GENOME ARCHAEOLOGY: DETECTING ANCIENT POLYPLOIDY IN CONTEMPORARY GENOMES

Todd J. Vision
Daniel G. Brown

Many present-day organisms are descendants of ancient *polyploids*. While recently generated polyploids are easily detected due to the presence of a complete set of duplicated chromosomes, the footprint of ancient polyploidy may be more subtle. After many generations of gene deletion, divergence of paralogous genes, and chromosomal rearrangements, an ancient polyploid comes to resemble a nonpolyploid genome that has experienced multiple local duplication events. Here we simulate genomes evolving under these conditions and monitor the resultant distances between paralogous pairs of genes, as measured by the number of intervening genes in each genome. Some measures of this nearest neighbor distribution decay very slowly following polyploidy and thus may offer a means of experimentally testing alternative hypothesis concerning ancestral ploidy. The nearest neighbor distribution for yeast appears consistent with the proposed ancient tetraploidy of this genome. We briefly review empirical data relevant to these processes and suggest areas for future research.

1. Introduction

Many genomes, even famously compact ones such as those of yeast and *Arabidopsis*, contain duplicated chromosomal segments that span multiple genes. It has been proposed that such segments derive from ancient global genome duplication events (Grant et al., 2000; Wolfe and Shields, 1997). Global genome duplication events generate two *symmetric* sets of chromosomes in each nucleus; gene content and order are initially identical between paired members of each set. This symmetry degrades over time as translocation between chromosome arms, inversions, transpositions, and small-scale deletion and duplication events occur. Duplicated genes, or paralogs, also gradually diverge from one another until homology is no longer recognizable: some duplicates acquire new functions while others decay to pseudogenes. As a result of these processes, ancient polyploids will have many small blocks of divergent, though recognizably duplicated, chromosomal segments

D. Sankoff et al. (eds.), *Comparative Genomics*

dispersed throughout the genome. Since such a pattern also arises from the accumulation of many subchromosomal duplications, it is of interest to explore how the footprint of ancient polyploidy may be detected.

1.1. Genome similarity matrix and nearest neighbor statistic

Our basic study object in this work is the genome similarity matrix (GSM). Consider a genome of n genes; the GSM is the $n \times n$ matrix G, where $G_{i,j} \neq 0$ if genes i and j share common ancestry and 0 otherwise. We use sequence similarity as a proxy for common ancestry, and set $G_{i,j}$ to be the time since common ancestry between i and j. If i and j share no ancestor, then $G_{i,j}$ is set to 0. The time since common ancestry between i and itself is defined to be 1. In an idealized genome with one chromosome in which all genes are different from all others, $G = I_n$, since the only match for gene i is with itself. Immediately after a full-genome duplication, if the new genes are indexed from $n+1$ to $2n$, the new GSM will be the block-identity matrix, $G' = \left(\begin{smallmatrix} I_n & I_n \\ I_n & I_n \end{smallmatrix}\right)$. The off-diagonal elements represent paralogous pairs. The ℓ_1, or Manhattan, distance between two paralogous pairs G_{i_1,j_1} and G_{i_2,j_2} is $|i_2 - i_1| + |j_2 - j_1|$.

Figure 1a shows that the ℓ_1 distance between each paralogous pair and its nearest neighbor in the GSM is 2 immediately following global genome duplication. Subsequently, the initial symmetry of the newly polyploid genome is degraded by the recurrent processes of chromosomal rearrangement, gene deletion and sequence divergence between paralogous genes. Yet, even after a considerable amount of genomic change, many rows of the GSM will contain at least one paralogous pair for which the nearest neighbor is only a short distance away. Chromosomal rearrangements only break connections between, at most, four paralogous pairs (Figure 1b). Gene deletion, insertion of new duplications, and divergence of paralogs break the connections between only two (Figure 1c,d). Thus, the footprint of global duplication will be detectable in the nearest neighbor distribution long after the duplication event itself.

