

# Using integrative estimates of deleteriousness and regulatory potential to identify causal eQTLs



UW Medicine  
SCHOOL OF MEDICINE

Ronald J. Hause Jr.<sup>1</sup>, Benjamin Weaver<sup>2,3</sup>, Martin Kircher<sup>1</sup>, Jay A. Shendure<sup>1</sup>, and Gregory Cooper<sup>2</sup>  
<sup>1</sup>Department of Genome Sciences, The University of Washington, Seattle, WA, 98195, USA, <sup>2</sup>HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, USA, <sup>3</sup>Department of Genetics, University of Alabama-Birmingham, 35233, USA.



## Introduction

Expression quantitative trait locus (eQTL) mapping has identified DNA variation associated with differences in mRNA expression and has previously been used as an effective strategy for prioritizing genetic variation that may play a role in complex phenotypes. However, current eQTL analyses are limited in their ability to identify causal variants, because most single nucleotide variants (SNVs) identified as eQTLs are likely tag SNVs in linkage disequilibrium (LD) with the true functional variant altering gene expression. Furthermore, eQTLs can be context specific, and understanding the determinants distinguishing “global” from tissue-specific eQTLs is critical to better interpret GWAS results in their most relevant contexts. Recent studies have demonstrated that incorporating regulatory annotations can improve power to identify putatively functional candidate variants and reproducible eQTLs across studies and between tissues<sup>1,2</sup>. We have recently developed a framework, Combined Annotation Dependent Depletion<sup>3</sup>, that integrates 63 different genomic annotations into a single C-score predicting the deleteriousness of any particular SNV. These annotations include metrics of chromatin accessibility, conservation scores, and gene-model-based information, among others. C-scores are associated with annotations of functionality, pathogenicity, and experimentally measured regulatory effects. Despite the substantial improvements made to prioritize eQTL discovery and effectively model replication across studies and tissues, many strong eQTLs still fail to replicate in independent cohorts. Therefore, we hypothesized that prioritizing variants on our composite metric of genome functionality could help to further enrich for causal variants involved in gene regulation.

## Project Outline

The goals of this project were to (1) further examine metrics related to eQTL reproducibility across studies and tissues and (2) to assess whether C-scores can improve power to detect reproducible and/or tissue-specific eQTLs along with cell type specific cis-regulatory element information from ENCODE. To achieve this goal, we systematically analyzed six datasets where eQTLs have previously been performed:

