8-10-2017 12:00 AM

Unravelling Organelle Genome Transcription Using Publicly Available RNA-Sequencing Data

Matheus Sanita Lima
The University of Western Ontario

Supervisor
David Roy Smith
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Abstract

The study of organelles helped forge theories of genome evolution because of their unconventional genomes and gene expression regimes. The organelle genomics field (~35 years old) has seen the development of next generation sequencing (NGS) techniques and the consequent skyrocketing of genomic and transcriptomic data. However, these data are being underused in the studies of organelle genome transcription. My thesis investigates how NGS has affected the field of organelle genomics at both the DNA and RNA levels. First, I demonstrate that although organelle genomes are being sequenced as never before, they are un-characterized as they are published mostly as "organelle genome reports". Then, I show that publicly available RNA-sequencing data represent an untapped datasource to study organelle genome transcription. I uncover the widespread pervasive transcription of organelle genomes across eukaryotes and speculate that this mechanism might have influenced the evolution of land plant terrestrialization and trophic mode determination in mixotrophs.

Keywords

RNA-seq, organelle gene expression, mitochondrial genome, plastid genome, nucleomorph genome, apicoplast genome, pervasive transcription

Co-Authorship Statement

Chapter 2 of this thesis has been published in Molecular Ecology Resources. Laura C. Woods, Matthew W. Cartwright, David R. Smith and I (Matheus Sanitá Lima) collaborated in this work. David conceived and designed the study. Laura, Matthew and I collected and analysed the data. David, Laura, Matthew and I wrote and revised the manuscript.

Chapter 3 of this thesis has been submitted to G3 (G3/2017/045096). David R. Smith and I collaborated in this work. David and I designed the study. I performed the analyses and wrote the manuscript. David helped interpret the data and revise the manuscript.

Chapter 4 of this thesis has been submitted to Genome Biology Evolution (GBE-170722). David R. Smith and I collaborated in this work. David and I designed the study. I performed the analyses and wrote the manuscript. David helped interpret the data and revise the manuscript.

Appendix A of this thesis has been published in Briefings in Bioinformatics. David R. Smith and I collaborated in this work. David designed the study. David and I collected data, David wrote the manuscript and I helped revise the manuscript.

Acknowledgements

I thank David, my supervisor, for the support, guidance and freedom given to me while I have been pursuing my Master's degree. I thank André Lachance and Ryan Austin, my advisors, for their help and guidance as well. Thanks to André Lachance, Jeremy Mcneil and Anne Simon too for the great conversations and pearls of wisdom.

Thanks to Curtis, Badru and Oscar for helping me navigating the challenging, but rewarding, path of graduate school. Finally, I thank everyone (either from the Department of Biology or from elsewhere) that helped me in anyhow in these two last years. The list would be endless, but each of those people knows how grateful I am to them. Thank you!

Table of Contents

Abstract i
Co-Authorship Statement ii
Acknowledgements iii
Table of Contents iv
List of Tables vi
List of Figures vii
List of Appendices viii
Chapter 1 1
1. Next Generation Sequencing (NGS) and organelle genomics 1
1.1 Introduction 1
The impact of NGS on organelle genomics 1
1.2 Thesis rationale and objectives 3
1.3 References 5
Chapter 2 8
2 The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes 8
2.1 Introduction 9
2.2 A snapshot of the experimental methods used in contemporary organelle genome papers 11
2.3 The good, the bad, and the ugly of organelle genomics 14
2.4 Limitations and implications of a "sequence-only" approach to organelle genomics 17
2.5 Concluding remarks 19
2.6 References 21
Chapter 3 27
3 Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists 27
3.1 Introduction 28
3.2 Materials and Methods 30

3.3 Results 30
Little genome, big RNA: genome-wide, polycistronic transcription in algal organelle DNAs 30
3.4 Discussion 38
RNA-seq: an untapped resource for organelle research 38
3.5 Conclusions 42
3.6 References 43
Chapter 4 49
4 Pervasive transcription of mitochondria, chloroplasts, cyanelle and nucleomorphs across plastid bearing protists 49
4.1 Introduction 50
4.2 Materials and Methods 51
4.3 Results 52
Pervasive organelle transcription is a widespread feature across eukaryotes 52
4.4 Discussion 62
4.5 References 67
Chapter 5 73
5. Organelles, revolutionary model systems 73
5.1 Concluding remarks 73
From endosymbiosis to land plant terrestrialization 73
5.2 References 76
Appendices 79
Curriculum Vitae 92

List of Tables

Table 3.1 Diverse organelle genomes and their RNA mapping statistics..... 32
Table 4.1 Mitochondrial, plastid and nucleomorph genomes from the species studied and their RNA mapping statistics..... 55

List of Figures

Figure 2.1 A survey of organelle genome papers published in the last half decade.....13
Figure 3.1 Pervasive organelle genome transcription across the eukaryotic tree of life... 34
Figure 3.2 Full transcription of small mitochondrial genomes in Apicomplexa.....35
Figure 3.3 Polycistronic transcription in mitochondrial genomes of chlorophytes, raphidophytes, and glaucophytes. 37
Figure 3.4 Entire and near entire transcriptional coverage of diverse plastid genomes.... 38
Figure 4.1 Occurrence of pervasive organelle and nucleomorph genome transcription across plastid-bearing prostists.... 54
Figure 4.2 Full transcription of bloated mitochondrial genomes in land plants.... 59
Figure 4.3 Full transcription of nucleomorph genomes in cryptophytes... 61
Figure 1A Available data in GenBank for exploring organelle transcription in plastid-bearing eukaryotes.. 80

List of Appendices

Appendix A: Unraveling chloroplast transcriptomes with ChloroSeq, an organelle RNA-Seq bioinformatics pipeline..... 79

Chapter 1

1. Next Generation Sequencing (NGS) and organelle genomics

1.1 Introduction

The impact of NGS on organelle genomics

Although the contribution of mitochondria to the origin of eukaryotes is still debatable (Martin et al. 2015; Pittis and Galbadón 2016), it is agreed that mitochondria came from the endosymbiosis between an archeaon and an alphaproteobacterium (Ku et al. 2015). It is also widely accepted that the origin of mitochondria was a single event that happened between 1.5 and 1.8 billion years ago, according to the fossil record (Javaux et al. 2001; Parfrey et al. 2011; Martin et al. 2017). Chloroplasts were established later, between 1.5 and 1.2 billion years, but they emerged through the very same process as mitochondria – an endosymbiotic event (Dyall et al. 2004). This time, the endosymbiotic relationship was between a heterotrophic protist (already mitochondriate) and a cyanobacterium. This single event marked the emergence of eukaryotic photosynthesis and the monophyletic lineage Archaeoplastida (Gould et al. 2015). Since then, eukaryotic photosynthesis has been laterally acquired through a series of secondary and even tertiary endosymbioses (Burki et al. 2014), which gave rise to the so-called "complex" plastids (Keeling 2013).

Organelles carry their own DNA inherited from their once free-living bacterial counterparts (Allen and Martin 2016). However, the transition from an endosymbiont to a fully-fledged organelle is primarily characterized by the loss of genetic material from the endosymbiont to the host, a process called endosymbiotic gene transfer (EGT) (Timmis et al. 2004). EGT culminates in genome reduction and consequent dependence of the endosymbiont on the host (Embley and Martin 2006). In other words, current organelles should carry genomes (if any) much smaller than their bacterial relatives.

Surprisingly, organelle genomes exhibit a genome size variation of orders of magnitude, reaching genome sizes larger than those of some bacterial genomes (Smith and Keeling 2015). Most of this size variation comes from the expansion of noncoding DNA that was very likely fixed by nonadapative mechanisms such as genetic drift and differences in mutation rates (Lynch et al 2006). Organelle genomes also show immeasurable diversity in structure and content. Gene and chromosome number variation (Shao et al. 2012; Janouškovec et al. 2013), amount of foreign DNA uptake (Smith 2011; Straub et al. 2013) and variable genome topologies (Nosek and Tomáska 2003; Smith et al. 2010) are just a few examples of how eccentric and diverse organelle genomes can be. These peculiarities have helped researchers forge theories of molecular evolution as they tried to make sense of such genomic features (Lynch et al. 2006; Lynch 2007; Gray et al. 2010).

The expression of organelle genomes is similarly convoluted (Smith and Keeling 2016). Noncanonical genetic codes (Jukes and Osawa 1990; Matsumoto et al. 2011), translational bypassing (Masuda et al. 2010; Lang et al. 2014), trans-splicing guided by anti-sense RNAs (Vlcek et al. 2011) and heavy RNA editing (Simpson et al. 2006) exemplify how unconventional organelle gene expression can be. On top of this, organelles respond to the nucleus and juggle with organellar and nuclear expression machineries (Cahoon and Stern 2001; Barkan 2011). Therefore, the expression of their genes is governed by the interaction(s) between nucleus and organelles (via retrograde and anterograde signalling) and between cellular compartments and environmental stimuli (Woodson and Chory 2008).

Most of what we know about organelle genomes and their transcription comes from single gene studies that took years of hard molecular biology work (Sanitá Lima et al. 2016). After all, organelle genomics established as a field only 36 years ago with the sequencing of the human (Anderson et al. 1981) and mouse mitochondrial genomes (Bibb et al. 1981), followed by the tobacco (Shinozaki et al. 1986) and *Marchantia polymorpha* (Ohyama et al. 1986) plastid genomes. Since then, sequencing technologies have improved (Metzker 2010) and organelle genomes currently are one of the most sequenced types of chromosomes (Smith 2016). That is not only because of their relatively small sizes (with a few exceptions), but also because of their importance to

fields such as phylogenetics (Daniell et al. 2016), forensics (van Oven and Kayser 2009), medicine (Picard et al. 2016) and archaeology (Pérez-Zamorano et al. 2017). More recently, the advent of next generation sequencing (NGS) techniques has contributed to the explosion of sequenced organelle genomes (Sanitá Lima et al. 2016). However, how was this contribution? What are the impacts and implications of NGS to the investigation of organelle genomes at both DNA and RNA levels?

1.2 Thesis rationale and objectives

NGS revolutionized Biology (Goodwin et al. 2016); it brought Biology to the realm of big data sciences (Mattmann 2013) and helped establishing the "-omics" approach to biological questions. Genomics (Hawkins et al. 2010), transcriptomics (Breschi et al. 2017), epigenomics (Orlando et al. 2015) and metagenomics (Kelley et al. 2016) are a few examples of areas of study that have been inundated with data coming from NGS projects. Organelle genomics is no exception (Smith and Keeling 2015). As already mentioned, organelle genomes are one of the most sequenced types of chromosomes and certainly NGS has contributed to that (Sanitá Lima et al. 2016). However, how much of the organelle genomes are being sequenced through NGS techniques? Is NGS equally applied to mitochondrial and plastid genomes, or do their size differences play a role in how we sequence them? I sought to investigate the impact of NGS on organelle genomics by trying to answer these questions first.

My colleagues and I analysed over 2,500 organelle genome papers published in the last five years (Chapter 2). We sorted them according to their sequencing techniques, the organisms studied and the types of journals that published those findings. With that, we identified trends within the field of organelle genomics and potential gaps to be filled, such as the underuse of RNA-seq data to study organelle genome transcription.

Therefore, knowing that public databases such as the Sequence Read Archive (SRA) from NCBI are ballooning with genomic and transcriptomic data (Smith and Sanitá 2017), I sought to test the utility of whole cell RNA-sequencing (RNA-seq) data to study organelle genome transcription. I predicted to find publicly available transcriptomic data

from species of all major eukaryotic groups, but I decided to sample only plastid-bearing taxa to make this project feasible. I chose organisms for which I could find RNA-seq datasets and full organelle genomes sequenced. Then, I performed RNA mapping analyses to determine how much of each genome is being transcribed. I hypothesized that small and compact organelle genomes (i.e. poor in noncoding DNA) would be fully covered by transcripts, whereas large and bloated genomes (i.e. rich in noncoding DNA) would have coding regions covered by transcripts interspersed with "deserts" of no transcription (i.e. noncoding DNA). Small and compact organelle genomes were first analysed and followed our expectations that they are fully transcribed (Chapter 3). However, big and bloated genomes exhibited full transcription as well, probably producing several noncoding RNAs (ncRNAs) with potential regulatory functions (Chapter 4). In the light of organelle genome size variation, I speculate that such ncRNAs might have played a role in the evolution of land plant terrestrialization and trophic mode determination in mixotrophs. I underscore the utility of publicly available RNA data to study organelle genome transcription and to determine organelle genomes not yet sequenced (Chapters 3 and 4). Finally, I explain the limitations of my approach and discuss future avenues of research in organelle genomics focusing in the ncRNA sphere (Chapter 5). Together with David, I also point to alternative analyzes of plastid genome transcription using ChloroSeq, a bioinformatics pipeline that employs RNA-seq data to investigate RNA editing, splicing efficiency and expression patterns in plastid genomes (Appendix A).

1.3 References

Allen JF, Martin WF. 2016. Why have organelles retained genomes? Cell Systems. 2:70-72.

Anderson S, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457-465.

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell. 26:167-180.

Breschi A, Gingeras TR, Guigó R. 2017. Comparative transcriptomics in human and mouse. Nat Rev Genet. 18:425-440.

Burki F, et al. 2014. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. Eukaryot Cell. 13:246-255.

Daniell H, Lin CS, Yu M, Chang WJ. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 17:134.

Dyall SD, Brown MT, Johnson PJ. 2004. Ancient invasions: from endosymbionts to organelles. Science. 304:253-257.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623-630.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 17:333-351.

Gould SB, Maier UG, Martin WF. 2015. Protein import and the origin of red complex plastids. Curr Biol. 25:R515-R521.

Hawkins RD, Hon GC, Ren B. 2010. Next-generation genomics: an integrative approach. Nat Rev Genet. 11:476-486.

Javaux EJ, Knoll AH, Walter MR. 2001. Morphological and ecological complexity in early eukaryotic ecosystems. Nature. 412:66-69.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 64:583–607.

Kelley JL, Brown AP, Therkildsen NO, Foote AD. 2016. The life aquatic: advances in marine vertebrate genomics. Nat Rev Genet. 17:523-534.

Ku C, et al. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. Nature. 524:427-432.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science. 311:1727-1730.

Martin WF, et al. 2017. Late mitochondrial origin is an artefact. Genome Biol Evol. 9:373-379.

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. Phil Trans R Soc B. 370:20140330.

Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. Microbiol Mol Biol Rev. 81:e00008-17.

Mattmann CA. 2013. Computing: a vision for data science. Nature. 493:473-475.

Ohyama K, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature. 322:572-574.

Orlando L, Gilbert MTP, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 16:395-408.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA. 108:13624-12629.

Pérez-Zamorano B, et al. 2017. Organellar genomes from a ~5,000-year-old archaeological maize sample are closely related to NB genotype. Genome Biol Evol. 9:904-915.

Picard M, Wallace DC, Burelle Y. 2016. The rise of mitochondria in medicine. Mitochondrion. 30:105-116.

Pittis AA, Galbadón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 531:101-104.

Sanitá Lima M, Woods CL, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Resour. 16:1279-1286.

Smith DR. 2016. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? Brief Funct Genomics. 15:47-54.

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.

Smith DR, Keeling PJ. 2016. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. Annu Rev Microbiol. 70:161-178.

Smith DR, Sanitá Lima M. 2017. Unraveling chloroplast transcriptomes with ChloroSeq, an organelle RNA-seq bioinformatics pipeline. Brief Bioinform. bbw088.

Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5:2043-2049.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123-135.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 30:E386-E394.

Wang L, et al. 2013. Complete sequence and analysis of plastid genomes of two economically important red algae: *Pyropia haitanensis* and *Pyropia yezoensis*. PLoS One. 8:e65902.

Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. Nat Rev Genet. 9:383-395.

Chapter 2

2 The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes

Published as: Sanitá Lima M, Woods CL, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Resour. 16:1279-1286.

Abstract

Not long ago, scientists paid dearly in time, money, and skill for every nucleotide that they sequenced. Today, DNA sequencing technologies epitomize the slogan "faster, easier, cheaper, and more," and in many ways sequencing an entire genome has become routine, even for the smallest laboratory groups. This is especially true for mitochondrial and plastid genomes. Given their relatively small sizes and high copy numbers per cell, organelle DNAs are currently among the most highly sequenced kind of chromosome. But accurately characterizing an organelle genome and the information it encodes can require much more than DNA sequencing and bioinformatics analyses. Organelle genomes can be surprisingly complex and can exhibit convoluted and unconventional modes of gene expression. Unraveling this complexity can demand a wide assortment of experiments, from pulsed-field gel electrophoresis to Southern and Northern blots to RNA analyses. Here, we show that it is exactly these types of "complementary" analyses that are often lacking from contemporary organelle genome papers, particularly short "genome announcement" articles. Consequently, crucial and interesting features of organelle chromosomes are going undescribed, which could ultimately lead to a poor understanding and even a misrepresentation of these genomes and the genes they express. High-throughput sequencing and bioinformatics have made it easy to sequence and

assemble entire chromosomes, but they should not be used as a substitute for or at the expense of other types of genomic characterization methods.

2.1 Introduction

Sequencing an entire organelle genome was once a long and arduous task. Now it is commonplace (Smith 2016a). With the advent of next-generation sequencing (NGS) technologies and sophisticated user-friendly bioinformatics software, scientists of all stripes can sequence and assemble dozens of organelle genomes in a few days or less, and often for very little money (Gan et al. 2014; Mariac et al. 2014; Tang et al. 2014). This kind of progress is great. More sequences mean more data for comparative studies and a better understanding of organelle genome evolution. Organelle sequences are used in a wide range of disciplines and analyses (Smith 2016a), from medicine to anthropology to phylogenetics, and have helped resolve major scientific questions, including the origins and diversification of eukaryotic life (Gray 2012; Keeling 2013). But accurately characterizing a genome and the information it encodes requires much more than just DNA sequencing and bioinformatics analyses, and organelle genomes are no exception.

Mitochondria and plastids harbour some of the most complex genomes and gene-expression systems of any genetic compartment (Smith and Keeling 2015). Take, for instance, the mitochondrial DNA (mtDNA) of the ichthyosporean *Amoebidium parasiticum*, which comprises several hundred small (0.3–8.3 kb) linear chromosomes (Burger et al. 2003), or the plastid DNAs (ptDNAs) of peridinin dinoflagellate algae, such as *Symbiodinium minutum*, which are distributed across multiple minicircular (~2.5 kb) molecules that can differ in copy number throughout the life cycle (Mungpakdee et al. 2014; Dorrell and Howe 2015). Equally as impressive is the giant (>11,000 kb) multi-chromosomal mtDNA of the flowering plant *Silene conica* (Sloan *et al.* 2012) and the tiny 6 kb mtDNA of *Plasmodium falciparum* (Feagin 1992), which is organized as a linear concatemer (Wilson and Williamson 1997).

In addition to being structurally diverse, organelle genomes can undergo massive amounts of post-transcriptional processing (Smith and Keeling 2016). In the euglenozoan

Diplonema papillatum, for example, *cox1* is transcribed from nine different mitochondrial chromosomes, giving nine partial transcripts that come together through trans-splicing to form a mature and intact mRNA (Vlcek *et al.* 2010). In the organelles of dinoflagellates, eleven of the twelve possible types of substitutional RNA editing (A-to-C, A-to-G, etc.) have been observed as well as a slew of other types of transcriptional modifications (Waller and Jackson 2009; Mungpakdee et al. 2014; Dorrell and Howe 2015). And this is to say nothing about nonstandard genetic codes (Knight et al. 2001), translational slippage (Masuda et al. 2010), and ribosomal jumping (Lang et al. 2014) within organelle systems.

Given this complexity, DNA sequencing data alone are often not sufficient to infer the true architecture and the resulting gene products of organelle genomes (Smith 2016a). Consequently, some of the most informative organelle genome analyses use a combination of different techniques, in addition to DNA sequencing and bioinformatics, to characterize the chromosome(s). For example, determining the mitochondrial genomic architecture of *D. papillatum* involved cloning, Sanger sequencing, high-throughput DNA and RNA sequencing, traditional and reverse-transcription PCR, DNA digestions, pulsed-field gel electrophoresis, and Southern and Northern blotting experiments, and still some of the chromosomes, coding regions, and gene products remain undefined (Marande et al. 2005; Vlcek et al. 2010; Valach et al. 2014). A similar array of techniques was used to describe the mitochondrial and plastid genomes of dinoflagellates (Nash et al. 2007; Barbrook et al. 2012; Jackson et al. 2012), and new organelle genomic features and peculiarities are still being uncovered within this lineage (Mungpakdee et al. 2014; Dorrell and Howe 2015). Although the *P. falciparum* mtDNA was completely sequenced more than twenty years ago (Feagin 1992; Wilson and Williamson 1997), it has taken another twenty years of detailed RNA work to resolve the large and small subunit rRNA genes, which are fragmented and scrambled into ~ 25 distinct coding modules (Feagin et al. 2012).

Improvements to traditional molecular biology techniques and the development of new technologies have only made it easier to characterize complex organelle genomes and their modes of repair, replication, and expression. State-of-the-art microscopes and

cameras can now provide ultra-high-resolution images of organelles and their nucleoids, which in turn is giving new insights into mitochondrial and plastid DNA maintenance (Golczyk et al. 2014; Oldenburg and Bendich 2015). Advanced PCR, gel-electrophoresis, and blotting methods are exposing the dynamic and multifarious nature of organelle chromosomes (Lewis et al. 2015) and their resulting transcripts (Wende et al. 2014). High-throughput transcriptomics and proteomics are also helping to disentangle the genetic information within organelles (Jedelský et al. 2011; Marková et al. 2015), as are new methods for exploring DNA-protein interactions, such as chromatin immunoprecipitation (Yagi et al. 2012). But many of these methods are technically challenging, time-consuming, and expensive, and unlike NGS they cannot be easily outsourced. Nevertheless, as the rate of organelle genome sequencing increases, one might expect the use of "complementary" characterization techniques, such as pulsed-field or two-dimensional gel electrophoresis (Slater et al. 1998), to also increase. However, this does not appear to be true. As described below, a scan of the recent literature reveals that apart from DNA sequencing and bioinformatics there is a paucity of experimental data in many contemporary organelle genome studies, with some notable exceptions.

