BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i318–i325
DOI: 10.1093/bioinformatics/bth934

Genomic features in the breakpoint regions between syntenic blocks

Phil Trinh[1], Aoife McLysaght[2] and David Sankoff[3,*]

[1]Hillcrest High School, Ottawa K1G 2L7, Canada, [2]Genetics Department, Trinity College, University of Dublin, Dublin 2, Ireland and [3]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa K1N 6N5, Canada

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: We study the largely unaligned regions between the syntenic blocks conserved in humans and mice, based on data extracted from the UCSC genome browser. These regions contain evolutionary breakpoints caused by inversion, translocation and other processes.

Results: We suggest explanations for the limited amount of genomic alignment in the neighbourhoods of breakpoints. We discount inferences of extensive breakpoint reuse as artefacts introduced during the reconstruction of syntenic blocks. We find that the number, size and distribution of small aligned fragments in the breakpoint regions depend on the origin of the neighbouring blocks and the other blocks on the same chromosome. We account for this and for the generalized loss of alignment in the regions partially by artefacts due to alignment protocols and partially by mutational processes operative only after the rearrangement event. These results are consistent with breakpoints occurring randomly over virtually the entire genome.

Contact: sankoff@uottawa.ca

1 INTRODUCTION

Complex alignment protocols developed independently by two research groups (Pevzner and Tesler, 2003; Kent *et al.*, 2003) have reconstructed the chromosomal segments conserved in the evolution of the genome sequences of both mouse and man, without recourse to an intermediate stage of orthologous gene identification. The protocols use somewhat different strategies to combine short regions of elevated similarity to construct the conserved segments, bridging singly- or doubly-gapped regions where similarity does not attain a threshold criterion and ignoring short inversions and transpositions that have rearranged one sequence or the other. The difficulty of this reconstruction task cannot be overemphasized and its accomplishment is a testimony to the scientific judgement and computational skills of the participating researchers.

One aspect of the reconstruction that is of particular interest is the nature of the DNA sequence in the neighbourhood of the breakpoints between two ‘conserved’ (or ‘syntenic’) blocks adjacent on an autosome in the human genome, say, but remote or even on different autosomes in the mouse genome. (For clarity, we will continue our exposition treating the human and mouse genomes asymmetrically in this way, though their roles could be reversed without materially affecting the discussion or results.) Generally, the two syntenic blocks on either side of the breakpoint do not abut directly, but are rather separated by a short region where there is little similarity with the mouse genome. These regions (or ‘spaces’) do generally contain a number of smaller fragments of homology with the same mouse chromosomes as the two adjacent syntenic blocks (the ‘archipelago’), with other mouse autosomes sharing syntenic blocks with the same human chromosome (the ‘compatriots’) and with mouse chromosomes, including the X, having no such syntenic blocks (the ‘foreigners’). Figure 1 depicts these categories.

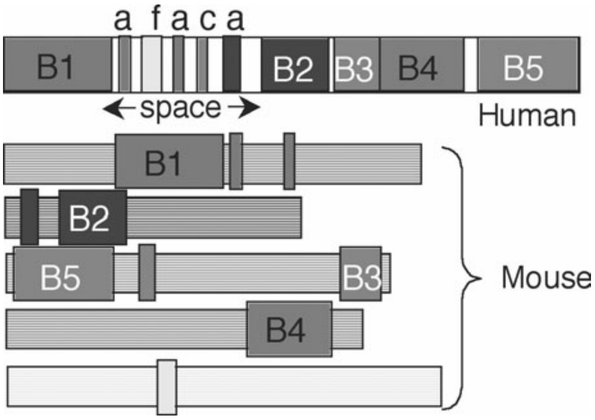

Fig. 1. Hypothetical human chromosome with ‘syntenic’ blocks B1–B5 and small fragments, with shading keyed to aligned portions of mouse chromosomes. a, ‘archipelago’; c, ‘compatriot’; and f, ‘foreigner’.

The breakpoints are created by chromosomal rearrangement process such as inversion and translocation of various kinds that drive the evolution of genomic structure. Where in the genome these breakpoints can and do occur is a fundamental

*To whom correspondence should be addressed.

question in the evolution of species, and it is in the hope that the small fragments within the breakpoint regions contain some hints about this question that we undertake a statistical assessment of the three types.

