BIOINFORMATICS

Vol. 30 ISMB 2014, pages i9–i18
doi:10.1093/bioinformatics/btu259

Evaluating synteny for improved comparative studies

Cristina G. Ghiurcuta* and Bernard M. E. Moret

Laboratory for Computational Biology and Bioinformatics, EPFL-IC-LCBB INJ 230, Station 14, CH-1015 Lausanne, Switzerland

ABSTRACT

Motivation: Comparative genomics aims to understand the structure and function of genomes by translating knowledge gained about some genomes to the object of study. Early approaches used pairwise comparisons, but today researchers are attempting to leverage the larger potential of multi-way comparisons. Comparative genomics relies on the structuring of genomes into *syntenic blocks*: blocks of sequence that exhibit conserved features across the genomes. Syntenic blocs are required for complex computations to scale to the billions of nucleotides present in many genomes; they enable comparisons across broad ranges of genomes because they filter out much of the individual variability; they highlight candidate regions for in-depth studies; and they facilitate whole-genome comparisons through visualization tools. However, the concept of syntenic block remains loosely defined. Tools for the identification of syntenic blocks yield quite different results, thereby preventing a systematic assessment of the next steps in an analysis. Current tools do not include measurable quality objectives and thus cannot be benchmarked against themselves. Comparisons among tools have also been neglected—what few results are given use superficial measures unrelated to quality or consistency.

Results: We present a theoretical model as well as an experimental basis for comparing syntenic blocks and thus also for improving or designing tools for the identification of syntenic blocks. We illustrate the application of the model and the measures by applying them to syntenic blocks produced by three different contemporary tools (DRIMM-Synteny, i-ADHoRe and Cyntenator) on a dataset of eight yeast genomes. Our findings highlight the need for a well founded, systematic approach to the decomposition of genomes into syntenic blocks. Our experiments demonstrate widely divergent results among these tools, throwing into question the robustness of the basic approach in comparative genomics. We have taken the first step towards a formal approach to the construction of syntenic blocks by developing a simple quality criterion based on sound evolutionary principles.

Contact: cristinagabriela.ghiurcuta@epfl.ch

1 BACKGROUND

Comparative studies have long been the mainstay of knowledge discovery in biology. With the advent of inexpensive sequencing tools, pairwise sequence comparison became a major research tool; programs such as BLAST (Altschul *et al.*, 1990) are used to identify regions with similar sequences in order to study problems in genetics and genomics by using knowledge from better characterized organisms. Such comparisons have been carried out on relatively short sequence fragments—usually up to the length of a protein transcript, i.e. a few thousand nucleotides. Such work continues at a great pace today, but the rapidly increasing availability of complete genome sequences has led to the desire to compare entire genomes at once, the better to understand the large-scale architectural features of genomes and the evolutionary events that have shaped these features, such as segmental and whole-genome duplication, horizontal transfer, recombinations of various types and rearrangements.

Comparing entire genomes is not new: almost a century ago, Thomas Morgan and his students used chromosomal banding to build genetic maps of various strains of *Drosophila melanogaster*. What is new today is the possibility of comparing complete genome sequences to each other. Comparing even just two genomes is a major computational challenge when the two genomes have several billion nucleotides and when most of the sequence (>90% in humans) is poorly understood and so lacks a suitable evolutionary model. Consequently, researchers have approached the problem by defining (or searching for) conserved sequence markers (mostly belonging to the better understood coding regions of the genome). These markers are then used to form large-scale patterns that can be evaluated for similarity and conservation. Such large-scale patterns, when used systematically, can be viewed as alternative representations of the genomes. The simplest such representation uses the concept of *syntenic blocks* (SBs), large blocks of sequence that are well conserved (as testified by commonality of markers and similarity of high-level patterns) across the species (or within a genome). Working with such blocks facilitates comparative studies: (i) it confers robustness against variability across individuals and against various sources of error; (ii) it reduces the dependence on an accepted model of sequence evolution for each region and is less likely to suffer from homoplasy; (iii) it reduces the complexity of the analysis of the genomic structures; (iv) it provides high-level features for further evolutionary studies; and (v) it identifies specific regions of interest for detailed studies and possible bench experiments.

In this article, we provide a concise overview of the existing notions of synteny in the literature and propose a formal, principled definition of SBs based on homologies. We discuss how the quality of SBs can be measured against this definition and illustrate our approach with a comparison of three current tools for the construction of SBs—Cyntenator (Roedelsperger and Dieterich, 2010), DRIMM-Synteny (Pham and Pevzner, 2010) (DRIMM) and i-ADHoRE 3.0 (Proost *et al.*, 2012) (i-ADHoRe). We investigate the underlying heuristics and evaluate the results on a dataset of eight full genomes of various species of yeasts from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005), pointing out the issues that arise when working with SBs.

1.1 Early notions of synteny

Little has been done so far towards a formal definition of SBs and/or SB families, nor have developers of algorithms and

*To whom correspondence should be addressed.

software for producing SBs given any quantifiable goals. Instead, identifying SBs has been a matter of application-dependent heuristics, lacking any serious attempt at evaluating the quality of the approaches—something that in any case would have proved difficult in absence of quality criteria. The first mention of synteny as it is understood today was in an article of Renwick (1971) on human chromosome mapping, where the term is introduced to denote collocation of markers on the same chromosome. Nadeau and Taylor (1984) gave an informal definition of syntenic segments, in a paper that has since been cited by most researchers concerned with synteny. Nadeau and Taylor gave a list of features viewed as supporting inclusion of markers in an SB, a list that includes conserved orientation, conserved adjacency and conserved position of homologous markers associated with the corresponding mapped chromosomes, a collection of features that loosely defines what is more commonly called today *collinearity*.

The study of rearrangements led to the definition of *common intervals* (Bergeron *et al.*, 2002; Jahn, 2011), conserved regions of a chromosome within which the same set of genes can be observed, albeit not necessarily in the same order. The concept is formally and precisely defined and captures many of the properties informally associated in the literature with SBs. The definition is given in terms of families of non-duplicated genes (or other families of unique sequences) and their ordering. It does not take into account precise locations on the genome, nor the actual nucleotide sequences of these genes.

