The Role of Mutational Dynamics in Genome Shrinkage

Milan J. A. van Hoek and Paulien Hogeweg

Theoretical Biology/Bioinformatics Group, Utrecht University, Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands

Genome shrinkage occurs after whole genome duplications (WGDs) and in the evolution of parasitic or symbiotic species. The dynamics of this process, whether it occurs by single gene deletions or also by larger deletions are however unknown. In yeast, genome shrinkage has occurred after a WGD. Using a computational model of genome evolution, we show that in a random genome single gene deletions cannot explain the observed pattern of gene loss in yeast. The distribution of genes deleted per event can be very well described by a geometric distribution, with a mean of 1.1 genes per event. In terms of deletions of a stretch of base pairs, we find that a geometric distribution with an average of 500–600 base pairs per event describes the data very well. Moreover, in the model, as in the data, gene pairs that have a small intergenic distance are more likely to be both deleted. This proves that simultaneous deletion of multiple genes causes the observed pattern of gene deletions, rather than deletion of functionally clustered genes by selection. Furthermore, we found that in the bacterium *Buchnera aphidicola* larger deletions than in yeast are necessary to explain the clustering of deleted genes. We show that the excess clustering of deleted genes in *B. aphidicola* can be explained by the clustering of genes in operons. Therefore, we show that selection has little effect on the clustering of deleted genes after the WGD in yeast, while it has during genome shrinkage in *B. aphidicola*.

Introduction

Massive gene loss has been a major evolutionary driving force in the evolution of many different species. After a whole genome duplication (WGD), most duplicate genes are lost and only a relatively small percentage of genes is kept in duplicate. WGD has been shown in the ancestry of yeast (Wolfe and Shields 1997; Kellis et al. 2004), plants (e.g., in *Arabidopsis thaliana* up to 3 WGD's can be distinguished, Initiative 2000; Bowers et al. 2003), vertebrates (Dehal and Boore 2005) (although this is still debated), the teleost fishes (Amores et al. 1998; Taylor et al. 2001), and the ciliate *Paramecium tetraurelia* (Aury et al. 2006). Massive gene loss occurred in all these lineages.

Genome shrinkage has first been studied in *mycoplasma*, prokaryotic parasites that lack a cell wall and have very small genomes. Contrary to initial expectations, these bacteria evolved from free-living bacteria with larger genomes (Maniloff 1983; Woese 1987). Because these bacteria are intracellular parasites, many metabolic functions can be performed by their hosts, which renders many genes nonfunctional. This caused the dramatic shrinkage of their genomes. This process of genome shrinkage also occurred in, for example, *Buchnera aphidicola*, an endosymbiont of aphids (Moran and Mira 2001).

An open question concerning the mutational dynamics of genome shrinkage is whether genes are lost one by one, via point mutations and subsequent pseudogenization, or that multiple neighboring genes are lost in one event, for example, via unequal crossing over. Here we study whether the pattern of gene deletions, as observed in different yeast species and *B. aphidicola* can be explained by single gene deletions or whether simultaneous deletion of neighboring genes is needed.

When large stretches of subsequent deleted genes are observed in an alignment between genomes, from which one experienced massive gene loss, this could be due to simultaneous deletion of several neighboring genes or to subsequent deletions of single genes. We constructed a computational model describing genome shrinkage. We use different probability distributions of deletion sizes (in number of genes) to simulate genome shrinkage. We find that nor in yeast, nor in *B. aphidicola*, single gene deletions in a nonstructured genome can explain the observed pattern of gene deletions. We find that when we allow for deletions of multiple adjacent genes, we can explain the pattern of gene loss satisfactorily.

An important issue in genome shrinkage is whether selection, neutral evolution, and/or the mutational dynamics determine which and how many genes are deleted. For endosymbionts, it has been proposed that gene deletion is much more frequent than gene duplication and that this causes the reduction of genome size in these bacteria (Mira et al. 2001) and selection does not play a large role. However, another study reveals that selection is also important (Delmotte et al. 2006). Gene evolution and genome reduction after WGD (in particular which genes are kept in duplicate and which are deleted) are most often explained by selection (Blomme et al. 2006; Lin et al. 2006; Thomas et al. 2006). In any case, it seems obvious that genome shrinkage is not a completely neutral process because some genes will be more important to retain (in duplicate) than others. Here we study whether the pattern in gene deletions both in yeast and *B. aphidicola* can be explained by the mutational mechanism alone or that selection is needed to explain this pattern.

Because gene order is not random, selection could cause clustering of gene deletions. This nonrandom gene order is very well known for prokaryotes, in which genes are clustered in operons. Also for eukaryotes it has been shown that genes are functionally clustered (Hurst et al. 2004). To distinguish between these 2 explanations, we constructed a second model for genome shrinkage. In this model, deletions are on the basis of base pairs, so instead of genes, a stretch of base pairs is deleted. We find that we can explain the pattern of gene loss satisfactorily when large deletions (in the order of 100–1000 bp) are frequent enough, both for yeast and *B. aphidicola*. This model predicts that genes which have a short intergenic distance are more likely to be simultaneously deleted. Furthermore, small genes are more likely

Key words: whole genome duplication, genome shrinkage, *Saccharomyces cerevisiae*, *Buchnera aphidicola*, gene deletion, computational model.

E-mail: m.j.a.vanhoek@bio.uu.nl.

Mol. Biol. Evol. 24(11):2485–2494. 2007
doi:10.1093/molbev/msm183
Advance Access publication September 3, 2007

to be deleted simultaneously with other genes than large genes. We show that both these predictions hold in yeast and therefore find strong evidence that deletion of stretches of base pairs underlie gene deletions in yeast.

We checked this result by studying gaps of base pairs in pseudogenes in *Saccharomyces cerevisiae*, in a similar way as has previously been done for *B. aphidicola* (Gomez-Valero et al. 2007). We used the distribution of gap sizes (in base pairs) that we observed in these pseudogenes in our model. We find that, using this distribution, we can satisfactorily explain the amount of clustering of deleted genes in *S. cerevisiae*. This all shows that the clustering of gene deletions after the WGD in *S. cerevisiae* is caused by the mutational dynamics and not by the clustering of functionally related genes.

For *B. aphidicola*, we observed that on average longer deletions in terms of base pairs were needed (≈1000 vs 500-600, which corresponds to on average 1.6 gene vs. 1.1 gene in *S. cerevisiae*). We found that these longer gaps are most likely due to the operon structure that is present in prokaryotes, which causes deleted genes to be clustered in the genome.

B. aphidicola is an endosymbiont of aphids. It is a relative of the Enterobacteriaceae, like *Escherichia coli*. *B. aphidicola*'s genome is practically a subset of the genome of *E. coli* (Moran and Mira 2001). It has 550 genes, whereas *E. coli* has 4488 genes. It is believed that *B. aphidicola* lost all these genes after becoming an endosymbiont. Delmotte et al. (2006) constructed an alignment of the genome of *B. aphidicola* with respect to the reconstructed last common ancestor of endosymbionts and their free-living relatives. They also measured gap sizes of deleted genes, which we use here.

