Genome Biology

RESEARCH Open Access

Comparative genomics reveals birth and death of fragile regions in mammalian evolution

Max A Alekseyev[1*], Pavel A Pevzner[2*]

Abstract

Background: An important question in genome evolution is whether there exist fragile regions (rearrangement hotspots) where chromosomal rearrangements are happening over and over again. Although nearly all recent studies supported the existence of fragile regions in mammalian genomes, the most comprehensive phylogenomic study of mammals raised some doubts about their existence.

Results: Here we demonstrate that fragile regions are subject to a birth and death process, implying that fragility has a limited evolutionary lifespan.

Conclusions: This finding implies that fragile regions migrate to different locations in different mammals, explaining why there exist only a few chromosomal breakpoints shared between different lineages. The birth and death of fragile regions as a phenomenon reinforces the hypothesis that rearrangements are promoted by matching segmental duplications and suggests putative locations of the currently active fragile regions in the human genome.

Background

In 1970 Susumu Ohno [1] came up with the Random Breakage Model (RBM) of chromosome evolution, implying that there are no rearrangement hotspots in mammalian genomes. In 1984 Nadeau and Taylor [2] laid the statistical foundations of RBM and demonstrated that it was consistent with the human and mouse chromosomal architectures. In the next two decades, numerous studies with progressively increasing resolution made RBM the *de facto* theory of chromosome evolution.

RBM was refuted by Pevzner and Tesler [3] who suggested the Fragile Breakage Model (FBM) postulating that mammalian genomes are mosaics of fragile and solid regions. In contrast to RBM, FBM postulates that rearrangements are mainly happening in fragile regions forming only a small portion of the mammalian genomes. While the rebuttal of RBM caused a controversy [4-6], Peng *et al.* [7] and Alekseyev and Pevzner [8] revealed some flaws in the arguments against FBM. Furthermore, the rebuttal of RBM was followed by many studies supporting FBM [9-31].

Comparative analysis of the human chromosomes reveals many short adjacent regions corresponding to parts of several mouse chromosomes [32]. While such a surprising arrangement of synteny blocks points to potential rearrangement hotspots, it remains unclear whether these regions reflect genome rearrangements or duplications/assembly errors/alignment artifacts. Early studies of genomic architectures were unable to distinguish short synteny blocks from artifacts and thus were limited to constructing large synteny blocks. Ma *et al.* [33] addressed the challenge of constructing high-resolution synteny blocks via the analysis of multiple genomes. Remarkably, their analysis suggests that there is limited breakpoint reuse, an argument against FBM, that led to a split among researchers studying chromosome evolution and raised a challenge of reconciling these contradictory results. Ma *et al.* [33] wrote: 'a careful analysis [of the RBM vs FBM controversy] is beyond the scope of this study' leaving the question of interpreting their findings open. Various models of chromosome evolution imply various statistics and thus can be verified by various tests. For example, RBM implies exponential distribution of the synteny block sizes, consistent

* Correspondence: maxal@cse.sc.edu; ppevzner@cs.ucsd.edu
[1]Department of Computer Science & Engineering, University of South Carolina, 301 Main St., Columbia, SC 29208, USA
[2]Department of Computer Science & Engineering, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA
Full list of author information is available at the end of the article

with the human-mouse synteny blocks observed in [2]. Pevzner and Tesler [3] introduced the 'pairwise breakpoint reuse' test and demonstrated that while RBM implies low breakpoint reuse, the human-mouse synteny blocks expose rampant breakpoint reuse. Thus RBM is consistent with the 'exponential length distribution' test [2] but inconsistent with the 'pairwise breakpoint reuse' test [34]. Both these tests are applied to *pairs* of genomes, not taking an advantage of multiple genomes that were recently sequenced. Below we introduce the 'multispecies breakpoint reuse' test and demonstrate that both RBM and FBM do not pass this test. We further propose the *Turnover Fragile Breakage Model* (TFBM) that extends FBM and complies with the multispecies breakpoint reuse test.

Technically, findings in [33] (limited breakpoint reuse between different lineages) are not in conflict with findings in [3] (rampant breakpoint reuse in chromosome evolution). Indeed, Ma *et al.* [33] only considered reuse between different branches of the phylogenetic tree (*inter-reuse*) and did not analyze reuse within individual branches (*intra-reuse*) of the tree. TFBM reconciles the recent studies supporting FBM with the Ma *et al.* [33] analysis. We demonstrate that data in [33] reveal rampant but elusive breakpoint reuse that cannot be detected via counting repeated breakages between various pairs of branches of the evolutionary tree. TFBM is an extension of FBM that reconciles seemingly contradictory results in [9-31] and [33] and explains that they do not contradict to each other. TFBM postulates that fragile regions have a limited lifespan and implies that they can migrate between different genomic locations. The intriguing implication of TFBM is that few regions in a genome are fragile at any given time raising a question of finding the currently active fragile regions in the human genome.

While many authors have discussed the causes of fragility, the question what makes certain regions fragile remains open. Previous studies attributed fragile regions to segmental duplications [35-38], high repeat density [39], high recombination rate [40], pairs of tRNA genes [41,42], inhomogeneity of gene distribution [7], and long regulatory regions [7,17,26]. Since we observed the birth and death of fragile regions, we are particularly interested in features that are also subject to birth and death process. Recently, Zhao and Bourque [38] provided a new insight into association of rearrangements with segmental duplications by demonstrating that many rearrangements are flanked by *Matching Segmental Duplications* (MSDs), that is, a pair of long similar regions located within a pair of breakpoint regions corresponding to a rearrangement event. MSDs arguably represent an ideal match for TFBM among the features that were previously implicated in breakpoint reuses. TFBM is consistent with the hypothesis that MSDs promote fragility since the similarity between MSDs deteriorates with time, implying that MSDs are also subjects to a 'birth and death' process.

Results and Discussion

Rearrangements and breakpoint graphs

For the sake of simplicity, we start our analysis with *circular genomes* consisting of circular chromosomes. While we use circular chromosomes to simplify the computational concepts discussed in the paper, all analysis is done with real (linear) mammalian chromosomes (see Alekseyev [43] for subtle differences between circular and linear chromosome analysis). We represent a circular chromosome with synteny blocks $x_1,\ldots, x_n$ as a cycle (Figure 1a) composed of n directed labeled edges (corresponding to the blocks) and n undirected unlabeled edges (connecting adjacent blocks). The directions of the edges correspond to *signs* (strands) of the blocks. We label the *tail* and *head* of a directed edge x_i as x_i^t and x_i^h respectively. We represent a genome as a *genome graph* consisting of disjoint cycles (one for each chromosomes). The edges in each cycle alternate between two colors: one color reserved for undirected edges and the other color (traditionally called 'obverse') reserved for directed edges.

Let P be a genome represented as a collection of *alternating* black-obverse cycles (a cycle is alternating if the colors of its edges alternate). For any two black edges $(u; v)$ and $(x; y)$ in the genome (graph) P , we define a *2-break* rearrangement (see [44]) as replacement of these edges with either a pair of edges (u, x), (v, y), or a pair of edges (u, y), (v, x) (Figure 2). 2-breaks extend the standard operations of reversals (Figure 2a), fissions (Figure 2b), or fusions/translocations (Figure 2c) to the case of circular chromosomes. We say that a 2-break on edges (u, x), (v, y) *uses* vertices u, x, v and y.

Let P and Q be 'black' and 'red' genomes on the same set of synteny blocks X. The *breakpoint graph* $G(P, Q)$ is defined on the set of vertices $V = \{x^t, x^h \mid x \in \chi\}$ with black and red edges inherited from genomes P and Q (Figure 1b). The black and red edges form a collection of alternating *black-red cycles* in $G(P, Q)$ and play an important role in analyzing rearrangements (see [45] for background information on genome rearrangements). The *trivial cycles* in $G(P, Q)$, formed by pairs of parallel black and red edges, represent common adjacencies between synteny blocks in genomes P and Q. Vertices of the non-trivial cycles in $G(P, Q)$ represent *breakpoints* that partition genomes P and Q into (P, Q)-synteny blocks (Figure 1c). The *2-break distance* $d(P, Q)$ between circular genomes P and Q is defined as the minimum number of 2-breaks required to transform one genome into the other (Figure 1d). In contrast to

