Modified penetrance of coding variants by *cis*-regulatory variation shapes human traits

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

1. New York Genome Center, NY, USA
2. Department of Systems Biology, Columbia University, NY, USA
3. Systems biology of drug resistance in cancer, Research Programs Unit, Genome-Scale Biology, University of Helsinki, Finland
4. Broad Institute of MIT and Harvard, Cambridge, USA
5. Centre for Genomic Regulation, Barcelona, Spain

* Corresponding authors

Abstract

Coding variants represent many of the strongest associations between genotype and phenotype, however they exhibit inter-individual differences in effect, known as variable penetrance. In this work, we study how *cis*-regulatory variation modifies the penetrance of coding variants in their target gene. Using functional genomic and genetic data from GTEx, we observed that in the general population, purifying selection has reduced haplotype combinations that lead to higher penetrance of pathogenic coding variants. Conversely, in the germline genomes of individuals with cancer, we observed an increase in predicted penetrance of pathogenic coding variants in disease relevant genes. Finally, we experimentally demonstrated that a regulatory variant can modify the penetrance of a coding variant by introducing a Mendelian SNP using CRISPR-Cas9 on distinct expression haplotypes and using the transcriptome as a phenotypic readout. Our results demonstrate that joint effects of regulatory and coding variants are an important part of the genetic architecture of human traits, and contribute to modified penetrance of disease-causing variants.

Main Text

Variable penetrance is a common phenomenon that causes individuals carrying the same variant to often display highly variable symptoms, even in the case of Mendelian and other severe diseases driven by rare variants with strong effects on phenotype [1]. This is a key challenge in understanding of how genetic variants manifest in human traits, and a major practical caveat in clinical genetics. However, the causes and mechanisms of variable penetrance are poorly understood. One potential cause of variable penetrance involves genetic variants with additive or epistatic modifier effects [2]. While some studies have successfully mapped genetic modifiers of, for example, BRCA [3] and RETT [4] mutations, genome-wide analysis of pairwise interactions between variants has proven to be challenging in humans [5]. Here, we use the term variable penetrance as a joint description of both variable expressivity (severity of phenotype) and penetrance (proportion of carriers with phenotype).

In this study, we analyze 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. (Figs. 1, S1). This phenomenon has been demonstrated to affect penetrance of disease-predisposing variants in individual loci [6-9], and explored in early functional genomic datasets [10,11]. In this work, we use large-scale functional genomics and disease cohort data sets as well as genome editing with CRISPR-Cas9 to demonstrate the role of regulatory variants affecting gene expression and splicing as modifiers of coding variant penetrance. We focus on rare pathogenic coding variants from exome and genome sequencing data that provide the best characterized group of variants with strong phenotype 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.

First, we analyzed data from the general human population to test the hypothesis that natural selection should favor haplotype configurations that reduce the penetrance of pathogenic coding variants. Throughout this study, we defined the predicted pathogenicity of variants using their CADD score (see Material and Methods – Variant Annotation) [12]. Using genotype and RNA-sequencing data of 7,051 samples from 449 individuals of the Genotype Tissue Expression (GTEx) project v6p [13,14], we first measured the regulatory haplotype of coding variants using allelic expression data, which captures *cis* effects of both splice and expression regulatory variation [15] (Fig. 2a). Supporting our hypothesis, missense variants showed reduced allelic expression that was proportional to their predicted pathogenicity, suggesting that they are enriched on lower expressed or exon-skipping regulatory haplotypes (Fig. 2b). When compared to derived allele frequency (DAF) matched benign synonymous variants, rare (DAF < 1%) potentially pathogenic missense variants had significantly reduced expression ($p = 1.80e-6$), but rare benign missense variants did not ($p = 0.203$) (Fig. 2c), consistently across the 44 GTEx tissues (Storey's $\Pi_1 = 0.74$; Fig. S2a). This effect remained in an analysis of a small subset of variants that were in constitutively included exons in the individual harboring the variant ($p = 0.0135$; Fig. S2b), suggesting that pathogenic variants are enriched in lower expressed haplotypes. Next, we analyzed if splice regulatory variation specifically might reduce coding variant penetrance (Fig. 2d). We quantified exon inclusion in each sample by percent spliced in (PSI) [16] (Fig. S2c), and analyzed exons showing inter-individual inclusion variability (Fig. S2d-e). We observed that in the individual harboring the missense variant, the probability that its exon was spliced in was proportional to its predicted pathogenicity (Fig. 2e). When compared to DAF matched benign synonymous variants, rare potentially pathogenic missense variants had significantly reduced exon inclusion ($p = 1.60e-4$), but rare benign missense variants did not ($p = 0.465$) (Fig. 2f). This suggests that pathogenic variants are more likely to accumulate in haplotypes where the corresponding exon is less likely to be included in transcripts.

While allelic expression and splice quantification provide powerful functional readouts 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 higher expression of the haplotype carrying the "wildtype", major coding allele reduces the penetrance of the rare allele, and vice versa (Figs. 3a, S1). To study this, we developed a test for regulatory modifiers of penetrance that uses phased genetic data. For each rare coding variant heterozygote we test whether the major coding allele is on the higher expressed eQTL haplotype (Fig. S3a) and determine if this occurs more or less frequently than would be expected based on eQTL frequencies in the population studied (Fig. S3b). Using simulated data, we found that our test was well calibrated under the null while still being sensitive to changes in haplotype configuration (Fig. S3c-d).

To analyze whether the distribution of coding variants on *cis*-eQTL haplotypes in GTEx showed signs of selection towards reduced penetrance, we produced a large set of haplotype phased genetic data from GTEx v7, where 30x whole genome sequencing of 620 individuals was available. This was obtained from population based phasing paired with read-backed phasing using DNA-seq reads [17] and RNA-seq reads [18] from up to 38 tissues for a single individual. This allowed us to analyze the haplotypes of 211,575 rare (MAF < 1%) coding variants at thousands of genes with known common (MAF > 5%) eQTLs [14] (Fig. S4a, Table S1). Using our test for regulatory modifiers of penetrance to analyze all protein coding genes in the GTEx data set we did not observe any signs of reduced penetrance of rare potentially pathogenic missense variants ($p = 0.682$). However, hypothesizing that selection may be acting primarily at genes that are associated to a phenotype, we focused on a broad set of genes with known phenotypic association [19]. For rare potentially pathogenic missense variants at these genes, we observed a

significant (p = 0.0230) increase of 0.85% in the frequency of haplotypes where the major coding allele was more highly expressed than would be expected under the null, while no effect was seen for benign missense (p = 0.480) or benign synonymous variants (p = 0.470) (Fig. 3b). Similarly, we also observed a significant reduction of predicted penetrance of rare potentially pathogenic missense variants but not controls in genes with a strong eQTL (p = 9.92e-3) and the most loss-of-function intolerant genes (p = 9.60e-4) (Fig. 3b). Altogether, combined with observations from functional data of allelic expression and exon inclusion, these results suggest that joint effects between regulatory and coding variants have shaped human genetic variation through natural selection favoring haplotype configurations where *cis*-regulatory variants reduce the penetrance of pathogenic coding variants (Fig. S1).