Around the same time, the need to compare entire genomes of the newly sequenced model species led The Mouse Genome Sequencing Consortium (2002) to propose SBs as sets of adjacent *syntenic fragments* (possibly shuffled in order and orientation) belonging to the same chromosome, where a syntenic fragment consisted of markers arranged in a conserved order. In this view, syntenic fragments obey collinearity, whereas SBs need not do so. Calabrese *et al.* (2003), authors of the FISH synteny tool, defined their model based on *segmental homology*, in which the ordering of features belonging to two homologous segments is roughly conserved, some variation being allowed. Pevzner and Tesler (2003) and later Bourque *et al.* (2004), both working on the GRIMM-Synteny tool, removed constraints on conserved segments, thereby implicitly defining an SB in terms of conserved segments that can be disrupted by internal rearrangements—rearrangements that the authors found to be far more common than expected and that therefore had to be largely ignored in constructing SBs. In contrast, Van de Peer and his group, authors of the ADHoRe tool (Vandepoele *et al.*, 2002), chose to emphasize collinearity and to break larger blocks into smaller ones as necessary to maintain this property. These and other early tools are briefly reviewed in (Deonier *et al.*, 2005).

1.2 Markers, syntenic blocks and genomic alignment

Identifying SBs and aligning whole genomes both rely on identifying markers, i.e. short sequences that are highly conserved across the genomes and long enough to make their conservation statistically significant. SB construction uses subsets from the set of markers: if a sufficiently dense region is identified in most of the genomes, those regions can be viewed as SBs. Genomic alignment uses the markers as anchors, i.e. fixed references in the alignment. Most SB finders use genes as markers; a few use k-mers, for a fixed value of k, to define a de Bruijn graph on the k-mers. [de Bruijn graphs are widely used for genome assembly—see Compeau *et al.* (2011) for an excellent introduction in this context. In such a graph, every k-mer found in the input sequences is represented by an edge connecting two vertices that are the $k-1$ prefix and $k-1$ suffix of the k-mer. Thus a path of j edges through such a graph corresponds to an assembled sequence of length $k+j-1$ formed by ordering j k-mers, with each consecutive pair presenting a perfect overlap of length $k-1$; in particular, an Eulerian path through the graph corresponds to an assembly of all k-mers into a single sequence.] Genomic alignment may use a richer pool of markers, such as scaffold data, maximum unique matches (perfectly conserved sequence fragments of maximal length), genes and even assembly contigs. Those that use markers in the sense of highly conserved sequence fragments define markers through a variety of criteria, such as Bayesian statistics in Pecan (Paten *et al.*, 2009) or sequence similarity iterated through a refinement pipeline in ProgressiveMauve (Darling *et al.*, 2010).

Just as most work on defining SBs focuses on two genomes at a time, so is whole-genome alignment usually done pairwise. Biologists have long known that multi-way comparisons provide more information than pairwise comparisons, especially multi-way comparisons within a phylogenetic context. However, comparing several genomes at once introduces problems: finding good markers that are present in all, or almost all, genomes; choosing or inferring a number of parameters related to attributes difficult to measure, such as the level of evolutionary divergence among the genomes or the quality of the genome sequences used; assigning one-to-one correspondences among similar blocks so as to minimize the number of evolutionary events needed to explain the architecture of the modern genomes; whether to insist on the transitivity of relationships such as homology and orthology (among markers, among genes, among SBs, etc.); and many others.

1.3 Work to date

Nadeau and Taylor (1984) defined synteny in terms of two or more pairs of homologous genes occupying the same chromosomal segment, where homologous loci are based on similarity of function of the products of the corresponding genes. They carefully distinguished synteny, which they were basing on conservation of function, from conserved segments, based on conservation of sequence. More recent work has typically used conservation of sequence rather than conservation of function, but has also made use of orthology, presumably because orthology is viewed as a stronger indicator of conserved function than homology.

Zeng *et al.* (2008) based their *Orthocluster* tool strictly on gene orthology and used many parametric constraints, such as position, overall number of genes in a block, allowed number of genes per block without orthologs, etc. Their tool handles large-scale genomic events such as translocation, transposition, indels and duplication. The restriction to orthology, however, means that the applicability of the tool is limited to collections of closely related organisms.

Cassis (Baudet *et al*., 2010), also based on orthology relationships, prunes considerably the list of orthologous gene pairs provided as input, eliminating those that disrupt collinearity. The remaining pairs are used to form blocks based on a statistical evaluation of their match to the collinear model.

Modern tools all attempt to handle the loss of collinearity, in recognition of the fact that collinearity (absence of rearrangements) is unlikely to be observed in collections of genomes of any significant size or degree of divergence. Equally important and still challenging is the ability to deal with varying marker (most often gene) content: given reasonably divergent genomes, markers will have been variously lost or acquired over time.

In the multiple alignment tool ProgressiveMauve, Darling *et al*. (2010) focused on a very principled approach to define and then to use the markers for the alignment process. Its strategy is to identify highly conserved, sufficiently long sequences (anchors) throughout a concatenated multi-chromosomal genome and then, for each interval between consecutive anchors that exceeds a certain length, to search recursively for additional, less perfectly conserved anchors. This recursive refinement continues until the anchor coverage has reached a sufficient density or the heuristic cannot retrieve any additional anchors. ProgressiveMauve was designed as an alignment tool, not a synteny tool, but it generates a list of homologous, locally collinear regions that can be used as a basis for defining SBs.

Cyntenator (Roedelsperger and Dieterich, 2010) uses genes as markers and is based on a progressive alignment of profiles of gene-order data. It allows gene duplication and loss and thus, in order to distinguish between orthologs and paralogs, takes into account gene family information as part of its scoring scheme. Pairwise alignments produced at each stage are refined before being used in the next stage. As is the case for most such tools, the blocks identified by Cyntenator are not formally characterized, but indirectly defined through the algorithm.

i-ADHoRe 3.0 (Proost *et al*., 2012) also uses genes as markers; it includes heuristics to deal with rearrangement and duplication. Duplicated genes are mapped onto a representative of the gene family. The tool produces profiles of collinear regions based on homology maps of pairs of genomic regions and uses heuristics based on network flow to resolve conflicting relations between pairs of genes. The tool provides three constraint models for generating SBs: collinear (conserving both order and orientation), cloud (conserving neither order nor orientation, but content) and a sequential mixture of the two.

