Abundant Raw Material for *Cis*-Regulatory Evolution in Humans

Matthew V. Rockman and Gregory A. Wray

Department of Biology, Duke University

Changes in gene expression and regulation—due in particular to the evolution of *cis*-regulatory DNA sequences—may underlie many evolutionary changes in phenotypes, yet little is known about the distribution of such variation in populations. We present in this study the first survey of experimentally validated functional *cis*-regulatory polymorphism. These data are derived from more than 140 polymorphisms involved in the regulation of 107 genes in *Homo sapiens*, the eukaryote species with the most available data. We find that functional *cis*-regulatory variation is widespread in the human genome and that the consequent variation in gene expression is twofold or greater for 63% of the genes surveyed. Transcription factor-DNA interactions are highly polymorphic, and regulatory interactions have been gained and lost within human populations. On average, humans are heterozygous at more functional *cis*-regulatory sites ($>$16,000) than at amino acid positions ($<$13,000), in part because of an overrepresentation among the former in multiallelic tandem repeat variation, especially $(AC)_n$ dinucleotide microsatellites. The role of microsatellites in gene expression variation may provide a larger store of heritable phenotypic variation, and a more rapid mutational input of such variation, than has been realized. Finally, we outline the distinctive consequences of *cis*-regulatory variation for the genotype-phenotype relationship, including ubiquitous epistasis and genotype-by-environment interactions, as well as underappreciated modes of pleiotropy and overdominance. Ordinary small-scale mutations contribute to pervasive variation in transcription rates and consequently to patterns of human phenotypic variation.

Introduction

Variation in noncoding *cis*-regulatory DNA sequences has been advanced as a major component of the genetic basis for phenotypic evolution (Britten and Davidson 1969; King and Wilson 1975; Stern 2000; Tautz 2000; Carroll, Grenier, and Weatherbee 2001; Davidson 2001; Stone and Wray 2001; Enard et al. 2002). This view is supported by interspecific analyses documenting dramatic changes in patterns of gene expression coupled with functional and structural conservation of proteins. Models of regulatory rewiring and co-option have proliferated to explain the evolution of developmental and morphological diversity. However, although intraspecific protein sequence polymorphism has been studied for decades, we lack even basic information about the level and nature of functional *cis*-regulatory variation within populations. Heritable variation in gene expression has been described (Damerval et al. 1994; Cavalieri, Townsend, and Hartl 2000; Jin et al. 2001; Brem et al. 2002), but the genetic basis of this variation represents, with very few exceptions (Stam and Laurie 1996; Crawford, Segal, and Barnett 1999; Schulte et al. 2000), a major lacuna in the literature of molecular evolution. We have turned to the literature of human medical genetics to fill this gap. Characterization of functional *cis*-regulatory variation, its effects on transcription, and its consequent influence on phenotypes is critical for a complete understanding of the genetic basis of phenotypic evolution. Molecular evolutionary studies of regulatory polymorphisms represent a promising avenue for the synthesis of the quantitative models of molecular population genetics and for the mechanistic and descriptive studies of developmental and macroevolutionary phenomena.

Variation in coding sequence is classified as nonsynonymous and synonymous, reflecting nucleotide variants with and without an effect on phenotype at the level of protein primary structure. Currently, however, we lack the means to distinguish functional regulatory variants in noncoding DNA from nucleotide sequences alone. Although noncoding variation has been implicated as the genetic basis for phenotypic variation by means of QTL mapping and association studies (e.g., Sucena and Stern 2000), such methods are typically unable to attribute effects or modes of action to specific nucleotides because of potential complications from linkage disequilibrium. We therefore assembled from the literature a data set of human *cis*-regulatory polymorphisms meeting stringent experimental criteria for functionality. Specifically, the functional import of each polymorphism has been validated by allele-specific reporter construct assays in cell culture. From this data set, we estimate the basic characteristics of functional *cis*-regulatory variation and describe the implications of this variation for regulatory and phenotypic evolution.

Materials and Methods

Literature Survey

We searched the literature (through December 2001) for articles relating to regulatory polymorphisms in humans. We used data from these articles to construct several nested data sets relating to sequence variation, reporter assays, allele frequencies, phenotypic associations, and protein binding. The final data sets, and the full list of the more than 400 references from which they are extracted, are available as supplementary material. The study of *cis*-regulatory variants introduces some unique challenges with consequences for our analyses. Unlike coding sequence, *cis*-regulatory function is inherently context dependent. In particular, the functional

Key words: promoter, polymorphism, gene regulation, evolution of development, gene network, transcription.

Address for correspondence and reprints: Matt Rockman, Department of Biology, Duke University, Box 90338, Durham, North Carolina 27708. E-mail: mrockman@duke.edu.

Mol. Biol. Evol. 19(11):1991–2004. 2002

consequences of a *cis*-regulatory polymorphism depend on cell type, temperature, the distribution of exogenous inducers, and covariation at other sites in the genome. Moreover, experimental methods used to identify functional variants introduce additional variation; for example, a variant may have no effect on transcription unless the construct includes introns or downstream elements that physically interact with it to transduce the transcriptional output. The spatial extent of relevant DNA is not defined, and important regulatory elements can lie hundreds of kilobases from transcription start sites. The consequence is that whereas coding variants are readily classified as synonymous or nonsynonymous, assignment of functional consequences to a *cis*-regulatory polymorphism requires analysis of the variants over a vast multidimensional range of conditions. Standardized, high-throughput methods will fail to detect many, if not most, functional variants. We therefore view literature survey, in which each variant has been studied in depth by specialists, as the best approach presently available to study the dynamics of functional *cis*-regulatory variants. Our reliance on a sample of convenience introduces certain deviations from a perfectly representative sample of variants from the genome. Any study of a selected subset of the genome, including recent large-scale studies of coding sequence variants in medically interesting genes (e.g., Cambien et al. 1999; Cargill et al. 1999; Halushka et al. 1999), relies on a nonrandom sample. We discuss the consequences of our sampling strategy below. Our survey stands in the tradition of other early analyses of human genetic samples of convenience that, despite their lack of strict random sampling, expanded our understanding of human genetic variation and evolution, including population subdivision (Lewontin 1972), molecular evolution (King and Wilson 1975), and nucleotide diversity (Li and Sadler 1991).

Inclusion Criteria

The primary data set consists of protein coding loci whose transcription is influenced by *cis*-regulatory polymorphisms. These genes meet two inclusion criteria. First, sequence variants have a statistically significant effect on transcription rate in the context of allelic reporter constructs transfected into a physiologically relevant cell line. Second, the responsible variants have rare allele frequencies greater than 1% in a human population. Genes are included if a reporter construct experiment shows allelic differences in transcription rate that the authors of the study characterize as statistically significant, even if other experiments show no significant differences. This approach is necessitated by the pervasive context-dependence of regulatory DNA function. The 1% frequency criterion eliminates a large number of rare variants whose pathological consequences brought their carriers to clinical attention. Functional noncoding variants that affect RNA stability and localization, splicing, and translational initiation or efficiency, though important components of the functional noncoding genome, are excluded (e.g., UTR variants in *LPA*, *TYMS*, and *F12*). In addition, we have excluded somatically unstable repeats (‘‘dynamic mutation’’) that affect transcription (e.g., repeats near *SIX5*, *FMR1*, *CSTB*, and *FRDA*). These noncoding variants may cause disease in their hyperexpanded forms, but the effect of ordinary, nonpathological variation in repeat number at these loci is unknown. Our experimental criterion excludes many likely functional variants for which the appropriate experiments have not yet been performed. For example, several polymorphic nucleotide variants are known to affect the binding of transcription factors and have also been implicated in variation in transcription rate by way of fine-scale linkage mapping and association studies, but in the absence of reporter construct assays these genes (e.g., *COL1A1* and *VWF*) are omitted from our data set. The primary data set includes 107 genes, of which 106 are separately named genes; the *GCK* gene is counted twice because it has two tissue-specific promoters and first exons, separated by 16 kb, each with functional *cis*-regulatory polymorphisms.

Fold-Difference Estimates

We tabulated the fold-difference in transcription level between alleles at each locus. These data represent the fold-difference observed in the experimental design giving the largest average difference between alleles. This maximum difference may be justified as likely the most physiologically important situation, but its use is also a practical necessity. Because expression level in vivo is a function of cell type, inducer concentration, and genetic background, and because experimental study introduces the additional variable of reporter construct design, there is no way to characterize a ‘‘typical’’ expression difference between alleles at a locus. In the extreme case, for instance, alleles will show no difference in transcription rate when present in cell types that do not express the gene. Use of the maximum difference introduces two opposing ascertainment biases. First, the statistical power to detect small differences is a function of sample size (the number of independent transfection experiments) and the magnitude of the experimental error, so small differences between alleles may be underrepresented relative to their true frequency. Second, many variants have been studied in only one or a few cell types and conditions, and none has been observed in embryos, so for most loci the true maximum allelic difference in transcription rate may not have been observed, leading to an underrepresentation of large differences.

