BIOINFORMATICS LETTER TO THE EDITOR Vol. 21 no. 24 2005, pages 4320–4321
doi:10.1093/bioinformatics/bti769

Beware of mis-assembled genomes

Steven L. Salzberg[1,*] and James A. Yorke[2]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA and [2]Institute for Physical Sciences and Technology, University of Maryland, College Park, MD 20742, USA

With hundreds of genomes now in GenBank, researchers might be forgiven for assuming that genome sequence data are correct, at least at a large scale. Certainly there might be errors at some small rate, perhaps 1 in 50 000 or 100 000 bases (Schmutz *et al*., 2004; Read *et al*., 2002), but at a large scale these genomes are put together correctly, are not they? Well, not always.

We have been looking at the assemblies of large genomes for several years now, and for every 'draft' genome we look at, we find hundreds—and sometimes thousands—of mis-assemblies. These include regions where a genome is incorrectly re-arranged as well as places where large chunks of DNA sequence are simply deleted and the surrounding sequences just crunched together.

The source of most mis-assemblies is, as it has always been, repeats. Genomes vary in their repeat content, but we have learned that large genomes are filled with repeats of all shapes and sizes. To illustrate how these repeats result in sequences being 'lost' by an assembler, consider the situation in Figure 1.

In the figure, we see that the genome has two copies, R1 and R2, of a sequence that lie near one another, separated by a unique region shown in red. If R1 and R2 are long enough, then the assembler will not have any individual sequences ('reads') containing the entire repeat and its unique flanking sequences (the green and blue regions). The result will be that the genome assembly looks like the lower half of the figure, with a contiguous stretch of DNA (a contig) that has just one copy of the repeat, incorrectly jamming together the blue and green regions, and the red region will have no place to go.

If this seems like a made-up example, it is not: we have observed that even the best assemblers today make exactly this mistake when assembling the *Drosophila* species currently being sequenced. Compressions such as this can easily total 1% or more of the genome, and the 'orphan' regions can be quite long, 5000–10 000 bp or more. And we would note that *Drosophila* is not a particularly difficult genome as compared with many others currently under way. To those who might think (or argue) that the assembler they are using is not prone to such errors, we can only reply that we have seen these types of errors in all the major assemblers in use today (e.g. Arachne (Batzoglou *et al*., 2002; Jaffe *et al*., 2003), Celera Assembler (Myers *et al*., 2000), Jazz (Aparicio *et al*., 2002), Phusion (Mullikin and Ning, 2003), PCAP (Huang *et al*., 2003) and Atlas (Havlak *et al*., 2004)), in some cases after running the assemblers ourselves and in other cases after carefully examining the results of assemblies created by others.

We have developed software for improving assemblies that can detect at least some situations like the one shown above, although there is still no automated way of fixing these problems. However, the problem is often made much more difficult by the diploid nature of most large genomes, particularly the many mammalian genomes currently being sequenced by the NIH. The problem is this: the two copies of a chromosome are always slightly divergent, and this has led assembly groups (including ours) to develop methods for separating the two haplotypes from one another. But wherever there are tandem repeats in two or more copies, it can become extremely difficult to distinguish an incorrectly collapsed repeat (including situations such as that shown in Fig. 1) from true polymorphisms between the haplotypes.

