Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions

Lucas D. Ward[1,2] and Manolis Kellis[1,2]*

Although only 5% of the human genome is conserved across mammals, a substantially larger portion is biochemically active, raising the question of whether the additional elements evolve neutrally or confer a lineage-specific fitness advantage. To address this question, we integrate human variation information from the 1000 Genomes Project and activity data from the ENCODE Project. A broad range of transcribed and regulatory nonconserved elements show decreased human diversity, suggesting lineage-specific purifying selection. Conversely, conserved elements lacking activity show increased human diversity, suggesting that some recently became nonfunctional. Regulatory elements under human constraint in nonconserved regions were found near color vision and nerve-growth genes, consistent with purifying selection for recently evolved functions. Our results suggest continued turnover in regulatory regions, with at least an additional 4% of the human genome subject to lineage-specific constraint.

Initial sequencing of the human genome revealed that 98.5% of human DNA does not code for protein (*1*), raising the question of what fraction of the remaining genome is functional. Mammalian conservation suggests that ~5% of the human genome (*2*, *3*) is conserved due to noncoding and regulatory roles, but more than 80% is transcribed, bound by a regulator, or associated with chromatin states suggestive of regulatory functions (*4*–*6*). This discrepancy may result from nonconsequential biochemical activity or lineage-specific constraint (*7*, *8*). Similarly, evolutionary turnover in regulatory regions (*9*–*11*) may be due to nonconsequential activity in neutrally evolving regions in each species or turnover in functional elements associated with turnover in activity. To resolve these questions, we need new methods for measuring constraint within a species, rather than between species.

Single-nucleotide polymorphisms (SNPs) within human populations have been identified only every 153 bases on average (*12*), compared with 4.5 substitutions per site among the genomes of 29 mammals (*2*), making it impossible to detect individual constrained elements (*13*). Instead, aggregate measures of human diversity across thousands of dispersed elements are needed. Such

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: manoli@mit.edu

measures have been used to show that human constraint correlates with mammalian conservation (*4*, *14*–*17*), mRNA splice sites (*18*), and regulatory elements (*19*), and that similar selective pressures act in humans and across mammals (*2*). However, differences between mammalian and human constraint remain unresolved. Recent positive selection has been detected by unexpectedly many recent substitutions (*20*) or extreme pat-

