the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **77,** 795–805 doi: 10.1111/tpj.12412

TECHNICAL ADVANCE

Comparative BAC-based physical mapping of *Oryza sativa* ssp. *indica* var. 93–11 and evaluation of the two rice reference sequence assemblies

Yonglong Pan[1], Ying Deng[1], Haiyan Lin[1], David A. Kudrna[2], Rod A. Wing[2], Lijia Li[3], Qifa Zhang[1] and Meizhong Luo[1,*]
[1]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China,*
[2]*Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA, and*
[3]*State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan 430072, China*

Received 20 April 2013; revised 3 December 2013; accepted 9 December 2013; published online 14 December 2013.
*For correspondence (e-mail mzluo@mail.hzau.edu.cn).

SUMMARY

Reference sequences are sequences that are used for public consultation, and therefore must be of high quality. Using the whole-genome shotgun/next-generation sequencing approach, many genome sequences of complex higher plants have been generated in recent years, and are generally considered reference sequences. However, none of these sequences has been experimentally evaluated at the whole-genome sequence assembly level. Rice has a relatively simple plant genome, and the genome sequences for its two sub-species obtained using different sequencing approaches were published approximately 10 years ago. This provides a unique system for a case study to evaluate the qualities and utilities of published plant genome sequences. We constructed a robust BAC physical map embedding a large number of BAC end sequences forrice variety 93–11. Through BAC end sequence alignments and tri-assembly comparisons of the 93–11 physical map and the two reference sequences, we found that the Nipponbare reference sequence generated using the clone-by-clone approach has a high quality but still contains small artifact inversions and missing sequences. In contrast, the 93–11 reference sequence generated using the whole-genome shotgun approach contains many large and varied assembly errors, such as inversions, duplications and translocations, as well as missing sequences. The 93–11 physical map provides an invaluable resource for evaluation and improvements toward completion of both Nipponbare and 93–11 reference sequences.

Keywords: BAC, physical map, rice, Nipponbare, 93–11, reference sequence, technical advance.

INTRODUCTION

Reference genome sequences are the bedrock for functional and comparative genomics studies. A high-quality reference sequence is essential to fully understand the biology of the organism that the reference sequence represents. An error-filled and fragmented genome sequence may not be effectively used as a reference sequence because it does not provide reliable and complete sequence information for users and may lead to incorrect conclusions and mis-guided follow-up experimentation. If error-filled genome sequences are used for multiple genome comparisons across relatively long phylogenetic distances, such as in the *Oryza* genus or cereal species, the errors may be amplified and the true genetic divergence may be mistaken. Reference sequences have been used to guide the assembly of newly sequenced genomes with lower coverage and short sequence reads (Schneeberger *et al.*, 2011) or to identify missing sequences and insertions in individual genomes (Kidd *et al.*, 2010). In these cases, the accuracy and completeness of the trusted reference sequences are critical for determining the outcomes. In addition, gaps in reference sequences have been shown to contain functional elements and important genes (Ammiraju *et al.*, 2005; Minocherhomji *et al.*, 2012). Church *et al.* (2009) reported that a comprehensive understanding of biology in the mouse (*Mus musculus*) is only possible with the availability of a finished, high-quality genome

sequence assembly. Recently, it was reported that 80% of the human genome encodes functions (ENCODE Project Consortium, 2012). However, although desired, a perfect reference genome sequence is difficult to obtain (Church *et al.*, 2009, 2011; Dolgin, 2009; Lewin *et al.*, 2009; Alkan *et al.*, 2011; Hamilton and Buell, 2012). The human reference genome sequence has the highest quality of all the mammalian genome sequences but still contains many errors and gaps (Church *et al.*, 2011). An international Genome Reference Consortium has been formed to complete the human, mouse and zebrafish (*Danio rerio*) genome sequences (Church *et al.*, 2011).

Rice is a staple food crop worldwide, and feeds over half of the world's population. Due to a compact genome, the availability of dense genetic maps and high transformation efficiency, rice represents an excellent model for functional genomics of grass species. The Asian cultivated rice *Oryza sativa* has two major sub-species: *japonica* and *indica*. The two sub-species have different origins (Kovach *et al.*, 2007; Huang *et al.*, 2012), display distinct divergence in important agronomic traits, and are maintained through a 'killer and protector' mechanism that regulates hybrid embryo survival (Yang *et al.*, 2012). Genome sequences for the *japonica* representative variety, Nipponbare, and the *indica* representative variety, 93–11 genome sequence finished by Beijing Institute of Genome (93–11 BGI), were obtained using different approaches and are recognized as rice reference sequences.

Sequencing of the Nipponbare genome utilized a map-based, clone-by-clone approach (primarily using BAC clones), similar to the human genome, and required an international effort over 11 countries. In this approach, a robust BAC physical map was first constructed and anchored to a dense genetic map (over 2000 mapped markers) to provide high accuracy for location of the BAC contigs (Harushima *et al.*, 1998; Chen *et al.*, 2002). Then, a minimal tile of BAC clones was individually sequenced, finished and assembled. Finally, the whole genome sequence was assembled from the completed individual BAC sequences (International Rice Genome Sequencing Project, 2005). The resulting Nipponbare reference sequence is of high quality and is considered a 'gold standard'. The Nipponbare reference sequence and the sequence-tagged BAC clones have been widely used as a rice research platform. In contrast, sequencing of the 93–11 genome used a whole-genome shotgun (WGS) approach. In this approach, the same Sanger sequencing method was used as for sequencing of the Nipponbare genome, but no BAC library and physical map were used, and the whole genome sequence was assembled directly from global sequence reads, such that the resulting 93–11 BGI has a relatively lower quality (International Rice Genome Sequencing Project, 2005; Yu *et al.*, 2006). Although the 93–11 BGI has played important roles in some fields such as providing SNP marker candidates (Feltus *et al.*, 2004; Huang *et al.*, 2009), it appears not to be fully accepted by the community, probably because of both the lower quality, which imposes an extra burden on users to distinguish errors and complete the sequences, and because of a lack of sequence-tagged clones that are required to confirm the sequence and to perform functional complementation. 93–11 is a parent of the popular super-hybrid rice, LYP9, and an elite representative of *indica* varieties, which account for more than 70% of the world's rice production. Obtaining a high-quality 93–11 genome sequence as a reliable *indica* reference sequence is imperative for world food security.

