Interactions between genetic variation and cellular environment in skeletal muscle gene expression

D. Leland Taylor[1,2], David A. Knowles[3], Laura J. Scott[4], Andrea H. Ramirez[5], Franceso Paolo Casale[2], Brooke N. Wolford[6], Li Guan[6], Arushi Varshney[7], Ricardo D'Oliveira Albanus[6], Stephen C.J. Parker[6,7], Narisu Narisu[1], Peter S. Chines[1], Michael R. Erdos[1], Ryan P. Welch[4], Leena Kinnunen[14], Jouko Saramies[8], Jouko Sundvall[14], Timo A. Lakka[9,10,11], Markku Laakso[12,13], Jaakko Tuomilehto[14,15,16,17], Heikki A. Koistinen[14,18,19], Oliver Stegle[2], Michael Boehnke[4], Ewan Birney[2]*, Francis S. Collins[1]*

1. National Human Genome Research Institute, National Institutes of Health, Bethesda, USA
2. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK
3. Department of Computer Science, Stanford University, Stanford, California, USA
4. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA
5. Department of Internal Medicine, Vanderbilt University, Nashville, Tennessee, USA
6. Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA
7. Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA
8. South Karelia Social and Health Care District, Lappeenranta 53130, Finland.
9. Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio FI-00100, Finland.
10. Kuopio Research Institute of Exercise Medicine, Kuopio FI- 00100, Finland.
11. Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, University of Eastern Finland, Kuopio FI-00100, Finland.
12. Department of Medicine, University of Eastern Finland, Kuopio FI-00100, Finland.
13. Kuopio University Hospital, Kuopio FI-00100, Finland.
14. Department of Public Health Solutions, National Institute for Health and Welfare, P.O. Box 30, Helsinki FI-00271, Finland.
15. Department of Neurosciences and Preventive Medicine, Danube University Krems, Krems 3500, Austria.
16. Diabetes Research Group, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

17. Dasman Diabetes Institute, Dasman 15461, Kuwait.
18. Department of Medicine and Abdominal Center: Endocrinology, University of Helsinki and Helsinki University Central Hospital, P.O. Box 340, Haartmaninkatu 4, Helsinki FI-00029, Finland.
19. Minerva Foundation Institute for Medical Research, Biomedicum 2U, Tukholmankatu 8, Helsinki FI-00290, Finland.

* Correspondence and requests for materials should be addressed to E.B. (birney@ebi.ac.uk) or F.S.C. (email: Francis.Collins3@nih.gov)

Abstract

From whole organisms to individual cells, responses to environmental conditions are influenced by genetic makeup, where the effect of genetic variation on a trait depends on the environmental context. RNA-sequencing quantifies gene expression as a molecular trait, and is capable of capturing both genetic and environmental effects. In this study, we explore opportunities of using allele-specific expression (ASE) to discover *cis* acting genotype-environment interactions (GxE) - genetic effects on gene expression that depend on an environmental condition. Treating 17 common, clinical traits as approximations of the cellular environment of 267 skeletal muscle biopsies, we identify 10 candidate interaction quantitative trait loci (iQTLs) across 6 traits (12 unique gene-environment trait pairs; 10% FDR per trait) including sex, systolic blood pressure, and low-density lipoprotein cholesterol. Although using ASE is in principle a promising approach to detect GxE effects, replication of such signals can be challenging as validation requires harmonization of environmental traits across cohorts and a sufficient sampling of heterozygotes for a transcribed SNP. Comprehensive discovery and replication will require large human transcriptome datasets, or the integration of multiple transcribed SNPs, coupled with standardized clinical phenotyping.

Introduction

A substantial fraction of variability in gene expression is controlled by changes in transcription rates, mainly mediated by transcription factor (TF) proteins binding to specific DNA sequence motifs that define regulatory elements [1,2]. The abundance of such proteins and their regulatory co-factors may in turn be controlled by intrinsic mechanisms inherent to a cell, such as an individual's genetic makeup or regulatory programs specific to a cell type, as well as cellular responses to environmental cues. A regulatory element, defined by the DNA region recognized by a DNA-binding TF and other required transcriptional machinery, may be either intrinsic or environment-dependent. In intrinsic elements, the TF and binding machinery is controlled by cell-intrinsic mechanisms that operate within a closed system and are unresponsive to environment. By contrast, in environment-dependent elements the TF and binding machinery is responsive to an environmental stimulus. Both regulatory element types are susceptible to perturbation by genetic variation because the region recognized by the TF is encoded in the DNA sequence.

Many genetic studies document the effects of genetic perturbations of regulatory elements on gene expression - expression quantitative trait loci (eQTL) [3–6]. Although it is possible to detect *trans* (different physical chromosome) effects, eQTLs are typically identified within a local window, centered on the transcription start site (TSS), and assumed to act via *cis* (on the same physical chromosome) mechanisms. Variation in intrinsic regulatory programs is expected to give rise to such “standard eQTLs”, identified by modeling genetic effects on gene expression. However, it is also likely that variation in environment-dependent elements will be detected in standard eQTL studies, as it is unlikely for a variant to change the relationship between gene expression and environment without altering the mean

gene expression levels for each genotype. Therefore we would expect a subset of eQTLs detected by modeling only genetic effects to also have effects unique to an environmental context. If one were to model the combined environmental and genetic effects on gene expression, such variants would exhibit interaction effects between genotype and environment (GxE) and could be described as GxE interaction quantitative trait loci (abbreviated as iQTLs in this paper), a specific type of eQTL whose effect changes according to an environmental context. To date, the overlap between standard eQTL and iQTL in human is largely unknown, as few studies have co-measured environmental and genetic effects at scale, and the technology for mapping such iQTLs is in its infancy.

In human populations, several GxE signals have been reported across diseases for various quantitative traits (reviewed in [7]), but few have mapped transcriptional iQTLs on a large scale, treating gene expression as a molecular quantitative trait [8–17]. Indeed transcriptional GxE effects have primarily been studied in model organisms where the environment and genotype can be controlled [18–23]. The challenge of mapping iQTLs using transcriptomic data outside of controlled laboratory settings lies in the confounding effects of environmental, biological, and technical factors on gene expression data, and the difficulty in isolating and/or accounting for such effects while preserving effects of the environment of interest.

However, such limitations may be mitigated if a study quantifies gene expression using RNA-seq technology because RNA-seq enables the measurement of allele specific expression (ASE), an alternative readout less prone to the confounders of gene level measurements [10,24]. By quantifying differences in expression between haplotypes in samples heterozygous for a transcribed allele

(abbreviated tSNP in this paper), ASE provides an internally controlled measurement where biological and technical exposures on the cells are essentially identical for both haplotypes. This makes ASE ideal for iQTL mapping since it minimizes batch effects while preserving *cis*-mediated environmental effects.

Furthermore, when integrated with standard gene expression data between individuals (abbreviated to gene-level expression in this paper), the two data types can serve as orthogonal forms of signal to validate iQTLs. In cases of true *cis* regulation of gene expression, when a TF preferentially binds to one allele, we would expect to observe increased ASE in participants heterozygous for the regulatory SNP. As an example, Fig 1 shows the different types of potential regulatory elements and the impact of different polymorphisms in schematic form. At the gene expression level, we would expect an iQTL to have different effects across environmental contexts in a genotype specific manner. In the ASE data, we would expect correlation between ASE and the environment only in individuals heterozygous for both the iQTL-SNP and tSNP. As opposed to standard eQTLs, which can be summarized by box-plots stratified by genotype, we believe a 6-panel regression plot is the most informative, and examples of expected behavior are shown in Fig S1.

In this study, we explore the opportunities and challenges for iQTL mapping and replication using gene-level expression and ASE data. We illustrate our approach using RNA-seq from 267 skeletal muscle biopsies from the Finland-United States Investigation of NIDDM Genetics (FUSION) tissue biopsy study [25], as this dataset features RNA-seq co-measured with rich clinical phenotypes spanning blood metabolites, anthropometric measurements, and medication (S1 Table). Collectively, we treat all clinical phenotypes as “environmental traits” since we model skeletal muscle gene expression and therefore the response of a population of cells to the

surrounding cellular environment - adjacent cells, extracellular matrix, blood plasma, and interstitial fluid - approximated by each phenotype.

As one clear limitation is sample size, we reduce the multiple testing burden by only testing eQTLs for GxE signals, based on the assumption outlined above that at least some of the strongest iQTLs will also show effects on mean gene expression when stratified by genotype and be detected also as eQTLs. With a well-calibrated statistical test, we identify 12 GxE signals that span 10 candidate iQTLs at a trait-specific FDR of 10%. Replication of such findings is challenging because of the lack of human studies on equivalent tissues with equivalent environmental measurements; however, two of the three testable traits shared with the larger GTEx study show non-random aggregate replication, although the need to restrict to heterozygous individuals limits the extent of this replication. This study highlights the utility of ASE based GxE analysis in observational studies, and emphasizes the need for large RNA-seq cohorts with standardized clinical phenotypes to enable study comparison and replication.

Results and Discussion

iQTL Results

As candidate iQTLs for each gene, we tested the most significant skeletal muscle eQTL per gene for the 19,455 autosomal, protein coding genes with at least one significant eQTL from our previous study of 267 Finnish muscle samples [25]. We tested for interaction of these SNP-gene pairs with 17 clinical phenotypes (S1 Table) by jointly modeling the impact of genotype effects on gene level expression and ASE levels (Methods). The resulting p-value distributions are well calibrated (S2

Fig), with the vast majority of tested SNPs consistent with the null distribution. Using a 10% FDR per trait, we identify 10 candidate iQTLs across 6 traits (12 unique gene-environment trait pairs) (Fig 2; Table 1; S2 Table). Of the clinical variables considered, sex is unique in that GxE sex signals could be due to environmental (for example, circulating sex hormones) or intrinsic, within cell, effects due to differences in gene expression from the sex chromosomes.

Table 1. iQTL results FDR 10%.

