Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia

Jean-Marc Aury1*, Olivier Jaillon1*, Laurent Duret2, Benjamin Noel1, Claire Jubin1, Betina M. Porcel1, Béatrice Séguernes1, Vincent Daubin1, Véronique Anthouard1, Nathalie Aich1, Olivier Arnaiz2, Alain Billaut1, Janine Beisson3, Isabelle Blanc3, Khaled Bouhouche4, Francisco Cámara5, Sandra Duhamel1, Roderic Guigo5, Delphine Gogendeau5, Michael Katinka1, Anne-Marie Keller3, Roland Kissmehl6, Catherine Klotz1, France Koll1, Anne Le Mouël1, Gersende Lepère1, Sophie Malinsky1, Mariusz Nowacki4, Jacek K. Nowak2, Helmut Plattner6, Julie Poulin1, Françoise Ruiz3, Vincent Serrano4, Marek Zagulski5, Philippe Dessen6, Mireille Bétermier3,4, Jean Weissbach1, Claude Scarpelli1, Vincent Schächter1, Linda Sperling3, Eric Meyer4, Jean Cohen3 & Patrick Wincker1

The duplication of entire genomes has long been recognized as having great potential for evolutionary novelties, but the mechanisms underlying their resolution through gene loss are poorly understood. Here we show that in the unicellular eukaryote Paramecium tetraurelia, a ciliate, most of the nearly 40,000 genes arose through at least three successive whole-genome duplications. Phylogenetic analysis indicates that the most recent duplication coincides with an explosion of species that give rise to the P. aurelia complex of 15 sibling species. We observed that gene loss occurs over a long timescale, not as an initial massive event. Genes from the same metabolic pathway or protein complex have common patterns of gene loss, and highly expressed genes are over-retained after all duplications. The conclusion of this analysis is that many genes are maintained after whole-genome duplication not because of functional innovation but because of gene dosage constraints.

Ciliates are unique among unicellular organisms in that they separate germ line and somatic functions1. Each cell harbours two kinds of nucleus, namely silent diploid micronuclei and highly polyploid macronuclei. The latter are unusual in that they contain an extensively rearranged genome streamlined for expression and division by a non-mitotic process. Only micronuclei undergo meiosis to perpetuate genetic information; the macronuclei are lost at each sexual generation and develop anew from the micronuclear lineage.

In Paramecium the exact number of micronuclear chromosomes (more than 50) and the structures of their centromeres and telomeres remain unknown. During macronuclear development, these chromosomes are amplified to about 800 copies and undergo two types of DNA elimination event. Tens of thousand of short, unique copy elements (internal eliminated sequences) are removed by a precise mechanism that leads to the reconstitution of functional genes2. Transposable elements and other repeated sequences are removed by an imprecise mechanism leading either to chromosome fragmentation and de novo telomere addition or to variable internal deletions3. These rearrangements occur after a few rounds of endoreplication, leading to some heterogeneity in the sequences abutting the imprecisely eliminated regions3. The sizes of the resulting, acentric macronuclear chromosomes range from 50-1,000 kilobases (kb) as measured by pulsed-field gel electrophoresis. Because the sexual process of autogamy results in an entirely homozygous genotype4, the macronuclear DNA that was sequenced was genetically homogeneous.

The Paramecium genome sequence

The Paramecium macronuclear genome sequence was established with the use of a whole-genome shotgun and assembly strategy. Paired-end sequencing of plasmid and bacterial artificial chromosome (BAC) clones provided a coverage of 13 genome equivalents (Supplementary Table S1). We assembled the sequence reads with Arachne in 1,907 contigs connected in 697 scaffolds of size greater than 2 kb, giving a total coverage of 72 Mb. Half of the assembly is contained in 64 scaffolds larger than 413 kb (the N50 size), and 96% of the assembly (69 Mb) in 188 scaffolds larger than 45 kb. A majority of the 188 largest scaffolds are complete macronuclear chromosomes, as determined by analysis of telomere repeats, and on average they contain one gap every 120 kb with a mean estimated size of 0.66 kb (Supplementary Table S2). The largest scaffold, 981 kb long, corresponds to the largest chromosome observed by pulsed-field analysis. Consistent with the estimated gap frequency and sizes, alignment of this scaffold with the finished, independently determined, reference
sequence of that chromosome revealed only four small gaps (less than 250 bases), and less than one difference per 6 kb. If this chromosome is representative, then the present assembly contains more than 99% of the genome sequence with more than 99.95% accuracy.