DRIMM-Synteny (Pham and Pevzner, 2010), the multi-way successor of the pairwise *GRIMM-Synteny*, is, like most synteny tools, based on genes, but follows an entirely different approach, as it is based on de Bruijn graphs. A somewhat different version of de Bruijn graphs, called A-Bruijn graphs, is used in order to take into account the different characteristics of the problem, such as the use of gene orders rather than overlaps. Thus a gene adjacency becomes an edge of the graph and is weighted by the number of its occurrences across the genomes. SBs correspond to paths through the graph.

Sibelia (Minkin *et al*., 2013) follows up on DRIMM, in that it is also based on de Bruijn graphs, but, being designed for bacterial genomes, it works directly from sequence data and so builds standard de Bruijn graphs from sequence *k*-mers. It also adds an iterative refinement procedure that provides a range of granularity for the blocks. The pipeline is executed individually for increasing sizes of the *k*-mers, until the output block is the whole genome. At each iteration, a different set of blocks is generated and is placed as a node into a tree structure, with the root of the tree corresponding to the whole genome.

Table 1. Major features or constraints of five synteny tools: ProgressiveMauve (PM), OrthoCluster (OC), Cyntenator (Cy), i-ADHoRe (i-A) and DRIMM (DR); presence is denoted by +, absence by − and options by o

	PM	OC	Cy	i-A	DR
Collinearity	−	o	−	o	−
Framed blocks	+	−	−	−	−
Overlapping content	−	+	+	+	−
Selective content	−	+	−	+	+
Across chromosomes	+	+	−	+	o
Duplicated regions	−	+	+	+	+

Table 1 lists the main features of the synteny tools we used.

1.4 Syntenic blocks, homology and granularity

That blocks generated from the same data by different tools may differ enormously is due mostly to the lack of a formal definition for SBs: with no verifiable constraints and no measurable optimality criterion, one cannot meaningfully compare two collections of SBs for the same data. In part, the lack of such constraints and criteria can be attributed to the very different uses to which SBs are put. For instance, using SBs to pinpoint a region of interest in the genomes works best if the blocks are small and highly conserved, whereas using SBs to study the evolution of the architecture of genomes does better with larger blocks and can tolerate much larger divergence in any given block among the genomes. (Indeed, the larger the evolutionary divergence, the larger and sparser the SBs should be, to account for the lower number of high-quality markers.)

When large-scale (segmental or whole-genome) duplications are present, multiple instances of the same SB will be found within the same genome, as well as throughout other genomes—that is, SBs, like genes, can be grouped into families of homologs. Identifying orthologies among the markers or genes is thus intertwined with identifying SBs—arguing for a simultaneous construction, which can take into account positions, rearrangements and duplications and losses of markers and of blocks all at once. Thus homology is at the root of any principled definition of SBs: the process of construction of SBs is simply the process of extending homologies among markers to homologies among blocks under a suitable model of evolution. In such a manner, partitioning the genomes into SBs defines the necessary higher-level homology relationships that relate such blocks within and across genomes.

Since all genomes share a common ancestor, every single genome is trivially an SB by itself, albeit with a very low degree of conservation across a collection of genomes. At the other extreme, if we had available a detailed history of all

evolutionary events at the sequence level, we could construct SBs consisting of a single nucleotide position. In a similar vein, two or more adjacent SBs can be viewed as single, larger SBs, presumably at the cost of some loss in conservation. In other words, *granularity* is an important attribute and one can construct a hierarchy of decompositions into SBs, taking the form of a rooted directed acyclic graph where the trivial decomposition into a single block sits at the root and the equally trivial decomposition into individual nucleotide positions sits at the single leaf. Children of a node in this dag are associated with decompositions of finer granularity than that associated with the node itself. Under some mild constraints, this dag is in fact a lattice (or partially ordered set).

It is important to note that the lattice is determined by constraints resulting from the definition of an SB, but the selection of a particular node in the lattice (a particular decomposition into blocks) is driven by other criteria (such as granularity) and thus determined by the application. (Of all the various tools reviewed here, only Sibelia makes explicit mention of a hierarchy of SBs.)

2 METHODS

2.1 Homology, orthology and synteny

Any definition of synteny must use homology or orthology. Most synteny tools today use both—homology as a matter of principle and orthology as a result of practical constraints. In evolutionary biology, two structures (character positions in a sequence, markers of various types, genes, SBs) are *homologous* if they are descended from a common ancestral structure (Fitch, 2000); if, in addition, the branching at the last common ancestor was a speciation, the structures are also *orthologous*. Thus homology is an equivalence relationship and, as such, determines equivalence classes, the homologous families of structures. Orthology, in contrast, depends on the speciation point and so is context-dependent; in particular, it is generally not transitive. (For instance, two gene duplicates within the same genome cannot be orthologous, but these two duplicates and a homologous gene in another species are orthologous if the duplication followed the speciation.) Instead, it must be specified through hierarchies structured through the phylogeny (see Gabaldon and Koonin, 2013).

Homology and orthology cannot be observed, but only inferred. In practice, homology for markers and genes is determined on the basis of sequence similarity, using tools such as BLAST. Orthology is also initially determined through sequence similarity, but often verified through phylogenetic analysis or by ascertaining functional similarity. However, only rarely is position along the genome taken into account—exceptions are the database OrthoDB (Waterhouse *et al.*, 2011), which also provides a hierarchy of orthology relationships, and the orthology tool MSOAR (Fu *et al.*, 2007). In practice, therefore, identifying homologies is much easier than identifying orthologies.

Synteny is defined both through families of homologous markers and through placement within the genome. Therefore identifying SBs, in addition to prior knowledge of homologies, requires taking into account rearrangements and duplications that disperses the members of a homologous family throughout the genome. (Conversely, of course, producing SBs makes direct statements about the evolutionary history of the genomes by ruling out some of the possible scenarios.) Therefore, in principle, the identification of SBs should proceed from homologies (which have little direct dependence on location) rather than from orthologies inferred without regard to location. Computing gene clusters, for instance, is best done based on families of homologous genes instead of relations derived from orthologous groups (Jahn, 2011).

