Title:

Systematic assessment of regulatory effects of human disease variants in pluripotent cells

Authors:

Marc Jan Bonder[*1,2], Craig Smail[*3], Michael J. Gloudemans[3], Laure Frésard[4], David Jakubosky[5,6], Matteo D'Antonio[7], Xin Li[4,8], Nicole M. Ferraro[3], Ivan Carcamo-Orive[9], Bogdan Mirauta[1], Daniel D. Seaton[1], Na Cai[1,10], Danilo Horta[1], YoSon Park[11], HipSci Consortium, iPSCORE Consortium, GENESiPS Consortium, PhLiPS Consortium, Erin N. Smith[11], Kelly A. Frazer[7,12], Stephen B. Montgomery[+,4], Oliver Stegle[+,1,2,13]

Affiliations:

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, CB10 1SD Cambridge, UK
2. European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
3. Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, US
4. Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
5. Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA
6. Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093, USA
7. Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093 USA
8. Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China
9. Division of Cardiovascular Medicine and Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
10. Wellcome Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA Cambridge, UK
11.
12. Department of Pediatrics and Rady Children's Hospital, University of California, San Diego, La Jolla, CA, 92093, USA
13. DKFZ: Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

* These authors contributed equally to this work

[+] Corresponding authors (smontgom@stanford.edu, oliver.stegle@embl.de)

Abstract

Identifying regulatory genetic effects in pluripotent cells provides important insights into disease variants with potentially transient or developmental origins. Combining existing and newly-generated data, we established a population-scale resource of 1,367 induced pluripotent stem cell lines derived from 948 unique donors, with matched RNA-sequencing (RNA-seq) and genetic information. The sample size of our study allowed us to significantly expand our knowledge of quantitative trait loci (QTL) in pluripotent human cells and their broad trait and disease relevance.

We identified *cis*-QTL for conventional gene expression levels but also transcript-ratio, exon expression levels, alternative-splicing and alternative polyadenylation usage, cumulatively yielding *cis* regulatory effects for 18,556 genes. We assessed the effects of rare variants identified from whole-genome sequencing data, by relating these to gene expression outliers, observing an enrichment for rare, deleterious variants in iPSCs with outlying expression profiles, which exceeds previous observations in differentiated tissues. Finally, we assessed distal QTL on a genome-wide scale, identifying 193 *trans*-eQTL, linked to 191 *trans*-eGenes (FDR <10%), 38% of which were replicated in independent samples.

By linking our regulatory map of rare and common QTL to comprehensive GWAS data, we identified high-confidence colocalization events for 4,336 individual GWAS loci, which includes physical traits such as height and coronary artery disease. In addition, we identify rare variant associations for metabolic rate and *trans*-eQTL linked to both cancer and height. Collectively, our data greatly expand the regulatory landscape in human pluripotent cells, and catalogues trait-associated variants that have potential developmental or transient contexts.

Introduction

Genome-wide association studies (GWAS) have yielded a compendium of genetic variants that are associated with human traits and diseases. However, the majority of these variants are in intergenic regions[1], and identifying their molecular function remains an ongoing challenge. Molecular quantitative trait loci (QTL) studies have been performed to systematically identify the function of genetic variants via their effects on molecular intermediates such as gene expression levels. However, existing QTL resources are largely focused on differentiated cell types, including blood[2–4] and post-mortem collected tissues[5]. In contrast, much less is known about the impact of genetic variants on molecular intermediates in pluripotent cells, which have only recently become accessible via the first population-based iPSC studies with moderate sample sizes[6–10]. Due to their embryonic-like state, iPSCs can provide unique insights into genetic regulation of expression in cell states that mimic early development, with relevance to diseases that manifest *in utero* or in transient states throughout development[9,11,12].

We characterized 1,367 iPSC lines from 948 unique donors, collectively analyzed within the "Integrated iPSC QTL" (i2QTL) Consortium, and provide a comprehensive map of expression regulatory variants in human iPSC. We mapped the effects of nearby common genetic variants on five expression phenotypes, identifying *cis*-QTL at gene-, exon-level and transcript-, splicing-, alternative polyadenylation-ratio (APA). For gene-level, we further quantified the effects of rare and singleton variants, and the effect of distal variants that act in *trans* (*trans*-eQTL). Our data are a valuable community resource, uncovering novel regulatory effects that have not previously been described in differentiated cells and tissues. Building on this regulatory map, we functionally explore GWAS signals for over 4,336 trait loci, finding evidence for colocalization with iPSC QTL for traits such as height and BMI, and diseases, such as cancer and coronary artery disease.

Results

Within the i2QTL Consortium we collected previously published and novel data from human iPSC lines (**Table S1 & S2**), providing genotype and RNA-seq data from a total of 1,367 iPSC lines, derived from 948 unique donors. The data were obtained from five iPSC consortia; iPSCORE[6] (N=206), GENESiPS[10], (N=85), PhLiPS[7] (N=65), Banovich *et al.*[8] (N=49) and HipSci[9], which has been greatly expanded in size for this study (N=543, original study N=166). Additionally, data from the Choi et al[13] study was included as reference, adding a total of 92 samples comprising fibroblast, ESC and iPSC, used as reference (**Table S1 & S2**).

In order to mitigate technical differences, we uniformly reprocessed all genotype data, considering both genotyping array and whole genome sequencing data, as well as uniformly reprocessed all RNA-seq data (see Methods). Joint multidimensional scaling (MDS) of i2QTL and GTEx[5] (v7) samples revealed a remarkable level of homogeneity between the iPSC studies compared to tissue-to-tissue and between-sample differences observed in GTEx, with the remaining variability in iPSCs being primarily due to differences in sequencing methods between iPSC consortia (**Fig. 1a**). Using samples from Choi et al, we were able to confirm both the minimal processing effects, using the fibroblast lines, and the embryonic-like state of our iPSC lines.

Mapping *cis* regulatory effects in iPSCs

We mapped *cis*-QTL, considering proximal (within the gene body +/- 250kb) common variants (MAF >1%) and paired-end stranded RNA-seq data available for 936 samples (N=682 donors) of European ancestry (see Methods).

Among 27,046 tested Ensembl genes, we identified at least one *cis*-eQTL for 18,430 genes (68% of tested; FDR<5% permutation-based, **Table S3**, Methods; in the following denoted eGenes), which corresponds to a 2.5 fold increase in the number of *e*Genes compared to the largest previous *cis*-eQTL map in iPSCs[9] (**Fig. 1b**). Conversely, the vast majority of the eQTL effects (between 62-86%) described in the previous studies were detected in our data (replication defined at nominal P<0.05 with matched effect direction) **(Fig. S1-4)**. We also performed iterative eQTL mapping using stepwise regression (see Methods), which identified two or more independent effects for 39.0% of eGenes, with up to 12 independent *cis*-eQTL for *PTGR1* **(Fig. S5)**.

To gain insights into iPSC-specific regulation, we contrasted our *cis*-eQTL map with somatic eQTL maps from GTEx[5] tissues and blood eQTL from BIOS[4]. We observed a larger number of eGenes in iPSC as compared to any of the single GTEx tissues (**Fig. 1b-c**), and approximately equal numbers as observed in BIOS, despite the substantially larger sample size of this study (**Fig 1b)**. The increase in power to detect eGenes in iPSC is in line with previous results[9,14], likely reflecting the homogeneity of the iPSC lines[15]. We identified 994 eGenes without prior evidence for regulatory effects in neither GTEx nor BIOS (methods; **Table S3**), 117 of which are protein-coding (**Fig 1c**). This set of iPSC-specific eGenes are enriched for known cancer genes curated in the COSMIC database (p<2.2×10^-16, 84 novel eGenes are in COSMIC), and we identified eQTL for 7 DECIPHER disease-implicated genes database (**Fig 1c**).

We extended the widely considered gene-level eQTL mapping to an extended set of expression phenotypes (transcript-ratio, exon-level, splicing-ratio (sQTL) and APA-ratio) (**Table S4-7**). To aid comparability, we used the same gene-based test region to test for QTL variants followed by gene-based FDR adjustment (Methods). In aggregate across all five expression phenotypes, this identified genetic effects for of 18,556 Ensembl genes (FDR<5% permutation-based, Methods; **Fig 1d**), where 77% of the eGenes discovered in one phenotype were also detected in at least one other expression-phenotypes. Despite this substantial overlap globally, we observed 3,118 genes (9.8% of all tested genes) with an exon-, transcript-, splicing- or APA-QTL but no gene-level eQTL, highlighting the benefits of QTL mapping using multiple expression phenotypes. Assessment of the pairwise replication between QTL maps at variant level confirmed substantial sharing of QTL maps between phenotypes with the largest overlap between the exon-level and gene-level effects, and the most distinct QTL identified at APA-ratio (**Fig 1E**).

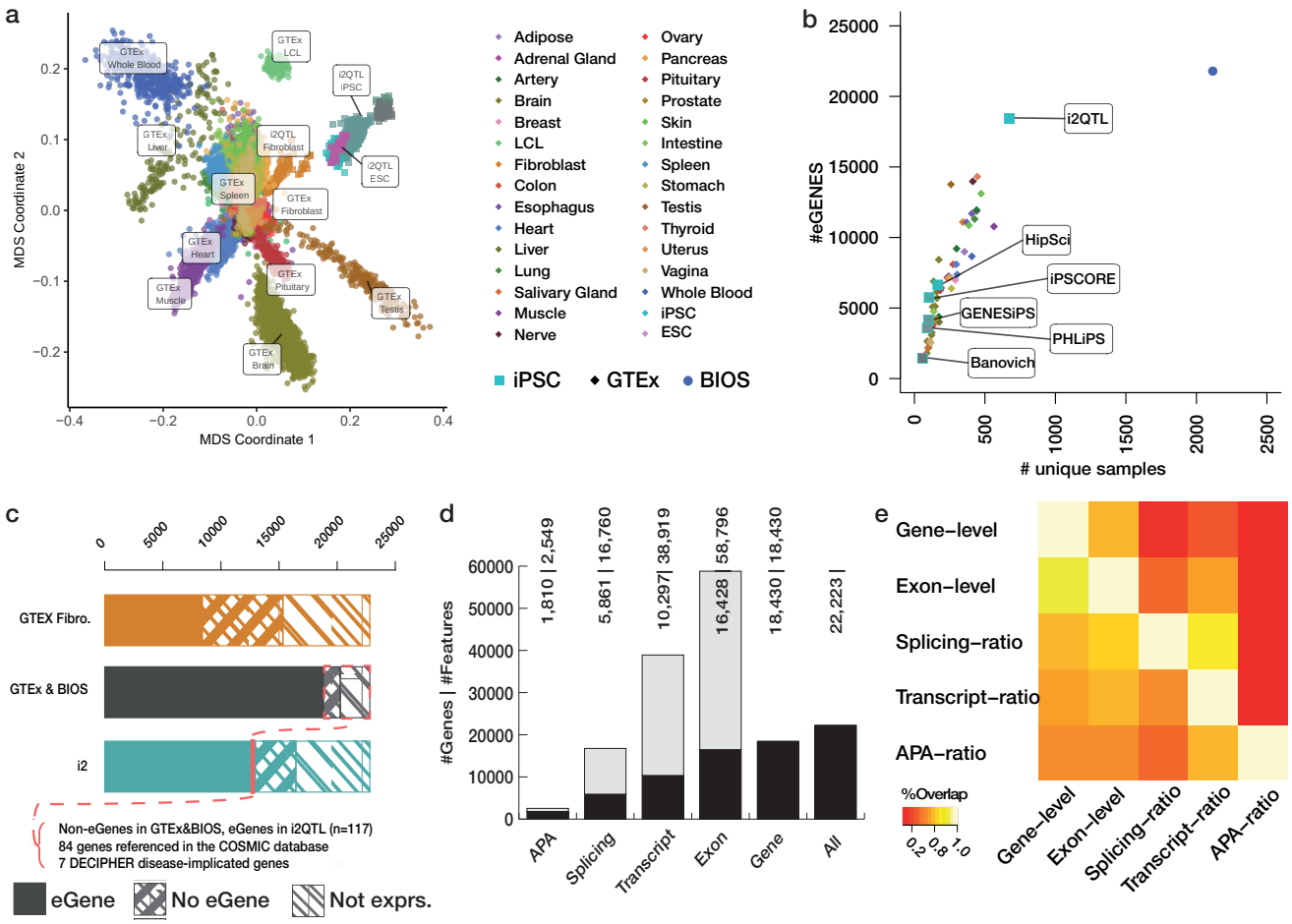
a
MDS Coordinate 1
MDS Coordinate 2
GTEx Whole Blood
GTEx LCL
i2QTL iPSC
GTEx Liver
i2QTL Fibroblast
i2QTL ESC
GTEx Spleen
GTEx Fibroblast
GTEx Heart
GTEx Pituitary
GTEx Muscle
GTEx Testis
GTEx Brain
Adipose
Adrenal Gland
Artery
Brain
Breast
LCL
Fibroblast
Colon
Esophagus
Heart
Liver
Lung
Salivary Gland
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Skin
Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
iPSC
ESC
iPSC
GTEx
BIOS
b
#eGENES
unique samples
i2QTL
HipSci
iPSCORE
GENESiPS
PHLiPS
Banovich
c
GTEX Fibro.
GTEx & BIOS
i2
Non-eGenes in GTEx&BIOS, eGenes in i2QTL (n=117)
84 genes referenced in the COSMIC database
7 DECIPHER disease-implicated genes
eGene
No eGene
Not exprs.
d
#Genes | #Features
1,810 | 2,549
5,861 | 16,760
10,297| 38,919
16,428 | 58,796
18,430 | 18,430
22,223 |
APA
Splicing
Transcript
Exon
Gene
All
e
Gene–level
Exon–level
Splicing–ratio
Transcript–ratio
APA–ratio
%Overlap