Clinical Trait	Gene	Chr	tSNP position	iQTL alleles (ref/alt)	iQTL position	p-value ASE	p-value gene	p-value combined	q-value
Age	*PCNT*	21	47786817	G/T	47823229	4.29x10-6	1.25x10-1	8.28x10-6	0.0735
Sex	*BSG*	19	582775	T/C	572878	1.75x10-5	1.00x10-1	2.50x10-5	0.0567
Sex	*NRAP*	10	115412793	C/T	115385650	1.65x10-7	5.61x10-1	1.59x10-6	0.0136
BMI	*DAGLB*	7	6449272	C/T	6476915	3.54x10-2	1.55x10-5	8.48x10-6	0.0753
SBP	*ELP2*	18	33750046	T/G	33743660	3.24x10-5	3.58x10-2	1.70x10-5	0.0607
SBP	*FHOD3*	18	34324091	T/C	33970347	2.82x10-4	5.07x10-3	2.06x10-5	0.0607
SBP	*IGF2R*	6	160453978	T/C	160379096	1.34x10-3	9.18x10-4	1.80x10-5	0.0607
TC, fasting	*AGMAT*	1	15909850	T/C	15918676	2.52x10-3	8.60x10-5	3.54x10-6	0.0315
LDLc, fasting	*AGMAT*	1	15909850	T/C	15918676	1.20x10-3	4.82x10-4	8.88x10-6	0.0501
LDLc, fasting	*DEPTOR*	8	121061879	G/T	120930135	4.43x10-2	1.69x10-5	1.13x10-5	0.0501
LDLc, fasting	*FHOD3*	18	34232657	T/C	33970347	6.78x10-3	4.54x10-4	4.21x10-5	0.0623
LDLc, fasting	*TMEM261*	9	7799653	A/G	7830189	8.31x10-5	1.39x10-2	1.69x10-5	0.0501

Summary of most significant tSNP for each iQTL-gene pair. Coordinates based on GRC37/hg19. The three p-value columns record the ASE, whole gene expression

level, and combined p-value respectively. The combined p-values are used for q-value calculation. Results with all iQTL-tSNP pairs are recorded in S2 Table.

GTEx Replication

We sought to replicate these results using skeletal muscle data from the GTEx study (http://www.gtexportal.org). Shared across studies, four traits were available for this purpose: age, sex, body mass index (BMI), and type 2 diabetes (T2D) status. Three of these variables: sex, BMI, and T2D status, had similar distributions in the GTEx and FUSION cohorts (S1 Table).

Despite significant differences in cohort populations, laboratory techniques, and analysis pipelines, we observe a trend in the replication rate of BMI and sex that increases with the significance of the iQTL in the FUSION discovery dataset (Fig 3). This trend was not observed in T2D, perhaps due to different criteria for inclusion of individuals with T2D. The FUSION tissue study only included individuals with newly diagnosed T2D, not yet treated with antihyperglycemic medications (described in [25]). In contrast, GTEx individuals may have had longstanding and heavily treated T2D [26,27].

Although this bulk replication is reassuring, closer inspection of the BMI and sex trends revealed that two pairs of genes are driving the observed trend in both BMI and sex, highlighting the need of large sample sizes for such GxE analyses. To this point, only two significant iQTL-tSNP pairs from FUSION met the tSNP filtering criteria in GTEx (Methods), neither of which showed similar GxE effects, potentially indicating false positives (S3 Fig).

Specific iQTL example: *FHOD3*

Despite the small number of reported hits and replication challenges, we observe some putative iQTLs with clear, consistent GxE effects in both gene expression and ASE data. The most clear, consistent example is *FHOD3*, formin homology 2 domain containing 3. *FHOD3* is essential for myofibril formation and repair, forming a doughnut shaped dimer, capable of moving along and extending actin filaments (reviewed in [28–30]). *FHOD3* is critical for heart development and function in mouse [31,32] and fly [33] and exhibits tissue specific splicing patterns [34,35] shown to enable myofibril targeting in striated muscle [34,36].

We observed a GxE effect for *FHOD3* with both low-density lipoprotein cholesterol (LDLc) levels and systolic blood pressure (SBP) (Figure 4; S4 Fig). The LDLc association was discovered separately in the ASE of two tSNPs, spanning different exons (S Table 2; Figure 4; S4 Fig), while the SBP association was discovered with an additional tSNP, falling in an exon separate from the LDLc tSNPs. In addition, although not significant in the FUSION dataset, a GxE effect with BMI and *FHOD3* was one of the main drivers of the observed GTEx BMI replication trend (2.47×10^{-4} FUSION and 8.40×10^{-4} GTEx - minimum combined p-value across tSNPs). Evaluation of the raw data showed modest replication of the *FHOD3*-BMI signal between the FUSION and GTEx datasets (S5 Fig).

We previously calculated a muscle expression specificity index (mESI), comparing skeletal muscle expression to a reference panel of 16 diverse tissues, and binned these scores into deciles such that genes in the 1st decile are uniformly, lowly expressed and genes in the 10th decile are highly, specifically expressed in skeletal muscle [25]. We found *FHOD3* expression to be highly specific to skeletal muscle (mESI decile of 9). The iQTL tag SNP, rs17746240, and rs2037043, an

additional SNP in high linkage disequilibrium (R^2 = 0.99 in Finns from the GoT2D reference panel), overlap a skeletal muscle stretch enhancer (Fig 5A), a regulatory element shown to be a signature of tissue-specific active chromatin [37]. In addition, these variants fall in two distinct ATAC-seq peaks unique to skeletal muscle, an indicator of open chromatin (Fig 5B).

Both SNPs affect predicted TF binding sites, as measured by the delta score (Methods). rs17746240 disrupts motifs for the GATA protein family, TBX5, and EP300 (Fig 5C). Within our skeletal muscle data, we find GATA2, GATAD1, GATAD2A, GATAD2B, and EP300 to be expressed (median FPKM > 1). The other variant, rs2037043, disrupts many motifs (Fig 5C) of which ZNF263, YY1AP1, YY1, SMAD4, SIN3A, RXRA, RAD21, NR2C2AP, NR2C2, NFIC, HES1, ESRRA, CTCF, and BDP1 are expressed in skeletal muscle (median FPKM > 1), making it difficult to identify a specific TF.

Conclusion

Understanding the genetic regulators of molecular responses to environment, both at the cellular and organismal level, is essential for a complete understanding of the relationship between genotype and phenotype. Environmental influences are a critical part of human disease etiology, but are far harder to study than intrinsic genetic factors. RNA-seq technology provides an information-dense molecular readout that includes ASE, an internally controlled experiment that minimizes technical artifacts by comparing read counts *within* samples instead of *between* samples [10,24]. Because ASE reduces confounding effects present in gene-level data that are difficult to distinguish from environmental effects, ASE is an ideal molecular readout for probing GxE effects. This study, which is amongst the first to leverage ASE in humans to map GxE effects [10,13], demonstrates both the

potential and the limitations for using ASE to unravel complex gene-environment regulatory structures. Using a well-calibrated model, we find a handful of iQTLs and show some level of bulk replication. Despite the low level of discovery in this study, which we believe is primarily limited by sample size, our success suggests that at least some eQTLs are likely to be in fact iQTLs.

This study highlights several challenges associated with using ASE signal for mapping regulatory loci. Such analyses require sufficient sampling of double heterozygotes of the iQTL and tSNP, and therefore large sample sizes are required for a well-powered study. Another limitation of ASE is that it can only be used to identify *cis*-effects. Previous studies indicate that many iQTLs operate distally, in *trans*, on highly regulated genes with more opportunities in the regulatory chain for genetic perturbation [8,15,22,23]. Because our method requires ASE, we could only assay local, *cis*-effects, and therefore may miss many large *trans*-effects.

In the future, we will need larger studies of specific human tissues with co-measured genetic, molecular, and clinical information. The possibility of mapping iQTLs underscores the importance of detailed characterization of study participants, especially when integrating molecular and genetic data with detailed clinical information. This becomes particularly relevant for replication studies, and argues for the standardization of a core set of phenotypes and environmental exposures between large cohorts. In addition, further development of statistical models to boost power will be needed - for instance by simultaneously modeling total gene expression and ASE, as well as accommodating technology developments, such as the integration of perfectly phased tSNP allele counts within a gene, made possible by long reads.

Materials and Methods

Sample recruitment, muscle biopsy procedures, genotype processing, and RNA sequencing have been previously described [25].

Ethics Statement

The study was approved by the coordinating ethics committee of the Hospital District of Helsinki and Uusimaa. A written informed consent was obtained from all the subjects.

Phenotype Processing

Metabolites were measured after a 12-hour overnight fast, during a 4-point (0, 30, 60, 120 min) oral glucose tolerance test (OGTT) [25]. Serum triglycerides, total and HDL cholesterol were measured by enzymatic methods with Abbott Architect analyzer (Abbott Laboratories, Abbott Park, IL, USA). LDL cholesterol concentration was calculated using the Friedewald formula [38]. Serum insulin and serum C-peptide concentrations were assayed by chemiluminescent microparticle immunoassays using Architect analyzer. Patient medications were also recorded at time of OGTT. Patient medications were analyzed and categorized by physician review. All phenotypes considered are listed in S Table1.

We inverse normalized all continuous traits. Blood pressure measurements were missing from 2 participants, whose samples were dropped when analyzing blood pressure traits. Prior to fitting models, we regressed all continuous traits on age, age^2, and sex, except for age where we regressed only on sex.

