Resource

Recovering genome rearrangements in the mammalian phylogeny

Hao Zhao and Guillaume Bourque[1]

Computational and Mathematical Biology, Genome Institute of Singapore, Singapore 138672, Singapore

The analysis of genome rearrangements provides a global view on the evolution of a set of related species. We present a new algorithm called EMRAE (efficient method to recover ancestral events) to reliably predict a wide-range of rearrangement events in the ancestry of a group of species. Using simulated data sets, we show that EMRAE achieves comparable sensitivity but significantly higher specificity when predicting evolutionary events relative to other tools to study genome rearrangements. We apply our approach to the synteny blocks of six mammalian genomes (human, chimpanzee, rhesus macaque, mouse, rat, and dog) and predict 1109 rearrangement events, including 831 inversions, 15 translocations, 237 transpositions, and 26 fusions/fissions. Studying the sequence features at the breakpoints of the primate rearrangement events, we demonstrate that they are not only enriched in segmental duplications (SDs), but that the enrichment of matching pairs of SDs is even stronger within the pairs of breakpoints associated with recovered events. We also show that pairs of L1 repeats are frequently associated with ancestral inversions across all studied lineages. Together, this substantiates the model that regions of high sequence identity have been associated with rearrangement events throughout the mammalian phylogeny.

[Supplemental material is available online at www.genome.org. EMRAE source code and predictions are available online at http://www.gis.a-star.edu.sg/~bourque/.]

The genomes of extant species can be viewed as patchworks of synteny blocks, or contiguous ancestral regions (CARs) (Ma et al. 2006), with the ordering of blocks being the result of a series of rearrangement events in particular lineages. Such orderings have been used for several years to study phylogenetic relationships since they enable a whole-genome view on the history of a set of species (Sankoff et al. 1992; Hannenhalli et al. 1995; Cosner et al. 2000). Specifically, a number of reconstruction algorithms have been developed seeking either a most parsimonious rearrangement scenario using different heuristics (Moret et al. 2001; Bourque and Pevzner 2002) or based on a maximum likelihood framework (Miklos 2003; Larget et al. 2005). The application of these reconstruction algorithms to vertebrate genomes in particular has led to new insights into the evolution of these species (Bourque et al. 2005; Murphy et al. 2005).

Phylogenetic reconstruction algorithms are typically evaluated based on three criteria: (1) their ability to recover the correct tree topology (Blanchette et al. 1999), (2) the total number of rearrangements in the scenario recovered (Moret et al. 2001; Bourque and Pevzner 2002), and (3) the quality of the ancestral reconstructions (Bourque et al. 2006; Froenicke et al. 2006; Ma et al. 2006; Rocchi et al. 2006). One of the major challenges faced by such algorithms is the nonuniqueness, for most realistic instances of the problem, of both the ancestral reconstructions and of the rearrangement scenarios (Bourque et al. 2006; Rocchi et al. 2006; Darling et al. 2008). For this reason, we focus here on a different criterion: the accuracy of the rearrangements recovered. The rationale is that the analysis of these highly reliable rearrangement events is likely to bring new insights into our understanding of the evolutionary forces associated with such changes.

In a recent paper (Zhao and Bourque 2007), we have developed an algorithm to trace back ancestral rearrangement events on a fixed phylogenetic tree. The approach relies on the identification of adjacencies shared by a significant fraction of the genomes in the phylogeny. The algorithm, called EMRAE (efficient method to recover ancestral events), was initially designed to recover reversals (or inversions) and transpositions and was restricted to uni-chromosomal genomes. In this study, we significantly extended this algorithm to study a wider range of rearrangement events (reversals, transpositions, translocations, fusions, and fissions) and, importantly, such that it is applicable to multichromosomal genomes. Using simulated data sets, we also compared EMRAE to MGR (Bourque and Pevzner 2002), another tool to study genome rearrangement in multichromosomal genomes, and showed that EMRAE achieves comparable sensitivity but significantly higher specificity when predicting evolutionary events.

Finally, we applied our new approach to the synteny blocks of six mammalian genomes (human, chimpanzee, rhesus macaque, mouse, rat, and dog) to recover a set of highly reliable rearrangement events and to explore the underlying evolutionary mechanisms that drive such rearrangement events. Going beyond the fact that breakpoint regions are enriched in segmental duplications (SDs) (Bailey et al. 2004a,b), we showed that there is an even stronger enrichment of pairs of SDs in the pairs of breakpoint regions associated with primate rearrangement events. Similarly, we also showed that pairs of L1 repeats are frequently associated with inversion events across all studied lineages. This substantiates that intra-genome homologous regions have been linked to rearrangement events throughout the mammalian phylogeny.

Results

EMRAE: An algorithm to predict ancestral rearrangement events

We present the general concepts behind our approach to recover ancestral rearrangement events on a fixed phylogeny using a simple example. Assume that T is a phylogenetic tree with six extant

[1]Corresponding author.
E-mail bourque@gis.a-star.edu.sg; fax 65-6478-9058.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.086009.108.

19:934–942 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

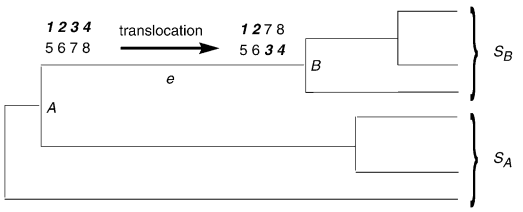

Figure 1. Schematic representation showing an ancestral translocation event on the edge $e = (A, B)$. The two simple ancestors A and B have two chromosomes with four synteny blocks labeled from 1 to 8.

genomes and assume further that A and B are two ancestral nodes on this tree and that we want to infer ancestral events on the edge $e = (A, B)$ of T (see Fig. 1). Note that the removal of e from T partitions the extant genomes into two subsets, S_A and S_B, which contain three genomes each. Assume that A has two chromosomes with only four blocks each (Chr1 = 1 2 3 4 and Chr2 = 5 6 7 8), and that the only rearrangement event on e is a translocation that exchanges the segment 3 4 in Chr1 with the segment 7 8 in Chr2. This translocation will transform A into B, where B is Chr1 = 1 2 7 8 and Chr2 = 5 6 3 4. Define an "adjacency" $a(c_i, c_{i+1})$ as an ordered pair of integers $c_i\ c_{i+1}$ or its inverse $-c_{i+1}\ -c_i$ found in a given genome. By comparing the adjacencies of A and B, we observe that the translocation changes two adjacencies in A, $a_1 = a(2, 3)$ and $a_2 = a(6, 7)$, and leads to two new adjacencies in B, $b_1 = a(2, 7)$ and $b_2 = a(6, 3)$, while the other adjacencies in A are left unchanged. If a_1 and a_2 are not disrupted further on the paths from A to the genomes in S_A and on the paths from B to the genomes in S_B, then a_1 and a_2 will be found in every genome of S_A, and neither of them will be found in a genome of S_B. We call the adjacencies a_1 and a_2 the "conserved adjacencies" of S_A. Similarly, b_1 and b_2 are the conserved adjacencies of S_B. Finally, we call the conserved adjacencies of S_A and S_B the conserved adjacencies of the edge e.