Local duplications involve small contiguous regions containing one or more genes rather than the whole genome. Unlike a global duplication, a local duplication creates only a small number of paralogous pairs. However, neighboring paralogous pairs are still created when two or more genes are duplicated. Collectively, the regions of paralogy created by local duplication differ from those generated by global duplication in a number of ways. For example, sets of paralogous pairs created by different local duplication events will have been subject to deletion, divergence and rearrangement for differing lengths of time. In fact, the coincidence of estimated divergence times among different paralogous pairs has been used as evidence for polyploidy in maize (Gaut and Doebley, 1997) and yeast (Keogh et al., 1998). Also if local duplications events occur randomly or are are aggregated over the genome, the distribution of genes with paralogous pairs that have near neighbors will be highly uneven. Since it is likely that local duplications

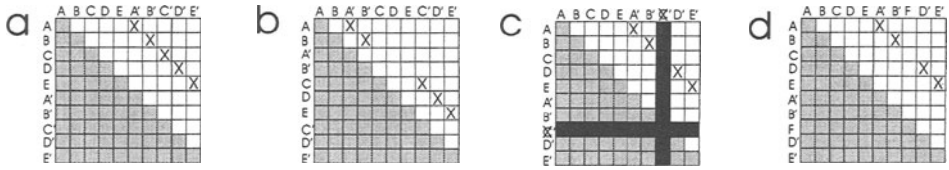

Figure 1. Evolution of the GSM. Paralogous pairs are marked with an X. The main diagonal and the lower triangular matrix are redundant and are shaded out. (a) Immediately after a global duplication. Each paralogous pair is one cell over and one cell up or down from its nearest neighbor. (b) Following a rearrangement event that switches the order of blocks CDE and $A'B'$. (c) Following deletion of C. (d) Following divergence of C and C', the latter now denoted by F. Note the minimal effect the latter three operations have on the distances between neighboring paralogous pairs.

are a recurrent process in all genomes, we must be able to spot such differences in order to detect the pattern of ancient global duplication through the fog created by more recent local duplications.

2. Genome evolution model

In this section, we specify more precisely the properties of the Markov chain model on which our simulations are based. The chain allows for local duplication, gene deletion, gene divergence, and simple gene order rearrangement. Global duplication events are outside the normal operation of the chain but can be externally introduced. Finally, we define nearest neighbor statistics for a GSM.

2.1. The GSM

Our model of genome evolution is a discrete-time Markov chain over the space of symmetric integral matrices. At a given time t, the $n_t \times n_t$ GSM $G(t)$ defines the similarity between the n_t genes found in the genome at that time. The basic intuition, as described above, is that the (i, j) entry of $G(t)$ is greater than zero if it is apparent that genes i and j are related by ancestry, and zero otherwise. For simplicity, we assume that the genome has no boundaries between chromosomes and only one correct linear order.

We initialize the process at time $t = 0$ with the $n_0 \times n_0$ identity matrix I_{n_0}, where n_0 is a parameter of the model; thus, genes initially are related by ancestry only to themselves.

2.2. Local duplication, deletion, divergence and rearrangement

We first model recurrent local duplication. Here, we assume that some number of contiguous genes are duplicated and that the copy is inserted intact into a random position. One local duplication occurs at each timestep provided the genome is smaller than some upper bound. What little is known about the size distribution of local duplications in actual genomes suggests that they usually involve only two or three genes (Semple and Wolfe, 1999). Here we assume duplicated block sizes are drawn from an exponential distribution (Figure 2a). The left end of the block is chosen uniformly from among the possible positions within the genome and the insertion point for the duplicate is chosen uniformly from among all positions in the genome. Rarely, the new copy of the block may, by chance, be interposed inside the original block. As a result of this process, the size of the GSM grows by the number of duplicated genes. Also, paralogous pairs are created; duplicated genes are similar to themselves, paralogous to their originals and inherit the paralogous pairs to which their originals belong.

