than the PCR detection method. After co-culture, leaf discs were assayed for GUS activity (using 0.5 mg ml^{-1} X-glcA) or transferred to regeneration media containing antibiotics.

Rice (*Oryza sativa* L. cv. Millin) was transformed using seed-derived callus based on the protocol in ref. 28, with modifications (see Supplementary Information).

Arabidopsis thaliana L. cv. C24 was transformed using a floral-dip protocol[15] with bacteria resuspended in infiltration medium containing Murashige and Skoog Basal Medium, 5.0% (w/v) sucrose, 0.05% (v/v) Silwet L-77 and 0.1% (w/v) MES, pH 5.7. The modified media contained Murashige and Skoog Basal Medium, 1.0% (w/v) sucrose, 0.02% (v/v) Silwet L-77 and 0.1% (w/v) MES, pH 7.0.

Analysis of transformed plants

Genomic DNA was extracted from the regenerated plants using DNAzol (Invitrogen), digested with *Eco*RI and used for Southern blotting with a ^{32}P-labelled *GUSPlus* probe. Plant DNA sequences flanking the T-DNA insertion site(s) were determined by restriction digest/adaptor ligation[29].

Received 26 November; accepted 29 December 2004; doi:10.1038/nature03309.

1. Gelvin, S. B. *Agrobacterium*-mediated plant transformation: the biology behind the "gene-jockeying" tool. *Microbiol. Mol. Biol. Rev.* **67,** 16–37 (2003).
2. Roa-Rodríguez, C. & Nottenburg, C. *Agrobacterium*-mediated transformation of plants. *CAMBIA technology landscape paper* ⟨http://www.bios.net/Agrobacterium⟩ (2003).
3. Van Montagu, M. Jeff Schell (1935–2003): steering *Agrobacterium*-mediated plant gene engineering. *Trends Plant Sci.* **8,** 353–354 (2003).
4. Wood, D. W. *et al.* The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294,** 2317–2323 (2001).
5. Goodner, B. *et al.* Genome sequencing of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294,** 2323–2328 (2001).
6. Klein, D. T. & Klein, R. M. Transmittance of tumor-inducing ability to avirulent crown-gall and related bacteria. *J. Bacteriol.* **66,** 220–228 (1953).
7. Hooykaas, P. J. J., Klapwijk, P. M., Nuti, M. P., Schilperoort, R. A. & Rörsch, A. Transfer of the *Agrobacterium tumefaciens* Ti plasmid to avirulent Agrobacteria and to *Rhizobium ex planta*. *J. Gen. Microbiol.* **98,** 477–484 (1977).
8. van Veen, R. J. M., den Dulk-Ras, H., Bisseling, T., Schilperoort, R. A. & Hooykaas, P. J. J. Crown gall tumor and root nodule formation by the bacterium *Phyllobacterium myrsinacearum* after the introduction of an *Agrobacterium* Ti plasmid or a *Rhizobium* Sym plasmid. *Mol. Plant Microbe Interact.* **1,** 231–234 (1988).
9. van Veen, R. J. M., den Dulk-Ras, H., Schilperoort, R. A. & Hooykaas, P. J. J. Ti plasmid containing *Rhizobium meliloti* are non-tumorigenic on plants, despite proper virulence gene induction and T-strand formation. *Arch. Microbiol.* **153,** 85–89 (1989).
10. Goethals, K., Vereecke, D., Jaziri, M., Van Montagu, M. & Holsters, M. Leafy gall formation by *Rhodococcus fascians*. *Annu. Rev. Phytopathol.* **39,** 27–52 (2001).
11. Lambert, B. *et al.* Identification and plant interaction of a *Phyllobacterium* sp., a predominant rhizobacterium of young sugar beet plants. *Appl. Environ. Microbiol.* **56,** 1093–1102 (1990).
12. Young, J.M., Kuykendall, L.D., Martínez-Romero, E., Kerr, A. & Sawada, H. Classification and nomenclature of *Agrobacterium* and *Rhizobium*—a reply to Farrand *et al.* (2003). *Int. J. Syst. Evol. Microbiol.* **53**, 1689–1695 (2003).
13. Galibert, F. *et al.* The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293,** 668–672 (2001).
14. Hood, E. E., Gelvin, S. B., Melchers, L. S. & Hoekema, A. New *Agrobacterium* helper plasmids for gene transfer to plants. *Transgenic Res.* **2,** 208–218 (1993).
15. Pueppke, S. G. & Broughton, W. J. *Rhizobium* sp. strain NGR234 and *R. fredii* USDA257 share exceptionally broad, nested host ranges. *Mol. Plant Microbe Interact.* **12,** 293–318 (1999).
16. Jefferson R. A., Harcourt R. L., Kilian A., Wilson, K. J. & Keese, P. K. Microbial β-glucuronidase genes, gene products and uses thereof. US patent 6,391,547 (2003).
17. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743 (1998).
18. Watts, R. A. *et al.* A hemoglobin from plants homologous to truncated hemoglobins of microorganisms. *Proc. Natl Acad. Sci. USA* **98,** 10119–10124 (2001).
19. Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. & Fukatsu, T. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc. Natl Acad. Sci. USA* **99,** 14280–14285 (2002).
20. Brown, J. R. Ancient horizontal gene transfer. *Nature Rev. Genet.* **4,** 121–132 (2003).
21. Martin, W. *et al.* Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA* **99,** 12246–12251 (2003).
22. Suzuki, K., Yamashita, I. & Tanaka, N. Tobacco plants were transformed by *Agrobacterium rhizogenes* infection during their evolution. *Plant J.* **32,** 775–787 (2002).
23. Reimmann, C. & Haas, D. in *Bacterial Conjugation* (ed. Clewell, D. B.) 137–188 (Plenum, New York, 1993).
24. Nair, G. R., Liu, Z. & Binns, A. N. Reexamining the role of the accessory plasmid pATC58 in the virulence of *Agrobacterium tumefaciens* strain C58. *Plant Physiol.* **133,** 989–999 (2003).
25. Dennis, C. Biologists launch 'open-source movement'. *Nature* **431,** 494 (2004).
26. Stabb, E. V. & Ruby, E. G. RP4-based plasmids for conjugation between *Escherichia coli* and members of the Vibrionaceae. *Methods Enzymol.* **358,** 413–426 (2002).
27. Svab, Z., Hajdukiewicz, P. & Maliga, P. in *Methods in Plant Molecular Biology* (eds Maliga, P., Klessig, D. F., Cashmore, A. R., Gruissem, W. & Varner, J. E.) 55–77 (Cold Spring Harbor Laboratory Press, New York, 1995).
28. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6,** 271–282 (1994).
29. Cottage, A., Yang, A., Maunders, H., de Lacy, R. C. & Ramsay, N. A. Identification of DNA sequences flanking T-DNA insertions by PCR walking. *Plant Mol. Biol. Rep.* **19,** 321–327 (2001).

