OXFORD

Briefings in Functional Genomics, 2024, 1–7

https://doi.org/10.1093/bfgp/elae026

Review Paper

Long-read RNA sequencing can probe organelle genome pervasive transcription

Matheus Sanita Lima [1,*], Douglas Silva Domingues [2], Alexandre Rossi Paschoal [3], David Roy Smith [1,*]

[1]Department of Biology, Western University, 1151 Richmond Street, London, Ontario N6A 5B7, Canada
[2]Department of Genetics, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Avenida Padua Dias 11, Piracicaba, SP 13418-900, Brazil
[3]Department of Computer Science, Bioinformatics and Pattern Recognition Group (BIOINFO-CP), Federal University of Technology – Paraná – UTFPR, Avenida Alberto Carazzai 1640, Cornélio Procópio, PR 86300000, Brazil

*Corresponding authors. Sanita Lima, M. (msanital@uwo.ca); Smith, D. R. (dsmit242@uwo.ca), Laboratory website: https://www.arrogantgenome.com/, Twitter handle: twitter.com/arrogantgenome

Abstract

40 years ago, organelle genomes were assumed to be streamlined and, perhaps, unexciting remnants of their prokaryotic past. However, the field of organelle genomics has exposed an unparallel diversity in genome architecture (*i.e.* genome size, structure, and content). The transcription of these eccentric genomes can be just as elaborate – organelle genomes are pervasively transcribed into a plethora of RNA types. However, while organelle protein-coding genes are known to produce polycistronic transcripts that undergo heavy posttranscriptional processing, the nature of organelle noncoding transcriptomes is still poorly resolved. Here, we review how wet-lab experiments and second-generation sequencing data (*i.e.* short reads) have been useful to determine certain **types of organelle RNAs, particularly noncoding RNAs**. We then explain how third-generation (long-read) RNA-Seq data represent the new frontier in organelle transcriptomics. We show that public repositories (*e.g.* NCBI SRA) already contain enough data for inter-phyla comparative studies and argue that organelle biologists can benefit from such data. We discuss the prospects of using publicly available sequencing data for organelle-focused studies and examine the challenges of such an approach. We highlight that the lack of a comprehensive database dedicated to organelle genomics/transcriptomics is a major impediment to the development of a field with implications in basic and applied science.

Keywords: long-read RNA sequencing; pervasive transcription; noncoding RNA; organelle genome; mitochondrial transcription; plastid transcription

Organelle genomes, veritable RNA machines

The sequencing of the human and mouse mitochondrial genomes ignited the field of organelle genomics 40 years ago [1, 2]. These compact genomes might have given the initial impression that organelle chromosomes provide very little genetic fodder for the scientific community to analyze. Indeed, with only 37 genes (encoding 13 proteins, 22 tRNAs, and 2 rRNAs), no introns, and only 1122 bp of noncoding DNA (forming the D-loop region), the human mitochondrial genome (NC_012920.1) is the epitome of a streamlined gene expression system. According to the tRNA-punctuation model, the human mtDNA generates two genome-wide, strand-specific polycistronic transcripts, which are processed into monocistronic transcripts [3]. What else could be expected from a mitochondrial chromosome after more than a billion years of reductive evolution *via* endosymbiotic gene transfer and gene loss? [4].

Today, we now know that 'anything goes' in the world of organelle genomics [5]. Noncanonical genetic codes, RNA editing, twintrons, and gene splintering are just a few of the eccentric traits that make mitochondrial and plastid chromosomes so unique and fascinating [6]. The massive diversity in size, structure, and content of organelle genomes continues to puzzle scientists and has spurred various evolutionary theories [7]. The expression of these molecules is, expectably, just as convoluted. A myriad of RNAs (both coding and noncoding) of various sizes and configurations are expressed in the mitochondria and plastids of eukaryotes [8]. These organelle RNAs, in turn, interact with nuclear-encoded proteins for posttranscriptional processes pertaining to all aspects of RNA metabolism (maturation, stability, translation, and degradation) [9]. The complexity of organelle genomes and transcriptomes (both coding and noncoding) is such that some of these systems have been compared to Rube Goldberg machines, **which are machines designed in profligate ways to perform a rather simple function** [10].

Initial investigations of organelle genome expression relied on painstaking piecemeal wet-lab analyses. Protein-coding genes (and their respective mRNAs) were the main focus, and little attention was given to the possible occurrence of regulatory and structural noncoding RNAs (rRNAs and tRNAs notwithstanding). A similar protein-centric trend carried on in the studies of nuclear

