Insights into Three Whole-Genome Duplications Gleaned from the *Paramecium caudatum* Genome Sequence

Casey L. McGrath,* Jean-Francois Gout,* Thomas G. Doak,*,† Akira Yanagi,§ and Michael Lynch*,1
*Department of Biology, Indiana University and †National Center for Genome Analysis Support at Indiana University, Bloomington, Indiana 47405, and §Department of Human Education Faculty of Human Studies Ishinomaki Senshu University Ishinomaki 986-8580, Japan

ABSTRACT *Paramecium* has long been a model eukaryote. The sequence of the *Paramecium tetraurelia* genome reveals a history of three successive whole-genome duplications (WGDs), and the sequences of *P. biaurelia* and *P. sexaurelia* suggest that these WGDs are shared by all members of the *aurelia* species complex. Here, we present the genome sequence of *P. caudatum*, a species closely related to the *P. aurelia* species group. *P. caudatum* shares only the most ancient of the three WGDs with the *aurelia* complex. We found that *P. caudatum* maintains twice as many paralogs from this early event as the *P. aurelia* species, suggesting that post-WGD gene retention is influenced by subsequent WGDs and supporting the importance of selection for dosage in gene retention. The availability of *P. caudatum* as an outgroup allows an expanded analysis of the *aurelia* intermediate and recent WGD events. Both the Guanine+Cytosine (GC) content and the expression level of preduplication genes are significant predictors of duplicate retention. We find widespread asymmetrical evolution among *aurelia* paralogs, which is likely caused by gradual pseudogenization rather than by neofunctionalization. Finally, cases of divergent resolution of intermediate WGD duplicates between *aurelia* species implicate this process acts as an ongoing reinforcement mechanism of reproductive isolation long after a WGD event.

THE genus *Paramecium* has been used as a model unicellular eukaryotic system for over a century, beginning with research by Jennings (1908) and leading to some of the earliest derivations of mathematical population-genetics properties (Jennings 1916, 1917). The subsequent discovery of mating types in members of the *Paramecium aurelia* complex (Sonneborn 1937) permitted the first systematic crossbreeding of different genotypes in any unicellular eukaryote. Later work provided fundamental insights into major issues in biology, including mutagenesis (Igarashi 1966), molecular and developmental genetics (Sonneborn 1947), symbiosis (Beale *et al.* 1969), mitochondrial genetics (Adoutte and Beisson 1972), aging (Siegel 1967), nuclear differentiation (Berger 1937), and gene regulation (Allen and Gibson 1972). More recently, *Paramecium* has been used to study maternal inheritance (Nowacki *et al.* 2005), programmed genome rearrangements and transposon domestication (Arnaiz *et al.* 2012), epigenetic inheritance (Singh *et al.* 2014), and whole-genome duplication (Aury *et al.* 2006; Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results). Although not as well-studied as the *P. aurelia* complex of species, *P. caudatum* also has a long history of research (Calkins 1902; Sonneborn 1933). Recent studies on *P. caudatum* involve investigations into quorum sensing (Fellous *et al.* 2012), thermal adaptation (Krenek *et al.* 2012), learning (Armus *et al.* 2006), endosymbiosis and parasite-mediated selection (Duncan *et al.* 2010, 2011), and ecotoxicology (Rao *et al.* 2007; Kawamoto *et al.* 2010; Hailong *et al.* 2011). *P. caudatum* cells are larger than those of the *aurelia* complex and have a single, larger micronucleus, whereas *aurelia* species have two smaller micronuclei (Fokin 2010). Like the *P. aurelia* species, *P. caudatum* appears to be widespread

doi: 10.1534/genetics.114.163287
Manuscript received February 20, 2014; accepted for publication May 13, 2014; published Early Online May 19, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163287/-/DC1.
This Whole-Genome Shotgun project has been deposited at DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank under the accession no. JMSD00000000. The version described in this article is version JMSD01000000. The assembly and annotation files are also available for download from parameciumDB at http://paramecium.cgm.cnrs-gif.fr/download/pcaudatum/.
[1]Corresponding author: Department of Biology, 1001 E. Third St., Indiana University, Bloomington, IN 47405. E-mail: milynch@indiana.edu

and cosmopolitan, with isolates identified on every continent except Antarctica (Fokin 2010).

With the discovery of at least three successive whole-genome duplications (WGDs) in the history of the *P. tetraurelia* lineage (Aury *et al.* 2006), there is renewed interest in the evolution and genetics of *Paramecium*. WGDs can be found in the ancestry of many model organisms, including zebrafish (Postlethwait *et al.* 2000), yeast (Wolfe and Shields 1997), *Xenopus* (Morin *et al.* 2006), and *Arabidopsis* (Simillion *et al.* 2002). Ancient WGDs are also likely to have occurred in the ancestor of all seed plants and the ancestor of all angiosperms (Doyle *et al.* 2008; Jiao *et al.* 2011), as well as the ancestor of all vertebrates (Panopoulou and Poustka 2005; Hughes and Liberles 2008; Putnam *et al.* 2008; Decatur *et al.* 2013). The evolutionary consequences of WGDs are far-reaching and include the possibilities of neofunctionalization and subfunctionalization of duplicated genes, as well as speciation through the process of divergent resolution of duplicates (Oka 1988; Werth and Windham 1991; Lynch and Conery 2000; Lynch and Force 2000). Retention of duplicate genes tends to be high after WGD compared to single-gene duplications, with 25–75% retention being typical (reviewed in Lynch 2007; Otto 2007). One explanation for this pattern is selection against breaking dosage balance among duplicate genes acting in the same pathway following WGD and opposing imbalances resulting from single-gene duplications (Yang *et al.* 2003; Davis and Petrov 2005; Veitia *et al.* 2008).

The sequencing of two additional *aurelia* species—*P. biaurelia* and *P. sexaurelia*—provided a detailed study of the consequences of the most recent WGD (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results), including the divergent resolution of duplicate genes between *aurelia* species. To further investigate all three WGDs (referred to here as recent, intermediate, and ancient) in the *Paramecium* lineage, we have sequenced the macronuclear genome of *P. caudatum* for use as a potential outgroup. Our results show that *P. caudatum* diverged from the *aurelia* complex before the two most recent *Paramecium* WGDs. We find that, although fewer duplicates from the intermediate WGD survive compared to the recent WGD, the forces governing retention have been largely similar for the two WGD events. In addition, we reveal that, despite the presence of the recent WGD, continuing divergent resolutions of intermediate WGD duplicates have occurred between species of the *aurelia* complex, further enforcing the reproductive isolation between these species.

Materials and Methods

DNA preparation and sequencing

To generate a fully homozygous stock of *P. caudatum*, a fresh isolate was derived from the wild and given the strain name 43. Cytogamy, a process similar to autogamy in *aurelia* species (Wichterman 1939), was then induced with 1.25% methyl cellulose (100 cP, Wako Pure Chemical Industries, Japan) for 5–8 hr (Yanagi and Haga 1995, 1998). Multiple lines were generated and screened for homozygosity, and line 43c3d was selected for sequencing. Roughly 2 liters of cells were then grown in Wheat Grass Powder (Pines International, Lawrence, KS) medium (Aury *et al.* 2006). Cells were subsequently starved, and *Paramecium* cells purified away from bacteria by filtration over a 10-μm Nitex membrane. Macronuclei were isolated away from other cellular debris by gentle lysis of the cell membrane and sucrose density separation (Aury *et al.* 2006). DNA was extracted and purified using a CTAB protocol (Doyle and Doyle 1987). We obtained 8-kb insert 454 FLX (454 Life Sciences, Branford, CT) mate-pair reads (24× coverage) and Illumina (Illumina, Inc., San Diego, CA) paired-end reads (162× coverage).

RNA preparation and sequencing

Roughly 1 liter of *P. caudatum* cells grown in Wheat Grass Powder medium to mid-log phase were purified away from bacteria by filtration over a 10-μm Nitex membrane. Whole-cell RNA was isolated using TRIzol (Ambion, Life Technologies, Carlsbad, CA) and the manufacturer's suggested protocol for tissue culture cells. We then obtained Illumina paired-end reads to a coverage of ~2000 reads/gene. RNAseq reads were mapped to the genome using Bowtie/TopHat (Langmead *et al.* 2009; Kim *et al.* 2013), and transcripts were predicted using Cufflinks (Trapnell *et al.* 2012). The abundance of each messenger RNA was predicted from a logarithm transformation of the FPKM (fragments per kilobase of exon per million fragments mapped) values reported by Cufflinks. Prior to the log transformation, we added 0.1 to the FPKM value of all genes to avoid log transformation of a null value.

Genome assembly and annotation

Illumina and 454 reads were co-assembled using the Celera assembler (Miller *et al.* 2008). Predicted transcripts from RNAseq reads as well as *ab initio* gene predictions from Augustus (Stanke *et al.* 2008), which had been previously trained on *P. tetraurelia* (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results), were used as hints for the final *ab initio* gene prediction by EuGene (Schiex *et al.* 2001; Foissac *et al.* 2008), which had also been previously trained on *P. tetraurelia* (Olivier Arnaiz, unpublished results). PANTHER (Mi *et al.* 2012) was used to predict gene functions for 9250 genes. Identification of genes with no homology to *P. aurelia* or *Tetrahymena thermophila* genes was conducted via BLASTP search (*E*-value cutoff = 0.05) against *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia* proteins (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results) or *Tetrahymena* proteins (Eisen *et al.* 2006).

