High-throughput allele-specific expression across 250 environmental conditions

Gregory A. Moyerbrailean,[1,4] Allison L. Richards,[1,4] Daniel Kurtz,[1,4] Cynthia A. Kalita,[1,4] Gordon O. Davis,[1] Chris T. Harvey,[1] Adnan Alazizi,[1] Donovan Watza,[1] Yoram Sorokin,[2] Nancy Hauff,[2] Xiang Zhou,[3] Xiaoquan Wen,[3] Roger Pique-Regi,[1,2] and Francesca Luca[1,2]

[1]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan 48201, USA; [2]Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan 48201, USA; [3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA

Gene-by-environment (GxE) interactions determine common disease risk factors and biomedically relevant complex traits. However, quantifying how the environment modulates genetic effects on human quantitative phenotypes presents unique challenges. Environmental covariates are complex and difficult to measure and control at the organismal level, as found in GWAS and epidemiological studies. An alternative approach focuses on the cellular environment using in vitro treatments as a proxy for the organismal environment. These cellular environments simplify the organism-level environmental exposures to provide a tractable influence on subcellular phenotypes, such as gene expression. Expression quantitative trait loci (eQTL) mapping studies identified GxE interactions in response to drug treatment and pathogen exposure. However, eQTL mapping approaches are infeasible for large-scale analysis of multiple cellular environments. Recently, allele-specific expression (ASE) analysis emerged as a powerful tool to identify GxE interactions in gene expression patterns by exploiting naturally occurring environmental exposures. Here we characterized genetic effects on the transcriptional response to 50 treatments in five cell types. We discovered 1455 genes with ASE (FDR < 10%) and 215 genes with GxE interactions. We demonstrated a major role for GxE interactions in complex traits. Genes with a transcriptional response to environmental perturbations showed sevenfold higher odds of being found in GWAS. Additionally, 105 genes that indicated GxE interactions (49%) were identified by GWAS as associated with complex traits. Examples include *GIPR*–caffeine interaction and obesity and include *LAMP3*–selenium interaction and Parkinson disease. Our results demonstrate that comprehensive catalogs of GxE interactions are indispensable to thoroughly annotate genes and bridge epidemiological and genome-wide association studies.

[Supplemental material is available for this article.]

For complex traits, a mismatch between genotype and environment can cause a higher disease risk. However, it is generally difficult to determine the relevant environmental factors to measure in order to study gene-by-environment (GxE) interactions. Consequently, some of the genetic effect sizes measured in GWAS may be underestimated when the relevant environmental factors are not controlled. Molecular phenotypes measured in tightly controlled cellular environments provide a more tractable setting in which we can study GxE interactions simplifying both complex phenotypes and environments (Fig. 1A). The cellular environment is determined by the complex of stimuli (e.g., hormonal and metabolic) to which a cell is exposed, can be defined as an agent that can potentially change the state of the cell, and is measurable at the molecular level. Examples include, agents secreted by nearby cells, hormones and metabolites secreted by other organs, pollutants, drugs, or micronutrients absorbed by the organisms. For example, physical or emotional environmental stressors alter blood glucocorticoid levels, which induce significant cellular changes in global gene expression patterns mediated through glucocorticoid receptor (GR) activation (Grundberg et al. 2011; Luca et al. 2013). Response expression quantitative trait loci (reQTL) mapping studies found that SNPs associated with specific immune traits are enriched for infection reQTL and for expression quantitative trait loci (eQTL) identified only in infected cells (Barreiro et al. 2012; Fairfax et al. 2014; Lee et al. 2014). However, eQTL mapping requires a large number of samples, thus limiting the number of cellular environments that can be analyzed in parallel. While association mapping compares genotypic effects across individuals that have different genetic backgrounds, allele-specific expression (ASE) approaches compare allelic effects within individuals, thereby controlling for genetic background and cellular environment. Currently, ASE approaches represent the most effective assay to quantify *cis*-regulatory variants within a defined cellular environment and to control for *trans*-acting modifiers of gene expression, such as genotype at other loci (Kasowski et al. 2010; McDaniell et al. 2010; Pastinen 2010; Skelly et al. 2011; Cowper-Sal·lari et al. 2012; Reddy et al. 2012; McVicker et al. 2013; Buil et al. 2014; Hasin-Brumshtein et al. 2014; Kukurba et al. 2014; Knowles et al. 2015; Kumasaka et al. 2016).

Here we developed a high-throughput in vitro system to characterize the response to tightly controlled environmental exposures. We then used ASE analysis to identify GxE interactions in

[4]These authors contributed equally to this work.
Corresponding authors: fluca@wayne.edu, rpique@wayne.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.209759.116. Freely available online through the *Genome Research* Open Access option.

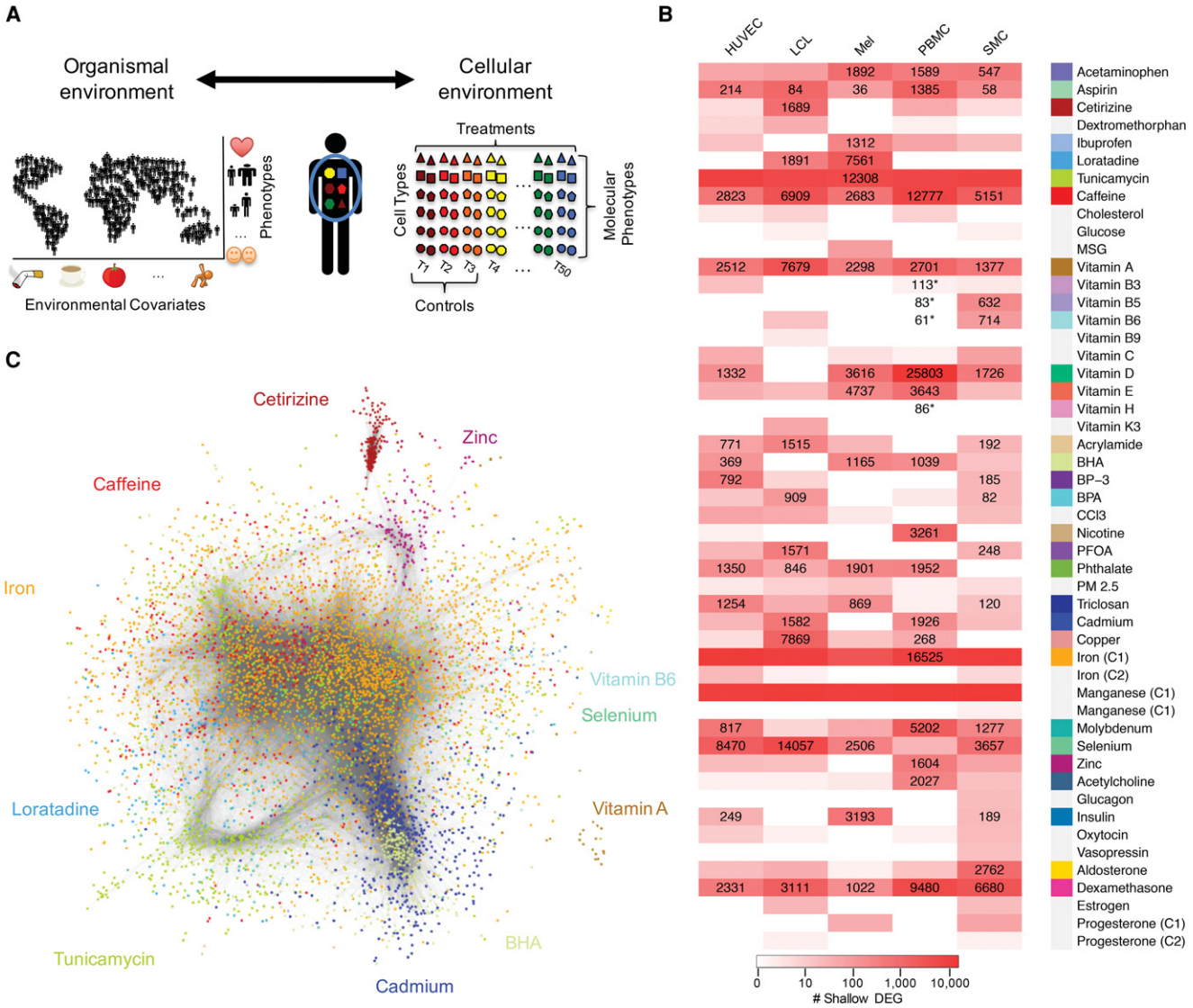

Figure 1. Overview of gene expression response. (*A*) Schematic of experimental design and rationale. Our approach uses specific treatment conditions as tightly controlled proxy for the organism environment and measures molecular phenotypes, such as gene expression, to infer genetic and molecular mechanisms for complex traits. (*B*) Heatmap of differential gene expression. Shown for each cell type (columns) and treatment (rows) combination are the number of differentially expressed genes (10% FDR and $|\log_2 FC| > 0.25$). The shade of red indicates the number of differentially expressed genes from an initial screening step (see Supplemental Texts 5 and 8.1). Cellular environments with a strong response were further sequenced to a higher depth (>58 M reads, 113 M on average), and the number of differentially expressed genes is indicated by the text. Environments marked with an asterisk were chosen to confirm that treatments with a small response from the shallow sequencing data similarly have a small response when deep sequenced. Colors next to treatment names represent treatments chosen for deep sequencing. Gray indicates treatments that were not deep sequenced. (*C*) Global coexpression network inferred using weighted gene correlation network analysis (WGCNA) on 14,535 genes. Each dot represents a gene. Each module is assigned a color based on the treatment with the highest eigengene *t*-value. Note that colors representing treatments are consistently used across all the figures.

