Mechanisms for Generating Mosaic Structures of Segmental Duplications

July 5, 2013

1 Introduction

Genome-scale evolutionary events include genome rearrangements, duplications and deletions. Genome rearrangements, which shuffle gene orders and change the gene orientations, includes inversions, transpositions, translocations, fusions, and fissions.

In this paper, we focus on special type of duplications called segmental duplications. Segmental duplications are segments of DNA (typically range at least 1kb in length), occur at more than one site within the genome, and typically share a high level of sequence identity ($> 90\%$) [1]. Segmental duplications have long been recognized as driving forces of evolution [2]. For example, segmental duplications in human genomes are hotspots for nonallelic homologous recombination, copy-number variations, genes and transcripts innovations [3]. Segmental duplications in human genomes generally form mosaic structures combining or juxtaposing original segments that range from 1kb to 200 kb from disparate regions of the genome [4]. Subsequent duplications duplicate portions of mosaic structures to other regions of the genome—create additional copies of the initial segments, in which the order of the duplicated segments is generally preserved [5]. Pevzner *et al.* proposed to use A-Bruijn graphs to represent repeats and sub-repeats of mosaic structures in a genome [6]. Jiang *et al.* applied A-Bruijn graphs to study ancestral reconstruction of segmental duplications that emerged within the last 40 million years of human genome evolution, based on a collection of 28,856 pairwise alignments (with length $>$1kbp and sequence identity $>$90%) [3].

The formation of mosaic structures of segmental duplications seems to involve complex rearrangements and duplications, but the mechanism behind it is still unclear. Recent studies on genomic disorders from clinical chromosomal microarray analysis revealed similar complex patterns of rearrangements and duplications [7, 8]. The analyses of sequences at breakpoint junctions by array comparative genomic hybridization (aCHG) [9], led to the proposal of a replication based mechanism, the Fork Stalling and Template Switching (FoSTeS) model [7] (inspired by a similar mechanism proposed for amplification in E. coli [10]) and the Microhomology Mediate Break-Induced Replication (MMBIR) model [11]. According to this mechanism, during DNA replication, long-distance template-switching occurs between replication forks in physical proximity (not necessary adjacent in primary sequences) through the microhomology. Depending on whether the position and orientation of the new fork was invaded and copied, and the direction of fork progression, template-switching results in a duplication, deletion, inversion or translation [11]. This procedure of long-distance template-switching could occur multiple times in series, due to the poor processivity of the involved DNA polymerase, causing the complex rearrangements and duplications [12].

We hypothesize that the formation of mosaic structures of segmental duplications involves a replication based mechanism for the following reasons.

- The replication based mechanism can account for complex rearrangements and duplications from the observed clinical genomic data, including juxtaposing of distantly distributed segments, a phenomenon resembling mosaic structures of segmental duplications.

- The original segments are usually 1kb to 200 kb away from the duplicated segments in the mosaic structures [4], and such distances are below the long distances in the observed template-switchings (e.g., 120kb to 550kb [12]).

We will use de Bruijn graphs built from genome sequences to check whether the formation of mosaic structures of segmental duplications involves a replication based mechanism.

1. For human genomes, we first need to deal with the high-multiplicity mobile elements (e.g. SINE and LINE repeats). The most common SINE repeats in primates are called Alu elements, each approximately 260 base pairs long. The human genome contains estimated over one million Alu elements—about 11% of the genome [13], and about half a million LINE repeats—about 17% of the genome [14]. We could mask all known the Alu elements and LINE repeats, by replacing them with random sequences of same lengths, but still record their position information for later analysis, since these high-multiplicity mobile elements may trigger complex

long-distance template-switchings and serve as footprints at the breakpoint junctions [15]. For E. coli genomes, we may not need this preprocess.

2. We build a de Bruijn graph from the input genome, by transforming that genome into perfect sequencing reads of (k+1)-mer (overlapping and uniform coverage). Note that this de Bruijn graph can only "glue" all perfect repeats without any gaps and mismatches. We could handle imperfect repeats by the techniques in SPAdes [16] (e.g., error corrections in reads and simplification of A-Bruijn graphs). Note that we may first focus on near-perfect repeats.