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

Acknowledgements This work was carried out within the Molecular Technologies Group of CAMBIA. We thank M. Irwin, J. Ward and H. Kilborn for their assistance with the plant transformation work. We thank P. Oger for sharing unpublished sequences of the pTiBo542 Ti plasmid, R. Wagner for comparisons of insertion sites with unpublished tobacco sequences, and M. Connett-Porceddu, P. Wenz and S. Hughes for discussions on the manuscript. This work was supported by grants from the Rockefeller Foundation, Horticulture Australia and Rural Industries R&D Corporation (RIRDC).

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

Correspondence and requests for materials should be addressed to R.A.J. (r.jefferson@cambia.org).

A universal trend of amino acid gain and loss in protein evolution

I. King Jordan[1], Fyodor A. Kondrashov[2], Ivan A. Adzhubei[3], Yuri I. Wolf[1], Eugene V. Koonin[1], Alexey S. Kondrashov[1] & Shamil Sunyaev[3]

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

Amino acid composition of proteins varies substantially between taxa and, thus, can evolve. For example, proteins from organisms with (G+C)-rich (or (A+T)-rich) genomes contain more (or fewer) amino acids encoded by (G+C)-rich codons[1–4]. However, no universal trends in ongoing changes of amino acid frequencies have been reported. We compared sets of orthologous proteins encoded by triplets of closely related genomes from 15 taxa representing all three domains of life (Bacteria, Archaea and Eukaryota), and used phylogenies to polarize amino acid substitutions. Cys, Met, His, Ser and Phe accrue in at least 14 taxa, whereas Pro, Ala, Glu and Gly are consistently lost. The same nine amino acids are currently accrued or lost in human proteins, as shown by analysis of non-synonymous single-nucleotide polymorphisms. All amino acids with declining frequencies are thought to be among the first incorporated into the genetic code; conversely, all amino acids with increasing frequencies, except Ser, were probably recruited late[5–7]. Thus, expansion of initially under-represented amino acids, which began over 3,400 million years ago[8,9], apparently continues to this day.

It is routinely assumed that extant proteins are in detailed equilibrium and their evolution is a stationary and reversible process: reciprocal fluxes of amino acid substitutions are equal, amino acid frequencies are constant, and nothing would change if time were to flow backwards[10–12]. Accordingly, symmetrical substitution matrices are used for protein sequence alignment[13].

We analysed 15 sets of three-way alignments of orthologous proteins encoded by triplets of closely related genomes (Table 1). At sites where the outgroup and one of the 'sister' genomes carry the same amino acid, whereas the other sister genome carries a different one, the amino acid ancestral for the sister genomes can be inferred. Accordingly, a specific, directed substitution can be assigned to the lineage of the sister that differs from the outgroup. For each set, we compared the 75 pairs of fluxes of forward and backward amino acid substitutions that can be caused by single-nucleotide replacements. The reciprocal fluxes differed by a factor of $\geq$1.5 for 575 of the 1,125 pairs (51%), and the differences were statistically signifi-

cant ($P \leq 0.05$, with sequential rejective correction for multiple tests[14]) for 175 pairs (16%). For example, after the divergence of human and chimpanzee lineages, there were 102 Leu → Phe substitutions but only 57 Phe → Leu substitutions in the 7,552 compared proteins. Thus, proteins are not in detailed equilibrium and their evolution is irreversible. Asymmetrical substitution matrices should be used for phylogeny-based alignment of multiple protein sequences[15].

Table 1 **Frequency changes of the strong gainer and loser amino acids in the 15 taxa**

Taxon	Number of proteins	Amino acid distance*	Useful sites†	G+C content¶	Pu/Py imbalance#
Hominidae	7,552	0.4, 7.9, 7.9	8,549 (83)	−0.098	+0.051
Muridae	12,356	4.2, 11.3, 11.3	242,639 (73)	−0.003	+0.007
Saccharomyces	3,462	6.3, 10.4, 9.3	90,496 (79)	+0.135	+0.011
Pyrococcus	1,283	11.7, 15.7, 15.0	35,300 (75)	+0.160	−0.138
Escherichia	3,337	0.7, 1.2, 1.3	7,511 (95)	−0.188	−0.005
Salmonella	2,623	0.6, 8.8, 8.9	4,560 (75)	−0.162	−0.052
Buchnera	386	14.1, 24.4, 23.0	10,794 (60)	+0.504	−0.097
Vibrio	755	0.5, 13.5, 13.5	1,063 (36)	+0.057	−0.031
Pseudomonas	2,507	15.0, 19.1, 17.4	95,260 (69)	−0.508	+0.046
Bordetella	2,815	0.3, 0.7, 0.8	2,822 (99)	−0.453	+0.016
Helicobacter	768	1.7, 23.7, 23.5	2,333 (54)	+0.356	−0.073
Chlamydia	750	7.1, 20.8, 20.1	13,033 (61)	+0.272	+0.021
Bacillus	3,728	3.0, 4.8, 4.7	30,874 (87)	+0.160	−0.006
Streptococcus	1,065	0.6, 15.9, 15.9	1,589 (68)	+0.312	−0.015
Staphylococcus	1,750	0.4, 15.3, 15.3	1,754 (69)	+0.385	−0.049

