LETTERS

Expression profiling in primates reveals a rapid evolution of human transcription factors

Yoav Gilad[1]†, Alicia Oshlack[2], Gordon K. Smyth[2], Terence P. Speed[2,3] & Kevin P. White[1]

Although it has been hypothesized for thirty years that many human adaptations are likely to be due to changes in gene regulation[1], almost nothing is known about the modes of natural selection acting on regulation in primates. Here we identify a set of genes for which expression is evolving under natural selection. We use a new multi-species complementary DNA array to compare steady-state messenger RNA levels in liver tissues within and between humans, chimpanzees, orangutans and rhesus macaques. Using estimates from a linear mixed model, we identify a set of genes for which expression levels have remained constant across the entire phylogeny (~70 million years), and are therefore likely to be under stabilizing selection. Among the top candidates are five genes with expression levels that have previously been shown to be altered in liver carcinoma. We also find a number of genes with similar expression levels among non-human primates but significantly elevated or reduced expression in the human lineage, features that point to the action of directional selection. Among the gene set with a human-specific increase in expression, there is an excess of transcription factors; the same is not true for genes with increased expression in chimpanzee.

A number of recent studies have used DNA microarrays to compare patterns of gene expression between closely related species[2–9]. Within primates, the focus has been primarily on human–chimpanzee comparisons, estimating gene expression profiles for a number of tissues, including liver, brain and heart[2,6,7,10]. The aim has been to characterize general trends in the evolution of gene expression rather than to identify specific genes of interest. To date, conclusions about the selection pressures acting on gene expression have been conflicting[2,3,6,11–13].

These studies have all relied on data collected from arrays using gene probes that were designed on the basis of human sequences only. However, sequence mismatches affect hybridization intensity and can therefore bias estimates of gene expression differences between species[14]. This limitation of single-species arrays is especially problematic when the goal is to study how expression changes over evolutionary time. To make comparisons between more distantly related primate species, we generated a multi-species cDNA array that allows comparison of gene expression between species without the confounding effects of sequence divergence[14]. This cDNA array contains probes for 1,056 orthologous genes from four species (see Supplementary Methods)[14].

We used this array to compare gene expression profiles in the livers of humans, chimpanzees (*Pan troglodytes*), orangutans (*Pongo pygmaeus*) and rhesus macaques (*Macaca mulatta*), the phylogeny of which represents approximately 70 million years (Myr) of evolution. By assigning expression changes in the liver to particular lineages, we were able to identify the first set of genes for which regulation seems to be under lineage-specific selection pressures. In order to measure gene expression levels within and between species, we extracted RNA from liver samples of five adult males from each of the four species. A common reference design was used, with a sixth human liver sample serving as the reference. We performed four technical replicates of each comparison, for a total of 80 hybridizations. Results from all species were obtained for 907 genes, used in subsequent analyses (Supplementary Table S1).

After image analysis, background correction and normalization, the log expression values were analysed using a linear mixed model with fixed effects for species and sequence mismatches, and a random effect for individuals within species (see Methods). For each gene, we used residual maximum likelihood[15] to estimate the fixed effects and variances. Hypothesis testing was performed using likelihood ratio tests (see Methods).

As a first step, we identified genes that are differentially expressed between species (Table 1). A phylogenetic tree based on the number of differentially expressed genes between species[16] recapitulates their known phylogeny (Supplementary Fig. S1). However, the number of significantly differentially expressed genes does not always increase with evolutionary time.

Focusing on human and chimpanzee, we found 110 genes (12%) to be differentially expressed at a false discovery rate (FDR)[17] of 0.01, with a mean absolute log ratio of 1.56-fold difference (Supplementary Table S2). Our observation is in general agreement with a statistical meta-analysis[11] of the data from ref. 2. In contrast to this meta-analysis, however, we find that equal numbers of genes have elevated (55) or reduced (55) expression levels in humans compared to chimpanzees.

