COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 1, pp. 1-22, 2010

DECODING THE GENOMIC ARCHITECTURE OF MAMMALIAN AND PLANT GENOMES: SYNTENY BLOCKS AND LARGE-SCALE DUPLICATIONS[*]

QIAN PENG[¶][†], MAX A. ALEKSEYEV[‡], GLENN TESLER[§], AND PAVEL A. PEVZNER[¶]

Abstract. MOTIVATION: The existing synteny block reconstruction algorithms use *anchors* (e.g., orthologous genes) shared over *all* genomes to construct the synteny blocks for multiple genomes. This approach, while efficient for a few genomes, cannot be scaled to address the need to construct synteny blocks in many mammalian genomes that are currently being sequenced. The problem is that the number of anchors shared among *all* genomes quickly decreases with the increase in the number of genomes. Another problem is that many genomes (plant genomes in particular) had extensive duplications, which makes decoding of genomic architecture and rearrangement analysis in plants difficult. The existing synteny block generation algorithms in plants do not address the issue of generating non-overlapping synteny blocks suitable for analyzing rearrangements and evolution history of duplications.

RESULTS: In this paper we present a new synteny block generation algorithm based on the A-Bruijn graph framework that overcomes these difficulties. We applied our algorithm to derive non-overlapping synteny blocks in *Arabidopsis thaliana*. We also generalized this approach to synteny block generation for multiple genomes. The algorithm was applied to human-mouse-rat-dog-chicken genomes and it is able to recover synteny blocks missed by algorithms requiring 5-way anchors.

1. Introduction. Plant genomes exhibit an unusually large proportion of duplicated regions. In particular, up to 70% of all plant species have polyploid origin [2]. Many plant genomes are believed to be the result of extensive duplications followed by mutations, gene losses, and rearrangements [3, 4, 5]. The large number of duplications makes decoding of genomic architecture and rearrangement analysis in plants difficult. In particular, segmental duplications represent a major obstacle to reconstruction of *synteny blocks* (i.e., conserved regions across the genomes), resulting in relatively few published results on synteny blocks in plant genomes as compared to vertebrate genomes (and especially to mammalian genomes) where segmental duplications are less prevalent and can therefore be largely ignored while constructing synteny blocks.[1] It is estimated that segmental duplications account for less than

[*] Dedicated to Michael Waterman on the occasion of his 67th birthday. A preliminary version of this work appeared in the WABI'09 proceedings [1].

[†] To whom correspondence should be addressed. E-mail: qpeng@cs.ucsd.edu

[‡] Department of Computer Science and Engineering, University of South Carolina, 315 Main St., Columbia, SC 29208.

[§] Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093.

[¶] Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093.

[1] For example, in the human genome, segmental duplications are usually represented as a set of pairwise alignments that are masked out in the synteny block generation algorithms.

10% of the human genome [6, 7] and 2.9% of the mouse genome [8]. By contrast, in plant genomes, duplications are prevalent (e.g., duplications account for more than 70% of the *Arabidopsis thaliana* genome [5]), and ignoring duplicated regions would render a synteny block analysis meaningless. This represents an intrinsic difficulty in constructing synteny blocks in plant genomes.

From an algorithmic perspective, the problems of finding synteny blocks between two genomes and duplicated blocks are very similar. In fact, finding synteny blocks of multiple genomes can be converted into the problem of finding duplicated (or multi-copy) blocks within a single genome by concatenating the multiple genomes in an arbitrary order. This illustrates the challenge one faces while reconstructing synteny blocks in multiple mammalian genomes: indeed this problem is not unlike the difficult and still poorly addressed problem of reconstructing the synteny blocks in highly duplicated plant genomes. In the past, this problem of reconstructing synteny blocks in k mammalian genomes was addressed by constructing k-way anchors shared between all genomes [9]. However, this approach is limited to small k since with the growing number of genomes, the number of k-way anchors sharply decreases. The disappearing k-way anchors may lead to disappearing synteny blocks. Short synteny blocks (which are important in studies of chromosome evolution [10, 9]) are particularly vulnerable to this effect. In this paper, we propose a unified approach to synteny block reconstruction for two or multiple genomes, synteny block reconstruction for genomes with large duplications, and duplicated block reconstruction within a genome.

A typical synteny block generation algorithm takes as an input a set of *alignment anchors* (i.e., local alignments or pairs of similar genes) between two genomes (or two copies of the same genome) and outputs a set of synteny blocks (or duplicated blocks) that cover (without overlaps) most of each participating genome. As a result, each genome is represented as a shuffled sequence of the constructed synteny blocks that enables further rearrangement analysis of the genomes (e.g., computing the rearrangement distance between them). For two genomes, most existing synteny blocks generation algorithms employ a 2-dimensional *genomic dot-plot* where two genomes (or two copies of the same genome) are placed along the axes on the plane and their alignment anchors are represented as dots (Fig. 1.1(a)). These algorithms further decompose the dot-plot into a collection of "long" diagonal-like segments constituting *2-D synteny blocks* (Fig. 1.1(b)). The conventional (1-D) synteny blocks for each genome can be obtained as projections of the 2-D synteny blocks onto a corresponding axis (Fig. 1.1(b)). The notions of 2-dimensional dot-plots and synteny blocks generalize to k-dimensions when there are k genomes. This simple description hides a number of computational details that make the problem of synteny block generation non-trivial [11]. In particular, it was demonstrated in [12] that some synteny block generation algorithms may produce biologically inadequate results and emphasized the important differences between 2-D and 1-D synteny blocks.

