

SOFTWARE

Open Access



EqualTDRL: illustrating equivalent tandem duplication random loss rearrangements

Tom Hartmann^{1*} , Matthias Bernt² and Martin Middendorf¹

Abstract

Background: To study the differences between two unichromosomal circular genomes, e.g., mitochondrial genomes, under the tandem duplication random loss (TDRL) rearrangement it is important to consider the whole set of potential TDRL rearrangement events that could have taken place. The reason is that for two given circular gene orders there can exist different TDRL rearrangements that transform one of the gene orders into the other. Hence, a TDRL event cannot always be reconstructed only from the knowledge of the circular gene order before a TDRL event and the circular gene order after it.

Results: We present the program `EqualTDRL` that computes and illustrates the complete set of TDRLs for pairs of circular gene orders that differ by only one TDRL. `EqualTDRL` considers the circularity of the given genomes and certain restrictions on the TDRL rearrangements. Examples for the latter are sequences of genes that have to be conserved during a TDRL or pairs of genes that frame intergenic regions which might represent remnants of duplicated genes. Additionally, `EqualTDRL` allows to determine the set of TDRLs that are minimum with respect to the number of duplicated genes.

Conclusion: `EqualTDRL` supports scientists to study the complete set of TDRLs that possibly could have taken place in the evolution of mitochondrial genomes. `EqualTDRL` is implemented in C++ using the `ggplot2` package of the open source programming language R and is freely available from <http://pacosy.informatik.uni-leipzig.de/equaltdrl>.

Keywords: Circular permutation, Gene order, Genome rearrangement, Mitochondria, Tandem duplication random loss

Background

The genetic information of species is stored in DNA (or RNA) molecules. These molecules are called chromosomes and can either be *linear* or *circular*. The set of all these molecules of a species forms its *genome*. The genome consists of *genes* which are DNA segments with certain functions. Mutations can modify the arrangement or the multiplicity of the genes within the genome. Such mutations are called *rearrangements*. The research field of genome rearrangement analysis tries to explain the differences between two genomes that are represented by the arrangement of their genes, in order to infer phylogenetic information. The following two central problems exist in this research field [1]. The *distance problem* aims

to determine the *distance* between two genomes, i.e., the smallest number of rearrangements (of certain types) that are needed to transform one gene order into the other. The *sorting problem* asks for a shortest sequence of rearrangements for such a transformation. Such a sequence is called a *shortest scenario*. In case that costs can be assigned to different types of rearrangements the distance problem (the sorting problem) asks for the minimum total cost for the transformation from one gene order to the other (respectively, for a corresponding minimum cost sequence of rearrangements).

One important type of rearrangements is the *tandem duplication random loss* (TDRL) rearrangement. A TDRL consists of a tandem duplication of a contiguous set of genes followed by a random loss of one copy of each duplicated gene. TDRLs have occurred several times in the evolution of mitochondrial genomes, e.g., in gulper eels [2] and in millipedes [3]. Several mechanisms that explain a TDRL have been discussed in the biological

*Correspondence: thartmann@informatik.uni-leipzig.de

¹Swarm Intelligence and Complex Systems Group, Faculty of Mathematics and Computer Science, Leipzig University, Augustusplatz 10, D-04109 Leipzig, Germany

Full list of author information is available at the end of the article

literature, e.g., slipped strand mispairing during replication or imprecise termination [4]. It has been shown that TDRLs are a major factor of gene order evolution for mitochondrial genomes [4–6].

The TDRL rearrangement was initially studied formally for linear genomes in [7]. The cost of a TDRL rearrangement is defined as α^k , where $\alpha \geq 1$ is a parameter and $k \in \mathbb{N}$ is the number of genes that are influenced by the TDRL. For the cases $\alpha = 1$ and $\alpha \geq 2$ polynomial time algorithms that solve the sorting problem (and therefore the distance problem) have been presented in [7]. In addition, it has been shown for $\alpha = 1$, that it is sufficient to consider TDRLs that duplicate the whole genome, because the cost of every TDRL is the same in this case. Here we consider the case $\alpha = 1$ as well. Note that this definition of a TDRL considers explicitly only whole genome duplications but implicitly also partial genome duplications, as explained in detail in “Implementation” section.

The authors of [3] studied an alternative version of TDRL rearrangement where genes with the same orientation or genes that belong to the same transcript are lost jointly, i.e., the loss is not completely random but depends on gene orientation or transcript structure.

In order to reconstruct genome rearrangements reliably it is important to consider the possibility of alternative shortest rearrangement scenarios and equivalent rearrangements, i.e., rearrangements that when applied to the same gene order lead also to the same resulting gene order. This has been studied for the inversion, the transposition and the double cut and join (DCJ) rearrangement [8, 9]. Equivalent inversions for signed circular permutations, which represent circular genomes, have been discussed in [10] and all shortest scenarios for the sorting of signed permutations have been studied in depth in [11–14]. Since every transposition can be represented by a TDRL, the set of equivalent transpositions has been studied in [15]. The set of all shortest scenarios of the problem to sort a multichromosomal genome by DCJ rearrangements has been investigated in [16]. In this work the authors also determined the exact number of optimal shortest scenarios for a particular set of problem instances.

An analysis of TDRL rearrangements on circular genomes has been presented in [15]. It has been shown that the circularity of the genomes should be considered, since the TDRL distance for an unfavorable choice of linear representatives (of the circular genomes) may lead to an overestimation of the distance. In addition, it has been shown that it is not always possible to uniquely reconstruct a TDRL only from the knowledge of the two circular gene orders before and after the application of the TDRL because there exist several TDRLs that can explain the change from one circular genome to the other.

Software tools that regard TDRL rearrangements for the construction of evolutionary scenarios are CREX [17], CREX2 [18], and TreeREX [19]. However, these tools present only a single shortest scenario in the case that a TDRL is present in a shortest scenario. Hence, when a scenario contains a TDRL, the tools do not show all possible alternative TDRLs. Therefore, it is important to further analyze the solutions that are generated and the software EqualTDRL that we present in this paper is designed to support scientists to do such an analysis.

EqualTDRL uses results from [15] to consider the circularity of the given gene orders. It provides figures that illustrate all equivalent TDRL operations for two given gene orders that differ by only a single TDRL. Therefore, EqualTDRL supports scientists to study the whole set of TDRLs that possibly could have taken place. This is beneficial in several aspects of inferring a most reliable TDRL. For example: i) for identifying whether the gene loss of a TDRL is random or is dependent on gene orientation or transcript structure, ii) for finding a (partial duplication) TDRL that duplicates only a minimum number of genes, iii) to determine a subset of TDRLs that satisfy certain conditions, e.g., to preserve certain sequences of genes that cannot be broken during a TDRL, and iv) for identifying the positions of potential duplication remnants for a further analyses of the nucleotide sequence.

