Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 86 (2005) 692 – 700
GENOMICS
www.elsevier.com/locate/ygeno

Murine segmental duplications are hot spots for chromosome and gene evolution

Lluís Armengol [a], Tomàs Marquès-Bonet [b], Joseph Cheung [c], Razi Khaja [c], Juan R. González [a], Stephen W. Scherer [c], Arcadi Navarro [b], Xavier Estivill [a,*]

[a] *Genes and Disease Program, Center for Genomic Regulation, Passeig Marítim 37–49, 08003 Barcelona, Catalonia, Spain*
[b] *Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain*
[c] *Program in Genetics and Genomic Biology, The Hospital for Sick Children, M5G 1X8 Toronto, ON, Canada*

Received 12 June 2005; accepted 23 August 2005
Available online 26 October 2005

Abstract

Mouse and rat genomic sequences permit us to obtain a global view of evolutionary rearrangements that have occurred between the two species and to define hallmarks that might underlie these events. We present a comparative study of the sequence assemblies of mouse and rat genomes and report an enrichment of rodent-specific segmental duplications in regions where synteny is not preserved. We show that segmental duplications present higher rates of molecular evolution and that genes in rearranged regions have evolved faster than those located elsewhere. Previous studies have shown that synteny breakpoints between the mouse and the human genomes are enriched in human segmental duplications, suggesting a causative connection between such structures and evolutionary rearrangements. Our work provides further evidence to support the role of segmental duplications in chromosomal rearrangements in the evolution of the architecture of mammalian chromosomes and in the speciation processes that separate the mouse and the rat.

Keywords: Chromosomal evolution; Mouse genome; Rat genome; Segmental duplications; Synteny breakpoint; Molecular evolution; Evolutionary breakpoints

Although the *random-breakage model* [1,2] has largely been accepted as the paradigm for chromosomal evolution, data from the study of newly available genomic sequences and the possibility of performing multispecies comparisons of genomes question this theory. For instance, clustering of evolutionary breakpoints that result in a large number of small syntenic blocks in certain genomic regions is a major argument in favor of the *fragile-breakage model* [3–7]. The fragile-breakage theory states that evolutionary breakpoints would not be randomly distributed throughout the genomes but would accumulate into relatively short fragile regions [3]. Nevertheless, some authors have proven that the available sequence data do not support a model in which only a discrete collection of hot spots is responsible for the rearrangement breakpoints [8]. The fragile-breakage theory is also supported by the observed recurrence of human chromosomal rearrangements that are the cause of several disorders [9,10] and the existence of fragile sites in the genomes [11–13]. So far, the nature and composition of such fragile sites in mammals, as well as the relationship between evolutionary and disorder-causing breakpoints, remain unclear although several studies have attempted to identify sequence elements involved in such rearrangements [12,14–16].

Previous studies have shown that regions where evolutionary chromosomal rearrangements have occurred (also called breaks of synteny and abbreviated BOS) between mouse and human are significantly enriched in primate-specific segmental duplications (SDs) [17,18]. Although SDs might not necessarily be the cause of such evolutionary rearrangements, it is tempting to speculate about a putative role for these low-copy repeat sequences in the evolution and plasticity of genomes, in much the same manner in which they trigger rearrangements in genomic disorders [10,19,20]. Indeed, data from studies in *Drosophila* show that repetitive elements have generated rearrangements separating different species [21]. In addition to the presence of low-copy repeats, other repeat sequences have been found to be present in regions surrounding evolutionary breakpoints [14,22,23]. An unusual composition

* Corresponding author. Fax: +34 93 224 0899.
E-mail address: xavier.estivill@crg.es (X. Estivill).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.08.008

of repeats in regions where evolutionary rearrangements have taken place might provide clues to a better understanding of the molecular mechanisms by which these events occur as well as point to putatively responsible sequences.

Chromosomal rearrangements are also thought to have a role in speciation, acting as genetic barriers to gene flow and thus increasing the time of divergence of genes linked to them. Previous studies have reported an association between rates of chromosomal rearrangement and genic evolution [24,25]. The issue, however, is far from settled since contradictory evidence has also been reported [26,27] and therefore, alternative hypotheses must be examined [28]. For example, genes within segmental duplications present higher rates of sequence and gene-expression divergence than single-copy genes [29,30] which, given their association with rearrangements [17,18], might help to explain the association between chromosomal rearrangements and higher evolutionary rates.

Current drafts of genomic sequences from mouse [31] and rat [32] are an invaluable resource for a detailed sequence-level study seeking to unravel the features involved in evolutionary chromosomal breakpoints between these two closely related species. We present here a comparative study of the sequences of these two organisms in which we identify synteny blocks caused by large-scale rearrangements, study the nature and composition of regions where synteny is not preserved, and analyze the genomic distribution of evolutionary rates.

Results

Identification of synteny blocks

For the identification of synteny blocks between the mouse and the rat genomes we used the publicly available alignments between mouse and rat genomic sequences obtained from UCSC Genome Bioinformatics Group (http://www.genome.ucsc.edu).

We started from a set of over 1.2 million genomic sequence anchors that were connected to give a total of 4117 synteny segments of length >25 kb. These segments were further grouped into 102 synteny blocks with a length of over 250 kb shared by the two species (see Materials and methods) and with an average size of 23.9 Mb in the mouse and 25.6 Mb in the rat genome (see Supplementary Table 1).