Corroborative Association Studies

We sought corroboration of functionality for the experimentally validated variants by identifying published associations between the variants and expected phenotypes. At the biochemical level, we tabulated statistically significant associations between the functional *cis*-regulatory variants and in vivo measurement of gene expression. At the organismal phenotype level, we tabulated statistically significant associations between the variants and any phenotype that the authors of the study expect to be associated with the expression of the gene. The phenotypes include morphometric, physiometric, and psychometric traits, disease risk and outcome, and pathogen susceptibility and transmission risk.

Functional Variants

Data sets of mutation types and positions were assembled from 101 genes in the 107-gene primary data set. Variants are included in these secondary data sets only if the specific nucleotide variant is implicated in allelic variation in transcription by reporter construct experiments. Consequently, six of the 107 genes in the primary data set are not represented because allelic variation is experimentally attributable only to haplotypic variation; which of the varying nucleotides and how many of then contribute to the transcription rate variation are unknown for these genes. Other genes are represented by multiple, individually tested functional variants. The data set of mutation types includes 144 observations. Biallelic tandem repeat polymorphisms are included in the count of indels, whereas multiallelic tandem repeat polymorphisms are classified separately as variable number tandem repeats (VNTRs). We separated single nucleotide polymorphisms (SNPs) into categories without regard to strand orientation (for example, a T/G SNP is counted together with A/C SNPs because they represent the same base-pair exchange, differing only in strand). This approach allows comparison with published tallies of genomic SNP types and is additionally justified by the fact that *cis*-regulatory DNA function is mediated by the structure of double-stranded DNA, i.e., there is no *cis*-regulatory ‘‘sense’’ strand. The data set of positions of variants relative to transcription start sites includes 141 observations from 100 genes, after excluding genes whose starts of transcription are inadequately mapped. For first exons with variable start sites (i.e., multiple mapped starts within a 100-bp region), we use the 5′-most major start site. For variants occupying multiple positions (e.g., indels and VNTRs), we calculated distance from the nucleotide of the variant that is nearest to start of transcription.

Protein-DNA Interactions

We assembled a data set of functional nucleotide variants (i.e., those in the mutation types data set) whose interactions with transcription factors have been experimentally determined to show allelic variation. The experimental criterion for inclusion in this data set is the demonstration of differential transcription factor binding in an electromobility shift assay using each allele as a probe for binding to proteins from nuclear extracts from relevant cell types. We use this experimental criterion because although transcription-factor binding sites can be predicted on the basis of sequence similarity to a consensus sequence, binding affinity often varies within the consensus, many consensuses are ill-defined, and many binding sites are as yet wholly uncharacterized. We have not counted inferred binding polymorphisms without experimental verification; for example, several functional variants that alter TATAA boxes (adjacent to the *CYP2A6*, *HSD17B2*, and *UGT1A1* genes) are excluded from this tally because differential protein binding has not been demonstrated by gel shift assays.

Frequency Data

We collected published allele frequency data for 129 of the functional variants. These data include observations of 241,008 chromosomes, a mean of 1,868 per variant (median 1,049). Among the sets of variants showing complete haplotypic association, we counted only a single representative site; however, in cases of incomplete or unknown linkage disequilibrium, we count each site individually. Some functional variants are unrepresented in the frequency data set because population samples are not available, although in each case the data are sufficient to indicate that the functional variants meet the 1% rare allele frequency threshold. Frequency data are derived from random populations or from the control populations of clinical studies; the homogeneity and geographic origins of the populations vary among the studies. As a representative statistic, we have used the midpoint of the range (over populations) of rare allele frequencies for each locus. For example, we have frequency data for *APOE* −491 from nine populations. The rare allele, −491T, has a minimum frequency of 0.113 (among Finns) and a maximum frequency of 0.305 (among African-Americans). The midpoint of this range is 0.209. For each population, we estimated the expected heterozygosity as one minus the sum of the squared allele frequencies. We then calculated the midpoint of the range of heterozygosities as a representative statistic for each variant. Note that the use of the minimum heterozygosities and frequencies for each variant (including zeroes for the 10 private polymorphisms) has little effect on our overall conclusions.

Results and Discussion

Abundant Functional *cis*-Regulatory Variation

Experimentally verified functional *cis*-regulatory polymorphisms influence the rate of transcription from 107 genes spread over 20 autosomes and the X chromosome (fig. 1). This number represents 1% of all officially named human genes (Locuslink, December 19, 2001). Because these named genes include all genes that have been subject to even minimal study, 1% represents a firm, minimum estimate of the fraction of genes with functional *cis*-regulatory polymorphism. Genes from a range of functional categories are represented, including metabolic and regulatory enzymes, intercellular transporters, transcription factors, signal transducing receptors and their ligands, proteinase inhibitors, extracellular matrix proteins, components of the immune system, and cell adhesion molecules.

The functional variants have large effects on transcription (fig. 2*A*). In reporter construct assay experiments, 63% of the 107 genes have allelic differences of twofold or greater in their rates of transcription, and 20-fold differences are not uncommon.

To corroborate the experimental evidence for functionality of these variants in cell culture conditions, we

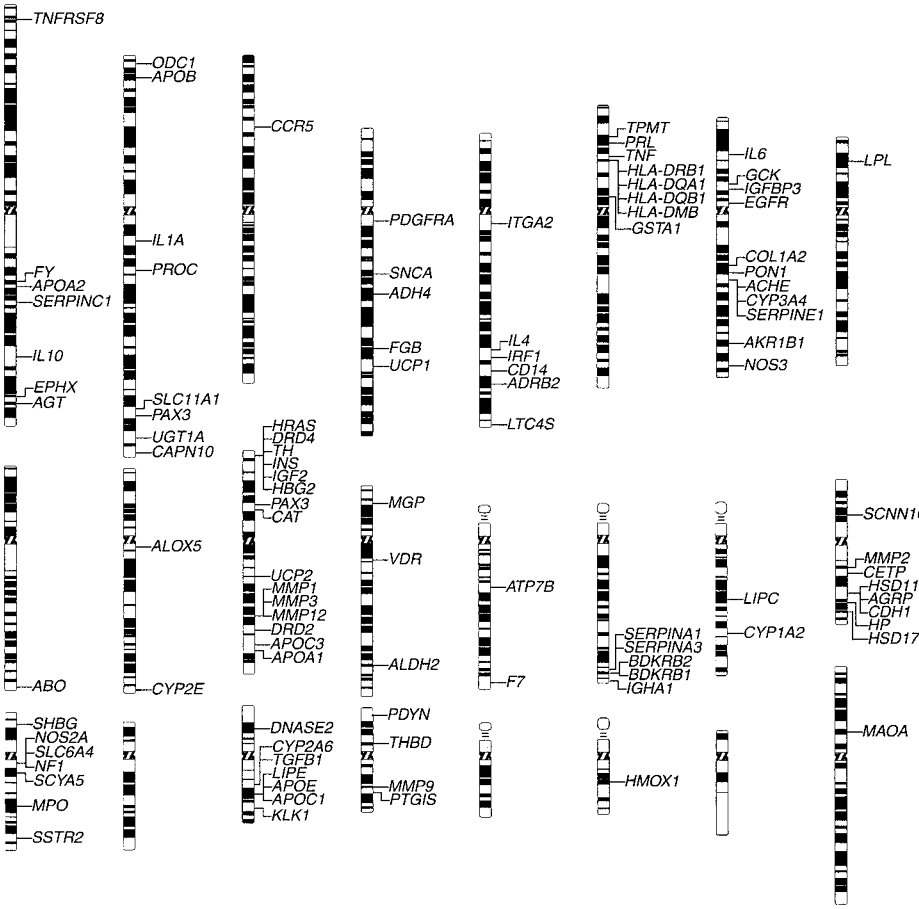

FIG. 1.—Chromosomal locations of human genes with experimentally validated functional *cis*-regulatory polymorphisms. The idiogram is modified from David Adler's Idiogram Album (http://www.pathology.washington.edu/cytopages/idiograms/human/), and the genes are assigned to positions following the annotation in the human genome working draft (http://genome.ucsc.edu/). Gene symbols are according to the Human Genome Organization Gene Nomeclature Committee (http://www.gene.ucl.ac.uk/nomenclature/).

sought support from studies documenting associations in vivo between the polymorphisms in their native chromosomal context and proximal biochemical phenotypes. We found such published support for 59% of the 107 genes. We next sought corroborative evidence from associations with predicted higher level phenotypes, including morphometric traits and disease susceptibilities. We obtained such evidence for 71% of the genes. At least one of the two lines of corroboration was found for 82% of the genes, and both lines were found for 47%. In most cases where corroboration is absent, the relevant studies have not yet been undertaken or published.

