Human Variation in Short Regions Predisposed to Deep Evolutionary Conservation

Gabriela G. Loots[1] and Ivan Ovcharenko*[,2]

[1]Biology and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA

[2]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

***Corresponding author:** E-mail: ovcharei@ncbi.nlm.nih.gov.

Associate editor: Naruya Saitou

Abstract

The landscape of the human genome consists of millions of short islands of conservation that are 100% conserved across multiple vertebrate genomes (termed "bricks"), the majority of which are located in noncoding regions. Several hundred thousand bricks are deeply conserved reaching the genomes of amphibians and fish. Deep phylogenetic conservation of noncoding DNA has been reported to be strongly associated with the presence of gene regulatory elements, introducing bricks as a proxy to the functional noncoding landscape of the human genome. Here, we report a significant overrepresentation of bricks in the promoters of transcription factors and developmental genes, where the high level of phylogenetic conservation correlates with an increase in brick overrepresentation. We also found that the presence of a brick dictates a predisposition to evolutionary constraint, with only 0.7% of the amniota brick central nucleotides being diverged within the primate lineage—an 11-fold reduction in the divergence rate compared with random expectation. Human single-nucleotide polymorphism (SNP) data explains only 3% of primate-specific variation in amniota bricks, thus arguing for a widespread fixation of brick mutations within the primate lineage and prior to human radiation. This variation, in turn, might have been utilized as a driving force for primate- and hominoid-specific adaptation. We also discovered a pronounced deviation from the evolutionary predisposition in the human lineage, with over 20-fold increase in the substitution rate at brick SNP sites over expected values. In addition, contrary to typical brick mutations, brick variation commonly encountered in the human population displays limited, if any, signatures of negative selection as measured by the minor allele frequency and population differentiation (F-statistical measure) measures. These observations argue for the plasticity of gene regulatory mechanisms in vertebrates—with evidence of strong purifying selection acting on the gene regulatory landscape of the human genome, where widespread advantageous mutations in putative regulatory elements are likely utilized in functional diversification and adaptation of species.

Key words: gene regulation, enhancer evolution, selection and adaptation, sequence conservation.

Research article

Introduction

One of the major goals in human genomics is understanding the genetic basis of phenotypic variation in human populations and translating these differences into susceptibility indexes to develop early diagnosis and disease prevention tools. In recent years, both evolutionary sequence analysis and global gene expression surveys have contributed to regulatory element detection and investigations of variation in gene regulation, but we have yet to understand the nature of genetic variants that functionally affect gene expression, as well as *in silico* predict single-nucleotide changes that would cause phenotypic variation or alter gene expression profiles. Single-nucleotide polymorphisms (SNPs) are the most prevalent form of genetic variation within populations and underlie variations that can influence phenotype or function (Stranger et al. 2007). Functional SNPs have the ability to influence the structure of DNA, RNA, or proteins and their interactions with each other, and these cumulative effects characterize an organism's phenotype as defined by morphology, physiology, metabolism, reproductive fitness, susceptibility to disease and environmental factors. By combining evolutionary genomics and population genetics, it is possible to address the evolutionary contribution from coding SNPs, and a recent study has shown that around 15% of amino acid human mutations are highly deleterious or lethal, an estimate that correlates well with the divergence rate of primate species (Boyko et al. 2008).

Whereas coding mutations reshape the protein anthology of a cell, mutations in noncoding regulatory elements affect the dynamics of gene expression and are the major driving force in establishing species-specific gene expression patterns (Wilson et al. 2008). Mechanistically, functional noncoding mutations disrupt binding of transcription factors to their cognate DNA attachment regions by modifying DNA-binding affinity (Lapidot et al. 2008; Rahimov et al. 2008). Divergence of gene expression driven by regulatory mutations provides an evolutionary mechanism not only for species differentiation and adaptation (Mancini 1991; Zhang et al. 1991) but also for susceptibility to disease (Dermitzakis 2008). For example, in a recent genomewide

association study that compared 3,000 patients with early-onset myocardial infarction to an equivalent cohort of healthy controls, a strong disease association was ascertained for nine SNPs, all of which were present in noncoding regions of the human genome (Kathiresan et al. 2009).

Evolutionary noncoding sequence conservation has been used successfully to identify regulatory elements in vertebrates (Loots et al. 2000; Dermitzakis et al. 2003; de la Calle-Mustienes et al. 2005; Woolfe et al. 2005). Elevated degree of phylogenetic conservation has been used as a tool to parse large intergenic intervals in search of enhancers (Nobrega et al. 2003; Woolfe et al. 2005), and *in vivo* testing of deeply conserved noncoding elements for enhancer activity has demonstrated a high rate of functional validation (Pennacchio et al. 2006). Therefore, we hypothesized that similar methods of analyzing evolutionary sequence conservation can be used to screen for functional noncoding SNPs. However, up to date, reports on the effectiveness of such screens have been conflicting (McCauley et al. 2007; Montgomery et al. 2007; Andersen et al. 2008). To quantify the abundance of and selection against mutations in deeply conserved noncoding sequences, we identified a set of noncoding blocks in the human genome that are perfectly conserved in vertebrates (100% identity; termed bricks) and profiled the occurrence and population genetic characteristics of SNPs in these blocks. Our findings suggest that bricks are significantly overrepresented in the promoters of transcription factors and developmental genes such that the deeper the phylogenetic conservation, the more restricted the brick distribution is in favor of these gene categories. When we examined SNP distribution and frequency across different levels of brick conservation, we found that only a minority of bricks with one diverged central human (CH) nucleotide exhibit human population variation, suggesting that these alleles favor rapid fixation in the primate lineage and may possibly contribute to hominoid and primate adaptation. We also found that SNPs located in bricks are virtually not affected by evolutionary conservation and display little, if any, indication of selective pressure. Comprehensively, our observations argue for the plasticity of gene regulatory mechanisms in vertebrates—and we bring forth evidence of gene regulatory remodeling that has occurred extensively in the human genome. Although we cannot directly study selection on gene regulation in primates and humans, our findings allow us to highlight putative regulatory regions that computationally appear consistent with the action of human-specific (and/or primate-specific) natural selection on gene regulation. Our observations therefore can aid formulate new hypotheses about the prevalence of widespread mutations in putative regulatory elements, which can be addressed experimentally through future profiling, to determine if changes in gene regulation are vital to human evolution and adaptive changes.