This article is organized as follows. In the next “Implementation” section a formal background is given on circular gene orders and TDRL rearrangements. Further, an overview on EqualTDRL is presented. In “Results and discussion” section the benefits of EqualTDRL are shown for a biological example of mitochondrial gene orders. The article ends with a conclusion in “Conclusion” section.

Implementation

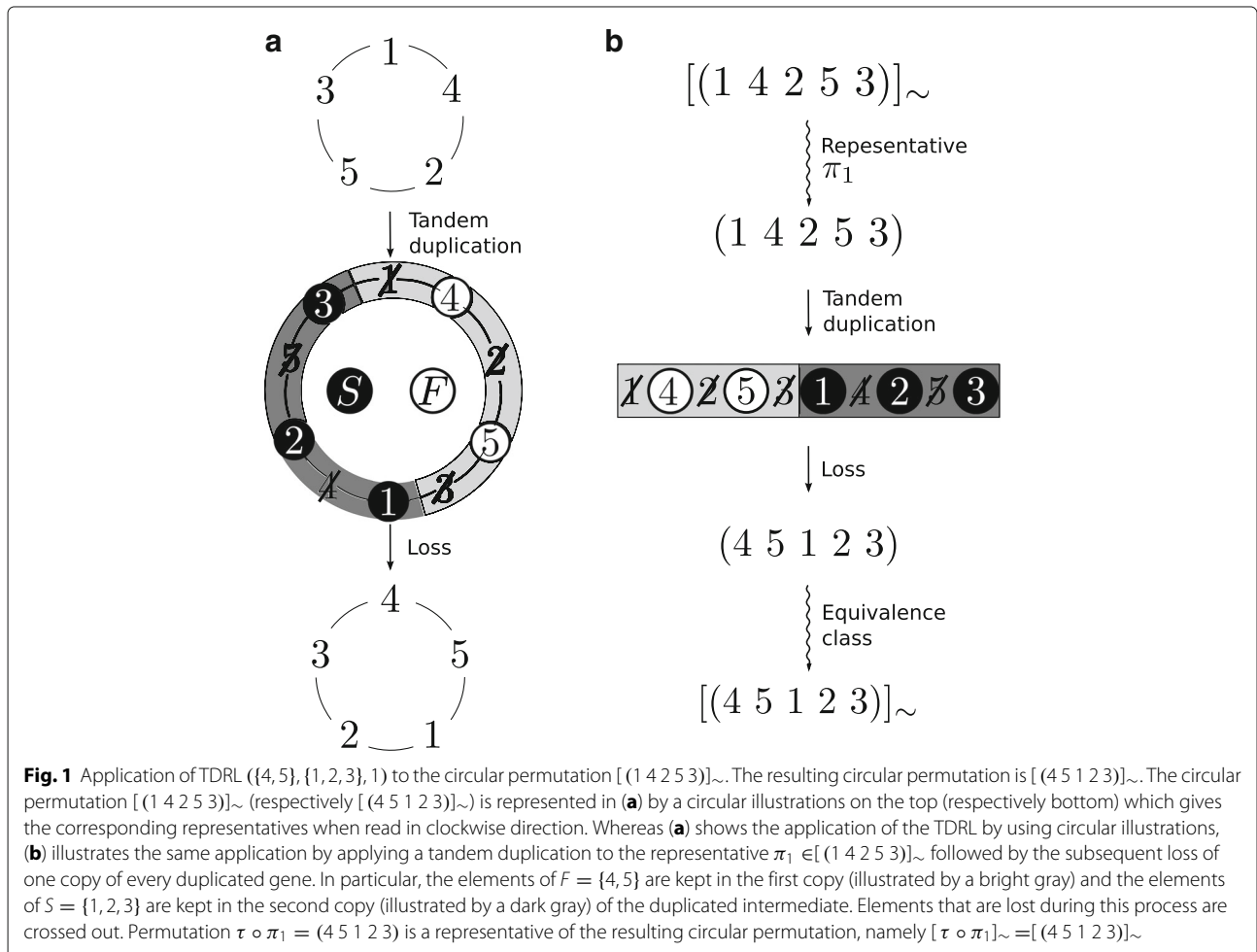
Methods

In this article it is assumed that the genes in a genome (before and after a TDRL) are not duplicated and that genes do not overlap. Therefore, a gene order of a (circular) genome can be represented by a (circular) permutation. A *permutation* π of length n , denoted by $\pi = (\pi(1) \dots \pi(n))$, is a bijection $\pi: [1 : n] \rightarrow [1 : n]$. The set of all permutations of length n is denoted by \mathcal{S}_n . The *shift operation* $\phi: \mathcal{S}_n \rightarrow \mathcal{S}_n$ is defined by $\pi = (\pi(1) \dots \pi(n)) \mapsto (\pi(2) \dots \pi(n) \pi(1))$ and for $k \in \mathbb{N}_{>0}$ the *k-shift* is $\phi^k \circ \pi$, where $\phi^k := \phi \circ \phi^{k-1}$ and $\phi^1 := \phi$. Note that with $f \circ g$ the composition of two functions f and g is denoted, i.e., $(f \circ g)(x) := f(g(x))$. With \sim we denote the equivalence relation on \mathcal{S}_n , where $\pi, \pi' \in \mathcal{S}_n$ are equivalent, denoted by $\pi \sim \pi'$, if and only if there exists an $m \in [1 : n]$ with $\phi^m(\pi) = \pi'$. The equivalence class $\pi^\circ := [\pi]_\sim$ of \sim on \mathcal{S}_n is called a *circular permutation* of length n and the set of all circular permutations of length

n is denoted by S_n° . In other words, a circular permutation $\pi^\circ \in S_n^\circ$ is the set of all permutations that are equivalent with respect to \sim , i.e., $\pi^\circ = \{\pi, \rho(\pi), \dots, \rho^{n-1}(\pi)\}$ or, less formally, a circular permutation is the set of all permutations that become equivalent when the first and the last element of a permutation are considered to be adjacent. Figure 1a shows an illustration of a circular permutation. Each $\pi \in \pi^\circ$ is called a *representative* of π° and the representative which starts with element $p \in [1 : n]$ is denoted by π_p . Circular permutations are used as a formal model for unichromosomal circular genomes, e.g., mitochondrial genomes, in which each element represents a gene and each representative stands for a possible linearization of the considered genome.