The concept of conserved adjacencies is important because in contrast to the adjacencies of A and B, the adjacencies of the extant genomes in S_A and S_B are observable, and thus conserved adjacencies can be directly computed. Moreover, under a parsimony assumption, different types of ancestral rearrangements (reversals, transpositions, translocations, fusions, and fissions) will leave distinctive signatures in the conserved adjacencies, and it will be possible to trace back events that have occurred (see Methods). For instance, in the example above, because of the specific structure of the conserved adjacencies a_1, a_2, b_1, and b_2, it will be possible to recover the precise translocation that occurred on the edge e. The algorithm EMRAE implements these basic principles along with several additional rules to provide sufficient flexibility to detect rearrangements even when the same breakpoints are reused in a limited number of genomes (see Methods).

EMRAE is implemented in Java, and its application to the two mammalian genome data sets described below took only 38 and 7 sec on a PC with a 2-GHz CPU and 2 GB of RAM. The source code and our detailed predictions are available online at http://www.gis.a-star.edu.sg/~bourque.

Performance of EMRAE on simulated data sets

We now describe experiments performed on simulated data sets involving multichromosomal genomes to evaluate the ability of EMRAE to recover reversals, transpositions, translocations, fusions, and fissions and to compare it to other tools used to study genome rearrangements. In contrast to the work done in Zhao and Bourque (2007), the sensitivity and accuracy of the predictions will only be compared to MGR (Bourque and Pevzner 2002) because GRAPPA (Moret et al. 2001) is only applicable to uni-chromosomal genomes. We experimented with two rearrangement models:

1. Reversals and translocations (Rev+Tloc model). In this model, both types of events are assumed to be equally likely.
2. Reversals, transpositions, translocations, and fusions/fissions (All events model). In this model, the events are randomly selected according to the ratios 10:2:2:0.1. These ratios were empirically estimated from the mammalian genome data set that is presented below (see Methods).

For both rearrangement models, we generated simulated instances using a phylogenetic tree with seven genomes, 100 ancestral blocks, and various evolutionary rates (see Methods). The average sensitivity (percentage of correct events that are predicted) and specificity (percentage of predicted events that are correct) for EMRAE and MGR are reported in Figure 2. We observe that EMRAE achieves high specificity without compromising the sensitivity in the prediction of ancestral events. For instance, when the average number of events per edge is five and for the rearrangement model with reversals and translocations only, both EMRAE and MGR recover a significant proportion of the actual events (~45%), but the specificity of these predictions is much higher with EMRAE (85%) as compared to MGR (44%).

To test the robustness of EMRAE in predicting events, we also evaluated the approach using more balanced rearrangement ratios in the All event models (see Supplemental Fig. 1). Overall, we found that EMRAE consistently achieved comparable sensitivity and higher specificity as compared to MGR.

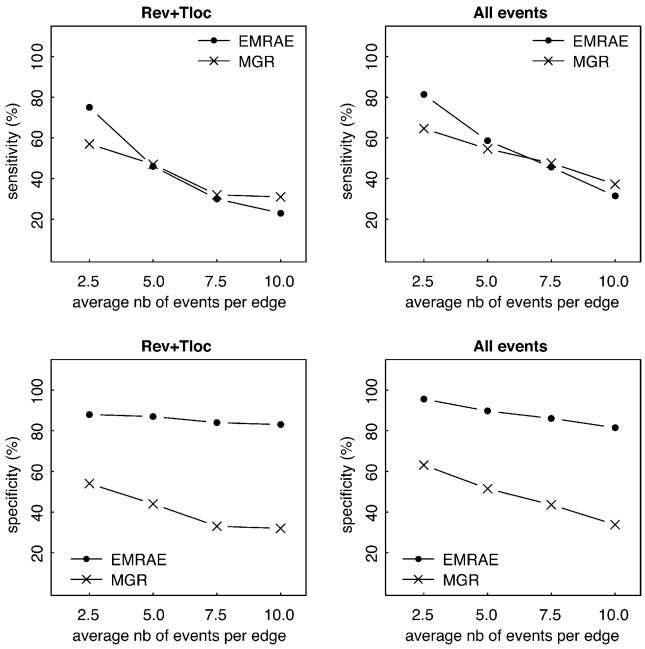

Figure 2. Sensitivity and specificity of EMRAE and MGR in predicting rearrangement events based on simulated data sets using two rearrangement models: Rev+Tloc (reversals and translocations) and All events (reversals, transpositions, translocations, and fusions/fissions). The x-axis corresponds to the evolutionary rate as defined by the average number of events simulated on each edge of the tree. nb, number.

Recovering genome rearrangements in the mammalian phylogeny

For the main analysis, we selected six mammalian genomes with high-quality assemblies: human, chimpanzee (chimp), rhesus macaque (rhesus), mouse, rat, and dog. Using an approach previously described (Ma et al. 2006), we constructed a set of contiguous ancestral regions (CARs) for these genomes at two different levels of resolution: 10 kb and 50 kb (see Methods). Regions falling between CARs are called "breakpoint regions." At the 10-kb resolution, 3356 synteny blocks were identified, and these blocks covered 90.1% of the human genome. At the 50-kb resolution, 1360 blocks were identified for a total coverage of 91.9%.

We applied EMRAE to both the 10-kb and the 50-kb data sets (see Table 1). At the 10-kb resolution, we recovered 1109 ancestral events including 831 reversals, 15 translocations, 237 transpositions, and 26 fusions/fissions (see Fig. 3). The majority of these predicted events were reversals (74.9%), followed by transpositions (21.4%) and with only a limited number of interchromosomal events (3.7%). We found that the proportion of conserved adjacencies that are successfully associated to events (i.e., used) by the prediction algorithm is relatively high (46% ~79%; see Table 1). This indicates that EMRAE recovers a significant portion of the ancestral events.

