Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk

Castel, Stephane E.

2018-09

Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk

Stephane E. Castel [1,2*], Alejandra Cervera [1,3], Pejman Mohammadi [1,2,4,5], François Aguet [6], Ferran Reverter[7], Aaron Wolman[1], Roderic Guigo [7,8], Ivan Iossifov[1,9], Ana Vasileva[1,2] and Tuuli Lappalainen[1,2*]

Coding variants represent many of the strongest associations between genotype and phenotype; however, they exhibit inter-individual differences in effect, termed 'variable penetrance'. Here, we study how cis-regulatory variation modifies the penetrance of coding variants. Using functional genomic and genetic data from the Genotype-Tissue Expression Project (GTEx), we observed that in the general population, purifying selection has depleted haplotype combinations predicted to increase pathogenic coding variant penetrance. Conversely, in cancer and autism patients, we observed an enrichment of penetrance increasing haplotype configurations for pathogenic variants in disease-implicated genes, providing evidence that regulatory haplotype configuration of coding variants affects disease risk. Finally, we experimentally validated this model by editing a Mendelian single-nucleotide polymorphism (SNP) using CRISPR/Cas9 on distinct expression haplotypes with the transcriptome as a phenotypic readout. Our results demonstrate that joint regulatory and coding variant effects are an important part of the genetic architecture of human traits and contribute to modified penetrance of disease-causing variants.

Variable penetrance and variable expressivity are common phenomena that often cause individuals carrying the same variant to display highly variable symptoms, even in the case of Mendelian and other severe diseases driven by rare variants with strong effects on phenotype[1]. For our purposes, we use the term 'variable penetrance' as a joint description of both variable expressivity (severity of phenotype) and penetrance (proportion of carriers with phenotype). These phenomena are a key challenge for understanding how genetic variants manifest in human traits and a major practical caveat for the prognosis of an individual's disease outcomes based on their genetic data. However, the causes and mechanisms of variable penetrance are poorly understood. In addition to environmental modifiers of genetic effects, a potential cause of variable penetrance involves other genetic variants with additive or epistatic modifier effects[2]. While some studies have successfully mapped genetic modifiers of, for example, BRCA variants in breast cancer[3] and RETT variants in Hirschsprung's disease[4], genome-wide analysis of pairwise interactions between variants has proven to be challenging in humans. In part, this is because exhaustive pairwise testing of genome-wide interactions typically lacks power and is easily affected by confounders[5], and a targeted analysis of a specific variant or gene that is strongly implicated in a rare disease typically suffers from a low number of carriers. However, emerging large data sets with functional genomic and genetic data from disease cohorts now enable the genome-wide study of mechanistically justified hypotheses of how combinations of genetic variants may have joint effects on disease risk.

In this study, we analyzed how regulatory variants in cis may modify the penetrance of coding variants in their target genes via the joint effects of these variants on the final dosage of functional gene product, depending on their haplotype combination (Fig. 1 and Supplementary Fig. 1). This phenomenon has been demonstrated to affect penetrance of disease-predisposing variants in individual loci[6–9] and explored in early functional genomic data sets[10,11], and expression modifiers are known in model organisms[12]. However, genome-wide evidence of regulatory modifiers of disease risk driven by coding variants has been lacking, alongside a generally applicable theoretical framework and analytical methods to study this phenomenon. This means that while potentially important, this phenomenon is often not addressed in genome-wide association studies of common disease. In this work, we use population-scale functional genomics and disease cohort data sets to show that genetic regulatory modifiers of pathogenic coding variants affect disease risk. Furthermore, we use genome editing with CRISPR/Cas9 to demonstrate an experimental approach to studying the role of regulatory variants as modifiers of coding variant penetrance. We focus on rare pathogenic coding variants from exome and genome sequencing data, which provide the best-characterized group of variants with strong phenotypic effects, and common regulatory variants affecting gene expression or splicing. Thus, our analysis integrates these traditionally separate fields of human genetics by considering joint effects that different types of mutations have on gene function.

Results

Purifying selection acts on haplotype combinations. First, we tested the hypothesis that purifying selection should deplete haplotype combinations that increase the penetrance of pathogenic

coding variants from the general population. To accomplish this, we analyzed data from the GTEx project, which is representative of the general population in that it lacks individuals with severe genetic disease[13]. This consists of genotype and RNA-sequencing (RNA-seq) data of 7,051 samples across 44 tissues from the 449 individuals with exome sequencing and SNP array data of the GTEx v6p release[14]. Throughout this study, we defined the predicted pathogenicity of variants using their Combined Annotation Dependent Deletion (CADD) score, which incorporates a wide breadth of annotations, including conservation and protein structure[15]. We used the authors' suggested cutoff of 15 for defining potentially pathogenic variants; this is the median CADD score across all possible canonical splice site and missense variants in the human genome (see Methods).

We first measured the regulatory haplotype of coding variants using allelic expression data, which captures cis effects of both expression and splice regulatory variation at the individual level. We employed multiple approaches to account for issues of mapping bias, which often affect allelic expression studies (see Methods)[16]. In the modified penetrance model, purifying selection should result in a depletion of pathogenic variants on higher-expressed or exon-including haplotypes. For each of the 44 GTEx tissues, we calculated the expression of coding variant minor alleles using allelic fold change (aFC)[17] and compared the expression of missense variants with allele frequency (AF)-matched synonymous controls. Supporting our hypothesis, the minor alleles of missense variants showed reduced allelic expression, which was proportional to their predicted pathogenicity (Fig. 2a). Across tissues, rare (AF < 1%) pathogenic (CADD > 15) missense variants showed a significant ($P = 4.57 \times 10^{-9}$) 0.70% reduction of allelic expression compared with synonymous controls, but rare benign (CADD < 15) missense variants did not ($P = 0.388$) (Fig. 2b and Supplementary Fig. 2a,b). This suggests that in the general population, pathogenic variants are depleted from higher-expressed or exon-including regulatory haplotypes. We also performed this analysis using polyPhen alone to define coding variant pathogenicity to ensure that our results were not biased by the additional features used by CADD, and found that they were consistent (Supplementary Fig. 2c).

To study whether this pattern is driven by regulatory variation affecting expression or splicing, both of which are manifest in allelic expression, we partitioned the coding variants into two groups. To accomplish this, we quantified exon inclusion in each GTEx sample using RNA-seq reads spanning exon junctions to produce a measure of percent spliced in (PSI) for each exon in each sample[18]. To isolate the effects of regulatory variation, we analyzed allelic expression only for variants that were found in an exon with 100% inclusion in that individual. As before, rare pathogenic missense variants had significantly reduced expression as compared with synonymous controls ($P = 5.94 \times 10^{-6}$; 1.56% reduction), but rare benign variants did not ($P = 0.521$), suggesting that pathogenic variants are less likely to accumulate on higher-expressed regulatory haplotypes (Fig. 2c). To isolate the effects of splice regulatory variation, we analyzed allelic expression of variants in exons where the sample had substantial deviation in exon inclusion from the population mean. To define these exons, for each exon, a population normalized PSI z-score was produced for each sample, allowing exon inclusion at the sample level to be compared with others (Supplementary Fig. 2e). When measuring allelic expression of variants found in the top 10% of sample exons by absolute PSI z-score, we again observed that rare pathogenic missense variants had significantly reduced expression as compared with synonymous controls ($P = 1.3 \times 10^{-3}$; 2.00% reduction), but rare benign variants did not ($P = 0.191$). This suggests that pathogenic variants are less likely to accumulate on haplotypes where the corresponding exon is more likely to be included in transcripts (Fig. 2d). In all analyses, pathogenic variants had significantly reduced expression versus allele frequency-matched synonymous controls as compared with benign variants (Supplementary
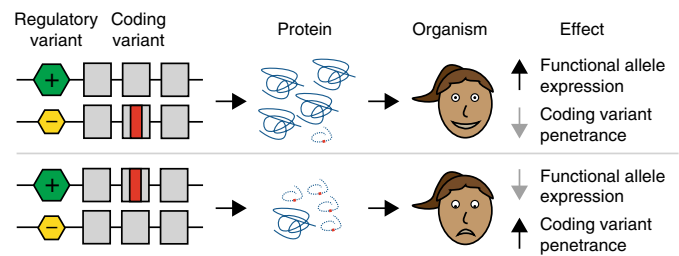

Fig. 1 | Regulatory variants as modifiers of coding variant penetrance. The hypothesis of this study is illustrated with an example in which an individual is heterozygous for both a regulatory variant and a pathogenic coding variant. The two possible haplotype configurations would result in either decreased penetrance of the coding variant, if it was on the lower-expressed haplotype, or increased penetrance of the coding variant, if it was on the higher-expressed haplotype. See Supplementary Fig. 1 for a quantitative description of the model.

