Rice (2009) 2:35–43
DOI 10.1007/s12284-009-9025-z

De Novo Next Generation Sequencing of Plant Genomes

Steve Rounsley · Pradeep Reddy Marri · Yeisoo Yu · Ruifeng He · Nick Sisneros · Jose Luis Goicoechea · So Jeong Lee · Angelina Angelova · Dave Kudrna · Meizhong Luo · Jason Affourtit · Brian Desany · James Knight · Faheem Niazi · Michael Egholm · Rod A. Wing

Received: 26 December 2008 / Accepted: 16 February 2009 / Published online: 7 March 2009

Abstract The genome sequencing of all major food and bioenergy crops is of critical importance in the race to improve crop production to meet the future food and energy security needs of the world. Next generation sequencing technologies have brought about great improvements in sequencing throughput and cost, but do not yet allow for de novo sequencing of large repetitive genomes as found in most crop plants. We present a strategy that combines cutting edge next generation sequencing with "old school" genomics resources and allows rapid cost-effective sequencing of plant genomes.

Keywords Next generation sequencing · *Oryza* · Genome sequencing · Genome assembly

S. Rounsley (✉) · P. R. Marri · Y. Yu · R. He · N. Sisneros · J. L. Goicoechea · S. J. Lee · A. Angelova · D. Kudrna · M. Luo · R. A. Wing (✉)
BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA
e-mail: rounsley@email.arizona.edu
e-mail: rwing@ag.arizona.edu

Y. Yu · R. He · N. Sisneros · J. L. Goicoechea · S. J. Lee · A. Angelova · D. Kudrna · M. Luo · R. A. Wing
Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA

Y. Yu · R. He · N. Sisneros · J. L. Goicoechea · S. J. Lee · A. Angelova · D. Kudrna · M. Luo · R. A. Wing
Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

J. Affourtit · B. Desany · J. Knight · F. Niazi · M. Egholm
454 Life Sciences,
20 Commercial Street,
Branford, CT 06405, USA

Introduction

To meet the food and bioenergy security needs of the future, farmers must double or even triple crop yields on less land, with less water, on poorer soils, and with less pesticides. Information gained from sequenced genomes in crops, coupled with genetic association studies, will allow us to identify key genes/quantitative trait loci and networks that can lead to higher yielding crops that can grow in extreme conditions but with reduced environmental impact. We therefore need to develop efficient and cost-effective methods to sequence major food crop genomes and their wild relatives.

Plant genome sequencing has progressed rapidly since the first genome (*Arabidopsis thaliana*) was completed in 2000 (Arabidopsis Genome Initiative 2000). The 389-Mb rice genome was completed in 2004 (International Rice Genome Sequencing Project 2005), and recently, a draft sequence of the 2,300-Mb maize genome was released (Pennisi 2008). All were sequenced using "traditional" sequencing approaches in which sequencing libraries are constructed from individual segments of the genome (such as those contained within bacterial artificial chromosomes (BAC) clones) and are sequenced via gel electrophoresis and dideoxy terminator chemistry (i.e., "Sanger sequencing"). A whole-genome shotgun (WGS) strategy, made possible with improved assembly algorithms, has been used for several recent plant genomes (Tuskan et al. 2006; Jaillon et al. 2007; Paterson et al. 2009) in which the sequencing libraries are made directly from genomic DNA.

Recently, next generation sequencing (NGS) technologies have promised to further accelerate progress, with huge increases in sequencing throughput and, perhaps more importantly, the ability to avoid the handling of individual clones from shotgun libraries. There are currently four

commercially available NGS technologies: 454 Life Sciences (acquired by Roche), Solexa (acquired by Illumina), ABI SOLID (acquired from Agencourt Biosciences), and Helicos Biosciences. Although all have their specific features, generally, they can be grouped into two classes based on the lengths of the sequence reads produced. Solexa, ABI SOLID, and Helicos all produce very short reads in very large quantities, while the 454 platform can produce a more moderate amount of sequence, but with much longer read lengths. Several of the platforms have already gone through multiple rounds of upgraded specifications, and improvements are likely to continue.

Not unexpectedly, applications for these platforms have up until now been those better suited to the short read lengths they produce. These include resequencing reference genomes (Wheeler et al. 2008), de novo sequencing of small bacterial genomes (McCutcheon and Moran 2007), assessing microbial diversity (Sogin et al. 2006), and gene expression, small RNA, and methylation analyses (Lister et al. 2008).