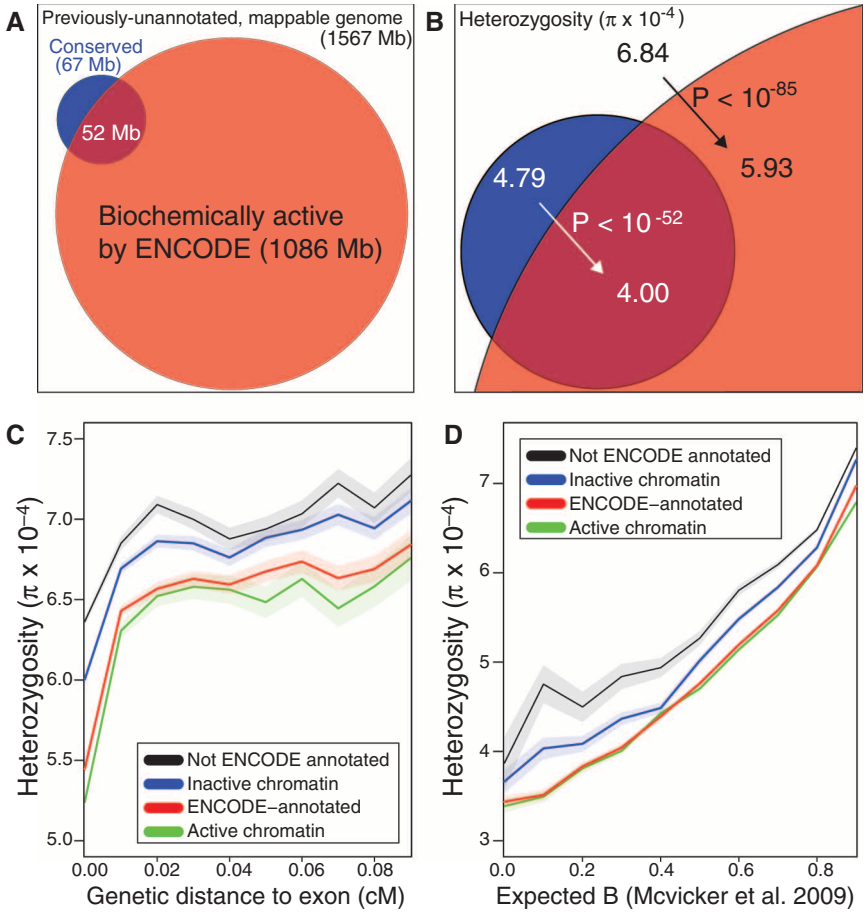

Fig. 1. (**A**) Only a small fraction (purple) of biochemically active regions (red) overlaps conserved elements (blue). (**B**) Active regions (red) show reduced heterozygosity relative to inactive regions outside conserved elements (white), suggesting lineage-specific purifying selection (black arrow). Conserved elements that lack activity (blue) show increased human heterozygosity relative to active conserved regions (purple), suggesting recent loss of selective constraint (white arrow). (**C** and **D**) Comparison of mean heterozygosity for ENCODE-annotated elements (red) versus non-ENCODE elements (black) and active chromatin (green) versus inactive (blue) shows a consistent reduction at varying genetic distances from exons (C) and varying expected background selection (D), confirming that the heterozygosity reduction is due to purifying selection. Shaded regions represent a 95% confidence interval on the mean heterozygosity if one assumes independence between bases.

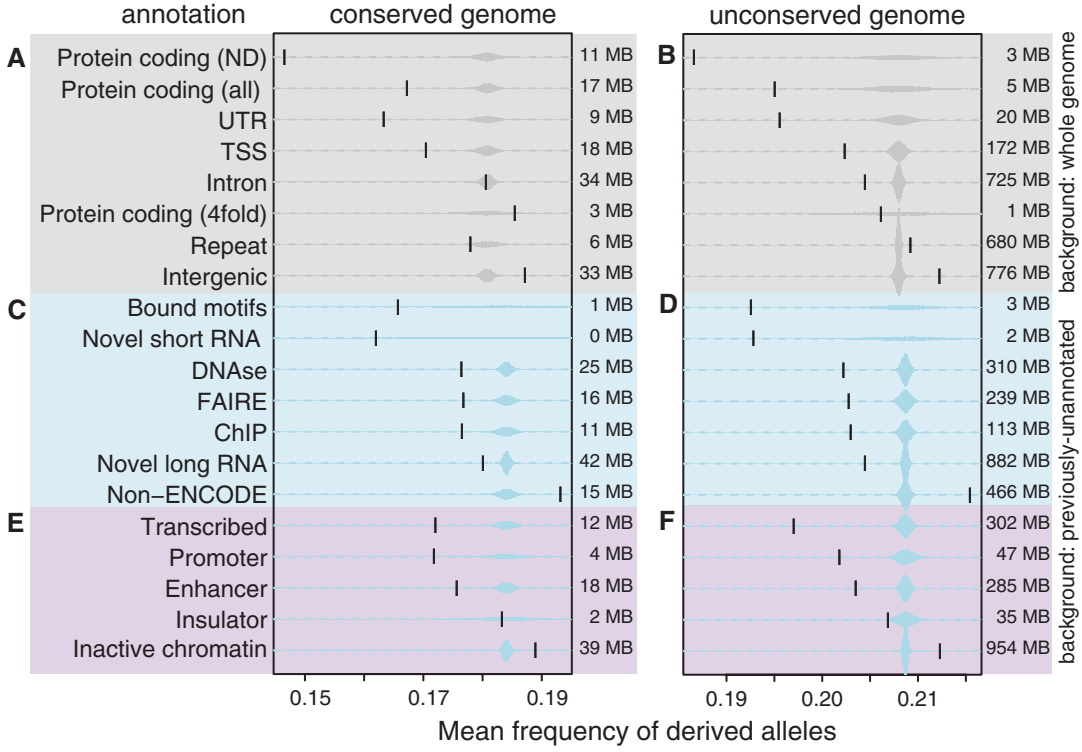

Fig. 2. Each row shows the mean frequency of derived alleles (vertical bar) found within the specified annotation type, relative to genomic samples (histogram) from the specified background (right column). These are shown for previously annotated features (**A** and **B**), ENCODE-defined features (**C** and **D**), and chromatin states (**E** and **F**), in both conserved regions [(A), (C), and (E)] and nonconserved regions [(B), (D), and (F)]. Region sizes are specified.

terns of linkage disequilibrium (LD) and population differentiation (*21*). However, recent negative selection has not been investigated, as the paucity of variants segregating in the global population makes a selective decrease in the diversity of any given locus indistinguishable from a fortuitous one.

