17 Bishop, G.J. *et al.* (1996) *Plant Cell* 8, 959–969
18 Nomura, T. *et al.* (1997) *Plant Physiol.* 113, 31–37
19 Yokota, T. *et al.* (1997) *Proceedings 24th Annual Meeting The Plant Growth Regulation Society of America* 8–12 August 1997, Atlanta, p. 94
20 Klahre, U. *et al.* (1997) *Plant Cell* 10, 1677–1690
21 Azpiroz, R. *et al.* (1998) *Plant Cell* 10, 219–230
22 Feldmann, K.A. *et al.* (1989) *Science* 243, 1351–1354
23 Mushegian, A.R. and Koonin, E.V. (1995) *Protein Sci.* 4, 1243–1244
24 Hartmann, M-E. (1998) *Trends Plant Sci.* 3, 170–175
25 Fujioka, S. *et al.* (1997) *Plant Cell* 9, 1951–1962
26 Li, J. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 3554–3559
27 Mathur, J. *et al. Plant J.* (in press)
28 Bishop, G.J. *et al.* (1998) *J. Exp. Bot.* 49 (Suppl.), 66
29 Cosgrove, D.J. (1997) *Plant Cell* 9, 1031–1041
30 Behringer, F.J. *et al.* (1990) *Plant Physiol.* 94, 166–173
31 Zurek, M.D. and Clouse, S.D. (1994) *Plant Physiol.* 104, 161–170
32 Zurek, D.M. *et al.* (1994) *Plant Physiol.* 104, 505–513
33 Mayumi, K. and Shibaoka, H. (1995) *Plant Cell Physiol.* 36, 173–181
34 Chory, J. *et al.* (1989) *Cell* 58, 991–999
35 Deng, X-W. *et al.* (1991) *Genes Dev.* 5, 1172–1182
36 Chory, J. *et al.* (1991) *Plant Cell* 3, 445–459
37 Hewitt, F.R. *et al.* (1985) *Aust. J. Plant Physiol.* 12, 201–211
38 Clouse, S.D. and Zurek, D.M. (1991) in *Brassinosteroids: Chemistry, Bioactivity and Applications* (Cutler, H.G., Yokota, T. and Adam, G., eds), pp. 122–140, ACS Symp. Series, American Chemical Society, Washington, DC
39 Iwasaki, T. and Shibaoka, H. (1991) *Plant Cell Physiol.* 32, 1007–1014
40 Fukuda, H. (1997) *Plant Cell* 9, 1147–1156
41 Clouse, S.D. *et al.* (1996) *Plant Physiol.* 111, 671–678
42 Li, J. and Chory, J. (1997) *Cell* 90, 929–938
43 Fantl, W.J. *et al.* (1993) *Annu. Rev. Biochem.* 62, 453–481
44 Beato, M. *et al.* (1995) *Cell* 83, 851–857
45 Thummel, C.S. (1995) *Cell* 83, 871–877
46 Thummel, C.S. (1996) *Trends Genet.* 12, 306–310
47 McEwen, B.S. (1991) *Trends Pharmacol. Sci.* 12, 141–147
48 Revelli, A. *et al.* (1998) *Endocrine Rev.* 19, 3–17
49 Clark, S.E. *et al.* (1997) *Cell* 89, 575–585
50 Torii, K.U. *et al.* (1996) *Plant Cell* 8, 735–746
51 Song, W-Y. *et al.* (1995) *Science* 270, 1804–1806
52 Kobe, B. and Deisenhofer, J. (1994) *Trends Biochem. Sci.* 19, 415–421
53 Koncz, C. (1998) *Trends Plant Sci.* 3, 1–2

T. Altmann *is at the Max-Planck-Institut für molekulare Pflanzenphysiologie, Karl-Liebknecht-Strasse 25, 14476 Golm, Germany.*

Counting on comparative maps

JOSEPH H. NADEAU (jhn4@po.cwru.edu)

DAVID SANKOFF (sankoff@ere.umontreal.ca)

Comparative maps record the history of chromosome rearrangements that have occurred during the evolution of plants and animals. Effective use of these maps in genetic and evolutionary studies relies on quantitative analyses of the patterns of segment conservation. We review the analytical methods that have been developed for characterizing these maps and evaluate their application to existing comparative maps mainly for plants and animals.

