

Prevalence of small inversions in yeast gene order evolution

Cathal Seoighe*, Nancy Federspiel†, Ted Jones†, Nancy Hansen†, Vesna Bivolarovic†, Ray Surzycki†, Raquel Tamse†, Caridad Komp†, Lucas Huizar†, Ronald W. Davis†, Stewart Scherer‡, Evelyn Tait§¶, Duncan J. Shaw§, David Harris§, Lee Murphy¶, Karen Oliver¶, Kate Taylor¶, Marie-Adele Rajandream¶, Bart G. Barrell¶, and Kenneth H. Wolfe*,**

*Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland; †Stanford DNA Sequencing and Technology Center, 855 California Avenue, Palo Alto, CA 94304; ‡3938 Paseo Grande, Moraga, CA 94556; §Department of Molecular and Cell Biology, Institute of Medical Science, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, United Kingdom; and ¶Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Contributed by Ronald W. Davis, September 28, 2000

Gene order evolution in two eukaryotes was studied by comparing the *Saccharomyces cerevisiae* genome sequence to extensive new data from whole-genome shotgun and cosmid sequencing of *Candida albicans*. Gene order is substantially different between these two yeasts, with only 9% of gene pairs that are adjacent in one species being conserved as adjacent in the other. Inversion of small segments of DNA, less than 10 genes long, has been a major cause of rearrangement, which means that even where a pair of genes has been conserved as adjacent, the transcriptional orientations of the two genes relative to one another are often different. We estimate that about 1,100 single-gene inversions have occurred since the divergence between these species. Other genes that are adjacent in one species are in the same neighborhood in the other, but their precise arrangement has been disrupted, probably by multiple successive multigene inversions. We estimate that gene adjacencies have been broken as frequently by local rearrangements as by chromosomal translocations or long-distance transpositions. A bias toward small inversions has been suggested by other studies on animals and plants and may be general among eukaryotes.

The order and transcriptional orientation of genes along a chromosome can change during evolution by DNA inversions and transpositions or by chromosomal translocations. In bacteria, long strings of genes show conserved order and orientation between closely related species or strains, so that most rearrangements involve large pieces of DNA (1–4). Comparative genetic mapping in vertebrates and plants has shown that large syntenic regions are conserved, with breakpoints corresponding to interchromosomal translocations (5, 6). More detailed mapping of some of these syntenic regions, however, has revealed several examples where the local gene order is not conserved (7–11).

The extent of gene order conservation among ascomycete fungi previously has been estimated by comparing the *S. cerevisiae* genome sequence (12) to DNA sequences from other species, using either random “genome survey” sequences from both ends of small clones (13–15) or existing European Molecular Biology Laboratory database sequences (16). Between *Saccharomyces cerevisiae* and *Candida albicans*, two species separated by 140–330 million years (17, 18), only one example of conserved gene order and orientation has been reported so far (*STE6–UBA1*; ref. 19), whereas there are three cases of gene pairs that are adjacent in both species but where one gene has been inverted (*RAD16–LYS2*, *NFS1–LEU2* and *RPS31–SEC10*; refs. 15, 16, and 20–22). To investigate whether the apparently high frequency of inversions is general throughout the *C. albicans* genome we analyzed the genome sequence data currently available.

Methods

The 1,680 contig DNA sequences in the May 2000 data release (7× coverage) from the *C. albicans* whole genome shotgun sequencing project at Stanford University were downloaded from <http://www-sequence.stanford.edu/group/candida>. The contigs range from 2 to 151 kb and total 16.2 Mb, which is approximately equal to the estimated genome size. The *S. cerevisiae* proteome was searched against these contigs by using gapped TBLASTN (23) with the *seg* filter (24) and a cutoff *E* value of 10^{-10} . Reanalysis of the data using other cutoffs (10^{-6} , 10^{-20}) did not change the results significantly. Gene locations in the Stanford contigs were estimated solely on the basis of TBLASTN hits to *S. cerevisiae* proteins, and we did not annotate them further. The sequences of six cosmids completely sequenced at the Sanger Centre were obtained from GenBank (accession numbers AL033391, AL033396, AL033497, AL033501, AL033502, and AL033503). Gene identifications for the Sanger Centre cosmids shown in Fig. 4 were made manually, and some of these involve matches weaker than *E* = 10^{-10} . *C. albicans* genes without orthologs in *S. cerevisiae* were ignored in the analysis of Stanford contigs, but are shown for the cosmids in Fig. 4.