Having observed reduced penetrance by analyzing signals of selection in the general population, 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 penetrance of putatively disease-causing coding variants – the opposite pattern to that seen in GTEx. To this end, we investigated the role of regulatory modifiers of penetrance in germline cancer risk using genetic data from the Cancer Genome Atlas (TCGA) [20]. Cancer is a strong candidate for study, due to its well understood genetic basis, large accessible data sets, and the established role that tumor suppressor gene dosage plays in disease [21]. For 925 individuals across 15 cancers where whole genome sequencing reads were available to us (Table S2), we called germline variants and phased these using population [22] and read-backed phasing [18], and analyzed haplotypes of coding variants and common regulatory variants annotated based on the most significant eQTL variant for each gene in GTEx v6p (Fig. S4b). Again, we applied our test, and found that at tumor suppressor genes whose expression was downregulated in tumor versus normal TCGA samples [23], TCGA individuals had a significant decrease of major coding alleles of rare potentially pathogenic variants found on higher expressed haplotypes (-2.68%; p = 0.0319), suggesting increased penetrance of potential germline cancer risk variants in cancer patients (Fig. 3c). Using GTEx individuals as a control, we observed an increase in major alleles of rare potentially pathogenic missense variants found on higher expressed haplotypes (+2.11%, p = 0.0383), indicating reduced penetrance. This is consistent with the analysis of a larger class of phenotype-associated and loss-of-function intolerant genes in GTEx (Fig. 3b), providing additional evidence that selection may have favored haplotype configurations that reduce coding variant penetrance in genes associated to disease. In both TCGA and GTEx individuals, no significant effect was seen for benign missense or synonymous variants in tumor suppressor genes, or any variants in matched control genes. Altogether, this suggests that increased penetrance of pathogenic germline coding variants by regulatory variation increases cancer risk.

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 experimentally validate this observation by using CRISPR-Cas9 to introduce a coding variant on distinct regulatory haplotypes, followed by quantification of its penetrance from a cellular readout. Our finding that modified penetrance of germline variants by eQTLs may be involved in cancer risk lead us to study a missense SNP (rs199643834, K>R) in the tumor suppressor gene folliculin (*FLCN)* that has a common eQTL in most GTEx v6p tissues [14], and causes Mendelian autosomal dominant disease Birt-Hogg-Dubé Syndrome [24]. This disease results in characteristic benign skin tumors, lung cysts, and cancerous kidney tumors and shows variable penetrance [25]. We edited the SNP in a fetal embryonic kidney cell line (293T), which is triploid at the *FLCN* gene and harbors a single copy of a loss of expression eQTL (rs1708629) located in the 5' UTR of the gene [14,26]. 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 [27], thus being a strong candidate for the causal regulatory variant of the FLCN eQTL. We recovered monoclonal cell lines, genotyped them by targeted DNA-seq and performed targeted RNA-seq of the edited SNP (Fig. 4a). Allelic expression analysis showed that the haplotypes in the cell line are indeed expressed at different levels, likely driven by rs1708629 or another causal variant tagged by it, and the allelic

expression patterns allowed phasing of the coding variant with the eQTL (Figs. 4c, S5a). In this way, we obtained four clones with a single copy of the Mendelian variant on the lower expressed haplotype (snpLOW, Fig. 4c), three clones with a single copy on the higher expressed haplotype (snpHIGH, Fig. 4c), two monoallelic clones with three copies of the alternative allele, and four with only the reference allele (WT) of rs199643834. 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 (FDR < 10%) differentially expressed in clones monoallelic for the SNP versus wildtype controls (Fig. S5b, Table S3). Gene set enrichment analysis [28] of differential expression test results revealed significant (FDR < 10%) enrichment of pathways related to cell cycle control, DNA replication, and metabolism, consistent with the annotation of FLCN as a tumor suppressor, and the occurrence of tumors in patients with the mutation (Table S4). To study the joint effect of the eQTL and Mendelian variant, we quantified the differential expression of these 664 genes in low and high edited SNP expression clones separately (Fig. 4b). As we predicted, clones with higher expression of the SNP showed a significantly stronger differential expression of both downregulated ($p = 8.60e-14$, Fig. 4d) and upregulated (4.40e-11, Fig. 4e) genes compared to lower SNP expression clones. These results provide experimental evidence that an eQTL can modify the penetrance of a disease-causing coding variant, and suggests a genetic regulatory modifier mechanism as a potential explanation of variable penetrance of rs199643834 in Birt-Hogg-Dubé Syndrome.

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, population and disease cohorts, and experimental approaches that together provide strong evidence for our model of modified penetrance due to joint functional effects of regulatory and coding variants. A key component of our analysis was integrated analysis of rare coding variants and common regulatory variants, which are too often considered as separate domains in human genetics. This work provides 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, supported by solid empirical analysis of genome-wide genetic data.

Our work opens important areas for future research. Larger data sets are needed to enable computational analysis at the level of individual genes to characterize how joint effects of regulatory and coding variants vary as a function of their effect size and type, as well as gene function. Analysis of regulatory modifier effects in diverse diseases will be of interest, as well as the study of modified penetrance of somatic variants in cancer. Furthermore, the dynamics of natural selection on haplotype combinations will be an interesting area of population genetic analysis. We note 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. Recently, analysis of loss-of-function variant interactions in humans has suggested that they act synergistically to impact fitness, supporting a role for epistatic interactions between coding variants contributing to human traits as well [29]. Altogether, 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.

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,** i11–7 (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. *Journal of Translational Medicine 2011 9:1* **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. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
13. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
14. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. *bioRxiv* 074450 (2016). doi:10.1101/074450
15. 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).
16. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–1523 (2014).
17. 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).
18. 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).
19. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucl Acids Res* **45,** D833–D839 (2017).
20. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
21. Payne, S. R. & Kemp, C. J. Tumor suppressor genetics. *Carcinogenesis* **26,** 2031–2045 (2005).
22. Abecasis, G. R. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
23. 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).
24. 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).
25. Khoo, S. K. *et al.* Clinical and genetic studies of Birt-Hogg-Dubé syndrome. *J. Med. Genet.* **39,** 906–912 (2002).

26. 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).
27. 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).
28. 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).
29. Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic epistasis. *Science* **356,** 539–542 (2017).
30. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
31. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
32. 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).
33. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *bioRxiv* (2016). doi:10.1101/078717
34. 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).
35. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48,** 1–8 (2016).
36. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).
37. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11,** 122–123 (2014).
38. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11,** 783–784 (2014).
39. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
40. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
41. 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).
42. 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).

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.

Acknowledgements

We would like to thank members of the Lappalainen lab and Ivan Iossifov for discussion surrounding the project, and both Kristin Ardlie and Sampsa Hautaniemi who supervised F.A. and A.C. respectively. We thank the GTEx donors for their contributions to science, the GTEx Laboratory, Data Analysis, and Coordinating Center (LDACC), and the GTEx analysis working group (AWG) for their work generating the resource. In particular we would like to thank Ayellet Segre and Xiao Li at the Broad for their work performing WGS variant calling and phasing of GTEx v7 data. The Genotype-Tissue Expression (GTEx)

Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Additionally, we would like to acknowledge the contribution of TCGA specimen donors, and The Cancer Genome Atlas Research Network for their analyses. Funds for the TCGA were provided by Cancer Institute and the National Human Genome Research Institute. T.L. and S.E.C. were supported by the NIGMS grant R01GM122924 and NIMH grant R01MH101814, T.L., S.E.C., and P.M. were supported by the NIH contract HHSN2682010000029C, T.L. and P.M. were supported by NIMH grant R01MH106842, and T.L. was supported by the NIH grant UM1HG008901 and 1U24DK112331. A.C. was supported by the Cancer Society of Finland and Academy of Finland grant 284598. The authors declare that they have no competing financial interests.