We do not specifically model tandem gene duplication, in which a single gene is copied and inserted in close proximity to the original. Tandem, or near-tandem, duplication appears to happen at a high frequency in actual genomes (Rubin et al., 2000; Semple and Wolfe, 1999), but tandem duplicates will seldom ever be remnants of ancient polyploidy and can only serve to confuse nearest neighbor analysis. The reason is that where tandem duplicates do occur, they result in spuriously close paralogous pairs. Therefore, we do not consider tandem duplication in our model and we take pains to remove tandem duplicates from actual data in Section 4.

Following local duplication, we then consider gene deletion via two modes. In the first mode, deletion is equally likely for all genes. This may not be very realistic. In particular, it seems reasonable to suppose that the deletion process tends to remove only functionally redundant genes and seldom removes the last remaining copy of a gene family. Therefore, for the second deletion mode, we assume that the probability of deletion is weighted by the number of paralogous pairs in a row (Figure 2c). In both cases, the number of genes deleted in each timestep is a Poisson distributed random variable provided the genome size is below some lower bound. Otherwise, it is zero. Mode two deletion weights are recalculated following each deletion event.

Next, we consider divergence, in which the sequence similarity between two paralogous genes declines below the point at which homology can be recognized. In the GSM, if i and j are sufficiently divergent, then $G_{i,j}$ and $G_{j,i}$ are set to zero. We attempt to capture some biological realism by allowing the divergence probability between two paralogs to be function of the time since common ancestry. We assume the form of the function to be linear until t_d, the time at which the probability reaches 1 (Figure 2d). In our model, divergence affects both $G_{i,j}$ and

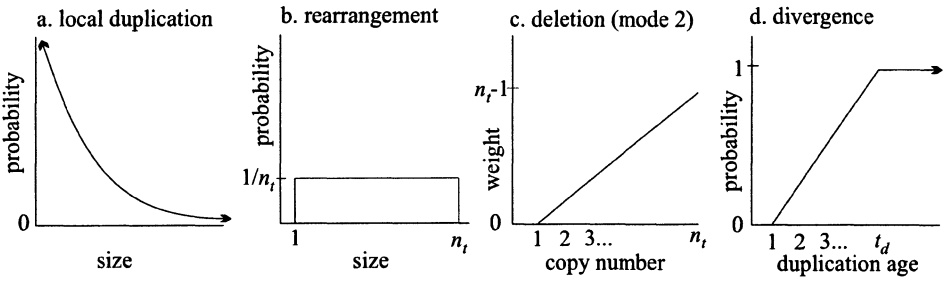

Figure 2. Probabilities of different events in the model: (a) exponentially distributed local duplication sizes, (b) uniformly distributed rearrangement packet size, (c) linear weights for mode 2 deletion probability as a function of copy number, (d) divergence probability for each paralogous pair as a function of duplication age (linearly increasing until t_d).

$G_{j,i}$, thus enforcing symmetry in the GSM.

Finally, we consider genomic rearrangements. Rather than model the large number of possible gene rearrangements in real genomes (inversions, reciprocal translocations, etc.), we allow a simple operation analogous to removing a packet of cards from a deck and reinserting it randomly into a new position. The size of the packet is chosen from a uniform distribution over the size of the genome. The first card in the packet and the insertion position of the packet are then chosen uniformly from among the possible choices.

2.3. Global duplication

We model local duplication, rearrangement, deletion and divergence as recurrent processes but we introduce a single global duplication event into the genome at a specified time point so that we may monitor the process as the symmetric genome decays. Global duplication of the genome at time t is a transformation of the GSM from $G(t) = A$ into $G(t+1) = \left(\begin{smallmatrix} A & A \\ A & A \end{smallmatrix}\right)$. This models a particular form of polyploidy, termed autotetraploidy, in which a single genome is duplicated exactly once.

