DNA Research, 2017, 24(3), 327–332
doi: 10.1093/dnares/dsw044
Advance Access Publication Date: 17 April 2017
Full Paper

OXFORD

Full Paper

ChloroMitoCU: Codon patterns across organelle genomes for functional genomics and evolutionary applications

Gaurav Sablok[1,†], Ting-Wen Chen[2,†], Chi-Ching Lee[2], Chi Yang[2], Ruei-Chi Gan[2], Jill L. Wegrzyn[3], Nicola L. Porta[4,5], Kinshuk C. Nayak[6], Po-Jung Huang[2], Claudio Varotto[1], and Petrus Tang[2,7,*]

[1]Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy, [2]Bioinformatics Core Laboratory, Molecular Medicine Research Center, Chang Gung University, Kweishan, Taoyuan 333, Taiwan, [3]Department of Ecology and Evolutionary Biology, University 10 of Connecticut, 75 North Eagleville Road, Storrs, CT 06269-3043 USA, [4]Department of Sustainable Agrobiosystems and Bioresources, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy, [5]MOUNTFOR Project Centre, European Forest Institute, Via E. Mach 1, 38010 San Michele all'Adige, Trento, Italy, [6]Bioinformatics Centre, Institute of Life Sciences, Department of Biotechnology, Govt. India, Nalco Square, Bhubaneswar - 751 023, India, and [7]Molecular Infectious Diseases Research Center, Chang Gung Memorial Hospital, Kweishan, Taoyuan 333, Taiwan

*To whom correspondence should be addressed. Tel: 001-860990-3742. Email: sablokg@gmail.com (Gaurav Sablok) or petang@mail.cgu.edu.tw (Petrus Tang)

[†]These authors are contributed equally to the work.

Edited by Prof. Kenta Nakai

Received 18 March 2016; Editorial decision 13 September 2016; Accepted 14 September 2016

Abstract

Organelle genomes are widely thought to have arisen from reduction events involving cyanobacterial and archaeal genomes, in the case of chloroplasts, or α-proteobacterial genomes, in the case of mitochondria. Heterogeneity in base composition and codon preference has long been the subject of investigation of topics ranging from phylogenetic distortion to the design of overexpression cassettes for transgenic expression. From the overexpression point of view, it is critical to systematically analyze the codon usage patterns of the organelle genomes. In light of the importance of codon usage patterns in the development of hyper-expression organelle transgenics, we present ChloroMitoCU, the first-ever curated, web-based reference catalog of the codon usage patterns in organelle genomes. ChloroMitoCU contains the pre-compiled codon usage patterns of 328 chloroplast genomes (29,960 CDS) and 3,502 mitochondrial genomes (49,066 CDS), enabling genome-wide exploration and comparative analysis of codon usage patterns across species. ChloroMitoCU allows the phylogenetic comparison of codon usage patterns across organelle genomes, the prediction of codon usage patterns based on user-submitted transcripts or assembled organelle genes, and comparative analysis with the pre-compiled patterns across species of interest. ChloroMitoCU can increase our understanding of the biased patterns of codon usage in organelle genomes across multiple clades. ChloroMitoCU can be accessed at: http://chloromitocu.cgu.edu.tw/

Key words: chloroplast, codon usage, functional genomics, evolution, mitochondria, web application

