LETTERS

Natural selection on protein-coding genes in the human genome

Carlos D. Bustamante[1], Adi Fledel-Alon[1], Scott Williamson[1], Rasmus Nielsen[1,2], Melissa Todd Hubisz[1], Stephen Glanowski[3], David M. Tanenbaum[3], Thomas J. White[4], John J. Sninsky[4], Ryan D. Hernandez[1], Daniel Civello[4], Mark D. Adams[5], Michele Cargill[4]* & Andrew G. Clark[6]*

Comparisons of DNA polymorphism within species to divergence between species enables the discovery of molecular adaptation in evolutionarily constrained genes as well as the differentiation of weak from strong purifying selection[1–4]. The extent to which weak negative and positive darwinian selection have driven the molecular evolution of different species varies greatly[5–16], with some species, such as *Drosophila melanogaster*, showing strong evidence of pervasive positive selection[6–9], and others, such as the selfing weed *Arabidopsis thaliana*, showing an excess of deleterious variation within local populations[9,10]. Here we contrast patterns of coding sequence polymorphism identified by direct sequencing of 39 humans for over 11,000 genes to divergence between humans and chimpanzees, and find strong evidence that natural selection has shaped the recent molecular evolution of our species. Our analysis discovered 304 (9.0%) out of 3,377 potentially informative loci showing evidence of rapid amino acid evolution. Furthermore, 813 (13.5%) out of 6,033 potentially informative loci show a paucity of amino acid differences between humans and chimpanzees, indicating weak negative selection and/or balancing selection operating on mutations at these loci. We find that the distribution of negatively and positively selected genes varies greatly among biological processes and molecular functions, and that some classes, such as transcription factors, show an excess of rapidly evolving genes, whereas others, such as cytoskeletal proteins, show an excess of genes with extensive amino acid polymorphism within humans and yet little amino acid divergence between humans and chimpanzees.

Of considerable interest to medical geneticists is the extent to which weak negative selection has shaped the human genome. Sensitivity to weak selection may be important in identifying candidate genes for association mapping studies, because weakly deleterious mutations can reach appreciable frequencies in local populations and, thus, may contribute significantly to genetic variance in disease susceptibility. A team at Celera Genomics sequenced by exon-specific polymerase chain reaction (PCR) amplification 20,362 loci in 20 European Americans, 19 African Americans and one male chimpanzee with the initial intention of finding novel non-synonymous single nucleotide polymorphisms (SNPs) based on their 2001 build of the human genome. Here we analyse 11,624 genes with complementary DNA support, strong evidence of orthology between humans and chimpanzee, and location in non-repetitive elements of the genome. The distributions of the total number of synonymous and non-synonymous SNPs and fixed differences for these loci are presented in Fig. 1a. We found that 10,767 genes (92.6%) showed some form of coding nucleotide variability either within human subjects or between humans and a chimpanzee. A total of 34,099 fixed synonymous differences between all humans in our sample and the chimpanzee yield a genomic average synonymous divergence of $\overline{d_S} = 1.02\%$. Correspondingly, we found 20,467 non-synonymous differences ($\overline{d_N} = 0.242\%$) across 11.81 megabases (Mb) of aligned coding DNA. We also discovered 15,750 synonymous and 14,311 non-synonymous SNPs among the human subjects, yielding average synonymous and non-synonymous SNP densities of $\overline{p_S} = 0.470\%$ and $\overline{p_N} = 0.169\%$. We note that the ratio of non-synonymous to synonymous differences (23.76%) is smaller than the ratio of non-synonymous to synonymous polymorphisms (38.42%), indicating a highly significant excess of amino acid variation relative to divergence ($\chi^2 = 816.03$, $P < 2 \times 10^{-16}$). This is consistent with previous studies of human polymorphism that suggest that a large proportion of amino acid variation in the human genome is slightly to moderately deleterious[11–16].