Results

Changes in Gene Order and Orientation. We used TBLASTN searches (23) with all *S. cerevisiae* proteins to estimate the locations of genes in the *C. albicans* contigs assembled at Stanford, thereby ignoring any *C. albicans* genes that do not have *S. cerevisiae* counterparts. The contigs contain 3,188 pairs of genes that appear to be adjacent in *C. albicans* (i.e., either they are adjacent, or any intervening genes do not have *S. cerevisiae* orthologs). For 298 pairs (9%), the *S. cerevisiae* orthologs are also adjacent. Despite remaining as neighbors, 103 of these pairs (35%) have different gene orientation or order in the two species. Ninety-one pairs can be explained by inversions of one gene, and 12 pairs require two inversions each (Fig. 1). From this observation (115 single-gene inversions among 298 intergenic links), we estimate that the total number of single-gene inversions that have occurred in their genomes following the divergence of these species is about 1,100 (= 5,800 genes in the genome \times 115/298 \div 2 links broken per inversion).

The set of 298 adjacent pairs includes 21 runs of three genes that have conserved gene order in the two species. Among

¶Present address: Department of Anatomy and Physiology, University of Dundee, Dundee DD1 4HN, United Kingdom.

**To whom reprint requests should be addressed. E-mail: khwolfe@tcd.ie.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: Proc. Natl. Acad. Sci. USA, 10.1073/pnas.240462997. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.240462997

<i>Candida albicans</i>					
<i>Saccharomyces cerevisiae</i>	parallel ➡➡	convergent ➡➡	divergent ➡➡	alternative parallel ➡➡	
	parallel ➡➡	conserved 71 pairs	1 inversion 14 pairs	1 inversion 18 pairs	2 inversions 6 pairs
	convergent ➡➡	1 inversion 32 pairs	conserved 70 pairs	2 inversions 5 pairs	n/a
	divergent ➡➡	1 inversion 27 pairs	2 inversions 1 pairs	conserved 54 pairs	n/a

Fig. 1. Order and orientation relationships between 298 gene pairs that are adjacent in both *S. cerevisiae* and *C. albicans*. All 10 possible relationships between two adjacent genes are shown, with the number of inversions needed to convert any combination into any other. The names of gene pairs in each category are listed at www.gen.tcd.ie/khwolfe/candida. The categories labeled as “2 inversions” also could be explained by one gene leapfrogging over the other, but we consider this unlikely.

these, 16 examples of apparent single-gene inversions are seen (Fig. 2). The most dramatic example is the cluster *SLU7–RRP1–SSS1*, where the order is conserved but all three genes have reversed orientations. This example could be explained either by three independent single-gene inversions or by two short-distance transpositions, both of which seem quite improbable.

Other pairs of adjacent *C. albicans* genes have *S. cerevisiae* orthologs that are physically close to each other but are not immediate neighbors (Fig. 3). The Stanford contig data include 97 pairs of adjacent *C. albicans* genes whose *S. cerevisiae* orthologs are separated by 1–5 intervening genes. Gene orientation and relative order are conserved in 28 of these pairs, which is only slightly more than the 24.25 expected by chance. These findings suggest that multigene inversions may have occurred, moving genes over short distances.

