



## Uncovering methylation-dependent genetic effects on regulatory element function in diverse genomes

Rachel M. Petersen, Christopher M. Vockley and Amanda J. Lea

*Genome Res.* 2025 35: 1781-1793 originally published online July 14, 2025

Access the most recent version at doi:[10.1101/gr.279957.124](https://doi.org/10.1101/gr.279957.124)

---

**References** This article cites 76 articles, 15 of which can be accessed free at:  
<http://genome.cshlp.org/content/35/8/1781.full.html#ref-list-1>

**Open Access** Freely available online through the *Genome Research* Open Access option.

**Creative Commons License** This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

## Method

# Uncovering methylation-dependent genetic effects on regulatory element function in diverse genomes

Rachel M. Petersen,<sup>1,2</sup> Christopher M. Vockley,<sup>3</sup> and Amanda J. Lea<sup>1,2,4</sup>

<sup>1</sup>Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee 37232, USA; <sup>2</sup>Evolutionary Studies Initiative, Vanderbilt University, Nashville, Tennessee 37232, USA; <sup>3</sup>The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; <sup>4</sup>Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA

A major goal in evolutionary biology and biomedicine is to understand the complex interactions between genetic variants, the epigenome, and gene expression. However, the causal relationships between these factors remain poorly understood. mSTARR-seq, a methylation-sensitive massively parallel reporter assay, is capable of identifying methylation-dependent regulatory activity at many thousands of genomic regions simultaneously and allows for the testing of causal relationships between DNA methylation and gene expression on a region-by-region basis. Here, we develop a multiplexed mSTARR-seq protocol to assay naturally occurring human genetic variation from 25 individuals from 10 localities in Europe and Africa. We identify 6957 regulatory elements in either the unmethylated or methylated state, and this set was enriched for enhancer and promoter chromatin annotations, as expected. The expression of 58% of these regulatory elements is modulated by methylation, which is generally associated with decreased transcription. Within our set of regulatory elements, we use allele-specific expression analyses to identify 8020 sites with genetic effects on gene regulation; further, we find that 42.3% of these genetic effects vary in direction or magnitude between methylated and unmethylated states. Sites exhibiting methylation-dependent genetic effects are enriched for GWAS and EWAS annotations, implicating them in human disease. Compared with data sets that assay DNA from a single European ancestry individual, our multiplexed assay is able to uncover more genetic effects and methylation-dependent genetic effects, highlighting the importance of including diverse genomes in assays that aim to understand gene regulatory processes.

[Supplemental material is available for this article.]

A major goal in evolutionary biology, human genomics, and biomedicine is to understand how the genome generates complex traits, which at the molecular level often involves complex interactions between genetic variants, the epigenome, and gene expression. DNA methylation (DNAm) refers to the addition of methyl groups to cytosine bases, usually within CpG dinucleotide contexts (the most commonly methylated motif in mammals). DNAm is an epigenetic gene regulatory mechanism that controls gene expression programs and plays important roles in the development and differentiation of tissues (Luo et al. 2018; Greenberg and Bourc'his 2019; Loyfer et al. 2023), phenotypic and developmental plasticity in response to environmental exposures (Waterland and Jirtle 2003; Weaver et al. 2004; Garg et al. 2018; Perera et al. 2020; Anderson et al. 2024), and aging-related physiological decline and disease (Jin and Liu 2018; Salameh et al. 2020; Yim et al. 2020).

Although DNAm is canonically associated with repressed transcription of nearby genes, for example, through alterations in transcription factor (TF) recruitment and/or binding affinity (Holliday and Pugh 1975; Riggs 1975; Han et al. 2011; Yin et al. 2017; Kaluscha et al. 2022), recent studies have shown that this relationship varies across the genome. For example, in many regions, DNAm is functionally negligible, having little to no effect on TF binding or downstream mRNA production (Hu et al. 2013; Pacis et al. 2015; Kribelbauer et al. 2020; Kreibich et al. 2023). In addition to many regions of the genome being functionally insensitive

to methylation, others display DNAm-expression relationships that are opposite the canonical pattern. For example, the positioning of methylation surrounding TF binding motifs can result in increased binding affinity or altered chromatin conformation, which can ultimately result in higher levels of DNAm correlating with increased expression of nearby genes (Luo et al. 2021; Monteagudo-Sánchez et al. 2024). Given the complexities of DNAm-expression relationships, it is therefore essential to directly evaluate whether and how DNAm impacts local gene transcription, rather than assuming a functional (typically repressive) relationship.

In addition to DNAm, it is also clear that genetic variation impacts gene expression, often through similar processes such as variable TF recruitment and binding (Kasowski et al. 2010; Deplancke et al. 2016). Accordingly, expression quantitative trait loci (eQTL) studies have documented a myriad of associations between *cis* genetic variants and gene expression in human populations (Kim-Hellmuth et al. 2020; The GTEx Consortium 2020; Kerimov et al. 2021; Zeng et al. 2022), with high powered studies now reporting one or more *cis* eQTL for most genes in the genome (Gong et al. 2018; Vösa et al. 2021). Further, recent work suggests that genotype-expression relationships may in some cases depend on the local DNAm context; for example, Zeng et al. (2023) mapped 1063 “methylation-dependent eQTL” (in which the eQTL effect depends on local DNAm levels): The vast majority of these eQTL had not previously been reported as standard eQTL (Zeng et al. 2023). Although studies such as these provide correlational evidence that DNAm may alter genotype-expression

**Corresponding author:** [amanda.j.lea@vanderbilt.edu](mailto:amanda.j.lea@vanderbilt.edu)

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.279957.124>. Freely available online through the *Genome Research* Open Access option.

© 2025 Petersen et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

relationships in important ways (Banovich et al. 2014; Singh et al. 2021), experimental studies demonstrating cause and effect relationships are much rarer.

To address these gaps, massively parallel reporter assays (MPRAs) have become a useful tool for experimentally testing the causal effects of *cis* genetic variation and DNAm on transcriptional activity. The original MPRA design involved cloning a potential regulatory region of interest into a plasmid backbone upstream of a reporter gene and a unique sequence-associated barcode (Patwardhan et al. 2009, 2012; Kwasiński et al. 2012; Melnikov et al. 2012; Sharon et al. 2012). Plasmids are transfected into a cell line and incubated, and regulatory activity of the insert is then quantified via barcode sequencing. Building upon these methods, self-transcribing active regulatory region sequencing (STARR-seq) was developed as a type of MPRA; in this assay, regions of interest are inserted within the 3' UTR, such that active regulatory elements drive expression of transcripts that include the focal sequence itself. Within a cellular environment, regulatory activity thus directly scales with plasmid-derived mRNA levels of the focal sequence (Arnold et al. 2013). Finally, methyl-STARR-seq (mSTARR-seq) pairs the STARR-seq approach with methylation manipulation of query fragments to test for the effects of DNAm on regulatory activity (Lea et al. 2018). Consistent with other techniques, mSTARR-seq studies have shown that experimental manipulation of DNAm does not always impact gene expression in the canonical direction: about one-half to three-fourths of the tested regulatory elements were found to be insensitive to changes in DNAm, and 2%–14% of regions with methylation-dependent (MD) regulatory function actually exhibit *increased* transcriptional activity following experimental methylation (Lea et al. 2018; Johnston et al. 2024).

Although MPRA methods excel in directly testing for causal relationships between genetic or epigenetic variation and regulatory activity, studies of genetic variation have largely focused on testing synthesized sequences (Tewhey et al. 2016; Ulirsch et al. 2016; Kalita et al. 2018; Siraj et al. 2024). Although synthesized oligos are preferred to interrogate the effects of rare genetic variants and/or to isolate the effects of individual SNPs that occur in linkage disequilibrium (LD), they are expensive to produce, are commonly short (<200 bp), and may lack naturally occurring genetic patterns. In turn, studies of epigenetic variation have exclusively focused on DNA from a single European ancestry individual (Lea et al. 2018; Johnston et al. 2024). The common practice of benchmarking experimental assays on a single European ancestry genome limits our knowledge of how genetic variation impacts functional genomic mechanisms and perpetuates biases in genomic research (Gurdasani et al. 2019; Mills and Rahal 2020). To improve upon these shortcomings, we developed multiplexed mSTARR-seq, an extension of mSTARR-seq that can incorporate genetically diverse input and is thus capable of experimentally testing for interactions between DNAm and genotype in high throughput.

In this study, we apply multiplexed mSTARR-seq to identify allelic differences in regulatory function, in both methylated and unmethylated contexts, for hundreds of thousands of regions simultaneously. To do so, we create a genetically heterogeneous, barcoded DNA input library from 25 individuals from the 1000 Genomes Project originating from 10 different populations throughout Europe and Africa (The 1000 Genomes Project Consortium 2015). Using this design, we aimed to experimentally identify genomic regions where methylation and genetic variation interact to modify regulatory activity (i.e., MD genetic effect sites). After identifying these sites, we then explored where they are locat-

ed throughout the genome, their proximity to TF binding motifs, and their overlap with genome-wide association study (GWAS) and epigenome-wide association study (EWAS) hits. Lastly, we asked how much the multiplexing of DNA from genetically diverse individuals can increase mSTARR-seq's ability to identify genetic effects as well as interactions between genetic and epigenetic effects. As the field of genomics works toward characterizing human genetic diversity and its impacts on regulatory function, multiplexed assays such as this will be an important tool to maximize the breadth of sequence variation that can be investigated and contribute to a more comprehensive understanding of genome regulation.

