Original Research

Genes Coding for GPI Biosynthesis in *Pneumocystis* Experienced Relaxed Selection: A Follow-up Study

Luis Delaye [*]

Departamento de Ingeniería Genética, CINVESTAV Irapuato, Km. 9.6 Libramiento Norte Carretera Irapuato-León, 36821 Irapuato, Guanajuato, Mexico; E-Mail: luis.delaye@cinvestav.mx

[*] **Correspondence:** Luis Delaye; E-Mail: luis.delaye@cinvestav.mx

Abstract

***Background*:** Phylogenetic analysis shows that *Pneumocystis* have coevolved with their mammalian hosts for millions of years. As a result, infection by *Pneumocystis* is species specific (i.e. a given *Pneumocystis* species can infect only a single species of host). This specificity suggests the existence of a Red-Queen dynamic between *Pneumocystis* and its host. Evidence of this dynamic is provided by the molecular diversity and the elevated rate of non-synonymous versus synonymous (dN/dS) codon substitutions among *Pneumocystis* major surface glycoproteins (MSG). It was previously suggested that this evolutionary dynamic also affected the evolution of genes coding for the biosynthesis of glycosylphosphatidylinositol (GPI). This was based on the fact that MSG are targeted to the cell surface by GPI, and GPI biosynthetic genes showed a statistically significant increase in their dN/dS rate when compared to other groups of genes in *Pneumocystis* genome; however, the average rate of dN/dS of GPI biosynthetic genes remained < 1.0. Here we test whether GPI biosynthetic genes increased their omega rate by positive or relaxed natural selection.

***Methods*:** Genes coding for the biosynthesis of glycosylphosphatidylinositol (GPI) in *Pneumocystis* and closely related fungi were retrieved from Genbank. Each one of the

retrieved gene families were analysed with RELAX software to discriminate between positive or relaxed selection. In addition, we searched for evidence of episodic selection in these genes by using aBSREL, and analysed their dN/dS rate by using the branch model as implemented in CodeML.

Results: Molecular phylogenetic analyses showed that the set of genes coding for the biosynthesis of GPI experienced relaxed selection *Pneumocystis*. In addition, just 4 of the 19 gene families studied here were able to detect evidence of episodic selection. Finally, the results of the CodeML branch model are consistent with evolution by relaxed selection.

Conclusions: The relaxation of selection in GPI biosynthetic genes is not consistent with the previous proposal suggesting that positive selection increased the dN/dS rate of these genes due to a Red-Queen dynamic between *Pneumocystis* and its host. Despite the statistical significance previously reported, the increase in the dN/dS rate in this group of genes still requires biological interpretation. An intriguing possibility is that the relaxation of selection in GPI biosynthetic genes is related to depletion of genes coding for transporters in *Pneumocystis*.

1. Introduction

In 1974, Van Valen published an article to suggest that "*over long intervals, the directional selection intensity on any protein is proportional (or at least positively related) to that of the phenotype as a whole*" [1]. In this article, Van Valen was trying to give an explanation for the constancy of the rate of molecular evolution described by Kimura [2]. As predicted by the Red Queen Hypothesis (RQH), the biotic environment of any species degrades at a proximately constant rate [3]. For any species to survive, it has to adapt to this ever-changing environment. Because the rate of environmental change is approximately constant, the rate of phenotypic adaptive change will be approximately constant, too (according to Van Valen). From this, it follows that genes coding for phenotypic traits involved in biotic interactions will show this constancy in their rate of change (i.e. they will follow the rate of environmental change).

This hypothesis was tested a few years ago. Paterson and co-workers [4] compared the accumulation of genetic change in co-evolving versus evolving phages. They found that phages that were allowed to coevolve with their bacterial host accumulated more genetic changes than those phages that infected a bacterial population that was not allowed to coevolve. Interestingly, they found that the genes that accumulated most genetic changes where those coding for proteins functioning at the "*interface of biotic interactions.*"

Recently, it was suggested that genes coding for the biosynthesis of glycosylphosphatidylinositol (GPI) in the fungi *Pneumocystis* showed an increase of non-synonymous (dN) over synonymous substitutions (dS) due to positive natural selection [5]. GPI is used by the cell to target proteins to the periplasmic membrane and the cell wall. In *Pneumocystis*,

major surface glycoproteins (MSG) are targeted to the cell surface by GPI. MSG are in turn used by *Pneumocystis* to attach to the alveoli and escape the immune system of the host [6].