Characteristics of Functional Variants

Figure 2*B* shows the spectrum of mutation types among 144 functional variants (see *Materials and Methods* for an explanation of the sample size, and see *Supplementary Material* for a detailed list). The spectrum is largely typical of segregating human genetic variation and of substitutions fixed in the human lineage, including the ratio of SNPs to biallelic indels, and the pattern of base pair exchanges among the SNPs (fig. 2*C*) (Taillon-Miller et al. 1998; Cambien et al. 1999; Nachman and Crowell 2000; Taillon-Miller and Kwok 2000; Venter et al. 2001; Yu et al. 2001). Ordinary small-scale mutations are thus major contributors to heritable variation in transcription (Stone and Wray 2001).

The spectrum of mutation types is atypical, however, in its proportion of multiallelic variants (those with more than two alleles, such as polymorphic microsatellites). At 20% (29/144), the proportion of multiallelic variants among functional *cis*-regulatory variants is higher than the proportion found in surveys of random human variation, typically around 5%, though precise numbers are scarce (Taillon-Miller et al. 1998; Cambien et al. 1999). Nineteen of the functional variants (13%) involve VNTRs with repeat unit length less than 10 nucleotides; 12 involve dinucleotide repeats. Among these dinucleotide microsatellites, $(AC)_n$-type repeats are overrepresented (two-tailed binomial $P = 0.006$); these constitute 50% of the dinucleotide microsatellites genome wide (International Human Genome Sequencing Consortium 2001) but 92% (11/12) in our data set.

The first 500 bp upstream of the transcription start sites contain 58.9% (83/141; see *Materials and Methods*) of the variants, but a substantial fraction are found further afield; 12.8% are more than 1 kb upstream, and

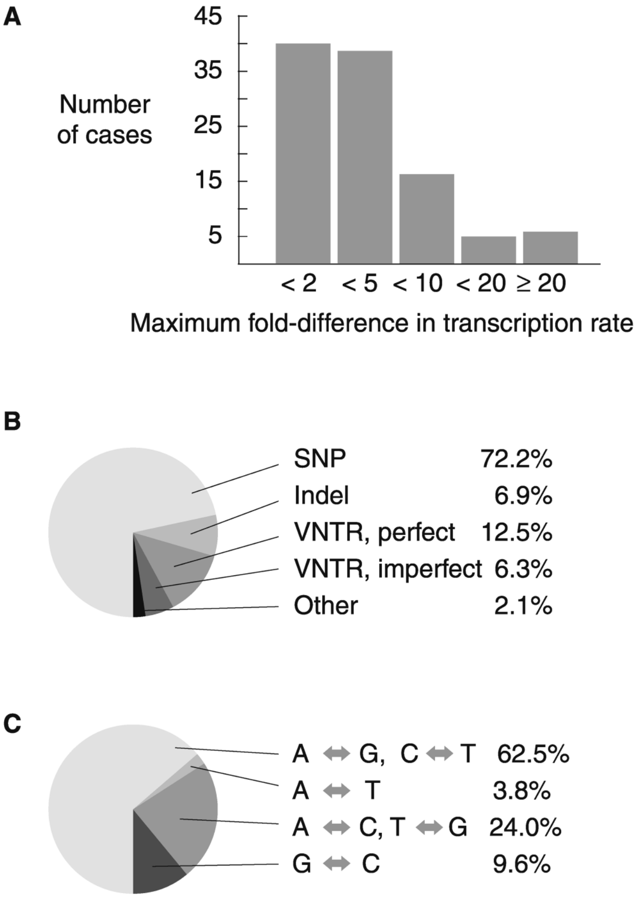

FIG. 2.—Functional polymorphisms in human *cis*-regulatory DNA. *A,* Maximum experimentally determined fold-difference in transcription rate between alleles of each gene. *B,* Distribution of mutation types. VNTR includes only multiallelic tandem repeat polymorphisms. ''Other'' includes one multiallelic hypervariable region (*KLK1*), one multiallelic complex of repeats, snps, and inversions (*IGHA1*), and one biallelic locus at which the alleles differ by multiple overlapping changes of length and sequence (*SERPINC1*). *C,* Distribution of SNP types without regard to orientation (A-G transitions on one strand are C-T transitions on the other strand; see *Materials and Methods*).

another 12.8% fall 3′ to the start of transcription. Two variants (1.4%) occur more than 10 kb upstream of their start sites. Given the ascertainment bias favoring discovery of variants in the immediate 5′ flanking sequence, these data indicate the broad spatial distribution of variation influencing transcription, as expected, given the broad distribution of functional regulatory elements (e.g., Carey and Smale 2000, pp. 61–62).

Polymorphic DNA–Protein Interactions

Allelic variation in transcription rate may be due to allelic variation in affinity for transcription factor binding. We sought published biochemical evidence for such variation for the functional variants in our data set.

For 51 variants, we inferred whether the allele with experimentally determined higher affinity for transcription factor binding is associated with the activation of transcription or its repression. We found 26 instances of polymorphic activator binding and 25 instances of polymorphic repressor binding. An additional nine loci showed transcription factor switching, with each allele having higher affinity for a different transcription factor.

Our inference of activation or repression is based on the evidence from reporter construct assays in cell lines and ignores the context-dependence of *cis*-regulatory function. For example, the influence of a polymorphic Sp1-Sp3 binding site in the *CD14* promoter depends on the ratio of Sp1 to Sp3 in the nucleus; because this ratio varies among cell types, protein binding to the site may result in either repression or activation (Levan et al. 2001). Nevertheless, our inference may be justified by noting that experimental studies typically focus on the cell types in which expression of the gene is thought to be of the greatest physiological consequence, and in most cases only a single effect, activation or repression, is documented for each polymorphic interaction.

We inferred ancestral states for 20 variants for which both experimental protein binding data and nonhuman primate sequences are available. Of these, the derived state is a gain of transcription factor binding in seven cases and a loss in 10 cases. The other three cases involve transcription factor switching. Two of the seven gains are human-specific expansions of tandem repeat sequences (*IGF2/INS* and *TH*). Two functional variants from the *HLA-DQB1* locus in the MHC complex are transspecific polymorphisms, i.e., both alleles segregate in both humans and chimpanzees. Considering more distant out-groups, we record these changes as one gain and one loss of transcription factor binding in the human-chimpanzee common ancestor.

Transcription factors from a diversity of structural families are involved in polymorphic binding, and many, including USF, Sp1, NFκB, GATA, and OCT, are involved in polymorphic binding to the *cis*-regulatory regions of more than one gene. Sp1 alone is involved in seven experimentally confirmed polymorphic functional DNA-protein interactions in our data set.

Five categories of regulatory change are represented in our small sample of human polymorphisms: gains and losses of repressor binding and of activator binding, and transcription factor switching. If we think of gene regulation as a network of linkages between transcription factors and *cis*-regulatory DNA, these data illustrate the extent to which the structure of the network is variable even within populations.

Population Genetics

The effects of *cis*-regulatory variation in human populations depend both on the number of segregating polymorphisms and on the heterozygosities of these polymorphisms. We collected frequency data from the literature for 129 functional variants from our 107 genes (table 1 and fig. 3*A*; see *Materials and Methods*). The distribution of rare allele frequencies is fairly uniform (fig. 3*B*) and thus deviates substantially, with an excess of intermediate frequency variants, from the neutral expectation (Hartl and Clark 1997, pp. 294–304) and from observed frequencies for both coding and other noncoding SNPs, which are skewed toward low-frequency alleles in humans (Cargill et al. 1999; Halushka et al.

Table 1
Mean Heterozygosities and Rare Allele Frequencies for Functional *Cis* -Regulatory Polymorphisms

Polymorphism Type	Number of Polymorphisms	Mean Sample Size Per Polymorphism	Mean Heterozygosity (SD)			Mean Rare Allele Frequency (SD)		
			Minima	Midpoints	Maxima	Minima	Midpoints	Maxima
All.	129	1868	0.3211 (0.2191)	0.3731 (0.1913)	0.4252 (0.1861)			
Biallelic	106	2092	0.2568 (0.1723)	0.3133 (0.1409)	0.3698 (0.1424)	0.1823 (0.1444)	0.2509 (0.1417)	0.3196 (0.1954)
Multiallelic.	23	839	0.6171 (0.1613)	0.6489 (0.1461)	0.6808 (0.1466)			
$(AC)_n$-type	9	787	0.6647 (0.1633)	0.6952 (0.1496)	0.7257 (0.1465)			
non–$(AC)_n$-type.	14	872	0.5865 (0.1582)	0.6192 (0.1411)	0.6519 (0.1445)			

NOTE.—For many loci, allele frequencies and heterozygosities were calculated from many populations. Minima, maxima, and midpoints indicate statistics calculated from the frequency or heterozygosity estimated from the population with the minimum or maximum for each locus or from the midpoint of the range of estimates for each locus. Frequency data were available for only 129 or the 144 functional polymorphisms (see *Materials and Methods*).