The *Paramecium* macronuclear genome has a composition of 28% G + C and is devoid of highly repeated sequences (Supplementary Figure S1). The CpG dinucleotide frequency is severely depressed (6.42-fold the expected value) and this cannot be accounted for by codon usage (Supplementary Table S3). In plant and vertebrate genomes, CpG depression is thought to be a consequence of cytosine methylation, but cytosine methylation has never been observed in *Paramecium*.

The genome sequence was annotated by using a combination of *ab initio* predictions, comparative genomics and complementary DNA alignments. A specific resource of about 90,000 expressed sequence tags (ESTs) from six physiological states or developmental stages of *Paramecium* was generated for this purpose. The current protein-coding catalogue is composed of 39,642 genes (Table 1). About 55% of predicted *Paramecium* proteins have no currently referenced InterPro domain, a surprisingly high percentage that probably reflects the great evolutionary distance between *Paramecium* and other eukaryotic genomes sequenced so far. The macronuclear genome is compact (78% coding density), with extremely small intergenic regions (352 bases on average) and introns (25 bp on average).

Remarkably, in contrast with other compact genomes, *Paramecium* genes contain many introns and have an intron/exon ratio closer to that of some invertebrates such as *D. melanogaster* than to that of other sequenced unicellular eukaryotes (Table 1). When looking at the ESTs mapped to each annotated gene, we found no evidence of exon skipping. Adaptation of the splicing machinery to the reduced intron sizes may be incompatible with alternative splicing.

The ciliate clade is generally considered to have evolved from an ancestor with a red algal endosymbiont that gave rise to all extant species of the chromist and alveolate groups (the 'chromalveolate' hypothesis). We failed to find any clear case of genes characteristic of the algal lineage, or of genes of chloroplast origin (see Supplementary Information). This implies that ciliates may have lost all traces of this ancient event because they no longer need photosynthesis or other algal-associated functions. Alternatively, endosymbiosis might not have occurred at the root of the chromalveolate lineage. In this hypothesis, the plastid-derived genes observed in apicomplexan parasites may be the result of a symbiotic event that occurred after their separation from the ciliate lineage.

A surprising result of the annotation process was the prediction of 39,642 protein-coding genes, a number unlikely to be an artefact of the gene prediction procedure (see Supplementary Table S5). *P. tetraurelia* is among the most gene-rich organisms yet sequenced, including many metazoans and plants. Using all-against-all comparisons, we observed that many of the protein-coding genes in *Paramecium* are present in closely related gene families (not shown). This redundancy may explain the high gene number.

### Genome duplications

We aligned each protein against the whole proteome to identify duplicated genes (see Methods). This yields a complete picture of a whole-genome duplication (WGD) linking two by two all of the 188 large scaffolds (example in Figure 1 and global view in the outer circle of Figure 2).

One remarkable characteristic of this WGD is the conservation of synteny: only 8 translocations and 76 internal inversions of chromosomal segments could be deduced. We found 10 instances in which two scaffolds ending in telomeres were paired with a single, longer scaffold, indicating that the germline fragmentation regions (for example transposons) might have moved since the WGD. However, validation of this will require examination of the corresponding micronuclear regions, because DNA elimination events that lead to chromosome fragmentation can also resolve to internal deletions. A second striking feature of this recent duplication is the number of genes retained in duplicate. About 68% of the proteome is composed of two-gene families originating from this event, whereas only 32% corresponds to genes returned to a single state by the loss of one of the two paralogues. This translates into 51% of the preduplication genes still present in two copies, corresponding to 12,026 ancestral genes, a number far greater than for any previously studied WGD. The recent *Paramecium* WGD thus contrasts with the WGDs discovered in yeast, fish or higher plants in recent years, in which synteny, or duplicate gene conservation, is greatly eroded and the WGD events could not always be detected without comparison with a non-duplicated reference genome.