Practice may dictate otherwise. Inferred homologies are neither symmetric nor transitive in practice, as they depend on similarity thresholds. In addition, since orthology is the stronger relationship, it is often preferred, at least for pairwise synteny, as it may provide higher quality markers and because it simplifies the task. (Some synteny finders simply transform orthologous relationships into bijections, in spite of the fact that orthology is a many-to-many relation.) When moving from pairwise to multi-way syntenies, orthologies become problematic: the more diverse the group of genomes, the more difficult it becomes to identify orthologies. In practice, therefore, synteny tools rely on both homology and orthology, viewed largely as different degrees of sequence similarity.

2.2 Towards a formal definition for syntenic blocks

Here we propose a fundamental constraint on the makeup of SBs, based on an evolutionary perspective. We first formalize that constraint for pairwise synteny, then extend it to multi-way synteny. We also propose a second constraint, which provides added refinement for bacterial genomes and also helps narrow searches when looking for conserved regions of interest.

Our definitions are made in terms of markers and homology statements among them. Thus we regard each genome as a multi-set of markers—a multi-set rather than a set, as the same marker may occur more than once in the same genome. Associated with each marker is a set of homology statements relating that marker to its homologs in other genomes or in its own genome; a homology statement is just an unordered pair of markers. Ideally, these homology statements define an equivalence relation on the set of markers; in practice, of course, these statements come from a variety of sources (databases, direct analysis of sequence similarity, etc.) and are unlikely to obey all the requirements of an equivalence relation.

Viewed abstractly, identifying SBs is a clustering problem: how do we partition the multi-set of markers into smaller multi-sets, so as to maximize the similarity (as attested by multiple homology statements) between some of the smaller multi-sets, while minimizing their similarity to others? Because our definition rests on homologies rather than orthologies, we expect to find homology statements connecting related SBs as well as some connecting unrelated SBs—by and large, the first are more likely to be orthologies, while the second are more likely to be paralogies. Our main constraint, then, is that, in order for two blocks to be homologous SBs, they must be connected through homology statements and that neither includes markers that, while unconnected in this manner to anything in the other blocks, are connected to markers in unrelated SBs.

We now formalize our definition for the basic version of SBs: SBs for two genomes, in which we restrict each to be a contiguous range of positions within a chromosome.

Definition 1. We are given two genomes, G_A with a set A of n_A markers and G_B with a set B of n_B markers; the markers of G_A are ordered along the chromosomes, as are the markers of G_B. Let H be a set of pairs of distinct elements of $A \cup B$—the homology statements. We assume that every marker in A and B is part of at least one homology statement.

Let S_A be a set of contiguous markers on one chromosome of G_A and S_B a set of contiguous markers on one chromosome of G_B. We say that S_A and S_B are homologous SBs if and only if, for any marker $x \in S_A$, there exists a marker $y \in S_B$ such that $\{x, y\}$ is a homology statement, and, for any marker $u \in S_B$, there exists a marker $v \in S_A$ such that $\{u, v\}$ is a homology statement.

We can further require that the two end markers form a conserved *frame*, thereby setting defined boundaries on the range of positions forming an SB.

Definition 2. Let S_A and S_B be homologous SBs as per Definition 1. If the first marker of S_A is a homolog of one of the two endmarkers (the first or last marker) of S_B and the last marker of S_A is a homolog

of the other endmarker of S_B, we say that S_A and S_B are (homologous) framed SBs.

Many of the existing tools require that the homology between markers respect the ordering of the markers along the blocks—a property usually referred to as *collinearity*. Because genomes are subject to rearrangements, we do not require collinearity, but we can define it as follows using our notation.

Definition 3. Let S_A and S_B be two homologous SBs as per Definition 1. We say that S_A and S_B are collinear SBs if the following condition, stated in the direction from S_A to S_B, holds in both directions: for any markers x and y in S_A with x appearing before y, there exist markers u and v in S_B, with u appearing before v, such that both $\{x, u\}$ and $\{y, v\}$ are homology statements.

Our requirement that each block be fully contained with a chromosome may require that some evolutionary events, such as translocation, fusion and fission, all of which can move genomic material between chromosomes, be treated as block-splitting events. For instance, if prior to such an operation, we would have identified regions A and B as homologous SBs, but the operation moved part of region A, call it A_t (tail) to another chromosome, leaving only A_h (head) in the original location, then after the operation we may be unable to associate either of A_h or A_t with B, but we may be able to associate A_h with a first subregion B_h of B and A_t with a second subregion B_t of B, thereby producing two pairs of smaller SBs.

We extend pairwise synteny to multi-way synteny by taking advantage of the transitive nature of true homology: we simply require transitive closure of pairwise relationships.

Definition 4. We say that blocks $A_1, A_2, \ldots, A_k$ are homologous SBs if and only if, for any i and j, $1 \leq i < j \leq k$, A_i and A_j are pairwise homologous SBs.

This definition is unambiguous whenever our set of homology statements defines an equivalence relation, since this property ensures transitivity. In practice, however, neither transitivity nor symmetry will hold: our set of homology statements will typically be incomplete as not all homologies among markers are detectable and homology defined through sequence similarity (the most common type in practice) need not be symmetric.

The output of a synteny tool is a collection of families of homologous SBs (henceforth SBFs), each family tied together with homology statements. We illustrate our definitions with a few cartoons. Figure 1 shows the building blocks for our cartoons and also demonstrates the additional structure present in framed SBs. Figure 2 illustrates the main characteristics used in our definitions. The first two cartoons in the figure show SBs defined through one-to-one (Fig. 2A) and one-to-many (Fig. 2B) homology statements. Homology statements may connect markers in non-homologous SBs, as long as other homology statements connect these markers to markers in homologous SBFs. The third cartoon (Fig. 2C) gives an example of invalid blocks: the red marker has a homolog in a non-homologous SB, but none in the putative homologous SBs.