Matheus Sanita Lima is a researcher at the University of Western Ontario where he investigates (in silico) the pervasive transcription of organelle genomes.
Douglas Silva Domingues is a Professor (Genetics) at ESALQ and the head of GeTransPlant where he investigates plant noncoding RNAs.
Alexandre Rossi Paschoal is a Professor (Bioinformatics) at UTFPR where he develops computational tools for genomic and transcriptomic analyses with a focus on noncoding RNAs.
David Roy Smith is a Full Professor (Biology) at the University of Western Ontario where investigates genome evolution of eukaryotic microbes. He can be found online at www.arrogantgenome.com and @arrogantgenome.
Received: March 28, 2024. **Revised:** May 20, 2024. **Accepted:** May 30, 2024

genome expression, and most of the noncoding DNA in both the nucleus and organelles was assumed to be transcriptionally inactive 'junk DNA' [11]. Yet, hints to the transcriptional (and functional) potential of organelle noncoding sequences surfaced with the first cases of novel noncoding RNAs in the plastomes of *Chlamydomonas reinhardtii* and *Nicotiana tabacum* [12, 13]. With the advent of next-generation sequencing (NGS) technologies, such as 454 and Illumina, the noncoding portions of nuclear and organelle transcriptomes have finally taken center stage. Nuclear genomes can produce a constellation of regulatory and structural ncRNAs that function in the nucleus, cytosol, and elsewhere [14]. Such a 'Kuhnian paradigm shift' has consolidated the notion of nuclear genomes being veritable 'RNA machines' **(*i.e.* the sheer amount of potentially functional noncoding DNA in nuclear genomes makes these genomes real factories of RNAs)** [11, 15]. Organelle genomes have followed suit. The poster child of organelle transcriptional efficiency, the human mitochondrial transcriptome was shown to boast a plethora of small noncoding RNAs, including those arising from tRNAs and the antisense portions of protein-coding genes [16]. NGS experiments also allowed the unravelling of multiple transcription start sites (TSSs) embedded within plastid gene clusters in barley and Arabidopsis [17, 18], pointing to the richness and complexity of organelle noncoding transcriptomes.

Third-generation long-read sequencing by PacBio and Oxford Nanopore Technologies (ONT) are now transforming NGS studies. **PacBio sequencing was made available already in 2011** [19]**, whereas ONT platform joined the market in 2014** [20]**. Their initial high error rates (between 11% - 33% depending on the platform) and much lower yields** [21] **prevented the widespread adoption of these technologies a decade ago. Fast-forward to now, error rates have dropped to below 1% in some protocols** [22]**, but long-read RNA sequencing still exhibit higher base-calling inaccuracies that demand computational mitigation** [23]**).** Nevertheless, long and ultra-long (1Kb - >10Kb) DNA/RNA reads span sequences that were hard to assemble with short reads alone [24]. Repetitive noncoding segments of genomes and transcriptomes can now be captured in uninterrupted long reads, which allows for telomere-to-telomere (T2T) assemblies [25]. Public repositories, such as the NCBI Sequence Read Archive (SRA), still have mostly short-read data, but a considerable number of long-read experiments can already be found. As expected, most of these datasets were produced to study nuclear genomes and transcriptomes, but here we argue that organelle biologists should not leave these data untapped.

In this Review, we discuss how publicly available, third-generation (*i.e.* long-read) RNA-Seq can be used to investigate the pervasive transcription of organelle genomes. Our experience with short-read (*i.e.* second-generation) RNA-Seq data informs our recommendations and substantiate our predictions; long RNA reads have the potential to unravel new layers of the transcriptional (particularly noncoding) architecture of organelle genomes. Insights pertaining several aspects of organelle biology, from the nature of pervasive transcription itself to the inter-organellar communication in the eukaryotic cell, ought to be ushered in.

Long-read RNA-Seq data: A goldmine for organelle research 2.0

Short-read RNA-Seq data were once called 'a goldmine for organelle research' [26]**; this is because RNA-Seq datasets derived from whole-cell experiments are typically replete with organelle transcripts that can be mined for organelle-focused studies** [27, 28]**.** In fact, it was publicly available short RNA reads that helped the elucidation of pervasive organelle transcription across eukaryotes [29, 30]. **Pervasive transcription is the transcription (above background levels) of RNAs outside the canonical boundaries of protein-coding, rRNA, and tRNA genes** [31]. **The 'pervasively transcribed' RNAs are noncoding RNAs that can have structural and/or regulatory roles depending on their size, nature, and origin of transcription** [14]**. The corollary is that noncoding RNAs (ncRNAs) are RNAs without any detectable coding capacity (*i.e.* a conventional open reading frame cannot be found within the ncRNA). However, there already are cases of long noncoding RNAs (lncRNAs) that exert their noncoding (structural and/or regulatory) roles and encode micropeptides thanks to hidden small open reading frames (sORFs)** [32].

By mining RNA-Seq datasets produced from whole-cell studies and deposited at the NCBI Sequence Read Archive (SRA), various teams have demonstrated that organelle genomes are fully or near-fully transcribed, independent of organelle genome content, structure, and taxonomic origin [33, 34]. Straightforward mapping analyses easily demonstrated that long stretches of noncoding DNA (both intronic and intergenic) are transcribed, and at levels beyond that of merely 'background noise'. These analyses were made possible because of both the high throughput of next-generation sequencing technologies and the popularity of short-read sequencing (particularly Illumina) across the life sciences. Second-generation RNA-Seq datasets from disparate species, including cryptic marine eukaryotes [35], abounded in public repositories, allowing for large-scale inter-domain comparative analyses.