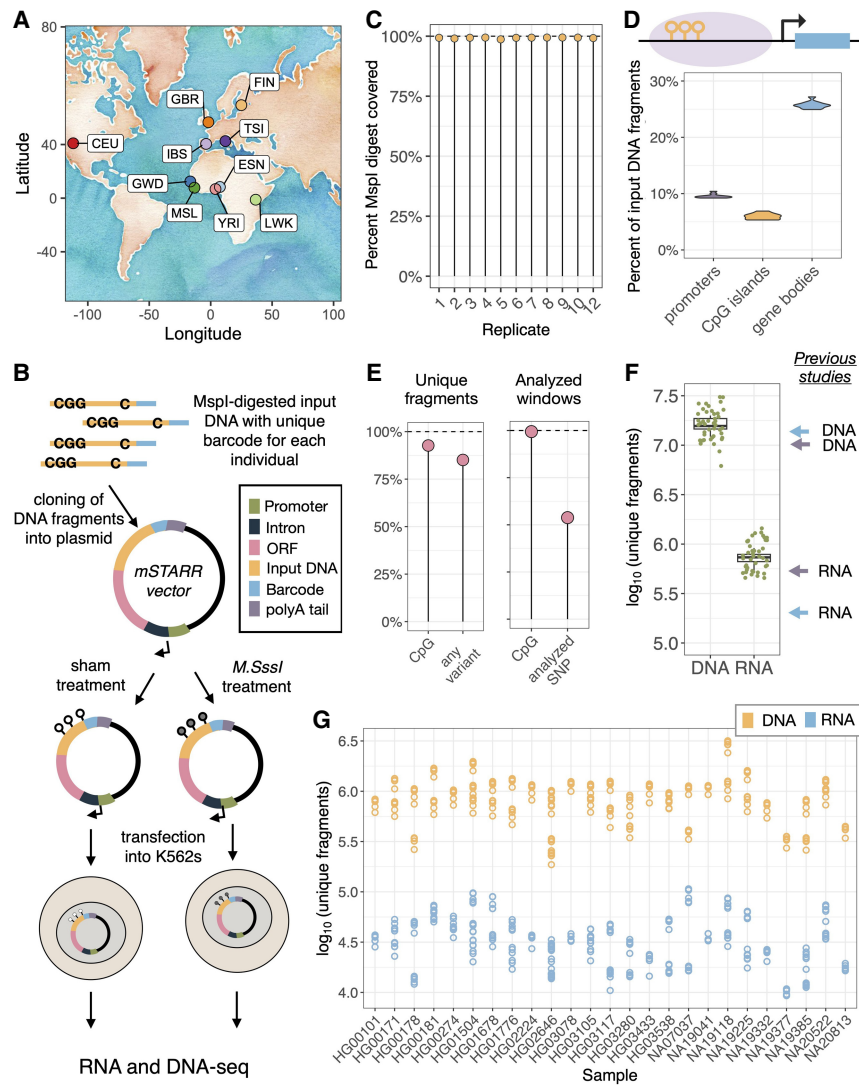
## Results

### Multiplexed mSTARR-seq produces a genetically diverse input library

We extracted DNA from lymphoblastoid cell lines (LCLs) of 25 different 1000 Genomes individuals originating from 10 different populations (Fig. 1A; Supplemental Table S1; The 1000 Genomes Project Consortium 2015). Following the methods of previous mSTARR-seq experiments (Lea et al. 2018, Johnston et al. 2024), we generated adaptor-ligated, size-selected DNA input fragments with a new addition of an individual-specific, CpG-free barcode (see Supplemental Methods). We pooled barcoded DNA libraries from different individuals to create two diverse input pools, with many individuals ( $n = 16$ ) present in both pools but harboring differing barcodes (Supplemental Fig. S1B; Supplemental Table S3). We cloned our genetically diverse input libraries into the mSTARR vector and confirmed the diversity of the resulting pool (Supplemental Figs. S1C, S2, S3). We enzymatically methylated half of the plasmids and exposed the other half to a mock/sham control (Supplemental Fig. S1D) and transfected them into the K562 myeloid cell line (Supplemental Fig. S1E). Following a 48 h incubation, we coextracted DNA and RNA, sequenced the resulting replicate-level DNA and RNA libraries (Supplemental Table S4), and assigned reads to individuals within each replicate through the demultiplexing of individual-specific barcodes (Fig. 1B).

In total, the assay included 36.1 million unique DNA fragments, covering 1.95 Gb of the human genome and including 23.2 million unique CpG sites. These DNA input fragments covered the majority of an *in silico* MspI digest: Each replicate covered >98% of the expected fragments (Fig. 1C), and each individual within each replicate covered on average 91.3% of the expected fragments; 84.9% of reads started or ended with an MspI cut site. As expected, DNA input fragments were enriched for functionally important areas of the genome, with 41% located within protein coding sequences, 18% in promoter regions, and 17% in CpG islands (Fig. 1D). As observed in RRBS (Gu et al. 2011), these DNA input fragments also covered at least one CpG in the vast majority of genes (99.1%), promoters (99.7%), CpG islands (84.1%), and enhancers (93.8%) in the human genome (Supplemental Fig. S4).

The input DNA fragments also contained significant genetic diversity, including 19.6 million unique genetic variants, 737,394 of which were biallelic SNPs with a minor allele frequency (MAF) >0.05; 92.7% of fragments contained at least one CpG site, and 85.1% of fragments contained at least one SNP (Fig. 1E). After restricting to genomic windows with sufficient coverage across replicates to facilitate statistical modeling and to SNPs called within our libraries (described below and in the Methods), 99.4% of windows contained at least one CpG site, totaling 6.2 million



**Figure 1.** Multiplexed mSTARR-seq assays a diverse input library. (A) Sampling locations of the 25 individuals included in the assay: (CEU) Utah residents (CEPH); (ESN) Esan in Nigeria; (FIN) Finnish in Finland; (GBR) British in England and Scotland; (GWD) Gambian in Western Division in the Gambia; (IBS) Iberian population in Spain; (LWK) Luhya in Webuye, Kenya; (MSL) Mende in Sierra Leone; (TSI) Toscani in Italy; and (YRI) Yoruba in Ibadan, Nigeria. (B) Multiplexed mSTARR-seq design: Sample-specific barcodes are added to MspI-digested input DNA and inserted into the mSTARR vector downstream from a promoter, intron, and open reading frame (ORF). Plasmids are exposed to a methylation treatment or sham control and transfected into K562 cells and incubated for 48 h, and DNA and RNA are extracted and sequenced. (C) Percentage of an *in silico* MspI digest of the human genome that is represented in the DNA input of each replicate. (D) Percentage of input DNA fragments located within promoters, CpG islands, and gene bodies in each replicate (note these are not mutually exclusive annotations). (E) Percentage of unique DNA fragments that contain at least one CpG site and at least one SNP (left), and the percentage of analyzed windows ( $n = 525,074$ ) that contain at least one CpG site and at least one analyzable SNP (i.e., biallelic,  $>0.05$  MAF, and was called in our joint genotyping analysis; right). (F) Number of unique DNA and RNA fragments observed in each replicate; the mean number of fragments included in each replicate in Lea et al. (2018) (purple arrows) and Johnston et al. (2024) (blue arrows) is shown for comparison. (G) Number of unique DNA and RNA fragments included in each replicate from each of the 25 individuals included in the assay.

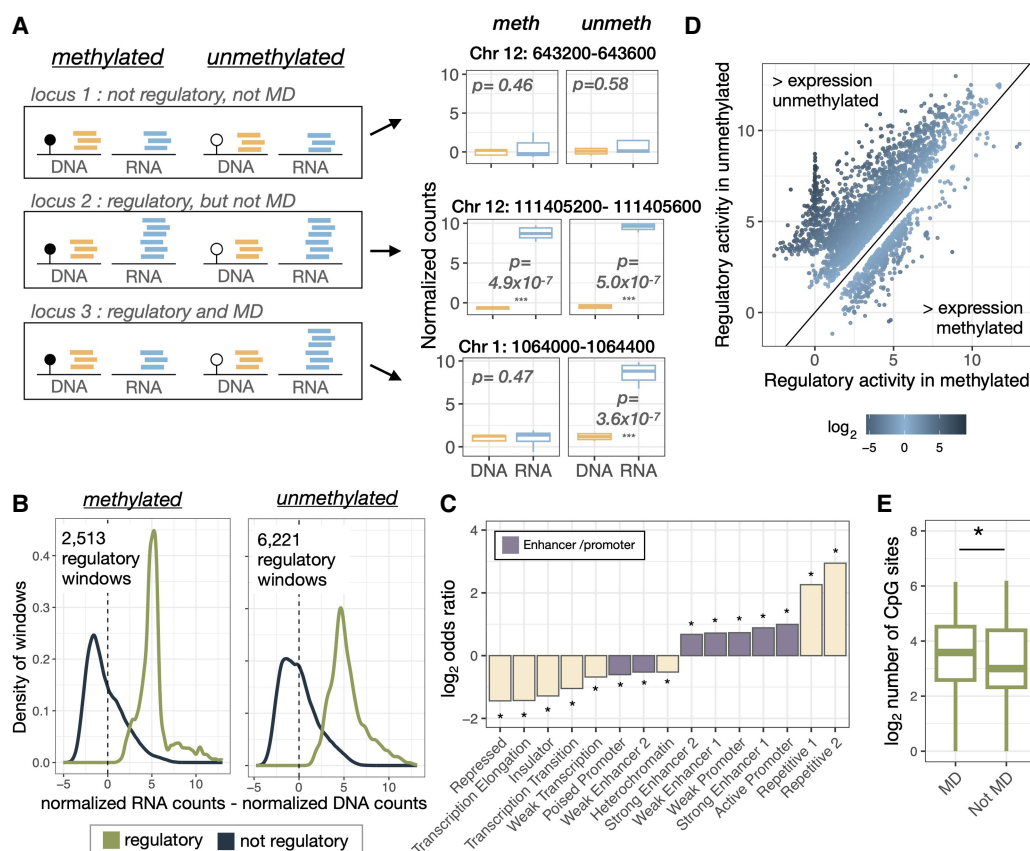
CpGs sites, and 53.8% of windows contained at least one analyzable SNP (i.e., biallelic,  $>0.05$  MAF, and called in our joint genotyping analysis), totaling 282,403 SNPs (Fig. 1E). Detailed information on sequencing depth can be found in Supplemental Table S4.

As observed in previous mSTARR-seq and STARR-seq work (Arnold et al. 2013; Vockley et al. 2015; Lea et al. 2018; Johnston

et al. 2024), many input DNA fragments do not generate plasmid RNA output, potentially indicative of a lack of endogenous regulatory activity. We found that each replicate contained  $736,006 \pm 32,831$  unique RNA fragments (generated from  $15.9 \pm 0.95$  million unique DNA fragments). Importantly, the average complexity of both our RNA and DNA libraries slightly exceeds that observed in previous mSTARR experiments (Fig. 1F; Lea et al. 2018; Johnston et al. 2024). All individuals achieved comparable representation across replicates, with a mean of  $35,047 \pm 2707$  unique RNA fragments generated from  $896,899 \pm 58,144$  unique DNA fragments per individual per replicate (Fig. 1G). The number of windows/variants tested at each data analysis step can be found in Supplemental Figure S5.