Lengths of genes and intergenic regions are downloaded from National Center for Biotechnology Information, except for the gene lengths and intergenic lengths of *A. gossypii*, which we downloaded from the "Ashbya Genome Database" (Hermida et al. 2005). For finding gaps in pseudogenes, we used a previously published list of identified pseudogenes in (Lafontaine et al. 2004). As was done in this article, we aligned the pseudogenes with their homologous open reading frame (ORF) using DIALIGN 2.2.1 (Morgenstern 2004) and subsequently we counted the gaps in the alignment.

Materials and Methods

Data

We obtained the distribution of gap sizes (in genes) for a number of species. For the yeast species, we used the Yeast Gene Order Browser (YGOB), version 1.0. (Byrne and Wolfe 2005). This is an alignment between 3 yeast species that underwent a WGD (*S. cerevisiae*, *Saccharomyces castellii*, and *Candida glabrata*) and 4 that did not (*Ashbya gossypii*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Saccharomyces kluyveri*).

The alignment in YGOB is local. The data in the YGOB consists of "pillars." A pillar represents one ancestral gene, which is duplicated in the WGD. For every post-WGD species, 0, 1, or 2 duplicates of this gene may be conserved. Every pre-WGD species may or may not have this gene. Therefore, a pillar consists of 10 slots that may or may not be filled. If only one slot in a pillar is filled, we discarded that pillar because then the most parsimonious explanation is that this gene is a recent duplication instead of already present at the WGD.

We downloaded the alignment in a region of 50 pillars around every gene of every pre-WGD species. As mentioned by Byrne and Wolfe (2006), focusing on pre-WGD species is the best way to view the syntenic context. We looked which post-WGD genes were in the same pillar as the in-focus pre-WGD genes. Then we counted the gap size left and right of these post-WGD genes. In this way of course gaps are counted twice, and we divided the resulting distribution by 2.

Focusing on different pre-WGD species sometimes gives a different result (as is noted by Byrne and Wolfe 2005). Therefore, we focused on every pre-WGD species (except *S. kluyveri* because it is only sequenced to 4× coverage and it is clear from YGOB that many genes are missing) and averaged the outcome. However, the 3 resulting distributions were very similar, which confirms the consistency of our results.

Model

In all species we studied, we observed that gap sizes of deleted genes are not geometrically distributed, which would be the case if gene loss occurred due to single gene deletions in a random way. To study how the pattern of gap sizes is influenced by the sizes of allowed deletions, we constructed a computational model to simulate massive gene loss. Only gene deletion is taken into account. A genome is represented by a list of genes. We impose a certain probability distribution of deletion sizes. We can, for example, only allow for single gene deletions or also for larger deletions (of more genes) using a geometrical distribution. A simulation is terminated when the number of deleted genes equals the number of deleted genes in the data set we are interested in. The number of genes a simulation starts with is simply the number of conserved genes plus the observed number of deleted genes in the data set.

When reconstructing genome evolution after a WGD, we take both duplicates of the genome into account. Thus, the starting genome then consists of 2 homologous "genomes." We assume selection against deletion of both genes of an ohnolog pair (an ohnolog is a paralog that arose through a WGD). To account for the fact that in yeast, in a relatively small number of cases both ohnologs are deleted, we assume a certain probability r that a certain ohnolog pair can be deleted, such that the actual number of double deletions corresponds to the actually observed number. For *S. cerevisiae*, we used $r = 0.12$, for *S. castellii* and *C. glabrata* $r = 0.17$.

At the end of a simulation, we count the distribution of gap sizes in the evolved genome. We do this 1000 times and take the average of these 1000 simulations. We now treat this average distribution as a theoretical distribution, to which we can compare the observed distribution. We use a χ^2 test to determine the probability that a certain

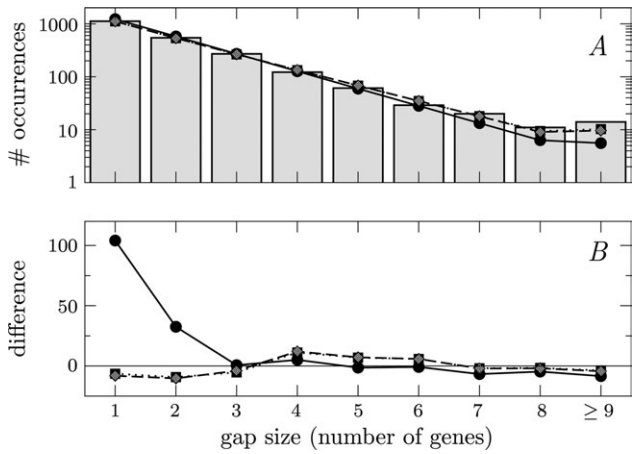

FIG. 1.—The observed and computationally obtained gap size distributions for *Saccharomyces cerevisiae*. (*A*) Gray bars: the observed distribution of gap sizes in YGOB. Solid line, circles: computationally obtained distribution, if only single gene deletions are allowed. Dashed line, squares: computationally obtained distribution, deletions are gene based. Deletion sizes are geometrically distributed ($q = 0.92$). Dotted line, diamonds: computationally obtained distribution when deletions occur on the basis of base pairs. Base pair deletions are geometrically distributed ($q = 0.0019$). (*B*) The difference between the computationally obtained and the observed distribution.

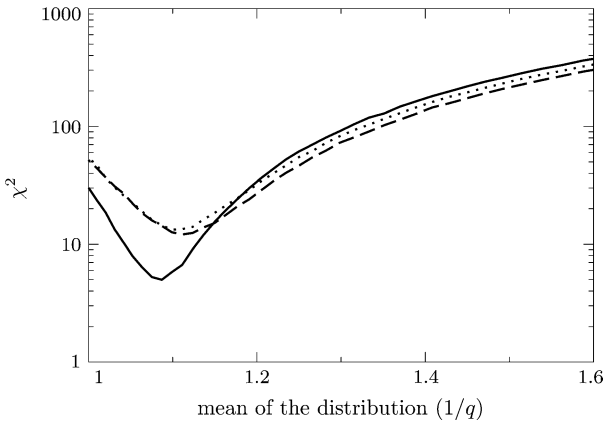

FIG. 2.—χ^2 as a function of the mean of the distribution (mean = $1/q$). Solid line, *Saccharomyces cerevisiae*; dashed line, *Candida glabrata*; and dotted line, *Saccharomyces castellii*.

computationally obtained distribution can explain the observed distribution.