1. Introduction

Organelle genomes have been the focus of much research for several decades because of their biological activity linked to photosynthetic efficiency, energy storage, and the regulation of gene expression.[1] They have evolved independently, and their genetic architecture has been shaped as a consequence of evolutionary forces.[2] One such evolutionary force is codon usage, which has evolved independently in organelle genomes. Codon usage, a pattern that describes the degeneracy of the genetic code and illustrates differential usage of synonymous codons, is thought to be under the influence of several factors such as mutational pressure,[3] translational selection,[4] and tRNA dependencies.[5,6] In model organisms, the evolutionary forces shaping codon usage have been assessed using either genome-wide sets of predicted coding regions or subsets of highly expressed ribosomal proteins.[7] In non-model organisms for which complete genome sequences are not available, codon usage has often been estimated using expressed sequence tags[4,8,9] or organelle genomes,[3] providing clues about the mechanistic evolution of the nuclear and organelle genomes, respectively. From an evolutionary point of view, codon usage plays an important role in defining phylogenies and within-taxon or among-taxa variance.[10] Previous studies demonstrated the clade-specific molecular selection and codon adaptation of *rbcL*, which encodes the large subunit of ribulose 1,5-bisphosphate carboxylase, or RuBisCO.[11] Furthermore, chloroplast genomes replicate through a D-loop mechanism and transcribe genes in both the sense and antisense orientation, like the nuclear genome. Codon usage patterns and strand asymmetry have been linked to transcription selection. The highly expressed genes tend to be localized on the leading strand, thus creating a coding bias, which orients the directionality of replication and transcription.[12]

Recently, there has been considerable focus on the exploitation of organelles as a source of pharmaceuticals and as biofactories,[13] in which foreign genes are reverse engineered and overexpressed.[14] This paves the way for hyper-expression-based genetic engineering using either a single gene, such as Toc cyclase or γ-Toc methyltransferase in lettuce,[15] or an entire pathway, such as the mevalonate pathway in tobacco.[14,16] For efficient transgenic overexpression in chloroplasts, well-established protocols have been widely demonstrated[17] based on the alternations of codon usage patterns between organelle genomes, which have predominantly A-ending and/or U-ending codons, and nuclear genomes, which have predominantly G-ending, C-ending, or U-ending codons.[18] Despite the relevance of codon usage patterns in phylogenomics and the design of optimal gene-expression experiments, relatively few open-source repositories have been developed for the wide exploration of synonymous codon usage across organelle genomes. Only a few of the repositories in existence encompass even a modest range of taxa; most are either organism-centric, with compiled codon usage tables for microbial organisms,[19] or simple, single-organism, web-based displays of codon usage statistics (available at: http://www.cmbl.uga.edu/software/codon_usage.html).

In the case of eukaryotes, the majority of the resources available to date are focused on either a single model organism, such as http://crumb.stanford.edu/community/codon_usage.shtml for *Saccharomyces cerevisiae*, which limits the exploration of codon usage patterns across multiple species. It is worthwhile to mention that the Kazusa Codon Usage Database (http://www.kazusa.or.jp/codon) is the only repository that hosts the pre-compiled codon usage counts for a relatively large number of species, including the organelle genomes.[20] The latest update of the Kazusa database dates back to 2007 (NCBI-GenBank Flat File Release), which presents a challenge for the exploration and cross-species comparison of codon patterns. The Kazusa database also lacks features such as the cross-comparison of organelle genomes and the comparison of codon usage patterns between user-submitted genes and the pre-compiled organelle genomes. Additionally, the Kazusa database contains a relatively small number of organelle genomes, limiting its ability to explore the recently published organelle genomes and the applicability of those genomes as transgenic vehicles. In a web report on the NCBI Genbank release, Richmond stated that one of the top priorities in codon-usage research was a necessary update of the pre-compiled codon tables, urging that such tables include data from non-model species.[21] In light of the limited resources available for the study of organelle genome-specific patterns of codon usage and the critical importance of those patterns to applications to enhance organelle-based bio-factories, multi-gene targeting, and recombinant protein expression[22], we developed ChloroMitoCU, which provides simple access to patterns of codon usage in 328 chloroplast genomes (29,960 CDS) and 3,502 mitochondrial genomes (49,066 CDS). ChloroMitoCU allows the comparison of organelle-specific codon patterns across multiple species with the same translational table and also allows comparative assessments of the codon usage patterns between user-submitted transcripts or genes and the pre-compiled organelle genomes.