Alignment of P. caudatum and P. aurelia orthologs

The *P. caudatum* genome was aligned with each of the three previously sequenced *aurelia* genomes using a modified version of the method used to align *P. aurelia* genomes with each other (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results). Specifically, an all-*vs.*-all BLASTP comparison (*E*-value cutoff = 10^{-5}) between *P. caudatum* protein-coding genes and those from each *P. aurelia* genome identified best reciprocal hits (RBHs) between *P. caudatum* and each *P. aurelia* species. At this point, it became apparent that there were two rounds of WGD in the *P. aurelia* species after they diverged from *P. caudatum*, as there were up to four co-orthologs in each *P. aurelia* species for each *P. caudatum* protein. We counted a BLAST match as a RBH if a *P. caudatum* protein was the best hit for an *aurelia* protein and if that *P. aurelia* protein was also among the top four best hits for the *P. caudatum* protein. We used the alignments of the most recent *P. aurelia* WGD (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results) to predict an ancestral, prerecent WGD gene order for each *aurelia* species. We then used the RBH information to delineate orthologous regions between this ancestral (*i.e.*, prerecent WGD) genome and the *P. caudatum* genome using a sliding-window analysis. Windows of 20 genes were examined at a time; if >60% of the *P. aurelia* genes in a window had a RBH to a single *P. caudatum* scaffold, we considered this an orthologous block. Contiguous windows that matched the same *P. caudatum* scaffold were merged. We then added additional non-RBH BLAST matches between *P. caudatum* and *P. aurelia* genes within the orthologous blocks that did not have RBH BLAST hits.

The above procedure produced alignments between *P. caudatum* scaffolds and prerecent WGD *aurelia* scaffolds. Because there was an additional round of WGD in the *aurelia* lineage, this produced two ancestral co-orthologous *aurelia* scaffolds (each one representing two present-day post-WGD scaffolds) aligned with each *P. caudatum* scaffold. We therefore identified *P. caudatum* regions that had two co-orthologous ancestral *aurelia* regions aligned and merged the information to allow alignment across the intermediate *aurelia* WGD. We further identified additional BLAST matches between *aurelia* genes within these regions that did not have a *P. caudatum* ortholog to align intermediate WGD paralogs even in the absence of a *P. caudatum* ortholog.

Alignment of ancient WGD paralogs in P. caudatum

The above procedure was adapted to align *P. caudatum* paralogs from the ancient WGD, with the only difference being that the percentage of genes required to have a RBH on the same scaffold was decreased to 30% due to the reduced retention rate following the ancient WGD.

Factors affecting intermediate and ancient WGD duplicate retention

Gene Ontology (GO) terms for *P. caudatum* genes [assigned by PANTHER (Mi *et al.* 2012)] and the gene-retention data from above were used to analyze functional categories of genes with an over- or underretention of intermediate WGD duplicates. The percentage of ancestral genes still duplicated for each GO term was compared to the percentage of ancestral genes still duplicated for all other GO terms within the same GO category ("Molecular Function," "Biological Process," or "Cellular Component") with a χ^2 test. A correction for multiple testing (Benjamini and Hochberg 1995) was applied to the resulting *P*-values to reduce the false discovery rate (FDR) to 5%.

To determine whether there was a correlation between retention of WGD duplicates and expression level or GC content, *P. caudatum* genes were sorted into bins based on their expression level or GC content, and then the fractional retention for each bin was calculated by the following equation: number of pairs still duplicated/(number of pairs still duplicated + number of singleton genes). For each of the *aurelia* species, we performed separate weighted least-squares regressions between this retention level and the expression level or GC content for duplicates arising at the intermediate and recent WGD events. In the latter case, we calculated retention across both WGDs and all species by adding up the number of extant orthologs for each *P. caudatum* gene (up to 4 in each of 3 species for a possible total of 12). We also used this last retention measure to determine whether both expression level and GC content were independent predictors of retention by performing a multiple regression of this retention level on both expression level and GC content of the *P. caudatum* gene.

Asymmetry of evolutionary rates between aurelia recent WGD paralogs

For all cases in which we could identify two *aurelia* intraspecific paralogs from the recent WGD and their *P. caudatum* ortholog, we produced protein alignments of the three genes using MUSCLE (Edgar 2004). We then performed Tajima's relative-rate test (Tajima 1993) to determine the level of significance of rate asymmetry observed in each alignment and corrected for multiple tests to control the FDR at 5% (Benjamini and Hochberg 1995). We used codeml from the PAML package (Yang 2007) to compute synonymous and nonsynonymous substitution rates on the branches leading to the two paralogs. Expression data from RNA sequencing were used to determine whether the faster-evolving paralog was also the lower-expressed paralog and to determine candidates for neofunctionalization.

Results

The macronuclear genome sequence of P. caudatum

The DNA sequence of the macronuclear genome of *P. caudatum* was determined to a final mean depth of 186× coverage with a combination of Illumina and 454 mate-pair reads (see *Materials and Methods*). The genome assembly that comprises 30.5 Mb (including estimated gaps) is composed of 1202 scaffolds, 274 of which have lengths >2 kb and 96 of

Table 1 Genome statistics for *P. caudatum* as compared to *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia*

	P. caudatum	*P. biaurelia*	*P. tetraurelia*	*P. sexaurelia*
Genome size (Mb)	30.5	77.0	72.1	68.0
Genes	18,509	39,242	39,521	34,939
Gene length (exons + introns) (bp)	1,445.3	1,456.4	1,431.3	1,460.6
Exons/gene	3.5	3.6	3.3	3.6
Average exon length (bp)	399.0	377.9	418.8	379.3
Average intron length (bp)	24.7	31.4	24.2	30.3
Intergenic length (bp)	110.0	335.9	261.3	418.3
Genomic GC content (%)	28.2	25.8	28.0	24.1

Lengths given for genes, exons, introns, and intergenic regions are genome averages. Regions containing gaps were removed from the analysis before calculating averages.

which are >50 kb (Supporting Information, Table S1). This is less than half the genome size of all members of the *P. aurelia* complex that have been sequenced (68–77 Mb) (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results) and is consistent with *P. caudatum* having diverged from the *P. aurelia* complex before the most recent genome duplication events.

The average lengths of exons and introns are similar between *caudatum* and the *aurelias*. However, the average intergenic length in *caudatum* is reduced and is one of the shortest for any eukaryote studied (Table 1). It is likely that the *aurelia* intergenic regions are longer because parts of these regions are composed of the remains of pseudogenized duplicates from the WGDs (Aury *et al.* 2006). The shorter average intergenic length of *caudatum*, therefore, likely represents the ancestral, preduplication intergenic length.

The GC content of the *P. caudatum* genome is 28.2%, which is more similar to the GC content of *P. tetraurelia* (28.0%) than that of *P. biaurelia* (25.8%) or *P. sexaurelia* (24.1%). Similar differences in GC content are found in both coding and noncoding regions. Given the relationships between species (Figure 1), this complex pattern implies multiple changes of the GC content within the different lineages of *Paramecium*. Such changes could reflect a difference in the mutation spectrum or in processes such as biased gene conversion. In the future, sequencing the genomes of more *P. aurelia* species as well as additional *Paramecium* species outside of the *aurelia* complex should help reveal the complex evolution of genomic GC content in *Paramecium*.

The *P. caudatum* genome contains 18,509 annotated protein-coding genes, about half the number of genes found in each *P. aurelia* species. PANTHER (Mi *et al.* 2012) functional annotations are available for 9250 (50.0%) of these genes. The genome sequence and annotations are available at ParameciumDB (http://paramecium.cgm.cnrs-gif.fr) (Arnaiz and Sperling 2011). Among the 18,509 *P. caudatum* predicted proteins, 1053 have no identifiable *P. aurelia* homolog, and 8456 have no identifiable homolog in another oligohymenophorean ciliate, *T. thermophila* (Eisen *et al.* 2006).

Retention and divergent resolution of aurelia intermediate WGD paralogs

Although previous data indicated that *P. caudatum* likely did not share the most recent WGD with the *aurelia* species (Aury *et al.* 2006), it was initially unclear whether *P. caudatum* shared the intermediate WGD. Upon aligning the *P. caudatum* genome with each of the *P. aurelia* genomes, it is apparent that there are up to four co-orthologs in each *aurelia* species for each *P. caudatum* protein (Figure 1), indicating that the intermediate WGD was also limited to the *P. aurelia* lineage. Although large-scale synteny has broken down to an extent between *P. caudatum* and the *P. aurelia* species, a substantial degree of local synteny remains (Figure 2). We were able to align 5781 *P. caudatum* genes with all 12 of their syntenic *aurelia* orthologs (when present), which includes 10,907 *P. biaurelia*, 10,970 *P. tetraurelia*, and 10,024 *P. sexaurelia* genes (File S1 and File S2). This represents ~30% of the genes in each of the aligned genomes, as missing data in any of one of the 12 scaffolds being aligned caused a region to be excluded. We also identified 305 ancestral genes that existed in the *P. aurelia* ancestor before the intermediate WGD but that do not appear to have syntenic *P. caudatum* orthologs. These genes have either been lost in *P. caudatum*, were acquired by the *aurelia* ancestor before the intermediate WGD, or have evolved too rapidly to identify homology between *P. caudatum* and *P. aurelia*.

From these alignments, we estimate that 24% of duplicates that originated from the intermediate WGD were still present in two copies in the *P. aurelia* ancestor at the time of the recent WGD. Note that, when no descendant copies of an intermediate WGD duplicate remain in any *P. aurelia* species, we infer that this duplicate was lost before the recent WGD. This estimate of intermediate WGD retention is identical to that inferred from the *P. tetraurelia* genome alone (Aury *et al.* 2006) and is much lower than the 42–52% of recent WGD duplicates retained in each *aurelia* lineage (Lynn McGrath, Jean-Francois Gout, Parul Johri, Thomas Graeme Doak, and Michael Lynch, unpublished results). Using parsimony, we assigned gene losses to the *P. aurelia* phylogenetic tree for a complete picture of duplicate gene loss following the intermediate and recent WGDs (Figure 3).