To further examine local gene order we studied six *C. albicans* cosmids (25) that were completely sequenced at the Sanger Centre. These sequence comparisons point to numerous rearrangements, both interchromosomal (translocations) and intrachromosomal (small inversions). Most of the long *C. albicans* sequences contain small clusters of genes whose *S. cerevisiae* orthologs also are physically clustered (Fig. 4). These clusters are generally shorter than 10 genes in *C. albicans* and often are interspersed with genes from other *S. cerevisiae* chromosomes. The ends of the clusters probably correspond to sites of chromosomal translocations (16, 26, 27). In some cases a cluster of genes in *C. albicans* is related to two *S. cerevisiae* genomic regions (blocks) that are paired by whole-genome duplication in the *S. cerevisiae* lineage (26), as predicted by our model (16, 27). The relationships shown in Fig. 4 comprise 32 orthologous genes and at least 11 independent inversions. It is not possible to estimate the exact sizes of these inversions (i.e., the numbers of genes involved) because, in all cases, the genes immediately upstream and downstream of the inverted ones are different in the two species. For example, the inversion of *YLR423C* in cosmid Ca49C10 might have included some of the four genes downstream of it in *S. cerevisiae*. However, the inversions must be relatively small because gene order is conserved at a coarser level (e.g., *YLR423C* is in-between *YLR418C* and *YLR424W* in both species). Similar scrambling of local gene order recently was reported by Mallet *et al.* (28) for the region around the *CHS6* gene compared between *S. cerevisiae* and *C. albicans*.

The conservation of small neighborhoods of genes, without absolute conservation of order or orientation, suggests that small

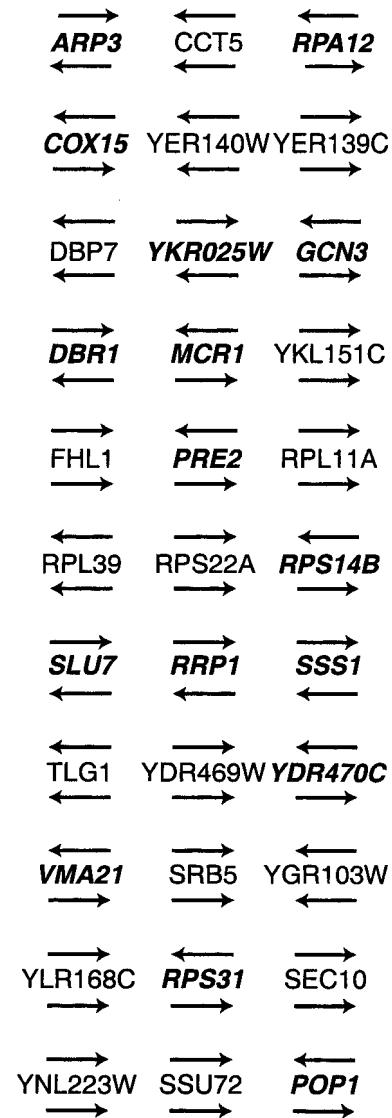


Fig. 2. Examples of single-gene inversions. The three genes in each set are adjacent (ignoring any *C. albicans* genes without homologs) and in the same order in the two species. Directions of transcription in *S. cerevisiae* and *C. albicans* are shown above and below gene names, respectively. Genes named in bold italics have different orientations in the two species. *S. cerevisiae* gene names are used.

DNA inversions have contributed significantly to the evolution of ascomycete genomes. A further example is seen in cosmid Ca49C4 (Fig. 4), which contains a pseudogene related to the *C. albicans* oligopeptide transporter gene *OPT1* (29) and its *S. cerevisiae* homolog *YJL212C*. The pseudogene has 98% DNA sequence identity over 2 kb to part of *OPT1*, but a 0.3-kb internal segment has been inverted relative to *OPT1* and other members of this gene family. There is also evidence for small inversions within the *S. cerevisiae* genome itself, where 11 of 655 duplicated genes associated with whole-genome duplication now show inverted orientation with respect to the flanking chromosomal regions (30).

