.	n.d.
H055E14-H106A20	7.7	20	7	0.91	2.6	0.35
H034C13-H098L02	1.29	0.8	1.17	0.91	0.62	1.46
H021K07-telomere	1.11	n.d.	1.44	1.3	n.d.	n.d.

#### Repeat and gene distribution in eu- and heterochromatic domains of chromosome 6

An additional advantage of the BAC FISH on pachytene chromosomes is the heterochromatin differentiation of the distal ends and the pericentromere [173]. The short arm displayed a relatively clear and distinct border of highly condensed heterochromatin and less condensed euchromatin, whereas the long arm shows a gradual transition of denser heterochromatin to euchromatin. We therefore focused on these borders to establish the boundaries of the repeat rich heterochromatin. The short arm euchromatin spanned between LE\_HBa\_016K14 just below the telomere region and LE\_HBa\_304P16 just north of the pericentromeric region (Figure 4.2). On the genetic map these BACs were mapped at 0 cM and 10 cM respectively. On the long arm SL\_Mbo\_082G10 landed just south of the pericentromeric domain mapping at 18.5 cM (Figure 4.2). LE\_HBa\_315H13 and LE\_HBa\_021K07 (Figures 4.1 and 4.2) were localized near the long arm telomeric domain mapping at 98 cM. Using these borders, the assembled sequence data were divided into three domains based on the local chromatin status: the short arm euchromatin (the first four supercontigs in Figure 4.1 measuring approximately 2.0 Mb), the pericentromeric heterochromatin (1.8 Mb) and the long arm euchromatin (the final fifteen supercontigs in Figure 4.1 spanning 6.9 Mb).



**Figure 4.2:** Predicted repeat and gene content of the assembled chromosome 6 sequence. On the left a straightened DAPI stained chromosome 6 of tomato cv. Heinz 1706, with a selected set of BACs and corresponding genetic markers to which is referred to in the text. Intense white and a reduced DAPI fluorescence correspond to condensed heterochromatin and less-condensed euchromatin, respectively. The superimposed colored dots correspond to the BAC FISH position taken from Fig. 1 and Fig. S1. Right of the chromosome is a cartoon representation of the genetic map on which the assembled supercontigs have been anchored. The three histograms on the right reflect repeats, genes and aligned transcript (EST) contents of the supercontigs. Each bar in the histograms represents a 50 kb interval of the assembled sequence.

Figure 4.2 displays the gene and repeat content of the assembled sequence contigs divided into bins of 50 kb and clearly identify the different chromatin domains based on their repeat and gene content. The short arm euchromatin had an average repeat content of 13.4%. In contrast, the repeat content of the pericentromeric heterochromatin measured 22.4%, yet a clear transition in repeat density between euchromatin and heterochromatin was not detected. The long arm euchromatin was almost devoid of repetitive sequences; however, the first Mb of the long arm, which contains BAC contigs that map to the transition between the tightly condensed heterochromatin and the more relaxed euchromatin as observed in pachytene chromosomes, clearly had a higher repeat content compared to the rest of the long arm (Figure 4.2). We also detected a single 20 kb insert of a *Ty3/Gypsy* type retrotransposon similar to Ogre [188] in an otherwise repeat-poor region (the large peak on the long arm in Figure 4.2).

The repeat content as a whole, but also *Ty1/Copia* and *Ty3/Gypsy* retrotransposons, were not uniformly distributed across the different chromosome domains. The short arm euchromatin contained more *Ty1/Copia* than *Ty3/Gypsy* related sequences (7.7% and 5.2%, respectively), whereas *Ty3/Gypsy* retrotransposons were more prevalent in the pericentromeric heterochromatin (14.7% *Ty3/Gypsy* versus 9.0% *Ty1/Copia*). The long arm euchromatin contained only 2.6% *Ty1/Copia* and 1.5% *Ty3/Gypsy* related sequences. The amount and distribution of DNA transposons also differed between the three domains (Supplementary Table 4.2). Approximately one quarter of the DNA transposons in the short arm euchromatin were identified as hAT and hAT-Ac transposons [189], whereas these only comprised one sixth and one ninth of the heterochromatin and long arm DNA transposon content, respectively. While the overall DNA transposon density of the pericentromeric heterochromatin was similar to that of the short arm, this domain contained a higher amount of En-Spm and MuDR transposons [190, 191]. The distribution of several previously identified tomato repeats also differed between the chromosome domains. For example, SINEs (the SolSINE family) and TRIMs (Tork2) were absent from the pericentromeric heterochromatin but prevalent in the euchromatin (Supplementary Table 4.3).

In total, 970 putative protein coding genes were predicted and their distribution of genes varied remarkably between the chromosome domains. The short arm euchromatin contained one gene per 15.3 kb, whereas the gene density in the long arm euchromatin was almost twice as high with one gene per 8.8 kb. Forty-eight genes were predicted in the pericentromeric heterochromatin yielding an unexpectedly high gene content of one gene per 36.7 kb. With on average 4.0 exons per gene, the genes in this domain differed from those in the short and long arm euchromatin, which had 5.0 and 5.2 exons per gene, respectively. This difference was also reflected by a higher relative abundance of single- and two-exon genes and a lower abundance of genes with nine or more exons in the heterochromatin. Several heterochromatic contigs contained 'gene islands' that were separated by long stretches of retrotransposons, whereas others showed a more even distribution of genes. For example, six genes were predicted on the 160 kb supercontig consisting of seed BACs LE\_HBa\_003K02 and LE\_HBa\_271L05, which were cytogenetically mapped close to the centromere. For three of these genes ample EST evidence was found, indicating that the gene-containing areas extend well into the pericentromeric heterochromatin. On the other hand, no genes were predicted on BAC LE\_HBa\_057J04, which was located nearest to the centromere. This BAC contained several copies of the centromere-associated TGRIV retrotransposon but we could not identify any large CAA blocks as found near the chromosome 12 centromere, nor any PCRT-related sequences [192, 173]. The BAC contigs that were mapped nearest to the telomeres did not show a decline in gene content, nor did they contain long stretches of subtelomeric TGRI repeats [173], likely because the outermost BAC contigs are not sufficiently close to the telomere to delimit the euchromatin (Table 4.2). Strikingly, a small cluster of TGRI repeats (23 copies and several fragments; Supplementary Table 4.3) was observed on the long arm, likely corresponding to the interstitial site previously identified through FISH [193].

Six clusters of functionally related genes were identified in the long arm euchromatin based on GO annotation. The contig harboring the P27 locus held two such clusters: one cluster of four putative GDSL-motif lipases upstream of P27 and another cluster of four predicted Tospovirus resistance genes just downstream of P27. Close to marker TG472 we found a cluster of five (potentially six, when considering *ab initio* gene predictions and BLASTX alignments) cytochrome P450 genes. Four genes resembling ABC2 transporters were found clustered together near marker T0534; a cluster of four (potentially five) single-exon Agenet homologs was predicted in a markerless region on BAC LE\_HBa\_036J15; and a group of four putative wound-inducible carboxypeptidases was identified between markers T1114 and T1124.

## Discussion

### Physical mapping accuracy

We used BAC FISH for cytogenetic confirmation of the chromosomal position of seed and extension BACs that were selected for sequencing the euchromatin parts of chromosome 6. Sequence data progressively became available in the course of this study and 31% of the markers we identified had not been previously mapped on the chromosome. We identified seven EST-derived markers that mapped in addition to chromosome 6 primarily on chromosome 3 and 9, suggesting gene duplications between these chromosomes. Overall, chromosomal positions of the BACs were generally in agreement with the genetically mapped marker order from the EXPEN 2000 map although we observed several types of inconsistency. These markers may have been erroneously mapped, or the BACs that were picked up with these markers have been aberrantly identified in the overgo screening process as result of spurious hybridization. Alternatively, discrepancies with the genetic map may be the result of different genotypes that have been used for the EXPEN 2000 map construction. This genetic map of tomato is based on a segregating population of *S. lycopersicum* LA 295 x *S. pennellii* LA716 and may be biased by small chromosomal rearrangements between the parents. Furthermore, rearrangements may exist between these lines and Heinz 1706. Nevertheless, the integrated map presents a crosslink between genetic markers, sequences and (cyto)genetic locations of BACs on tomato chromosome 6 and as such is a very valuable resource to the tomato research community.

We assessed the accuracy of the ‘local’ and ‘global’ fingerprint mapping approaches. In contrast to a global map-based approach, the STC approach bypasses the need for a global physical map and requires fewer BACs to be fingerprinted. However, the reduced fingerprinting effort comes at the expense of time-saving massive parallel fingerprinting and mapping as it needs successive rounds of fingerprinting for each individual BAC extension. In addition, continuous rounds of BLASTN similarity searches are needed for each successive bidirectional extension. Our ‘local’ mapping approach involved fingerprinting a single seed BAC together with candidate extension BACs and thereby the seed BAC was isolated from repetitive sequences in other BACs. In this way a drastic reduction of complexity circumvented the adverse effect of repeats that caused repetitive BACs to cluster. Whereas ‘global’ mapping was less accurate and suffered from repetitive BACs which were displaced in the FPC map, SNaPshot fingerprinting combined with local FPC mapping produced more reliable results and was more robust when mapping high repeat containing BACs. Thus, while the existing ‘global’ FPC map provides a valuable resource for selecting candidate extension BACs, ‘local’ FPC maps can resolve repeat-rich contigs, which are abundant in large plant genomes such as tomato.

## Chromosome structure and organization

The higher abundance of *Ty3/Gypsy* in chromosome 6 supports earlier observations on the unequal ratio of *Ty3/Gypsy* and *Ty1/Copia* retrotransposons in the tomato genome. A survey of more than 300,000 tomato BES presented a *Ty3/Gypsy*: *Ty1/Copia* ratio between 2:1 and 3:1 (Chapter 3) and in the current study we found that this ratio varied between the gene-poor heterochromatin (roughly 3:2) and gene-rich euchromatin (roughly 2:3) domains of chromosome 6 (Supplementary Table 4.2). An insertion bias for retrotransposons has previously been described for other plant genomes including *A. thaliana*, *Cestrum* spp., members of the genus *Helianthus*, conifers and maize [121, 194-198] and several families of retrotransposons have been identified in maize that preferentially associate with gene-rich or gene-poor regions [198]. Tam *et al.* [199] showed that the distribution of *Ty1/Copia* retrotransposons in tomato and related wild species are determined by factors such as genetic drift and mating system, but not recombination rate. We observed an enrichment of *Ty1/Copia* retrotransposons in the short arm euchromatin and *Ty3/Gypsy* retrotransposons in the pericentromeric heterochromatin and we know from previous research that the TGRIV *Ty3/Gypsy* retrotransposon is preferentially located in the structural centromeres of tomato [173]. Consistent with previous genome-wide cytogenetic studies we also observed a preferential localization of TGRII and TGRIII in the pericentromeric heterochromatin (Supplementary Table 4.3 and [173]). Taken together these data strongly suggest that insertion preferences of different types of retrotransposons also occurred in the tomato genome. It has been suggested that transposable elements play a major role in genome organization, evolution, gene regulation and function [200], however it is not clear how mobile elements have affected the evolution of genes and function in the tomato genome. Additional studies on the insertion distributions are needed in order to understand better the role of transposable elements on tomato genome evolution.

Functional annotation of the predicted genes revealed 15 putative cytochrome P450 genes, thereby reinforcing the observation of a large number of cytochrome P450 domains in the BES data representing 19% of the tomato genome (Chapter 3). Five P450 genes were found clustered together and another six of these genes were present in three pairs. In total we found six clusters of four or more genes that overlapped in their GO annotation. Recently a conserved cluster of four genes was identified in *Arabidopsis* and oat that plays a role in triterpene synthesis [201] and similar metabolic gene clusters have been identified in maize and rice [202, 203]. The clusters of lowly transcribed genes we found on chromosome 6 could also indicate a functional relationship between these genes [204]. While we only studied one chromosome, these findings may hold true for the complete tomato genome. Using Fisher's exact test we could not identify any significantly over- or underrepresented GO terms between the genes annotated on chromosome 6 and the GO terms found in the genome-wide study of the tomato BES (Chapter 3 and data not shown), indicating that the annotation of chromosome 6 can be considered an informative sample of the genome as a whole. Manual curation of the sequence annotation as well as exploiting the hierarchical relationships between different GO terms can also be used to identify more and larger

clusters of functionally related genes in tomato and in this way shed light on the molecular evolution of the genome [205].

### Genetics of tomato chromosome 6

The higher abundance of retrotransposons and repeats in the short arm euchromatin co-localized with disease resistance gene (*R* gene) loci near markers T1188 and *Mi* in BACs LE\_HBa\_019E05 and LE\_HBa\_250I21, respectively. Analysis of DAPI-stained tomato pachytene chromosomes showed the morphological differentiation into euchromatic and heterochromatic parts. BAC FISH analysis showed the repeat rich regions to be of euchromatic nature. These findings indicate that the distribution of relatively gene poor and repeat rich domains are not necessarily confined to the heterochromatin, but also extend to the less condensed short arm euchromatin. The repeat rich nature coincides with a suppression of recombination that has been noticed for these regions. The accumulation of retrotransposons as a result of recombinatorial suppression in repeat-rich regions has been observed in the maize genome and is suggested to be a general property of interstitial gene-poor domains intermixed with euchromatin from maize [198]. Currently, we do not know whether there is a relationship in tomato between repeat content and recombination frequencies.

A previous study revealed a chromosomal inversion in the *Mi-1* region between *S. lycopersicum* and *S. peruvianum* (donor of the *Mi-1* gene), which might explain the severe recombination suppression in this region [206]. Molecular markers flanking two different alleles of the *Mi*-homologues were in the same relative orientation, but markers between the two clusters were in an inverse orientation. Interestingly, a macro-synteny study between tomato (including Cherry VFNT and Heinz 1706) and potato using cross-species BAC-FISH painting revealed a paracentric inversion in the short arm of chromosome 6 that covers the whole euchromatic part. This inversion may have reshuffled the gene order and affected gene function [22]. If this inversion occurred in wild tomato species, we would expect suppression in recombination frequency for interspecific crosses. Another cluster of *R* genes was found on the long arm near marker P27 (47 cM). For this region a suppression of recombination has been observed (Y. Bai, unpublished results). Interestingly, some of the markers that map in this region (including P27 and C2\_At4g10030) also appear in reverse orientation compared to the order of the associated BACs in the FISH map.

### Distribution of the repeat and gene space

Many genetic markers, such as EST-derived markers, COS markers and cDNA-derived RFLPs, which anchor BACs to the tomato EXPEN 2000 genetic map correspond to genes. This has resulted in a low number of anchored BACs in gene poor and repeat rich regions and has led, amongst others, to a discontinuous BAC minimal tiling path for several repeat-rich regions of tomato chromosome 6. Tanksley *et al.* [128] noted large gaps in the molecular map of tomato and suggested these gaps might represent areas which were deficient in genes and low copy sequences. Alternatively, the large genetic distance

combined with a relative short physical size, as also observed for the 73 to 93 cM interval on the long arm of chromosome 6, might be explained as a result of a high recombination frequency in this region. In contrast, we observed a relative cold spot of recombination for the 97 to 98 cM interval which was reflected by a fivefold increase in the basepair / centimorgan ratio. The inconsistent order of markers on the genetic map might be explained by the low rate of recombination in this region.

The mapping and sequence analysis effort presented here is aimed at getting an overview of the gene-rich space of tomato chromosome 6. Since the bulk of all non-transposon related genes are currently thought to reside in the euchromatin [149, 25], we delineated the euchromatin borders by combining BAC FISH on pachytene chromosomes, sequence annotation and genetic mapping data. While we indeed identified a high repeat content in the pericentromeric heterochromatin, 48 genes were predicted in the 1.8 Mb that represent the various regions of heterochromatic sequence and transcription for many of these genes was detected from the EST data. The high gene density of these regions corresponds well to that found for six heterochromatic BACs from chromosomes 2, 7, 8, 9 and 10 [25] as well as that of the *jointless-2* locus in the pericentromeric region of chromosome 12 [192]. These regions contain one gene per 56 and 65 kb, respectively; however, they were selected for sequencing using gene-based markers or other single-copy sequences. Similarly, the heterochromatic BACs in this study could also present a biased view, yet the seed BACs of four of these contigs (BACs LE\_HBa\_003K02, LE\_HBa\_057J04, LE\_HBa\_040F08 and LE\_HBa\_308F14) were selected on the basis of AFLP markers and not on gene-based markers (data not shown). These four contigs span 679 kb and contain 22 genes, implying that the gene content of these contigs is not substantially different from those identified by gene-based markers. Considering the 28 Mb of heterochromatic sequence of chromosome 6, a substantial fraction of tomato genes may in fact reside in the pericentromeric domain.

Recent discoveries have reported on large amounts of heterochromatic genes in plants, mammals and *Drosophila* [207]; in the latter, the expression of such genes has been shown to be dependent on the heterochromatin environment [208]. The genes we found in the pericentromeric heterochromatin contained on average fewer exons than euchromatic genes and were often grouped into ‘gene islands’ separated by stretches of retrotransposons, as was previously reported for the FER locus [209]. While we currently do not know whether these findings imply an adaptation of gene structure and organization to this chromosome domain, our observations confirm that tomato heterochromatin cannot merely be regarded as being functionally inactive regions with respect to gene expression.

The predicted gene content and the amount of transcriptional evidence were generally in good agreement with each other. However, several bins had high predicted gene content yet a low amount of EST evidence (Figure 4.2). One bin in BAC LE\_HBa\_107H05 contained three putative leucine-rich repeat protein kinases. Two other bins corresponded to the *Mi* locus which harbored 3 putative NBS-ARC-LRR disease resistance genes, amongst other predicted genes. We were unable to delineate the proposed topology of the *Mi* locus [206],

although we predicted one additional disease resistance gene approximately 300 kb upstream and a second one close by based on *ab initio* gene predictions and BLASTX alignments. This chromosomal region is a hot spot of *R* genes. These genes are often clustered in tandem arrays including *Cf-2/Cf-5*, *Ol-4/Ol-6*, *Mi-1/Mi-9* and *Ty-1* genes conferring resistance to several unrelated pathogens [210]. Another two bins of high putative gene content and low expression were found in the long arm euchromatin of which one bin mapped between markers P27 and C2\_At4g10030 and contained the cluster of putative Tospovirus resistance genes. Interestingly, *R* genes for virus resistance have been mapped in this region [211]. Some of the genes in these bins could be non-functional, or were simply not be transcribed in detectable amounts under the conditions in which the EST libraries were generated. We also identified a number of gene-poor bins in the pericentromeric heterochromatin, in the short arm euchromatin and the first Mb of the long arm euchromatin that contained a large amount of aligned transcripts. A considerable amount of ESTs in these bins matched to retrotransposon-related genomic sequences, providing further circumstantial evidence for the presence of transcriptionally active retrotransposons in the tomato genome [212].

### BAC walking and finishing the tomato genome

We observed chromosomal regions which are not yet targeted with markers and this has resulted in an assembly containing mega base sized gaps between BAC supercontigs. Without the identification of new seed BACs the bridging of these gaps by BAC-walking will be difficult and time-consuming. To complement the chromosome 6 sequencing project and indeed the whole tomato sequencing project we will consider additional approaches. Sequence comparison between tomato and BACs from other *Solanum* species combined with cross-species multicolor BAC-FISH painting may allow identification of new candidate seed BACs. Whole genome shotgun sequencing with Next Generation Sequencing platforms undoubtedly will speed up and help to complete the sequencing. Yet, while the NGS platforms represent powerful technologies that produce large amounts of sequences, chromosome based assemblies including the vast amount of repetitive sequences will be a major challenge. Nevertheless, such a whole genome sequencing approach can benefit from the sequence islands that have been produced and which may serve as a backbone for whole chromosome assembly.

## **Methods**

### Chromosome preparations

All FISH experiments were performed on tomato *S. lycopersicum* cv. Heinz 1706 ( $2n=2x=24$ ). The pachytene preparations from young anthers containing pollen mother cells and the spreads of extended DNA fibres from young leaves were made following the protocols of Zhong *et al.* [213] and Budiman *et al.* [214].

### Fluorescence in situ hybridization

Two-color and multi-color FISH of BAC clones to pachytene chromosomes were performed according to the FISH protocols [215]. Slides were examined under a Zeiss Axioplan 2 Imaging Photomicroscope equipped with epifluorescence illumination, filter sets for DAPI (4', 6-diamino-2-phenylindole), FITC, Cy3, Cy5, DEAC, and Cy3.5 fluorescence. To determine  $\mu\text{m}$  distances between BACs each distance was measured on 10 straightened chromosomes and then averaged. Capturing of selected images, image processing, and distance measurements were performed as previously described [28].

### BAC clones

BAC clones were inoculated in 48 well blocks containing 2.5 ml 2 x LB medium with 12.5  $\mu\text{g}/\text{ml}$  chloramphenicol and were grown for 20 hours at 37°C and 175 rpm. Overnight cultures were centrifuged at 5796 rcf for 10 minutes. Cell pellets were used for high throughput BAC DNA isolation in a 96-well plate format using a Biorobot 9600 and the R.E.A.L prep kit (Qiagen). BAC DNA was dissolved overnight at 4°C in 50  $\mu\text{l}$  milliQ. Quantification of BAC DNA yield was carried out by fluorescence detection (Tecan XFluor4) in the presence of Quant iT PicoGreen (Molecular Probes).

### Fingerprinting reaction

On average 1  $\mu\text{g}$  BAC DNA digested by *Bam*HI, *Eco*RI, *Xba*I, *Xho*I, and *Hae*III (Invitrogen) according to the protocol described by Luo *et al.* [27] and labeled with the SNaPshot Multiplex Ready Reaction Mix (ABI). Sedimented and labeled DNA was washed with 100  $\mu\text{l}$  70% ethanol and centrifuged again for 10 min. Air-dried pellets were dissolved in 10  $\mu\text{l}$  of Hi-Di formamide and 0.5  $\mu\text{l}$  of internal size standard LIZ-600 (Applied Biosystems). DNA was denatured at 95°C for 5 min and chilled on ice and analyzed on a ABI 3730 Genetic Analyzer (Applied Biosystems) using POP-7 polymer. Peak detection, intensity, and collection were executed with the Any5 Dye-set by the data-collection software version 3.0 (Applied Biosystems).

