Research

Author for correspondence:
Kevin P. Johnson
e-mail: kpjohnso@illinois.edu

Rates of genomic divergence in humans, chimpanzees and their lice

Kevin P. Johnson[1], Julie M. Allen[1], Brett P. Olds[2,4], Lawrence Mugisha[5,6,7], David L. Reed[8], Ken N. Paige[2] and Barry R. Pittendrigh[3]

[1]Illinois Natural History Survey, University of Illinois, 1816 South Oak Street, Champaign, IL 61820, USA
[2]Department of Animal Biology, and [3]Department of Entomology, University of Illinois, Urbana, IL 61801, USA
[4]Department of Biology Sciences, University of Notre Dame, Notre Dame, IN 46556, USA
[5]College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, Kampala, Uganda
[6]College of Veterinary Medicine, University of Minnesota, St. Paul, MN 55108, USA
[7]Conservation and Ecosystem Health Alliance, Kampala, Uganda
[8]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

The rate of DNA mutation and divergence is highly variable across the tree of life. However, the reasons underlying this variation are not well understood. Comparing the rates of genetic changes between hosts and parasite lineages that diverged at the same time is one way to begin to understand differences in genetic mutation and substitution rates. Such studies have indicated that the rate of genetic divergence in parasites is often faster than that of their hosts when comparing single genes. However, the variation in this relative rate of molecular evolution across different genes in the genome is unknown. We compared the rate of DNA sequence divergence between humans, chimpanzees and their ectoparasitic lice for 1534 protein-coding genes across their genomes. The rate of DNA substitution in these orthologous genes was on average 14 times faster for lice than for humans and chimpanzees. In addition, these rates were positively correlated across genes. Because this correlation only occurred for substitutions that changed the amino acid, this pattern is probably produced by similar functional constraints across the same genes in humans, chimpanzees and their ectoparasites.

1. Introduction

Understanding differences between species in the rate of molecular evolution is of considerable interest in the fields of evolution, molecular biology, population genetics and systematics. However, estimating this rate variation is often difficult because of uncertainties regarding the timing of diversification. Host and parasites provide a system in which the relative rates of molecular evolution can be directly estimated, provided some degree of congruence (codivergence) between host and parasite evolutionary trees exists [1,2]. Several such relative rate estimates have been made for a variety of host and parasite systems, and parasites are usually observed to evolve more rapidly than their hosts at the molecular level [2–9]. Some previous explanations for this phenomenon have been shorter generation times, population bottlenecks or relaxed selection for mutation repair in parasite lineages when compared with their hosts.

However, most of these studies compare only relative rates of genetic substitution for a single gene, typically mitochondrial genes because of the ease of sequencing homologous loci in host and parasite taxa. There has been no study of the variation in the relative rates of DNA substitution between hosts and parasites across the genome. The codivergence event between humans, chimpanzees and their lice (genus *Pediculus*) is well documented and provides an excellent calibration point for comparing rates of genetic divergence [10–12]. The genomes of humans (*Homo sapiens*), common chimpanzees (*Pan troglodytes*) and the human body louse (*Pediculus humanus*) have already been sequenced [13–15]. We sequenced across the genome of the chimpanzee louse (*Pediculus schaeffi*) using next-generation sequencing of a 500 bp paired-end library with the Illumina HiSeq2000 platform in a single lane and achieved approximately

120× coverage. Because these data were not collected for complete genome assembly as a goal, for the chimpanzee louse, we used reference-based assembly of individual nuclear encoded genes that are orthologues between humans, chimpanzees and the human body louse. These individual gene assemblies were then used to compare the relative rate of divergence across these genes between these ectoparasites and their hosts.

2. Material and methods

Lice were collected from chimpanzees (*Pan troglodytes schweinfurthii*) from Ngamba Island Chimpanzee Sanctuary during annual health checks. Total genomic DNA was extracted by grinding five adult female specimens of the chimpanzee louse (*P. schaeffi*) in 300 μl of saline EDTA with 5 μl of lysozyme, incubating at 37°C for 1 h. The 5 μl of proteinase K and 10 μl 25% SDS solution were added and incubated at 55°C for 1 h. After incubation, 1 : 1 volumes of phenol : chloroform were added to the solution. The mix was centrifuged (10 min at 13 000*g*) and 200 μl of chloroform was added to the aqueous layer and centrifuged (5 min at 13 000*g*). The supernatant was combined with 1/10 volume of 3 M sodium-acetate and 0.7 volumes of isopropanol and placed in −20°C for 2 h. DNA was pelleted by centrifuging for 15 min at 13 000*g* at 4°C. Supernatant was removed and DNA washed in cold 80% ethanol and resuspended in nuclease-free water. Additional voucher specimens are stored at the University of Florida.