2.2 A snapshot of the experimental methods used in contemporary organelle genome papers

The first completely sequenced mitochondrial genomes (human and mouse) were published more than thirty years ago, using a Sanger-sequencing approach (Anderson et al. 1981; Bibb et al. 1981). These feats were soon followed by the entire plastid genome sequencing of tobacco and the liverwort *Marchantia polymorpha* (Ohyama et al. 1986; Shinozaki et al. 1986). Over the ensuing years, organelle genome data steadily accumulated from diverse species and by the turn of the millennium, which brought improvements to automated capillary Sanger sequencing, new organelle DNA sequences were being published every month or faster (Smith 2016a). Around 2010, following the advent of massively parallel high-throughput sequencing (NGS), the production and

publication rate of organelle genome data skyrocketed, with hundreds—and more recently thousands—of sequences appearing annually (Smith 2016a).

Indeed, a PubMed search of scientific articles indexed in MEDLINE retrieved 2,601 organelle genome papers published between 1 January 2010 and 1 November 2015 (Figure 2.1; Additional File 2.1). About 92% of these papers describe mtDNAs, and 8% represent plastid genomes; these sequence data span a large breadth of eukaryotic diversity, but there is nonetheless an over representation of metazoan mtDNAs and land plant ptDNAs, and a lack of data from many protist lineages (Figure 2.1; Additional File 2.1). Although some of these trends have been documented and discussed before (Smith and Keeling 2015; Smith 2016a), no one has yet surveyed the range of methods commonly employed in organelle genome studies.

We scanned the materials and methods from organelle genome papers published since 2010 (Figure 2.1), recorded the techniques used to characterize the chromosomes, and then placed these techniques into one of the following three broad categories. (I) "DNA extraction, amplification, and sequencing." (II) "Bioinformatics," which includes, for example, genome assembly and annotation, molecular sequence alignments, phylogenetic analyses, and estimations of genetic diversity. And (III) "complementary experiments," comprising any experiments not related to DNA sequencing or bioinformatics, such as restriction endonuclease digestion, gel electrophoresis, nucleotide blotting, real-time PCR, RNA analyses/sequencing, or DNA imaging. Preparatory experiments for DNA sequencing, such as cloning or gel electrophoresis of PCR products prior to Sanger sequencing, were not considered complementary techniques.

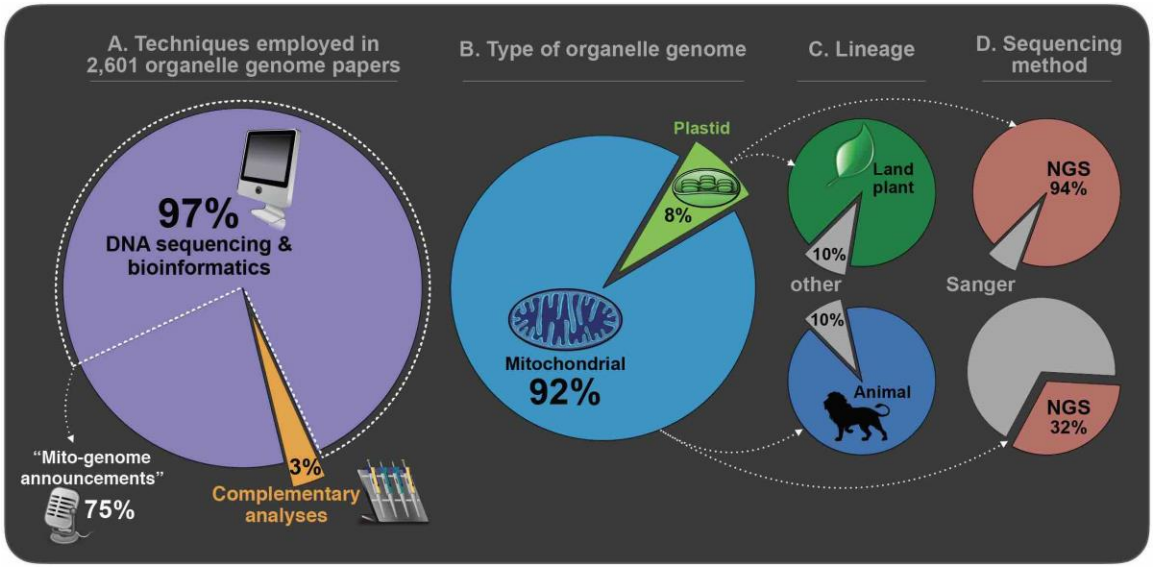

Figure 2.1 A survey of organelle genome papers published in the last half decade. Organelle genome papers indexed in MEDLINE were collected via the PubMed Advanced Search Builder at the National Center for Biotechnology Information website using the following keyword combinations: "entire chloroplast/plastid/mitochondrial DNA/genome", "complete chloroplast/plastid/mitochondrial DNA/genome", "whole chloroplast/plastid/mitochondrial DNA/genome", and "full chloroplast/plastid/mitochondrial DNA/genome". We linked the different keyword combinations with OR (instead of AND), and did not use quotation marks, in order to retrieve as many hits as possible. We limited the search field to "title/abstract," and the date range from 1 January 2010 to 1 November 2015. We scanned the results by eye, removing any obviously spurious hits. Altogether, we retrieved 2,601 organelle genome papers (including 1,781 Mitogenome Announcements), only 3% of which included complementary analyses (A). Approximately 92% and 8% of the collected articles were mitochondrial and plastid genome papers, respectively (B). The former comprised mostly animal mtDNAs, and the latter were primarily plant ptDNAs (C). Most of ptDNAs were sequenced using NGS methods (or a combination of NGS and Sanger), whereas two thirds of the mtDNAs were sequenced using a Sanger-sequencing-only approach (D). Note: "Lineage" (C) and "Sequencing Method" (D) statistics do not include Mitogenome Announcements. See Additional File 2.1 for further details.

Only a small fraction (3%) of organelle genome studies carried out over the past five years employed complementary experiments. In other words, most of the studies (97%) used only DNA sequencing and bioinformatics to characterize the chromosomes. Among the papers that did contain additional analyses, quantitative PCR was one of the most commonly employed experiments. Rarely did any of the papers include a detailed examination of organelle gene expression or chromosome structure. Instead, analyses relied upon bioinformatics software for RNA and protein predictions and for determining the size, conformation, and number of chromosomes.

The compiled articles stem from an eclectic list of mostly life-science journals, spanning an assortment of sub-disciplines (e.g., genomics, evolution, and molecular biology) and impact factors (Additional File 2.1). However, more than three-quarters of the papers come from a single journal: *Mitochondrial DNA* (formerly called *DNA Sequence*, 1990–2008), which is published by Taylor & Francis and has a Thomson Reuters impact factor of 1.2 (2014). Most of the articles collected from *Mitochondrial DNA* are "Mitogenome Announcements", short (~500 words) fast-tracked reports describing organelle genome sequences, which do not contain complementary analyses and mostly describe animal mtDNAs (Additional File 2.1). Other papers that we collected were similar to "Mitogenome Announcements" in that they were brief reports highlighting a genome sequence and its GenBank accession, including papers from the journal *Genome Announcements*, published by the American Society for Microbiology, as well as Genome Reports from the journals *Genome Biology and Evolution*. Altogether, short genome announcement-type articles (<2,000 words) represented ~75% of the papers that we surveyed.

2.3 The good, the bad, and the ugly of organelle genomics

The publication of more than 2,600 organelle genome articles over the past half-decade is an impressive achievement and a testament to how far and fast the field of genomics has progressed. (This number is likely even larger given that we could not feasibly capture every organelle genome paper using our PubMed search methods.) Together, these organelle genome data have helped to progress the field of genetics. For example, they have improved our understanding of genomic diversity and gene expression (Fitzgerald et al. 2011; Segovia et al. 2011), and yielded new insights into the mutational and population-genetic processes impacting mtDNA and ptDNA (Hardouin and Tautz, 2013). They have also advanced our understanding and/or treatment of human disease (Govindaraj et al. 2013), migration (Ning et al. 2016), and forensics (Just et al. 2015), and led to methodological advancements (Dong et al. 2013). But perhaps more than anything else, these data have provided the raw material for countless phylogenetic and

population-level studies (Njuguna et al. 2013; Taylor et al. 2013), refining our view of the origins, evolution, and diversity of eukaryotic life.

The efforts of the organelle research community to generate, annotate, and describe these genomic data are laudable. And no matter what your opinion about the impact or level of detail to which the authors analyzed these genomes, we are better off for having these data. There is no denying, however, that aside from bioinformatics analyses many published organelle genomes have not been characterized in great detail, including some of those published by the corresponding author of this perspectives piece (e.g., Smith et al. 2012; Del Vasto et al. 2015). This lack of information about organelle DNA architecture is unfortunate given that some of the most interesting aspects of these genomes are found at the structural rather than the sequence level. The paucity of detailed data on organelle chromosome structure (as discussed further below) has also likely contributed to the popular misconception that mitochondrial and chloroplast genomes typically exist as intact circular molecules, which is known to be an oversimplification (Bendich 2004, 2010; Oldenburg and Bendich 2015).

What is driving the rapid growth in organelle genomics, and why are some researchers failing to include even the most straightforward experiments in their studies? NGS techniques have streamlined genomics (Gan et al. 2014; Mariac et al. 2014; Tang et al. 2014) and certainly contributed to the massive rise in organelle DNA sequencing and publishing over the past five years (Smith 2016a). But despite these advancements, the majority of the articles examined here (>65%), including many published in the past year, employed Sanger sequencing rather than "next-generation" methods (Figure 2.1; Additional File 2.1). The continued popularity of Sanger sequencing can be partly explained by the fact that most newly sequenced organelle genomes are animal mtDNAs, which are generally small (<25 kb) and easily amplified using PCR, sometimes with a single set of primers (Cheng et al. 1994). In contrast, large organelle genomes (>50 kb), which are not amenable to PCR amplification, are now almost entirely sequenced using next-generation techniques or a combination of NGS and Sanger sequencing (Figure 2.1; Additional File 2.1).

Improved sequencing technologies may partly account for the large number of organelle DNAs being sequenced, but they cannot account for why so many investigators are ignoring traditional methods of genome characterization. One reason for the absence of additional analyses could be the growing popularity of "genome announcement" articles, which serve to highlight a DNA sequence and little else, and by their very nature are too short to permit a thorough description of the sequence (Smith 2016b). These kinds of papers are also fast to prepare and are usually accepted within a few weeks or sooner after the initial submission, thereby catering to the increasing pressure within academia to publish more and publish often (Smith 2016b). In fact, from 2009–2015 the proportion of Mitogenome Announcements in the journal *Mitochondrial DNA* rose from 50% to 80% (DeSalle 2016a), leading to the creation in 2016 of a new open-access journal called *Mitochondrial DNA Part B: Resources*, which is devoted almost entirely to short reports on whole mitochondrial genomes (DeSalle 2016b).

In defence of studies that do not include complementary analyses, many researchers who sequence and publish organelle genomes are not directly interested in or concerned with organelle genome structure or gene expression. Instead, their primary goal is to sequence organelle DNA for use in phylogenetic or population-level studies. In such cases, it might be unreasonable to expect the authors to perform a slew of complementary analyses unrelated to the questions that are being addressed—for instance, evolutionary relationships. Likewise, organelle genome sequences are sometimes generated as part of large studies, such as nuclear genome sequencing projects or broad-scale genetic diversity analyses. Again, in these instances it might be asking too much for the researchers to carry out additional analyses that are not directly connected to the project at hand. But whatever the reasons for the lack of complementary experiments in contemporary organelle genome papers, they could be negatively impacting the field of mitochondrial and plastid genomics. Soon, it might become increasingly important to incentivize more thorough analyses of organelle genomes in order to offset some of these potential negative effects.

2.4 Limitations and implications of a “sequence-only” approach to organelle genomics

There are obvious limitations and drawbacks to characterizing an organelle genome using only DNA sequencing data. Yeast mitochondrial genomes, for example, typically assemble as genome-sized circular chromosomes, leading some to assume that these chromosomes have circular conformations *in vivo*. However, it is now well established that the mtDNAs of yeast, as well as those from other groups, can have much more complex and dynamic conformations than DNA assemblies may suggest, existing (at least in part) as complex multigenomic branched structures (Bendich 1996, 2010; Gerhold et al. 2010). Similar findings have come from the ptDNAs of land plants, which typically map as circles but in many instances are found in complex linear-branched forms larger than the size of the genome, similar to those of yeast mtDNAs (Bendich 2004; Oldenburg and Bendich 2016). And there is an assortment of protists that have linear mtDNAs with elaborate telomeres: for example, the linear mitochondrial genomes of the green algae *Chlamydomonas reinhardtii* and *Polytomella capuana* end in single-stranded 3’ overhangs and covalently closed hairpin loops, respectively (Vahrenholz et al. 1993; Smith and Lee 2008). The misrepresentation of organelle chromosome conformation is so widespread that some modern biology textbooks still describe mtDNAs and ptDNAs as unit-sized circular genomes (Hartwell et al. 2014). Moving forward, elucidating the dynamic structures of organelle chromosomes will require, in the very least, extensive gel-electrophoresis work (Oldenburg and Bendich 2016).

On top of providing minimal details about genome architecture, DNA-sequencing data give limited insights into organelle transcription and translation. Mitochondria and plastids are veritable circus acts of gene expression (Smith and Keeling 2016). The mtDNAs of most metazoans, fungi, and protists have undergone one or more changes to the standard genetic code (Knight et al. 2001). Many groups undergo organelle RNA editing, whereby nucleotides are substituted, inserted, and/or deleted from transcripts. In the mitochondria of kinetoplastids, such as *Trypanosoma brucei*, uracil insertion/deletion editing can affect up to 90% of the codons in a single protein-coding transcript (Simpson

and Shaw 1989). Post-transcriptional editing can be nearly as extreme in the mitochondria and plastids of various land plants and dinoflagellates where nucleotide substitution editing is often rampant (Waller and Jackson 2009; Mungpakdee et al. 2014; Dorrell and Howe 2015). Other elaborate types of post-transcriptional processing, such as trans-splicing, transcriptional cleavage, and polyadenylation, are also widespread in mitochondria and plastids, and new idiosyncrasies are continually being uncovered (Masuda et al. 2010; Lang et al. 2014). Sometimes the levels of post-transcriptional editing and processing are so severe that given the DNA sequence alone it is not possible to distinguish coding from noncoding DNA. In such cases, data at the RNA and/or protein level are crucial to understanding the information encoded in the organelle DNA.

With notable exceptions (e.g. Mercer et al. 2011), we still have a poor understanding of organelle gene expression, especially in non-model species. But this is poised to change in the near future. There are now thousands of eukaryotic RNA-sequencing projects in GenBank's Sequence Read Archive. These publically available data abound with mitochondrial- and plastid-derived reads, most of which are unanalyzed and represent an excellent untapped resource for exploring organelle transcription (Smith 2013). Already, scientists have started publishing organelle transcriptome papers (Bundschuh et al. 2011; Kolondra et al. 2015; Wu et al. 2015; Tian and Smith 2016) or begun to include next-generation RNA sequencing data alongside whole organelle genome analyses (Fang et al. 2011; Margam et al. 2011; Jackson et al. 2012). RNA sequencing data may not be a substitute for more sophisticated transcript detection technologies, but they certainly add an additional layer of understanding and well-needed depth to any organelle genome paper. Moving forward, organelle genome studies need to combine high-throughput sequencing with molecular-biology-focused methods. This combined with information on population genetics and mutation rates, as well as a more unified understanding of cytonuclear interactions will result in some very exciting analyses. And even if these additional data are not of immediate interest to all researchers who sequence organelle genomes, then perhaps a central resource database linking the different types of experimental information for each genome would be useful.

2.5 Concluding remarks

The last thing we want to do is discourage scientists from sequencing and publishing organelle genomes, even if they are in the form of a genome announcement. Rather, we want to encourage authors to include more in-depth information about those genomes. And, again, we support the view that more genome sequence data, even if the genomes from which they are derived are not characterized in great detail, are still a scientific asset and better than no data at all. The editor-in-chief of the journal *Mitochondrial DNA*, Rob DeSalle, recently took such a stance in an eloquent commentary article defending mitochondrial genome papers:

"Publications announcing mtDNA genomes serve an important purpose in science. Access to information should be enhanced whenever we can [sic] and it seems to me that having the information about a newly sequenced mtDNA genome in the literature is an enhancing element. More importantly, an announcement can link the specimen's archival data to a sequence and clarify the provenance of a sequence. In addition, if phylogenetic analysis of the generated sequence is required (as the journal *mtDNA* requires) then the validity of the sequence can be determined by its phylogenetic placement with other known sequences" (DeSalle 2016a). These are all valid points. DeSalle (2016a) ultimately concludes: "If the incentive of publishing the findings from a novel mtDNA genome is removed … I fear that the generation of these genomes will be severely slowed and in essence a reachable goal of a mitochondrial/chloroplast DNA genomic database for all organisms on the planet with these genomes will not be realized."

A database of organelle genome sequences for all eukaryotes is an admirable goal and one that would undoubtedly contribute to the barcoding and resolution of life on Earth. Future innovations in DNA sequencing and bioinformatics will only make it easier to achieve such a goal. But these innovations should not be used as a substitute for or come at the expense of other types of genomic characterization methods.

It is important to remember that most of the greatest contributions from the field of organelle genetics have not necessarily come from the raw genome sequence data

themselves but from the complete picture of the organelle, its genome and chromosome(s), and mode of expression, including knowledge of mutation rates, population-genetic landscapes, and nuclear-encoded organelle targeted proteins. If researchers had not been striving towards this "complete" understanding we may not have seen the development of leading evolutionary theories, such as constructive neutral evolution, which was based largely on studies of organelle post-transcriptional editing and processing (Covello and Gray 1993; Stoltzfus 1999).

We will have to wait and see if the next five years bring as many new mtDNA papers as the previous five, and if those studies are short genome reports or detailed investigations. Whatever the outcome, the choice to include or not include complementary experiments will likely have a major impact on where the study ultimately gets published. Of the small fraction of papers in our survey that included additional techniques, three-quarters were published in a journal with an impact factor greater than 3. Conversely, the vast majority (>80%) of papers that contained only DNA sequencing and bioinformatics data were published in a journal with an impact factor less than 2. So if you are planning to write an organelle genome paper there is a lot to think about—or not.

Additional Files

Additional File 2.1: Table S2.1. Methodological survey of organelle genome papers indexed in MEDLINE from 1 January 2010 to 1 November 2015. (XLSX 265KB)

2.6 References

Anderson S, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457–465.

Barbrook AC, et al. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. Plant Mol Biol. 79:347–357.

Bendich AJ. 1996. Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. J Mol Biol. 255:564–588.

Bendich AJ. 2004. Circular chloroplast chromosomes: the grand illusion. Plant Cell. 16:1661–1666.

Bendich AJ. 2010. The end of the circle for yeast mitochondrial DNA. Mol Cell. 39:831–832.

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell. 26:167–180.

Bundschuh R, Altmüller J, Becker C, Nürnberg P, Gott JM. 2011. Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA. Nucleic Acids Res. 39:6044–6055.

Burger G, Forget L, Zhu Y, Gray MW, Lang BF. 2003. Unique mitochondrial genome architecture in unicellular relatives of animals. Proc Natl Acad Sci USA. 100:892–897.

Cheng S, Higuchi R, Stoneking M. 1994. Complete mitochondrial genome amplification. Nat Genet. 7:350–351.

Covello PS, Gray MW. 1993. On the evolution of RNA editing. Trends Genet. 9:265–268.

Del Vasto M, et al. 2015. Massive and widespread organelle genomic expansion in the green algal genus *Dunaliella*. Genome Biol Evol. 7:656–663.

DeSalle R. 2016a. Comments on Smith (2015)–'The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs'. Brief Funct Genomics. 15:373.

DeSalle R. 2016b. To new authors and readers of Mitochondrial DNA Part B: Resources. Mitochondrial DNA B Resour. 1:1.

Dong W, Xu C, Cheng T, Lin K, Zhou S. 2013. Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. Genome Biol Evol. 5:989–997.

Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: Lessons learned from dinoflagellates. Proc Natl Acad Sci USA. **112**:10247–10254.

Fang Y, et al. 2012. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. PloS One. 7:e37164.

Feagin JE. 1992. The 6 kb element of *Plasmodium falciparum* encodes mitochondrial cytochrome genes. Mol Biochem Parasitol. 52:145–148.

Fitzgerald TL, et al. 2011. Genome diversity in wild grasses under environmental stress. Proc Natl Acad Sci USA. 108:21140–21145.

Gan HM, Schultz MB, Austin CM. 2014. Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. BMC Evol Biol. 14:19.

Gerhold JM, Aun A, Sedman T, Jõers P, Sedman J. 2010. Strand invasion structures in the inverted repeat of *Candida albicans* mitochondrial DNA reveal a role for homologous recombination in replication. Mol Cell. 39:851–861.

Golczyk H, et al. 2014. Chloroplast DNA in mature and senescing leaves: a reappraisal. Plant Cell. 26:847–854.

Govindaraj P, et al. 2013. Mitochondrial DNA variations in Madras motor neuron disease. Mitochondrion. 13:721–728.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harb Perspect Biol. 4:a011403.

Hardouin EA, Tautz D. 2013. Increased mitochondrial mutation frequency after an island colonization: positive selection or accumulation of slightly deleterious mutations? Biol Lett. 9:20121123.

Hartwell LH, et al. 2014. Genetics: From Genes to Genomes. McGraw-Hill, Toronto, Canada.

Jackson CJ, Gornik SG, Waller RF. 2012. The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome Biol Evol. 4:59–72.

Jedelský PL, et al. 2011. The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. PloS One. 6:e17285.

Just RS, et al. 2015. Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three US populations. Forensic Sci Int Genet. 14:141–155.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 64:583–607.

Knight RD, Freeland SJ, Landweber LF. 2001. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet. 2:49–58.

Kolondra A, Labedzka-Dmoch K, Wenda JM, Drzewicka K, Golik P. 2015. The transcriptome of *Candida albicans* mitochondria and the evolution of organellar transcription units in yeasts. BMC Genomics. 16:827.

Lang BF, et al. 2014. Massive programmed translational jumping in mitochondria. Proc Natl Acad Sci USA. 111:5926–5931.