Over the last ten years, third-generation sequencing technologies have advanced and undergone consolidation [24]. PacBio and ONT are currently the leading platforms for long-read studies and have been used to explore species from all well-studied eukaryotic groups (*i.e.* land plants, green algae, metazoans, and fungi), including some protist lineages. As of March 2024, long-read RNA-Seq datasets for >1000 species are publicly available at NCBI SRA (Fig. 1). PacBio is the most popular sequencing platform (~79% of the long-read RNA data come from PacBio experiments), which can indicate that the portability of MinION by ONT has not yet borne fruit. The 'patchy distribution' of ONT datasets (*i.e.* these datasets seem to be clustered around certain taxonomic groups) might be explained by project consortia that aim to study entire groups of organisms - see below The Darwin Tree of Life and the Vertebrate Genomes projects. Independent of sequencing platform, land plants overshadow even metazoans mostly because of the bias towards economically important crop species. Ecologically relevant protists and fungi are poorly represented in absolute numbers but are relatively well sampled in terms of diversity within each group.

While the mitochondrial and plastid transcriptomes of some model species (*e.g. Arabidopsis thaliana*, *C. reinhardtii*, and *Drosophila melanogaster*) have been extensively studied, these investigations used mostly, if not only, short-read data. To our knowledge, long RNA reads have been used to study the organelle transcriptomes of only *Saccharomyces cerevisiae*, *Mus musculus*, *Homo sapiens*, *Erthesina fullo*, *Caulerpa lentillifera* and *Nymphaea* 'Joey Tomocik' [36–41]. Conversely, hundreds of organelle genomes, mostly mitochondrial genomes, have been sequenced with long DNA reads. Large-scale projects, such as the Darwin Tree of Life [42] and Vertebrate Genomes Project [43], have generated long-read genomic data and assembled the mitochondrial genomes of hundreds of species through tailored pipelines [44, 45]. We identified a similar trend (*i.e.* organelle genomes are studied more readily than organelle

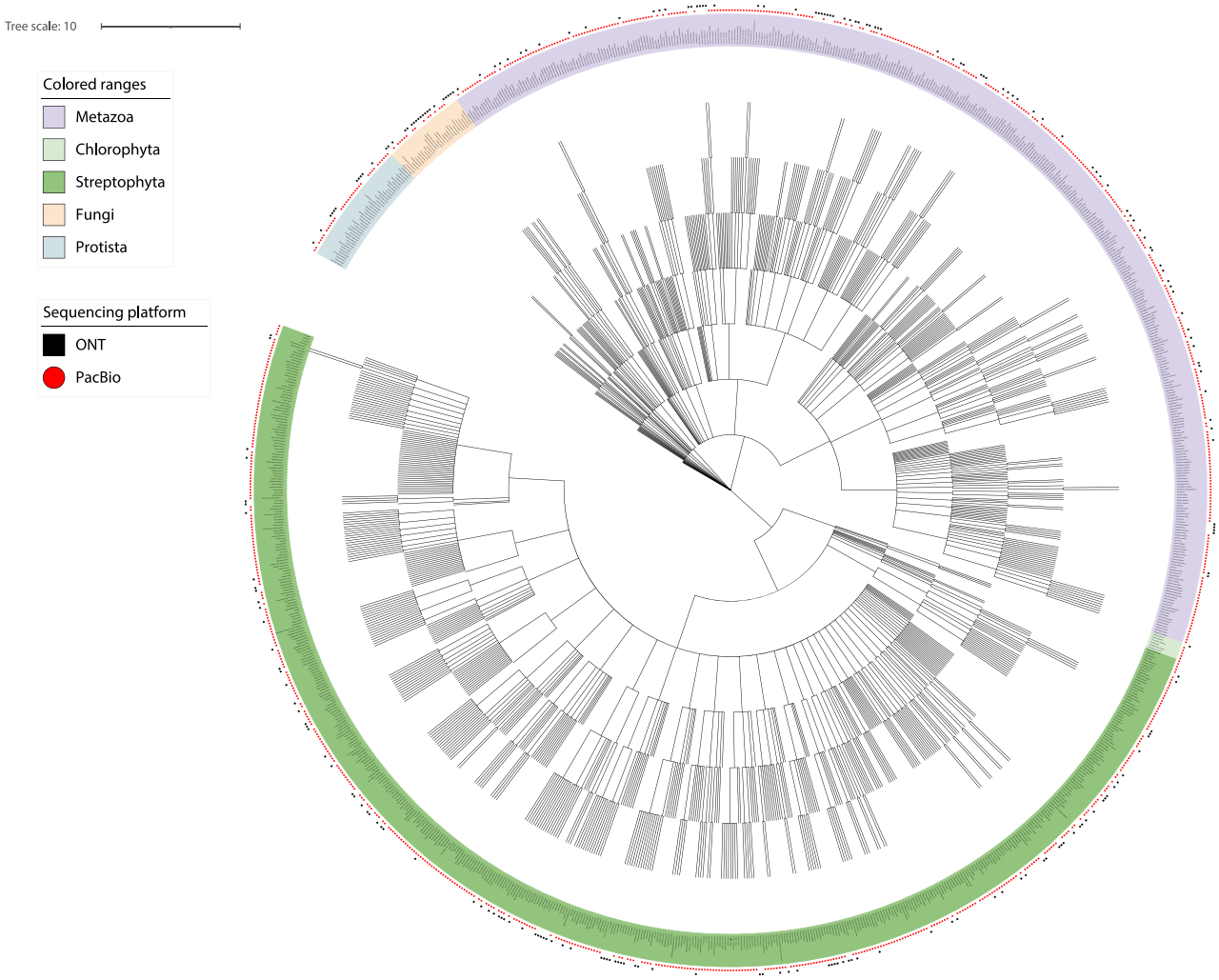

Figure 1. **Distribution of publicly available long-read RNA-Seq datasets for eukaryotes in NCBI SRA.** As of March 2024, NCBI SRA has long-read RNA-Seq datasets for at least 1091 eukaryotic species. Fully (or near fully) sequenced organelle genomes from most of these species are deposited in public repositories (such as NCBI Nucleotide - https://www.ncbi.nlm.nih.gov/nucleotide/). These data represent an untapped resource for the study of organelle noncoding transcriptomes, particularly long noncoding RNAs. The 1091 species were manually compiled after a keyword search in the NCBI SRA Advanced Search page (https://www.ncbi.nlm.nih.gov/sra/advanced). The search builder was filled as 'All fields: PacBio' OR 'All fields: Oxford Nanopore'. Search results for RNA datasets were downloaded and filtered by taxon. The depicted tree was created *via* iTOL – Interactive Tree of Life – v6 (https://itol.embl.de/) using a phylip tree obtained from the NCBI Taxonomy Common Tree tool (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi). Species name incongruences between tools (cause by either heterotypic or homotypic synonyms) were manually corrected. Varieties and subspecies (*e.g. Solanum lycopersicum var. cerasiforme*) were lumped within the species name, whenever possible. The tree is not meant to represent the most up-to-date phylogenetic relationships among and within groups. Branch lengths are ignored, and groups (*i.e.* 'colored ranges') are coloured by clades. Groups were partitioned according to data availability and may not always represent a phylogenetically informed classification (*e.g.* 'protists' is a paraphyletic group). Outer rings represent the sequencing platform of the RNA-Seq dataset(s) from each species. Some species have PacBio and ONT transcriptomic datasets.

transcriptomes) with second-generation sequencing data as well. NGS genomic data have made mitochondrial genomes the most sequenced type of chromosomes, while NGS transcriptomic data have been largely underused by organelle biologists [46]. Given the available data (*i.e.* long RNA reads from numerous species) and the understudied system (*i.e.* organelle noncoding transcriptomes), we highlight that publicly available long-read RNA-Seq represents a new iteration of a goldmine for organelle research.

The challenges and opportunities of using third-generation RNA-Seq data for organelle transcriptomics

Pervasive, polycistronic transcription of organelle genomes is near ubiquitous across the eukaryotic domain. Organelle polycistronic transcription is, in many ways, a relic from the prokaryotic progenitors of mitochondria and plastids. The organellar novelty is the intricate posttranscriptional processing mechanisms, which are typically much simpler in bacteria. This posttranscriptional regulation of organelle genes is heavily reliant on nuclear-encoded factors and could be considered to represent the 'eukaryotization' of organelles [47, 48]. In these intricacies lies the potential of pervasive transcription beyond the canonical boundaries of protein-coding regions, which can generate ncRNAs (excluding tRNAs and rRNAs) with potential biological functions. Short-read sequencing data have repeatedly confirmed the presence of numerous small (< 200 nt) ncRNAs from both mitochondria and plastids of diverse species [49, 50]. Long-read sequencing data can now unravel the widespread production of long (> 200 nt) ncRNAs of various natures (*i.e.* intergenic, intronic, and antisense) in organelles.

As third-generation sequencing technologies become mainstream and populate public repositories, there are four main challenges (or opportunities, depending on the viewpoint) that organelle genomicists should consider: i) NUMTs and NUPTs, ii) metadata information and protocol details, iii) the need for benchwork validation, and iv) the lack of tailored databases.

Nuclear-mitochondrial-like and nuclear-plastid-like DNAs (NUMTS and NUPTS, respectively) have always been a thorny issue for organelle transcriptomic studies. These nuclear sequences are derived from recent organelle-to-nucleus DNA transfer events. Most NUMTs and NUPTs are in the process of pseudogenization, so their transcripts (when mapped to organelle genomes) should have moderate to low sequence similarity (<90%) to their organelle counterparts [50] – but see [51] for kilobase-long NUMT segments that are 100% identical to mitochondrial genome sequences. The problem is accentuated in species rich with NUMTs/NUPTs, such as *A. thaliana*, because nuclear genomes are pervasively transcribed, including their pseudogenes [52]. It can be difficult to pinpoint whether short RNA reads come from NUMTs/NUPTs or actual organelle genomes, as mapped short reads can fall within NUMTs/NUPTs and away from junction regions. Still, analyses of small RNA data from six vertebrate species have shown that the amount of NUMTs does not correlate with the amount of small (mitochondrial) RNAs [53], which indicates that transcripts from NUMTs/NUPTs might not always be a concern in mapping analyses. Long RNA reads have the capacity to alleviate this issue, as they can span entire NUMTs/NUPTs (including flanking nuclear sequences) and be filtered out from organelle read mapping analyses. If long-read RNA-Seq data and the corresponding organelle genome(s) are available, investigations similar to Pozzi and Dowling [53] comparing species with varying ratios of NUMTs/NUPTs would be timely and informative.

