HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2012 November 19.

Published in final edited form as:

Nature. ; 482(7385): 390–394. doi:10.1038/nature10808.

DNaseI 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,*], and **Jonathan K. Pritchard**[1,4,*]

[1]Department of Human Genetics, University of Chicago

[2]Committee on Genetics, Genomics and Systems Biology, University of Chicago

[3]BioMiningLabs, Lyon, France

[4]Howard Hughes Medical Institute, University of Chicago

[5]Institute for Genome Sciences and Policy, Duke University

[6]Departments of Pediatrics, Division of Medical Genetics, Duke University

[7]Department of Statistics, University of Chicago

Abstract

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. To address this gap, we used DNaseI sequencing to measure chromatin accessibility in 70 Yoruba lymphoblastoid cell lines (LCLs), 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-seq reads, which allowed us to produce genome-wide maps of chromatin accessibility for each individual. We identified 9,595 locations at which DNase-seq read depth correlates significantly with genotype at a nearby SNP or indel (FDR=10%). We call such variants "DNaseI 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, (namely, these loci are also

[*]To whom correspondence should be addressed: jdegner@uchicago.edu, athma@uchicago.edu, rpique@uchicago.edu, gilad@uchicago.edu, pritch@uchicago.edu.

[†]These authors contributed equally.

Author Contributions A.A.P led the data collection with assistance from S.D.L, K.M, and N.L. The data analysis was performed jointly by J.F.D and R.P.R., with contributions from A.A.P., J.B.V., D.J.G, and J.K.Pi. G.E.C and M.S. provided technical assistance and discussion of methods and results. The manuscript was written by J.F.D., A.A.P., R.P.R., Y.G., and J.K.Pr. The project was jointly supervised by Y.G. and J.K.Pr.

Author information All raw data and tables of all dsQTL are available under GEO accession number GSE31388.

classified as eQTLs). Conversely, we estimate that as many as 55% of eQTL SNPs 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 important contributors to phenotypic variation[1-5]. However, the underlying regulatory mechanisms by which eQTLs impact gene expression remain poorly understood. One mechanism that may be important is when the alternative alleles at a particular 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 high-depth DNaseI-sequencing (DNase-seq) in a panel of 70 individuals and find that indeed a large fraction of eQTLs are likely caused by this type of mechanism.

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

We collected DNase-seq data for 70 HapMap Yoruba lymphoblastoid cell lines (LCLs) for which gene expression data and genome-wide genotypes were already available[6-8]. We obtained an average of 39M 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 DNaseI cuts within individual hypersensitive sites at extremely high resolution. As expected, the DHSs coincide to a great extent with previously annotated regulatory regions, and DNaseI sensitivity is positively correlated with the expression levels of nearby genes (Figures S6&7). Overall, the locations of hypersensitive sites are highly correlated across individuals (Supplementary Information)[11].

We tested for genetic variants that impact local chromatin accessibility. To do this, we divided the genome into non-overlapping 100 bp windows, and then focused our analysis on the 5% of windows with the highest DNaseI 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 estimates the level of chromatin accessibility. We then tested for association between individual-specific DNaseI sensitivity in each window and genotypes of all SNPs and indels in a cis candidate region of 40kb centered 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 FDR=10% (corresponding to 8,902 distinct DHSs, once we combine adjacent windows whose hypersensitivity data is associated

with the same SNP or indel; Figure 1A). We refer to these 8,902 loci as "DNaseI sensitivity QTLs", or dsQTLs. We additionally considered a much smaller cis-candidate region of just 2kb around each target window, and found that the majority of the dsQTLs are 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 find only weak evidence of trans-acting dsQTLs, likely because our experiment is 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 correlates well with the dsQTL effect sizes (Figure 1B, correlation coefficient r=0.72, $P<<10^{-16}$).

