Evidence of the Red-Queen Hypothesis from Accelerated Rates of Evolution of Genes Involved in Biotic Interactions in *Pneumocystis*

Luis Delaye[1,*], Susana Ruiz-Ruiz[2], Enrique Calderon[3,4], Sonia Tarazona[5,6], Ana Conesa[5,7], and Andrés Moya[2,8,*]

[1]Departamento de Ingeniería Genética, CINVESTAV Irapuato, Guanajuato, México

[2]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO)-Salud Pública, València, Spain

[3]Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/Consejo Superior de Investigaciones Científicas/Universidad de Sevilla

[4]Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

[5]Centro de Investigacion Principe Felipe, València, Spain

[6]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Spain

[7]Microbiology and Cell Science, University of Florida

[8]Institute for Integrative Systems Biology, Universitat de València, Spain

*Corresponding authors: E-mails: luis.delaye@cinvestav.mx; andres.moya@uv.es.

Accepted: June 5, 2018

Abstract

Pneumocystis species are ascomycete fungi adapted to live inside the lungs of mammals. These ascomycetes show extensive stenoxenism, meaning that each species of *Pneumocystis* infects a single species of host. Here, we study the effect exerted by natural selection on gene evolution in the genomes of three *Pneumocystis* species. We show that genes involved in host interaction evolve under positive selection. In the first place, we found strong evidence of episodic diversifying selection in Major surface glycoproteins (Msg). These proteins are located on the surface of *Pneumocystis* and are used for host attachment and probably for immune system evasion. Consistent with their function as antigens, most sites under diversifying selection in Msg code for residues with large relative surface accessibility areas. We also found evidence of positive selection in part of the cell machinery used to export Msg to the cell surface. Specifically, we found that genes participating in glycosylphosphatidylinositol (GPI) biosynthesis show an increased rate of nonsynonymous substitutions (d*N*) versus synonymous substitutions (d*S*). GPI is a molecule synthesized in the endoplasmic reticulum that is used to anchor proteins to membranes. We interpret the aforementioned findings as evidence of selective pressure exerted by the host immune system on *Pneumocystis* species, shaping the evolution of Msg and several proteins involved in GPI biosynthesis. We suggest that genome evolution in *Pneumocystis* is well described by the Red-Queen hypothesis whereby genes relevant for biotic interactions show accelerated rates of evolution.

Key words: stenoxenism, majors surface glycoproteins, glycosylphosphatidylinositol, natural selection.

Introduction

Pneumocystis is an ascomycete adapted to live inside the lungs of mammals. It was initially described erroneously as a new trypanosomal life form independently by Carlos Chagas and Antonio Carinii (Chagas 1909; Carinii 1910). A few years later, researchers from the Pasteur Institute realized that the microorganism was a different species of parasite and named it *Pneumocystis carinii* (Delanoë and Delanoë 1912). However, it was not until 1988 when *P. carinii* was correctly identified as a member of the fungal kingdom based on phylogenetic analyses of its ribosomal RNA (Edman et al. 1988; Stringer et al. 1989).

One of the peculiarities of this ascomycete is its stenoxenism, such that each *Pneumocystis* species can infect only a single host species (Gigliotti et al. 1993; Aliouat-Denis et al. 2008). The association is so close that it is possible to recover the host phylogeny by using *Pneumocystis* genes; as demonstrated for *Pneumocystis* isolated from diverse primate species (Demanche et al. 2001; Derouiche et al. 2009). This suggests

a process of adaptation and coevolution between *Pneumocystis* and its host.

The species infecting humans, *P. jirovecii*, causes pneumonia in immunocompromised hosts (Catherinot et al. 2010). It is among the top 10 invasive fungal infections worldwide (Brown et al. 2012). Regarding its lifestyle, *Pneumocystis* has been described as a biotroph (Cushion et al. 2007; Hauser 2014); that is, a parasite that feeds from living cells without killing them. In contrast, necrotrophs kill host cells for feeding. It has been suggested too that *Pneumocystis* acquired its biotrophy by massive gene loss (Cissé et al. 2014).

The genomes from *P. carinii*, *P. murina*, and *P. jirovecii* were sequenced recently (Ma et al. 2016). These species parasitize rats, mice, and humans, respectively. The genome sizes varied between 7.50 and 8.40 Mb, showed low G + C content and a relatively small number of protein coding genes (from 3,623 in *P. murina* to 3,761 in *P. jirovecii*). The genome from *P. jirovecii* showed several rearrangements when compared with those from *P. carinii* and *P. murina*.

Pneumocystis genomes encodes a multicopy family of *msg* genes coding for major surface glycoproteins (Msg). These genes are coded in tandem near telomeres which promotes recombination (Keely et al. 2005). Msg proteins are exported to the cell membrane where they are used to attach to host structures and probably to evade the host immune system (Thomas and Limper 2007). In addition, Msg are the most abundant proteins in the cell membrane (Kutty, Shroff, et al. 2013). Based on sequence similarity, Msg proteins were classified into five families: Msg-A, -B, -C, -D, and -E (Ma et al. 2016). The largest family by far (Msg-A), is further subdivided into the: Msg-A1, -A2, and -A3 subfamilies. High quality *msg* gene sequences were obtained recently by a combination of PacBio sequencing technology and a clustering based analysis pipeline (Liang et al. 2016). More recently, Msg proteins from *P. jirovecii* were classified into six families (msg-I to -VI) (Schmid-Siegert et al. 2017).

Here, we study the effect exerted by natural selection on functionally related sets of genes in *P. carinii*, *P. murina*, and *P. jirovecii*. Specifically, we asked whether there are GO categories enriched in genes under negative or positive selection. We also investigated the pattern of natural selection and recombination in Msg protein coding genes from *P. jirovecii*. We show that Msg protein coding genes and part of the cell machinery used to export proteins to the cell surface show increased rates of nonsynonymous substitutions (d*N*). We suggest that genome evolution in *Pneumocystis* is well described by the Red-Queen hypothesis whereby genes involved in biotic interactions show accelerated rates of evolution.

Materials and Methods

Ortholog Identification

The genome assembly of *P. carinii* (LFVZ01.1), *P. jirovecii* (LFWA01.1), and *P. murina* (AFWA02.1) were retrieved from Genbank. Predicted proteins were annotated by using Blast2GO version 4.0.7 (Götz et al. 2008). Ortholog protein families were identified by using GET_HOMOLOGUES.pl (Vinuesa and Contreras-Moreira 2015) with the MCL algorithm (Li et al. 2003) and default parameters. For the rest of the analysis, we considered only single copy orthologs (i.e., protein families having one sequence per *Pneumocystis* sp.).

Natural Selection Analyses

Protein families were aligned with MUSCLE (Edgar 2004). Then, by using protein alignments as templates, we aligned gene families with exonerate (https://github.com/nathanweeks/exonerate, last accessed 2017). Next, we estimated the rate of synonymous (d*S*) and nonsynonymous (d*N*) substitutions and the omega = d*N*/d*S* rate of each gene family. For that purpose, we used a branch model implemented in CodeML from the PAML package (Yang 2007).

Because the rate of molecular evolution varies between genes within gene families, we tested whether molecular evolution is best explained by one or three omega rates (i.e., one rate for each of the three *Pneumocystis* species). The single omega rate model corresponds to the null hypothesis (H_0) and the three omega model to the alternative (H_1) hypotheses. The two models of evolution are compared by using a likelihood ratio test (LRT). LRT is defined as $2(\ln LH_1 - \ln \ln LH_0)$ and the significance of the test is obtained from a chi-square (χ^2) distribution with two degrees of freedom. The $\ln LH_0$ and $\ln LH_1$ stands for the logarithm of the likelihood of null (H_0) and alternative (H_1) hypotheses, respectively.

Next, the result of each LRT was transformed to *p*-values by using the R function *pchisq* (The R Project for Statistical Computing, https://www.R-project.org/; last accessed 2018). Since we have multiple tests (one for each family of proteins), we applied a false discovery rate (FDR) correction to the above list of *P* values. We considered as significant only those *P* values associated to a FDR <0.05. Accordingly, we assigned omega values to branches calculated under the alternative hypothesis only if FDR < 0.05. Otherwise, we assigned omega values to branches calculated under the null hypothesis. By this, we ended up with three lists (one for each *Pneumocystis* species), with the estimated omega (d*N*/d*S*) rates of each gene in all gene families. We tested the reliability of estimating the rate d*N*/d*S* from only three sequences by a simulation analysis. Details of this simulation are found in supplementary protocol S1 from additional file 1, Supplementary Material online.

Gene Set Enrichment Analysis

Our analysis was inspired by the work of Serra et al. (2011). We aimed at identifying functional modules enriched in genes sharing extreme d*N*/d*S* ratios. For this, we used GO categories to define functional modules (Ashburner et al. 2000). Then, by using Blast2GO version 4.0.7 (Götz et al. 2008), we performed GSEA (Subramanian et al. 2005) to ask whether

there are GO categories enriched in genes with extreme omega ratios in each one of the *Pneumocystis* species. We considered as significant only those enriched categories having an FDR < 0.05.

Molecular Evolution of Msg Coding Genes

The 179 Msg protein coding genes were retrieved from the original publication (Ma et al. 2016). Msg proteins were aligned with MUSCLE (Edgar 2004); and *msg* genes were codon aligned with exonerate by using corresponding protein alignments as templates. The multiple sequence alignment containing Msg coding genes was manually curated to eliminate small sequences. By this, we ended with an alignment of 137 sequences. We used the curated alignment for the rest of the analysis.

To detect recombination, we used GARD from HyPhy 2.220 package (Kosakovsky et al. 2006). We used the HKY85 model (010010) of evolution and homogeneous rate of evolution across sites. Results from GARD were processed with GARDProcessor. Then episodic diversifying selection was detected with MEME on each of the recombinant segments previously identified with GARD (Murrell et al. 2012). The options used were: [Model Options] GTR model (012345); we used the tree inferred with GARD; [d*N*/d*S* bias parameter options] option 5, fast; [Ancestor Counting Options] option 11, corresponding to MEME analysis; [Significance level for Likelihood Ratio Tests] we used default value = 0.1. GARD and MEME were run on the cluster *mazorka* from Langebio-UGA facilities (Cluster de Cómputo de Alto Rendimiento Mazorka, http://mazorka.langebio.cinvestav.mx, last accessed 2018).

Phylogenetic trees from nonrecombinant segments from *P. jirovecii* Msg-A protein coding genes were reconstructed by maximum-likelihood with PhyML (Guindon et al. 2010) and by MEGA7 (Kumar et al. 2016). Best fitting models of molecular evolution were identified with MEGA7 and with HyPhy 2.220 package (Pond et al. 2005). And 100 bootstrap replicas were used to evaluate the confidence of nodes in the tree. For tree manipulation, we used ETE Toolkit (Huerta-Cepas et al. 2016) and *ape* package version 4.1 from R (Paradis et al. 2004).

Protein secondary structure prediction and surface accessibility was predicted by using NetSurfP ver 1.1 (Petersen et al. 2009). All statistical analyses were conducted in R (The R Project for Statistical Computing, https://www.R-project.org/; last accessed 2018).

Results

Functional Modules Enriched in Genes with Extreme d*N*/d*S* Ratios

We identified 2,989 families of ortholog genes among *P. carinii*, *P. murina*, and *P. jirovecii*. Of these, 2,967 families were single copy orthologs (i.e., each *Pneumocystis* species was represented by a single gene). We estimated the rate of nonsynonymous (d*N*) versus synonymous (d*S*) substitutions (d*N*/d*S* = omega) of all these 2,967 gene families by using a branch model implemented in CodeML (Yang 2007). Because the rate of molecular evolution varies, we tested whether the evolution of genes within gene families is best described by one rate of evolution (H_0), or three rates of evolution (H_1). The null hypothesis (H_0) assumes a single omega rate, while the alternative hypothesis (H_1) assumes that each one of the three branches in the phylogeny has its own omega rate. To assess the significance, we used a likelihood ratio test (LRT) with two degrees of freedom. Since we have multiple tests (one test for each of the 2,967 gene families), we applied a false discovery rate (FDR or Benjamini–Hochberg procedure) on the list of *P* values derived from the 2,967 tests and rejected the null hypothesis only in those cases showing a $Q < 0.05$. We found that 2,781 gene families were better described by a single rate (hypothesis H_0) and 186 gene families were better described by three rates (hypothesis H_1).

Next, for each *Pneumocystis* species, we performed a gene set enrichment analysis (GSEA) to identify gene categories enriched in genes with similar omega rates. GSEA is a statistical technique widely used to identify functional categories enriched in genes showing extreme expression levels. GSEA can identify subtle differences in gene expression, or other quantitative properties of genes (Subramanian et al. 2005). In our case, we assigned GO categories to gene families with Blast2GO (Götz et al. 2008) and then performed the GSEA to identify GO categories enriched in genes with extreme omega rates.

To perform the GSEA, we used only GO terms common to all three genes within each family of orthologous genes. This allowed us to interpret the results of GSEA only in terms of differences in rates of evolution between genes. Of the 2,967 families of orthologs, 2,638 were composed of genes sharing at least one GO term. These were divided in 2,467 gene families that did not rejected the null hypothesis of a single omega rate and 171 genes that were better described by three omega rates (table 1). The rest of the families (329) were not used in the GSEA. These correspond to those where one or two genes in the family failed to have any GO annotation (315 cases); or by families composed by genes that do not share any GO term (14 cases).

There were 3,689 GO terms associated to these 2,638 gene families. However, the distribution of GO terms in gene families was highly skewed toward small values (i.e., there were many GO categories represented by very few gene families). Therefore, we restricted our analysis to all GO categories with >15 and <500 gene families. This is because very large GO categories are not informative when subjected to enrichment analysis. This resulted in 2,638 gene families associated to 227 GO terms.

We found several GO terms showing significant normalized enrichment scores (NES) (fig. 1 and supplementary figs. S1 and S2, Supplementary Material online; table 2 and

Table 1

Selection Analysis of Gene Families in *Pneumocystis* Species

Species	Protein Coding Genes	Single Copy Gene Families	H_0 Gene Families	H_1 Gene Families	H_0 Gene Families Sharing GO Terms	H_1 Gene Families Sharing GO Terms
P. carinii	3,646					
P. murina	3,623	2,967	2,781	186	2,467	171
P. jirovecii	3,761					

NOTE.—Of the 2,967 single copy gene families (orthologs), 2,781 did not reject the null hypothesis of a single omega rate and 186 did reject the hypothesis. From these gene families (2,781 + 186), we selected for gene set enrichment analysis (GSEA), those endowed with genes sharing GO terms. These correspond to the last two columns with 2,467 + 171 gene families. At the same time, only within family shared GO terms were used for GSEA.

H_0, those families not rejecting the null hypothesis of a single omega rate of evolution; H_1, those rejecting the null hypothesis ($Q < 0.05$).

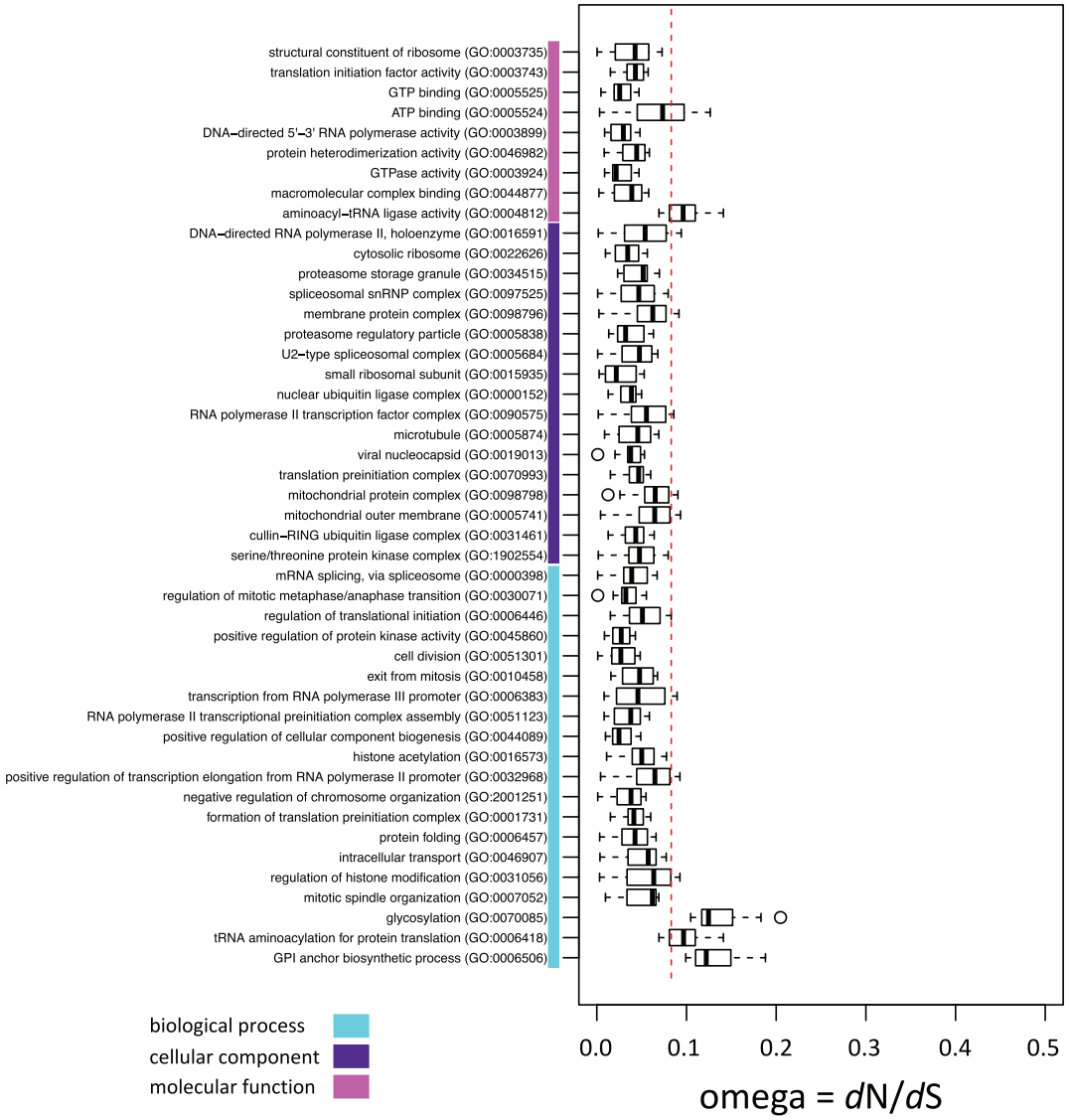

FIG. 1.—GO categories enriched in gene families showing high or low omega (d*N*/d*S*) values for *Pneumocystis jirovecii*. The red line indicates the median (d*N*/d*S*) for all genes. GO categories are ordered from small to large NES (normalized enrichment scores) within each GO type (biological process, cellular component, and molecular function). Each box denotes the median and the 25% and 75% percentiles.

Table 2

Number of GO Terms Significantly Enriched ($Q < 0.05$) in Genes Showing Distinctive Omega Rates

Species	Number of GO Terms Used for GSEA[a]	Number of GO Terms with (−) NES	Number of GO Terms with (+) NES
Pneumocystis carinii	227	50	1
P. murina		52	6
P. jirovecii		42	4

NOTE.—GO terms showing negative NES correspond to those enriched in genes showing lower omega rates; and GO terms showing positive NES correspond to those enriched in genes showing higher omega rates.

NES, normalized enrichment score.

[a]Number of GO terms associated to 2,638 gene families (see text for details).

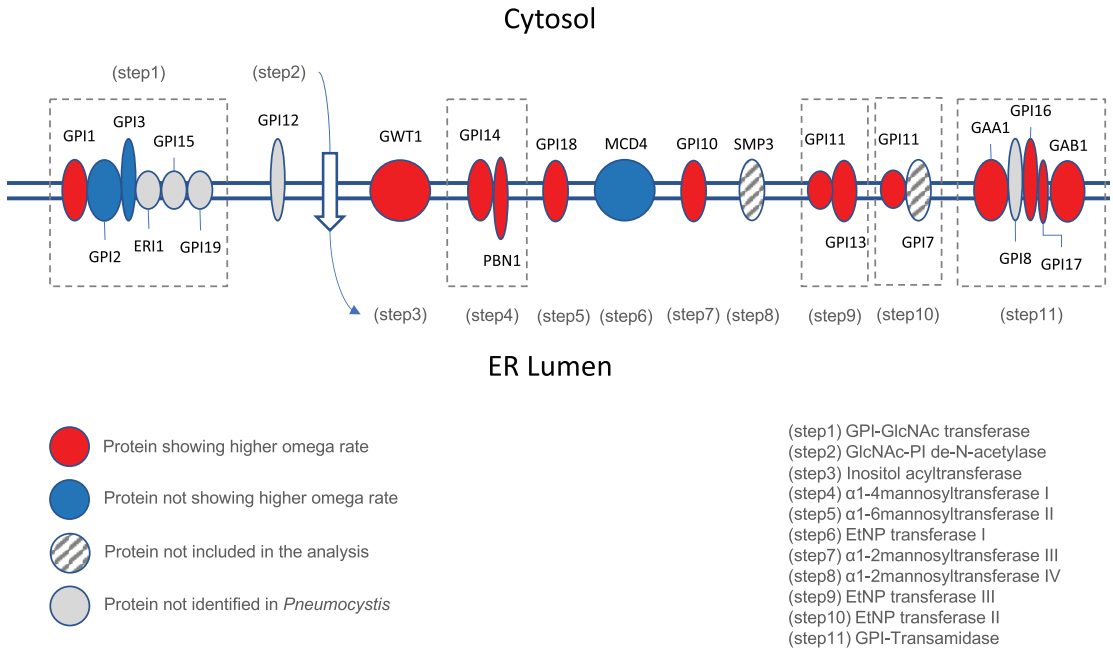

FIG. 2.—Most genes in the biosynthesis of GPI show elevated omega rates. In red, we show those proteins having higher omega rates. These proteins contribute to the statistical significance of the GO: 0006506 term. In blue, we show those proteins not having higher rates of evolution. Hatched proteins were not included in GSEA. Gray proteins were not identified in *Pneumocystis*. Protein complexes are shown within dashed boxes.

supplementary table S1, Supplementary Material online). In GSEA, the NES is the primary statistic used to identify GO terms enriched in genes having extreme values. In our case, negative NES correspond to GO terms enriched in genes with the lowest omega rates; while positive NES, correspond to terms enriched in genes with the highest omega rates.

In our results, the most significant enriched GO terms have negative NES. Moreover, the same GO terms show the smallest NES in *P. jirovecii*, *P. carinii*, and *P. murina* (supplementary table S1, Supplementary Material online), indicating that similar selective pressures operate in all three genomes irrespective of the host species. Purifying or negative selection in these functional categories is strongest.

We also found GO categories with positive NES. Positive NES indicate categories enriched in genes with higher omega rates. Notably, the GPI anchor biosynthetic process (GO: 0006506) is statistically significant in the three genomes and ranks as the GO term with the highest NES (supplementary table S1, Supplementary Material online). GPI is a molecule synthesized in the endoplasmic reticulum (ER) that is used to anchor proteins to the outer leaflet of cell membranes (Ferguson et al. 2009). GPI is synthesized by at least eleven enzymatic steps (Pittet and Conzelmann 2007; Fujita and Kinoshita 2010). These steps are performed by transmembrane proteins in the ER, some of which conform multimeric complexes. In addition, most of the genes participating in GPI biosynthesis contribute to the statistical significance of the term (fig. 2 and supplementary table S2, Supplementary Material online). Interestingly, major surface glycoproteins (Msg) are GPI-anchored to membranes (Kutty, England, et al. 2013).

Other terms showing positive NES were: aminoacyl-tRNA ligase activity (GO: 0004812) and tRNA aminoacylation for protein translation (GO: 0006418) in *P. jirovecii* and *P. murina*;

Golgi apparatus (GO: 0005794), mating projection tip (GO: 0043332) and protein glycosylation (GO: 0006486) in *P. murina*; and finally, Glycosylation (GO: 0070085) in *P. jirovecii*. None of these GO terms reached statistical significance in the three species coincidentally.

Selection and Recombination in Msg Protein Coding Genes

Episodic selection occurs when a site in the multiple alignment evolves by positive selection only on a subset of branches on a phylogenetic tree. This contrasts with pervasive selection, where for a given site, most (or all) branches on a phylogenetic tree evolve by positive selection. Episodic selection is more common than pervasive selection (Murrell et al. 2012). We studied the pattern of episodic selection and recombination among sites in Msg protein coding genes. Msg proteins were previously classified into five families (A, B, C, D, and E) (Ma et al. 2016). Family A is further subdivided into A1, A2, and A3. Family A2 is present in *P. carinii* and *P. murina*, but not in *P. jirovecii*. As mentioned earlier, Msg proteins from *P. jirovecii* were classified more recently into families six families (msg-I to -VI) (Schmid-Siegert et al. 2017). Both classifications are fairly consistent. Families A1, B, D, and E from Ma et al. (2016) correspond to families I, IV, V, and VI from Schmid-Siegert et al. (2017). The only difference is that family A3 from Ma et al. (2016) is correctly divided into families II and III by Schmid-Siegert et al. (2017). We will refer to these as families A3(II) and A3(III). The correspondence between both classifications is shown in supplementary figure S3, Supplementary Material online. We note that some of the sequences originally classified by Ma et al. (2016) as A3, belong to family A1.

We first reconstructed a phylogenetic tree with *msg* gene sequences. After careful manual curation to exclude small sequences, we aligned 137 out of the 179 sequences reported by Ma et al. (2016). These include 75, 12, 16, 11, and 18 sequences from families A1, A3(II), A3(III), B, and D, respectively. Family E was not included in this tree due to the large number of gaps introduced by these relatively small sequences. With the exception of five A3 sequences that grouped with A1 genes, our phylogenetic reconstruction is in general agreement with those from Ma et al. (2016) and Schmid-Siegert et al. (2017) (supplementary fig. S4, Supplementary Material online).

We next proceeded to investigate the pattern of natural selection. For this, we estimated the d*N*/d*S* ratio by using a free-rate branch model as implemented in CodeML. We first eliminated recombining sites from the multiple alignment by implementing a GARD analysis from HyPhy 2.220 package (Kosakovsky et al. 2006). The free-rate model implemented in CodeML estimated a single d*N*/d*S* ratio for each one of the branches on the phylogentic tree.

We found a striking pattern where some internal branches (and a few external ones), show large d*N*/d*S* values ($\gg 1$) and most external branches show d*N*/d*S* < 1 (fig. 3*A* and *B*). This pattern indicates that purifying as well as positive selection play important roles during the evolution of *msg* genes. Based on this result, we envision a scenario where: 1) the evolution of a given variant driven by positive selection (pink or reddish branches); 2) is followed by expansion of the variant by gene duplication; and 3) stasis by purifying selection (blue branches). We suggest that the pattern of selection depicted in figure 3, is consistent with the proposal of Schmid-Siegert et al. (2017). Accordingly, the strategy of antigenic variation used by *P. jirovecii* consists in a "*continuous segregation of subpopulations with a new mixture of glycoproteins at the cell surface.*"

Msg proteins are proposed to be involved in the evasion of the host immune system by providing epitope diversity. This diversity should be located preferentially on the surface of Msg proteins. To investigate if this is the case, we identified sites predicted to have large surface accessibility to solvent and asked whether these sites tend to be positively selected (table 3). We found a significant statistical correlation (*P* value < 0.05) between both variables in families A1, A3(III), and D. This relation also holds for a sample of 11 sequences from family A1. The identity of these 11 sample sequences are shown in table S3. This result adds to the hypothesis of Msg proteins as central players in parasite–host interaction.

A Distinctive Rich G + C Pattern in Msg Genes Suggests Biased Gene Conversion

Meiotic recombination hot-spots in yeast are associated with high G + C regions (Gerton et al. 2000; Petes 2001; Mancera et al. 2008). Because recombination plays a role in the evolution of *msg* genes (Schmid-Siegert et al. 2017), we asked whether they also have a high G + C content. We found that *msg* genes have a higher G + C content than the rest of the genes in the genome (*P* value < 0.001, Wilcoxon test). To investigate how this high G + C content is distributed along the sequence of *msg* genes, we performed a meta-gene analysis. In a meta-gene analysis, each genetic sequence is divided into a fixed number of segments (50 in this case) and the G + C content is measured at segments with the same cardinality across all genes.

The result of the meta-gene analysis for *P. jirovecii* is shown in figure 4. We found that in the A1, D, and E families, there is an increase in the G + C content toward the 3′ end of the gene (fig. 4). We interpret this finding as evidence of biased gene conversion (Galtier et al. 2001). The fact that gene members from the A1 family can be expressed only by recombining at their 5′ end with the conserved recombination junction element of the upstream conserved element and the increase of G + C at this position; strongly support this hypothesis.

We also investigated the frequency of recombination on *msg* gene families. For this, we used a GARD (genetic algorithm for recombination detection) analysis from the HyPhy

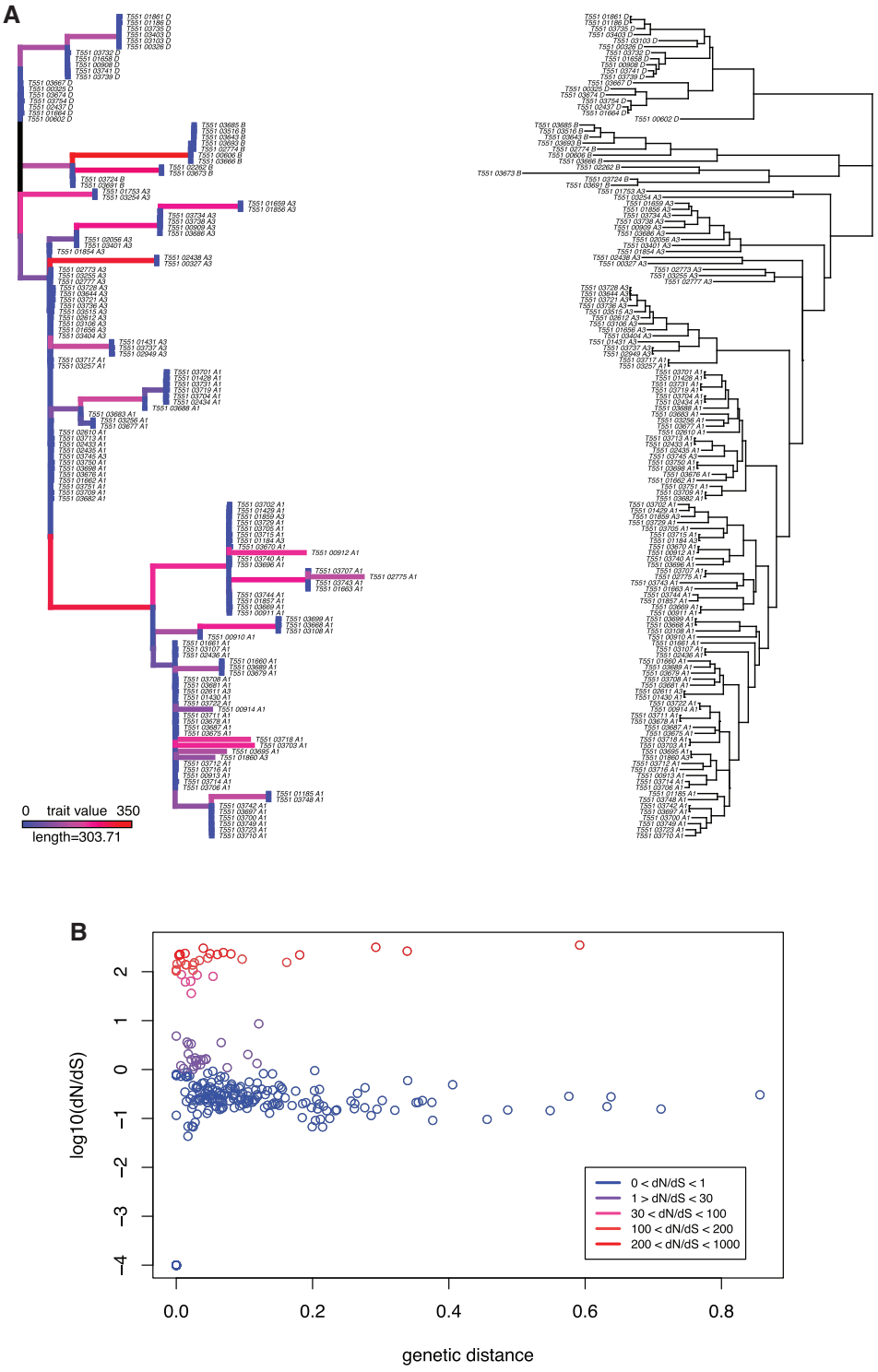

Fig. 3.—Episodic selection on Msg protein coding genes from *Pneumocystis jirovecii*. (*A*) Left tree shows variation in d*N*/d*S*. The length of the branches as well as the color is proportional to d*N*/d*S*. Right tree shows branch length proportional to genetic distance as estimated by Maximum-Likelihood (GTR + G+I model). Both trees have identical topologies. (*B*) Genetic distance versus d*N*/d*S*. While most branches evolve by negative selection (log10(d*N*/d*S*) < 0) a few branches evolve by strong positive selection (log10(d*N*/d*S*) > 1).

Table 3
Selection on Exposed Residues

Family	N of Seq	A[a]	B[b]	C[c]	D[d]	Odds	*P* Value
Msg A1	78	83	11	337	195	4.36	3.6e-07
Msg A1 (sample)	11	51	13	471	224	1.86	0.031
Msg A3(II)	14	22	5	437	207	2.08	0.097
Msg A3(III)	9	83	11	337	195	2.45	0.011
Msg B	11	28	9	233	121	1.61	0.151
Msg D	17	53	7	402	194	3.65	3.0e-04
Msg E	5	10	3	286	66	0.77	0.78

NOTE.—Association between codons under positive selection and residue surface accessibility to solvent among *msg* gene families. For categories A, B, D, and D, we included only free-gap sites. The *P* value is calculated with a one-sided Fisher's exact test.

[a]Codons under positive selection coding for residues showing surface accessibility >0.2 (i.e., exposed).

[b]Codons under positive selection coding for residues showing surface accessibility under 0.2 (i.e., buried).

[c]Codons not under positive selection coding for residues showing surface accessibility >0.2 (i.e., exposed).

[d]Codons not under positive selection coding for residues showing surface accessibility <0.2 (i.e., buried).

2.220 package (Kosakovsky et al. 2006). For each one of the variable sites, GARD divides the multiple alignment in two halves. Then, the algorithm inquires if the evolution of the sequences in the multiple alignment is better described by one or two trees. If it is better described by two trees, a recombination point is inferred. As shown in table 4 and supplementary figure S5, Supplementary Material online, we found evidence of recombination in families A1, A3(II), A3(III), and D. We did not find evidence of recombination in families B and E. However, detected recombination points are not preferentially located near the 3′ of *msg* genes (supplementary fig. S5, Supplementary Material online).

Msg as a Novel Gene

Msg genes are specific to the *Pneumocystis* genus. BLAST searches of Msg proteins against the nonredundant Genbank protein database finds no homologs outside *Pneumocystis* (e-value threshold < 0.1 and filtering out low complexity regions). A search for Msg protein domains in the Pfam database (Finn et al. 2016) reveals only one similar sequence in the arthropod *Daphnia pulex*. However, this similarity could be due to convergence. Therefore, the most parsimonious explanation is that Msg proteins originated concomitantly with *Pneumocystis* and its peculiar stenoxenism.

Discussion

Footprints of Positive Selection in Pneumocystis Genome Evolution

To evade the immune system of the host, pathogens evolve mechanisms to generate antigenic diversity; one such mechanism is recombination by gene conversion (Palmer and

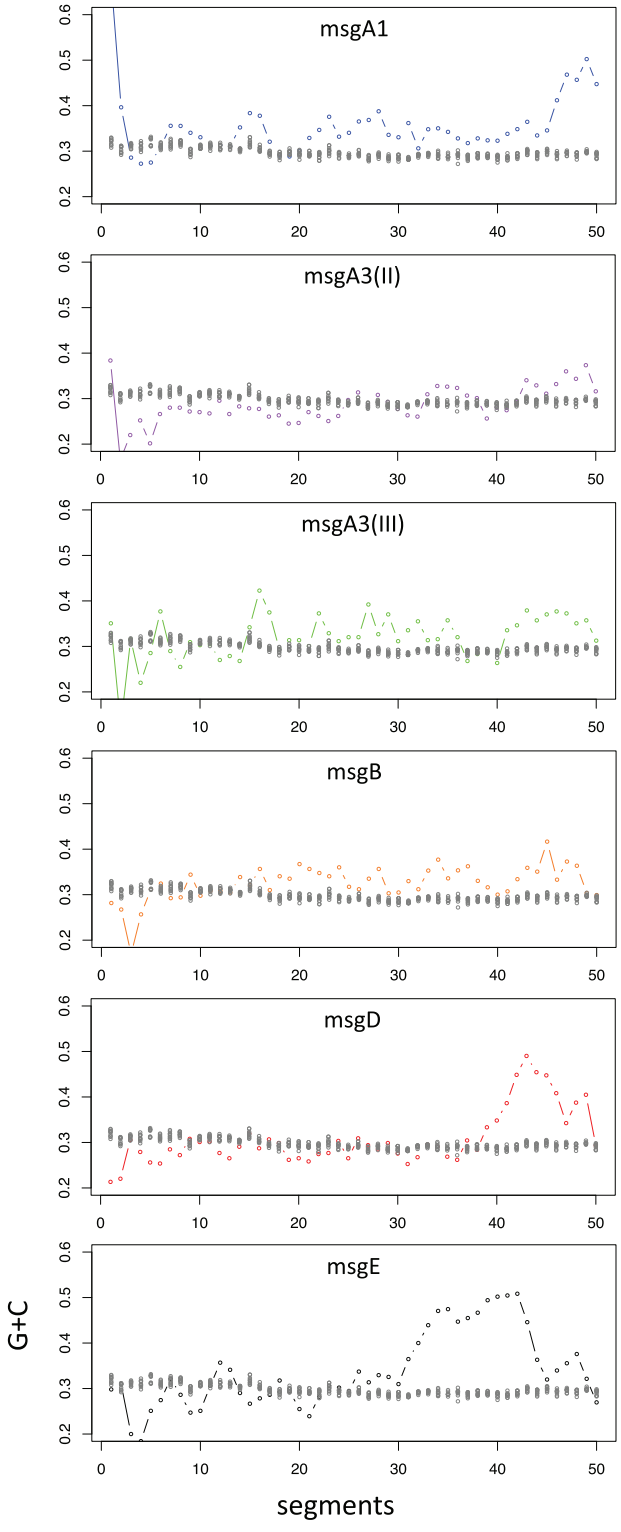

FIG. 4.—*msg* protein coding genes show a distinctive high G + C pattern in *Pneumocystis jirovecii*. Each gene from *P. jirovecii* was divided in 50 segments; and the G + C content of each segment was measured. In gray, we show the G + C content of 10 random samples of genes from *P. jirovecii*. Each random sample consists of 100 genes. The number of *msg* genes from families A1, A3(II), A3(III), B, D, and E is: 78, 9, 14, 11, 17, and 5, respectively.

Table 4
Recombination Events on *msg* Gene Families

Family[a]	Family[b]	*N* of Seq	Recombination Events
Msg A1	I	78	9
Msg A1 (sample)	I	11	4
Msg A3(II)	II	14	4
Msg A3(III)	III	9	3
Msg B	IV	11	0
Msg D	V	17	2
Msg E	VI	5	0

NOTE.—We show the result of applying GARD (a genetic algorithm for recombination detection) from HyPhy 2.220 package (Kosakovsky et al. 2006) on *msg* gene families. Only recombination events showing a P value < 0.01 are reported.

[a]Nomenclature by Ma et al. (2016).

[b]Nomenclature by Schmid-Siegert et al. (2017).

Brayton 2007; Deitsch et al. 2009; Vink et al. 2012). Gene conversion is the mechanism by which one DNA sequence, or more precisely a segment of a DNA sequence, replaces a homologous sequence during recombination. Well known cases of eukaryotes using this mechanism to generate antigenic diversity include: 1) the variant erythrocyte surface antigens (VESA) in *Babesia* spp. (Jackson et al. 2014); 2) the *var* genes in *Plasmodium falciparum* (Kyes et al. 2007); 3) the genes coding for *trans*-sialidases (TcTC) in *Trypanosoma cruzi*, which are also glycosylphosphatidylinositol-anchored proteins (Weatherly et al. 2016); and 4) the variant surface glycoproteins (VSG) in *T. brucei* (Hall et al. 2013). The results shown here, add to the hypothesis that *msg* genes are used by *Pneumocystis* to generate antigenic diversity by gene conversion (Keely et al. 2005; Stringer 2007; Kutty et al. 2008; Keely and Stringer 2009; Vink et al. 2012; Schmid-Siegert et al. 2017).

The pattern of recombination in *msg* genes has been studied recently (Schmid-Siegert et al. 2017). Accordingly, recombination is more frequent in families I, II, III, and IV than in families V and VI. These correspond to families A1, A3(II), A3(III), B, D, and E, respectively (table 4). Although we used a different method, our analysis of recombination is generally consistent with that reported by Schmid-Siegert et al. (2017). Accordingly, we detect from three to nine recombination events in families I, II, III; two recombination events in family V and none in VI. However, differing from Schmid-Siegert et al. (2017) that detected two recombination events in family IV, we detected none.

The pattern that emerges is that recombination is more common in families I, II, and III. However, we do not discard the existence of recombination in the other families. In the first place, Schmid-Siegert et al. (2017) reports several putative recombination events in families V and VI by using a more sensitive method. In the second place, the increase of G + C content toward the 3′ end of families V and VI (D and E in fig. 4) suggest biased gene conversion by recombination. Unfortunately, the distribution of the recombination events along *msg* genes does not support the above hypothesis (supplementary fig. S5, Supplementary Material online). Only in the case of the sample of 11 sequences from family A1 there seems to be a higher frequency of recombination events toward the 3′ of *msg* genes. A more in depth analysis is needed to clarify this discrepancy.

The *msg* genes belonging to different families show important differences, implying different functionalities (Schmid-Siegert et al. 2017). For instance, only A1 members have the recombination junction element at the beginning of their exon, suggesting that they can only be expressed by recombining with the recombination junction element of the upstream conserved element, which is present at a single copy in the genome. However this is not the case of the other *msg* families which present presumptive TATA boxes, the ATG initiation codon and a Cap signal at their sites of initiation of transcription, suggesting that they can be expressed independently of one another. Another important difference is that all families, except members from family IV, show a GPI anchor signal at its C terminus. This suggests that members from this family are attached to the cell wall by another mechanism or are secreted to the environment. Regarding their location along the subtelomeres, members from family A1 are located proximal to the telomere while members from family VI are located in a distal position. Here, we also show that members of different families are subject to different selective pressures. Residues predicted to be exposed are more likely to be evolving by positive selection in proteins coded by *msg* genes from families I, III, and V, suggesting that these ones are more often targets of the host immune system.

Here, we show that most genes involved in GPI biosynthesis have a significantly larger omega rate than other groups of functionally related genes. Although the average omega rate of these genes is <1.0, we argue that positive natural selection is the most likely cause of this increase. In principle, positive selection can increase the overall omega rate of a gene by acting only on a few codons, while maintaining most of the other codons under negative selection. In addition, there is a small probability (< 0.05) that the observed increase of the omega rate is due to chance. An alternative possibility, would be that the observed increase of the omega rate is due to relaxation of negative selection (Hughes 2007). However, the functional link between GPI biosynthesis and Msg proteins suggest that the higher omega rate of GPI biosynthetic genes is related to the role played by Msg proteins on parasite to host adaptation. This adaptation is guided by the selection imposed by the immune system of the host. Positive natural selection drives the evolution of these cellular subsystems that are functionally linked.

Overall, we suggest that genome evolution in *Pneumocystis* species can be described by the Red-Queen hypothesis (RQH). The RQH states that evolution is driven mostly by biotic interactions. Accordingly, the biotic environment of

any species is constantly evolving; therefore, species are under a continuous pressure to evolve adaptations to survive. One of the prediction of this dynamic process is that "*fast-evolving genes are commonly those at the interface of biotic interactions*" (Brockhurst et al. 2014). This prediction of the RQH has received empirical support by studying coevolving populations of *Pseudomonas fluorescens* SBW25 and the phage Phi2, its viral parasite (Paterson et al. 2010). Here, we suggest that the set of genes coding for GPI-biosynthesis and Msg proteins are at "*the interface of biotic interactions.*" The RQH also implies that evolutionary novelties in the parasite trigger evolutionary changes in the host. Consistent with this prediction, genes involved in the immune system show strong evidence of selection among humans and great apes (Cagan et al. 2016; Daub et al. 2017).

Conclusion

Here, we show that Msg protein coding genes evolve by positive episodic selection. We also show that most genes in the GPI-biosynthetic pathway show an increase of the omega rate in *P. jirovecii*, *P. murina*, and *P. carinii*. Overall, we suggest that this pattern is consistent with the Red-Queen hypothesis that predicts that genes involved in biotic interactions will show accelerated rates of molecular evolution.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Author's Contribution

A.M. and E.C. planed the study. L.D. and S.R. conducted the bioinformatics analysis. S.T., L.D., and A.C. performed the statistical analysis. L.D. wrote the first draft of the manuscript. All authors read and approved the manuscript.

Availability of data and material

All data to perform the analysis reported here is available upon request.

Acknowledgments

L.D. wishes to thank Eugenia Flores and Ana Fayos for support provided. This project has received funding from the Marie Curie International Research Staff Exchange Scheme within the 7th European Community Framework Program under grant agreement No 612583-DEANN. Part of this work was done during an internship of L.D. as invited professor at the Universidad de Valencia. Support from CONACYT (grant 454938) is gratefully acknowledged. This work was supported by grants to A.M. from the Spanish Ministry of Science and Competitivity (projects SAF 2012-31187, SAF2013-49788-EXP, SAF2015-65878-R), Carlos III Institute of Health (projects PIE14/00045, AC 15/00022 and AC15/00042), Generalitat Valenciana (project PrometeoII/2014/065) and cofinanced by FEDER.

Literature Cited

Aliouat-Denis CM, et al. 2008. *Pneumocystis* species, co-evolution and pathogenic power. Infect Genet Evol. 8(5):708–726.
Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25(1):25–29.
Brockhurst MA, et al. 2014. Running with the Red Queen: the role of biotic conflicts in evolution. Proc Biol Sci. 281(1797):20141382.
Brown GD, et al. 2012. Hidden killers: human fungal infections. Sci Transl Med. 4(165):165rv13.
Cagan A, et al. 2016. Natural selection in the Great Apes. Mol Biol Evol. 33(12):3268–3283.
Carinii A. 1910. Formas de eschizogonia do *Trypanozoma lewisi*. Commun Soc Med Sao Paolo 16:204.
Catherinot E, et al. 2010. *Pneumocystis jirovecii* Pneumonia. Infect Dis Clin N Am. 24(1):107–138.
Chagas C. 1909. Nova tripanozomiata humana. Mem Inst Oswaldo Cruz 1(2):159–218.
Cissé OH, Pagni M, Hauser PM. 2014. Comparative genomics suggests that the human pathogenic fungus *Pneumocystis jirovecii* acquired obligate biotrophy through gene loss. Genome Biol Evol. 6(8):1938–1948.
Cushion MT, et al. 2007. Transcriptome of *Pneumocystis carinii* during fulminate infection: carbohydrate metabolism and the concept of a compatible parasite. PLoS One 2(5):e423.
Daub JT, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M. 2017. Detection of pathways affected by positive selection in primate lineages ancestral to humans. Mol Biol Evol. 34(6):1391–1402.
Deitsch KW, Lukehart SA, Stringer JR. 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. Nat Rev Microbiol. 7(7):493–503.
Delanoë P, Delanoë M. 1912. Sur les rapports des kystes de *Carinii* du poumon des rats avec le trypanosoma *Lewisii*. C R Acad Sci (Paris) 155:658–660.
Demanche C, et al. 2001. Phylogeny of *Pneumocystis carinii* from 18 primate species confirms host specificity and suggests coevolution. J Clin Microbiol. 39(6):2126–2133.
Derouiche S, et al. 2009. *Pneumocystis diversity* as a phylogeographic tool. Mem Inst Oswaldo Cruz 104(1):112–117.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.
Edman JC, et al. 1988. Ribosomal RNA sequence shows *Pneumocystis carinii* to be a member of the fungi. Nature 334(6182):519–522.
Ferguson MAJ, Kinoshita T, Hart GW. 2009. Glycosylphosphatidylinositol anchors. In: Varki A, Cummings RD, Esko JD, et al. editors. Essentials of glycobiology. 2nd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.
Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.
Fujita M, Kinoshita T. 2010. Structural remodeling of GPI anchors during biosynthesis and after attachment to proteins. FEBS Lett. 584(9):1670–1677.
Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159(2):907–911.
Gerton JL, et al. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 97(21):11383–11390.

Gigliotti F, Harmsen AG, Haidaris CG, Haidaris PJ. 1993. *Pneumocystis carinii* is not universally transmissible between mammalian species. Infect Immun. 61:2886–2890.
Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.
Hall JP, Wang H, Barry JD. 2013. Mosaic VSGs and the scale of *Trypanosoma brucei* antigenic variation. PLoS Pathog. 9(7):e1003502.
Hauser PM. 2014. Genomic insights into the fungal pathogens of the genus *Pneumocystis*: obligate biotrophs of humans and other mammals. PLoS Pathog. 10(11):e1004425.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33(6):1635–1638.
Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity (Edinb) 99(4):364–373.
Jackson AP, et al. 2014. The evolutionary dynamics of variant antigen genes in *Babesia* reveal a history of genomic innovation underlying host-parasite interaction. Nucleic Acids Res. 42(11):7113–7131.
Keely SP, et al. 2005. Gene arrays at *Pneumocystis carinii* telomeres. Genetics 170(4):1589–1600.
Keely SP, Stringer JR. 2009. Complexity of the MSG gene family of *Pneumocystis carinii*. BMC Genomics 10(1):367.
Kosakovsky PSL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22(24):3096–3098.
Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.
Kutty G, England KJ, Kovacs JA. 2013. Expression of *Pneumocystis jirovecii* major surface glycoprotein in *Saccharomyces cerevisiae*. J Infect Dis. 208(1):170–179.
Kutty G, Maldarelli F, Achaz G, Kovacs JA. 2008. Variation in the major surface glycoprotein genes in *Pneumocystis jirovecii*. J Infect Dis. 198(5):741–749.
Kutty G, Shroff R, Kovacs JA. 2013. Characterization of *Pneumocystis* major surface glycoprotein gene (msg) promoter activity in *Saccharomyces cerevisiae*. Eukaryot Cell 12(10):1349–1355.
Kyes SA, Kraemer SM, Smith JD. 2007. Antigenic variation in *Plasmodium falciparum*: gene organization and regulation of the var multigene family. Eukaryot Cell 6(9):1511–1520.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.
Liang M, et al. 2016. Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. BioData Min. 9(1):13.
Ma L, et al. 2016. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nat Commun. 7:10740.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454(7203):479–485.
Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8(7):e1002764.
Palmer GH, Brayton KA. 2007. Gene conversion is a convergent strategy for pathogen antigenic variation. Trends Parasitol. 23(9):408–413.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of Phylogenetics and Evolution in R language. Bioinformatics 20(2):289–290.
Paterson S, et al. 2010. Antagonistic coevolution accelerates molecular evolution. Nature 464(7286):275–278.
Petersen B, Petersen N, Andersen P, Nielsen M, Lundegaard C. 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol. 9(1):51.
Petes TD. 2001. Meiotic recombination hot spots and cold spots. Nat Rev Genet. 2(5):360–369.
Pittet M, Conzelmann A. 2007. Biosynthesis and function of GPI proteins in the yeast *Saccharomyces cerevisiae*. Biochim Biophys Acta 1771(3):405–420.
Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676–679.
Schmid-Siegert E, et al. 2017. Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. MBio 8(6):e01470-17.
Serra F, Arbiza L, Dopazo J, Dopazo H. 2011. Natural selection on functional modules, a genome-wide analysis. PLoS Comput Biol. 7(3):e1001093.
Stringer JR. 2007. Antigenic variation in *Pneumocystis*. J Eukaryot Microbiol. 54(1):8–13.
Stringer SL, Stringer JR, Blase MA, Walzer PD, Cushion MT. 1989. *Pneumocystis carinii*: sequence from ribosomal RNA implies a close relationship with fungi. Exp Parasitol. 68(4):450–461.
Subramanian A, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 102(43):15545–15550.
Thomas CF Jr, Limper AH. 2007. Current insights into the biology and pathogenesis of *Pneumocystis* pneumonia. Nat Rev Microbiol. 5(4):298–308.
Vink C, Rudenko G, Seifert HS. 2012. Microbial antigenic variation mediated by homologous DNA recombination. FEMS Microbiol Rev. 36(5):917–948.
Vinuesa P, Contreras-Moreira B. 2015. Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids. Methods Mol Biol. 1231:203–232.
Weatherly DB, Peng D, Tarleton RL. 2016. Recombination-driven generation of the largest pathogen repository of antigen variants in the protozoan *Trypanosoma cruzi*. BMC Genomics 17(1):729.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Associate editor: Laurence Hurst