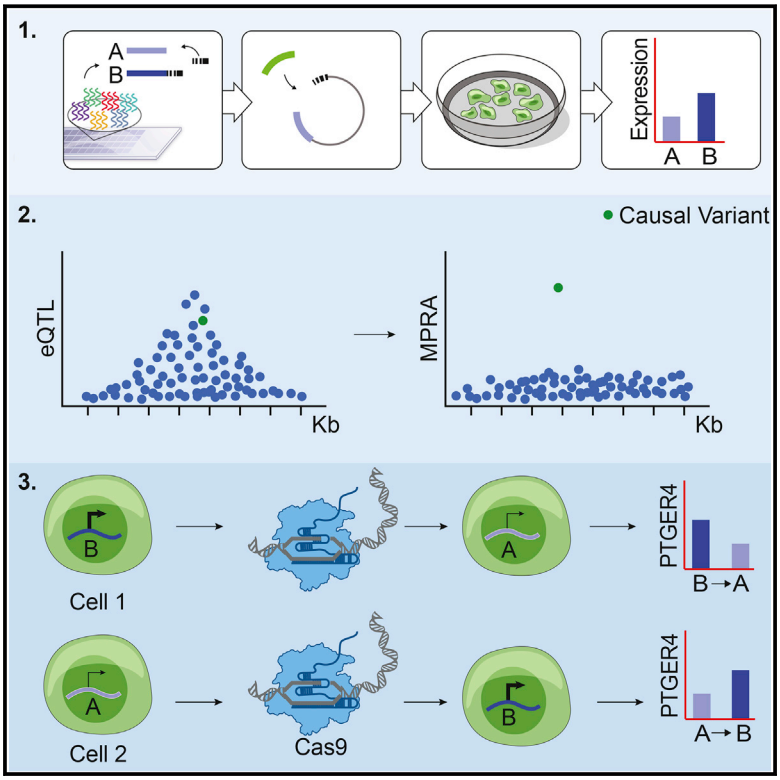
Resource

Cell

Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay

Graphical Abstract



Authors

Ryan Tewhey, Dylan Kotliar, Daniel S. Park, ..., Eric S. Lander, Stephen F. Schaffner, Pardis C. Sabeti

Correspondence

rtewhey@broadinstitute.org (R.T.), pardis@broadinstitute.org (P.C.S.)

In Brief

A massively parallel reporter assay analyzes thousands of human polymorphisms to identify alleles that impact gene expression, providing a tool with which to move from disease-associated GWAS hits to the identification of functional variants.

Highlights

- A new version of MPRA with greater throughput and sensitivity
- Evaluation of 32,373 variants associated with eQTLs in lymphoblastoid cell lines
- 842 variants showed differential gene expression between alleles
- Use of CRISPR/cas9 to identify a distal eQTL causal allele for *PTGER4*

Accession Numbers

GSE75661

Tewhey et al., 2016, Cell *165*, 1519–1529
June 2, 2016
http://dx.doi.org/10.1016/j.cell.2016.04.027

CellPress

Resource

Cell

Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay

Ryan Tewhey,[1,2,*] Dylan Kotliar,[1,2] Daniel S. Park,[2] Brandon Liu,[2] Sarah Winnicki,[2] Steven K. Reilly,[1,2] Kristian G. Andersen,[1,2] Tarjei S. Mikkelsen,[2] Eric S. Lander,[2] Stephen F. Schaffner,[2] and Pardis C. Sabeti[1,2,*]

[1]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

[2]Broad Institute, Cambridge, MA 02142, USA

*Correspondence: rtewhey@broadinstitute.org (R.T.), pardis@broadinstitute.org (P.C.S.)

http://dx.doi.org/10.1016/j.cell.2016.04.027

SUMMARY

Although studies have identified hundreds of loci associated with human traits and diseases, pinpointing causal alleles remains difficult, particularly for non-coding variants. To address this challenge, we adapted the massively parallel reporter assay (MPRA) to identify variants that directly modulate gene expression. We applied it to 32,373 variants from 3,642 cis-expression quantitative trait loci and control regions. Detection by MPRA was strongly correlated with measures of regulatory function. We demonstrate MPRA's capabilities for pinpointing causal alleles, using it to identify 842 variants showing differential expression between alleles, including 53 well-annotated variants associated with diseases and traits. We investigated one in detail, a risk allele for ankylosing spondylitis, and provide direct evidence of a non-coding variant that alters expression of the prostaglandin EP_4 receptor. These results create a resource of concrete leads and illustrate the promise of this approach for comprehensively interrogating how non-coding polymorphism shapes human biology.

INTRODUCTION

The genomic era has enormously increased our knowledge of human genetic variation, but our understanding of the functional consequences of that variation has not kept pace (Cooper and Shendure, 2011). Although genome-wide association studies (GWAS) and whole-genome scans for natural selection have identified numerous loci linked to human traits and diseases, correlation between nearby polymorphisms (linkage disequilibrium or LD) within individual associations often leaves dozens to hundreds of potential causal variants to be interrogated (Grossman et al., 2013; Schaub et al., 2012). Mounting evidence suggests that at the majority of these loci, the causal variant(s) is a non-coding regulatory change rather than an amino acid substitution (Farh et al., 2015; Maurano et al., 2012). Indeed, regulatory changes drive some of the best understood examples of phenotypic diversity and adaptive evolution (Claussnitzer et al., 2015; Musunuru et al., 2010; Tishkoff et al., 2007). Therefore, it is critical that we be able to test whether a variant affects gene regulation.

Current approaches for measuring a variant's effect on gene expression fall into two categories, each with its own limitation. Indirect methods, such as whole-genome epigenetic assays, can only identify the broader regulatory state of a region, not necessarily the effect of a particular variant (Andersson et al., 2014; ENCODE Project Consortium, 2012; Kasowski et al., 2013; McVicker et al., 2013). Direct methods, ones that measure the impact of individual alleles in an episomal or native context on gene expression, are currently low throughput and require substantial resources for comprehensive evaluation of a region.

We adopted the massively parallel reporter assay (MPRA) as a solution and modified it so that we could carry out large-scale, sensitive, and direct testing of potential regulatory variants. This assay is based upon the well-established reporter gene assay, in which a vector containing a reporter gene (e.g., luciferase or green fluorescent protein [GFP]), a promoter, and a potential regulatory sequence is inserted into a plasmid, which is transfected into a cell; sequences that regulate gene expression then alter the amount of luciferase/GFP expressed (Arnold et al., 2013; Melnikov et al., 2012; Ow et al., 1986; Patwardhan et al., 2012; Vockley et al., 2015; Kwasnieski et al., 2014). Through the use of unique barcodes in the 3′ UTR of the reporter to differentiate expression of individual oligos, MPRA can test many different sequences simultaneously, and it has been shown to reproducibly detect segments of the genome that change expression levels (Kheradpour et al., 2013; Mogno et al., 2013). We aimed to incorporate single-nucleotide and small-insertion/deletion polymorphisms (referred to below as single-nucleotide variants or SNVs) into these assays to see whether we could detect subtle differences in how each allele drives expression. Because we used only a minimal promoter, with very low baseline expression, in this iteration of the assay, we intended it primarily as a test of regulatory sequence that increases (i.e., enhancers and promoters), rather than decreases, expression; the latter will be difficult to detect because baseline expression is already low.

Ideally, one would test the assay for sensitivity and specificity by applying it to a set of "gold-standard" variants previously identified as expression quantitative trait loci (eQTLs) that act on enhancer and promoter elements. However, there is a dearth of such known variants. As the best available alternative, we studied

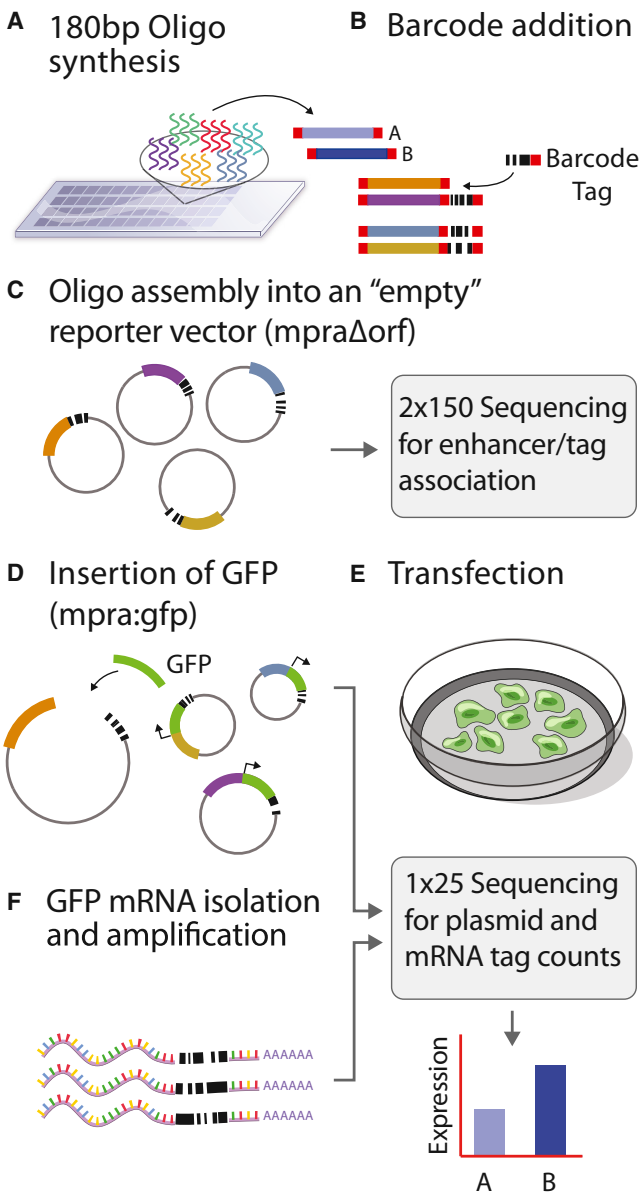


Figure 1. Overview of the MPRA Workflow

(A) Oligos are synthesized as 180 mer followed by cleavage off of the array. (B–F) The single-stranded DNA is amplified, barcoded, and converted to double-stranded DNA by emulsion PCR (B), which is then cloned into a reporter vector that has had the reporter gene removed to create the mpra:Δorf library (C). The plasmid library is linearized between the barcode and oligo sequences by restriction digest and a minimal promoter, and GFP open reading frame is inserted by gibson assembly to create the final mpra:gfp library (D), which is used for transfection into the desired cell type (E). RNA is harvested from the transfected cells, mRNA is captured and sequenced (F), and barcode counts are compared to the count estimates from the sequencing of the mpra:orf library (D).

a set of thousands of *candidate* eQTL variants in regions associated with differential gene expression in the population. There are important considerations in interpreting the results of such a test. First, we test *multiple* candidate variants (sometimes dozens) within each eQTL region that are in LD with one another. We generally expect that at most *one* of these variants will be causal, and sometimes the true causal variant will not appear in the set because the degree of LD falls below the cutoff used for inclusion. For these reasons only a minority of the variants tested are expected to give a signal in the MPRA assay. Conservative estimates suggest that in 34%–41% of eQTL peaks (dependent on the population tested), the causal variant will be the top-associated allele. Second, only a subset of the true variants responsible for eQTLs (23%–64%, according to recent estimates) will act on enhancers or promoters, which are the functional classes that will be detected in the MPRA assay (Farh et al., 2015; Gymrek et al., 2015; Lappalainen et al., 2013). Variants that work by other processes, such as microsatellites or post-transcriptional regulation, would not be expected to score.

Because we expect only a minority (8%–26% based on the prior estimates) of the variants in our test set to score by MPRA, we must evaluate the assay by comparison with the performance on control sets of common polymorphisms. We used two control sets: the first chosen randomly from the genome, and the second containing variants near but not associated with eQTL variants.

We estimated the specificity, sensitivity, and reproducibility of MPRA to localize causal alleles within large genomic loci linked to variation in gene expression. In addition, we comprehensively interrogated GWAS loci that overlap with eQTLs to identify and characterize potential regulatory variants underlying human diseases and traits.

RESULTS

Adapting MPRA to Test ~30k Candidate Variants

We modified MPRA to increase its throughput while also improving its reproducibility and sensitivity (Figure 1; Experimental Procedures); the latter is crucial because we aim not merely to find genetic elements that regulate genes but to detect *differences* in regulation caused by single variants within those elements. To accommodate our large library size and to increase the sensitivity of the assay, we added 20 nucleotide barcodes to the oligos by emulsion PCR and cloned the fragments to generate a library; in this manner, each oligo is represented by an average of a thousand tags within the plasmid library. Following transfection, we captured the GFP mRNA by hybridization and performed high-throughput sequencing of the UTR barcode to determine the effects of the oligos on the transcription level of the reporter gene. This new experimental approach decreased inter-experimental noise and allowed us to apply a parametric statistical framework during analysis (Supplemental Experimental Procedures).

Selection of Variants Tested

For benchmarking and discovery, we first examined nearly thirty thousand SNVs within a set of eQTLs. We identified eQTLs from the Geuvadis RNA-seq dataset of lymphoblastoid cell lines (LCLs) from individuals of European (EUR) and West African (YRI) ancestry due to the availability of both genome sequences and immortalized cell lines for these individuals (Abecasis et al., 2012; Lappalainen et al., 2013). For each of the 3,642 eQTLs, we

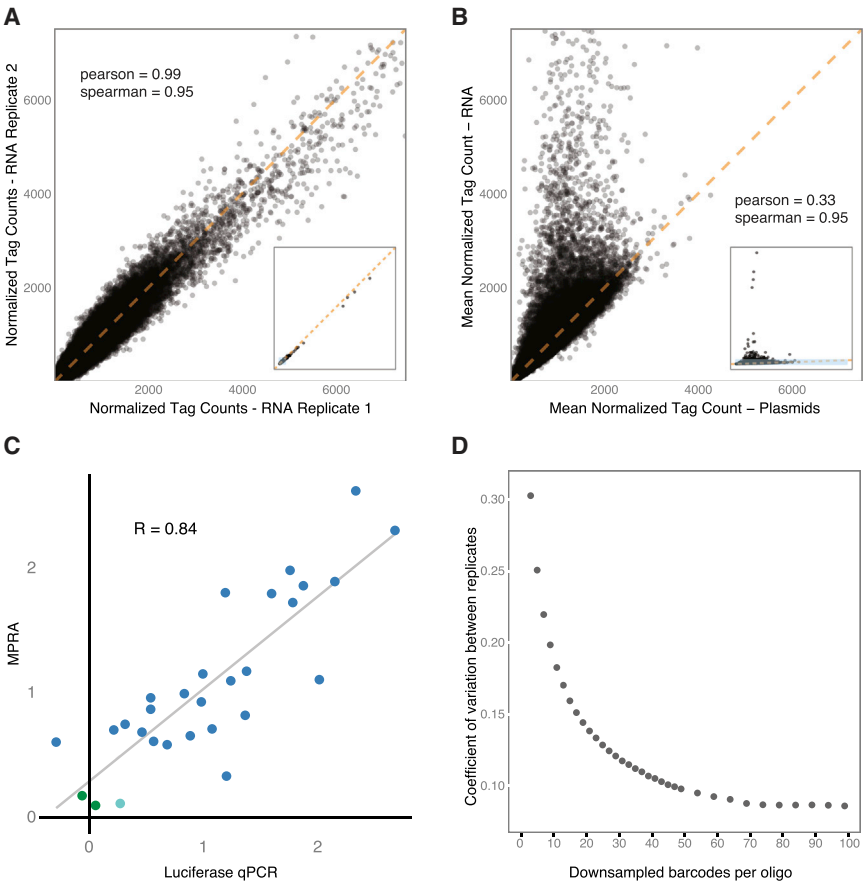


Figure 2. Experimental Reproducibility

(A and B) Correlation of normalized oligo counts between two transfection replicates of NA12878 (A). (B) Average normalized oligo counts for all five plasmid replicates compared to normalized counts for the five replicates from NA12878 RNA. Axes across all plots were kept constant with subplots added when additional data points were excluded from the main plot (A and B). Blue boxes within the inserts signify the displayed areas of the main plots.

