Inferring Selective Constraint from Population Genomic Data Suggests Recent Regulatory Turnover in the Human Brain

Daniel R. Schrider[1,*] and Andrew D. Kern[1,2]

[1]Department of Genetics, Rutgers University, Piscataway

[2]Human Genetics Institute of New Jersey, Piscataway, New Jersey

*Corresponding author: E-mail: dan.schrider@rutgers.edu.

Accepted: November 11, 2015

Abstract

The comparative genomics revolution of the past decade has enabled the discovery of functional elements in the human genome via sequence comparison. While that is so, an important class of elements, those specific to humans, is entirely missed by searching for sequence conservation across species. Here we present an analysis based on variation data among human genomes that utilizes a supervised machine learning approach for the identification of human-specific purifying selection in the genome. Using only allele frequency information from the complete low-coverage 1000 Genomes Project data set in conjunction with a support vector machine trained from known functional and nonfunctional portions of the genome, we are able to accurately identify portions of the genome constrained by purifying selection. Our method identifies previously known human-specific gains or losses of function and uncovers many novel candidates. Candidate targets for gain and loss of function along the human lineage include numerous putative regulatory regions of genes essential for normal development of the central nervous system, including a significant enrichment of gain of function events near neurotransmitter receptor genes. These results are consistent with regulatory turnover being a key mechanism in the evolution of human-specific characteristics of brain development. Finally, we show that the majority of the genome is unconstrained by natural selection currently, in agreement with what has been estimated from phylogenetic methods but in sharp contrast to estimates based on transcriptomics or other high-throughput functional methods.

Key words: population genetics, natural selection, human evolution.

Introduction

Although computational and experimental approaches have identified the majority of protein-coding genes in humans, these coding sequences only account for ~1% of the genome. Determining the extent to which the remaining ~99% of the genome may be functional remains a major challenge for biology. To this end, recent experimental advances have facilitated the identification of regulatory regions (Johnson et al. 2007), noncoding RNAs (Guttman et al. 2010), histone modifications (Barski et al. 2007), and accessible chromatin (Boyle et al. 2008). Collectively, these experiments suggest that a substantial number of functional genomic elements reside in noncoding regions.

Although these experimental approaches represent a promising avenue toward identifying noncoding functional elements in the genome, many of the putatively functional noncoding regions they identify may be inconsequential to the organism. For example, the ENCODE project (Dunham et al. 2012) integrated data from a variety of genome-wide experiments assessing expression, transcription factor binding, and other biochemical activities and concluded that 80.4% of the human genome is functional. However, if we define function as biochemical activity with fitness consequences for the organism, then evolutionary analyses tell a very different story (Graur et al. 2013). Under this definition, which we adopt here, functional regions of the genome will experience purifying (or negative) selection, which removes deleterious mutations from populations. Comparative genomic studies have identified regions of the human genome where substitutions occur less often than expected in the absence of selection, and have concluded that on the order of 5% of the human genome is functional (Chinwalla et al. 2002; Siepel et al. 2005; Lunter et al. 2006; Birney et al. 2007; Pollard et al. 2010)—far less than estimated by ENCODE. This disparity demonstrates that knowledge of purifying selection is essential for identifying functional regions of the genome.

One limitation of purely comparative genomic approaches to detect purifying selection is that selective constraint may

not be detected if it is present in only a small portion of the phylogenetic tree being examined. A particularly interesting class of elements is therefore missed by these techniques: Elements that have acquired selective constraint only recently in a single species (e.g., human-specific gains of function [GOFs]). Conversely, genomic regions experiencing a recent loss of selective constraint in only a single lineage may be misidentified as conserved throughout the phylogeny (e.g., human-specific losses of function [LOFs]). Identifying these species-specific gain and loss of function events is critical to illuminating the genetic bases for species-specific biology. Yet while comparative genomic data may not be able to detect these events, population genetic data can be used to infer the current action of purifying selection within a single species. Within a population, purifying selection will confine deleterious mutations to relatively rare allele frequencies or eliminate them altogether. This process will also reduce variation at linked sites via background selection (Charlesworth et al. 1993). Together negative and background selection decrease the number of polymorphisms and the average derived allele frequencies of polymorphisms within and surrounding functional elements (fig. 1). Indeed, the marked reduction in diversity seen within and around coding regions in the human genome is consistent with the effects of background selection (McVicker et al. 2009; Hernandez et al. 2011; Lohmueller et al. 2011).

Here we describe a method exploiting the impact of negative selection on genetic diversity within populations to identify functional regions of the human genome. Although recent studies have been able to leverage population genetic data to identify differences in the amount of purifying selection acting on different classes of sites (Pierron et al. 2012; Ward and Kellis 2012; Somel et al. 2013), we attempt to classify individual genomic regions as constrained or unconstrained by selection. In principle, this could be accomplished by comparing observed patterns of diversity with theoretical expectations. However, these expectations depend on the demographic history of the populations examined as well as the distribution of selection coefficients encountered by new mutations. Given that there is considerable uncertainty surrounding these selective and demographic parameters (Marth et al. 2003; Stajich and Hahn 2005; Eyre-Walker and Keightley 2007; Boyko et al. 2008), and given the extensive heterogeneity in recombination rates (McVean et al. 2004), as well as variation in mutation rate and data quality across the genome (Green and Ewing 2013), here we adopt a supervised machine learning approach to classification—where genomic windows of known class (i.e., functional or not) are used to algorithmically learn a set of criteria to predict the classes of genomic windows whose class membership is unknown.

In particular, we use a support vector machine (SVM) approach to classify sliding windows of the human genome as either experiencing purifying/background selection or as unconstrained based on the density and allele frequencies of single nucleotide polymorphisms (SNPs) in the 1000 Genomes data set (Altshuler et al. 2012). SVMs are trained by finding the hyperplane that optimally separates two classes of data points from a training set (where the true class of each datum is known) (Vapnik and Lerner 1963), with each data point represented by a vector of multiple measured attributes or "features." The SVM can then be used to classify data points whose classes are not known a priori according to the side of the hyperplane on which their feature vectors are located. This classification is often performed after implicitly mapping feature vectors to a higher dimensional space where the two classes are easier to separate (the "kernel trick"; Aizerman et al. 1964; Boser et al. 1992), allowing for nonlinear discrimination. Modern SVMs can also learn hyperplanes that do not perfectly separate the entire training set (Cortes and Vapnik 1995)—a necessity when some of the training data themselves may have been misclassified. SVMs have proven highly effective in a variety of biological applications (Byvatov and Schneider 2003), yet have only begun to be applied to evolutionary questions (Pavlidis et al. 2010; Lin et al. 2011; Ronen et al. 2013; Schrider et al. 2015).

Because we use genomic variation data (shaped by demographic history) to train our classifier, it will be robust to nonequilibrium demographic events provided they typically have a similar effect on patterns of variation in constrained and unconstrained regions. Thus, this supervised machine learning approach allows us to sidestep the problem of learning a parameter-rich model of demography and selection. This is a particular strength of our method in that we can use the most comprehensive data set on genomic variation, the 1000 Genomes collection, without having to fit a model consisting of dozens if not hundreds of parameters. Importantly, using real population genetic data to train our classifier will expose it to heterogeneity in mutation rate, recombination rate, and read depth.

Our resulting SVM is very effective on both simulated data and human population genetic data. Examining regions classified with high confidence, we find that the majority of the genome is unconstrained. Finally, by contrasting our classifications with phylogenetic conservation (fig. 1), we identify regions that appear to have experienced human-specific changes in selective constraint. Such regions are disproportionately found near genes involved in the development of the central nervous system (CNS), and may point to important regulatory changes affecting the human brain. These results underscore the utility of population genetic data for revealing function within the human genome.

Methods

Single Nucleotide Polymorphism Data

We downloaded SNP genotypes from Phase 1 of The 1000 Genomes Project (Altshuler et al. 2012); we ignored SNPs discovered in the exome and/or trio data but not the

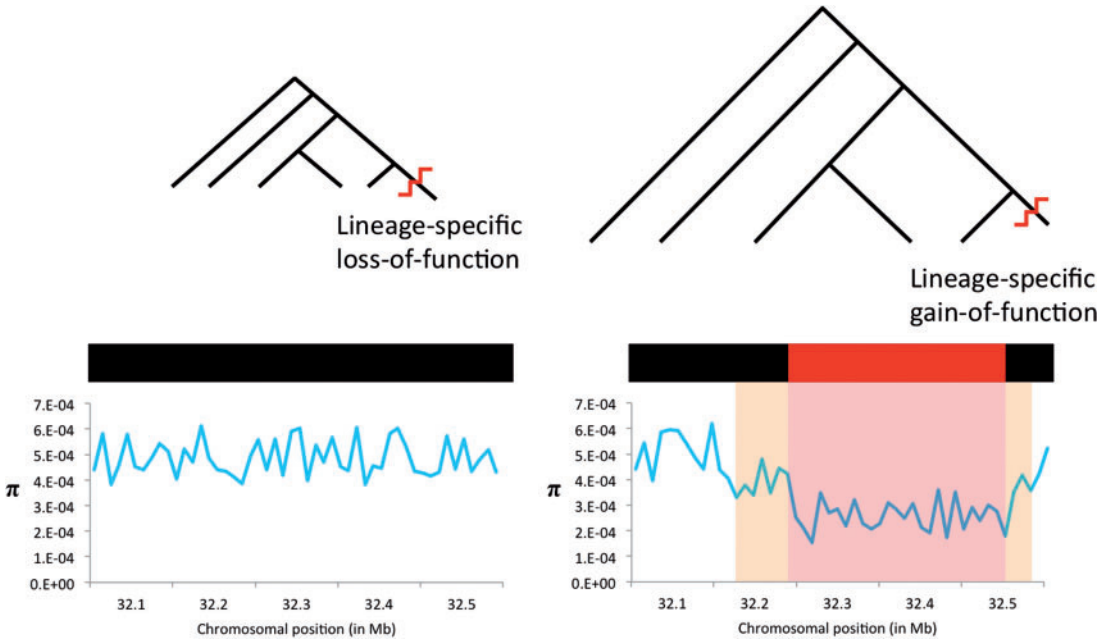

Fig. 1.—Using phylogenetic and population genetic data to find lineage-specific changes in selective constraint. In a genomic region (black bar) experiencing a lineage-specific loss of function (left), the presence of purifying selection in the majority of the phylogeny reduces divergence (short branch lengths). However, because the genomic region no longer performs a function with fitness consequences in one species, population genetic data from this species shows no reduction in diversity (as measured by nucleotide diversity, π) in this region. In the case of a lineage-specific gain of function, the majority of the phylogeny has experienced no purifying selection, and therefore divergence is higher (long branch lengths). In the species experiencing the gain of function, purifying selection reduces genetic variation in the functional region (red portion of the black bar), and background selection lowers diversity at flanking sites.

low-coverage whole-genome data in order to minimize variation in read depth across the genome, which affects the probability of discovering a polymorphism (Ajay et al. 2011). This data set contains 1,092 low-coverage genomes; however, 28 pairs of individuals in this set are close relatives to one another. We removed one individual from each of these 28 pairs leaving a set of 1,064 unrelated individuals. These individuals and their populations of origin are listed in supplementary table S1, Supplementary Material online.