Combining population genomic information from the 1000 Genomes Project (*12*) and biochemical data of the Encyclopedia of DNA Elements (ENCODE) Project (*5*), we estimated constraint associated with diverse genomic functions in aggregate over 1567 Mb of "previously unannotated" regions encompassing 4.7 million SNPs, excluding exons, proximal promoter regions, and artifact-prone regions (*22*) (Fig. 1A). On the basis of SNP density, heterozygosity, and derived allele frequency (DAF), we developed a statistical procedure for measuring genome-wide constraint accounting for mutation rate biases and interdependence of allele frequencies due to LD (*22*). All *P* values are derived from this test unless otherwise noted. To distinguish whether the increased human constraint in active regions (*5*) could be due solely to mammalian conservation (Fig. 1B and fig. S1), rather than lineage-specific constraint, we specifically studied regions not conserved across mammals.

Remarkably, nonconserved active regions showed significant evidence of purifying selection: SNP density was 10% lower ($P < 10^{-64}$), heterozygosity 13% ($P < 10^{-85}$), and DAF 5% ($P < 10^{-65}$), compared with reductions of 28%, 33%, and 16%, respectively, for conserved regions. Because nonconserved regions cover a >10-fold larger fraction of the genome, this suggests that a substantial fraction of human constraint lies outside mammalian-conserved regions. The observed decrease in diversity is not due to undetected conserved regions or the threshold used to defined conserved elements (fig. S2), or to background selection (*23*) (Fig. 1, C and D), biased gene conversion (table S1), or decreased mapping to nonreference alleles (*22*) (table S2).

The level of human-specific constraint varies with the observed biochemical activity (Fig. 2, figs. S3 to S5, and tables S3 and S4). Short noncoding RNAs are as strongly constrained as protein-coding regions. Long noncoding RNAs (lncRNAs) are significantly constrained in humans, even though they lack significant mammalian conservation (*5*), suggesting primarily lineage-specific functions. These results are not explained by local mutation rate variation or transcription-mediated repair, as DAF is robust to both.

We also found human-specific constraint across nonconserved regulatory features (Fig. 2, C and D). Regulatory motifs bound by their regulators show constraint similar to coding regions, and consistently higher than for nonbound instances ($P = 9.5 \times 10^{-7}$, binomial test) (Fig. 3). Regulatory regions defined by different assays, including deoxyribonuclease hypersensitivity and transcription factor binding, show significant and similar levels of human constraint. Different chromatin states (*5*, *24*) show levels of constraint according to their roles (Fig. 2, E and F), with promoter states similar to previously annotated transcription start site–proximal regions, enhancer states significant but weaker, and insulators similar to background regions, consistent with enhancer and promoter regions requiring a larger number of motifs than insulator regions. In contrast, regions that do not overlap with active ENCODE elements and inactive chromatin states show even lower constraint than ancestral repeats (Fig. 2, B, D, and F), suggesting that they may provide a more accurate neutral reference than repeats that can have exapted functions (*25*).

Comparison with primate constraint suggests evolutionary turnover. Mammalian-conserved regions lacking ENCODE activity show reduced human constraint relative to active regions (SNP density $P < 10^{-41}$, heterozygosity $P < 10^{-52}$, DAF $P < 10^{-14}$) (Fig. 1B and fig. S1), suggesting recent loss in function and activity. These also show higher primate divergence relative to active regions, suggesting that some loss of constraint likely predates human-macaque divergence. Conversely, a fraction of lineage-specific elements likely arose in the common ancestor of primates, as human-macaque divergence mirrors human diversity for both active and inactive nonconserved regions (fig. S6).

To gain insights into the functional adaptations likely involved in this turnover, we applied our aggregation approach to regulatory regions