1999; Zwick, Cutler, and Chakravarti 2000; Stephens et al. 2001). The observed distribution is attributable, in part, to an ascertainment bias—clinical studies of common health factors may ignore rare variants and consequently experimental validation of functional effects will focus on intermediate-frequency variants. Nevertheless, the unusual shape of the distribution may suggest that natural selection contributes to *cis*-regulatory polymorphism in humans. Directional selection with geographically heterogeneous selection regimes accounts for variation in the *cis*-regulatory region of the *FY* locus (Hamblin and Di Rienzo 2000) and possibly the *CCR5* locus as well (Schliekelman, Garner, and Slatkin 2001). Overdominant balancing selection may maintain variation in the *cis*-regulation of MHC genes (Guardiola et al. 1996), whereas variation at *UGT1A* (Beutler, Gelbart, and Demina 1998) and *TNF* (Wilson et al. 1997) may be maintained by balancing selection because of antagonistic pleiotropy. Natural selection operating according to Neel's (1962) ''thrifty genotype'' hypothesis has been invoked to explain variation at the *AGT* (Inoue et al. 1997), *CAPN10* (Baier et al. 2000), and *INS* (McCarthy 1998) loci. Although these models are at present little more than verbal scenarios (with the exception of the *FY* and *CCR5* variants), distinctive properties of *cis*-regulatory variants discussed below, such as tissue-type overdominance, genotype-by-environment interactions, and interactions between epistasis and linkage, may make *cis*-regulatory polymorphisms especially prone to maintenance by natural selection.

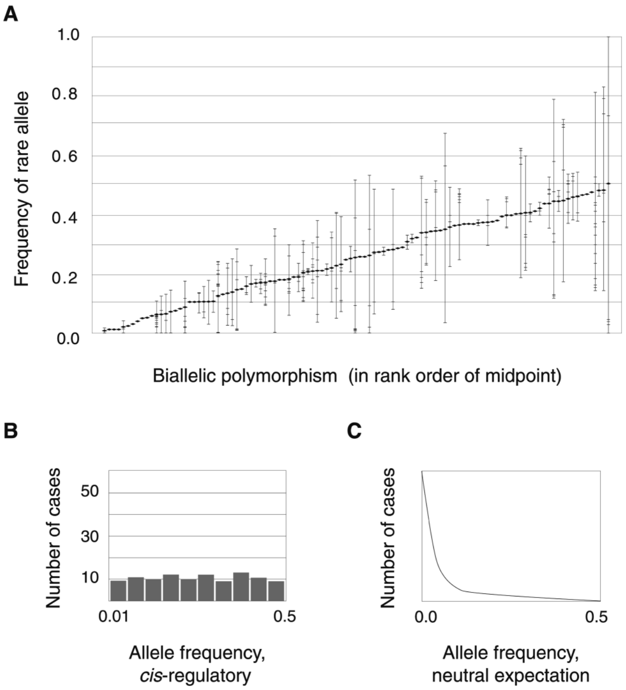

FIG. 3.—Rare allele frequencies for biallelic functional *cis*-regulatory polymorphisms in humans. *A*, Loci are ordered along the *x*-axis by the midpoints of the observed frequency range, as indicated by the diamond symbols. Each cross hatch represents a frequency estimate for a single study population, typically derived from a sample of hundreds of chromosomes. Complete sample details are available as supplementary material. *B*, The histogram of midpoint frequencies differs in shape from *C*, that expected under neutrality.

We have inferred ancestral states for 21 biallelic variants for which we have frequency data. Derived states have higher frequencies than do ancestral states at seven of these loci, at least in some populations. There are 11 additional variants for which ancestral states are unknown but that have frequencies on either side of 0.5 in different populations. These data illustrate the fact that functional *cis*-regulatory polymorphisms are contributors not only to transient human variation but likely also to divergence of humans from our ancestors.

Population Genomics

Our data permit some preliminary and cautious estimates of the extent of *cis*-regulatory variation in the genome. Because of the way variants are sampled in our data set and because of the inherent context-dependence of the *cis*-regulatory function, we cannot address the usual molecular population genomics parameters such as π, the per-site heterozygosity at functional *cis*-regulatory sites, or *S*, the number of segregating functional

Table 2
Distribution of Functional Variation Within the Human Genome

Type of Variant	Expected Number of Heterozygous Sites	Percent of Genes Heterozygous
Cis-Regulatory		
Biallelic	10,700	30.0
SNPs	9,700	27.6
Indels	900	3.0
Other	100	0.3
Multiallelic	6,500	19.5
$(AC)_n$-type	2,600	8.3
non–$(AC)_n$-type	3,900	12.2
Total *cis*-regulatory	17,200	43.6
Protein Coding		
Amino acid variants	7,900–12,900	23.2–39.9

NOTE.—All numbers are based on 30,000 genes and can be rescaled to accommodate other estimates of total gene number. The percent of genes heterozygous is per individual assuming a Poisson distribution of variants across all genes.

cis-regulatory sites in a sample of specified size. We concentrate instead on the total number of segregating functional *cis*-regulatory polymorphisms (i.e., variants with a rare allele population frequency greater than 1%) genome wide and on the expected number of heterozygous sites genome wide, considering only segregating functional *cis*-regulatory polymorphisms.

A recent analysis of upstream sequences from 180 loci (Stephens et al. 2001) found that the total frequency of SNPs in the proximal 850 bp of the 5′ flanking sequence averages 5.73/kb (this number is basically a large sample estimate of *cis*-regulatory *S* because it includes rare variants with frequencies less than 1%). Genes in our data set average 0.94 functional SNPs per kilobase in the proximal 850 bp 5′ to the start of transcription. (Independently functional nucleotides have been both identified and mapped in relation to the start of transcription in 100 of 107 genes; we thus calculate on the basis of 80 SNPs in a sample of 85 kb, i.e., 100 genes times 850 bp per gene.) A crude comparison suggests that, if our genes are typical, more than 16.4% of proximal promoter SNPs influence transcription. Although not directly comparable, large sample estimates of the percentage of human coding sequence variants that alter an amino acid range from 44% to 56% (Cargill et al. 1999; Halushka et al. 1999; Stephens et al. 2001; Venter et al. 2001).

Our estimate of 0.94 functional SNPs per kilobase is affected by two ascertainment biases, one among genes and one among nucleotide variants. First, our genes may be unrepresentative; they were included specifically because they contain functional variants. We suspect that this bias is minor: our genes are drawn from a wide range of functional categories and chromosomes, and in the vast majority of the studies from which our data are drawn, genes were selected with no a priori expectation of functional *cis*-regulatory variation. In most cases, the polymorphisms were identified by medical researchers who first selected genes known from model system biology as candidate genes for medical conditions (for example, apolipoproteins as candidate genes for cardiovascular phenotypes). The genes selected were then searched for polymorphisms, some of which were experimentally tested for function. Moreover, genes with functional *cis*-regulatory variants are included in our data set, irrespective of whether the variants are SNPs or whether they occur in the proximal 850 bp. Nevertheless, bias among genes, in particular a publication bias favoring genes with functional polymorphisms, may result in an inflation of the density of functional SNPs in our data set. A second ascertainment bias, bias among nucleotide variants, affects the estimate in the opposite way, by undercounting functional variants. Specifically, our estimate assumes that every functional variant in the proximal 850 bp of each of the genes in our data set has been identified. In fact, in many of the genes, little or none of the proximal region has yet been studied. Only a fraction of the identified variants have been tested experimentally, and of those many have not been tested under relevant conditions. On balance, we expect that our estimate of the number of functional *cis*-regulatory SNPs will prove low. As shown below, our primary conclusions regarding the distribution of human *cis*-regulatory variation are largely insensitive to our assumptions.

Given 0.94 functional SNPs/kb in the proximal promoter of each gene, and then by assuming an average heterozygosity for biallelic variants derived from our analysis (0.313; see table 1) and an estimate of 30,000 genes in the genome, we can extrapolate that on average a human is heterozygous for 7,500 functional SNPs in the proximal 850 bp of the 5′ flanking sequence. If we consider the 9.9% of biallelic variants in our data set that are not SNPs and the 22.5% of biallelic variants that fall outside the proximal 850 bp, we arrive at an estimate of about 10,700 functional biallelic *cis*-regulatory variants heterozygous in a typical human (table 2).

The preceding extrapolations are based on the best-studied variants, proximal promoter SNPs. We can also use the data set without breaking it down by spatial location. We found a total of 115 biallelic variants in 101 genes for which we have the appropriate data. If we extrapolate genewise, 1.1386 variants/gene times 30,000 genes times an average heterozygosity of 0.313 again yields an estimate of 10,700 functional biallelic *cis*-regulatory variants heterozygous in a typical human.

