Drawing mitochondrial genomes with circularMT

Simon J. Goodman[1] and Ian M. Carr[2]

1) School of Biology, Faculty of Biological Science, University of Leeds, Leeds, LS2 9JT, UK

2) Leeds Institute of Medical Research at St James's, School of Medicine, University of Leeds, Leeds, LS9 7TF, UK

Abstract

Mitochondrial DNA sequences are used extensively in phylogeographic and phylogenetic studies for a wide range of organisms. With the advent of low-cost, high throughput 'next generation' DNA sequencing, and user-friendly bioinformatics pipelines for generating and annotating whole mitochondrial genome assemblies, the analysis of whole mitochondrial genomes has become an important component of phylogenomic studies for taxa with high species diversity but limited coverage for other genomic resources. An important step in characterising de novo mitochondrial genome assemblies is to evaluate and describe structural rearrangements relative to reference taxa. Accessible tools are needed to help visualise gene and noncoding feature complement, their order and strand orientation. However, there are few dedicated applications that generate high quality genome diagrams. Here we present circularMT and circularMT-console that allow users to create highly customisable, publication quality images, of linear and circular mitochondrial genome maps, either individually, or integrated into an analysis pipeline.

Availability and Implementation

Both applications are implemented in C#, with binaries, source code and user guides available on GitHub (https://github.com/msjimc/circularMT). An archive of the published version is available on Zenodo (https://zenodo.org/records/10912319).

Contact: I.M.Carr@leeds.ac.uk

Introduction

Mitochondria derive from the assimilation of symbiotic bacteria during the early evolution of eukaryotes (Sagan l. 1967) and are present in the majority of eukaryotes as the main site for aerobic metabolism (Voet et al, 2006). Although, for multicellular organisms, many of the genes present in the ancestral mitochondrion have transferred to the nuclear genome, mitochondria still possess a small circular genome encoding transcribed sequences, that in animals ranges from 11 to 28 kb (Andersson et al. 2002, Kolesnikov et al. 2012). With few exceptions, animal mitochondria have 37 transcribed sequences, comprising 13 protein coding genes, 22 tRNAs, and 2 mitochondrion specific rRNAs, although the order and orientation of features can vary.

In higher organisms, mitochondrial genomes are typically only inherited through the maternal line and so, from a population genetic perspective, behave as single haploid locus in linkage disequilibrium, without opportunity for recombination (Wei and Chinnery. 2020). The transcribed and non-transcribed regions of the mitochondrial genome experience different levels of purifying selection, and therefore have varying substitution rates, creating both highly conserved and hypervariable sequence regions. Metabolically active cells may have several hundred mitochondria, each containing 2 to 10 copies of their genome. As a consequence, mitochondrial genome copy number exceeds that of nuclear DNA (Robin and Wong 1988), making them easier to study than nuclear loci, especially when dealing with potentially degraded material, for example environmental samples.

These properties of mitochondrial genomes have long made them popular choices as a source of markers for phylogeographic and phylogenetic studies. With the advent of low-cost, high throughput massively parallel DNA sequencing, and user-friendly bioinformatics pipelines for generating and annotating whole mitochondrial genome assemblies (e.g. Jin et al. 2020; Bernt et al. 2013), their analysis has become an important component of phylogenomic studies (Tang et al. 2014). This is particularly relevant for the study of taxa such as invertebrates that have high

species diversity, but are sparsely covered by other genomic resources (Tang et al. 2019; Jasso-Martínez et al. 2020). As a result, the comparison of mitochondrial genomes is an increasingly important component of biodiversity discovery, taxonomic characterisation and monitoring workflows (Crampton-Platt et al. 2015).

