Opinion

On the importance of being finished

Lance E Palmer and W Richard McCombie

Address: Genome Research Center, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard, Woodbury, NY 11797, USA.

Correspondence: W Richard McCombie. E-mail: mccombie@cshl.edu

Published: 27 September 2002

Genome ***Biology*** 2002, **3(10)**:comment2010.1–2010.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/10/comment/2010

Abstract

The publication of an increasing number of draft genome sequences presents problems that will only be resolved by improved search tools and by complete finishing of the sequences - and their deposition in publicly accessible databases.

A growing number of draft genome sequences are being generated [1-3]. This has largely resulted from the realization by sequencing groups and the community at large that such drafts provide much of the information that is available within a genome at a greatly reduced cost compared to a fully finished sequence. Recently, draft sequences of the genomes from two different strains of rice were published by Syngenta (*Oryza sativa* L. ssp. *japonica*) [4] and the Beijing Genomics Institute (BGI; *Oryza sativa* L. ssp. *indica*) [5]. While these sequences provide a wealth of data for researchers, they also point to the limitations of draft sequence versus complete sequence. This article seeks to shed light on these limitations and the problems they may create, as well as to discuss the limitations of less than complete submission of sequences to public repositories.

The limitations of draft sequence can be grouped into three main areas. First are the problems relating to the incompleteness of the genome sequence. Second are problems relating to the discontinuity of the data. And third are problems caused by the greater likelihood of errors in a draft sequence.

Incompleteness of the genome sequence

Syngenta reported that they had assembled a total sequence of length 389,809,244 base pairs (bp; making 93% of a predicted 420 Mbp rice genome) [4], while the BGI assembly length was reported to be 361 Mbp (out of a predicted 466 Mbp genome) [5]. A number of techniques were used to measure the completeness of the respective assembled sequences. Both groups matched known rice genes (with evidence from assembled expressed sequence tags (ESTs) or cDNAs) to the assembled sequence. Approximately 92% of the length of all genes were found in the BGI assembly and 99.2% of the genes searched by Syngenta were found in their assembly. The major caveat to this approach, however, is the extensive amount of gene duplication within rice. Goff *et al.* found that approximately 60% of 2,000 cDNA markers could be mapped to more than one locus; they also found that of 25,728 genes found on 791 contiguous sequences (contigs) assembled from clones within bacterial artificial chromosomes (BACs), the fraction of locally duplicated genes ranged from 15.4% to 30.3%. It is thus not clear whether each of the genes matched in the coverage analysis by Syngenta and the BGI was a true ortholog of the query cDNA or if a number of supposed matches were in fact simply highly conserved paralogs. The problem this creates is that one can not be entirely certain that a BLAST search will correctly identify the true ortholog of the query.

In addition to potentially lacking known genes, the assembled contigs from the Syngenta and BGI draft sequences will not contain a number of features that are usually screened out during the draft assembly process but would be incorporated in a final functional sequence: these include repeat sequences and any potential organellar DNA inserts. To prevent misassemblies, sequencing reads containing repeat sequences were either masked or completely removed before assembly into contigs. The BGI assembly is estimated to have excluded an equivalent of 78 Mb of fully masked reads and 26 Mb of partially masked reads from the assembly, while the assembled Syngenta data excluded an equivalent

of 38 Mb of repeats from the 390 Mb of assembled sequences. Also screened from assemblies are chloroplast- and mitochondrion-related sequences, but reads with sequence similarity to the mitochondrial and chloroplast genomes may have derived from insertions of organellar DNA into the nuclear genome. This is entirely possible, as a large portion of the mitochondrial genome of *Arabidopsis* has been inserted into *Arabidopsis* chromosome 2 [6].

Discontinuity of the data

To study the problems that may arise from discontinuity, we examined the sequence of the Indica strain of rice from the BGI [5]. BGI divided the draft rice genome into 127,550 contigs with an N50 size of 6.69 (that is, 50% of nucleotides are in contigs of 6.69 kb or larger). These contigs were assembled into 103,044 scaffold sequences (N50 of 11.76) using clone-end pairing information. Examination of the contig and scaffold sizes, however, revealed that 22.8% of the entire contig length and 18.2% of the entire scaffold length are comprised of contigs and scaffolds that are less than 2,000 bp long; this is less than half the mean gene size of rice (4,500 bp) as reported by the BGI [5], and suggests that many genes may not be represented on a single contig.

