Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations

Arjun Bhutkar[1,2] arjunb@morgan.harvard.edu
Susan Russo[1] russo@morgan.harvard.edu
Temple F. Smith[2] tsmith@darwin.bu.edu
William M. Gelbart[1] gelbart@morgan.harvard.edu

[1] Department of Molecular and Cellular Biology, Harvard University, Cambridge MA 021383, USA
[2] BioMolecular Engineering Research Center, Boston University, 36 Cummington St., Boston MA 02215, USA

Abstract

Genome scale synteny analysis, the analysis of relative gene-order conservation between species, can provide key insights into evolutionary chromosomal dynamics, rearrangement rates between species, and speciation analysis. With the rapid availability of multiple genomes, there is a need for efficient solutions to aid in comparative syntenic analysis. Current methods rely on homology assessment and multiple alignment based solutions to determine homologs of genetic markers between species and to infer syntenic relationships. One of the primary challenges facing multi-genome syntenic analysis is the uncertainty posed by genome assembly data with un-sequenced gaps and possible assembly errors. Currently, manual intervention is necessary to tune and correct the results of homology assessment and synteny inference. This paper presents a novel automated approach to overcome some of these limitations. It uses a graph based algorithm to infer sub-graphs denoting synteny chains with the objective of choosing the best locations for homologous elements, in the presence of paralogs, in order to maximize synteny. These synteny chains are expanded by merging sub-graphs based on various user defined thresholds for micro-syntenic scrambling. This approach comprehends and accommodates for contig and scaffold gaps in the assembly to determine homologous genetic elements that might either fall in unsequenced assembly gaps or lie on the edges of sequenced segments or on small fragments. Furthermore, it provides an automated solution for breakpoint analysis and a comparative study of chromosomal rearrangements between species. This approach was applied to a comparative study involving *Drosophila.melanogaster* and *Drosophila.pseudoobscura* genomes, as an example, and has been useful in analyzing inter-species syntenic relationships.

Keywords: genome assembly gaps, synteny, breakpoints, comparative genomics

1 Introduction

The analysis of synteny, the study of relative gene-order conservation between genomes, can provide key insights into evolutionary chromosomal dynamics, rearrangement rates between species, and speciation analysis. With the rapid availability of multiple genomes, there is a need for efficient solutions to aid in comparative syntenic analysis. Current methods [2, 3, 5] rely on homology assessment and multiple alignment based solutions to determine homologs of genetic markers between species and to infer syntenic relationships. One of the primary challenges facing multi-genome syntenic analysis is the uncertainty posed by genome assembly data with un-sequenced gaps and possible assembly errors. Currently, manual intervention is necessary to tune and correct the results of homology assessment and synteny inference. This paper presents a novel automated approach to overcome these limitations. It