An important step in characterising de novo mitochondrial genome assemblies is to evaluate and describe any structural rearrangements relative to reference taxa. Accessible tools are needed to help visualise a mitochondrion's gene and noncoding feature complement, their order and strand orientation, through circular or linear schematic plots of the genome. Options for generating such plots include proprietary software such as Geneious Prime (Geneious, 2024), or open-source resources including the OGDRAW webserver (Lohse et al. 2013), or various implementations of the CIRCOS package (Krzywinski et al. 2009). However, there are obstacles to their use. For instance, Geneious Prime is a commercially available software application that is able to perform a wide range sequence analysis and visualisation processes, of which the production of mitochondrial genome maps is a small part. As a consequence, the required licensing fee may make the application too expensive for the infrequent generation of mitochondrial genome images. Conversely, OGDRAW is a simple, free to use web application that is dedicated to the production of organelle genome images, however, it is limited to the processing of Genbank and EMBL-EBI annotation files and offers few options for the modification of the image, with the user encouraged to save the image as postscript file that can be further edited in a dedicated graphics application such as CorelDraw (CorelDraw 2024) or Inkscape (Inkscape 2024). Finally, CIRCOS is a fully featured application, initially created as a Perl module, for the visualisation of genomic data, typically in a circular fashion. While CIRCOS is very flexible and feature-rich, it requires a level of programming/command line expertise that many biological researchers do not possess and so its use may present a daunting learning curve to infrequent users. An implementation of CIROS hosted on Galaxy (The Galaxy Community 2022) has resolved some of these issues, but its use can be slow due to the multi-

user nature of Galaxy, and currently has limited documentation. Here we introduce circularMT, a desktop and command line application for producing high quality circular and linear maps of mitochondrial genomes from a wide range of input file formats. While simple to use, the application allows the user to modify the genomic maps, such as correcting erroneous annotation pipeline artefacts (for example non-coding features assigned with low confidence), and a high level of customisation for the graphical presentation of features of interest.

Design and Implementation

The object orientated language C# was used to create the desktop application: circularMT and the command line programme: circularMT-console. While primarily a Windows based program, they can be run on several POSIX-compliant operating systems, such as Linux, macOS, & BSD using Wine (Wine, 2024) as described in the user guide. The primary interface of circularMT consists of a single window containing controls on the righthand panel that modify the displayed genome, drawn on the plot area to the left, either directly or via a series of process specific dialogue boxes. Whereas circularMT-console is a command line driven implementation that can be used for batch processing files or as part of an analysis pipeline.

To aid useability, both programs are able to process a wide range of file formats most notably GenBank annotation files and those exported by MITOS (Bernt et al. 2013), a widely used organelle annotation web application currently hosted on Galaxy. While these files, and others like GTF and GFF formatted files, have a rigid structure, other file types are more loosely defined and must conform to the requirements outlined in the user guides.

As circularMT parses the input file, it determines each feature's genomic coordinates and orientation that are required to draw the map, along with the optional data of the feature's type (i.e. tRNA, rRNA or protein coding gene), and any names it has been given. This information is then used to create either a linear or circular map of the genome in which features are represented as arrows (pointing 5' to 3'). For circular maps the arrows form two rings with the

outer circle representing the forward strand and the inner circle the reverse strand, while linear maps are arranged as two rows of arrows with the upper and lower rows representing the forward and reverse strands respectively (Figure 1).