Although the Nipponbare reference sequence has a high quality, it still contains several tens of physical gaps and many assembly errors (Ammiraju *et al.*, 2005; International Rice Genome Sequencing Project, 2005; Yu *et al.*, 2006; Lin *et al.*, 2012). Previously, we constructed a fivefold coverage BAC physical map for *japonica* rice variety ZH11, and showed that a BAC physical map with BAC end sequences (BESs) was an effective tool to detect structural variations on the reference sequences (Lin *et al.*, 2012). Using BES alignments, we assigned four ZH11 contigs that spanned four Nipponbare gaps. Forty-six inversely matched BESs (termed abnormally oriented BESs by Lin *et al.*, 2012) flagged 17 Nipponbare reference sequence locations that most likely contained spurious inversions (i.e. inversions by assembly errors). Of the 17 locations, 10 were inversely matched by two or more BESs, and thus their detection should be more reliable. One such location was confirmed to be a spurious inversion by Yu *et al.* (2006). When another three such locations were further analyzed by our group, they were all confirmed to be spurious inversions (Lin *et al.*, 2012). To date, the widely used early version (IRGSP build 4) has been updated twice [build 5 (http://rgp.dna.affrc.go.jp) and Os Nipponbare Reference IRGSP–1.0 (referred to here as IRGSP1.0, http://rapdb.dna.affrc.go.jp/)]. In the latest update of IRGSP1.0, the sequence was improved by a revision that used optical map data and by whole-genome re-sequencing using next-generation sequencing (NGS; Kawahara *et al.*, 2013).

Although rice is important for world food security, and its genome sequence is therefore as important as that of human, no similar consortium to the Genome Reference Consortium has been proposed. Detecting more assembly errors and finally completing the Nipponbare and 93–11 reference sequences will require additional resources and efforts. Genome sequence comparisons between the *japonica* rice variety Nipponbare and *indica* rice variety 93–11 will provide basic information to link the DNA sequences to the phenotypes and evolution of the two sub-species. However, reliable sequence-based information may be obtained only if both genome sequences are of high quality. Moreover, the two rice reference sequences were obtained using different approaches, and

the feasibility of the WGS approach to produce a high-quality genome sequence has been debated (International Rice Genome Sequencing Project, 2005; Yu *et al.*, 2006). Distinguishing actual sequence variations from artifacts and evaluating sequence qualities between the Nipponbare and 93–11 sequences requires a third independent assembly as a reference. Furthermore, many genome sequences of complex higher plants have been published in recent years, with a rapid increase in those generated using the WGS/NGS approach, and these are generally considered reference sequences. However, none of these reference sequences was experimentally evaluated at the whole-genome sequence assembly level.

Rice has a relatively simple genome among plants. The two reference sequences were generated using the best sequencing method (Sanger method), and have been available for approximately 10 years. This provides a unique system for a case study to evaluate the qualities and utilities of the published plant genome sequences. For this purpose, we constructed a robust BAC physical map embedding a large number of BESs for 93–11, and we compared the 93–11 physical map with the two reference sequences at the structural level through BES alignments. This study provides not only a reference and invaluable resource for evaluation and improvements toward completion of both the Nipponbare and 93–11 reference sequences, but also a general picture of the overall completeness of many other plant genome sequences when obtained through the WGS/NGS approach.

RESULTS

Construction of a BAC physical map for 93–11

To evaluate and improve the quality of the Nipponbare and 93–11 BGI reference sequences, we generated a BAC-based physical map of the 93–11 genome. Briefly, an 8.8-fold 93–11 BAC library was constructed (36 864 clones, mean insert size 113 kb), sequenced at BAC ends (65 014 qualified sequences with quality value >16, length ≥100 bp) and fingerprinted using SNaPshot (Luo *et al.*, 2003; 32 205 of hight-quality fingerprints, bands/clone = 112 ± 19.56). The fingerprints were assembled into a physical map using FPC (Soderlund *et al.*, 1997), which resulted in a phase I assembly of 659 contigs and 600 singleton clones. The mean BAC clone number per contig was 48.87, with a range of 2–468. The mean size per contig was 517 consensus bands (CBs), with a range of 68–4148 CBs, which, based on the mean size of 1009 bp per CB unit (mean insert size of clones/mean CBs of clones, 113 kb/112 CBs), corresponded to a mean of 521.6 kb, with a range of 68.6 kb–4.18 Mb. The cumulative length of the phase I FPC assembly was approximately 344 Mb (340 880 CBs).

The 93–11 phase I physical map was manually edited as described previously (Kim *et al.*, 2007; Lin *et al.*, 2012), which resulted in a phase II FPC assembly that consisted of 287 contigs (30 938 clones) and 1267 singletons. The cumulative length of the 93–11 phase II FPC assembly was approximately 303 Mb (300 122 CB units). The 93–11 phase II FPC assembly was named '93–11 PM' to distinguish it from the '93–11 BGI' reference sequence.

Alignment of 93–11 PM to the Nipponbare and 93–11 BGI reference sequences

On comparative maps aligned by BESs, the links appear like steps on a ladder. Putative assembly discrepancies may be detected if BES alignments are not parallel. To detect and visualize putative assembly discrepancies between 93–11 PM and the two reference sequences, we aligned the 93–11 PM contigs with the BESs whose repeat- and organelle sequences were masked to the three versions of the Nipponbare reference sequence and to 93–11 BGI using BLAT (Kent, 2002), and displayed the alignments using SYMAP software (Soderlund *et al.*, 2006). Scalable figures and detailed information for all local alignments are available at http://gresource.hzau.edu.cn/resource/symap_93–11.html. We primarily report the results for comparison with the latest version IRGSP1.0 in this paper because the discrepancies detected in IRGSP1.0 are those that remain from the early versions. We compared the three versions for large differences (clone-size level). Ten differences were found between build 4 and build 5, including nine inversion corrections and one location change. Four differences were found between build 5 and IRGSP1.0, including three inversion corrections and one location change (Figure S1). Improvements between build 4 and IRGSP1.0 at the sequence level were not assessed. IRGSP1.0 corrected many sequence errors (Kawahara *et al.*, 2013). However, our results show that improvement at the structural level in IRGSP1.0 as a result of whole-genome re-sequencing using NGS is limited.

Table 1 shows a summary of BES mapping results for comparison of 93–11 PM to IRGSP1.0 and 93–11 BGI. For IRGSP1.0, BLAT matched 33 659 BESs (51.69% of the total

Table 1 Alignment results for the clones and BESs of the 93–11 PM contigs to the Nipponbare and 93–11 BGI reference sequences

93–11 PM	IRGSP1.0	93–11 BGI
BES with hits	33 659	35 804
Clones in anchored contigs	30 834	30 790
Without BES hits	6923	8479
With BES hits	23 911	22 311
With paired-end BES hits	9748	9116
With single-end BES hits	14 163	13 195
With inversely matched BES hits	1589	1806
With normally matched BES hits	22 322	20 505
Clones in unanchored contigs	103	148
Anchored contigs	274	277
Unanchored contigs	13	10

93–11 BESs), and SYMAP anchored 274 93–11 PM contigs comprising 30 834 BAC clones spanning a total length of 297 713 CB units (300.40 Mb). Of the 30 834 BAC clones in the anchored contigs, 23 911 (77.5%) were directly anchored by BESs: 9748 by paired ends and 14 163 by single ends. Thirteen small contigs consisting of 103 BAC clones and spanning a total length of 2549 CBs (approximately 2532 kb) could not be anchored to IRGSP1.0. Some of these contigs contained no or few BESs for anchoring, whereas the others may be from unique regions of the 93–11 genome relative to Nipponbare or may correspond to the gap regions of IRGSP1.0.

