This model can also explain the more recent observation that healthy children with microscopically detectable *P. falciparum* infections at the end of the dry season in Kenya are significantly more likely to recognize a heterologous parasite than those without parasites[14] despite similar levels of cumulative exposure, as summarized in Fig. 4b. Within our model framework, this can be ascribed to a difference in individual ability to mount an immune response to minor epitopes and therefore conforms to the relationship between duration of infection and γ shown in Fig. 4a. In other words, individuals who are better able to respond to the minor epitopes (as reflected in their ability to recognize heterologous parasites) are, paradoxically, more likely to sustain a chronic infection.

In summary, by proposing that transient heterologous responses to VSAs have evolved to coordinate the sequential expression of the associated multigene families within the host, we suggest a novel mechanism for antigenic variation in *P. falciparum* that also resolves several conflicting epidemiological observations. It will be interesting to see whether this paradigm extends to other antigenically variable pathogens such as African trypanosomes in which—in addition to structured switch rates and other mechanisms that might initiate an early order of expression[15,16]—a shared network of minor epitopes between antigenic variants might enable the parasite to exploit the host immune response to achieve chronicity. □

Methods

The dynamics of variant i can be given by

$$dy_i/dt = \phi y_i - \alpha z_i y_i - \alpha' w_i y_i \quad (1)$$

under the assumption that each variant has a net growth rate ϕ and can be destroyed at a rate α by the specific long-lasting immune response (z_i) and at a rate α' by the transient immune responses (w_i) against minor shared epitopes. The relative efficacy (at the effector level) of the transient immune response can be measured as $\gamma = \alpha'/\alpha$.

The dynamics of the specific response, z_i, against strain i can, in its simplest form, be represented as

$$dz_i/dt = \beta y_i - \mu z_i \quad (2)$$

where the proliferation rate β is in direct correlation with the amount of antigen (y_i), and μ is the rate of decay of the immune response. Functions that explicitly incorporate clonal expansion of the relevant immune cells (see Supplementary Information) can be substituted for βy in the proliferation term.

The dynamics of the transient cross-reactive immune response against the minor shared epitopes can be represented in the same form as equation (2) with the appropriate parameter changes:

$$dw_i/dt = \beta' \Sigma y_j - \mu' w_i \quad (3)$$

where j refers to all variants that share these epitopes with i. The relative efficacy of induction can be measured as β'/β. An equivalence between this measurement and γ ($=\alpha'/\alpha$) can be demonstrated analytically.

Note that this analytical framework intentionally excludes switching between variants as a means of structuring the appearance of antigenic variants.

Received 17 November 2003; accepted 9 March 2004; doi:10.1038/nature02486.

