nature genetics

Conserved noncoding sequences are selectively constrained and not mutation cold spots

Jared A Drake[1,2,10], Christine Bird[3,10], James Nemesh[1,2,10], Daryl J Thomas[4,10], Christopher Newton-Cheh[2,5,6], Alexandre Reymond[7], Laurent Excoffier[8], Homa Attar[7], Stylianos E Antonarakis[7], Emmanouil T Dermitzakis[3] & Joel N Hirschhorn[1,2,9]

Noncoding genetic variants are likely to influence human biology and disease, but recognizing functional noncoding variants is difficult. Approximately 3% of noncoding sequence is conserved among distantly related mammals[1–4], suggesting that these evolutionarily conserved noncoding regions (CNCs) are selectively constrained and contain functional variation. However, CNCs could also merely represent regions with lower local mutation rates. Here we address this issue and show that CNCs are selectively constrained in humans by analyzing HapMap genotype data. Specifically, new (derived) alleles of SNPs within CNCs are rarer than new alleles in nonconserved regions ($P = 3 \times 10^{-18}$), indicating that evolutionary pressure has suppressed CNC-derived allele frequencies. Intronic CNCs and CNCs near genes show greater allele frequency shifts, with magnitudes comparable to those for missense variants. Thus, conserved noncoding variants are more likely to be functional. Allele frequency distributions highlight selectively constrained genomic regions that should be intensively surveyed for functionally important variation.

Evolutionary conservation is a potentially powerful method to identify functional regions of the genome[5,6]. Coding regions are strongly conserved across species but comprise only ~1.5% of the genome. Comparisons of mainly mammalian but also nonmammalian genomes demonstrate that a substantial fraction of noncoding sequence (~3%) is also strongly conserved[1–4]. If CNCs are functionally important in humans, these regions would become a key hunting ground for genetic variants that contribute to human disease. To date, some studies have supported a functional role for CNCs in model systems[5–7]. However, deletions of large genomic regions containing numerous CNCs had no appreciable phenotype[8], suggesting that many CNCs might not be functionally important. Furthermore, there is little direct evidence that CNCs contain elements that are functionally relevant in humans[1–4].

Notably, the mere existence of CNCs does not prove their functional relevance in humans. The mutational or evolutionary forces that produced CNCs could still be acting on the human genome, or they could have acted in other species (past or present) and may no longer be relevant for humans. To distinguish between these possibilities, we analyzed human polymorphism data, which reflects events within the human lineage. Specifically, we compared the polymorphism densities in CNCs and in the remainder of the noncoding human genome using SNPs identified by a single genome-wide resequencing effort (see Methods)[9]. SNPs in CNCs are present at only 82% of the density seen in noncoding portions of the genome that are not evolutionarily conserved (3.9×10^{-4} versus 4.8×10^{-4}; $\chi^2 = 686$, 1 degree of freedom (d.f.); $P < 10^{-99}$). These results are consistent using two different criteria to define CNCs, are not due to differences in CNC sequence composition (see Methods) and are consistent with other recently reported results[10].

Although this marked difference in SNP density proves that the forces shaping CNCs acted on the human genome, it does not demonstrate that CNCs are functionally important. Conservation could be explained if new mutations arising in CNCs were often deleterious and less likely to become fixed in the population, thereby decreasing sequence divergence in CNCs. However, conservation and decreased SNP density could equally be explained by lower local mutation rates in CNCs relative to the remainder of the noncoding genome. Allele frequency distributions can distinguish between these possibilities because mutation rate differences do not affect frequency spectra, but selective constraint shifts the frequency distribution of constrained alleles downward[11]. Analyzing allele frequency data on human polymorphisms could determine whether CNCs are functionally relevant in humans.

[1]Program in Genomics and Divisions of Endocrinology, Children's Hospital, Boston, Massachusetts 02115, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK. [4]Department of Biomolecular Engineering, University of California Santa Cruz, California 95064, USA. [5]Division of Cardiology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, 01702, USA. [7]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [8]Zoological Institute, University of Bern, Bern 3012, Switzerland. [9]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to J.N.H. (joelh@broad.mit.edu) and E.T.D. (md4@sanger.ac.uk).