The random-breakage model of chromosome evolution [1,33] predicts that the length of syntenic segments approximates an exponential distribution with density function $f(x) = (1/L)^{-x/L}$, where L is the average length of all syntenic segments. In concordance with previous synteny analyses using older assemblies of the mouse and rat genomes [4,5], the lengths of the synteny segments we obtained from our study were not in agreement with the distribution predicted by the random-breakage model, even when we centered the study on large synteny segments (Fig. 1). We observed an enrichment of small segments (<5 Mb, $p = 6.57 \times 10^{-6}$), which would support the fragile-breakage model [3] and an increased frequency of some long fragments (Fig. 1).

Following $N_b = N_{sb} - N_c$ (where N_b is the number of breakpoints, N_{sb} is the number of synteny blocks, and N_c is the number of chromosomes) [34], 82 evolutionary breakpoints were identified in the mouse genome and 81 in the rat genome. Two synteny blocks in the second genome flank each breakpoint in the first, so we distinguish between multichromosomal (when the synteny blocks in the second genome correspond to different chromosomes) and unichromosomal breakpoints. In both genomes, unichromosomal breakpoints occur more than twice as often as multichromosomal breakpoints (data not shown). The lengths of synteny breakpoints range from hundreds of base pairs to millions of base pairs and were found to span around 4–5% of each genome (see Supplementary Table 1).

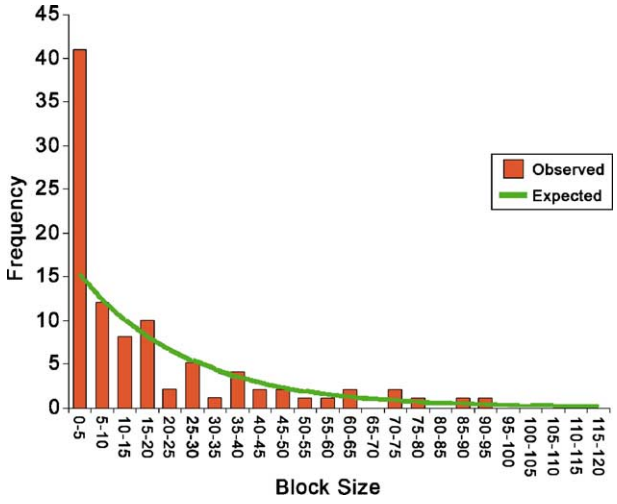

Fig. 1. Distribution of mouse–rat synteny block lengths. Frequency histogram of the lengths of the 102 synteny blocks observed in our analysis fitted with the distribution of expected fragment lengths in a random distribution. The observed data do not fit well the curve predicted by the density function describing the random-breakage model of chromosomal evolution, especially because an enrichment of small fragments (<5 Mb) is observed, together with an enrichment of some larger segments.

Segmental duplications correlate with regions of BOS

We previously identified all large and recent rodent-specific SDs (>90% sequence identity, >5 kb in length) corresponding to mm5 mouse and rn3 rat UCSC assemblies as described in [35]. Data are publicly available at http://www.projects.tcag.ca/xenodup. To obtain a visual overview of the synteny segments, the BOS, and the regions containing SDs, we drew dot plots of the shared synteny blocks between the two genomes and superimposed coordinates of SDs of each genome. We observed that duplicons were present in a large number of the regions where the synteny was lost between the two species (Fig. 2). Using coordinates of both SDs and synteny blocks, we performed a more detailed analysis.

By simply counting, we found an average of 13 SDs per megabase in syntenic regions of the mouse and rat genomes and, in contrast, we counted 27 SDs on average per megabase in regions occupied by synteny breakpoints (Table 1). Due to the known clustering of SDs in relatively short chromosomal regions in the two genomes [35,36] and to avoid bias produced by this fact, we decided to simplify our approach and verify the presence or absence of SDs in these regions. We identified SDs in 49 (60%) of 82 breakpoints in conserved synteny in the mouse

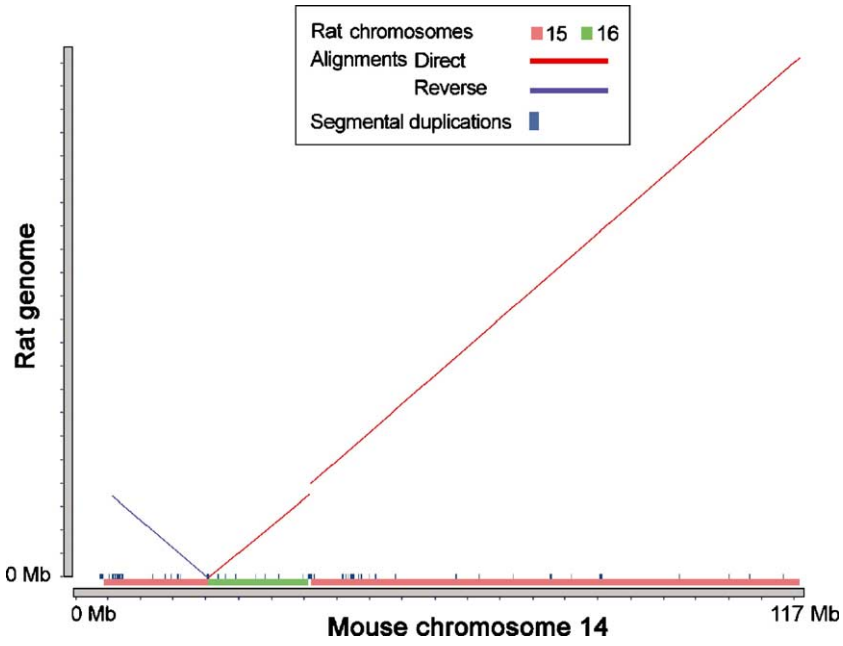

Fig. 2. Segmental duplications correlate with mouse/rat breaks of synteny. Dot-plot representations of alignments between mouse chromosome 14 and the rat genome. Direct and reverse alignments appear as red and blue lines, respectively. On the *x* axis, information on the corresponding rat chromosomes is depicted according to the color code in the legend. Positions of SDs in the mouse genome are represented as bluish rectangles in the *x* axis. In this image, the correlation between synteny breaks and SDs in the mouse chromosome is observed. A complete set of dot plots can be obtained on demand.

genome and in 35 (43%) of 81 in the rat genome (Table 2). No SDs were present in breakpoints from mouse chromosomes 12, 15, and 19 or rat chromosome 6.

