

High-throughput localization of functional elements by quantitative chromatin profiling

Michael O Dorschner¹, Michael Hawrylycz¹, Richard Humbert¹, James C Wallace¹, Anthony Shafer¹, Janelle Kawamoto¹, Joshua Mack¹, Robert Hall¹, Jeff Goldy¹, Peter J Sabo¹, Ajay Kohli², Qiliang Li², Michael McArthur¹ & John A Stamatoyannopoulos¹

Identification of functional, noncoding elements that regulate transcription in the context of complex genomes is a major goal of modern biology. Localization of functionality to specific sequences is a requirement for genetic and computational studies. Here, we describe a generic approach, quantitative chromatin profiling, that uses quantitative analysis of *in vivo* chromatin structure over entire gene loci to rapidly and precisely localize *cis*-regulatory sequences and other functional modalities encoded by DNase I hypersensitive sites. To demonstrate the accuracy of this approach, we analyzed ~300 kilobases of human genome sequence from diverse gene loci and cleanly delineated functional elements corresponding to a spectrum of classical *cis*-regulatory activities including enhancers, promoters, locus control regions and insulators as well as novel elements. Systematic, high-throughput identification of functional elements coinciding with DNase I hypersensitive sites will substantially expand our knowledge of transcriptional regulation and should simplify the search for noncoding genetic variation with phenotypic consequences.

Understanding the human genome will require comprehensive delineation of functional elements within the 98% of genomic terrain that does not encode protein. *In vivo*, *cis*-regulatory modalities colocalize with focal alterations in chromatin structure^{1–4}, and this governs the accessibility of genomic sequences to critical regulatory factors. Exploitation of the close connection between functional elements and chromatin structure should offer a powerful and generic approach for *de novo* identification of *cis*-regulatory sequences in the context of complex gene domains.

Active regulatory elements within complex genomes are distinguished by pronounced sensitivity to the nonspecific endonuclease DNase I^{3–5} when exposed in the context of intact nuclei. DNase I hypersensitive sites are the *sine qua non* of a diverse spectrum of classical transcriptional and chromosomal regulatory activities including enhancers, promoters, silencers, insulators, boundary elements and locus control regions^{1,3,6}. Indeed, in the human

genome, many functional elements were first identified as major DNase I hypersensitive sites and only later were found to have specific regulatory roles. Analysis of chromatin structure may enable generic delineation of functional elements across the genome, provided it exhibits direct sequence specificity, quantitative data output that permits automated analysis, and adaptability to a high-throughput format.

DNase I hypersensitive sites in native genomic domains have traditionally been localized by an approach relying on Southern transfer followed by indirect end-labeling⁵. Although widely applied, this technique is not quantitative and has numerous technical and resource-related limitations that prevent its application on a genome-wide scale. The major limitations of conventional hypersensitivity assays are the low throughput and the lack of sequence specificity. Conventional Southern blot-based hypersensitivity assays are tremendously time-consuming, and comprehensive mapping of DNase I hypersensitive sites over an extended gene locus (50–200 kilobase, kb) requires months, or in some cases years, of dedicated effort. Because it is an indirect assay, another critical drawback of this conventional approach is its lack of sequence specificity. Identification of the core ~50–200 base pair (bp) elements over which cooperative *trans* factor binding occurs often requires extensive additional functional analysis, such as deletion studies⁷.

We therefore sought to develop a high-throughput, sequence-specific approach to mapping hypersensitive sites. We showed that the continuous distribution of DNase I sensitivity across a gene locus can be accurately sampled in a high-throughput format. These data were analyzed using a rigorous algorithm to construct a quantitative ‘chromatin profile’ and to identify and score hypersensitive outliers via a signal-to-noise ratio (SNR). Application of this quantitative chromatin profiling (QCP) approach to ~300 kb of diverse human genomic terrain demonstrated that peaks in hypersensitivity SNR are highly sensitive and specific for DNase I hypersensitive site core elements and the associated *cis*-regulatory sequences. This approach may also reveal new elements even in the context of well-explored genomic loci.

¹Department of Molecular Biology, Regulome, 2211 Elliott Avenue, Suite 600, Seattle, Washington 98121, USA. ²Division of Medical Genetics, University of Washington, K-253 Health Sciences Building, Box 357720, 1705 NE Pacific Street, Seattle, Washington 98105, USA. Correspondence should be addressed to J.A.S. (jstam@regulome.com).

RESULTS

QCP

Sensitivity of chromatin to DNase I *in vivo* is expected to vary as a continuous function of genomic position. We reasoned that if this variation could be captured quantitatively, it would be possible to distinguish DNase I hypersensitive sites by performing outlier detection using rigorous statistical methods. We hypothesized that continuous quantitative profiles of DNase I sensitivity could be obtained by (i) tiling PCR amplicons across a candidate gene locus, (ii) quantifying the relative DNase I sensitivity of genomic DNA corresponding to each amplicon, and (iii) plotting replicate determinations as a function of genomic position (Fig. 1). We further expected that such profiles could be analyzed to reveal a DNase I sensitivity ‘baseline’ (together with 95% confidence bounds), relative to which statistically significant outliers could be reliably detected using a generic algorithm. DNase I hypersensitive sites typically comprise a core element of roughly 250 bp, where regulatory factors bind and exclude a canonical nucleosome. However, the chromatin structural alterations produced by such complexes vary considerably and may extend to flanking sequences⁷. We therefore predicted that core DNase I hypersensitive site elements would be situated at the local maxima of hypersensitivity, which could be identified rigorously by computing its SNR as a function of genomic position (Fig. 1).