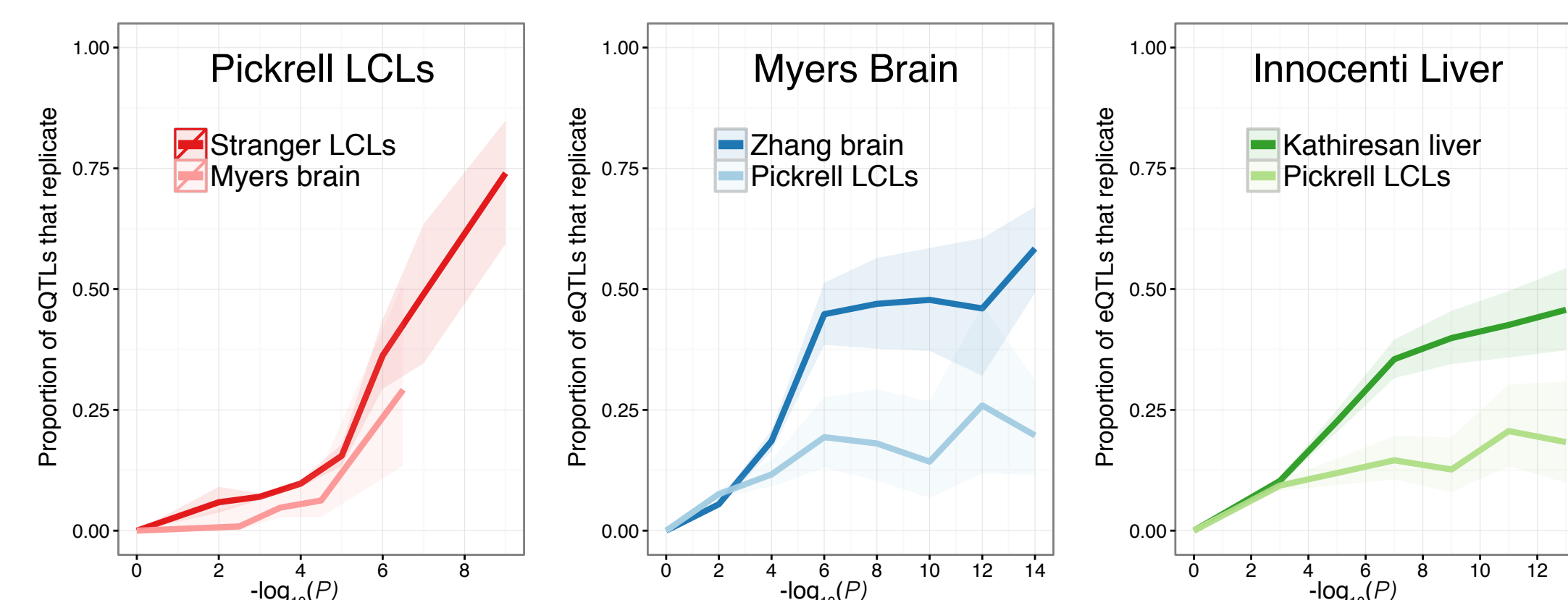
Study	Tissue	Samples	Genes	Platform
Pickrell <sup>4</sup>	LCL	68	20861	RNA-Seq
Stranger <sup>5</sup>	LCL	60	15757	Illumina WG-6 v1
Myers <sup>6</sup>	Brain	193	11707	Sentrix HumanRef-8
Zhang <sup>7</sup>	Brain	153	17377	Affy HG-U96 v2
Innocenti <sup>8</sup>	Liver	206	16236	Agilent 4x44K
Kathiresan <sup>9</sup>	Liver	60	10877	Illumina HumanRef-8 v2

### Data:

- DNase hypersensitivity sites from GM12878, SK-N-SH, and HepG2 cell lines
- TFBS clusters (V3) from ENCODE data uniformly processed by ENCODE Analysis Working Group
- GWAS SNPs from the NHGRI and curated variants from the NIH ClinVar database

### Methods:

- Expression and genotype data were QCed and SNPs imputed against 1000 Genomes (1kG) variants
- Association mapping was performed by regressing surrogate variable-corrected, inverse normal-transformed gene-level expression measurements against minor allele frequency (MAF) copy number for each individual for all SNPs within 1Mb of the transcription start site (TSS)
- Pickrell, Myers, and UC studies were used as eQTL discovery cohorts, and the Zhang, UC, and UW studies were used as the corresponding eQTL replication cohorts
- eQTL replication is here defined as concordant effect direction and nominal significance for the subset of genes and common SNPs overlapping between studies and tissues
- Fisher's exact tests were used for enrichment analyses and wilcoxon rank sum tests for differences between CADD score distributions