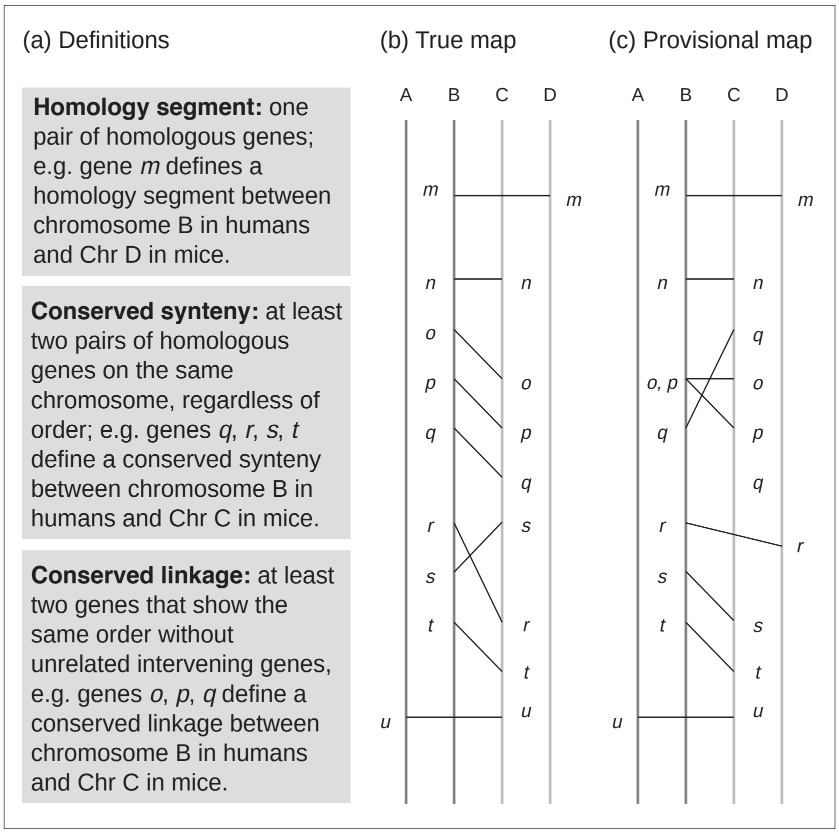
Comparative gene mapping is based on the observation that genes that are closely linked in one species tend to be closely linked in other species, whereas loosely linked genes in one species tend to be unlinked in related species. In the absence of chromosome rearrangements, linkages and gene order are preserved. During evolution, chromosome rearrangements invariably occur, disrupting some, but not all, ancestral linkages. The probabilities that linkage or synteny will be conserved or disrupted depend on rearrangements that have accumulated since divergence of the lineages leading to the two species being considered. Closely related species have usually accumulated fewer rearrangements and therefore have many long conserved segments, whereas distantly related species have usually accumulated more rearrangements and have many short conserved segments. These tendencies have exceptions, however, with some phylogenetic lineages showing remarkable conservation and others extensive chromosome rearrangements. This mosaic of conserved and rearranged segments, which are revealed by comparison of the chromosomal location of homologous genes in different species, provide insight into some of the forces guiding genome organization and evolution[1–3].

The challenge has been to develop objective, algorithmic methods for identifying the content of conserved segments, measuring the extent of segment conservation, estimating the rates and patterns of chromosome rearrangements in various lineages during evolution, and reconstructing the organization of ancestral genomes. The purpose of this review is to raise awareness of these quantitative issues and to assess the progress that has been made in developing and applying analytical methods.

Identifying conserved genes

Homologous genes are landmarks needed to relate corresponding chromosome segments in different species and, as such, they are one of the two key elements in comparative mapping. Standard criteria for identifying homologous genes have been used for many years[4]. Recently, the sequence similarity for nearly 3000 genes in humans, mice and rats was determined[5,6]. Homologies over longer evolutionary periods are also being recognized[7]. However, for certain genes the judgement of

0168-9525/98/$ – see front matter © 1998 Elsevier Science All rights reserved.
PII: S0168-9525(98)01607-2



Homology segment: one pair of homologous genes; e.g. gene *m* defines a homology segment between chromosome B in humans and Chr D in mice.

Conserved synteny: at least two pairs of homologous genes on the same chromosome, regardless of order; e.g. genes *q*, *r*, *s*, *t* define a conserved synteny between chromosome B in humans and Chr C in mice.

Conserved linkage: at least two genes that show the same order without unrelated intervening genes, e.g. genes *o*, *p*, *q* define a conserved linkage between chromosome B in humans and Chr C in mice.

FIGURE 1. Comparative maps for human chromosomes A and B, and mouse chromosomes C and D involving genes *m*, ..., *u*. (a) Definitions of relevant comparative mapping terminology with examples from the true map illustrated in (b). (c) Shows the provisional map with correct locations, errors and ambiguities. In the true map, the conserved segment marked by genes *n*, ..., *t* are bounded by genes m and u. In the evolving map, *q* has been mapped to the wrong location in humans or mice, *r* has been mapped to the wrong mouse chromosome (D instead of C), and the relative position of [*o*, *p*] is ambiguous. These anomalies can result either from legitimate new mapping information that correctly reflects the true genetic and comparative maps or alternatively from incorrect homology determinations and mapping errors that place the gene on the wrong chromosome or at the wrong location with respect to other genes on the correct chromosome. In the absence of definitive information, e.g. complete DNA sequence of the relevant genomes, some anomalies are difficult to resolve.

experts is critical in homology determination. This combination of objective measures of homology together with expert knowledge has led to high-quality lists of homologous genes in various databases[8–14].