Methods

Brick Identification and Gene Annotation

ECR Browser pairwise alignments (Loots and Ovcharenko 2007) of nine vertebrate genomes (human, chimp, macaque, dog, mouse, rat, chicken, frog, and fugu) produced using the blastz local alignment utility (Schwartz et al. 2003) were utilized to construct human and mouse brick sets and to bin those into brick-level groups according to the phylogenetic conservation profile.

Combined RefSeq (Maglott et al. 2000) and University of California–Santa Cruz (UCSC) Known (Hsu et al. 2006) Gene annotations have been utilized to demarcate coding exons, untranslated regions (UTRs), introns, promoters, and intergenic regions. Multiple overlapping gene transcripts have been combined into a single gene locus, and promoters have been defined as 1.5-kb regions immediately upstream of the outermost transcription start site. Promoters extending into the next gene have been truncated to span intergenic interval only. The annotation of repetitive elements was downloaded from the UCSC Genome Browser (Karolchik et al. 2003).

Gene Ontology (GO) analysis (Ashburner et al. 2000) has been performed using well-populated GO categories, to which at least five genes have been annotated. A hypergeometric distribution test has been utilized assessing a probability to observe a given number of genes assigned to a particular GO category given the total number of genes in the test, the total number of genes belonging to the GO category and 18,587 unique genes in the human genome. Bonferroni multiple testing correction applied to redefine significance thresholds has been also implemented.

SNP Mapping and Gene Expression Analysis

Positional information for human and mouse SNPs was downloaded from the National Center for Biotechnology Information (NCBI) SNP database (dbSNP) database (Sherry et al. 2001). Human and mouse releases b126 and b128 SNP sets were utilized, respectively. Human SNPs were overlaid with the HapMap II release 23a data (Frazer et al. 2007) providing detailed genotype information for 3.02 million human SNPs with minor allele frequency (MAF) 1% and higher. F-statistic measuring the ratio of the variation between populations to the variation within populations was utilized to compute the degree of human population differentiation F-statistical measure (Fst) as described in Weir and Cockerham (1984).

Identified SNPs and genotypes were analyzed in the context of population-specific gene expression data generated by Stranger et al. (2007). We downloaded normalized gene expression levels profiled for a panel of 210 HapMap II individuals from European ancestry, Chinese, Japanese, and Yorubian populations. Starting with an SNP within a brick, we examined genes with expression level strongly correlated with the SNP genotype. Up to three gene associations were tested for each SNP—two with flanking genes and one more with the gene containing the SNP (the latter only in case of intronic SNPs). A linear regression fit was used to extrapolate the genotype–phenotype association (Stranger et al. 2007), in which up to three possible genotypes were numbered 1–3. Only SNPs with a major allele frequency of 85% or less were selected for the analysis. Analysis of variance (ANOVA) variation analysis (Davenport 1940) performed using the R statistical package was utilized to

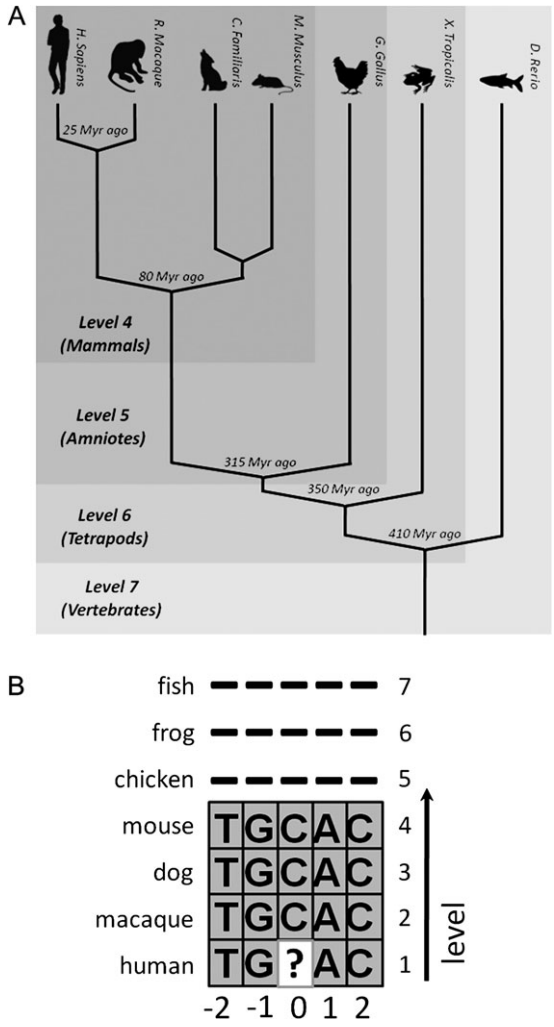

FIG. 1. Defining brick evolutionary levels across the vertebrate lineage through the analysis of seven species (*A*). Schematic structure of a (conservation) brick (*B*). Filled squares depict invariable nucleotides, and nucleotide at position zero was allowed to vary in humans only. Relative evolutionary time is in My.

compute the probability of nonhorizontal linear fit being explained by the variation in gene expression alone.

Results

Defining "Conservation Bricks" as a Template for the Genome Conservation Architecture

To investigate the correlation between recently acquired variation and phylogenetic sequence conservation patterns in the human genome, we first introduced blocks of conservation that serve as an indicator of local negative selection. Specifically, all human 5-mers were identified, for which there exists an orthologous counterpart in another primate and at least two other mammalian genomes. Conservation bricks (referred to as bricks throughout the article) were built using a set of 5-mers requiring sequence identity across a set of species at all positions, except the central nucleotide, which was allowed to differ from the consensus in humans only (fig. 1). By permitting variability at the central brick position in humans only, we could assess the rate of change and predisposition to change, under conditions of perfect evolutionary constraint within the flanking nucleotides. Seven vertebrate genomes served as anchors and allowed us to define the extent of brick evolutionary conservation assessed by the "brick height level" [which could range from four to seven genomes, where level 4 describes ideal conservation in mammals; level 5, amniotes (all land-dwelling vertebrates: human, macaque, dog, mouse, and chicken); level 6, tetrapods (+amphibians: frog); level 7, all vertebrates (+fish) (fig. 1). The number of bricks decreases notably with an increase in brick height, with 51 million mammalian and 406,000 vertebrate bricks. As the majority of level 4 bricks are noncoding in nature (see below), the rapid decrease in the number of bricks based on evolutionary separation of species is in agreement with previously reported analogous rapid decrease in the number of conserved noncoding elements in vertebrates (Loots and Ovcharenko 2007).

