EMBO reports

Revisiting published genomes with fresh eyes and new data

Revising old sequencing data can yield unexpected insights and identify errors

David R Smith

Growing up, I hated leftovers. Nothing, in my teenage culinary opinion, was worse than sitting down to a big plate of last night's spaghetti. All the excitement and novelty of the meal had been sapped. And so, with my shoulders slouched, I would shovel soggy noodles towards my forlorn face, praying that tomorrow's dinner would bring something new.

> *". . . the thrill of research is in the unexplored [. . .] and old data, especially those that are already published, are the equivalent of a three-day-old dinner bun with a heaping scoop of expired peanut butter."*

Now that I am a biologist, I often feel the same way about old data as I once did about leftover food, which I am sure is a sentiment to which other scientists can relate. For many, the thrill of research is in the unexplored—the yet to be discovered—and old data, especially those that are already published, are the equivalent of a three-day-old dinner bun with a heaping scoop of expired peanut butter.

Perhaps we should not be so picky. Maybe more scientists should embrace their leftover, picked-through data, squeezing the last bit of insight out of them. Or, as my dad (the family chef) would have done, make them feel new again by throwing in some fresh bits and bobs. With respect to my own area of research—organelle genomics—I am beginning to realize that a lot can be learned by revisiting previously sequenced mitochondrial and chloroplast genomes (mtDNAs and ptDNAs).

Organelle genomics, a victim of its own success

The field of organelle genomics is almost as old as genomics itself. In fact, mtDNAs and ptDNAs were among the very first genomes to be completely decoded [1]. They are still one of the most commonly sequenced types of genome, with examples coming from every corner of the eukaryotic tree of life [2]. Together, these data have been instrumental in helping scientists understand the evolution and diversification of complex life, not to mention the pivotal roles of organelles in cell biology, biotechnology and medicine.

> *"Looking back at much older genome sequences, including those of well-studied model species, can yield unforeseen and important insights."*

Unfortunately, organelle genomics has also become a victim of its own success. The arrival of powerful sequencing technologies and user-friendly bioinformatics software during the past decade has made it easy to quickly sequence mtDNAs and ptDNAs en masse. Consequently, the scientific literature has been flooded with organelle genome papers, so much so that there are journals specializing in organelle genome reports—short (~500-word) articles presenting newly sequenced mtDNAs and ptDNAs and their GenBank accession numbers [3]. Although a quick avenue to a peer-reviewed publication, these miniature papers are usually lacking in scientific substance. In some cases, little effort is made to properly characterize the genomes that make up genome reports, meaning GenBank has a growing surplus of poorly annotated mtDNAs and ptDNAs [3].

For instance, the chloroplast genome of the green alga *Haematococcus lacustris*, which is currently the largest on record (1.35 Mb), was recently published in the journal *Genome Announcements* [4]. Sadly, the GenBank entry accompanying this paper lacked even the most straightforward annotations, such as ribosomal RNAs, and contained a number of mislabelled genes. I say this not to pick on the authors of this paper (we are now collaborators), but to highlight the current state of organelle genomics and to argue that it is time we stopped sequencing new genomes so hastily and started re-examining the data that are already available. Case in point: a reassessment of the *H. lacustris* plastome revealed that it has a non-standard genetic code, an unprecedented GC content, and a repetitive element that has also spread throughout the mtDNA [5]. It is not just data from genome reports that warrant revisiting. Looking back at much older genome sequences, including those of well-studied model species, can yield unforeseen and important insights.

When the reference is wrong

In the world of plant research, the angiosperm *Arabidopsis thaliana* and green alga

Department of Biology, University of Western Ontario, London, ON, Canada. E-mail: dsmit242@uwo.ca

DOI 10.15252/embr.201949482 | EMBO Reports (2019) 20: e49482 | Published online 3 November 2019

Chlamydomonas reinhardtii are model organisms of popstar status. As such, their organelle genome sequences have been available for years, been used as reference sequences in hundreds of studies and been updated on multiple occasions [6,7]. Therefore, it stands to reason that the *A. thaliana* and *C. reinhardtii* organelle DNA data should be relatively free of errors. Surprisingly, that is not true.

A recent study found persistent errors in the standard mitogenome reference sequence of *A. thaliana*—that is, from the Col-0 ecotype [6]. Using publicly available sequencing data, Sloan *et al* [6] showed that the Col-0 mtDNA in GenBank contained, on average, a sequencing error every 2.4 kb, including "57 single-nucleotide polymorphisms (SNPs), 96 indels (up to 901 bp in size) and a large repeat-mediated rearrangement" [6]. My goodness! And to think, this mtDNA was supposed to represent one the more highly polished organelle genomes. What's more, some of the errors in the Col-0 mtDNA have been carried over to the mtDNA sequences of other *A. thaliana* ecotypes through reference-based assembly approaches. Ultimately, these mistakes have misled subsequent studies on plant mitochondrial mutation by "giving the false impression that the errors are naturally occurring variants present in multiple ecotypes" [6]. Thankfully, the revised and now highly accurate mitochondrial genome of Col-0 can be found in GenBank under accession number BK010421.

> "... *if the widely used reference organelle DNAs of these two species were shown to contain numerous mistakes, what does that suggest about the quality of other available mtDNAs and ptDNAs...*"

In a similar turn of events, a team of Chlamydomonas researchers discovered a number of small and large errors in the reference mtDNA and ptDNA sequences of the most commonly used laboratory strain of *C. reinhardtii* (CC-503) [7]. Employing previously published data, they reassembled *de novo* the *C. reinhardtii* organelle genomes and identified dozens of SNPs and indel errors in both the mitogenome and the plastome, including a 2.4-kb inversion in the latter. By incorporating newly generated RNA-seq data into their reanalysis of these genomes, they demonstrated polycistronic organelle gene expression, quantified splicing of organelle introns and characterized cytosine-rich polynucleotide tails on mitochondrial transcripts [7]. Not bad for what were thought to be completed genomes.