To estimate lineage-specific changes in expression levels, we used the expression profiles from orangutan and rhesus macaques as outgroups for 84 of the genes that show significantly different expression between human and chimpanzee (Fig. 1a; see Methods). Using this approach, we found similar numbers of genes for which expression has been altered in either the human or the chimpanzee lineage. Moreover, in both species, the numbers of genes that show increased or decreased expression levels relative to the estimated ancestral expression level is similar (45 and 43 of the genes are upregulated in humans and chimpanzees, respectively). In addition, the average or median fold change in gene expression level is similar regardless of the lineage or the trend (that is, up or down)

Table 1 | Inter-species differentially expressed genes

	Chimpanzee	Orangutan	Rhesus macaque
Human	110	128	176
Chimpanzee	–	150	141
Orangutan	–	–	129

[1]Department of Genetics and Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06510, USA. [2]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [3]Department of Statistics and Program in Biostatistics, University of California, Berkeley, California 94720, USA. †Present address: Department of Human Genetics, University of Chicago, Chicago, Illinois 60605, USA.

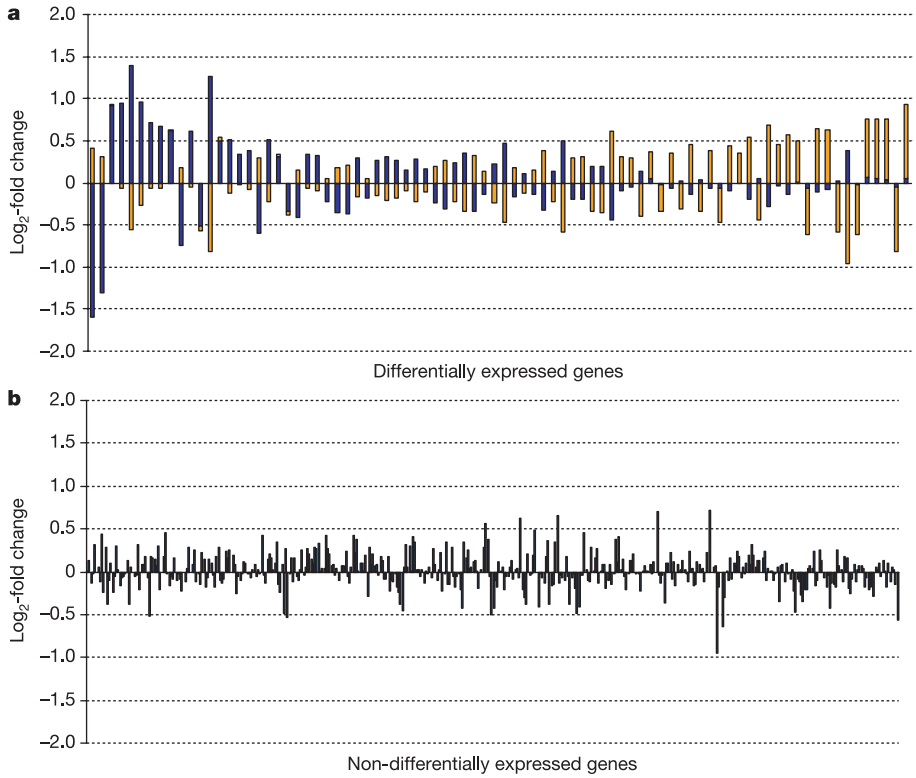

Figure 1 | Expression changes in specific lineages. a, For 84 genes that are differentially expressed between human and chimpanzee, the $\log_2$-fold change relative to the common ancestor is given for the human (blue) and chimpanzee (orange) lineages. Genes are ordered by the ratio of their expression changes in the human lineage compared to the chimpanzee lineage. **b**, For 446 genes that are not differentially expressed between human and chimpanzee, and for which an ancestral state could be estimated (see Methods), the $\log_2$-fold change relative to the common ancestor is shown for the human lineage.

(Supplementary Fig. S2). The pattern also holds for expression changes in the human lineage in genes that are not differentially expressed between human and chimpanzee (Fig. 1b; 52% of the genes were upregulated). These observations do not agree with previous studies[2,10]. Possible explanations for the discrepancy are the use of human microarrays for inter-primate comparisons[2], or the assignment of expression changes to lineages in the absence of outgroup data[10].

Our approach also allows the identification of genes for which regulation is likely to have evolved under stabilizing selection. Previous studies have done this by testing for deviations from neutrality or stabilizing selection[13,16,18]. Such an approach requires a model for the evolution of expression, and thus relies on a number of parameter estimates about which there is considerable uncertainty in primates (for example, the neutral expression change per generation, and the environmental and mutational variance for each gene). Instead of specifying an explicit model, we used statistical analyses to rank genes according to their pattern of evolutionary change among the four species, and focused on those at the top of the list as the most promising candidates.