Relative Rates of Intrachromosomal Versus Interchromosomal Rearrangements. Small rearrangements keep genes within a local neighborhood, so we can use the *C. albicans*/*S. cerevisiae* comparisons from the Stanford dataset (Figs. 1 and 3) to estimate the

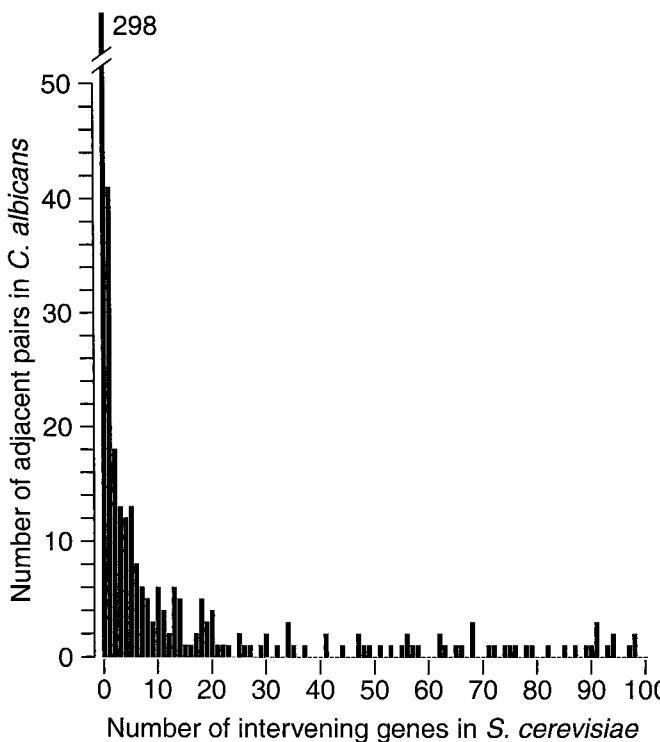


Fig. 3. Histogram showing the distance apart in *S. cerevisiae* of the orthologs of gene pairs that are adjacent in *C. albicans*. The distance between two genes is expressed in terms of the number of other genes between them on the chromosome.

rate of small rearrangements (which we suggest are mostly inversions) relative to large rearrangements (translocations, larger inversions, and long-distance transpositions). Even if there had been no other chromosomal rearrangements, we would expect about half of the links between immediate neighbors in *S. cerevisiae* and *C. albicans* to have been broken by the process of random gene loss due to differential silencing after genome duplication in the *S. cerevisiae* lineage (26, 31). The remaining breaks are the combined result of inversions, translocations, and transpositions. The fraction of links that has been conserved is under 10%, but this fraction has been reduced by a factor of 2 by genome duplication in *S. cerevisiae*. Consequently, chromosomal rearrangements are responsible for breaking over 80% of the links between neighbors. Assuming that breakpoints are made randomly, and using a Poisson distribution to correct for multiple hits, this implies that there have been an average of 1.6 breaks per link, or approximately 9,000 breakpoints in total since speciation. This argument assumes that the *S. cerevisiae* genome duplication occurred recently, but an identical conclusion is reached if the genome duplication is assumed to have occurred shortly after speciation. It also assumes that no other genome duplications have occurred in either lineage.

Statistical methods have been developed previously to estimate relative numbers of intrachromosomal and interchromosomal rearrangements between species (32), but these methods are not adaptable to the current problem because the kind of data being considered is local (the Stanford contigs are short relative to chromosomes) and because the number of rearrangements is close to saturation. It is problematic to model the small inversions directly because not enough is known about their size distribution. Instead, to model the combined processes of large and small chromosomal rearrangements,

adjacent genes in *C. albicans* having orthologs on the same chromosome in *S. cerevisiae* were divided into two categories: gene pairs that are also adjacent in *S. cerevisiae* (state A), and gene pairs that are “near-neighbors” (syntenic but separated by a small number of genes) in *S. cerevisiae* (state B).