If the re-examination of the *A. thaliana* and *C. reinhardtii* organelle genomes can teach us anything, it is that revisiting published genomes is a worthwhile endeavour and one that should be encouraged. Moreover, if the widely used reference organelle DNAs of these two species were shown to contain numerous mistakes, what does that suggest about the quality of other available mtDNAs and ptDNAs, particularly those from non-model species? It is probably best to assume that they contain errors. I would add that organelle genomes with complex architectures, which might make them prone to sequencing and/or assembly errors, should be approached with particular caution.

Revisiting the reference sequences of complex organelle genomes

One factor that likely contributed to the persistent errors in the *A. thaliana* mtDNA and *C. reinhardtii* ptDNA is the expanded architecture of these two genomes. Both are big (> 200 kb), bloated (> 50% non-coding) and repeat-rich, which made them challenging to accurately assemble in the era of low-coverage, Sanger-based sequencing [6,7]. Other organelle DNAs, however, can be much larger than those of *A. thaliana* and *C. reinhardtii*, and it is the reference sequences of these genomes that merit close reassessment.

For example, there are at least twenty sequenced land-plant mtDNAs longer than 700 kb, some of which were generated using low-coverage Sanger sequencing [8] and others based solely on short-read data from early next-generation sequencing [9]. Likewise, of all the available plastomes with lengths of more than 300 kb, more than half were assembled with only Sanger or short-read data. My sense is that resequencing and *de novo* assembly of these DNAs using long-read single-molecule real-time (SMRT) sequencing in conjunction with modern Illumina methods would uncover a large number of errors. Such an approach would also be useful for revising the slew of partially assembled mitochondrial and chloroplast genomes.

Indeed, many large mtDNAs and ptDNAs have proved so hard to assemble using short-read data that their published sequences are deposited in GenBank as a series of fragmented contigs [10]. These types of assemblies typically include a complete or near-complete coding repertoire, but the gene order and sequences of the intergenic regions are unresolved. Although the availability of long-read sequencing techniques means that it is now possible to bridge the gaps and complete these disjointed genomes, very few researchers seem interested in returning to their old data—myself included.

Apart from size, organelle DNAs can have other features that make sequencing and assembly difficult. For instance, certain species harbour linear mtDNAs with complex telomeres, which are renowned for being hard to characterize [2]. Organelle genomes can also be multipartite whereby the DNA sequence is distributed across a few or even hundreds of chromosomes [2]. These "unconventional" organelle genomes represent excellent candidates for resequencing, which in some cases might turn up unexpected insights, such as previously unidentified chromosomes.

> "*It is hard to finish a project, but it is even harder to revisit and revise one that is already completed.*"

My old and grizzled PhD supervisor used to say: "Smitty, it is always easier to start a project than it is to finish one". I would now add the following corollary to that statement. It is hard to finish a project, but it is even harder to revisit and revise one that is already completed. The rate at which we sequence genomes will only increase in the coming months and years, and as unprecedented amounts of data are deposited into GenBank, it is important that we not forget to look back and update older sequences. The immediate rewards of such work might seem small, but the long-term impact could be massive. It is high time we started devouring our genomic leftovers.

Acknowledgements

DRS is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F *et al* (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290: 457 – 465
2. Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci USA* 112: 10177 – 10184
3. Smith DR (2017) Goodbye genome paper, hello genome report: the increasing popularity of 'genome announcements' and their impact on science. *Brief Funct Genomics* 16: 156 – 162
4. Bauman N, Akella S, Hann E, Morey R, Schwartz AS, Brown R, Richardson TH (2018) Next-generation sequencing of *Haematococcus lacustris* reveals an extremely large 1.35-megabase chloroplast genome. *Genome Announc* 6: e00181-18
5. Zhang X, Bauman N, Brown R, Richardson TH, Akella S, Hann E, Morey R, Smith DR (2019) The mitochondrial and chloroplast genomes of the green alga *Haematococcus* are made up of nearly identical repetitive sequences. *Curr Biol* 29: R736 – R737
6. Sloan DB, Wu Z, Sharbrough J (2018) Correction of persistent errors in Arabidopsis reference mitochondrial genomes. *Plant Cell* 30: 525 – 527
7. Gallaher SD, Fitz-Gibbon ST, Strenkert D, Purvine SO, Pellegrini M, Merchant SS (2018) High-throughput sequencing of the chloroplast and mitochondrion of Chlamydomonas reinhardtii to generate improved *de novo* assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. *Plant J* 93: 545 – 565
8. Goremykin VV, Salamini F, Velasco R, Viola R (2008) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Mol Biol Evol* 26: 99 – 110
9. Kersten B, Rampant PF, Mader M, Le Paslier MC, Bounon R, Berard A, Vettori C, Schroeder H, Leple JC, Fladung M (2016) Genome sequences of *Populus tremula* chloroplast and mitochondrion: implications for holistic poplar breeding. *PLoS One* 11: e0147209
10. Hu Y, Xing W, Song H, Liu G, Hu Z (2019) Evolutionary analysis of unicellular species in Chlamydomonadales through chloroplast genome comparison with the colonial volvocine algae. *Front Microbiol* 10: 1351