### FPC data processing, BAC selection and mapping accuracy assessment

ABI 3730XL Genetic Analyzer data was processed with Genemapper 4.0 (Applied Biosystems). Typically, 115 fragments  $\pm$  52 per fingerprinted BAC with a size range of 100 to 600 bp were selected with GeneMapper at a peak intensity threshold of 100. Genemapper sizing results were processed by Genoprofiler 2.1 [216] and converted into a FPC [217] usable format.

The procedure to identify minimal overlapping clones and maximal extending inserts was as previously described [184]. SNaPshot fingerprinted tomato BACs were mapped with FPC with an optimal probability of coincidence of  $10^{-7}$  and a tolerance level of 0.4 bp, which accounted for the highest amount of true overlapping extension BACs. The mapping

result was compared against the contig order within 2.0 Mb of assembled BAC sequence and BAC extension overlaps found with FPC were compared to the overlap positions and linear order of BES of candidate extensions assembled onto the seed BAC insert sequence, and subsequently the number of displaced BACs was determined.

### Sequencing and assembly

Sequencing and assembly were performed as described by Peters *et al.* [184]. BAC sequences are available at [137].

### Sequence annotation

The sequence data was annotated using the Cyrille2 system [116]. Interspersed repeats were identified using RepeatMasker 3.2.5 [94] and `cross_match` 0.990319 [161] with the *Magnoliophyta* section of RepBase (release 2008-08-01) [95], the TIGR Lycopersicon repeats v3.1 and Solanaceae repeats v3.1 [218] and a tomato-specific retrotransposon library (E. Datema, unpublished results).

Genes were predicted using the linear combiner option of JIGSAW 3.2.8.1 [84] integrating data from *ab initio* gene predictors Augustus 2.0.2 [219], Genscan [75], Geneid 1.3.7 [220] and GlimmerHMM 3.0.1 [76], each using *A. thaliana* gene models); alignments of *S. lycopersicum*, *S. tuberosum* and *N. tabacum* ESTs (obtained from [164], [221] and [137], respectively) generated using BLASTN 2.2.17 [103] and `sim4` [222]; and alignments of proteins from the plant division of UniProt [223] and *A. thaliana* TAIR8 [150] generated using BLASTX 2.2.17 and GeneWise 2.2.0 [119]. The predicted genes were annotated through BLASTX alignments to the NCBI nr (2008-03-30 release) and RefSeq (2008-05-28 release) databases and InterProScan 4.3.1 [165] domain searches using version 15.0 of the databases. Marker sequences downloaded from SGN were aligned to the genomic sequences using BLASTN and `sim4`.

To identify clusters of functionally related genes, a list of non-redundant GO terms per gene was produced. Subsequently, overlapping windows of ten genes were created and tested for overrepresented GO terms using Fisher's exact test. The number of occurrences of each GO term per window was compared to the frequency of that GO term in the full set of 970 genes. A GO term was found to be significantly overrepresented in a window at a P value smaller than 0.01 applying the Holm-Bonferroni correction.

## **Acknowledgements**

This project was (co)financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research and by the European Commission (EU-SOL project PL 016214).

## Supplementary materials

**Supplementary Table 4.1:** Physical, cytogenetic and genetic distances for tomato chromosome 6 BACs and contigs. Contigs are ordered according to their FISH position and contig sizes are derived from assembled BACs and are depicted in light grey. Dark grey boxes represent distances between FISHed BACs in  $\mu\text{m}$ . Genetic positions of markers identified with BLASTN are derived from the EXPEN 2000 genetic map and BACs are shown on which the corresponding marker was found.

	Marker ID	Position (cM)	BAC ID			Size (kb)	Size ( $\mu\text{m}$ )
1			H016K14			35	
2	T0687	2	H147H20	H158P14		132	0
	C2_At3g46780	4	H147H20	H158P14			
	CT216	0	H158P14	H147H20			
	C2_At5g59520		H147H20				
			E52O17				
3	T1646		E008G21			750	0.31
	T1082	3	H054K13	E008G21			
	T0335		H107A05	H144J05			
	T1188	3	H107A05	H144J05			
	C2_At2g39690	5.3	H250I21				
	Mi	5.5	H250I21				
	T1921		H250I21				
4			H225A15			93	
5	C2_At5g48655		H251G05	H061O08		149	1.22
	C2_At3g25120	5.2	H251G05				
	T1198	5	H251G05	H185J10			
	T1636	5	H185J10				
6	TG436		M009E16	M115P13	H024L21	943	0.2
	SSR47	6.5	M009E16	M115P13			
	T1151	7	H095I22				
	C2_At4g01900-3		H095I22				
	C2_At3g62360		H304P16	H288L16			
	cLET-2-H1	10	H304P16	H288L16			
			M134P07				
7	T1063	12	H242H19			380	4.27
	C2_At4g32060		H242H19				
8	cLET-5-A4	10	H309K01	H295L11		263	13.1
	T0244	10	H309K01	H295L11			
centromere							
9	TG178	10	H003K02	H271L05		162	
10			H057J04			96	
11			H040F08			118	
12	C2_At5g05330		H116O16			302	
	T1563	16	H308F14				
	C2_At3g56290		M067G18				

13	T0774	18	H302A23			319	0	0	0.58	2.48
			H020O17							
	SSR129		M026P18							
14	C2_At3g56230	18.5	M082G10			116				
15	C2_At5g05690	24.5	M005H10			480		1.92	5.61	
	C2_At2g40080		H192J17							
	C2_At5g05660		H023B17							
	C2_At2g28160		H076N05							
16	C2_At1g06110	28	M047K24			247				
	TG118		M047K24							
	C2_At3g13940		M047K24							
17	T0834		H301C21	H307J13		164				
18	T0929	32	E123G17			201				
19	T1735		H024F02			149		1.36		
	CT184		H024F02							
	TG177	43	H024F02	M047M23						
	cLET-2-L2	43.5	M047M23							
20			M074G21			269	0.92			
	cLEC-15-N20	44	E093P23	H120H21						
21	TG356		E097J14			329	0		1.4	
	TG352	33.5	E097J14							
	P27	47	H026E06							
	C2_At4g10030	44	H106K23							
	C2_At1g22850	46	M143P12	H106K23						
22	C2_At2g05170		M123E21	H286N17		328		3.65		3.65
	cLET-5-C8	45	H286N17	H194N16	M132E21					
	C2_At1g03150-2		H286N17	M132E21	H194N16					
	C2_At1g71990		H286N17	H194N16	M132E21					
	C2_At1g71950	46.5	H177K13							
	TG240		H177K13	M118M03						
	CT246		M118M03							
	CT76		M118M03							
	cLET-8-I22	44.3	M118M03							
TG54		M118M03								
23	C2_At5g25150		H046G10			509				
	C2_At4g32260	43	M135H21							
	T0805	43	M135H21							
	T1666	43.7	H010E05							
	C2_At5g62030		H042L06							
	TG472	44	H042L06							
	C2_At1g32130		H042L06							
24	CT247		H066A20			318	0	0.91	1.25	2.15
	TG365	50	H109C03	H309D09						
	C2_At5g51020-2		H309D09							

	C2_At5g20165		H309D09						
25	T1734		H012O10						
	C2_At1g73885-2		H012O10			80			
26	C2_At5g61140		M062L07	M065P01			0.35		
	CT204	54	H066I09						
	T0534		H052N09	H128E05	E084C11				
	C2_At5g62530-3		H052N09	H128E05	E084C11			0.9	
	C2_At5g62530-2		H052N09	H128E05	E084C11				0.49
	T1639	57	H052N09	H128E05	H125P18				
	TG253	55	H052N09	H128E05	H125P18		0.03		
	CT174	56	H023D12	H002C17					0.06
	C2_At5g62590		H153G23	H002C17					
	TG364		H153G23						
	C2_At1g18340		H310B09	H197N20			0.54		
	T1556	59	H310B09	H197N20					
	T0969		H082B05						
	C2_At5g17990		H082B05						
	cLET-5-M3		H082B05						
	C2_At5g63100		H246P13						
	T1376		H246P13						
	C2_At4g24690-2		H246P13						
	cLER-1-K4	54.5	H246P13						
	C2_At5g63200		M127C05						
	C2_At1g24360	62.7	M127C05						
	cTOA-4-O17	63	H036J15	M074O13					
	TG292	64	M074O13	H036J15					
	C2_At1g12060	64.5	H222J18	M074O13					
	TG552		H222J18	M074O13					
	CT90	66	H222J18						
27			H084A18						
	T0798	69	H117B06	H304I22					
	C2_At5g24490-2		H117B06	H304I22					
	cLEY-15-N18	69	H117B06						
	C2_At2g43540		H246E15	H117B06					
	T1835		H246E15	H117B06					
	U144275		H246E15	H117B06					
	T1462	69	H246E15	H117B06					
	C2_At4g34215	69	H246E15	H117B06					
	C2_At1g09340	67.3	H246E15	H117B06					
	T0585	71	H246E15	H117B06					
	T1515	69	H246E15	H117B06					
	C2_At4g33680		H246E15	H117B06					
	T1707		H246E15	H117B06					
	C2_At3g59400		H117B06						
	T0447		H117B06	H059D21					
	C2_At3g11710	68.3	H059D21						
	TG162		H059D21						
	C2_At3g11710-2		H059D21						
	C2_At3g11710-3		H059D21						
	T1272		H059D21						
	C2_At2g43360	67	H038F03						
T1066	69	H038F03							
C2_At2g37560		H038F03							
									0.69
									0.06
									0.9
									0.49
									0.03
									0.54
									1.69
									1.5
									3.92
									0.61
									0.1
									0.71
									0.93
									0.76

28	T0405	73	H203N09	H055E14		181				
	C2_At2g27170	73	H055E14	H203N09						
	TG579	73	E131F12	H055E14						
	C2_At5g09920-2		E131F12	H055E14						
29	C2_At4g21105-2		M099A03			153				
	T0140		H116G14							
30	C2_At2g28490		M012J12			128				
31	C2_At1g18640-2		H169D11			89				
32	C2_At5g58470	92.5	M106A20			722	0.7	1.48	3.07	
	C2_At1g16870		M106A20							
	T0123		E014B21							
	T1090		E014B21							
	cTOE-11-C4	101	H060A01							
	SSR122		H060A01							
	C2_At5g01910		H060A01							
	C2_At2g39780		H060A01							
	C2_At5g48380		H060A01	M059G17						
	C2_At5g48380-2		H060A01	M059G17						
	T1449	5 & 93	H060A01	M059G17						
	C2_At1g20050	101	H060A01	M059G17						
	T1114		H060A01	M059G17						
	T1124	95	H182D16	M059G17						
	T1789	95	H182D16							
TG642		H182D16								
U146140	97.2	H034C13	H182D16							
T1785	5 & 100	H034C13								
33	TG314	101	H315H13			105	0.7			
	TG215		H315H13							
	TG193	97.8	H315H13							
34	TG482	98	H098L02			430	0.33			
	C2_At2g42450		H098L02							
	C2_At1g69523		H098L02	H217M17	H028D14					
	TG115	97	H217M17	H021K07	H028D14					
	cLEY-13-J2	98	H028D14	H021K07						

**Supplementary Table 4.2:** Classification and distribution of known plant repeats in tomato chromosome 6. Numbers represent percentages of nucleotides that show similarity to a repeat of the indicated category. An ‘x’ represents the absence of a repeat family; ‘0.00’ indicates that the repeat is present, but at a frequency lower than 0.005 % of the total sequence length. Chr6: chromosome 6; short: short arm euchromatin; hetero: pericentromeric heterochromatin; long: long arm euchromatin.

	Chr6 short	Chr6 hetero	Chr6 long
<b>Class I retrotransposons</b>	<b>13.48</b>	<b>24.37</b>	<b>4.81</b>
<b>LTR retrotransposons</b>	<b>12.95</b>	<b>23.85</b>	<b>4.13</b>
Ty1/Copia	7.65	8.98	2.63
Ty3/Gypsy	5.19	14.71	1.45
Unclassified	0.11	0.16	0.05
<b>Non-LTR retrotransposons</b>	<b>0.53</b>	<b>0.52</b>	<b>0.68</b>
LINE	0.11	0.37	0.20
SINE	0.42	0.15	0.48
<b>Class II DNA transposons</b>	<b>2.07</b>	<b>2.05</b>	<b>1.59</b>
En-Spm	0.12	0.73	0.21
Harbinger	0.03	0.06	0.08
Helitron	0.08	0.08	0.03
Mariner	x	0.00	0.01
MuDR	0.13	0.33	0.11
Pogo	0.09	0.02	0.11
Stowaway	0.03	0.03	0.02
hAT	0.13	0.09	0.04
hAT-Ac	0.39	0.25	0.13
hAT-Tip100	x	0.01	0.06
Unclassified	1.07	0.45	0.79
<b>Satellites</b>	<b>0.00</b>	<b>0.01</b>	<b>0.00</b>
Centromeric	x	x	x
Unclassified	x	0.01	0.00
<b>Ribosomal genes</b>	<b>0.01</b>	<b>0.00</b>	<b>0.01</b>
rRNA	0.01	x	0.01
<b>Unclassified</b>	<b>0.24</b>	<b>0.21</b>	<b>0.07</b>
RC/Helitron	0.24	0.21	0.06
Unknown	0.00	x	0.01
<b>Total</b>	<b>13.35</b>	<b>22.40</b>	<b>5.66</b>

**Supplementary Table 4.3:** Number of (near-) complete BLASTN matches to several known tomato repeats in the chromosome 6 sequence data. Matches to members of the same family (e.g. Sol3, SolSINE, TAPIR) may overlap with each other. SE: short arm euchromatin; H: pericentromeric heterochromatin; LE: long arm euchromatin.

Repeat	Genbank	Description	SE	H	LE
TGRI	X87233	Telomere-associated satellite	0	0	23
TGRII	X90770	Ty3/Gypsy (fragment)	5	11	4
TGRIII	AY880063	Ty3/Gypsy (fragment)	1	11	2
TGRIV	EU526907	Ty3/Gypsy, centromere-associated	0	4	0
Jinling	DQ445619	Ty3/Gypsy	5	9	3
Rider	EF094939	Ty1/Copia	5	1	2
Sol3	AF043122	Non-autonomous transposable element	1	9	8
LeFTB-Sol3	U75644	Sol3-type non-autonomous transposable element	0	0	0
Les0.5-1	U91989	Sol3-type non-autonomous transposable element	1	9	7
Les0.5-2	U91990	Sol3-type non-autonomous transposable element	1	8	6
Les0.6	U91991	Sol3-type non-autonomous transposable element	1	9	7
SoFT1	U75644	Foldback transposon	0	0	1
SoFT2	X80908	Foldback transposon	0	0	0
SolSINE1	U75644	SINE	2	0	4
SolSINE2	U75644	SINE	4	0	10
SolSINE3	U75644	SINE	4	0	11
TAPIR_chr06	AY678298	MITE	17	15	31
TAPIR_chr07	AJ439079	MITE	12	12	25
TAPIR_chr10_1	EF094939	MITE	9	12	22
TAPIR_chr10_2	AJ583670	MITE	15	15	34
TAPIR_chr11_1	AF275345	MITE	15	15	31
TAPIR_chr11_2	AF275345	MITE	17	15	34
TAPIR_chr11_3	AF275345	MITE	16	15	34
Tork2	EU090224	TRIM	3	0	9



## Chapter 5

# Genome sequence and analysis of the tuber crop potato

*The Potato Genome Sequencing Consortium*

A modified version is published in *Nature* (2011) 475, 189-195

### Author contribution

Erwin Datema performed the assembly, super-scaffolding and base error correction of the BAC tiling paths of the RH genotype; performed the analysis and preliminary assembly of the Illumina data derived from the RH genotype; contributed to the *k*-mer analyses of the DM and RH genotypes; performed the sequence comparison between the DM and RH genotypes, and between the RH linkage phases; contributed to the study of haplotype diversity between the genotypes; was contributing author for the sections of Chapter 5 relevant to these analyses; and was lead author for the sections of Supplementary materials included in this chapter.

### Summary

Potato (*S. tuberosum* L.) is the world's most important non-grain food crop and central to global food security. It is a clonally propagated, highly heterozygous autotetraploid, and suffers acute inbreeding depression. We exploited a homozygous doubled-monoploid potato clone to sequence and assemble 86% of the 844 Mb genome. We predict 39,031 protein-coding genes and present evidence for at least two genome duplication events indicative of a paleopolyploid origin. As the first genome sequence of an Asterid, the potato genome reveals 2,642 genes specific to this large Angiosperm clade. We also sequenced a heterozygous diploid clone and show that gene presence/absence variants and other potentially deleterious mutations occur frequently and are a likely cause of inbreeding depression. Gene family expansion, tissue specific expression, and recruitment of genes to new pathways contributed to the evolution of tuber development. Access to the potato genome provides a platform for improvement of this vital crop.

## Introduction

Potato (*S. tuberosum* L.) is a member of the Solanaceae, an economically important family that includes tomato, pepper, eggplant, petunia, and tobacco. Potato belongs to the Asterid clade of Eudicot plants that represents ~25% of flowering plant species and from which a complete genome sequence is not yet published. Potato occupies a wide eco-geographical range [224] and is unique amongst the major world food crops in producing stolons (underground stems) that under suitable environmental conditions swell to form tubers. Its worldwide importance, especially within the developing world, is growing rapidly with production in 2009 reaching 330 million tons [17]. The tubers are a globally important dietary source of starch, protein, antioxidants, and vitamins [225], serving the plant as both a storage organ and vegetative propagation system. Despite their importance, the evolutionary and developmental mechanisms that lead to the initiation and growth of tubers remain elusive.

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 autotetraploid ( $2n=4x=48$ ), 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 19<sup>th</sup> century. Together, these attributes present a significant barrier to potato improvement using classical breeding approaches and a challenge to the scientific community is to obtain a genome sequence that will ultimately facilitate advances in breeding programs.

To overcome the key issue of heterozygosity and allow us to generate a high quality draft potato genome sequence, we used a unique homozygous form of potato called a doubled-monoploid derived using classical tissue culture techniques [226]. 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 both genotypes, allowed us to explore potato genome structure and organization, as well as key aspects of the biology and evolution of this important crop.

## Genome sequencing, assembly and annotation

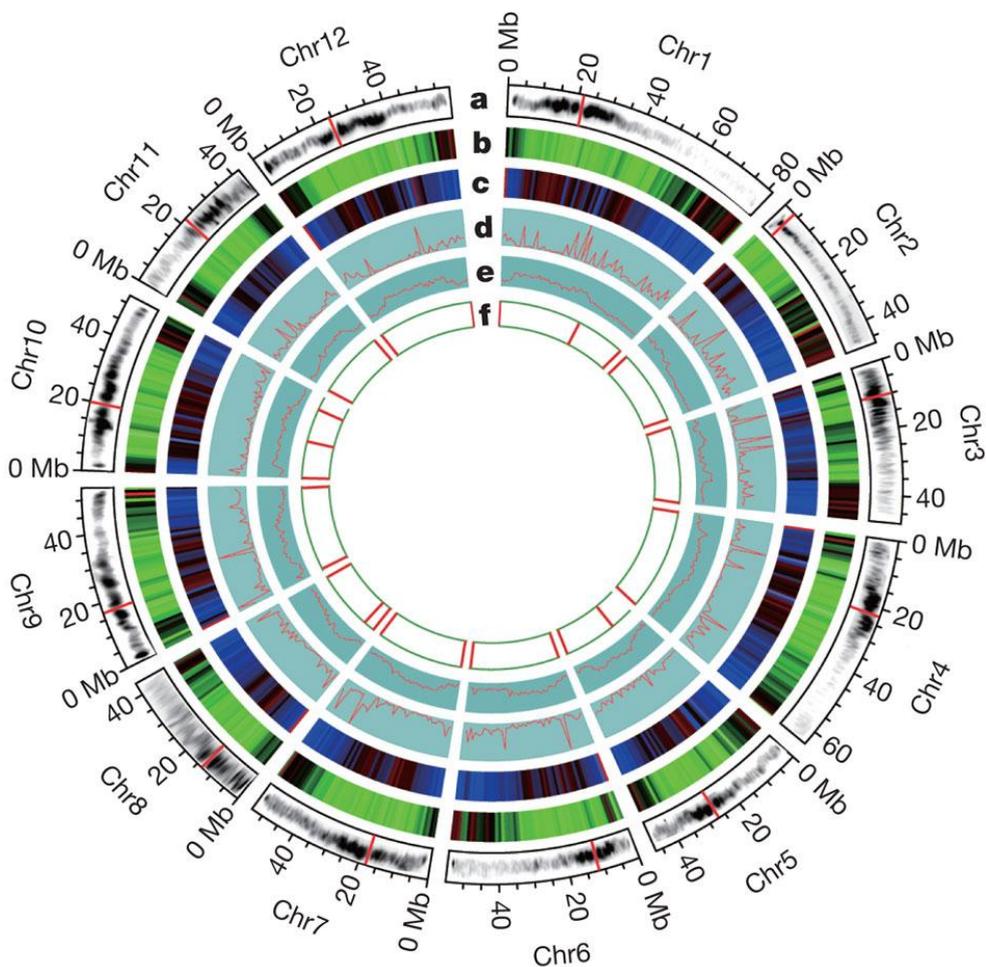
We sequenced the nuclear and organellar genomes of DM employing a whole genome shotgun sequencing approach. We generated 96.6 Gb of raw sequence from two next

generation sequencing platforms, Illumina Genome Analyzer and Roche Pyrosequencing, as well as conventional Sanger sequencing technologies. The genome was assembled using SOAPdenovo [52] resulting in a final assembly of 727 Mb, of which, 93.9% is non-gapped sequence. Ninety percent of the assembly falls into 443 superscaffolds larger than 349 kb. The 17-mer depth distribution suggests a genome size of 844 Mb, consistent with estimates from flow cytometry [227]. Our assembly of 727 Mb is 117 Mb less than the estimated genome size. Analysis of the DM scaffolds indicate 62.2% repetitive content in the assembled section of the DM genome, less than the 74.8% estimated from BAC and fosmid end sequences, suggesting that much of the unassembled genome is composed of repetitive sequences.

