BMC Bioinformatics

RESEARCH ARTICLE Open Access

PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees

Joseph MEX Lucas[1,2,3], Matthieu Muffato[4] and Hugues Roest Crollius[1,2,3*]

Abstract

Background: Extant genomes share regions where genes have the same order and orientation, which are thought to arise from the conservation of an ancestral order of genes during evolution. Such regions of so-called conserved synteny, or synteny blocks, must be precisely identified and quantified, as a prerequisite to better understand the evolutionary history of genomes.

Results: Here we describe PhylDiag, a software that identifies statistically significant synteny blocks in pairwise comparisons of eukaryote genomes. Compared to previous methods, PhylDiag uses gene trees to define gene homologies, thus allowing gene deletions to be considered as events that may break the synteny. PhylDiag also accounts for gene orientations, blocks of tandem duplicates and lineage specific de novo gene births. Starting from two genomes and the corresponding gene trees, PhylDiag returns synteny blocks with gaps less than or equal to the maximum gap parameter gap_{max}. This parameter is theoretically estimated, and together with a utility to graphically display results, contributes to making PhylDiag a user friendly method. In addition, putative synteny blocks are subject to a statistical validation to verify that they are unlikely to be due to a random combination of genes.

Conclusions: We benchmark several known metrics to measure 2D-distances in a matrix of homologies and we compare PhylDiag to i-ADHoRe 3.0 on real and simulated data. We show that PhylDiag correctly identifies small synteny blocks even with insertions, deletions, incorrect annotations or micro-inversions. Finally, PhylDiag allowed us to identify the most relevant distance metric for 2D-distance calculation between homologies.

Keywords: Comparative genomics, Synteny, Synteny block, Segmental homologies, Homology, Gene order, Rearrangement, Ancestral genome, Gene tree

Background

Changes in the order of genes in a genome are caused by two categories of mutational events: genic events, which include de novo gene births, deletions, duplications, and genomic rearrangements, which include chromosome fusions and fissions, segmental translocations or segmental inversions. Synteny blocks are composed of those genes that retain an ancestral organisation despite these events, and one way to understand how genic events and genomic rearrangements affect genome evolution is to identify such synteny blocks. The extremities of synteny blocks also define the positions of breakpoints where rearrangements took place. Precisely defining synteny blocks thus allows, in turn, an accurate definition of breakpoints [1], which has important implications from ancestral genome reconstruction [2] to the understanding of genome mutational processes in healthy and disease states [3]. In addition, it has been shown in eukaryotes that some synteny blocks may be under negative selection due to long-range functional constraints between genes and regulatory elements [4,5].

Several methods have been developed to identify synteny blocks from extant chromosomes comparisons. In the field of bacterial genome evolution, algorithms tend to focus on the notion of "gene team" [6], which denotes a

*Correspondence: hrc@ens.fr
[1]Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, 46 rue d'Ulm, 75005 Paris, France
[2]CNRS, UMR 8197, 75005 Paris, France
Full list of author information is available at the end of the article

set of genes that stay in the vicinity of each other with no constraint on gene order. Such methods include TEAM [7], HomologyTeams [8], CCCPart [9], CloseUp [10] and MCMuSeC [11].

However, because gene order conservation in eukaryotes is stronger [12] compared to bacteria, algorithms that infer synteny blocks in eukaryotes tend to account for this extra constraint. GRIMM-synteny [13], i-ADHoRe 3.0 (often just called ADHoRe later) [14-17], DiagHunter [18], LineUp [19], FISH [20], DAGchainer [21], SyMAP [22], ColinearScan [23], Cinteny [24], OrthoCluster [25], Syntenator [26] and Cyntenator [27], MCScan [28] and MCScanX [29], Enredo [30], and DRIMM-Synteny [31] are the main algorithms developed to infer synteny blocks in eukaryotes. Many were applied to model species such as *Arabidopsis thaliana* and rice, among plants, and mammals such as human, mouse, dog and rat, among metazoans. These algorithms can be broadly classified according to their heuristic and features.

Four distinct heuristics are used to infer synteny blocks. The first builds two-dimensional matrices filled with homologies [13,17,18,20,22,24]. The algorithms analyse the matrices with procedures that resemble those developed in the field of image analysis.

A second heuristic uses optimisation techniques and dynamic programming [19,21,28]. Many of the methods that fall in this category are greedy, although with the benefit of often providing more flexibility. Indeed, the choice of the cost parameters in the objective function, allows the user to accurately account for different synteny block characteristics. A third heuristic is based on a modification of the Smith-Waterman [32] approach [23,26] while the last type of heuristic relies on graph editing [30,31].

Some algorithms compare genomes by performing pairwise comparisons of genomes whereas others perform multi-genomes comparisons. Combining pairwise comparisons does not capture the additional significance of genes that are conserved in more than two regions, resulting in under-estimation of cluster significance [33]. Multi-genomes comparisons are especially relevant for highly diverged synteny blocks and Whole Genome Duplication (WGD) analysis. However, multi-genomes comparisons usually require genomes to be reduced to a set of markers shared between all genomes, thus limiting the resolution of the analysis.

The transcriptional orientations of genes on the chromosome are used by some algorithms and provide informations about micro-rearrangements and may contribute to making the correct choice when there are several possibilities to extend a synteny block. In addition, accounting for gene orientations increases the statistical relevance of small synteny blocks, see [Additional file 1: Section 11].

Gene duplications increase the complexity of identifying synteny blocks. Duplications can be dispersed, or in tandem when the two copies are adjacent. Tandem duplications create blocks of tandem duplicates that disrupt local gene adjacencies without strictly breaking the synteny. In order to overcome blocks of tandem duplicates, algorithms may propose to collapse tandem duplicates into one occurrence by remapping their coordinates [17,20] or by performing ad hoc editions of the graph of adjacencies [30,31]. WGDs complicate matters further when new genes copies have been randomly inactivated throughout the genome. Yet some algorithms identify highly diverged synteny blocks or double conserved syntenies caused by WGDs [17,19].

Once an algorithm has returned putative synteny blocks, a statistical validation can assess their relevance given the input data. A putative synteny block is more likely to be found by chance in a comparison involving a large number of homologies than when few homologies are available. A putative synteny block is also less likely to have occurred by chance if it is composed of a large number of ordered adjacent homologies than if it is composed of a few unordered homologies separated by gaps. Statistical validation may involve either a p-value, an e-value or a score. The analytical calculation is not a simple task [8,33-36] and there is no standard p-value yet established in the field. Simulations are often used to bypass this difficulty, although they are usually time consuming and not very realistic.

To infer a synteny block, each algorithm uses parameters such as the maximum gap gap_{max} to define the maximum allowed distance between two genes in a synteny block. The gap_{max} parameter value can be optimised through a theoretical exploration, saving the need to test numerous different values before finding the optimal value [23].

Another important variable is the metric used to allow gaps between genes within a synteny block. Some algorithms use the Diagonal Pseudo Distance [17,18] whereas others use the Manhattan Distance [13,20,22,24].

Finally, a useful feature is to represent synteny blocks graphically, such as diagonals in a matrix [18], circular views [29] or alignments [14,29].

Here, we are interested in reconstructing synteny blocks to capture the signals of ancestral gene order and gene orientations in eukaryotic genomes. To this end, we developed PhylDiag, a user-friendly method to identify synteny blocks between two genomes using reconstructed phylogenetic gene trees. The full evolutionary history of each ancestral gene is taken into account in the form of those phylogenetic gene trees, which include in particular gene losses, duplications, 1:1 but also 1:many and many:many homology relationships. All PhylDiag parameters can either be set automatically or be specified by the user. A p-value calculation provides a statistical basis to select significant blocks and a utility provides graphical

representations of identified synteny blocks. Users may also chose among several metrics to allow gene gaps within a synteny block. PhylDiag accounts for tandem duplications and gene orientations, and is thus able to accurately identify small synteny blocks. Among algorithms that already account for gene order and gene orientations, only i-ADHoRe 3.0, FISH and Enredo also handle tandem duplications, although they do not use gene trees reconstructions. Here we compare PhylDiag to i-ADHoRe 3.0 [14] (version i-ADHoRe 3.0.2a) using both real data and simulations.

By introducing the concepts of "tandem blocks" and "homology packs", PhylDiag overcomes the disruption of gene adjacencies caused by blocks of tandem duplicates. As in other existing methods, PhylDiag allows gaps between genes within synteny blocks up to a customizable maximum gap parameter, and thus bypasses small genic indels (insertions and deletions) and annotation errors. In this study, we also benchmark different metrics used to allow these gaps within a synteny block on simulated data, and show that the choice of the metric has a direct impact on performances.

Methods

After providing basic definitions, we describe the PhylDiag algorithm, which consists of four main parts. First, PhylDiag filters extant genomes. Second, PhylDiag rewrites the genomes from lists of genes to lists of tandem blocks. Third, PhylDiag extracts synteny blocks as diagonals with no gaps by considering the order and orientations of tandem blocks on the chromosomes and then merge these diagonals as long as merges do not generate gaps longer than gap_{max}. Finally, PhylDiag computes a p-value to remove diagonals that are likely to be produced by chance rather than being a signature of an ancestral gene order. Before performing these tasks PhylDiag also calculates a recommended value for the maximum gap gap_{max} to free the user from testing multiple values before finding the appropriate one.