3. We find out all the candidates for the mosaic structures of segmental duplications in the input genomes. A candidate is a continuous region that duplicates and juxtaposes distantly distributed segments on the same arm of the chromosome. Figure 1 illustrates an example of a candidate for a mosaic structure of segmental duplications in the de Bruijn graph. Note that these duplicated segments are linked by edges of approximate length k (in terms of the number of k-mers, see brown edges in Figure 1).

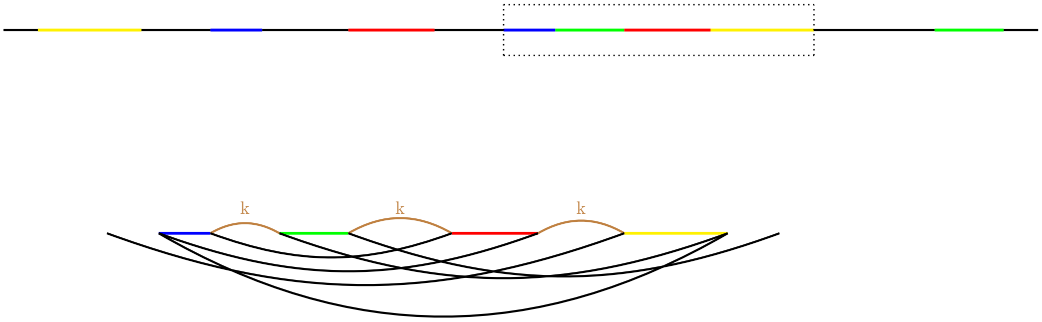


Figure 1: An example for a mosaic structure of segmental duplications and the corresponding part of the de Bruijn graph

4. For each candidate mosaic structure, we retrieve the sequence information from the de Bruijn graph, and check whether there are microhomology [11] or novel sequence insertions from neighboring regions at the breakpoint junctions [8]. Note that microhomology of m bps at the breakpoint junctions may result in the $(k-m)$-length edges linking adjacent segments in a mosaic structure.

2 Followup: the Breakpoint Graph and the Colored de Bruijn Graph

Assume that a genome is a set of strings. Given genomes *A* and *B*, we classify (*k*+1)-mers in the genomes into 3 classes: *A*(occur only in *A*), *B*(occur only in *B*) and AB(occur in both *A* and *B*).

2.1 the Breakpoint Graph

The breakpoint graph is a data structure first introduced by Bafna and Pevzner to study the inversion distance [17], which has formed the basis for much algorithmic research on rearrangements over the last twenty years. Alekseyev and Pevzner extened the pairwise breakpoint graphs [17] to the multiple breakpoint graphs to overcome some limitations in the study of ancestral genome reconstructions [18]. However, it is difficult to incorporate segmental duplications into breakpoint graphs, since breakpoint graphs are usually defined on genomes with equal gene content and no duplicates.

Given a set of strings *S*, the breakpoint graph $BP(S)$ is defined as follows.

- Each *k*-mer in the strings is a directed edge with two vertices, *u* as its start and *v* as its end, and each *k*+1-mer in the strings is an undirected edge which connects the end of its prefix *k*-mer and the start of its suffix *k*-mer.
- Identically labeled edges (*k*-mers) are glued together.

The colored breakpoint graph is simply coloring the edges of the breakpoint graph $BP(A\bigcup B)$ into 3 colors: *A*(red), *B*(Blue) and *AB* (purple). There are two kinds of non-branching and unicolored paths in the colored breakpoint graphs. Type *I* paths are from an end of a *k*-mer to a start of another *k*-mer, and type *II* paths are from a start of a *k*-mer to an end of another *k*-mer. We can condense the colored breakpoint graphs by replacing type *I* paths with undirected colored edges, and type *II* paths with directed colored edges (e.g., see Figure 2 (d) and (e)). Note that the traditional pairwise breakpoint graphs [17] correspond to the condensed (colored) breakpoint graphs (e.g., see Figure 2 (e)).

2.2 the Colored de Bruijn Graph

The de Bruijn graph is a data structure first brought to bioinformatics by Pevzner as a method to assemble k-mers generated by sequencing by hybridization [19], and is later used in sequence assembly [20, 21].

Given a set of strings S, the de Bruijn graph $DB(S)$ is defined as follows.

- Each k-mer in the strings is a vertex, and each k+1-mer is a directed edge from its prefix k-mer to its suffix k-mer.