Metadata information and sequencing protocol details are of uttermost importance when performing organelle mapping analyses with publicly available data. Certain aspects of the organelle transcriptomes can only be understood if specific types of data are available. For example, strand-specific data are needed for the investigation of antisense RNAs. RNA isolation/enrichment procedures (*i.e.* oligo dT primers versus rRNA degradation) will also dictate which pool of long RNAs is being sequenced. As polyadenylation has numerous consequences (degradation or stabilization) on organelle transcripts in different species [8, 9], investigators should be mindful of what function polyadenylation has on their organism(s)/genome(s) of study. Metadata information pertaining to species/strain name, tissue of origin (*i.e.* root versus leaf; muscle versus bone; single-cell versus bulk) will be needed to interpret the mapping results adequately. The same idea applies to experimental conditions (*e.g.* light versus dark; high-salt versus low-salt), as organelle genomes readily sense environmental stimuli [48], and ncRNAs show condition/cell-specific expression patterns [54]. Organelle biologists who are producing their own long-read data, should be mindful of the FAIR principles when depositing their raw reads in public repositories [55].

Although thousands of ncRNAs have been found in the nuclear genomes of humans (and other species) and hundreds in organelle genomes [14, 49, 50], most of these transcripts have only putative functions. In only few cases have ncRNA had thorough benchtop validation, but those that have been validated have shown essential functions in health and disease states with profound phenotypic consequences [15]. Given that diverse species can now be investigated with long-read data, techniques for targeted loss-of-function mutation and organelle genome manipulation of noncoding regions represent the next frontier in the functional investigations of (organelle) ncRNAs [56].

The importance of bioinformatics databases is hard to overstate, particularly in the emerging field of ncRNAs. International initiatives, such as ENCODE [57] and FANTOM [58], have created unparallel amounts of data and incentivized the development of specialized databases. Chiefly amongst such databases is RNAcentral (https://rnacentral.org/), a catch-all network containing 53 'expert databases' dedicated solely to ncRNAs [59]. At present, there is not a single expert database within RNAcentral dedicated to organelle ncRNAs. Even more despairing is that only mitochondrial rRNAs (labelled as mt rRNA) can be found amongst the numerous database categories. Only one sequence (out of more than 36 million sequences – version 23) has the label 'mitochondrial DNA'. Databases dedicated to organelle genomes have always fared poorly when compared to their nuclear counterparts, but the current state-of-affairs for organelle databases is alarming. The NCBI Organelle Genome Resources database [60] has recently become a legacy database and will cease to exist in May 2024. GOBASE [61] was the gold-standard for organelle genome databases, but no longer exists. ChloroplastDB [62] seems to have been terminated as well, and the sole remaining organelle database is OGDA [63] – a database of organelle genomes from algae only. **The fact that these databases have been discontinued, despite the broad-range implications of organelle genome studies, is worrisome and perhaps an indication that (more) money and time should be invested in the field. Alternatively, augmenting existing repositories (such as NCBI Nucleotide) with better tagging of fully sequenced organelle genomes and related data could be a quick(er) solution. However, the databases listed above** were/are dedicated to the compilation of organelle genomes, and not of organelle transcriptomic data. Given the numerous cases of organelle ncRNAs found in a handful of species [49, 50], the potential for a dedicated organelle ncRNA database is tangible. The creation of such a database will allow comparative studies to unravel common trends and singularities across the eukaryotic Tree of Life. These comparative studies can provide insights into countless aspects of organelle biology, some of which are discussed below.

Significance and conclusions

All eukaryotic life (most of which is microbial) depends on organelles in one way or another. Organelles are information processing hubs that control cellular processes far beyond energy production [48, 64]. Being the hallmark of eukaryotes (and of eukaryogenesis, in the case of mitochondria [65]), studying any aspect of organelle biology is poised to have far-reaching implications and undisputed significance. Here we hold that the noncoding portion of organelle transcriptomes represents an understudied layer of organelle biology that can have influence on at least four mechanisms: i) anterograde/retrograde pathways, ii) plastid differentiation, iii) organelle genome topology, and iv) metabolic regulation.

Having separate genetic compartments represents a real challenge for eukaryotes, as these compartments can only function adequately with a fine-tuned communication system between them [66]. Beyond that, organelles represent an extra point of entry for environmental stimuli, which creates more levels of communication and regulation [48]. Anterograde signalling is the communication pathway from the nucleus to the organelles, whereas retrograde signalling is the reverse channel

– from organelles to the nucleus. The nature of the signalling molecules in both directions has been hard to determine, but most inter-compartmental communication seems to happen *via* molecules such as mitochondrial metabolites and reactive oxygen species – ROS [66]. The clue to the role of ncRNAs in antero-grade/retrograde signalling came from NGS mapping analyses demonstrating the nuclear localization of mitochondrial ncRNAs [67]. More recently, the production and export of (long) circular mitochondrial RNAs has been uncovered using third-generation RNA-Seq data as well [40]. Although yet not fully elucidated, the possibility of a consolidated communication pathway based on large(r) molecules opens a new avenue of research, not to mention the systems of import/export and transcriptional regulation necessary for this RNA-based communication to happen [68].

Adjacent to the inter-organellar communication lies the obscure mechanisms of organelle differentiation and fate coordination. Whereas mitochondrial shape, size, and number vary considerably depending on cell type and species [69], plastid differentiation in land plants is arguably unmatched. The factors that determine the transition from a proplastid to a chloroplast (or from a chloroplast to a chromoplast) are poorly known, but undoubtedly rely on the nucleus-organelle communication systems. As plastid gene repertoire does not vary much across land plants, the noncoding segments of plastomes might hold the clues to the regulation of organelle fate determination. What triggers a chloroplast to turn into a chromoplast during fruit ripening might be hidden in regulatory/structural noncoding RNAs acting in and outside organelles.

Nuclear lncRNAs are remarkable for their structural roles in the nucleus, the cytoplasm and elsewhere [14]. These versatile molecules could be seen as Swiss Army knives capable of regulating nuclear gene expression by changing local and global chromatin configurations. The topology of organelles genomes is just as important, yet vastly understudied. Plastomes and mitogenomes are organized into (highly compact) nucleoids attached to the membranes of thylakoids and inner mitochondrial membrane [70, 71]. Such configuration might be reminiscent of their bacterial ancestry, but part of the proteome that shapes such organelle nucleoids is of eukaryotic origin (which is another example of the 'eukaryotization' of organelles) [72]. If nuclear long lncRNAs are shaping nuclear genomes (and nuclear speckles) [14], what could nuclear and organellar lncRNAs do to organelle nucleoids? Organelle genome topology can be a new line of inquiry birthed by long-read sequencing technologies.

Mitochondria and chloroplasts are known for their energy-producing capabilities, but how these cellular hubs communicate with one another and determine trophic modes in mixotrophic organisms is basically unknown. Capable of switching from autotrophy to heterotrophy (and *vice-versa*), mixotrophs are unique model systems for the study of interorganellar communication and sensing. As different environmental conditions determine what trophic mode is turned on [73], the noncoding transcriptomes of organelles might play a role in gene expression regulation and trophic mode determination.

The common theme among these proposed research arenas is the overlooked transcriptional and regulatory potential of non-coding sequences in organelle genomes. Organelle protein-coding gene repertoires are well known, and the function of organelle proteins have been firmly established. The treasure trove lies in the noncoding component of organelle genomes, be it intergenic and intronic nucleotides or antisense strands to protein-coding genes. These segments have proven transcriptional capacity, but the nature of their transcripts is yet mostly unknown. We defend that organelle ncRNAs, both small and long, have the potential to exert regulatory and structural roles in the organelles, cytosol, nucleus, and outside of cells. One major challenge in the study of ncRNAs, particularly lncRNAs, is their generally low expression levels and cell-specific transcriptional patterns [54]. Direct RNA long-read sequencing [39] combined with 'single-organelle' (analogous to single-cell) protocols would represent the next frontiers in the study of organelle noncoding transcriptomes.

Should the function of organelle noncoding sequences be corroborated, the discourse about organelle genome evolution would be, yet again, reshaped. Although we do not object to the view that organelle genomes might be hoarding non-functional sequences, we certainly do not ignore the possibility of exaptation of some of those noncoding sequences. As organelle genomes expand, their transcriptomes also expand due to pervasive transcription. If evolution is a tinkerer [74], the tinkering has ended up producing organelle Rube Goldberg machines. They might not be the most efficient way to run a eukaryotic cell, but they certainly keep organelle biologists occupied.

Key Points

- Third-generation long-read RNA-Seq data already abound in public repositories, such as NCBI SRA.
- Publicly available long RNA reads can be used to investigate the pervasive transcription of organelle genomes.
- Organelle noncoding transcriptomes should contain numerous long noncoding RNAs with putative function within and outside organelles.
- The fields of organelle genomics and transcriptomics lack a comprehensive database to enable comparative evolutionary studies.

Author contributions

M.S.L. wrote the first draft and created the figures. D.R.S. revised the manuscript. D.S.D. and A.R.P. provided feedback in the drafts. All authors have discussed the ideas here presented and contributed to the conceptualization of this piece.

Matheus Sanitá Lima (Conceptualization [supporting], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [equal], Project administration [lead], Writing—original draft [lead], Writing—review & editing [equal]), Douglas Silva Domingues (Conceptualization [equal], Data curation [supporting], Formal analysis [supporting], Methodology [supporting], Resources [equal], Validation [equal], Writing—review & editing [equal]), Alexandre Paschoal (Conceptualization [equal], Funding acquisition [equal], Investigation [supporting], Methodology [supporting], Project administration [supporting], Resources [supporting], Supervision [equal], Writing—review & editing [supporting]), and David Smith (Conceptualization [lead], Data curation [supporting], Formal analysis [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Project administration [lead], Resources [lead], Supervision [lead], Writing—original draft [lead], Writing—review & editing [lead]).

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [Discovery Grant to D.R.S.]; and the Fundação Araucária [NAPI Bioinformatica 66.2021 to A.R.P.]