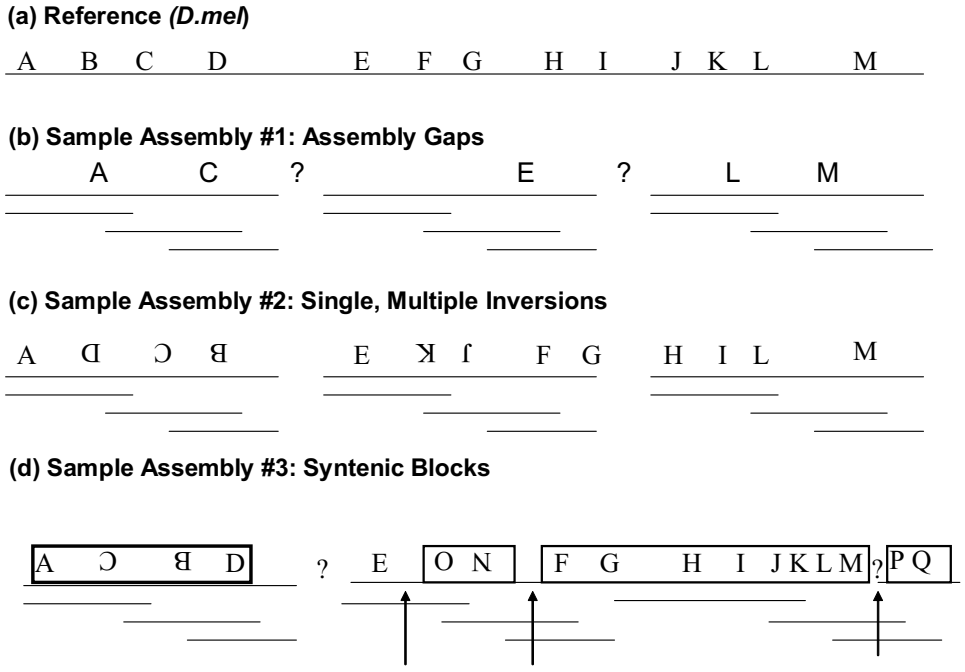


Figure 1: Sample genome assembly gaps, gene-order inversions, and synteny compared to a reference genome. (a) The order of genes along a chromosome of the reference species (in this case *D.melanogaster*). (b) Sample genome assembly #1: Overlapping reads resulting in contigs for a sample species other than *D.mel* are shown. Homologs of *D.mel* genes are shown on the contigs. Question marks represent inter-contig gaps (similar issue with scaffolds) that might contain the true homologs of missing genes (D, F, G, H, I, J, K). (c) Sample genome assembly #2: Inversions are illustrated in a species in comparison to the D.mel gene order. A single inversion flips the segment containing genes B,C,D and two inversions result in the inversion and translocation of the segment containing genes J, K. (d) Sample genome assembly #3: Syntenic blocks with respect to the gene order along the reference species in (a). Inversions cause gene-order changes leaving E as a singleton. Localized scrambling (micro-synteny) is seen in the block A, C, B, D. Arrows point to some syntenic breaks either due to synteny interruption or assembly gaps.

uses a graph based algorithm to infer sub-graphs denoting synteny chains with the objective of choosing the best locations for homologous elements, in the presence of alternates, in order to maximize synteny. These synteny chains are expanded by merging sub-graphs based on various user defined thresholds for micro-syntenic scrambling. This approach comprehends and accommodates for contig and scaffold gaps in the assembly to determine homologous genetic elements that might either fall in un-sequenced assembly gaps or lie on the edges of sequenced segments or on small fragments. Furthermore, it provides an automated solution for breakpoint analysis and a comparative study of chromosomal rearrangements between species. This approach was applied to *Drosophila* genomes and has been useful in analyzing inter-species syntenic relationships and in uncovering assembly errors in other cases.

There are a number of challenges in dealing with assemblies that have un-sequenced gaps between contigs or scaffolds. The analysis is typically performed for the genome assembly of a newly sequenced species with respect to the gene order on the chromosome(s) of a reference species. In this study,

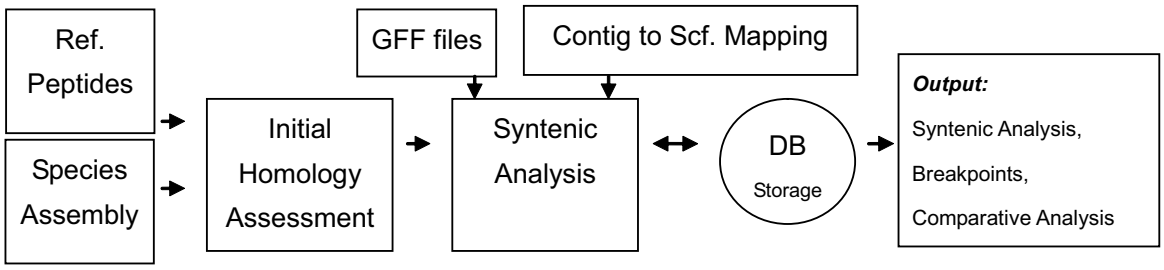


Figure 2: Simplified system and data-flow overview. The reference peptide set is from an annotated reference species (in this case *D.mel*). This is used along with the genome assembly for the species to be analyzed as input to the initial homology assessment stage. The syntenic analysis process (see text and figure 3) refines the homology assessment, performs synteny analysis, and accounts for assembly gaps. GFF files with gene coordinate information from the reference species and contig-to-scaffold assembly mapping files are used as supplementary material.

