Chapter 16
Advances in Sequencing the Barley Genome

Nils Stein and Burkhard Steuernagel

Contents

16.1 Introduction 392
16.2 Next Generation Sequencing: New Perspectives for Barley Genome Analysis 392
16.2.1 Genome Survey and Genome Composition 393
16.2.2 Survey Sequencing of Sorted Chromosomes—A Gene Index of Barley 394
16.2.3 Next Generation Sequencing of BAC Clones 394
16.2.4 Genotyping by Sequencing 396
16.3 Whole Genome Sequencing of Barley—the Challenge Shifts from Sequencing to Assembly 397
16.4 Outlook 400
References 401

Abstract Barley genome sequencing is lagging behind the status achieved for many other crop genomes although barley is ranking worldwide as fifth most important crop species. Whole genome sequencing of barley with classical Sanger sequencing technology was long meant to be too costly due to the very large genome size of more than 5 Gigabases. By the introduction of Next Generation Sequencing technology this situation has changed and fascinating new possibilities opened up for in depth barley genome analysis and whole genome sequencing. Genome composition has been revealed at unprecedented resolution. A linear gene order map comprising two thirds of all barley genes could be developed and the approach is currently adopted for other related and important cereal genomes like wheat and rye. Important technical limitations have been solved making even whole genome sequencing in barley a feasible endeavor. Provided these new possibilities, it is becoming obvious that soon sequencing per se is no longer the limiting factor but sequence assembly remains the challenge. This review will provide a brief summary of the recent developments in barley genome sequencing achieved since the introduction of Next Generation Sequencing.

Keywords Barley · Next generation sequencing · Genome sequencing · BAC clones · EST · SNP · Haplotype · Synteny · Retrotransposons · Heterochromatin · Hordeum vulgare · Triticeae

N. Stein (✉) · B. Steuernagel
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
Corrensstr. 3, 06466 Gatersleben, Germany
e-mail: stein@ipk-gatersleben.de

R. Tuberosa et al. (eds.), *Genomics of Plant Genetic Resources,*
DOI 10.1007/978-94-007-7572-5_16,

16.1 Introduction

The barley (*Hordeum vulgare*) genome comprises over 5 billion base pairs which equals about double the size of the maize (*Zea mays*) genome or twelve times the genome of rice (*Oryza sativa*). If one assumes an overall investment of 100 million dollars for the sequencing of the rice genome it is obvious that sequencing the barley genome was not fundable as long scenarios were based on classical Sanger sequencing (Sanger et al. 1977). Until recently the main sequence information of the barley genome has been obtained by sequencing expressed genes (Expressed Sequence Tags, ESTs, Zhang et al. 2004) or contiguous stretches of genomic DNA (Bacterial Artificial Chromosome (BAC) clone contigs) selected in frame of map-based cloning of important genes (reviewed by: Stein and Graner 2004; Krattinger et al. 2009; Eversole et al. 2009). Even at this limited access to the barley genome, important features of genome organization could be revealed and this has been reviewed before (Stein 2007). More recently, due to the availability of novel high-throughput sequencing technologies (Next Generation Sequencing, NGS, Mardis 2008; Holt and Jones 2008) costs for sequencing have tremendously decreased. This has stimulated and facilitated genome sequencing and genome analysis in many organisms including the major crop species. The present chapter reviews recent achievements in barley genome analysis that were exclusively triggered by the technical innovation provided by NGS.

16.2 Next Generation Sequencing: New Perspectives for Barley Genome Analysis

NGS originally referred to novel sequencing technology concepts that provided an alternative, cost efficient platform to previous "gold-standard" Sanger sequencing (Sanger et al. 1977), which relies on fluorescent-di-desoxy-terminator chemistry in combination with capillary electrophoresis and laser-detection devices. The first commercial systems entered the market as "short read" sequencing technologies providing sequence read lengths of between 30 and 100 bp (Service 2006; Holt and Jones 2008). Keeping in mind that these sequencing technologies were especially designed for "re-sequencing" of high-quality reference genome sequences (i.e. the human genome), it was unclear how useful such technologies would proof to be for de novo sequencing of large crop genomes like barley which used to lack any kind of reference sequence. This concern was based on the difficulties of the barley genome carrying 80 % repetitive DNA (Flavell et al. 1974) with the main constituent repetitive elements expanding each over several kilobases (i.e. the BARE1 element, Manninen and Schulman 1993). De novo sequencing and assembly of such structures would instantly lead to false assemblies and gaps. It was therefore important to critically assess the true potential of short read technologies and revisit regularly if their use would impact the strategy of barley genome sequencing.

