LETTER

DNase I sensitivity QTLs are a major determinant of human expression variation

Jacob F. Degner[1,2]*, Athma A. Pai[1]*, Roger Pique-Regi[1]*, Jean-Baptiste Veyrieras[1,3], Daniel J. Gaffney[1,4], Joseph K. Pickrell[1], Sherryl De Leon[4], Katelyn Michelini[4], Noah Lewellen[4], Gregory E. Crawford[5,6], Matthew Stephens[1,7], Yoav Gilad[1] & Jonathan K. Pritchard[1,4]

The mapping of expression quantitative trait loci (eQTLs) has emerged as an important tool for linking genetic variation to changes in gene regulation[1–5]. However, it remains difficult to identify the causal variants underlying eQTLs, and little is known about the regulatory mechanisms by which they act. Here we show that genetic variants that modify chromatin accessibility and transcription factor binding are a major mechanism through which genetic variation leads to gene expression differences among humans. We used DNase I sequencing to measure chromatin accessibility in 70 Yoruba lymphoblastoid cell lines, for which genome-wide genotypes and estimates of gene expression levels are also available[6–8]. We obtained a total of 2.7 billion uniquely mapped DNase I-sequencing (DNase-seq) reads, which allowed us to produce genome-wide maps of chromatin accessibility for each individual. We identified 8,902 locations at which the DNase-seq read depth correlated significantly with genotype at a nearby single nucleotide polymorphism or insertion/deletion (false discovery rate = 10%). We call such variants 'DNase I sensitivity quantitative trait loci' (dsQTLs). We found that dsQTLs are strongly enriched within inferred transcription factor binding sites and are frequently associated with allele-specific changes in transcription factor binding. A substantial fraction (16%) of dsQTLs are also associated with variation in the expression levels of nearby genes (that is, these loci are also classified as eQTLs). Conversely, we estimate that as many as 55% of eQTL single nucleotide polymorphisms are also dsQTLs. Our observations indicate that dsQTLs are highly abundant in the human genome and are likely to be important contributors to phenotypic variation.

It is now well established that eQTLs are abundant in a wide range of cell types and in diverse organisms, and recent studies have implicated human eQTLs as being important contributors to phenotypic variation[1–5]. However, the underlying regulatory mechanisms by which eQTLs affect gene expression remain poorly understood. One mechanism that may be important is when the alternative alleles at a particular single nucleotide polymorphism (SNP) lead to different levels of transcription factor binding or nucleosome occupancy at regulatory sites; this in turn may lead to allele-specific differences in transcription rates[9–12]. In this study we used DNase-seq in a panel of 70 individuals and found that a large fraction of eQTLs are indeed probably caused by this type of mechanism.

DNase-seq is a genome-wide extension of the classical DNase I footprinting method[13–15]. This assay identifies regions of chromatin that are accessible (or 'sensitive') to cleavage by the DNase I enzyme. Such regions are referred to as DNase I-hypersensitive sites (DHSs). DNase I sensitivity provides a precise, quantitative marker of regions of open chromatin and is well correlated with a variety of other markers of active regulatory regions including promoter-associated and enhancer-associated histone marks. Furthermore, bound transcription factors protect the DNA sequence within a binding site from DNase I cleavage, often producing recognizable 'footprints' of decreased DNase I sensitivity[13,15–17].

We collected DNase-seq data for 70 HapMap Yoruba lymphoblastoid cell lines for which gene expression data and genome-wide genotypes were already available[6–8]. We obtained an average of 39 million uniquely mapped DNase-seq reads per sample, providing individual maps of chromatin accessibility for each cell line (see Supplementary Information for all analysis details). Our data allowed us to characterize the distribution of DNase I cuts within individual hypersensitive sites at extremely high resolution. As expected, the DHSs coincided to a great extent with previously annotated regulatory regions, and DNase I sensitivity was positively correlated with the expression levels of nearby genes (Supplementary Figs 6 and 7). Overall, the locations of hypersensitive sites were highly correlated across individuals (Supplementary Information)[11].