The median synonymous site divergence (*dS*) between intermediate WGD duplicates (*e.g.*, between genes along scaffold *tetraurelia* A and scaffold *tetraurelia* C in Figure 1) is much higher (median *dS* saturated at ~8) than between recent WGD duplicates (*P. tetraurelia* A and *P. tetraurelia* B in Figure 1; median $dS = 1.7$; McGrath *et al.*,

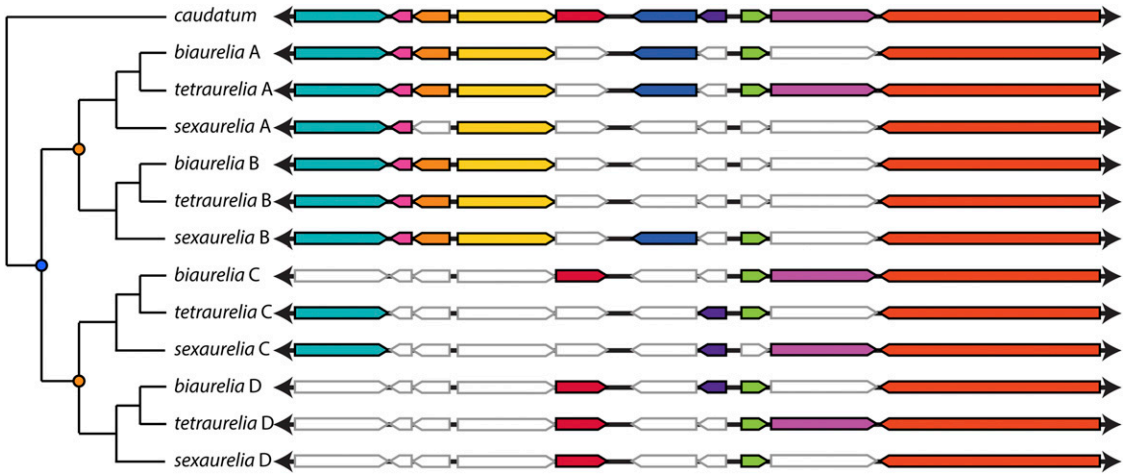

Figure 1 Example alignment of a region from a *caudatum* scaffold and 12 co-orthologous scaffolds from *biaurelia*, *tetraurelia*, and *sexaurelia*. Scaffolds designated with the same letter (A, B, C, or D) are orthologous to each other. Homologous genes are displayed in matching colors, and genes that have been lost on *aurelia* scaffolds are in white with gray outlines. Intergenic regions are not shown to scale to demonstrate the locations of gene losses. A phylogeny is given to demonstrate relationships between scaffolds with orange dots denoting the location of the recent WGD and a blue dot denoting the intermediate WGD. Regions illustrated include the following: *caudatum*—scaffold_0002:1786-13531; *biaurelia*—scaffold_0010:297344-307146, scaffold_0012:192506-199800, scaffold_0312:38737-46840, and scaffold_0297:20187-26050; *tetraurelia*—scaffold_15:289275-399069, scaffold_25:343442-351219, scaffold_30:236161-243107, and scaffold_31:210331-217756; *sexaurelia*—scaffold_015:487286-494871, scaffold_017:247245-256846, scaffold_026:298395-306376, and scaffold_027:313613-319815.

unpublished results), indicating that the intermediate WGD occurred long before the recent WGD. It was previously observed that intraspecific paralogs, on average, have lower *dS* values than interspecific paralogs for the recent WGD, which was interpreted as evidence for frequent gene conversion between paralogs (McGrath *et al.*, unpublished results). For duplicated genes derived from the intermediate WGD, intraspecific paralogs do not appear to be more similar to each other than interspecific intermediate WGD duplicates (Figure S1), suggesting that gene conversion between intermediate WGD duplicates does not occur frequently. However, because the *dS* between paralogs from the intermediate WGD is highly saturated, it is possible that gene conversion occurred immediately following the intermediate WGD and then gradually stopped, as observed for the most recent WGD (McGrath *et al.*, unpublished results). The action of gene conversion on WGD duplicates therefore appears to be limited to a finite period of time following WGD.

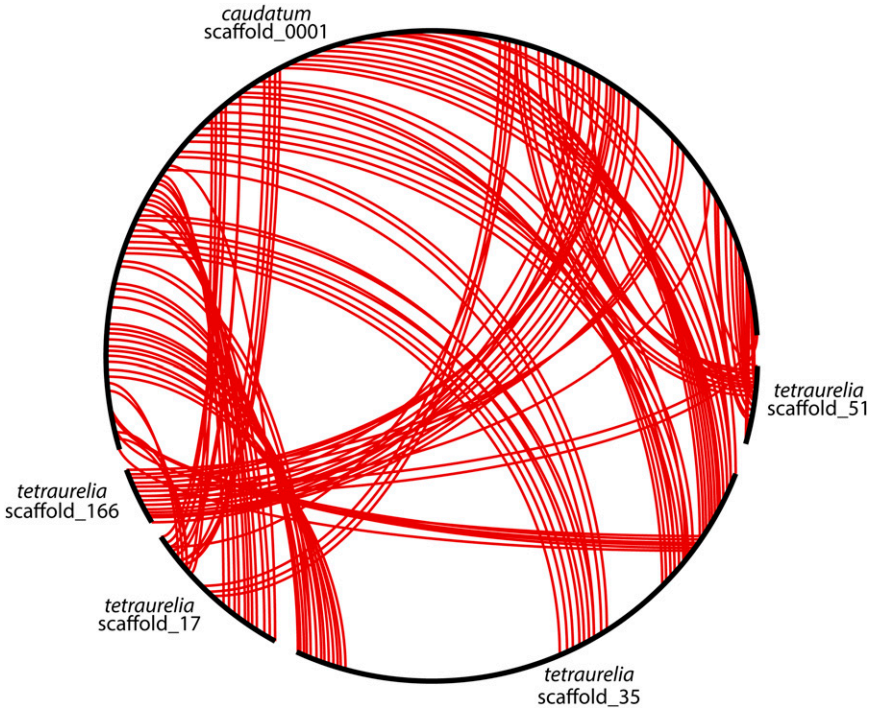

Figure 2 Overview of synteny between a *caudatum* scaffold (scaffold_0001) and the four *tetraurelia* scaffolds that together make up one set of orthologs (*e.g.*, "*tetraurelia* A" in Figure 1). Orthologs between *caudatum* and *tetraurelia*, as identified by our orthologous alignment process (see *Materials and Methods*), are connected by red lines. Scaffolds are represented by black lines around the outside of the circle.

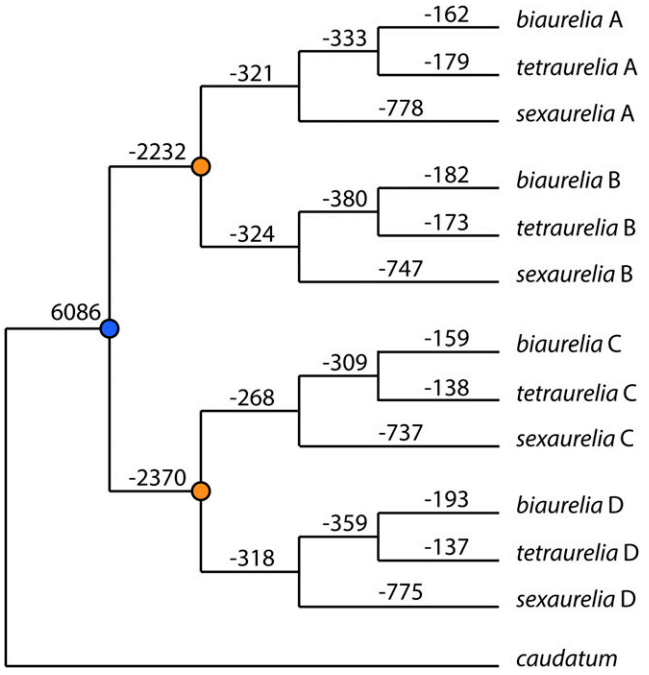

Figure 3 Gene losses following the intermediate and recent WGDs along the phylogenetic tree. The number of ancestral genes included in the analysis (6086) is given, followed by the number of losses occurring along each branch. Blue dot: intermediate WGD; orange dots: recent WGD. Note that branch lengths are not to scale.

Estimating the age of the intermediate WGD

Because synonymous substitutions are usually neutral, their accumulation should directly reflect the time of divergence between two homologous sequences. We previously relied on this principle to estimate the age of the recent WGD, based on the median *dS* between paralogs (1.7) and the mutation rate measured in *P. tetraurelia* [2.64×10^{-11} mutations per site per cell division (Sung *et al.* 2012)], and obtained an estimation of ~230 million years (McGrath *et al.*, unpublished results). Using the same strategy, and given that the median *dS* between paralogs from the intermediate WGD is ~8 (indicating that each site in the *P. caudatum* genome has been hit by an average of approximately four mutations per site since the intermediate WGD), we derive an estimate of ~150 billion cell divisions ($4/2.64 \times 10^{-11}$) since the intermediate WGD, which translates to ~1.5 billion years, assuming 100 cell divisions per year (Pianka 2000). Because the estimation of such high *dS* values lacks precision and because both the mutation rate and the number of cell divisions per year could have varied substantially within this long period of time, these estimates need to be taken cautiously. However, it is clear that the intermediate WGD is much older than the recent one and that both events are at least hundreds of millions of years old.