(C) Luciferase assay validation of estimated effect sizes for individual oligos tested by MPRA. Each point represents the average of eight MPRA and four qPCR replicates. qPCR values were normalized to two non-significant sequences (green points) as determined by MPRA. Blue points: significantly expressed sequences from MPRA; cyan point: marginally significant sequence. Correlation is provided as Pearson's R.

(D) Coefficient of variation between experimental replicates as a product of the number of barcodes tagging an oligo.

designed and synthesized DNA oligonucleotides (oligos) representing the top-associated variant and all variants in perfect LD with it. This approach selects an average of 3 SNVs per eQTL peak. As noted above, we expect that (1) this set will often fail to contain the true causal variant, and (2) when the set does contain the causal variant, the other two variants will not be causal. We also included 209 eQTLs that overlapped GWAS hits for deeper investigation; for these, we tested all alleles in moderately strong LD (r^2 > = 0.9) with the lead variants. After inclusion of several smaller sets of variants, and accounting for neighboring variants and orientation of the variant when associated with multiple genes, this first 79k oligo library included 39,479 oligo pairs, originating from 29,173 unique variants (see Supplemental Experimental Procedures for a complete breakdown).

We also performed a second MPRA experiment that assayed 264 positive control variants (sites identified in the first 79k MPRA library) and 3,200 negative control variants. The negative controls included 2,700 variants chosen at random across the genome matching the minor allele frequency distribution of the larger 79k library. To select a set of negative control variants with a similar biological profile, we included 500 SNVs that were in close proximity to an eQTL peak (within 250–1000 bp) and not in LD with the lead variant. We incorporated all variants into a 7.5k oligo library. In total, across the two experiments we evaluated 85,358 oligos (42,679 reference/alternate pairs), centering the variant of interest in 150 bp of its genomic sequence.

Technical Performance of MPRA

We transfected the original 79k MPRA library into two separate lymphoblastoid cell lines (NA12878 and NA19239) from the 1000 Genomes project as well as into a hepatocarcinoma cell line (HepG2). We performed eight and five technical replicates for the lymphoblastoid and hepatocarcinoma cell lines, respectively. We observed high coverage and excellent reproducibility in the assay, capturing 98.4% of the 79K oligos tested at a depth of 20 reads or greater. Reproducibility was excellent between experimental replicates of identical cell lines, with an average correlation of 0.99 (Figures 2A, 2B, and S1), and expression values were strongly correlated with a traditional single-plexed luciferase assay for the 29 sequences we examined (R = 0.84, Figure 2C).

Each oligo in our initial 79k library was captured by an average of 73 unique barcodes per replicate during sequencing (per sample range: 34–117 barcodes), with an average total read count of 1,102 (Figure S2A). A key feature of our approach is the use of additional barcodes to reduce variability between replicates; reducing this variability is crucial for achieving the sensitivity required to detect subtle differences between alleles because detection requires distinguishing the distributions for the two alleles. To evaluate how this variability depended on the number of barcodes, we downsampled barcodes for each oligo. We observed little fluctuation in the estimated mean oligo count, as long as they were captured with greater than 20 barcodes (Figure S2B). The variance between the individual replicates, on the other hand, continued to improve throughout the range from 20 to 50 barcodes, indicating that power to detect small

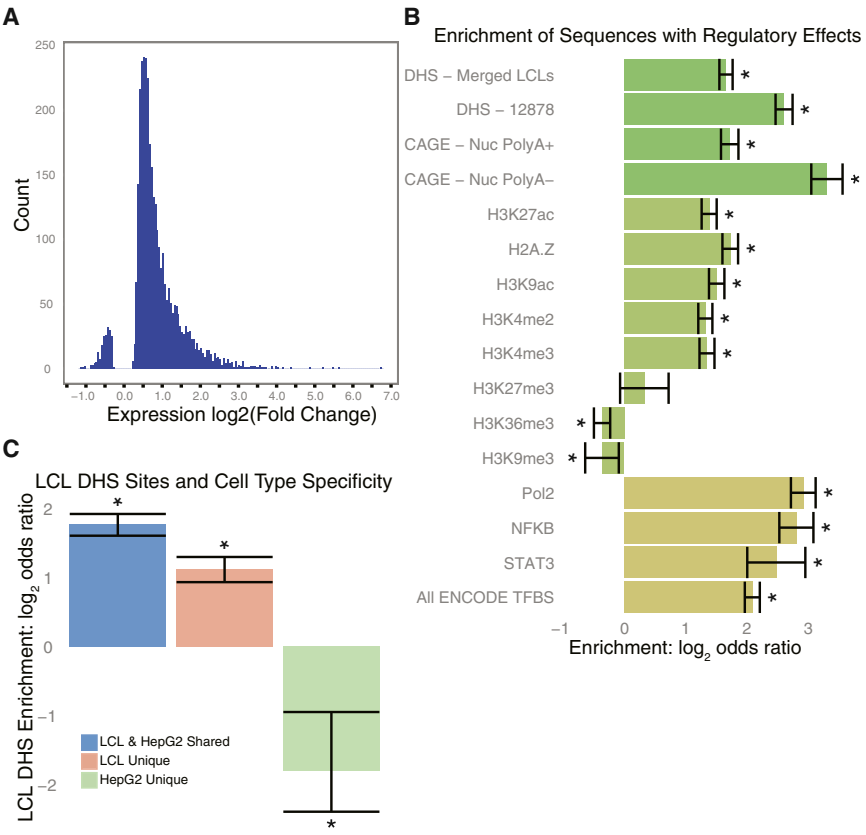


Figure 3. Validation of Expression-Modifying Sequences Discovered by MPRA

(A) Distribution of effect sizes (log_2 of the RNA/plasmid ratio) for oligos that were detected as being under- or overexpressed.

(B) Log_2(odds ratio) for the enrichment of regulatory annotations in the 3,432 MPRA active sequences within LCLs relative to non-active sequences.

(C) Log_2(odds ratio) for the enrichment in LCL DHS sites for active sequences shared between LCLs and HepG2s (blue), active in only LCLs (red), and active in only HepG2 cells (green). Asterisk: fisher's test p value < 0.01.

Error bars represent one standard deviation.

differences between oligos is substantially affected by barcode count (Figure 2D). This effect is highlighted in the second 7.5k library: its smaller size allowed us to tag each oligo with an average of 350 barcodes. This resulted in a greater sensitivity to detect weak expression changes, illustrating the impact of the number of barcodes tagging each variant and also highlighting the requirement for normalization when comparing between libraries (Figures S3A–S3D; Supplemental Experimental Procedures).

Evaluating Regulatory Activity of the Oligos

Before looking for allelic effects, we first identified the subset of sequences for which either or both variants altered the expression of the reporter. Of the 29k variants evaluated in the original assay, 12% (3,432) had an effect on the reporter for at least one of the two alleles (Table S1); these we call "active" sequences. Of these, 95% enhanced expression of the reporter (Figures 3A and S3E). Because the assay uses a weak basal promoter, it is more sensitive to increases in expression than to decreases. It is conceivable, however, that the result also reflects differences in the proportion of activating and silencing elements in the genome. We found that active sequences were shared between LCLs from different individuals more often than between different cell types (74% between NA12878 and NA19239 compared to 52% between NA12878 and HepG2). This difference in overlap is likely an underestimate, as only three replicates were performed in NA19239 compared to five in HepG2.

The active sequences were reproducible, with the effect sizes being highly correlated when we re-tested 274 active variants in the 7.5k MPRA experiment (R = 0.95) (Figure S3F).

The sequences that scored as active in the assay are significantly enriched for markers associated with regulation in the genome, including open chromatin, biochemical marks denoting active promoters and enhancers, and individual transcription factors. We first evaluated overlap with open chromatin, as identified by DNase hypersensitivity sites (DHS). We found that 43.3% of the active sequences were marked as DHS compared to only 19.4% for the non-active sequences, a 2.2-fold enrichment (odds ratio [OR] = 3.2, $p = 1.8 \times 10^{-191}$; Figure 3B). Histone marks associated with active promoters and enhancers (H3k4me3, H2az, H327ac, CAGE) were both similarly enriched, whereas marks associated with heterochromatin and/or the blocking of transcription initiation (H3k9me3 and H3k36me3) were significantly depleted, as expected (Figure 3B). The strongest enrichments were seen with individual transcription factor (TF) binding locations, with increases ranging from 3- to 39-fold for all TFs surveyed in LCLs by ENCODE, except for the repressor element Ezh2. Enrichment was cell type specific, again as expected: sequences active only in LCLs and not HepG2 cells were enriched for DHS sites unique to LCLs (OR = 2.2, $p = 6.6 \times 10^{-32}$). Similarly, sequences that were active only in HepG2 cells were depleted in LCL DHS sites (OR = 0.29, $p = 8.7 \times 10^{-8}$) compared to all other sites tested (Figure 3C).

Identifying Alleles with Differential Activity

Focusing on those sequences for which at least one allele affected the expression of the reporter, we identified those that showed *differential* expression between the reference and alternate allele ("allelic skew"). Of the 3,432 active sequences, 842 showed allelic skew; we call these "expression-modulating variants" (emVars) (Table S1). Most of the emVars exhibited modest expression differences between alleles: only 46 had more than a 2-fold change (Figures 4A and 4B). The changes were, however,

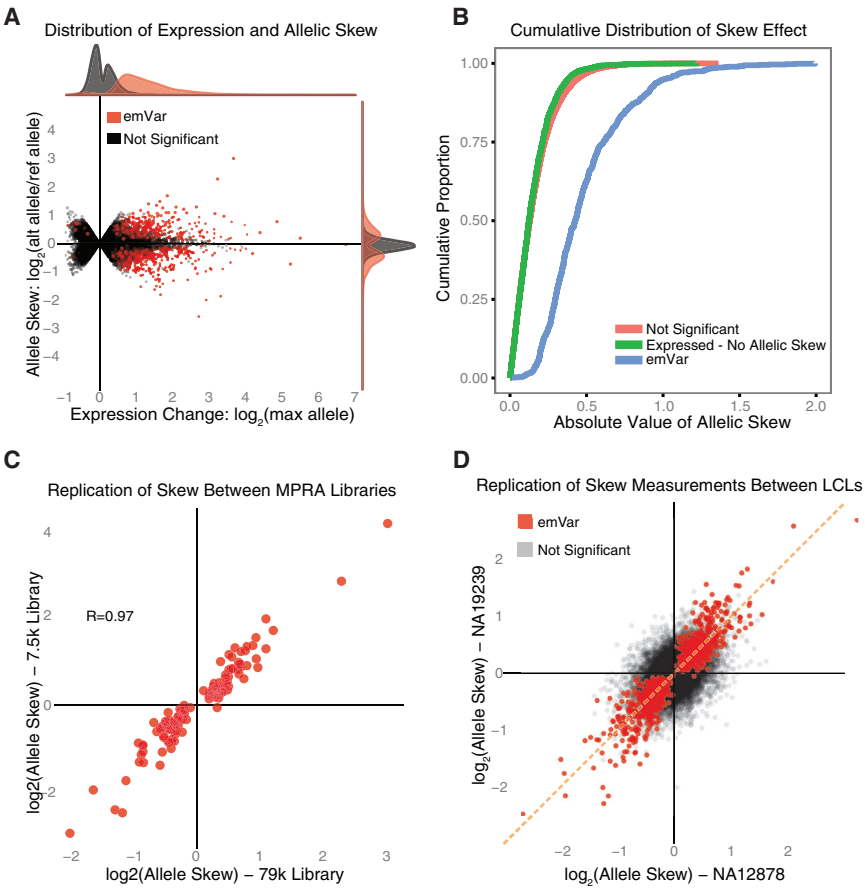


Figure 4. emVar Reproducibility and Effect-Size Distribution

(A) Distribution of expression strength (x axis) and allelic skew (y axis) for all 29k variants.

(B) Cumulative distribution of the absolute difference of the $\log_2$-fold change between the reference and alternate alleles for emVars (blue), active variants that were not detected as emVars (green), and all other variants (red).

(C) Allelic skew as measured by MPRA for 127 positive control values that were discovered in the original 79k library (x axis) and were tested in the 7.5k library (y axis).

(D) Comparison of allelic skew as estimated from the mean of two independent LCLs (NA12878 and NA19239). Red points in both plots denote variants called as emVars from the joint LCL analysis. Correlation is provided as Pearson's R.

highly reproducible. We randomly selected 127 emVars for testing in the second, 7.5k MPRA experiment and observed strong correlation for allelic skew to that of the 79k experiment (R = 0.97) (Figure 4C). For all 842 emVars, the effect size was highly correlated between the two LCLs tested (R = 0.92) and moderately correlated between LCLs and HepG2 cells (R = 0.63) (Figures 4D and S4A). Concordant with observations that eQTLs are associated with promoter regions (Veyrieras et al., 2008), we saw a 13.6-fold enrichment of emVars within core promoters (+100/−50 bp) relative to our test set of 29k variants and a 113.7-fold enrichment relative to all common variation (RR = 14.8, p = 2.7 × 10^{-52} and RR = 113.7, p = 1.2 × 10^{-121}, respectively). Despite this enrichment, many emVars fell outside promoters, with 59% lying at a distance of 10 kb or more from the nearest transcription start site (TSS), suggesting a prominent role for distal regulatory elements.

Like the overall set of active sequences, our emVars were enriched for markers associated with regulation, such as TF binding. We therefore examined whether the presence of allelic skew correlated with predicted disruption of a TF motif. Of the emVars that overlapped a ChIP-seq peak for a given TF and that contained the corresponding TF motif, the predicted strength of TF binding (based on position-weight matrices) differs significantly between alleles in 76% of cases (35 out of 46 had a difference of at least 3 in log-likelihood binding score based on the position-weight matrices). This was 4-fold greater than the difference for active sequences that did not show allelic skew (OR = 4.1, p = 8.1 × 10^{-8}) and 41-fold greater than for the non-active sequences (OR = 42.7, p = 1.9 × 10^{-36}; Figure 5A). The quantitative change in predicted binding also correlated with the magnitude of allelic skew observed by MPRA, supporting a direct relationship between predicted binding dynamics and regulatory activity within MPRA (R = 0.47, p = 6.4 × 10^{-10}; Figure 5B).