For the 93–11 BGI, BLAT matched 35 804 BESs (54.98% of the total 93–11 BESs), and SYMAP anchored 277 93–11 PM contigs comprising 30 790 BAC clones spanning a total length of 297 489 CB units (300 Mb). Of the 30 790 BAC clones in the anchored contigs, 22 311 (72.46%) were directly anchored by BESs: 9116 by paired ends and 13 195 by single ends. Ten small contigs consisting of 148 BAC clones and spanning a total length of 2348 CBs (approximately 2369 kb) cannot be anchored to the 93–11 BGI. Some of these contigs contained no or few BESs for anchoring, whereas the others may correspond to missing regions of the 93–11 BGI.

Discrepancies among the three assemblies

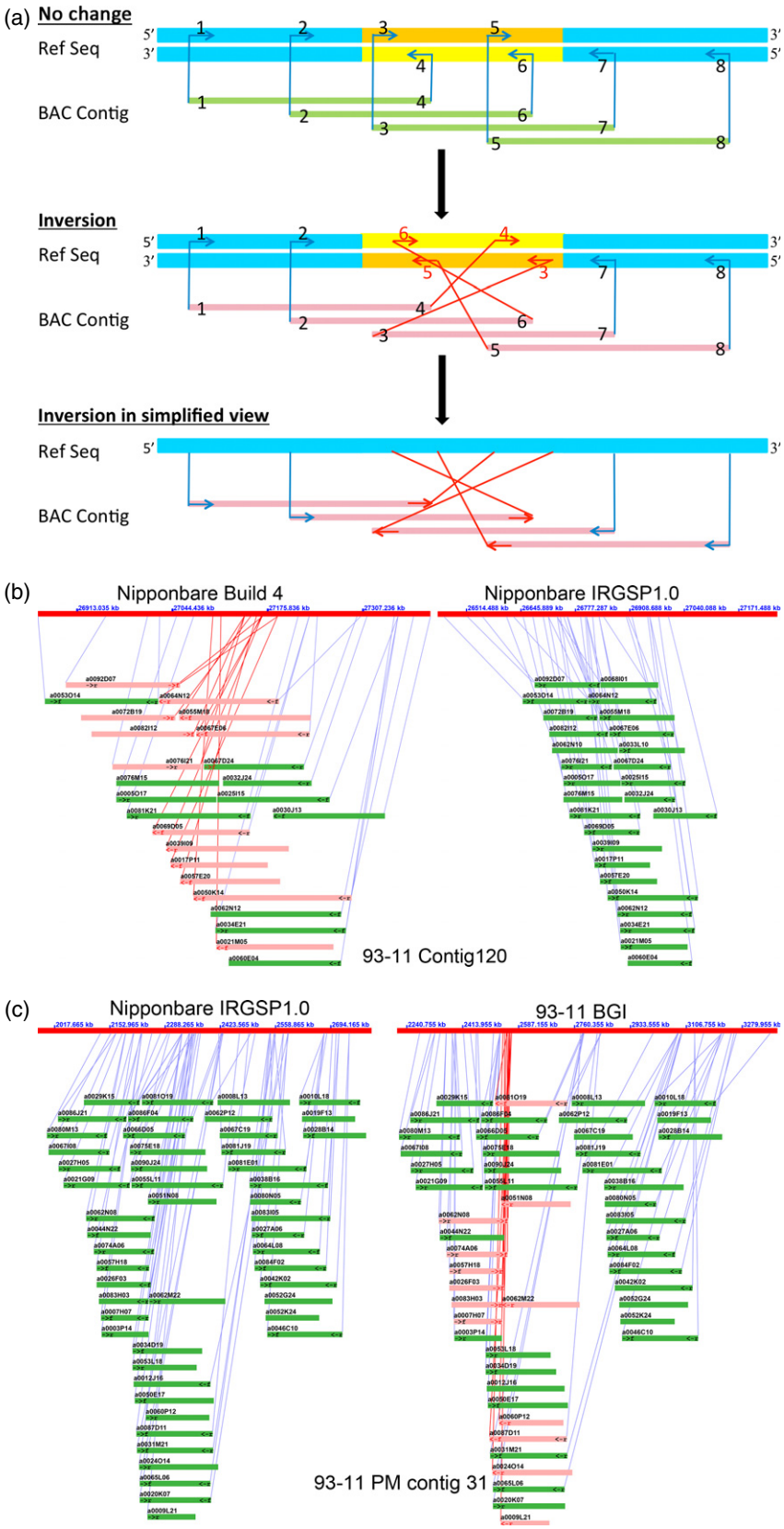
Discrepancies between the 93–11 PM contigs and the Nipponbare reference sequence must be due to one of three reasons: genetic variations between the two varieties, assembly errors and missing sequences in the Nipponbare reference sequence, or assembly errors in the 93–11 PM contigs (the missing regions between the contigs of 93–11 PM are excluded in comparisons). In contrast, discrepancies between the 93–11 PM contigs and the 93–11 BGI may be due only to either assembly errors in the 93–11 PM contigs or assembly errors and missing sequences in the 93–11 BGI. In tri-assembly comparisons, every assembly is used as an outgroup for comparison of the other two assemblies. We found many discrepancies among the three assemblies.

Inversely matched BES hits

BESs on the contigs may be used to detect inversions, as illustrated in Figure 1(a). If the two assemblies in the comparison are from the same genome, the inversion must come from a spurious event that reflects an assembly error in either assembly. If the two assemblies in the comparison are from two different genomes, then the inversion may come from either a genetic event or a spurious event. In these cases, one BES of each related BAC clone inversely matches the reference sequence (i.e. matches the complementary strand).

Our analysis showed that 1589 BESs from the anchored 93–11 PM contigs inversely matched 930 locations on IRGSP1.0. Of these, 791 BESs inversely matched 265 locations on IRGSP1.0 in groups of two or more BESs (Tables 1 and S1). Theoretically, the more BESs that inversely match the same location, the more reliable the detection will be. Because all 93–11 PM contigs have high qualities and high coverage of randomly produced BAC clones (manually inspected), it is almost impossible that spurious events occur simultaneously in a group of two or more overlapping individual clones inside the 93–11 PM contigs. Therefore, many of the locations inversely matched by multiple BESs may contain genetic inversions between Nipponbare and 93–11, or spurious inversions caused by assembly errors on IRGSP1.0. In fact, four locations of known spurious inversions of IRGSP build 4, one found by Yu *et al.* (2006) and three found in our previous work (Lin *et al.*, 2012), were also detected by the 93–11 PM BESs on IRGSP build 4 (Table S2), but three were not detected on the corrected IRGSP build 5 and IRGSP1.0. Figure 1(b) shows that the same 93–11 PM contig 120 detected an inversion on IRGSP build 4 but not on IRGSP1.0. We retrieved the corresponding Nipponbare BAC contig 40 (Chen *et al.*, 2002), and aligned this contig to the same region on both IRGSP build 4 and IRGSP1.0. A similar result to that in Figure 1(b) confirmed the assembly error in IRGSP build 4 at this location (Figure S2). These analyses demonstrate the power of BES alignments for detection of spurious inversions.