Fig. 2d). Altogether, these analyses of allelic expression data suggest that in a cohort representative of the general population, pathogenic coding variants exist less frequently in high-penetrance regulatory haplotype combinations, as would be expected under the modified penetrance model.

While allelic expression paired with splice quantification provides a powerful functional readout of latent regulatory variants acting on a gene in each individual, the phenomenon of modified penetrance can also be studied from genetic data alone by analyzing phased haplotypes of coding variants and regulatory variants identified by expression quantitative trait locus (eQTL) mapping in cis. Our hypothesis is that in pathogenic coding variant heterozygotes, eQTL-mediated lower expression of the haplotype carrying the 'wild-type' (WT) major coding allele increases the penetrance of the rare allele, and vice versa (Fig. 3a and Supplementary Fig. 1). To study this, we developed a test for regulatory modifiers of penetrance that uses phased genetic data (see Methods). Briefly, for each rare coding variant heterozygote, we test whether the major coding allele is on the lower-expressed eQTL haplotype (Supplementary Fig. 3a) and determine whether this occurs more or less frequently than would be expected under the null hypothesis based on eQTL frequencies in the population studied (Supplementary Fig. 3b). Using simulated data, we found that our test was well calibrated under the null hypothesis while still being sensitive to changes in haplotype configuration (Supplementary Fig. 3c,d).

To analyze whether the distribution of coding variants on cis-eQTL haplotypes in GTEx showed signs of selection against increased penetrance, we produced a large set of haplotype phased genetic data from GTEx v7, where 30× whole genome sequencing of 620 individuals was available. This was obtained from population-based phasing paired with read-backed phasing using DNA sequencing (DNA-seq) reads[19] and RNA-seq reads[20] from up to 38 tissues for a single individual. This allowed us to analyze the haplotypes of 221,487 rare (minor allele frequency (MAF) < 1%) coding variants at thousands of genes with known common (MAF > 5%) eQTLs from GTEx v6p[14] (Supplementary Table 1). Using our test for regulatory modifiers of penetrance, we did not observe any significant evidence of reduced penetrance of rare potentially pathogenic variants when all protein coding genes were analyzed together ($P = 0.268$). However, we hypothesized that genes might be under differing selective pressure with respect to this phenomenon, so we stratified our analysis based on eQTL effect size, gene conservation and coding constraint. We observed a significant negative correlation between the predicted penetrance of rare potentially pathogenic variants and both eQTL effect size ($\rho = -0.229$, $P = 0.0224$) and gene conservation ($\rho = -0.217$, $P = 0.0304$), while no significant
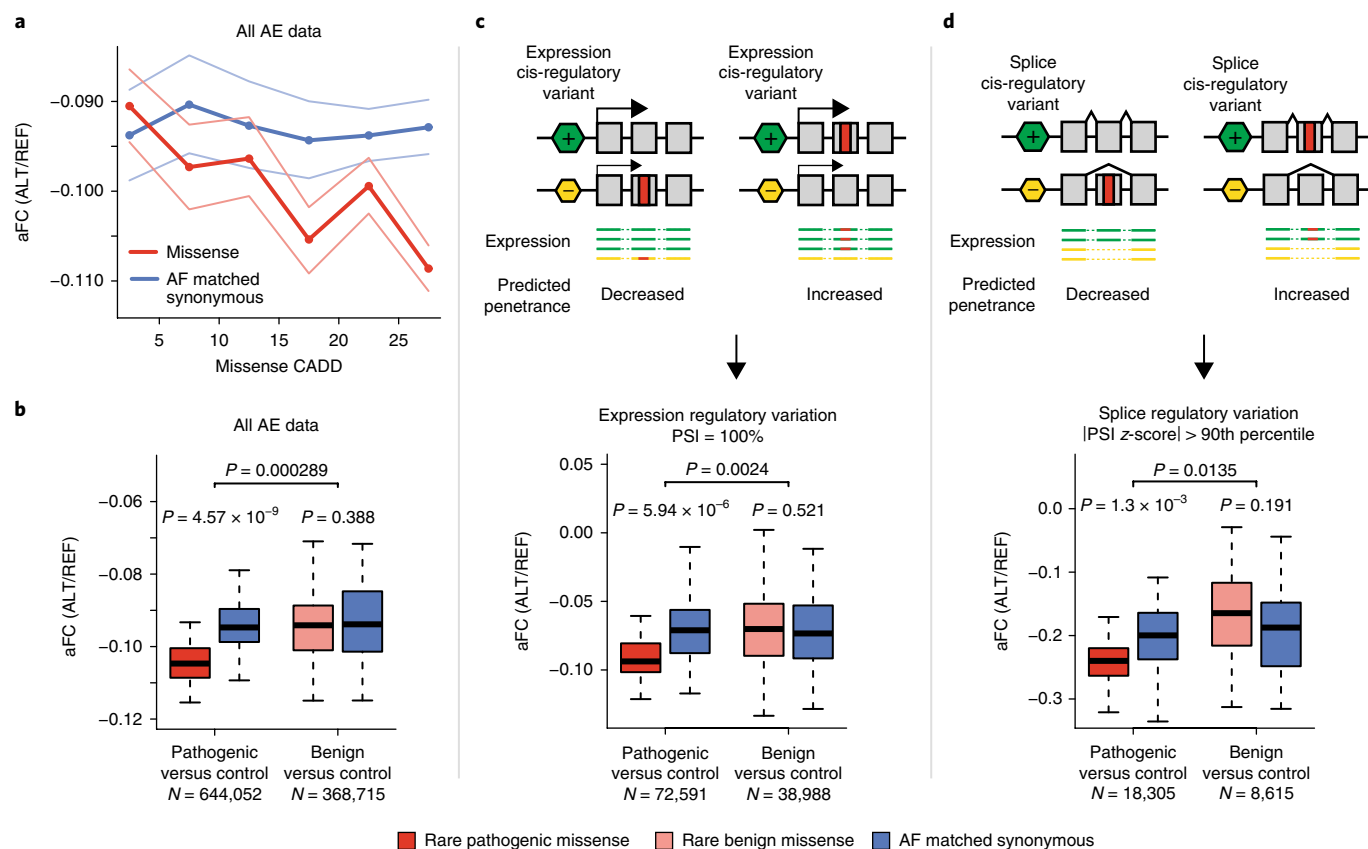
Fig. 2 | Analysis of regulatory effects at the individual level shows that pathogenic coding variants are depleted from higher-expressed and exon-including regulatory haplotypes in the general population. a, Allelic fold change (aFC) of rare (allele frequency (AF) <1%) missense and allele frequency-matched synonymous variants in bins of 5 CADD PHRED with 95% confidence interval across GTEx tissues. $N = 1,012,767$ independent allele-specific expression (ASE) measurements. **b**, Boxplot of mean aFC across each of the 44 GTEx tissues calculated for rare pathogenic missense (CADD >15), rare benign missense (CADD <15) and allele frequency-matched synonymous controls. **c**, Mean aFC across tissues calculated using only variants found in exons where the sample has 100% exon inclusion, as measured by percent spliced in (PSI), which removes allelic effects arising from splice regulatory variation. **d**, Mean aFC across tissues calculated using only variants found in exons where the sample has substantial variation in exon inclusion compared with the population, as defined by |PSI z-score| >90th percentile across all exons, which enriches for allelic effects caused by splice regulatory variation. The total number (N) of variant aFC measurements across all tissues for pathogenic and benign variants is indicated. P-values are generated by comparing mean aFC of missense variants with allele frequency-matched synonymous controls across tissues using a two-sided paired Wilcoxon signed rank test. For boxplots, bottom whisker: Q1 − 1.5*interquartile range (IQR), top whisker: Q3 + 1.5*IQR, box: IQR, center: median, and outliers are not plotted for ease of viewing. See Supplementary Fig. 2 for related analyses.

correlation was observed for benign variants (Supplementary Fig. 4). We quantified this effect and found that pathogenic variants in genes with strong eQTLs (top 25% by |effect size|) had a significant ($P = 6.08 \times 10^{-3}$) decrease of 1.32% in the frequency of haplotypes where the major coding allele was on the lower-expressed haplotype expressed than would be expected under the null hypothesis, while no effect was seen for benign variants ($P = 0.703$) (Fig. 3b). Similarly, we also observed a significant reduction of predicted penetrance of rare potentially pathogenic variants ($P = 0.0186$) but not benign variants ($P = 0.412$) in conserved genes (top 25% by median exon base conservation). Finally, we observed the strongest effect at genes that were defined using the Exome Aggregation Consortium (ExAC) as loss of function and missense intolerant (-2.20%, $P = 0.0221$)[21,22], while no effect was seen for benign variants ($P = 0.333$).