In order to identify particular genes with evidence of non-neutral evolution, we applied a statistical approach that makes efficient use of comparative population genomic data[4,7,9,17]. The method quantifies the extent and directionality of selection operating on a given gene in terms of the population genetic selection parameter $\gamma = 2N_e s$ (where N_e is the effective population size and s is the selection coefficient in a Wright–Fisher genic selection model) as estimated from contingency tables comparing polymorphism (P) versus divergence (D) at synonymous (S) versus non-synonymous (N) sites[3,4]. The parameter is negative if a gene shows an excess of amino acid polymorphism (or paucity of amino acid divergence) and positive if a gene has an excess of amino acid divergence relative to the genomic average for synonymous sites. Invariant sites in an alignment do not affect the estimate of γ, so it is independent of strong purifying selection operating at the locus. One consequence of this robustness is that genes with little or no variation contain little information regarding γ. We therefore concentrate on two subsets of the data: the $n = 3{,}277$ genes with at least four variable non-synonymous sites in the alignment (that is, $P_N + D_N \geq 4$), which are potentially informative about positive selection (IPS data), and the set with at least two variable non-synonymous sites, which are potentially informative only about negative selection (INS data; $n = 6{,}033$). In order to classify genes, we use the posterior distribution of the selection coefficient to calculate a test statistic, such that $P^+ = P_r\{\gamma > 0|Data\}$ is the posterior probability that a gene is subject to positive darwinian selection and $P^- = P_r\{\gamma < 0|Data\}$ is the posterior probability that a gene harbours excess amino acid variation. (Negative

[1]Department of Biological Statistics and Computational Biology, 101 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [2]Center for Bioinformatics, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. [3]Applied Biosystems, 45 West Gude Drive, Rockville, Maryland 20850, USA. [4]Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, California 94502, USA. [5]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [6]Department of Molecular Biology and Genetics, 227 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA.
*These authors contributed equally to this work.