Table 3
Distribution of Known Functional Polymorphic $(AC)_n$ Microsatellites

Gene	Position	Orientation	Number of Repeats
SNCA	−8896	AC	10–13
AKR1B1	−2036	AC	19–28
COL1A2	−1383	AC	11–27
PAX6	−965	AC	18–31
PAX3	−337	AC	13–30
SLC11A1	−274	GT	15–23 (imperfect)
HMOX	−117	GT	15–41
MMP9	−90	CA	14–27
COL1A2	+1420 (intron 1)	GT	11–19
EGFR	+1429 (intron 1)	AC	14–22
HSD11B2	+2571 (intron 1)	AC	14–24

NOTE.—Position is relative to the start of transcription. Orientation refers to the sequence of the repeat unit on the sense strand. The *SNCA* repeats are part of a larger compound repeat, $(TC)_{10–11}TT(TC)_{8–11}(TA)_{7–9}(CA)_{10–13}$.

A Functional Role for Microsatellites

Multiallelic variants comprise 20% (29/144) of the variants in our data set, and among these, $(AC)_n$ microsatellites are overrepresented. Because these microsatellites have usually been considered nonfunctional neutral markers, we discount investigator bias as an explanation. We suggest instead that $(AC)_n$ repeats are major contributors to transcriptional regulation and hence to phenotypic variation in humans. Evidence for their role in transcriptional regulation is accumulating. In reporter constructs, $(AC)_n$ repeat number variation alone influences rates of transcription (Hamada et al. 1984), and their influence in the context of specific genomic sequences has been demonstrated experimentally for each of the 11 examples in our data set (table 3). At least one unidentified human transcription factor binds specifically to $(AC)_n$ sequences (Epplen, Kyas, and Maueler 1996; Shimajiri et al. 1999), and $(AC)_n$ sequences may influence transcription rates by their effects on DNA conformation even in the absence of specific protein binding (Naylor and Clark 1990; Wolfl et al. 1996; Rothenburg et al. 2001).

We can extrapolate from our data to the number of functional, polymorphic $(AC)_n$ repeats in the genome. The prevalence of $(AC)_n$ repeat elements is approximately 27.7/Mb, for more than 88,000 in the 3.2 Gb human genome, on the basis of stringent criteria which do not count short or imperfect repeats (International Human Genome Sequencing Consortium 2001). A survey of 2,506 $(AC)_n$ repeat elements found 93% to be polymorphic (Weissenbach et al. 1992), or about 82,000 for the whole genome. Note that this estimate is based on a study of unpreselected $(AC)_n$ elements with 12 or more repeats, such that the proportion polymorphic is likely a good estimate of the proportion genome wide. The variants in our data set occur in either orientation with respect to the sense strand and are widely distributed with respect to the start of transcription (table 3). We infer the presence of 3,800 functional, polymorphic $(AC)_n$ microsatellites in human populations, given three conservative assumptions: first, that repeats within 2.5 kb of the start of transcription (a 5-kb window) are functional, as implied by experimental data (Hamada et al. 1984; Epplen, Kyas, and Maueler 1996; Shimajiri et al. 1999; table 3) and that others are not (conservative because some repeats further afield are known to be functional); second, that $(AC)_n$ elements are randomly distributed throughout the genome (conservative because their distribution may actually be biased toward the vicinity of transcription start sites, Schroth, Chou, and Ho 1992); and third, that there are 30,000 starts of transcription (conservative because many genes have multiple start sites). The average heterozygosity for the $(AC)_n$ repeats in our data set is 0.695, which is typical for polymorphic dinucleotide repeats (Bowcock et al. 1994; Dib et al. 1996). Thus, we infer that a typical human is heterozygous for 2,600 functional $(AC)_n$ variants.

A functional role for $(AC)_n$ microsatellites has important consequences. Their high heterozygosity and high ($5–12 \times 10^{-4}$) mutation rate (Weissenbach et al. 1992; Weber and Wong 1993) may account for the rapid generation of phenotypic variation and its maintenance through bottleneck episodes or in the face of stabilizing or directional selection (Kashi, King, and Soller 1997). In addition, microsatellite length may itself be under stabilizing selection for *cis*-regulatory function, explaining their unexpectedly high level of conservation among apes and low repeat-number variance among humans (Bowcock et al. 1994; Deka et al. 1994).

Only 37.9% of the multiallelic variants in our data set are $(AC)_n$. If our sample is representative of functional multiallelic variants, we may infer the presence of an additional 6,200 functional multiallelic polymorphisms segregating in human populations. In our data set, the non-$(AC)_n$ multiallelic variants have an average heterozygosity of 0.619, implying that each human is heterozygous at 3,900 non-$(AC)_n$ multiallelic functional *cis*-regulatory sites or 6,500 multiallelic sites in total (table 2).

An alternative approach to this extrapolation, abandoning the assumption of a ±2.5 kb functional domain, uses the observation that in 101 genes with appropriate data, we found 29 multiallelic variants; the average heterozygosity of functional multiallelic variants, based on the 23 for which we have frequency data, is 0.649. Extrapolating to 30,000 genes, we arrive at an estimate of 8,600 functional multiallelic variants, of which 5,600 will be heterozygous on average.

Genome-Wide Functional Variation: *cis*-Regulatory Versus Coding

From the separate extrapolations for bi- and multiallelic *cis*-regulatory variants, we approach a range of 16,300–17,200 functional variants that are heterozygous on average. If we extrapolate from the combined data, we can use the observations of 1.43 functional variants per gene (115 bi- and 29 multiallelic variants in 101 genes), an average heterozygosity of 0.373, and a total of 30,000 genes, to estimate 16,000 functional *cis*-regulatory sites heterozygous on average.

In contrast, a human is likely to be heterozygous at only about 7,900–12,900 amino acid positions, given estimates of heterozygosity at nonsynonymous sites (0.196/kb [Cargill et al. 1999] and 0.32/kb [Sunyaev et al. 2000]), an average coding sequence length of 1,340 bp per gene (International Human Genome Sequencing Consortium 2001), and 30,000 genes. Only if our data set is dramatically unrepresentative, such that one-fifth of all genes have no segregating *cis*-regulatory variation in any human population, would the average number of heterozygous *cis*-regulatory sites be less than 12,900. The greater number of heterozygous functional *cis*-regulatory than coding sites is consistent with the results of recent interspecific analyses, which suggest that the number of noncoding nucleotides conserved between human and mouse exceeds the number of coding nucleotides (Frazer et al. 2001; Shabalina et al. 2001). In short, the size of the functional noncoding genome may simply exceed the size of the coding genome.

Four sources of bias contribute to these estimates. First, we assume that our genes are representative, although they are included specifically because of the presence of functional polymorphisms. We doubt, however, that this constitutes a major source of bias in practice; a consideration of the size of each gene's *cis*-regulatory region and the extent of human nucleotide diversity suggests that nearly all genes will have at least one *cis*-regulatory site segregating as a polymorphism in some human population. We can circumvent this potential bias by noting that our extrapolations for multiallelic variants do not rely on the assumption that our genes are representative of all genes, only that $(AC)_n$ microsatellites are randomly distributed and functional when near transcription start sites. If we accept the extrapolation for multiallelic variants, we can back-calculate heterozygosity at biallelic sites. We found previously that 10,000 functional multiallelic sites (3,800 $[AC]_n$ and 6,200 others) are segregating variants in human populations, and we found that multiallelic variants make up 20.1% of functional variants in our data set. Thus, by assuming that the proportion of multiallelic functional variants in our data set is a good estimator of the proportion among all functional variants, which we have no reason to doubt, we arrive at an estimate of 40,000 functional biallelic variants segregating in human populations. Given the estimated heterozygosity for such variants (0.313), we estimate an average of 12,500 heterozygous biallelic functional sites, slightly higher than the 10,700 estimated by assuming that the genes in the data set are representative. Second, our extrapolations are based on the assumption that all variants within our genes are known. In fact, for most of these genes only small segments of the *cis*-regulatory sequence have been searched for polymorphisms, and few of the known polymorphisms have been experimentally tested for function. Even then, the context-dependence of *cis*-regulatory function dictates that some functional variants will be missed and we will have underestimated the number of functional variants per gene. Third, we assume that there are 30,000 genes (i.e., sets of cotranscribed coding exons) and 30,000 starts of transcription (equivalently, first exons or promoters). These specific numbers are unimportant because our estimates of heterozygosities can be rescaled to accommodate any other number. But the ratio of genes to promoters is important. Many genes have multiple first exons and hence *cis*-regulatory regions, which increase the number of possible sites for functional *cis*-regulatory variants without affecting the number of possible sites for nonsynonymous coding variants. For example, *UGT1A* has at least nine separate first exons, each with its own promoter and start of transcription, spread over hundreds of kilobases (Gong et al. 2001). By assuming a 1:1 ratio of genes to start of transcription, we underestimate the ratio of functional *cis*-regulatory sites to amino acid sites that are heterozygous on average, given in this study as greater than 16,000:12,900. A correction for this bias would involve multiplying the estimated number of heterozygous functional *cis*-regulatory sites by the average number of promoters per gene, a number that is unfortunately unknown. Finally, we consider only polymorphisms, variants with rare allele frequencies greater than 1%, whereas surveys of amino acid polymorphism consider all variants, including rare singleton variants, which usually comprise about a third of polymorphisms identified in most surveys (e.g., Stephens et al. 2001). This difference again contributes to an underestimate of the ratio of *cis*-regulatory to amino acid sites that are heterozygous.