	Substitutions to and from an amino acid‡								
	Cys	Met	His	Ser	Phe	Gly	Glu	Ala	Pro
Hominidae	137/72, + 0.31‖	324/204, + 0.23‖	297/206, + 0.18‖	633/586, + 0.04	231/115, + 0.34‖	294/342, − 0.08	232/386, − 0.25‖	517/606, − 0.08§	204/437, − 0.36‖
Muridae	2,547/1,729, + 0.19‖	5,920/4,205, + 0.17‖	5,702/4,752, + 0.09‖	1,9904/1,8108, + 0.05‖	4,461/3,895, + 0.07‖	8,238/8,677, − 0.03§	8,056/11,269, − 0.17‖	1,559/1,733, − 0.05‖	7,350/9,883, − 0.15‖
Saccharomyces	778/430, + 0.29‖	1,768/1,455, + 0.10‖	1,715/1,606, + 0.03	7,776/7,567, + 0.01	1,824/1,546, + 0.08‖	3,370/2,511, + 0.15‖	4,006/4,554, − 0.06‖	5,336/4,950, + 0.04‖	2,188/2,290, − 0.02
Pyrococcus	35/21, + 0.25	772/426, + 0.29‖	277/244, + 0.06	1,756/1,222, + 0.18‖	680/598, + 0.06	727/757, − 0.02	1,663/2,836, − 0.26‖	1,241/1,598, − 0.13‖	211/416, − 0.33‖
Escherichia	143/57, + 0.43‖	238/155, + 0.21‖	284/194, + 0.19‖	738/549, + 0.15‖	151/122, + 0.11	239/334, − 0.17‖	360/476, − 0.14‖	531/1,129, − 0.36‖	138/294, − 0.36‖
Salmonella	67/11, + 0.79‖	127/81, + 0.22§	122/74, + 0.25§	365/216, + 0.26‖	84/57, + 0.19	126/205, − 0.24‖	176/202, − 0.07	285/519, − 0.29‖	53/167, − 0.52‖
Buchnera	75/35, + 0.36‖	166/102, + 0.25‖	191/105, + 0.29‖	654/587, + 0.05	249/164, + 0.21‖	99/149, − 0.20§	327/263, + 0.11§	403/396, + 0.01	65/124, − 0.31‖
Vibrio	4/0, + 1.00	22/15, + 0.19	16/10, + 0.23	64/52, + 0.10	18/17, + 0.03	23/24, − 0.02	45/70, − 0.22	45/93, − 0.35‖	10/21, − 0.36
Pseudomonas	501/339, + 0.19‖	2,985/1,090, + 0.47‖	1,737/1,252, + 0.16‖	6,289/4,277, + 0.19‖	1,956/1,623, + 0.09‖	2,759/3,186, − 0.07‖	3,780/5,448, − 0.18‖	6,623/9,261 − 0.17‖	1,054/2,287, − 0.37‖
Bordetella	89/14, + 0.73‖	104/66, + 0.22§	113/75, + 0.20§	301/153, + 0.33‖	50/51, − 0.01	157/247, − 0.22‖	102/125, − 0.10	248/575, − 0.40‖	83/188, − 0.39‖
Helicobacter	15/11, + 0.15	48/20, + 0.41§	55/24, + 0.39§	93/75, + 0.11	31/28, + 0.05	45/42, + 0.03	44/67, − 0.21	93/119, − 0.12	14/27, − 0.32
Chlamydia	107/59, + 0.29‖	212/137, + 0.22‖	249/150, + 0.25‖	911/761, + 0.09‖	238/160, + 0.20‖	174/198, − 0.07	307/512, − 0.25‖	697/790 − 0.06	116/224, − 0.32‖
Bacillus	235/126, + 0.30‖	960/722, + 0.14‖	910/739, + 0.10‖	2,044/1,861, + 0.05§	802/692, + 0.07§	898/932, − 0.02	1,702/2,258, − 0.14‖	1,911/2,117, − 0.05§	342/619, − 0.29‖
Streptococcus	20/3, + 0.74§	27/26, + 0.02	38/28, + 0.15	114/61, + 0.30‖	28/21, + 0.14	28/50, − 0.28	61/101, − 0.25§	71/128, − 0.29‖	13/47, − 0.57‖
Staphylococcus	10/1, + 0.83	32/23, + 0.16	30/25, + 0.09	97/75, + 0.13	44/31, + 0.17	33/53, − 0.23	81/93, − 0.07	76/103, − 0.15	6/47, − 0.77‖
Average *D*, (standard, error of *D*)	+0.452 (0.07)	+0.219 (0.03)	+0.178 (0.03)	+0.135 (0.03)	+0.120 (0.02)	−0.098 (0.03)	−0.150 (0.03)	−0.163 (0.04)	−0.362 (0.04)

*Percentage of mismatches in alignments of orthologous proteins between sister genome 1 and sister genome 2, sister genome 1 and the outgroup, and sister genome 2 and the outgroup, respectively.
†The number of sites where the two sister proteins contain different amino acids, and the percentage (in parentheses) of such sites where the outgroup has the same amino acid as one of the sisters.
‡The number of all substitutions creating (*C*) and removing (*R*), respectively, a particular amino acid are presented together with their normalized difference $D = (C - R)/(C + R)$. Out of 135 values of *D*, 89 (73 after the sequential rejective correction for multiple tests[14]) deviate from 0 significantly.
§The values of *D* that deviated from 0 significantly ($P \leq 0.01$) are indicated.
‖The values of *D* that deviated from 0 significantly ($P \leq 0.05$), even after the multiple tests correction[14], are indicated.
¶The normalized difference between the number of substitutions creating and removing a G-C pair within a coding region.
#The normalized difference between the number of substitutions creating and removing a purine within the coding strand. Pu, purine; Py, pyrimidine.