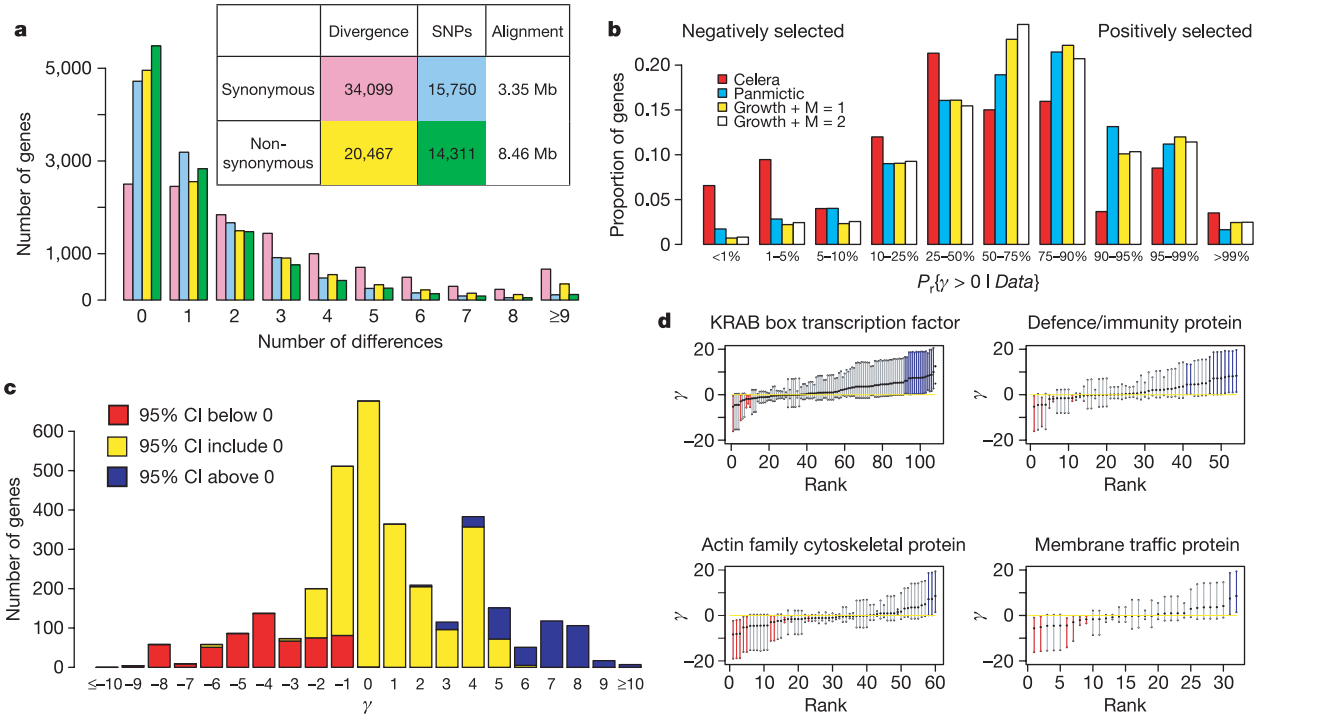

Figure 1 | Summary distributions of McDonald–Kreitman cell entries and mkprf analyses. **a**, Distributions of synonymous and non-synonymous SNPs and fixed differences across 11,624 genes. **b**, Distribution of the posterior probability that a gene is positively selected for simulated data under three neutral demographic scenarios and for the Celera data conditioning on a gene having at least four variable amino acid positions (IPS data). **c**, Distribution of the estimated average selection coefficient for the 3,277 genes in the Celera IPS data set. **d**, Distribution of γ for four molecular functions showing an excess of non-neutral loci. Bars represent 95% CIs with blue and red denoting CIs completely above or below $\gamma = 0$.

Table 1 | Molecular functions and biological processes showing as an excess of positively or negatively selected genes

Category	*P*-value	Number where CI > 0	Number where CI < 0	*N*
Biological process				
Apoptosis	0.00336	12	10 (5)	99 (53)
Cell structure and motility	**0.00008**	**4**	**27 (8)**	**176 (101)**
Ectoderm development	**0.02805**	**1**	**12 (7)**	**98 (61)**
Gametogenesis	0.03411	5	4 (1)	41 (23)
General vesicle transport	**0.00016**	**0**	**14 (4)**	**40 (20)**
Intracellular protein traffic	**0.01151**	**8**	**32 (10)**	**159 (83)**
mRNA transcription	0.00002	29	34 (17)	333 (185)
Natural-killer-cell-mediated immunity	0.03299	1	3 (2)	19 (9)
Nucleoside, nucleotide and nucleic acid metabolism	0.00467	38	65 (28)	568 (311)
Sensory perception	0.04577	9	9 (4)	101 (56)
Molecular function				
Actin family cytoskeletal protein	**0.00008**	**4**	**23 (10)**	**104 (60)**
Cytoskeletal protein	**0.00000**	**7**	**36 (12)**	**205 (118)**
Defence/immunity protein	0.00965	10	12 (5)	89 (54)
Extracellular matrix	**0.01478**	**3**	**15 (11)**	**103 (78)**
Homeotic transcription factor	0.02586	2	2 (0)	28 (8)
Immunoglobulin receptor family member	0.04558	7	3 (1)	36 (27)
Kinase modulator	**0.00538**	**0**	**9 (3)**	**28 (14)**
KRAB box transcription factor	0.00004	17	10 (5)	168 (108)
Membrane traffic protein	**0.02701**	**2**	**17 (6)**	**70 (32)**
Microtubule-binding motor protein	**0.03109**	**1**	**7 (1)**	**31 (19)**
Microtubule family cytoskeletal protein	**0.01373**	**2**	**9 (1)**	**52 (34)**
Non-motor actin-binding protein	**0.01691**	**3**	**12 (3)**	**58 (27)**
Nuclear hormone receptor	0.00143	3	0 (0)	10 (7)
Protein kinase	0.04366	6	4 (1)	94 (36)
Receptor	0.00211	31	37 (18)	343 (207)
RNA helicase	**0.03948**	**0**	**8 (4)**	**33 (18)**
Transcription factor	0.00000	39	41 (19)	428 (240)
Voltage-gated potassium channel	**0.03943**	**0**	**5 (2)**	**23 (11)**
Zinc finger transcription factor	0.00074	20	19 (9)	229 (141)

Classification is based on Panther classification. The *P*-value is the uncorrected value from the Mann–Whitney *U*-test. A total of 139 different molecular functions and 133 biological processes were tested; non-significant categories are listed in Supplementary Table 1, as are significant categories with considerable overlap with those shown in the Table. Parentheses for all count data denote the IPS data set. Bold text indicates negatively selected genes; non-bold text indicates positively selected genes.

selection on slightly deleterious mutations increases the proportion of non-synonymous polymorphisms relative to the proportion of non-synonymous fixed differences.)