3 RESULTS AND DISCUSSION

Our goal is to enable evaluations and comparisons of decompositions into SBs. Such evaluations and comparisons have mostly been missing and, when present, have typically been limited to aspects such as coverage of the genome or number of blocks, neither of which has much to do with quality. Our first step was to propose formal constraints that any decomposition into SBs should satisfy. These constraints are not likely to be met except in ideal cases, so our second step is to measure compliance with the constraints, which is to say, to measure quality. We therefore assemble a dataset of whole genomes to use in testing various methods; devise specific measurements of compliance with our definitions; and provide other insights and measures regarding the various tools tested.

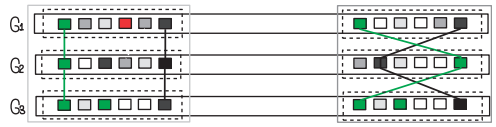

Fig. 1. A cartoon for SBFs among three genomes G_1, G_2 and G_3. The horizontal strips correspond to the genomes; small colored boxes denote markers; each SBF is framed by a dashed rectangular outline; and homologous SBFs are aligned vertically and enclosed in a thin solid box. Colored lines between horizontal strips connect markers and denote selected homology statements. Shown are an SBF of three framed homologous SBFs (on the left) and, using the same homology statements, an SBF of three ordinary homologous SBFs (on the right)

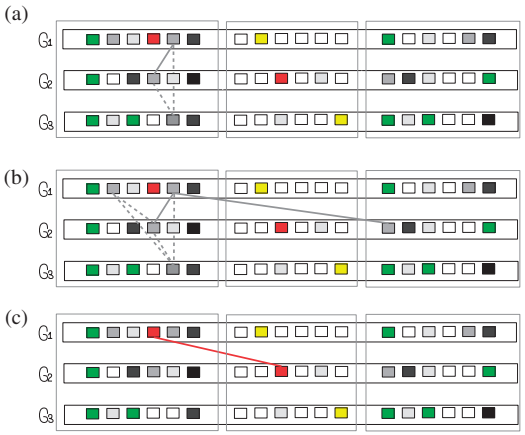

Fig. 2. Cartoons illustrating SBF structures on three genomes. Colors at marker level denote families of homologous units. (**a**) Three SBFs; in the SBF on the left, three markers are in one-to-one homology. (**b**) Three SBFs; in the SBF on the left, three markers are in one-to-many homology, including an additional homologous marker in another SBF. (**c**) Three putative SBFs; as shown, the red marker violates our definition, since it has a homology statement, but that homology connects it to a marker in a different SBF, while there is no homology connecting it to any marker within its own putative SBF

3.1 The data

Because we chose to include DRIMM in our evaluation, but could not reproduce its authors' results, we decided to use their results directly. Of the datasets used in the DRIMM study, only the yeasts combined complete results from the authors and public availability of the genomic data. We thus used the gene data from the *Yeast Gene Order Browser* (version of April 2009) (Byrne and Wolfe, 2005) for the following eight yeast genomes: *Candida glabrata* (c), *Eremothecium gossypii* (g), *Kluyveromyces lactis* (l), *Lachancea thermotolerans* (t), *Saccharomyces cerevisiae* (s), *Zygosaccharomyces rouxii* (r), *Kluyveromyces waltii* (w) and *Saccharomyces kluyveri* (k). The _genome.tab files were used to retrieve the complete list of genes for each of the organisms and

Table 2. Characteristics of the data from YGOB

Genomes	Genes/genome	Homolog pairs
C.glabrata	5211	106 291
E.gossypii	4725	104 817
K.lactis	5086	113 075
L.thermotolerans	5111	94 262
S.cerevisiae	6600	140 851
Z.rouxii	5006	135 707
K.waltii	10 825	194 234
S.kluyveri	5340	166 835

The 'genes' for *K.waltii* are often contigs with various functions (ORFs, short complements with intron/exon annotation), which explains their abnormally high number.

the associated NT.fsa file was processed in order to retrieve the sequences for these genes. Table 2 summarizes the characteristics of the data. All four tools require a list of homology statements—orthology statements for OrthoCluster. We used Fasta36 (Pearson, 1998), with a cutoff of 10^{-5}, to compile homology statements for each gene, reflecting common practice. We discarded any gene for which no homology statement was produced and, because Cyntenator does not scale well with large gene-family sizes, we retained only the 10 best matches (homolog candidates) for each gene. Computational constraints imposed by the tools meant that the number of markers could not be too large; moreover, a number of tools assume that the markers are genes; thus we used genes as markers.

3.2 The tools

We used the results of the DRIMM study and ran OrthoCluster, Cyntenator and i-ADHoRe on the yeast dataset. We had chosen DRIMM because it represented a very different approach to the problem (using de Bruijn graphs) and chose the other three because all are of recent design and maintained, all support multiway comparisons and all have clear statements about their design in the respective original publications. Unfortunately, in spite of prompt support from the developers, *OrthoCluster* (Zeng *et al.*, 2008) could not run within reasonable time on our dataset without removing so many genes and homology statements as to invalidate the exercise, so we had to exclude it from the study. (We ran the tool for 2 weeks on a 48-core, 256 GB Dell Poweredge 815 without results.)

We ran *Cyntenator* with the parameter setting used by the authors in the original article: gap = 0.3, mismatches = 0.3, threshold = 2 and filter = 10 000. The final output depends on, in effect, a guide tree (a phylogeny of the eight species), as it is obtained by running the tool on pairs of intermediate results—the tool ran well on pairs, but not so well on triples, and almost never on larger subsets of genomes. We eventually settled on the pattern described by the tree ((r, (w, (g, (k, (c, s))))), (l, t)).

We ran *i-ADHoRe* in collinear mode, with the following parameters: gap size = 15, cluster gap = 35, q value = 0.9, probability cutoff = 0.001, anchor points = 3, gg2 heuristic, no level 2 only and FDR as multiple hypothesis correction.