### Multiplexed mSTARR-seq identifies MD regulatory elements

To identify generalizable regulatory and MD regulatory elements, we combined uniquely mapped reads from all individuals within a replicate, overlapped these with 400 bp nonoverlapping genomic windows to accommodate our mean insert size, and filtered for windows with adequate coverage (one or more DNA read in half of replicates from both conditions, one or more RNA read in half of replicates in either condition) (see Lea et al. 2018; Johnston et al. 2024). Following filtering for DNA and RNA coverage, we were left with 525,074 400 bp windows, of which 522,294 (99.4%) contained at least one CpG site. Nearly all windows (95.9%) were represented across all replicates (Supplemental Fig. S6). We tested for regulatory activity at each 400 bp window in methylated and unmethylated conditions separately by asking whether the abundance of the reporter gene-derived mRNA exceeded the amount of input DNA for that window (Fig. 2A). We identified 6221 windows with regulatory activity in the unmethylated replicates and 2513 windows with regulatory activity in the methylated replicates (at a 1% FDR) (Fig. 2B). We found a strong enrichment for shared identification of regulatory windows and a significant correlation in effect sizes for windows tested in this study versus previous mSTARR-seq studies (Supplemental Table S5; Supplemental Figs. S7, S8; Supplemental Results). As expected, the regulatory windows we identified were significantly enriched for strong enhancer and active promoter ChromHMM annotations in K562s (Fig. 2C; Supplemental Table S6). When comparing



**Figure 2.** Multiplexed mSTARR-seq identifies regulatory and methylation-dependent (MD) regulatory activity. (A) Heuristic patterns of read pileups associated with the identification of regulatory activity and methylation dependence, with an example of the normalized read counts for a window falling into each of these categories. (B) Density of regulatory and nonregulatory windows (area under each curve normalized to one) in relation to the difference between normalized RNA and DNA counts for that window. (C) Fisher's exact test for enrichment in ChromHMM genomic annotations when comparing windows with regulatory activity (combined across conditions) versus nonregulatory windows. Bars *above*  $y=0$  indicate annotations that are overenriched in regulatory windows, and bars *below*  $y=0$  indicate annotations that are under enriched in regulatory windows; purple bars indicate enhancer and promoter annotations; and stars indicate significant over/underenrichment of that annotation type (for full results, see Supplemental Table S6). (D) Regulatory activity in the methylated versus unmethylated condition for MD windows colored by the logFC between conditions: 84.8% of MD windows have greater activity in the unmethylated condition, with the clustering of sites at  $x=0$  representing windows whose expression is entirely repressed by methylation. (E) Windows with MD regulatory activity have a greater number of CpG sites compared with regulatory windows that are not modulated by methylation.

regulatory windows that were found in one condition but not the other, we found that those specific to the unmethylated condition were particularly enriched for promoter and enhancer annotations, whereas those that were specific to the methylated condition were more enriched for heterochromatin and regions of weak transcription (Supplemental Fig. S9; Supplemental Table S7).

To test for methylation dependence in regulatory activity, we used a multivariate adaptive shrinkage approach (mashr) (Urbut et al. 2019), which enables testing for effect size sharing versus heterogeneity across conditions. We defined MD regulatory elements as those exhibiting regulatory activity in at least one condition (local false sign rate: LFSR < 0.05) and with a log fold change (logFC) in regulatory activity > 1.5 between conditions (Fig. 2A). Of the 6957 unique windows with regulatory activity in one condition, 4052 displayed MD activity, meaning that methylation modulated the degree to which the fragment drove gene expression for 58.2% of tested fragments. After rerunning previous mSTARR-seq data sets using this pipeline, we find a similar proportion of regulatory windows exhibiting methylation dependence (Johnston et al. 2024: 67.6%; Lea et al. 2018: 80.2%; Supplemental Table S5). As expected, most windows (84.8%) had greater transcriptional activ-

ity in the unmethylated relative to the methylated condition (Fig. 2D). As noted in previous work, we found that windows displaying MD regulatory activity had a greater number of CpG sites relative to regulatory elements that were not modulated by methylation (Student's  $t$ -test:  $t=6.38$ ,  $df=6352.2$ ,  $P=1.9 \times 10^{-10}$ ) (Fig. 2E), and the strength of MD activity was positively associated with the number of CpG sites in the region (linear model: estimate = 0.02, SE = 0.002,  $P < 2.2 \times 10^{-16}$ ) (Supplemental Fig. S10). For quality control, we also confirmed that MD windows were very unlikely to be CpG-free relative to the background set of all regulatory windows tested (Fisher's exact test: OR = 0.38,  $P = 5.76 \times 10^{-6}$ ).

To understand how often MD regions experience variation in DNAm levels in the endogenous genome, we used a cell type-specific whole-genome bisulfite sequencing data set (Loyfer et al. 2023) to ask whether CpG sites located within MD regulatory windows exhibit variable DNAm (1) between individuals in a single blood cell type (monocytes) and (2) between blood cell types within a single individual (comparing monocytes, granulocytes, NK cells, helper T cells, and B cells). We found that 75.1% of MD regulatory windows contained one or more CpGs exhibiting interindividual DNAm variation, and 78.3% contained one or more CpGs

exhibiting intraindividual DNAm variation. These findings confirm that in vivo epigenetic variation is common at sites where we implicate methylation as functionally important in a relevant cell type.

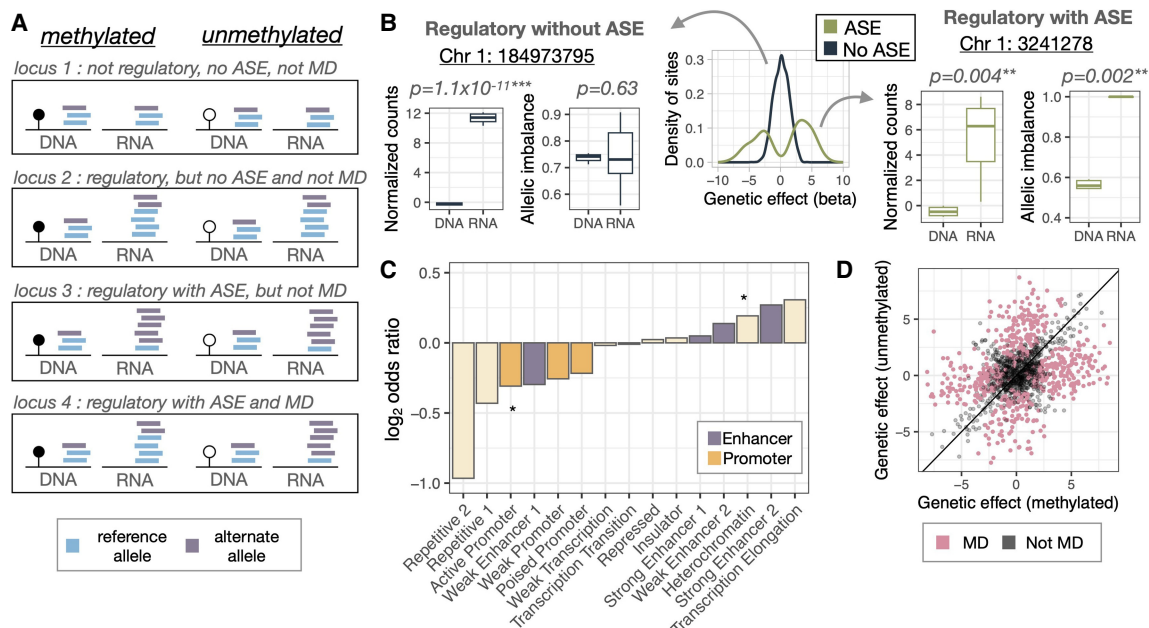
### Genetic variation impacts regulatory function

To interrogate the effect of genotype on regulatory activity, we estimated allele-specific RNA output versus DNA input in the methylated and unmethylated conditions separately, using standard pipelines in GATK and quality score cutoffs (mean quality score = 358) (Supplemental Fig. S11; Van der Auwera et al. 2013). After filtering for active regulatory windows with only one SNP or haplotype, we retained 8570 and 5853 variants for the unmethylated and methylated conditions for downstream analyses. As quality control, we confirmed these windows generally had high coverage across both alleles at the DNA level (mean coverage =  $1191 \pm 743$  reads) (Supplemental Fig. S12), with the difference between reference and alternate allele DNA coverage scaling with MAF, as expected (Supplemental Fig. S13).