Under standard buffer conditions, DNase I introduces both single-stranded (ss) nicks and double-stranded (ds) cuts into chromatin-assembled genomic DNA templates. Given a population of templates in which some undergo modification whereas others do not, the sensitivity of a given amplicon to DNase I will be reflected by a reduction in its apparent amplification efficiency from a DNase I-treated as compared to an untreated template population. Such differences can be readily detected using replicate quantitative real-time PCR measurements, which may be performed in a high-throughput format. The feasibility of using this protocol to measure the kinetics of DNase I digestion at known hypersensitive sites has been demonstrated previously⁸. However, accurate *de novo* localization of DNase I hypersensitive sites against a moving baseline of *in vivo* DNase I sensitivity presents a considerably greater challenge requiring the development of a quantitative methodology as outlined above.

QCP accurately delineates DNase I hypersensitive sites

To test these concepts, we produced a quantitative chromatin profile over a 21 kb interval spanning the human *HBB* (β -globin) locus control region (LCR), a paradigmatic complex *cis*-regulatory element⁹. The *HBB* LCR comprises an array of functional elements that *in vivo* give rise to five major DNase I hypersensitive sites (designated HS1–HS5) upstream of the *HBE1* (ϵ -globin) gene on chromosome 11 (refs. 10–12). The core elements of these DNase I hypersensitive sites have been localized at the sequence level through extensive functional and chromatin structural studies¹³.

To produce a quantitative chromatin profile of the human *HBB* LCR, we prepared genomic DNA from DNase I-treated and untreated nuclei of K562 cells, an erythroid cell line in which the lineage-specific DNase I hypersensitive sites of the LCR are known to be active¹⁰. We designed a series of 89 contiguous ~225 bp amplicons spanning the LCR, and then obtained nine replicate DNase I sensitivity measurements (ratios of the copy number in treated versus untreated samples) for each amplicon (801 total

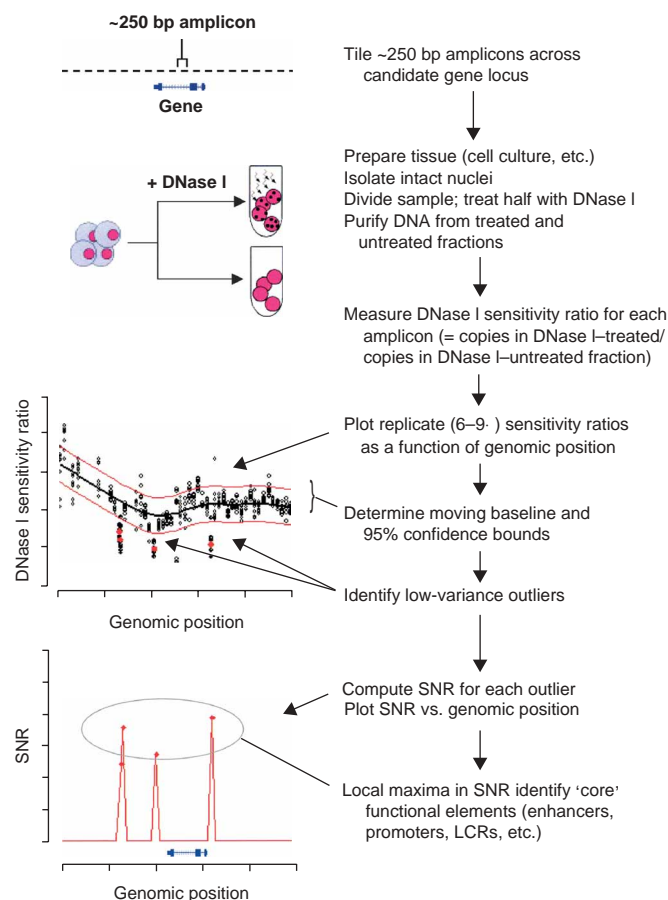


Figure 1 | Overview of quantitative chromatin profiling. To study a candidate gene of interest in the context of a given cell type, intact nuclei are isolated, and DNase I-treated and untreated samples are prepared. Primers are designed to amplify tiled contiguous ~250 bp amplicons spanning the candidate gene locus. The relative number of intact copies of the genomic DNA sequence corresponding to each amplicon is then determined (using a suitable modality, such as real-time PCR) in DNase I-treated and untreated samples in replicate (6–9 \times). The resulting DNase I sensitivity ratios (relative copies in treated / relative copies in untreated samples) are then plotted on a genomic axis. Mean values for 6–9 \times replicates from each amplicon are computed and define a moving DNase I sensitivity ‘baseline,’ relative to which 95% confidence bounds are determined empirically. A robust statistical algorithm is then applied to identify outliers, which signify hypersensitive amplicons. A hypersensitivity SNR is then computed for each outlier amplicon and plotted as a function of genomic position. For a given outlier feature, the SNR provides a quantification of the number of standard deviations by which the feature exceeds the expected background variability, given its empirical distribution. Note that SNR plots contain only values for significant outliers. In a final step, local maxima in SNR are determined, defining the peaks in hypersensitivity that overlie core functional elements.

