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

Jun Yu[1,2,*], Peixiang Ni[1,3,*], Gane Ka-Shu Wong[1,2,4,‡].

[1]*Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing 101300, China.*

[2]*James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Hangzhou 310007, China.*

[3]*T-life Research Center, Fudan University, Shanghai 200433, China.*

[4]*University of Washington Genome Center, Department of Medicine, Seattle WA 98195, USA.*

[*]These authors contributed equally to this paper.

[‡]Corresponding author: gksw@genomics.org.cn

Abstract

The rice genome has now been sequenced using whole-genome-shotgun and map-based methods. Much as in the human genome project, there are debates over the relative merits of the two methods. We will show that to the extent there are serious discrepancies between the resultant sequences, they are mostly found in the large transposable elements like *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 human share the dubious distinction of having their genomes sequenced by multiple competing 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 a 2005 paper published in *PLoS* [3], and 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]. 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, and to put matters in perspective, we recount that history in Box 1. 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.

Box 1: The human genome debates

The formation of the private company, Celera, 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, as neither sequence was of particularly high quality, especially compared to the finished product that would be published 3 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. In fact, 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, as 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.* the fact that shotgun reads can 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 pros [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, like for example *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.* TEs, recent segmental duplications, gene families).

- Exactly identical repeats matter. Approximately identical repeats do not. Since all repeats diverge with time, especially 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. The point is 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.* 2 or 3) repeats are difficult to detect, and may therefore 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 can therefore argue that there is no long-term cost advantage to a WGS; but one can 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.

Intergenic repeats in plant genomes

Plant genomes, especially those of cereals crops like rice and maize, are known to be full of TEs – even more so than vertebrate genomes. One might thus think that a WGS would be disastrous; but in fact (and partly because plant genes are so small), the method works better in plants than in vertebrates. The reasons have to do with the nature of 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 like *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 left out of the WGS assembly, as can be seen in Table S2 of the *PLoS* paper. Some 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. Neither are the genes 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 the fact 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 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, 1.7 SNP/kb [25] to 7.1 SNP/kb [26]. The problem was that every paper used a different criterion to reject SNPs in repetitive sequences, to avoid confusing SNPs with paralogs. In contrast, for the *PLoS* paper, 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 were 3.0, 6.1, and 27.6 SNP/kb, respectively. Note the 10-fold increase from coding exons to TEs. It explains why the overall rate is so sensitive to how SNPs in repetitive sequences are rejected.

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 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 6.5×10^{-9} a year for grasses [28], 3% divergence corresponds to a duplication from 2.3 Mya . It is of course 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. One should however realize that, at some point, even the physical maps will fail to assemble correctly – just not as soon as the 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 paper. Nevertheless, this is something to be aware of, as sequencing moves to less-studied organisms.

Completeness of the WGS sequence

All of the WGS sequences analyzed in this paper are taken from the data in *PLoS*, which are at 6x coverage, versus 4x in the original *indica* draft. The map-based sequence released by IRGSP is now at 10x 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 to 4-kb, one would expect a nontrivial fraction of the genes to be fragmented (*i.e.* split between different contigs) at 4x, but much less so at 6x or 10x. However, one should not confuse a statistical sampling issue that is easy to fix by spending more money to an intrinsic problem with the WGS method.

To determine gene content, 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, was the fact that when they saw a discrepancy they rejected the entire sequence, as opposed to only the discrepant parts. Supplementary Table 19 said 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 find only 68.3% of the genes. 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 that 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 to the assembled sequence, as in the *PLoS* paper. On the unlikely chance that the data in GenBank is corrupt, we redid this assessment using the same 19,079 non-redundant full-length cDNAs (nr-KOME) [31]. If we ask that the genes be aligned in one piece, without fragmentation, both WGS are at least 91.2% complete. The other genes are not missing. All of 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 IRGSP, 98.1% of the genes are found.

Accuracy of the sequence assembly

The fact that essentially all of the genes are found in the WGS sequences does not prove that they are correctly assembled. IRGSP addressed this issue with a comparison to Syngenta *japonica* on the first megabase of chromosome 1, 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 there is a remarkable agreement in the positions of the 2685 genes on this chromosome. IRGSP reported on the presence of 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 is perfect. They did not, for example, mask repetitive sequences in assembling the shotgun data from their bacterial-artificial-chromosomes (BACs). In 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. 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 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 labs, as was the case in the IRGSP, can lead to inconsistent quality standards. We did a string search for unfinished gaps (*i.e.* strings of 50 or more N's) and found 575 examples, with 287 in chromosome 11 alone, or 288 times worse than the best chromosome. To assay single-base error rate, we compared IRGSP to the most reliable bases in Syngenta *japonica* (*i.e. RePS* estimated error rate less than 10^{-4}). The overall discrepancy rate is 0.020%, but for chromosome 4 it is 0.040%, or 3.2 times worse than the best chromosome. Just for fun, we combined quality measures to generate the chromosome rankings of Table 1.

