Theoretical Population Biology **61**, 497–501 (2002)
doi:10.1006/tpbi.2002.1599

TPB

Chromosomal Distributions of Breakpoints in Cancer, Infertility, and Evolution

David Sankoff, Melanie Deneault, and Pascal Turbis
Centre de Recherches Mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec, Canada H3C 3J7

and

Chris Allen
Department of Molecular Genetics and Microbiology, University of New Mexico, Albuquerque, New Mexico 87131

Received January 15, 2002

We extract 11 genome-wide sets of breakpoint positions from databases on reciprocal translocations, inversions and deletions in neoplasms, reciprocal translocations and inversions in families carrying rearrangements and the human–mouse comparative map, and for each set of positions construct breakpoint distributions for the 44 autosomal arms. We identify and interpret four main types of distribution: (i) a uniform distribution associated both with families carrying translocations or inversions, and with the comparative map, (ii) telomerically skewed distributions of translocations or inversions detected consequent to births with malformations, (iii) medially clustered distributions of translocation and deletion breakpoints in tumor karyotypes, and (iv) bimodal translocation breakpoint distributions for chromosome arms containing telomeric proto-oncogenes.

Key Words: **genome rearrangement; chromosome aberrations; translocation; inversion; Nadeau–Taylor hypothesis.**

1. INTRODUCTION

Chromosome rearrangement has important consequences for speciation and phylogenetic divergence. It is also implicated in tumorigenesis and infertility. Much is known about the origins of rearrangement, for example in ionizing radiation (Friedberg *et al*., 1995), fragile sites (Smith *et al*., 1998), incorrectly repaired replication errors (Caldecott, 2001), malfunctions in topoisomerase activity (Aplan, 1999), faulty meiotic recombination (Bishop and Schiestl, 2000; Emanuel and Shaikh, 2001), and somatic recombination mistakes in B-cells (Kuppers *et al*., 1999).

Implicit in much statistical and algorithmic analysis (Sankoff and Nadeau, 2000; Sankoff and El-Mabrouk, 2002) of genome evolution is the Nadeau–Taylor hypothesis (Nadeau and Taylor, 1984), that genome rearrangements such as translocations and inversions occur at random sites along the length of the chromosome, but this can generally only be tested indirectly, using the pattern of breakpoints on a comparative map (Sankoff *et al*., 1997; International Human Genome Sequencing Consortium, 2001, Figs. 46–48). It is generally difficult to reconstruct *in an unambiguous manner* the actual series of overlapping rearrangements that produced this pattern, despite great progress in solving minimum rearrangements problems (Sankoff and El-Mabrouk, 2002).

On the other hand, there is much data available on rearrangements in clinical contexts. Can the patterns of pathological constitutional or somatic rearrangements in human genomes provide insight into evolutionary processes, despite the vastly different time scales

0040-5809/02 $35.00

involved? There have been many studies quantitatively relating the frequency of clinically significant breakpoints to local or intensional properties of chromosomal maps or genome sequence (Cohen, *et al.*, 1996), but here we initiate the extensional characterization of genomic breakpoints, namely their spatial distribution along chromosomal arms.

In this note, we analyze data on rearrangement breakpoints resulting from individual real-time cytogenetic events in order to help understand the distribution of multiple breakpoints in comparative maps. We compare breakpoint positions from four different databases, on reciprocal translocations, inversions and deletions in neoplasms, reciprocal translocations and inversions in families carrying rearrangements and the human–mouse comparative map. For each set of positions, we construct breakpoint distributions for as many as possible of the 44 autosomal arms. We identify and interpret four main types of distribution:

- the uniform distribution associated both with families carrying translocations or inversions, and with the comparative map,
- telomerically skewed distributions of translocations or inversions detected consequent to births with malformations,
- medially clustered distributions of translocation and deletion breakpoints in tumor karyotypes, and
- bimodal translocation breakpoint distributions for chromosome arms containing telomeric proto-oncogenes.

2. METHODS

We use data drawn from four sources, the Mitelman Database of Chromosome Aberrations in Cancer (Mitelman *et al.*, 2001), based on some 40,000 individual cases or associations manually culled from the literature, the human cytogenetics (HC) Forum data (Cohen *et al.*, 2001) on over 4000 families carrying constitutional and other genetic rearrangements, another data set we constructed on constitutional rearrangements from studies on sperm abnormalities (Guttenbach *et al.*, 1997; Shi and Martin, 2001) and the Human–Mouse Homology Map (2001).