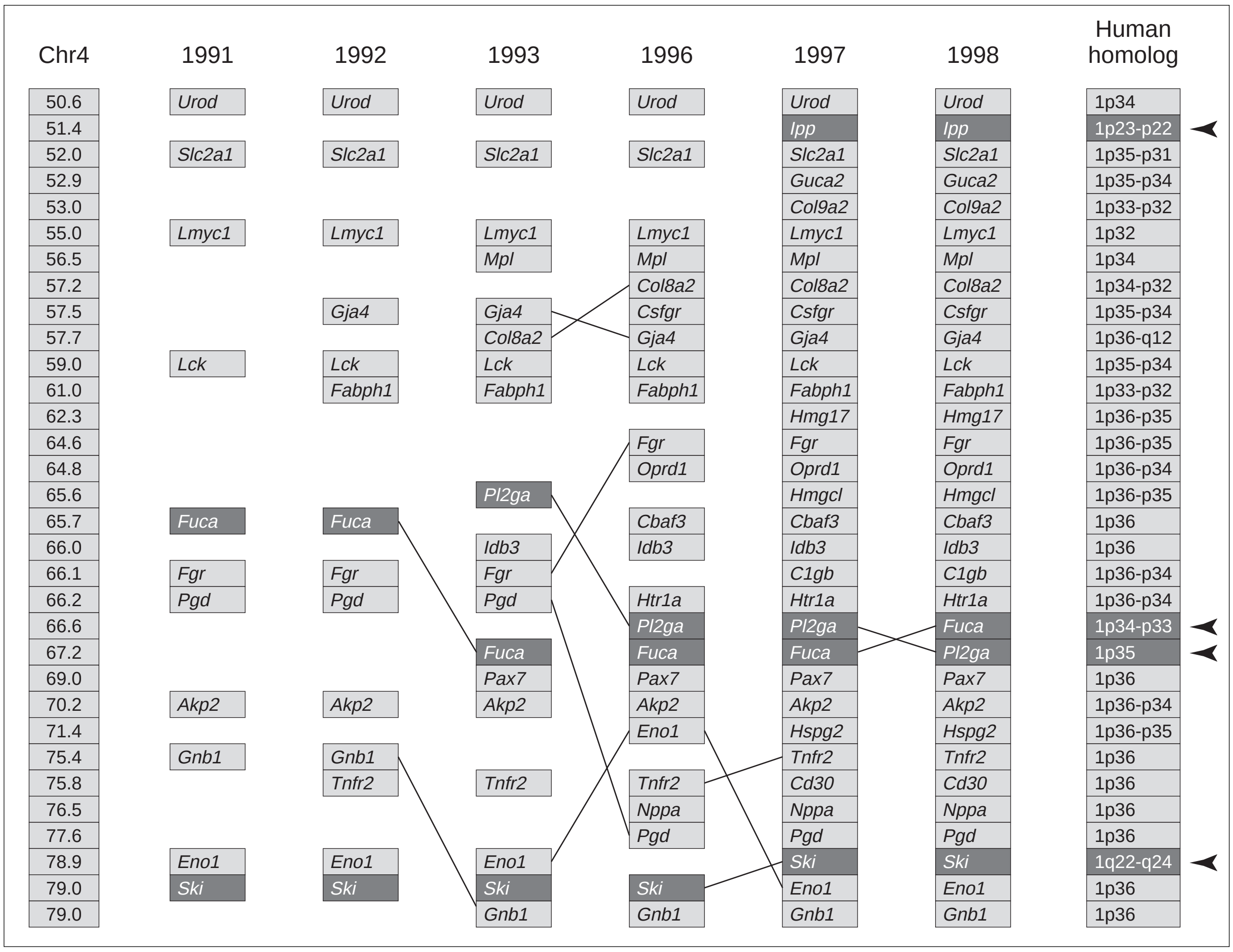
The advantages of genes as landmarks in comparative mapping include homologies that are usually unambiguous, large numbers of markers (e.g. as many as ~75 000 to ~100 000 genes in the genomes of humans and mice), and precise localizations. The resolution of comparative maps is illustrated by the observation that the mapping of a pair of homologous genes in at least two species can suffice as evidence for a conserved segment (the homology segment in Fig. 1). As more genes are mapped to the segment, the extent of map conservation can be determined. The disadvantage of genes to track patterns of genome evolution is that each must be individually mapped in the species of interest. An important caveat is that genome polyploidization, which, though rare in comparison with other genomic changes, is phylogenetically widespread[15–17], especially in fish, amphibian and plant lineages, has complicated identification of true homologs among families of unlinked genes.

Addition of new genes to comparative maps has largely depended on the vagaries of mapping interests among geneticists. Because most genes have not been mapped in more than one species, few landmarks are available for identifying homologous segments in different species. At one time for example, 4344 genes had been mapped in humans, 5554 in mice, but only 1889 in both species[18]. However, important steps are being made towards defining comparative maps more systematically[19]. Genetic markers, such as comparative anchored tagged sites or CATS (Ref. 20) and ESTs (Ref. 21), are being developed that are easy to map and whose homology in different species is readily established.

Identifying conserved segments

The second key element in comparative mapping is identifying conserved segments. Translocations, inversions, transpositions and other less common kinds of chromosome rearrangements disrupt ancestral linkages and syntenies. By comparing the location of homologous genes in different species, we can try to determine whether a particular chromosome segment has been conserved or disrupted during evolution. Despite the simplicity of the notion that a conserved segment is a maximally contiguous chromosomal region with identical gene content and order in the two species that are being compared, the operational identification of conserved segments is not simple.

Conserved segments are evident among diverse evolutionary lineages. The evidence for mammals is considerable[11]. Evidence for conservation has also been found among certain species of fish as well as between these fish and mammals[22–24]. For pufferfish, homologs of many human and mouse genes have been mapped with some segments showing strong conservation[25], other segments showing small local rearrangements[26], and still others showing disrupted linkage[27]. Examples of segment conservation extending from humans to fruit flies and nematodes have been reported[28]. Comparative maps have also been made for many plants[29–33], including several grasses[31,32], tomato and potato[34], maize, rice and wheat[35,36], sorghum and maize[37], pea and lentil[38], oat and wheat[39,40], pepper and tomato[41], and *Arabidopsis* and *Brassica*[42]. Although some of these maps are composed of a modest number of genes, their definition is rapidly improving and general patterns are beginning to emerge. Both long and short conserved segments as well as rearranged segments are evident. A remarkable