We assessed the quality of the WGS assembly through alignment to Sanger-derived phase 2 BAC sequences. In an alignment length of ~1 Mb (99.4% coverage), no gross assembly errors were detected. Alignment of fosmid and BAC paired-end sequences to the WGS scaffolds revealed limited ( $\leq 0.12\%$ ) potential misassemblies. Extensive coverage of the potato genome in this assembly was confirmed using available Expressed Sequence Tag (EST) data; 97.1% of 181,558 available Sanger-sequenced *S. tuberosum* ESTs (>200bp) were detected. 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 elements (TE) classes, representing 29.4% of the genome. In addition, subtelomeric repeats were identified at or near chromosomal ends (Figure 5.1). Using a newly constructed genetic map based on 2,603 polymorphic markers, we were able to genetically anchor 623 Mb (86%) of the assembled genome, and construct pseudomolecules for each of the 12 chromosomes (Figure 5.1) which harbour 90.3% of the predicted genes.

To aid annotation and address a series of biological questions, we generated 31.5 Gb of RNA-seq data from 32 DM and 16 RH libraries representing all major tissue types, developmental stages, and responses to abiotic and biotic stresses. For annotation, reads were mapped against the DM genome sequence (90.2% of 824,621,408 DM reads and 88.6% of 140,375,647 RH reads) and in combination with *ab initio* gene prediction, protein and EST alignments, we annotated 39,031 protein-coding genes. RNA-seq data revealed abundant alternative splicing; 23,617 genes (60.5%) contained two or more isoforms, indicative of substantially more functional variation than represented by the gene set alone. Overall, 87.9% of the gene models were supported by transcript and/or protein similarity with only 12.1% derived solely from *ab initio* gene predictions.

Karyotypes of RH and DM suggested similar heterochromatin content [228] with large blocks of heterochromatin located at the pericentromeric regions (Figure 5.1). As observed in other plant genomes, there was an inverse relationship between gene density and repetitive sequences (Figure 5.1). However, many predicted genes in heterochromatic regions are expressed, consistent with observations in tomato (Chapter 4) that genic ‘islands’ are present in the heterochromatic ‘ocean’.



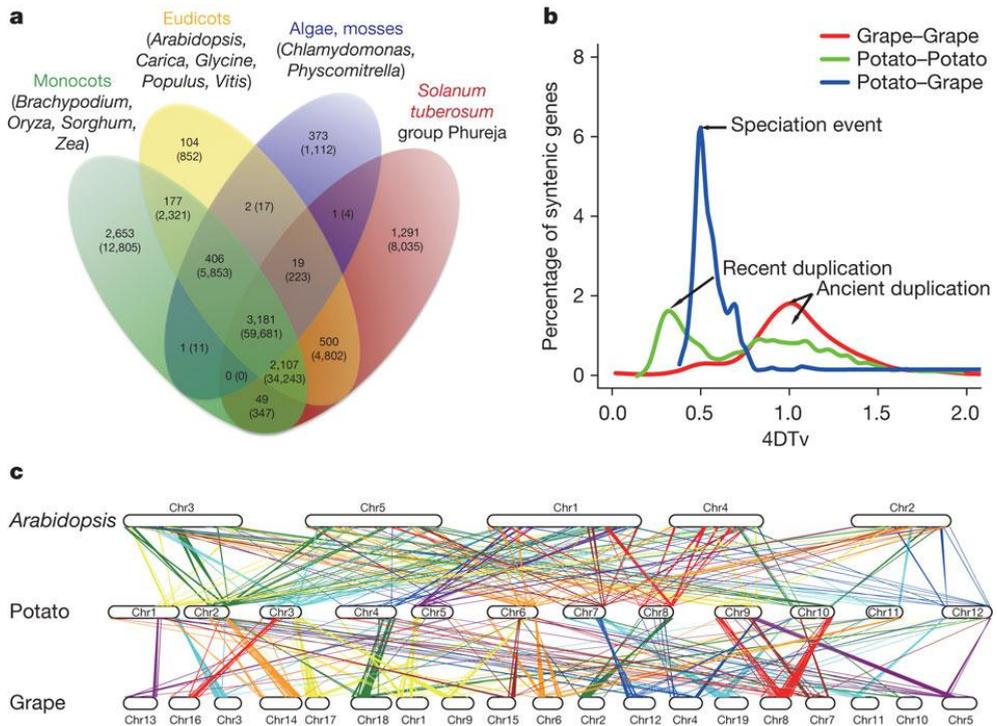
- a** Chromosome karyotype | centromere
- b** Gene density (genes per Mb) 0 76
- c** Repeats coverage (%) 0 100
- d** Transcription state  
 (x10,000 FKPM, bin = 1 Mb)
- e** GC content  
 (GC%, bin = 1 Mb)
- f** Subtelomeric repeats distribution |

**Figure 5.1:** The potato genome. a, Ideograms of the 12 pseudochromosomes of potato (in Mb scales). Each of the 12 pachytene chromosomes from DM was digitally aligned with the ideogram (the amount of DNA in each unit of the pachytene chromosomes is not in proportion to the scales of the pseudochromosomes). b, Gene density represented as number of genes per Mb (non-overlapping, window size = 1 Mb). c, Percentage of coverage of repetitive sequences (nonoverlapping windows, window size = 1 Mb). d, Transcription state. The transcription level for each gene was estimated by averaging the fragments per kb exon model per million mapped reads (FPKM) from different tissues in non-overlapping 1 Mb windows. e, GC content was estimated by the percent G+C in 1 Mb non-overlapping windows. f, Distribution of the subtelomeric repeat sequence CL14\_cons.

## Genome evolution

Potato is the first sequenced genome of an Asterid, a clade within Eudicots that encompasses nearly 70,000 species characterised by unique morphological, developmental and compositional features [229]. Orthologous clustering of the predicted potato proteome with 11 other green plant genomes revealed 4,479 potato genes in 3,181 families in common (Figure 5.2a); 24,051 potato genes clustered with at least one of the 11 genomes. Filtering against transposable elements and 153 non-asterid and 57 asterid publicly available transcript sequence sets yielded 2,642 high confidence asterid-specific and 3,372 potato lineage-specific genes; both sets were enriched in genes with no known function and had less expression support than the core Viridiplantae genes. Genes encoding transcription factors, self-incompatibility, and defence-related proteins were evident in the asterid-specific gene set and presumably contributing to unique characteristics of the Asteridae.

Structurally, we identified 1,811 syntenic gene blocks involving 10,046 genes in the potato genome. Based on these pairwise paralogous segments, we calculated an age distribution based on the number of transversions at four-fold degenerate sites (4DTv) for all duplicate pairs. In general, two significant groups of blocks are seen in the potato genome (4DTv  $\sim$ 0.36 and  $\sim$ 1.0, Figure 5.2b) suggestive of two whole genome duplication (WGD) events. We also identified collinear blocks between potato and three rosid genomes (Grape, *A. thaliana* and *Populus*) that also suggest both events (Figure 5.2c). The ancient WGD corresponds to the ancestral hexaploidization ( $\gamma$ ) event in grape (Figure 5.2b), consistent with a previous report based on EST analysis that the two main branches of eudicots, the Asterids and Rosids, may share the same palaeo-hexaploid duplication event [230]. The  $\gamma$  event likely occurred after the divergence between dicots and monocots about  $185 \pm 55$  million years ago (Mya) [13]. The recent duplication can therefore be placed at  $\sim$ 67 Mya, consistent with the WGD that occurred near the Cretaceous-Tertiary boundary ( $\sim$ 65 Mya) [231]. The divergence of potato and grape occurred at  $\sim$ 89 Mya (4DTv  $\sim$ 0.48), which is likely to represent the split between the Rosids and Asterids.



**Figure 5.2:** Comparative analyses and evolution of the potato genome. a, Clusters of orthologous and paralogous gene families in 12 plant species as identified by OrthoMCL [232]. Gene family number is listed in each of the components; the number of genes within the families for all of the species within the component is noted within parentheses. b, Genome duplication in dicot genomes as revealed through 4DTV analyses. c, Syntenic blocks between *A. thaliana*, potato, and *Vitis vinifera* (grape) demonstrating a high degree of conserved gene order between these taxa.

## Haplotype diversity and inbreeding depression

High heterozygosity and inbreeding depression are inherent to potato, a species that predominantly outcrosses and propagates by means of vegetative organs. Indeed, the phenotypes of DM and RH differ, with RH more vigorous than DM (Figure 5.3a). To explore the extent of haplotype diversity and possible causes of inbreeding depression, we sequenced and assembled 1,644 RH BAC clones generating 178 Mb of non-redundant sequence from both haplotypes (~10% of the RH genome with uneven coverage). After filtering to remove repetitive sequences, we aligned 99 Mb of RH sequence (55%) to the DM genome. These regions were largely collinear with an overall sequence identity of 97.5%, corresponding to one SNP every 40 bp and one InDel every 394 bp (average length 12.8 bp). Between the two RH haplotypes, 6.6 Mb of sequence could be aligned with 96.5% identity, corresponding to 1 SNP per 29 bp and 1 InDel per 253 bp (average length

10.4 bp). Additional information on RH BAC assembly and comparison with the DM genome is found in Supplementary materials, Supplementary Tables 5.1 through 5.3 and Supplementary Figure 5.1.

Current algorithms are of limited use in *de novo* whole genome assembly or haplotype reconstruction of highly heterozygous genomes such as RH, as shown by *k*-mer-count histograms (Figure 5.3b). To complement the BAC-level comparative analysis and provide a genome-wide perspective of heterozygosity in RH, we mapped 1,118 million whole genome NGS reads from RH (84× coverage) onto the DM assembly. A total of 457.3 million reads uniquely aligned providing 90.6% (659.1 Mb) coverage. We identified 3.67 million SNPs between DM and one or both haplotypes of RH, with an error rate of 0.91% based on evaluation of RH BAC sequences. We used this dataset to explore the possible cause of inbreeding depression by quantifying the occurrence of premature stop, frameshift, and PAV [233] as they disable gene function and contribute to genetic load. We identified 3,018 SNPs predicted to induce premature stop codons in RH, with 606 homozygous (in both haplotypes) and 2,412 heterozygous. In DM, 940 premature stop codons were identified. In the 2,412 heterozygous RH premature stop codons, 652 were shared with DM and the remaining 1,760 were found in RH only (Figure 5.3c). Frameshift mutations were identified in 80 loci within RH, 31 heterozygous and 49 homozygous, concentrated in seven genomic regions (Figure 5.3c). Finally, we identified PAV for 275 genes; 246 were RH-specific (absent in DM) and 29 were DM-specific, with 125 and 9 supported by RNA-seq and/or Gene Ontology [234] annotation for RH and DM, respectively. Collectively, these data suggest that the complement of deleterious alleles in DM may be responsible for its reduced level of vigor (Figure 5.3a).

The divergence between potato haplotypes is similar to that reported between out-crossing maize accessions [235] and coupled with our inability to successfully align 45% of the BAC sequences, intra- and inter-genome diversity appears to be a significant feature of the potato genome. A detailed comparison of the three haplotypes (DM, and two haplotypes of RH) at two genomic regions (334 kb in length) using RH BAC sequence (Figure 5.3d) revealed considerable sequence and structural variation. In one region ('euchromatic' Figure 5.3d) we observed one instance of copy number variation, five genes with premature stop codons, and seven RH-specific genes. These observations suggest that the plasticity of the potato genome is greater than revealed from the unassembled RH NGS. Improved assembly algorithms, increased read lengths, and *de novo* sequence of additional haplotypes will reveal the full catalogue of genes critical to inbreeding depression.

## **Tuber biology**

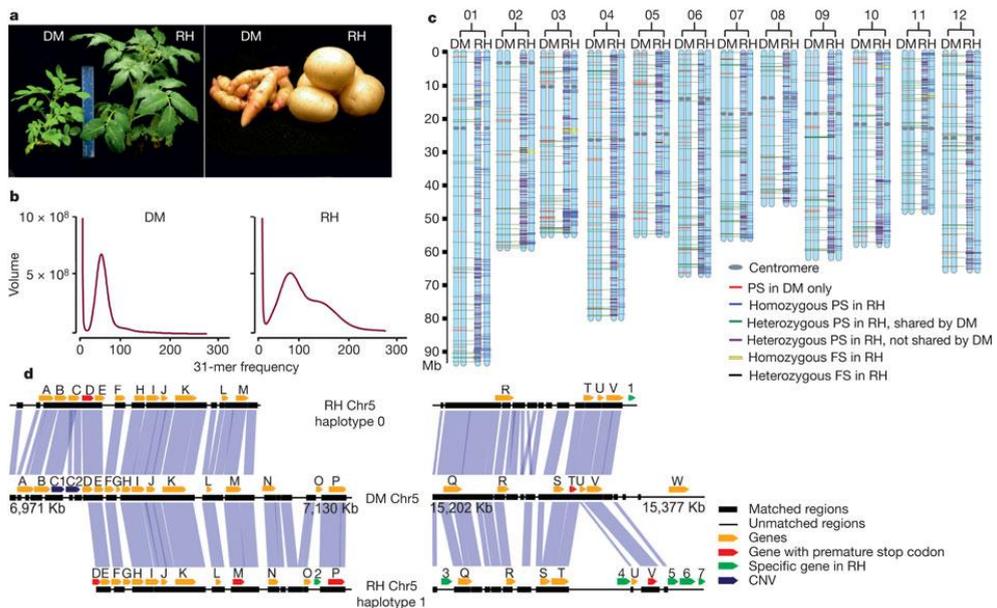
In developing DM and RH tubers, 15,235 genes were expressed in the transition from stolons to tubers, with 1,217 transcripts exhibiting >5-fold expression in stolon vs. five RH

tuber tissues (young tuber, mature tuber, tuber peel, cortex and pith). Of these, 333 transcripts were up-regulated during the transition from stolon into tuber, with the most highly up-regulated transcripts encoding storage proteins. Foremost amongst 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 [236]. 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. The KTI genes are distributed across the genome with individual members exhibiting specific expression patterns (Figure 5.4a, b). KTIs are frequently induced following pest and pathogen attack and act primarily as inhibitors of exogenous proteinases [237], therefore the expansion of the KTI family may provide resistance to biotic stress for the newly evolved vulnerable underground organ.

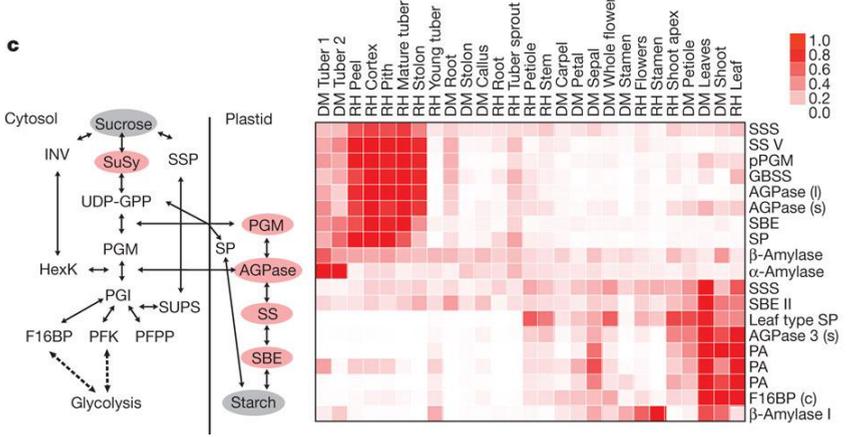
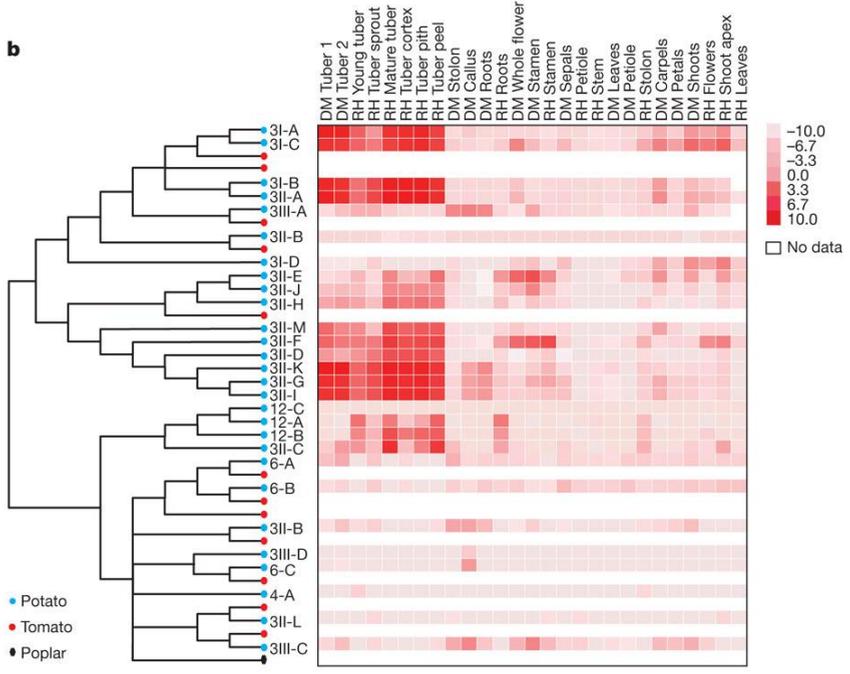
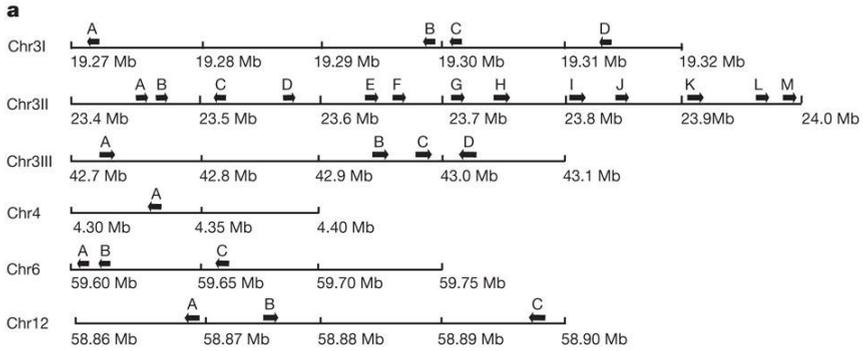
The stolon to tuber transition also coincides with a strong up-regulation of genes associated with starch biosynthesis (Figure 5.4c). We observed several starch biosynthetic genes that were 3-8 fold more highly expressed in tuber tissues of RH compared to DM (Figure 5.4c). Together this suggests a stronger 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. This contrasts with the cereal endosperm where carbon is transported into the amyloplast in the form of ADP-glucose via a specific transporter (Brittle 1 protein; [238]). Carbon transport into the amyloplasts of potato tubers is primarily in the form of glucose-6-phosphate [239] although recent evidence suggests glucose-1-phosphate is quantitatively important under certain conditions [240]. The transport mechanism for glucose 1-phosphate is unknown and the genome sequence contains six genes for hexose-phosphate transporters with two highly and specifically expressed in stolons and tubers. Furthermore, an additional 23 genes encode proteins homologous to other carbohydrate derivative transporters such as triose phosphate, phosphoenolpyruvate, or UDP-glucuronic acid transporters and two loci with homologues for the Brittle 1 protein. By contrast, in leaves, carbon fixation specific genes such as plastidic aldolase, fructose 1,6-biphosphatase and distinct leaf isoforms of starch synthase, starch branching enzyme, starch phosphorylase and ADP-glucose pyrophosphorylase were up-regulated. Of particular interest is the difference in tuber expression of enzymes involved in the hydrolytic and phosphorolytic starch degradation pathways. Considerably greater levels of  $\alpha$ -amylase (10-25 fold) and  $\beta$ -amylase (5-10 fold) mRNAs were found in DM tubers compared to RH whereas  $\alpha$ -1,4 glucan phosphorylase mRNA was equivalent in DM and RH tubers. These gene expression differences between the breeding line RH and the more primitive DM are consistent with the concept that increasing tuber yield may be partially attained by selection for decreased activity of the hydrolytic starch degradation pathway.

Recent studies using a potato genotype strictly dependent upon short days for tuber induction (*S. tuberosum* Group Andigena) identified a potato homologue (*StSP6A*) of *A. thaliana* FLOWERING TIME LOCUS T (FT) as the long distance tuberisation inductive

signal. *StSP6A* is produced in the leaves consistent with its role as the mobile signal (Navarro et al., pers. comm.). *SP/FT* is multi-gene family and expression of a second FT homologue, *StSP5G*, in mature tubers suggests a possible function in the control of tuber sprouting, a photoperiod-dependent phenomenon [241]. Likewise, expression of a homologue of the *A. thaliana* flowering time MADS box gene *SOC1*, acting downstream of *FT* [242], is restricted to tuber sprouts. Expression of a third FT homologue, *StSP3D*, does not correlate with tuberisation induction but instead with transition to flowering which is regulated independently of day length (Navarro et al., pers. comm.). These data suggest that neofunctionalization of the day-length dependent flowering control pathway has occurred in potato to control formation and possibly sprouting of a novel storage organ, the tuber.