Genomes, Gene Annotations, and Other Genomic Features

For the purposes of counting SNPs and monomorphic sites in an unbiased manner, creating training sets, and performing various downstream analyses, we downloaded a variety of data from version hg19 of the UCSC Genome Browser database (Kent et al. 2002; Meyer et al. 2013). These data included version GRCh37 of the human genome (Lander et al. 2001; Collins et al. 2004) with bases masked by RepeatMasker (http://www.repeatmasker.org) appearing in lower case, the UCSC gene annotation (Hsu et al. 2006), human–chimpanzee and human–macaque pairwise whole-genome alignments generated by BLASTZ (Schwartz et al. 2003), "mappability" scores for 50 bp reads (Derrien et al. 2012), regulatory regions from ORegAnno (Montgomery et al. 2006; Griffith et al. 2008), transcription factor binding sites from ENCODE (Dunham et al. 2012), lincRNAs (Trapnell et al. 2010; Cabili et al. 2011), small noncoding RNAs from miRBase (Griffiths-Jones et al. 2006; Lestrade and Weber 2006), gene-disease associations from the Genetic Association Database (Becker et al. 2004), disease-associated SNPs from genome-wide association studies compiled by Hindorff et al. (2009), and phastCons elements (Siepel et al. 2005). We also used phastCons elements called from an alignment of 29 mammalian genomes but ignoring the human state (Lindblad-Toh et al. 2011). Most of these data were downloaded using the UCSC Table Browser (Karolchik et al. 2004). We also downloaded the GENCODE v7 annotation including noncoding RNAs (Harrow et al. 2012) from www.gencodegenes.org, and Gene Ontology (GO) data from www.geneontology.org, and used the set of regulatory elements inferred to be gained or lost on the human lineage by Cotney et al. (2013).

Inferring Ancestral States and Removing Uninformative Sites

Because we sought to use the derived (or "unfolded") site frequency spectrum (SFS), we attempted to determine the ancestral state of each site containing an SNP. This was done by parsimony using whole-genome alignments of human and chimpanzee (Mikkelsen et al. 2005) and human and rhesus macaque (Gibbs et al. 2007). For each SNP, we

compared the chimpanzee and macaque genomes. If both genomes exhibited the same nucleotide as one another and as one of the two human alleles, we inferred that this nucleotide was the ancestral state. Otherwise, we considered the ancestral state to be ambiguous and ignored the SNP. If only one of the chimpanzee or macaque genomes had a base call at the site, we inferred that this base was the ancestral state if it agreed with either human allele and considered the ancestral state to be ambiguous otherwise. We also considered the ancestral state to be ambiguous if neither chimpanzee nor macaque had a base call at the site. All SNPs whose ancestral state could not be inferred unambiguously according to these rules were considered as uninformative. Although our ancestral state inferences may contain errors, our machine learning strategy should be robust if such misorientation errors also appear in our training set.

We aimed to use not only SNP allele frequencies, but also the fraction of monomorphic sites in a given region in order to classify it as constrained or unconstrained. Thus, eliminating biases affecting the fraction of sites within a genomic region inferred to be polymorphic was essential for our analysis. Because we eliminated SNPs with ambiguous ancestral states, we therefore eliminated monomorphic sites with ambiguous ancestral states to prevent the failure of ancestral state reconstruction from biasing the density of polymorphisms. This was done by attempting to infer the ancestral state at each site in the genome using rules similar to those used for SNPs as described above, but with no requirement that the sole human allele equal the chimpanzee/macaque allele(s). We considered sites where chimpanzee and macaque alleles were both found but differed from one another, or where neither were found, as having ambiguous ancestral states and therefore uninformative.

In order to prevent biases related to accuracy of mapping short-read sequences from affecting our analysis, we examined "mappability" scores calculated by Derrien et al. (2012). The mappability score for a given site is $1/n$, where n is the number of distinct positions in the genome from which a read mapped to this site could be derived (allowing two mismatches). For example, a site lying in a sequence motif occurring three times in the genome would have a score of 1/3, while a site in unique sequence would have a score of 1. We examined all adjacent 1 kb windows across the human genome and found a significant positive correlation with average mappability score and the number of SNPs called from the 1000 Genomes data ($\rho = 0.068$; $P < 2.2 \times 10^{-16}$). Windows in the lowest mappability score bin contained 7.9 SNPs on average, while windows with a mappability score of one averaged 13.6 SNPs (supplementary fig. S1, Supplementary Material online). The lack of SNP calls within regions of low mappability shows that poor mapping quality prevents high-confidence SNP detection—this underscores the importance of accounting for mappability when examining the density of SNPs or other polymorphisms. We therefore considered only sites with mappability scores of 1 to be informative. Similarly, sites masked by RepeatMasker were considered uninformative. All uninformative sites were ignored when calculating the SFS for a given window as described in the following section, and therefore had no impact on SVM training or classification.

Estimating a Modified Site Frequency Spectrum in Genomic Windows

Our goal in this study was to accurately classify genomic windows of a given size as constrained or unconstrained by purifying selection. The practical utility of this approach depends on the size of the windows: Small windows may be difficult to classify accurately as they have fewer informative sites, while larger windows provide lower resolution. To find an appropriate balance between accuracy and resolution, we attempted to train classifiers using 5, 10, and 20 kb windows; windows of these sizes contain 65, 130, and 260 SNPs and 2,176, 4,352, and 8,703 informative sites on average in the 1000 Genomes data, respectively.

We represented each window with the same modified version of the SFS used for simulated data set (as described in supplementary text S1, Supplementary Material online): $\xi = [\xi_0\ \xi_1\ \xi_2 \ldots \xi_{n-1}]$, where ξ_i is the fraction of informative sites in the window having an SNP whose derived allele is present in i chromosomes, and n is the number of chromosomes in the sample (i.e., twice the number of diploid individuals). As with the simulated data, sites containing a fixed derived allele were included in ξ_0, as our goal was to use only polymorphism data to perform classification. However, we did experiment by including derived fixations during training (as described below), finding that the gains in accuracy were quite modest (typically on the order of 1% or less; supplementary table S2, Supplementary Material online).

We estimated the modified SFS for each window only from informative sites as defined above. As a consequence, for some windows the SFS was estimated from only a small number of sites. To prevent elevated uncertainty around these SFS estimates from confounding our classifier, we arbitrarily removed windows comprised of $\leq$25% informative sites. We refer to the remaining windows as informative windows.

Because SVMs allow for a large number of features, we are able to use the complete SFS rather than a small number of summary statistics to perform classification—this is an important advantage of our method insofar as condensing the entire SFS into a summary statistic such as Tajima's D (Tajima 1989) might remove valuable information. However, the full SFS in the 1000 Genomes data is quite sparse, containing 2,128 frequency bins but only ~130 SNPs per 10 kb window on average. We therefore experimented with grouping the SFS into different numbers of bins: 10, 25, 50, 100, 250, 500, 1,000, and 2,128 (no binning), in addition

to the different genomic window sizes listed above. We found that classification was most effective with 1,000 bins, and that 10 kb windows yielded a good balance between resolution and accuracy (supplementary table S2, Supplementary Material online).

Training a Support Vector Machine Classifier

For the purposes of extracting a training set from the human genome, we subdivided the genome into adjacent windows. We then labeled windows as constrained if they were composed of >25% sites conserved across vertebrates according to phastCons (Siepel et al. 2005), or unconstrained if they contained zero base pairs within vertebrate phastCons elements, GENCODE v7 exons including noncoding RNAs (Harrow et al. 2012), UCSC exons (Hsu et al. 2006), ENCODE transcription factor binding sites (Dunham et al. 2012), or ORegAnno regulatory elements (Montgomery et al. 2006; Griffith et al. 2008). Although the >25% phastCons cutoff for functional training data is arbitrary, only ~5% of the human genome is conserved across species; windows that are 25% conserved according to phastCons are thus very likely to encode important functions. Because the amount of observed divergence on the human branch will correlate with the amount of observed polymorphism within humans due to ascertainment bias (Kern 2009), when building our training set we used phastCons conserved elements obtained from examining only nonhuman mammals (Lindblad-Toh et al. 2011). The 25% conserved sequence cutoff was adjusted for 5 and 20 kb window sizes to achieve appropriate sized training sets (supplementary table S2, Supplementary Material online). To construct an unbiased training set, we included the same number of conserved and unconserved windows. Because for each training set examined below there were more unconserved than conserved windows, windows meeting the unconserved criteria were randomly selected until a set matching the conserved set in size was obtained (i.e., a balanced training set). For 10 kb windows, this training set contained 1,482 windows in total—741 windows met the criterion for inclusion in the functional set, and 741 of the 11,439 that met the nonfunctional criteria were randomly selected for inclusion in the nonfunctional set.

For each combination of bin size and window size, we conducted a grid search of the C and γ hyperparameters and assessed the accuracy of the resulting SVMs in the same manner as for our simulated data sets. The results of these grid searches are shown in supplementary table S2, Supplementary Material online. Prior to training the SVM, we used LIBSVM's svm-scale to rescale the training data (with default parameters), saving the scalars for reuse prior to prediction. We then used LIBSVM's svm-train to learn an SVM from the entire training data set using the optimal number of bins (1,000) for 10 kb windows. The -b 1 option was used to allow estimation class membership probabilities during prediction. We used LIBSVM's plotroc.py python script to generate the receiver operating characteristic (ROC) curve (supplementary fig. S2, Supplementary Material online) for this SVM using 10-fold cross-validation. We also used plotroc.py to generate the ROC curve on a balanced independent test set and calculated the area under the curve. For this test set windows with between 20% and 25% phastCons elements were labeled as functional, while only windows with no phastCons conservation were labeled as nonfunctional.

Predictions and Element Calls

After training the SVM, we formatted all overlapping 10 kb windows (100 bp step size) for classification, and rescaled these windows using the same scalars used for the training set. We then used svm-pred to perform classification, using the -b 1 option to perform class probability estimates for each window. Next, we then combined all overlapping windows assigned to a given class with probability >0.95; LIBSVM calculates these probability estimates using Algorithm 2 from Wu et al. (2004). We refer to these regions as popCons elements when made up of windows classified as constrained, and as popUncons elements when made up of windows classified as unconstrained. We imposed this 95% probability cutoff in order to focus on windows classified with high confidence. Finally, we removed elements having ≥20% of informative sites masked by the 1000 Genomes Project for having elevated or reduced read depth or low mapping quality in order to limit the effect of these sources of error on our predictions. This was done using the strictMask files which impose stringent filters devised for population genetic analysis (available at http://www.1000genomes.org/). Note that because we performed classification on overlapping windows, it was possible for popCons elements and popUncons elements to overlap.