Received 14 October; accepted 4 November; published online 25 December 2005; doi:10.1038/ng1710

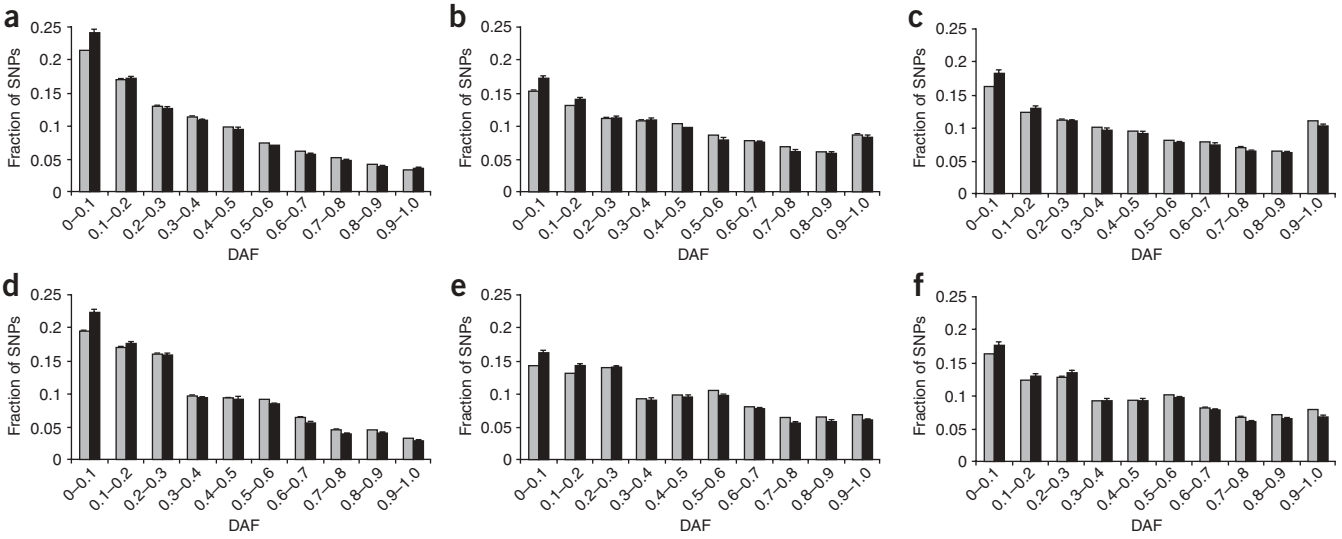

Figure 1 DAFs (derived allele frequencies) are lower for SNPs within CNCs. Shown are the distributions of DAFs for all polymorphic noncoding HapMap SNPs by frequency bins of width 0.1. Data in **a**–**c** are for CNCs ascertained using criteria set 1, and **d**–**f** are for CNCs ascertained using criteria set 2 (see Methods). DAFs were determined in unrelated individuals from three HapMap samples[12]: **a**,**d**: YRI (120 chromosomes); **b**,**e**: CEU (120 chromosomes); **c**,**f**: CHB and JPT (180 chromosomes). Light gray and black bars represent data for SNPs outside CNCs and within CNCs, respectively. The number of SNPs on which the comparisons are based are 521,719 versus 16,383 (**a**); 515,988 versus 16,284 (**b**); 487,973 versus 15,543 (**c**); 714,061 versus 29,720 (**d**); 658,495 versus 27,619 (**e**); and 628,088 versus 26,226 (**f**).