Supplementary Materials

Materials and Methods
Table S1-S5
Figure S1-S5
References (30-42)

Figures

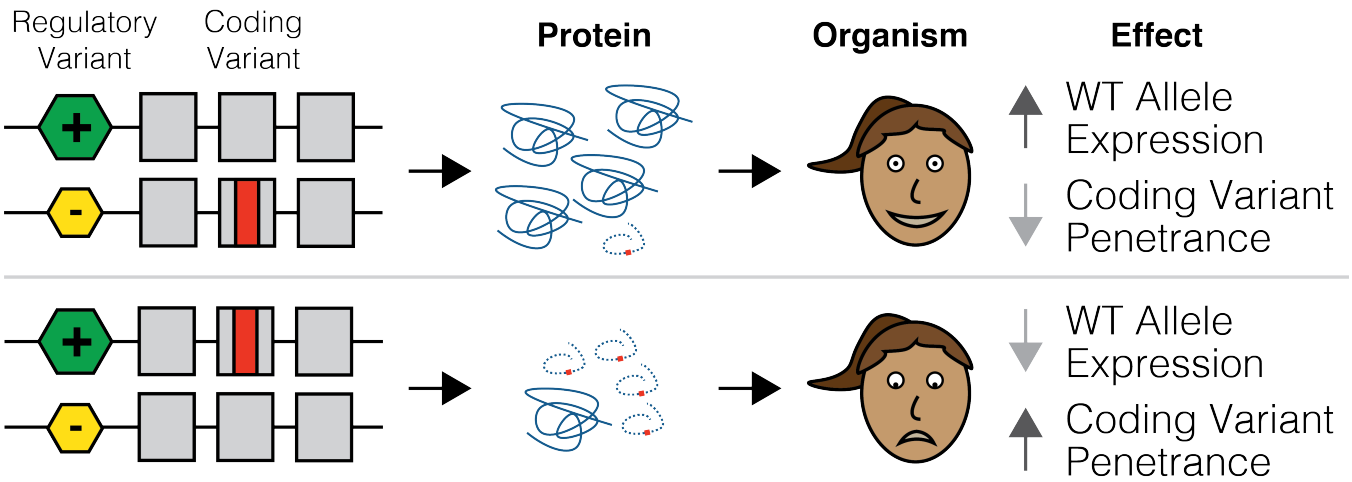

Figure 1. Regulatory variants as modifiers of coding variant penetrance. The hypothesis of this study is illustrated with an example where an individual is heterozygous for both a regulatory variant and a pathogenic coding variant. The two possible haplotype configurations would result in either decreased 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.

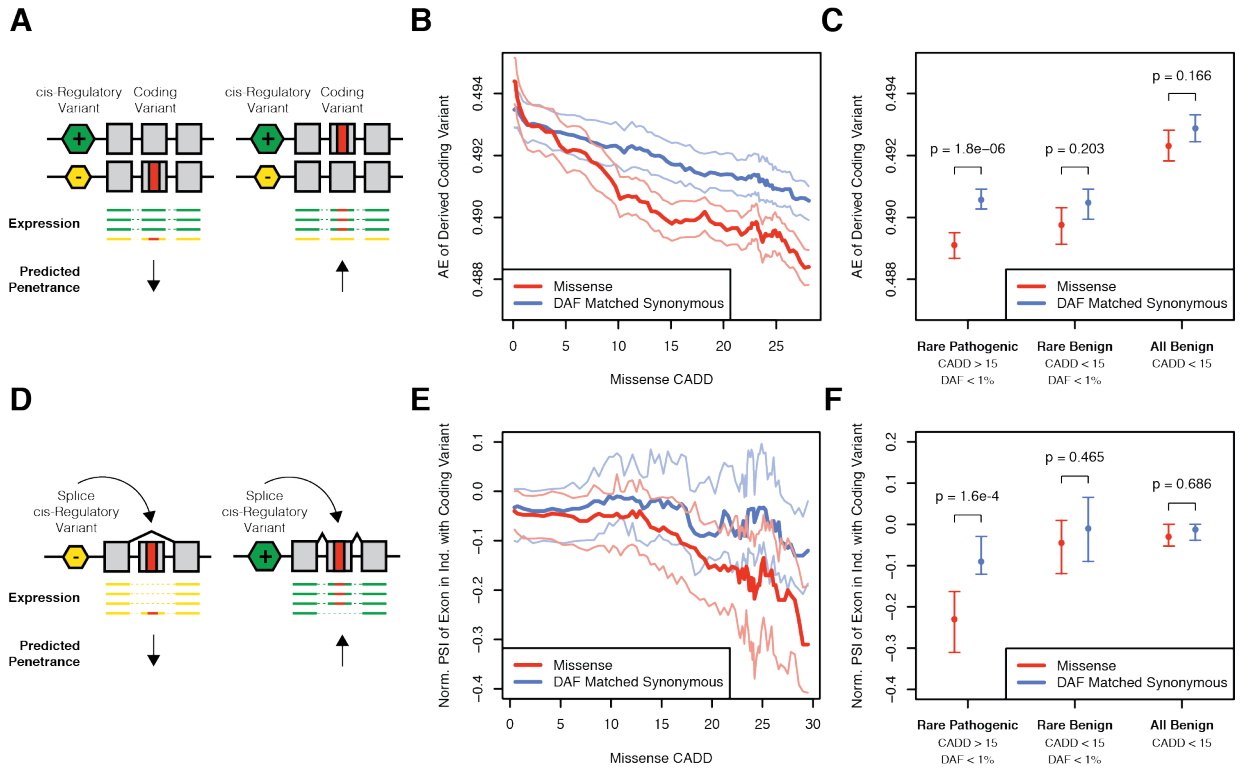

Figure 2. Functional genomic data reveals that pathogenic variants are enriched on lower expressed, and exon-skipped haplotypes. A) Allelic expression (AE) data can be used to measure the expression of a derived coding variant relative to the ancestral variant in heterozygous individuals. Reduced expression of the derived variant, observed as a decrease of AE suggests reduced penetrance by either splice or expression regulatory variation in that individual. **B)** Median GTEx v6p cross-tissue derived missense variant AE (red) as a function of predicted pathogenicity measured by CADD score, with derived allele frequency (DAF) matched benign synonymous variants (blue) as a control, and 95% confidence intervals. **C)** Comparison of median AE between missense and DAF matched benign synonymous variants with 95% confidence intervals. **D)** Percent spliced in (PSI) measures inclusion of a given exon in each individual. Reduced inclusion of the exon a derived variant is found in suggests reduced penetrance by splice regulatory variation in that individual. **E)** Median cross-tissue PSI in GTEx v6p for the exon where an individual carries a derived missense variant (red) as a function of predicted pathogenicity measured by CADD score, with derived allele frequency (DAF) matched benign synonymous variants (blue) as a control, and 95% confidence intervals. **F)** Comparison of median exon PSI between missense and DAF matched benign synonymous variants with 95% confidence intervals. All 95% confidence intervals generated with 1000 bootstrap samples, and p-values calculated using paired Wilcoxon signed rank test.

