Comparing the whole-genome-shotgun and map-based sequences of the rice genome

Jun Yu1,2*, Peixiang Ni1,3* and Gane Ka-Shu Wong1,2,4

1Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing 101300, China
2James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Hangzhou 310007, China
3T-life Research Center, Fudan University, Shanghai 200433, China
4University of Washington Genome Center, Department of Medicine, Seattle, WA 98195, USA

The rice genome has now been sequenced using whole-genome-shotgun and map-based methods. The relative merits of the two methods are the subject of debate, as they were in the human genome project. In this Opinion article, we will show that the serious discrepancies between the resultant sequences are mostly found in the large transposable elements such as copia and gypsy, that populate the intergenic regions of plant genomes. Differences in published gene counts and polymorphism rates are similarly resolved by considering how transposable elements affect the sequence analysis.

Two methods used to sequence rice

Rice and humans share the dubious distinction of having their genomes sequenced by multiple competing research groups. In the case of rice, draft sequences were first published in 2002 by the Beijing Genomics Institute [1] and the Syngenta Corporation [2]. Both used a whole-genome-shotgun (WGS) method. These sequences were updated in 2005 [3], and are referred to as Beijing indica and Syngenta japonica. Later that year, a map-based sequence of japonica was published by the International Rice Genome Sequencing Project (IRGSP) [4]. The IRGSP questioned the quality of the WGS data, arguing that they were incomplete and misassembled. This situation is reminiscent of the debates that followed the publication of the human genome; to put matters in perspective, we recount history in Box 1 [5–20]. The details are different for rice because of the subtle ways by which transposable elements (TEs) in plant intergenic regions interfere with the WGS assembly and with the interpretation of the resultant sequences. Different species have to be treated differently. Once these issues are factored in, there is remarkable agreement in the sequences produced by these two methods.

Intergenic repeats in plant genomes

Plant genomes, particularly those of cereals crops such as rice and maize, are known to be full of TEs – even more so than in vertebrate genomes. One might think that a WGS would be disastrous, but (partly because plant genes are so small) the method works better in plants than it does in vertebrates. The reasons have to do with the nature of the plant TEs and where they lie in relation to the genes. Misassembly problems are due to exactly repeated sequences that are longer than the nominal read length of 500-bp. Some TEs are too small to do much damage. MITEs for example are a few hundred bases. It is the large TEs such as copia and gypsy that cause all the problems. These are almost exclusively found in the intergenic regions between genes [21,22]. Most of them are high copy number and were left out of the WGS assembly, as can be seen in Table S2 in Ref. [3]. Some TEs might be of sufficiently low copy number, counted as exact repeats, to go undetected. These can cause misassembly problems. So, one way or another, large intergenic TEs are sacrificed. However, most other sequences are correctly assembled. Regulatory regions flanking the genes are not expected to be sacrificed because they are not exactly repeated. More importantly, the different members of a gene family are generally easy to distinguish from each other because of their introns.

What is so fortuitous about a plant WGS is that the sacrificed sequences are unlikely to be functional. Large intergenic TEs evolve rapidly. For example, in only 3 million years, they expanded the maize genome from 1200-Mb to 2400-Mb [23]. TEs are also extensively methylated, so deamination of the 5-methyl deoxycytidine to deoxythymidine leads to an elevated mutation rate relative to genes. Similarly fast evolution is reported in indica and japonica rice [24]. Lack of appreciation for the difference in the rate of single-nucleotide-polymorphisms (SNPs) in genes and intergenic TEs has led to wildly different estimates of rice genetic variation ranging from 1.7 SNP/kb [25] to 7.1 SNP/kb [26]. The problem was that every research group used a different criterion to reject SNPs in repetitive sequences to avoid confusing SNPs with paralogs. By contrast, when the sequences were updated by J. Yu et al. [3], entire chromosomes were aligned, with the aid of 34 190 anchor points. Hence there is no confusion. Resultant rates for coding exons, introns and TEs
Box 1. The human genome debates

