Plant Physiology and Biochemistry 137 (2019) 53–61

Contents lists available at ScienceDirect

Plant Physiology and Biochemistry

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

Review

RNA editing in plants: A comprehensive survey of bioinformatics tools and databases

Claudio Lo Giudice[a], Irene Hernández[b], Luigi R. Ceci[a], Graziano Pesole[a,c], Ernesto Picardi[a,c,*]

[a] *IBIOM-CNR, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Italy*
[b] *Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009, Zaragoza, Spain*
[c] *Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari A. Moro, Bari, Italy*

ARTICLE INFO

Keywords:
RNA editing
C-To-U change
Organelle
Mitochondrion
Chloroplast
Database

ABSTRACT

RNA editing is a widespread epitranscriptomic mechanism by which primary RNAs are specifically modified through insertions/deletions or nucleotide substitutions.

In plants, RNA editing occurs in organelles (plastids and mitochondria), involves the cytosine to uridine modification (rarely uridine to cytosine) within protein-coding and non-protein-coding regions of RNAs and affects organelle biogenesis, adaptation to environmental changes and signal transduction.

High-throughput sequencing technologies have dramatically improved the detection of RNA editing sites at genomic scale. Consequently, different bioinformatics resources have been released to discovery and/or collect novel events.

Here, we review and describe the state-of-the-art bioinformatics tools devoted to the characterization of RNA editing in plant organelles with the aim to improve our knowledge about this fascinating but yet under investigated process.

1. Introduction

RNA editing is an important co/post transcriptional process that modifies primary RNAs through insertions/deletions or base substitutions. It has been observed in a wide range of organisms, including basal eukaryotes, fungi, land plants, vertebrates and viruses.

Molecular targets of RNA editing comprise mRNAs (translated/untranslated regions), introns, tRNAs, rRNAs, microRNAs and long noncoding RNAs (Licht and Jantsch, 2016; Picardi et al., 2014), located in the nucleus, in the cytoplasm as well as in energy-producing organelles (mitochondria and plastids).

RNA editing is widespread in almost all land plants, including liverworts, mosses, hornworts, lycopods, ferns and flowering plants, and affects only organellar transcripts. Strangely, no instance of RNA editing has yet been reported in green algae and, thus, it could have evolved during the transition from water to land (Tillich et al., 2006) with some loss as in the liverwort *Marchantia polymorpha*. However, the observed species specificity of RNA editing suggests that it could have followed multiple origins (Yura et al., 2009).

RNA editing in plant organelles involves mainly the deamination of cytidines into uridines (C-to-U) by specific pentatricopeptide repeat (PPR) proteins that are encoded in the nuclear genome (Takenaka et al., 2019). PPR proteins are mainly characterised by 30–40 amino acid repeated motifs that form a binding surface capable of sequence-specific recognition of RNA target molecules (Takenaka et al., 2019). As demonstrated by Cheng et al. (2016), deciphering the code through which these proteins find their targets, is one of the main potential approaches for finding new RNA editing sites.

Reverse U-to-C editing has also been observed even though it appeared prominent in hornworts, lycophytes and ferns (monilophytes) (Knie et al., 2016) and rare in higher plants.

The vast majority of plant organellar RNA editing occurs in protein coding regions (generally at the first or second position of codons) and leads to amino acid changes that appear to be conserved along the evolution, suggesting a consolidated biological function (Edera et al., 2018).

RNA editing in plants is commonly seen as a repair mechanism to correct genomic point mutations at RNA level and alter the substitutional rate that is extremely low in organellar genes. Moreover, since many genes (eg. *rps12*, *ndhD-2*, *rpoA*) exhibit different editing levels under different conditions (Phreaner et al., 1996; Okuda et al., 2007; Chateigner-Boutin et al., 2008), multiple protein isoforms with different characteristics (stable/unstable) or functions (active/inactive) could be generated at the same time, leading to fine regulatory mechanisms that

[*] Corresponding author. IBIOM-CNR, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Italy.
E-mail address: ernesto.picardi@uniba.it (E. Picardi).

https://doi.org/10.1016/j.plaphy.2019.02.001
Received 5 October 2018; Received in revised form 30 January 2019; Accepted 2 February 2019
Available online 02 February 2019

many plants may use to control the amount of active protein complexes and modify their composition.

Additionally, RNA editing occurs also in non-protein coding regions affecting splicing, transcript stability, translation efficiency and other important cellular and organellar processes. Interestingly, the RNA editing extent varies among different tissues, organs, developmental stages, mutant lines (Ichinose and Sugita, 2016) or environmental factors (Xiong et al., 2017) and, in general, its pattern changes between different plant species and between genes of the same species (Takenaka et al., 2013).

Several studies have demonstrated the involvement of RNA editing in various plant developmental processes, including organelle biogenesis (Sosso et al., 2012), adaptation to environmental changes (Yuan and Liu, 2012), and signal transduction (Tang et al., 2010).

RNA editing events can be easily identified by the concomitant sequencing of the transcript and its cognate genomic region. However, the advent of high-throughput sequencing (HTS) technologies has largely improved and facilitated the detection of RNA editing events at genomic scale. Indeed, the number of complete plants organellar genomes and related transcriptome data have considerably grown in the last decade. The explosion of genomic data, in turn, has led to the development of new bioinformatics tools for the accurate analysis and interpretation of such datasets.

Hereafter, we review main bioinformatics resources to investigate RNA editing in plant organelles, focusing on specialized databases, web services and stand-alone tools based on primary sequences or deep transcriptome (RNAseq) and genome sequencing datasets.

2. Bioinformatics approaches for studying RNA editing in plants

In order to provide a comprehensive overview of bioinformatics tools for detecting and cataloguing plant RNA editing events, we group computational resources in four main categories: databases, prediction tools, HTS-based tools and machine learning (ML) approaches. The complete list is shown in Table 1.

2.1. Databases

Databases are extremely useful resources because represent collections of specific biological objects but need to be continuously updated and revised.

GOBASE (O'Brien et al., 2003) and ChloroplastDB (Cui et al., 2006) can be considered the first relational databases containing information about RNA editing in plants, even though not specifically designed for this purpose. Soon after, many other resources collecting plants RNA editing events have been released. However, most of them have been dismissed or not updated (such as dbRES (He et al., 2007)) and, thus, we will focus only on active projects.