Searching for Evidence of Human-Specific Gain and Loss of Function

In order to find genomic regions experiencing gain or loss of selective pressure in humans only, we contrasted phylogenetic evidence for selective constraint from phastCons with population genetic evidence from popCons and popUncons elements. To find human-specific LOFs, we examined popUncons elements made up of at least 15% vertebrate phastCons elements and cross-referenced this list with UCSC genes (Hsu et al. 2006) to search for compelling candidates. For human-specific gains of selective constraint, we examined popCons elements composed of <1% vertebrate phastCons elements, cross-referencing this list with UCSC genes and ORegAnno elements to find candidate regions. For this analysis, we only included elements with informative windows (on which classification was performed) within at most 100 kb of the element in each direction. Thus, the element must be flanked by regions that contain enough

informative sites to be classified but do not exhibit a strong enough signal of selective constraint to be classified as popCons elements. This step is necessary to ensure that the target of purifying selection resides within the GOF candidate element itself rather than some flanking functional element lacking enough informative sites to be classified. Candidate GOF regions singled out in the text were also examined manually via the UCSC Genome Browser (Kent et al. 2002) to ensure that no flanking, but unclassified element, appeared to be the true target of selection. Patterns of phylogenetic conservation among primates, mammals, and vertebrates were examined using the phastCons (Siepel et al. 2005) and Genomic Evolutionary Rate Profiling (GERP; Davydov et al. 2010) tracks in the UCSC Genome Browser.

Testing for Enrichment of Element Calls with Various Genomic Features

To ask whether popCons elements overlapped more often than expected by chance with exons and other features listed in supplementary table S3, Supplementary Material online, we first counted the number of base pairs lying within both a popCons element and within one of the features being tested for enrichment. Next, we permuted the popCons coordinates such that no two elements in the permuted data set overlapped (just as in the true set). For our popCons permutations, we ensured that every permuted element consisted entirely of windows that were classified one way or another by our SVM (i.e., "informative windows"); this step ensures that any systematic differences between informative and uninformative regions (e.g., repeat content or read mappability) will not produce spurious enrichment/depletion results. We were unable to meet this constraint when permuting popUncons elements, as our permutation algorithm of randomly placing the largest remaining element in an unoccupied portion of the genome and repeating would run out of available room to randomly place elements before terminating. Fortunately, this limitation likely makes our depletion results conservative, as our informative windows are enriched for many of the functional annotation categories listed in supplementary tables S3 and S4, Supplementary Material online. For both popCons and popUncons permutations, we also ensured that no permuted elements had fewer than 80% of base pairs passing the 1000 Genomes Project's coverage and quality cutoffs in the same manner as described above for our filtering of popCons and popUncons elements.

We constructed 1,000 such permuted data sets, and then compared each of these permuted sets with each of the data sets listed in supplementary table S3, Supplementary Material online. For each comparison we counted the total number of base pairs lying within both sets. The *P*-value for each enrichment test was simply the number of permuted data sets exhibiting equal or greater overlap with the genomic feature being examined than the real popCons data set. For popUncons elements, we performed a similar test but counted permuted data sets exhibiting lesser or equal overlap to obtain a *P*-value for depletion.

We performed similar tests for GOF and LOF candidate regions and sets of genomic features listed in supplementary table S4, Supplementary Material online. These sets were obtained by applying the phastCons cutoffs we used to define GOF and LOF regions to our permuted sets. Specifically, each permuted GOF set was constructed by removing all elements from the corresponding permuted popCons set except those with $<1\%$ phastCons bases. Similarly, each permuted LOF set was constructed by removing all elements from the corresponding permuted popUncons set but those with $>15\%$ phastCons bases. In each case, the permuted set yielded more regions than the true candidate set, so we randomly sampled permuted sets of the correct size. Before testing our GOF candidates for enrichment of the genomic features in supplementary table S4, Supplementary Material online, we removed from these sets of features all elements comprised of $\geq 1\%$ phastCons bases. Similarly, we removed all genomic features comprised of $\leq 15\%$ phastCons bases before testing for LOF candidates for enrichment.

We also used version 2.0.2 of GREAT (McLean et al. 2010) to ask whether GOF and LOF elements were preferentially located near genes of particular functional categories, relative to the set of all popCons or popUncons elements, respectively. We then repeated these tests on our permuted data, asking how often terms significantly enriched in our true data were enriched in the permuted data sets.

Synonymous and Nonsynonymous Variation within PopCons and PopUnCons Elements

For orthogonal evidence that popCons and popUncons elements were correctly classified as conserved or unconserved, respectively, we examined coding SNPs within genes found in these regions. We counted the number of nonsynonymous and synonymous SNPs in each gene using the GENCODE annotation. Singleton SNPs were omitted from this analysis to limit the influence of sequencing/genotyping error.

Recombination Rates in PopCons Elements, PopUncons Elements, and Training Data

We downloaded sex-averaged recombination rates calculated by Kong et al. (2010) from the UCSC Genome Browser Database (Meyer et al. 2013). These data show the average recombination rates within 10 kb windows. These rates are adjusted so that a rate value of 1 is the genome-wide average. We calculated the average rate for each element as the sum of the rates of each 10 kb window overlapping the element, with the rate of each window weighted by the fraction of the element overlapped by the window.

Human-Specific Substitutions from a Four-Way Ape Whole-Genome Alignment

To locate human-specific substitutions and indels we first obtained an alignment consisting of human (hg19), chimpanzee (panTro2), gorilla (gorGor1), and orangutan (ponAbe2). To do this we obtained the multiz46way alignment from the UCSC genome browser (http://genome.ucsc.edu) and then extracted only these four sequences. Using this four-way alignment we then located human-specific changes using parsimony criteria requiring invariance in the other three great apes. To obtain counts of substitutions or indels per window of a given size throughout the genome, we used the featureBits tool from the Kent source tree available from the UCSC Genome Browser group.

Data Availability

Our popCons, popUncons, GOF, and LOF predictions are available in BED format on GitHub (https://github.com/kern-lab/popCons). We have also made these data accessible as a UCSC Genome Browser track hub (http://kerndev.rutgers.edu/~dan/popCons/hub.txt).

Results and Discussion

Detecting Negative Selection in Simulated Data

We assessed the effectiveness of our SVM-based approach to detect selective constraint by performing forward simulations of functional 10 kb windows containing constrained elements of various sizes, and experiencing varying strengths of negative selection, as well as nonfunctional 10 kb windows evolving entirely under drift. Each simulation utilized one of the three different mutation and recombination rates (supplementary text S1, Supplementary Material online). These simulations were performed under the demographic model learned from Tennessen et al. (2012) as described in the supplementary text S1, Supplementary Material online. This scenario models the divergence of Europeans and Africans and their subsequent population size dynamics. This demographic model is not meant to perfectly match the demographic history of our data set, which contains samples from a variety of subpopulations across the globe. Rather, it was chosen simply because it models some events common to many human subpopulations (e.g., migration out of Africa, and recent exponential population size expansion). For the purposes of training and testing our SVMs, we represented the output from each simulation by a feature vector consisting of the window's SFS (supplementary text S1, Supplementary Material online).

After training, we assessed the accuracy of each SVM using an independent test set; this estimate typically closely matched that obtained from cross-validation during the grid search (1.6% lower on average; supplementary table S5, Supplementary Material online), showing that our grid search does not lead to substantial overfitting. This important result implies that our cross-validation accuracies estimated from real data (see below) are probably reliable indicators of our method's effectiveness, even though the demographic and selective history of the 1000 Genomes population sample differs from that of our simulated populations. Moreover, we found that after imposing a >95% posterior probability cutoff (as we did when calling putative constrained and unconstrained elements from the 1000 Genomes data as discussed below), classification accuracy typically well exceeded 95% (supplementary table S5, Supplementary Material online).

Next, we assessed the effectiveness of a single SVM classifier on test sets with varying selection coefficients, selected element lengths, mutation rates, and recombination rates. For this analysis, we used the classifier learned from regions evolving under drift from those with 75% of sites under selection with a selection coefficient of $2Ns$ of 100, and with variable mutation and recombination rates; we chose to test the classifier learned from these data because this SVM's cross-validation accuracy closely mirrored that of the SVM we learned from real genomic data (see below). Perhaps unsurprisingly, we found that accuracy with which we could discriminate between simulated functional and nonfunctional windows varied according to the fraction of the 10 kb window experiencing selective constraint. When a 2.5 kb subset of the region was under negative selection, accuracy was quite low, ranging from 50% to 60% and varying only slightly according to the strength of selection (supplementary fig. S3*A*, Supplementary Material online). When the entire window was constrained accuracy was much higher (supplementary fig. S3*A*, Supplementary Material online), typically ~90% or greater (supplementary table S6, Supplementary Material online), with the one exception being cases where the mutation rate was low—in these cases unselected regions were often misclassified as selected. However, for these and other parameter combinations, the number of unselected regions misclassified as constrained decreases dramatically after imposing a 95% confidence cutoff (supplementary fig. S3*B*, Supplementary Material online). On the other hand, we find that regions with a smaller number of selected sites may often be classified as unconstrained even after imposing this cutoff (supplementary fig. S3*C*, Supplementary Material online). Thus, it may be difficult, using our approach, to confidently assert that a genomic window contains no functional sequence—a window experiencing selection at relatively few selected sites will be difficult to distinguish from an unconstrained window.

Because our SVM classifies every genomic window as either evolving under selective constraint or under drift, we reasoned that regions experiencing positive selection might be classified as constrained, as positive selection reduces diversity at linked sites (Maynard Smith and Haigh 1974). We thus simulated regions experiencing adaptive mutations (supplementary text

S1), and asked how often each SVM described above classified such regions as experiencing selective constraint. The fraction of positively selected regions classified as constrained exhibited considerable variation across SVMs, governed in part by the extent to which diversity within the negatively selected regions used to train the SVM mirrored that within positively selected simulations: The absolute difference in average π in the positively and negatively selected simulations was negatively correlated with the fraction of positively selected regions classified as constrained (Spearman's $\rho = -0.87$; $P < 2.2 \times 10^{-16}$). Thus, it appears that, depending on the strength and amount of negative selection acting on putatively functional windows used to train our SVM and the strength of recent selective sweeps occurring in the human genome, positively selected regions may often be classified as constrained.

In summary, extensive forward population genetic simulations show that our SVM approach is able to detect negative selection even in the face of the confounding effects of nonequilibrium demography. Although we have greater ability to classify as functional those windows composed of a greater number of selected sites and sites under stronger selection, we have very high specificity when detecting functional windows after imposing a strict 95% posterior probability cutoff, although we may classify windows with smaller numbers of functional base pairs as unconstrained. With these encouraging results in hand, we turn attention to empirical human data.