individual measurements). To arrive at a common DNase sensitivity scale that could be applied to other loci, relative copy ratios were normalized to a standardized reference amplicon from the rhodopsin gene locus on chromosome 3, which is transcriptionally inactive and resistant to DNase I in K562 and other cell types used herein (data not shown). We performed nine replicate measurements for each amplicon and determined both the trend and ‘baseline’ behavior of DNase I sensitivity across the locus. The measurement errors for DNase I sensitivity values clustered around

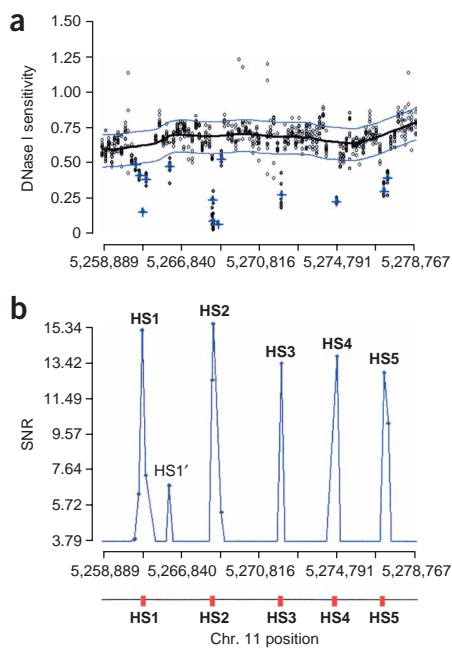


Figure 2 | Peaks in hypersensitivity SNR identify core functional elements. (a) Relative DNase I sensitivity measurements (DNase I-treated versus untreated; y-axis) over 25.6 kb spanning the *HBB* LCR (x-axis: chromosome 11 coordinates). (b) DNase I hypersensitivity in K562 cells expressed as computed SNR. HS1–HS5 are clearly evident; HS1' identifies a previously described non-erythroid-specific HS³¹. Peaks in SNR coincide with the core functional elements of HS1–HS5 (red boxes) as localized through extensive *in vivo* studies. Notably, the amplicons corresponding to SNR peaks localize precisely to the core regulatory factor-binding regions of HS1–HS5 as determined by *in vivo* footprinting and other functional studies (see **Supplementary Fig. 1** online).

All of the DNase I hypersensitive sites of the human *HBB* LCR were correctly and rigorously identified using this method. A weaker hypersensitive site described previously¹³ in K562 cells was also found (designated HS1', **Fig. 2**). We then examined the correlation between the peaks in hypersensitivity SNR and the DNase I hypersensitive site core sequence. In each case, we found that the amplicon with the highest SNR precisely corresponded with sequences that had been determined previously to encompass the core regulatory factor binding domains of HS1–HS5 (**Supplementary Fig. 1** online). These findings demonstrated the ability of QCP to rapidly recognize DNase I hypersensitive sites and to localize their core elements in a sequence-specific fashion.

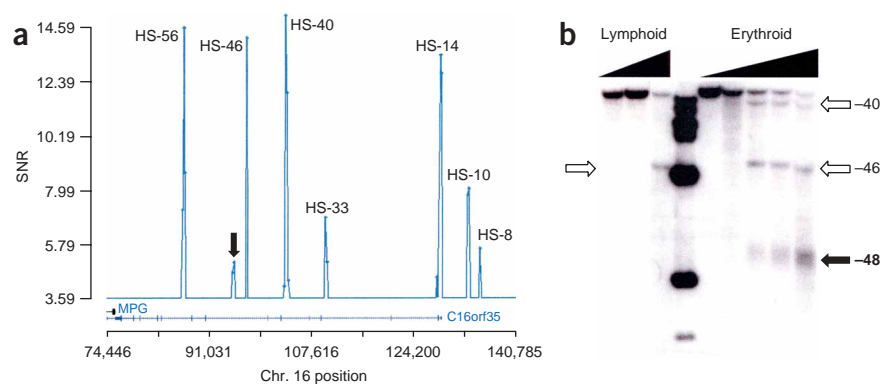
High-throughput mapping of human gene regulatory regions

We then asked whether QCP could similarly expose the *cis*-regulatory architecture of other complex regulatory regions. Chromatin profiles were produced for the *HBA@* (α -globin) locus upstream regulatory region on chromosome 16, the *ADA* locus (adenosine deaminase) on chromosome 20, *CD2* locus (CD2 antigen, p50) on chromosome 1, the *MYC* locus (also known as *c-myc*) on chromosome 8, and the *TCRA* (T-cell receptor- α , TCR- α) locus downstream regulatory region on chromosome 14 (**Figs. 3–6**).

We produced a chromatin profile spanning 66.4 kb upstream of the *HBZ* (ξ -globin) gene in the context of K562 cells (**Fig. 3a**). This region contains the *HBA@* major regulatory region, which is situated in an intron of an unrelated upstream gene¹⁴. The hypersensitive sites identified by QCP encompass all of the major transcriptional

the baseline, and hence confidence bounds on outliers and extreme values for this distribution. We then identified amplicons yielding replicate measurements having low variance (with respect to the mean measurement error) and a mean sensitivity ratio outside of the 95% confidence intervals about the baseline. To visualize the chromatin profile, we computed SNRs for each outlier relative to variability about the baseline and plotted these values on a genomic axis. The SNR is a broadly applied instrument in quantitative assays, in which numerical values reflect the number of standard deviations by which the observed measurement exceeds the expected background variability. Computed hypersensitivity SNRs across the *HBB* LCR are shown in **Figure 2**. Note that because SNRs are only shown for amplicons with significant (outlier) sensitivity ratios, the numerical range on the y-axis is arbitrary; in practice this range is scaled according to the distribution of SNR scores in the plotted segment.