We tested for genetic variants that affect local chromatin accessibility. To do this, we divided the genome into non-overlapping 100-base-pair (bp) windows, and then focused our analysis on the 5% of windows with the highest DNase I sensitivity (see Supplementary Information). For each individual we treated the number of DNase-seq reads in a given window, divided by the total number of mapped reads, as a quantitative trait that estimated the level of chromatin accessibility. We then tested for association between individual-specific DNase I sensitivity in each window and genotypes of all SNPs and insertions/deletions (indels) in a cis-candidate region of 40 kilobases (kb) centred on the target window.

Using this procedure, we identified associations between genotypes and inter-individual variation in DNase-seq read depth in 9,595 windows at a false discovery rate (FDR) of 10% (corresponding to 8,902 distinct DHSs, once we combined adjacent windows whose hypersensitivity data were associated with the same SNP or indel; Fig. 1a). We refer to these 8,902 loci as 'DNase I sensitivity QTLs', or dsQTLs, and show an example in Fig. 1c–f. We additionally considered a much smaller cis-candidate region of only 2 kb around each target window and found that most of the dsQTLs were detected within this smaller region (7,088 associated windows in 6,070 DHSs), suggesting that most dsQTLs lie close to the target DHS. In contrast, we found only weak evidence of trans-acting dsQTLs, probably because our experiment was underpowered for detecting these (Supplementary Information). For dsQTLs with enough DNase-seq reads overlapping the most significant SNP ($n = 892$), we confirmed that the fraction of reads carrying each allele in heterozygotes was well correlated with the dsQTL effect sizes (correlation coefficient $r = 0.72$, $P \ll 10^{-16}$; Fig. 1b).