16.2.1 Genome Survey and Genome Composition

Two early studies were designed to test the usefulness of 30 bp (Illumina Solexa) or 100 bp (Roche/454 GS20) short read technology, respectively, for skim sequencing the barley genome (Wicker et al. 2008; Wicker et al. 2009). Sequence depth was shallow in both cases reaching to between 1 % (100 bp GS20) and 10 % (30 bp Solexa) haploid genome equivalents. At this coverage any attempt of genome sequence assembly was rather meaningless. But the statistical properties of the datasets revealed interesting characteristics of the barley genome.

In the initial study (Wicker et al. 2008) the 30 bp short reads were utilized to generate an index of mathematically defined repeats (MDR index, Kurtz et al. 2008). For this purpose all high quality bases of the almost 270 Mio sequence reads were used to count the occurrence of all 20 mers (a sequence word of 20 consecutive nucleotides). 159 million discrete 20 mers (sequence differs at least at one nucleotide from all other 20 mers) could be determined and 88 % of the discrete 20 mers occurred only once in the dataset. Almost 99 % of the discrete 20 mers occurred only between 1 and 10 times in the barley genome. One percent of the discrete 20 mers occured eleven times or more often and thus constituted 30 % of the barley genome. One of these frequent 20 mers was present in almost 170,000 copies. Together this analysis provided a good reflection of the repetitiveness of the barley genome (Wicker et al. 2008).

In the second study the genome was sequenced to 1 % haploid genome coverage with 100 bp reads of the early Roche/454 GS20 technology platform (Wicker et al. 2009). The resulting 570,000 sequence reads were systematically compared against databases in order to determine the fraction of sequences that could either be related to genes or known repetitive DNA elements of Triticeae genomes. Interestingly but not unexpectedly, similarity to genes was detected only in less than 1 % of the entire dataset. 50 % of the sequence data could be assigned to only 14 families of transposable elements (TE) with the BARE1 family alone representing about 13 % of the barley genome. This is a 5–10 times higher frequency compared to its original copy number estimates (Manninen and Schulman 1993; Vicient et al. 1999). Sequences that could not be immediately related to any known class of DNA (TE, genes, SSR, organellar genomes) were assembled. Based on the shallow sequence depth resulting sequence clusters could per se be classified as repetitive sequence. Altogether 70 % of this "snapshot" sequence dataset comprised repetitive DNA. Based on the assumption that the dataset was representative for the overall composition of the barley genome it was compared to the sequence composition of individual BAC clone sequences from public databases. Such "gene-containing" BAC clones exhibited significantly different sequence compositions. Caspar transposons were over- and BAGY2 retrotransposons were underrepresented on the analysed set of BACs. The genome-wide distribution of these two TE classes was monitored by fluorescent in-situ hybridization (FISH) to barley metaphase spreads. The Caspar elements were predominantly clustered at the telomeric ends of chromosomes whereas BAGY2 elements produced almost a mirror image labeling of chromosomes covering all the pericentromeric regions but avoiding subtelomeric parts. This analysis of sequence element frequencies in comparison to a genomic index thus confirmed the subtelomeric origin of the analysed BAC clones (Wicker et al. 2009).