Figure 3 | Identification of functionally diverse *cis*-regulatory sequences with QCP. Peaks in SNR (vertical axes) define functional sequences relative to genomic positions (horizontal axes). Genes (blue) are shown below genomic positions. (a) Quantitative chromatin profile over 66 kb spanning the *HBA@* upstream regulatory region in erythroid cells (K562) using 2,439 DNase I sensitivity measurements from 271 amplicons. Analysis of the resulting profile revealed eight hypersensitive sites at –8 kb, –10 kb, –14 kb, –33 kb, –40 kb, –46 kb, –48 kb and –56 kb relative to the *HBZ* cap site. (b) Analysis of the newly identified –48 kb element (arrow) with conventional hypersensitive site assays in erythroid (K562) and lymphoid (Jurkat) cells. The parental band is a 10 kb *Xba*I fragment, and the probe (508 bp fragment, chromosome 16 94,343–94,922) is situated at the 5' end. The new element shares the property of erythroid specificity with the two major elements of the locus (HS-40 and HS-33) and may therefore be important in α -globin regulation.



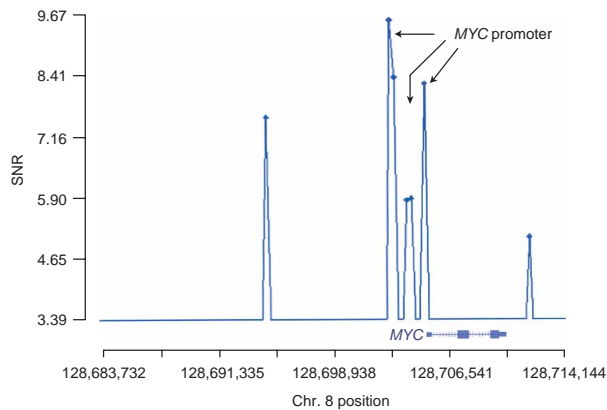


Figure 4 | QCP of the *MYC* locus (30.4 kb) on chromosome 8. Shown is a 30.4 kb chromatin profile derived from 855 measurements over 95 amplicons spanning the *MYC* locus, assayed in HepG2 hepatocellular carcinoma cells, in which *MYC* is transcriptionally active. Chromatin structural studies have identified three functional elements designated HS1–HS3¹⁷. The *MYC* P1/2 promoter is contained in HSIII and the P0 promoter in HS2. We observed three clustered central sites, the SNR peaks of which corresponded precisely with the core elements of HS1–HS3. Two additional sites are evident, one at approximately –11 kb upstream and a second about 1.5 kb downstream of the gene; both correspond to previously documented elements³².

control elements of the *HBA@* upstream regulatory domain and all major hypersensitive sites previously reported following laborious studies of this region with conventional hypersensitivity assays¹⁴.

One of the major questions surrounding *HBA@* locus regulation has been the location of additional erythroid-specific regulatory sequences that complement the activity of the major element at HS-40 and the auxiliary element at HS-33. Using QCP, we identified a site at –48 kb (Fig. 3a) that had not been detected in previous extensive surveys¹⁴. We confirmed that this site could be visualized with conventional assays (Fig. 3b), suggesting that it had been overlooked or was not detectable for technical reasons (such as a lack of appropriately positioned restriction sites). Like HS-40 and HS-33, this new element is erythroid specific, suggesting that it may contain the sought-after regulatory role that complements the other erythroid sites.

ADA is under control of a regulatory region positioned centrally within the 18 kb first intron¹⁵. The active core of this region is distinguished by the presence of a strong central tissue-specific DNase I hypersensitive site, which encodes a powerful transcriptional enhancer that also acts as a locus control region and is the dominant regulatory element of *ADA*. In the thymus, and in cell lines with a T-cell phenotype, this site is the only strongly evident hypersensitive site in the region. We performed QCP over the 58 kb *ADA* locus in Jurkat T cells and readily delineated this site, as well as the *ADA* promoter, which had not been previously analyzed at the chromatin level (Supplementary Fig. 2a online). Using the same preparation of Jurkat T cells, we applied QCP to a 27 kb region encompassing the *CD2* locus. The resulting profile localized the *CD2* promoter and the 3' enhancer–LCR, as well as disclosing a previously unknown intronic element (Supplementary Fig. 2b online).