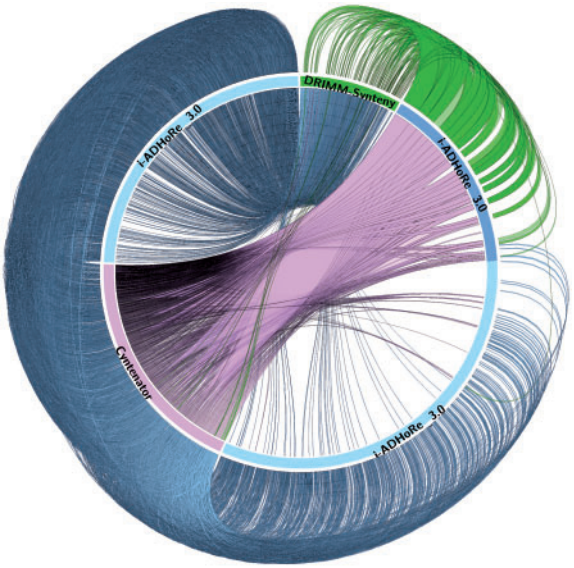

Fig. 3. SBFs defined by Cyntenator (purple), i-ADHoRe (blue) and DRIMM (green), mapped to each other in terms of gene content. Each link bears the color of the tool, the output of which is mapped through the link onto the outputs of the other tools. There are six pairwise comparisons between the SBFs produced by the three tools. The thickness of a link shows the level of similarity, measured by the overlap between the gene content of two SBFs relative to the SBF being mapped. Each sector of the diagram is an ordering by size of all blocks generated by the corresponding tool

3.3 The output

The output of all three tools is in the form of families of homologous SBFs, where each family has at most eight blocks, each belonging to one of the eight genomes under comparison. That we get no more than eight is due to the use of genes as markers: a large fraction of the genes are singletons (have no homolog within their own genome), thereby making it highly unlikely that a particular block structure would be found repeated within the genome. A family has fewer than eight blocks when no homologous SB in that family can be identified in a particular genome.

Figure 3 gives an overall feel for the results of the study, showing how the blocks from one tool map onto those of another. A very clear mapping pattern can be observed from both Cyntenator and DRIMM to a specific, small subset of the blocks generated by i-ADHoRe, as highlighted by the dark blue section on the ring of i-ADHoRe. The number of blocks generated by i-ADHoRe is considerably higher than those generated by Cyntenator or DRIMM, so the blocks are smaller and the (blue) links thinner. (This kind of mapping also illustrates the lattice concept discussed earlier: the thin links bind smaller blocks to a larger block made of these smaller blocks.)

3.4 Evaluation against our definitions

Our main requirement is that markers within an SB have homologs within each of the other SBs in the family. As we saw, this simple constraint is unlikely to be satisfied in practice, so we

Table 3. Characteristics of the SBFs generated by the tools

	SBFs	w/o homologs in the SBF	Content overlap	Selective content
DRIMM	509	509	0	455
Cyntenator	1106	583	39	0
i-ADHoRe	8088	278	2	7247

relax the transitivity requirement and measure compliance with the resulting weakened constraint.

Our first measure relates to the families of SBFs: we compute the number of SBFs that include within one of their SBs a marker with no homolog within any block of the SBF. This count is reported in the second column of Table 3. Since this measure tolerates failures in transitivity, the number of SBFs not in perfect compliance with our definition may be much larger.

This first measure is an absolute count, although different tools produce different numbers of SBFs; moreover, it counts an SBF as a failure no matter how many markers in that SBF fail the test. To address the first issue, we compute the percentage of 'failing' markers in an SBF—i.e. markers that have homologs in other SBFs, but none in their own SBF. We use two different base counts for normalization, to reflect fundamental differences between the tools with respect to selective use of markers: the first count is the total number of markers present in the SBF as generated by the tool, denoted $E(X)$, while the second is the total number of markers present in the genome within the coordinates of the generated blocks, denoted $E(X')$. Because DRIMM and i-ADHoRe eliminate markers from within SBs (within the coordinates of the block), something that Cyntenator does not do, the values of $E(X)$ for DRIMM and i-ADHoRe may be significantly smaller than those of $E(X')$. Figure 4 shows that i-ADHoRe generates more, and Cyntenator fewer, blocks with a very small fraction of markers lacking any homolog within their own SBF.

Definition 5. We define two scores, the first more forgiving than the second.

Relaxed Scoring uses a pairwise view of SBFs; for each block from an SBF, it counts the number of markers in that block that have *at least one* homolog within the SBF and normalizes it by the total number of markers present in the SBF.

Weighted Scoring attempts to quantify the deviation from our formal definition; for each block in an SBF, we count the number of markers in that block that have at least one homolog *in each of the other blocks* in the SBF and normalize this result by the number of blocks (minus 1) in the SBF and again by the total number of markers present in the SBF.

A perfect weighted score is 1, yet an SBF of n blocks with a weighted score of $1/(n-1)$ gets a perfect relaxed score. These scores allow us to estimate the robustness of the homology statements, as they show how densely interconnected the SBs are through their homology statements. A reduction from the first score to the second indicates that the tool has removed markers (to place them in other blocks) that fell within the block—so that the block produced is not contiguous.

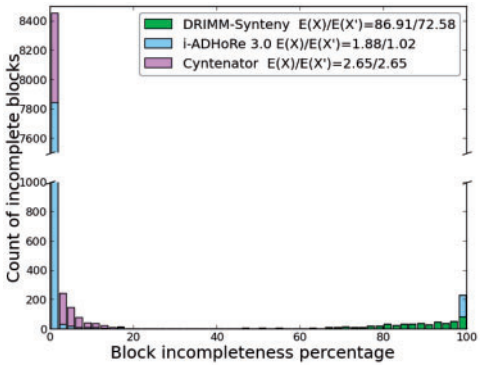

Fig. 4. Histogram showing the percentage of markers from an SBF that do not have any homolog in that SBF. The percentage is computed with respect to the total number of markers present in the SBF as generated by the tool and is supplemented by the $E(X)/E(X')$ ratio

Figure 5 gives histograms of the two measures for our experiments. Since i-ADHoRe explicitly produces non-contiguous blocks, its two scores predictably differ significantly (by a third). Like i-ADHoRe, DRIMM ignores many markers within a block, but in most cases it does not use them elsewhere—instead, it eliminates them from the list of markers it uses. As a result, its two base counts remain very close, but its two scores are very different.