In sum, a bias among genes, which can be circumvented, contributes to an overestimate of the ratio, whereas biases among variants and frequencies, as well as our assumed gene:promoter ratio, contribute to an underestimate. The magnitudes of these biases are impossible to estimate, but on the whole, the data suggest that humans are on average heterozygous for more functional *cis*-regulatory sites than for amino acid sites. When additional functional noncoding variants are considered, such as those influencing mRNA stability and transport, splicing, and translation, the average number of heterozygous functional noncoding sites surely exceeds the number of heterozygous amino acid sites.

The Absence of Transposons

Although transposition is often considered an engine of regulatory evolution (Britten 1997; Brosius 1999; Carroll, Grenier, and Weatherbee 2001, p. 180), we found no example in humans of a segregating polymorphism of this type that has been experimentally implicated in variation in transcription. The disparity between the well-documented role for transposition in interspecific regulatory evolution (Britten 1997; Brosius 1999) and the absence of transposition mutations in our data set may be due to a dramatic slowdown in transposition rates in humans (International Human Genome Sequencing Consortium 2001). According to this model, transposition, although once important, no longer plays a major role in regulatory evolution in our species. A second possibility is that transposons typically have dramatic fitness consequences, such that they are either quickly fixed in populations or lost but rarely segregate as polymorphisms. Although rare transposition events with severe clinical phenotypes are documented in humans (Deininger and Batzer 1999), supporting the latter scenario, neither explanation is complete: segregating LINE-1 and Alu polymorphisms are well documented and likely number more than 2,000 (Sheen et al. 2000). Their absence from our data set may simply indicate that researchers have not yet focused on their functional consequences or that the numbers of transcriptionally relevant polymorphisms due to transposition are so small relative to other classes of polymorphisms that we would not expect to find them in our sample. Proportionally, polymorphisms due to transposition are clearly minor contributors to transcriptional variation in humans.

Distinctive Genotype-Phenotype Dynamics of *Cis*-Regulatory Variants

Our survey of known functional *cis*-regulatory variants has revealed a number of distinctive characteristics of the mode of action of these variants, with implications for their evolution and their role in phenotypic variation. Although differences among the studies in our survey prevent us from drawing strict quantitative conclusions about these characteristics, we believe that by outlining them and providing some empirical examples we can help focus future research on *cis*-regulatory variation.

First, while transcription factors typically regulate many downstream target genes, the identities of the target genes will differ among individuals on the basis of the variation in *cis*-regulatory DNA sequences. Consequently, changes in expression or structure of a transcription factor will not only be highly pleiotropic, influencing many downstream loci, but also highly epistatic, that is, dependent on genetic background. That this sort of *cis-trans* epistasis is common is suggested by the occurrence of multiple polymorphic binding sites for individual transcription factors, especially Sp1, and by the occurrence in our data set of four transcription factors whose *cis*-regulation is polymorphic (*IRF1, PAX3, PAX6,* and *VDR*). Moreover, *cis-trans* epistasis is corroborated by experimental studies of genes in our data set. For example, Rutter et al. (1998), using 4 kb reporter constructs of the *MMP1* gene, differing only by a single nucleotide indel, found repeatable differences in the effect on transcription when the constructs were transfected into fibroblast cells from four different donors. At the level of organismal phenotype as well, background effects are well documented for *cis*-regulatory variants; for example, national origin interacts with variants at the *SCYA5* and *CCR5* loci to influence disease progression (Gonzalez et al. 1999, 2001).

Second, epistasis in *cis,* whereby the effect on transcription of one *cis*-regulatory variant depends on covariation at a linked *cis*-regulatory site, is common in our data set, including, for example, among SNPs at the *LIPC* locus (Botma, Verhoeven, and Jansen 2001), among SNPs and a VNTR at *IL6* (Terry, Loukaci, and Green 2000), and among VNTRs at *COL1A2* (Akai, Kimura, and Hata 1999). A second form of *cis*-epistasis, in which organismal phenotype depends on linkage phase between *cis*-regulatory variants and amino acid polymorphisms, is also common. For example, this haplotype-dependent dosage-by-structure interaction is documented for *PON1* (James et al. 2000) and *APOE* (Lambert et al. 1998). The ubiquity of these two classes of epistasis in *cis* lends credence to the classical idea of the coadapted gene complex (Dobzhansky 1951, p. 278) and the Lewontinian notion of the genome as the unit of selection (Lewontin 1974, pp. 273–318). It may also point to the viability of epistatic selection as an explanation for irregular patterns of linkage disequilibrium in humans (Stephens et al. 2001). At minimum, the frequent occurrence of linked, interacting functional sites mandates that models for the evolution of the human gene sequences incorporate the interaction of linkage and epistasis.

Third, we note that many of the genes with regulatory polymorphisms interact with one another in regulatory, metabolic, and physiological networks. As a consequence, moderate levels of variation at any one locus will be amplified by the number of genes in the network to yield high levels of polymorphism at the network level. For example, the genes *APOA1, APOA2, APOB, APOC1, APOC3, IGFBP3 IGF2, IL1, IL4, IL6, IL10, INS, FGB, LPL, LIPC, LIPE, TNF,* and *TGFB* all have functional *cis*-regulatory polymorphisms and interact with one another directly or indirectly. Analytical models and simulations have shown that polymorphism distributed through a network of interacting genes may result in a nonlinear genotype-phenotype relationship (that is, one characterized by dominance and epistasis) as an emergent property of lower level additive gene action (Nijhout and Paulsen 1997; Gilchrist and Nijhout 2001), and such multilocus phenomena are now being reported in the literature on human *cis*-regulatory variation (Jansen et al. 2001).

Fourth, a large proportion of the variants in our data set exhibit differential responses to exogenous inducers. In some cases, this means that alleles differ in the magnitude of inductive response—for example, phorbol ester stimulation interacts with an SNP at the *MGP* locus (Farzaneh-Far et al. 2001) and with a VNTR at the *PDYN* locus (Zimprich et al. 2000)—but in other cases each allele shows a different direction of response, as at the *SLC11A1* locus, at which the combination of interferon-γ and bacterial lipopolysaccharide leads to the upregulation of one allele relative to the basal transcription rate but leads to repression of transcription from another allele (Searle and Blackwell 1999). Because environmental cues are often mediated through transcriptional regulation and because inducibility is a pervasive property of human genes (Iyer et al. 1999), *cis*-regulatory polymorphism likely constitutes a major genetic basis for genotype-by-environment interaction effects. At the organismal phenotypic level, such effects are well documented, as for example at the *CETP* locus, where a *cis*-regulatory SNP and alcohol consumption interact in the modulation of HDL-cholesterol levels (Corbex et al. 2000).

Fifth, although *cis*-regulatory sequences are often assumed to influence discrete aspects of transcription of a single gene and so to be the basis for developmental modularity (Stern 2000; Carroll, Grenier, and Weatherbee 2001, pp. 91–92), our data set includes two instances of the opposite phenomenon: single variants influencing transcription from multiple genes. The imperfect VNTR upstream of the *INS* gene influences transcription from both *INS* and the downstream *IGF2* locus (Kennedy, German, and Rutter 1995; Paquette et al. 1998), and SNPs upstream of *APOC3* influence transcription of that gene in the liver (Li et al. 1995) and transcription of the *APOA1* gene, which is transcribed convergently to *APOC3,* in the colon (Naganawa et al. 1997). The prevalence of this sort of multifunctional noncoding DNA variation and its attendant pleiotropic consequences have likely been underestimated, in part because most experimental studies of regulatory variants focus only on the effects on transcription of the nearest gene.

Sixth, we found many loci at which alleles produce spatially or temporally nonnested gene expression patterns such that heterozygotes have patterns of expression beyond the range of either homozygote. Although such alleles may act additively at the level of number of transcripts produced, they exhibit overdominance at the level of the number of nuclei or tissues with a given rate of transcription. This single-locus tissue-type overdominance may represent a selective mechanism for the maintenance of *cis*-regulatory variation (Guardiola et al. 1996). For example, one allele at the *HLA-DQB1* locus shows higher expression than another in primary skin cells, whereas in peripheral blood mononuclear cells and B-lymph cells the situation is reversed (Beaty, Sukiennicki, and Nepom 1999). Such behavior is not limited to the MHC complex; similar dynamics are evident at such loci as *AGT* (Zhao et al. 1999), *INS* (Pugliese et al. 1997), and *NOS2A* (Morris et al. 2001).