Of all the reports in the Mitelman database of recurrent aberrations as revised July 20, 2001, we extract data only on single translocations, inversions and deletions, plus compound operations involving derived chromosomes whenever we can unambiguously identify breakpoints of their component translocations, inversions and deletions. Each report indicates a number of cases, and our data consist of sums of all the cases for each breakpoint.

To complement the HC Forum data, we assembled a small, but independent dataset based on literature surveys by Guttenbach *et al.* (1997) and by Shi and Martin (2001) of chromosomal abnormalities in the sperm of 62 carriers of constitutional translocations. We do not analyze the abnormalities *per se* but simply use the 62 translocations themselves as our sample.

We used the comparison between the current NCBI assembly of the human genome (Build 25) and the MGD genetic map (Blake *et al.*, 2001) for the mouse, as available on the NCBI website (Human–Mouse Homology Map, 2001). We use the breakpoints as reconstructed on the NCBI site. Because these breakpoints are sometimes situated between two genes which are only approximately mapped cytogenetically, there remains some uncertainty in a few of our assignments, but this has negligible effect on our estimation of the breakpoint distribution parameters.

Within each chromosome arm, we count the number of breakpoints per band, since only this level of resolution is available for all the data sets. The band positions and widths were obtained courtesy of the NCBI information service. In estimating the parameters, the breakpoints in each band are treated as if they all fell at the mid-point of the band.

3. RESULTS

Following Cohen *et al.* (1996), we constructed the distribution of rearrangement breakpoints on each autosome arm for each source of data and for each type of rearrangement event, 11 sets of distributions in all. To illustrate, Fig. 1 shows the distributions of translocation breakpoints on chromosome 9, for the data drawn from the Mitelman and HC Forum databases.

According to the Nadeau–Taylor hypothesis (Nadeau and Taylor 1984), rearrangement breakpoints should occur according to a uniform probability density along the length of the chromosome. Identifying the chromosome arm from the centromere to the telomere with the interval $[0, 1]$, this hypothesis establishes a mean breakpoint position $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12}$.

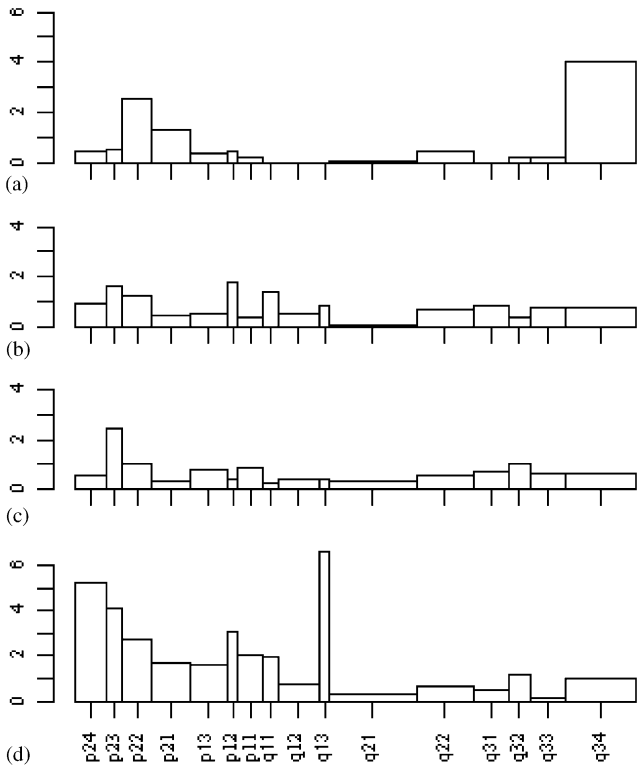

FIG. 1. Distributions of breakpoints on chromosome 9: (a) reciprocal translocations in neoplasms; (b) in infertility; (c) discovered fortuitously and (d) malformed births. Length of band x proportional to estimated length in base pairs. Number of breakpoints n proportional to area of histogram bar: $n = xy \times 10^{-6}$, where y is the vertical axis score.

For each of the 11 data sets, Table I presents average $\hat{\mu}$ and $\hat{\sigma}^2$ over all 44 autosomal arms.

3.1. The Mitelman Database

For the cancer data, $\bar{\hat{\mu}}$ for each type of rearrangement is near $\frac{1}{2}$, consistent with a uniform distribution, but $\bar{\hat{\sigma}}^2 < \frac{1}{12}$, significantly so for deletions and translocations. This indicates that the breakpoints tend to be clustered in the median region of the arms, as in the p arm of chromosome 9 in Fig. 1a. Indeed, 35 of 36 deletion distributions have variances less than the uniform, as do 27 of 43 translocation distributions. There is no obvious explanation of the arm–medial concentration, either as a data acquisition bias, or as a genuine biological effect. Since oncogenesis often involves the disruption or establishment of gene adjacencies, breakpoint frequency might be positively correlated with gene density, but the opposite is true (International Human Genome Sequencing Consortium, 2001, p. 908), and there is no clear evidence of greater gene density in the arm–medial region in any case (Venter *et al.*, 2001, Fig. 11).