However, de novo sequencing of large repetitive eukaryotic genomes has yet to be accomplished with a NGS platform and remains an important and significant goal. There are many species for which a genome sequence is still a valuable and desirable resource and that are too distant from a nearest sequenced relative to take advantage of resequencing strategies. In fact, the emerging picture of the "pan genome" (Morgante et al. 2007) suggests that even sequencing genomes within a species may benefit from a de novo approach rather than a resequencing approach. However, the challenge of de novo sequencing with larger genomes is that assembly becomes difficult as repeat content increases, and many larger genomes, particularly those of crop plants, have significant repetitive content. These assembly challenges sometimes impact even traditional sequencing approaches, but are particularly problematic with short read technologies.

Here, we present a strategy (Fig. 1) designed to use the 454 Life Sciences platform to sequence the ten genome types of the genus *Oryza*, which range in size from 350 to 1,285 Mb and includes the world's most important food crop—rice. We also present the results of a proof of principle experiment—the sequence of an entire chromosome arm of *Oryza barthii*. Key to our approach is reduction of assembly complexity by subdividing the genome into smaller units—not individual BAC clones as in the past, but pools of clones selected from a physical map of fingerprinted and end sequenced BAC clones (Kim et al. 2008). This has the multiple advantages of reducing assembly complexity, limiting the number of sequencing libraries required, and being able to take advantage of the large sequencing capacity of the 454 sequencing platform. We also take advantage of the latest version of the 454 platform called Titanium, which has improved read lengths. As implemented here, multiple pools of BAC clones from a minimum tiling path (MTP) across a region can be sequenced, assembled, and combined to generate a high quality draft sequence of a chromosome arm and, by extension, can be applied to a whole genome in a straightforward manner.

Results

Physical map and MTP

As a pilot experiment, we selected an MTP of 166 BAC clones across the short arm of chromosome 3 of *O. barthii*, the progenitor of the West African cultivated rice—*Oryza glaberrima*. The physical map had previously been generated with fingerprinted contigs (FPC; Soderlund et al. 2000) and SyMAP (Soderlund et al. 2006) using SNaPshot fingerprints and BAC end sequencing (BES) of a ~10× BAC library from *O. barthii* accession #105608 (see supplementary materials for further data on the BAC library). When aligned to the *Oryza sativa*, reference genome via available BES, the selected MTP spanned over 19.4 Mb of the reference genome with an average predicted clone size of 159 kb, and an average predicted overlap between BACs of 42 kb.

Sequencing

Two distinct sets of sequence data were collected. (a) Single-end reads using the Roche/454 extra-long read Titanium platform generated independently from each pool. The combined collection of Titanium reads contained 1,071,556 reads and 393,338,082 bp with an average read length of 367 bp. Size distributions for the Titanium reads are shown in Fig. 2a. Individual pools had between 53 and 78 Mb of raw Titanium data generated. (b) Long read paired ends using the Roche/454 FLX platform. Paired-end libraries were generated independently from each pool, but three pools (1–3 and 4–6) were combined for sequencing. Overall, 348,926 reads were generated for a total of 85.6 Mb. Subsequently, generated pool-specific datasets contained between 10.6 and 15.4 Mb of sequence. Subsequent analysis of paired-end data showed that the paired-end libraries had a very consistent insert size between pools—with mean insert sizes ranging from 3,132 to 3,164 bp across the six pools (Fig. 2b).

Assembly phases

Assembly of sequence data proceeded in four steps. First, the data were preprocessed to maximize the quality of the