To measure the significance of this association and exclude the possibility that our results were incidental, we performed a computer simulation in which the positions of the synteny breakpoints were randomly assigned but their size and the positions of the SDs were kept constant. We then evaluated the presence of SDs in the randomly located breaks of synteny. Comparing these results with our own, we concluded that the amount of SDs in the BOS regions is significantly higher than the expected in a random distribution of evolutionary breakpoints for those mouse chromosomes in which SDs were found, except for chromosome X, and for all rat chromosomes, except chromosomes 14 and 17 (Table 2). We, therefore, conclude that the association of SDs with the synteny breakpoints is not due to chance.

Two breakpoints flank each synteny block, except those that contain the telomeres. To refine our study, we looked for SDs in the 50 kb flanking these breakpoints. We found SDs in 55 of 164 regions explored in the mouse genome and in 58 of 162 in the rat genome, which corresponds to ~35% of regions flanking breakpoints containing SDs in both genomes (Table 2). The number of breakpoint-flanking regions containing SDs was found to be significantly higher than expected for all mouse chromosomes except 12, 16, and X, compared to a random distribution of synteny breakpoints. The same observation was made for rat chromosomes, with the exception of chromosomes 3 and X. Interestingly, mouse chromosomes 15 and 19, and rat chromosome 14, which did not contain SDs within the synteny breakpoint regions (see above and Table 2), were found to contain more SDs than expected in the breakpoint-flanking regions.

Repeat and GC composition of synteny breakpoint regions

We analyzed the GC and repeat content in breakpoint regions to verify whether there are sequence features that could facilitate rearrangements in the rodent lineages.

Table 1
Segmental duplications in block and in BOS regions in mouse and rat

	Block regions			BOS regions		
	Size (Mb)[a]	Number of SDs[b]	Density (SDs/Mb)	Size (Mb)	Number of SDs	Density (SDs/Mb)
Mm[c]	2442.47	32089	13.14	66.60	1844	27.69
Rn[d]	2613.30	33027	12.64	86.57	2392	27.63

Synteny break regions correspond to genomic regions where the synteny criteria are not met.
[a] Megabases.
[b] Segmental duplications.
[c] *Mus musculus*.
[d] *Rattus norvegicus*.

Table 2
Segmental duplications in breaks of synteny (BOS) and breakpoints in mouse and rat

Chromosome	BOS with SD/total[a]	*p* value[b]	BP with SD/total[c]	*p* value
Mus musculus				
1	5/5	0.005	4/10	0.003
2	2/2	0.003	3/4	0.006
3	1/3	0.023	2/6	0.015
5	7/8	0.003	6/16	0.001
8	4/4	0.018	4/8	0.015
10	7/15	<0.001	9/30	0.009
11	1/1	0.021	1/2	0.007
12	0/7	0.999	0/14	0.999
13	2/5	0.010	3/10	0.030
14	2/2	0.007	3/4	0.001
15	0/1	0.999	2/2	0.022
16	2/3	<0.001	1/6	0.100
17	10/13	<0.001	11/26	0.018
18	2/4	0.003	2/8	0.006
19	0/2	0.999	1/4	0.049
X	4/7	0.240	3/14	0.378
Total	49/82		55/164	
Rattus norvegicus				
1	4/11	0.036	9/22	0.047
2	3/7	0.003	7/14	0.023
4	2/3	0.010	3/6	0.013
5	1/1	0.022	1/2	0.013
6	0/10	0.999	8/20	0.003
7	2/5	0.028	2/10	0.034
9	2/5	0.033	4/10	0.023
10	2/3	0.023	2/6	0.013
11	1/2	0.007	2/4	0.005
12	3/3	0.015	3/6	0.021
13	1/2	0.030	1/4	0.015
14	1/2	0.096	1/4	0.043
15	1/1	0.028	1/2	0.038
16	1/2	0.045	1/4	0.041
17	1/5	0.216	3/10	0.199
18	1/2	0.018	1/4	0.019
19	1/1	0.016	1/2	0.042
20	3/9	0.003	5/18	0.022
X	5/7	<0.001	3/14	0.197
Total	35/81		58/162	

Only chromosomes that have synteny breaks are shown.
[a] Number of BOS containing segmental duplications (SDs)/total number of BOS regions.
[b] Permutation *p* value.
[c] Number of breakpoint regions containing segmental duplications/total number of breakpoints.

As a first approach, we generated graphics containing chromosomal representations of synteny blocks together with density plots of GCs and repeats, looking for a consistent pattern that could correlate these elements (data not shown). *De visu*, we did not observe any consistent pattern of increased repeat or GC content in the breakpoint regions or within 50 kb surrounding them (see above). Total repeat content in breakpoint regions ranged between 37 and 71% in the mouse and between 31 and 59% in the rat. Since the amount of different types of repeats varies among the different chromosomes, instead of comparing with the genome average we compared the observed amounts with the expected in a random distribution of synteny breakpoints (Supplementary Table 2). A few breakpoint regions contained increased amounts of overall repeat content compared to the simulations, which was attributable to different types of repeats in different chromosomes. Nevertheless, this increase in repeat content was not observed in the majority of chromosomes nor was it specific to a type of repeat. Furthermore, we could not decipher any pattern that is followed by a majority of the break of synteny regions. Finally, no abnormal GC composition was observed for the breakpoint regions in any chromosome (Supplementary Table 2).