2.1.1. REDIdb

REDIdb (Lo Giudice et al., 2018) (Fig. 1A) was the first specialized database of plant RNA editing events to be released involving expert-based data curation. Its latest version includes 26,618 RNA editing sites distributed among 281 organisms, 85 complete organellar genomes and 3467 sequences.

RNA editing events stored in REDIdb have been extracted from GenBank and literature using a semi-automated procedure in which each annotation was manually checked for potential errors or inconsistencies. Thanks to its modular web interface, REDIdb searches can be performed by organism, location, gene or a combination of them, and results can be refined by activating optional filters (RefSeq, Exons, Full ORFs). Query outcomes are summarised in a tabular format which includes relevant information such as the GenBank accession number, the organism name, the link to an interactive taxonomy chart, the organelle type, the link to the complete genome, the gene name and a flag indicating its partial or full nature, the editing type and details, and the total number of events.

REDIdb events are organized in specific flat-files that comprise four main sections: 1) a header containing main record features (organism, GenBank accession, intracellular location, gene name, PubMed references and so on); 2) a gene box describing its ontologies; 3) a feature table summarizing all editing events per gene; 4) a sequence box including the genomic sequence and the corresponding edited transcript and/or protein.

REDIdb has been developed to facilitate the investigation of edited cDNA and protein sequences in their evolutionary context. Indeed, it implements a specific module to visualize multiple alignments of orthologue sequences (if any). Additionally, REDIdb allows the study of the RNA editing impact on protein coding genes because C-to-U or U-to-C changes are shown along the edited sequence containing the annotations of functional domains and predicted secondary protein structures. Finally, REDIdb allows the visualization of RNA editing events at genomic level by means of an *ad hoc* graphic module.

2.1.2. RESOPS

RESOPS (Yura et al., 2009) (Fig. 1B) is a database specialized in displaying plant organellar RNA editing sites on protein three-dimensional structures. C-to-U or U-to-C events derive from GenBank by means of in-house parsers followed by manual correction of errors and sequence discrepancies (employing literature or contacting entry submitters). The latest RESOPS release stores 5754 RNA editing sites and each query provides: 1) a link to unedited and edited cDNA sequences in GenBank-like style with notes about target nucleotides of RNA editing; 2) the conceptual translation of the edited cDNA in the pseudo-Uniprot format; 3) the multiple-alignment of homologous edited proteins; 4) the location of edited amino acids in the protein three-dimensional structure (if any). RESOPS uses different colours to mark edited amino acids that reside in the protein structural core or in others protein structural elements, providing precious functional information about the effect of specific RNA editing events at protein level.

For the visualization of three-dimensional structures, RESOPS relies on an old version of the Jmol library that may not be longer supported by modern browsers.

2.1.3. PED

To date different types of proteins have been identified as component of plant RNA editosome, including pentatricopeptide repeat (PPR) proteins, RNA editing factor interacting proteins (RIPs), multiple organellar RNA editing factors (MORFs), organelle RNA recognition motif (ORRM) proteins, organelle zinc finger (OZ) proteins, short DYW proteins and protoporphyrinogen oxidase 1 (PPO1) (Takenaka et al., 2019; Sun et al., 2016a).

Since an accurate study of RNA editing cannot be separated from these factors, Li et al. (2019) recently released a curated Plant Editosome Database (PED), containing both RNA editing events and their related proteins (Fig. 1C).

Current version of PED includes a total of 98 experimentally validated RNA editing factors and 20,836 editing events, covering 203 organelle genes and 1621 plant species.

All editing events stored in PED derive from 2651 organelle flatfiles downloaded from NCBI Organelle Genome Resources (https://www.ncbi.nlm.nih.gov/genome/organelle), parsed by mean of a custom python script and integrated with literature and manual curation.

For each specific editing factor PED summarizes all related editing events and provides a set of additional information such as the editing region, the editing type, the aminoacidic change (if any), the editing effect (recoding or synonymous) and supporting experimental evidences.

PED represents a good starting point for systematic investigations on the RNA editing machinery in a variety of plant species.

Table 1
List of available bioinformatics tools for investigating RNA editing in plant organelles.