All *Paramecium* chromosomes seem to be conserved after the duplication, and the 11,451 genes lost since the WGD seem to have disappeared through local decay mechanisms acting on a single gene or small region (Supplementary Figure S11). The fact that we could detect a wide range of decay states (Supplementary Table S8 and Supplementary Figure S12), indicates that pseudogenization might not have occurred abruptly after the WGD event, and we cannot exclude the possibility that some of these regions are still present in the micronuclear genome but are no longer amplified during macronuclear development. We estimate the number of recent pseudogenes present in the annotated genes to be about 1,500 (see Supplementary Information).

We inferred the ancient gene order on the basis of the actual chromosomes (see Methods), revealing that two other WGD events had occurred before the recent WGD (Figure 2, second and third circles).

### Table 1 | Comparison of *P. tetraurelia* gene characteristics with selected sequenced eukaryotes

| Cladistic | Chromalveolates | Alveolates | Red algae | Streptophytes | Stramenopiles | Aquatic ciliates | Animals | Unikonts | Opisthokonts |
|-----------|----------------|----------|-----------|-------------|-------------|----------------|---------|---------|-------------|
| Species   | *P. tetraurelia* | *P. tetraurelia* | *C. elegans* | *D. melanogaster* | *D. discoideum* | *D. discoideum* | *C. elegans* | *H. sapiens* | *M. musculus* |
| Species   | *P. tetraurelia* | *P. tetraurelia* | *C. elegans* | *D. discoideum* | *D. discoideum* | *D. discoideum* | *C. elegans* | *H. sapiens* | *M. musculus* |
| Number of genes | 39,642 | 25,687 | 5,233 | 25,498 | 5,317 | 437 | 516 | 572 | 509 |
| Mean coding size (amino acids) | 454 | 561 | 53,177 | 437 | 79 | 81 | 85 | 509 |
| Genes with introns (%) | 80 | 54 | 0.5 | 79 | 69 | 52 | 93 | 81 | 85 |
| Mean intron size (bp) | 179 | 248 | 170 | 146 | 207 | 445 | 1,558 | 568 | 3,365 |
| Mean no. of introns (in spliced genes) | 2.9 | 2.6 | 1.0 | 5.4 | 1.9 | 1.0 | 5.1 | 3.7 | 7.3 | 8.1 |

* Most data are from ref. 45.
† According to ref. 46.
‡ Computed with BDGP4 annotation reported at ensembl34.
§ Computed with SGD1 annotation reported at ensembl34.
‖ Computed with CL100 annotation reported at ensembl34.
*Estimated value from ref 13.
from the outside). In comparison with the recent WGD, the number of genes conserved is smaller, as is the average similarity between paralogues (Table 2 and Fig. 3). We calculate the number of protein-coding genes in the ancestral genome before the three WGD events to be 19,552 by extrapolation of the number of genes present in the paralogons (paralogons are pairs of paralogous blocks that could be recognized as deriving from a common ancestral region). Therefore, taken together, the three WGDs have led to an approximate doubling of the gene count, explaining most of the gene number paradox in _P. tetraurelia_.

Application of the reconstruction and matching method a third time revealed a significant number of duplicated blocks with 436 duplicated genes (Fig. 2, innermost circle), and yet lower similarity between paralogues (Fig. 3a). It is therefore probable that the _Paramecium_ lineage underwent a fourth, more ancient, WGD. The paralogous genes obtained in the final round of reconstruction, near detection limits, are called 'ancient duplicates', with no assumption as to their origin.

Nuclear dimorphism in _Paramecium_ may have helped in tolerating WGD, because the ploidy of the expressed macronucleus is subject to physiological regulation to maintain a constant nucleocytoplasmic ratio; it is therefore possible that a sudden duplication of the micronuclear genome does not alter macronuclear ploidy. However, it is not clear whether this influences the recurrent appearance of WGDs, which have been found in other taxa without nuclear dimorphism.

Retention and evolution of duplicate genes

The nature of the duplicated genes that are retained or not after a WGD can shed light on the still debated forces shaping the genome in the wake of these drastic molecular events. Most studied WGDs are ancient and have provided clues about their long-term effects. In particular, over-retention of signalling molecules and transcription factors has been observed, linking WGD with the evolution of regulatory complexity. Short-term effects have not been easy to track. Very different theories on the forces that drive the retention or loss of gene duplicates include buffering for essential genes, enhancement of metabolic fluxes, protein dosage effects and rapid divergence of gene pairs. Because these models predict different retention outcomes according to gene classes, we analysed the predicted gene categories in _Paramecium_ for their propensity for duplicate retention.