- Identically labeled vertices (k-mers) are glued together.

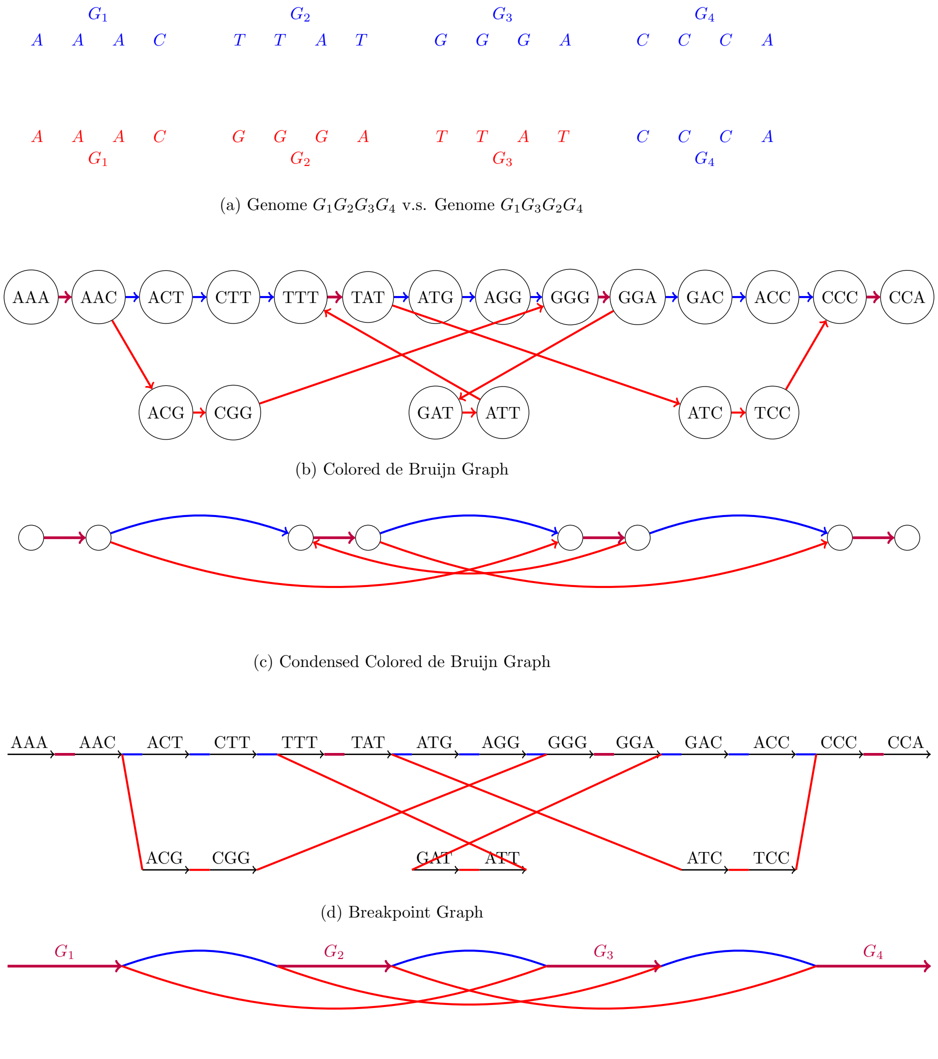
A non-branching paths in de Bruijn graph can be replaced by a directed edge of which the direction is from its directed edges (k+1-mers). The condensed de Bruijn graph replaces all non-branching paths by single edges.

The colored de Bruijn graph $G(A,B)$ is simply coloring the edges of the de Bruijn graph $DB(A \bigcup B)$ into 3 colors: A(red), B(blue) and AB (purple). We can condense the colored de Bruijn graph by replacing all non-branching and unicolored paths by directed colored edges (e.g., see Figure 2 (b) and (c)).

The bi-directed de Bruijn graphs [22] can be viewed as gluing reverse complement k-mers (vertices) in the de Bruijn graph. This notation can also easily extended to the colored de Bruijn graphs and their condensed forms (e.g., see Figure 4 (b) and (c)).