16.2.2 Survey Sequencing of Sorted Chromosomes—A Gene Index of Barley

Barley genome size is a major disadvantage for genome sequencing. On the other hand, the size of its genome is an advantage for cytogenetic applications. This feature could be exploited for PCR-based detection of markers from micro-dissected chromosomes (Sorokin et al. 1994). The first "physical" map of all barley chromosomes was produced by PCR amplification of genetic markers from micro-dissected chromosome arms or deletion chromosomes (Künzel et al. 2000). The usefulness of the size of Triticeae chromosome was further demonstrated by the feasibility of purifying mitotic chromosomes. Chromosome suspensions obtained from synchronized root tip meristems can be utilized for flow-cytometric sorting of over 90 % pure fractions of individual chromosomes (reviewed by Doležel et al. 2007). The Roche/454 GSFLX system was utilized to test whether such purified chromosomal DNA could be used for direct shotgun sequencing (Fig. 16.1). Chromosome 1H of barley was sequenced to about 1-fold coverage and the dataset provided partial sequence access to up to 80 % of all genes located on this respective chromosome (Mayer et al. 2009). By exploiting the extent of conserved synteny between barley and the sequenced genomes of rice and sorghum (International Rice Genome Sequencing Project 2005; Paterson et al. 2009), a potential linear order model of almost 2,000 barley genes detected in the shotgun sequences could be proposed (Fig. 16.1). These results stimulated a genome-wide analysis and all remaining barley chromosomes were sequenced to 1-fold coverage by using Roche/454 GSFLX Titanium (Mayer et al. 2011). The strategy previously applied to chromosome 1H was further reinforced by including into the analysis the information of a third sequenced model grass genome, *Brachypodium distachion* (The International Brachypodium Initiative 2010). A linear gene order map of more than 21,000 genes—perhaps two thirds of all barley genes—could be constructed with sequence tag access to all of these genes (Mayer et al. 2011). The strong enabling potential of survey sequencing of flow-sorted barley chromosomes was very convincing and stimulated similar studies on wheat chromosomes 1A, 1B, 1D (Wicker et al. 2011), 4A (Hernandez et al. 2011), 5A (Vitulo et al. 2011), 7BS and 7DS (Berkman et al. 2011a; Berkman et al. 2011b). The International Wheat Genome Sequencing Consortium has furthermore adopted the concept for wheat genome survey sequencing (www.wheatgenome.org).

16.2.3 Next Generation Sequencing of BAC Clones

Shotgun sequencing of larger stretches of genomic DNA (i.e. the insert of an average sized BAC clone) of barley can be challenging even if Sanger sequencing is applied. Seldom would an entire insert of a BAC immediately assemble into one single contig. This most often is hampered due to the presence of multiple copies of highly conserved members of the same class of repetitive elements. Therefore, it was uncertain how sequencing and assembly of barley BAC clones would perform on the basis of NGS. It could be shown by re-sequencing a few BACs (previously