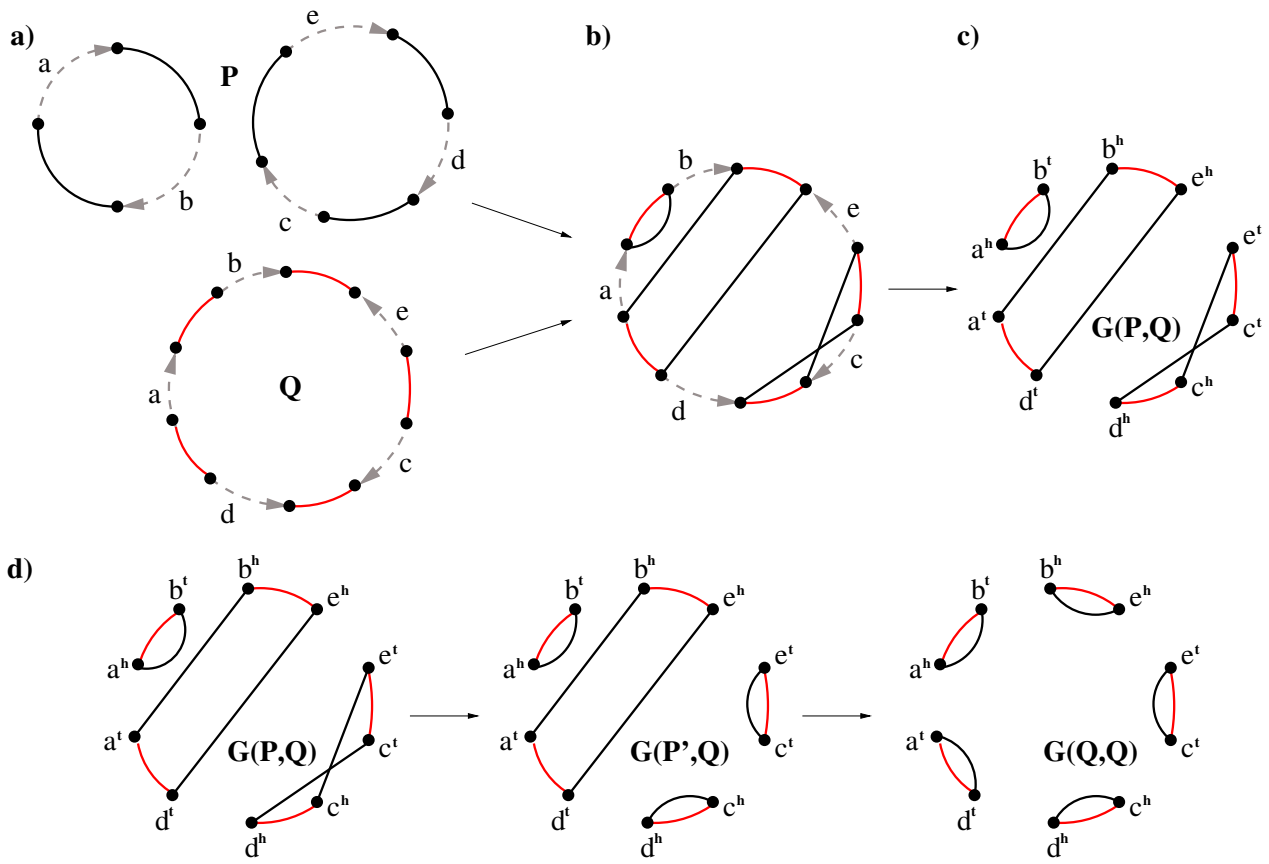

Figure 1 An example of the breakpoint graph and its transformation into an identity breakpoint graph. **(a)** Graph representation of a two-chromosomal genome $P = (+a + b)(+c + e + -d)$ as two black-obverse cycles and a unichromosomal genome $Q = (+a + b - e + c - d)$ as a red-obverse cycle. **(b)** The superposition of the genome graphs P and Q. **(c)** The breakpoint graph $G(P, Q)$ of the genomes P and Q (with removed obverse edges). The black and red edges in $G(P, Q)$ form $c(P, Q) = 2$ non-trivial black-red cycles and one trivial black-red cycle. The trivial cycle (a^h, b^t) corresponds to a common adjacency between the genes a and b in the genomes P and Q. The vertices in the non-trivial cycles represent breakpoints corresponding to the endpoints of $b(P, Q) = 4$ synteny blocks: ab, c, d, and e. By Theorem 1, the distance between the genomes P and Q is $d(P, Q) = 4 - 2 = 2$. **(d)** A transformation of the breakpoint graph $G(P, Q)$ into the identity breakpoint graph $G(Q, Q)$, corresponding to a transformation of the genome P into the genome Q with two 2-breaks. The first 2-break transforms P into a genome $P' = (+a + b)(+c\ d - e)$, while the second 2-break transforms P' into Q. Each 2-break increases the number of black-red cycles in the breakpoint graph by one, implying this transformation is shortest (see Theorem 1).

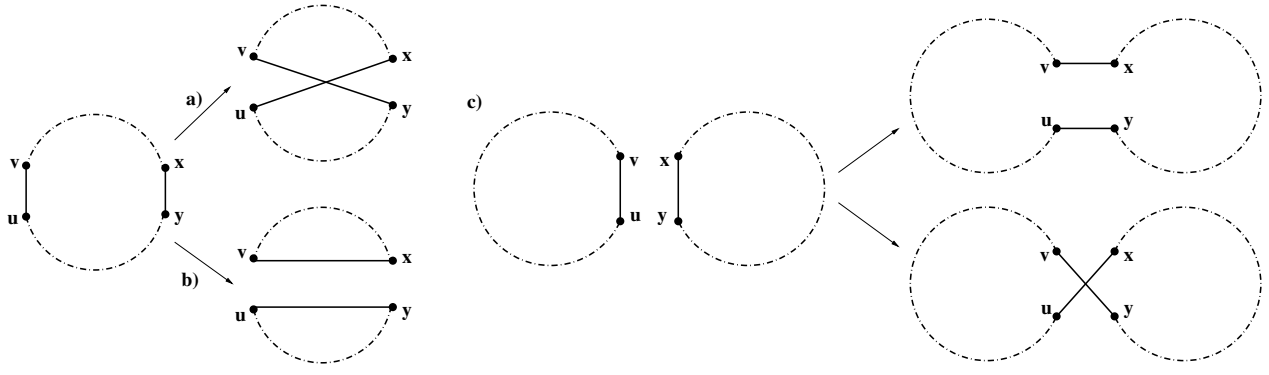

Figure 2 A 2-break on edges (u, v) and (x, y) corresponding to (a) reversal, (b) fission, (c) translocation/fusion.

the genomic distance [46] (for linear genomes), the 2-break distance for circular genomes is easy to compute [47]:

Theorem 1 *The 2-break distance between circular genomes P and Q is d(P, Q) = b(P, Q) - c(P, Q), where b (P, Q) and c(P, Q) are respectively the number of (P, Q)-synteny blocks and non-trivial black-red cycles in G (P, Q).*

Inter- and intra-breakpoint reuse

Figure 3 shows a phylogenetic tree with specified rearrangements on its branches (we write $\rho \in e$ to refer to a 2-break ρ on an edge e). We represent each genome as a genome graph (that is, a collection of cycles) on the same set V of $2n$ vertices (corresponding to the endpoints of the synteny blocks). Given a set of genomes and a phylogenetic tree describing rearrangements between these genomes, we define the notions of inter- and intra-breakpoint reuses. A vertex $v \in V$ is *inter-reused* on two distinct branches e_1 and e_2 of a phylogenetic tree if there exist 2-breaks $\rho_1 \in e_1$ and $\rho_2 \in e_2$ that both use v. Similarly, a vertex $v \in V$ is *intra-reused* on a branch e if there exist two distinct 2-breaks ρ_1, $\rho_2 \in e$ that both use v. For example, a vertex c^h is inter-reused on the branches (Q_3, P_1) and (Q_2, P_3), while a vertex f^h is intra-reused on the branch (Q_3, Q_2) of the tree in Figure 3. We define $br(e_1, e_2)$ as the number of vertices inter-reused on the branches e_1 and e_2, and $br(e)$ as the number of vertices intra-reused on the branch e. An alternative approach to measuring breakpoint intra-reuse is to define *weighted intra-reuse* of a vertex v on a branch e as $\max\{0, use(e, v) - 1\}$ where $use(e, v)$ is the number of 2-breaks on e using v. The weighted intra-reuse $BR(e)$ on the branch e is the sum of weighted intra-reuse of all vertices. We remark that if no vertex is used more than twice on a branch e then $BR(e) = br(e)$.

Given simulated data, one can compute $br(e)$ for all branches and $br(e_1, e_2)$ for all pairs of branches in the phylogenetic tree. However, for real data, rearrangements along the branches are unknown, calling for alternative ways for estimating the inter- and intra-reuse.

Cycles in the breakpoint graphs provide yet another way to estimate the inter- and intra-reuse. For a branch $e = (P, Q)$ of the phylogenetic tree, one can estimate $br(e)$ by comparing the 2-break distance $d(P, Q)$ and the number of breakpoints $2 \cdot b(P, Q)$ between the genomes P and Q. This results in the lower bound $bound(e) = 4 \cdot d(P, Q) - 2 \cdot b(P, Q)$ for $BR(e)$ [34] that also gives a good approximation for $br(e)$. On the other hand, one can estimate $br(e_1, e_2)$ as the number $bound(e_1, e_2)$ of vertices shared between non-trivial cycles in the breakpoint graphs corresponding to the branches e_1 and e_2 (similar approach was used in [48] and later explored in [12,33]). Assuming that the genomes at the internal nodes of the phylogenetic tree can be reliably reconstructed [33,49-51], one can compute $bound(e)$ and $bound(e_1, e_2)$ for all (pairs of) branches. Below we show that these bounds accurately approximate the intra- and inter-reuse.