A 500 bp shotgun library was constructed using this extract and sequenced using pair-end reads on a single lane of an Illumina HighSeq2000 Analyzer with 100 bp reads. Raw sequences from the chimpanzee louse (*P. schaeffi*) are deposited in GenBank Short Read Archive (accession SRX390495). We searched OrthoDB (http://cegg.unige.ch/orthodb6, [16]) for genes that are 1 : 1 : 1 single copy orthologues between humans (*Homo sapiens*), common chimpanzee (*Pan troglodytes*) and human body lice (*P. humanus*). This initial search recovered 3026 potential orthologues. We used CLC Genome Workbench (CLCbio) to assemble the Illumina reads for the chimpanzee louse against coding DNA sequences of the 3026 orthologues from the human body louse (*P. humanus*) genome [15] in Vector-Base (https://www.vectorbase.org). In addition to protein sequences from OrthoDB, we also retrieved coding DNA sequences for the 3026 putative orthologues for humans and chimpanzees from the Ensembl database [17].

As a further check on orthology, we used a BLAST (NCBI Blast 2.2.27) search of the chimpanzee genes against the human protein sequence database (from Ensembl). We also conducted a BLAST search of the chimpanzee louse assemblies against the protein database for human body louse (from VectorBase). Only those genes for which the top BLAST hit corresponded to the putative orthologue from OrthoDB were included in subsequent analyses (1724 genes). Gene sequences from humans and chimpanzees and from human body lice and chimpanzee lice were pairwise aligned in MUSCLE [18]. In 164 genes, stop codons were detected in the *Pan troglodytes* sequences and in 11 cases stop codons were detected in the *P. schaeffi* sequences. These genes were removed from further analyses. A statistical test for outliers using *Z*-value scores was performed on the per cent divergence values (see below), and after Bonferroni correction 11 genes were determined to be outliers for humans/chimpanzees and three genes for the lice. These outliers were removed leaving a final dataset of 1534 orthologous genes (alignment deposited in Dryad doi:10.5061/dryad.9fk1s).

For each of the 1534 orthologous loci, we used custom Perl scripts (deposited in GitHub, www.github.com/juliema/publications) to calculate uncorrected per cent sequence divergence between humans and chimpanzees and between the human body louse and chimpanzee louse. These DNA sequences were also translated to protein sequences and the same comparisons were made.

These comparisons did not include any sites in which gaps were introduced because of the alignment. In addition, D_n and D_s values were calculated for all these comparisons using the codeml program in the PAML package (v. 4.4b, [19]) and Nei & Gojobori's calculations of D_n and D_s [20]. To estimate the genome-wide relative rate, all sites were pooled for each species across genes and the average genome-wide sequence divergence was calculated by dividing the total number of substitutions by the total number of sites. All statistical analyses of these values were performed in the R statistics package [21].

To estimate the relative rates of sequence divergence for mitochondrial genes, the mitochondrial genomes for humans, chimpanzees and *P. humanus* were downloaded from GenBank. Because of the extremely high divergences between *P. humanus* and *P. schaeffi*, reference-based assemblies could not be used. Rather, we used a combination of Target Restricted Assembly [22] and BLAST searches of a de novo partial genome assembly of *P. schaeffi* constructed using SOAPDENOVO [23] to obtain sequences of mitochondrial protein-coding genes of *P. schaeffi*. The sequences for two short genes (ATPase8 and ND4L) could not be recovered using these methods. The sequences for the 11 genes that were recovered were aligned against sequences for *P. humanus* and genetic distances and relative rates calculated in the same way as for nuclear genes.