Basic notations and definitions

Genomic conventions

S is a species. Given two species S_a and S_b, $\mathsf{LCA}(S_a, S_b)$ is the Last Common Ancestor of S_a and S_b. A species S_a has a genome G_a composed of chromosomes. $c_a = [\mathsf{g}_{a,1}, \ldots, \mathsf{g}_{a,N_a}] = [\mathsf{g}_{a,k}]_{k\in[1,N_a]}$ is a chromosome of G_a with N_a oriented genes $\mathsf{g}_{a,k}$. The chromosome is chosen to be ordered from $\mathsf{g}_{a,1}$ to $\mathsf{g}_{a,N}$ and not the reverse, thus defining a reference orientation. The orientation of a gene is determined by the orientation of transcription into RNA, and the orientation of $\mathsf{g}_{a,k}$, denoted $o(\mathsf{g}_{a,k})$, is equal to $+1$ if transcription is performed in the same direction as $\overrightarrow{\mathsf{g}_{a,1}\mathsf{g}_{a,N}}$ otherwise $o(\mathsf{g}_{a,k}) = -1$. A sub-list of c_a is often denoted $c_a[\, i_s \rightarrow i_e]$ where i_s (respectively i_e) is the index of the starting (respectively ending) gene in the sub-list.

Synteny block, intuitive definition

Intuitively (a formal definition is given in 'Synteny block, formal definition') we define a Synteny Block (sb, plural sbs) between two species S_a and S_b as a set of neighbouring genes with gene content, gene order and gene orientations conserved during the evolution from $\mathsf{LCA}(S_a, S_b)$ to S_a and S_b. Two genes are neighbours if they are separated by less than a user-defined parameter gap_{max}. During evolution we consider that a set of neighbouring genes remains a synteny block until:

- a chromosomal rearrangement creates a breakpoint within the sb and changes the order or the orientations of genes
- the gap between any two neighbouring genes, caused by gene insertions and/or gene deletions, exceeds gap_{max} genes (see the formal definition of gap_{max} in 'Synteny block, formal definition' and see 'Step 1: Filter extant genomes' for the choice of the type of gene insertions or gene deletions that may break the synteny)

An ancestral sequence of genes remains a sb even if tandem duplications occur within the synteny block.

Gene family and homology

The evolution of a gene can be represented by a rooted binary tree called a gene tree. The root of a gene tree is the first ancestral gene, the nodes correspond to events of speciations or duplications that occurred during the evolutionary history of the descending genes, and the leaves of the gene tree correspond to extant genes originating from the first gene.

Two genes are homologs if they are in the same gene tree. Two genes are orthologs if they are in the same gene tree and if their last common event is a speciation. Two genes are paralogs if they are in the same gene tree and if their last common event is a duplication. The homology relationship between two genes g_a and g_b is denoted $\mathsf{g}_a \mathcal{H} \mathsf{g}_b$. A homology relation defines classes of homologs, called families. An issue in comparative genomics is to define gene families and gene trees. Sequence comparison algorithms provide measures (such as BLASTP [37] scores) that make it possible to quantify the similarity between two sequences which may, in turn, be used to cluster genes that show high similarity, thus defining gene families. Gene families can then be organised in phylogenetic gene trees using a vast choice of tree reconstruction methods. Here, we use gene trees from Ensembl [38], built using the TreeBest pipeline [39]. Since in this study we are interested in finding synteny blocks conserved from $\mathsf{LCA}(S_a, S_b)$ to S_a and S_b, we

pruned all gene trees to define a gene family as a set of genes that come from a unique gene of LCA(S_a, S_b). Families are defined with these genes, so that two genes are in the same family if and only if they come from the same ancestral gene of LCA(S_a, S_b). We note that, depending on the purpose of the analysis, PhylDiag offers the possibilty to prune gene trees at an ancestor that precedes LCA(S_a, S_b), so that more paralogy relationships are included in the gene family, see [Additional file 1: Section 1].

Considering the species tree of Figure 1A and the original gene tree of Figure 1B, the Figure 1C describes how we pruned the original gene trees to define our families. Ultimately, the roots of the gene trees correspond to a unique gene of LCA(S_a, S_b).

Step 1: Filter extant genomes

When comparing two species S_a and S_b, the first step of PhylDiag is to propose a filtering of extant genomes. There are two filters:

- *InBothSpecies* removes genes that have no homolog in the other genome. This only retains genes that previous algorithms call "anchor genes" and it is the classical way of filtering extant genomes. This filter is well suited for finding functional clusters of genes.
- *InCommonAncestor* removes genes that arose de novo specifically after LCA(S_a, S_b). The removed genes are those that have no ancestral gene in LCA(S_a, S_b) and they are called "lineage specific genes". The selective removal is possible using the pre-computed phylogenetic gene trees. This step is equivalent to retaining "anchor genes", but here, using gene trees, the procedure also keeps genes that have lost their ortholog in the other species because of a deletion since LCA(S_a, S_b). This filtering is well suited for reconstructing ancestral gene orders.

Both filtering get rid of the noise introduced by lineage specific genes. PhylDiag using the *InBothSpecies* filter does not consider ancestral gene deletions as events that break the synteny whereas PhylDiag using the *InCommonAncestor* filter does consider ancestral gene deletions as events that break the synteny.

It may also be advantageous in some specific cases of functional studies of synteny blocks to avoid filtering extant genomes thus considering de novo births of lineage specific genes as events that break the synteny.

Depending on the desired purpose, PhylDiag offers the possibility to easily choose between no filtering at all, the *InBothSpecies* filter or the *InCommonAncestor* filter. Since in this study we are interested in reconstructing the ancestral gene order, the *InCommonAncestor* filter is applied and extant genomes should now be considered to only be composed of genes that have an ancestral gene in LCA(S_a, S_b).

Step 2: Build the matrix of homology packs

Extracting sbs conserved in G_a and G_b corresponds to extracting sbs for each comparison of chromosomes c_a of G_a and c_b of G_b. Indeed, genes in two different chromosomes, if they were in synteny before, have been separated by a chromosomal rearrangement and the synteny is broken anyway. Thus it is justified to limit the search to pairs of chromosomes rather than pairs of genomes.

Tandem blocks, an abstraction of genes

In a chromosome, under a parsimonious reasoning, homologous and adjacent genes are tandem duplicates. Here, we refer to such blocks as "tandem block". Formally,

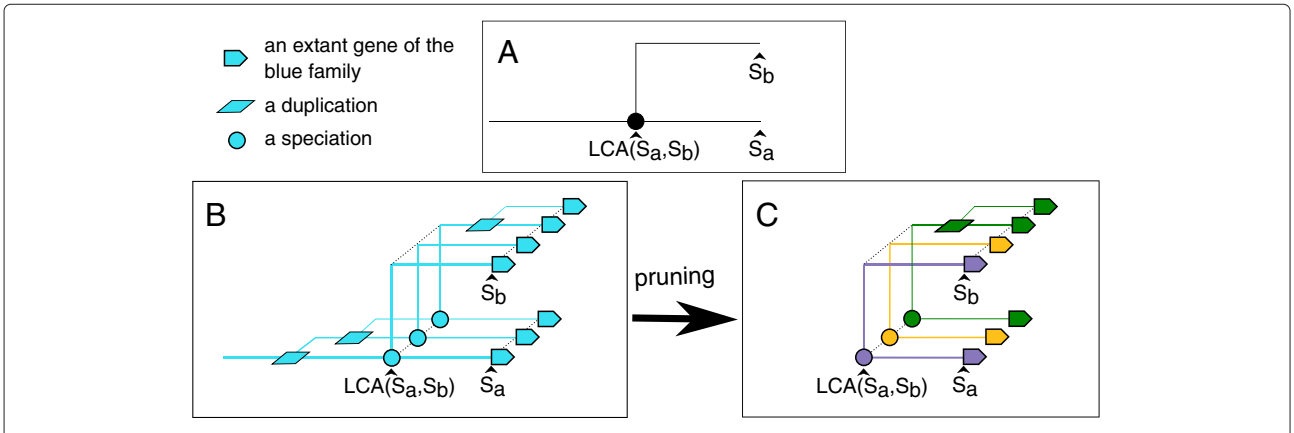

Figure 1 Definition of gene families based on gene tree pruning. Figure **A** represents a species tree with two extant species S_a and S_b, and their last common ancestor LCA(S_a, S_b). Figure **B** represents a gene tree within the species tree. This gene tree is represented in simple 3D schema for better visualisation. In a gene tree, squares represent duplication events and circles represent speciation events. Figure **C** shows how the original gene tree of figure B is pruned in order to define families that correspond to a unique gene of LCA(S_a, S_b).

a tandem block (tb, plural tbs) of a chromosome c is an uninterrupted sub-list of c that contains paralogous genes. For instance, if the 3 paralogous genes g_4, g_5 and g_6 are in an uninterrupted row in c, the corresponding tb is the sub-sequence $c[4 \rightarrow 6] = [\mathsf{g}_4, \mathsf{g}_5, \mathsf{g}_6]$. The size of a tb is equal to the number of tandem gene copies that it contains, for instance the last tb has a size 3. A gene which has no tandem duplicate is in a tb of size 1. By convention tbs are always maximum, i.e. a given tb cannot be contained within another tb. Like genes, tbs also have an orientation. However, in a tb, tandem duplicates may or may not all have the same orientation. When they all share the same orientation, the tb itself is oriented with the same orientation as the orientation of the genes thus, either $o(\mathsf{tb}) = +1$ or $o(\mathsf{tb}) = -1$. When tandem duplicates have different orientations, the orientation of the tb is considered to be *unknown*, and $o(\mathsf{tb}) = \varnothing$.