individual samples at heterozygous sites. We focused on 50 treatments (Supplemental Table S1) in five cell types (250 conditions) across 15 individuals (three samples per cell type), including paired vehicle-controls. These treatments represent the cellular counterparts of a range of organismal exposures (Supplemental Table S1). We broadly grouped the treatments into six categories: steroid hormones, peptide hormones, metal ions, dietary components, common drugs, and environmental contaminants. For each treatment, we used the metabolically active form detected in the bloodstream at the highest physiological concentration reported by the Mayo Clinic (http://www.mayomedicallaboratories.com) or the CDC (http://www.cdc.gov/biomonitoring/), as available. Our goal is to identify GxE interactions across these conditions and characterize their roles in complex traits.

Results

High-throughput characterization of transcriptional responses

Literature reports on transcriptional responses for many treatments across cell types are often contradictory or nonexistent. To characterize transcriptional responses to 50 environmental

perturbations (Supplemental Table S1) in five cell types (250 conditions), we utilized a high-throughput two-step RNA-seq approach (Supplemental Fig. S2; Moyerbrailean et al. 2015). In step one, we used shallow RNA-seq (8.2 million reads/sample on average) (Supplemental Table S2) and DESeq2 (Love et al. 2014) to coarsely characterize global changes in gene expression (Supplemental Fig. S3; Supplemental Table S3). We considered only treatment-by-cell-type combinations with more than 80 differentially expressed (DE) genes detected at 10% FDR and corresponding to $|\log_2 FC| > 0.25$. We found eight treatments that induced gene expression changes across all cell types, such as dexamethasone and vitamin A, while other treatments had a cell-type–specific effect, such as vitamin B6 in peripheral blood mononuclear cells (PBMCs) (Fig. 1B). Of the 50 treatments, 16 did not induce significant changes in gene expression in any cell type. We excluded a few outlier response conditions (see Supplemental Methods 8.2). By using these criteria, we selected 89 conditions (35 treatments across five cell types and three individuals) corresponding to 297 RNA-seq libraries and resequenced them to 130 M reads per sample on average (Supplemental Table S4) in step two.

Treatment-specific gene coexpression network

In step two, we used deep sequencing data and weighted gene correlation network analysis (WGCNA) (Langfelder and Horvath 2008) to infer the global gene coexpression network across all samples and environments (Fig. 1C). This network comprised 87 modules, which grouped genes with similar expression patterns that may be coregulated in a concerted way. The largest module contained 1456 genes, and the median module size was 42 genes. We assigned a representative treatment to each module based on the most significant treatment effect size on gene expression (Supplemental Fig. S11), and this allowed us to identify clusters of genes that represent treatment-specific responses. For example, two modules were each associated with treatment conditions in opposite directions: module 30 (vitamin D in HUVECs and PBMCs) and module 22 (caffeine and aspirin in smooth muscle cells [SMCs]) (Supplemental Figs. S12, S13). These results suggest that analysis of transcriptional responses to a large number of treatments in parallel can identify gene regulatory networks that mediate divergent effects that depend on specific cell type or treatment conditions.

Analysis of ASE

We used QuASAR (quantitative allele-specific analysis of reads) to identify genes with evidence of ASE. QuASAR (Harvey et al. 2015) identifies heterozygous genotypes and uses a beta binomial distribution to infer ASE in RNA-seq data. In the 89 treatment conditions, we identified 11,305 instances of ASE (10% FDR) (Fig. 2), corresponding to 1455 unique ASE genes. In an individual sample, 0.92% of expressed genes with heterozygous SNPs are ASE genes, on average.

The ASE analysis was performed on all expressed genes and was not limited to DE genes as some genes may not change total expression level. When we consider all ASE genes in our data set, 92% were also identified with eQTL in the Genotype-Tissue Expression (GTEx) Project. Thus, similar to other ASE studies (Buil et al. 2014), we were able to capture genes with regulatory variants that were previously identified at baseline. Additionally, many of the genes identified in GTEx with eQTL may have unknown latent environmental components modulating their expression.

Analysis of GxE interactions

Next, we characterized GxE interactions on gene expression by analyzing ASE differences between treatment and control. Reliably estimating ASE effect sizes required a significant amount of reads for detecting condition-specific ASE (cASE) (see Fig. 3A; Supplemental Fig. S15). However, some genes had very low expression levels in the control condition and high expression with ASE following treatment, suggesting that expression of these genes would be induced by a specific treatment. For these cases, ASE in the control condition cannot be defined accurately. We denoted this phenomenon as induced ASE (iASE) (see Fig. 3B), which

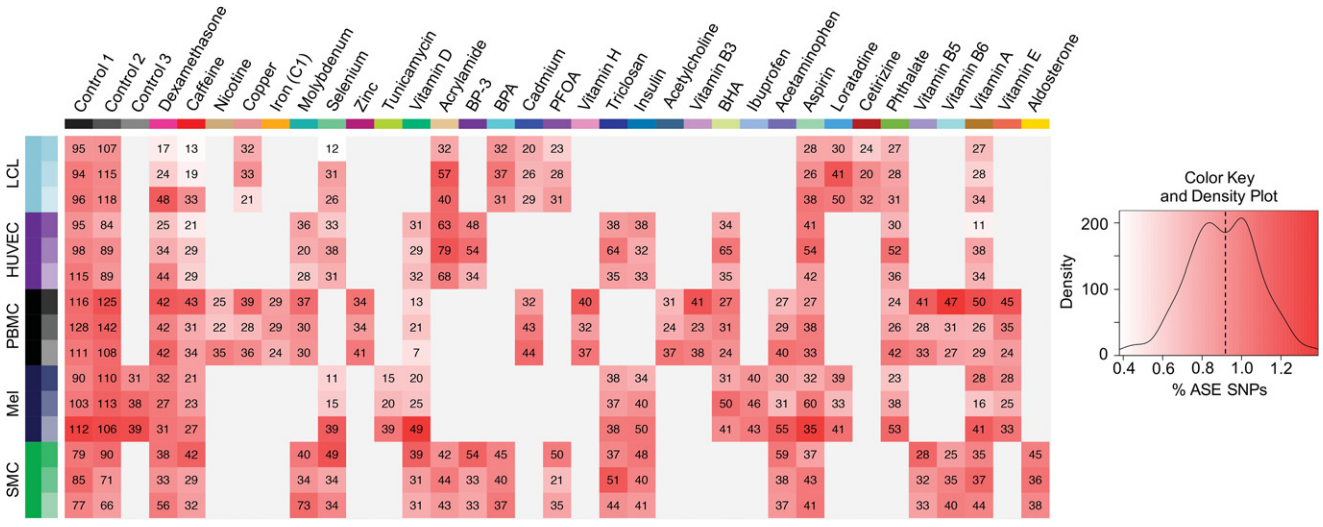

Figure 2. Heatmap of allele-specific expression (ASE). For each individual (row) and treatment (column) we list the number of SNPs displaying ASE (as determined in QuASAR [quantitative allele-specific analysis of reads] at 10% FDR). The shade of red represents the fraction of ASE SNPs to the number of heterozygous SNPs tested (% ASE SNPs) in a given sample and condition. The dotted line on the density plot indicates the average % ASE SNPs across all individual samples and conditions.