The relative expression levels of proteins involved in the same complex (or metabolic pathway; see below) must be balanced to

Table 2 | Gene content of the paralogons for each duplication

| Duplication event | Genes in paralogons | Ancestral genes in paralogons | Ancestral genes with paralogues | Ancestral genes without paralogues | Percentage of genes in the reconstruction |
|-------------------|---------------------|-----------------------------|-------------------------------|------------------------------------|------------------------------------------|
| Recent WGD        | 35,503              | 35,503                      | 24,052                        | 11,451                             | 90                                       |
| Intermediary WGD  | 31,129              | 20,578                      | 7,996                         | 12,582                             | 79                                       |
| Old WGD           | 18,792              | 9,999                       | 1,530                         | 8,469                              | 47                                       |
| Ancient duplicates| 9,735               | 4,830                       | 436                           | 4,394                              | 25                                       |

* All genes present in a paralgon, whether a duplicated copy is retained or not.
* An ancestral gene is a gene present before a given duplication that could be represented by one or two retained paralogous copies.
* Percentage of genes from the annotated genome present in the ancestral blocks reconstructed from the paralogous relationships.
ensure proper stoichiometry within the complex. Thus, if duplicated genes are involved in a protein complex, the loss of one gene copy is expected to be counterselected unless it is compensated for by the upregulation of the other copy or by gene losses of other partners. To test this hypothesis we defined *Paramecium* orthologues of *Saccharomyces cerevisiae* genes involved in known complexes (Comprehensive Yeast Genome Database (CYGD) on the MIPS website (http://mips.gsf.de/), and ref. 29). Because we retained only clear orthologues between the two species, we expect most of the detected proteins to have ancient conserved functions and therefore a high probability of belonging to conserved complexes. We observed a strong over-retention of genes from the recent WGD involved in known complexes (Fig. 4a). The absence of any common functional theme among the complexes analysed here makes it unlikely that there has been enrichment of specific functions. In contrast, the copy number of each protein within a complex is significantly correlated with that of the other components of the complex (P-value details are given in Supplementary Tables S9 and S10). We concluded that stoichiometry of the proteins constituting most complexes tends to be conserved. It has been previously shown that imbalance in the

Figure 2 | Representation of the successive duplications of the *Paramecium* genome. The exterior circle displays all chromosome-sized scaffolds, and the three interior circles show the reconstructed sequences obtained by fusion of the paired sequences from each previous step. Red lines link pairs of genes with a BRH match, and blue lines link pairs of genes with a non-BRH match that were added on the basis of syntenic position. Only the BRHs that link two genes in the same paralogon are represented. The position of an ancestral block is unrelated to the position of its constituents in the previous circle.
comparisons with inter-species distances. A key function for stoichiometry in this covariation is supported by each protein (Supplementary Table SI). Second, proteins involved as a function of the percentage of cases for paralogues originating from the recent (black), intermediary (red) and old (green) WGDs and from the concentration of complex components can frequently be harmful. A key function for stoichiometry in this covariation is supported by two observations. First, the conservation of gene duplicates in a complex sometimes arises from the retention of genes originating from different duplication events, resulting in identical copy numbers for each protein (Supplementary Table S1). Second, proteins involved in many different complexes are highly retained in the recent WGD but are completely absent as retained duplicates from the old WGD (Supplementary Fig. S20). This indicates that gene loss for complex constituents might be a two-step process, in which proteins with many interactions are preferentially lost as more genes are inactivated, after a first period of preferential conservation to maintain the initial stoichiometry.

A striking result came from an analysis of central metabolism genes, which showed a clear decline in retention over time. All central pathways evolve with the same trend, although at different rates (Fig. 4b). Most pathways are still significantly overamplified with respect to the recent duplication, an effect that has not been observed in WGDs analysed in other organisms. As with protein complexes, the genes involved in a common metabolic pathway show a tendency to covary across the duplications, behaviour that can probably be attributed to stoichiometry, because imbalance in metabolic pathways may have a similar negative effect to that in multiprotein complexes.

