Personal and population genomics of human regulatory variation

Benjamin Vernot, Andrew B. Stergachis, Matthew T. Maurano, Jeff Vierstra, Shane Neph, Robert E. Thurman, John A. Stamatoyannopoulos,1 and Joshua M. Akey1

Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

The characteristics and evolutionary forces acting on regulatory variation in humans remains elusive because of the difficulty in defining functionally important noncoding DNA. Here, we combine genome-scale maps of regulatory DNA marked by DNase I hypersensitive sites (DHSs) from 138 cell and tissue types with whole-genome sequences of 53 geographically diverse individuals in order to better delimit the patterns of regulatory variation in humans. We estimate that individuals likely harbor many more functionally important variants in regulatory DNA compared with protein-coding regions, although they are likely to have, on average, smaller effect sizes. Moreover, we demonstrate that there is significant heterogeneity in the level of functional constraint in regulatory DNA among different cell types. We also find marked variability in functional constraint among transcription factor motifs in regulatory DNA, with sequence motifs for major developmental regulators, such as HOX proteins, exhibiting levels of constraint comparable to protein-coding regions. Finally, we perform a genome-wide scan of recent positive selection and identify hundreds of novel substrates of adaptive regulatory evolution that are enriched for biologically interesting pathways such as melanogenesis and adipocytokine signaling. These data and results provide new insights into patterns of regulatory variation in individuals and populations and demonstrate that a large proportion of functionally important variation lies beyond the exome.

[Supplemental material is available for this article.]
creation of genome-scale maps of diverse functional noncoding elements marked by DHSs. For example, in the ENCODE Project, ~3 million DHSs have now been mapped across 138 cell types (Thurman et al. 2012). In addition, genomic DNase I footprinting of 41 cell diverse cell types has resulted in the localization of 8.4 million DNase I footprints (Neph et al. 2012b).

Here, we describe a comprehensive analysis into patterns of genetic variation in regulatory DNA marked by DHSs and DNase I footprints (The ENCODE Project Consortium 2012). By analyzing whole-genome sequence data, we are able to directly compare characteristics of regulatory and protein-coding variation and find that individuals harbor considerably more regulatory compared to protein-coding variants. Moreover, we demonstrate that significant heterogeneity of functional constraint exists across regulatory DNA between cell types and that regulatory DNA present in multiple broad categories of cell types is significantly more constrained. Finally, we quantify patterns of population structure in regulatory DNA and identify several hundred loci that contain signatures of local adaptation. In summary, these analyses represent the most comprehensive assessment of human regulatory variation described to date and have important implications for personal genomics, disease mapping studies, and human evolution.

Results and Discussion

Overview of DNase I and whole-genome sequence data

A schematic illustration of the classes of data used in our analyses is shown in Figure 1A. Within DHSs, DNase I “peaks” correspond to ~150-bp regions of maximum hypersensitivity (Fig. 1A; see The ENCODE Project Consortium 2012). Embedded within peaks, are much smaller 6- to 20-bp DNase I footprints, which identify regions bound by regulatory factors (Fig. 1A). We also obtained publicly available whole-genome sequence data for 53 unrelated individuals that encompass five geographically diverse populations (Fig. 1B) from Complete Genomics. The average sequencing depth per individual was ~40×. Variants were filtered for deviations from Hardy-Weinberg equilibrium, and partial genotype calls were set to missing data (see Methods). The high-coverage whole-genome sequence data are ideal for population genetics analyses as they are free from the confounding effects of ascertainment bias present in genotypes obtained from SNP chips (Tennessen et al. 2011).

Pervasive regulatory variation across the human genome

Across all cell types, over 2.9 million DNase I peaks and 8.4 million DNase I footprints were identified across sampled cell types that collectively span 577 Mb and 156 Mb of sequence, respectively (18.7% and 5.1% of the genome for peaks and footprints, respectively). By use of the whole-genome sequence data, we observed 3.85 million, 1.01 million, and 0.15 million variants in DNase I peaks, DNase I footprints, and the exome, respectively. The large number of variants in peaks and footprints relative to exomes is a function of the total amount of sequence they encompass. For example, the number of variants per kilobase in peaks, footprints, and the exome is 6.7, 6.5, and 4.2, respectively.