Figure 4A shows the localization of the reversals recovered on the path from the primate-rodent ancestor to the human. These reversals are human–chimp–rhesus (HCR) specific, human–chimp (HC) specific, or human specific, and we call them "primate reversals" in the following for simplicity. Most of these reversals flipped interstitial regions of the chromosomes with only two HC-specific reversals flipping the centromeres of Chr3 and Chr11. Figure 4B illustrates in more detail another HC-specific reversal together with the UCSC Net tracks (Kent et al. 2002) in that region. From the Net tracks, it is easy to see that in rhesus, mouse, rat, and dog, the synteny block 398 is in opposite orientation to its flanking blocks 397 and 399. This implies that the corresponding region was reversed in the human–chimp lineage. Interestingly, one of the exons of the transcript *AK126351* is embedded in the inverted segment, suggesting a human–chimp innovation. See Supplemental Table 1 for a full listing of genes overlapping the boundaries of reversal events.

Finally, Figure 4C shows the size distribution of the inverted and transposed segments on the leaf edges of the mammalian phylogeny. We observed in particular that, even though we detect 1, 4, 7, 7, 2, and 2 large reversals (>500 kb) in the human, chimp, rhesus, mouse, rat, and dog lineages, respectively, the vast majority of the reversals detected (between 44% and 80%) are <50 kb in size. We note that the length of the inverted and transposed fragments is estimated based on the current

Table 1. Genome rearrangement predictions of EMRAE on the mammalian genome data sets at two different resolution levels (10 kb and 50 kb)

	10-kb predictions								
	EMRAE							MGR	
	Cons Adj	Used Adj (%)	Rev	Tloc	Tran	Fus/Fis	Total events	Total events	Overlap (%)
Human	157	45.9	12	0	4	0	16	51	91.7
HC	419	49.1	29	0	15	0	44	136	100
HCR	590	65	83	0	8	2	93	163	95.3
Chimp	161	68.9	17	0	7	1	25	42	94.4
Rhesus	554	78.8	49	0	40	1	90	174	98
Mouse	750	60.1	90	3	10	5	108	220	99.0
Rat	2390	71.7	227	0	127	3	357	836	98.7
MR	1130	55	140	2	9	0	151	354	95.1
Dog	1226	73.2	184	10	17	14	225	376	95.2
Total	—	—	831	15	237	26	1109	2352	96.8
	50-kb predictions								
	EMRAE							MGR	
	Cons Adj	Used Adj (%)	Rev	Tloc	Tran	Fus + Fis	Total Events	Total Events	Overlap (%)
Human	25	56	2	0	1	0	3	8	100
HC	171	62.6	19	0	4	1	24	54	100
HCR	268	52.2	27	0	5	2	34	77	86.2
Chimp	63	87.3	12	0	1	1	14	15	92.3
Rhesus	173	72.2	22	0	6	1	29	48	95.7
Mouse	243	48.1	25	3	0	5	33	74	93.9
Rat	1249	76.4	128	0	65	5	198	432	96.2
MR	628	29.3	41	2	2	0	45	224	81.4
Dog	500	55.8	46	7	8	13	74	172	90.9
Total	—	—	322	12	92	28	454	1104	92.5

We report the number of conserved adjacencies (Cons Adj); the proportion of the adjacencies successfully used by EMRAE to infer events (Used Adj); the number of predicted reversals, translocations, transpositions, fusions, and fissions; and the total number of events on each edge. The conserved adjacencies on a given edge are the conserved adjacencies in the two sets of genomes partitioned by that particular edge. HC are human–chimp-specific events, HCR are human–chimp–rhesus specific, and MR are mouse–rat specific. The last column is the percentage of the predicted events (excluding transpositions) that are also found in the MGR scenario.

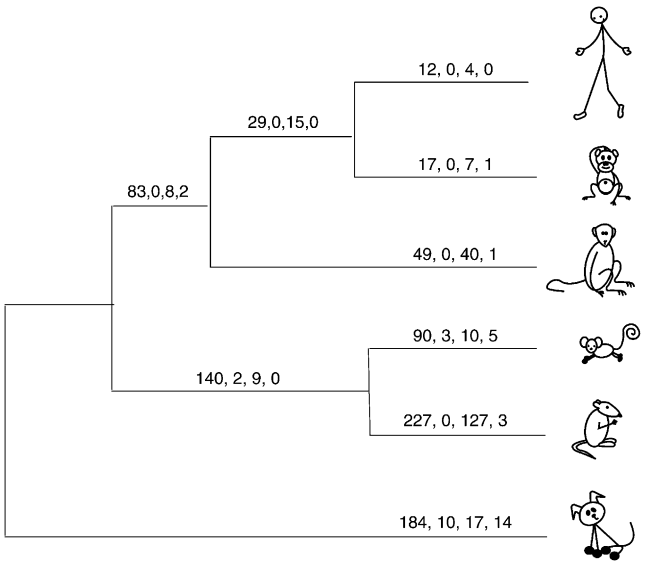

Figure 3. Genome rearrangement events predicted by EMRAE in the mammalian phylogeny of six species (human, chimpanzee, rhesus macaque, mouse, rat, and dog) at a 10-kb resolution. The four numbers on each edge represent the number of predicted reversals, translocations, transpositions, and fusions/fissions, respectively.

arrangement of blocks in the extant genomes (see Supplemental Material).

Robustness and comparison with other reconstruction methods

It is expected that the resolution cutoff will have an impact on the construction of the synteny blocks, and we wanted to evaluate how these changes affected the predictions of EMRAE. As seen in the previous section, many of the predicted events involve segments of moderate size (see Fig. 4C), and thus it is not surprising that the total number of events predicted using the 50-kb resolution is lower than the one at 10 kb (454 vs. 1109; see Table 1). On the human lineage, for instance, although EMRAE predicted 12 reversals at the 10-kb resolution, only two were found in the 50-kb data set. We note, however, that out of the 10 predicted reversals missing from the 50-kb data set, nine are shorter than 50 kb, and one is ~60 kb. Overall, we find that most of the 50-kb events (85.5%) are matching events predicted in the 10-kb data set (see Supplemental Fig. 2). This high overlap confirms that the EMRAE approach is robust to changes in the parameters used to construct synteny blocks.