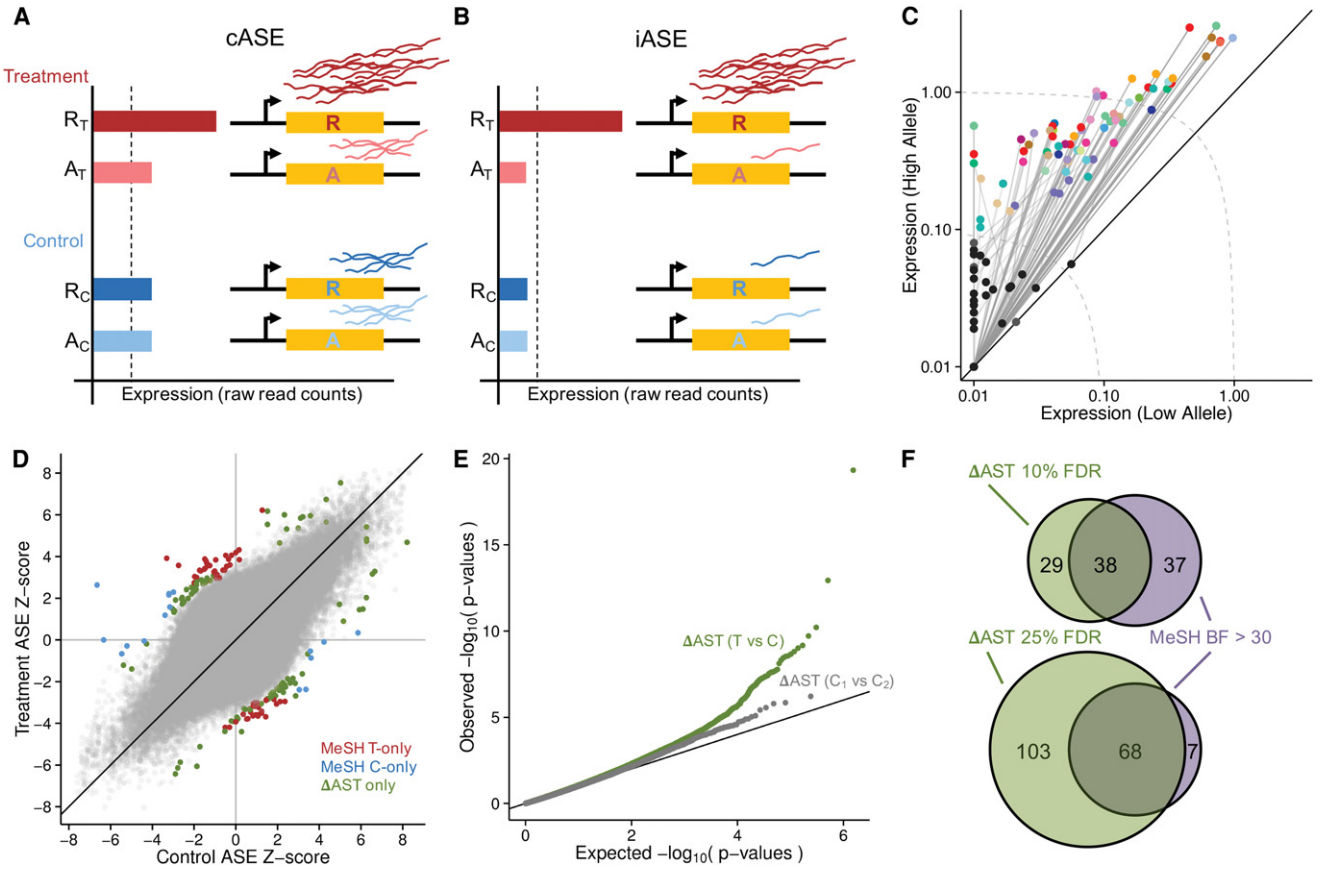

Figure 3. Gene–environment interactions. (*A*,*B*) Two types of gene–environment interactions: conditional ASE (*A*) and induced ASE (*B*). Treatment conditions are in red and control conditions in blue, with the shade (dark/light) representing the allele (reference/alternate). In this example of cASE, there is an imbalance of expression between the two alleles in the treatment condition, while the control shows balanced expression. iASE is defined by an imbalance of expression between the two alleles in the treatment condition and by expression below detectable levels (dotted line) in the control condition. (*C*) Plot of all iASE SNPs detected. Each iASE SNP is represented as two points (representing treatment and control expression) connected by a line (representing the fold-change between conditions). SNPs are plotted based on the expression (TPM [tags per million]) of each allele, with the higher expressed allele in the treatment on the *y*-axis and the lower allele on the *x*-axis. Points are colored by treatment (controls are black and gray), and the dotted lines represent constant expression levels 0.1, 1, and 10. For ease of display, expression of SNPs <0.01 have been set to 0.01. (*D*) Scatter plot of the *Z*-scores in the paired treatment and control samples for all SNPs tested for cASE. Colored points indicate those displaying cASE: Red is SNPs identified by meta-analysis of subgroup heterogeneity (MeSH) as having cASE in the treatment, blue is SNPs identified by MeSH as having cASE in the control, and green is SNPs identified by ΔAST (differential allele-specific test) that were not identified by MeSH. (*E*) QQ-plot of *P*-values for cASE identified with the ΔAST method for treatment versus control (green line) and Control 1 versus Control 2 (gray line). (*F*) Venn diagrams showing the number of cASE SNPs identified by two methods: MeSH and ΔAST at different empirically estimated FDR thresholds.

indicated cases when the ASE was only observed in genes induced by the treatment. Studies that only consider baseline eQTL or ASE may fail to characterize or may mischaracterize genes with iASE if the relevant environmental stimulus is present as latent exposure. We identified 75 iASE SNPs (10% FDR) corresponding to 60 unique genes (Fig. 3C). The genetic effect in these iASE SNPs is slightly stronger than that of baseline ASE (Supplemental Fig. S17).

When we can reliably measure ASE in both treatment and control conditions for the same SNP and individual, we can contrast the amount of ASE between the conditions to determine cASE. ASE across cell types was never contrasted because the samples correspond to different individuals. Here we used two approaches to identify cASE (see Fig. 3D): (1) a *qualitative* "on/off" approach using a meta-analysis framework, and (2) a new *quantitative* approach to detect ASE changes even when ASE is present in both conditions.

For the qualitative analysis of cASE, we used meta-analysis of subgroup heterogeneity (MeSH) (Wen and Stephens 2014), a Bayesian meta-analysis approach that has been previously used in eQTL studies to contrast effect sizes across conditions (Maranville et al. 2011) and tissues (Flutre et al. 2013). Here, for each SNP, individual, and treatment/control experiment pair, we assume four different mutually exclusive models: (1) no ASE in either condition, (2) ASE in both conditions, (3) ASE in treatment only, or (4) ASE in control only. Configurations 3 and 4 represent cASE, while configuration 2 represents shared ASE accommodating for random effects in the genetic effect size. This results in a stringent test for cASE.

For each of the QuASAR treatment/control measurement pairs, MeSH calculated a Bayes factor (BF) for each configuration. We observed that the majority of genes had ASE shared between the treatment and control conditions (Fig. 3D). We identified 75 SNPs with cASE (difference in the $BF_{treatment}$ or $BF_{control}$ and the $BF_{shared} > 30$) corresponding to 71 unique genes. We observed a larger proportion of treatment-only cASE compared with control-only cASE (59 vs. 16) (Figs. 3D, 4). These proportions are consistent

with observations from eQTL studies contrasting individual treatments and tissues (Maranville et al. 2011; Fairfax et al. 2012; Flutre et al. 2013; Mangravite et al. 2013).

MeSH detects qualitative interactions and strictly requires ASE only in one condition analyzed, either treatment or control, while showing no ASE in the alternate condition. However, these extreme on/off ASE cases are rare. Other cASE models may exist. For example, a prior report identified eQTL with genetic effects in opposite directions in the treatment and control conditions in stimulated monocytes (Fairfax et al. 2014). Additionally, the majority of GxE interactions can arise in cases where the genetic effects differ significantly between treatment and control conditions, but they are nonzero in both conditions.

To capture cASE genes that may have ASE in both conditions but of a significantly different magnitude, we developed an alternate approach named ΔAST (differential allele-specific test). The goal was to quantitatively detect GxE interactions. For each heterozygous site, we compared the QuASAR-derived ASE estimates following treatment to those observed in the matched control for each individual. We calculated a *P*-value for the difference in ASE between the two conditions (Fig. 3E).

A key component of our experimental approach is the inclusion of two sets of vehicle controls in each experimental batch, which empirically estimates the true underlying FDR for identifying cASE, equivalent to permutation-based approaches used in eQTL studies (see Supplemental Methods 10.3). By use of ΔAST, we identified 67 cASE SNPs corresponding to an FDR of 10%, 38 of these cASE SNPs were also identified by MeSH. When we relaxed the FDR threshold (25%), we found a total of 178 cASE SNPs corresponding to 160 genes. Of these genes, 65 were identified with both methods (Figs. 3D–F, 4).

Features of GxE interactions

When we considered all cASE SNPs, we observed a significant positive correlation between the gene expression $\log_2$ (fold change) after treatment and differences in the genetic effect in treatment and control samples (Fig. 5A). This result could be a consequence of increased power to detect ASE with more highly expressed genes. However, we observed a negative correlation between gene expression levels and the difference in the genetic effect in the treatment and the control samples (Fig. 5B). This finding suggests that if a gene has sufficiently high expression to test for ASE in both conditions, stronger cASE occurs at genes with stronger positive responses to treatment.