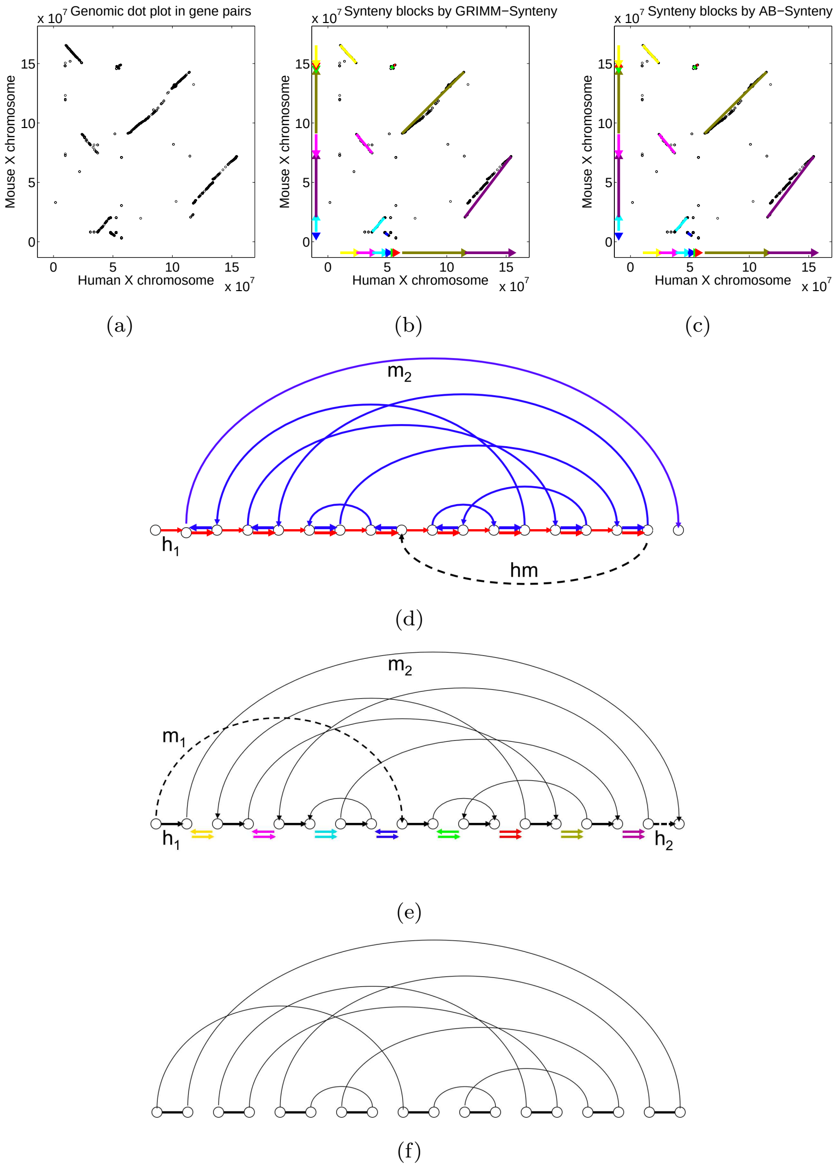


FIG. 1.1. *(a) Genomic dot-plot between human (x axis) and mouse (y axis) X chromosomes for all 714 orthologous gene pairs represented in Ensembl. Each dot represents a pair of "similar" genes between the two species. (b,c) Synteny blocks in 2-D (diagonals) and 1-D (vector arrow along x or y axis) produced by (b) GRIMM-Synteny [13] and (c) AB-Synteny. Each color represents a synteny block (duplication in the concatenated genome). (d) A-Bruijn graph after simplifications (with symmetrical nodes/edges removed). Edges are red (human), blue (mouse), and black (transition edge from human to mouse). Synteny blocks of (c) are illustrated as parallel pairs of red+blue edges in (d), but are actually single edges of multiplicity 2 in the A-Bruijn graph. Single red or blue edges represent breakpoint regions between blocks of one genome. (e,f) Transformation of A-Bruijn graph to breakpoint graph of [14]. The dashed edge hm in (d) is split and replaced by edges h_2 and m_1 in (e). The A-Bruijn graph has a transition edge hm between the genomes, while the breakpoint graph joins the beginning of each genome to the same start vertex (by edges h_1 in human and m_1 in mouse) and the end of each genome to the same end vertex (by edges h_2 and m_2). Each synteny block is given a unique color in (e). Removing the synteny blocks gives the breakpoint graph (f).*

Nadeau *et al.* [15] introduced the notion of *conserved segments* defined as segments with preserved gene orders without disruption by rearrangements. The early study of conserved segments were mainly based on comparative genetic maps [16, 17, 18]. In contrast, genomic sequences reveal substantially more rearrangements including *micro-rearrangements* (i.e., relatively short rearrangements) that were invisible in mapping data, requiring new algorithms to adequately deal with micro-rearrangements. Waterston *et al.* [19] and Pevzner *et al.* [11] described two approaches to synteny block generation, that produce similar results. Yet another approach based on *syntenic chains and nets* was proposed by [20]. There are many other studies describing different methods of "synteny block" generation [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. While these approaches proved to be adequate for small sets of mammalian genomes,[2] and in some cases prokaryotic genomes, they do not particularly address issues that stem from extensive duplications in plant genomes. In the presence of duplications, the 2-D synteny blocks may overlap in 1-D, i.e., along one of the genomes. A natural way to overcome synteny blocks overlapping in 1-D is to decompose them into smaller blocks that either do not overlap or overlap entirely in 1-D. As soon as such synteny blocks are constructed, each genome can be represented as a shuffled sequence of these blocks with some of them appearing in multiple copies.

There are a few previous efforts to generate synteny blocks for genomes with large duplications. Kellis *et al.* [33] constructed *Doubly Conserved Syntenies* (DCS) between *Saccharomyces cerevisiae* and *Kluyveromyces waltii* yeast genomes and used them to argue that the *Saccharomyces cerevisiae* genome underwent a *Whole Genome Duplication* (WGD) in the course of evolution. Duplicated synteny blocks for the *A. thaliana* genome were independently generated by [34] and [5]. Bowers' construction was done by connecting two anchors in the dot-plot as soon as the Manhattan distance between them was less than 20 genes (which is equivalent to GRIMM-Synteny's gap threshold $G = 20$ [11]). Then adjacent blocks with the opposite orientation that could be explained by a local inversion were combined. A gap threshold of 6 genes was used in [5][3], resulting in a larger number of duplicated blocks of a shorter average length. Baldi *et al.* [35, 36] developed algorithms LineUp and CloseUp that applied statistical methods to generate *chromosomal homology* (which may overlap even in 2-D, in contrast to synteny blocks) based on maize mapping data. Haas *et al.* [37] detects syntenic or duplicated regions by identifying chains of gene pairs sharing conserved order between genomic regions. Vandepoele *et al.* [38] introduced the ADHoRe algorithm to determine synteny blocks (which may overlap in 1-D) between *A. thaliana* genome and rice BACs. The improved tool i-ADHoRe combines gene content and gene order information to detect highly degenerated homology within and between

[2] Since the number of duplications in mammalian genomes is small, the 2-D synteny blocks usually do not overlap in 1-D.

[3] Details of the algorithm in [5] are sketchy.

genomes [39]. SyMAP was introduced by [40] to identify syntenic chains between *Zea mays*, *Sorghum bicolor*, and *Sorghum propinquum* FPC maps and the *Oryza sativa* genomic sequences. None of the aforementioned studies directly addressed the issue of generating *non-overlapping synteny blocks in 1-D*, which are more suitable for analyzing rearrangements and evolution history of duplications.

Pevzner *et al.* [41] introduced the *A-Bruijn graph* approach to repeat classification, representing all repeats in a genome as a mosaic of sub-repeats. Later, A-Bruijn graphs were also found useful in other problems such as multiple alignment [42], composite repeat analysis [43], and *de novo* protein sequencing [44]. In this paper we demonstrate that the A-Bruijn graph framework can be also applied to the problem of synteny block generation for genomes with large duplications. Our algorithm produces non-overlapping synteny blocks in both 2-D and 1-D representations.

By simply concatenating multiple genomes, we can generalize this approach to synteny block generation for multiple genomes. Previous efforts to generate synteny blocks for k genomes often required k-way alignment anchors, e.g., orthologous genes present in all k genomes [45]. As k increases, the number of k-way anchors decreases, making the methods hard to scale. For example, in Ensembl database version 44, there are 14903 one-to-one orthologous human-mouse gene pairs (anchors), 13452 human-mouse-rat anchors, 12359 human-mouse-rat-dog anchors, and only 8735 human-mouse-rat-dog-chicken anchors.[4] While missing anchors are allowed in [30] when constructing conserved segments between multiple genomes, it requires a reference genome to which all other genomes are aligned. Mercator [46] detects cliques of various sizes to use as potential anchors. While it allows anchors from a subset of the genomes, it also requires all-vs-all alignments (unless some preprocessing is performed). Our approach is not subject to such constraints as it uses pairwise anchors as an input, and it does not require a reference genome.

While there is no gold standard to what constitutes "correct" synteny blocks, all synteny block generation algorithms are parameter-dependent and may produce different synteny blocks on the same input data. To evaluate the performance of synteny block generation algorithms, we simulated genomes with large duplications and known synteny blocks and analyzed how well our algorithm reconstructs the underlined synteny blocks. We further benchmarked our algorithm on five vertebrate genomes to reconstruct 5-way syntenys, and on the plant genome *A. thaliana* to find duplicated blocks. We compared the results to published syntenys or duplicated blocks.