Chr4	1991	1992	1993	1996	1997	1998	Human homolog	
50.6	*Urod*	*Urod*	*Urod*	*Urod*	*Urod*	*Urod*	1p34	
51.4					*Ipp*	*Ipp*	1p23-p22	◄
52.0	*Slc2a1*	*Slc2a1*	*Slc2a1*	*Slc2a1*	*Slc2a1*	*Slc2a1*	1p35-p31	
52.9					*Guca2*	*Guca2*	1p35-p34	
53.0					*Col9a2*	*Col9a2*	1p33-p32	
55.0	*Lmyc1*	*Lmyc1*	*Lmyc1*	*Lmyc1*	*Lmyc1*	*Lmyc1*	1p32	
56.5			*Mpl*	*Mpl*	*Mpl*	*Mpl*	1p34	
57.2				*Col8a2*	*Col8a2*	*Col8a2*	1p34-p32	
57.5		*Gja4*	*Gja4*	*Csfgr*	*Csfgr*	*Csfgr*	1p35-p34	
57.7			*Col8a2*	*Gja4*	*Gja4*	*Gja4*	1p36-q12	
59.0	*Lck*	*Lck*	*Lck*	*Lck*	*Lck*	*Lck*	1p35-p34	
61.0		*Fabph1*	*Fabph1*	*Fabph1*	*Fabph1*	*Fabph1*	1p33-p32	
62.3					*Hmg17*	*Hmg17*	1p36-p35	
64.6				*Fgr*	*Fgr*	*Fgr*	1p36-p35	
64.8				*Oprd1*	*Oprd1*	*Oprd1*	1p36-p34	
65.6			*Pl2ga*		*Hmgcl*	*Hmgcl*	1p36-p35	
65.7	*Fuca*	*Fuca*		*Cbaf3*	*Cbaf3*	*Cbaf3*	1p36	
66.0			*Idb3*	*Idb3*	*Idb3*	*Idb3*	1p36	
66.1	*Fgr*	*Fgr*	*Fgr*		*C1gb*	*C1gb*	1p36-p34	
66.2	*Pgd*	*Pgd*	*Pgd*	*Htr1a*	*Htr1a*	*Htr1a*	1p36-p34	
66.6				*Pl2ga*	*Pl2ga*	*Fuca*	1p34-p33	◄
67.2			*Fuca*	*Fuca*	*Fuca*	*Pl2ga*	1p35	◄
69.0			*Pax7*	*Pax7*	*Pax7*	*Pax7*	1p36	
70.2	*Akp2*	*Akp2*	*Akp2*	*Akp2*	*Akp2*	*Akp2*	1p36-p34	
71.4				*Eno1*	*Hspg2*	*Hspg2*	1p36-p35	
75.4	*Gnb1*	*Gnb1*			*Tnfr2*	*Tnfr2*	1p36	
75.8		*Tnfr2*	*Tnfr2*	*Tnfr2*	*Cd30*	*Cd30*	1p36	
76.5				*Nppa*	*Nppa*	*Nppa*	1p36	
77.6				*Pgd*	*Pgd*	*Pgd*	1p36	
78.9	*Eno1*	*Eno1*	*Eno1*		*Ski*	*Ski*	1q22-q24	◄
79.0	*Ski*	*Ski*	*Ski*	*Ski*	*Eno1*	*Eno1*	1p36	
79.0			*Gnb1*	*Gnb1*	*Gnb1*	*Gnb1*	1p36	

FIGURE 2. The evolving nature of the mouse Chromosome 4 comparative map. Data are from the Mouse Chromosome 4 Committee Reports (http://www.informatics.jax.org). Arrowheads indicate genes (*Ipp*, *Ski*, *Pl2ga* and *Fuca*) that obviously disrupt putative conserved linkages. This map illustrates the consequences of adding new genes to the map and the resulting reassessment of the cumulative mapping information. Mouse Chromosome Committee Reports can be found in Ref. 71.

observation is the conservation of chromosome segments despite large differences in genome size[43].

Several sources of error and ambiguity are evident when identifying conserved segments (Fig. 1). Gross errors occur when homologous genes are incorrectly identified. More subtle errors occur when quantitative differences in recombination frequencies place genes in the wrong location. Errors can also occur when integrating mapping results from different sources. Ambiguity arises from small inversions that alter gene order and blur the pattern of conserved segments created by reciprocal translocations. According to a strict definition of conserved segments, a series of short inversions can create numerous small adjacent segments, whereas longer inversions tend to create several segments scattered along the length of the chromosome. When only a few genes were mapped, these ambiguities and errors were not a serious problem[44]. But with increasingly dense genetic maps, these problems become more serious because arbitrary decisions must be made about the order of genes at the boundaries of conserved segments[9].

The first problem, that of mapping errors, is illustrated by recent work on the human–mouse comparative map[45,46]. In April 1996, mapping evidence in the Mouse Genome Database suggested the existence of ~110 conserved segments between humans and mice. However, for 28 of these segments, the sole evidence for homology was single homologous gene pairs. By August 1996, five of these 28 genes had been removed from the mouse or human gene lists, four had been reassigned to another location in the human or mouse genome, and the existence of only two segments was confirmed by the mapping of additional genes. In addition, six new homology segments marked by a single gene were reported. Obviously, until genetic maps are completely defined, ambiguities will be an integral part of comparative mapping. An example of the evolving nature of comparative map composition and construction is the distal portion of mouse Chromosome (Chr) 4 (Fig. 2).

The confounding effect of inversions on the analysis of translocations is evident in the distribution of segments on a single mouse chromosome with respect to the corresponding human chromosomes. Occasionally, a translocation coincidentally reunites two segments in one species that were originally syntenic (and still are in the other species) and that had been separated by a previous translocation. On mouse Chr 4, the Davis Human/Mouse Homology Map shows four distinct but adjacent conserved segments with counterparts on human Chr 9. The likelihood that random translocations among ~20 chromosomes would achieve this level of coincidence is infinitesimal. That there are four segments instead of just one is most likely to be the result of inversions or other intra-chromosomal rearrangements, in one lineage or both, scrambling the gene order after creation of a single segment by translocation.