We also constructed a more detailed model of genome shrinkage. In this model, we delete base pairs, instead of genes. The genome now consists of genes and intergenic regions of certain lengths. In the case of yeast, the length of the genes and intergenic regions are randomly drawn from the observed genic and intergenic length distributions in the pre-WGD species *K. lactis*, whereas for *B. aphidicola*, we use lengths of *E. coli*, a close relative in which genome shrinkage did not occur.

This model works identically as the previously explained model. Randomly a base pair is picked from the genome and a deletion length from a certain distribution and, if possible, the deletion is carried out. If a gene has lost some base pairs due to a deletion, we assume that gene becomes nonfunctional and is degraded instantly and all base pairs from that gene are deleted. If, however, a part of an intergenic region is deleted, the rest of the intergenic region stays intact. This model is consistent with the observation that very few pseudogenes belonging to deletions after the WGD in *S. cerevisiae* can still be identified (Lafontaine et al. 2004). We also tried a rule in which a gene, of which a certain part is deleted, becomes a pseudogene (intergenic region). This model gave almost identical results to the instant gene degradation model.

Results

Yeast

First we look at the distribution of gap sizes in *S. cerevisiae*. The observed distribution is shown by the gray histogram in figure 1*A*. First we simulated genome shrinkage when only single gene deletions are allowed. This is done first with deletions occurring on the gene level. We then find the distribution indicated by the solid line. This distribution is simply a geometric distribution $f(n) = P(1 - P)^n$, where P equals the probability that a gene is conserved (the number of conserved genes divided by the total number of conserved genes plus the total number of deleted genes) and n the observed gap size.

From figure 1*A*, it is clear that there are too many large gaps in the observed distribution to be described by single gene deletions alone. In figure 1*B*, the difference between the observed and the computationally obtained distribution is shown. From this picture, it is clear that there are also too few small gaps. If we compare the distribution caused by single gene deletion and the experimentally observed distribution, we find that $\chi^2 = 31$, which gives a P value of $P = 0.0001$. Note that in the last bin in figure 1, we added all gaps larger than 9 genes. Also when calculating χ^2, this binning was used, to avoid very small numbers of observations per category, which is necessary for a χ^2 test. The exact location for this cut-off, however, does not essentially change the results.

Allowing for larger deletions may explain the data better. A natural assumption for a distribution of deletions is a geometric distribution. If a deletion event has a fixed probability to terminate after every gene, the resulting distribution of sizes of deletion events is geometric. Note the important difference between the distribution of deletion events and the distribution of observed gap sizes. Assuming only single gene deletion leads to a geometric distribution of observed gap sizes. Assuming a geometric distribution of deletion events leads to a different distribution of observed gap sizes, due to clumping of deletion events into larger gaps.

So we assume that sizes of deletion events are geometrically distributed with parameter q. This parameter indicates the probability that a deletion event is a single gene deletion. We can try to fit the observed distribution by varying q and calculating χ^2 between the observed and the simulated distribution. In figure 2, we show how χ^2 depends on the mean of the distribution used (which is equal to $1/q$). For the interpretation of the results, it is crucial that χ^2 has a single, definite minimum when changing q, which is indeed the case. The best P value is found for

Table 1
Statistics of Fit for All Yeast Species

	$q = 1$		Geometric Gene Deletion			Geometric Base Pair Deletion		
Species	χ^2	P Value	q	χ^2	P Value	q	χ^2	P Value
Saccharomyces cerevisiae	31	0.0001	0.92	4.8	0.78	0.0019	5.8	0.67
Saccharomyces castellii	56	<0.0001	0.90	13	0.11	0.0017	16	0.05
Candida glabrata	58	<0.0001	0.90	12	0.15	0.0016	14	0.09

$q = 0.92$. Indeed this distribution describes the data very well, with $\chi^2 = 4.8$, $P = 0.78$ (dashed line in fig. 1).

To check in a reverse way whether the single gene deletion model can describe the data, we fitted all 1000 distributions, obtained from the simulations using only single gene deletions, against q, thus treating these as observed distributions. In 0.7% of the cases, the best-fit value of q was 0.95, in all other cases, the resulting q was higher. A q value of 0.92 was therefore never found, and it is therefore very improbable that such a mechanism can lead to a distribution which is best fitted by $q = 0.92$.

For the other 2 post-WGD yeast species available, *S. castellii* and *C. glabrata*, we find very similar results. The results of all 3 post-WGD yeast species are summarized in Table 1.

Interestingly, we find almost equal q values for the different species, which indicates the consistency of the result. However, the fits are always best for *S. cerevisiae*, maybe due to the better annotation of genes in this species.

Deletion of Base Pairs

From the previous section, we conclude that gene deletions in *S. cerevisiae* are clustered within the genome. However, from this model, we cannot conclude that the mutational dynamics are responsible for this clustering. The results found in the previous section could just as well reflect the fact that genes are functionally clustered in the genome of *S. cerevisiae* (Hurst et al. 2004). When a certain gene is deleted, the chance of neighboring genes to be deleted might increase and this effect would lead to the clustering of deleted genes. When this explanation would be correct, selection would cause the clustering of deleted genes, instead of the mutational dynamics.

Therefore, we constructed a model in which the mutational dynamics are incorporated more realistically. Because mutations obviously act on the level of base pairs instead of on the level of genes, we modified the gene deletion model to a base pair deletion model. In this model, adjacent genes can be deleted simultaneously due to a deletion of a long stretch of base pairs. In this way, we hope to find out whether the mutational dynamics or the clustering of functionally related genes cause the observed clustering of deleted genes.

We tested whether this model, using base pair deletions, could also cause the observed pattern in gap sizes. Again we assume a geometric distribution of base pair deletions, corresponding to a fixed probability of termination of a deletion event after each base pair.

Again we fitted the resulting distribution of gap sizes (still in genes) to the observed distribution, by varying q, the probability that a deletion has a length of only one base pair and therewith the mean length of the deletions. We found that for *S. cerevisiae*, a mean length of 530 bp fits the data best. The obtained distribution is also shown in figure 1. For *S. castellii* and *C. glabrata*, we found a mean deletion length of 600 and 620 bp, respectively.

As expected, a geometric distribution of base pair deletion also describes the data nicely. This gives us a mechanistic explanation for the observed gap sizes. More importantly, this model also provides predictions about which genes are more likely to be deleted than others. Therefore, we can check whether these predictions hold in the data and in this way prove whether or not the above mechanism is responsible for the large gaps. For example, neighboring genes that have a short intergenic region in between are more likely to be both deleted than genes that have long intergenic regions in between because a deletion has to reach the next gene in order to delete both genes simultaneously.