The formation of Celera, a private company, to sequence the human genome [5] triggered one of the most raucous episodes in recent science history. This came to fruition in 2001 with the release of two draft sequences, one by the International Human Genome Sequencing Consortium (IHGSC) [6] and the other by Celera [7]. The IHGSC trumpeted the superiority of their map-based strategy over the WGS favored by Celera – an issue on which we concur, given our own experiences in physical mapping [8]. But at that time, it was difficult to prove which method was superior because neither sequence was of that high a quality, particularly compared with the finished product that was published three years later by the IHGSC [9]. In the acrimonious subsequent debates, the IHGSC did not argue that the WGS per se failed, but rather that Celera failed to prove the efficacy of the WGS because Celera used ordering information from the IHGSC sequence [10-13]. This subtle distinction was important because a WGS can be made to work. Indeed, every vertebrate sequenced since that time, from mouse [14] to dog [15], has used one variation or another of a WGS. One might therefore conclude that both methods are valid. That is not entirely correct because the quality of the resultant sequences must also be considered.

Prominent members of the IHGSC were arguing that the sequence had to be near-perfect [16]. What they were worried about was not the completeness of the sequence or the single-base error rate. Both could conceivably be achieved by WGS. The concern was about misassemblies (i.e. that shotgun reads would be assembled in the wrong place because of the repetitive sequences that are known to be abundant in vertebrate and plant genomes). A year before the formation of Celera, the pross [17] and cons [18] of the WGS versus map-based methods were debated. WGS is faster and cheaper. The problem is that the likelihood of a misassembly increases with the size of the shotgun segment. Contrary to popular perception, map-based methods do not eliminate this problem. All they can do is localize it to the smaller regions defined by the mapped clones. The essential issues are understood but not well known outside of the small community of researchers who work on the sequence assembly algorithms, such as RePS [19]. Here, we discuss these issues. A ‘repeat’ is defined as any sequence that occurs more than once in a genome. No assumption is made about the underlying biology responsible for such a repeat (e.g. TE, recent segmental duplications and gene families).

- Exactly identical repeats matter. Approximately identical repeats do not. Given that all repeats diverge with time, particularly in non-protein-coding sequences, this is another way of saying that ancient repeats are harmless.
- Lengths matter. Long is bad. Short is good. Exact repeats shorter than nominal read lengths of 500-bp are harmless.
- Copy numbers matter. We can estimate copy numbers. High copy repeats are easy to detect, so even if they cannot be incorporated into the final assembly, one can at least avoid misassemblies. Low copy (i.e. two or three) repeats are difficult to detect and, therefore, might cause problems.

Unfortunately, one cannot know the severity of the misassembly problem without knowing the sequence itself. Given the resources that the IHGSC had already put into the map-based method, the WGS was thought to be an unacceptable risk. Moreover, because of their ability to localize misassembly problems, mapped clones would still be needed to finish the genome. One could therefore argue that there is no long-term cost advantage to a WGS; but one could also argue that having a draft sequence three years before the finished product is of value to the community. In any case, Celera took that risk. Notwithstanding how they did their original assembly, they have now redone their assembly without using the IHGSC sequence [20]. Their comparisons to the map-based sequence revealed a 97% agreement in order and orientation. To understand the differences, one has to consider the biology of the repeats, which is species-specific.

The perils of segmental duplications

For vertebrates, the experience is that misassembly problems are caused by recent segmental duplications rather than by large intergenic TEs. These do affect the genes, but no such problems were observed in the rice analysis, despite the existence of a segmental duplication 21 million years ago (Mya) and a whole genome duplication before the origin of the grasses 55 Mya to 70 Mya. Analysis of the human WGS assembly showed that, to cause problems, the repeats must be at least 97% identical over a 15-kb segment [27]. Given the neutral substitution rate of $6.5 \times 10^{-9}$ a year for grasses [28], 3% divergence corresponds to a duplication from 2.3 Mya. It is possible that problems will arise in other plant species. There is evidence that this might be the case with soybean, a recently diploidized tetraploid in which many (or most) genetic markers assign to more than one physical map contig [29]. Without question, for a sufficiently recent segmental duplication there will be misassembly problems. However, one should recognize that, at some point, even the map-based method will fail, although probably not as soon as a WGS.