2 THE RANDOM HYPOTHESIS AND THE ALTERNATIVE

At a sufficiently low level of resolution, one might hypothesize that breakpoints could occur randomly along the lengths of chromosome, in analogy to recombination sites. Indeed, this hypothesis is implicit in the prophetic work of Nadeau and Taylor (1984) in their early estimation of the number of conserved segments in the human–mouse comparison. Again in analogy with recombination sites, we may weaken this hypothesis by allowing some variation of breakage susceptibility. And of course at higher levels of resolution, we would expect selection to disfavour breakage at gene-internal sites (in introns and especially within exons) or occasionally between neighbouring genes co-expressed for functional reasons, while breakage is known to be endemic in eukaryotes in subtelomeric regions (Mefford and Trask, 2002; Kellis *et al.*, 2003; Katinka *et al.*, 2001) and, at least in primates, much rearrangement seems to occur in pericentromeric regions (Bailey *et al*., 2002). Nevertheless, with specific exceptional regions, accounting for perhaps 5% of the genome, the idea that evolutionary rearrangements can break chromosomes anywhere in the genome cannot be rejected with current data. Indeed, the only data not of the historical inference type bearing directly on this question, namely the location of breakpoints in (non-sterile) human carriers of translocations, suggests a uniform distribution the length of the chromosome, contrasting with breakpoints in somatic cell (tumour) genomes, which are non-uniformly concentrated arm-centrally on chromosomes, or in subtelomeric bands (Sankoff *et al.*, 2002).

Documentation of evolutionary subtelomeric translocational hotspots and pericentromeric duplication and/or transpositional hotspots lead, nonetheless, to an alternate hypothesis, that potential breakpoints are largely restricted to a limited number (e.g. <500) of very small regions in the genome, and that this regional susceptibility is conserved over considerable evolutionary time scales. This position has been argued most forcefully by Pevzner and Tesler (2003), who advanced the hypothesis that the observed spaces between syntenic blocks correspond to ‘fragile’ breakpoint regions and these are conserved, at least across the mammals. The main evidence offered for this claim is that an algorithmic reconstruction of rearrangement history, based on the current positions of the syntenic blocks in the two species, requires almost the same number of rearrangements (mostly inversions and reciprocal translocations) as the number of blocks, implying that each breakpoint region contains almost two breakpoints, on the average (since each inversion or reciprocal translocation involves two breakpoints). Were the two breakpoints for each rearrangement situated at random chromosomal sites, on the other hand, it would be rare that any two points would fall in the same small region.

Pevzner and Tesler (2003) interpret the lack of sustained human–mouse similarity in the breakpoint regions as suggestive of frequent rearrangement affecting these regions, which we may term ‘churning’.

Further lines of evidence for this viewpoint include the high rates of recurrence of certain breakpoints in the clinical study of tumour cell karyotypes, and the existence of certain physically fragile regions in human chromosomes under laboratory conditions.

3 A CONCERTED FOCUS ON THE BREAKPOINT REGIONS

In this paper, we take issue with the Pevzner–Tesler interpretation of all these lines of evidence and suggest different explanations for the limited amounts of similarity in the neighbourhood of breakpoints. We review arguments against their breakpoint recurrence, or reuse, results as reflecting artefacts introduced during their reconstruction of syntenic blocks, and show how these techniques artificially inflate reuse rates even when breakpoints are uniformly distributed across the genome. We study the number, size and distribution of archipelago, compatriot and foreigner fragments in the breakpoint regions and account for them and for the reduction of similarity in terms not only of possible artefacts from the alignment protocols, but also in terms of several biological processes, only one of which is specific to this type of region. The latter process is operative only *after* the rearrangement event, and is consistent with breakpoints occurring randomly over virtually the entire genome.