Pneumocystis is an ascomycete that lives in the lungs of mammals. In humans, it causes a severe pneumonia in immunocompromised patients. One of the peculiarities of this fungi is that each species of *Pneumocystis* is capable of infecting a single species of host [7]. This property of *Pneumocystis* is called stenoxenism and its biological basis is still unknown. As of today, the genomes of three different *Pneumocystis* species have been sequenced [8].

The increase of dN over dS in the genes coding for the biosynthesis of GPI was interpreted in the light of the RQH by Delaye L, et al [5]. Accordingly, there is a Red-Queen dynamic between *Pneumocystis* and the host immune system. This Red-Queen dynamic promotes the evolution of molecular diversity among MSG by positive natural selection. Because MSG are carried to the cell surface by GPI (i.e. they are functionally linked), [5] hypothesized that positive selection also drove the evolution of genes coding for GPI biosynthesis. However, the average dN/dS of evolution of GPI biosynthetic genes is < 1 and an increase of dN over dS can be due to positive as well as relaxed negative selection. Here we test this hypothesis by asking whether the increase of dN over dS of the genes coding for GPI biosynthesis is due to positive or relaxed selection.

2. Materials and Methods

2.1 Phylogenetic Analysis

2.1.1 Search for Homologs

To identify close homologs to each one of the genes participating in the biosynthesis of GPI, we used BLASTp as provided by NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi). We first compared each one of the 16 protein coding genes of the GPI biosynthetic pathway from *P. jirovecii* (Table 1) against the non-redundant (nr) protein sequence database. For each one of the BLASTp searches, we retrieved all hits showing an E-value < 0.01. We modified the parameters in order to allow BLASTp to report a maximum of 500 hits per sequence search (the default value is 100 hits). Next, by using the bitscores, we built histograms for each one of the BLASTp searches. The distribution of bitscores often showed a bimodal distribution. This allowed us to easily identify, in most cases, the set of sequences most closely related to those of *P. jirovecii* (Figure S1).

2.1.2 Multiple Sequence Alignment

For each set of homologs identified in the previous step, we selected only those sequences belonging to species shared by all 16 protein families. This makes the phylogenetic and natural selection analysis (see below) more comparable between protein families. Each protein family was aligned with MAFFT (https://mafft.cbrc.jp/alignment/software/) [9] and then the corresponding gene sequences were codon aligned with pal2nal (http://www.bork.embl.de/pal2nal/) [10]. The multiple sequence alignment was visually inspected in Jalview (http://www.jalview.org) [11] and those sequences that were too short and introduced an excessive number of gaps were manually removed. Having done this, we ended up with gene families between 53 and 67 sequences.

2.1.3 Phylogenetic Inference

We inferred phylogenies for each one of the 16 gene families by using PhyML (http://www.atgc-montpellier.fr/phyml/) [12] with GTR+G+I as the model of sequence evolution and the following parameters inferred from the data: Frequency of bases, proportion of invariable sites, gamma parameter, and transition versus transversion ratio. We also used four substitution rate categories and we searched for the best tree topologies by using both NNI and SPR. Finally, unrooted trees were prepared by using the ape module in R [13]. Multiple sequence alignments and phylogenetic trees are provided as supplementary material (Figure S2 and supplementary material).

2.2 Selection Analysis

2.2.1 Analysis of Relaxed Selection

To study whether genes coding for the biosynthesis of GPI in *Pneumocystis* experienced relaxation or intensification of positive natural selection, we used the software RELAX [14]. Briefly, RELAX introduces a k parameter (known as the selection intensity parameter) to determine the intensity of selection along a given set of branches (the test branches) in the phylogenetic tree. A significant ($p < 0.05$) $k > 1$ indicates that selection strength has intensified, and a significant $k < 1$ indicates that selection has relaxed. We applied RELAX to each one of the 16 gene families studied here. For this, we used the multiple sequence alignments and the phylogenetic trees described in section 2.1.

2.2.2 Analysis of Episodic Selection

To analyse if episodic selection has occurred along the branches leading to *Pneumocystis* in each of the 16 gene families studied here, we used aBSREL [15]. This method allows us to test if a proportion of sites has evolved by positive natural selection in a branch of interest in a phylogenetic tree. As such, aBSREL is useful to detect episodic selection. For this, we used the multiple sequence alignments and the phylogenetic trees described in section 2.1.

