Genome analysis

IRscope: an online program to visualize the junction sites of chloroplast genomes

Ali Amiryousefi[1,*], Jaakko Hyvönen[1,2] and Peter Poczai[2]

[1]Organismal Evolutionary Biology Research Program, Faculty of Biology and Environmental Sciences, Viikki Plant Science Centre and [2]Finnish Museum of Natural History, University of Helsinki, FI-00014 Helsinki, Finland

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on November 2, 2017; revised on March 16, 2018; editorial decision on April 3, 2018; accepted on April 4, 2018

Abstract

Motivation: Genome plotting is performed using a wide range of visualizations tools each with emphasis on a different informative dimension of the genome. These tools can provide a deeper in-sight into the genomic structure of the organism.

Results: Here, we announce a new visualization tool that is specifically designed for chloroplast genomes. It allows the users to depict the genetic architecture of up to ten chloroplast genomes in the vicinity of the sites connecting the inverted repeats to the short and long single copy regions. The software and its dependent libraries are fully coded in R and the reflected plot is scaled up to realistic size of nucleotide base pairs in the vicinity of the junction sites. We introduce a website for easier use of the program and R source code of the software to be used in case of preferences to be changed and integrated into personal pipelines. The input of the program is an annotation GenBank (.gb) file, the accession or GI number of the sequence or a DOGMA output file. The soft-ware was tested using over a 100 embryophyte chloroplast genomes and in all cases a reliable out-put was obtained.

Availability and implementation: Source codes and the online suit available at https://irscope.shinyapps.io/irapp/ or https://github.com/Limpfrog/irscope. Contact: ali.amiryousefi@helsinki.fi

1 Introduction

Considering the availability of genome sequences, a class of tools are developed to assist making valid hypotheses about homology be-tween different lineages. Visualizing methods as a subclass of these tools help us to unravel the structural composition of the genomes from synteny analysis (Lyons and Freeling, 2008) to comparative genome plots (Frazer et al., 2004). While many of these are applic-able to organellar genomes the very structure of these genomes calls for more specialized packages to reflect their unique features (e.g. Lohse et al., 2007). Due to the high level of conservation and com-pactness of plastomes, they have commonly been used in various evolutionary studies. In angiosperms plastid genomes are 35– 217 kb long, circular/linear (Oldenburg and Bendich, 2004) and comprise a total of 40–179 genes. The composition of plastomes with limited number of functional genes typically under purifying selection leaves only a fraction of the genome for intergenic spacer regions. Another interesting feature of plastid genomes is the exist-ence of two inverted repeats (IRs) of 15–32 kb. These regions are attached within the long single-copy (LSC) and short single-copy (SSC) regions on four distinct junction sites (JSs). The structural or-ganization of the genome along these sites is of crucial importance that can show sweep or evolutionary drift in lineages. The circular plastid molecule undergoes interconversion into a dumbbell-shaped conformation facilitated by the IRs, which are copy corrected via concerted evolution (Kolodner et al., 1976). The size variation of angiosperm plastid genomes is primarily due to expansion and con-traction of the IR and SSC boundary regions. The importance of IR boundaries can be inversely confirmed by the number of chloroplast genome publications presenting manually produced plots with the genes located around the JSs (e.g. Li et al., 2013).

2 Materials and methods

IRscope and its dependencies are coded in R. Besides the online usage of the software @ https://irscope.shinyapps.io/IRapp/, we release the source code of the software in the same directory. This allows the modification of the code and its more versatile use for more advanced users of R. We tested the program on over a 100 embryophytes sampled widely of dif-ferent orders. The accession number of this test data plus the detailed instructions for the use of the IRscope source code are available within the downloadable file on our online pages.

2.1 Input data

IRscope accepts the chloroplast genome sequence GenBank (GB) files as input ('.gb' suffixed files). The number of input genome files (max. 10) defines the number of the horizontal tracks for the plot. The files should follow the standard structure of the GB files for the program to function properly. For example, the species and gene names are read from the definition and gene lines of the GB files, respectively. Gene annotations should be as reliable as possible since the program is dependent on their corresponding coordinates. Files can be also uploaded manually in a DOGMA output format (Wyman et al., 2004) accompanied with a fasta file and/or the coordinates of the IR junctions. For more compact visualizations, we recommend the use of the shortened gene names in annotations. While the program can handle redundant bases or non-identical IRs resulting from sequencing errors, it is recommended that users assure that plastomes are assembled correctly.

