Current Plant Biology 7–8 (2016) 34–38

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Chloroplast genomics: Expanding resources for an evolutionary conserved miniature molecule with enigmatic applications☆

Gaurav Sablok[a,*], Suresh B. Mudunuri[b], David Edwards[c], Peter J. Ralph[a]

[a] Climate Change Cluster, University of Technology Sydney, New South Wales 2007, Australia
[b] Centre for Bioinformatics Research, SRKR Engineering College, Chinna Amiram, Bhimavaram, West Godavari District, Andhra Pradesh 534204, India
[c] School of Plant Biology, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form 7 December 2016
Accepted 7 December 2016

ABSTRACT

Chloroplast, methylation depreived uniparental organelle genome is the most studied organelle genome from the perspective of evolution and functional omics. Recent advances in organelle genome sequencing both in terms of genome or transcriptome sequencing has opened a wide range of opportunities to understand the transcriptional and translational role of the genes mainly involved in the light harvesting apparatus and the evolution of the inverted repeats across the lineage. However, as compared to the nuclear genome, limited resources are available in case of organelle genome. In this review, we discuss the recent advances in the chloroplast genomics and the resources that have been developed for understanding the evolution, repeat patterns, functional genomics of this miniature molecule with enigmatic applications.

1. Chloroplast genomes: dynamic organization and evolutionary fluctuations

Photosynthesis is critical to all aspects of plant life and to combat the environmental fluctuations. Chloroplast, evolutionary conserved and endoymbiotically originated molecule play a major role in photosynthesis by acting as host to three major complex such as photosystem II (PSII), the cytochrome b6f complex (Cytb6f), and photosystem I (PSI) [1]. Evolutionary conservation of these complexes in chloroplast genome thylakoid membrane represents the main sites of the light capture and the oxygen production as well as playing a major role in the light state transitions with plastid division apparatus responsible for the binary fission spatially distributed between the stromal and cytosolic space [2]. Among the spatially distributed genes in circular fashion, chloroplast represents a set of genes, which are vital for controlling the photosynthetic efficiency and to determine the dynamic organization of the thylakoid membrane and cyclic electron flow [3]. Evolutionary conserved organization of chloroplast genomes, which is circular in nature and follows a D-loop replication model is structured in a quadripartite structure, which is partitioned into two repeat regions, which are defined by the differences in the length as large single copy (LSC) regions spanning across a length of 80–90 kb and a short single copy region (SSC), representing a 16–27 kb region. Organization of these LSC and SSC regions is a dynamic process and has been widely reported to undergo expansion and contraction [4]. Although the evolutionary conservation of the chloroplast genic regions has been widely reported as exemplified by their use as molecular barcodes, few instances of rapidly evolving genes such as rbcL, which encodes the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO), and plays a major role in carbon assimilation and other genes such as matK (maturase K), ndhB and psbA-J, which are involved in modulating the state transitions has also been seen [5]. In contrast, the repeat organization is very dynamic and although conserved across the angiosperms, dynamic loss of the inverted repeat copies has been widely documented amongst the gymnosperms [6]. Taking all these structural variations with in the size of (150–160 kb), it worth to highlight the role of evolutionary conserved, distinct and model genome to understand the genome fluctuations (Fig. 1). From the view point of regulatory genomics, transcriptional and, transcriptional flux, chloroplast genomes have been widely explored in addition to point mutants and also the identification of the RNA Editing events.

☆ This article is part of a special issue entitled "Genomic resources and databases", published in the journal Current Plant Biology 7–8, 2016.

* Corresponding author.
E-mail address: sablokg@gmail.com (G. Sablok).