Figure 1. Discovery of *cis* genetic regulation across a comprehensive set of RNA phenotypes. A. Comparison of gene expression profiles of iPSC vs the GTEx (v7) tissues. Shown are the first 2 MDS components derived from gene expression levels. **B**. Comparison of the number of discovered eGenes as a function of sample size for existing iPSC studies, GTEx (v7) and BIOS. **C**. Comparison of eQTL coverage for protein-coding genes in Fibroblasts (orange), the combination of all GTEx tissues & BIOS (black) and i2QTL (blue). The fraction of novel protein coding eGenes as compared to BIOS & GTEx is shown in red (see Results), disease association is shown at the bottom. **D**. Breakdown of the identified *cis*-QTL for different expression phenotypes. For each phenotype category, shown are the number of associations at feature-level (grey), depending on the phenotype gene, transcript, exon, intron-exon junction or APA site, and the number of corresponding genes (black). On top the number of assessed eGenes and eFeatures (#eGenes | #eFeature). **E**. Pairwise replication of genetic signals between expression phenotypes, showing the fraction of replicating QTL. *cis*-QTL are discovered at FDR 5% (row info) and replication is performed at FDR 10% (column info).

Mapping large-effect rare variants impacting gene expression

Rare coding and non-coding variants are associated with comparatively larger effects on gene expression than common variants[16,17], which contributes to increased disease risk[18,19]. As the frequency of variants that can be assessed using QTL mapping is limited by sample size, we sought to identify the landscape of rare genetic variants linked to outlying expression profiles of individual genes within iPSCs.

We considered the subset of lines with both RNA-seq and WGS available (N = 425 cell lines after filtering - Methods, **Table S2**), and focused on variants within, or up to 10 kb around, protein-coding and long non-coding RNA genes. Gene expression outliers were defined as the sample with minimum gene expression Z-score (Z-score < -2, under-expression outlier) or maximum gene expression Z-score (Z-score > 2, over-expression outlier) for a given gene. We computed variant burden scores, equivalent to relative risk, across genes and assessed differences in variant burden in outliers compared to non-outliers, aggregating across gnomAD MAF thresholds (both common and rare) and deleterious annotation (CADD - where available). This score was computed separately for gene-level under-expression outliers and over-expression outliers. Notably, for both SNPs and indels, we observed an enrichment in under-expression outliers for rare (gnomAD MAF 0<MAF<=0.01%), highly-deleterious (CADD>25) variants (5-fold and 40-fold increase in comparison to non-outlier in SNP and indels respectively) (**Fig 2A**) (**Table S8**). For structural variants (SVs), a 9-fold increase in rare (study MAF 0<=MAF<1%) duplications, and 18-fold increase in rare multi-allelic copy number variants (mCNVs), was observed in over-expression outliers compared with non-outliers (**Fig 2B**) (**Table S8**). The findings for SVs agrees with recent consortium findings detailed previously[20].

To further identify the extent to which the frequency of a variant in the studied population is associated with outliers compared to non-outliers, we computed enrichment scores for singleton SNPs (where a singleton is defined as a SNP not observed in gnomAD (**Fig 2C**)). In this way, the full frequency spectrum of SNPs could be studied, from ultra-rare to common. The analysis was repeated for increasing outlier stringency, from Z-score -2 to -6. Singleton, high-CADD SNPs (CADD>25) were found at a 12-fold higher rate in under-expression outliers than non-outliers at a Z-score of Z <- 2 (up to 60-fold when Z < -6). The estimate for singleton, high-CADD (CADD>25) SNPs (12-fold increase) is substantially higher than that observed for the next variant bin at the same CADD and Z-score threshold (5-fold increase) (**Fig 2A**).

To contextualize these enrichments with somatic tissues, we repeated the outlier analysis for singleton, high-CADD (CADD>25) SNPs in the 35 tissues from GTEx (v7) with 50 or more samples remaining after PEER factor correction[21] and global outlier removal (see Methods) (**Fig 2D**). To control for variation in sample size across tissues, we calculated the enrichment across 10,000 random draws without replacement of equal sample size for each tissue (N samples = 50). In this comparison, iPSC had the greatest enrichment score (median ~9), followed by GTEx fibroblasts (enrichment ~7) and GTEx testis (enrichment ~5) (**Table S9**). This enrichment was moderately correlated with number of genes expressed in each tissue (Pearson correlation = 0.34); however, when randomly subset to a fixed number of genes across tissues, the maximum median enrichment was still observed for iPSC compared to GTEx tissues (**Sup Fig 6**).

Motivated by the hypothesis that rare variants are driving observed gene outlier effects, we chose the class of variants found to have the largest burden scores (i.e. gnomAD singletons, CADD>25 SNPs) and selected the intersection of these variants present in both GTEx and i2QTL samples (N variants = 305) to assess the relative discovery of outlier effects in genes linked with a shared set of rare, deleterious variants. In total, 38 SNPs were identified in outlier samples in iPSC, with the next closest being GTEx skeletal muscle and lung outlier samples (N variants = 15) (**Table S10**). Furthermore, 20 of the 38 SNPs identified in iPSC outlier genes were observed in iPSC but not in any of the 35 GTEx tissues selected (i.e. the variants were observed in outlier samples in iPSC only). In GTEx, these variants were associated with an overall median absolute Z-score across tissues of 0.49 (range 0.27-1). Four of these i2QTL-only outlier genes have Developmental Process annotations in Gene Ontology (GO), and six have known OMIM disease associations (**Table 1**).

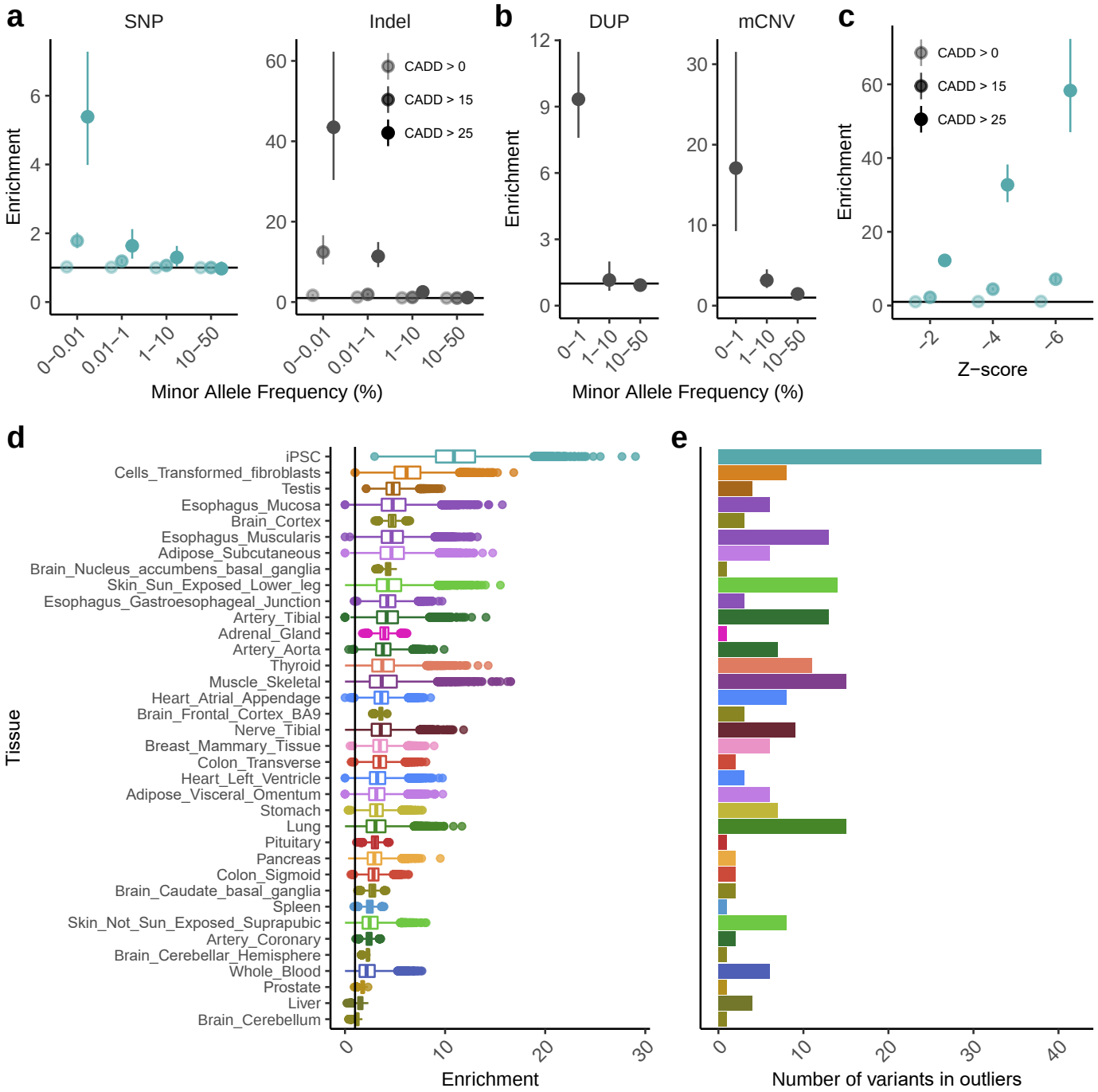
a
SNP
Indel
b
DUP
mCNV
c
CADD > 0
CADD > 15
CADD > 25
Enrichment
0–0.01
0.01–1
1–10
10–50
Minor Allele Frequency (%)
0–1
Z−score
d
e
Tissue
iPSC
Cells_Transformed_fibroblasts
Testis
Esophagus_Mucosa
Brain_Cortex
Esophagus_Muscularis
Adipose_Subcutaneous
Brain_Nucleus_accumbens_basal_ganglia
Skin_Sun_Exposed_Lower_leg
Esophagus_Gastroesophageal_Junction
Artery_Tibial
Adrenal_Gland
Artery_Aorta
Thyroid
Muscle_Skeletal
Heart_Atrial_Appendage
Brain_Frontal_Cortex_BA9
Nerve_Tibial
Breast_Mammary_Tissue
Colon_Transverse
Heart_Left_Ventricle
Adipose_Visceral_Omentum
Stomach
Lung
Pituitary
Pancreas
Colon_Sigmoid
Brain_Caudate_basal_ganglia
Spleen
Skin_Not_Sun_Exposed_Suprapubic
Artery_Coronary
Brain_Cerebellar_Hemisphere
Whole_Blood
Prostate
Liver
Brain_Cerebellum
Enrichment
Number of variants in outliers