Figure 1. Finding inversions through BES alignments.
(a) Schematic view. Top panel: BESs of the contig are normally matched to the reference sequence. Both strands are shown for the reference sequence. Arrows indicate the directions (5′→3′) of the sequences matched by BESs. Lines represent sequence alignments. Middle panel: inversion indicated by inversely matched BESs of contig. Bottom panel: simplified view of the middle panel. Only the plus strand is shown for the reference sequence. The arrows are moved onto the BAC ends to confer the BES function of inversion detection.
(b) Discovery and confirmation of the spurious inversion in the Nipponbare reference sequence by BES alignments. Left panel: a group of BESs of 93–11 PM contig 120 are inversely matched to IRGSP build 4 chromosome 04, indicating that a genetic or spurious inversion exists on IRGSP build 4 chromosome 04. Right panel: the group of BESs of the same 93–11 PM contig 120 are normally matched to the corresponding location of IRGSP1.0 chromosome 04, excluding the possibility of the genetic inversion and confirming the existence of the spurious inversion in IRGSP build 4.
(c) Discovery and confirmation of the spurious inversion of the 93–11 BGI by BES alignments. A group of BESs of 93–11 PM contig 31 are normally aligned to IRGSP1.0 chromosome 06 (left panel) but inversely to the 93–11 BGI chromosome 06 (right panel), confirming the spurious inversion of the 93–11 BGI.
In (b) and (c), the reference sequence is shown at the top of each panel in red. Pink bars represent BAC clones with inversely matched BESs. Green bars represent BAC clones with normally matched BESs. Red arrows on pink bars represent inversely matched BESs, and red lines indicate inverse alignments. Blue arrows on green bars represent normally matched BESs, and blue lines indicate normal alignments.



Our analysis also showed that 1806 BESs from the anchored 93–11 PM contigs inversely matched 1041 locations on the 93–11 BGI. Of these, 1091 BESs inversely matched 325 locations in groups of two or more BESs (Tables 1 and S3). For the same reason explained above, many of these locations may contain spurious inversions that reflect assembly errors of the 93–11 BGI. We inspected the region between nucleotides 2 540 228 and 2 567 193 of 93–11 BGI chromosome 06 (highlighted rows in Table S3). Figure 1(c) shows the alignment results for the corresponding region of 93–11 PM contig 31 to both 93–11 BGI and Nipponbare IRGSP1.0. All BESs from this contig region were normally matched to the IRGSP1.0 (Figure 1c, left panel), whereas 13 BESs from the same contig region inversely matched the 93–11 BGI (Figure 1c, right panel), suggesting that the discrepancy in this region is due to the assembly error of the 93–11 BGI. Our subsequent experiment confirmed the assembly error at this location on 93–11 BGI by PCR (Appendix S1, Figures S3 and S4, and Tables S4 and S5).

Contig alignment discrepancies

Alignments of the 93–11 PM contigs to IRGSP1.0 and 93–11 BGI displayed many contig alignment discrepancies. Figure 2 shows the comparative results for chromosome 04, and comparative results for the other 11 rice chromosomes are shown in Figure S5. Some representative large and distinct discrepancies are indicated in the figures, and are further analyzed below.

Contigs aligned to unique regions on IRGSP1.0 but spanning two or more chromosomes on 93–11 BGI. We did not find any 93–11 PM contigs spanning two or more chromosomes on IRGSP1.0. However, several 93–11 PM contigs were found to align to unique regions on IRGSP1.0 but span two or more chromosomes on 93–11 BGI (spurious translocations; Figures 3 and S6). Contig 40 aligned to chromosome 12 on IRGSP1.0, but half aligned to chromosome 12 and half to chromosome 05 on 93–11 BGI. Contig 79 aligned to chromosome 11 on IRGSP1.0, but different parts of this contig aligned to chromosome 11 and 06 on 93–11 BGI. Contig 177 aligned to chromosomes 11 on IRGSP1.0, but different parts of this contig aligned to chromosomes 11 and 04 on 93–11 BGI. Contig 285 aligned to chromosome 12 on IRGSP1.0 but the main part aligned to chromosome 12 with two small parts aligned to chromosome 11 on 93–11 BGI. Four reasons support our conclusion that mis-assembly of 93–11 BGI in these regions is the only explanation for these discrepancies: (i) all PM contigs were assembled well and were re-inspected to verify their reliability; (ii) each contig contains many BESs that align to reference sequences as a single unit with multiple anchors; (iii) only the best alignments were displayed; (iv) all these contigs aligned well to IRGSP1.0. Chromosome 11 and 12 have high homology in an approximately 2 Mb region, with interruptions at the tip of the chromosomes (International Rice Genome Sequencing Project, 2005; and our analysis). The stringency that we used distinctly aligns all contigs in this region to either chromosomes 11 or 12. Contig 285 locates in this region but is clearly from chromosome 12. The BES alignments of two small parts to chromosome 11 rather than chromosome 12 on the 93–11 BGI must be caused by mis-use of chromosome 12 sequences in assembly of chromosome 11 on the 93–11 BGI.

To further confirm our above analysis, we performed FISH experiments for 93–11 PM contig 79 (Figure 3). The BAC clones 38L08 and 51D19 from 93–11 PM contig 79 that aligned to chromosome 06 on 93–11 BGI were selected as probes. These clones were labeled with digoxigenin and co-hybridized with chromosome 11-specific 5S rDNA labeled with biotin to chromosomes prepared from 93–11. The results showed that both clones hybridized to chromosome 11, which indicates mis-assembly of 93–11 BGI in this region.

Contigs with disorderly BES alignments to 93–11 BGI. We did not find any 93–11 PM contigs with large regions disorderly aligned to IRGSP1.0 but found many to 93–11 BGI. Contigs 79 (Figure 3), 91, 177 and 186 (Figure S6) were all normally aligned to IRGSP1.0 but disorderly aligned to 93–11 BGI in large regions. For the same reasons discussed above, the discrepancies in these regions among the three assemblies must also be caused by assembly errors of 93–11 BGI. The contig 186-aligned region on 93–11 BGI appeared to contain a spurious large fragment inversion, and flipping of that region normalized the alignments. However, flipping the contig 79-, 91- and 177-aligned regions on 93–11 BGI did not make the alignments normal.

Contigs aligned to Nipponbare and 93–11 BGI reference sequences in different orders. We found that some 93–11 PM contigs aligned to IRGSP1.0 and 93–11 BGI in different contig orders. Figure 2 shows that 93–11 PM contigs aligned to IRGSP1.0 chromosome 04 in the order 280–91 but to 93–11 BGI in the order 91–280. Contigs 280 and 91 together span approximately 4.68 Mb of chromosome 04 of IRGSP1.0, and approximately 3.95 Mb of chromosome 04 of 93–11 BGI (Figure 2b). The almost 0.7 Mb difference in this region between IRGSP1.0 and 93–11 BGI may be explained by both insertions being present in IRGSP1.0 and missing sequence in 93–11 BGI. The corresponding region of contig 91 on 93–11 BGI has already displayed disorderly BES alignments (Figure S6). Taken together, these results indicate that the sequence of this large region of 93–11 BGI chromosome 04 is mis-assembled.