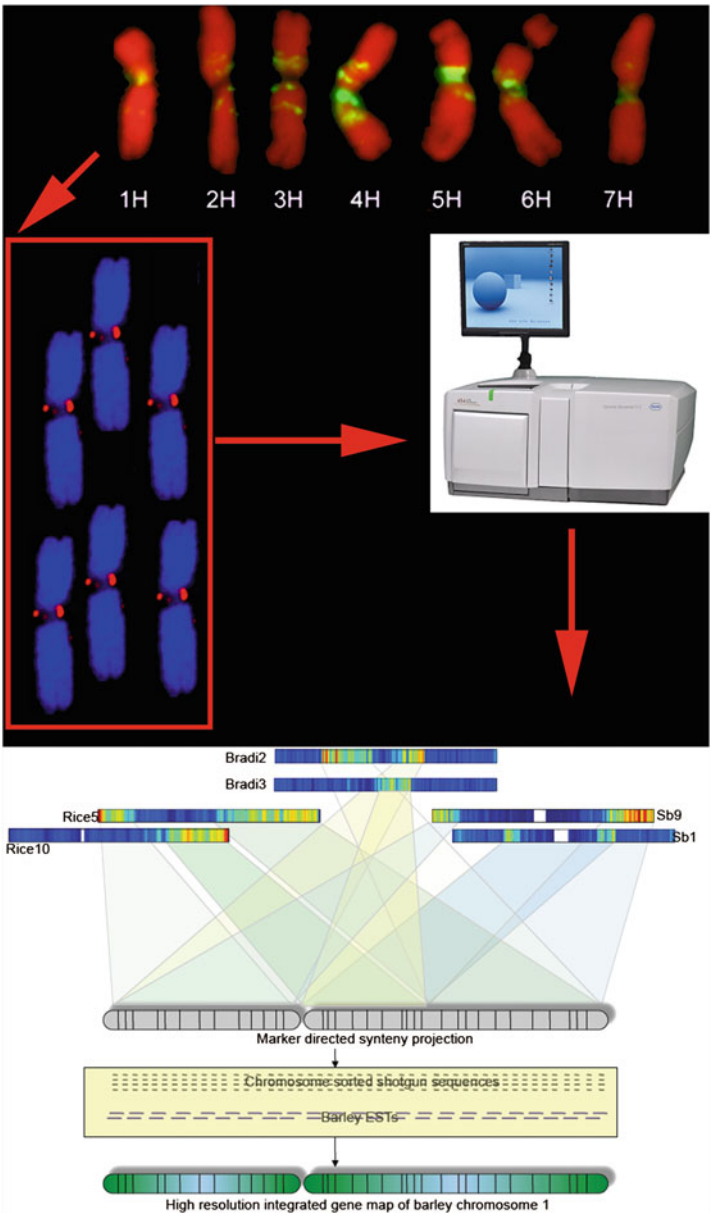

Fig. 16.1 Chromosomal genomics in barley. Barley chromosomes differ in size. This size difference can be exploited for the purification of individual chromosomes which provide, after enzymatic amplification of the minute amounts of DNA, an excellent template for Next Generation Sequencing. Sequencing to around 1-fold chromosomal coverage is providing access to about 70–80 % of all genes present on the respective chromosome. This data can be integrated on the basis of a dense gene-based marker map of barley and gene order information from syntenic regions of sequenced model grass genomes into linear gene order maps of individual barley chromosomes—so-called "genome zippers" (Mayer et al. 2011; Mayer et al. 2009)

sequenced by Sanger technology) that Roche/454 GS20 sequencing could provide more even sequencing coverage due to the lack of cloning bias (Wicker et al. 2006). Assemblies were more fragmented but sequence accuracy in the assembled regions was comparable to Sanger sequencing. The even read coverage helped to detect misassemblies in the originally Sanger sequenced and assembled BACs. Importantly, the genes present on the sequenced BACs almost exclusively assembled into single contigs comprising the entire gene structure (Wicker et al. 2006). This feature was exploited in a recent study of sequencing 400 gene-containing BAC clones of chromosome 3H from barley cultivar Haruna Nijo. Pools of 10 or 20 BACs each were sequenced and the resulting sequences were combined in a mixed assembly. This produced 7,512 contigs larger than 500 bp. In contrast to the 444 ESTs originally used for the identification of BAC clones overall 1,239 open reading frames (ORF) could be detected in the assembled sequence and thus assigned to chromosome 3H (Sato et al. 2011).

The large sequencing capacity provided by NGS platforms per individual sequencing run requires for multiplexing of samples if the advantage of cost efficiency should not be compromised. Several procedures were established all aiming at the incorporation of unique sequence identifiers (multiplex identifier, MID; barcode tags) into the fragments obtained from different samples while ligating specific sequencing adaptors (i.e. Meyer et al. 2008). Ninety-six barley BAC clones were sequenced in parallel to average 20-fold coverage in a single Roche/454 GSFLX run. Eighty percent of the analysed clones assembled into less than 10 contigs at N50 of 50 kb for all 96 clones. Superior assembly results may be expected if BAC clones would not just be shotgun sequenced but if paired-end or mate-pair sequencing strategies would be applied which provide two sequences per DNA fragment and their physical linkage information. This strategy has been introduced in the early times of genome sequencing (Roach et al. 1995). It can be implemented today also on all NGS platforms, however, together with individual sample bar-coding it comes at the disadvantage of substantially more labor. The effect of implementing information obtained from mate-pair sequencing has been simulated for barley by combining the *de novo* shotgun 454 sequence assemblies of the above mentioned 96 BACs with additional 2×36 bp sequence reads from a non-barcoded 2.5 kb mate-pair library of the same 96 BACs which allowed to scaffold 80 % of the assembled contig length (Taudien et al. 2011). Based on the outcome of Roche/454 barcoded BAC pool sequencing two projects for sequencing chromosome 3H in barley (http://barleygenome.org) and 3B (http://urgi.versailles.inra.fr/Projects/3BSeq) in wheat have been initiated.