Lewis SC, et al. 2015. A rolling circle replication mechanism produces multimeric lariats of mitochondrial DNA in *Caenorhabditis elegans*. PLoS Genet. 11:e1004985.

Marande W, Lukeš J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell. 4:1137–1146.

Margam VM, et al. 2011. Mitochondrial genome sequence and expression profiling for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). PloS One. 6:e16444.

Mariac C, et al. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour. 14:1103–1113.

Masuda I, Matsuzaki M, Kita K. 2010. Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of *Perkinsus marinus* (Alveolata; Dinoflagellata). Nucleic Acids Res. 38:6186–6194.

Marková S, Filipi K, Searle JB, Kotlík P. 2015. Mapping 3′ transcript ends in the bank vole (*Clethrionomys glareolus*) mitochondrial genome with RNA-Seq. BMC Genomics. 16:870.

McPherson H, et al. 2013. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. 13:8.

Mercer TR, et al. 2011. The human mitochondrial transcriptome. Cell. 146:645–658.

Mungpakdee S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genome Biol Evol. 6:1408–1422.

Nash EA, et al. 2007 Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. Mol Biol Evol. 24:1528–1536.

Njuguna W, Liston A, Cronn R, Ashman TL, Bassil N. 2013. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. Mol Phylogenet Evol. 66:17–29.

Ning C, et al. 2016. Ancient mitochondrial genome reveals trace of prehistoric migration in the east Pamir by pastoralists. J Hum Genet. 61:103-108.

Oldenburg DJ, Bendich AJ. 2015. DNA maintenance in plastids and mitochondria of plants. Front Plant Sci. 6:883.

Oldenburg DJ, Bendich AJ. 2016. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. Curr Genet. 62:431-442.

Ohyama K, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature. 322:572–574.

Segovia R, Pett W, Trewick S, Lavrov DV. 2011 Extensive and evolutionarily persistent mitochondrial tRNA editing in velvet worms (phylum Onychophora). Mol Biol Evol. 28:2873–2881.

Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5:2043–2049.

Simpson L, Shaw J. 1989. RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa. Cell. 57:355–366.

Slater GW, Kist TB, Ren H, Drouin G. 1998. Recent developments in DNA electrophoretic separations. Electrophoresis. 19:1525–1541.

Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PloS Biol. 10:53.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. Brief Funct Genomics. 12:454–456.

Smith DR. 2016a. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? Brief Funct Genomics. 15:47–54.

Smith DR. 2016b. Goodbye genome paper, hello genome report: the increasing popularity of "genome announcements" and their impact on science. Brief Funct Genomics. 16:156-162.

Smith DR, et al. 2012. First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. Genome Biol Evol. 4:52–58.

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.

Smith DR, Keeling PJ. 2016. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. Annu Rev Microbiol. 70:161-178.

Smith DR, Lee RW. 2008. Mitochondrial genome of the colorless green alga *Polytomella capuana*: a linear molecule with an unprecedented GC content. Mol Biol Evol. 25:487–496.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. J Mol Evol. 49:169–181.

Tang M, et al. 2014. Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res. 42:e166.

Taylor JE, et al. 2013. The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. Mol Biol Evol. 30:2050–2064.

Tian Y, Smith DR. 2016. Recovering complete mitochondrial genome sequences from RNA-Seq: A case study of *Polytomella* non-photosynthetic green algae. Mol Phylogenet Evol. 98:57–62.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet. 24:241–247.

Valach M, Moreira S, Kiethega GN, Burger G. 2014. Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res. 42:2660–2672.

Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. 2010. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 39:979–988.

Waller RF, Jackson CJ. 2009. Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. BioEssays. 31:237–245.

Wende S, et al. 2014. Biological evidence for the world's smallest tRNAs. Biochimie. 100:151–158.

Wilson RJ, Williamson DH. 1997. Extrachromosomal DNA in the Apicomplexa. Microbiol Mol Biol Rev. 61:1–16.

Wu Z, Stone JD, Štorchová H, Sloan DB. 2015. High transcript abundance, RNA editing, and small RNAs in intergenic regions within the massive mitochondrial genome of the angiosperm *Silene noctiflora*. BMC Genomics. 16:938.

Yagi Y, Ishizaki Y, Nakahira Y, Tozawa Y, Shiina T. 2012. Eukaryotic-type plastid nucleoid protein pTAC3 is essential for transcription by the bacterial-type plastid RNA polymerase. Proc Natl Acad Sci USA. 109:7541–7546.

Chapter 3

3 Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists

Submitted as: Sanitá Lima M, Smith DR. 2017. Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists. G3. (G3/2017/045096).

Abstract

Organelle genomes are among the most sequenced kinds of chromosome. This is largely because they are small and widely used in molecular studies, but also because next-generation sequencing (NGS) technologies made sequencing easier, faster and cheaper. However, studies of organelle RNA have not kept pace with those of DNA, despite huge amounts of freely available eukaryotic RNA-sequencing (RNA-seq) data. Little is known about organelle transcription in non-model species, and most of the available eukaryotic RNA-seq data have not been mined for organelle transcripts. Here, we use publicly available RNA-seq experiments to investigate organelle transcription in 30 diverse plastid-bearing protists with numerous organelle genomic architectures.

Mapping RNA-seq data to organelle genomes revealed pervasive, genome-wide transcription, regardless of the taxonomic grouping, gene organization, or non-coding content. For every species analyzed, transcripts covered at least 85% of the mitochondrial and/or plastid genomes (all of which were ≤105 kb), indicating that most of the organelle DNA—coding and non-coding—is transcriptionally active. These results follow earlier studies of model species showing that organellar transcription is coupled and ubiquitous across the genome, requiring significant downstream processing of polycistronic transcripts.

Our findings suggest that non-coding organelle DNA can be transcriptionally active, raising questions about the underlying function of these transcripts and underscoring the utility of publicly available RNA-seq data for recovering complete genome sequences. If pervasive transcription is also found in bigger organelle genomes (>105 kb) across a

broader range of eukaryotes, this could indicate that non-coding organelle RNAs are regulating fundamental processes within eukaryotic cells.

3.1 Introduction

Mitochondrial and plastid DNAs (mtDNA and ptDNAs) are among the most sequenced and best-studied types of chromosome (Smith 2016a). This is not surprising given the widespread use of organelle genome data in forensics, archaeology, phylogenetics, biotechnology, medicine, and other scientific disciplines. Unfortunately, investigations of organelle RNA have not kept pace with those of the DNA, and for most non-model species there are little or no published data on organelle transcription (Sanitá Lima et al. 2016). But this is poised to change.

Next generation sequencing (NGS) technologies, ballooning genetic databanks, and new bioinformatics tools have made it easier, faster, and cheaper to sequence, assemble, and analyze organelle transcriptomes (Smith 2016a). The National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), for example, currently houses tens of thousands of freely available eukaryotic RNA sequencing (RNA-seq) datasets (Kodam et al. 2012), hundreds of which come from non-model species and/or poorly studied lineages (Keeling et al. 2014). Among their many uses, these data have proven to be a goldmine for mitochondrial and plastid transcripts (Smith 2013; Shi et al. 2016; Tian and Smith 2016).

Recently, researchers have started mining the SRA for organelle-derived reads, and already these efforts have yielded interesting results, such as pervasive organelle transcription—i.e., transcription of the entire organelle genome, including coding and non-coding regions (Shi et al. 2016; Tian and Smith 2016). This kind of research has been further aided by a range of new bioinformatics software catered to assembling and analyzing organelle genomes and transcriptomes from NGS data (Castandet et al. 2016; Dierckxsens et al. 2016; Soorni et al. 2017). Nevertheless, most of the eukaryotic RNA-seq data within the SRA have not been surveyed for organelle transcripts, particularly those from plastid-bearing protists, and it is not known if pervasive organelle

transcription is a common theme among diverse eukaryotic groups. If it is, then RNA-seq could presumably be used to glean complete or near-complete organelle genomes in the presence *or* absence of DNA data, which would be particularly useful, for example, in cases where there are abundant RNA-seq data but no available DNA information.

It goes without saying that the complexities of organelle transcription cannot be unravelled solely via *in silico* RNA-seq analyses (Sanitá Lima et al. 2016). Indeed, organelle gene expression is surprisingly complex and often highly convoluted (Moreira et al. 2012), as anyone who has studied the mtDNA of *Trypanosome* spp. (Feagin et al. 1988) or the ptDNA of *Euglena* spp. (Copertino et al. 1991) can attest. If organelle transcriptional research has taught us anything over the past few decades, it is that even the seemingly simplest mtDNAs and ptDNAs can have unexpectedly complicated transcriptomes and/or modes of gene expression (Feagin et al. 1988; Copertino et al. 1991; Marande and Burger 2007; Masuda et al. 2010; Vlcek et al. 2011; Lang et al. 2014; Valach et al. 2014; Smith and Keeling 2016). Moreover, accurately and thoroughly characterizing organelle transcriptional architecture can take years of detailed laboratory work using an assortment of techniques (Marande et al. 2005; Nash et al. 2007; Barbrook et al. 2012; Feagin et al. 2012; Jackson et al. 2012; Mungpakdee et al. 2014; Dorrell and Howe 2015). That said, RNA-seq is a quick and cost-effective starting point for early exploratory work of organelle transcription, and it can help identify lineages or species with particularly bizarre or unconventional transcriptional architectures.

Here, we use publically available RNA-seq data to survey mitochondrial and plastid transcription in a variety of eukaryotic algae. To streamline and simplify our analyses, we focus specifically on species for which the mitochondrial and/or plastid genomes have been completely sequenced and are not overly long (≤105 kb). Our explorations reveal pervasive, genome-wide organelle transcription among disparate plastid-bearing protists and highlight the potential of publically available RNA-seq data for organelle research.

3.2 Materials and Methods

By scanning the SRA (using NCBI's Taxonomy Browser), we identified 30 plastid-bearing species for which there are complete mitochondrial and/or plastid genome sequences and abundant RNA-seq data. We downloaded the RNA-Seq reads from the SRA (https://www.ncbi.nlm.nih.gov/sra) and the organelle DNAs from the Organelle Genome Resources section of NCBI (https://www.ncbi.nlm.nih.gov/genome/organelle/) or GenBank (https://www.ncbi.nlm.nih.gov/genbank/). See Additional File 3.1 for detailed information on the RNA-seq and organelle genome data we downloaded, including accession numbers, sequencing technologies, read counts, organelle DNA features, and the strains used for genome and RNA sequencing.

We mapped the RNA-Seq reads to the corresponding organelle genomes using Bowtie 2 (Langmead and Salzberg 2012) implemented through Geneious v9.1.6 (Biomatters Ltd., Auckland, NZ), a user-friendly, commercial bioinformatics software suite, which contains a graphical user interface (Kearse et al. 2012). All mapping experiments were carried out using default settings, the highest sensitivity option, and a min/max insert size of 50nt/750nt; we also allowed each read to be mapped to two locations to account for repeated regions, which are common in organelle genomes (Smith and Keeling 2015). The mapping histograms shown in Figures 3.2–3.4 were extracted from Geneious.

3.3 Results

Little genome, big RNA: genome-wide, polycistronic transcription in algal organelle DNAs

After an exhaustive search of GenBank and the SRA, we identified 30 plastid-bearing protists for which there were abundant RNA-seq data as well as complete mtDNA and/or ptDNA sequences with lengths of ~100 kb or smaller. We did not include larger organelle DNAs because we wanted to reconstruct entire organelle genomes from the transcript data alone and assumed that it would be easier to do so using RNA from small to moderately sized organelle genomes. Moreover, organelle DNAs greater than 100 kb are typically repeat rich (Smith and Keeling 2015), making RNA-seq mapping much

more challenging and error-prone (Treangen and Salzberg 2011). Nonetheless, the 30 species we analyzed span the gamut of plastid-containing eukaryotic diversity, and include taxa with primary and "complex" plastids (Keeling 2013) as well as nonphotosynthetic species, such as apicomplexan parasites (Table 3.1; Figure 3.1; Additional Files 3.1 and 3.2). The organelle genomic architectures of these species vary in structure (e.g., linear- vs. circular-mapping), size (5.8–105 kb), gene repertoire (e.g., gene rich vs. gene poor), gene arrangement (e.g., intact vs. fragmented genes), and coding content (e.g., ~7.5-95%) (Table 3.1; Figures 3.2–3.4; Additional Files 3.1 and 3.2). We made sure that the RNA-seq and corresponding organelle genome data always came from the same species, but, in a few instances, they were from different strains of the same species (Additional File 3.1). It should be stressed that most of the RNA-seq experiments we sourced were generated under stress-related conditions and often using very different protocols (Additional File 3.1). But these caveats did not seem to impede the mapping experiments.

Indeed, for each of the species and genomes we explored, the raw RNA-seq reads covered the entire or nearly entire organelle DNA, regardless of taxonomic grouping, organelle type (i.e., mtDNA vs. ptDNA), or underlying genomic architecture (Table 3.1, Figure 3.1, Additional Files 3.1 and 3.2). Not only was the overall read coverage high across the various mitochondrial and plastid genomes (85-100%), but the mean read depth (reads/nt), with few exceptions, was consistently high, ranging from 5 to >23,000 (Table 3.1). Assuming the RNA-seq reads that mapped correspond to bona fide organelle-derived transcripts (see below), these findings suggest that transcription is pervasive, spanning most or all of the organelle genome, including non-coding regions, in a diversity of plastid-bearing protists.

Table 3.1 Diverse organelle genomes and their RNA mapping statistics

TAXONOMIC GROUP AND SPECIES	ORGANELLE	GENBANK ENTRY	GENOME SIZE (bp)	MEAN COVERAGE (reads/nt)	% REFSEQ[a]	% CODING[b]
API - *Theileria parva*	mt	NC_011005.1	5,895	710.934	99.7	67.5
API - *Plasmodium berghei*	mt	LK023131.1	5,957	3,111.87	100	92.4
API - *Plasmodium falciparum*	mt	AY282930.1	5,959	368.286	100	55.7
API - *Plasmodium vivax*	mt	NC_007243.1	5,990	693.631	100	56.3
API - *Babesia bovis*	mt	NC_009902.1	6,005	614.848	99.9	63.5
	api	NC_011395.1	35,107	71.60	90.2	54.1
API - *Babesia microti*	mt	LN871600.1	10,547	5.188	93.4	37
CP - *Chlamydomonas leiostraca*	mt	NC_026573.1	14,029	136.967	95.8	86.4
DF - *Symbiodinium minutum*	mt	LC002801	19,577	2,763.05	100	7.43
CP - *Chlamydomonas moewusii*	mt	NC_001872.1	22,897	59.767	86.7	55.4
CP - *Pycnococcus provasolii*	mt	GQ497137	24,321	2,942.35	99.8	87.7
PP - *Fucus vesiculosus*	mt	NC_007683.1	36,392	98.866	97.9	90
RP - *Porphyra purpurea*	mt	NC_002007.1	36,753	1,250.44	98.7	81.5
RP - *Pyropia haitanensis*	mt	NC_017751.1	37,023	24.413	85.6	63.2
PP - *Undaria pinnatifida*	mt	NC_023354.1	37,402	165.098	92.8	89.9
PP - Saccharina japonica	mt	NC_013476.1	37,657	145.915	100	89.4
EP - *Nannochloropsis oceanica*	mt	NC_022258.1	38,057	118.754	95.8	88.8

RH - *Heterosigma akashiwo*	mt	NC_016738.1	38,690	205.219	98.5	81.3
RP - *Pyropia yezoensis*	mt	NC_017837.1	41,688	16.205	88	56.6
DT - *Pseudo-nitzschia multiseries*	mt	NC_027265.1	46,283	1,261.27	96.4	71.5
CP - *Micromonas commoda*	mt	NC_012643.1	47,425	180.623	94	82.5
CP - *Helicosporidium sp.*	mt	NC_017841.1	49,343	147.453	94.7	65
	pt	NC_008100.1	37,454	103.633	98	94.9
GP - *Cyanophora paradoxa*	mt	NC_017836.1	51,557	3,355.88	94.6	58.9
CP - *Chlorella sorokiniana*	mt	NC_024626.1	52,528	23,494.23	86.6	63
CA - *Chara vulgaris*	mt	NC_005255.1	67,737	24.862	94.2	52.3
CP - *Micromonas commoda*	pt	NC_012575.1	72,585	2,854.087	93.7	67.8
CP - *Picocystis salinarum*	pt	NC_024828.1	81,133	142.060	85.5	90.6
CR - *Vitrella brassicaformis*	pt	HM222968	85,535	5,523.59	100	88.5
HP - *Emiliana huxleyi*	pt	NC_007288.1	105,309	789.915	97	85.8
HP - *Pavlova lutheri*	pt	NC_020371.1	95,281	2,771.83	99.4	81
API - *Toxoplasma gondii*	apic	NC_001799.1	34,996	1,501.45	95	80.7

mt – mitochondrion; pt – plastid; api – apicoplast; API – Apicomplexa; CP – Chlorophyta; DF – Dinoflagellates; PP – Phaeophyta; RP – Rhodophyta; EP – Eustigmatophytes; RH – Raphidophyta; DT – Diatoms; GP – Glaucophyta; CA – Charophyta; CR – Chromerida; HP – Haptophyta

[a] Percentage of the reference genome sequence that is covered by one or more reads in the mapping contig

[b] Percentage of the coding region (tRNA-, rRNA- and protein-coding genes) in the organelle genome. The "% coding" of each genome was determined for this study using the function "extract annotation" in Geneious. We extracted tRNA-, rRNA- and protein-coding (CDS) gene annotations, then excluded spurious annotations and calculated the final length of coding sequences altogether.

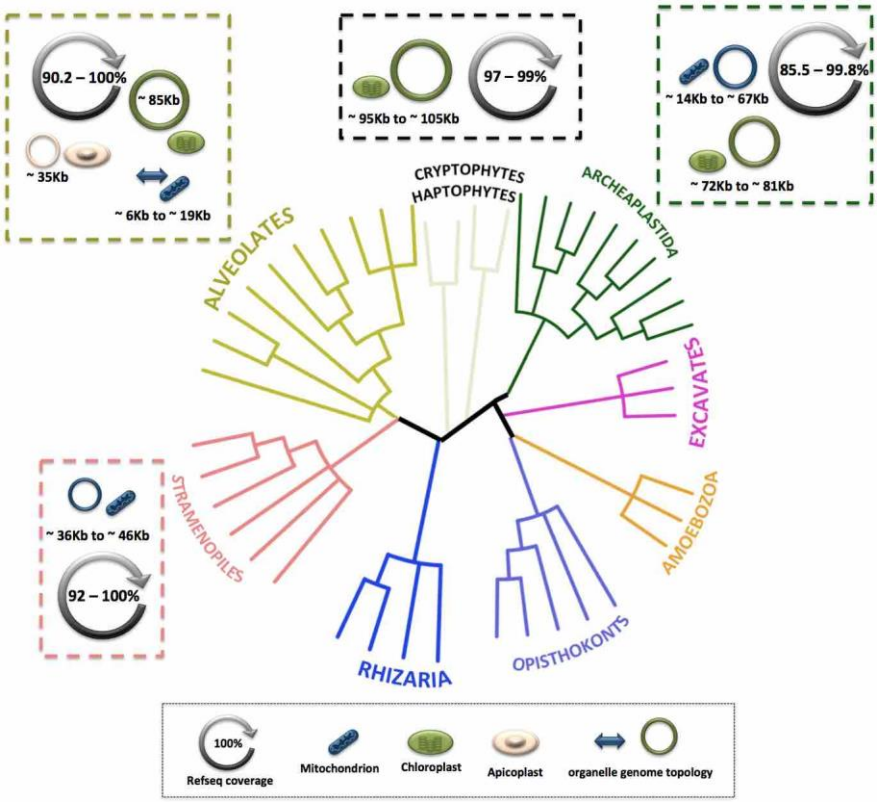

Figure 3.1 Pervasive organelle genome transcription across the eukaryotic tree of life. Organelle genomes ≤105 kb are fully or almost fully transcribed in diverse eukaryotic groups, regardless of their coding content and structure. Outer dashed boxes summarize the breadth of organelle genomes analysed within each major eukaryotic group. Representation of organelle genomes and organelles are not to scale. Refseq coverage represents the percentage of the reference genome sequence that was covered by one or more RNA-seq reads in the mapping analyses. Phylogenetic tree is adapted from (Burki 2014) for the relationships among major groups; branches within groups are merely illustrative and not based on sequence analyses. Tree was generated using NCBI Common Tree taxonomy tool (Federhen 2012) and iTOL v3.4.3 (Letunic and Bork 2016).

Close inspection of the RNA-seq mapping results revealed some interesting trends within and among the various lineages and genomes (Figures 3.2–3.4). As expected, the overall RNA read coverage was particularly high (93–100% of the reference genome) for the miniature and highly compact mtDNAs of the five apicomplexan parasites in our dataset (Figure 3.2), and when applicable (e.g., *Babesia bovis*) it extended into and encompassed the entire mitochondrial telomeres, as has been observed for linear mtDNAs from other lineages (Tian and Smith 2016). These results are consistent with earlier work on

apicomplexans showing that their mitochondrial genomes are transcribed in a polycistronic manner (Ji et al. 1996; Rehkopf et al. 2000), and reinforce the notion that mitochondrial telomeres are involved in gene expression.