We observed that dsQTLs typically affected chromatin accessibility for about 200–300 bp (Fig. 2a). Of the DHSs affected by dsQTLs, 77%
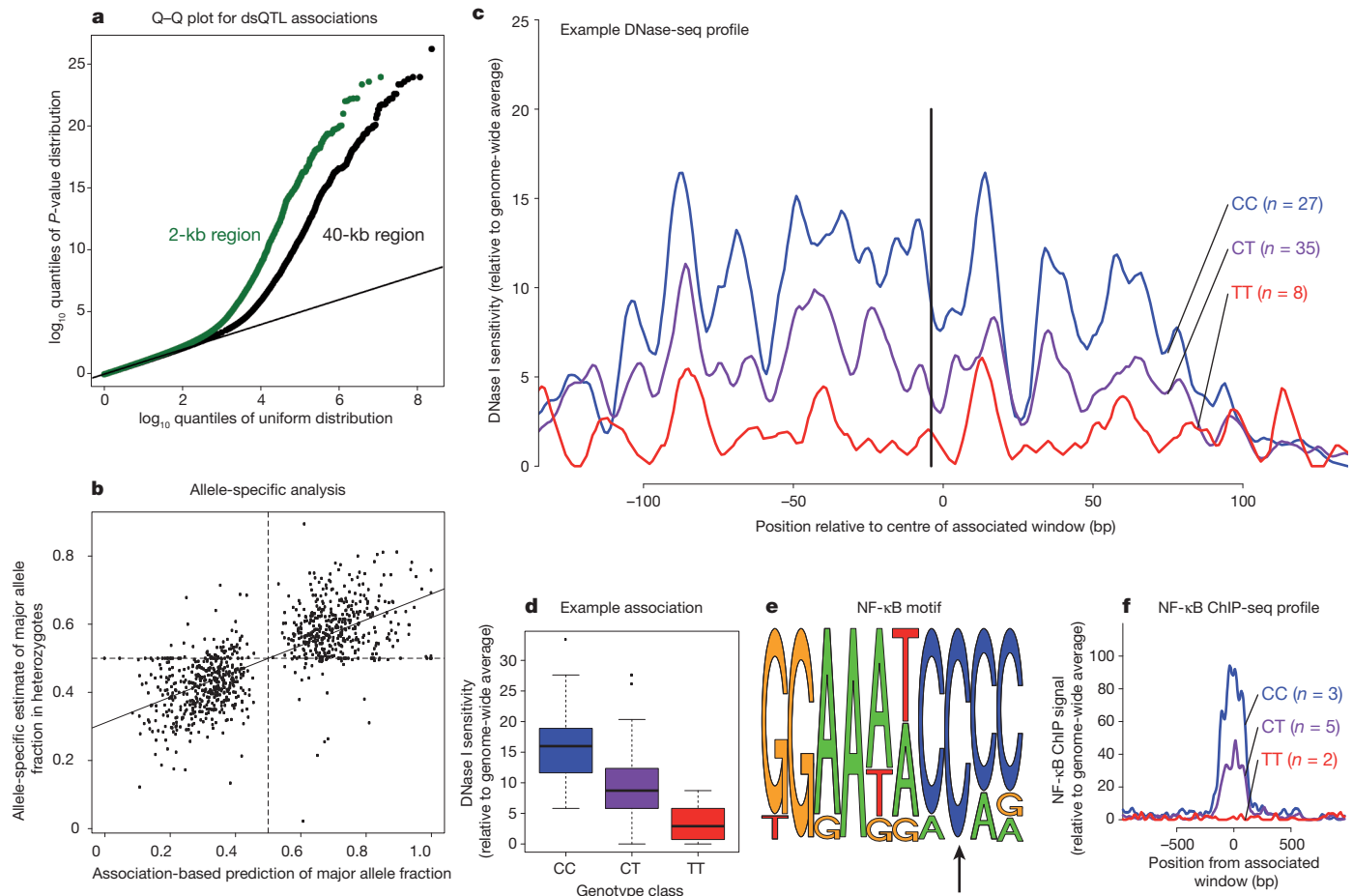
Figure 1 | Genome-wide identification of dsQTLs and a typical example.
a, Q–Q plots for all tests of association between DNase I cut rates in 100-bp windows, and variants within 2-kb (green) and 40-kb (black) regions centred on the target DHS windows. **b**, Allele-specific analysis of dsQTLs in heterozygotes. Plotted are the predicted (x axis) and observed (y axis) fractions of reads carrying the major allele based on the genotype means. **c**, Example of a dsQTL (rs4953223). The black line indicates the position of the associated SNP. **d**, Box plot showing that rs4953223 is strongly associated with local chromatin accessibility ($P = 3 \times 10^{-13}$). **e**, The T allele, which is associated with low DNase I sensitivity, disrupts the binding motif of a previously identified NF-κB-binding site at this location[14]. **f**, NF-κB ChIP-seq data from ten individuals[7] indicates a strong effect of this SNP on NF-κB binding.

lie in chromatin regions previously predicted[18] to be functional in lymphoblastoid cell lines: 41% in predicted enhancers, 26% in promoters, and 10% in insulators, even though those chromatin states together cover only 6.7% of the genome overall (and 38% of our hypersensitive sites).

We next studied the properties of *cis*-acting variants that generated dsQTLs, with the use of a Bayesian hierarchical model that accounted for the uncertainty about which sites are causal[19] (Supplementary Information). This model obtained unbiased estimates of the average properties of causal sites even though, because of linkage disequilibrium, it was typically uncertain which site was causal for any individual dsQTL (Supplementary Information). As shown in Fig. 2b, c, most dsQTLs were generated by variants close to the target window. We estimate that 56% of the dsQTLs were due to variants that lay within the same DHSs and that 67% lay within 1 kb of the target window. dsQTLs that lay more than 1 kb from the target window were themselves significantly enriched in non-adjacent DHS windows (2.4-fold compared with matched random SNPs) and were often associated with changes in sensitivity in multiple non-adjacent DHS windows (Supplementary Fig. 15).