References

1. Anderson S, Bankier AT, Barrell BG *et al.* Sequence and organization of the human mitochondrial genome. *Nature* 1981;**290**:457–65.
2. Bibb MJ, Van Etten RA, Wright CT *et al.* Sequence and gene organization of mouse mitochondrial DNA. *Cell* 1981;**26**:167–80.
3. Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. *Nature* 1981;**290**:470–4.
4. Timmis JN, Ayliffe MA, Huang CY *et al.* Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 2004;**5**:123–35.
5. Burger G, Gray MW, Franz Lang B. Mitochondrial genomes: anything goes. *Trends Genet* 2003;**19**:709–16.
6. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci U S A* 2015;**112**:10177–84.
7. Smith DR. The mutational hazard hypothesis of organelle genome evolution: ten years on. *Mol Ecol* 2016;**25**:3769–75.
8. Smith DR, Keeling PJ. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. *Annu Rev Microbiol* 2016;**70**:161–78.
9. Stern DB, Goldschmidt-Clermont M, Hanson MR. Chloroplast RNA metabolism. *Annu Rev Plant Biol* 2010;**61**:125–55.
10. Gray MW, Lukes J, Archibald JM *et al.* Irremediable complexity? *Science* 2010;**330**:920–1.
11. Mattick JS. A Kuhnian revolution in molecular biology: most genes in complex organisms express regulatory RNAs. *Bioessays* 2023;**45**:e2300080.
12. Goldschmidt-Clermont M, Choquet Y, Girard-Bascou J *et al.* A small chloroplast RNA may be required for trans-splicing in Chlamydomonas reinhardtii. *Cell* 1991;**65**:135–43.
13. Vera A, Sugiura M. A novel RNA gene in the tobacco plastid genome: its possible role in the maturation of 16S rRNA. *EMBO J* 1994;**13**:2211–7.
14. Mattick JS, Amaral PP, Carninci P *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 2023;**24**:430–47.
15. Amaral PP, Dinger ME, Mercer TR *et al.* The eukaryotic genome as an RNA machine. *Science* 2008;**319**:1787–9.
16. Mercer TR, Neph S, Dinger ME *et al.* The human mitochondrial transcriptome. *Cell* 2011;**146**:645–58.
17. Castandet B, Germain A, Hotto AM *et al.* Systematic sequencing of chloroplast transcript termini from *Arabidopsis thaliana* reveals >200 transcription initiation sites and the extensive imprints of RNA-binding proteins and secondary structures. *Nucleic Acids Res* 2019;**47**:11889–905.
18. Zhelyazkova P, Sharma CM, Forstner KU *et al.* The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plant Cell* 2012;**24**:123–36.
19. Quail MA, Smith M, Coupland P *et al.* A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;**13**:341.
20. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014;**14**:1097–102.
21. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. *Nat Rev Genet* 2023;**24**:627–41.
22. Amarasinghe SL, Su S, Dong X *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:30.
23. Su Y, Yu Z, Jin S *et al.* Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. *Nat Commun* 2024;**15**:3972.
24. Kovaka S, Ou S, Jenike KM *et al.* Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nat Methods* 2023;**20**:12–6.
25. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.
26. Smith DR. RNA-Seq data: a goldmine for organelle research. *Brief Funct Genomics* 2013;**12**:454–6.
27. Forsythe ES, Grover CE, Miller ER *et al.* Organellar transcripts dominate the cellular mRNA pool across plants of varying ploidy levels. *Proc Natl Acad Sci U S A* 2022;**119**:e2204187119.
28. Tian Y, Smith DR. Recovering complete mitochondrial genome sequences from RNA-Seq: a case study of *Polytomella* non-photosynthetic green algae. *Mol Phylogenet Evol* 2016;**98**:57–62.
29. Sanita Lima M, Smith DR. Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists. *G3 (Bethesda)* 2017;**7**:3789–96.
30. Sanita Lima M, Smith DR. Pervasive transcription of mitochondrial, plastid, and nucleomorph genomes across diverse plastid-bearing species. *Genome Biol Evol* 2017;**9**:2650–7.
31. Sanita Lima M, Rossi Paschoal A, Silva Domingues D *et al.* Pervasive transcription of plant organelle genomes: functional noncoding transcriptomes? *Trends Plant Sci*. IN PRESS.
32. Sruthi KB, Menon A, Akash P *et al.* Pervasive translation of small open reading frames in plant long non-coding RNAs. *Front Plant Sci* 2022;**13**:975938.
33. Castandet B, Hotto AM, Strickler SR *et al.* ChloroSeq, an optimized chloroplast RNA-Seq Bioinformatic pipeline, reveals Remodeling of the Organellar transcriptome under heat stress. *G3 (Bethesda)* 2016;**6**:2817–27.
34. Forni G, Puccio G, Bourguignon T *et al.* Complete mitochondrial genomes from transcriptomes: assessing pros and cons of data mining for assembling new mitogenomes. *Sci Rep* 2019;**9**:14806.
35. Keeling PJ, Burki F, Wilcox HM *et al.* The Marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 2014;**12**:e1001889.
36. Gao S, Ren Y, Sun Y *et al.* PacBio full-length transcriptome profiling of insect mitochondrial gene expression. *RNA Biol* 2016;**13**:820–5.
37. Gao S, Tian X, Chang H *et al.* Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. *Mitochondrion* 2018;**38**:41–7.
38. He Z-S, Zhu A, Yang J-B *et al.* Organelle genomes and transcriptomes of *nymphaea* reveal the interplay between intron splicing and RNA editing. *Int J Mol Sci* 2021;**22**:9842.
39. Koster CC, Kleefeldt AA, van den Broek M *et al.* Long-read direct RNA sequencing of the mitochondrial transcriptome of *Saccharomyces cerevisiae* reveals condition-dependent intron abundance. *Yeast* 2023;1–23.
40. Zhang J, Hou L, Zuo Z *et al.* Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat Biotechnol* 2021;**39**:836–45.
41. Zou X, Verbruggen H, Li T *et al.* Identification of polycistronic transcriptional units and non-canonical introns in green algal