1. Newbold, C. Antigenic variation in *Plasmodium falciparum*: mechanisms and consequences. *Curr. Opin. Microbiol.* **2,** 420–425 (1999).
2. Scherf, A. *et al.* Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.* **17,** 5418–5426 (1998).
3. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
4. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4,** 358–360 (1998).
5. Giha, H. A. *et al.* Antibodies to variable *Plasmodium falciparum*-infected erythrocytes surface antigens are associated with protection from novel malaria infections. *Immunol. Lett.* **71,** 117–126 (2000).
6. Dodoo, D. *et al.* Antibodies to variant antigens on the surfaces of infected erythrocytes are associated with protection from malaria in Ghanian children. *Infect. Immun.* **69,** 3713–3718 (2001).
7. Tebo, A. E., Kremsner, P. G., Piper, K. P. & Luty, A. J. Low antibody responses to variant surface antigens of *Plasmodium falciparum* are associated with severe malaria and increased susceptibility to malaria attacks in Gabonese children. *Am. J. Trop. Med. Hyg.* **67,** 597–603 (2002).
8. Marsh, K. & Howard, R. Antigens induced on erythrocytes by *P. falciparum*: expression of diverse and conserved determinants. *Science* **231,** 150–153 (1986).
9. Gupta, S., Trenholme, K., Anderson, R. M. & Day, K. P. Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science* **263,** 961–963 (1994).
10. Bull, P. C., Lowe, B. S., Kortok, M. & Marsh, K. Antibody recognition of *Plasmodium falciparum* erythrocyte surface antigens in Kenya: evidence for rare and prevalent variants. *Infect. Immun.* **67,** 733–739 (1999).
11. Ofori, M. F. *et al.* Malaria-induced acquisition of antibodies to *Plasmodium falciparum* variant surface antigens. *Infect. Immun.* **70,** 2982–2988 (2002).
12. Giha, H. A. *et al.* Nine-year longitudinal study of antibodies to variant antigens on the surface of *Plasmodium falciparum*-infected erythrocytes. *Infect. Immun.* **67,** 4092–4098 (1999).
13. Kinyanjui, S., Bull, P. C., Newbold, C. I. & Marsh, K. Kinetics of antibody responses to *Plasmodium falciparum*-infected erythrocyte variant surface antigens. *J. Infect. Dis.* **187,** 667–674 (2003).
14. Bull, P. C. *et al. Plasmodium falciparum* infections are associated with agglutinating antibodies to parasite-infected erythrocyte surface antigens among healthy Kenyan children. *J. Infect. Dis.* **185,** 1688–1691 (2002).
15. Kosinski, R. J. Antigenic variation in trypanosomes: a computer analysis of variant order. *Parasitology* **80,** 343–357 (1980).
16. Agur, Z., Abiri, D. & Van der Ploeg, L. H. T. Ordered appearance of antigenic variants of African trypanosomes explained in a mathematical model based on a stochastic switch process and immune-selection against putative switch intermediates. *Proc. Natl Acad. Sci. USA* **86,** 9626–9630 (1989).
17. Frank, S. A. A model for the sequential dominance of antigenic variants in African trypanosome infections. *Proc. R. Soc. Lond. B* **266,** 1397–1401 (1999).
18. Antia, R., Nowak, M. A. & Anderson, R. M. Antigenic variation and the within-host dynamics of parasites. *Proc. Natl Acad. Sci. USA* **93,** 985–989 (1996).
19. Nowak, M. A. *et al.* Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature* **375,** 606–611 (1995).
20. Haraguchi, Y. & Sasaki, A. Evolutionary pattern of intra-host pathogen antigenic drift: effect of cross-reactivity in immune response. *Phil. Trans. R. Soc. Lond. B* **352,** 11–20 (1997).
21. Gog, J. R. & Grenfell, B. Dynamics and selection of many-strain pathogens. *Proc. Natl Acad. Sci. USA* **99,** 17209–17214 (2002).
22. Molineaux, L. *et al. Plasmodium falciparum* parasitaemia described by a new mathematical model. *Parasitology* **122,** 379–391 (2001).
23. Paget-McNichol, S., Gatton, M., Hastings, I. & Saul, A. The *Plasmodium falciparum var* gene switching rate, switching mechanism and patterns of parasite recrudenscence described by mathematical modelling. *Parasitology* **124,** 225–235 (2002).
24. Gupta, S. *et al.* The maintenance of strain structure in populations of recombining infectious agents. *Nature Med.* **2,** 437–442.
25. Chattopadhyay, R. *et al. Plasmodium falciparum* infection elicits both variant-specific and cross-reactive antibodies against variant surface antigens. *Infect. Immun.* **71,** 597–604 (2003).
26. Gamain, B., Miller, L. H. & Baruch, D. I. The surface variant antigens of *Plasmodium falciparum* contain cross-reactive epitopes. *Proc. Natl Acad. Sci. USA* **98,** 2664–2669 (2001).
27. Molineaux, L. & Gramiccia, G. *The Garki Project* (World Health Organisation, Geneva, 1980).
28. Beck, H.-P. *et al.* Analysis of multiple *Plasmodium falciparum* infections in Tanzanian children during the trial of the malaria vaccine SPf66. *J. Infect. Dis.* **175,** 921–926 (1997).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank A. McLean, G. Rudenko and D. Barry for their valuable comments, and the MRC and The Wellcome Trust for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.G. (sunetra.gupta@zoology.ox.ac.uk).

Positive selection at sites of multiple amino acid replacements since rat–mouse divergence

Georgii A. Bazykin[1], Fyodor A. Kondrashov[2], Aleksey Y. Ogurtsov[3], Shamil Sunyaev[4] & Alexey S. Kondrashov[3]

[1]*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Section of Evolution and Ecology, University of California at Davis, Davis, California 95616, USA*
[3]*National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894, USA*
[4]*Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA*

New alleles become fixed owing to random drift of nearly neutral mutations or to positive selection of substantially advantageous mutations[1–3]. After decades of debate, the fraction of fixations driven by selection remains uncertain[4–9]. Within 9,390 genes, we analysed 28,196 codons at which rat and mouse differ from each other at two nucleotide sites and 1,982 codons with three differences. At codons where rat–mouse divergence involved two non-synonymous substitutions, both of them occurred in

the same lineage, either rat or mouse, in 64% of cases; however, independent substitutions would occur in the same lineage with a probability of only 50%. All three non-synonymous substitutions occurred in the same lineage for 46% of codons, instead of the 25% expected. Furthermore, comparison of 12 pairs of prokaryotic genomes also shows clumping of multiple non-synonymous substitutions in the same lineage. This pattern cannot be explained by correlated mutation or episodes of relaxed negative selection, but instead indicates that positive selection acts at many sites of rapid, successive amino acid replacement.

We aligned 9,390 triplets of orthologous genes from rat, mouse and human. Among the 2,999,920 homologous rat and mouse codons within these genes, 83.30% were identical, and 15.70%, 0.94% and 0.07% differed at one, two and three nucleotide sites, respectively (no-, one-, two- and three-substitution codons). The average evolutionary distances between mouse and rat are 0.22 at synonymous sites (K_s) and 0.04 at non-synonymous sites (K_n), in agreement with previous estimates[10]. We assume that at an *i*-substitution codon, exactly *i* substitutions occurred after the divergence of rat and mouse lineages from their last common ancestor (the rat–mouse last common ancestor (RMCA)), because non-parsimonious evolutionary paths between such a close pair of species must be rare[2].

The RMCA codon is revealed exactly by the homologous human codon 'H' if no substitutions occurred on the path connecting these codons. Even after synonymous substitutions, H still reveals the amino acid encoded by the RMCA codon. As the K_s and K_n between human and rat or human and mouse is ~0.5 and ~0.1, respectively[10], we expect ~60% of human codons to coincide with the RMCA exactly and ~80% to encode the same amino acid. In fact, among no-substitution codons, H coincides with the codon present in rat 'R' and mouse 'M' in 69% of cases, and encodes the same amino acid, with or without synonymous substitutions, in 90% of cases.

At 71% of one-substitution codons, H coincides with either M or R, and at 74% of one-substitution codons, H encodes the same amino acid as M and/or R. In such cases we assume that H reveals the RMCA codon or, at least, the amino acid it encodes. Otherwise, RMCA remains unknown. We assume that the nucleotide (amino acid) substitution occurred in the rat lineage if H coincides (encodes the same amino acid) with M, and in the mouse lineage if H coincides (encodes the same amino acid) with R (Table 1).

Let us now consider the 28,196 two-substitution codons (Table 2). Among them rat and mouse differ from each other by: two synonymous substitutions (such codons encode either arginine or leucine; for example, TTA versus CTG) at 1,635 codons; one synonymous and one non-synonymous substitution (for example, CCC versus CAT) at 14,935 codons; none or one synonymous substitution and one or two non-synonymous substitutions depending on their order (for example, ACG versus AAT) at 4,417 codons; two non-synonymous substitutions or two synonymous substitutions depending on their order (for example, AGG versus CGT) at 715 codons; and two non-synonymous substitutions (for example, AAA versus AGT) at 6,146 codons. The two substitutions at a codon could occur in the rat lineage (pattern 0), the mouse lineage (pattern 2), or in both lineages (pattern 1). Accordingly, the RMCA codon would coincide with mouse codon M (pattern 0), rat codon R (pattern 2) or with one of the two intermediate codons I_1 and I_2 (pattern 1; for example, if the rat codon is AAG and the mouse codon is CCG, the intermediate codons are ACG or CAG; for some rat–mouse codon pairs, only one intermediate codon is possible because the other one is a stop codon). When H coincides (or encodes the same amino acid) with M, R, I_1 or I_2, we assume that it reveals the RMCA codon or, at least, the amino acid it encoded. This was the case for 57% and 62% of codons, respectively. Otherwise, RMCA and the pattern remain unknown.

If the two substitutions were independent (implying that neither of the intermediate codons is a stop codon) and equally common in rat and mouse lineages (Table 1), frequencies of patterns 0, 1 and 2 (P_0, P_1 and P_2) would be 25%, 50% and 25%, respectively. This is approximately the case when one or both substitutions at a codon were synonymous (Table 2). In contrast, when both substitutions were non-synonymous, we observed a large excess of the frequencies of patterns 0 and 2; that is, of codons where both substitutions occurred within the mouse or the rat lineage. This excess is significant both in comparison with the expected 25:50:25 ratio (chi-square, $P < 0.001$) and in comparison with the pattern in codons with two synonymous substitutions (chi-square, $P < 0.001$). Substantial clumping of non-synonymous substitutions within the same lineage only exists when both substitutions affect the same codon (Supplementary Fig. 1).

Analysis of 1,982 three-substitution codons reveals an even more marked clumping effect (Table 3). For each such codon we need to consider, in addition to R and M, six intermediate codons, three of which differ from R by one substitution (J_1, J_2 and J_3) and three of which differ from R by two substitutions (K_1, K_2 and K_3). When the RMCA codon coincides with M, a K codon, a J codon or R the corresponding number of substitutions that occurred in the rat lineage are 3, 2, 1 and 0, respectively (patterns α, β, γ and δ). If the substitutions occurred independently, the ratio of the numbers of codons with patterns α, β, γ and δ would be 1:3:3:1. However, a twofold excess of patterns δ and α is observed, which increases with the contribution of non-synonymous substitutions into rat–mouse divergence. Indeed, this excess is significantly higher when only 0–3 possible paths involve synonymous substitutions than when 4–6 paths involve them (chi-square, $P < 0.001$).