To determine if we can validate some of the few previously known reQTL with cASE, we compiled a list of 134 genes reported to have GxE effects from prior reQTL and ASE studies. Of these genes, 83 were heterozygous and could be tested for cASE in our data (Maranville et al. 2011, 2013; Idaghdour et al. 2012; Franco et al. 2013; Mangravite et al. 2013; Çalışkan et al. 2015; Knowles et al. 2015). Sixty-three out of 83 genes were identified as cASE genes in our analysis ($P < 0.05$) (Supplemental Table S13). For example, gene *IRF5* has a rhinovirus-reQTL (Çalışkan et al. 2015) and showed cASE in response to two treatments: phthalate ($P = 0.04$) and vitamin B5 ($P = 0.04$). Additionally, *IRF5* is also linked to autoimmune responsivity through its association to lupus (Sigurdsson et al. 2005; Graham et al. 2006, 2007). Interestingly, phthalates may play a role in lupus etiology since they induce anti-DNA antibodies (Lim and Ghosh 2005), while vitamin B5 deficiency is found in lupus patients (Leung 2004).

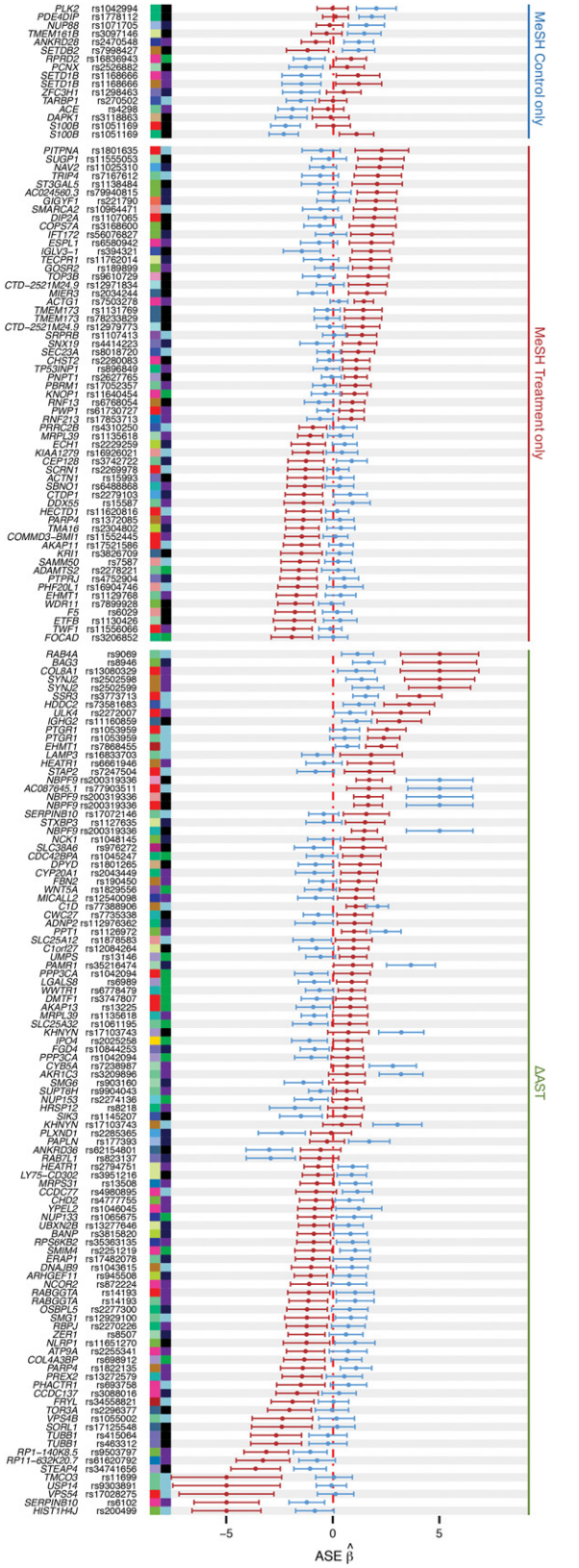

Figure 4. Forest plot of all cASE SNPs. Each row shows the ASE $\hat{\beta}$ for paired treatment (red) and control (blue) conditions. Defined as in Figure 2, colored squares indicate the treatment (*left*) and cell type (*right*) in which cASE was identified, along with the gene name and SNP rsID.

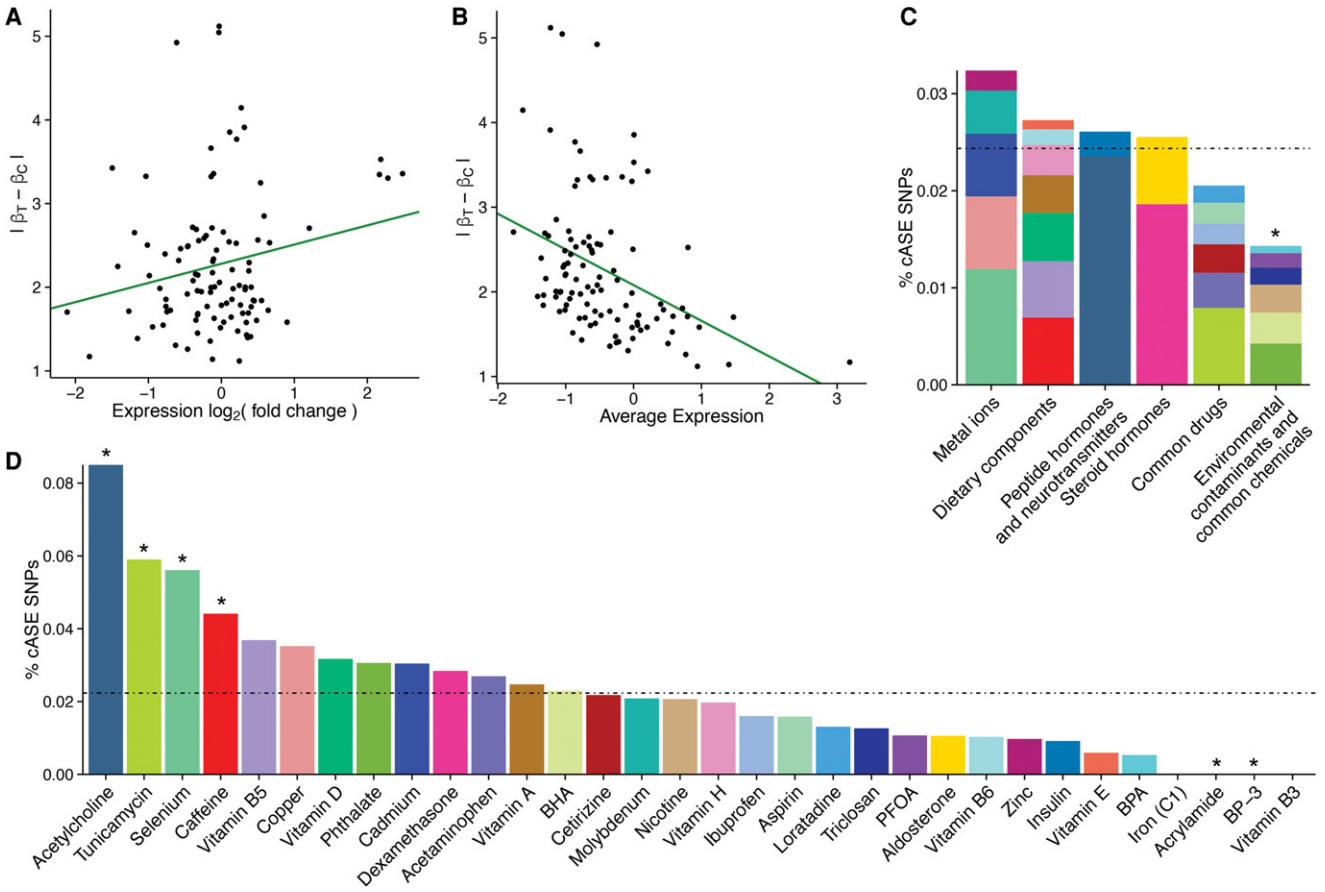

Figure 5. Features of cASE SNPs. (*A*,*B*) Scatter plot comparing the absolute difference in ASE $\hat{\beta}$ between treatment and control (*y*-axis) and the average $\log_2$ (expression; *A*) or $\log_2$ (fold change; *B*) between treatment and control samples for cASE SNPs. The green line indicates the trendline from a linear model fit on the points. (*C*,*D*) Percentage of cASE SNPs identified in each treatment category (*C*) or treatment (*D*). For each group, plotted is the percentage of cASE SNPs identified, relative to the number of SNPs tested for that group. The dotted black line represents the average percentage of cASE SNPs across all groups. Groups with an asterisk are significantly enriched or depleted (binomial *P*-value <0.05) relative to the average. The colors in *C* represent the relative proportion of cASE SNPs for each treatment in a treatment category.