ASE Processing

We quantified ASE in autosomal, protein coding genes as described previously [25], removing tSNPs that showed mapping bias based on simulated

reads. To obtain a high confidence ASE dataset, we removed tSNPs per sample with < 30 total reads. We subsequently required that tSNPs were heterozygous in >= 20 samples. From the remaining 25,913 autosomal tSNPs, we discarded 1,254 tSNPs where one or more sample exhibited near mono-allelic expression, defined as | 0.5 - (count $^{\text{alternate SNP}}$ / count $^{\text{total}}$) | > 0.4. Altogether, we considered 24,659 tSNPs to map candidate iQTLs.

iQTL Discovery

Using 19,455 autosomal, protein coding skeletal muscle eQTLs published in [25], we tested for GxE effects in the ASE and gene expression data across all clinical traits. For ASE data, we used EAGLE [10], which models count overdispersion using a random effect term with per tSNP variance vs with an inverse gamma prior *IG(a, b)*. We learned the hyperparameters *a*, *b* for this distribution across all tSNPs after filters, estimating them to be 1.80, 0.0024 respectively. For sample *i* and tSNP *s*, we mapped GxE signals by fitting the model:

$$min(y_{is}, n_{is} - y_{is}) \mid \beta, \mu_s, \epsilon_{is} \sim Binomial[n_{is}, \sigma(e_{is}\gamma_s^e + h_{is}\gamma_s^h + e_{is}h_{is}\beta_s^{eh} + \mu_s + \epsilon_{is})]$$

Here n_{is} and y_{is} denote the total and alternative read count for individual *i* at tSNP *s*, e_{is} the environment, h_{is} the indicator that the eQTL is heterozygous, μ_s an intercept term to take into account unexplained allelic imbalance unrelated to the environment, $\sigma(x) = 1/(1 + e^{-x})$ the logistic function, $\epsilon_{is}/v \sim N(0, v_s)$ a per individual per locus random effect modeling overdispersion, and, γ_s^e, γ_s^h, and β_s^{eh} the effect sizes of the environment, eQTL heterozygosity status, and SNP*environment interaction, respectively. We test the null hypothesis $\beta_s^{eh} = 0$ using a likelihood ratio test. As covariates, we included the first two principal components (PCs) calculated across all genotypes, consistent with Scott *et al.* [25]. In our analyses we required ≥ 15

homozygous and ≥ 15 heterozygous samples for the eQTL tag SNP and, in the case of dichotomous variables, no group was formed with < 5 samples. With these filters, we could only test for iQTL effects in a subset of genes that differed according to clinical trait in the case of discrete variables where the total sample size was not constant due to missing data (S6 Fig).

We also mapped GxE interaction effects for each candidate iQTL in total gene expression data using a linear model for expression levels, testing interactions for each gene-environment pair. Let y_j be a vector of inverse normalized FPKMs for gene *j* across individuals. We consider the following linear genetic model of gene expression:

$$y_j = Z\alpha_j + e\gamma_j^e + g\gamma_j^g + (g \odot e)\beta_j + \psi_j, \quad \sim N(0, \sigma_e^2)$$

Here *Z* denotes the matrix design of fixed effect confounding covariates, *e* and *g* the environment and genotype vector, $g \odot e$ their element-wise product, ψ_j Gaussian noise, and α_j, γ_j^e,γ_j^g, and β_j the effects of covariates, environment, genotype, and the genotype*environment interaction respectively.

To capture hidden variation in gene expression data, we used PEER [39,40] as described previously [25] to learn latent factors. For covariates in the GxE interaction model, we included sequencing batch, the first two genotype PCs, and the first two PEER factors, as a recent report suggests two PEER factors capture the majority of technical variation, preserving biological effects [41]. We additionally include age and sex as covariates when either trait was not considered as an environmental trait. We implemented the GxE model using the linear mixed model framework LIMIX (v0.7.6) [42,43].

We combined the ASE p-values and gene expression p-values using Fisher's combined test. We controlled for FDR per environment using the Benjamini–Hochberg procedure [44]. Our method assumes 1) ASE and gene expression are independent measurements for GxE and 2) we have enough double heterozygous individuals to map the iQTL.

GTEx Replication

We conducted a replication study using genotype, gene expression, and ASE from the GTEx v6 dbGaP release (phs000424.v6.p1). ASE was calculated across imputed genotypes of 360 skeletal muscle samples. The GTEx samples were collected post-mortem and do not have available many of the traits assayed in the FUSION samples. Of the clinical variables measured in the FUSION dataset, four were also recorded in the GTEx dataset - age, sex, BMI, and T2D status - from which we excluded age as the distribution was significantly different between FUSION and GTEx (S Table 1).

Notably, besides the differences in collected phenotype information and age distribution, the GTEx data differ from the FUSION data in four other relevant ways: 1) FUSION is drawn from a more genetically homogenous population (Finland); 2) FUSION is sequenced to mean depth of 91.3M reads per sample compared to 82.1M reads per sample in GTEx; 3) FUSION uses a 100bp strand specific, paired-end read protocol for RNA-seq and GTEx uses 76bp non-strand specific, paired-end RNA-seq; and 4) the computational analysis pipelines are different for read mapping, expression abundance quantification, and ASE calculations [45].

Within the GTEx dataset, we tested for GxE effects with the FUSION eQTL SNPs, using the ASE interaction and gene expression interaction models described above. Because our goal was replication of the FUSION genotype-environment

interactions we did not require the eQTL to be significant. For the GTEx ASE interaction model, we including the first three genotype PCs as covariates, as was used previously by the GTEx consortium [45], and for the gene expression interaction model, we included age, sex, expression batch, the first three genotype PCs, and the first two PEER factors from the GTEx data release as covariates. We tested iQTL-tSNP pairs in GTEx with sufficient double heterozygotes to pass the filters described above. For genes with multiple tSNPs, we selected the minimum iQTL p-value per gene for the GTEx and FUSION datasets separately. Treating the FUSION data as a discovery dataset, we calculated the replication rate across varying p-value threshold cutoffs. We selected n FUSION hits at a given p-value cutoff from N total shared iQTLs without replacement, stopping when $n < 10$. At each cutoff, we calculated k, the number of FUSION hits that replicate in GTEx (GTEx p-value < 0.01), out of the total number of nominally significant GTEx hits, K. Using the mean, K/N, and the hypergeometric distribution, we estimated two standard deviations from the null distribution. Because we select the minimum iQTL-tSNP pair per gene it is possible that genes with more tSNPs will be more likely to show significant results. We calculated the average tSNPs for the replicated and not replicated iQTL sets to explore if sampling from a larger number of transcribed SNPs was responsible for the observed trends (S7 Fig).

Chromatin States

We performed integrative chromatin state analyses as reported previously [37]. Briefly, we collected cell/tissue ChIP-seq (chromatin immunoprecipitation followed by sequencing) reads from a diverse set of publicly available data. Chromatin states were learned jointly by applying the ChromHMM (v1.10) algorithm at 200bp resolution to six data tracks (Input, K27ac, K27me3, K36me3, K4me1,

K4me3) from each of the cell/tissue types [46,47]. We elected a 13 state model as it provided sufficient resolution to identify biologically meaningful patterns in a reproducible way [37].

ATAC-seq Footprinting

Assay for transposase-accessible chromatin (ATAC-seq) generates detailed maps of open, active chromatin and TF binding dynamics [48]. We used previously published ATAC-seq data in skeletal muscle [25].

Transcription Factor Binding Predictions

To identify potential transcription factor binding sites (TFBS), with particular attention to those that may be affected by variants, we generated short sequence fragments around each of the biallelic SNPs and short indels discovered in 1000 Genomes Phase 3 (release 5), by embedding each allele in flanking sequence (29bp on each side) from the hg19 human reference genome. We scanned the entire reference sequence, as well as these variant fragments, with a library of position weight matrices (PWMs) compiled from JASPAR [49], ENCODE [50], and Jolma *et al.* [51], using FIMO [52] from the MEME suite [53]. FIMO was executed using the background nucleotide frequency of the human reference (40.9% GC) and the default p-value cutoff, 10^{-4}.

To quantify the effect of SNPs on these motifs, we calculated a delta score, $-\log10(p^{\text{alternate allele}}) - -\log10(p^{\text{reference allele}})$, for each SNP where at least one of the alleles passed our p-value cutoff of 10^{-4}. In cases where a PWM hit was not detected for the second allele by FIMO at a threshold of 0.01, we use a value of 0.01 for that allele, so that the delta score will be conservative in these cases.

Acknowledgements

We thank Anthony Kirilusha, John Didion, Daniel Bar, and Lori Bonnycastle for helpful comments and feedback. We also thank Julia Fekecs for help in designing Fig 5.

Funding

This research was supported in part by US National Institutes of Health grants 1-ZIA-HG000024 (to F.S.C.), U01DK062370 (to M.B.), R00DK099240 (to S.C.J.P.), the American Diabetes Association Pathway to Stop Diabetes Grant 1-14-INI-07 (to S.C.J.P.), Academy of Finland Grants 271961, 272741 (to M.L.), 258753 (to H.A.K.), and the European Molecular Biology Laboratory (O.S. and E.B.).

References

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. doi:10.1038/nature11247.
2. Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev 14: 2551–2569. doi:10.1101/gad.831000.
3. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224. doi:10.1038/ng2142.
4. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250. doi:10.1126/science.1174148.
5. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369. doi:10.1038/nature04244.
6. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214. doi:10.1371/journal.pgen.1000214.
7. Hunter DJ (2005) Gene-environment interactions in human diseases. Nat Rev Genet 6: 287–298. doi:10.1038/nrg1578.
8. Smirnov DA, Morley M, Shin E, Spielman RS, Cheung VG (2009) Genetic analysis of radiation-induced changes in human gene expression. Nature 459:

587–591. doi:10.1038/nature07940.

9. Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, et al. (2015) Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet 47: 88–91. doi:10.1038/ng.3162.

10. Knowles DA, Davis JR, Raj A, Zhu X, Potash JB, et al. (2015) Allele-specific expression reveals interactions between genetic variation and environment. BioRxiv. doi:10.1101/025874.

11. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, et al. (2012) Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A 109: 1204–1209. doi:10.1073/pnas.1115761109.

12. Maranville JC, Luca F, Stephens M, Di Rienzo A (2012) Mapping gene-environment interactions at regulatory polymorphisms: insights into mechanisms of phenotypic variation. Transcription 3: 56–62. doi:10.4161/trns.19497.

13. Moyerbrailean GA, Richards AL, Kurtz D, Kalita CA, Davis GO, et al. (2016) High-throughput allele-specific expression across 250 environmental conditions. Genome Res 26: 1627–1638. doi:10.1101/gr.209759.116.

14. Casale FP, Horta D, Rakitsch B, Stegle O (2016) Joint genetic analysis using variant sets reveals polygenic gene-context interactions. bioRxiv. Available: http://biorxiv.org/content/early/2016/12/31/097477.

15. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, et al. (2010) Systems genetics analysis of gene-by-environment interactions in human cells. Am J Hum Genet 86: 399–410. doi:10.1016/j.ajhg.2010.02.002.

16. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, et al. (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343: 1246949. doi:10.1126/science.1246949.

17. Idaghdour Y, Quinlan J, Goulet J-P, Berghout J, Gbeha E, et al. (2012) Evidence for additive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci U S A 109: 16786–16793. doi:10.1073/pnas.1204945109.

18. Gagneur J, Stegle O, Zhu C, Jakob P, Tekkedil MM, et al. (2013) Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet 9: e1003803. doi:10.1371/journal.pgen.1003803.

19. Landry CR, Oh J, Hartl DL, Cavalieri D (2006) Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. Gene 366: 343–351. doi:10.1016/j.gene.2005.10.042.

20. Sambandan D, Carbone MA, Anholt RRH, Mackay TFC (2008) Phenotypic

plasticity and genotype by environment interaction for olfactory behavior in Drosophila melanogaster. Genetics 179: 1079–1088. doi:10.1534/genetics.108.086769.

21. Runcie DE, Garfield DA, Babbitt CC, Wygoda JA, Mukherjee S, et al. (2012) Genetics of gene expression responses to temperature stress in a sea urchin gene network. Mol Ecol 21: 4547–4562. doi:10.1111/j.1365-294X.2012.05717.x.

22. Smith EN, Kruglyak L (2008) Gene-environment interaction in yeast gene expression. PLoS Biol 6: e83. doi:10.1371/journal.pbio.0060083.

23. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, et al. (2006) Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet 2: e222. doi:10.1371/journal.pgen.0020222.

24. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T (2015) Tools and best practices for data processing in allelic expression analysis. Genome Biol 16: 195. doi:10.1186/s13059-015-0762-6.

25. Scott LJ, Erdos MR, Huyghe JR, Welch RP, Beck AT, et al. (2016) The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun 7: 11764. doi:10.1038/ncomms11764.

26. Keen JC, Moore HM (2015) The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. J Pers Med 5: 22–29. doi:10.3390/jpm5010022.

27. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580–585. doi:10.1038/ng.2653.

28. Paul AS, Pollard TD (2009) Review of the mechanism of processive actin filament elongation by formins. Cell Motil Cytoskeleton 66: 606–617. doi:10.1002/cm.20379.

29. Goode BL, Eck MJ (2007) Mechanism and function of formins in the control of actin assembly. Annu Rev Biochem 76: 593–627. doi:10.1146/annurev.biochem.75.103004.142647.

30. Campellone KG, Welch MD (2010) A nucleator arms race: cellular control of actin assembly. Nat Rev Mol Cell Biol 11: 237–251. doi:10.1038/nrm2867.

31. Rosado M, Barber CF, Berciu C, Feldman S, Birren SJ, et al. (2014) Critical roles for multiple formins during cardiac myofibril development and repair. Mol Biol Cell 25: 811–827. doi:10.1091/mbc.E13-08-0443.

32. Kan-O M, Takeya R, Abe T, Kitajima N, Nishida M, et al. (2012) Mammalian formin Fhod3 plays an essential role in cardiogenesis by organizing myofibrillogenesis. Biol Open 1: 889–896. doi:10.1242/bio.20121370.

33. Wooten EC, Hebl VB, Wolf MJ, Greytak SR, Orr NM, et al. (2013) Formin homology 2 domain containing 3 variants associated with hypertrophic

cardiomyopathy. Circ Cardiovasc Genet 6: 10–18. doi:10.1161/CIRCGENETICS.112.965277.

34. Iskratsch T, Lange S, Dwyer J, Kho AL, dos Remedios C, et al. (2010) Formin follows function: a muscle-specific isoform of FHOD3 is regulated by CK2 phosphorylation and promotes myofibril maintenance. J Cell Biol 191: 1159–1172. doi:10.1083/jcb.201005060.

35. Kanaya H, Takeya R, Takeuchi K, Watanabe N, Jing N, et al. (2005) Fhos2, a novel formin-related actin-organizing protein, probably associates with the nestin intermediate filament. Genes Cells 10: 665–678. doi:10.1111/j.1365-2443.2005.00867.x.

36. Iskratsch T, Reijntjes S, Dwyer J, Toselli P, Dégano IR, et al. (2013) Two distinct phosphorylation events govern the function of muscle FHOD3. Cell Mol Life Sci 70: 893–908. doi:10.1007/s00018-012-1154-7.

37. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, et al. (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A 110: 17921–17926. doi:10.1073/pnas.1317023110.

38. Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 18: 499–502.

39. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol 6: e1000770. doi:10.1371/journal.pcbi.1000770.

40. Stegle O, Parts L, Piipari M, Winn J, Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 7: 500–507. doi:10.1038/nprot.2011.457.

41. Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, et al. (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol 32: 888–895. doi:10.1038/nbt.3000.

42. Lippert C, Casale FP, Rakitsch B, Stegle O (2014) LIMIX: genetic analysis of multiple traits. BioRxiv. doi:10.1101/003905.

43. Casale FP, Rakitsch B, Lippert C, Stegle O (2015) Efficient set tests for the genetic analysis of correlated traits. Nat Methods 12: 755–758. doi:10.1038/nmeth.3439.

44. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300. Available:

http://www.jstor.org/stable/2346101.

45. GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648–660. doi:10.1126/science.1262110.

46. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49. doi:10.1038/nature09906.

47. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9: 215–216. doi:10.1038/nmeth.1906.

48. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10: 1213–1218. doi:10.1038/nmeth.2688.

49. Mathelier A, Fornes O, Arenillas DJ, Chen C-Y, Denay G, et al. (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44: D110–5. doi:10.1093/nar/gkv1176.

50. Kheradpour P, Kellis M (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res 42: 2976–2987. doi:10.1093/nar/gkt1249.

51. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, et al. (2013) DNA-binding specificities of human transcription factors. Cell 152: 327–339. doi:10.1016/j.cell.2012.12.009.

52. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018. doi:10.1093/bioinformatics/btr064.

53. Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Res 43: W39–49. doi:10.1093/nar/gkv416.

Supporting Information

S1 Table. Clinical traits. Phenotype information used as traits from the FUSION tissue biopsy study participants and GTEx skeletal muscle participants. For T2D status in GTEx, only T2D status available, non-T2D participants presumed to be NGT. In some cases, the GTEx T2D status was missing (NA), therefore T2D fraction calculated over non-missing data.

S2 Table. All iQTL-tSNP pairs FDR 10%. All candidate iQTLs (FDR 10%).

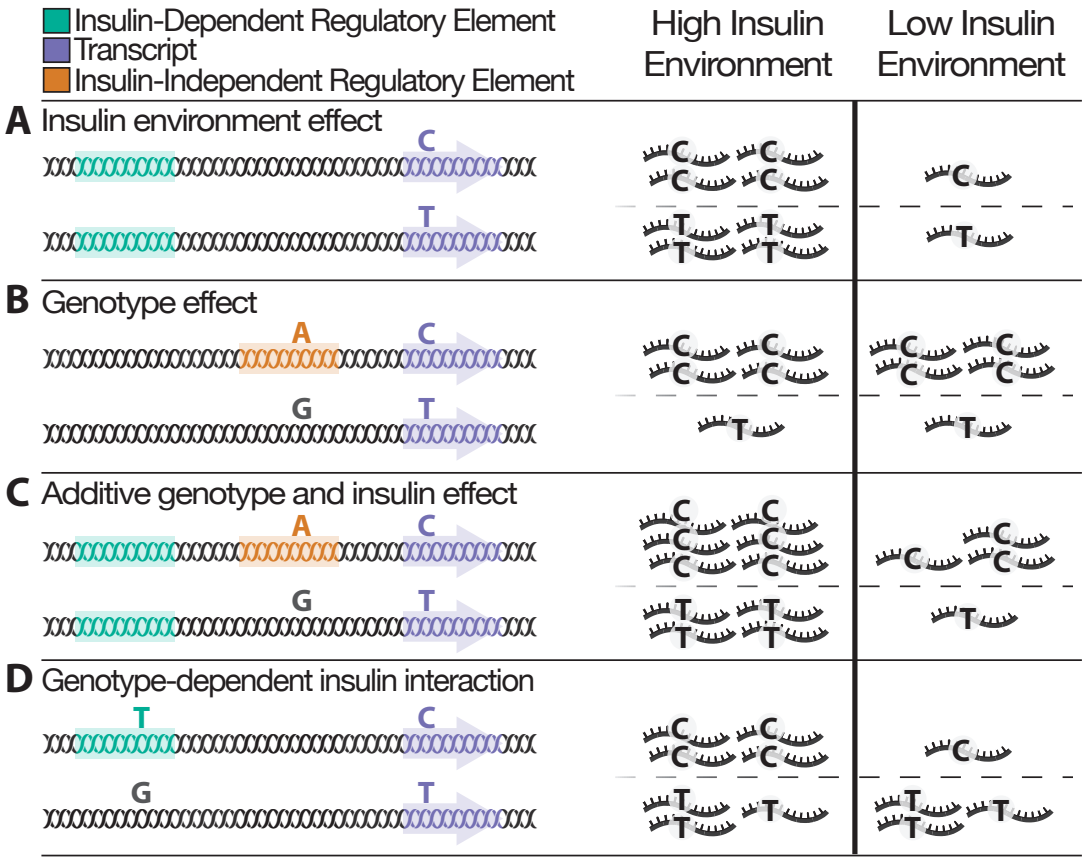


Fig 1. Genetic and environmental effects on gene expression. Blood insulin levels represent a cellular environment for tissues such as skeletal muscle. The left panel depicts a single genome with color-coded genomic elements and various heterozygous sites. The right panel shows the relative transcript abundance for the corresponding locus on the left panel. Some genomic elements contain genetic variants. When the variant is the same color as the element, the element is active. In some cases the variant is black, indicating that the variant renders the regulatory element nonfunctional and only basal transcription occurs. The purple element represents a gene with a transcribed SNP (tSNP), shown in the transcripts. Allele specific expression is calculated across both chromosomes and compared to the high and low environment. (A) When regulated by an insulin-responsive element (green), gene expression changes according to insulin concentrations in the extracellular environment. (B) When regulated by an insulin-independent element (orange) containing genetic variation, gene expression changes according to the presence of a genetic variant (eQTL), but not to insulin levels. The tSNP shows allelic bias due to the eQTL effect, but is not associated with the insulin environment. (C) When regulated by both an insulin-responsive element and an insulin-independent element containing genetic variation, the effects of the insulin environment and the genetic variation on gene expression may be additive, although more complex relationships are possible. The tSNP shows some imbalance due to the eQTL effect and is associated to insulin levels. Such cases may be identified as weak iQTLs. (D) When regulated by an insulin-responsive element containing genetic variation, there may exist an interaction effect between the genetic variant and insulin levels such that changes in gene expression across insulin environments depend on the genetic variant. The tSNP shows allelic imbalance associated with insulin levels due to the iQTL effect. One of several possible interaction effects depicted.