16.2.4 Genotyping by Sequencing

Access to high-throughput sequencing has not only stimulated research towards whole genome sequencing. It can be predicted that many of today's marker technologies will be replaced by one or the other kind of NGS application in the near/midterm future. Whole genome skim sequencing was used in rice for high-density SNP mapping in a population of 150 recombinant inbred lines (RILs). Each

individual was sequenced to average 0.02-fold genome coverage which allowed scoring of 1.2 million SNPs at a density of 3.2 SNP/kb (Huang et al. 2009). To apply this approach economically to large genome species it is appropriate to implement steps of reducing genome complexity. In restriction-associated DNA (RAD) sequencing a fraction of restriction fragments produced with specific endonucleases is size selected and sequences are generated adjacent to these restriction sites (Baird et al. 2008). A high-density haplotype map was developed by such approach for maize (Gore et al. 2009) and other species (reviewed by Rowe et al. 2011). Two studies showed the applicability to barley. 10,000 RAD fragments were generated between the parental genotypes of the Oregon Wolfe Barley (OWB) mapping population. This included 530 fragments with codominant polymorphism between both genotypes of which 436 could be genetically mapped (Chutimanitsakun et al. 2011). This approach was extended further by including presence/absence polymorphisms to the linkage analysis. For the same OWB population ~24,000 sequence tags could be mapped into the existing framework map comprising already 2,382 markers (Elshire et al. 2011).

16.3 Whole Genome Sequencing of Barley—the Challenge Shifts from Sequencing to Assembly

Whole genome sequencing in barley is feasible now because of the possibilities provided by NGS. The pure sequencing costs for the entire barley genome likely amount to less than 3 million EURO for a multiplex barcoded BACpool sequencing project with Roche/454 GS FLX + (60,000 BACs, average insert 100 kb, 50 EURO/BAC) or to as little as 50,000 EURO for 100-fold whole genome shotgun sequencing with Illumina Hiseq2000 technology. Costs of labor and bioinformatics required for assembly and annotation have not been considered and would add up. However, having these different options in mind it is important to consider the quality of sequence that is aimed for. Recently published genome sequences produced by NGS often suffer from limited quality, hence they do allow only for limited genome wide studies and conclusions (Chain et al. 2009; Feuillet et al. 2011). The International Barley Genome Sequencing Consortium (IBSC, http://barleygenome.org) has proposed a multistep genome sequencing strategy which implements the advantages and quality provided by whole genome shotgun and hierarchical BAC-by-BAC sequencing utilizing a densely anchored physical map as a template (Schulte et al. 2009).

Since generating sequence data is principally no longer a limitation, the main challenge in genome sequencing shifted to sequence assembly. In any shotgun sequencing approach (WGS and hierarchical clone-by-clone shotgun sequencing) a genome needs to be reconstructed from all sequence reads by connecting overlapping reads (Miller et al. 2010). An assembly can be divided into three phases (Fig. 16.2): The first phase is the construction of contigs where a contig is a set of reads that are inter-related by overlap of their sequences (Staden 1980). All reads belong to one and only one contig. Each contig contains at least one read. In the second phase these unordered contigs are then ordered and directed. Results of the second phase are called scaffolds where a scaffold is a set of directed and ordered contigs, but gaps

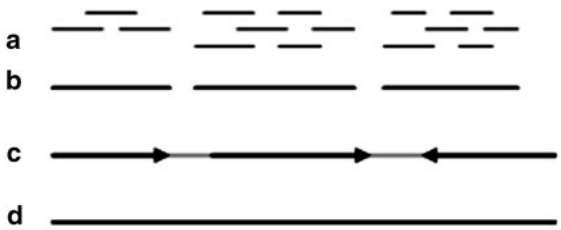

Fig. 16.2 Schematic figure of **a** reads, **b** contigs, **c** a scaffold and **d** a finished sequence. Sequence reads are provided as an output of a sequencing machine. They are used in the first phase of an assembly to produce unordered contigs. The second phase of the assembly results in scaffolds. The order and mutual direction of contigs within one scaffold is known, but gaps between the contigs are allowed. Ideally the distance of contigs is also known. In a third phase the gaps between contigs are closed

Table 16.1 Selection of current state-of-the-art NGS assembly softwares for de novo genome assembly

Assembler	DataStructure	Large Datasets[10)]	Platforms[11)]	Comment
MIRA[1)]	Overlap	–	all	Very good on small but highly repetitive datasets
Newbler[2)]	Overlap	+	454/ fasta	Distributed together with 454 Platform
CABOG[3)]	Overlap	–	454/ fasta	Celera improved for 454
SOAPdenovo[4)]	de Bruijn	+	Illumina	Used for various large genomes
CLC Assembly Cell[5)]	de Bruijn	++	Illumina/ 454/ fasta	Commercial! Very efficient memory usage.
Velvet[6)]	de Bruijn	–	Illumina/ fasta	Pipelines for large datasets in preparation (Cortex, Curtain)
AbySS[7)]	de Bruijn	+	Illumina/ fasta	Designed for cluster computing
ALLPATH-LG[8]	de Bruijn	+	Illumina	Specific variety of paired-end-libraries required
SGA[9]	Overlap (String-graph)	++	Illumina/ 454/ fasta	Burrows-Wheeler transform for overlap detection