**Figure 5.3:** Haplotype diversity and inbreeding depression. a, Plants and tubers of DM and RH showing that RH has greater vigor. b, Illumina *k*-mer volume histograms of DM and RH. The volume of *k*-mers (y-axis) is plotted against the frequency at which they occur (x-axis). The leftmost truncated peaks at low frequency and high volume represent *k*-mers containing essentially random sequencing errors, whereas the distribution to the right represents proper (putatively error-free) data. In contrast to the single modality of DM, RH exhibits clear bimodality caused by heterozygosity. c, Genomic distribution of premature stop (PS), frame-shift (FS), and presence/absence variation (PAV) contributing to inbreeding depression. The hypothetical RH pseudomolecules were solely inferred from the corresponding DM ones. Due to the inability to assign heterozygous PS and FS of RH to a definite haplotype, all heterozygous PS and FS were arbitrarily mapped to the left haplotype of RH. d, A zoom-in comparative view of the DM and RH genomes. The left and right alignments are derived from the euchromatic and heterochromatic regions of chromosome 5, respectively. Most of the gene annotations, including PS and RH-specific genes, are supported by transcript data.



**Figure 5.4:** Gene expression of selected tissues and genes. a, Kunitz trypsin inhibitor (KTI) gene organization across the potato genome. Black arrows indicate the location of individual genes on six scaffolds located on four chromosomes. b, Phylogenetic tree and KTI gene expression heat map. The KTI genes were clustered using all potato and tomato genes available with the *Populus* KTI gene as an out-group. The tissue specificity of individual members of the highly expanded potato gene family is shown in the heat map. Expression levels are indicated by shades of red, where white indicates no expression or lack of data for tomato and poplar. c, A model of starch synthesis showing enzyme activities is shown on the left. AGPase, ADP-glucose pyrophosphorylase; F16BP, fructose-1,6-biphosphatase; HexK, hexokinase; INV, invertase; PFK, phosphofructokinase; PFPP, pyrophosphatefructose-6-phosphate-1-phosphotransferase; PGI, phosphoglucose isomerase; PGM, phosphoglucomutase; SBE, starch branching enzyme; SP, starch phosphorylase; SPP, sucrose phosphate phosphatase; SS, starch synthase; SuSy, sucrose synthase; SUPS, sucrose phosphate synthase; UDP-GPP, UDP-glucose pyrophosphorylase. The grey background denotes substrate (sucrose) and product (starch) and red background indicates genes that are specifically up-regulated in RH versus DM. On the right, a heat map of the genes involved in carbohydrate metabolism is shown. ADP-glucose pyrophosphorylase large subunit, AGPase (l); ADP-glucose pyrophosphorylase small subunit, AGPase (s); ADP-glucose pyrophosphorylase small subunit 3, AGPase 3 (s); cytosolic fructose-1,6-biphosphatase, F16BP (c); granule bound starch synthase, GBSS; leaf type L starch phosphorylase, Leaf type SP; plastidic phosphoglucomutase, pPGM; starch branching enzyme II, SBE II; soluble starch synthase, SSS; starch synthase V, SSV; three variants of plastidic aldolase, PA.

## Disease resistance

Potato is susceptible to a wide range of pests and pathogens and the identification of genes conferring disease resistance has been a major focus of the research community. Most cloned disease resistance genes in the Solanaceae encode nucleotide-binding site (NBS) and leucine-rich-repeat (LRR) domains. The DM assembly contains 408 NBS-LRR-encoding genes, 57 Toll Interleukin-like receptor (TIR) and 351 non-TIR types, similar to the 402 resistance (*R*) gene candidates in *Populus* [243]. Highly related homologues of the cloned potato late blight resistance genes *R1*, *RB*, *R2*, *R3a*, *Rpi-blb2* and *Rpi-vnt1.1* were present in the assembly. In RH, the chromosome 5 *R1* cluster contains two distinct haplotypes; one is collinear with the *R1* region in DM, yet neither the DM or RH *R1* regions are collinear with the *R1* region in cultivated potato or *Solanum demissum* [244, 157]. Comparison of the DM potato *R* gene sequences with well-established gene models (functional *R* genes) indicates that many (39.4%) of NBS-LRR genes are pseudogenes due to InDels, frame shift mutations, or premature stop codons including the *R1*, *R3a* and *Rpi-vnt1.1* clusters which contain extensive chimeras and exhibit evolutionary patterns of Type I *R* genes [245]. This high rate of pseudogenization parallels the rapid evolution of effector genes observed in the potato late blight pathogen, *Phytophthora infestans* [246]. High haplotype diversity coupled with a high pseudogenization rate suggests that tetraploid potato may contain thousands of *R* genes.

## Conclusions and future directions

We sequenced a unique doubled monoploid potato clone to overcome the problems associated with genome assembly due to high levels of heterozygosity and were able to generate a high quality draft potato genome sequence that provides new insights into Eudicot genome evolution. Using a combination of data from the vigorous, heterozygous diploid RH and relatively weak, doubled monoploid DM, we could 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 tuberisation. This evolutionary innovation evolved exclusively in the *Solanum* section *Petota* that encompasses ~200 species distributed from the southwestern United States to central Argentina and Chile. Neighboring *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 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. While most 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.

## Methods

DM1-3 516 R44 (DM) resulted from chromosome doubling of a monoploid ( $1n=1x=12$ ) derived by anther culture of a heterozygous diploid ( $2n=2x=24$ ) *S. tuberosum* Group Phureja clone (PI 225669) [247]. RH89-039-16 (RH) is a diploid clone derived from a cross between a *S. tuberosum* ‘dihaploid’ (SUH2293) and a diploid clone (BC1034) generated from a cross between two *S. tuberosum* × *S. tuberosum* Group Phureja hybrids. Sequence data from three platforms, Sanger, Roche 454 Pyrosequencing, and Illumina Sequencing-by-Synthesis, were used to assemble the DM genome using the SOAPdenovo assembly algorithm [52]. The RH genotype was sequenced using shotgun sequencing of BACs and WGS in which reads were mapped to the DM reference assembly. Superscaffolds were anchored to the 12 linkage groups using a combination of *in silico* and genetic mapping data. Repeat sequences were identified through sequence similarity at the nucleotide and protein level [248, 249]. Genes were annotated using a combined approach [250] on the repeat masked genome with *ab initio* gene predictions, protein similarity, and

transcripts to build optimal gene models. Illumina RNA-seq reads were mapped to the DM draft sequence using Tophat [251] and expression levels from the representative transcript were determined using Cufflinks [252]. Genome sequence and annotation can be obtained and viewed at [135]. The DM genome assembly is available in GenBank under ProjectID 63145.

## Acknowledgements

We acknowledge the assistance of Walter Amoros, Beata Babinska, Roman V. Baslerov, Boris K. Bumazhkin, Martín Federico Carboni, Tony Conner, Joseph Coombs, Loretta Daddiego, Juan Martín D'Ambrosio, Gianfranco Diretto, Silvina B. Divito, David Douches, Malgorzata Filipiak, Giulio Gianese, Ronald Hutten, Evert Jacobsen, Ewa Kalinska, Sophien Kamoun, Donna Kells, Helena Kossowska, Loredana Lopez, Maria Magallanes-Lundback, Tomas Miranda, P. S. Naik, Angela N. Panteleeva, D. Pattanayak, Ekaterina O. Patutina, Marina Portantier, Shashi Rawat, Reinhard Simon, B. P. Singh, Brajesh Singh, Willem Stiekema, Marina V. Sukhacheva and Christopher Town in providing plant material, generating data, annotation, analyses, and discussions. We are indebted to additional faculty and staff of the BGI-Shenzhen, J Craig Venter Institute and MSU Research Technology Support Facility who contributed to this project. Background and preliminary data was kindly provided by the Centre for BioSystems Genomics (CBSG), EU-project (APOPHYS EU-QLRT-2001-01849) and U.S. Department of Agriculture National Institute of Food and Agriculture SolCAP project (2008-55300-04757 and 2009-85606-05673). We acknowledge the funding made available by the “863” National High Tech Research Development Program in China (2006AA100107), “973” National Key Basic Research Program in China (2006CB101904, 2007CB815703, 2007CB815705, 2009CB119000), Board of Wageningen University and Research Centre, CAPES - Brazilian Ministry of Education, Chinese Academy of Agricultural Sciences (seed grant to Sanwen Huang), Chinese Ministry of Agriculture (The "948" Program), Chinese Ministry of Finance (1251610601001), Chinese Ministry of Science and Technology (2007DFB30080), China Postdoctoral Science Foundation (20070420446 to Zhonghua Zhang), CONICET (Argentina), DAFF Research Stimulus Fund (07-567), CONICYT-Chile (PBCT-PSD-03), Danish Council for Strategic Research Programme Commission on Health, Food and Welfare (2101-07-0116), Danish Council for Strategic Research Programme Commission on Strategic Growth Technologies (Grant 2106-07-0021), FINCyT ((099-FINCyT-EQUIP-2009)/(076-FINCyT-PIN-2008), Préstamo BID N° 1663/OC-PE, FONDAP and BASAL-CMM)), Fund for Economic Structural Support (FES), HarvestPlus Challenge Program, Indian Council of Agricultural Research, INIA-Ministry of Agriculture of Chile, Instituto Nacional de Innovacion Agraria-Ministry of Agriculture of Peru, Instituto Nacional de Tecnología Agropecuaria (INTA), Italian Ministry of Research (Special Fund for Basic Research), International Potato Center (CIP-CGIAR core funds), LBMG of Center for Genome Regulation and Center for Mathematical

Modeling, Universidad de Chile (UMI 2807 CNRS), Ministry of Education and Science of Russia (contract 02.552.11.7073), National Nature Science Foundation of China (30671319, 30725008, 30890032, 30971995), Natural Science Foundation of Shandong Province in China (Y2006D21), Netherlands Technology Foundation (STW), Netherlands Genomics Initiative (NGI), Netherlands Ministries of Economic Affairs (EZ) and Agriculture (LNV), New Zealand Institute for Crop & Food Research Ltd. Strategic Science Initiative, Perez Guerrero Fund, Peruvian Ministry of Agriculture-Technical Secretariat of coordination with the CGIAR, Peruvian National Council of Science and Technology (CONCYTEC), Polish Ministry of Science and Higher Education (47/PGS/2006/01), Programa Cooperativo para el Desarrollo Tecnológico Agroalimentario y Agroindustrial del Cono Sur (PROCISUR), Project Programa Bicentenario de Ciencia y Tecnología - Conicyt, PBCT - Conicyt PSD-03, Russian Foundation for Basic Research (09-04-12275), Secretaría de Ciencia y Tecnología (SECyT) actual Ministerio de Ciencia y Tecnología (MINCyT), ARGENTINA, Shenzhen Municipal Government of China (CXB200903110066A, ZYC200903240077A, ZYC200903240076A), Solexa project (272-07-0196), Special Multilateral Fund of the Inter-American Council for Integral Development (FEMCIDI), Teagasc, Teagasc Walsh Fellowship Scheme, The New Zealand Institute for Plant & Food Research Ltd. Capability Fund, U.K. Potato Genome Sequencing grant (Scottish Government Rural and Environment Research and Analysis Directorate (RERAD), Department for Environment, Food and Rural Affairs (DEFRA), Agriculture and Horticulture Development Board - Potato Council), U.K. Biotechnology and Biological Sciences Research Council (Grant BB/F012640), U.S. National Science Foundation Plant Genome Research Program (DBI-0604907 /DBI-0834044), Virginia Agricultural Experiment Station USDA Hatch Funds (135853), and Wellcome Trust Strategic award (WT 083481).

### **Authors (listed alphabetically by affiliation)**

**BGI-Shenzhen:** Xun Xu<sup>1</sup>, Shengkai Pan<sup>1</sup>, Shifeng Cheng<sup>1</sup>, Bo Zhang<sup>1</sup>, Desheng Mu<sup>1</sup>, Peixiang Ni<sup>1</sup>, Gengyun Zhang<sup>1</sup>, Shuang Yang (Principal Investigator)<sup>1</sup>, Ruiqiang Li (Principal Investigator)<sup>1</sup>, Jun Wang (Principal Investigator)<sup>1</sup>; **Cayetano Heredia University:** Gisella Orjeda (Principal Investigator, PGSC Steering Committee Member)<sup>2</sup>, Frank Guzman<sup>2</sup>, Michael Torres<sup>2</sup>, Roberto Lozano<sup>2</sup>, Olga Ponce<sup>2</sup>, Diana Martinez<sup>2</sup>, Germán De la Cruz<sup>2</sup>; **Central Potato Research Institute:** S. K. Chakrabarti (Principal Investigator)<sup>3</sup>, Virupaksh U. Patil<sup>3</sup>; **Centre Bioengineering RAS:** Konstantin G. Skryabin (Principal Investigator)<sup>4</sup>, Boris B. Kuznetsov<sup>4</sup>, Nikolai V. Ravin<sup>4</sup>, Tatjana V. Kolganova<sup>4</sup>, Alexey V. Beletsky<sup>4</sup>, Andrei V. Mardanov<sup>4</sup>; **CGR-CMM, Universidad de Chile:** Alex Di Genova<sup>5</sup>; **College of Life Sciences, University of Dundee:** Daniel M. Bolser<sup>6</sup>, David M. A. Martin (Principal Investigator)<sup>6</sup>; **High Technology Research Center, Shandong Academy of Agricultural Sciences:** Guangcun Li<sup>7</sup>, Yu Yang<sup>7</sup>; **Huazhong Agriculture University:** Hanhui Kuang<sup>8</sup>, Qun Hu<sup>8</sup>; **Hunan Agricultural University:** Xingyao Xiong<sup>9</sup>;

**Imperial College London:** Gerard J. Bishop<sup>10</sup>; **Instituto de Investigaciones Agropecuarias:** Boris Sagredo (Principal Investigator)<sup>11</sup>, Nilo Mejía<sup>11</sup>; **Institute of Biochemistry and Biophysics:** Wlodzimierz Zagorski (Principal Investigator)<sup>12</sup>, Robert Gromadka<sup>12</sup>, Jan Gawor<sup>12</sup>, Pawel Szczesny<sup>12</sup>; **Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences:** Sanwen Huang (Principal Investigator, PGSC Steering Committee Member)<sup>13</sup>, Zhonghua Zhang<sup>13</sup>, Chunbo Liang<sup>13</sup>, Jun He<sup>13</sup>, Ying Li<sup>13</sup>, Ying He<sup>13</sup>, Jianfei Xu<sup>13</sup>, Youjun Zhang<sup>13</sup>, Binyan Xie<sup>13</sup>, Yongchen Du<sup>13</sup>, Dongyu Qu (Principal Investigator)<sup>13</sup>; **International Potato Center:** Merideth Bonierbale<sup>14</sup>, Marc Ghislain<sup>14</sup>, Maria del Rosario Herrera<sup>14</sup>; **Italian National Agency for New Technologies, Energy and Sustainable Development:** Giovanni Giuliano (Principal Investigator)<sup>15</sup>, Marco Pietrella<sup>15</sup>, Gaetano Perrotta<sup>15</sup>, Paolo Facella<sup>15</sup>; **J Craig Venter Institute:** Kimberly O'Brien<sup>16</sup>; **Laboratorio de Agrobiotecnología, Instituto Nacional de Tecnología Agropecuaria:** Sergio E. Feingold (Principal Investigator)<sup>17</sup>, Leandro E. Barreiro<sup>17</sup>, Gabriela A. Massa<sup>17</sup>; **Laboratorio de Biología de Sistemas, Universidad Nacional de LaPlata:** Luis Diambra<sup>18</sup>; **Michigan State University:** Brett R. Whitty<sup>19</sup>, Brienne Vaillancourt<sup>19</sup>, Haining Lin<sup>19</sup>, Alicia N. Massa<sup>19</sup>, Michael Geoffroy<sup>19</sup>, Steven Lundback<sup>19</sup>, Dean DellaPenna<sup>19</sup>, C. Robin Buell (Principal Investigator, PGSC Steering Committee Member)<sup>19</sup>; **Scottish Crop Research Institute:** Sanjeev Kumar Sharma<sup>20</sup>, David F. Marshall<sup>20</sup>, Robbie Waugh<sup>20</sup>, Glenn J. Bryan (Principal Investigator, PGSC Steering Committee Member)<sup>20</sup>; **Teagasc Crops Research Centre:** Marialaura Destefanis<sup>21</sup>, Istvan Nagy<sup>21</sup>, Dan Milbourne (Principal Investigator)<sup>21</sup>; **The New Zealand Institute for Plant & Food Research Ltd:** Susan J. Thomson<sup>22</sup>, Mark Fiers<sup>22</sup>, Jeanne M.E. Jacobs (Principal Investigator, PGSC Steering Committee Member)<sup>22</sup>; **University of Aalborg:** Kåre L. Nielsen (Principal Investigator)<sup>23</sup>, Mads Sønderkær<sup>23</sup>; **University of Wisconsin:** Marina Iovene<sup>24</sup>, Giovana A. Torres<sup>24</sup>, Jiming Jiang (Principal Investigator)<sup>24</sup>; **Virginia Polytechnic Institute and State University:** Richard E. Veilleux<sup>25</sup>; **Wageningen University & Research Centre:** Christian W. B. Bachem (Principal Investigator, PGSC Steering Committee Member)<sup>26</sup>, Jan de Boer<sup>26</sup>, Theo Borm<sup>26</sup>, Bjorn Kloosterman<sup>26</sup>, Herman van Eck<sup>26</sup>, Erwin Datema<sup>27</sup>, Bas te Lintel Hekkert<sup>27</sup>, Aska Goverse<sup>28,29</sup>, Roeland C. H. J. van Ham<sup>27,28</sup>, & Richard G. F. Visser<sup>26,27</sup>

<sup>1</sup>BGI-Shenzhen, Chinese Ministry of Agricultural, Key Lab of Genomics, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China; <sup>2</sup>Cayetano Heredia University, Genomics Research Unit, Av Honorio Delgado 430, Lima 31, Peru and San Cristobal of Huamanga University, Biotechnology and Plant Genetics Laboratory, Ayacucho, Peru; <sup>3</sup>Central Potato Research Institute, Shimla 171001, Himachal Pradesh, India; <sup>4</sup>Centre Bioengineering RAS, Prospekt 60-letya Oktyabrya, 7-1, Moscow, Russia, 117312; <sup>5</sup>Center for Genome Regulation and Center for Mathematical Modeling, Universidad de Chile (UMI 2807 CNRS); <sup>6</sup>College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH UK; <sup>7</sup>High Technology Research Center, Shandong Academy of Agricultural Sciences, 11 Sangyuan Road, Jinan, 250100, P. R. China; <sup>8</sup>Huazhong Agriculture University, Ministry of Education, College of Horticulture and Forestry, Department of

Vegetable Crops, Key Laboratory of Horticulture Biology, Wuhan, P.R. China, 430070; <sup>9</sup>Hunan Agricultural University, College of Horticulture and Landscape, Changsha, Hunan, 410128 China; <sup>10</sup>Imperial College London, Division of Biology, South Kensington Campus, London SW7 1AZ; <sup>11</sup>Instituto de Investigaciones Agropecuarias, Avda. Salamanca s/n, Km 105 ruta 5 sur, sector Los Choapinos. Rengo, Región del Libertador Bernardo O'Higgins, Chile, Código Postal 2940000; <sup>12</sup>Institute of Biochemistry and Biophysics, DNA Sequencing and Oligonucleotides Synthesis Laboratory, PAS ul. Pawinskiego 5a, 02-106 Warsaw, Poland; <sup>13</sup>Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Beijing 100081, China; <sup>14</sup>International Potato Center, P.O. Box 1558, Lima 12, Peru; <sup>15</sup>Italian National Agency for New Technologies, Energy and Sustainable Development (ENEA), Casaccia Research Center, Via Anguillarese 301, 00123 Roma, Italy and Trisaia Research Center, S.S. 106 Ionica - Km 419,50 75026 Rotondella (Matera), Italy; <sup>16</sup>J Craig Venter Institute, 9712 Medical Center Dr, Rockville, MD, 20850; <sup>17</sup>Laboratorio de Agrobiotecnología, Estación Experimental Agropecuaria Balcarce, Instituto Nacional de Tecnología Agropecuaria (INTA) cc276 (7620) Balcarce, Argentina; <sup>18</sup>Laboratorio de Biología de Sistemas, CREG, Universidad Nacional de LaPlata, Argentina; <sup>19</sup>Michigan State University, East Lansing MI 48824 USA; <sup>20</sup>Scottish Crop Research Institute, Genetics Programme, Invergowrie, Dundee, DD2 5DA, UK; <sup>21</sup>Teagasc Crops Research Centre, Oak Park, Carlow, Ireland; <sup>22</sup>The New Zealand Institute for Plant & Food Research Ltd., Private Bag 4704, Christchurch 8140, New Zealand; <sup>23</sup>University of Aalborg (AAU), Department of Biotechnology, Chemistry and Environmental Engineering, Sohngaardsholmsvej 49, 9000 Aalborg, Denmark; <sup>24</sup>University of Wisconsin-Madison, Department of Horticulture, 1575 Linden Drive, Madison, WI 53706, USA; <sup>25</sup>Virginia Polytechnic Institute and State University, Department of Horticulture, 544 Latham Hall, Blacksburg VA 24061; <sup>26</sup>Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Plant Breeding, Droevendaalsesteeg 1, 6708PB Wageningen, the Netherlands; <sup>27</sup>Wageningen University and Research Centre, Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands; <sup>28</sup>Centre for BioSystems Genomics, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands; <sup>29</sup>Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Nematology, Droevendaalsesteeg 1, 6708PB Wageningen, the Netherlands.