To demonstrate the ability of QCP to resolve complex, densely co-located elements, we studied the *MYC* locus on chromosome 8. *MYC* belongs to a class of highly regulated genes for which multiple

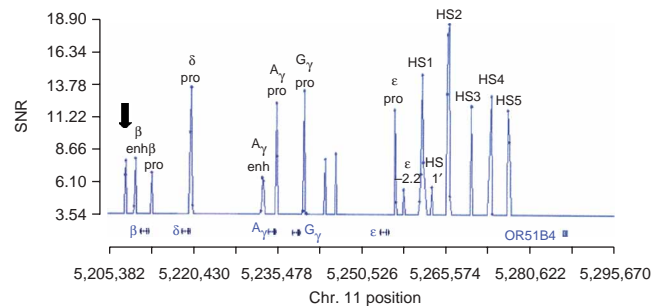


Figure 5 | The 90.4 kb quantitative chromatin profile of the human β -globin locus in K562 cells. Shown (3' to 5') on the horizontal axis are the genomic positions of the *HBE* (ϵ), *HBG1* ($^G\gamma$), *HBG2* ($^A\gamma$), *HBD* (δ) and *HBB* (β) genes, as well as an olfactory receptor–like gene (*OR51814*) located 5' of the LCR. The 90.4 kb (3,393 measurements over 377 amplicons) *HBB* quantitative chromatin profile from K562 cells revealed, in a single experiment, all of the major *cis*-regulatory elements of the globin locus, including the LCR (HS1–HS5); the *HBE* promoter together with an upstream element; the *HBG1* promoter; the *HBG2* promoter; the *HBG2* 3' enhancer; the *HBD* promoter; the *HBB* promoter; and the *HBB* 3' enhancer. The profile identified novel features (unlabeled peaks), including an element downstream of the *HBB* enhancer (arrow, far left) that coincides with a focus of intergenic transcription in erythroid cells¹⁸. Note the prominence of the *HBD* promoter, which is consistent with active δ -globin transcription in K562 cells³³.

distinct promoter elements have been described¹⁶, which are contained within hypersensitive sites *in vivo*¹⁷. A 30.4 kb chromatin profile performed in hepatocellular carcinoma (HepG2) cells revealed three clustered central sites, the SNR peaks of which corresponded precisely to the three core elements of the *MYC* promoter complex (Fig. 4).

Delineation of *cis* elements across a multigene locus

We next asked whether the approach could resolve correctly, in a single experiment, the *cis*-active elements over a large multigene locus. To test this, we profiled the entire human β -like globin gene domain on chromosome 11, which comprises five genes (*HBE*, *HBG2*, *HBG1*, *HBD* and *HBB*, encoding ϵ -, $^G\gamma$ -, $^A\gamma$ -, δ - and β -globin, respectively) organized in a 5'-to-3' fashion that corresponds to the timing of their expression during development and differentiation¹³. In addition to the LCR, known *cis*-active elements of the locus comprise the promoters of *HBE*, *HBG2*, *HBG1*, *HBD* and *HBB*, the *HBG2* enhancer, and the 3' *HBB* enhancer. We performed QCP over a ~90 kb region encompassing the locus in K562 cells and localized all of the aforementioned elements (Fig. 5), as well as new elements. In particular, we observed an element distal to the *HBB* 3' enhancer whose presence we also confirmed with conventional hypersensitivity assays (data not shown). This element lies within a ~650 bp gap between two large repetitive tracts and coincides with a reported focus of intergenic transcription in erythroid cells¹⁸, suggesting a new regulatory role.

Analysis of the human *TCRA* regulatory domain

The human *TCRA* locus is embedded in a highly conserved segment of chromosome 14 and has not been analyzed systematically at the chromatin level. Knowledge of the *cis*-regulatory features of this locus derives largely from functional studies of the homologous mouse domain¹⁹, including the identification of a locus control region²⁰ and an insulator activity²¹. In human cells,