[1)] Chevreux et al. 1999, [2)] www.my454.com, [3)] Miller et al. 2010, [4)] Li et al. 2010, [5)] www.clcbio.com, [6)] Zerbino and Birney 2008, [7)] Simpson et al. 2009, [8)] Gnerre et al. 2011, [9)] Simpson and Durbin 2011, [10)] describes its performance on genomes larger than 2 Gb: − = not possible, + = possible on a high-end supercomputer (∼1 TB RAM), ++ = possible on standard computer (∼100 GB RAM); [11)] only Illumina GAIIx and Hiseq, 454 and general fasta are concerned.

between such contigs are allowed. In a third phase, gaps between contigs within scaffolds are closed and the number of scaffolds is minimized. The result of phase three is a reference sequence ideally reflecting the sequence of the real genome.

The main challenges of whole genome sequence assembly in barley are imposed by genome size and content of repetitive DNA. Different algorithms for assembly may not necessarily cope equally well with both issues. A collection of current assembly tools for NGS data is listed in Table 16.1. Assembly softwares usually run a

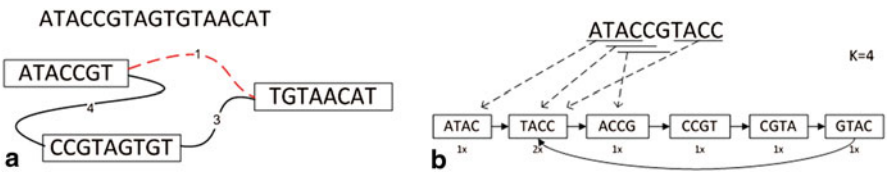

Fig. 16.3 Examples for an overlap graph **a** and a de Bruijn graph **b**. In an overlap graph every node represents one read from a sequencing machine. An edge between two nodes is drawn if the reads overlap. The edge can be weighted according to the degree of overlap. In a de Bruijn graph each occurring k-sized substring of any read is represented as a node but the substring of each node is unique. Only the number of occurrences of the substring can be tracked for each node. Thus the complexity of this graph does not increase if sequencing with a larger depth. Numbers in the graphs help to extract consensus contigs. In **a** the weight of a node represents the degree of overlap between reads. Circular structures in the graph are resolved by deleting edges with the least overlap. In **b** the numbers below each node track the number of occurrence in the input sequence. In this simple example the sequence can be constructed from the graph by following the directed graph until each node is passed as many times as its occurrence is denoted

graph-based data structure internally to compute contigs. The traditional graph structures are overlap-graphs: a node in a graph represents a sequence read and an edge between two nodes represents an overlap between the two reads (Fig. 16.3a). The overlap-graph is read-centric, containing as many nodes as reads are used. The assembled consensus contig sequence is then extracted from the graph. This may be achieved by applying weight to edges according to overlap length and include edges subsequently starting with the largest overlap and discarding every edge that would result in a circular sub-graph (Myers et al. 2000). Graph-based assembly tools like MIRA or NEWBLER proved to be efficient for highly accurate assembly of sequenced barley BAC clones (Steuernagel et al. 2009; Taudien et al. 2011; Wicker et al. 2006). At a size of 5 Gbp, barley whole genome sequencing to 50-fold coverage with 100 bp reads would produce 2.5 billion reads. Since the overlap-graph algorithms require every-read to every-other-read comparisons this results in a quadratic problem of necessary computing steps. Additionally all reasonable read-pairings have to be stored. This quickly exceeds the memory capacity of even a super-computer ($\sim$1 Terabyte RAM) and basically excludes overlap-graph based assembly algorithms for whole genome assembly in barley. This bottleneck may not occur if the construction algorithm and storage data-structure is optimized dramatically. An example for such an optimized solution is maybe provided by the optimized SGA-assembler (Simpson et al. 2009). Another alternative is provided by algorithms using a de Bruijn graph data structure (Fig. 16.3b). Here a reduction of sequence dataset complexity is achieved by extracting k-mers (a short sequence of length k, where k is a positive integer). Every node in a de Bruijn graph is a unique k-mer that occurs at least once in the input sequence dataset. An edge is always drawn between two nodes if the first node's suffix of length k-1 is equal to the second node's prefix of length k-1. Since each k-mer must only occur once in the graph, the number of nodes depends on the number of different k-mers in the genome but not on the number of input reads. This diminishes the dilemma of NGS to produce short reads and high coverage. However,