A TDRL $\tau^\circ : S_n^\circ \rightarrow S_n^\circ$ is a bijection that is denoted by a triple (F, S, p) , where (F, S) is a bipartition of $[1 : n]$, i.e., $F, S \subset [1 : n]$, $F \cap S = \emptyset$, and $F \cup S = [1 : n]$, and $p \in [1 : n]$. Element p is called the *origin* of (F, S, p) and it denotes the position where the whole genome duplication of π° starts resulting in a duplicated intermediate. The sets F and S are a bipartition of $[1 : n]$ and denote the

sets of genes that are kept (i.e., they are not deleted) in the first and second copy of π_p in the duplicated intermediate, respectively. Here “first” and “second” are defined with respect to π_p . The elements that are deleted in the duplicated intermediate are called *lost elements*. The effect of a TDRL $\tau^\circ = (F, S, p)$ on π° is defined as $\tau^\circ \circ \pi^\circ := [\tau \circ \pi_p]_\sim$, where $\tau \circ \pi_p$ is a rearranged permutation of π_p such that the elements of F are moved in front of the elements of S and the relative order of all elements of F (respectively S) of π_p is unchanged. More precisely, $\tau \circ \pi_p$ is the (linear) permutation for which it holds that 1) if $i \in F, j \in S$, then i is to the left of j in $\tau \circ \pi_p$ and 2) if $i, j \in F$ or $i, j \in S$, then i is to the left of j in $\tau \circ \pi_p$ if and only if i is to the left of j in π_p . Note that by this definition, a TDRL (F, S, p) , which maps a set π° of permutations to another set of permutations, can be visualized by reordering the elements of the representative π_p according to the sets F and S , see Fig. 1b for an example. For the combinatorics of TDRLs it is irrelevant whether F or S is the set of genes that are preserved in the original part of the duplicated intermediate, since both applications result in the same circular permutation



[15], i.e., $(F, S, p) \circ \pi^\circ = (S, F, p) \circ \pi^\circ$. Note that the strand-
 edness of a gene (also called its orientation) is not relevant
 for this paper since a TDRL does not change the strand-
 edness of genes. This is also the reason why we represent
 a gene order as an (unsigned) permutation.

Two TDRLs are called *equivalent* if their application
 to the same (circular) permutation results in the same
 (circular) permutation. It was shown that the number of
 equivalent TDRLs is $2n^2$ if the TDRLs are identity maps
 and $2n$ otherwise, where n is the length of permutation
 π° [15]. The TDRL distance of π° and σ° is $d(\pi^\circ, \sigma^\circ) =$
 $\min(\{d \in \mathbb{N}_0 \mid \exists \text{ TDRLs } \tau_1^\circ, \dots, \tau_d^\circ : \tau_d^\circ \circ \dots \circ \tau_1^\circ \circ \pi^\circ = \sigma^\circ\})$.

The definitions related to TDRL rearrangements and
 circular permutations are exemplified for the circular per-
 mutation $\pi^\circ = [(1\ 4\ 2\ 5\ 3)]_\sim$. See Fig. 1 for an illustration.
 The representatives of π° are $\pi_1 = (1\ 4\ 2\ 5\ 3)$, $\pi_2 =$
 $(2\ 5\ 3\ 1\ 4)$, $\pi_3 = (3\ 1\ 4\ 2\ 5)$, $\pi_4 = (4\ 2\ 5\ 3\ 1)$, and
 $\pi_5 = (5\ 3\ 1\ 4\ 2)$, i.e., $\pi^\circ = \{\pi_1, \dots, \pi_5\}$. It holds that $\pi_2 =$
 $\phi^2(\pi_1)$, $\pi_3 = \phi^2(\pi_2)$, $\pi_4 = \phi^2(\pi_3)$, $\pi_5 = \phi^2(\pi_4)$, and $\pi_1 =$
 $\phi^2(\pi_5)$. The application of TDRL $\tau^\circ = (\{4, 5\}, \{1, 2, 3\}, 1)$
 to π° gives $[(4\ 5\ 1\ 2\ 3)]_\sim$. Since genes 1 and 2 are in S
 and 1 is to the left of 2 in π_1 , it holds that 1 is to the
 left of 2 in $(4\ 5\ 1\ 2\ 3)$. Also, since $5 \in F$ and $3 \in S$, it
 holds that 3 is to the right of 5 in $(4\ 5\ 1\ 2\ 3)$. Figure 1
 illustrates the application of τ° to $[(1\ 4\ 2\ 5\ 3)]_\sim$. Since
 $(\{3, 4\}, \{1, 2, 5\}, 5) \circ \pi^\circ = (\{4, 5\}, \{1, 2, 3\}, 1) \circ \pi^\circ$ it holds
 that TDRLs $(\{3, 4\}, \{1, 2, 5\}, 5)$ and $(\{4, 5\}, \{1, 2, 3\}, 1)$ are
 equivalent.

To see that the formal model also covers partial duplica-
 tion TDRLs (i.e., TDRLs where not all genes are duplicated)
 consider a partial duplication TDRL that transforms π°
 into σ° . For every element $e \in [1 : n]$ it holds that either
 $e \in F'$ (i.e., e is kept in the first copy), $e \in S'$ (i.e., e is kept
 in the second copy), or $e \in N$ (i.e., e is not duplicated).
 Then the TDRL (F, S, p) , where $F = S'$, $S = N \cup F'$, and
 p being the unique non-duplicated element adjacent to
 one element of the second copy in the partially duplicated
 intermediate, gives the same circular output permutation
 (see Fig. 2 for an example). Consequently, the same rear-
 rangement can be achieved by a TDRL that duplicates all
 elements [15].