Altogether, combined with observations from functional data of allelic expression, these results suggest that joint effects between regulatory and coding variants have shaped human genetic variation in the general population through purifying selection depleting haplotype combinations whereby cis-regulatory variants increase the penetrance of pathogenic coding variants (Supplementary

Fig. 1). These patterns are significant and consistent, although the genome-wide magnitude of their effects is not strong. However, since our results indicate that regulatory modifiers of penetrance affect primarily pathogenic coding variants, stronger cis-regulatory variants, and both conserved and constrained genes, genome-wide analysis probably ends up diluting a signal that may be strong and phenotypically relevant for a subset of genes and variants.

Regulatory modifiers of penetrance affect disease risk. We next sought to investigate whether regulatory modifiers of penetrance affect disease risk in patients. This would manifest as patients having an overrepresentation of regulatory haplotype configurations that increase the penetrance of putatively disease-causing coding variants as compared with controls, in whom an enrichment of low-penetrance combinations is expected. Importantly, our test is calibrated to eQTL allele frequencies separately in case and control individuals, so that it measures only differences in haplotype configurations and not eQTL frequency between the populations. To test this hypothesis, we applied our genetic test for regulatory modifiers of penetrance to two large disease cohorts in cancer and autism. These diseases have a known contribution from rare coding
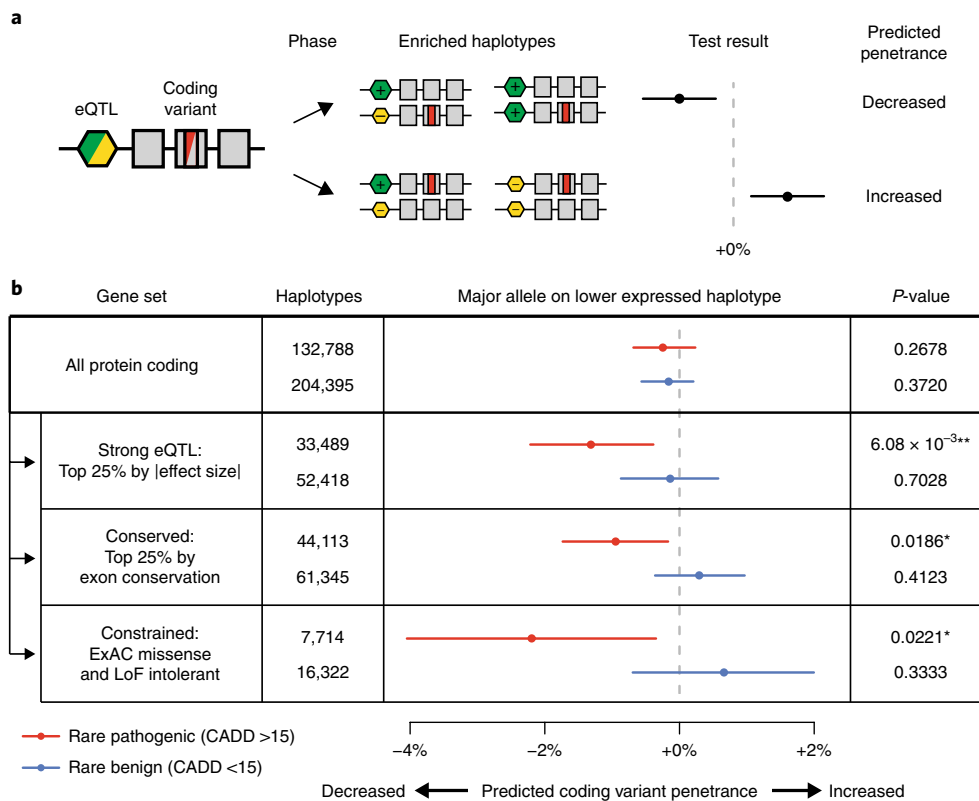
Fig. 3 | eQTL haplotype configurations that are predicted to increase pathogenic coding variant penetrance are depleted in the genomes of GTEx individuals. a, Phased genetic data can be used to produce haplotype configurations between regulatory variation identified using expression quantitative trait locus (eQTL) mapping and coding variant heterozygotes. Decreased expression of major coding alleles mediated by an eQTL could result in increased penetrance of the minor coding allele and vice versa (Supplementary Fig. 1). The observed frequency at which major coding variant alleles are on lower-expressed eQTL haplotypes is tested against a null distribution, which accounts for eQTL frequencies and assumes that coding variants occur on random haplotypes and in random individuals. The statistic indicates the percentage increase or decrease compared with the null hypothesis, where a positive value suggests increased penetrance and a negative value decreased penetrance. **b**, Test for regulatory modifiers of coding variant penetrance using 620 GTEx v7 population and read-backed phased whole genomes and GTEx v6p eQTLs, applied to rare (MAF <1%) pathogenic (CADD >15, including missense, splice and stop gained) and rare benign (CADD <15, including synonymous and missense) SNPs for different gene sets. Median estimates, 95% confidence intervals and two-sided empirical *P*-values were generated using 100,000 bootstraps. *$P < 0.05$, **$P < 0.01$. See Methods—Test for Regulatory Modifiers of Penetrance Using Phased Genetic Data and Supplementary Fig. 3 for description and benchmarking of the test and a description of the gene sets used.

variants in hundreds of disease-implicated genes, as well as large accessible genomic data sets that include genome-wide data of rare coding variants and common variants.

To study the role of regulatory modifiers of penetrance in germline cancer risk, we used population and read-backed phased germline variants (Supplementary Fig. 5b) from whole genome sequencing of 615 Cancer Genome Atlas (TCGA) individuals (Supplementary Table 2)[23]. Whole genome-sequenced, population and read-backed phased genotypes from 620 GTEx v7 individuals were used as controls (Supplementary Fig. 5a). We analyzed tumor suppressor genes (see Methods) that are known to harbor germline risk variants for cancer, often with a dosage-sensitive disease mechanism[24]. To study autism spectrum disorder (ASD), we used transmission phased exome and imputed SNP array genotype data (Supplementary Fig. 5c,d) from the Simons Simplex Collection (SSC) of 2,600 simplex families with one child with autism, their parents and any unaffected siblings[25–27]. When available, one unaffected sibling per family was used as a control. We analyzed a broad set of genes spanning multiple sources that have been previously implicated in ASD[27,28] (see Methods).