Figure 1b, c shows histograms of the distribution of P^+ and γ among genes in the IPS data. In order to assess how deviations from the assumptions of our model may influence our conclusions, we simulated 10,000 replicate data sets under each of three different demographic models (See Supplementary Data 1 for details) and plotted the distribution of P^+ for the corresponding IPS data sets in these simulations. The Celera data have a clear excess of genes with high and low posterior probabilities (that is, there are too many genes in the <1%, 1–5% and >99% categories) regardless of which demographic model is used as the null hypothesis. The signature is particularly strong for weak negative selection. Figure 1c shows the distribution of γ among genes in the IPS data set (estimated as the posterior mean of the selection parameter for each gene) and Fig. 1d shows the posterior mean and credibility intervals (CIs) for genes in four Panther classifications enriched for positively or negatively selected genes (see Table 1 and Supplementary Table 1 for details).

In Fig. 2 we present an amino acid selection map of the human genome with each locus mapped onto its genomic location and coloured according to the strength of evidence for positive or

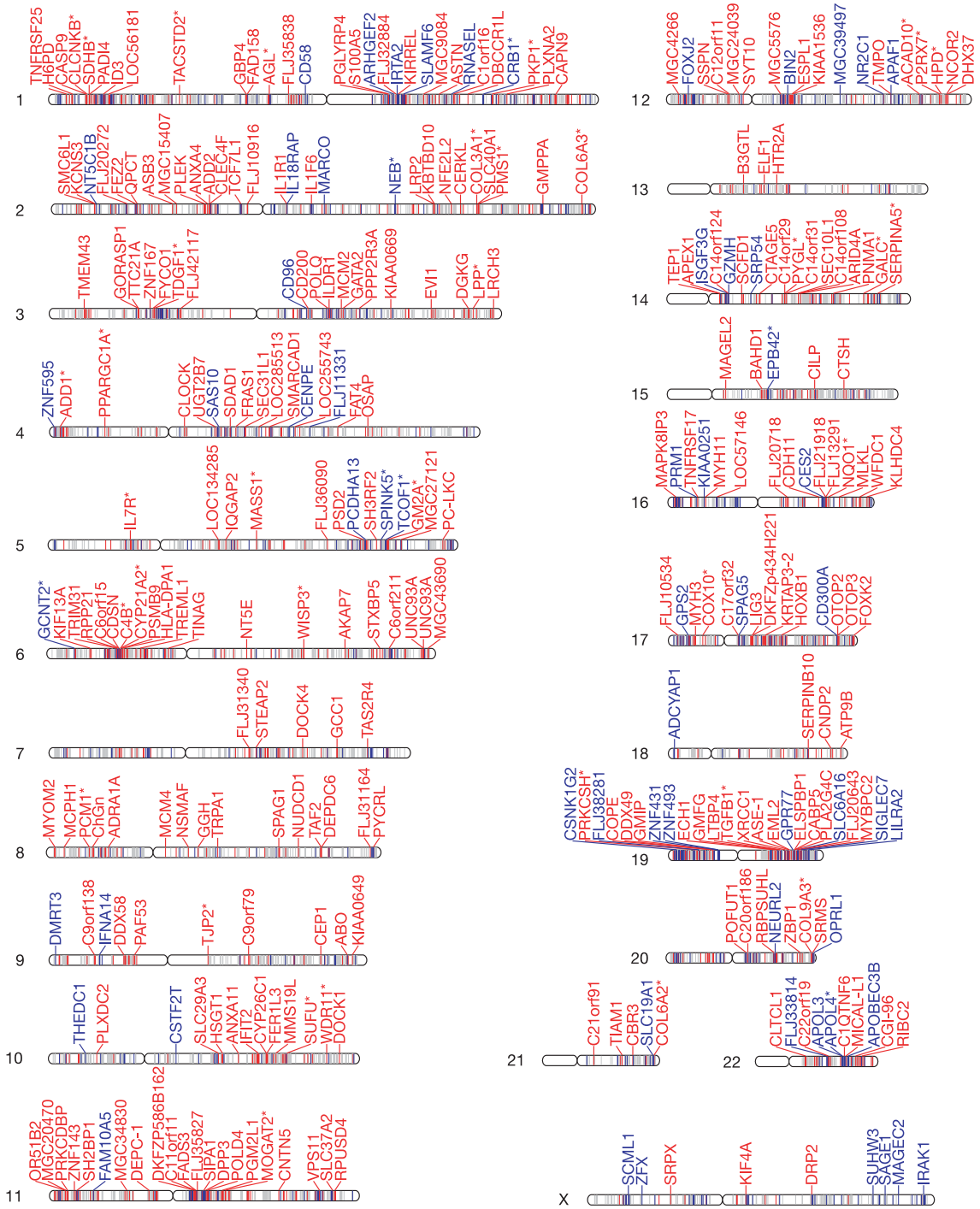

Figure 2 | A selection map of the human genome. Red bars indicate loci under negative selection and blue bars are loci under positive selection at 95% credibility level. Loci with very strong evidence of selection (>99% credibility) are denoted by their HUGO name, and within this category genes with a morbidity entry in the OMIM database are denoted by an asterisk.

negative selection. Although we find that most of the genes in the informative data sets ($n = 4{,}916$) show no evidence of selection according to our methods, we do classify 813 loci as significantly negatively selected and 304 as positively selected at a 5% cutoff. Of the 50 loci identified by ref. 18 as rapidly evolving, 45 were informative about positive selection in our data set. Of these, 14 (31.1%) had more than 95% of their posterior mass above $\gamma = 0$, and 37 (82.2%) had a majority of their posterior mass above neutrality, indicating good agreement between population genetic and phylogenetic approaches