biomolecules

MDPI

Review

Small Genomes and Big Data: Adaptation of Plastid Genomics to the High-Throughput Era

Christen M. Klinger [1] and Elisabeth Richardson [2,*]

[1] Division of Infectious Diseases, Department of Medicine, University of Alberta, Edmonton, AB T6G 2R3, Canada
[2] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2R3, Canada
* Correspondence: ehrichar@ualberta.ca

Received: 27 June 2019; Accepted: 16 July 2019; Published: 24 July 2019

Abstract: Plastid genome sequences are becoming more readily available with the increase in high-throughput sequencing, and whole-organelle genetic data is available for algae and plants from across the diversity of photosynthetic eukaryotes. This has provided incredible opportunities for studying species which may not be amenable to in vivo study or genetic manipulation or may not yet have been cultured. Research into plastid genomes has pushed the limits of what can be deduced from genomic information, and in particular genomic information obtained from public databases. In this Review, we discuss how research into plastid genomes has benefitted enormously from the explosion of publicly available genome sequence. We describe two case studies in how using publicly available gene data has supported previously held hypotheses about plastid traits from lineage-restricted experiments across algal and plant diversity. We propose how this approach could be used across disciplines for inferring functional and biological characteristics from genomic approaches, including integration of new computational and bioinformatic approaches such as machine learning. We argue that the techniques developed to gain the maximum possible insight from plastid genomes can be applied across the eukaryotic tree of life.

Keywords: plastid biology; bioinformatics; biotechnology; next-generation sequencing

1. Introduction

1.1. Plastid Genomes

The endosymbiotic theory, proposed by Lynn Margulis in 1966, has provided the theoretical underpinnings of over five decades of endosymbiotic organelle biology research [1]. Through the interpretations of mitochondria and plastids as the enslaved remnants of free-living bacteria, scientists gained critical insight into the membrane biology, protein complement and genetics of these enigmatic organelles. One of the key traits supporting the identification of endosymbioses was the presence of organellar genomes, similar in structure, replication and expression to their prokaryotic cousins [2]. Critical insights into plastids and plastid-like organelles have arisen from close study of the plastid genome, aided by advances in sequencing technology. For example, gene sequencing of the 23S ribosomal gene in the apicoplast of the eponymous Apicomplexa were crucial to its identification as a highly reduced plastid-like organelle [3].

Use of genes and genomics in biology has exploded in the past few decades. The phylogenetic species concept, where species were delimited based on % divergence in DNA sequence, introduced DNA as a unit of biological classification, and plant researchers increasingly turned to plastid genes to unravel the tangled web of algal and plant diversity [4]. The large subunit of the RuBisCO protein, essential for carbon fixation, is a commonly used marker gene for phylogenetic analysis of the

relationships between plants and algae [5]. Plastids are relatively easy to extract from plant tissues, and contain an abundance of DNA; this, along with their uniparental maternal inheritance, makes them an ideal target for taxonomic gene sequencing [6]. To date, over 1000 plastid genomes have been sequenced from across eukaryotic diversity, including representatives from all known plastid-bearing phyla [7]. The array of genetic complements and structures revealed by this sequencing effort is astounding; from the tiny, highly-fragmented plastid genomes of dinoflagellates to the bloated, largely non-coding plastid genomes of Chlamydomonadales, plastid genomes can also have circular, linear, or multi-chromosomal structures [8], and the wide spread of plastids across eukaryotic diversity is shown in Figure 1. However, it is worth noting at this stage that plastids, generally, exhibit much less diversity in their genetic rearrangement than the other major genome-containing organelle, the mitochondria. The reasons for this are not known, though there is speculation that it is due to the near-ubiquity of mitochondria, all of which are derived from a single source long before the endosymbiosis of the first plastid, allowing greater genetic divergence to emerge over evolutionary time, also shown in Figure 1 [8].

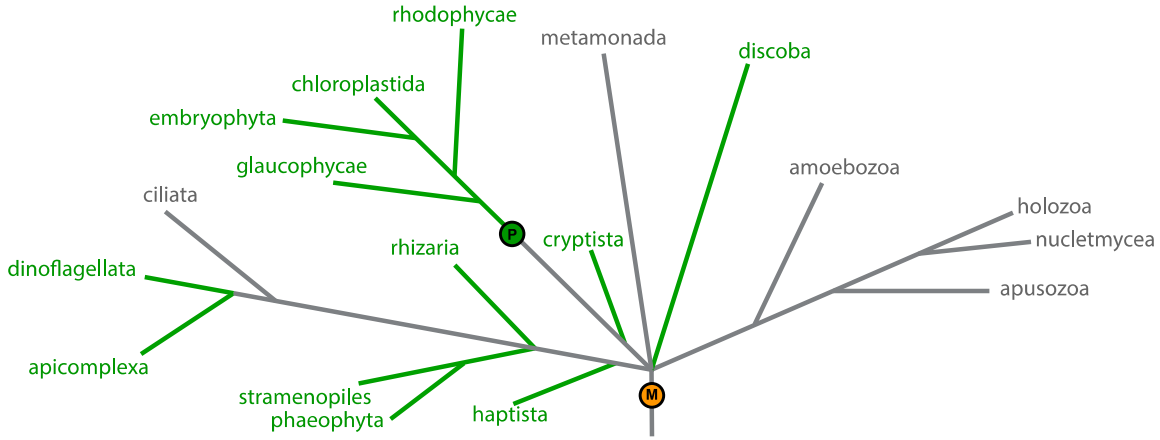

Figure 1. The diversity of eukaryotes, with lineages containing plastids identified in green. The origin of the cyanobacterial plastid present in the majority of photosynthetic organisms (excluding *Paulinella* spp., which has a plastid derived from a second primary endosymbiosis) and the origin of the alphaproteobacterial mitochondria are indicated by the green P and orange M, respectively.