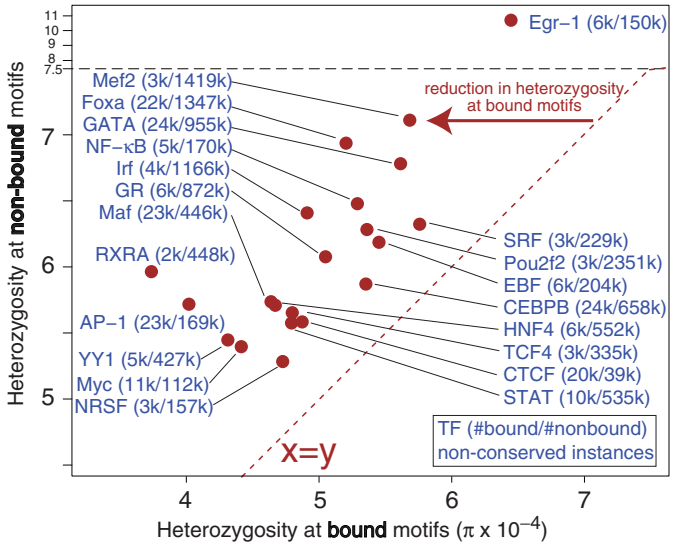

Fig. 3. Average heterozygosity for bound regulatory motif instances (*x* axis) and nonbound regulatory motif instances (*y* axis), evaluated in nonconserved regions of the genome to estimate lineage-specific constraint. Shown are all transcription factors with at least 30 kb of bound instances (red points). Numbers in parentheses indicate number of bound and number of nonbound instances, respectively.

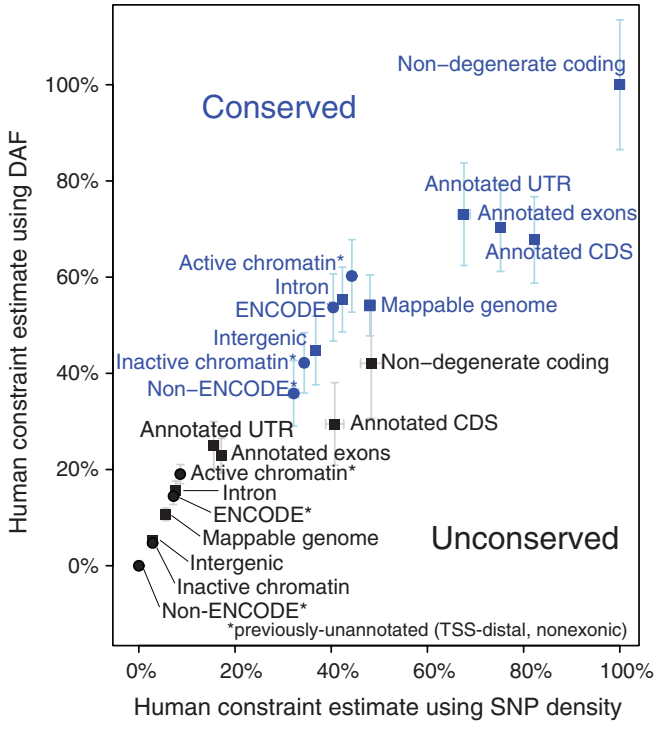

Fig. 4. Estimated PUC in human using SNP density (*x* axis) and DAF (*y* axis), across previously annotated elements (squares) and newly annotated ENCODE elements (circles), in both conserved (blue) and nonconserved (black) regions. Error bars denote 95% confidence intervals on the estimates. Each metric was linearly scaled between 0% for non-ENCODE nonconserved regions and 100% for conserved nondegenerate coding positions in each background selection bin separately (fig. S8).

associated with genes of different functions (*22*). We found that highly constrained nonconserved enhancers are associated with retinal cone cell development [$P < 10^{-4}$ in gene ontology (GO)] and nerve growth ($P < 10^{-5}$ in GO, Reactome, and the Kyoto Encyclopedia of Genes and Genomes) (fig. S7). This evidence of recent purifying selection for regulation of the nervous system and color vision is intriguing given their accelerated evolution in primates (*20*, *26*, *27*).

We next studied how the number of aggregated regions affects the ability to discriminate functional elements based on their increased human constraint (fig. S8). We found no discriminative power for individual elements, despite a significant global reduction in heterozygosity ($P < 10^{-20}$, Mann-Whitney-Wilcoxon test on heterozygosity of individual elements), but discriminative power increased significantly as the sample size grew (*22*).

We estimated the proportion of the human genome under constraint (PUC) after correcting for background selection (fig. S9) and found remarkable agreement between our orthogonal metrics (Fig. 4A). We estimate that an additional 137 Mb (4%) of the human genome is under lineage-specific purifying selection (table S6), consistent with a recent cross-species extrapolation (*28*).

Our results suggest that almost half of human constraint lies outside mammalian-conserved regions, even though the strength of human constraint is higher in conserved elements. Protein-coding constraint occurs primarily in conserved regions, whereas regulatory constraint is primarily lineage-specific (fig. S10), as proposed during mammalian radiation (*29*). Although differences in activity between mammals (*10*, *11*) can be interpreted as lack of functional constraint (*30*), our results suggest instead that turnover in activity is accompanied by turnover in selective constraint. A minority of new regulatory elements lies in recently acquired primate-specific regions (*5*), but the bulk lies in mammalian-aligned regions that provided raw materials for regulatory innovation.