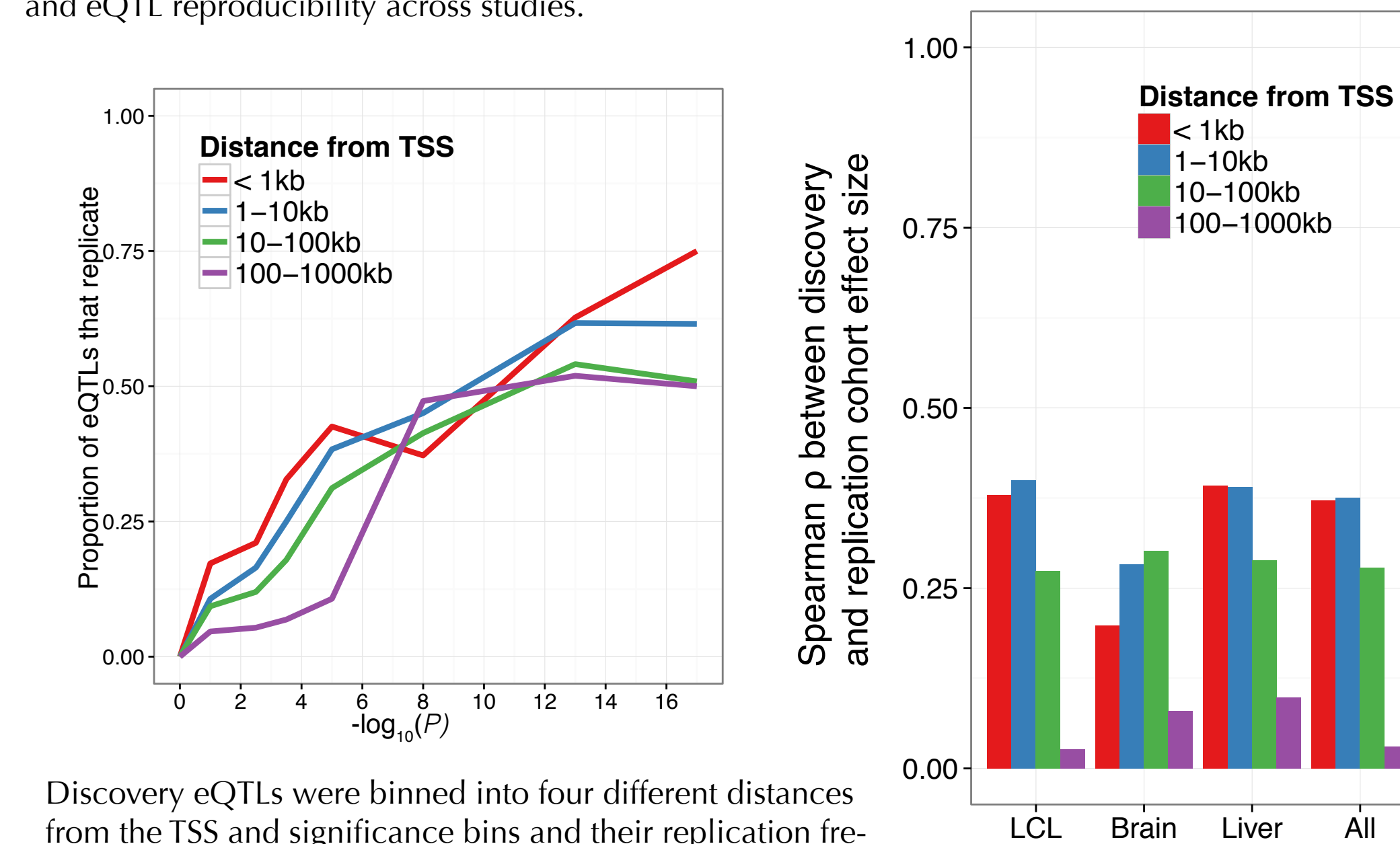


cis-eQTL replication frequency (y-axis) as a function of discovery significance (x-axis) for Pickrell LCL eQTLs in Stranger LCLs (dark red) or Myers brain (light red) (left); Myers brain eQTLs in Zhang brain (dark blue) or Pickrell LCLs (light blue) (middle); and Innocenti liver eQTLs in Kathiresan liver (dark green) or Pickrell LCLs (light green) (right). Ribbons depict 95% confidence intervals.