1.2. Next-Generation Sequencing

The number of available plastid genomes has been massively increased by the advent of next-generation sequencing and the corresponding explosion in publicly accessible genetic information [7,9]. Experimental techniques using genetic sequencing, or a combination of methods including gene sequencing, have also become more and more prevalent [10]. Most journals have a requirement for genetic data produced during an experiment to be deposited in a curated sequence archive, and it is now common for candidates for new species to be identified not from microscopy or culture but from DNA sequence extracted from environmental samples. One of the most notable examples of this is the identification of the Asgardarchea, the closest extant archeal prokaryotic relatives of the eukaryotes [11,12]. Despite the extensive work that has been done to identify the genetic complement and architecture of Asgardarchea, none of them have yet been cultured or even photographed in detail [13]. It is also becoming more cost-effective to amplify an entire genome or transcriptome to identify a particular gene or pathway than to try to isolate the genetic components responsible for the phenotype [14]. With the flood of sequencing of DNA and RNA from environmental or RNA-seq experiments and subsequent publication of this data with minimal annotation, it is relatively easy to extract organelle genomes from these datasets [15]. Plastid genes can be extracted from RNA-seq datasets by searching for bacterial-like genes and plastid targeting sequences. It is therefore possible to assemble large, pan-taxon organelle genome datasets using just publicly accessible

data [15]. These techniques have already been applied successfully to identify new plastid genomes from the Marine Microbial Eukaryote Transcriptome Sequence Project (MMETSP), which have provided insights for multiple photosynthetic lineages and their plastid gene complement [16].

The research potential of these datasets should not be underestimated; there are still many aspects of plastid biology that are known only in individual lineages, or the mechanisms for which have only been elucidated in selected model organisms. It is now possible for researchers, with extremely low up-front cost, to determine how widespread these traits are, and to use this information to inform their future research. Mechanisms that have been suggested to explain traits in specific species can now be evaluated across a much broader range of data and allow researchers to draw general conclusions on overarching plastid functional hypotheses across the entire diversity of eukaryotes—or, conversely, determine which functions are lineage-specific adaptations. A selection of common methods of obtaining DNA sequencing data relevant to plastids, and how the data can be analysed to extract both plastid genomes and plastid-targeting genes, is illustrated in Figure 2. There are dozens of genome and transcriptome assemblers available for research use, all of which use slightly different algorithms and assembly parameters and will be optimal for different experimental approaches. The development of genome assembly algorithms has been reviewed by Simpson and Pop [17], and more recently by Sohn and Nam [18]. Metagenomic and metatranscriptomic studies require an additional step before genome assembly where the combined environmental reads are separated by sequence; similarly, many programmes are publicly available and optimised for various types of data, reviewed in Ladoukakis et al. [19]. Finally, genes and open reading frames (ORFs) can be predicted from high-throughput sequencing data by codon analysis, as regions with long stretches of coding amino acids between a START and STOP codon can be further analysed for similarity to known genes by using algorithms such as BLAST [20]

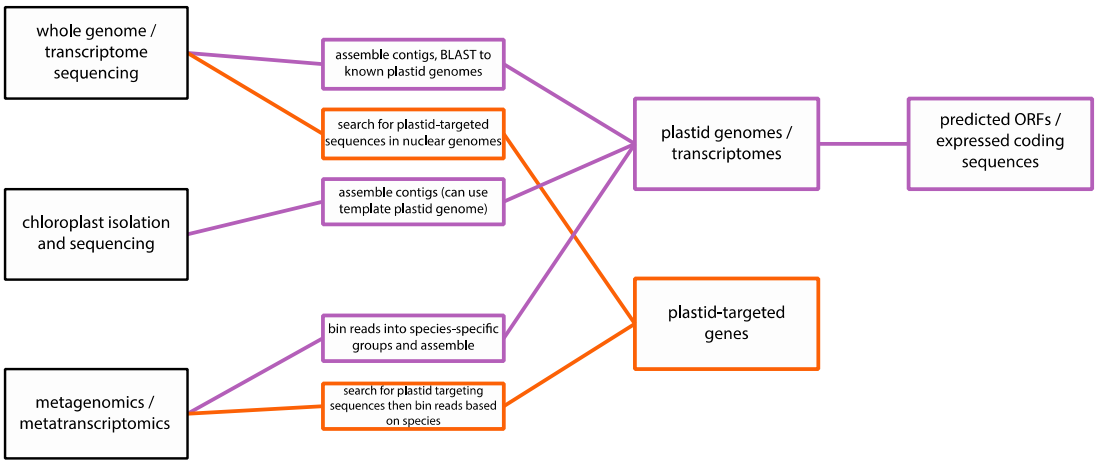

Figure 2. Methodology for identifying plastid genome sequences and plastid-targeted genes from various types of sequencing: genome/transcriptome sequencing projects, where the DNA is extracted from a cultured organism or single cell; chloroplast isolation and sequencing, where DNA is extracted from purified chloroplasts; and metagenomics/metatranscriptomics, where DNA from a microbial community is extracted and sequenced directly from an environmental sample.

Here, we present two case studies where researchers have evaluated lineage-specific plastid genetic traits over a large dataset assembled from publicly available data. This has allowed functional hypotheses based on specific plastid genomes to be evaluated over plastids generally within that lineage, or across photosynthetic diversity. We then use these examples to examine how bioinformatic tools created to analyse enormous genetic datasets have been invaluable for this effort. We also suggest other research techniques which could add to bioinformatic re-analysis of publicly available data in fields as diverse as evolutionary cell biology, ecology, and biotechnology.

2. Case Studies

2.1. Case Study 1: Codon Usage

Codon usage bias arises as a consequence of redundancy within the genetic code for almost all amino acids; those that can be encoded for by between two and six codons (two- and six-fold redundancy, respectively), are said to have synonymous codons. Under a random null model, all synonymous codons would be expected to be present in protein coding genes at equal frequency. Codon usage bias therefore refers to deviations from this null [21]. Implicit in the genetic code is the fact that third position substitutions are frequently synonymous, a fact which allows overall genome composition (GC-richness) to play a substantial role in shaping increased representation of NNA/NNT over NNC/NNG codons when the amino acid coded is synonymous [22,23]. This 'mutational hypothesis' is offset by the 'translational hypothesis' that posits that codon usage is optimized in highly expressed genes to improve either translation speed, translation efficiency, or both [21,24].

Support for the translational efficiency hypothesis comes from a multitude of observations. In bacteria, although overall patterns follow compositional biases, growth rate was found to correlate with codon usage of highly expressed genes [25]. Similarly, codon composition was found to differ in highly expressed genes in diverse plastids across multiple lineages [26,27]. However, the relevance of translational efficiency in endogenous genes has been questioned by the notion that transcription initiation, rather than elongation, is typically rate-limiting for protein synthesis [24]. With expression of transgenes, where the overexpression of heterologous mRNA often exceeds endogenous mRNAs by several-fold, codon optimization has been shown to increase expression by up to 1000-fold [28].