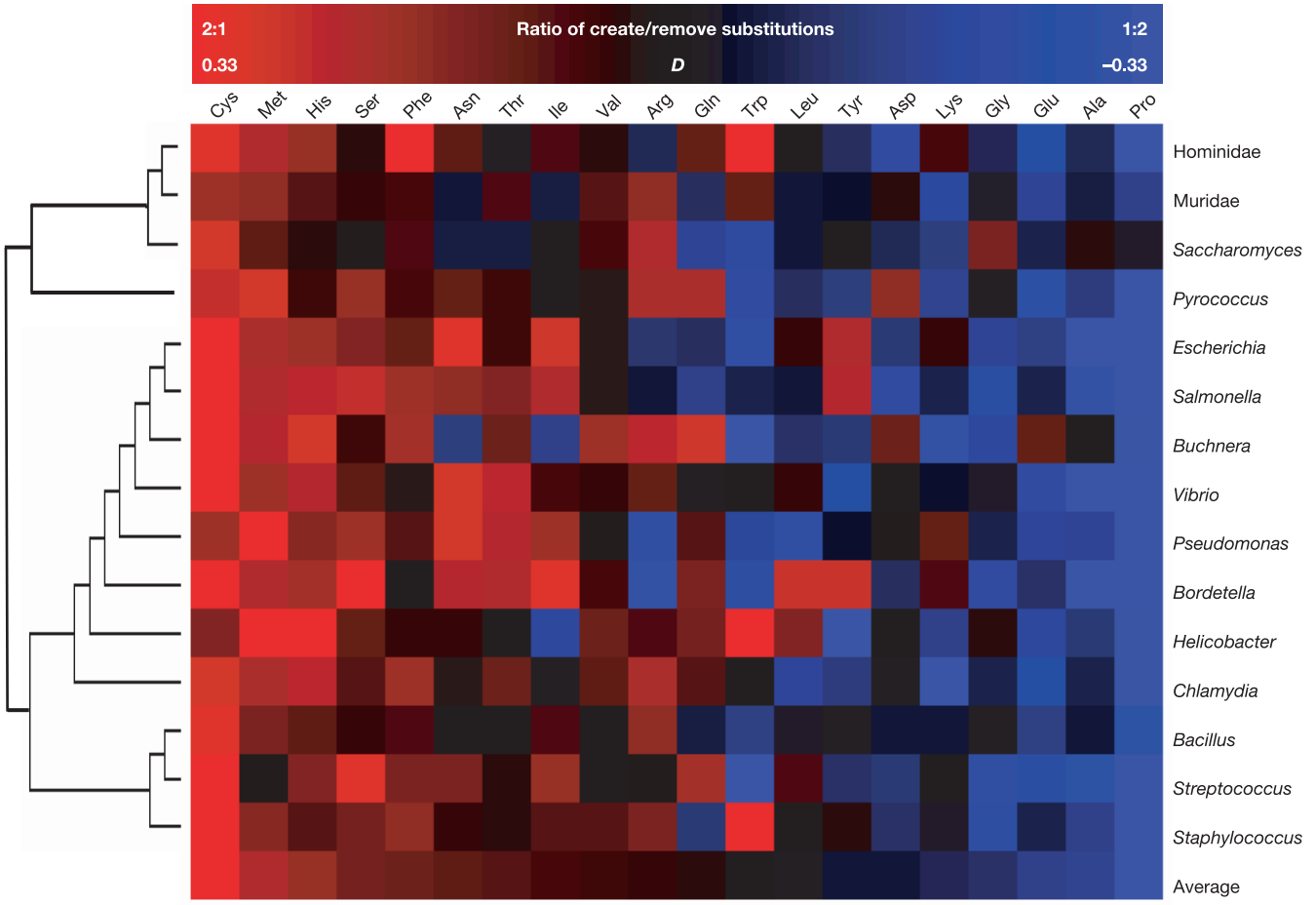

Figure 1 Normalized differences (D) between the number of substitutions creating (C) and removing (R) each amino acid ($D = (C - R)/(C + R)$) in 15 genome triplets. $|D| = 0.33$ corresponds to a twofold difference between C and R. A tentative phylogenetic tree of the analysed taxa is shown to the left of the matrix.

Table 2 **Gain and loss of amino acids in human SNPs**

Amino acid	Number of polymorphisms*	Number of substitutions†	Relationship between divergence and polymorphism‡
Cys	1,046/484, + 0.367	137/72, + 0.311	0.88
Met	1,219/811, + 0.201	324/204, + 0.227	1.06
His	1,616/835, + 0.319	297/206, + 0.181	0.75
Ser	2,589/2,321, + 0.055	633/586, + 0.039	0.97
Phe	1,046/572, + 0.293	231/115, + 0.335	1.10
Asn	1,199/940, + 0.121	340/277, + 0.102	0.96
Thr	2,109/1,839, + 0.068	504/524, − 0.019	0.84
Ile	1,660/1,245, + 0.143	525/450, + 0.077	0.88
Val	2,459/2,038, + 0.094	725/674, + 0.036	0.89
Arg	2,321/3,665, − 0.225	570/670, − 0.081	1.34
Gln	1,510/1,151, + 0.135	359/291, + 0.105	0.94
Trp	557/221, + 0.432	61/25, + 0.419	0.97
Leu	2,178/1,690, + 0.126	394/397, − 0.004	0.77
Tyr	609/440, + 0.161	96/114, − 0.086	0.61
Asp	927/1,300, − 0.167	197/308, − 0.220	0.90
Lys	1,370/1,032, + 0.141	336/292, + 0.070	0.87
Gly	1,374/1,954, − 0.174	294/342, − 0.075	1.22
Glu	926/1,729, − 0.302	232/386, − 0.249	1.12
Ala	1,505/2,740, − 0.291	517/606, − 0.079	1.55
Pro	1,118/2,331, − 0.352	204/437, − 0.363	0.97

*For each amino acid, listed as in Fig. 1 in the order of the decreasing D values, the number of human non-synonymous SNPs where the amino acid is the derived allele (p_+), the number of SNPs where the amino acid is the ancestral allele (p_-), and the normalized difference D between these numbers is presented (all the normalized differences deviate from 0 significantly, $P < 10^{-9}$, except for Ser).

†For each amino acid, the number of substitutions creating (s_+) and removing (s_-) it since the human–chimpanzee divergence is shown, together with the normalized difference.