Divergent resolution of intermediate WGD-derived paralogs is still ongoing

An important consequence of WGDs is their potential to cause reproductive isolation and speciation through divergent resolution of duplicate genes (Oka 1988; Werth and Windham 1991; Lynch and Conery 2000; Lynch and Force 2000). Although the intermediate WGD occurred long ago and there has been another round of WGD since, it is possible that divergent resolution of intermediate WGD duplicates is still ongoing between *aurelia* species (Figure 4). We found two instances of divergent resolution of intermediate WGD duplicates between *P. tetraurelia* and *P. sexaurelia*, four between *P. biaurelia* and *P. sexaurelia*, and two between *P. biaurelia* and *P. tetraurelia* (Table S2). Among these, there are two cases in which each species retained two recent copies of alternative intermediate duplicates (Figure 4A), three cases in which each species retained one copy (Figure 4C), and three cases in which one species retained two copies and the other species retained only one (Figure 4B). Although these represent only a handful of cases, it is surprising to observe that divergent resolution still occurs for such an old WGD event. Therefore, WGDs may still promote speciation even hundreds of millions of years after they arose.

GC content, expression level, and functional category influence duplicate-gene retention

Previously, we found that expression level, GC content, and functional category influenced the retention of duplicate genes from the recent *aurelia* WGD (Gout *et al.* 2010; McGrath *et al.*, unpublished results). We investigated whether the same factors are correlated with retention of intermediate WGD duplicates. In the case of the recent WGD, we previously found that 10–17 functional categories were significantly overretained after the recent WGD in the three different *P. aurelia* species investigated, whereas five to seven functional categories were significantly underretained (McGrath *et al.*, unpublished results). For the intermediate WGD, only three functional categories are overrepresented among duplicates (ribosomal proteins, proteins involved in amino acid phosphorylation, and proteins involved in response to stress), and no categories are significantly underrepresented (File S3). These three overretained categories were also found to be overretained after the recent WGD (McGrath *et al.*, unpublished results). The relative scarcity of over- and underrepresented functional categories from the intermediate WGD could be due to the fact that biased retention of functional classes decreases over time, although the fact that fewer intermediate WGD duplicates survive reduces the power of the analysis.

We previously found that the GC content and expression level of *P. aurelia* genes are positively correlated with duplicate retention from the most recent WGD (McGrath *et al.*, unpublished results). However, both these factors could be influenced by duplication itself. Frequent gene conversion between paralogs could increase their GC content via GC-biased gene conversion (Galtier *et al.* 2001; Duret *et al.* 2008; Pessia *et al.* 2012), making it unclear whether high-GC content drives gene retention or is a secondary consequence of gene retention. With the *P. caudatum* genome

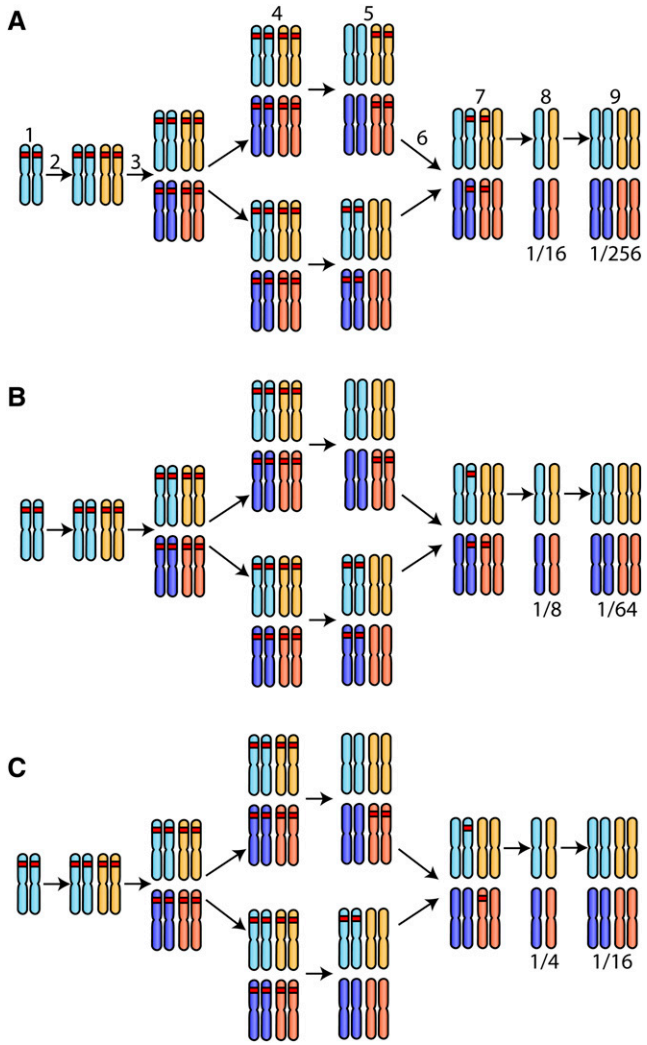

Figure 4 Divergent resolution of intermediate WGD duplicates. (A) In the preduplication ancestor (1), one gene (red) is shown on an ancestral, diploid chromosome. Following the intermediate WGD (2), the gene is present in two copies. Following the recent WGD (3), the gene is present in four copies. Two subpopulations become isolated from each other (4) and divergent resolution of the intermediate WGD duplicate occurs, with one population losing the gene from both the light and dark blue chromosomes and one population losing the gene from both the light- and dark-orange chromosomes (5). Hybridization then occurs between the two subpopulations (6). Hybridization leads to the creation of F_1 hybrids (7). One-sixteenth of the gametes produced by the F_1 hybrids have no functional copy of the gene (8). Of the F_2 offspring, 1/256 have no functional copy of the gene (9). The amount of reproductive isolation increases when one (B) or both (C) subpopulations retain only one copy, instead of two copies, of the gene at stage 5.

providing a proxy for the ancestral, preduplication genome, it is now possible to eliminate this ambiguity.

We separated *P. caudatum* genes into bins based on their GC content and then within each *P. aurelia* species determined the retention level for orthologous duplicates from the recent WGD using the following equation: retention rate = number of pairs of retained paralogs/(number of pairs of retained paralogs + number of singleton genes). A weighted least-squares regression of the retention levels on the *P. caudatum* GC content revealed a significant positive correlation for all three *aurelia* species (Figure 5A) (*P. biaurelia* $r^2 = 0.95$, *P. tetraurelia* $r^2 = 0.92$, *P. sexaurelia* $r^2 = 0.96$; $P < 0.0001$ in each case).

We repeated this procedure for the intermediate WGD, this time calculating a retention level for each bin based on the retention of intermediate WGD duplicates using the same formula as above (Figure 5B). Again, we found a significant positive correlation between GC content of *P. caudatum* genes and the retention of their orthologs following the intermediate WGD, although the correlation is weaker than for recent WGD duplicates ($r^2 = 0.32$, $P < 0.0001$). Finally, we calculated a retention level for the *P. caudatum* GC bins across both the recent and intermediate WGDs (see *Materials and Methods*). We obtained a similar correlation between GC content and retention of duplicates across both WGDs ($r^2 = 0.88$, $P < 0.0001$) (Figure 5C).

We applied the same analysis to expression levels (Figure 5, D–F). Like GC content, the retention rate of *P. aurelia* genes is positively correlated with the expression level of their preduplication orthologs in *P. caudatum* after both the intermediate ($r^2 = 0.21$, $P < 0.0001$) and the recent WGD (*P. biaurelia* $r^2 = 0.88$, *P. tetraurelia* $r^2 = 0.85$, *P. sexaurelia* $r^2 = 0.88$; $P < 0.0001$ for all three species) and after both WGDs combined ($r^2 = 0.86$, $P < 0.0001$). Again, the correlation with the intermediate WGD is weaker than with the recent WGD, suggesting that the selective forces preventing gene loss weaken over time.

Because GC content and expression level are correlated with each other (Figure S2), we conducted a multiple regression analysis to determine whether both were independent predictors of duplicate retention. We calculated the number of extant *aurelia* orthologs across all species for each *P. caudatum* gene. Because there are up to four co-orthologs of each *P. caudatum* gene in each of three *aurelia* species, this retention value ranges from 1 to 12. We then determined whether this retention value was correlated with GC content and/or expression level of the *P. caudatum* genes. Although the overall correlation is low for this analysis ($r^2 = 0.05$), the number of retained orthologs is significantly correlated with both GC content and expression level (GC-content partial regression coefficient = 5.43, $P < 0.0001$; expression-level partial regression coefficient = 0.15, $P < 0.0001$), indicating that the two variables, although correlated with each other, do have independent predictive power.