One intuitive mechanism for dsQTLs is that these may be caused by variants that strengthen or weaken individual transcription factor binding sites, thereby changing transcription factor affinity and local nucleosome occupancy[20–22] and hence DNase I cut rates. Consistent with this model, an aggregated plot of DNase I sensitivity at dsQTLs showed a distinct drop in chromatin accessibility around putatively

causal SNPs that was reminiscent of transcription factor binding footprints, especially in the genotypes associated with high sensitivity[15–17].

To test the importance of disruption of transcription factor binding sites as a mechanism underlying dsQTLs, we again turned to the Bayesian hierarchical model. We used the union of all published footprint locations in lymphoblastoid cell lines[16,17] and a set of footprints that we identified from the DNase-seq data reported in this study (Supplementary Methods). Analysis using the hierarchical model indicated a 3.6-fold enrichment of dsQTLs within transcription factor binding footprints ($P \ll 10^{-16}$), controlling for the overall enrichment within DHSs. In addition, the allele associated with a higher score of the position weight matrix is typically associated with higher chromatin accessibility ($P \ll 10^{-16}$), which is consistent with the expectation that higher transcription factor binding affinity leads to more open chromatin (Fig. 2d). Of the dsQTLs that fell within DNase-seq footprints tied to specific transcription factor motifs (using CENTIPEDE[17]), CCCTC binding factor (CTCF), cAMP-response element (CRE) and interferon-stimulated response element (ISRE) were the most enriched, whereas MADS box transcription enhancer factor 2 (MEF2) was significantly depleted.

To further understand the functional consequences of dsQTLs, we examined ChIP-seq data for nine transcription factors collected by the ENCODE Project in one or more lymphoblastoid cell lines[10,23]. Overall, the alleles that were associated with increased DNase I sensitivity were highly associated with increased transcription factor
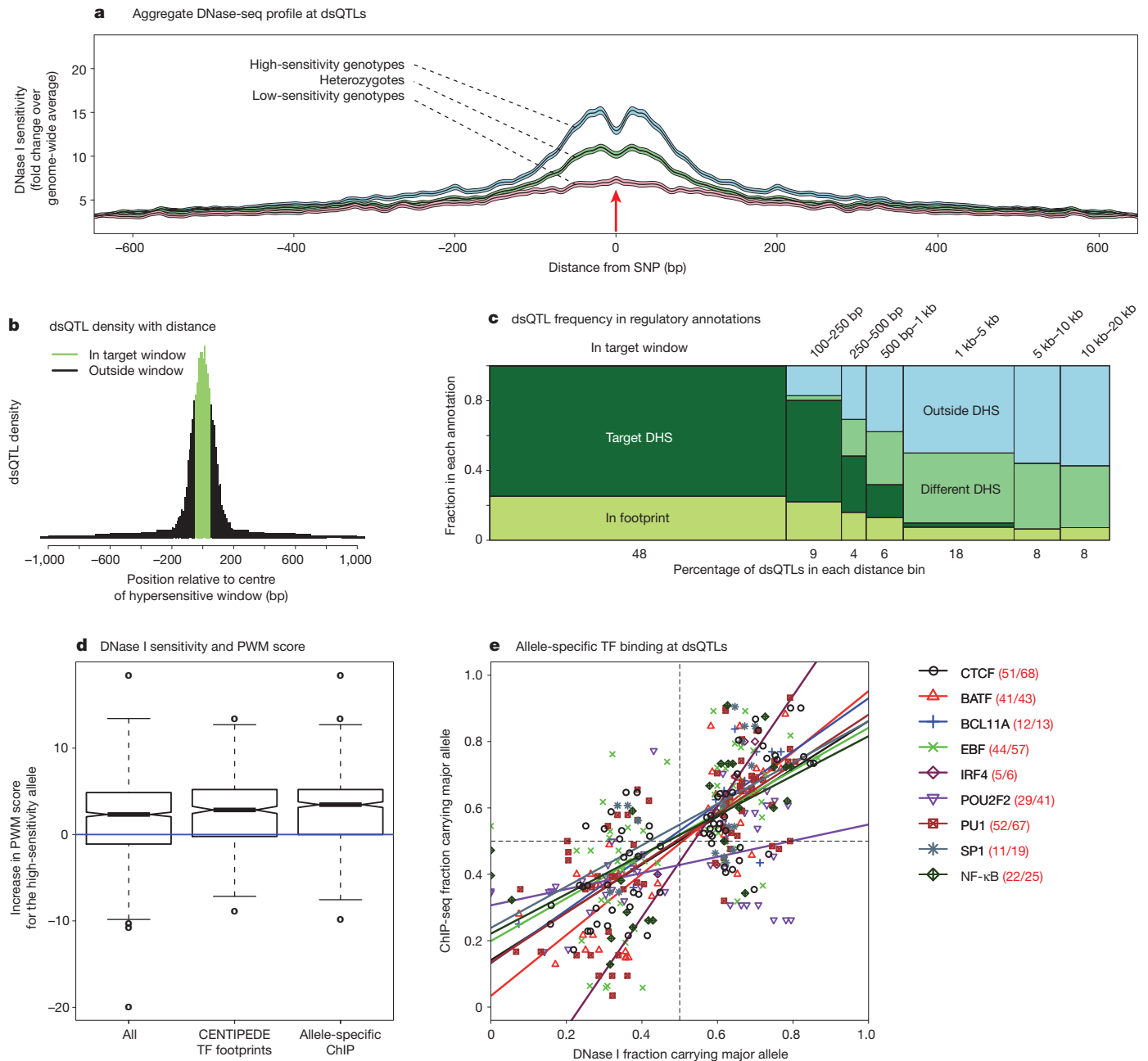
Figure 2 | Properties of dsQTLs. a, Aggregated plot of DNase I sensitivity for high-confidence dsQTLs that lie within the target DHS. Individuals were separated into the high-sensitivity (blue), heterozygote (green), and low-sensitivity (red) classes. The shading indicates the bootstrap 95% confidence intervals. **b**, The peak density of dsQTLs is very tightly focused around the target DHS window. **c**, Total fraction of *cis*-dsQTLs that fall into different categories of distance from the target window (*x* axis) and different annotations (*y* axis). The total area of each rectangle is proportional to the estimated number of dsQTLs in that category. **d**, Box plot showing distribution of position weight matrix (PWM) score differences between high-sensitivity and low-sensitivity dsQTL alleles, respectively. Notches indicate 95% confidence intervals for