Gaps in synteny breakpoints

The generation of rat and, especially, mouse genome sequences involved a lot of shotgun sequencing. It is known that this methodology is prone to cause misassemblies due to the presence of repeat sequences [36,37]. On the other hand, SDs are also known to lead to misassemblies and gaps in the sequences [18,38–40], and the inability to map them unambiguously to an orthologous position might also lead to synteny gaps. To discard the possibility that sequence gaps were confounding our analyses, we used restrictive synteny criteria (see Materials and methods) and tested whether gaps were present in synteny breakpoints and if this presence was significantly higher than the expected if evolutionary breakpoints were randomly distributed in the chromosomes. Due to the huge amount of gaps present in both genomes, most synteny breakpoints were found to contain gaps. For all chromosomes, except for the mouse X chromosome, we report that the presence of sequence gaps was not significantly higher than the expected in a random distribution of breakpoints (Supplementary Table 3).

Genes at regions of BOS

In the mouse, 16,725 genes were found to be located in syntenic fragments (2442.47 Mb in size) and 654 in nonsyntenic regions (66.60 Mb). This means 6.8 genes/Mb in syntenic regions and 8.6 genes/Mb in BOS regions. For the rat genome, we found that 6573 genes were in syntenic regions (2613.30 Mb in size), while 203 were located in breakpoint regions (86.57 Mb). Overall, 1.59 rat genes/Mb were found in syntenic regions and 1.87 genes/Mb in nonsyntenic regions. Although the RefSeq gene sets are still incomplete (especially for the rat) and may not reflect the total number of genes, with the available data we conclude that both syntenic and nonsyntenic regions have similar amounts of genes.

We used the Gene Ontology Tree Machine (http://www.genereg.ornl.gov/gotm/) to obtain a comparison of the functional profile of genes located in break of synteny regions with the genome average. We found enrichment of genes corresponding to different GO categories; including genes related to pheromone biology and sensory organ development (Supplementary Table 4). Interestingly, it is

known that these types of genes are implicated in biological adaptation and speciation processes.

Genomic distribution of rates of molecular evolution

We first examined the possibility of different evolutionary rates in different chromosomes and found that they are clearly heterogeneous (Kruskal–Wallis, $p < 0.001$). The potential causes of these differences are multiple. First, as previously shown in other species [41,42], the X chromosome presents lower divergence than the average for autosomes (Table 3, dS = 0.1581 vs 0.1981, permutation test, $p < 0.001$). We, therefore, removed sex chromosomes from subsequent analyses. A second potential cause of chromosomal heterogeneity is linked to telomeres, which have also been shown to be associated with factors affecting evolutionary rates such as either higher or lower recombination rates or higher GC content [25]. In the current dataset, genes located in telomeres (within 3 Mb of any end of the chromosome) showed lower synonymous divergence than genes elsewhere in the genome and higher GC content (Table 3, dS = 0.1841 vs 0.1991, $p < 0.001$; GC 45.94 vs 46.40, $p < 0.05$). Therefore, these genes were excluded from further analysis, producing a dataset of 12,139 genes with average evolutionary rates of dN = 0.0331, dS = 0.1991, and dN/dS = 0.1690.

To test whether the reported acceleration in rates of evolution in SDs in other species [29,30,43,44] can also be detected between rat and mouse, we compared evolutionary rates of genes involved in SDs with genes that are not in SDs, regardless of their chromosomal position. Genes in SDs present significantly higher synonymous and nonsynonymous rates of substitution than single-copy genes. Interestingly, they also present higher rates of protein evolution, as indicated by their significantly higher dN/dS ratio (Table 3).

To test for the effects of rearrangements on rates of evolution we excluded all genes involved in SDs and compared all genes in regions of break of synteny to all genes in syntenic regions. Genes located in no-block regions (regions where synteny between mouse and rat cannot be reconstructed) were not found to evolve faster than genes in syntenic regions (Table 3). In fact, the dN/dS ratio is marginally significant but in the opposite direction. We decided to remove these genes, for which synteny could not be defined, from further analysis, producing a final dataset of 11,364 genes. With this curated dataset, we compared genes located inside inversions with those located outside inversions. We found that genes within inversions present significantly higher synonymous and nonsynonymous rates of substitution (Table 3). In addition to the regions within or outside inversions, it is also possible to study genes in regions surrounding any BOS corresponding to inversions and translocations. We compared genes within 2.5 Mb of the breakpoints of such rearrangements with genes located in colinear regions (zones beyond 2.5 Mb from any breakpoint) and found that genes in these regions present a statistically significant increase in dS (Table 3).