We observed that dsQTLs typically affect chromatin accessibility for about 200-300 bp (Figure 2A). Of the DHSs affected by dsQTLs, 77% lie in chromatin regions predicted by Ernst *et al*. to be functional in LCLs[18]: 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 generate dsQTLs, using a Bayesian hierarchical model that accounts for the uncertainty in which sites are causal[19](Supplementary Information). This model obtains unbiased estimates of the average properties of causal sites even though, because of linkage disequilibrium, it is typically uncertain which site is causal for any individual dsQTL (Supplementary Information). As shown in Figures 2B&C, most dsQTLs are generated by variants that are close to the target window. We estimate that 56% of the dsQTLs are due to variants that lie within the same DHSs and that 67% lie within 1 kb of the target window. dsQTLs that lie more than 1kb from the target window are themselves significantly enriched in non-adjacent DHS windows (2.4-fold compared to matched random SNPs), and are often associated with changes in sensitivity in multiple non-adjacent DHS windows (Figure S15).

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 DNaseI cut rates. Consistent with this model, an aggregated plot of DNaseI sensitivity at dsQTLs shows a distinct drop in chromatin accessibility around putatively causal SNPs that is 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 using 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 ($P<<10^{-16}$), controlling for the overall enrichment within DHSs. Additionally, the allele associated with a higher score of the position weight matrix (PWM) is typically associated with higher chromatin accessibility ($P<<10^{-16}$), consistent with the expectation that higher transcription factor binding affinity leads to more open chromatin (Figure 2D). Of the dsQTLs that fall within DNase-seq footprints tied to

specific transcription factors motifs (using CENTIPEDE[17]), CTCF, CRE and ISRE are the most enriched while MEF2 is 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 are associated with increased DNaseI sensitivity are highly associated with increased transcription factor binding ($P<10^{-16}$; Figure 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 hypothesized 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 using the 2kb window size and expression levels of nearby genes (i.e., genes with transcription start sites, TSSs, within 100kb) estimated by sequencing RNA from the same cell lines[8]. Using this approach, we found that 16% of dsQTL SNPs are 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, $P<<10^{-16}$; Figure 3). One example of a joint dsQTL-eQTL is illustrated in Figure 3A, in which a SNP disrupts an interferon-sensitive response element (ISRE) located in the first intron of the *SLFN5* gene, leading to both a strong dsQTL and an eQTL for *SLFN5*. Conversely, out of 1,271 eQTLs detected using RNA-seq data from these cell lines[8], 23% of the most significant SNPs are also dsQTLs (FDR=10%). Using the method of Storey *et al.*[24] for estimating the proportion of tests where 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. Hence dsQTLs are a major mechanism by which genetic variation may impact gene expression levels.

We observed that for most (70%) of the joint dsQTL-eQTLs, the allele that is associated with increased chromatin accessibility is also associated with increased gene expression levels (Figure 3B). Since higher DNaseI-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 GABP/ETS-box were the most enriched motifs among repressors and enhancers respectively. The dsQTLs that are also eQTLs (FDR=10%) are highly enriched around the TSSs of the target genes: for 23% of the joint dsQTL-eQTLs, the associated DHS is within 1kb, and for 39% it is within 10kb of the TSS (Figure 4A). This is consistent with previous work showing strong clustering of eQTLs around TSSs[19,25-26]. Nonetheless, there is a significant signal of long-range regulation as far as 100kb. Additionally, 14% of the joint dsQTL-eQTLs are significant eQTLs for two or more genes, suggesting that some regulatory regions affect more than one gene.