Reason for gene count discrepancy

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 *PLoS* analysis, 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. 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 prior to the gene predictions. Applying this procedure to Syngenta *japonica* gave us 38,133 genes directly, comparable to the 37,544 genes reported by IRGSP. More detailed analysis (unpublished data) 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 paper to explain how *ReAS* is incorporated into the gene predictions; but for the interested readers, our updated gene sets are available at http://rice.genomics.org.cn/rice/link/download.jsp.

Cost-benefits for finished sequence

Let there be no mistake that we acknowledge 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 data released in *PLoS* 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 [34], it requires ~40-fold less reagents and ~10-fold less personnel effort, while producing results of very 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 commend IRGSP for at least trying 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 wish to 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

1. Yu, J. *et al*. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**, 79-92.

2. Goff, S.A. *et al*. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100.

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

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

5. Venter, J.C. *et al*. (1998) Shotgun sequencing of the human genome. *Science* **280**, 1540-1542.

6. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

7. Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* **291**, 1304-1351

8. Wong, G.K. *et al*. (1997) Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proc. Natl. Acad. Sci. USA* **94**, 5225-5230.

9. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

10. Waterston, R.H. *et al*. (2002) On the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA* **99**, 3712-3716.

11. Myers, E.W. *et al*. (2002) On the sequencing and assembly of the human genome. *Proc. Natl. Acad. Sci. USA* **99**, 4145-4146.

12. Waterston, R.H. *et al*. (2003) More on the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA* **100**, 3022-3024.

13. Adams, M.D. *et al*. (2003) The independence of our genome assemblies. *Proc. Natl. Acad. Sci. USA* **100**, 3025-3026.

14. Waterston, R.H. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562

15. Lindblad-Toh, K. *et al*. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819.

16. Olson, M.V. and Green, P. (1998) A quality-first credo for the Human Genome Project. *Genome Res.* **8**, 414-415.

17. Weber, J.L. and Myers, E.W. (1997) Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401-409.

18. Green, P. (1997) Against a whole-genome shotgun. *Genome Res.* **7**, 410-417.

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

20. Istrail, S. *et al*. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101**, 1916-1921.

21. SanMiguel, P. *et al*. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765-8.

22. Bennetzen, J.L. (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**, 1021-1029.

23. SanMiguel, P. *et al*. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43-45.

24. Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404-12410.

25. Feltus, F.A. *et al*. (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* **14**, 1812-1819.

26. Shen, Y.J. *et al*. (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* **135**, 1198–1205.

27. She, X. *et al*. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930.

28. Gaut, B.S. *et al*. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. USA* **93**, 10274-10279.

29. Wu, C. *et al*. (2004) A BAC- and BIBAC-based physical map of the soybean genome. *Genome Res.* **14**, 319-326.

30. Lander, E.S. and Waterman, M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239.

31. Kikuchi, S. *et al*. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301**, 376-379.

32. Bennetzen, J.L. *et al*. (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7**, 732-736.

33. Li, R. *et al*. (2005) ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.* **1**, e43.

34. Blakesley, R.W. *et al*. (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res.* **14**, 2235-2244.

Figures and Tables

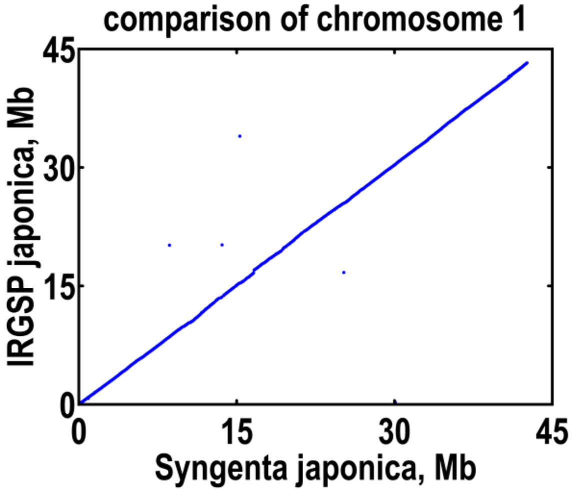

Figure 1: Map positions of 2685 nr-KOME cDNAs on chromosome 1, comparing IRGSP to the Syngenta *japonica* WGS assembly.