Discussion

Two decades ago, Nadeau and Taylor proposed the random-breakage model of chromosomal evolution based on statistical arguments and the synteny data between human and mouse available at the time [1]. With the availability of genome sequence data for several mammalian genomes, analyses that are more detailed can now be performed to examine chromosome evolution and dynamics at the DNA sequence level. Given the resolution of our study, the inability to fit the lengths of the observed synteny segments with the expected ones in the random-breakage model suggests that this theory

Table 3
Evolutionary rates of genes in relation to SDs and evolutionary rearrangements

	N[a]	dN		dS		dN/dS	
		Mean	SE	Mean	SE	Mean	SE
Genes within SDs[b]	322	0.0578	0.0032	0.2120	0.0036	0.2622	0.0138
Genes not located in SDs	11,817	0.0324	0.0003	0.1988	0.0006	0.1665	0.0017
p value[c]		<0.001		<0.001		<0.001	
Genes in no-blocks	256	0.0295	0.0020	0.2046	0.0042	0.1444	0.0009
Genes in synteny blocks	11,364	0.0324	0.0003	0.1986	0.0006	0.167	0.0017
p value[c]		0.174		0.134		0.05	
Inside inversions	2,138	0.0343	0.0008	0.2122	0.0014	0.1669	0.0040
Outside inversions	9,226	0.0318	0.0004	0.1956	0.0007	0.1658	0.0019
p value[c]		0.002		<0.001		0.795	
<2.5 translocation breakpoint	506	0.0329	0.0014	0.2054	0.0030	0.1602	0.0064
>2.5 any breakpoint	10,203	0.0325	0.0003	0.1977	0.0006	0.1682	0.0018
p value[c]		0.805		0.011		0.337	
<2.5 inversion breakpoint	546	0.0310	0.0012	0.2046	0.0025	0.1551	0.0062
>2.5 any breakpoint	10,203	0.0325	0.0003	0.1977	0.0006	0.1682	0.0018
p value[c]		0.316		0.014		0.100	

Averages of evolutionary rates for different categories of rearrangements are shown.
[a] N, number of genes.
[b] SDs, segmental duplications.
[c] Permutation p value comparing the averages for each category of genes.

may not be the most appropriate to describe the occurrence of the evolutionary breakpoints. This observation is in agreement with previous reports on synteny using sequences from different organisms and older assembly versions of the mouse genome [4,5]. We observed an enrichment of small syntenic segments (<5 Mb, $p = 6.57 \times 10^{-6}$) and some long syntenic segments. Bearing in mind the observation of a significant enrichment of SDs in regions that coincide with synteny breakpoints, one could speculate about a connection between small synteny regions and the clustering of SDs in several regions of these two genomes [36,38]. The short synteny segments identified in our study could be attributable to the clustering of breakpoints in relatively short fragile regions, as proposed by the fragile-breakage model, while the long ones are likely to be attributable to the short time of divergence since the mouse/rat common ancestor. Clustering of SDs in discrete genomic regions would lead to a number of synteny blocks undetectable at the resolution of this study. The higher number of intrachromosomal SDs in both genomes could also be an explanation for a higher occurrence of evolutionary inversions, which resulted in the twofold higher amount of intrachromosomal evolutionary breakpoints observed.

Since we focused on synteny segments longer than 250 kb, our study did not have the potential to detect all synteny breakpoints between the mouse and the rat. Using the current mouse and rat genome assemblies, there are several factors that could interfere with the identification of the exact positions of synteny block boundaries: (i) the existence of unfinished regions (sequence gaps) in both genome sequences, (ii) the presence of SDs creating gaps and confounding the correct genome assembly [18,36], and (iii) the presence of large clusters of masked repeats. To override the possibility that local assembly errors interfere with our analysis, we used relatively conservative criteria to define synteny segments (see Materials and methods). The possibility that misassemblies and gaps were confounding our results (due to the presence of SDs in breakpoint regions) was excluded since these regions are not significantly different in terms of presence of sequence gaps compared to random regions chosen from both genomes.

Different types of repeat sequences are thought to play a role in chromosomal rearrangements in mammalian genomes [14,22,45,46], as well as in other eukaryotic organisms [47–49]. In our study, some break of synteny regions were found to be significantly different from the rest of the genome regarding repeat content although no differences were found regarding GC content or gene density.

Comparisons between human and mouse revealed that primate-specific SDs are significantly enriched in regions where evolutionary chromosomal breakpoints occur [17,18]. The presence of SDs has also been shown, by different methods, in BOS between human and other great apes [7,50–54]. Interestingly, human SDs have also been found in regions where recurrent chromosomal rearrangements, which lead to either structural polymorphisms or genomic disorders, occur. In this study, the majority of breaks of synteny contain more SDs than expected in a random distribution, either within nonsyntenic regions or their boundaries or both. Only mouse chromosome 12 was not found to contain SDs in breaks of synteny.

"Classical" repeats (i.e., DNA elements, LINEs, SINEs, and LTRs) are not systematically found in excess in synteny breakpoints. This raises two different possibilities: either a different type of repeat drives each rearrangement event and, by averaging the breakpoints, we have missed this information or classical repeats are not directly responsible for these events. In the first case, a detailed analysis of each individual breakpoint will be necessary. In the second scenario, SDs by themselves, or other unknown epigenetic phenomena related to them, could act as the driving force of evolutionary rearrangements. Since previous studies have proven the presence as well as the absence of different types of classical repeats in these regions, we propose that the hypotheses above are not mutually exclusive and that both elements could be acting either alone or synergistically.

In a recent paper, Murphy et al. [7] demonstrated a significant enrichment of human genes close to evolutionary breakpoints; we were not able to assess such enrichment in the mouse rat/synteny breakpoint regions, either due to the primate specificity of such gene enrichment or because we used a more restricted set of genes (i.e., only RefSeq genes instead of RefSeq+ gene predictions that were used in that paper).

In addition, our results show that genes mapping to SDs have undergone accelerated rates of sequence evolution. The analysis performed after exclusion of SDs allowed us to assess a higher divergence rate for genes inside inversion rearrangements and close to breakpoint regions compared to genes in colinear regions. These results are consistent with a role for chromosomal rearrangements in the speciation processes that separated rat and mouse. This being said, further research making use of outgroups will be necessary to clarify this issue.