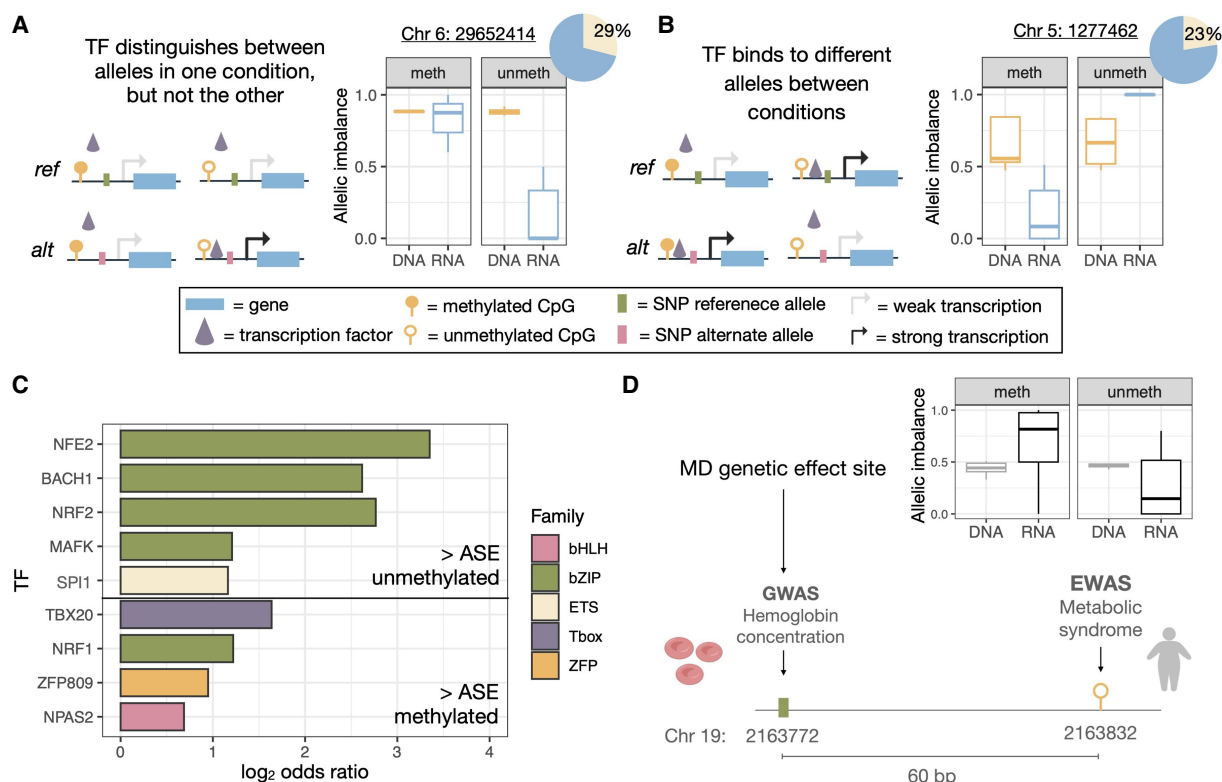
When we analyzed allelic bias at the transcriptional level, we found that 3636 SNPs in the unmethylated condition and 2816 SNPs in the methylated condition were monoallelically represented, despite robust representation of both alleles at the DNA level; these sites thus exhibited extremely strong ASE (see Methods) (Supplemental Table S8). For sites that were not monoallelically expressed, we used beta-binomial modeling (implemented in the aod package in R, <https://cran.r-project.org/package=aod>) to test whether the proportion of reference to alternate allele counts sig-

nificantly differed between the DNA input and the mRNA output at each site (using a 1% FDR threshold) (Fig. 3A). With this method, we discovered 2807 sites with ASE: 1858 sites in the unmethylated state (37.6% of tested sites) and 1150 sites in the methylated state (37.9% of tested sites) (Fig. 3B). Sequencing depth did not significantly differ between ASE and non-ASE sites, suggesting that the identification of ASE is not owing to variability or noise associated with lower DNA read counts (Student's *t*-test; unmethylated replicates:  $t = -0.278$ ,  $df = 4112.2$ ,  $P = 0.78$ ; methylated replicates:  $t = 1.18$ ,  $df = 2732.4$ ,  $P = 0.24$ ) (Supplemental Fig. S14). Furthermore, although allele-specific mapping bias should not impact our biological conclusions, which are based on RNA/DNA and unmethylated/methylated condition comparisons in which mapping biases should be identical, we nevertheless performed a parallel analysis implementing the WASP mappability filtering pipeline, which retained 78% and 81% of sites from the original pipeline in the methylated and unmethylated conditions, respectively (Supplemental Methods; Van De Geijn et al. 2015).

Recent work has highlighted that phenotypically relevant variants identified via GWAS are typically located in noncoding regions distal to transcription start sites (TSSs) and promoter regions (Maurano et al. 2012; Mostafavi et al. 2023). Conversely, eQTL signals commonly cluster within promoter regions near TSSs and may be less relevant to complex traits as they often fail to colocalize with GWAS hits (Chun et al. 2017; The GTEx Consortium 2020; Mostafavi et al. 2023). Thus, we aimed to explore where in the genome variants exerting an allele-specific effect on regulatory activity were located. Similar to GWAS sites, we found that ASE variants were located farther from the nearest TSS (Student's



**Figure 3.** Multiplexed mSTARR-seq identifies allele-specific regulatory activity that is modulated by methylation. (A) Patterns of read pileups associated with the identification of regulatory activity, ASE, and MD ASE. Reference alleles are indicated in blue; alternate alleles, in purple. (B) Density of tested sites with ASE versus without ASE in the unmethylated condition, with an example plot of each showing the regulatory activity (normalized DNA and RNA counts) and ASE (the ratio of reference allele to total counts, i.e., allelic imbalance) present in each replicate. (C) Fisher's exact test for enrichment in ChromHMM genomic annotations comparing ASE sites (combined across conditions) versus non-ASE sites. Bars above  $y=0$  indicate annotations that are overenriched in ASE sites, and bars below  $y=0$  indicate annotations that are underenriched in ASE sites; purple bars indicate enhancer annotations; orange bars indicate promoter annotations; and stars indicate significant ( $P < 0.05$ ) over/underenrichment of that annotation type (for full results, see Supplemental Table S9). (D) The genetic effect in the methylated condition plotted against the genetic effect in unmethylated condition for ASE sites, with the 575 MD ASE sites highlighted in pink.



**Figure 4.** Insights into potential mechanisms involved in MD genetic effects. (A) One potential mechanism leading to MD genetic effects wherein TFs distinguish between alleles in one condition but not the other with an example site showing allelic imbalance (the ratio of reference allele to total counts) in the methylated versus unmethylated condition and a pie chart showing the percentage of MD genetic effect sites following this pattern. (B) Another potential mechanism leading to MD genetic effects wherein TFs bind to different alleles in alternate conditions with an example site and a pie chart showing the percentage of MD genetic effect sites following this pattern. (C) TF motifs that are enriched within  $\pm 200$  bp of MD genetic sites that display increased ASE in the unmethylated condition (*top*) or methylated condition (*bottom*) colored by TF family. (D) Example of a MD genetic effect site that directly overlaps with a GWAS hit and is 60 bp away from an EWAS hit, which are associated with hemoglobin concentration and MetS, respectively. Created with bioRender (<https://www.biorender.com/>).

*t*-test:  $t = -4.19$ ,  $df = 8284.8$ ,  $P = 2.83 \times 10^{-5}$  (Supplemental Fig. S15) and were less likely to be located within active promoter regions (Fisher's exact test:  $OR = 0.81$ ,  $P = 0.0004$ ) (Fig. 3C) compared with non-ASE regulatory variants. Instead, sites with allele-specific regulatory effects were more likely to be located in enhancer annotations (although this trend does not reach statistical significance) (Fig. 3C; Supplemental Table S9). These results are robust to the exclusion of monoallelic sites in our test data set (Supplemental Fig. S16; for full results, see Supplemental Results).

### Methylation impacts allele-specific regulatory function

To understand how methylation impacts ASE, we again used *mashr* to compare effect size estimates from our beta-binomial model between conditions and asked whether the genetic effect on regulatory activity is modified by methylation manipulation. We tested 1359 sites with ASE for methylation dependence and found 575 sites with MD genetic effects (43.2% of tested sites) (Fig. 3D). Two hundred eight of the tested sites directly overlapped with either the C or G of a CpG site in the hg38 reference genome, such that methylation could only occur for one genotype. Of these 208 CpG disrupting sites, 74% exhibited MD genetic effects ( $n = 154$  sites).

Using the same publicly available whole-genome bisulfite sequencing data set as described above (Loyfer et al. 2023), we asked

whether MD genetic effect sites were located within 200 bp of sites exhibiting variable *in vivo* DNAm. We found that 97% ( $n = 558$ ) of MD genetic effect sites were near CpGs exhibiting interindividual DNAm variation, and 99.5% ( $n = 572$ ) were near CpGs exhibiting intraindividual DNAm variation between blood cell types, providing strong evidence that *in vivo* epigenetic variation is pervasive at sites where methylation may also play a functional role. Additionally, we tested whether genetic variation impacts methylation *in vivo* at MD genetic effect sites by overlapping with meQTL identified in whole blood from the Framingham Heart Study cohort (Huan et al. 2019). We found that the majority of our ASE variants (71.6%) were meQTL in blood, suggesting that genetic variation can generate methylation differences between individuals at sites where we identify this variation as functionally meaningful.

Next, we investigated regulatory mechanisms that could generate MD genetic effects. Of the 421 sites that did not disrupt a CpG site, the majority ( $n = 239$ ) had greater regulatory activity in the unmethylated condition compared with the methylated condition (as was observed in the overall data set; binomial test,  $P = 6.28 \times 10^{-3}$ ). The most common pattern suggested a potential mechanism of TF binding with limited regulatory activity in the methylated condition and allele-specific binding in the unmethylated condition (29% of sites) (Fig. 4A). We also investigated whether the direction of ASE could vary between conditions. We

found that 95 out of the 421 MD genetic effect sites not directly abolishing a CpG site have enhanced reference allele expression in one condition and enhanced alternate allele expression in the other (Fig. 4B), suggesting mechanisms that reverse the direction of ASE depending on methylation status.

Using the genomic regions enrichment of annotations tool (GREAT) (McLean et al. 2010), we did not find enrichment of MD genetic effects near particular genes or biological pathways, suggesting that these regulatory elements occur near genes involved in diverse biological functions. Next, we tested for overrepresentation of MD genetic effects in TF binding motifs and found significant enrichment for multiple TF motifs consistent with our mechanistic hypotheses (Fig. 4A). Specifically, sites that displayed greater ASE in the methylated condition were enriched for T-box binding motifs ( $q=0.024$ ) and a moderate enrichment for basic leucine zipper (bZIP), zinc-finger protein (ZFP), and basic helix-loop-helix (bHLH) TFs ( $q=0.182$ ). Further, sites that displayed greater ASE in the unmethylated condition were enriched for bZIP and erythroblast transformation-specific (ETS) binding motifs (bZIP:  $q=0.027$ , ETS:  $q=0.026$ ) (Fig. 4C; Supplemental Table S10). These two TF families have both been previously shown to preferentially bind to unmethylated sequences in mSTARR-seq (Lea et al. 2018) and other assays (Hernandez-Corchado and Najafabadi 2022). To assess *in vivo* TF binding, we used the ChIP-Atlas database (Zou et al. 2024) and found 37 TFs that were more likely to be bound at ASE variants in the methylated condition and nine TFs that were more likely to be bound at ASE variants in the unmethylated condition (Supplemental Table S11; Supplemental Methods). TFs associated with ASE in the methylated condition included 11 from the ZFP family, including CTCF, and four from the bZIP family, similar to our motif results.

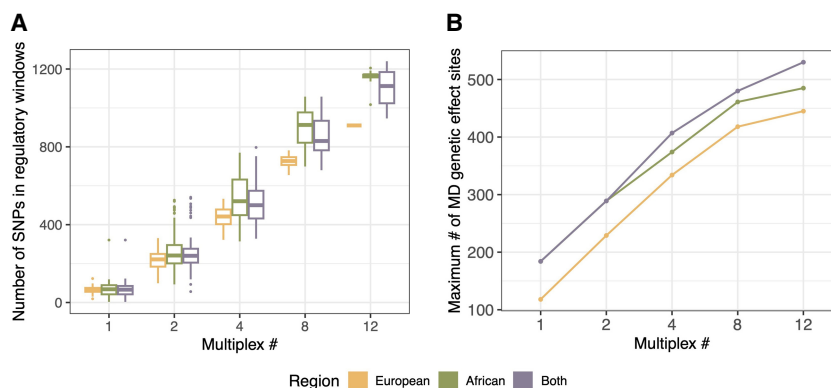
### MD genetic effects are linked to diverse phenotypes

To investigate the potential phenotypic relevance of MD genetic effect sites, we first assessed the enrichment of MD genetic effect sites near GWAS hits identified within the Pan-UK Biobank (Karczewski et al. 2024). We focused on 20 quantitative blood phenotypes relevant to the experimental cell type and did not find significant enrichment for MD genetic effect sites within the set of variants implicated in any of these individual complex traits (Supplemental Table S12).

Next, we explored GWAS and EWAS trait associations more broadly to better understand how both genetic and epigenetic variation may contribute to diverse, non-blood-related phenotypes. We found that 98 MD genetic effect sites (17%) were located within 400 bp of a GWAS hit (75 of which directly overlapped with a GWAS hit), representing 151 different site-trait associations and 114 unique GWAS hits (Supplemental Table S13). Of these 114 GWAS hits, 52 were determined to be the leading causal variant in the site-trait association using fine-mapping approaches (data accessed using the Open Targets Genetics Resource) (Ghossaini et al. 2021). The most common traits associated with MD genetic effect sites included physical attributes like height and body mass, as well as other blood-related traits like white blood cell, red blood cell, and platelet count (although we did not observe a statistical enrichment of these blood-related traits in our formal enrichment tests described above). When we asked whether MD genetic effect sites were associated with a greater number of GWAS hits or trait associations (as each GWAS hit can be associated with multiple traits) compared with random variants matched for gene density, LD, and MAF (see Supplemental Methods), we found that MD genetic effect sites were associated with a greater number of GWAS hits (permutation test:  $P=0.002$ ) (Supplemental Fig. S17A) and a comparable number of trait associations (permutation test:  $P=0.3$ ) (Supplemental Fig. S17B) compared with our matched control SNP sets. These results emphasize the utility of multiplexed mSTARR-seq in identifying variants that are important for organism-level phenotypes.

Likewise, we found that 159 MD genetic effect sites (28%) were located within 400 bp of at least one EWAS probe, totaling 340 unique probes. Of these 340 probes, 134 were associated with one or more phenotypes (39.4%), resulting in 278 different site-trait associations (Supplemental Table S14). Again, some traits were associated with multiple sites in our data set, with the most common being aging and smoking, which were associated with 18 and 16 sites, respectively. To understand whether we observe more phenotype associations than would be expected by chance, we again utilized matched control SNP data sets and found that MD genetic effect sites were not associated with more EWAS hits (permutation test:  $P=0.12$ ) (Supplemental Fig. S18A) but were associated with a greater number of traits (permutation test:  $P=0$ ) (Supplemental Fig. S18B), suggesting that MD genetic effect sites contribute broadly to phenotypes, through pleiotropy or other mechanisms.

Finally, we found that 31 of our MD genetic effect sites were located within 400 bp of both a GWAS and an EWAS hit (Supplemental Table S15), some of which were associations for related phenotypes; for example, a MD genetic effect site directly overlapped with a GWAS hit associated with hemoglobin concentration that was 60 bp away from an EWAS hit associated with metabolic syndrome (MetS) (Fig. 4D). Additionally, we found one example for which the GWAS and EWAS hits were associated with the same trait: We identified a MD genetic effect site that directly overlapped with a GWAS hit and was 37 bp away from an EWAS hit, which were both associated with smoking.



**Figure 5.** Multiplexing increases detection of MD genetic effects. (A) Number of variant sites located in regulatory regions that could potentially be tested for ASE when subset for different multiplexing regimes and samples originating from different geographical regions. (B) Maximum number of MD genetic effect sites that would have been present in our assay when subset for different multiplexing regimes and geographical regions.

## Multiplexing increases capacity to identify allele-specific and MD genetic effects

To gauge the potential benefits of multiplexing DNA from genetically diverse populations, we assessed both the number of SNPs located within regulatory regions and the number of MD genetic effect sites that could be identified in our data set by subsampling different numbers of individuals included in the assay. We found that, on average, approximately 15× more unique SNPs can be analyzed when multiplexing 12 individuals as opposed to performing allele-specific analyses on only a single individual (as has been done previously in STARR-seq experiments) (Johnson et al. 2018). This is augmented especially by the inclusion of African ancestry individuals, with a nearly 20% increase in the number of SNPs found in regulatory regions when sampling individuals of African origin in comparison to European origin (Fig. 5A). For example, when we limit the tested genomic regions to the 6,957 400 bp windows with significant regulatory function in either the methylated or unmethylated condition and subsample to sequences from a single European ancestry individual, we find that we can assay a mean of 65.4 (range: 19–123) variants. However, when we include sequences from all 12 of the European individuals in our assay, we find that we are able to assay 910 variants; these same numbers rise to 75.4 (range: 3–321) variants for a single African individual and 1104.9 (range: 1017–1205) for 12 multiplexed African individuals, respectively.

In addition to exploring allele-specific effects on expression, we aimed to describe MD genetic effects, and thus, we investigated how many of the 575 identified MD genetic effect sites would have been present and variable in our assay under different subsampling regimes. For example, sequences from a single European individual contained a maximum of 118 MD genetic effect sites; from a single African individual, a maximum of 184 MD genetic effect sites. Notably, multiplexing 12 samples originating from both continents resulted in the greatest representation of the MD genetic effect sites identified in our experiment (530 sites) (Fig. 5B), arguing for inclusion of global genetic diversity whenever possible.

## Discussion

Here, we present the first demonstration of multiplexed mSTARR-seq and its capacity to jointly assay the impact of methylation and genetic variation on regulatory activity. When identifying regulatory and MD regulatory regions, multiplexed mSTARR-seq performs similarly to previously, nonmultiplexed iterations (Lea et al. 2018, Johnston et al. 2024), confirming that the method is repeatable and robust to the addition of genetic variation and multiplexing barcodes. Multiplexed mSTARR-seq measures the gene regulatory consequences of methylation in bulk cell populations experimentally forced to ~0% or ~100% methylation, an approach that may seem to oversimplify the complex continuum of *in vivo* methylation. However, it precisely models the fundamental binary state observed at the single-cell level, at which most CpG sites are either completely methylated or unmethylated. This framework is particularly powerful because nearly all MD genetic effect sites identified in our assay exhibit variable methylation *in vivo*, either across individuals within a single cell type or between different cell types within an individual. New to this assay, we can simultaneously assess how naturally occurring genetic variation interacts with DNAm to contribute to variation in regulatory activity. Through this, we identified several hundred instances in which the genotype modulated expression differently

in methylated and unmethylated contexts, consistent with recent work exploring this mechanism *in vivo* (Zeng et al. 2023).

eQTL studies have now identified expression-associated variants for nearly every gene in the genome (The GTEx Consortium 2020); however, it is unclear how greatly these sites contribute to organism-level phenotypes as many GWAS hits do not colocalize with eQTL in any tissue (Umans et al. 2021). In a recent study exploring the genomic contexts in which eQTL and GWAS hits are found, researchers observed that eQTL were more likely to be located within promoter regions close to TSSs, whereas phenotypically relevant GWAS hits were more likely to be depleted in promoter regions and located distally to TSSs (Mostafavi et al. 2023). Here, we find that in a causal, experimental assay, SNPs with allele-specific impacts on expression are found in more similar genomic contexts to GWAS hits: They are located farther away from TSSs and are less likely to be located within promoter regions in comparison to SNPs with no impact on expression. Although GWAS studies are effective at linking genomic regions to traits, they often lack the ability to pinpoint causal variants. Although fine mapping can improve the ability to identify causal variants among linked sites, MPRA, including multiplexed mSTARR-seq, can be useful tools for experimentally nominating causal variants (Siraj et al. 2024). In the case of multiplexed mSTARR-seq, testing naturally occurring genetic variation (instead of synthetic sequences) allows for the inclusion of longer fragments that are more likely to exhibit regulatory activity (Lea et al. 2018) while also enabling the disentanglement of variant effects when they are separated by more than the insert size. Notably, previous MPRA experiments that focused on genetic effects (e.g., Tewhey et al. 2016; Ulirsch et al. 2016) were conducted with all fragments in an entirely unmethylated state: Our work suggests that an appreciable proportion of SNPs may have different effects on expression when nearby CpGs are methylated.