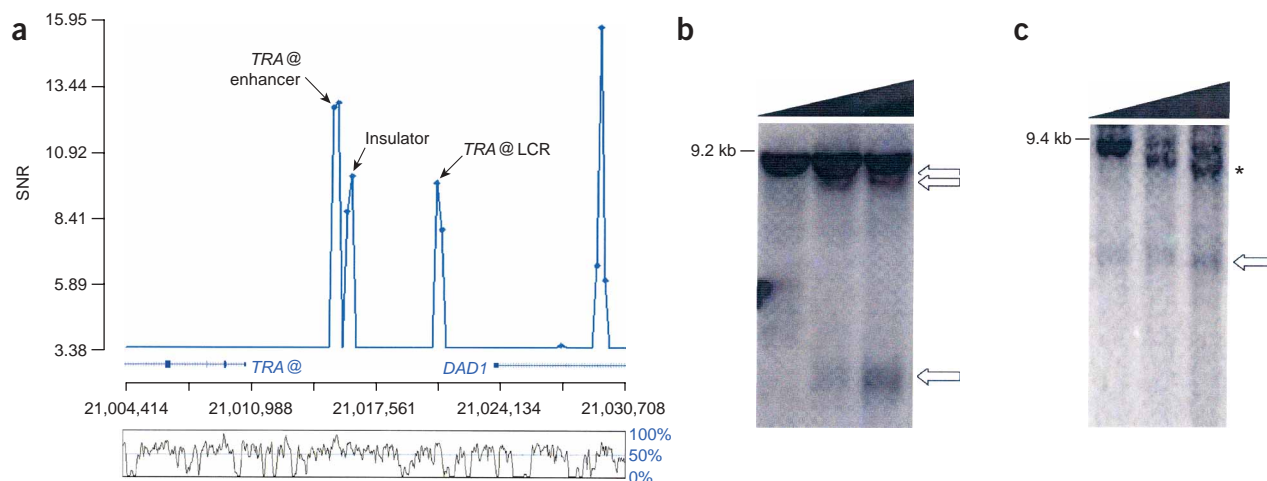


Figure 6 | Human *TCRA* (TCR- α) downstream regulatory region. **(a)** Quantitative chromatin profile (26.3 kb; 864 measurements over 96 amplicons assayed in Jurkat T cells) shows the spatial organization of the human *TCRA* regulatory region to be similar to that of the mouse domain, including the previously described human *TCRA* enhancer²², the human homolog of the mouse T cell α LCR, and an insulator element. A prominent hypersensitive site is also detected within an intron of the widely expressed *DAD1* gene. Alignment of orthologous human and mouse sequences reveals extensive conservation across the *TCRA* regulatory region (below, % human-mouse conservation). Specific *in vivo* elements defined by QCP show varying degrees of conservation, but discrimination from other sequences in the locus showing similar levels of conservation is difficult. **(b)** Detection of DNase I hypersensitive sites (arrows) within the *TCRA* regulatory region in Jurkat cells. Parental band is a 9.2 kb *Bsp*HI fragment. **(c)** Detection of intronic DNase I hypersensitive site in *DAD1* (arrow). The parental band is the 9.4 kb *Nde*I fragment. An additional hypersensitive site evident in the figure (“*”) lies further downstream of the region profiled in **a**.

functional studies have disclosed an enhancer located ~ 4.5 kb 3' to the *TCRA*²².

We applied QCP to examine whether the spatial organization of the human regulatory region parallels that of the mouse. Comparative genomic analyses provide little insight into this question because of the high level of background noncoding sequence conservation (Fig. 6). We detected four prominent hypersensitive site core elements, which cleanly delineated the *TCRA* enhancer²², a juxtaposed insulator element, and the human homolog of the *TCRA* locus control region. A previously unidentified human element was evident within an intron of the downstream widely expressed *DAD1* gene (Fig. 6). This example illustrates the power of a functionally based approach to illuminate and inform phylogenetic analyses.

Accuracy of QCP

Of the 23 previously described classical *cis*-regulatory sequences expected to be active in the study cell types, 100% were successfully detected by QCP (Supplementary Table 1 online). Moreover, 29 of 29 previously described major human DNase I hypersensitive sites were also detected. We surveyed a total 1,167 amplicons in the best explored human genomic domains and identified a total of 41 SNR peaks, of which 33 coincided with previously defined sequence elements. Therefore, on the basis of this result alone (that is, not considering any additional validation studies, such as those presented in Figs. 4 and 6), the specificity of QCP (defined as the ratio of true-negative results to true-negative plus false-positive results) is at least 99.3% ($= [1,167 - 41] / \{[1,167 - 41] + [41 - 33]\}$).

DISCUSSION

Sequence-specific localization of functional elements in the noncoding 98% of the human genome is a basic requirement for

studies of gene regulation, for computational analysis of genomic sequences, for the interpretation of comparative genomic data and for systematic identification of a major class of functional genetic variation that has heretofore been largely obscure. We have shown that quantitative chromatin profiling may be applied to achieve precise localization of DNase I hypersensitive sites and, thereby, a diverse array of functional noncoding elements. QCP also revealed previously unidentified elements, which, surprisingly, were detected even in the context of the most heavily explored human genomic domains, the *HBB* and *HBA@* loci.

Functional noncoding elements are rare in the human genome, and the critical test of any methodology applied to their detection over genomic distances is the specificity of the result. The specificity of QCP exceeds 99% and thereby substantially eclipses the performance of any described molecular, phylogenetic or computational approach. In addition to the accuracy and sequence specificity, a major strength of QCP is its applicability to diverse functional elements. This contrasts sharply with methodologies such as chromatin immunoprecipitation, which can only detect the presence of individual factors for which suitable antibodies are available.

The vast majority of human genetic variation resides in noncoding regions²³. Genetic alterations in functional noncoding elements are expected to play a central role in the modulation of quantitative traits, including those that characterize the phenotypes of major common diseases. The search for such variation, however, has been complicated by the lack of methodologies for systematic exposition of functional noncoding elements within candidate gene loci. The ability of QCP to discriminate rapidly and accurately functional regions of the noncoding genome should have a substantial impact on our ability to discern functionally important alleles, identify heretofore elusive regulatory polymorphisms and dissect *cis*-acting contributors to phenotypic variation.

METHODS

Cell culture and DNase I digestion. See **Supplementary Methods** online.

Primer selection. We designed primers to amplify contiguous or minimally overlapping ~250 bp amplicons across target genomic regions. Primers were designed using Primer3 (ref. 24) restricting several parameters, including target amplicon size (250 bp \pm 50 bases), primer T_m (optimal, 60 °C \pm 2 °C), %GC (50% optimal, range 40–80%) and length (optimal 24, range 19–27), and the poly X (maximum 4). Primers were then scanned for repetitive sequences by alignment with a database of repetitive sequences.

Conventional DNase I assays. Conventional DNase I hypersensitivity studies were performed using the indirect end-label technique⁵ according to a standard protocol as described previously²⁵.