Next, we were interested in comparing the predictions of EMRAE to the predictions of MGR (Bourque and Pevzner 2002) on the same data sets. Applying MGR to the 10-kb and 50-kb data sets leads to the recovery of evolutionary scenarios with 2352 and 1104 events, respectively. Because MGR does not model transpositions, we will only focus on the predictions of EMRAE for other types of events. As expected from our simulations, we find that the majority of the predictions of EMRAE are included in the MGR scenario for both data sets (96.8% for 10 kb and 92.5% for 50 kb; see Supplemental Fig. 3) with the major difference being that in the case of EMRAE, these highly reliable events are not mixed with more ambiguous events as with MGR. We also note that, at the 10 -kb resolution, 28 of the predictions by EMRAE are not recovered by MGR. This shows that one set of predictions is not simply the subset of the other.

Finally, because primate genomes have similar karyotypes, large-scale primate-specific rearrangement events have been studied extensively in the past. For instance, a recent genome-wide comparison between human and chimp revealed nine centromeric reversals (Newman et al. 2005). We found that of these nine reversals, four are now also predicted by EMRAE (see Supplemental Table 2). This sensitivity level is comparable to the one expected from our simulations (see Fig. 2).

Sequence features associated with mammalian breakpoints and rearrangement events

It is well known that primate-specific breakpoint regions are significantly enriched in segmental duplications (SDs) defined as regions within the same genome that are at least 1 kb and 90% homologous (Samonte and Eichler 2002; Bailey et al. 2004a). It is also believed that SDs might be one of the driving forces that trigger rearrangement events (for review, see Bailey and Eichler 2006). Having for the first time access to an extensive list of high-quality primate-specific rearrangement events, we wanted to explore further the prevalence of this association.

We started by performing a test to measure the enrichment of SDs in the breakpoint regions identified in the present study. On the human lineage there are 157 conserved adjacencies (see Table 1), and 74 of these correspond to adjacencies observed in human and not in any other genome. Using these human-specific breakpoint regions, we showed with simulations that they are significantly enriched in SDs (P-value < 0.001; see Fig. 5A and Methods). Indeed, we observed that 93.2% of the human-specific breakpoint regions (69 out of 74) contain SDs, a property found in only ~60% of size-matched random regions. We further studied the relationship between SDs and breakpoints by looking for the presence of homologous matching pairs of SDs within the same set of human-specific breakpoints. Interestingly, we observed 100 pairs of regions with matching pairs of SDs instead of an average of 25 pairs observed in the random simulated data sets (see Fig. 5B and Methods). This shows that the human-specific breakpoint regions are not only enriched in SDs, but that the enrichment in matching pairs of SDs is even stronger.

Next, we were also interested in testing the association between SDs and the reversals predicted by EMRAE. Specifically, we wanted to assess whether predicted reversals were preferentially associated with supporting pairs of SDs. Although such pairing had been observed previously in a limited number of cases (Kehrer-Sawatzki and Cooper 2008), the prevalence of this phenomenon has not been fully explored. Studying the 12 predicted human-specific reversals, we found that seven of them (58%) are supported by pairs of SDs, while randomly selecting 12 pairs of comparable breakpoint regions would lead to, at most, two random reversals with SD support (see Fig. 5C and Methods).

Extending these analyses to the list of predicted primate reversals, we find that the enrichment is retained with 34 of the 118 (28.9%) primate reversals having SD support (see Fig. 6A and Methods). Interestingly, we also found that the average percent identity of the SDs that are associated with reversals correlates to the relative age of these events (see Fig. 6B). This, combined with the strong enrichment of SD support in the predicted reversals, substantiates the link between SDs and rearrangements events.

Mammalian reversal events are enriched in pairs of LI repeats

The previous analyses were restricted to primate-specific events; we extended this work by performing a BLAST search between all

