RESEARCH ARTICLE

HUMAN GENOMICS

Multiple causal variants underlie genetic associations in humans

Nathan S. Abell[1]*, Marianne K. DeGorter[2], Michael J. Gloudemans[3], Emily Greenwald[1], Kevin S. Smith[2], Zihuai He[4,5], Stephen B. Montgomery[1,2]*

Associations between genetic variation and traits are often in noncoding regions with strong linkage disequilibrium (LD), where a single causal variant is assumed to underlie the association. We applied a massively parallel reporter assay (MPRA) to functionally evaluate genetic variants in high, local LD for independent cis-expression quantitative trait loci (eQTL). We found that 17.7% of eQTLs exhibit more than one major allelic effect in tight LD. The detected regulatory variants were highly and specifically enriched for activating chromatin structures and allelic transcription factor binding. Integration of MPRA profiles with eQTL/complex trait colocalizations across 114 human traits and diseases identified causal variant sets demonstrating how genetic association signals can manifest through multiple, tightly linked causal variants.

Genome-wide association studies (GWASs) have emerged as an important tool with which to assess the effect of individual genetic variants on phenotypes, ranging from gene expression to complex traits and diseases (*1*, *2*). However, because of linkage disequilibrium (LD), it is challenging to identify a single causal variant among multiple correlated variants. To address this challenge, statistical and functional fine-mapping approaches have been developed to identify credible sets of variants that contain the causal variant (*3*). However, these approaches often cannot distinguish between proximal or highly linked variants and lack systematic prior information on the number of causal variants underlying association signals.

One approach to systematically identify causal variants while controlling for LD is applying massively parallel reporter assays (MPRAs). MPRAs measure the effects of synthetic DNA libraries on the expression of a reporter gene, typically luciferase or green fluorescent protein (GFP), containing a 3′ untranslated region (UTR) barcode (*4*). Such assays have screened potential regulatory elements in diverse cellular contexts and also have applications in saturation mutagenesis or tiling along regulatory regions of interest (*5*–*7*).

Beyond tests of regulatory function, MPRAs have also been applied to assay the differential regulatory effects of genetic variants (*8*–*10*). However, existing studies have either targeted variants with the strongest trait associations and/or applied extensive prior filtering limiting resolution of linked causal variants (*8*, *9*, *11*, *12*). In the yeast *Saccharomyces cerevisiae*, quantitative trait loci (QTL) mapping has identified loci containing multiple causal variants in tight LD, suggesting that the same genetic architecture may also underlie many human traits (*13*, *14*).

[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA. [2]Department of Pathology, School of Medicine, Stanford University, Stanford, CA 94305, USA. [3]Biomedical Informatics Program, Stanford University, Stanford, CA 94305, USA. [4]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA. [5]Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: nsabell@stanford.edu (N.S.A.); smontgom@stanford.edu (S.B.M.)

Results

Functional fine-mapping of eQTL reproducibly identifies regulatory and allelic hits

We applied an MPRA to systematically characterize causal variants underneath multiple expression QTL (eQTL) and GWAS loci. We selected independent, common, and top-ranked eQTL across 744 eGenes identified in the CEU cohort (which comprises Utah residents of Northern and Western European ancestry). Each eQTL had a median of six lead associated variants (range of 1 to 472) in perfect LD. For each lead variant, we identified all additional variants with a correlation coefficient (r^2) ≥ 0.85 that were associated with the same gene, as well as a set of variants (n = 2114 non-eQTLs) that were not associated with any gene's expression. Our final library included 30,893 variants, with a median of 50 variants per eQTL (range of 2 to 2824) (Fig. 1A).

For each variant, we identified 150–base pair (bp) sequences (centered on the variant) and generated a MPRA library by random barcoding (Fig. 1B). For allelic pairs, the fragment lengths and surrounding sequence were held constant to allow measurement of allele-specific effects. For indels, fragment lengths between allelic pairs differed by less than 9 bp. Furthermore, in sequences with multiple variants, distinct oligonucleotides (oligos) were designed for each possible haplotype, resulting in an average of 3.19 oligos per variant. Overall, this resulted in an assay of 49,256 total allelic pairs. After reporter gene insertion, the library was transfected into lymphoblastoid cell lines (LCLs) in triplicate, sequenced, and then quantified for each oligo.

To measure regulatory effects from oligo counts, we used negative binomial regression. For each variant, we computed the allele-independent regulatory effects of an oligo ("expression" effects) and the difference in regulatory effects between reference and alternative allele-containing oligos ("allelic" effects). We detected 8502 expression effects and 1264 allelic effects across all tested variants.

We observed a modest increase in the total number of MPRA hits in eQTLs relative to non-eQTLs (27 versus 26% for expression hits and 9 versus 8% for allelic hits), reflecting the low proportion of eQTL variants overall that are expected to be causal (Fig. 1, C and D). We observed a larger increase in allelic effect sizes among hits that are also eQTL versus non-eQTL (fig. S1D). This was the case when comparing MPRA hits between eQTL and non-eQTL for both expression effects [Kolmogorov-Smirnov (K-S) $P = 1.704 \times 10^{-4}$) and allelic effects (K-S $P = 0.0116$). Taken together, we obtained for each eQTL gene (eGene), a profile of allele-independent and -dependent effects across all highly associated proximal variants (Fig. 1E).

By design, a subset of tested variants (n = 782) were previously identified as expression-modulating variants in (*8*). This overlapping subset was highly enriched for expression and allelic effects (Fig. 1, C and D). Further, we observed that 89.6% of allelic MPRA hits in both datasets were directionally concordant (fig. S2A). From these results, we constructed a concordant, high-confidence "MPRA-positive" variant set that contains 250 variants with expression effects and 120 with allelic effects (fig. S2, B and C).

Diverse transcription factor programs contribute at eQTL

The large number of MPRA expression effects enabled identification of transcription factors (TFs) that affect gene expression within eQTLs. We observed widespread positive enrichment of chromatin immunoprecipitation–sequencing (ChIP-seq) peaks for multiple TFs in MPRA expression effects (n = 160 total TFs). Moreover, applying a more stringent filter (adjusted $P \leq 5 \times 10^{-10}$) increases these enrichments in most TFs (Fig. 2A and table S5). Although enrichments vary across a broad range (1.2- to 17-fold), many enriched TFs are members of the same family and exhibit highly correlated genome-wide binding profiles. This demonstrates the wide range of regulatory element

effects captured in our assay and pinpoints specific TFs that drive the regulatory effects of genetic variation.

