RNA-Seq data: a goldmine for organelle research

David Roy Smith

Advance Access publication date 18 January 2013

Abstract

GenBank is bursting with eukaryotic RNA sequencing (RNA-Seq) results. These data are transforming our view of nuclear transcriptional architecture, but many scientists are ignoring a major component of the data: mitochondrial- and chloroplast-derived sequences. Indeed, organelle transcripts typically represent a significant proportion of the reads generated from eukaryotic RNA-Seq experiments. Here, I argue that these data are an excellent and untapped resource for investigating many aspects of organelle function and evolution.

Keywords: *gene expression; mitochondrial genome; next-generation sequencing; plastid genome; transcriptomics*

Next-generation sequencing has reshaped how we address questions at the genomic level. Recently, I argued that the data coming from nuclear genome sequencing projects could also be used to explore organelle genomes [1]. In fact, eukaryotic next-generation sequencing results are riddled with mitochondrial and, for plants and algae, chloroplast sequences [1,2]. Here, I want to draw attention to another untapped resource: RNA sequencing (RNA-Seq) data. Researchers around the world are using high-throughput technologies, such as Illumina and 454, to sequence the transcriptomes from thousands of diverse species, and are depositing the results from these experiments in public online databases, such as the National Center for Biotechnology Information Sequence Read Archive (SRA) [3]. There are over 20 000 RNA-Seq projects from eukaryotes in the SRA, and soon there will be even more. The National Center for Genome Resources (NCGR) in collaboration with the Gordon and Betty Moore Foundation is sequencing the transcriptomes from approximately 750 marine unicellular eukaryotes [4], and an international collaboration headed by Drs Gane Ka-Shu Wong and Michael Deyholos from the University of Alberta is doing the same for 1000 different plant species [5]. The raw Illumina reads generated from both of these projects will be deposited in the SRA. Eukaryotic RNA-Seq data are revolutionizing nuclear genome sequence analysis and transforming our view of nuclear transcriptional architecture [6], but another major component of the data—organelle-derived transcripts—should not be overlooked.

The high copy number and elevated expression levels of organelle genomes mean that their transcripts represent a significant proportion (up to 25%) of the reads generated from RNA-Seq experiments [7,8]. And because organelle transcripts are typically AT rich and often polyadenylated their contribution to the overall number of RNA-Seq reads has been shown to go up with increased poly-A RNA selection; moreover, their concentrations also remain high when employing other types of enrichment protocols, such as ribosomal RNA depletion [8]. My collaborators and I recently received from the NCGR the RNA-Seq results for one of the marine algae that we nominated for transcriptome sequencing: *Pyramimonas parkeae* (project ID MMETSP0059). As expected, the data are teeming with organelle-derived transcripts, so much so that almost complete mitochondrial and plastid transcriptomes (and genomes) could be assembled from the reads.

In addition to providing information on organelle gene expression, RNA-Seq data can give insight into a variety of other questions. Organelle transcription

Corresponding author. David Roy Smith, Department of Botany, Canadian Institute for Advanced Research, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4. E-mail: smithdr@dal.ca

David Roy Smith is an Izaak Walton Killam postdoctoral scholar in the Botany Department at the University of British Columbia, where he studies the genomes of protists.

is remarkably complex [9] but poorly studied in the majority of eukaryotes. The huge quantity of available RNA-Seq reads from across the eukaryotic domain represents an excellent and untapped resource for investigating organelle transcriptomic and genomic architecture. For instance, the mitochondrial genomes of certain land plants and protists undergo massive and perplexing amounts of post-transcriptional editing and processing [10], the study of which has helped spur paradigm-shifting theories of genome evolution [11]. RNA-Seq has proven to be an excellent tool for examining RNA editing and processing within organelles [12,13]—but see [14–16] for critical discussion. Finally, given that most regions in an organelle genome are transcribed [9,13], RNA-Seq results can be mined for coding and noncoding sequences from species for which there are few or no existing organelle data, facilitating phylogenetic, comparative genomic, and genetic barcoding analyses.

Some researchers have already started mining organelle transcripts from the large stores of RNA-Seq data in the SRA. For example, a recent study used available RNA-Seq experiments to compare expression levels of mitochondrial- versus nuclear-encoded genes across metazoans [17]. Others have combined RNA-Seq results with genome sequencing projects to measure genetic diversity and post-transcriptional editing within the plastids of plants and algae [12,13,18]. However, the bulk of the organelle-derived RNA-Seq reads in public databases are waiting to be analyzed.