In figure 3*A*, the frequency distribution of intergenic regions in *K. lactis* is shown. We are interested whether small intergenic regions are overrepresented if both neighboring genes are deleted after the WGD. Therefore, we calculated the difference between the frequency distribution for intergenic regions for which both neighboring genes are deleted and the frequency distribution of all intergenic regions. We did this for each post-WGD species and averaged the 3 curves. This gives the dashed curve in figure 3*B*. This curve indicates the over or underrepresentation of genes of a certain length if both neighboring genes are deleted. For the simulations, we performed exactly the same procedure, except that we averaged over 1000 simulations, for each post-WGD species, which gives the solid curve in figure 3*B*.

We find that the simulations predict the observed outcome surprisingly well. Both the overrepresentation of small intergenic regions (<500 bp) and the underrepresentation of intergenic regions of intermediate length are captured well by the model.

A second prediction of the model is that when we compare genes in gaps that consist of only one gene with genes in gaps that consist of more genes, small genes will be relatively overrepresented in large gaps compared with gaps consisting of only one gene. This is because when a large gene is picked to be deleted, the probability is higher that the whole base pair deletion will not reach the end of the gene. Therefore, large genes will be relatively more often deleted on their own.

In the simulations, we clearly observe this phenomenon (see fig. 4*B*). These curves are made in a similar way as the curves in figure 3*B*. Small genes are underrepresented in the set of genes in gaps of length one, whereas genes of

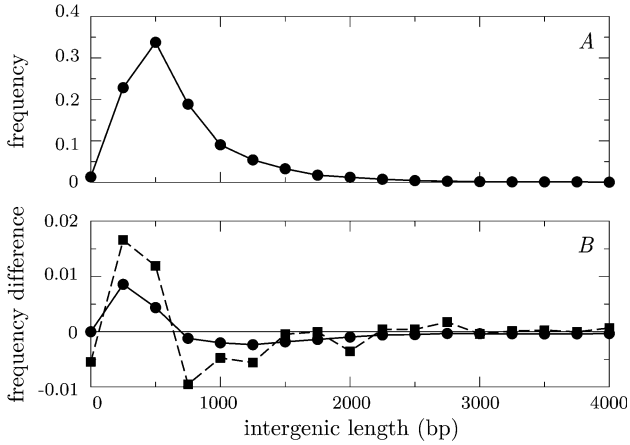

FIG. 3.—Intergenic length between genes of *Kluyveromyces lactis*. (*A*) Frequency distribution of all intergenic regions of *K. lactis* in YGOB. (*B*) Dashed line, squares: difference between the distribution of intergenic lengths between 2 genes in *K. lactis*, which are both deleted in a post-WGD species and the distribution of all intergenic regions (the distribution in fig. 3*A*). Solid line, circles: the resulting curve of the simulations.

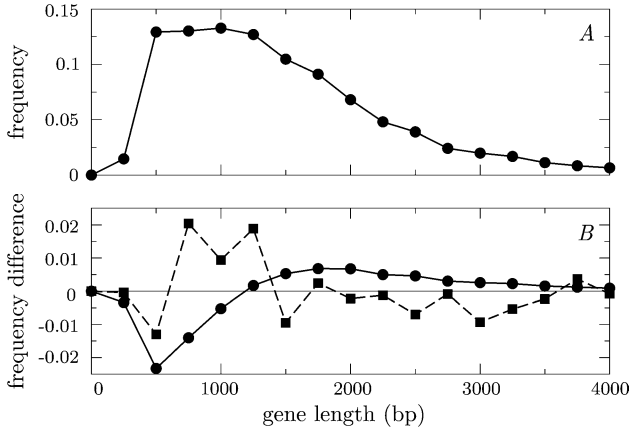

FIG. 4.—Length of genes in *Kluyveromyces lactis*. (*A*) Frequency distribution of gene lengths in *K. lactis*. (*B*) Dashed line, squares: the difference between the gene length of genes that are deleted in gaps of size one and genes that are deleted in gaps of a larger size. Solid line, circles: the resulting curve of the simulations.

intermediate length are overrepresented. Although the data are very noisy, the underrepresentation of small genes (<500 kbp) is precisely in correspondence with the model. The genes that are overrepresented in gaps of size one in the data are, however, smaller than we expect from the model (see below).

For both these observations to be meaningful, it is necessary that intergenic length and gene length are conserved between pre-WGD and post-WGD species. We checked this, and we indeed found good correlation for genic and intergenic length between *S. cerevisiae* and *K. lactis* (for intergenic length, we found a correlation coefficient of 0.65, and for gene length, a correlation coefficient of 0.98).

Influence of Gene Length on Deletion Probability

Inspired by the above result, we studied the effect of gene size on the probability to be deleted after a WGD in more detail. Previously, it has been found that genes which are duplicated in the human genome are on average shorter (Nembaware et al. 2002). The authors believe that this is likely a mutational effect.

In the same article, it is also mentioned that in yeast, this relationship between gene length and the probability of being kept as duplicate after a WGD was not found. On the contrary, the length of genes belonging to an ohnolog pair was even a bit larger than the average gene length.

We more precisely studied the relationship between gene length and being kept in duplicate after a WGD (see fig. 5). We divided all genes of *S. cerevisiae* in 3 groups, genes that are kept in duplicate after WGD, genes that are conserved on their own after WGD, and recent duplications (genes that have no homolog in a pre-WGD species). On average, ohnologs are slightly larger. Furthermore, it is very clear that recent duplications are on average shorter, just as was found for humans (Nembaware et al. 2002). Therefore, we find that there is a correlation between gene length and the probability of being kept as a duplicate after a WGD, but it is reversed with respect to single gene duplications. These results also hold for genes of *C. glabrata*.

The fact that ohnologs are longer than on average indicates that shorter genes might be more often deleted after a WGD than longer genes. To check whether this is correct, we looked at the length of pre-WGD species, instead of post-WGD. For each gene in *K. lactis*, we counted how many homologs, from 0 to 6, are present in *S. cerevisiae*, *S. castellii*, and *C. glabrata*. Figure 6 clearly shows that genes that are often deleted after the WGD are on average smaller than genes that are seldom deleted. For *A. gossypii*, we found a very similar pattern.

All this confirms that indeed, as found by Nembaware et al. (2002), short genes are more often duplicated, but in contrast, after a WGD, where all genes are duplicated, small genes are also more often lost. Note that the use of a multiple alignment allows us to distinguish between gene deletions and duplications.

Although this may appear contradictory, it is very reasonable that genes that are more often duplicated also need to be more often deleted. Otherwise, the genome would gradually evolve to contain smaller and smaller genes. Therefore, a possible explanation is that because gene duplications are mutationally more likely for shorter genes, deletions of short genes are favored by selection. In contrast, in our neutral model, we observe that longer genes are more often deleted because the base pair where a deletion starts is randomly picked from the whole genome. Therefore, we believe that selection causes short genes to be more often deleted.