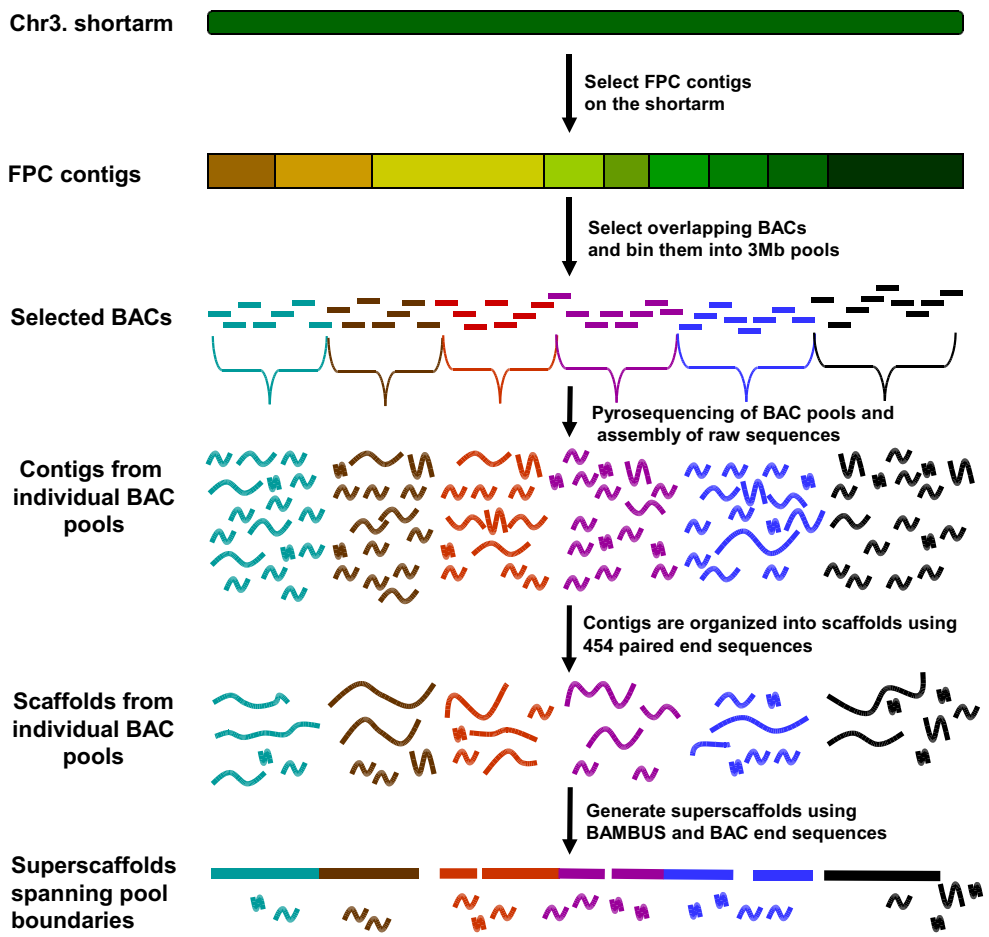

Fig. 1 Schematic view of the sequencing strategy. Based on the *O. barthii* physical map, nine FPC contigs span the ~18-Mb short arm of Chr3. A minimum tiling path of BACs across the FPC contigs was divided into six regions (each approximately 3-Mb) from which six pools of BACs were created. Each pool was sequenced and assembled to create contigs and scaffolds. Superscaffolds across pool boundaries were generated using the BAC end sequence information.

data provided to the assembler. This involved prefiltering the paired-end reads to remove those that did not contain both ends of the pair and screening for *Escherichia coli* and BAC vector contamination in both Titanium and paired-end reads. After preprocessing, between 15.1× (pool 2) and 22.2× (pool 4), coverage of sequence (paired-end and Titanium combined) was available for assembly. Physical coverage of paired ends was between 13.9× (pool 2) and 24.3× (pool 5). Second, the preprocessed data for each individual pool were assembled by the Newbler assembler from 454/Roche. The contiguity of these assemblies varied from pool to pool with contig N50 ranging from 10.8 (pool 6) to 19.9 kb (pool 1), and scaffold N50 ranging from 242 (pool 5) to 518 kb (pool 1). A clear trend of declining assembly metrics was seen as proximity to the centromere increased.

Since regions between neighboring pools overlapped, we implemented a third step to the assembly protocol in order to make these regions nonredundant and to improve their quality. This third step entailed comparing the assemblies from neighboring pools and identifying the contigs and underlying reads in the overlapping regions. These reads from both neighboring pools were combined and reassembled with Newbler. As can be seen from Tables 1 and 2, this step resulted in an overall reduction of 207 kb in total contig length for the six pools and an increase in both contig N50 and scaffold N50.

Finally, to take advantage of the available BAC end sequences and to create scaffolds that could span across pools, we performed additional scaffolding using the BAMBUS scaffolding software (Pop et al. 2004). The resulting final assembly had a total length of 18.4 Mb and was composed of just 44 scaffolds. Two facts are of particular note. First, 90% of the chromosome arm was contained in just six scaffolds, with the largest spanning more than a third of the arm (6.6 Mb). Secondly, the scaffolding with BAMBUS led to a nearly ninefold increase in the scaffold N50. Detailed assembly statistics are provided in Tables 1 and 2.