binding ($P < 10^{-16}$; Fig. 2e), indicating that dsQTLs are strong predictors of changes in occupancy by a range of DNA-binding proteins.

Given that dsQTLs produce sequence-specific changes in chromatin accessibility and, frequently, changes in transcription factor binding, we speculated that a fraction of the dsQTL variants might also affect expression levels of nearby genes. We examined this by testing for associations between the most significant variant at each of the dsQTLs detected by using the 2 kb window size and expression

median. **e**, The *x* axis shows the fraction of sequence reads predicted to carry the major allele based on the DNase I genotype means; the *y* axis shows the observed fraction in ChIP-seq data. The lines show the regression fits for each factor separately; the numbers in the key show the fraction of sites that are in a concordant direction for each factor. CTCF, CCCTC binding factor; BATF, basic leucine zipper transcription factor; BCL11A, B-cell CLL/lymphoma 11A zinc-finger protein; EBF, early B-cell factor 1; IRF4, interferon regulatory factor 4; POU2F2, POU class 2 homeobox 2; PU1, proviral integration oncogene spi1; SP1, Sp1 transcription factor; NF-κB, nuclear factor of κ light polypeptide gene enhancer in B-cells 1.

levels of nearby genes (that is, genes with transcription start sites (TSSs) within 100 kb) estimated by sequencing RNA from the same cell lines[8]. Using this approach, we found that 16% of dsQTL SNPs were also significantly associated with variation in expression levels of at least one nearby gene (FDR = 10%). This represents a huge enrichment over random expectation (450-fold, $P \ll 10^{-16}$; Fig. 3). One example of a joint dsQTL–eQTL is illustrated in Fig. 3a, in which a SNP disrupts an ISRE located in the first intron of the *SLFN5* gene,
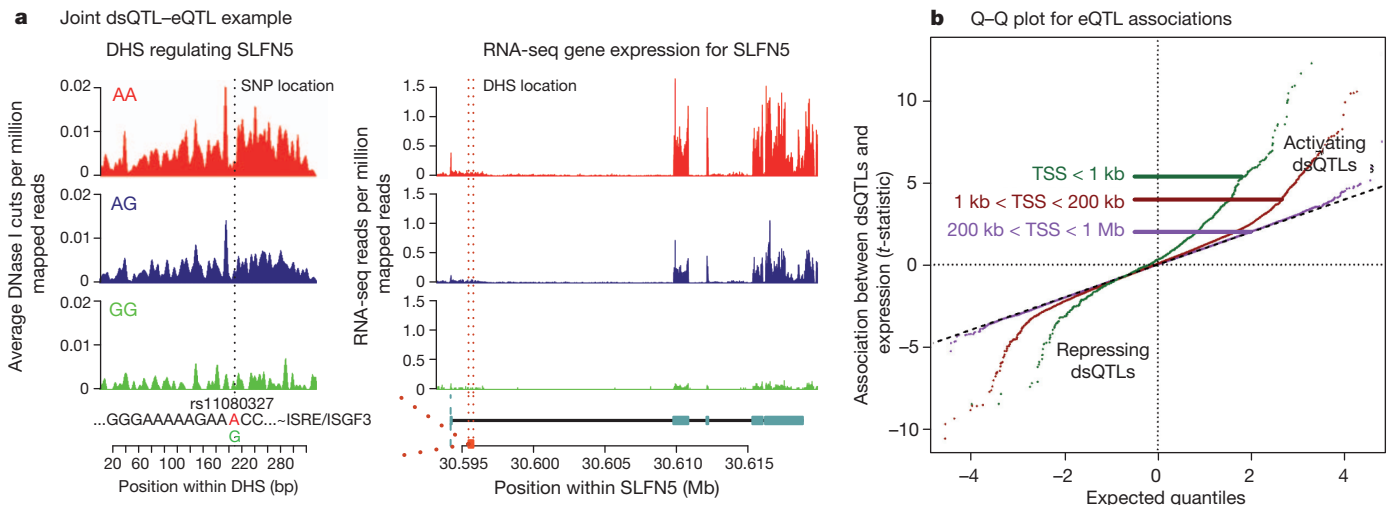
a Joint dsQTL–eQTL example

b Q–Q plot for eQTL associations

Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene *SLFN5*. The SNP disrupts an interferon-sensitive response element, thereby changing local chromatin accessibility within the first intron of *SLFN5*. Expression of *SLFN5* has been shown to be inducible by interferon α in melanoma cell lines. DNase-seq (left) and RNA-seq (right) measurements from DNase-seq and RNA-seq are plotted, stratified by genotype at the putative causal SNP. **b**, Q–Q plot of the *t*-statistic for association with gene expression changes (eQTL) of dsQTL SNPs. The sign of the eQTL *t*-statistic is with respect to the genotype that increases DNase sensitivity.

leading to both a strong dsQTL and an eQTL for *SLFN5*. Conversely, out of 1,271 eQTLs detected by using RNA-seq data from these cell lines[8], 23% of the most significant SNPs were also dsQTLs (FDR = 10%). Using the method in ref. 24 for estimating the proportion of tests in which the null hypothesis is false (while accounting for incomplete power), we estimate that 55% of the most significant eQTL SNPs are also dsQTLs and that 39% of the dsQTLs are also eQTLs. dsQTLs are therefore a major mechanism by which genetic variation may affect gene expression levels.