‡The ratio $r = (s_+/p_+)/(s_-/p_-)$ was computed for each amino acid. For a gainer, $r > 1$ if a large fraction of the substitutions that created it were driven by positive selection; conversely, if many substitutions removing a loser were driven by positive selection, $r < 1$ for it[24]. Clearly, the values of r are not higher for gainers than for losers, arguing against a major role of positive selection.

Moreover, amino acid composition of proteins is far even from a total equilibrium (Fig. 1 and Table 1). With all substitutions taken into account, regardless of the number of nucleotide replacements required, five strong 'gainers' (Cys, Met, His, Ser and Phe) accrue in 14 or all 15 taxa, and frequencies of four strong 'losers' (Pro, Ala, Glu and Gly) decline in at least 13 taxa. Four amino acids are weak gainers, which either accrue in 11 taxa (Asn, Thr and Ile) or accrue slowly in all taxa (Val); in contrast, Lys is a weak loser, lost in ten taxa (Table 1). Frequencies of the remaining six amino acids (Arg, Gln, Trp, Leu, Tyr and Asp) evolve more erratically, although many individual changes are highly significant, reflecting taxon-specific mutational and selective pressures[1–4] (Supplementary Table 1).

The only conceivable source of systematic error in this analysis could be failure to identify the ancestral amino acid at a site where multiple substitutions occurred. This could disproportionately reduce the number of detected substitutions that remove evolvable and/or rare amino acids, thus creating a false impression that the frequencies of such amino acids increase[16]. However, a conservative (for our purposes) correction for multiple substitutions did not change the results substantially (Supplementary Methods and Supplementary Table 1). Also, the pattern of amino acid gain and loss in three taxa where the outgroup was close enough to the sister genomes to make misidentification of the ancestral state unlikely (*Escherichia*, *Bordetella* and *Bacillus*) was the same as in the remaining 12 taxa (Fig. 1 and Table 1). Furthermore, limiting the analysis to highly conserved regions of proteins did not affect this

pattern (data not shown). Finally, we examined human single-nucleotide polymorphisms (SNPs) polarized unambiguously by orthologous chimpanzee sequences. Because a majority of even non-synonymous human SNPs are effectively neutral[17], they are expected to reveal a pattern of ongoing amino acid gain and loss in the human proteins similar to that during the course of human–chimpanzee divergence. Indeed, this was the case (Table 2), which corroborates our inferences from interspecies comparisons.

Because the sister species in all the 15 taxa are closely related, the observed pattern directly reflects only the last ~10 million years of evolution. We cannot conceive of a global external factor that could cause, during this time, parallel evolution of amino acid compositions of proteins in 15 diverse taxa that represent all three domains of life and span a wide range of lifestyles and environments. Thus, currently, the most plausible hypothesis is that we are observing a universal, intrinsic trend that emerged before the last universal common ancestor of all extant organisms. A comparison of extant proteins and reconstructed ancient proteins[8,9] implied a long-term trend consistent with the trend observed in recent evolution (Supplementary Table 2), despite several substantial limitations of the deep reconstructions.

This trend in protein evolution is not driven by any simple trend at the DNA level. Among the losers, Pro, Ala and Gly are encoded by (G+C)-rich codons, so declining G+C content might, potentially, drive their loss[1–4]. However, Arg and Trp, also encoded by (G+C)-rich codons, are not consistently lost, whereas Tyr and Lys, encoded by (A+T)-rich codons, are not gainers. Furthermore, frequencies of strong gainers or losers encoded by two codons, one containing two Gs or Cs and the other containing one G or C (Cys, His and Glu), cannot be directly affected by G+C content. Most importantly, there was no consistent pattern in evolution of the G+C content (which increases more often than it declines, in nine and six taxa, respectively; Table 1), purine/pyrimidine strand imbalance (Table 1), or several other DNA-level parameters (data not shown), in marked contrast to the uniformity observed at the protein level.

Proteins currently gain, on average, ~0.007 Cys and lose ~0.014 Pro residues per substitution (Table 3). Using a simple linear model to extrapolate the observed dynamics backwards, we estimate that, since the time when the nine gainers first appeared, about two substitutions have been accepted per protein site, and proteins at that time mostly consisted of today's losers (Table 3). This epoch probably antedates the last universal common ancestor, which probably lived ~3.5 billion years ago[18]. Amino acid substitutions in murid rodents[19], insects[20], vertebrates[20], diatoms[21], Enterobacteriaceae[22] and *Buchnera*[23] occur with average rates of 1.5, 1.5, 1, 2, 0.2 and 1.5 substitutions per site per billion years, respectively. Thus, the average number of substitutions per site since the last universal common ancestor is probably between 0.7 (the number suggested by Enterobacteriaceae; the proteins of the endosymbiont *Buchnera* evolve unusually quickly for bacteria[23]) and 7, a range that includes the above estimate.

A forward extrapolation shows that the asymptotic frequencies of the gainers (losers), to be reached in the distant future, are much higher (lower) than their current frequencies. In extant proteins, gainers tend to be under-represented (except for Met), and losers are over-represented (except for Pro), relative to the frequencies corresponding to equal usage of all codons, but these differences are predicted to decline (Table 3).

Gainers (or losers) do not seem to be unusually similar to each other in terms of their physico-chemical properties, metabolic pathways, or codon assignments. However, a loser-rich and gainer-poor amino acid composition of primordial proteins may be explained historically by the order in which amino acids were recruited into the genetic code[5–8]. All four strong losers (but only one of the five strong gainers, Ser) are among the ten amino acids produced in spark experiments imitating conditions for pre-biotic synthesis[6]. All strong losers, but none of the strong gainers, are among the eight amino acids found in the Murchison meteorite and thought to be abiogenic[7]. A consensus chronology of incorporation of amino acids into the genetic code, based on 60 criteria[5], suggests that all strong losers are among the six most ancient amino acids, whereas four of the five strong gainers (except for Ser) are among the seven recruited last (Table 3; see also Supplementary Table 3). Notable exceptions are Tyr and Trp, assumed to have been recruited among the last but not accumulating consistently. Perhaps these

Table 3 **Inferred evolutionary trajectories of amino acid frequencies**

Amino acid*	Rate of gain/loss†	Current frequency	Asymptotic frequency‡	Frequency with unbiased usage of codons§	Number of substitutions since origin‖	Initial frequency¶	Consensus order[5] of recruitment into the genetic code
Cys	+0.0067 ± 0.00171	0.012	0.032	0.033	1.363	–	16
Met	+0.0088 ± 0.00126	0.025	0.038	0.016	1.595	–	19
His	+0.0073 ± 0.00107	0.022	0.032	0.033	1.592	–	14
Ser	+0.0167 ± 0.00356	0.062	0.080	0.098	1.597	–	6
Phe	+0.0042 ± 0.00091	0.042	0.053	0.033	4.067	–	17
Asn	+0.0073 ± 0.00268	0.040	0.048	0.033	1.985	–	13
Thr	+0.0091 ± 0.00282	0.051	0.061	0.066	1.957	–	9
Ile	+0.0089 ± 0.00500	0.068	0.077	0.049	2.120	–	12
Val	+0.0098 ± 0.00182	0.069	0.078	0.066	1.927	–	4
Arg	+0.0038 ± 0.00433	0.051	0.056	0.098	3.195	–	10
Gln	+0.0020 ± 0.00184	0.038	0.041	0.033	3.821	–	11
Trp	+0.0002 ± 0.00041	0.011	0.012	0.016	–	–	20
Leu	−0.0017 ± 0.00346	0.103	0.100	0.098	–	0.109	8
Tyr	−0.0005 ± 0.00118	0.032	0.031	0.033	–	0.034	18
Asp	−0.0039 ± 0.00229	0.052	0.047	0.033	–	0.074	3
Lys	−0.0065 ± 0.00333	0.059	0.051	0.033	–	0.092	15
Gly	−0.0063 ± 0.00253	0.072	0.059	0.066	–	0.093	1
Glu	−0.0137 ± 0.00244	0.065	0.048	0.033	–	0.136	7
Ala	−0.0239 ± 0.00862	0.081	0.058	0.066	–	0.242	2
Pro	−0.0139 ± 0.00262	0.043	0.022	0.066	–	0.101	5

*Assuming that u and v, the per site rates of substitution from and to an amino acid, respectively, are constant, x, the frequency of the amino acid, obeys the equation $dx/dt = v(1 - x) - ux$. Thus, $x(t) = (x_0 - v/(u + v))\exp(-(u + v)t) + v/(u + v)$, with time measured in the units of the number of amino acid substitutions per site. We used values of u and v that are their arithmetic means across the 15 taxa. Obviously, forward and backward extrapolations of x obtained in this way should be treated with caution.

†The average number of gained or lost residues per one amino acid substitution per protein site for the 15 taxa (mean ± standard error; $n = 15$).

‡In the distant future, the frequency of an amino acid will approach $v/(u + v)$.

§Frequency of an amino acid attained if all non-stop codons were used equally; that is, $n/61$, where n is the number of codons that encode the amino acid.

‖The estimated average number of substitutions per site since the moment when the frequency of a gainer was 0.

¶The frequency of a loser at the moment in the past since which two substitutions per average site occurred.

amino acids reached equilibrium concentrations more quickly than other late recruitments.

The initial frequency of a late-coming amino acid must be low because too many simultaneous changes in all proteins of the cell must be lethal. In particular, four of the five strong gainers (except for Ser) are thought to have been recruited by capturing codons that previously encoded other amino acids[5]. Such recruitments would need time to gain a foothold, at the expense of the primordial amino acids. Indeed, reconstructed ancient proteins were poor in Cys residues and some other late-coming amino acids, suggesting that they still accrued for some time after the last universal common ancestor (refs 8, 9 and Supplementary Table 2). However, it is remarkable that accumulation of late-coming amino acids continues to this day throughout the entire range of cellular life, affecting, among other things, the human–chimpanzee divergence and the genetic variability within human populations.

Accumulation of the gainers could not be driven solely by random drift because selectively neutral equilibrium would have been reached a long time ago. If gainers were favoured by positive selection, the rate of fixation of amino acid substitutions would be increased but the level of polymorphism would not. However, an analysis similar to the McDonald–Kreitman test[24] did not reveal any excess, in the course of human–chimpanzee divergence, in the rate of substitutions that add gainers or remove losers, relative to the density of corresponding human SNPs (Table 2). Probably, accumulation of the gainers was caused mostly by relaxation of negative selection if, for example, an ever-increasing fraction of sites where only Gly conferred high fitness initially became tolerant to Cys. Opening of extra sites to the gainers, if caused by epistatic interactions with other evolving sites[25,26], can be very slow. □

Methods

All 12 available triplets of complete prokaryotic genomes—for which the divergence of the amino acid sequences of orthologous proteins from the sister genomes (the closest pair) was within 0.3–15%, and the outgroup–sisters divergence was <25%—were extracted from the NCBI genome database using the Entrez retrieval engine (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). There are also three suitable triplets of eukaryotic genomes, resulting in the following 15 triplets, with sister genomes shown in parentheses. Hominidae: (*Homo sapiens*, *Pan troglodytes*), *Mus musculus*; Muridae: (*M. musculus*, *Rattus norvegicus*), *H. sapiens*; *Saccharomyces*: (*S. cerevisiae*, *S. paradoxus*), *S. mikatae*; *Pyrococcus*: (*P. horikoshii*, *P. abyssi*), *P. furiosus* DSM 3638; *Escherichia*: (*E. coli* K12, *E. coli* O157:H7), *E. coli* CFT073; *Salmonella*: (*S. typhimurium* LT2, *S. enterica enterica* serovar Typhi Ty2), *Shigella flexneri* 2a strain 301; *Buchnera*: (*B. aphidicola* strain APS, *B. aphidicola* strain Sg), *B. aphidicola* strain Bp; *Vibrio*: (*V. vulnificus* YJ016, *V. vulnificus* CMCP6), *V. parahaemolyticus* RIMD 2210633; *Pseudomonas*: (*P. syringae* pv. tomato strain DC3000, *P. putida* KT2440), *P. aeruginosa* PAO1; *Bordetella*: (*B. bronchiseptica* RB50, *B. parapertussis*), *B. pertussis* Tohama I; *Helicobacter*: (*H. pylori* 26695, *H. pylori* J99), *H. hepaticus* ATCC 51449; *Chlamydia*: (*C. trachomatis*, *C. muridarum*), *C. caviae* GPIC; *Bacillus*: (*B. anthracis* strain Ames, *B. cereus* ATCC 10987), *B. cereus* ATCC 14579; *Streptococcus*: (*S. pyogenes* M1 GAS, *S. pyogenes* MGAS315), *S. agalactiae* NEM316; *Staphylococcus*: (*S. aureus aureus* Mu50, *S. aureus aureus* MW2), *S. epidermidis* ATCC 12228.