We next wanted to determine the difference in the number of cASE SNPs across cellular environments (Fig. 5C,D; Supplemental Fig. S18). We found that acetylcholine, selenium, and caffeine had significantly higher numbers of cASE SNPs compared with the mean number of cASE SNPs per treatment, while acrylamide and BP-3 had significantly fewer (binomial, $P < 0.05$). In addition, we found that environmental contaminants and common chemicals have a significantly lower proportion of cASE SNPs ($P < 0.003$).

To assess the extent of GxE interactions on gene regulation in different cellular environments, we developed an index that aggregates all cASE tests for each condition and cell type and determines how much the environment can globally perturb ASE. To achieve this, we can compare the correlation of the standardized effect size between treatment and control (as shown in Fig. 3D) across all SNPs tested for a given cell type and condition. We denote this genome-wide measurement as environmental displacement of genetic effect (EDGE) index (for more details, see Methods). Specifically, the EDGE index is the ratio between the pair-wise correlation observed between the two control sets and the correlation observed between the treatment and control conditions. The EDGE index is one for the control conditions (Supplemental Fig. S19A) and will have higher values for treatments that can affect ASE for a large number of genes (Supplemental Fig. S19B). We find significant differences in the EDGE index across many treatments (Supplemental Fig. S20). As expected, we found a high correlation between the proportion of cASE SNPs and the EDGE index (Spearman $\rho = 0.717$, $P = 3.8 \times 10^{-6}$) (Supplemental Fig. S19C).

GxE interactions and complex traits

We then used the GxE interactions we identified in vitro to characterize putative molecular mechanisms for risk or protective environmental factors for complex traits (Fig. 6A). We found that 22% of DE genes overlap with those identified in GWAS analyses (Welter et al. 2014) compared with 4% of nondifferentially expressed genes expressed in our samples (Fig. 6B). This overlap corresponds to a sevenfold enrichment ($P < 2.2 \times 10^{-16}$). These results suggest that genes responsive to our treatments are more likely involved in organismal traits.

To investigate the role of GxE interactions in complex traits directly, we analyzed genes containing iASE or cASE. Forty-nine percent of genes (105 out of 215) that contain either iASE or cASE were identified by GWAS as associated with various complex traits; this corresponds to a 3.5-fold enrichment ($P < 2.2 \times 10^{-16}$) compared with DE genes without iASE or cASE. Importantly, genes

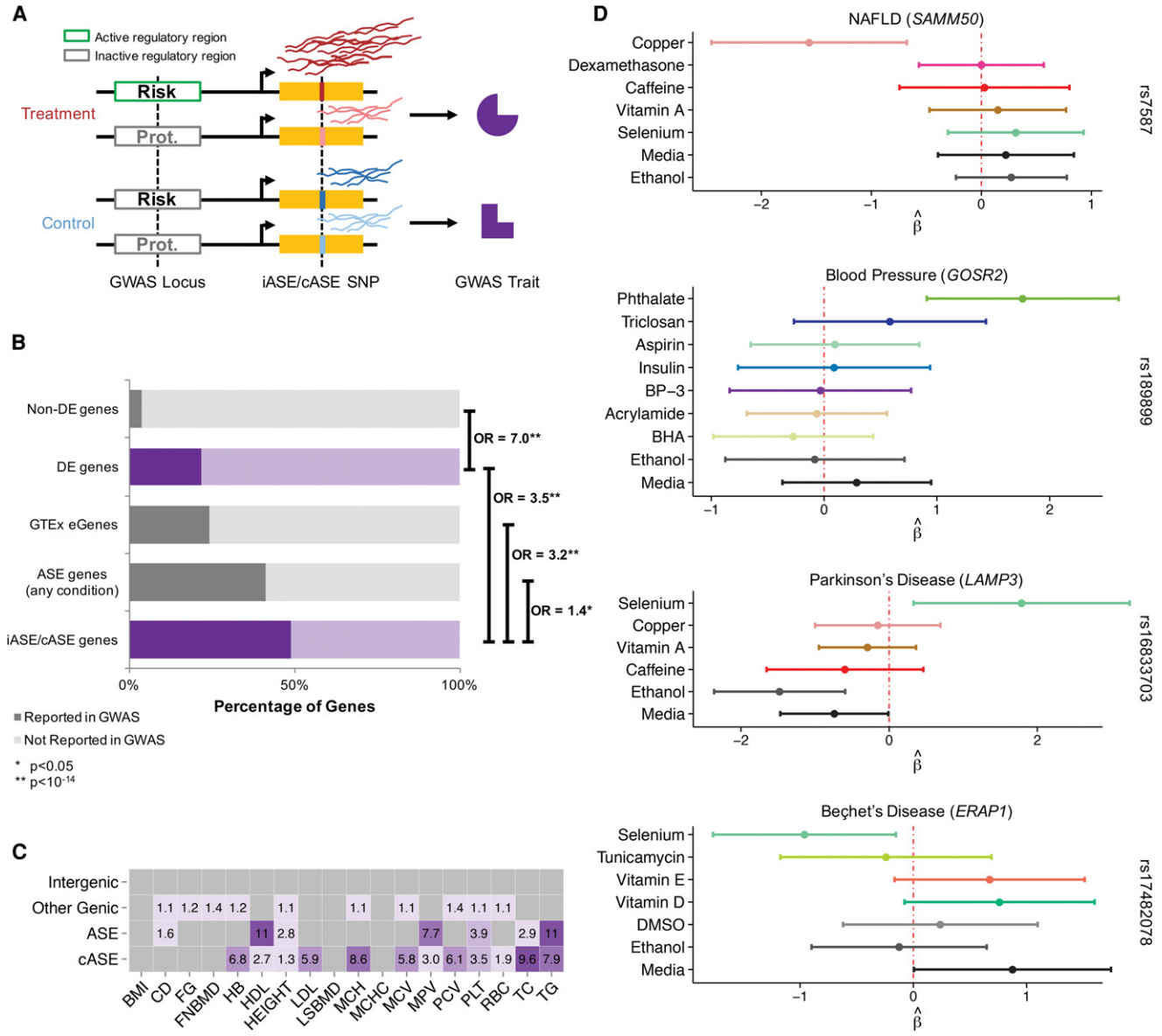

Figure 6. Integration with GWAS. (*A*) Hypothetical model detailing the use of GxE interactions to characterize putative molecular mechanisms for risk or protective environmental factors for complex traits. In the treatment environment, a regulatory region is either active or inactive depending on the haplotype, therefore resulting in different levels of gene expression. In the control environment, the regulatory region is inactive regardless of haplotype. Risk and protective haplotypes are identified in GWAS. (*B*) Enrichment analysis of GWAS genes. Reported genes from the GWAS catalog (version 1.0.1) were compared to different gene sets analyzed in this study: (1) genes that were not differentially expressed in any condition, (2) genes that were differentially expressed in any condition, (3) genes previously associated with an eQTL in GTEx (eGenes) (The GTEx Consortium 2015), (4) genes containing ASE in any condition, and (5) genes containing either iASE or cASE. The percentage of genes in these data sets that were found in the GWAS catalog is indicated by a darker shade. Genes that can be perturbed by our environments are highlighted in purple and indicate a GxE mechanism for the GWAS association. Odds ratios and enrichment *P*-values were calculated using a Fisher's exact test and are shown on the *right* for each pair of gene categories contrasted. (*C*) Genome-wide efficient mixed model association (GEMMA) per SNP heritability estimates relative to the genomic average for cASE (SNPs in genic regions with cASE or iASE), ASE (SNPs in genic regions with ASE), other genic (SNPs in genic regions), and intergenic (SNPs <100 kb from any gene). Only significant enrichment values are reported, with a darker tone of purple indicating a higher enrichment odds ratio relative to the genome average. (*D*) Forest plots of four cASE SNPs in genes associated with a GWAS trait. For each SNP, shown is the ASE $\hat{\beta}$ for each treatment in which the SNP was tested. The 95% CI bars are colored for each treatment as in Figure 2.

with iASE or cASE also have a 1.4-fold increased relevance for complex traits ($P = 0.025$) compared with ASE genes and a 3.2-fold enrichment ($P < 4.3 \times 10^{-15}$) compared with genes with eQTL identified at baseline (eGenes from GTEx). Note that by design much of our detected ASE may have an environmental component, but we may lack the power to claim cASE/iASE. These results suggest that GxE interactions represent an important mechanism for inter-individual variation in complex traits.