Conserved segments, as strictly conceived, are invaluable primary data. But as evidence for evolutionary process, they reflect distinct events, inversions and translocations, with different rates of occurrence. It seems important then to devise objective methods that reduce the arbitrary nature of comparative map construction, first to diminish the effect of mapping error noise, and second to provide insights into the rates of inversions and translocations. An algorithm has been devised to partition each of the two genomes under study into a predetermined number of conserved segments[46], with input consisting of the locations of pairs of homologous genes in their respective genomes[22]. The 'quality' of each segment is measured in terms of the number of genes it contains in proportion to its length (the compactness factor) and the number of other segments whose genes are interspersed with its own, thereby disrupting an otherwise intact segment (the integrity factor). Given the relative weightings for these two factors, the algorithm finds the optimal partition of the genomes into segments. A key problem is to determine appropriate weightings of these factors. This was done by fitting the output of the algorithm to published maps prepared by experts[9,47]. A fit with appropriate weighting factors was readily found. These weighting factors are 'global' – they apply to all autosomal segments and therefore represent a general solution to the problem of identifying conserved segments. The critical issue now is to test the robustness of the proposed method with comparative maps for other species and to find true values for the weighting factors so that comparative maps can be prepared more objectively. Even when complete genomic sequence data are available, identification of conserved segments may pose methodological difficulties due to loss of detectable homology for most genes with the segments. Notions related to integrity and compactness were used to determine chromosome segments conserved in yeast after extensive genomic rearrangement following an ancient genome duplication[48].

The lengths of conserved linkages and the numbers of conserved syntenies

The lengths of conserved segments can be estimated by examining recombination distances in genetic maps[44]. This method is based on using known segment lengths to estimate the average length for all conserved segments in the genome, including segments that are already known as well as those that remain to be discovered. Moreover, with the estimated average segment length (8.1 ± 1.6 cM) and an estimate of genome size (1400 cM) for humans and mice, the true total number of conserved linkages can be estimated[44]. Currently, ~180 of the estimated total of ~200 conserved linkages are known[9]. The number of comparative genetic maps is currently small, but the average conserved segment lengths for many other pairs of species should soon be available, thereby enabling more detailed comparisons of the patterns of linkage conservation. Recently, the Nadeau–Taylor theory has been reformulated in terms of the distribution of the number of genes per conserved segment[49], bypassing the technical problems in estimating segment lengths.

Another important measure for characterizing comparative maps is the number of conserved syntenies. While conserved segments are potentially a more precise measure of genome conservation, the number of conserved syntenies has advantages, not the least of which is substantial existing data for analysis. Several measures of synteny conservation have been proposed[50–54], which do not, however, take into account potentially numerous unobserved segments. Sankoff and Nadeau[55] proposed a method to estimate the true total number of conserved syntenies, including both observed and unobserved. With 87 known conserved syntenies for humans and mice, the true total number of conserved syntenies was estimated to be ~95, suggesting that eight conserved syntenies remain to be discovered.

Rates of chromosome rearrangement

With estimates of the total number of conserved linkages[44] and conserved syntenies in the genome[55], rates of linkage disruption, including both intra- and inter-chromosomal rearrangements, can be calculated[18]. With an estimated total of ~180 rearrangements, the average rearrangement rate is ~0.8 disruptions per million years for humans and mice[44,47,56]. If data are available for only two species, the average rate of change can be estimated for the two lineages, but there is insufficient information to estimate lineage-specific rates. But, when data are available for more than two species, lineage-specific rate estimates can be calculated[18,57].

Rates of chromosomal rearrangement have been estimated for several monocot and dicot plants[29]. Remarkably, the distribution of rate estimates suggests bimodality, with some pairs of species showing lower rates of ~0.15–0.41 rearrangement per million years, while other pairs of species show much higher rates of 1.1–1.3 rearrangements per million years. Whether bimodality reflects the sampling bias of modest linkage conservation data for a small number of species, or a general characteristic of chromosome evolution remains to be determined. It is known that some evolutionary lineages are prone to particular kinds and rates of chromosome rearrangements[58], so multiple modes of rearrangement rates are expected.

Much more data are available for conserved syntenies. Lineage-specific rates, which have been estimated for seven mammalian lineages, range from approximately four rearrangements for the lineage leading to cats, to 18, 23 and 26 for lineages leading to mink, chimps and cattle, and to 52, 57 and 69 for lineages leading to humans, rats and mice, respectively. Unlike rates of nucleotide substitution, which seem to be fairly constant among lineages, generally varying less than twofold[6], rates for synteny disruption vary more than 15-fold. Why it is that chromosomal rearrangement rates are relatively more variable than nucleotide substitution rates remains unclear.

Original synteny – ancestral genetic maps

An important problem involves the use of comparative maps for reconstructing the inferred ancestral genetic map for a diverse group of species. For many species, synteny assignments exceed linkage determinations, although recent technological developments are having a profound impact on the pace and nature of gene mapping. At present, however, comparative mapping often involves analysing sets of syntenic genes. More formally, a genome becomes a family of sets, with one set for each chromosome. Intra-chromosomal rearrangements, such as inversions and local transpositions, do not affect chromosomal assignment and are therefore not detectable in these data sets. Only inter-chromosomal events, such as reciprocal and Robertsonian (fusions and fissions) translocations, affect synteny and can be detected in synteny data.

Ferretti *et al.*[45] explored the following questions: given synteny data from existing organisms, to what extent can we infer 'original synteny' – the synteny sets of ancestral species? How many chromosomes did these ancestors possess and what genes were on each chromosome?