Regardless of the exact mechanisms governing codon bias, a recent study highlighted the extent to which the phenomenon exists across 103 publicly available plastomes by using a combination of previously described metrics and resampling simulations. Their results were variable, but despite the lack of overall strongly observed trends, codon bias was found to be highly lineage-dependent and to slightly favour highly expressed genes [23]. Other studies have shown that codon usage patterns can be a confounding factor in identifying trends if not considered properly [29]. Within the photosynthetic dinoflagellates, the ancestral plastid is of red algal origin and contains the pigment peridinin. However, this plastid has been serially replaced on multiple occasions, including through the putative engulfment of a haptophyte giving rise to a 19′ hexanoyloxyfucoxanthin-type plastid (commonly referred to simply as a fucoxanthin plastid). As such, it is expected that haptophyte and fucoxanthin plastid genes should place as sisters in phylogenetic analysis. However, DNA-based phylogenies of *psaA* and *psbA* appeared to suggest that haptophytes instead formed an outgroup to all dinoflagellate sequences, with the fucoxanthin plastid as ancestral, rather than secondarily acquired. Extensive re-analysis of datasets was conducted where Leu, Ser, and Arg codons, the only amino acids with codon variation at the first codon position, were either omitted or recoded. These results, alongside additional protein-based phylogenies, showed that the strong relationship between peridinin dinoflagellates and haptophytes was a result of codon bias. This suggests that the haptophyte + all dinoflagellate relationship in DNA phylogenies was artefactual and was due to shared codon composition bias between a subset of peridinin-containing dinoflagellates and haptophytes [29]. Hence, codon composition bias will be important to study moving forward, both as a means to better understand the evolutionary implications of plastid evolution, but also as a means for the expression of heterologous gene products in algae and beyond.

2.2. Case Study 2: RNA Editing

RNA editing is a common post-transcriptional modification frequently found in mitochondria of diverse eukaryotes, but occasionally in the nuclear and plastid genomes as well [30]. Some plastomes of land plants are heavily edited [31], but the majority of species show a comparatively lower editing rate, at or below 1% of positions within coding sequences. Organellar C-U interconversions are observed across land plants, but not in green algae. Although the complete machinery of the editing complexes

and the nature of the editing mechanism are still being elucidated, it appears that editing sites are specified by pentatricopeptide repeat (PPR) proteins [32]. This specificity of PPR proteins has allowed the construction of databases and the development of predictive tools for PPR binding sites.

Plastid RNA editing is also found in alveolates, in a case of what appears to be convergent evolution (shown in Figure 3A,B respectively). The alveolate group dinoflagellates possess often extensive plastid RNA editing, with overall rates up to ~5% [33]. Functional explanations for these editing events mostly evoke the removal of premature STOP codons or some effect on base and/or codon composition [34–38]. Studies focusing on individual lineages in isolation over a decade failed to find consistent patterns across dinoflagellates. In the case of *Symbiodinium minutum*, it was also suggested that changing the hydrophobicity of encoded proteins might be important. A systematic study, incorporating both novel and publicly available data and making use of computer simulations, found general trends such as clustering of editing events and propensity for editing to improve biochemical similarity of homologous amino acids to those from lineages that do not undergo editing, but failed to identify any putative signals directing editing events [33]. The mechanism of RNA editing in dinoflagellates remains to be uncovered, although the low-level nuclear RNA editing detected in *S. microadriaticum* was associated in some cases with significantly enriched motifs and genes encoding PPR proteins were detected in the genome [39]. This suggests an editing protein complement similar to that of land plants. RNA editing has also been identified in Apicomplexa, a sister group to dinoflagellates, which retain a remnant plastid, along with a single apicoplast-targeted PPR protein [40].

Mechanistically, RNA editing can be envisioned either as a directed process, in which an encoded factor (for example, protein and/or RNA) directs a specific nucleotide alteration at a specific position (call this a "editing-directing factor", EDF) based on the genomic context, which could be the adjacent sequence, higher-level structure, or some other defining feature. Alternatively, editing could be promiscuous, either by reducing the stringency with which cellular machinery recognizes editing sites or by having a large number of EDFs that are capable of duplicating, accumulating mutations and novel features, and, hence, altering their specificity.

RNA editing presents a challenge to adaptive evolutionary theory, given the presence of silent editing events (i.e., those that do not change the encoded amino acid), editing events that move away from biochemical consensus, and editing events in non-coding regions, which may or may not have any functional implications [29–33,37–41]. In land plants, it has been suggested that changes in the subunit composition, and hence subunit–subunit interactions, of the plastid NDH complex may have been eased by RNA editing, while in *S. microadriaticum* it appears that nuclear RNA editing may be responsive to stress conditions [39,42]. In dinoflagellates in general, plastid RNA editing has been suggested as a mechanism for correcting deleterious mutations induced by the high level of sequence evolution in the plastid genome [43].