## **Supplementary materials**

### RH BAC-by-BAC sequencing and assembly

The majority of the RH BACs were assigned to the corresponding chromosomal location on the basis of Amplified Fragment Length Polymorphism markers from the ultra-high

density genetic map, with additional anchor points coming from simple sequence repeat mapping, fluorescence *in situ* hybridizations, or tomato sequences [18, 253]. Sequence overlaps between BACs within the same physical tiling path were identified using megablast from BLAST 2.2.21 [103] and merged with megamerger from the EMBOSS 6.1.0 package [73], and several kilobase-sized gaps were closed through alignment of a preliminary RH whole-genome assembly. The resulting non-redundant contigs were scaffolded by mapping the RH whole-genome Illumina and 454 matepairs against these contigs with SOAPalign 2.20 [254] and subsequently processing these mapping results with a custom Python script. The scaffolds were then ordered into super-scaffolds based on the BAC order in the tiling paths of the FPC map. In total, 1,644 BACs were sequenced, assembled and super-scaffolded into 674 tiling paths spanning 178 Mb of non-redundant sequence from both linkage phases combined (Supplementary Tables 5.1 and 5.2). These tiling paths represent approximately 21% of the 850 Mb RH genome, and due to the sequencing approach have an uneven distribution over the chromosomes.

Consensus base calling errors in the BAC sequences were corrected using custom Python and C scripts using an approach similar to the one described in [60]. Briefly, 31-mers were extracted from the RH whole-genome Illumina GA2 paired-end data and 31-mers with a frequency lower than three were discarded. The remaining 31-mers were subsequently aligned to the BAC sequences, after which erroneous base-calls in the BAC sequences were detected as positions that were covered by less than ten distinct, overlapping 31-mers. Errors were corrected by changing the nucleotide at that position through substitution, insertion or deletion, such that valid 31-mer coverage was maximized. In total, 17,114 base-calling errors were repaired, corresponding to one corrected error every 12 kb (Supplementary Table 5.3). Approximately 45% of these involved deletions at contig edges, which is likely the result of incorrect base calls due to low sequence coverage in these regions.

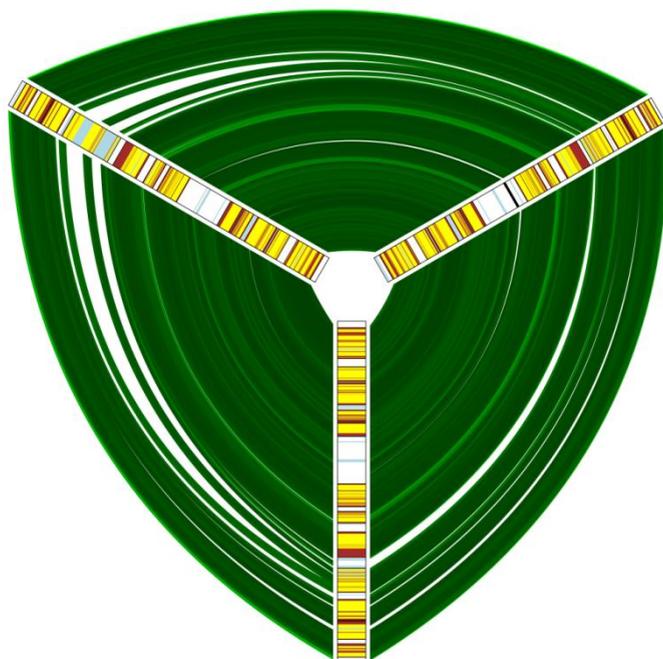
### Comparison of DM and RH assemblies

Superscaffolds of DM and RH were aligned using megablast with a minimum sequence identity of 90% to identify potential collinear segments. Repetitive matches between a single RH superscaffold and multiple DM superscaffolds, and matches spanning less than 20 kb, were discarded using custom Python scripts. The resulting groups of potentially collinear superscaffolds were subjected to a more accurate alignment using lastz [111] with the "--gextend --chain --gapped" parameters. These alignments were filtered for a minimum match length of 1,000 bp and a minimum sequence identity of 90%, and repetitive matches were discarded using custom Python scripts. Sequence polymorphism and InDel frequencies were derived from the lastz alignments using custom python scripts. To avoid overestimation of small InDels due to sequencing errors, InDels of 1 and 2 bp were not considered if the distance between them was less than 30 bp or the identity of 10 bp of flanking sequence on each side was smaller than 95%. The same methodology was

applied to investigate the sequence diversity between the corresponding linkage-specific superscaffolds of RH.

### Haplotype diversity

After filtering to remove repetitive sequences, 99 Mb out of 178 Mb of the RH sequence aligned to the DM genome. These regions consisted primarily of low-repetitive sequence from the euchromatic distal ends of chromosomes and were largely collinear between the two genomes. The majority of sequence polymorphism was present outside protein-coding regions: whereas the overall identity between DM and RH sequence measured 96.5%, in coding regions there was up to 98.4% identity. There were only few regions of the genome for which the sequence was present from all three potato haplotypes, and these were concentrated primarily on chromosomes 1 and 5. As an example, Supplementary Figure 5.1 displays a three-way alignment of a region of 100 kb on chromosome 5 for which sequence from all three haplotypes was available.



**Supplementary Figure 5.1:** Three-way alignment of an 100 kb region on chromosome 5 in DM (bottom) and the corresponding regions in RH linkage phase 0 (top left) and RH linkage phase 1 (top right) with genes as red (exon) and yellow (intron) bars, repeat elements as blue bars and sequence gaps as black bars. The alignments between the three genotypes are represented in green with the intensity indicating the percent identity, ranging from 90% (bright green) to 100% (dark green). Unaligned regions are represented in white. The gene order is fully conserved between the three allelic potato sequences, but one gene is interrupted by repeat insertions in RH linkage phase 0.

**Supplementary Table 5.1:** RH sequence data of BACs, organized per chromosome. All BACs not assigned to any chromosome were combined into a virtual chromosome 0.

	Chromosome												
	1	2	3	4	5	6	7	8	9	10	11	12	0
Sequences	3,390	174	430	767	5,810	2,111	313	167	1,366	374	600	752	974
Total length (Mb)	39.4	2.2	6.0	10.3	69.0	22.5	8.0	1.6	13.6	4.0	8.1	6.6	11.8
Minimum length	35	507	286	198	112	84	364	528	29	199	3	32	33
Maximum length	155,460	79,841	120,722	107,847	134,177	112,064	175,778	149,534	133,864	76,465	151,103	146,171	122,038
Average length	11,627	12,358	14,032	13,437	11,884	10,676	25,542	9,573	9,944	10,658	13,531	8,774	12,126
Median length	6,382	6,691	6,960	8,519	6,884	5,103	9,781	4,698	4,299	5,968	5,052	1,718	6,285
N50 length	23,393	27,542	29,007	23,411	21,303	22,057	69,670	18,160	21,987	21,596	37,414	33,061	25,332
GC content	35.95	35.10	34.46	34.63	35.51	35.16	34.88	35.10	34.82	37.39	34.55	34.74	36.96
N content	0.00	0.00	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.39	0.08

66

**Supplementary Table 5.2:** RH sequence data of the superscaffolds generated from the BACs, organized per chromosome. All superscaffolds consisting of BACs not assigned to any chromosome were combined into a virtual chromosome 0.

	Chromosome												
	1	2	3	4	5	6	7	8	9	10	11	12	0
Sequences	464	55	92	117	1,016	508	94	38	442	109	151	387	295
Total length (Mb)	35.2	2.1	6.2	8.6	55.3	21.3	7.7	1.6	12.2	3.6	7.6	5.5	11.1
Minimum length	35	567	286	198	296	84	366	955	29	199	3	32	33
Maximum length	642,579	358,478	294,220	690,852	643,428	448,350	502,327	262,308	367,793	198,610	292,582	312,227	435,238
Average length	75,851	39,032	67,397	73,530	54,427	41,839	81,712	43,006	27,653	33,277	50,448	14,200	37,622
Median length	28,259	5,336	37,150	30,589	13,181	5,452	58,981	22,820	1,762	9,278	14,948	1,252	11,671
N50 length	174,853	123,371	140,403	150,765	167,097	133,651	147,381	99,089	126,089	111,174	133,018	124,670	115,411
GC content	34.51	33.93	33.38	33.49	34.22	33.73	34.43	32.76	33.62	35.14	33.23	33.81	35.28
N content	3.90	3.37	3.20	3.09	3.69	3.90	1.23	6.76	3.34	4.73	3.47	3.30	3.42

**Supplementary Table 5.3:** Corrections made to RH BAC sequences using whole-genome Illumina data.

Change	Number
A > T substitution	1,223
C > G substitution	322
A > C substitution	505
A > G substitution	1,108
C > A substitution	734
G > A substitution	910
N substitution	230
A / T insertion	3,854
C / G insertion	569
A/T deletion	4,789
C/G deletion	2,763
N deletion	107
Total substitutions	5,032
Total insertions	4,423
Total deletions	7,659
Total changes	17,114

# Chapter 6

## Discussion

*Erwin Datema, Roeland C.H.J. van Ham*

DNA sequencing has evolved radically in the past two decades, starting with the automation of Sanger sequencing and culminating in the development of novel sequencing technologies that have raised the throughput by orders of magnitude while dramatically lowering the costs. These developments were driven by the aim to produce a complete human genome sequence for less than 1,000 US dollars and so to make genome sequencing an affordable tool in disease diagnostics. The large number of genome publications in recent years underlines that the application of these technologies is not limited to biomedical research alone. The genome sequences of a considerable number of food animals, crop plants and pathogenic fungi, bacteria and viruses have become available or are being generated at the moment, and many more can be expected in the near future. As of May 2011, the Genomes OnLine Database [255] lists more than 10,000 completed and on-going genome sequencing projects. The current thesis has shown the successful application of Sanger, Illumina and 454 sequencing to unravel the genomes of potato and tomato, and has documented the changes in sequencing technology and the associated developments in genome bioinformatics for complex plant genomes.

### **Outstanding challenges in plant genome sequencing**

The high repetitiveness of plant genomes hinders the assembly of large, megabase-sized sequence contigs from the short sequence reads produced by current NGS platforms. Matepair sequences generated from a collection of libraries that together cover a wide fragment length distribution have proven invaluable for generating large scaffolds that span substantial portions of chromosomes (Chapter 5). The contigs in these scaffolds represent the majority of unique sequences in the genome, whereas the gaps correspond mostly to unassembled repetitive elements. The fragment size from which matepair libraries can be generated ranges currently from two to approximately twenty kilobases, implying that the majority of large retrotransposable elements can be spanned effectively [256]. Nonetheless, retrotransposons are often found nested inside each other in plant genomes, resulting in large stretches of repetitive sequence [257] that result in loss of sequence connectivity. The combination of genome sequencing with physical maps, genetic maps and FISH encompasses a powerful tool for the proper placement and orientation of distinct sequence

scaffolds into partial or complete chromosomal pseudomolecules (Chapter 4). The physical evidence obtained through FISH can be exploited to anchor sequence scaffolds to their correct chromosomal location and to resolve chimerisms in the assembly. Moreover, it can visualize the organization of genomic regions that are not readily assembled due to their repetitive nature [173]. High-density sequence-based genetic maps can also be used to anchor and orient sequence scaffolds on their respective chromosomal locations; however, both technologies have their limitations. On the one hand, integration of FISH signals from multiple experiments is difficult due to variable stretching of chromatin between experiments [258]. On the other hand, the genome structure of the parents from which a genetic mapping population is constructed may differ from that of the sequenced genome. To resolve the discrepancies that may arise between the assembled sequence, cytogenetic information and genetic maps, a good understanding of the power and limitations of all these technologies is required (Chapter 4).

Another property of many crop plants that hinders the efficient assembly of their genomes is their high level of heterozygosity combined with their polyploidy. The sequence variation between alleles in a heterozygous polyploid requires a lower identity threshold for overlap-layout-consensus assembly, which conflicts with the accurate assembly of low-repetitive sequences and multi-copy gene families. In De Bruijn-graph assembly of such genomes, the assembly graph will contain many bubbles that can only be resolved by either extensive bubble pinching, resulting in a loss of heterozygosity, or by breaking the graph, resulting in a loss of sequence contiguity.

In order to reconstruct the two distinct haplotypes from the diploid RH potato genome, initially a clone-based sequencing strategy was combined with a linkage-phase specific physical map (Chapter 5). As a result of the introduction of novel sequencing technologies, the strategy was adapted to first obtain a reference genome sequence from a homozygous diploid accession of potato through whole-genome shotgun sequencing. This reference genome then served as a basis to study the genomic variation in the heterozygous genotype through re-sequencing (Chapter 5). This approach is however unable to yield large haplotype blocks from the individual linkage phases. Moreover, it may be infeasible to generate homozygous diploid accessions for every interesting polyploid crop species due to accumulation of lethal alleles. In these cases, the clone-based strategy combined with the current sequencing platforms is likely to be the only viable strategy in the near future. Physical mapping technologies such as Whole-Genome Profiling (WGP) [259] may accelerate such clone-based strategies. In WGP, the fingerprints of clones consist of short sequence tags rather than restriction patterns, allowing for a direct match between clone contigs and sequence scaffolds (Chapter 5). Another promising technology is optimal mapping [260], which has recently been combined with genome sequencing to generate a single molecule scaffolds for the highly repetitive maize genome [261, 262]. These mapping approaches are however costly and laborious. Computational tools that can resolve the problem of *de novo* heterozygous genome assembly from the sequence data

alone are starting to emerge [263], but will need to be adapted to the complexities that underlie plant genomes.

## **Opportunities for plant breeding and genome research**

The general availability of genome sequences for crop plant species is having a tremendous impact on the genetics and breeding of these organisms [264]. The release of the potato genome sequence (Chapter 5) and the imminent completion of the tomato genome (Chapter 4; [137]) will serve as a foundation for their research and breeding communities. Future comparative sequence analyses of the completed tomato and potato genome sequences will address many of the unresolved questions from Chapters 3, 4 and 5. It has now become technologically and financially feasible to study sequence variation on a genome-wide scale through re-sequencing of hundreds to thousands of individuals, both in naturally occurring accessions of a species as is currently done in the 1001 *Arabidopsis* genomes project [265] and in large breeding populations consisting of parents and offspring. The short reads produced from the high-throughput technologies can readily be mapped to the reference genome sequence to identify variation on both the nucleotide and the structural level. Extensive re-sequencing of wild and cultivated potato accessions will reveal whether the extensive sequence variation that was observed between the DM and RH potato genotypes (Chapter 5) is a typical feature of potato. The same question can soon be addressed for tomato, and moreover the self-pollinating nature of many domesticated tomato cultivars makes them excellent candidates for Genome-Wide Association Studies (GWAS), in which identified sequence variation can be linked to traits of interest through GWAS [266, 267]. This variation can subsequently be exploited to develop molecular markers tightly linked to these traits. In plant breeding, these markers can be employed to select lines with desirable genetic properties from large populations through Marker-Assisted Selection (MAS) [268]. These studies are becoming increasingly more powerful and are complemented by recent developments in high-throughput plant phenotyping hardware and software [269].

Other emerging applications of NGS combine high throughput sequencing with the detection of DNA regions bound to histones using Chromatin ImmunoPrecipitation Sequencing (ChiP-seq), or the methylated regions of a genome using Methylated DNA ImmunoPrecipitation (MeDIP-seq) in order to map the epigenome in single-base resolution [270]. In addition to producing genomic sequences, the advances in sequencing technology have also enabled deep sequencing of complete transcriptomes (RNA-seq) from large collections of tissues, conditions and time points to study the temporal and spatial distribution of gene activity [271]. Through these developments the focus of DNA sequencing has shifted from a primary data generation tool for the production of genome sequences to a functional genomics tools for both fundamental and applied research.

## **Beyond second-generation sequencing**

The throughput and read lengths of current sequencing technologies continue to increase, resulting in an ever decreasing cost per sequenced base. The initial Roche/454 GS20 instrument had a throughput of approximately 30 Mb per run with an average read length of 100 bp [39], which was upgraded through several improvements in hardware and chemistry to the current throughput and length of 400 Mb per run and 500 bp per read, respectively. Additional improvements to the sequencing chemistry are expected to extend the read length up to 800 bp in the near future, thereby approaching the read length of capillary Sanger sequencing. Whereas the advances in Roche/454 sequencing have been focused on the increase in read length to match that of Sanger sequencing, the developments in Illumina/Solexa and AB/SOLiD sequencing have resulted in a tremendous increase in throughput on both these platforms. For example, the Illumina GA apparatus was introduced with a throughput of approximately 1 Gb per run with reads measuring 35 bp each [272]. On the current HiSeq 2000 platform, the read length has increased to 100 bp and the throughput has become 400 Gb, and it is expected to soon break the Terabase barrier for a single instrument run. A similar development is observed for the AB/SOLiD technology.

In addition to the improvements made to the current sequencing platforms, novel technologies are being developed and implemented in third-generation sequencing platforms that sequence individual molecules without need for amplification. The PacBio RS machine from Pacific Biosciences [273] was commercially released in early 2011 and is based on single-molecule, real-time (SMRT) sequencing [274]. While this technology currently has a substantial error rate for single-pass reads, it produces reads of up to 1,000 bp at a throughput of 100,000 reads per instrument run [275]. The high error rate can be negated through application of the circular consensus sequencing protocol, in which a small template DNA molecule is read multiple times in both forward and reverse orientation. It furthermore distinguishes itself from second-generation technologies by offering strobe sequencing, in which short sequence tags are generated from a long template DNA molecule at known intervals. Looming on the horizon is nanopore sequencing, a novel technology that relies on the electrochemical properties of nucleotides for detection rather than light emission [276]. No commercial implementation is yet available but the principle of the technology has been successfully demonstrated by Oxford Nanopore Technologies [277]. A key feature of this technology is that it can distinguish methylated cytosine from unmethylated cytosine, allowing for direct detection of genome-wide methylation.

## **Bioinformatics for high throughput sequencing**

The novel properties of the second and third generation technologies create exciting opportunities for new sequencing applications, but also demand for robust software tools to

process the data produced through them efficiently. The raw image data captured by the second-generation platforms is often discarded after basecalling as it is orders of magnitudes larger than the text version of the sequence data. Nonetheless the vast amount of sequence data produced by these technologies has outpaced the capacity improvements in hard disk storage [278], which has spurred the development of tools that reduce the required storage space for unprocessed sequence data [279] and mapped sequence variation that is derived from these data [280]. A large variety in software tools currently exists for the assembly and mapping of sequence reads (Chapter 1) and as a result of the rapid improvements and changes in sequencing technologies there are few widely used, reproducible and automated analysis pipelines for these data. The Galaxy system [281] has become popular in automating read mapping to reference genomes and manipulation of the derived coordinate-based data. It features a web-based graphical interface and does not require computational resources from the user, but instead runs the required analyses on a dedicated server. It has recently been updated to also process the raw sequence data prior to mapping [282], but does not include data-intensive or computationally intensive tasks such as *de novo* genome assembly or read mapping due to its remote-execution nature. Existing local workflow management systems such as Cyrille2 [116] may be adapted to accelerate automation of these primary data analyses, but the mismatch in data volume between genome annotation on the one hand, and genome assembly and read mapping on the other hand may require more specialized systems.

Not only the tools for data processing, but also those for data visualization, interpretation and interrogation must be adapted to the flood of new sequence data that is being produced. Many tools for the visualization of genome assemblies, annotations and comparisons have been developed in order to aid researches in formulating hypotheses about their data [283]. These tools are often centered on data browsing, in which the user selects a certain region of the genome for closer inspection based on a pre-existing hypothesis. However, the amount of detail that can be visualized in a single image is limited and the same applies to the number of data elements that a user can visually process. The ever increasing amount of available genome sequences, annotations and sequence variation data create the need for an intuitive platform for the automated interrogation of these data in order to formulate new biologically relevant questions on datasets spanning hundreds or thousands of genome sequences.

## **Lessons learned from two Solanaceous genomes**

In this thesis, genome sequence analysis has been combined with genetics, cytogenetics and physical mapping in order to shed more light on the structure and organization of the tomato (Chapter 4) and potato (Chapter 5) genomes. FISH of Cot 100 DNA (the repetitive fraction of the genome) on the tomato genome has demonstrated that repeat elements occur most abundantly in the heterochromatin domains around the centromeres, in chromomeres,

and at the telomeres of chromosomes [173]. These regions are predominantly populated by retrotransposable elements of the *Gypsy* and *Copia* families, and their expansion after the divergence between tomato and potato may provide an explanation for the larger genome size of tomato compared to potato (Chapter 3). A similar mechanism of genome expansion appears to underlie the increased size of the pepper genome [284], where *Gypsy* elements have preferentially expanded in the euchromatin. Retrotransposons were also identified in the euchromatin of tomato chromosome 6, but there *Copia* elements were more abundant than *Gypsy* elements (Chapter 4). Remarkably, while the heterochromatin was generally believed to be devoid of protein-coding genes, both in tomato (Chapter 4) and potato (Chapter 5) heterochromatin a considerable number of genes was identified, the expression of which was supported by transcript sequences. The presence of these genes, which may very well be responsible for agriculturally interesting traits, in the heterochromatin regions of the genome combined with the suppressed recombination rate that is generally observed there presents novel challenges to tomato and potato breeding.

To overcome the difficulties associated with *de novo* assembly of a highly heterozygous diploid or polyploid potato genome, a unique doubled monoploid potato clone was sequenced. Comparisons with both short reads obtained from WGS and a large collection of linkage-phase specific BAC tiling paths revealed that the haplotype diversity within a heterozygous diploid potato is at least as large as, if not bigger than, the variation between different potato accessions. Within such diploids, PAVs and other potentially deleterious mutations such as SNPs and InDels occur frequently and are a probable cause of inbreeding depression. Sequence analysis of tetraploid potato cultivars in the future will further develop our understanding of the extent of this variation, and the molecular mechanisms that drive it. In contrast to potato, the nucleotide sequence variation within tomato, and in particular between domesticated *S. lycopersicum* and wild *S. pimpinellifolium*, is markedly lower. Preliminary comparisons of the draft genome sequences of these species have revealed only a minor variation in gene content and genome organization. Nonetheless the SNPs identified in such comparisons can aid plant breeders in generating new molecular markers to introduce novel, commercially valuable traits from the wild tomato into the domesticated variant. Taken together, the research presented in this thesis has contributed significantly to the understanding of the contents, structure and organization of the tomato and potato genomes, and the resulting genome sequences will be of great value to plant breeders and researchers in the years to come.