The translocation distributions on several of the chromosomal arms (e.g. 6q, 8q, 14q, 19q, 20p, 21q, 9q as seen in Fig. 1a) are characterized by concentrations of breakpoints at the most distal band (which includes the telomere), giving the distributions a bi-modal appearance. This can be explained by the location of several proto-oncogenes near telomeric breakpoints. Oncogene activation via reciprocal translocation at these sites provides a powerful stimulus in neoplastic transformation and tumorigenesis, e.g., ABL in 9q34, which through t(9;22)(q34;q11) gives rise to the "Philadelphia chromosome" implicated in acute myeloid leukemia and chronic myelogenous leukemia (Apaln, 1999). Neoplastic transformation can also occur by bringing proto-oncogenes to loci that result in inappropriately high levels of gene expression. In Burkitt's lymphoma, reciprocal translocations appose the transcription factor *C-MYC* at 8q24 and various immunoglobulin gene loci (Kuppers *et al.*, 1999), resulting in increased cell proliferation, and ultimately, transformation.

3.2. The HC Forum Database

That a family carries a constitutional rearrangement can be discovered fortuitously, such as through routine screening, or as a result of infertility testing or after a malformation at birth. It is known (Cohen *et al.*, 1996) that for translocations detected as a result of malformed births, there is a bias toward a distal, or telomeric distribution. This is seen in Fig. 1d for translocations of chromosome arm 9p and in Table I for translocations and inversions. As illustrated in Fig. 1b and c, no such bias occurs for rearrangements discovered fortuitously or through fertility testing, and Table I confirms that the apparently uniform distribution of these classes (with the exception of inversions discovered fortuitously), with mean close to $\frac{1}{2}$ and variance almost $\frac{1}{12}$, is consistent with the Nadeau–Taylor hypothesis.

3.3. The Guttenbach–Martin Carriers

Data from the Guttenbach *et al.* (1997) and Shi and Martin (2001) surveys on a total of 62 normal male carriers, confirm the uniformity found in Section 3.2.

3.4. The Comparative Map

The effect of chromosomal location on breakpoint frequency cannot generally be ascertained from com-

TABLE I

Average Means and Variances for 11 Breakpoint Distributions

	$\bar{\mu}$	Breakpoints	Arms	$\bar{\sigma}^2$	Arms
Mitelman					
Deletion	0.436 ± 0.105	4464	36	0.044 ± 0.019	36
Inversion	0.494 ± 0.212	176	22	0.058 ± 0.069	21
Translocation	0.471 ± 0.153	1794	44	0.069 ± 0.037	43
HC forum					
Translocation					
Infertility	0.516 ± 0.083	2086	44	0.073 ± 0.019	44
Fortuitous	0.501 ± 0.086	1702	44	0.077 ± 0.022	43
Malformations	0.608 ± 0.118	2480	44	0.065 ± 0.027	44
Inversion					
Infertility	0.424 ± 0.162	370	38	0.077 ± 0.076	30
Fortuitous	0.384 ± 0.172	708	44	0.056 ± 0.041	40
Malformations	0.647 ± 0.206	204	36	0.034 ± 0.035	28
Guttenbach–Martin	0.538	124	36	0.107	32
Human–Mouse					
Homology	0.491	190	37	0.075	33

Note. Because of sparse data in the Guttenbach–Martin and homology databases, normalized breakpoint positions on All arms were carried to a common $[0, 1]$ interval and a single estimate of μ and σ^2 was made.

parative maps because the current location of an ancient breakpoint will not be the same as when it actually occurred. Nevertheless, there may well be a relevant signal in current breakpoint positions; e.g., were near-telomeric position a necessary condition for fixation of a translocation in a population, we would expect a high density of breakpoints in a comparative map near the telomeres and a low density elsewhere. Table I shows, however, that the data are quite consistent with the uniform distribution.