Cyntenator and DRIMM yield similar distributions in both cases, but i-ADHoRE, which scores nearly perfectly under pairwise scoring, scores poorly under weighted scoring. i-ADHoRe does not place much emphasis on multi-way homologies: it keeps markers in its blocks even if these markers have just one homology with one other block. In contrast, Cyntenator progressively eliminates markers with few homology statements, therefore yielding blocks with strongly related markers. DRIMM has much the same behavior under both scoring schemes, but its score drops by 80% when moving from pairwise to weighted scores, due to its dropping large numbers of markers from its working list. That DRIMM scores poorly under both schemes, however, is due to a different set of goals: as stated by the authors, DRIMM aims at maximum genome coverage and simply ignores discordant homologies and other conditions that would cause Cyntenator or i-ADHoRe to break a block.

The yeast dataset contains several genes and ORFs that overlap. Such overlaps are discarded by DRIMM, but not by the other two tools; consequently, Cyntenator and i-ADHoRe occasionally output SBs with overlapping content (see Table 3).

Although we do not require collinearity, it remains desirable because it greatly simplifies the interpretation of the blocks. Cyntenator makes this a formal constraint; in contrast, most of the blocks produced by DRIMM and i-ADHoRe are interrupted intervals—between the leftmost marker and the rightmost one, both tools 'pick and choose' what to keep in the block. The last column of Table 3 indicates the number of blocks affected by this selection. The high proportion of blocks with selected content explains in part the good scoring of i-ADHoRe. In contrast, the very high proportion of such blocks, together with the 100% rate of homology violation, in DRIMM confirm the very

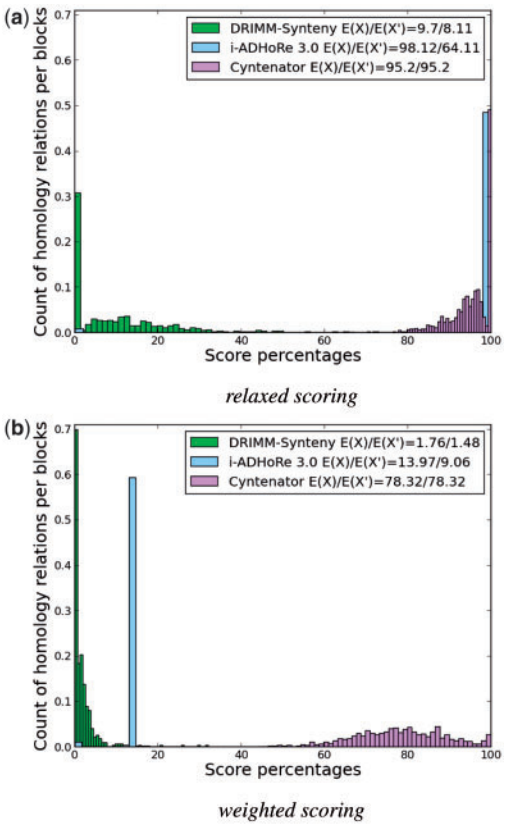

Fig. 5. Histograms of the two scores of Definition 5, illustrating the refinement over the simple score used in Figure 4

different aim driving the tool. A related issue is the handling of interchromosomal blocks: since genomic recombinations of various types can move parts of a conserved region to a different chromosome, one has to decide whether to split the conserved region into two SBs or to keep it as a single block. Our definition requires a split, since it assumes that each block is contained within a chromosome; DRIMM and Cyntenator do the same, but i-ADHoRe allows blocks to span multiple chromosomes.

3.5 Quantifying the features of the blocks

Comparing the blocks to each other is difficult, since explicit features of the blocks have not been defined *a priori* for any of the tools. We chose to focus on three features: genome coverage in terms of used markers (the one measure commonly used in the original papers), overlap of blocks for each tool and agreement among blocks in terms of marker content. We define marker coverage as the ratio of the total number of markers present in the blocks generated by a tool to the total number of markers present in the input within the generated block boundaries. Figure 6 illustrates (qualitatively, not quantitatively) how the blocks generated by each tool cover a certain genomic area. Figures 3 and 6 were generated using Circos (Krzywinski *et al.*, 2009). The three inner rings correspond to the three tools; each genome from our dataset corresponds to a cone in the figure, as indicated by the thin, labeled color indicator enclosing the

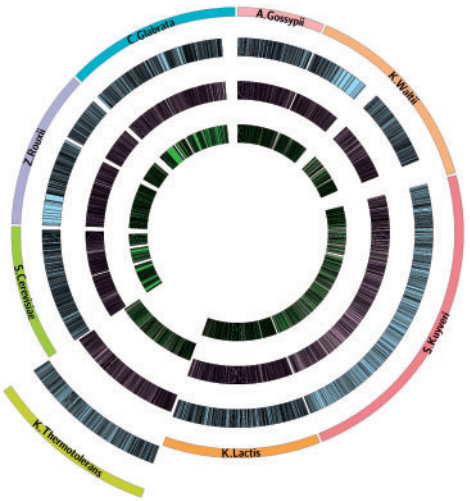

Fig. 6. SBFs generated by DRIMM (inside ring), Cyntenator (middle ring) and i-ADHoRe (outside ring). Each ring segment is a yeast genome. Dark regions include many block boundaries—these SBFs have few markers—while white regions have no identified SBFs. Note the many contrasting outcomes from ring to ring: where one tool breaks a region into many small blocks, another produces a single block

diagram. Block boundaries are drawn in thin black lines, so that dark areas represent short marker sets, thus small blocks and highly fragmented coverage. Uncovered areas are white.

Our definition does not preclude using overlapping SBs, since it sets conditions on one SBF at a time. In the lattice of decompositions into SBFs, one may then choose to impose additional conditions to select good blocks. DRIMM produces no overlapping blocks, because it does not reuse markers, whereas Cyntenator and (especially) i-ADHoRe do, which allows them to flag regions with ambiguous homologies or complex evolutionary histories. Figure 7 illustrates the degree to which markers are reused by Cyntenator and i-ADHoRe. While Cyntenator just reuses a few markers and not more than twice, i-ADHoRe reuses several of them up to 10 times, as depicted by the shape of the histograms.