Genome-wide association studies suggest that 85% of disease-associated variants are noncoding (*8*), a fraction similar to the proportion of human constraint that we estimate lies outside protein-coding regions (table S6). This suggests that mutations outside conserved elements play important roles in both human evolution and disease and that large-scale experimental assays in multiple individuals, cell types, and populations can provide a means to their systematic discovery.

References and Notes

1. E. S. Lander *et al.*; International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. K. Lindblad-Toh *et al.*; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University, *Nature* **478**, 476 (2011).
3. C. P. Ponting, R. C. Hardison, *Genome Res.* **21**, 1769 (2011).
4. E. Birney *et al.*; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, *Nature* **447**, 799 (2007).
5. The ENCODE Project Consortium, *Nature* **489**, 57 (2012).
6. J. Ernst *et al.*, *Nature* **473**, 43 (2011).
7. M. R. Nelson *et al.*, *Science* **337**, 100 (2012).
8. L. A. Hindorff *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362 (2009).
9. C. B. Lowe *et al.*, *Science* **333**, 1019 (2011).
10. D. Brawand *et al.*, *Nature* **478**, 343 (2011).
11. D. Schmidt *et al.*, *Science* **328**, 1036 (2010).
12. 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).
13. S. R. Eddy, *PLoS Biol.* **3**, e10 (2005).
14. S. Asthana *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12410 (2007).
15. J. A. Drake *et al.*, *Nat. Genet.* **38**, 223 (2006).
16. D. G. Torgerson *et al.*, *PLoS Genet.* **5**, e1000592 (2009).
17. S. Katzman *et al.*, *Science* **317**, 915 (2007).
18. D. Lomelin, E. Jorgenson, N. Risch, *Genome Res.* **20**, 311 (2010).
19. X. J. Mu, Z. J. Lu, Y. Kong, H. Y. Lam, M. B. Gerstein, *Nucleic Acids Res.* **39**, 7058 (2011).
20. K. S. Pollard *et al.*, *Nature* **443**, 167 (2006).
21. P. C. Sabeti *et al.*, *Science* **312**, 1614 (2006).
22. Materials and methods are available as supplementary materials on *Science* Online.
23. G. McVicker, D. Gordon, C. Davis, P. Green, *PLoS Genet.* **5**, e1000471 (2009).
24. J. Ernst, M. Kellis, *Nat. Biotechnol.* **28**, 817 (2010).
25. G. Bejerano *et al.*, *Nature* **441**, 87 (2006).
26. S. Dorus *et al.*, *Cell* **119**, 1027 (2004).
27. G. H. Jacobs, *Adv. Exp. Med. Biol.* **739**, 156 (2012).
28. S. Meader, C. P. Ponting, G. Lunter, *Genome Res.* **20**, 1335 (2010).
29. T. S. Mikkelsen *et al.*; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, *Nature* **447**, 167 (2007).
30. X. Y. Li *et al.*, *PLoS Biol.* **6**, e27 (2008).

Acknowledgments: We thank the ENCODE Project Consortium data producers and the ENCODE Data Analysis Center for coordinating access and performing quality control and peak-calling analysis; the Analysis Working Group of the ENCODE Project Consortium for feedback throughout this project, especially E. Birney, I. Dunham, M. Gerstein, R. Hardison, J. Stamatoyannopoulos, J. Herrero, S. Parker, P. Sabeti, S. Sunyaev, R. Altshuler, P. Kheradpour, and J. Ernst; and other members of the Kellis laboratory for discussions. L.D.W. and M.K. were funded by NIH grants R01HG004037 and RC1HG005334 and NSF CAREER grant 0644282. Data from the ENCODE consortium are available from the UCSC Genome Browser at http://genome.ucsc.edu/ENCODE, and data from the 1000 Genomes Project is available at www.1000genomes.org/data. ENCODE annotations, mammalian constraint, human diversity, background selection, and filtering information for every SNP and every human nucleotide are available at http://compbio.mit.edu/human-constraint. L.D.W. and M.K. designed the study, analyzed data, and wrote the paper.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1225057/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S6
References (*31*–*43*)

22 May 2012; accepted 14 August 2012
Published online 5 September 2012;
10.1126/science.1225057