Accurate Classification of Functional and Nonfunctional Windows

We trained an SVM to classify 10 kb genomic windows as either constrained or unconstrained according to the same modified SFS used to classify simulated data (see Methods) using LIBSVM (Chang and Lin 2011). For this we used data from 1,064 unrelated whole-genome sequences included in Phase 1 of the 1000 Genomes Project (http://www.1000genomes.org; Altshuler et al. 2012; see Methods). This data set contains one SNP every 76.9 bp on average—we hypothesized that this high density of polymorphism would allow for the detection of regions under purifying selection at high enough resolution to be of practical utility. We then trained our SVM as described in the Methods section. Because cross-validation accuracies achieved on the X were relatively low, perhaps due to limited training data (supplementary table S2, Supplementary Material online), we only performed classification on the autosomal portion of the genome.

The optimal hyperparameter combination ($C = 2$; $\gamma = 0.125$) from the autosomal grid search resulted in a cross-validation accuracy of 87.79% (supplementary table S2, Supplementary Material online; area under ROC curve = 0.94; supplementary fig. S2, Supplementary Material online). The full results of this grid search are shown in supplementary figure S4, Supplementary Material online. That many of the other parameter values neighboring the optimal combination were nearly as accurate suggests that we did not significantly overfit our training data. Moreover, we achieve high accuracy on an independent test set not used in the selection of hyperparameter values or training (area under curve = 0.88). Furthermore, simulation results (see above) demonstrate that cross-validation accuracy for our SVM is reflective of true accuracies under a broad range of models, suggesting that we are not dramatically overestimating our accuracy due to overfitting. Moreover, we achieve these high accuracies despite the fact that levels of genetic diversity are impacted by forces other than natural selection such as drift and variation in mutation and recombination rates, supporting the notion that population genetic data can be used to distinguish constrained from unconstrained DNA (Schrider and Kern 2014).

We then used the optimal hyperparameters to train an SVM from the entire training set; this SVM was in turn used to classify every 10 kb window (with 100 bp step size) in the genome comprised of at least 25% informative sites as either constrained or unconstrained. Of 22,358,126 such genomic windows covering a total of 86.5% of the genome, the majority (16,836,483 or 75.3% of windows) were classified as unconstrained, in general agreement with comparative genomic studies (Shabalina et al. 2001; Chinwalla et al. 2002; Siepel et al. 2005; Lunter et al. 2006; Birney et al. 2007; Pollard et al. 2010). LIBSVM can be used to estimate posterior probabilities for classifications according to the distances between the classified feature vector and the discriminating hyperplane during cross-validation. In order to focus on windows classified with high confidence, we imposed a 95% probability cutoff for windows assigned as constrained or unconstrained, a cutoff that we show to be quite conservative in our simulation study (see above). Overlapping windows classified as constrained with high confidence were merged together into regions we refer to as popCons elements, and overlapping high-confidence unconstrained windows were merged into popUncons elements.

Because we trained our SVM to discriminate between regions with a fairly large fraction of conserved sites according to phastCons (>25%) and regions with zero conservation according to phastCons, regions with lower levels of conservation may not be properly classified. Indeed, this appears to often be the case in simulated data as discussed above. We therefore sought to directly assess our method's accuracy on regions with fewer functional sites by constructing several test sets with different amounts of selective constraint. We found that windows with between 0% and 5% conserved sites according to phastCons are classified as popUncons elements by our classifier 33.2% of the time, while 16.4% of windows with 5–10% conservation are classified as popUncons elements, versus 8.9% of windows with 10–15% conservation and 5.3% of windows with 15–20% conservation (table 1).

Table 1
SVM Accuracies When Discriminating between Simulated Constrained and Unconstrained Genomic Regions in Independent Test Sets

Fraction of Selected Sites	Overall Accuracy	Accuracy of popCons Calls (95% confidence)	Fraction of Unconstrained Windows Classified as popCons Elements	Accuracy of popUncons Calls (95% confidence)	Fraction of Constrained Windows Classified as popUncons Elements
0–5% ($n = 2000$)	54.45%	29/46 = 63.04%	17/1,000 = 1.70%	495/827 = 59.85%	332/1,000 = 33.20%
5–10% ($n = 2000$)	62.70%	54/67 = 80.60%	13/1,000 = 1.30%	475/639 = 74.33%	164/1,000 = 16.40%
10–15% ($n = 2000$)	69.45%	114/125 = 91.20%	11/1,000 = 1.10%	472/561 = 84.13%	89/1,000 = 8.90%
15–20% ($n = 2000$)	76.25%	184/201 = 91.50%	17/1,000 = 1.70%	477/530 = 90.00%	53/1,000 = 5.30%
20–25% ($n = 1652$)	81.17%	219/231 = 94.81%	12/826 = 1.45%	385/413 = 93.22%	28/826 = 3.39%

These results imply that many of our popUncons elements may have a relatively small number of selected sites. Although we do not have power to classify 10 kb windows as completely unconstrained by negative selection, the results from table 1 imply that our popCons elements probably contain a substantially greater density of selected sites than popUncons on average.

Crucially, we sought to minimize the impact of variation in read depth and mapping quality on our predictions. We therefore only retained elements for which >80% of all informative sites met the strict read depth and mapping quality constraints imposed by the 1000 Genomes Consortium (Altshuler et al. 2012) for population genetic analyses using these data (see Methods); these criteria enforce both strict minimum and maximum read depth as well as minimum mapping quality thresholds. This step may not be sufficient to completely eliminate the impact of variation in read depth on our predictions (Green and Ewing 2013). Such variation may thus contribute to the error rates that we have measured in our empirical test data sets.

We examined the amount and spectrum of genetic variation found in popCons and popUncons elements. Consistent with purifying and background selection acting on popCons elements, popCons elements exhibit a much greater skew in the SFS toward lower frequency variants than do popUncons elements (fig. 2*A*, Supplementary Material online) as well as much lower nucleotide diversity ($\pi = 4.02 \times 10^{-4}$ in popCons elements and $\pi = 1.12 \times 10^{-3}$ in popUncons elements; fig. 2*B*). Thus, our classifier is segmenting the genome based on the amount and spectrum of genetic diversity, as expected.

PopCons Elements Are Enriched for Features Indicative of Functionality

To test if our predictions recover previously known functional elements, we asked whether popCons elements were enriched for various genomic features that may experience selective constraint, including coding sequences, phylogenetically conserved regions of the genome (phastCons elements), regulatory elements gained or lost in the human lineage, transcription factor binding sites and other oRegAnno regulatory elements, small noncoding RNAs, lincRNAs, disease-associated genes, and candidate SNPs from GWAS studies (see Methods). The results of these enrichment tests are shown in supplementary table S3, Supplementary Material online. After Bonferroni correction, PopCons elements were significantly enriched for, and popUncons elements depleted of, all these features except of lincRNAs, GWAS SNPs, and regulatory elements lost in humans. These results show that our classifier correctly identifies constrained and unconstrained genomic regions as expected from current annotations, providing further evidence that our approach is not severely confounded by nonselective factors that impact genetic diversity. Moreover, these results confirm that our predictions have practical utility despite their relatively coarse resolution in comparison with phylogenetic methods such as GERP (Davydov et al. 2010) and phastCons (Siepel et al. 2005).

As stated above, many more genomic windows were classified as unconstrained than constrained. When using only high-confidence windows, more than half of the genome lies within popUncons elements (50,378 elements; 53.8% of the autosomes); far more than in popCons elements (17,551 elements; 11.1% of the autosomes). popUncons elements are also much larger than popCons elements on average (28,695.2 vs. 16,999.4 bp; $P < 2.2 \times 10^{-16}$; Mann–Whitney *U*-test; fig. 2*C*). At face value this result seems to strongly reject the possibility that 80% of the human genome is functional (Dunham et al. 2012). However, our classifier does not have enough resolution to predict precisely which base pairs are functional and which are not—popUncons elements may be experiencing purifying selection weak enough to go undetected, and popCons elements probably contain many base pairs not directly under purifying selection but instead linked to sites undergoing negative selection (or recent positive selection; see simulation results). Nonetheless, our results suggest that only a small fraction of the genome is experiencing strong purifying selection, again in general agreement with comparative genomic analyses (Shabalina et al. 2001; Chinwalla et al. 2002; Siepel et al. 2005; Lunter et al. 2006; Birney et al. 2007; Pollard et al. 2010; Gulko et al. 2015).

Identifying Human-Specific Loss of Function

Comparative genomic studies have identified many genes lost in humans but present in other primates (Wang et al. 2006);

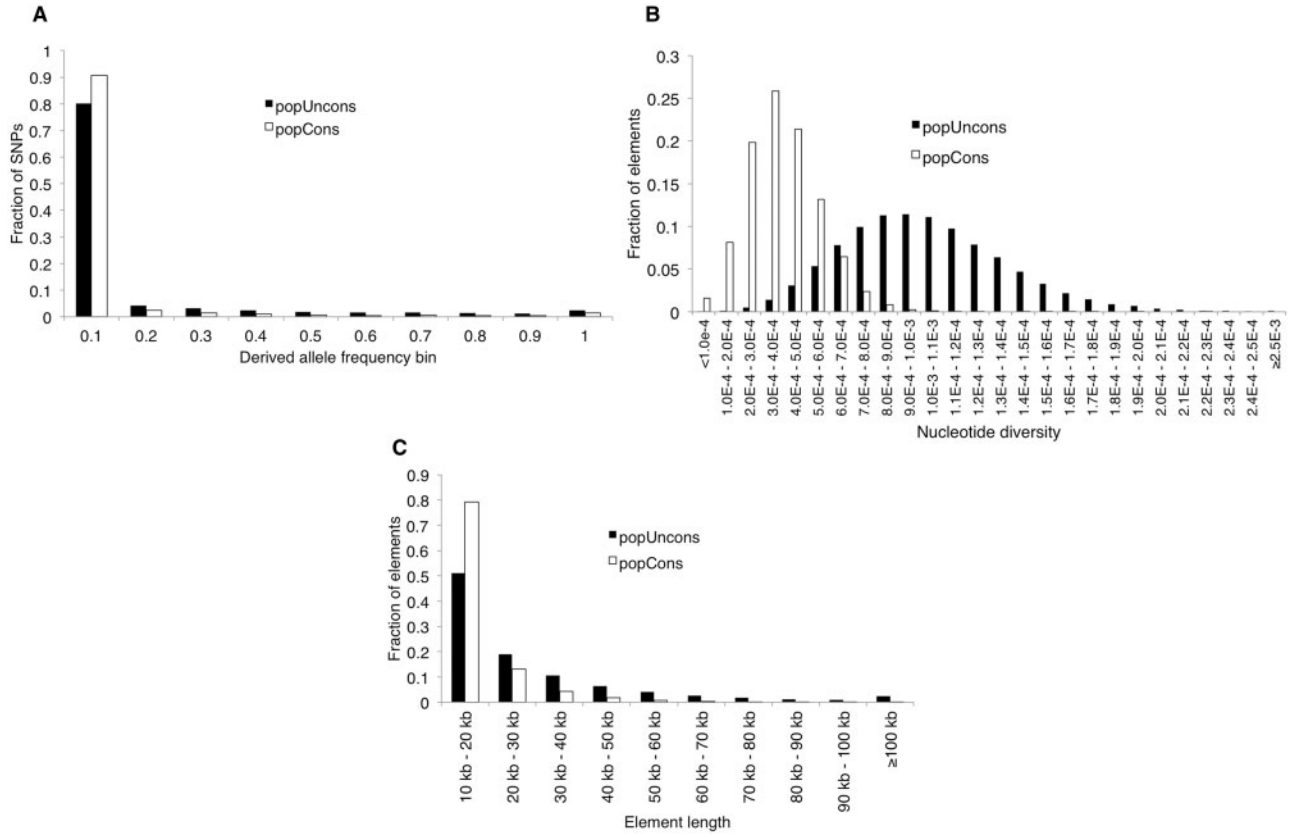

Fig. 2.—Reduced genetic variation in popCons versus popUncons elements. (*A*) SFS of popCons (white) and popUncons elements (black). The bars show the fraction of SNPs in a given element type found within each derived allele frequency bin. (*B*) Histogram of values of π within popCons (white) and popUncons (black) elements. (*C*) Histogram of lengths of popCons (white) and popUncons (black) elements.

these loss events are typically caused by a missense or other inactivating mutation and leave behind a pseudogene remnant (Schrider et al. 2009). It has been hypothesized that these loss of function events often confer fitness advantages (Olson 1999), and there are several examples of putative adaptive losses occurring since the human–chimpanzee split (Hayakawa et al. 2006; Wang et al. 2006; Xue et al. 2006).