We observed that for most (70%) of the joint dsQTL–eQTLs, the allele that was associated with increased chromatin accessibility was also associated with increased gene expression levels (Fig. 3b). Because higher DNase I sensitivity generally correlates with higher transcription factor occupancy, this suggests that transcription factors that are bound to DHSs usually act as enhancers. CRE-box and ETS-box were the most enriched motifs among repressors and enhancers, respectively. The dsQTLs that were also eQTLs (FDR = 10%) were highly enriched around the TSSs of the target genes: for 23% of the joint dsQTL–eQTLs, the associated DHS was within 1 kb of the TSS, and for 39% it was within 10 kb (Fig. 4a). This is consistent with previous work showing strong clustering of eQTLs around TSSs[19,25,26]. Nonetheless, there was a significant signal of long-range regulation as far as 100 kb. In addition, 14% of the joint dsQTL–eQTLs were significant eQTLs for two or more genes, suggesting that some regulatory regions affect more than one gene.

We sought to identify additional factors that might influence whether a dsQTL regulates gene expression of nearby genes, while controlling for the very strong effect of distance from TSS (Fig. 4b). We observed that a dsQTL was more likely to be an eQTL for the gene with the nearest TSS (1.6-fold, $P = 3 \times 10^{-4}$) and was more likely to be an eQTL if it was located within the transcribed region of the gene (2.7-fold, $P = 2 \times 10^{-9}$). Further, a dsQTL was 2.6-fold more likely to be an eQTL if it was associated with a DHS that overlapped a DNA methylation QTL[27] ($P = 4 \times 10^{-4}$), and showed a 2.4-fold increase if the associated DHS overlapped a RNA polymerase II ChIP-seq peak[10] ($P = 4 \times 10^{-4}$). Conversely, a dsQTL was significantly less likely to be an eQTL for a gene if an active binding site for the insulator protein CTCF[17] lay between the dsQTL and the gene's TSS (2.4-fold decrease, $P = 10^{-12}$). Finally, the presence of the enhancer mark P300 (from ENCODE ChIP-seq data[28]) in the dsQTL window increased the probability that a distal dsQTL (TSS > 1.5 kb) was an eQTL (1.7-fold, $P = 10^{-5}$).

We have shown here that common genetic variants affect chromatin accessibility at thousands of hypersensitive regions across the human genome. The putative causal variants most often lie within or very near the hypersensitive regions, and frequently act by changing the binding affinity of transcription factors. Mapping of dsQTLs provides
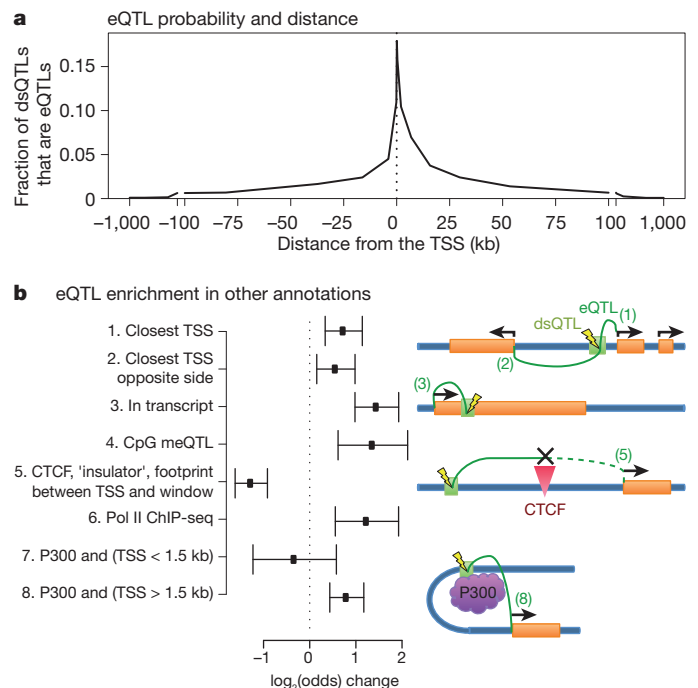
a eQTL probability and distance