We find similar results when we analyze per SNP heritability for 18 complex traits using genome-wide efficient mixed model association (GEMMA) (Zhou and Stephens 2012; Zhou 2016). Similar to the LD-score regression method that partitions

heritability estimates across SNPs functional categories (Gusev et al. 2014), we contrasted SNPs in genes with cASE/iASE, genes with ASE, genes without ASE, and inter-genic regions. The per SNP heritability for each of these categories is then compared with the genome average. A higher value of per SNP heritability for one of these categories indicates a higher number of causal SNPs, higher effect sizes, or both in that category. We found that the per SNP heritability estimate for SNPs in genes with ASE is 11.1 times higher than the genome average for high-density lipoprotein (HDL). For 13 of the 18 traits analyzed, per SNP heritability estimates for SNPs in genes with cASE, iASE, or ASE were significantly higher than the genome average. For seven of these, the cASE and iASE category estimates were even higher than any other partition (Fig. 6C), thus indicating that GxE interaction for these traits are particularly relevant. The highest values for cASE and iASE were observed for blood total cholesterol level (TC; 9.7-fold), triglicerides (TG; 7.9-fold), and mean corpuscular hemoglobin levels (MCH; 8.6-fold). Overall, these results suggest an important role for GxE interaction in a large number of traits.

When we isolated genes with iASE, we found 28 genes associated with a phenotype in the GWAS catalog (Supplemental Table S18). Additional investigation into these genes may yield insights not only on the GxE role in specific traits but also on the underlying molecular mechanisms. For example, previous reports show that caffeine prevents and treats obesity presumably through mitotic clonal expansion effects (Li et al. 2015; Kim et al. 2016; Ohara et al. 2016). Our work suggests that caffeine activates the GIPR pathway, which regulates insulin production. *GIPR* is linked to obesity and several obesity-related traits, including body mass index and type 2 diabetes (Saxena et al. 2010; Speliotes et al. 2010; Fox et al. 2012; Okada et al. 2012; Wen et al. 2012, 2014; Berndt et al. 2013; Mahajan et al. 2014). We identified a SNP, rs5390, in *GIPR* that demonstrates iASE following caffeine treatment. Specifically, we found higher *GIPR* expression and ASE favoring the rs5390 reference allele following caffeine treatment and low expression in controls. The rs5390 reference allele is located on the same haplotype as the nonrisk allele for body mass index in the individual sample used here (Wen et al. 2012, 2014; Okada et al. 2012). These results suggest that caffeine may reduce obesity through its effect on gene expression and ASE in *GIPR*.

Among the genes with cASE, 79 are associated with complex traits in the GWAS catalog (Supplemental Table S18). Figure 6D shows four examples of cASE genes associated with complex traits. These include cASE in *SAMM50* in response to copper, in *ERAP1* in response to selenium treatment, in *GOSR2* following treatment with mono-n-butyl phthalate, and in *LAMP3* in response to selenium. This last example may explain the influence of selenium on Parkinson's disease (PD). Previous studies found reduced selenium levels in PD patients (Shahar et al. 2010). In addition, selenium reduces bradykinesia, a well-described symptom of PD, in rats (Ellwanger et al. 2015), suggesting that higher selenium levels would be beneficial for PD patients. A GWAS hit for PD (Do et al. 2011; International Parkinson Disease Genomics Consortium 2011) was an eQTL for *LAMP3*, where the reference/PD-risk allele led to increased expression of *LAMP3* (The GTEx Consortium 2015). In our data, cASE at rs16833703 in *LAMP3* preferentially expressed the alternative allele at this SNP, located on the same haplotype as the nonrisk allele at the GWAS SNP. These data suggest that selenium is beneficial for PD patients through its influence on allelic expression in *LAMP3*. Overall, these cASE examples illustrate genes associated with complex traits, with a plausible biological association with the treatment (for details on the other three genes, see the Supplemental Material).

Discussion

We presented a scalable high-throughput approach to characterize the effect of environmental and genetic perturbation on gene expression levels. In this study, we tightly controlled environmental exposure using in vitro treatments in different cell types and analyzed the transcriptional response for hundreds of conditions that were previously uncharacterized. These results will be highly valuable to many researchers interested in changes to specific genes or pathways following various treatments and cell types. Among the DE genes, we found that 22% have been associated with complex traits in GWAS. These results strongly suggest a key environmental component for many complex traits and should assist the design of future studies. For example, this resource can help in selecting relevant environmental variables that should be considered in animal models for human complex traits, in patient studies, and in reanalyzing GWAS data when the relevant exposure variable was collected as part of the study.

One of the main advantages of our approach is that it can be used to detect GxE interaction in a single individual for many treatments and cell types using ASE analysis. Compared to model organisms and transgenic models, studying GxE interaction in humans poses significant challenges. In clinical trials, a limited number of exposures can be tested, while in large-scale epidemiological studies many exposures convolve together. In this study, we have analyzed three individuals per cell type in order to explore a large number of environmental conditions that would not be practical for a reQTL study design. While some GxE interactions may be detected in a conventional baseline eQTL study, this would only occur if all or a subset of the samples was exposed to a relevant environment. However, in eQTL studies, the specific exposure would likely remain uncharacterized as a latent variable that may be unknown or difficult to model. Though we do not require many individuals, our approach is limited by the requirement of having two heterozygous variants: at the causal regulatory variant and at the variant for which ASE is measured. A small fraction of the ASE we detected may actually correspond to low-frequency variants that are sampled in three individuals, but the majority will correspond to common variants. The requirement that at least one of the three individuals is a double heterozygote means that we are missing instances of GxE interaction, especially those at low allele frequencies. Nonetheless, the 215 instances of GxE described here represent a lower bound to the amount of GxE signal that can be identified by applying our approach to additional treatment panels, cell types, and/or larger sample sizes.

Our catalog of GxE interactions, and future ones expanding on the one generated here, will be a necessary resource to thoroughly annotate genes and create a bridge between epidemiological and genome-wide association studies. Here we showed that 49% of genes with GxE interactions are GWAS genes. Although limited by our false-negative rate, we compiled the most comprehensive list to date of GxE and relevant environmental conditions that can aide in the interpretation of specific GWAS findings. We provided some examples of candidate GxE mechanisms for complex traits and released our results as a browsable web-resource. Mining of our results by other researchers has the potential to inform new GWAS findings and identify latent variables in

GWAS that are important risk/protective factors for human complex traits and diseases.

In future research, we anticipate that the approach we developed will potentially aid in precision medicine to tailor medication exposures using patient cells for improved patient outcomes. Indeed, when considering potential translation of these findings to clinical practice, in vitro measurements of ASE for a large panel of cell types, extracted and derived from single-patient stem cells, are a promising solution to studying rare disease variants and individualized outcomes of combinatorial interactions of common genetic variants.

Methods

Cell culture and treatments

Experiments were conducted using the following cell types: lymphoblastoid cell lines (LCLs), PBMCs, human umbilical vein endothelial cells (HUVECs), human SMCs, and melanocytes. LCLs (GM18507, GM18508, and GM19239) were purchased from Coriell Cell Repository, cultured, and treated as described previously (Moyerbrailean et al. 2015). PBMCs were derived from whole human blood purchased from Research Blood Components. Blood specimens were obtained from three individual donors. Primary HUVECs and SMCs were isolated from human umbilical cord tissue collected shortly following birth. Additionally, cryopreserved HUVECs (CC-2517-0000315288) and SMCs (CC-2579-7F3794) were purchased from Lonza. For additional details on HUVEC and SMC preparation, see Supplemental Methods 1. Primary melanocytes (NHEM) isolated from neonatal foreskin were purchased from Lonza (CC-2504 lot no. 252410 and 5F0885J) and from Promocell (C-12400 lot no. 3052103.1). Details on cell culturing are provided in the Supplemental Methods 2. Supplemental Table S1 shows the concentrations used for each treatment. For each treatment panel and cell type, cells derived from three individuals were treated at the same time on a 96-well plate. A schematic of the study design is provided in Supplemental Figure S1.

RNA-seq library preparation and sequencing

We used a two-step approach to gene expression analysis that we recently developed (Moyerbrailean et al. 2015). A 96-library pooling and shallow sequencing strategy (<10 M reads per library) (Supplemental Table S2) were used to minimize the amount of resources used in the first step. For the second step, we repooled a selection of the initial libraries (Fig. 1B; Supplemental Methods 8.1) to achieve a more uniform allocation of sequencing reads across samples (130 M reads/sample on average) (Supplemental Table S4). Pools of 96 samples from step 1 were sequenced on two lanes of an Illumina HiSeq2500 in fast mode to obtain 50-bp paired-end reads at the University of Chicago and at the Michigan State University Genomics Cores or were sequenced on one lane of the Illumina NextSeq500 for 75 cycles of paired-end in HO mode in the Luca/Pique-Regi laboratory. Step 2 resequencing was performed on the NextSeq500 in the Luca/Pique-Regi laboratory. The number of reads collected for each sample in step 1 and step 2 is reported in Supplemental Tables S2 and S4, respectively.