[4] Note that "anchors" in the context of synteny block generation usually refer to "unique anchors." We use the term more loosely in this paper to represent any aligned elements between or within genomes, where an element may align to one or more other elements.

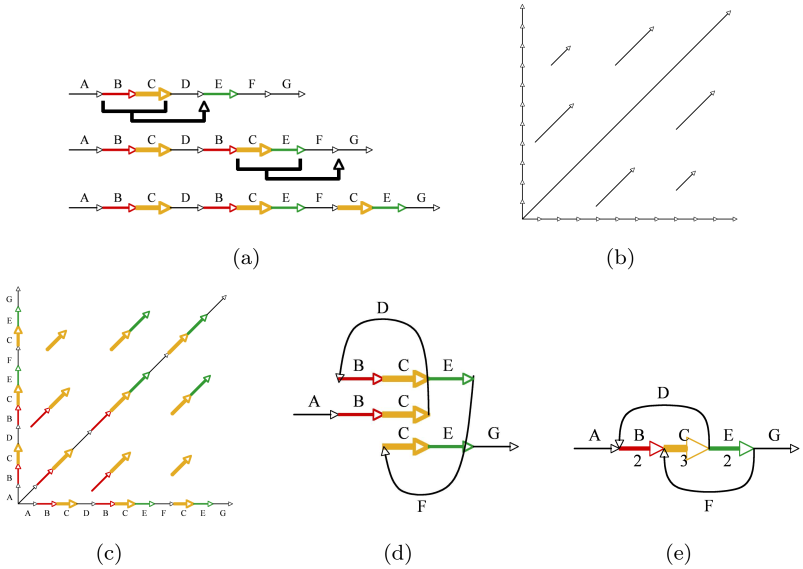


Fig. 2.1. *(a) Hypothetical sequence with multiple duplications. (b,c) Genomic dot-plot and the resulting synteny blocks of the sequence with multiple duplications. The 2-D representations overlap in 1-D. (d) Generate A-Bruijn graph of the sequence. (e) A-Bruijn graph of the sequence. Edges with multiplicity greater than 1 are synteny blocks. Blocks B, E each have two copies, and C has three copies. The algorithm outputs B,C,E as separate paths/blocks.*

2. Approach.

Fig. 2.1(a) shows a hypothetical sequence of genes, resulting from multiple segmental duplications. In reality, we are given only the resulting genomic sequence and know nothing about the structure of its segments (marked by the colors in Fig. 2.1(a)). It is natural to ask what evolutionary events (including rearrangements and duplications) created the given genomic sequence. Before answering this question, we need to understand the duplication structure of the given genome, i.e., to represent it as a sequence of non-overlapping blocks, each of which may appear one or more times.

The diagonals in Fig. 2.1(b) are what conventional synteny block construction methods would produce as synteny blocks from the genomic dot-plot of a genome against itself. Since these blocks overlap along the sequence, the duplication structure is unclear. Ideally, we would like to see diagonals that do not overlap along the sequence (Fig. 2.1(c)). One natural approach is for every pair of partially overlapped blocks along each axis to cut the overlapping region off these blocks into two new entirely overlapping blocks. As newly created blocks may partially overlap with other blocks, to eliminate all such partial overlaps a number of subsequent cuts may be required. The problem with such an approach, however, is that in some cases the initial synteny blocks might result in the iterative fragmenting and shrinking of synteny blocks. While this phenomenon is well known in repeat classification (e.g., the RECON algorithm [47] follows a similar scheme), it has not been addressed yet in synteny block reconstruction. This simple and seemingly sensible approach does not work well in complex cases [41]. For example, early attempts to use a similar approach for constructing "duplication subunits" (analogs of synteny blocks for segmental duplications) failed and new more elaborated techniques were used to resolve this challenge [48]. While the complexity of synteny reconstruction so far is nowhere close to the complexity of the repeat analysis, the addition of every new species will soon make the synteny reconstruction more difficult, thus calling for techniques to overcome the limits of tools based on iterative splitting. In addition, synteny blocks (different from repeats) are subject to microrearrangements, thus further complicating the problem.

Although repeats and duplicated synteny blocks result from different biological events, they both represent sub-sequences appearing multiple times in the genomes. Repeats and duplicated synteny blocks differ mostly in length and in the number of occurrences in the genomes. Therefore, the problem of constructing non-overlapping synteny blocks for genomes with duplications is similar to the problem of *de novo* repeat classification and can be solved accordingly.

The same approach can also be used for generation of synteny blocks across multiple genomes by simply concatenating them into a single genome. If there are no duplications in the original genomes, then a k-copy synteny block in the concatenated genome corresponds to a k-way synteny block of the original genomes.

3. Methods.

3.1. AB-Synteny. Without loss of generality, we will generate synteny blocks for a single genome: for two or more genomes, we will simply concatenate them to obtain a combined genomic sequence that will constitute a "single genome" as input to our synteny block generation algorithm. Suppose that the given genome is represented as a sequence of elements (base pairs or genes) $v_1, v_2, \ldots, v_n$. These elements form the vertices of a *path-graph* P where every pair of consecutive vertices v_i and v_{i+1} (for $i = 1, \ldots, n-1$) are connected with a directed edge. To obtain an A-Bruijn graph [41] A from the graph P, one needs to "glue" all vertices of P belonging to the same anchor into a single vertex (Fig. 2.1(d)). The resulting A-Bruijn graph A inherits all edges from the path-graph P, counting multiplicity of each edge as its weight, (hence, the edges in A are weighted and there are no parallel edges (Fig. 2.1(e)).

The A-Bruijn graph A has one source and one sink such that the original genome can be read along some path from the source to the sink. In practice, it is convenient to include the inverted sequence of the same genome, representing the reverse DNA strand, in which case there are two sources and two sinks. Every edge with weight greater than one corresponds to a *syntenic region* (i.e., a region that may belong to at most one synteny block), and its weight gives the number of copies of this syntenic region in the genome.

Unfortunately, such an interpretation of the A-Bruijn graph meets a number of obstacles. Inconsistencies in alignments and tandem duplications may create *whirls* (short directed cycles) in A, while gaps in alignments may create *bulges* (short undirected cycles) in A [41]. As a result, the constructed A-Bruijn graph can be exceedingly complicated. For example, the A-Bruijn graph constructed from *A. thaliana* gene pairs has 6394 vertices and 12761 edges. To overcome these difficulties, [41] suggested a heuristic routine simplifying an A-Bruijn graph, which we partially apply to the graph A. In the process, we simplify the A-Bruijn graph by substituting every simple path in the graph by a single edge with its length equal to the length of the path.

Overall, our synteny block generation algorithm AB-Synteny has the following parameters (in the corresponding unit, e.g., number of genes if genes are used as anchors, or number of nucleotides if anchors are sequence alignments):

- the *girth* g specifies a distance threshold for removing whirls;
- B specifies the threshold for bulge removal;
- L is the block size threshold (minimum number of elements in the blocks);
- G and C are gap and block size thresholds used in pre-processing (to eliminate noisy anchors from the input data).