2.2.3 Analysis of Selection along the Lineages Leading to *Pneumocystis*

In addition, we analysed whether selection is different in the lineages leading to *Pneumocystis* as compared to the rest of the branches in the phylogenetic trees. For this we used the branch model as implemented in CodeML from the PAML package [16]. For this, we used the multiple sequence alignments and the phylogenetic trees described in section 2.1.

3. Results

3.1 Genes Coding for GPI Biosynthesis in Pneumocystis *Experienced Relaxed Selection*

The rate of non-synonymous substitutions over synonymous substitutions (dN/dS) is used to study evolution by natural selection on protein coding genes. Positive selection, neutral evolution, and negative selection is inferred if dN/dS is > 1, = 1, or < 1, respectively [17]. As mentioned above,

the increase of dN over dS can be due to positive selection as well as by relaxation of purifying (or negative) selection.

To test whether the increase of dN over dS on the genes coding for GPI biosynthesis in *Pneumocystis* is due to positive or relaxed selection we used RELAX [14]. This method is useful for identifying changes in the stringency of selection. In particular, RELAX allowed us to identify whether the strength of selection has intensified or relaxed along the branches leading to the three *Pneumocystis* species. The RELAX algorithm reports the strength of selection by a k parameter. If $k < 1$ then relaxation of negative selection is inferred while $k > 1$ indicates an intensification of positive selection on the specified branches. The RELAX analysis provided the main results of this study.

In addition to the RELAX analysis, we applied two different tests to study the pattern of dN/dS on all genes coding for the biosynthesis of GPI in the three *Pneumocystis* species. These two tests were used to better understand the role of natural selection on the evolution of these genes. The first one of these accessory tests is aBSREL (an adaptive Branch-Site REL test for episodic diversification) [15]. This test was used to detect if positive natural selection occurred along the branches leading to the three *Pneumocystis* species by implementing a branch-site model. aBSREL is a very sensitive method to detect episodic selection (i.e. whether selection occurred only in a set of sites in the multiple alignment and in a set of branches of the phylogenetic tree). aBSREL is complementary to RELAX because RELAX is not suitable for explicitly testing for positive selection.

The second of these accessory tests was the branch model as implemented in CodeML [16]. This test allowed us to estimate gene wide selection along the branches leading to the three *Pneumocystis* species. Differing from aBSREL, the estimates of selection obtained by the branch model in CodeML are an average of the dN/dS along the whole gene on the specified branches. In principle, the result of CodeML should be consistent with that of RELAX.

As shown in Table 1, all genes participating in the biosynthesis of GPI in *Pneumocystis* experienced relaxed selection when compared to their homologs in closely related species. In all cases the k parameter is < 1 and the test is statistically significant (p-value < 0.05). This is consistent with the estimated dN/dS when using the branch model in CodeML. In this last case, the lineages leading to the three *Pneumocystis* species have significantly larger (p-value < 0.05) dN/dS values than the rest of the branches in the phylogenetic trees in all but one of the genes (Figure 1). This indicates that purifying selection has been stronger in the lineages of species other than *Pneumocystis*. Finally, positive episodic selection was detected in only four genes (GAB1, GPI11, GPI17, and THO) by using the highly sensitive method aBSREL.

The above results are not consistent with the hypothesis that the increase in dN/dS in genes coding for the biosynthesis of GIP (when compared to the dN/dS of other genes in the genome of *Pneumocystis spp.*) is due to positive natural selection [5].

Table 1 Selection analysis of genes coding for the biosynthesis of GPI in *Pneumocystis*. The results of three analyses are shown. RELAX tests whether selection has intensified or relaxed. aBSREL tests for episodic selection. Branch model in CodeML tests whether the dN/dS rate in the branches leading to *Pneumocystis* is different from the rest of the branches in the tree. Statistical significance: (*) p-value < 0.05; (**) p-value < 0.01. After the name of the gene, we show in parenthesis the number of homologs used to make the phylogenetic trees.