for identifying rapidly evolving genes. There is a high degree of overlap in the types of genes classified as positively selected by both studies (see Table 1 and Supplementary Table 1), including defence/immunity proteins ($P = 0.00965$), gametogenesis ($P = 0.03411$), apoptosis ($P = 0.00336$) and sensory perception ($P = 0.04577$). One interesting result of our analysis is that transcription factors as a group seem to be rapidly evolving ($P < 0.0001$), with 39 out of 240 in the informative data sets (16.25%) having P^+ greater than 97.5% as compared to 9.6% of all loci in the IPS data set. Similarly, we find evidence that the categories of nuclear hormone receptors ($P = 0.00143$) and genes involved in nucleoside, nucleotide and nucleic acid metabolism ($P = 0.00467$) have an excess of rapidly evolving genes.

We also used our approach to identify loci and classes of genes that show a paucity of amino acid divergence between humans and chimpanzees, yet have moderate to high levels of amino acid polymorphism, which we believe to harbour an excess of mildly deleterious variation. For example, loci involved in actin binding ($P = 0.00013$; Supplementary Table 1) and cytoskeletal formation ($P < 0.00001$) contain an over-representation of negatively selected loci, with 36 out of 205 cytoskeletal proteins in the informative set (17.6%) having P^- greater than 97.5% (Table 1). For example, 6 out of the 9 myosin heavy chain loci exhibit a large excess of amino acid polymorphism, including non-muscle myosin (*MYH9*, $P^- > 0.983$), embryonic (*MYH3*, $P^- > 0.999$), perinatal (*MYH8*, $P^- > 0.962$) and adult skeletal (*MYH4*, $P^- > 0.946$; *MYH13*, $P^- > 0.957$) myosin as well as the smooth muscle (*MYH11*, $P^- > 0.999$) form. Other cytoskeletal proteins with excess amino acid polymorphism within human populations include myomesin 2 and 3 (*MYOM2*, *MYOM3*), dystrophin related protein 2 (*DRP2*), α- and β-adducin (*ADD1*, *ADD2*), sarcospan (*SSPN*) and scinderin (*SCIN*). These results are consistent with the fact that as a group, genes involved in cell structure and motility (Table 1) show a signature of negative or purifying selection ($P = 0.00008$), with 27 out 176 (15.3%) loci exhibiting excess amino acid polymorphism relative to divergence.