Notice too that every genome sequenced to date, with a handful of exceptions, has been done on inbreds. Outbreds are known to introduce a plethora of other complications that are beyond the scope of this Opinion article. Nevertheless, this is something to be aware of as sequencing moves to less-studied organisms.

Completeness of the WGS sequence

All the WGS sequences analyzed in this paper are taken from Ref. [3], which are at 6× coverage, versus 4× in the original indica draft. The map-based sequence released by the IRGSP is now at 10× coverage. Differences in assembly quality as a function of increased shotgun coverage are reliably predicted by Lander-Waterman statistics [30]. For example, at nominal gene sizes of 3–4 kb, one would expect a nontrivial fraction of the genes to be fragmented (i.e. split between different contigs) at 4×, but much less so at 6× or 10×. However, one should not confuse a statistical sampling issue that is easy to fix by spending more money with an intrinsic problem with the WGS method.

To determine gene content, the IRGSP used an unorthodox approach. They started by aligning the map-based and WGS sequences to each other. What was odd, although fully documented in their Supplementary Notes [4], was that when they saw a discrepancy they rejected the entire sequence, as opposed to only the discrepant parts. Supplementary Table 19 in Ref. [4] indicates that only 258-Mb and 290-Mb of the indica and japonica WGS sequences were retained, respectively. It is one thing to reject a discrepancy in a large intergenic TE, but it is another thing to also reject the neighboring genes. For example, in Beijing indica they considered 258-Mb of the available 411-Mb and, not surprisingly, found only 68.3% of the genes. The IRGSP argued that the problem was due to the small size of the indica contigs, whose mean was only 8.2-kb. This betrayed two misunderstandings. First, the contigs are linked together, with the correct order and orientation, to create much larger scaffolds and super-scaffolds. Second, most of the genome is in a small
number of large sequences, but there are also a large number of small sequences. It makes more sense to use N50 size, or the size above which half of the total length is found. Thus, the indica contigs and super-scaffolds become 23-kb and 8.3-Mb, respectively.

Gene content is more reliably determined by direct comparison of experimentally confirmed genes with the assembled sequence, as shown in Ref. [3]. On the unlikely chance that the data in GenBank are corrupt, we redid this assessment using the same 19,079 non-redundant full-length cDNAs (nr-KOME) [31]. If we align the genes in one piece, without fragmentation, both WGS are at least 91.2% complete. The other genes are not missing. All the exons are present but the genes are fragmented across the introns, as expected from Lander-Waterman. However, 98.1% of the genes can be found intact in one or the other WGS. Requiring that the genes be anchored to the map brings us down to 97.7%. Applying the same rules to the IRGSP, 98.1% of the genes are found.

**Accuracy of the sequence assembly**

Although essentially all the genes are found in the WGS sequences, this does not prove that they are correctly assembled. The IRGSP addressed this issue by comparing the first megabase of chromosome 1 of Syngenta japonica with IRGSP japonica, finding several discrepancies. What they did not ask was if the discrepancies were genic or intergenic. We checked, and none of the structures for the 73 nr-KOME defined genes in this region are affected. Two genes are at slightly different positions but, overall, Figure 1 shows that there is a remarkable agreement in the positions of the 2685 genes on this chromosome. We also did an explicit scan for recent segmental duplications in IRGSP japonica, generously defined as regions of 95% identity.