The loss of one paralogue would be a slow process, involving over-expression of the other copy to maintain stoichiometric constraints. A strong correlation is observed between expression levels and retention rates of WGD paralogues across all duplications (Fig. 4c). When we consider only the proteins previously defined as associated within complexes, the level of retention is greater (Supplementary Fig. S22), showing that retention is affected both by expression levels and by association in complexes. High expression is therefore a genuine effect in duplicate retention, whether or not the protein belongs to a complex. This correlation has already been observed on a smaller scale in the yeast genome duplication, indicating that it might be a general trait of WGD.

We noticed the existence of a small group of paralogous genes that retained high expression levels, low amino-acid replacement rates and low rates of silent substitution of nucleotides. These genes also have biased codon usage (Supplementary Fig. S21), with many codons apparently optimized for high expression, because their anti-codons correspond to the most abundant tRNA genes. Sequence homology allows us to propose a function for 40 of them (Supplementary Table S1). Most are involved in very basic cellular processes: ribosomal proteins, histones, cytoskeleton components, translation elongation factor 2 and succinate dehydrogenase, and many of them have conserved all eight possible copies. The resulting concentration of complex components can frequently be harmful. A key function for stoichiometry in this covariation is supported by two observations. First, the conservation of gene duplicates in a complex sometimes arises from the retention of genes originating from different duplication events, resulting in identical copy numbers for each protein (Supplementary Table S1). Second, proteins involved in many different complexes are highly retained in the recent WGD but are completely absent as retained duplicates from the old WGD (Supplementary Fig. S20). This indicates that gene loss for complex constituents might be a two-step process, in which proteins with many interactions are preferentially lost as more genes are inactivated, after a first period of preferential conservation to maintain the initial stoichiometry.

A striking result came from an analysis of central metabolism genes, which showed a clear decline in retention over time. All central pathways evolve with the same trend, although at different rates (Fig. 4b). Most pathways are still significantly overamplified with respect to the recent duplication, an effect that has not been observed in WGDs analysed in other organisms. As with protein complexes, the genes involved in a common metabolic pathway show a tendency to covary across the duplications, behaviour that can probably be attributed to stoichiometry, because imbalance in metabolic pathways may have a similar negative effect to that in multiprotein complexes.

The loss of one paralogue would be a slow process, involving over-expression of the other copy to maintain stoichiometric constraints. A strong correlation is observed between expression levels and retention rates of WGD paralogues across all duplications (Fig. 4c). When we consider only the proteins previously defined as associated within complexes, the level of retention is greater (Supplementary Fig. S22), showing that retention is affected both by expression levels and by association in complexes. High expression is therefore a genuine effect in duplicate retention, whether or not the protein belongs to a complex. This correlation has already been observed on a smaller scale in the yeast genome duplication, indicating that it might be a general trait of WGD.

We noticed the existence of a small group of paralogous genes that retained high expression levels, low amino-acid replacement rates and low rates of silent substitution of nucleotides. These genes also have biased codon usage (Supplementary Fig. S21), with many codons apparently optimized for high expression, because their anti-codons correspond to the most abundant tRNA genes. Sequence homology allows us to propose a function for 40 of them (Supplementary Table S1). Most are involved in very basic cellular processes: ribosomal proteins, histones, cytoskeleton components, translation elongation factor 2 and succinate dehydrogenase, and many of them have conserved all eight possible copies. The resulting
indicating that the old WGD occurred before the divergence of Paramecium and Tetrahymena, and the recent WGD before the divergence of P. tetraurelia and P. octoauraria. There are currently not enough data to date the intermediary WGD.

Relative enhancement of these activities should be a significant effect of WGD. Indeed, retention of duplicates for some of these genes, such as those encoding ribosome assembly proteins, is also observed in the WGDs of yeast and Arabidopsis.