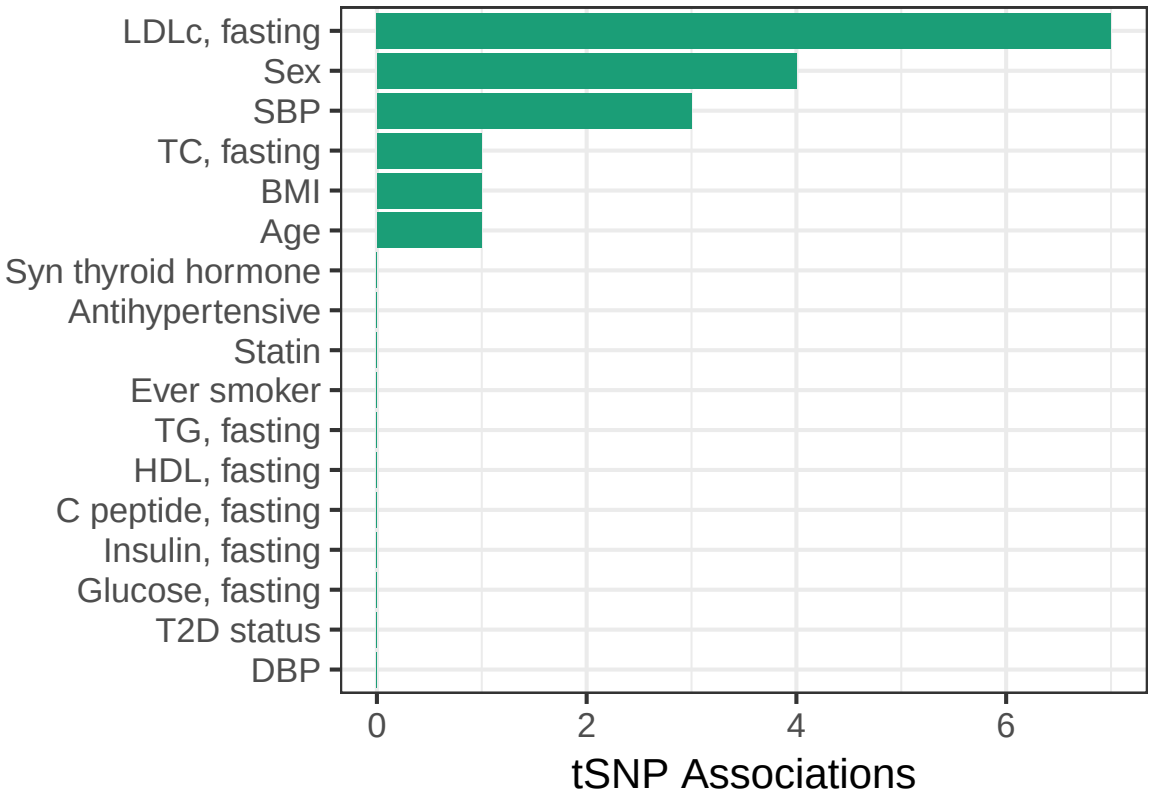


Fig 2. GxE signals. Number of tSNP-environment associations per clinical variable at a 10% FDR.

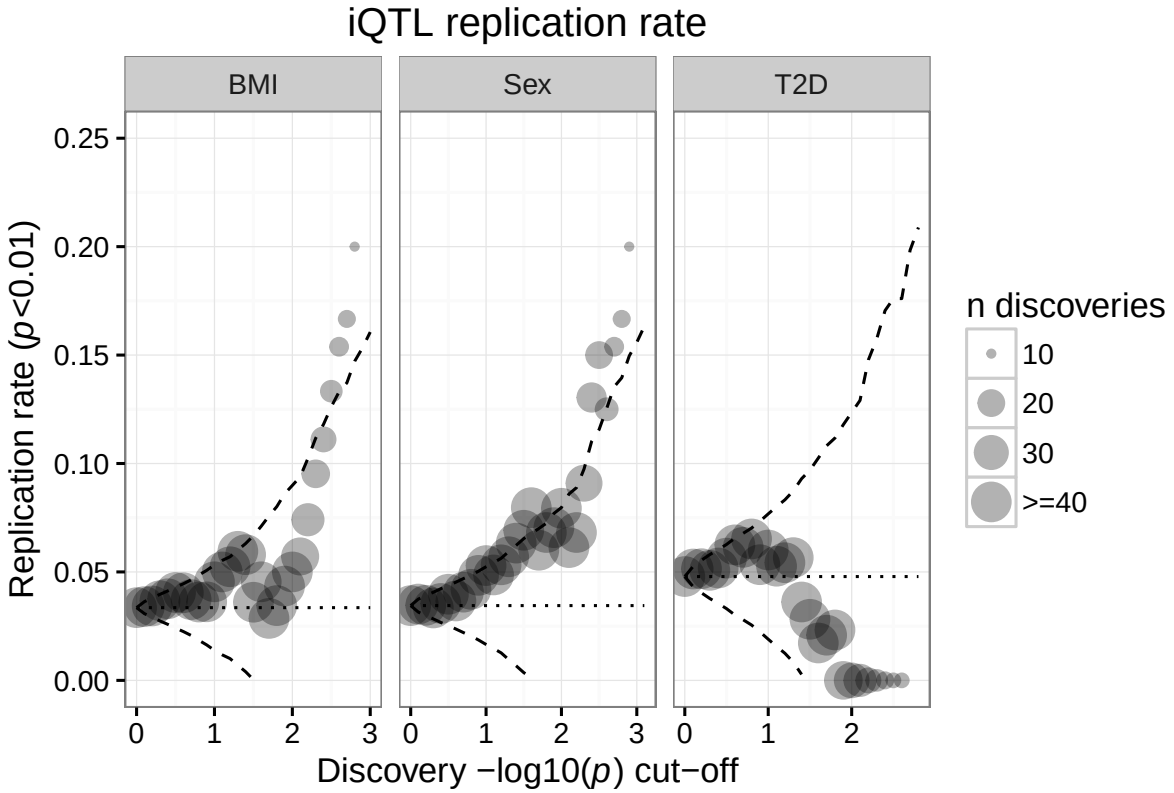


Fig 3. GTEx Replication. Replication rate (y axis) as a function of FUSION iQTL p-value cutoff (x axis). Dashed line represents two standard deviations from the null distribution, calculated using the hypergeometric distribution.

(A) FHOD3 LDLc-iQTL **(B)** FHOD3 SBP-iQTL

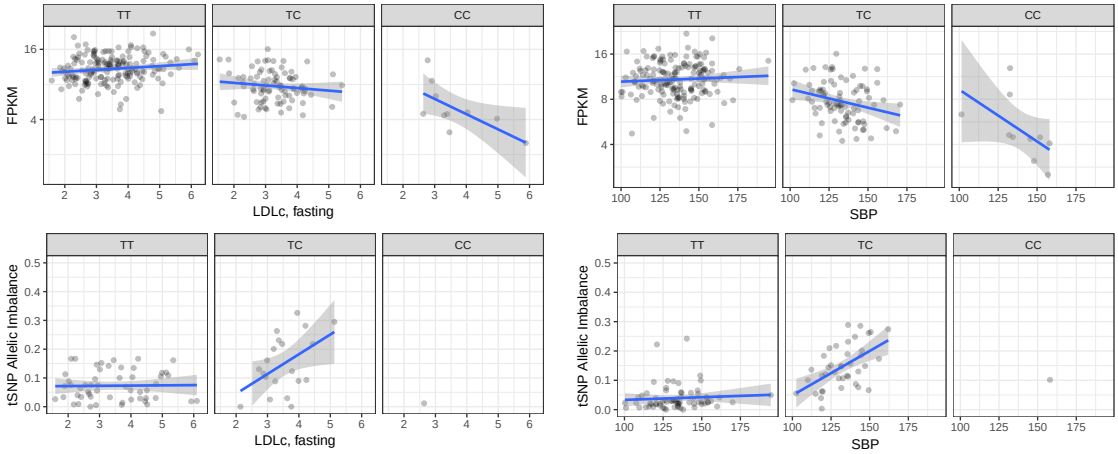


Fig 4. FHOD3 iQTL, rs17746240 (18:33970347). The data for each of the three possible iQTL genotypes are presented in separate plots (columns). The top row plots show the relationship between gene expression (y axis) and the clinical variable (x axis). The bottom row plots show the relationship between the allelic imbalance of the tSNP and the clinical variable (x axis). Note the bottom row has fewer samples because it is limited to samples heterozygous for the tSNP. (A) LDLc GxE effect with rs72895597 (18:34232657) as the tSNP (B) SBP GxE effect with rs2303510 (18:34324091) as the tSNP.