Could this clumping be an artefact? There are two possible sources of error. More than two substitutions may have occurred at a two-substitution codon since rat–mouse divergence. However, if the true number of substitutions at a codon was three, treating it as a two-substitution codon only underestimates the clumping. Indeed, we record pattern 0 (or 2) when H coincides with M (or R), and the presence of an extra substitution on the rat–mouse evolutionary path in such cases implies that three (instead of just two) substitutions occurred within the rat (or mouse) lineage. Clumping can be overestimated only with four or more substitutions at a two-substitution codon, but high degrees of non-parsimony must be very rare for rat and mouse. Furthermore, biased misidentification of RMCA is feasible. We compared data on false excess codons (where evolution on the RMCA–human path may inflate the observed P_0 and P_2) and on false deficit codons (where this evolution may cause underestimation of P_0 and P_2 (Supplementary Information)). The excess of patterns 0 and 2

Table 1 **Divergence at codons where rat and mouse differ at one nucleotide site**

Parameter	Substitution in rat lineage	Substitution in mouse lineage	RMCA unknown
Synonymous substitution	143,405 (51.2%)	136,475 (48.8%)	88,153 (24.0%)
Non-synonymous substitution			
Nucleotide-level pattern	27,305 (50.4%)	26,868 (49.6%)	48,692 (47.3%)
Amino-acid-level pattern	38,019 (50.3%)	37,583 (49.7%)	27,263 (26.5%)

Frequencies in the first two columns are only within codons where the RMCA is known, either at the nucleotide level or, at least, at the amino acid level (see text).

Table 2 **Divergence at codons where rat and mouse differ at two nucleotide sites**

Parameter	Both substitutions in rat lineage	One substitution in each lineage	Both substitutions in mouse lineage	RMCA unknown
Two synonymous substitutions	401 (29.0%)	638 (46.1%)	346 (25.0%)	250 (15.3%)
One synonymous and one non-synonymous substitution, neither intermediate codon is a stop	2,449 (27.4%)	4,237 (47.4%)	2,258 (25.3%)	5,900 (39.7%)
One synonymous and one non-synonymous substitution, one intermediate codon is a stop	16 (30.8%)	20 (38.5%)	16 (30.8%)	57 (52.3%)
Two synonymous or two non-synonymous substitutions	130 (28.6%)	197 (43.4%)	127 (28.0%)	261 (36.5%)
None or one synonymous substitution, two or one non-synonymous substitutions	731 (32.3%)	815 (36.0%)	719 (31.7%)	2,152 (48.7%)
Two non-synonymous substitutions, one intermediate codon is a stop, amino-acid-level pattern	78 (38.1%)	57 (27.8%)	70 (34.2%)	125 (37.9%)
Two non-synonymous substitutions, neither intermediate codon is a stop				
Nucleotide-level pattern	875 (30.2%)	1,047 (36.2%)	972 (33.6%)	3,252 (52.9%)
Amino-acid-level pattern at codons:				
All	1,266 (30.3%)	1,543 (36.9%)	1,371 (32.8%)	1,966 (32.0%)
Possible false excess	129 (30.0%)	161 (37.4%)	140 (32.6%)	163 (27.5%)
Possible false deficit	184 (31.5%)	243 (41.5%)	158 (27.0%)	274 (31.9%)
1,3-substitution	120 (28.4%)	163 (38.6%)	139 (32.9%)	193 (31.4%)
CpG-free	822 (30.6%)	971 (36.2%)	891 (33.2%)	1,210 (31.1%)
Convergence-free	165 (25.3%)	254 (39.0%)	233 (35.7%)	304 (31.8%)
Amino-acid-level pattern within regions:				
Very strong conservation	47 (35.6%)	28 (21.2%)	57 (43.2%)	23 (14.8%)
Strong conservation	314 (32.1%)	334 (34.1%)	331 (33.8%)	266 (21.4%)
Moderate conservation	428 (30.7%)	503 (36.1%)	464 (33.3%)	595 (29.9%)
All others	477 (28.5%)	678 (40.5%)	519 (31.0%)	1,082 (39.3%)
Amino-acid-level pattern at genes:				
Low K_n	82 (35.8%)	63 (27.5%)	84 (36.7%)	52 (18.5%)
Medium K_n	341 (30.6%)	400 (35.9%)	372 (33.4%)	404 (26.6%)
High K_n	843 (29.7%)	1,080 (38.1%)	915 (32.2%)	1,510 (34.7%)
Low G+C contents	515 (30.6%)	614 (36.4%)	556 (33.0%)	814 (32.6%)
Medium G+C contents	435 (30.1%)	523 (36.2%)	487 (33.7%)	683 (32.1%)
High G+C contents	316 (30.1%)	406 (38.7%)	328 (31.2%)	469 (30.9%)

Frequencies in the first three columns are only within codons where the RMCA is known. Possible false excess (false deficit) codons are those in which misidentification of RMCA is likely to cause overestimation (underestimation) of the frequency of patterns 0 and 2 (see Supplementary Information). 1,3-substitution codons are those where rat and mouse differ from each other at the first and third nucleotide sites. CpG-free codons are those in which neither of the two possible intermediate states between rat and mouse codons includes CpG context, neither inside the codon nor on its boundary. Convergence-free codons are those where the difference between properties of rat and mouse amino acids is greater than any of the four differences between one of them and one of the two possible intermediate amino acids. Regions with very strong, strong and moderate conservation are those in which the codon under consideration is flanked by gapless rat–mouse–human alignments of length 10 or more each with 10, 8 or 9, and 6 or 7 invariant amino acids, respectively. Genes were split into three equally large bins, according to their rat–mouse K_n or to their frequency of G and C. Average values of K_n within the bins are 0.006 (low), 0.026 (medium) and 0.081 (high). Average G+C contents within the bins are 0.463 (low), 0.530 (medium) and 0.592 (high).

is only marginally different in these two opposite cases (Table 2; chi-square, $P > 0.1$), ruling out systematic bias in misidentification of RMCA.

In contrast, random misidentification of RMCA will mask the excess of patterns 0 and 2. Even if both substitutions always occur in the same lineage ($P_1 = 0$), we would observe $P_1 \approx$one-sixth if H randomly deviates from RMCA in one-third of cases (assuming that a deviating H coincides with I_1 or I_2, instead of M or R, with a probability of 50%). Indeed, P_1 is lowest within genes with low K_n (chi-square, $P < 0.01$) and within conservative regions of all genes (chi-square, $P < 0.001$; Table 2), because misidentification of RMCA is rarer where the divergence is slow.

Data on two-substitution codons within 12 pairs of prokaryotic genomes show that the excess of patterns 0 and 2 for non-synonymous substitution is a universal phenomenon, and is more pronounced when the two sister genomes and the out-group are close to each other (Table 4). Thus, the clumping of non-synonymous substitutions within the same lineage must be real, and misidentification of the ancestral state only obscures it.

Three factors can cause substitutions to occur mostly in the same lineage. First, the corresponding mutations can be correlated. Such correlations can arise if one mutational event simultaneously affects two successive nucleotides; however, such double substitutions are rare[11], and the excess of patterns 0 and 2 remains high at codons where rat differs from mouse at the first and the third nucleotides (Table 2). Also, the first substitution can facilitate the second one by increasing the mutation rate. The only known mechanism of such facilitation[12] is the creation of a CpG hypermutable context by the first substitution. However, the excess of patterns 0 and 2 remains high for rat–mouse codon pairs where neither of the intermediate codons involves CpG, either inside the codon or on its boundary. Also, this excess does not depend on the G+C content of a gene (Table 2; chi-square, $P > 0.3$). Peculiarities of the mutation process can hardly be universal, and thus cannot explain a phenomenon observed in many diverse genomes (Table 4). Second, negative selection at a codon can affect one lineage less than the other. This can happen if such selection randomly switches off and on, perhaps due to substitutions at other codons of the same gene

Table 3 **Divergence at codons where rat and mouse differ at three nucleotide sites**

Number of paths involving synonymous substitutions	All substitutions in rat lineage	Two substitutions in rat and one in mouse	One substitution in rat and two in mouse	All substitutions in mouse lineage	RMCA unknown
Six	114 (20.2%)	160 (28.3%)	167 (29.6%)	124 (21.9%)	378 (40.1%)
Five	38 (24.8%)	45 (29.4%)	38 (24.8%)	32 (20.9%)	113 (42.5%)
Four	18 (17.0%)	30 (28.3%)	31 (29.2%)	27 (25.5%)	103 (49.3%)
Three	32 (32.7%)	20 (20.4%)	16 (16.3%)	30 (30.6%)	100 (50.5%)
Two	13 (24.5%)	13 (24.5%)	11 (20.8%)	16 (30.2%)	46 (46.5%)
One	8 (22.2%)	10 (27.8%)	8 (22.2%)	10 (27.8%)	21 (36.8%)
None	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (50.0%)

There are six evolutionary paths between two codons that differ from each other at all three sites, depending on the order in which the substitutions occur. Frequencies in the first four columns are only within codons where the nucleotide-level RMCA is known. Codons for which an intermediate stop codon is possible are not considered.