As suggested previously (Gout *et al.* 2010), it is likely that expression level promotes duplicate-gene retention through selection for increased dosage of the gene product. Because we observe a correlation between the preduplication GC content and the retention rate, we can rule out the hypothesis that the higher GC content of retained genes is simply a consequence of frequent GC-biased gene conversion between paralogs. Given the fact that the mutational spectrum of *Paramecium* is strongly biased toward GC→AT mutations (Sung *et al.* 2012), it is likely that a high-GC content in the

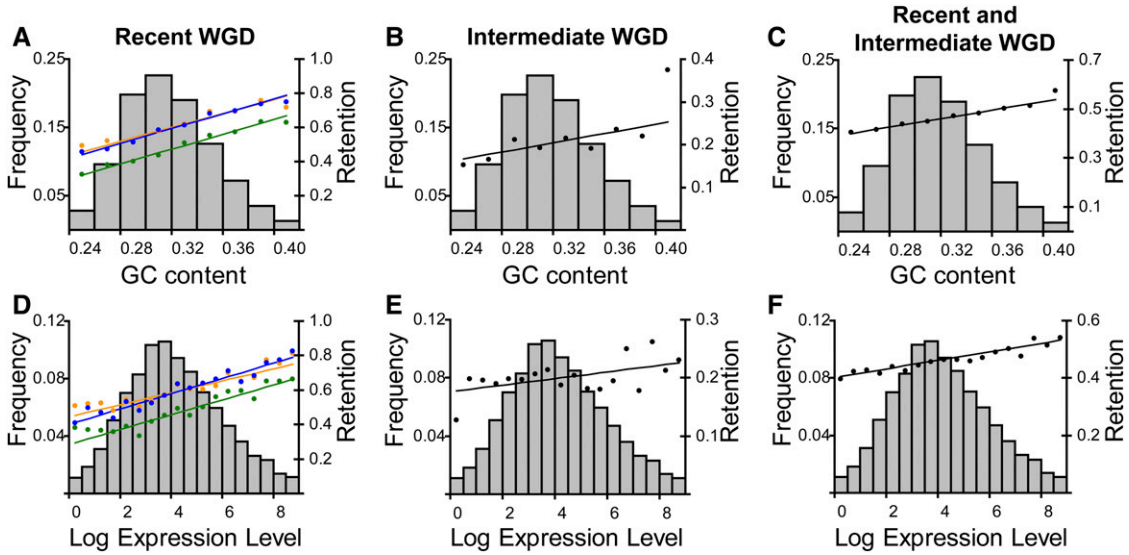

Figure 5 Relationships between duplicate retention and *P. caudatum* GC content or expression level. *P. caudatum* genes are sorted into bins based on GC content (A–C) or log-expression level (D–F). The frequency distribution of genes that fall into each GC content or expression level bin is shown (gray bars), along with the retention level of genes within each bin (dots) and the corresponding regression lines for retention on GC content or expression level (lines) (see *Materials and Methods* for more details). A and D show the retention of recent WGD duplicates in *biaurelia* (blue), *tetraurelia* (orange), and *sexaurelia* (green). B and E show the retention of intermediate WGD duplicates, and C and F show the retention of duplicates across both WGDs.

coding region of ancestral genes reflects strong selective pressures at the sequence level, leading to the suggestion that genes under strong selective pressure at the sequence level are also under strong selective pressure against post-WGD changes in dosage (*i.e.*, against gene loss). In agreement with this hypothesis, we observed a strong negative correlation between the speed of evolution of ancestral genes [measured as nonsynonymous site divergence (*dN*) between *P. caudatum* and *P. multimicronucleatum* orthologs] and post-WGD retention rate (Figure S3).

Asymmetry of evolutionary rate between recent aurelia paralogs

Although gene duplicates are 100% identical immediately after a WGD event, early mutational events may set them on differential courses of evolution, resulting in asymmetrical rates of evolution between the two copies. Using Tajima's (1993) relative rate test, we found that 14% of *P. biaurelia*, 13% of *P. tetraurelia*, and 20% of *P. sexaurelia* paralogs exhibit significant rate asymmetry after correction for multiple testing (see *Materials and Methods*). This represents a significantly greater percentage of asymmetrical trees for *P. sexaurelia* than for *P. biaurelia* and *P. tetraurelia* ($P < 10^{-15}$, χ^2 test). Such asymmetries in the rate of evolution are often interpreted as evidence for neofunctionalization (Aury *et al.* 2006; Johnson and Thomas 2007; Han *et al.* 2009). However, both positive selection (expected in the case of neofunctionalization) and relaxed purifying selection (expected in the case of pseudogenization) can produce a signature of asymmetric evolution between the paralogs. Therefore, it is possible that a substantial fraction of duplicates showing asymmetrical evolutionary rates represent cases of ongoing pseudogenization, rather than neofunctionalization.

Because *P. sexaurelia* has a higher duplicate-gene loss rate than both *P. biaurelia* and *P. tetraurelia*, it is likely that the relative abundance of asymmetrical trees in this species is due to greater levels of pseudogenization occurring in *P. sexaurelia*, rather than to an increased level of neofunctionalization. To test this hypothesis, we compared the difference in expression level to the difference in the rate of evolution of the protein sequence between paralogs. While there is no reason to believe that genes undergoing neofunctionalization should preferentially become less expressed, genes undergoing pseudogenization are likely to experience decreased expression level due to random mutation and relaxed evolutionary constraints. We observed a significant negative correlation between the difference in expression level and the difference in *dN* between paralogs from the recent WGD in all three *P. aurelia* species (*P. biaurelia*: $r = -0.32$, *P. tetraurelia*: $r = -0.26$, *P. sexaurelia*: $r = -0.30$; $P < 0.001$ for all three species).

Further evidence that a pseudogenization process is responsible mainly for the sequence asymmetry comes from the fact that the faster-evolving paralog is also the lower-expressed paralog 77% of the time among significantly asymmetrically evolving pairs in *P. biaurelia* and *P. tetraurelia* and 81% of the time among *P. sexaurelia* asymmetrical pairs. Among symmetrically evolving pairs, on the other hand, this is true only slightly more than half the time (58% in *P. biaurelia*, 56% in *P. tetraurelia*, and 60% in *P. sexaurelia*), representing a significant difference between asymmetrical and symmetrical pairs for all three species (all *P*-values: $<10^{-16}$, χ^2 test). The proportion for symmetrical

pairs, while lower than for asymmetrical pairs, is nonetheless significantly different from 50% (all *P*-values: $<10^{-16}$, χ^2 test), indicating that this process of increased sequence evolution and reduced expression level for one paralog of a pair has begun even for some of the pairs that do not yet demonstrate significant sequence evolution asymmetry. Finally, the percentage of both asymmetrically evolving and symmetrically evolving pairs where the faster-evolving paralog is also the lower-expressed paralog is greater in *P. sexaurelia* than in *P. biaurelia* or *P. tetraurelia* ($P < 0.02$ for all three species, χ^2 test), consistent with the pseudogenization process happening more rapidly in *P. sexaurelia*, where the gene-loss rate is higher.

These observations do not preclude the possibility of some neofunctionalized paralogs among the set of asymmetrical pairs. To search for candidate genes possibly undergoing neofunctionalization, we identified paralog pairs in each species with significant asymmetry of coding sequence evolution, where the expression level difference between the two paralogs was above the 80th percentile, and where the more highly expressed paralog was the faster-evolving paralog. This analysis yielded 35 *P. biaurelia* genes, 18 *P. tetraurelia* genes, and 29 *P. sexaurelia* genes that are candidates for neofunctionalization (File S4). Interestingly, we found one gene (Pterin-4–α-carbinolamine) that is a candidate for neofunctionalization in both *P. tetraurelia* and *P. sexaurelia*. Because these species diverged soon after the recent WGD, it is likely that if neofunctionalization happened, it happened independently in the two species. Indeed, the faster-evolving copy in *P. tetraurelia* is the recent WGD paralog, not the ortholog, of the faster-evolving copy in *P. sexaurelia*. Similarly, a sodium-dependent phosphate transporter is a candidate in both *P. sexaurelia* and *P. biaurelia*, although in this case the faster-evolving copies are orthologs of each other, making it possible that only one event of neofunctionalization happened before the *P. sexaurelia–P. biaurelia* speciation. Because this analysis is based solely on sequence data, additional biochemical experiments will be necessary to test the function of these genes in the different *aurelia* species and confirm (or refute) the possibility of neofunctionalization.

Analysis of the ancient WGD in P. caudatum vs. P. aurelia

Although *P. caudatum* does not share the intermediate or recent WGDs with the *aurelia* complex, our genome alignments indicate that all species share the ancient WGD. We used a method similar to that of Aury *et al.* (2006) to identify ancient paralogous blocks within *P. caudatum* for comparison to the ancient paralogous blocks identified in *P. tetraurelia*. We hypothesized that, if gene dosage constraints play an important role in post-WGD retention, the subsequent rounds of WGD within the *P. aurelia* lineage may relax selective pressure for retention of paralogs derived from the more ancient WGD. Under this scenario, the retention rate for the ancient WGD is expected to be higher in *P. caudatum* compared to the *P. aurelia* species. Alternatively, if changes of function are responsible for most duplicate-gene retention, subsequent duplications should not affect the probability of retention, so that we would not expect to observe a difference in the post-ancient WGD retention rate between *P. caudatum* and the *P. aurelia* species. We find the post-ancient WGD retention rate in *P. caudatum* to be 16% (File S5), which is about twice the retention rate observed in *P. tetraurelia* (Aury *et al.* 2006), supporting the idea that gene dosage constraints play an important role in post-WGD retention for a majority of genes.

Like the intermediate and recent WGDs, *P. caudatum* genes retained in duplicate from the ancient WGD have significantly higher GC content ($P < 10^{-6}$, *t*-test) and expression level ($P < 10^{-8}$, *t*-test) than those whose paralogs have been lost (Table S3). Again, the most overretained functional category is ribosomal proteins, although due to the small number of genes included, the power of the GO analysis is diminished and no category remains significant after correcting for multiple tests (File S6).