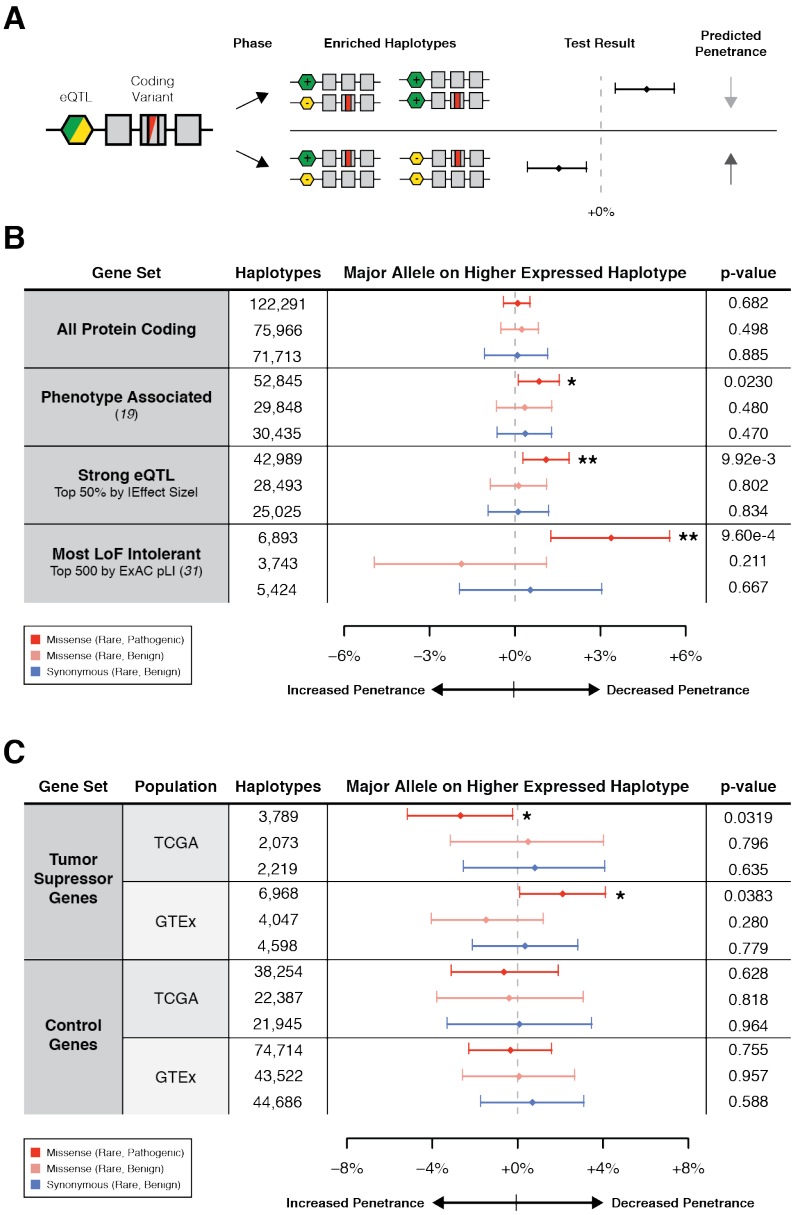

Figure 3. Regulatory haplotype configurations that reduce pathogenic variant penetrance are enriched in the general population, and depleted in individuals with disease. A) Test using phased genetic data (see Fig. S3, Materials and Methods) for haplotype configuration patterns indicating modified penetrance of rare (MAF < 1%) coding variants due to common (MAF > 5%) regulatory variation (GTEx v6p eQTLs). **B)** In GTEx individuals, representative of the general population, natural selection has favored haplotype configurations that reduce potentially pathogenic coding variant penetrance at relevant genes. **C)** Conversely, individuals from TCGA who developed cancer show increased predicted penetrance of potentially pathogenic germline variants in tumor suppressor genes [23]. GTEx haplotypes were generated from 620 population and read-back phased whole genomes. TCGA haplotypes were generated from 925 population and read-back phased whole genomes. Control genes were selected to have within ± 5% the number of coding variants, coding variant frequency, and number of eQTL coding variant haplotypes as tumor suppressor genes, and had a matched number of haplotypes sampled from them. 95% confidence intervals and empirical p-values were generated using 100,000 bootstraps. * $p < 0.05$, ** $p < 0.01$.

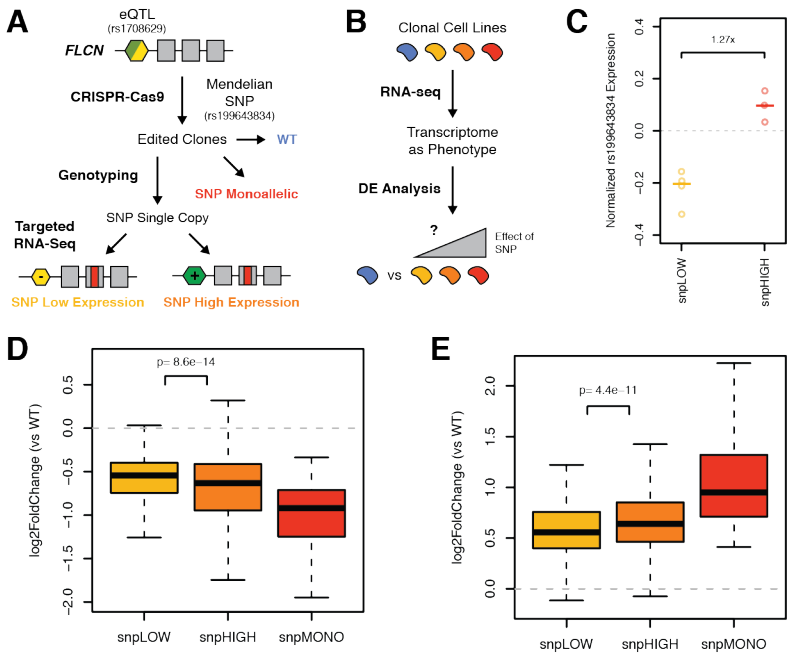

Figure 4. Distinct haplotypes of regulatory and coding variants in *FLCN* created by CRISPR-Cas9 differ in penetrance based on a cellular phenotypic readout. A) Illustration of the experimental study design, where CRISPR-Cas9 was used to edit a Mendelian missense SNP in *FLCN* (rs199643834) that causes Birt-Hogg-Dubé Syndrome into 293T cells that harbor a single copy loss of expression eQTL for the gene (rs1708629). Monoclonal cell lines were produced, genotyped using targeted DNA-seq of the edit site, and classified as monoallelic for the edit SNP (snpMONO), or as having a single copy. Targeted RNA-seq and AE analysis of the edit SNP was performed for single copy clones, allowing the phase of the SNP with respect to the eQTL to be determined. **B)** Using the transcriptome as a phenotype, changes in gene expression compared to wild-type should be stronger in snpHIGH clones versus snpLOW clones if SNP penetrance is modified by the eQTL. **C)** Copy number normalized expression of the edited SNP as measured by targeted RNA-seq (allelic expression, log2(ALT/REF)) in snpLOW (allelic expression < 0, p-value < 0.01, derived from binomial distribution without imbalance) and snpHIGH (allelic expression > 0, p-value < 0.01) clones. **D-E)** Change in expression of genes that were significantly downregulated (D, 277 genes) or upregulated (E, 387 genes) in clones monoallelic for the edited SNP versus wild-type controls. Single copy edit SNP clones are stratified by haplotype configuration. P-values were calculated using a paired Wilcoxon signed rank test.

Supplementary Materials

Materials and Methods

Variant Annotation

Variant annotations for SNPs were retrieved from CADD v1.3 [12]. As per guidelines by the CADD authors, missense variants with a CADD PHRED score of > 15 were defined as potentially pathogenic. This corresponded to approximately the top 60% of missense variants by CADD score in the GTEx haplotype dataset. To define benign synonymous controls, a threshold that included the same proportion of the bottom synonymous variants by CADD score was used. This corresponded to < 10 for the GTEx and TCGA datasets. To be considered rare, variants had to have a MAF < 1% across GTEx v7, 1000 Genomes Phase 3 [30], and gnomAD r2.0.1 [31].

GTEx Allelic Expression Analysis

GTEx v6p allelic expression data generated from whole exome sequencing genotypes were used [14]. Variants that were in low mapability regions (UCSC mapability track < 1), showed mapping bias in simulations [32], had strong allelic expression (>= 99% of reads from one allele), or had less than 8 reads were excluded [15]. To minimize the probability that the observed allelic imbalance was due to effects of the AE variants themselves on splicing, only variants farther than 10 bp from an annotated splice site [12] were used. To collapse AE data for a given variant across individuals and tissues, first, cross-tissue AE was calculated for each individual by summing reference and alternative allele reads across tissues, and then the median reference ratio was calculated across individuals. For each variant, the derived ratio was reported by defining the evolutionary derived allele using the CADD annotation [12]. For each rare (DAF < 1%) potentially pathogenic missense variant a matched rare benign synonymous variant was randomly selected with replacement controlling for DAF within 25%.