In figure 4, we observed that in the data, the strongest overrepresentation in gene length in the data coincides with the maximum in the length distribution (the curve in fig. 4*A*). Because long genes in our model have a larger probability to be deleted, the maximum in the model prediction (solid line in fig. 4*B*) is shifted to longer gene length, which explains the inconsistency between the model prediction and the data.

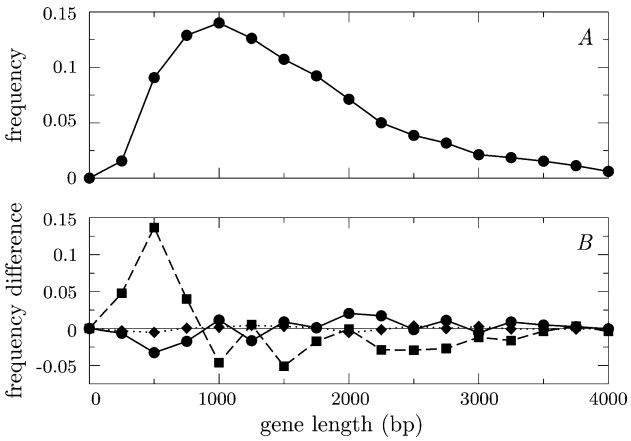

FIG. 5.—Length of genes in *Saccharomyces cerevisiae*. (*A*) Frequency distribution of gene lengths of all genes of *S. cerevisiae* in YGOB. (*B*) Solid line, circles: difference in gene length distribution between genes that are conserved both (ohnologs) and the gene length distribution of all genes (fig. 5*A*). Dotted line, diamonds: idem for genes that are conserved once. Dashed line, squares: recent duplications.

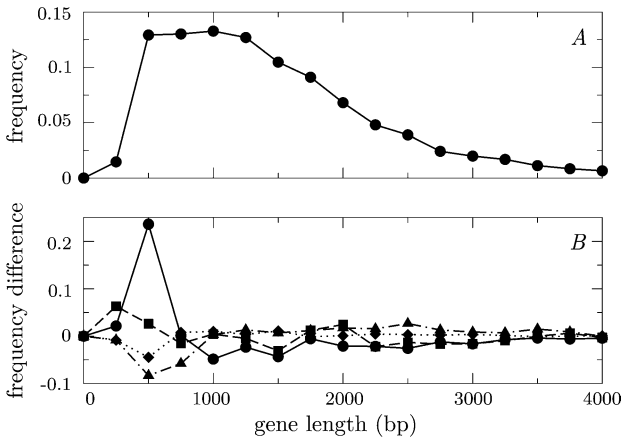

FIG. 6.—Conserved genes in post-WGD species are on average longer in *Kluyveromyces lactis*. (*A*) Frequency distribution of gene lengths of all genes of *K. lactis* in YGOB. (*B*) Solid line, circles: difference in gene length distribution between genes with 0 homologs in the post-WGD species and all genes in YGOB (fig. 6*A*). Dashed line, squares: idem for genes with 1 or 2 homologs in the post-WGD species. Dotted line, diamonds: idem for genes with 3 or 4 homologs in the post-WGD species. Dot-dashed line, triangles: idem for genes with 5 or 6 homologs in the post-WGD species.

Gaps in Pseudogenes of *S. cerevisiae*

We found that in our model, base pair deletions of 100–1000 bp are responsible for the clustering of deleted genes in *S. cerevisiae*. We wondered whether we could find evidence that deletions of this size occur frequently enough during evolution of *S. cerevisiae*. Inspired by Gomez-Valero et al. (2007), we looked at pseudogenes in *S. cerevisiae*. We aligned pseudogenes in *S. cerevisiae* with their homologous ORF's, which are identified in Lafontaine et al. (2004) and counted the gaps in the alignments. In 230 relics, we observed 744 gaps. Also gaps at the beginning or the end of a pseudogene are counted. However, these gaps will in reality have been larger and therefore some gaps are underestimated in size. In figure 7, the distribution of gap sizes is shown.

We observe that the distribution of gap sizes is not geometric. It looks more like a power-law distribution, although it is statistically different. We can see from figure 7 that gaps with sizes as large as 500 bp do occur, although not very often. We wondered whether we could explain the clustering of deleted genes in *S. cerevisiae*, using precisely this distribution of base pair deletions, without any fitting. The result is shown in figure 8.

Indeed, the computational obtained distribution fits the data remarkably well, considered there is no fitting involved. We find a χ^2 value of 9.69, which corresponds to a P value of 0.29. We only overestimate the number of single gene deletions, although considerably less than when only single gene deletions are allowed. It must be noted, however, that the distribution of gap sizes as shown in figure 7 strongly depends on the alignment algorithm used. We also tried ClustalW (Thompson et al. 1994), using the default settings, and we obtained a different distribution, with less large gaps. The fit with the data was also less good than when using DIALIGN 2.2.1 (Morgenstern 2004). However, it remains that base pair deletions of 100–1000 bp are relatively frequent and can therefore be responsible for simultaneous deletion of neighboring genes. We also performed all other simulations in the previous section using the distribution from figure 7. Quantitatively there were differences, but qualitatively the results remained unchanged.

Buchnera aphidicola

We also used this approach to study gene loss in *B. aphidicola*. The computational models we used are the same as the ones used to study the evolution of the yeast genome, except that every gene is only present once, instead of twice (WGD did not occur in *B. aphidicola*).

Again we find that single gene deletions cannot explain the data (see Fig. 9), as was previously noted (Delmotte et al. 2006). However, when deletion sizes (gene based) are geometrically distributed, the fit again becomes very good for an average deletion length of 1.6 gene, which is much larger than for yeast (1.1 gene).

We used gene lengths and intergenic lengths from *E. coli* as initial condition for the base pair based simulations. Both the genes and the intergenic regions of *E. coli* are much shorter than for yeast. This, however, only partly explains the larger gaps found in *B. aphidicola*. The mean deletion length that fits the data best for *B. aphidicola* is 1010 bp. Again, the fit is very well, even better than for the gene-based simulations. Unfortunately, we cannot study the influence of intergenic length and gene length on the deletion probability for *B. aphidicola*, as we did for yeast. For this, we would need an exact alignment between *E. coli* and *B. aphidicola*, which is not available.

Operon Structure

The average number of genes deleted per event is higher for *B. aphidicola* than for *S. cerevisiae*. Correspondingly, on the level of base pairs, larger deletions are required

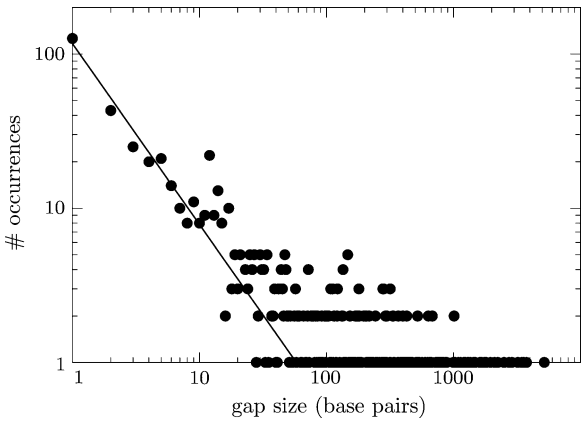

FIG. 7.—Distribution of gap sizes as observed in intergenic relics in *Saccharomyces cerevisiae*. The black lines give a power-law fit, using a maximum likelihood estimation (Goldstein et al. 2004).