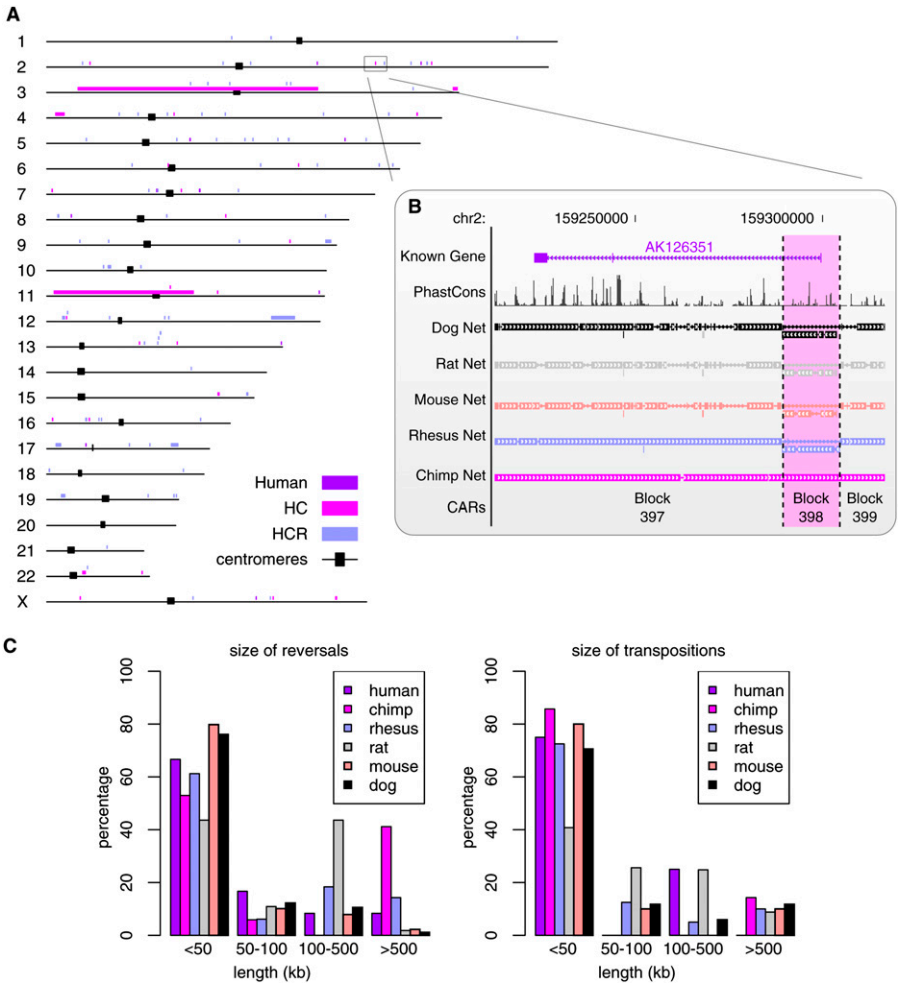

Figure 4. Localization on the human genome of the primate reversals and length distribution of the predicted rearrangement events. (*A*) Reversals recovered on the path from the primate-rodent ancestor to the human. These reversals are human–chimp–rhesus (HCR) specific, human–chimp (HC) specific, or human specific. Some reversals are displayed at different heights if they are too close (e.g., the three small regions in Chr13). (*B*, highlighted in pink) Example of an HC-specific reversal recovered by EMRAE. The reversal is shown on human chromosome 2 along with the UCSC net tracks for the other genomes. Synteny block (or CARs) 398 has opposite orientation in rhesus, mouse, rat, and dog as compared to human and chimp. (*C*) Length distribution of the reversed and of the transposed regions for the events predicted for the extant genomes.

pairs of breakpoints associated with predicted mammalian reversals (see Methods). We found that, similarly to the primate-specific events that are supported by pairs of SDs, many events in other mammalian lineages are also associated with regions of high sequence identity (see Supplemental Fig. 4). More specifically, we found that 58.3%, 29.4%, 24.4%, 42,7%, 47.4%, and 20.6% of the human, chimp, rhesus, rat, mouse, and dog reversals are supported by regions with BLAST scores greater than 1000.

To identify the source of this homology, we restricted our analysis to the 550 reversals with breakpoints defined within 100 kb and assessed the overlap between the regions of high sequence identity and repeats. For primate reversals, this mostly excluded reversals already supported by SDs (see Supplemental Table 3). We annotated each reversal to a particular repeat family when the overlap between the homologous segment identified and a repeat instance was >50% and compared the results to matched simulated data sets (see Fig. 7 and Methods). Interestingly, we found an overrepresentation of L1 repeats across all lineages and in particular in the mouse and rat lineages, where pairs of L1 repeats are found in the breakpoints of 81.5% and 65.3% of the reversals.

Discussion

We presented a new method to infer partial rearrangement scenarios on a given phylogenetic tree that is applicable to both uni- and multichromosomal genomes. In contrast to previous approaches (Moret et al. 2001; Bourque and Pevzner 2002; Ma et al. 2006), we focused on the quality of the rearrangements recovered with the rationale that downstream analyses from these predicted events would improve our understanding of underlying evolutionary mechanisms that are shaping genomes at a global scale. Using simulated data sets, we confirmed that our algorithm, called EMRAE, successfully achieved high specificity without compromising sensitivity in the prediction of ancestral events. In the current implementation, the evolutionary events that we considered were reversals, translocations, transpositions, and fusions/fissions. But this list could easily be extended to take into account other types of events, for instance the "double-cut and join" (DCJ) operation (Yancopoulos et al. 2005), since only the inference rules defined on the list of conserved adjacencies would need to be adjusted.

As a first application, we used EMRAE on a set of six mammalian genomes with good quality assemblies (human, chimpanzee, rhesus macaque, mouse, rat, and dog). We traced back over a thousand ancestral rearrangement events, identified several genes that are likely to have been affected by these events (see Fig. 4B and Supplemental Table 1), and demonstrated the robustness of the approach for different choices of input parameters. We note that the highest number of predicted events was on the rat edge, a phenomenon observed previously that might be the consequence of regions of problematic assembly or of a higher rate of rearrangements in that lineage (Bourque et al. 2004). Overall, for this data set, the vast majority of predicted events were reversals and transpositions of moderate size (see Fig. 4C) with very few predicted translocations (15 in the 10-kb and 12 in the 50-kb data sets; see Table 1). But, because EMRAE only recovers partial scenarios with a preference for events in regions of limited breakpoint reuse, it is also possible that some types of events are harder to recover than others. This suggests that either interchromosomal events are relatively rare or that translocations are more likely to be associated with secondary events than reversals and transpositions.

Studying the sequence features of breakpoint regions, we showed that segmental duplications (SDs) were not only enriched

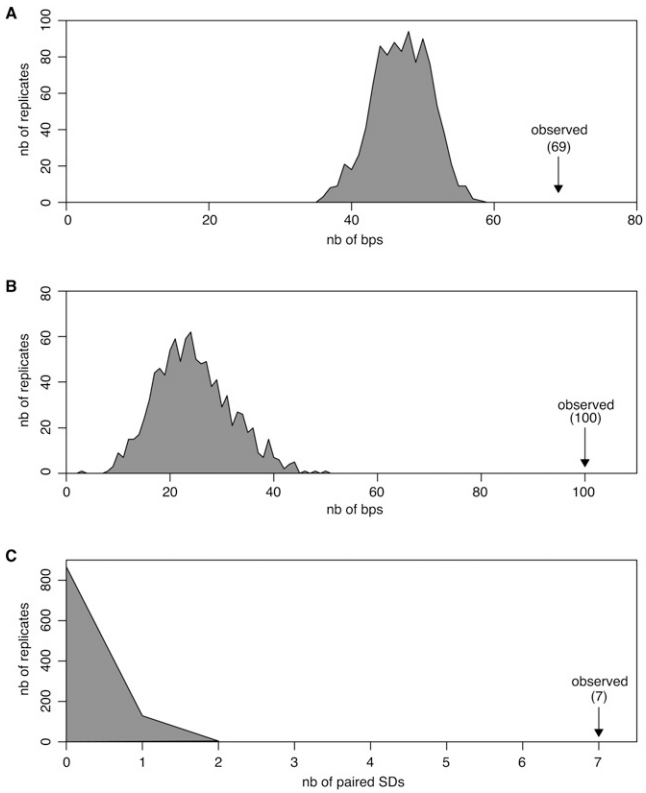

Figure 5. Association between segmental duplications (SDs) and human-specific breakpoint regions. (*A*) Simulations showing the enrichment of SDs in the human-specific breakpoints. (*B*) Simulations showing the enrichment of pairs of SDs in the human-specific breakpoints. (*C*) Simulations showing the enrichment of pairs of SDs associated with human-specific reversals. (Shaded areas) The distribution of values observed in 1000 matched random data sets. nb, number.

in evolutionary breakpoints, but that pairs of SDs were associated with many primate rearrangement events. The fact that the age of these supporting pairs of SDs matched the timing of their associated events (see Fig. 6B) substantiates a direct link between these regions and the rearrangement themselves. Extending the analysis by performing BLAST searches between pairs of breakpoints at the edge of mammalian reversals revealed that a large fraction of the predicted events contained pairs of L1 repeats (see Fig. 7). This is consistent but significantly broadens an observation made in two recent studies that L1 repeats are enriched in regions of structural variation in the human genome (Korbel et al. 2007; Kim et al. 2008). Although *Alu* repeats are very common in the primate breakpoints and have been linked to structural variation (Bailey et al. 2003), we did not observe pairs of *Alu* to be overrepresented in breakpoints associated with primate reversals. This could be because we are focusing on the best BLAST alignment for each pair of regions.