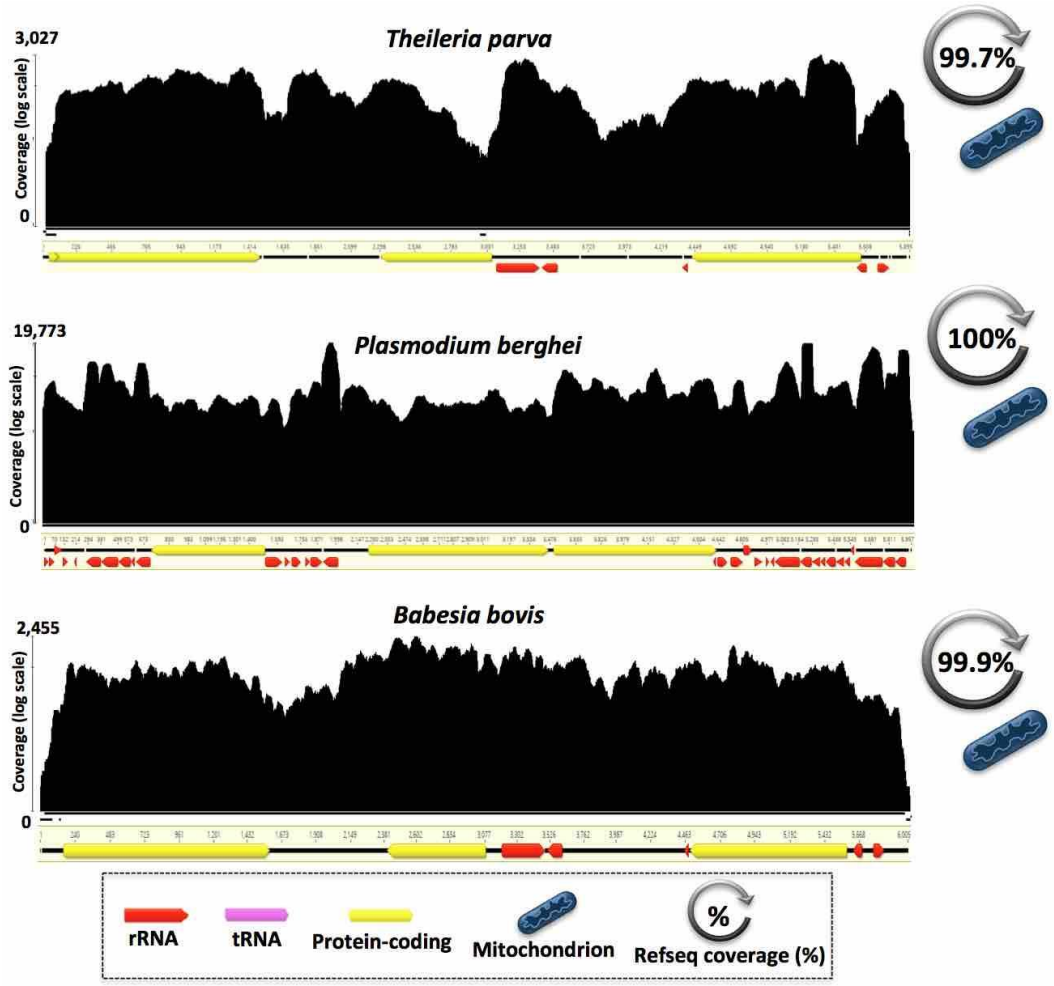

Figure 3.2 Full transcription of small mitochondrial genomes in Apicomplexa. Mapping histograms (or transcription maps) depict the coverage depth – number of transcripts mapped per nucleotide – on a log scale. We used the organelle genome annotations already present in the genome assemblies deposited in GenBank (accession numbers provided in Table 3.1 and additional file [see Additional File 3.1]). Mapping contigs are not to scale and direction of transcription is represented by the direction of the arrows – annotated genes. Mapping histograms were obtained from Geneious v9.1.6 (Kearse et al 2012).

The RNA-seq data of the circular-mapping mtDNAs from the green alga *Chlamydomonas moewusii*, the glaucophyte alga *Cyanophora paradoxa*, and the stramenopile alga *Heterosigma akashiwo* are also consistent with a polycistronic mode of

transcription, revealing deep, genome-wide RNA coverage across most of the chromosomes, including intergenic regions (Figure 3.3). Full transcription also appears to be occurring in the mtDNAs from other major algal groups, including brown algae (e.g., *Fucus vesiculosus*), red algae (e.g., *Porphyra purpurea*), dinoflagellate algae (e.g., *Symbiodinium minutum*), and diatom algae (e.g., *Pseudo-nitzschia multiseries*), as well as in both compact and moderately bloated mtDNAs (57–90% coding) (Table 3.1; Additional Files 3.1 and 3.2).

Almost identical trends were observed for the plastid genome data, all of which showed 85.5–100% RNA coverage and a mean read depth of 72–5,524 (Table 3.1, Figure 3.4). Like with the mtDNAs, the overall RNA-seq read coverage was especially high for small, compact ptDNAs, such as those from apicomplexan parasites (e.g., *Toxoplasma gondii*) (Table 3.1) and that of the nonphotosynthetic green alga *Helicosporidium* sp. (~37 kb; ~95% coding), 98% of which was represented at the RNA level (Figure 3.4). The secondary, red-algal-derived plastid genomes of the photosynthetic chromerid *Vitrella brassicaformis* and the haptophyte *Emiliana huxleyi* were also well represented in the RNA reads (100% and 97% coverage, respectively), as were those of *C. moewusii* and *H. akashiwo* (Figure 3.4). Overall, these data, alongside previous experiments (Mercer et al. 2011; Zhelyazkova et al. 2012; Shoguchi et al. 2015; Shi et al. 2016; Tian and Smith 2016), show that pervasive polycistronic transcription is the norm rather than the exception among mtDNAs and ptDNAs, and underscore the usefulness of RNA-seq for recovering whole organelle genomes, which can then be used in an array of downstream applications, such as for phylogenetic analyses, barcoding, or measuring nucleotide diversity within and among populations.

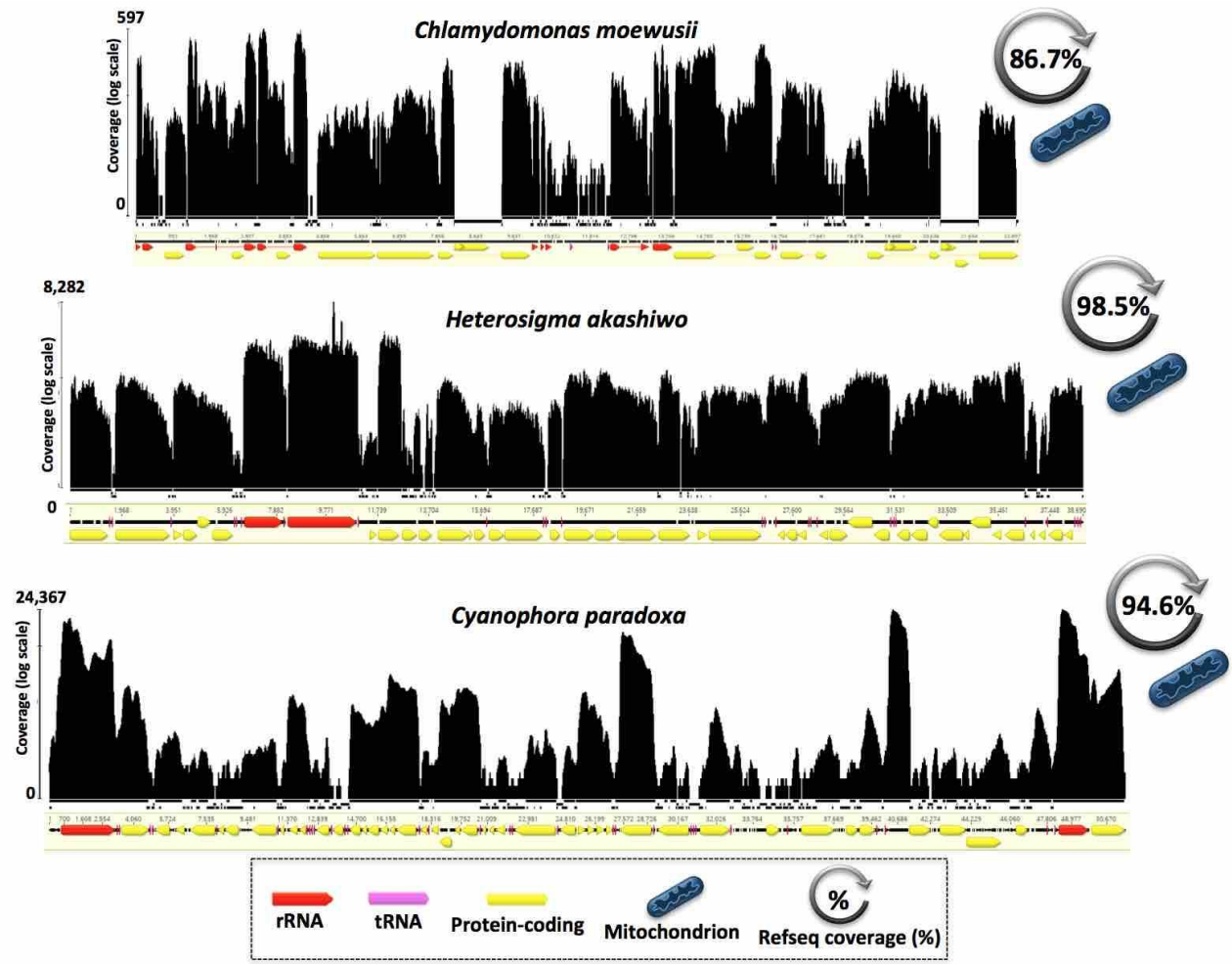

Figure 3.3 Polycistronic transcription in mitochondrial genomes of chlorophytes, raphidophytes, and glaucophytes. *Chlamydomonas moewusii* (Chlorophyta), *Heterosigma akashiwo* (Raphidophyta) and *Cyanophora paradoxa* (Glaucophyta) exhibited clear drops of transcript coverage in some potentially non-coding regions (intergenic regions, intros and hypothetical proteins). Mapping histograms follow the same structure as in Figure 3.2 and mapping contigs are not to scale.

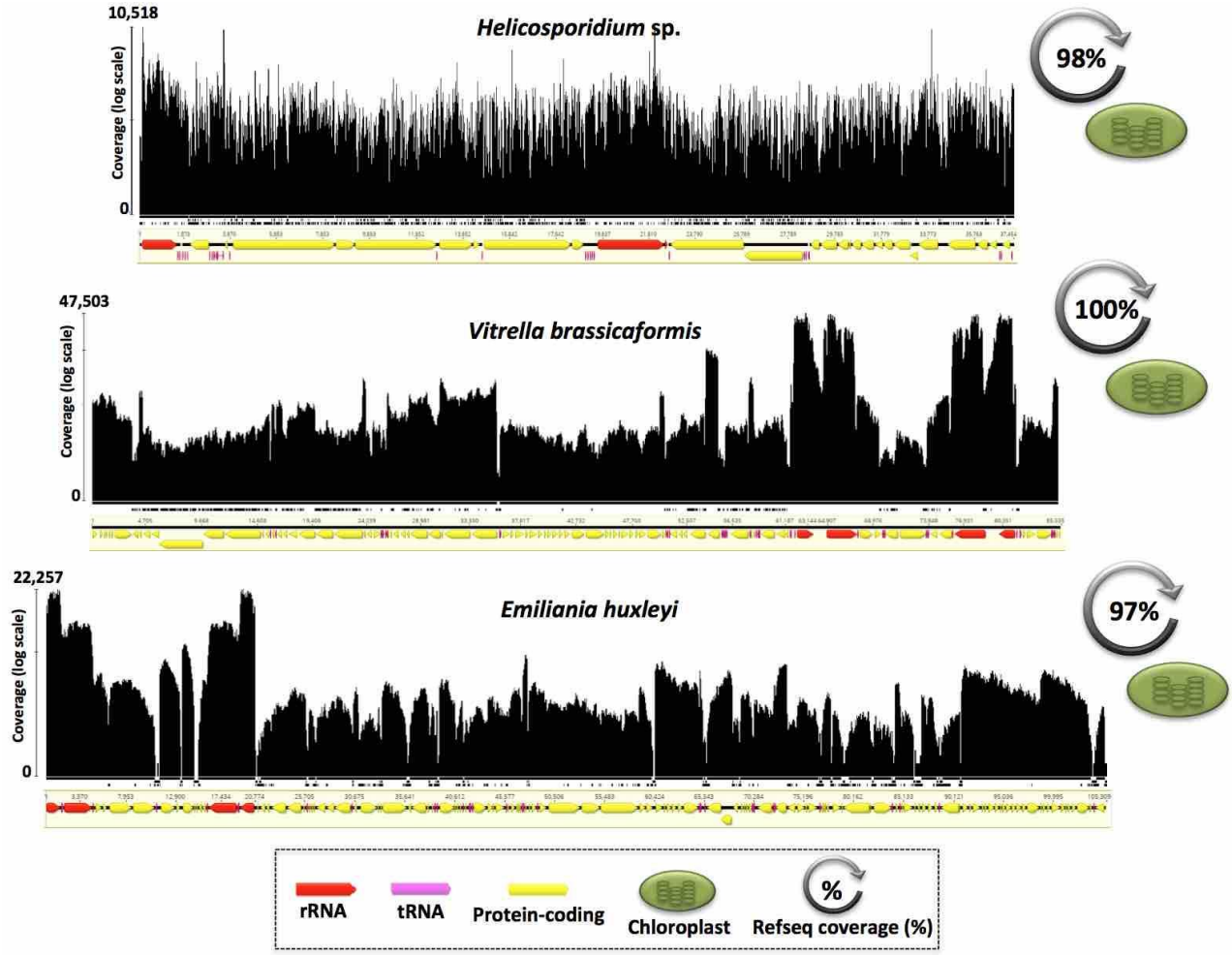

Figure 3.4 Entire and near entire transcriptional coverage of diverse plastid genomes. *Vitrella brassicaformis* (Chromerida) exhibited entire genome transcription, whereas *Helicosporidium* sp. (Chlorophyta) and *Emiliana huxleyi* (Haptophyta) had near entire genome transcriptional coverage. Drops in coverage happened mostly in intergenic regions of the *E. huxleyi* plastid genome. Mapping histograms follow the same structure as in Figures 3.2 and 3.3; mapping contigs are not to scale.

3.4 Discussion

RNA-seq: an untapped resource for organelle research

None of the RNA-seq datasets employed here were initially generated with the intent of studying organelle transcription, and to the best of our knowledge we are the first group to mine organelle transcripts from these experiments. Most, if not all, of the NGS data

used here were produced for investigating nuclear gene expression. For instance, the stramenopile alga *Nannochloropsis oceanica* is a model candidate for harvesting biofuels and, thus, the currently available RNA-seq experiments for this species are aimed at better understanding its growth and lipid production, and maximizing its economic potential (Li et al. 2014). The same can be said for many of the other species we investigated, such as the seaweeds *Undaria pinnatifida* and *Saccharina japonica*, which are harvested for food (Shan et al. 2015, Ye et al. 2015), and the apicomplexans *Babesia* sp. and *Theileria* sp., which parasitize livestock (Gardner et al. 2005; Brayton et al. 2007).

Most scientists do not have the time, resources, or expertise to explore every aspect of an NGS dataset, especially when considering the prodigious amount of information that can be contained within one. But if more scientists knew how easy it was to mine organelle transcriptomes from RNA-seq data, they might be more inclined to study various aspects of organelle genetics, even if it was merely collecting a few sequences for building a phylogenetic tree or for barcoding. And one cannot forget that organelle biology is intimately tied to that of the nucleus—to fully understand the latter one needs to study the former, and vice versa.

As shown here, and elsewhere (Shi et al. 2016; Tian and Smith 2016), complete organelle genomes can be easily and quickly reconstructed from NGS experiments, provided that these experiments were generated in a way that did not exclude organelle transcripts from the sequencing libraries. In some instances, only a single RNA-seq dataset was needed to successfully recover an entire organelle transcriptome—we recovered 99.4% of the *Pavlova lutheri* plastid genome from one 6.7 Gb paired-end RNA-seq experiment. In other cases, we had to source multiple transcriptomic experiments to recover the complete organelle genome [Additional File 3.1], suggesting that the libraries used for the cDNA sequencing were depauperate in organelle-derived transcripts. This could be because RNA-Seq libraries are often filtered for polyadenylated transcripts (mRNA) and in some lineages organelle RNA can become unstable upon polyadenylation (Rorbach et al. 2014). Other library preparation techniques, however, are much more organelle

friendly, including those that target non-coding nuclear RNAs (Di et al. 2014) as well as those catered to total cellular RNA (Hotto et al. 2011).

One must be careful not to overstate or exaggerate the usefulness of online RNA-seq data for organelle research. There are limitations to what can be deduced about gene expression from the mapping or *de novo* assembly of sequencing reads. Moreover, NGS data downloaded from public databanks can have little or no accompanying information about how they were generated, leaving users guessing about the underlying experimental conditions. And this is to say nothing about the problems of combining and comparing RNA-seq data that were generated by different laboratory groups and/or using different protocols. There is also a danger of confusing the transcripts of nuclear mitochondrial-like sequences (NUMTs) and nuclear plastid-like sequences (NUPTs) for genuine organelle RNA, but this is less of an issue for protists than it is for animals and land plants (Smith et al. 2011). Finally, there is always the possibility of genomic DNA contamination within the cDNA library, even after multiple rounds of DNAse treatment (Haas et al. 2012), but this is an issue affecting all types of RNA-seq analyses, not just those exploring organelle RNA.

Despite these drawbacks, scouring RNA-seq databases can reveal important features about organelle transcriptional architecture, such as splice variants, post-transcriptional processing, and RNA editing (Castandet et al. 2016) — or the absence of such features. For example, there were no signs of substitutional or insertion/deletion RNA editing in any of the organelle genomes we investigated, but we did detect putative polycistronic processing sites (Figures 3 and 4). RNA-seq has also helped identify transcriptional start sites in the plastid genome of barley (Zhelyazkova et al. 2012) and whole-genome transcription in land plant ptDNAs (Shi et al. 2016). Although not employed in this study, differential (d)RNA-seq and strand-specific (ss)RNA-seq can provide an even deeper resolution of organelle transcription, exposing antisense RNAs and small non-coding RNAs (Mercer et al. 2011; Zhelyazkova et al. 2012). As more dRNA-seq and ssRNA-seq experiments are deposited in the SRA (mostly from model species), they can be used to examine fine-tuned features of organelle gene expression using a similar approach to that taken here.

An emerging and recurring theme from organelle transcriptional studies (including this one) is that mitochondrial and plastid genomes are pervasively transcribed (Mercer et al. 2011; Zhelyaskova et al. 2012; Dietrich et al. 2015; Shoguchi et al. 2015; Shi et al. 2016; Tian and Smith 2016). This is also true for the genomes of alphaproteobacteria and cyanobacteria (Landt et al. 2008; Schlüter et al. 2010; Mitschke et al. 2011; Mitschke, Vioque et al. 2011; Shi et al. 2016), suggesting that pervasive organelle transcription is an ancestral trait passed down from the bacterial progenitors of the mitochondrion and plastid (Shi et al. 2016). Many nuclear genomes also show pervasive transcription (Berretta and Morillon 2009), including those of *Saccharomyces cerevisiae* (David et al. 2006), *Drosophila melanogaster* (Stolc et al. 2004), *Oryza sativa* (Li et al. 2006), and *Mus musculus* (Carninci et al. 2005). It is estimated that up to ~75% of the human nuclear genome can be transcriptionally active when looking across tissues and subcellular compartments (Djebali et al. 2012). In fact, the more we study genome-wide transcription, the more we realize that few regions in a genome are entirely exempt from transcription and that genomes are real 'RNA machines' producing multiple types of RNA from end to end (Amaral et al. 2008; Wade and Grainger 2014). Some have suggested that pervasive transcription can provide raw RNA material for new regulatory pathways (Libri 2015). However, certain bacteria can repress pervasive transcription (Lasa et al. 2011; Singh et al. 2014), so obviously it is not a good strategy all of time, at least in some systems.

It remains to be seen if big (>>100 kb) organelle genomes, such as land plant mtDNAs (Sloan et al. 2012) and chlamydomonadalean ptDNAs (Featherston et al. 2016), are fully transcribed, but preliminary work suggests that they are. RNA-seq analyses revealed complete transcription of the *Symbiodinium minutum* mtDNA (~327 kb) (Shoguchi et al. 2015), *Chlamydomonas reinhardtii* ptDNA (~204 kb), and other bloated organelle DNAs (Shi et al. 2016). If pervasive transcription is shown to be widespread in small and giant organelle genomes throughout the eukaryotic domain, then it could indicate that non-coding organelle RNAs have important, undescribed functions, or alternatively that they are transcriptional noise (Struhl 2007)—or both, depending on the RNA in question. One should be careful not to mistake transcription for function (Doolittle 2013), but non-coding organelle RNAs (both long and short) are known to carry out crucial regulatory

functions (Hotto et al. 2011; Small et al. 2013; Dietrich et al. 2015). Perhaps having more non-coding DNA and therefore more non-coding RNA leads to increased regulatory control of certain metabolic pathways within organelles (e.g., those for the development of different plastids in land plants [Jarvis and López-Juez 2013]) or more fine-tuned responses to environmental conditions (e.g., changing trophic strategies in mixotrophic algae [Worden et al. 2015]). But if so, why is there such a massive variation in organelle genome size (and transcriptome size) within and among lineages (Khaitovich et al. 2004; Lynch et al. 2006; Smith and Keeling 2015; Smith 2016b; Figueroa-Martinez et al. 2017a; Figueroa-Martinez et al. 2017b)? Alas, there is still a lot to be learned about organelle gene expression, and thankfully online RNA-seq data are there to help pave the way.

3.5 Conclusions

The primary goal of this study was to show that entire organelle genome sequences from diverse plastid-containing species can be reconstructed from publically available RNA-seq datasets within the SRA, as has been previously argued (Smith 2013). On this front, we were successful: algal mtDNAs and ptDNAs from disparate lineages consistently undergo full or nearly full transcription. Thus, available RNA-seq data are an excellent starting point and an untapped resource for exploring transcriptomic and genomic architecture from poorly studied species. Nevertheless, online RNA-seq experiments have their limitations and drawbacks, and one should be mindful when employing such data. It will be interesting to see if the major trends reported here will be borne out by future investigations, specifically those of larger organelle genomes. Ultimately, a deep understanding of organelle gene expression requires a multi-pronged approach, employing both traditional molecular biology techniques as well as more modern high-throughput methods (Sanitá Lima et al. 2016).

Additional Files

Additional File 3.1: Table S3.1. Mapping analyses details containing accessions numbers of the datasets used. (XLSX 51KB)

Additional File 3.2: Figure S3.1. Transcription maps for all 30 species analysed. (PDF 4.1MB)

3.6 References

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. Science. 319:1787-1789.

Barbrook AC, et al. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. Plant Mol Biol. 79:347–357.

Berretta J, Morillon A. 2009. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep. 10:973-982.

Brayton KA, et al. 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3:1401-1413.

Burki, F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol. 6:a016147.

Carninci P, et al. 2005. The transcriptional landscape of the mammalian genome. Science. 309:1559-1563.

