Sequencing the Breakpoint Regions of Genome Rearrangements

July 11, 2013

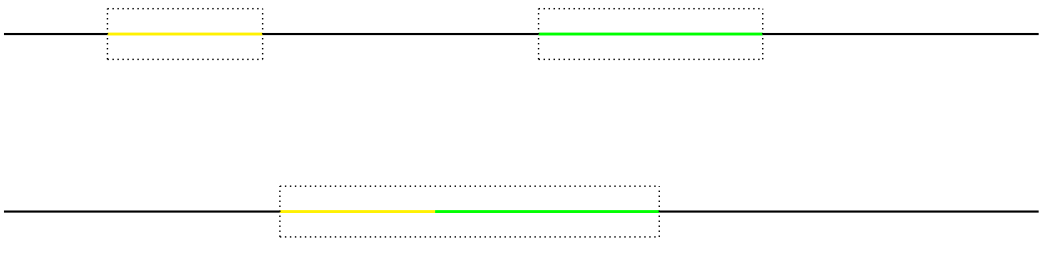
1 Introduction

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.

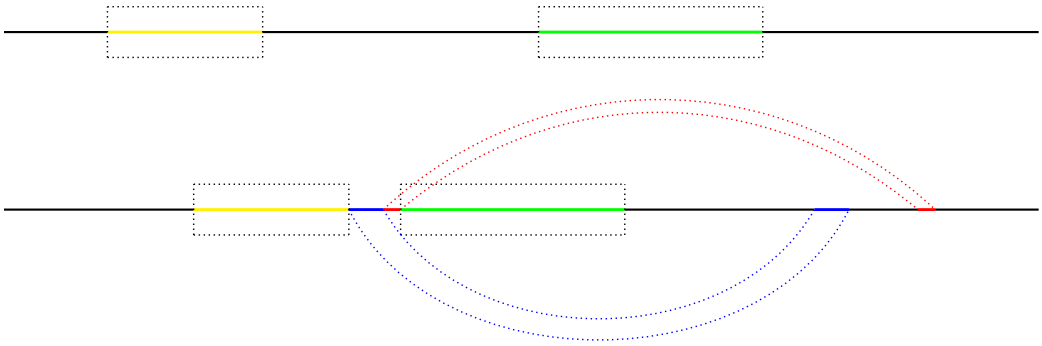
In this paper, we focus on the breakpoint regions of genome rearrangements. Genome rearrangements represent large-scale evolutionary events that are clearly distinct from the small-scale changes as mutations and indels. Genome rearrangements in primates evolution have been studied at different levels of resolution, such as comparative chromosome banding [1], gene mapping [2], cross-species chromosomal painting [3, 4], comparative genome hybridization painting [5], fluorescent in situ hybridization (FISH) [6] and bacterial artificial chromosome (BAC) [7]. Current approaches (except BAC [7]) can not offer base-pair resolution to study the exact sequences at the breakpoint regions of genome rearrangements, while BAC [7] has only been applied to sequence 24 breakpoint regions between human and gibbon genomes.

The reliable and accurate assembly of large genomes from next-gen sequencing techniques provides an opportunity to study breakpoint regions systematically between different genomes. Detailed sequencing analysis of human-gibbon rearrangements revealed precise breakage positions for 60% breakpoints (see Figure 1(a)), and new insertion sequences at 40% breakpoint regions, most of which consisted of common repeat families (e.g. Alu and LINE) as well as duplicate sequences that originate from neighborhood of the breakpoints [7] (see Figure 1(b)). Studies on genomic disorders from clinical chromosomal microarray analysis

[8, 9] also revealed similar complex breakpoint regions of rearrangements, with the analysis of sequences at breakpoint junctions by array comparative genomic hybridization (aCHG) from clinical chromosomal microarray analysis [10]. These observations support a replication based mechanism, the Fork Stalling and Template Switching (FoSTeS) model [8] (inspired by a similar mechanism proposed for amplification in E. coli [11]) and the Microhomology Mediate Break-Induced Replication (MMBIR) model [12]. 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 [12]. 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 breakpoint regions of rearrangements [13].



(a) No insertion at the breakpoint region



(b) Insertion of duplicated segments from two blocks located near the breakpoint

Figure 1: Examples of no insertion and complex insertions at breakpoint regions

We hypothesize that the formation of complex breakpoint regions of rearrangements involves a replication based mechanism for the following reasons.

- The replication based mechanism can account for complex breakpoint regions from both the human-gibbon rearrangements [7] and the observed clinical rearrangement data of genomic disorders [8, 9]. This mechanism may expand breakpoint regions by juxtaposing and inserting distantly distributed segments.

- The original segments are usually less than 100kb away from the duplicated segments in the breakpoint region [7], and such distances are below the long distances in the observed template-switchings (e.g., 120kb to 550kb [13]).

- Microhomology (2-6 bp) was observed in 50% of the human-gibbon breakpoint regions [7], as well as many breakpoint regions from clinical data of genomic disorders [8, 9].

We will use colored de Bruijn graphs to study the breakpoint regions of rearrangements between input genomes.

1. For primate 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 [14], and about half a million LINE repeats—about 17% of the genome [15]. 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 contribute into the complex breakpoint regions [7] and serve as footprints at the breakpoint junctions [16].