We compiled the allele frequencies for derived (new) alleles, using the chimpanzee genome to define the ancestral allele, for 990,418 SNPs from phase I of HapMap[12]. We compared the derived allele frequency (DAF) distribution for SNPs within CNCs and SNPs outside of CNCs in the HapMap populations representing three continents: Yoruba from Nigeria (YRI), American of European ancestry (CEU) and Han Chinese from Beijing combined with Japanese from Tokyo (CHB+JPT). We defined CNCs in two different ways (see Methods). With both definitions, we observed a highly significant excess of rare (<10%) derived alleles of SNPs within CNCs in all three populations (Mann-Whitney, χ^2 and Kolmogorov-Smirnoff $P \ll 0.0001$; **Fig. 1**). For the YRI population, 24% of SNPs within CNCs had DAF <10%, compared with 21% of SNPs outside of CNCs ($\chi^2 = 76$, 1 d.f.; $P < 10^{-17}$; see also **Supplementary Table 1** online). Similar results were obtained with different DAF cut points. The effect cannot be a result of population bottlenecks or other population-specific effects on the lower end of the allele frequency distribution[13], which are more marked in non-African populations (CEU, CHB+JPT)[14], because we observe if anything a slightly stronger shift in the YRI population (**Supplementary Table 1**).

We next considered the possibility that ascertainment bias in HapMap could explain the shift in DAF distributions, as could result if different criteria were used to select SNPs within and outside of CNCs. Because the targeted SNP spacing in HapMap phase I is large compared with the average size of CNCs (~360 bp), and because most major SNP discovery efforts that preceded HapMap were not focused on conserved regions, the impact of ascertainment bias is likely to be negligible. Nevertheless, we sought to control for this possible bias. First, we restricted our analysis of HapMap SNPs to those ascertained by genome-wide unbiased SNP discovery efforts (see Methods), which should have been unaffected by the presence of CNCs. When we restricted our analyses to these SNPs, we still observed a downward shift in DAF distribution in CNCs (**Fig. 2**). The level of significance was reduced because of smaller sample size, but it remained very strong in the YRI sample (χ^2 P values of 8×10^{-7} and 1×10^{-8} for CNCs selected according to the two different CNC definitions; **Table 1**). Furthermore, the downward shift in DAF distribution of SNPs within CNCs was similar to the magnitude observed for coding SNPs (**Fig. 2**). We also examined the ten 500-kb ENCODE regions across which resequencing in 48 individuals and genotyping of all SNPs in the 269 HapMap samples had been attempted[12]. In these regions, we found a similar downward trend in DAF distribution for CNCs, although the effect did not achieve statistical significance owing to small sample size (only 225 polymorphic markers fall within CNCs in the ENCODE regions; **Supplementary Table 1**). Finally, we resequenced 95 CNCs (from ref. 3) and eight exons from chromosome 21 in ten European-Americans and ten residents of West Africa. Although there were only 32 SNPs observed, the DAF spectrum showed a similar excess of rare alleles as was seen in exonic sequences (data not shown), further supporting our observations with ENCODE data. Furthermore, recent results using a differently ascertained, smaller set of SNPs were consistent with our findings[15]. Thus, ascertainment bias does not explain the shift towards rarer derived alleles in CNCs; rather, CNCs have likely been under purifying selection in humans, as has been similarly demonstrated for missense SNPs[11,16].

To provide further evidence of purifying selection in CNCs, we looked for a signature of background purifying selection at CNCs—a suppression of allele frequencies in variants close to regions under constraint[17]. For SNPs within 5 kb of a CNC, DAF was correlated with distance from a CNC (Pearson's $r = 0.022$–0.025, $P < 10^{-5}$ for all HapMap populations combined), as expected if CNCs are under purifying selection. We observed a similar correlation for SNPs within 5 kb of exons (Pearson's $r = 0.027$–0.029, $P < 10^{-5}$ for all HapMap populations combined).