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 percent spliced in (PSI) of each exon in each individual [16]. For each exon in each tissue for all exons with at least 10 PSI measurements, the cross-sample PSI median absolute deviation was calculated, and the deviation from the median of each sample was calculated to produce a normalized PSI value (Figure S2c). Within a tissue, only exons where the median absolute deviation was greater than 0 were used, corresponding to approximately 30% of exons with PSI data (Figure S2d-e). For each coding variant, the median cross-tissue normalized PSI of the exon was calculated per individual, and then the median cross-tissue value was calculated across individuals to produce a single PSI measure for each coding variant. To minimize the probability that any observed PSI changes were due to effects of the coding variants themselves on splicing, only variants farther than 10 bp from an exon start or stop site were used. For each rare (DAF < 1%) potentially pathogenic missense variant a matched rare benign synonymous variant was randomly selected with replacement controlling for DAF within 25%.

GTEx Expression Quantitative Trait Loci (eQTL)

The official set of GTEx v6p top significant (FDR < 5%) eQTLs by permutation p-value were used for all analyses such that each gene by tissue had at most a single eQTL [14]. Those eQTLs where the 95% confidence interval of eQTL effect size overlapped 0, representing weak eQTLs, were discarded [33]. 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 (Table S1).

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 [17]. Following this, phASER v1.0.0 was used to perform read-backed phasing using RNA-seq reads [18] from all samples for each individual, which was a median of 17 tissues, and ranged 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.

TCGA – Paired tumor and normal WGS reads from 925 individuals were used to call germline and somatic variants with Bambino v1.06 [34]. The resulting germline genotypes were population phased with EAGLE2 v2.3 [35] using the 1000 Genomes Phase 3 panel [30] and read-back phased with phASER v1.0.0 [18]. For read-backed phasing, only reads with MAPQ ≥ 30 and with a base quality of ≥ 10 overlapping heterozygous sites were used, and all other phASER settings were left as default. For eQTL genotypes only, the resulting phased genotypes were imputed into 1000 Genomes Phase 3 [30] with Minimac3 v2.0.1 [36].

Test for Regulatory Modifiers of Penetrance Using Phased Genetic Data

Here we test the hypothesis that in loss-of-function coding variant heterozygotes, increased expression of the major, or "wild type" coding allele mediated by an eQTL can reduce the penetrance of the mutant allele by increasing the dosage of functional gene transcript, and vice-versa (Fig. S1). 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 that the major allele is observed on the higher expressed eQTL haplotype (Fig. S3a). 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, the frequency of observed major alleles on higher expressed haplotypes would simply be equal to the frequency of the higher expressed eQTL allele in the population. Alternatively, an increased frequency indicates an enrichment of haplotype configurations that decrease coding variant penetrance in the population studied, and vice-versa (Fig. S3b). 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 higher expressed haplotype (Fig. S3a). Alongside this binary measure the frequency of the higher 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 alleles, respectively, and B and b denote the major and minor coding variant alleles, 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 higher 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:

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

Where $f(A)$ is the population frequency of the higher expressed eQTL allele included in the tested haplotype g.

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

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

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

Confidence intervals 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 1000 individuals at 500 genes with 1000 replicates. The higher 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 higher expressed haplotypes and strengths of this bias. For the test, 1000 bootstrap samples were used. We found that at 5% significance threshold, 5% of simulation replicates were significant, suggesting that the test is well calibrated under the null. For real world data, reported in the study, we used 100,000 bootstrap samples to calculate p-values and derive confidence intervals.

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 confidence intervals and accounting for differences in sample size between control genes and genes of interest. We compared p-values derived from our test to those derived from a Poisson-Binomial distribution with parameters $\mathrm{E}[\beta(g_1)] \ldots \mathrm{E}[\beta(g_N)]$. In practice, our p-

values are very similar to that generated using the Poisson-Binomial distribution (Pearson correlation = 0.996, slope = 0.997, Fig. S3e).

Gene Sets

A list of phenotype associated genes was produced by downloading all gene to phenotype associations from DisGeNET [19] v5.0 on 06/08/17, and selecting genes with at least 2 phenotypes associated to them. Genes with strong eQTLs were selected as the top 50% of eGenes by absolute eQTL effect size [33]. Extremely loss-of-function intolerant genes were selected as the top 500 by ExAC pLI [31]. A list of 983 down-regulated tumor suppressor genes in tumor samples versus normal tissue in TCGA expression data was downloaded from the Tumor Suppressor Gene Database [23] website (https://bioinfo.uth.edu/TSGene/) on 08/24/17.

CRISPR/Cas9 Guide Selection and Cloning

Prior to 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 the above regions using forward and reverse genotyping primers FLCN_genot and FLCNeQTL_genot, respectively (Table S5). Amplicons were sequenced by both Sanger sequencing and on the Illumina MiSeq. The 293T cell genotype was Ref/Ref at rs199643834 and Ref/Alt at rs1708629. There were no single nucleotide changes close to rs199643834 that may affect sgRNA activity or require modified homologous template.

Using computational algorithms with prioritization for on-target efficiency and reduced off-target effects (available online: CRISPR Design tool (crispr.mit.edu) and E-CRISPR [37] 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 (Table S5). 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 [38]. Plasmids were transformed into chemically competent E. coli (One Shot Stbl3 Chemically Competent E. coli, ThermoFisher Scientific, cat#: C737303), 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 IDT DNA in the form of lyophilized ultramer (Table S5).

Transfections and T7 Endonuclease I(T7E1) Assays

Human 293T cell line (ATCC, cat. # 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 6-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 (ThermoFisher Scientific, cat. # L3000008). At 24-hours post-transfection selective pressure in the form of 5 µg/ml 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 two 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 base pair region flanking rs199643834 was PCR-amplified from the crude extracts using FLCN_genot primers and purified using Ampure XP beads (Beckman-Coulter, part #: A63880). Purified products were heteroduplexed, digested with T7 endonuclease 1 (NEB, cat # M0302L), and run on a 2% agarose gel. Cleavage patterns from editing experiments conducted with each sgRNA were qualitatively analyzed to determine each Cas9/sgRNA cutting efficiency 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 were selected for single-cell sorting in 96-well format on SONY SH800 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 1920 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, 4 wild-type (Ref/Ref), 7 heterozygous (Ref/Alt) and 2 homozygous mutant (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 (Table S5). Targeted amplicons were dual-indexed using custom Nextera indexes and sequenced on the Illumina MiSeq with 2x150 bp reads.

For all the 13 lines the genotype determined by DNA-sequencing was confirmed by RNA-seq reads. For the 7 lines with a single copy of the edited SNP, we performed allelic expression analysis. Reads were aligned to hg19 using STAR [39]. The number of reads mapping to the reference and alternative alleles was quantified using allelecounter requiring MAPQ = 255 and BASEQ ≥ 10 [15]. Across samples, there was a median of 34,870 reads passing filters overlapping the site. A binomial test using reads containing the edit SNP allele against a null of 1/3 (corresponding to a single copy of the edit SNP) was performed. Copy number normalized allelic expression of the edit SNP was calculated as log2((ALT_COUNT/REF_COUNT)/(1/3)). Samples with allelic expression < 0 and binomial $p < 0.01$ were categorized as snpLOW (edit SNP on lower expressed eQTL haplotype), and those with allelic expression > 0 and binomial $p < 0.01$ were categorized as snpHIGH (edit SNP on higher expressed eQTL haplotype).

RNA-seq and Gene Expression Analysis of Edited 293T Cells

RNA sequencing libraries were prepared using the TruSeq Stranded mRNA Library Sample Preparation Kit in accordance with manufacturer's instructions. Briefly, 500ng 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 10 cycles).

Final libraries were evaluated using fluorescent-based assays including PicoGreen (Life Technologies) and Fragment Analyzer (Advanced Analytics), and were sequenced on the Illumina NovaSeq Sequencing System using 2 x 100bp cycles to a median depth of 52.8 million reads. Trimmomatic [40] v0.36 was used to clip Illumina adaptors and quality trim, and reads were aligned to hg19 using STAR [39] in 2 pass mode. A median of 98% of reads mapped to the human genome, with a median of 95.2% reads mapping uniquely. featureCounts [41] 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 [42] v1.16.1 and R v3.4.0 on genes with a mean of greater than 5 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 [28] with Wikipathway enrichment categories.