Our genetic test for regulatory modifiers of penetrance was applied to these data sets, first separately and then jointly, since

we were testing the same hypothesis in both the cancer and autism cohorts. We stratified our analysis by the sharing of coding variants between cases and controls, with coding variants observed only in cases likely to have the highest proportion of true disease-contributing variants, and with a decreasing proportion of variants contributing to disease among those observed in both cases and controls and those only in controls (Fig. 4). Using this approach, we found that in disease-associated genes, case-specific rare pathogenic variants were significantly enriched for haplotype configurations where the major allele was on the lower-expressed haplotype ($P = 9.53 \times 10^{-3}$), with control-specific variants showing no enrichment, as expected ($P = 0.597$). When analyzing shared variants, we found that in control individuals these were enriched for haplotype configurations where the major allele was on the higher-expressed haplotype ($P = 7.28 \times 10^{-3}$—suggesting a potentially decreased penetrance of some disease-contributing variants—but no consistent or significant effect was observed in cases for this group of variants ($P = 0.284$). No significant haplotype configuration enrichment in either cases or controls was found for rare benign variants at disease-associated genes (Supplementary Fig. 6a) or pathogenic variants at control genes matched for coding variant frequency (Supplementary Fig. 6b). All individual cohort results are
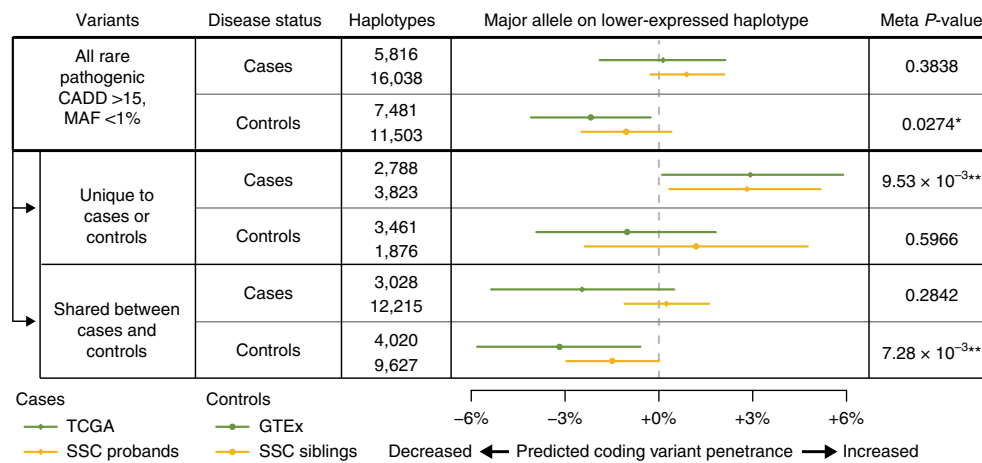
Fig. 4 | eQTL haplotype configurations that are predicted to increase pathogenic coding variant penetrance are enriched in individuals with cancer and autism spectrum disorder. Analysis of eQTL and coding variant haplotype configurations in cases and controls for autism spectrum disorder (ASD) and cancer, using the top GTEx v6p eQTL per gene by *P*-value across all tissues. For ASD analysis, haplotype configurations generated from transmission phased genetic data of 2,304 Simon's Simplex Collection (SSC) ASD-affected probands (cases) and 1,712 of their unaffected siblings (controls) were used, and haplotypes were analyzed at ASD-implicated genes. For cancer analysis, haplotype configurations generated from population and read-back phased germline whole genomes of 615 TCGA individuals (cases) and 620 whole genomes of v7 GTEx individuals (controls) were used, and haplotypes were analyzed at tumor suppressor genes. To enrich for putatively disease-causing variants, results were stratified based on whether variants were restricted to cases or controls or shared between both. Median estimates and 95% confidence intervals were generated using 100,000 bootstraps, and two-sided empirical *P*-values were generated from these confidence intervals and combined between cohorts using Fisher's method to produce meta p-values (*P < 0.05, **P < 0.01). See Methods for description of gene sets used, Supplementary Fig. 5 for description of eQTL coding variant haplotypes used for the analysis, Supplementary Fig. 6 for results from benign variants and control genes, and Supplementary Table 3 for the full table of results, including individual cohort-level *P*-values. MAF, minor allele frequency.

presented in Supplementary Table 3, with generally consistent patterns between the autism and cancer cohorts. Altogether, these results suggest that individuals with disease have an enrichment of harmful expression haplotype configurations that are predicted to increase coding variant penetrance, whereas unaffected individuals have an enrichment of protective configurations predicted to decrease coding variant penetrance. We expect that the true magnitude of the biological effect is diluted in our analysis due to false positives in the disease gene sets, only a subset of the potentially pathogenic variants studied being disease relevant, and modified penetrance affecting only a subset of genes. Nevertheless, the significant disease association of specific regulatory and coding variant configurations across two independent disease cohorts indicates a role for modified penetrance of coding variants by regulatory variation in both cancer and ASD.

Experimental demonstration of a regulatory modifier effect. Our population-scale analyses provide observational evidence that regulatory modifiers of penetrance play a role in the genetic architecture of human traits. We next sought to demonstrate an experimental approach to testing this hypothesis for a specific gene by using CRISPR/Cas9 to introduce a coding variant on distinct regulatory haplotypes, followed by quantification of its penetrance from a cellular readout. Such a framework will be useful for future studies that aim to validate single candidate genes from genome-wide analyses. Our finding that modified penetrance of germline variants by eQTLs may be involved in cancer risk led us to study a missense SNP (rs199643834, p.Lys508Arg) in the tumor suppressor gene *FLCN*, which codes for the protein folliculin and has a common eQTL in most GTEx v6p tissues[14]. This SNP causes the Mendelian autosomal dominant disease Birt–Hogg–Dubé syndrome[29], which results in characteristic benign skin tumors, lung cysts and cancerous kidney tumors and shows variable penetrance[30]. We edited the SNP in a fetal embryonic kidney cell line (293T), which is triploid and harbors a single copy of a common (1000 Genomes AF = 0.428)

loss of expression eQTL (rs1708629) located in the 5′ UTR of the gene[14,31]. This variant is among the most significant variants for the *FLCN* eQTL signal, overlaps promoter marks across multiple tissues and alters motifs of multiple transcription factors[32], therefore being a strong candidate for the causal regulatory variant of the *FLCN* eQTL (Supplementary Fig. 7a). We recovered monoclonal cell lines, genotyped them by targeted DNA-seq and performed targeted RNA-seq of the edited SNP (Fig. 5a, Supplementary Fig. 6b, and Supplementary Table 4). Allelic expression analysis showed that the haplotypes in the cell line are indeed expressed at different levels, probably driven by rs1708629 or another causal variant tagged by it, and the allelic expression patterns allowed phasing of the coding variant with the eQTL (Fig. 5b). In this way, we obtained four clones with a single copy of the Mendelian variant on the lower-expressed haplotype (snpLOW), three clones with a single copy on the higher-expressed haplotype (snpHIGH) and two monoallelic clones with three copies of the alternative allele of rs199643834 (Supplementary Fig. 7d). In addition, four clones that had been exposed to the CRISPR/Cas9 machinery but were WT at the *FLCN* locus were included as controls. As a phenotypic readout, we performed RNA-seq on all monoclonal lines.