Castandet B, Hotto AM, Strickler SR, Stern DB. 2016. ChloroSeq, an optimized chloroplast RNA-seq bioinformatics pipeline, reveals remodelling of the organellar transcriptome under heat stress. G3. doi:10.1534/g3.116.030783.

Copertino DW, Christopher DA, Hallick RB. 1991. A mixed group II/group III twintron in the *Euglena gracilis* chloroplast ribosomal protein S3 gene: evidence for intron insertion during gene evolution. Nucleic Acids Res. 19:6491-6497.

David L, et al. 2006. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA. 103:5320-5325.

Di C, et al. 2014. Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. Plant J. 80:848-861.

Dierckxsens N, Mardulyn P, Smits G. 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45:e18.

Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F. 2015. Organellar non-coding RNAs: emerging regulation mechanisms. Biochimie. 117:48-62.

Djebali S, et al. 2012. Landscape of transcription in human cells. Nature. 489:101-108.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA. 110:5294-5300.

Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: lessons learned from dinoflagellates. Proc Natl Acad Sci USA. 112:10247–10254.

Feagin JE, Abraham JM, Stuart K. 1988. Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. Cell. 53:413-422.

Feagin JE, et al. 2012. The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. PLoS One. 7:e38320.

Featherston J, Arakaki Y, Nozaki H, Durand PM, Smith DR. 2016. Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in volvocine green algae. Eur J Phycol. 51:369-377.

Federhen, S. 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40:D136-D143.

Figueroa-Martinez F, Nedelcu AM, Reyes-Prieto A, Smith DR. 2017. The plastid genomes of nonphotosynthetic algae are not so small after all. Commun Integr Biol. 10:e1283080.

Figueroa-Martinez F, Nedelcu AM, Smith DR, Reyes-Prieto A. 2017. The plastid genome of *Polytoma uvella* is the largest known among colorless algae and plants and reflects contrasting evolutionary paths to nonphotosynthetic lifestyles. Plant Physiol. 173:932-943.

Gardner MJ, et al. 2005. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science. 309:134-137.

Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics. 13:734.

Hotto AM, Schmitz RJ, Fei Z, Ecker JR, Stern DB. 2011. Unexpected Diversity of Chloroplast Noncoding RNAs as Revealed by Deep Sequencing of the *Arabidopsis* Transcriptome. G3. 1:559-570.

Jackson CJ, Gornik SG, Waller RF. 2012. The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome Biol Evol. 4:59–72.

Jarvis P, López-Juez E. 2013. Biogenesis and homeostasis of chloroplast and other plastids. Nat Rev Mol Cell Biol. 14:787-802.

Ji YE, Mericle BL, Rehkopf DH, Anderson JD, Feagin JE. 1996. The *Plasmodium falciparum* 6 kb element is polycistronically transcribed. Mol Biochem Parasitol. 81:211-23.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28:1647-1649.

Keeling PJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12:e1001889.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 64:583-607.

Khaitovich P, et al. 2004. A neutral model of transcriptome evolution. PLoS Biol. 2:e132.

Kodam Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40:D54-D56.

Landt SG, et al. 2008. Small non-coding RNAs in *Caulobacter crescentus*. Mol Microbiol. 68:600-614.

Lang BF, et al. 2014. Massive programmed translational jumping in mitochondria. Proc Natl Acad Sci USA. 111:5926-5931.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357-359.

Lasa I, et al. 2011. Genome-wide antisense transcription drives mRNA processing in bacteria. Proc Natl Acad Sci USA. 108:20172-20177.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44:W242-W245.

Li J, et al. 2014. Choreography of transcriptomes and lipidomes of *Nannochloropsis* reveals the mechanisms of oil synthesis in microalgae. Plant Cell. 26:1645-1665.

Li L, et al. 2006. Genome-wide transcription analyses in rice using tilling microarrays. Nat Genet. 38:124-129.

Libri, D. 2015. Sleeping beauty and the beast (of pervasive transcription). RNA. 21:678-679.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science. 311:1727-1730.

Marande W, Burger G. 2007. Mitochondrial DNA as a genomic jigsaw puzzle. Science. 318:415.

Marande W, Lukes J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell. 4:1137–1146.

Masuda I, Matsuzaki M, Kita K. 2010. Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome *c* oxidase subunit 1 gene of *Perkinsus marinus* (Alveolata; Dinoflagellata). Nucleic Adics Res. 38:6186-6194.

Mercer TR, et al. 2011. The human mitochondrial transcriptome. Cell. 146:645-658.

Mitschke J, et al. 2011a. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. Proc Natl Acad Sci USA. 108:2124-2129.

Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM. 2011b. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. Proc Natl Acad Sci USA. 108:20130-20135.

Moreira S, Breton S, Burger G. 2012. Unscrambling genetic information at the RNA level. Wiley Interdiscip Rev RNA. 3:213-228.

Mungpakdee S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genome Biol Evol. 6:1408–1422.

Nash EA, et al. 2007. Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. Mol Biol Evol. 24:1528–1536.

Rehkopf DH, Gillespie DE, Harrell MI, Feagin JE. 2000. Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial mRNAs. Mol Biochem Parasitol. 105:91-103.

Rorbach J, Bobrowicz A, Pearce S, Minczuk M. 2014. Polyadenylation in bacteria and organelles. Methods Mol Biol. 1125:211-227.

Sanita Lima M, Woods LC, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Resour. 16:1279-1286.

Schlüter JP, et al. 2010. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. BMC Genomics. 11:245.

Shan TF, Pang SJ, Li J, Li X. 2015. De novo transcriptome analysis of the gametophyte of *Undaria pinnatifida* (Phaeophyceae). J Appl Phycol. 27:1011.

Shi C, et al. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. Sci Rep. 6:30135.

Shoguchi E, Shinzato C, Hisata K, Satoh N, Mungpakdee S. 2015. The large mitochondrial genome of *Symbiodinium minutum* reveals conserved noncoding sequences between dinoflagellates and apicomplexans. Genome Biol Evol. 7:2237-2244.

Singh SS, et al. 2014. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 28:214-219.

Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10:e1001241.

Small ID, Rackham O, Filipovska A. 2013. Organelle transcriptomes: products of a deconstructed genome. Curr Opin Microbiol. 16:652-658.

Smith DR, Crosby K, Lee RW. 2011. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biol Evol. 3:365-371.

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.

Smith DR, Keeling PJ. 2016. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. Annu Rev Microbiol. 70:161-78.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. Brief Funct Genomics. 12:454-456.

Smith DR. 2016a. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? Brief in Funct Genomics. 15:47-54.

Smith DR. 2016b. The mutational hazard hypothesis of organelle genome evolution: 10 years on. Mol Ecol. 25:3769-3775.

Soorni A, Haak D, Zaitlin D, Bombarely A. 2017. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics. 18:49.

Stolc V, et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science. 306:655-660.

Struhl, K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol. 14:103-105.

Tian Y, Smith DR. 2016. Recovering complete mitochondrial genome sequences from RNA-seq: a case study of Polytomella non-photosynthetic green algae. Mol Phylogenet Evol. 98:57-62.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13:36-46.

Valach M, Moreira S, Kiethega GN, Burger G. 2014. Trans-spicling and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res. 42:2660-2672.

Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. 2011. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 39:979-988.

Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol. 12:647-653.

Worden AZ, et al. 2015. Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 347:1257594.

Ye N, et al. 2015. *Saccharina* genomes provide novel insight into kelp biology. Nat Commun. 6:6986.

Zhelyazkova P, et al. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell. 24:123-136.

Chapter 4

4 Pervasive transcription of mitochondria, chloroplasts, cyanelle and nucleomorphs across plastid bearing protists

Submitted as: Sanitá Lima M, Smith DR. 2017. Pervasive transcription of mitochondria, chloroplasts, cyanelle and nucleomorphs across plastid bearing protists. Genome Biol Evol. (GBE-170722).

Abstract

Organelle genomes exhibit remarkable diversity in content, structure, and size, and in their modes of gene expression, which are governed by both organelle- and nuclear-encoded machinery. Next generation sequencing (NGS) has generated unprecedented amounts of genomic and transcriptomic data, which can be used to investigate organelle genome transcription. However, most of the available eukaryotic RNA-sequencing (RNA-seq) data are used to study nuclear transcription only, even though large numbers of organelle-derived reads can typically be mined from these experiments. Here, we use publicly available RNA-seq data to assess organelle genome transcription in 59 diverse plastid-bearing species. Our RNA mapping analyses unravelled pervasive (full or near-full) transcription of mitochondrial, plastid, and nucleomorph genomes. In all cases, 85% or more of the organelle genome was recovered from the RNA data, including non-coding (intergenic and intronic) regions. These results reinforce the idea that organelles transcribe all or nearly all of their genomic material and are dependent on post-transcriptional processing of polycistronic transcripts. We explore the possibility that transcribed intergenic regions are producing functional non-coding RNAs, and that organelle genome non-coding content might provide raw material for generating regulatory RNAs.

4.1 Introduction

Organelle genomes can be extreme at both the DNA and RNA levels (Smith and Keeling 2015; Smith and Keeling 2016). Gene fragmentation (Barbrook et al. 2010), gene and chromosome number variation (Shao et al. 2012; Janouškovec et al. 2013), diverse genome topology (e.g., circular or linear with telomeres) (Bendich 2007), and genome size range (Sloan et al. 2012) are some of the many examples of organelle genomic diversity. Similarly, the expression of organelle genomes can be unconventional, including non-canonical genetic codes (Burger et al. 2003), substitutional or insertion/deletion RNA-editing (Castandet and Araya 2011), trans-splicing followed by polyadenylation (Vlcek et al. 2011), and even translational bypassing (Masuda et al. 2010; Lang et al. 2014). In many instances, unravelling these complicated genomic and transcriptional architectures took years of laborious investigation, using a wide range of molecular biology techniques (Sanitá Lima et al. 2016).

More recently, next generation sequencing (NGS) has allowed researchers to take a genome-wide approach to investigating organelle genomes and transcriptomes (Ruwe et al. 2013). For instance, NGS RNA sequencing (RNA-seq) of isolated organelles helped uncover pervasive transcription in the human mitochondrial genome and barley plastid genome (Mercer et al. 2011; Zhelyazkova et al. 2012). Given the popularity of NGS, organelle genome transcription can now easily be explored using publicly available RNA-seq data from whole cell experiments (Smith 2013). Such an approach revealed full transcription of plastid DNAs (ptDNAs) from various land plants and in the mitochondrial DNAs (mtDNAs) of *Polytomella* green algae (Tian and Smith 2016; Shi et al. 2016).

Most of the researchers that generate whole-cell eukaryotic RNA-seq data are not necessarily interested in organelle transcription, and many treat the organelle-derived reads as contamination, filtering them out before downstream analyses. Consequently, public databases, such the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), are increasingly becoming an untapped source for mining

organelle transcriptomic data from eukaryotic RNA-seq studies, regardless of the NGS sequencing protocol that was used (Smith and Sanitá Lima 2017).

RNA-seq data alone are rarely enough to uncover the full complexity of organelle gene expression, but they are a fast, efficient, and cost-effective first approach to studying transcription (Dietrich et al. 2015). Although pervasive transcription has been extensively demonstrated in nuclear and bacterial systems (Berretta and Morillon 2009; Wade and Grainger 2014), it is not yet known how common it is among organelle genomes. Most of the reports of genome-wide transcription in organelles come solely from model species (Hotto et al. 2012; Ro et al. 2013; Ross et al. 2016), suggesting that this strategy is the norm, rather than the exception, in mitochondria and plastids, and perhaps inherited from their bacterial progenitors (Shi et al. 2016). Here, by taking advantage of publicly available eukaryotic RNA-seq data, we investigate the transcriptional architecture of diverse plastid-bearing species, and show that pervasive transcription is a widespread phenomenon across the eukaryotic domain, including in very large organelle genomes with high non-coding contents. We speculate about the potential function roles (if any) of organelle non-coding RNAs (ncRNAs), particularly with respect to land plants and mixotrophs. If anything, these data highlight the utility of freely accessible RNA-seq data for organelle gene expression studies.

4.2 Materials and Methods

Using the NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/taxonomy), we identified 59 plastid-bearing species for which complete mitochondrial, plastid, and/or nucleomoprh genome sequences (>100 kb) and ample RNA-seq datasets were available. The RNA-Seq data were downloaded from the NCBI SRA (Kodama et al. 2011), and the genome sequences from GenBank. See Additional File 4.1 for detailed information on the RNA-seq and organelle genome data we collected, including accession numbers, read counts, sequencing technologies, organelle genome features (e.g., GC content, genome topology, and percent protein-coding), and the strains used for genome and transcriptome sequencing.

Mapping analyses were performed using Geneious v9.1.6 (Biomatters Ltd., Auckland, NZ) (Kearse et al. 2012). Briefly, raw whole-cell RNA-seq reads were mapped to the corresponding organelle genomes with Bowtie 2 (Langmead and Salzberg 2012) using the default settings, the highest sensitivity option, and a min/max insert size of 50 nt/750 nt. We allowed each read to be mapped up to two locations to account for repeated regions, which are common in organelle genomes (Smith and Keeling 2015). The mapping histograms were extracted from Geneious.

4.3 Results

Pervasive organelle transcription is a widespread feature across eukaryotes

Is organelle transcription primarily restricted to coding regions or does it extend to intergenic regions as well? Do compact versus bloated organelle genomes differ in their transcriptional patterns? Is pervasive transcription a common theme among mtDNAs and ptDNAs across the eukaryotic domain? To address these and other questions about organelle gene expression, we identified 59 diverse plastid-bearing eukaryotes for which there were abundant RNA-seq data as well as complete mtDNA and/or ptDNA sequences (and, when applicable, nucleomorph DNAs). We limited our search to species with organelle genomes that were 100 kb or greater. Previously, we explored the prevalence of pervasive transcription in small and compact organelle genomes (≤105 kb) (Sanitá Lima and Smith 2017, submitted), and here we wanted to see if the same trends held for larger organelle DNAs with long intergenic regions.

The 59 species we identified include land plants and other members of the Archaeplastida as well as various species with “complex” plastids, such as cryptophytes and stramenopiles (Figure 4.1 and Additional Files 4.1 and 4.2). The organelle genomic architectures of these species span the gamut of size (~104-980 kb), coding content (~0.6-82%;), structure (circular versus linear), and chromosome number (intact versus fragmented). We ensured that the RNA-seq and corresponding organelle genome data came from the same species, but sometimes they came from different strains of the same species (Additional File 4.1). Also, the RNA-seq experiments we sourced were often

generated using very different protocols and experimental conditions (Additional File 4.1). Nevertheless, these caveats did not hinder the mapping analyses.

For each of the organelle genomes studied here, RNA-seq reads covered 85% or more of the reference sequence (RefSeq), regardless of the genome size, non-coding content, or taxonomic grouping (Figure 4.1, Table 4.1, Additional Files 4.1 and 4.2). In 24 cases, >99% the organelle DNA sequence was present at the RNA level. In other words, all of the genomes exhibited pervasive, genome-wide transcription. The mean RNA-seq read coverage was consistently high across the different genomes, varying from ~30 to >2,300,000 reads/nt.

Together, these data indicate that non-coding regions from disparate organelle genomes are broadly transcribed, which can be clearly deduced from the RNA-seq mapping histograms (Additional File 4.2). This was true for relatively compact genomes, such as the ptDNA of the stramenopile alga *Nannochloropsis oceanica* (82% coding; RefSeq coverage 94%) as well as for the highly bloated organelle genomes (Figure 4.1 and Additional Files 4.1 and 4.2). For instance, RNA-seq coverage exceeded 90% for the very large mitochondrial genomes of the land plants *Salvia miltiorrhiza* (~499 kb, ~9.5% coding), *Capsicum annum* (~507kb, ~12% coding), *Rhazya stricta* (~548 kb, ~8% coding), *Asclepias syriaca* (~682 kb, ~5% coding), *Phoenix dactylifera* (~715 kb, ~5% coding), and *Cucurbita pepo* (~982 kb, ~15% coding) (Figure 4.2). This implies that hundreds of thousands of nucleotides of ncRNAs are being generated in these mitochondria, and within distinct groups of angiosperm (e.g., asterids, commelinids, and rosids).

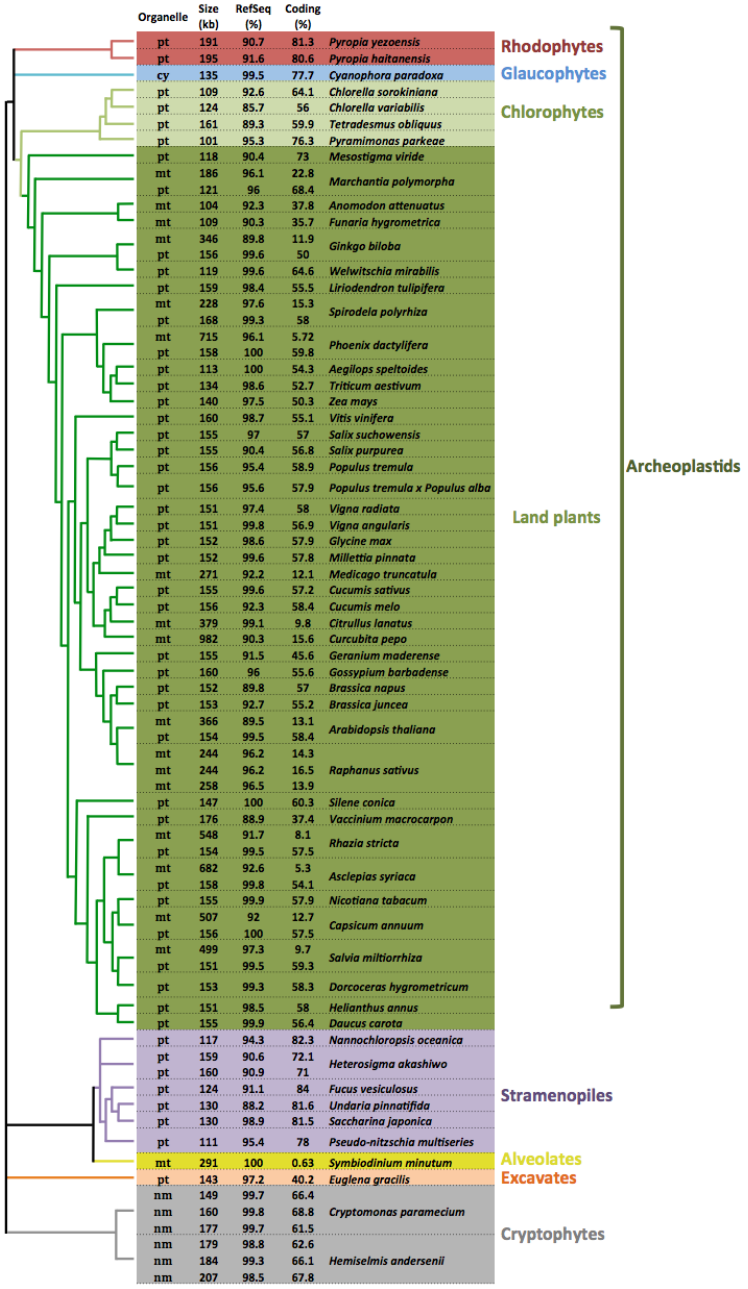

Figure 4.1 Occurrence of pervasive organelle and nucleomorph genome transcription across plastid-bearing prostists. Unscaled phylogenetic relationships were extracted from: (Stevens 2001; Wojciechowski 2006; Burki 2014; Plackett et al. 2015; Renner and Schaefer 2016). mt, mitochondrion; pt, plastid; nm, nucleomorph; RefSeq %, percentage of the reference organelle genome covered by one or more transcripts; Coding %, percentage of the amount of coding sequences (tRNA-, rRNA- and protein coding genes) in the organelle genome. The coding % was manually determined by extracting tRNA-, rRNA- and coding sequences (CDS) annotations and then subtracting spurious annotations using Geneious v9.1.6 (Kearse et al. 2012).