We predicted that if emVars were true regulatory variants, then the allele associated with higher expression would also be associated with greater chromatin accessibility as measured by DHS in their native context. By counting the number of DHS reads attributed to each allele at heterozygous sites in LCLs, we examined whether there was an allelic skew in DHS status for emVars. We found that emVars were significantly more likely to show DHS skew than active sequences that were not emVars (OR = 2.5, p = 0.003). Furthermore, 89% of variants shared the same direction of effect with a strong correlation in the magnitude of allelic skew of the emVar activity and the number of reads at DHS sites (R = 0.78, p = 1.0 × 10^{-8}; Figure 5C). We also predicted the same effect would be observed for TF binding as measured by ChIP-seq. We observed that emVars were more likely to show allelic skew in the binding of at least one overlapping TF than active sequences that were not emVars (OR = 1.9, p = 0.03). For emVars that showed allelic skew in TF occupancy for at least one TF, there was a substantial concordance of direction and magnitude between the allelic skew in TF binding and the allelic skew in activity (77% agreement in directionality, R = 0.60, p = 2.1 × 10^{-7}; Figures S5A and S5B).

Estimating Specificity of the MPRA Assay

We next set out to estimate the specificity of the MPRA assay. Because many of the variants tested are not actually drivers of eQTLs, we focused on a set that is likely to be enriched: the

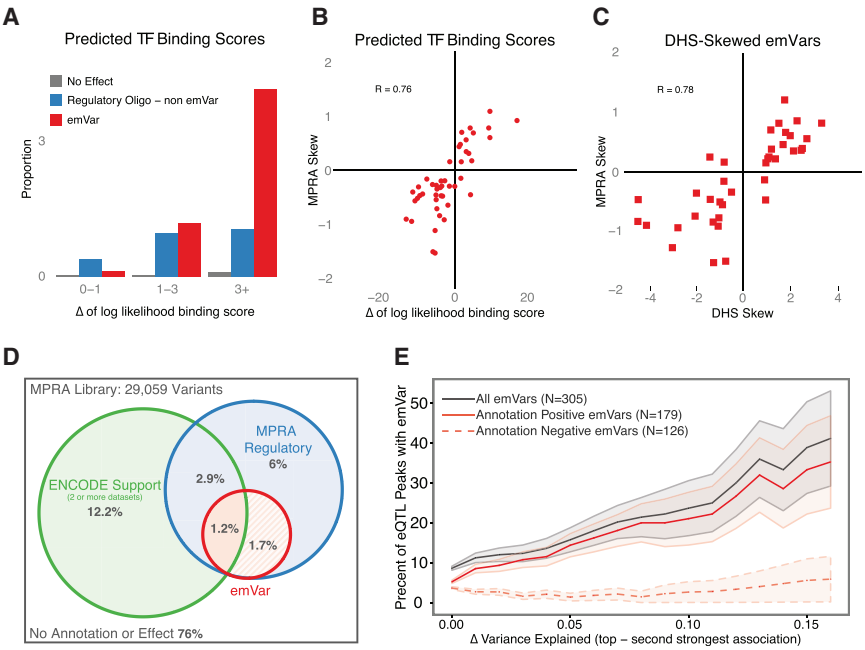


Figure 5. emVar Concordance with Existing Measures of Allelic Effect

(A) Proportion of variants by their MPRA classification that fall in an ENCODE TF ChIP-seq peak and contain the predicted motif. Variants are binned according to the difference in predicted binding strength between the two alleles. (For multiple TF associations, the one with the largest delta is used.)

(B) MPRA skew estimates for LCL emVars with TF motif/ChIP annotations compared to the predicted change in binding between the two alleles.

(C) Comparison between skew seen in MPRA and that in DHS for all emVars passing stringent filters for high-confidence DHS skew sites (Experimental Procedures). Skew is calculated as $\log_2$(Alt-allele counts/Ref-allele counts).

(D) Overlap between annotation-positive sites (Experimental Procedures), sequences detected as regulatory by MPRA, and emVars.

(E) Proportion of EUR eQTLs explained by an emVar plotted against the difference in variance explained between the top variant and the second strongest association in the EUR eQTL analysis. Gray line: all emVars; solid red line: annotation-positive emVars; dashed red line: annotation-negative emVars.

All Correlations are provided as Pearson's R.

top-associated variant for each eQTL and all variants in perfect LD. We carried out analysis on these 11,213 variants from 3,642 eQTLs taking into consideration that only one, if any, of the variants in each region may be a true regulatory variant and only a fraction of these may drive transcriptional regulation discoverable by MPRA.

We observed allelic skew in activity for 27% of the active sequences associated with an EUR eQTL and 26% of the YRI variants. In contrast, only 9% of the active sequences for the location-matched controls exhibited allelic skew, suggesting that two-thirds of the emVars associated with an eQTL peak are true causal variants for that peak. Randomly selected controls gave a similar result: 11% of active sequences contained emVars. Based on these results, we estimate the positive predictive value (PPV) to be 58%–68%. Because the controls were tested independently in a second library, we were careful to normalize the tag-counts to match those of the first experiment. To validate our normalization, we compared the proportion of active sequences in the two libraries. In the second, 7.5k library, 13.4% of the 500 location-matched and 9.7% of the 2,700 randomly selected controls scored as active sequences, compared to 12% for the 79k library (Supplemental Experimental Procedures).

As an alternative approach to explore the false discovery rate, we compared how often the direction of effect agrees between MPRA and the eQTL analysis. If we focus first on those emVars that reside in regions biologically annotated as likely to be related to enhancer function (by virtue of being marked by at least two of the following: DHS, CAGE, histone ChIP, or TF-ChIP), we find 80% agreement in directionality, corresponding to a PPV of 59% ($R = 0.61$, $p = 7.5 \times 10^{-15}$; Figures 5D and S5C; Supplemental Experimental Procedures). Notably, when we examine sites *not* supported by annotation, we observe a level of agreement consistent with random chance (48%, $R = 0.06$, $p = 0.61$). When we consider all emVars together (regardless of whether they are annotated as related to enhancer function), the concordance is 67% (PPV of 34% [$R = 0.33$, $p = 4.8 \times 10^{-7}$]). (We note that MPRA may not always correctly model the direction or magnitude of a variant's effect because the assay isolates a sequence from potential cofactors that may modify the effect; this has been observed for the genes yy1 and dorsal [Dubnicoff et al., 1997; Ip et al., 1991; Shi et al., 1991].) Finally, we explored whether some of the discordant emVars might represent false eQTL discoveries by removing the weakest one-third of eQTL associations; we observed a further increase in agreement for annotation-supported emVars of 84%, suggesting that false-positive eQTLs indeed contribute to the discordance.

Estimating Sensitivity of the MPRA Assay

We next estimated the sensitivity of MPRA to identify a causal eQTL variant when it is present in our study. Based on previous estimates, we expect the causal allele to be in our enriched set (top eQTL variant and all variants in perfect LD) 34% of the time in the EUR population. Given that our assay identified emVars in 8.6% (273/3171) of eQTL peaks, and 3%–6% after accounting for false positives, we estimate a sensitivity of 9%–18% in the EUR population. In the YRI population, lower LD makes it more likely that a top-scoring eQTL variant will be the causal allele, an estimated 41%. As expected we observe a larger fraction for YRI eQTLs with 13.8% (65/471) containing an emVar, 5%–10% after accounting for false positives, giving an estimated sensitivity of 12%–24%. When only taking into account variants supported by functional annotation, the estimated sensitivity is nearly equal by virtue of an increased PPV, emphasizing the value of filtering MPRA emVars with existing

Table 1. High-Confidence emVars Associated with Known GWAS Loci

GWAS Trait	Gene(s)	Sites Tested by MPRA			ENCODE and emVar[c]	Chr	Pos (hg19)	r^2 with Lead GWAS SNP
		All[a]	ENCODE[b]	emVar				
Mean platelet volume	*KIF1B*	26	4	3	rs4240912	1	10437778	0.92
					rs6670157	1	10458439	0.92
Wilms tumor	*DDX1*	79	3	1	rs60016948	2	15728544	1
Renal function-related traits	*PAX8*	18	5	1	rs7576384	2	113993385	0.96
Ankylosing spondylitis	*PTGER4*	5	4	1	rs9283753	5	40490609	0.99
Crohn's disease	ERAP2	147	25	2	rs1363974	5	96293816	0.91
Nasopharyngeal carcinoma	*IFITM4P*, *HLA-H,HCG4P5*, *HLA-J,HLA-G*	73	22	5	rs116025516	6	29910189	0.98
Beta-2-M plasma levels	*HCG27*, *HLA-L*	41	39	1	rs116587107	6	31239227	0.92
Systemic lupus erythematosus	*BLK*, *FAM167A*	16	14	1	chr8:11353110:D	8	11353110	1
Narcolepsy with cataplexy	*UBXN2B*	12	3	1	rs56316188	8	59323811	0.95
IgG glycosylation	*B4GALT1*	12	5	1	rs12342831	9	33124872	1
Inflammatory bowel disease	*MAP3K8*	31	2	3	rs306587	10	30722908	0.98
Crohn's disease	*CREM*	241	22	5	rs16935880	10	35415468	0.99
					rs4934730	10	35415555	0.99
Body mass index	*C1QTNF4*	26	1	1	rs35184771	11	47475189	0.97
Atopic dermatitis	*AP5B1*, *OVOL1*	2	1	1	rs10791824	11	65559266	0.91
Mean corpuscular hemoglobin	*PTPLAD1*	60	7	1	rs28640237	15	66070962	0.99
Body mass index, obesity, weight	*EIF3CL*, *EIF3C*, *SPNS1*, *CDC37P1*	137	33	4	rs7198606	16	28875122	1
Parkinson's disease	*STX4*	50	10	2	rs58726213	16	31044683	0.95
					rs11865038	16	31095171	1
Bone mineral density	*C17orf53*	56	15	1	rs227578	17	42210189	1
Coronary heart disease	*UBE2Z*	105	18	8	rs4378658	17	46993370	0.99
Liver enzyme levels (ALP)	*GINS1*, *ABHD12*	319	45	5	rs2258769	20	25276680	0.99

[a]All variants tested in this study by MPRA with an r^2 of 0.9 or greater to the lead eQTL variant.
[b]Variants within the tested subset classified as having strong ENCODE support (Supplemental Experimental Procedures).
[c]emVars that were classified as having strong encode support.

annotations. The estimate that MPRA can identify the causal allele for an eQTL for 9%–24% of peaks when tested is in line with the previous observations that 23%–64% of eQTLs are driven by promoter or enhancer modifications, the processes we expect MPRA to capture.

We further performed two alternate estimates for sensitivity focusing on regions where the causal allele is likely to be captured in our dataset. We first partitioned peaks based on the difference in variance (Δr^2) between the lead variant (the variant tested by MPRA) and the second strongest association. The top eQTL variant is most likely to be causal when the Δr^2 is large; accordingly we see an increase of emVars in these regions (Figure 5E). Modeling this relationship using a logistic regression that also controlled for the effect size of the eQTL, we derived a sensitivity of 16%–21%. Second, we identified eQTL peaks where the same top-associated variant occurred in both EUR and YRI. Differing LD structure between the two populations decreases the number of linked variants and increases the confidence that the top variant is causal. Of the 34 shared variants, 8 were identified as emVars, suggesting a 24% sensitivity of MPRA to correctly identify the causal allele when it is tested. Both orthogonal approaches are consistent with our initial estimate of 9%–24%.

GWAS-Associated Regions

We next investigated in greater depth regions that were previously associated with a trait or disease in human studies. For 209 eQTLs overlapping 163 GWAS loci, we tested all alleles in strong LD ($r^2 > 0.9$) of an eQTL variant, a total of 9,664 variants. We identified 248 emVars in 99 eQTLs (Table S2). Based on our previous findings, we prioritized the emVars that also carry annotations associated with an enhancer or promoter; we identified 53 emVars in 56 eQTLs (a subset of these, further restricted by LD, is shown in Table 1). This represents a highly promising set of candidates and greatly reduced testing burden compared to current approaches. For example, applying only our stringent ENCODE annotation criteria identifies 1,302 variants across 171 of the 209 eQTL peaks.

Candidates identified through MPRA still require experimental validation. We pursued a striking example, in a distal enhancer for prostaglandin E receptor 4 (*PTGER4*). The emVar rs9283753 sits 190 kb away from the gene and is in strong LD with the top-associated risk allele for ankylosing spondylitis (with moderate LD to risk alleles for Crohn's disease and multiple sclerosis) (Figures 6A–6C) (Barrett et al., 2008; Evans et al., 2011;

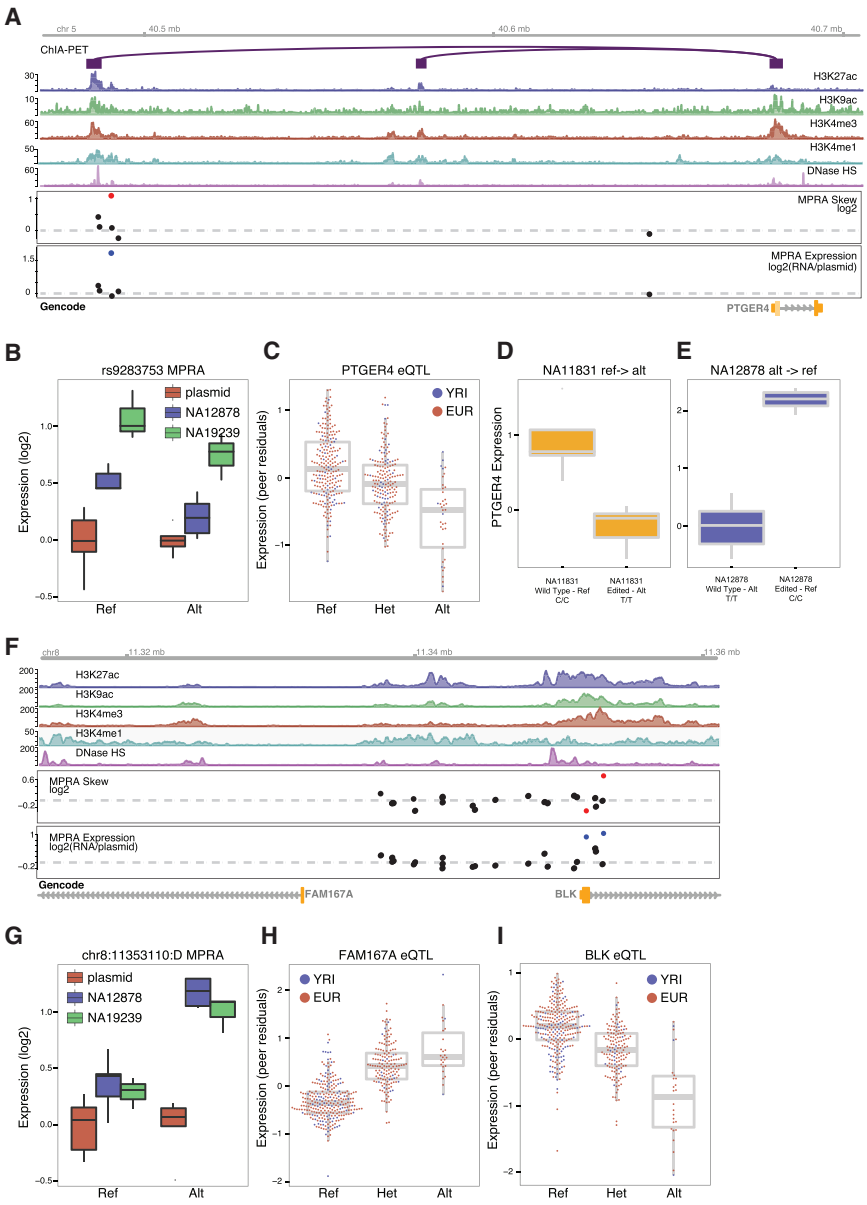


Figure 6. emVars Associated with Ankylosing Spondylitis and SLE

(A) Plot of the *PTGER4* locus, which overlaps a GWAS peak for ankylosing spondylitis displaying ChIA-PET and ENCODE annotations (top 6 tracks), observed allelic skew (track 7), and expression strength (track 8) from MPRA. Significant variants for expression (blue) and skew (red) in the MPRA data are indicated by color; black: non-significant.

(B) MPRA expression values of the *PTGER4* variant rs9283753 in LCLs normalized to the plasmid library.

(C) LCL eQTL results in EUR and YRI populations for the *PTGER4* with rs9283753.

(D and E) *PTGER4* expression as measured by qPCR for two LCLs that underwent allelic replacement at rs9283753.

(F) Plot of the *FAM167A-BLK* locus associated with SLE.

(G) MPRA expression values of the chr8:11353110 deletion variant in LCLs normalized to the plasmid library.

(H and I) LCL eQTL results in EUR and YRI populations for the *FAM167A* and *BLK* associations.

De Jager et al., 2009). The variant resides in a distal enhancer clearly defined by strong DHS and H3K27ac marks, with a CREB motif residing over rs9283753. The allele change is not predicted to alter binding of CREB, however, and further work will be needed to elucidate the mechanism of regulation.

To validate the *PTGER4* emVar, we used homology-directed repair with CRISPR/Cas9 to perform allelic replacement. We edited two cell lines, a homozygous-ancestral (NA12878) and a homozygous-derived (NA11831) individual for the variant to test the effect of the allele in a controlled isogenic background. As expected from the MPRA and eQTL data, switching NA12878 to be homozygous for the derived allele caused an increase in expression for *PTGER4*, whereas the replacement with the ancestral allele decreased expression of NA11831 (Figures 6D–6E). The concordant MPRA, eQTL, and CRISPR data support the hypothesis that the risk allele is associated with decreased expression of *PTGER4*.

Our finding of a regulatory variant in the distal enhancer in *PTGER4* is consistent with prior observations that identify elements outside core promoters as significant contributors to the heritable component of complex diseases (Farh et al., 2015; Gjoneska et al., 2015; Parker et al., 2013). Looking more broadly, the 188 emVars in strong LD ($r^2 >= 0.9$) with GWAS variants tend to lie further from promoters than randomly chosen eQTL variants: 78% (147) reside greater than 10 kb from an active TSS, compared to 53% for all other emVars. We observed a corresponding depletion in strongly linked GWAS emVars within core promoters compared to all non-GWAS emVars (RR = 5.3, p = 0.0015).

There are many other promising candidates to pursue, including both a core promoter and intronic emVar in the *BLK* locus associated with systemic lupus erythematosus (SLE). The locus has previously been characterized by Guthridge and colleagues, who reported the promoter variant rs922483 (Guthridge et al., 2014). We replicated this finding via MPRA while also observing a second emVar within the first intron of the gene. This is a one-base deletion at chr8:11353110 that introduces a novel NF-kB-binding site. Notably, we found that this emVAR decreased expression of *BLK* while increasing expression of the nearby gene FAM167A and was validated with the traditional luciferase assay (Figures 6F–6I). Moreover, this emVAR is in perfect LD with the top-associated SLE risk variant

among Europeans, rs2618476 (Graham et al., 2008; Guthridge et al., 2014).

DISCUSSION

Our findings demonstrate that MPRA can be an invaluable tool for localizing individual causal variants influencing phenotypic traits. We have discovered hundreds of variants as putative causal alleles for gene expression, many of which are linked to known disease-causing loci. Furthermore we directly demonstrate causality by allelic replacement of an ankylosing spondylitis risk allele, rs9283753, which modulates expression of *PTGER4* from a distal enhancer.

As with any assay, it is important to understand the limitations of MPRA. The sensitivity of our current assay, which can identify an estimated 9%–24% of the eQTL causal alleles, is limited in several distinct ways. (1) Causal alleles of weak effect may fall below MPRA's limit of detection. (2) Regulatory processes may require additional sequence context not captured on the oligo, for example, when transcription depends on nearby DNA-binding co-factor(s) or chromatin structure. (3) Transcription-repressing effects might be undetectable due to the low basal activity of the minimal promoter used. (4) Causal alleles may regulate post-transcriptional events such as mRNA processing and stability.

The first three categories represent limitations of the current assay design and may be overcome in subsequent iterations of MPRA. Analysis of the proportion of active variants suggests that for one-third of the 79k library, we were underpowered due to a low abundance of the plasmid pool, something that could be overcome by increased sequencing and library uniformity (Figure S2C). In addition, further improvements, such as longer oligo sequences to capture greater contextual information and the use of a stronger constitutive promoter to detect repressive elements, may provide substantial gains in sensitivity. Nevertheless, there is undoubtedly contextual information, such as long-distance interactions, that will never be captured by an episomal assay.

One of the largest influences to the current sensitivity is the substantial role of post-transcriptional effects driving eQTLs; these are not targeted by our assay. For example, a recent analysis by Farh and colleagues of eQTL causal variants estimated that 36% of sites fall within transcripts themselves, and only 23% are attributed to known promoter/enhancer elements, suggesting a substantial role for post-transcriptional activities (Farh et al., 2015). This implies that, at best, MPRA would have a maximum sensitivity of 23%–64% for detecting an eQTL causal allele, as it is not designed to detect variants acting post-transcriptionally. In contrast, the same study reported a very different picture for autoimmunity GWAS hits: only 19% of causal alleles fell in transcripts, and 67% resided in known promoter/enhancers, with the remainder associated with unannotated non-coding sequence. The discrepancy in the predicted mechanisms of eQTL and GWAS causal sites suggests that the sensitivity of MPRA may well be higher for disease-associated variants than reported here (Ulirsch et al., 2016 [this issue of *Cell*]).

Although the sensitivity may be increased through further technical development, the positive predictive value of 34%–68% is likely an inherent property of the assay. This suggests that a substantial segment of the genome has the potential to change gene expression but is repressed from doing so through modulating interactions or heterochromatin silencing. Endogenously silenced sequences likely also explain a proportion of the active sequences we observed by MPRA; we note that this proportion was unexpectedly high. As a result some variants discovered by MPRA will be of little biological value. However, the assay still identifies 1–2 true causal allele for every 3 variants that score, which provides an enrichment and throughput unparalleled by alternative approaches. Although MPRA does not prove causality, it does substantially reduce the test space of alleles linked to a trait locus and provides a concise list of high-priority targets for follow-up. Furthermore, the improved agreement with eQTL directionality when subsetting those emVars with supporting biological annotation demonstrates the strength of a combined approach when searching for non-coding causal alleles.

Regardless of the high-throughput approach taken to identify variants influencing gene regulation, whether it is computational or experimental, it is critical that the results are interpreted as the product of a discovery tool and not as a test for causality; this is a first step in the difficult task of linking a genetic loci to a physiological phenotype. By example, we demonstrate for *PTGER4* how we can readily identify and validate an allele that influences gene expression, and extending this observation further to a disease causation will require further work. Being able to identify and validate expression-modulating variants from tens of thousands of sites will ultimately greatly aid in our ability to translate non-coding regulatory code and will bring us a step closer to the difficult task of linking human genetic variation to specific phenotypic traits.

EXPERIMENTAL PROCEDURES

Variant Selection

To construct the 79k oligo library, eQTLs were identified by reanalysis of the Geuvadis RNA-seq dataset of LCLs from individuals of EUR and YRI ancestry (Supplemental Experimental Procedures). We used significance thresholds corresponding to a 0.1% false-positive rate within permutated samples to identify 3,642 eQTLs within EUR and YRI. Using the selection and design criteria described in the Supplemental Experimental Procedures, we included 29,173 variants to test by MPRA. After accounting for both the reference and alternate alleles, neighboring variants, and in some instances orientation of the oligo relative to the promoter, we designed a total of 78,956 oligos with the variant of interest centered within 150 bp of genomic sequence.

The 7.5k oligo library was constructed by selecting variants representing four different classes: (1) variants called as expression positive in the 79k oligo experiment; (2) variants called as expression positive and having allelic skew (emVars) in the 79k oligo experiment; (3) location-matched controls, selected for being between 250 and 1,000 bp of a lead eQTL association and not in LD with the lead candidate ($r^2 \leq 0.25$) and for not having an appreciable eQTL signal in the Geuvadis or GTEx datasets; (4) randomly selected variants from across the genome matching only to the minor allele frequency spectrum of EUR eQTL variants. A subset of the randomly selected variants were further filtered for having no detectable eQTL signal in the Geuvadis and GTEx datasets. The two sets of randomly selected sites behaved similarly by MPRA and were combined as a single set during analysis.

MPRA

Oligos were synthesized (Agilent Technologies) as 180 bp sequences containing 150 bp of genomics sequence and 15 bp of adaptor sequence on either end. Unique 20 bp barcodes were added by emulsion PCR along with

additional constant sequence for subsequent incorporation into a backbone vector by gibson assembly (Table S3). The oligo library was expanded by electroporation into *E. coli*, and the resulting plasmid library was sequenced by Illumina 2 × 150 bp chemistry to acquire barcode/oligo pairings. The library underwent restriction digest, and GFP with a minimal TATA promoter was inserted by gibson assembly resulting in the 150 bp oligo sequence positioned directly upstream of the promoter and the 20 bp barcode falling in the 3′ UTR of GFP. After expansion within *E. coli* the final MPRA plasmid library was sequenced by Illumina 1 × 30 bp chemistry to acquire a baseline representation of each oligo within the library.

Libraries were electroporated into LCLs using the Neon system (Life Technologies). We performed multiple independent replicates for NA12878 (5 replicates) and NA19239 (3 replicates) with each replicate consisting of ~5 × 10^8 cells. Transfections for 5 independent replicates of HepG2 cells were performed using Lipofectamine 3000 (Life Technologies). For both cell types RNA was harvested 24 hr post-transfection followed by DNA digestion, capturing of the GFP transcripts, and cDNA synthesis. Sequencing libraries were constructed by adding adapters by PCR and sequenced using Illumina 1 × 30 bp chemistry. Detailed experimental conditions as well as oligo and primer sequences are provided in the Supplemental Experimental Procedures.

Allelic Replacement at *PTGER4* Locus

Cas9-GFP vector, guide RNA (gRNA) targeting rs9283753, and a 150 bp homology oligo with either the reference (C) or alternate (T) allele were transfected into 5 × 10^6 LCLs. Cells were sorted for GFP expression 24 hr post-transfection and expanded for 2 weeks in bulk. Single-cell dilutions of each bulk population were performed and after 2 weeks of growth genotyped using Illumina sequencing to identify mutations of interest. All clones were confirmed by Sanger sequencing. To quantify changes in expression of *PTGER4*, qPCR was performed on clonal colonies identified as either HDR or wild-type. RNA was collected from ~7.5 × 10^6 cells, and cDNA was synthesized. qPCR was performed with technical triplicates for each reaction. Detailed transfection and qPCR conditions as well as gRNA, homology oligo, and primer sequences are provided in the Supplemental Experimental Procedures.

Data Analysis

The sum of the barcode counts for each oligo within replicates was normalized, and oligos showing differential expression relative to the plasmid input were identified by modeling a negative binomial with DESeq2 and applying a threshold of 0.01 for the Bonferroni corrected p value. For sequences that displayed regulatory activity, we applied a t test on the log-transformed RNA/plasmid ratios for each experimental replicate to test whether the reference and alternate allele had similar activity (Figures S4B–S4E). Combining independent results from NA12878 and NA9239 using Fisher's method generated a final LCL-specific call set. We used an FDR (Benjamini–Hochberg) cutoff of 5% as a threshold for calling emVars. Detailed procedures for calculating enrichments, sensitivity/specificity, and concordance with established measures of regulatory activity are provided in the Supplemental Experimental Procedures and Table S4.

ACCESSION NUMBERS

The accession number for the sequencing data reported in this paper is GSE75661.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.04.027.

AUTHOR CONTRIBUTIONS

All authors contributed in the development of this work. R.T. and P.C.S. conceptualized and initiated the study; R.T., D.S.P., and S.W. performed experimental work; R.T., D.K., B.L., S.K.R., K.G.A., T.S.M., E.S.L., S.F.S., and P.C.S. analyzed the data; and R.T., S.F.S., and P.C.S. wrote the paper.

ACKNOWLEDGMENTS

We would like to thank A. Melnikov, members of the Sabeti lab, the Broad walk-up sequencing platform, and C. Daly and K. Groglio of Harvard's Bauer Core Facility for their expertise and advice. This work was supported by NIH grants DP2OD006514 to P.C.S. and K99HG008179 to R.T. T.S.M. is supported by NIH grant R01HG006785. Broad Institute and Harvard have filed patent applications directed to the work described in this paper.

Received: July 17, 2015
Revised: November 25, 2015
Accepted: April 12, 2016
Published: June 2, 2016

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.

Claussnitzer, M., Dankel, S.N., Kim, K.-H.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C., et al.; International MS Genetics Consortium (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet. *41*, 776–782.

Dubnicoff, T., Valentine, S.A., Chen, G., Shi, T., Lengyel, J.A., Paroush, Z., and Courey, A.J. (1997). Conversion of dorsal from an activator to a repressor by the global corepressor Groucho. Genes Dev. *11*, 2952–2957.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Evans, D.M., Spencer, C.C., Pointon, J.J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., Stone, M.A., et al.; Spondyloarthritis Research Consortium of Canada (SPARCC); Australo-Anglo-American Spondyloarthritis Consortium (TASC); Wellcome Trust Case Control Consortium 2 (WTCCC2) (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat. Genet. *43*, 761–767.

Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Gjoneska, E., Pfenning, A.R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H.H., and Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature *518*, 365–369.

Graham, R.R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C.J., Leon, J.M., Burtt, N.P., Guiducci, C., Parkin, M., Gates, C., et al. (2008). Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat. Genet. *40*, 1059–1061.

Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., et al.; 1000 Genomes Project (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.