Human–chimpanzee–mouse and *S. cerevisiae–S. paradoxus–S. mikatae* alignments were obtained from refs 27 and 28, respectively, and mouse–rat–human alignments were constructed as described in ref. 19. Triplets of orthologues, identified using the reciprocal best-hit approach[29], were aligned by ClustalW[30], except for pre-aligned primate and yeast proteins. To eliminate erroneous alignments and misidentified orthologues, triplets in which the divergence between two sister sequences significantly ($P < 0.001$) exceeded the minimal divergence between a sister and the outgroup sequence were discarded. To eliminate incorrectly aligned regions, which could originate from errors in genome assemblies or gene structure prediction, we analysed only sites embedded within a frame of length ten with at least six matches between the sister sequences and at least four matches between each of the sisters and the outgroup. All alignments are available at ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine.

In each three-sequence alignment, only sites where the outgroup and one of the sisters carries the same amino acid but the two sisters carry different amino acids were analysed. Sites where the outgroup differed from both sisters were ignored because the ancestral state of the sisters cannot be inferred. The 15 20 × 20 matrices of amino acid substitutions are available at ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine. Fluxes of reciprocal substitutions caused by single-nucleotide replacements were compared and the statistical significance of the observed deviations from detailed equilibrium was estimated for each of the 75 pairs of reciprocal fluxes using a proportion test, after which sequential rejective Bonferroni procedure[14] for multiple tests was applied to the whole set of results. Owing to the close relatedness of the sister genomes, substitutions that require two or three nucleotide replacements comprised, on average, only 6.8% of all substitutions (from 0.7% in Hominidae to 17% in *Pseudomonas*) and, thus, were too uncommon to compare individual fluxes. The total gain or loss of an amino acid was calculated by counting all substitutions to and from it (Table 1; see also Supplementary Table 1).

This approach relies on close similarity between the sister and the outgroup sequences. Otherwise, it becomes prone to systematic error stemming from two sources: (1) the probability of identical substitutions in the outgroup and one of the sister lineages or of multiple substitutions in a lineage becomes substantial, leading to erroneous inference of the ancestral state and, thus, to misidentification of the substitution; and (2) the fraction of sites where both sisters differ from the outgroup becomes substantial, and exclusion of such sites biases the count of substitutions. To deal with these effects, we used a statistical correction for multiple substitutions (Supplementary Methods). The correction was devised to be conservative with respect to the asymmetry of reciprocal substitutions; that is, it tends to underestimate gain or loss of amino acids.

Human non-synonymous SNPs were extracted from dbSNP build 120 (http://ncbi.nlm.nih.gov/SNP/). The ancestral amino acid for each SNP was inferred using the orthologous chimpanzee sequence from the UCSC chimpanzee genome assembly, version PanTro1 (http://hgdownload.cse.ucsc.edu/goldenPath/panTro1/bigZips/). The 29,262 SNP sites were analysed as described for substitutions between species.

Received 14 September; accepted 15 December 2004; doi:10.1038/nature03306.
Published online 19 January 2005.

1. Sueoka, N. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl Acad. Sci. USA* **47,** 1141–1149 (1961).
2. Gu, X., Hewett-Emmett, D. & Li, W.-H. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* **103,** 383–391 (1998).
3. Knight, R. D., Freeland, S. J. & Landweber, L. F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **4,** Research0010.1 (2001).
4. Wang, H.-C., Singer, G. A. C. & Hickey, D. A. Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **21,** 90–96 (2004).
5. Trifonov, E. N. The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22,** 1–11 (2004).
6. Miller, S. L. Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb. Symp. Quant. Biol.* **52,** 17–27 (1987).
7. Cronin, J. R. & Pizzarello, S. Amino acids in meteorites. *Adv. Space Res.* **3,** 5–18 (1983).
8. Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19,** 1645–1655 (2002).
9. Brooks, D. J. & Fresco, J. R. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Mol. Cell. Proteom.* **1,** 125–131 (2002).
10. Muller, T. & Vingron, M. Modeling amino acid replacement. *J. Comp. Biol.* **7,** 761–776 (2000).
11. Goldman, N. & Whelan, S. A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.* **19,** 1821–1831 (2002).
12. Veerassamy, S., Smith, A. & Tillier, E. R. M. A transition probability model for amino acid substitutions from blocks. *J. Comp. Biol.* **10,** 997–1010 (2003).
13. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices. *Adv. Prot. Chem.* **54,** 73–97 (2000).
14. Holm, S. A simple sequentially rejective multiple test procedure. *Stand. J. Stat.* **6,** 65–70 (1979).
15. Feng, D. F. & Doolittle, R. F. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* **266,** 368–382 (1996).
16. Eyre-Walker, A. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47,** 686–690 (1998).
17. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10,** 591–597 (2001).
18. Tice, M. M. & Lowe, D. R. Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature* **431,** 549–552 (2004).
19. Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–535 (2004).
20. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298,** 149–159 (2002).
21. Sorhannus, U. & Fox, M. Synonymous and nonsynonymous substitution rates in diatoms: A comparison between chloroplast and nuclear genes. *J. Mol. Evol.* **48,** 209–212 (1999).
22. Ochman, H. & Wilson, A. C. Evolution in bacteria—evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26,** 74–86 (1987).
23. Clark, M. A., Moran, N. A. & Baumann, P. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol. Biol. Evol.* **16,** 1586–1598 (1999).
24. Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
25. 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).
26. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
27. Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302,** 1960–1963 (2003).
28. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 233–234 (2003).
29. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
30. 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).