2. Material and methods

2.1 Genome data retrieval and codon usage patterns

We retrieved all of the chloroplast (328) and mitochondrial (3,502) genomes from the NCBI organelle genomes repository.[23] The coding regions of the respective genomes were downloaded and subsequently parsed for a length threshold. For the identification of the codon usage patterns, we maintained a threshold of 300 bp. We used CodonW (available from codonw.sourceforge.net) to identify the codon usage patterns in each of the genes with a minimum length of 300 bp. To identify the strand-specific codon usage patterns, genes localized on the positive (+) and negative (−) strands were separated based on the organelle genome annotation available in NCBI RefSeq. The current version of ChloroMitoCU contains 29,960 complete (full-length) protein-coding genes (CDSs) from chloroplasts and 49,066 CDSs from mitochondria. The relative synonymous codon usage (RSCU) index was calculated as previously described.[3] RSCU values represent the ratio between the observed usage frequency of a codon in a sample and the expected usage frequency if all of the codons in the synonymous codon family for the particular amino acid are used equally.[3] To predict potential CDSs in assembled transcripts or organelle genomes, we deployed a modified version of the Transdecoder prediction pipeline (available from https://transdecoder.github.io), which incorporates all of the available translational tables for wider predictions of CDSs across many organisms.

3. Results and discussion

3.1 Development and applications of ChloroMitoCU

For efficient access, ChloroMitoCU was developed and is hosted on a 64-bit Linux server pre-installed with Python version 2.7.3, Apache (http://www.apache.org), and PHP (http://www.php.net/). ChloroMitoCU runs all analyses in real time via Python modules (version 2.7.10), and the parsed results integrate with PHP modules to provide a web-based display. For each organelle genome, the codon information is further sorted into hierarchal levels of information such as the genomic GC content of the first, second, and third codon positions, the number of encoded proteins, the genome length, the number

Figure 1. The efficient browsing system of ChloroMitoCU.

of tRNAs, and the usage patterns and relative frequencies of codons (Fig. 1). Figure 1 shows the user-friendly interface of ChloroMitoCU, which offers a two-way search for browsing chloroplast-specific codon usage patterns (Browse ChloroCU) and mitochondria-specific codon usage patterns (Browse MitoCU). In addition, a dropdown menu provides the alphabetically sorted list of the chloroplast and mitochondrial genomes with associated translational codes. Lastly, all of the alphabets in the quick-search forms are hyperlinked to the corresponding codon usage-pattern page of the selected species.

For each organism and organelle, ChloroMitoCU hosts organism-specific and organelle-specific web layouts, which display the species name and corresponding translational table, the amino acid organization, information about synonymous codons from the species-specific translational table (Fig. 1). Organism specific webpages in both Browe ChloroCU and Browse MitoCU displays the species-specific distribution of GC content in the CDSs at the first, second, and third positions, and color-coded codons display of the RSCU values, with stop codons indicated in white and marked by asterisk (*) (Fig. 2). Additionally, it displays the summary of the codon usage statistics with information on the synonymous codons as per the species specific translational table (Fig. 2). Finally, the last module in each organelle-specific or organism-specific web page in Browse ChloroCU and BrowseMitoCU provides real-time comparisons of the codon usage-frequency profiles of the user-selected plastome and any phylogenetically related chloroplast genome (Fig. 3). The inference of RSCU values can help the end-user to investigate the codon usage independently of the amino acid composition. RSCU values > 1 (red horizontal bars), as compared with those < 1 (green horizontal bars), for a codon imply that the codon is used more frequently than expected based on the amino acid content of the gene or genome (Fig. 2).

An additional module (Fig. 4) implemented in ChloroMitoCU allows the user to either analyze the CDSs for codon usage patterns or predict the CDSs from stranded or non-stranded assemblies or assembled genes. It further allows the user to select the translational table according to the organelle and also to select between complete CDS predictions or all reported CDSs within the assembled regions. The results page of the prediction module allows the comparative assessment of the codon usage patterns across other pre-embedded organelle genomes using the same codon translational table. The Python module (version 2.7.10) allows the user to simultaneously view the codon usage patterns of the user-input gene of interest and the organelle-specific codon patterns of the selected species in order to understand the codon-specific patterns according to the specific organelle of interest.