2.3 Colored de Bruijn Graph v.s. Breakpoint Graph

Figure 2 illustrates an example of a transposition. Note only one strand is shown in the figure. A cycle in an edge-colored graph is called alternating if the colors of every two consecutive edges of this cycle are distinct. Figure 2 (c) and (e) contain alternating cycles.


G_1 G_2 G_3 G_4
A A A C T T A T G G G A C C C A
A A A C G G G A T T A T C C C A
G_1 G_2 G_3 G_4
AAA AAC ACT CTT TTT TAT ATG AGG GGG GGA GAC ACC CCC CCA
ACG CGG GAT ATT ATC TCC
AAA AAC ACT CTT TTT TAT ATG AGG GGG GGA GAC ACC CCC CCA
ACG CGG GAT ATT ATC TCC
G_1 G_2 G_3 G_4

(a) Genome $G_1G_2G_3G_4$ v.s. Genome $G_1G_3G_2G_4$

(b) Colored de Bruijn Graph

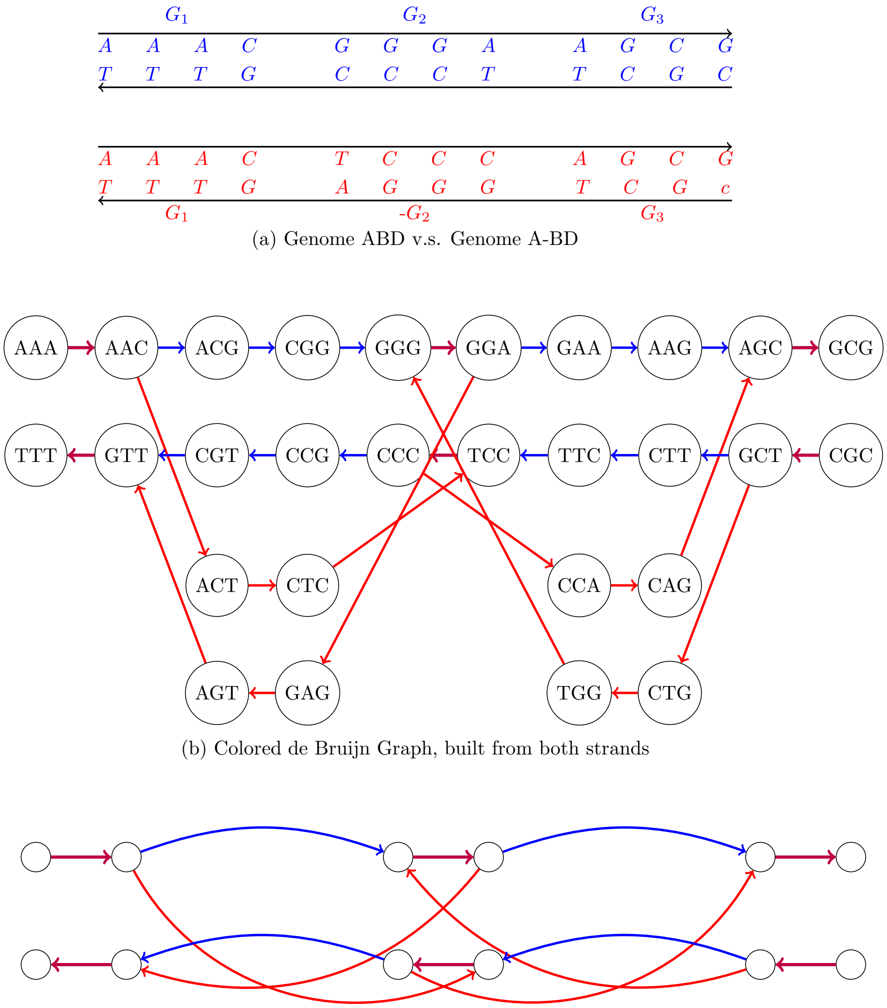
(c) Condensed Colored de Bruijn Graph

(d) Breakpoint Graph

(e) Condensed Breakpoint Graph

Figure 2: A toy example of a transposition.

Figure 3 illustrates an example of an inversion on the de Bruijn graph of both strands. Note the both strands are used to the formulate of alternating cycles.