We next evaluated histone modifications and observed enrichments for activating histone modifications but not for repressive marks such as histone H3 lysine 36 trimethylation (H3K36me3) (Fig. 2B). We also observed the strongest enrichments in chromatin accessibility regions that were tissue invariant or specific to the stromal A (representing JDP2 and other AP-1 TF families), lymphoid, and erythroid/myeloid tissue clusters, demonstrating detection of cell-type information encoded in accessible chromatin (Fig. 2C).

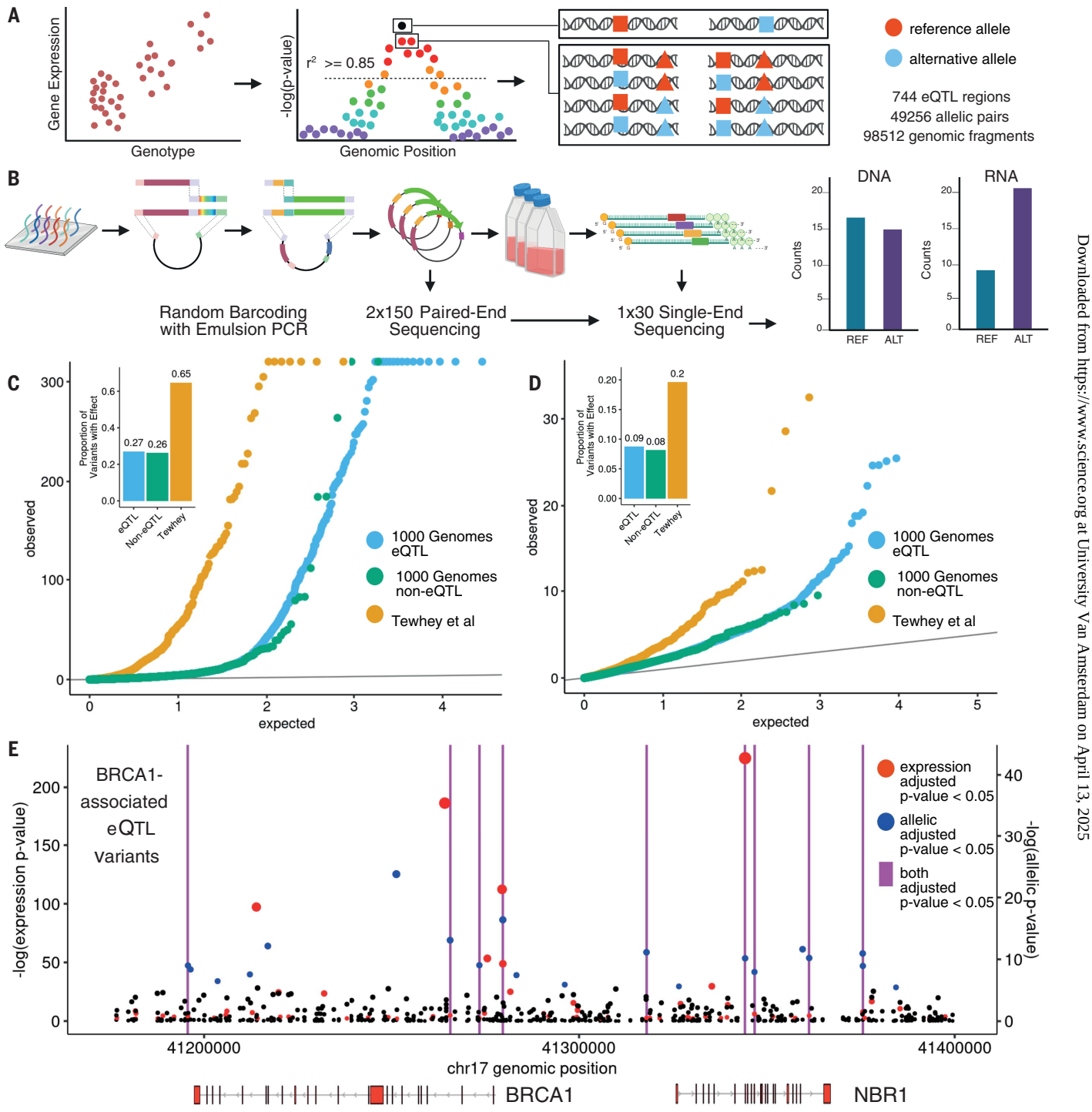

Fig. 1. Design and implementation of a variant-based MPRA. (**A**) Variant selection and oligonucleotide sequence design. (**B**) Random barcoding, sequencing, and expression of the MPRA library. (**C**) Distribution of eQTLs (orange) and non-eQTLs (blue) from the 1000 Genomes Project compared with Tewhey *et al*. (green) (*8*) variant expression *P* values (negative binomial regression) and relative effect proportions. (Inset) Proportion of tested variants that are significant MPRA hits. (**D**) Same as in (C) but with allelic *P* values (negative binomial regression). (**E**) Genomic position and unadjusted *P* values for all tested breast cancer 1 (BRCA1)–associated variants, with colors indicating Benjamini-Hochberg (BH) adjusted $P \leq 0.05$. Vertical magenta lines indicate positions of variants that are both expression and allelic MPRA hits.

Identifying regulatory variants through allelic transcription factor binding and chromatin accessibility

To identify specific TFs affected by regulatory variation, we first characterized whether the direction of allelic MPRA hits was concordant with single-nucleotide polymorphism evaluation by systematic evolution of ligands by exponential enrichment (SNP-SELEX) scores, a set of allele-specific binding models created from in vitro TF binding affinities. We observed global concordance (Fisher's exact $P = 3.43 \times 10^{-15}$) that was absent in other tested sites (Fisher's exact $P = 0.63$) (Fig. 2D).