Guthridge, J.M., Lu, R., Sun, H., Sun, C., Wiley, G.B., Dominguez, N., Macwana, S.R., Lessard, C.J., Kim-Howard, X., Cobb, B.L., et al. (2014). Two functional lupus-associated BLK promoter variants control cell-type- and developmental-stage-specific transcription. Am. J. Hum. Genet. *94*, 586–598.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlick, Y. (2015). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.

Ip, Y.T., Kraut, R., Levine, M., and Rushlow, C.A. (1991). The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in *Drosophila*. Cell *64*, 439–446.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., and Cohen, B.A. (2014). High-throughput functional testing of ENCODE segmentation predictions. Genome Res. *24*, 1595–1602.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747–749.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Jr., Callan, C.G., Jr., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Mogno, I., Kwasnieski, J.C., and Cohen, B.A. (2013). Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Res. *23*, 1908–1915.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

Ow, D.W., DE Wet, J.R., Helinski, D.R., Howell, S.H., Wood, K.V., and Deluca, M. (1986). Transient and stable expression of the firefly luciferase gene in plant cells and transgenic plants. Science *234*, 856–859.

Parker, S.C., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al.; NISC Comparative Sequencing Program; National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. USA *110*, 17921–17926.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Shi, Y., Seto, E., Chang, L.S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. Cell *67*, 377–388.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.

Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., and Sankaran, V.G. (2016). Systemic functional dissection of common genetic variation affecting red blood cell traits. Cell *165*, this issue, 1530–1545.

Veyrieras, J.-B.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.

Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, G.M., Lowe, W.L., and Reddy, T.E. (2015). Massively parallel quantification of the regulatory effects of non-coding genetic variation in a human cohort. Genome Res. *25*, 1432–1441. Published online June 17, 2015. http://dx.doi.org/10.1101/gr.190090.115.

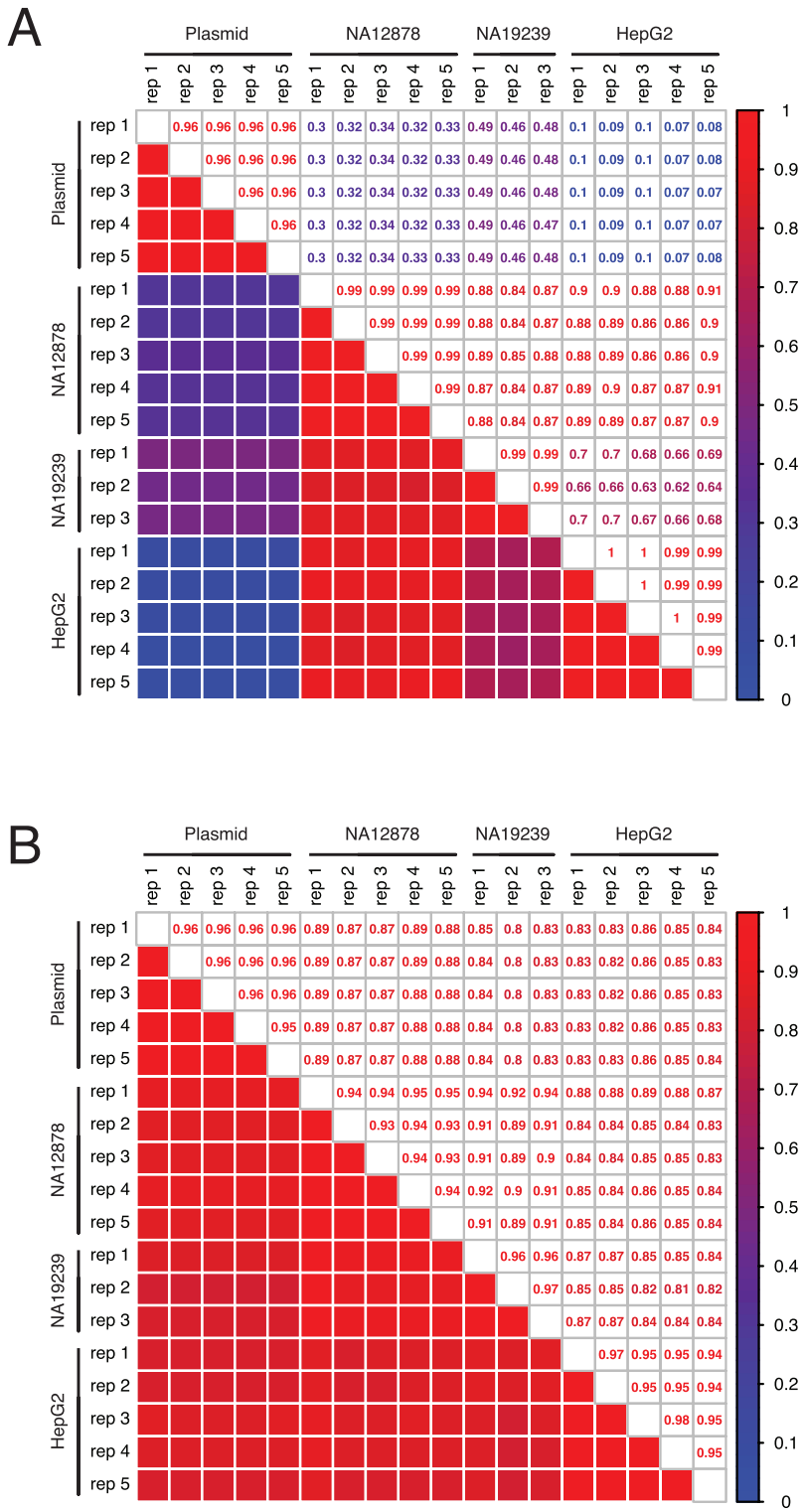


Figure S1. Correlation Matrices of Normalized Read Counts, Related to Figure 2

(A) Coefficient of determination of normalized oligos counts between all replicates.

(B) Correlation of counts for oligo with values of 7,500 or less, which demonstrates the difference between replicates that are largely driven by highly expressed sequences.

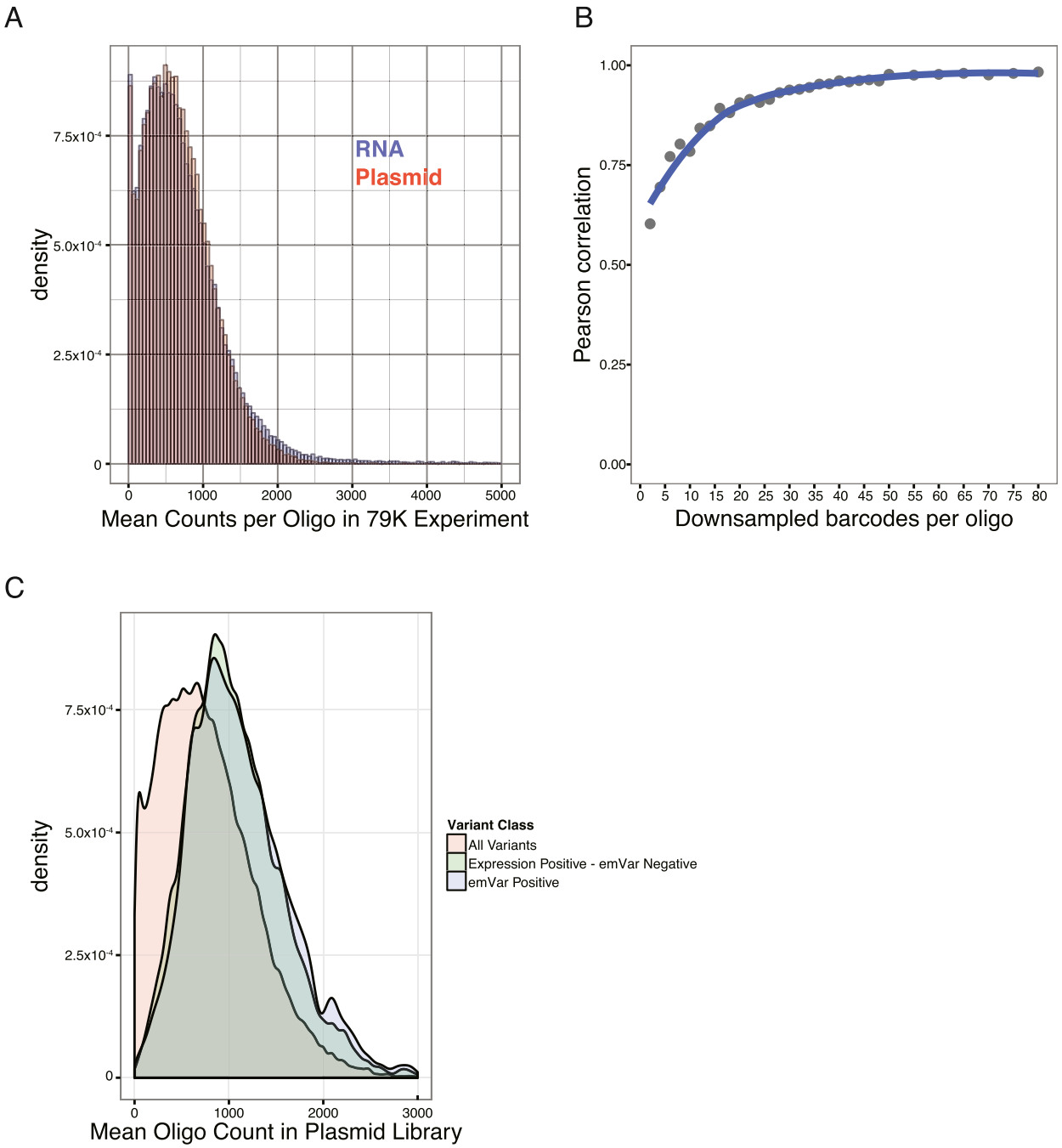


Figure S2. Read and Tag Counts per Oligo, Related to Figure 2

(A) Unnormalized mean reads from the sum of all barcodes counts for the five plasmid replicates and the 13 experimental samples.

(B) Effect of barcode abundance on the oligo mean. The 79k MPRA library was downsampled to select the specified number of barcodes for each oligo. Barcodes were summed to obtain oligo counts for normalization in DESeq. Replicates for NA12878 were averaged and compared to oligo counts calculated from all barcodes.

(C) Density plot of unnormalized plasmid counts for all variants, expression-positive, skew-negative, and skew-positive variants tested in the 79k assay.

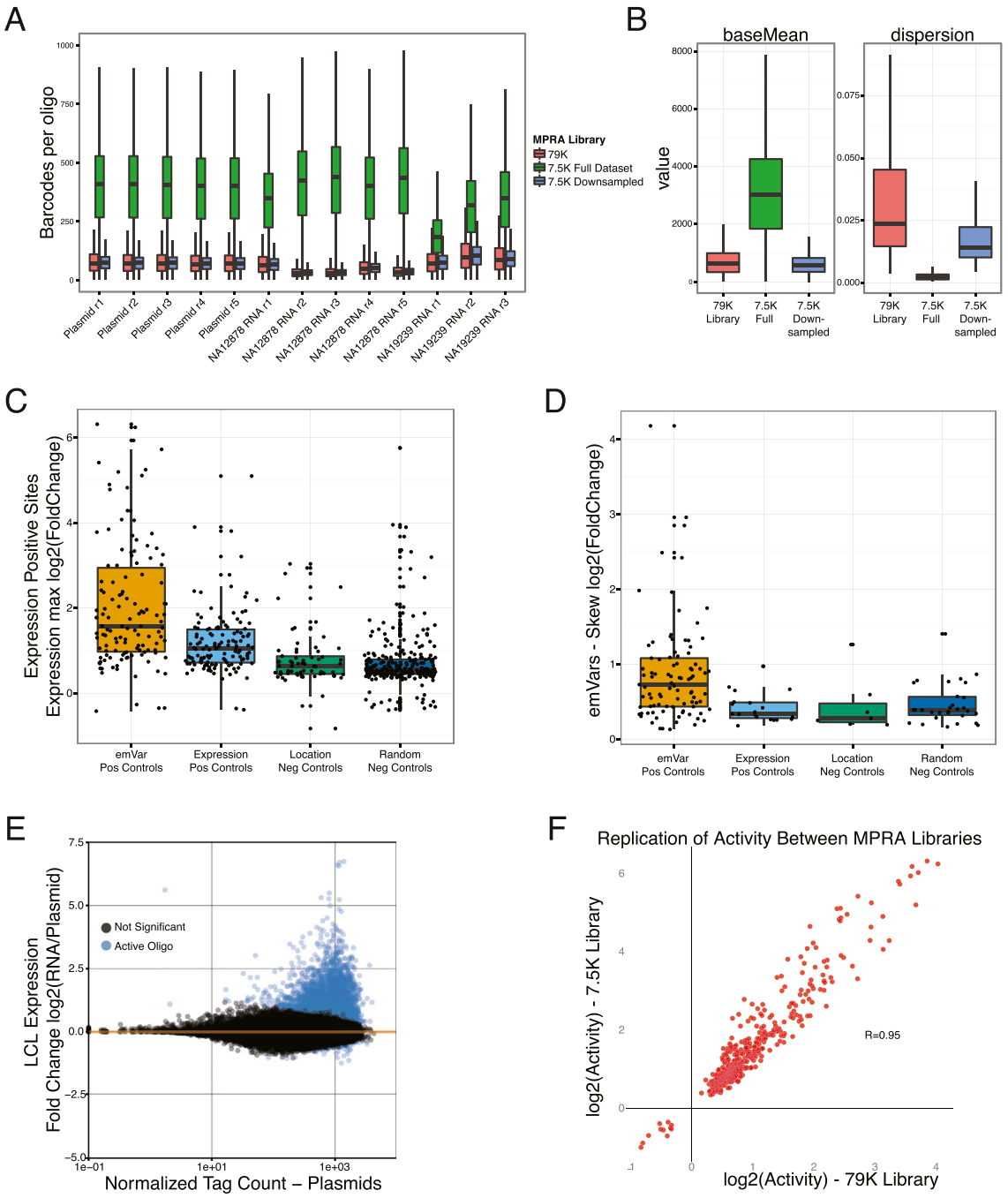


Figure S3. Performance Measures of the 7.5k and 79k MPRA Experiments, Related to Figure 3

(A) Bar plots of the number of barcodes tagging each oligo for the initial 79k MPRA library and the 7.5k libraries at before and after downsampling.

(B) Bar plots of the base mean value and dispersion (variance) as estimated in DESeq2 for the 79k MPRA library and the 7.5k library using all barcodes and a downsampled subset.

(C and D) Effect sizes from 7.5k MPRA experiment for expression (C) and skew effect size (D) of positive and negative control variants from the downsampled 7.5k library.

(E) Log_2-fold change of oligo expression in RNA data plotted against the normalized abundance of each oligo in the plasmid library for the combined LCL analysis.

(F) Replication of MPRA active sequences for positive controls oligos discovered in the original 79k library (x axis) and tested in the 7.5k library (y axis). Only sequences that past significance thresholds in the 79k library are shown.