Table 4.1 Mitochondrial, plastid and nucleomorph genomes from the species studied and their RNA mapping statistics

TAXONOMIC GROUP AND SPECIES	ORGANELLE	GENBANK ENTRY	GENOME SIZE (bp)	MEAN COVERAGE (reads/nt)	% REFSEQ[a]	% CODING[b]
LP - *Anomodon attenuatus*	mt	NC_021931.1	104,252	30.312	92.3	37.8
LP - *Funaria hygrometrica*	mt	NC_024523.1	109,586	128.046	90.3	35.7
LP - *Marchantia polymorpha*	mt	NC_001660.1	186,609	124.778	96.1	22.8
	pt	NC_001319.1	121,024	1,690.900	96	68.4
LP - *Spirodela polyrhiza*	mt	NC_017840.1	228,493	12,523.76	97.6	15.3
	pt	NC_015891.1	168,788	38,525.506	99.3	58
LP - *Raphanus sativus*	mt	AB694743	244,036	2,701.11	96.2	14.3
	mt	KJ716484	244,054	2,713.51	96.2	16.5
	mt	AB694744	258,426	2,655.455	96.5	13.9
LP - *Medicago truncatula*	mt	NC_029641.1	271,618	327.497	92.2	12.1
DF - *Symbiodinium minutum*	mt	LC002802	291,416	2,128.72	100	0.63
LP - *Ginkgo biloba*	mt	NC_027976.11	346,544	92.582	89.8	11.9
	pt	NC_016986.1	156,988	5,666.88	99.6	50
LP - *Arabidopsis thaliana*	mt	NC_001284.2	366,924	1,659.35	89.5	13.1
	pt	NC_000932.1	154,478	39,032.50	99.5	58.4
LP - *Citrullus lanatus*	mt	NC_014043.1	379,236	556.984	99.1	9.8
LP - *Capsicum annuum*	mt	KJ865409	507,452	1,321.22	92	12.7
	pt	NC_018552.1	156,781	4,005.96	100	57.5
LP - *Rhazia stricta*	mt	NC_024293.1	548,608	56.55	91.7	8.1
	pt	NC_024292.1	154,841	264.182	99.5	57.5

LP - *Asclepias syriaca*	mt	NC_022796.1	682,498	1,241.26	92.6	5.3
	pt	NC_022432.1	158,719	12,971.22	99.8	54.1
LP - *Phoenix dactylifera*	mt	NC_016740.1	715,001	3,457.245	96.1	5.72
	pt	NC_013991.2	158,462	29,039.188	100	59.8
LP - *Curcubita pepo*	mt	NC_014050.1	982,833	1,480.88	90.3	15.6
CP - *Pyramimonas parkeae*	pt	NC_012099.1	101,605	776.192	95.3	76.3
CP - *Chlorella sorokiniana*	pt	NC_023835.1	109,811	12,424.93	92.6	64.1
DT - *Pseudo-nitzschia multiseries*	pt	NC_027721.1	111,539	29,671.42	95.4	78
LP - *Aegilops speltoides*	pt	NC_022135.1	113,536	130,214.80	100	54.3
EP - *Nannochloropsis oceanica*	pt	NC_022263.1	117,557	1,444.152	94.3	82.3
CA - *Mesostigma viride*	pt	NC_002186.1	118,360	6,314.017	90.4	73
LP - *Welwitschia mirabilis*	pt	NC_010654.1	119,726	817.69	99.6	64.6
CP - *Chlorella variabilis*	pt	NC_015359.1	124,579	2,344.05	85.7	56
PP - *Fucus vesiculosus*	pt	NC_016735.1	124,986	71.946	91.1	84
PP - *Undaria pinnatifida*	pt	NC_028503.1	130,383	1,915.687	88.2	81.6
PP - *Saccharina japonica*	pt	NC_018523.1	130,584	421.388	98.9	81.5
LP - *Triticum aestivum*	pt	NC_002762.1	134,545	21,753.04	98.6	52.7
LP - *Zea mays*	pt	KP966114	140,447	11,443.27	97.5	50.3

EG - *Euglena gracilis*	pt	NC_001603.2	143,171	7,918.18	97.2	40.2
LP - *Silene conica*	pt	NC_016729.1	147,208	51,767.34	100	60.3
LP - *Helianthus annus*	pt	NC_007977.1	151,104	458.647	98.5	58
LP - *Vigna radiata*	pt	NC_013843.1	151,271	372.165	97.4	58
LP - *Salvia miltiorrhiza*	mt	NC_023209.1	499,236	2,141,919	97.3	9.7
	pt	NC_020431.1	151,328	3,418,651	99.5	59.3
LP - *Vigna angularis*	pt	NC_021091.1	151,683	20,760.909	99.8	56.9
LP - *Glycine max*	pt	NC_007942.1	152,218	2,735.90	98.6	57.9
LP - *Brassica napus*	pt	NC_016734.1	152,860	1,584.530	89.8	57
LP - *Millettia pinnata*	pt	NC_016708.2	152,968	12,444.57	99.6	57.8
LP - *Brassica juncea*	pt	NC_028272.1	153,483	13,516.298	92.7	55.2
LP - *Dorcoceras hygrometricum*	pt	NC_016468.1	153,493	950.679	99.3	58.3
LP - *Salix suchowensis*	pt	NC_026462.1	155,214	1,739.18	97	57
LP - *Cucumis sativus*	pt	NC_007144.1	155,293	1,458.78	99.6	57.2
LP - *Salix purpurea*	pt	KP019639.1	155,590	448.062	90.4	56.8
LP - *Geranium maderense*	pt	NC_029999.1	155,694	350.685	91.5	45.6
LP - *Daucus carota*	pt	NC_008325.1	155,911	689.940	99.9	56.4
LP - *Nicotiana tabacum*	pt	NC_001879.2	155,943	2,328,505	99.9	57.9
LP - *Cucumis melo*	pt	NC_015983.1	156,017	96.536	92.3	58.4
LP - *Populus tremula*	pt	NC_027425.1	156,067	877.749	95.4	58.9
LP - *Populus tremula x Populus alba*	pt	NC_028504.1	156,641	499.792	95.6	57.9
RH - *Heterosigma akashiwo*	pt	NC_010772	159,370	708.891	90.6	72.1
	pt	EU168191	160,149	705.806	90.9	71

LP - *Liriodendron tulipifera*	pt	NC_008326.1	159,886	115.344	98.4	55.5
LP - *Gossypium barbadense*	pt	NC_008641.1	160,317	1,540.45	96	55.6
LP - *Vitis vinifera*	pt	NC_007957.1	160,928	137.518	98.7	55.1
CP - *Tetradesmus obliquus*	pt	DQ396875	161,452	32,109.500	89.3	59.9
LP - *Vaccinium macrocarpon*	pt	NC_019616.1	176,045	590.047	88.9	37.4
RP - *Pyropia yezoensis*	pt	NC_007932.1	191,952	193.022	90.7	81.3
RP - *Pyropia haitanensis*	pt	NC_021189.1	195,597	5,755	91.6	80.6
GP - *Cyanophora paradoxa*	cy	NC_001675.1	135,599	24,515.36	99.5	77.7
CT - *Cryptomonas paramecium*	nm	NC_015331	149,539	676.688	99.7	66.4
	nm	NC_015330	160,189	821.75	99.8	68.8
	nm	NC_015329	177,338	991.703	99.7	61.5
CT - *Hemiselmis andersenii*	nm	CP000883	179,593	283.158	98.8	62.6
	nm	CP000882	184,755	457.806	99.3	66.1
	nm	CP000881	207,524	360.808	98.5	67.8

mt – mitochondrion; pt – plastid; cy – cyanelle; nm – nucleomorph; CP – Chlorophyta; DF – Dinoflagellates; PP – Phaeophyta; RP – Rhodophyta; EP – Eustigmatophytes; RH – Raphidophyta; DT – Diatoms; GP – Glaucophyta; CA – Charophyta; EG – Euglenids; CT – Cryptomonads

[a] Percentage of the reference genome sequence that is covered by one or more reads.

[b] Percentage of the amount of coding sequences (tRNA-, rRNA-, and protein-coding genes) in the organelle genome. We determined this percentage by first extracting tRNA-, rRNA- and protein-coding gene annotations from the respective genome. Then, we excluded spurious annotations and calculated the resultant final length of coding sequences. We used the "extract annotation" function in Geneious v9.1.6 (Kearse et al. 2012) for that.

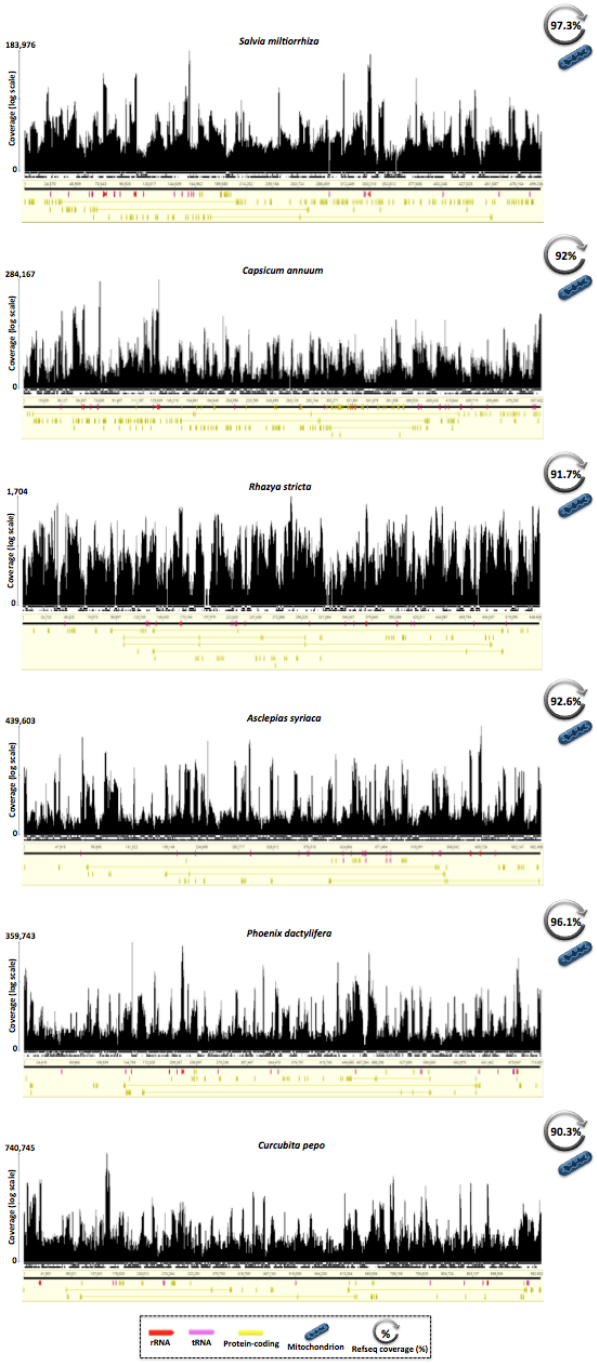

Figure 4.2 Full transcription of bloated mitochondrial genomes in land plants. Mapping histograms show coverage depth (transcripts mapped per nucleotide) on a log scale. Organelle genome annotations are from genome assemblies deposited at GenBank (accession numbers provided in Table 4.1 and Additional File 4.1). Mapping contigs are not to scale and direction of transcription is given by the direction of the arrows of the annotated genes. Mapping histograms were extracted from Geneious v9.1.6 (Kearse et al. 2012).

In fact, pervasive transcription of mitochondrial and plastid genomes appears to be the norm rather than the exception across plastid-bearing species as a whole. We found that it was common throughout the Archaeplastida, including in land plants, green algae, red algae, and glaucophytes, as well as in species with eukaryote-eukaryote derived plastids. Complete or nearly complete transcription is also found in organisms coming from very different habitats and ecosystems, such as deserts (e.g., *Welwitschia mirabilis*), irrigated cultures (e.g., *Zea Mays* and *Glycine max*), freshwater (e.g., *Tetradesmus obliquus*) and seawater (e.g., *Pyropia* spp.).

Among the most impressive examples of pervasive organelle transcription comes from the mtDNA of the dinoflagellate alga *Symbiodinium minutum* (a coral symbiont). This ~326 kb genome is made up of more than 99% non-coding DNA, all of which appears to be transcriptionally active (Figure 4.1, Additional Files 4.1 and 4.2). This result is consistent with a previous report of full mitochondrial transcription of the *S. minutum* mitochondrial genome using a different dataset (Shoguchi et al. 2015). We also observed full transcription in the nucleomorph genomes of *Cryptomonas paramecium* and *Hemiselmis andersenii* (Figure 4.3).

In some instances, organelle genome intergenic regions were not completely represented in the RNA-seq data (i.e., RefSeq coverage <100%). This is possibly a consequence of post-transcriptional processing resulting in the cleavage of those regions, thus, preventing them from being captured in the transcriptomic sequencing experiment. But even when considering these few missing regions, there is no denying that organelle genomes typically go full transcription no matter their structure, size, or content, or taxonomic grouping.

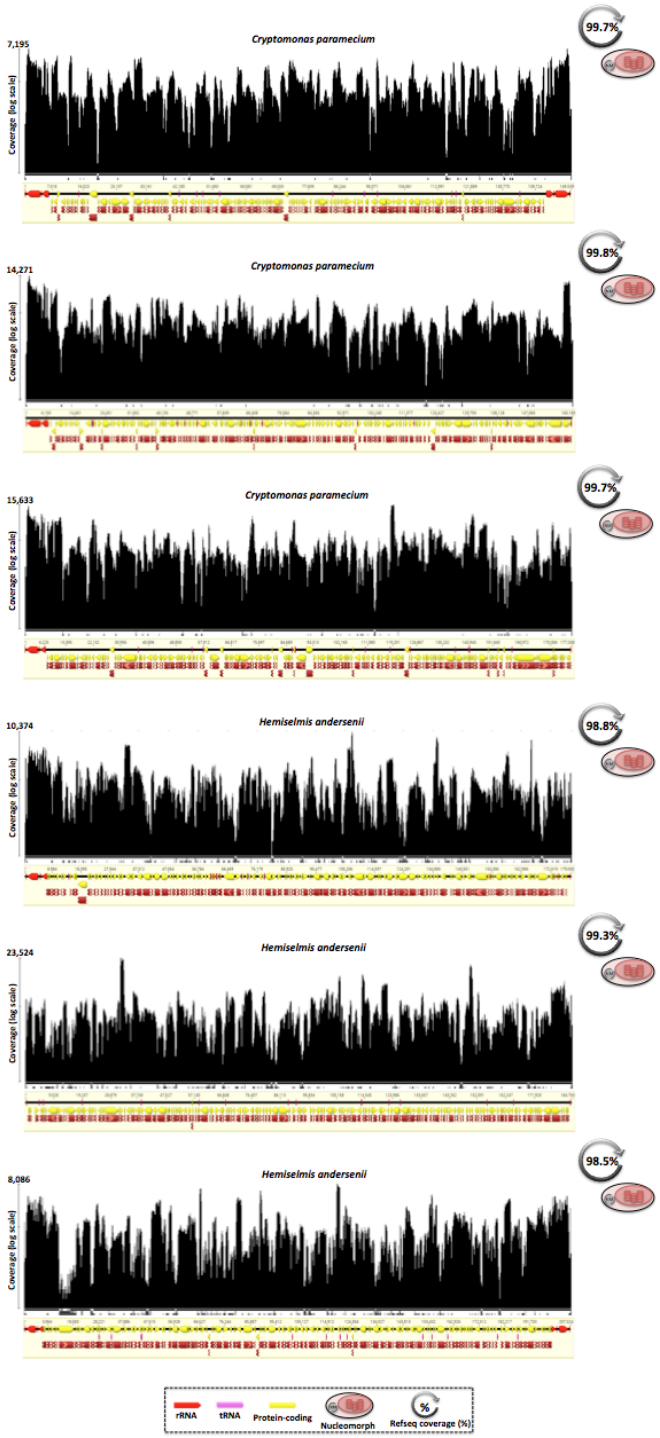

Figure 4.3 Full transcription of nucleomorph genomes in cryptophytes. *Cryptomonas paramecium* and *Hemiselmis andersenii* had full transcription in every chromosome of their nucleomorph genomes. Mapping histograms follow the same structure as in Figure 4.2 and mapping contigs are not to scale.

4.4 Discussion

Our RNA mapping analyses provide various insights into organelle transcription and how it can be investigated using publically available RNA-seq data. First, the size of the RNA-seq datasets we employed did not always positively correlate with the overall organelle genome read coverage (Additional File 4.1). This was to be expected given that the RNA-seq data we used derive from different experiments and laboratory groups and were produced under varying conditions and sequencing protocols. Poly-A selection, for example, can lead to an enrichment in highly AT-rich organelle transcripts, and in some lineages, including land plants, organelle polyadenylation is a target for transcript degradation (Small et al. 2013). But we quickly overcame any issues associated with biased or underrepresentation organelle reads by combining multiple RNA-seq datasets from different experiments (Additional File 4.1).

We also found differences in the RNA-seq coverage statistics for plastid and mitochondrial genomes. For the species which we had complete sequence data for both the mitochondrial and plastid genomes, the latter tended to have higher overall and mean coverage rates than the former. This could be connected to transcript abundance or genome copy number in plastids versus mitochondria, or perhaps the half-life of mitochondrial transcripts is shorter than that of plastid RNAs, or merely that mitochondria are responding to the experimental treatments differently than the plastid.

Many of the genomes we analyzed undergo minor to moderate amounts of substitutional RNA editing (Shoguchi et al. 2015; Shi et al. 2016). We did not set out to specifically study post-transcriptional editing, but we were able to easily identify edited sites from our mapping analyses, reinforcing the utility of freely available RNA-seq for quantifying and categorizing RNA editing in organelle systems (Smith 2013; Moreira et. al. 2016; Shi et al. 2016). Micro-RNA (miRNA) analyses were also beyond the scope of our work, but nevertheless we covered 4.5% of the *Citrullus lanatus* (watermelon) mitochondrial genome with few micro-RNA NGS datasets (data not shown). Telomeric RNA can be studied using RNA-seq: we found widespread telomeric transcription of the nucleomorph genomes from *C. paramecium* and *H. andersenii*, which is in line with previous work on

the mitochondrial telomeres of *Polytomella* spp. (Tian and Smith 2016) and apicomplexan parasites (Raabe et al. 2010). The significance of organelle telomeric transcription is not unknown, but in the nuclei of humans, mice, yeast, and zebrafish, telomeres can be transcribed into regulatory long ncRNAs called TERRA (telomeric repeat-containing RNA) (Maicher et al. 2012; Arora et al. 2012; Cusanelli and Chartrand 2015).

The utility of RNA-seq for scrutinizing organelle gene expression has its limitations and drawbacks. For example, nuclear mitochondrial-like and nuclear plastid-like DNA (NUMTs and NUPTs)—and even mitochondrial plastid-like DNA (MTPTs)—could be mistaken as bona fide organelle genome sequences in RNA-seq mapping experiments, and this is of particular concern for species with multiple mitochondria and/or plastids per cell (Smith 2011; Smith et al. 2011). Another downside to the approach used here is contamination. Genomic DNA (local or foreign) can persist in RNA-seq libraries even after treatments to eliminate it (Haas et al. 2012), but this is an issue affecting all types of RNA-seq analyses and not just those focusing on organelle transcription. Even RNA-seq data derived from isolated organelles can have contamination: we were able to recover ~97% of the *Euglena gracilis* plastid genome with RNA-seq datasets produced from isolated mitochondria (Table 4.1, Additional Files 4.1 and 4.2). Clearly, plastids and plastid RNA passed through the isolation protocol.

While accepting the shortcomings of RNA-seq, the mapping data presented here do support the idea that organelle genomes are pervasively transcribed in wide array of species. Again, this is not the first report of genome-wide organelle transcription. More than 25 years ago, Finnegan and Brown (1990) characterized the transcription of noncoding DNA in maize mitochondria. More recently, organelle ncRNAs have been described from animals and plants, some of which are candidates for gene regulation (Hotto et al. 2012; Ro et al. 2013; Ross et al. 2016). And every month brings more and more examples of complete organelle genome transcription from disparate groups throughout the eukaryotic tree of life, but the functional relevance of this is poorly understood (Vendramin et al. 2017). Similar trends are emerging from studies of nuclear genomes, where accounts of pervasive transcription are widespread, so much so that the

expressions “noncoding RNA revolution” and “eukaryotic genome as an RNA machine” are now commonplace (Amaral et al. 2008; Cech and Steitz 2014). However, there are ongoing and heated debates about whether noncoding RNAs are functional (Struhl 2007; Ponjavic et al. 2007; Doolittle 2013). No matter where you stand on the debate, there is no denying that at least some noncoding RNAs are functional, and participate in major biological process (Louro et al. 2009; Cabili et al. 2011; Esteller 2011), from synaptic plasticity (Smalheiser 2014) to cancer development (Fang and Fullwood 2016).

Given the prevalence of pervasive transcription, many are questioning/exploring the evolutionary origins of such a strategy (Ulitsky 2016). As any undergraduate genetics textbook will tell one day, pervasive genome-wide transcription is standard fare for bacteria, including alphaproteobacteria and cyanobacteria (Landt et al. 2008; Georg et al. 2009; Schlüter et al. 2010; Mitschke et al. 2011a; Mitschke et al. 2011b; Voigt et al. 2014). Thus, its widespread occurrence in organelles is arguably an ancestral trait (Shi et al. 2016). But the prevalence of full genome transcription in organelles is made more impressive by the fact that it can occur in systems with massive non-coding DNA contents (>90%), much larger than those of most bacteria. Could some of this non-coding organelle RNA have a regulatory role? And, if so, do large and bloated organelle genomes have more regulatory RNAs than their smaller, more compact counterparts?

Recent data have supported the hypothesis that ncRNAs (both long and short) carry out crucial functions within mitochondria and plastids (Vendramin et al. 2017). For example, mitochondria can produce miRNAs (Smalheiser et al. 2011) and act as a reservoir for nuclear-encoded ones (Bandiera et al. 2011), which can respond to environmental cues and regulate both cytosolic and organellar transcription (Duarte et al. 2014). Likewise, nuclear long noncoding RNAs appear to mediate crosstalk between the nucleus and mitochondrion (Vendramin et al. 2017). The nature and function of plastid and nuclear-encoded plastid-targeted noncoding RNAs are poorly understood (Zhelyazkova et al. 2012), but likely perform similar roles to those in the mitochondrion. That ncRNAs can move between organelles raises interesting questions about the transport machinery mediating this movement, most of which remain a mystery (Dietrich et al. 2015; Vendramin et al. 2017). The transport of RNA is even more complicated in the case of

complex plastids (Keeling 2013), cyanelles (Steiner and Löffelhardt 2002), and nucleomorphs (Moore and Archibald 2009).

Pervasive organelle transcription might also be involved in plastid development (and its putative link to land plant terrestrialization) as well as in trophic mode determination in mixotrophs. Plastid-specific traits, such as high-light tolerance and ptDNA architectural features, might have had a fundamental role in the evolutionary transition from water to land (de Vries et al. 2016). If true, variation in the number and types of ncRNA could have helped shape and regulate the characteristics that allowed for the terrestrialization of land plants. Land plants, for example, have an array of plastids (e.g., proplastids, chloroplasts, chromoplasts, and amiloplasts) (Jarvis and López-Juez 2013), which could likely be generated and regulated in part by ncRNAs. Similar arguments can be made for the evolution of mixotrophic algae, which can switch between heterotrophy and photoautotrophy (Jassey et al. 2015). Although speculative, the mechanisms for trophic mode determination could be partly controlled by organelle (or nuclear) ncRNAs generated via pervasive transcription. It would be interesting to explore the hypothesis that organelle genome size variation (together with organelle number) played a role in the evolution of mixotrophy. After all, non-coding sequences can be used as the raw material for generating new regulatory pathways (Libri 2015).

Although not the first account on pervasive organelle transcription, this is the first report of such widespread occurrence of this phenomenon. Most of the data used in our work came from whole-cell RNA-seq experiments in which the organelle reads were ignored. That we could use these data to assemble complete or near-complete organelle transcriptomes highlights the value of publicly available RNA-seq experiments (and the SRA) for organelle research. This work also emphasizes the ease with which one can assemble a complete organelle genome from RNA-seq data alone. A quick scan through the SRA reveals many species for which there are whole-cell RNA-seq data but no or minimal organelle DNA sequence data (Smith and Sanitá Lima 2017). Some of these species are poorly studied marine protists of great ecological importance, which had their transcriptomes sequenced as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014). As a proof of concept, fourteen

land plant plastid genomes were recently *de novo* assembled from transcriptomic data coming from SRA (Shi et al 2016). Clearly, publicly available whole cell RNA-seq data are a goldmine for organelle genomics and transcriptomics (Smith 2013). We just need to start digging.