Data Use, Availability, and Accessions

GTEx v6p eQTLs are publically available through the GTEx Portal (https://gtexportal.org/). GTEx genotype data, AE data, and RNA-seq reads are available to authorized users through dbGaP (study accession phs000424.v6.p1, phs000424.v7.p2). TCGA data is available to authorized users through dbGap (study accession phs000178.v9.p8). HEK293T RNA-seq data generated in this study is available on the SRA under accession TBD.

Supplementary Tables

Table S1. Top cross tissue GTEx v6p eQTLs per gene.

Table S2. TCGA individuals and respective cancer types used for analysis.

Table S3. Differentially expressed genes in CRISPR-Cas9 edited rs199643834 monoallelic versus wildtype 293T cells.

Table S4. Pathway based gene set enrichment analysis of rs199643834 differential expression data.

Table S5. Oligonucleotides used in this study.

Supplementary Figures

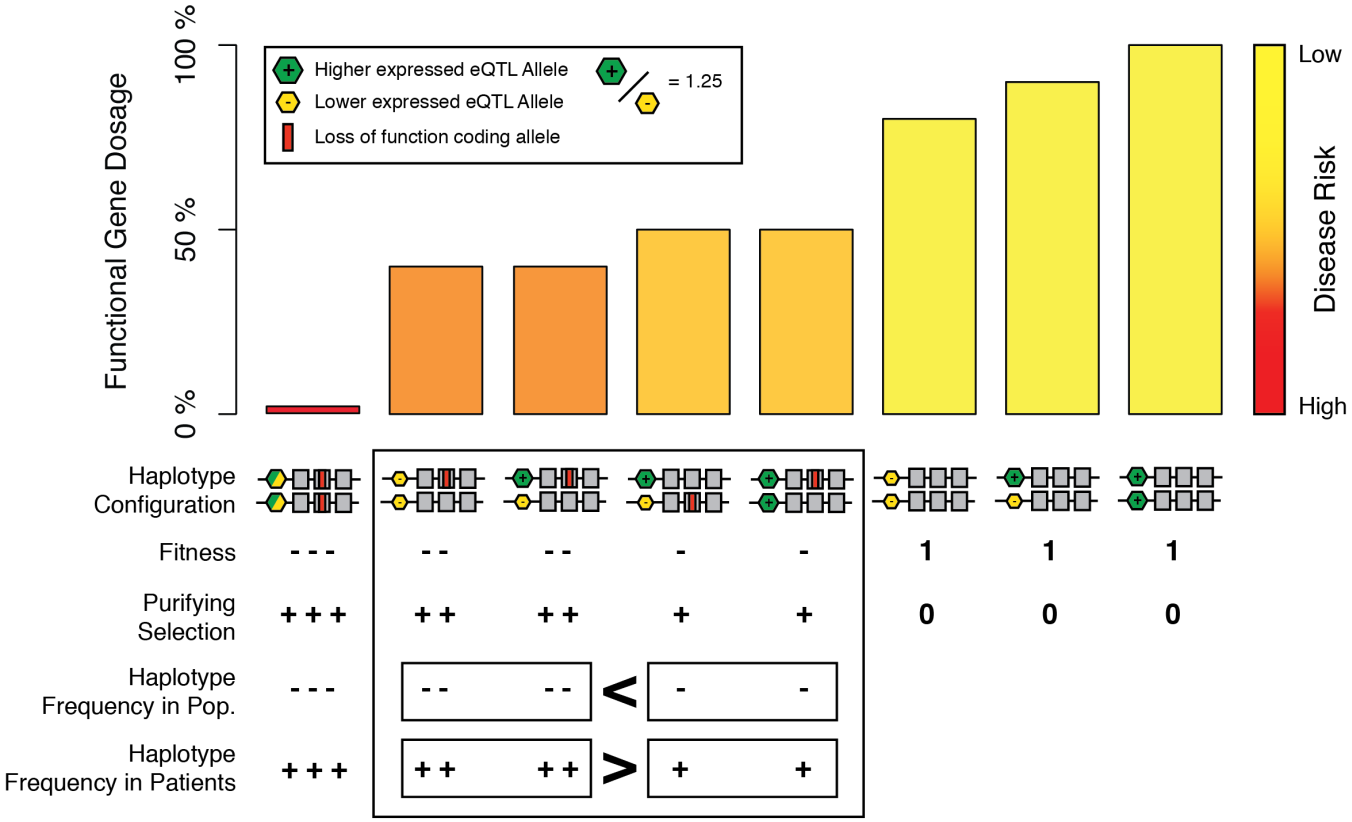

Figure S1. Illustration of the key features of our model of joint effects of regulatory and coding variants on functional gene dosage and selection. Under the model, regulatory variation altering functional gene dosage is particularly important in loss-of-function heterozygotes, where the dosage of functional protein is already reduced to half. Our general assumption is that common regulatory variants typically have such low effects on gene dosage that in the absence of coding variants, they do not cause severe disease or substantial reduction of fitness. Accordingly, in this example, under an additive model of gene expression, the higher expressed eQTL allele increases expression by 1.25x, and disease risk increases non-linearly with decreasing gene dosage, there are potentially large disease risk differences for loss-of-function heterozygotes depending on eQTL haplotype configuration. This results in purifying selection acting more strongly against haplotype configurations that decrease functional gene dosage, while acting more weakly on those that increase functional gene dosage. At the population level this differential strength of purifying selection would result in haplotype configurations that increase functional gene dosage being present at higher frequencies than those that decrease dosage. We note that while we believe that this general model is plausible for many genes with dosage-sensitivity, other scenarios are likely to exist, and for example, fully recessive genes or gain-of-function coding variants would not follow this model. Future work and larger data sets are needed to elucidate the full picture of relative importance of different types of joint effects of regulatory and coding variants.