Improving our understanding of the evolutionary forces driving large-scale rearrangement events has been a promise only partially fulfilled by previous phylogenetic reconstruction analyses. We have now shown that focusing on ancestral events can provide new insights into the sequence features associated with these global changes. Ultimately, we should be able to use this knowledge to also feed back into the design of more accurate rearrangement models and scenarios.

Methods

EMRAE for multichromosomal genomes

A chromosome X can be represented by a signed permutation $c_1\ c_2 \ldots c_n$, where each integer c_i corresponds to a synteny block or a contiguous ancestral region. The sign of c_i represents its orientation. We view the chromosome $c_1\ c_2 \ldots c_n$ the same as its reverse $-c_n - c_{n-1} \ldots - c_2 - c_1$. Our main methodological contribution is an extension of EMRAE (Zhao and Bourque 2007) that makes it applicable to multichromosomal genomes and allows recovery of a wider range of rearrangement events, specifically, reversals, translocations, transpositions, and fusions/fissions. These rearrangements are defined as follows:

- A "reversal" $r(i, j)$, where $i \leq j$, transforms X into $c_1\ c_2 \ldots - c_j - c_{j-1} \ldots - c_{i+1} - c_i\ c_{j+1} \ldots c_n$ by inverting both the order of $c_i\ c_{i+1} \ldots c_{j-1}\ c_j$ and the sign of each gene.
- A "translocation" $tloc(i, j, X, Y)$ acts on two chromosomes $X = X_1\ X_2$ and $Y = Y_1\ Y_2$, where $X_1 = x_1\ x_2 \ldots x_{i-1}$, $X_2 = x_i\ x_{i+1} \ldots x_m$, and $Y_1 = y_1\ y_2 \ldots y_{j-1}$, $Y_2 = y_j\ y_{j+1} \ldots y_n$. A translocation $tloc(i, j, X, Y)$ exchanges X_1 and Y_1 and leads to two new chromosomes X' and Y', where $X' = Y_1\ X_2$ and $Y' = X_1\ Y_2$ or exchanges X_1 and $-Y_2$ and leads to two new chromosomes $X'' = -Y_2\ X_2$ and $Y'' = X_1\ -Y_1$.
- A "transposition" $t(i, j, k)$ picks up a segment $c_i \ldots c_j$ of a chromosome and then reinserts it immediately after c_k. If c_k is on the same chromosome ($k > j$ or $k < i$), then the transposition $t(i, j, k)$ is intrachromosomal; otherwise, it is interchromosomal.
- A "fission" breaks a chromosome $X = X_1\ X_2$ and leads to two new ones, X_1 and X_2 (where X_1 and X_2 are nonempty segments). A "fusion" is the opposite of a fission: It connects two chromosomes X_1 and X_2 and leads to a new one, $X_1\ X_2$ or $X_1 - X_2$.

Predicting ancestral events

Given a set of genomes G and their phylogenetic tree T, the idea is that for each edge $e = (A, B)$ on the tree T, we partition the genomes

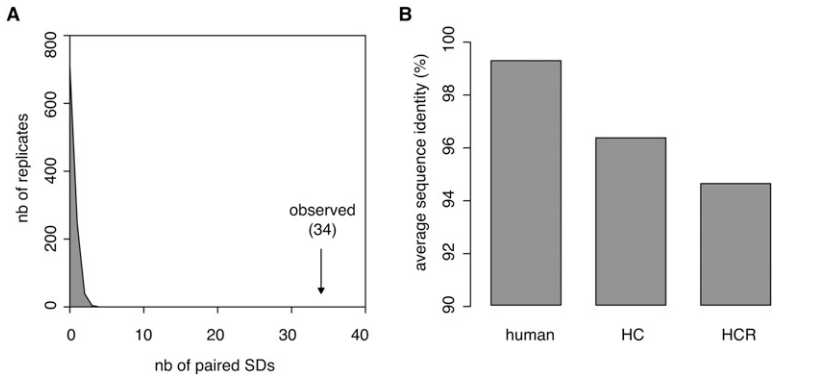

Figure 6. Association between pairs of SDs and primate reversals. (*A*) Simulations showing the enrichment of pairs of SDs within the primate breakpoints; (shaded area) the distribution of values observed in 1000 matched random data sets. (*B*) Average sequence identity of pairs of SDs associated with the primate reversals that are human specific, human–chimp (HC) specific, or human–chimp–rhesus (HCR) specific. nb, number.

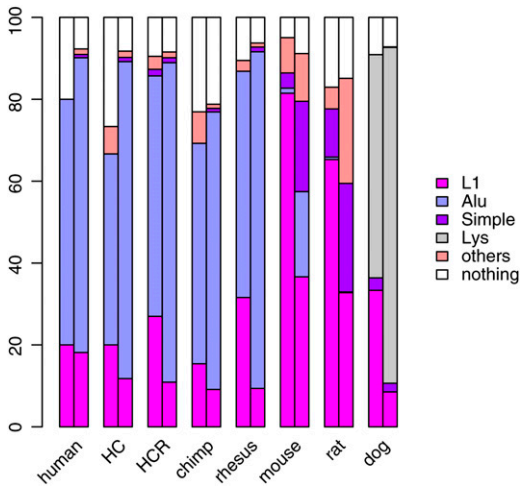

Figure 7. Overrepresentation of alignable pairs of L1 repeats in the breakpoints of mammalian reversals. In each lineage, the first bar represents the proportion of reversals for which the best BLAST alignment overlaps a particular repeat family, and the second bar represents the proportion of reversals that is expected to overlap a particular repeat family based on size-matched random simulations. This analysis is restricted to the 550 reversals with breakpoints defined within 100 kb.