Additional Files

Additional File 4.1: Table S4.1. Mapping analyses details containing accessions numbers of the datasets used. (XLSX 97 KB)

Additional File 4.2: Figure S4.1. Transcription maps for all 59 species analysed. (PDF 16.2 MB)

4.5 References

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. Science. 319:1787-1789.

Arora R, Brun CM, Azzalin CM. 2012. Transcription regulates telomere dynamics in human cancer cells. RNA. 18:684-693.

Bandiera S, et al. 2011. Nuclear outsourcing of RNA interference components to human mitochondria. PLoS ONE. 6:e20746.

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2010. Organization and expression of organelle genomes. Philos Trans R Soc Lond B Biol Sci. 365:785-797.

Bendich AJ. 2007. The size and form of chromosomes are constant in the nucleus, but highly variable in bacteria, mitochondria and chloroplasts. BioEssays. 29:474-483.

Berretta J, Morillon A. 2009. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep. 10:973-982.

Burger G, Gray MW, Lang BF. 2003. Mitochondrial genomes: anything goes. Trends Genet. 19:709-716.

Burki, F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol. 6:a016147.

Cabili MN, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915-1927.

Castandet B, Araya A. 2011. RNA editing in plant organelles. Why make it easy? Biochemistry. 76:924-931.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution – trashing old rules to forge new ones. Cell. 157:77-94.

Cusanelli E, Chartrand P. 2015. Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telome biology to genome integrity. Front Genet. 6:143.

de Vries J, Stanton A, Archibald JM, Gould SB. 2016. Streptophyte terrestrialization in light of plastid evolution. Trends Plant Sci. 21:467-476.

Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F. 2015. Organellar non-coding RNAs: emerging regulation mechanisms. Biochimie. 117:48-62.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA. 110:5294-5300.

Duarte FV, Palmeira CM, Rolo AP. 2014. The role of microRNAs in mitochondria: small players acting wide. Genes. 5:865-886.

Esteller M. 2011. Non-coding RNAs in human disease. Nat Rev Genet. 12:861-874.

Fang Y, Fullwood MJ. 2016. Roles, functions, and mechanisms of long non-coding RNAs in cancer. Genomics Proteomics Bioinformatics. 14:42-54.

Finnegan PM, Brown GG. 1990. Transcriptional and post-transcriptional regulation of RNA levels in maize mitochondria. Plant Cell. 2:71-83.

Georg J, et al. 2009. Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. Mol Syst Biol. 5:305.

Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics. 13:734.

Hotto AM, Germain A, Stern DB. 2012. Plastid non-coding RNAs: emerging candidates for gene regulation. Trends Plant Sci. 17:737-744.

Janouškovec J, et al. 2013. Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. PLoS ONE. 8:e59001.

Jarvis P, López-Juez E. 2013. Biogenesis and homeostasis of chloroplasts and other plastids. Nat Rev Mol Cell Biol. 14:787-802.

Jassey VEJ, et al. 2015 An unexpected role for mixotrophs in the response of peatland carbon cycling to climate warming. Sci Rep. 5:16931.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28:1647-1649.

Keeling PJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12:e1001889.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 64:583-607.

Kodama Y, Shumway M, Leinonen R. 2011. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40:D54-D56.

Landt SG, et al. 2008. Small non-coding RNAs in Caulobacter crescentus. Mol Microbiol. 68:600-614.

Lang BF, et al. 2014. Massive programmed translational jumping in mitochondria. Proc Natl Acad Sci USA. 111:5926-5931.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357-359.

Libri, D. 2015. Sleeping beauty and the beast (of pervasive transcription). RNA. 21:678-679.

Louro R, Smirnova AS, Verjovski-Almeida S. 2009. Long intronic noncoding RNA transcription: expression noise or expression choice? Genomics. 93:291-298.

Maicher A, Kastner L, Dees M, Luke B. 2012. Deregulated telomere transcription causes replication-dependent telomere shortening and promotes cellular senescence. Nucleic Acids Res. 40:6649-6659.

Masuda I, Matsuzaki M, Kita K. 2010. Extensive framshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of Perkinsus marinus (Alveolata; Dinoflagellata). Nucleic Acids Res. 38:6186-6194.

Mercer TR, et al. 2011. The human mitochondrial transcriptome. Cell. 146:645-658.

Mitschke J, et al. 2011a. An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. Proc Natl Acad Sci USA. 108:2124-2129.

Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM. 2011b. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120. Proc Natl Acad Sci USA. 108:20130-20135.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251-264.

Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G. 2016. Novel modes of RNA editing in mitochondria. Nucleic Acids Res. 44:4907-4919.

Plackett ARG, Di Stilio VS, Langdale JA. 2015. Ferns: the missing link in shoot evolution and development. Front Plant Sci. 6:972.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17:556-565.

Raabe CA, et al. 2010. A global view of the nonprotein-coding transcriptome in Plasmodium falciparum. Nucleic Acids Res. 38:608-617.

Renner SS, Schaefer H. 2016. Phylogeny and evolution of the Curcubitaceae. In: Grumet R, Katzir N, Garcia-Mas J, editors. Genetics and genomics of Curcubitaceae. Springer International Publishing.

Ro S, et al. 2013. The mitochondrial genome encodes abundant small noncoding RNAs. Cell Res. 23:759-774.

Ross E, Blair D, Guerrero-Hernández C, Sánchez Alvarado A. 2016. Comparative and transcriptome analyses uncover key aspects of coding- and long noncoding RNAs in flatworm mitochondrial genomes. G3 (Bethesda). 6:1191-1200.

Ruwe H, Castandet B, Schmitz-Linneweber C, Stern DB. 2013. Arabidopsis chloroplast quantitative editotype. FEBS Lett. 587:1429-1433.

Sanitá Lima M, Smith DR. 2017. Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists. G3. (G3/2017/045096).

Sanitá Lima M, Woods LC, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Resour. 16:1279-1286.

Schlüter JP, et al. 2010. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium Sinorhizobium meliloti. BMC Genomics. 11:245.

Shao R, Zhu XQ, Barker SC, Herd K. 2012. Evolution of extensively fragmented mitochondrial genomes in the lice of humans. Genome Biol Evol 4:1088-1101.

Shi C, et al. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. Sci Rep. 6:30135.

Shoguchi E, Shinzato C, Hisata K, Satoh N, Mungpakdee S. 2015. The large mitochondrial genome of Symbiodinium minutum reveals conserved noncoding sequences between dinoflagellates and apicomplexans. Genome Biol Evol. 7:2237-2244.

Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10:e1001241.

Smalheiser NR, Lugli G, Thimmapuram J, Cook EH, Larson J. 2011. Mitochondrial small RNAs that are up-regulated in hippocampus during olfactory discrimination training mice. Mitochondrion. 11:994-995.

Smalheiser NR. 2014. The RNA-centred view of the synapse: non-coding RNAs and synaptic plasticity. Philos Trans R Soc Lond B Biol Sci. 369:20130504.

Small ID, Rackham O, Filipovska A. 2013. Organelle transcriptomes: products of a deconstructed genome. Curr Opin Microbiol. 16:652-658.

Smith DR. 2011. Extending the limited transfer window hypothesis to inter-organelle DNA migration. Genome Biol Evol. 3:743-748.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. Brief Funct Genomics. 12:454-456.

Smith DR, Crosby K, Lee RW. 2011. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biol Evol. 3:365-71.

Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genomes architecture: reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci USA. 112:10177-10184.

Smith DR, Keeling PJ. 2016. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. Annu Rev Microbiol. 70:161-178.

Smith DR, Sanitá Lima M. 2017. Unraveling chloroplast transcriptomes with ChloroSeq, an organelle RNA-seq bioinformatics pipeline. Brief Bioinform. bbw088.

Steiner JM, Löffelhardt W. 2002. Protein import into cyanelles. Trends Plant Sci. 7:72-77.

Stevens PF. Angiosperm phylogeny website. http://www.mobot.org/MOBOT/Research/APweb/. (2001).

Struhl, K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol. 14:103-105.

Tian Y, Smith DR. 2016. Recovering complete mitochondrial genome sequences from RNA-seq: a case study of Polytomella non-photosynthetic green algae. Mol Phylogenet Evol. 98:57-62.

Ulitsky, I. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 17:601-614.

Vendramin R, Marine JC, Leucci E. 2017. Non-coding RNAs: the dark side of nuclear-mitochondrial communication. EMBO J. 36:1123-1133.

Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. 2011. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 39:979-988.

Voigt K, et al. 2014. Comparative transcriptomics of two environmentally relevant cyanobacteria reveals unexpected transcriptome diversity. ISME J. 8:2056-2068.

Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol. 12:647-653.

Wojciechowski MF. Millettioid sensu lato clade. http://tolweb.org/Millettioid_sensu_lato_clade/60341/2006.06.14. (2006).

Zhelyazkova P, et al. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell. 24:123-136.

Chapter 5

5. Organelles, revolutionary model systems

5.1 Concluding remarks

From endosymbiosis to land plant terrestrialization

Organelles have been intriguing scientists at least since the mid 19th century, when Swiss and German botanists found that plastids themselves go through division (Martin and Kowallik 1999). Since then, organelles have proved to be real revolutionary model systems. From the first account of the endosymbiotic origin of plastids, given by the Russian botanist Mereschkowski (Mereschkowski 1905), passing through Lynn Margulis' seminal paper "On the origin of mitosing cells" (Sagan 1967), organelles still provide scientists with mysteries that change the way we understand Biology. Although the endosymbiotic origin of organelles is textbook knowledge today (Martin 2017), the incommensurable diversity of organelle genome size, structure and content is still a puzzle (Smith and Keeling 2015). Not to mention the debate between mitochondrion-early and mitochondrion-late models of the origin of eukaryotes (Martin et al. 2017) and the discussions around the impact of endosymbiosis on evolution (Lane and Martin, 2010; Booth and Doolittle 2015; Lane and Martin 2015). In the attempt to understand those mechanisms, researchers have used organelles to forge and test new hypotheses on evolution and molecular biology (Lynch et al. 2006; Lynch 2007; Gray et al. 2010).

Organelle genomics started 36 years ago, when the human and mouse mitochondrial genomes were fully sequenced (Anderson et al. 1981; Bibb et al. 1981). By that time, a lot had happend to the field of molecular biology – the central dogma of molecular biology had been proposed (Crick 1958), tRNA, rRNA and mRNA were already described (Brenner et al. 1961; Gros et al. 1961; Scherrer and Darnell 1962; Scherrer et al. 1963; Holley et al. 1965) and the class of noncoding RNAs started to expand (Busch et al. 1982). Organelle DNA replication and transcription was also already documented

(Berk and Clayton 1974; Battey and Clayton 1978; Schwarz and Kössel 1980; Kearsey and Craig 1981; Ojala et al. 1981), but all this knowledge was scattered around several labs worldwide and based mostly on gene level experiments (Eddy 2001; Scherrer 2003; Cobb 2015).

36 years later, ncRNAs fully meet organelle genomes. Since the 80s, not only organelle genome diversity has been fairly documented (Smith and Keeling 2015), but also ncRNAs have taken over the field of molecular biology. Although the numerous types of ncRNAs have been gradually characterized through the three last decades, we came to realize how widespread they are only after the advent of next generation sequencing (NGS) techniques (Cech and Steitz 2014). Pervasive transcription across entire bacterial and nuclear genomes is now uncontested (Amaral et al. 2008; Wade and Grainger 2014), as most of the RNA-seq studies were devoted to study whole cell transcription (Smith 2013), be it prokaryotic or eukaryotic.

Conversely, the study of pervasive transcription in organelle genomes is still incipient and pervaded by uncertainties about the occurrence and significance of this transcriptional phenomenon (Dietrich et al. 2015; Vendramin et al. 2017). The few studies reporting pervasive transcription in organelles mostly employed NGS and provided different lines of evidence for full transcription of organelle genomes; they characterized multiple transcriptional start sites (Zhelyazkova et al. 2012), novel small RNAs (Mercer et al. 2011) and the transcription of entire plastid genomes in some land plants (Shi et al. 2016), for instance. But, how widespread the full (and consequently pervasive) transcription of organelle genomes was unknown, until now. Here, I demonstrated that organelle genomes are fully transcribed independent of their size, structure, content and taxonomic origin. My analyses, despite not identifying candidate ncRNAs, show high levels of transcription for both coding and noncoding organelle DNA and therefore, point to the existence of numerous ncRNAs pervasively transcribed. The functions of those ncRNAS, from the regulation of organelle genome transcription and translation (Dietrich et al. 2015) to the communication between organelle and nucleus (Vendramin et al. 2017), are just now being unraveled, but hold big promises. Under the light of organelle genome size variation, I pointed to the fact that those

organellar ncRNAs might have played a role in the terrestrialization of land plants and consequent evolution of plastid biogenesis. Because organelles themselves sense environmental stimuli (Woodson and Chory 2008), I argued that ncRNAs also might regulate trophic mode determination in mixotrophs, organisms of which are capable of switching between autotroph and heterotroph (Worden et al. 2015).

Initially, my collegues and I found that organelle genomes are being sequenced at unprecedented rates, but are not being further explored (Sanitá Lima et al. 2016). Knowing that NGS techniques not only helped to increase the number of organelle genomes sequenced, but also inundated public databases with genomic and transcriptomic data (Smith 2013), I sought to fill this gap. Then, as I determined the widespread occurence of genome-wide pervasive transcription in organelles, I demonstrated that publicly available RNA-seq data coming from whole cell experiments represent an untapped datasource to organelle genomics. Further exploration on the nature of organellar ncRNAs should not only unravel their regulatory functions, but also give insights onto their impact on evolution on Earth (Ulitsky 2016).

5.2 References

Aloni Y, Attardi G. 1971. Expression of the mitochondrial genome in HeLa cells II. Evidence for complete transcription of mitochondrial DNA. J Mol Biol. 55:251-270.

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. Science. 319:1787-1789.

Anderson S, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457-465.

Battey J, Clayton DA. 1978. The transcription map of mouse mitochondrial DNA. Cell. 14:143-156.

Berk AJ, Clayton DA. 1974. Mechanism of mitochondrial DNA replication in mouse L-cells: asynchronous replication of strands, segregation of circular daughter molecules, aspects of topology and turnover of an initiation sequence. J Mol Biol. 86:801-824.

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell. 26:167-180.

Booth A, Doolittle WF. 2015. Eukaryogenesis, how special really? Proc Natl Acad Sci USA. 112:10278-10285.

Brenner S, Jacob F, Meselson M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature. 190:576-581.

Busch H, Reddy R, Rothblum L, Choi YC. 1982. SnRNAs, SnRNPs, and RNA processing. Annu Rev Biochem. 51:617-654.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution – trashing old rules to forge new ones. Cell. 157:77-94.

Cobb M. 2015. Who discovered messenger RNA? Curr Biol. 25:R526-R532.

Crick FH. 1958. On protein synthesis. Symp Soc Exp Biol. 12:138-163.

Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F. 2015. Organellar non-coding RNAs: emerging regulation mechanisms. Biochimie. 117:48-62.

Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2:919-929.

Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolitlle WF. 2010. Irremediable complexity? Science. 330:920-921.

Gros F, et al. 1961. Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*. Nature. 190:581-585.

Holley RW, et al. 1965. Structure of a ribonucleic acid. Science. 147:1462-1465.

Kearsey SE, Craig IW. 1981. Altered ribosomal RNA genes in mitochondria from mammalian cells with chloramphenicol resistance. Nature. 290:607-608.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature. 467:929-934.

Lane N, Martin WF. 2015. Eukaryotes really are special, and mitochondria are why. Proc Natl Acad Sic USA. 112:E4823.

Lynch M. 2007. The origins of genome architecture. Sinauer, Sunderland, MA.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science. 311:1727-1730.

Martin WF. 2017. Physiology, anaerobes, and the origin of mitosing cells 50 years on. J Theor Biol. http://dx.doi.org/10.1016/j.jtbi.2017.01.004.

Martin W, Kowallik KV. 1999. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. Eur J Phycol. 34:287-295.

Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. Microbiol Mol Biol Rev. 81:e00008-17.

Mercer TR, et al. 2011. The human mitochondrial transcriptome. Cell. 146:645–658.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593-604. English translation in Martin W, Kowallik KV. 1999. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. Eur J Phycol. 34:287-295.

Ojala D, Montoya J, Attardi G. 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature. 290:470-474.

Sagan L. 1967. On the origin of mitosing cells. J Theor Biol. 14:225-274.

Sanita Lima M, Woods LC, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Resour. 16:1279-1286.

Scherrer K. 2003. Historical review: The discovery of ‘giant’ RNA and RNA processing: 40 years of enigma. Trends Biochem Sci. 28:566-571.

Scherrer K, Darnell JE. 1962. Sedimentation characteristics of rapidly labelled RNA from HeLa cells. Biochem Biophys Res Commun. 7:486-490.

Scherrer K, Latham H, Darnell JE. 1963. Demonstration of an unstable RNA and of a precursor to ribosomal RNA in HeLa cells. Proc Natl Acad Sci USA. 49:240-248.

Schwarz Zs, Kössel H. 1980. The primary structure of 16S rDNA from *Zea mays* chloroplast is homologous to E. coli 16S rRNA. Nature. 283:739-743.

Shi S, et al. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. Sci Rep. 6:30135.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. Brief Funct Genomics. 12:454-6.

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.

Ulitsky, I. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 17:601-614.

Vendramin R, Marine JC, Leucci E. 2017. Non-coding RNAs: the dark side of nuclear-mitochondrial communication. EMBO J. 36:1123-1133.

Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol. 12:647-653.

Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. Nat Rev Genet. 9:383–395.

Worden AZ, et al. 2015. Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 347:1257594.

Zhelyazkova P, et al. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell. 24:123-136.

Appendices

Appendix A: Unraveling chloroplast transcriptomes with ChloroSeq, an organelle RNA-Seq bioinformatics pipeline.

Published as: Smith DR, Sanitá Lima M. 2016. Unraveling chloroplast transcriptomes with ChloroSeq, an organelle RNA-seq bioinformatics pipeline. Brief Bioinform. bbw088.

Abstract

Online sequence repositories are teeming with RNA-Seq data from a wide range of eukaryotes. Although most of these datasets contain large numbers of organelle-derived reads, researchers tend to ignore these data, focusing instead on the nuclear-derived transcripts. Consequently, GenBank contains massive amounts of organelle RNA-Seq data that are just waiting to be downloaded and analyzed. Recently, a team of scientists designed an open-source bioinformatics program called ChloroSeq, which systemically analyzes an organelle transcriptome using RNA-Seq. The ChloroSeq pipeline uses RNA-Seq alignment data to deliver detailed analyses of organelle transcriptomes, which can be fed into statistical software for further analysis and for generating graphical representations of the data. In addition to providing data on expression levels via coverage statistics, ChloroSeq can examine splicing efficiency and RNA editing profiles. Ultimately, ChloroSeq provides a well-needed avenue for researchers of all stripes to start exploring organelle transcription and could be a key step towards a more thorough understanding of organelle gene expression.

Introduction

Massively parallel high-throughput sequencing of cDNA (RNA-Seq) has become a preeminent technique in plant research, and life science investigations as a whole (Wang et al. 2009). Consequently, open-access sequence repositories, such as GenBank, are expanding with RNA-Seq data from diverse land plants and algae (Fig. 1). As of 17 June 2016, GenBank's Sequence Read Archive (SRA) (Kodama et al. 2012) contained over

39,000 RNA-Seq datasets from streptophytes, and the Marine Microbial Eukaryotic Transcriptome Sequencing Project (Keeling et al. 2014) recently sequenced and made publically available the transcriptomes from hundreds of plastid-bearing protists.

RNA-Seq datasets from land plants and algae are obviously a great resource for investigating nuclear gene expression (Wang et al. 2009), but they are also an excellent but untapped means for exploring plastid and mitochondrial transcription (Smith 2013). Given that organelle genomes are present in many copies per cell and are highly expressed, organelle transcripts can represent a significant proportion of plant cellular RNA (Loening and Ingle 1967). Thus, eukaryotic RNA-Seq libraries typically contain large numbers (1–30%) of organelle-derived transcripts (Raz et al. 2011; Castandet et al. 2016), so much so that nearly complete organelle genome sequences can sometimes be assembled from RNA-Seq data alone (Shi et al. 2016; Tian and Smith 2016).

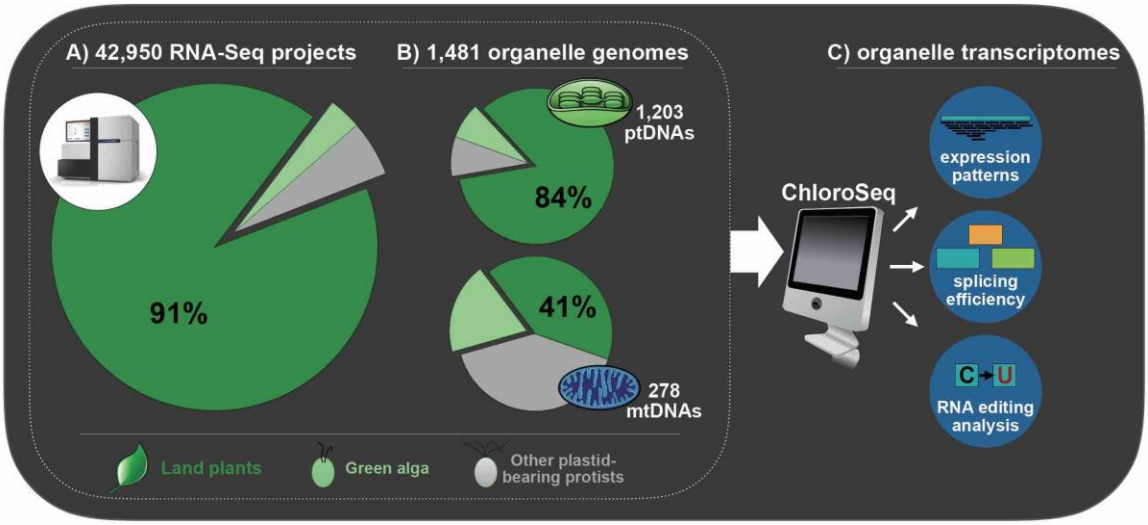

Figure 1A Available data in GenBank for exploring organelle transcription in plastid-bearing eukaryotes. A) As of June 17, 2016, GenBank's Sequence Read Archive (SRA) [http://www.ncbi.nlm.nih.gov/sra] contained 42,950 publically available RNA-Seq datasets from plastid-bearing species, 91% of which came from land plants. B) Similarly, the most recent Refseq release of mitochondrial and plastid organelle genome sequences (accessed June 17, 2016) [http://www.ncbi.nlm.nih.gov/genome/organelle/] included 1,481 organelle genomes from land plants and algae, 1,203 and 278 of which were plastid DNAs (ptDNAs) and mitochondrial DNAs (mtDNAs), respectively. This is an underestimate of the total number of available organelle genome sequences in GenBank because the Refseq database often does not include genomes from different strains of the same species or nearly complete organelle DNAs. C) These freely accessible RNA-Seq and organelle genome data can be used with the bioinformatics program ChloroSeq (Castandet et al. 2016) to systematically analyze organelle transcriptomes.