It is possible to rewrite chromosomes as a unique ordered list of oriented tbs. For instance $c_a = [\mathsf{g}_{a,1}, \ldots, \mathsf{g}_{a,N_a}]$ can be rewritten $c_a = [\mathsf{tb}_{a,1}, \ldots, \mathsf{tb}_{a,n_a}]$ where n_a is the number of tbs in c_a. $n_a \leq N_a$ and $n_a = N_a$ if and only if there is no tandem duplicate in c_a.

A tandem block tb_a of S_a is said to be in a homology relation with a tandem block tb_b of S_b if the genes of the two tbs are in the same family. We will also say that in this case tb_a and tb_b are homologs or even that tb_a and tb_b are homologous tandem blocks. Using the same notation as for genes, $\mathsf{tb}_a \mathcal{H} \mathsf{tb}_b$ means that tb_a and tb_b are homologs. If tb_a and tb_b are homologs, they share a Last Common Ancestral gene in $\mathsf{LCA}(S_a, S_b)$ and we note $\mathsf{LCAg}(\mathsf{tb}_a, \mathsf{tb}_b)$ the Last Common Ancestral gene of tb_a and tb_b. $\mathsf{LCAg}(\mathsf{tb}_a, \mathsf{tb}_b)$ is defined as soon as it is observed that $\mathsf{tb}_a \mathcal{H} \mathsf{tb}_b$. Of note, two homologous tandem blocks tb_a and tb_b are not necessarily of the same size if deletions or tandem duplications took place specifically in the branches of S_a or S_b after $\mathsf{LCA}(S_a, S_b)$.

Matrix of homologies

The classic Matrix of Homologies $\mathsf{MH} \in \mathfrak{M}_{N_a,N_b}$ of two chromosomes $c_a = [g_{a,k}]_{k\in[1,N_a]}$ and $c_b = [g_{b,k}]_{k\in[1,N_b]}$ is defined such that:

$$\mathsf{MH}[i,j] = \begin{cases} g_{a,i} \bullet g_{b,j}, & \text{if } g_{a,i} \mathcal{H} g_{b,j} \\ 0, & \text{otherwise} \end{cases} \qquad \forall (i,j) \in [1, N_a] \times [1, N_b]$$

Where $g_a \bullet g_b$ is the "sign" of the homology of g_a and g_b

$$g_a \bullet g_b = \begin{cases} +1, & \text{if } o(g_a) = o(g_b) \\ -1, & \text{if } o(g_a) = -o(g_b) \end{cases}$$

A MH can be represented as an array of values equal to $+1, -1$ or 0. Non-0 values correspond to homologies.

Homology packs, an abstraction of homologies

A Homology Pack (hp, plural hps) is the set of homology relationships between the tandem duplicates of two homologous tandem blocks tb_a (in c_a) and tb_b (in c_b). A hp is always maximum, i.e. a hp cannot be contained within another hp. Graphically, a hp appears as a rectangle of non-0 values in a MH. Each hp has a last common ancestral gene in $\mathsf{LCA}(S_a, S_b)$ denoted $\mathsf{LCAg}(\mathsf{hp})$ and equal to $\mathsf{LCAg}(\mathsf{tb}_a, \mathsf{tb}_b)$. Tandem duplications generate vertical, horizontal, or rectangular hps in a MH, making it difficult to identify sbs as diagonals. However, the rewriting of a chromosome in a way that collapses these hps to unique values in the MH, as described above, greatly simplifies this problem. Indeed, once c_a and c_b are rewritten as ordered lists of tbs, it becomes possible to define a matrix whose non-0 values correspond to hps of the two chromosomes c_a and c_b.

Matrix of homology packs

Given that c_a is rewritten in $[\mathsf{tb}_{a,k}]_{k\in[1,n_a]}$ and c_b is rewritten in $[\mathsf{tb}_{b,k}]_{k\in[1,n_b]}$, we introduce the Matrix of Homology Packs $\mathsf{MHP} \in \mathfrak{M}_{n_a,n_b}$ of the two chromosomes $c_a = [\mathsf{tb}_{a,k}]_{k\in[1,n_a]}$ and $c_b = [\mathsf{tb}_{b,k}]_{k\in[1,n_b]}$ defined such that:

$$\mathsf{MHP}[i,j] = \begin{cases} \mathsf{tb}_{a,i} \bullet \mathsf{tb}_{b,j}, & \text{if } \mathsf{tb}_{a,i} \mathcal{H} \mathsf{tb}_{b,j} \\ 0, & \text{otherwise} \end{cases} \qquad \forall (i,j) \in [1, n_a] \times [1, n_b]$$

Whith $\mathsf{tb}_a \bullet \mathsf{tb}_b$ the "sign" of the hp of tb_a and tb_b

$$\mathsf{tb}_a \bullet \mathsf{tb}_b = \begin{cases} +1, & \text{if } o(\mathsf{tb}_a) = o(\mathsf{tb}_b) \\ -1, & \text{if } o(\mathsf{tb}_a) = -o(\mathsf{tb}_b) \\ \varnothing, & \text{if } o(\mathsf{tb}_a) = \varnothing \text{ or } o(\mathsf{tb}_b) = \varnothing \end{cases}$$

In other words, the matrix construction is the same as for the MH of c_a and c_b, with tbs instead of genes and hps instead of gene homologies. The only difference is that while genes always have a known orientation, tbs can have *unknown* orientations that generate hps with signs equal to $\varnothing$. Similarly, the MHP can be represented as an array of values equal to $+1, -1, \varnothing$ or 0. Non-0 values correspond to hps. The X-axis corresponds to c_a ordered from $\mathsf{tb}_{a,1}$ to tb_{a,n_a} and the Y-axis corresponds to c_b ordered from $\mathsf{tb}_{b,1}$ to tb_{b,n_b}. With this convention $\mathsf{MHP}[0,0]$ corresponds to the bottom-left corner, $\mathsf{MHP}[n_a, 0]$ corresponds to the bottom-right corner, $\mathsf{MHP}[0, n_b]$ corresponds to the top-left corner and $\mathsf{MHP}[n_a, n_b]$ corresponds to the top right corner of the array.

[Additional file 1: Section 2] gives a graphical representation of the transition between the MH and the MHP via rewriting chromosomes with tbs.

Distances and gaps

The "gap between two tbs" on the same chromosome is the number of tbs between them.

The "distance between two tbs" is equal to the gap between these two tbs *plus one*. Thus two adjacent tbs are at a distance one from each other.

As in definition 2.1 in [35], a set of tbs forms a "chain" with gaps $\leq gap_{max}$ if all consecutive tbs are separated by gaps $\leq gap_{max}$ tbs.

Now, given a MHP, we define the "distance between two hps" as the 2D-distance between hps coordinates which depends on a distance metric. Several distance metrics can be used in PhylDiag: the Euclidean Distance (ED), the Chebyshev Distance (CD), the Manhattan Distance (MD), or the Diagonal Pseudo Distance (DPD) (Figure 2). Equations for each distance metric can be found in [Additional file 1: Section 3]. The CD yields the maximum of the distances on c_a and c_b, the ED yields the classical geometric distance, the DPD yields smaller distances between hps sitting close to the diagonal axis and therefore tends to provide a higher distance as the distance from the diagonal axis increases. In contrast, the MD tends to yield smaller distances between hps sitting close to the vertical and horizontal axis.

We define the "gap between two hps" as the distance between these hps *minus one*, thus a gap between two hps depends on the distance metric used. A gap of 0 between two hps means that there is no gap and this corresponds to a distance equal to 1. Given a maximum gap gap_{max}, a set of hps forms a "cluster" if no gap between them is longer than gap_{max}.

Step 3: Extract putative synteny blocks as consistent diagonals

In the following section, we define the notion of consistent diagonals in a MHP and we formally define synteny blocks. Then, we explain how synteny blocks generate consistent diagonals in MHPs, and we describe how PhylDiag extracts consistent diagonals. Because some consistent diagonals may be due to chance, we next describe how they are validated as synteny blocks after succeeding a statistical test.

Diagonals

In a MHP, a list of m hps $[\,\mathrm{MHP}[\,x_k, y_k]]_{k\in[0,m-1]}$ forms a:

– "slash" diagonal if $\begin{cases} x_{k+1} \geq x_k \\ y_{k+1} \geq y_k \end{cases} \quad \forall k \in [0, m-2].$

– "backslash" diagonal if $\begin{cases} x_{k+1} \geq x_k \\ y_{k+1} \leq y_k \end{cases} \quad \forall k \in [0, m-2].$

In both cases, x_k (respectively y_k) is the index of the homologous tb on c_a (respectively c_b) corresponding to the k^{th} hp. In a MHP, a "slash" diagonal is thus a list of non-0 cells that goes up according to a direction from bottom-left to top-right and a "backslash" diagonal is a list of non-0 cells that goes down according to a direction from top-left to bottom-right. A "diagonal" is either a slash diagonal or a backslash diagonal. A diagonal with gaps $\leq gap_{max}$ is a diagonal where all consecutive hps are separated by gaps $\leq gap_{max}$.

We define a "strict" diagonal as a diagonal that has no gap between its hps. Thus, m hps form a strict slash diagonal if the list of m hps can be written $[\mathrm{MHP}[s_a + k, s_b + k]]_{k\in[0,m-1]}$. Similarly, m hps form a strict backslash diagonal if the list of m hps can be written $[\mathrm{MHP}[s_a + k, s_b - k]]_{k\in[0,m-1]}$. In both cases, (s_a, s_b) is the position of the first hp of the diagonal.