Next, we computed the concordance proportion (how often a SNP-SELEX score for a specific TF was concordant with the MPRA allelic effect) for all TFs that overlapped at least three tested variants (Fig. 2E). The mean concordance proportion across TFs ($n = 59$) was 0.733 when using allelic MPRA hits and 0.505 when using other tested sites ($n = 91$) [binomial logistic generalized linear model (GLM) $P = 8.46 \times 10^{-12}$]. Although allelic MPRA hits were enriched in SNP-SELEX variants, only 13.5% of SNP-SELEX variants had an MPRA effect. This suggests that many allelic effects can be explained by altered TF binding, but altered binding itself does not typically affect transcription.

A similar pattern emerged when comparing allelic imbalance in accessible chromatin with MPRA allelic hits. We observed significant

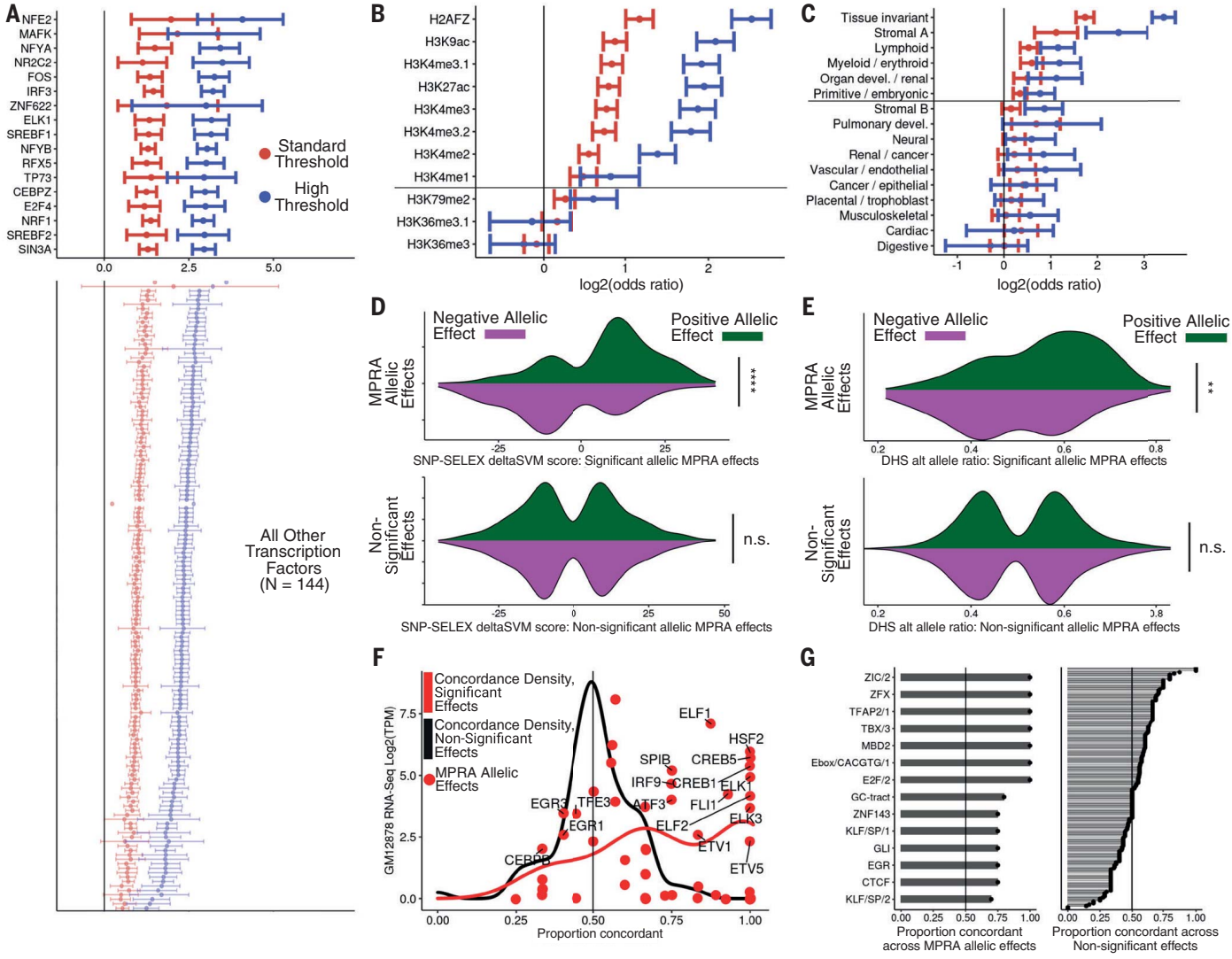

Fig. 2. General and allele-specific functional properties of regulatory variants. (**A**) Odds ratios and 95% confidence intervals for enrichment of peaks from 160 ENCODE ChIP-seq datasets within expression MPRA hits. Standard and high thresholds required an expression BH-adjusted $P < 5 \times 10^{-2}$ or $P < 5 \times 10^{-10}$, respectively. Only TFs with an enrichment adjusted $P < 0.005$ are shown, and listed TFs have BH-adjusted enrichment $P < 0.05$ and odds ratio > 5 (Fisher's exact test). (**B**) Same as in (A) but for histone modifications. Marks above the horizontal line have a BH-adjusted $P < 0.05$ at both thresholds. (**C**) Same as in (A) but for clustered chromatin accessibility regions in fragments with expression effects. (**D**) Distribution of SNP-SELEX deltaSVM scores at allele-specific binding variants stratified by MPRA allelic hit direction (color) and significance category (top and bottom). MPRA allelic hits have BH-adjusted expression and allelic $P \leq 0.05$, whereas nonsignificant variants have $P > 0.75$ (negative binomial regression). (**E**) Same as in (D) but for allelic imbalance in chromatin accessibility from ENCODE. (**F**) For all TFs evaluated in (D), comparison of the concordance proportion across MPRA variants with the expression of each included TF in GM12878 cells. Points indicate significant effect concordances. (**G**) Comparison of directional concordances within accessible chromatin motifs for (left) significant and (right) nonsignificant MPRA effects. Significance values for (C) and (D) were calculated with Fisher's exact test; $^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$, $^{****}P < 5 \times 10^{-5}$.

concordance between allelic imbalance and MPRA allelic effect directions for allelic MPRA hits but not other variants (Fisher's exact $P = 7.33 \times 10^{-3}$ and $P = 0.839$, respectively) (Fig. 2F). Separation by functional footprints found within accessible chromatin regions revealed that several motifs, including Gli and a canonical E-box, were concordant across all allelic MPRA hits (Fig. 2G).

To further assess the relationship between regulatory variants and chromatin accessibility, we integrated chromatin accessible QTL (caQTL) data to identify variant annotations that increased MPRA signals. Using Encyclopedia of DNA Elements (ENCODE) allelic imbalance data, we separated all variants by whether they were inside or outside an associated peak. MPRA allelic hits were strongly concordant with allelic imbalance when inside their peaks but not when adjacent to them (Fisher's exact $P = 3.2 \times 10^{-5}$ and $P = 0.055$, respectively) (fig. S3A). Separately, in a set of caQTLs assessed across 10 population groups, variants that were caQTLs in multiple populations were more enriched in MPRA allelic hits than caQTLs shared in only a few populations (fig. S3B). Taken together, MPRA allelic hits were concordant with in vitro and in vivo measures of allelic regulatory activity, whereas other tested sites were directionally random.

MPRAs inform noncoding variant effect prediction

An ongoing challenge is to summarize and predict the regulatory effect of noncoding variants by using sequence and annotation alone. We evaluated whether genome-wide variant effect predictors could identify allelic MPRA hits. Using scores from Enformer, a neural network that predicts variant effects by incorporating sequence information, we observed significant enrichment of allelic MPRA hits in the top percentiles of Enformer scores (K-S test) (Fig. 3; A, inset, and B) (*15*). We next assessed all tested variants with their annotation principal components (aPCs) from FAVOR, an integrated variant effect prediction tool (*16*). We again observed enrichment of allelic MPRA hits for multiple aPCs. These enrichments were strongest for the TF and epigenetics-based aPCs, whereas others such as distance from transcription start site/transcription end site (TSS/TES) were similarly enriched in both allelic MPRA hits and all other tested sites (K-S test) (Fig. 3; C, inset, and D).

Both predictors could distinguish eQTL regions from genomic background; we also observed a positive enrichment in allelic MPRA hits up to the 50th percentile of these scores, with increasing enrichment at very high percentiles (Fig. 3, A and C). Despite this overlap, when comparing allelic MPRA hits with other tested variants, the distributions of all Enformer PCs and FAVOR aPCs, except for Distance-to-TSS/TES, were significantly different (K-S test) (Fig. 3, B and D). This suggests why functional fine-mapping approaches have not always benefitted from noncoding variant effect predictions while also showing that the highest genome-wide percentiles of these scores identify variants enriched for MPRA allelic effects.

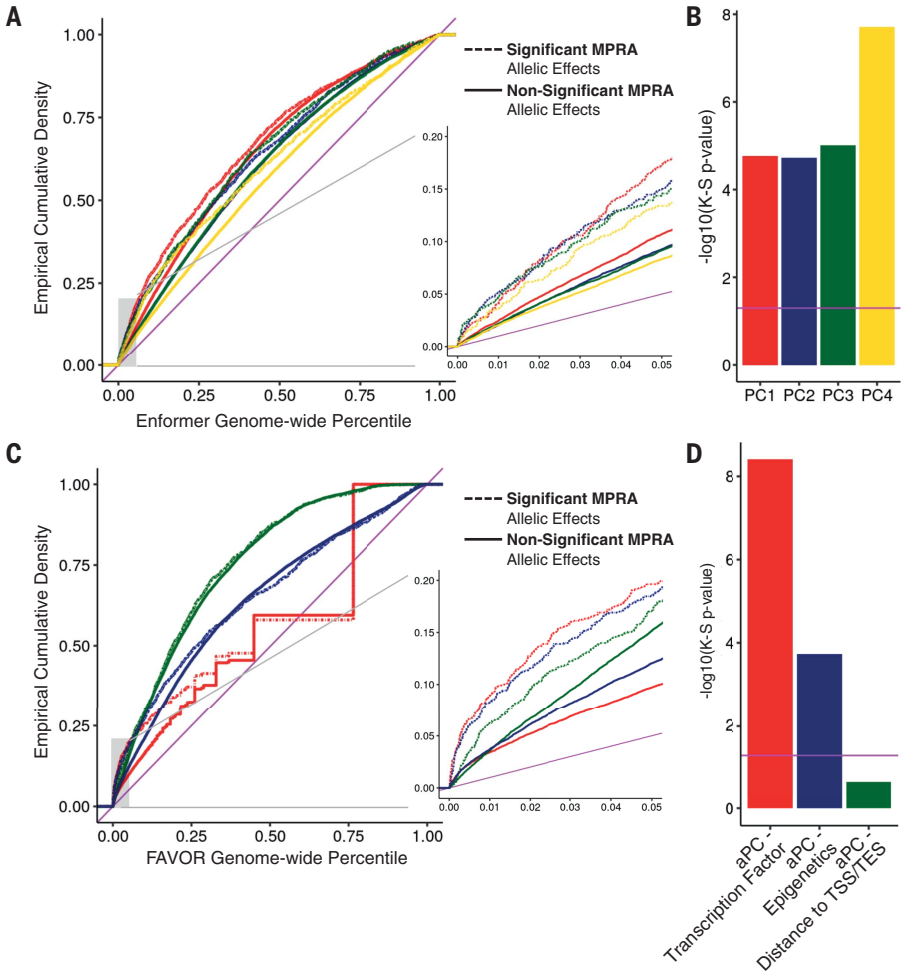

Fig. 3. Integrative noncoding variant effect prediction. (**A**) Empirical cumulative probability distribution of the first through fourth principal components (PC) scores from Enformer for allelic MPRA hits and other tested variants significant and nonsignificant MPRA allelic hits. Genome-wide percentiles were computed across all common variants in 1000 Genomes Phase 3. (Inset) A blow up of lower genome-wide percentile curves. (**B**) Significance of a K-S test comparing the empirical distributions of Enformer scores for significant and nonsignificant allelic MPRA hits v; magenta horizontal line indicates significance by K-S test ($P < 0.05$). (**C**) Same as in (A) except showing annotation principle components from FAVOR. Genome-wide percentiles were computed across all variants in TOPMed Freeze5. (**D**) Same as in (B), except testing FAVOR aPCs.