Table 4 **Patterns of divergence between 13 pairs of genomes**

Taxon	No. of genes	Fraction of amino acid differences*	All codons with synonymous substitutions†	All codons with non-synonymous substitutions†	Synonymous substitutions at two-substitution codons‡	Non-synonymous substitutions at two-substitution codons‡
Muridae	9,390	4.3% 11.6% 11.6%	143,955:136,963	49,194:48,743	401:638:346 (84.7%) 29.0:46.1:25.0§	1,266:1,543:1,371 (68.0%) 30.3:36.9:32.8§
Bacillus	2,260	3.5% 39.6% 39.5%	18,193:18,065	4,434:4,175	56:128:60 (39.1%) 23.0:52.5:24.6	190:172:151 (42.9%) 37.0:33.5:29.4§
Bordetella	2,878	0.4% 1.3% 1.2%	3,361:2,420	2,066:1,210	1:0:0 (100.0%) 100.0:0.0:0.0	36:3:79 (92.2%) 30.5:2.5:67.0§
Buchnera	421	20.2% 35.2% 35.2%	7,932:7,576	6,187:6,186	67:140:72 (70.6%) 24.0:50.2:25.8	380:484:408 (50.3%) 29.9:38.1:32.1§
Chlamydia	802	11.0% 31.5% 31.4%	22,215:21,908	7,088:6,835	275:502:263 (53.9%) 26.4:48.3:25.3	224:255:296 (42.9%) 28.9:32.9:38.2§
Escherichia	3,590	1.2% 3.4% 3.1%	25,412:21,197	7,151:4,116	255:41:88 (95.5%) 66.4:10.7:22.9§	419:56:164 (91.5%) 65.6:8.8:25.7§
Helicobacter	994	3.0% 42.7% 42.8%	4,500:4,954	1,478:1,606	16:43:19 (24.5%) 20.5:55.1:24.4	50:74:47 (36.7%) 29.2:43.3:27.5
Pseudomonas	2,714	16.9% 26.0% 25.3%	93,293:67,191	45,116:40,321	207:190:81 (65.7%) 43.3:39.8:17.0§	3,422:1,927:3,220 (59.2%) 39.9:22.5:37.6§
Pyrococcus	1,413	13.7% 20.1% 20.4%	43,107:48,275	17,067:18,063	365:698:390 (59.6%) 25.1:48.0:26.8	845:705:1,099 (66.1%) 31.9:26.6:41.5§
Salmonella	2,760	0.8% 12.4% 12.5%	9,230:9,147	2,276:2,466	52:35:37 (80.0%) 41.9:28.2:29.8§	172:37:67 (81.9%) 62.3:13.4:24.3§
Staphylococcus	1,873	0.5% 23.5% 23.5%	2,131:2,163	812:712	5:5:5 (60.0%) 33.3:33.3:33.3	32:17:15 (62.1%) 50.0:26.6:23.4§
Streptococcus	1,197	0.7% 27.3% 27.2%	1,779:1,802	732:638	6:16:9 (60.8%) 19.4:51.6:29.0	12:16:11 (50.0%) 30.8:41.0:28.2
Vibrio	796	0.7% 21.7% 21.7%	2,258:2,339	400:435	6:15:4 (61.0%) 24.0:60.0:16.0	8:9:6 (46.0%) 34.8:39.1:26.1

The following triplets of genomes were analysed, with sister genomes shown in parentheses. Muridae: (*Rattus norvegicus*, *Mus musculus*), *Homo sapiens*; *Bacillus*: (*B. cereus* ATCC 14579, *B. anthracis* strain Ames), *B. subtilis subtilis* strain 168; *Bordetella*: (*B. parapertussis*, *B. bronchiseptica* RB50), *B. pertussis* Tohama I; *Buchnera*: (*B. aphidicola* strain APS, *B. aphidicola* strain Sg), *B. aphidicola* strain Bp; *Chlamydia:* (*C. trachomatis*, *C. muridarum*), *C. caviae* GPIC; *Escherichia*: (*E. coli* O157:H7, *E. coli* K12), *E. coli* CFT073; *Helicobacter*: (*H. pylori* 26695, *H. pylori* J99), *H. hepaticus* ATCC 51449; *Pseudomonas*: (*P. syringae* pv. tomato strain DC3000, *P. putida* KT2440), *P. aeruginosa* PAO1; *Pyrococcus*: (*P. horikoshii*, *P. abyssi*), *P. furiosus* DSM 3638; *Salmonella*: (*S. typhimurium* LT2, *S. enterica enterica* serovar Typhi Ty2), *Shigella flexneri* 2a strain 301; *Staphylococcus*: (*S. aureus aureus* Mu50, *S. aureus aureus* MW2), *S. epidermidis* ATCC 12228; *Streptococcus*: (*S. pyogenes* M1 GAS, *S. pyogenes* MGAS315), *S. agalactiae* NEM316; *Vibrio*: (*V. vulnificus* YJ016, *V. vulnificus* CMCP6), *V. parahaemolyticus* RIMD 2210633.

*Fraction of mismatches in alignments of orthologous proteins between sister genome 1 and sister genome 2, sister genome 1 and the out-group, and sister genome 2 and the out-group, respectively.

†Total number of codons at which at least one synonymous, or at least one non-synonymous, substitution occurred in the first and the second sister lineage after their divergence.

‡Data on two-substitution codons where both substitutions were either synonymous or non-synonymous. The top line shows the number of codons where both substitutions occurred in sister lineage 1, one substitution occurred in each lineage, and both substitutions occurred in lineage 2, respectively. The fraction of codons for which the out-group reveals the last common ancestor is given in parentheses. The bottom line shows the frequencies of the corresponding patterns of substitution among those codons for which the common ancestor is known.

§Significance of deviation from 25:50:25 ($P < 0.01$).

(covarion model[13,14]) or of other genes. However, quantitative analysis of fluctuating negative selection shows that it is unlikely to generate the observed excess of patterns 0 and 2 (Supplementary Fig. 2).

