Assemblies: the good, the bad, the ugly

Ewan Birney

The low cost of short-read sequencing has motivated the development of *de novo* assemblies from only short-read data; impressively, assemblies for large mammalian genomes are now available. However, this is still a developing field, and these *de novo* assemblies have many artifacts, as do all *de novo* assemblies.

In this issue, Can Alkan, Saba Sajjadian and Evan Eichler[1] provide an in-depth critique of two new *de novo* assemblies of the human genome[2] by comparing them to the human reference sequence. These 'next-generation assemblies' had been generated from high coverage data consisting of short, paired-end reads (36–50 base pairs, spanning insert sizes of 200–10,000 base pairs)[2]. As someone who works in the genome assembly field, I believe that just the fact that these assemblies are feasible and correct for the majority of their bases (over 99% of the reported bases are correct in a local context) is an impressive algorithmic and, in particular, engineering success. Five years ago such data would have been considered simply impossible to assemble because the short-length reads did not allow the previous overlap-based assembly methods to work sensibly[3]. The continued exponential decrease in cost for short-read, high-throughput sequencing means that genomes of many more species can be sequenced. However, Alkan *et al.*[1] provide an appropriate reminder of the large and complex artifacts that occur in nearly every assembly and, in particular, in this generation of assemblies from short-read data.

Alkan *et al.*[1] comprehensively lay out deficiencies in the assemblies, some of which had been clearly mentioned in the original paper[2], whereas others had been mentioned only in supplementary materials or not noted. None of the deficiencies—a reasonably high rate of missing sequences, contamination, order errors, orientation errors and systematic collapse of segmental duplications—would surprise anyone involved in generating large-scale genome assemblies and analyses. Each of these artifacts is also present to varying extent in 'traditional' long read–based whole-genome shotgun assemblies. Unsurprisingly, given the technology and 'youth' of the algorithms, new assemblies certainly seem to have more deficiencies than the state-of-the-art traditional assemblies, though the complexity of assessing assemblies and the lack of a perfect assembly in large genomes makes any quantification of relative error rates difficult. Scientific papers on previous draft assemblies in other species were not always completely thorough in discovering, characterizing and appropriately highlighting artifacts in the sequence[4]. If there is one message to remember from the Perspective by Alkan *et al.*[1], it is to not fully trust any assembly, in particular one relying on new technology early on in the development of new methods.

One should be especially cautious about the absence of a particular sequence or gene, which can easily be an assembly artifact rather than a genuine lineage-specific deletion. Also, the analysis of segmental duplications, a particular focus of research in the Eichler laboratory, has to be very carefully performed, preferably directly on raw read data rather than on assemblies. Knowing these potential pitfalls in the assemblies allows one to be more confident in using these *de novo* assemblies for other aspects, such as for determination of the precise exonic sequence of single-copy genes or as a framework in which to generate polymorphic sites for genetic analysis.

It is important to chart a path forward for the genome-sequencing community, genome analysts and the downstream biology community for handling next-generation assemblies. One clear aspect is that genome assembly is not a solved problem. New sequencing technologies may provide better cost profiles for data generation but do not change some of the extremely hard problems in generating end-to-end correct sequence. This clearly requires the development of both new experimental technologies that create, for example, longer reads and better computational algorithms to use this information optimally. There are some very promising new approaches for creating assembly algorithms in development, in particular from the group of David Jaffe at the Broad Institute (personal communication), as well as from the developers of the assemblies critiqued by Alkan *et al.*[1] at the Beijing Genome Institute and from other groups worldwide. It seems likely that there will be solid improvement over the coming years resulting from innovation in both experimental and computational methods.

Alkan *et al.*[1] also highlight the need to not dismiss older technologies, such as large-insert cloning; innovative ways of using, for example, older cloning techniques with next-generation sequencing will hopefully result in continued use of these technologies. As with many areas of science, there is a group of innovative, accomplished experimentalists whose skills might be dismissed if *de novo* genome assembly is considered a solved problem; new techniques will come from both the 'wet' and 'dry' components of the process. But to succeed these scientists must be funded, in particular to approach the complex, as-yet unresolved areas of the

Ewan Birney is at the European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK.
e-mail: birney@ebi.ac.uk

genome or to improve existing (perhaps far less than optimal) assemblies. As the process of generating assemblies has become less fashionable, the pool of scientists interested in performing them has naturally become smaller. This means that more careful thought has to be made by funding agencies to ensure appropriate progress.

A particularly hard task is to assess the cost-to-benefit ratio of different assembly improvements. It is clear that a 'platinum-grade' error-free, complete sequence of one individual's genome is currently both technologically unfeasible and too costly. Nevertheless, at least for the human genome, in which potentially every base pair might be involved in disease and for which there will be millions of researchers performing experiments over the next 50 years, error-free, complete sequence should remain our goal. As Eichler and colleagues emphasize, this will require both experimental and computational innovation[1]. Coupled to this, there should be an effective, sequence-level resolution of the majority of the structural variants in human populations. Projects such as 1,000 genomes are a starting point for this analysis[5], and organizations such as the Genome Reference Consortium are the practical implementation of parts of that goal. However, the devil is in the details of both the science of making good assemblies and good processes for funding them.

Given the fundamental importance of genome assemblies for many aspects of biology, the critique from Alkan *et al.*[1] provides a clear reminder of the complexity and artifacts in assemblies and is worth reading for any biologist working with genomic data. This critique, in my view, should not be seen as criticism of the members of the Beijing Genome Institute team, who developed these first large-scale *de novo* assemblies using only short reads—but rather an indication of the need for continued development in this area. However, every scientist should have an appropriate skepticism of all analyses, in particular at genome-wide scale, and must critically examine any conclusions, in particular surprising ones, drawn on assemblies of any type.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Alkan, C., Sajjadian, S. & Eichler, E.E. *Nat. Methods* **8**, 61–65 (2011).
2. Li, R. *et al. Genome Res.* **20**, 265–272 (2009).
3. Miller, J.R., Koren, S. & Sutton, G. *Genomics* **95**, 315–327 (2010).
4. Salzberg, S.L. & Yorke, J.A. *Bioinformatics* **21**, 4320–4321 (2005).
5. 1000 Genomes Project Consortium. *Nature* **467**, 1061–1073 (2010).