The number of gene pairs in the sequenced sample that are in state A is P_A . In a time interval Δt the change in population of state A is

$$\Delta P_A = -P_A(L + S)\Delta t \dots, \quad [1]$$

where L and S are the rates at which single intergenic links are broken by large and small rearrangements, respectively. Let I be the mean number of intervening genes for gene pairs that are near-neighbors in *S. cerevisiae*, so that $I + 1$ is the mean number of intervening links. If we make the assumption that the average separation of this category of gene pair has been similar throughout the evolutionary history then

$$\Delta P_B = SP_A\Delta t - (I + 1)LP_B\Delta t \dots. \quad [2]$$

This assumption is justified because gene pairs in state B are unlikely to drift too far apart before their linkage is broken by a translocation. Large rearrangements (translocations) are taken to be the only way in which gene pairs leave state B because the number of gene pairs that are syntenic but not near-neighbors is small (Fig. 3).

Eq. 1 and 2 above can be treated as differential equations and solved, giving

$$P_A = C_1 e^{-(L+S)t} \dots. \quad [3]$$

$$P_B = C_2 e^{-(I+1)Lt} + \frac{SC_1}{IL - S} e^{-(L+S)t}. \quad [4]$$

At time 0 $P_A = J/2$, where J is the number of gene pairs in the sample that are adjacent in *C. albicans* and have orthologs in *S. cerevisiae*, because *S. cerevisiae* has undergone genome duplication followed by differential silencing. Therefore $C_1 = J/2$. At time 0 $P_B = 0$, therefore

$$C_2 = -\frac{1}{2} \frac{SJ}{(IL - S)}.$$

Eq. 3 and 4 provide an estimate of the proportion of all rearrangements that are small ($S/S+L$), given values for the number of conserved adjacent gene pairs (P_A , which is 298; Fig. 3), the number of pairs that are adjacent in *C. albicans* but near-neighbors in *S. cerevisiae* (P_B) and the average spacing between near-neighbors (I). The values of P_B and I can be calculated from the data in Fig. 3 but depend on the maximum number of intervening genes that is permitted in the definition of near-neighbors (I_{\max}). In Fig. 3 there appears to be an excess of conserved linkages over short distances, up to a limit of at least five intervening genes and possibly as many as 20. The relationship between the estimated proportion of small rearrangements and I_{\max} is shown in Fig. 5. Allowing a maximum of five genes between near-neighbors, 38% of broken links are attributed to small rearrangements. This increases to 67% for $I_{\max} = 20$ genes. These results suggest that approximately equal numbers of linkages have been broken by small and large rearrangements.

A limit of $I_{\max} = 5$ also was suggested by an experiment where we compared the number of adjacent pairs in *C. albicans* whose homologs are syntenic in *S. cerevisiae* to those whose homologs are located on specific pairs of different chromosomes, as a way