3. Results and discussion

For any relative rate comparison, it is important that orthologous genes are being compared, in this case between humans, chimpanzees, human lice and chimpanzee lice. First, we used the Ortholog Database (OrthDB, 16) to determine which protein-coding genes were strict 1 : 1 : 1 orthologues between humans, chimpanzees and the human body louse. Using reference-based assembly against the human body louse and reciprocal best BLAST analyses, we compiled sequences for these same 1534 orthologue genes in the chimpanzee louse genome (Material and methods). Comparisons of the per cent sequence divergence across these 1534 orthologues between human lice (*P. humanus*) and chimpanzee lice (*P. schaeffi*) to those for humans and chimpanzees revealed that lice are evolving 14.8 times faster at the DNA sequence level than their hosts (Wilcoxon signed-rank test, $p < 0.0001$). However, there was considerable variation in DNA sequence divergence across genes for both humans and chimpanzees (s.d. = 0.003) and for lice (s.d. = 0.02) (figure 1).

Some of this variation could be explained by a correlation between sequence divergences for a gene between lice and between primates (figure 1). Regression of pairwise divergences across the 1534 orthologous genes revealed a significantly positive correlation (slope = 1.08, Student's $t = 7.38$, $p < 0.0001$), though this correlation accounted for only a small fraction of the variation in divergence values across genes ($r^2 = 0.034$).

This correlation in rates of divergence across the genome could potentially be explained by similar selective constraints on the same gene in different taxa. Regression of protein divergences between lice against those between humans and chimpanzees were also significantly positive (slope = 1.98, Student's $t = 12.56$, $p < 0.0001$) and this correlation explained a larger fraction of the variation ($r^2 = 0.093$) than did comparisons of DNA substitutions. As with DNA divergence, at the protein level lice were evolving 8.6 times faster than primates. The average protein divergence between the lice was much higher (4.8%) than between humans and chimpanzees (0.6%).
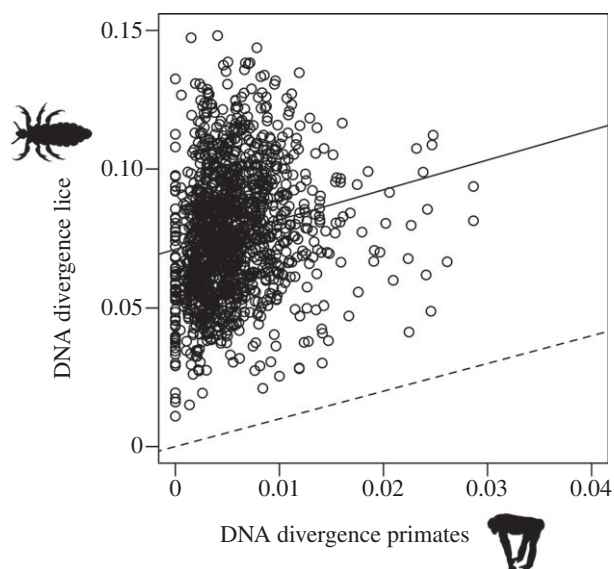
Figure 1. Plot of pairwise uncorrected sequence divergence across 1534 nuclear protein-coding genes between human and chimpanzee lice against that for the orthologous gene in humans and chimpanzees. The solid line indicates least-squares regression line (slope = 1.08, Student's $t = 7.38$, $p < 0.0001$), whereas the dashed line indicates expectation if genes evolve at the same rate in both groups. Points above the dashed line are genes that evolved faster in lice and points below the line are genes that evolved faster in humans and chimpanzees (none in this case).

To evaluate whether functional constraints on amino acids could fully explain the correlation across the genome in rates of DNA substitution, we used estimates of D_n (non-synonymous divergence) and D_s (synonymous divergence) calculated for pairwise comparisons for all 1534 genes between the two louse species as well as and between humans and chimpanzees. Regression of D_n for lice against D_n for humans and chimpanzees (figure 2a) revealed a significantly positive correlation (slope = 1.86, Student's $t = 11.20$, $p < 0.0001$). By contrast, regression of D_s for lice against D_s for humans and chimpanzees (figure 2b) showed no correlation (Student's $t = 0.82$, $p = 0.41$), even though synonymous substitutions accumulate 25.13 times faster in lice than their hosts.

The D_n/D_s ratio is often used as an estimate of the relative level of functional constraint on protein evolution and has also been used to detect positive selection. D_n/D_s ratios much less than 1 usually indicate purifying selection against amino acid changes, whereas D_n/D_s ratios greater than 1 are often taken as evidence for adaptive or positive selection [24]. Interestingly, none of the D_n/D_s ratios for these genes in lice were greater than 1, while D_n/D_s ratios for 24 genes were greater than 1 for humans and chimpanzees. In fact, the D_n/D_s ratios for primates (mean = 0.11) were significantly higher than those for lice (mean = 0.08) for the same genes (Wilcoxon signed-rank test $p < 0.00001$). The correlation between D_n values across genes between lice and primates is also reflected in the positive correlation between the D_n/D_s ratios (slope = 0.058, t-value = 8.65, $p < 1 \times 10^{-15}$).