**Genome duplication and speciation**

To allow dating of the intermediary and old duplications, we sampled 27 gene families for which homologues were found in Tetrahymena thermophila and several outgroup eukaryotes. Phylogenetic analyses indicate that the old duplication may have occurred shortly before the divergence between Tetrahymena and Paramecium lineages, whereas the intermediary duplication is specific to Paramecium (see Supplementary Information). The recent duplication was dated by analysing 12 gene families for which orthologues in species of the Paramecium genus were available. Phylogenetic trees indicate, with strong bootstrap support, that the recent event occurred after the divergence of P. tetraurelia from P. caudatum and P. multimicronucleatum (Fig. 5). P. tetraurelia belongs to a complex of 15 sibling species (the Paramecium aurelia complex)\(^2\). Data from other species of the P. aurelia complex indicate that the recent WGD occurred before their divergence from P. tetraurelia, which is indicative of a possible link between genome duplication and speciation. Indeed, after a genome duplication, many gene losses occur independently in different populations. This process is therefore expected to lead rapidly to reproductive isolation by Dobzhansky–Muller incompatibility. Interestingly, species of the P. aurelia complex are extremely similar both morphologically and in terms of ecological environment (and hence they were initially thought to correspond to a single species). We therefore propose that the explosion of speciation events that gave rise to the P. aurelia complex is not the result of adaptive evolutionary events (for example the colonization of new ecological niches) but the neutral consequence of the genome duplication\(^2\).

**WGDs in genome evolution**

Although there is no current consensus on gene fates after a WGD, many current models consider that gene loss is rapid after the duplication, and that functional divergence (either by sub-functionalization or neo-functionalization) can explain most of the retention pattern. The Paramecium data point to another model, which is summarized in Fig. 6. Some genes, representing a fraction of the genome, probably quickly lose their duplicated state, because they can never be detected as duplicates in any of the three WGDs. Duplication of these genes may provide no advantage or may even be counter-selected. However, most of the gene duplicates did not go through this rapid elimination. Their maintenance may be driven at least partly by the maintenance of high expression and stoichiometric constraints. Indeed, many of the duplicates of the recent WGD are functionally redundant. Patterns of retention of more ancient WGDs indicate that most of these duplicates will be progressively lost. Yet the distribution of $K_s/K_a$ ratios shows that almost all recent duplicates are under strong purifying selection (Supplementary Fig. S18), indicating that deleterious mutations affecting the coding sequence of one
gene are not masked by the expression of the other copy. As gene regulation evolves to permit loss without affecting dosage, most genes then gradually resolve to a single state. Another model for duplicate retention favours a buffering effect, in which genes are retained in two copies to provide robustness against mutation for important functions\cite{2}, which would be especially important for a homozygous population. It is worth noting that many plants and *Paramecium* can undergo self-fertilization and that natural populations of *P. tetraurelia* are probably homozygous most of the time. This model is consistent with high retention for basic physiological processes in *Paramecium*. However, it is difficult to consider buffering as the main effect of WGD, because it does not explain why so many genes have been retained in many more than two copies.

The appearance of new derived functions may be a secondary product of this basic model. Entire gene modules are retained initially as duplicates with redundant functions, but some of them will evolve to derived ones if there is no need other than stoichiometry for maintaining the redundancy. One example of neo-functionalization is provided by α- and β-tubulins, paralogues of the old WGD, which have distinct functions in the assembly (β-tubulin\cite{1}) or duplication (α-tubulin\cite{1}) of ciliary basal bodies. β-Tubulin is known only in *Paramecium*, whereas α-tubulin is conserved in eukaryotes that perpetuate the centriolar structure.

To gain insight into the acquisition of new functions through WGD, we looked for cases of marked sequence divergence between paralogues, because the neo-functionalization model predicts that the two copies will exhibit asymmetric rates of molecular evolution. After correction for multiple testing, we find that 10.9% of recent duplicates and 16.2% of intermediary duplicates show asymmetric evolutionary rates. This difference is significant ($\chi^2 = 7.55, P<0.01$; see Supplementary Information) and can probably be attributed to the progressive nature of duplicate gene loss: a larger proportion of functionally redundant genes remain among recent duplicates than among the intermediary ones.

Among the 62 intermediary duplicates with asymmetric evolutionary rate, we noticed that the rapidly evolving copy tends to be less retained after the recent WGD than the slowly evolving one (26% versus 60%; $\chi^2 = 2.0, P<10^{-4}$). This indicates that neo-functionalized copies might be more prone to pseudogenization at subsequent WGDs.

The sub-functionalization model predicts that a gene that has been preserved by sub-functionalization at a given WGD is less likely to be retained in two copies at a subsequent WGD\cite{1}. We observed the opposite pattern: 57% of genes already retained in the intermediary duplication are retained after the recent WGD, whereas only 47% of genes not retained after the intermediary duplication are retained after the recent WGD ($P<10^{-4}$). Because this might simply reflect different retention rates between functional classes, only genes from the same functional class were considered. We selected families resulting from the ancient duplication for which one member has been retained in two copies after the intermediary WGD but the other has not (n = 343 families). In agreement with the prediction of the sub-functionalization model, the frequency of retention after the recent WGD was lower for genes retained after the intermediary WGD than for genes that had lost their duplicate (60% versus 67%; $\chi^2 = 4.58, P<.05$). Although significant, this difference is not very strong, which is consistent with the fact that sub-functionalization is an unlikely evolutionary pathway in species with large population sizes\cite{1}.

**Conclusion**

The *Paramecium* duplications have many traits that render them particularly attractive for studying the fates of duplicate genes. Both the high level of gene retention and the strong conservation of syteny helped to define an unusually large collection of paralogues, with high confidence in their exact origin. In particular, the most recent WGD proved informative on the early evolution of paralogues. We predict a pattern of gene loss that is consistent with progressive decay, acting predominantly at the gene level. The initial maintenance of a complete double set of genes is therefore probably an essential difference between WGD and other large-scale duplications.

We can see parallel evolution for whole protein complexes, metabolic pathways, and known interacting protein groups. All these observations are consistent with a central function of dosage constraints on the evolution of gene duplicates after WGD. We propose a new model for WGD evolution, in which expression and stoichiometry are important in the retention of many duplicate genes, that can thereafter evolve in a few cases towards new derived functions. WGD is therefore fundamentally different from single-gene duplication, in which the genetic consequences of sub-functionalization mechanisms are significantly different\cite{2}. The fact that WGDs have been postulated or demonstrated at important converging points in eukaryotic evolution\cite{2,3} makes plausible the direct link between these amplifications and the appearance of some eukaryote-specific novelties. *Paramecium* supports different mutational systems for genetic analysis, including the use of RNA-mediated interference, that can be implemented at the genome scale. We expect that the present study will lead to further work exploring the functional fate of WGDs in detail.

**METHODS**

**Sequencing and assembly.** The whole-genome shotgun and assembly method is described in detail in Supplementary Information.

**Gene annotation.** Protein-coding genes were predicted by combining ab initio models, *Paramecium* cDNA alignments, and alignments of proteins and genomic DNA from other species. The integration of the data was realized using GAZE\cite{2}. Details are given in Supplementary Information.

**Genome duplication.** A global all-against-all comparison of every predicted protein was performed with the Smith–Waterman algorithm (e value <0.01) and a set of 12,504 best reciprocal hits (BRHs) based on alignment scores. Each family was scanned for a best reciprocal hit (w) of 10, and 60% of genes were paired (parameter p). Paralogous blocks were obtained with this windowing strategy and by merging contiguous windows associated with a common target scaffold. In the next step, 1,477 non-BRH syntenic matches found among the ten best matches of any protein with no BRH were added to the parologue list. These paralogons were fused to create a single molecule with the assumption that every gene without a parologue corresponds to a gene whose parologue was lost after the duplication. This creates a succession of genes approximating the gene order of the pre-duplication region, which we called an ancestral block.

To describe the older duplications, the same strategy was iterated three times with the following parameters: intermediary WGD, w = 10 and p = 41%; old WGD and ancient duplicates, w = 20 and p = 50%. The paralogous links for each duplication are presented as Supplementary Tables S17–S20, and the composition of all ancestral blocks is given in Supplementary Table S21. Figure 2 was constructed using Circos (http://m 
available as Supplementary Information. Data on the Tetrahymana proteomes were from The Institute for Genomic Research (http://www.tigr.org/dsb/ezk/; tigr/).

Received 4 May 2006; accepted 31 August 2006.