The questions were approached in terms of a syntenic edit distance between two genomes, based on reciprocal translocations and Robertsonian translocations. Formally, these operations can all be represented by the exchange of subsets (possibly null) between chromosomes. The syntenic distance between two genomes thus becomes the minimum number of operations necessary to transform one genome into another, and is a lower bound on the true translocation distance, because not every such subset exchange represents a translocation of two contiguous chromosome fragments. Ferretti *et al.*[45] provided an efficient algorithm for estimating synteny distance. DasGupta *et al.*[59] showed that finding the exact solution is NP-hard. (NP-hard means that the problem is unlikely to be solved by an efficient algorithm because even for moderate-sized instances of this class of problem, an exact computational solution is prohibitively expensive. For realistic numbers of chromosomes, however, this is not a major obstacle.)

The synteny distance algorithm was then used to analyse the median problem for synteny, namely the construction of an ancestral genome the sum of whose distances to three given genomes is minimal. The original synteny problem becomes one of optimizing synteny sets at the internal vertices of a given phylogenetic tree, given synteny data from species at each of the terminal vertices. Ferretti *et al.*[45] devised algorithms for these problems and applied them to known synteny sets for 11 mammalian genomes spanning primates (3), rodents (3), carnivores (2) and artiodactyls (3). A total of 109 translocations sufficed to account for the observed synteny sets on a mammalian phylogeny with 11 terminal taxa and nine ancestral nodes. Synteny sets were reconstructed for the ancestral nodes consistent with current notions based on the fossil record and patterns of DNA sequence variation, but more data for more species are required before confidence can be placed in the details of the reconstruction.

Random breakage model: number and lengths of conserved segments

Several models have been proposed to account for the patterns of segment conservation during evolution. Two of these models are based on selection either preserving particular combinations[31,60] or favoring novel rearrangements that affect gene expression patterns[61]. An alternative to both models is that rearrangement breakpoints are randomly distributed over the genome[2,44]. Testing these hypotheses is difficult: the general patterns of conservation, rather than anecdotal examples, are needed because of uncertainty whether particular segments are representative or exceptional.

Several kinds of evidence are being evaluated. Nadeau and Taylor[44] used the genetic lengths of known conserved linkages to estimate the average length of all conserved segments in humans and mice. With this estimate, the expected frequency distribution of segments lengths was calculated, assuming a random distribution of rearrangement breakpoints – a truncated negative exponential distribution[44]. The fit between the expected and empirical distributions was remarkably good even though the original analysis was based on only 84 genes in 13 segments. Although the Nadeau–Taylor analysis made many assumptions and was based on modest data, their results have held up with increasing amounts of data[47,56], with analyses based only on the distribution of the number of genes per segment[49], and with alternative approaches[62]. Together, these results suggest ~180 rearrangements and ~200 conserved autosomal segments. The consistency of these results, with respect to both larger data sets for analysis and alternative analytical methods, supports the random breakage model.

Questions have been raised, however, about the adequacy of current maps to estimate the true total number of conserved segments between humans and mice[9]. Several recent gene mapping efforts have discovered conserved segments that some believe to be smaller than expected for the random breakage model[31,32,63]. To address the significance of these short segments, the average expected lengths of conserved segments that remain to be discovered was calculated within the random breakage model[63]. With 2097 genes in the comparative map for humans and mice, the expected segment length is ~0.6 cM, a result that is remarkably consistent with the lengths of many new segments that are being discovered[63]. This consistency between expected and empirical lengths for newly discovered segments provides additional support for the random breakage model.

Another test for the nature of conserved segments involves direct evaluation of the functional relation between genes in conserved segments. An obvious limitation of this approach is that known genes are usually only a small sample of all genes present in a segment. Nevertheless, efforts have largely failed to find shared functional characters among genes in conserved segments[60]. Functionally related genes, such as the Hox complexes, the keratin gene families and T-cell receptor complexes, are usually embedded within larger conserved segments that contain genes with many diverse functions.

Outstanding issues

The Human Genome Project is providing a wealth of information about genome organization and evolution, patterns of sequence alterations, and changes in genome composition during evolution are being systematically documented. Inventories of genes are being rapidly developed for human, mice and many other species. Comparative mapping information is increasing rapidly and progress is being made in the development of methods for quantitative analysis of these new data. Algorithms devised for analysing single-chromosome genomes[64,65] are being extended to the multi-chromosomal case[62]. Phylogenetics can now be based on gene order as well as sequence data[66–70].

One of the most important questions involves understanding the factors that contribute to the variable rates of nucleotide substitution and chromosome rearrangement among various evolutionary lineages. Variation might result from cellular factors controlling the occurrence and repair of mutations and rearrangements, or from population factors that determine the probabilities of fixing mutations and rearrangement during evolution. Another important question involves whether the extent to which the location of rearrangement breakpoints is guided by natural selection to preserve particular gene combinations, by structural DNA features that promote or restrict chromosome breakage and repair, or by simply random processes. The proliferation of mapping and sequencing information is beginning to provide unique opportunities to answer these fundamental questions about genome organization and evolution.

Acknowledgements

We thank B. Gaut and A. Paterson for discussing plant comparative mapping and for suggesting pertinent references. This work was supported by a Howard Hughes Medical Institute grant to Case Western Reserve University School of Medicine and by grants from the Natural Sciences and Engineering Research Council of Canada, and the Canadian Genome Analysis and Technology Program. D.S. is a fellow of the Canadian Institute for Advanced Research.