http://dx.doi.org/10.1016/j.cpb.2016.12.004

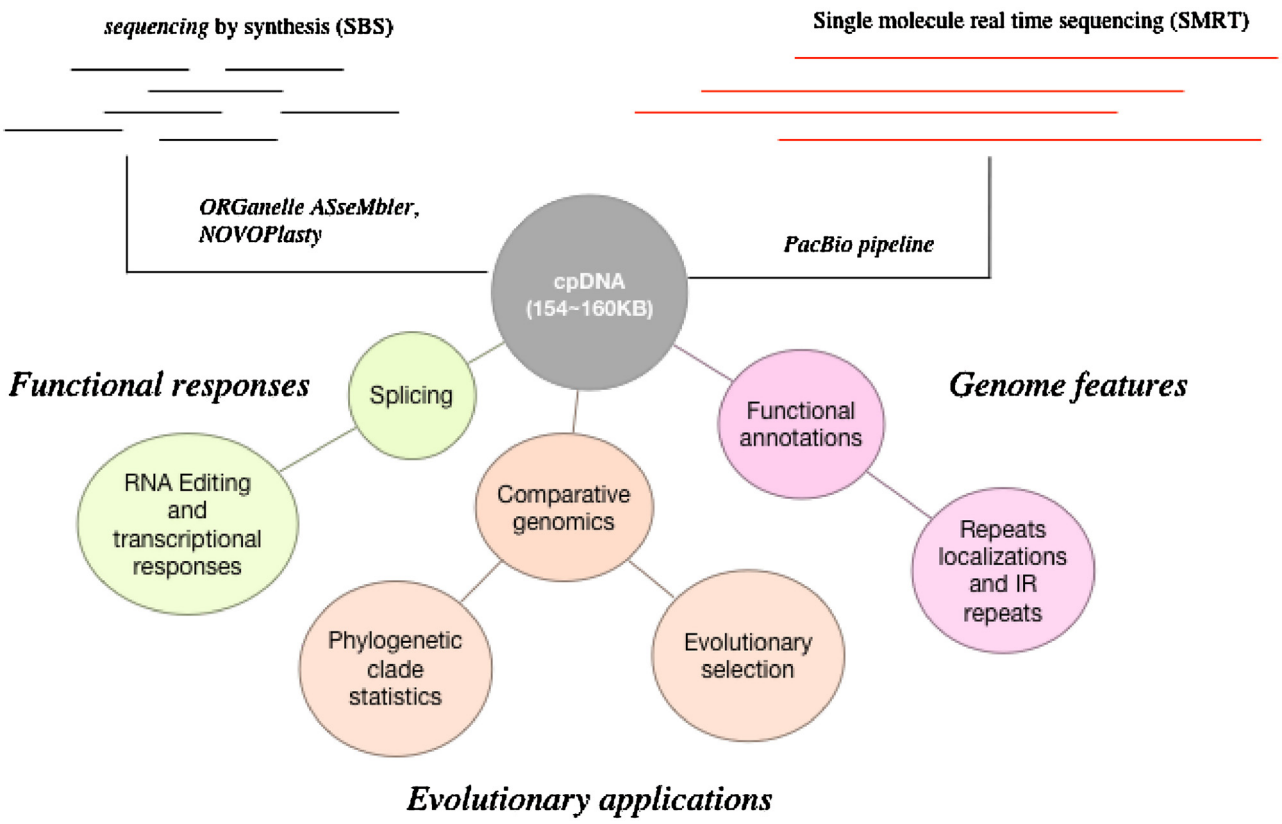

Fig. 1. Genomics and applications of chloroplast genomics.

2. Resources and tools for chloroplast genomics and functional genomics

In the past few years, considerable focus has been leveraged on the development of tools and genomic resources for advancing the chloroplast genomics. Here we outline the recent tools that have been developed for the chloroplast genomics from the viewpoint of genome assembly, annotation, evolutionary aspects, repeats, markers, and functional genomics (Table 1).

2.1. Organelle genome assembly

ORGanelle ASseMbler [7] is a useful tool for assembling of organelle genome sequences such as mitochondria and plant plastid genomes. The tool can be used to assemble small sequences that are over-represented in a whole genome shotgun sequence dataset. ORGanelle ASseMbler is a command line open source software tool developed using Python libraries and works in Linux and MacOSX systems. The implemented algorithm is linearized in a three step model: 1. The first step indexes the sequencing reads and then assembles the organelle genome with an option of the assembling graph in GML format. The assembling graph can be visualized using any graph visualization tools. The last step involves the extraction of the organelle specific sequences from the graph in a single FASTA file. The limitation of the algorithm is that the implemented algorithm is not capable of reorienting the circular structure of the chloroplast genome and thus manual curation is required to circularize the genome.

NOVOPlasty [8] is a recently published algorithm, which uses seed-extend based assembler and perform the organelle genome assembly by hashing the sequences from the whole genome sequencing (WGS) runs into a table, which allows for the rapid sorting of the sequences. Assembling of the organelle genomes starts with the seed sequence, which acts as an anchor to the extend the seed bi-directionally. The unique feature of the NOVOPlasty is that the user can specify the sequences, which can be a single organelle genome reads, a closely related organelle gene or an evolutionary related or distinct chloroplast genome. The unique feature of NOVOPlasty is that using the seed-extend approach, the assembler assembles the genome in circularized format if both the ends of seed-extend overlap by 200 bp [8]. The tool works on both Linux and MacOSX operating system.