Multiple causal regulatory variants in high LD underlie eQTL

To fine-map regulatory variants, we assessed MPRA hits within eQTLs. Across all loci, 76.7% (571 of 744) and 45.6% (339 of 744) had at least one expression or allelic MPRA hit, respectively (Fig. 4, A and B); 17.7% (132 of 744) had more than one allelic MPRA hit, indicating that an appreciable number of genetic associations contain multiple regulatory variants in high LD (Fig. 4B). Of allelic hits, 69% were in perfect LD in Europeans from the 1000 Genomes Project, which limited the use of statistical approaches (Fig. 4C). Even when additionally requiring a strong MPRA expression effect ($|\log 2 \text{ effect size}| > 1.4$), 6.3% of all eQTL contained multiple regulatory variants.

The degree to which eQTLs are composite products of multiple causal variants is unknown because of high LD. We assessed whether allelic MPRA hits found within eQTL were more likely to be concordant with eQTL effect direction than other tested sites. We found

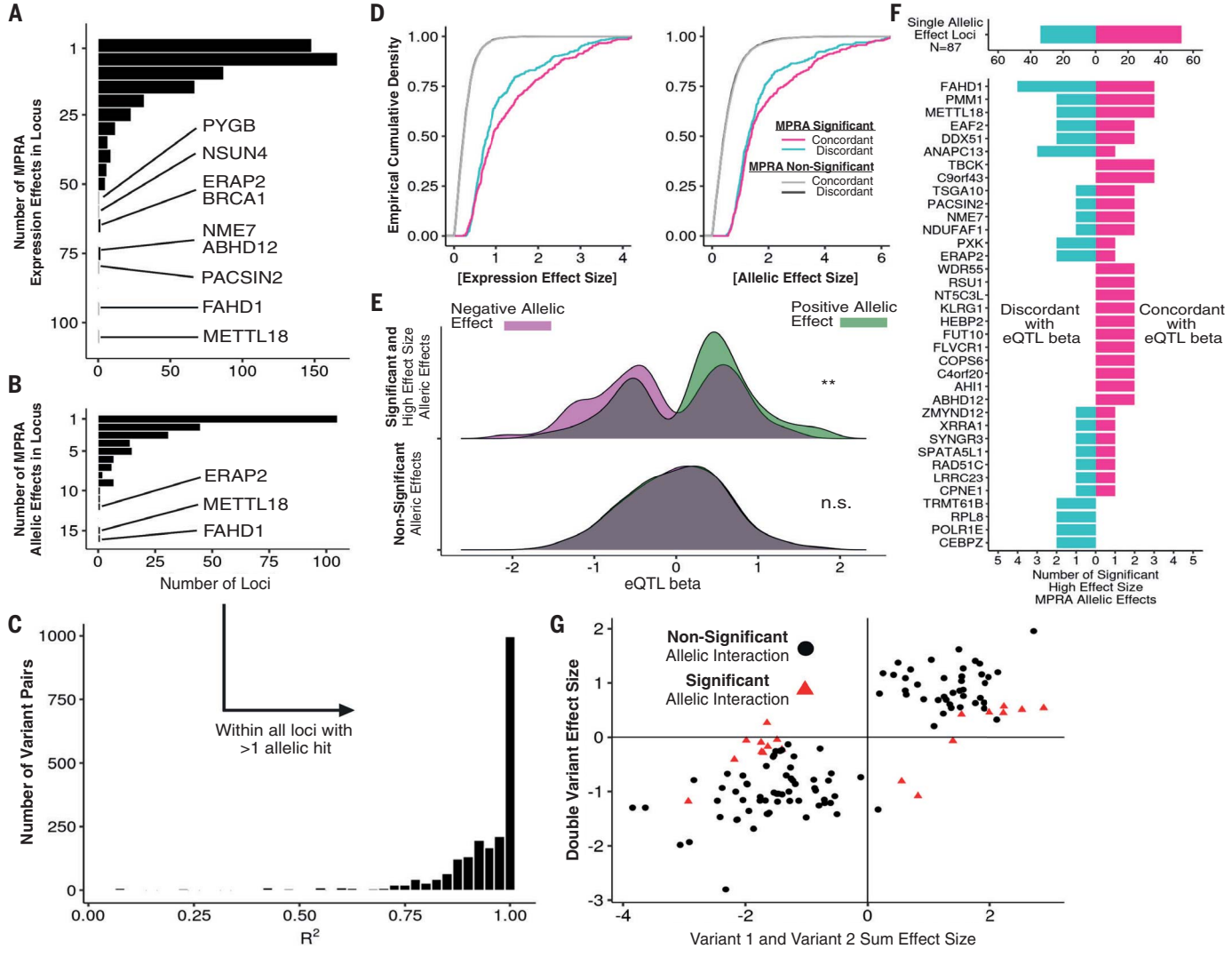

Fig. 4. Decomposition of allelic heterogeneity within regulatory loci. (**A**) Histograms of the number of expression MPRA hits per locus with BH-adjusted $P \leq 0.05$ (negative binomial regression). (**B**) Same as in (A) but requiring BH-adjusted $P \leq 0.05$ for allelic MPRA hits. (**C**) Distribution of LD r^2 values between all pairs of allelic MPRA hits within genes with multiple hits. (**D**) Cumulative distribution of effect sizes stratified by concordance. Concordance is defined as the sign of the allelic effect size matching the sign of eQTL beta. (**E**) Distribution of eQTL betas measured in GTEx v8 LCLs for strong MPRA hits (log expression effect size ≥ 1.4), stratified by MPRA allelic effect direction and significance from negative binomial regression. (**F**) Using the same variants as (E), counts of directionally concordant and discordant allelic MPRA hits across all loci. (**G**) Comparison of haplotype regression coefficients for variants tested individually or jointly; red points indicate allelic interaction BH-adjusted $P \leq 0.05$ (negative binomial regression). The *x* axis displays the sum of effect sizes associated with oligos containing each variant individually, and the *y* axis displays the effect size associated with the oligo containing both variants.

that expression and allelic MPRA effect sizes were larger for concordant variants compared with discordant variants (Fig. 4D). Across strong allelic MPRA hits ($|\log 2$ effect size$| > 1.4$), we observed significant concordance with eQTL effect direction (Fisher's exact $P = 4.75 \times 10^{-3}$) (Fig. 4E) and the strongest examples of allelic heterogeneity (Fig. 4F).