AB-Synteny(G, C, g, B, L)

1. For two or more genomes, concatenate all genomes forming a single genome.
2. Pre-processing: run GRIMM-Synteny(G, C) [13] to produce non-overlapping syntenic blocks in 2-D (blocks may overlap in 1-D). GRIMM-Synteny removes all anchors within "small" blocks (smaller than C).
3. Construct an A-Bruijn graph: run A-Bruijn(g, B)—a simpler version of the graph clean up routines detailed in [41]—on the remaining anchors, removing whirls shorter than g and simple bulges shorter than B.
4. Output non-overlapping paths whose multiplicities are greater than one and whose length are equal or greater than L. These are syntenic regions.
5. Post-processing: merge neighboring syntenic regions of same orientation interrupted only by short gaps into synteny blocks. Assign each synteny block a unique ID.

We remark that since the constructed paths (syntenic regions) do not overlap in the A-Bruijn graph, they also do not overlap in 1-D (both before and after the post-processing step). As a result, AB-Synteny produces a number of synteny blocks non-overlapping in 1-D and a representation of the given genome as a mosaic of these blocks (each block may appear in multiple copies). In other words, an entire genome is represented as a *word* over the alphabet of synteny blocks, which facilitates further duplication and rearrangement studies.

Table 3.1

Number of orthologous gene-pairs used by AB-Synteny to construct synteny blocks

	human	mouse	rat	dog	chicken
# genes	31101	28157	27264	22602	15936
mouse	14917				
rat	14257	16006			
dog	14943	14517	13940		
chicken	10990	10717	10304	10850	

3.2. Data sets.

3.2.1. Vertebrate genomes.. As an illustration and a validation of the AB-Synteny algorithm, we extracted and analyzed 714 gene pairs between human and mouse X chromosomes from Ensembl database version 39. The gene pairs are described as "orthologs" by Ensembl. After highly repetitive gene pairs (present in more than 10 copies) are removed (as they do not normally contribute to synteny blocks but would instead increase noises), 606 gene pairs remain. The human X chromosome has a total of 1360 genes and mouse X chromosome has 1267 genes.

We further constructed 5-way synteny blocks for human, mouse, rat, dog and chicken genomes using all available pairwise orthologous (one-to-one) genes from Ensembl 44 (Table 3.1).

3.2.2. Plant genome.. We analyzed 5700 paralogous gene pairs in *A. thaliana* from [34], selected from about 30503 *A. thaliana* genes. We compared our results to the published *A. thaliana* duplicated blocks generally accepted by the plant research community.

3.2.3. Simulated genomes with segmental duplications.. To construct a simulated genome with duplications we started with a sequence of unique elements (genes) and performed a number of segmental duplications over it. A segmental duplication over a sequence $x_1, ..., x_N$ is defined by three parameters: starting/ending positions $1 \le p \le q \le N$ and a target position $1 \le t < p$ or $t < q \le N$, and results in the sequence $x_1, x_2, \dots, x_{t-1}, x_p, x_{p+1}, \dots, x_q, x_t, x_{t+1}, \dots, x_N$. Each duplicated region then becomes subject to gene loss.

Procedure Simulate_SEG generates a simulated genome with segmental duplications with the following parameters:

- N is the number of unique elements (e.g. genes) in the genome prior to any duplication;
- n is the number of duplications performed;
- W is the maximum span of each duplication;
- R is the gene loss rate, i.e., the percentage of the duplicated elements (both original and duplicated copies) to be deleted;

- m is the total number of inversions performed within duplicated regions.

Simulate_SEG(N, n, W, R, m)

1. Generate genome $x_1...x_N$, where $x_i \neq x_j$ for all $1 \leq i, j \leq N$.
2. Randomly select locations i, j (duplication breakpoints), and a width w, such that $1 \leq w \leq W$, $1 \leq i, j \leq N$ and $i < j$ or $j + w \leq i$. Duplicate $x_j, x_{j+1}, ...x_{j+w-1}$ at location i. Repeat n times. Record all duplicated regions.
3. Randomly delete k elements from duplicated regions, where $k = (R/100)*$total number of elements in the duplicated regions.
4. Perform m inversions at random breakpoints within duplicated regions.

The breakpoints of each duplication are recorded and tracked. As the duplication breakpoints are randomly selected, some regions of the genome may be duplicated multiple times. By performing inversions within duplicated regions, the duplication breakpoints do not move and non-duplicated regions are not mixed with duplicated regions. Since we know exactly where the duplicated regions are, we can compare them to the duplicated regions recovered by the algorithm. Note that it is possible that duplicated regions are beyond recognition in case of severe gene loss (when R is large).

3.2.4. Simulated genomes with whole genome duplications.. As many plants have undergone Whole Genome Duplication, we also simulated this process. Procedure Simulate_WGD generates a simulated genome with whole genome duplication with the following parameters:

- N is the number of unique elements (e.g. genes) in the genome prior to the whole genome duplication;
- M is the total number of macro-inversions performed on the genome after a WGD. Inversion breakpoints are randomly selected;
- m is the total number of micro-inversions performed on the genome with randomly selected breakpoints subject to the constrain of a maximum span;
- w is the maximum inversion span of each micro-inversion;
- R is the gene loss rate, i.e., the percentage of the duplicated elements (both original and duplicated copies) to be deleted.

Simulate_WGD(N, M, m, w, R)

1. Generate genome $x_1...x_N, x_1...x_N$, where $x_i \neq x_j$ for all $1 \leq i, j \leq N$.
2. Perform M macro-inversions. Both breakpoints are randomly selected.
3. Perform m micro-inversions. For each micro-inversion, randomly select a breakpoint i and a width d such that $1 \leq d \leq w$; perform the inversion between breakpoints i and $i + d$.
4. Randomly delete k elements from the genome, where $k = (R/100) * 2N$.

By mapping all inversion breakpoints back to the genome $(x_1...x_N)$ prior to the whole genome duplication, we may obtain all duplicated regions. After gene deletions, however, some of the regions may be deleted entirely or have lost many of their genes such that a pair of duplicated regions share few common genes (anchors).

4. Results.

4.1. Syntenic analysis of vertebrate genomes. We concatenated 1360 genes from the human X chromosome and 1267 genes from the mouse X chromosome, forming a genome of 2627 genes.[5] During concatenation, a number of elements (larger than the gap threshold) were inserted between two chromosomes to prevent synteny blocks from forming across boundaries of chromosomes or genomes. An A-Bruijn graph was constructed on the concatenated genomes using the 606 gene pairs between human and mouse X chromosomes as gluing instructions. The A-Bruijn graph has 906 vertices and 1636 edges. The graph was further simplified with the parameters $g = 10, B = 20$ and $L = 4$. The simplfied graph includes both forward and inverted sequences. After all simplificiations, we extract the forward sequence, shown in Fig. 1.1(d). Figs. 1.1(e) and 1.1(f) illustrate that the A-Bruijn graph is actually equivalent to the breakpoint graph for analyzing rearrangement scenarios [14]. After joining the neighboring synteny blocks of the same orientation, a total of 8 strips of synteny blocks emerged between human and mouse X chromosomes 1.1(c), covering 85.64% of human and 89.72% of mouse X chromosomes. These synteny blocks are similar to the published results [11] with small differences mainly caused by correcting fragment assembly errors in the latest versions of the human and mouse genomic sequences. The GRIMM-Synteny results on this dataset are shown in 1.1(b). The blocks from AB-Synteny and GRIMM-Synteny largely coincide.