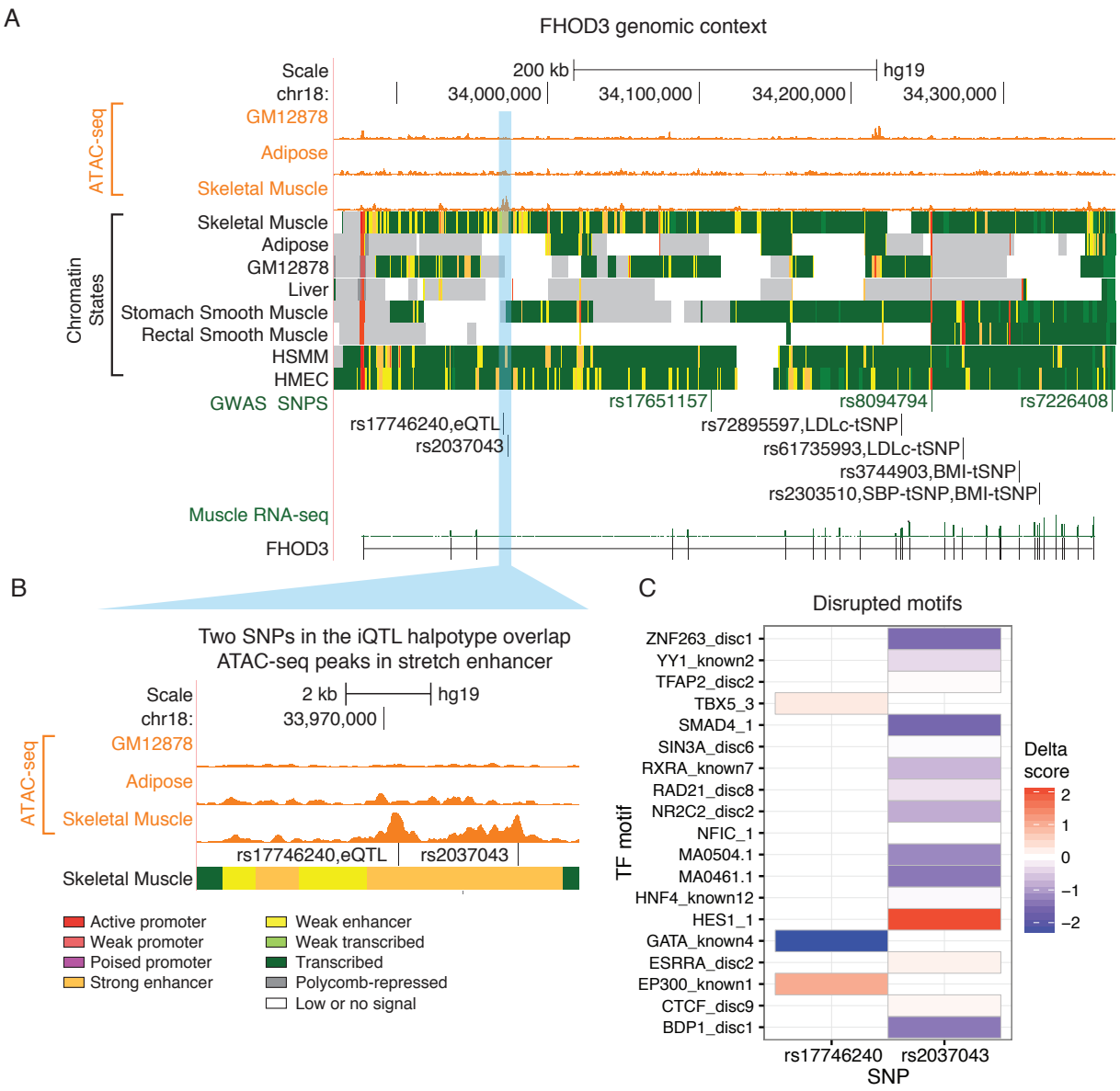
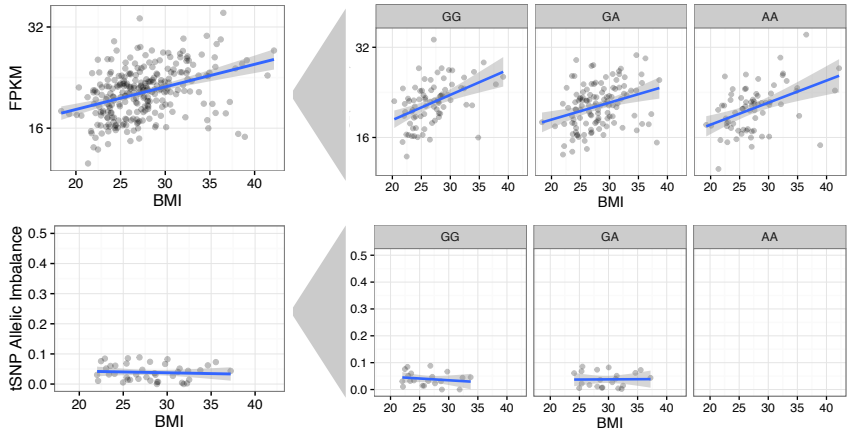
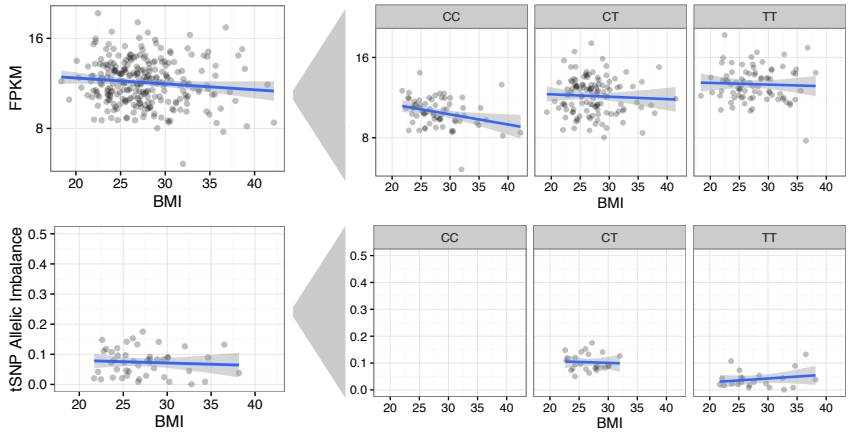


Fig 5. FHOD3 locus. (A) Top wiggle tracks show ATAC-seq signal in multiple cell types, followed by ChromHMM chromatin state tracks. Beneath are FHOD3 GWAS loci and the SNPs from this study (iQTL and tSNP). The bottom track shows the FUSION FHOD3 RNA-seq signal. (B) ATAC-seq signal highlights potential regulatory regions with the skeletal muscle stretch enhancer. (C) Effects of SNPs overlapping ATAC-seq peaks in the iQTL haplotype on in silico predicted TF binding.