Using evidence of phylogenetic conservation in conjunction with our population, genetic-based predictions of conservation should allow for discovery of LOF events in the genome. That is, LOF events should have strong signatures of phylogenetic conservation but also reside within popUncons elements. Indeed, our classifier was able to recover several previously identified cases of putatively adaptive pseudogenization events. For example, *MYH16*, which encodes a protein that is found in the temporalis and masseter muscles and increases bite strength, has been inactivated in the human lineage (Stedman et al. 2004). It has been hypothesized that the loss of this protein has allowed for cranial expansion in humans (Stedman et al. 2004). This gene exhibits strong phylogenetic evidence for conservation within primates according to phastCons, but is largely contained within a popUncons element, consistent with human-specific loss of selective constraint. Additional human-specific losses of *CASP12* (Fischer et al. 2002) and *CMAH* (Chou et al. 1998; Irie et al. 1998), both of which appear to have been fixed by positive selection (Hayakawa et al. 2006; Wang et al. 2006; Xue et al. 2006), occur in regions conserved across species according to phastCons but are contained entirely in popUncons elements.

Perhaps the most striking pattern to emerge from studies of human-specific pseudogenization events is the large number of nonfunctional olfactory receptors (ORs) in the human genome (Rouquier et al. 1998). ORs appear to have experienced diminished selective constraint in primates (Rouquier et al. 1998; Young et al. 2002; Zhang and Firestein 2002), perhaps due to reduced dependence on olfaction after the gain of trichromatic vision (Gilad et al. 2004). This reduction appears to be particularly pronounced in humans (Gilad et al. 2003), with roughly two-thirds of human ORs being pseudogenes (Glusman et al. 2001). Many of these inactivation events are still segregating in human populations (Menashe et al. 2003), suggesting that the loss of these genes is ongoing.

We asked whether there was greater than expected overlap between popUncons elements and OR genes and found substantial and significant enrichment (1.23-fold enrichment;

$P < 0.001$, one-tailed permutation test; see Methods). In fact, 272 of the 395 autosomal ORs not annotated as pseudogenes by GENCODE were contained entirely within a popUncons element (versus 144.25 expected; $P < 0.001$; one-tailed permutation test), while only 17 OR genes reside even partially within popCons elements (versus 46.43 expected; $P < 0.001$; one-tailed permutation test). Given that background selection may cause a gene to exhibit reduced diversity even if it is not itself the target of purifying selection, our results imply that vast majority of OR genes in the human genome are currently experiencing little if any selective constraint. This is consistent with the elevated fraction of nonsynonsymous SNPs predicted to disrupt protein function in OR genes recently observed by Pierron et al. (2012).

We searched for previously unknown cases of human-specific LOF by examining popUncons elements with strong phylogenetic evidence for conservation. We identified a total of 496 popUncons elements of which at least 15% was conserved across vertebrates according to phastCons; we refer to this set of elements and candidate LOF regions. This heuristic cutoff of 15% conservation is three times the genome-wide average and four times the average within popUncons elements (supplementary fig. S4*A*, Supplementary Material online), implying that these regions were subject to considerable selective constraint for the majority of vertebrate evolution. As discussed above, many of our popUncons elements may contain a small fraction of sites under selective constraint. This hinders our ability to detect complete loss of function with high confidence. However, given that we have defined LOF candidates as having >15% conservation across vertebrates (and they exhibit 18.56% conservation on average; supplementary fig. S4*B*, Supplementary Material online), and that our classifier labels <5% of regions with this level of conservation as popUncons elements (table 1), many of our 496 LOF candidates may have lost selective constraint at some of these previously conserved sites. This finding suggests that the loss of selective constraint on the human branch may have been a common occurrence, as suggested by Olson (1999).

Because we defined LOF candidates as regions where phylogenetic and population genetic signatures of purifying selection disagree (phastCons and popCons, respectively), they may be enriched for false positives, especially if functional turnover is a rare event. It is necessary to seek orthogonal evidence that these candidates may represent true losses of functional constraint. For this reason, we asked whether these candidates were enriched for any ontology categories. Such information can also aid in the separation of biologically meaningful candidates from spurious ones (i.e., candidates associated with an enriched functional category may more often represent true positives). This same line of reasoning also holds for gain of function candidates (discussed below).

First, we used GREAT (McLean et al. 2010) to determine whether these candidate LOF regions were enriched for particular functional categories compared with the set of all popUncons elements (although the results described below hold qualitatively when using the entire human genome as a background). Because GREAT examines genes and their flanking regions, it is able to identify the enrichment of elements within *cis*-regulatory regions of genes with a particular annotation (McLean et al. 2011) as well as the genes themselves. Using GREAT, we found that a variety of annotation terms were significantly enriched after correcting for multiple testing using *q*-values (false discovery rates). However, the most striking result was the enrichment of candidate LOF regions near genes expressed in the nervous system during various developmental stages in mice, including the developing forebrain, telencephalon, diencephalon, medulla oblongata, and optic stalk (all enriched structures shown in supplementary table S7, Supplementary Material online). We repeated this analysis on our permuted data sets (see Methods) and found that most of these terms very rarely, if ever, exhibited significant enrichment (at $q < 0.05$) in the permuted data (supplementary table S7, Supplementary Material online). The enrichment of these categories is driven largely by a set of transcription factors annotated with the zinc finger, C2H2-type/integrase, DNA-binding InterPro domain, which is also enriched for the presence of nearby LOF candidates (2.27-fold enrichment; false discovery rate $q = 2.54 \times 10^{-4}$). This result suggests that changes in the transcriptional regulation of genes may have been a common feature on the lineage leading to humans (King and Wilson 1975), with regulators of brain development playing an especially important role. We also found that LOF candidates were significantly depleted of various genomic features, including exons, disease-associated mutations, noncoding RNAs, and transcription factor binding sites (supplementary table S4, Supplementary Material online; Methods). Together these results provide additional evidence that at least a portion of sites within many of our LOF candidates have recently lost selective constraint.

Several interesting candidate loci emerged from the GREAT analysis. For example, we found a LOF candidate located <150 bp downstream of the homeobox gene *EMX2* (fig. 3*A*). This gene is expressed in the cerebral cortex during embryonic development in mice (Simeone et al. 1992), where it is required for the proper assignment of area identity to neocortical cells, as is *PAX6* (Bishop et al. 2000), another homeobox gene which itself has two upstream LOF candidates. *EMX2* also plays a role in the development of the sensory and motor regions (Hamasaki et al. 2004). The gene is one of the two human homologs of the *Drosophila* gene *empty spiracles* (or *ems*) which is required for development of the head as well as the posterior spiracles (Walldorf and Gehring 1992). We also find a LOF candidate region overlapping the 3′ exon of *SIM1* (fig. 3*B*), the homolog of *sim* (*single-minded*), which is essential for proper neurogenesis in *Drosophila* (Thomas et al. 1988). *SIM1* is associated with obesity in humans (Holder et al. 2000) and in mice (Michaud et al. 2001) where it is required for the development of the paraventricular nucleus, which is

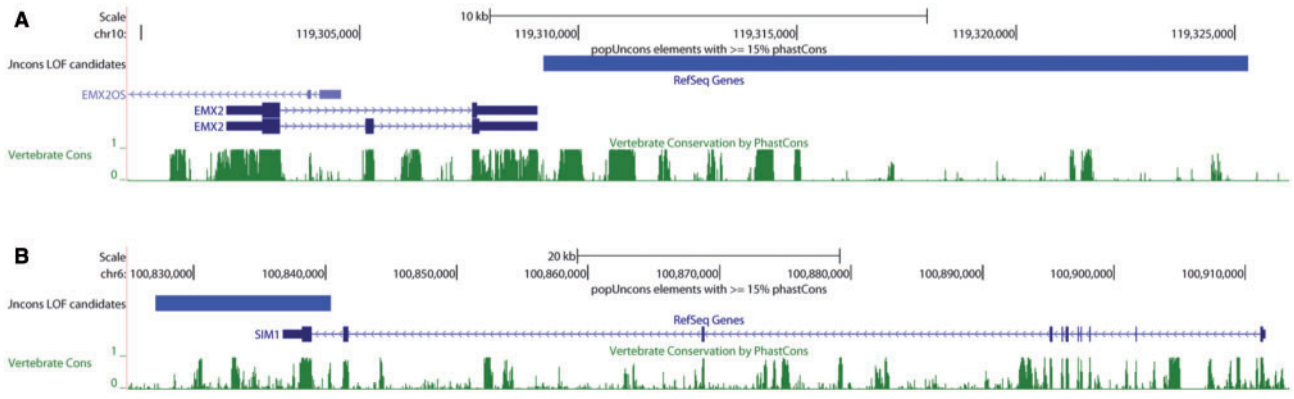

Fig. 3.—Candidate LOF regions. (*A*) A diagram of *EMX2* and the downstream flanking region generated by the UCSC Genome Browser shows a popUncons LOF candidate region (large blue bar) with a strong phylogenetic signal of conservation (high phastCons posterior probabilities, green). (*B*) A diagram of *SIM1* and its downstream flanking region.

responsible for appetite regulation among other functions (Michaud et al. 1998). Another candidate LOF region lies 7.5 kb downstream of *NR4A2* (also known as *NURR1*), a transcription factor expressed in the brain (Law et al. 1992) where it is involved in the production of dopamine neurons in mice (Saucedo-Cardenas et al. 1998). Mutations in this gene have been implicated in schizophrenia (Chen et al. 2001), Parkinson's disease (Le et al. 2002), and bipolar disorder (Buervenich et al. 2000). Intriguingly, *NR4A2* has experienced a human-specific change in the expression pattern it exhibits over the course of the lifespan in the lateral cerebellar cortex (Liu et al. 2012), which may be involved in language and other cognitive functions (Rilling 2006).