First, we identified genes that best fitted a model of constant expression level throughout the phylogeny, reasoning that these represent promising candidates for stabilizing selection. A majority of the genes on the array (60%) do not show significant inter-species expression differences. However, failure to reject the null hypothesis of no expression difference between species can result from constant expression level in all individuals in all species (Fig. 2a) or large within-species variance (Fig. 2b)—especially as primate tissues cannot be staged[10]. As our aim is to identify genes under stabilizing selection, we are only interested in the former scenario. We therefore ranked genes by their expression variation among individuals across all species (see Methods). Genes at the top of our list are not significantly differentially expressed between species, and also have low within-species variance (Fig. 2). The expression levels of these genes seem to have remained constant for ~70 Myr[19], suggesting that their regulation is under evolutionary constraint. Among the first 100 genes on our list (Supplementary Table S3), the most significant enrichment ($P < 10^{-8}$; uncorrected for multiple tests) is for genes from the category 'regulation of cellular physiological process' (Gene Ontology ID 0051244; http://www.geneontology.org). As we expect transcription of such genes to be similar across individuals and species, this finding serves as a validation of the approach.

A number of recent papers have argued that the majority of expression differences observed between primates are neutral, based primarily on the observation that the mean square fold change in expression levels in liver and brain increases linearly with species divergence time[6,12]. Having found no clear increase in the number of significantly differentially expressed genes with time (Table 1), we re-examined the mean square fold change for our data. This revealed no linear increase over time (Supplementary Fig. S3). Moreover, our observation that many genes show stable expression levels over 70 Myr suggests that, rather than evolving mostly neutrally, expression levels are often under stabilizing selection, consistent with findings in *Drosophila*[16,18] and in *C. elegans*[20].

This finding has implications for studies of human disease. Indeed, our observations suggest that many changes in gene regulation may be deleterious and hence influence disease susceptibility. Consistent with this, among the top 100 genes for which regulation is probably evolving under stabilizing selection, genes associated with human cancer are slightly enriched (9% compared to 5% in the total gene sample; $P = 0.10$, one-tailed Fisher's exact test). Moreover, the expression levels of five genes (*MBD4*, *WWOX*, *ING1*, *ATP7B* and *IGFBP2*; ranked 5th, 12th, 28th, 58th and 66th, respectively) have been shown to be altered specifically in liver carcinoma[21–24]. These

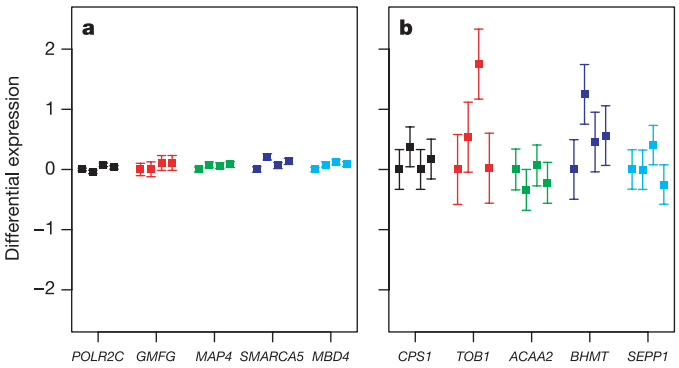

Figure 2 | Genes that are not differentially expressed across species. In each plot, different genes (*x*-axis) are represented by different colours. For each gene, the estimated expression level (±s.e.m.) is shown for humans, chimpanzees, orangutans and rhesus macaque (left to right). **a**, The five highest-ranked genes (see Methods). These genes have constant expression levels in all species, suggesting that their expression levels are under stabilizing selection. **b**, Examples of genes that are not differentially expressed across species, probably due to high within-species variance (gene rankings 489–493).