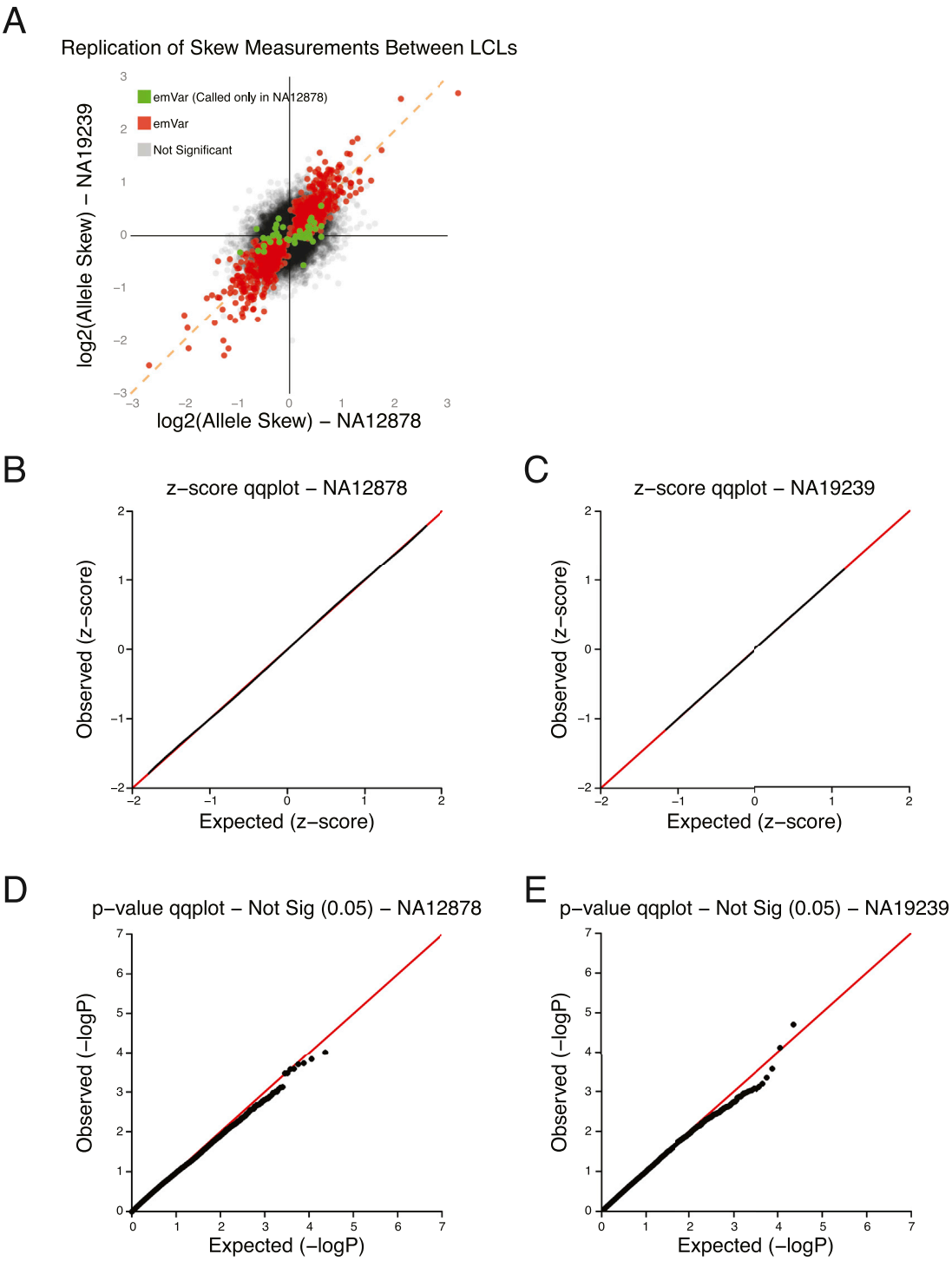


Figure S4. Performance of the Allelic Skew Measurements in the 79k Library, Related to Figure 4

(A) Comparison of allelic skew as estimated from the mean of two independent LCL experiments. Points colored in red were called in the joint LCL analysis. Green points represent variants called significant when analyzing only NA12878 but failed to reach significance in the joint analysis.

(B and C) Z-score qq-plots of the expression differences between alleles. Differences between log ratios of the reference and alternate allele were converted to Z scores for each oligo pair using all sequence in (B) NA12878 and (C) NA19283.

(D and E) qq-plots of p values from the allelic skew t test at non-expressed sequences. Ref/alt pairs in (D) NA12878 and (E) NA19283 were considered not expressed if both oligos had uncorrected expression p values of 0.05 or greater.

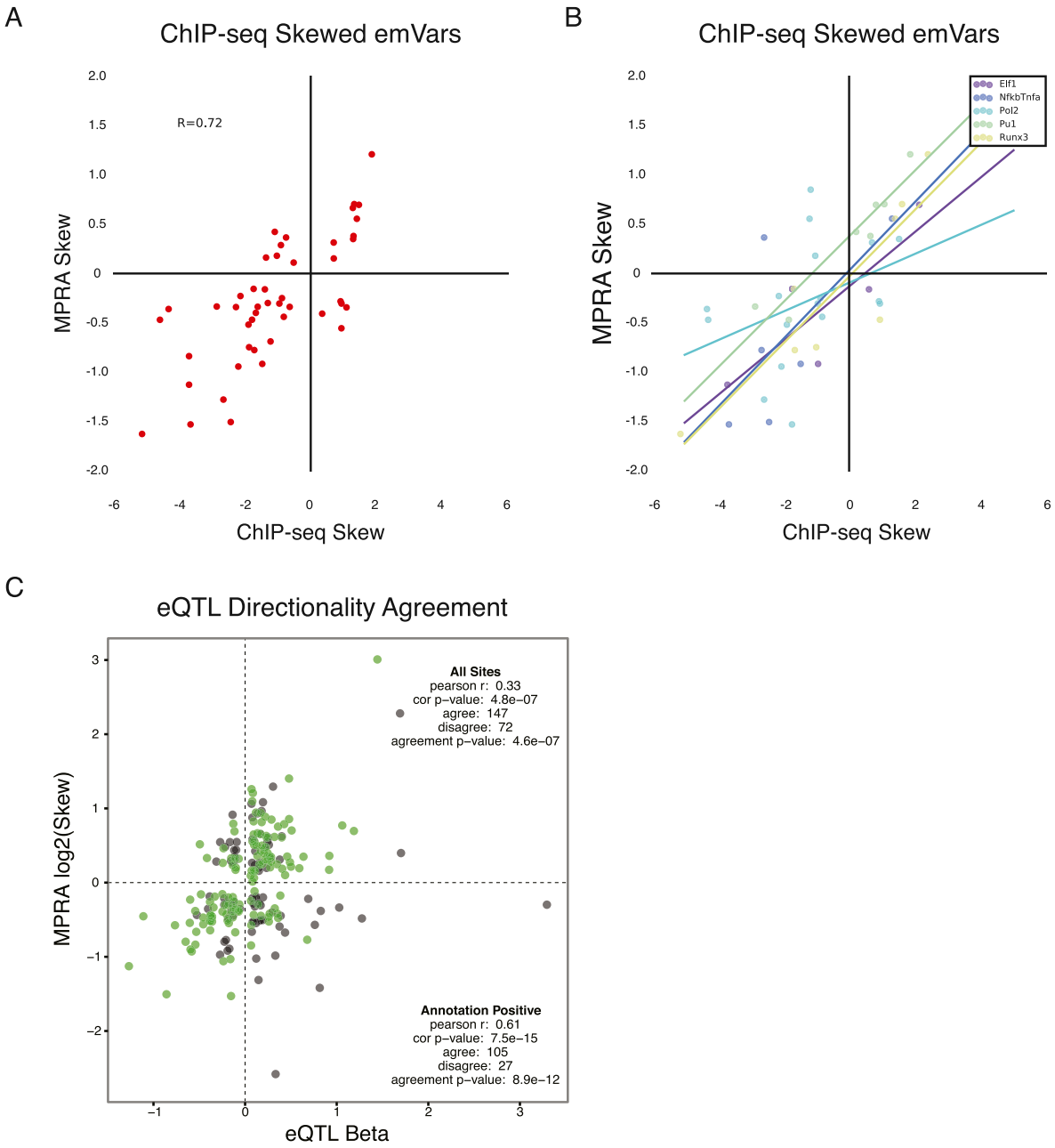


Figure S5. Correlation of MPRA Allelic Skew with Existing Datasets, Related to Figure 5

(A and B) Correlation of skew in TF binding with skew on the MPRA.

(A) Plotted are all emVars that passed stringent filters for high-quality TF-binding sites demonstrating allelic skew (Supplemental Experimental Procedures) with MPRA allelic skew FDR < 0.05, and with DHS skew p < 0.1. For variants that were overlapped by multiple TF-binding sites, we picked the TF with the most statistically significant skew. Skew is calculated as $\log_2$(Alt-allele counts/Ref-allele counts) and correlation is provided as Pearson's R.

(B) Like (A) but with points labeled by the TF showing allelic skew. Points are only plotted for TFs with detectable allelic skew at 3 or more emVars. emVars overlapping more than one TF are plotted as multiple points. A simple linear regression line for each transcription factor is plotted through the points.

(C) Shared directionality between MPRA and eQTL results. MPRA skew (y axis) plotted against beta values from the EUR eQTL analysis. Variants with significant associations to multiple genes in the eQTL analysis were discarded if all genes failed to share directionality. Green points denote annotation-positive emVars; annotation-negative sites are plotted as gray points.

Cell, Volume *165*

Supplemental Information

Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay

Ryan Tewhey, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, Tarjei S. Mikkelsen, Eric S. Lander, Stephen F. Schaffner, and Pardis C. Sabeti

Supplemental Experimental Procedures

eQTL Calling and SNP Selection

RNA Mapping
Illumina 100 bp paired-end RNA-seq reads from 446 lymphoblastoid cell lines (LCLs) sequenced by the Geuvadis consortium were downloaded from http://www.ebi.ac.uk/ena/data/view/ERP001942 (Lappalainen et al., 2013). Samples were mapped with Tophat v2.0.9 without coverage searching and an edit distance of 3 against human genome build 37 using Gencode v13 as a transcriptome guide (Harrow et al., 2012; Kim et al., 2012). Coverage across genes were estimated with Cufflinks v2.1.1 using multi read correction and masking of rRNA and tRNA loci (Trapnell et al., 2010). Fragments per kilobase of transcripts (FPKM) estimates for each gene were acquired from Cufflinks and genes were filtered for having at least one individual with an FPKM of 0.5 or greater. Expression values were log2 transformed and normalized for both known and hidden covariates using the PEER software package (Stegle et al., 2010). Gender, population and the sequencing laboratory for each sample were provided to PEER as known covariates and the number of unknown covariates was set to 10.

Sample Imputation
Genotypes were obtained for 420 RNA-seq samples from the phase 1 release of the 1000 genomes (1KG) project (Consortium). For the remaining 26 samples where 1KG phase1 data was not available, we obtained Illumina OMNI 2.5M, Affymatrix Axiom 5M and HapMap data. Merged genotypes for the 26 individuals were imputed against the full phase 1 collection of genotypes using IMPUTE2 with the requirement of an imputation confidence score of 0.9 or greater for keeping the imputed call. All indels greater than 35 bp were removed from subsequent analysis (Howie et al., 2009).

eQTL Calling
Genotypes were separated into three groups by population: all (1KG), Yoruba (YRI) and European (EUR) individuals, and filtered to include only variants at greater than 5% minor allele frequency within each population group. PEER residuals and genotypes were provided to matrix eQTL to calculate cis eQTLs (SNP/gene distance less than 1Mb) using an additive linear model. To set a significance threshold we ran 1000 permutations for chromosomes 1,7,16 and 19 in each population group and calculated the p-value that corresponded to an empirical 0.1% false positive rate and used these values as significance thresholds in our eQTL analysis. This resulted in p-value cutoffs of $6.3x10^{-11}$ and $4.1x10^{-10}$ for YRI and EUR respectively (eQTL count: 471 in YRI & 3171 in EUR). For every gene with a significant snp/gene association, we performed a conditional analysis for all other SNPs within the 1Mb window of the gene using the top associated variant as a covariate in the linear model and the same p-value thresholds used in the initial pass. We iterated through consecutive rounds of conditional analysis adding each new top associated SNP to the linear model until no other variants in the region showed a significant association with gene expression (conditional eQTL count: 315 in EUR & 8 in YRI).

SNP Selection for Reporter Assay
To select variants for testing in the 79k reporter assay we first selected the top associated variant for each significant eQTL in both the primary and conditional eQTL analysis of EUR and YRI (3,965). We calculated linkage disequilibrium (LD) for every top association within the discovery population and selected all variants that were in perfect LD (r^2=1) with it (12,321 variants). In addition, for each gene significantly associated in EUR we selected the top associated variant and all variants in perfect LD within the YRI and 1KG analysis regardless of the strength of the association (12,230 variants, 4,177 redundant with prior selections).

We then selected a subset of eQTL peaks to characterize comprehensively, beyond only the top associated variants. We chose 209 eQTLs for testing of all variants having an r^2 of 0.9 or greater with the most highly associated variant (9,921 variants, 1,122 redundant with prior selections). We selected these 209 peaks based on their intersection with SNPs in the NHGRI's catalog of published GWAS hits, capturing a total of 163 GWAS SNPs. To determine overlapping peaks, we calculated r^2 within the 1KG EUR supergroup to top associated variants in the eQTL and GWAS analysis. We called loci overlapping when a shared variant was within 0.8 r^2 of the GWAS loci and 0.9 r^2 of the eQTL peak.

To create oligonucleotides (oligos) for the 29,173 sites we centered the variant within 150 bp of flanking hg19 sequence (74 bp on the 5' and 75 bp on the 3' end for SNPs). To determine the orientation of the oligo we chose the direction of the variant relative to the transcription start site of the gene it was associated with in the eQTL analysis. If the variant was associated with multiple genes in different orientations, we designed oligos in both the forward and reverse direction. For variants within the test set where other variants fell within the 150 bp oligo, we created an additional alternative haplotype oligo testing the alternate and reference allele of the centered variant with the alternative allele(s) inserted into the flanking sequence. Finally, there were 7 variants that contained an AsiI restriction site within the flanking sequence that would cause these oligos to be lost during construction of the MPRA library. To rescue the oligos we made a single base change in the restriction site; none of these changes altered the variant site or 20 bp flanking either side. In total we designed 78,956 oligos for synthesis, testing a total of 29,173 variants.

The smaller 7.5k oligo reporter library was constructed of 5 subsets of variants (2 positive controls and 3 negative controls). For positive controls we randomly selected 137 variants that were expression positive but emVar negative and 127 emVar positive variants from the 79k oligo experiment. For the location matched negative controls we compiled a list of all variants with a MAF >= 5%, residing 150-1000 bp away from a top eQTL association in the EUR analysis. Variants were filtered for having low LD with the lead association (<= 0.25 r^2) and for having no detectable signal of eQTL association in both the Geuvadis LCL dataset (p-value >= 0.001) and 13 tissues from the GTEx consortium (p-value >= 0.0001). If multiple variants met this criteria at any one loci, only a single site was selected at random for testing. For the randomly selected control variants we picked 2700 sites at random from the EUR population matching the allele frequency distribution of lead variants in the primary EUR eQTL analysis. For 1200 of the 2700 variants we set an additional criteria on requiring eQTL signal in LCLs and GTEx at the same thresholds as the location matched controls. Downstream analysis suggested not differences between the two randomly selected control sets prompting us to combine all 2700 sites together as a single set. All sites were tested in the forward orientation with the flanking sequence taken from hg19. There were 572 additional variants designed on the assay that were unrelated to the positive and control sets that were discarded from the primary analysis.