Taken together, these observations suggest a relationship between SDs, chromosomal rearrangements, and the speciation process. Given the significant correlation of SDs with synteny breakpoints, one could speculate that duplicons themselves, or sequences located in SDs, could play a key role (very likely acting as catalyzers of nonallelic homologous recombination) as a driving force for evolutionary chromosomal rearrangements, which, in turn, could promote the chromosomal evolution leading to speciation. A recent paper by Zhou et al. [55] demonstrates that SDs are flanked by nonrepeat sequences that possess physical features that coincide with those in fragile sites (decreased DNA helix stability and increased flexibility). This would provide these regions with increased liability to breakage and thus increase the possibility of being involved in rearrangements.

The work presented here supports a potential role for SDs in evolution, since these repeats are consistently found in genomic sequences from different species correlated to a large number of evolutionary breakpoints. Although no intrinsic sequence similarities between SDs from different species have so far been discovered, it is possible that similar mechanisms involving equivalent genomic structures have occurred in different evolutionary lineages, providing the appropriate hallmarks for both chromosomal and gene evolution.

Materials and methods

Identification of synteny blocks

Alignments between mouse (mm5) and rat (rn3) sequences were downloaded from the University of California at Santa Cruz (http://www.genome.ucsc.edu/goldenPath/mm5/vsRn3/axtNet). Information regarding the filtering of the alignments is publicly available at the UCSC Web site. Alignments were done using BLASTZ and filtered as described elsewhere [56]. To construct the synteny blocks, we proceeded as previously described in [57]. Briefly, we transformed genomic alignment information into anchors, which are two-dimensional diagonals formed by the start position in each genome and the length of the alignment. The distance between two anchors in the same chromosome is calculated as the Manhattan distance between their closest ends. To avoid interference from small isolated paralogous anchors, the original anchors were grouped into higher order structures called *synteny clusters* if the calculated distance between them was smaller than a given threshold (G = 500 kb). Clusters smaller than a given threshold (C = 250 kb) were discarded to overcome possible assembly errors. In genomic regions that contained gaps or that had been difficult to assemble, it was possible that two consecutive anchors were more apart than the maximum allowed distance (G), so a single synteny block would appear in consecutive synteny clusters. To overcome this problem, if a sequence of consecutive clusters in the first genome appeared either in the same or in the reverse order in the second genome they were merged into the same *synteny block*. Regions between synteny blocks are referred as break of synteny regions. Biologically relevant BOS were associated with evolutionary rearrangements between two species. In such cases, the rearrangement causing the BOS could be identified as an inversion or a translocation. The positions of such BOS were further confirmed using a complementary approach based on the positions of genes (see below). Several scripts written in Java were used for this approach.

Although a huge effort was made during the assembly process to obtain the most accurate picture of the real genomic sequences, it is known that the released sequences were not free of assembly errors. These errors, together with small synteny regions produced by microrearrangements and clusters of indels, could generate incorrect assemblies and produce artifactual BOS. Such regions, where no synteny blocks could be constructed, were called no-blocks. To minimize their presence in our analysis, we used conservative G and C criteria that restricted the study to large synteny blocks derived from either orientation changes within a single chromosome or translocations between chromosomes, but allowed us to overcome partially the corruption caused by assembly errors.

Segmental duplications in breakpoint regions

To obtain a graphical overview, we wrote several Perl scripts that allowed us to draw dot plots in which we included information on SDs. We also developed other scripts that allowed us to evaluate the presence of SDs in break of synteny regions and in the regions flanking them. The presence of SDs was also assessed in nonoverlapping windows of 50 kb around the BOS.

Gaps in the assemblies

We obtained the files containing the position and size of gaps in the mouse and rat assemblies from publicly available files in the annotation databases of the UCSC.

Simulation studies

To obtain the random distributions of synteny breakpoints that were used for most of the analyses in the present work, we performed a stochastic reassignment of the positions of each of the breakpoints without replacement using Perl's random number generator through the rand() function. Sizes of breakpoints were kept constant. To test the significance of the observations, we compared the number of times that an event was observed (presence of segmental duplications, repeats, etc.) in our set of data with the number of times that the same event was observed in the random distribution. A total of 1000 permutations were carried out in each experiment. The p value was calculated as the number of times that the observed equaled or exceeded the expected divided by the total number of permutations plus 1 (observed $\geq$ expected)/(permutations + 1) [58].

The p value for the enrichment of observed small synteny blocks versus the expected by the random-breakage model was calculated using a goodness of fit (χ^2) test, after a 2 × 2 contingency table.

Genes, ontology, and repeat content

Gene information was extracted from RefFlat tables corresponding to the mm5 and rn3 assemblies available at the UCSC. RefFlat contains essentially the same information as RefGene plus an extra column with the gene symbol that was used for the GO analysis. We used Awk and Perl scripts to count and obtain information on genes in target regions. Information on gene function and processes in which genes are involved was obtained from the Gene Ontology Tree Machine database at http://www.genereg.ornl.gov/gotm/.

Information about repeats in regions of interest was obtained by parsing repeat content tables, available at UCSC (chr*_rmsk.txt), corresponding to the mm5 and rn3 mouse and rat assemblies. We used Perl scripts to generate reports of repeat content in each region.