	Selection (RELAX)	Episodic selection (aBSREL)		Branch model (CodeML)		
Gene (N)	*k*	Selection in *Pneumocystis*	Lineage under positive episodic selection	dN/dS *Pneumocystis*	dN/dS other branches	*Statistical significance*
GAA1 (64)	0.41 (**)	No		0.39	0.06	**
GAB1 (58)	0.20 (**)	Yes (*)	*P. jirovecii*	0.37	0.06	**
GPI1 (59)	0.60 (**)	No		0.20	0.04	**
GPI10 (65)	0.56 (*)	No		0.56	0.11	**
GPI11 (53)	0.66 (**)	Yes (**)	*P. carinii*	0.61	0.09	**
GPI13 (67)	0.35 (**)	No		0.34	0.06	**
GPI14 (62)	0.49 (**)	No		0.36	0.07	**
GPI16 (64)	0.16 (**)	No		0.36	0.06	**
GPI17 (65)	0.43 (**)	Yes (*)	(*P. murina, P. carinii*)	0.30	0.05	**
GPI18 (59)	0.51 (**)	No		0.34	0.09	**
GPI3 (51)	0.69 (**)	No		0.05	0.02	**
GWT1 (60)	0.18 (**)	No		0.48	0.06	**
MCD4 (54)	0.36 (**)	No		0.26	0.05	**
PBN1 (62)	0.22 (**)	No		0.59	0.13	**
PER1 (62)	0.61 (**)	No		0.29	0.07	**
THO (67)	0.47 (**)	Yes (*)	*P. jirovecii*	0.35	0.40	**
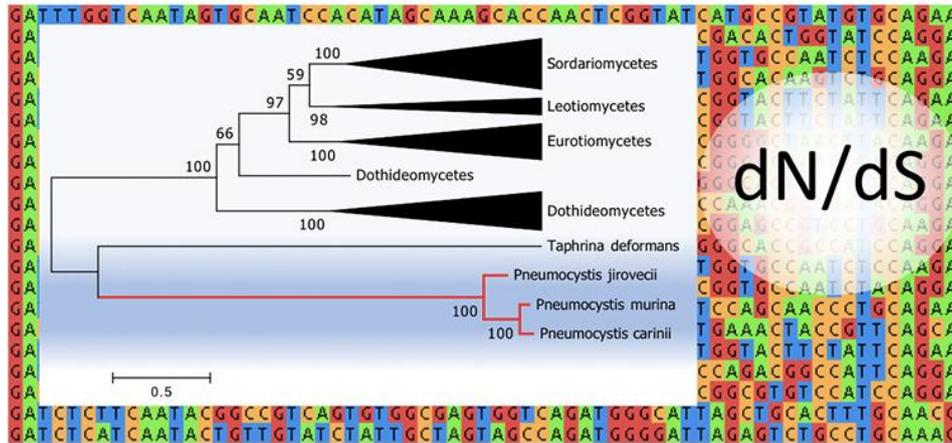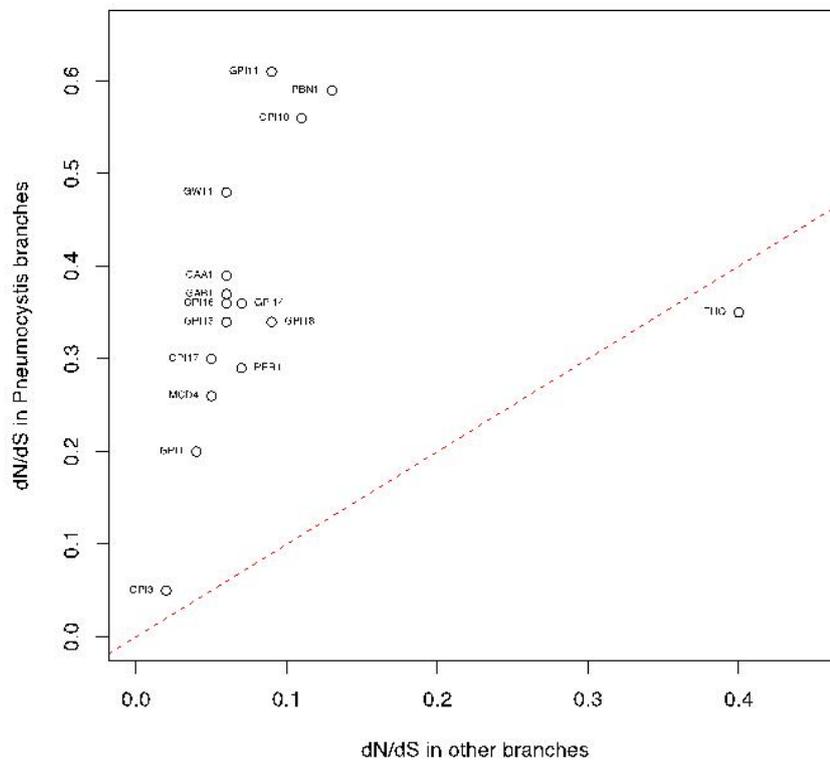
Figure 1 GPI biosynthetic genes in *Pneumocystis* show an elevated dN/dS compared to homologs in closely related fungi. A) The phylogeny of GAA1 is shown as a reference. The CodeML branch model compares whether the dN/dS rate is the same between the red (*Pneumocystis*) and the rest of the branches. B) In all cases, *Pneumocystis* genes show an elevated dN/dS compared to their homologs in other fungi, with the exception of the gene THO. The red line indicates the cases where the rate of evolution (dN/dS) is the same in both black and red branches.