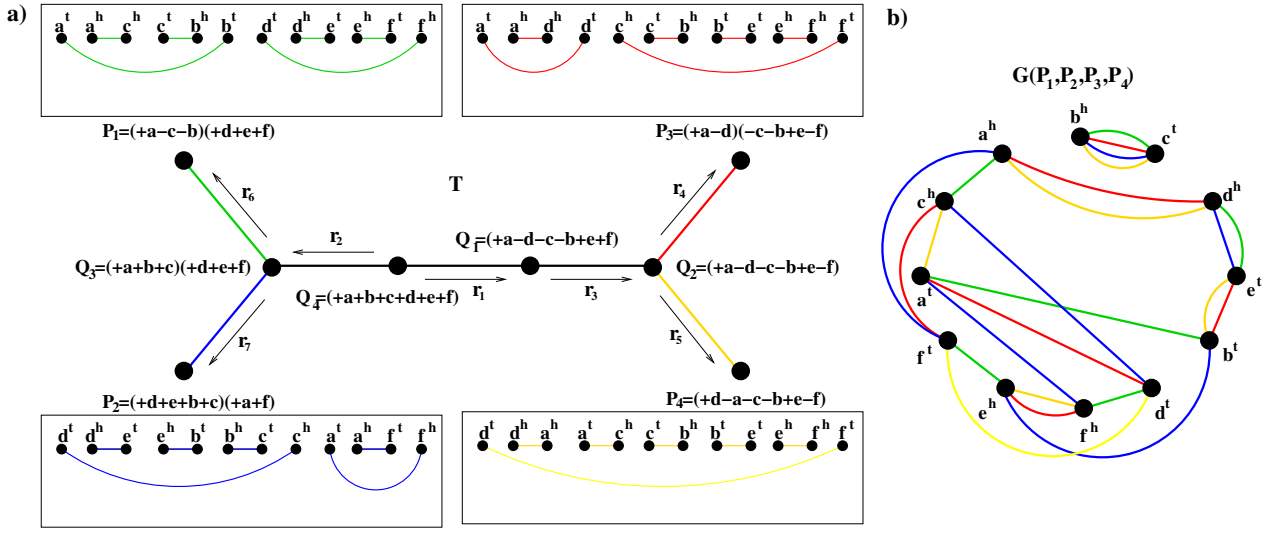

Figure 3 An example of four genomes with a phylogenetic tree and their multiple breakpoint graph. **(a)** A phylogenetic tree with four circular genomes P_1, P_2, P_3, P_4 (represented as green, blue, red, and yellow graphs respectively) at the leaves and specified intermediate genomes. The obverse edges are not shown. **(b)** The multiple breakpoint graph $G(P_1, P_2, P_3, P_4)$ is a superposition of graphs representing genomes P_1, P_2, P_3, P_4.

Analyzing breakpoint reuse (simulated genomes)

We start from analyzing simulated data based on FBM with n fragile regions present in k genomes that evolved according to a certain phylogenetic tree (for the varying parameter n). We represent one of the leaf genomes as the genome with 20 random circular chromosomes and simulate hundred 2-breaks on each branch of the tree.

Figure 4 represents a phylogenetic tree on five leaf genomes, denoted *M*, *R*, *D*, *Q*, *H*, and three ancestral genomes, denoted *MR*, *MRD*, *QH*. Table in Figure 5 presents the results of a single FBM simulation and illustrates that $bound(e_1, e_2)$ provides an excellent approximation for inter-reuses $br(e_1, e_2)$ for all 21 pairs of branches. While $bound(e)$ (on the diagonal of table in Figure 5) is somewhat less accurate, it also provides a reasonable approximation for $br(e)$. We remark that $bound(e_1, e_2) = br(e_1, e_2)$ if simulations produce the shortest rearrangement scenarios on the branches e_1 and e_2. Table in Figure 5 illustrates that this is mainly the case for our simulations.

Below we describe analytical approximations for the values in table in Figure 5. Since every 2-break uses four out of $2n$ vertices in the genome graph, a random 2-break uses a vertex v with the probability $\frac{2}{n}$. Thus, a sequence of t random 2-breaks does not use a vertex v with the probability $(1-\frac{2}{n})^t \approx e^{-\frac{2t}{n}} (for\ t \ll n)$. For branches e_1 and e_2 with respectively t_1 and t_2 random 2-breaks, the probability that a particular vertex is inter-reused on e_1 and e_2 is approximated as $(1-e^{-\frac{2t_1}{n}})\cdot(1-e^{-\frac{2t_2}{n}})$. Therefore, the expected number of inter-reused vertices is approximated as $2n\cdot(1-e^{-\frac{2t_1}{n}})\cdot(1-e^{-\frac{2t_2}{n}})$. Below we will compare the observed inter-reuse with the expected inter-reuse in FBM to see whether they are similar thus checking whether FBM represents a reasonable null hypothesis. We will use the term *scaled inter-reuse* to refer to the observed inter-reuse divided by the expected inter-reuse. If FBM is an adequate null hypothesis we expect the scaled inter-reuse to be close to one.

Similarly, a sequence of t random 2-breaks uses a vertex v exactly once with the probability $t\cdot\frac{2}{n}\cdot\left(1-\frac{2}{n}\right)^{t-1} \approx \frac{2t}{n}e^{\frac{2(t-1)}{n}}$. Therefore, the probability of a particular vertex being intra-reused on a branch with t random 2-breaks is approximately $1-e^{-\frac{2t}{n}}-\frac{2t}{n}e^{\frac{2(t-1)}{n}}$, implying that the expected intra-reuse is approximately $2n\cdot\left(1-e^{-\frac{2t}{n}}-\frac{2t}{n}e^{\frac{2(t-1)}{n}}\right)$. We will use the term *scaled intra-reuse* to refer to the observed n^e intra-reuse divided by the expected intra-reuse. Table S1 in Additional file 1 shows the scaled intra- and inter-reuse for 21 pairs of branches (averaged over 100 simulations) and illustrates that they all are close to one.

We now perform a similar simulation, this time varying the number of 2-breaks on the branches according

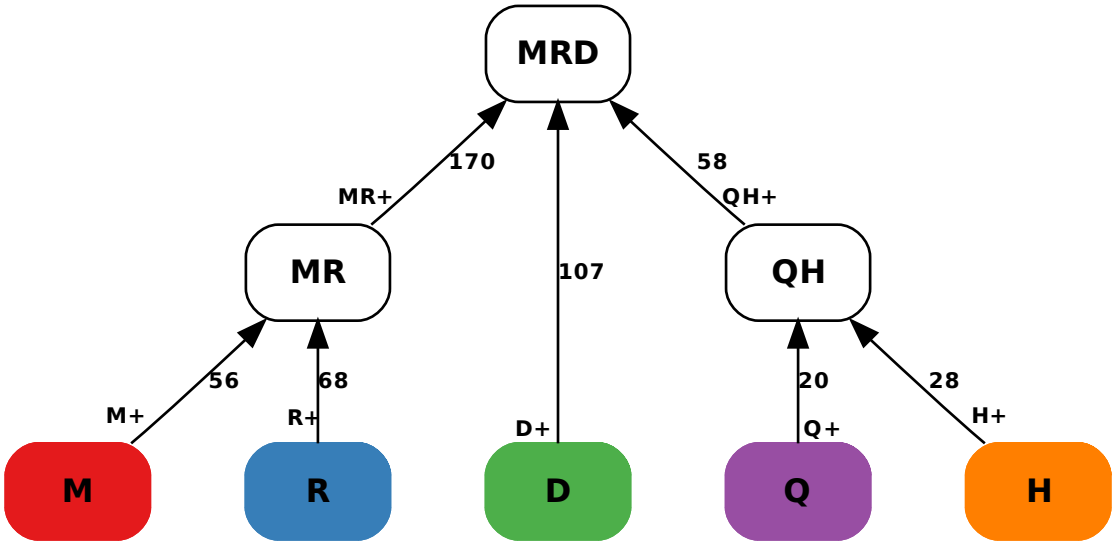

Figure 4 The phylogenetic tree *T* on five genomes *M*, *R*, *D*, *Q*, and *H*. The branches of the tree are denoted as *M+*, *R+*, *D+*, *Q+*, *H+*, *MR+*, and *QH+*.

$n = 500$	M+	R+	D+	Q+	H+	MR+	QH+
M+	63:70	106:106	103:103	97:97	108:108	98:98	113:113
R+		57:70	103:103	108:108	98:98	102:102	122:122
D+			65:74	104:104	125:125	104:104	106:106
Q+				58:68	126:126	120:120	120:120
H+					56:62	113:113	116:116
MR+						71:84	104:104
QH+							54:60

$n = 900$	M+	R+	D+	Q+	H+	MR+	QH+
M+	37:38	70:70	83:83	90:90	72:72	76:76	87:87
R+		47:50	67:67	63:63	74:74	68:68	49:49
D+			37:38	69:69	62:62	78:78	84:84
Q+				32:36	76:76	75:75	94:94
H+					40:44	64:64	68:68
MR+						42:44	64:64
QH+							28:28

$n = 1300$	M+	R+	D+	Q+	H+	MR+	QH+
M+	42:46	46:46	52:52	51:51	47:47	62:62	39:39
R+		31:34	53:53	66:66	54:54	48:48	56:56
D+			25:26	64:64	62:62	60:60	64:64
Q+				22:22	58:58	50:50	50:50
H+					30:30	57:57	72:72
MR+						31:34	42:42
QH+							19:20

Figure 5 The number of intra- and inter-reuses between seven branches of the tree in Figure 4, each of length 100, for simulated genomes with *n* fragile regions (*n* = 500, 900, 1, 300). The diagonal elements represent intra-reuses while the elements above diagonal represent inter-reuses. In each cell with numbers *x* : *y*, *x* represents the observed reuse while *y* represents the corresponding lower bound. The cells of the table are colored red (for adjacent branches like *M*+ and *R*+), green (for branches that are separated by a single branch like *M*+ and *D*+ separated by *MR*+), and yellow (for branches that are separated by two branches like *M*+ and *H*+ separated by *MR*+ and *QH*+).

to the branch lengths specified in Figure 4. Table S2 in Additional file 1 (similar to Table S1 in Additional file 1) illustrates that the lower bounds also provide accurate approximations in the case of varying branch lengths. Similar results were obtained in the case of evolutionary trees with varying topologies (data are not shown). We therefore use only lower bounds to generate table in Figure 6 rather than showing both real distances and the lower bounds as in table in Figure 5.

In the case when the branch lengths vary, we find it convenient to represent data in Table S2 in Additional file 1 in a different way (as a plot) that better illustrates variability in the scaled inter-use. We define the *distance* between branches e_1 and e_2 in the phylogenetic tree as the distance between their midpoints, that is, the overall length of the path, starting at e_1 and ending at e_2, minus $\frac{d(e_1)+d(e_2)}{2}$. For example, $d(M+,H+) = 56+170+58+28-\frac{56+28}{2} = 270$ (see Figure 4). The *x*-axis in Figure S1 in Additional file 1, 2 represents the distances between pairs of branches (21 pairs total), while *y*-axis represents the scaled inter-reuse for pairs of branches at the distance *x*.