G into two separate subsets S_A and S_B. Formally, denote by CA(e, A) and by CA(e, B) the sets of "conserved adjacencies" in S_A and S_B, respectively, for the edge e. A translocation can be viewed as a reversal if we concatenate the two affected chromosomes in a proper way. Thus, we use a similar rule to infer reversals and translocations. Because a translocation affects two chromosomes, we count the inferred event as a translocation only if the adjacencies used are on different chromosomes in most genomes. We use the following "Inference Rules" (see Supplemental Material for a description of the algorithm):

- *Reversal and translocation*. Suppose we have $a_1 = a(c_{i-1}, c_i)$, $a_2 = a(c_j, c_{j+1})$ in CA(e, A), and $b_1 = a(c_{i-1}, -c_j)$, $b_2 = a(c_i, -c_{j+1})$ in CA(e, B). If the genomes are uni-chromosomal, we infer a reversal $r(i, j)$ from A to B. Otherwise, a_1, a_2 and b_1, b_2 may also result from a translocation $tloc(c_{i-1}, c_i; c_j, c_{j+1})$. If there is at least one genome G_m in S_A and G_n in S_B, such that a_1, a_2 are on the same chromosome of G_m and b_1, b_2 are on the same chromosome of G_n, then we infer a reversal $r(i, j)$. Otherwise, we infer a translocation $tloc(c_{i-1}, c_i; c_j, c_{j+1})$. Similarly, given $a_1 = a(c_{i-1}, c_i)$, $a_2 = a(c_j, c_{j+1})$ in CA(e, A), and $b_1 = a(c_{i-1}, c_{j+1})$, $b_2 = a(c_j, c_i)$ in CA(e, B) for some genomes, we infer a translocation $tloc(c_{i-1}, c_i; c_j, c_{j+1})$, or a reversal that transforms $a_1 = a(c_{i-1}, c_i)$, $a_2 = a(-c_{j+1}, -c_j)$ into $b_1 = a(c_{i-1}, c_{j+1})$, $b_2 = a(-c_i, -c_{j-1})$ based on the same criteria.
- *Transposition*. Suppose we have $a_1 = a(c_{i-1}, c_i)$, $a_2 = a(c_j, c_{j+1})$, $a_3 = a(c_k, c_{k+1})$ in CA(e, A), and $b_1 = a(c_{i-1}, c_{j+1})$, $b_2 = (c_k, c_i)$, $b_3 = (c_j, c_{k+1})$ in CA(e, B). If the genomes are uni-chromosomal, then we infer a transposition $t(i, j, k)$ from A to B. Otherwise, suppose a_1 and a_2 appear in m genomes in S_A; then we infer a transposition $t(i, j, k)$ if on at least $m/2$ of the genomes, the four genes of a_1 and a_2 are on the same chromosome, or if on at least $m/2$ of the genomes, the four genes of a_2 and a_3 are on the same chromosome. This condition is applicable to predict interchromosomal transpositions. It suggests that the transposed segment must be from the same chromosome and makes sure that we recover a true transposition.
- *Fusion/fission*. Suppose we have $a = a(c_i, c_j)$ in CA(e, A); if for each genome G_k in S_B, a is sign-compatible, then we infer a fission that breaks $a = a(c_i, c_j)$. A fusion from A to B can be viewed as a fission from B to A.

We note that for each event predicted by EMRAE, the method only identifies the adjacencies associated with this event. This implies, for instance, that if EMRAE recovers a reversal on a given edge e, it does not predict the precise content of the region that is flipped (see Supplemental Material and Supplemental Table 4).

Refinement step to identify events involving limited breakpoint reuse

If there is no breakpoint reuse, it will be straightforward to identify all the conserved adjacencies and to recover a rearrangement history using the inference rules described above. However, breakpoint reuse is common (Pevzner and Tesler 2003), and when it occurs, the affected conserved adjacencies could be missing from the correct edges and slide to wrong places. For this reason, we have designed a refinement step (Zhao and Bourque 2007) to detect potentially sliding adjacencies and associate them to the correct edges. The simulations showed that this refinement step can partially recover such adjacencies and help to infer events that are associated with breakpoint reuse (Zhao and Bourque 2007). We have now adapted this refinement step to also be applicable to multichromosomal genomes.

Simulated data sets

In our experiments, we generated a random rooted tree T and assigned a random number k to each edge, where k is in [1, 2*μ] and the evolutionary rate μ is the average number of events per edge of a tree T. In our simulations, the random trees have seven multichromosome genomes, each genome has 100 blocks, and we varied μ from 2.5 to 10. We evolved the tree starting from the root by performing k random events to each edge until we get the block orders of all the leaf genomes. In the process, we stored all the events that are performed and recorded the "true" evolutionary scenario. Finally, we removed the root and took the permutations of the leaf genomes and the tree topology as the input for EMRAE and MGR. For each choice of parameter μ, we repeated the experiment 100 times and computed the average sensitivity and specificity of the two methods.

For the Rev+Tloc model, we tested the ability of EMRAE and MGR to recover reversals and translocations only; in the All events model, we tested their performance in inferring reversals, transpositions, translocations, and fusions/fissions. Although MGR reconstructs trees and ancestral genomes, it does not directly provide a detailed rearrangement scenario as part of the output. As a surrogate, we used GRIMM (Tesler 2002), which relies on the same set of operations, to produce a most parsimonious scenario on each edge of the trees recovered by MGR. Note that because transpositions are not directly considered in MGR, we used three consecutive reversals to mimic intrachromosomal transpositions as in Zhao and Bourque (2007). Similarly, we used two consecutive translocations to mimic interchromosomal transpositions (see Supplemental Material).

To estimate biologically realistic ratios in the All events model, we looked at the number of events predicted between six pairs of genomes in the mammalian data set at the 10-kb resolution. These pairs of genomes were selected to cover all the branches of the tree, and the results are shown in Supplemental Table 5. These results substantiate that in the mammalian data set, reversals are overrepresented and fusions/fissions are underrepresented. Based on this, we have used ratios 10:2:2:0.1 in the simulations when selecting randomly between reversals, transpositions, translocations, and fusions/fissions.

