Scaffold assembly based on genome rearrangement analysis

Sergey Aganezov and Max A. Alekseyev*

George Washington University, Washington, DC, USA. *Email: maxal@gwu.edu

Introduction. Genome sequencing technology has evolved over time, increasing availability of sequenced genomic data. Modern sequencers are able to identify only short subsequences (*reads*) in the supplied genomic material, which then become an input to genome assembly algorithms aimed at reconstruction of the complete genome. Such reconstruction is possible (but not guaranteed) only if each genomic region is covered by sufficiently many reads. Lack of comprehensive coverage (particularly severe in single-cell sequencing) and presence of long similar subsequences (*repeats*) in genomes pose major obstacles for existing assembly algorithms. They therefore often are able to reliably reconstruct only long subsequences of the genome (inter-spaced with low-coverage regions and repeats), called *scaffolds*.

The challenge of reconstructing a complete genomic sequence from scaffolds is known as the *scaffolds assembly* problem. It is often addressed technologically by generating so-called long-jump libraries or by using a related complete genome as a reference. Unfortunately, the technological solution may be expensive and inaccurate, while the reference-based approach is obfuscated with structural variations across the genomes.

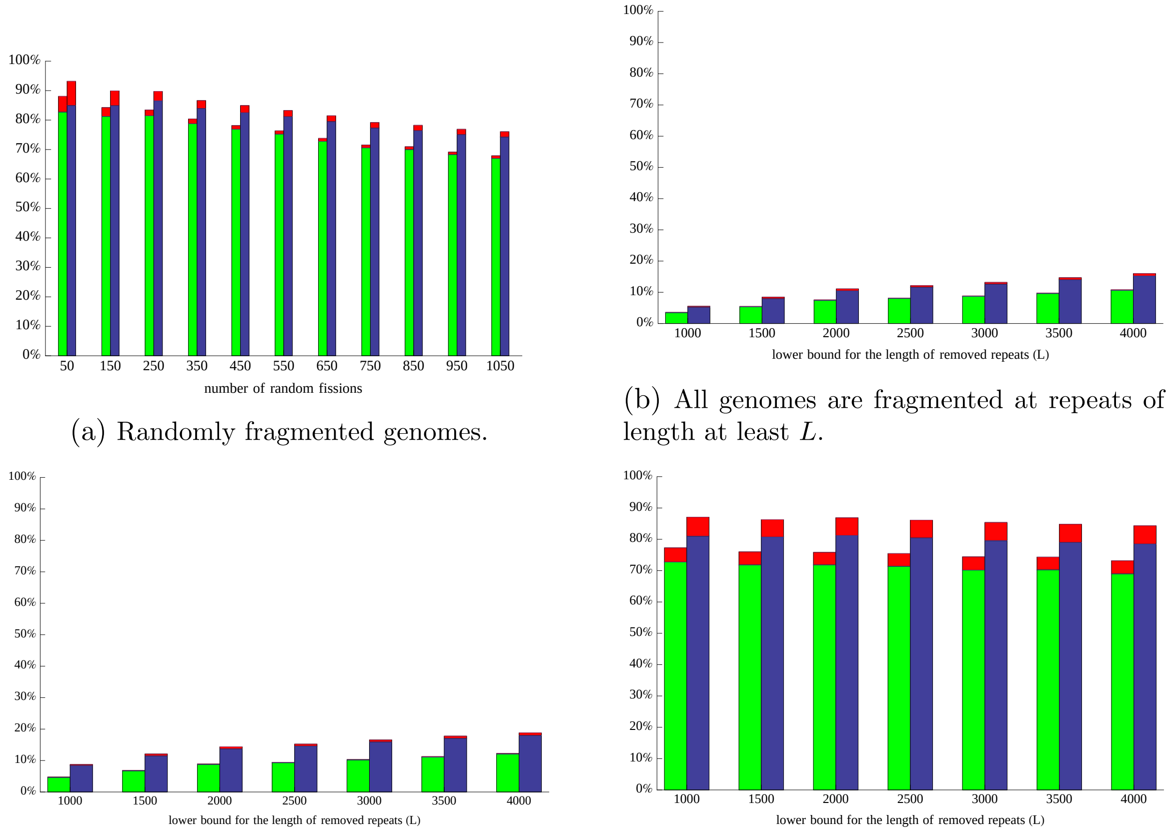
Methods. In the current work, we assume that the constructed scaffolds are accurate and long enough to allow identification of orthologous genes. The scaffolds then can be represented as ordered sequences of genes and we pose the scaffolds assembly problem as the reconstruction of the global gene order (along genome chromosomes) from the gene sub-orders defined by the scaffolds. We view such gene sub-orders as the result of both evolutionary events and technological fragmentation in the genome. Evolutionary events that change gene orders are *genome rearrangements*, most common of which are *reversals*, *fusions*, *fissions*, and *translocations*. Technological fragmentation can be modeled by artificial "fissions" that break genomic chromosomes into scaffolds. Scaffold assembly can therefore be reduced to the search for "fusions" that revert technological "fissions" and glue scaffolds back into chromosomes. This observation inspires us to employ the genome rearrangement analysis techniques for scaffolding purposes.

Rearrangement analysis of multiple genomes relies on the concept of the *breakpoint graph*. While traditionally the breakpoint graph is constructed for complete genomes, it can also be constructed for fragmented genomes, where we treat scaffolds as "chromosomes". We demonstrate that the breakpoint graph of multiple genomes possesses an important property that its connected components are robust with respect to genome fragmentation. In other words, connected components of the breakpoint graph mostly retain information about the complete genomes, even when the breakpoint graph is constructed on their scaffolds. Our method utilizes connected components of the breakpoint graph for the scaffold assembly of fragmented genomes.