Similarly, on the comparative maps of chromosome 11 (Figure S5), the 93–11 PM contigs 79 and 230 aligned

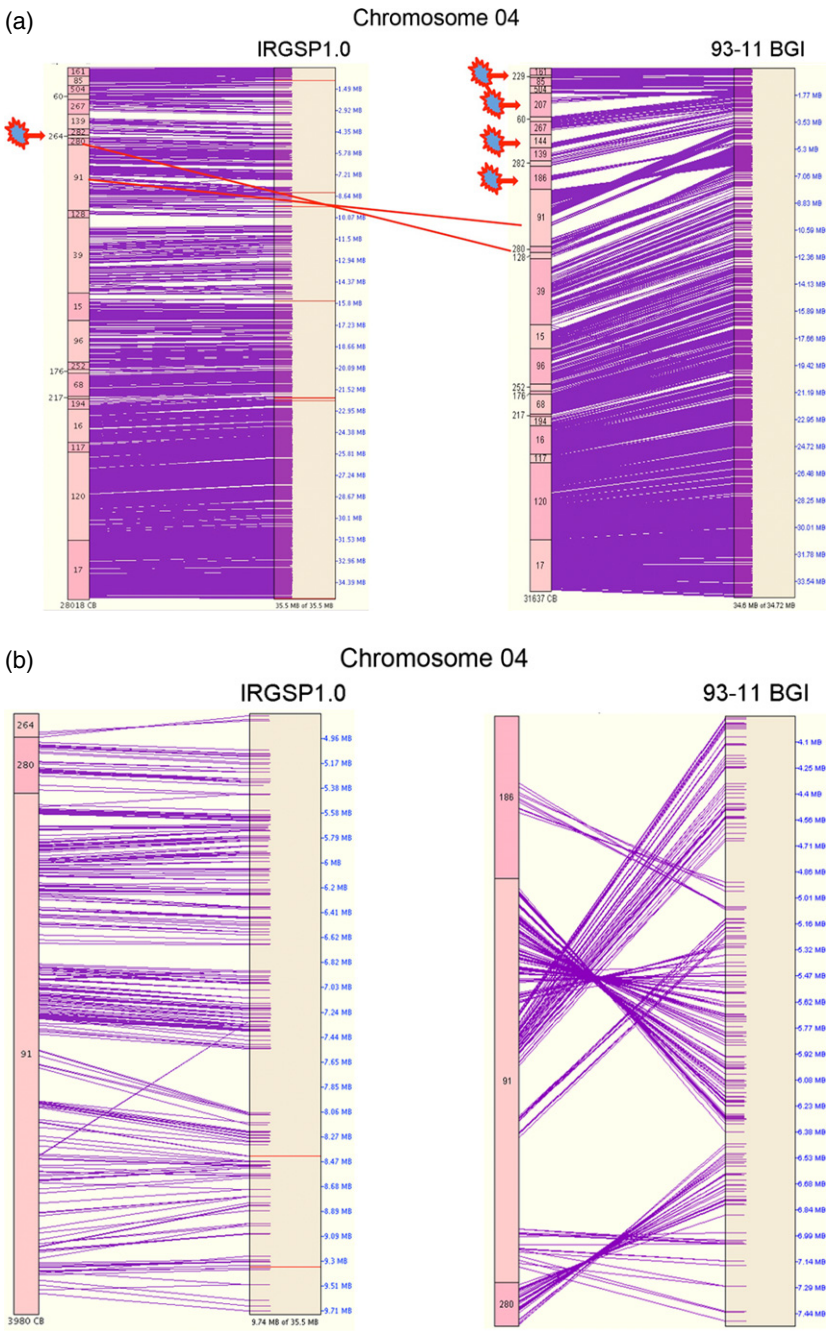


Figure 2. SYMAP alignments of the 93–11 PM contigs to chromosome 04 of IRGSP1.0 (left panel) and 93–11 BGI (right panel).
(a) Whole-chromosome view. The blue/red splodges indicate the inserted contigs.
(b) Enlarged view of the indicated region.
The boxes on the left of each panel represent the 93–11 PM contigs. The purple lines represent the 93–11 PM contig BES alignments. The distinct discrepant contigs are indicated. Red lines link the two contigs that are differently ordered on IRGSP1.0 and 93–11 BGI. Red bars on IRGSP1.0 represent physical gaps. Zoomable figures and detailed information for local regions are available at http://GResource.hzau.edu.cn.

to IRGSP1.0 in the order 79–230 but to 93–11 BGI in the order 230–79. Contig 93 also aligned to a different region on IRGSP1.0 from that on 93–11 BGI. Our results (Figures 3, S5 and S6) show that assembly of an approximately 3.8 Mb region of 93–11 BGI Chromosome 11, spanned by 93–11 PM contigs 93, 186, 230, 79, 260 and177, is disordered.

A few contigs aligned to either IRGSP1.0 or 93–11 BGI. However, these contigs were usually small and aligned with two or three BESs. More experiments are required to confirm their correct alignments.

Contigs spanning physical gaps of IRGSP1.0. The 93–11 BGI contains several tens of thousands of gaps, and 93–11 PM may be used as a robust resource to close the majority of those gaps. To date, IRGSP1.0 contains 44 physical gaps. We found that 35 IRGSP1.0 physical gaps are spanned by 30 93–11 PM contigs (Figure S7). The gap-corresponding regions of the 23 93–11 PM contigs (contigs 34, 168, 29, 4, 248, 232, 42, 85, 194, 157, 127, 108, 109, 274, 154, 261, 129, 190, 255, 77, 207, 148 and 27) are small, and are encompassed by single BAC clones with paired-end BESs. The side regions of these contigs aligned well to the flanking