A tremendous amount of genome analysis is built upon the framework of the DNA sequence itself: not only are genes and regulatory sites anchored in the sequence, but analyses of synteny, duplications and evolutionary relationships among species all depend on having the correct structure of the genome. We need to devote more effort to making sure the basis for all these analyses does not turn out to be a house of cards. Our group has created a website (http://cbcb.umd.edu/research/benchmark.shtml) for depositing reference assemblies: genomes for which the sequence is finished, and for which we can demonstrate how all the original data map to that finished sequence. The site also distinguishes the original whole-genome shotgun reads from any additional finishing reads. This small set of genomes, which thus far only includes bacteria, should be just the beginning: all assemblies need to be available so that others can check them and, if necessary, correct them. Fortunately, NCBI has created a much larger resource to capture both draft and finished assemblies, the Assembly Archive (Salzberg *et al*., 2004). This archive captures the complete information about how a set of raw sequences maps to a genome assembly, whether that assembly is 'draft' or 'finished'. After spending fifteen years and hundreds of millions of dollars on the human genome, the community has a near-complete draft sequence, but the evidence for that sequence—the underlying raw data and the assembly itself—is, amazingly, not available. Indeed, many of the original assemblies of parts of the human genome were done in the mid- and late-1990s, and are now lost. We can only hope that future genomes would not be needlessly lost now that there is a place to deposit them.

Are we arguing that all genomes should be finished? Actually, finishing does not necessarily address this problem at all. Finishing efforts are usually directed at closing gaps, not at fixing mis-assemblies, and therefore 'finished' genomes are very likely to contain errors of the type we are discussing. A better term for such genomes is 'closed': gaps are closed but sequence is not confirmed. We strongly suspect that many of the already-published finished genomes in GenBank today contain assembly errors.

*To whom correspondence should be addressed. E-mail: salzberg@umd.edu

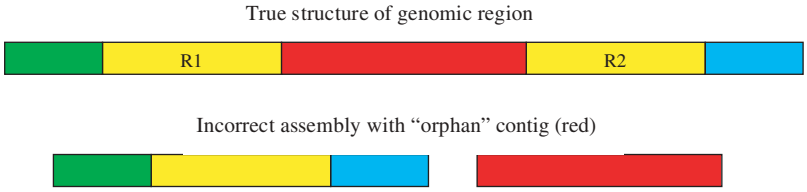

Fig. 1. Assemblies can collapse around repetitive sequences. R1 and R2, in yellow, represent near-identical copies of the same DNA sequence.

Clearly we also need new, well-defined methods for comparing assemblies. The most popular metrics right now all seem to emphasize size: size of contigs, size of scaffolds, and especially N50 sizes. (The N50 size is computed by sorting all contigs from largest to smallest and by determining the minimum set of contigs whose sizes total 50% of the entire genome. The N50 size is the smallest contig in that set.) The standard of judging assembly quality by size of contigs is questionable. Large contigs may simply reflect overly aggressive joining of contigs, thereby creating larger contigs with mis-assemblies. As a consequence, genome scientists who are not experts at assembly can be completely misled by statistics about contig sizes, and as a result might prefer the 'larger' but incorrect assembly when given a choice.

We need to start capturing assemblies and looking at them with a more skeptical eye. This need has become even greater in the face of a growing number of 'draft' assemblies, many of which will never be finished. Before launching lengthy projects based on these genomes, we need to be confident that they are assembled correctly. The bioinformatics community should take the lead in this effort, by developing standards for quality control and by devoting more time and energy to careful evaluations of genome assemblies.

REFERENCES

Aparicio,S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu* rubripes. *Science*, **297**, 1301–1310.

Batzoglou,S. *et al*. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res.*, **12**, 177–189.

Havlak,P. *et al*. (2004) The Atlas genome assembly system. *Genome Res.*, **14**, 721–732.

Huang,X. *et al*. (2003) PCAP: a whole-genome assembly program. *Genome Res.*, **13**, 2164–2170.

Jaffe,D.B. *et al*. (2003) Whole-genome sequence assembly for Mammalian genomes: ARACHNE 2. *Genome Res.*, **13**, 91–96.

Mullikin,J.C. and Ning,Z. (2003) The PHUSION assembler. *Genome Res.*, **13**, 81–90.

Myers,E.W. *et al*. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Read,T.D. *et al*. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033.

Salzberg,S.L. *et al*. (2004) The genome assembly archive: a new public resource. *PLoS Biol.*, **2**, E285.

Schmutz,J. *et al*. (2004) Quality assessment of the human genome sequence. *Nature*, **429**, 365–368.