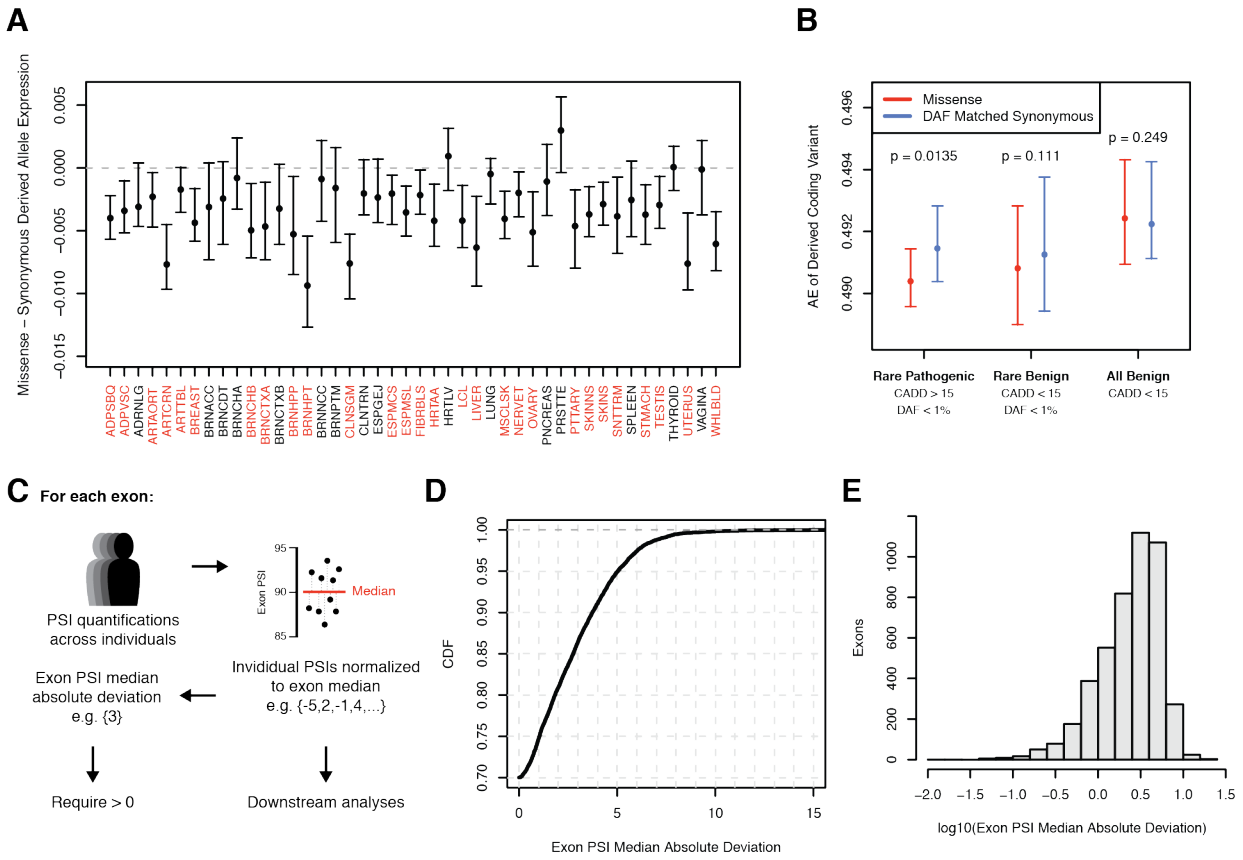

Figure S2. Using GTEx allelic expression (AE) and percent spliced in (PSI) to estimate the penetrance of coding variants at the individual level. A) Difference in allelic expression between rare potentially pathogenic missense variant and DAF matched benign synonymous variants across GTEx tissues. A negative difference indicates reduced expression of missense variants compared to synonymous controls. Bars show the 95% confidence interval of the difference calculated using a paired Wilcoxon signed rank test, and tissues labeled in red have a significant difference (FDR < 10%). B) Comparison of median AE between missense and DAF matched benign synonymous variants in exons where inclusion in that individual was 100% (PSI = 100), with p-values generated using a paired Wilcoxon signed rank test and 95% confidence intervals of AE generated by 1000 bootstraps. This indicates that reduction in allelic expression of potentially pathogenic coding variants occurs through regulatory variation affecting expression level, likely in addition to variation affecting splicing. C) Workflow for generating quantifications of individual level exon inclusion (see Materials and Methods). D) CDF function of calculated cross-tissue exon PSI median absolute deviation, which illustrates that 30% of exons show robust variation in PSI across individuals. E) Histogram of cross-tissue exon PSI median absolute deviation for all exons with non zero median absolute deviation.

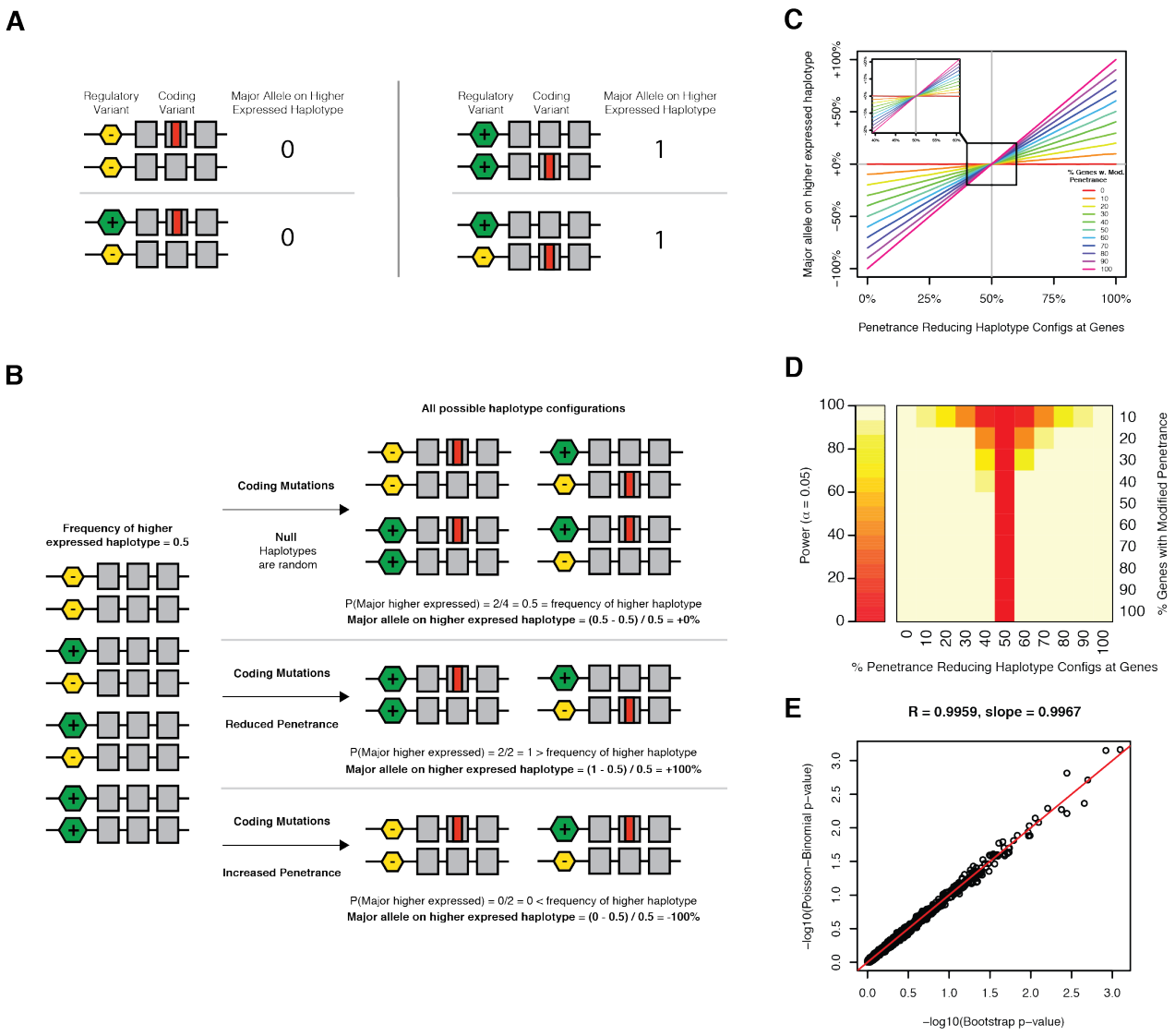

Figure S3. Test for regulatory modifiers of coding variant penetrance using phased genetic data. A) As input the test takes phased genotypes of coding variants and the eQTL for that gene. For each individual heterozygous for a coding variant a binary measure is produced to indicate if the major (wild-type) allele is on the higher expressed eQTL haplotype. **B)** Across a population of individuals, the null expectation is that the observed haplotype configurations are a random sampling of all possible configurations, and thus the proportion of observed major alleles on the higher expressed haplotype is equal to the frequency of the higher expressed haplotype in the population. The diagram depicts a single gene example, but observations are aggregated across genes, and the difference between the observed frequency of major alleles on the higher expressed haplotype and the higher expressed haplotype frequency across those genes is calculated. **C)** Results of test performed on simulated haplotype data from 1000 individuals at 500 genes with 1000 replicates using a higher expressed haplotype frequency of 50% and coding variant frequencies observed in GTEx, across a range of genes exhibiting joint effects between regulatory and coding variants and effect size. The simulated effect size is described by the x-axis in terms of the percentage of observed haplotype configurations that decrease penetrance. **D)** Power to detect significant (α = 0.05) regulatory modifiers of penetrance from simulation data in (C) is robust across a range of effect sizes. **E)** Comparison of p-values calculated using either the bootstrap approach or with the Poisson Binomial distribution from 1000 simulations of 1000 haplotypes generated under the null shows that they are extremely similar. The equality line is shown in red. See “Materials and Methods – Test for Regulatory Modifiers of Penetrance Using Phased Genetic Data” for more information.