1. Jahn, C. L. & Klisbaucher, L. A. Genome remodeling in eukaryotic cells. Annu. Rev. Microbiol. 56, 489–530 (2002).

2. Beteta, M. Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res. Microbiol. 155, 399–408 (2004).

3. Le Mouel, A., Butler, A., Cano, F. & Meyer, E. Developmentally regulated chromosome fragmentation linked to impressive elimination of repeated sequences in paramecia. Isuoyot. Cell 2 1076–1090 (2003).

4. Sonneborn, T. M. in Handbook of Genetics (ed. King, R.) 469–594 (Plenum, New York, 1974).

5. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13, 91–96 (2003).

6. Zagulski, M. et al. Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res. Microbiol. 155, 399–408 (2004).

7. Bhattacharya, D., Yoon, H. S. & Hackett, J. D. Photolytic synthesis of eukaryotic uracil: endosymbiosis connects the dots. Bioss. To 50–60 (2004).

8. Vision, T. J., Brown, D. G. & Tansley, S. D. The origins of genomic duplications in Arabidopsis. Science 290, 211–2117 (2000).

9. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploid superimposed on older large-scale duplications in the Arabidopsis genome. Proc. Natl Acad. Sci. USA 101, 178–183 (2004).

10. Wolfe, K. H. A recent polyploid superimposed on older large-scale duplications in the Arabidopsis genome. Proc. Natl Acad. Sci. USA 101, 178–183 (2004).

11. Einsele, C. in Handbook of Genetics (ed. King, R.) 469–594 (Plenum, New York, 1974).

12. Betermier, M. Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res. Microbiol. 155, 399–408 (2004).

13. Wang, J., Yoshida, Y. & Hackett, J. The developmentally programmed elimination of repeated sequences in paramecia. Isuoyot. Cell 2 1076–1090 (2003).

14. Bhattacharya, D. & Hackett, J. D. Photolytic synthesis of eukaryotic uracil: endosymbiosis connects the dots. Bioss. To 50–60 (2004).

15. Vision, T. J., Brown, D. G. & Tansley, S. D. The origins of genomic duplications in Arabidopsis. Science 290, 211–2117 (2000).

16. Wolfe, K. H. A recent polyploid superimposed on older large-scale duplications in the Arabidopsis genome. Proc. Natl Acad. Sci. USA 101, 178–183 (2004).

17. Wolfe, K. H. A recent polyploid superimposed on older large-scale duplications in the Arabidopsis genome. Proc. Natl Acad. Sci. USA 101, 178–183 (2004).

18. Bhattacharya, D., Yoon, H. S. & Hackett, J. D. Photolytic synthesis of eukaryotic uracil: endosymbiosis connects the dots. Bioss. To 50–60 (2004).

19. Vision, T. J., Brown, D. G. & Tansley, S. D. The origins of genomic duplications in Arabidopsis. Science 290, 211–2117 (2000).

20. Wolfe, K. H. A recent polyploid superimposed on older large-scale duplications in the Arabidopsis genome. Proc. Natl Acad. Sci. USA 101, 178–183 (2004).

21. Bhattacharya, D., Yoon, H. S. & Hackett, J. D. Photolytic synthesis of eukaryotic uracil: endosymbiosis connects the dots. Bioss. To 50–60 (2004).

22. Vision, T. J., Brown, D. G. & Tansley, S. D. The origins of genomic duplications in Arabidopsis. Science 290, 211–2117 (2000).

23. Chapman, B. A., Bowers, J. E., Rong, K., Chapman, B. A. & Rong, K. Subfunction partitioning, the telost radiation and the annotation of the human genome. Trends Genet. 20, 481–490 (2004).

24. De Bodt, S., Maere, S. & Van De Peer, Y. Genome duplication and the origin of angiosperms. Trends Ecol. Evol. 20, 591–597 (2005).

25. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene prediction data by dynamic programming. Genome Res. 12, 1418–1427 (2002).

26. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–566 (1997).

27. Claude, R.-C., Chevalet, C., Faraut, T. & Kahn, D. Vertebrate multiple genes by complementary, degenerative mutations. Genetics 151, 1531–1545 (1999).

28. Gehring, W. & Gehring. W. A. The altered evolutionary trajectories of gene duplicates. Trends Ecol. Evol. 20, 670–676 (2005).

29. Koning, P. J. et al. The tree of eukaryotes. Trends Ecol. Evol. 20, 670–676 (2005).