the problem to extract consensus contigs from a de Bruijn graph structure is far more complex than from an overlap graph. Tracking the number of occurrences of each k-mer in the input reads helps to resolve loops in the graph, since it is used to infer the number of occurrences of each k-mer in the consensus contigs.

Still the number of assemblers that can compute barley genome sized datasets is small. The commercial software CLC assembly cell (www.clcbio.com) is able to process such a dataset with less than 250 Gb of Memory (A standard personal computer is equipped with 4 Gb). Such an assembly of the barley genome has been made recently available providing direct access to most of the barley genes and for over 20,000 genes with transcript evidence also a genetic and physical position in the context of the barley genome physical map (The International Barley Genome Sequencing Consortium 2012). SOAP Denovo (v. 1.05) is an alternative assembler that can process barley data on a comparable computer up to a sequencing depth of 30 fold coverage (own unpublished data).

The repetitive nature of the barley genome causes the second challenge to sequence assembly since highly conserved multiple copies of repetitive DNA are a major cause of mis-assemblies. A mis-assembled sequence differs from the original DNA and mainly two classes of mis-assemblies occur. In the first case two pieces of sequence are tied together on the basis of partial sequence identity although they do not truly overlap physically. This may apply to two low copy sequences situated adjacently to highly conserved copies of the same class of TE. The second case results from collapsing two or more (almost) identical sub-sequences producing a shorter consensus sequence than in the original DNA. This situation may occur at highly conserved tandemly repeated sequences (i.e TE or genes). The main strategy of preventing repetitive DNA based mis-assemblies builds on the use of paired end sequencing. All advanced assemblers have implemented the inclusion of paired-end information for correct contig construction and most times also its incorporation for immediate scaffolding. Additionally paired-end data can be exploited to validate finished assemblies and detect mis-assemblies (Phillippy et al. 2008). These strategies could successfully be implemented in mammalian whole genome shotgun sequencing projects. The genome of giant panda (size: 2.6 Gb) was sequenced using Illumina including paired-end (and mate pair) libraries of varying insert sizes from 150 bp to 10 kb. Half of the assembled sequence could be covered with scaffolds (N50) larger than 1.3 Mb (Li et al. 2010). The assembly software ALLPATH-LG was introduced and tested on human genome shotgun data sequenced on Illumina (Gnerre et al. 2011). The software requires a specific variety of paired-end libraries reaching from overlapping read-pairs to large insert mate pairs. The N50 scaffold even reached a size of more than 11 Mb which is comparable to the quality previously obtained by Sanger-based shotgun sequencing of the human genome.

16.4 Outlook

Progress in barley genome analysis was accelerated by the availability of the different Next Generation Sequencing technology platforms. It could be demonstrated that sequencing the barley genome is feasible now due to the availability of NGS and the

decrease in sequencing costs coming along with. Access to a relatively good quality draft sequence of barley has been provided recently. How much further such draft sequences can be developed into high-quality reference sequences will be depending on, however, the introduction of a further generation of improved sequencing technology (Metzker 2010). Such platforms will allow single molecule, real time, very long read sequencing thus helping to overcome the limitations of sequence assembly due to high repetitive DNA content. Until then, the new knowledge base obtained recently for the barley genome by high-throughput NGS is providing novel opportunities and tools to the research community for addressing questions of Triticeae biology and performance much more efficiently.