# Bibliography

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
2. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT *et al*: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**(7189):872-876.
3. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
4. Swami M: **Whole-genome sequencing identifies Mendelian mutations**. *Nat Rev Genet* 2010, **11**(5):313.
5. Eisenstein M: **Sequencing firms vie for diagnostics market, tiptoe round patents**. *Nat Biotechnol* 2010, **28**(7):635-636.
6. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
7. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS *et al*: **A whole-genome assembly of the domestic cow, *Bos taurus***. *Genome Biol* 2009, **10**(4):R42.
8. Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MA, Harlizius B, Lee KT, Milan D, Rogers J, Rothschild MF *et al*: **Pig genome sequence--analysis and publication strategy**. *BMC Genomics* 2010, **11**:438.
9. Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW, Wade C *et al*: **The sheep genome reference sequence: a work in progress**. *Anim Genet* 2010, **41**(5):449-453.
10. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)**. *Science* 2002, **296**(5565):92-100.
11. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**(5565):79-92.
12. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P *et al*: **The genome of the cucumber, *Cucumis sativus* L.** *Nat Genet* 2009, **41**(12):1275-1281.
13. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**(7161):463-467.
14. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
15. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL *et al*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.

16. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al*: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**(7278):178-183.
17. **FAOSTAT** [<http://faostat.fao.org/>]
18. van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P *et al*: **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* 2006, **173**(2):1075-1087.
19. Arumuganathan K, Earle E: **Estimation of nuclear DNA content of plants by flow cytometry.** *Plant Molecular Biology Reporter* 1991, **9**(3):229-241.
20. Bonierbale MW, Plaisted RL, Tanksley SD: **RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato.** *Genetics* 1988, **120**(4):1095-1103.
21. Iovene M, Wielgus SM, Simon PW, Buell CR, Jiang J: **Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato.** *Genetics* 2008, **180**(3):1307-1317.
22. Tang X, Szinay D, Lang C, Ramanna MS, van der Vossen EA, Datema E, Lankhorst RK, de Boer J, Peters SA, Bachem C *et al*: **Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements.** *Genetics* 2008, **180**(3):1319-1328.
23. Lou Q, Iovene M, Spooner DM, Buell CR, Jiang J: **Evolution of chromosome 6 of Solanum species revealed by comparative fluorescence in situ hybridization mapping.** *Chromosoma* 2010, **119**(4):435-442.
24. Szinay D, Bai Y, Visser R, de Jong H: **FISH applications for genomics and plant breeding strategies in tomato and other solanaceous crops.** *Cytogenet Genome Res* 2010, **129**(1-3):199-210.
25. Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J, Tanksley SD: **Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome.** *Genetics* 2006, **172**(4):2529-2540.
26. Soderlund C, Longden I, Mott R: **FPC: a system for building contigs from restriction fingerprinted clones.** *Comput Appl Biosci* 1997, **13**(5):523-535.
27. Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis.** *Genomics* 2003, **82**(3):378-389.
28. Szinay D, Chang SB, Khrustaleva L, Peters S, Schijlen E, Bai Y, Stiekema WJ, van Ham RC, de Jong H, Klein Lankhorst RM: **High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6.** *Plant J* 2008, **56**(4):627-637.
29. Ross MT, LaBrie S, McPherson J, Stanton VP: **Screening Large-Insert Libraries by Hybridization.** In: *Current Protocols in Human Genetics*. John Wiley & Sons, Inc.; 2001.
30. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94**(3):441-448.

31. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage [phi]X174 DNA.** *Nature* 1977, **265**(5596):687-695.
32. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB: **Nucleotide sequence of bacteriophage lambda DNA.** *J Mol Biol* 1982, **162**(4):729-773.
33. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
34. Metzker ML: **Emerging technologies in DNA sequencing.** *Genome Res* 2005, **15**(12):1767-1776.
35. Collins FS, Green ED, Guttmacher AE, Guyer MS: **A vision for the future of genomics research.** *Nature* 2003, **422**(6934):835-847.
36. Kircher M, Kelso J: **High-throughput DNA sequencing – concepts and limitations.** *Bioessays* 2010, **32**(6):524-536.
37. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11**(1):31-46.
38. **454 Life Sciences** [<http://454.com/>]
39. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**(7057):376-380.
40. **Illumina, Inc.** [<http://www.illumina.com/>]
41. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**(7218):53-59.
42. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**(16):e105.
43. Kircher M, Stenzel U, Kelso J: **Improved base calling for the Illumina Genome Analyzer using machine learning strategies.** *Genome Biol* 2009, **10**(8):R83.
44. **Applied Biosystems** [<http://www.appliedbiosystems.com/>]
45. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome.** *Science* 2005, **309**(5741):1728-1732.
46. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**(6):315-327.
47. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing.** *Genome Res* 2010, **20**(9):1165-1173.
48. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**(5223):496-512.
49. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: A Whole-Genome Shotgun Assembler.** *Genome Res* 2002, **12**(1):177-189.
50. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24**(24):2818-2824.

51. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: **ABYSS: A parallel assembler for short read sequence data.** *Genome Res* 2009, **19**(6):1117-1123.
52. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K *et al*: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20**(2):265-272.
53. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821-829.
54. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proceedings of the National Academy of Sciences* 2001, **98**(17):9748-9753.
55. Koren S, Miller J, Walenz B, Sutton G: **An algorithm for automated closure during assembly.** *BMC Bioinformatics* 2010, **11**(1):457.
56. Tsai I, Otto T, Berriman M: **Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps.** *Genome Biol* 2010, **11**(4):R41.
57. Bonfield JK, Staden R: **The application of numerical estimates of base calling accuracy to DNA sequencing projects.** *Nucleic Acids Res* 1995, **23**(8):1406-1410.
58. Tammi MT, Arner E, Kindlund E, Andersson B: **Correcting errors in shotgun sequences.** *Nucleic Acids Res* 2003, **31**(15):4663-4672.
59. Gajer P, Schatz M, Salzberg SL: **Automated correction of genome sequence errors.** *Nucleic Acids Res* 2004, **32**(2):562-569.
60. Chaisson M, Pevzner P, Tang H: **Fragment assembly with short reads.** *Bioinformatics* 2004, **20**(13):2067-2074.
61. Zhao X, Palmer LE, Bolanos R, Mircean C, Fasulo D, Wittenberg GM: **EDAR: an efficient error detection and removal algorithm for next generation sequencing data.** *J Comput Biol* 2010, **17**(11):1549-1560.
62. Kelley D, Schatz M, Salzberg S: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biol* 2010, **11**(11):R116.
63. Yang X, Dorman KS, Aluru S: **Reptile: representative tiling for short read error correction.** *Bioinformatics* 2010, **26**(20):2526-2533.
64. Ilie L, Fazayeli F, Ilie S: **HiTEC: accurate error correction in high-throughput sequencing data.** *Bioinformatics* 2011, **27**(3):295-302.
65. Choi J-H, Kim S, Tang H, Andrews J, Gilbert DG, Colbourne JK: **A machine-learning approach to combined evidence validation of genome assemblies.** *Bioinformatics* 2008, **24**(6):744-750.
66. Phillippy A, Schatz M, Pop M: **Genome assembly forensics: finding the elusive mis-assembly.** *Genome Biol* 2008, **9**(3):R55.
67. Zimin AV, Smith DR, Sutton G, Yorke JA: **Assembly reconciliation.** *Bioinformatics* 2008, **24**(1):42-45.
68. Nijkamp J, Winterbach W, van den Broek M, Daran J-M, Reinders M, de Ridder D: **Integrating genome assemblies with MAIA.** *Bioinformatics* 2010, **26**(18):i433-i439.
69. Meader S, Hillier LW, Locke D, Ponting CP, Lunter G: **Genome assembly quality: Assessment and improvement using the neutral indel model.** *Genome Res* 2010, **20**(5):675-684.
70. Do JH, Choi DK: **Computational approaches to gene prediction.** *J Microbiol* 2006, **44**(2):137-144.

71. Picardi E, Pesole G: **Computational methods for ab initio and comparative gene finding.** *Methods Mol Biol* 2010, **609**:269-284.
72. Fickett JW, Tung CS: **Assessment of protein coding measures.** *Nucleic Acids Res* 1992, **20**(24):6441-6450.
73. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**(6):276-277.
74. Guigo R: **Assembling genes from predicted exons in linear time with dynamic programming.** *J Comput Biol* 1998, **5**(4):681-702.
75. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**(1):78-94.
76. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20**(16):2878-2879.
77. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A *et al*: **mGene: accurate SVM-based gene finding with an application to nematode genomes.** *Genome Res* 2009, **19**(11):2133-2143.
78. Majoros W, Salzberg S: **An empirical analysis of training protocols for probabilistic gene finders.** *BMC Bioinformatics* 2004, **5**(1):206.
79. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**(1):59.
80. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucleic Acids Res* 2005, **33**(20):6494-6506.
81. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24**(5):637-644.
82. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms.** *Information and Software Technology* 2005, **47**(15):965-978.
83. Gascuel O, Sagot M-F, Schiex T, Moisan A, Rouzé P: **Eugène: An Eukaryotic Gene Finder That Combines Several Sources of Evidence.** In: *Computational Biology*. vol. 2066: Springer Berlin / Heidelberg; 2001: 111-125.
84. Allen JE, Salzberg SL: **JIGSAW: integration of multiple sources of evidence for gene prediction.** *Bioinformatics* 2005, **21**(18):3596-3603.
85. Allen J, Majoros W, Pertea M, Salzberg S: **JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions.** *Genome Biol* 2006, **7**(Suppl 1):S9.
86. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2**(12):919-929.
87. Wu Y, Wei B, Liu H, Li T, Rayner S: **MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences.** *BMC Bioinformatics* 2011, **12**:107.
88. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAMmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35**(9):3100-3108.
89. Lowe TM, Eddy SR: **A Computational Screen for Methylation Guide snoRNAs in Yeast.** *Science* 1999, **283**(5405):1168-1171.
90. Lowe TM, Eddy SR: **tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.** *Nucleic Acids Res* 1997, **25**(5):0955-0964.

91. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009.
92. Eichler EE, Sankoff D: **Structural Dynamics of Eukaryotic Chromosome Evolution.** *Science* 2003, **301**(5634):793-797.
93. Vitte C, Panaud O: **LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model.** *Cytogenet Genome Res* 2005, **110**(1-4):91-107.
94. **RepeatMasker** [<http://www.repeatmasker.org/>]
95. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Rebase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**:462-467.
96. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.** *BMC Bioinformatics* 2008, **9**(1):18.
97. Rho M, Choi J-H, Kim S, Lynch M, Tang H: **De novo identification of LTR retrotransposons in eukaryotic genomes.** *BMC Genomics* 2007, **8**(1):90.
98. Bao Z, Eddy SR: **Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes.** *Genome Res* 2002, **12**(8):1269-1276.
99. Saha S, Bridges S, Magbanua ZV, Peterson DG: **Empirical comparison of ab initio repeat finding programs.** *Nucleic Acids Res* 2008, **36**(7):2284-2294.
100. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573-580.
101. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G: **Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes.** *Genome Res* 2004, **14**(10A):1861-1869.
102. Kofler R, Schlötterer C, Lelley T: **SciRoKo: a new tool for whole genome microsatellite search and investigation.** *Bioinformatics* 2007, **23**(13):1683-1685.
103. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
104. Chen R, Jeong SS: **Functional prediction: identification of protein orthologs and paralogs.** *Protein Sci* 2000, **9**(12):2344-2353.
105. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2011, **39**(Database issue):D32-37.
106. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database* 2011, **2011**.
107. Mulder NJ, Apweiler R: **The InterPro database and tools for protein domain analysis.** *Curr Protoc Bioinformatics* 2008, **Chapter 2**:Unit 2 7.
108. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.
109. Gene Ontology Consortium: **The Gene Ontology in 2010: extensions and refinements.** *Nucleic Acids Res* 2010, **38**(Database issue):D331-335.
110. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**(2):R12.

111. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**(1):103-107.
112. Trapnell C, Salzberg SL: **How to map billions of short reads onto genomes**. *Nat Biotechnol* 2009, **27**(5):455-457.
113. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
114. Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
115. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads**. *Bioinformatics* 2009, **25**(21):2865-2871.
116. Fiers MW, van der Burgt A, Datema E, de Groot JC, van Ham RC: **High-throughput bioinformatics with the Cyrille2 pipeline system**. *BMC Bioinformatics* 2008, **9**:96.
117. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database**. *Genome Res* 2002, **12**(10):1599-1610.
118. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
119. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**(5):988-995.
120. Berman P, Zhang Z, Wolf YI, Koonin EV, Miller W: **Winnowing sequences from a database search**. *J Comput Biol* 2000, **7**(1-2):293-302.
121. **BioPerl** [<http://bioperl.org/>]
122. Zhang J, Madden TL: **PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation**. *Genome Res* 1997, **7**(6):649-656.
123. Cantalloube H, Chomilier J, Chiusa S, Lonquety M, Spadoni JL, Zagury JF: **Filtering redundancies for sequence similarity search programs**. *J Biomol Struct Dyn* 2005, **22**(4):487-492.
124. **Python Programming Language** [<http://www.python.org/>]
125. **Numerical Python** [<http://sourceforge.net/projects/numpy/>]
126. **Biopython** [<http://www.biopython.org/>]
127. **The GNU General Public License** [<http://www.gnu.org/copyleft/gpl.html>]
128. Tanksley SD, Ganai MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB *et al*: **High density molecular linkage maps of the tomato and potato genomes**. *Genetics* 1992, **132**(4):1141-1160.
129. Yano K, Watanabe M, Yamamoto N, Tsugane T, Aoki K, Sakurai N, Shibata D: **MiBASE: A database of a miniature tomato cultivar Micro-Tom**. *Plant Biotechnology* 2006, **23**:195-198.
130. D'Agostino N, Aversano M, Frusciante L, Chiusano ML: **TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species**. *Nucleic Acids Res* 2007, **35**(Database issue):D901-D905.
131. **Wageningen UR Plant Breeding CBSG Potato & Tomato Genomics Database** [<http://potatodbase.dpw.wau.nl/>]

132. **PotatEST DB** [<http://biosrv.cab.unina.it/potatestdb/>]
133. Mueller LA, Tanksley SD, Giovannoni JJ, Van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C *et al*: **The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL)**. *Comparative and Functional Genomics* 2005, **6**(3):153-158.
134. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y *et al*: **The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond**. *Plant Physiol* 2005, **138**(3):1310-1317.
135. **Potato Genome Sequencing Consortium** [<http://www.potatogenome.net/>]
136. Budiman MA, Mao L, Wood TC, Wing RA: **A Deep-Coverage Tomato BAC Library and Prospects Toward Development of an STC Framework for Genome Sequencing**. *Genome Res* 2000, **10**(1):129-136.
137. **SOL Genomics Network** [<http://solgenomics.net/>]
138. Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganai MW, Spivey R, Wu T, Earle ED, Tanksley SD: **Map-based cloning of a protein kinase gene conferring disease resistance in tomato**. *Science* 1993, **262**(5138):1432-1436.
139. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF *et al*: **Sequence composition and genome organization of maize**. *Proc Natl Acad Sci U S A* 2004, **101**(40):14349-14354.
140. Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I *et al*: **A Survey of the Brassica rapa genome by BAC-end sequence analysis and comparison with Arabidopsis thaliana**. *Mol Cells* 2006, **22**(3):300-307.
141. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
142. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.
143. Lai CW, Yu Q, Hou S, Skelton RL, Jones MR, Lewis KL, Murray J, Eustice M, Guan P, Agbayani R *et al*: **Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome**. *Mol Genet Genomics* 2006, **276**(1):1-12.
144. Katti MV, Ranjekar PK, Gupta VS: **Differential distribution of simple sequence repeats in eukaryotic genome sequences**. *Mol Biol Evol* 2001, **18**(7):1161-1167.
145. Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA: **The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean**. *Theoretical and Applied Genetics* 2007, **114**(6):1081-1090.
146. Cheung F, Town CD: **A BAC end view of the Musa acuminata genome**. *BMC Plant Biol* 2007, **7**(29).
147. Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J, Denny R, Endré G, Saurat O, Duzde AM *et al*: **Distribution of microsatellites in the genome of Medicago truncatula: a resource of genetic markers that integrate genetic and physical maps**. *Genetics* 2006, **172**(4):2541-2555.
148. Areshchenkova T, Ganai MW: **Long tomato microsatellites are predominantly associated with centromeric regions**. *Genome* 1999, **42**(3):536-544.
149. van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S: **Deductions about the number, organization, and evolution of genes in the tomato genome**

- based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *The Plant Cell* 2002, **14**(7):1441-1456.
150. TAIR [<http://www.arabidopsis.org/>]
  151. Schuler MA, Werck-Reichhart D: **Functional genomics of P450s**. *Annu Rev Plant Biol* 2003, **54**:629-667.
  152. Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S: **Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot**. *Plant Physiol* 2004, **135**(2):756-772.
  153. Xu Y, Ishida H, Reisen D, Hanson MR: **Upregulation of a tonoplast-localized cytochrome P450 during petal senescence in *Petunia inflata***. *BMC Plant Biol* 2006, **6**(8).
  154. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**(6930):433-438.
  155. Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R: **Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, Arabidopsis, and *Capsella* genomes**. *The Plant Cell* 2001, **13**(4):979-988.
  156. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny**. *Proc Natl Acad Sci U S A* 2000, **97**(16):9121-9126.
  157. Ballvora A, Jöcker A, Viehöver P, Ishihara H, Paal J, Meksem K, Bruggmann R, Schoof H, Weisshaar B, Gebhardt C: **Comparative sequence analysis of *Solanum* and Arabidopsis in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments**. *BMC Genomics* 2007, **8**(112).
  158. Gebhardt C, Walkemeier B, Henselewski H, Barakat A, Delseny M, Stüber K: **Comparative mapping between potato (*Solanum tuberosum*) and Arabidopsis thaliana reveals structurally conserved domains and ancient duplications in the potato genome**. *The Plant Journal* 2003, **34**(4):529-541.
  159. NCBI dbGSS [<http://www.ncbi.nlm.nih.gov/dbGSS/>]
  160. Daniell H, Lee SB, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK: **Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes**. *Theoretical and Applied Genetics* 2006, **112**(8):1503-1518.
  161. Green Group [<http://www.phrap.org/>]
  162. **EST-SSRs From Wheat, Barley And Rice** [<http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/>]
  163. GenBank [<http://www.ncbi.nlm.nih.gov/Genbank/>]
  164. Micro-Tom database [<http://www.kazusa.or.jp/jsol/microtom/index.html>]
  165. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R *et al*: **Pfam: clans, web tools and services**. *Nucleic Acids Res* 2006, **34**(Database issue):D247-D251.
  166. Mi H, Guo N, Kejariwal A, Thomas PD: **PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways**. *Nucleic Acids Res* 2007, **35**(Database issue):D247-D252.

167. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R *et al*: **New developments in the InterPro database.** *Nucleic Acids Res* 2007, **35**(Database issue):D224-D228.
168. **The R Project For Statistical Computing** [<http://www.r-project.org/>]
169. **Joint Genome Institute** [[ftp://ftp.jgi-psf.org/pub/JGI\\_data/Poplar/](ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/)]
170. **International Solanaceae Project** [<http://solgenomics.net/solanaceae-project/index.pl>]
171. Ganal MW, Lapitan NL, Tanksley SD: **Macrostructure of the tomato telomeres.** *Plant Cell* 1991, **3**(1):87-94.
172. De Jong JH, Zhong XB, Fransz PF, Wennekes-van Eden J, Jacobsen E, Zabel P: **High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences of individual tomato chromosomes.** In: *Chromosomes today*. Edited by Olmo E, Redi CA: Birkhauser Verlag; 2000: 267 - 275.
173. Chang SB, Yang TJ, Datema E, van Vugt J, Vosman B, Kuipers A, Meznikova M, Szinay D, Lankhorst RK, Jacobsen E *et al*: **FISH mapping and molecular organization of the major repetitive sequences of tomato.** *Chromosome Res* 2008, **16**(7):919-933.
174. Khush GS, Rick CM, Robinson RW: **Genetic Activity in a Heterochromatic Chromosome Segment of the Tomato.** *Science* 1964, **145**(3639):1432-1434.
175. Peterson DG, Stack SM, Price HJ, Johnston JS: **DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes.** *Genome* 1996, **39**(1):77-82.
176. Peterson DG, Pearson WR, Stack SM: **Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation.** *Genome* 1998, **41**(3):346-356.
177. **International Tomato Genome Sequence Project** [[http://solgenomics.net/organism/solanum\\_lycopersicum/genome](http://solgenomics.net/organism/solanum_lycopersicum/genome)]
178. **Tomato FPC map** [<http://www.genome.arizona.edu/fpc/tomato/>]
179. Meyers BC, Scalabrin S, Morgante M: **Mapping and sequencing complex genomes: let's get physical!** *Nat Rev Genet* 2004, **5**(8):578-588.
180. Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD: **Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants.** *Plant Cell* 2002, **14**(7):1457-1467.
181. Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG *et al*: **Genetic, physical, and informatics resources for maize. On the road to an integrated map.** *Plant Physiol* 2002, **130**(4):1598-1605.
182. Venter JC, Smith HO, Hood L: **A new strategy for genome sequencing.** *Nature* 1996, **381**(6581):364-366.
183. Batzoglou S, Berger B, Mesirov J, Lander ES: **Sequencing a genome by walking with clone-end sequences: a mathematical analysis.** *Genome Res* 1999, **9**(12):1163-1174.
184. Peters SA, van Haarst JC, Jesse TP, Woltinge D, Jansen K, Hesselink T, van Staveren MJ, Abma-Henkens MH, Klein-Lankhorst RM: **TOPAAS, a tomato and potato assembly assistance system for selection and finishing of bacterial artificial chromosomes.** *Plant Physiol* 2006, **140**(3):805-817.
185. **The overgo plating process** [[http://solgenomics.net/maps/physical/overgo\\_process\\_explained.pl](http://solgenomics.net/maps/physical/overgo_process_explained.pl)]
186. **SGN marker search** [<http://solgenomics.net/search/markers>]

187. Koo DH, Jo SH, Bang JW, Park HM, Lee S, Choi D: **Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2.** *Genetics* 2008, **179**(3):1211-1220.
188. Macas J, Neumann P: **Ogre elements--a distinct group of plant Ty3/gypsy-like retrotransposons.** *Gene* 2007, **390**(1-2):108-116.
189. Rubin E, Lithwick G, Levy AA: **Structure and evolution of the hAT transposon superfamily.** *Genetics* 2001, **158**(3):949-957.
190. Bennetzen JL: **The Mutator transposable element system of maize.** *Curr Top Microbiol Immunol* 1996, **204**:195-229.
191. Gierl A: **The En/Spm transposable element of maize.** *Curr Top Microbiol Immunol* 1996, **204**:145-159.
192. Yang TJ, Lee S, Chang SB, Yu Y, de Jong H, Wing RA: **In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotransposons.** *Chromosoma* 2005, **114**(2):103-117.
193. Zhong XB, Fransz PF, Wennekes-Eden J, Ramanna MS, van Kammen A, Zabel P, Hans de Jong J: **FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato.** *Plant J* 1998, **13**(4):507-517.
194. Friesen N, Brandes A, Heslop-Harrison JS: **Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers.** *Mol Biol Evol* 2001, **18**(7):1176-1188.
195. Pereira V: **Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome.** *Genome Biol* 2004, **5**(10):R79.
196. Natali L, Santini S, Giordani T, Minelli S, Maestrini P, Cionini PG, Cavallini A: **Distribution of Ty3-gypsy- and Ty1-copia-like DNA sequences in the genus Helianthus and other Asteraceae.** *Genome* 2006, **49**(1):64-72.
197. Fregonezi JN, Vilas-Boas LA, Fungaro MHP, Gaeta ML, Vanzela ALL: **Distribution of a Ty3/gypsy-like retroelement on the a and B-chromosomes of Cestrum strigilatum Ruiz & Pav. and Cestrum intermedium sendtn. (Solanaceae).** *Genet Mol Biol* 2007, **30**(3):599-604.
198. Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL: **A GeneTrek analysis of the maize genome.** *Proc Natl Acad Sci U S A* 2007, **104**(28):11844-11849.
199. Tam SM, Causse M, Garchery C, Burck H, Mhiri C, Grandbastien MA: **The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species.** *J Evol Biol* 2007, **20**(3):1056-1072.
200. Kazazian HH, Jr.: **Mobile elements: drivers of genome evolution.** *Science* 2004, **303**(5664):1626-1632.
201. Field B, Osbourn AE: **Metabolic diversification--independent assembly of operon-like gene clusters in different plants.** *Science* 2008, **320**(5875):543-547.
202. Gierl A, Frey M: **Evolution of benzoxazinone biosynthesis and indole production in maize.** *Planta* 2001, **213**(4):493-498.
203. Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A: **A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants.** *Proc Natl Acad Sci U S A* 2004, **101**(21):8233-8238.
204. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5**(4):299-310.