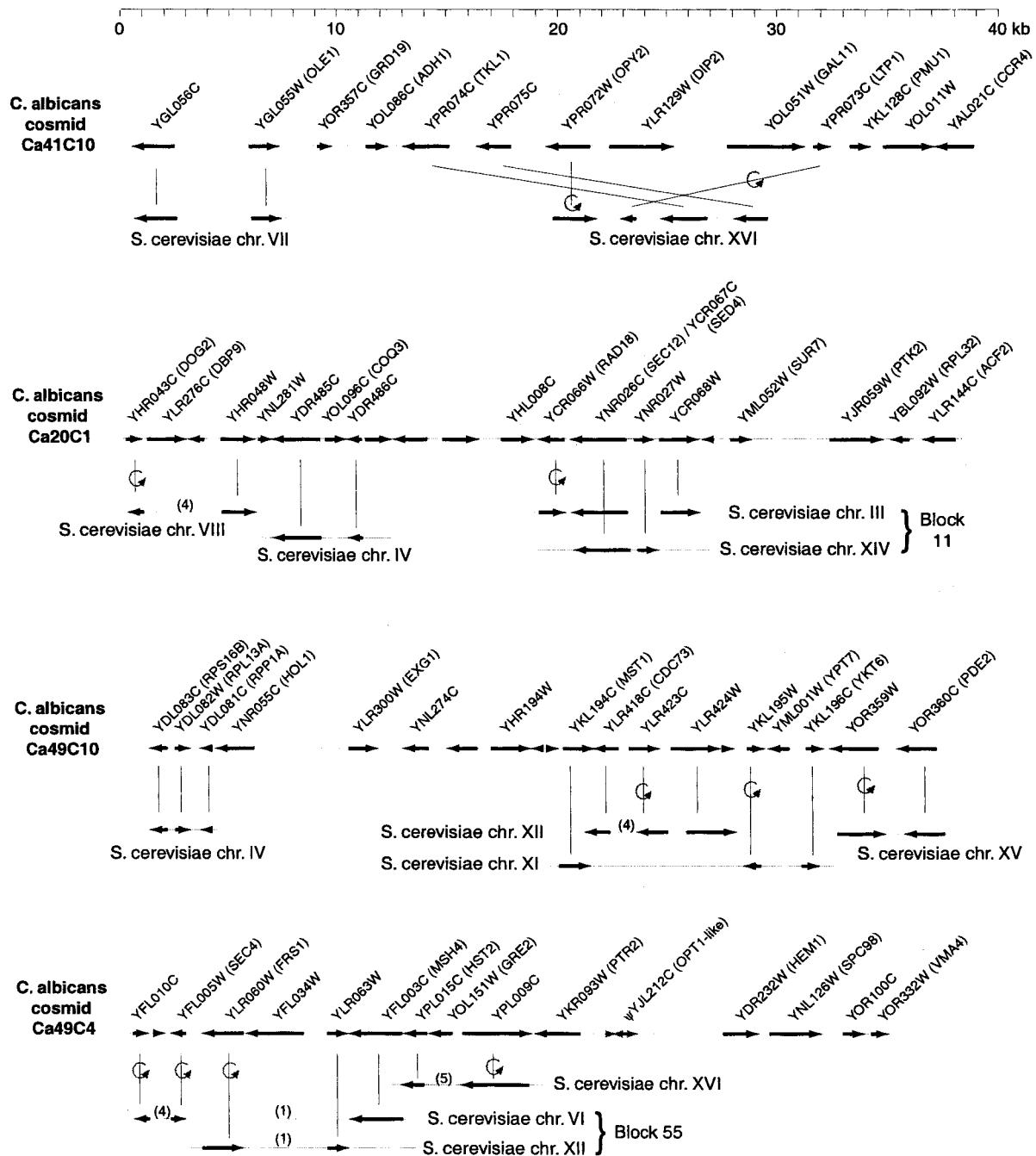


Fig. 4. Gene order relationships of four *C. albicans* cosmids, sequenced at the Sanger Centre, to parts of the *S. cerevisiae* genome. Vertical lines connect orthologous genes. Curved arrows indicate genes with inverted orientations. *C. albicans* genes are named after their *S. cerevisiae* orthologs; unnamed genes have no close relative in *S. cerevisiae*. Numbers in parentheses indicate numbers of intervening genes in *S. cerevisiae* that are not shown. *S. cerevisiae* regions in or near duplicated chromosomal blocks (26) are labeled. The scale at the top refers to *C. albicans* only.

of estimating the “background” level of random gene associations in Fig. 3 (data not shown).