4 BREAKPOINT REUSE

Pevzner and Tesler goal was to infer evolutionary history without having to deal with gene finding and orthologue identification, using only the order of syntenic blocks constructed solely from sequence data as input to a genome rearrangement algorithm. Their method focuses on major evolutionary events by glossing over small block-internal rearrangements, and neglecting intervening blocks smaller than a threshold length. We have previously shown, however, that setting aside short blocks and small rearrangements may blur important parts of the historical derivation of the genomes (Sankoff and Trinh, 2004). We modelled the effects of eliminating and amalgamating short blocks, concentrating on the summary statistic of breakpoint reuse, which can vary from 2.0, in the random model, down to 1.0, the minimum inferable from the reconstruction algorithm. We used analytical and simulation methods to investigate this statistic as a function of threshold size and of rearrangement parameters. We showed that breakpoint reuse of the same magnitude as found by Pevzner and

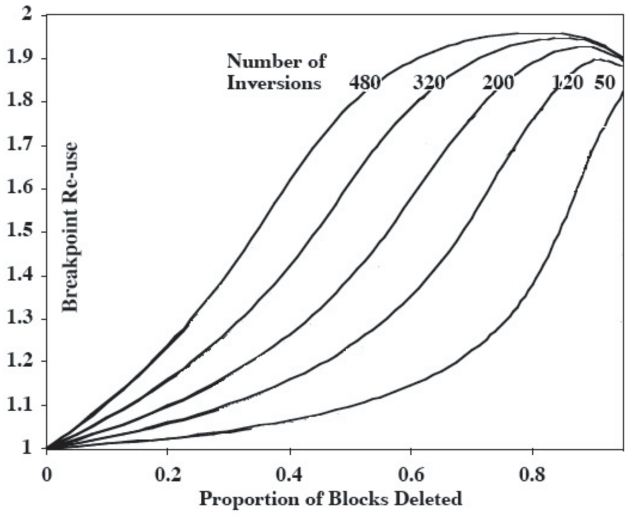

Fig. 2. Effects of deleting blocks on inferred reuse of breakpoints.

Tesler (2003), i.e. close to 2 and can be artefacts of setting aside small blocks in the purely random context, i.e. where no reuse actually occurred. For example, in Figure 2, we show the results of simulating many inversions, each with two random breakpoints but with *no* reuse, on a single chromosome and then deleting a proportion of both initial and final chromosomes, prior to applying the reconstruction algorithm. The deletion step simulates the setting aside of small blocks and the proportion deleted is analogous to the threshold criterion.

Figure 2 shows that breakpoint reuse rises rapidly as the proportion of the chromosome deleted is increased, especially for highly rearranged genomes. Further simulations showed that this effect is amplified when the effects of short rearrangements were systematically overridden when reconstructing syntenic blocks.

Granted our experiment is a rather abstract analogy to the rearrangement divergence of human and mouse. Nevertheless, the number of rearrangements found by Pevzner and Tesler was in the order of a few hundred, and the proportion of blocks they discarded in constructing conserved syntenic blocks was at least as large, and probably much larger, suggesting that, as in Figure 2, an inferred breakpoint reuse close to 2 is likely to be an artefact. We concluded that in the context where they use it, the statistic Pevzner and Tesler invented does not measure breakpoint reuse, but instead effectively assesses the amount of noise affecting a genomic rearrangement inference.

5 SIMILARITY LOSS NEAR BREAKPOINTS

Given that the reuse statistics cannot be considered solid evidence of breakpoint reuse, how can we assess the notions of evolutionarily conserved fragility of breakpoint regions and the lack of human–mouse similarity in these regions? Indeed, there is some inherent contradiction proposing both of these simultaneously: if conserved fragility is based on some substantial primary sequence signal, why is this not picked up by the alignment protocol and how is it conserved if the region is being churned by rearrangements? There are of course, many possible answers: the signals may be too short, they may be removed by the repeat masking prior to the reconstruction of the syntenic blocks, they may involve conserved secondary but not primary structures, they may involve GC-poorness or other gross sequence characteristics, or they may even be determined by unknown epigenetic considerations. There is no evidence, however, for any of these, nor for that matter, for the notion that the breakpoint regions contain multiple breakpoints.

We attribute the lack of similarity in the breakpoint region not to some aspect of an a priori proclivity for breakage, but rather to some combination of the following three factors.

- The algorithms that reconstruct the syntenic blocks bridge gaps as long as appropriate similarity exists at *both* ends of the gap. A rearrangement event with one breakpoint within a gap destroys the match between the homologies at each end.
- To the extent that breakage occurs disproportionately in intergenic regions, these tend to undergo more rapid sequence evolution than regions containing exons and introns.
- Though we are unaware of any pertinent molecular cytogenetic evidence, we hypothesize the increase in aberrant processes, such as recombination errors, deletion, duplication or retroposition in the vicinity of breakpoints in quadrivalent meiotic figures (in the case of reciprocal translocations) and in looped figures (in the case of inversions), during the period of heterokaryotypy before the rearrangement becomes fixed in a population, as depicted in Figure 3.