Using the transcriptomes of these clones, we carried out differential expression analysis. Introduction of the Mendelian SNP had a genome-wide effect on gene expression, with 664 of 20,507 tested genes being significantly (false discovery rate (FDR) < 10%) differentially expressed in clones monoallelic for the SNP versus WT controls (Supplementary Fig. 7c and Supplementary Table 5). Gene set enrichment analysis[33] of differential expression test results showed significant (FDR < 10%) enrichment of pathways related to cell cycle control, DNA replication and metabolism, consistently with the annotation of *FLCN* as a tumor suppressor gene and the occurrence of tumors in patients with the mutation (Supplementary Table 6). To study the joint effect of the eQTL and Mendelian variant, we quantified the differential expression of these 664 genes separately in clones with low and high expression of the edited SNP (Fig. 5a).
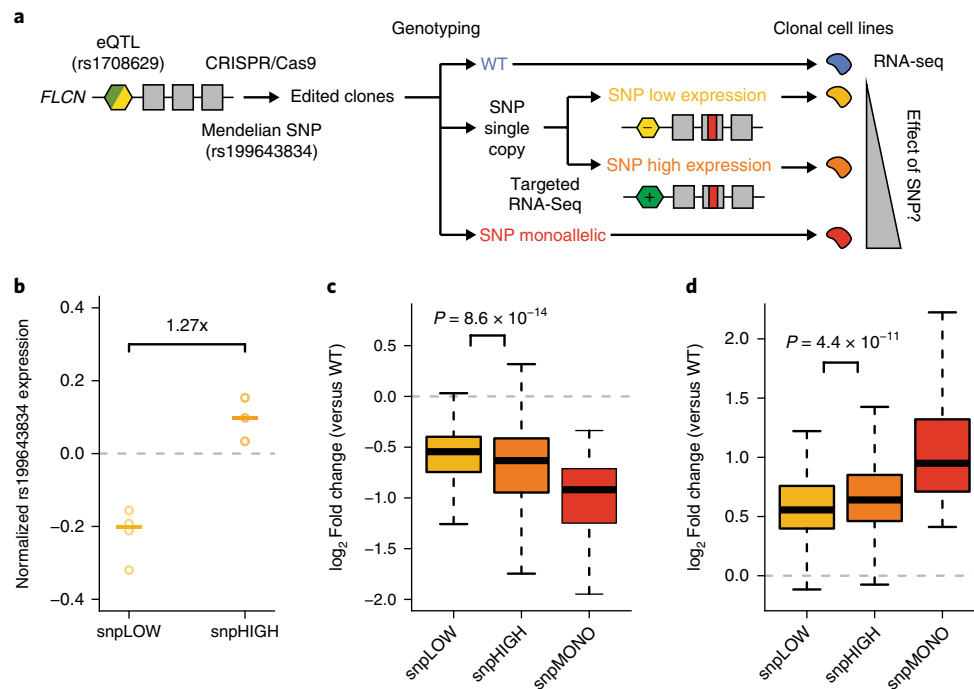
Fig. 5 | Haplotype-aware genome editing of a Mendelian disease SNP in *FLCN* demonstrates that expression regulatory variation can modify its penetrance. a, Illustration of the experimental study design. Briefly, a SNP that causes Birt–Hogg–Dubé syndrome was edited on distinct eQTL haplotypes in 293T cells using CRISPR/Cas9. Monoclonal cell lines were genotyped and classified as monoallelic for the edited SNP (snpMONO) or as having a single copy. Using targeted RNA-seq, single-copy clones were classified according to whether the edited SNP was on the lower- (snpLOW) or higher- (snpHIGH) expressed haplotype. The global transcriptome was used as a cellular phenotype to assess SNP penetrance. **b**, Copy number normalized expression of the edited SNP as measured by targeted RNA-seq (AE, log2(ALT/REF)) in snpLOW (AE < 0, $P < 0.01$, binomial test versus 0.5) and snpHIGH (AE > 0, $P < 0.01$) clones. Center lines represent median. **c,d**, Change in expression of genes that were significantly downregulated (**c**, 277 genes) or upregulated (**d**, 387 genes) in clones monoallelic for the edited SNP versus wild-type controls. Single-copy edited SNP clones are stratified by haplotype configuration. P-values were calculated using a two-sided paired Wilcoxon signed rank test. See Supplementary Fig. 7 and Supplementary Tables 4–6 for experimental details and additional analyses. For all plots, $N = 4$ snpLOW, 3 snpHIGH, 2 snpMONO and 4 WT biologically independent samples. For boxplots, bottom whisker: $Q1 - 1.5 * IQR$, top whisker: $Q3 + 1.5 * IQR$, box: IQR, center: median, and outliers are not plotted for ease of viewing.

As we predicted, clones with higher expression of the SNP showed significantly stronger differential expression of both downregulated (median = 8.10% increase; 95% confidence interval (CI) = 5.93% to 10.36%; $P = 8.60 \times 10^{-14}$) and upregulated (median = 6.52% increase; 95% CI = 4.76% to 8.22%; $P = 4.40 \times 10^{-11}$) genes compared with clones having lower SNP expression (Fig. 5c,d). Supporting this, 350 of the 664 genes affected by the Mendelian variant were significantly (FDR < 10%) differentially expressed in clones with high SNP expression, compared with only 186 in clones with low SNP expression. These results provide experimental demonstration that an eQTL can modify the penetrance of a disease-causing coding variant, and suggest a genetic regulatory modifier mechanism as a potential explanation for the variable penetrance of rs199643834 in Birt–Hogg–Dubé syndrome. While further animal models or analyses of large patient cohorts would be needed to fully describe how the cellular transcriptome effects may translate to modified penetrance at a complex phenotype level, the use of genome editing in relevant cell lines and the transcriptome as a molecular phenotype will be an important and scalable approach for studying effects at individual genes of clinical importance.

Discussion

In conclusion, we have studied the hypothesis that regulatory variants in cis can affect the penetrance of pathogenic coding variants. We used diverse data types from population and disease cohorts, and experimental approaches, which together provide strong evidence of modified penetrance due to joint functional effects of

regulatory and coding variants. Our functional genomic and genetic analysis of the general population provides evidence that purifying selection is acting on joint regulatory and coding variants haplotypes. Importantly, this suggests that the combination of an individual's regulatory and coding variant genotypes has an effect on phenotype, since purifying selection acts only on traits that affect fitness. Notably, we observed a weaker signal when analyzing eQTL haplotype configurations from genetic data alone as compared with allele-specific expression (ASE) data. This difference could arise because the genetic analysis inferred expression haplotypes using the top common regulatory variant per gene as opposed to directly measuring them using expression data. Such an approach does not capture the combinatorial effects of independent common regulatory variants or the effects of rare regulatory variation, both of which might make significant contributions to modified penetrance.

Our case–control analyses of autism and cancer cohorts provide direct evidence that regulatory modifiers of coding variants contribute to disease risk, which is jointly driven by the combination of an individual's eQTL and coding variant genotypes. Furthermore, our experimental approach provides an indication of potential regulatory modifiers in the Mendelian Birt–Hogg–Dubé syndrome. The approaches developed and introduced in this work can be applied to additional disease data sets, with GTEx data providing an essential resource of regulatory variants to empower these analyses. In individual genes, finding regulatory modifiers will require relatively large data sets, and studies of large families with segregating coding variants may be a particularly powerful approach. Genome editing

experiments, as we demonstrated for *FLCN*, will be important for functionally validating results from computational analysis.

A key component of our work was the integrated analysis of rare coding variants and common regulatory variants, which are too often considered as separate domains in human genetics, despite the fact that their interplay is gaining increasing interest[34]. Currently, rare coding variants are studied largely by exome sequencing in relatively rare diseases, and common regulatory variant analyses are focused on applications in genome-wide association studies of common diseases. Setting the stage for future studies, our work supports one of the few concrete and generalizable models of modified penetrance of genetic variants in humans, with a clear biological mechanism based on the net effect of variants on the dosage of functional gene product, and is backed by solid empirical analysis of genome-wide genetic data.

This work opens additional important areas for future research. Our results demonstrate that the strength of modified penetrance depends on the functional importance and dosage sensitivity of the gene, the effect size of the regulatory variants that affect expression or splicing, and the type of coding variant. Larger data sets are needed to uncover this full spectrum at the level of individual genes instead of the gene classes analyzed here. In this work, we focused on loss-of-function analysis, in which the expression level of the *non-mutant* haplotype matters, but it is likely that for less common gain-of-function germline and somatic variants, modified penetrance may depend on the expression of the *mutant* haplotype instead. This may be an important consideration for potential future work on variable penetrance of somatic variants in cancer. The dynamics of natural selection on haplotype combinations will be an interesting area of population genetic analysis, in which an individual's fitness depends on multiple variants on different homologs as well as linkage disequilibrium between these variants.

Finally, we highlight that while other mechanisms are also likely to contribute to variable penetrance of coding variants, analysis of cis-regulatory modifiers is particularly tractable, with multiple practically feasible approaches introduced in this work. Our findings highlight the importance of considering coding variation in the context of regulatory haplotypes in future studies of modified penetrance of genetic variants affecting disease risk.

URLs. SFARI gene database, https://gene.sfari.org; Tumor Suppressor Gene Database, https://bioinfo.uth.edu/TSGene/; GTEx Portal, https://gtexportal.org/; CRISPR Design Tool, http://crispr.mit.edu.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0192-y.