We explore several mechanisms that may generate MD genetic effects: (1) MD genetic effects may involve a SNP that changes either the C or G base of a CpG site; (2) TFs may exhibit genotype based binding preferences, and if the effects of methylation outweigh genotype preferences, these effects may only be observed in one condition (Kribelbauer et al. 2017); and (3) TFs may preferentially bind to distinct sequences in methylated versus unmethylated contexts (Hu et al. 2013). To understand which TFs may be particularly sensitive to variation in both sequence and methylation status, we investigated which TF binding motifs were proximal to MD genetic effect sites. Among sites with greater ASE in the unmethylated condition, we found three significantly overrepresented motifs: SPI1 from the ETS family and BACH1 and NFE2, both from the bZIP TF family. These TF families are both known to avoid binding to methylated sequences determined both from prior mSTARR-seq assays (Lea et al. 2018) and through joint accessibility-methylation-sequence modeling (Hernandez-Corchado and Najafabadi 2022). These results were corroborated by ChIP-seq data, in which we find NFE2, as well as two other TFs within the bZIP family (FOSL1 and ATF2), are more likely to be bound at sites exhibiting ASE in the unmethylated condition compared with matched background sites. Among sites with greater ASE in the methylated condition, we found a significant enrichment for TBX20 binding motifs, a TF within the Box family that is known to play a pivotal role in heart development and preferentially bind to methylated sequences (Hu et al. 2013). Just above our significance threshold, both NRF1 and ZFP809 (FDR adjusted *P*-values=0.18) were also enriched near sites with greater ASE in methylated condition. NRF1 is part of the bZIP TF family, which has been demonstrated via SELEX binding assays to be particularly

sensitive to the position of CpG methylation in the flanking regions of binding motifs (Kribelbauer et al. 2017). ZFP809 is a KRAB ZFP, a family of TFs containing a canonical methyl-binding domain (Hashimoto et al. 2015). The ability of ZFPs to engage in allele-specific binding in methylated contexts is further bolstered by ChIP-seq data, in which 30% of the TFs found to be enriched in ASE in the methylated condition were members of the ZFP family. Prior MPRA studies have also found an enrichment for these same TF families near to areas of the genome containing multiple variants in strong LD associated with eQTL and GWAS hits (Abell et al. 2022). In the future, leveraging more empirical *in vivo* TF binding data sets, with the focal cell type and potentially paired with *in vivo* DNAm information, will provide a clearer picture of how genetic and epigenetic variation disrupt TF binding.

By developing a multiplexed design, we could assay genetic variation from 25 individuals for regulatory and MD regulatory effects in a single experiment, allowing us to test a much greater number of variants for both ASE and MD genetic effects than previously possible (Lea et al. 2018, Johnston et al. 2024). An overrepresentation of European ancestry individuals is a current reality of most “omic” data sets, with >80% of GWAS and EWAS participants reporting European ancestry (Mills and Rahal 2020; Breeze et al. 2022). However, the inclusion of non-European individuals in association studies not only improves our ability to identify disease-related variants specific to underrepresented populations but also improves statistical power to identify disease-relevant variants across all populations (Pulit et al. 2010). Thus, the development of new technologies that support the inclusion of diverse genetic material into experimental assays will be necessary to develop a more comprehensive understanding of the complex interactions between genetic variants, the epigenome, and gene expression. Moving forward, incorporating diverse cell types, cellular environments, and genetic backgrounds into MPRA studies will be key to understanding the epistatic effects of *cis* and *trans* variation across ancestries. This could involve querying DNA fragments with the same SNPs but on ancestry-specific haplotypes or transfecting multiplexed libraries into cell lines derived from individuals of diverse ancestries. Expanding these studies to non-human species (e.g., Mohammed et al. 2022; Hansen et al. 2024) will also provide valuable insights into the role of gene regulation in shaping complex traits, from both biomedical and evolutionary perspectives.

## Methods

### Input library generation

We generated input fragments by extracting genomic DNA from LCLs from 25 individuals included in the 1000 Genomes Project. We included 12 individuals of European origin and 13 individuals of African origin, sampled from five different locations each (Supplemental Table S1). For each sample, we performed an MspI digest that has a CCGG recognition motif and thus enriches for ~5% of the genome containing a high density of CpG sites (Gu et al. 2011). We size-selected the resultant MspI-digested fragments by gel electrophoresis, focusing on the 300–700 bp range. We ligated an adaptor to each size-selected DNA fragment pool and performed PCR enrichment of adaptor-ligated DNA using a reverse primer with an individual-specific, CpG-free barcode (mSTARR\_primerR2\_\*wbarcode). All relevant mSTARR primer sequences can be found in Supplemental Table S2.

We created two pools of barcoded libraries that were used as input for our experiments and derived from independent library preparations. Pool 1 contained 20 individuals, and pool 2 con-

tained 21 individuals (Supplemental Table S3). Sixteen individuals are represented in both pools, and nine individuals are included in only one of the two pools. When a sample was included in both pools, we used different multiplexing indices in pool 1 versus pool 2 to ensure that results were reproducible across independent library preparations. We also included a subset of samples twice within the same pool with two different indices ( $n=2$  in each pool). By sequencing most individuals twice with different indices, our goal was to limit any potential impacts that the barcode sequence we included to differentiate individual libraries may have on our results. Unlike some MPRA designs (e.g., Kwasniewski et al. 2012; Patwardhan et al. 2012; Sharon et al. 2012), multiplexed mSTARR-seq relies on barcodes only to pool DNA from multiple individuals and does not rely upon barcodes to quantify the regulatory signal. The inclusion of barcodes is therefore only to ensure representation of all individuals within the assay; however, in future work, one might consider including more than two barcodes per sample.

### mSTARR-seq experiments

The resulting pools of barcoded libraries were used as input into the mSTARR-seq assay, closely following the protocol of Lea et al. (2018). Briefly, we linearized the mSTARR plasmid and inserted the multiplexed MspI-digested libraries using a Gibson assembly. To replicate the assembled plasmid libraries, we transformed each pool into electrocompetent GT115 cells, incubated overnight, and extracted the replicated plasmid libraries. We then exposed half of each plasmid pool to water (mock/sham control) and the other half to a CpG methyltransferase enzyme (*M.SssI*) to experimentally induce methylation. We transfected the treated plasmids into K562 cells, performing three replicates of each condition (three methylated and three unmethylated replicates) for each of the two pools, creating 12 replicates total with 15 million cells each. We incubated the transfected cells for 48 h before performing DNA/RNA extraction. We extracted DNA from about 3 million cells and extracted RNA from the remaining approximately 12 million cells from each of the 12 replicates. We created mRNA sequencing libraries, conducted low-level data processing, and generated replicate-level window-based counts matrices following the method of Lea et al. (2018) (see Supplemental Methods).

### Identification of regulatory and MD regulatory elements

All modeling and data visualization were conducted in R version 4.4.3 (R Core Team 2025). We used linear mixed effects modeling to test for regulatory activity at each 400 bp window in methylated and unmethylated conditions separately. We normalized total read counts for the 11 DNA and 11 RNA replicates using “voomWithQualityWeights,” fit a linear model for each window within each condition using “lmFit,” and calculated test statistics using “eBayes,” all of which are functions available in the R package “limma” (Ritchie et al. 2015). We included sample type (DNA vs. RNA) as our predictor variable, pool (1 vs. 2) as a covariate, and normalized counts as the response variable in each linear model. Using the *P*-value output for each condition, we extracted *P*-values for the sample type term and performed a FDR correction using the R function “qvalue” (implemented in the qvalue package, <https://bioconductor.org/packages/qvalue>). We considered a region to have regulatory activity in a given condition if (1) the sample type (DNA vs. RNA) was a significant predictor of read count at an FDR-corrected *P*-value < 0.01, and (2) the effect size for sample type was more than zero, meaning that the normalized counts of mRNA were greater than the normalized counts of DNA for that

window. This left us with 6221 and 2513 regulatory regions in the unmethylated and methylated conditions, respectively.

To both validate our results and generate a better understanding of the genetic contexts in which mSTARR-seq identifies regulatory activity, we ascertained which chromatin state the regulatory windows are located in by overlapping significant windows with chromatin state annotations generated for K562 cells, lifted over from the hg19 to hg38 genome build (The ENCODE Project Consortium 2012; Hoffman et al. 2013). We ran a Fisher's exact test to quantify enrichment of different chromatin states within our regulatory windows (Supplemental Table S6). We also confirmed our results by reprocessing two previous mSTARR-seq data sets using the same data analysis pipeline outlined above and tested for a correlation in effect sizes and an enrichment in shared regulatory activity (for more detailed description, see Supplemental Results).

We estimated MD regulatory activity within our data set using the "mashr" package in R (Urbut et al. 2019). This package provides tools to generate more accurate effect size estimates across multicondition data sets by exploiting correlations between conditions; it also provides a framework for testing for effect size sharing versus heterogeneity across conditions. Using the results from our linear modeling, we followed the authors' recommendations and input the 6957 regions that met our regulatory activity criteria in at least one condition as the set of "strong sites" that mashr uses to learn the correlation structure between nonnull effects in different conditions; we also included a random set of windows ( $n=20,000$ ) to estimate signals in the data associated with null effects and to distinguish these from nonnull effects. After running mashr with this pipeline, we generated a posterior mean and LFSR for each of our set of 6957 tested regions. We subtracted the posterior mean of the sample type effect estimated in the methylated condition from the unmethylated condition to generate a logFC estimate, which represents the difference in regulatory activity between conditions. We considered a window to have MD activity if the LFSR was  $<0.1$  and the logFC between conditions was  $>1.5$ . To determine if CpG density was associated with methylation dependence, we overlapped MD windows with all CpG sites in the human genome and ran a Student's *t*-test comparing the number of CpG sites found in MD versus non-MD windows. To determine whether CpG number increased MD in a linear manner, we ran a linear model testing the effect of CpG number on the strength of MD, measured as the logFC in expression between the methylated and unmethylated conditions.

### Analysis of allele-specific expression patterns

To conduct ASE analyses, we performed SNP calling, joint genotyping, and quality filtering of variants identified from the DNA extracted from our assay (see Supplemental Methods; Supplemental Fig. S11; Supplemental Table S5). For the 3037 sites in the methylated condition and 4931 sites in the unmethylated condition that passed our filtering criteria, we ran beta-binomial models using the function "betabin" in the "aod" package in R to test whether the ratio of reference to alternate allele counts significantly differed between the DNA input and the mRNA output at each site in the methylated and unmethylated conditions separately (implemented in the aod package in R, <https://cran.r-project.org/package=aod>). We included sample type (DNA vs. RNA) as the predictor variable, pool as a covariate, and the number of reference and alternate allele counts as the response. We explored the genomic context of genetic effect sites utilizing a data set containing both sites with ASE and monoallelic/nearly monoallelic sites (although results are robust to the inclusion of only ASE sites) (see Supplemental Results). We ascertained chromatin states for these

sites by overlapping with the same chromatin state annotations for K562 cells as mentioned above and performed a Fisher's exact test to assess the enrichment of enhancer states.