Discussion

This study compares gene order between two eukaryotes based on whole genome sequence data. Our results suggest that successive random small inversions frequently cause a gene’s chromosomal position and orientation to drift during its evolution. This process would alter gene order and orientation without moving any genes very far from their starting points. It also would tend to blur the endpoints of interchromosomal

translocations. The mechanism by which small inversions occur is unknown, and our data are uninformative in this regard because intergenic sequences are highly diverged between *C. albicans* and *S. cerevisiae*. Our results also suggest that gene order in yeasts is relatively unconstrained by natural selection. The orientations of some pairs of adjacent genes, particularly those that are transcribed divergently from a shared regulatory region (such as the histone pair *HTA1*–*HTB1*) may be under selection, but the high frequency of rearrangement indicates that this type of constraint is the exception rather than the rule (cf. ref. 33). It is notable that

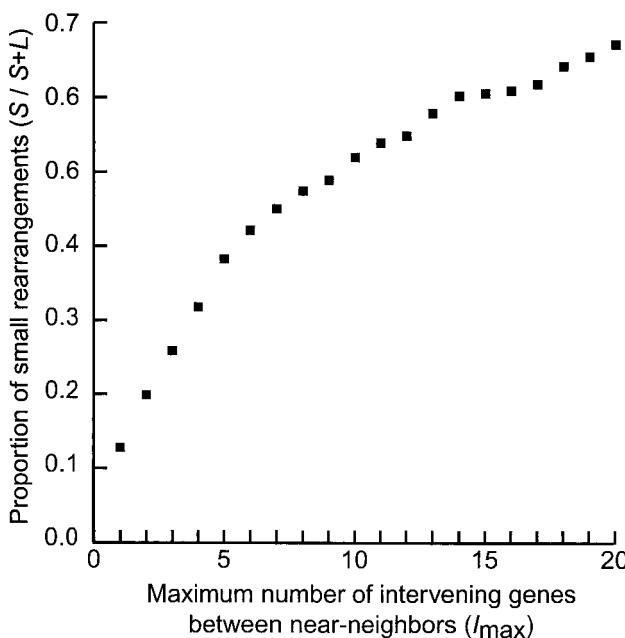


Fig. 5. Relationship between the maximum permitted number of intervening genes (l_{\max}) between near-neighbors (gene pairs in state B), and the estimate of the proportion of rearrangements that are small ($S/S+L$). Calculated numerically from Eqs. 3 and 4 using data for P_B at different values of l_{\max} from Fig. 3.

divergently transcribed adjacent gene pairs are broken up at approximately the same frequency as pairs transcribed convergently or in parallel (Fig. 1).

- Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. (1997) *Nucleic Acids Res.* **25**, 701–712.
- Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., et al. (1999) *Nature (London)* **397**, 176–180.
- Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R. W., Olinger, L., Grimwood, J., Davis, R. W. & Stephens, R. S. (1999) *Nat. Genet.* **21**, 385–389.
- Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Utterback, T., Berry, K., et al. (2000) *Nucleic Acids Res.* **28**, 1397–1406.
- Carver, E. A. & Stubbs, L. (1997) *Genome Res.* **7**, 1123–1137.
- Gale, M. D. & Devos, K. M. (1998) *Science* **282**, 656–659.
- Johansson, M., Ellegren, H. & Andersson, L. (1995) *Genomics* **25**, 682–690.
- Lindsay, E. A., Botta, A., Jurecic, V., Carattini-Rivera, S., Cheah, Y. C., Rosenblatt, H. M., Bradley, A. & Baldini, A. (1999) *Nature (London)* **401**, 379–383.
- Yang, Y. P. & Womack, J. E. (1998) *Genome Res.* **8**, 731–736.
- Gilley, J. & Fried, M. (1999) *Hum. Mol. Genet.* **8**, 1313–1320.
- Cavell, A. C., Lydiate, D. J., Parkin, I. A., Dean, C. & Trick, M. (1998) *Genome* **41**, 62–69.
- Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L., Allen, E., Alt-Mörbe, J., André, B., Andrews, S., et al. (1997) *Nature (London)* **387**, Suppl., 5–105.
- Altmann-Jöhl, R. & Philipsen, P. (1996) *Mol. Gen. Genet.* **250**, 69–80.
- Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F. & Dujon, B. (1998) *Nucleic Acids Res.* **26**, 5511–5524.
- Hartung, K., Frishman, D., Hinnen, A. & Wolf, S. (1998) *Yeast* **14**, 1327–1332.
- Keogh, R. S., Seoighe, C. & Wolfe, K. H. (1998) *Yeast* **14**, 443–457.
- Berbee, M. L. & Taylor, J. W. (1993) in *The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics* eds. Reynolds, D. R. & Taylor, J. W. (CAB International, Wallingford, U.K.), pp. 67–78.
- Pesole, G., Lotti, M., Alberghina, L. & Saccone, C. (1995) *Genetics* **141**, 903–907.
- Raymond, M., Dignard, D., Alarco, A. M., Mainville, N., Magee, B. B. & Thomas, D. Y. (1998) *Mol. Microbiol.* **27**, 587–598.
- Suvarna, K., Seah, L., Bhattacherjee, V. & Bhattacharjee, J. K. (1998) *Curr. Genet.* **33**, 268–275.
- Plant, E. P., Becher, D. & Poulter, R. T. (1998) *Yeast* **14**, 287–295.
- Roig, P., Martinez, J. P., Gil, M. L. & Gozalbo, D. (2000) *Yeast* **16**, 1413–1419.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
- Tait, E., Simon, M. C., King, S., Brown, A. J., Gow, N. A. & Shaw, D. J. (1997) *Fungal Genet. Biol.* **21**, 308–314.
- Wolfe, K. H. & Shields, D. C. (1997) *Nature (London)* **387**, 708–713.
- Seoighe, C. & Wolfe, K. H. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 4447–4452.
- Mallet, L., Renault, G. & Jacquinet, M. (2000) *Yeast* **16**, 959–966.
- Lubkowitz, M. A., Hauser, L., Breslav, M., Naider, F. & Becker, J. M. (1997) *Microbiology* **143**, 387–396.
- Seoighe, C. & Wolfe, K. H. (1999) *Gene* **238**, 253–261.
- Seoighe, C. & Wolfe, K. H. (1999) *Curr. Opin. Microbiol.* **2**, 548–554.
- Ehrlich, J., Sankoff, D. & Nadeau, J. H. (1997) *Genetics* **147**, 289–296.
- Kruglyak, S. & Tang, H. (2000) *Trends Genet.* **16**, 109–111.
- Ladrière, J. M., Georis, I., Guerineau, M. & Vandenhoute, J. (2000) *Gene* **255**, 83–91.
- Ruddle, F. H., Bentley, K. L., Murtha, M. T. & Risch, N. (1994) *Development (Cambridge, U.K.) Suppl.*, 155–161.
- Trachtulec, Z., Hamvas, R. M., Forejt, J., Lehrach, H. R., Vincek, V. & Klein, J. (1997) *Genomics* **44**, 1–7.
- Pébusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. (1998) *Mol. Biol. Evol.* **15**, 1145–1159.
- Ruvkun, G. & Hobert, O. (1998) *Science* **282**, 2033–2041.
- Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C., Mastwal, S., Zhu, X., Scheel, J., et al. (2000) *Science* **287**, 989–994.
- Robin, G. C. d. Q., Claudianos, C., Russell, R. J. & Oakeshott, J. G. (2000) *J. Mol. Evol.* **51**, 149–160.
- Semple, C. & Wolfe, K. H. (1999) *J. Mol. Evol.* **48**, 555–564.
- Lin, X., Kaul, S., Rounseley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., et al. (1999) *Nature (London)* **402**, 761–768.
- Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoff, A., Stiekema, W., Entian, K. D., Terryn, N., et al. (1999) *Nature (London)* **402**, 769–777.
- Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. (2000) *Plant Cell* **12**, 1093–1102.