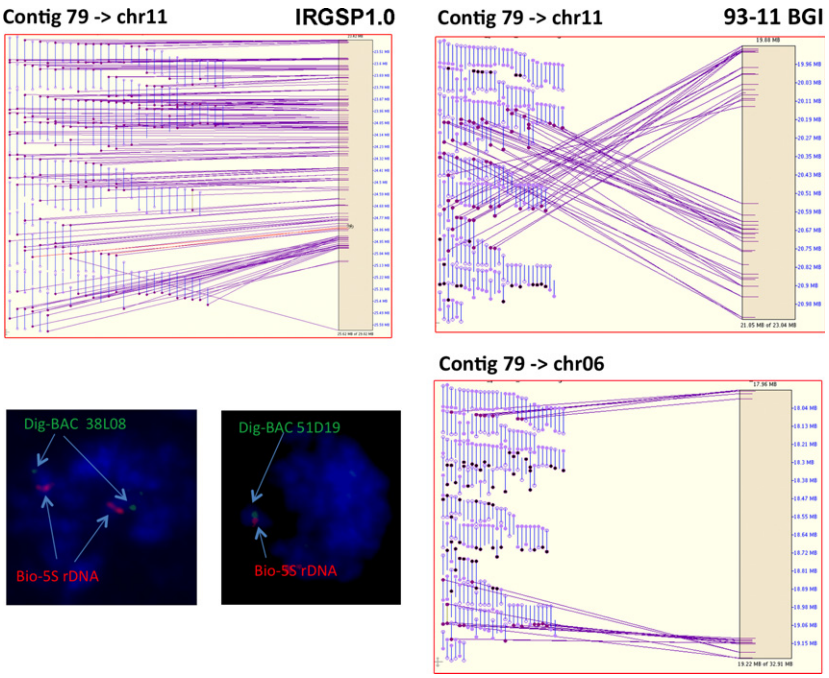


Figure 3. Mapping of 93–11 PM contig 79.
Top left: the 93–11 PM contig 79 is aligned to chromosome 11 on IRGSP1.0. Right: different parts of the 93–11 PM contig 79 are disorderly aligned to chromosome 11 and chromosome 06 of the 93–11 BGI, respectively. The vertical blue lines represent BAC clones. The solid purple circles, solid pink circles and empty pink circles represent BESs that match unique places, multiple places and masked repetitive sequences/missing sequences (no matches) of the reference sequences, respectively. Bottom left: FISH mapping of BAC clones 38L08 and 51D19 of 93–11 PM contig 79 on the 93–11 chromosome. The 5S rDNA is a rice chromosome 11-specific probe.

regions of the corresponding IRGSP1.0 gaps. These gaps may contain unique or low-copy genomic sequences and may be easily closed by designed sequencing. In contrast, gaps aligning to (or near) the ends of the remaining contigs may contain repetitive sequences or have complex sequence structures.

DISCUSSION

We constructed a deep-coverage BAC physical map for the *indica* rice variety 93–11 to provide a robust resource to improve both the Nipponbare and 93–11 reference sequences. Using BESs associated with 93–11 PM contigs, we flagged 265 regions on IRGSP1.0 to which groups of two or more BESs inversely aligned. These regions may be considered candidates for genomic inversions between Nipponbare and 93–11 or sequence assembly errors (spurious inversions) of the Nipponbare reference sequence. However, the agreement between the discordant regions detected on IRGSP build 4 by the 93–11 PM BESs (Table S2) and the known spurious inversions reported by Yu *et al.* (2006) and Lin *et al.* (2012), or corrected in IRGSP build 5 and IRGSP1.0, suggests that the majority of discordant regions detected by BESs in this study are caused by assembly errors of the Nipponbare reference sequence rather than genetic inversions between Nipponbare and 93–11. Thirty-five of the 44 euchromatic physical gaps on IRGSP1.0 were spanned with 93–11 PM contigs (Figure S7), and thus these contigs may provide insight into the sequence content within these gaps once sequenced. For example, nine Nipponbare gaps (gap 3 on chromosome 03; gaps 4 and 7 on chromosome 04; gaps 2 and 3 on chromosome 10; gaps 2 and 5 on chromosome 11; gaps 2 and 4 on chromosome 12) that were spanned by 93–11 PM contigs were also spanned by *japonica* ZH11 contigs (Lin *et al.*, 2012; see http://gresource.hzau.edu.cn/resource/symap_93–11.html for the updated alignment between ZH11 contigs and IRGSP1.0). The sizes and patterns of the gap-corresponding regions on the 93–11 PM contigs and the ZH11 contigs are similar, suggesting that these regions may be conserved between the two *O. sativa* sub-species. Therefore, these gap-spanning contigs may be used as evidence to help close the majority of the gaps that remain on IRGSP1.0.

By a parallel comparison, we detected various types of assembly errors on the 93–11 BGI, such as missing sequences, spurious inversions, multiple assignments of identical sequence contigs (i.e. spurious duplication) and mis-assignment of sequence contigs (i.e. spurious translocation). Much fewer 93–11 PM BESs matched to the 93–11 BGI (22 311) than to IRGSP1.0 (23 911; Table 1). This is not surprising because the 93–11 BGI still contains more than 50 000 gaps (Yu *et al.*, 2005) and approximately 106 Mb of 93–11 sequences were not mapped (according to the downloaded 93–11 BGI sequence reads). Using BESs associated with 93–11 PM contigs, we also flagged 325 locations on the 93–11 BGI to which groups of two or more BESs were inversely matched (Table S3). These locations most likely contain spurious inversions. In fact, when such a location was further analyzed, mis-assembly was implied (Figure 1c). Large stretches of chromosomes 04 and 11 of the 93–11 BGI are mis-assembled.

Our work has provided a platform to evaluate different genome sequencing approaches. The quality of a genome

sequence is primarily affected by the genome size and complexity, sequence read accuracy and length, and the sequencing and assembly strategies. Nipponbare and 93–11 have similar genome sizes and complexities, and both used Sanger sequencing reads; therefore, the quality differences between the IRGSP build 4 and 93–11 BGI were mainly caused by use of different sequencing and assembly strategies. Nipponbare IRGSP build 4 was obtained using a costly map-based BAC-by-BAC approach, whereas the 93–11 BGI was achieved using a relatively cost-saving WGS approach. A deep-coverage and manually edited physical map is considered essential for obtaining a high-quality genome sequence (Meyers *et al.*, 2004; Lewin *et al.*, 2009). In the map-based BAC-by-BAC approach, assembly errors are limited to individual BACs. In contrast, many large and long-range mis-assemblies on various chromosomes were detected on the 93–11 BGI. Recently, we analyzed five locations on IRGSP build 5 detected by groups of two or more inversely matched paired-end BESs from the Nipponbare physical map (Chen *et al.*, 2002), and found that they were all small spurious inversions (within BAC size assemblies). Dot-plot analysis and PCR validation found that four spurious inversions were caused by the flanking reverse repetitive sequences, and one may be caused by assembly using low-coverage or low-quality sequence reads (Deng *et al.*, 2013).

The larger and more complex the genome and the shorter the sequence reads, the more difficult it is to assemble a genome sequence. In recent years, the scientific community has seen a rapid increase in the number of genome sequences for higher-plant species generated using WGS/NGS technologies (Hamilton and Buell, 2012). These sequences play important roles in accelerating functional genomics studies of the respective species. However, many such assemblies may be far from suitable as reference genome sequences. Most of these species have larger and more complex genomes than rice. NGS sequence reads are much shorter than Sanger sequence reads. High-quality genome sequence assembly with NGS reads is difficult to achieve (Alkan *et al.*, 2011; Birney, 2011; Schatz *et al.*, 2012). Therefore, it may be anticipated that the quality of these genome sequences is much lower than that of the 93–11 BGI. A separate evaluation of the quality of most of these plant genome sequences is lacking.