Drosophila melanogaster (D.mel) was used as the reference species, given the wealth of annotated data for this genome. Assembly gaps in a newly sequenced genome might interrupt syntenic blocks with reference to a known gene order and might contain homologs for one or more genes. Gene order can also be shuffled due to paracentric inversions along chromosome arms that disrupt gene order between species and breakup or translocate syntenic blocks. Figure 1 outlines some of these issues. In this study, a comparative analysis between *D.melanogaster* and *D.pseudoobscura* [5] was performed. Given that translocation of genes between chromosome arms is rare in *Drosophila* [5], most genes have orthologs on homologous chromosome arms. This simplifies the analysis somewhat.

This study addressed the problem of inferring syntenic blocks in a candidate species (in this case *D.pseudoobscura*) in the presence of assembly gaps and paralogs. Alternate placements for orthologous genes were selected in order to maximize synteny. In the presence of gene duplication and paralogs, the best candidate placement that maximized the syntenic relationship was chosen. In a number of cases, the syntenic context of neighboring genes allowed the inference of a gene being in an assembly gap or at the edge of an assembly contig. This paper describes the method and algorithm employed to automate the determination of syntenic blocks within the assembly of a newly sequenced species while overcoming some of the limitations posed by gaps in the assembly data. The *D.mel* gene order was used as the reference order. Results showing the syntenic inference for a sample species are shown.

2 Method and Results

2.1 Simplified System Overview

Syntenic analysis is done with reference to the gene order of a reference species whose set of annotated peptides form the primary input along with the genome assembly of the species under investigation. Initial homology assessment can be performed with a tool such as tBLASTn [1]. The syntenic analysis process, described later, refines the homology assessment and goes through various scenarios to maximize the synteny for this species in the presence of inter-contig and inter-scaffold assembly gaps. Annotation details (via GFF files) for the reference species provide reference synteny information and chromosome arm location coordinates for genes in the reference species. Contig to scaffold mapping files for the genome assembly provide information on location of inter-contig gaps and scaffold sizes. A backend database is used to store preliminary and final results. The primary output of this system consists of: a list of syntenic blocks and the placement of genes therein, a list of reciprocal breakpoints, a list of linked breakpoints via reuse of breakpoint boundaries, and a comparative analysis (example: shared breakpoints) between various lineages with respect to a reference species. Figure 2 presents an overview of the data flow.

2.2 Syntenic Analysis

The syntenic analysis process starts with initial homology assessment for genes from a reference species, as described before. A graph based algorithm builds sub-graphs of syntenic blocks and merges these sub-graphs to form longer syntenic blocks depending on user-defined thresholds to allow localized scrambling (micro-synteny). Key steps in this process, including collision resolution and breakpoint inference, are shown in Figure 3 and described below:

- An initial vertex set for the graph is obtained from preliminary homology assignments as described before. This is shown in Figure 3(a).

- Edges are then added to the graph to link neighboring vertices (genes) in the same syntenic order as in the reference species, where possible. This is shown in Figure 3(b). This results in a set of disconnected sub-graphs and leaves some genes as singletons, i.e. these vertices do not have an edge incident upon them. These sub-graphs can be further linked through special edges denoting the sequence and orientation of syntenic blocks along a scaffold of the candidate assembly.

- Allowing for missing genes in this species and a user-defined threshold for localized scrambling (micro-synteny, or allowing genes to be out-of-order with respect to the reference species), sub-graphs are merged to form longer chains of synteny as seen in Figure 3(c). The default threshold used for localized scrambling was 10 genes, where a set of ten or less neighboring genes could appear together and be allowed to be out of order with respect to the reference species.

- Refining the placement of singletons is the next step in this process. Two problems dominate: the inference of homologous genes in the presence of paralogs, and the lack of information due to gaps in the assembly. The primary objective is to refine the initial homology placement using knowledge of genome assembly gaps and to maximize synteny in comparison to the reference species. A number of different scenarios are considered to see if singletons could be placed in a better syntenic location. Table 1 shows the various scenarios that are considered in the presence of paralogs, assembly gaps, or missing data. The location of syntenic neighbors could lead to the inference of a gene actually having an orthologous location that is within an assembly gap. In some cases, this inference might be reinforced by the presence of partial hits to the edge on an assembly contig or partial hits on multiple intermediate contigs between syntenic neighbors. In the presence of paralogs, better syntenic placement for a gene can be inferred through alternate locations next to one of its syntenic neighbors, if it exists. These and other cases are listed in Table 1 along with the actions taken in each case. The process is also shown in Figure 3(d).

- The homology assignment process and singleton placement refinement process can lead to "collisions" – the overlapping of homologous gene assignment areas due to similar coding domains between genes. This is shown if Figures 3(d) and 3(e). In order to untangle these cases as far as possible, reference species' homology to itself (in this case *D.mel* to *D.mel*) is used. This gives the expected cases of collisions due to similar coding domains. Using this information, collisions are classified as those due to clusters of multiple neighboring genes with sequence similarity, genes with known overlapping coding domains, and singletons placed incorrectly where they collide with paralogs in syntenic blocks. As far as possible, these collisions are resolved where they can be assigned neighboring locations that do not overlap (in the case of clusters), marked as expected collisions (overlapping gene cases), or dropped from the analysis (singletons hitting paralogs).

- Breakpoint identification can then be performed on this graph consisting of disconnected paths denoting syntenic blocks (as noted before, special edges are used to link blocks in order, on a scaffold of the candidate assembly). Reciprocal inversion breakpoints are identified using genes

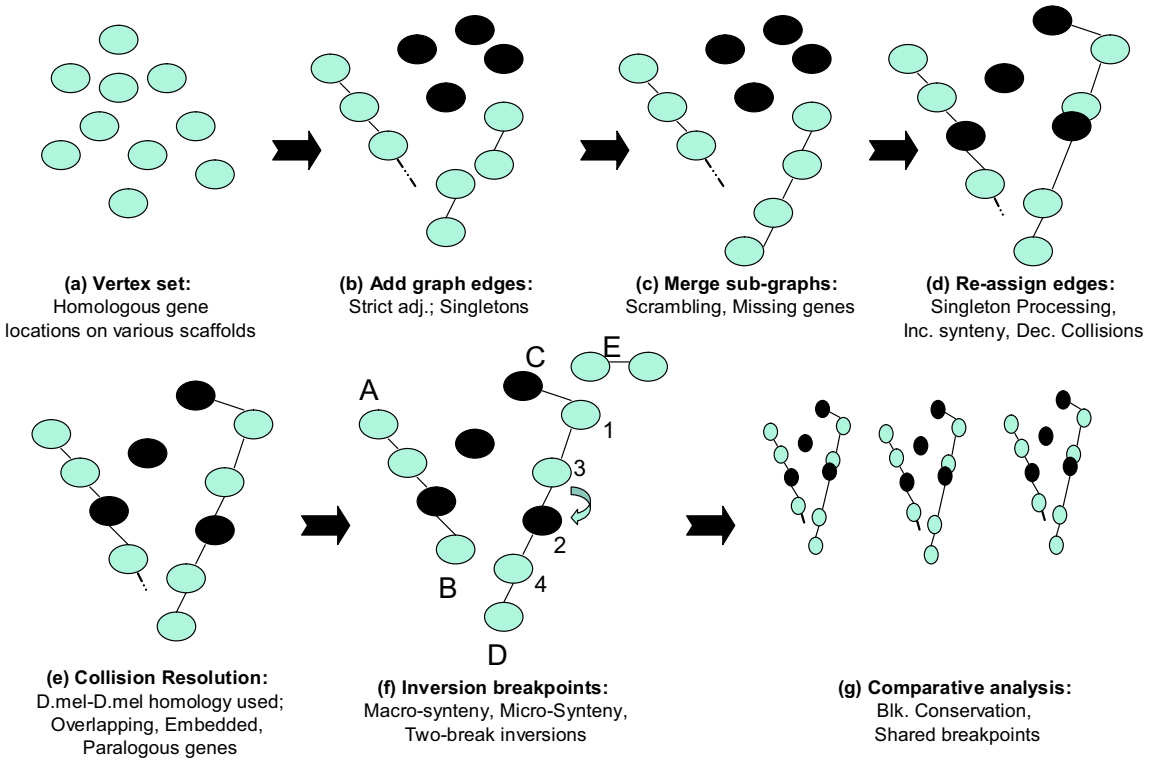


Figure 3: Graphical representation of the syntenic process. (a) The vertex set of homologous gene locations. (b) Chained sub-graphs and singletons (dark circles). (c) Merged sub-graphs to form longer chains of synteny. (d) Singleton relocation to maximize synteny. (e) Collision resolution. (f) Inversion breakpoint identification where there is a breakpoint between the syntenic blocks A-B and C-D. The reversal of the 2-3 segment is an example of micro-synteny. (g) Three sample genomes that have been processed in this manner which can now be analyzed from a comparative standpoint. See text for further details.

at the edges of these syntenic blocks. These are two-break events where an inversion would reconstruct the gene-order in the reference species. This process accommodates for missing genes or genes in assembly gaps.

□ In the case where multiple candidate species' assemblies are evaluated with respect to a reference species, comparative analysis across the graphs for various species is possible. Determining shared conserved blocks, longest common paths of synteny, and shared breakpoint events are some of the areas that can be easily analyzed using this approach.

2.3 Results

This method was applied to a comparison between *D.melanogaster* and *D.pseudoobscura*, using the former as the reference species. In order to illustrate the results, two homologous chromosome arms (*Drosophila* Muller elements [4]), E and C, were chosen. For each arm, dot plots mapping the location of homologous genes are presented in two parts. In the first plot the cases where singletons were moved to a better syntenic location by this method are omitted and their location is represented by an empty circle. In the second plot, these singletons are included. This demonstrates the cases where singletons are indeed moved to a better syntenic context, circles overlapping syntenic segments,

Table 1: Singleton placement to maximize syntenic location in the presence of assembly gaps and paralogs.

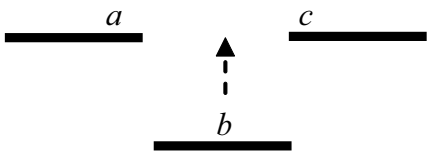
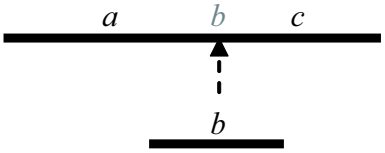
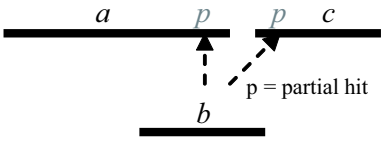
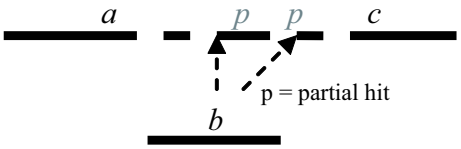
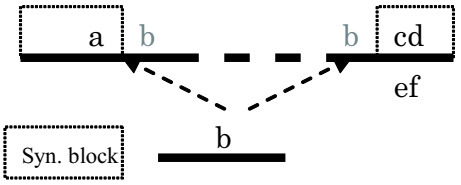
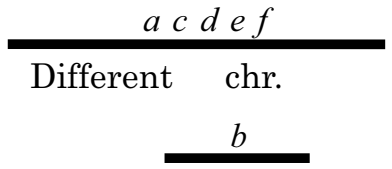
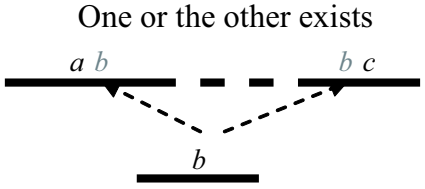
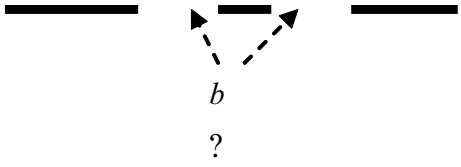
#	Type/Conditions	Example
1	*In assembly gap:* Syntenic neighbors on adjacent contigs and an inter-contig gap exists. Lack of alternate hits that maximize synteny. Gene is assumed to lie in assembly gap. Similar logic for scaffold gaps. **Action:** Ortholog location is classified as such and information is utilized in syntenic analysis.	
2	*Maximize syntenic neighborhood on same contig:* Syntenic neighbors on same contig and alternate hit for this gene on that contig. **Action:** Ortholog placement is moved to better syntenic location	
3	*Partial hits:* Syntenic neighbors on adjacent contigs and partial hit(s) to edges of either (or both) contig(s) **Action:** Ortholog placement is moved to the edge of the corresponding contig(s)	
4	*Intermediate hits:* Syntenic neighbors on different contigs on the same scaffold with other contigs in between. Partial hits exist on these intermediate contig(s). **Action:** Ortholog placement is moved to lie in between the two syntenic neighbors across intermediate contigs	
5	*Rearrangement edge:* Syntenic neighbors on the edges of two different syntenic blocks that are not adjacent. Gene has lower hit on same contig as one of the syntenic neighbors (on edge of rearrangement). **Action:** Ortholog placement is moved to the edge of the corresponding syntenic block	
6	*Non-collision hit on correct chromosome arm:* Best location on different chromosome arm when a non-collision hit exists on a scaffold mapped to the correct arm. **Action:** Ortholog placement is moved to the original chromosome arm, ignoring the paralog. This is based on observation of Drosophila arm conservation.	

Table 1: (continued)

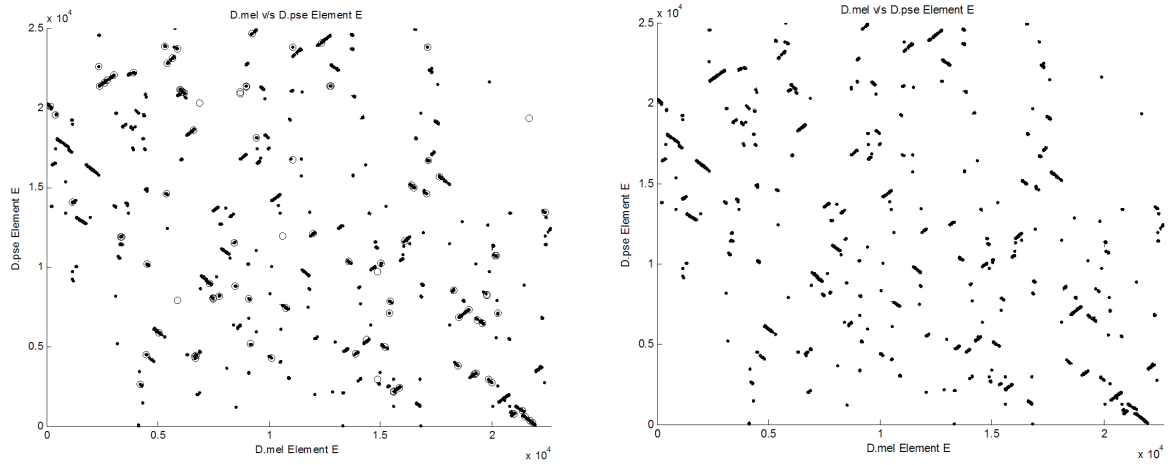
7	*Translocated/Transposed genes:* These are candidates for gmoved genesh where syntenic neighbors are adjacent without any gap between them (on same contig or on adjacent contig edges without any inter-contig gap). Two possibilities exist: translocated on the same arm or transposed to a different arm. **Action:** Ortholog placement is moved to the translocated / transposed location	
8	*Maximize synteny:* Probable placement next to a syntenic neighbor irrespective of where that neighbor is located and whether or not it is in a syntenic block. **Action:** Ortholog placement at the alternate location	
9	*In unknown assembly gap or missing:* Placements that cannot be categorized in any of the above categories are bucketed as either falling in some un-sequenced region of the assembly or missing in this species. **Action:** Ortholog location is classified as such and information is utilized in syntenic analysis	