To compare the number of putatively functional variants across peaks, footprints, and the exome we obtained GERP scores for each variant, which is a measure of evolutionary constraint with positive values indicating greater conservation (Cooper et al. 2005). Peaks and footprints not only have an overall larger number of variants relative to exomes but also manifest more high GERP variants compared with protein-coding regions (Fig. 2A), although the differences between categories becomes less dramatic. For example, at a threshold of GERP $\geq 3$ (Cooper et al. 2005) 146,570, 61,933, and 36,935 variants are observed in peaks, footprints, and the exome, respectively. It is interesting to note that protein-coding DNA contains proportionally more putatively functional variation compared with noncoding DNA (i.e., 24.6%, 6.1%, and 3.8% of variants have a GERP $\geq 3$ for exomes, footprints, and peaks, respectively), but the absolute number of functional variants in

Figure 1. Overview of data used in the analyses. (A) Schematic of the DNase I data. Binding of regulatory proteins to DNA (blue rectangle) results in nucleosome (open circles) displacement and local chromatin remodeling, and these regions are susceptible to cleavage with the endonuclease DNase I. High-throughput sequencing of libraries made from digested nuclei reveals DNase I hypersensitive sites, detectable by increased depth of coverage. Peaks are defined as 150-bp windows centered on the area of maximum cleavage (The ENCODE Project Consortium 2012). Within hypersensitive sites, footprints of regulatory factor binding are observed as decreased cleavage. (B) Unrooted neighbor-joining tree of the 53 unrelated individuals colored by population. Abbreviations are described in Supplemental Table 2.
noncoding regions is larger because of the greater amount of genomic sequence they encompass. Thus, regulatory variation is pervasive across the human genome, and a substantial proportion of functional variation exists in noncoding DNA.

Next, we investigated the distribution of putative regulatory and protein-coding variation across individuals. As expected, the average number of variants (±SD) per individual in peaks and footprints is dramatically higher than that found in the exome (741k ± 72k in peaks, 192k ± 18k in footprints, and 24.4k ± 2.2k in the exome) (Fig. 2B). A more interesting comparison, however, is the number of putatively functional regulatory and protein variants per individual. Therefore, we also determined the number of variants per individual with GERP ≥ 3 in peaks, footprints, and exomes. Therefore, we also determined the number of variants per individual with a GERP ≥ 3 in peaks, footprints, and the exome (Fig. 2C). On average, individuals contain 24.2k ± 2.3k, 10.1k ± 0.92k, and 4.7k ± 0.40k high GERP variants in peaks, footprints, and the exome, respectively (Fig. 2C). Although evolutionary constraint is not a perfect proxy for function, these results suggest that individuals possess more regulatory versus protein-coding variants. Assuming the probability that a variant is functional is the same between coding and noncoding DNA, we estimate that individuals contain up to five times as many regulatory compared with protein-coding variants. This assumption, however, is dubious (McVicker et al. 2009), and more definitive inferences on the proportion of functional variants in noncoding versus coding DNA will ultimately require further experimental data. In addition, it is interesting to note that, as expected, the average number of variants per individual in peaks and footprints is significantly higher for individuals of African ancestry compared to non-Africans (859k vs. 710k in Africans and non-Africans, respectively; \(P < 9.95 \times 10^{-16}\)) (Fig. 2B).

Patterns of nucleotide diversity in regulatory DNA sequence motifs

The unique scope of the data sets analyzed here allows us for the first time to systematically investigate genomic patterns of variation in DNA sequence motifs. To this end, we scanned DNase I footprints for 732 known motifs (see Methods), and for each motif, we calculated nucleotide diversity, \(\pi\), averaged across all instances of the motif in these regions. To facilitate interpretation of motif diversity, we also calculated \(\pi\) for fourfold synonymous sites, a proxy for neutrally evolving DNA, and protein-coding sequences. As shown in Figure 3A, average diversity varies by over sevenfold across known regulatory motifs, ranging from 2.67 \(\times 10^{-4}\) to 2.0 \(\times 10^{-3}\). Approximately 60% of motifs have average diversities significantly lower than fourfold synonymous sites (Fig. 3A), indicative of purifying selection.