Our analysis demonstrates that completing a high-quality genome sequence of a higher eukaryotic organism is not a trivial task and requires multiple efforts, supporting the statement that 'every genome sequence needs a good map' (Lewin *et al.*, 2009). The Nipponbare reference sequence was declared as the golden standard for a crop genome. Even this 'gold standard' assembly has many small spurious inversions and several tens of physical gaps, but still has the best quality of any existing crop genome sequence. The cost of generating a high-quality genome sequence is a major consideration when embarking on such an endeavor, whether or not a physical map is required. Although many basic questions may be answered using a low-cost WGS/NGS assembly, it is important to make a distinction between a WGS-NGS draft assembly and a reference sequence. It is our opinion that the majority of crop genomes and key species should have the goal of having reference-quality genomes. If a high-quality reference genome sequence is required, the map-based BAC-by-BAC approach may remain the best method for a high cost/benefit ratio. First, even though the sequencing cost dramatically drops with WGS-NGS technology, the overall cost to obtain a genome sequence is still not low (Sboner *et al.*, 2011). Second, because of the low assembly quality in the WGS approach, especially with NGS reads, continuous improvement of the low-quality genome sequence is required, resulting in overall costs that are not less than *de novo* BAC-by-BAC genome sequencing. Third, the lost cost and time resulting from incorrect conclusions and pointless experiments performed on the basis of the errors in the reference sequence is immeasurable. Recently, we greatly improved the vector and technology for BAC library construction (Shi *et al.*, 2011; Wang *et al.*, 2013), and established a new method for simultaneous *de novo* physical mapping and genome sequencing using the same set of NGS sequences (Y. Pan, X. Wang, L. Liu, H. Wang, G. Wang and M. Luo, unpublished results). At present, NGS physical maps combined with NGS assemblies of BACs are a viable and cost-effective strategy to obtain the highest-quality genome assembly for the lowest cost. In the future, as longer sequence reads become more available and more accurate, it may be possible to avoid the requirement for physical maps.

In conclusion, in this study, we evaluate the quality of the two rice reference sequences as the first example for all plant genome sequences. Our results demonstrate that the Nipponbare reference sequence is of high quality, but still contains small spurious inversions in addition to the reported several tens of gaps. However, many large and varied assembly artifacts were found in the 93–11 BGI, such as inversions, duplications and translocations, as well as missing sequences. Our work provides a reference and invaluable resource for improvements toward completion of both Nipponbare and 93–11 reference sequences. Because most plant species that have been sequenced have a more complex genome than rice, and because their whole-genome sequence assemblies usually used short NGS reads compared with the two rice reference sequences (which used long Sanger reads), we anticipate that the quality of the WGS-NGS sequences may be much lower even than that of the 93–11 BGI, and may not qualify as reference sequences.

EXPERIMENTAL PROCEDURES

BAC library construction

BAC library construction was performed as described previously (Luo *et al.*, 2001; Luo and Wing, 2003). The BAC vector was prepared from the pAGIBAC1 vector (Luo *et al.*, 2006). Megabase genomic DNA was isolated from young 93–11 seedlings, restricted with *Hin*dIII, cloned into the vector and used to transform *Escherichia coli* DH10B cells. Insert containing clones were arrayed into 384-well microtiter plates. The BAC library was named OSI9Ba and contained 36 864 clones. The BAC library copies were stored at both the Arizona Genomics Institute (http://www.genome.arizona.edu) and the Genome Resource Laboratory of Huazhong Agricultural University (http://GResource.hzau.edu.cn). Details of the BAC clones and filters are available on these websites. Eighty-seven BAC clones were randomly selected from the library, and plasmids were restricted using *Not*I before analysis on 1% agarose CHEF gels (Bio–Rad, http://www.bio-rad.com) using a 5–15 sec linear ramp time at 6 V cm^{-1} and 14°C in 0.5× TBE buffer for 16 h.

BAC end sequencing, fingerprinting, FPC assembly, contig alignment and manual editing

BAC end sequencing, fingerprinting, FPC assembly, contig alignment and manual editing were performed exactly as described previously (Lin *et al.*, 2012). The 93–11 PM contig BESs were masked using Repeatmask (http://www.repeatmasker.org/) to remove the repeat and organelle sequences, and the remaining sequences were used to align the 93–11 PM contigs to each of the three versions of the Nipponbare reference sequence and the 93–11 BGI using BLAT (Kent, 2002). The alignments were then displayed using SYMAP (Soderlund *et al.*, 2006). The parameters used for alignments the SYMAP default parameters. The 93–11 BGI was downloaded from the website http://rice.genomics.org.cn. In SYMAP, BESs on physical contigs were assigned by global locations of alignments. The orientations of BESs were determined by directions of sequence alignments. Normally, the paired-end BESs are aligned in different directions, as shown in Figure 1(a). If the assigned orientations of BESs did not agree with the directions of sequence alignments, discrepancies between the physical map and reference sequence exist. We re-drew all the physical contigs that inversely matched the reference sequences and labeled all discrepant BESs.

FISH experiment

Root tips of the *indica* rice variety 93–11 were pre-treated at 0°C for 24 h, and fixed in 100% ethanol/acetic acid (3:1 v/v) at 4°C overnight. Mitotic chromosome preparation was performed using the routine protoplast technique as described by Gustafson and Dille (1992). The BAC DNA and rice chromosome 11-specific 5S rDNA were labeled using a digoxigenin and biotin nick translation kits, respectively (Hoffmann La Roche, http://www.rocheusa.com).

The *in situ* hybridization protocol was modified slightly from that described by Jiang *et al.* (1995). The hybridization mixture contained 50% deionized formamide, 20% sodium dextran sulfate, 10% 2× SSC, 1 mg ml^{-1} salmon sperm DNA, 30 ng 5S rDNA probes and 40 ng BAC DNA probes, and co-denatured at 75°C for 5 min. Hybridization was performed overnight at 37°C. The biotinylated probes were detected using fluorescein isothiocyanate-conjugated avidin, and the dig-labeled probes were detected using a rhodamine-conjugated anti-dig antibody. Chromosomes were counterstained using 5 mg ml^{-1} 4′,6–diamidino-2–phenylindole in Vectashield (Vector Laboratories, http://www.vectorlabs.com). Chromosome preparations were examined using an Olympus BX–60 fluorescence microscope (http://www.olympus-global.com). Images were captured using a cooled CCD camera (SenSys 1401E B0; Photometrics, http://www.photometrics.com) with METAMORPH software version 4.6r5 (Universal Imaging, http://www.universalimaginginc.com/). The final images were adjusted using Adobe Photoshop (http://www.adobe.com/).

Accession numbers

The BES sequences reported in this paper have been deposited in the GenBank database under accession numbers JY187038–JY252051.

Note added in proof

Just before re-submission of the revised manuscript for final consideration, Gao *et al.* (2013) reported an update of the 93–11 genome sequence. This work filled approximately 3.8 Mb of new sequences into 1493 gaps, corrected 62 650 single-base errors, and removed 23.6 Mb of falsely assembled sequences. We therefore performed a new alignment between 93–11 PM and the updated 93–11 BGI sequence (Figure S8). Compared with the alignment between 93–11 PM and the early 93–11 BGI sequence above, the new alignment showed three improvements on large sequence fragments, three contig shifts and differences in 16 contig inserts/deletes.