  This seems a likely consequence of the breakpoint neighbourhood being unaligned in these figures during the meiotic pairing of homologous chromosomes. The length of sequence affected may well be of the same order as those we observe in the regions between reconstructed syntenic blocks. The resultant reduction in similarity would be highly variable, depending on population size, generation time and other factors. Similarly, any template-assisted DNA repair or conversion processes depending on homology between chromosome pairs is subject to disruption in the neighbourhood of breakpoints, resulting in the acceleration of sequence divergence.

Of these three factors, the first is basically an analytical artefact and the second applies widely across the genome. Only the third is a biological process specific to the breakpoint regions, and this would only apply *after* a rearrangement. We claim there is as yet no direct evidence for assuming any pre-existing properties of a site predisposing it to a rearrangement

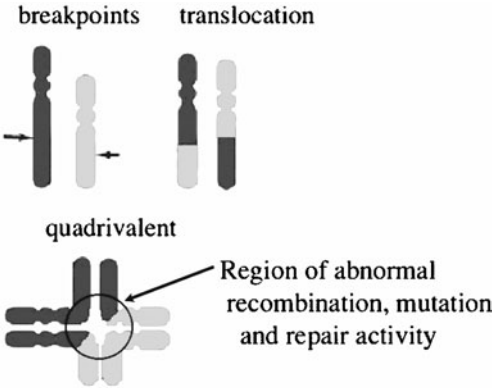

Fig. 3. Proposed effect of meiotic non-alignment of regions surrounding breakpoints in heterokaryotypes.

event and fixation. The characteristics of breakpoint regions appear, biologically and/or analytically, after and not before the rearrangement.

6 DIVERSE ORIGIN OF SEQUENCE FRAGMENTS IN BREAKPOINT REGIONS

6.1 The data

For the purpose of the present work we use the UCSC Genome Browser http://genome.UCSC.edu/, July 2003 freeze for the human assembly and February 2003 freeze for the mouse assembly. The ingenious 'net' constructions featured in this browser provide us with well documented first-level, non-overlapping, syntenic blocks, with second- or third-level blocks nested in the gaps in the first level, but also allow us to zoom in as closely as desired to sequence details. Whereas Pevzner and Tesler (2003) used a 1 Mb threshold for blocks, work in the same laboratory (Bourque *et al*., 2004) subsequently lessened this to 300 kb and lower. We adopt a 100 kb threshold, closer to the scale of the breakpoint regions. Our selection of the blocks is thus influenced by the parameters and conventions used in the net construction, and may contain a small number of non-existent blocks and may be missing others due to assembly errors and other artefacts. The risk of such errors has presumably been greatly reduced in successive improvements of the genome sequences.

Thus, we extracted all first-level blocks of length 100 kb or larger in the human genome. In addition, where any such block contained gaps of 100 kb or larger, containing one or more nested blocks >100 kb syntenic in different mouse chromosomes, we split the first level block in two, as long as these remained larger than the threshold, and included these new items in our final set of large, non-overlapping blocks. This resulted in 364 blocks on all 23 chromosomes, based on 318 first level blocks on the mouse net, some of which we divided in two and separated in order to consider nested blocks >100 kb. There are thus, $341 = 364 - 23$ spaces between the blocks. Of these spaces, we set aside 21 pericentromeric spaces[1] subject to repetitive segmental duplication and/or transposition (Bailey *et al*., 2002), leaving 320 spaces for our analysis.

Our resulting data for chromosome 20 then contained only one interblock space and those for chromosome 21 contained only two, but 17 of the 23 chromosomes contained 10 or more spaces, even without the discarded subtelomeric and pericentromeric exceptions. Chromosome 2 contained the maximum, 28 spaces. The number of different mouse chromosomes with at least one syntenic block on a given human chromosome ranged from one, for human chromosomes 20 and X, to nine, for chromosomes 2 and 10, with a mean of 4.8.

We found that in eight cases, the two adjacent syntenic blocks abutted directly, so that the space had length zero, while three other spaces were <100 bp. The median length was 120 kb, the longest 4.5 Mb and 18 others longer than 1 Mb.