To explore whether genetic effect sites identified using mSTARR-seq are more comparable to GWAS or eQTL hits based on their genomic location, we utilized the data set containing both ASE and monoallelic/nearly monoallelic sites and summarized the distance to the nearest TSS and promoter annotations for each site using the "annotatePeaks" function in HOMER v 4.11.1 (Heinz et al. 2010). We performed a Student's *t*-test to compare the distance to the nearest TSS and a Fisher's exact test to examine underenrichment of promoter annotations in ASE sites compared with the matched background.

### Analysis of MD genetic effects

To understand how methylation impacts allele-specific expression, we again used mashr to compare effect size estimates between conditions. We tested the 1359 sites that had adequate coverage/variability between replicates to be modeled in both conditions (i.e., were not removed owing to monoallelic/nearly monoallelic expression). We input 783 sites that had strong ASE (at a 1% FDR) in at least one condition and a random set of 783 sites with null effects to generate posterior mean and LFSR values. Beta-binomial modeling uses raw reference and alternate allele counts that are not normalized across libraries, so we  $\log_2$ -transformed the posterior mean of the unmethylated and methylated conditions before calculating the logFC between conditions (unmethylated condition divided by the methylated condition). We considered a SNP to have MD genetic effects if the logFC between conditions was  $>1.5$ .

To explore whether variation in methylation that impacts expression in our assay also occurs naturally in vivo, we tested whether MD genetic effect sites were within  $\pm 200$  bp of variably methylated sites in the genome, classifying variable methylation in two ways: (1) between three individuals within a single blood cell type (monocytes) and (2) within a single individual between five blood cell types (monocytes, NK cells, granulocytes, helper T cells, and B cells). To do so, we used a publicly available cell-sorted whole-genome bisulfite sequencing data set (Loyfer et al. 2023). We classified CpG sites as being variably methylated in vivo if they did not show consistent hyper ( $>90\%$ ), hypo ( $<10\%$ ), or equal methylation across all individuals or cell types. Additionally, we tested whether genotype influences in vivo methylation at MD genetic effect sites by intersecting these sites with meQTL identified in whole blood from the participants of the Framingham Heart Study (Huan et al. 2019).

One potential mechanism by which MD genetic effects might arise is owing to variable binding in TFs. To better understand how methylation may impact allele-specific TF binding, we used the "findMotifsGenome" function in the program HOMER to identify TF binding motifs enriched within a  $\pm 200$  bp range centered around the variant site of interest (Heinz et al. 2010). We split our MD genetic effect sites into two sets: those with a more pronounced genetic effect in the methylated condition ( $n=226$ ) and those with a more pronounced genetic effect in the unmethylated condition ( $n=195$ ). We used sites with ASE whose effects were not modified by methylation status as the background comparison set (methylated condition:  $n=776$ , unmethylated condition:  $n=780$ ). Additionally, we assessed whether the abundance of TF binding motifs translated to differences in TF binding by incorporating ChIP-seq peaks curated by ChIP-Atlas 3.0 (Zou et al. 2024). Using this program, we again asked whether there was enrichment for TF binding among sites with a more pronounced genetic effect in the methylated condition or

unmethylated condition compared with a background set of SNPs that were matched for MAF, LD, and gene density (see [Supplemental Methods](#)). We also required that SNPs included in our background were ones that could have potentially been included in our assay by filtering for SNPs within the 400 bp windows that possessed at least one DNA read in at least half of the replicates in both conditions of our assay, so as not to bias our results toward TFs that bind to GC-rich regions. We set our threshold of significance to include peaks with a  $Q$ -value of less than  $1 \times 10^{-50}$  (generated using MACS2 peak-caller), filtered results for K-562 cells, and report enriched TFs at a 1% FDR.

We further explored the functional significance of MD genetic effect sites by testing for their enrichment near genes and phenotypically relevant GWAS and EWAS sites. We used GREAT (McLean et al. 2010) to test for the enrichment of specific genes as well as their ontology pathways, using the basal plus extension setting and curated regulatory domains. We conducted a Fisher's exact test for enrichment of GWAS hits within  $\pm 400$  bp of our MD genetic effect sites. Using the Pan-UK Biobank and a GWAS association  $P$ -value cutoff of  $5 \times 10^{-8}$ , we tested for sites significantly associated with 20 different blood-specific phenotypes relevant to our experimental cell type (Karczewski et al. 2024). A list of tested traits can be found in [Supplemental Table S12](#). Furthermore, we intersected our MD genetic effect sites with GWAS and EWAS hits linked to diverse phenotypes, using the GWAS catalog v1.0.2 associations (Sollis et al. 2023) and EWAS Atlas associations accessed through the EWAS open platform data hub (Li et al. 2019), downloaded in March 2024. We determined the number of MD genetic effect sites located within  $\pm 400$  bp of GWAS hits and EWAS probes and compared these results to the number of associations that would be found using 1000 comparison data sets of SNPs that have been matched for MAF, LD, and gene density (see [Supplemental Methods](#)). To determine whether MD genetic effect sites are enriched in GWAS or EWAS hits or are associated with a greater number of traits (as each GWAS and EWAS hits can be associated with multiple traits), we performed a permutation test and report the  $P$ -value as the proportion of matched sets having a greater number of associations compared to our test set (MD genetic effect sites).

### Assessing the benefits of multiplexing

To determine how the multiplexing of genetically diverse input fragments impacts our capacity to identify MD genetic effects, we performed two analyses. First, we determined the number of SNPs that would be identified in regulatory fragments when downsampling our data set to combinations of one, two, four, eight, or 12 individuals. When downsampling to a single individual, we extracted the number of heterozygous sites located within a regulatory window for each of the 25 individuals. When downsampling to two, four, eight, or 12 individuals, we created 100 unique combinations of individuals (or the maximum unique combinations possible) for each multiplexing regime and counted the number of variable sites that would have had both the reference and alternate allele present in the assay. To better understand how geographic origin may impact the amount of genetic diversity assayed, we ran this analysis three times, using genotype information from all European sampled individuals, all African sampled individuals, or both.

Second, we compared the maximum number of MD genetic effect sites identified in the current study that would have been included in the assay within different multiplexing regimes. To do so, we summed the number of unique MD genetic effect sites that would have had both alleles represented in the assay when downsampling to one, two, four, eight, or 12 individuals, creating a distribution of the number of MD genetic effect sites present by

sampling 100 unique combinations of individuals to include in each multiplex regime. We performed this analysis three times, subsampling to individuals of European origin, African origin, or all individuals.

### Data access

The DNA and RNA sequencing reads generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA1137064. Count matrices, modeling results, and other data sets used in the text are available at Zenodo (<https://zenodo.org/records/15116803>) and as [Supplemental Data](#). Code is available at GitHub ([https://github.com/rachpetersen/multiplexed\\_mSTARRseq](https://github.com/rachpetersen/multiplexed_mSTARRseq)) and as [Supplemental Code](#).

### Competing interest statement

The authors declare no competing interests.

### Acknowledgments

We thank members of the Lea laboratory, the Hodges laboratory, and the Evolutionary Studies Initiative at Vanderbilt University for their support in completing this work. We specifically thank Emily Hodges, Tyler Hansen, Sai Han Presley, Nicholas Ryan, Rachel Johnston, and Jenny Tung for their advice and input. This work was supported by the Searle Scholars Program and National Institutes of Health/National Institute of General Medical Sciences (1R35GM147267).

*Authors contributions:* Conceptualization was by A.J.L. and C.M.V. Obtaining resources was by A.J.L. Investigation was by all of the authors. Data curation was by R.M.P. Formal analysis was by all of the authors. Visualization was by R.M.P. Writing of the original draft was by R.M.P. and A.J.L. Review and editing were by all of the authors.

### References

- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68. doi:10.1038/nature15393
- Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, Montgomery SB. 2022. Multiple causal variants underlie genetic associations in humans. *Science* **375**: 1247–1254. doi:10.1126/science.abbj5117
- Anderson JA, Lin D, Lea AJ, Johnston RA, Voyles T, Akinyi MY, Archie EA, Alberts SC, Tung J. 2024. DNA methylation signatures of early-life adversity are exposure-dependent in wild baboons. *Proc Natl Acad Sci* **121**: e2309469121. doi:10.1073/pnas.2309469121
- Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077. doi:10.1126/science.1232542
- Banovich NE, Lan X, McVicker G, Van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. 2014. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* **10**: e1004663. doi:10.1371/journal.pgen.1004663
- Breeze CE, Beck S, Berndt SI, Franceschini N. 2022. The missing diversity in human epigenomic studies. *Nat Genet* **54**: 737–739. doi:10.1038/s41588-022-01081-4
- Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, Sunyaev SR, Cotsapas C. 2017. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet* **49**: 600–605. doi:10.1038/ng.3795
- Deplancke B, Alpern D, Gardeux V. 2016. The genetics of transcription factor DNA binding variation. *Cell* **166**: 538–554. doi:10.1016/j.cell.2016.07.012

- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. doi:10.1038/nature11247
- Garg P, Joshi RS, Watson C, Sharp AJ. 2018. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genet* **14**: e1007707. doi:10.1371/journal.pgen.1007707
- Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A, Carvalho-Silva D, Buniello A, et al. 2021. Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* **49**: D1311–D1320. doi:10.1093/nar/gkaa840
- Gong J, Mei S, Liu C, Xiang Y, Ye Y, Zhang Z, Feng J, Liu R, Diao L, Guo A-Y, et al. 2018. PancaQTL: systematic identification of *cis*-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res* **46**: D971–D976. doi:10.1093/nar/gkx861
- Greenberg MV, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**: 590–607. doi:10.1038/s41580-019-0159-6
- The GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330. doi:10.1126/science.aaz1776
- Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**: 468–481. doi:10.1038/nprot.2010.190
- Gurdasani D, Barroso I, Zeggini E, Sandhu MS. 2019. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**: 520–535. doi:10.1038/s41576-019-0144-0
- Han H, Cortez CC, Yang X, Nichols PW, Jones PA, Liang G. 2011. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum Mol Genet* **20**: 4299–4310. doi:10.1093/hmg/ddr356
- Hansen TJ, Fong SL, Day JK, Capra JA, Hodges E. 2024. Human gene regulatory evolution is driven by the divergence of regulatory element function in both *cis* and *trans*. *Cell Genomics* **4**: 100536. doi:10.1016/j.xgen.2024.100536
- Hashimoto H, Zhang X, Vertino PM, Cheng X. 2015. The mechanisms of generation, recognition, and erasure of DNA 5-methylcytosine and thymine oxidations. *J Biol Chem* **290**: 20723–20733. doi:10.1074/jbc.R115.656884
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589. doi:10.1016/j.molcel.2010.05.004
- Hernandez-Corchado A, Najafabadi HS. 2022. Toward a base-resolution panorama of the *in vivo* impact of cytosine methylation on transcription factor binding. *Genome Biol* **23**: 151. doi:10.1186/s13059-022-02713-y
- Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827–841. doi:10.1093/nar/gks1284
- Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development: developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. *Science* **187**: 226–232. doi:10.1126/science.187.4173.226
- Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, et al. 2013. DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2**: e00726. doi:10.7554/eLife.00726
- Huan T, Joehanes R, Song C, Peng F, Guo Y, Mendelson M, Yao C, Liu C, Ma J, Richard M, et al. 2019. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat Commun* **10**: 4267. doi:10.1038/s41467-019-12228-z
- Jin Z, Liu Y. 2018. DNA methylation in human diseases. *Genes Dis* **5**: 1–8. doi:10.1016/j.gendis.2018.01.002
- Johnson GD, Barrera A, McDowell IC, D'Ippolito AM, Majoros WH, Vockley CM, Wang X, Allen AS, Reddy TE. 2018. Human genome-wide measurement of drug-responsive regulatory activity. *Nat Commun* **9**: 5317. doi:10.1038/s41467-018-07607-x
- Johnston RA, Aracena KA, Barreiro LB, Lea AJ, Tung J. 2024. DNA methylation-environment interactions in the human genome. *eLife* **12**: RP89371. doi:10.7554/eLife.89371.3
- Kalita CA, Moyerbrailean GA, Brown C, Wen X, Luca F, Pique-Regi R. 2018. QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* **34**: 787–794. doi:10.1093/bioinformatics/btx598
- Kaluscha S, Domcke S, Wirbelauer C, Stadler MB, Durdu S, Burger L, Schübeler D. 2022. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nat Genet* **54**: 1895–1906. doi:10.1038/s41588-022-01241-6
- Karczewski KJ, Gupta R, Kanai M, Lu W, Tsuo K, Wang Y, Walters RK, Turley P, Callier S, Shah NN, et al. 2024. Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. medRxiv doi:10.1101/2024.03.13.24303864
- Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328**: 232–235. doi:10.1126/science.1183621
- Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samoviča M, Sakthivel MP, Kuzmin I, Trevanion SJ, et al. 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**: 1290–1299. doi:10.1038/s41588-021-00924-w
- Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, Castel SE, Hamel AR, Viñuela A, Roberts AL, et al. 2020. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**: eaaz8528. doi:10.1126/science.aaz8528
- Kreibich E, Kleinendorst R, Barzagli G, Kaspar S, Krebs AR. 2023. Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. *Mol Cell* **83**: 787–802.e9. doi:10.1016/j.molcel.2023.01.017
- Kribelbauer JF, Laptenko O, Chen S, Martini GD, Freed-Pastor WA, Prives C, Mann RS, Bussemaker HJ. 2017. Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep* **19**: 2383–2395. doi:10.1016/j.celrep.2017.05.069
- Kribelbauer JF, Lu X-J, Rohs R, Mann RS, Bussemaker HJ. 2020. Toward a mechanistic understanding of DNA methylation readout by transcription factors. *J Mol Biol* **432**: 1801–1815. doi:10.1016/j.jmb.2019.10.021
- Kwasniewski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc Natl Acad Sci* **109**: 19498–19503. doi:10.1073/pnas.1210678109
- Lea AJ, Vockley CM, Johnston RA, Del Carpio CA, Barreiro LB, Reddy TE, Tung J. 2018. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* **7**: e37513. doi:10.7554/eLife.37513
- Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, Li R, Xia L, Zhang T, Niu G, et al. 2019. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* **47**: D983–D988. doi:10.1093/nar/gky1027
- Loyfer N, Magenheimer J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T, et al. 2023. A DNA methylation atlas of normal human cell types. *Nature* **613**: 355–364. doi:10.1038/s41586-022-05580-6
- Luo C, Hajkova P, Ecker JR. 2018. Dynamic DNA methylation: in the right place at the right time. *Science* **361**: 1336–1340. doi:10.1126/science.aat6806
- Luo X, Zhang T, Zhai Y, Wang F, Zhang S, Wang G. 2021. Effects of DNA methylation on TFs in human embryonic stem cells. *Front Genet* **12**: 639461. doi:10.3389/fgene.2021.639461
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195. doi:10.1126/science.1222794
- McLean CY, Bristol D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28**: 495–501. doi:10.1038/nbt.1630
- Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–277. doi:10.1038/nbt.2137
- Mills MC, Rahal C. 2020. The GWAS diversity monitor tracks diversity by disease in real time. *Nat Genet* **52**: 242–243. doi:10.1038/s41588-020-0580-y
- Mohammed J, Hansen K, Claes P, Weinberg S, Selleri L, Swigut T, Wysocka J. 2022. Making the human face: elucidating the role of enhancers in hominid craniofacial evolution. *FASEB J* **36**. doi:10.1096/fasebj.2022.36.S1.01629
- Monteagudo-Sánchez A, Noordermeer D, Greenberg MV. 2024. The impact of DNA methylation on CTCF-mediated 3D genome organization. *Nat Struct Mol Biol* **31**: 404–412. doi:10.1038/s41594-024-01241-6
- Mostafavi H, Spence JP, Naqvi S, Pritchard JK. 2023. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**: 1866–1875. doi:10.1038/s41588-023-01529-1
- Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, Dumaine A, Danckaert A, Luca F, Grenier J-C, et al. 2015. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res* **25**: 1801–1811. doi:10.1101/gr.192005.115

- Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**: 1173–1175. doi:10.1038/nbt.1589
- Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265–270. doi:10.1038/nbt.2136
- Perera BP, Faulk C, Svoboda LK, Goodrich JM, Dolinoy DC. 2020. The role of environmental exposures and the epigenome in health and disease. *Environ Mol Mutagen* **61**: 176–192. doi:10.1002/em.22311
- Pulit SL, Voight BF, de Bakker PI. 2010. Multiethnic genetic association studies improve power for locus discovery. *PLoS One* **5**: e12600. doi:10.1371/journal.pone.0012600
- R Core Team. 2025. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>.
- Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenet Genome Res* **14**: 9–25. doi:10.1159/000130315
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47. doi:10.1093/nar/gkv007
- Salameh Y, Bejaoui Y, El Hajj N. 2020. DNA methylation biomarkers in aging and age-related diseases. *Front Genet* **11**: 480672. doi:10.3389/fgene.2020.00171
- Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521–530. doi:10.1038/nbt.2205
- Singh A, Zhong Y, Nahlawi L, Park CS, De T, Alarcon C, Perera MA. 2021. Incorporation of DNA methylation into eQTL mapping in African Americans. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pp. 244–255. World Scientific, Singapore.
- Siraj L, Castro RI, Dewey H, Kales S, Nguyen TTL, Kanai M, Berenzy D, Mouri K, Wang QS, McCaw ZR, et al. 2024. Functional dissection of complex and molecular trait variants at single nucleotide resolution. bioRxiv doi:10.1101/2024.05.05.592437
- Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2023. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**: D977–D985. doi:10.1093/nar/gkac1010
- Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2016. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**: 1519–1529. doi:10.1016/j.cell.2016.04.027
- Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS. 2016. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**: 1530–1545. doi:10.1016/j.cell.2016.04.048
- Umans BD, Battle A, Gilad Y. 2021. Where are the disease-associated eQTLs? *Trends Genet* **37**: 109–124. doi:10.1016/j.tig.2020.08.009
- Urbut SM, Wang G, Carbonetto P, Stephens M. 2019. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**: 187–195. doi:10.1038/s41588-018-0268-8
- Van De Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061–1063. doi:10.1038/nmeth.3582
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**: 11.10.11–11.10.33. doi:10.1002/0471250953.bi1110s43
- Vockley CM, Guo C, Majoros WH, Nodzinski M, Scholtens DM, Hayes MG, Lowe WL, Reddy TE. 2015. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res* **25**: 1206–1214. doi:10.1101/gr.190090.115
- Vösa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, et al. 2021. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**: 1300–1310. doi:10.1038/s41588-021-00913-z
- Waterland RA, Jirtle RL. 2003. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol* **23**: 5293–5300. doi:10.1128/MCB.23.15.5293-5300.2003
- Weaver IC, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. 2004. Epigenetic programming by maternal behavior. *Nat Neurosci* **7**: 847–854. doi:10.1038/nn1276
- Yim YY, Teague CD, Nestler EJ. 2020. In vivo locus-specific editing of the neuroepigenome. *Nat Rev Neurosci* **21**: 471–484. doi:10.1038/s41583-020-0334-y
- Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al. 2017. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**: eaaj2239. doi:10.1126/science.aaj2239
- Zeng B, Bendt J, Kosoy R, Fullard JF, Hoffman GE, Roussos P. 2022. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat Genet* **54**: 161–169. doi:10.1038/s41588-021-00987-9
- Zeng Y, Jain R, Lam M, Ahmed M, Guo H, Xu W, Zhong Y, Wei G-H, Xu W, He HH. 2023. DNA methylation modulated genetic variant effect on gene transcriptional regulation. *Genome Biol* **24**: 285. doi:10.1186/s13059-023-03130-5
- Zou Z, Ohta T, Oki S. 2024. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Res* **52**: W45–W53. doi:10.1093/nar/gkac358

Received August 23, 2024; accepted in revised form June 17, 2025.