CNCs include intronic sequences, promoter elements, untranslated regions (UTRs) and sequences distant from known genes. CNCs within introns showed a signal of purifying selection at least as strong

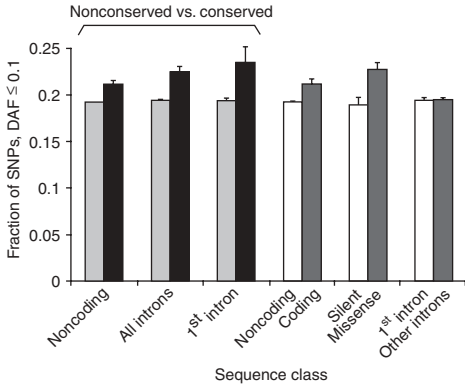

Figure 2 Fraction of evenly ascertained HapMap SNPs with DAF ≤ 0.1 in the YRI HapMap samples within and outside of CNCs and for selected functional classes. The three pairs of bars at left represent the fraction of SNPs with DAF ≤ 0.1 for all SNPs, intronic SNPs alone, and for SNPs that fall within first introns. Light gray and black bars represent data for SNPs outside CNCs and within CNCs, respectively. The three pairs of bars at right represent data for all SNPs in a sequence class, independent of conservation status. The three pairs of bars at right represent data for all SNPs in the pairs of sequence classes listed below the axis, independent of conservation status. The enrichment in SNPs with DAF ≤ 0.1 observed for noncoding SNPs within CNCs compared with non-CNCs is comparable to that observed for all coding SNPs, whereas the enrichment for intronic SNPs within CNCs is comparable to that observed for missense SNPs. SNPs in first-intron CNCs seem to have greater enrichment for rare alleles than do SNPs within CNCs that fall in other introns, whereas the set of all first-intron SNPs show no enrichment in rare alleles compared with SNPs that fall in other introns. To correct for ascertainment, SNPs are a subset of those in the HapMap database (see Methods). The number of SNPs on which the comparisons are based are, from left to right, 350,227 versus 10,676; 121,447 versus 3,532; 31,768 versus 726; 360,903 versus 6,145; 2,469 versus 3,676; and 32,494 versus 92,485.

as that seen for the entire set of CNCs and comparable to that seen for missense SNPs; the signal was more prominent for SNPs within first introns, but it was also observed in the remaining introns (**Table 1**, **Fig. 2**). The difference between conserved and nonconserved sequences persisted when we controlled for distance from exons (**Supplementary Table 1**), indicating that the decreased DAF in intronic CNCs cannot be explained by background purifying selection acting on nearby exons[17]. Furthermore, because CNCs >1 kb away from exons show stronger constraint (**Supplementary Table 1**), the functional elements in intronic CNCs are not simply canonical splicing signals but are likely to serve some as yet undetermined function. CNCs in UTRs also showed a signal of constraint; the number of HapMap SNPs that were both in promoters and in CNCs was too small to permit meaningful analysis (**Table 1**). The shift in DAF was less pronounced for CNCs located >10 kb away from genes, although the trend was still towards rarer alleles (**Table 1**). Because the overall heterozygosity was reduced in CNCs >10 kb from genes compared with similarly located nonconserved sequences (data not shown), CNCs >10 kb from genes are likely to be selectively constrained, although perhaps less strongly so than conserved elements more proximal to genes. Our previous analysis has shown that conservation is not correlated with distance from genes or presence in intergenic versus intronic regions[18]. However, our prior analysis was driven predominantly by very large intergenic regions and may not have detected the relatively subtle effects that we observed in this study at closer distances to genes. Furthermore, our prior analysis was based on interspecific variation (that is, fixed differences between species) in mammals, which has different properties than intraspecific variation (polymorphism).

Finally, we explored the relationship between conservation and selective constraint measured by DAF. We found a highly significant correlation ($P \ll 10^{-4}$; Pearson's $r = 0.08$) between human-chimpanzee rate of fixed differences[19] (eliminating the effect of polymorphism to the estimate of divergence) and DAF in the HapMap samples. In addition, exploratory analyses demonstrated that the shift in DAF distribution towards rare alleles increases as CNCs are defined using increasingly stringent conservation thresholds (see Methods). At the extreme, we analyzed 'ultraconserved' elements[20], which have remained unchanged since the common ancestor of humans and rodents. As expected, these very highly conserved regions had even lower numbers of variants than the other CNCs: within 457 ultra-conserved regions spanning 117,842 bases on the 22 autosomes, there were only 16 SNPs from a uniformly ascertained data set[9] (density = 1.4×10^{-4}/bp, compared with 3.9×10^{-4} for CNCs; $P = 8.8 \times 10^{-6}$). We genotyped 75 SNPs within ultraconserved elements and, as expected, most (52) were monomorphic in the YRI HapMap sample. Notably, the derived alleles for 11 of these SNPs were quite common (DAF > 10%; **Supplementary Table 2** online). Some individuals were homozygous for the derived alleles, indicating that despite absolute