Evolutionary rates

We obtained approximately 13,000 orthologous genes from Ensembl (http://www.ensembl.org). To avoid false orthologous gene pairs, several filters were applied. We kept only the unique Best reciprocal hit and those genes with a cutoff of twice the median value of all dS (cutoff dS = 0.4074). We determined the positions of these genes relative to the synteny blocks and rearrangements we had previously determined. Single genes located in a synteny block that did not belong to it (that is, whose position was not in accordance with that of their surrounding genes) were potentially misplaced in the genomic assembly and, thus, were conservatively removed from the analysis.

To study patterns of molecular divergence we used several classical indexes of molecular evolution: the number of nonsynonymous substitutions per nonsynonymous site (dN), the number synonymous substitutions per silent site (dS), and their ratio (dN/dS). These indexes were obtained using the Codeml program included in the PAML package. The p values of comparisons of genes in different locations of the genome were obtained by means of pair-wise permutation tests (based of 1000 permutations). The significance level was the proportion of times that the difference of class averages after permutation was equal to or larger than the observed difference.

Acknowledgments

We are grateful to J.F. Abril and O. González, at the GRIB, for bioinformatics support and discussion and to H. Howard for helpful comments in the preparation of the manuscript. This work was supported in part by the Departament d'Universitats Recerca i Societat de la Informació and the GENOMA ESPAÑA and GENOME CANADA joint R+D+I projects in human health, plants, and aquaculture.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.08.008.

References

[1] J.H. Nadeau, B.A. Taylor, Lengths of chromosomal segments conserved since divergence of man and mouse, Proc. Natl. Acad. Sci. U. S. A. 81 (1984) 814–818.

[2] J.H. Nadeau, D. Sankoff, The lengths of undiscovered conserved segments in comparative maps, Mamm. Genome 9 (1998) 491–495.
[3] P. Pevzner, G. Tesler, Transforming men into mice: the Nadeau–Taylor chromosomal breakage model revisited, "RECOMB 2003," April 10–13, 2003, Max Planck Institute for Molecular Genetics and the Berlin Center for Genome-Based Bioinformatics, Berlin, Germany, 2003,
[4] G. Bourque, P.A. Pevzner, G. Tesler, Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes, Genome Res. 14 (2004) 507–516.
[5] S. Zhao, J. Shetty, L. Hou, A. Delcher, B. Zhu, K. Osoegawa, et al., Human, mouse, and rat genome large-scale rearrangements: stability versus speciation, Genome Res. 14 (2004) 1851–1860.
[6] G. Bourque, E.M. Zdobnov, P. Bork, P.A. Pevzner, G. Tesler, Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages, Genome Res. 15 (2005) 98–110.
[7] W.J. Murphy, D.M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler, L. Auvil, et al., Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps, Science 309 (2005) 613–617.
[8] P. Trinh, A. McLysaght, D. Sankoff, Genomic features in the breakpoint regions between syntenic blocks, Bioinformatics 20 (Suppl. 1) (2004) I318–I325.
[9] E. Kolomietz, M.S. Meyn, A. Pandita, J.A. Squire, The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors, Genes Chromosomes Cancer 35 (2002) 97–112.
[10] C.J. Shaw, J.R. Lupski, Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease, Hum. Mol. Genet. 1 (2004) R57–R64.
[11] G.R. Sutherland, E. Baker, R.I. Richards, Fragile sites still breaking, Trends Genet. 14 (1998) 501–506.
[12] A. Ruiz-Herrera, M. Ponsa, F. Garcia, J. Egozcue, M. Garcia, Fragile sites in human and Macaca fascicularis chromosomes are breakpoints in chromosome evolution, Chromosome Res. 10 (2002) 33–44.
[13] D. Sankoff, M. Deneault, P. Turbis, C. Allen, Chromosomal distributions of breakpoints in cancer, infertility, and evolution, Theor. Popul. Biol. 61 (2002) 497–501.
[14] P. Dehal, P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, et al., Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution, Science 293 (2001) 104–111.
[15] R. Puttagunta, L.A. Gordon, G.E. Meyer, D. Kapfhamer, J.E. Lamerdin, P. Kantheti, et al., Comparative maps of human 19p13.3 and mouse chromosome 10 allow identification of sequences at evolutionary breakpoints, Genome Res. 10 (2000) 1369–1380.
[16] M. Kost-Alimova, H. Kiss, L. Fedorova, Y. Yang, J.P. Dumanski, G. Klein, et al., Coincidence of synteny breakpoints with malignancy-related deletions on human chromosome 3, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 6622–6627.
[17] L. Armengol, M.A. Pujana, J. Cheung, S.W. Scherer, X. Estivill, Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements, Hum. Mol. Genet. 12 (2003) 2201–2208.
[18] J.A. Bailey, R. Baertsch, W.J. Kent, D. Haussler, E.E. Eichler, Hotspots of mammalian chromosomal evolution, Genome Biol. 5 (2004) R23.
[19] P. Stankiewicz, J.R. Lupski, Molecular-evolutionary mechanisms for genomic disorders, Curr. Opin. Genet. Dev. 12 (2002) 312–319.
[20] P. Stankiewicz, C.J. Shaw, J.D. Dapper, K. Wakui, L.G. Shaffer, M. Withers, et al., Genome architecture catalyzes nonrecurrent chromosomal rearrangements, Am. J. Hum. Genet. 72 (2003) 1101–1116.
[21] M. Caceres, J.M. Ranz, A. Barbadilla, M. Long, A. Ruiz, Generation of a widespread Drosophila inversion by a transposable element, Science 285 (1999) 415–418.
[22] M.T. Pletcher, B.A. Roe, F. Chen, T. Do, A. Do, E. Malaj, et al., Chromosome evolution: the junction of mammalian chromosomes in the formation of mouse chromosome 10, Genome Res. 10 (2000) 1463–1467.
[23] H. Kehrer-Sawatzki, C. Sandig, N. Chuzhanova, V. Goidts, J.M. Szamalek, S. Tanzer, et al., Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes), Hum. Mutat. 25 (2005) 45–55.
[24] A. Navarro, N.H. Barton, Chromosomal speciation and molecular divergence—Accelerated evolution in rearranged chromosomes, Science 300 (2003) 321–324.
[25] T. Marques-Bonet, M. Caceres, J. Bertranpetit, T.M. Preuss, J.W. Thomas, A. Navarro, Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees, Trends Genet. 20 (2004) 524–529.
[26] J. Zhang, X. Wang, O. Podlaha, Testing the chromosomal speciation hypothesis for humans and chimpanzees, Genome Res. 14 (2004) 845–851.
[27] E.J. Vallender, N.M. Pearson, B.T. Lahn, The X chromosome: not just her brother's keeper, Nat. Genet. 37 (2005) 343–345.
[28] J. Lu, W.H. Li, C.I. Wu, Comment on "Chromosomal speciation and molecular divergence—Accelerated evolution in rearranged chromosomes, " Science 302 (2003) 988 (author reply 988).
[29] X. Gu, Evolution of duplicate genes versus genetic robustness against null mutations, Trends Genet. 19 (2003) 354–356.
[30] Z. Gu, L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, W.H. Li, Role of duplicate genes in genetic robustness against null mutations, Nature 421 (2003) 63–66.
[31] M.G.S. Consortium, R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[32] R.G.S. Consortium, R.A. Gibbs, G.M. Weinstock, M.L. Metzker, D.M. Muzny, E.J. Sodergren, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, Nature 428 (2004) 493–521.
[33] S. Ohno, Ancient linkage groups and frozen accidents, Nature 244 (1973) 259–262.
[34] P. Pevzner, G. Tesler, Genome rearrangements in mammalian evolution: lessons from human and mouse genomes, Genome Res. 13 (2003) 37–45.
[35] J. Cheung, M.D. Wilson, J. Zhang, R. Khaja, J.R. MacDonald, H.H. Heng, et al., Recent segmental and gene duplications in the mouse genome, Genome Biol. 4 (2003) R47.
[36] E. Tuzun, J.A. Bailey, E.E. Eichler, Recent segmental duplications in the working draft assembly of the brown Norway rat, Genome Res. 14 (2004) 493–506.
[37] E.E. Eichler, Masquerading repeats: paralogous pitfalls of the human genome, Genome Res. 8 (1998) 758–762.
[38] J. Cheung, X. Estivill, R. Khaja, J.R. MacDonald, K. Lau, L.C. Tsui, et al., Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence, Genome Biol. 4 (2003) R25.
[39] X. She, Z. Jiang, R.A. Clark, G. Liu, Z. Cheng, E. Tuzun, et al., Shotgun sequence assembly and recent segmental duplications within the human genome, Nature 431 (2004) 927–930.
[40] J.A. Bailey, A.M. Yavor, H.F. Massa, B.J. Trask, E.E. Eichler, Segmental duplications: organization and impact within the current human genome project assembly, Genome Res. 11 (2001) 1005–1017.
[41] F.C. Chen, W.H. Li, Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees, Am. J. Hum. Genet. 68 (2001) 444–456.
[42] J. Castresana, Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content, Nucleic Acids Res. 30 (2002) 1751–1756.
[43] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.
[44] V.E. Prince, F.B. Pickett, Splitting pairs: the diverging fates of duplicated genes, Nat. Rev. Genet. 3 (2002) 827–837.
[45] W.J. Murphy, R. Stanyon, S.J. O'Brien, Evolution of mammalian genome organization inferred from comparative gene mapping, Genome Biol. 2 (2001) (REVIEWS0005).
[46] S.J. O'Brien, M. Menotti-Raymond, W.J. Murphy, W.G. Nash, J. Wienberg, R. Stanyon, The promise of comparative genomics in mammals. Science 286 (1999) 458–462, 479–481.
[47] J.M. Carlton, S.V. Angiuoli, B.B. Suh, T.W. Kooij, M. Pertea, J.C. Silva,