EqualTDRL

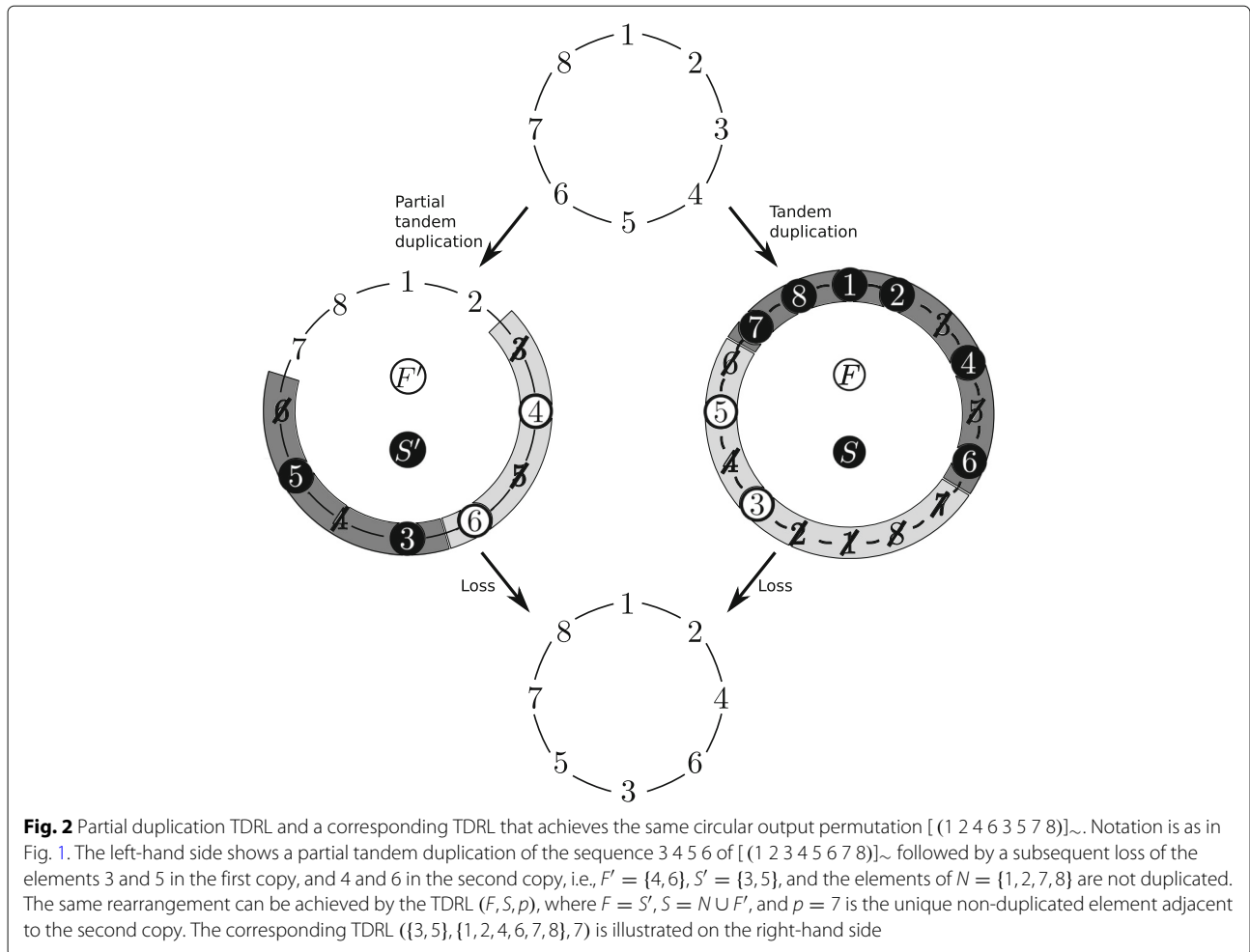
The software tool EqualTDRL calculates for two
 circular permutations π° and σ° that represent two
 circular unichromosomal genomes the distances
 $d(\pi^\circ, \sigma^\circ)$ and $d(\sigma^\circ, \pi^\circ)$. In the case that $d(\pi^\circ, \sigma^\circ) = 1$
 (respectively $d(\sigma^\circ, \pi^\circ) = 1$) EqualTDRL produces
 an illustration which shows all equivalent TDRL rear-
 rangements, i.e., all triples (F, S, p) that transform π°
 into σ° or vice versa. An example of such an illustra-
 tion is given in Fig. 3 for the circular permutations
 $\iota^\circ = [(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)]_\sim$ and $\pi^\circ = [(1\ 2\ 4\ 6\ 3\ 5\ 7\ 8)]_\sim$ with
 $d(\iota^\circ, \pi^\circ) = 1$. Figure 3 illustrates all equivalent TDRLs

that transform ι° into π° : $(\{3, 5, 7, 8\}, \{1, 2, 4, 6\}, 1)$,
 $(\{1, 3, 5, 7, 8\}, \{2, 4, 6\}, 2)$, $(\{4, 6\}, \{1, 2, 3, 5, 7, 8\}, 3)$, $(\{1, 2, 5, 7, 8\},$
 $\{3, 4, 6\}, 4)$, $(\{3, 6\}, \{1, 2, 4, 5, 7, 8\}, 5)$, $(\{1, 2, 4, 7, 8\}, \{3, 5, 6\}, 6)$,
 $(\{3, 5\}, \{1, 2, 4, 6, 7, 8\}, 7)$, $(\{3, 5, 7\}, \{1, 2, 4, 6, 8\}, 8)$, and
 all TDRLs that can be obtained from the listed TDRLs
 by interchanging the sets F and S . Recall that TDRL
 $(\{3, 5\}, \{1, 2, 4, 6, 7, 8\}, 7)$ is illustrated in Fig. 2.

Moreover, EqualTDRL can illustrate all partial TDRLs
 in addition to the whole genome TDRLs. This is shown in
 Fig. 3 where a black square surrounds the circle of a gene
 if and only if the gene is part of the duplicated sequence of
 a partial duplication TDRL. Note that for partial TDRLs
 it is not important whether genes that are excluded from
 the duplication are considered to be in F or in S since they
 are not duplicated and therefore no copy of these genes is
 lost. As an example, consider the origin 3 in Fig. 3. The
 figure shows that TDRL $(\{4, 6\}, \{1, 2, 3, 5, 7, 8\}, 3)$, which
 considers the whole permutation to be duplicated, can be
 replaced by a partial duplication TDRL that only dupli-
 cates the sequence 3 4 5 6 and contains the genes 4, 6 in
 the – with respect to the origin – first copy of the dupli-
 cated intermediate and genes 3, 5 in the second copy of the
 duplicated intermediate. See Fig. 2 for an illustration of
 the specified partial duplication TDRL and a correspond-
 ing TDRL that duplicates the whole permutation.

Finally, EqualTDRL is also able to illustrate only those
 TDRLs that satisfy the following types of conditions which
 can be given by the user: i) specific sets of sequences of
 genes that are conserved by a TDRL and ii) intergenic
 regions that are framed by specific pairs of genes. Both
 conditions and the type of input that they require are
 explained in the following.