Large (several hundred base pairs long) evolutionary conserved regions (ECRs) usually display a variable profile of conservation across their sequence (a combination of multiple blocks of high and low conservation that sum to a cumulative restrictive criteria established by the search; most published records refer to ECRs as >100 bp and >70% identity). We were interested in selectively analyzing only the invariable short islands of conservations within these larger ECRs, which potentially represent the most important functional regions or "cores." To test the association of bricks with regions of evolutionary constraint, we overlapped the distribution of bricks at different levels of conservation with the distribution of highly conserved elements measured using an independent phastCons phylogenetic conservation approach (Siepel et al. 2005). We found that bricks coincide with ECRs significantly more frequently than expected by chance (12- to 35-fold higher number of overlaps than expected; hypergeometric distribution P value $< 10^{-100}$; supplementary fig. S2, Supplementary Material online). This approach allowed us to select regions that are highly conserved independent of a particular multigenome phylogenetic conservation model. We reasoned that 5-mers correspond to the average length of a typical core of a transcription factor binding site (TFBS), such that 5-mer bricks may potentially correspond to TFBS cores that remained invariable in earlier mammals and vertebrates but were possibly mutated in closely related species such as hominoids. The short length of these blocks permitted us to identify a sufficient number of ideally conserved 5-mers abundantly distributed in different functional regions of the genome to provide a uniform representation of the entire genomic landscape (fig. 2).

The Majority of Bricks are Invariable in the Human Genome

We found that 4.9 million bricks are conserved in amniotes (level 5), and 62% of them correspond to noncoding regions. By subjecting amniota bricks to a CH nucleotide variability analysis, it was observed that the CH nucleotide is invariable in 99.3% of level 5 bricks. The fraction of bricks

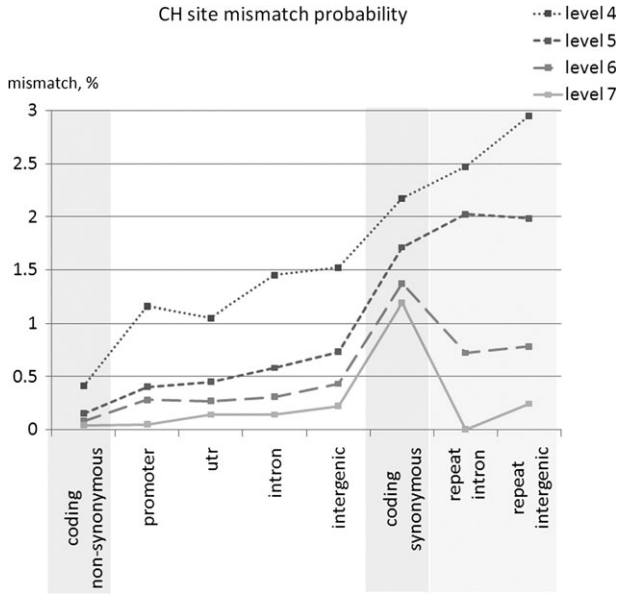

FIG. 2. The fraction of bricks with a mismatch at the CH nucleotide profiled in eight genic categories.

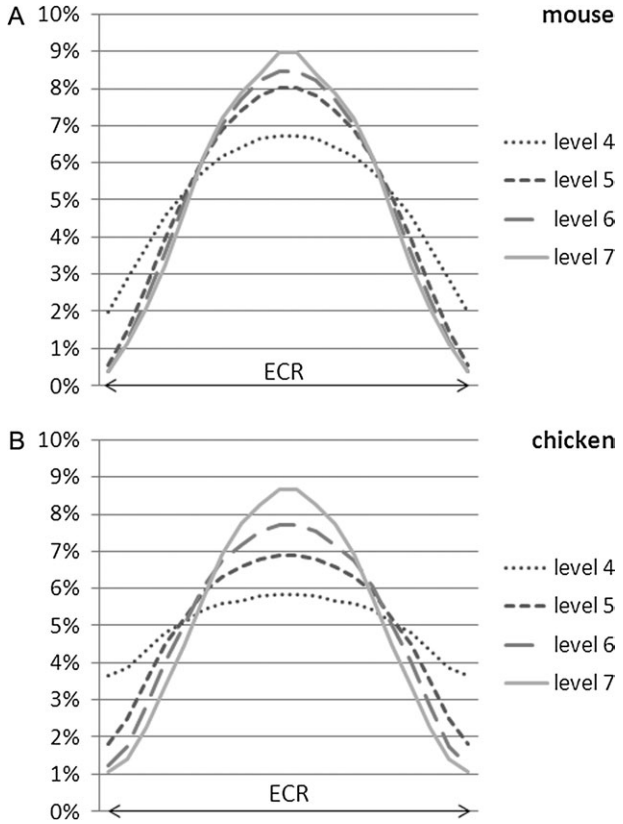

FIG. 3. The histogram of brick distribution within human/mouse (*A*) and human/chicken (*B*) ECRs. The horizontal axis represents a combined set of all ECRs. The position on the horizontal axis depicts one of the 20 ECR relative position bins: 0–5%, 5–10%, . . . 95–100%.

with variability at the CH nucleotide position was found to depend strongly on the genomic context or type of brick (fig. 2). As expected, a much higher CH nucleotide variation rate was observed for synonymous than for nonsynonymous substitutions—an 11.4-fold increase (1.71% vs. 0.15%, respectively)—in coding regions. The strength of purifying selection acting on promoter, UTR, nonrepetitive intronic and nonrepetitive intergenic (referred to as intronic and intergenic, respectively, later on) bricks was higher than at synonymous and repetitive sites, suggesting that nonrepetitive and noncoding amniota (and higher level) bricks are indicative of negative selection, and thus likely to be associated with noncoding functional elements. The mutation rate at synonymous sites, 2.0%, was among the three highest mutation rates in the eight profiled genomic categories (coding synonymous and nonsynonymous, repetitive intronic and intergenic, promoter, UTR, intronic, and intergenic), accompanied by the anticipated high mutation rate in intronic and intergenic repetitive elements. To establish a baseline for the expected mutation rate, we profiled all human–macaque gapless alignment segments that are at least 10 bp long. On average, 8% of nucleotides from these segments were found to be diverged. Thus, the mutation rate at nonsynonymous, promoter, and UTR CH nucleotides is >16-fold lower than the baseline, whereas the mutation rate at repetitive and synonymous positions is ~4-fold lower. This suggests that, although repeats are generally assumed to bear little biological relevance and are expected to diverge neutrally, some bricks residing in repetitive elements might be under negative selective pressure. Additionally, using synonymous sites as the threshold of minimal selective pressure, the analysis of level 6 and 7 brick CH mutations displayed elevated selective pressure in repetitive regions containing bricks. The latter could thus serve as an indicator of functional signals residing within corresponding repetitive DNA sequences that are ancestral in nature.