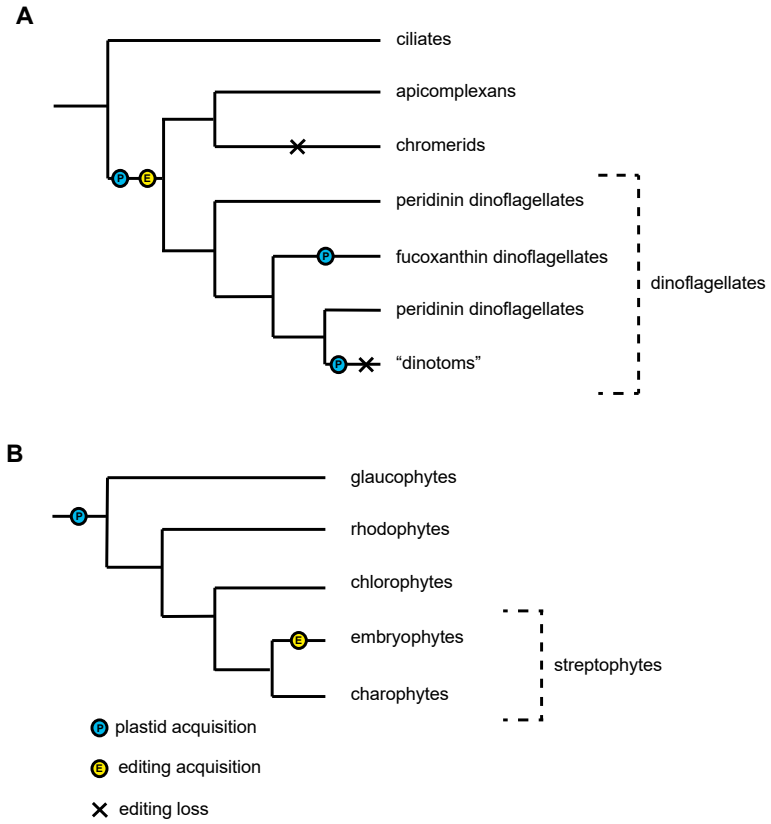

Figure 3. Convergent evolution of RNA editing in plastids. (**A**) RNA editing in the alveolates, including the Apicomplexa and dinoflagellates. There have been two additional plastid acquisitions in dinoflagellates, indicated by the blue Ps: the uptake of a haptophyte-derived plastid in the fucoxanthin dinoflagellates and the uptake of a diatom-derived plastid in the "dinotoms". The emergence of transcript editing at the base of the Apicomplexa and dinoflagellates is indicated with a yellow E. Dinotoms do not appear to exhibit transcript editing, while the trait has been retained in fucoxanthin dinoflagellates. This trait also does not appear to be present in chromerids, the closest apicomplexan relative with a retained photosynthetic plastid. These losses of transcript editing are indicated with black crosses. (**B**) RNA editing in the Archaeplastida, with the primary endosymbiosis of a cyanobacteria at the base of the clade indicated with a blue P. Restricted (U to C and C to U) RNA editing in this clade appears to be restricted to the land plants, and the emergence of this trait is indicated by a yellow E.

In general though, the origin of such adaptive hypotheses for RNA editing is difficult to conceive, at least in cases where a nucleotide change that could be subject to editing induces a lethal alteration, as this would require both general editing machinery and an EDF capable of reverting this change be present and expressed prior to the change actually occurring for the organism to survive. As discussed above, this could occur either through the relative promiscuity of each individual EDF, through the presence of a large complement of EDFs in an organism's genome, or a combination thereof. Three potential editing scenarios are indicated in Figure 4. Figure 4A shows a method of RNA editing which would result in specific edits with a relatively small number of base pair conversions, which corresponds to the editing observed in land plants. Figure 4B,C indicate two possible scenarios for more promiscuous editing, both in terms of editing site selection and potential base pair conversions. Promiscuous editing could be problematic, as the advent of an EDF may itself induce an undesirable change (i.e., from a nucleotide to one that causes a nonsynonymous change or encodes a premature STOP codon), but it is likely that such a change would not be driven to fixation and hence, on the whole, the system would retain only generally positive events. This is consistent with the pattern of editing found in dinoflagellates. In *Plasmodium*, editing appears to be stage-specific, which implies additional developmental functions [40].

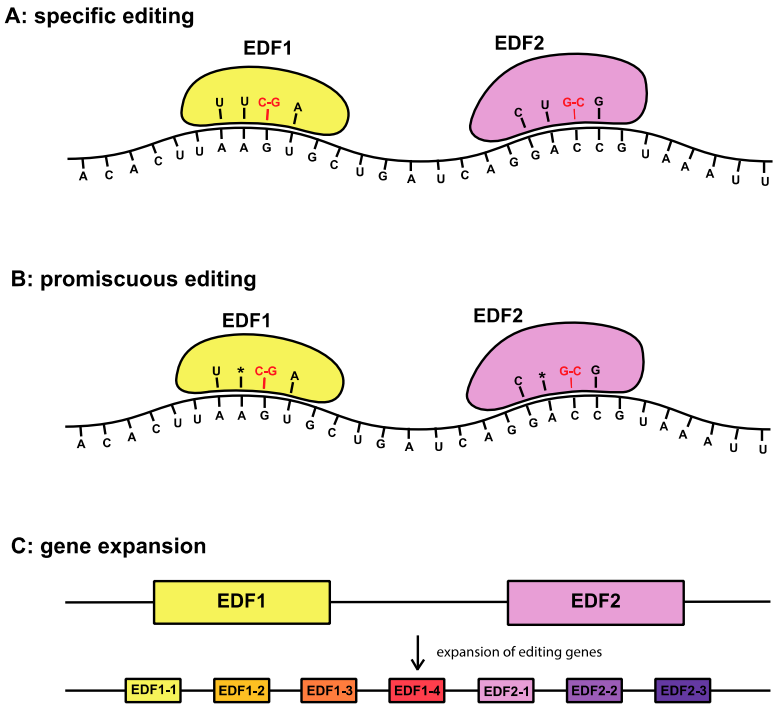

Figure 4. Possible models of transcript editing with editing factors (EDFs). (**A**) Specific editing, where specific EDFs are responsible for facilitating edits on matching regions of the transcript. (**B**) Promiscuous editing, where there is some flexibility in the possible matches that any one EDF can make, or what conversions are possible. (**C**) Gene expansion, where a small group of ancestral EDFs expand within the genome and evolve different specificities to transcript sequences or possible base conversions.

Regardless of the exact nature of the editing landscape in plastids, it is likely an irremediable genetic feature. Over time, the ability of editing to correct multiple deleterious nucleotides at the genomic level would lead to an increasingly low probability that all such bases could revert to the ancestral, or otherwise synonymous, base. Failing this, editing becomes an essential function for the organism, despite arguably greater complexity and the energetic requirements to maintain and express editing machinery. Additionally, the presence of editing machinery, especially as a promiscuous system, would allow for additional deleterious changes at the genomic level over time, especially if the evolutionary rate were high, as has been described in dinoflagellates [33]. This non-adaptive or neutral theory is consistent with the concept of an "evolutionary ratchet", which suggests that the dependence of a cellular system that has accumulated mutations (here, the expression of genes) on another system (here, RNA editing) means that reversal to independence is unlikely [44]. Continued study of RNA editing using larger datasets and more sophisticated analytical and statistical tools will likely improve our understanding of this fascinating molecular feature.