Mutations in cytoskeletal protein-coding genes are known to cause a number of mendelian diseases and have been implicated in various complex disorders. For example, with the dystrophin gene many different types of mutation are known to cause both Duchenne and Becker types of muscular dystrophy (*DMD*, $P^- > 0.969$). Also in this set is myosin VIIA (*MYO7A*, $P^- > 0.99$), a gene implicated in Usher syndrome (1B), the most common cause of congenital deafness and blindness in developed countries. Similarly, the α- and β-adducin genes ($P^- > 0.99$ for both) are associated with hypertension and cardiovascular disease, and one known causative variant (*ADD1* G460W (refs 19, 20)) is found at moderate frequencies in both African Americans (9.7%) and European Americans (28.9%) in our sample.

Another interesting group of genes that show excess amino acid polymorphism ($P = 0.02805$) are those involved in ectoderm development, with 12 loci out of 98 exhibiting significantly elevated levels of amino acid polymorphism relative to amino acid divergence (Table 1). These include three loci in which mutations are known to cause disease: *GLI3* (polydactyly), *NOTCH3* (*Drosophila* homologue implicated in cerebral arteriopathy) and *DCC* (colorectal carcinoma). Interestingly, all three of these genes are also involved in neurogenesis according to the Panther molecular function classification. Genes involved in general vesicle transport also show excess amino acid polymorphism as a class ($P = 0.00016$). Sixteen loci have posterior distributions with more than 95% and four with more than 99% mass above $\gamma = 0$ (for example, *COPE*, *HD*, *KIF4A*, *SEC31L1* and *STX11*). The most familiar of these is *HD*—the gene that causes Huntington's disease.

To quantify the strength of association between non-neutral evolution and genetic disease more formally, we undertook two analyses. The first is a logistic regression of Online Mendelian Inheritance in Man (OMIM) morbidity status (1 = disease; 0 = non-disease) on the square-root of observed species non-synonymous substitution rate $\sqrt{d_N}$ for all loci in the IPS data contained in the OMIM database of disease and non-disease genes. This approach (which is independent of the Poisson random field (PRF) analysis and summarized in Fig. 3a) yields a highly significant negative association ($\beta_{\sqrt{d_N}} = -7.6941; \beta_0 = -0.4555$) between rate of amino acid evolution and disease status for moderate to highly polymorphic loci (likelihood-ratio test (LRT) = 14.8; $P_{\chi^2_1} = 1.1 \times 10^{-4}$; $n = 1{,}438$ loci), suggesting that mendelian disorders may have discernable darwinian consequences. A second test of the same hypothesis, a Mann–Whitney *U*-test comparing disease versus non-disease loci, also indicates a marginally significant excess of genes with high P^- among genes contributing to mendelian disease ($P = 0.0374$; see Fig. 3b).

The results of analysis of human polymorphism at this scale may help to guide our thinking about human evolution at a broad level, and they also help sharpen our targets for exploration of the genetic basis for medical conditions. Examination of Figs 1 and 2 shows that only a small minority of non-neutral genes are facing positive selection. In this group are genes that might be responsible for