Surprising irregularities in breakpoint reuse in mammalian genomes

The branch lengths shown in Figure 4 actually represent the approximate numbers of rearrangements on the branches of the phylogenetic tree for *M*ouse, *R*at, *D*og, maca*Q*ue, and *H*uman genomes (represented in the alphabet of 433 'large' synteny blocks exceeding 500, 000 nucleotides in human genome [50]). For the mammalian genomes, *M*, *R*, *D*, *Q*, and *H*, we first used MGRA [50] to reconstruct genomes of their common ancestors (denoted *MR*, *MRD*, and *QH* in Figure 4) and further estimated the breakpoint inter-reuse between pairs of branches of the phylogenetic tree. The resulting table in Figure 7 reveals some striking differences from the simulated data (Figure 6) that follow a peculiar pattern: the larger is the distance between two branches, the smaller is the amount of inter-reuse between them (in contrast to RBM/FBM where the amount of inter-reuse does not depend on the distance between

$n = 500$	M+	R+	D+	Q+	H+	MR+	QH+
M+	23	48	71	16	22	99	41
R+		34	83	19	25	116	49
D+			78	26	37	171	74
Q+				2	9	39	16
H+					6	51	22
MR+						186	102
QH+							25

$n = 900$	M+	R+	D+	Q+	H+	MR+	QH+
M+	13	30	44	9	13	67	25
R+		20	53	11	16	79	31
D+			46	17	24	121	45
Q+				1	4	24	9
H+					4	34	13
MR+						113	70
QH+							14

$n = 1300$	M+	R+	D+	Q+	H+	MR+	QH+
M+	8	21	33	7	9	52	19
R+		13	39	8	11	60	24
D+			34	12	17	91	34
Q+				1	3	19	7
H+					2	25	10
MR+						81	51
QH+							9

Figure 6 The estimated number of intra- and inter-reuses *bound*(*e*) and *bound*(e_1, e_2) between seven branches with varying branch length specified in Figure 4 (data simulated according to FBM). The cells are colored as in Figure 5.

branches). The statement above is imprecise since we have not described yet how to compare the amount of inter-reuse for different branches at various distances. However, we can already illustrate this phenomenon by considering branches of similar length that presumably influence the inter-reuse in a similar way (see below).

We notice that branches *M*+, *R*+, and *QH*+ have similar lengths (varying from 56 to 68 rearrangements) and construct subtables of Figure 6 (for n = 900) and Figure 7 with only three rows corresponding to these branches (Figure 8). Since the lengths of branches *M*+, *R*+, and *QH*+ are similar, FBM implies that the elements

	M+	R+	D+	Q+	H+	MR+	QH+
M+	84	68	20	4	5	58	15
R+		96	22	3	6	60	17
D+			174	17	19	98	64
Q+				12	10	25	18
H+					22	23	18
MR+						292	80
QH+							70

Figure 7 The estimated number of intra- and inter-reuses *bound*(*e*) and *bound*(e_1, e_2) between seven branches of the phylogenetic tree in Figure 4 of five mammalian genomes (real data). The cells are colored as in Figure 5.

	M+	R+	D+	Q+	H+	MR+	QH+
M+	13	30	44	9	13	67	25
R+	30	20	53	11	16	79	31
QH+	25	31	45	9	13	70	14
M+	84	68	20	4	5	58	15
R+	68	96	22	3	6	60	17
QH+	15	17	64	18	18	80	70

Figure 8 Subtables of Figure 6 for n = 900 (top part) and Figure 7 (bottom part) featuring branches *M+*, *R+*, and *QH+* as one element of the pair. The cells are colored as in Figure 5.

belonging to the same columns in table in Figure 8 should be similar. This is indeed the case for simulated data (small variations within each column) but not the case for real data. In fact, maximal elements in each column for real data exceed other elements by a factor of three to five (with an exception of the *MR+* column). Moreover, the peculiar pattern associated with these maximal elements (maximal elements correspond to red cells) suggests that this effect is unlikely to be caused by random variations in breakpoint reuses. We remind the reader that red cells correspond to pairs of adjacent branches in the evolutionary tree suggesting that breakpoint reuse is maximal between close branches and is reducing with evolutionary time. A similar pattern is observed for the other pairs of branches of similar length: adjacent branches feature much higher inter-reuse than distant branches. We also remark that the most distant pairs of branches (*H+* and *M+*, *H+* and *R+*, *Q+* and *M+*, *Q+* and *R+* in the yellow cells) feature the lowest inter-reuse. The only branch that shows relatively similar inter-reuse (varying from 58 to 80) with the branches *M+*, *R+*, and *QH+* is the branch *MR+* which is adjacent to each of these branches.

Below we modify FBM to come up with a new model of chromosome evolution, explaining the surprising irregularities in the inter-reuse across mammalian genomes.

Turnover fragile breakage model: birth and death of fragile regions

We start with a simulation of 100 rearrangements on every branch of the tree in Figure 4. However, instead of assuming that fragile regions are fixed, we assume that after every rearrangement x fragile regions 'die' and x fragile regions are 'born' (keeping a constant number of fragile regions throughout the simulation). We assume that the genome has m potentially 'breakable' sites but only n of them are currently fragile ($n \le m$) (the remaining n - m sites are currently solid). The dying regions are randomly selected from n currently fragile regions, while the newly born regions are randomly selected from m - n solid regions. The simplest TFBM with a fixed rate of the 'birth and death' process is defined by the parameters m, n, and *turnover* rate x. FBM is a particular case of TFBM corresponding to x = 0 and $n < m$, while RBM is a particular case of TFBM corresponding to x = 0 and $n = m$. While this over-simplistic model with a fixed turnover rate may not adequately describe the real rearrangement process, it allows one to analyze the general trends and to compare them to the trends observed in real data. We further remark that the goal of this paper is to develop a test for distinguishing between TFBM and FBM/RBM rather than a test for distinguishing between FBM and RBM. Thus, our simulations do not distinguish between FBM (x = 0 and $n < m$) and RBM (x = 0 and $n = m$) since they do not affect m - n inactive breakpoints in FBM. To distinguish FBM from RBM, one has to analyze the long cycles in the breakpoint graph and the distribution of synteny block sizes (see [3,8]).

The leftmost subtable of Figure 9 with x = 0 represents an equivalent of table in Figure 5 for FBM and reveals that the inter-reuse is roughly the same on all pairs of branches (approximately 110 for n = 500, approximately 70 for n = 900, approximately 50 for n = 1, 300). The right subtables of Figure 9 represent equivalents of the leftmost subtable for TFBM with the turnover rate x = 1, 2, 3 and reveal that the inter-reuse in yellow cells is lower than in green cells, while the inter-reuse in green cells is lower than in red cells.

Figure 10 shows the scaled inter-reuse averaged over yellow, green, and red cells that reveals a different behavior between FBM and TFBM. Indeed, while the scaled inter-reuse is close to 1 for all pairs of branches in the case of FBM, it varies in the case of TFBM. For example, for n = 900, m = 2, 000, and x = 3, the inter-reuse in yellow cells is approximately 40, in green cells is approximately 45, and in red cells is approximately 56. Table S3 in Additional file 1 presents the differences in

	x = 0 (FBM)							x = 1							x = 2							x = 3						
n = 500	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+
M+	67	110	109	109	110	111	108	64	93	75	66	65	92	77	63	78	57	45	46	79	57	58	69	47	36	36	68	46
R+		69	110	110	108	109	107		67	76	65	65	92	78		63	57	46	45	78	58		60	47	36	37	69	46
D+			69	109	108	109	109			66	76	77	91	90			62	56	58	78	77			58	46	47	68	69
Q+				68	108	109	110				65	92	77	93				61	79	57	76				60	68	48	66
H+					71	107	109					66	78	94					61	58	79					60	48	67
MR+						70	109						65	91						62	77						57	69
QH+							68							65							61							59
n = 900	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+
M+	42	71	72	71	71	71	71	40	64	58	53	54	65	60	39	60	51	45	45	59	51	37	55	45	39	39	56	44
R+		41	72	71	72	72	73		39	59	53	54	65	58		38	51	45	45	61	51		38	45	39	39	56	45
D+			40	73	72	70	72			41	60	59	65	65			38	50	51	58	60			38	46	46	56	54
Q+				39	73	71	73				39	64	58	64				38	59	49	60				37	55	46	56
H+					41	71	71					38	59	64					38	49	59					37	46	55
MR+						40	74						40	66						40	61						37	55
QH+							41							39							40							37
n = 1300	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+	M+	R+	D+	Q+	H+	MR+	QH+
M+	28	54	52	54	52	53	55	27	48	46	45	44	49	47	27	46	44	40	39	48	41	28	45	40	37	38	45	40
R+		28	53	53	54	53	52		29	45	44	44	48	48		28	43	40	41	46	44		27	41	38	37	44	39
D+			31	52	51	53	54			28	46	47	50	49			29	42	42	47	46			27	39	41	44	46
Q+				28	52	55	53				29	50	46	50				28	49	42	47				27	46	39	45
H+					29	53	52					28	47	49					27	42	46					27	41	44
MR+						27	53						29	49						27	48						27	46
QH+							29							28							28							27

Figure 9 The breakpoint intra- and inter-reuse (averaged over 100 simulations) for five simulated genomes *M*, *R*, *D*, *Q*, *H* under TFBM model with *m* = 2, 000 synteny blocks, *n* fragile regions, the turnover rate *x*, and the evolutionary tree shown in Figure 4 with the length of each branch equal 100. The cells are colored as in Figure 5.