3.2 Functional genomics and evolutionary applications of ChloroMitoCU

Functional genomics is at the forefront of the accelerated use of plants as models for antibody expression and oral vaccine development.[24] Recent reports demonstrate the role of codon optimization in the enhancement of chloroplast gene expression and translation rates.[25] Taking that perspective into account, ChloroMitoCU provides comparative analysis of the codon usage patterns of a wide range of chloroplast genomes, which holds great potential for the optimization of gene expression and for heterologous gene expression. Recently, the investigation of a hypothesis coupling codon usage and

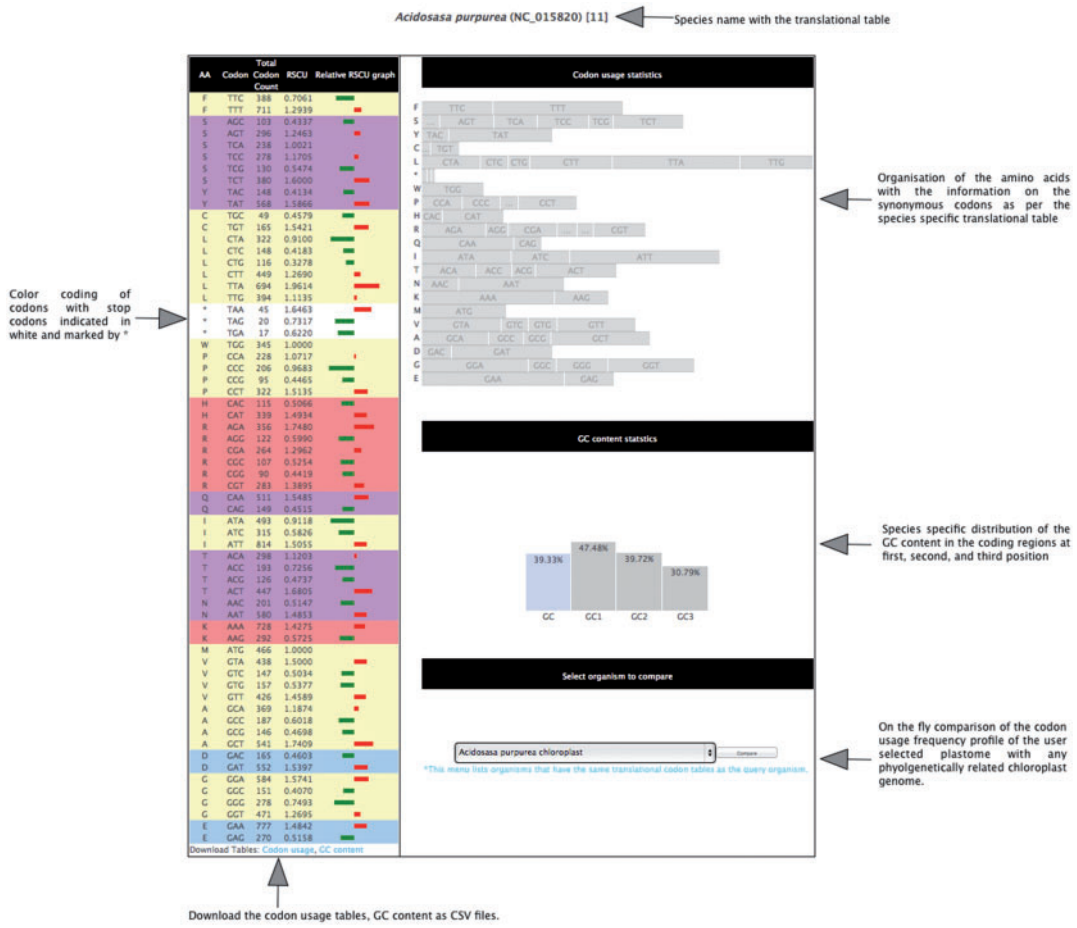

Figure 2. Graphical visualization of the codon counts and real time comparison of the codon counts across several species for comparative assessment across phylogenies.