b eQTL enrichment in other annotations

Figure 4 | Relationship between dsQTLs and eQTLs. a, Most joint dsQTL–eQTLs lie close to the gene TSS. **b**, Effect of various factors on the log odds that a given dsQTL is also an eQTL, while controlling for the strong distance relationship observed in **a**. In annotations (1) and (2) we do not consider the direction of transcription. In annotations (6–8) ChIP-seq is measured on the dsQTL window. In annotations (4) and (6), 'meQTL' refers to a dsQTL that is also associated with methylation levels of a nearby CpG site[27] and 'Pol II' refers to the presence of an RNA polymerase II ChIP-seq peak overlapping the DHS associated with the dsQTL[23]. One of the most significant annotations in delineating the regulatory regions is defined by the presence of the CTCF insulator element, which decreases 2.4-fold the probability that a dsQTL is an eQTL. Error bars represent 95% confidence intervals.

a powerful tool for detecting potentially functional changes in a variety of different types of regulatory element, and roughly 50% of eQTLs are also dsQTLs. Furthermore, analysis of significantly associated SNPs from genome-wide association studies additionally implicates some of these dsQTLs as potentially underlying a variety of genome-wide association study hits (Supplementary Information). Changes in chromatin accessibility may be a major mechanism linking genetic variation to changes in gene regulation and, ultimately, organismal phenotypes.

METHODS SUMMARY

DNase-seq libraries were created as described previously[29], with small modifications. Each library was sequenced on at least two lanes of an Illumina GAIIx. Resulting 20-bp sequencing reads were mapped to the human genome sequence (hg18) using an algorithm that we designed specifically to eliminate mappability biases between sequence variants. We divided the genome into 100-bp windows and selected the top 5% in terms of total DNase I sensitivity. DNase I sensitivity for each individual in each window was normalized by the total number of mapped reads for that individual. For QTL mapping, the data were further rescaled within and across individuals, and we adjusted the data for an observed individual $\times$ GC interaction, as well as for the top four principal components of the DNase I sensitivity matrix. Genotypes for all available SNPs and indels were obtained from HapMap and 1,000 Genomes data and imputed where necessary[6,7,30]. We performed DNase-seq association mapping by regressing the adjusted sensitivity in each window against the genotypes at variants in a 40-kb region centred on each DHS. As validation, we used our DNase-seq reads as well as ChIP-seq reads and DNase-seq reads from ENCODE to confirm that allele-specific reads spanning heterozygous sites at dsQTLs were consistent with the association analysis. We also used RNA-seq data from the same cell lines[8] to study the links between dsQTLs and eQTLs. Finally, we explored the properties of dsQTLs that made them more or less likely to influence gene expression by fitting a logistic model on all dsQTLs, where the eQTL status of each dsQTL–eQTL test was modelled as a function of distance from the TSS and a variety of other annotations. For full details of all methods see Supplementary Information.

1. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
2. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437,** 1365–1369 (2005).
3. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).
4. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6,** e1000895 (2010).
5. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
6. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
7. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
8. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
9. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature Genet.* **42,** 255–259 (2010).
10. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328,** 232–235 (2010).
11. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (2010).
12. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464,** 1187–1191 (2010).
13. Galas, D. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5,** 3157–3170 (1978).
14. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
15. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6,** 283–289 (2009).
16. Boyle, A. P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).
17. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–455 (2011).
18. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
19. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4,** e1000214 (2008).
20. Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl Acad. Sci. USA* **107,** 22534–22539 (2010).
21. Wasson, T. & Hartemink, A. J. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* **19,** 2101–2112 (2009).
22. Raveh-Sadka, T., Levo, M. & Segal, E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* **19,** 1480–1496 (2009).
23. Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).
24. Storey, J. D., Taylor, J. E. & Siegmund, D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc., B* **66,** 187–205 (2004).
25. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39,** 1202–1207 (2007).
26. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature Genet.* **39,** 1217–1224 (2007).
27. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12,** R10 (2011).
28. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
29. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protocols.* doi:10.1101/pdb.prot5384 (2010).
30. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* **4,** e1000279 (2008).