Additional transcription factors expressed in the mouse brain and involved in nervous system development and that are flanked or overlapped by candidate LOF regions include two zinc finger homeobox genes involved in neuronal differentiation, *ZFHX3* (Miura et al. 1995), whose first coding exon overlaps a LOF region, and *ZFHX4* (Hemmi et al. 2006); myelin transcription factor 1 (*MYT1*), which is important for oligodendrocyte differentiation (Nielsen et al. 2004); *LMX1B*, which plays a role in hindbrain roof plate development (Mishima et al. 2009); *NEUROG3*, a gene that is important for neuronal determination (Sommer et al. 1996); and *PAX2*, which can result in brain defects in mice when deleted (Favor et al. 1996), and whose first three exons are contained within a LOF candidate. The presence of LOF candidate regions near these transcription factors suggests recent functional turnover at their regulatory regions. *NR4A2*'s human-specific expression pattern in the brain is consistent with this hypothesis.

One notable LOF candidate region not associated with an enriched category is found within the protocadherin β (PCDHB) cluster on chromosome 5, containing most of *PCDHB14* and *PCDHB18* pseudogene. In addition to this LOF region, the PCDHB cluster contains four additional popUncons elements, three of which contain a fair amount of conserved sequence according to phastCons, although less than our 15% cutoff for LOF candidates: One element containing *PCDHB4* is made up of 10.3% conserved sequence (across vertebrates); a second element encompassing *PCDHB6* and *PCDHB17* pseudogene is 8.3% conserved; and a third element covering most of *PCDHB15* is 7% conserved. In total, 6 of the 19 PCDHB genes are mostly contained within these 5 popUnCons elements which encompass over one-third of the nearly 200 kb gene cluster.

Protocadherin genes, including the PCDHB cluster, encode cell–cell adhesion molecules that are believed to play a role in the formation of synaptic connections (Frank and Kemler 2002). The large number of and functional diversity among these genes may contribute to the complexity of the network of synapses in the human brain (Shapiro and Colman 1999). Despite strong phylogenetic evidence of purifying selection—each of the 19 PCDHB genes is largely composed of vertebrate phastCons elements—there is a fairly high rate of gene turnover in this cluster among mammals (Vanhalst et al. 2001). Indeed, 3 of the 19 genes in this cluster in humans are known to be pseudogenes. The prevalence of popUncons elements and pseudogenes among the PCDHB genes implies that their selective constraint is considerably reduced in humans. Such a change in selective pressure may have allowed for changes to the neural network in the human brain.

Candidate Human-Specific Gain of Function Events

As our extensive simulation and cross-validation experiments show, we should have excellent specificity for detecting human-specific gains of function. We use a complementary approach to that described above to find GOFs—by searching the genome for those regions that show no signs of phylogenetic conservation but are contained within popCons elements. Unfortunately, there are relatively few well-studied examples of previously nonfunctional sequences acquiring function recently in humans. We examined three known human-specific de novo genes identified by Knowles and

McLysaght (2009), *CLLU1*, *C22orf45*, and *DNAH10OS*, to see if our approach could identify these candidates. Two of these genes, *C22orf45* and *DNAH10Os*, were largely contained within popCons elements. However, these genes are found on the opposite strand of more ancient and conserved genes, thus negative selection on these older genes may be responsible for the popCons classification. Interestingly, the other gene, *CLLU1*, was found within a popUncons element and exhibits a ratio of nonsynonymous to synonymous SNPs in the 92nd percentile among all genes (see Methods), suggesting that it may not be experiencing strong selective constraint.

Although there are not enough known examples of de novo human functional elements for us to systematically assess our strategy, we can identify candidate GOF regions in a similar vein as our search for LOF regions. To this end we searched for popCons elements with little phylogenetic evidence of conservation and found 700 popCons elements that composed of <1% of base pairs within phastCons elements. On average, 0.54% of nucleotides within these regions are conserved across vertebrates versus 9.54% of nucleotides lying within the full set of popCons elements (supplementary fig. S5*C* and *D*, Supplementary Material online). These candidate GOF regions are enriched for promoters/enhancers identified by Cotney et al. (2013) as present in humans but absent from mice, as well as small noncoding RNAs, with the latter remaining significant after Bonferonni correction (supplementary table S4, Supplementary Material online; see Methods).

As before for LOF regions, we ran GREAT to identify functional categories of genes either overlapping or neighboring candidate GOF regions more often than expected by chance (using the set of all popCons elements as the background, though again we recover similar terms when using the whole genome as the background). Here we found a striking pattern: We observed significant enrichment of genes annotated with the GO molecular function term "extracellular ligand-gated ion channel activity" (false discovery rate $q = 0.045$). Indeed all enriched molecular function terms were related to GABA (γ-aminobutyric acid) or other neurotransmitters. This enrichment was driven primarily by GOFs near genes annotated with the GO molecular function "GABA-A receptor activity" ($q = 0.022$). GABA is the nervous system's primary inhibitory neurotransmitter (Petroff 2002), and GABA receptor expression patterns are known to play a key role in brain development (Lujan et al. 2005). As for LOFs, we found that these two terms were enriched at $q < 0.05$ in only a small fraction of our permuted data sets (0.3% and 2.2% of permuted sets, respectively). Human-specific changes in function affecting either GABA sequences themselves or their flanking regions could thus have profound effects on the CNS. We therefore examined these GOF candidates more closely for evidence that they may have affected the human CNS after the split with chimpanzees.

We found five GOF regions within a cluster of three GABA receptor subunit genes (*GABRB3*, *GABRA5*, *GABRG3*) on chromosome 15. Three of these GOF candidates are located downstream of *GABRB3*, which Liu et al. (2012) identified as having evolved a human-specific temporal expression pattern in the prefrontal cortex (PFC) after the human-chimpanzee divergence. *GABRB3* alleles have also been associated with autism (Buxbaum et al. 2002; Kim et al. 2007), savant skills (Nurmi et al. 2003), and epilepsy (Tanaka et al. 2008). The other two GOF candidates are located within introns of *GABRG3*. These GOFs contain several transcription factor binding sites identified by ENCODE ChIP-seq, including one ~400 bp peak observed in brain cancer cell lines among other tissues and containing 7 human-specific substitutions in an alignment of great apes (see Methods). This is a relatively high density of changes occurring on the human branch: Fewer than 2.5% of adjacent 500 bp windows in a whole-genome great ape alignment exhibit seven or more human-specific substitutions or indels. We also observed three GOF regions within a cluster of four GABA receptors on chromosome 5. One of these appears within an intron of *GABRB2*, while the other two flank either side of *GABRG2*, which evolved a novel temporal expression pattern in the human PFC according to Liu et al. (2012). Dysfunction of *GABRG2* appears to play a role in epilepsy (Hirose 2006; Tanaka et al. 2008) and alcohol dependence (Radel et al. 2005). Another GOF candidate is located upstream of *GABRA2* on chromosome 4, which like *GABRG2* and *GABRB3* experienced a human-specific change in PFC temporal expression pattern (Liu et al. 2012). *GABRA2* is also upregulated following neuronal stimulation via exposure to potassium chloride (Liu et al. 2012), and has been associated with alcohol dependence (Edenberg et al. 2004; Dick et al. 2006). It is also worth noting that we found a LOF candidate within an intron of *GABBR2*, also singled out by Liu et al. (2012) as having evolved a human-specific expression pattern in the PFC.

The proximity of GOF and LOF candidates around GABA receptor genes implies that these candidate regions may be the site of regulatory turnover responsible for human-specific expression patterns of these genes in the prefrontal cortex. Moreover, the association of these genes with neurological phenotypes such as autism suggests that they play a crucial role in CNS development. Thus our findings, combined with Liu et al.'s (2012) observation that GABA receptors have experienced an unusually high rate of such changes in expression, strongly suggest that human-specific changes in selective pressure in these candidate regions may underlie important developmental differences between the brains of humans and chimpanzees.

The signal of GOFs near neurotransmitter receptors is not limited to GABA receptors—we also find several GOFs near subunits of receptors of glutamate, the primary excitatory neurotransmitter in the CNS. Glutamate is a GABA precursor (Petroff 2002), and glutamate signaling is vital for CNS development (Lujan et al. 2005). For example, we observe a GOF

candidate upstream of *GRIK1* which encodes a glutamate receptor subunit. This gene has been associated with autism (Haldeman-Englert et al. 2010), Down syndrome (Ghosh et al. 2009), and juvenile absence epilepsy (Sander et al. 1997), and its expression levels are altered in patients with schizophrenia and bipolar disorder (Woo et al. 2007). We also find a GOF region within an intron and another downstream of *GRIK2*, a glutamate receptor subunit (fig. 4*A*). *GRIK2* has been linked to mental retardation (Motazacker et al. 2007), autism (Jamain et al. 2002), and schizophrenia (Bah et al. 2004), suggesting an important developmental role in the CNS. In addition, we find five GOF candidates in the vicinity of *GRID2* (two upstream and three intronic; fig. 4*B*), another glutamate receptor subunit which interacts directly with *GRIK2* (Kohda et al. 2003). Deletions in *GRID2* can result in cerebellar ataxia and related motor deficits (Utine et al. 2013) and delays in cognition and speech (Hills et al. 2013). Both *GRID2* and *GRIK2* were identified by Liu et al. (2012) as evolving a human-specific temporal expression profile in the lateral cerebellar cortex.

Although zinc finger genes were not enriched for GOF candiates according to GREAT, two GOFs located upstream of the brain-expressed *ZNF131* (Trappe et al. 2002) are notable because they harbor regulatory elements that may modulate its expression (fig. 5). The GOF candidate closest to the gene, ~9 kb upstream, encompasses a 1,051 bp ORegAnno element. Examining the great ape alignment we find 11 human-specific substitutions or indels within the ORegAnno element—this number is within the upper 2.5% tail of the empirical distribution of all adjacent 1 kb windows in the genome. These substitutions may have created regulatory features unique to humans. A second GOF element is located another 19 kb further upstream containing another ORegAnno element along with two noncoding RNAs with no annotated function. In addition, *ZNF131* is predicted by UNIPROT to function in the brain.

Overall, our results suggest the possibility that a substantial number of regions flanking or overlapping genes functioning in the CNS may have gained selective constraint specifically in humans. We see this pattern from not only the compelling individual cases presented above, but also from genome-wide enrichments of our predicted GOF elements. This pattern could result from the gain or modification of regulatory regions bringing about novel expression patterns. Such changes could in part be responsible for the dramatic differences in structure and function between the human brain and that of other primates. The fact that many of these genes have recently changed expression patterns in the human brain, combined with the significant enrichment of our GOF candidates for human-specific regulatory elements, shows the power of our approach of contrasting phylogenetic and population genetic data to find human-specific change of function.