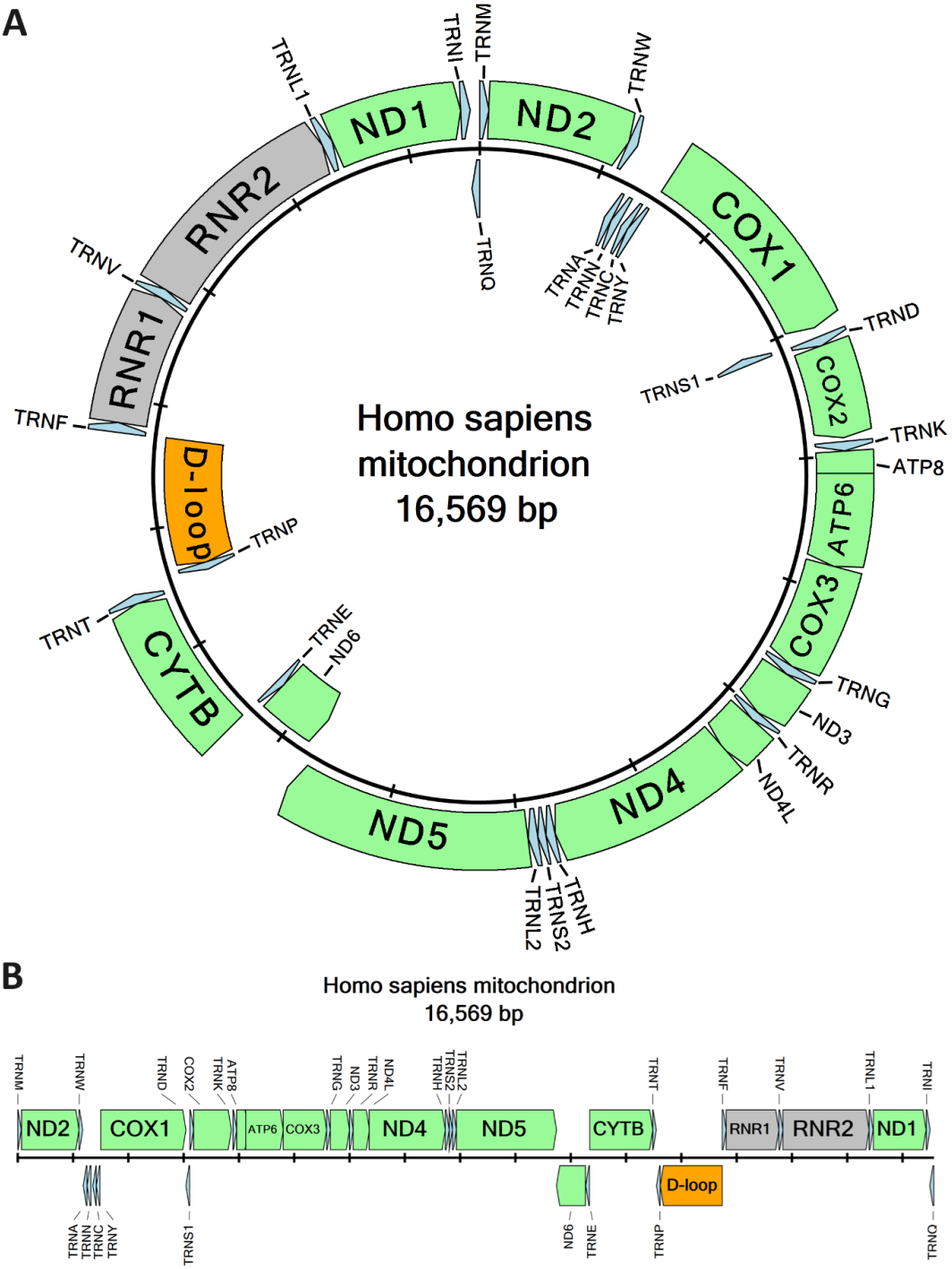

Figure 1: The human mitochondrial genome (NC_012920.1) draw as: A) circular and B) linear maps by circularMT using the sequence's Genbank annotation file.

Where possible, feature names are drawn within the arrow, but if the text is too long it is written by the side of and at 90˚ to the arrow. tRNAs often form tandem arrays of up to five genes, which due to their short sequence length may result in externally drawn text clashing, consequently circularMT will automatically modify the text's position to limit this overlap, while also offering a method for the user to manually adjust their position. For circular diagrams, the map will resize itself such that the names of features do not extend beyond the edge of the image. It also allows the user to manually modify other aspects of the map, for instance, the user can add, delete and rename features to correct errors in the annotation. It is also possible to customise the colour scheme of one or all the features of a specific feature type to produce a more informative colour coded image to meet the user's requirements.

The *de novo* assembly of mitochondrial genomes results in contigs in both orientations that start at random points in their genome. Consequently, circularMT allows the user to switch the annotation's strand orientation and change its start point such that a set of images conform to a standardised style, for instance all maps start at the beginning of the methionine encoding tRNA gene, which is placed on the forward strand.