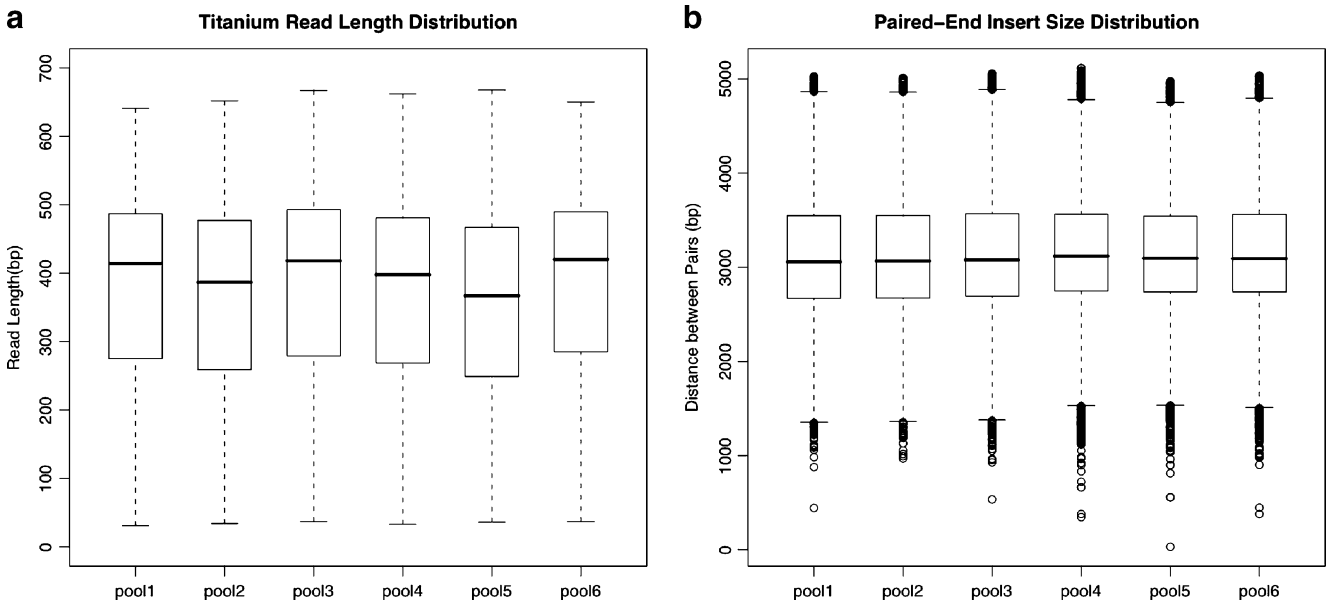

Fig. 2 *Box plots* showing size distributions for **a** Titanium shotgun reads and **b** distances between FLX paired-end reads for each of the six BAC pools. The *box* for each pool represents the interquartile range, and the *dark horizontal line with the box* represents the median.

Pooling assessment

The success of a pooling strategy depends on the quality of the pools that are constructed—it is important that each BAC be present in sufficient quantity to be sequenced effectively and that no individual BAC dominates the pool. We assessed the level of coverage of individual BAC clones in the pools by mapping of individual reads onto the assembled chromosome arm. The BAC end sequences of the pooled BACs allowed the definition of overlap and nonoverlap regions. Considering read mapping results in just the nonoverlapping regions, it was clear that there was substantial variability in the relative amounts of each BAC clone in the pools (see Table 3). This was likely due to the crude quantification employed prior to pooling the BAC DNA. Despite these wide ranges, no BAC had less than 4.5× coverage and contributed to the overall impressive assembly statistics. We are currently testing more precise quantification methods for construction of the BAC pools and this may have a positive effect upon the assembly.

Accuracy assessment

Since *O. barthii* has not been previously sequenced, we assessed the quality of our assembly via two independent approaches. Firstly, on the macroscale, an alignment of

Table 1 Assembly Statistics (Commonly Reported Assembly Statistics for Each of the Six BAC Pools)

	Individual pool statistics					
	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6
# Scaffolds	15	25	22	9	21	31
N50 scaffold size	518,306	353,691	309,466	407,073	242,868	360,621
Largest scaffold	692,300	727,204	628,772	1,195,083	728,801	769,231
# Contigs in scaffolds	251	328	290	290	361	346
# Contigs	349	483	484	369	506	527
N50 contig size	19,948	12,338	15,511	14,645	11,029	10,877
Largest contig	64,379	61,596	76,790	62,004	45,971	54,004
# Titanium reads used in assembly	171,538	134,113	164,319	185,005	133,321	142,414
# Paired ends used in assembly	17,191	13,694	14,713	20,551	23,047	19,544
Total base pair used	62,526,876	46,881,469	60,814,619	66,641,838	47,621,595	53,998,008
Total size	3,422,263	3,100,984	3,354,957	2,990,942	2,978,383	2,786,664

Contig or scaffold N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value

Table 2 Assembly Statistics (Commonly Reported Assembly Statistics for the Combined Data Covering the Entire Chromosome Arm)