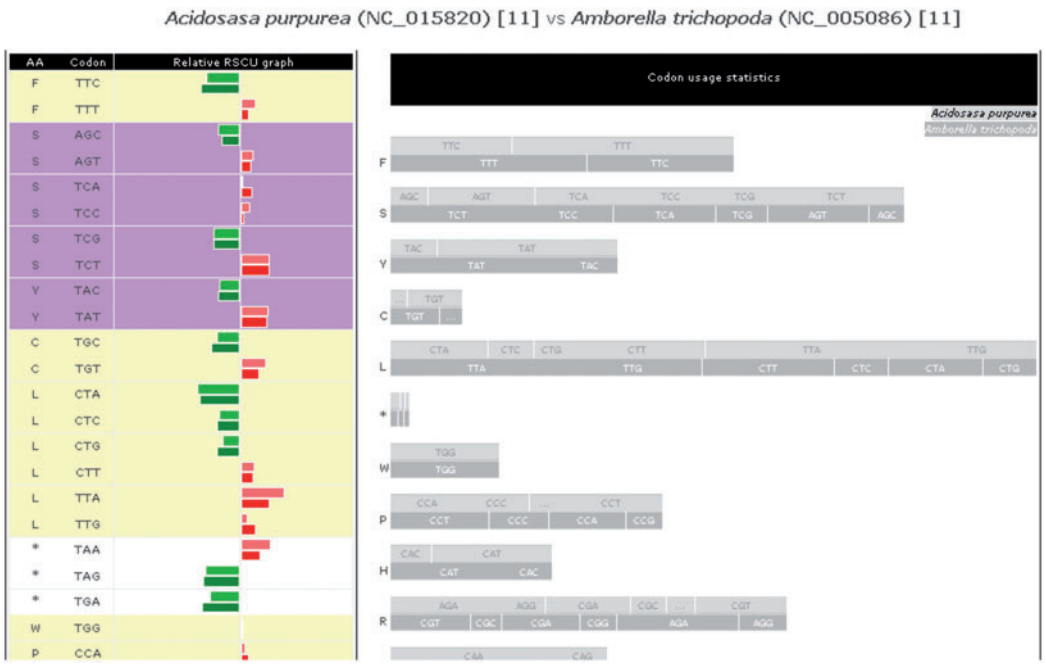

Figure 3. Comparative profiling of the codon usage tables across two organelle genomes.

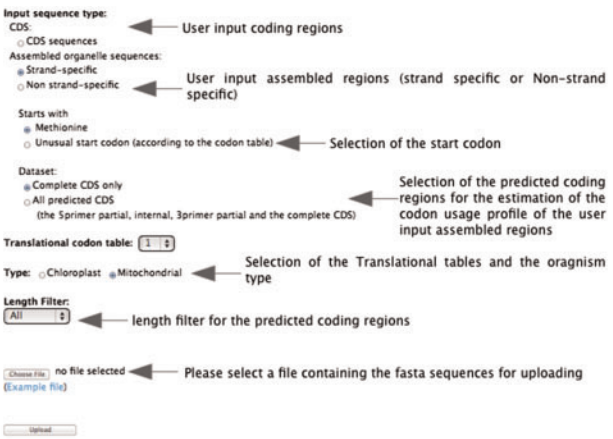

Figure 4. On-the-fly prediction of CDSs from assembled organelle genome fragments and estimation of the codon counts, taking into account the translational codon table.

tRNA import in the unicellular model algae, *Chlamydomonas reinhardtii*, suggested that co-evolution is the mechanism underlying the dynamic import of tRNAs and that the hypothesis could be adapted to explain the composition of the mitochondrial genome.[26] Comparative analyses of base composition revealed a varying level of mutational pressure acting upon chloroplast and nuclear genes[18] however, base frequencies at fourfold-degenerate sites in mitochondrial genomes displayed a strong context-dependent mutational pattern, indicating correlations between pairs of neighbouring bases.[27] Last, taking into account the relative differences in the strand-specific codon usage patterns in organelle genomes, especially in mitochondrial genomes,[27] ChloroMitoCU, provide the strand specific browsing of the codon usage patterns in both organelle genomes (chloroplast and mitochondria) (Fig. 1) to understand the evolutionary role of the strand asymmetry in shaping the codon usage of the user interested organelle and to infer the difference in strand-specific mutational biases.