The command line application circularMT-console is intended for the batch creation of images from multiple annotation files, or to be integrated into an analysis pipeline. As it runs in an unsupervised manner, it has few user set parameters and is primarily intended to create a first pass image for use in an analysis pipe line report, with circularMT used to create bespoke maps used in a more formal setting.

Once a genomic map has been produced by circularMT, it can be saved as one of several image formats (TIFF, JPEG, PNG or BMP) at a resolution between 100 to 1,000 dpi allowing the images to be used in publications, theses or reports (see Figure 1). Images created by circularMT-console are saved as 300 dpi, TIFF, JPEG, PNG or BMP files.

Summary

Due to their comparatively high copy number and their simple and relatively conserved structure, mitochondrial genomes are often one of the first de novo DNA sequences to be determined for a species. The ease by which they can then be amplified and sequenced from low quality environmental and bulk samples has led to them being routinely used in a wide range of biological fields such as biodiversity and evolutionary studies. An important first step in the study of these sequences it to determine their genomic organisation and any structural variation when compared to reference taxa. While a wide range of genome assembly and annotation workflows have been designed, there are fewer applications that visualise the genomes. Consequently, we have developed two free to use, versatile yet simple applications, that are able to draw high quality mitochondrial genome maps, for use in a wide range of specialisations.

Funding

No external funding supported this work.

References

Andersson SG, Karlberg O, Canbäck B, Kurland CG. On the origin of mitochondria: a genomics perspective. Philos Trans R Soc Lond B Biol Sci. 2003;358:165-77

Bernt, M, Donath, A, Jühling, F, Externbrink, F, Florentz, C, Fritzsch, G, Pütz, J, Middendorf, M, and Stadler, PF. MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution 2013;69:313–319.

Cantelli G, Bateman A, Brooksbank C, Petrov AI, Malik-Sheriff RS, Ide-Smith M, Hermjakob H, Flicek P, Apweiler R, Birney E, McEntyre J. The European Bioinformatics Institute (EMBL-EBI) in 2021. Nucleic Acids Res. 2022;50:D11-D19.

CoralDraw, https://www.coreldraw.com/en/ (27 March 2024)

Crampton-Platt A, Timmermans MJ, Gimmel ML, Kutty SN, Cockerill TD, Vun Khen C, Vogler AP. Soup to Tree: The Phylogeny of Beetles Inferred by Mitochondrial Metagenomics of a Bornean Rainforest Sample. Mol Biol Evol. 2015;32:2302-16

Geneious, https://www.geneious.com/features/ (27 March 2024)

Inkscape, https://inkscape.org/ (27 March 2024)

Jasso-Martínez JM, Quicke DLJ, Belokobylskij SA, Santos BF, Fernández-Triana JL, Kula RR, Zaldívar-Riverón A. Mitochondrial phylogenomics and mitogenome organization in the parasitoid wasp family Braconidae (Hymenoptera: Ichneumonoidea). BMC Ecol Evol. 2022;22:46.

Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639-45.

Kolesnikov AA, Gerasimov ES. Diversity of mitochondrial genome organization. Biochemistry (Mosc). 2012:77:1424-35.

Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41:W575-81.

Robin ED, Wong R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol. 1988;136:507-13

Sagan L. On the origin of mitosing cells. Journal of Theoretical Biology. 1967:14:225-274

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20-D26.

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X. Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res. 2014;42:e166.

Tang P, Zhu JC, Zheng BY, Wei SJ, Sharkey M, Chen XX, Vogler AP. Mitochondrial phylogenomics of the Hymenoptera. Mol Phylogenet Evol. 2019;131:8-18.

The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, 2022;50:W345–W351

Voet D, Voet JC, Pratt CW (2006). Fundamentals of Biochemistry (2nd ed.). John Wiley and Sons, Inc. pp. 547, 556. ISBN 978-0471214953.

Wei W, Chinnery PF. Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. J Intern Med. 2020;287:634-644.

Wine, https://www.winehq.org/ (22nd May 2024)