The third (and the only remaining) explanation is positive selection. One possibility is that at a two-substitution codon, the first substitution was slightly deleterious and only the successive, compensatory substitution was advantageous. Such situations can be expected when the amino acid encoded after two substitutions is more similar to the original one than the amino acid encoded after the first substitution. We tested this possibility by only considering those two-substitution codons where the difference in properties (according to Miyata matrix[15]) between amino acids encoded by M and R is greater than any of the differences between amino acids encoded by I_1 and M, I_1 and R, I_2 and M, and I_2 and R. The excess of patterns 0 and 2 at such convergence-free codons was essentially the same as at other codons (Table 2; chi-square, $P > 0.2$).

Thus, positive selection usually favours every successive non-synonymous substitution at a codon. Indeed, when selection favours a new amino acid, there is no reason why this amino acid should always be reachable by just a single nucleotide substitution. This positive selection is probably different between the rat and mouse lineages, as identical selection would lead to homoplasy and their parallel evolution. However, the overall rate of evolution of a protein is principally determined not by positive selection but by selective constraint, because the combined excess of patterns 0 and 2 is the highest for slowly evolving proteins (Table 2).

A rather small (3–4%) excess of patterns 0 and 2 exists even when one (chi-square, $P < 0.001$) or both (chi-square, $P < 0.005$) substitutions at a two-substitution codon were synonymous (Table 2). Such excesses are also present in several pairs of bacterial genomes (Table 4). This is consistent with very weak selection on silent sites in mammalian[16,17] and in bacterial[2] genomes.

To generate the observed excesses of two-substitution codons with patterns 0 and 2, and of three-substitution codons with patterns α and δ, positive selection must act on a substantial fraction of non-synonymous substitutions at these codons. As there were over 2,500 of such substitutions in 9,390 genes, and the total number of amino acid replacements between mouse and rat is approximately 25 per protein (Table 4), the fraction of replacements driven by positive selection is at least about 0.5%. A tendency of neighbouring amino acid replacements to occur within the same lineage, which slowly declines when the distance between them increases (Supplementary Fig. 1), suggests that positive selection also affects some one-substitution codons. Thus, the overall contribution of positive selection may be considerably higher[4–9,18]. □

Methods

Analysis of alignments

Human, mouse and rat coding sequences were obtained from version 32 of the human genome[19], version 30 of the mouse genome[20] and version 2 of the rat genome[21] from the National Center for Biotechnology Information (NCBI). Orthologues were identified according to the two-directional best-hit approach tailored for three species[22] using protein BLAST[23]. Alignments of the amino acid sequences for each of the orthologue triplets were made using ClustalW[24] and reverse transcribed to get the nucleotide triple alignments. K_s and K_n were estimated using pairwise nucleotide alignments taken from the triple alignments for mouse–rat, mouse–human and rat–human pairs using the codeml program of the PAML package[25]. To eliminate erroneous alignments and non-orthologous gene triplets, those triplets that showed $K_n > 0.2$ and/or $K_s > 0.45$ in the mouse–rat comparison were removed from the analysis. To eliminate erroneous regions of alignments, which may originate from errors in genome assemblies or annotations, we only analysed those codons that were flanked by gapless alignments of length ten or more with at least five amino-acid matches between rat and mouse sequences (on each side) and at lease three matches between human and rat and/or mouse sequences. Difference between a pair of amino acids was taken as the corresponding term from Miyata matrix of

amino-acid pair distances[15]. Subdivision of rat–mouse–human codon triplets into classes represented in Tables 1–3 easily follows from the genetic code table. All suitable triplets of bacterial genomes were obtained from the NCBI Entrez database and processed analogously. All alignments are available at ftp://ftp.ncbi.nih.gov/pub/kondrashov/RatMouse/.

Fluctuating negative selection

We assumed that negative selection at a codon switches off and on at random moments. The expected waiting times (in units of time since rat–mouse divergence) for off to on and on to off switches are T and bT, respectively. Thus, negative selection is present with a probability of $b/(1 + b)$. If the total duration of episodes of absent negative selection in rat and mouse lineages were f_r and f_m, respectively, at a two-substitution codon $P_0 = r^2$, $P_1 = 2r(1 - r)$ and $P_2 = (1 - r)^2$, where $r = f_r/(f_r + f_m)$. This model was studied by Monte–Carlo simulations. For each pair of values of T and b, we performed 1,000,000 runs. Initially, negative selection in both rat and mouse lineages was off with probability $1/(1 + b)$ and on with probability $b/(1 + b)$. Then, switches of negative selection and accumulation of substitutions in the two lineages occurred independently. For each run, we calculated the probability that at a codon exactly two substitutions took place in rat and mouse lineages, assuming that substitutions occur independently, only when negative selection is off, with the instant rate 0.2 (the value of rat–mouse K_s). After this, the probability of pattern 1 was calculated within two-substitution codons.

Received 5 January; accepted 26 April 2004; doi:10.1038/nature02601.

1. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
2. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
3. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99,** 16134–16137 (2002).
4. Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
5. Fay, J. C., Wyckoff, G. J. & Wu, C.-I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415,** 1024–1026 (2002).
6. Bustamante, C. D. *et al.* The cost of inbreeding in *Arabidopsis*. *Nature* **416,** 531–534 (2002).
7. Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162,** 2017–2024 (2002).
8. Yang, Z. H. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12,** 688–694 (2002).
9. Anisimova, M., Nielsen, R. & Yang, Z. H. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164,** 1229–1236 (2003).
10. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95,** 9407–9412 (1998).
11. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21,** 12–27 (2003).
12. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
13. Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4,** 579–593 (1970).
14. Huelsenbeck, J. P. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* **19,** 698–707 (2002).
15. Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12,** 219–236 (1979).
16. Bustamante, C. D., Nielsen, R. & Hartl, D. L. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19,** 110–127 (2002).
17. Hellman, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13,** 831–837 (2003).
18. Wilbur, W. J. On the PAM matrix model of protein evolution. *Mol. Biol. Evol.* **2,** 434–447 (1985).
19. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
20. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
21. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
22. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
23. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
24. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
25. Yang, Z. H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank N. Bierne for a number of suggestions. G.A.B. was supported by a BWF graduate fellowship. S.S. was supported by Genome Canada Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.A.B. (gbazykin@princeton.edu).

Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body

Dae Sung Hwangbo, Boris Gersham, Meng-Ping Tu, Michael Palmer & Marc Tatar

Division of Biology and Medicine, Box G-W, Brown University, Providence, Rhode Island 02912, USA

In *Drosophila melanogaster*, ageing is slowed when insulin-like signalling is reduced: life expectancy is extended by more than 50% when the insulin-like receptor (*InR*) or its receptor substrate (*chico*) are mutated, or when insulin-producing cells are ablated[1–3]. But we have yet to resolve when insulin affects ageing, or whether insulin signals regulate ageing directly or indirectly through secondary hormones. *Caenorhabditis elegans* lifespan is also extended when insulin signalling is inhibited in certain tissues, or when repressed in adult worms[4,5], and this requires the forkhead transcription factor (FOXO) encoded by *daf-16* (ref. 6). The *D. melanogaster* insulin-like receptor mediates phosphorylation of dFOXO, the equivalent of nematode daf-16 and mammalian FOXO3a[7,8]. We demonstrate here that dFOXO regulates *D. melanogaster* ageing when activated in the adult pericerebral fat body. We further show that this limited activation of dFOXO reduces expression of the *Drosophila* insulin-like peptide *dilp-2* synthesized in neurons, and represses endogenous insulin-dependent signalling in peripheral fat body. These findings suggest that autonomous and non-autonomous roles of insulin signalling combine to control ageing.

To investigate whether activated dFOXO affects ageing in *D. melanogaster* we conditionally expressed *dFOXO* in specific adult tissues. Without ligand binding at the insulin-like receptor, dFOXO remains unphosphorylated and is transported to the nucleus where it promotes factors that retard cell growth and proliferation[7,8]. We transformed *D. melanogaster* with UAS-constructs, containing either a wild-type full-length complementary DNA of *dFOXO* (UAS–*dFOXO*) or *dFOXO* with the three protein kinase B (PKB) phosphorylation sites mutated to permit insulin-insensitive nuclear transport (UAS–*dFOXO*-TM). Expression of these constructs in the eye disc reduced growth (Supplementary Fig. S1), as has previously been reported for independent transformants of UAS–*dFOXO* and for a phosphorylation-site mutant of human *FOXO3a*[7,8]. The constitutive expression of UAS–*dFOXO* or UAS–*dFOXO*-TM killed larvae when promoted from *actin*-GAL4, or when expressed from fat body (*adh*-GAL4) or neurons (*ELAV*-GAL4) (Supplementary Table S1). Therefore, conditional expression of *dFOXO* is required to bypass developmental lethality as well as to study its impact on ageing exclusively in the adult stage.

We used the mifepristone inducible-GAL4 system (annotated P{Switch}[9] and GeneSwitch[10]) to drive the expression of UAS-constructs in defined adult tissues. Ingested mifepristone strongly induced reporter expression at all ages (Supplementary Fig. S1), and the compound alone had no effect on adult survival (Supplementary Fig. S2). Adult survival was not improved when UAS–*dFOXO*-TM was induced by a pan-neuronal driver (ELAV-GeneSwitch), or in glial cells (P{Switch} MB221) or neurolemma (P{Switch} $S_1$13) (Fig. 1; Supplementary Table S2). Thus, broadly activated dFOXO in neuron-associated cells is not sufficient to slow ageing; however, it may do so if expressed in subsets of cells within these tissues. Similarly, expression of UAS–*dFOXO*-TM or UAS–*dFOXO* did not affect survival when induced with the P{Switch} strain $S_1$106, an efficient promoter in the fat body[9]. In contrast, survival was significantly increased in both sexes when *dFOXO* was induced