Figure 3A also highlights motif diversity for several important classes of transcriptional regulators. For example, HOX-, POU-, and FOX-domain factors are heavily enriched in developmental regulators and controllers of cellular differentiation. Motifs for transcription factors belonging to these classes are markedly shifted toward lower diversity, and motifs for several individual factors exhibit levels of diversity that are reduced beyond that of protein-coding sequences (Fig. 3A). In contrast, diversity in motifs for tandem zinc finger transcription factors, which comprise the largest and most diverse class of human transcription factors, is distributed relatively evenly across the diversity spectrum (Fig. 3A). Members of this group include core regulatory factors such as CTCF and YY1, developmental regulators such as PRDM1 and ZIC3, and numerous chromatin repressors such as RREB1, REST, and the KRAB-ZNF family of proteins. Because many of the canonical motifs for these factors contain one or more CG dinucleotides, we hypothesized that the increased average diversity for these factors might be a consequence of higher mutation rates at CpG sites. To explore this hypothesis, we identified factors for which >50% of the motif instances in regulatory DNA contained CpGs, which revealed that the ubiquitous presence of CpG sites is a common characteristic of motifs with high levels of diversity (Fig. 3A).

To more systematically control for mutation rate heterogeneity, we also calculated \(\pi\) normalized for divergence (see Methods) for each motif. As shown in Figure 3B, normalized diversity has the
Figure 3. Significant variation of diversity between 732 cis-regulatory motifs. (A) For each motif, average diversity is plotted as a black circle, and 95% confidence intervals obtained by bootstrapping are shown as gray lines. The light blue and yellow rectangles denote the 95% confidence intervals of diversity in fourfold synonymous sites (FFSs) and the exome, respectively. (Red vertical lines) Motifs that belong to the indicated class of transcription factor. (Black vertical lines) Motifs where at least 50% of all instances of that motif contain a CpG dinucleotide. (B) Normalized diversity in motifs versus non-normalized diversity. Motifs with a CpG (defined as above) are plotted in red. (Dashed line) Best fit for non-CpG motifs ($r = 0.70, P < 10^{-10}$).

Heterogeneity of functional constraint across cell types

We next tested the hypothesis that levels of functional constraint acting on regulatory DNA varied across cell types. To this end, we calculated normalized $\pi$ averaged across all DNase I peaks for each of the 138 cell lines. We found marked differences in normalized diversity between cell lines ($P < 10^{-4}$) (Fig. 4A), which ranged from a low of $5.52 \times 10^{-4}$ in primary hepatocytes to a high of $6.15 \times 10^{-4}$ in the immortalized B-lymphoblastoid cell line GM12864. The majority of cell types exhibited average levels of normalized diversity that are within the range of fourfold degenerate sites (Fig. 4A). Note, as we are averaging over many megabases of sequence in each cell type, this does not mean that specific sites, such as motifs embedded within peaks, are evolving neutrally. Six cell types (retinal pigment epithelial, neuroblastoma, primary liver, skeletal muscle myoblast, umbilical vein endothelial, and prostate adenocarcinoma cells, corresponding to cell lines HRPEpiC, SK-N-SH, Hepatocyte, Hsmm, Huvec, and LNCaP, respectively) exhibited average levels of normalized diversity that are significantly lower (ranging from $P = 0.024$ to $P < 10^{-4}$) than fourfold degenerate sites, indicative of stronger functional constraint.

It is important to note that the magnitude of reduced diversity in these six cell types is much less than that observed for protein-coding genes. Specifically, normalized diversity across the exome is $4.04 \times 10^{-4}$, a reduction of 31.2% compared with fourfold degenerate sites. The stronger signature of purifying selection on exonic sequence relative to regulatory regions defined by DNase I hypersensitivity is likely attributable to both the higher proportion of functionally important variants in protein-coding versus non-coding DNA and that, on average, mutations in exonic sequences are more deleterious than mutations in regulatory regions. Indeed, numerous studies have found that regulatory mutations tend to be mildly deleterious (Asthana et al. 2007; Chen et al. 2007; Ronald and Akey 2007).