For about half the spaces, the two adjacent syntenic blocks were from different mouse chromosomes. This was true for about half of the very short spaces and half of the very long ones, but their median length was somewhat higher, at 148 kb.

From each of the 320 spaces, we extracted all the 12 930 smaller aligned fragments identified by the browser (by definition <100 kb, but overwhelmingly much smaller), and categorized them by length, origin (i.e. which position on which mouse chromosome) and position within the space. Consequently, by taking into account the syntenic blocks adjacent to each space as well as the rest of the blocks on the same chromosome, we labelled each fragment as archipelago ($N = 4139$), compatriot ($N = 2706$) or foreigner ($N = 6085$).

6.2 Statistical analysis

Our null hypothesis will be that the fragments contained within any given space are chosen at random from anywhere in the genome, i.e. from any chromosome, that their sizes are drawn from some common distribution, independent of the two syntenic blocks surrounding the space, and what is elsewhere on the same chromosome, and that they are randomly ordered within the space. That is, we assume that there is no statistical difference between the archipelago, the compatriots and the foreigners.

The alternate hypothesis, derived from biological considerations explained below, is that compatriot fragments will be bigger and proportionately more numerous than foreigners and, especially, that fragments belonging to the archipelago in a space will be bigger and more numerous than other compatriots (and, ipso facto, than the foreigners). Moreover, the archipelago fragments will be 'chips off the old block', closer

[1] This includes pericentromeric spaces in chromosomes 1p and q, 2p, 5p, 7–12 p and q, 16–19 p and q and X p and q. By our definitions some spaces spanned an entire centromere, but most of these were among the excluded spaces. We did not have to deal with subtelomeric phenomena as none of the syntenic blocks extended to the telomere.

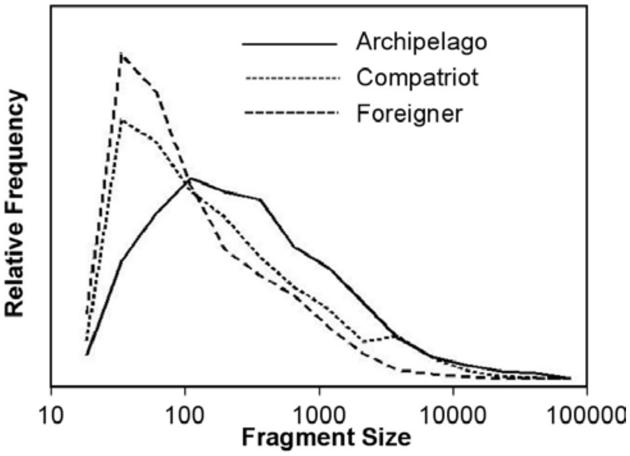

Fig. 4. Length distribution for fragment categories.

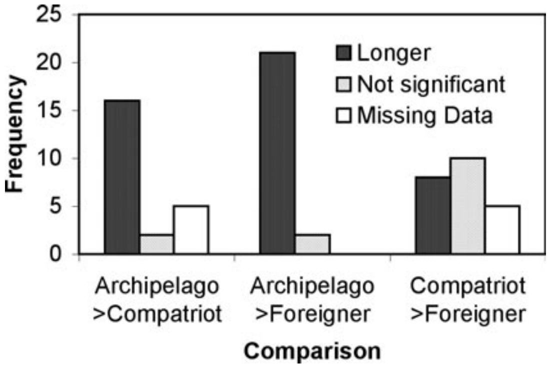

Fig. 5. Number of chromosomes for which the null hypotheses of identical size fragments is rejected or accepted.

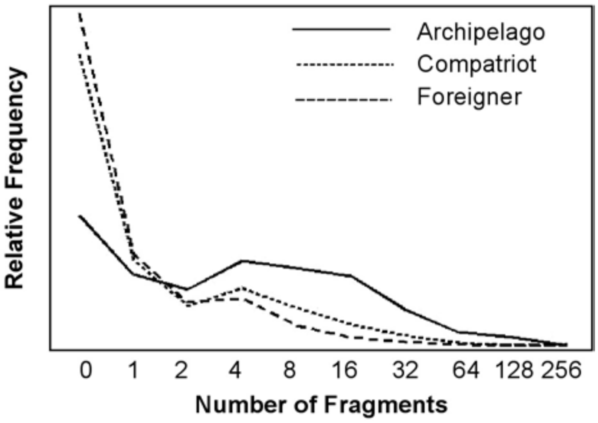

Fig. 6. Distributions of fragment numbers per space.