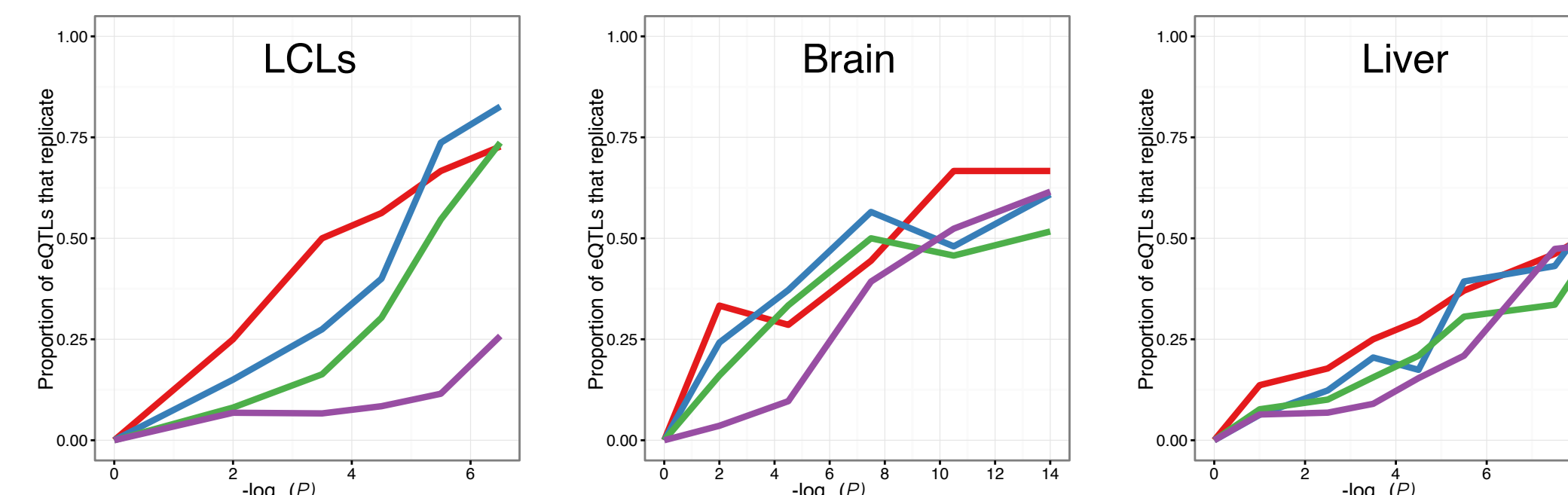
Consistent with previous observations, cis-eQTLs are more than two times as likely to replicate across studies within the same cell type than they are to replicate between different cell types.

## eQTL reproducibility

There is a well-established relationship between eQTL effect size and proximity to the TSS of the associated gene. We here demonstrate that a relationship also exists between eQTL distance to gene TSS and eQTL reproducibility across studies.

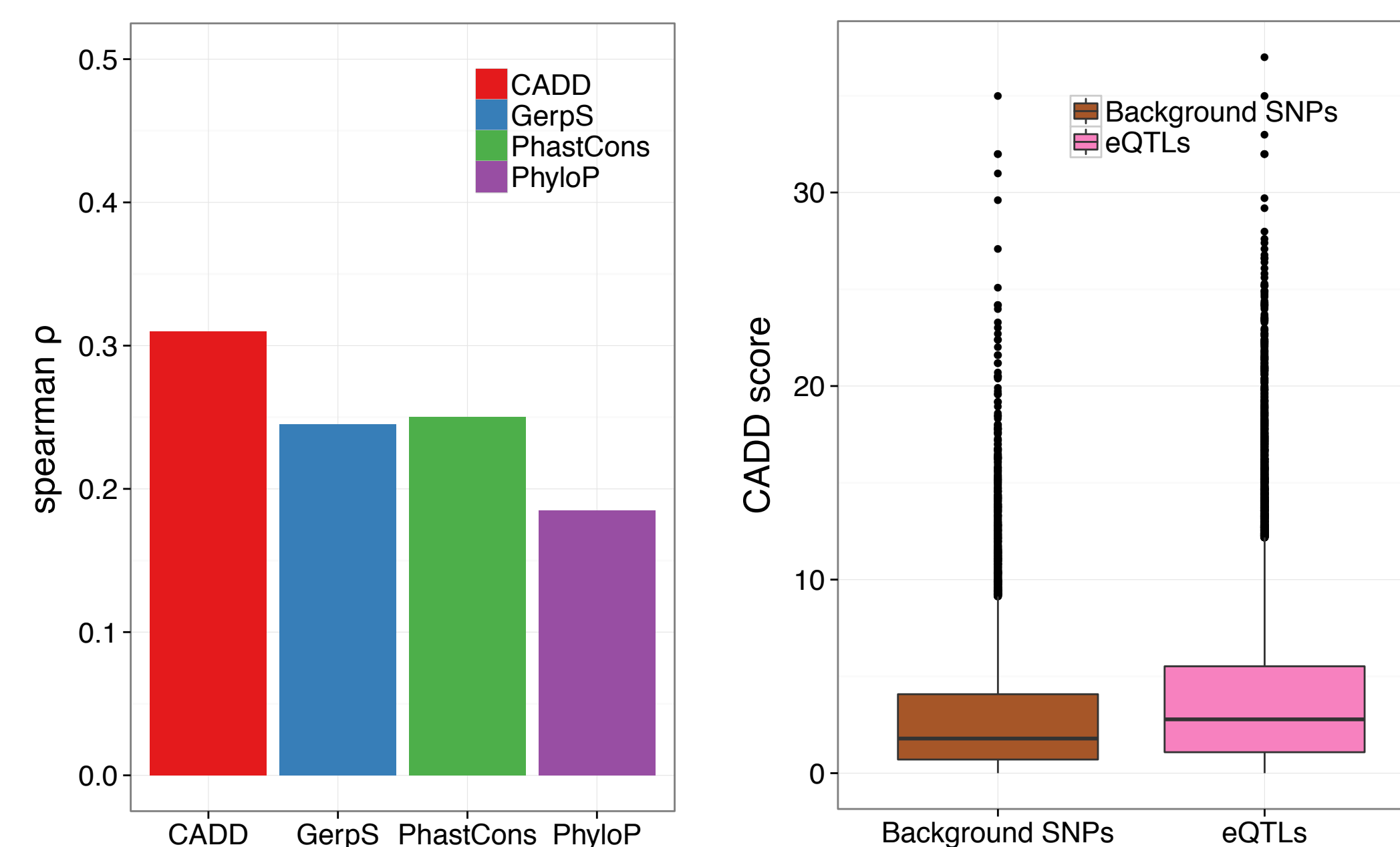


Discovery eQTLs were binned into four different distances from the TSS and significance bins and their replication frequencies plotted for all (upper left), LCL (lower left), brain (lower middle), and liver eQTLs (lower right).

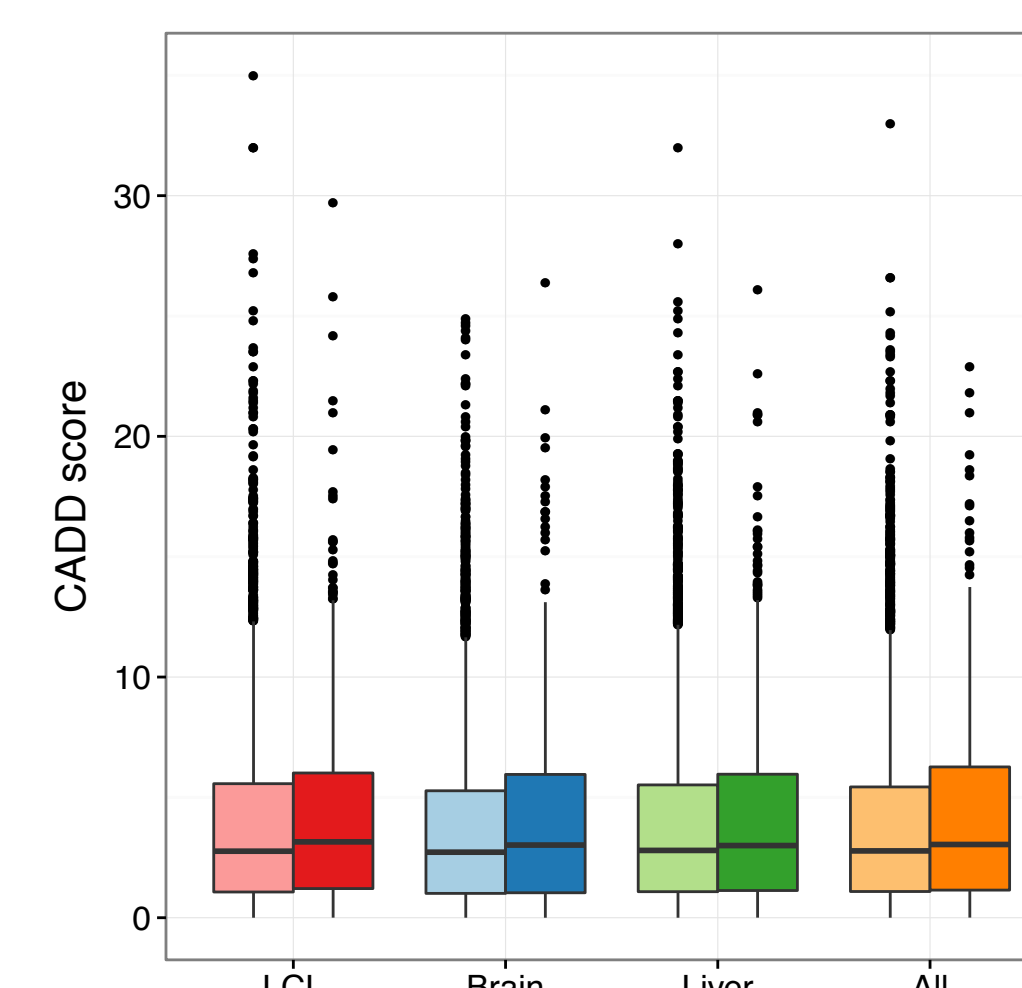


We also observed that the correlation between effect sizes of reproducible SNPs across studies was higher as a function of distance to TSS (upper right). We appreciate that the elevated replication rate in LCLs is because the Pickrell and Stranger studies are measuring the expression levels of the same individuals, providing a metric for cis-eQTL reproducibility across labs and platforms. The reduced replication rates for liver eQTLs are likely because of the smaller sample size (and thereby lower power) of the UW replication cohort.

## C-scores and eQTLs



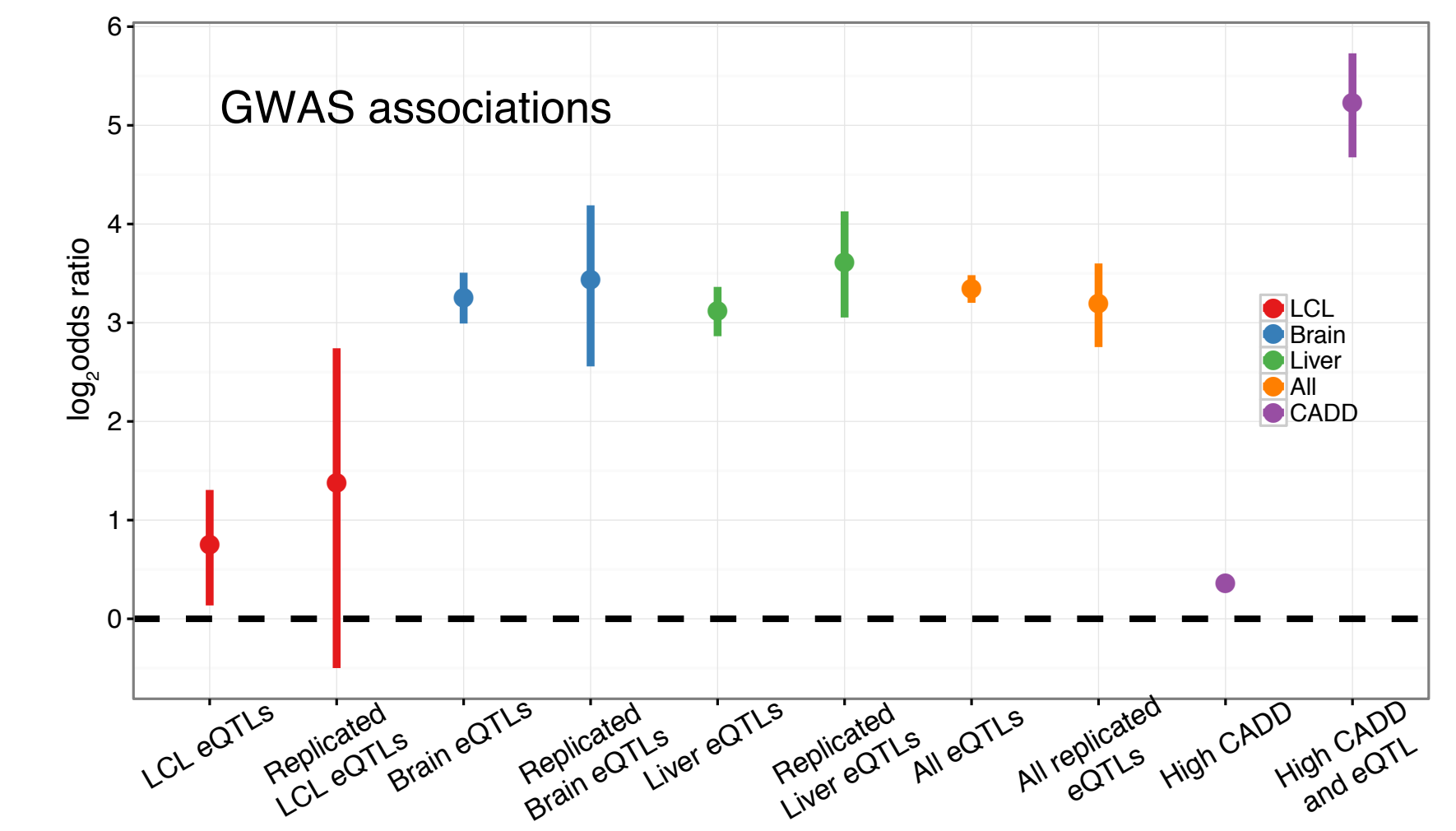
Previously, our lab has demonstrated that CADD scores are more strongly correlated with experimentally measured effects on transcriptional regulatory function of two enhancers and one promoter as measured by saturation mutagenesis and massively parallel reporter assays<sup>10</sup> than other scores (upper left), indicating their relevance in potentially predicting genome functionality. This functional relationship was also observed for common eQTLs, with C-scores for eQTLs in LCLs are approximately 1.6 times higher than randomly selected, MAF- and location-matched SNVs from 1000 Genomes ( $P < 2.2 \times 10^{-16}$ ) (upper right).



We also observed a significant inflation of C-scores for reproducible eQTLs relative to eQTLs that did not reproduce across studies or tissues. This inflation yielded an overall 1.3% increase in eQTL reproducibility, a ~7% increase for eQTLs overlapping DNase sites or TFBS for the relevant cell line, and a 16.3% increase in reproducibility for eQTLs within 1-kb of the TSS.

## eQTLs and CADD in GWAS and ClinVar

Although the effect sizes were modest, CADD scores were significantly associated with eQTLs. Therefore, we compared their relationships to variants implicated in human phenotypic variation. We first quantified the relative enrichments of eQTLs and variants with high CADD scores (C-score > 10) to be either GWAS associations or ClinVar variants.



Notably, we observed stronger enrichments for eQTLs across all tissues to be GWAS hits than high CADD variants (left). Variants that were both eQTLs and predicted to be deleterious by CADD were dramatically enriched in GWAS hits ( $P = 1.42 \times 10^{-38}$ ,  $OR = 37.47$ ). Conversely, eQTLs were depleted in ClinVar “pathogenic” variants, whereas CADD-predicted deleterious variant were highly enriched (right,  $P < 10^{-300}$ ,  $OR = 8.35$ ).