We next investigated differences in normalized diversity between “core” DHS and DHS found in only one category of cell types. To this end, all of the cell types can be grouped into one of three categories: normal/primary, iPS/ES, and malignant. To minimize potential contributions from experimental noise, we focused on a subset of 92 cell lines with high-quality DNase I data in which >40% all sequence tags map within DHSs (equivalent to average signal-to-noise of ~100:1) (Thurman et al. 2012) and calculated normalized $\pi$ in DNase I peaks that are shared and unique to each cell type category (Fig. 4B). Eight percent of peaks are found in all three categories, whereas 6.4%, 31.1%, and 28.2% of peaks are unique to iPS/ES, malignant, and normal/primary cell types, respectively (Fig. 4B). Overall, there is significant variation ($P < 10^{-4}$) in normalized diversity among peaks shared between cell type categories versus those found in a single category (Fig. 4B). In particular, DNase I peaks shared by two or three categories of cell types exhibit the lowest levels of normalized diversity (Fig. 4B), consistent with stronger selective constraint. Conversely, peaks found in only one cell category have significantly higher normalized diversity than shared peaks, (Fig. 4B). These results suggest that the “core” set of DHSs, present in more than one cell type category, is subject to stronger purifying selection because they are necessary for proper transcriptional programs in multiple cell types.
Evidence for ectopic activation of DHSs in malignant cell types

Many cancers are characterized by disruptions in chromatin maintenance pathways (Wang et al. 2007; Morin et al. 2010; Jiao et al. 2011). Additionally, many immortalized cells express different complements or ratios of transcriptional factors than are found in normal differentiated cells (Zaidi et al. 2007). These observations suggest that immortalized and malignant cell lines may experience increased "ectopic" activation of DHSs. To explore the potential ectopic activation of DHSs in malignant and immortalized cell types further, we calculated the proportion of DNase I peaks that are present in only one cell line, as noncanonical chromatin remodelling would be expected to result in an excess of cell type–restricted DHSs. Again, we used the same 92 cell types as described above.

We found that 54% of peaks specific to normal/primary cells are present in a single cell type. In contrast, 81% of malignantspecific peaks and 86% of iPS/ES-specific peaks are present in a single cell line. However, these percentages are not directly comparable, because of sample size differences between categories (n = 58, 29, five for normal/primary, malignant, and iPS/ES, respectively). When we correct for the number of cell types per category (see Methods), we find that iPS/ES cells are not enriched for singleton DHSs compared to normal/primary cells (P = 0.236), whereas malignant cell types are significantly enriched (P < 10^-4) for singleton DHSs compared to normal/primary cells (Fig. 5). These data raise the intriguing possibility that the DHSs found in malignant cells, though not increased significantly in number (data not shown), are enriched in elements resulting from ectopic cooperative transcription factor binding within neutrally evolving sequences.

Signatures of positive selection

A large number of genome-wide scans for recent positive selection have been performed in humans (for review, see Akey 2009). Typically, these studies focus only on patterns of DNA sequence variation and are not informed by functional genomics data, although genome-wide analyses have been pursued on computa-

Figure 4. Heterogeneity of polymorphism across cell types. (A) Distribution of normalized nucleotide diversity (black points) across DNase I peaks in 138 cell types. Vertical bars around peaks indicate 95% confidence intervals obtained by bootstrapping. (B) Venn diagram showing the amount of shared and unique sequence for DNase I peaks among normal/primary, malignant, and iPS/ES cell types. The barplot on the left shows average normalized diversity for several categories of peaks in the Venn diagram. Shared all and shared two denote peaks shared among all three categories and between any two categories, respectively. N, M, and SC denotes peaks specific to normal/primary, malignant, and stem cell (iPS/ES) cell types, respectively.