Resource	Web link	Reference	Comment	[a] Required input	
DATABASES				**QUERY**	
RESOPS	http://cib.cf.ocha.ac.jp/RNAEDITING/	Yura, K. et al. (Yura et al., 2009)	Database useful to visualize RNA editing events in 3D protein structures. Requires an old Jmol library version.	● Clickable list of proteins	
REDIDB 3.0	http://srv00.recas.ba.infn.it/redidb/index.html	Lo Giudice, C. et al. (Lo Giudice et al., 2018)	Database suitable to query known RNA editing events in plants. It includes a comparative genomics module to study editing in its evolutionary context. Editing events derive from GenBank and are manually curated.	● Organism; ● Location; ● Gene; ● Additional filters (RefSeq, Exons, FullOrfs)	
PED	http://bigd.big.ac.cn/ped	Li M. et al. (Li et al., 2019)	Database useful to interrogate known RNA editing factors and events. It contains functional effects of editing factors in regulating plant phenotypes and includes detailed experimental evidence.	● Editing factor ● Edited gene ● Organism	
PREDICTION TOOLS				**INPUT**	**OPTIONS**
PREPACT 3.0	http://www.prepact.de/prepact-main.php	Lenz, H. et al. (Lenz et al., 2018)	Web-based tool for analysing, predicting and cataloguing plant-type RNA editing. It includes EdiFacts and TargetScan modules. EdiFacts integrates information on PPR proteins while TargetScan enables the identification of sequence motifs. May be useful to improve the RNA editing annotation status of novel organelle genomes.	● DNA for prediction (gene or organellar genome ● Protein reference(s) (BLASTX)	● Alignment prediction mode; ● cDNA analysis mode; ● BLASTX prediction mode; ● Additional filters (Forward editing CU, Reverse editing UC)
CURE-Chloroplast	http://bioinfo.au.tsinghua.edu.cn/software/pure/	Du, P. et al. (Du et al., 2009)	Web-based tool to predict RNA editing in chloroplast sequences. Although fast, it is outdated.	BASIC MODE ● (Single raw genomic sequence for prediction)	ADVANCED MODE ● Micro-analyser parameters; ● Blast parameters
PREP suite	http://prep.unl.edu/	Mower, J. P. (Mower, 2009)	Web-based tool to identify RNA editing in plant organelle sequences. It works only on coding sequences and employs conservation to detect potential events. Pre-computed alignments used to perform predictions may not be completely updated.	● (Prep-Aln) Protein-guided codon alignment in FASTA format ● (Prep-Mt and Prep-Cp) **Batch mode** Tab-delimited batch file containing (sequence names, gene names, codon position of the first nucleotide, prediction cutoff, DNA sequence). **Single sequence** Protein-guided codon alignment in FASTA format	● Codon position of the first nucleotide; ● Gene name; ● Prediction cutoff
HTS-BASED TOOLS				[a] **REQUIRED INPUT**	**MAIN OUTPUT**
REDItools	https://sourceforge.net/projects/reditools/	Picardi, E. and Pesole G (Picardi and Pesole, 2013)	A suite of scripts to investigate RNA editing in RNA-Seq experiments. The REDItoolDenovo.py script has been successfully used to profile RNA editing in mitochondria of *Vitis vinifera* (Picardi et al., 2010). It is portable and requires only the installation of the external pysam module.	● Multiple read alignments in BAM format; ● Reference genome in FASTA format.	Delimited tables containing putative RNA editing positions, coverage depth, mean quality score, observed base distribution, strand (if available) and a list of observed substitutions with their frequencies.
ChloroSeq	https://github.com/BenoitCastandet/chloroseq	Castandet, B. et al. (Castandet et al., 2016)	Is a pipeline to identify RNA editing in chloroplast genomes. Requires perl and external software such as Samtools and Bedtools. Bioinformatics skills are also required.	● Multiple read alignments in BAM format; ● Reference genome in FASTA format..	Delimited tables including gene expression profiles, splicing efficiency and RNA editing events.
RED	https://github.com/REDetector/RED	Sun, Y. et al. (Sun et al., 2016b)	Java software to detect and visualize RNA editing events at genomic scale using next-generation	● Multiple read alignments in BAM format[b]; ● List of RNA variants in VCF format; ● Reference genome in FASTA format.	Delimited tables including chromosome name, position, reference and alternative base, nucleotide

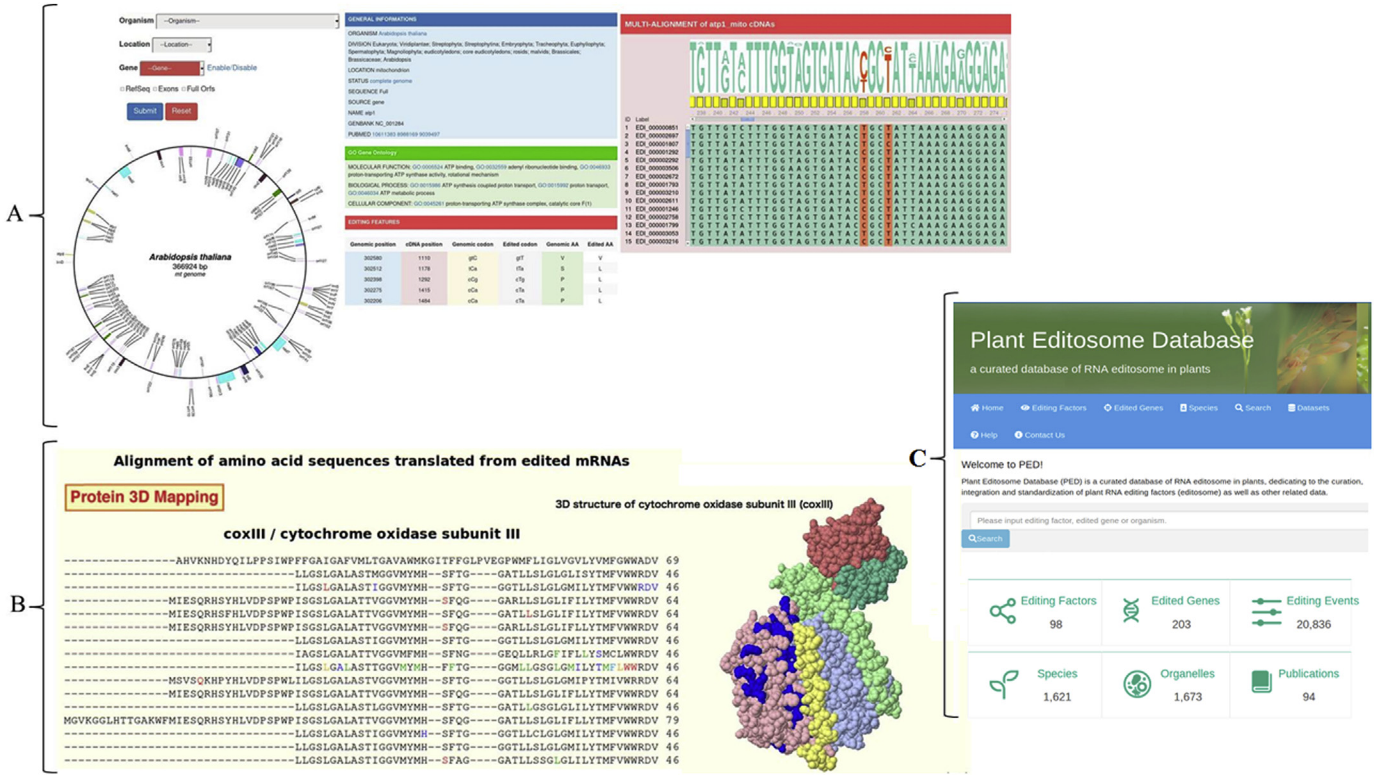
(continued on next page)

Table 1 (*continued*)