To rule out study-specific effects, we verified that eQTL effect sizes were consistent across multiple studies (fig. S4A) (*17*, *18*). We found consistent patterns of concordance (fig. S4B). Additionally, to ensure that concordance patterns were not driven by individual eQTL with many concordant MPRA hits, we applied binomial count logistic regression to test whether concordance proportions were shifted between allelic MPRA hits and other tested sites. We found that allelic MPRA hits, but not other sites, were significantly concordant ($P = 2.85 \times 10^{-3}$) (fig. S4C). We further found that concordance persists through the top four ranked variants per eQTL, with the set of third-strongest MPRA hits across all eQTL having a concordance rate of 0.67 (fig S4D). Altogether, these results indicate that several eQTL regions contain multiple, concordant allelic MPRA hits.

Haplotype decomposition identifies allelic regulation that is unlikely to be observed by population sampling

A major advantage of synthetic library design is separation of extremely proximal variants that are unlikely to be naturally separated through recombination. Our library included 2097 pairs of eVariants within 75 bp. For these variants, we extended our statistical model to account for four haplotypes at each pair of variants and computed summary statistics for each of the three nonreference haplotypes (fig. S5A). We then selected all variants included in at least one haplotype allelic MPRA hit.

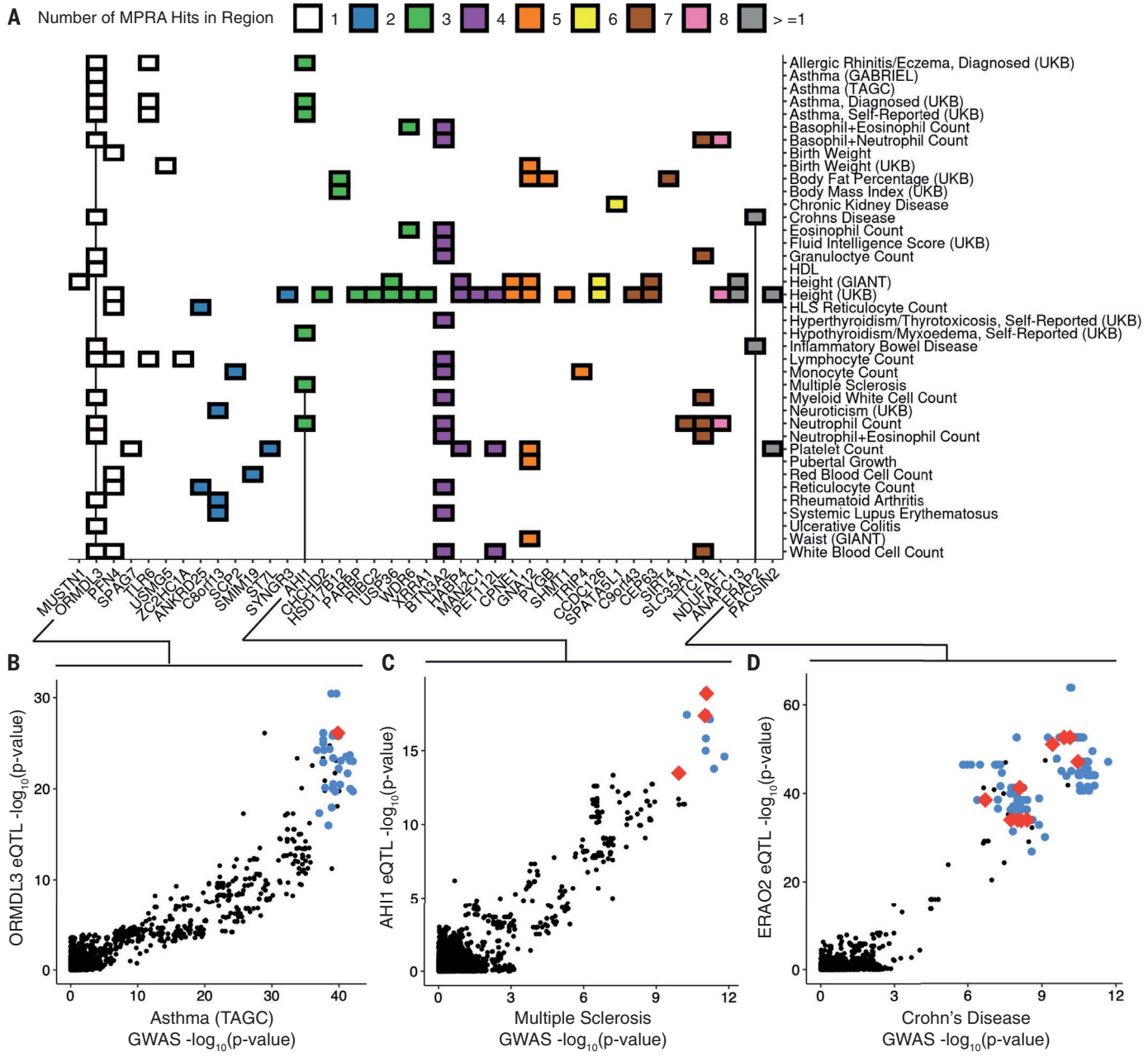

Fig. 5. Resolving complex trait associations with multiple causal variants. (**A**) Heatmap of significant colocalizations between eQTL loci and selected GWAS. Color indicates the number of allelic MPRA hits within the colocalized regions. (**B**) Comparison of genetic associations for asthma and *ORMDL3* expression in GTEx v8 LCLs. Red and blue points indicate allelic MPRA hits and other tested variants significant and nonsignificant allelic MPRA hits, respectively; black points indicate untested variants not included in our library. (**C**) Same as in (B) for associations with multiple sclerosis and *AHI1* expression. (**D**) Same as in (B) and (C) for associations with Crohn's disease and *ERAP2* expression, which was also colocalized with inflammatory bowel disease (fig. S5A). All GWAS and eQTL colocalizations are retrieved from (*22*), and lead variants were required to be genome-wide significant (reported GWAS $P \leq 5 \times 10^{-8}$ and reported eQTL $P \leq 5 \times 10^{-5}$) even if colocalization probability was high.