To determine what percentage of known proteins are not encoded on a single contig, we analyzed 100 rice proteins in the SWISS-PROT database, using BLASTP [7] searches against a set of proteins predicted from the BGI contig sequences using Fgenesh [8]; TBLASTN searches of the SWISS-PROT proteins against the contigs were also performed, in case Fgenesh failed to predict the corresponding protein. We found that 33 out of 100 SWISS-PROT proteins analyzed were not encoded on a single contig (see Additional data file 1 for a list of the proteins analyzed). This is consistent with the finding of Yu *et al.* [5] that out of 75,659 predicted genes 22,261 (29%) are incomplete (do not contain an initial and a terminal exon). This presents a number of problems when doing automated analysis on a set of predicted proteins.

The major caveat is that in homology searches against a set of predicted proteins, a full-length paralog will be likely to have a higher score than an ortholog that is split among multiple contigs; this occurred in 17 of the 33 SWISS-PROT proteins that were not encoded on a single contig, and an example is shown in Figure 1. The ACT1 (actin 1) protein from the SWISS-PROT database was queried against a set of Fgenesh predicted proteins from the BGI data. The top hit was the protein encoded by the first predicted gene from contig 17097; the two proteins are 94% identical. But a better match is split into two contigs: contig 23050 contains the amino-terminal half of ACT1 (99% identical; Figure 1b). While the Fgenesh prediction matches only up to amino acid 210, a TBLASTN search against contig 23050 showed that it has similarity to the ACT1 sequence up to amino acid 243 (Figure 1c). The carboxyl terminus of ACT1 is contained in contig 2561 (96% identical). Figure 1d shows a TBLASTN search of ACT1 against contig 2561; the similarity between ACT1 and the two contigs (23050 and 2561) extends to within three nucleotides of the ends of the contigs. This analysis shows that because the orthologous actin 1 gene is split amongst two contigs, one can incorrectly identify a paralog as an ortholog.

The fact that one can incorrectly identify a true ortholog is an important factor to take into consideration when performing automated analysis that relies on using predicted protein sequences from whole-genome shotgun sequence data. This is particularly important in the case of rice, where there is extensive duplication of genes.

The greater likelihood of errors in a draft sequence

Errors in assembled genomes can occur for a number of reasons, including sequencing and misassembly errors. Both kinds of error can create problems for gene prediction programs and subsequent analysis. For example, the BGI reported [5] that 94.2%, 90.8% and 83.5% of their sequence had error rates (based on Phred/Phrap estimates) better than 10^{-2}, 10^{-3} and 10^{-4} respectively. These values improved to 97.3%, 96.1% and 92.5%, respectively, when only contigs larger than 3 kb were used in the analysis and 500 bases from the ends of contigs were trimmed to remove lower quality sequence. This applies to only 221 Mb (61%) of BGI's contig sequence, however. When Syngenta compared their sequence to six finished rice BACs, they found that there was, on average, a single base pair difference once out of every 1,000 bases and an insertion/deletion difference once out of every 2,000 bases.

We attempted to simulate the effect of draft-quality sequencing data on the ability to predict encoded genes and protein domains. We took 75 randomly chosen finished rice BACs or phage artificial chromosomes (PACs) from GenBank for *in silico* mutagenesis (see Additional data file 2 for a list of BAC accession numbers and results of this analysis). Each base was given a 1 in 1,000 chance of having a base substitution and a 1 in 2,000 chance of having a 1-nucleotide insertion placed before it. While these conditions do not exactly match the differences between draft and finished sequence, they can give us an approximation. Both the 'mutagenized' and unmutagenized BAC sequences were run through the Fgenesh gene prediction program [8], and gene products predicted to contain transposon and retrotransposon elements were eliminated from the set of predicted proteins. We found that the total number of Fgenesh-predicted genes was similar between normal and *in silico* mutagenized samples (1205 versus 1193). The total number of domains predicted by HMM-Pfam [9] was also similar (522 versus 499 in the mutagenized sample).