The central part of an ECR is under increased selective pressure as compared with its flanking regions according to the observed preferential conservation of the ECR central part through the evolution of vertebrates (Ovcharenko et al. 2004; Prabhakar et al. 2006). We utilized ECRs defined as regions that display at least 70% sequence identity in pairwise sequence comparisons over at least 100-bp sequence intervals (Loots et al. 2000) and quantified the positional preference of bricks within these larger segments of conservation. In the case of human/mouse ECRs, we observed a >14-fold increase in the density of bricks in the center of ECRs as compared with the outermost flanks for level 5, 6, and 7 bricks and a 3.4-fold increase for the most abundant level 4 bricks (hypergeometric distribution P value $< 10^{-100}$ in all four cases; fig. 3). This result suggests the presence of elevated purifying selection pressure applied to bricks, with a gradual increase in brick positional bias toward the center of the ECR, which correlates with an increase in the degree of phylogenetic conservation. Assuming that bricks are likely to correspond to key functional elements within ECRs, it remains unclear whether the presence of a brick defines the "core" of a larger functional element (encompassed by an ECR) that is resistant to mutations or whether the central part of a functional

element imposes a stronger selection and thus correlates with an increased density of bricks.

Bricks are Strongly Associated with Transcription Factors and Developmental Genes

To address the function of bricks residing in regulatory regions, we performed a GO functional classification of genes with bricks mapped to their promoter regions (defined as 1.5-kb regions upstream of the transcription start site; supplementary table 1, Supplementary Material online). We observed a strong association of bricks with proximal regulatory elements of transcription factor and developmental (so-called trans-dev) genes (Woolfe et al. 2005). Specificity of functional association, which was assessed by fold enrichment in a functional category, increased with an increase in brick level. The association of bricks with active TFBS within proximal regulatory elements is possible but has not been investigated. Notably, less than 1% of bricks at any brick level were detected in promoter regions, and although the function of some of the remaining bricks could be related to gene regulation, the extent of this relationship is unclear.

Rapid Fixation of Brick Mutations in the Primate Lineage

To estimate the fraction of diverged CH nucleotides that have been fixed in the hominoid lineage since the human–macaque radiation, we aligned diverged bricks and human SNPs (Sherry et al. 2001). Over 97% of diverged CH sites (at any level) do not correspond to a known SNP, thus representing a hominoid-specific mutation fixation event, sequencing error, or rare/missing (as of yet unidentified) polymorphism. Human sequencing errors should have minimal impact on these measurements, as their extent could be upper bounded at 0.05%–the fraction of diverged CH nucleotides at coding nonsynonymous sites (level 7 graph in fig. 2). A recent estimate of the total number of common SNPs in humans amounting to ~11 million (Sabeti et al. 2006) suggests that SNP undersampling is only partially responsible for the limited overlap of diverged CH nucleotides with SNPs. Therefore, we conclude that the observed large fraction of diverged CH sites independent of known SNPs advocates for a widespread lineage-specific fixation of mutations at CH sites, in hominoids and possibly other primates. This observation is in line with the known extensive mutation accumulation followed by rapid fixation in one of the most conserved sets of vertebrate regions—the ultraconserved elements (Chen et al. 2007; Ovcharenko 2008).

By excluding chimp from the comprehensive mammalian brick analysis, it was possible to use its genome sequence to study the recent evolution of bricks in closely related lineages. Specifically, to profile the time line of CH site mutation fixations in hominoids (represented by human and chimp), we compared the human and ancestral (defined using macaque, mouse, and dog at level 4) CH nucleotides with the corresponding nucleotide in the chimp genome. We were able to identify a corresponding chimp nucleotide for over 91% of human diverged CH sites (at all brick levels), where the remainder 9% CH sites corresponded either to gaps or deletions spanning CH sites in the chimp genome (the latter are less likely to be the main contributor than sequencing gaps). For the chimp nucleotides corresponding to CH sites of level 4 bricks, we observed a match to the human nucleotide in 75.3% of the cases and a match to the ancestral nucleotide in 23.5% of bricks. The chimp nucleotide diverged from both human and ancestral nucleotides in only 1.2% of bricks. This provides an estimate of the collective effect of independent mutations in both human and chimp lineages at the same site as well as an impact of potential sequencing errors in the chimp genome. The small overall fraction of inconsistency cases supports the validity of the chimp nucleotide use for the evolutionary change timing. The observation that over three-quarters of human CH nucleotide changes are supported by a corresponding change in the chimp indicates that the fixation of the divergent nucleotide preceded the hominoid radiation (this also serves as indirect validation of human sequence change at the CH site as evolutionary divergence, not sequencing error). About the same fraction of human diverged CH nucleotides with a match to chimp was observed in bricks of different conservation levels (supplementary table 2, Supplementary Material online), which indicates that the effect does not depend on the preceding history of brick conservation but depicts the profile of mutation fixation in the primate lineage. The 3.2-fold higher number of fixed mutations in primates preceding the hominoid radiation as compared with human-specific post-hominoid radiation mutations relates well to the 6 My and 25 My of evolutionary separation of humans and chimps and humans and macaques, respectively. Collectively, these results argue for the pronounced mutation fixation events throughout the evolution of primates, with the rate of mutation fixation in hominoids being comparable with that of earlier primates.

