Chromosome rearrangements in evolution: From gene order to genome sequence and back

David Sankoff* and Joseph H. Nadeau†‡

**Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, ON, Canada K1N 6N5; and †Department of Genetics, Center for Computational Genomics, and Center for Human Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106*

Since Sturtevant's 1926 genetic proof of inversions (1) and Wright and Haldane's discovery of conserved linkages (2), classical geneticists have compared pairs of linkage maps to infer chromosomal rearrangements and define evolutionarily conserved chromosomal segments. The data have been limited to whatever sample of the gene complement has been mapped in both sequences, and by statistical uncertainty of gene order and map distances. It is a far cry from the beads-on-a-string maps underlying, for example, the first systematic comparative mapping in mammals (3) to today's comprehensive primary data for documenting the structural evolution of mammalian genomes: billions of base pairs of nearly complete genomic sequences, with their vast intergenic distances, highly dispersed upstream and downstream regulatory elements, overlapping genes, alternative exons, somatic rearrangements, paralogs, gene families large and small, apparently randomly scattered pseudogenes, massive assembly-defeating arrays of repetitive sequences of all sorts, great volumes of transposon-inserted material, and confusing recent accretions of highly paralogous material in subtelomeric (4) and pericentromeric regions (5). Making sense of these diverse elements in the genome, understanding how they are organized within the genome of each species, and characterizing the changes in genome organization during evolution are critical problems in comparative genomics. Kent *et al*. (6), in this issue of PNAS, and Pevzner and Tesler (7, 8) now add new insight into our knowledge of evolutionarily conserved genome structure.

We are obliged to greatly expand the neat repertoire of classical evolutionary processes affecting genome structure: inversion and reciprocal translocation; chromosome fusion and fission; gene, segment, and chromosomal duplication and loss; polyploidization [even in mammals (9)] and return to diploidy, to account not only for the various highly productive mechanisms for inserting external material, for the proliferation of repetitive sequencing, for massive ongoing sequence conversion, e.g., in the Y chromosome (10), but also for the complex patterning of frequency and size of the chromosomal fragments involved in the more traditional processes. Even if we wish to describe only the major structural changes that differentiate two species, genomic sequence alignment to find corresponding orthologous segments must overcome a bewildering inventory of very short segments and local rearrangements, multiply aligned regions, and unaligned zones, which represent the "noise" in this analysis, although they are, of course, of vital interest in their own right. Existing methodology for discerning conserved segments in gene orders despite small rearrangements and error (11) are grossly inadequate for "cleaning up" comparisons at the genome sequence level.

Making sense of diverse elements in the genome is a critical problem in comparative genomics.

Kent *et al.* (6) have carried out a careful alignment study of the human genome sequence with the draft sequence for the mouse, building up syntenic blocks from alignments while allowing long overlapping gaps in the genome sequence for each species, a simple but very clever device that minimizes the methodological difficulties caused by embedded inversions, tandem repeats, short transpositions, and transposed pseudogenes. Some 344 of these blocks longer than 100 kb represent the "primary units of conserved synteny" between human and mouse. Previously, Pevzner and Tesler (7, 8) adopted another stratagem: leap-frogging the comprehensive alignment step by relying on almost 600,000 relatively short [average length, 340 bp (7)] anchors of highly aligned sequence fragments as a starting point for building blocks of conserved synteny, and amalgamating neighboring subblocks by using a variety of criteria to avoid disruptions due to "microrearrangements" <1 Mb. This procedure eventually succeeds in inferring a set of 281 blocks >1 Mb, which is comparable to the result in ref. 6 and indeed to the most recent results [somewhat more than 200 in the current National Center for Biotechnology Information Human-Mouse Homology Map (www.ncbi.nlm.nih.gov/Homology/ComMapDoc.html)] of the comparative mapping and other "pregenome sequence" approaches. Pevzner and Tesler go a step further and use the order of these large blocks on the chromosomes as input to an improved adaptation (12) of the gene order rearrangement algorithms, originally devised by Hanennhalli and Pevzner (13), to reconstruct aspects of the actual sequence of inversions and translocations that account for the divergent structures of the two genomes.

A common emphasis of Kent *et al.* and Pevzner and Tesler is the high frequency of small sequence level rearrangements: inversions, transpositions, and duplications of both local and remote small segments, although the two analyses actually refer to different levels of resolution. The latter refers only to some 3,000 microrearrangments of fragments <200 kb in size, many of which they attribute to assembly errors, whereas the former counts $>100{,}000$ apparent disruptions not accessible to the anchor-based technology of ref. 8, involving much smaller fragments (length from 100 to 1,000 bases), carefully controlled by comparison of draft with finished mouse sequence. These authors also report 255 additional rearrangements larger than 100 kb, mostly inversions within the 344 syntenic blocks. The apparent proliferation of small rearrangements ties in with previous results of Wolfe and colleagues on short inversions in gene order in eukaryotes such as yeast (14) and *Caenorhabditis* (15).

Although the levels of resolution of the two human–mouse comparisons are very different, both suggest that the random breakpoint model implicit in ref. 3 should be modified, although both agree that the length distributions of the major syntenic blocks are consistent with this model. In ref. 6, it is suggested that the distribution

See companion articles on pages 11484 in this issue and 7672 in issue 13 of volume 100.