ACKNOWLEDGEMENTS

We thank So-Jeong Lee, Nicholas B. Sisneros, Xiang Song, Fusheng Wei, José L. Goicoechea, HyeRan Kim, Yeisoo Yu, Jetty S.S. Ammiraju, and other members of the Arizona Genomics Institute for production of the BAC library, fingerprints and BESs. We also thank Jun Li (College of Life Sciences, Wuhan University, China) for his technical assistance in FISH experiments. This work was supported by the National Natural Science Foundation of China (grant number. 30971748) and the Chinese 111 Project (grant number B07041).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Comparisons of three Nipponbare genome sequence versions at the structural level.

Figure S2. Alignments of Nipponbare contig 40 to Nipponbare genome sequence build 4 and IRGSP1.0.

Figure S3. PCR product validation using primers listed in Table S4.

Figure S4. Alignments between BESs on contig 31 of 93–11 and the two references.

Figure S5. SYMAP alignments of 93–11 PM contigs to chromosomes of Nipponbare IRGSP1.0 and 93–11 BGI.

Figure S6. Comparative view of 93–11 PM contigs aligning to unique regions on IRGSP1.0 but spanning two or more chromosomes on the 93–11 BGI.

Figure S7. The comparative view of the 93–11 PM contigs spanning the physical gaps on IRGSP1.0.

Figure S8. Differences between the earlier and updated version of the 93–11 reference sequence at the structural level.

Table S1. Locations on Nipponbare IRGSP1.0 flagged by inversely matched BESs of the 93–11 PM contigs.

Table S2. The locations on Nipponbare IRGSP build 4 flagged by inversely matched BESs of the 93–11 PM contigs.

Table S3. Locations on the 93–11 BGI flagged by inversely matched BESs of the 93–11 PM contigs.

Table S4. Primers used for experimental validation.

Table S5. PCR results for experimental validation.

Appendix S1. Experimental validation of 93–11 contig 31.

REFERENCES

Alkan, C., Sajjadian, S. and Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods*, **8**, 61–65.

Ammiraju, J.S., Yu, Y., Luo, M. *et al.* (2005) Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa* spp. Nipponbare) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. *Theor. Appl. Genet.* **111**, 1596–1607.

Birney, E. (2011) Assemblies: the good, the bad, the ugly. *Nat. Methods*, **8**, 59–60.

Chen, M., Presting, G., Barbazuk, W.B. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell*, **14**, 537–545.

Church, D.M., Goodstadt, L., Hillier, L.W. *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112.

Church, D.M., Schneider, V.A., Graves, T. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091.

Deng, Y., Pan, Y. and Luo, M. (2013) Detection and correction of assembly errors of rice Nipponbare reference sequence. *Plant Biol.* doi: 10.1111/plb.12090.

Dolgin, E. (2009) Human genomics: the genome finishers. *Nature*, **462**, 843–845.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Feltus, F.A., Wan, J., Schulze, S.R. *et al.* (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* **14**, 1812–1819.

Gao, Z.Y., Zhao, S.C., He, W.M. *et al.* (2013) Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc. Natl Acad. Sci. USA*, **110**, 14492–14497.

Gustafson, J.P. and Dille, J.E. (1992) Chromosome location of *Oryza sativa* recombination linkage groups. *Proc. Natl Acad. Sci. USA*, **89**, 8646–8650.

Hamilton, J.P. and Buell, C.R. (2012) Advances in plant genome sequencing. *Plant J.* **70**, 177–190.

Harushima, Y., Yano, M., Shomura, A. *et al.* (1998) A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics*, **148**, 479–494.

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

Huang, X., Kurata, N., Wei, X. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.

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

Jiang, J., Gill, B.S., Wang, G.L. *et al.* (1995) Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc. Natl Acad. Sci. USA*, **92**, 4487–4491.

Kawahara, Y., de la Bastide, M., Hamilton, J.P. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 10.

Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

Kidd, J.M., Sampas, N., Antonacci, F. *et al.* (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods*, **7**, 365–371.

Kim, H., San Miguel, P., Nelson, W. *et al.* (2007) Comparative physical mapping between *Oryza sativa* (AA genome type) and *O. punctata* (BB genome type). *Genetics*, **176**, 379–390.

Kovach, M.J., Sweeney, M.T. and McCouch, S.R. (2007) New insights into the history of rice domestication. *Trends Genet.* **23**, 578–587.

Lewin, H.A., Larkin, D.M., Pontius, J. *et al.* (2009) Every genome sequence needs a good map. *Genome Res.* **19**, 1925–1928.

Lin, H., Xia, P., Wing, R.A. *et al.* (2012) Dynamic intra-japonica subspecies variation and resource application. *Mol. Plant*, **5**, 218–230.

Luo, M. and Wing, R.A. (2003) An improved method for plant BAC library construction. *Methods Mol. Biol.* **236**, 3–20.

Luo, M., Wang, Y.H., Frisch, D. *et al.* (2001) Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon *Fusarium* wilt (Fom–2). *Genome*, **44**, 154–162.

Luo, M.C., Thomas, C., You, F.M. *et al.* (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.

Luo, M., Kim, H., Kudrna, D. *et al.* (2006) Construction of a nurse shark (*Ginglymostoma cirratum*) bacterial artificial chromosome (BAC) library and a preliminary genome survey. *BMC Genomics*, **7**, 106.

Meyers, B.C., Scalabrin, S. and Morgante, M. (2004) Mapping and sequencing complex genomes: let's get physical!. *Nat. Rev. Genet.* **5**, 578–588.

Minocherhomji, S., Seemann, S., Mang, Y. *et al.* (2012) Sequence and expression analysis of gaps in human chromosome 20. *Nucleic Acids Res.* **40**, 6660–6672.

Sboner, A., Mu, X.J., Greenbaum, D. *et al.* (2011) The real cost of sequencing: higher than you think!. *Genome Biol.* **12**, 125.

Schatz, M.C., Witkowski, J. and McCombie, W.R. (2012) Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol.* **13**, 243.

Schneeberger, K., Ossowski, S., Ott, F. *et al.* (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl Acad. Sci. USA*, **108**, 10249–10254.

Shi, X., Zeng, H., Xue, Y. *et al.* (2011) A pair of new BAC and BIBAC vectors that facilitate BAC/BIBAC library construction and intact large genomic DNA insert exchange. *Plant Methods*, **7**, 33.

Soderlund, C., Longden, I. and Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535.

Soderlund, C., Nelson, W., Shoemaker, A. *et al.* (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* **16**, 1159–1168.

Wang, C., Shi, X., Liu, L. *et al.* (2013) Genomic resources for gene discovery, functional genome annotation, and evolutionary studies of maize and its close relatives. *Genetics*, **195**, 723–737.

Yang, J., Zhao, X., Cheng, K. *et al.* (2012) A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science*, **337**, 1336–1340.

Yu, J., Wang, J., Lin, W. *et al.* (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38.

Yu, J., Ni, P. and Wong, G.K. (2006) Comparing the whole-genome-shotgun and map-based sequences of the rice genome. *Trends Plant Sci.* **11**, 387–391.