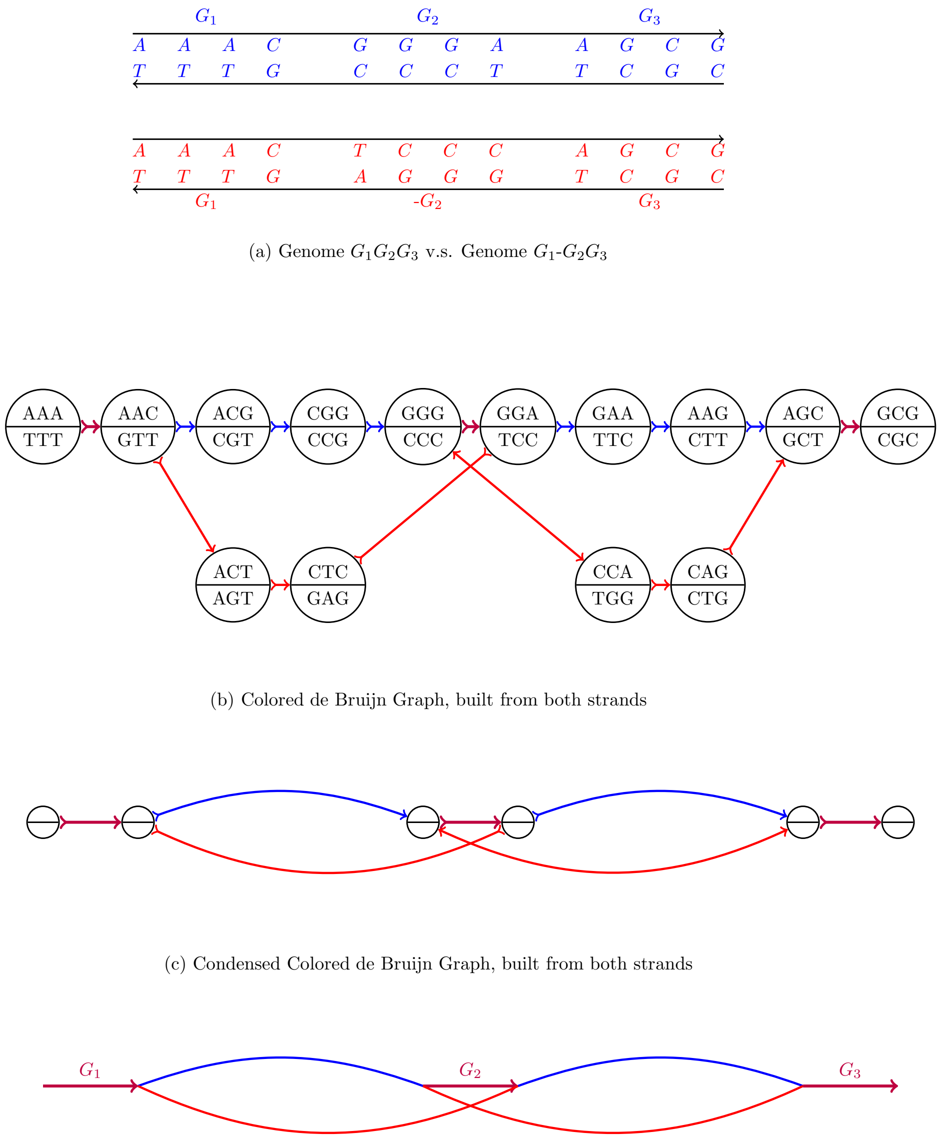
(a) Genome ABD v.s. Genome A-BD

(b) Colored de Bruijn Graph, built from both strands

(c) Condensed Colored de Bruijn Graph, built from both strands

Figure 3: A toy example of an inversion on the de Bruijn graph of both strands.

Figure 4 illustrates the same example of an inversion by gluing each k-mer with its reverse complement as in the bi-directed de Bruijn graph [22].



(a) Genome $G_1G_2G_3$ v.s. Genome G_1-G_2G_3

(b) Colored de Bruijn Graph, built from both strands

(c) Condensed Colored de Bruijn Graph, built from both strands

(d) Condensed Breakpoint Graph

Figure 4: A toy example of an inversion on the bidirected de Bruijn graph.

The above toy examples show that the simplified colored de Bruijn graph is very similar to the corresponding breakpoint graph. But the following properties are worth noticing.