‡To whom correspondence should be addressed. E-mail: jhn4@cwru.edu.

of short (<100 Kb) blocks not part of the primary synteny correspondence are generated by some process other than random placement of breakpoints. Unfortunately, this initial report did not provide data that could indicate why or what type of process. In ref. 8, it is conjectured that a small proportion of the genome is made up of fragile regions with a much higher susceptibility to rearrangement than the rest of the genome, and, in ref. 6, it is reported that the 344 long chains are often separated from each other in a genome by long runs of shorter segments (20,000 in all) presumably aligning with a number of different chromosomes in the other genome. Again, we must await fuller presentations of the data to get a detailed quantitative impression of the content, organization, and evolution of these particular segments.

Large-scale data analyses are facilitated by the adoption of arbitrary thresholds dividing macro- (>1 Mb) from microrearrangements (<1 Mb) (8), or long chains (>100 Kb) from short ones (<100 Kb) (6). Whereas these choices may be imposed for methodological considerations, it should be remembered that they have direct consequences on the key results of the analysis. Lower thresholds immediately increase the number of syntenic blocks considered significant and decrease the amount of the genome that is considered highly rearranged. Hopefully, further experience with genome comparison at the sequence level will enable us to express these central notions in a less arbitrary manner.

Nevertheless, the discovery of short genomic regions that align successively to several chromosomes in the other genome is as fundamental a finding as the documentation of the many minor disruptions in the long blocks of conserved synteny. That many of these short regions separate long blocks is suggestive (and this is the main point of ref. 8) that chromosomal neighborhoods around evolutionary breakpoints may be particularly prone to rearrangements although no independent type of evidence has been developed to distinguish these regions from the rest of the genome. However, the data presented in ref. 6 indicate that these regions contain sites for high rates of retroposon activity (much of it involving processed pseudogenes), incorporation of non-tandemly duplicated genes, and more general expansion of lineage-specific gene families. Statistically, there will be occasional regions where orthology decays unusually rapidly, allowing a variety of paralogies to be picked up by the alignment. Indeed, whether anything more than a minuscule proportion of the 20,000 short chains reflects major interchromosomal exchanges such as reciprocal translocation is most unlikely, given what is more directly observable about such processes among closely related genomes. Of course, this only displaces the questions about these regions to another level: Why are they receptive to such activity? Why are they associated with evolutionary breakpoints, if indeed they are? (According to ref. 6, there are many more such regions embedded *within* syntenic blocks.)

This work establishes computational technology for studying the evolution of gene order in genome sequences.

One of the most exciting aspects of this new work is the prospect of adapting the established computational technology for studying the evolution of gene order so that it is applicable to genome sequence. Until recently, these methods could be applied only to small genomes (mitochondria, chloroplasts, and prokaryotes), the difficulty with eukaryotic nuclear genomes being not so much the computational cost but rather the absence of comprehensive lists of genes and their orthologs. Pevzner and Tesler's work shows how to effectively bypass the gene finding and ortholog identification steps, by using the order of syntenic blocks as input to the rearrangement algorithm. In ref. 8, however, the use of large unresolved blocks and the (methodologically largely unavoidable) neglect of blocks shorter than 1 Mb may raise doubts about the interpretation of the output rearrangements because important parts of the historical derivation of the genomes may be blurred. In particular, the consequences of these practices for the breakpoint "re-use" calculations are clearly an important area for further mathematical and statistical research. In the future, as the genome assemblies become increasingly refined, the threshold size for syntenic blocks may be lowered, or threshold-independent measures may be developed (as in ref. 3) so that a more complete proportion of the small segments and hence of the evolutionary breakpoints is considered. The methodological difficulty will then be to ensure that duplicated genes, converted segments, pseudogenes, and other retrotranspositions are not forced into the same mold as inversions and reciprocal translocations.

Comparative genomics can look forward to further results of the research projects for which refs. 6 and 8 are among the initial publications. The extension from gene maps to the sequence samples (the anchors in ref. 8) and to the complete genome sequences in ref. 6 represents a breakthrough, but one that leads to more questions than answers. Quantitative inferences about all of the processes disrupting the alignment of the two genomes should be a research priority, beyond statistics on alignment lengths. Another question will be the relationship between sequence rearrangement and gene order rearrangement. Almost all of the short chains identified in ref. 6 and some of the long syntenic blocks likely contain no genes, so whether they are inverted, transposed, duplicated, or deleted will have little consequence for gene order. Until these questions are investigated, the pertinence of sequence rearrangements to gene order rearrangements and the functionality of gene adjacencies will remain open questions.

1. Sturtevant, A. H. (1926) *Biol. Zentralbl* **46,** 697–702.
2. Ohno, S. (1973) *Nature* **244,** 259–262.
3. Nadeau, J. H. & Taylor, B. A. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 814–818.
4. Mefford, H. C. & Trask, B. J. (2002) *Nat. Rev. Genet.* **3,** 91–102.
5. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. & Eichler, E. E. (2002) *Science* **297,** 1003–1007.
6. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 11484–11489.
7. Pevzner, P. & Tesler, G. (2003) *Genome Res.* **13,** 37–45.
8. Pevzner, P. & Tesler, G. (2003) *Proc. Natl. Acad. Sci. USA*. (2003) **100,** 7672–7677.
9. Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A. & Kohler, N. (1999) *Nature* **401,** 341.
10. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., *et al.* (2003) *Nature* **423,** 825–837.
11. Sankoff, D., Ferretti, V. & Nadeau, J. H. (1997) *J. Comp. Biol.* **4,** 559–565.
12. Tesler, G. (2002) *Bioinformatics* **18,** 492–493.
13. Hanennhalli, S. & Pevzner, P. A. (1995) in *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science* (Washington, DC), pp. 581–592.
14. Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14433–14437.
15. Coghlan, A. & Wolfe, K. H. (2002) *Genome Res.* **12,** 857–867.