2.2. Functional annotation

Functional annotation of chloroplast genome is an important process, as the rate of molecular evolution depends on the well annotated genome. Previously DOGMA [9] has been developed and has been the gold standard for the annotation of chloroplast genomes. Here, we describe some of the recent tools that have been developed for the functional annotation:

PLANN (Plastome Annotator) [10] is a command-line tool developed for the automated annotation of the assembled chloroplast genomes by comparing the user given chloroplast genome to a well annotated chloroplast genome. The tool has been designed to work on unix-based operating systems including Linux and MacOSX. PLANN uses NCBI tools BLASTN, tbl2asn and Sequin to perform the annotation. The graphical user interface application Sequin is first used to generate a template file for the output. The input files include a new plastome fasta sequence, a reference plastome file in GenBank format and a Sequin template file. PLANN annotates the new plastome by matching sequences with the gene sequences of reference genome using BLASTN and tbl2asn, and then transforms them into corresponding genomic locations of the new plastome in Sequin format.

Table 1
List of the recently developed resources for chloroplast genomics.

Resource	Type	Reference/Link
ORGanelle ASseMbler	Assembler	http://pythonhosted.org/ORG.asm/ [7]
NOVOPlasty	Assembler	Dierckxsens et al. [8]
PLANN	Annotation	Huang and Cronk [10]
CpGAVAS	Annotation	Liu et al. [11]
Verdant	Annotation and Evolution	McKain et al. [14]
CGAP	Comparative genomics	Cheng et al. [35]
ChloroMitoSSRDB and ChloroMitoSSRDB 2.00	Marker profiling	Sablok et al. [23,24]
ChloroSSRdb	Marker profiling	Kapil et al. [28]
ChloroMitoCU	Codon usage	Sablok et al. (2016)
Chloro-Seq	Transcriptional profiling	Castandet et al. [30]
PREPACT	RNA Editing	Lenz et al. [33]
OrganellarGenomeDRAW	Visualization	Lohse et al. [36]

CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission) [11] is an online webserver meant to provide standard functions to annotate and analyse chloroplast genome sequences. Additionally, it can generate circular genome maps, summary statistics of annotated genome and creation of files for GenBank submission. CpGAVAS has been developed to overcome the limitations of DOGMA, the popularly used chloroplast genome annotation web server. CpGAVAS has been implemented using Perl Catalyst Web Application Framework and a combination of Perl programs. The CpGAVAS server accepts a completely sequenced chloroplast genome as input and predicts its protein coding regions, rRNA genes, tRNA genes and inverted repeats through the comparative annotation with well annotated chloroplast genomes. It also includes tRNAScan [12] for the prediction of the tRNA in the chloroplast genomes. Protein coding regions are predicted using *ab initio* gene prediction tools and similarity based approaches. GenBank annotations of chloroplast genomes are first used to cluster the protein, CDS and rRNA gene sequences into homologous groups and formed into one blast-able database. The database is further used to create reference protein and cDNA/rRNA gene dataset for each input genome sequence. The reference protein, cDNA and rRNA gene sequences are searched using Blastx, Blastn, protein2genome, est2genome programs and corresponding best hits are used to annotate the input genome sequence. Further, inverted repeats of the input genome are identified using the Vmatch [13] software tool and tRNAs are identified using tRNAscan [12].

Verdant [14] is a new developed database driven suite of tools specifically designed for annotation, alignment and tree generation of chloroplast genomes. It is a web-based software connected to a database that provides accurate annotation of chloroplast genomes without manual intervention. Verdant uses different programs namely annoBTD (unpublished), MAFFT [15], RaxML [16], Circos [17] and JBrowse [18] to perform defined functions. AnnoBTD has been implemented to automate annotation without any manual editing. Protein coding regions of the input genome are identified by using *de novo* ORF identification, which is a novel feature of AnnoBTD. rRNAs and tRNAs are detected and annotated by blastn. Very small exons which are missed by other annotation programs are also detected by AnnoBTD. The extensive features of Verdant not only includes the annotation of the chloroplast genome but also allows for the automated alignment of the annotated genes, rRNAs and tRNAs, introns and intergenic regions using progressive and iterative refinement algorithms as implemented in MAFFT. Alignments done using the MAFFT can be passed to RaxML for phylogeny estimation. For the visualization of annotations, Circos and JBrowse has been implemented, which allows the visualization of the circular features. Verdant is developed using PHP, MySQL, Perl, JS, HTML and CSS. Users can create their own projects and can perform taxon selection, feature selection, alignment and phylogenetic tree reconstruction.

CGAP (Chloroplast Genome Analysis Platform) [19] is a comprehensive resource developed for comparative analysis of chloroplast genomes. CGAP is an interactive web-based tool with features like genome collection, visualization, phylogenetic analysis, content comparison and annotation of complete chloroplast genomes. It contains a back-end database of hundreds of complete chloroplast genomes including their annotation features such as genes, CDS, tRNA, rRNA, promoter, exon/intron regions, and repeats. The visualization module of CGAP can be used to create high quality genome maps to visualize circular complete genomes, linear regional genomes, modified published genomes and user unpublished genomes. CGAP can also be used to compare the similarities and differences of the feature content between different chloroplast genomes. CGAP is also integrated with phylogeny tools that uses an alignment free method for tree generation and comparison. The Genome Annotation module of CGAP can be used to annotate new chloroplast genomes based on the reference chloroplast genomes in the CGAP database by using BLAST programs. CGAP has been developed using Python language and Web2py web framework.

2.3. Visualization

Visualization of the chloroplast genome characteristics is an important feature that allows the display of chloroplast genes in circular fashion with additional features such as rRNAs, tRNAs, genic regions and IR boundaries. For the visualization of the chloroplast genome, OrganellarGenomeDraw (OGDRAW) [20,21] has been developed, which allows for the display of the genomic features and allows the user to create high quality circular and linear graphs. An important feature present in OGDRAW allows for the visualization of the expression data from the transcript profiling, polysome profiling or from the proteomics experiments. The software allows to display the transcriptional and translational status of the chloroplast encoded genes. Besides, web-based, OGDRAW is also available as a Perl module, which can be integrated into the annotation pipelines. CpGAVAS [22] and Verdant [14] provides inbuilt visualization of the annotated chloroplast genome using the OGDRAW. Colour sets in the OGDRAW for the clockwise and counter clockwise genes can be easily edited using the configuration file and java enabled OGDrawConfig [20].

2.4. Markers and codon usage

The main application of the chloroplast genome has been attributed to the development of the molecular markers primarily due to the conserved gene regions and the ease of the development of the polymorphic markers.

ChloroMitoSSRDB [23] and ChloroMitoSSRDB2.00 [24] is the first repository that has been developed for the large-scale visualization of the simple sequence repeats (SSRs) across the chloroplast genomes. The developed platform offers several features such

as the visualization of the distribution of repeat patterns using dynamic graphs, and the cross-linking of the identified repeats to the genic or non-genic regions. The developed platform also offers a comparative assessment of the two repeat mining algorithms IMEx [25,26] and MISA [27] and allows the repeat mining using the commonly used tools under one comparative framework. ChloroSSRdb [28] is a repeat mining framework, which is focused primarily on green plants.

Another aspect that made the chloroplast genomes distinct is the use of the chloroplast genomes for functional genomics by over-expressing the gene of interest. Chloroplast based plant functional factories has been widely exploited to develop over-expression of immunogenic vaccines. Recently developed ChloroMitoCU [29] offers a comparative assessment of the codon usage profiles across the chloroplast genomes. Currently, ChloroMitoCU contains 29,960 complete (full-length) protein-coding genes (CDSs) from all reference clades of chloroplasts genomes. The unique features of ChloroMitoCU involves the comparative assessment of the codon usage profiles across phylogenetic distant and related chloroplast genomes. Additionally, ChloroMitoCU allows for the comparative assessment of the codon usage patterns across the previously analysed chloroplast genome and the user submitted chloroplast genes.

2.5. Transcriptional profiling and RNA-editing

RNA polymerases and association of six sigma factors play a major role in maintaining the transcriptional based expression profiling [30]. Associative role of these polymerases and sigma factors provide important understanding of the role of splicing, gene editing and expression profiling of mutants in response to environmental stresses. Recently developed ChloroSeq [30] presents an optimized pipeline, which combines the spliced alignment tool tophat, bowtie, and bedtools, which have been developed previously for the genomic architecture visualization to process chloroplast RNA-seq expression profiles and allows for the estimation of the expression quantification across the exon and introns, splicing efficiency and the putative RNA Editing sites. RNA editing is a post-transcriptional process, which mainly involves the conversion of cytidine-to uridine and forms an important part of the RNA maturation process [31,32]. Across the plant lineage, evolutionary conserved RNA editing factors such as CRR28 and RARE1 have been widely shown to affect the cytodine-to-uridine conversion in sites mainly associated with ndh genes such as ndhBeU467PL, ndhDeU878SL and accDeU794SL [32]. PREPACT [33] has been widely used for the estimation of the RNA-Editing events from angiosperms like *Arabidopsis thaliana*, *Oryza sativa* to the early branching angiosperms such as *Amborella*. PREPACT allows the detection of the RNA Editing by comparative analysis across the 17 pre-implemented chloroplast transcriptomes and provides a user adjustable stringency threshold of 90 −70% for the detected RNA editing events. PREPACT operates in three distinct modes which allows for the prediction of the RNA-Editing events based on the alignment prediction, cDNA or BLASTx predictions. Recent implication of the RNA-Editing events allowed for the construction of the first synthetic operon in chloroplast genomes [34].

3. Conclusion

Chloroplast genomics is at the forefront of biology with the advent of the next generation sequencing, coupled with the advances in the assembling strategies applying novel seed-extend approaches to allow the assembly of the completely circularized genome. Recent advances have mainly focussed on the annotation and the comparative genomics of chloroplast genomes, thus allowing a better development of chloroplast genomes as plant factories.

Authors Contribution

GS conceived, desgined the research and wrote the MS; SBM contributed to the MS writing; DE and PJR provided the revisions and edits to the MS.

References

[1] John A. Raven, F. John, Allen Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome Biol. 4 (no. 3) (2003) 1.
[2] Indranil Basak, Simon Geir Møller, Emerging facets of plastid division regulation, Planta 237 (no. 2) (2013) 389–398.
[3] Poul Erik Jensen, Dario Leister, Chloroplast evolution, structure and functions, F1000Prime Rep. 6 (no. 40) (2014) 10–12703.
[4] Guillaume Martin, Franc-Christophe Baurens, Céline Cardi, Jean-Marc Aury, Angélique D'Hont, The complete chloroplast genome of banana (Musa acuminata, Zingiberales): insight into plastid monocotyledon evolution, PLoS One 8 (no. 6) (2013) e67350.
[5] H. Li, D. Li, S. Yang, J. Xie, J. Zhao, The state transition mechanism – simply depending on light-on and –off in Spirulina platensis, Biochim. Biophys. Acta 1757 (2006) 1512–1519.
[6] Chung-Shien Wu, Ya-Nan Wang, Chi-Yao Hsu, Ching-Ping Lin, Shu-Miaw Chaw, Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny, Genome Biol. Evol. 3 (2011) 1284–1295.
[7] ORGanelle ASseMbler Tool: http://pythonhosted.org/ORG.asm/ (Accessed 07 November 2016).
[8] Dierckxsens, Nicolas, Patrick Mardulyn, Guillaume Smits, NOVOPlasty: de novo assembly of organelle genomes from whole genome data, Nucleic Acids Res. (2016) kw955.
[9] Stacia K. Wyman, Robert K. Robert Jansen, Jeffrey L. Boore, Automatic annotation of organellar genomes with DOGMA, Bioinformatics 20 (no. 17) (2004) 3252–3255.
[10] Daisie I. Huang, Quentin C.B. Cronk, PLANN: a command-line application for annotating plastome sequences, Appl. Plant Sci. 3 (2015).
[11] Chang Liu, Linchun Shi, Yingjie Zhu, Haimei Chen, Jianhui Zhang, Xiaohan Lin, Xiaojun Guan, CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences, BMC Genomics 13 (no. 1) (2012) 715.
[12] Todd M. Lowe, Sean R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (no. 5) (1997) 955–964.
[13] Stefan Kurtz, The Vmatch large scale sequence analysis software, Ref Type: Computer Program (2003) 4–12.
[14] R. Michael McKain, H. Ryan Hartsock, M. Molly Wohl, A. Elizabeth Kellogg, Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes, Bioinformatics (2016) tw583.
[15] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (no. 14) (2002) 3059–3066.
[16] Alexandros Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 21 (no. 22) (2006) 2688–2690.
[17] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, Marco A. Marra, Circos: an information aesthetic for comparative genomics, Genome Res. 19 (no. 9) (2009) 1639–1645.
[18] E. Mitchell Skinner, V. Andrew Uzilov, D. Lincoln Stein, J. Christopher Mungall, H. Ian Holmes, JBrowse: a next-generation genome browser, Genome Res. 19 (no. 9) (2009) 1630–1638.
[19] Jinkui Cheng, Xu Zeng, Guomin Ren, Zhihua Liu, CGAP: a new comprehensive platform for the comparative analysis of chloroplast genomes, BMC Bioinformatics 14 (no. 1) (2013) 1.
[20] Marc Lohse, Oliver Drechsel, Ralph Bock, OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes, Curr. Genet. 52 (no. 5–6) (2007) 267–274.
[21] Marc Lohse, Oliver Drechsel, Sabine Kahlau, Ralph Bock, OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets, Nucleic Acids Res. (2013) gkt 289.
[22] Chang Liu, Linchun Shi, Yingjie Zhu, Haimei Chen, Jianhui Zhang, Xiaohan Lin, Xiaojun Guan, CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences, BMC Genomics 13 (no. 1) (2012) 715.
[23] Gaurav Sablok, Suresh B. Mudunuri, Sujan Patnana, Martina Popova, Mario A. Fares, Nicola La Porta, ChloroMitoSSRDB: open source repository of perfect

and imperfect repeats in organelle genomes for evolutionary genomics, DNA Res. 20 (no. 2) (2013) 127–133.
[24] Gaurav Sablok, G.V. Padma Raju, Suresh B. Mudunuri, Ratna Prabha, Dhananjaya P. Singh, Vesselin Baev, Galina Yahubyan, Peter J. Ralph, Nicola La Porta, ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection, Database (2015) (2015) bav084.
[25] Suresh B. Mudunuri, Hampapathalu A. Nagarajaram, IMEx: imperfect microsatellite extractor, Bioinformatics 23 (no. 10) (2007) 1181–1187.
[26] Suresh Babu Mudunuri, Pankaj Kumar, Allam Appa Rao, S. Pallamsetty, H.A. Nagarajaram, G-IMEx: a comprehensive software tool for detection of microsatellites from genome sequences, Bioinformation 5 (no. 5) (2010) 221–223.
[27] T. Thiel, MISA—Microsatellite identification tool, 2003. http://pgrc.ipk-gatersleben.de/misa/ (Acessed 07 November 2016).
[28] Aditi Kapil, Piyush Kant Rai, Asheesh Shanker, ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants, Database (2014) (2014) bau107.
[29] http://chloromitocu.cgu.edu.tw (Sablok et al. in press).
[30] Benoît Castandet, Amber M. Hotto, Susan R. Strickler, David B. Stern, ChloroSeq, an optimized chloroplast RNA-Seq bioinformatic pipeline, reveals remodeling of the organellar transcriptome under heat stress, G3: Genes| Genomes| Genetics 6 (no. 9) (2016) 2817–2827.
[31] Ting Chen, Zhoubin Li, Weiguo Zhu, Junhua Ge, Xiaoye Zheng, Xiaoping Pan, Hui Yan, Jianhua Zhu, MicroRNA-146a regulates the maturation process and pro-inflammatory cytokine secretion by targeting CD40L in oxLDL-stimulated dendritic cells, FEBS Lett. 585 (no. 3) (2011) 567–573.
[32] Anke Hein, Monika Polsakiewicz, Volker Knoop, Frequent chloroplast RNA editing in early-branching flowering plants: pilot studies on angiosperm-wide coexistence of editing sites and their nuclear specificity factors, BMC Evol. Biol. 16 (no. 1) (2016) 1.
[33] Henning Lenz, Volker Knoop, 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, Bioinf. Biol. Insights 7 (2013) 1.
[34] Elena Martin Avila, Martin F. Gisby, Anil Day, Seamless editing of the chloroplast genome in plants, BMC Plant Biol. 16 (no. 1) (2016) 168.
[35] GAP: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-95.
[36] OrganelleGenomeDraw:https://www.ncbi.nlm.nih.gov/pubmed/23609545.