**Figure 1.** Map positions for 2685 nr-KOME cDNAs on chromosome 1, comparing the IRGSP japonica map-based assembly with the Syngenta japonica WGS assembly.

**Figure 2.** (a) A local misassembly on the IRGSP japonica chromosome 11. This 80.4-kb region contains two nr-KOME cDNAs (AK108542, AK066746) that lie in opposite directions on the map-based and WGS assemblies. The IRGSP japonica assembly is contradicted by four pairs of BAC-ends (OSJNba0018B21, OSJNbb0084N04, K0472B04, OSJNba0044O21). (b) A simple flip will fix everything. The problem is due to a pair of unknown transposons that create flanking inverted repeats of 5.3-kb and 95% identity.
identity across 10-kb. A mere 6.5-Mb qualified, and these duplications are distributed across all chromosomes. In these regions, gene density is 3.5 times smaller, but gene positions with respect to Syngenta japonica remain in good agreement. The IRGSP has mentioned that they have also discovered centromeric repeats in Beijing indica. That is technically correct. However, if we consider all 19 079 nr-KOME defined genes, such contaminations affected only 72-bp (0.0004%) of the exons and 0.02% of the introns.

We would also caution against a presumption that the IRGSP data are perfect. They did not, for example, mask repetitive sequences in assembling the shotgun data from their bacterial-artificial-chromosomes (BACs). By contrast, all WGS assembly algorithms mask repetitive sequences, in one way or another. As surprising as this might seem, the IRGSP sequence has misassembly problems that are localized to individual BACs. For example, Figure 2 shows a region of chromosome 11 with two nr-KOME cDNAs. The IRGSP japonica is contradicted by 4 pairs of BAC-ends. Inverted repeats of size 5.3-kb and 95% identity flank the misassembly. Flipping this 80.4-kb region around fixes everything. These errors can escape detection by restriction enzyme (RE) fingerprint analysis because only one or two RE fragments are changed by any such error. Unfortunately, it is difficult to estimate the magnitude of this problem given the limited number of BAC-ends.

One of the advantages of a WGS is that the same effort is put into every chromosome and, indeed, even genes buried in heterochromatic DNA (often not clonable in BACs) can be recovered. Distributing the project among different laboratories, as was the case in the IRGSP, can lead to inconsistent quality standards. We did a string search for laboratories, as was the case in the IRGSP, can lead to discrepancies also disappear upon closer examination. Computational gene predictions can mistakenly identify TEs as genes [32]. In the updated analysis by J. Yu et al. [3], gene predictions were scanned for known TEs and 20-mers of high copy number. Beyond a 50% threshold, genes were rejected. For Syngenta japonica, this procedure gave 56 885 and 45 824 genes before and after removing likely TEs, respectively. Using a combination of EST confirmation and indica–japonica overlap, the gene count estimate was lowered to 37 794. In the IRGSP analysis, likely TEs were removed before the gene predictions. Applying this procedure to Syngenta japonica gave us 38 133 genes directly, comparable to the 37 544 genes reported by the IRGSP.

Cost–benefits for finished sequence
We acknowledge that the map-based method is superior, if the objective is a near-perfect finished sequence, as produced for the human genome. The issue is how much the improvement costs and how much it is worth. The updated data published in Ref. [3] by J. Yu et al. are similar, in spirit, to the ‘intermediate grade of finished genomic sequence’ that has been proposed by the National Human Genome Research Institute (NHGRI). By their estimates