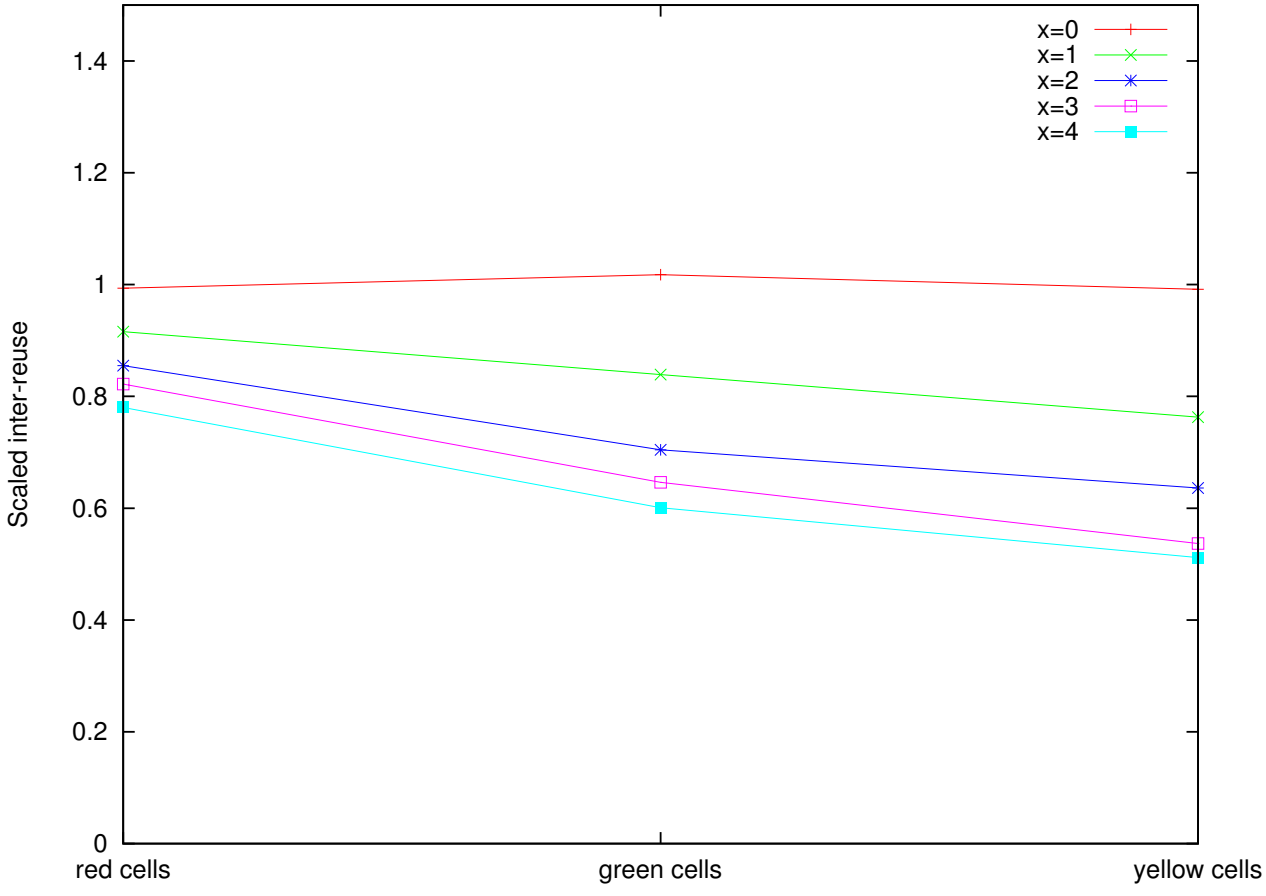

Figure 10 The scaled inter-reuse for five simulated genomes *M*, *R*, *D*, *Q*, *H* on *m* = 2,000 synteny blocks, *n* = 900 fragile regions, and the turnover rate *x* varying from zero to four with the phylogenetic tree and branch lengths shown in Figure 4. The simulations follow FBM (*x* = 0) and TFBM (*x* varies from one to four). The plot shows the scaled inter-reuse for only three reference points (corresponding to red, green, and yellow cells) that are somewhat arbitrarily connected by straight segments for better visualization.

the inter-reuse between red, green, and yellow cells as a function of m and x (for n = 900). In Methods we describe a formula for estimating the breakpoint inter-reuse in the case of TFBM that accurately approximates the values shown in Figure 10.

Table S3 in Additional file 1 demonstrates that the distribution of inter-reuses among green, red, and yellow cells differs between FBM and TFBM. We argue that this distribution (for example, the slope of the curve in Figure 10) represents yet another test to confirm or reject FBM/TFBM. However, while it is clear how to apply this test to the simulated data (with known rearrangements), it remains unclear how to compute it for real data when the ancestral genomes (as well as the parameters of the model) are unknown. While the ancestral genomes can be reliably approximated using the algorithms for ancestral genome reconstruction [33,49-51], estimating the number of fragile regions remains an open problem (see [3]). Below we develop a new test (that does not require knowledge of the number of the fragile regions n) and demonstrate that FBM does not pass this test while TFBM does, explaining the surprisingly low inter-reuse in mammalian genomes.

Multispecies breakpoint reuse test

Given a phylogenetic tree describing a rearrangement scenario, we define the multispecies breakpoint reuse on this tree as follows. For two rearrangements ρ_1 and ρ_2 in the scenario, we define the distance $d(\rho_1, \rho_2)$ as the number of rearrangements in the scenario between ρ_1 and ρ_2 plus one. For example, the distance between 2-breaks r_4 and r_6 in the tree in Figure 3 is four. We define the (actual) multispecies breakpoint reuse as a function

$$R(\ell) = \frac{\sum_{\rho_1,\rho_2 \,:\, d(\rho_1,\rho_2)=\ell} br(\rho_1,\rho_2)}{\sum_{\rho_1,\rho_2 \,:\, d(\rho_1,\rho_2)=\ell} 1}$$

that represents the total breakpoint reuse between pairs of rearrangements ρ_1, ρ_2 at the distance l divided by the number of such pairs. Here $br(\rho_1, \rho_2)$ stands for the number of vertices used by both 2-breaks ρ_1 and ρ_2.

Since the rearrangements on branches of the phylogenetic tree are unknown, we use the following sampling procedure to approximate $R(l)$. Given genomes P and Q, we sample various shortest rearrangement scenarios between these genomes by generating random 2-break transformations of P into Q. To generate a random transformation we first randomly select a non-trivial cycle C in the breakpoint graph $G(P, Q)$ with the probability proportional to $|C|/$ = 2 - 1, that is, the number of 2-breaks required to transform such a cycle into a collection of trivial cycles ($|C|$ stands for the length of C). Then we uniformly randomly select a 2-break ρ from the set of all $\binom{|C|/2}{2} = \frac{|C|(|C|-2)}{8}$ 2-breaks that splits the selected cycle C into 2 8 two and thus by Theorem 1 decreases the distance between P and Q by one (that is, $d(\rho P, Q) = d(P, Q)$ -1). We continue selecting non-trivial cycles and 2-breaks in an iterative fashion for genomes $\rho \cdot P$ and Q and so on until P is transformed into Q.

The described sampling can be performed for every branch $e = (P, Q)$ of the phylogenetic tree, essentially partitioning e into $length(e) = d(P, Q)$ sub-branches, each featuring a single 2-break. The resulting tree will have $\sum_e length(e)$ sub-branches, where the sum is taken over all branches e.

For each pair of sub-branches, we compute the number of reused vertices across them and accumulate these numbers according to the distance between these sub-branches in the tree. The *empirical multispecies breakpoint reuse* (the average reuse between all sub-branches at the distance l) is defined as the actual multispecies breakpoint reuse in a sampled rearrangement scenario. Figure S2 in Additional file 1 represents this function for five simulated genomes on m = 2, 000 synteny blocks, n = 900 fragile regions, and the turnover rate x varying from zero to four, with the same phylogenetic tree and distances between the genomes (averaged over 100 random samplings, while individual samplings produce varying results, we found that the variance of the $R(l)$ estimates across various samplings is rather small). Figure S3 in Additional file 1 demonstrates that our sampling procedure, while imperfect, accurately estimates the theoretical $R(l)$ curve (see [52] for other approaches to sampling rearrangement scenarios). Similar tests on phylogenetic trees with varying topologies demonstrated a good fit between actual, empirical, and theoretical $R(l)$ curves (data are not shown).

For the five mammalian genomes, the plot of $R(l)$ is shown in Figure 11. From this empirical curve we estimated the parameters $n \approx 196$, $x \approx 1{:}12$, and $m \approx 4,017$ (see Methods) and displayed the corresponding theoretical curve. We remark that the estimated parameter n in TFBM is expected to be larger than the observed number of synteny blocks (since not all potentially breakable regions were broken in a given evolutionary scenario). Figure S4 in Additional file 1 represents an analog of Figure 11 for the same genomes in higher resolution and illustrates that all three parameters n, x, and m depend on the data resolution.

We argue that the empirical multispecies breakpoint reuse curve $R(l)$ complements the 'exponential length distribution' [2] and 'pairwise breakpoint reuse' [3] tests

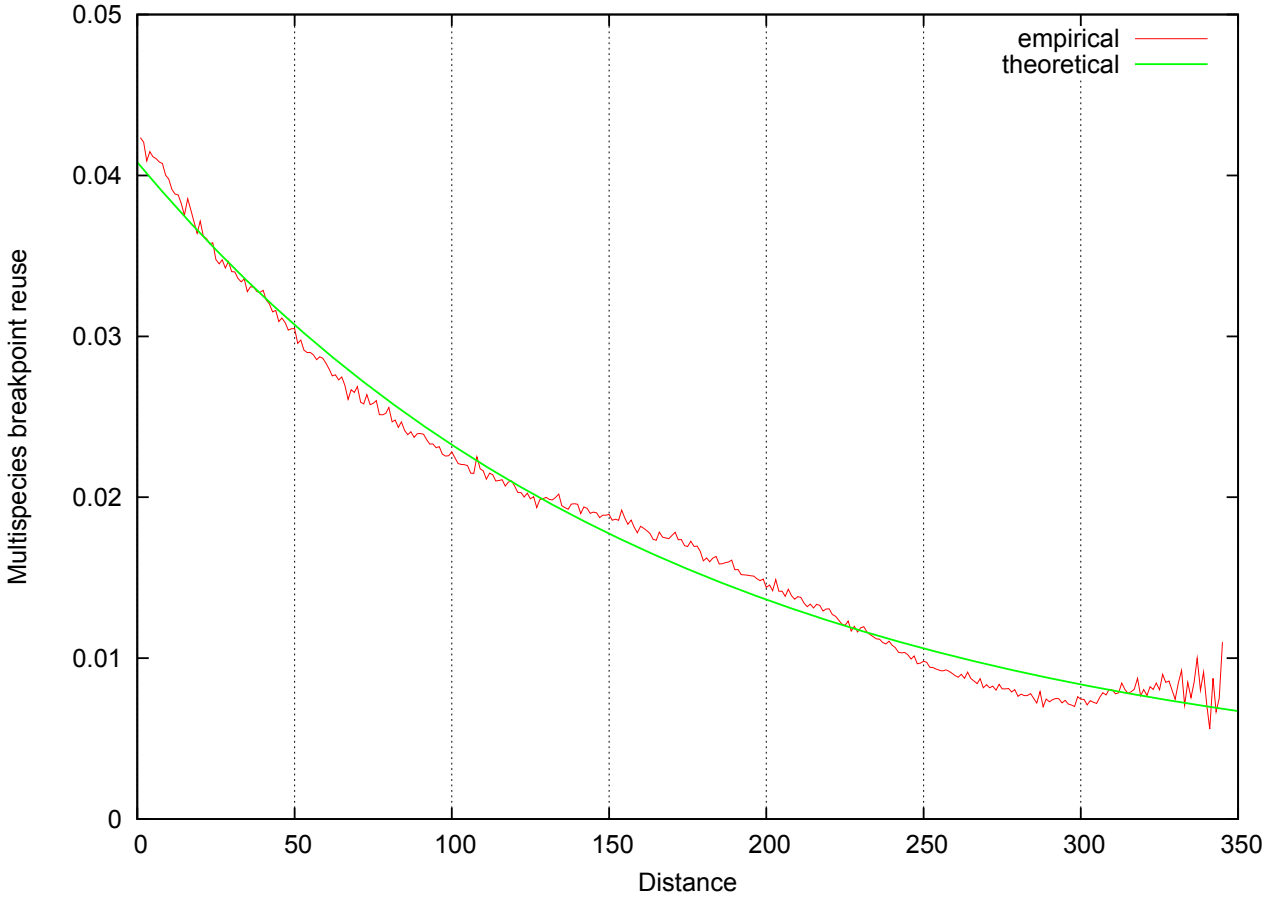

Figure 11 Empirical and theoretical curves representing the number of reuses *R*(*l*) as a function of distance *l* between pairs of subbranches of the tree in Figure 4 of the five mammalian genomes (ancestral genomes were computed using MGRA [50]). The empirical curve is averaged over 1, 000 random samplings of shortest rearrangement scenarios, while the theoretical curve represents the best fit with parameters $n \approx 196$, $x \approx 1{:}12$, and $m \approx 4, 017$ (see Methods).

as the third criterion to accept/reject RBM, FBM, and now TFBM. One can use the parameters n and x (estimated from empirical $R(l)$ curve) to evaluate the extent of the 'birth and death' process and to explain why Ma *et al.* [33] found so few shared breakpoints between different mammalian lineages. In practice, the 'multispecies breakpoint reuse test' can be applied in the same way as the Nadeau-Taylor 'exponential length distribution test' was applied in numerous papers. The Nadeau-Taylor test typically amounted to constructing a histogram of synteny blocks and evaluating (often visually) whether it fits the exponential distribution. Similarly, the 'multispecies breakpoint reuse test' amounts to constructing $R(l)$ curve and evaluating whether it significantly deviates from a horizontal line suggested by RBM and FBM. The estimated parameters of the TFBM model (see Methods) can be used to quantify the extent of these deviations.

TFBM also raises an intriguing question of what triggers the birth and death of fragile regions. As demonstrated by Zhao and Bourque [38], the disproportionately large number of rearrangements in primate lineages are flanked by MSDs. TFBM is consistent with the Zhao-Bourque hypothesis that rearrangements are triggered by MSDs since MSDs are also subject to the 'birth and death' process. Indeed, after a segmental duplication the pair of matching segments becomes subjected to random mutations and the similarity between these segments dissolves with time (a pair of segmental duplications 'disappears' after approximately 40 million years of evolution if one adopts the parameters for defining segmental duplications from [53]).

The mosaic structure of segmental duplications [53] provides an additional explanation of how MSDs may promote breakpoint re-uses and generate long cycles typical for the breakpoint graphs of mammalian genomes. The future studies of the correlation between fragile regions and MSDs in the human genome will benefit from the algorithms for precise detection of rearrangement breakpoints [54] and will be described elsewhere.

Fragile regions in the human genome

Imagine the following gedanken experiment: 25 million years ago (time of the human-macaque split) a scientist sequences the genome of the human-macaque ancestor (*QH*) and attempts to predict the sites of (future) rearrangements in the (future) human genome. The only other information the scientist has is the mouse, rat, and dog genomes. While RBM offers no clues on how to make such a prediction, FBM suggests that the scientist should use the breakpoints between one of the available genomes and *QH* as a proxy for fragile regions. For example, there are 552 breakpoints between the mouse genome (*M*) and *QH* and 34 of them were actually used in the human lineage, resulting in only 34 = 552 ≈ 6% accuracy in predicting future human breakpoints (we use synteny blocks larger than 500 K from [50]).

TFBM suggests that the scientist should rather use the *closest* genome to *QH* to better predict the human breakpoints. That can be achieved by first reconstructing the common ancestor (*MRD*) of mouse, rat, dog, and human-macaque ancestor and then using the breakpoints between *MRD* and *QH* as a proxy for the sites of rearrangements in the human lineage. 18 out 162 breakpoints between *MRD* and *QH* were used in the human lineage, resulting in 18 = 162 ≈ 11% accurate prediction of human breakpoints, nearly doubling the accuracy of predictions from distant genomes.

Now imagine that the scientist somehow gained access to the extant macaque genome. There are 68 breakpoints between *Q* and *QH* and 10 of them were used in the human lineage, resulting in 10 = 68 ≈ 16% accurate prediction of human breakpoints, again improving the accuracy of predictions. These estimates indicate that TFBM can be used to improve the prediction accuracy of *future* rearrangements in various lineages and demonstrate that the sites of *recent* rearrangements in the human and other primate lineages represent the best guess for the currently active fragile regions in the human genome.

We therefore focus on the incident branches *H*+, *Q*+, and *QH*+ and construct the breakpoint graphs *G*(*H*, *QH*), *G*(*Q*, *QH*), and *G*(*QH*, *MRD*). Figure S5 in Additional file 1 superimposes these three graphs and (together with Table S4 in Additional file 1) illustrates breakpoints that were inter-reused on the branches *H*+, *Q*+, and *QH*+. Figure 12 shows the positions of these recently affected breakpoints (projected to the human genome) that, according to TFBM, represent the best proxy for the currently active fragile regions in the human genome. Various ongoing primate genome sequencing projects will soon result in an even better estimate for the fragile regions in the human genome.

Conclusions

Since every species on Earth (including *Homo sapiens*) may speciate into multiple new species, one can ask a question: 'How will the human genome evolve in the *next* million years?' TFBM suggests the putative sites of *future* rearrangements in the human genome. The answer to the question 'Where are the (future) fragile regions in the human genome?' may be surprisingly simple: they are likely to be among the breakpoint regions that were used in various primate lineages.

Nadeau and Taylor [2] proposed RBM based on a single observation: the exponential distribution of the human-mouse synteny block sizes. There is no doubt that jumping to this conclusion was not fully justified: there are many other models (for example, FBM) that lead to the same exponential distribution of the 'visible' synteny block sizes. Currently, there is no single piece of evidence that would allow one to claim that RBM is correct and FBM is not.

While Pevzner and Tesler [3] revealed large breakpoint reuse (supporting FBM and contradicting RBM), Ma *et al.* [33] revealed low breakpoint inter-reuse (contradicting FBM). This discovery calls for yet another generalization of FBM. The proposed TFBM model not only passes both 'exponential length distribution' test (motivation for RBM) and 'pairwise breakpoint reuse' test (motivation for FBM) but also explains the puzzling discovery of limited breakpoint inter-reuse in [33]. We therefore argue that TFBM is a more accurate model of chromosome evolution, allowing one to approximate the currently active fragile regions in the human genome.

Needless to say, TFBM, similarly to RBM and FBM (or various models of point mutations, for example, Jukes-Cantor model), is a simplistic model of chromosome evolution that is only an approximation of the real evolutionary process. Moreover, in the current paper we considered TFBM only for the case of 2-breaks and did not include other rearrangements such as transpositions. However, it is fair to assume that transpositions are as likely to happen on incident branches as on distant branches, implying that they cannot possibly cause the reduced breakpoint inter-reuse on distant branches. In addition to limitations of TFBM as a model, there exists a concern whether computation of empirical multispecies breakpoint reuse (that requires reconstruction of ancestral genomes) may be affected by errors in reconstruction of ancestral genomes. While various tools for ancestral genome reconstruction (such as MGRA [50] and inferCARs [33]) were shown to be quite accurate (in particular, they produce nearly identical results while using very different algorithms), it is a challenging open problem to evaluate the multispecies breakpoint reuse without explicitly computing ancestral genomes.