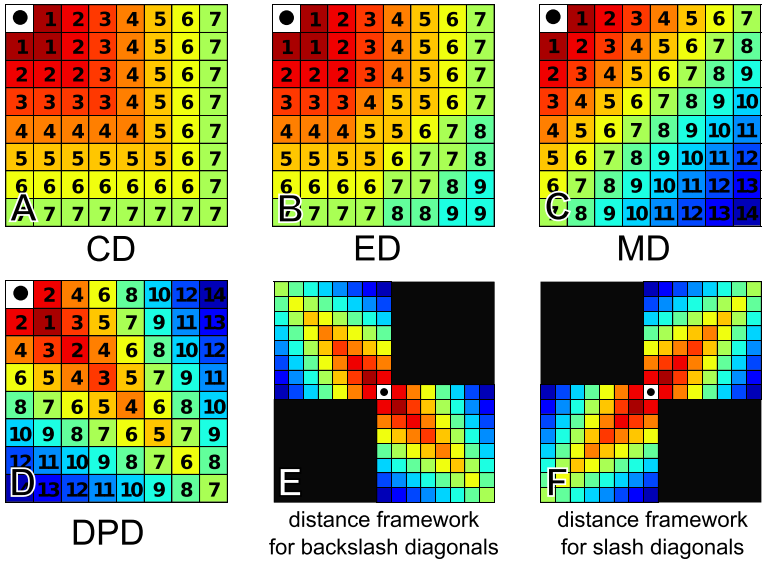

Figure 2 Distance metrics and frameworks used for the distance calculation. Figures **A**, **B**, **C** and **D** represent the metrics available in PhylDiag. Distance values are calculated starting from the black dot. The warmer the colour, the closer the point from the black dot. Considering that the user chose the DPD, when backslash diagonals are merged, the backslash framework of figure **E** is used for the distance calculation. For a slash diagonal merge, the framework of figure **F** is used. Frameworks would have been built in the same fashion if another metric had been chosen.

We also define a "consistent" diagonal as a diagonal composed of hps with signs consistent with hps order:

- either a slash diagonal only composed of hps with signs equal to either $+1$ or $\varnothing$
- or a backslash diagonal only composed of hps with signs equal to either -1 or $\varnothing$

In addition, we consider that the distance between two diagonals corresponds to the distance between their closest extremities.

Synteny block, formal definition

We formally define a "synteny block" of m tbs with gaps $\leq gap_{max}$ of a comparison of two genomes G_a and G_b, as a chain of m tbs with gaps $\leq gap_{max}$ that, during the evolution from $\mathsf{LCA}(S_a, S_b)$ to S_a and S_b, remains a chain of m tbs with gaps $\leq gap_{max}$. Within a synteny block, tbs order is conserved and tbs orientations either remain conserved or change from a known to an unknown orientation. Synteny blocks are chosen maximal, i.e. not included in another synteny block.

In addition we define a "strict synteny block" as a synteny block with no gaps between tbs ($gap_{max} = 0$).

In [Additional file 1: Section 4], we show that a strict synteny block generates a strict and consistent diagonal in a MHP. Following the reasoning in [Additional file 1: Section 4], using the CD distance metric, it is also possible to show that a synteny block with gaps $\leq gap_{max}$ generates a consistent diagonal with gaps $\leq gap_{max}$. However, using the ED, MD or the DPD distance metrics, a synteny block with gaps $\leq gap_{max}$ may generate a consistent diagonal with gaps $> gap_{max}$, although every consistent diagonal with gaps $\leq gap_{max}$ always represents a putative synteny block with gaps $\leq gap_{max}$. It should be noted that, given the CD distance metric and a gap_{max}, our definition of a diagonal is similar to the definition 4.1 of a "max-gap cluster" in [35] with constraints on gene order and gene orientations.

Extract strict consistent diagonals

Algorithm 1 describes how PhylDiag finds strict and consistent diagonals of hps in the MHP. First, chromosomes are rewritten with tbs and the MHP is built. Then the MHP is scanned from left to right and from bottom to top. Algorithm *findDiagType* in [Additional file 1: Section 5], sets the diagonal type at the beginning of a strict and consistent diagonal extraction using the sign of the first hp if the sign is known or using the position of the second hp if there is a second hp.

Strict and consistent diagonals are recorded as chains of ordered and oriented (whenever it is possible) ancestral genes. By convention the orientation of an ancestral gene $\mathsf{LCAg}(\mathsf{tb}_a, \mathsf{tb}_b)$ is chosen equal to the orientation of tb_a. However, if the orientation of tb_a is *unknown*, the orientation of $\mathsf{LCAg}(\mathsf{tb}_a, \mathsf{tb}_b)$ may still be inferred using the diagonal type of the current diagonal and a known orientation of tb_b, see Equation 1.

$$o\,(\mathsf{LCAg}\,(\mathsf{tb}_a, \mathsf{tb}_b)) = \begin{cases} o(\mathsf{tb}_a), & \text{if } o(\mathsf{tb}_a) \neq \varnothing \\ o(\mathsf{tb}_b), & \text{else if } o(\mathsf{tb}_b) \neq \varnothing \text{ and diagType} = \text{slash} \\ -o(\mathsf{tb}_b), & \text{else if } o(\mathsf{tb}_b) \neq \varnothing \text{ and diagType} = \text{backslash} \\ \varnothing, & \text{otherwise} \end{cases} \quad (1)$$

Algorithm 1 *extractSbs*(c_a, c_b)

inputs
  $c_a = [\mathsf{g}_{a,i}]_{i\in[1,N'_a]}$: a filtered chromosome of G_a
  $c_b = [\mathsf{g}_{b,j}]_{j\in[1,N'_b]}$: a filtered chromosome of G_b
rewrite c_a into $c_a = [\mathsf{tb}_{a,i}]_{i\in[1,n_a]}$
rewrite c_b into $c_b = [\mathsf{tb}_{b,j}]_{j\in[1,n_b]}$
define the matrix of homology packs $\mathsf{MHP} \in \mathfrak{M}_{n_a,n_b}$ of c_a and c_b
initialize diag $\leftarrow$[ ]: an empty diagonal
initialize listOfDiags $\leftarrow$[ ]: an empty list of diagonals
for all $i \in [1, n_a]$ **do**
  $i_{old} \leftarrow i$
  for all $j \in [1, n_b]$ **do**
    $i \leftarrow i_{old}$ // after extracting a diagonal, need to restart scanning from the next position
    while $\mathsf{MHP}[i,j] \neq 0$ **do**
      if diag is empty **then**
        diagType $\leftarrow$ *findDiagType*(MHP, (i,j))
      add $\mathsf{LCAg}(\mathsf{MHP}[i,j]) = \mathsf{LCAg}(\mathsf{tb}_{a,i}, \mathsf{tb}_{b,j})$, oriented using equation 1, **to** diag
      $\mathsf{MHP}[i,j] \leftarrow 0$ // for the following scanning process of MHP
      if *diagType* $=$ *slash* and $\mathsf{MHP}[i+1, j+1] = +1$ or $\varnothing$ **then**
        $i \leftarrow i+1$
        $j \leftarrow j+1$
      else if *diagType* $=$ *backslash* and $\mathsf{MHP}[i+1, j-1] = -1$ or $\varnothing$ **then**
        $i \leftarrow i+1$
        $j \leftarrow j-1$
      else
        add diag **to** listOfDiags
        diag $\leftarrow$[ ]
        break while
listOfDiags $\leftarrow$ *mergeDiags*(listOfDiags)
listOfSbs $\leftarrow$ *statisticalValidation*(listOfDiags)
return listOfSbs

Merge strict consistent diagonals

Once strict diagonals have been returned, it is advantageous to merge diagonals which have the same diagonal type, as long as their extremities are in close proximity. Depending on the allowed gap size gap_{max}, a limited number of errors of annotation and indels are thus allowed, and longer sbs are found that still reflect an ancestral arrangement of genes. It should be noted that this step possibly introduces micro-inversions within gaps of a diagonal, which will however always remain shorter than gap_{max} tbs. As we will see, the choice of the distance metric used to merge diagonals is crucial to limit or allow such micro-inversions, see [Additional file 1: Section 14].

The merging process is simple: diagonals are merged iteratively, starting by those separated by the shortest gap to those separated by the longest gap, as long as the gap remains below gap_{max}. For a given diagonal extremity, more than one other extremity may be situated at exactly the same distance. In this case, PhyDiag chooses to fuse the diagonals that maximise the number of hps in the diagonal that results from the fusion.

As described in the introduction, the DPD is used in ADHoRe and DiagHunter whereas the MD is used in GRIMM-Synteny, FISH, Cinteny and SyMAP. Although the CD and the ED have never been used to our knowledge in the context of synteny block inference we still included them in the benchmark presented in the 'Results' section.

Figure 3 shows an example of a merge between two strict backslash diagonals spaced by a distance 5 if the user chose the DPD, 4 if the user chose the MD and 3 if the user chose the CD or the ED.

Given a maximum gap gap_{max}, users should be aware that, with reference to the formal definition of sbs given in section 'Synteny block, formal definition', choosing another distance metric than the CD may return non-maximum sbs in the MHP. Another reason that may lead to non-maximum sbs may come from the fusion of diagonals. As mentioned before, if during the fusion process more than one diagonal extremity is available to extend the current diagonal, PhylDiag choses the extremity of the longest diagonal. However, it may be that fusing with a shorter one ultimately would lead to a longer diagonal once the iterative fusion process is complete.