We sought to identify additional factors that may influence whether a dsQTL regulates gene expression of nearby genes, while controlling for the very strong effect of distance from TSS (Figure 4B). We observed that a dsQTL is more likely to be an eQTL for the gene with the nearest TSS (1.6-fold, $P=3\times10^{-4}$) and is more likely to be an eQTL if it is located within the

transcribed region of the gene (2.7-fold, $P=2\times10^{-9}$). Further, a dsQTL is 2.6 fold more likely to be an eQTL if it is associated with a DHS that overlaps a DNA methylation QTL[27] ($P=4\times10^{-4}$), and shows a 2.4-fold increase if the associated DHS overlaps a PolII ChIP-seq peak[10] ($P=4\times10^{-4}$). Conversely, a dsQTL is significantly less likely to be an eQTL for a gene if an active binding site for the insulator protein CTCF[17] lies between the dsQTL and the gene's TSS (2.4-fold decrease, $P=1\times10^{-12}$). Finally, the presence of the enhancer mark P300 (from ENCODE ChIP-seq data[28]) in the dsQTL window increases the probability that a distal dsQTL (TSS>1.5kb) is an eQTL (1.7-fold, $P=1\times10^{-5}$).

In summary, we have shown that common genetic variants impact 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 DNaseI sensitivity QTLs provides a powerful tool for detecting potentially functional changes in a variety of different types of regulatory elements, 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 GWAS 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 previously described[29], with small modifications. Each library was sequenced on at least two lanes of an Illumina GAIIx. Resulting 20bp 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 DNaseI sensitivity. DNaseI 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 × GC interaction, as well as for the top four principal components of the DNaseI sensitivity matrix. Genotypes for all available SNPs and indels were obtained from HapMap and 1000 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 centered 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 are 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 make 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 is modeled as a function of distance from the TSS and a variety of other annotations. For full details of all methods see the Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by grants from the National Institutes of Health to YG (HG006123) and JKP (MH090951), by the Howard Hughes Medical Institute, by the Chicago Fellows Program (RPR), by the American Heart Association (AAP), and by the NIH Genetics and Regulation Training grant (AAP and JFD). We thank members of the Pritchard, Przeworski, Stephens and Gilad labs as well as three anonymous reviewers for many helpful comments or discussions, and the ENCODE Project for publicly available ChIP-seq data.

References

[1]. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296:752–5. [PubMed: 11923494]

[2]. Cheung VG, et al. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005; 437:1365–9. [PubMed: 16251966]

[3]. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]

[4]. Nica AC, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010; 6:e1000895. [PubMed: 20369022]

[5]. Allen, H. Lango, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–8. [PubMed: 20881960]

[6]. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–61. [PubMed: 17943122]

[7]. Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]

[8]. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010

[9]. Gaulton KJ, et al. A map of open chromatin in human pancreatic islets. Nat Genet. 2010; 42:255–9. [PubMed: 20118932]

[10]. Kasowski M, et al. Variation in transcription factor binding among humans. Science. 2010

[11]. McDaniell R, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010

[12]. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. Genetic analysis of variation in transcription factor binding in yeast. Nature. 2010; 464:1187–91. [PubMed: 20237471]

[13]. Galas D, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978; 5:3157–70. [PubMed: 212715]

[14]. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–22. [PubMed: 18243105]

[15]. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009; 6:283–9. [PubMed: 19305407]

[16]. Boyle AP, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011; 21:456–64. [PubMed: 21106903]

[17]. Pique-Regi R, et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011; 21:447–55. [PubMed: 21106904]

[18]. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011

[19]. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]

[20]. Mirny LA. Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A. 2010; 107:22534–9. [PubMed: 21149679]

[21]. Wasson T, Hartemink AJ. An ensemble model of competitive multi-factor binding of the genome. Genome Res. 2009; 19:2101–12. [PubMed: 19720867]

[22]. Raveh-Sadka T, Levo M, Segal E. Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res. 2009; 19:1480–96. [PubMed: 19451592]

[23]. Myers RM, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9:e1001046. [PubMed: 21526222]

[24]. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society, Series B. 2004; 66:187–205.

[25]. Dixon AL, et al. A genome-wide association study of global gene expression. Nat Genet. 2007; 39:1202–7. [PubMed: 17873877]

[26]. Stranger BE, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–24. [PubMed: 17873874]

[27]. Bell JT, et al. DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. Genome Biol. 2011; 12:R10. [PubMed: 21251332]

[28]. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–8. [PubMed: 19212405]

[29]. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010; 2010 pdb.prot5384.