	Entire chromosome arm		
	Stage 1: initial assembly	Stage 2: merge overlaps between pools	Stage 3: BAMBUS scaffolding
Number scaffolds	123	119	44
N50 scaffold size	360,621	370,878	3,165,134
Largest scaffold	1,195,083	1,195,083	6,645,942
Total scaffold length	18,635,055	18,401,654	18,496,608
Number contigs	2,718	2,677	
N50 contig size	14,195	14,307	
Largest contig	76,790	76,903	
Total contig length	18,634,193	18,427,025	

the assembled *O. barthii* chromosome arm against the reference genome sequence (RefSeq) of cultivated Asian rice, *O. sativa* ssp. *japonica* (International Rice Genome Sequencing Project 2005) shows no large rearrangements (Fig. 3). In fact, the most visible difference in this alignment is due to a physical gap in the rice RefSeq. Secondly, we assessed nucleotide accuracy by comparing those regions of *O. barthii* that lay in the overlapping regions between neighboring pools and were therefore sequenced twice. For each pair of neighboring pools, contigs from the initial Newbler assembler (prior to combining overlaps) were aligned to each other using BLASTN (Altschul et al. 1990). A total of 42 alignments were obtained from four overlaps (pools 5 and 6 have a small gap between them). Of 214,166 bp of aligned sequence, the total number of non-matching nucleotides was 95. If one assumes that for each position of disagreement between two pools, one is correct, then the overall nucleotide accuracy for the overlap regions is 99.978% or 2.2 errors for every 10 kb sequenced—very close to the 1 in 10,000 Bermuda standard for accuracy of finished sequence (http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml). Further analysis of the sequencing errors revealed that 26% are nucleotide mismatches, and 74% are indels, and, as expected, the majority (69%) of the latter ones are in homopolymer runs of 5 bp or greater. There is also a bias in base composition of the homopolymer run errors with 50% occurring in polyT runs.

Table 3 Pooling Efficiency

Pool	Mean coverage in pool	Lowest coverage in pool	Highest coverage in pool	Standard deviation
Pool 1	15.9×	5.4×	41.7×	7
Pool 2	12.1×	4.5×	27.1×	5.4
Pool 3	14.5×	6.8×	24.1	4.2
Pool 4	16.8×	6.9×	37.9×	6.8
Pool 5	11.7×	6.5×	20.7×	3.2
Pool 6	13.8×	5.3×	23.8×	4.9

Data summarizing the range of coverage obtained from individual BACs within each of the six pools. Assembly was high despite these large ranges of coverage

Discussion

Strategies for genome sequencing have generally taken one of two approaches: WGS (Adams et al. 2000; Venter et al. 2001) or hierarchical clone-by-clone sequencing (AGI 2000; IRGSP 2005). For the major plant genomes sequenced to date, both strategies have been employed. The genomes for *A. thaliana* (AGI 2000), *O. sativa* ssp. *japonica* (IRGSP 2005), and, most recently, *Zea mays* (Pennisi 2008) were generated by a clone-by-clone approach using BAC clones. Conversely, others including *O. sativa* ssp. *indica* (Yu et al. 2002), *Populus trichocarpa* (Tuskan et al. 2006), *Vitis vinifera* (Jaillon et al. 2007), and *Sorghum* (Paterson et al. 2009) have been sequenced by WGS. Generally, the advantages of WGS are that it needs fewer sequencing libraries, has a simpler workflow, and is therefore faster while clone-by-clone approaches are usually preferred when the genome is sufficiently large and complex that WGS assembly is problematic. With the advent of NGS technologies, the ease of generating large quantities of sequence has made WGS strategies even more appealing. However, the reality is that the short read lengths produced by early generations of NGS platforms have limited their application in de novo genome sequencing to smaller genomes—primarily bacteria.

The primary reason for this is the increasing abundance of repetitive sequence in larger genomes. Prior to the arrival of NGS platforms, WGS on larger genomes was possible because Sanger-based sequencing platforms generated paired-end reads of 700 bp or greater and a range of insert sizes from 3 to 40 kb (Weber and Myers 1997). This

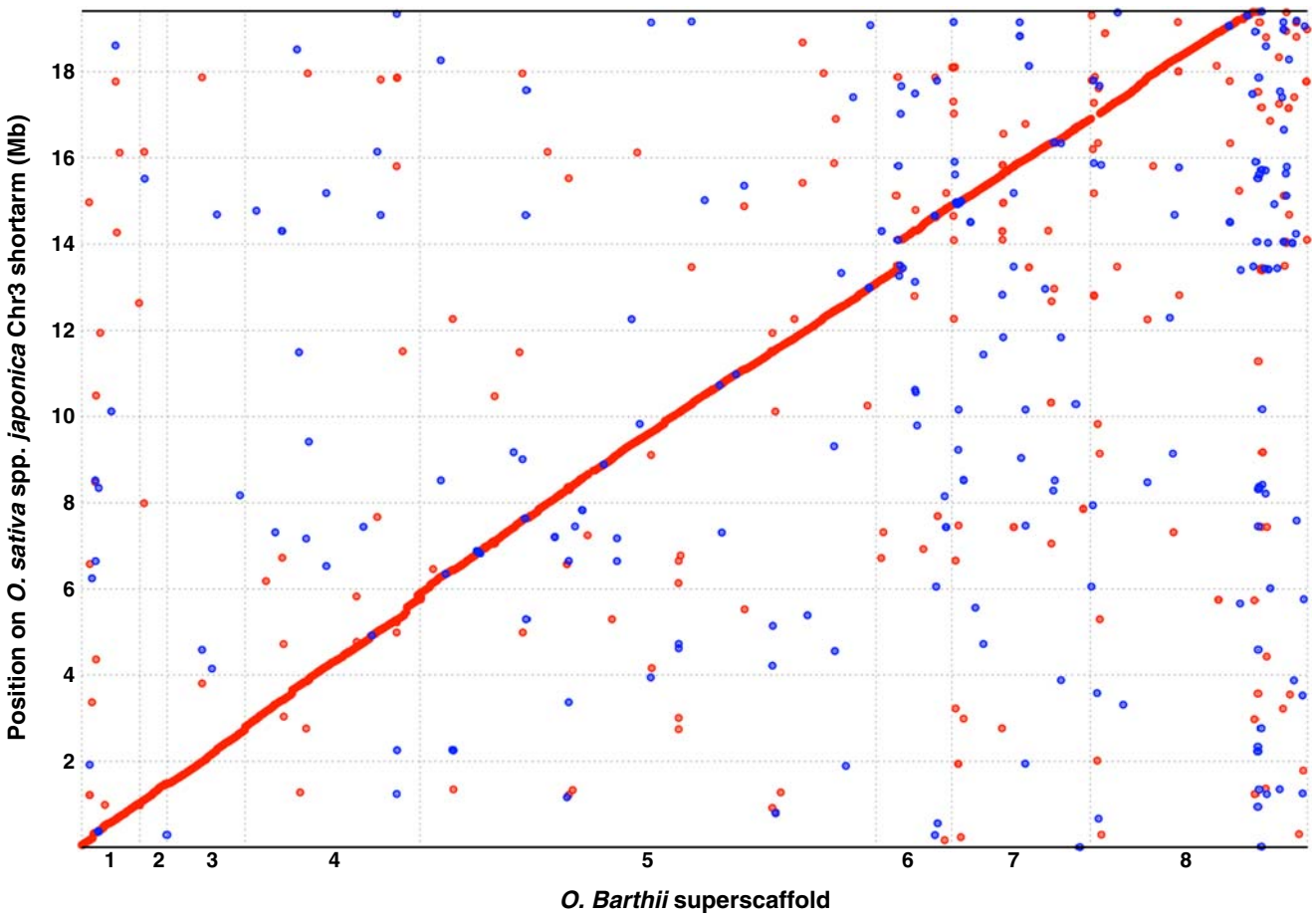

Fig. 3 *O. barthii* vs *O. sativa* alignment. A *dot plot* showing the alignment of the *O. barthii* scaffolds against the *O. sativa* ssp. *japonica* RefSeq (build 4.0) for the short arm of chromosome 3. The apparent gap in scaffold 6 represents a physical gap in *japonica* that contains 500,000 N. The *dot plot* was generated by the program MUMmerplot (Delcher et al. 2002).

combination of read length and paired reads spanning quite large distances can be used effectively by assembly algorithms to resolve many repeats and reconstruct a draft genome sequence (Myers et al. 2000; Batzoglou et al. 2002; Jaffe et al. 2003). However, current NGS platforms are unable to deliver both these features and thus cannot effectively span repeats.

In the absence of a realistic WGS strategy with NGS platforms, the alternative is a clone-based approach. The traditional single clone-by-clone approach has the advantage of reduced assembly complexity, but the disadvantage of requiring large numbers of sequencing libraries and all of the logistical challenges associated with such a large-scale effort. Pooling multiple BACs together reduces the number of libraries required and, with reasonable size pools, keeps assembly complexity sufficiently manageable.

There are several significant factors that contributed to the success of this strategy in *O. barthii*. The first is the availability of high quality physical map contigs from which we could select contiguous pools of BAC clones. The second is the availability of the Titanium platform from 454 Life Sciences. The increased read lengths that this platform produces, compared to its predecessor (GS FLX), are invaluable in producing a high quality assembly—particularly when combined with paired-end reads from the GS FLX platform. Recently, there was a report of unsatisfactory results from attempts at pooling as few as eight BAC clones from the salmon genome for sequencing with the 454 GS FLX platform (Quinn et al. 2008). It seems likely that the major factor in the difference between our contrasting experiences with BAC pooling is the increased read lengths we were able to obtain with the newer platform. Lastly, the use of BAC end sequences to combine multiple pools together allowed our sequence scaffolds to grow larger than our initial BAC pools.

Even with these positive factors in place, it is possible that other factors may negatively impact the success of this strategy. The most obvious one is the nature of the

repetitive sequence in a given pool. This not only will certainly vary between species (and could have contributed to the experience with the salmon genome) but will also vary across a given genome, as more repetitive regions such as centromeres are approached or traversed. In fact, our data demonstrate the very effect with pool 6 (closest to the centromere) having a contig N50 of only 54% of that of pool 1 (furthest from the centromere) and more than twice the number of scaffolds.

Given the factors involved, selection of pool size requires compromise between assembly quality and sequencing costs/logistics. The larger the pool, the more efficient the sequencing but the lower the expected quality of the assembly. For our proof of principle experiment, we selected 3-Mb pools after we had initial success with 1.5-Mb pools (data not shown). However, it is very possible that even larger pools could provide adequate assembly quality for many needs. For example, preliminary data shows that combining data from all 6.3 Mb pools together for assembly gives only a 35% decrease in contig N50 to 9.3 kb and a 55% decrease in scaffold N50 to 161 kb. Given the expected variation in the content and nature of repeats around a genome and between species, it would be useful to be able to predict repeat content prior to deciding upon a pool size. In fact, with a physical map and BAC end sequences in hand, it may be possible to predict such regions and modify the pool size appropriately.

Two other aspects of our method are important to emphasize. Firstly, with the sole exception of physical map editing in the first step, our results presented here did not utilize the *O. sativa* RefSeq in any way other than in assessment of results. Thus, the method is a true de novo assembly and could easily be applied to any species—no matter what related genomes have been sequenced. Secondly, the quality of the sequence obtained with this strategy should be emphasized. Our results clearly show that overall nucleotide accuracy approaches that of the Bermuda standard for accuracy of finished sequence—(i.e., 1 error in 10 kb). The results also show that the assembled sequence has very good contiguity—partly not only due to successful assembly but also due to a feature inherent in the BAC pooling strategy—that every pool is from a defined region of the genome. However, it is also important to recognize that there are diverse uses for a genome sequence, and many researchers do not require a high standard for both accuracy and contiguity. However, plant researchers, due to the nature of their work, often place a premium on knowing where in the genome a sequence resides. This is critical for trait mapping and breeding applications. For this reason, the genome sequencing strategy we present here is of particular utility to plant genomes.

To summarize, we utilized a fusion of "old school" physical mapping with ultrahigh throughput pyrosequencing to sequence an ~18-Mb chromosome arm in a single experiment. The resulting sequence of the chromosome arm has contiguity and accuracy that rivals any genome sequencing approach. In addition to being useful for targeting specific regions of a genome, scaling of this approach could generate the genome of all 12 chromosomes (~400 Mb) of *O. barthii* in a month or two. Future improvements in NGS technologies may continue to reduce the time and cost of such projects, but this strategy will likely always be valuable for addressing the complexities of sequencing large repetitive genomes with NGS. It provides an immediate and practical solution to the rapid generation of genome sequences from large and complex eukaryotic genomes for accelerated biological discovery to address some of our most critical food and bioenergy challenges.

Methods

Generation of a BAC library for *O. barthii*

Seed from *O. barthii* accession #105608 was obtained from the International Rice Research Institute, Los Banos, Philippines. Megabase-size DNA, isolated from nuclei from young 3-week-old seedling tissue, was partially digested with *Hin*dIII, and size-selected fragments were ligated into pAGIBAC1 for BAC library construction, using previously described protocols (Luo and Wing 2003; Ammiraju et al. 2006). The resultant BAC library (OB_ABa) contained 36,864 clones with an average insert size of 136 kb and represented about ten times coverage of the 420-Mb *O. barthii* genome. The BAC library was deposited in the AGI BAC/EST Resource Center (www.genome.arizona.edu/orders) and is publically available.