Two points to keep in mind when collecting organelle transcripts from whole transcriptome RNA-Seq projects: First, because of intercompartmental gene transfer, nuclear genomes often contain mitochondrial- and chloroplast-like sequences, dubbed NUMTs and NUPTs [19], which, if transcriptionally active, can be mistaken for organelle-derived RNA. If a putative organelle transcript is identical to a region in the nuclear genome and/or shows differences to the organelle genome (assuming these data are available) then that could indicate it is a NUMT/NUPT. Second, RNA-Seq libraries are sometimes contaminated with genomic DNA, which is normally due to incomplete DNA digestion during RNA purification. If so, a large fraction of the DNA contamination can come from organelle genomes because their high AT content helps them pass through the poly-A selection step during mRNA enrichment [20]. Poly-A selection will also skew the organelle data towards AT-rich regions of the organelle genome. Most of the RNA-Seq projects deposited within the SRA contain details about the sample preparation and library construction, including if poly-A purification or ribosomal RNA depletion protocols were used.

Despite the large amounts of eukaryotic RNA-Seq data, most researchers disregard organelle transcripts focusing entirely on those from the nucleus. Consequently, there are vast and accumulating amounts of unexplored mitochondrial and chloroplast sequences in the SRA—data that could be used to examine organelle function and evolution. Scientists should take advantage of these data as well as those from species soon to be sequenced as part of the Marine Microbial Eukaryote and 1000 Plant Transcriptome Initiatives. And remember not to overlook the organelle sequences.

Key Points

- There are massive amounts of unexplored mitochondrial and chloroplast RNA-Seq data in public online sequence repositories.
- These data represent an excellent and untapped resource for investigating the transcription, function and evolution of organelle genomes.
- Soon we will be inundated with RNA-Seq data from hundreds of marine microbial eukaryotes and diverse land plants, providing researchers with an unprecedented opportunity to explore nuclear and organelle transcription in some of the world's most diverse and fascinating species.
- I encourage researchers to take advantage of the available RNA-Seq data in GenBank, and not to overlook the mitochondrial- and chloroplast-derived sequences.

Acknowledgements

I thank Aurora Nedelcu for critical reading of the article.

FUNDING

Postdoctoral fellowship (to D.R.S.) from the Izaak Walton Killam Memorial Trusts.

References

1. Smith DR. Not seeing the genomes for the DNA. *Brief Funct Genomics* 2012;**11**:289–90.
2. Iorizzo M, Senalik D, Szklarczyk M, *et al*. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biol* 2012;**12**:61.

3. Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;**40**:D54–6.
4. Marine Microbial Eukaryote Transcriptome Project. http://marinemicroeukaryotes.org (13 December 2012, date last accessed).
5. The 1KP Project. http://www.onekp.com (13 December 2012, date last accessed).
6. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
7. Neira-Oviedo M, Tsyganov-Bodounov A, Lycett GJ, *et al*. The RNA-Seq approach to studying the expression of mosquito mitochondrial genes. *Insect Mol Biol* 2010;**20**:141–52.
8. Raz T, Kapranov P, Lipson D, *et al*. Protocol dependence of sequencing-based gene expression measurements. *PLoS One* 2011;**6**:e19287.
9. Mercer TR, Neph S, Dinger ME, *et al*. The human mitochondrial transcriptome. *Cell* 2011;**146**:645–58.
10. Knoop V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci* 2011;**68**:567–86.
11. Covello PS, Gray MW. On the evolution of RNA editing. *Trends Genet* 1993;**9**:265–8.
12. Tangphatsornruang S, Uthaipaisanwong P, Sangsrakru D, *et al*. Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. *Gene* 2011;**475**:104–12.
13. Fang Y, Wu H, Zhang T, *et al*. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. *PLoS One* 2012;**7**:e37164.
14. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;**335**:1302–c.
15. Lin W, Piskol R, Tan MH, *et al*. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;**335**:1302–e.
16. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;**335**:1302–d.
17. Nabholz B, Ellegren H, Wolf JBW. High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. *Mol Biol Evol* 2012;**30**:272–84.
18. Smith DR, Hua J, Lee RW. Relative rates of evolution among the three genetic compartments of the red alga *Porphyra* differ from those of green plants and do not correlate with genome architecture. *Mol Phylogenet Evol* 2012;**65**:339–44.
19. Kleine T, Maier UG, Leister D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol* 2009;**60**:115–38.
20. Schliesky S, Gowik U, Weber APM, *et al*. RNA-seq assembly—are we there yet? *Front Plant Sci* 2012;**3**:220.