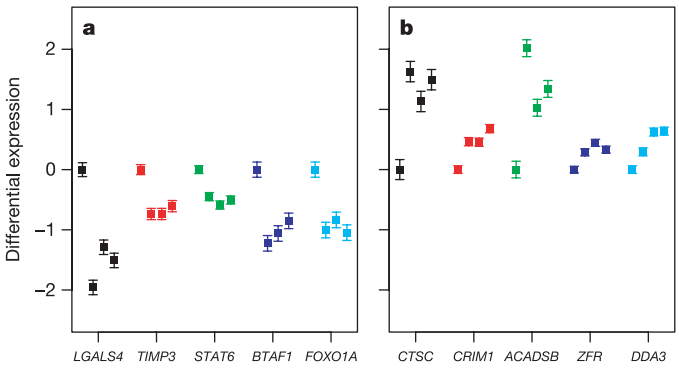

Figure 3 | Genes with distinct expression pattern in humans. Different genes (*x*-axis) are represented by distinct colours. For each gene, the $\log_2$ expression levels for humans are set to zero. Estimated gene expression level relative to human (±s.e.m.) is shown for humans, chimpanzees, orangutans and rhesus macaque (left to right). Shown are examples of five genes that are not differentially expressed in the non-human primates but are upregulated (**a**) or downregulated (**b**) in humans. The expression levels of these genes seem to have been under stabilizing selection in the non-human primates and under directional selection in the human lineage.

findings suggest that focusing on genes with conserved expression levels among primates may be helpful in identifying promising candidates for disease-association studies, much like phylogenetic shadowing of DNA sequences[25] can aid in the identification of non-coding elements of functional importance.

Using this general approach, we also identified genes for which expression levels are not significantly different among non-human primates but are significantly elevated or reduced in humans relative to each of the three other species (see Methods and Supplementary Table S4). In other words, the expression level of the gene has remained similar over ~65 Myr of evolution and then changed over the ~5 Myr of the human lineage, indicative of directional selection in humans. Our analysis revealed 14 genes with significantly higher expression levels in humans and five with lower expression (Fig. 3). We note that we are likely to be missing a number of targets of positive selection: gene expression varies across tissues and developmental stages[26], and as a result, the absence of support for selection in primate expression data is weak evidence against it.

Notably, among the genes with higher expression in humans, we find a significant excess of transcription factors (5/12, 42% compared with 10% representation on the array; $P = 0.003$ by Fisher's exact test, including all genes for which GO annotation was available), whereas no transcription factors were found among genes with unusually low expression in humans. We repeated this analysis using a less stringent criterion to identify genes for which the mean expression level in humans differed significantly from that of non-human primates (see Methods). Again, transcription factors were overrepresented among the 30 genes with elevated expression in humans (30%; $P = 0.001$, Fisher's exact test), and no transcription factors were found among 19 genes with reduced expression. In contrast, when these analyses were applied to chimpanzee (Supplementary Table S5), the number of transcription factors was equivalent among genes with elevated (9%) or reduced (9%) expression levels (for the less stringent cutoff), and neither proportion was significantly different from the overall representation on the array (that is, 10%). It is unlikely that these observations can be explained by differential degradation of transcripts encoding specific classes of proteins[27], as no difference in RNA quality was observed between human and non-human primate samples during sample preparation (on the basis of electrophoretic analyses).

In addition to the rapid evolution of expression levels, genes encoding transcription factors have also been shown to evolve rapidly in the human lineage at the coding sequence level[28]. Together, these findings raise the possibility that the function and regulation of transcription factors have been substantially modified in the human lineage, potentially affecting many downstream targets over a short evolutionary time frame. Notably, the opposite finding emerged from studies of closely related *Drosophila* species, in which the expression levels of transcription factors were shown to evolve slower than genes encoding other types of proteins[16,18]. Given the large number of phenotypic changes in the human lineage[1], it is tempting to speculate that relative rates of transcription factor evolution may serve as an indicator of rates of phenotypic evolution at the organismal level.

Finally, to examine the extent to which evolution of protein-coding regions mirrors gene expression level changes in the liver, we considered three sets of genes: those for which expression levels seem to be under directional selection in humans (set A), the top 100 candidates for stabilizing selection (set B) and the remaining genes (set C). To assess the evidence for natural selection acting on coding regions, we used estimates of the posterior probability that a gene is subject to positive or negative selection based on synonymous and non-synonymous nucleotide polymorphism and divergence levels at genes on our array[28]. Using this approach (with a posterior probability of 0.05), only 6% of genes in set C and 4% in set B are inferred to evolve under positive selection. In contrast, among set A, significantly more genes (25%) are inferred to evolve under positive selection ($P = 0.03$, one-tailed Fisher's exact test). These observations suggest that genes with expression levels under directional selection in humans are somewhat more likely to show accelerated amino acid evolution.