Resource	Web link	Reference	Comment	[a] Required input	
DATABASES				**QUERY**	
			sequencing data. Never applied to plant RNA editing. Requires bioinformatics skills.		quality, editing levels with their p-value and related FDR.
REDO	https://sourceforge.net/projects/redo/	Wu, S. et al. (Wu et al., 2018)	Is a comprehensive application tool for identifying RNA editing events in organelles based on variant call results files (VCF). Requires perl and R as well as bioinformatics skills. Can detect RNA editing events in multiple samples simultaneously.	● Multiple read alignments in BAM format; ● List of RNA variants in VCF format; ● Reference genome in FASTA format; a features table file in tabular format.	Delimited tables including the genomic position of editing candidates as well as all relevant related annotations (derived from the feature table file).
RES-Scanner	https://github.com/ZhangLabSZ/RES-Scanner	Wang, Z. et al. (Wang et al., 2016)	Software package for genome-wide identification and annotation of RNA-editing sites for any species with matched RNA-Seq and DNA-Seq data. Requires perl, BWA and bioinformatics skills. Never applied to plant RNA editing	● Unaligned single or paired-end Illumina transcriptome reads in FASTQ format: ● Multiple read alignments in BAM format; ● Reference genome in FASTA format; ● Feature table file in tabular format.	Delimited tables including the identified RNA-editing sites with a variety of genomic features, such as exon, intron, coding sequence, etc. Codon and amino acid change are also reported for the editing events targeting CDSs.
ML approaches					
ML strategy	N/A	Cummings, M. P. and D. S. Myers (Cummings and Myers, 2004)	*Ab initio* method useful in the first steps of genome annotation to have look at the RNA editing potential. Requires a set of known events to train the algorithm and very good bioinformatics skills.	Algorithm training set containing known edited sites vs a null set of non-edited sites.	
ML strategy (REGAL)	N/A	Thompson, J. and S. Gopal (Thompson and Gopal, 2006)	*Ab initio* method useful in the first steps of genome annotation to have look at the RNA editing potential. Requires a set of known events to train the algorithm and very good bioinformatics skills.	Algorithm training set containing known edited sites vs a null set of non-edited sites.	
ML strategy	N/A	Du, P. et al. (Du et al., 2007)	*Ab initio* method useful in the first steps of genome annotation to have look at the RNA editing potential. Requires a set of known events to train the algorithm and very good bioinformatics skills.	Algorithm training set containing known edited sites vs a null set of non-edited sites.	

[a] Minimum input required by the application to produce an output. Depending on the application, accessory files can be provided for specific outputs.

[b] HTS reads in BAM format are mandatory for the graphical representation of the edited sites.



Alignment of amino acid sequences translated from edited mRNAs

coxIII / cytochrome oxidase subunit III

```
---------------AHVKNHDYQILPPSIWPFFGAIGAFVMLTGAVAWMKGITFFGLPVEGPWMFLIGLVGVLYVMFGWWADV 69
-------------------------------LLGSLGALASTMGGVMYMH--SFTG----GATLLSLGLGLISYTMFVWWRDV 46
-------------------------------ILGSLGALASTIGGVMYMH--SFTG----GGTLLSLGLGMILYTMFVWWRDV 46
--------------MIESQRHSYHLVDPSPWPISGSLGALATTVGGVMYMH--SFQG----GARLLSLGLIFILYTMFVWWRDV 64
--------------MIESQRHSFHLVDPSPWPISGSLGALATTVGGVMYMH--SFQG----GATLLSLGLIFILYTMFVWWRDV 64
--------------MIESQRHSYHLVDPSPWPISGSLGALATTVGGVMYMH--SFQG----GARLLSLGLIFLLYTMFVWWRDV 64
-------------------------------LLGSLGALASTIGGVMYMH--SFTG----GGTLLSLGLGMILYTMFVWWRDV 46
-------------------------------IAGSLGALATTVGGVMFMH--SFNG----GEQLLRLGFIFLLYSMCLWWRDV 46
-------------------------------ILGSLGALASTTGGVMYMH--FFTG----GGMLLGSGLGMILYTMFLWWRDV 46
--------------MSVSQKHPYHLVDPSPWLILGSLGALASTVGGVMYMH--SFTG----GGTLLSLGLGMIPYTMIVWRRDV 64
--------------MIESQRHSYHLVDPSPWPISGSLGALATTVGGVMYMH--SFQG----GATLLSLGLIFLLYTMFVWWRDV 64
-------------------------------LLGSLGALASTLGGVMYMH--SFTG----GATLLGSGLGLILYTMFVWWRDV 46
MGVKGGLHTTGAKWFMIESQRHSYHLVDPSPWPISGSLGALATTVGGVMYMH--SFQG----GATLLSLGLIFLLYTMFVWWRDV 79
-------------------------------LLGSLGALASTIGGVMYMH--SFTG----GGTLLCLGLGMILYTMFVWWRDV 46
-------------------------------LLGSLGALASTIGGVMYMH--SFTG----GGTLLSLGLGMILYTMFVWWRDV 46
-------------------------------LLGSLGALASTIGGVMYMH--SFAG----GATLLSSGLGLILYTMFVWWRDV 46
```

Welcome to PED!

Plant Editosome Database (PED) is a curated database of RNA editosome in plants, dedicating to the curation, integration and standardization of plant RNA editing factors (editosome) as well as other related data.

Please input editing factor, edited gene or organism.

Fig. 1. Screen shots adapted from **(A)** REDIdb **(B)** RESOPS and **(C)** PED databases.

2.1.4. Additional resources

Recently, Edera et al. (2018) published a work aimed to provide a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria using publicly available RNAseq data.

Although results have not been released in a specific database, they are freely available as supplementary material and contain 10,217 editing sites from 17 mitochondrial genomes of five angiosperms lineages: magnoliids, monocots, basal eudicots, rosids, and asterids.

2.2. Prediction tools

RNA editing in plants tends to increase the complexity of the organellar transcriptome and proteome. Indeed, the majority of known events are non-synonymous, leading to different protein isoforms whose functions are yet undetermined. Detecting plant RNA editing is therefore extremely useful and important. To this aim, several bioinformatics tools have been specifically developed. Below we describe main computational resources starting from online prediction tools.

2.2.1. PREPACT

The Plant RNA Editing Prediction and Analysis Computer Tool (PREPACT) (Lenz et al., 2018) is a web resource which allows prediction, analysis, annotation and visualization of C-to-U or U-to-C editing events in plants.

RNA editing events are reported using the standardized nomenclature proposed by Rüdinger et al., in 2009 (Rudinger et al., 2009). According to this convention, each editing event is indicated with an "e" (for editing) preceded by gene name and followed by "U or C" for the respective nucleotide created by RNA editing, the nucleotide position in the reading frame and the induced codon change (if any) using the single-letter amino acid code. As an example, "atp9eU56 PL" indicates a C-to-U RNA editing of the *atp9* mRNA at position 56 with the consequential change of a genomically encoded proline codon into a leucine codon on mRNA level.

The web interface of PREPACT allows the RNA editing detection in three operative ways: 1) editing site prediction in DNA sequences; 2) editing site investigation using cDNA sequences; 3) editing site identification using BLASTX searches against known protein databases.

The prediction of RNA editing in DNA sequences is based on the comparison between a reference DNA (comprising the coding sequence of a non-edited gene) and a target DNA (comprising the coding sequence of the gene where editing should be predicted). In particular, the target sequence is split in triplets and all C-to-U or U-to-C changes are done in order to identify base modifications restoring the amino acid identities encoded by the reference sequence. Such changes are returned as potential editing candidates.

The identification of editing events using cDNA sequences employs the same strategy but with the important restriction that a codon is considered as edited only if exactly matching its counterpart on the reference sequence.

BLASTX prediction, instead, is specifically conceived for the analysis of novel organelle genome sequences. It doesn't require any prior knowledge on the query sequence and can be done against several set of proteins encoded by selectable organelle genomes. Target sequences for comparison include the chloroplast and mitochondrial proteomes from more than 20 species.

The latest version, PREPACT 3.0, includes new features, such as the possibility to restrict searches only to U-to-C editing sites or limit candidate editing sites to positions conserved in at least one orthologous. PREPACT 3.0 introduced also the "EdiFacts" and "TargetScan" modules. EdiFacts is a local database containing information on pentatricopeptide repeat (PPR) proteins, a large family of modular RNA-*proteins* characterised by tandem *repeats* of a degenerate 35 amino acid motifs (Small and Peeters, 2000). TargetScan, instead, allows position-weighted searches for sequence motifs in the PREPACT's reference sequences.

PREPACT service accepts as input DNA/cDNA sequences in FASTA format or GenBank accession numbers. As output, PREPACT prints out lists of RNA editing candidates as well as multiple sequence alignments. PREPACT allows user-defined colour schemes for highlighting editing events and binary matrices of edited positions that can be used for phylogenetic analyses.

2.2.2. CURE-chloroplast

CURE-Chloroplast (Du et al., 2009) is an online service devoted to the prediction of C-to-U RNA editing sites in chloroplast genes of seed plants. Its algorithm is an extension of the CURE method, originally used to predict RNA editing in plant mitochondria (Du and Li, 2008), and achieves over 80% sensitivity and over 99% specificity.

CURE-Chloroplast is based on known RNA editing events stored in the REDIdb database. Indeed, all C-to-U chloroplast RNA editing sites of seed plants are retrieved from REDIdb and orthologous edited genes or whole genomes are aligned by ClustalW (Larkin et al., 2007) or TBA (Blanchette et al., 2004), respectively. Such alignments are employed to train the algorithm collecting the so called Evolutionary Potential Editing Sites (EPES), defined as columns of multiple sequence alignment containing RNA editing events. EPES are analysed to extract three sequence descriptors (based on biochemical and evolutionary information) and saved in a local database. Every time an input sequence is provided, the prediction algorithm uses BLAST to map all EPES consensus sequences against this sequence, checking for the editing status of each cytidine. CURE-Chloroplast can predict RNA editing in single genes or whole chloroplast genomes and the web interface can accept single or multi-fasta sequences.

One of the major limitations of CURE-Chloroplast is that the training dataset does not contain information about RNA editing events in non-coding regions. Moreover, due to its content, the performance in non-seed plants is less accurate.

2.2.3. The PREP suite

The PREP (predictive RNA editors for plants) suite (Mower, 2009) is a collection of web resources specifically designed for the fast and accurate prediction of RNA editing sites in plant organellar genes. It offers predictive tools for plant mitochondrial genes (PREP-Mt), chloroplast genes (PREP-Cp) and custom alignments submitted by the user (PREP-Aln). The basic idea behind PREP SUITE is that RNA editing in plant organelles tends to increase the conservation of proteins across species. According to this idea, PREP-Mt and PREP-Cp translate and align an input sequence to a pre-calculated alignment of mitochondrial or plastidial homologs.

Then, the resulting multi-alignment is screened column-by-column to identify potential RNA editing events (C-to-U changes) that could increase the similarity of the input sequence to the sequences in the pre-defined multi-alignment.

For each editing candidate, a conservation score, representing a raw indicator of the prediction confidence is provided in output.

In all cases in which pre-defined alignments are not satisfactory, PREP-Mt and PREP-Cp tools can be applied to a custom submitted alignment through the PREP-Aln module.

2.3. HTS-based tools

RNA editing events can be easily discovered comparing cDNA sequences and their corresponding genomic *loci*. Such naïve approach is feasible at genomic scale by high-throughput sequencing (HTS) technologies, paying attention to sequencing or read-mapping errors and false predictions due to genome-encoded single nucleotide polymorphisms (SNPs) (Picardi et al., 2012). Nowadays, several bioinformatics tools devoted to the identification of RNA editing events in deep transcriptome datasets (RNA-Seq) have been released. However, only ChloroSeq and REDItools have been developed to take into account C-to-U and U-to-C changes in plant organelles. Indeed, the majority of HTS-based tools for RNA editing have been conceived to discover A-to-I events in humans. Although they could be adapted to plant RNA editing, their suitability has never been tested.

2.3.1. REDItools

REDItools (Picardi and Pesole, 2013) (Supplementary Fig. S1) is a suite of python scripts to perform genome wide investigations of RNA editing using HTS data. They are organism and editing type independent and comprise three main scripts, REDItoolDnaRna.py, REDItoolKnown.py and REDItoolDenovo.py as well as several utilities for pre- and post-data processing.

REDItoolDnaRna.py identifies RNA editing events by comparing multiple read alignments of RNA-Seq and DNA-Seq data (WGS or WXS) from the same sample. It explores a complete genome positions by position and looks at RNA variants supported by a minimum number of reads (generally ≥ 10) that are homozygous at DNA level. To mitigate the effect of false predictions, REDItoolDnaRna.py implements a variety of user tuneable filters.

Differently from REDItoolDnaRna.py, REDItoolKnown.py is specialized in the extraction of known RNA editing events from RNA-Seq experiments and does not require DNA-Seq data.

Finally, REDItoolDenovo.py is specific for finding potential RNA editing events in RNA-Seq data alone without any a priori knowledge of the RNA editing type. Its algorithm calculates the distribution of expected/observed bases at all covered genomic positions and uses the Fisher exact test to provide a p-value per site (corrected by Bonferroni or Benjamini-Hochberg). REDItoolDenovo.py has been successfully applied to unveil the RNA editing landscape of the grapevine mitochondrial genome (Picardi et al., 2010).

REDItools scripts requires multiple read alignments in BAM format (it is the binary version of SAM format released in output from many aligners) and return in output tab delimited tables including the putative RNA editing positions, the coverage depth, the mean quality score, the observed base distribution, the strand (if available) and a list of observed substitutions with their frequencies. Optionally, REDItools scripts can filter out sites in homopolymeric regions of predefined length or close to splice sites and can exclude multi-mapping reads or PCR duplicates. In addition, REDItools can work on RNA-Seq from strand-oriented libraries to mitigate biases introduced by antisense transcription or mapping errors.

2.3.2. ChloroSeq

ChloroSeq (Castandet et al., 2016) is a bioinformatic pipeline to systematically analyse plastid transcriptomes using RNA-Seq data. It comprises command line Perl scripts that can be implemented on a variety of modern computers. Like REDItools, ChloroSeq requires multiple read alignments in the standard binary format (BAM) and begins the analysis extracting and indexing reads mapping onto the plastid genome. Although the main output consists in count tables useful for gene expression analyses, ChloroSeq can examine splicing efficiency and RNA editing profiles through well-established external software such as SAMtools (the first package for manipulating sequence read alignments in SAM and BAM format) and BEDtools (a collection of scripts for comparing large datasets of genomic features in BED format).

RNA editing detection in ChloroSeq is performed by the get_editing_efficiency.sh script in which SAMtools are invoked to create a pileup file (a multiple sequence alignment from a group of related sequences) containing the putative variants. Differently from REDItoolDenovo.py, ChloroSeq does not assign a p-value to RNA editing candidates and a very limited repertoire of filters is available.

Many ChloroSeq functionalities depend on the availability of organellar genome annotations that are commonly extracted from GenBank. Since such annotations are error prone is mandatory a pre-check before whatever ChloroSeq analysis.

Although initially conceived for plastids, ChloroSeq has been successfully applied to plant and algal mitochondrial transcriptomes (Castandet et al., 2016), making this tool useful to investigate transcriptional properties of plant organellar genomes.

2.3.3. RED

RED (RNA Editing Site detector) (Sun et al., 2016b) is a platform-independent Java-based tool designed for the identification of RNA editing events at genome scale.

RED core algorithm uses rule-based and statistical filters to remove spurious RNA editing sites. It also provides a graphical user interface (GUI) to visualize edited positions in their genomic context. RNA editing sites can be detected using the *de novo* mode or the *DNA-RNA* mode if DNA sequencing data are available (Supplementary Fig. S2). All HTS reads need to be in BAM format and are mandatory for the graphical representation of edited sites. In addition, RED requires a pre-calculated VCF (Variant Calling Format) file containing the list of all RNA variants and accessory files for filtering purposes such as a repeat region masked file generated by RepeatMasker, a gene annotation file in GTF format, a VCF file containing known SNPs and a file containing known RNA editing sites from multiple databases (Kiran and Baranov, 2010; Ramaswami and Li, 2014).

RNA editing candidates are detected by filtering the list of pre-calculated RNA variants (generally by using GATK (McKenna et al., 2010)) and stored in an internal MySQL database (as well as positions at each filtering step) to speed up queries and facilitate data integration. Output images generated by the RED GUI can be exported in Scalable Vector Graphics (SVG) or Portable Network Graphics (PNG) formats, while potential RNA editing candidates are saved in a tab-delimited text file that includes the chromosome name, the position, the reference and alternative base, the nucleotide quality, the editing level, the p-value and its related False Discovery Rate (FDR). Differently from similar software, RED installation requires multiple environment dependencies (jre or jdk 1.6.0_43 or later, MySQL 5.1.73 or later, R 3.0.1 or later) and has never been used to detect RNA editing in plants. So, potential results from plant organelles need careful inspection and further experimental validations.

2.3.4. REDO

REDO (Wu et al., 2018) is a bioinformatic tool designed for the identification of RNA editing events in plant organelles from HTS data. Like RED, it requires a VCF file containing all genomic positions supported by transcriptome data, the reference genome in FASTA format and gene annotations (Supplementary Fig. S3).

RNA variants representing RNA editing candidates are filtered according to several parameters (as read quality or sequencing depth) in order to exclude potential false positives.

REDO incorporates also two statistical tests, the likelihood ratio (*LLR*) test (Benjamini et al., 2001) and the Fisher's exact test to provide a p-value for each detected RNA editing site.

Output results are printed out in tabular format and include the genomic position of editing candidates as well as all relevant annotations.

A peculiar feature of REDO relies in its ability to process multiple samples simultaneously.

Similar to RED, RNA editing candidates are detected by filtering a list of pre-calculated RNA variants in VCF format. This greatly speeds up the analysis compared to tools such as REDItools in which the variant calling is carried out directly by the program.

2.3.5. RES-scanner

RNA editing site scanner (RES-scanner) (Wang et al., 2016) is a Perl tool devoted to genome-wide identification and annotation of RNA-editing sites. It integrates transcriptome mapping, homozygous genotype calling, *de novo* RNA-editing site identification and annotation for any species provided with matching RNA-seq and DNA-seq data (Supplementary Fig. S4). RES-scanner accepts single or paired-end Illumina transcriptome reads in FASTQ format and supports both stranded and unstranded RNA-seq data.

RES-scanner embeds the alignment tool BWA for mapping transcriptome reads against a reference genome and a combination of exonic sequences surrounding splicing junctions. Nonetheless, it can accept pre-aligned RNA and DNA reads from other mappers (eg. Bowtie 2 (Langmead and Salzberg), TopHat2 (Kim et al., 2013), GSNAP (Wu et al., 2016), HISAT2 (Kim et al., 2015)).

Bioinformatics tools similar to RES-scanner detect candidate RNA-editing events separating them from genome-encoded variants by applying arbitrary thresholds on the frequency of the alternative allele recovered from DNA-Seq data. This approach may lead to an underestimation of heterozygous sites, especially in case of low sequencing depth or not diploid genomes. RES-scanner overcomes this limitation introducing statistical models (Bayesian and Binomial) to infer the reliability of homozygous genotypes.

Like other tools RNA editing sites can be discriminated from sequencing errors by assigning a p-value to each candidate. RES-Scanner can be used as a standalone program from raw reads up to final editing sites and appears useful in non-model species. However, no application has been done in plant organelles and, thus, same recommendations as for similar tools are still valid.

2.4. Machine learning-based approaches

Machine learning (ML) is an emerging field of computer science that uses statistical techniques to give computer systems the ability to improve their performances on a specific task, without being specifically programmed (Samuel, 1959). The process strongly depends on initial training datasets, which are used by consolidated learning algorithms to operate decisions on different data without human intervention.

The first attempts at using machine learning to detect RNA editing events in plants was published in 2004 by Cummings and Mayers (Cummings and Myers, 2004). They proposed two *ab initio* methods to predict C-to-U RNA editing sites using a tree-based statistical model (Kingsford and Salzberg, 2008) and random forests (Breiman, 2001).

Tree-based statistical models are predictive models that are constructed by a branching series of boolean tests on a set of known training data. They are then applied to classify previously unseen examples and, if trained on high-quality data, they are capable of making very accurate predictions. A random forest, instead, attempts to improve the concept of a single tree-based statistical model by generating a collection of trees and using them in aggregate.

The philosophy behind the random forest is that since the random trees could have some overlap, data with the various trees can be analysed redundantly looking for trends and patterns that better support a given outcome, thus providing more sophisticated analyses.

Cummings and Mayers models require a non-random distribution of nucleotides close to known edited sites (20 nucleotides upstream and downstream) for training purposes and a null-set of non-edited sites derived from known complete mitochondrial genomes.

By applying both models on the same test dataset, Cummings and Mayers found better performances (in terms of accuracy and sensitivity) for random forests than the tree-based statistics.

Soon after Cummings and Mayers, Thompson and Gopal released a second *ab initio* approach, named REGAL (Thompson and Gopal, 2006), to identify C-to-U editing sites in plant mitochondrial genomes.

REGAL is based on a genetic algorithm (Mitchell, 1998) (a heuristic search strategy that reflects the process of natural selection where the fittest individuals are selected for reproduction at each generation) trained on a subset of known editing sites derived from *Arabidopsis thaliana* and tested against the mitochondrial genomes of *A. thaliana, Brassica napus* and *Oryza sativa*.

REGAL predictions have a comparable sensitivity and higher specificity than approaches relying on sequence similarity (e.g. Prep-Mt), without the need of specific conservation thresholds to detect reliable predictions of edited sites.

Another interesting approach for RNA editing site identification by mean of ML algorithms has been proposed by Du et al. (2007). Such approach is based on a Support Vector Machine (SVM) algorithm

(Vapnik, 1995) combined with a triplet scoring model. SVM is used to identify regions flanking C-to-U editing sites, while the triplet scoring model (an extended version of the commonly used positional weighted matrix model) is employed to improve the prediction and provide a detection score. The algorithm was trained on a RNA editing dataset from *A. thaliana*, *B. napus* and *O. sativa* mitochondria. A null set of non-edited sites was derived from the same mitochondrial genomes by applying the same strategies used by both REGAL (Thompson and Gopal, 2006) and the classification tree based model by Cummings and Meyers (Cummings and Myers, 2004).

Using complete mitochondrial genome from *A. thaliana*, *B. napus* and *O. sativa*, for which RNA editing events have been experimentally validated, the SVM based method achieved better performance than classification three based methods and REGAL. Similar results were found comparing the SVM based method and PREP-Mt which is based on protein sequence similarity.

3. Conclusions

Detecting RNA editing events at genomic scale by using computational tools is yet a challenging task even though several bioinformatics resources have been developed. Some of them are based on sequence properties and employs machine learning technics. Although accurate, such tools do not capture the entire RNA editing landscape of an organellar genome since limited by reliable training datasets. Similarity-based resources, instead, provide RNA editing predictions with high specificity but are biased towards events in coding protein genes.

HTS technologies allow transcriptome investigations at single nucleotide resolution and enable genome-wide identification of RNA-editing sites with high reliability. Indeed, they are not limited to protein coding genes or depends on training datasets.

Choosing the correct strategy to investigate plant organellar RNA editing is not trivial. For this reason, our review comes with the aim to facilitate that choice, providing an overview of the state-of-the-art bioinformatics resources devoted to solving this task in order to better understand functional and physiological roles of RNA editing in plants.

Author contribution

Claudio Lo Giudice: performed bibliographic searches, tested bioinformatics tools and drafted the manuscript;

Irene Hernández: carried out bibliographic searches and helped in testing bioinformatics tools;

Luigi R. Ceci: revised the manuscript and tested databases;

Graziano Pesole: revised the manuscript and helped with HTS based methods;

Ernesto Picardi: supervised the work and revised the manuscript.

Acknowledgements

We kindly thanks Luigi Mansi for critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.plaphy.2019.02.001.

Funding

This work was supported by ELIXIR IIB (CNR), the Italian Infrastructure for Bioinformatics.

References

Benjamini, Y., et al., 2001. Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. 125 (1–2), 279–284.

Blanchette, M., et al., 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14 (4), 708–715.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Castandet, B., et al., 2016. ChloroSeq, an optimized chloroplast RNA-seq bioinformatic pipeline, reveals remodeling of the organellar transcriptome under heat stress. G3 (Bethesda) 6 (9), 2817–2827.

Chateigner-Boutin, A.L., et al., 2008. CLB19, a pentatricopeptide repeat protein required for editing of rpoA and clpP chloroplast transcripts. Plant J. 56 (4), 590–602.

Cheng, S., et al., 2016. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. Plant J. 85 (4), 532–547.

Cui, L., et al., 2006. ChloroplastDB: the chloroplast genome database. Nucleic Acids Res. 34 (Database issue), D692–D696.

Cummings, M.P., Myers, D.S., 2004. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. BMC Bioinf. 5 (1), 132.

Du, P., Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J. Theor. Biol. 253 (3), 579–586.

Du, P., He, T., Li, Y., 2007. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. Biochem. Biophys. Res. Commun. 358 (1), 336–341.

Du, P., Jia, L., Li, Y., 2009. CURE-Chloroplast: a chloroplast C-to-U RNA editing predictor for seed plants. BMC Bioinf. 10, 135.

Edera, A.A., Gandini, C.L., Sanchez-Puerta, M.V., 2018. Towards a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria. Plant Mol. Biol. 97 (3), 215–231.

He, T., Du, P., Li, Y., 2007. dbRES: a web-oriented database for annotated RNA editing sites. Nucleic Acids Res. 35 (Database issue), D141–D144.

Ichinose, M., Sugita, M., 2016. RNA editing and its molecular mechanism in plant organelles. Genes 8 (1).

Kim, D., et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14 (4), R36.

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12 (4), 357–360.

Kingsford, C., Salzberg, S.L., 2008. What are decision trees? Nat. Biotechnol. 26 (9), 1011–1013.

Kiran, A., Baranov, P.V., 2010. DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics 26 (14), 1772–1776.

Knie, N., et al., 2016. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns - a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. BMC Evol. Biol. 16 (1), 134.

Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9(4): p. 357-359.

Larkin, M.A., et al., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23 (21), 2947–2948.

Lenz, H., Hein, A., Knoop, V., 2018. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0. BMC Bioinf. 19 (1), 255.

Li, M., et al., 2019. Plant editosome database: a curated database of RNA editosome in plants. Nucleic Acids Res. 47 (D1), D170–D174.

Licht, K., Jantsch, M.F., 2016. Rapid and dynamic transcriptome regulation by RNA editing and RNA modifications. J. Cell Biol. 213 (1), 15–22.

Lo Giudice, C., Pesole, G., Picardi, E., 2018. REDIdb 3.0: a comprehensive collection of RNA editing events in plant organellar genomes. Front. Plant Sci. 9, 482.

McKenna, A., et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (9), 1297–1303.

Mitchell, M., 1998. An Introduction to Genetic Algorithms. MIT Press, pp. 209.

Mower, J.P., 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 37 (Web Server issue), W253–W259.

O'Brien, E.A., et al., 2003. GOBASE–a database of mitochondrial and chloroplast information. Nucleic Acids Res. 31 (1), 176–178.

Okuda, K., et al., 2007. Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. Proc. Natl. Acad. Sci. U. S. A. 104 (19), 8178–8183.

Phreaner, C.G., Williams, M.A., Mulligan, R.M., 1996. Incomplete editing of rps12 transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. Plant Cell 8 (1), 107–117.

Picardi, E., Pesole, G., 2013. REDItools: high-throughput RNA editing detection made easy. Bioinformatics 29 (14), 1813–1814.

Picardi, E., et al., 2010. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. Nucleic Acids Res. 38 (14), 4755–4767.

Picardi, E., et al., 2012. A Novel Computational Strategy to Identify A-to-I RNA Editing Sites by RNA-Seq Data: de Novo Detection in Human Spinal Cord Tissue. PLoS One 7 (9), e44184.

Picardi, E., et al., 2014. Uncovering RNA editing sites in long non-coding RNAs. Front Bioeng Biotechnol 2, 64.

Ramaswami, G., Li, J.B., 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 42 (Database issue), D109–D113.

Rudinger, M., et al., 2009. RNA editing: only eleven sites are present in the Physcomitrella patens mitochondrial transcriptome and a universal nomenclature proposal. Mol. Genet. Genom. 281 (5), 473–481.

Samuel, A.L., 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3 (3), 210–229.

Small, I.D., Peeters, N., 2000. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem. Sci. 25 (2), 46–47.

Sosso, D., et al., 2012. PPR2263, a DYW-Subgroup Pentatricopeptide repeat protein, is required for mitochondrial nad5 and cob transcript editing, mitochondrion biogenesis, and maize growth. Plant Cell 24 (2), 676–691.

Sun, T., Bentolila, S., Hanson, M.R., 2016a. The unexpected diversity of plant organelle RNA editosomes. Trends Plant Sci. 21 (11), 962–973.

Sun, Y., et al., 2016b. RED: a java-MySQL software for identifying and visualizing RNA editing sites using rule-based and statistical filters. PLoS One 11 (3), e0150465.

Takenaka, M., et al., 2013. RNA editing in plants and its evolution. Annu. Rev. Genet. 47, 335–352.

Takenaka, M., et al., 2019. RNA editing mutants as surrogates for mitochondrial SNP mutants. Plant Physiol. Biochem. 135, 310–321.

Tang, J., et al., 2010. The mitochondrial PPR protein LOVASTATIN INSENSITIVE 1 plays regulatory roles in cytosolic and plastidial isoprenoid biosynthesis through RNA editing. Plant J. 61 (3), 456–466.

Thompson, J., Gopal, S., 2006. Genetic algorithm learning as a robust approach to RNA editing site prediction. BMC Bioinf. 7, 145.

Tillich, M., et al., 2006. The evolution of chloroplast RNA editing. Mol. Biol. Evol. 1912–1921. https://doi.org/10.1093/molbev/msl054. 23(10 %M).

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, pp. 188.

Wang, Z., et al., 2016. RES-Scanner: a software package for genome-wide identification of RNA-editing sites. GigaScience 5 (1), 37.

Wu, T.D., et al., 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol. Biol. 1418, 283–334.

Wu, S., et al., 2018. REDO: RNA editing detection in plant organelles based on variant calling results. J. Comput. Biol. 25 (5), 509–516.

Xiong, J., et al., 2017. RNA editing responses to oxidative stress between a wild abortive type male-sterile line and its maintainer line. Front. Plant Sci. 8, 2023.

Yuan, H., Liu, D., 2012. Functional disruption of the pentatricopeptide protein SLG1 affects mitochondrial RNA editing, plant development, and responses to abiotic stresses in Arabidopsis. Plant J. 70 (3), 432–444.

Yura, K., et al., 2009. RESOPS: a database for analyzing the correspondence of RNA editing sites to protein three-dimensional structures. Plant Cell Physiol. 50 (11), 1865–1873.