In Algorithm 1, the merging process is encapsulated in the function *mergeDiags* that takes a list of strict and consistent diagonals and returns a list of consistent diagonals with gaps $\leq gap_{max}$.

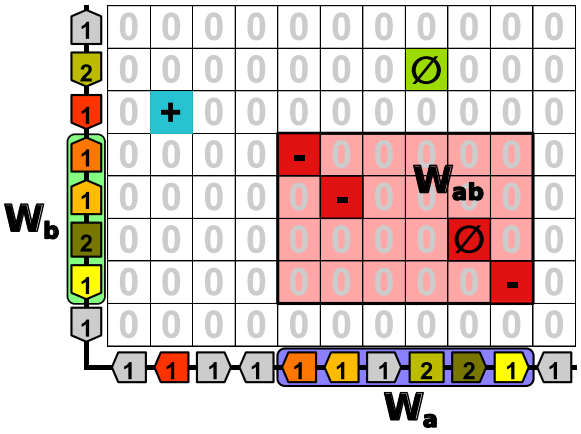

Figure 3 Example of a merge between two diagonals. Two chromosomes, c_a of $n_a = 11$ tbs and c_b of $n_b = 8$ tbs, are compared. The number in each tb is its size, arrows indicate tbs orientations and if a tb is represented as a rectangle it means that it has an *unknown* orientation. The MHP contains $n_{ab} = 6$ hps. During step 2, PhylDiag finds two strict backslash diagonals and 2 single hps. Each strict diagonal contains 2 hps. If $gap_{max} = 4$, during the diagonal merging process, diagonals separated by a distance of 2 to $gap_{max} + 1 = 5$ are merged. Consider that the user chooses the DPD metric, when reaching a distance of 5, two possible fusions are theoretically possible, one between the two extremities of the strict backslash diagonals and another fusion between the single hp (coordinates $[2, 6]$) and the extremity (coordinates $[5, 5]$) of the leftmost strict diagonal. However since the sign of the single hp is not consistent with the diagonal type of the left-most diagonal, the second fusion is not performed. Around the resulting consistent diagonal (in red) three windows are drawn: W_a (purple) on c_a contains 6 tbs and W_b (green) on c_b contains 4 tbs, and at last the window W_{ab} (pink) contains 6×4 cells and 4 hps. These windows are useful in section 'Step 4: Statistical validation of consistent diagonals as synteny blocks' for the p-value calculation.

Step 4: Statistical validation of consistent diagonals as synteny blocks

We compare two chromosomes, c_a and c_b. c_a has a length of n_a tbs, c_b has a length of n_b tbs and the comparison involves n_{ab} hps. During the comparison, PhylDiag returns many consistent diagonals that correspond to putative synteny blocks, each characterized by its number of hps, its window W_{ab} and the maximum gap g between its tbs (note that $g \neq gap_{max}$). Figure 3 shows an example of a consistent diagonal of 4 hps contained in the window W_{ab} with a maximum gap $g = 2$ tbs reached on c_a. The window W_{ab} has a size 6×4. The chromosomal windows W_a and W_b are the projections of W_{ab} on each chromosome. W_a has a length of $l_a = 6$ tbs and W_b has a length of $l_b = 4$ tbs. As in previous works [33-35], here distances and gaps between hps are calculated with the Chebyshev Distance metric which allows the most relaxed and method-independent sb definition.

A given consistent diagonal is a statistically significant signature of a sb if it cannot be obtained from a random distribution of tbs (null-hypothesis) up to a fixed probability threshold α. This is equivalent to selecting consistent diagonals that are unlikely to be the result of chance, which we wish to quantify here by a probability, a p-value.

We calculate the p-value of each consistent diagonal in five steps. Considering a consistent diagonal of m hps contained in a window W_{ab} of size $l_a \times l_b$ with a maximum gap between hps equal to g, the probability that such a consistent diagonal (or an even more improbable consistent diagonal with gaps $\leq g$) arises by chance is denoted $pVal(m,g,l_a,l_b,n_{ab},n_a,n_b)$. To compute this value we first compute $p_d(k,l_a,l_b,n_{ab},n_a,n_b)$ the probability of obtaining exactly k hps in the window W_{ab}, knowing the MHP density in terms of non-0 values. We next compute $p_{g,2\mathrm{D}}(k,g,l_a,l_b)$, the probability that k hps in W_{ab} are spaced with gaps $\leq g$, knowing that there is at least k hps in W_{ab}. We also calculate $p_{o,o}(k)$ the probability that k hps have consistent order and signs. By summing and multiplying these probabilities in an appropriate manner we calculate $p_w(m,g,l_a,l_b,n_{ab},n_a,n_b)$, the probability corresponding to a window sampling search. Finally, we use the former probability to compute the p-value $pVal(m,g,l_a,l_b,n_{ab},n_a,n_b)$ corresponding to a whole genome comparison. The formulas of the two first probabilities are based on [33,35] respectively and the passage from p_w to the $pVal$ is based on [34]. Here we combine these probabilities and add a last probability, $p_{o,o}(k)$, to account for tbs order and orientations.

Probability accounting for the density

Using the reasoning of [33], in a MHP of size $n_a \times n_b$ without dispersed paralogy (see Discussion), involving n_{ab} hps, the probability of obtaining exactly k hps in a window W_{ab} of size $l_a \times l_b$ is:

$$p_d\left(k,l_a,l_b,n_{ab},n_a,n_b\right) = \frac{\binom{n_{ab}}{k}\sum_{i=0}^{min(l_a-k,n_{ab}-k)}\binom{n_{ab}-k}{i}\binom{n_a-n_{ab}}{l_a-(k+i)}\binom{n_b-(k+i)}{l_b-k}}{\binom{n_a}{l_a}\binom{n_b}{l_b}} \quad (2)$$

The subscript d stands for *density* because this probability takes into account the density of the MHP. The demonstration of this formula is in [Additional file 1: Section 6].

Probability accounting for the maximum gap between hps

Using the reasoning of [35], the probability that k marked tbs (in any order) form a chain with gaps $\leq g$ *anywhere* within a window composed of l tbs is:

$$p_{g,1\mathrm{D}}\left(k,g,l\right) = \frac{1}{\binom{l}{k}}\begin{cases}\left(l+1-\frac{w_{kg}+k}{2}\right)(g+1)^{k-1}, & \text{if } w_{kg} \leq l+1\\ d_0\left(k,g,l\right), & \text{otherwise}\end{cases} \quad (3)$$

Where $w_{kg} = k + (k-1)g$ is the maximum length of a chain containing k tbs with a maximum gap g, and

$$d_0\left(k,g,l\right) = \sum_{i=0}^{|(l-k)/(g+1)|}(-1)^i\binom{k-1}{i}\binom{l-i(g+1)}{k} \quad (4)$$

the number of ways of arranging k tbs so that they form a chain with gaps shorter or equal to g anywhere within a window of l tbs even if $w_{kg} > l+1$, to address edge effects.

Thus, knowing that W_{ab} contains at least k hps, the probability that W_{ab} contains k marked hps spaced with gaps $\leq g$ is:

$$p_{g,2\mathrm{D}}\left(k,g,l_a,l_b\right) = p_{g,1\mathrm{D}}\left(k,g,l_a\right) \times p_{g,1\mathrm{D}}\left(k,g,l_b\right) \quad (5)$$

Probability accounting for hps order and signs

Then, if k hps are close enough, the probability that they form a consistent slash diagonal with gaps $\leq g$ is:

$$p_{slash}\left(k\right) = \frac{1}{k!}\left[P(sign = +1 \text{ or} \varnothing)\right]^k \quad (6)$$

Where $P(sign = +1 \text{ or } \varnothing) = P(sign = +1) + P(sign = \varnothing)$ and $P(sign = s)$ is the probability that one hp sign equals s, this probability calculation is explained in [Additional file 1: Section 7]. $\frac{1}{k!}$ is the probability that k homologous tbs of chromosome c_b have the *same* order as the corresponding k homologous tbs of chromosome c_a and $[P(sign = +1 \text{ or } \varnothing)]^k$ is the probability that the k signs of the hps are consistent with a slash diagonal. $p_{backslash}(k)$ is defined similarly.

Thus, if k hps are close enough, the probability that they form a consistent diagonal with gaps $\leq g$ is:

$$p_{o,o}(k) = \begin{cases}1, & \text{if } k = 1\\ p_{slash}(k) + p_{backslash}(k), & \text{otherwise}\end{cases} \quad (7)$$

The subscript *o,o* stands for consistent tbs *Order* and tbs *Orientations*. The demonstration of the $p_{o,o}$ formula can be found in [Additional file 1: Section 8].

Probability for a window sampling scenario

Now, in a MHP of size $n_a \times n_b$ without dispersed paralogy (see Discussion), involving n_{ab} hps, the probability that in a window W_{ab} of size $l_a \times l_b$ there is *at least* one consistent diagonal containing *at least* m hps spaced with gaps $\leq g$ is:

$$p_w(m,g,l_a,l_b,n_{ab},n_a,n_b) = \sum_{k=m}^{min(n_{ab},l_a,l_b)} p_d(k)\sum_{i=m}^{k} p_{g,2\mathrm{D}}(i)p_{o,o}(i) \quad (8)$$

The subscript w stands for *Window* because this probability corresponds to a window sampling [34] scenario. Only varying parameters are shown in the right-hand side

of the equation in the preceding formula. This formula is explained in [Additional file 1: Section 9].