In summary, the use of a new multi-species cDNA array has allowed us to identify a set of genes with regulation under natural selection in humans. In particular, the over-representation of transcription factors among the genes with modified expression levels in the human lineage is consistent with the suggestion that most differences between human and chimpanzee are due to changes in gene regulation[1], and might provide insight into their genetic architecture.

METHODS

Study design and analysis. The 80 arrays were scanned using a GenePix Axon scanner and data were extracted using GenePix 6 (Molecular Devices) to give Cy5 and Cy3 foreground and background fluorescence intensities. Analysis was done in the R computing environment (http://www.r-project.org). Background-corrected Cy5 and Cy3 intensities were produced using the 'normexp' method

with an offset of 50, implemented in the limma software package[29]. Lowess curves for intensity-dependent normalization were generated in a way similar to ref. 14, where probes from the two species involved in the hybridization were used to fit the curves. All probes on the array were adjusted by the fitted lowess curve (see Supplementary Methods). We concentrated on the 907 genes on the array for which successful polymerase chain reaction (PCR) products were obtained from all species[14]. The expression log ratios for each gene were analysed using the linear mixed model:

$$y_{tijp} = \mu_t + \kappa_{tp} - \kappa_{hp} + \alpha_{ti} + \varepsilon_{tijp}$$

in which we have suppressed the gene labels. Here, y_{tijp} is the normalized $\log_2$ ratio measured for target species t for replicate j of individual i on species probe p. The term μ_t is the expected log ratio of the expression level of the gene in target species t relative to the human reference, and κ_{tp} and κ_{hp} are parameters corresponding to the reduction in the log expression levels caused by reduced affinity owing to target and probe sequence mismatches. As each hybridization has target species t on the red channel and the human reference on the green channel, there are two κ terms for each measurement. We assume that κ_{tt} is equal to 0, and that the affinity adjustments are symmetrical in target and probe (that is, $\kappa_{tp} = \kappa_{pt}$). The term α_{ti} is the random effect for individual i of species t, assumed to be uncorrelated with mean zero and variance σ^2_α. Finally, ε_{tijp} is the residual error term, and these are assumed to be uncorrelated with mean zero and variance σ^2_ε. We also considered models that included random effects for probes within arrays and a crossed term for an array × probe interaction, but found that the contributions from these terms were substantially smaller than the error term and therefore did not warrant inclusion in the model. (See Supplementary Information for further details on the parameters and model.) For each gene, the model was fitted by residual maximum likelihood using statmod and lme software packages[30].

Hypothesis testing. Likelihood ratio tests were used for hypothesis testing. Under the full model, for each gene, 12 parameters (4 μ_t parameters, 6 κ_{tp} parameters, σ^2_α and σ^2_ε) were estimated by maximum likelihood. Genes deemed to be under stabilizing selection were those for which the fit of a reduced model with $\mu = \mu_h = \mu_c = \mu_o = \mu_r$ was adequate (h, human; c, chimpanzee; o, orangutan; r, rhesus macaque). Such genes were selected on the basis of the likelihood ratio test statistic comparing the fit under this sub-model to that under the full model. Under the null hypothesis, −2(log-likelihood ratio) has an approximate χ^2 distribution on 3 degrees of freedom, and genes for which this statistic was less than 12.4 ($P = 6.1 \times 10^{-3}$) were chosen. We then ranked these genes according to the magnitude of the between-to-within individual ratio mean squares $\left(16\hat{\sigma}^2_\alpha + \hat{\sigma}^2_\varepsilon\right)/\hat{\sigma}^2_\varepsilon$ starting with genes for which this was small. We note that the latter process alone would not suffice to identify genes that are not differentially expressed between species (Supplementary Fig. S4).

To select genes that were different in human compared to the other three species, we combined three criteria. First, we used a likelihood ratio statistic to exclude genes that were differentially expressed in the non-human primates. We maximized the likelihood under the constraints $\mu = \mu_c = \mu_o = \mu_r$, constructed the ratio of this likelihood compared to the full model, and removed genes where we estimated significant differences. Second, we used a likelihood ratio statistic to rank genes on the basis of differences between human and the other species (that is, $\mu \neq \mu_h$). We chose a cutoff statistic of 16 ($P = 6.3 \times 10^{-5}$) to select genes, but also investigated genes selected under a more relaxed cutoff of 12, which corresponds to ~1% FDR[17]. Third, we restricted the list to genes with small between relative to within individual variance. Pairwise differences between species were also constructed using a likelihood ratio statistic with a cutoff chosen to give 1% FDR, assuming a χ^2_1 distribution (numbers are given in Table 1). We found by simulations that the null likelihood ratio test statistic was well approximated by a χ^2 distribution, implying that our assumptions are accurate (data not shown).