References

Baird NA, Etter PD, Atwood TS et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376

Berkman P, Skarshewski A, Manoli S et al (2011a) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. Theor Appl Genet. in press

Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. Plant Biotechnol J 9:768–775

Chain PSG, Grafham DV, Fulton RS et al (2009) Genome project standards in a new era of sequencing. Science 326:236–237

Chevreux B, Wetter T, Suhei S (1999) Genome sequence assembly using signals and additional sequence information. Computer science and biology: proceedings of the German conference on bioinformatics (GCB), 99:45–56

Chutimanitsakun Y, Nipper R, Cuesta-Marcos A et al (2011) Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. BMC Genomics 12:4

Doležel J, Kubaláková M, Paux E et al (2007) Chromosome-based genomics in the cereals. Chromosome Res 15:51–66

Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Eversole K, Graner A, Stein N (2009) Wheat and barley genome sequencing. In: Feuillet C, Muehlbauer GJ (eds) Genetics and genomics of the Triticeae. Springer pp 713–742

Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88

Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12:257–269

Gnerre S, Maccallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108:1513–1518

Gore MA, Chia J-M, Elshire RJ et al (2009) A first-generation haplotype map of maize. Science 326:1115–1117

Hernandez P, Martis M, Dorado G et al (2011) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69:377–386

Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. Genome Res 18:839–846

Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068–1076

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800
Krattinger S, Wicker T, Keller B (2009) Map-based cloning of genes in Triticeae (wheat and barley). In: Feuillet C, Muehlbauer GJ (eds) Genetics and genomics of the Triticeae. Springer pp 337–357
Künzel G, Korzun L, Meister A (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. Genetics 154:397–412
Kurtz S, Narechania A, Stein J, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517
Li R, Fan W, Tian G et al (2010) The sequence and de novo assembly of the giant panda genome. Nature 463:311–317
Manninen I, Schulman A (1993) BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22:829–846
Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141
Mayer KFX, Taudien S, Martis M et al (2009) Gene content and virtual gene order of barley chromosome 1H. Plant Physiol 151:496–505
Mayer KFX, Martis M, Hedley P et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46
Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. Nat Protoc 3:267–278
Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327
Myers EW, Sutton GG, Delcher AL et al (2000) A whole-genome assembly of Drosophila. Science 287:2196–2204
Paterson AH, Bowers JE, Bruggmann R et al (2009) The sorghum bicolor genome and the diversification of grasses. Nature 457:551–556
Phillippy A, Schatz M, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 9:R55
Roach JC, Boysen C, Wang K, Hood L (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26:345–353
Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. Mol Ecol 20:3499–3502
Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467
Sato K, Motoi Y, Yamaji N, Yoshida H (2011) 454 sequencing of pooled BAC clones on chromosome 3H of barley. BMC Genomics 12:246
Schulte D, Close TJ, Graner A et al (2009) The international barley sequencing consortium–at the threshold of efficient access to the barley genome. Plant Physiol 149:142–147
Service RF (2006) Gene sequencing: the race for the $1000 genome. Science 311:1544–1546
Simpson JT, Durbin R (2011) Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22:549–556
Simpson J, Wong K, Jackman S et al (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123
Sorokin A, Marthe F, Houben A et al (1994) Polymerase chain reaction mediated localization of RFLP clones to microisolated translocation chromosomes of barley. Genome 37:550–555
Staden R (1980) A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res 8:3673–3694
Stein N (2007) Triticeae genomics: advances in sequence analysis of large genome cereal crops. Chromosome Res 15:21–31
Stein N, Graner A (2004) Map-based gene isolation in cereal genomes. In: Gupta P, Varshney R (eds) Cereal genomics. Kluwer Academic Publishers, Dordrecht, pp 331–360

Steuernagel B, Taudien S, Gundlach H et al (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. BMC Genomics 10:547

Taudien S, Steuernagel B, Ariyadasa R et al (2011) Sequencing of BAC pools by different next generation sequencing platforms and strategies. BMC Res Notes 4:411

The International Barley Genome Sequencing Consortium (IBSC) (2012) A physical, genetical and functional sequence assembly of the barley genome. Nature 491:711–716

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Vicient CM, Suoniemi A, Anamthawat-Jonsson K et al (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769–1784

Vitulo N, Albiero A, Forcato C et al (2011) First survey of the wheat chromosome 5A composition through a next generation sequencing approach. PLoS One 6:e26421

Wicker T, Schlagenhauf E, Graner A et al (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275

Wicker T, Narechania A, Sabot F et al (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9:518

Wicker T, Taudien S, Houben A et al (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Wicker T, Mayer KFX, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23:1706–1718

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang H, Sreenivasulu N, Weschke W et al (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. Plant J 40:276–290