Human Population Variation in Bricks Suggests Release in Evolutionary Pressure on SNP Brick Nucleotides

To determine the rate of ongoing mutations at CH sites within the human population, we examined the distribution of SNPs associated with a brick of a given level centered at the SNP location and evaluated whether the occurrence of a common SNP at a CH site depends on the evolutionary depth of the brick. For this purpose, we collected 51,371 HapMap SNPs (Frazer et al. 2007), which coincide with CH sites of bricks level 4 and higher. These SNPs represent 1.7% of all HapMap SNPs and were termed "brick SNPs." One would expect strong purifying selection acting on CH sites to result in either lower MAF or lower population differentiation (assessed here using the Fst; Weir and Cockerham 1984) for brick SNPs. Contrary to our expectations, we observed only a minimal decrease in Fst and MAF for brick SNPs level 4 to level 6 (level 7 MAF was higher than the HapMap average and that might be attributed to a very small number of level 7 brick SNPs); additionally, the observed MAF/Fst decrease is much lower than the standard deviation (table 1).

Table 1. Average MAF and Fst of brick SNPs.

	HapMap SNPs	Level 4	Level 5	Level 6	Level 7
MAF	0.133 (0.120)	0.132 (0.119)	0.128 (0.119)	0.124 (0.114)	0.144 (0.136)
	0.000	0.001	0.002	0.004	0.011
Fst	0.207 (0.145)	0.199 (0.146)	0.186 (0.145)	0.175 (0.144)	0.189 (0.138)
	0.000	0.001	0.003	0.005	0.011
SNP count	3.0 million	47,045	3,329	842	155

NOTE.—Standard deviation is given in parenthesis. Standard error of the mean is under the standard deviation.

Although unlikely, it could be that the observed wide distribution of MAF and Fst values precluded us from detecting the purifying selection acting on brick SNPs. To measure the variability of brick SNPs relative to the evolutionary history of bricks, we compared the fraction of mutated CH sites within the brick SNP data set with the average CH site mutation rate. Depending on the brick level, we observed a 21- to 59-fold increase in the CH site mutation rate at brick SNPs as compared with all bricks (table 2). Although brick SNPs are polymorphic by definition, and variation of at least 1% is expected to be observed at brick SNP sites in any human genome due to chance, the brick SNP mutation rate can be estimated close to 1.6% (or the average level 4 CH site mutation rate). However, for level 4 brick SNPs, the mutation rate was found to be significantly higher at 33.1% change. As this analysis was performed using a particular human genome assembly (NCBI Build 36.1)—the same assembly as for the previous analysis of the CH site variability in all bricks—and resulted in about one-third of all brick SNP sites being diverged, this indicates that the selective pressure on polymorphic brick SNP sites in the human lineage is much lower, if present at all, than the selective pressure acting on an average brick CH site. Interestingly, the observed CH site variability in SNP bricks is over 10-fold higher than the CH site variability of bricks centered on synonymous alleles (table 2 and fig. 2). In summary, we found a disproportionally large number of SNP bricks with the CH nucleotide diverged in a single human genome. The rate of CH nucleotide mutation was found consistently high across all brick levels and did not follow the evolutionary imposed constraints on CH sites but was similar to the average mutation rate at regular SNP sites.

Mouse Brick Population Variation Supports the Hypothesis of Lineage-Independent Brick Profiling

It was unclear whether the observed elevated mutation rate at brick SNP sites reflects effects specific to these particular sites or a local selective pressure relaxation in a larger region encompassing the SNP position. To address the extent of this effect, we measured the CH site mutation rate at positions flanking the SNP location. This was accomplished using bricks centered on each of the four nucleotides immediately adjacent to the SNP site (two on each side), if the bricks were available for a particular SNP. To accomplish this, we scanned 5-bp windows centered on each human SNP to determine if the SNP overlaps with a brick (not necessarily the CH nucleotide). If an overlap was identified, we noted the variation at the brick CH site and the distance separating the CH site and the SNP. We found bricks centered on SNPs to have a significantly higher mutation rate at CH position than bricks located in the immediate vicinity of SNPs, which had a mutation rate at the neighboring nucleotides corresponding to the average CH brick mutation rate level (fig. 4A). We also included +3/−3 flanking sites into the analysis to address a possible codon-like three-nucleotide periodicity in nucleotide variability. No mutation rate spikes were observed at either +3 or −3 positions indicating that if the codon-like relaxation of the selective pressure on synonymous sites had any impact on the observed effect, its contribution was negligible (fig. 4A). This finding suggests that the elevated negative

Table 2. Mismatch frequency at brick CH nucleotide sites.

	Level 4 (%)	Level 5 (%)	Level 6 (%)	Level 7 (%)
All bricks	1.6	0.7	0.5	0.5
Brick SNPs	33.1	31.8	29.7	23.2

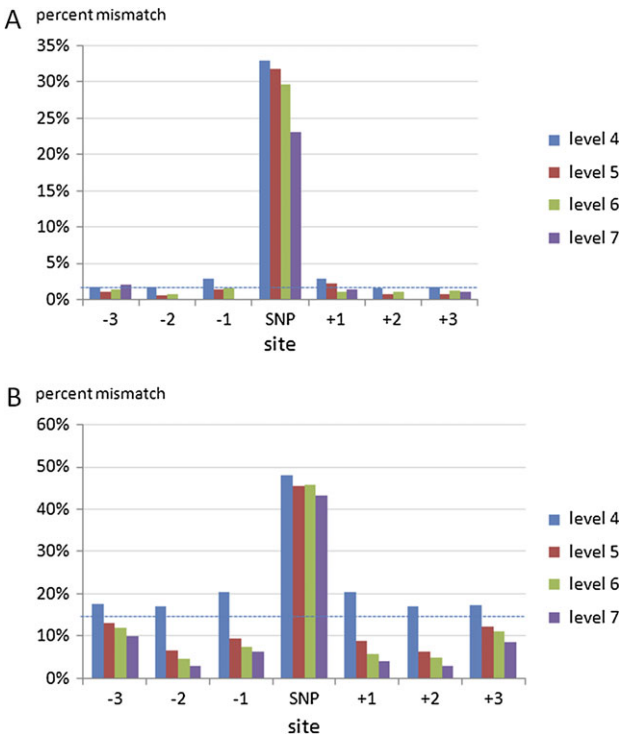

FIG. 4. Percent mismatch at the brick SNP site and three flanking nucleotides around the SNP site. Level 4 average mismatch rate in the human genome is plotted by a blue dashed line. (*A*) Human and (*B*) mouse reference sequence.

selection pressure remains almost intact at non-SNP sites, whereas the reversal or release from selective pressure is specific and distinct at the CH brick sites corresponding to known SNPs. Given the evolutionary history of the CH site conservation in multiple species, it is likely that the observed mutations at brick SNP sites are suggestive of recent evolutionary events or possibly adaptive effects.

Recent reports have ignited a new debate centered on whether there is a loss of purifying selection in the primate lineage or not (Keightley et al. 2005; Kryukov et al. 2005; Ovcharenko 2008). It is unclear whether a possible loss of purifying selection in hominoids could be the primary explanation for the observed high variability rate in brick SNP nucleotides. To address this question, we compared the human brick variability profile surrounding SNP sites with the corresponding evolutionary profile in the murine lineage. The availability and large size of the mouse SNP data set consisting of 14.9 million SNPs deposited into the dbSNP database (Sherry et al. 2001) that were mapped to the mouse genome (Karolchik et al. 2003) facilitated this analysis. Human and mouse genomes were permuted in the original brick definition (fig. 1) to identify bricks in reference to the mouse genome. The number of mouse bricks varied from 56.7 million in mammals (level 4) to 459,000 in vertebrates (level 7)—displaying an evolutionary trend highly similar to the observed frequencies of human bricks. The brick variability surrounding mouse SNP sites resembled the results obtained for humans (a 3-fold mutation rate increase at level 4; fig. 4*B*). Effective recapitulation of the elevated brick SNP divergence mutation rate in human and mouse lineages argues against the hypothesis that a relaxed primate-specific selection is the primary contributor to an increase in the mutation rates but rather argues that these events are widespread across different lineages. This also suggests that individual brick variation could potentially result in mild phenotypic effects, whereas collectively, brick mutations could synergistically be utilized in evolutionary adaptations, in accord with a previously proposed hypothesis (Kryukov et al. 2005). Following the same line of reasoning, one would expect brick SNPs originating from the most conserved bricks to have the most profound phenotypic impact.

Unlike results obtained in reference to the human genome, the brick mutation rate at non-SNP sites was noticeably elevated in mouse bricks (fig. 4). It could be argued that the presence of the macaque genome in the set of mammalian species played a key role in ensuring a low average human brick mutation rate as the conservation of the CH nucleotide and flanking sites in the macaque genome (which is closely related to the human genome) might have biased their conservation in the human genome. To test the impact of this effect, we repeated the mouse brick analysis after substituting macaque with rat (because the rat genome is the closest to the mouse genome) in the mouse brick definition. As expected, the average level of divergence measured for level 4 bricks decreased from 14.6% to 4.0% (supplementary fig. S1, Supplementary Material online), confirming that the presence of a closely related species defines the absolute level of divergence at non-SNP sites. However, replacing macaque with rat did not have a qualitative impact on the spike in mutation rate at the central brick site. In contrast, this effect was further amplified by increasing the mutation rate at SNP sites from 3- to 10-fold (supplementary fig. S1, Supplementary Material online).

Population-Specific Differences in Brick SNPs

By comparing CH nucleotide divergence across different levels of brick conservation, we found that level 6 and 7 brick divergence rates for all tested noncoding nonrepetitive categories are lower than the level 4 coding nonsynonymous divergence rates (fig. 1). This trend is intermediate for level 5 bricks, with CH nucleotide divergence rate in noncoding nonrepetitive categories found to be 2- to 3-fold lower that the level 4 brick CH nucleotide divergence rate in the corresponding categories, and about the same as the level 4 brick CH nucleotide divergence rate at nonsynonymous sites. This evidence of negative selection suggests that mutations at CH sites of level 5 and higher bricks are more likely to have phenotypic effects than level 4 CH site mutations. We investigated the genomic distribution and population specifics of brick SNPs corresponding to the set of noncoding nonrepetitive level 5 and higher bricks (termed "hotSNPs"), using these SNPs as proxy to gene regulatory mutations with an increased likelihood of a phenotypic effect in the human lineage. As an example, the derived G allele of one of the hotSNPs, rs12469063 (16.5% MAF, Fst = 0.14), resides in a deeply conserved region of the *MEIS1* intron and is known to have a highly significant association with the restless legs syndrome (Schimmelmann et al. 2009). In total, we identified 2,863 hotSNPs, 31% of which correspond to a brick with a human-specific mismatch at the CH site. We found ~1% of hotSNPs to reside in promoter regions, ~5% in UTRs, whereas the majority of hotSNPs (~51%) were found to reside more than 10 kb away from the nearest gene. Thirty-three percent of hotSNPs were found in regions devoid of genes known as gene deserts (noncoding regions $\geq$ 500 kb; Nobrega et al. 2003; Ovcharenko et al. 2005). This represents a significant 1.3-fold higher density of hotSNPs in gene deserts than the density of level 5 bricks ($P = 2.7 \times 10^{-15}$, hypergeometric distribution). These findings, in addition to previously published observations that deeply conserved noncoding elements in gene deserts often correspond to transcriptional regulatory elements (Nobrega et al. 2003; de la Calle-Mustienes et al. 2005), would suggest that evolutionary profiling SNPs in distant elements is likely to yield functional relationships at higher rates than those in proximal elements, at least in primates. In addition, increased population differentiation (Fst) could serve as an indicator of positive selection. We identified 609 (21.2%) hotSNPs with high Fst (Fst > 0.2), which is comparable with 22.1% of all brick SNPs with high Fst. As an indicator of positive selection acting at high-Fst sites, high-Fst hotSNPs displayed much higher CH mutation rates than the remaining set of hotSNPs—37% versus 29% (*P* value

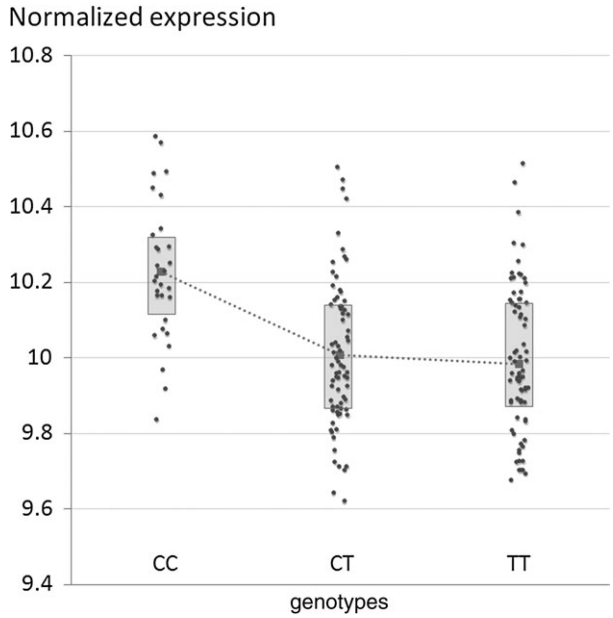

FIG. 5. *MAP3K7* gene expression level sampled in 210 HapMap individuals correlates with the rs9342216 SNP genotype. Gray boxes encompass two central quartiles of the distribution, and dotted lines connect average expression levels of CC, CT, and TT phenotypes ($P = 3.2 \times 10^{-6}$).

of 1.52×10^{-7}). This further strengthens the hypothesis that positive selection with rapid fixation acts on a subset of brick SNPs.