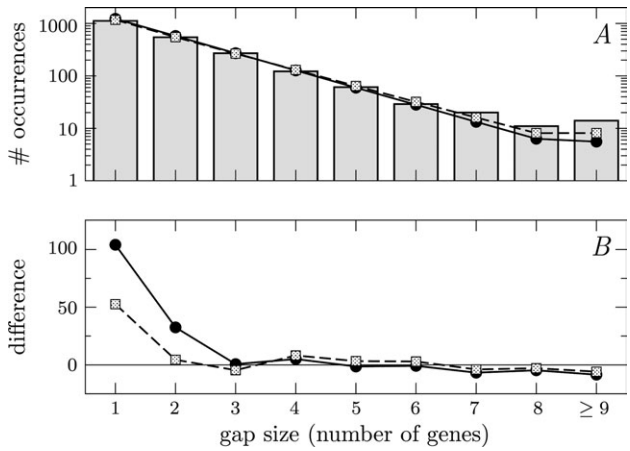

FIG. 8.—The computationally obtained gap size distributions for *Saccharomyces cerevisiae*, when we use the observed gap distribution in relics to delete base pairs. (*A*) Gray bars: the observed distribution of gap sizes in YGOB. Solid line, circles: computationally obtained distribution using only single gene deletions (from fig. 1*A*). Dashed line, squares: computationally obtained distribution when base pair deletions are according to figure 7. (*B*) The difference between the computationally obtained and the observed distribution.

to fit the data (≈1000 vs. 500–600 bp). Gap sizes in pseudogenes have also been measured for bacteria, in *Rickettsia* (Andersson JO and Andersson SG 1999a, 1999b) and *B. aphidicola* (Gomez-Valero et al. 2007). As we found for yeast, many small gaps (<10 bp) and some large gaps (up to 1500 bp) were observed. However, too few gaps were observed for us to use as a distribution in our model. It, however, appears that the gaps in pseudogenes are somewhat smaller than those we found in *S. cerevisiae* (Andersson JO and Andersson SG 1999b). This is surprising because it has been argued that these larger gaps in *B. aphidicola* are because *B. aphidicola* lacks recA, a DNA repair protein (Gomez-Valero et al. 2007). It appears, therefore, that larger base pair deletions are not a plausible explanation for the difference we observe.

However, it is known that in bacteria, the genome is organized in operons. Therefore, the clustering of functional genes is much more pronounced in bacteria than in eukaryotes. We wondered whether the clustering of functional genes in bacteria could explain the larger gaps we observed. It is very well conceivable that operons are deleted more or less as a whole, instead of gene by gene. When a certain gene from an operon is deleted, the pathway in which this gene functions may become nonfunctional and the deletion of the rest of the operon becomes very likely. This is somewhat similar to the "Domino Theory" of gene death (Dagan et al. 2006), which claims that genome reduction in endosymbionts is gradual at first, but when a crucial gene renders a pathway nonfunctional, the entire pathway will be deleted very fast.

To study whether this effect can also explain the observed large gaps, we counted how many operons of a certain length are known in *E. coli*. These data are available from the RegulonDB (version 5.5) database (Salgado et al. 2006). Also "operons" consisting of only one gene are considered. We now assume that operons in our starting genome have the same length distribution (in genes) as in the operon database. We use our gene-based model and furthermore assume that if a certain gene is deleted, there is a certain probability q that the whole operon is deleted. If this is not so, then we assume that this operon will also not be deleted as a whole by following gene deletions in that operon. Furthermore, the deletion will then always be a single gene deletion.

This gives us, for $q = 0.5$, the best fit. This distribution fits the data reasonably well, much better than the single gene assumption, but also less well than the geometric deletion size assumption (both gene based and base pair based).

If we, however, assume that, when a gene is deleted from an operon, only the genes downstream are deleted, the fit becomes as good as the (gene based) geometric distribution (using $q = 0.80$). Here we assume that the orientation of the operons in the genome is random but fixed. This scenario assumes that genes are ordered in the operon and that sometimes only a part of the operon will be nonfunctional if one gene is deleted. So we conclude that the amount of genome organization in operons suffices to explain the excess of large gaps in *B. aphidicola*.

Discussion

We have shown that single gene deletions in a nonstructured genome cannot explain the pattern of gene deletions in yeast, nor in *B. aphidicola*, because larger gaps are observed than can be explained by this random model. There are 2 possible explanations for these larger gaps. The first is that the mutational dynamics are responsible for the large gaps. The second possible explanation is selection: if functional genes are clustered in the genome, this can cause clustering of gene deletions.

Genome organization by operons is well known for bacteria, but for eukaryotes, it has been assumed that gene order is random. In recent years, it has become clear that this assumption is false (Hurst et al. 2004). Operons have been identified in some eukaryotes (mostly in *Caenorhabditis elegans* (Blumenthal 2004)). Furthermore, coexpressed genes in the cell cycle (Cho et al. 1998) or stress-related genes (Burhans et al. 2006) are shown to

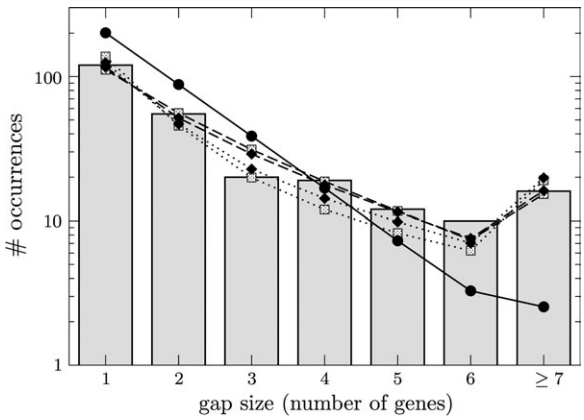

FIG. 9.—Histograms of observed and computationally obtained gap sizes for *Buchnera aphidicola*. Gray bars: observed distribution of gap sizes (Delmotte et al. 2006). Solid line, circles: computationally obtained distribution, if only single gene deletions are allowed ($\chi^2 = 142$, $P < 0.0001$). Dashed line, squares: computationally obtained distribution, deletions are gene based, deletion sizes are geometrically distributed ($q = 0.64$, $\chi^2 = 5.6$, $P = 0.47$). Dashed line, diamonds: computationally obtained distribution, deletions are base pair based and geometrically distributed ($q = 0.0010$, $\chi^2 = 4.0$, $P = 0.67$). Dotted line, squares: computationally obtained distribution, assuming operon structure and a probability of deleting operons as a whole ($q = 0.5$, $\chi^2 = 13$, $P = 0.04$). Dotted line, diamonds: computationally obtained distribution, assuming operon structure and a probability of deleting all genes downstream in that operon ($q = 0.80$, $\chi^2 = 6.0$, $P = 0.43$).

be clustered in the genome of yeast. The amount of gene clustering in eukaryotes is, however, not well known, although it is clearly less than in prokaryotes and it is hard to predict whether it is sufficient to explain for the amount of clustering of deletions that is observed.

For yeast, we found convincing evidence that the mutational dynamics, instead of clustering of single gene deletions, are responsible for the observed large gaps. Firstly, we showed that genes which have a short intergenic region in between are more likely to be both deleted. However, having a short neighboring intergenic region in itself does not increase the probability for a gene to be deleted, which excludes the possibility that this phenomenon is caused by selection. Secondly, small genes are more often found in large gaps. Thirdly, a prediction of our model is that when 2 neighboring pre-WGD genes are deleted, the probability is higher that they are deleted on the same chromosome. We find that this is also the case in the data. For *S. cerevisiae*, *S. castellii*, and *C. glabrata*, we find that, respectively, 53.0%, 53.5%, and 54.3% of the cases genes are deleted in parallel. In our models, both the gene based and base pair based, we find very similar percentages. The base pair–based deletion model gives 53.9%, 55.0%, and 55.2%, respectively, whereas the gene-based model gives 53.7%, 54.8%, and 54.5%. Finally, we confirmed that the size of base pair deletions that we need to fit the data compares well with the size of base pair deletions we observed in pseudogenes in *S. cerevisiae*.

Because short genes have a higher probability to be deleted after WGD, a natural explanation for the occurrence of large gaps would then be that short genes are clustered in the genome. We checked this, but only very limited clustering of short genes can be observed in *K. lactis* and the amount of clustering does not appear to be sufficient to account for the amount of large gaps.

In *A. thaliana*, it has been observed that after a WGD, genes were preferentially lost from one chromosome (Thomas et al. 2006). This phenomenon has been called biased fractionation. Biased fractionation could also result in large gap sizes because in the chromosome from which most genes are deleted, the gaps will become larger than expected. We checked whether biased fractionation occurs in yeast, but in yeast, genes are lost equally from both chromosomes and biased fractionation therefore can also not explain the large gap sizes.

The model we use for yeast is almost entirely neutral. The only selection we implemented is that duplicate genes have only a small probability to be both deleted. Therefore, we conclude that the observed pattern in gene loss in yeast is caused by the mutational dynamics and not by selection. Indeed, it has been shown that the selective constraints on duplicate genes is very modest shortly after the duplication (Lynch and Conery 2000) and selection increases approximately 10-fold later on. Given that most of the duplicates are lost very shortly after WGD (Scannell et al. 2006), it is to be expected that most genes are lost in a relatively neutral way. In later stages of genome reduction, we expect that selection plays a larger role. The fact that we find that in yeast selection does not cause the clustering of deleted genes does not mean, however, that selection is not important during genome shrinkage after a WGD, but more probably that the amount of clustering in functionally related genes in *S. cerevisiae* is very moderate.

In bacteria, the situation is quite different. We found that deleted genes are much more clustered in *B. aphidicola* than in yeast. We propose that, in contrast to yeast, clustering of genes in operons, and hence selection, explains the large gap sizes in *B. aphidicola*. However, also in *B. aphidicola*, we expect that the mutational dynamics partially cause the large gaps because large base pair deletions are also observed in bacteria.

Interestingly, the gap size distribution of base pairs in pseudogenes we found for *S. cerevisiae* (see fig. 7) was very similar to the distribution already found for *Rickettsia* (Andersson JO and Andersson SG 1999a, 1999b). There it was found that 35% of the gaps were of one base pair, whereas 6% of the gaps had a size larger than 500 bp (Andersson JO and Andersson SG 1999a, 1999b). We found in *S. cerevisiae* that 17% of the gaps were single base pair gaps, whereas 10% of the gaps had a size larger than 500 bp. Such a distribution, where most gaps are very short and some are very long, is indicative of a power-law distribution. We assumed a geometric distribution of base pair deletions. Which distribution is used in the model is, however, not crucial for the outcome. What is crucial is that most deletions are only a few base pairs long (and hence cause single gene deletions) and a few are large (causing deletion of multiple genes).

Some studies concerning very recent deletions, using experimental evolution or different natural isolates of a certain species, observed much larger gaps than studies that focused on the comparison of different species. For example in Ochman and Jones (2000), different strains of *E. coli* were compared and many large gaps (>10 kbp) were

observed. A similar study for *Mycobacterium tuberculosis* (Kato-Maeda et al. 2001) also found such large gaps. In Nilsson et al. (2005), such large gaps were observed after experimental evolution of *Salmonella enterica* of mutants that are defective in mismatch repair. Finally, in *S. cerevisiae*, chromosomal deletions are not uncommon (Dunham et al. 2002). All this might indicate that these large deletions do occur but are only selectively advantageous in specific environments, and over a longer evolutionary timescale, these deletions are lost by purifying selection.

Previously, it was observed (Scannell et al. 2006) that, when one ohnolog of a pair is lost, then most often the same gene is lost in different lineages, even if corrected for the fact that many of these cases are because that gene is already lost in the ancestor. This is called convergent gene loss. Functional divergence of genes after WGD is mentioned as a possible explanation for convergent gene loss. We hypothesized that, when deletions would span several genes, another explanation might also play a role. During evolution after WGD, 2 ohnologs will get different neighboring genes. If a certain gene becomes a neighbor of 2 essential genes in one chromosome, but not in the other, this gene cannot be deleted by a gene deletion of length 2 or larger, whereas its ohnolog can. Therefore, its ohnolog will have a higher probability to be deleted. However, because the frequency of single gene deletions is very large in yeast ($q = 0.92$), deletions of multiple neighboring genes cannot explain the observed amount of convergent gene loss.

Our results for *S. cerevisiae* are in contrast with the findings of Byrnes et al. (2006), who study the pattern of gene deletions in the alignment between *S. cerevisiae* and *A. gossypii* (which is found in Dietrich et al. 2004). They compare whether a single gene deletion model or models with larger deletions (a uniform distribution with maximum size 2 or a Poisson distribution with mean 1 or 2) can best describe the data. It was found that the single gene deletion describes the data best ($P = 0.11$).

We find a much lower P value for the single gene deletion model. This is because we found a different gap size distribution, which is based on the alignment of 7 yeast species which each other in the YGOB instead of one. In this way, genes that were deleted in one of the species will be present in others, and this will give a better alignment. This is particularly important because *A. gossyppii* has the least annotated protein-coding genes of all the yeast species we take into account (Byrne and Wolfe 2005).

In addition, instead of using distributions with arbitrarily chosen parameters, we fitted q and therewith the mean of the distribution to the data. Therefore, we were able to find that distributions with a mean size of approximately 1.1 gene fit the data best.

There are more species that have undergone a WGD and from which the pattern of gene deletions is available. For *A. thaliana*, it has been shown that its genome is shaped by 3 WGD's (Initiative 2000; Bowers et al. 2003). The histograms of gap sizes are given in Thomas et al. (2006). *Paramecium tetraurelia* has also undergone several WGD's, and the pattern of gene deletions is available from the supplementary materials in Aury et al. (2006). For these 2 species, there is, however, no alignment available with a pre-WGD ancestor. Only an alignment between both chromosomes is available, but from this, we cannot infer the gap sizes.

Very recently, an alignment between *Tetraodon nigroviridis*, *Danio rerio*, and several post-WGD species was used to study WGD in these teleost fishes (Semon and Wolfe 2007). Unfortunately, this alignment was not given in the article. Moreover, this alignment is not made for the whole genome, because, unlike in yeast, extensive rearrangements have frequently caused loss of synteny. Therefore, we did not use this method for the teleost fishes.

Whether genome reduction is driven by selection or by neutral evolution is a much debated question. Noncoding DNA (like introns, transposons, etc.) is much more abundant in eukaryotes than in prokaryotes. It has been proposed that, by changing the amount of nuclear DNA, the cell size is changed, which has selective power (Cavalier-Smith 1978, 2005). A totally opposite hypothesis states that noncoding DNA is slightly deleterious, because of the potential mutational burden, but only species with high effective population sizes and high per nucleotide mutation rates are capable of keeping their genome devoid of this noncoding DNA (Lynch 2006; Lynch et al. 2006).

In this study, however, we look at loss of genes, instead of loss of noncoding DNA. In *B. aphidicola*, we find evidence that selection is an important factor in determining which genes are lost. Although genome size reduction in *B. aphidicola* might very well be caused by a lack of selection pressure, due to the small effective population size (Mira et al. 2001; Moran and Mira 2001), this does not mean that selection is not important in determining which genes are lost. In *S. cerevisiae*, we find strong evidence that selection does not influence the pattern of gene deletions after its WGD. However, we do not believe that this means that selection is not important during genome shrinkage in yeast.

In summary, we show that gene loss not only occurs in a gradual, gene by gene manner, but that larger base pair deletions can cause simultaneous loss of several neighboring genes. Furthermore, we show that this mechanism is responsible for clustering of deleted genes in *S. cerevisiae*. In *B. aphidicola*, however, we argue that the excess amount of large gaps is due to the clustering of functional genes in operons.

Acknowledgments

We thank A. Crombach for his help with data retrieval and B. Snel for giving an inspiring seminar, which lead us to do this research. This work has been supported by the Faculty of Biology at Utrecht University.

Literature Cited

Amores A, Force A, Yan YL, et al. (13 co-authors). 1998. Zebrafish hox clusters and vertebrate genome evolution. Science. 282: 1711–1714.

Andersson JO, Andersson SG. 1999a. Genome degradation is an ongoing process in Rickettsia. Mol Biol Evol. 16:1178–1191.

Andersson JO, Andersson SG. 1999b. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev. 9:664–671.

Aury JM, Jaillon O, Duret L, et al. (43 co-authors). 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 444:171–178.
Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.
Blumenthal T. 2004. Operons in eukaryotes. Brief Funct Genomic Proteomic. 3:199–211.
Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422:433–438.
Burhans DT, Ramachandran L, Wang J, Liang P, Patterton HG, Breitenbach M, Burhans WC. 2006. Non-random clustering of stress-related genes during evolution of the S. cerevisiae genome. BMC Evol Biol. 6:58.
Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.
Byrne KP, Wolfe KH. 2006. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. Nucleic Acids Res. 34:D452–D455.
Byrnes JK, Morris GP, Li WH. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol Biol Evol. 23:1136–1143.
Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci. 34:247–278.
Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot (Lond). 95:147–175.
Cho RJ, Campbell MJ, Winzeler EA, et al. (11 co-authors). 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 2:65–73.
Dagan T, Blekhman R, Graur D. 2006. The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Mol Biol Evol. 23:310–316.
Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.
Delmotte F, Rispe C, Schaber J, Silva FJ, Moya A. 2006. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC Evol Biol. 6:56.
Dietrich FS, Voegeli S, Brachat S, et al. (14 co-authors). 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science. 304: 304–307.
Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D. 2002. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natl Acad Sci USA. 99:16144–16149.
Goldstein ML, Morris SA, Yen GG. 2004. Problems with fitting to the power-law distribution. Eur Phys J B. 41:255–258.
Gomez-Valero L, Silva FJ, Christophe Simon J, Latorre A. 2007. Genome reduction of the aphid endosymbiont Buchnera aphidicola in a recent evolutionary time scale. Gene. 389: 87–95.
Hermida L, Brachat S, Voegeli S, Philippsen P, Primig M. 2005. The Ashbya Genome Database (AGD)—a tool for the yeast community and genome biologists. Nucleic Acids Res. 33: D348–D352.
Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5:299–310.
Initiative AG. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408:796–815.
Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species Mycobacterium tuberculosis. Genome Res. 11:547–554.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 428:617–624.
Lafontaine I, Fischer G, Talla E, Dujon B. 2004. Gene relics in the genome of the yeast Saccharomyces cerevisiae. Gene. 335:1–17.
Lin YS, Byrnes JK, Hwang JK, Li WH. 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc Natl Acad Sci USA. 103:14412–14416.
Lynch M. 2006. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol. 60:327–349.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.
Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science. 311:1727–1730.
Maniloff J. 1983. Evolution of wall-less prokaryotes. Annu Rev Microbiol. 37:477–499.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.
Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2:RESEARCH0054
Morgenstern B. 2004. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res. 32:W33–W36.
Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res. 12:1370–1376.
Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC, Andersson DI. 2005. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci USA. 102:12112–12116.
Ochman H, Jones IB. 2000. Evolutionary dynamics of full genome content in Escherichia coli. EMBO J. 19:6637–6643.
Salgado H, Gama-Castro S, Peralta-Gil M, et al. (12 co-authors). 2006. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 34:D394–D397.
Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 440:341–345.
Semon M, Wolfe KH. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet. 23:108–112.
Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos Trans R Soc Lond B Biol Sci. 356:1661–1679.
Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Woese CR. 1987. Bacterial evolution. Microbiol Rev. 51:221–271.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 387:708–713.

William Martin, Associate Editor

Accepted August 22, 2007