We compute block similarity based on marker content: the markers of an SBF as generated by each tool are viewed as a single set and we compute the ratio between the overlap of two such sets relative to each of the sets, thereby yielding an asymmetric measure and six comparisons among the three tools. Figure 8 shows that the distribution is skewed towards small values—most SBFs have a small overlap with other families. Figure 8 also explains the types of links seen in Figure 3: most of the weight of the distribution is in the 10–40% region, corresponding to overlaps with the many small blocks produced by i-ADHoRe and thus to the thin blue links of Figure 3, while the same small blocks are also responsible for the large spike at 100%, since many will completely overlap with the larger blocks.

4 DISCUSSION AND CONCLUSIONS

We presented a review of the work to date on the definition and construction of SBs, pointing out the lack of a formal definition

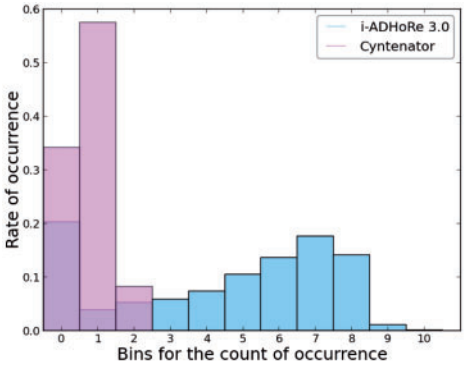

Fig. 7. Histogram of the reuse rate per marker for Cyntenator (green) and i-ADHoRe (yellow) on *C.Glabrata*. The *x*-axis shows the number of times a marker is reused and the *y*-axis shows the corresponding rate

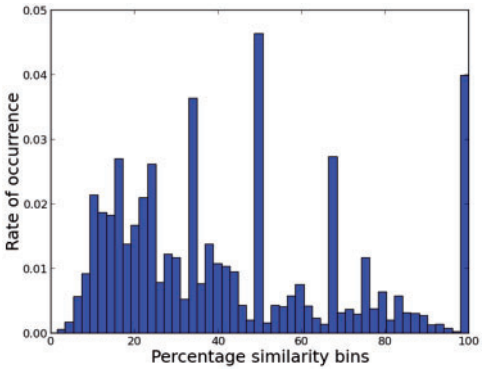

Fig. 8. Distribution of the similarity values for all pairwise comparisons between the SBFs generated by the three tools

of SBs as well as the lack of clear objectives for the tools designed to construct these blocks. The latter prevents us from evaluating each tool in terms of its own performance; the former prevents us from establishing a gold standard for evaluating the quality of SBs.

To remedy this situation, we proposed a simple set of homology-based criteria that SBs should satisfy. These criteria do not identify unique solutions—we argued that a range of solutions should remain, since the specifics of the application should influence the selection of good blocks. We based our definitions on homologies, because SBs are aimed at decomposing a genome into conserved regions (one of the few points on which all researchers agree) and conservation is embodied in homologies.

Since evaluating the quality of a decomposition into SBs is our main short-term goal, we defined new quality measures applicable to all decompositions into SBs and applied them to the output of several synteny tools run on a dataset of eight yeast genomes. This evaluation revealed very different behavior, as well as some reassuring commonalities, among the tools on the same dataset.

Almost all existing synteny tools use genes as markers. Not only does such a choice restrict the usable range of granularity, but, at least in the case of most eukaryotic genomes, it discards most of the sequence data (close to 98% in the case of the human genome). A sequence-based approach to the identification of markers, in the style of progressiveMauve or Sibelia, makes more sense in today's data environment. Among choices that a user should be able to make are: (i) permissible degree of overlap of blocks; (ii) acceptable percentage of dropped markers; and (iii) granularity. In addition, since the level of confidence in markers will vary, these choices should be further refined by taking into account the contribution of each shared, dropped or included marker. Clearly, then, the next generation of tools needs a hierarchical organization of blocks, a measure of significance for blocks based on strong connections between markers in the same SBF, and user-defined (and application-motivated) constraints and parameters.

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Baudet,C. *et al.* (2010) Cassis: Detection of genomic rearrangement breakpoints. *Bioinformatics*, **26**, 1897–1898.

Bergeron,A. (2002) Common intervals and sorting by reversals: a marriage of necessity. *Bioinformatics*, **18**, S54–S63.

Bourque,G. *et al.* (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, **14**, 507–516.

Byrne,K. and Wolfe,K. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.

Calabrese,P. *et al.* (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19** (**Suppl 1**), i74–i80.

Compeau,P. *et al.* (2011) How to apply de Bruijn graphs to genome assembly. *Nat. Biotech.*, **29**, 987–991.

Darling,A. *et al.* (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.

Deonier,R. *et al.* (2005) *Computational Genome Analysis: An Introduction*. Springer-Verlag, Berlin.

Fitch,W. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.

Fu,Z. *et al.* (2007) MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *J. Comput. Biol.*, **14**, 1160–1175.

Gabaldon,T. and Koonin,E. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.

Jahn,K. (2011) Efficient computation of approximate gene clusters based on reference occurrences. *J. Comput. Biol.*, **18**, 1255–1274.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Minkin,I. *et al.* (2013) Sibelia: A scalable and comprehensive synteny block generation tool for closely related microbial genomes. In *Proceedings of the 13th Workshop Algorithms in Bioinformatics (WABI'13)*, Vol. 8126 of *Lecture Notes in Computer Science*, pp. 215–229. Springer Verlag, Berlin.

Nadeau,J. and Taylor,B. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA*, **81**, 814–818.

Paten,B. *et al.* (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, **25**, 295–301.

Pearson,W. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Pevzner,P. and Tesler,G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, **13**, 37–45.

Pham,S. and Pevzner,P. (2010) DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.

Proost,S. *et al.* (2012) i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, **40**, e11.

Renwick,J.H. (1971) The mapping of human chromosomes. *Ann. Rev. Gen.*, **5**, 81–120.

Roedelsperger,C. and Dieterich,C. (2010) CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS One*, **5**, e8861.

The Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Vandepoele,K. *et al.* (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res.*, **12**, 1792–1801.

Waterhouse,R. *et al.* (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.*, **39**, D283–D288.

Zeng,X. *et al.* (2008) Orthocluster: A new tool for mining synteny blocks and applications in comparative genomics. In *Proceedings of the 11th Conference of Extending Database Technology EDBT'08*, pp. 656–667. ACM Press, New York.