Figure 2. Linking rare variants to gene expression outliers. A. Enrichment for increasingly deleterious rare SNPs and indels in gene expression outliers; **B.** Enrichment for rare duplications and mCNVs in gene expression outliers. **C.** Enrichment for singleton, deleterious SNPs in gene expression outliers (enrichment increases in more extreme under-expression outliers). **D.** Comparing enrichment for singleton, highly-deleterious (CADD > 25) SNPs in iPSC (panel C, Z-score = -2, dark-green point) against GTEx v7 tissues (adjusting for difference in sample size; Methods). Enrichment score from 10,000 permutations. Enrichment is strongest in iPSC. **E.** Highest proportion of outlier effects for shared set of rare, deleterious variants (i.e. variants passing thresholds and found in both i2QTL iPSC and GTEx) are in iPSC.

Table 1. Genes linked to singleton, highly-deleterious SNPs with transcriptome outlier effects in pluripotent cells only

Gene	GO Annotation (developmental process)	OMIM Disease Association	Absolute Z-Score (i2QTL)	Maximum Absolute Z-Score (GTEx)
CDCA7	-	Immunodeficiency-Centromeric Instability-Facial Anomalies Syndrome 3	2.22	1.76
DUOX2	Cuticle development	Thyroid Dyshormonogenesis	2.66	0.62
ELMOD3	-	Deafness, Autosomal Recessive 88	2.84	0.90
PRSS12	Nervous tissue development	Mental Retardation	2.57	1.85
RASGRP2	Regulation of cell growth	Bleeding disorder	2.47	1.44
TRIP11	Connective tissue development	Achondrogenesis	3.50	0.84

Distal genetic effects linked to gene expression

Next, we leveraged all samples from European ancestry (samples N=1,123, donors N=759, Table S2 see Methods) to probe for downstream regulatory effects that act in *trans* (defined as >2.5 mb away from the gene body). The identification of *trans*-eQTL is challenging due to the large number of tests and the typically smaller effect sizes, and hence only feasible in large cohorts. To mitigate the multiple testing burden we considered protein-coding genes with an average TPM>1 (16,451 genes) and *cis*-QTL variants (N=93,146), originating from the *cis*-QTL maps described above, as well as known GWAS variants ((N=23,798) NHGRI-EBI GWAS catalog[22] v92), considering a total of 115,709 unique variants in our analysis.

Genome-wide, this identified 193 independent *trans*-eQTL comprising 191 unique genes (FDR<10%; permutation-based, **Table S11 & S12**, Methods). The majority of *trans*-eQTL were exclusively linked to *cis*-QTL variants (163, 85.5%), with 21 *trans*-eQTL linked to both GWAS and *cis*-QTL variants and seven *trans*-eQTL exclusively linked to a GWAS variants. Notably, 138 of the 184 *cis*-QTL with *trans* effects had *cis* effects originating from more than one gene phenotype in *cis*. The majority of the effects that are not linked to gene-level *cis*-QTL are linked to transcript-ratio *cis*-QTL (33 of the 46). Notably, 46 *trans*-eGenes (24% of the *trans*-eGenes) would not have been identified when restricting the analysis to conventional gene-level *cis*-eQTL and/or GWAS variants. Collectively, these results highlight the relevance and the potential for downstream consequences of *cis*-QTL that do not act on a gene level.

To assess the functional relevance of the identified *trans*-eQTL, we focused on four *trans*-eQTL hotspots, defined as *trans*-eVariants, or blocks of *trans*-eVariants within 250Kb, associated with five or more unique genes. The hotspot with the largest number of downstream genes affected was located near *ELF2*, a transcription factor linked to reduced proliferation[23]. The hotspot comprises effects of 16 *cis*-QTL variants affecting *ELF2* and other genes around it; the *cis*-QTL effects could be split-out into three distinct blocks (between block LD < 0.67) (**Fig. 3C**). Variants in the different blocks were linked to multipe *cis*-QTL types mostly on *ELF2*, but also to the APA ratio change of *NAA15* (**Fig. 3D**). Despite moderate LD between these variants, we see a high degree of downstream *trans*-eGenes sharing between the *cis*-eVariants (**Fig. 3E**), and no effect exclusively mapped to *NAA15*. The 37 downstream genes were enriched for sequence motifs from the ELF transcription factors, which have a strong sequence resemblance. Significant associations were observed for *ELF3*, *ELF4* and *ELF5* (p-adj: 3.3×10^-7, 6.8×10^-5 and 2.2×10^-6 respectively, g:Profiler[24]), and 20 out of the 37 *trans*-eGenes were also known match the *ELF2* sequence motif. A second hotspot, in *cis* linked to a *CREB3L2* and in *trans* linked to 8 eGenes, was interesting as we observed an enrichment for the reactome pathway: "ER to Golgi Anterograde Transport" (p-adj<5.6×10^-5), *CREB3L2* was previously linked to both the ER and Golgi complex[25].

We identified 28 *trans*-eQTL variants linked to GWAS. Among these discoveries, we found *trans*-eQTL of GWAS variants linked to breast cancer, congenital craniofacial abnormalities, age of menopause and metabolic traits (Supplementary note 1). For example, the most significant *trans*-eQTL in iPSC linked the expression of NBPF14 to a GWAS hit (rs11249433) on breast cancer; NBPF14 was previously identified as one of the top mutated genes in breast cancer[26]. The same GWAS variant was also linked to *cis*-QTL to three genes of which the gene expression (FAM72[27], RP11–439A17.7[28]) or DNA-methylation levels (HIST2H2BA[29]) are themselves directly linked to (breast) cancer. Another interesting example was found for cleft palate, cleft lip and hemifacial microsomia, for which we identified a set of four variants (rs11072494, rs2289187, rs10459648, rs6495117) that are linked to six downstream *trans*-eGenes (*GPR160*, *SEMA3A*, *MDFIC*, *GRIK2*, *FAM169A* and *GLB1L3*). Among these genes we observed a significant enrichment for the JNK cascade and the MAPK cascade biological processes (g:Profiler p-adj: 1.0×10^-2, 2.1×10^-2 respectively) which are themselves implicated in congenital craniofacial abnormalities[30,31].

To assess the replication of *trans*-eQTL we used the remaining iPSCs from non-European samples (N=187), in which 17% of the effects could be replicated (nominal P<0.05 and matched

effect direction; Methods); a replication rate that exceeds chance expectation (median replication rate 5.3% assessed by testing random *trans*-eQTL pairs (see Methods)). We also sought to replicate our trans association using existing data, however could not directly replicating events in either blood[2] or the GTEx[5] *trans*-eQTL maps. However, we did identify a set of 12 *trans*-eQTL SNPs that are in strong LD (LD>0.8) with previously published *trans*-eVariants, here affecting different genes (**Table S12**). This result is consistent with reports from GTEx showing the difficulty to replicate *trans*-eQTL effects across tissues[5]. To further assess the robustness of the *trans*-eQTL, we tested the cross-omic replication of *trans*-eQTL in DNA-methylation data, available on a subset of the samples (N=572 donors, N=841 lines), finding evidence for cross-omic replication for 27% of the identified *trans*-eQTL in *trans*-meQTL and for DNA-methylation (**methods**). This was higher than expected by chance with median replication rate of *trans*-meQTL: 15.3 (see Methods). When combined, we find replication evidence for 38% of the identified *trans*-eQTL. Within the hotspots we observe increased replication 45.3% versus 25.6%, Fisher exact P-value 0.01 (**Table S12**).

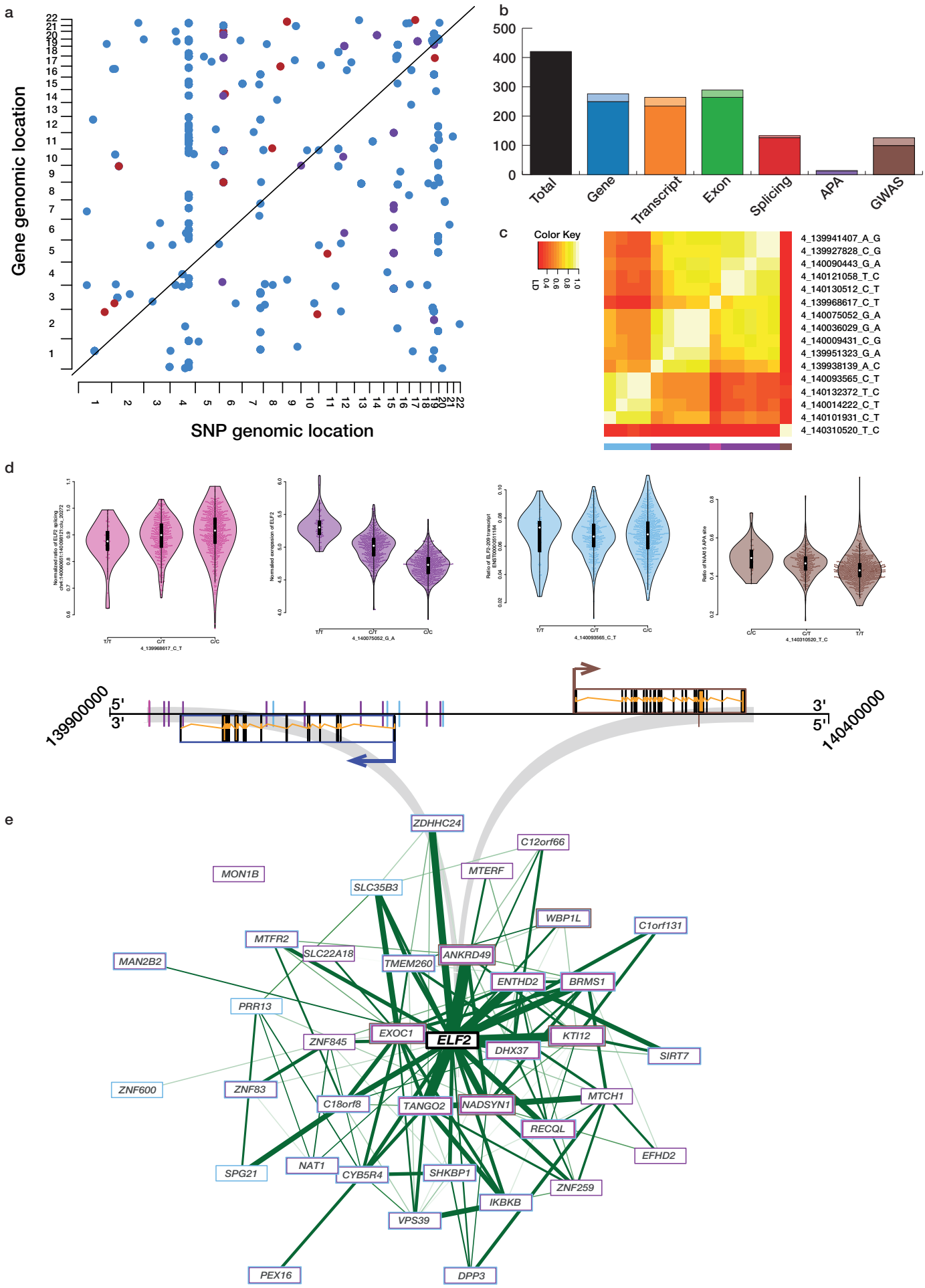
a
Gene genomic location
SNP genomic location
b
Total
Gene
Transcript
Exon
Splicing
APA
GWAS
c
Color Key
LD
4_139941407_A_G
4_139927828_C_G
4_140090443_G_A
4_140121058_T_C
4_140130512_C_T
4_139968617_C_T
4_140075052_G_A
4_140036029_G_A
4_140009431_C_G
4_139951323_G_A
4_139938139_A_C
4_140093565_C_T
4_140132372_T_C
4_140014222_C_T
4_140101931_C_T
4_140310520_T_C
d
e
ZDHHC24
C12orf66
MON1B
MTERF
SLC35B3
WBP1L
C1orf131
MTFR2
SLC22A18
ANKRD49
MAN2B2
TMEM260
ENTHD2
BRMS1
PRR13
EXOC1
ELF2
KTI12
ZNF845
DHX37
SIRT7
ZNF600
ZNF83
C18orf8
TANGO2
NADSYN1
MTCH1
RECQL
EFHD2
SPG21
NAT1
CYB5R4
SHKBP1
ZNF259
IKBKB
VPS39
PEX16
DPP3