2.4. Nearest-neighbor statistics

Consider the GSM $G(t)$ resulting from t steps of the above process, when $G(0) = I_{n_0}$, the $n_0 \times n_0$ identity matrix. We define the distance matrix, $D(t)$, whose entries are the distances between nearest paralogous pairs in the GSM. If the (i, j) entry of $G(t)$ is nonzero, let $D_{i,j}(t)$ be the minimum number of cells in $G(t)$ that one must move from the (i, j) entry to find another nonzero entry, where our movements are restricted to the right and down. (In genomes for which inversion is an allowable operation, we must also look to the right and up). We exclude the nonzero entries

along the diagonal; these all (except $G_{n_t,n_t}(t)$) have a neighbor one row over and one column down. For zero entries of the GSM, as well as for nonzero entries in the last row or column, we set $D_{i,n_t}(t)$ equal to some large constant.

This distance matrix captures information concerning the past history of the evolutionary process. In particular, if many pairs of neighboring genes both have copies which are themselves near neighbors, this is evidence for either global duplication or for a high rate of local duplication relative to deletion and divergence. To extract this information, we define the s-nearest neighbor array $N(t)$ to be the $n_t \times s$ matrix whose ith row consists of the s smallest entries in the ith row of $D(t)$. In the simulations described below, we explore how $N(t)$ differs between genomes that undergo global duplication and those that do not for a range of local duplication, rearrangement, deletion and divergence parameters and for a range of time intervals following global duplication.

2.5. A simpler null model

Here, we consider a much simpler null model in which the distribution of nearest neighbor distances can be easily derived. Consider a GSM of size $n_t \times n_t$, where each entry is nonzero with probability α; for small values of α and a large matrix, this corresponds essentially to choosing αn_t^2 points uniformly from an $n_t \times n_t$ grid.

It is possible to compute the distribution of the distance from a point to its nearest neighbor under some limiting assumptions. First, we assume that a nearest neighbor always is less far away from the point than its nearest boundary. Second, that the probability of a point being found at some distance away from the starting point is proportional to the distance. This is valid when the probability at any one site is extremely small. Under these assumptions, the probability of finding a nearest neighbor a distance d away from a given starting point is $\alpha(d+1) \times (1-\alpha)^{(n+2)(n-1)/2}$ if we only look to the lower right of our starting point and it is $\alpha(2d+1)(1-\alpha)^{n^2-1}$ if we look to the right both above and below. These density functions may be approximated by $\alpha(d)(1-\alpha)^{n^2/2}$ and $\alpha(2d)(1-\alpha)^{n^2}$, which are very nearly exact probability density functions. Their expectations may be computed by integration or truncated sums, yielding the result that the mean distance to a neighbor is proportional to $\sqrt{\alpha}$. The constants of proportionality are approximately 1.2 under the first model and .89 under the second.

3. Simulations

Here we simulate genomes run through time under the influence of recurrent local duplication, deletion, divergence and rearrangement. We run each process until t_c, the timestep when the number of paralogous pairs in the GSM reaches a stable plateau. The form of $N(t_c)$ is studied under a variety of plausible parameter values. At t_c, we introduce a global duplication event, reset the time step t to

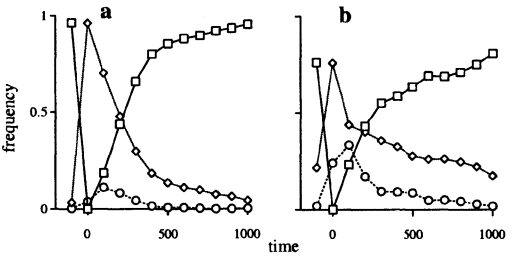

Figure 3. The frequencies of genes with zero (solid line, squares), one (dotted line, diamonds) or two (dashed line, circles) near neighbors within the 5% distance threshold for two different expected local duplication sizes: (a) $\hat{d} = 2$, (b) $\hat{d} = 10$. Global duplication is introduced at $t = 0$; the prior timestep shows the pre-global duplication equilibrium. The deletion rate is the inverse of the rate at which new genes are added by duplication, thus is five times higher in (b) than in (a). A rearrangement event occurs with probability 0.1 at each timestep, $t_d = 10^5$, and $n_0 = 1000$. Shown are the averages of three replicates.

zero, and calculate $N(t)$ at set values of t until the number of paralogous pairs again plateaus.

One local duplication occurs at each timestep; deletion, divergence and rearrangement rates are varied relative to this fixed reference. The number of deletions per time step is a random variable with expectation approximately equal to the expected size of a duplication. The maximum divergence time t_d is a run-dependent constant, as is the number of rearrangement events per timestep. The match of a gene with itself never diverges; these matches are not included when calculating $N(t)$, as they do not represent paralogous pairs.

Figure 3 shows the effect of the size distribution of local duplications on the frequencies of rows containing zero, one or two paralogous pairs with near neighbors. For each run, the distance threshold below which a near neighbor is counted is based on the density of the GSM at the pre-global duplication equilibrium. Specifically, it is chosen to require that in a random matrix of the same sparsity, only 5% of nonzero elements will have a nearest neighbor below the threshold. The results are shown for timesteps immediately prior to, at and following a global duplication event. When the expected size of duplicated blocks, $\hat{d}$, is 2, only 4% of the rows have one or two near neighbors. For a larger expected duplication size, $\hat{d}$=10, almost one quarter of the rows have either one or two near neighbors. Yet even this second case is easily distinguished from the pattern after a global duplication event. Initially, all rows have at least one near neighbor. This number decays to 50% in 300–400 timesteps. The frequency of rows with two near neighbors does not return to its pre-duplication state until after 500 timesteps. Surprisingly, the decay of the nearest neighbor distribution after global duplication appears to proceed at a slower rate for higher $\hat{d}$.

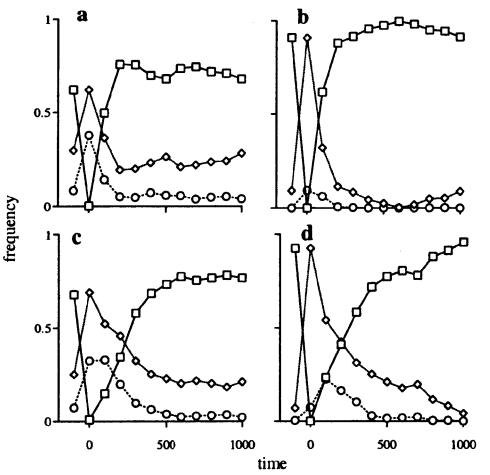

Figure 4. The frequencies of genes with zero (solid line, squares), one (dotted line, diamonds) or two (dashed line, circles) near neighbors within the 5% distance threshold for varying divergence rates and deletion modes. $t_d = 10^5$ in (a) and (b) and $t_d = 10^6$ in (c) and (d). Deletion is unweighted (mode 1) in (a) and (c) and weighted (mode 2) in (b) and (d). Global duplication occurs at $t = 0$; the prior timestep shows the pre-global duplication equilibrium state. A rearrangement occurs with probability 0.1 at each timestep, $\hat{d} = 2$ and $n_0 = 1000$. The deletion rate is the inverse of the rate at which new genes are introduced by duplication. One replicate for each parameter combination is shown.

Figure 4 shows the effect of the mode of gene deletion and the rate of divergence on the frequency of rows with zero, one or two near neighbors before and after global duplication where the 5% threshold distance is calculated as before. The results indicate that divergence is the major factor in the rate at which the nearest neighbor distribution decays following duplication while the mode of gene deletion plays the major role in determining the steady state properties of that distribution. The perturbation caused by global duplication mostly decays within 200 timesteps at $t_d = 10^5$, while it requires several hundred timesteps more at $t_d = 10^6$. When the probability of gene deletion is weighted by the number of duplications in a row, the frequency of rows containing zero paralogous pairs with near neighbors is considerably higher. In fact, there are no rows at all with two near neighbors prior to global duplication with mode 2 deletion for either value of t_d. Curiously, it appears that the equilibrium frequencies following duplication do not correspond to those prior in the case of mode 1 deletion.

4. Remnants of polyploidy in yeast?

It has been proposed that the genus *Saccharomyces*, which includes yeast, is derived from a tetraploid ancestor. Several lines of evidence support this hypothesis. The chromosome number of *Saccharomyces* is double that of the nearest related genera. Approximately half of the genome can be accounted for by nonoverlapping regions containing several closely spaced putative paralogs with conserved orientation relative to each other and to their centromeres. Gene order in a putative outgroup is consistent with the pattern reconstructed for the pre-duplication, pre-deletion ancestral yeast genome (Wolfe and Shields, 1997; Keogh et al., 1998; Seoighe and Wolfe, 1998, 1999). Thus, the evidence for a global duplication is quite strong; however, the possibility of multiple chromosomal duplications cannot be excluded.

Here we examine the nearest neighbor distribution for the GSM of *Saccharomyces cereviseae*, using a dataset prepared by the Saccharomyces Genome Database (Cherry et al., 1998). These data are the Smith-Waterman alignment scores with associated P values less than 0.01 for all pairwise alignments of the 6,210 open reading frames in the annotation database as of March 18, 1998. The resulting GSM is slightly asymmetric as the alignment scores are themselves asymmetric.

To reduce the complications due to tandem duplications, we collapse tandem and nearly-tandem duplicates into single composite genes. We join any two genes within 6 genes of each other that are connected by one or two matches. This seems reasonable, since duplicates spaced at this distance are very unlikely to have been duplicated during a global event and tandem arrays of duplicated genes in yeast tend to be quite small (Goffeau et al., 1996). Non-tandem duplications in actual genomes may tend to be more closely spaced than expected by this model (Semple and Wolfe, 1999), thus creating spurious near neighbors. But these artifacts are probably not counted in our analysis due to the strict distance threshold employed. The elements of the composite gene are assigned the maximum value in the corresponding column of all the collapsed genes. This seems reasonable given that the rows of the GSM tend to be very similar for tandem copies. This procedure removes 158 genes, approximately 2.5% of the total.

A further processing step is required to correct for the minority of genes that have very large numbers of matches throughout the genome. It is likely that such promiscuously matching genes do not reflect the local duplication process but rather possess highly conserved domains of atypical sequence. Therefore, we have required that each row have 5 or fewer nonzero entries and have kept only those with the highest scores.

Whereas the nearest neighbor distance of a paralogous pair in the simulated data may be computed by only counting neighbors to the right and down, we now must compute the nearest neighbor distance looking over a broader area. For, contrary to the model, gene order may be inverted between duplicate regions in the yeast genome. As a result, two paralogous pairs duplicated together may not

Table 1. The frequency of rows containing x nonzero entries with near neighbors at or closer than the 5% distance threshold for the GSM of yeast and for a random matrix of the same size (6052) and density (3×10^{-4}).

	0	1	2	3	4	5+
yeast	79.5	16.8	2.7	0.6	0.2	0.1
random	91.2	8.4	0.4	0	0	0

be necessarily be oriented to the upper left and lower right of one another. Thus, we allow nearest neighbors to be found either up and to the right or down and to the right of a nonzero element.

After completion of the above processing steps, there are 11,076 matches among the 6,052 remaining genes, an average of 1.83 per gene. For a random matrix of this size, where cells are nonzero with probability $11,076/6,052^2$, the mean distance of a cell in the matrix to its nearest neighbor is approximately 50.9. Approximately 95% of paralogous pairs have a neighbor within a distance of 100 cells. Only 5% have a neighbor within 17 cells, the value used as the threshold for defining a near neighbor. Note that since chromosomes are concatenated for analysis, near neighbors may, on rare occasions, span chromosome boundaries.

The results for yeast clearly deviate from the expected pattern that would be obtained by counting near neighbors in a random matrix (Table 1). Most notably, over 20% of the yeast genes have one or more near neighbors, whereas less than 10% would be expected in a random matrix. The frequency of genes with with one or more near neighbors is higher for yeast in each category.

Comparison by eye of the results for yeast with those obtained for the simulations (Figures 3 and 4) indicate that the best fit to the data would be for values of $\hat{d}$ not much greater than 2 and for values of t_d not much less than 10^6. While the data are consistent with the apparent steady state values for deletion mode 1 in Figure 4c, we favor the alternative deletion mode for its biological realism. For deletion mode 2, the yeast data most closely resemble the simulated genome 400–800 timesteps after global duplication (Figure 4d). We conclude that we may be detecting the footprint of ancient polyploidy in yeast, but more empirical data are needed to inform the model before such hypotheses can be rigorously put to the test.

5. Empirical data for parameterizing the model

We have presented a relatively simple null model for the evolution of genomic self-similarity matrices evolving under the basic forces of recurrent local duplication, deletion, divergence and rearrangement. In addition, we have explored the consequences of introducing unique global duplication events to such a process. Here, we consider what is known regarding the forces under consideration to determine

where the empirical holes remain to be filled.

Rearrangement rate estimates, based primarily on application of the Nadeau and Taylor (1984) model to sparse mapping datasets, appear to be bimodal and to fall in the relatively well-defined range of 0.15 to 1.3 rearrangements per million years for species of the same ploidy (Lagercrantz, 1998; Nadeau and Sankoff, 1998). Rearrangement rates show significant variation among lineages (Ehrlich et al., 1997) and are apparently elevated after global duplication events (Lear and Bailey, 1997). We suspect, however, that the nearest neighbor distribution is relatively insensitive to natural variation in rearrangement rate; a number of rearrangements within an order of magnitude of the number of genes is necessary to cause significant decay to the pattern established by a global duplication event. This would require several hundred rearrangement events even for the relatively tiny yeast genome, yet it is estimated that yeast has experienced fewer than 100 rearrangements in the 100 million years since duplication (Seoighe and Wolfe, 1998; El-Mabrouk et al., 1999).

The extent gene deletion following global duplication has been estimated by a number of studies using different pairs of organisms and different methodologies. One study has estimated, from the size distribution of gene families in humans and mice, that approximately half of the duplicates still present from a putative ancient global duplication have since been deleted in one of the two lineages, which diverged approximately 250 million years ago (Nadeau and Sankoff, 1997). Another study estimated that less than 28% of duplicated genes have been deleted since the allopolyploidy event that gave rise to maize just over 10 million years ago (Ahn and Tanksley, 1993; Gaut and Doebley, 1997). And a third study estimated that 92% of duplicated genes have been deleted over the past 100 million years since that duplication of the yeast genome (Seoighe and Wolfe, 1998). These suggest that the deletion probability for one of a duplicate pair ranges from approximately 0.003 (mouse-human) to 0.03 (maize) per million years. It has been suggested that gene deletion events are concentrated shortly after genome duplication (Matzke et al., 1999). If that were true, the human-mouse estimate would be the only one with relevance to the recurrent gene deletion process; but the grossly different proportions of deleted genes between maize and yeast appear to contradict this hypothesis.

The rate of divergence between duplicate pairs depends not only upon the rate of sequence evolution, but also on the methodology for detecting duplicates. The rate of amino acid substitution is approximately 1–5 per codon per billion years (Marti and Binns, 1998; Li, 1997), but varies greatly both within and among protein sequences. The use of substitution matrices such as those of the BLOSUM and PAM series allow the commonly used protein sequence matching algorithms to detect ancient relationships among proteins even when alignable sequences are short and identical amino acids are few. However, the quantitative strength of a match, as determined by pattern matching or optimal alignment algorithms, may not be a reliable measure of the relative phylogenetic distances between the

sequence under consideration and the set of sequences that show a detectable similarity. Studies need to be done on the distribution of divergence times represented by the matches seen in datasets such as this one.

The rate and size distribution of local duplication is perhaps the most important parameter to understand and the one for which the least empirical data exist. Small numbers of collinear duplicates comprising large numbers of genes, as is seen in *Arabidopsis thaliana*, suggest a past history of global duplication. Very large local duplications could also conceivably result in such a pattern, but there is little evidence for their existence. For example, in a study of the two-thirds completed genomic sequence of *Caenorhabditis elegans*, three duplications involving three genes apiece were the largest that were observed (Semple and Wolfe, 1999). If this is a typical sample of local duplication events, then one would hardly expect a large number of neighboring paralogous pairs to be generated by local duplication alone. Thus, it may be that what eventually allows us to detect the footprint of ancient polyploidy in contemporary genomes will be as much the lightness of the footprint left behind by local duplication as the left of that left behind by global duplication.

Acknowledgments

The authors thank Prof. R. Durrett (Cornell University) for helpful discussions concerning the ideas in this work. Research of the second author has been supported by an NSF Graduate Research Fellowship, NSF grants CCR-970029, DMS-9805602, DBI-9872617, and ONR grant N0014-96-1-00500.

References

AHN, S. AND TANKSLEY, S. D. 1993. Comparative linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences USA* 90:7980–7984.

CHERRY, J. M., ADLER, C., BALL, C., CHERVITZ, S. A., DWIGHT, S. S., ET AL. 1998. SGD: Saccharomyces genomes database. *Nucleic Acids Research* 26:73–80.

EHRLICH, J., SANKOFF, D., AND NADEAU, J. H. 1997. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics* 147:289–296.

EL-MABROUK, N., BRYANT, B., AND SANKOFF, D. 1999. Reconstructing the pre-doubling genome. *In* Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99), pp. 154–163. ACM, New York.

GAUT, B. S. AND DOEBLEY, J. F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences USA* 94:6809–6814.

GOFFEAU, A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J. D., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H., AND OLIVER, S. G. 1996. Life with 6000 genes. *Science* 274:546, 563–567.

GRANT, D., CREGAN, P., AND SHOEMAKER, R. C. 2000. Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proceedings of the National Academy of Sciences USA* 97:4168–4173.

KEOGH, R., SEIOGHE, C., AND WOLFE, K. H. 1998. Evolution of gene order and chromosome number in *Saccaromyces*, *Kluyverromyces* and related fungi. *Yeast* 14:443–457.

LAGERCRANTZ, U. 1998. Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* 150:1217–1228.

LEAR, T. L. AND BAILEY, E. 1997. Localization of the U2 linkage group of horses to eca 3 using chromosome painting. *Journal of Heredity* 88:162–164.

LI, W.-H. 1997. Molecular Evolution. Sinauer, Sunderland MA.

MARTI, E. AND BINNS, M. 1998. Horse genome mapping: a new era in horse genetics? *Equine Veterinary Journal* 30:13–17.

MATZKE, M. A., SCHEID, O. M., AND MATZKE, A. J. M. 1999. Rapid structural and epigenetic changes in polyploid and aneuploid genomes. *BioEssays* 21:761–767.

NADEAU, J. H. AND SANKOFF, D. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147:1259–1266.

NADEAU, J. H. AND SANKOFF, D. 1998. Counting on comparative maps. *Trends in Genetics* 14:495–501.

NADEAU, J. H. AND TAYLOR, B. A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA* 81:814–818.

RUBIN, G., YANDELL, M. D., WORTMAN, J. R., MIKLOS, G. L. G., NELSON, C. R., ET AL. 2000. Comparative genomics of the eukaryotypes. *Science* 287:2204–2215.

SEMPLE, C. AND WOLFE, K. H. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *Journal of Molecular Evolution* 48:555–56.

SEOIGHE, C. AND WOLFE, K. H. 1998. Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences USA* 95:4447–4452.

SEOIGHE, C. AND WOLFE, K. H. 1999. Updated map of duplicated regions in the yeast genome. *Gene* 238:253–261.

WOLFE, K. H. AND SHIELDS, D. C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

USDA-ARS CENTER FOR BIOINFORMATICS AND COMPARATIVE GENOMICS, CORNELL UNIVERSITY, ITHACA NY 14853
E-mail address: tv23@cornell.edu

DEPARTMENT OF COMPUTER SCIENCE, CORNELL UNIVERSITY, ITHACA NY 14853
E-mail address: snowman@cs.cornell.edu