For the first type of condition the user has to specify
 the corresponding sets of genes. Then, EqualTDRL pro-
 ceeds as explained in the following. Consider a set of genes
 $G \subset [1 : n]$, two circular permutations π° and σ° such that
 $d(\pi^\circ, \sigma^\circ) = 1$, and a TDRL $\tau^\circ = (F, S, p)$ with $\tau^\circ \circ \pi^\circ =$
 σ° . When G is used as a condition EqualTDRL only con-
 sideres equivalent TDRLs (F', S', p') of τ° such that either
 $G \subseteq F'$ or $G \subseteq S'$. Therefore, if there exist such an equiva-
 lent TDRL all genes of G are lost in the same copy and the
 loss of this TDRL depends on G . In the case that set G
 contains exactly all genes that belong to the same transcript
 the first type of condition is used to determine whether
 or not the loss of a TDRL depends on a given transcript
 structure. Note that EqualTDRL can also use multiple
 gene sets. For an example consider Fig. 3 that illustrates all
 equivalent TDRLs that transform $\iota^\circ = [(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)]_\sim$
 into $\pi^\circ = [(1\ 2\ 4\ 6\ 3\ 5\ 7\ 8)]_\sim$. Let 1 2 and 7 8 be a sequence
 of genes (of π° and ι°) that shall be conserved by a TDRL,
 hence $G_1 = \{1, 2\}$ and $G_2 = \{7, 8\}$. If these conditions are
 specified, then EqualTDRL provides an illustration simi-
 lar to Fig. 3 but with the difference that it does not show
 the TDRLs for origin 2 and 8, i.e., $(\{1, 3, 5, 7, 8\}, \{2, 4, 6\}, 2)$



and $(\{3, 5, 7\}, \{1, 2, 4, 6, 8\}, 8)$. These TDRLs are not illustrated since for $i \in [1 : 2]$ neither $G_i \subseteq F$ nor $G_i \subseteq S$. However, for all other origins $p \in \{1, 3, 4, 5, 6, 7\}$ it holds that G_1 and G_2 are subsets of either F or S .

The second type of condition requires to find all possible TDRLs that allow to deduce intergenic regions between pairs of genes that are given to EqualTDRL as an input from the user. To see this, consider two circular permutations π° and σ° , a TDRL $\tau^\circ = (F, S, p)$ such that $\tau^\circ \circ \pi^\circ = \sigma^\circ$, and a pair of two distinct genes $x, y \in [1 : n]$ that is given by the user. Such a pair of genes should frame an intergenic region in the genome that is represented by σ° . Then EqualTDRL considers only equivalent TDRLs (F', S', p') of τ° where at least one gene of the duplicated intermediate that is with respect to $\pi_{p'}$ between x and y is deleted, i.e., there exists at least one gene $z \in [1 : n] \setminus \{x, y\}$ such that z is between x and y in the duplicated intermediate, z is lost, and either $x \in F$ and $y \in S$ or if $x, y \in F$ (respectively $x, y \in S$) then $z \in S$ (respectively $z \in F$). For such a TDRL the intergenic region between x and y in the genome that is represented by σ° can be explained

by an incomplete deletion of the gene (or the set of genes) that were lost between x and y . Therefore, if such a TDRL exists EqualTDRL provides evidence that an intergenic region might be a remnant of a gene (or the set of genes) that is formed by an incomplete deletion of the same. For an example consider Fig. 3 that illustrates all equivalent TDRLs that transform $\iota^\circ = [(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)]_{\sim}$ into $\pi^\circ = [(1\ 2\ 4\ 6\ 3\ 5\ 7\ 8)]_{\sim}$. Assume that an intergenic region is between genes 3 and 5 in π° and that one is interested to know which TDRLs (and which corresponding gene losses) applied to ι° could possibly result into this arrangement. Hence $x = 3$ and $y = 5$ are chosen. If this condition is given, EqualTDRL would produce Fig. 3. Hence, all illustrated TDRLs allow to deduce the intergenic region between 3 and 5. This holds because in all TDRLs of Fig. 3 either $3 \in F$ and $5 \in S$ (e.g., origin 5), $5 \in F$ and $3 \in S$ (e.g., origin 4), or $3, 5 \in F$ (respectively $3, 5 \in S$) and the gene 4, which is between 3 and 5 in the duplicated intermediate, is lost. If, for example, condition $x = 1$ and $y = 2$ is used, then EqualTDRL provides a figure that illustrates only TDRL $(\{1, 3, 5, 7, 8\}, \{2, 4, 6\}, 2)$ of Fig. 3 (i.e., the row of

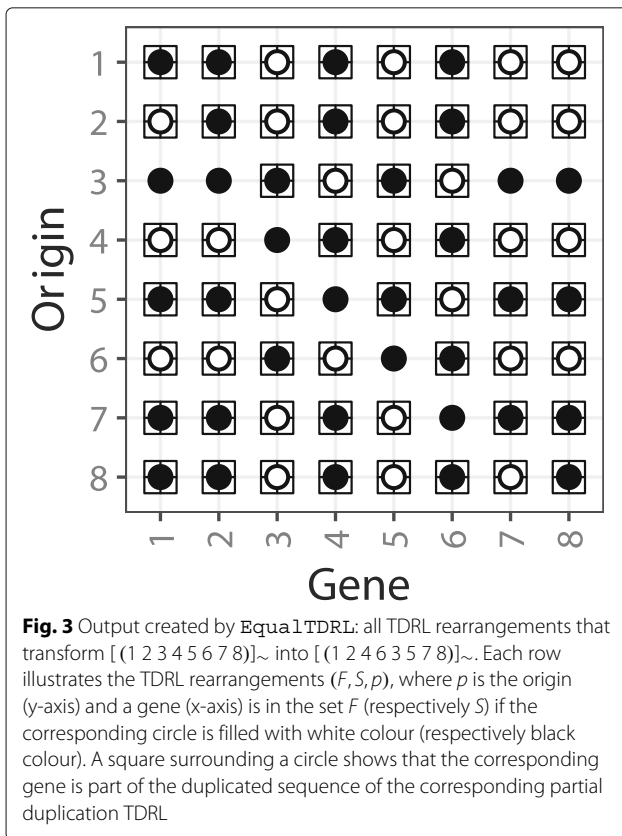


Fig. 3 with origin 2). This holds since $1 \in F$, $2 \in S$, and 3 is between 1 and 2 in π_2 . For all other origins $p \in [1 : 8] \setminus \{2\}$ holds that $1, 2 \in F$ (respectively $1, 2 \in S$) and there does not exist an element z between 1 and 2 in the duplicated intermediate that is lost.

Results and discussion

Experiment

In this section we show on the basis of mitochondrial gene orders how EqualTDRL can be used to find a plausible TDRL rearrangement. Specifically, the following two mitochondrial gene orders are considered: i) the gene order of *Prionoglaris stygia* (Genbank, accession: MG255141.1), which represents the ancestral mitochondrial gene order of the Pancrustacea [20], and ii) the mitochondrial gene order of the Trogiomorpha species *Lepidopsocidae sp.* (RefSeq, accession: NC_004816.1) and *Dorypteryx domestica* (Genbank, accession: MG255136.1) that has been published at the NCBI RefSeq release 84 [21] and the NCBI Genbank release 221 [22]. Note that the mitochondrial genomes of *Lepidopsocidae sp.* and *Dorypteryx domestica* comprise the same gene order. Both gene orders have recently been discussed in [20]. In this publication the authors studied the phylogenetic relationships between different barklice species of *Psocoptera* (a

taxon which includes all considered species) and estimated the history of rearrangements using the condition-based coding algorithm MacClade [23], which does not support TDRL rearrangements, and the TreeREx software [19]. For both given mitochondrial gene orders the TreeREx analysis resulted in a large TDRL rearrangement which transforms the gene order of *Prionoglaris stygia* into the gene order of *Lepidopsocidae sp.* by duplicating 34 of 38 genetic markers (37 genes plus the control region). The corresponding TDRL is illustrated in Fig. 4 for the origin *cox2*.

Since for circular gene orders several equivalent TDRLs exist [15], we investigate the set of all possible TDRL rearrangements in the following. Since we also want to consider the intergenic sequences the mitochondrial genomes of the Trogiomorpha species *Lepidopsocidae sp.* and *Dorypteryx domestica* have been reannotated with an extended version of MITOS (unpublished <http://mitos2.bioinf.uni-leipzig.de>) [24]. In order to examine whether an intergenic region contains remnants of genes caused by an incomplete gene loss, the following data analysis was executed for *Lepidopsocidae sp.* and *Dorypteryx domestica*. For every tRNA and ribosomal gene (respectively protein-coding gene) the Covariance Models (respectively Hidden Markov Models) that are used in MITOS were applied i) to search in every intergenic region for every gene sequence and ii) if a gene (or a remnant of a gene) has been found to (locally) align the gene sequence to the corresponding intergenic region. The analysis was carried out for the tRNA and ribosomal genes with CMsearch and CMAalign from the Infernal 1.1rc4 software package [25] and for protein-coding genes with HMMsearch and HMMalign from the HMMER 3.1b1 software package [26]. Moreover, for both mitochondrial gene orders the conserved sequences of genes (i.e., the maximal – with respect to inclusion – sequences of genes that occur in both gene orders) were calculated.

Then EqualTDRL was used to determine TDRLs that transform the mitochondrial gene order of *Prionoglaris stygia* into the mitochondrial gene order of *Lepidopsocidae sp.* under four different objectives: A) to show all possible TDRLs, B) to show all TDRLs that do not break any conserved sequence of genes, C) to highlight a subset of the TDRLs of (B) that provide additional evidence for the origin of intergenic regions outside of conserved sequences by an incomplete gene loss, and D) to find all TDRLs that minimize the number of duplicated genes.

Results

The gene orders which result from the annotation that has been done with MITOS for the mitochondrial genomes of *Prionoglaris stygia*, *Lepidopsocidae sp.*, and *Dorypteryx domestica* are equal to the gene orders that have been discussed in [20].

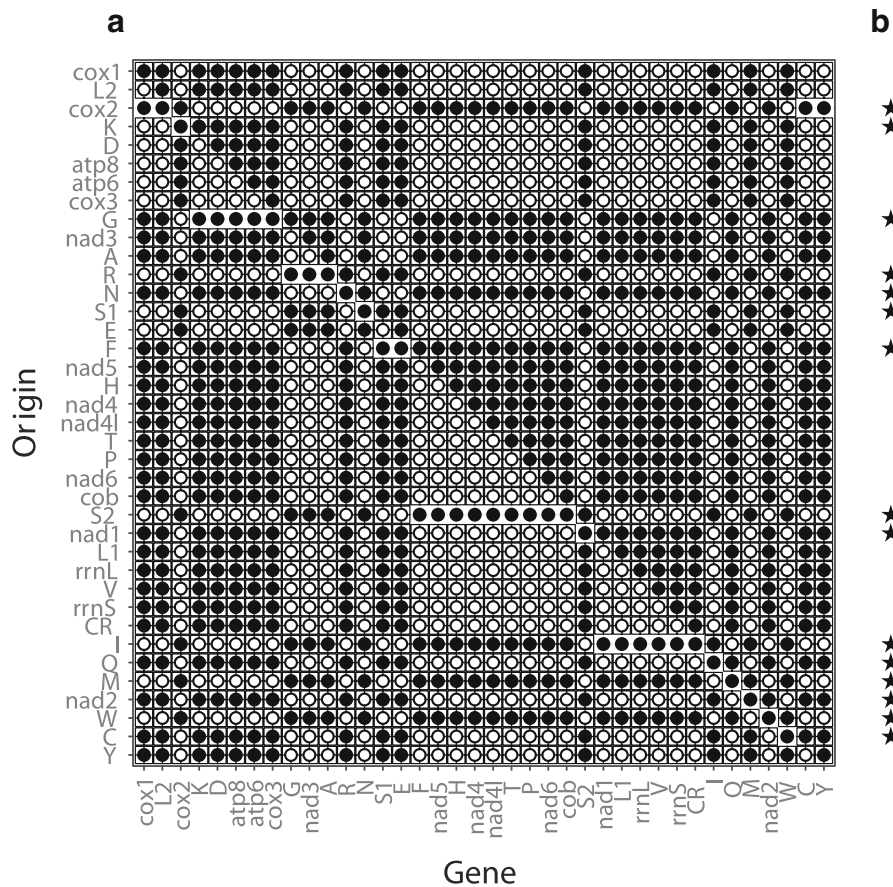
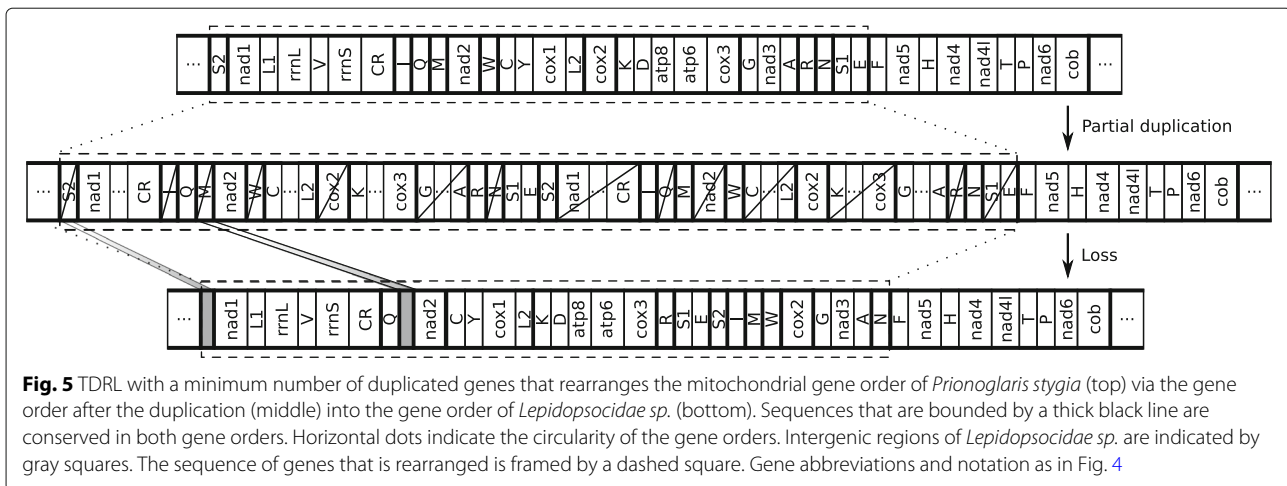


Fig. 4 a Output created by Equal TDRL: complete set of TDRLs that rearrange the mitochondrial gene order of *Prionoglaris stygia* into the mitochondrial gene order of *Lepidopsocidae sp.* Protein-coding and ribosomal genes are denoted by their names and one capital letter indicates the amino acid for the tRNAs. **b** A star denotes a TDRL that does not break a conserved gene sequence. A pentagon highlights the TDRL that has been presented in [20]. A diamond highlights the TDRL which duplicates the minimum number of genes (the corresponding minimum TDRL is illustrated in Fig. 5)

A linear representation of the mitochondrial gene orders of *Prionoglaris stygia* and *Lepidopsocidae sp.* (respectively *Dorypteryx domestica*) that are used in this study is shown in Fig. 5. It can be seen that both gene orders contain the following conserved sequences: i) *trnC trnY cox1 trnL2*, ii) *trnK trnD atp8 atp6 cox3*, iii) *trnG nad3 trnA*, iv) *trnS1 trnE*, v) *trnF nad5 trnH nad4 nad4l trnT trnP nad6 cob*, and vi) *nad1 trnL1 rrnL trnV rrnS CR*, where CR denotes the control region.

The analysis with MITOS shows that the genome of *Lepidopsocidae sp.* contains 10 intergenic regions with a length of at least 10 base pairs, whereby only the three regions between *trnQ* and *nad2* (79 base pairs), *nad2* and *trnC* (105 base pairs), and *cob* and *nad1* (58 base pairs) are not contained in a conserved sequence. It is worth to mention that the intergenic region between *nad2* and *trnC* cannot be found in a less restrictive annotation with MITOS (when E-value exponent of 1 is used instead of the default value 2 for BLAST [27] and CMsearch).

In this less restrictive annotation the putative intergenic region is completely assigned to *nad2*. The mitochondrial genome of *Dorypteryx domestica* contains 4 intergenic regions with a length of at least 10 base pairs: between *trnQ* and *nad2* (50 base pairs), *trnC* and *trnY* (31 base pairs), *trnN* and *trnF* (25 base pairs), and *cob* and *nad1* (21 base pairs). Note that the intergenic region between *trnC* and *trnY* is contained in a conserved interval. Since *Lepidopsocidae sp.* and *Dorypteryx domestica* have the same gene order, which is different to the ancestral *Pancrustacea* gene order, we assume that the TDRL studied in the following has happened before the speciation of the Trogiomorpha species. Therefore, we consider only intergenic regions that occur in both mitochondrial genomes of *Lepidopsocidae sp.* and *Dorypteryx domestica*. Thus in the analysis that follows we consider the intergenic regions between *trnQ* and *nad2*, and between *cob* and *nad1*. These intergenic regions are particularly interesting for further analysis since the presence of a gene remnant in those



intergenic regions can be used to identify the subset of TDRLs that can explain the presence of the remnants as a result of an incomplete gene loss of a TDRL.

Figure 4a was generated with EqualTDRL and shows all equivalent TDRLs that transform the gene order of *Prionoglaris stygia* into the gene order of *Lepidopsocidae sp.* It also shows the set of duplicated genes for every TDRL. In Fig. 4b symbols were (manually) added to the right of a TDRL if and only if the TDRL exhibits specific characteristics: those that do not break any of the conserved sequences of genes and those that supports conjectures that can explain the presence of the intergenic regions as a result of an incomplete gene loss. These indications can be used to further search for gene remnants. Interestingly, it can be seen in Fig. 4 that the TDRL which has been presented in [20] preserves all conserved sequences of genes and it also explains the presence of both considered intergenic regions.

The HMMER and Infernal software package were used to identify potential remnants of genes. The experiments showed that *trnM* and *trnS1* can be found in the intergenic region of *Lepidopsocidae sp.* between *trnQ* and *nad2* with an E-value of 0.005 and 0.016, respectively. In addition several protein-coding genes hit the intergenic region between *trnQ* and *nad2*. However, the E-values of these hits are larger than 0.2 and therefore less reliable. Interestingly, no hit with an E-value smaller than 1 was found for both intergenic regions of *Dorypteryx domestica* and the intergenic region between *cob* and *nad1* of *Lepidopsocidae sp.* A table that summarizes all hits in the intergenic regions, its E-values, and the corresponding alignments can be found in the Additional file 1: Table S3 and Figures S1–S10.

Figure 4 shows that there exists a TDRL which duplicates only 29 of 38 genetic markers, which are 5 less than the TDRL that has been presented in [20]. Interestingly,

this TDRL also preserves all conserved gene sequences and provides evidence for the presence of the intergenic regions between *trnQ* and *nad2*, and *cob* and *nad1* as a result of an incomplete gene loss of the genes *trnM* and *trnS2*, respectively. While the origin of the former intergenic region by the loss of *trnM* is supported weakly by our analysis, the origin of the latter intergenic region has no support from our sequence/structural similarity based analysis. The corresponding TDRL is illustrated in Fig. 5.

Altogether, the scenario presented in [20] partially agrees with the results that are presented in this article. However, due to the circularity of the mitochondrial gene orders it is important to consider the complete set of TDRL rearrangements that can explain the gene order of *Lepidopsocidae sp.* When it is considered to be relevant that a TDRL duplicates only a minimum number of genes, the results computed with EqualTDRL weaken the support for the TDRL rearrangement that has been presented in [20] in favor of the TDRL rearrangement that is shown in Fig. 5. Moreover, the TDRL presented in Fig. 5 gets additional support by the detection of a putative remnant of the *trnM* gene in the intergenic region of *Lepidopsocidae sp.* between *trnQ* and *nad2*.

Conclusion

In this article we have presented the tool EqualTDRL, which illustrates all equivalent TDRLs for a pair of gene orders that differ by one TDRL. EqualTDRL considers the circularity of genomes and helps to study their differences in consideration of all TDRL predictions that possibly could have taken place. Thereby, it helps to identify TDRLs that satisfy different biological constraints. For example, a requirement might be that the TDRL duplicates only a minimum number of genes or that the TDRL allows to explain the presence of intergenic regions. In

addition, EqualTDRL supports scientists to determine whether the gene loss of a TDRL is random or might dependent on gene orientation or transcript structure. It has been shown for two example mitochondrial gene orders how EqualTDRL can be used to identify more plausible TDRLs that possibly could have taken place.

Availability and requirements

Project name: EqualTDRL

Project home page: <http://pacosy.informatik.uni-leipzig.de/equaltdrl>

Operating system(s): Linux distribution

Programming language: C++,R

Other requirements: ggplot2 package of R

License: none

Any restrictions to use by non-academics: none

Additional file

Additional file 1: Supplement. Additional file that contains the MITOS2 annotations and alignments of the mitochondrial genomes analyzed in the current study. (PDF 222 kb)

Abbreviations

CR: Control region; TDRL: Tandem duplication random loss

Funding

TH was funded by a PhD student fellowship from the Leipzig University. We acknowledge support from the German Research Foundation (DFG) and Leipzig University within the program of Open Access Publishing. The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The dataset analyzed in the current study is available in the NCBI database [21]. The code of EqualTDRL and examples can be found at <http://pacosy.informatik.uni-leipzig.de/equaltdrl>. The remainder of the data generated during this study is included in this published article and its supplementary information file.

Authors' contributions

TH and MB wrote and tested the EqualTDRL code. MM made substantive intellectual contributions to this paper. TH wrote the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Swarm Intelligence and Complex Systems Group, Faculty of Mathematics and Computer Science, Leipzig University, Augustusplatz 10, D-04109 Leipzig, Germany. ²Helmholtz Centre for Environmental Research - UFZ, Permoserstraße 15, D-04318 Leipzig, Germany.

Received: 22 January 2018 Accepted: 30 April 2018

Published online: 30 May 2018

References

- Fertin G, Labarre A, Rusu I, Tannier E, Vialette S. *Combinatorics of Genome Rearrangements*. Cambridge: MIT Press; 2009.
- Inoue JG, Miya M, Tsukamoto K, Nishida M. Evolution of the deep-sea gulper eel mitochondrial genomes: Large-scale gene rearrangements originated within the eels. *Mol Biol Evol*. 2003;20(11):1917–24.
- Lavrov DV, Boore JL, Brown WM. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. *Mol Biol Evol*. 2002;19(2):163–9.
- Boore JL. The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In: *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. Dordrecht: Springer; 2000. p. 133–47.
- Bernt M, Braband A, Schierwater B, Stadler PF. Genetic aspects of mitochondrial genome evolution. *Mol Phyl Evol*. 2012;69:328–38.
- San Mauro D, Gower DJ, Zardoya R, Wilkinson M. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. *Mol Biol Evol*. 2005;23(1):227–34.
- Chaudhuri K, Chen K, Mihaescu R, Rao S. On the tandem duplication-random loss model of genome rearrangement. In: *Proc. 17th Ann. ACM-SIAM Symp. Discrete Algorithm (SODA '06)*. Philadelphia: Society for Industrial and Applied Mathematics; 2006. p. 564–70.
- Yancopoulos S, Attie O, Friedberg R. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*. 2005;21(16):3340–6.
- Bergeron A, Mixtacki J, Stoye J. A unifying view of genome rearrangements. In: *Proc. 6th Int'l Workshop Algorithms in Bioinformatics (WABI '06)*. LNCS, vol. 4175. Berlin: Springer; 2006. p. 163–73.
- Meidanis J, Walter M, Dias Z. Reversal distance of signed circular chromosomes. Technical report IC-00-23. 2000.
- Bergeron A, Chauve C, Hartman T, St-Onge K. On the properties of sequences of reversals that sort a signed permutation. In: *Proceedings of JOBIM*, vol. 2; 2002. p. 99–108.
- Siepel AC. An algorithm to enumerate sorting reversals for signed permutations. *J Comput Biol*. 2003;10(3-4):575–97.
- Braga MD, Sagot M-F, Scornavacca C, Tannier E. The solution space of sorting by reversals. *Proc. 3rd Int'l Symp. Bioinforma Res Appl (ISBRA '07)*. 2007;4463:293–304.
- Braga MDV, Sagot M-F, Scornavacca C, Tannier E. Exploring the solution space of sorting by reversals, with experiments and an application to evolution. *IEEE ACM T Comput Bi*. 2008;5(3):348–56.
- Hartmann T, Chu A-C, Middendorf M, Bernt M. Combinatorics of tandem duplication random loss mutations on circular genomes. *IEEE ACM T Comput Bi*. 2016;15(1):83–95.
- Braga MD, Stoye J. The solution space of sorting by DCJ. *J Comput Biol*. 2010;17(9):1145–65.
- Bernt M, Merkle D, Ramsch K, Fritzsich G, Perseke M, Bernhard D, Schlegel M, Stadler P, Middendorf M. CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics*. 2007;23(21):2957–8.
- Hartmann T, Bernt M, Middendorf M. An Exact Algorithm for Sorting by Weighted Preserving Genome Rearrangements. *IEEE ACM T Comput Bi*. 2018. In press.
- Bernt M, Merkle D, Middendorf M. An algorithm for inferring mitogenome rearrangements in a phylogenetic tree. In: *Proc. 6th Int'l Workshop Comparative Genomics (RECOMB-CG '08)*. Berlin: Springer; 2008. p. 143–57.
- Yoshizawa K, Johnson KP, Sweet AD, Yao I, Ferreira RL, Cameron SL. Mitochondrial phylogenomics and genome rearrangements in the barklice (insecta: Psocodea). *Mol Phyl Evol*. 2017.
- Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(Database issue):61–5.
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. Genbank. *Nucleic Acids Res*. 2012;41(D1):36–42.
- Maddison WP, Maddison DR. *MacClade: analysis of phylogeny and character evolution*. version 3.0. Sunderland: Sinauer Associates; 1992.
- Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsich G, Pütz J, Middendorf M, Stadler PF. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol Phyl Evol*. 2013;69(2):313–9.

25. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29(22):2933–5.
26. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39(suppl_2):29–37.
27. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–402.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