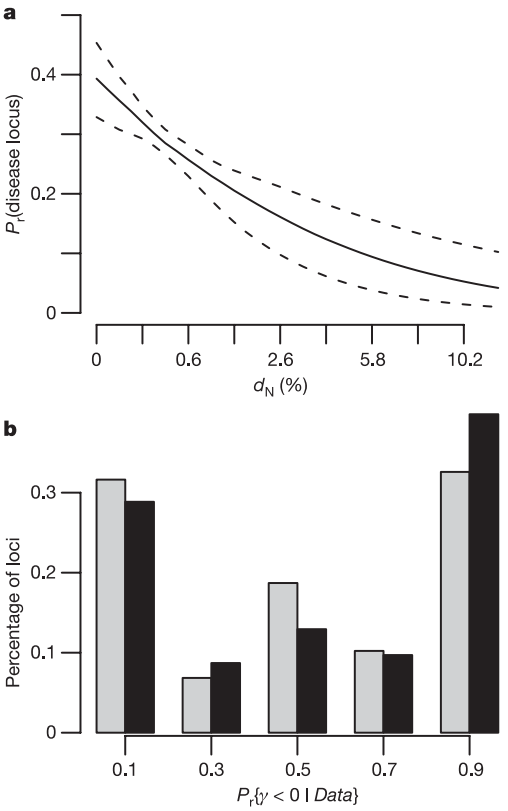

Figure 3 | Association between negative selection and disease. a, Results of a logistic regression analysis of OMIM status (disease versus non-disease) in the IPS data set as a function of the per site amino acid substitution rate. 95% confidence intervals are found via non-parametric bootstrap re-sampling. **b**, Histograms for $P^- = P_r\{\gamma < 0|Data\}$, the probability that non-lethal mutations at a gene are negatively selected, across disease (black bars) and control (non-disease; grey bars) genes.

adaptive differences between humans and chimpanzees, or for which molecular evolution might be driven by pathogen pressure. Each of these genes is a candidate for a deeper analysis of the evolutionary past of our species. Most genes in Fig. 2 show a signature of weak negative selection, a feature that is detected only if there are segregating variants that have deleterious consequences. This implies that this group constitutes a set of genes of keen interest in medical genetics. Notably, a substantial portion of these genes is also already known to be mutable to a mendelian disorder. These genes remain prime candidates in our quest for understanding the genetic basis for a diverse array of complex disorders that show weak familial tendencies.

METHODS

Sequencing and bioinformatics. Celera Genomics applied exon-specific PCR amplification to 20,362 loci in 39 humans and one male chimpanzee in order to obtain sequence variants in these regions (details are provided in the Supplementary Methods). Of these, 14,032 genes shared at least 95% sequence identity to a unique accession in the NM series of the NCBI Reference Sequence (RefSeq) 9.0 database (3/1/2005) and mapped onto build 34 of the human genome at the same genomic location as the accession (mapping was done using BLAT[21] v. 29). The loci sequenced by Celera were aligned with the best hit in RefSeq. Regions of the alignment that were not in the RefSeq portion of the alignment were omitted from further analyses as were loci where the chimpanzee sequence had an internal stop codon. In order to exclude paralogous genes, we restricted our selection analysis to loci with $d_S < 0.1$. Of the remaining 11,624 genes, 8,292 had at least one non-synonymous SNP or fixed difference, and could thus be used for further study using the mkprf method (McDonald–Kreitman analysis using Poisson random field; http://cbsuapps.tc.cornell.edu/mkprf.aspx). We also define two sets of informative loci ($n = 3{,}277$) with at least four variable non-synonymous sites in the alignment (that is, $P_N + D_N \geq 4$; IPS) as well as a set with at least two variable non-synonymous sites that are informative only about negative selection ($n = 6{,}033$; INS). Loci with less than four amino acid variable sites in the alignment cannot be informative about positive selection, and loci with 0 or 1 amino acid variable sites cannot inform a hypothesis regarding weak negative selection. The reason for this is that we make a conservative assumption that some (unknown) fraction of mutations are lethal. While inferences of selection based only on variable sites in the alignment reduces power, it does not require *a priori* assumptions about the distribution of selective effects among moderately to highly deleterious mutations.