Figure 5. Malignant cell lines exhibit significantly more singleton DNase I peaks than normal cell lines. (Triangles) Observed proportion of singleton peaks. (Blue and green lines) Distribution (density histograms) of singleton peaks when randomly sampling 29 (blue) or five (green) cell types; this is the distribution of the number of singleton peaks we would expect if malignant or stem cells were similar to normal cells, respectively. Note the malignant category (blue) shows significantly more singleton peaks than expected given its sample size, but the stem cell category (green) falls within the expected range.
tionally predicted motifs. The large compendium of experimentally characterized regulatory regions provides a unique data set to interrogate for signatures of recent positive selection.

To this end, we performed a population genomics analysis to identify DNase I peaks that contain variants with large allele frequency differences between populations relative to the genome-at-large, which is a signature of geographically restricted selection (Akey et al. 2002; Akey et al. 2004). Specifically, we calculated locus-specific branch lengths (LSBLs) (Shriver et al. 2004) for variants in DNase I peaks in Africans, Asians, and Europeans. LSBL is a function of pairwise $F_{ST}$ between populations (see Methods) and helps isolate the direction of allele frequency change (Shriver et al. 2004). To reduce the stochasticity inherent in summary statistics of population differentiation, we averaged LSBL across all variants in a peak. We excluded X-linked variants from our analysis due to its different effective population size.

The genome-wide distributions of population structure in DNase I peaks in the African, Asian, and European populations are shown in Figure 6A. We pursued two distinct approaches to interpret these data. First, to obtain general insights into the characteristics of DNase I peaks that exhibit large allele frequency differences between populations, we focused on peaks in the 1% tail of the empirical distribution of LSBLs in each population (Fig. 6A). Next, we identified all genes within 50 kb of these peaks ($n = 3372, 3224$, and $3099$ such genes in Africans, Asians, and Europeans, respectively) and tested for enrichment of KEGG pathways. As shown in Table 1, this set of genes is significantly enriched for 15 KEGG pathways, seven of which are shared between two or more populations (including pathways related to cancer, axon guidance, and WNT signaling). Interestingly, the most significantly enriched pathway in Europeans is melanogenesis (Table 1), suggesting that in addition to protein-coding variants (Lamason et al. 2005), regulatory polymorphisms influencing pigmentation phenotypes have also been a target of recent positive selection. Moreover, our African sample is significantly enriched for chemokine and adipocytokine signaling pathways (Table 1), which is particularly interesting given the known differences in prevalence of insulin resistance and type 2 diabetes in individuals of African ancestry (Reimann et al. 2007).

![Figure 6](image_url)

**Figure 6.** Genome-wide distribution of population structure in regulatory DNA. (A) Genome-wide distribution of locus-specific branch lengths (LSBLs) for Africans, Asians, and Europeans, respectively. Note that the valley of uniform LSBL on chromosome 17 in Europeans corresponds to the MAPT region that is segregating a large chromosomal inversion (Zody et al. 2008). (B) Distribution of the proportion of highly differentiated DNase I peaks found for different categories of cell types. (SC) Stem cells (iPS/ES); (I) immortalized; (M) malignant; (N) normal/primary cell types. (C) Distribution of African LSBL across intron 1 of VDR. (D) Distribution of European LSBL across intron 4 of FTO. In panels C and D, peaks are shown as red rectangles and exons as black rectangles.
This set of peaks is the well-documented promoter variant in peanuts, respectively (Supplemental Text 1). Notably, included in identified genes located within 50 kb of each of these peaks and these criteria in Africans, Asians, and Europeans, respectively. We reported signatures of selection collected by Akey (2009), which is in Africans, Asians, and Europeans, respectively, overlap previously et al. 2002), which demonstrates the potential power of this data that results in malaria resistance in African populations (Hamblin...