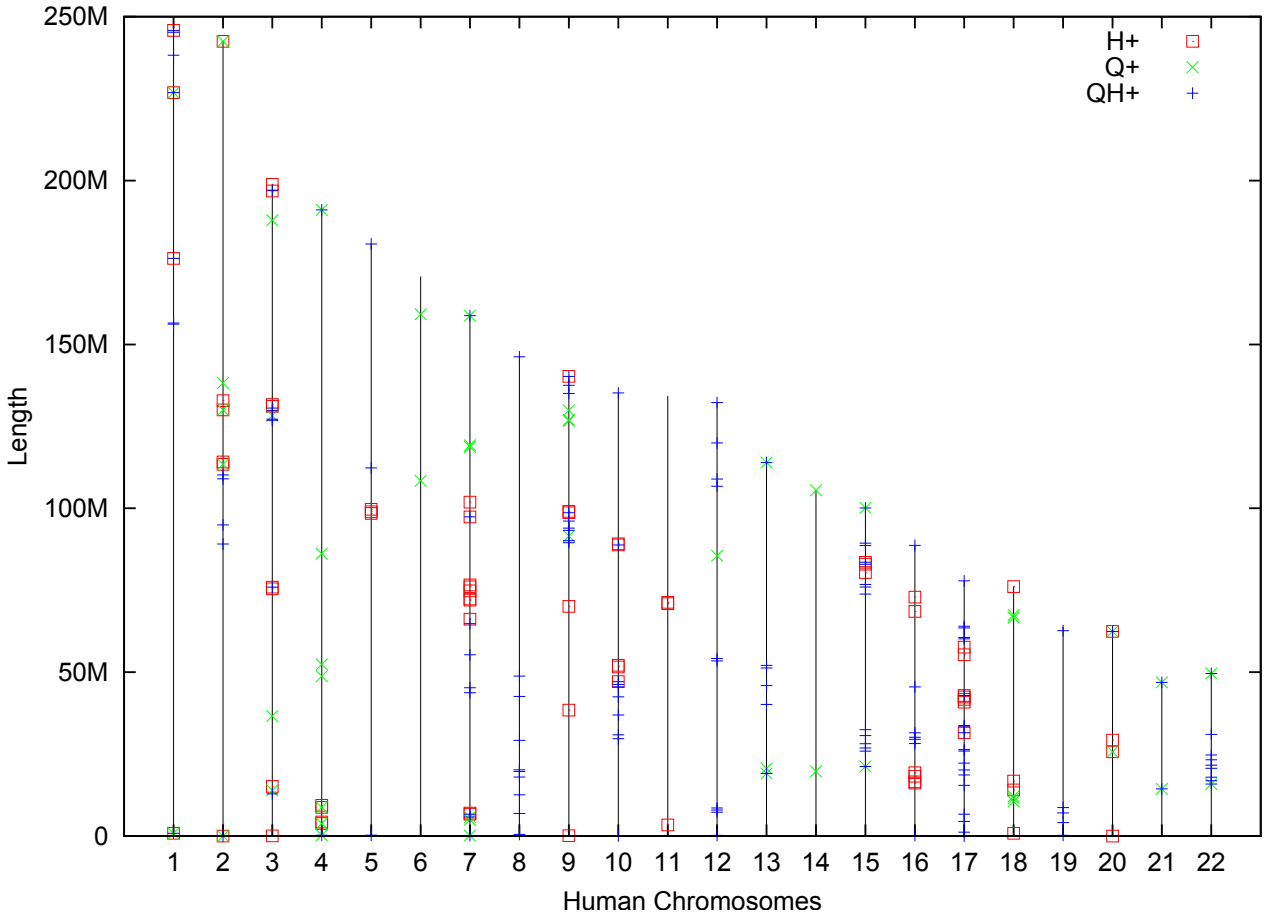

Figure 12 Positions of regions broken on the evolutionary path from the rodent-primate-carnivore ancestor (that is, on *H+*, *Q+*, and *QH+* branches) projected to the human chromosomes.

The key point of this paper is the birth and death process of fragile regions rather than a specific model aimed at estimating the hidden parameters of this process. TFBM is merely an initial and over-simplistic attempt to estimate these parameters. The parameters predicted by TFBM (for example, the number of active fragile regions) are currently difficult to superimpose with scarce information about rearrangements in only seven reliably completed mammalian genomes, not unlike the parameters of RBM derived in 1984 when no high-resolution comparative mammalian genomic architectures were available. However, similarly to comparative mapping efforts in early 1990 s that confirmed the Nadeau-Taylor estimates, we believe that imminent sequencing of over 400 primate species will soon provide the detailed information about chromosomal fragility in human genome and will allow one to verify the TFBM parameters.

Similarly to the discovery of breakpoint reuse in 2003 [3], there is currently only indirect evidence supporting the birth and death of fragile regions in chromosome evolution. However, we hope that, similarly to FBM (that led to many follow-up studies supporting the existence of fragile regions), TFBM will trigger further investigations of the fragile regions longevity.

Materials and methods

Computing multispecies breakpoint reuse in the TFBM model

Let *Fragile* and *Solid* be the sets of n initial fragile regions and m - n initial solid regions respectively. In TFBM, the sets *Fragile* and *Solid* change in accordance with the turnover rate x, that is, after every 2-break x fragile regions (corresponding to $2x$ vertices in the breakpoint graph) from *Fragile* are moved to *Solid* and vice versa.

For a vertex in the set *Fragile*, we evaluate the probability $P(l)$ that this vertex still belongs to *Fragile* after l 2-breaks. After every 2-break, a vertex from *Fragile* moves to *Solid* with the probability $\frac{x}{n}$, while a vertex

from *Solid* moves to *Fragile* with the probability $\frac{x}{m-n}$. Therefore,

$$P(\ell+1) = P(\ell)\cdot(1-\frac{x}{n}) + (1-P(\ell))\cdot\frac{x}{m-n}$$
$$= (1-\frac{xm}{n(m-n)})\cdot P(\ell) + \frac{x}{m-n}.$$

Solution to this recurrence with the initial condition $P(0) = 1$ is $P(\ell) = \frac{m-n}{m}\left(1-\frac{xm}{n(m-n)}\right)^{\ell} + \frac{n}{m}$. We now compute the expected reuse between 2-breaks ρ_1 and ρ_2 separated by l other 2-breaks. Since every 2-break uses 4 vertices, the probability that it uses a particular vertex in *Fragile* is $\frac{2}{n}$. Since the 2-break used 4 vertices, the expected reuse between ρ_1 and ρ_2 is:

$$R(\ell) = 4\cdot\frac{2}{n}\cdot P(\ell) = \frac{8\cdot(m-n)}{n\cdot m}\left(1-\frac{xm}{n(m-n)}\right)^{\ell} + \frac{8}{m}.$$

Figure S6 in Additional file 1 demonstrates that this formula fits simulated data well, thus opening a possibility to determine the parameters m, n, and x for given real genomes.

We remark that if $\frac{xml}{n(m-n)} \ll 1$ is approximated by a line $\frac{8\cdot(m-n)}{n\cdot m}\left(1-\frac{xm}{n(m-n)}\ell\right) + \frac{8}{m} = \frac{8}{n} - \frac{8x}{n^2}\ell$ that does not depend on m.

The difference between empirical and theoretical estimates for *R*(*l*)

Figure S3 in Additional file 1 illustrates the results of simulating of 400 2-breaks according to TFBM with parameters m = 2, 000, n = 900, x = 1. As expected, the theoretical curve and the curve derived from simulated data (without sampling of various rearrangement scenarios) are nearly identical. We now assume that only five out of 401 simulated genomes are available (after 0, 100, 200, 300, and 400 rearrangements) and use sampling of rearrangement scenarios to compute the empirical $R(l)$ (Figure S3 in Additional file 1). One can see that empirical $R(l)$ differs from the theoretical $R(l)$, particularly for small'. To understand why the empirical curve (obtained via sampling of rearrangement scenarios) differs from the theoretical curve, one has to realize that the multi-species breakpoint reuse test requires *multiple* genome to reveal the 'birth and death' of fragile regions. Indeed, it is impossible to detect this process from only two genomes: for example, sampling of rearrangement scenarios on a single branch (simulated with TFBM with parameters described above) produces a nearly horizontal curve $R(l) \approx 0.0083$ with TFBM signal lost. The green curve follows the same horizontal trend for small l (for example $l < 100$) that typically represent pairs of 2-breaks on the *same* branch. However, for distances larger than the shortest branches, the theoretical curve approximates the empirical $R(l)$ curve well. The reason this 'horizontal trend' is not seen in Figure 11 most likely explained by the fact that H+ and Q+ branches in the corresponding phylogenetic tree are rather short thus masking this effect.

Additional material

Additional file 1: Supplementary tables and figures. Additional file 1 contains supplementary Tables S1, S2, S3, S4 and Figures S1, S2, S3, S4, S5, S6.

Abbreviations

FBM: fragile breakage model: MSDs: matching segmental duplications: RBM: random breakage model: TFBM: turnover fragile breakage model.

Acknowledgements

The authors thank Glenn Tesler and Jian Ma for many helpful comments.

Author details

[1]Department of Computer Science & Engineering, University of South Carolina, 301 Main St., Columbia, SC 29208, USA. [2]Department of Computer Science & Engineering, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA.

Authors' contributions

Both authors participated in data analysis and writing the manuscript. MA also performed the simulations and prepared illustrations. Both authors read and approved the final manuscript.