[30]. Guan Y, Stephens M. Practical issues in imputation-based association mapping. PLoS Genet. 2008; 4:e1000279. [PubMed: 19057666]

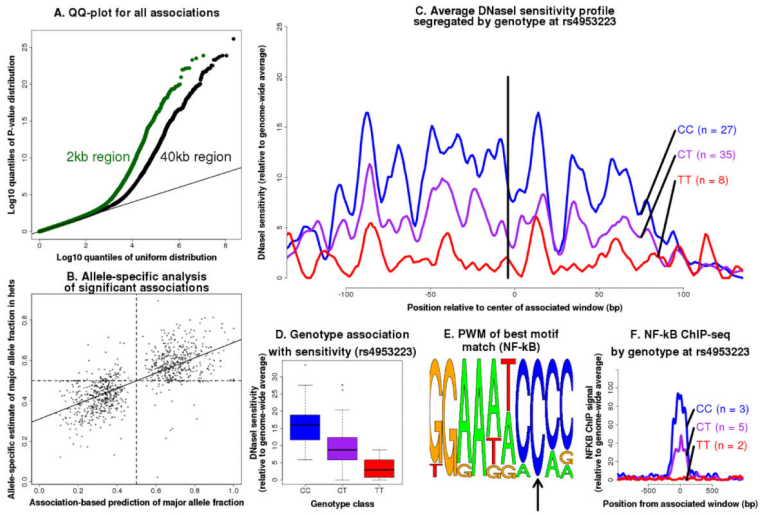

Figure 1. Genome-wide identification of dsQTLs and a typical example

(A) QQ-plots for all tests of association between DNaseI cut rates in 100bp windows, and variants within 2kb (green) and 40kb (black) regions centered 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)** Boxplot showing that rs4953223 is strongly associated with local chromatin accessibility ($P=3\times10^{-13}$). **(E)** The T allele, which is associated with low DNaseI sensitivity, disrupts the binding motif of a previously identified NF-κB binding site at this location[14] **(F)**. NF-κB ChIP-seq data from 10 individuals[7] indicates a strong effect of this SNP on NF-κB binding.