205. Yi G, Sze SH, Thon MR: **Identifying clusters of functionally related genes in genomes.** *Bioinformatics* 2007, **23**(9):1053-1060.
206. Seah S, Yaghoobi J, Rossi M, Gleason CA, Williamson VM: **The nematode-resistance gene, Mi-1, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato.** *Theor Appl Genet* 2004, **108**(8):1635-1642.
207. Yasuhara JC, Wakimoto BT: **Oxymoron no more: the expanding world of heterochromatic genes.** *Trends Genet* 2006, **22**(6):330-338.
208. Weiler KS, Wakimoto BT: **Heterochromatin and gene expression in Drosophila.** *Annu Rev Genet* 1995, **29**:577-605.
209. Guyot R, Cheng X, Su Y, Cheng Z, Schlagenhauf E, Keller B, Ling HQ: **Complex organization and evolution of the tomato pericentromeric region at the FER gene locus.** *Plant Physiol* 2005, **138**(3):1205-1215.
210. Bai Y, van der Hulst R, Huang CC, Wei L, Stam P, Lindhout P: **Mapping OI-4, a gene conferring resistance to Oidium neolycopersici and originating from Lycopersicon peruvianum LA2172, requires multi-allelic, single-locus markers.** *Theor Appl Genet* 2004, **109**(6):1215-1223.
211. Ji Y, Schuster D, Scott J: **Ty-3, a begomovirus resistance locus near the Tomato yellow leaf curl virus resistance locus Ty-1 on chromosome 6 of tomato.** *Molecular Breeding* 2007, **20**(3):271-284.
212. Manetti ME, Rossi M, Costa AP, Clausen AM, Van Sluys MA: **Radiation of the Tnt1 retrotransposon superfamily in three Solanaceae genera.** *BMC Evol Biol* 2007, **7**:34.
213. Zhong X-b, Fransz P, Wennekes-van Eden J, Zabel P, van Kammen A, Hans de Jong J: **High-resolution mapping on pachytene chromosomes and extended DNA fibres by fluorescence in-situ hybridisation.** *Plant Molecular Biology Reporter* 1996, **14**(3):232-242.
214. Budiman MA, Chang SB, Lee S, Yang TJ, Zhang HB, de Jong H, Wing RA: **Localization of jointless-2 gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping.** *Theor Appl Genet* 2004, **108**(2):190-196.
215. Zhong XB, Hans de Jong J, Zabel P: **Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence in situ hybridization (FISH).** *Chromosome Res* 1996, **4**(1):24-28.
216. You FM, Luo MC, Gu YQ, Lazo GR, Deal K, Dvorak J, Anderson OD: **GenoProfiler: batch processing of high-throughput capillary fingerprinting data.** *Bioinformatics* 2007, **23**(2):240-242.
217. Soderlund C, Humphray S, Dunham A, French L: **Contigs built with fingerprints, markers, and FPC V4.7.** *Genome Res* 2000, **10**(11):1772-1787.
218. **J. Craig Venter Institute** [<http://www.jcvi.org/>]
219. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19 Suppl 2**:ii215-225.
220. Parra G, Blanco E, Guigo R: **GeneID in Drosophila.** *Genome Res* 2000, **10**(4):511-515.
221. **PotatEST database** [<http://biosrv.cab.unina.it/potatestdb/>]
222. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8**(9):967-974.

223. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot**. *Methods Mol Biol* 2007, **406**:89-112.
224. Hijmans RJ: **Global distribution of the potato crop**. *Am J Potato Res* 2001, **78**:403-412.
225. Burlingame B, Mouill B, Charrondi R: **Nutrients, bioactive non-nutrients and anti-nutrients in potatoes**. *J Food Compos Anal* 2009, **22**:494-502.
226. Paz MM, Veilleux RE: **Influence of culture medium and in vitro conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids**. *Plant Breeding* 1999, **118**(1):53-57.
227. Arumuganathan K, Earle E: **Nuclear DNA content of some important plant species**. *Plant Mol Biol Rep* 1991, **9**:208-218.
228. Tang X, de Boer JM, van Eck HJ, Bachem C, Visser RGf, de Jong H: **Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology**. *Chromosome Res* 2009, **17**(7):899-915.
229. Albach DC, Soltis, P. S., Soltis, D. E. : **Patterns of embryological and biochemical evolution in the Asterids**. *Syst Bot* 2001, **26**:242-262.
230. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps**. *Genome Res* 2008, **18**(12):1944-1954.
231. Fawcett JA, Maere S, Van de Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event**. *Proc Natl Acad Sci U S A* 2009, **106**(14):5737-5742.
232. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.
233. Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M *et al*: **Genome-wide patterns of genetic variation among elite maize inbred lines**. *Nat Genet* 2010, **42**(11):1027-1030.
234. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
235. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J *et al*: **A first-generation haplotype map of maize**. *Science* 2009, **326**(5956):1115-1117.
236. Prat S, Frommer WB, Hofgen R, Keil M, Kossmann J, Koster-Topfer M, Liu XJ, Muller B, Pena-Cortes H, Rocha-Sosa M *et al*: **Gene expression during tuber development in potato plants**. *FEBS Lett* 1990, **268**(2):334-338.
237. Glaczinski H, Heibges A, Salamini R, Gebhardt C: **Members of the Kunitz-type protease inhibitor gene family of potato inhibit soluble tuber invertase in vitro**. *Potato Res* 2002, **45**(163-176).
238. Shannon JC, Pien FM, Liu KC: **Nucleotides and Nucleotide Sugars in Developing Maize Endosperms (Synthesis of ADP-Glucose in brittle-1)**. *Plant Physiol* 1996, **110**(3):835-843.
239. Tauberger E, Fernie AR, Emmermann M, Renz A, Kossmann J, Willmitzer L, Trethewey RN: **Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate**. *Plant J* 2000, **23**(1):43-53.
240. Fettke J, Albrecht, T., Hejazi, M., Mahlow, S., Nakamura, Y., Steup, M.: **Glucose 1-phosphate is efficiently taken up by potato (*Solanum tuberosum*) tuber**

- parenchyma cells and converted to reserve starch granules. *New Phytol* 2010, **185**:663-675.
241. Sonnewald U: **Control of potato tuber sprouting.** *Trends Plant Sci* 2001, **6**(8):333-335.
242. Yoo SK, Chung KS, Kim J, Lee JH, Hong SM, Yoo SJ, Yoo SY, Lee JS, Ahn JH: **CONSTANS activates SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 through FLOWERING LOCUS T to promote flowering in Arabidopsis.** *Plant Physiol* 2005, **139**(2):770-778.
243. Kohler A, Rinaldi C, Duplessis S, Bauchet M, Geelen D, Duchaussoy F, Meyers BC, Boerjan W, Martin F: **Genome-wide identification of NBS resistance genes in *Populus trichocarpa*.** *Plant Mol Biol* 2008, **66**:619-636.
244. Kuang H, Wei F, Marano MR, Wirtz U, Wang X, Liu J, Shum WP, Zaborsky J, Tallon LJ, Rensink W *et al*: **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 J* 2005, **44**(1):37-51.
245. Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW: **Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce.** *Plant Cell* 2004, **16**(11):2870-2894.
246. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T *et al*: **Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*.** *Nature* 2009, **461**(7262):393-398.
247. Haynes FL: **The use of cultivated diploid *Solanum* species in potato breeding.** In: *Prospects for the potato in the developing world: an international symposium on key problems and potentials for greater use of the potato in the developing world: 1972; Lima, Peru*: International Potato Center (CIP); 1972: 100-110.
248. Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2004, **Chapter 4**:Unit 4 10.
249. Jiang Z, Hubley R, Smit A, Eichler EE: **DupMasker: a tool for annotating primate segmental duplications.** *Genome Res* 2008, **18**(8):1362-1368.
250. Elsieck CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set.** *Genome Biol* 2007, **8**(1):R13.
251. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
252. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**(5):511-515.
253. Tang X, de Boer JM, van Eck HJ, Bachem C, Visser RG, de Jong H: **Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology.** *Chromosome Res* 2009, **17**(7):899-915.
254. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25**(15):1966-1967.
255. **Genomes OnLine Database** [<http://www.genomesonline.org/>]

256. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O *et al*: **A unified classification system for eukaryotic transposable elements**. *Nat Rev Genet* 2007, **8**(12):973-982.
257. Kronmiller BA, Wise RP: **TENest: automated chronological annotation and visualization of nested plant transposable elements**. *Plant Physiol* 2008, **146**(1):45-59.
258. Ekong R, Wolfe J: **Advances in fluorescent in situ hybridisation**. *Curr Opin Biotechnol* 1998, **9**(1):19-24.
259. van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R *et al*: **Sequence-based physical mapping of complex genomes by whole genome profiling**. *Genome Res* 2011, **21**(4):618-625.
260. Cai W, Jing J, Irvin B, Ohler L, Rose E, Shizuya H, Kim U-J, Simon M, Anantharaman T, Mishra B *et al*: **High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping**. *Proceedings of the National Academy of Sciences* 1998, **95**(7):3390-3395.
261. Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W *et al*: **The physical and genetic framework of the maize B73 genome**. *PLoS Genet* 2009, **5**(11):e1000715.
262. Zhou S, Wei F, Nguyen J, Bechner M, Potamouis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S *et al*: **A single molecule scaffold for the maize genome**. *PLoS Genet* 2009, **5**(11):e1000711.
263. Kang SH, Jeong IS, Cho HG, Lim HS: **HapAssembler: a web server for haplotype assembly from SNP fragments using genetic algorithm**. *Biochem Biophys Res Commun* 2010, **397**(2):340-344.
264. Varshney RK, Nayak SN, May GD, Jackson SA: **Next-generation sequencing technologies and their implications for crop genetics and breeding**. *Trends Biotechnol* 2009, **27**(9):522-530.
265. **1001 genomes: A catalog of Arabidopsis thaliana genetic variation** [<http://www.1001genomes.org/>]
266. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT *et al*: **Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines**. *Nature* 2010, **465**(7298):627-631.
267. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z *et al*: **Genome-wide association studies of 14 agronomic traits in rice landraces**. *Nat Genet* 2010, **42**(11):961-967.
268. Varshney RK, Graner A, Sorrells ME: **Genomics-assisted breeding for crop improvement**. *Trends Plant Sci* 2005, **10**(12):621-630.
269. Vankadavath RN, Hussain AJ, Bodanapu R, Kharshiing E, Basha PO, Gupta S, Sreelakshmi Y, Sharma R: **Computer aided data acquisition tool for high-throughput phenotyping of plant populations**. *Plant Methods* 2009, **5**:18.
270. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S: **Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription**. *Nat Genet* 2007, **39**(1):61-69.
271. Wang L, Li P, Brutnell TP: **Exploring plant transcriptomes using ultra high-throughput sequencing**. *Briefings in Functional Genomics* 2010, **9**(2):118-128.

272. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics.** *Genomics* 2008, **92**(5):255-264.
273. Pacific Biosciences [<http://www.pacificbiosciences.com/>]
274. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules.** *Methods Enzymol* 2010, **472**:431-455.
275. Thompson J, Milos P: **The properties and applications of single-molecule DNA sequencing.** *Genome Biol* 2011, **12**(2):217.
276. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al*: **The potential and challenges of nanopore sequencing.** *Nat Biotechnol* 2008, **26**(10):1146-1153.
277. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing.** *Nature Nanotechnology* 2009, **4**(4):265-270.
278. Kahn SD: **On the future of genomic data.** *Science* 2011, **331**(6018):728-729.
279. Deorowicz S, Grabowski S: **Compression of DNA sequence reads in FASTQ format.** *Bioinformatics* 2011, **27**(6):860-862.
280. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression.** *Genome Res* 2011, **21**(5):734-740.
281. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol* 2010, **Chapter 19**:Unit 19 10 11-21.
282. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A: **Manipulation of FASTQ data with Galaxy.** *Bioinformatics* 2010, **26**(14):1783-1785.
283. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T: **Visualizing genomes: techniques and challenges.** *Nat Methods* 2010, **7**(3 Suppl):S5-S15.
284. Park M, Jo S, Kwon JK, Park J, Ahn JH, Kim S, Lee YH, Yang TJ, Hur CG, Kang BC *et al*: **Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements.** *BMC Genomics* 2011, **12**:85.

# Summary

In the past two decades genome sequencing has developed from a laborious and costly technology employed by large international consortia to a widely used, automated and affordable tool used worldwide by many individual research groups. Genome sequences of many food animals and crop plants have been deciphered and are being exploited for fundamental research and applied to improve their breeding programs. The developments in sequencing technologies have also impacted the associated bioinformatics strategies and tools, both those that are required for data processing, management, and quality control, and those used for interpretation of the data.

This thesis focuses on the application of genome sequencing, assembly and annotation to two members of the *Solanaceae* family, tomato and potato. Potato is the economically most important species within the *Solanaceae*, and its tubers contribute to dietary intake of starch, protein, antioxidants, and vitamins. Tomato fruits are the second most consumed vegetable after potato, and are a globally important dietary source of lycopene, beta-carotene, vitamin C, and fiber. The chapters in this thesis document the generation, exploitation and interpretation of genomic sequence resources for these two species and shed light on the contents, structure and evolution of their genomes.

**Chapter 1** introduces the concepts of genome sequencing, assembly and annotation, and explains the novel genome sequencing technologies that have been developed in the past decade. These so-called Next Generation Sequencing platforms display considerable variation in chemistry and workflow, and as a consequence the throughput and data quality differs by orders of magnitude between the platforms. The currently available sequencing platforms produce a vast variety of read lengths and facilitate the generation of paired sequences with an approximately fixed distance between them. The choice of sequencing chemistry and platform combined with the type of sequencing template demands specifically adapted bioinformatics for data processing and interpretation. Irrespective of the sequencing and assembly strategy that is chosen, the resulting genome sequence, often represented by a collection of long linear strings of nucleotides, is of limited interest by itself. Interpretation of the genome can only be achieved through sequence annotation – that is, identification and classification of all functional elements in a genome sequence. Once these elements have been annotated, sequence alignments between multiple genomes of related accessions or species can be utilized to reveal the genetic variation on both the nucleotide and the structural level that underlies the difference between these species or accessions.

**Chapter 2** describes BlastIf, a novel software tool that exploits sequence similarity searches with BLAST to provide a straightforward annotation of long nucleotide sequences. Generally, two problems are associated with the alignment of a long nucleotide sequence to a database of short gene or protein sequences: (i) the large number of similar

hits that can be generated due to database redundancy; and (ii) the relationships implied between aligned segments within a hit that in fact correspond to distinct elements on the sequence such as genes. BlastIf generates a comprehensible BLAST output for long nucleotide sequences by reducing the number of similar hits while revealing most of the variation present between hits. It is a valuable tool for molecular biologists who wish to get a quick overview of the genetic elements present in a newly sequenced segment of DNA, prior to more elaborate efforts of gene structure prediction and annotation.

In **Chapter 3** a first genome-wide comparison between the emerging genomic sequence resources of tomato and potato is presented. Large collections of BAC end sequences from both species were annotated through repeat searches, transcript alignments and protein domain identification. In-depth comparisons of the annotated sequences revealed remarkable differences in both gene and repeat content between these closely related genomes. The tomato genome was found to be more repetitive than the potato genome, and substantial differences in the distribution of *Gypsy* and *Copia* retrotransposable elements as well as microsatellites were observed between the two genomes. A higher gene content was identified in the potato sequences, and in particular several large gene families including cytochrome P450 mono-oxygenases and serine-threonine protein kinases were significantly overrepresented in potato compared to tomato. Moreover, the cytochrome P450 gene family was found to be expanded in both tomato and potato when compared to *Arabidopsis thaliana*, suggesting an expanded network of secondary metabolic pathways in the *Solanaceae*. Together these findings present a first glimpse into the evolution of Solanaceous genomes, both within the family and relative to other plant species.

**Chapter 4** explores the physical and genetic organization of tomato chromosome 6 through integration of BAC sequence analysis, High Information Content Fingerprinting, genetic analysis, and BAC-FISH mapping data. A collection of BACs spanning substantial parts of the short and long arm euchromatin and several dispersed regions of the pericentromeric heterochromatin were sequenced and assembled into several tiling paths spanning approximately 11 Mb. Overall, the cytogenetic order of BACs was in agreement with the order of BACs anchored to the Tomato EXPEN 2000 genetic map, although a few striking discrepancies were observed. The integration of BAC-FISH, sequence and genetic mapping data furthermore provided a clear picture of the borders between eu- and heterochromatin on chromosome 6. Annotation of the BAC sequences revealed that, although the majority of protein-coding genes were located in the euchromatin, the highly repetitive pericentromeric heterochromatin displayed an unexpectedly high gene content. Moreover, the short arm euchromatin was relatively rich in repeats, but the ratio of *Gypsy* and *Copia* retrotransposons across the different domains of the chromosome clearly distinguished euchromatin from heterochromatin. The ongoing whole-genome sequencing effort will reveal if these properties are unique for tomato chromosome 6, or a more general property of the tomato genome.

**Chapter 5** presents the potato genome, the first genome sequence of an Asterid. To overcome the problems associated with genome assembly due to the high level of heterozygosity that is observed in commercial tetraploid potato varieties, a homozygous doubled-monoploid potato clone was exploited to sequence and assemble 86% of the 844 Mb genome. This potato reference genome sequence was complemented with re-sequencing of a heterozygous diploid clone, revealing the form and extent of sequence polymorphism both between different genotypes and within a single heterozygous genotype. Gene presence/absence variants and other potentially deleterious mutations were found to occur frequently in potato and are a likely cause of inbreeding depression. Annotation of the genome was supported by deep transcriptome sequencing of both the doubled-monoploid and the heterozygous potato, resulting in the prediction of more than 39,000 protein coding genes. Transcriptome analysis provided evidence for the contribution of gene family expansion, tissue specific expression, and recruitment of genes to new pathways to the evolution of tuber development. The sequence of the potato genome has provided new insights into Eudicot genome evolution and has provided a solid basis for the elucidation of the evolution of tuberisation. Many traits of interest to plant breeders are quantitative in nature and the potato sequence will simplify both their characterization and deployment to generate novel cultivars.

The outstanding challenges in plant genome sequencing are addressed in **Chapter 6**. The high concentration of repetitive elements and the heterozygosity and polyploidy of many interesting crop plant species currently pose a barrier for the efficient reconstruction of their genome sequences. Nonetheless, the completion of a large number of new genome sequences in recent years and the ongoing advances in sequencing technology provide many exciting opportunities for plant breeding and genome research. Current sequencing platforms are being continuously updated and improved, and novel technologies are being developed and implemented in third-generation sequencing platforms that sequence individual molecules without need for amplification. While these technologies create exciting opportunities for new sequencing applications, they also require robust software tools to process the data produced through them efficiently. The ever increasing amount of available genome sequences creates the need for an intuitive platform for the automated and reproducible interrogation of these data in order to formulate new biologically relevant questions on datasets spanning hundreds or thousands of genome sequences.



# Samenvatting

Genoomsequencing heeft zich gedurende de laatste twee decennia ontwikkeld van een bewerkelijke en dure technologie, die slechts gebruikt werd door grote internationale consortia, tot een veelgebruikte, geautomatiseerde en betaalbare toepassing die wereldwijd benut wordt door vele onderzoeksgroepen. De genoomsequenties van vele landbouwdieren en gewassen zijn intussen ontrafeld en worden geëxploiteerd voor fundamenteel onderzoek en toegepast voor het verbeteren van veredelings- en fokprogramma's. De ontwikkelingen in sequenceringstechnologie hebben ook hun uitwerking gehad op de benodigde strategieën en toepassingen binnen de bioinformatica, zowel voor het verwerken als voor het interpreteren van de data.

Het onderwerp van dit proefschrift is de toepassing van sequencing, assemblage en annotatie op de genomen van aardappel en tomaat, twee leden van de *Solanaceae* familie. De aardappel is economisch gezien de meest belangrijke soort binnen de *Solanaceae*, ende consumptie ervan draagt bij aan de dagelijkse behoefte aan zetmeel, eiwitten, antioxidanten en vitamines. Tomaten zijn de meest geconsumeerde groenten na aardappelen, en ze vormen wereldwijd een belangrijke bron van lycopene, beta-carotene, vitamine C en vezels. De hoofdstukken in dit proefschrift beschrijven de generatie, het gebruik en de interpretatie van genomische sequenties van deze twee soorten teneinde inzicht te geven in de inhoud, structuur en organisatie van deze genomen.

**Hoofdstuk 1** introduceert de concepten rondom genoomsequencing, assemblage en annotatie en beschrijft de nieuwe technologieën rondom genoomsequencing die het afgelopen decennium ontwikkeld zijn. De onderliggende chemie en werkwijze van deze zogenaamde "Next Generation Sequencing" platformen vertonen aanzienlijke onderlinge verschillen. Als gevolg hiervan is er een aanzienlijke variatie in de productiesnelheid en kwaliteit tussen de platformen. De huidige sequencersapparaten produceren sequenties van uiteenlopende lengtes en maken het mogelijk om sequentieparen te genereren met een bekende afstand tussen deze paren van sequenties. De keuze van sequencerschemie en technologisch platform in combinatie met het type sequencers-template vereist specifieke bioinformatica-toepassingen om de data te verwerken en te interpreteren. Welke sequencingstechnologie en strategie ook gekozen wordt, de resulterende genoom sequentie zelf –uitgedrukt als een lineaire reeks van letters- is slechts matig interessant. Interpretatie van de genoomsequentie kan pas plaatsvinden na sequentie annotatie – dat wil zeggen, de identificatie en classificatie van alle functionele elementen in de sequentie. Wanneer deze elementen eenmaal geannoteerd zijn, kan de genetische variatie die ten grondslag ligt aan de verschillen tussen soorten, danwel tussen accessies binnen een soort, in kaart gebracht worden door middel van sequentievergelijkingen.

In **Hoofdstuk 2** wordt Blastf beschreven, een nieuwe software toepassing waarin gebruik gemaakt wordt van het BLAST programma om op eenvoudige wijze lange

nucleotidesequenties te annoteren. In het algemeen spelen er twee problemen bij het aligneren van een lange nucleotidesequentie tegen een databank van korte gen- of eiwitsequenties: (i) het grote aantal resultaten dat gevonden kan worden door de redundantie van de sequentiedatabank; en (ii) de relatie die geïmpliceerd wordt tussen gealigneerde sequentiesegmenten die eigenlijk betrekking hebben op verschillende elementen zoals genen. BlastIf genereert een overzichtelijke BLAST-uitvoer voor lange nucleotidesequenties door het aantal op elkaar gelijkende resultaten te reduceren, terwijl de grootst mogelijke variatie tussen de resultaten behouden blijft. Het is een waardevolle tool voor moleculaire biologen die op zoek zijn naar een snel overzicht van de elementen in een nieuw gesequeneerd stuk DNA, alvorens meer uitvoerige annotatie plaatsvindt.

**Hoofdstuk 3** presenteert een eerste genoombrede vergelijking tussen de sequenties van tomaat en aardappel. Grote collecties van “BAC end” sequenties van beide soorten zijn geannoteerd aan de hand van databanken van transcriptsequenties, eiwitdomein en repetitieve elementen. Door middel van diepgaande vergelijkingen van de geannoteerde sequenties werden opmerkelijke verschillen gevonden in zowel genen als repetitieve sequenties tussen deze nauw verwante soorten. Het tomaatgenoom bevatte meer repetitieve sequenties dan het aardappelgenoom, en er waren aanzienlijke verschillen tussen de twee genomen in de distributie van *Gypsy* en *Copia* retrotransposons en microsatellieten. In aardappel werd een hogere hoeveelheid genen waargenomen, en bepaalde grote genfamilies zoals cytochroom P450 mono-oxygenases en serine-threonine kinases waren significant overgerepresenteerd in aardappel in vergelijking tot tomaat. Er werd bovendien vastgesteld dat de cytochroom P450 familie uitgedijd is in tomaat en aardappel in vergelijking tot *Arabidopsis thaliana*, hetgeen zou kunnen duiden op een uitgebreider netwerk van secundaire metabolieten in de *Solanaceae*. De bevindingen in dit hoofdstuk hebben een eerste inzicht gegeven in de genomevolutie van de *Solanaceae*, zowel binnen de familie als in relatie tot andere plantensoorten.

De fysieke en genetische organisatie van chromosoom 6 van tomaat wordt verkend in **Hoofdstuk 4** door middel van de integratie van BAC sequentie analyse, “High Information Content Fingerprinting”, genetische analyse, en BAC-FISH karteringsdata. In deze studie is een collectie BACs gesequeneerd en geassembleerd die grote delen van het euchromatine in de korte en lange arm en diverse verspreide gebieden in het heterochromatine bestrijkt, met een totale lengte van ongeveer 11 Mb. De cytogenetische volgorde van de BACs kwam grotendeels overeen met de volgorde zoals deze BACs op de EXPEN 2000 genetische kaart zijn verankerd, hoewel een klein aantal opmerkelijke verschillen gevonden werd. Door de integratie van BAC-FISH, sequenties en genetische karteringsdata ontstond er een helder beeld van de fysieke grenzen tussen het euchromatine en het heterochromatine op chromosoom 6. Annotatie van de BAC sequenties toonde aan dat, hoewel de meeste eiwit-coderende genen in het euchromatine lagen, er een onverwacht hoge gendichtheid in het hoog-repetitieve heterochromatine bestond. Daarnaast bevatte het euchromatine van de korte arm relatief veel repetitieve sequenties, maar er kon een goed onderscheid gemaakt worden tussen euchromatine en heterochromatine door de verhouding tussen de dichtheid

van *Gypsy* en *Copia* retrotransposons in de verschillende delen van het chromosoom. De lopende inspanning om het gehele tomaatgenoom te sequencen zal uitwijzen of deze eigenschappen uniek zijn voor chromosoom 6, of algemeen gelden voor het tomaatgenoom.

In **Hoofdstuk 5** wordt het aardappelgenoom gepresenteerd, de eerste genomesequentie van een Asterid. Om problemen te voorkomen tijdens de genomassemblage door het hoge niveau van heterozygotie dat gevonden wordt in commerciële tetraploïde aardappelrassen, is gebruik gemaakt van een homozygote dubbel-haloïde aardappelkloon om 86% van het 844 Mb grote genoom te sequencen en te assembleren. Dit referentiegenoom voor aardappel is vergeleken met sequentiedata van een heterozygote diploïde kloon, waardoor de verdeling en mate van sequentiepolymorfisme zowel tussen verschillende genotypen als binnen een enkel heterozygoot genotype in kaart gebracht kon worden. Tussen de drie genotypen werd een hoge frequentie gevonden van verschil in aan-/afwezigheid van genen en andere potentieel schadelijk mutaties. Deze hoge mate van variatie vormt een waarschijnlijke verklaring voor de inteeltdepressie in aardappel. De annotatie van het genoom werd ondersteund door het diep sequencen van zowel het transcriptoom van de dubbel-monoploïde aardappel als dat van de heterozygote aardappel, en resulteerde in de voorspelling van ruim 39,000 eiwit-coderende genen. Tijdens de analyse van de beide transcriptomen werd bewijs gevonden voor de bijdrage aan de evolutie van knolontwikkeling door de expansie van genfamilies, weefsel-specifieke expressie en de rekrutering van genen in nieuwe reactiepaden. Op deze manier heeft de sequentie van het aardappelgenoom nieuwe inzichten in de evolutie van de genomen van de Eudicotylen opgeleverd, en een solide basis gelegd voor de opheldering van knolvorming. Veel eigenschappen die voor veredelaars interessant zijn, zijn van nature kwantitatief, en de aardappelgenoomsequentie zal de karakterisering en toepassing van deze eigenschappen in nieuwe aardappelvariëteiten sterk vereenvoudigen.

In **Hoofdstuk 6** worden de openstaande uitdagingen in het sequencen van plantengenomen behandeld. De hoge dichtheid van repetitieve elementen en de heterozygotie en polyploïdie van veel interessante gewassen vormen op dit moment een barrière voor het efficiënt reconstrueren van hun genomesequenties. Desalniettemin bieden het grote aantal recentelijk gecompleteerde genomesequenties en de voortdurende vooruitgang in sequentietechnologie een groot scala aan nieuwe mogelijkheden voor plantenveredeling en genomonderzoek. De huidige sequentieplatformen worden continu bijgewerkt en verbeterd, en nieuwe technologieën zijn reeds in ontwikkeling en worden geïmplementeerd in de derde generatie van sequentieplatforms. Deze zullen in staat zijn om individuele moleculen te sequencen zonder dat er amplificatie nodig is. Hoewel deze technologieën vele mogelijkheden creëren voor nieuwe toepassingen, vereisen ze ook robuuste softwarepakketten om de geproduceerde data te verwerken. De steeds maar toenemende hoeveelheid beschikbare genomesequenties schept de behoefte aan een intuïtief platform voor de automatische en reproduceerbare interrogatie van deze data, teneinde nieuwe, biologisch relevante vragen te formuleren aan de hand van datasets die honderden of zelfs duizenden genomesequenties bevatten.



## Curriculum vitae

Erwin Datema was born on August 15<sup>th</sup> 1980 in Delft, the Netherlands. After finishing his Gymnasium education at the Esdal College in Emmen in 1998 he started his Bachelor in Plant Biotechnology at Larenstein International Agricultural College in Velp. During his internships on construction of cDNA libraries from small RNA molecules and purification of small proteins at Plant Research International (PRI) in Wageningen he first came into contact with bioinformatics. After obtaining his BSc degree in 2002 he started a Master in Bioinformatics at Wageningen University. He completed his Master education with a thesis research on the construction of a plant transcriptome database at the Centre for Molecular and Biomolecular Informatics (CMBI) in Nijmegen under supervision of Prof. Dr. Gert Vriend, and another thesis research at PRI in Wageningen where he evaluated the state-of-the-art in gene prediction for Solanaceous plants. There appeared to be a good match between Erwin and PRI, and after earning his MSc degree in 2005 he accepted a PhD position there on the annotation of tomato chromosome 6 under supervision of Dr. Roeland van Ham and Prof. Dr. Willem Stiekema. This project was embedded in the international Tomato Genome Sequencing Consortium; however the turbulent developments in genome sequencing technologies lead to a re-definition of the tomato genome sequencing project, and therefore Erwin's thesis expanded to include not only work on the tomato genome but also two potato genomes. Starting September 1<sup>st</sup> 2011, Erwin has taken up a position as Researcher Sequencing Applications at Keygene N.V. in Wageningen.



## Publications

The International Tomato Genome Sequencing Consortium. **The genomes that make tomatoes.** *In preparation.*

The Potato Genome Sequencing Consortium. **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011; 475:189–95.

Peters SA, Datema E, Szinay D, van Staveren MJ, Schijlen EGWM, van Haarst JC, Hesselink T, Abma-Henkens MHC, Bai Y, de Jong H, Stiekema WJ, Klein Lankhorst RM, van Ham RCHJ. **Solanum lycopersicum cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements.** *Plant J.* 2009; 58:857-69.

Tang X, Szinay D, Lang C, Ramanna MS, van der Vossen EAG, Datema E, Klein Lankhorst RM, de Boer J, Peters SA, Bachem CWB, Stiekema WJ, Visser RGF, de Jong H, Bai Y. **Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements.** *Genetics.* 2008; 180:1319-28.

Chang SB, Yang TJ, Datema E, van Vugt J, Vosman B, Kuipers A, Meznikova M, Szinay D, Klein Lankhorst RM, Jacobsen E, de Jong H. **FISH mapping and molecular organization of the major repetitive sequences of tomato.** *Chromosome Res.* 2008;16:919-33.

Datema E, Mueller LA, Buels R, Giovannoni JJ, Visser RGF, Stiekema WJ, van Ham RCHJ. **Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato.** *BMC Plant Biol.* 2008; 8:34.

Fiers MWEJ, van der Burgt A, Datema E, de Groot JCW, van Ham RCHJ. **High-throughput bioinformatics with the Cyrille2 pipeline system.** *BMC Bioinformatics.* 2008; 9:96.



# Acknowledgements

During the course of this thesis research I have had the opportunity to work together with a large group of amazing people. First and foremost of these are my direct co-workers from the Applied Bioinformatics group at Plant Research International and the Laboratory of Bioinformatics at Wageningen University. The many discussions I had with you, aided by our whiteboards, have helped me to order my thoughts and ideas into the chapters of this thesis. You have provided me with an amazing professional and personal environment and the successful completion of my thesis research is in no small part thanks to all of you. Indeed you made me feel at home so much that I chose not to hurry up and finish my thesis in time, but instead stick around for a couple of more years!

Whereas many PhD students often end up doing their own little research on their own little islands, much of the work in this thesis has a solid foundation in collaboration with others, both within and outside Wageningen. Through collaborations with the Plant Breeding, Phytopathology and Cytogenetics groups I have met many great new people, and learned many new and interesting things that have broadened my horizon. I also feel privileged to have been a part of two large, globe-spanning collaborations on the tomato and potato genomes. These collaborations have provided me with both an excellent platform to expand my knowledge and perform my research, as well as a great environment to meet interesting new people and share ideas. The fruits (and tubers) produced from the work within these consortia are indeed greater than the sum of their parts, and I want to thank all of you for making me feel welcome and at home.

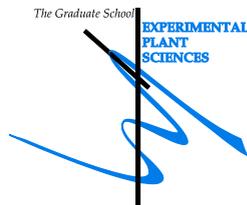
I have tried to avoid naming individual people as much as possible in order to avoid leaving anyone out, but there are a few persons whose contributions to my thesis deserve a special mention. The work on tomato chromosome 6 forms one of the most important parts of the work presented here and the successful completion of this can only be attributed to the willingness of all people involved to work together. I am very grateful to Sander in particular for seizing this opportunity. Together with the people at Greenomics and the Laboratory of Cytogenetics we poured our ideas and findings into a nice story, which eventually even inspired our international collaborators to write a strikingly similar story. I am also very pleased that this work lead to even more collaborations between Applied Bioinformatics and Cytogenetics, as is clear from my other publications outside this thesis.

I would like to thank Roeland for his excellent day-to-day supervision, for providing me with all the opportunities to travel around the world to share my work and ideas with so many other amazing scientists, and for challenging me to always improve myself. I feel that under his auspices I have grown from a young, naïve and stubborn master student into a somewhat more mature, slightly less naïve and unfortunately just as stubborn researcher.

Most of all, I want to thank Evelyn for loving me and showing me her endless patience.



# Education statement of the graduate school Experimental Plant Sciences



**Issued to:** Erwin Datema  
**Date:** 11 November 2011  
**Group:** Laboratory of Bioinformatics  
 Wageningen University & Research Centre

1) Start-up phase	<u>date</u>
▶ <b>First presentation of your project</b> Functional Annotation of Tomato Chromosome 6	Sep 09, 2005
▶ <b>Writing or rewriting a project proposal</b> Functional Annotation of Tomato Chromosome 6	Sep 07, 2005
▶ <b>Writing a review or book chapter</b>	
▶ <b>MSc courses</b>	
▶ <b>Laboratory use of isotopes</b>	

*Subtotal Start-up Phase*      4.5 credits\*

2) Scientific Exposure	<u>date</u>
▶ <b>EPS PhD student days</b> EPS Student day, Wageningen University	Sep 19, 2006
EPS Student day, Wageningen University	Sep 13, 2007
▶ <b>EPS theme symposia</b>	
EPS Theme 4 symposium 'Genome Plasticity', Wageningen University	Dec 12, 2008
EPS Theme 4 symposium 'Genome Plasticity', Radboud University Nijmegen	Dec 11, 2009
▶ <b>NWO Lunteren days and other National Platforms</b>	
NBIC 2006 (Netherlands BioInformatics Conference), Ede	Apr 24, 2006
BBC 2006 (Benelux Bioinformatics Conference), Wageningen	Oct 17&18, 2006
NBIC 2007 (Netherlands BioInformatics Conference), Amsterdam	Apr 17&18, 2007
BBC 2007 (Benelux Bioinformatics Conference), Leuven (Belgium)	Nov 12&13, 2007
BioRange 2008, Lunteren	Mar 05&06, 2008
BBC 2008 (Benelux Bioinformatics Conference), Maastricht	Dec 15&16, 2008
▶ <b>Seminars (series), workshops and symposia</b>	
Workshop Bioinformatics: From Genomes to Systems Biology, Munich (Germany)	Nov 09&10, 2006
WUR BioInformatics Day, Wageningen	Nov 13, 2006
Evolutionary Bioinformatics Symposium, Utrecht	Nov 16, 2006
CBSG Summit, Wageningen	Feb 06&07, 2007
CBSG workshop data mining, Wageningen	Feb 20, 2007
PGSC meeting, Wageningen	Apr 18&19, 2007

Cytoscape symposium, Amsterdam	Nov 09, 2007
EU-SOL BioMoby Workshop, Cologne (Germany)	Mar 11-14, 2008
CBSG Summit, Wageningen	Mar 17&18, 2008
Potato workshop, Wageningen	Apr 21&22, 2008
Tomato finishing workshop, Wageningen	Apr 24&25, 2008
DAS workshop, Welcome Trust Genome Campus, Hinxton (UK)	Mar 09&10, 2009
Next Generation Sequencing user meeting, Utrecht	Jul 07, 2009
CBSG Summit, Wageningen	Sep 24, 2009
CBSG Summit, Wageningen	Mar 15&16, 2010
ESF Next Generation Sequencing meeting, Leiden	Aug 30-31, & Sep 01, 2010
▶ <b>Seminar plus</b>	
▶ <b>International symposia and congresses</b>	
Plant & Animal Genomes 2006, San Diego (USA)	Jan 14-18, 2006
PAA/Solanaceae 2006, Madison (USA)	Jul 23-27, 2006
SOL Annotation meeting, Ghent (Belgium)	Oct 23-25, 2006
EU-SOL kickoff meeting, Wageningen	Nov 21&22, 2006
ECCB/ISMB 2007, Vienna (Switzerland)	Jul 21-25, 2007
EU-SOL 2007, Rome (Italy)	Nov 19&20, 2007
SOL 2008, Cologne (Germany)	Oct 13-16, 2008
EU-SOL 2008, Toulouse (France)	Nov 13-17, 2008
EU-SOL 2009, Toledo (Spain)	Oct 05&06, 2009
Plant & Animal Genomes 2010, San Diego (USA)	Jan 09-13, 2010
EU-SOL 2010, Rome (Italy)	Mar 17-19, 2010
SOL 2010, Dundee (Scotland)	Sep 06-08, 2010
EU-SOL / Lat-SOL 2010, Natal (Brazil)	Nov 13-16, 2010
▶ <b>Presentations</b>	
Poster @ BBC 2006	Oct 2006
Presentation @ Plantenveredeling group	May 2007
Presentation @ Fytopathologie group	Jun 2007
Poster @ BBC 2007	Nov 2007
Poster @ BioRange 2008	Mar 05, 2008
Presentation @ Natural variation in plants course	Aug 28, 2008
Presentation @ EPS Theme 4 symposium	Dec 12, 2008
Presentation @ Next Generation Sequencing course	Mar 11, 2009
Presentation @ Next Generation Sequencing user meeting	Jul 07, 2009
Presentation @ ESF Next Generation Sequencing meeting	Aug 31, 2010
▶ <b>IAB interview</b>	Sep 14, 2007
▶ <b>Excursions</b>	

*Subtotal Scientific Exposure*

*37.4 credits\**

<b>3) In-Depth Studies</b>	<i>Date</i>
	▶ <b>EPS courses or other PhD courses</b>
	Multivariate analysis Natural variation in plants Introduction to next-generation sequencing: technologies and data analysis Managing Life Science Information
	▶ <b>Journal club</b> once every 2 weeks, Bioinformatics
▶ <b>Individual research training</b>	

*Subtotal In-Depth Studies*      7.5 credits\*

<b>4) Personal development</b>	<i>date</i>
	▶ <b>Skill training courses</b>
	Project and time management Mobilizing your Scientific Network Information Literacy PhD including EndNote Introduction
	▶ <b>Organization of PhD students day, course or conference</b>
▶ <b>Membership of Board, Committee or PhD council</b>	

*Subtotal Personal Development*      3.1 credits\*

<b>TOTAL NUMBER OF CREDIT POINTS*</b>	<b>52.5</b>
---------------------------------------	-------------

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits.

\* A credit represents a normative study load of 28 hours of study.

The research described in this thesis was financially supported by (in alphabetical order) Centre for BioSystems Genomics (CBSG), EU-SOL through the 6<sup>th</sup> Framework Programme of the European Commission, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) and Stichting voor de Technische Wetenschappen (STW).

This thesis was printed at Wöhrmann Print Service B.V.

The cover image was generated using AndreaMosaic.