Massively Parallel Reporter Assay

Oligo Synthesis

Oligos were synthesized by Agilent Technologies as 180 bp sequences containing 150 bp of genomic context and 15 bp of adapter sequence at either end (5'ACTGGCCGCTTGACG [150 bp oligo] CACTGCGGCTCCTGC3') (Figure 1A). Post synthesis (Figure 1B), 20 bp barcodes and additional adapter sequences were added by performing 28 emulsion PCR reactions each 50 µL in volume containing 1.86 ng of oligo, 25 µL of Q5 NEBNext MasterMix (NEB, M0541S), 1 unit Q5 HotStart polymerase (NEB, M0493S), 0.5 uM MPRA_v3_F and MPRA_v3_20I_R primers and 2 ng BSA (NEB, B9000). PCR master mix was emulsified by vortexing with 220 µL Tegosoft (Evonik), 60 µL ABIL WE (Evonik) and 20 µL Mineral Oli (Sigma, M5904) per 50 µL PCR reaction at 4°C for 5 min. 100 µL of Emulsion mixture was plated per well across a 96 well plate and cycled with the following conditions; 95°C for 30 sec, 15 cycles of (95°C for 20 sec, 60°C for 10 sec, 72°C for 15 sec), 72°C for 5min. Amplified emulsion mixture was broken and purified by adding 1 mL of 2-butanol (VWR, AA43315-AK), 50 µL of AMPure XP SPRI (Beckman Coulter, A63881) and 80 µL of binding buffer (2.5M NaCl, 20% PEG-8000) per 350 µL of Emulsion mix and vigorously vortexing followed by incubation for 10 minutes at room temperature. Broken emulsion/butanol mixture was spun at 2900 rcf for 5 min and the butanol phase was discarded. The aqueous phase was placed on a magnetic rack for 20 minutes prior to aspiration. Remaining beads were washed once with 2-butanol, three times with 80% EtOH and eluted in EB (Qiagen, 19086).

MPRA Vector Assembly

To create our mpraΔorf library (Figure 1C), barcoded oligos were inserted into sfiI digested pGL4:23:ΔxbaΔluc by gibson assembly (NEB, E2611) using 1.1 µg of oligos and 1 µg of digested vector in a 40 µL reaction incubated for 60 min at 50°C followed by SPRI purification and elution in 20 µL of EB. Half of the ligated vector was then transformed into 100 µL of 10-beta e.coli (NEB, C3020K) by electroporation (2kV, 200 ohm, 25 µF). Electroporated bacteria were immediately split into eight 1 mL aliquots of SOC (NEB, B9020S) and recovered for 1 hour at 37°C then independently expanded in 20 mL of LB supplemented with 100 µg/mL of carbenicillin (EMD, 69101-3) on a floor shaker at 37°C for 6.5 hours. After outgrowth aliquots were pooled prior to plasmid purification (Qiagen, 12963). For each of the aliquots we plated serial dilutions after SOC recovery and estimated a library size of $>10^8$ CFUs.

To create our final mpra:gfp library (Figure 1D), 20 µg of mpra:Δorf plasmid was linearized with 200 units of AsiSI (NEB, R0630) and 1x cutsmart buffer (NEB) in a 500 µL volume for 3.5 hours at 37°C. An amplicon containing a minimal promoter, GFP open reading frame and a partial 3' UTR was then inserted by gibson assembly using 10 µg of AsiSI linearized mpraΔorf plasmid, 33 µg of the GFP amplicon in 400 µL of total volume for 90 minutes at 50°C followed by a 1.5x beads/sample SPRI purification. The total recovered volume was digested a second time to remove remaining uncut vectors by incubation with 50 U of AsiSI, 5 U of RecBCD (NEB, M0345), 10 µg BSA, 1 mM ATP, 1x NEB Buffer 4 in a 100 µL reaction for 6 hours at 37 C followed by SPRI purification and elution with 55 µL of EB.

To generate transfection ready MPRA libraries 10 µL of mpra:gfp plasmid was electroporated (2kV, 200 ohm, 25 µF) into 220 µL of 10-beta cells. Electroporated bacteria was split across 6 tubes and each recovered in 2 mL of SOC for 1 hour at 37°C then added to 500 mL of LB with 100 µg/mL of carbenicillin and grown for 9 hours at 37°C prior to plasmid purification (Qiagen, 12991). We repeated this same electroporation protocol 3 additional times, each time with an estimated transformation efficiency of $>10^8$ cfu. All plasmid preps were then pooled and normalized to 1 µg/uL to generate our final mpra:gfp library used in all subsequent transfections.

MPRA Transfections

Lymphoblastoid cells were grown in RPMI (Life Technologies, 61870) supplemented with 15% FBS (Life Technologies, 26140) maintaining a cell density of 2-10x10^5 cells per mL. For all 8 transfections (5 x NA12878 and 3 x NA19239) cells were grown to a

density of ~$1x10^6$ cells/mL prior to the removal of $5x10^8$ cells. Cells were collected by centrifugation at 120x g and eluted in 4 mL of RPMI with 500 µg of mpra:gfp library. Electroporation was performed in 100 µL volumes with the Neon transfection system (Life Technologies) applying 3 pulses of 1200 V for 20 ms each. Using separate control transfections we achieved transfection efficiencies of 40-60% for all replicates. Cells were allowed to recover in 180 mL in RPMI with 15% FBS for 24 hours then collected by centrifugation, washed once with PBS, collected and frozen at -80°C (Figure 1E).

Hepatocytes were grown in MEM alpha (Life Technologies, 32561) supplemented with 10% FBS. Cells were plated across ten 15 cm cell culture plates and grown to 60-70% cell density. On the day of transfection media was replaced with 30 mL fresh MEM/FBS followed by transfection with 87.5 µL of Lipofectamine 3000 (Life Technologies, L3000015) and 35 µg of DNA using the manufacturer's protocol. Cells were incubated with transfection reagents for 24 hours, then washed with 15 mL of PBS followed by dissociation with 0.05% trypsin-EDTA (Life Technologies, 25300), centrifugation, PBS wash and a final collection at 300x g prior to storage at -80°C.

RNA Extraction and cDNA Synthesis

Total RNA was extracted from cells using Qiagen Maxi RNeasy (Qiagen, 75162) following the manufacturer's protocol including the on-column DNase digestion. A second DNase treatment was performed on the purified RNA using 5 µL of Turbo DNase (Life Technologies, AM2238) in 750 µL of total volume for 1 hour at 37°C. The digestion was stopped with the addition of 7.5 µL 10% SDS and 75 µL of 0.5M EDTA followed by a 5 minute incubation at 70°C. The total reaction was then used for pulldown of GFP mRNA. Water was added to the DNase digested RNA to bring the total volume to 898 µL to which 900 µL of 20X SSC (Life Technologies, 15557-044), 1800 µL of Formamide (Life Technologies, AM9342) and 2 µL of 100 uM biotin-labeled GFP probe (GFP_BiotinCapture_1-3, Table S3) were added and incubated for 2.5 hours at 65°C. Biotin probes were captured using 400 µL of pre-washed Streptavidin beads (Life Technologies, 65001) eluted in 500 µL of 20X SSC. The hybridized RNA/probe bead mixture was agitated on a nutator at room temperature for 15 minutes. Beads were captured by magnet and washed once with 1x SSC and twice with 0.1x SSC. Elution of RNA was performed by the addition of 25 µL water and heating of the water/bead mixture for 2 minutes at 70°C followed by immediate collection of eluent on a magnet. A second elution was performed by incubating the beads with an additional 25 µL of water at 80°C. A final DNase treatment was performed in 50 µL total volume using 1 µL of Turbo DNase incubated for 60 minutes at 37°C followed by inactivation with 1 µL of 10% SDS and purification using RNA clean SPRI beads (Beckman Coulter, A63987).

First-strand cDNA was synthesized from half of the DNase-treated GFP mRNA with SuperScript III and a primer specific to the 3' UTR (MPRA_v3_Amp2Sc_R) using the manufacturer's recommended protocol, modifying the total reaction volume to 40 µL and performing the elongation step at 47°C for 80 minutes. Single-stranded cDNA was purified by SPRI and eluted in 30 µL EB.

Tag-seq Library Construction

To minimize amplification bias during the creation of cDNA tag sequencing libraries, samples were amplified by qPCR to estimate relative concentrations of GFP cDNA using 1 µL of sample in a 10 µL PCR reaction containing 5 µL Q5 NEBNext master mix, 1.7 µL Sybr green I diluted 1:10,000 (Life Technologies, S-7567) and 0.5 uM of TruSeq_Universal_Adapter and MPRA_Illumina_GFP_F primers (Table S3). Samples were amplified with the following conditions: 95°C for 20 seconds, 40 cycles (95°C for 20 sec, 65°C for 20 sec, 72°C for 30 sec), 72°C for 2 min. All LCL cDNA samples had a cycle threshold of approximately 11 while HepG2s showed an earlier cycle threshold corresponding to the larger amount of RNA recovered. To add Illumina sequencing adapters, cDNA samples and 5 mpra:gfp plasmid controls were diluted to match the replicate with the lowest concentration and 10 µL of normalized sample was amplified using the reaction conditions from the qPCR scaled to 50 ul, excluding Sybrgreen I and using only 10 amplification cycles. Amplified cDNA was SPRI purified and eluted in 40 µL of EB. Individual sequencing barcodes were added to each sample by amplifying the entire 40 µL elution in a 100 µL Q5 NEBNext reaction with 0.5 uM of TruSeq_Universal_Adapter primer and a reverse primer containing a unique 8 bp index (Illumina_Multiplex) for sample demultiplexing post-sequencing. Samples were amplified at 95°C for 20 seconds, 6 cycles (95°C for 20 sec, 64°C for 30 sec, 72°C for 30 sec), 72°C for 2 minutes. Indexed libraries were SPRI purified and pooled according to molar estimates from Agilent TapeStation quantifications. Samples were sequenced using 1x30 bp chemistry on an Illumina HiSeq through the Broad Institute's walk-up sequencing service.

To determine oligo/barcode combinations within the mpra pool, Illumina libraries were prepared from the mpraΔorf plasmid library by performing 4 separate amplifications with 200 ng of plasmid in a 100 µL Q5 NEBNext PCR reaction containing 0.5 uM of TruSeq_Universal_Adapter and MPRA_v3_TruSeq_Amp2Sa_F primers with the following conditions: 95°C for 20 sec, 6 cycles (95°C for 20 sec, 62°C for 15 sec, 72°C for 30 sec), 72°C for 2 minutes. Amplified material was SPRI purified using a 0.6x bead/sample ratio and eluted with 30 µL of EB. Sequencing indexes were then attached using 20 µL of the eluted product and the same reaction conditions as for the tag-seq except the number of enrichment cycles were lowered to 5. Samples were molar pooled and sequenced using 2x150 bp chemistry on Illumina HiSeq and NextSeq instruments through the Broad Institute's walk-up sequencing service.

7.5k Oligo MPRA Experiment
To perform the MPRA experiment of the 7.5k library we adjust the experimental conditions to 1/10th the scale used for the 79k library with the following exceptions. The two gibson assembly steps were performed at ¼ scale of the original library and RNA extraction was performed using Qiagen Midi RNeasy (Qiagen, 75142) followed by ½ scale reactions for the GFP pulldown and cDNA synthesis. Library preparation was performed as described above but diluting the samples to the cycle threshold specific for the 7.5k library.

Single Oligo Validation
To validate the expression values obtained by MPRA we selected 29 oligos consisting of 13 ref/alt pairs and 3 additional oligos. We selected the oligos based upon their association with an eQTL/GWAS peak while maintaining diverse representation of regulatory strength. Two of the oligos were selected as no-expression controls for having uncorrected p-values of greater than 0.01. We designed the same 150 bp sequence that was tested by MPRA as a gBlock (IDT) and cloned each into the pGL4.23 firefly luciferase reporter vector (Promega, E8411). We initially performed a standard luciferase reporter assay, co-transfecting $1x10^6$ LCLs (NA12878) with 1ug of the cloned firefly luciferase vector and 200 ng of a renilla luciferase control vector (pGL4.74, Promega, E6921) and recovered for 24 hour in a 96 well plate. We performed three separate experimental replicates each with two transfection replicates per experiment for a total of six replicates per oligo. Firefly luciferase luminescence for each well was normalized to the renilla luciferase luminescence for the same well and each experiment was normalized as a log-ratio value relative to the mean of the two no-expression control oligos.

Initial analysis of the luciferase expression strength to the MPRA expression values showed moderate correlation for a portion of the oligos but also displayed discordant values for many sites. Under the hypothesis that some oligos carry novel transcription start sites causing out of frame transcription of the luciferase loci, we performed qPCR as the end point for the luciferase assay instead of luminescence. Using the same transfection protocol as the luciferase assay, we performed two transfection replicates for each oligo and extracted mRNA 24 hours post transfection using the MagMAX 96 Total RNA isolation kit (Life Technologies, AM1830). DNA was digested by incubation with 2U of Turbo DNase for 60 minutes followed by RNA SPRI purification. qPCR was performed in replicate according to the manufacturer's recommendations on the purified RNA using 1-step Sybr-to-Ct and gene-specific qPCR primers (Fluc_F/Fluc_R or Rluc_F/Rluc_R) to measure both the firefly and renilla luciferase RNA levels, with pGL4.23 and pGL4.74 plasmids serving as standards for copy number calculations. For each replicate a firefly/renilla ratio was calculated as well as a log-ratio value relative to the mean of the 2 negative control samples.

Data Analysis

Barcode/Oligo Reconstruction
Paired-end 150 bp reads from the sequencing of the mpraΔorf library were merged into single amplicons using Flash v1.2.7 (flags: -r 150, -f 220, -s 10) (Magoč and Salzberg, 2011). Amplicon sequences were kept if the 5' adapter matched with a levenshtein distance of 3 or less and 2 bp at the edges of both the 5' and internal constant sequences matched perfectly. Oligo sequences from the passing reads were then mapped back to the expected oligo sequences using BWA mem version 0.7.9a (flags: -L 100 -k 8 -O 5) (Li, 2013). Alignment scores were calculated as matching bases divided by the expected oligo size and reads with alignment scores of less than 0.95 were discarded. Remaining oligo/barcode pairs were then merged and barcodes attributable to multiple oligo sequences were marked as conflicting and removed from further analysis. In total we observed 90.2 million unique barcodes in the sequencing data.