For human-mouse-rat-dog-chicken, we did two sets of synteny block generations. In the first set, we concatenated all 31101 human genes, 28157 mouse genes, 27264 rat genes, 22602 dog genes, and 15936 chicken genes into a single genome, and applied AB-Synteny $(G = 30, C = 3, g = 10, B = 20, L = 4)$ to the resulting genome using total of 125347 available gene pairs (1-to-1 orthologs) between any two genomes as gluing instructions. In the second set, we removed all genes that do not belong to any gene pair, and concatenated the remaining 16196 human genes, 17196 mouse genes, 16464 rat genes, 15792 dog genes, and 10908 chicken genes into a single genome, and applied AB-Synteny with the same parameters. The results are very similar and we report the results from the second set. After the vertices with 1-in and 1-out edges are merged, the A-Bruijn graph has 23228 vertices and 41833 edges. After the simplification, 3564 vertices and 6814 edges remained. They resulted in 666 5-way

[5] The previous analysis [45] revealed that adding nongenetic to genetic similarities hardly affects the synteny blocks. We therefore limit our analysis to genes only rather than arbitrary regions of similarity between multiple genomes.

Table 4.1
5-way synteny blocks constructed by AB-Synteny and GRIMM-Synteny

	Genome	AB-Synteny (666)		GRIMM-Synteny (466)		Shared coverage	
	length (Mb)	length (Mb)	%	length (Mb)	%	length (Mb)	%
human	3080	2017	65.47	2091	67.90	1837	59.63
mouse	2644	1792	67.79	1810	68.45	1604	60.66
rat	2719	1884	69.30	1915	70.43	1692	62.24
dog	2445	1669	68.26	1776	72.62	1527	62.44
chicken	1032	790	76.52	790	76.57	702	68.03

synteny blocks. We also extracted 8735 5-way orthologous genes from the gene pairs and applied GRIMM-Synteny($G = 100, C = 4$), where G is the total gap threshold. The results from the two algorithms are compared and listed in Table 4.1. The shared coverage refers to regions of a genome that belong to synteny blocks reconstructed by both algorithms. Fig. 4.1 compares the human-mouse X chromosome portion of the 5-way synteny blocks to those synteny blocks derived from human-mouse data alone as shown in Fig. 1.1(c). As expected, the 5-way blocks are shorter and more fragmented.

GRIMM-Synteny requires k-way anchors when there are k species. Some of the synteny blocks recovered by AB-Synteny but missed by GRIMM-Synteny are due to a reduced number of k-way (5-way) anchors as the number of species increases. Fig. 4.2 illustrates such an example. The region is on chromosome 1 of the five species and it consists of 13 genes from human, 9 genes from mouse, 9 genes from rat, 10 genes from dog and 5 genes from chicken. There are three 5-way anchors, two 4-way anchors and one 3-way anchor. Only the 5-way anchors can be used as inputs to GRIMM-Synteny or any other algorithms that require k-way anchors for k genomes. Since the number of such anchors is below the block threshold, the syntenic region is missed by GRIMM-Synteny. On the other hand, AB-Synteny requires only pairwise anchors between any two genomes. All six anchors therefore can be used as inputs. With equivalent parameter settings, AB-Synteny is able to recover the block as a result of more supporting anchors. This feature of AB-Synteny allows the algorithm to scale more easily to a large number of genomes.

4.2. Duplication analysis of plant genome. We applied AB-Synteny ($G = 20$, $C = 6$, $g = 10$, $B = 100$, $L = 4$) to the *Arabidopsis thaliana* genome with 30503 genes and 5700 anchors (gene pairs). It generated 223 non-overlapping segments in 1-D, making up 103 synteny blocks. Tables 4.2 and 4.3 compare AB-Synteny results with the synteny blocks from [34] and [5].[6] Almost all our synteny blocks are inside

[6] There is a discrepancy between the numbers reported here and those reported by [34] due to the total number of genes considered. We used the *Arabidopsis thaliana* gene set released by NCBI on 9 June 2005 (and by TIGR 5.0), with 30503 genes, which includes more predicted genes than what

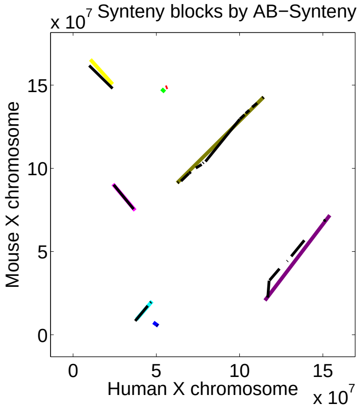


FIG. 4.1. *Synteny blocks between human and mouse X chromosomes generated by AB-Synteny. Colored diagonals are generated by using human and mouse data alone (same as in Figure 1.1(c)). Black diagonals are the human-mouse part of the human-mouse-rat-dog-chicken 5-way syntenic blocks.*

blocks of Bowers *et al.*, and about 2.66% of our blocks (covering 812 genes) are outside blocks of Blanc *et al.*

To estimate the quality of a synteny block, we define two measures. First, we define the *anchor density* of synteny block i as $d_i = |A|/|B_1 \cup B_2|$, where A is the set of anchors contained in the synteny block, and B_1 and B_2 are the set of elements in the two syntenic regions respectively. The anchor density of all synteny blocks is defined as $\frac{1}{n}\sum_{i=1}^{n} d_i$, where n is the number of synteny blocks.

As a second measure, we define the *anchor-synteny distance* (ds_i) of synteny block i between two species as follows. Let X_1, X_2 be the starting and ending points of synteny block i on the first species and Y_1, Y_2 on the second species. Let n_i be the total number of anchors in this block. Let x_k, y_k be the center of the kth anchor in the block (where the centers are computed by averaging the start and end coordinates of the anchor within each species). Then

$$ds_i = \frac{|\sum_{k=1}^{n_i}(X_2 - X_1)(Y_1 - y_k) - (X_1 - x_k)(Y_2 - Y_1)|}{\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}}\,.$$

In other words, ds_i is the sum of the signed distance from each anchor to its synteny block represented as a diagonal line. The anchor-synteny distance of all synteny blocks is simply $\sum_{i=1}^{n} ds_i$ where n is the total number of synteny blocks.

While there is currently no objective criteria to evaluate the quality of synteny block generation algorithms, the anchor density (while imperfect and biased in favor of short synteny blocks) and anchor-synteny distance allow one to compare different synteny block generation approaches.

was considered in [34]. Since the blocks reported by [34] contain only anchors, the size of each block is determined by the number of genes it covers.

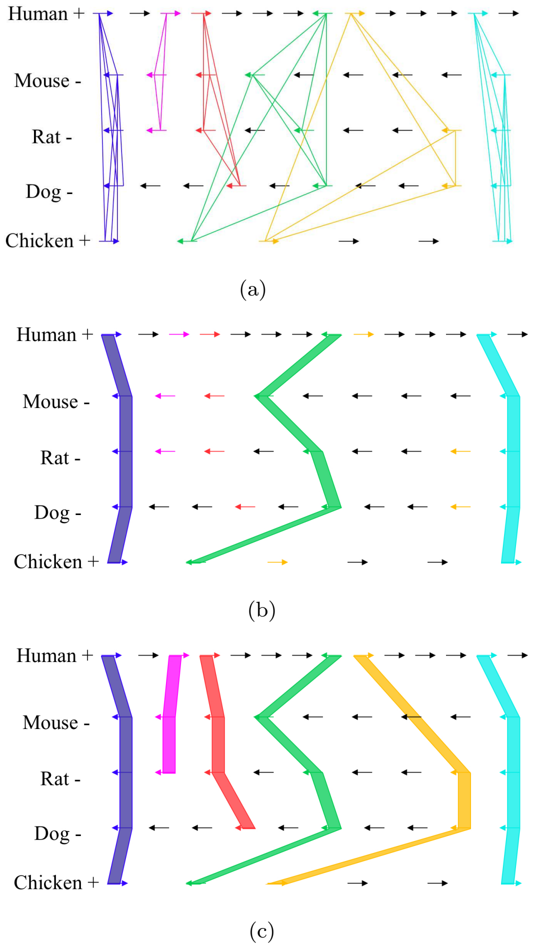


FIG. 4.2. *A region of chromosome 1 of five species that is recovered by AB-Synteny but missed by GRIMM-Synteny due to the small number of 5-way anchors. (a) one-to-one orthologous gene pairs from Ensembl database. (b) 5-way anchors input to GRIMM-Synteny: the block is missed. (c) All available anchors are used by AB-Synteny: the block is recovered.*