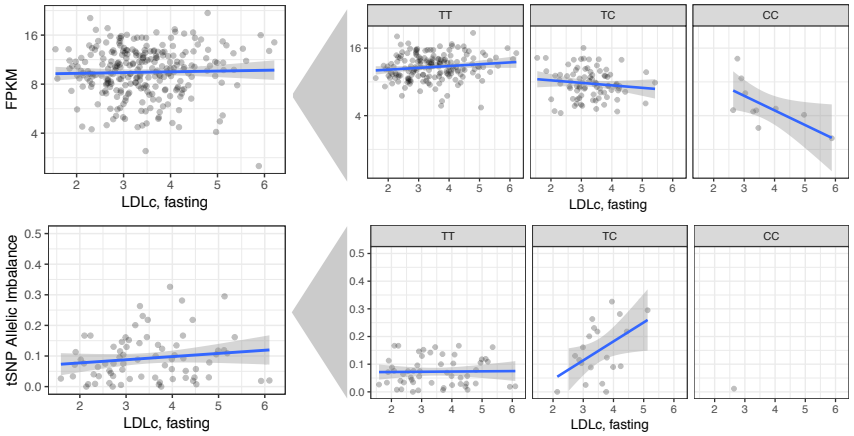
(A) Environment effect



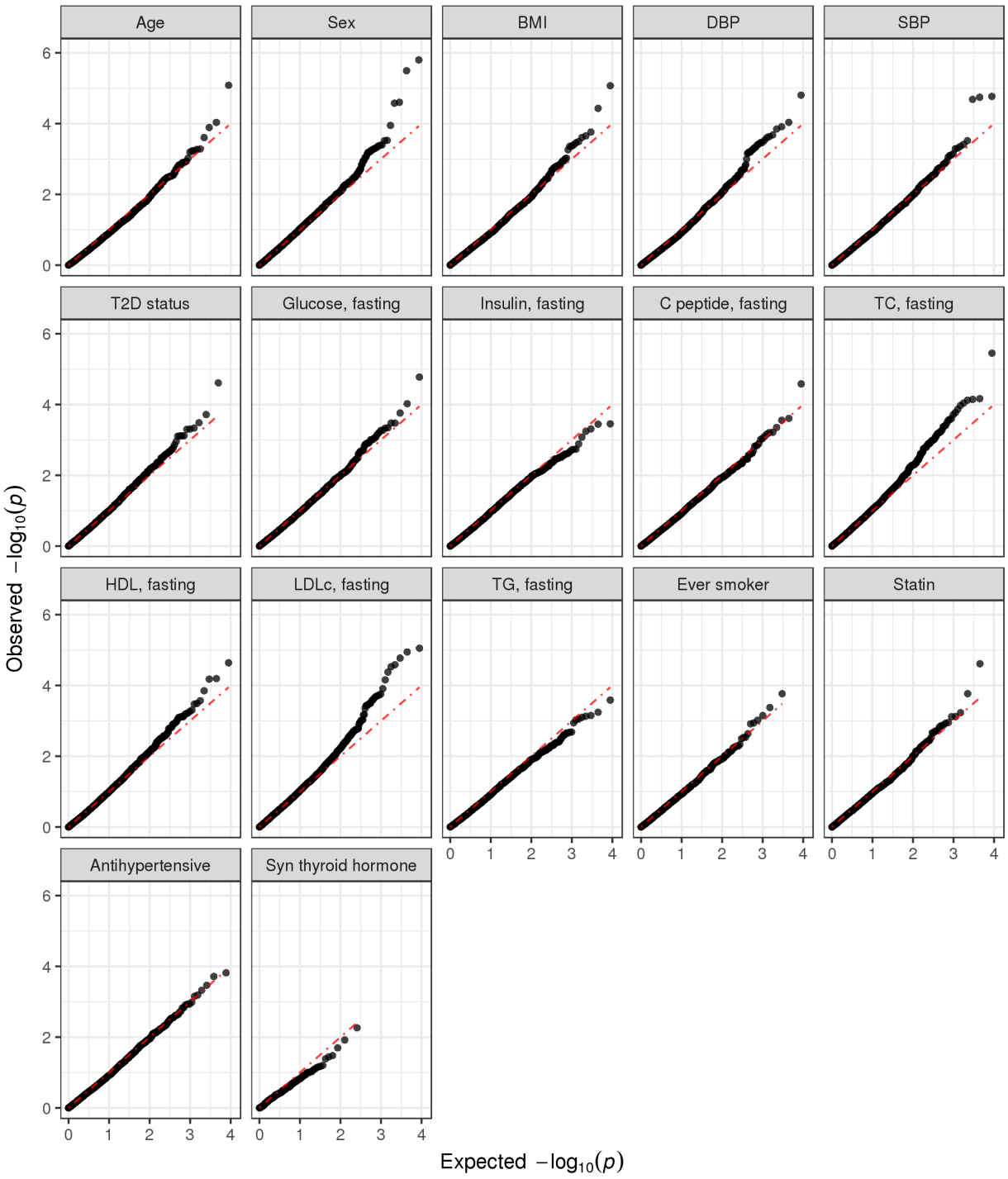
(B) Genotype effect



(C) Genotype-environment interaction effect

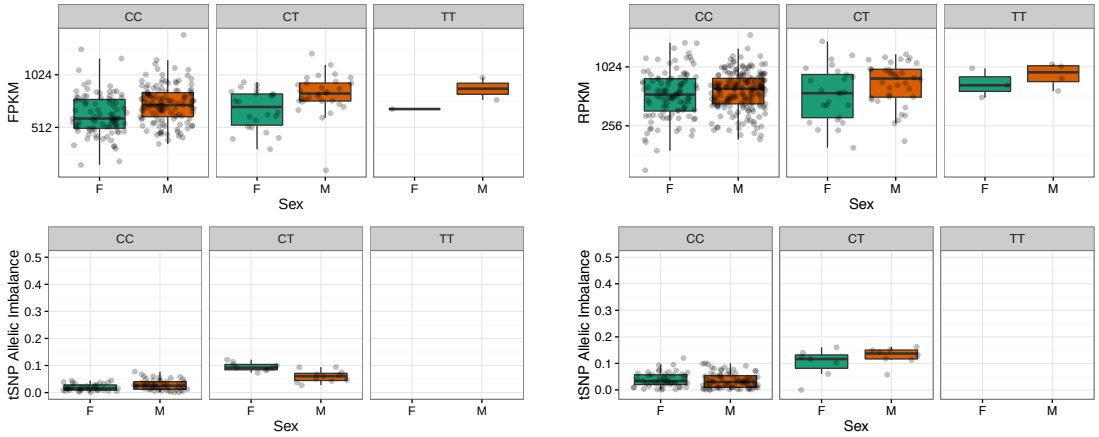


S1 Fig. Examples of genetic and environmental effects. (A) Example of a pure environment effect in SZRD1 - rs12568938 regulatory SNP (rSNP) and rs7529767 transcribed SNP (tSNP). SZRD1 expression is associated with BMI, and the rSNP does not affect gene expression. The relationship between SZRD1 and BMI does not change across the rSNP alleles, and BMI is not associated with allelic imbalance. (B) Example of a pure genetic effect in RBM6 - rs9881008 regulatory locus and rs2023953 tSNP. BMI is not associated with RBM6 expression or allelic imbalance. The rSNP alleles are associated with RBM6 expression and allelic imbalance is increased in samples heterozygous for the rSNP. (C) Example of a GxE effect in FHOD3 - rs17746240 regulatory locus and rs72895597 tSNP. The relationship between LDLc and FHOD3 expression changes according to the rSNP allele as well as the overall expression abundance levels. LDLc is only associated with allelic imbalance in heterozygous individuals, where preferential TF binding could occur.

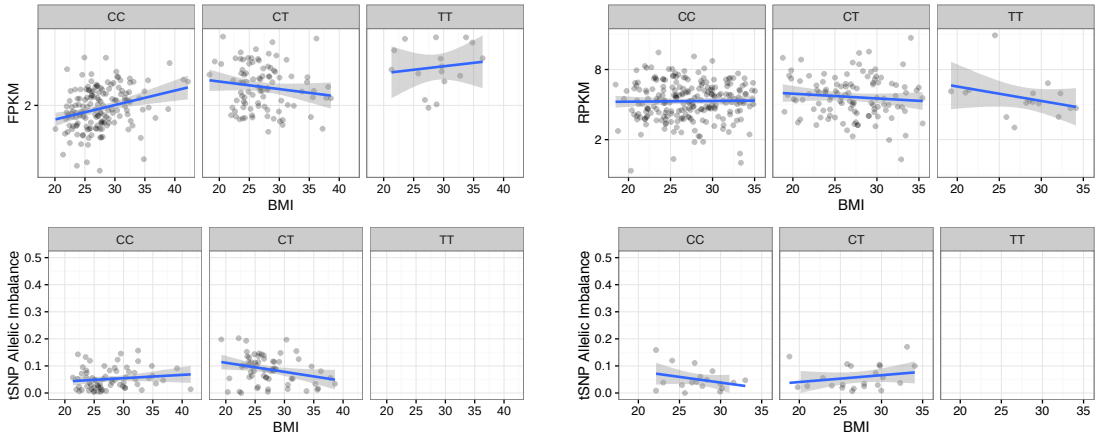


S2 Fig. QQ-plots across traits. QQ-plots of GxE signal discovery across clinical traits.

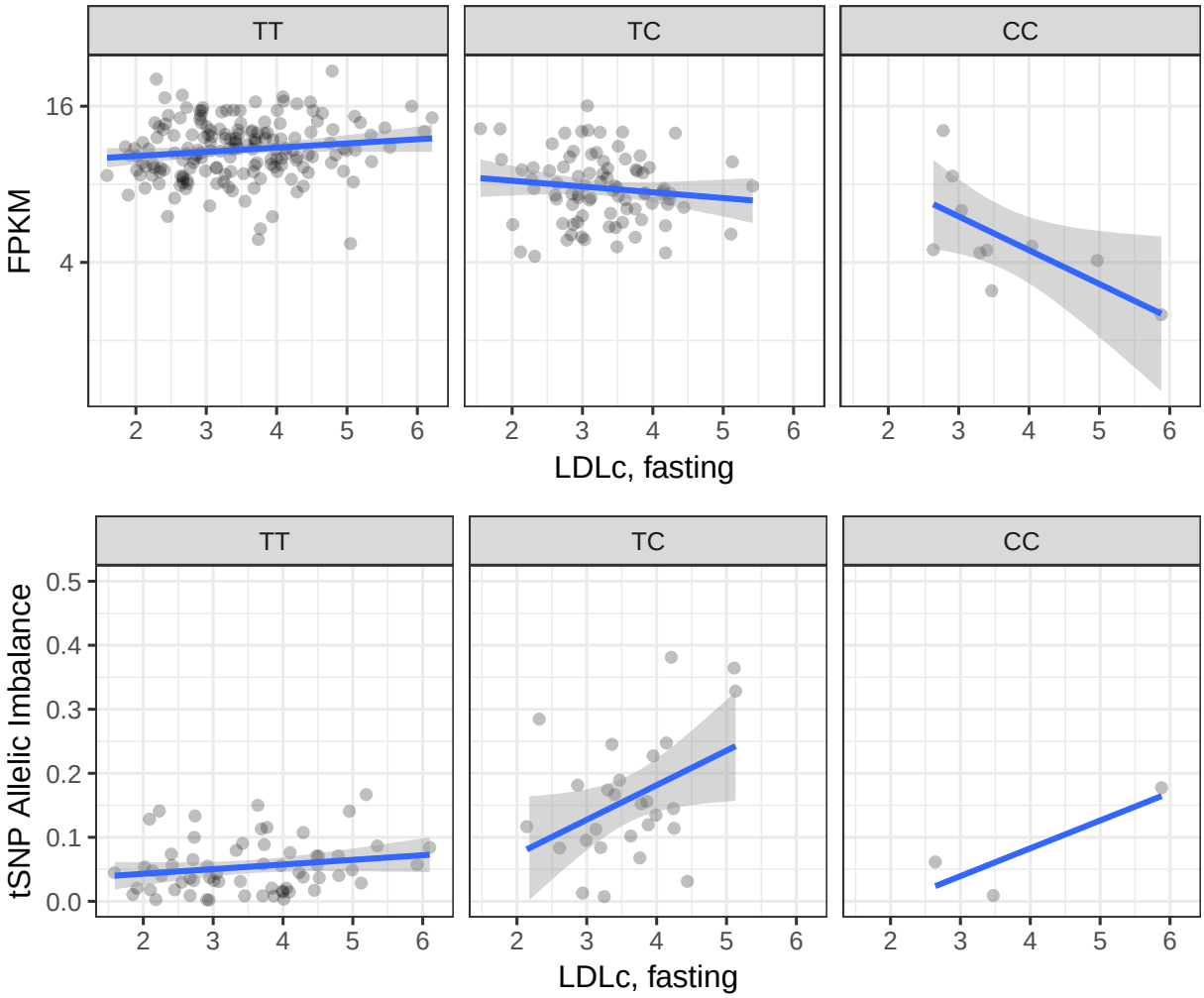
(A) NRAP sex-iQTL FUSION (tSNP 115412793) **(B)** NRAP sex-iQTL GTEx (tSNP 115412793)



(C) DAGLB BMI-iQTL FUSION (tSNP 6449272) **(D)** DAGLB BMI-iQTL GTEX (tSNP 6449272)

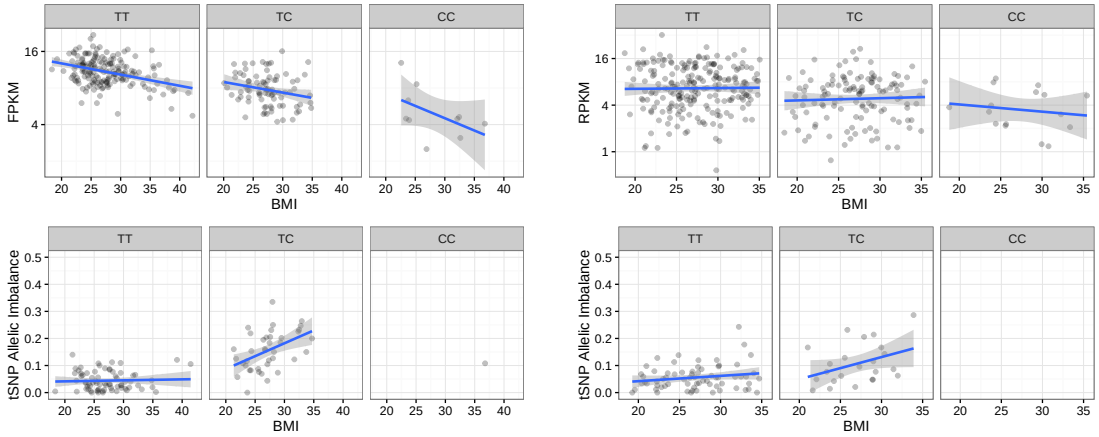


S3 Fig. Comparison of candidate FUSION iQTLs to GTEx.(A) NRAP sex-iQTL in FUSION (B) NRAP sex-iQTL in GTEx (C) DAGLB BMI-iQTL in FUSION (D) DAGLB BMI-iQTL in GTEx.