3. Future Possibilities

3.1. Evolutionary Biology

Some of the most interesting debates in evolutionary cell biology revolve around the evolutionary origins of plastids. The history of plastid transfer, gain, and loss is an unbelievably complex web of serial endosymbiosis that can often seem hopelessly tangled, particularly within the group historically identified as the "chromalveolates", or organisms containing red algal-derived plastids [45,46]. The exact relationships between these organisms and their chosen endosymbionts are an area of lively debate—one which we have no desire to re-open—but has resulted in extremely active interest in and development of sophisticated tools to detect the evolutionary origins of specific plastid genes [47]. Endosymbiosis, and the associate gene transfer, is a unique way of introducing genetic plasticity and

many hypotheses for the evolution of extant photosynthetic diversity which incorporate endosymbiosis as an evolutionary mechanism have not yet been tested across the diversity of eukaryotes, or deeply within that diversity. For example, the "limited transfer window" hypothesis for horizontal gene transfer between organelles within an organism predicts that the fewer plastids an organism has, the less likely transfer is to the nucleus or other organelles. This hypothesis was proposed in 2006 by Barbrook et al. [48] and was tested extensively across plastid diversity in 2011 by Smith et al. [49,50]. Since 2011, nuclear and plastid genome pairs have been acquired from a multitude of organisms, including those adapted to extreme environments such as the High Arctic and from additional eukaryotic groups such as the excavates and chlorarachniophytes [51–53]. Research on the genomes of the chlorarachniophyte *Bigelowiella natans* and the excavate *Euglena gracilis* both reported gene transfer between the organelles consistent with the limited transfer window hypothesis [52,53]. Expanding the techniques used in Smith et al. [49,50] into this wider range of organisms across eukaryotic diversity and incorporating bioinformatic models of different selective pressures on this trait could provide additional insight into this long-held hypothesis.

3.2. Ecological Research

Plastids are an essential aspect of photosynthesis, and therefore an essential part of global nutrient cycling [54]. Some of the clades identified as most important for carbon fixation in the open ocean, such as the dinoflagellates algae and diatoms, have rich evolutionary histories for both their nuclear and their plastid genomes [43,55]. Modern molecular ecology questions have increasingly focused on the functional consequences of identifying a particular taxon or genetic pathway in an environment. For example, if many parasitic lineages are identified in a metatranscriptomic study of soils, what does this say about the nutrient cycling within an environment [56]? If we are able to determine the microbial community present in an environment and have previously determined which functional traits are associated with the plastids of these organisms, then researchers can begin to make conclusions about the ecological consequences of those traits being active. We also briefly discussed in the introduction the potential identification of organelle sequences and genes from environmental DNA sequence runs; in most cases, organelle-derived DNA is discarded from eukaryotic microbial community assessments as the 18S rRNA gene is used as a proxy for diversity [57]. However, this genetic data may contain additional information that can be incorporated into diversity assessments. For example, the most commonly used primers for amplifying eukaryotic DNA from environmental DNA, the universal eukaryotic primers for the V4 region of the 18S rRNA developed in Stoeck et al. [58], have been demonstrated to amplify excavate DNA relatively poorly [59]. However, some excavates, including the group Euglenozoa, contain functional plastids [7]. Incorporating organelle DNA into environmental surveys may assist in identifying excavates whose presence may otherwise be overlooked. This has already been used successfully in a survey of the fungal and plant communities of a protected wetland in Wood Buffalo National Park, Alberta, Canada. Porter et al. [60] extracted nuclear ITS spacer region sequences and plastid *rbcL* sequences from metagenomic samples and used them to determine the diversity of local fungi and plants, respectively.

Plastid genomics, particularly from environmental DNA studies, can also be used to identify if a plastid genome itself is adapted to an environment. Genome-level evolutionary adaptation to environments has long been observed. For example, patterns of gene loss and genome reduction in parasitic lineages have long been observed across the diversity of eukaryotes [61]. One could expect to identify adaptations that may provide a selective advantage in the organelles of organisms extracted from extreme environments. We attempted, albeit unsuccessfully, to use this approach when examining the genomes of obligate psychrophile algae isolated from the High Arctic. Using the reasoning that the genomes of these organisms would be exposed to higher than usual amounts of UV radiation, we hypothesised that their plastid genomes would use fewer codons containing thymine repeats, which could result in the production of thymine dimers. An automated methodology for detecting thymine dimers was initially proposed to detect internal poly(T) gene sequences in Klinger et al. [33]. Due to

technical difficulties, we were unable to identify any definitive trends, but this illustrates the lines of research that can be tested through bioinformatic analysis of plastid genome data. Mining public data archives for plastid genomes may also result in the identification of previously unheard-of plastid traits; the discovery of the cercozoan *Paulinella* spp. and the sequencing of its plastid genome revealed that a secondary primary endosymbiosis had occurred in this lineage, where previous hypotheses suggested a single primary endosymbiotic origin of both the plastid, in the same way the mitochondria is accounted for by a single primary endosymbiosis [62].

3.3. Incorporation of New Computational Techniques and Biotechnology

With the amazing possibilities generated by bioinformatic analysis of plastid genomes across eukaryotic diversity in front of us, it is worth considering how this approach could develop in the light of rapidly advancing technology [10]. Machine learning and artificial intelligence are increasingly being incorporated into genomic analyses, with machine learning algorithms and tools being developed which are able to, for example, identify ORFs, enhancer and suppressor elements, and protein binding sites [63,64]. In these techniques, machine learning algorithms compare large datasets of different states: for example, gene/not gene, diseased/not diseased, or regulatory element/not regulatory element. The algorithm can identify features that are common to only one type of dataset or the other, and then use these to classify new examples. The major benefit of these techniques is that they are affected to a much lesser extent by the biases inherent in the identification of these sites by manually developed tools; machine learning algorithms are able to identify one condition versus the other often with more sophistication than trained experts [65]. One example from Case Study 2: RNA editing shows an obvious potential application of this technology: there are already algorithms based on machine learning that are able to identify putative editing sites in plant plastids [66,67]. If given enough examples of edited and non-edited sites in dinoflagellates, where the editing machinery likely evolved independently, would these algorithms be able to correctly classify edited versus non-edited transcripts? Machine learning could also be incorporated in examination of the "limited transfer window" hypothesis discussed in the previous paragraph: The studies in Smith et al. [49,50] provided a wealth of examples of horizontally-derived nuclear and mitochondrial sequences which could be used to train a classifier to identify transfers automatically across a much larger dataset, or identify more cryptic transfers. This is an exciting line of enquiry, particularly in organisms where molecular tools for genetic manipulation are absent or newly developed, such as dinoflagellates.