chloroplast based on long-read RNA sequencing data. *BMC Genomics* 2021;**22**:298.
42. The Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin tree of life project. *Proc Natl Acad Sci U S A* 2022;**119**:e2115642118.
43. Rhie A, McCarthy SA, Fedrigo O *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021;**592**:737–46.
44. Formenti G, Rhie A, Balacco J *et al.* Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol* 2021;**22**:120.
45. Uliano-Silva M, Ferreira JGRN, Krasheninnikova K *et al.* MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* 2023;**24**:288.
46. Sanita Lima M, Woods LC, Cartwright MW *et al.* The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes. *Mol Ecol Resour* 2016;**16**:1279–86.
47. Weihe A, Liere K, Borner T. 2012. Transcription and transcription regulation in chloroplasts and mitochondria of higher plants. In Bullerwell CE (ed). *Organelle genetics: evolution of organelle genomes and gene expression*. Heidelberg: Springer Berlin. 297–325.
48. Woodson J, Chory J. Coordination of gene expression between organellar and nuclear genomes. *Nat Rev Genet* 2008;**9**:383–95.
49. Anand A, Pandi G. Noncoding RNA: an insight into chloroplast and mitochondrial gene expressions. *Life (Basel)* 2021;**11**:49.
50. Dietrich A, Wallet C, Iqbal RK *et al.* Organellar non-coding RNAs: emerging regulation mechanisms. *Biochimie* 2015;**117**:48–62.
51. Fields PD, Waneka G, Naish M *et al.* Complete sequence of a 641-kb insertion of mitochondrial DNA in the *Arabidopsis thaliana* nuclear genome. *Genome Biol Evol* 2022;**14**:evac059.
52. Xie J, Li Y, Liu X *et al.* Evolutionary origins of pseudogenes and their association with regulatory sequences in plants. *Plant Cell* 2019;**31**:563–78.
53. Pozzi A, Dowling DK. The genomic origins of small mitochondrial RNAs: are they transcribed by the mitochondrial DNA or by mitochondrial pseudogenes within the nucleus (NUMTs)? *Genome Biol Evol* 2019;**11**:1883–96.
54. Unfried JP, Ulitsky I. Substoichiometric action of long noncoding RNAs. *Nat Cell Biol* 2022;**24**:608–15.
55. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
56. Wang J, Kan S, Liao X *et al.* Plant organellar genomes: much done, much more to do. *Trends Plant Sci* 2023online ahead of print. https://doi.org/10.1016/j.tplants.2023.12.014.
57. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
58. Abugessaisa I, Ramilowski JA, Lizio M *et al.* FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res* 2020;**49**:D892–8.
59. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res* 2021;**49**:D212–20.
60. Wolfsberg TG, Schafer S, Tatusov RL *et al.* Organelle genome resources at NCBI. *Trends Biochem Sci* 2001;**26**:199–203.
61. O'Brien EA, Zhang Y, Wang E *et al.* GOBASE: an organelle genome database. *Nucleic Acids Res* 2009;**37**:D946–50.
62. Cui L, Veeraraghavan N, Richter A *et al.* ChloroplastDB: the chloroplast genome database. *Nuclei Acids Res* 2006;**34**:D692–6.
63. Liu T, Cui Y, Jia X *et al.* OGDA: a comprehensive organelle genome database for algae. *Database* 2020;**2020**:baaa097.
64. Picard M, Shirihai OS. Mitochondrial signal transduction. *Cell Metab* 2022;**34**:1620–53.
65. Embley T, Martin W. Eukaryotic evolution, changes and challenges. *Nature* 2006;**440**:623–30.
66. Vendramin R, Marine J-C, Leucci E. Non-coding RNAs: the dark side of nuclear-mitochondrial communication. *EMBO J* 2017;**36**:1123–33.
67. Landerer E, Villegas J, Burzio VA *et al.* Nuclear localization of the mitochondrial ncRNAs in normal and cancer cells. *Cell Oncol* 2011;**34**:297–305.
68. Ren B, Guan M-X, Zhou T *et al.* Emerging functions of mitochondria-encoded noncoding RNAs. *Trends Genet* 2023;**39**:125–39.
69. Bereiter-Hahn J, Voth M. Dynamics of mitochondria in living cells: shape changes, dislocations, fusion, and fission of mitochondria. *Microsc Res Tech* 1994;**27**:198–219.
70. Chen XJ, Butow RA. The organization and inheritance of the mitochondrial genome. *Nat Rev Genet* 2005;**6**:815–25.
71. Nishimura Y. Plastid nucleoids: insights into their shape and dynamics. *Plant Cell Physiol* 2023;**65**:551–9.
72. Sakai A, Takano H, Kuroiwa T. Organelle nuclei in higher plants: structure, composition, function, and evolution. *Int Rev Cytol* 2004;**238**:59–118.
73. Millette NC, Gast RJ, Luo JY *et al.* Mixoplankton and mixotrophy: future research priorities. *J Plankton Res* 2023;**45**:576–96.
74. Jacob F. Evolution and tinkering. *Science* 1977;**196**:1161–6.