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

Acknowledgements S.S. and I.A.A. were supported by the 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 A.S.K. (kondrashov@ncbi.nlm.nih.gov) or S.S. (ssunyaev@rics.bwh.harvard.edu).

Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song

Mimi H. Kao, Allison J. Doupe & Michael S. Brainard

Keck Center for Integrative Neuroscience, Departments of Physiology and Psychiatry, University of California, San Francisco, California 94143-0444, USA

Cortical–basal ganglia circuits have a critical role in motor control and motor learning[1]. In songbirds, the anterior forebrain pathway (AFP) is a basal ganglia–forebrain circuit required for song learning and adult vocal plasticity but not for production of learned song[2–5]. Here, we investigate functional contributions of this circuit to the control of song, a complex, learned motor skill. We test the hypothesis that neural activity in the AFP of adult birds can direct moment-by-moment changes in the primary motor areas responsible for generating song. We show that song-triggered microstimulation in the output nucleus of the AFP induces acute and specific changes in learned parameters of song[6,7]. Moreover, under both natural and experimental conditions, variability in the pattern of AFP activity is associated with variability in song structure. Finally, lesions of the output nucleus of the AFP prevent naturally occurring modulation of song variability. These findings demonstrate a previously unappreciated capacity of the AFP to direct real-time changes in song. More generally, they suggest that frontal cortical and basal ganglia areas may contribute to motor learning by biasing motor output towards desired targets or by introducing stochastic variability required for reinforcement learning.

Song is a complex, learned motor skill that involves the precise coordination of vocal and respiratory musculature in order to produce highly stereotyped renditions of a memorized song model. Two pathways have been identified in the songbird forebrain that contribute to this behaviour (Fig. 1a): (1) a motor pathway that is required throughout life for normal song production[8]; and (2) a basal ganglia–forebrain circuit[9] (the AFP) that is necessary for song learning and plasticity but not for the production of adult song[2–5]. The AFP converges with the motor pathway at the premotor nucleus RA (robust nucleus of the arcopallium). Although lesion studies have demonstrated that the AFP is required for vocal motor plasticity[2–5], its function in motor control and learning remains unclear.

Previous modelling work has hypothesized that a critical contribution of the AFP to song plasticity relies on its capacity to modulate ongoing song[10,11]. Such a modulatory influence might serve to introduce into song the stochastic variability that is necessary for reinforcement learning[11], or to systematically bias song towards desired targets in a manner that eventually becomes encoded in the song motor pathway[10]. Consistent with this hypothesis, the AFP exhibits patterned neural activity in singing birds[12,13]. However, previous studies have failed to demonstrate any direct contribution of the AFP to the production of adult birdsong. In adult zebra finches, lesions of LMAN (lateral magnocellular nucleus of the anterior nidopallium), the output nucleus of the AFP (Fig. 1a), have not been reported to affect ongoing song. Moreover, whereas stimulation in the motor pathway of quiescent birds can elicit vocalizations[14,15], comparable microstimulation in LMAN does not. Finally, a previous study did not report any gross effects of LMAN microstimulation on song[14]. These findings indicate that LMAN is not an obligatory premotor structure for song, but they leave unresolved whether, and how, patterned activity from the AFP modulates song production.

Here, we examine the hypothesis that neural activity in the AFP contributes to vocal control by using microstimulation triggered by real-time song to manipulate activity in LMAN during precisely targeted parts of a song (Fig. 1b). We report that artificially altering the pattern of activity in LMAN induces acute changes in the structure of individual song elements, or 'syllables', without altering the order or structure of ensuing syllables. Such changes include increases and decreases in sound frequency (a learned parameter of song; ref. 6), as well as increases and decreases in sound amplitude (a parameter of song that is precisely controlled; refs 7, 16). Figure 1c illustrates a representative experiment. Every time the bird sang a rendition of a stereotyped sequence of syllables, or 'motif' ('abcdef'), syllable 'b' was detected using a real-time template-matching algorithm. After detection, microstimulation was delivered in LMAN on randomly selected trials (in this case, during the first motif rendition, but not the second). In this experiment, LMAN stimulation caused a systematic downward shift in the fundamental frequency of syllable 'c'. Figure 1d illustrates a second experiment using the same paradigm in which LMAN microstimulation caused a systematic decrease in the loudness of the targeted syllable. Such acute, stimulation-induced changes in syllable structure were observed for 18 of 20 sites in LMAN of five birds (Supplementary Tables 1 and 2).

Evoked changes in syllable structure were tightly locked to the delivery of stimuli. In cases where latency could be assessed, the mean latency between the onset of a stimulus and an effect on syllable structure was 50 ms, but it could be as short as 35 ms (see Supplementary Fig. 1). In addition, the effects of stimulation typically terminated within 60–70 ms after the end of the stimulus train (Fig. 1d). LMAN projects directly to the song motor nucleus RA, which is thought to provide motor commands that control the precise structure of individual song elements with a latency of 40–45 ms[14,17]. Although the exact mechanisms and pathways underlying the influence of LMAN on song remain to be determined, both the rapidity with which LMAN stimulation could drive changes in syllable structure and the rapid termination of its effects are consistent with a direct modulation of RA by LMAN.

In a given experiment, stimulation in LMAN typically induced systematic changes in syllable structure, rather than a general degradation of song structure or an enhancement of song variability. When stimulation elicited a significant shift in the mean fundamental frequency of a syllable, it had no effect on the variability of the fundamental frequency in 59% of cases (note the standard deviation (s.d.) of the histograms in Fig. 2b), increased the variability in 30% of cases (Fig. 1c), and reduced the variability in the remaining 11% of cases. Thus, the predominant effect of artificially imposing a fixed pattern of activity in LMAN during singing was to shift systematically the mean value of syllable parameters, rather than to grossly disrupt song.

There was significant specificity to the changes elicited by microstimulation in LMAN. For 13 of 18 sites, a fixed pattern of stimulation had qualitatively different effects when delivered during different syllables (Supplementary Table 2). Figure 2a shows a case in which stimulation delivered at a fixed site in LMAN caused