4. Discussion

Red-Queen dynamics are expected to be stronger between closely dependent species, as in the case of parasite-host interactions. The stenoxenism shown by *Pneumocystis* suggests that it should be subject to a Red-Queen dynamic. The large number of sites showing dN/dS > 1 in their MSG suggests that this is the case. However, RELAX analysis shows that the acceleration of the rate of dN/dS among genes coding for GPI biosynthesis is not due to positive natural selection. On the contrary, the acceleration is due to relaxation of natural selection.

The analysis published by Delaye L, et al [5] indicates that this particular group of genes showed an increase of the omega rate (dN/dS) as compared to other groups of genes. Here we show that this increase is due to a relaxation of natural selection. Is there any rational to explain this increase in the omega rate? One possibility is that despite the statistical significance of the results (*p*-value < 0.01 and Q < 0.05) provided by Delaye L, et al [5], the increase in the omega rate of GPI biosynthetic genes is not significant from the biological point of view. However, this is unlikely because the same group of genes (coding for the biosynthesis of GPI) showed an increase in their dN/dS in the three analysed genomes (those of *P. jirovecii, P. carinii,* and *P. murina*). In addition, other categories of genes related to protein secretion and maturation showed an increase in their dN/dS in one or two of the genomes. For instance, in *P. jirovecii* the category of Glycosylation (GO: 0070085) and in *P. murina*, the categories of Protein glycosylation (GO: 0006486) and Golgi apparatus (GO: 0005794) showed an increase of their dN/dS [5]. Moreover, the increase in the evolutionary rate is not due to a local change in base composition in the genome, since GPI biosynthetic genes are scattered along the genome of the three *Pneumocystis* species (Figure S3 and Table S1).

Based on the above, it is more likely that the relative increase in the dN/dS rate of the genes coding for GPI biosynthesis is a real phenomenon; however, the cause is not positive natural selection. Then, the question that remains is why purifying selection has been less strong in GPI biosynthetic genes in the three species despite its relevance in the protein secretion apparatus and the biology of *Pneumocystis*? One possibility is that most genes in *Pneumocystis* experienced relaxed selection. After all, *Pneumocystis* has a reduced genome of approximately 7.5 to 8.4 Mb and between 3,675 to 3,812 genes [8]. As a comparison, *Taphrina deformans* (the closest relative to *Pneumocystis*), has a genome size of 13.4 Mb and about 4,661 genes [8]. And species undergoing genome reduction due to symbiosis (whether parasitic or mutualist) are known to experience acceleration in their rate of evolution [18, 19]. Extending the analysis presented here to all genes in the genome of *Pneumocystis* would reveal whether relaxation of selection is a widespread phenomenon in these endosymbiotic fungi.

Species living in similar environments often show character convergence [20]. The microsporidian *Encephalitozoon cuniculi* is an obligate intracellular parasite that infects many animal groups. Paralleling *Pneumocystis*, *E. cuniculi* also has a reduced genome [21]. Comparative genome analysis revealed that both species converged in losing similar gene categories [8]. In particular, the genome of these species showed extensive reduction in the following gene categories: transporters; transcription factors; and enzymes including oxidoreductases, hydrolases, transferases, and coenzymes. One possibility is that the reduction in the number of transporters is related to the relaxation of selection of GPI biosynthetic genes in *Pneumocystis*. After all, *Pneumocystis* retain < 25 transporters while other ascomycetes have between 131 to 217. It would

be interesting to study whether GPI biosynthetic genes in *E. cuniculi* also experienced relaxation of selection. A comparative analysis between *Pneumocystis* and *E. cuniculi* could reveal which convergences exist in the pattern of natural selection at the genome level in intracellular obligate parasitic eukaryotes.