Mammalian genome data set

The six mammalian genome assemblies used are human (UCSC build hg18), chimpanzee (panTro2), rhesus macaque (rheMac2), mouse (mm9), rat (rn4), and dog (canFam2). We used the program described in Ma et al. (2006) to construct the contiguous ancestral regions for these six genomes. Starting from the two-way "nets" (Kent et al. 2002; Schwartz et al. 2003) between human and the other genomes, we built the six-way shared orthologous blocks. Next, we discarded blocks smaller than a specific cutoff (either 10 kb or 50 kb) and merged consecutive orthologous blocks that had the same order and orientation in all the six genomes. For the breakpoint content analyses, we extended small breakpoint regions (<10 kb) by 10 kb on each side.

Enrichment of SDs and paired SDs in the breakpoint regions

To detect SDs, we used the UCSC human self-chains, which are BLASTZ alignments of the human genome to itself (Kent et al. 2002; Schwartz et al. 2003). A chained alignment is allowed to have much larger gaps than traditional alignments. In our analysis, we use alignments larger than 1 kb, and we split the human chains into traditional alignments by removing gaps larger than 300 bp. See Supplemental Figure 5 for an example of splitting chains.

For the first analysis, we generated 1000 sets of random regions and counted the number of regions containing SDs. Here each set of random regions consisted of 74 regions having the same length distribution as the human-specific breakpoints. For the second analysis, we used the same random data sets and counted the number of pairs of SDs. Here a paired SD was a segment (>1 kb) in one random region that was aligned with a segment (>1 kb) in another region from the same random set. Finally, for the third analysis, we kept the first breakpoint region of each human-specific reversal intact, we generated 1000 random regions of the size of the second breakpoint region, and we calculated the number of these random pairs that were supported by a paired SD. Similar simulations were performed for the 118 primate reversals for which the corresponding breakpoints in human could be unambiguously identified (see Supplemental Material).

Pairwise BLAST analysis of breakpoint regions associated with predicted events

To analyze the homology between pairs of breakpoints associated with rearrangements in other mammalian lineages, we used BLAST (Altschul et al. 1990) since few of the self-chains of other genomes are available from the UCSC genome website. For each pair of breakpoints associated with a rearrangement, we recorded the best BLAST score under standard parameter settings. On each edge, we extracted the k reversals that had breakpoints defined within 100 kb (see Supplemental Table 3) and generated 1000 simulated data sets, where each set consisted of k pairs of size-matched random regions on the corresponding genome. Then, we used BLAST to identify the homologous regions between each pair of random regions and detect the possible pairs of repeats supporting them. Formally, a pair of breakpoint regions is said to be supported by a particular repeat if more than 50% of both homologous regions are covered by a repeat instance from the same family.

Acknowledgments

We thank Jian Ma for help with the CARs and the gene order data and the reviewers for helpful recommendations. This work is supported by funds from the Biomedical Research Council (BMRC) of Singapore.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
Bailey, J.A. and Eichler, E.E. 2006. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7:** 552–564.
Bailey, J.A., Liu, G., and Eichler, E.E. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73:** 823–834.
Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. 2004a. Hot spots of mammalian chromosomal evolution. *Genome Biol.* **5:** R23. http://genomebiology.com/2004/5/4/R23.
Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M., and Eichler, E.E. 2004b. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14:** 789–801.
Blanchette, M., Kunisawa, T., and Sankoff, D. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* **49:** 193–203.
Bourque, G. and Pevzner, P.A. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.* **12:** 26–36.
Bourque, G., Pevzner, P.A., and Tesler, G. 2004. Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Res.* **14:** 507–516.
Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P.A., and Tesler, G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15:** 98–110.
Bourque, G., Tesler, G., and Pevzner, P.A. 2006. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res.* **16:** 311–313.
Cosner, M.E., Jansen, R.K., Moret, B.M., Raubeson, L.A., Wang, L.S., Warnow, T., and Wyman, S. 2000. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 104–115.
Darling, A.E., Miklos, I., and Ragan, M.A. 2008. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* **4:** e1000128. doi: 10.1371/journal.pgen.1000128.
Froenicke, L., Caldes, M.G., Graphodatsky, A., Muller, S., Lyons, L.A., Robinson, T.J., Volleth, M., Yang, F., and Wienberg, J. 2006. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res.* **16:** 306–310.
Hannenhalli, S., Chappey, C., Koonin, E.V., and Pevzner, P.A. 1995. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics* **30:** 299–311.
Kehrer-Sawatzki, H. and Cooper, D.N. 2008. Molecular mechanisms of the chromosomal rearrangement. *Chromosome Res.* **16:** 41–56.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.
Kim, P.M., Lam, H.Y., Urban, A.E., Korbel, J.O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M.B. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* **18:** 1865–1874.
Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.
Larget, B., Simon, D.L., Kadane, J.B., and Sweet, D. 2005. A Bayesian analysis of metazoan mitochondrial genome arrangements. *Mol. Biol. Evol.* **22:** 486–495.
Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., and Miller, W. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **16:** 1557–1565.
Miklos, I. 2003. MCMC genome rearrangement. *Bioinformatics* (Suppl. 2) **19:** ii130–ii137.
Moret, B.M., Wyman, S., Bader, D.A., Warnow, T., and Yan, M. 2001. A new implementation and detailed study of breakpoint analysis. *Pac. Symp. Biocomput.* **2001:** 583–594.
Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309:** 613–617.
Newman, T.L., Tuzun, E., Morrison, V.A., Hayden, K.E., Ventura, M., McGrath, S.D., Rocchi, M., and Eichler, E.E. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15:** 1344–1356.
Pevzner, P. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci.* **100:** 7672–7677.

Rocchi, M., Archidiacono, N., and Stanyon, R. 2006. Ancestral genomes reconstruction: An integrated, multi-disciplinary approach is needed. *Genome Res.* **16:** 1441–1444.

Samonte, R.V. and Eichler, E.E. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3:** 65–72.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., and Cedergren, R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci.* **89:** 6575–6579.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Tesler, G. 2002. GRIMM: Genome rearrangements web server. *Bioinformatics* **18:** 492–493.

Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21:** 3340–3346.

Zhao, H. and Bourque, G. 2007. Recovering true rearrangement events on phylogenetic trees. In *Comparative Genomics, RECOMB 2007 International Workshop, RECOMB-CG 2007* (eds. G. Tesler and D. Durand), pp. 149–161. Springer, San Diego, CA.

Received September 1, 2008; accepted in revised form January 13, 2009.

Recovering genome rearrangements in the mammalian phylogeny

Hao Zhao and Guillaume Bourque

Genome Res. 2009 19: 934-942
Access the most recent version at doi:10.1101/gr.086009.108