Combined, we identified 120 variant pairs (6.15% of all tests) with at least one haplotype allelic MPRA hit relative to all-reference sequence (negative binomial adjusted $P < 0.05$).

Most of the haplotype effects appeared additive, with a small number displaying nonadditivity (Fig. 4G). Our linear contrast test allowed us to identify these nonadditive interactions between allelic hits (fig. S5, B to D, and table S7). Of the variant pairs with at least one haplotype hit, 19 pairs also had a significant haplotype interaction effect (negative binomial adjusted $P < 0.05$; 14.7% of all significant haplotype effects and 0.91% of all tested variant pairs). Significant interactions were weaker than additive effects (fig. S5E) and rarely reversed the direction of individual allelic effects. These results support other studies that have identified nonadditive regulatory effects (*14*, *19*–*21*) and find that 14.7% of haplotype effects (only 0.91% of all tests) have evidence of nonadditivity.

Experimental fine-mapping of complex trait associations

To identify loci with shared genetic architecture between eQTL and human traits, we retrieved

all genes tested in our dataset that had both an allelic MPRA hit and at least one LCL eQTL/GWAS colocalization (*22*). Out of 744 eGenes, 5.51% colocalized with at least one trait and contained at least one allelic MPRA hit. Most colocalizations contained more than one allelic MPRA hit (71.9% of colocalizations and 82.9% of eGenes), with some loci containing as many as 13 (Fig. 5A). This suggests that the default assumption of one causal variant, often used in fine mapping or GWAS colocalization, does not reflect causal variant biology at many regulatory regions. Traits with high-confidence colocalization were diverse, including blood-cell traits such as *ZC2HC1A*/lymphocyte count or *PACSIN2*/platelet count and highly polygenic traits such as *GNA12*/height (fig. S6).

The 17q21 locus contains the most extensively replicated genetic association with asthma, which colocalizes with *ORMDL3* eQTLs (Fig. 5B). This region contains a haplotype block with dozens of linked variants, flanked by two variants (rs4065275 and rs12936231) that induce loss and gain of CCCTC-binding factor (CTCF) binding, respectively. Further, other variants located between these two variants display allele-specific chromatin accessibility, histone modification, and CpG methylation (*23*, *24*). Altogether, the risk haplotype results in increased *ORMDL3* expression, which in turn negatively regulates interleukin-2 production in $CD4^+$ T cells. We identified a single allelic MPRA hit, rs12950743, that is linked to and located between the two CTCF variants (Fig. 5B). When tested by means of luciferase assay, this variant displayed a nominally significant but weak effect in the same direction as the MPRA (luciferase unpaired *t* test $P = 0.035$) (fig. S7A). Taken together, this suggests that two variants on the risk haplotype alter *CTCF* binding, leading to distinct regulatory contacts with their own allelic specificity.

By contrast, a different colocalization that included three active variants was *AHI1*, a well-characterized gene strongly associated with multiple sclerosis (MS) (Fig. 5C) (*25*). This region contains a strong eQTL and colocalization signal in LCLs; however, its causal variant(s) are unknown. We identified rs6908428, rs9399148, and rs761357 as allelic MPRA hits. We validated the allelic effects of these variants by means of luciferase assay and found that rs6908428 (luciferase unpaired *t* test $P = 5.1 \times 10^{-6}$) and rs761357 (unpaired *t* test $P = 7.6 \times 10^{-3}$) showed allelic differences consistent with the MPRA, whereas rs9399148 (unpaired *t* test $P = 0.18$) did not (fig. S7B). The first two of these variants have been highlighted by annotation overlap in prior studies of the role of *AHI1* in MS pathology, particularly interferon-γ production and $CD4^+$ T cell differentiation, but were severely limited by linkage across the risk haplotype (*25*). When screened against known TF binding motifs, we found that rs6908428 and rs761357 overlapped predicted binding motifs for SMAD3/4 and HNF1A, respectively. Unlike HNF1A, SMAD3/4 are expressed in LCLs, suggesting that rs6908428 may function to create a SMAD3/4 binding site (fig. S8A).

A complex multivariant colocalization was identified at *ERAP2*, an aminopeptidase functionally implicated in both inflammatory bowel disease and Crohn's disease (*26*). We detected 13 active variants that span a strongly linked haplotype. Although both eQTL and GWAS suggest a single top SNP, that top SNP differs between eQTL and GWAS, and neither are MPRA hits (Fig. 5D and fig. S6A). Prior work has shown that a common splice variant in *ERAP2* results in nonsense-mediated decay (NMD) and allele-specific expression, which can cause an eQTL signal (*27*). However, the haplotype with this variant contains hundreds of other linked variants and harbors a second conditional *ERAP2* eQTL in Genotype-Tissue Expression (GTEx) LCLs. We evaluated eight of the 13 active variants from our MPRA by means of luciferase assay and found significant allelic differences at four of the eight loci (luciferase unpaired *t* test $P < 0.05$; rs1757538970, rs2549785, rs27298, and rs7713127) (fig. S7C). This suggests that *ERAP2* is regulated by a complex allelic structure that operates through gene expression and splicing.

Another colocalization was *PACSIN2*, which contained 13 variants and whose eQTL colocalized with platelet count. *PACSIN2* is an F-BAR domain protein involved in vascular and platelet homeostasis (*28*). We evaluated eight of the 13 allelic variants by means of luciferase assay and found significant effects at six of the eight loci (luciferase unpaired *t* test $P < 0.05$) (fig. S7D). Two of the variants (rs5751402 and rs9607970) were predicted to disrupt known TF binding motifs in directions consistent with their luciferase assay result (fig. S8B). The two TF were PAX5 and NFKB1, both of which are very highly expressed in LCLs, suggesting that rs5751402 and rs9607970 may function through disruption of NFKB1- and PAX5-binding sites.