- Rearrangements create edges of center size in the condensed colored de Bruijn graph. For example, each read or blue edge caused by rearrangements in alternating cycles should have length around k (in terms of the number of (k+1)-mer edges).

- Assume there is no large duplications and losses, for each alternating cycle caused by rearrangements in the condensed colored de Bruijn graph, the sum of lengths of blue edges and that of red edges should have similar values.

2.4 Why Colored de Bruijn Graphs?

- Extended model for both genome rearrangements, duplications and deletions.

- Direct use of the sequence information

- Convenient incorporation of reference genomes for comparative study (multicolored de Bruijn graph)

- Accurate prediction of boundaries of the evolutionary events (thanks to the high resolution)

- Detection of interesting genomic patterns/features (e.g. the mosaic structure in segmental duplications)

References

[1] Evan E Eichler. Recent duplication, domain accretion and the dynamic mutation of the human genome. *TRENDS in Genetics*, 17(11):661–669, 2001.

[2] J.A. Bailey and E.E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics,* 7(7):552–564, 2006.

[3] Zhaoshi Jiang, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A Pevzner, and Evan E Eichler. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature genetics*, 39(11):1361–1368, 2007.

[4] Rhea Vallente Samonte and Evan E Eichler. Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, 3(1):65–72, 2002.

[5] Jeffrey A Bailey, Amy M Yavor, Luigi Viggiano, Doriana Misceo, Juliann E Horvath, Nicoletta Archidiacono, Stuart Schwartz, Mariano Rocchi, and Evan E Eichler. Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *The American Journal of Human Genetics*, 70(1):83–100, 2002.

[6] Paul A Pevzner, Haixu Tang, and Glenn Tesler. De novo repeat classification and fragment assembly. *Genome Research*, 14(9):1786–1796, 2004.

[7] Jennifer A Lee, Claudia Carvalho, and James R Lupski. A dna replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131(7):1235–1247, 2007.

[8] Pengfei Liu, Ayelet Erez, Sandesh C Sreenath Nagamani, Shweta U Dhar, Katarzyna E Kołodziejska, Avinash V Dharmadhikari, M Lance Cooper, Joanna Wiszniewska, Feng Zhang, Marjorie A Withers, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, 146(6):889–903, 2011.

[9] Daniel Pinkel and Donna G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37:S11–S17, 2005.

[10] Andrew Slack, PC Thornton, Daniel B Magner, Susan M Rosenberg, and PJ Hastings. On the mechanism of gene amplification induced under stress in escherichia coli. *PLoS genetics*, 2(4):e48, 2006.

[11] PJ Hastings, Grzegorz Ira, and James R Lupski. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genetics*, 5(1):e1000327, 2009.

[12] Wenli Gu, Feng Zhang, and James R Lupski. Mechanisms for human genomic rearrangements. *Pathogenetics*, 1(1):4, 2008.

[13] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[14] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

[15] Claudia MB Carvalho, Melissa B Ramocki, Davut Pehlivan, Luis M Franco, Claudia Gonzaga-Jauregui, Ping Fang, Alanna McCall, Eniko Karman Pivnick, Stacy Hines-Dowell, Laurie H Seaver, et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature genetics*, 43(11):1074–1081, 2011.

[16] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 19(5):455–477, 2012.

[17] V. Bafna and P.A. Pevzner. Genome rearrangements and sorting by reversals. In *Proc. 34th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'93)*, pages 148–157, 1993.

[18] Max A Alekseyev and Pavel A Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, 19(5):943–957, 2009.

[19] P.A. Pevzner. l-tuple dna sequencing: computer analysis. *J. Biomol. Struct. Dyn.*, 7:63–73, 1989.

[20] R.M. Idury and M.S. Waterman. A new algorithm for dna sequence assembly. *J. Comput. Biol.*, 2(2):291–306, 1995.

[21] P.A. Pevzner, H. Tang, and M.S. Waterman. An eulerian path approach to dna fragment assembly. *Proc. Nat'l Acad. Sci., USA*, 98(17):9748, 2001.

[22] Paul Medvedev, Konstantinos Georgiou, Gene Myers, and Michael Brudno. Computability of models for sequence assembly. In *Proc. 7th Workshop*

Algs. in Bioinf. (WABI'07), volume 4645 of *Lecture Notes in Comp. Sci.*, pages 289–301. Springer, 2007.