Unfortunately, researchers carrying out RNA-Seq on eukaryotes often ignore the organelle data, focusing instead on nuclear-derived transcripts (Smith 2013). In other words, GenBank contains a treasure trove of organelle RNA-Seq data that are just waiting to be examined (Figure 1A). But there has not been a sophisticated bioinformatics pipeline designed for analyzing organelle reads from eukaryotic RNA-Seq studies. That is, until now.

ChloroSeq: an Organelle RNA-Seq Bioinformatics Pipeline

Recently, a team of scientists from the Boyce Thompson Institute at Cornell University designed a new bioinformatics program called ChloroSeq, which systemically analyzes a plastid transcriptome using RNA-Seq (Castandet et al. 2016). ChloroSeq is open-source and freely available from GitHub (https://github.com/BenoitCastandet/chloroseq). The program operates through command-line-driven Perl scripts, which can be easily implemented on most laptop computers, provided the user has some experience with Unix.

Once installed, ChloroSeq uses RNA-Seq alignment data (i.e., a BAM file) to deliver a detailed analysis of the plastid transcriptome. The program first indexes and then extracts the plastid reads from the alignment BAM file, and uses these data for executing a variety of downstream analyses. The final output of ChloroSeq is in the form of text files (count tables), and it is important to emphasize that the program itself does not perform any statistical analyses on the transcriptional data; however, the count tables can be easily fed to other statistical software, such as R, for further investigations and for generating graphical representations of the data. Although most people associate transcriptomics with studies on differential gene expression, organelle genomes can undergo an assortment of other types of transcriptional modifications (Moreira et al. 2012; Smith and Keeling 2016). Accordingly, in addition to providing data on expression levels via coverage statistics, ChloroSeq can examine splicing efficiency and RNA editing profiles.

To help carry out these different analyses, the ChloroSeq pipeline relies upon other free, open-source bioinformatics programs, including the popular genomic software suites SAMtools (Li et al. 2009) and BEDtools (Quinlan 2014), which need to be installed on the host computer for the complete ChloroSeq workflow to run properly. And, again, users must provide an alignment BAM file, which can be generated using most read mapping software, such as Bowtie2 and TopHat2 (Kim et al. 2013).

Not surprisingly, much of the RNA-Seq data within the SRA come from paired-end libraries that were enriched for polyadenylated transcripts and/or were depleted of rRNAs. These types of datasets can be used with ChloroSeq, but the software has been optimized for single-end, strand-specific RNA-Seq. Moreover, the creators of ChloroSeq advise against using data from poly(A)-enriched libraries. This is because plant organelle transcripts become unstable upon polyadenylation (Rorbach et al. 2014) and are grossly underrepresented in these kinds of libraries. By comparing available RNA-Seq data from *Arabidopsis thaliana*, Castandet *et al.* (2016) showed that around 1% of the reads from oligo(dT)-selected libraries mapped to the plastid genome, whereas when generated from poly(A)-depleted total RNA followed by rRNA subtraction an astounding 30% of the reads came from the plastid. Nevertheless, if only 1% of RNA-Seq data are plastid-derived that still provides thousands and thousands of organelle reads for analysis, and means that researchers should be open to using ChloroSeq to explore any eukaryotic RNA-Seq dataset for organelle reads, no matter the protocol used to generate the library.

If you do decide to use poly(A)-enriched RNA for organelle studies it is important to keep in mind that different types of organelle transcripts could be differentially represented in the data. Unlike the near-ubiquity of polyadenylation of nuclear mRNAs, organelle transcripts are not necessarily polyadenylated (Small et al. 2013; Rorbach et al. 2014), and even when polyadenylation does occur, the transcripts for the various genes are often not polyadenylated at the same frequency. Moreover, polyadenylation is often a degradation signal in organelles (Hayes et al. 1999), meaning that researchers using poly(A)-selected RNA-Seq for measuring differential expression in organelle systems may, in some instances, be measuring the opposite: differential degradation.

Putting it to the test

To demonstrate the utility of ChloroSeq, Castandet *et al.* (2016) applied the software to various *A. thaliana* RNA-Seq projects from the SRA for which the plastid transcript data had not been mined or studied. By comparing RNA-Seq information from plants grown under control and abiotic stress conditions, the authors showed that heat stress can result in a global reduction in plastid RNA splicing and editing efficiency as well as an increase in plastid transcript abundance, including transcripts from coding, noncoding, and antisense regions of the genome. For instance, the authors used ChloroSeq to measure the ratio of spliced to un-spliced plastid RNAs and found that 12 hours of heat stress greatly inhibited the splicing efficiency of nearly all the plastid-encoded introns from *A. thaliana*, suggesting that organelle intron structure might be sensitive to temperature in a functionally significant manner (Castandet et al. 2016).

By searching other available data in the SRA, one can easily identify a variety of interesting experiments to run with ChloroSeq. Members of the land plant genus *Selaginella*, for example, are known to undergo extremely high levels of organelle RNA editing (Hecht et al. 2011; Oldenkott et al. 2014). Indeed, transcriptome sequencing of *Selaginella uncinata* uncovered 3,415 C-to-U RNA-editing sites in the plastid genome, which is one of the highest levels of post-transcriptional editing ever observed for a ptDNA. But detailed plastid RNA analyses have not yet been performed on any other members of the genus, even though the data needed to do so are available in GenBank. For *Selaginella moellendorffii* there exists a complete plastid genome sequence (accession NC_013086) and more than 15 different RNA-Seq datasets (e.g., SRA accessions SRX828740–5). Similarly, data from at least 4 RNA-Seq projects are available for *Selaginella kraussiana* (SRA accessions SRX1043962–5), and although the plastid genome of this species remains to be sequenced, one could easily generate a complete ptDNA from freely available whole genome shotgun sequencing data for *S. kraussiana* (SRA accession SRX1036537). Together, these datasets could be used in conjunction with ChloroSeq to generate complete RNA-editing profiles for the ptDNAs of *S. moellendorffii* and *S. kraussiana* and provide insights into the evolution, conservation, and diversity of plastid RNA-editing in the *Selaginella* lineage.

If extreme RNA-editing doesn't impress you, then widespread and bizzare intron splicing might. Expression of the *Euglena gracilis* plastid genome is a veritable circus act, requiring the removal of ~160 introns, including 15 twintrons (introns within introns), which need to be subtracted sequentially for accurate splicing (Hallick et al. 1993). Despite its record-breaking number of introns, RNA processing and intron splicing in the *E. gracilis* plastid remains poorly understood and poorly characterized. However, given that there are 22 freely available RNA-Seq datasets for this alga (e.g., SRA accessions ERX1051903–4) as well as a complete ptDNA sequence (accession NC_001603) one could easily employ ChloroSeq to investigate the plastid transcriptional architecture of *E. gracilis*.

Although designed with plastid transcriptomics in mind, ChloroSeq can also be used for studying plant and algal mitochondrial transcription (Castandet et al. 2016)—or transcription from any organelle system for that matter (e.g., animal mitochondria). In fact, many of the same transcriptional modifications and peculiarities found in plastids can also occur in mitochondria, such as RNA editing (Smith et al. 2012) and trans-splicing (Smith and Keeling 2016). Thus, the key features of ChloroSeq are equally as applicable to mitochondrial studies as they are to those on chloroplasts. Because of this, the software could help stimulate more thorough and extensive investigations of organelle gene expression.

Like with plants and algae, there is a plethora of publically available RNA-Seq data from metazoans, which can be used for addressing interesting questions in organelle genetics. Medusozoans (jellyfish and hydras), for instance, can have linear or linear fragmented mitochondrial genomes (Kayal et al. 2012) with elaborate telomere structures and homogenized gene sequences (Smith et al. 2012). Although there exist dozens of completely sequenced mtDNAs and more than 200 RNA-Seq datasets for medusozoans, very few researchers have studied mitochondrial transcription in this lineage (Kayal et al. 2015). Using ChloroSeq to examine these mtDNA and RNA-Seq data (e.g., GenBank accessions JN593332 and SRX315373) could lead to an interesting synthesis.

Bringing Organelle Transcriptomics to the Forefront

Plastids and mitochondria harbour some of the most extreme and unconventional modes of gene expression identified from across the tree of life (Smith and Keeling 2016). As noted above, posttranscriptional editing is rampant within the organelles of many plants and some algae. For instance, eleven of twelve possible types of substitution RNA editing (A-to-C, A-to-G, A-to-U, etc.) have been identified in the plastids of dinoflagellate algae (Mungpakdee et al. 2014), and both the plastid and mitochondrial transcripts of vascular plants can undergo moderate to severe C-to-U and/or U-to-C editing (Knoop 2011). Similarly, various plastid-bearing protists employ non-standard genetic codes in their plastid and/or mitochondrion (Matsumoto et al. 2011), and the organelle genomes of plants and algae often contain an abundance of introns, which in certain cases are *trans*-spliced or have unusual arrangements (Glanz and Kück 2009). More recently, organelle non-coding RNAs have been shown to be possible regulators of gene expression, and certain cases might be integral components for nuclear gene regulation (Dietrich et al. 2015). And organelle gene expression is integral to various aspects of cell signaling and cell physiology in plants, algae, and eukaryotes as a whole, including animals (Woodson and Chory 2008).

Despite being so remarkable, organelle transcription remains a relatively poorly studied topic. In the past five years more than 2,500 organelle DNAs were sequenced, resulting in thousands of organelle genome papers (Sanitá Lima et al. 2016). But in the same time period only a few dozen high-quality organelle transcriptome analyses were published, most of which came from model species (Mercer et al. 2011; Zhelyazkova et al. 2012). Although the human mitochondrial genome was sequenced more than thirty-five years ago, it has only been in past half-decade that a detailed human mitochondrial transcriptome was published (Mercer et al. 2011). But with over 300,000 RNA-Seq datasets from diverse eukaryotes currently sitting in the SRA and with new software like ChloroSeq arriving, the time is ripe for investigating organelle transcriptomes, and if the research community takes advantage of these freely available assets (Figure 1A), we might soon uncover novel and critical facets of organelle gene expression.

One of the major limitations of ChloroSeq is that it requires the input of alignment data based on a reference organelle genome sequence upon which RNA-Seq reads have been mapped. This means that RNA-Seq data for which there do not exist a corresponding organelle genome sequence (or one from a very close relative) cannot be used with ChloroSeq. But with thousands of complete organelle DNAs available in GenBank, and hundreds more arriving each month, this should not be a hurdle for much longer. Moreover, there is always the strong possibility that researchers can reconstruct a near-complete organelle genome sequence from the RNA data itself and then use it as a ChloroSeq reference sequence (Shi et al. 2016; Tian and Smith 2016).

Although not mandatory, most of the key functions of ChloroSeq are dependent on the existence of a proper annotation file for the organelle genome of interest. One might assume that the organelle genome data in GenBank are completely and properly annotated, but there are a surprising number of mtDNA and ptDNA sequences that are poorly and/or incorrectly annotated, and some lack annotations altogether (Smith 2012). Thus, it would be smart to verify the organelle annotation files prior to using them with ChloroSeq.

RNA-Seq and ChloroSeq might be great starting points for investigating transcription, but a complete picture of organelle gene expression will likely require a broad range of techniques and experiments, in addition to sequencing and bioinformatics. If past work has proven anything, it is that a deep understanding of organelle transcription can entail years of painstaking experiments, and can involve everything from advanced PCR, gel-electrophoresis, and blotting methods to high-throughput transcriptomics and proteomics. For example, it has taken more than twenty years of detailed RNA work to resolve the large and small subunit rRNA genes from the *Plasmodium falciparum* mitochondrial genome, which are fragmented and scrambled into ~25 distinct coding modules (Feagin et al. 2012). ChloroSeq is not a panacea for organelle transcriptional studies, but it is certainly a well-needed tool in an environment where there are too few bioinformatics programs devoted to organelle research.

The Growth of Bioinformatics Software for Organelle Research

ChloroSeq is among a handful of free bioinformatics software packages dedicated to studying plastid and mitochondrial genetics. Other popular programs include RNAweasel and MFannot (http://megasun.bch.umontreal.ca/RNAweasel/), which predict and model complex organelle RNAs and annotate introns and exons, as well as the webservers MITOFY (Alverson et al. 2010) and Organellar Genome Draw (Lohse et al. 2013), which respectively annotate and graphically map organelle genomes. The ORGanelle ASseMbler (ORGASM) (https://git.metabarcoding.org/org-asm/org-asm/wikis/home) is an open-source program designed to assemble complete organelle DNAs (and other small genomes) from whole genome shotgun sequencing data. Similar to ChloroSeq, the programs PREP-Mt (Mower 2005) and PREPACT 2.0 (Lenz and Knoop 2013) predict RNA editing sites in organelle genomes by searching against databases of known sequences, but unlike ChloroSeq they cannot make use of raw RNA-Seq data and next-generation sequencing read mappers.

Together, these and other software suites (Picardi et al. 2011) have helped streamline the study of organelle genomics, saving researchers time and energy. Yet, it is disappointing that there are not more bioinformatics programs specifically designed for analyzing organelle genomes. Organelle genetic data are used in a surprisingly wide variety of scientific disciplines, including medicine, forensics, genetic engineering, and archeology, to name but a few, and they have yielded countless fundamental insights into our understanding of the origins, evolution, and diversification of eukaryotic life, and continue to do so (Gray 2012; Keeling 2013).

As scientists, it is paramount that we employ the data that are available to us now and that will become available in the near and distant future. For researchers that study organelles, ChloroSeq will help make this possible. As more bioinformatics programs devoted to plastid and mitochondrial genetics arise, we could soon find ourselves in a position where many (even most) aspects of organelle genomic and transcriptomic analyses are automated—in fact, we have arguably nearly reached this point. Likewise, it will soon be

possible to outsource nearly all of the laboratory and bioinformatics work required to generate, assemble, annotate, and analyze an organelle genome. I recently received an email from a company called Phyzen (http://www.phyzen.com), advertising complete plastid genome assemblies, including annotations and GenBank submission files, for a few thousand US dollars. With ChloroSeq now freely available, I am betting that they will soon add plastid transcriptome analyses to their list of services.

Key points

- High-throughput sequencing of cDNA (RNA-Seq) has become a preeminent technique in life science research and, consequently, open-access sequence repositories are expanding with RNA-Seq data from diverse eukaryotes.
- Eukaryotic RNA-Seq datasets typically contain large numbers of organelle-derived reads, but researchers tend to ignore these data, focusing instead on the nuclear-derived transcripts. Moreover, there is a paucity of bioinformatics software for analyzing organelle transcriptomes.
- Recently, researchers designed a freely available bioinformatics program called ChloroSeq, which systemically analyzes an organelle transcriptome using RNA-Seq.
- The ChloroSeq pipeline uses RNA-Seq alignment data to deliver detailed analyses of organelle transcriptomes, including splicing efficiencies and RNA editing profiles.
- Our understanding of organelle transcription is surprisingly limited, despite the fact that mitochondria and chloroplast harbor some of the most unusual modes of gene expression ever identified. ChloroSeq provides a well-needed avenue for researchers of all stripes to start exploring organelle transcription.

References

Alverson AJ, et al. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol. 27:1436–1448.

Castandet B, Hotto AM, Strickler SR, Stern DB. 2016. ChloroSeq, an optimized chloroplast RNA-Seq bioinformatic pipeline, reveals remodeling of the organellar transcriptome under heat stress. G3. 6:2817-2827.

Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F. 2015. Organellar non-coding RNAs: emerging regulation mechanisms. Biochimie. 117:48-62.

Feagin JE, et al. 2012. The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. PLoS One. 7:e38320.

Glanz S, Kück U. 2009. Trans-splicing of organelle introns—a detour to continuous RNAs. Bioessays. 31:921–934.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harb Perspect Biol. 4:p.a011403.

Hallick RB, et al. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res. 21:3537–3544.

Hayes R, Kudla J, Gruissem W. 1999. Degrading chloroplast mRNA: the role of polyadenylation. Trends Biochem Sci. 24:199–202.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biol Evol. 3:344–358.

Kayal E, et al. 2015. Phylogenetic analysis of higher-level relationships within Hydroidolina (Cnidaria: Hydrozoa) using mitochondrial genome data and insight into their mitochondrial transcription. PeerJ. 3:e1403.

Kayal E, et al. 2012. Evolution of linear mitochondrial genomes in medusozoan cnidarians. Genome Biol Evol. 4:1–12.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Ann Rev Plant Biol. 64:583-607.

Keeling PJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12:p.e1001889.

Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. Cell Mol Life Sci. 68:567–586.

Kodama Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40:D54–D56.

Lenz H, Knoop V. 2013. PREPACT 2.0: predicting C-to-U and U-to-C RNA editing in organelle genome sequences with multiple references and curated RNA editing annotation. Bioinfor Biol Insights. 7:1.

Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079.

Loening UE, Ingle J. 1967. Diversity of RNA components in green plant tissues. Nature. 215:363–367.

Lohse M, Drechsel O, Kahlau S, Bock R. 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 41:W575-W581.

Matsumoto T, Ishikawa SA, Hashimoto T, Inagaki Y. 2011. A deviant genetic code in the green alga-derived plastid in the dinoflagellate *Lepidodinium chlorophorum*. Mol Phylogenet Evol. 60:68–72.

Mercer TR, et al. 2011. The human mitochondrial transcriptome. Cell. 146:645–658.

Moreira S, Breton S, Burger G. 2012. Unscrambling genetic information at the RNA level. Wiley Interdiscip Rev RNA. 3:213–228.

Mower JP. 2005. PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics. 6:1.

Mungpakdee S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genome Biol Evol. 6:1408–1422.

Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. RNA. 20:1499–1506.

Picardi E, Regina TM, Verbitskiy D, Brennicke A, Quagliariello C. 2011. REDIdb: an upgraded bioinformatics resource for organellar RNA editing sites. Mitochondrion. 11:360-365.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 47:11.12.1-11.12.34.

Raz T, et al. 2011. Protocol dependence of sequencing-based gene expression measurements. PLoS One. 6:e19287.

Rorbach J, Bobrowicz A, Pearce S, Minczuk M. 2014. Polyadenylation in bacteria and organelles. Methods Mol Biol. 1125:211–227.

Sanitá Lima M, Woods LC, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. Mol Ecol Res. 16:1279-1286.

Shi S, et al. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. Sci Rep. 6:30135.

Small ID, Rackham O, Filipovska A. 2013. Organelle transcriptomes: products of a deconstructed genome. Curr Opin Microbiol. 16:652–658.

Smith DR, et al. 2012. First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. Genome Biol Evol. 4:52–58.

Smith DR, Keeling PJ. 2016. Protists and the wild, wild west of gene expression: new frontiers, lawlessness, and misfits. Ann Rev Microbiol. 70:161–178.

Smith DR. 2012. Making your GenBank entry count. Front Genet. 3:123.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. Brief Funct Genomics. 12:454-456.

Tian Y, Smith DR. 2016. Recovering complete mitochondrial genome sequences from RNA-Seq: A case study of *Polytomella* non-photosynthetic green algae. Mol Phylogenet Evol. 98:57–62.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57–63.

Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. Nat Rev Genet. 9:383–395.

Zhelyazkova P, et al. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell. 24:123-136.

Curriculum Vitae

Name: Matheus Sanitá Lima

Post-secondary Education and Degrees:
University of São Paulo
Ribeirão Preto, São Paulo, Brazil
2008-2013 BSc Hons

Western University
London, Ontario, Canada
2015-present MSc

Honours and Awards:
Honerkamp-Smith travel award (£380)
2015

Department of Biology (Western University) travel award (CAN$200)
2017

PSAC 610 Outstanding Research Contributions Scholarship (CAN$400)
2017

Related Work Experience
Teaching Assistant
Western University
2015-2017

Publications:

Submitted: Sanitá Lima M, Smith DR. 2017. Pervasive transcription of mitochondria, chloroplasts, cyanelle and nucleomorphs across plastid bearing protists. Genome Biol Evol. (GBE-170722).

Submitted: Sanitá Lima M, Smith DR. 2017. Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists. G3. (G3/2017/045096).

Smith DR, Sanitá Lima M. 2017. Unravelling chloroplast transcriptomes with ChloroSeq, an organelle RNA-Seq bioinformatics pipeline. Brief Bioinform. bbw088.

Sanitá Lima MS, Woods LC, Cartwright MW, Smith DR. 2016. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid genomes. Mol Ecol Resour. 16:1279-86.

Sanitá Lima M, et al. 2016. Co-cultivation of *Aspergillus nidulans* recombinant strains produces an enzymatic cocktail as alternative to alkaline sugarcane bagasse pretreatement. Front Microbiol. 7:583.

da Silva TM, Pessela BC, da Silva JCR, Sanitá Lima M, Jorge JA, Polizeli MLTM. 2014. Immobilization and high stability of an extracellular β-glucosidase from *Aspergillus japonicus* by ionic interactions. J Mol Catal B: Enzym. 104:95-100.

Benassi VM, da Silva TM, Pessela BC, Guisan JM, Mateo C, Sanitá Lima M, Jorge JA, Polizeli MLTM. 2012. Immobilization and biochemical properties of β-xylosidase activated by glucose/xylose from *Aspergillus niger* USP-67 with transxylozylation activity. J Mol Catal B: Enzym. 89:93-101.