It is possible that some brick SNPs are associated with changes in gene expression. To examine this hypothesis, we analyzed expression data generated by Stranger et al. (2007) for 210 unrelated individuals from the HapMap panel to identify significantly correlated hotSNP genotypes and expression phenotypes of neighboring genes. Linear regression extrapolation was followed by an ANOVA analysis of the variation in the linear fit as previously described (Stranger et al. 2007). The genotype–phenotype association was tested between a hotSNP and the closest transcript with recorded expression on each side of the hotSNP. For hotSNPs located within a transcript with recorded expression, additional association with the transcript containing the hotSNP was measured as well (for a total of three profiled transcripts per such hotSNP). Thirty-five percent of the 2,194 common hotSNP genotypes were significantly correlated with a gene expression phenotype with a *P* value of 0.05. Sixty-five (3%) of common hotSNP genotype–phenotype associations maintained significance after multiple testing corrections using Bonferroni adjustment ($P < 0.000023$; supplementary table 3, Supplementary Material online). In 10 of the 65 strong SNP–gene associations, the hotSNP was found to reside in the gene the SNP was associated with, but in the majority of cases, the association was with a neighboring gene. One of the examples of the latter cases is the rs9342216 hotSNP. No significant association was found between this SNP and the expression level of the *BACH2* gene spanning it, whereas a strong association was observed with the neighboring *MAP3K7* (*TAK1*) oncogene, located 480 kb away from the SNP (fig. 5). In particular, there is an increased expression level associated with the derived C/C rs9342216 genotype, and it is interesting that its effect is population specific—the C/C genotype is almost uniquely associated with the European population, in which the frequency of this genotype is 54%. Due to the nature of the experimental data that is specific to lymphoblastoid cell lines, these results describe only a limited set of transcripts highly specific to a particular cellular state. Therefore, there are transcripts for which genotype–phenotype associations are inaccessible using this approach. Also, there are multiple regulatory elements affecting gene expression at different time points, distant gene regulatory activities, and other effects that collectively hinder a comprehensive analysis. The absence of genotype–phenotype association in the reported tests should therefore be treated with caution and cannot be used to conclusively reject a functional link. For example, no significant association was detected between the previously mentioned rs12469063 SNP and the *MEIS1* gene linked to this SNP, whereas both of them are independently associated with and linked to the restless leg syndrome.

Discussion

There are two schools of thought in computational and evolutionary genomics that are rapidly joining forces in characterizing the gene regulatory architecture of the human genome, both of them aiming to determine the contribution of gene regulation to phenotypic differences in human and animal populations. Studies in evolutionary genomics have been successful in arguing the case for strong associations between deep evolutionary conservation and functional importance of noncoding DNA segments, including gene regulatory regions (Woolfe et al. 2005; Pennacchio et al. 2006). In parallel, a large body of work in population genetics, including whole genome association studies, has presented compelling evidence for the correlation between noncoding polymorphisms and levels of gene expression and disease phenotypes (Emison et al. 2005; De Gobbi et al. 2006; Zhou et al. 2008). The cross junction of these two fields prompts formulating hypotheses on the phenotypic impact of polymorphisms present in noncoding regions that have been deeply conserved in vertebrates. Up to date, no robust arguments have been presented that strengthen the claim that polymorphisms present in deeply conserved region are more likely to have a phenotypic impact (Montgomery et al. 2007; Stranger et al. 2007). The question that we addressed in this study is whether predisposition to conservation, depicted by perfect (i.e., 100%) local sequence identity in multiple species that is associated with signatures of strong purifying selection, when compromised by variation in human populations could be indicative of a recently emerged phenotypic variability. In other words, from the statistical point of view, how probable is it to observe a SNP in

a deeply conserved region and how likely is it that such an observation is associated with a biological consequence.

We found that about 0.5% of the human genome consists of 5-mers perfectly conserved in amniotes (represented by human, mouse, dog, and chicken genomes in this study; fig. 1), if we permit, but not require, a single substitution only in the human sequence. These 5-mers, here termed "bricks," reside primarily in noncoding regions (62% are noncoding) and provide a reliable proxy to identify regions under purifying selection in the human genome—the average 0.5% human–chimp mutation rate at the central nucleotide of nonrepetitive noncoding amniota bricks is much lower than the average 8.0% human–chimp mutation rate. We observed a strong correlation of bricks with trans-dev genes, and their prevalence at sites remote from transcription start sites correlates with the known abundance of distant regulatory elements in loci of developmental genes (Woolfe et al. 2005).

To address the importance of a SNP occurring in the center of a brick and the likelihood of this observation to infer a phenotypic association, we aligned the distribution of HapMap SNPs (MAF $\geq$ 1%) with the distribution of bricks. At an average density of 1.4 SNPs/kb in noncoding nonrepetitive regions of the human genome, the SNP density at a brick CH site was significantly lower at all levels of conservation (1.4-, 2.0-, 2.8-, and 4.0-fold decrease at levels 4–7, respectively; *P* value of any of these observations assessed using binomial distribution is less than 10^{-10}). The correlation of SNP density fold-decrease with the degree of evolutionary conservation, no inherited bias in SNP profiling toward the location of bricks in experimental SNP discovery protocols, and the exclusion of repetitive regions from the analysis (in which the ambiguity in SNP mapping could potentially inadvertently affect SNP density) reject the possibility of ascertainment bias as the dominant cause of these results. Instead, the pronounced depletion of SNPs at CH sites reflects strong negative selection acting on CH sites in the human lineage, with the degree of brick evolutionary conservation being reflected in the strength of purifying selection on a CH site.