Statistics. To maximize our power to identify non-neutrally evolving loci, we pool genomic polymorphism and divergence data using the approach of ref. 9 who used a Poisson random field setting[4] for bayesian population genetic inference of selection. The chief advantage of the mkprf method is an increase in power to detect selection without an increase in type I error as long as per locus mutation rates are low[9]. One slight modification of the model used here is that we set a gaussian prior distribution for γ_i with mean 0 and an arbitrary standard deviation of 8 (as opposed to a hierarchical model as in ref. 9). This simplification is made so that the marginal posterior distributions of the selection coefficients are conditionally independent of one another and can be pooled for further analysis. Full details of the likelihood function used here can be found in Supplementary Data 1. Briefly, our method estimates posterior distributions of genomic parameters such as the species divergence time between humans and chimpanzees ($\tau =$ number of generations/$2N_e$, where N_e is the effective population size of humans) using all synonymous SNPs and synonymous fixed differences in the sample. Conditional on genomic parameters, posterior distributions for individual gene parameters are informed only by the non-synonymous cell entries in a conventional McDonald–Kreitman table (that is, P_N and D_N). Evidence in our model for non-neutral evolution at a particular gene comes from the equal tail credibility intervals (a form of bayesian confidence intervals) on the selection parameter for the locus. That is, if the whole of the CI for γ_i is above 0, there is strong evidence that the locus i has an excess of amino acid fixed differences given the observed level of amino acid variability within humans at the gene, and is therefore likely to be rapidly evolving due to positive darwinian selection. Similarly, if the whole of the CI for γ_i is below 0, there are too few fixed amino acid differences in gene i; this we interpret as evidence that the locus is subject to negative selection, and that many amino acid polymorphisms at the locus are weakly deleterious. An alternative explanation is that the locus is subject to strong balancing selection (or over-dominance) that prevents fixation of non-lethal amino acid mutations between species. Recent analytical and simulation work[22,23] suggests that over-dominance needs to be very strong for the McDonald–Kreitman cell entries to deviate in the direction of excess amino acid variation relative to amino acid fixed differences (such as in self-incompatibility alleles[24]). In fact, under weak over-dominance, the McDonald–Kreitman cell entries show the same deviation (that is, an excess of amino acid fixed differences relative to polymorphism) as under positive selection[23].

Received 24 April; accepted 14 September 2005.

1. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
2. Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
3. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
4. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132,** 1161–1176 (1992).
5. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397,** 344–347 (1999).
6. 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).
7. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
8. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
9. Bustamante, C. D. *et al.* The cost of inbreeding in *Arabidopsis*. *Nature* **416,** 531–534 (2002).
10. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3,** e196 (2005).
11. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22,** 239–247 (1999).
12. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
13. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293,** 489–493 (2001).
14. Livingston, R. J. *et al.* Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14,** 1821–1831 (2004).
15. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10,** 591–597 (2001).
16. Williamson, S. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci. USA* **102,** 7882–7887 (2005).
17. Barrier, M., Bustamante, C. D., Yu, J. & Purugganan, M. D. Selection on rapidly evolving proteins in the *Arabidopsis* genome. *Genetics* **163,** 723–733 (2003).
18. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **6,** e170 (2005).
19. Manunta, P. *et al.* Alpha-adducin polymorphisms and renal sodium handling in essential hypertensive patients. *Kidney Int.* **53,** 1471–1478 (1998).
20. Morrison, A. C., Bray, M. S., Folsom, A. R. & Boerwinkle, E. ADD1 460W allele associated with cardiovascular disease in hypertensive individuals. *Hypertension* **39,** 1053–1057 (2002).
21. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
22. Weinreich, D. M. & Rand, D. M. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics* **156,** 385–399 (2000).
23. Williamson, S., Fledel-Alon, A. & Bustamante, C. D. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **168,** 463–475 (2004).
24. Ioerger, T. R., Clark, A. G. & Kao, T. H. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc. Natl Acad. Sci. USA* **87,** 9732–9735 (1990).

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

Acknowledgements We thank K. Thornton and B. Payseur for suggestions during the analysis. Some of the analysis was supported by NIH grants to C.D.B., R.N. and A.G.C. We also acknowledge the help of J. Pillardy and the Cornell University Theory Center Computational Biology Service Unit.

Author Contributions S.G., D.M.T., D.C., T.J.W., J.J.S., M.D.A. and M.C. conceived, designed and performed the experiments. C.D.B., A.F.-A., A.G.C., S.W., R.N. and M.J.H. analysed the data.

Author Information Accession numbers for the SNP markers analysed in this study are dbSNP numbers ss48401226–ss48429818 and ss48429821–ss48431291, submitted under the handle APPLERA_GI. 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 C.D.B. (cdb28@cornell.edu).