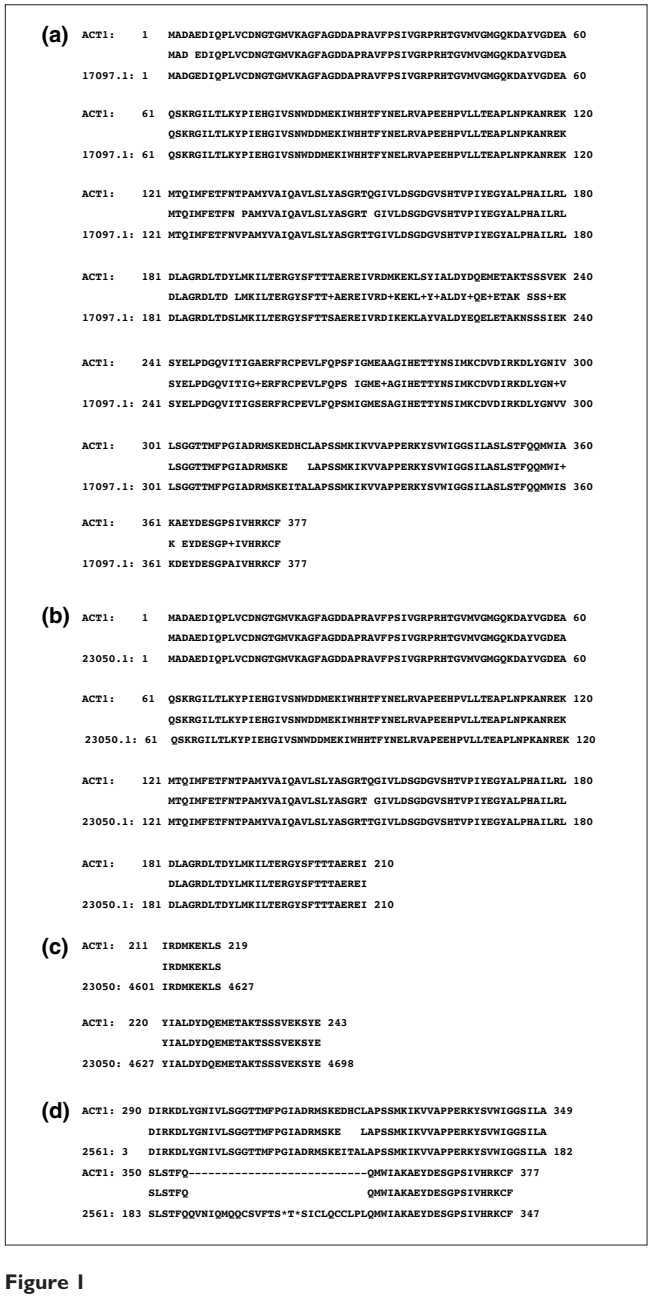

Figure 1
Complete paralogs have higher scores in homology searches than incomplete orthologs. The ACT1 (actin 1) protein from rice was used as a query in a BLASTP search against a set of Fgenesh-predicted proteins from *Oryza sativa* L. ssp. *indica* (see text for further details). **(a)** The top match was the first predicted protein from contig 17097, but this predicted protein is likely to be a paralog of ACT1 as ACT1 also matched to a protein encoded on contig 23050 with a lower score **(b)**, but with a higher degree of similarity in the region that did match (amino acids 1-210). **(c)** A TBLASTN search revealed that amino acids 211-243 are encoded on contig 23050. **(d)** The TBLASTN search also revealed that the carboxyl terminus (amino acids 290-377) of ACT1 is encoded on contig 2561.

Although the total number of predicted genes and domains did not change dramatically, analysis of individual predicted proteins showed that many gene predictions were altered with the 'draft quality' data. For example, 12% of all predicted genes were sufficiently disrupted in the mutagenized sample that less than 50% of the length of their sequence could be found to correspond to a single predicted protein using a FASTA [10] search. And 23% of all predicted proteins had at least a minor disruption, such that less than 90% of a predicted protein's length could be found encoded in a single sequence in the mutagenized sample.

So, draft-quality sequence does provide useful information on the number of genes and the number and types of domains, but analysis of specific genes may be flawed because of inaccurate sequencing information. These problems can be somewhat alleviated by using TBLASTX or TBLASTN searches against the whole genome instead of BLASTP searches against a set of predicted proteins. If a gene with a frame-shift mutation is found in TBLASTX or TBLASTN searches, however, it would be impossible from the sequencing data alone to determine whether the mutation is real or simply a sequencing error.