resulting in longer syntenic chains. In the second plot large blocks of conserved gene order are seen. This analysis permitted a maximum localized scrambling of ten genes. A lower threshold would further break apart these blocks whereas a higher threshold would result in some larger blocks of synteny. Figures 4 and 5 show the relevant plots for homologous chromosomes representing Muller elements E and C, respectively.

In this analysis between the two genomes, 24 genes were inferred as probably being in un-sequenced gaps in the *D.pseudoobscura* assembly, 47 genes were placed close to assembly gaps near their syntenic neighbors based on weak signals at the end of contigs bordering an assembly gap, 156 gene placements were moved to alternate locations within a better syntenic context, and 118 gene placements were moved to a syntenic location that put them adjacent to one of their syntenic neighbors when they are on the border of a rearrangement. Out of a total of over 13,000 *D.mel* genes used in this analysis, 384 orthologous placements in *D.pseudoobscura* were moved to a better syntenic location, improving the measure of overall synteny between the two species.

3 Discussion

The analysis of syntenic information derived from this study provides insights into the evolutionary chromosomal rearrangements between species. It provides a rich dataset for further analysis and biological discovery. The correct placement of homologous genes between species is important in order to prevent incorrect inferences of gene transpositions, as an example. The analysis of chromosomal rearrangements between species depends upon the correct identification of syntenic relationships and conserved blocks. This analysis aids in minimizing the impact of assembly gaps and paralogous relationships between species. It is especially useful in analyzing early draft assemblies of a candidate species that might have low coverage. In addition to syntenic analysis, this approach was also successful



(a) (b)

Figure 4: *D.melanogaster* v/s *D.pseudoobscura* synteny for homologous chromosomes (**Element E** [4]). Contiguous segments show blocks of synteny. (a) Circles represent placeholders for locations where singletons were moved in order to maximize overall synteny. (b) Plot after including the moved singletons at their new location as part of syntenic blocks. Compressed scale to accommodate all homologous genes.

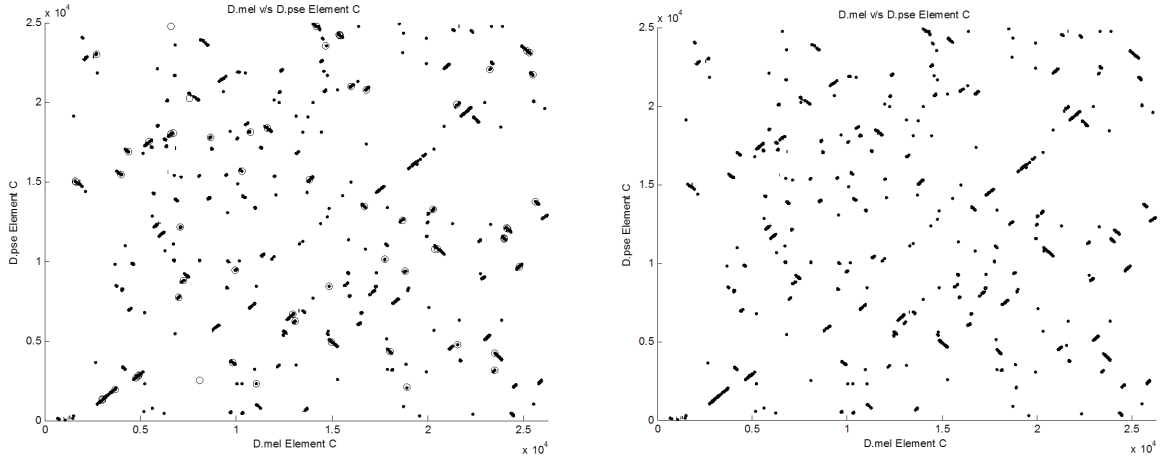
in uncovering genome assembly errors that were useful in improving the quality of the assembly (data not shown). While there are single gene transpositions in *Drosophila*, transposition of large segments is rare, other than a few known cases of pericentric inversions involving the centromere, and some fusions. This makes it easier to use a synteny based approach to uncover possible assembly mis-joins and distinguish them from rearrangements due to paracentric inversions that involve elements from the same arm.