Comparative analyses with nuclear genes revealed the dominance of variable patterns of evolutionary constraints, particularly in chloroplast and mitochondrial genes, which were found to be influenced by the translation level.[28] The deciphering of species-specific patterns of codon usage has effectively paved the way for the enhancement of heterologous gene translation in several model and non-model organisms[29,30]. Comparative assessments of codon usage patterns are important for understanding patterns of codon preferences. For this reason, we implemented an additional module in ChloroMitoCU that allows assessment of two organelle genomes if the genomes share the same codon table. When compared with the previously developed Kazusa database (http://www.kazusa.or.jp/codon/), ChloroMitoCU provides an easier way to access pre-compiled codon usage tables and has several added functionalities.

4. Conclusion

ChloroMitoCU provides curated information of codon usage patterns in organelle genomes and allows for the comparative assessment of those patterns across organelle genomes. ChloroMitoCU also offers the prediction of CDSs and the comparative assessment of codon usage between the predicted CDSs and the embedded organelle genomes. In the future, ChloroMitoCU will be updated as per the NCBI RefSeq release, and the newly reported chloroplast and mitochondrial genes, and their corresponding patterns of codon usage, will be incorporated. The enhancement of expression through the selection of appropriate codon usage patterns has been used to transform several key traits to manage diseases, including viral diseases and to create edible vaccines.[29] We believe that ChloroMitoCU will serve as a standard platform to enhance organelle genomics and gene expression and to understand the evolution of the various codon usage patterns in relation to the genomic compositions across the Kingdoms of life.

Acknowledgements

GS thanks Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 S. Michele all' Adige (TN), Italy for providing the facility for computational analysis and GS thanks Petrus Tang for supporting for the open access fees.

Authors' contributions

G.S, C.V. conceived and designed the research; G.S. and T.W.C. performed the re- search; C.C.L., R.C.G., P.J.H., and P.T. contributed to the design of the web server. G.S. wrote the article, and K.C.N, J.W and N.L.P provided revisions. All authors have read and approved the article.

References

1. Allen, J.F. 2015, Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proc. Natl. Acad. Sci. USA*, **112**:10231–8.
2. Smith, D.R. and Keeling, P.J. 2015, Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. USA*, **112**:10177–84.
3. Sablok, G., Nayak, K.C., Vazquez, F. and Tatarinova, T.V. 2011, Synonymous codon usage, GC(3), and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. *Mol. Biotechnol*., **49**:116–28.
4. Sablok, G., Wu, X., Kuo, J., Nayak, K.C., Baev, V., Varotto, C. and Zhou, F. 2013, Combinational effect of mutational bias and translational selection for translation efficiency in tomato (Solanum lycopersicum) cv. Micro-Tom. *Genomics*, **101**: 290–5
5. Gilchrist, M.A., Shah, P. and Zaretzki, R. 2009, Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, **183**:1493–505.
6. Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C. and Zhang, J. 2012, Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet*., **8**: e1002603.
7. Faux, N.G., Huttley, G.A., Mahmood, K., Webb, G.I., de la Banda, M.G., and Whisstock, J.C. 2007, RCPdb: an evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res*., **17**:1118–27
8. Serres-Giardi, L., Belkhir, K., David, J. and Glémin, S. 2012, Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell*, **24**: 1379–97.
9. Ahmad, T., Sablok, G., Tatarinova, T.V., Xu, Q., Deng, X.X. and Guo, W.W. 2013, Evaluation of codon biology in citrus and Poncirus trifoliata based on genomic features and frame corrected expressed sequence tags. *DNA Res*., **2**:135–50
10. Christianson, M.L. 2005, Codon usage patterns distort phylogenies from or of DNA sequences. *Am. J. Bot*., **92**:1221–33.
11. Wall, D.P. and Herbeck, J.T. 2003, Evolutionary patterns of codon usage in the chloroplast gene rbcL. *J. Mol. Evol*., **56**: 673–88.
12. Morton, B.R. 1999, Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. *Proc. Natl. Acad. Sci. USA*, **27**:5123–8.
13. Verma, D. and Daniell, H. 2007, Chloroplast vector systems for biotechnology applications. *Plant Physiol*., **145**:1129–43.