Fig. 4.3 shows several synteny blocks generated by AB-Synteny for the *A. thaliana* genome. Notice that the blocks appearing in more than two copies (blue colored blocks) are delineated from the 2-copy blocks (magenta blocks). The single red block is one of the synteny blocks (referred to as *chromosomal segment pairs*) reported in [34]. Careful inspection of the genomic segment shown in Fig. 4.3 reveals a large gap of 499 genes in chromosome 3 (x axis) and a corresponding gap of 12 genes in chromosome 4 (y axis). We therefore argue that the AB-Synteny representation provides a more accurate view of the *A. thaliana* genomic architecture.

The last step in the synteny block generation algorithm in [34] combines adjacent syntenic regions with opposite orientation and order that may be explained by local inversions, although it is not clear which inversions are considered local. The separa-

TABLE 4.2
Comparison of AB-Synteny results to published A. thaliana synteny blocks

	# of Synteny	Coverage		Overlap in 1-D		Anchor	Anchor-synteny
Methods	blocks	# genes	%	# genes	%	density	distance
1. [34]	34	26034	85.35	5069	16.62	0.158	107.70
2. [5]	91	24370	79.89	7118	23.34	0.119	116.74
3. AB-Synteny	103	21862	71.67	**0**	**0**	0.193	84.07

TABLE 4.3
Synteny block coverage shared between methods in Table 4.2

Methods	# genes	%
1 & 2	24089	78.97
1 & 3	21847	71.62
2 & 3	21050	69.01

tion of segments in such cases can partially explain the comparatively low coverage of AB-Synteny as shown in Table 4.2.

There is partial agreement of our synteny blocks with those generated by LineUp [35] (data not shown). While the syntenic regions reported by LineUp do in general overlap with the regions generated by AB-Synteny, LineUp reports all statistically significant syntenic regions without trying to define the boundaries of the regions. As these regions overlap significantly, they cannot be used for reconstruction of rearrangement and duplication scenarios.

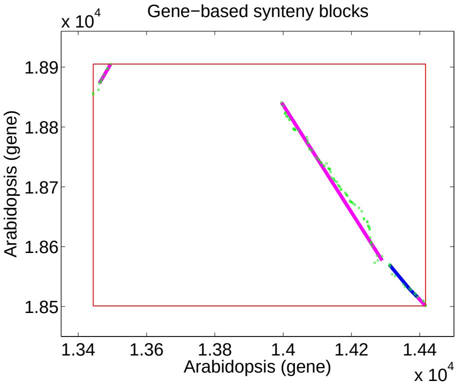


FIG. 4.3. *A local view of synteny blocks of A. thaliana generated by AB-Synteny. Gene pairs (anchors) are in green; the synteny block from [34] is a red box; and other colored diagonals are synteny blocks generated by AB-Synteny: magenta: 2 copies, blue: 3 copies (the extra copy is not shown).*

4.3. AB-Synteny on simulated genomes. While AB-Synteny produces non-overlapping synteny blocks in their 1-D representations, the purpose of simulation is to examine how well the algorithm recovers known (from simulation) duplicated regions in a genome. The simulated genomes with segmental duplications are generated by running Simulate_SEG with following parameters:

$N = 4000$, $n = 100, 200$, $W = 40, 60, 80, 100$, $R = 30, 35, 40, 45, 50, 60$, $m = 50, 100, 150, 200$

We applied AB-Synteny$(.,.,10,40,4)$ without pre-processing (hence a . in place of G and C) to the simulated genomes. In general, the duplicated portion of the genome grows as n and W increase and as R decreases, and m has no effect as inversions only happen within duplicated regions in the simulation. Fig. 4.4 shows the percentage of the originally duplicated regions recovered by AB-Synteny (*recovery rate*) as a function of the gene loss rate R. The recovery rate is defined as

$$\frac{|\cup S^i_{AB}| - |\cup S^i_{AB} \setminus \cup S^i_{orig}|}{|\cup S^i_{orig}|},$$

where S^i_{AB} are the sets of synteny blocks generated by AB-Synteny, S^i_{orig} are the original duplicated blocks, and the function $|S|$ gives the size (the number of genes) of a block or a union of blocks.

When the number of duplications is small, the duplicated regions contain a small number of genes and the same number of inversions cause more disruption to the gene order, making the duplication detection harder, resulting in a reduced recovery rate. As W increases, the duplicated regions involve more and more genes, making the detection of duplicated regions easier, thus increasing the recovery rate. The recovery rate decreases obviously as the gene loss rate grows. The number of inversions within duplicated regions does not seem to have an obvious effect on the recovery rate.

The performance of AB-Synteny on simulated genomes with segmental duplications is shown in Fig. 4.5. Fig. 4.5(a) shows the proportional size of true duplications and false duplications identified by AB-Synteny, and Fig. 4.5(b) shows the recovery rate (true positives) of AB-Synteny as a function of its false positives. The formulas for these in terms of genome length N' are

$$\begin{aligned}
\text{true duplications} &= \frac{|\cup S^i_{AB}| - |\cup S^i_{AB} \setminus \cup S^i_{orig}|}{N'} \\
\text{false duplications} &= \frac{|\cup S^i_{AB} \setminus \cup S^i_{orig}|}{N'} \\
\text{recovery rate} &= \frac{|\cup S^i_{AB} \setminus \cup S^i_{orig}|}{N' - |\cup S^i_{orig}|}
\end{aligned}$$

The simulated genomes with whole genome duplication are generated by running Simulate_WGD with following parameters: $N = 4000$; $M = 50, 100, 150, 200$; $m = 0, 500$; $w = 5$; $R = 0, 30, 35, 40, 45, 50, 60$

We applied AB-Synteny$(.,.,10,40,3)$ without pre-processing to the simulated genomes. Of all the originally recorded duplicated blocks in a simulated genome, we discard a pair of duplicated blocks if they have only one anchor or if both blocks

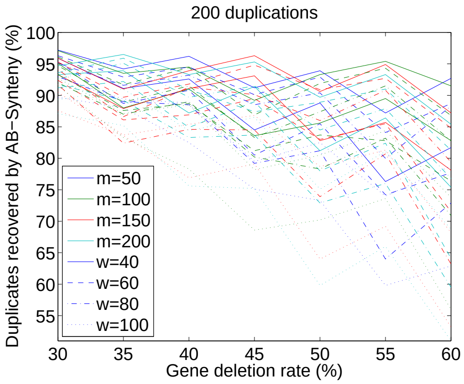


FIG. 4.4. *Recovery rate of duplicated regions by AB-Synteny on simulated genomes with $n = 200$ segmental duplications. Colors represent various total numbers of inversions (m). Line-styles represent various maximum spans of duplications (W).*