We note that correlation between species due to shared phylogeny is not expected to influence our results, as no structure is imposed on the parameters for the means of the different species and no model is fitted to them across species.

Received 5 November; accepted 29 December 2005.

1. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
2. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296,** 340–343 (2002).
3. Caceres, M. *et al.* Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl Acad. Sci. USA* **100,** 13030–13035 (2003).
4. Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. & Hartl, D. L. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300,** 1742–1745 (2003).
5. Karaman, M. W. *et al.* Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Res.* **13,** 1619–1630 (2003).
6. Khaitovich, P. *et al.* A neutral model of transcriptome evolution. *PLoS Biol.* **2,** E132 (2004).
7. Khaitovich, P. *et al.* Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14,** 1462–1473 (2004).
8. Rise, M. L. *et al.* Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res.* **14,** 478–490 (2004).
9. Nuzhdin, S. V., Wayne, M. L., Harmon, K. L. & McIntyre, L. M. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* **21,** 1308–1317 (2004).
10. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309,** 1850–1854 (2005).
11. Hsieh, W. P., Chu, T. M., Wolfinger, R. D. & Gibson, G. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* **165,** 747–757 (2003).
12. Khaitovich, P., Paabo, S. & Weiss, G. Toward a neutral evolutionary model of gene expression. *Genetics* **170,** 929–939 (2005).
13. Lemos, B., Meiklejohn, C. D., Caceres, M. & Hartl, D. L. Rates of divergence in gene expression profiles of primates, mice, and flies: stablizing selection and variability among functional categories. *Evolution Int. J. Org. Evolution* **59,** 126–137 (2005).
14. Gilad, Y., Rifkin, S. A., Bertone, P., Gerstein, M. & White, K. P. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* **15,** 674–680 (2005).
15. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545–554 (1971).
16. Rifkin, S. A., Kim, J. & White, K. P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genet.* **33,** 138–144 (2003).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
18. Rifkin, S. A., Houle, D., Kim, J. & White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **438,** 220–223 (2005).
19. Glazko, G. V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20,** 424–434 (2003).
20. Denver, D. R. *et al.* The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nature Genet.* **37,** 544–548 (2005).
21. Park, S. W. *et al.* Frequent downregulation and loss of WWOX gene expression in human hepatocellular carcinoma. *Br. J. Cancer* **91,** 753–759 (2004).
22. Sugeno, H. *et al.* Expression of copper-transporting P-type adenosine triphosphatase (ATP7B) in human hepatocellular carcinoma. *Anticancer Res.* **24,** 1045–1048 (2004).
23. Zhu, Z. *et al.* Inhibitory effect of tumour suppressor $p33^{ING1b}$ and its synergy with p53 gene in hepatocellular carcinoma. *World J. Gastroenterol.* **11,** 1903–1909 (2005).
24. Chiba, T. *et al.* Identification and investigation of methylated genes in hepatoma. *Eur. J. Cancer* **41,** 1185–1194 (2005).
25. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
26. Ludwig, M. Z. *et al.* Functional evolution of a *cis*-regulatory module. *PLoS Biol.* **3,** e93 (2005).
27. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13,** 1863–1872 (2003).
28. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437,** 1153–1157 (2005).
29. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) 397–420 (Springer, New York, 2005).
30. Pinheiro, J. C. & Bates, D. M. *Mixed-Effects Models in S and S-PLUS* (Springer-Verlag, New York, 2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. E. Holt for pointing out a possible technical explanation for the excess of upregulated transcription factors in humans, the Yale hospitals, S. Paabo and the Yerkes Primate Center for providing samples used in the study, and A. Clark and M. Przeworski for comments on the manuscript. This research was supported by grants to K.P.W. from the W. M. Keck Foundation, the Arnold and Mabel Beckman Foundation and the National Human Genome Research Institute of the National Institutes of Health, and by NHMRC grants to G.K.S., T.P.S. and A.O. Y.G. was supported by an EMBO fellowship.

Author Information Expression data from this study have been deposited in the GEO database under the series accession number GSE2569. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.G. (gilad@uchicago.edu) or K.P.W. (kevin.white@yale.edu).