2. We build a colored de Bruijn graph from the input genomes, by transforming that genomes into perfect sequencing reads of (k+1)-mer (overlapping and uniform coverage) and coloring the reads of different genomes respectively. Note that this colored de Bruijn graph can only "glue" all perfect repeats of the same color without any gaps and mismatches. We could handle imperfect repeats by the techniques in SPAdes [17] (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 breakpoints of rearrangements and focus on candidates for complex breakpoint regions. A candidate is a continuous insertion sequence at the breakpoint region that mainly consists of common repeat families as well as distributed segments duplicated from neighborhood of the breakpoint(see Figure 1(b)). Note that these duplicated segments are usually smaller than the rearranged segments, and different parameters k should be applied to study such regions at different resolution in the colored de Bruijn graph.

4. For each candidate breakpoint region, we retrieve the sequence information from the colored de Bruijn graph, and check whether there are microhomology [7, 12, 9] at the breakpoint junctions.

The study of complex breakpoint regions between input genomes may also indicate the most "plausible" rearrangement scenario. Take two genomes of 4 synteny blocks for example, say $G_1 = 1\ 2\ 3\ 4$ and $G_2 = 1\ {-3}\ 2\ 4$. The breakpoint graph [18] of these two genomes consists of an alternating cycle of length 6, and there are two possible inversion scenarios shown in Figure 2. The detailed analysis at breakpoint regions may help to find out the most "plausible" rearrangement scenario. For example, in Figure 2, if we find insertion segments from the start region of the blue block or the end region of the red block at the first breakpoint region (between the yellow and blue syteny blocks), we can infer that the left scenario is more "plausible" than the right one, since it is less likely to include such insertion segments at the first breakpoint from the right scenario. Similarly, if we find the insertion segment from the start region of the green block at the second breakpoint region (between the blue and red syteny blocks), the right scenario is more likely to occur.

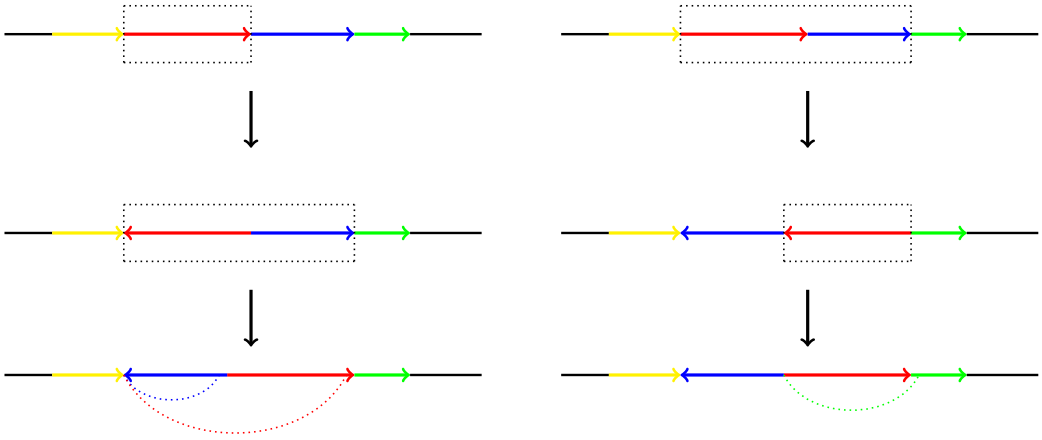


Figure 2: Two possible inversion scenarios. Colored dotted curves represent several possible insertions at the breakpoint region, which can be used to infer the more “plausible” scenario of the two.

References

[1] Jorge J Yunis and Ora Prakash. The origin of man: a chromosomal pictorial legacy. *Science*, 215(4539):1525–1530, 1982.

[2] Catherine Turleau, Nicole Créau-Goldberg, Chantal Cochet, and J De Grouchy. Gene mapping of the gibbon. its position in primate evolution. *Human Genetics*, 64(1):65–72, 1983.

[3] Anna Jauch, Johannes Wienberg, Roscoe Stanyon, N Arnold, S Tofanelli, T Ishida, and Thomas Cremer. Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proc. Nat’l Acad. Sci., USA*, 89(18):8611–8615, 1992.

[4] William J Murphy, Denis M Larkin, Annelie Everts-van der Wind, Guillaume Bourque, Glenn Tesler, Loretta Auvil, Jonathan E Beever, Bhanu P Chowdhary, Francis Galibert, Lisa Gatzke, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617, 2005.

[5] Lucia Carbone, Gery M Vessere, Boudewijn FH ten Hallers, Baoli Zhu, Kazutoyo Osoegawa, Alan Mootnick, Andrea Kofler, Johannes Wienberg,

Jane Rogers, Sean Humphray, et al. A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS genetics*, 2(12):e223, 2006.

[6] J Wienberg. Fluorescence in situ hybridization to chromosomes as a tool to understand human and primate genome evolution. *Cytogenetic and genome research*, 108(1-3):139–160, 2005.

[7] Santhosh Girirajan, Lin Chen, Tina Graves, Tomas Marques-Bonet, Mario Ventura, Catrina Fronick, Lucinda Fulton, Mariano Rocchi, Robert S Fulton, Richard K Wilson, et al. Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome research*, 19(2):178–190, 2009.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.