Figure 3. Discovery of distal genetic regulation on gene expression. A. Dot plot of the significant independent *trans*-eQTL in iPSC (FDR<10%). Each dot corresponds to one *trans*-eQTL, colored by variant category (*cis*-QTL in blue, GWAS variant in red, shared effects in purple). **B**. Total number of *trans*-eQTL per variant category, the numbers shown are the total number of *trans*-eQTL effects. The darker shaded colors denote *trans*-eQTL linked to multiple variant types, the lighter shades show the number of unique effects for that category of variant. **C-E** Exploration of the *trans*-eQTL hotspot near *ELF2*. In blue the transcript-ratio effect on *ELF2*, in purple the *cis*-eQTL on EFL2, in pink the splice effect in *ELF2* and in brown the APA effect on NAA15 in the same region. **C.** LD between the multiple included variants around *ELF2*. **D.** An illustration of the *ELF2* (left) and NAA15 (right) gene regions, including the four top *cis*-QTL for the four effects we separate in this hotspot. SNPs linked to effects in the regions are shown on the illustration colored based on the effect they tag. **E**. The co-expression network of the genes that are linked to the *trans*-eQTL linked to the hotspot around *ELF2*, including *ELF2* (in the center). Highlighting of the genes names correspond to the effect a gene is linked to. Genes can be linked to multiple effects, this is depicted by multiple colored rings around the gene name. (Color is matched to the effects depicted in D).

Colocalization of GWAS signals and molecular QTL

A major opportunity provided by eQTL maps in pluripotent cells is the identification of disease-associated variants that colocalize with eQTL, providing evidence for a developmental or transient mechanism. Using a combination of FINEMAP[32] and eCAVIAR[33] (see Method), we systematically assessed colocalization between the presented *cis*-QTL map, comprised out of QTL on the five QTL types, and a broad range of previously reported genetic associations obtained from diverse GWAS studies (N=350), collected from the Phenome Scanner Database V2[34], the NHGRI-EBI GWAS catalog[22], GWAS curated in LocusCompare[35], and 1,740 traits from the UKBB phase 1 GWAS[36].

This identified a total of 4,336 significant colocalization events (Methods), linking 608 diseases and phenotypes to 10,794 *cis*-QTL from the full range of expression phenotypes (**Table S13**). The colocalization events covered a broad range of traits and disease, with the largest number of colocalizations linked to physical traits, such as height and BMI, lab-measurements, like blood-cell composition traits, and diseases, such as Crohn's disease, and coronary artery disease (CAD) (**Fig 4A**). Importantly, when restricting our analysis to gene-level eQTL, over 40% of colocalization events are lost. The largest number of eQTL-type specific colocalization events were observed for the exon-level QTL, followed by colocalization events with gene-level and transcript-ratio QTL(**Fig 4B**). As an illustration for CAD, 36 out of 93 GWAS loci were found to colocalize with an iPSC QTL, with at least one QTL-type specific colocalization event that would not be detected by any of the remaining QTL types (**Sup Fig S7**). Next to the *cis*-QTL colocalization we also colocalized the *trans*-eQTL with the GWAS loci, and found amongs others significant overlaps between the *ELF2* and *CREB3L2* hotspots with height (**Supplementary note 1**; **Table S14**).

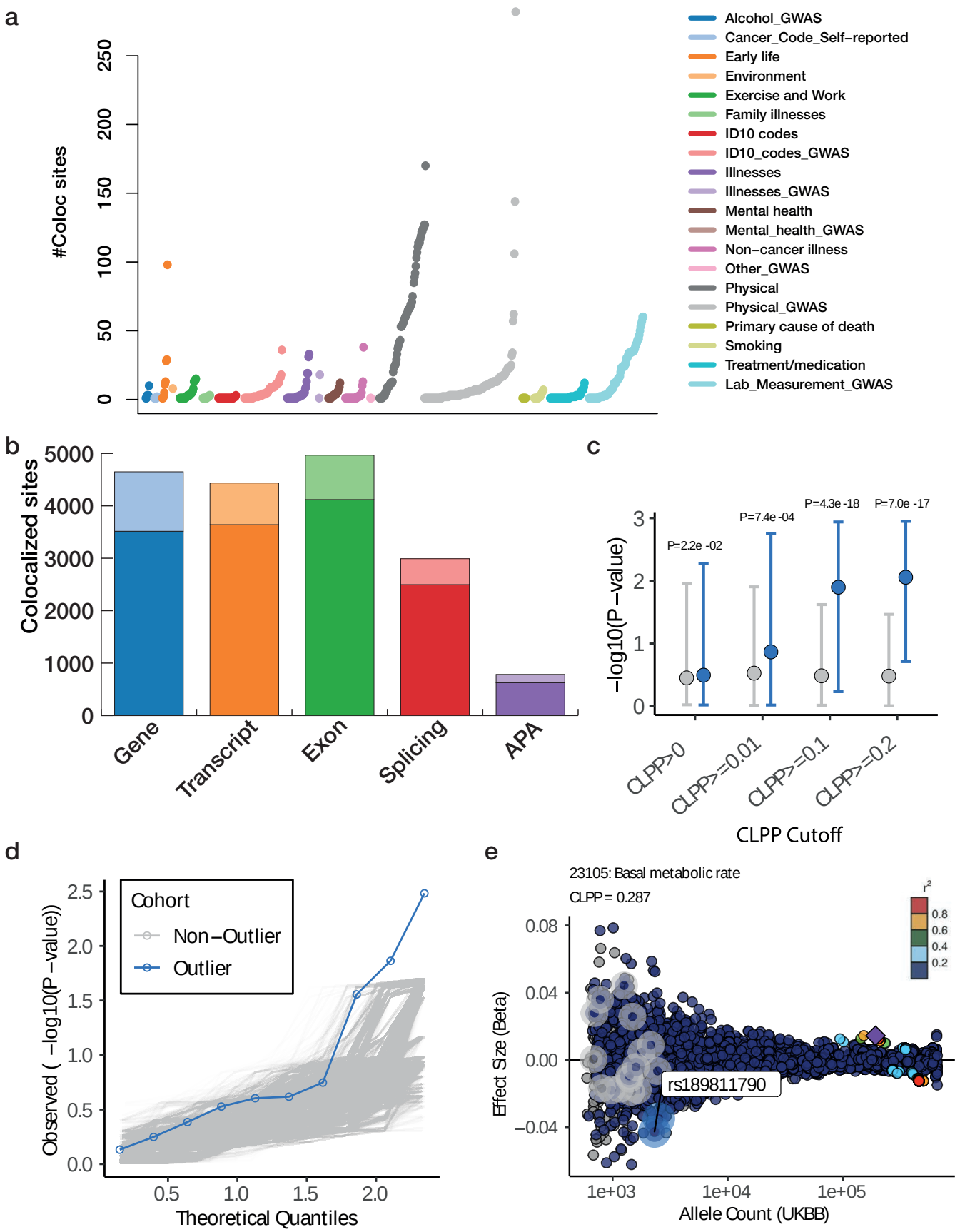
a
#Coloc sites
0
50
100
150
200
250
Alcohol_GWAS
Cancer_Code_Self−reported
Early life
Environment
Exercise and Work
Family illnesses
ID10 codes
ID10_codes_GWAS
Illnesses
Illnesses_GWAS
Mental health
Mental_health_GWAS
Non−cancer illness
Other_GWAS
Physical
Physical_GWAS
Primary cause of death
Smoking
Treatment/medication
Lab_Measurement_GWAS
b
Colocalized sites
0
1000
2000
3000
4000
5000
Gene
Transcript
Exon
Splicing
APA
c
−log10(P−value)
0
1
2
3
P=2.2e -02
P=7.4e -04
P=4.3e -18
P=7.0e -17
CLPP>0
CLPP>=0.01
CLPP>=0.1
CLPP>=0.2
CLPP Cutoff
d
Observed (−log10(P−value))
0.0
0.5
1.0
1.5
2.0
2.5
Cohort
Non−Outlier
Outlier
0.5
1.0
1.5
2.0
Theoretical Quantiles
e
23105: Basal metabolic rate
CLPP = 0.287
r2
0.8
0.6
0.4
0.2
Effect Size (Beta)
0.08
0.04
0.00
−0.04
rs189811790
1e+03
1e+04
1e+05
Allele Count (UKBB)

Figure 4. Colocalization of disease and traits variants with iPSC QTL and the link to expression outliers. A. Overview of the number of colocalization events for each GWAS, colors correspond to trait categories. **B**. Colocalization events for each *cis*-QTL type, displaying the number of GWAS loci with colocalization events that are specific to a given QTL type (light colour; not detected by any other QTL type) or shared with at least one other QTL type (dark colour) which were affected. **C**. Mean P-values (+/- 95% CI) for iPSC outlier- and non-outlier-associated variants available in UKBB GWAS results when stratifying by colocalization posterior probability (CLPP) score. Outlier-associated variants had overall greater effects in the GWAS in which they are also colocalized, compared to matched non-outlier variants, increasing with colocalization probability. **D**. Plotting outlier- (blue line) and non-outlier-associated (gray lines) variant p-values for basal metabolic rate (UKBB GWAS ID: 23105) against a uniform distribution. Non-outlier-associated variants were chosen randomly to match the number of outlier-associated variants (repeated 10,000 times). Effects of outlier-associated variants were more significant than any non-outlier combination sampled. **E**. Example of a gene locus showing outlier variants (blue highlight) with the greatest protective effect sizes among all outlier and non-outlier samples (gray highlight). Points are colored by LD (1000 Genomes) relative to lead variant (top P-value) (purple diamond) in gene locus.