Sequence alignment and post-processing

RNA-seq data for step 1 was processed as described previously (Moyerbrailean 2015). For step 2, reads were aligned to the hg19 human reference genome using STAR (https://github.com/alexdobin/STAR/releases, version STAR_2.4.0h1) (Dobin et al. 2013) and the Ensembl reference transcriptome (version 75). Details are provided in Supplemental Methods 7.1. We did not realign the reads to GRCh38 because hg19 is the version of the reference human genome used in the release of the 1000 Genomes Project that we used for pileup. The 1000 Genomes Project data were not available in GRCh38 coordinates until October 26, 2016. Realigning the reads should not affect the conclusions as any problematic region of the genome is excluded from any analysis as detailed in the Supplemental Material. To correct for potential alignment biases, we used the WASP suite of tools for allele-specific read mapping (https://github.com/bmvdgeijn/WASP, downloaded 09/15/15) (Van de Geijn et al. 2015). Note that we do not use the WASP combined haplotype test (CHT) as we tested each SNP separately using QuASAR (Harvey et al. 2015). Retained read counts per sample after filtering can be found in Supplemental Table S4.

Differential gene expression

To identify DE genes, we used the method implemented in the software DESeq2 (R version 3.2.1, DESeq2 version 1.8.1) (Love et al. 2014). DE genes were determined as genes with at least one transcript having a Benjamini-Hochberg controlled FDR (BH-FDR) (Benjamini and Hochberg 1995) of 10% and an absolute $\log_2$ (fold-change) >0.25. The same procedure was used for step 1 and step 2. A summary of differential expression for both steps can be found in Supplemental Tables S3 and S5, and a full set of differential expression results from step 2 can be found in Supplemental Table S6.

Network analysis with WGCNA

For network analysis, we used gene expression data normalized as described in Supplemental Methods 8.4. We combined all the data across cell types, treatments, and individuals, resulting in a matrix with 14,527 rows (genes) and 297 columns (samples). We then used WGCNA (Langfelder and Horvath 2008), version 1.47, implemented in R to build an unsigned network. A soft thresholding power of six was chosen, and the network was built using the automated block-wise modules pipeline using Pearson correlations, a signed topological overlap matrix, and a minimum module size of 10. Modules were cut from the network dendrogram with the dynamic hybrid tree cut method, and the module eigengene was calculated as the first principal component of each module's expression matrix. A measure of module membership was calculated for each gene by correlating the gene's expression profile with its module's eigengene. More details on the network module analysis are in Supplemental Methods 8.6.

Joint genotyping and ASE inference

To create a core set of SNPs for ASE analysis, we started with all the SNPs from the phase 3 release of the 1000 Genomes Project Consortium (www.1000Genomes.org, v5b.20130502, downloaded on 08/24/15) (The 1000 Genomes Project Consortium 2015) and first removed SNPs with low minor allele frequency (MAF <5%). We also removed SNPs within the regions of annotated copy number variation and ENCODE blacklisted regions (Degner et al. 2012), leaving a total of 7,340,521 SNPs in the core set. Counts of reads covering each allele at selected SNPs (Supplemental Methods 9.1) were obtained by "piling up" aligned reads for each sample over SNPs using samtools mpileup (Li et al. 2009) and the hg19 human reference genome. All sample pileups for a given individual across all treatment conditions and the two treatment plates were processed together (not merged) using

the QuASAR package (Harvey et al. 2015) for joint genotyping. ASE inference was performed for each sample separately. Heterozygous SNPs with a read coverage greater than 40 were tested for ASE using QuASAR (Harvey et al. 2015). A summary of the amount of ASE detected in each sample is in Supplemental Figure S10 and Supplemental Table S9. A full list of SNPs tested can be found in Supplemental Table S10.

Identification of induced ASE

To identify genes with iASE, we selected SNPs that were well covered in the treatment (i.e., more than 40 reads) and had ASE (10% FDR) but had little to no expression in the matched control. We used a coverage threshold in the control of $10 \times (D_C/D_T)$, where D_C and D_T represent the sequencing depth of the control and treatment libraries, respectively, in TPM (see Supplemental Methods 9.3). This equates to a ratio of 40 reads to 10 (expression in the control is fourfold lower than the minimum required for a gene to be considered expressed in the ASE analysis) while accounting for sequencing depth differences. Finally, we required the SNP-based $\log_2$ (fold change) (Supplemental Methods 9.3) to be $>\log_2(5)$.

Meta-analysis of subgroup heterogeneity

We used MeSH to model potentially heterogeneous cASE effects across multiple subgroups contained within the data. The input to MeSH is a pair of ASE observations derived from QuASAR summarized by the parameter β measuring the allelic imbalance and a standard error of the parameter. To look specifically at conditional ASE, a BF for cASE is calculated as $BF_{treatment} - BF_{shared}$ (treatment-only cASE) and $BF_{control} - BF_{shared}$ (control-only cASE). All the cASE BFs are then used to rank and select the observations with strongest evidence for cASE.

ΔAST: a novel method to measure cASE

Differential Z-scores (Z_Δ) were calculated from QuASAR β parameters using the following formula. For each SNP,

$$Z_\Delta = \frac{\beta_T - \beta_C}{\sqrt{se_T^2 + se_C^2}}, \quad (1)$$

where β_T and se_T represent the estimates for the ASE parameter and its standard error for the treatment condition, and β_C and se_C represent the corresponding estimates for the control condition. The Z_Δ scores were then normalized by the standard deviation across Z_Δ scores corresponding to control versus control (CO1 vs. CO2). Finally, P-values (P_Δ) are calculated from the Z_Δ scores as $P_\Delta = 2 \times \text{pnorm}(-|z|)$. Under the null, Z_Δ are asymptotically normally distributed, and Figure 3C shows that when contrasting the two sets of controls the P_Δ-values are almost uniformly distributed as expected. To further correct for this small deviation, we used the control versus control P-values to empirically estimate the FDR (see Supplemental Methods 10.3). The list of significant cASE SNPs is in Supplemental Table S11.

Analysis of EDGE

Within each treatment and cell line subgroup, we examined the Pearson's correlation of the treatment standardized effect size (ASE $Z_T = \beta_T/se_T$) to the matched control one (ASE $Z_C = \beta_C/se_C$) across all SNPs tested (see Supplemental Fig. S19). This correlation measures the consistency of the genetic effect between the treatment and control, and therefore, a lower correlation indicates a higher perturbation or displacement of the genetic effects. We define this as the EDGE index, which is formally defined as

$$\text{EDGE}_{s,t} = \frac{\text{Pearson}(Z_{s,CO1}, Z_{s,CO2})}{\text{Pearson}(Z_{s,t}, Z_{s,c})}, \quad (2)$$

where $\text{Pearson}(Z_{s,t}, Z_{s,c})$ is the sample Pearson correlation coefficient between treatment t and control c ASE Z-scores across all observed SNPs in cell type s. Equivalently, $\text{Pearson}(Z_{s,CO1}, Z_{s,CO2})$ is the Pearson correlation coefficient between the two control sets ASE Z-scores across all observed SNPs in cell type s. The EDGE index values for each cell type and condition can be found in Supplemental Figure S20 and Supplemental Table S14.

Analysis of heritability enrichment using GEMMA

To run GEMMA (Zhou and Stephens 2012; Zhou 2016), we partitioned SNPs genome wide to create a category file. Each SNP was assigned to one of the following categories: cASE (genic regions with conditional ASE) or iASE (genic regions with induced ASE), ASE (genic regions with ASE), other genic (genic regions that do not fall into any of the categories above), and intergenic (>100 kb from any gene). We then used GEMMA to compute the SNP correlations among different categories from a reference panel (502 individuals of European ancestry from the 1000 Genomes Project). This was followed by summing the Z^2 statistics from the GWAS meta-analysis within the categories. Finally, we computed the proportion of variance and the fold enrichment of heritability explained by each category. A table of the results can be found in Supplemental Table S17.

Data access

The RNA sequencing data from this study have been submitted to the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap) under the accession number phs001176.v1.p1. Additional browsing tools for exploring the data in this project are available at http://genome.grid.wayne.edu/gxebrowser.

Acknowledgments

We thank K. Zhang for providing tunicamycin, pantothenic acid, and PM2.5 reagents; the University of Chicago and the Michigan State Functional Genomics Core for providing sequencing service for part of this study; and the Wayne State University High Performance Computing Grid for all computational support. We thank Jonathan Pritchard, Anna Di Rienzo, Stephen Krawetz, Joseph Maranville, Russ Finley, Luis Barreiro, and three anonymous reviewers for helpful comments and suggestions that helped to improve the quality of our manuscript. We also thank all the anonymous individuals that contributed biological samples to this project. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (5R01GM109215 to F.L. and R.P.) and the American Heart Association (14SDG20450118 to F.L.).