Finally, because *cis*-regulatory sequences respond to spatially and temporally regulated transcription factors, *cis*-regulatory polymorphism represents an important genetic basis for the evolution of development by way of heterotopy and heterochrony. Our data set includes several examples of polymorphic temporal and spatial regulation. For example, a polymorphism at the *FY* locus affects expression in erythroid cells but not in other tissues (Tournamille et al. 1995). *Cis*-regulatory variants at the *HBG2* locus alter the time course of expression, resulting in variable fetal hemoglobin expression among adults (Labie et al. 1985). Developmentally important variants are likely vastly underrepresented in our data set because of the impossibility of carrying out the necessary experiments in developing human embryos.

Conclusions

The human genome is brimming with functional variation that influences transcription and thus many aspects of phenotype. The *cis*-regulatory polymorphisms analyzed in this study, scattered across a diversity of interacting genes, point to the surprising extent of regulatory variation in humans and underscore the complexity of the genotype-phenotype relationship. *Cis*-regulatory polymorphism represents the intersection of the central themes of development and evolution: the role of differential gene expression in development (Davidson 2001) and the evolutionary origin of species differences as variation within populations (Haldane 1932; Dobzhansky 1951). Our species is depauperate in sequence variation compared with many others (Zwick, Cutler, and Chakravarti 2000); we expect that comparable or greater levels of functional variation exist within the *cis*-regulatory regions of other eukaryotes.

Supplementary Material

Tables enumerating the genes, variants, and frequencies used in our study, along with references for source studies, are available at the Molecular Biology and Evolution website http//www.smbe.org.

Acknowledgments

We thank the National Science Foundation and NASA for support and E. Davidson, G. Gibson, M. Hahn, M. Rausher, J. Stone, K. Zigler, R. Zufall, and members of the Wray lab for helpful suggestions.

LITERATURE CITED

Akai, J., A. Kimura, and R. I. Hata. 1999. Transcriptional regulation of the human type I collagen α2 (*COL1A2*) gene by the combination of two dinucleotide repeats. Gene **239**: 65–73.

Baier, L. J., P. A. Permana, X. Yang et al. (13 co-authors). 2000. A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. J. Clin. Investig. **106**:R69–R73.

Beaty, J. S., T. L. Sukiennicki, and G. T. Nepom. 1999. Allelic variation in transcription modulates MHC class II expression and function. Microbes Infect. **1**:919–927.

Beutler, E., T. Gelbart, and A. Demina. 1998. Racial variability in the UDP-glucuronosyltransferase 1 (*UGT1A1*) promoter: a balanced polymorphism for regulation of bilirubin metabolism? Proc. Natl. Acad. Sci. USA **95**:8170–8174.

Botma, G. J., A. J. Verhoeven, and H. Jansen. 2001. Hepatic lipase promoter activity is reduced by the C-480T and G-216A substitutions present in the common *LIPC* gene variant, and is increased by Upstream Stimulatory Factor. Atherosclerosis **154**:625–632.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**:455–457.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science **296**:752–755.

Britten, R. J. 1997. Mobile elements inserted in the distant past have taken on important functions. Gene **205**:177–182.

Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. Science **165**:349–357.

Brosius, J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica **107**: 209–238.

Cambien, F., O. Poirier, V. Nicaud et al. (16 co-authors). 1999. Sequence diversity in 36 candidate genes for cardiovascular disorders. Am. J. Hum. Genet. **65**:183–191.

Carey, M., and S. T. Smale. 2000. Transcriptional regulation in eukaryotes. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Cargill, M., D. Altshuler, J. Ireland et al. (18 co-authors). 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22**:231–238.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2001. From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science, London.

Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. Proc. Natl. Acad. Sci. USA **97**:12369–12374.

Corbex, M., O. Poirier, F. Fumeron, D. Betoulle, A. Evans, J. B. Ruidavets, D. Arveiler, G. Luc, L. Tiret, and F. Cambien. 2000. Extensive association analysis between the *CETP* gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction. Genet. Epidemiol. **19**: 64–80.

Crawford, D. L., J. A. Segal, and J. L. Barnett. 1999. Evolutionary analysis of TATA-less proximal promoter function. Mol. Biol. Evol. **16**:194–207.

Damerval, C., A. Maurice, J. M. Josse, and D. de Vienne. 1994. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics **137**:289–301.

Davidson, E. H. 2001. Genomic regulatory systems: development and evolution. Academic Press, San Diego.

Deininger, P. L., and M. A. Batzer. 1999. Alu repeats and human disease. Mol. Genet. Metab. **67**:183–193.

Deka, R., M. D. Shriver, L. M. Yu, L. Jin, C. E. Aston, R. Chakraborty, and R. E. Ferrell. 1994. Conservation of human chromosome 13 polymorphic microsatellite $(CA)_n$ repeats in chimpanzees. Genomics **22**:226–230.

Dib, C., S. Faure, C. Fizames et al. (14 co-authors). 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152–154.

Dobzhansky, T. 1951. Genetics and the origin of species. Columbia University Press, New York.

Enard, W., P. Khaitovich, J. Klose et al. (13 co-authors). 2002. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340–343.

Epplen, J. T., A. Kyas, and W. Maueler. 1996. Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins. FEBS Lett. **389**:92–95.

Farzaneh-Far, A., J. D. Davies, L. A. Braam, H. M. Spronk, D. Proudfoot, S. W. Chan, K. M. O'Shaughnessy, P. L. Weissberg, C. Vermeer, and C. M. Shanahan. 2001. A polymorphism of the human matrix gamma-carboxyglutamic acid protein promoter alters binding of an activating protein-1 complex and is associated with altered transcription and serum levels. J. Biol. Chem. **276**:32466–32473.

Frazer, K. A., J. B. Sheehan, R. P. Stokowski, X. Chen, R. Hosseini, J. F. Cheng, S. P. Fodor, D. R. Cox, and N. Patil. 2001. Evolutionarily conserved sequences on human chromosome 21. Genome Res. **11**:1651–1659.

Gilchrist, M. A., and H. F. Nijhout. 2001. Nonlinear developmental processes as sources of dominance. Genetics **159**: 423–432.

Gong, Q. H., J. W. Cho, T. Huang et al. (11 co-authors). 2001. Thirteen UDPglucuronosyltransferase genes are encoded at the human *UGT1* gene complex locus. Pharmacogenetics **11**:357–368.

Gonzalez, E., M. Bamshad, N. Sato et al. (22 co-authors). 1999. Race-specific HIV-1 disease-modifying effects associated with *CCR5* haplotypes. Proc. Natl. Acad. Sci. USA **96**:12004–12009.

Gonzalez, E., R. Dhanda, M. Bamshad et al. (16 co-authors). 2001. Global survey of genetic variation in *CCR5*, *RANTES*, and *MIP-1*α*:* impact on the epidemiology of the HIV-1 pandemic. Proc. Natl. Acad. Sci. USA **98**:5199–5204.

Guardiola, J., A. Maffei, R. Lauster, N. A. Mitchison, R. S. Accolla, and S. Sartoris. 1996. Functional significance of polymorphism among MHC class II gene promoters. Tissue Antigens **48**:615–625.

Haldane, J. B. S. 1932. The causes of evolution. Longmans, Green and Co., London.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22**:239–247.

Hamada, H., M. Seidman, B. H. Howard, and C. M. Gorman. 1984. Enhanced gene expression by the poly(dT-dG)●poly(dC-dA) sequence. Mol. Cell. Biol. **4**:2622–2630.

Hamblin, M. T., and A. Di Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66**:1669–1679.

Hartl, D. L., and A. G. Clark. 1997. Principles of population genetics. Sinauer, Sunderland, Mass.

Inoue, I., T. Nakajima, C. S. Williams et al. (12 co-authors). 1997. A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. J. Clin. Investig. **99**: 1786–1797.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Iyer, V. R., M. B. Eisen, D. T. Ross et al. (14 co-authors). 1999. The transcriptional program in the response of human fibroblasts to serum. Science **283**:83–87.

James, R. W., I. Leviev, J. Ruiz, P. Passa, P. Froguel, and M. C. Garin. 2000. Promoter polymorphism T(–107)C of the paraoxonase *PON1* gene is a risk factor for coronary heart disease in type 2 diabetic patients. Diabetes **49**:1390–1393.

Jansen, H., D. M. Waterworth, V. Nicaud, C. Ehnholm, and P. J. Talmud. 2001. Interaction of the common apolipoprotein C-III (*APOC3*−482C > T) and hepatic lipase (*LIPC*−514C > T) promoter variants affects glucose tolerance in young adults. Ann. Hum. Genet. **65**:237–243.

Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29**:389–395.

Kashi, Y., D. King, and M. Soller. 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. **13**:74–78.

Kennedy, G. C., M. S. German, and W. J. Rutter. 1995. The minisatellite in the diabetes susceptibility locus *IDDM2* regulates insulin transcription. Nat. Genet. **9**:293–298.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.

Labie, D., J. Pagnier, C. Lapoumeroulie, F. Rouabhi, O. Dunda-Belkhodja, P. Chardin, C. Beldjord, H. Wajcman, M. E. Fabry, and R. L. Nagel. 1985. Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. Proc. Natl. Acad. Sci. USA **82**:2111–2114.

Lambert, J. C., F. Pasquier, D. Cottel, B. Frigard, P. Amouyel, and M. C. Chartier-Harlin. 1998. A new polymorphism in the *APOE* promoter associated with risk of developing Alzheimer's disease. Hum. Mol. Genet. **7**: 533–540.

LeVan, T. D., J. W. Bloom, T. J. Bailey, C. L. Karp, M. Halonen, F. D. Martinez, and D. Vercelli. 2001. A common single nucleotide polymorphism in the *CD14* promoter decreases the affinity of Sp protein binding and enhances transcriptional activity. J. Immunol. **167**:5838–5844.

Lewontin, R. C. 1972. The apportionment of human diversity. Evol. Biol. **6**:381–398.

———. 1974. The genetic basis of evolutionary change. Columbia University Press, New York.

Li, W. W., M. M. Dammerman, J. D. Smith, S. Metzger, J. L. Breslow, and T. Leff. 1995. Common genetic variation in the promoter of the human apo CIII gene abolishes regulation by insulin and may contribute to hypertriglyceridemia. J. Clin. Investig. **96**:2601–2605.

Li, W.-H., and L. A. Sadler. 1991. Low nucleotide diversity in man. Genetics **129**:513–523.

McCarthy, M. 1998. Weighing in on diabetes risk. Nat. Genet. **19**:209–210.

Morris, B. J., C. L. Glenn, D. E. Wilcken, and X. L. Wang. 2001. Influence of an inducible nitric oxide synthase promoter variant on clinical variables in patients with coronary artery disease. Clin. Sci. (Lond.) **100**:551–556.

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297–304.

Naganawa, S., H. N. Ginsberg, R. M. Glickman, and G. S. Ginsburg. 1997. Intestinal transcription and synthesis of apolipoprotein AI is regulated by five natural polymorphisms upstream of the apolipoprotein CIII gene. J. Clin. Investig. **99**:1958–1965.

Naylor, L. H., and E. M. Clark. 1990. $d(TG)_n \bullet d(CA)_n$ sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. Nucleic Acids Res. **18**:1595–1601.

Neel, J. V. 1962. Diabetes mellitus: a ''thrifty'' genotype rendered detrimental by ''progress''? Am. J. Hum. Genet. **14**: 353–362.

Nijhout, H. F., and S. M. Paulsen. 1997. Developmental models and polygenic characters. Am. Nat. **149**:394–405.

Paquette, J., N. Giannoukakis, C. Polychronakos, P. Vafiadis, and C. Deal. 1998. The *INS* 5′ variable number of tandem repeats is associated with *IGF2* expression in humans. J. Biol. Chem. **273**:14158–14164.

Pugliese, A., M. Zeller, A. Fernandez Jr., L. J. Zalcberg, R. J. Bartlett, C. Ricordi, M. Pietropaolo, G. S. Eisenbarth, S. T. Bennett, and D. D. Patel. 1997. The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the *INS* VNTR-*IDDM2* susceptibility locus for type 1 diabetes. Nat. Genet. **15**:293–297.

Rothenburg, S., F. Koch-Nolte, A. Rich, and F. Haag. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc. Natl. Acad. Sci. USA **98**:8985–8990.

Rutter, J. L., T. I. Mitchell, G. Buttice, J. Meyers, J. F. Gusella, L. J. Ozelius, and C. E. Brinckerhoff. 1998. A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter creates an Ets binding site and augments transcription. Cancer Res. **58**:5321–5325.

Schliekelman, P., C. Garner, and M. Slatkin. 2001. Natural selection and resistance to HIV. Nature **411**:545–546.

Schroth, G. P., P. J. Chou, and P. S. Ho. 1992. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. J. Biol. Chem. **267**:11846–11855.

Schulte, P. M., H. C. Glemet, A. A. Fiebig, and D. A. Powers. 2000. Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. Proc. Natl. Acad. Sci. USA **97**:6597–6602.

Searle, S., and J. M. Blackwell. 1999. Evidence for a functional repeat polymorphism in the promoter of the human *NRAMP1* gene that correlates with autoimmune versus infectious disease susceptibility. J. Med. Genet. **36**:295–299.

Shabalina, S. A., A. Y. Ogurtsov, V. A. Kondrashov, and A. S. Kondrashov. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. **17**: 373–376.

Sheen, F. M., S. T. Sherry, G. M. Risch, M. Robichaux, I. Nasidze, M. Stoneking, M. A. Batzer, and G. D. Swergold. 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. **10**:1496–1508.

Shimajiri, S., N. Arima, A. Tanimoto, Y. Murata, T. Hamada, K. Y. Wang, and Y. Sasaguri. 1999. Shortened microsatellite $d(CA)_{21}$ sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett. **455**: 70–74.

Stam, L. F., and C. C. Laurie. 1996. Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in *Drosophila melanogaster*. Genetics **144**:1559–1564.

Stephens, J. C., J. A. Schneider, D. A. Tanguay et al. (28 co-authors). 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science **293**:489–493.

Stern, D. L. 2000. Evolutionary developmental biology and the problem of variation. Evolution **54**:1079–1091.

Stone, J. R., and G. A. Wray. 2001. Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol. Biol. Evol. **18**:1764–1770.

Sucena, E., and D. L. Stern. 2000. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by *cis*-regulatory evolution of *ovo/shaven-baby*. Proc. Natl. Acad. Sci. USA **97**:4530–4534.

Sunyaev, S., J. Hanke, D. Brett, A. Aydin, I. Zastrow, W. Lathe, P. Bork, and J. Reich. 2000. Individual variation in protein-coding sequences of human genome. Adv. Protein Chem. **54**:409–437.

Taillon-Miller, P., Z. Gu, Q. Li, L. Hillier, and P. Y. Kwok. 1998. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res. **8**:748–754.

Taillon-Miller, P., and P. Y. Kwok. 2000. A high-density single-nucleotide polymorphism map of Xq25-q28. Genomics **65**:195–202.

Tautz, D. 2000. Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**:575–579.

Terry, C. F., V. Loukaci, and F. R. Green. 2000. Cooperative influence of genetic polymorphisms on interleukin 6 transcriptional regulation. J. Biol. Chem. **275**:18138–18144.

Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat. Genet. **10**:224–228.

VENTER, J. C., M. D. ADAMS, E. W. MYERS et al. (271 co-authors). 2001. The sequence of the human genome. Science **291**:1304–1351.

WEBER, J. L., and C. WONG. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

WEISSENBACH, J., G. GYAPAY, C. DIB, A. VIGNAL, J. MORISSETTE, P. MILLASSEAU, G. VAYSSEIX, and M. LATHROP. 1992. A second-generation linkage map of the human genome. Nature **359**:794–801.

WILSON, A. G., J. A. SYMONS, T. L. MCDOWELL, H. O. MCDEVITT, and G. W. DUFF. 1997. Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. Proc. Natl. Acad. Sci. USA **94**:3195–3199.

WOLFL, S., C. MARTINEZ, A. RICH, and J. A. MAJZOUB. 1996. Transcription of the human corticotropin-releasing hormone gene in NPLC cells is correlated with Z-DNA formation. Proc. Natl. Acad. Sci. USA **93**:3664–3668.

YU, N., Z. ZHAO, Y. X. FU et al. (11 co-authors). 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18**:214–222.

ZHAO, Y. Y., J. ZHOU, C. S. NARAYANAN, Y. CUI, and A. KUMAR. 1999. Role of C/A polymorphism at −20 on the expression of human angiotensinogen gene. Hypertension **33**:108–115.

ZIMPRICH, A., J. KRAUS, M. WOLTJE, P. MAYER, E. RAUCH, and V. HOLLT. 2000. An allelic variation in the human prodynorphin gene promoter alters stimulus-induced expression. J. Neurochem. **74**:472–477.

ZWICK, E. M., D. J. CUTLER, and A. CHAKRAVARTI. 2000. Patterns of genetic variation in mendelian and complex traits. Annu. Rev. Genomics Hum. Genet. **1**:387–407.

STEPHEN PALUMBI, reviewing editor

Accepted July 17, 2002