Poster A-7

Breakpoint Re-use and Generalized Transpositions Depend on Synteny Block Size in Mammalian Genome Rearrangements

Authors:
Oliver Attie (*Department of Infectious Diseases, Mount Sinai School of Medicine, NY, NY 10029*)
Aaron E. Darling (*Department of Computer Science, University of Wisconsin-Madison, Madison, WI 53706*)
John Sikorski3 (*Regeneron Pharmaceuticals, Tarrytown, NY10591*)
Sophia Yancopoulos (*Chiorazzi Lab, Feinstein Institute for Medical Research, Manhasset, NY 11030*)

Short Abstract: We analyzed human (NCBI 35), mouse (NCBI 33), and rat (RGSC 3.4) genome assembly versions using Mauve to determine syntenic regions, GRIMM and GRITT (Genome Rearrangement by Inversion, Translocation and generalized Transposition) for genome rearrangement. Transpositions occur in GRITT at high resolution and breakpoint re-use varies with scale.

Long Abstract:

Breakpoint Re-use and Generalized Transpositions Depend on Synteny Block Size in Mammalian Genome Rearrangements

Oliver Attie1, Aaron E. Darling2, John Sikorski3, Sophia Yancopoulos4

1Department of Infectious Diseases, Mount Sinai School of Medicine, NY, NY 10029.
2Department of Computer Science, University of Wisconsin-Madison, Madison, WI 53706.
3Regeneron Pharmaceuticals, Tarrytown, NY10591.
4Chiorazzi Lab, Feinstein Institute for Medical Research, Manhasset, NY 11030.

Motivation

Relatively high values of the breakpoint re-use statistic in large scale chromosomal comparisons of entire genomes, sparked lively debate as to whether evolution repeats cataclysmic genomic events[1, 2]. Such analyses depend on which genomic operations are admissible. For rearrangements based on generalized inversions (including translocations, fissions and fusions) the maximal number of breakpoints incurred (or healed) after performing the operation is two. Since inversions create at most two breakpoints at each genome step, dividing genomic distance (or total number of genomic steps taken) by half the number of breakpoints for a generalized inversion distance measure arrives at a breakpoint re-use of "1". For transpositions, three breakpoints are incurred. Since it takes a minimum of 3 inversions to undo a transposition, the breakpoint re-use statistic for transpositions is 2. However, for a rearrangement model [3] incorporating transpositions (with genomic step-size 2) the breakpoint re-use for rearrangements consisting of transpositions would be only 4/3. Similarly, breakpoint re-use could depend on resolution. For example if larger scale

transformations are inversions, but at a higher resolution transpositions become important, the scale at which the analysis is carried out would affect the value of breakpoint re-use, especially if small scale transpositions become rampant at high resolution.

Approach

To examine the effect of resolution and genome conversion scenario on the current controversy in models of large-scale chromosomal evolution, we examined the dependence of genomic distance and breakpoint re-use on resolution, for human (NCBI 35), mouse (NCBI 33), and rat (RGSC 3.4) genome assembly versions.

From sequence to synteny: Mauve

Although previous mammalian genomes alignments have been constructed [4], none have used a method sensitive to small, trans-chromosome, rearrangements of genomic sequence. Mauve [5] implements a novel method to identify genomic micro-rearrangements by anchoring alignments in conserved regions of non-repetitive sequence.

Data Crunch

Three-way synteny blocks for human, mouse, and rat were constructed using Mauve. The sequences were analyzed* to determine syntenic regions, called "Locally Collinear Blocks" (LCBs), based on sequence similarity. Each LCB has a "weight" equal to the sum of the lengths of its ungapped aligned components. Resolution is controlled by the minimum LCB weight accepted. At a given resolution, average LCB chromosomal length is typically 100-1000 times the minimum weight.

Construction of the initial homology map consumed 12 hours on a Linux PC with two disks. Subsequent whole-genome alignment based on the initial homology map consumed another 12 hours on a 96-CPU Orion Multisystems deskside workstation. The full alignments used for our analyses used a minimum LCB weight of 56 which resulted in 6351 LCBs in the initial homology map.

To evaluate the effect of micro-rearrangements of various sizes, we constructed alignments with increasing minimum LCB weights between 56 and 100,000. Several thousand LCBs in the initial homology map are less than 10Kbp in length-truly micro-rearrangements!

Analyzing Human-Mouse-Rat Rearrangements: Onset of Transpositions

Genomic distances were computed for all pairwise genome comparisons using both GRIMM [6] and GRITT [3] (Genome Rearrangement by Inversion, Translocation and generalized Transposition based on DCJs). GRIMM and GRITT agree at low resolution but differ at high resolution, necessitating the use of generalized transposition (or block interchange) [BI] [3] in rearrangement scenarios by GRITT. (Lacking this option causes the GRIMM distance to exceed GRITT approximately by the number of "hurdles". [7]) The onset of BI occurs at minimum LCB weights of 2153, 6284 and 6284 respectively for h-m, m-r and h-r, roughly the scale of the 300kb blocks reported by Bourque et al [8].

Dependence of Breakpoint Re-use on Resolution

Breakpoint re-use (2d/b, d=genomic distance, b=#breakpoints) [1] was computed for both distance measures. Although the maximum value of breakpoint re-use (using the GRIMM distance measure) is consistent with values reported elsewhere (1.65 for m-h [1]) values less than these obtain either by using a different rearrangement scenario, or by going to high or low resolution.

Wrap-up

Our analysis finds new phenomena arising at high resolution with the onset of generalized transpositions via GRITT. Breakpoint re-use is dynamic, depending on resolving scale as well as the rearrangement scenario. An elevated value of the statistic at one resolution may decline at others and may not be a decisive indication of genome "fragility". Although we elucidated the dependence of breakpoint re-use on scale, it requires further analysis to decide between fragile and random breakage models of chromosome evolution.

* data can be downloaded from http://gel.ahabs.wisc.edu/~koadman/orion_results/

References

[1] Peng Q, Pevzner PA, Tesler G, (2006) The Fragile Breakage versus Random Breakage Models of Chromosome Evolution. PLOS Computational Biology 2(2):e14.
[2] Sankoff D, (2006) The Signal in the Genome. PLOS Computational Biology 2(4): e35.
[3] Yancopoulos S, Attie O, Friedberg R. (2005) Efficient Sorting of Genomic Permutations by Translocation, Inversion & Block Interchange. Bioinformatics 21: 3340-3346.
[4] Automated whole-genome multiple alignment of rat, mouse, and human.* Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglou S, Dubchak I. Genome Res. 2004 Apr;14(4):685-92.
[5] Darling ACE, Mau B, Blattner FR, Perna NT.(2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research 14(7):1394-403.
[6] Tesler G. (2002) GRIMM: Genome Rearrangements Web Server. Bioinformatics 18: 492-493.
[7] Hannenhalli S and Pevzner PA (1995) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In Proceedings of the 27th Annual ACM Symposium on the Theory of Computing, pp. 178-189 (full version appeared in J. ACM, 46, 1-27, 1999).
[8] Bourque G, Pevzner P, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes.Â Genome Research 14 507-516.