Concluding Remarks

Understanding which portions of the human genome are functional is a central goal in modern biology. Here we have developed a supervised machine learning framework to detect purifying selection from population genetic data alone. Because our approach does not examine phylogenetic evidence for sequence conservation, it can be used to detect recent lineage-specific changes in selective pressure. We found through extensive simulations and cross-validation on the 1000 Genomes data set that our method is highly accurate and can be used to identify candidate regions experiencing either gain or loss of function occurring after the human–chimpanzee divergence, successfully recovering known examples of the latter. Moreover, because our supervised machine learning approach does not depend on heavily parameterized models of human demographic history and selection, we are able to leverage all available human sequence data in our search.

Although it has many advantages, our method does come with some caveats. Because we utilize the fraction of segregating sites in a region as well as their allele frequencies, variation in the spontaneous mutation rate across the genome could impact predictions. However, because we used supervised learning our classifier should be robust to such variation if it is well represented in our training set or if its effect is modest compared with the impact of purifying selection. Our high accuracy rates show that this is the case.

On the other hand, our method does appear to be confounded by balancing selection, which is expected to increase variability within the population. For example, the *HLA* loci, the *ABO* locus, and the hemoglobin *HBB* gene, which are all highly polymorphic and believed to be experiencing balancing selection (Allison 1954; Hedrick and Thomson 1983; Saitou and Yamamoto 1997; Stajich and Hahn 2005), are all classified as unconstrained by our method. This limitation of our method is probably a minor one, as balancing selection in the human genome appears to be the exception rather than the rule (Bubb et al. 2006; Leffler et al. 2013).

Our method may also be confounded by selective sweeps, which we suspect will be classified as constrained because sweeps reduce the number of segregating sites and skew the SFS away from intermediate-frequency variants (although an excess of high-frequency variants is also observed at flanking sites; Fay and Wu 2000). This issue may not greatly affect accuracy as regions experiencing selective sweeps must contain functional DNA, and as with balancing selection, such sweeps seem to have little impact on human polymorphism genome wide in any case (Hernandez et al. 2011; Lohmueller et al. 2011). However, strong selective sweeps can reduce diversity in large regions, potentially greatly inflating the inferred size of the functional region. Given sufficient numbers of examples of targets of positive or balancing selection, one could in principle train an SVM to identify these types of loci as

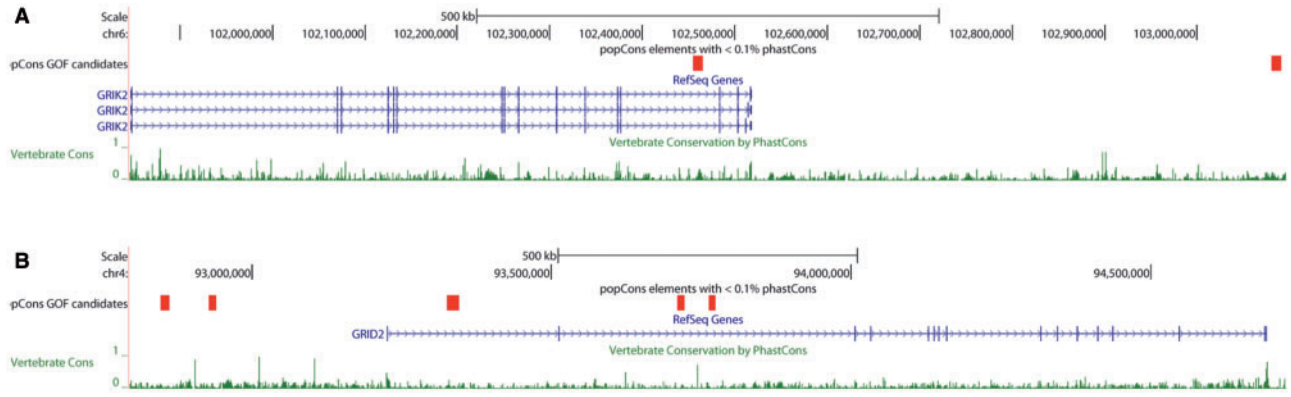

Fig. 4.—GOF candidates near glutamate receptor genes. (*A*) A diagram of *GRIK2* and its downstream flanking region generated by the UCSC Genome Browser. PopCons GOF candidate regions, shown in red, show little evidence for selective constraint across vertebrates (low phastCons posterior probabilities, green). (*B*) A diagram of GRID2 and its upstream region.

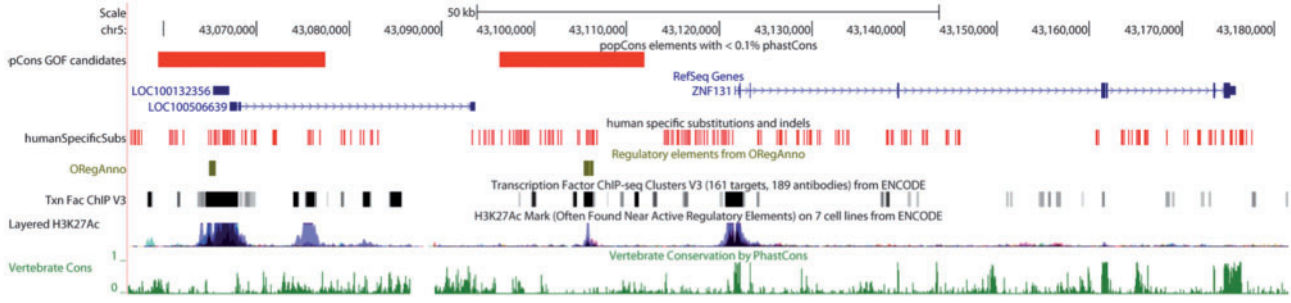

Fig. 5.—GOF candidates upstream of *ZNF131*. A diagram of *ZNF131* and its upstream flanking region generated by the UCSC Genome Browser. PopCons GOF candidate regions are shown in red. Each of these GOF regions contains an ORegAnno regulatory element, with the element closer to ZNF131 having a high density of human-specific substitutions (red tick marks). ChIP-seq peaks indicative of transcription factor binding sites are also shown (black and gray bars), as are H3K27Ac peaks (blue graph), both from ENCODE.

well. Finally, our approach may not be able to differentiate recent changes in selective pressure affecting multiple lineages (e.g., occurring prior to the human–chimpanzee split) from truly lineage-specific changes. Dense polymorphism data from multiple species would allow us to discriminate between these two cases.

Despite these limitations, our approach appears to be quite useful for identifying candidate human-specific gains and losses of function. Indeed, while we cannot directly show that these candidate regions have experienced recent changes in selective constraint, the clustering of such candidates in loci affecting CNS development and exhibiting novel expression patterns in the human brain suggest that many of these candidates represent true gains or losses of function responsible for key human-specific traits, and that such functional turnover is common on evolutionary timescales. Furthermore, our method is complementary to previous strategies for identifying lineage-specific changes in selective pressure. For example, searches for sequences highly conserved in other species but evolving rapidly in humans reveals regions likely responsible for important human-specific adaptations (Pollard, Salama, King, et al. 2006; Pollard, Salama, Lambert, et al. 2006; Kostka et al. 2012); however, the acquisition of new functional elements need not occur in previously conserved regions or be accompanied by a burst of substitution. Our approach does not depend on either of these two assumptions. Unfortunately, our resolution is currently limited by the relatively low density of polymorphism in humans. Nonetheless, the results presented here demonstrate the promise of leveraging population genetic data to detect selective constraint, an approach whose power will improve as more human genomes are sequenced.

Supplementary Material

Supplementary figures S1–S5, text S1, and tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank K. Pollard for the phastCons elements from Linbald-Toh et al. (2011), and M.W. Hahn and D.J. Begun for comments on the manuscript. D.R.S. was supported by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award F32 GM105231. A.D.K. was supported in part by Rutgers University and National Science Foundation Award MCB-1161367.

Literature Cited

Aizerman A, Braverman EM, Rozoner L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automat Remote Control 25:821–837.
Ajay SS, Parker SC, Abaan HO, Fajardo KVF, Margulies EH. 2011. Accurate and comprehensive sequencing of personal genomes. Genome Res. 21:1498–1505.
Allison AC. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J. 1:290–294.
Altshuler DM, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.
Bah J, et al. 2004. Maternal transmission disequilibrium of the glutamate receptor GRIK2 in schizophrenia. Neuroreport 15:1987–1991.
Barski A, et al. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.
Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet. 36:431–432.
Birney E, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.
Bishop KM, Goudreau G, O'Leary DD. 2000. Regulation of area identity in the mammalian neocortex by *Emx2* and *Pax6*. Science 288:344–349.
Boser BE, Guyon IM, Vapnik VN. 1992. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh: ACM Press. p. 144–152.
Boyko AR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083.
Boyle AP, et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell 132:311–322.
Bubb KL, et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics 173:2165–2177.
Buervenich S, et al. 2000. NURR1 mutations in cases of schizophrenia and manic-depressive disorder. Am J Med Genet. 96:808–813.
Buxbaum J, et al. 2002. Association between a GABRB3 polymorphism and autism. Mol Psychiatry. 7:311–316.
Byvatov E, Schneider G. 2003. Support vector machine applications in bioinformatics. Appl Bioinformatics 2:67–77.
Cabili MN, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915–1927.
Chang CC, Lin CJ. 2011. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2:27.
Charlesworth B, Morgan M, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.
Chen YH, Tsai MT, Shaw CK, Chen CH. 2001. Mutation analysis of the human NR4A2 gene, an essential gene for midbrain dopaminergic neurogenesis, in schizophrenic patients. Am J Med Genet. 105:753–757.
Chinwalla AT, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Chou HH, et al. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc Natl Acad Sci U S A. 95:11751–11756.
Collins F, Lander E, Rogers J, Waterston R, Conso I. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.
Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn. 20:273–297.
Cotney J, et al. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell 154:185–196.
Davydov EV, et al. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 6:e1001025.
Derrien T, et al. 2012. Fast computation and applications of genome mappability. PLoS One 7:e30377.
Dick DM, et al. 2006. Marital status, alcohol dependence, and GABRA2: evidence for gene-environment correlation and interaction. J Stud Alcohol. 67:185–194.
Dunham I, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.
Edenberg HJ, et al. 2004. Variations in GABRA2, encoding the α2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. Am J Hum Genet. 74:705–714.
Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610–618.
Favor J, et al. 1996. The mouse *Pax2*^{1Neu} mutation is identical to a human *PAX2* mutation in a family with renal-coloboma syndrome and results in developmental defects of the brain, ear, eye, and kidney. Proc Natl Acad Sci U S A. 93:13870–13875.
Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.
Fischer H, Koenig U, Eckhart L, Tschachler E. 2002. Human caspase 12 has acquired deleterious mutations. Biochem Biophys Res Commun. 293:722–726.
Frank M, Kemler R. 2002. Protocadherins. Curr Opin Cell Biol. 14:557–562.
Ghosh D, Sinha S, Chatterjee A, Nandagopal K. 2009. A study of GluK1 kainate receptor polymorphisms in Down syndrome reveals allelic non-disjunction at 1173 (C/T). Dis Markers. 27:45–54.
Gibbs RA, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.
Gilad Y, Man O, Pääbo S, Lancet D. 2003. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A. 100:3324–3327.
Gilad Y, Wiebe V, Przeworski M, Lancet D, Pääbo S. 2004. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. PLoS Biol. 2:e5.
Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. Genome Res. 11:685–702.
Graur D, et al. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 5:578–590.
Green P, Ewing B. 2013. Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". Science 340:682–682.
Griffith OL, et al. 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 36:D107–D113.
Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34:D140–D144.
Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 47:276–283.