Are the problems of incompleteness, discontinuity and errors that are apparent in the rough draft sequence of rice also problems for other genomes? This will depend on how complete each genome sequence is and how repetitive the genome is. For other grasses, such as wheat and sorghum, these complications will be important to consider. For less repetitive genomes, the problems exist but to a lesser extent, so they will probably not affect genome-wide analysis but do pose problems for analyzing specific genes, especially when trying to identify true orthologs. When doing BLAST searches with unfinished genome projects that may be in large number of contigs, one should be cautioned to not rely on the top BLAST hit. BLAST scores are ordered by e-value (or Expect value - the number of matches with the same score that are expected to occur by chance [11]). Longer matches with slightly lower percentage identity may have more significant e-values, so a full-length paralog may appear as a more significant match in a BLAST search. Perhaps our search tools for analyzing draft sequences need to be modified slightly to report scores by percent identity (assuming a minimum length) or to generate possible orthologs based on matches to one or more contigs.

Data access issues

Both Syngenta [4] and the BGI [5] have stated that their data will be available to the academic community. For academic users, Syngenta sequence can be queried through a web browser [12] or can be obtained by CD-ROM after a public-access agreement is submitted. The drawback to the web-based sequence retrieval is that it is limited to 100 kb per week. The BGI data can be queried or retrieved over the web [13] and includes masked reads, contigs and scaffold sequences. The scaffold sequences have been deposited into

GenBank (project accession number AAAA00000000) [14], but we have not found a way to perform BLAST searches using the BLAST services of GenBank.

Although the difficulties of accessing the data may not be a major inconvenience for bioinformaticians and researchers interested in rice genomics, the average biologist is not likely to be aware of any publicly accessible sequence database outside of GenBank and perhaps a few other specialized databases in their field of interest. Thus, although useful for the limited application of finding genes in a given species, data release on individual websites is extremely limiting. It has a major impact on cross-species analysis and on utilization of the data by those in other fields. This is becoming ever more important in biology as genome sequence provides a common currency for researchers to move from species to species with their inquiries. The days of looking at a single gene in a single species as a major focus of biological research are rapidly drawing to an end. Compartmentalized data release greatly inhibits the process of accessing data for the next generation of studies, and is also problematic for bioinformaticians. It is difficult to include such data in the automated analysis pipelines so crucial for much of current bioinformatics work, for example.

If the trend of not submitting sequence data to GenBank continues, it will be even more difficult for researchers to keep up with the locations of web pages for various genomes. Instead of doing a single BLAST search against a centralized database such as GenBank, researchers might end up performing numerous searches over the web. This has the potential of turning what should be a golden age of genomics into something far less - an age limited by what could be viewed as a bioinformatics Tower of Babel, in which a wealth of data cannot be integrated and fused to provide understanding of complex biological processes.

Additional data files

Additional data file 1, available with the online version of this article, provides the list of 100 proteins from the SWISS-PROT database that were used to estimate the proportion of protein-coding genes carried on a single contig. Additional data file 2 provides the list of 75 finished BACs and PACs that were used for *in silico* mutagenesis to simulate draft-quality sequence, and the results of their analysis.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
2. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
4. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*).** *Science* 2002, **296:**92-100.
5. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).** *Science* 2002, **296:**79-92.
6. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, *et al.*: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402:**761-768.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
8. Salamov AA, Solovyev VV: ***Ab initio* gene finding in *Drosophila* genomic DNA.** *Genome Res* 2000, **10:**516-522.
9. Eddy SR: **HMMER: Profile hidden Markov models for biological sequence analysis (2001)** [http://hmmer.wustl.edu/]
10. Pearson WR, Lipman DJ**: Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85:**2444-2448.
11. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264-2268.
12. **Torrey Mesa Research Institute (TMRI) - the rice genome project** [http://portal.tmri.org/rice/]
13. **Rice GD: genome database of Chinese super hybrid rice** [http://btn.genomics.org.cn/rice]
14. **GenBank** [http://www.ncbi.nih.gov/GenBank]