Across 1534 genes that could be confidently assigned as orthologues between humans, chimpanzees and their parasitic lice, these genes evolve just over 14 times faster in lice when compared with in their hosts. Interestingly, the genetic divergence between lice was correlated with the genetic divergence in these same genes between humans and chimpanzees. That is, genes that evolve more rapidly in humans and chimpanzees also evolve more rapidly in their parasitic lice, even though lice

and primates are separated by more than 600 million years of evolution [25]. This is, to our knowledge, the first evidence of correlated rates of evolution across the genome between hosts and their parasites. The explanation for this correlation appears to be the relative level of functional constraint on different genes. Comparisons of the rate of silent substitutions did not reveal any correlation, while comparisons of the rate of substitutions that changed the amino acid (replacement substitutions) were correlated.

This study also provides an estimate of the relative rate of substitution between hosts and parasites across their genomes. Previous comparisons have generally involved single genes, typically from the mitochondrion. These estimates of mitochondrial relative rates of substitution between humans, chimpanzees and their lice indicated that lice evolve 2.3 times faster based on only the cytochrome oxidase I and cytochrome b gene regions [10]. However, this prior comparison used slightly different methods and did not estimate rates of substitution across the entire mitochondrial genome. Applying the same methods used in this study to mitochondrial protein-coding genes (see Material and methods), we estimated that mitochondrial genes in lice evolve 2.9 faster than in humans and chimpanzees, generally in line with these prior estimates. However, there was no correlation across these mitochondrial genes in the degree of genetic divergence (Student's $t = 0.52$, $p = 0.62$) as there was for nuclear genes.

The fact that both nuclear and mitochondrial genes in parasitic lice evolve more rapidly than those of their hosts suggest a more universal explanation for the rate increase is needed, rather than explanations specific only to the mitochondrion. Mitochondrial genomes of lice are highly rearranged compared with other insects [25–29], and these rearrangements were hypothesized to be correlated with increased substitution rates. In fact, in the human body louse (P. humanus), the mitochondrion is divided into a number of minicircular chromosomes [15,30]. The lack of mitochondrial single-stranded binding protein (mtSSB) in lice was postulated to account for this correlation [29]. However, the lack of mtSSB cannot explain the rate increase for nuclear protein-coding genes. In fact, the relative rate of substitution for nuclear genes is even much higher ($14.8\times$) than it is for mitochondrial genes ($2.9\times$), further suggesting that factors not specific to the mitochondrion are involved.

Another hypothesis that has been used to explain the higher substitution rate in lice, compared with their vertebrate hosts, is an elevated rate of slightly deleterious substitutions because of repeated population bottlenecks in lice upon transmission between hosts and inbreeding on individual hosts [4]. However, the fact that the D_n/D_s ratio is significantly higher for primates when compared with lice is evidence against this hypothesis, which predicts elevated non-synonymous substitutions (assumed to be slightly deleterious) in lineages with stronger bottlenecks.

In addition, the elevated substitution rate in lice occurs for both synonymous and non-synonymous substitutions (figure 2), which suggests that underlying differences in mutation rates must be accounted for in the explanation. Some possible explanations include genome-wide relaxed selection for mutation repair [29], shorter generation times [3] or overall elevated mutation rates [4]. In fact, an increased mutation rate combined with purifying selection could potentially explain the reduced D_n/D_s ratio in lice when compared with primates. Under this hypothesis, synonymous mutations would accumulate in proportion to the mutation rate [31], whereas