Discussion

Previously, we examined the evolutionary consequences of the recent WGD predating the emergence of the *P. aurelia* species complex (McGrath *et al.*, unpublished results). The sequencing of an outgroup species, *P. caudatum*, has allowed us to further investigate the recent and intermediate WGDs, both of which occurred after the divergence of *P. caudatum* and the *aurelia* lineage, as well as the ancient WGD that is shared between them. Although the percentage of retained duplicates differs across all three WGDs, we found similarities in the factors influencing retention, namely, GC content, expression level, and to a lesser extent, functional category. By correlating duplicate-gene retention in the *aurelia* species with gene features in *P. caudatum*, we demonstrate that the properties of the preduplication genes influence their post-WGD fate.

Interestingly, the correlation between all three of these gene features and retention is lower for the intermediate WGD than for the recent WGD, indicating that over time, as more duplicates are lost, biases in which genes are retained become weaker. This suggests that many paralogs are eventually able to escape the selective pressures that caused them to be initially maintained. For example, if a gene is preserved initially by selection for increased dosage or due to dosage balance constraints, subsequent mutations altering the gene's expression could ultimately allow the gene to be lost. Moreover, these changes in preservational biases over time could complicate comparisons of WGDs across phyla, as such comparisons are generally not corrected for the amount of evolutionary time since the WGD.

It is still unclear what fraction of genes are retained for dosage constraints and what fraction are retained because of changes in function. However, our observation that *P. caudatum* has retained twice as many paralogs from the ancient WGD than the *aurelia* species suggests that subsequent

WGDs can promote the loss of ancient paralogs. Because subsequent duplications are expected to relax the level of dosage constraints on genes, we interpret this observation as evidence for the intermediate and recent WGDs lowering the dosage constraints on ancient-WGD-derived paralogs in *P. aurelia*, therefore allowing a higher loss rate of these ancient paralogs. Considering the large difference in post-ancient WGD paralog retention rates between *P. caudatum* and the *aurelia* species, we suspect that selection on gene dosage is the major evolutionary force acting on post-WGD gene retention.

With *P. caudatum* as an outgroup, we were also able to conduct evolutionary analyses of the *aurelia* paralogs from the recent WGD. We found that, compared to *P. biaurelia* and *P. tetraurelia*, *P. sexaurelia* has a higher percentage of paralog pairs that are evolving asymmetrically and that this is likely reflective of the higher rate of gene loss through pseudogenization in this species. This does not preclude the possibility of some neofunctionalized paralogs among the asymmetrical gene pairs, and we have identified a list of candidate genes for neofunctionalization in each species. However, it should be emphasized that *P. aurelia* is a complex of cryptic species that are morphologically indistinguishable (Sonneborn 1975), so that if any new gene functions have evolved, they are not apparent from our current knowledge of the biology of *P. aurelia*. Therefore, it seems likely that neofunctionalization played only a minor role in the evolution of the *P. aurelia* lineage.

The underlying cause for the higher gene loss rate in *P. sexaurelia* remains unclear. One possibility is that the mutation rate in *P. sexaurelia* is higher than in *P. tetraurelia* and *P. biaurelia*, leading to an overall faster evolutionary rate. It is also possible that *P. sexaurelia* has a smaller effective population size than the other *aurelia* species studied, which would result in less efficient selection against gene loss. Data on the mutation rate and the effective population sizes of the different *aurelia* species will be required to test these hypotheses.

Perhaps the most surprising finding is the discovery of divergent resolutions of intermediate WGD duplicates between *aurelia* species. We previously showed that divergent resolution slowed rapidly over time after the recent *aurelia* WGD, as the majority of losses between *P. biaurelia* and *P. tetraurelia* represented parallel, rather than divergent, resolutions (McGrath *et al.*, unpublished results). However, we now demonstrate that, although divergent resolutions may become increasingly rare, they can still occur after a large amount of time has elapsed following WGD. Divergent resolution of duplicates then has the potential to continue to contribute to reproductive isolation and speciation between lineages long after a whole-genome duplication event.

Finally, we point out that this study might help in understanding why the *Paramecium* lineage has been so permissive to WGDs. *Paramecium* cells contain two types of nuclei: the micronucleus (germline) is diploid and transcriptionally silent, while the macronucleus (soma) is generated *de novo* by amplification of a micronucleus after sexual reproduction. The macronucleus is highly polyploid (~800 copies of each chromosome in *P. aurelia*) and is the site of all gene expression. *Paramecium* cells can sense the amount of DNA present in their macronucleus, so that they "synthesize similar amounts of macronuclear DNA, regardless of the number of macronuclei or their prereplication DNA content" (Berger and Schmidt 1978, pp. 116–126). Therefore, it is possible that a WGD would be immediately compensated by a homeostatic maintenance to the macronuclear mass (halving the ploidy of newly generated macronuclei relative to the now tetraploid micronucleus), and only the micronucleus would see its size initially doubling, before gradual gene loss slowly drives it toward its preduplication size. Because gene expression happens only in the macronucleus, most of the immediate deleterious effects of a genome duplication (Otto and Whitton 2000) would be absent under this scenario, making successive WGDs in the *Paramecium* lineage largely neutral events at the time of their establishment and their fixation via genetic drift more likely than in other species. After the successive WGDs, gradual gene loss might lead to a slow decrease in the total amount of macronuclear DNA present in the cell [if the mechanism described in Berger and Schmidt (1978) is not sensitive enough to detect small decreases of DNA content, as expected in the case of gene-by-gene losses], causing a reduction in the size of the macronucleus and therefore a reduction in the total cell size. This scenario would solve the paradoxical observation that, despite having undergone two extra rounds of WGDs and containing about twice as many genes, *P. aurelia* cells are smaller than *P. caudatum* cells. Therefore, the peculiarities of the *Paramecium* genome organization might have enabled this lineage to tolerate multiple rounds of WGDs, resulting in a greatly expanded gene repertoire and offering new evolutionary opportunities to the *Paramecium* lineage.

Acknowledgments

The authors thank O. Arnaiz and L. Sperling for sharing the EuGene annotation pipeline trained on *P. tetraurelia*. Funding for this work has been provided by a National Science Foundation (NSF) Graduate Research Fellowship (to C.L.M.), National Institutes of Health Genetics, Cellular, and Molecular Sciences Training Grant T32 GM007757 (to C.L.M.), and NSF grant EF-0328516-A006 (to M.L.). This research includes work supported by the NSF under grant no. ABI-1062432 to the National Center for Genome Analysis Support at Indiana University.

Literature Cited

Adoutte, A., and J. Beisson, 1972 Evolution of mixed populations of genetically different mitochondria in *Paramecium aurelia*. Nature 235: 393–395.

Allen, S., and I. Gibson, 1972 Genome amplification and gene expression in the Ciliate macronucleus. Biochem. Genet. 5: 161–181.

Armus, H. L., A. R. Montgomery, and R. L. Gurney, 2006 Discrimination learning and extinction in Paramecia (*P. caudatum*). Psychol. Rep. 98: 705–711.

Arnaiz, O., and L. Sperling, 2011 ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. Nucleic Acids Res. 39: D635–D636.

Arnaiz, O., N. Mathy, C. Baudry, S. Malinsky, J.-M. Aury *et al.*, 2012 The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PLoS Genet. 8: e1002984.

Aury, J.-M., O. Jaillon, L. Duret, B. Noel, C. Jubin *et al.*, 2006 Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444: 171–178.

Beale, G. H., A. Jurand, and J. R. Preer, 1969 The classes of endosymbiont of *Paramecium aurelia*. J. Cell Sci. 5: 65–91.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57: 289–300.

Berger, J. D., 1937 Nuclear differentiation and nucleic acid synthesis in well-fed exconjugants of *Paramecium aurelia*. Chromosoma 42: 247–268.

Berger, J. D., and H. J. Schmidt, 1978 Regulation of macronuclear DNA content in Paramecium tetraurelia. J. Cell Biol. 76: 116–126.

Calkins, G. N., 1902 Studies on the life-history of Protozoa, I. The life-cycle of Paramaecium caudatum. Dev. Genes Evol. 15: 139–186.

Davis, J. C., and D. A. Petrov, 2005 Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet. 21: 548–551.

Decatur, W. A., J. A. Hall, J. J. Smith, W. Li, and S. A. Sower, 2013 Insight from the lamprey genome: glimpsing early vertebrate development via neuroendocrine-associated genes and shared synteny of gonadotropin-releasing hormone (GnRH). Gen. Comp. Endocrinol. 192: 237–245.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19: 11–15.

Doyle, J. J., L. E. Flagel, A. H. Paterson, R. A. Rapp, D. E. Soltis *et al.*, 2008 Evolutionary genetics of genome merger and doubling in plants. Annu. Rev. Genet. 42: 443–461.

Duncan, A. B., S. Fellous, R. Accot, M. Alart, K. C. Sobandi *et al.*, 2010 Parasite-mediated protection against osmotic stress for Paramecium caudatum infected by Holospora undulata is host genotype specific. FEMS Microbiol. Ecol. 74: 353–360.

Duncan, A. B., S. Fellous, and O. Kaltz, 2011 Reverse evolution: selection against costly resistance in disease-free microcosm populations of *Paramecium caudatum*. Evolution 65: 3462–3474.

Duret, L., J. Cohen, C. Jubin, P. Dessen, J. F. Gout *et al.*, 2008 Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline. Genome Res. 18: 585–596.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Eisen, J. A., R. S. Coyne, M. Wu, D. Wu, M. Thiagarajan *et al.*, 2006 Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4: e286.

Fellous, S., A. Duncan, A. Coulon, and O. Kaltz, 2012 Quorum sensing and density-dependent dispersal in an aquatic model system. PLoS ONE 7: e48436.