Discussion

LD is a major barrier to identifying causal variants in genetic association studies. Furthermore, functional genomic annotations can be useful to prioritize likely causal variants, but many annotations are also inconclusive, unattainable, or unknown (*2*). In this study, we demonstrate that MPRAs provide a scalable platform with which to separate and map the regulatory activities of expression- and complex trait–associated natural genetic variants and highlight the limitations of existing approaches to variant interpretation and computational fine mapping. Across positional annotations and variant scores, we observed that both allelic MPRA hits and other tested variants were shifted relative to the corresponding genome-wide distributions. This demonstrates how functional predictions may readily distinguish eQTL regions from the genomic background while struggling to discriminate regulatory activity between highly linked allelic MPRA hits within the same region.

We found that multiple, tightly linked causal variants could be found under eQTL and GWAS loci. We identified that at least 17.7% of eQTL had more than one allelic hit. We further observed that most haplotype combinations exhibited additive effects, with 0.91% exhibiting nonadditivity. Using these data, we demonstrate the power of MPRA-based experimental fine mapping and report likely causal variants underlying hundreds of molecular and complex trait phenotypes, including a single variant underlying *ORMDL3*/asthma, three variants underlying *AHI1*/multiple sclerosis, and up to 13 variants each underlying *PACSIN2*/platelet count and *ERAP2*/Crohn's disease/inflammatory bowel disease.

REFERENCES AND NOTES

1. A. Buniello *et al.*, *Nucleic Acids Res.* **47** (D1), D1005–D1012 (2019).
2. V. Tam *et al.*, *Nat. Rev. Genet.* **20**, 467–484 (2019).
3. D. J. Schaid, W. Chen, N. B. Larson, *Nat. Rev. Genet.* **19**, 491–504 (2018).
4. A. Melnikov, X. Zhang, P. Rogov, L. Wang, T. S. Mikkelsen, *J. Vis. Exp.* **90**, e51719 (2014).
5. J. Ernst *et al.*, *Nat. Biotechnol.* **34**, 1180–1190 (2016).
6. M. Kircher *et al.*, *Nat. Commun.* **10**, 3583 (2019).
7. R. P. Patwardhan *et al.*, *Nat. Biotechnol.* **27**, 1173–1175 (2009).
8. R. Tewhey *et al.*, *Cell* **165**, 1519–1529 (2016).
9. J. C. Ulirsch *et al.*, *Cell* **165**, 1530–1545 (2016).
10. C. V. Weiss *et al.*, *eLife* **10**, e63713 (2021).
11. J. Choi *et al.*, *Nat. Commun.* **11**, 2718 (2020).
12. J. C. Klein *et al.*, *Nat. Commun.* **10**, 2434 (2019).
13. R. She, D. F. Jarosz, *Cell* **172**, 478–490.e15 (2018).
14. K. Renganaath *et al.*, *eLife* **9**, e62669 (2020).
15. Ž. Avsec *et al.*, *Nat. Methods* **18**, 1196–1203 (2021).
16. X. Li *et al.*, *Nat. Genet.* **52**, 969–983 (2020).
17. T. Lappalainen *et al.*, *Nature* **501**, 506–511 (2013).
18. GTEx Consortium, *Science* **369**, 1318–1330 (2020).
19. J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, B. A. Cohen, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 19498–19503 (2012).
20. J. E. Powell *et al.*, *PLOS Genet.* **9**, e1003502 (2013).
21. V. Hivert *et al.*, *Am. J. Hum. Genet.* **108**, 786–798 (2021).
22. A. N. Barbeira *et al.*, *Genome Biol.* **22**, 49 (2021).
23. D. J. Verlaan *et al.*, *Am. J. Hum. Genet.* **85**, 377–393 (2009).
24. A. Rathod *et al.*, *Epigenet. Insights* **13**, 2516865720923395 (2020).
25. B. J. Kaskow *et al.*, *Neurol. Neuroimmunol. Neuroinflamm.* **5**, e414 (2017).
26. K. Christodoulou *et al.*, *Gut* **62**, 977–984 (2013).
27. A. M. Andrés *et al.*, *PLOS Genet.* **6**, e1001157 (2010).
28. A. J. Begonja *et al.*, *Blood* **126**, 80–88 (2015).
29. N. Abell, nsabell/mpra-v2: FineMapMPRA. *Zenodo* (2022); doi: 10.5281/zenodo.5921041.

ACKNOWLEDGMENTS

We thank members of the S. Montgomery laboratory for general guidance and feedback on this work and members of the M. Bassik laboratory for experimental advice. We also thank R. Tewhey and M. Love for experimental and statistical modeling advice, respectively, and N. Cyr for assistance with figures. **Funding:** N.S.A. is supported by the Stanford Department of Genetics T32 training grant and the Joint Institute for Metrology in Biology (JIMB) training program. E.G. is funded by the National Science Foundation Graduate Research Fellowship Program grant DGE-1656518. S.B.M. is supported by National Institutes of Health grants R01AG066490, R01MH125244, U01HG009431 (ENCODE), R01HL142015 (TOPMed), and R01HG008150 (NoVa). This work in

part used supercomputing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by National Institutes of Health S10 Instrumentation Grant S10OD023452.
Author contributions: N.S.A. and S.B.M. conceived and designed the study. N.S.A., M.K.D., E.G., and K.S.S. performed all experiments, including the MPRA and luciferase assays. N.S.A., M.K.D., and M.J.G. designed oligonucleotide libraries. N.S.A. and Z.H. conducted statistical and bioinformatic analyses of sequencing data. N.S.A. and S.B.M wrote the manuscript, with contributions from all authors. **Competing interests:** S.B.M. has consulting agreements with MyOme, Biomarin, and Tenaya Therapeutics. All other authors report no competing interests. **Data and materials availability:** Sequencing data are available through the Gene Expression Omnibus under accession no. GSE174534. All code and supplementary tables are available through Zenodo (*29*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj5117
Materials and Methods
Figs. S1 to S9
Tables S1 to S9
References (*30*–*47*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

17 May 2021; resubmitted 19 October 2021
Accepted 17 February 2022
10.1126/science.abj5117