et al., Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii, Nature 419 (2002) 512–519.
[48] A. Coghlan, K.H. Wolfe, Fourfold faster rate of genome rearrangement in nematodes than in Drosophila, Genome Res. 12 (2002) 857–867.
[49] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E.S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature 423 (2003) 241–254.
[50] E. Nickerson, D.L. Nelson, Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees, Genomics 50 (1998) 368–372.
[51] R.V. Samonte, E.E. Eichler, Segmental duplications and the evolution of the primate genome, Nat. Rev. Genet 3 (2002) 65–72.
[52] D.P. Locke, N. Archidiacono, D. Misceo, M.F. Cardone, S. Deschamps, B. Roe, et al., Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster, Genome Biol. 4 (2003) R50.
[53] V. Goidts, J.M. Szamalek, H. Hameister, H. Kehrer-Sawatzki, Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates, Hum. Genet. 115 (2004) 116–122.
[54] W.J. Murphy, R. Agarwala, A.A. Schaffer, R. Stephens, C., N.J. Crumpler Jr., et al., A rhesus macaque radiation hybrid map and comparative analysis with the human genome, Genomics (2005).
[55] Y. Zhou, B. Mishra, Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 4051–4056.
[56] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, et al., Human–mouse alignments with BLASTZ, Genome Res. 13 (2003) 103–107.
[57] P. Pevzner, G. Tesler, Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 7672–7677.
[58] P. Good, Permutation Tests, Springer-Verlag, New York, 2000.