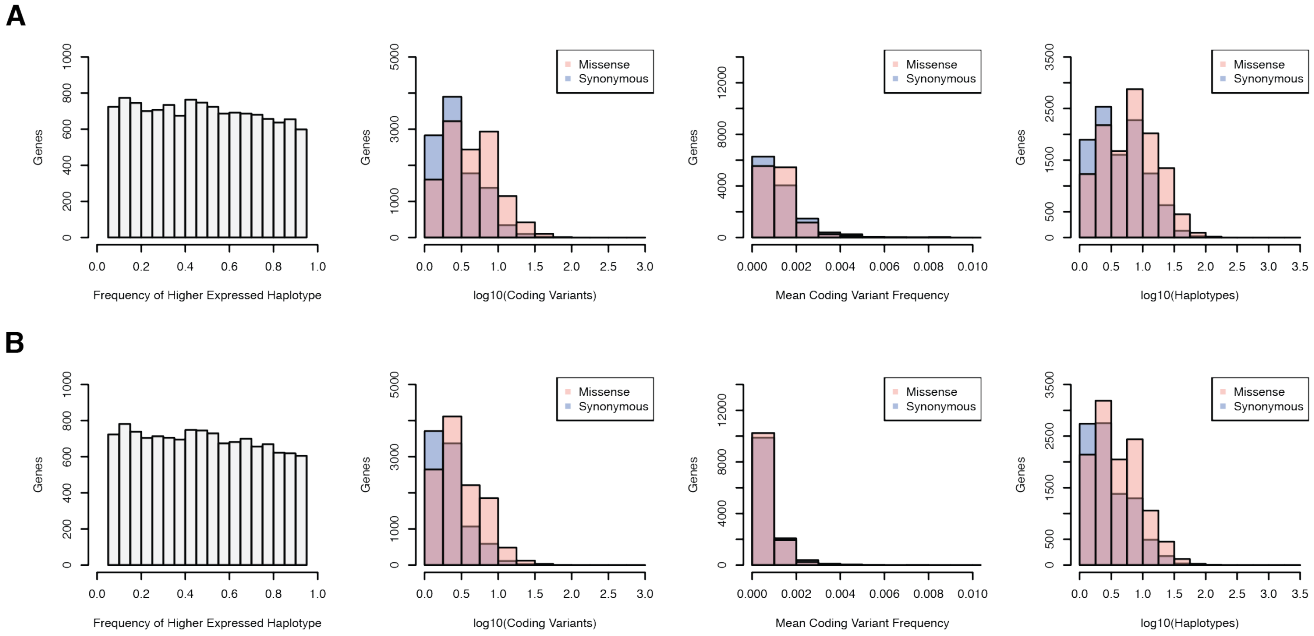

Figure S4. Gene level metrics of common (MAF > 5%) regulatory and rare (MAF < 1%) coding variant haplotypes. Haplotypes generated using potentially pathogenic missense (red) or benign synonymous (blue) coding variants and the top cross-tissue GTEx v6p eQTLs to define higher and lower expressed haplotypes. Histograms of higher expressed haplotype frequencies, number of coding variants with haplotype data, mean coding variant frequency, and number of haplotypes observed at the gene level for haplotypes from 620 phased GTEx v7 whole genomes (A) and 925 phased TCGA germline whole genomes (B).

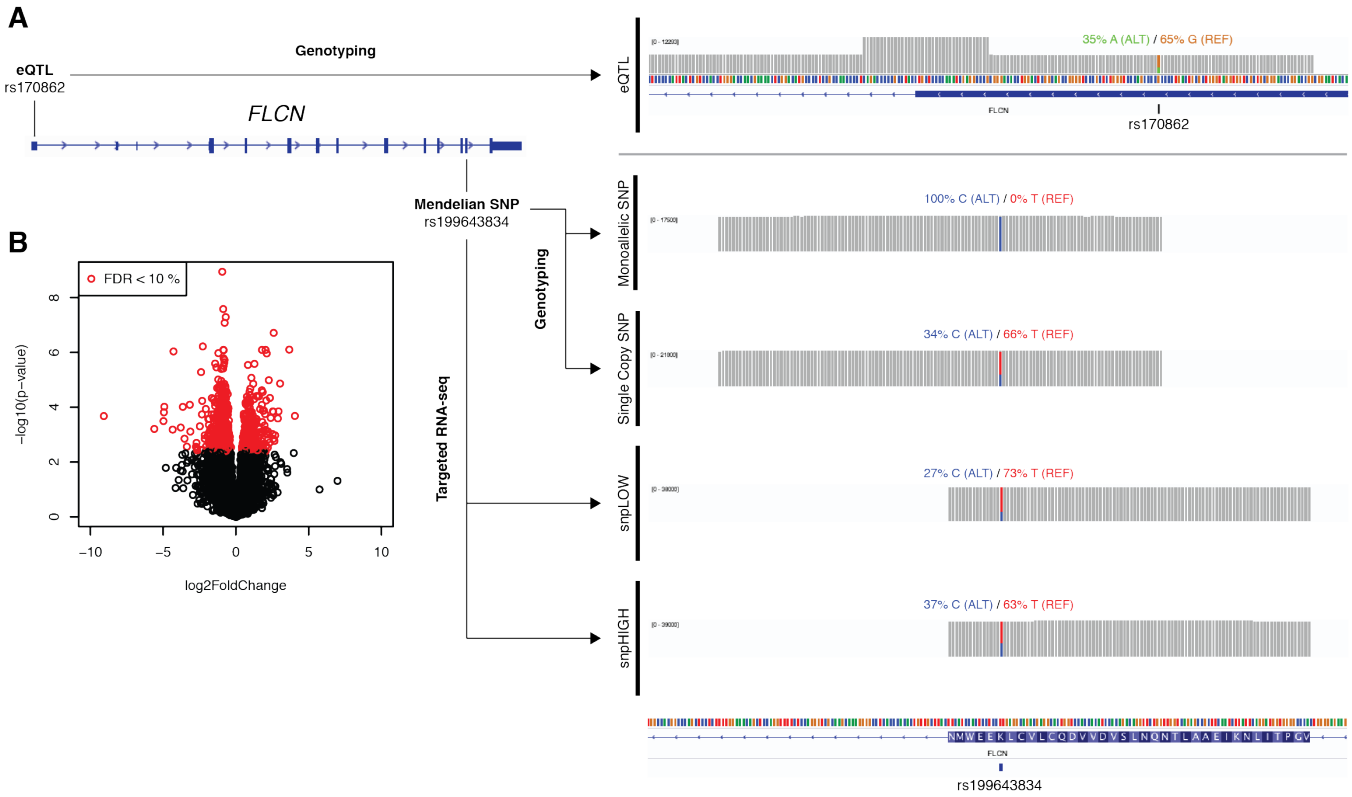

Figure S5. Expression haplotype aware editing of a Mendelian SNP in 293T cells using the transcriptome as a phenotypic readout. A) Visualization of eQTL SNP (rs170862) genotyping reads, edited SNP (rs199643834) genotyping reads from representative monoallelic and single copy clones, and targeted RNA-seq reads from representative low SNP expression (snpLOW) and high SNP expression (snpHIGH) clones. Ratios of the reference allele and alternative allele in targeted DNA and RNA sequencing are indicated. A full list of primers used for sequencing can be found in Table S5. **B)** Volcano plot of differential expression analysis comparing two clones monoallelic for the edit SNP versus four wildtype clones.