As some GWAS, such as BMI or height, had many GWAS hits reported, we examined the fraction of GWAS loci that colocalize with QTL. Here, we limited our analysis to traits with at least ten loci tested for colocalization. The highest relative overlap was found for primary biliary cirrhosis (PBC), followed by triglyceride (TG) levels. For PBC, 10 out of the 12 reported GWAS loci colocalize with *cis*-QTL data. The corresponding genes were enriched for MAPK, NF-kappa B and TNF-R1 signaling pathways (g:Profiler p-adj: 1.1×10^-2, 4.9×10^-2, 5.725×10^-3 respectively), all of which are involved in the immune system. For TG levels, 12 out of 15 reported GWAS loci overlap with iPSC QTL. Enrichment analysis for these QTL shows significant overlap with metabolic pathways (g:Profiler: alpha-Linolenic acid metabolism, p-adj: 4.899×10^-3; Biosynthesis of unsaturated fatty acids, p-adj: 5.729×10^-3, Fatty acid metabolism, p-adj: 2.585×10^-2). Furthermore, even when normalizing by the total number of loci linked to a GWAS trait, many physical traits still appeared as top colocalized traits.

Integration of transcriptomic data with GWAS summary statistics offered the opportunity to understand the distribution of large-effect rare variants across many diseases and traits. To assess the impact of rare variants on GWAS traits we focused on rare variants linked to iPSC outlier- and non-outlier samples which were available in UKBB Phase 1 GWAS. Given the sample size of large scale GWAS, such as UKBB, rare variants become testable in GWAS. Here we integrated information on gene colocalization and gene outlier gene-expression levels to annotate function to rare variants, not targeted by QTL mapping. We first selected the variants observed within one individual in i2QTL only; for outlier-associated variants, this has the effect of isolating the set of variants putatively driving observed outlier gene expression (i.e. if the same variant was found in both an outlier and non-outlier individual, this would suggest that the variant is less likely to be causing the observed expression aberration). Furthermore, we filtered for variants with a MAF of 0<MAF<=1% from gnomAD and linked to genes with a colocalization CLPP score > 0. For each outlier variant, we selected a non-outlier variant linked to the same gene and with the condition that the non-outlier variant have a CADD score in a range +/- 5 compared to the CADD score of the outlier variant. Given that UKBB Phase 1 GWAS used imputed genotypes from common variants measured via microarray for rare variant calling[37], the intersection of variants was considerably smaller than the set of donor-specific rare variants in i2QTL. After intersecting these datasets, we found 10,103 outlier- and non-outlier associated variants linked to 779 genes.

From this set of variants, we investigated genes with at least one outlier and one non-outlier variant in/near the gene (N traits = 543) and compared the GWAS p-values from outlier- and non-outlier-associated variants at different colocalization CLPP cutoffs (**Fig 4C**). Genes that colocalized with the QTL results suggest that variation in the gene locus is involved in modulating risk for a given trait. As colocalization CLPP score was increased from CLPP>0 to CLPP>=0.2, we observed overall greater GWAS p-values for outlier-associated variants than non-outlier-associated variants (CLPP>0 P=0.02; CLPP>=0.2 P<1e-16), a trend which was also observed for the effect size estimates (data not shown). For traits with >=10 outlier-associated variants in the overlapping set, variant GWAS p-values were plotted in a quantile-quantile plot to assess how likely the distribution of outlier-associated variant p-values would be observed from a matched set of non-outlier variant p-values. For example, for basal metabolic rate (UKBB GWAS ID: 23105) we observed an inflation in outlier variant p-values compared to 10,000 random, matched selections of non-outlier variants (**Fig 4D**). The top hit was in gene *HSD17B12*, a gene known to be involved in type 2 diabetes mellitus[38], and with a CLPP score of 0.29 in the present study (**Fig 4E**). Notably, the outlier variant has one of the largest effect sizes in a 1Mb locus, but the p-value does not pass a statistical significance threshold in conventional GWAS analysis. Overall, we find 48/543 (8.8%) traits with evidence of colocalization (CLPP>=0.01), comprising 58 unique outlier-associated variants.

Discussion

Access to molecular QTL in pluripotent cells provide insights into the spectrum of genetic effects with trait-relevance that can manifest during development and cell differentiation. To maximize the power to identify such QTL, we harmonized existing population-scale iPSC genetic and transcriptome datasets across five studies, and generated additional data. The scale of our resource, with close to a thousand individuals, has enabled the generation of a comprehensive map of *cis*-QTL across a broad range of expression phenotypes, but also *trans*-eQTL and rare variant effects that are typically power-limited in smaller studies.

We identified unique *cis*-QTL on each of the five gene phenotypes we considered in our genetic analysis, finding *cis*-QTL for a total of 58% of all quantified genes (N=18,556). This includes 994 *cis*-eGenes that were not previously detected in other tissues and 3,118 *cis*-QTL genes with QTL that are specific to ratio- or exon-level QTL effects. Next to the *cis*-QTL we identified 193 *trans*-eQTL; 24% of the *trans*-eQTL were driven by ratio- or exon-level *cis*-eVariants, variants that are mostly not considered for *trans*-eQTL. Even when a gene-level effect was observed for a *trans*-eVariant, we observe that 91 (out of 139) of these *trans*-eQTL are also linked to ratio *cis*-eQTL. This shows the existence of *trans* effects downstream of *cis*-QTL that do not affect gene expression level but instead act on other gene expression phenotypes.

To assess the impact of rare variants, we analyzed their role in gene expression outliers. Like previous studies in somatic tissues[39], rare SNP, indel and SV variants were enriched in outliers compared to non-outliers. By comparing iPSC against 35 GTEx tissues, we observed that iPSC provides the greatest power for detecting aberrant gene expression associated with rare, highly-conserved SNPs. Furthermore, when looking at genes where outlier effects were detected in iPSC but not in any GTEx tissue, we find a unique subset of genes important to developmental processes and linked in OMIM to rare developmental disorders, such as the gene *PRSS12* (mental retardation)[40], *RASGRP2* (bleeding disorder)[41] and *DUOX2* (thyroid dyshormonogenesis)[42].

The large scale QTL maps on the diverse set of gene expression phenotypes made it possible to generate an extensive colocalization map between the QTL effects and GWAS loci. We were able to annotate over 4,400 GWAS implicated loci (out of the 29,666 assessed loci), originating from over 600 traits, to expression changes. We observe unique colocalized loci to each of the *cis*-QTL types and a wide range of traits, from physical traits to diseases and lab-measurements. Over 40% of the GWAS colocalized QTL loci are found not to be linked to gene-level eQTLs, highlighting the importance of effects beyond gene expression level. Interestingly, we also observed colocalizations for two of the *trans*-eQTL hotspots; namely, the hotspots around *ELF2* and *CREB3L2*, which both colocalized with height GWAS signals. Furthermore, we found colocalization for developmental traits, such as congenital craniofacial abnormalities, and cancers. Lastly, by integrating our colocalization results and rare variants linked to expression outliers, we demonstrated prioritization of variants with large impacts on phenotype in the UK Biobank, thereby providing a means for understanding the spectrum of large-effect rare variants associated with a wide range of complex traits and diseases.

Overall, we present the most comprehensive map of genetic regulation of gene-expression in iPSC to date, and demonstrate the use of this map by generating large-scale colocalization between GWAS and iPSC QTL. Here we were able to link both the QTL map as well as the outlier linked rare variants to trait-associated genetic variants, and observe relevance for developmental traits, physical traits and cancer, among multiple others. Large-scale characterization of iPSC presented herein, combined with population-scale cohorts on differentiated tissues previously published[2,4,5], allows for detailed comparisons of the molecular functions of genetic variants throughout different stages, from a pluripotent to terminally-differentiated tissues. These complementary datasets, will be invaluable in ongoing research efforts such as gene therapies, by showing a more complete context in which genetic variants and genes function.

The genetic maps and colocalization catalogs generated by the i2QTL consortium form a unique reference catalog, further aiding interpretation of risk variants in a unique cell type relevant for both development, cellular differentiation and cancer research. We expect that the genetic maps presented here, in ongoing combination with GWAS results will reveal the molecular underpinnings of genetic disease and traits manifesting during early-development.

Online Methods

Dataset information

Within the “Integrated iPSC QTL” (i2QTL) consortium we reanalyzed iPSC data originating from five different studies. Below a short description is given on each of the analyzed studies, in **table S1** the references to the data sources are given.

HipSci

The Human Induced Pluripotent Stem Cells Initiative (HipSci)[9,43] was set up to generate a large, high-quality reference panel of human iPSC lines for the research community. Within the initiative a large population based iPSC panel was formed, next the population based inclusion there was a special inclusion for several diseases: Monogenic diabetes, Bardet-Biedl syndrome, hereditary cerebellar ataxia, hereditary spastic paraplegia, kabyki syndrome, usher syndrome and congenital eye defects, congenital hyperinsulinia, alport syndrome, hypertrophic cardiomyopathy, primary immune deficiency, bleeding and platelet disorders, macular dystrophy, retinitis pigmentosa, batten disease and childhood nerolog diseases. iPSC lines were generated through a non-integrative methodology (Sendai virus) using either fibroblasts or blood as a starting population. All of the lines were genotyped using the Illumina beadchip HumanCoreExome-12 genotyping chip. RNA-sequencing was performed, using either a pair-end stranded protocol, or a single end protocol, followed by Illumina sequencing. For a subset of the samples whole genome sequencing data was generated. Next to transcriptomic and expression data DNA-methylation information was generated using the Illumina 450K or Illumina EPIC array. For QC information on embryonic stem cells and the donor material (i.e. fibroblast or blood), was generated on a subset of the lines. In this study, we considered data from N=543 donors. A subset of (N=166) has previously been described in Kilpinen et al. More information about data generation, which is the same for the entire consortium, can be found in the original publications Kilpinen et al and Streeter et al.

iPSCORE

The iPSCORE project[6,44,45] was completed with the goal of assembling high quality iPSC lines derived from hundreds of individuals, and profiling them with an array of genomics assays, to provide a resource for studying functional genetics in the context of derived human cell lines. For 273 iPSCORE individuals, deep whole genome sequencing (median 48x) was performed and additionally 215 of these individuals had iPSCs created by reprogramming fibroblasts, on which RNA-sequencing and chip genotyping using the Illumina MEGA array was also performed. Notably, some iPSCORE individuals are related as part of families of 2-14 subjects, while 167 are unrelated. Further information about this study, its constituents, and the generation of the sequencing data can be found in several previous publications (Panopoulos et al., DeBoever et al., D'Antonio et al,).

GENESiPS

The original GENESiPS study included 201 subjects with insulin sensitivity measurement performed by a modified insulin suppression test in accordance with Knowles et al[46]. The aim of the study was to generate an iPSC library reflecting the broad spectrum of insulin sensitivity in human populations. iPSC lines were generated through a non-integrative methodology (Sendai virus) using erythroblasts as starting population. iPSC grown under feeder-free conditions from passage 8-11 were used for RNA-sequencing on the Illumina HiSeq 2500 system with 100 nucleotide single-end reads. Further information can be found at Carcamo-Orive et al[10].