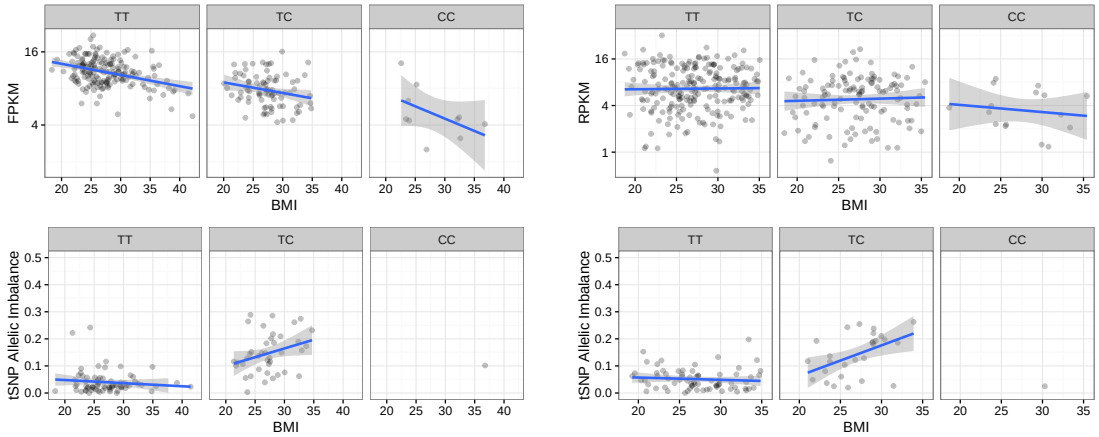


S4 Fig. Additional FHOD3 LDLc-iQTL. Additional LDLc GxE effect with rs61735993 (18:34273279) as the tSNP.

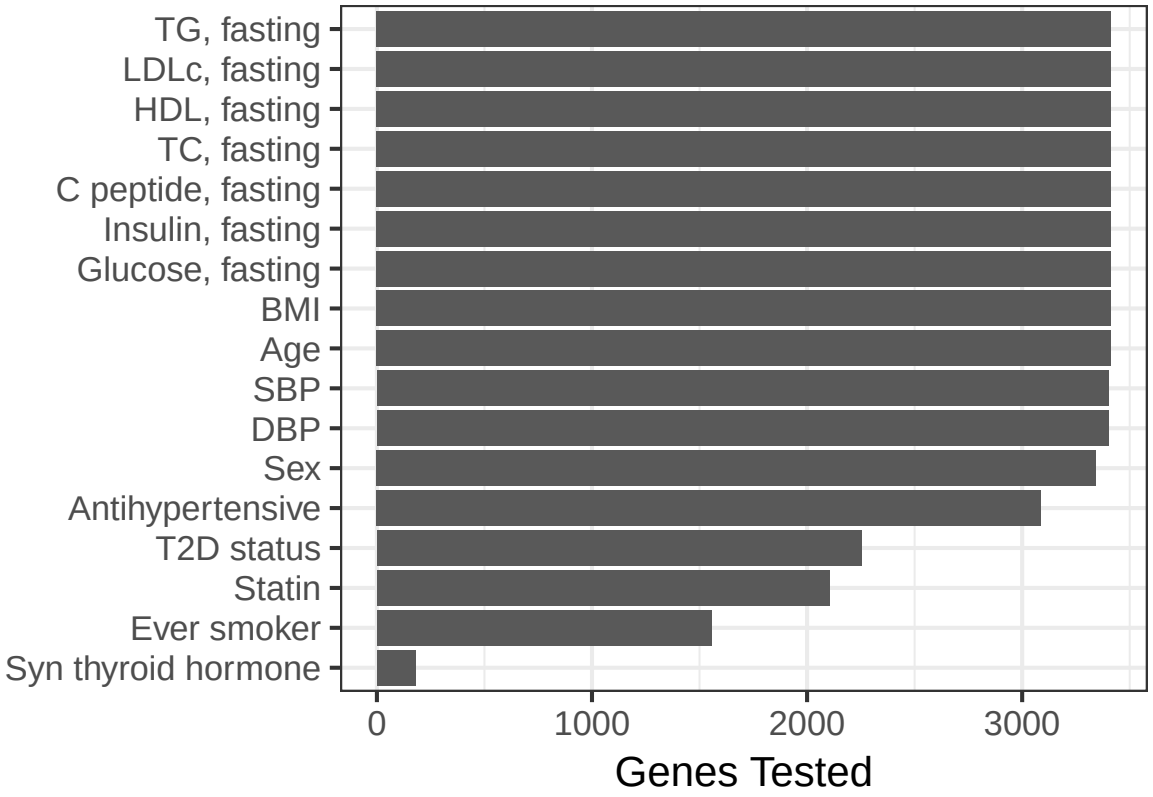
(A) FHOD3 BMI-iQTL FUSION (tSNP 34310668) **(B)** FHOD3 BMI-iQTL GTEx (tSNP 34310668)



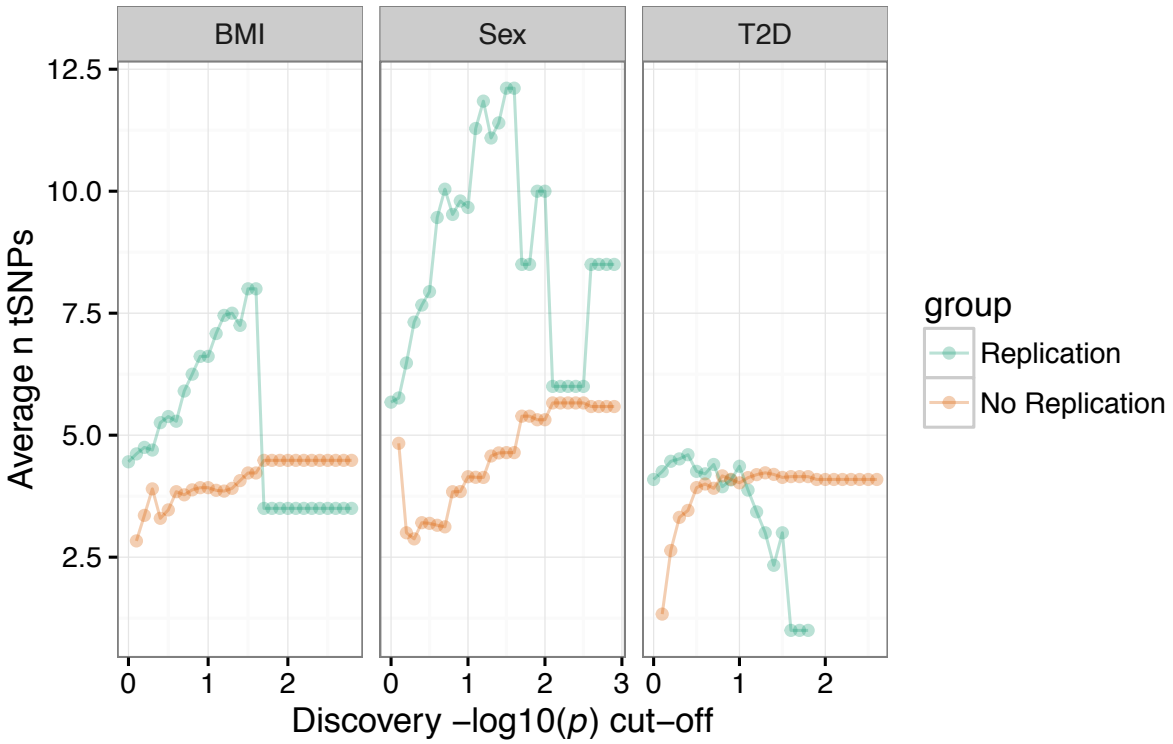
(C) FHOD3 BMI-iQTL FUSION (tSNP 34324091) **(D)** FHOD3 BMI-iQTL GTEx (tSNP 34324091)



S5 Fig. Comparison of FHOD3 BMI-iQTL in FUSION and GTEx. (A) FHOD3 BMI-iQTL in FUSION with rs3744903 (18:34310668) as the tSNP (B) FHOD3 BMI-iQTL in GTEx with rs3744903 (18:34310668) as the tSNP (C) FHOD3 BMI-iQTL in FUSION with rs2303510 (18:34324091) as the tSNP (D) FHOD3 BMI-iQTL in GTEx with rs2303510 (18:34324091) as the tSNP.



S6 Fig. Total number of tested genes across traits. Total number of genes in FUSION considered for each clinical trait.



S7 Fig. FUSION-GTEx Replication. Average number of tSNPs in the genes with signals that replicated (Replication group) and signals that did not replicate (No Replication).