We remark that our method can be integrated with the MGRA framework [2], which performs rearrangement analysis of multiple genomes, identifies reliable genome rearrangements and transforms their breakpoint graph into an *identity* breakpoint graph (of a single ancestral genome). In the process of this transformation MGRA can only break the connected components of the breakpoint graph into smaller ones, which often helps to process them more accurately with our scaffold assembly method.

Evaluation. We evaluated our method on artificially fragmented mammalian genomes as well as on incomplete highly fragmented anophelinae genomes. Namely, we used Ensembl to obtain a set of six complete mammalian genomes: *Homo sapiens* (GRCh38), *Mus musculus* (GRCm38.p2), *Rattus norvegicus* (Rnor_5.0), *Canis familiaris* (CanFam3.1), *Macaca mulatta* (MMUL_1.0), and *Pan troglodytes* (CHIMP2.1.4). From their orthologous gene mappings, we constructed gene families and filtered some of them so that each genome was represented as sequences of the same 11816 genes. The genomes were fragmented with *random fragmentation* and *repeat-based fragmentation.* The random fragmentation allows us to overcome the lack of information about genome fragmentation mechanism. However, we may have better insight in the fragmentation model, if we assume that genome scaffolds were obtained from a conventional genome assembler having difficulties in reconstruction of the order of long DNA repeats. In this case, it becomes realistic to fragment the genomes based on locations of such repeats.

Random fragmentation. To create instances of randomly fragmented genomes, we applied k random artificial "fissions" to each of the genomes. For each value of k, we created 10 different sets of fragmented genomes, executed our method on each of the sets (both with and without MGRA integration). Accuracy of our assembly results (Fig. 1a) allows us to conclude that while our method is not able to reconstruct complete genomes where fragmentation is high, it is able to reconstruct genomes almost completely, when the fragmentation is low. While the true positive rate of the assembly results decreases as fragmentation raises, the results still remain highly reliable with low false positive rate. This is not an unexpected property of our method since it is based on the robustness of the connected components of the breakpoint graph. Integration with the MGRA framework



(a) Randomly fragmented genomes.

(b) All genomes are fragmented at repeats of length at least L.

(c) All but one genome (dog) are fragmented at repeats of length at least L.

(d) Only one genome (dog) is fragmented at repeats of length at least L.

Figure 1: Accuracy of our scaffold assembly method on artificially fragmented mammalian genomes with (blue bars) and without (green bars) integration with MGRA. The blue and green bars give normalized true positive, while red bar gives normalized false positive rate for assembly results. **(a)** All genomes are randomly fragmented by applying a specified number of random "fissions" (with step 100). **(b–d)** Genomes are broken at the positions of repeats of length at least L (with step 500 bp).

further yields additional number of highly reliable fragment assemblies.

Repeat-based fragmentation. To create instances of repeat-based fragmented genomes, we removed all repeats longer that a fixed number of basepairs (from $1K$ to $4K$ bp with the step of $0.5K$) and partitioned the genomes into fragments with no long repeats. We used the same set of six mammalian genomes, for which we obtained the repeats locations from the RepeatMasker database. We performed the three experiments with (i) all, (ii) all but one, and (iii) only one of the genomes being fragmented (Fig. 1b-d). Experiments (i) and (ii) demonstrate that while in the presence of a reference genome our method yields more true fragment adjacencies, it still performs relatively well in the case, when no reference is known. Experiment (iii) shows that our method can be used as a highly reliable step for assembly of a single fragmented genome, when several complete reference

Genome	# assemblies	
	without MGRA	**integrated with MGRA**
An. gambiae	0	0
An. arabiensis	6	10
An. quadriannulatus	75	91
An. merus	466	550
An. dirus	30	45
A. albimanus	6	10

Table 1: Statistics on the number of reported scaffold assemblies, both with and without integration with MGRA.

genomes are known. Since DNA repeats are subject to genome rearrangements in the course of evolution, integration with MGRA yields additional true adjacencies.

Anophelinae genomes. The second evaluation of our scaffold assembly method was performed on highly fragmented genomes from anophelinae subfamily, followed by comparison of the results to a reference-based assembly approach. Namely, we considered six anophelinae genomes: *An. gambiae*, *An. arabiensis*, *An. quadriannulatus*, *An. merus*, *An. dirus*, and *An. albimanus*, represented as sequences of the same $6,837$ genes. The results of our scaffold assembly method on these genomes are reported in Table 1.

We compared our assembly results to those from another anophelinae study (*comparison study*) led by Dr. Igor Sharakhov at Virginia Tech University. The comparison study performed analysis of *An. gambiae*, *An. arabiensis* genomes from the same source, where *An. gambiae* represents a complete genome, while *An. arabiensis* exposes rather high fragmentation. The genome data preparation was similar to ours. The relationships between these genes and their order on scaffold were compared to the cytogenetic and physical maps identifying breakpoints of fixed reversals. Among 10 assemblies in *An. arabiensis* genome identified by our method, the comparison study was able to identify and confirm only 6.

Acknowledgements. The work was done in collaboration with the Anopheles Genomes Cluster Consortium [3] and supported by the National Science Foundation under Grant No. IIS-1462107. The extended abstract of our work is published in [1].

References

[1] Aganezov, S., N. Sydtnikova, AGC Consortium, and M. A. Alekseyev (2015). Scaffold assembly based on genome rearrangement analysis. *Computational Biology and Chemistry 57*, 46–53.

[2] Alekseyev, M. A. and P. A. Pevzner (2009). Breakpoint Graphs and Ancestral Genome Reconstructions. *Genome Research 19*(5), 943–957.

[3] Neafsey, D. E., R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, et al. (2015). Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science 347*(6217), 1258522.