Whereas a strong negative selection was recorded for brick CH sites (supported by a significant decrease of both mutation rates and SNP density), we find little evidence of negative selection acting on brick CH sites polymorphic in either the human or mouse population. There is no substantial decrease in either MAF or Fst measures of divergence for SNPs residing at brick CH sites (table 1), and the mutation rate at these sites is over 20-fold higher than what would be expected given the evolutionary predisposition to mutations at brick CH sites (table 2). There is an important practical outcome of these observations: Although the presence of a brick SNP might indicate a functional polymorphism (with the degree of association being proportional to the level of phylogenetic brick conservation), a high MAF value of a brick SNP or a notable population differentiation of its genotype is unlikely to differentiate that SNP from the pool of all other brick SNPs. Thus, population genetics data of an individual brick SNP is unlikely to serve as a good measure of functional prioritization or represent an indicator of positive selection.

We observed a large fraction of primate CH site mutations—up to 97%—which have been fixed in hominoids. Taken together, these data advocate for strong negative selection acting on bricks, which is being counterbalanced by widespread variation due to mutations at brick CH sites. This effectively permits profiling for advantageous mutations at functional sites, which is being followed by rapid fixation of either one or another allele. Additionally, these results suggest association of bricks with functional elements under negative selection and a plasticity of gene regulatory networks in vertebrates with an efficient use of mutations in bricks in the adaptation of species. With the majority of mammalian and amniota bricks being noncoding, it is fair to assume that this general trend is likely to represent an evolutionary adaptation of gene regulatory mechanisms, at least in the tetrapod lineage.

The identified groups of brick SNPs we introduced here provide a SNP selection template for phenotypic associations and medical diagnostics. We also describe here a smaller subset of 65 SNPs for which we can detect a significant association between a SNP genotype and a known gene expression phenotype. In summary, our data suggest that the strong local conservation surrounding an SNP could be useful for prioritizing candidate functional SNPs within a haplotype block; however, it is likely that only mild phenotypic effects will be associated with individual brick SNPs and, therefore, the presence of an SNP in a deeply conserved element in a locus might not be sufficient to suggest discovery of a regulatory SNP. Adaptive evolution of humans displays a multifaceted nature with a large number of low-impact contributors, which, in turn, leads to a conclusion of multiple regulatory mutations shaping the architecture of an average gene locus.

Supplementary Material

Supplementary tables S1–S3 and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

G.G.L. was supported by National Institutes of Health (NIH) grant HG003963. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. I.O. was supported by the Intramural Research Program of the National Library of Medicine, NIH.

References

Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J. 2008. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol*. 4(1):e5.

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 25(1):25–29.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 4(5):e1000083.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet*. 80(4):692–704.

Davenport CB. 1940. Analysis of variance applied to human genetics. *Proc Natl Acad Sci U S A*. 26(1):1–3.

De Gobbi M, Viprakasit V, Hughes JR, et al. (15 co-authors). 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312(5777):1215–1217.

de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res*. 15(8):1061–1072.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302(5647):1033–1035.

Dermitzakis ET. 2008. From gene expression to disease risk. *Nat Genet*. 40(5):492–493.

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. 2005. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* 434(7035):857–863.

Frazer KA, Ballinger DG, Cox DR, et al. (249 co-authors). 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC Known Genes. *Bioinformatics* 22(9):1036–1046.

Karolchik D, Baertsch R, Diekhans M, et al. (14 co-authors). 2003. The UCSC Genome Browser Database. *Nucleic Acids Res*. 31(1):51–54.

Kathiresan S, Voight BF, Purcell S, et al. (184 co-authors). 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*. 41(3):334–341.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol*. 3(2):e42.

Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet*. 14(15):2221–2229.

Lapidot M, Mizrahi-Man O, Pilpel Y. 2008. Functional characterization of variations on regulatory motifs. *PLoS Genet*. 4(3):e1000018.

Loots G, Ovcharenko I. 2007. ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics* 23(1):122–124.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288(5463):136–140.

Maglott DR, Katz KS, Sicotte H, Pruitt KD. 2000. NCBI's LocusLink and RefSeq. *Nucleic Acids Res*. 28(1):126–128.

Mancini GB. 1991. Hypertension, hypertrophy, and the coronary circulation. *Circulation* 83(3):1101–1103.

McCauley JL, Kenealy SJ, Margulies EH, Schnetz-Boutaud N, Gregory SG, Hauser SL, Oksenberg JR, Pericak-Vance MA, Haines JL, Mortlock DP. 2007. SNPs in Multi-species Conserved Sequences (MCS) as useful markers in association studies: a practical approach. *BMC Genomics*. 8:266.

Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJ. 2007. A survey of genomic properties for the detection of regulatory polymorphisms. *PLoS Comput Biol*. 3(6):e106.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302(5644):413.

Ovcharenko I. 2008. Widespread ultraconservation divergence in primates. *Mol Biol Evol*. 25(8):1668–1676.

Ovcharenko I, Stubbs L, Loots GG. 2004. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* 84(5):890–895.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res*. 15(1):137–145.

Pennacchio LA, Ahituv N, Moses AM, et al. (19 co-authors). 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444(7118):499–502.

Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*. 16(7):855–863.

Rahimov F, Marazita ML, Visel A, et al. (22 co-authors). 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet*. 40(11):1341–1347.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312(5780):1614–1620.

Schimmelmann BG, Friedel S, Nguyen TT, et al. (22 co-authors). 2009. Exploring the genetic link between RLS and ADHD. *J Psychiatr Res*. 43:941–945.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res*. 13(1):103–107.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 29(1):308–311.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15(8):1034–1050.

Stranger BE, Forrest MS, Dunning M, et al. (17 co-authors). 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813):848–853.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* 322(5900):434–438.

Woolfe A, Goodson M, Goode DK, et al. (16 co-authors). 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*. 3(1):e7.

Zhang XL, Bellett AJ, Hla RT, Braithwaite AW, Mullbacher A. 1991. Adenovirus type 5 E3 gene products interfere with the expression of the cytolytic T cell immunodominant E1a antigen. *Virology* 180(1):199–206.

Zhou Z, Zhu G, Hariri AR, et al. 2008. Genetic variation in human NPY expression affects stress response and emotion. *Nature* 452(7190):997–1001.