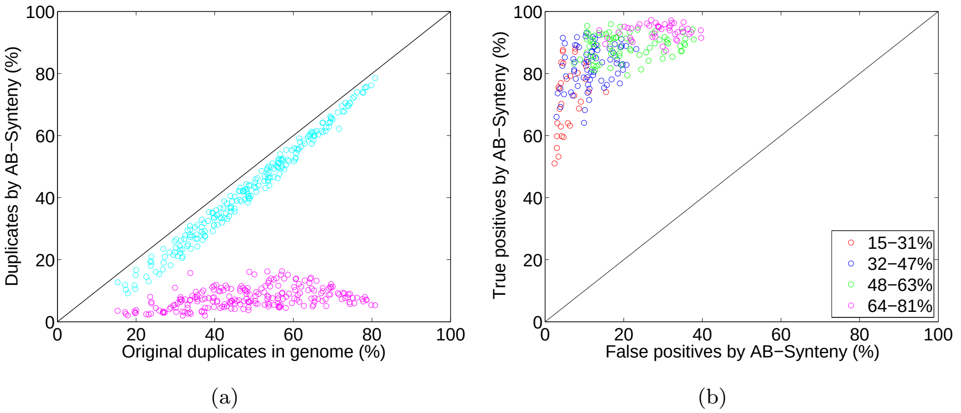
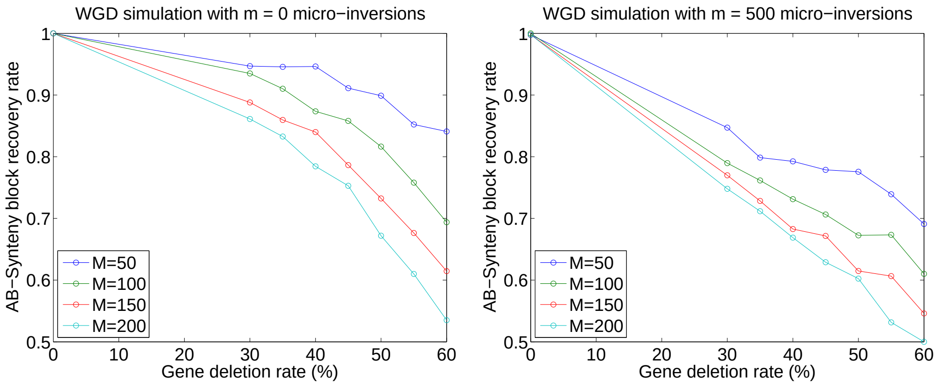


FIG. 4.5. *On simulated genomes with segmental duplications: (a) True (blue color) and false (magenta color) duplications as constructed by AB-Synteny. (b) True duplication recovery rate versus false duplication recovery rate by AB-Synteny. Colors represent the percentage of the simulated genomes that are originally duplicated.*

have fewer than three genes. The remaining blocks are considered as *real* duplicated blocks. They do not, however, necessarily represent recoverable duplications.

The block recovery rate of AB-Synteny is defined as $\frac{N_{ab}}{N_{orig}}$ where N_{ab} is the number of real duplicated blocks *covered* by AB-Synteny blocks, and N_{orig} is the total number of real duplicated blocks. A block is covered by another block if they overlap by at least two genes. The block recovery rate of AB-Synteny on the simulated genomes with WGD are shown in Fig. 4.6 as a function of gene loss rate R. At gene loss rate R, two duplicated regions are expected to share $\frac{(100-R)^2}{100}$ percent of their genes.

5. Discussion. The uniqueness of our new synteny block generation algorithm AB-Synteny stems from the fact that it produces synteny blocks that do not overlap



(a)

(b)

Fig. 4.6. *Block recovery rate of duplicated regions by AB-Synteny as a function of gene loss rate (R) on simulated genomes with WGD. Colors represent various numbers of macro-inversions (M). (a) Without micro-inversion (b) 500 micro-inversions.*

in 1-D representations. None of the conventional synteny block generation algorithms have this property, which is essential for further analysis of the synteny blocks to study rearrangement and duplication history.

AB-Synteny can also be used for generation of synteny blocks across multiple genomes. Given k genomes, one simply concatenates them into a single genome. If there are no duplications in the original genomes, then the edges with multiplicity k in the A-Bruijn graph correspond to synteny blocks shared by all k genomes.

Even though we treat synteny block generation for a single genome and for multiple genomes as the same problem, we remark that the alignment anchors in the two cases have different characteristics. Inside the underlying synteny blocks, alignment anchors across multiple genomes are usually better preserved than within a single genome. Since genes are often deleted or diverged over time such that the similarity is beyond recognition after duplication events, some of the underlying duplicated regions may not share many alignment anchors, making it difficult to detect the true extent of duplications.

Our AB-Synteny algorithm constructs A-Bruijn graphs using the the RepeatGluer code initially developed for repeat classification and DNA fragment assembly. Extending RepeatGluer to new research domains typically requires new application-specific algorithmic developments (e.g., constructing A-Bruijn graphs in mass spectrometry applications [44]). Similarly, the synteny block reconstruction may benefit from the modifications of the A-Bruijn graph approach that take into account the specific challenges of analyzing large highly duplicated genomes. We found that while most RepeatGluer steps (e.g., bulge removal) work well for synteny block generation, some steps need to be further optimized for the new application domain. In particular, we found that the *threading* heuristic from [41] (which worked well for fragment assembly) may lead to suboptimal results in synteny block reconstruction. Optimizing the A-Bruijn graph approach for synteny block generation represents the next challenge in analyzing the genomic architecture of the quickly increasing set of mammalian and plant genomes that are being sequenced using next generation sequencing technologies.

Acknowledgments. We thank Haixu Tang for providing the RepeatGluer code for A-Bruijn graph construction. We thank Joseph R. Ecker for insightful comments. Qian Peng was supported by NSF (Plant Systems Biology) IGERT Training Grant DGE0504645. Glenn Tesler was supported by NSF Grant DMS-0718810.

REFERENCES

[1] Q. Peng, M.A. Alekseyev, G. Tesler, and P.A. Pevzner, *Decoding synteny blocks and large-scale duplications in mammalian and plant genomes.* Proceedings of WABI'09 (2009)

[2] J.F. Wendel, *Genome evolution in polyploids.* Plant Molecular Biology, 42:1(2000), pp. 225–249.

[3] G. Blanc, A. Barakata, R. Guyota, R. Cookea, and M. Delseny, *Extensive duplication and reshuffling in the arabidopsis genome.* Plant Cell, 12(2000), pp. 1093–1102.

[4] T.J. Vision, D.G. Brown, and S.D. Tanksley, *The Origins of Genomic Duplications in Arabidopsis.* Science, 290:5499(2000), pp. 2114–2117.

[5] G. Blanc, K. Hokamp, and K.H. Wolfe, *A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome.* Genome Research, 13:2(2003), pp. 137–144.

[6] E. Lander, L. Linton, B. Birren, and C. Nusbaum et al., *Initial sequencing and analysis of the human genome.* Nature, 409(2001), pp. 860–921.

[7] J. Bailey, Z. Gu, R.A., C., Reinert, K., R.V., S., S. Schwartz, M. Adams, E. Myers, P. Li, and E. Eichler, *Recent segmental duplications in the human genome.* Science, 297(2002), pp. 1003–1007.

[8] J. Bailey, R. Baertsch, W. Kent, D. Haussler, and E. Eichler, *Hotspots of mammalian chromosomal evolution.* Genome Biology, 5:4(2004), pp. R23.

[9] G. Bourque, P.A. Pevzner, and G. Tesler, *Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons From Human, Mouse, and Rat Genomes.* Genome Res., 14:4(2004), pp. 507–516.