5. Conclusions

Here we show that the increase in the dN/dS rate of genes coding for GPI biosynthesis, reported by Delaye L, et al [5], is due to relaxation of natural selection. This relaxation of natural selection is not consistent with the RQH. Although the increase in the dN/dS rate reported by Delaye L, et al [5] is statistically significant, it is not clear why natural selection has been less strong in this group of genes. An intriguing possibility is that the relaxation of selection in GPI biosynthetic genes is related to depletion of genes coding for transporters in *Pneumocystis*.

Acknowledgments

Additional Materials

The following additional materials are uploaded at the page of this paper.
1. Figure S1: For each one of the 16 enzymes participating in the biosynthesis of GPI in *P. jirovecii* we show the distribution of bitscores (from BLASTp) for all homologous sequences showing an E-value < 0.01.
2. Figure S2: Maximum likelihood phylogenetic trees of genes coding for GPI biosynthesis in Pneumocystis and related species.
3. Figure S3: Distribution of genes coding for the biosynthesis of GPI in *P. carinii, P. jirovecii* and *P. murina*.
4. Table S1: Coordinates of genes coding for the biosynthesis of GPI in *P. carinii, P. jirovecii* and *P. murina*.
5. Multiple sequence alignments and phylogenetic trees.

Author Contributions

LD planed and conducted the bioinformatic analysis and wrote and approve the manuscript.

Funding

Competing Interests

The author has declared that no competing interests exist.

References

1. Van Valen L. Molecular evolution as predicted by natural selection. J Mol Evol. 1974; 3: 89-101.
2. Kimura M. Evolutionary rate at the molecular level. Nature. 1968; 217: 624-626.
3. Van Valen, L. A new evolutionary law. Evol Theo. 1973; 1: 1-30.
4. Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, et al. Antagonistic coevolution accelerates molecular evolution. Nature. 2010; 464: 275-278.
5. Delaye L, Ruiz-Ruiz S, Calderon E, Tarazona S, Conesa A, Moya A. Evidence of the red-queen hypothesis from accelerated rates of evolution of genes involved in biotic interactions in *Pneumocystis*. Genome Biol Evol. 2018; 10: 1596-1606.
6. Hauser PM. Is the unique camouflage strategy of *Pneumocystis* associated with its particular niche within host lungs? PLoS Pathog. 2019; 15: e1007480.
7. Demanche C, Berthelemy M, Petit T, Polack B, Wakefield AE, Dei-Cas E, et al. Phylogeny of Pneumocystis carinii from 18 primate species confirms host specificity and suggests coevolution. J Clin Microbiol. 2001; 39: 2126-2133.
8. Ma L, Chen Z, Huang da W, Kutty G, Ishihara M, Wang H, et al. Genome analysis of three Pneumocystis species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nat Commun. 2016; 7: 10740.
9. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program, Brief Bioinform. 2008; 9: 286-298.
10. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006; 34: W609-W612.
11. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--A multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009; 25: 1189-1191.
12. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 2009; 537: 113-137.
13. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. bioinformatics. 2004; 20: 289-290.
14. Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. RELAX: Detecting relaxed selection in a phylogenetic framework. Mol Biol Evol. 2015; 32: 820-832.
15. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Pond SLK. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015; 32: 1342-1353.
16. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24: 1586-1591.
17. Anisimova M, Liberles D. Detecting and understanding natural selection. Codon Evol. 2012: 73-96.
18. Bromham L, Cowman PF, Lanfear R. Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evol Biol. 2013; 13: 126.
19. Itoh T, Martin W, Nei M. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci USA. 2002; 99: 12944-12948.
20. Jonathan B. Losos. Improbable destinies: Fate, chance, and the future of evolution. First Ed. New York: Riverhead Books; 2017.

21. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature. 2001; 414: 450-453.

Enjoy *OBM Genetics* by:

1. Submitting a manuscript
2. Joining in volunteer reviewer bank
3. Joining Editorial Board
4. Guest editing a special issue

For more details, please visit:
http://www.lidsen.com/journals/genetics