Synthetic biology is another area where plastid genomics using pan-eukaryotic datasets may be extremely effective. Processes associated with plastids, and in particular carbon fixation, have become even more critical for us to understand in the context of ongoing climate change [68]. There is already considerable effort being placed into developing synthetic organisms capable of mitigating some forms of anthropogenic change, such as degradation of oil spills [69]. By identifying trends in codon usage and distribution of metabolic pathways across plastids, researchers can identify which genes, codons, or genetic motifs are essential—or even just most efficient—in a synthetic organism. Bioremediation and biofuel production, both of which involve manipulation of hydrocarbon chains by organisms to produce a useful output or degrade an undesirable one, already make great use of photosynthetic organisms and plastid pathways [70]. For example, plastid genome editing in the marine algae *Nannochloropsis oceanica* can increase fatty acid biosynthesis and therefore biofuel production [71].

4. Conclusions

In the early twenty-first century, all scientific disciplines are in some respect governed by the incredible acceleration of technological development. In this review, we discuss how the co-advancement of two aspects of computational biology, bioinformatics and next-generation sequencing, have opened up an entirely new approach for studying plastid genomics and cell biology. We illustrate two cases in which mining publicly available databases to isolate plastid genome sequences has allowed researchers

to implement bioinformatic approaches to test functional hypotheses across large-scale, pan-lineage datasets. We have also provided some examples of how this approach may open doors for further research in several disciplines, including ecology, synthetic biology and evolutionary biology. We eagerly await further developments in computational biology that will allow even deeper understanding of the genetic possibilities present in possibly the most intriguing cell structure: the plastid.

Author Contributions: E.R. made the figures; E.R. and C.M.K. both wrote and edited the manuscript.

Funding: E.R. is funded by a Vanier Canada Graduate Scholarship.

Acknowledgments: The authors would like to thank Ellen Nisbet for her assistance preparing the manuscript, and Joel B. Dacks and members of the Dacks lab for insightful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **1967**, *14*, 225–274. [CrossRef]
2. Wallace, D.C. Structure and evolution of organelle genomes. *Microbiol. Rev.* **1982**, *46*, 208–240. [PubMed]
3. Wilson, R.J.; Williamson, D.H.; Preiser, P. Malaria and other Apicomplexans: The "plant" connection. *Infect. Agents Dis.* **1994**, *3*, 29–37. [PubMed]
4. Chase, M.W.; Fay, M.F. Ancient flowering plants: DNA sequences and angiosperm classification. *Genome Biol.* **2001**, *2*, reviews1012.1. [CrossRef] [PubMed]
5. Ruhfel, B.R.; Gitzendanner, M.A.; Soltis, P.S.; Soltis, D.E.; Burleigh, J.G. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **2014**, *14*, 23. [CrossRef] [PubMed]
6. Lung, S.-C.; Smith, M.D.; Chuong, S.D. Isolation of Chloroplasts from Plant Protoplasts. *Cold Spring Harb. Protoc.* **2015**, *2015*, 074559. [CrossRef] [PubMed]
7. Tonti-Filippini, J.; Nevill, P.G.; Dixon, K.; Small, I. What can we do with 1000 plastid genomes? *Plant J.* **2017**, *90*, 808–818. [CrossRef] [PubMed]
8. Smith, D.R.; Keeling, P.J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10177–10184. [CrossRef]
9. Cheng, S.; Melkonian, M.; Smith, S.A.; Brockington, S.; Archibald, J.M.; Delaux, P.-M.; Li, F.-W.; Melkonian, B.; Mavrodiev, E.V.; Sun, W.; et al. 10KP: A phylodiverse genome sequencing plan. *GigaScience* **2018**, *7*, 1–9. [CrossRef]
10. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big Data: Astronomical or Genomical? *PLoS Biol.* **2015**, *13*, e1002195. [CrossRef]
11. Spang, A.; Saw, J.H.; Jørgensen, S.L.; Zaremba-Niedzwiedzka, K.; Martijn, J.; Lind, A.E.; Van Eijk, R.; Schleper, C.; Guy, L.; Ettema, T.J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **2015**, *521*, 173–179. [CrossRef] [PubMed]
12. Zaremba-Niedzwiedzka, K.; Caceres, E.F.; Saw, J.H.; Bäckström, D.; Juzokaite, L.; Vancaester, E.; Seitz, K.W.; Anantharaman, K.; Starnawski, P.; Kjeldsen, K.U.; et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **2017**, *541*, 353–358. [CrossRef] [PubMed]
13. Salcher, M.M.; Andrei, A.-Ş.; Bulzu, P.-A.; Keresztes, Z.G.; Banciu, H.L.; Ghai, R. Visualization of Loki- and Heimdallarchaeia (Asgardarchaeota) by fluorescence in situ hybridization and catalyzed reporter deposition (CARD-FISH). *bioRxiv* **2019**. [CrossRef]
14. Saliba, A.-E.; Westermann, A.J.; Gorski, S.A.; Vogel, J. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **2014**, *42*, 8845–8860. [CrossRef] [PubMed]
15. Smith, D.R. RNA-Seq data: A goldmine for organelle research. *Brief. Funct. Genom.* **2013**, *12*, 454–456. [CrossRef] [PubMed]
16. Keeling, P.J.; Burki, F.; Wilcox, H.M.; Allam, B.; Allen, E.E.; Amaral-Zettler, L.A.; Armbrust, E.V.; Archibald, J.M.; Bharti, A.K.; Bell, C.J.; et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* **2014**, *12*, e1001889. [CrossRef] [PubMed]
17. Simpson, J.T.; Pop, M. The Theory and Practice of Genome Sequence Assembly. *Annu. Rev. Genom. Hum. Genet.* **2015**, *16*, 153–172. [CrossRef] [PubMed]

18. Sohn, J.-I.; Nam, J.-W. The present and future of de novo whole-genome assembly. *Brief. Bioinform.* **2018**, *19*, 23–40.
19. Ladoukakis, E.; Kolisis, F.N.; Chatziioannou, A.A. Integrative workflows for metagenomic analysis. *Front. Cell Dev. Biol.* **2014**, *2*, 70. [CrossRef]
20. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
21. Plotkin, J.B.; Kudla, G. Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* **2011**, *12*, 32–42. [CrossRef] [PubMed]
22. Chen, S.L.; Lee, W.; Hottes, A.K.; Shapiro, L.; McAdams, H.H. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3480–3485. [CrossRef] [PubMed]
23. Suzuki, H.; Morton, B.R. Codon Adaptation of Plastid Genes. *PLoS ONE* **2016**, *11*, e0154306. [CrossRef] [PubMed]
24. Bulmer, M. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics* **1991**, *129*, 897–907.
25. Sharp, P.M.; Emery, L.R.; Zeng, K. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. B Boil. Sci.* **2010**, *365*, 1203–1212. [CrossRef]
26. Morton, B.R. Chloroplast DNA codon use: Evidence for selection at the psb A locus based on tRNA availability. *J. Mol. Evol.* **1993**, *37*, 273–280. [CrossRef]
27. Morton, B.R. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J. Mol. Evol.* **1998**, *46*, 449–459. [CrossRef]
28. Gustafsson, C.; Govindarajan, S.; Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **2004**, *22*, 346–353. [CrossRef]
29. Inagaki, Y.; Simpson, A.G.B.; Dacks, J.B.; Roger, A.J. Phylogenetic Artifacts Can be Caused by Leucine, Serine, and Arginine Codon Usage Heterogeneity: Dinoflagellate Plastid Origins as a Case Study. *Syst. Biol.* **2004**, *53*, 582–593. [CrossRef]
30. Gray, M.W. Evolutionary Origin of RNA Editing. *Biochemistry* **2012**, *51*, 5235–5242. [CrossRef]
31. Jackson, C.J.; Gornik, S.G.; Waller, R.F. A Tertiary Plastid Gains RNA Editing in Its New Host. *Mol Biol Evol* **2013**, *30*, 788–792. [CrossRef] [PubMed]
32. Sun, T.; Bentolila, S.; Hanson, M.R. The Unexpected Diversity of Plant Organelle RNA Editosomes. *Trends Plant Sci.* **2016**, *21*, 962–973. [CrossRef] [PubMed]
33. Klinger, C.M.; Paoli, L.; Newby, R.J.; Wang, M.Y.-W.; Carroll, H.D.; Leblond, J.D.; Howe, C.J.; Dacks, J.B.; Bowler, C.; Cahoon, A.B.; et al. Plastid Transcript Editing across Dinoflagellate Lineages Shows Lineage-Specific Application but Conserved Trends. *Genome Biol. Evol.* **2018**, *10*, 1019–1038. [CrossRef] [PubMed]
34. Zauner, S.; Greilinger, D.; Laatsch, T.; Kowallik, K.V.; Maier, U.-G. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellateCeratium horridum. *FEBS Lett.* **2004**, *577*, 535–538. [CrossRef] [PubMed]
35. Wang, Y.; Morse, D. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate Lingulodinium. *Nucleic Acids Res.* **2006**, *34*, 613–619. [CrossRef] [PubMed]
36. Dang, Y.; Green, B.R. Substitutional editing of Heterocapsa triquetra chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA. *Gene* **2009**, *442*, 73–80. [CrossRef] [PubMed]
37. Iida, S.; Kobiyama, A.; Ogata, T.; Murakami, A. Identification of transcribed and persistent variants of the psbA gene carried by plastid minicircles in a dinoflagellate. *Curr. Genet.* **2009**, *55*, 583–591. [CrossRef]
38. Mungpakdee, S.; Shinzato, C.; Takeuchi, T.; Kawashima, T.; Koyanagi, R.; Hisata, K.; Tanaka, M.; Goto, H.; Fujie, M.; Lin, S.; et al. Massive Gene Transfer and Extensive RNA Editing of a Symbiotic Dinoflagellate Plastid Genome. *Genome Biol. Evol.* **2014**, *6*, 1408–1422. [CrossRef]
39. Liew, Y.J.; Li, Y.; Baumgarten, S.; Voolstra, C.R.; Aranda, M. Condition-specific RNA editing in the coral symbiont Symbiodinium microadriaticum. *PLoS Genet.* **2017**, *13*, e1006619. [CrossRef]
40. Hicks, J.; Lassadi, I.; Carpenter, E.; Eno, M.; Vardakis, A.; Waller, R.F.; Howe, C.; Nisbet, R.E.R. A Pentatricopeptide Repeat Protein in the Plasmodium apicoplast is essential and shows sequence-specific RNA binding. *bioRxiv* **2018**. [CrossRef]
41. Oldenkott, B.; Yamaguchi, K.; Tsuji-Tsukinoki, S.; Knie, N.; Knoop, V. Chloroplast RNA editing going extreme: More than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte Selaginella uncinata. *RNA* **2014**, *20*, 1499–1506. [CrossRef] [PubMed]