4

rspb.royalsocietypublishing.org   *Proc. R. Soc. B* **281**: 20132174
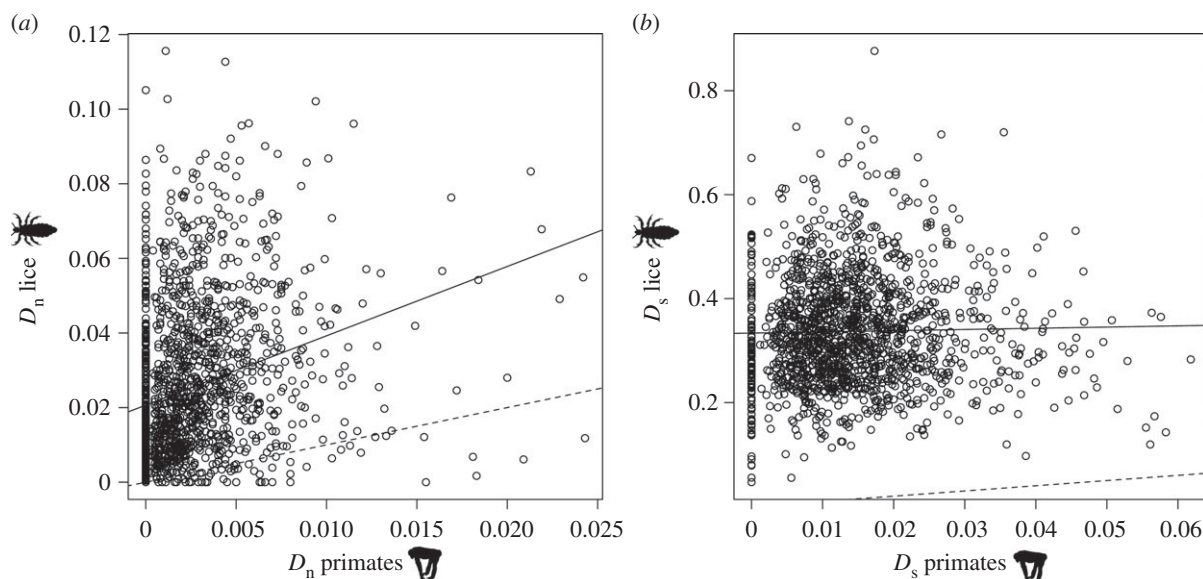
Figure 2. Plots of (*a*) non-synonymous divergence (D_n), the and (*b*) synonymous divergence (D_s) across 1534 nuclear protein-coding genes between human and chimpanzee lice against that for the orthologous gene in humans and chimpanzees. The solid line indicates the least-squares regression line (*a*: slope = 1.86, Student's $t = 11.20$, $p < 0.0001$; *b*: Student's $t = 0.82$, $p = 0.41$), while the dashed line indicates expectation if genes evolve at the same rate in both groups. Points above the dashed line are genes that evolved faster in lice and points below the line are genes that evolved faster in humans and chimpanzees.

non-synonymous mutations would be eliminated by purifying selection, which would lower the D_n/D_s ratio in the case of elevated mutation rates. Functional constraints on protein structure appear to play some role in determining the substitution rates for different genes; however, it explains very little (less than 3%) of the overall variation in substitution rates across genes for lice compared with humans and chimpanzees.

Research permit approvals were from Uganda Wildlife Authority (permit NS71) and Uganda Council of Science and Technology (UNCST).

References

1. Page RDM. 1994 Maps between trees and cladistics analysis of biogeography and cospeciation. *Syst. Zool.* **39**, 205–226. (doi:10.2307/2992182)

2. Page RDM. 1996 Temporal congruence revisited: comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and their chewing lice. *Syst. Biol.* **45**, 151–167. (doi:10.1093/sysbio/45.2.151)

3. Hafner MS, Sudman P, Villablanca F, Spradling T, Demastes J, Nadler S. 1994 Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265**, 1087–1090. (doi:10.1126/science.8066445)

4. Page RDM, Lee PLM, Becher A, Griffiths R, Clayton DH. 1998 A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. *Mol. Phylogenet. Evol.* **9**, 276–293. (doi:10.1006/mpev.1997.0458)

5. Ricklefs RE, Fallon SM. 2002 Diversification and host switching in avian malaria parasites. *Proc. R. Soc. Lond. B* **269**, 885–892. (doi:10.1098/rspb.2001.1940)

6. Johnson KP, Adams RJ, Page RDM, Clayton DH. 2003 When do parasites fail to speciate in response to host speciation? *Syst. Biol.* **52**, 37–47. (doi:10.1080/10635150390132704)