<table>
<thead>
<tr>
<th>Reason for gene count discrepancy</th>
</tr>
</thead>
</table>
| Gene count estimates vary wildly. However, these differences also disappear upon closer examination. Computational gene predictions can mistakenly identify TEs as genes [32]. In the updated analysis by J. Yu et al. [3], gene predictions were scanned for known TEs and 20-mers of high copy number. Beyond a 50% threshold, genes were rejected. For Syngenta japonica, this procedure gave 56 885 and 45 824 genes before and after removing likely TEs, respectively. Using a combination of EST confirmation and indica–japonica overlap, the gene count estimate was lowered to 37 794. In the IRGSP analysis, likely TEs were removed before the gene predictions. Applying this procedure to Syngenta japonica gave us 38 133 genes directly, comparable to the 37 544 genes reported by the IRGSP. More detailed analysis (R. Li et al., unpublished) reveals that all existing gene sets are still contaminated by TEs because none of the TE libraries were complete. This problem was only recently solved by the introduction of a new algorithm, ReAS, to reconstruct ancestral sequences for transposable elements from the unassembled reads of a WGS [33]. It is beyond the scope of this Opinion article to explain how ReAS is incorporated into the gene predictions but, for interested readers, our updated gene sets are available at http://rice.genomics.org.cn/rice/link/download.jsp.

<table>
<thead>
<tr>
<th>Cost–benefits for finished sequence</th>
</tr>
</thead>
</table>
| We acknowledge that the map-based method is superior, if the objective is a near-perfect finished sequence, as produced for the human genome. The issue is how much the improvement costs and how much it is worth. The updated data published in Ref. [3] by J. Yu et al. are similar, in spirit, to the ‘intermediate grade of finished genomic sequence’ that has been proposed by the National Human Genome Research Institute (NHGRI). By their estimates

<table>
<thead>
<tr>
<th>Table 1. Ranking of the IRGSP japonica chromosomes by quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Country</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>CHR02</td>
</tr>
<tr>
<td>CHR01</td>
</tr>
<tr>
<td>CHR06</td>
</tr>
<tr>
<td>CHR12</td>
</tr>
<tr>
<td>CHR10</td>
</tr>
<tr>
<td>CHR09</td>
</tr>
<tr>
<td>CHR03</td>
</tr>
<tr>
<td>CHR07</td>
</tr>
<tr>
<td>CHR05</td>
</tr>
<tr>
<td>CHR08</td>
</tr>
<tr>
<td>CHR04</td>
</tr>
<tr>
<td>CHR11</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td><strong>Max/Min</strong></td>
</tr>
</tbody>
</table>

*Q40 bases refer to the highest quality regions in Syngenta japonica, where estimated error rates are better than 10^-4, and which are then compared with the corresponding bases in IRGSP japonica. Ngap refers to strings of 50 or more N’s that are found in IRGSP japonica. Genomewide means are indicated at the bottom, and the two columns labeled ‘vs Mean’ refer to the ratio of the chromosome-specific quantity to the genomewide mean. Rankings are based on the sum of these two ratios. ‘Max/Min’ refers to the ratio of lowest/highest quality chromosomes within the indicated columns.
[34], it requires 1/40 of the reagents and 1/10 of the personnel effort, while producing results of high quality, with most of the residual gaps and errors falling in the repetitive sequences. To our knowledge, no genome other than mouse is funded to be finished to the same near-perfect standards as human (at least since the completion of the smaller Arabidopsis and Drosophila genomes). We do appreciate that the IRGSP tried to match the human standards but, with the reduced funding of current times, it is important for the community to understand cost–benefits.

Acknowledgements
We thank Jun Wang, Hongkun Zheng, Yiran Guo, Dongyuan Liu, Yue Li and Songgang Li for the analysis presented in this article. In addition, we thank Maynard Olson for a critical reading of the manuscript.

References
19 Wang, J. et al. (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. Genome Res. 12, 824–831

Reproduction of material from Elsevier articles
Interested in reproducing part or all of an article published by Elsevier, or one of our article figures? If so, please contact our Global Rights Department with details of how and where the requested material will be used. To submit a permission request online, please contact:

Elsevier
Global Rights Department
PO Box 800
Oxford OX5 1DX, UK
Phone: +44 (0)1865 843 830
Fax: +44 (0)1865 853 333
permissions@elsevier.com
Alternatively, please visit:
www.elsevier.com/locate/permissions