Identification of Regulatory Oligos
Reads from the tag sequencing were filtered for the inclusion of the constant sequence within the GFP 3' UTR. Specifically, a levenshtein distance of 4 or less was required within the constant sequence at the end of the tag-seq read with the two bases directly adjacent to the barcode (base 21 & 22) required to match perfectly. Barcodes were then matched with oligo sequences determined through sequencing of the mpraΔorf library and the sum of all barcodes counts within each of the 78,956 oligos were calculated.

Oligo counts from all 18 samples (5 plasmid controls, 5 NA12878, 3 NA19239 and 5 HepG2) were passed into DESeq2 and a median-of-ratios method was used to normalize samples for varying sequencing depths (Love et al., 2013). Normalized read counts of each oligo were then modeled by DESeq2 as a negative binomial distribution (NB). DESeq2 estimates variance for each NB by pooling all oligo counts across samples and fitting a trend line to model the relationship between oligo counts and observed dispersion. It then applies an empirical Bayes shrinkage by taking the observed relationship as a prior and performing a maximum a posteriori estimate of the dispersion for each oligo.The overall result is that DESeq2 can obtain an estimate for dispersion of each oligo with greatly reduced bias by pooling information from all oligos.

We then used DESeq2 to estimate the fold change estimation between the control condition (plasmid) and each of the three experimental conditions (NA12878, NA19239, and HepG2). Again, DESeq2 applies a Bayesian shrinkage on the log ratios to prevent

false positive results at the extreme ends of expression (low and high count oligos). We use Wald's test to estimate significance for expression differences between conditions and corrected for multiple hypothesis testing by Bonferroni's method accounting for 39,479 tests. We required a corrected p-value of 0.01 or less in either the reference or alternate allele in order to call a sequence as having a regulatory effect on expression.

Identification of Expression-Modulating Variants

For the identification of variants altering expression strength we considered only variants originating from sequences determined to have a regulatory effect. We calculated p-values for allelic skew by comparing the log ratios of the reference and alternate alleles using a paired t-test with independent estimation of variance and Welch's adjustment to the degrees of freedom. This test assumes normality; to evaluate normality, we calculated z-scores for the differences of the log ratios for all 39,479 ref/alt pairs and observed a distribution very similar to that expected from sampling ratios from a normal distribution (Figure S4B & C). We further validated our approach by evaluating the qq plot for variants that failed to show regulatory effects (uncorrected expression p-value > 0.01), which are not expected to have any allelic bias (Figure S4D & E). To collapse results from the two LCL experiments, we averaged both the expression ratio and allelic skew weighted by the number of replicates from each sample; p-values were combined using Fisher's method. For all samples as well as the combined LCL analysis FDR was calculated for the skew p-value using the Benjamini-Hochberg procedure.

Downsampling and Analysis of the 7.5k Oligo Library

Analysis of the 7.5k oligo MPRA experiment was performed as described for the initial 79k library applying FDR cutoffs for expression and skew calling that matched those originally used with the 79k library. Initial analysis of the full dataset showed the smaller library had greater power to detect weaker expression changes than the original library due to a 2.5-14 fold increase in the median number of barcodes tagging each oligo (Figure S3A). As expected from our previous observation of the effect of barcodes, we detected lower dispersion in the 7.5k oligo pool as estimated by DESeq2 (Figure S3B). Therefore, to better match the two libraries, we downsampled the 7.5k dataset to match the median number of barcodes representing each oligo in the 79k pool while maintaining the rank and distribution for each 7.5k experimental replicate. Specifically, we paired replicates between the two pools, calculated the ratio of the median number of barcodes per oligo in the old to its pair in the new, and sampled without replacement that proportion of the original number of barcodes for each oligo. We applied this to each replicate and repeated 500 times, each time calculating summary statistics including the number of expression positive variants and the number of emVars per subsampling. In the text, we report the mean of these 500 experiments.

This procedure caused both the dispersion and raw counts to better reflect the 79k experiment. After downsampling, we saw no loss in our ability to call both expression and emVar positive controls included in the new library compared to the full dataset with 96% of expression variants and 72% of emVars detected (96% and 71% detected in the full dataset). In addition, Correlation was strong between predicted expression effects between the two experiments. We note that despite extensive downsampling, the dispersion was still lower for the 7.5k library than in the 79k library. As a result, it is probable that the 7.5k library still maintains a slightly higher sensitivity to detect expression effects than the original 79k library resulting in conservative false positive estimates. This conclusion is supported by the lower effect size of expression positive sites detected in all three negative control sets relative to the 264 expression positive controls.

Annotations Used for Enrichment Analysis

For enrichment analysis we downloaded narrowPeak files from the ENCODE project's FTP server (Table S4). All enrichment analysis for regulatory oligos required an overlap of 1 bp or greater with an annotation at any position along the oligo. For analysis using LCL DHS regions, unless otherwise specified we took the union of all LCL regions from UW, Duke and the Unified analysis. For the annotation positive designation we required the oligo to be overlap 2 of the following 4 categories; all TF-ChIP data (including POL), LCL DHS regions, CAGE regions for NA12878 (Nucleus, Cytosol and Cell) and chromatin marks (H3k27ac,H2A.Z,H3K4me2,H3K4me3,H3K9ac). All fold enrichments are reported as odds-ratios unless otherwise stated.

For enrichments involving promoter proximity we defined transcription start sites (TSS) by analyzing the predicted transcript abundance within our mapped RNA-seq reads. Using the cufflinks generated estimation of transcript abundance we identified genes with an average FPKM of 0.5 or greater across all LCLs and selected TSSs from transcripts within 50% of the most abundant transcript's FPKM. We counted an overlap with the core promoter when the variant fell within 100 bp upstream and 50 bp downstream of the TSS. A hypergeometric test was used to evaluate significance of all emVars relative to either the proportion promoter sites in all 29k variants tested or genome wide (all variants >=5% minor allele frequency in EUR). For strongly linked GWAS associated eQTL peaks (>= 0.9 r^2) we first tested enrichment of promoter emVars relative to all other emVars falling outside these peaks. Significance was calculated by Fisher's exact test splitting all variants into two categories; those strongly linked to a GWAS SNP (r^2 of 0.9 or greater) and all non-gwas associated variants, and testing the number of emVars falling in a promoter compared to the total number of promoter variants. To verify our analysis is not confounded by the differences in how the variants were selected (variants with an r^2 of 0.9 or greater to an eQTL peak compared to perfect LD) we performed the same analysis but

using only the GWAS eQTL peaks. We split these variants into two groups based on the maximum LD value to a GWAS SNP. Using 0.9 r^2 as a cutoff we then used Fisher's exact test to calculate significance of enrichment with promoter sites.

Transcription Factor Binding Sites
We used FIMO and HOCOMOCO v9 to calculate binding scores for the reference and alternative allele in all 29k oligos (Grant et al., 2011; Kulakovskiy et al., 2012). We identified SNPs and indels where the motif had a binding p-value of $1x10^{-5}$ or less and the predicted TF showed binding in the analogous ENCODE ChIP-seq experiments. From this list we calculated the difference in binding score between the reference and alternate allele for each TF predicted to bind. Where sites had multiple predicted binding partners we selected the TF with the greatest change between the the alleles. We binned the variants based on the allele difference score and calculated the proportion these variants represented within the three classes of function (emVars, regulatory/non emVar, not significant).

Sensitivity Estimates
For the majority of eQTL peaks, only the top associated variant (and variants in perfect LD) were tested by MPRA. Therefore, an emVar may not be detected for a given eQTL peak for one of two reasons:

(1) The causal variant was not among the top associated variants for the eQTL peak and so was not tested
(2) The causal variant was tested but the MPRA assay gave a false negative

These two reasons for failure to detect an emVar in an eQTL peak correspond to two distinct sensitivity estimates, a technical sensitivity (2 alone), and the power of the study design to detect a causal variant (1 and 2 together). To estimate the power of experimental design, we first estimated the number of true positive emVars in EUR and YRI peak independently.

$$TP = (V) - N*(1-specificity)$$

Where V is the number of variants identified by the MPRA as emVars, N is the total number of variants tested across the peaks and a specificity of 99.04%. We then simulated the number of eQTL peaks explained within EUR and YRI when selecting the specified number of emVars (TP) randomly from the list of MPRA+ variants.

To estimate the MPRA's technical sensitivity, we first note that the probability of the causal variant being among those with the maximum r^2 in the peak should increase with the difference in the variance explained (Δr^2) between the top associated variant (and variants in perfect LD) and the second best association for each eQTL peak. To quantify this relationship, we fit a logistic regression to model the effect of a peak's Δr^2 and the effect size of the top associated SNP (ES) on the probability of detecting an emVar in that peak. Specifically, we fit the following regression:

$$\text{logit}(P(\text{emVar})) = \beta_2 \ln(\Delta r^2) + \beta_1 |ES| + \beta_0$$

The regression was fit using the statsmodels toolbox in Python 3.3 (https://github.com/statsmodels/statsmodels). The natural log-transformation was used to capture the expected convexity of the covariates. Using this model, we estimated the technical sensitivity as a function only of effect size alone by setting $\Delta r^2=1$, corresponding to the maximum possible separation between the top and second to top variant.

Detection of Allelic Skew in DHS and ChIP-seq Data
Detecting allelic skew in DHS and ChIP-seq datasets is challenging because mapping of short reads produces a bias toward the reference allele, confounding measurements of skew. To circumvent this problem, we constructed personal reference genomes for the maternal and paternal haplotype of each of the lymphoblastoid cell lines for which we had DHS or ChIP-seq data using the software vcf2diploid (Rozowsky et al., 2010). We then aligned DHS or ChIP-seq data to the corresponding personal reference genomes using BWA aln/samse with the default parameters (Li and Durbin, 2009). We then filtered out aligned reads with a mapping quality of 0 and obtained the set of reads overlapping heterozygous genomic variant using bedtools (Quinlan and Hall, 2010). Finally, we obtained allele counts for reads overlapping each variant by counting the reads that mapped only to the maternal or paternal reference, or that mapped with a better alignment score to one reference than the other. Alignment scores were calculated as a -1 penalty per mismatched base and a -1 penalty for each base-pair difference of an indel. Reads with an equivalent alignment score when aligned to the maternal and paternal reference were discarded.

For calculating skew in DHS, allele counts were pooled across all replicates and all LCLs that were heterozygous for a given variant. All variants with a pooled-coverage greater than 20 with a count of at least 1 for each allele were scored for DHS skew. A p-value cutoff of 0.1 was used. For figure 5C, a coverage cutoff of 10 and a p-value cutoff of 0.1 was used. In addition, variants that showed a substantial fraction of poorly mapped reads for one allele but not the other were discarded.

For calculating skew in transcription factor binding, allele counts were pooled across all LCL replicates for samples that were heterozygous for a given variant. All variants with a pooled-coverage greater than 20 for a transcription factor with a count of at least 1 for each allele were scored for TF skew. A p-value cutoff of 0.001 was used for calculating enrichment as well as for figures S5A & B.

Allelic Replacement of rs9283753

We performed CRISPR editing in LCL to alter the rs9283753 SNP at PTGER4. We used a modified version of the pX458 cas9 vector with U6:gRNA and F1 ori removed and delivered the gRNA as a 455 bp dsDNA amplicon (Mali et al., 2014; Ran et al., 2013). A guide sequence was designed that overlapped rs9283753 at positive 6 of the gRNA (Table S3). Two versions of the guide were created changing only position 6 to match the variant to be edited. Off-target cutting was assessed by *in silico* mapping with no appreciable off-target sites identified. Homology repair templates were synthesized as 150 bp PAGE purified Ultramer oligonucleotides (IDT) with a phosphorothioate bind at the predicted cutting location of cas9 (Table S3). We used two cell lines for allelic replacement; NA12878 which is homozygous for the ancestral allele (alt, T) and NA11831 which is homozygous for the derived (ref, C). We electroporated $5x10^6$ cells with 4 ug of cas9 vector, 400 ng of gRNA and 500 nM of the homology vector using the Neon system. Cells were sorted 24 hours post-transfection for GFP expression using a MoFlo Astrios EQ Cell Sorter (Beckman Coulter) at the Harvard University Bauer Core Facility.

Clonal populations of edited cells were isolated by single-cell dilution into 384-well plates. Genotypes of the clones were determined by Illumina sequencing of a the genomic regions surrounding the SNP using primers rs9283753_ILMN_F and rs9283753_ILMN_R (Table S3). Cell lysate were obtained for successful clonal populations after two weeks of growth by lysing 100 ul of cells in 100 ul of lysis buffer (100 mM KCl, 2mM EDTA, 20 mM Tris-HCl, 2% Triton-X 100, 2% TWEEN 20, 0.08 U/ul proteinase K) at 55 C for 60 minutes followed by 95 C for 10 minutes. PCR was performed on an aliquot of the lysate using Q5 Hot Start Master Mix (NEB). After amplification of the genomic region, each clonal amplicon had unique indices for sequencing added by PCR. Amplicons were sequenced on a MiSeq (Illumina) with 2x150 bp reads. Clones that showed the edited genotypes after analysis were confirmed by Sanger sequencing.

To quantify changes in expression of the PTGER4 gene after CRISPR editing, we performed qPCR comparing edited cells to wild-type cells that had undergone the same cas9/clonal expansion process. Cells were seeded at $2.5x10^5$ cells/mL 24 hours prior to RNA isolation. For RNA isolation, $7.5x10^5$ cells were collected and RNA was isolated with the MagMAX-96 Total RNA Isolation Kit (Life Technologies) according to manufacturer's instructions. cDNA for each sample was produced from 1 ug of isolated RNA using the SuperScript III First Strand Synthesis System (Life Technologies). qPCR were performed with PowerUp SYBR Green Master Mix (Life Technologies), 10 ng of cDNA, and forward and reverse primers at 500 nM each in a total volume of 10 ul. Technical triplicates were performed for each reaction. For each edited cell line two biological (independent seedings) were performed, for the negative control samples (wild-type) 4 and 3 independent clonal populations were tested for NA12878 and NA11831 respectively. Expression values for two seperate primer pairs for PTGER4 were averaged together and normalized using the ΔΔCt method using PPIA and TBP as references (Primers listed in Table S3).

References

Consortium, T. An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Grant, C., Bailey, T., and Noble, W. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics (Oxford, England) *27*, 1017–1018.

Harrow, J., Frankish, A., Gonzalez, J., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research *22*, 1760–1774.

Howie, B., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics *5*, e1000529.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2012). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology *14*, R36.

Kulakovskiy, I., Medvedeva, Y., Schaefer, U., Kasianov, A., Vorontsov, I., Bajic, V., and Makeev, V. (2012). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Research *41*, D195–202.

Lappalainen, T., Sammeth, M., Friedländer, M., Hoen, P., Monlong, J., Rivas, M., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Li (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754–1760.

Love, M., Huber, W., and Anders, S. (2013). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Magoč, T., and Salzberg, S. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics

(Oxford, England) *27*, 2957–2963.
Quinlan, A., and Hall, I. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841–842.
Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2010). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Molecular Systems Biology *7*, 522.
Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Computational Biology *6*, e1000770.
Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology *28*, 511–515.