PhLiPS

The PhLiPS Study (Phenotyping Lipid traits in iPS-derived hepatocytes Study) aimed to create a library of iPSC lines and iPSC-derived hepatocytes of diverse genotypes for metabolic profiling and lipid trait genetic screening. Detailed methods and data descriptions can be found elsewhere Pashos et al[7]. As a part of the Next Generation Genetic Association Studies (Next Gen) program, PhLiPS ascertained 91 subjects who were free of cardiovascular disease and in generally good health. Peripheral blood samples obtained from the subjects were used for genome-wide genotyping, blood lipid measurements and for generating iPSC lines. Infinium Human CoreExome-24 BeadChip (Illumina) was used for all sample genotyping. Extracted RNA with a minimum RNA integrity number (RIN) 7.5 were sequenced on HiSeq 2000/2500 systems (Illumina) with paired-end, 100-bp/125-bp read lengths with a target read-depth of 50 million reads per sample.

Banovich

The Banovich et al study[8] investigated the use of iPSCs to study the impact of genetic variation on gene regulation. A panel of iPSC lines were derived from 58 Yoruba (YR) lymphoblastoid cell lines (LCLs), originally generated from the YRI samples within the 1000 Genomes (1000G) projects. LCLs were reprogrammed using an episomal approach described in Okia et al[47]. RNA-sequencing was performed using fifty-base pair single-end libraries using the Illumina TruSeq kit, sequenced on an Illumina HiSeq 2500. More information on the iPSC generation and RNA-sequencing can be found in Banovich et al. Genotypes were taken from the 1000G samples[48] where the LCLs are originally derived from, for this we focused on the array genotypes to match the other datasets.

Genotyping information

Array based genotyping

Genotyping information for the individual studies was generated as described in the original publications referenced above. For the QTL analyses the genotyping information as generated by genotyping arrays are used and human genome assembly GRCh37 was used. The genotype information was imputed and phased per sample using IMPUTE2 v2.3.1[49] for imputation and SHAPEIT2 v2.r790[50] for phasing. The imputation and phasing was performed using a combined genotyping reference, encompassing the haplotypes from the UK10K cohorts and (1000G) Phase 1 data[50,51]. The imputation was run in chunks of average 5 Mb and 300 kb buffer regions on each side, and used the following MCMC options (-Ne 20000 -k 80) for autosomes. SHAPEIT2 was run without MCMC iteration (-no-mcmc) so that each sample is phased independently. Single-sample VCFs were merged and subsequent QC was performed using Genotype Harmonizer[52] and BCFtools[53]. After imputation each cohort was combined and cohort-, and therefore chip-, wise SNP-QC was performed using Genotype Harmonizer. SNPs were excluded if the call rate was below 90% and imputation quality (Mach R2) was below 0.4. After dataset wide QC we merged the datasets and performed a combined SNP QC filtering for a call rate of 1.0 and an imputation score of 0.4. Finally we were left with 2,533 samples and 7,188,631 variants for which array genotypes were available.

Super population assignment

All imputed samples were assigned to 1000G super-populations by projecting the samples with the genetic PCA based on the full 1000G reference (phase 3) dataset[48]. Next we selected the variants that had a minor allele frequency (MAF) over 5% within 1000G, were available within the imputed i2QTL genotype dataset and were retained after pruning of the remaining variants. We performed a principal component analysis (PCA) on the 1000G samples and projected the i2QTL samples on to the calculated PCs. Next we calculated the distance for each of the i2QTL samples to the centroid value for each super-population as calculated using the 1000G reference, based on the first 40 principal components (**Figure S8**). We chose to use 40 components to predict the super-population, as this was the optimal number based on re-assigning the 1000G samples back to their original super-populations.

Whole genome sequencing

For a subset of the samples, originating from HipSci and iPSCORE, whole genome sequencing data was available. Within the HipSci consortium whole genome sequencing data was available from fibroblasts obtained from 204 donors, many of which also had whole genomes available from derived iPSCs. From the iPSCORE consortium, 273 individuals had whole genome sequencing data available which was generated from blood (254 donors) or fibroblasts (19 donors). Genotype calling on the i2QTL WGS samples is described in detail in Jakubosky et al[54]. In short, we called both SNVs and structural variants within the joint dataset. To do so we: (1) called SNPs and short indels using GATK's best practices for genotype calling; (2) called duplications, deletions, inversions, reference mobile element insertions (rMEI) and other novel adjacencies referred to as "breakends" using SpeedSeq (LUMPY/CNVnator); (3) called duplications, deletions and multiallelic copy number variants (mCNVs) using Genome STRiP; and (4) called mobile element insertions using MELT. After calling the structural variant classes, subsequent quality control and a stitching and merging procedure was used to derive a non-redundant high quality set of variants. In the outlier analyses a total of 34,804,470 SNPs and 4,721,392 indels were called in 928 high-quality samples. 528 of these samples had iPSC RNA-seq.

RNA-sequencing

RNA-sequencing feature quantification

Raw RNA-seq data for the samples of each of the projects, including Choi et al[13] as reference, was collected from EGA, dbGaP and SRA (**Table S1**). Depending on the availability of data from each source, we started the reprocessing with CRAM, BAM or FASTQ files. To ensure uniform processing of all samples, the CRAM and BAM files were first transformed to FASTQ files. The reads in the FASTQ files were trimmed to remove adapters and low quality bases (using Trim Galore![55–57]). Next, trimmed reads were alignment using STAR (version: 020201)[58], using the two-pass alignment mode and the default parameters as proposed by ENCODE (c.f. STAR manual). All alignments were relative to the GRCh37 reference genome and Ensembl 75 genome annotations[59]. Based on the quality controlled aligned reads we quantified gene-level information on all samples. For the paired-end stranded data we also quantified transcript-ratios, exon-level, alternative polyadenylation ratios and splicing-ratios.

Gene-level RNA expression was quantified from the STAR alignments using featureCounts (v1.6.0)[60], which was applied to the primary alignments using the "-B" and "-C" options in stranded mode if applicable, using the Ensembl 75 GTF file. In case multiple RNA-seq runs per iPSC line were generated these were summed to one set of gene-counts per iPSC line. Read counts per iPSC line were normalized for gene length and the edgeR[61] normalized total expression count per sample, yielding corrected transcript per million counts (TPM). Using the same setup exon expression level were quantified. We chose to first do a strand-specific merge of overlapping exons and quantified these meta-exons as one entity. Subsequent normalization is the same as for the feature-count derived gene information.

To facilitate a comparison to GTEx v7[5] a second gene-level quantification was performed. The STAR aligned BAMs were used to quantify gene-level information using RNA-SeQC[62] (v.1.1.8) with the gencode v19 annotation, matched to the GTEx annotation. This limited the quantification differences to only read-QC and alignment. This RNA-SeQC quantification was used only when comparing the i2QTL data to GTEx tissues.

Transcript ratios were quantified by determining transcript isoform levels using Salmon[63] (version: 0.8.2). The Salmon transcript database was built based on transcript information from Ensembl 75. Salmon directly uses the QCed FASTQ data; in the transcript quantification "--seqBias", "--gcBias" and "VBOpt" options were used. The transcript count quantification as returned by Salmon were normalized as described for the gene-level quantification using feature count information. Subsequently, we transformed the transcript levels to ratios per gene, by dividing the count of an individual transcript by the transcript sum per gene.

Alternative polyadenylation (APA) quantification was performed following a similar approach to that used in Zhernakova et al[4]. In brief: first, overlapping 3'UTRs for transcript isoforms from the same gene were merged by taking the union of overlapping regions, such that the set of 3'UTRs for each gene was a set of disjoint genomic regions. Next, these regions were extended by considering the most distal available polyadenylation site for each 3'UTR region, as present in the APADB[64]. For each 3'UTR, the set of alternative polyadenylation sites within that region were identified by taking the union of all annotated APA sites from the APADB, in addition to the annotated end of the 3'UTR given by the Ensembl annotation. These APA sites then subdivide each 3'UTR into a set of windows. The depth of RNA-seq in each window was then computed by using the samtools bedcov function. Finally, the read depth in each window was expressed as a fraction of read depth in the window immediately upstream.

Lastly, we quantified splicing levels using leafcutter[65] (v0.2.8). Quantifications were of splice-events are calculated as described in the leafcutter manual. First the junction reads are extracted from the BAM files, after which leafcutter clusters junction reads into introns. Subsequently the intron counts are transformed into ratios per cluster and we linked the intron locations to genes.

RNA quality control

After feature quantification low quality RNA-seq samples were filtered out by applying filters on both Picard[66] and VerifyBamID[67] quality measures as well as gene expression levels. We defined high quality samples as those with > 15 million reads, > 30% coding bases, > 65% coding mRNA bases, a duplication rate lower than 75%, Median 5‘ bias below 0.4, a 3’ bias below 4, a 5’ to 3’ bias between 0.2 and 2, a median coefficient of variation of coverage of the 1000 most expressed genes below 0.8, a free-mix value below 0.05. Lastly, we removed samples that had low expression correlations (<0.6) to the average iPSC expression values across our study, as measured per chromosome. This resulted in 1,367 iPSC lines derived from 948 donors for analysis, all of which also have genetic information available. Next to this we had 98 samples from Choi et al[13] and HipSci[9,43] which are included as reference (table S2).

After RNA QC the sample linking was validated using VerifyBamID, for each of the RNA-samples the best matching genotype were computed and compared to the expected match. This resulted in the identification of 36 sample mixups (between 0-26 mismatches per study), and identified in total 33 unmatched RNA-seq samples. Some of the mix-ups were corrected (N=34) and others were removed (N=2).

PEER correction and optimization

We correct for transcriptome-wide confounders arising from the merging of multiple datasets, to do so we chose to estimate PEER[21] factors to capture these confounders. We chose not to include any known variation when estimating PEER factors, as meta-data was sparse and not standardized between studies. We ran PEER (v1.3) on the log-transformed gene-level quantifications, on genes that had a TPM > 2. We assessed the impact of the number of PEER factors on eQTL mapping in our data by calculating the number of eGenes (at least one eQTL at FDR < 5%) with incrementally increasing number of PEER factors (**Figure S9**). We chose to use 50 PEER factors to correct for confounders in out data. Assuming *trans*-eQTL impacting many transcripts may be included as latent factors in PEER analyses, we then tested eQTL effects against each of the 50 PEER factors. We did not find any likely *trans*-eQTL-associated PEER factors.

Quantitative trait loci mapping

For the quantitative trait loci mapping (QTL) mapping, both the *cis* and *trans*, we used a linear mixed model implemented in LIMIX[68]. This model allowed to control for both population structure and repeat lines from the same donor using kinship as a random effect component. The kinship matrix was estimated using the identity-by-descent function in PLINK (1.07)[69], on independent variants with a MAF >5%. Fifty PEER factors, derived on the gene-level data were included as fixed effect covariates in all analyses, optimization is described above. The QTL mappings are performed on the log transformed standardized expression levels in case of the gene-level and exon-level data; for the other expression phenotypes, the ratio based phenotypes, we used an arcsin-transformation.