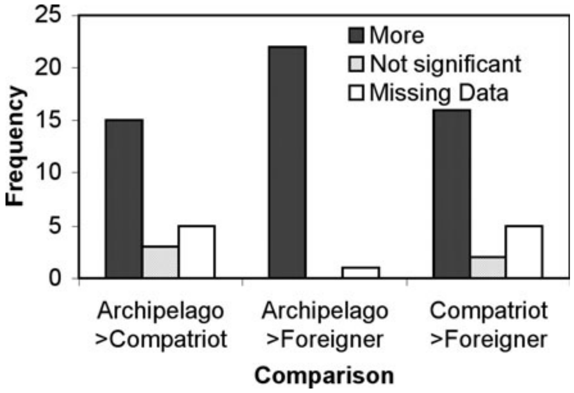

Fig. 7. Number of human chromosomes for which the null hypotheses of an identical distribution of the number of fragments aligned with all mouse chromosomes is rejected.

to the adjacent syntenic block from the same mouse chromosome than to the other block adjacent to the space, and will also be close to the homologous block in the mouse chromosome.

6.3 Data analysis

6.3.1 Distribution of block sizes The archipelago fragments are considerably longer than the compatriot and foreigner fragments as can be seen from the distributions of fragment length in Figure 4. The median length of the archipelago fragments is twice as large as either of the other two in most chromosomes.

Figure 5 plots the number of chromosomes for which a one-tailed Kolmogorov–Smirnov test rejects the null hypothesis that they have the same distribution.

Figures 4 and 5 also show that the compatriot fragments are systematically longer than the foreigner ones, though the difference is less marked than that between either of these categories and the archipelago. Of the 18 chromosomes for which there are sufficient data, 14 have longer mean fragment size for compatriots than foreigners, and 8 of these are significantly so at the 5% level.

6.3.2 Distribution of number of fragments of each type The null hypothesis is that each fragment in a space has the same chance of aligning with a fragment from any mouse chromosome. To correct for the different number of archipelago, compatriot and foreigner mouse chromosomes, we divide the numbers of fragments in each space by 2, c and $18-c$, respectively, where c is the number of compatriots.

This normalization reveals that there are of the order of 10 times as many fragments from each archipelago-associated chromosome as from each other compatriot-associated chromosome, and of the order of 100 times as many as from each foreigner chromosome. Figure 6 shows the distribution of the number (not normalized) of fragments of each type.

Figure 7 plots the number of chromosomes for which a one-tailed Kolmogorov–Smirnov test rejects the null hypothesis that all mouse chromosomes are aligned with the same distribution of number (normalized) of fragments in a space.

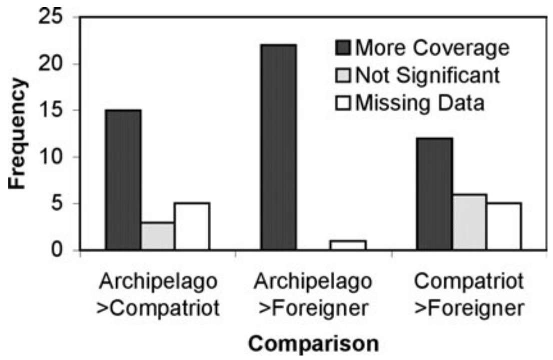

Fig. 8. Number of chromosomes for which the null hypotheses of identical coverage distribution for three categories of fragments is rejected.

6.3.3 Distribution of proportion of space covered by each type of block Archipelago fragments tend to cover considerably more of each space than the compatriot and foreigner fragments. We calculated the proportion, not of the total sequence in each space, but of the total aligned sequence, for each type of fragment, with the same normalization as in Section 6.3.2. Figure 8 plots the number of chromosomes for which a one-tailed Kolmogorov–Smirnov test rejects the null hypothesis that these proportions have the same distribution.

6.3.4 Interspersed left and right fragments in the archipelago The Mann–Whitney–Wilcoxon runs test confirms that for 45% of spaces separating blocks syntenic with two different mouse chromosomes, the archipelago fragments segregate into two clear groups, each group closer to the corresponding block (with a further 29% tending to segregate in the same sense, but not based on enough data to be statistically significant). This is consistent with the idea that it is high rates of local mutation that make it difficult to detect homology in this region, rather than additional rearrangement. Nevertheless the two groups usually overlap, and the non-rejection of the runs test in over half of the spaces is suggestive of some degree of local rearrangement after the two major blocks were concatenated.