Measurement of relative DNase I sensitivity by quantitative PCR. We assembled 15 μ l real-time quantitative PCR reactions using 0.9 μ M forward and reverse primers, 30 ng template DNA (untreated or DNase I treated) and master mix composed of 1 \times FastStart buffer (Roche), 200 μ M of each dATP, dCTP, dGTP, dTTP, 3 mM MgCl₂ and FastStart Taq DNA polymerase (0.033 U/ μ l). The reaction mixture was supplemented with 0.33 \times SYBR green I stain and 300 nM 6-ROX (Molecular Probes) to detect the accumulation of PCR product during amplification and to normalize fluorescence intensity, respectively. All quantitative PCR reactions were set up robotically with a Biomek FX (Beckman). Samples were run in triplicate on individual 384-well plates, and thermocycled with an ABI 7900HT Sequence Detection System (Applied Biosystems).

Primary data analysis. We analyzed primary data in four phases: (i) determination of cycle threshold (C_t) values, (ii) amplification efficiency correction, (iii) melting curve analysis, and (iv) calculation of DNase I sensitivity ratios. Normalized fluorescence data were obtained and an amplification curve and n^{th} -order polynomial fit were computed for each reaction. The C_t values were then determined for each curve. The amplification efficiency of a reference amplicon selected from the inactive and DNase-insensitive *RHO* (rhodopsin) locus (3q21–q24) was determined empirically for every reaction plate using a standard dilution series of DNA and the equation $E = 10^{-1/\text{slope}}$. We derived the efficiency of each test amplicon from the amplification curve²⁶. Efficiency corrections were performed on all test amplicons with respect to the reference amplicon, and then we calculated relative copy number differences using the comparative C_t method²⁷. Melting curve analysis was done for each amplicon to discard those yielding multiple products. Efficiency-corrected C_t values were then used to compute a relative copy number ratio by applying the formula $2^{-\Delta\Delta C_t}$ or $2^{-(\text{treated}(\text{target} - \text{reference}) - \text{calibrator}(\text{target} - \text{reference}))}$. Relative DNase I sensitivity ratios (relative copy ratios) were thus obtained. Ratios < 1 were indicative of relative copy loss resulting from preferential cleavage of chromatin by DNase I.

Statistical analysis of DNase I sensitivity and hypersensitivity. For a given genomic locus we expect, subject to inherent measurement error, a baseline response indicating general trends in DNase I sensitivity across the locus. The deviation of repeated DNase I

sensitivity measurements (at a given coordinate) from this baseline signals the presence of a hypersensitive site. The presence and significance of hypersensitive sites were quantified by using the following steps:

1. Identification of the trend or baseline behavior of a locus over an extended region.
2. Determination of the measurement error for DNase I sensitivity measurement values clustered around the baseline, and hence confidence bounds on outliers and extreme values for this distribution.
3. Identification of outliers that under repeated measurement have a clustering behavior or low variance with respect to the mean measurement error.
4. Assignment of a SNR to quantify the significance of this observation from the baseline.

An important observation that recurs throughout the analysis of DNase I sensitivity measurements (as well as other genomic data types such as DNA microarray data) is the non-Gaussian behavior of measurements. As the standard error increases, the ratio of measurements from Gaussian random variates approaches the Cauchy or Lorentz distribution. This has been shown to be the case for DNA microarray data²⁸, where more robust methods for treating outliers are often necessary.

To determine the baseline of DNase I sensitivity measurements across a locus, a linear pass through the locus dataset is first performed using a 20% trimmed mean to remove glaring outliers. The remaining data set is smoothed using a locally weighted least squares (LOWESS) smoother that is adapted for robust locally weighted time series and scatter plot smoothing, as described²⁹. A parameter (values 0–1) controls the size of the smoothing window and the standard value of 0.2 (which corresponding to using 20% of the locus in the fit) provides good results across a wide range of genomic loci. An important consideration in our choice of this method is its robustness in handling non-Gaussian time series.

To quantify the noise about the smooth baseline, data are mean centered about the moving baseline and about the outliers of this distribution using the median average deviation approach³⁰. According to this method, a value X is declared to be an outlier if $0.6745 |X - M| / MAD > 2.24$ where M is the median and MAD is the average median deviation. Trimming the data in this way removes both the lower and upper extremes of the distribution, thereby addressing the problem of masking resulting from low sample number breakdown.

Clusters of hypersensitive site scores whose trimmed means occur outside the lower noise threshold from the baseline were then identified. To achieve this while keeping the analysis robust with respect to measurement error, another linear pass of the data is performed identifying groups at a common genomic position whose 20% trimmed mean lies strictly below the interpolated value at the lower shifted baseline. A small correction factor eliminates from consideration groups with very high variance or those consisting of a single point (zero variance).

Outliers corresponding to hypersensitive sites are further scored using SNR. To compute the SNR we compute $S / N_i = |HS_i - B_i| / MAD_B (1 + \sigma_C)$, where the SNR at site i is measured as the absolute deviation of the trimmed mean of the hypersensitive site cluster (HS_i) from the interpolated baseline (B_i), divided by the median average deviation of the centered baseline. The remaining term in the denominator is a small correction factor that penalizes

larger variances in hypersensitive site clusters and rewards highly compact clusters. The scoring method described above has merit in that it makes few distribution assumptions about the data and is robust under a relatively broad set of profiling scenarios.