To control for multiple testing we used an approximate permutation scheme as presented in Fast-QTL[70], based on a parametric fit to the null distribution, to adjust for multiple testing across *cis* variants for each gene. Briefly, for each gene, we obtained p-values from 1,000 permutations on the genotypes while keeping covariates, kinship, and expression values fixed. Subsequently, we estimated an empirical null distribution by fitting a parametric Beta distribution to top P-values per gene per permutation round. Using this null model, we estimated *cis* region adjusted p-values for QTL lead variants. To control for the multiple genes-level tests we do per QTL run, we performed the Storey's Q-value procedure[71]. In cases where multiple features per gene where tested, i.e. for transcript-ratio, exon-level, splicing-ratio and APA-ratio *cis*-QTL (herein features), we chose to control the FDR at gene-level, to explicitly account for the number of features per gene. To do so, we applied a Bonferroni correction, for the number of features per gene, on the permutation-based *cis*-SNP corrected eQTL p-values. We then take the lead QTL effect per gene, i.e. the most significant associated variant-feature pair (*cis* region and across features adjusted), and apply the Storey procedure on these selected p-values. After the Storey procedure we find the Q-value (FDR) threshold and report all QTL at the selected level, i.e. potentially more per gene, that pass this threshold.

cis-eQTL Mapping

During the mapping of *cis*-quantitative trait loci (*cis*-QTL), we considered common variants (MAF>1%) in gene-proximal regions of 250k upstream and downstream of gene transcription start and end sites (GRCh37). We limited the analysis to include only paired-end stranded European samples, as determined from the super population assignment, (n=716 donors, n=932 lines) to limit variation between trait discovery and to have the same discovery sample size in all *cis*-QTL maps. We controlled for multiple testing at a gene-level FDR of < 5% for the *cis*-QTL maps.

For the different expression phenotypes, we tried to use matched inclusion criteria. We limited the analysis to features that were expressed, i.e. non-zero expression, in at least 25% of the samples. For the splicing and APA-QTL we required at least 50% of the samples to have a non-zero and non-NA ratio, the assessed genes per QTL type is summarized in **Table S15**.

After identification of the lead *cis*-eQTL for the gene-level effects, we searched for independent eQTL (secondary eQTL and higher order effects). For each of the significant eQTL gene the significant eQTL variant(s) from lower levels were added as a covariate to the model and a new QTL mapping round was run to identify independent *cis*-QTL. This search for independent effects was continued for each gene until no more independent *cis*-eQTL were identified.

Trans-eQTL mapping

During the mapping of *trans*-quantitative trait loci (*trans-e*QTL), we considered common variants (MAF>1%) in gene-distal regions defined as at least 2.5M upstream and downstream of the gene transcription start and end sites (GRCh37). Given the large number of tests in a full *trans*-map, we chose to limit the *trans*-eQTL map to both *cis*-QTL variants and GWAS implicated variants. This limited the number of variants that had to be tested to 115,709. In addition, we only considered protein coding genes with a TPM ≥ 1 (n=17,039) and expressed in at least 25% of the samples. Other settings were matched to those described above. However, to limit the chance of spurious associations we included an extra expression normalization step - we quantile normalized the normalized gene counts to match a normal distribution. During the *trans*-QTL phase we switched to using all samples with European ancestry (n=743, lines=1,120) as we only looked at gene-level effects.

To further limit the chance of spurious associations, specifically those linked to cross-mapping of RNA-reads[72], we set out to black-list combinations of genes a-priori. We created a black-list by using the information contained in the mapping of the RNA reads to the genome, i.e. using primary and secondary genome matches. The gene-level quantifications are based only on the primary mapping of a read but the secondary mappings that do pass the mapping thresholds are informative to identify genes with high sequence similarity. A benefit of this approach is that it is directly based on the reads that we also use for gene quantification and is influenced by the same biases. We built this cross-mapping map using all the paired end stranded data. When a gene combination would be in the black list we also removed the *trans*-window around the secondary gene from the test region. We also excluded the HLA region, due to its complex LD structure.

cis-eQTL annotation

The i2QTL genetic maps where annotated by overlapping the QTL signals with information taken from published QTL maps, gathered from the original iPSC studies[7–10,45], GTEx[5] and BIOS[4].

To assess the replication of the eQTL effects published in the original iPSC studies[7–10,45], we tested for replication of the published eQTL effects at p-value<0.05 and assessed effect direction **(Fig. S1-4)**. The original QTL maps were taken as provided, i.e. replication was assessed based on the published significance levels. When assessing the replication within the iPSCORE study we focused only on the effects attributed to SNPs, and excluded the effects of structural variants. Between 75% and 82% of the effects were assessed in both the original study and the i2QTL study, replication rates of the effects that are between 62% and 86% **(Fig. S1-4)**.

Using the GTEx[5] and BIOS[4] QTL maps we assessed the overlap of eGenes to identify effects specific to iPSC. Effects from the original studies where taken as provided, i.e. FDR 5% as calculated in the studies. Next we assessed replication in terms of eGenes within the individual tissues (**Fig 1B)** as well as looking at eGenes in any of the GTEx tissues or BIOS (**Fig 1B-1C**).

To test for similarity between genetic effects observed at the different gene expression phenotypes, we assessed replication between the signals at the different levels. The overlap was assessed for individual eQTL, i.e. SNP-gene combinations, by mapping the different QTL types to the respective gene of the effect and taking the most significant association (at gene-level FDR) if multiple were tested per gene. If the effect present at the first QTL type was replicated at FDR 10% in a second QTL type this was counted as a replicating effect (**Fig 1E**).

Replication of *trans*-eQTL

Using the expression samples from other inferred super populations (n=186 donors, n=253 lines) **Table S2**, we could assess replication of the *trans*-eQTL. Next to this we used DNA-methylation data available in a subset (n=572 donors, n=841 lines) to do cross-omic replication of the identified *trans*-eQTL.

The replication of the *trans*-eQTL in the left out expression data was performed using the same setup as for the main *trans*-eQTL analysis, with the exception of the PEER factor correction. To not induce effects based on PEER factors, we chose to only correct for the limited known factors across our datasets: inferred ancestry, sequencing type (paired-end (yes/no), stranded (yes/no)) and hot encoded vectors describing the dataset of origin. We deemed *trans*-eQTL as replicated in the left-out samples when the raw p-value of the association was below 0.05 & the effect direction was matching to the effect observed in the *trans*-eQTL map.

The replication within the DNA-methylation data was similarly matched as closely as possible. We started with a joint normalization of the Banovich et al[8] and HipSci[9,43] DNA-methylation arrays. As the data was generated on two different Illumina methylation arrays, the Illumina 450K and Illumina EPIC array, we started with sub-selecting the CpG probes that are present on both. After this we normalized the DNA-methylation profiles, as described in detail Bonder et al[3], based on the DASEN[73] normalization. After joint normalization, we correct the data for the first 20 PCs to account for batch effects in the DNA-methylation data. In order to replicate the *trans*-eQTL in DNA-methylation data, we first had to link CpG probes to genes. We chose to do so by linking a gene to all DNA-methylation probes inside the gene and CpG probes that are within within 250Kb of the gene TSS and TES. Given that we test a multitude of CpG's per *trans*-eQTL gene we determine the number of independent DNA-methylation probes per gene (R2<0.2) and used Bonferroni procedure to correct for the number of independent probes that were tested per gene. We deemed *trans*-eQTL as replicated in DNA-methylation when the CpG corrected replication p-value was below 0.05.

To check if we replicate more effects than expected by chance, we repeated the *trans*-eQTL and *trans*-meQTL replication on random *trans*-eQTL pairs. To try and match the trans-eQTL characteristics, we chose to link the *trans*-eQTL variant to a gene with matched expression characteristics, by selecting the ten genes per *trans*-gene with the closest average and variance levels. The same random pairs were used for the *trans*-meQTL replication.

Outlier analysis

In addition to variation linked to common variants (MAF >1%), rare variants were linked to transcriptomic outliers. Using the featureCount gene quantifications (log TPM), genes were subset to only include autosomal protein-coding and long non-coding RNA genes. Cell lines from donors with predicted ancestry other than European super-population were removed, and only paired-end stranded cell lines were selected. Genes were then filtered for minimal expression, defined as gene expression TPM>0 in 50% or more within each study.

To correct for technical effects, PEER[21] factor correction was run on the filtered data as described above. Following the correction of 50 PEER factors, as chosen based on the *cis*-eQTL PEER optimization, the resulting residuals were scaled and centered (Z-score). Samples were then checked for large gene outlier counts (i.e. cell line is found to be the most under- or over-expressed cell line across hundreds of genes), which suggests possibly incomplete correction for technical effects. Cell lines with expression abs(Z-score)>2 in more than 100 genes were removed from subsequent analysis (N=21 lines). Finally, cell lines were retained only if WGS was available in addition to RNA-seq, leaving data from the HipSci and iPSCORE projects only. After applying these steps, 17,514 genes and 425 cell lines remained in the dataset for further analysis.

To prepare the WGS genotype data SNP and indel variants for the analysis, variants were first filtered based on QVSR tranche 99%. The software vcfanno[74] was used to annotate the WGS VCF with minor allele frequency from gnomAD[75] (version r2.0.2), and CADD score from CADD[76] (version 1.3). Variants were filtered per sample to retain variants with at least one alternate allele only. Variants were then linked to genes using the bcftools[53] window command, with a maximum distance of 10 Kb window size and using the Ensembl 75 GTF as reference. A separate file was produced for each cell line comprised of the following columns: cell line ID; gene ID; chromosome; position; gnomAD MAF; CADD (phred); CADD (raw).

We performed a comparative analysis using GTEx v7 tissues, reprocessing i2QTL data to match the GTEx v7 processing pipeline to limit technical variation. For this specific analysis, RNA-SeQC expression quantification was used, and a separate PEER analysis was run to correct for technical variation, without known factors. As before, the top 50 PEER factors were chosen for i2QTL data. For GTEx v7 tissues with <= 150 samples, 15 PEER factors were used; for tissues with <= 250 samples, 30 PEER factors; for tissues with >250 samples, 35 PEER factors. GTEx v7 WGS variants were annotated with MAF from gnomAD (version r2.0.2) and CADD scores from

CADD (version 1.3) using vcfanno. GTEx v7 tissue samples with expression abs(Z)>2 in more than 100 genes were removed. Tissues were retained only if there was >=50 samples available (N tissues = 35). Residuals resulting from PEER factor correction were centered and scaled to generate expression Z-scores.

Outlier enrichment

An enrichment score was calculated to quantify the difference in proportions of outlier lines with variants across several MAF/CADD bins compared with non-outlier lines. Specifically, enrichment here refers to the relative risk:

$$RR = \frac{\left(\frac{OV}{O}\right)}{\left(\frac{O'V}{O'}\right)}$$

, where OV is the number of outlier lines with >=1 variant in or near (+/- 10 Kb upstream or downstream of gene body) a gene passing given MAF and CADD thresholds, O is the total number of outlier lines, $O'V$ is the number of non-outlier lines with >=1 variant in/near gene passing given MAF and CADD thresholds, and O' is the number of non-outlier lines. The relative risk is reported with 95% Wald confidence intervals derived from the asymptotic distribution of the log relative risk:

$$log(SE) = \sqrt{(1/OV) - (1/O) + (1/O'V) - 1/(O')}$$

The bounds on the confidence interval are defined as follows:

$$maxCI = RR * \exp(1.96 * logSE)$$
$$minCI = RR * \exp(-1.96 * logSE)$$

The analysis was performed separately for SNP, indel, SV variants, and across different MAF bins (from common to rare) and CADD bins (progressively more deleterious variants). Expression outlier direction (i.e. under-expression, over-expression) was tested separately. For example, for under-expression outliers, an outlier line was defined as the least-expressed line in a given gene that also has a Z-score < -2. In this way, a gene can have a maximum of one outlier line per outlier direction. Non-outliers were defined as lines with a Z-score between -1 and 1 for a given gene. Genes were discarded if there was not at least one outlier and one non-outlier line matching MAF and CADD thresholds. The outlier analysis was performed for both i2QTL and GTEx datasets.