6.3.5 Provenance of foreigner, compatriot and archipelago fragments Our data indicate no privileged source of foreigner fragments in the mouse genome. Of the 6085 foreigner fragments no mouse chromosome provided $<2.5\%$ and not $>10\%$, with the larger contributions coming mainly from the larger chromosomes. Indeed the correlation between mouse chromosome size and number of foreigner fragments aligned with it is a highly significant 0.6. This is consistent with the notion of random origins for the retroposed fragments.

Fragments in the archipelago come mostly from the same region in the mouse chromosome as the adjacent large syntenic block. More than half are within 1 Mb and a third within 100 kb. The compatriots come disproportionately from the same region of the mouse chromosome as one of the syntenic blocks on the human chromosome.

7 DISCUSSION

7.1 The origin of the fragments in the breakpoint regions

In rejecting the null hypotheses, we conclude that the fragments derive from at least three separate types of process. All or most of the foreigners but a smaller proportion of the compatriots and a much smaller proportion of the archipelago, probably come from some common processes such as retroposition of mRNA, or small jumping translocation or transposition events originating randomly across the genome and correlating roughly with chromosome size. Compatriots represent either a greater propensity for retroposition to the same chromosome originating, due to geometrical considerations (mRNA is more concentrated around the chromosome from which it is transcribed) or, in some lesser proportion, from some intrachromosomal shuffling process, such as inversion or transposition. Finally, the larger archipelago blocks seem to be ‘hived’ off the large syntenic blocks on either side, and are the results, in some proportion, of two types of process. One is the residual similarity exceeding whatever thresholds are required by the alignment algorithms. These islands of similarity ‘peeking through’ the noise may be either a natural consequence of the variable degree of similarity across all regions of the genome, or indicate the sporadic way the algorithms fail near breakpoints, or both. Second, these fragments may be chunks of the two surrounding syntenic blocks that have been thrown from near the ends of these blocks into the space by the same processes of local rearrangement that affect the interior of the blocks. That the pieces from two syntenic blocks are partially interspersed is evidence that such rearrangement continues to occur post-rearrangement, and that they are not solely the residues of decaying measures of similarity.

7.2 Predisposition for breakage versus rapid post-breakage divergence

Neither our study of the fragments in the breakpoint regions in Section 6 nor our simulation-based critique of the Pevzner–Tesler reuse inferences in Section 4 lend any credence to the idea that these regions are hotspots for major chromosomal rearrangements. Indeed, there is no *direct* evidence for the fragile regions hypothesis, aside from the well-documented tendencies for rearrangements in pericentromeric and subtelomeric regions. Clinically there may well be many recurrent sites of rearrangement, especially in somatic cell oncogenesis, but also in the germ line, generally leading to miscarriage, non-viable progeny, sterility or greatly reduced fertility. There is no systematic evidence, however, that it is these recurrent tendencies that are translated into fixed rearrangements, to say nothing about the re-usability of their breakpoints, on the evolutionary time scale, despite some suggested examples (Raphael *et al*., 2003). A few breakpoints on a region of the dog genome have been characterized as recurrent (Andelfinger *et al*., 2004), but not all of these are convincing nor are they statistically meaningful at the level of the whole genome. Bailey *et al*. (2004) have rigorously shown a high rate of co-occurrence of segmental duplication and evolutionary breakpoints and argue that this is evidence for pre-existing hotspots for rearrangement breakpoints. Their methods, however, cannot exclude the likelihood that this co-occurrence is one of cause-consequence in one direction or both (Eichler and Sankoff, 2003), and that segmental duplications and breakpoints, considered separately, are more or less randomly distributed across the genome.

If there is a paucity of direct evidence for the fragile regions hypothesis, this is even more the case for our suggestion of rapid post-breakage divergence due to an increase in various mutational processes around breakpoints in heterokaryotypic meiotic figures. Nevertheless, the latter is consistent with known mutational and population-level mechanisms.

Some of these mechanisms would involve the increased accessibility of unapposed chromosomal regions to retroposition and other mutational processes in the nuclear environment, and the decreased likelihood that these would be repaired.