This analysis led to proper placement of 384 genes in syntenic blocks where they were marked as singletons in the initial analysis. Overall, the number of syntenic blocks inferred between the two species (block size greater than one gene) was 983 blocks, up from 921 blocks reported previously [5].

Although this approach improves upon the placement of singletons in the right syntenic context, it suffers from some of the limitations inherent in such analyses. There is only so much one can do in the presence of missing data such as assembly gaps. For cases that fall into the last category in Table 1, one cannot ascertain whether a gene is missing from a given species or whether it is in some assembly gap that cannot be pinpointed. Further, there is always the case of ambiguity due to the presence of paralogs and gene duplications in such an approach. While this algorithm compensates for some of these problems by selecting the best length normalized hit for a gene in the best syntenic location, this might not be enough in the case of tandem duplications with localized scrambling. Also, this study involved two species that are closely related (25-55 MY separation) and hence share a large number of genes. In such cases syntenic relationships are more informative and can be used to compensate for missing information in the presence of assembly gaps. We have applied this approach to genomes with different divergence rates and have seen good performance over the range of 40-60 MY with lessening of predictive power to compensate for missing information or assembly gaps, due to large number of rearrangements, over the 250-300 MY range (data not shown).

4 Conclusions

For accurate syntenic analysis between genomes, it is essential to account for the impact of genome assembly gaps and for the presence of paralogs. This effort provides a classification of the various



(a) (b)

Figure 5: *D.melanogaster* v/s *D.pseudoobscura* synteny for homologous chromosomes (**Element C** [4]). Contiguous segments show blocks of synteny. (a) Circles represent placeholders for locations where singletons were moved in order to maximize overall synteny. (b) Plot after including the moved singletons at their new location as part of syntenic blocks. Compressed scale to accommodate all homologous genes.

scenarios for singleton placement refinement and an algorithm for synteny inference in the presence of paralogs, assembly limitations, collisions, and missing data. It speeds up the analysis process which can now be performed on multiple versions of the assembly as it is improved, rather than undergoing a manual process to refine the results, as before. This allows the focus to be on biologically relevant analysis rather than time being spent on manual data refinement. Further, the identification of probable assembly errors provides feedback to the sequencing centers for improving assembly quality by managing various assembler parameters. Numerous genes that would otherwise be considered missing, or in incorrect locations, can now be analyzed as part of syntenic blocks in their correct homologous locations in the newly sequenced species. This results in more precise downstream analysis, including rearrangement analysis between species, as an example.

Acknowledgments

The authors wish to acknowledge the support of various Drosophila sequencing centers, the Harvard FlyBase team (including former members Brian Bettencourt, Stan Letovsky, and Pavel Hradecky), Bio-Molecular Engineering Research Center (BMERC Boston University) computing and support staff, Harvard's Center for Genomic Research (CGR) computing staff, and the fly research community. This work was supported by a subcontract from Harvard University under NIH grant HG000739.

References

[1] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D., Basic local alignment search tool, *J. Mol. Biol.*, **215**:403–410, 1990.

[2] Kube, M., Beck, A., Zinder, S.H., Kuhl, H., Reinhardt, R., and Adrian, L., Genome sequence of the chlorinated compound-respiring bacterium Dehalococcoides species strain CBDB1, *Nat. Biotechnol.*, **10**:1269–1273, 2005.

[3] Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, *Nature*, **420**:520–562, 2002.

[4] Muller, H.J., The new systematics, ed. J. Huxley. Clarendon Press, Oxford, UK, 185–268, 1940.

[5] Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., *et al.*, Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution, *Genome Res.*, **1**:1–18, 2005.