Generation of a minimum tiling path

An *O. barthii* chr3 short-arm MTP was selected from a whole genome physical map that was generated by SNaP-shot fingerprinting and BAC end sequencing of a ~10× BAC library. The fingerprints were assembled into contigs using FPC (Soderlund et al. 2000). FPC contigs were aligned to the rice RefSeq (IRGSP 2005) using SyMAP (Soderlund et al. 2006) and adjacent contigs were merged when warranted. The MTP consisted of 166 BACs with eight gaps between FPC.

Sequencing library preparation and sequencing

DNA from each BAC clone was isolated, sheared to a size of 2–6 kb (HydroShear), purified, and quantified. Equal microgram amounts of DNA from the 168 BACs were then divided into six pools of 28 sequential BACs to form pools

1 through 5 and 26 BACs for pool 6. Each DNA pool was individually processed using 454/Roche Titanium shotgun and standard long paired-end sequencing kits and sequenced on 454/Roche GSFLX sequencers according to manufacturers specifications. Each pool was sequenced individually for the Titanium single-end read data, but three pools were combined for the paired-end read data. Pool-specific paired-end datasets were created by comparison to the pools-specific Titanium datasets using BLASTN and assigning each read to an individual pool.

Sequence assembly

Assembly was performed in four stages: (1) data preprocessing. Paired-end reads were prefiltered so that only reads containing both ends of the pair were included. (2) 454/Roche Titanium shotgun and standard long paired-end reads for each pool were combined and assembled using the 454 Newbler assembler (v2.0.00.10), after screening for *E. coli* and BAC vector sequence. (3) Contigs in overlap regions between neighboring pools were identified from both pools via alignments with MUMMER version 3.20 (Delcher et al. 2002). Underlying reads for these contigs were then combined and reassembled to remove duplication in overlap regions. Alignment against the rice RefSeq identified one misassembled contig caused by a single problematic read. This pool was reassembled after removal of the problem read. A composite assembly was constructed of the scaffolds and unscaffolded contigs from all six pools. (4) The BAMBUS scaffolding software (Pop et al. 2004) was then used to construct superscaffolds across pool boundaries using paired BAC end sequences from *O. barthii* combined with the above composite assembly.

Assembly analysis

Assembly accuracy of the initial 454 Newbler assembly (stage 1) was assessed by comparing the overlap regions of neighboring pools using BLASTN (Altschul et al. 1990). Mismatches and gaps were counted, and the latter gaps were categorized for the nature of the underlying sequence. Since the overlap regions were independently sequenced from different substrates (overlapping BAC clones), non-matching sequence could be of biological origin (e.g., mutation in the BAC clone, or heterozygous plant material) or could be due to sequencing or assembly error. For this reason, these accuracy estimates should be considered a minimum accuracy expected from this sequencing approach.

Sequence submission to NCBI

Assembled contigs and scaffolds for each of six 3-Mb pools, as well as the entire ~18.4-Mb *O. barthii* chromosome 3 short arm superscaffold have been deposited at DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank under the project accession ABRL00000000.

Acknowledgements This work was supported by National Science Foundation grants DBI-0638541 (to RAW and SR), DBI-0321678 (to RAW), and the Bud Antle Endowed Chair (to RAW)

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000;287(5461):2185–95.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

Ammiraju JS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, et al. The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus oryza. Genome Res. 2006;16(1):140–147.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408(6814):796–815.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002;12(1):177–89.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30(11):2478–83.

International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436(7052):793–800.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 2003;13(1):91–6.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463–7. Nature.

Kim H, Hurwitz B, Yu Y, Collura K, Gill N, SanMiguel P, et al. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza. Genome Biol. 2008;9(2):R45.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523–36.

Luo M, Wing RA. An improved method for plant BAC library construction. Methods Mol Biol. 2003;236:3–20.

McCutcheon JP, Moran NA. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A. 2007;104(49):19392–7.

Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol. 2007;10(2):149–55.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287(5461):2196–204.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457:551–6.

Pennisi E. Plant sciences. Corn genomics pops wide open. Science. 2008;319(5868):1333.

Pop M, Kosack D, Salzberg SL. A hierarchical approach to building contig scaffolds. Genome Res. 2004;14(1):149–59.

Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, et al. Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics. 2008;9:404.

Soderlund C, Humphray S, Dunham A, French L. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 2000;10 (11):1772–87.

Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: a system for discovering and viewing syntenic regions of FPC maps. Genome Res. 2006;16(9):1159–68.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A. 2006;103 (32):12115–20.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313(5793):1596–604.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291 (5507):1304–51.

Weber JL, Myers EW. Human whole-genome shotgun sequencing. Genome Res. 1997;7(5):401–9.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872–6.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science. 2002;296 (5565):79–92.