42. Shikanai, T. RNA editing in plants: Machinery and flexibility of site recognition. *Biochim. Biophys. Acta* **2015**, *1847*, 779–785. [CrossRef] [PubMed]
43. Dorrell, R.G.; Howe, C.J. Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10247–10254. [CrossRef] [PubMed]
44. Gray, M.W.; Lukeš, J.; Archibald, J.M.; Keeling, P.J.; Doolittle, W.F. Irremediable Complexity? *Science* **2010**, *330*, 920–921. [CrossRef] [PubMed]
45. Stiller, J.W.; Schreiber, J.; Yue, J.; Guo, H.; Ding, Q.; Huang, J. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* **2014**, *5*, 5764. [CrossRef] [PubMed]
46. Sevcikova, T.; Horák, A.; Klimes, V.; Zbránková, V.; Demir-Hilton, E.; Sudek, S.; Jenkins, J.; Schmutz, J.; Přibyl, P.; Fousek, J.; et al. Updating algal evolutionary relationships through plastid genome sequencing: Did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci. Rep.* **2015**, *5*, 10134. [CrossRef] [PubMed]
47. Dorrell, R.G.; Gile, G.; McCallum, G.; Méheust, R.; Bapteste, E.P.; Klinger, C.M.; Brillet-Guéguen, L.; Freeman, K.D.; Richter, D.J.; Bowler, C. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* **2017**, *6*, e23717. [CrossRef] [PubMed]
48. Barbrook, A.C.; Howe, C.J.; Purton, S. Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci.* **2006**, *11*, 101–108. [CrossRef]
49. Smith, D.R.; Crosby, K.; Lee, R.W. Correlation between Nuclear Plastid DNA Abundance and Plastid Number Supports the Limited Transfer Window Hypothesis. *Genome Biol. Evol.* **2011**, *3*, 365–371. [CrossRef] [PubMed]
50. Smith, D.R. Extending the Limited Transfer Window Hypothesis to Inter-organelle DNA Migration. *Genome Biol. Evol.* **2011**, *3*, 743–748. [CrossRef] [PubMed]
51. Dorrell, R.; Kuo, A.; Ibarbalz, F.; Rocha Jimenez Vieira, F.; Pierella Karlusich, J.J.; McFarlane, J.; Richardson, E.; Edgar, R.M.; Potvin, M.; Peng, Y.; et al. Arctic algae genomes libraries. *Apollo* **2018**. [CrossRef]
52. Ebenezer, T.E.; Carrington, M.; Lebert, M.; Kelly, S.; Field, M.C. Euglena gracilis Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. *Adv. Exp. Med. Biol.* **2017**, *979*, 125–140.
53. Curtis, B.A.; Tanifuji, G.; Burki, F.; Gruber, A.; Irimia, M.; Maruyama, S.; Arias, M.C.; Ball, S.G.; Gile, G.H.; Hirakawa, Y.; et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **2012**, *492*, 59–65. [CrossRef] [PubMed]
54. Arrigo, K.R. Marine microorganisms and global nutrient cycles. *Nature* **2005**, *437*, 349–355. [CrossRef] [PubMed]
55. Yu, M.; Ashworth, M.P.; Hajrah, N.H.; Khiyami, M.A.; Sabir, M.J.; Alhebshi, A.M.; Al-Malki, A.L.; Sabir, J.S.M.; Theriot, E.C.; Jansen, R.K. Chapter Five—Evolution of the Plastid Genomes in Diatoms. In *Advances in Botanical Research. Plastid Genome Evolution*; Chaw, S.-M., Jansen, R.K., Eds.; Academic Press: Cambridge, MA, USA, 2018; Volume 85, pp. 129–155.
56. Geisen, S. Thorough high-throughput sequencing analyses unravels huge diversities of soil parasitic protists. *Environ. Microbiol.* **2016**, *18*, 1669–1672. [CrossRef] [PubMed]
57. Bik, H.M.; Porazinska, D.L.; Creer, S.; Caporaso, J.G.; Knight, R.; Thomas, W.K. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* **2012**, *27*, 233–243. [CrossRef]
58. Stoeck, T.; Bass, D.; Nebel, M.; Christen, R.; Jones, M.D.M.; Breiner, H.; Richards, T.A. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **2010**, *19*, 21–31. [CrossRef]
59. Wylezich, C.; Herlemann, D.; Jürgens, K. Improved 18S rDNA amplification protocol for assessing protist diversity in oxygen-deficient marine systems. *Aquat. Microb. Ecol.* **2018**, *81*, 83–94. [CrossRef]
60. Porter, T.M.; Shokralla, S.; Baird, D.; Golding, G.B.; Hajibabaei, M. Ribosomal DNA and Plastid Markers Used to Sample Fungal and Plant Communities from Wetland Soils Reveals Complementary Biotas. *PLoS ONE* **2016**, *11*, e0142759. [CrossRef]
61. Woo, Y.H.; Ansari, H.R.; Otto, T.D.; Klinger, C.M.; Kolisko, M.; Michálek, J.; Saxena, A.; Shanmugam, D.; Tayyrov, A.; Veluchamy, A.; et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* **2015**, *4*, e06974. [CrossRef]
62. Nowack, E.C.; Melkonian, M.; Glöckner, G. Chromatophore Genome Sequence of Paulinella Sheds Light on Acquisition of Photosynthesis by Eukaryotes. *Curr. Biol.* **2008**, *18*, 410–418. [CrossRef] [PubMed]

63. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
64. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-Generation Machine Learning for Biological Networks. *Cell* **2018**, *173*, 1581–1592. [CrossRef] [PubMed]
65. Qu, K.; Guo, F.; Liu, X.; Lin, Y.; Zou, Q. Application of Machine Learning in Microbiology. *Front. Microbiol.* **2019**, *10*, 827. [CrossRef] [PubMed]
66. Barkan, A.; Rojas, M.; Fujii, S.; Yap, A.; Chong, Y.S.; Bond, C.S.; Small, I. A Combinatorial Amino Acid Code for RNA Recognition by Pentatricopeptide Repeat Proteins. *PLoS Genet.* **2012**, *8*, e1002910. [CrossRef] [PubMed]
67. Giudice, C.L.; Hernández, I.; Ceci, L.R.; Pesole, G.; Picardi, E. RNA editing in plants: A comprehensive survey of bioinformatics tools and databases. *Plant Physiol. Biochem.* **2019**, *137*, 53–61. [CrossRef]
68. Higgins, P.A.T.; Harte, J. Carbon Cycle Uncertainty Increases Climate Change Risks and Mitigation Challenges. *J. Clim.* **2012**, *25*, 7660–7668. [CrossRef]
69. Mapelli, F.; Michoud, G.; Aulenta, F.; Boon, N.; Borin, S.; Kalogerakis, N.; Scoma, A.; Daffonchio, D. Biotechnologies for Marine Oil Spill Cleanup: Indissoluble Ties with Microorganisms. *Trends Biotechnol.* **2017**, *35*, 860–870. [CrossRef]
70. Georgianna, D.R.; Mayfield, S.P. Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature* **2012**, *488*, 329–335. [CrossRef]
71. Gan, Q.; Jiang, J.; Han, X.; Wang, S.; Lu, Y. Engineering the Chloroplast Genome of Oleaginous Marine Microalga *Nannochloropsis oceanica*. *Front. Plant Sci.* **2018**, *9*. [CrossRef]