Foissac, S., J. Gouzy, S. Rombauts, C. Mathe, J. Amselem *et al.*, 2008 Genome annotation in plants and fungi: EuGene as a model platform. Curr. Bioinform. 3: 87–97.

Fokin, S. I., 2010 *Paramecium* genus: biodiversity, some morphological features and the key to the main morphospecies discrimination. Protistology 6: 227–235.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159: 907–911.

Gout, J.-F., D. Kahn, and L. Duret, 2010 The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 6: e1000944.

Hailong, Z., L. Guofeng, H. Renliang, C. Song, and D. Xiaoping, 2011 Initial study on acute toxicity of two typical persistent organic pollutanta (POPs) against Paramecium caudatum. Asian J. Ecotoxicol. 6: 37–42.

Han, M. V., J. P. Demuth, C. L. McGrath, C. Casola, and M. W. Hahn, 2009 Adaptive evolution of young gene duplicates in mammals. Genome Res. 19: 859–867.

Hughes, T., and D. A. Liberles, 2008 Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. J. Mol. Evol. 67: 343–357.

Igarashi, S., 1966 Temperature-sensitive mutation in Paramecium aurelia. I. Induction and inheritance. Mutat. Res. 3: 13–24.

Jennings, H. S., 1908 Heredity, variation and evolution in Protozoa. 1. The fate of new structural characters in Paramecium, in connection with the problem of the inhertiance of acquired characters in unicellular organisms. J. Exp. Zool 5: 577–632.

Jennings, H. S., 1916 The numerical results of diverse systems of breeding. Genetics 1: 53–89.

Jennings, H. S., 1917 The numerical results of diverse systems of breeding, with respect to two paris of characters, linked or independent, with special relation to the effects of linkage. Genetics 2: 97–154.

Jiao, Y. N., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97–100.

Johnson, D. A., and M. A. Thomas, 2007 The monosaccharide transporter gene family in Arabidopsis and rice: a history of duplications, adaptive evolution, and functional divergence. Mol. Biol. Evol. 24: 2412–2423.

Kawamoto, K., T. Oashi, K. Oami, W. Liu, Y. H. Jin *et al.*, 2010 Perfluorooctanoic acid (PFOA) but not perffluorooctane sulfonate (PFOS) showed DNA damage in comet assay on *Paramecium caudatum*. J. Toxicol. Sci. 35: 835–841.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley *et al.*, 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14: R36.

Krenek, S., T. Petzoldt, and T. U. Berendonk, 2012 Coping with temperature at the warm edge: patterns of thermal adaptation in the microbial eukaryote *Paramecium caudatum*. PLoS ONE 7: e30598.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Lynch, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Lynch, M., and A. G. Force, 2000 The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. 156: 590–605.

Mi, H., A. Muruganujan, and P. D. Thomas, 2012 PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41: D377–D386.

Miller, J. R., A. L. Delcher, S. Koren, E. Venter, B. P. Walenz *et al.*, 2008 Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24: 2818–2824.

Morin, R. D., E. Chang, A. Petrescu, N. Liao, M. Griffith *et al.*, 2006 Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals

post-tetraploidization transcriptome remodeling. Genome Res. 16: 796–803.
Nowacki, M., W. Zagorski-Ostoja, and E. Meyer, 2005 Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in *Paramecium tetraurelia*. Curr. Biol. 15: 1616–1628.
Oka, H. I., 1988 Functions and genetic bases of reproductive barriers, pp. 181–209 in *Origin of Cultivated Rice*, edited by H. I. Oka. Japan Scientific Societies Press/Elsevier, Tokyo.
Otto, S. P., 2007 The evolutionary consequences of polyploidy. Cell 131: 452–462.
Otto, S. P., and J. Whitton, 2000 Polyploid incidence and evolution. Annu. Rev. Genet. 34: 401–437.
Panopoulou, G., and A. J. Poustka, 2005 Timing and mechanism of ancient vertebrate genome duplications: the adventure of a hypothesis. Trends Genet. 21: 559–567.
Pessia, E., A. Popa, S. Mousset, C. Rezvoy, L. Duret *et al.*, 2012 Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol. Evol. 4: 675–682.
Pianka, E. R., 2000 *Evolutionary Ecology*. Benjamin Cummings, San Francisco.
Postlethwait, J. H., I. G. Woods, P. Ngo-Hazelett, Y. L. Yan, P. D. Kelly *et al.*, 2000 Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res. 10: 1890–1902.
Putnam, N. H., T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten *et al.*, 2008 The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071.
Rao, J. V., V. G. Gunda, K. Srikanth, and S. K. Arepalli, 2007 Acute toxicity bioassay using Paramecium caudatum, a key member to study the effects of monocrotophos on swimming behaviour, morphology and reproduction. Toxicol. Environ. Chem. 89: 307–317.
Schiex, T., A. Moisan, and P. Rouze, 2001 EuGene: an eucaryotic gene finder that combines several sources of evidence, in *Computational Biology*, edited by O. Gascuel and M. -F. Sagot. Springer Berlin Heidelberg 2066: 111–125.
Siegel, R. W., 1967 Genetics of ageing and the life cycle in ciliates. Symp. Soc. Exp. Biol. 21: 127–148.
Simillion, C., K. Vandepoele, C. Van Montagu, M. Zabeau, and Y. Van de Peer, 2002 The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 99: 13627–13632.
Singh, D. P., B. Saudemont, G. Guglielmi, O. Arnaiz, J. F. Gout *et al.*, 2014 Genome-defence small RNAs exapted for epigenetic mating-type inheritance. Nature.
Sonneborn, T. M., 1933 Mendelian methods applied to the ciliate protozoan, Paramecium caudatum. Am. Nat. 57: 72.
Sonneborn, T. M., 1937 Sex, sex inheritance and sex determination in Paramecium aurelia. Proc. Natl. Acad. Sci. USA 23: 378–385.
Sonneborn, T. M., 1947 Developmental mechanisms in *Paramecium*. Growth Symposia 11: 291–307.
Sonneborn, T. M., 1975 The *Paramecium aurelia* complex of fourteen sibling species. Trans. Am. Microsc. Soc. 94: 155–178.
Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24: 637–644.
Sung, W., A. E. Tucker, T. G. Doak, E. Choi, W. K. Thomas *et al.*, 2012 Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. Proc. Natl. Acad. Sci. USA 109: 19339–19344.
Tajima, F., 1993 Simple methods for testing molecular clock hypothesis. Genetics 135: 599–607.
Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn *et al.*, 2012 Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31: 46–53.
Veitia, R. A., S. Bottani, and J. A. Birchler, 2008 Cellular reactions to gene dosage imbalance: genomic, transcriptomic, and proteomic effects. Trends Genet. 24: 390–397.
Werth, C. R., and M. D. Windham, 1991 A model for divergent, allopatric speciation of polyploid Pteridophytes resulting from silencing of duplicate-gene expression. Am. Nat. 137: 515–526.
Wichterman, R., 1939 Cytogamy: a new sexual process in joined pairs of Paramecium caudatum. Nature 144: 123–124.
Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
Yanagi, A., and N. Haga, 1995 A simple method of induction of autogamy by methyl cellulose in *Paramecium caudatum*. Zoolog. Sci. 12: 26.
Yanagi, A., and N. Haga, 1998 Induction of conjugation by methyl cellulose in *Paramecium*. J. Eukaryot. Microbiol. 45: 87–90.
Yang, J., R. Lusk, and W.-H. Li, 2003 Organismal complexity, protein complexity, and gene duplicability. Proc. Natl. Acad. Sci. USA 100: 15661–15665.
Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.

Communicating editor: G. Stormo

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163287/-/DC1

Insights into Three Whole-Genome Duplications Gleaned from the *Paramecium caudatum* Genome Sequence

Casey L. McGrath, Jean-Francois Gout, Thomas G. Doak, Akira Yanagi, and Michael Lynch

DOI: 10.1534/genetics.114.163287

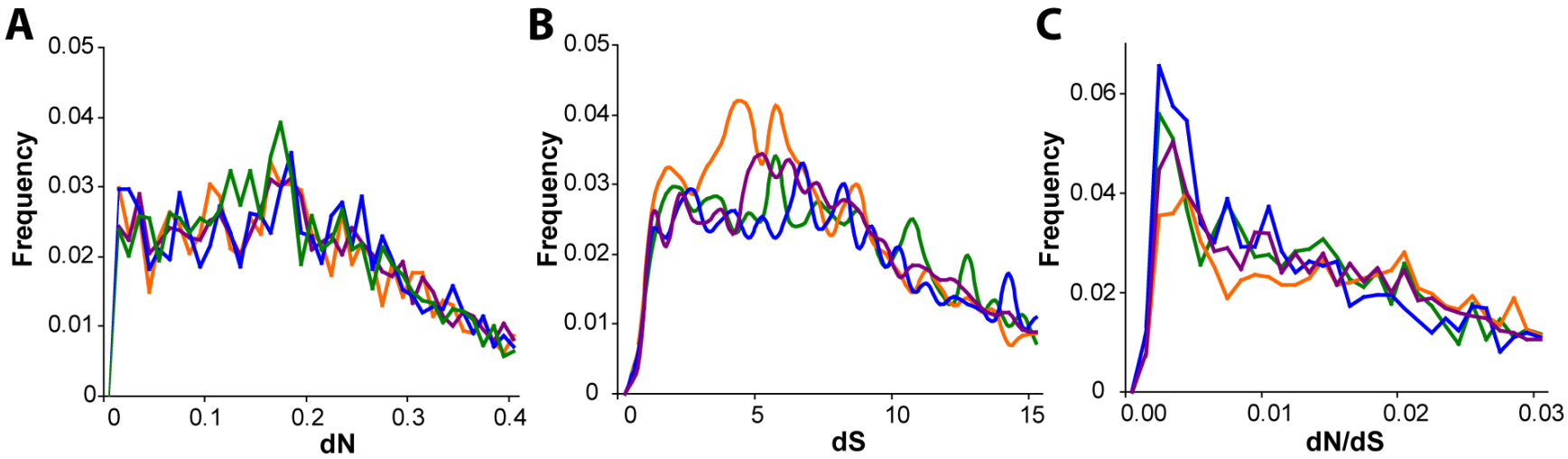

Figure S1 Distributions of dN (A), dS (B), and dN/dS (C) among intermediate WGD duplicates. Green = intermediate WGD duplicates within *biaurelia*, orange = intermediate WGD duplicates within *tetraurelia*, blue = intermediate WGD duplicates within *sexaurelia*, purple = intermediate WGD duplicates between *tetraurelia* and *sexaurelia*.

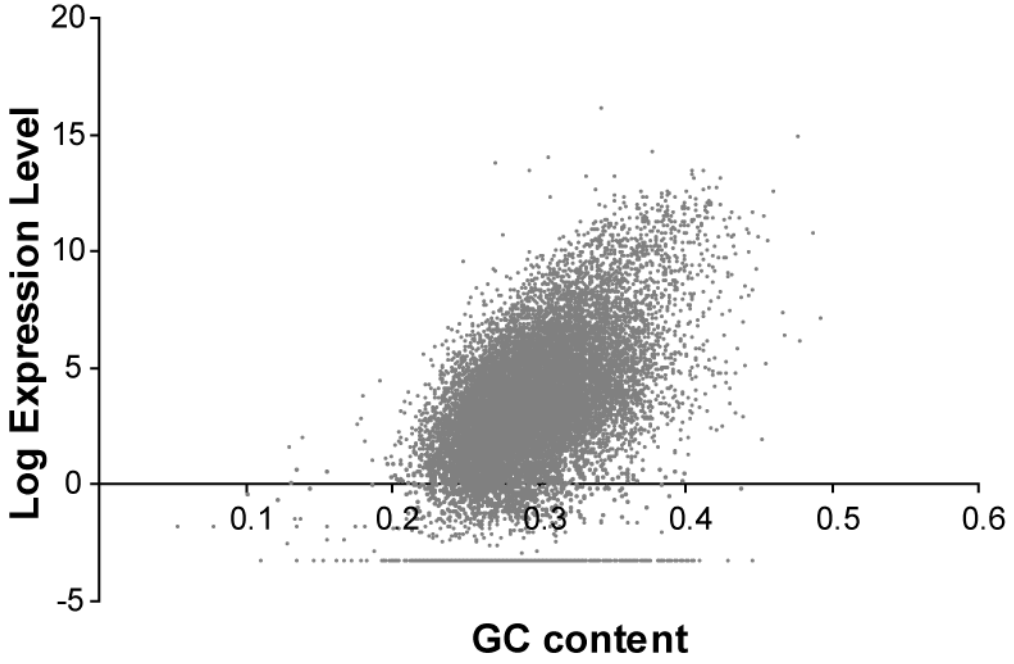

Figure S2 Relationship between log expression level and GC content of *caudatum* genes. R^2=0.23

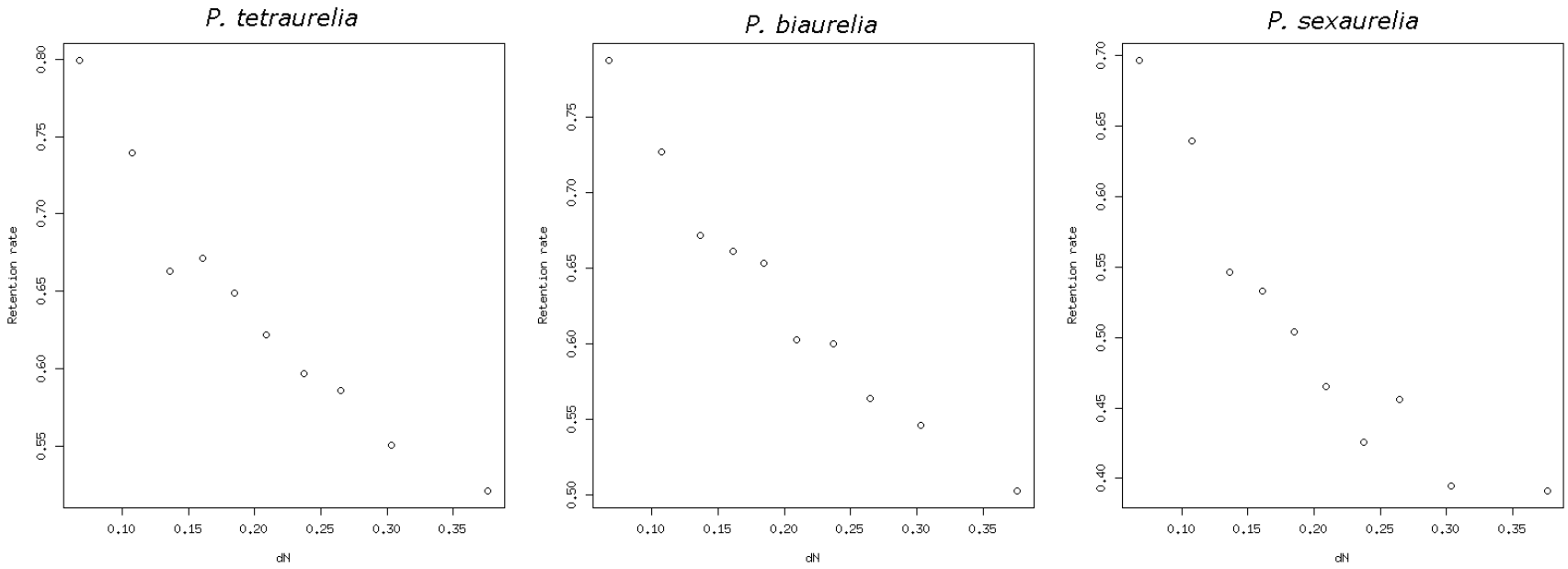

Figure S3 Relationship between evolutionary rate (dN computed between *P. caudatum* / *P. multimicronucleatum* orthologs) and post-recent-WGD retention rate in three *P. aurelia* species.

Files S1-S6

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163287/-/DC1

File S1 Excel file containing orthologous blocks between *caudatum* genes and recreated pre-recent WGD *aurelia* segments. Cells where one gene has been lost from a scaffold contain a dot (“.”).

File S2 Excel file containing orthologous blocks between *caudatum* genes and recreated pre-intermediate WGD *aurelia* segments. Cells where one gene has been lost from a scaffold contain a dot (“.”).

File S3 Excel file containing number of intermediate WGD duplicates vs. single-copy genes in each GO term functional category and *P*-values.

File S4 Excel file containing candidates for neofunctionalization from *biaurelia*, *tetraurelia*, and *sexaurelia*.

File S5 Excel file containing paralogous blocks within *caudatum* from the ancient WGD.

File S6 Excel file containing number of ancient WGD duplicates vs. single-copy genes in each GO term functional category and *P*-values.

Table S1 Genome assembly and annotation statistics for *P. caudatum*, as compared to *biaurelia* (McGrath *et al*., submitted), *tetraurelia* (AURY *et al.* 2006) and *sexaurelia* (McGrath *et al.,* submitted).

	caudatum	*biaurelia*	*tetraurelia*	*sexaurelia*
Average coverage	186X	45X	13X	42X
Number of scaffolds (total)	1,202	2,362	N/A*	547
Number of scaffolds (> 2 kb)	274	1,426	697	230
Average scaffold length (>2 kb scaffolds)	109,242	53,140	103,448	294,183
Largest scaffold length	793,585	1,048,449	980,760	1,303,432
Number of gaps	1,412	1,459	419	1,298
Assembly length with gaps (all scaffolds)	30,525,943	76,976,592	72,102,941	68,020,722
Number of genes	18,509	39,242	39,521	34,939

* Only scaffolds > 2 kb are included in *tetraurelia* assembly

Table S2 Divergent resolutions of intermediate WGD duplicates between *aurelia* species.

Intermediate Duplicate 1 descendant(s)		Intermediate Duplicate 2 descendants (s)		Function (if known)
Divergent resolutions between *tetraurelia* and *sexaurelia*				
GSPATP00020634001	GSPATP00027241001	PSEXGNP07757	PSEXGNP11110	Serine/threonine protein kinase NEK
GSPATP00001449001	GSPATP00002598001	PSEXGNP12568	--	KH domain containing RNA binding protein
Divergent resolutions between *biaurelia* and *sexaurelia*				
PBIGNP32737	PBIGNP21940	PSEXGNP01107	PSEXGNP03977	Serine/threonine protein kinase-related
PBIGNP15479	PBIGNP27623	PSEXGNP17612	--	Ribosomal protein L15
PBIGNP00423	PBIGNP01848	PSEXGNP12568	--	KH domain containing RNA binding protein
PBIGNP00575	--	PSEXGNP18664	--	
Divergent resolutions between *biaurelia* and *tetraurelia*				
PBIGNP34817	--	GSPATP00027407001	--	
PBIGNP00237	--	GSPATP00004767001	--	Cabriolet-related

Table S3 Average GC content and log expression level for duplicated vs. single-copy *caudatum* genes from the ancient WGD.

	Duplicated	Single-copy	*P*-value
GC content	0.306	0.296	$<10^{-6}$
Log expression level	4.33	3.50	$<10^{-8}$