Received: 15 July 2010 Revised: 5 October 2010
Accepted: 30 November 2010 Published: 30 November 2010

References

1. Ohno S: *Evolution by Gene Duplication* Berlin: Springer; 1970.
2. Nadeau JH, Taylor BA: **Lengths of chromosomal segments conserved since divergence of man and mouse.** *Proc Natl Acad Sci U S A* 1984, **81**:814-818.
3. Pevzner P, Tesler G: **Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.** *Proc Natl Acad Sci U S A* 2003, **100**:7672-7677.
4. Sankoff D, Trinh P: **Chromosomal breakpoint reuse in genome sequence rearrangement.** *Journal of Computational Biology* 2005, **12**:812-821.
5. Sankoff D: **The signal in the genome.** *PLoS Comput Biol* 2006, **2**:e35.
6. Bergeron A, Mixtacki J, Stoye J: **On computing the breakpoint reuse rate in rearrangement scenarios.** *Lecture Notes in Bioinformatics* 2008, **5267**:226-240.
7. Peng Q, Pevzner PA, Tesler G: **The fragile breakage versus random breakage models of chromosome evolution.** *PLoS Computational Biology* 2006, **2**:e14.
8. Alekseyev MA, Pevzner PA: **Are there rearrangement hotspots in the human genome?.** *PLoS Computational Biology* 2007, **3**:e209.
9. van der Wind AE, Kata SR, Band MR, Rebeiz M, Larkin DM, Everts RE, Green CA, Liu L, Natarajan S, Goldammer T, Lee JH, McKay S, Womack JE, Lewin HA: **A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates.** *Genome Research* 2004, **14**:1424-1437.
10. Bailey J, Baertsch R, Kent W, Haussler D, Eichler E: **Hotspots of mammalian chromosomal evolution.** *Genome Biology* 2004, **5**:R23.

11. Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM: **Human, mouse, and rat genome large-scale rearrangements: stability versus speciation.** *Genome Research* 2004, **14**:1851-1860.
12. Murphy WJ, Larkin DM, van der Wind AE, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers CN, Milan D, Ostrander EA, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'Brien SJ, Pevzner PA, Lewin HA: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative map.** *Science* 2005, **309**:613-617.
13. Webber C, Ponting CP: **Hotspots of mutation and breakage in dog and human chromosomes.** *Genome Research* 2005, **15**:1787-1797.
14. Hinsch H, Hannenhalli S: **Recurring genomic breaks in independent lineages support genomic fragility.** *BMC Evolutionary Biology* 2006, **6**:90.
15. Ruiz-Herrera A, Castresana J, Robinson TJ: **Is mammalian chromosomal evolution driven by regions of genome fragility?.** *Genome Biology* 2006, **7**: R115.
16. Yue Y, Haaf T: **7E olfactory receptor gene clusters and evolutionary chromosome rearrangements.** *Cytogenet Genome Res* 2006, **112**:6-10.
17. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** *Genome Research* 2007, **17**:545-555.
18. Mehan MR, Almonte M, Slaten E, Freimer NB, Rao PN, Ophoff RA: **Analysis of segmental duplications reveals a distinct pattern of continuation-of-synteny between human and mouse genomes.** *Human Genetics* 2007, **121**:93-100.
19. Caceres M, Sullivan RT, Thomas JW: **A recurrent inversion on the eutherian X chromosome.** *Proc Natl Acad Sci U S A* 2007, **104**:18571-18576.
20. Gordon L, Yang S, Tran-Gyamfi M, Baggott D, Christensen M, Hamilton A, Crooijmans R, Groenen M, Lucas S, Ovcharenko I, Stubbs L: **Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions.** *Genome Research* 2007, **17**:1603-1613.
21. Ruiz-Herrera A, Robinson T: **Chromosomal instability in Afrotheria: fragile sites, evolutionary breakpoints and phylogenetic inference from genome sequence assemblies.** *BMC Evolutionary Biology* 2007, **7**:199.
22. Misceo D, Capozzi O, Roberto R, Dell'Oglio MP, Rocchi M, Stanyon R, Archidiacono N: **Tracking the complex flow of chromosome rearrangements from the Hominoidea Ancestor to extant Hylobates and Nomascus Gibbons by high-resolution synteny mapping.** *Genome Research* 2008, **18**:1530-1537.
23. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM: **Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes.** *Genetics* 2008, **179**:1657-1680.
24. Ruiz-Herrera A, Robinson TJ: **Evolutionary plasticity and cancer breakpoints in human chromosome 3.** *BioEssays* 2008, **30**:1126-1137.
25. Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA: **Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories.** *Genome Research* 2009, **19**:770-777.
26. Mongin E, Dewar K, Blanchette M: **Long-range regulation is a major driving force in maintaining genome integrity.** *BMC Evolutionary Biology* 2009, **9**:203.
27. Kulemzina A, Trifonov V, Perelman P, Rubtsova N, Volobuev V, Ferguson-Smith M, Stanyon R, Yang F, Graphodatsky A: **Cross-species chromosome painting in Cetartiodactyla: reconstructing the karyotype evolution in key phylogenetic lineages.** *Chromosome Research* 2009, **17**:419-436.
28. Longo M, Carone D, Program NCS, Green E, O'Neill M, O'Neill R: **Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty.** *BMC Genomics* 2009, **10**:334.
29. Larkin D: **Role of chromosomal rearrangements and conserved chromosome regions in amniote evolution.** *Mol Gen Mikrobiol Virusol* 2010, **25**:3-8, [Article in Russian].
30. Mlynarski E, Obergfell C, O'Neill M, O'Neill R: **Divergent patterns of breakpoint reuse in Muroid rodents.** *Mammalian Genome* 2010, **21**:77-87.
31. von Grotthuss M, Ashburner M, Ranz JM: **Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila.** *Genome Research* 2010, **20**:1084-1096.
32. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci U S A* 2003, **100**:11484-11489.
33. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent JW, Blanchette M, Haussler D, Miller W: **Reconstructing contiguous regions of an ancestral genome.** *Genome Research* 2006, **16**:1557-1565.
34. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Research* 2003, **13**:37-45.
35. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X: **Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements.** *Human Molecular Genetics* 2003, **12**:2201-2208.
36. Koszul R, Dujon B, Fischer G: **Stability of large segmental duplications in the yeast genome.** *Genetics* 2006, **172**:2211-2222.
37. San Mauro D, Gower DJ, Zardoya R, Wilkinson M: **A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome.** *Mol Biol Evol* 2006, **23**:227-234.
38. Zhao H, Bourque G: **Recovering genome rearrangements in the mammalian phylogeny.** *Genome Research* 2009, **19**:934-942.
39. Myers S, Spencer CCA, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G: **The distribution and causes of meiotic recombination in the human genome.** *Biochemical Society Transactions* 2006, **34**:526-530.
40. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome.** *Science* 2005, **310**:321-324.
41. Lecompte O, Ripp R, Puzos-Barbe V, Duprat S, Heilig R, Dietrich J, Thierry JC, Poch O: **Genome evolution at the genus level: comparison of three complete genomes of Hyperthermophilic Archaea.** *Genome Research* 2001, **11**:981-993.
42. Eichler EE, Sankoff D: **Structural dynamics of eukaryotic chromosome evolution.** *Science* 2003, **301**:793-797.
43. Alekseyev MA: **Multi-break rearrangements and breakpoint re-uses: from circular to linear genomes.** *Journal of Computational Biology* 2008, **15**:1117-1131.
44. Alekseyev MA, Pevzner PA: **Multi-break rearrangements and chromosomal evolution.** *Theoretical Computer Science* 2008, **395**:193-202.
45. Fertin G, Labarre A, Rusu I, Tannier E: *Combinatorics of Genome Rearrangements* Cambridge, MA: The MIT Press,; 2009.
46. Hannenhalli S, Pevzner P: **Transforming men into mouse (polynomial algorithm for genomic distance problem).** *Proceedings of the 36th Annual Symposium on Foundations of Computer Science* 1995, 581-592.
47. Yancopoulos S, Attie O, Friedberg R: **Efficient sorting of genomic permutations by translocation, inversion and block interchange.** *Bioinformatics* 2005, **21**:3340-3346.
48. Larkin DM, Everts-van der Wind A, Rebeiz M, Schweitzer PA, Bachman S, Green C, Wright CL, Campos EJ, Benson LD, Edwards J, Liu L, Osoegawa K, Womack JE, de Jong PJ, Lewin HA: **A cattle-human comparative map built with cattle bac-ends and human genome sequence.** *Genome Research* 2003, **13**:1966-1972.
49. Ma J, Ratan A, Raney BJ, Suh BB, Miller W, Haussler D: **The infinite sites model of genome evolution.** *Proc Natl Acad Sci U S A* 2008, **105**:14254-14261.
50. Alekseyev MA, Pevzner PA: **Breakpoint graphs and ancestral genome reconstructions.** *Genome Research* 2009, **19**:943-957.
51. Swenson K, Moret B: **Inversion-based genomic signatures.** *BMC Bioinformatics* 2009, **10**:S7.
52. Miklos I, Darling AE: **Efficient sampling of parsimonious inversion histories with application to genome rearrangement in yersinia.** *Genome Biol Evol* 2009, **1**:153-164.
53. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE: **Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.** *Nature Genetics* 2007, **39**:1361-1368.
54. Lemaitre C, Zaghloul L, Sagot MF, Gautier C, Arneodo A, Tannier E, Audit B: **Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation.** *BMC Genomics* 2009, **10**:335.

doi:10.1186/gb-2010-11-11-r117
Cite this article as: Alekseyev and Pevzner: **Comparative genomics reveals birth and death of fragile regions in mammalian evolution.** *Genome Biology* 2010 **11**:R117.