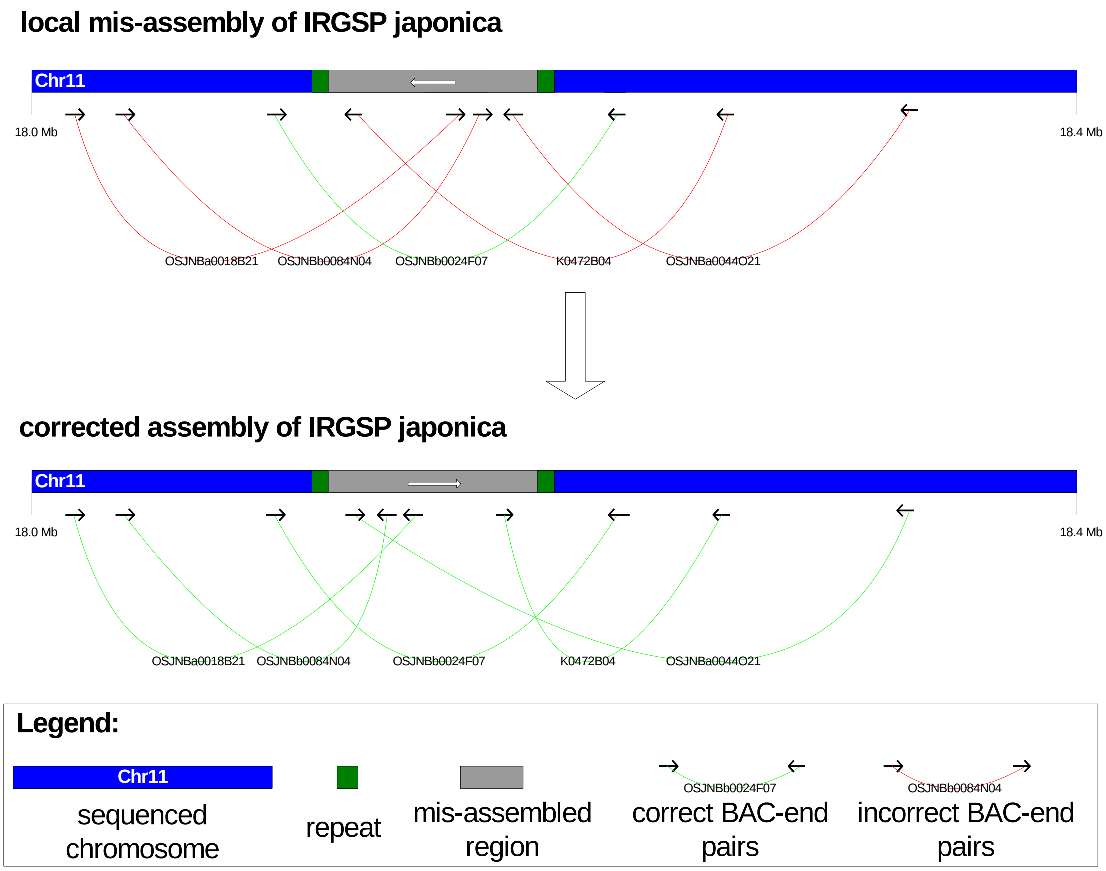

Figure 2: A local misassembly on IRGSP *japonica* chromosome 11. This 80.4-kb region contains two nr-KOME cDNAs (AK108542, AK066746) that lie in opposite directions on the map-based and WGS assemblies. IRGSP is contradicted by four pairs of BAC-ends (OSJNBa0018B21, OSJNBb0084N04, K0472B04, OSJNBa0044O21). A simple flip will fix everything. The problem is due to a pair of unknown transposons that create flanking inverted repeats of size 5.3-kb and 95% identity.

Table 1: Ranking of IRGSP chromosomes by quality. Q40 bases refer to the highest quality regions in Syngenta *japonica*, where error rates are better than 10^{-4}. These are then compared with IRGSP. Ngaps refer to strings of 50 or more N's that are found in IRGSP. For each chromosome, we compute ratios relative to the genome-wide mean. Overall ranking is based on a sum of ratios.

	Country	Size Mb	Q40 bases		strings of N		Rank
			different	vs mean	Ngap/Mb	vs mean	
CHR02	Japan	36.0	0.012%	0.630	0.25	0.15	0.78
CHR01	Japan	43.3	0.014%	0.729	0.16	0.10	0.83
CHR06	Japan	30.7	0.015%	0.743	0.22	0.13	0.88
CHR12	France	27.6	0.017%	0.847	0.07	0.04	0.89
CHR10	U.S.	22.7	0.023%	1.176	0.04	0.02	1.20
CHR09	Japan, Thailand, Korea, Brazil	22.7	0.015%	0.745	0.75	0.45	1.20
CHR03	U.S., Japan, France, China	36.2	0.016%	0.808	0.97	0.59	1.40
CHR07	Japan	29.6	0.019%	0.968	1.35	0.82	1.78
CHR05	Chinese Taipei	29.7	0.016%	0.813	1.72	1.04	1.85
CHR08	Japan	28.4	0.017%	0.854	2.34	1.42	2.27
CHR04	China	35.5	0.040%	2.036	1.81	1.10	3.14
CHR11	France, U.S., India	28.4	0.032%	1.652	10.12	6.14	7.79
	Average		**0.020%**		**1.65**		
	Max/Min		**3.23**		**288**		