References

1 Comings, D.E. (1972) *Nature* 238, 455–457
2 Ohno, S. (1973) *Nature* 244, 259–262
3 O'Brien, S.J., Seuanez, H. and Womack, J.E. (1988) *Annu. Rev. Genet.* 22, 323–351
4 O'Brien, S.J. and Marshall-Graves, J.A. (1994) in *Human Gene Mapping 1993* (Cuticchia, A.J. and Pearson, P.L., eds), pp. 846–892, Johns Hopkins Press
5 Duret, L., Mouchiroud, D. and Gouy, M. (1994) *Nucleic Acids Res.* 22, 2360–2365
6 Makalowski, W. *et al.* (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 9407–9412
7 Bassett, D.E. *et al.* (1997) *Trends Genet.* 15, 339–344
8 Nadeau, J.H. *et al.* (1995) *Nature* 373, 363–365
9 DeBry, R.W. and Seldin, M.F. (1996) *Genomics* 33, 337–351
10 Peters, J. and Searle, A.G. (1996) in *Genetic Variants and Strains of the Laboratory Mouse* (Lyon, M.F., Rastan, S. and Brown, S.D.M., eds), pp. 1256–1311, Oxford University Press
11 Andersson, L. *et al.* (1996) *Mamm. Genome* 7, 717–773
12 Moore, G. *et al.* (1995) *Curr. Biol.* 5, 737–739
13 McCouch, S. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 1983–1985
14 http://www.informatics.jax.org;
http://www.ncbi.nlm.nih.gov/homology;
http://www.hgmp.mrc.ac.uk;
http://www-iggi.bio.purdue.edu;
http://wheat.pw.usda.gov;
http://probe.nalusda.gov:8300
15 Helentjaris, T., Weber, D.L. and Wright, S. (1988) *Genetics* 118, 353–363
16 Nadeau, J.H. and Sankoff, D. (1997) *Genetics* 147, 1259–1266
17 El-Mabrouk, N., Nadeau, J.H. and Sankoff, D. (1998) in *Combinatorial Pattern Matching*, Ninth Ann. Symp. Lect. Notes Comp. Sci. 1448, pp. 235–250, Springer-Verlag
18 Ehrlich, J., Sankoff, D. and Nadeau, J.H. (1997) *Genetics* 147, 289–296
19 O'Brien, S.J. *et al.* (1993) *Nat. Genet.* 3, 103–112
20 Lyons, L.A. *et al.* (1997) *Nat. Genet.* 15, 47–56
21 Marra, M.A., Hillier, L. and Waterston, R.H. (1998) *Trends Genet.* 14, 4–7
22 Morizot, D.C. and Siciliano, M.J. (1983) *Curr. Top. Biol. Med. Res.* 10, 261–285
23 Elgar, G. *et al.* (1996) *Trends Genet.* 12, 145–150
24 Postlethwait, J.H. *et al.* (1998) *Nat. Genet.* 18, 345–349
25 Baxendale, S. *et al.* (1995) *Nat. Genet.* 10, 67–76
26 Venkatesh, B. and Brenner, S. (1995) *Adv. Exp. Med. Biol.* 395, 629–638
27 Gilley, J., Armes, N. and Fried, M. (1997) *Nature* 385, 305–306
28 Trachtulec, Z. *et al.* (1997) *Genomics* 44, 1–7
29 Paterson, A. *et al.* (1996) *Nat. Genet.* 14, 380–382
30 Moore, G. *et al.* (1995) *Curr. Biol.* 5, 737–739
31 Gale, M.D. and Devos, K.M. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 1971–1974
32 Bennetzen, J.L. *et al.* (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 1975–1978
33 Dean, C. and Schmidt, R. (1995) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46, 395–418
34 Tanksley, S.D. *et al.* (1992) *Genetics* 132, 1141–1160
35 Ahn, S., Anderson, J.A., Sorrells, M.E. and Tanksley, S.D. (1993) *Mol. Gen. Genet.* 241, 483–490
36 Ahn, S. and Tanksley, S.D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7980–7984
37 Whitkus, R., Doebley, J. and Lee, M. (1992) *Genetics* 132, 1119–1130
38 Weeden, N.R., Muehlbauer, F.J. and Ladizinisky, G. (1992) *J. Hered.* 83, 123–129
39 Van Deynze, A.E. *et al.* (1995) *Mol. Gen. Genet.* 248, 744–754
40 Van Deynze, A.E. *et al.* (1995) *Mol. Gen. Genet.* 249, 349–356
41 Tanksley, S.D., Bernatsky, R., Lapitan, N.L. and Prince, J.P. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 6419–6423
42 Kowalski, S.D., Lan, T-H., Feldmann, K.A. and Paterson, A.H. (1994) *Genetics* 138, 499–510
43 Moore, G., Gale, M.D., Kurata, N. and Flavell, R.B. (1993) *Biotechnology.* 11, 584–589
44 Nadeau, J.H. and Taylor, B.A. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 814–818
45 Ferretti, V., Nadeau, J.H. and Sankoff, D. (1996) in *Combinatorial Pattern Matching, Seventh Annual Symposium* (Hirschberg, D. and Myers, G., eds), Lect. Notes Comp. Sci. 1075, pp. 159–167, Springer-Verlag
46 Sankoff, D., Ferretti, V. and Nadeau, J.H. (1997) *J. Comp. Biol.* 4, 559–565
47 Copeland, N.G. *et al.* (1993) *Science* 262, 57–66
48 Wolfe, K.H. and Shields, D.C. (1997) *Nature* 387, 708–713
49 Sankoff, D., Parent, M.N., Marchand, I. and Ferretti, V. (1997) in *Combinatorial Pattern Matching. Eighth Annual Symposium* (Apostolico, A. and Hein, J., eds), Lect. Notes Comp. Sci. 1264, pp. 262–274, Springer-Verlag
50 Zakharov, I.A. and Valeev, A.K. (1988) *Proc. (Doklady) Acad. Sci. USSR* 301, 1213–1218
51 Zakharov, I.A., Nikiforov, V.I. and Stepanyuk, E.V. (1992) *Genetika* 28, 77–81
52 Zakharov, I.A., Nikiforov, V.I. and Stepanyuk, E.V. (1995) *Genetika* 31, 1163–1167
53 Bengtsson, B.O., Levan, K.K. and Levan, G. (1993) *Cytogenet. Cell Genet.* 64, 198–200
54 Zakharov, I.A. (1993) in *Bioinformatics, Supercomputing and Complex Genome Analysis* (Lim, H., Ficket, J. and Cantor, C., eds), World Scientific

55 Sankoff, D. and Nadeau, J.H. (1996) *Discr. Appl. Math.* 71, 247–257

56 Nadeau, J.H. (1989) *Trends Genet.* 5, 82–86

57 Sarich, V.M. and Wilson, A.C. (1967) *Science* 158, 1200–1203

58 White, M.J.D. (1978) *Modes of Speciation*, W.H. Freeman

59 DasGupta, B. *et al.* (1997) in *Proc. First Annu. Int. Comput. Mol. Biol. (RECOMB '97)*, pp. 99–108, ACM Press

60 Lundin, L-G. (1993) *Genomics* 16, 1–19

61 Wilson, A.C., Sarich, V.M. and Maxson, L.R. (1974) *Proc. Natl. Acad. Sci. U. S. A.* 71, 3028–3030

62 Hannenhalli, S. and Pevzner, P.A. (1995) *Proc. 36th Ann. Symp. Found. Comp. Sci.*, pp. 581–592, IEEE Computer Society Press

63 Nadeau, J.H. and Sankoff, D. (1998) *Mamm. Genome* 9, 491–495

64 Kececcioglu, J. and Sankoff, D.S. (1995) *Algorithmica* 13, 180–210

65 Hannenhalli, S. and Pevzner, P.A. (1995) *Proc. 27th Ann. ACM Symp. Theory Comput*, pp. 178–189, ACM Press

66 Sankoff, D. *et al.* (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 6575–6579

67 Boore, J.L. *et al.* (1995) *Nature* 376, 163–165

68 O'Brien, T.G. *et al.* (1998) *Nature* 392, 667–668

69 Kellogg, E.A. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 2005–2010

70 Sankoff, D. and Blanchette, M. *J. Comp. Biol.* (in press)

71 Silver, L.M., Nadeau, J.H. and Brown, S.D.M. (1998) *Mamm. Genome* 8, S1–S450

J.H. Nadeau *is in the Genetics Department, Case Western Reserve University School of Medicine, 10900 Euclid Avenue, Cleveland, OH 44106, USA.*

D. Sankoff *is at the Centre de recherches mathématiques, Université de Montréal, Montréal, Québec, Canada H3C 3J7.*

Fragile sites still breaking

GRANT R. SUTHERLAND (gsutherl@mad.adelaide.edu.au)

ELIZABETH BAKER (ebaker@medicine.adelaide.edu.au)

ROBERT I. RICHARDS (rrichard@medicine.adelaide.edu.au)

Rare fragile sites on chromosomes are the archetypal dynamic mutations. They involve large expansions of the microsatellite CCG or AT-rich minisatellites. The mutation process is an increase in repeat-unit number from within a normal range, through a premutation range, up to full mutation where the fragile site is expressed. Full mutations can inactivate genes and are regions of genomic instability. Common fragile sites, in particular, might have a role in oncogenesis by facilitating gene inactivation through chromosomal deletion or amplification, but this requires further exploration. The mechanisms behind the changes that give rise to the cytogenetic manifestation of chromosomal fragility are now beginning to be understood.

Fragile sites are points at which chromosomes break non-randomly under certain specific conditions and have become of increasing interest over the past two decades. At first they seemed to be a cytogenetic curiosity. Then, one fragile site became a marker for, and gave its name to, the most common form of familial mental retardation, fragile X syndrome. More recently, some fragile sites are seen as having a role in genome instability, which might contribute to oncogenesis. The cloning of the fragile X gave rise to the concept of dynamic mutation, a process now known to be responsible for a number of neurological disorders and all of the rare fragile sites that have been cloned.

We have reviewed the cytogenetic and molecular genetic aspects of fragile sites on several occasions. Since our most recent review[1] there have been a number of developments: new information on the structure of fragile sites that are not in the folate-sensitive category has become available; there have been some new cytogenetic findings; some light has been shed on the timing of expansion of the *FRAXA* CCG repeat; and further characterization of the *FMR2* gene associated with *FRAXE*. Also, the possible role of fragile sites in the oncogenic process has, again, come into prominence – the association between common fragile sites and cancer has recently been reviewed[2].

Classification

Fragile sites on chromosomes are classified as rare (carried by less than 1 in 20 people) or common (on virtually all chromosomes) and are subdivided further according to the conditions under which they are induced in tissue culture (Box 1). The major group of rare fragile sites is the folate-sensitive group that includes *FRAXA*, which is responsible for fragile X syndrome, and *FRAXE*, which is associated with non-specific mental retardation. Other rare fragile sites are induced by bromodeoxyuridine (BrdU) alone, distamycin A alone or by either of these compounds. The majority of common fragile sites are induced by aphidicolin. A complete listing of human fragile sites and their classification is available[3].

Structure

Before a physical structure like a fragile site has been identified it is difficult to know with certainty when the DNA sequences responsible will have been isolated. It is now clear that the rare fragile sites have large expansions of repeat sequences that co-segregate with the fragile sites, and that sequences that flank these repeats can be shown by fluorescence *in situ* hybridization to flank the fragile site. For the common fragile sites the situation is less clear and despite DNA sequence information across these fragile sites it is not known what element or elements of sequence are the minimal requirement for a fragile site. Or, indeed, if there is a specific sequence requirement for the common fragile sites.

PII: S0168-9525(98)01628-X