...ttribute to fitness differences among individuals, more definitive inferences will require an even broader sampling of cell types. Second, to develop a more refined list of putative targets of recent adaptive evolution, we focused on the most differentiated 1% of DHSs that also contain one or more highly differentiated variants with a GERP $\geq 3$. In total, 323, 349, and 313 DHSs meet these criteria in Africans, Asians, and Europeans, respectively. We identified genes located within 50 kb of each of these peaks and identified 187, 174, and 179 genes in Africans, Asians, and Europeans, respectively (Supplemental Text 1). Notably, included in this set of peaks is the well-documented promoter variant in DARCR that results in malaria resistance in African populations (Hamblin et al. 2002), which demonstrates the potential power of this data set to fine-scale map signatures of selection and identify selected alleles. Moreover, 61, 40, and 51 of these candidate selection genes in Africans, Asians, and Europeans, respectively, overlap previously reported signatures of selection collected by Akey (2009), which is...
nucleotide considered for a particular analysis. Normalized diversity was calculated by dividing the per nucleotide diversity by the total number of bases. Among these 69 individuals, 54 are reported to be unrelated. To verify that these 54 individuals are unrelated, we performed relationship inference with KING (Manichaikul et al. 2010). Two Maasai individuals (NA21732 and NA21737) who were not reported as being related were found to be either siblings or parent–child. We removed NA21737 from further analyses as this individual had more missing data than NA21732. Thus, our final data set consists of 53 individuals from five populations (Supplemental Table 2). Genotype data were filtered to remove partial genotypes (i.e., where one allele is called and the other is reported as missing), by coverage (>20% of individuals must have calls), and by extreme departures from Hardy-Weinberg Equilibrium ($P < 10^{-8}$, which corresponds to all individuals being heterozygous and therefore most likely a paralogous sequence variant). We defined fourfold degenerate sites using NCBI-called reading frames. We used the NimbleGen SeqCapEZ Exome, version 2.0, definition, downloaded from the NimbleGen website (http://www.nimblegen.com/products/seqcap/ez/v2/). Repeats were defined by RepeatMasker regions, obtained from the UCSC Genome Browser. A dinucleotide is conservatively called as CpG for NpG and CpN dinucleotides (where N = A,C,T, or G) in any of our 53 genomes, chimpanzee, orangutan, or rhesus macaque.

**Statistical analyses**

We calculated nucleotide diversity as: $\pi = \frac{1}{4N} \sum_{i=1}^{n} 2p_i(1-p_i)$, where $n$ is the number of chromosomes and $p_i$ is the frequency of the major allele for the $i$th segregating site. To obtain a per nucleotide estimate of $\pi$, we divided by the total number of bases considered for a particular analysis. Normalized diversity was calculated by dividing the per nucleotide $\pi$ estimate by the estimated neutral mutation rate. For exonic sequence, we used the mutation rate calculated at fourfold synonymous sites, as this sequence is less neutral mutation rate. For exonic sequence, we used the mutation rate calculated at the ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. 

**References**


Human regulatory variation


Received November 23, 2011; accepted in revised form May 10, 2012.

Genome Research 1697

www.genome.org
Personal and population genomics of human regulatory variation
Benjamin Vernot, Andrew B. Stergachis, Matthew T. Maurano, et al.

Genome Res. 2012 22: 1689-1697
Access the most recent version at doi:10.1101/gr.134890.111

Supplemental Material

http://genome.cshlp.org/content/suppl/2012/08/22/22.9.1689.DC1

Related Content

Decoding the human genome
Kelly A. Frazer
Genome Res. September, 2012 22: 1599-1601 What does our genome encode?
John A. Stamatoyannopoulos
Genome Res. September, 2012 22: 1602-1611 Toward mapping the biology of the genome
Stephen Chanock
Genome Res. September, 2012 22: 1612-1615 Linking disease associations with regulatory information in the human genome
Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, et al.
Alan P. Boyle, Eurie L. Hong, Manoj Hariharan, et al.
Genome Res. September, 2012 22: 1790-1797

References

This article cites 46 articles, 12 of which can be accessed free at:
http://genome.cshlp.org/content/22/9/1689.full.html#ref-list-1

Articles cited in:
http://genome.cshlp.org/content/22/9/1689.full.html#related-urls

Open Access

Freely available online through the Genome Research Open Access option.

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

© 2012, Published by Cold Spring Harbor Laboratory Press
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.