Colocalization of GWAS loci with iPS QTL

To choose the combinations of loci, features, and traits at which to test for colocalization, we started with two sets of curated GWAS summary stats: one including UK Biobank (UKBB) rapid GWAS results (for 1,740 traits) (cite Neale et al.), and the other including publicly available GWAS results gathered from the NHGRI-EBI GWAS catalog[22] and PHENOMESCANNER V2[34] for 303 studies, obtained through a wide variety of studies and consortia[35]. For every trait in this set, we selected loci having a GWAS association a p-value of < 5x10^-8 and located at least 1Mb away from all previously selected (more significant) loci for the same GWAS trait. We further narrowed this set of GWAS-relevant loci to with a significant *cis*-QTL, for any of the quantified expression phenotypes, within 10kb of the lead GWAS hit at a gene-level corrected FDR of less than 5%. Due to the vast number of SNPs and traits in the UKBB, we only tested UKBB GWAS hits for colocalization if the lead GWAS hit overlapped exactly with a known eQTL, for computational feasibility. The resulting loci comprised our colocalization test set, to be tested for colocalization with every eQTL phenotype having a significant association nearby. Given the presence of several different types of eQTL and abundant measured features for each of these eQTL types, it was possible for a single GWAS locus to be tested for colocalization with a number of QTL traits originating from single or multiple genes.

We then tested every resulting pair of GWAS locus and eQTL feature in our set. For each of these pairs, we narrowed our summary statistics to the set of SNPs tested for association with both the given GWAS trait and the given eQTL trait, and removed all sites containing less than 5 SNPs after this filter. Using the full 1000 Genomes dataset from phase 3 (2,504 individuals) as a reference population[48], we estimated LD between all of these SNP pairs. We then ran FINEMAP[32] independently on the GWAS and eQTL summary stats to obtain posterior probabilities of causality for each of the remaining SNPs, and we combined these probabilities to compute a colocalization posterior probability (CLPP) using the formula described in the eCAVIAR method[32,33,35].

Given the imperfect overlap between the variants in the GWAS summary statistics and the QTL data it can happen that the minimal corrected eQTL p-value or the GWAS p-values in a locus, after subsetting, were higher than the initial inclusion criteria of a locus. To exclude colocalizations in loci where only a very weak association, at either or both statistics, is observed, we excluded tests where the minimal GWAS p-value in the overlapping locus was higher than 5e-6 or the gene-level corrected eQTL p-value was greater than 0.05. We considered a locus to be colocalized if it

passed one of three CLPP thresholds; 1) minimal 5 variants with a CLPP of 0.5; 2) minimal 10 variants and a CLPP of 0.1; or 3) minimal 25 variants with a CLPP of 0.01.

Annotating i2QTL rare variants using UKBB GWAS summary statistics

To test for differences in outlier and non-outlier associated rare variants and risk for disease and traits, we overlapped i2QTL WGS variants (as described in Outlier Analysis above) with those measured or imputed in UKBB GWAS[77]. Specifically, we looked at variants with gnomAD MAF<1% and CADD>0. Multi-allelic variants were discarded. Variants were retained if they were observed in only one i2QTL individual, with the hypothesis being that these variants are driving observed outlier effects. From this list of unique variants (i.e. observed in only one i2QTL sample), outlier-associated variants were found separately for under-expression and over-expression outlier samples (therefore, there could be a maximum of two outlier samples per gene).

For each gene with >= 1 outlier sample, non-outlier-associated variants were chosen for each non-outlier sample if a variant had a CADD score within a range +/- 5 of outlier variants. A non-outlier sample was defined as a sample with expression Z-score > -1 and <1 for a given gene. If a non-outlier sample had a greater number of variants than the outlier sample, variants were randomly downsampled to match the number of outlier variants. If a non-outlier sample had a less or equal number of variants than the outlier sample, all variants were chosen for that sample. This process was performed separately for each gene and each outlier direction (i.e. under-expression, over-expression). The final list of variants was linked to UKBB GWAS effect size and p-values for each trait in UKBB GWAS Phase 1.

Data availability

All data used in the study is available via SRA, dbGaP or ENA; the full data availability is provided in **Table S1**. **Table S2** provides sample description on the samples used in the study.

Acknowledgements

The authors would like to thank the HipSci, iPSCORE, GENSIPS and PhLiPS cohorts for their input into the study and (early) access to the data. This work was supported by: the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant number 664726) (to MJB & DS); National Institutes of Health (T32 LM012409 to CS; T15 LM01127 to DJ; U01 HL107388-01, P30DK116074, SPO 130829 to IC-O; HL107442, DK105541, DK112155 to KAF); Stanford Graduate Fellowship (to MJG); National Natural Science Foundation of China (31970554 to XL); EBI–Sanger Postdoctoral Fellowship (to NC); National Science Foundation Graduate Research Fellowship (to NMF); California Institute for Regenerative Medicine (GC1R-06673 to KAF). Research in the Stegle lab is supported by the BMBF, the Volkswagen Foundation and the EU (ERC project DECODE). SBM is supported by NIH grants U01HG009431, R01HL142015, R01HG008150 and U01HG009080. This work in-part used supercomputing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by National Institutes of Health S10 Instrumentation Grant S10OD023452. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

The main analyses and data preparations were performed by MJB and CS; DJ and MD performed structural variant calling and analysis; NMF performed GTEx v7 data processing; LF and XL performed rare variant annotation and interpretation; MJG performed the colocalization analysis; MJB, DS, BM and DH developed the QTL calling software; NC, IC-O, YP assisted with data processing and analysis; MJB, CS, MJG, SBM, OS wrote the manuscript; IC-O, NC, NMF, KAF, LF, MJG, DJ, BM assisted in editing the manuscript; MJB, CS, ES, KAF, OS, SBM conceived and oversaw the study.

Conflicts of interest

SBM is on SAB of Prime Genomics Inc.

References

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
3. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
4. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
5. Consortium, G. & GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
6. Panopoulos, A. D. *et al.* iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**, 1086–1100 (2017).
7. Pashos, E. E. *et al.* Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell Stem Cell* **20**, 558–570.e10 (2017).
8. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).
9. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
10. Carcamo-Orive, I. *et al.* Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. *Cell Stem Cell* **20**, 518–532.e9 (2017).
11. Dubois, N. C. *et al.* SIRPA is a specific cell-surface marker for isolating cardiomyocytes derived from human pluripotent stem cells. *Nat. Biotechnol.* **29**, 1011–1018 (2011).
12. Sterneckert, J. L., Reinhardt, P. & Schöler, H. R. Investigating human disease using stem cell

models. *Nature Reviews Genetics* **15**, 625–639 (2014).

13. Choi, J. *et al.* A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. *Nat. Biotechnol.* **33**, 1173–1181 (2015).
14. Thomas, S. M. *et al.* Reprogramming LCLs to iPSCs Results in Recovery of Donor-Specific Gene Expression Signature. *PLoS Genet.* **11**, e1005216 (2015).
15. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants. *bioRxiv* 671040 (2019). doi:10.1101/671040
16. Li, X. *et al.* Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.* **95**, 245–256 (2014).
17. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
18. Consortium, T. U. & The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
19. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
20. Jakubosky, D. *et al.* Genomic properties of structural variants and short tandem repeats that impact gene expression and complex traits in humans. *bioRxiv* 714477 (2019). doi:10.1101/714477
21. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
22. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
23. Guan, F. H. X. *et al.* The antiproliferative ELF2 isoform, ELF2B, induces apoptosis in vitro and perturbs early lymphocytic development in vivo. *J. Hematol. Oncol.* **10**, (2017).
24. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191–W198 (2019).
25. Sampieri, L., Di Giusto, P. & Alvarez, C. CREB3 Transcription Factors: ER-Golgi Stress Transducers

as Hubs for Cellular Homeostasis. *Front. Cell Dev. Biol.* **7**, (2019).

26. Pongor, L. *et al.* A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. *Genome Med.* **7**, 104 (2015).
27. Rahane, C. S., Kutzner, A. & Heese, K. A cancer tissue-specific FAM72 expression profile defines a novel glioblastoma multiform (GBM) gene-mutation signature. *J. Neurooncol.* **141**, 57–70 (2019).
28. Wu, L. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* **50**, 968–978 (2018).
29. Makabe, T. *et al.* Genome-wide DNA methylation profile of early-onset endometrial cancer: its correlation with genetic aberrations and comparison with late-onset endometrial cancer. *Carcinogenesis* **40**, 611–623 (2019).
30. Bartzela, T. N., Carels, C. & Maltha, J. C. Update on 13 Syndromes Affecting Craniofacial and Dental Structures. *Front. Physiol.* **8**, (2017).
31. Iwata, J.-I. *et al.* Modulation of noncanonical TGF-β signaling prevents cleft palate in Tgfbr2 mutant mice. *J. Clin. Invest.* **122**, 873 (2012).
32. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
33. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
34. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz469
35. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
36. Website.
37. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
38. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).

39. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).

40. Molinari, F. *et al.* Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation. *Science* **298**, 1779–1781 (2002).

41. Canault, M. *et al.* Human CalDAG-GEFI gene (RASGRP2) mutation affects platelet function and causes severe bleeding. *J. Exp. Med.* **211**, 1349–1362 (2014).

42. Moreno, J. C. *et al.* Inactivating mutations in the gene for thyroid oxidase 2 (THOX2) and congenital hypothyroidism. *N. Engl. J. Med.* **347**, 95–102 (2002).

43. Streeter, I. *et al.* The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).

44. D'Antonio, M. *et al.* Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Rep.* **24**, 883–894 (2018).

45. DeBoever, C. *et al.* Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**, 533–546.e7 (2017).

46. Knowles, J. W., Hao, K., Xie, W., Weedon, M. N. & Zhang, Z. Genetic and Functional Analyses Identify NAT2 as a Human Insulin Sensitivity Gene. (2013).

47. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412 (2011).

48. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

49. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).

50. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

51. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

52. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).

53. *bcftools*. (Github).

54. Jakubosky, D. *et al.* Discovery and Quality Analysis of a Comprehensive Set of Structural Variants and Short Tandem Repeats. *bioRxiv* 713198 (2019). doi:10.1101/713198

55. Krueger, F. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. *URL http://www. bioinformatics. babraham. ac. uk/projects/trim_galore/. (Date of access: 28/04/2016)* (2012).

56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

57. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).

58. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

59. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

60. 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).

61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

62. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

63. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

64. Müller, S. *et al.* APADB: a database for alternative polyadenylation and microRNA regulation events. *Database* **2014**, (2014).

65. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

66. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard/. (Accessed: 24th September 2019)

67. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

68. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015).

69. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

70. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

71. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

72. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res.* **7**, 1860 (2018).

73. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).

74. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.* **17**, 118 (2016).

75. Hall, C. L. *et al.* Frequency of genetic variants associated with arrhythmogenic right ventricular cardiomyopathy in the genome aggregation database. *Eur. J. Hum. Genet.* **26**, 1312–1318 (2018).

76. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

77. Churchhouse, C. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank — Neale lab. *Neale lab* (2017). Available at: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank. (Accessed: 24th September 2019)