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

ACKNOWLEDGMENTS

This work was supported by Public Health Service/National Institutes of Health grant U01 HG003161-01 under the National Human Genome Research Institute Encyclopedia of DNA Elements (ENCODE) Project.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 10 August; accepted 19 October 2004

Published online at <http://www.nature.com/naturemethods/>

- Felsenfeld, G. Chromatin unfolds. *Cell* **86**, 13–19 (1996).
- Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448–453 (2003).
- Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
- Elgin, S.C. Anatomy of hypersensitive sites. *Nature* **309**, 213–214 (1984).
- Wu, C. The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854–860 (1980).
- Burgess-Beusse, B. *et al.* The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA* **99** (Suppl. 4), 16433–16437 (2002).
- Lowrey, C.H., Bodine, D.M. & Nienhuis, A.W. Mechanism of DNase I hypersensitive site formation within the human globin locus control region. *Proc. Natl. Acad. Sci. USA* **89**, 1143–1147 (1992).
- McArthur, M., Gerum, S. & Stamatoyannopoulos, G. Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI hypersensitivity of the mouse beta-globin LCR. *J. Mol. Biol.* **313**, 27–34 (2001).
- Li, Q., Peterson, K.R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).
- Tuan, D., Solomon, W., Li, Q. & London, I.M. The “beta-like-globin” gene domain in human erythroid cells. *Proc. Natl. Acad. Sci. USA* **82**, 6384–6388 (1985).
- Forrester, W.C., Thompson, C., Elder, J.T. & Groudine, M. A developmentally stable chromatin structure in the human beta-globin gene cluster. *Proc. Natl. Acad. Sci. USA* **83**, 1359–1363 (1986).
- Grosfeld, F. *et al.* The dominant control region of the human beta-globin domain. *Ann. NY Acad. Sci.* **612**, 152–159 (1990).
- Stamatoyannopoulos, G. & Grosfeld, F. Hemoglobin switching. in *The Molecular Basis of Blood Diseases* (eds. Stamatoyannopoulos, G., Majerus, P., Perlmutter, R. & Varmus, H.) 135–182 (W.B. Saunders, Philadelphia, 2001).
- Higgs, D.R. *et al.* A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes. Dev.* **4**, 1588–1601 (1990).
- Aronow, B. *et al.* Evidence for a complex regulatory array in the first intron of the human adenosine deaminase gene. *Genes. Dev.* **3**, 1384–1400 (1989).
- Bentley, D.L. & Groudine, M. Novel promoter upstream of the human c-myc gene and regulation of c-myc expression in B-cell lymphomas. *Mol. Cell. Biol.* **6**, 3481–3489 (1986).
- Siebenlist, U., Hennighausen, L., Battey, J. & Leder, P. Chromatin structure and protein binding in the putative regulatory region of the c-myc gene in Burkitt lymphoma. *Cell* **37**, 381–391 (1984).
- Gubin, A.N., Njoroge, J.M., Bouffard, G.G. & Miller, J.L. Gene expression in proliferating human erythroid cells. *Genomics* **59**, 168–177 (1999).
- Hong, N.A. *et al.* A targeted mutation at the T-cell receptor alpha/delta locus impairs T-cell development and reveals the presence of the nearby antiapoptosis gene *Dad1*. *Mol. Cell. Biol.* **17**, 2151–2157 (1997).
- Diaz, P., Cado, D. & Winoto, A. A locus control region in the T cell receptor alpha/delta locus. *Immunity* **1**, 207–217 (1994).
- Zhong, X.P. & Krangel, M.S. An enhancer-blocking element between alpha and delta gene segments within the human T cell receptor alpha/delta locus. *Proc. Natl. Acad. Sci. USA* **94**, 5219–5224 (1997).
- Ho, I.C., Yang, L.H., Morle, G. & Leiden, J.M. A T-cell-specific transcriptional enhancer element 3' of C alpha in the human T-cell receptor alpha locus. *Proc. Natl. Acad. Sci. USA* **86**, 6714–6718 (1989).
- Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
- Rozen, S. & Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. in *Bioinformatics Methods and Protocols* (eds. Misener, S. & Krawetz, S.A.) 365–386 (Humana Press, Totowa, N.J., 2000).
- Stamatoyannopoulos, J.A., Goodwin, A., Joyce, T. & Lowrey, C.H. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.* **14**, 106–116 (1995).
- Ramakers, C., Ruijter, J.M., Deprez, R.H. & Moorman, A.F. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66 (2003).
- Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{(-Delta Delta C(T))} Method. *Methods* **25**, 402–408 (2001).
- Brody, J.P., Williams, B.A., Wold, B.J. & Quake, S.R. Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl. Acad. Sci. USA* **99**, 12975–12978 (2002).
- Chambers, J.M., Cleveland, W.S. & Tukey, P.A. *Graphical Methods for Data Analysis* (Wadsworth, Belmont, California, USA, 1983).
- Rousseeuw, P.J. & van Zomeren, B.C. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* **85**, 633–639 (1990).
- Forrester, W.C., Takegawa, S., Papayannopoulou, T., Stamatoyannopoulos, G. & Groudine, M. Evidence for a locus activation region: the formation of developmentally stable hypersensitive sites in globin-expressing hybrids. *Nucleic Acids Res.* **15**, 10159–10177 (1987).
- Mautner, J. *et al.* Identification of two enhancer elements downstream of the human c-myc gene. *Nucleic Acids Res.* **23**, 72–80 (1995).
- Mookerjee, B., Arcasoy, M.O. & Atweh, G.F. Spontaneous delta- to beta-globin switching in K562 human leukemia cells. *Blood* **79**, 820–825 (1992).