2.2 Usage

The input data of the program should be either uploaded by users or NCBI accession numbers should be provided. Once the total number of GB files are gathered, the program performs a quick scan of the files and enters to the IR finding phase for each genome. Once IR coordi-nates are detected for each junction site for all sequences an optimum consensus radius to search and plot the genes will be detected. With the given coordinates of a JSs, the program plots the calculated relative positions of the genes within the corresponding radius on each track. Since chloroplast DNA within individual plants exhibits a form of heter-oplasmy in which the plastome exists in two equimolar states (inversion isomers) that differ in the relative orientation of the SSC region (see Palmer, 1983; Walker et al., 2015) we include a function, which allows the depiction of both forms.

3 Results

IRscope plots the specific information about the IR regions and the gene structures in the vicinity of the JSs. Each input GB file for each species is depicted as a horizontal track with their names and the corresponding chloroplast genome sequence length marked on the left. Four vertical lines representing different JSs partition the tracks into distinctly depicted sections of LSC, IRb, SSC and IRa with their marked corresponding lengths. The differently colored regions and genes, and their relative positions are plotted correspondingly for each site and track (Fig. 1). The output of IRscope is a 300ppi jpg file available for download through a personal link provided to user based on provided submissions.

4 Discussion and conclusion

IRscope is a generic local genomic visualizer aimed to reflect the scaled genetic structure of chloroplast genome sequences on their respective JSs. It can produce up to ten tracks for the selected entries. The program is ideally designed for visualization of the JSs of angiosperms and an optimum consensus radius for each gene is searched based on the selected species. Extending sampling to phylogenetically larger set of line-ages like embryophytes may cause complications in resolving the opti-mum consensus radius for each site followed by no or poor plot. In such cases, we recommend using the manual file upload section and dividing the data to smaller sets of closely related species and run the program once for each smaller set. The quality of the annotations is another factor affecting the output. Poor annotations may hinder software to function properly or may lead to low quality output. IRscope is coded in R and is freely available either for generic online use or for download. With consideration of the input data, IRscope is a reliable tool to help inspect evolutionary differences of the chloroplast genome of embryophytes.

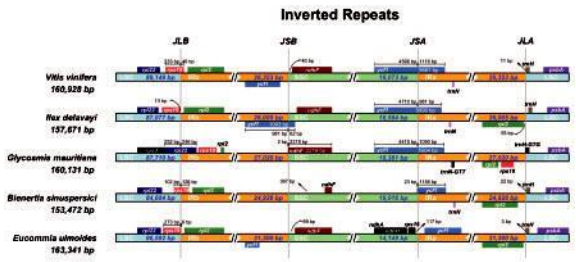

Fig. 1. IR plot of five angiosperms. Each species and their corresponding chloroplast genome sequence length are depicted to the left of each track. Genes transcribed in positive and negative strands are presented above and below of their corresponding tracks with from right-to-left and left-to-right directions, respectively. The arrows are showing the bp distance of the start or end coordinate of a given gene from the corresponding junction site. For the genes extending from a region to another, the T bar on the top or below shows the extent of their parts with their corresponding values. The genes in the vicinity of the junctions are the realistically scaled projections of the bp distances for each site. JLB (IRb /LSC), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (IRa/LSC) denote the JSs between each corresponding region in the genome

Acknowledgement

We thank University of Helsinki DPPS graduate school for providing the facilities and funding of the project.

Conflict of Interest: none declared.

References

Frazer,K.A. et al. (2004) VISTA: computational tools for comparative genom-ics. Nucleic Acids Res., 32, W273–W279.

Kolodner,R. et al. (1976) Physical studies on the size and structure of the cova-lently closed circular chloroplast DNA from higher plants. Biochim. Biophys. Acta, 447, 144–155.

Li,X. et al. (2013) Complete chloroplast genome sequence of Magnolia grandiflora and comparative analysis with related species. Sci. China. Life Sci., 56, 189–198.

Lohse,M. et al. (2007) OrganellarGenomeDRAW (OGDRAW)—a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet., 52, 267–274.

Lyons,E. and Freeling,M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J., 53, 661–673.

Oldenburg,D.J. and Bendich,A.J. (2004) Most chloroplast DNA of maze seed-lings in linear molecules with defined ends and branched forms. J. Mol. Biol., 335, 953–970.

Palmer,J.D. (1983) Chloroplast DNA exists in two orientations. Nature, 301, 92–93.

Walker,J.F. et al. (2015) Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. Am. J. Botany, 102, 1751–1752.

Wyman,S.K. et al. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics, 20, 3252–3255.