4. DISCUSSION

What do the four different patterns tell us about the relationship between cytogenetic processes and evolution? First, the cancer data establish a new and unexplained fact about arm–medial concentrations of breakpoints, and serve as a control to prove that our method can detect distributions other than uniform. But these somatic cell rearrangements do not have any direct implications for evolution. Neither do the telomeric distributions associated with malformed births, since these are a pathological, usually barely viable, consequence of incorrect segregation of quadrivalent and other abnormal meiotic figures. It is the uniform distribution of breakpoints in normal carriers of translocations and inversions which is of most importance for evolution, since it offers a principled justification of the Nadeau–Taylor hypothesis.

Whatever the mitotic barriers to propagation of translocations and inversions, and whatever the selective pressures at the fertilization and developmental levels, it seems clear that at the population level, there is no detectible departure from the uniform distribution of breakpoints in these carriers. Thus, in the extremely rare occurrence (once per 10^6 or 10^7 years), via small population size, inbreeding patterns, or other circumstance, that a rearrangement be fixed in a population, we have no reason to expect the breakpoints to occur other than at uniformly random positions on the chromosome arms affected.

ACKNOWLEDGMENTS

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. D.S. is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

Aplan, P. D. 1999. Mechanisms of leukemogenesis: Chromosomal translocations, Hematology 1999. American Society of

Hematology, pp. 77–82. http://www.hematology.org/education/hematology99.cfm.

Blake, J. A., Eppig, J. T., Richardson, J. E., Bult, C. J., Kadin, J. A., and the Mouse Genome Database Group 2001. The Mouse Genome Database (MGD): Integration nexus for the laboratory mouse, *Nucl. Acids Res.* **29**, 91–94.

Bishop, A. J., and Schiestl, R. H. 2000. Homologous recombination as a mechanism for genome rearrangements: Environmental and genetic effects, *Hum. Mol. Genet.* **9**, 2427–2434.

Caldecott, K. W. 2001. Mammalian DNA single-strand break repair: An X-ra(y)ted affair, *Bioessays* **23**, 447–455.

Cohen, O., *et al.* 1996. Cartographic study: Breakpoints in 1574 families carrying human reciprocal translocations, *Hum. Genet.* **97**, 659–667.

Cohen, O., Mermet, M. A., and Demongeot, J. 2001. HC Forum: A web site based on an international human cytogenetic database, *Nucl. Acids Res.* **29**, 305–307. http://HCForum.imag.fr/.

Emanuel, B. S., and Shaikh, T. H. 2001. Segmental duplications: An 'expanding' role in genomic instability and disease, *Nat. Rev. Genet.* **2**, 791–800.

Friedberg, E. C., Walker, G. C., and Siede, W. 1995. "DNA Repair and Mutagenesis" ASM Press, Washington, DC.

Guttenbach, M., Engel, W., and Schmid, M. 1977. Analysis of structural and numerical chromosome abnormalities in sperm of normal men and carriers of constitutional chromosome aberrations. A review, *Hum. Genet.* **100**, 1–21.

Human–Mouse Homology Map, 2001. http://www.ncbi.nlm.nih.gov/Homology/.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome, *Nature* **409**, 860–921.

Kuppers, R., Klein, U., Hansmann, M. L., and Rajewsky, K. 1999. Cellular origin of human B-cell lymphomas, *N. Eng. J. Med.* **341**, 1520–1529.

Mitelman, F., Johansson, B., and Mertens, F. (Eds.), 2001. Mitelman database of chromosome aberrations in cancer, http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Sankoff, D., and El-Mabrouk, N. 2002. Genome rearrangement. *in* "Current Topics in Computational Molecular Biology" (T. Jiang, Y. Xu, and M. Q. Zhang Eds.), pp. 135–155, MIT Press, Cambridge, MA.

Sankoff, D., and Nadeau, J. H. 2000. "Comparative Genomics," Kluwer Academic Press, Dordrecht, NL.

Sankoff, D., Parent, M.-N., Marchand, I., and Ferretti, V. 1977. On the Nadeau–Taylor theory of conserved chromosome segments, *in* "Combinatorial Pattern Matching" [(A. Apostolico, J. Hein, Eds.),] Eighth Annual Symposium, Lecture Notes in Computer Science, Vol. 1264, pp. 262–274, Springer-Verlag, Berlin.

Smith, D. I., Huang, H., and Wang, L. 1998. Common fragile sites and cancer, *Int. J. Oncol.* **12**, 187–196.

Shi, Q., and Martin, R. H. 2001. Aneuploidy in human spermatozoa: FISH analysis in men with constitutional chromosomal abnormalities, and in infertile men, *Reproduction* **121**, 655–666.

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.

Venter, J. C., *et al.* 2001. The sequence of the human genome, *Science* **291**, 1304–1351.