[10] P. Pevzner and G. Tesler, *Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.* PNAS, 100:13(2003), pp. 7672–7677.

[11] P. Pevzner and G. Tesler, *Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes.* Genome Research, 13(2002), pp. 37–45.

[12] Q. Peng, P. Pevzner, and G. Tesler, *The fragile breakage versus random breakage models of chromosome evolution.* PLoS Computation Biology, 2:2(2006), pp. e14.

[13] G. Tesler, *Grimm: genome rearrangements web server.* Bioinformatics, 18:3(2002), pp. 492–493.

[14] S. Hannenhalli and P. Pevzner, *Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals.* Journal of ACM, 46(1999), pp. 1–27.

[15] J. Nadeau and B. Taylor, *Lengths of chromosomal segments conserved since divergence of man and mouse.* Proceedings of the National Academy of Sciences USA, 81(1984), pp. 814–818.

[16] N. Copeland, N. Jenkins, D. Gilbert, J. Eppig, L. Maltais, J. Miller, W. Dietrich, A. Weaver, S. Lincoln, and R. Steen et al., *A genetic linkage map of the mouse: Current applications and future prospects.* Science, 262(1993), pp. 57–66.

[17] R. DeBry and M. Seldin, *Human/mouse homology relationships.* Genomics, 33(1996), pp. 337–351.

[18] S. O'Brien, M. Menotti-Raymond, W. Murphy, W. Nash, J. Wienberg, R. Stanyon, N. Copeland, N. Jenkins, J. Womack, and J. Graves, *The promise of comparative genomics in mammals.* Science, 286(1999), pp. 458–481.

[19] R. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. Abril, P. Agarwal, R. Agarwala, R. Ainscough, and M. P. A. Alexanderson et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 420(2002), pp. 520–562.

[20] W.J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, *Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes.* PNAS, 100:20(2003), pp. 11484–11489.

[21] W. Fujibuchi, H. Ogata, H. Matsuda, and M. Kanehisa, *Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping.* Nucl. Acids Res., 28:20(2000), pp. 4029–4036.

[22] I.B. Rogozin, K.S. Makarova, J. Murvai, E. Czabarka, Y.I. Wolf, R.L. Tatusov, L.A. Szekely, and E.V. Koonin, *Connected gene neighborhoods in prokaryotic genomes.* Nucl. Acids Res., 30:10(2002), pp. 2212–2223.

[23] M. Brudno, S. Malde, A. Poliakov, C. Do, and O. Couronne et al., *Glocal alignment: Finding rearrangements during alignment.* Bioinformatics, 19(2003), pp. i54–i62.

[24] P. Calabrese, S. Chakravarty, and T. Vision, *Fast identification and statistical evaluation of segmental homologies in comparative maps.* Bioinformatics, 19(2003), pp. i74–i80.

[25] A. Darling, B. Mau, F. Blattner, and P. NT, *Mauve: Multiple alignment of conserved genomic sequence with rearrangements.* Genome Research, 14(2004), pp. 1394–1403.

[26] A. Darling, B. Mau, F. Blattner, and N. Perna, *Gril: Genome rearrangement and inversion locator.* Bioinformatics, 20(2004), pp. 122–124.

[27] G. Bourque, Y. Yacef, and N. El-Mabrouk, *Maximizing synteny blocks to identify ancestral homologs.* Proc. of the 3rd RECOMB on Comparative Genomics RECOMB-CG'05, 3678(2005), pp. 21–34.

[28] F. Swidan, E.P.C. Rocha, M. Shmoish, and R.Y. Pinter, *An integrative method for accurate comparative genome mapping.* PLoS Computational Biology, 2(2006), pp. e75.

[29] C.N. Dewey, P.M. Huggins, K. Woods, B. Sturmfels, and L. Pachter, *Parametric alignment of drosophila genomes.* PLoS Comput Biol, 2:6(2006), pp. e73.

[30] J. Ma, L. Zhang, B.B. Suh, B.J. Raney, R.C. Burhans, W.J. Kent, and M. Blanchette, *Reconstructing contiguous regions of an ancestral genome.* Genome Research, 16(2006), pp. 1557–1565.

[31] A. Sinha and J. Meller, *Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms.* BMC Bioinformatics, 8:1(2007), pp. 82.

[32] T. Hachiya, Y. Osana, K. Popendorf, and Y. Sakakibara, *Accurate identification of orthologous segments among multiple genomes.* Bioinformatics, 25:7(2009), pp. 853–860.

[33] M. Kellis, B.W. Birren, and E.S. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae.* Nature, 428:6983(2004), pp. 617–624.

[34] J.E. Bowers, B.A. Chapman, J. Rong, and A.H. Paterson, *Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.* Nature, 422(2003), pp. 433–438.

[35] S. Hampson, A. McLysaght, B. Gaut, and P. Baldi, *LineUp: Statistical Detection of Chromosomal Homology With Application to Plant Comparative Genomics.* Genome Res.,

13:5(2003), pp. 999–1010.

[36] S. Hampson, B. Gaut, and P. Baldi, *CloseUp: Statistical Detection of Chromosomal Homology Using Shared-Gene Density Alone.* Bioinformatics, 21:8(2005), pp. 1339–1348.

[37] B.J. Haas, A.L. Delcher, J.R. Wortman, and S.L. Salzberg, *DAGchainer: a tool for mining segmental genome duplications and synteny.* Bioinformatics, 20:18(2004), pp. 3643–3646.

[38] K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van de Peer, *The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity between Arabidopsis and Rice.* Genome Res., 12:11(2002), pp. 1792–1801.

[39] C. Simillion, K. Janssens, L. Sterck, and Y. Van de Peer, *i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles.* Bioinformatics, 24:1(2008), pp. 127–128.

[40] C. Soderlund, W. Nelson, A. Shoemaker, and A. Paterson, *SyMAP: A system for discovering and viewing syntenic regions of FPC maps.* Genome Research, 16:9(2006), pp. 1159–1168.

[41] P.A. Pevzner, H. Tang, and G. Tesler, *De Novo Repeat Classification and Fragment Assembly.* Genome Res., 14:9(2004), pp. 1786–1796.

[42] B. Raphael, D. Zhi, H. Tang, and P. Pevzner, *A novel method for multiple alignment of sequences with repeated and shuffled elements.* Genome Res., 14:11(2004), pp. 2336–2346.

[43] D. Zhi, B. Raphael, A. Price, H. Tang, and P. Pevzner, *Identifying repeat domains in large genomes.* Genome Biology, 7:1(2006), pp. R7.

[44] N. Bandeira, K.R. Clauser, and P.A. Pevzner, *Shotgun Protein Sequencing: Assembly of Peptide Tandem Mass Spectra from Mixtures of Modified Pro teins.* Mol Cell Proteomics, 6:7(2007), pp. 1123–34.

[45] G. Bourque, E.M. Zdobnov, P. Bork, P.A. Pevzner, and G. Tesler, *Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages.* Genome Res., 15:1(2005), pp. 98–110.

[46] C.N. Dewey and L. Pachter, *Mercator: Multiple whole-genome-orthology map construction.* http://bio.math.berkeley.edu/mercator (2006)

[47] Z. Bao and S.R. Eddy, *Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes.* Genome Res., 12:8(2002), pp. 1269–1276.

[48] Z. Jiang, H. Tang, M. Ventura, M.F. Cardone, T. Marques-Bonet, X. She, P.A. Pevzner, and E.E. Eichler, *Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.* Nature Genetics, 11(2007), pp. 1361–1368.