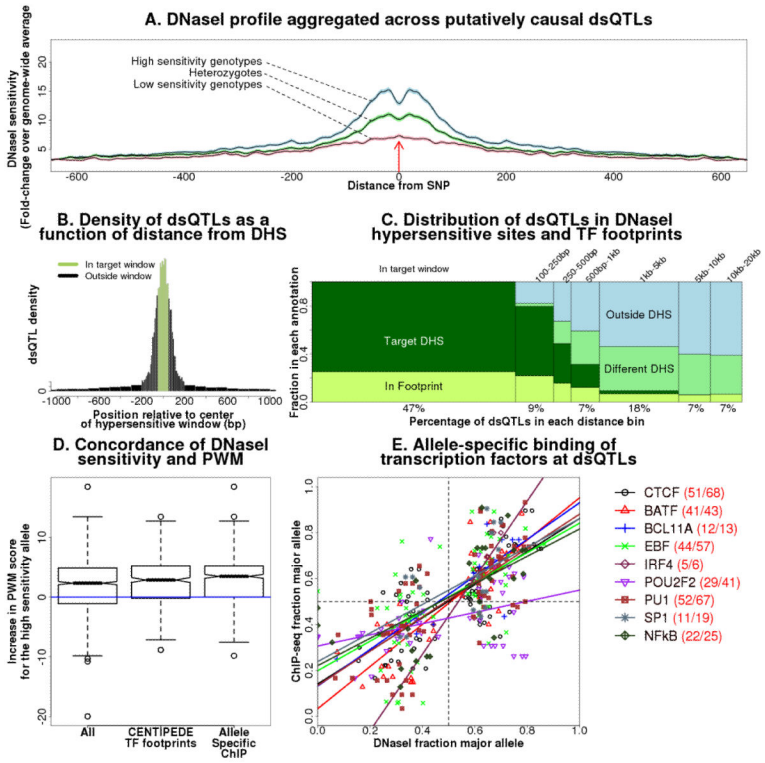

Figure 2. Properties of dsQTLs

(A) Aggregated plot of DNaseI-sensitivity for high-confidence dsQTLs that lie within the target DHS. Individuals were assigned 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)** Boxplot showing distribution of PWM score differences between high sensitivity and low sensitivity dsQTL alleles, respectively. Notches indicate 95% CI for median. **(E)** The x-axis shows the fraction of sequence reads predicted to carry the major allele based on the DNaseI 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 legend show the fraction of sites that are in a concordant direction for each factor.

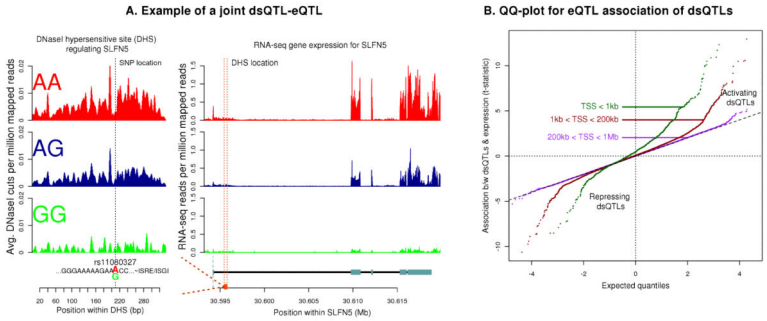

Figure 3. Relationship between dsQTLs and eQTLs

(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 column) and RNA-seq (right column) measurements from DNase-seq and RNA-seq are plotted, stratified by genotype at the putative causal SNP. **(B)** QQ-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.

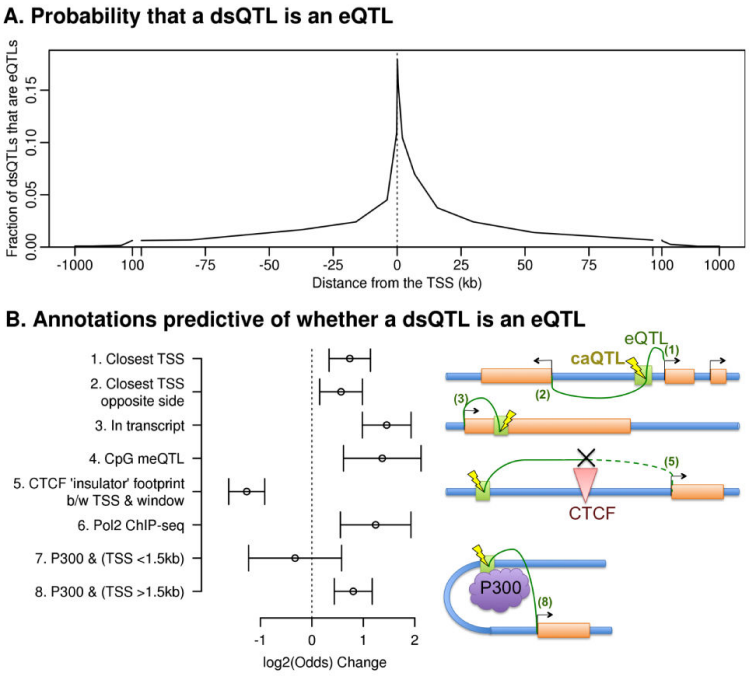

Figure 4. Relationship between dsQTLs and eQTLs

(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 panel 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. One of the most significant annotations in delineating the regulatory regions is defined by the presence of the CTCF insulator element, which reduces the probability that a dsQTL is an eQTL by 2.4-fold. Error bars represent 95% confidence intervals