Perhaps a more important role than lack of repair through recombination is the positive effect of recombinational processes in actually favouring mutation in this context. The processes leading to the formation of recombinational chiasma are complex and not completely understood. Nevertheless, these are not just accidents depending on the geometrical apposition of homologous chromosomes. The initiation of chiasma through double-stranded breaks requires the activity of several genes, the assembly of a specific protein complex in the region where the break eventually occurs and modifications in chromatin conformation (Nicolas, 2004, http://www.curie.fr/recherche/themes/detail_equipe.cfm/lang/_gb/id_equipe/23.htm). The influence of these regional processes does not necessarily stop short in the unapposed breakpoint region. Here, however, initiation of chiasma through double-stranded breaks could not lead to normally completed chiasma, greatly augmenting to the possibility of non-homologous recombination with similar sequence on the same or even a different chromosome. This in turn would result in segmental duplication, deletions and other sequence divergence.

This hypothesis of mutagenesis during heterokaryotypy predicts that the period of rapid sequence evolution occurs only in the same lineage as the genome rearrangement event, i.e. that the non-rearranged lineage should be more similar to the ancestral sequence in this region of genome. The fragile breakage model makes no such prediction and neither lineage is expected to have diverged at a faster rate. This should be testable from inferences on three or more comparable genomes.

Individuals heterozygous for the chromosomal rearrangement are likely to be partly sterile, and may therefore be under selective pressure for increased fertility. Under chromosomal models of speciation this pressure culminates in the isolation of the variant chromosome arrangements into separate lineages. Navarro and Barton (2003) recently uncovered evidence for a variant of this model that invokes positive selection acting on DNA linked to the chromosome rearrangement. Positive selection of this sort could provide a partial explanation for the decay of inter-species similarity in breakpoint regions in the cases where the rearrangement was involved in a speciation event.

8 CONCLUSION

Rearrangement breakpoints are not scattered across the genome according to a uniform probability distribution. But much as is the case of recombination sites, between the purely uniform abstraction and the concept of a very restricted number of hotspots, there lie more reasonable interpretations of the available data, where breakage is more or less likely in various local or regional contexts.

ACKNOWLEDGEMENTS

We thank Thomas Faraut and Donal Hickey for helpful discussions, and Alain Nicolas for suggesting the mutagenic possibilities of normal chiasma initiation in the breakpoint region in heterokaryotypes. Nevertheless, we assume responsibility for all the work and opinions reported here. Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

Andelfinger,G., Hitte,C., Etter,L., Guyon,R., Bourque,G., Tesler,G., Pevzner,P., Kirkness,E., Galibert,F. and Benson,D.W. (2004) Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* (in press).

Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W. and Eichler,E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Bailey,J.A., Baertsch,R., Kent,W.J., Rocchi,M., Archidiacono,N., Haussler,D. and Eichler,E.E. (2004) Hotspots of human chromosomal evolution. *Genome Biol.*, **5**, R23.

Bourque,G., Pevzner,P. and Tesler,G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, **14**, 507–516.

Eichler,E.E. and Sankoff,D. (2003) Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793–797.

Katinka,M.D., Duprat,S., Cornillot,E., Metenier,G., Thomarat,F., Prensier,G., Barbe,V., Peyretaillade,E., Brottier,P., Wincker,P. *et al*. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 450–453.

Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489

Mefford,H.C. and Trask,B.J. (2002) The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.*, **3**, 91–102.

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

Navarro,A. and Barton,N.H. (2003) Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science*, **300**, 321–324.

Nicolas,A. (2004) Molecular genetics of recombination.

Pevzner,P.A. and Tesler,G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl Acad. Sci. USA*, **100**, 7672–7677.

Raphael,B.J., Volik,S., Collins,C. and Pevzner,P.A. (2003) Reconstructing tumor genome architectures. *Bioinformatics*, (Suppl. 2), II162–II171.

Sankoff,D., Deneault,M., Turbis,P. and Allen,C. (2002) Chromosomal distributions of breakpoints in cancer, infertility, and evolution. *Theoret. Popul. Biol.*, **61**, 497–501.

Sankoff,D. and Trinh,P. (2004) Chromosomal breakpoint reuse in the inference of genome sequence rearrangement. *Proceedings of RECOMB'04, Eighth International Conference on Computational Molecular Biology*. ACM Press, New York, pp. 30–35.