References

The 1000 Genomes Project Consortium 2015. A global reference for human genetic variation. *Nature* **526:** 68–74.

Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. 2012. Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proc Natl Acad Sci* **109:** 1204–1209.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* **57:** 289–300.

Berndt SI, Gustafsson S, Mägi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR, et al. 2013. Genome-wide

meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45:** 501–512.
Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng H-F, Richards JB, Glass D, Small KS, Durbin R, et al. 2014. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet* **47:** 88–91.
Çalişkan M, Baker SW, Gilad Y, Ober C. 2015. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genet* **11:** e1005111.
Cowper-Sal·lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M. 2012. Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* **44:** 1191–1198.
Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482:** 390–394.
Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, et al. 2011. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet* **7:** e1002141.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.
Ellwanger JH, Molz P, Dallemole DR, dos Santos AP, Müller TE, Cappelletti L, da Silva MG, Franke SIR, Prá D, Henriques JAP, et al. 2015. Selenium reduces bradykinesia and DNA damage in a rat model of Parkinson's disease. *Nutrition* **31:** 359–365.
Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC, et al. 2012. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat Genet* **44:** 502–510.
Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, et al. 2014. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343:** 1246949.
Flutre T, Wen X, Pritchard J, Stephens M. 2013. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* **9:** e1003486.
Fox CS, Liu Y, White CC, Feitosa M, Smith AV, Heard-Costa N, Lohman K, Johnson AD, Foster MC, Greenawalt DM, et al. 2012. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* **8:** e1002695.
Franco LM, Bucasas KL, Wells JM, Niño D, Wang X, Zapata GE, Arden N, Renwick A, Yu P, Quarles JM, et al. 2013. Integrative genomic analysis of the human immune response to influenza vaccination. *eLife* **2:** e00299.
Graham RR, Kozyrev SV, Baechler EC, Reddy MVPL, Plenge RM, Bauer JW, Ortmann WA, Koeuth T, González Escribano MF, Pons-Estel B, et al. 2006. A common haplotype of interferon regulatory factor 5 (*IRF5*) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat Genet* **38:** 550–555.
Graham RR, Kyogoku C, Sigurdsson S, Vlasova IA, Davies LRL, Baechler EC, Plenge RM, Koeuth T, Ortmann WA, Hom G, et al. 2007. Three functional variants of IFN regulatory factor 5 (*IRF5*) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci* **104:** 6758–6763.
Grundberg E, Adoue V, Kwan T, Ge B, Duan QL, Lam KC, Koka V, Kindmark A, Weiss ST, Tantisira K, et al. 2011. Global analysis of the impact of environmental perturbation on *cis*-regulation of gene expression. *PLoS Genet* **7:** e1001279.
The GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660.
Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, et al. 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* **95:** 535–552.
Harvey CT, Moyerbrailean GA, Davis GO, Wen X, Luca F, Pique-Regi R. 2015. QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **31:** 1235–1242.
Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake TA. 2014. Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics* **15:** 471.
Idaghdour Y, Quinlan J, Goulet J-P, Berghout J, Gbeha E, Bruat V, de Malliard T, Grenier J-C, Gomez S, Gros P, et al. 2012. Evidence for additive and interaction effects of host genotype and infection in malaria. *Proc Natl Acad Sci* **109:** 16786–16793.
International Parkinson Disease Genomics Consortium. 2011. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377:** 641–649.
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.
Kim HJ, Yoon BK, Park H, Seok JW, Choi H, Yu JH, Choi Y, Song SJ, Kim A, Kim J-W, et al. 2016. Caffeine inhibits adipogenesis through modulation of mitotic clonal expansion and the AKT/GSK3 pathway in 3T3-l1 adipocytes. *BMB Rep* **49:** 111–115.
Knowles DA, Davis JR, Raj A, Zhu X, Potash JB, Weissman MM, Shi J, Levinson D, Mostafavi S, Montgomery SB, et al. 2015. Allele-specific expression reveals interactions between genetic variation and environment. *bioRxiv* doi: 10.1101/025874.
Kukurba KR, Zhang R, Li X, Smith KS, Knowles DA, How Tan M, Piskol R, Lek M, Snyder M, MacArthur DG, et al. 2014. Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet* **10:** 1–9.
Kumasaka N, Knights AJ, Gaffney DJ. 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48:** 206–213.
Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9:** 559.
Lee MN, Ye C, Villani A-C, Raj T, Li W, Eisenhaure TM, Imboywa SH, Chipendo PI, Ran FA, Slowikowski K, et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343:** 1246980.
Leung LH. 2004. Systemic lupus erythematosus: a combined deficiency disease. *Med Hypotheses* **62:** 922–924.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Li D-K, Ferber JR, Odouli R. 2015. Maternal caffeine intake during pregnancy and risk of obesity in offspring: a prospective cohort study. *Int J Obes* **39:** 658–664.
Lim S-Y, Ghosh SK. 2005. Autoreactive responses to environmental factors: 3. Mouse strain-specific differences in induction and regulation of anti-DNA antibody responses due to phthalate-isomers. *J Autoimmun* **25:** 33–45.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550.
Luca F, Maranville JC, Richards AL, Witonsky DB, Stephens M, Di Rienzo A. 2013. Genetic, functional and molecular features of glucocorticoid receptor binding. *PLoS One* **8:** e61654.
Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Horikoshi M, Johnson AD, Ng MCY, Prokopenko I, et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46:** 234–244.
Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI, Mecham BH, Howie B, Shim H, Naidoo D, et al. 2013. A statin-dependent QTL for *GATM* expression is associated with statin-induced myopathy. *Nature* **502:** 377–380.
Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, Baxter S, Stephens M, Rienzo A. 2011. Interactions between glucocorticoid treatment and *cis*-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet* **7:** e1002162.
Maranville JC, Baxter SS, Witonsky DB, Chase MA, Di Rienzo A. 2013. Genetic mapping with multiple levels of phenotypic information reveals determinants of lymphocyte glucocorticoid sensitivity. *Am J Hum Genet* **93:** 735–743.
McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328:** 235–239.
McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK, et al. 2013. Identification of genetic variants that affect histone modifications in human cells. *Science* **342:** 747–749.
Moyerbrailean GA, Davis GO, Harvey CT, Watza D, Wen X, Pique-Regi R, Luca F. 2015. A high-throughput RNA-seq approach to profile transcriptional responses. *Sci Rep* **5:** 14976.
Ohara T, Muroyama K, Yamamoto Y, Murosaki S. 2016. Oral intake of a combination of glucosyl hesperidin and caffeine elicits an anti-obesity effect in healthy, moderately obese subjects: a randomized double-blind placebo-controlled trial. *Nutr J* **15:** 6.
Okada Y, Kubo M, Ohmiya H, Takahashi A, Kumasaka N, Hosono N, Maeda S, Wen W, Dorajoo R, Go MJ, et al. 2012. Common variants at *CDKAL1* and *KLF9* are associated with body mass index in east Asian populations. *Nat Genet* **44:** 302–306.
Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11:** 533–538.
Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. 2012. Effects of sequence

variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22:** 860–869.
Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU, et al. 2010. Genetic variation in *GIPR* influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42:** 142–148.
Shahar A, Patel KV, Semba RD, Bandinelli S, Shahar DR, Ferrucci L, Guralnik JM. 2010. Plasma selenium is positively related to performance in neurological tasks assessing coordination and motor speed. *Mov Disord* **25:** 1909–1915.
Sigurdsson S, Nordmark G, Göring HHH, Lindroos K, Wiman A-C, Sturfelt G, Jönsen A, Rantapää-Dahlqvist S, Möller B, Kere J, et al. 2005. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am J Hum Genet* **76:** 528–537.
Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21:** 1728–1737.
Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R, et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42:** 937–948.
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.
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**(Database issue): D1001–D1006.
Wen X, Stephens M. 2014. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *Ann Appl Stat* **8:** 176–203.
Wen W, Cho Y-S, Zheng W, Dorajoo R, Kato N, Qi L, Chen C-H, Delahanty RJ, Okada Y, Tabara Y, et al. 2012. Meta-analysis identifies common variants associated with body mass index in east Asians. *Nat Genet* **44:** 307–311.
Wen W, Zheng W, Okada Y, Takeuchi F, Tabara Y, Hwang J-Y, Dorajoo R, Li H, Tsai F-J, Yang X, et al. 2014. Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum Mol Genet* **23:** 5492–5504.
Zhou X. 2016. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *bioRxiv* doi: 10.1101/042846.
Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44:** 821–824.

Received May 12, 2016; accepted in revised form October 13, 2016.

High-throughput allele-specific expression across 250 environmental conditions

Gregory A. Moyerbrailean, Allison L. Richards, Daniel Kurtz, et al.

Genome Res. 2016 26: 1627-1638 originally published online October 19, 2016
Access the most recent version at doi:10.1101/gr.209759.116