Table 1 Comparison of DAF distributions in HapMap YRI samples for SNPs within and outside of CNCs, by SNP functional class

		Within CNCs	Outside CNCs			
Class of SNPs	DAF	Number of SNPs (%)	Number of SNPs (%)	χ^2	χ^2 *P* value	K-S *P* value
All	≤0.1	2,257 (21.1%)	67,343 (19.2%)	24.3	8.1×10^{-7}	$<1.0 \times 10^{-4}$
	>0.1	8,419 (78.9%)	282,884 (80.8%)			
Intron	≤0.1	791 (22.4%)	23,626 (19.5%)	18.9	1.4×10^{-5}	$<1.0 \times 10^{-4}$
	>0.1	2,741 (77.6%)	97,821 (80.5%)			
UTR	≤0.1	100 (25.6%)	598 (20.4%)	5.6	0.018	0.152
	>0.1	291 (74.4%)	2,338 (79.6%)			
Promoter	≤0.1	11 (15.9%)	458 (21.6%)	1.3	0.257	0.095
	>0.1	58 (84.1%)	1,659 (78.4%)			
>10 kb from gene	≤0.1	1,219 (20.0%)	37,859 (19.0%)	4.1	0.043	0.116
	>0.1	4,871 (80.0%)	161,548 (81.0%)			

DAF, derived allele frequency; K-S, Kolmogorov-Smirnoff; UTR, untranslated region. SNPs were annotated by functional status and distance from the nearest gene, and divided into SNPs within and outside of CNCs. Two tests of significance were calculated for each class of SNP to compare the DAF distributions within and outside of CNCs: a χ^2 test with 1 d.f. to compare the fraction of SNPs with DAF ≤0.1 and a nonparametric K-S test to compare the entire frequency spectrum, with empirical significance based on 10,000 permutations (see Methods). Promoters are defined as sequence 1 kb upstream of the transcriptional start. For the data in this table, only a subset of SNPs in HapMap are considered, to control for ascertainment bias (see Methods).

conservation over hundreds of millions of years of evolution, some variation in these regions is tolerated in humans. Because of the incredibly strong evolutionary constraint, the few common SNPs within ultraconserved regions become excellent candidates for disease association studies.

We have shown that evolutionarily conserved noncoding regions have been subjected to purifying selection in humans and thus are likely to harbor functionally important variants. The signal of selection is present in several different classes of CNCs and is prominent in introns, but it is perhaps weaker in CNCs at greater distances from genes. The overall signal is highly significant and similar in magnitude to that observed when comparing coding and noncoding sequences. Our results are supported by previous studies that have reported weak selective effects in CNCs[3,20–22]. However, the conclusions of these studies were based mostly on interspecific divergence, which, unlike DAF distributions, may equally be influenced by variations in mutation rate. One recent study that examined DAF data[10] used a data set, J-SNP, that intentionally oversampled coding and other potentially functional regions[23], skewing the DAF distributions in these regions towards rarer alleles; this strongly biased ascertainment limits the interpretability of comparisons based on DAF distributions. By contrast, we corrected for ascertainment bias, so we can use DAF data to demonstrate unambiguously that CNCs are selectively constrained in humans; results consistent with ours from a smaller DAF dataset were reported recently[15].