PUBMED ID	Disease/Trait	Gene	SNPs	Context	CADD score	eQTL tissue	eQTL Pval	Rep?	Dnase LCL	Dnase Brain	Dnase Liver	TFBS LCL	TFBS Brain	TFBS Liver
22010049	Gamma glutamyl transferase levels	EPHA2	rs1497406	Intergenic	15.22	Liver	1.99E-10	1	0	0	0	0	0	0
19451621	Amyotrophic lateral sclerosis	SELL	rs3177980	missense	14.06	Brain	7.12E-04	1	0	0	0	0	0	0
21833088	Multiple sclerosis	AGAP2	rs12368653	intron	12.38	Brain	2.30E-05	0	0	0	0	0	0	3
23251661	Obesity-related traits	CCDC41	rs1520724	Intergenic	16.86	Liver	1.36E-04	1	1	0	0	0	0	0
19011631	Colorectal cancer	RHPN2	rs10411210	intron	10.37	Liver	5.16E-04	0	0	0	0	0	0	0
21378988	Coronary heart disease	TFPI	rs7586970	missense	14.64	Liver	2.20E-16	1	0	0	0	0	0	0
23770605	Chronic lymphocytic leukemia	FAKBP2	rs757978	missense	21	Brain	2.78E-15	1	0	0	0	0	0	0
21659334	Response to hepatitis C treatment	DDRK1	rs11697186	intron	10.68	LCL	2.00E-04	0	1	0	0	0	0	4
20173735	Chronic Hepatitis C infection	ITPA	rs2720101	intron	13.41	Liver	1.36E-12	1	0	0	0	1	0	0
21901158	Insulin resistance/response	CACNA1D	rs1401492	intron	12.5	LCL	4.26E-05	1	0	0	0	0	0	0
20881960	Height	SLBP	rs2247341	cds-synon	10.1	Brain	8.53E-04	1	0	0	0	0	0	0
23128233	Inflammatory bowel disease	RGS14	rs12654812	intron	10.5	Liver	3.00E-06	1	1	0	0	0	0	3

Incorporating C-scores along with metrics of eQTL reproducibility and regulatory element information will allow us to prioritize likely functional haplotypes (and eventually mutations) underlying complex phenotypes that be experimentally validated (see above table). For example, rs1497406, underlying an association with gamma glutamyl transferase levels, an enzyme predominantly expressed in liver, sits in a region where 24 TFs bind and therefore is not only a reproducible liver eQTL, but is also weighted highly by CADD to be deleterious.

## Conclusions

- In summary, we observed that eQTLs that reproduce across studies and tissues are enriched to lie close to the TSS, overlap with DNase sites and TFBS, and have overall higher integrative measures of deleteriousness as predicted by CADD.
- Additionally, prioritizing variants that are both eQTLs and predicted to be deleterious by CADD dramatically increased the odds of being GWAS-associated variant by four-fold relative to using eQTLs alone.
- After conditioning on distance to TSS or overlap with cis-regulatory elements, C-score improves eQTL reproducibility, indicating that the additional components incorporated into CADD can assist researchers in further informing not only likely eQTLs, but reproducible eQTLs that can be further dissected into their underlying eQTNs.
- Our approach will be useful for not only prioritizing the most likely causal variants underlying eQTLs, but also for better elucidating the functional mechanisms underlying GWAS results and eQTLs in a variety of contexts.

## Future Directions

- Incorporate C-score as a prior to inform posterior eQTN probabilities in a Bayesian framework
- Integrate median CADD-scores across LD-blocks associated with eQTLs
- Functionally validate predicted causal eQTNs using CRISPR/Cas9 or site-directed mutagenesis
- Jointly estimate global vs. tissue-specific and reproducible eQTLs rather than thresholding
- Create classifier incorporating C-score to predict global and tissue-specific eQTNs
- Extend into other endophenotypes (pQTLs, mQTLs, etc.)

## References and Acknowledgements

- Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* 13, R7 (2012).
- Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet* 9, e1003649 (2013).
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).
- Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772 (2010).
- Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–853 (2007).
- Myers, A. J. et al. A survey of genetic human cortical gene expression. *Nat. Genet.* 39, 1494–1499 (2007).
- Zhang, D. et al. Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *Am. J. Hum. Genet.* 86, 411–419 (2010).
- Innocenti, F. et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7, e1002078 (2011).
- Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* 40, 189–197 (2008).
- Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30, 265–270 (2012).

We would like to thank Darren Cusanovich, Aaron Mckenna, and Evan Boyle for helpful discussions. Our work was supported by grant DP1HG007811 from the NHGRI (J.S.).