7. Jousselin ES, van Noort S, Berry V, Rasplus J-Y, Rnsted N, Erasmus JC, Greeff JM. 2008 One fig to bind them all: host conservatism in a fig wasp community unraveled by cospeciation analyses among pollinating and nonpollinating fig wasps. *Evolution* **62**, 1777–1797. (doi:10.1111/j.1558-5646.2008.00406.x)

8. Ricklefs RE, Outlaw DC. 2010 A molecular clock for malaria parasites. *Science* **329**, 226–229. (doi:10.1126/science.1188954)

9. Mizukoshi A, Johnson KP, Yoshizawa K. 2012 Co-phylogeography and morphological evolution of sika deer lice (*Damalina sika*) with their hosts (*Cervus nippon*). *Parasitology* **139**, 1614–1629. (doi:10.1017/S0031182012000996)

10. Reed DL, Smith VS, Hammond SL, Rogers AR, Clayton DH. 2004 Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol.* **2**, e304. (doi:10.1371/journal.pbio.0020304)

11. Reed DL, Light JE, Allen JM, Kirchman JJ. 2007 Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice. *BMC Biol.* **5**, 7. (doi:10.1186/1741-7007-5-7)

12. Smith VS, Ford T, Johnson KP, Johnson PCD, Yoshizawa K, Light JE. 2011 Multiple lineages of lice pass through the K-Pg boundary. *Biol. Lett.* **7**, 782–785. (doi:10.1098/rsbl.2011.0105)

13. Venter JC *et al*. 2001 The sequence of the human genome. *Science* **291**, 1305–1351. (doi:10.1126/science.1058040)

14. The Chimpanzee Sequencing and Analysis Consortium 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87. (doi:10.1038/nature04072)

15. Kirkness EF *et al*. 2010 Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl Acad. Sci. USA* **107**, 12 168–12 173. (doi:10.1073/pnas.1003379107)

16. Kriventseva EV, Rahman N, Espinosa O, Zbodnov EM. 2008 OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36**, D271–D275. (doi:10.1093/nar/gkm845)

17. Flicek P *et al*. 2012 Ensembl. *Nucleic Acids Res.* **40**, D84–D90. (doi:10.1093/nar/gkr991)

18. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **3**, 1792–1797. (doi:10.1093/nar/gkh340)

19. Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

20. Nei M, Gojobori T. 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.

21. R Development Core Team 2012 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org/).

22. Johnson KP, Walden KKO, Robertson HM. 2013 Next-generation phylogenomics using a target restricted assembly method. *Mol. Phylogenet. Evol.* **66**, 417–422. (doi:10.1016/j.ympev.2012.09.007)

23. Li R *et al*. 2010 De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272. (doi:10.1101/gr.097261.109)

24. Paland S, Lynch M. 2006 Transitions to asexuality result in excess amino acid substitutions. *Science* **311**, 990–992. (doi:10.1126/science.1118152)

25. Benton MJ, Ayala FJ. 2003 Dating the tree of life. *Science* **300**, 1698–1700. (doi:10.1126/science.1077795)

26. Shao R, Campbell NJH, Barker S. 2001 Numerous gene rearrangements in the mitochondrial genome of wallaby louse, *Heterodoxus macropus* (Phthiraptera). *Mol. Biol. Evol.* **18**, 858–865. (doi:10.1093/oxfordjournals.molbev.a003867)

27. Covacin C, Shao R, Cameron SL, Barker SC. 2006 Extraordinary amounts of gene rearrangement in the mitochondrial genomes of lice (Insecta: Phthiraptera). *Insect Mol. Biol.* **15**, 63–68. (doi:10.1111/j.1365-2583.2005.00608.x)

28. Cameron SC, Johnson KP, Whiting MF. 2007 The mitochondrial genome of the screamer louse *Bothriometopus* (Phthiraptera: Ischnocera): effects of extensive gene rearrangements on the evolution of the genome as a whole. *J. Mol. Evol.* **65**, 589–604. (doi:10.1007/s00239-007-9042-8)

29. Cameron SC, Yoshizawa K, Mizukoshi A, Whiting MF, Johnson KP. 2011 Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). *BMC Genomics* **12**, 394. (doi:10.1186/1471-2164-12-394)

30. Shao R, Kirkness EF, Barker SC. 2009 The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res.* **19**, 904–912. (doi:10.1101/gr.083188.108)

31. Kimura M. 1979 The neutral theory of molecular evolution. *Sci. Am.* **241**, 94–104. (doi:10.1038/scientificamerican1179-98)