References

1. Chen, R. et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016).
2. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
3. Milne, R. L. & Antoniou, A. C. Genetic modifiers of cancer risk for *BRCA1* and *BRCA2* mutation carriers. *Ann. Oncol.* **22**(Suppl 1), i11–17 (2011).
4. Emison, E. S. et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).
5. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).
6. Snozek, C. L. H. et al. LDLR promoter variant and exon 14 mutation on the same chromosome are associated with an unusually severe FH phenotype and treatment resistance. *Eur. J. Hum. Genet.* **17**, 85–90 (2009).
7. Alberobello, A. T. et al. An intronic SNP in the thyroid hormone receptor β gene is associated with pituitary cell-specific over-expression of a mutant thyroid hormone receptor β2 (R338W) in the index case of pituitary-selective resistance to thyroid hormone. *J. Trans. Med.* **9**, 144 (2011).
8. Butt, C. et al. Combined carrier status of prothrombin 20210A and factor XIII-A Leu34 alleles as a strong risk factor for myocardial infarction: evidence of a gene-gene interaction. *Blood* **101**, 3037–3041 (2003).
9. Amin, A. S. et al. Variants in the 3' untranslated region of the KCNQ1-encoded Kv7.1 potassium channel modify disease severity in patients with type 1 long QT syndrome in an allele-specific manner. *Eur. Heart J.* **33**, 714–723 (2012).
10. Dimas, A. S. et al. Modifier effects between regulatory and protein-coding variation. *PLoS Genet.* **4**, e1000244–10 (2008).
11. Lappalainen, T., Montgomery, S. B., Nica, A. C. & Dermitzakis, E. T. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* **89**, 459–463 (2011).
12. Vu, V. et al. Natural variation in gene expression modulates the severity of mutant phenotypes. *Cell* **162**, 391–402 (2015).
13. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
14. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
15. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
16. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
17. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).
18. Irimia, M. et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).
19. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
20. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817 (2016).
21. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
22. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
23. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
24. Payne, S. R. & Kemp, C. J. Tumor suppressor genetics. *Carcinogenesis* **26**, 2031–2045 (2005).
25. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
26. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
27. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
28. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
29. Toro, J. R., Wei, M.-H., Glenn, G. M. & Weinreich, M. BHD mutations, clinical and molecular genetic investigations of Birt–Hogg–Dubé syndrome: a new series of 50 families and a review of published reports. *J. Med. Genet.* **45**, 321–331 (2008).
30. Khoo, S. K. et al. Clinical and genetic studies of Birt–Hogg–Dubé syndrome. *J. Med. Genet.* **39**, 906–912 (2002).
31. Lin, Y.-C. et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* **5**, 4767 (2014).
32. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucl. Acids Res.* **40**, D930–4 (2012).
33. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucl. Acids Res.* **45**, W130–W137 (2017).
34. Werling, D. M. *et al.* Limited contribution of rare, noncoding variation to autism spectrum disorder from sequencing of 2,076 genomes in quartet families. Preprint at https://doi.org/10.1101/127043 (2017).

Acknowledgements

Author contributions

S.E.C. and T.L. designed the study and wrote the manuscript. S.E.C., A.V., and T.L. designed analyses and experiments. S.E.C., A.C., F.A., A.W., and A.V. performed analyses and experiments. P.M. aided development of the test for regulatory modifiers of penetrance. F.R. and R.G. provided and assisted in analysis of GTEx PSI data. I.I. provided and assisted in the analysis of SSC data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0192-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.E.C. or T.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Variant annotation. Variant annotations for SNPs were retrieved from CADD v1.3[15]. As per guidelines by the CADD authors, missense variants with a CADD PHRED score of >15, which is the median CADD score across all possible canonical splice site and missense variants, were defined as potentially pathogenic. Synonymous variants with a CADD PHRED score <15 were used as controls. To be considered rare, variants were required to have a MAF <1% across GTEx v7, 1000 Genomes Phase 3[35] and gnomAD r2.0.1[22].

GTEx allelic expression analysis. GTEx v6p allelic expression data generated from whole exome sequencing genotypes were used[14]. Variants that were in low-mappability regions (University of California, Santa Cruz (UCSC) mappability track <1), showed mapping bias in simulations[36] or had significant (FDR <1%) evidence that the variant was monoallelic in that individual across all GTEx tissues were excluded to reduce mapping bias[16]. Only variants with at least ten reads were used. To minimize the probability that the observed allelic imbalance was due to effects of the allelic expression variants themselves on splicing, only variants farther than 10 base pairs (bp) from an annotated splice site[15] were used. Within each GTEx tissue, when allelic expression measurements from the same variant were present from different individuals, the measurement with the highest read coverage was used. Only variants where the alternative allele was the minor allele were used to ensure that mapping biases were consistent across variants. For missense variants, matched synonymous controls were selected, controlling for allele frequency within 25% of missense variants (for example, between 0.75% and 1.25% for a 1% frequency missense variant). Allelic fold change (aFC) was calculated as log2(alternative allele reads +1/ reference allele reads +1)).

GTEx exon inclusion quantification analysis. Individual-level quantifications of exon inclusion were generated for all GTEx v6p samples with the VAST-TOOLS pipeline, which measures the PSI of each exon in each individual[18]. Within a given tissue, for each exon with at least ten PSI measurements, PSI z-scores were generated for each sample. Individuals with substantial variation in exon inclusion compared with the population were defined as the top 10% of PSI z-scores across all sample exons (Supplementary Fig. 2d).

GTEx expression quantitative trait loci (eQTL). The official set of GTEx v6p top significant (FDR <5%) eQTLs by permutation P-value was used for all analyses, such that each gene by tissue had at most a single eQTL[14]. Those eQTLs for which the 95% CI of eQTL effect size overlapped 0, representing weak eQTLs, were discarded[17]. To produce a single set of cross-tissue top eQTLs, the top eQTL by FDR across tissues was selected for each eGene, with ties broken by choosing the eQTL with the larger effect size. This resulted in a set of 26,942 eGenes, each with a single eSNP (Supplementary Table 1).

Genetic data and haplotype phasing. GTEx: GTEx v7 genotypes from whole genome sequencing of the 620 individuals who had at least one RNA sample were used. These genomes were population and read-back phased using DNA-seq reads with SHAPEIT2[19]. Following this, phASER v1.0.0 was used to perform read-backed phasing using RNA-seq reads[20] from all samples for each individual: a median of 17 tissues, ranging from 1 to 38. For RNA-seq-based read-backed phasing, only uniquely mapping reads (STAR MAPQ 255) with a base quality of ≥10 overlapping heterozygous sites were used, and all other phASER settings were left as default. The resulting phased genotypes were imputed into 1000 Genomes Phase 3[35] with Minimac3 v2.0.1[37].

SSC: Genotypes of the SSC cohort from Sanders et al. consisting of data generated on Illumina 1Mv1, 1Mv3 and Omni2.5 arrays[26] were transmission phased using SHAPEIT2 with relatedness data[38] and then imputed into the 1000 Genomes Phase 3 panel using the Sanger Imputation Service with the positional Burrows–Wheeler transform[39,40]. Coding variants called from whole exome sequencing data in Iossifov et al.[27] were transmission phased on a per variant basis when possible using the genotypes of both parents. In total, genetic data from 2,304 ASD-affected probands and 1,712 unaffected siblings were used for the analysis. Expression haplotypes of coding variants were annotated on the most significant eQTL variant for each gene in GTEx v6p across all tissues. The top GTEx eQTLs from across all tissues were used for analysis instead of brain regions only due to the substantially lower sample sizes in GTEx brain tissues, which result in fewer eQTLs being discovered.

TCGA: Paired tumor and normal WGS reads from 925 individuals across 15 cancer types were used to call germline and somatic variants with Bambino v1.06[41]. The resulting germline genotypes were population phased with EAGLE2 v2.3[42] using the 1000 Genomes Phase 3 panel[35] and read-back phased with phASER v1.0.0[20]. For read-backed phasing, only reads with mapping quality (MAPQ) ≥30 and with a base quality of ≥10 overlapping heterozygous sites were used, and all other phASER settings were left as default. The resulting phased genotypes were imputed into 1000 Genomes Phase 3[35] with Minimac3 v2.0.1[37]. Due to the highly variable sequencing depth across TCGA whole genome libraries, from the 925 individuals, 615 individuals with high-quality genotyping and phasing were selected for downstream analysis by filtering the bottom 30% of samples by number of variants called and median EAGLE phase confidence across autosomes.