14. Jin, S. and Daniell, H. 2015, The engineered chloroplast genome just got smarter. *Trends Plant Sci*., **20**:622–40.
15. Yabuta, Y., Tanaka, H., Yoshimura, S., Suzuki, A., Tamoi, M., Maruta, T. and Shigeoka, S. 2013, Improvement of vitamin E quality and quantity in tobacco and lettuce by chloroplast genetic engineering. *Transgenic Res*., **22**:391–402.
16. Kumar, S., Hahn, F.M., Baidoo, E., et al. 2012, Remodeling the isoprenoid pathway in tobacco by expressing the cytoplasmic mevalonate pathway in chloroplasts. *Metab. Eng*., **14**: 19–28.
17. Verma, D, Samson, NP, Koya, V. and Daniell, H. 2008, A protocol for expression of foreign genes in chloroplasts. *Nat. Protoc*., **3**:739–58.
18. Liu, Q. and Xue, Q. 2005, Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J. Genet*., **84**:55–62.
19. Hilterbrand, A., Saelens, J. and Putonti, C. 2012, CBDB: The codon bias database. *BMC Bioinformatics*, **13**:62. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-62
20. Nakamura, Y., Gojobori, T. and Ikemura, T. 2000, Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*., **28**:292.
21. Richomond, T. 2000, Precompiled codon-usage tables. *Genome Biol*., 2000, 1:reports241.
22. Ma, J.K., Drake, P.M. and Christou, P. 2003, The production of recombinant pharmaceutical proteins in plants. *Nat. Rev. Genet*., **4**: 794–805.
23. Pruitt, K.D., Tatusova, T., Klimke, W. and Maglott, D.R. 2009, NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*., **37**:D32–36.
24. Shahaid, N. and Daniell, H. 2016, Plant based oral vaccines against zoonotic and non-zoonotic diseases. *Plant Biotechnol. J*., doi: 10.1111/pbi.12604.
25. Kwon, K.C., Chan, H.T., Leon, I.R., Williams-Carrier, R., Barkan, A. and Daniell, H. 2016, Codon-optimization to enhance expression yields insights into chloroplast translation. *Plant Physiol*., doi: 10.1104/pp.16.00981.
26. Salinas, T., Duby, F., Larosa, V., et al. 2012, Co-evolution of mitochondrial trna import and codon usage determines translational efficiency in the green alga chlamydomonas. *PLoS Genet*., **8**: e1002946.
27. Jia, W. and Higgs, P.G. 2007, Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol. Biol. Evol*., **25**: 339–51.
28. Zhang, W.J., Zhou, J., Li, Z.F., Wang, L., Gu, X. and Zhong, Y. 2007, Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in Triticum aestivum L. *J. Integr. Plant Biol*., **49**:246–54.
29. Daniell, H., Ruiz, G., Denes, B., Sandberg, L. and Langridge, W. 2003, Optimization of codon composition and regulatory elements for expression of human insulin like growth factor-1 in transgenic chloroplasts and evaluation of structural identity and function. *BMC Biotechnol*., **9**:33.
30. Daniell, H., Chan, H.-T. and Pasoreck, E.K. 2016, Vaccination through chloroplast genetics: affordable protein drugs for the prevention and treatment of inherited or infectious human diseases. *Annu. Rev. Genet*., **50**:1.