Our analysis suggests that interrogation of SNPs in CNCs could help guide the search for disease-causing variants. In addition to a proposed focus on missense SNPs[24], it would be reasonable to test SNPs within CNCs in association studies, either by genotyping CNC SNPs or SNPs in strong linkage disequilibrium (LD) with CNC variants. A systematic effort to genotype CNC variants in HapMap samples could help identify proxies for CNC SNPs. Because SNPs in CNCs are rarer on average, CNC variants, like missense variants, may be captured less well by standard LD-based approaches, so additional approaches, including directed resequencing of CNCs, may beneficially increase coverage of potentially functional SNPs in CNCs. In addition, examining CNCs is likely to uncover some causal mutations in single-gene and oligogenic disorders[25,26]. Finally, it should become possible to define computationally those classes of sequences within or in addition to CNCs that could be functionally important, such as sequences that are strongly conserved across many species and overlap particular motifs[27,28]. The increasing availability of DAF data and genome sequences should permit a more precise delineation of important noncoding elements, bringing us closer to deciphering the still largely unknown function of the noncoding portion of the human genome.

METHODS

Defining evolutionarily conserved noncoding regions (CNCs). We defined CNCs using two sets of criteria (one used by each of the two collaborating groups). For criteria set 1, evolutionarily conserved regions were gathered from the ECR Browser for a variety of parameters, including minimum lengths of 100, 150 and 200 bp and minimal identity of 70%, 75% and 80% (for mouse) and 80%, 85% and 90% (for dog). These were used to examine a pilot data set (chromosomes 1 and 2 from a previous release of HapMap data). We identified the two sets of parameters that yielded the strongest enrichment of rare alleles: a minimum of 80% identity between human and mouse with a minimum length of 200 bases (covering 3.9% of human sequence), and a minimum of 90% identity between human and dog with a minimum length of 200 bases (covering 2.9% of human sequence). The intersection of these two sets of conserved regions was used to produce the final set of CNCs (covering 2.2% of the human genome). For criteria set 2, we used conserved noncoding sequences defined by a phylogenetic Hidden Markov Model designed to extract the top 5% of the conserved genome. This is found under the track 'most conserved' from the University of California, Santa Cruz (UCSC) Browser, which captures the results of an analysis described in ref. 29. The sequences were taken from NCBI Build 35, filtered to exclude any overlap with Ensembl exons and lifted over to NCBI Build 34 using the program Liftover available at the UCSC browser. The overall proportion of sequence that is represented in the remaining set of CNCs is about 3.5% of the human genome. Except for **Figure 1**, which presents data based on both sets of criteria, the data presented are for CNCs selected according to criteria set 1. However, similar results were seen with each of the two criteria sets.

Annotating SNPs by functional class. Functional class annotations (intronic, first intron, missense, silent, UTR, promoter, intergenic) for each HapMap SNP were determined using RefSeq data from the UCSC refGene table and the National Center for Biotechnology Information (NCBI) human.rna.fna table. SNPs that fall within a transcript of the refGene table were mapped to mRNA sequence from the human.rna.fna table and their function annotated using custom software written for this purpose. In cases in which the refGene table coding start or stop coordinates did not match the correct codon in the mRNA sequence, coding start or stop coordinates from the NCBI human rna.gbk file were used. SNPs were annotated relative to each RefSeq gene within 10 kb. First introns were defined as the first intron following the first transcribed exon (regardless of the location of the start of translation). More than 99% of HapMap SNPs were successfully annotated against all nearby RefSeq genes using this method. These annotations are highly concordant with those found in dbSNP, and manual inspection of discrepancies showed this annotation method to be more accurate (data not shown).

Masking repeats and transcribed sequence. SNPs that fell in repeats from the UCSC chrN_rmsk tables were excluded from the analyses. For intronic and intergenic SNPs, SNPs in RefSeq exons or that fell within any spliced EST from the UCSC chrN_intronEST tables were excluded. Bases that fell within repeats were excluded from calculations of SNP density and heterozygosity. Bases that fell within spliced-ESTs were excluded from calculations of SNP density for intergenic regions.

Comparing SNP density and heterozygosity in CNC and non-CNC regions. To compare SNP density within and outside of CNCs, we counted the number of SNPs in these two classes of sequence that were present in dbSNP and were identified by a single unbiased genome-wide effort[9] (SNPs discovered in this effort carry the submitter handle TSC-CSHL in the dbSNP database). We divided by the number of bases within and outside of CNCs. Although we could not calculate heterozygosity directly because we did not know the depth of coverage at each nucleotide achieved in this effort, we were able to assume safely that the average depth of coverage was not different between CNCs and sequences outside of CNCs (note that repeats were masked for this analysis). Thus, we were able to compare the relative heterozygosity within CNCs and outside of CNCs by comparing the densities of TSC-CSHL SNPs.

To address possible biases stemming from differences in G+C or CpG content between CNCs and nonconserved noncoding sequence, we compared the G+C and CpG content of CNCs with that of nonconserved regions, and we found that they were essentially identical (CNCs: 39.6% G+C, 1.0% CpG; nonconserved regions, 39.6% G+C, 0.9% CpG). In addition, we repeated the SNP density test using a set of nonconserved sequences matched to the CNCs by G+C and CpG content, length and position. For each CNC, G+C and CpG content were calculated as the fraction of nucleotides that were G or C and that were C followed immediately by G, respectively. To find matched control regions, we divided the remainder of the noncoding, repeat-masked, nontranscribed genome into 100-bp bins and performed the same G+C and CpG calculations on these. CNCs were matched to nearby control regions with similar G+C and CpG content, which were trimmed or merged to match the length of the CNCs.

Comparison of derived allele frequency distributions. We restricted our analyses to the 22 autosomes. Chimpanzee alleles were used to determine the ancestral allele in calculating the derived allele frequency. Chimpanzee alleles were determined using the consensus of the UCSC chimpSimpleDiff table and an independently generated set of differences in human and chimp sequences (courtesy of Tarjei S. Mikkelsen, Broad Institute). SNPs were discarded

whenever there was a disagreement between the two chimp allele sets (7.9%) or when the chimp allele did not match either human allele (1.0%).

For allele frequency data, we used 'Phase 1' genotype data from HapMap release #16a for the allele frequency analyses[12]. For parent-offspring samples, only parents were considered to ensure independence of the chromosomes. Most of the analysis used the parents of the 30 YRI trios representing 120 independent chromosomes (see main text), but some analyses used the remaining three HapMap populations. To avoid false positive 'SNPs' that might have been genotyped by HapMap, we excluded SNPs that were monomorphic in the population being analyzed. SNPs with successful genotyping percentage below 90% were also excluded from the analysis.

For the derived allele frequency (DAF) analysis, frequency distributions were created for sets of SNPs based on functional class and whether they were within or outside of CNCs. We excluded from all analyses SNPs that fell in repeats, RefSeq exons, Ensembl exons (criteria set 2) or spliced ESTs (criteria set 1), except for analyses of missense silent, and UTR SNPs, in which only SNPs in repeats were excluded. For each functional class, 2×2 contingency tables were used to compare the DAF distribution for SNPs within and outside CNCs, using 10% as a frequency cutoff to separate rare from common SNPs. Significance was assessed using a χ^2 test with 1 d.f. We also used the Kolmogorov-Smirnoff (K-S) and Mann-Whitney tests[30] to test for differences across the entire allele frequency distribution. For the K-S analysis, 'jitter' was added to avoid ties in allele frequency, and we permuted the SNP distribution labels (CNC versus non-CNC) 10,000 times to assess empirical significance. Error bars in **Figures 1** and **2** are based on the standard deviation for estimated frequencies, using the fact that the variance of a binomial distribution is npq, where p is the frequency at which the events are observed, $q = 1 - p$, and n is the number of trials. For this data set, n is the number of SNPs and np is the number of SNPs with DAF ≤ 0.1, so the standard deviation for p is $((npq)^{1/2})/n = (pq/n)^{1/2}$.

Correction for possible ascertainment bias. To account for possible biases generated by the different ascertainment schemes that were used originally to discover the SNPs ultimately genotyped in the HapMap, we performed our analyses with both the entire set of HapMap genotype data and with a subset of data with uniform ascertainment. For the latter analysis, we restricted our analysis to SNPs discovered by BAC overlap (category 4) or the SNP Consortium (TSC) overlap (category 5) according to the HapMap allocation files. These ascertainment methods were agnostic to gene and conservation status and therefore unbiased with regard to SNP functional class or sequence conservation. We also analyzed DAF distributions and SNP density in ten ENCODE regions in which ascertainment of SNPs was much more complete. As described in the HapMap phase I paper[12], ten 500-kb ENCODE regions (5 Mb, total) in which complete resequencing was attempted in 16 Yoruba, 16 CEPH Utah (European-American), 8 Japanese and 8 Han Chinese unrelated individuals. All SNPs discovered by resequencing or already deposited in dbSNP were then genotyped in the 270 HapMap samples by the HapMap project. We also resequenced 95 CNCs from ref. 3 and eight exons on HSA21 (human chromosome 21) in 10 CEPH and 10 West African individuals using PCR-amplified fragments.

URLs. ECR Browser, http://ecrbrowser.dcode.org; UCSC Browser (including refGene and chimpSimpleDiff tables), http://www.genome.ucsc.edu; NCBI, http://ncbi.nlm.nih.gov; HapMap, http://www.hapmap.org; dbSNP, http://ncbi.nlm.nih.gov/SNP.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors wish to thank the International HapMap investigators, T.S. Mikkelsen for assistance determining derived alleles from chimp sequence comparisons, T. Bersaglieri for assistance genotyping SNPs, G. Rockwell for assistance implementing several statistical methods and analysis algorithms, A. Langane for DNA samples and M. Gagnebin and C. Rossier for sequencing. J.N.H. is a recipient of a Burroughs Wellcome Career Award in Biomedical Sciences, which supported this work. E.T.D. is supported by the Wellcome Trust and NIH. D.J.T. was supported by grants from the National Human Genome Research Institute. S.E.A. is supported by the Swiss National Science Foundation, NIH and EU. C.N.C. is supported by the GlaxoSmithKline Competitive Grants Award Program for Young Investigators and by a National Heart, Lung, and Blood Institute Mentored Patient Oriented Research Career Development Award.

AUTHORS' CONTRIBUTIONS

J.A.D., C.B., J.N. and D.J.T. contributed equally to this manuscript, and E.T.D. and J.N.H. co-directed this project.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
2. Thomas, J.W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
3. Dermitzakis, E.T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).
4. Dermitzakis, E.T., Reymond, A. & Antonarakis, S.E. Conserved non-genic sequences – an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6**, 151–157 (2005).
5. Loots, G.G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
6. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
7. Frazer, K.A. *et al.* Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14**, 367–372 (2004).
8. Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E.M. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988–993 (2004).
9. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
10. Kryukov, G.V., Schmidt, S. & Sunyaev, S. Small fitness effect of mutations in highly conserved non-coding regions. *Hum. Mol. Genet.* **14**, 2221–2229 (2005).
11. Fay, J.C., Wyckoff, G.J. & Wu, C.I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
12. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. Maruyama, T. & Fuerst, P.A. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111**, 675–689 (1985).
14. Marth, G.T., Czabarka, E., Murvai, J. & Sherry, S.T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
15. Keightley, P.D., Kryukov, G.V., Sunyaev, S., Halligan, D.L. & Gaffney, D.J. Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.* **15**, 1373–1378 (2005).
16. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
17. Charlesworth, B., Morgan, M.T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
18. Dermitzakis, E.T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
19. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
20. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
21. Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
22. Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, e42 (2005).
23. Hirakawa, M. *et al.* JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.* **30**, 158–162 (2002).
24. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33** (Suppl.), 228–237 (2003).
25. Beysen, D. *et al.* Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. *Am. J. Hum. Genet.* **77**, 205–218 (2005).
26. Emison, E.S. *et al.* A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).
27. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
28. King, D.C. *et al.* Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15**, 1051–1060 (2005).
29. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
30. Zar, J.H. *Biostatistical Analysis* 4th edn. (Prentice Hall, Upper Saddle River, New Jersey, 1999).