This resulted in an approximately equal number of TCGA (615) and GTEx (620) individuals for analyses. Expression haplotypes of coding variants were annotated on the most significant eQTL variant for each gene in GTEx v6p across all tissues.

The TCGA individuals analyzed and GTEx v7 individuals used as a control had very similar inferred ancestry compositions, although the TCGA individuals had a slightly higher proportion of individuals with Asian ancestry (Supplementary Table 7). To ensure that the results were robust to ancestry proportions, we performed our analysis after removing these individuals from the TCGA data set. We found that while the analysis was less powered, resulting in larger CIs for the TCGA cohort, the results were consistent and significant (Supplementary Fig. 6c).

Test for regulatory modifiers of penetrance using phased genetic data. Here we test the hypothesis that in loss-of-function coding variant heterozygotes, decreased expression of the major, or WT, coding allele mediated by an eQTL can increase the penetrance of the mutant allele by decreasing the dosage of functional gene transcript, and vice versa (Supplementary Fig. 1). The null hypothesis is that eQTL-mediated changes of major allele expression have no effect on the penetrance of mutant alleles. Since penetrance cannot be easily measured, we instead measure the frequency at which the major allele is observed on the lower-expressed eQTL haplotype (Supplementary Fig. 3a). Under the null hypothesis, a coding mutation would occur in random individuals in the population, and on random haplotypes in those individuals, irrespective of their eQTL genotype. Thus, under the null hypothesis, the frequency of observed major alleles on lower-expressed haplotypes would simply be equal to the frequency of the lower-expressed eQTL allele in the population. Alternatively, an increased frequency indicates an enrichment of haplotype configurations that increase coding variant penetrance in the population studied, and vice versa (Supplementary Fig. 3b). Importantly, the test is calibrated to the eQTL frequency in the specific population studied, so it is internally controlled for differences in, for example, eQTL allele frequencies between cases and controls.

To perform the test, for each observation of a heterozygous coding variant of interest, the phased genotypes of the coding variant and the top GTEx cross-tissue eQTL for that gene are used to produce a binary measure of whether the major coding allele is on the lower-expressed haplotype (Supplementary Fig. 3a). Alongside this binary measure, the frequency of the lower-expressed eQTL allele is recorded.

For each observation of a heterozygous coding variant in a single individual with genotype g, let A and a denote the higher- and lower-expressed eQTL allele, respectively, and B and b denote the major and minor coding variant allele, respectively. We assume that the minor allele is the non-functional allele. For a given haplotype g, we define the indicator function β such that it is 1 if the functional allele is on a lower-expressed eQTL haplotype, and 0 otherwise:

$$\beta(g) = \begin{cases} 1 \text{ if } g \in \{(ab/aB), (Ab/aB)\} \\ 0 \text{ if } g \in \{(Ab/AB), (ab/AB)\} \end{cases}$$

For a given haplotype, the expectation for β under the null model, where the haplotype configurations are random (H_0), is

$$E[\beta(g)] = \begin{cases} 0.5 \text{ if } g \in \{(A/a)\} \\ f(a)^2 / (f(a)^2 + (1-f(a))^2) \text{ if } g \in \{(A/A), (a/a)\} \end{cases}$$

where $f(a)$ is the population frequency of the lower expressed eQTL allele included in the tested haplotype g.

The indicator function β and its expectation under the null model are calculated across all individuals, genes and variants. The average relative deviation of the observed mean of β from its expectation is calculated as

$$\varepsilon = \frac{1}{N} \sum_{n=1}^{N} \frac{\beta(g_n) - E[\beta(g_n)]}{E[\beta(g_n)]}$$

where N is the total number of observed haplotype configurations consisting of an eQTL and coding variant, pooled over all individuals, variants and genes.

CIs for ε are generated by bootstrapping genotypes, and the two-sided empirical P-value against H_0 is calculated as

$$P(H_0) = 2 \min \left[\frac{\sum_{b=1}^{B} \varepsilon_b < 0}{B}, \frac{\sum_{b=1}^{B} \varepsilon_b > 0}{B} \right]$$

where B is the total number of bootstraps.

We ran the test on simulated haplotype data from 1,000 individuals at 500 genes with 1,000 replicates. The lower-expressed haplotype frequency was set to 50% and the coding variant frequencies as observed in GTEx. This was done across a range of genes exhibiting a bias of major coding alleles being found on lower-expressed haplotypes and strengths of this bias. For the test, 1,000 bootstrap samples were used. We found that at the 5% significance threshold, 5% of simulation replicates were significant, suggesting that the test is well calibrated under the null hypothesis. For real-world data, reported in the study, we used 100,000 bootstrap samples to calculate P-values and derive CIs.

This is a similar problem to that addressed by the Poisson binomial distribution, which describes the sum of successes in a set of independent Bernoulli trials with different success rates. However, the bootstrap approach is more convenient for calculating CIs and accounting for differences in sample size between control genes and genes of interest. We compared P-values derived from our test with those derived from a Poisson binomial distribution with parameters $E[\beta(g_1)] \ldots E[\beta(g_N)]$. In practice, our P-values are very similar to that generated using the Poisson binomial distribution (Pearson correlation = 0.996, slope = 0.997, Supplementary Fig. 3e).

A key part of our test is that, as opposed to simple linkage disequilibrium, it tests a specific directional hypothesis: the frequency that coding variant functional alleles are on lower-expressed regulatory haplotypes. Thus, in the absence of selection on regulatory haplotype configurations, differences in recombination rates between genes would not be expected to bias the results of our test. However, it is possible that the distribution of distances between the coding and regulatory variants tested could differ between test sets. To ensure that this was not the case, we compared the distance between coding and eQTL variants for each of the relevant tests performed and saw no significant difference in distance distribution for any of the relevant test pairs (Supplementary Fig. 8).

Gene sets. Genes with strong eQTLs were selected as the top 25% of eGenes by absolute eQTL effect size[17]. A conservation score was calculated for each eGene as the median UCSC hg19 placental mammal base conservation across all exons. Loss-of-function and missense intolerant genes were selected by requiring ExAC pLI ≥0.9 and significant missense constraint (FDR <10%)[22]. P-values for missense constraint were generated from ExAC missense z-scores using the R command 'p = 2*pnorm(-abs(mis_z))' and Benjamini–Hochberg corrected to control for FDR. A broad set of genes associated with ASD was produced by combining high-confidence Simons Foundation Autism Research Initiative (SFARI) database genes (see URLs, categories 1, 2 and S) downloaded on 20 October 2017, genes from Krumm et al. with nominally significant ($P < 0.05$) enrichment of de novo single-nucleotide variants (SNVs) in probands versus siblings, and genes with recurrent likely gene-disrupting and missense de novo mutations in probands but not in unaffected siblings in Iossifov et al.[27,28]. These were further filtered by removing genes that are highly tolerant to genetic variation, as defined by being in the top 10% of tolerant genes by Residual Variation Intolerance Score (RVIS) (v3_12Mar16)[43]. In total, this resulted in a list of 455 ASD-associated genes. A list of 983 downregulated tumor suppressor genes in tumor samples versus normal tissue in The Cancer Genome Atlas (TCGA) expression data was downloaded from the Tumor Suppressor Gene Database[44] website (see URLs) on 24 August 2017.

CRISPR/Cas9 guide selection and cloning. Before RNA design and editing, we verified the genotype at the regions of interest, namely the Mendelian variant rs199643834 and eQTL variant rs1708629. Crude extracts prepared from 293T cells were used to amplify these regions using forward and reverse genotyping primers FLCN_genot and FLCNeQTL_genot, respectively (Supplementary Table 8). Amplicons were sequenced both by Sanger sequencing and on the Illumina MiSeq. The 293T cell genotype was Ref/Ref/Ref at rs199643834 and Ref/Ref/Alt at rs1708629. There were no single-nucleotide changes close to rs199643834 that might have affected single-guide RNA (sgRNA) activity or required modified homologous template.