Probability for a whole chromosome comparison

Finally, since PhylDiag performs a whole chromosome comparison, it is not possible to use the probability of a window sampling method that would underestimate the probability to find a consistent diagonal by a factor of $O(n_a n_b)$. Thus, relying on the reasoning of section 4.2 of [34] we adjust the former probability to compute the probability corresponding to a whole chromosome comparison.

In a MHP of size $n_a \times n_b$ containing n_{ab} hps without dispersed paralogy (see Discussion), the probability of finding *at least* one window W_{ab} of size $l_a \times l_b$ containing *at least* a consistent diagonal of *at least* m hps spaced by gaps $\leq g$ can be approximated by:

$$pVal\left(m, g, l_a, l_b, n_{ab}, n_a, n_b\right) \simeq 1 - (1 - p_w)^{n_w} \qquad (9)$$

where $n_w = \frac{n_a n_b}{l_a l_b}$ is the number of windows of width l_a and height l_b in the MHP such that no window overlap with any other window. The underlying assumption of this formula is justified in [Additional file 1: Section 10] and examples of calculation are performed in [Additional file 1: Section 11].

In Algorithm 1, the statistical validation is encapsulated in the function *statisticalValidation* that takes a list of consistent diagonals as input and returns statistically validated sbs.

Estimation of a recommended maximum gap parameter

All algorithms designed to identify synteny blocks use a maximum gap parameter (gap_{max}) to allow gaps in sbs. However, the user may find it difficult to estimate the optimal value for this parameter. In order to avoid guessing or multiple trials before finding the optimal gap_{max} value, PhylDiag uses the dependency between the probability of finding a consistent diagonal of m hps spaced by gaps $\leq gap_{max}$ and the gap_{max} value. The complete reasoning used to calculate the recommended maximum gap parameter can be found in [Additional file 1: Section 12].

Viewer

PhylDiag includes a utility to visualise the MHP of a pairwise comparison of chromosomes with colours and surrounding black rectangles for sbs recognition. This viewer writes a vectorial image allowing an infinite zoom on details with no pixelisation. Figure 4A shows an example of the viewer during the comparison of the human X chromosome with the mouse X chromosome. If more information about a region of the MHP is required, a zoom can be performed by specifying the desired chromosomal regions. If these are small enough, more information is shown, such as hps signs, oriented tbs on each axis, the size of each tb and colours for homology recognition. Grey tbs represent tbs that do not have hps in the MHP, but they have hps elsewhere in the pairwise

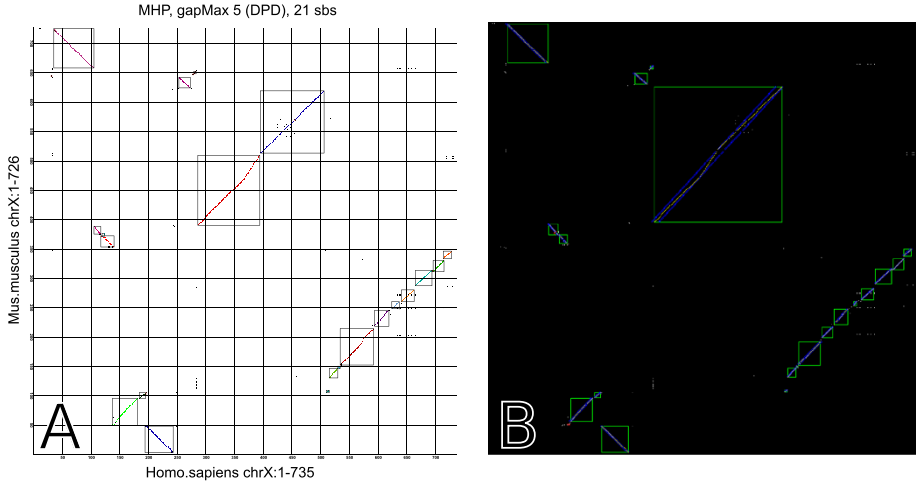

Figure 4 Representations of a comparison between the human and the mouse X chromosomes, produced from the same input data by PhylDiag (A) and i-ADHoRe 3.0.2a (B). The maximum gap parameter gap_{max} is equal to 5 and the merging process used the DPD metric in both cases. In figure A each axis displays explicitly the paths to the files containing the relevant genome data, the name of the chromosome and the chromosomal window range. As in ADHoRe sbs identified by PhylDiag are surrounded by a rectangle and each sb is drawn in a specific colour. In figure B blue dots represent a confidence interval around each sb drawn in yellow. The long synteny block in the middle of the ADHoRe MHP is in two parts in the PhylDiag MHP because the two extremities are spaced by a distance higher than 5 using the DPD metric. By default ADHoRe has a special feature using extremities of diagonals during its merging process, however this feature cannot be deactivated which may lead to undesired merges.

comparisons of genomes, in another pairwise comparison of chromosomes. Figure 3 was produced with the viewer and shows such informations. The user may also visualise the MH, for example to study the genic composition of a tb.

Implementation

The complete algorithm has been implemented in Python. Pairwise comparisons of chromosomes are performed in parallel since they are independent. In Algorithm 1, the MHP matrix is stored considering that it is a sparse matrix to reduce memory usage and the merging process is optimised. Combinations in probability formulas are computed using Pascal's rule and dynamic programming. On a single 3,0 GHz processor with 32 Gb RAM, loading the data in memory requires 3 seconds, and the running time for the pairwise analysis of the Human and Mouse genomes requires less than 3 seconds. Without any optimisation of the memory allocations the peak of RAM consumption is 221 Mb, thus a standard personal computer can run PhylDiag.

Results

To evaluate the performances of PhylDiag, we performed a comparative analysis with i-ADHoRe 3.0 [14], a state-of-the-art algorithm used in many recent studies. To make comparisons possible however, we used a version of the program provided by the authors. Indeed, i-ADHoRe 3.0 first rewrites genomes in tbs like PhylDiag, but allows a user-defined "tandem_gap" between genes in a tb. In version 3.0, the minimal tandem_gap is 2, and it is not possible to set the tandem_gap to 0, as in PhylDiag. In the version provided (i-ADHoRe 3.0.2a) this option is enabled.

When ADHoRe compares two chromosomes, it first generates "baseclusters" which correspond to PhylDiag's sbs. ADHoRe uses the DPD metric to build baseclusters containing gaps $\leq$"gap_size" in the matrix of homologies. ADHoRe also uses the "prob_cutoff" parameter for the statistical filtering and a last parameter is the "q_value", a real value between 0 and 1, indicating the minimum r^2 (a measure for the linearity of baseclusters in the matrix of homologies) that a cluster should display.

Comparison with i-ADHoRe 3.0.2a on real data

In a first comparison, we provided the same input based on real genomic data to PhylDiag and ADHoRe. We used the human genome (G_h) and the mouse genome (G_m) of Ensembl v72. As explained in section 'Gene family and homology', families correspond to genes that are descended from a unique gene of $\mathsf{LCA}(S_h, S_m)$ = Euarchontoglire.

PhylDiag computes a recommended gap_{max} of 5 tbs for the human-mouse comparison. We therefore set a gap_{max} parameter of 5 and we chose a probability threshold $\alpha = 1 \times 10^{-3}$ for PhylDiag. i-ADHoRe 3.0.2a was set with tandem_gap=0, gap_size=5, prob_cutoff=1×10^{-3} and q_value=0.9 (the default value). Figure 5 compares

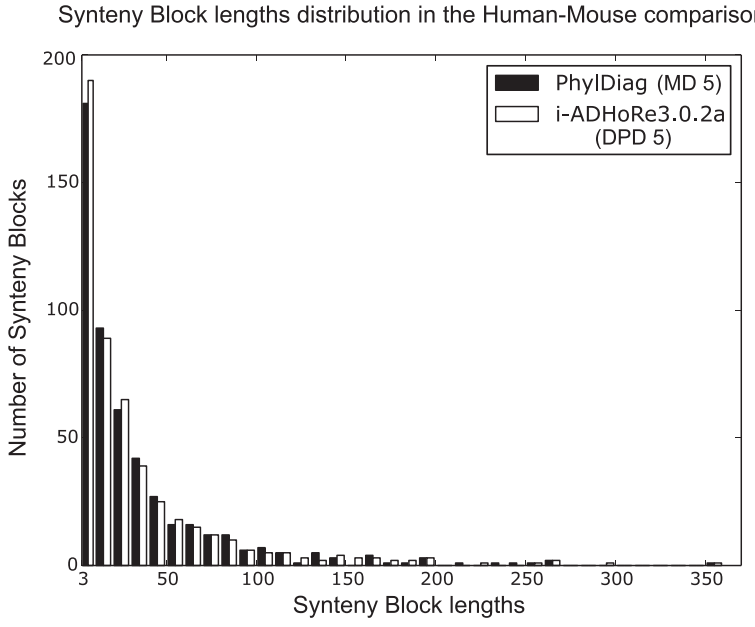

Figure 5 Synteny block lengths distribution as computed by PhylDiag (black) and ADHoRe (white). Each bin has a width of 10 units of length (hps), apart from the first bin that contains sb lengths from 3 to 10 hps. ADHoRe uses the DPD distance metric whereas PhylDiag uses the MD distance metric.

the distributions of synteny block lengths of ADHoRe and PhylDiag using the MD distance metric. The two distributions are not different from each other (Mann-Whitney U test: pval = 0.9791), and show that neither methods suffer from strong biases in over or under detection of synteny blocks in a given size range. Of note, PhylDiag returned 17 significant sbs of 2 hps out of 175 consistent diagonals of 2 hps. These are not shown in Figure 5 because ADHoRe does not report sbs of size 2. PhylDiag statistically validated all consistent diagonals containing more than 2 hps as significant synteny blocks.

Comparison with i-ADHoRe 3.0.2a on simulated data

Our simulator first designs an ancestral genome G_{anc} with a user defined number of genes and chromosomes. The lengths of chromosomes in G_{anc} are expressed in number of genes, and are determined randomly. Simulated evolution gives rise to the two extant genomes G_a and G_b of two extant species. The simulator performs genic events, which include de novo gene births, deletions, duplications (tandem and dispersed), and genomic rearrangements, which include chromosome fusions and fissions, segmental translocations or segmental inversions. The evolutionary scenario is calibrated so as to fit the known evolution of the human and the mouse genome from the Euarchontoglire genome using phylogenetic gene tree reconstructions from Ensembl Compara version 72. See [Additional file 1: Section 13] for a more detailed description of the Simulator.

We performed 100 simulations of the evolution of the human and the mouse genome, and analysed them with PhylDiag and ADHoRe to identify sbs. The PhylDiag merging process was performed with the 4 different distance metrics (ED, CD, DPD and MD). For ADHoRe the DPD is the only distance metric available. As in the comparison with real data, since the simulation is calibrated to fit real evolutionnary rates, the recommended gap_{max} found by PhylDiag is still 5. Results of PhylDiag with a $gap_{max} = 5$ and ADHoRe with $gap\_size = 5$ are shown in Table 1.

Coverage is the fraction of the number of gene families (each family corresponds to a single ancestral gene of Euarchontoglire) contained in sbs over the total number of ancestral genes conserved in both the simulated human genome and the simulated mouse genome. *N50* is the length of the sb such that all sbs of greater lengths represent 50% of the ancestral genes contained in sbs. *Sensitivity* is the fraction of the number of correctly inferred ancestral adjacencies over the total number of ancestral genes conserved in both the simulated human genome and the simulated mouse genome. *Specificity* is the fraction of the number of correctly inferred ancestral adjacencies over the total number of inferred ancestral adjacencies, false inferences included.

Table 1 Results of synteny block identification with PhylDiag and i-ADHoRe3.0.2a, both using a $gap_{max} = 5$

Algorithm	PhylDiag (without sbs of 2 hps)				ADHoRe
distance	ED	CD	DPD	MD	DPD
coverage	98.71%	98.74%	97.02%	98.55%	96.55%
N50	44.69	46.62	32.66	37.33	31.71
Analysis without gene orientations					
sensitivity	94.99%	95.06%	92.26%	94.32%	91.68%
specificity	99.92%	99.85%	99.90%	99.98%	99.83%
Analysis with gene orientations					
sensitivity	94.20%	94.26%	91.56%	93.54%	88.56%
specificity	99.08%	99.01%	99.13%	99.15%	96.43%

Since ADHoRe only returns sbs containing at least 3 hps, we only consider PhylDiag's sbs containing at least 3 hps.

Specificity and sensitivity are calculated twice: first by ignoring gene orientations (an inferred adjacency between two genes is considered correct if both genes are adjacent in G_{anc} even if their relative orientation is different compared to the ancestral relative orientation), and second by taking gene orientations into account (to be correct an inferred adjacency must contain genes with a relative orientation that is the same as in the G_{anc}).

Results show that PhylDiag with the DPD, and ADHoRe obtain similar results when we do not consider gene orientations during the analysis. Interestingly, simply using the ED, the CD or the MD metrics allows PhylDiag to achieve better sensitivity and specificity than ADHoRe (Mann-Whitney U test on sensitivity % and specificity % using the MD in PhylDiag and DPD in ADHoRe over 100 simulations: pval $\leq$ 2.2e-16 and pval $\leq$ 2.2e-16 respectively). In addition, as soon as gene orientations are considered in the analysis, PhylDiag improves substantially, in part because of Equation 1.

Discussion

We have compared PhylDiag to i-ADHoRe 3.0, a state-of-the-art algorithm including advanced features which are not present in PhylDiag, including the possibility to identify sbs in the "twilight zone", i.e. sbs highly diverged or separated by a WGD, where many gene deletions may have occurred. ADHoRe uses "profiles" across more than two genomes to identify poorly conserved sbs, for example due to long divergence times. These features were not exploited here because unlike ADHoRe, PhylDiag only performs pairwise comparisons of genomes since our primary interest is to identify sbs in closely related species.

We explored different distance metrics to measure distances in matrices of homology, and found that the DPD used in ADHoRe, which favours fusions of diagonals along $\pm 45°$ axes in the MHP (Figures 2D, 2E and 2F), is not

optimal. This has been discussed previously [23] and the simulations clearly show that exploring first laterally (i.e. vertically and horizontally), as with the MD (Figure 2C), improves results. Merging diagonals with the DPD distance metric allows more small inversions within sbs gaps while considering that genic/segmental indels and incorrect annotations break the synteny more easily than with the MD. Conversely, merging diagonals with the MD metric gives priority to lateral directions and this allows more small genic/segmental indels and annotation errors within sbs gaps and considers that inversions break the synteny more easily than with the DPD, see [Additional file 1: Section 14]. Interestingly the unusual ED or CD distance metrics also show improved results over the DPD (Table 1). It should be noted that a given distance may cover a different number of cells in the MHP depending on the metric chosen. For instance 9 cells are covered within a distance value of 3 with the MD whereas 7 cells are covered within the same distance value of 3 with the DPD (Figures 2C and 2D). Although this bias may play a role in the results, on chromosomes, gaps between tbs involved in pairs of chains corresponding to sbs are always smaller or equal to gap_{max} independently of the metric chosen. Thus comparing metrics is fair. Finally, contrary to ADHoRe, PhylDiag can return sbs containing 2 hps if their p-value is under the p-value threshold of the user.

PhylDiag includes a new statistical validation to estimate the probability that a putative sb may be due to chance. Unlike other tests, it accounts for gene orientations, thus providing increased sensitivity. It also accounts for tandem duplications but ignores the possibility that duplicate gene copies may be dispersed. Neglecting dispersed duplicates underestimates the p-values of sbs and the significance of sbs are thus overestimated. However models considering gene families exist [8,23] and in a future version it might be advantageous to implement the p-value proposed in [36], even if the calculation is based on an unrealistic assumption that all gene families are of fixed size. Nevertheless the error in the p-value calculation in PhylDiag is likely to be small for closely related species. For instance the analysis of phylogenetic trees described here shows that only 2.4% of tbs are dispersed duplicates in the human genome (3.2% in mouse) using our family definition (section 'Gene family and homology').

The p-value used by PhylDiag is relative to a comparison of two chromosomes, and therefore assumes that random consistent diagonals might arise based on the number of tbs and hps relevant to the two chromosomes only. In contrast, a global (i.e. genome wide) threshold α is chosen to distinguish significant sbs from non-significant sbs. This inconsistency represents an area of further development, in order to better account for heterogeneous densities of hps depending on which chromosomes are being compared.

Conclusion

PhylDiag is designed around a heuristic-independent formal definition of synteny blocks. Its implementation and benchmarking using real and simulated data allowed us to rank 2D-distance metrics in terms of sensitivity and specificity, and to evaluate its performance in comparison with ADHoRe. Results show that the DPD distance metric yields the poorest performances when identifying synteny blocks, both with ADHoRe and PhylDiag. In contrast, PhylDiag highlights the interesting sensitivity-specificity trade-off achieved by the MD distance metric, closely followed by the CD and the ED distance metrics. Compared to ADHoRe and other algorithms that infer synteny blocks, the definition of gene families in PhylDiag is based on gene trees. Most notably, this feature offers the opportunity to precisely group extant genes into families that descend from a unique gene in the last common ancestor of the two species being compared. Furthermore, a meticulous attention to tandem duplicates and gene orientations allow PhylDiag to reach a high resolution in the analysis of rearrangements, down to single gene inversion. Finally, the statistical validation of putative synteny blocks filters out putative false positives due to randomly convergent gene order. PhylDiag is a software for synteny block inference that benefits from extensive parameters, including gap_{max}, distance metric, p-value threshold, filtering of extant genomes and ancestor for the gene family definition. Their values can be set by PhylDiag (default values are based on previous benchmarks or set automatically based on the data) or set by the user. These features, together with post-processing graphical analysing tools and printed statistics (number of tandem duplicates in extant genomes, number of dispersed duplicates, number of homologies involved in the pairwise comparison) contribute to making PhylDiag a user-friendly method to find synteny blocks.

Availability and requirements

Project name: PhylDiag
Project home page: https://github.com/DyogenIBENS/PhylDiag
Operating system(s): Platform independent
Programming language: Python
Other requirements: Python 2.7 or higher
License: GNU GPL v3 or later, and the CeCiLL v2 license in France
Any restrictions to use by non-academics: No

Additional file

Additional file 1: Supplementary Data.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JMEXL designed the core of PhylDiag. MM designed the library framework for genome data manipulation and gene tree editing. HRC supervised the work. JMEXL and HRC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Alexandra Louis and Pierre Vincens for their help with computer facilities and Wassim Abou-Jaoudé for his assistance with mathematical problems. We also thank Yves Van de Peer, Klaas Vandepoele and Jan Fostier, the authors of ADHoRe who provided the version i-ADHoRe 3.0.2a. This work is funded by Centre National de la Recherche Scientifique (CNRS) and grants from the Agence Nationale de la Recherche (ANR) [Ancestrome Project ANR-10-BINF-01-03, ANR Blanc-PAGE ANR-2011- BSV6-00801].

Author details
[1]Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, 46 rue d'Ulm, 75005 Paris, France. [2]CNRS, UMR 8197, 75005 Paris, France. [3]Inserm, U1024, 75005 Paris, France. [4]The EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Düsternbrooker Weg 20, CB10 1SD Hinxton, Cambridge, UK.

Received: 21 March 2014 Accepted: 17 July 2014
Published: 8 August 2014

References

1. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers SN, Milan D, Ostrander Ea, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'brien SJ, Pevzner Pa, Lewin Ha: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps.** *Science* 2005, **309**(5734):613–7. doi:10.1126/science.1111387.
2. Chauve C, Tannier E: **A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes.** *PLoS Comput Biol* 2008, **4**(11):1000234. doi:10.1371/journal.pcbi.1000234. PMID:19043541.
3. Darai-Ramqvist E, Sandlund A, Müller S, Klein G, Imreh S, Kost-Alimova M: **Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions.** *Genome Res* 2008, **18**(3):370–9. doi:10.1101/gr.7010208.
4. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates A. C, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** *Genome Res* 2007, **17**(5):545–55. doi:10.1101/gr.6086307.
5. Irimia M, Tena JJ, Alexis M. S, Fernandez-Miñan A, Maeso I, Bogdanovic O, de la Calle-Mustienes E, Roy SW, Gómez-Skarmeta JL, Fraser HB: **Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints.** *Genome Res* 2012, **22**(12):2356–67. doi:10.1101/gr.139725.112.
6. Bergeron A, Corteel S, Raffinot M: **The algorithmic of gene teams.** *Algorithms Bioinformatics* 2002, **2452:**464–476. doi:10.1007/3-540-45784-4_36.
7. Luc N, Risler J-L, Bergeron A, Raffinot M: **Gene teams: a new formalization of gene clusters for comparative genomics.** *Comput Biol Chem* 2003, **27**(1):59–67. PMID:12798040.
8. He X, Goldwasser MH: **Identifying conserved gene clusters in the presence of homology families.** *J Comput Biology: J Comput Mol Cell Biol* 2005, **12**(6):638–656. doi:10.1089/cmb.2005.12.638. PMID:16108708.
9. Boyer F, Morgat A, Labarre L, Pothier J, Viari A: **Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data.** *Bioinformatics* 2005, **21**(23):4209–4215. doi:10.1093/bioinformatics/bti711. PMID:16216829. Accessed 2013-10-24.
10. Hampson SE, Gaut BS, Baldi P: **Statistical detection of chromosomal homology using shared-gene density alone.** *Bioinformatics (Oxford, England)* 2005, **21**(8):1339–1348. doi:10.1093/bioinformatics/bti168. PMID:15585535.
11. Ling X, He X, Xin D: **Detecting gene clusters under evolutionary constraint in a large number of genomes.** *Bioinformatics (Oxford, England)* 2009, **25**(5):571–577. doi:10.1093/bioinformatics/btp027. PMID:19158161.
12. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W: **Reconstructing contiguous regions of an ancestral genome.** *Genome Res* 2006, **16**(12):1557–1565. doi:10.1101/gr.5383506. Accessed 2013-06-18.
13. Tesler G: **GRIMM genome rearrangements web server.** *Bioinformatics* 2002, **18**(3):492–493. doi:10.1093/bioinformatics/18.3.492. PMID:11934753. Accessed 2013-06-18.
14. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de, Peer Y, Vandepoele K: **i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets.** *Nucleic Acids Res* 2012, **40**(2):11. doi:10.1093/nar/gkr955.
15. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y: **Building genomic profiles for uncovering segmental homology in the twilight zone.** *Genome Res* 2004, **14**(6):1095–1106. doi:10.1101/gr.2179004. Accessed 2013-07-08.
16. Van de Peer Y, Meyer A: **Chapter 6 - large-scale gene and ancient genome duplications.** In *The Evolution of the Genome*. Edited by Gregory TR. Burlington: Academic Press; 2005:340–344. http://www.sciencedirect.com/science/article/pii/B9780123014634500085. Accessed 2013-09-28.
17. Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y: **The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between arabidopsis and rice.** *Genome Res* 2002, **12**(11):1792–1801. doi:10.1101/gr.400202. PMID:12421767.
18. Cannon SB, Kozik A, Chan B, Michelmore R, Young ND: **DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization.** *Genome Biol* 2003, **4**(10):68. doi:10.1186/gb-2003-4-10-r68. PMID:14519203.
19. Hampson S, McLysaght A, Gaut B, Baldi P: **LineUp: statistical detection of chromosomal homology with application to plant comparative genomics.** *Genome Res* 2003, **13**(5):999–1010. doi:10.1101/gr.814403. PMID:12695327.
20. Calabrese PP, Chakravarty S, Vision TJ: **Fast identification and statistical evaluation of segmental homologies in comparative maps.** *Bioinformatics* 2003, **19**(suppl 1):74–80. doi:10.1093/bioinformatics/btg1008. PMID:12855440. Accessed 2013-06-21.
21. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny.** *Bioinformatics* 2004, **20**(18):3643–3646. doi:10.1093/bioinformatics/bth397. PMID:15247098. Accessed 2013-06-21.
22. Soderlund C, Nelson W, Shoemaker A, Paterson A: **SyMAP: a system for discovering and viewing syntenic regions ofFPC maps.** *Genome Res* 2006, **16**(9):1159–1168. doi:10.1101/gr.5396706. PMID:16951135. Accessed 2013-06-21.
23. Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J: **Statistical inference of chromosomal homology based on gene colinearity and applications to arabidopsis and rice.** *BMC Bioinformatics* 2006, **7:**447. doi:10.1186/1471-2105-7-447. Accessed 2013-06-20.
24. Sinha AU, Meller J: **Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms.** *BMC Bioinformatics* 2007, **8**(1):82. doi:10.1186/1471-2105-8-82. PMID:17343765. Accessed 2013-06-19.
25. Zeng X, Nesbitt MJ, Pei J, Wang K, Vergara IA, Chen N: **OrthoCluster.** In *Proceedings of the 11th International Conference on Extending Database Technology Advances in Database Technology - EDBT '08*. New York, USA: ACM Press; 2008:656. doi:10.1145/1353343.1353423. [http://portal.acm.org/citation.cfm?doid=1353343.1353423]
26. Rödelsperger C, Dieterich C: **Syntenator: Multiple gene order alignments with a gene-specific scoring function.** *Algorithms Mol Biol* 2008, **3**(1):14. doi:10.1186/1748-7188-3-14. PMID:18990215. Accessed 2013-06-21.
27. Rödelsperger C, Dieterich C: **CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes.** *PLoS ONE* 2010, **5**(1):8861. doi:10.1371/journal.pone.0008861. Accessed 2013-06-21.

28. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps.** *Genome Res* 2008, **18**(12):1944–1954. doi:10.1101/gr.080978.108. PMID:18832442. Accessed 2013-10-23.
29. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH: **MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** *Nucleic Acids Res* 2012, **40**(7):49. doi:10.1093/nar/gkr1293. PMID:22217600.
30. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E: **Enredo and pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs.** *Genome Res* 2008, **18**(11):1814–1828. doi:10.1101/gr.076554.108. Accessed 2013-09-10.
31. Pham SK, Pevzner PA: **DRIMM-Synteny: decomposing genomes into evolutionary conserved segments.** *Bioinformatics (Oxford, England)* 2010, **26**(20):2509–16. 0doi:10.1093/bioinformatics/btq465.
32. Smith TF, Waterman MS, Subsequences CM: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**(1):195–7.
33. Raghupathy N, Hoberman R, Durand D: **Two plus two does not equal three: statistical tests for multiple genome comparison.** *J Bioinform Comput Biol* 2008, **6**(1):1–22.
34. Durand D, Sankoff D: **Tests for gene clustering.** *J Comput Biol: J Comput Mol Cell Biol* 2003, **10**(3-4):453–82. doi:10.1089/10665270360688129.
35. Hoberman R, Sankoff D, Durand D: **The statistical analysis of spatially clustered genes under the maximum gap criterion.** *J Comput Biol: J Comput Mol Cell Biol* 2005, **12**(8):1083–1102. doi:10.1089/cmb.2005.12.1083. PMID:16241899.
36. Raghupathy N, Durand D: **Gene cluster statistics with gene families.** *Mol Biol Evol* 2009, **26**(5):957–68. doi:10.1093/molbev/msp002.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–10. doi:10.1016/S0022-28360580360-2.
38. Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DST, Humphrey J, Kerhornou A, Khobova J, Langridge N, McDowall MD, Maheswari U, Maslen G, Nuhn M, Ong CK, Paulini M, Pedro H, Toneva I, Tuli MA, Walts B, Williams G, Wilson D, Youens-Clark K, Monaco MK, Stein J, Wei X, Ware D, Bolser DM, Howe KL, et al: **Ensembl Genomes 2013 scaling up access to genome-wide data.** *Nucleic Acids Res* 2014, **42**(Database issue):546–52. doi:10.1093/nar/gkt979.
39. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19**(2):327–335. doi:10.1101/gr.073585.107. Accessed 2013-09-29.

doi:10.1186/1471-2105-15-268
Cite this article as: Lucas *et al.*: PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 2014 **15**:268.