Guttman M, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 28:503–510.
Haldeman-Englert CR, et al. 2010. A de novo 8.8-Mb deletion of 21q21.1–q21.3 in an autistic male with a complex rearrangement involving chromosomes 6, 10, and 21. Am J Med Genet A. 152:196–202.
Hamasaki T, Leingärtner A, Ringstedt T, O'Leary DD. 2004. EMX2 regulates sizes and positioning of the primary sensory and motor areas in neocortex by direct specification of cortical progenitors. Neuron 43:359–372.
Harrow J, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22:1760–1774.
Hayakawa T, Aki I, Varki A, Satta Y, Takahata N. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics 172:1139–1146.
Hedrick PW, Thomson G. 1983. Evidence for balancing selection at HLA. Genetics 104:449–456.
Hemmi K, et al. 2006. A homeodomain-zinc finger protein, ZFHX4, is expressed in neuronal differentiation manner and suppressed in muscle differentiation manner. Biol Pharma Bull. 29:1830–1835.
Hernandez RD, et al. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920–924.
Hills LB, et al. 2013. Deletions in GRID2 lead to a recessive syndrome of cerebellar ataxia and tonic upgaze in humans. Neurology 81:1378–1386.
Hindorff LA, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 106:9362–9367.
Hirose S. 2006. A new paradigm of channelopathy in epilepsy syndromes: intracellular trafficking abnormality of channel molecules. Epilepsy Res.. 70(Suppl 1):S206–S217.
Holder JL, Butte NF, Zinn AR. 2000. Profound obesity associated with a balanced translocation that disrupts the SIM1 gene. Hum Mol Genet. 9:101–108.
Hsu F, et al. 2006. The UCSC known genes. Bioinformatics 22:1036–1046.
Irie A, Koyama S, Kozutsumi Y, Kawasaki T, Suzuki A. 1998. The molecular basis for the absence of N-glycolylneuraminic acid in humans. J Biol Chem. 273:15866–15871.
Jamain S, et al. 2002. Linkage and association of the glutamate receptor 6 gene with autism. Mol Psychiatry. 7:302–310.
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502.
Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
Kern AD. 2009. Correcting the site frequency spectrum for divergence-based ascertainment. PLoS One 4:e5152.
Kim SA, Kim JH, Park M, Cho IH, Yoo HJ. 2007. Association of GABRB3 polymorphisms with autism spectrum disorders in Korean trios. Neuropsychobiology 54:160–165.
King MC, Wilson AC. 1975. Evolultion at two levels in humans and chimpanzees. Science 188:107–116.
Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome Res. 19:1752–1759.
Kohda K, et al. 2003. Heteromer formation of δ2 glutamate receptors with AMPA or kainate receptors. Mol Brain Research 110:27–37.
Kong A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099–1103.
Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol Biol Evol. 29:1047–1057.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Law SW, Conneely O, DeMayo F, O'malley B. 1992. Identification of a new brain-specific transcription factor, NURR1. Mol Endocrinol 6:2129–2135.
Le WD, et al. 2002. Mutations in NR4A2 associated with familial Parkinson disease. Nat Genet. 33:85–89.
Leffler EM, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578–1582.
Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. 34:D158–D162.
Lin K, Li H, Schlötterer C, Futschik A. 2011. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187:229–244.
Lindblad-Toh K, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.
Liu X, et al. 2012. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res. 22:611–622.
Lohmueller KE, et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7:e1002326.
Lujan R, Shigemoto R, Lopez-Bendito G. 2005. Glutamate and GABA receptor signalling in the developing brain. Neuroscience 130:567–580.
Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol. 2:e5.
Marth G, et al. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci U S A. 100:376–381.
Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.
McLean CY, et al. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 28:495–501.
McLean CY, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219.
McVean GA, et al. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471.
Menashe I, Man O, Lancet D, Gilad Y. 2003. Different noses for different people. Nat Genet. 34:143–144.
Meyer LR, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41:D64–D69.
Michaud JL, et al. 2001. Sim1 haploinsufficiency causes hyperphagia, obesity and reduction of the paraventricular nucleus of the hypothalamus. Hum Mol Genet. 10:1465–1473.
Michaud JL, Rosenquist T, May NR, Fan CM. 1998. Development of neuroendocrine lineages requires the bHLH–PAS transcription factor SIM1. Genes Dev. 12:3264–3275.
Mikkelsen TS, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.
Mishima Y, Lindgren AG, Chizhikov VV, Johnson RL, Millen KJ. 2009. Overlapping function of Lmx1a and Lmx1b in anterior hindbrain roof plate formation and cerebellar growth. J Neurosci 29:11377–11384.
Miura Y, et al. 1995. Cloning and characterization of an ATBF1 isoform that expresses in a neuronal differentiation-dependent manner. J Biol Chem. 270:26840–26848.

Montgomery S, et al. 2006. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 22:637–640.
Motazacker MM, et al. 2007. A defect in the ionotropic glutamate receptor 6 gene (*GRIK2*) is associated with autosomal recessive mental retardation. Am J Hum Genet. 81:792–798.
Nielsen JA, Berndt JA, Hudson LD, Armstrong RC. 2004. Myelin transcription factor 1 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells. Mol Cell Neurosci 25:111–123.
Nurmi EL, et al. 2003. Exploratory subsetting of autism families based on savant skills improves evidence of genetic linkage to 15q11-q13. J Am Acad Child Adolesc Psychiatry. 42:856–863.
Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 64:18–23.
Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185:907–922.
Petroff OA. 2002. Book review: GABA and glutamate in the human brain. Neuroscientist 8:562–573.
Pierron D, Cortés NG, Letellier T, Grossman LI. 2012. Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. Mol Phylogenet Evol. 66:558–564.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Pollard KS, Salama SR, King B, et al. 2006. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2:e168.
Pollard KS, Salama SR, Lambert N, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172.
Radel M, et al. 2005. Haplotype-based localization of an alcohol dependence gene to the 5q34 {gamma}-aminobutyric acid type A gene cluster. Arch Gen Psychiatry. 62:47–55.
Rilling JK. 2006. Human and nonhuman primate brains: are they allometrically scaled versions of the same design? Evol Anthropol 15:65–77.
Ronen R, Udpa N, Halperin E, Bafna V. 2013. Learning natural selection from the site frequency spectrum. Genetics 195:181–193.
Rouquier S, et al. 1998. Distribution of olfactory receptor genes in the human genome. Nat Genet. 18:243–250.
Saitou N, Yamamoto FI. 1997. Evolution of primate ABO blood group genes and their homologous genes. Mol Biol Evol. 14:399–411.
Sander T, et al. 1997. Allelic association of juvenile absence epilepsy with a GluR5 kainate receptor gene (GRIK1) polymorphism. Am J Med Genet. 74:416–421.
Saucedo-Cardenas O, et al. 1998. Nurr1 is essential for the induction of the dopaminergic phenotype and the survival of ventral mesencephalic late dopaminergic precursor neurons. Proc Natl Acad Sci U S A. 95:4013–4018.
Schrider DR, Costello JC, Hahn MW. 2009. All human-specific gene losses are present in the genome as pseudogenes. J Comput Biol. 16:1419–1427.
Schrider DR, Kern AD. 2014. Discovering functional DNA elements using population genomic information: a proof of concept using human mtDNA. Genome Biol Evol. 6:1542–1548.
Schrider DR, Mendes FK, Hahn MW, Kern AD. 2015. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200:267–284.
Schwartz S, et al. 2003. Human–mouse alignments with BLASTZ. Genome Res. 13:103–107.
Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. 17:373–376.
Shapiro L, Colman DR. 1999. The diversity of cadherins and implications for a synaptic adhesive code in the CNS. Neuron 23:427–430.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Simeone A, et al. 1992. Two vertebrate homeobox genes related to the *Drosophila* empty spiracles gene are expressed in the embryonic cerebral cortex. EMBO J. 11:2541–2550.
Somel M, et al. 2013. A scan for human-specific relaxation of negative selection reveals unexpected polymorphism in proteasome genes. Mol Biol Evol. 30:1808–1815.
Sommer L, Ma Q, Anderson DJ. 1996. *neurogenins*, a novel family of *atonal*-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor cell heterogeneity in the developing CNS and PNS. Mol Cell Neurosci 8:221–241.
Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 22:63–73.
Stedman HH, et al. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428:415–418.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Tanaka M, et al. 2008. Hyperglycosylation and reduced GABA currents of mutated *GABRB3* polypeptide in remitting childhood absence epilepsy. Am J Hum Genet. 82:1249–1261.
Tennessen JA, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69.
Thomas JB, Crews ST, Goodman CS. 1988. Molecular genetics of the *single-minded* locus: a gene involved in the development of the *Drosophila* nervous system. Cell 52:133–141.
Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.
Trappe R, et al. 2002. The murine BTB/POZ zinc finger gene *ZNF131*: predominant expression in the developing central nervous system, in adult brain, testis, and thymus. Biochem Biophys Res Commun. 296:319–327.
Utine GE, et al. 2013. A homozygous deletion in GRID2 causes a human phenotype with cerebellar ataxia and atrophy. J Child Neurol. 28:926–932.
Vanhalst K, Kools P, Vanden E, van Roy F. 2001. The human and murine protocadherin-beta one-exon gene families show high evolutionary conservation, despite the difference in gene number. FEBS Lett. 495:120–125.
Vapnik V, Lerner A. 1963. Pattern recognition using generalized portrait method. Automat Remote Control. 24:774–780.
Walldorf U, Gehring W. 1992. Empty spiracles, a gap gene containing a homeobox involved in *Drosophila* head development. EMBO J. 11:2247–2259.
Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. PLoS Biol. 4:e52.
Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337:1675–1678.
Woo TU, et al. 2007. Differential alterations of kainate receptor subunits in inhibitory interneurons in the anterior cingulate cortex in schizophrenia and bipolar disorder. Schizophr Res. 96:46–61.
Wu TF, Lin CJ, Weng RC. 2004. Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res. 5:975–1005.
Xue Y, et al. 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet. 78:659–670.
Young JM, et al. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum Mol Genet. 11:535–546.
Zhang X, Firestein S. 2002. The olfactory receptor gene superfamily of the mouse. Nat Neurosci. 5:124–133.

Associate editor: Gunter Wagner