Using computational algorithms with prioritization for on-target efficiency and reduced off-target effects from CRISPR Design tool (see URLs) and E-CRISP[45], we identified *Streptococcus pyogenes* Cas9 (SpCas9) guide RNAs that bind near variant rs199643834 (A>G). We selected three sgRNA sequences within 50 bp of the target SNP (rs199643834), which were predicted to result in maximum cleavage efficiency without off-target effects (Supplementary Table 8). Annealed oligomers inclusive of guide RNA sequences were sub-cloned into the lentiCRISPRv2 plasmid (Addgene plasmid #52961), which contains expression cassettes for the guide RNA, a human codon-optimized Cas9 and a puromycin resistance gene[46]. Plasmids were transformed into chemically competent *E. coli* and grown at 30°C; plasmid DNA was extracted and purified. A 150 bp single-stranded DNA template (ssODN) for precise editing by homologous recombination (HDR) carrying the rs199643834 A allele was designed and obtained from Integrated DNA Technologies (IDT) DNA in the form of lyophilized ultramer (Supplementary Table 8).

Transfections and T7 endonuclease I (T7E1) assays. Human 293T cell line (ATCC, cat. no. CRL-3216) was adapted to and subsequently routinely grown in Opti-MEM/5% CCS (newborn calf serum), 1% GlutaMAX, 1% Penicillin/Streptomycin and sodium pyruvate. For transfection with Cas9- and sgRNA-expressing plasmids as well as ssODN template, cells were harvested for seeding at a log growth phase (approximately 70% confluency). In a six-well format, 300,000 293T cells were seeded a day prior to transfection. The next day, 2 µg of each lentiCRISPR v2 plasmid and 0.5 µg of ssODN HDR template were delivered into the cells using Lipofectamine 3000 reagent. At 24 hours post-transfection, selective pressure in the form of 5 µg ml⁻¹ puromycin was applied for 8 hours to enrich for transfected cells. The short time-frame reduces the chances of selecting monoclonal lines with stable plasmid integration. Following 2 days of cell growth, cells were harvested and crude extracts prepared from a small fraction for

genotyping. The remainder of the cells were frozen for subsequent isolation of cell lines containing desired edits.

For T7E1 assays, a 362 bp region flanking rs199643834 was amplified by PCR from the crude extracts using FLCN_genot primers and purified using Ampure XP beads. Purified products were heteroduplexed, digested with T7 endonuclease 1 and run on a 2% agarose gel. Cleavage patterns from editing experiments conducted with each sgRNA were qualitatively analyzed to determine the cutting efficiency of each Cas9/sgRNA to guide further experiments. Subsequently, the crude cell lysates were used to prepare amplicon libraries containing ScriptSeq adapters, which were sequenced on the Illumina MiSeq instrument with paired-end 150 bp reads. Rates of indel mutations by non-homologous end joining (NHEJ) and precise SNP editing by homology-directed repair (HDR) were determined by an in-house analysis pipeline.

Generation and identification of monoclonal cell lines containing desired precise edits. The initial screening showed that editing of 293T polyclonal cell population at rs199643834 with sgRNA 1 resulted in the highest rate of HDR. This population was selected for single-cell sorting in a 96-well format to obtain monoclonal edited cell lines. Following 10 days of cell growth, individual wells were scored for the presence of healthy colonies, and altogether approximately 1,920 healthy colonies were screened. At first passage, a third of the cells from each well were collected for crude cell extracts and genotyping.

High-throughput genotyping was performed by preparing an amplicon library from each crude extract with Nextera adapters enabling differential custom dual-indexing. Screening for desired mutations was performed using in-house software. In total, four WT (Ref/Ref/Ref), seven heterozygous (Ref/Ref/Alt) and two homozygous mutant (Alt/Alt/Alt) clones with each desired mutation were expanded for downstream analyses.

Targeted RNA-seq of allelic series and eQTL phasing. Expanded lines were grown to 70–80% confluency, and RNA was isolated using the Qiagen RNAeasyMini kit. cDNA was synthesized from each RNA sample, and the region spanning the Mendelian variant rs199643834 was amplified using primers FLCN_exon9-10-F and FLCN_exon11-R2, containing Nextera adapters (Supplementary Table 8). Targeted amplicons were dual-indexed using custom Nextera indexes and sequenced on the Illumina MiSeq with 2 × 150 bp reads.

For all the 13 lines, the genotype determined by DNA-seq was confirmed by RNA-seq. For the seven lines with a single copy of the edited SNP, we performed allelic expression analysis. Reads were aligned to hg19 using STAR[47]. The number of reads mapping to the reference and alternative alleles was quantified using Allele Counter, requiring MAPQ = 255 and base quality ≥10 (Castel et al.[16]). Across samples, there was a median of 34,870 reads passing filters overlapping the site. A binomial test using reads containing the edited SNP allele against a null hypothesis of 1/3 (corresponding to a single copy of the edited SNP) was performed. Copy number-normalized allelic expression (AE) of the edited SNP was calculated as log2((ALT_COUNT/REF_COUNT)/ (1/3)). Samples with AE <0 and binomial $P < 0.01$ were categorized as snpLOW (edited SNP on lower-expressed eQTL haplotype), and those with AE >0 and binomial $P < 0.01$ were categorized as snpHIGH (edited SNP on higher-expressed eQTL haplotype).

RNA-seq and gene expression analysis of edited 293T cells. RNA-seq libraries were prepared using the TruSeq Stranded mRNA Library Sample Preparation Kit in accordance with manufacturer's instructions. Briefly, 500 ng of total RNA was used for purification and fragmentation of mRNA. Purified mRNA underwent first and second strand cDNA synthesis. cDNA was then adenylated, ligated to Illumina sequencing adapters and amplified by PCR (using ten cycles). Final libraries were evaluated using fluorescent-based assays including PicoGreen and Fragment Analyzer, and were sequenced on the Illumina NovaSeq Sequencing System using 2 × 100 bp cycles to a median depth of 52.8 million reads. Trimmomatic[48] v0.36 was used to clip Illumina adapters and quality trim, and reads were aligned to hg19 using STAR[47] in two-pass mode. A median of 98% of reads mapped to the human genome, with a median of 95.2% reads mapping uniquely. featureCounts[49] v1.5.3 was used in read counting and strand-specific mode (-s 2) with primary alignments only to generate gene-level read counts with Gencode v19 annotations used in GTEx v6p[14]. Differential expression analysis was performed using DESeq2[50] v1.16.1 and R v3.4.0 on genes with a mean of greater than five counts across samples. FDR correction of P-values was performed using Benjamini–Hochberg. Gene set enrichment analysis on differential expression data was performed using the Web-based Gene Set Analysis Toolkit[33] with Wikipathway enrichment categories.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. GTEx v6p eQTLs are publicly available through the GTEx Portal (see URLs). GTEx genotype data, allelic expression data and RNA-seq reads are available to authorized users through dbGaP (study accessions phs000424.v6.p1 and phs000424.v7.p2). TCGA data are available to authorized users through dbGaP

(study accession phs000178.v9.p8). 293T RNA-seq data generated in this study are available through Gene Expression Omnibus (GEO) under accession GSE116061.

References

35. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
36. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15**, 467 (2014).
37. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
38. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
39. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
40. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
41. Edmonson, M. N. et al. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* **27**, 865–866 (2011).
42. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1–8 (2016).
43. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
44. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucl. Acids Res.* **44**, D1023–D1031 (2015).
45. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11**, 122–123 (2014).
46. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

nature research

Corresponding author(s):   Castel SE, Lappalainen T

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	SHAPEIT v2, phASER v1.0.0, Minimac3 v2.0.1, EAGLE2 v2.3 , Trimmomatic v0.36, STAR v2.5.0a, featureCounts v1.5.3, DESeq2 v1.16.1, R 3.4.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

GTEx v6p eQTLs are publicly available through the GTEx Portal (https://gtexportal.org/). GTEx genotype data, AE data, and RNA-seq reads are available to authorized

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the availability of existing data (GTEx, TCGA, and SSC cohorts).
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	Unique biological materials
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology
☒	Animals and other organisms
☒	Human research participants

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging