Supplementary Information for:

Identification of rare and common regulatory variants in pluripotent cells using population scale transcriptomics.

M.J. Bonder & C. Smail et al.

Table of contents

Supplementary results

Extended *trans*-eQTL results

Trans-eQTL hotspots

We identified four trans-eQTL hotspots, i.e. *trans*-eSNPs (eQTL SNPs) that are associated to more than five genes in *trans* (**Supplementary Table 10)**. The hotspot with the largest number of downstream genes affected was located near *ELF2* (**Extended data 5A**), a transcription factor linked to reduced proliferation[1]. The hotspot comprises effects of 16 individual *cis*-QTL variants at *ELF2* and genes in the vicinity; these *cis*-QTL effects could be decomposed into three major blocks (between block LD < 0.67) (**Extended data 5B**). Variants in the different blocks were linked to multiple *cis*-QTL types, primarily affecting *ELF2* RNA traits, but also the APA ratio of *NAA15* (**Extended data 5C-5F**). Despite moderate LD between individual variants, we observed a high degree of sharing of downstream *trans*-eGenes between these *cis*-eSNPs (**Extended data 5G**). From the mediation analyses we find that 17 of the top effects were significantly linked to both *ELF2* expression and other RNA traits, and in five instances *ELF2* expression level was no longer significantly associated after accounting for the other RNA trait (i.e. splicing-ratio, APA-ratio, exon-level or transcript-ratio). The 37 downstream genes, linked to the *ELF2* hotspot, were enriched for sequence motifs from the ELF transcription factor family, 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[2]), and 20 out of the 37 *trans*-eGenes were also known target of *ELF2 (*g:Profiler[2]). A second hotspot, in *cis*- linked to a *CREB3L2* with *trans*- effects on 8 eGenes, stood out because of an enrichment for the reactome pathway: “ER to Golgi Anterograde Transport” (p-adj<5.6×10^{-5}), which is consistent with previous associations between *CREB3L2* and both the ER and Golgi complex[3]. The remaining two hotspots are linked to multiple genes in *trans* as well as *cis* but none of them displayed an enrichment in the linked genes.

Next, we assessed the replication rate of hotspots focusing on *ELF2*. First, we assessed the replication of the gene-level *cis*-eQTL that underpins the majority of the *trans*-eQTL effects. Even though *ELF2* was expressed in all GTEx tissues, the iPSC *cis*-eQTL was “replicated” in only three GTEx tissues (Esophagus Mucosa; Esophagus Muscularis and Skin (Sun Exposed Lower Leg). When assessing the tissue specificity of the cis-eQTL using MASHR we find that the iPSC *cis*-eQTL on *ELF2* is specific to iPSC (interms of sign and ratio of effect size (**Supplementary methods**). To gain additional insights into this *trans*-eQTL hotspot, we replicated the *cis*- and *trans*-eQTL in a single cell differentiation study, providing regulatory effects of differentiating iPS cells towards definitive endoderm[4]. We assessed the replication of *cis*- and *trans*-eQTL effects in the three distinct cell types: iPSC (day-0), 1 day of differentiation (mesendoderm) and following 3 days of differentiation (definitive endoderm). The gene-level *cis*-eQTL was not replicated in any of the cell types (nominal P>0.05), most likely due to the low expression levels of *ELF2*. However, we replicated 12 out of 37 *ELF2* linked *trans*-eQTL effects in iPSC (day-0), 6 *trans*- effects in mesendoderm and 1 effect in definitive endoderm cells- (**Extended data 5H**).

Relationship between GWAS variants and *trans*-eQTL drivers

Complementary to the *cis*-eQTL colocation analysis, we assessed the overlap between variants that drive *trans*-eQTL and GWAS variants. One of the four *trans*-eQTL hotspots was found to be in high LD (>0.95) to a GWAS variant for telomere length (rs412658:C>T), the hotspot on chromosome 19 around *ZNF257*. The GWAS variant is linked to eight genes, (*ZNF257*, *ZNF208*, *ZNF98*, *ZNF209P*, *RPL34P34*, *ZNF676*, *ZNF729* and *VN1R85P*) in *cis*- and seven *trans*-eGenes (*DNAH3*, *MATN4*, *RBPJL*, *GALNT13*, *BEST2*, *SLC30A8* and *S100A4*). *ZNF257* itself is implicated in telomere length[5], however none of the downstream genes have previously been associated to telomere length. Another example is the GWAS variant (rs2277339:T>G) for age of menopause, which is in *trans* associated to *PRIM2* expression and to *PRIM1* in *cis*. Consistent with this association, missense variants in *PRIM1* have previously been implicated with age of menopause[6], and our data suggest an additional role of *PRIM2*. *PRIM1* and *PRIM2* function in a heterodimer at the protein level., but our genetic analysis also identified an expression link between the two. Finally, we identified four GWAS variants (rs11072494:C>T, rs2289187:C>T, rs10459648:C>T, rs6495117:T>C) for cleft palate, cleft lip and hemifacial microsomia, which were associated to six *trans*-eGenes (*GPR160*, *SEMA3A*, *MDFIC*, *GRIK2*, *FAM169A* and *GLB1L3*). This set of genes is enriched for the JNK cascade and the MAPK cascade biological processes (g:Profiler p-adj: 1.0×10^{-2}, 2.1×10^{-2} respectively), with known implications in congenital craniofacial abnormalities[7,8].

Supplementary methods

Dataset information

HipSci

The Human Induced Pluripotent Stem Cells Initiative (HipSci)[11,12] was established as a large, high-quality reference panel of human iPSC lines for the research community. In addition to large component of healthy individuals sampled from the population, HipSci includes individuals from selected rare genetic disease: Monogenic diabetes, Bardet-Biedl syndrome, hereditary cerebellar ataxia, hereditary spastic paraplegia, Kabuki syndrome, Usher syndrome and congenital eye defects, congenital hyperinsulinism, Alport syndrome, hypertrophic cardiomyopathy, primary immune deficiency, bleeding and platelet disorders, macular dystrophy, retinitis pigmentosa, Batten disease and childhood neurological diseases. iPSC lines were generated using a non-integrative methodology (Sendai virus), and either derived from fibroblasts or blood samples. All lines were genotyped using the Illumina beadchip HumanCoreExome-12 genotyping chip. RNA-sequencing was performed, either using a paired-end stranded protocol, or a single-end protocol, followed by Illumina[11] sequencing. DNA methylation information was generated using the Illumina 450K or Illumina EPIC array. For a subset of the lines whole genome sequencing was perfomed. 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, more information on the lines and donors can be found in **Supplemental Table 2** (a subset of (N=166) has been described previously in Kilpinen et al[11]). Further information on data generation can be obtained from Kilpinen et al[11] and Streeter et al[12].

iPSCORE

The iPSCORE project[13–15] was completed with the goal of assembling high quality iPSC lines derived from hundreds of ethnically diverse 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 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 previous publications: Panopoulos et al[15], DeBoever et al[14], D'Antonio et al[13]. Data was downloaded using the NCBI SRA download tool.

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[16]. 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 a 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[17]. Data was downloaded using the NCBI SRA download tool.

PhLiPS

The PhLiPS Study (Phenotyping Lipid traits in iPSC-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 in Pashos et al[18]. 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. Data was downloaded using the NCBI SRA download tool.

Banovich

The Banovich et al study[19] 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.[20]. RNA-sequencing was performed using 50-bp, 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[19]. Genotype information was taken from the illumina arrays generated in the 1000G project[21]. Data was downloaded using the NCBI SRA download tool.

Stanford Center for Undiagnosed Diseases

Fibroblast culture: primary skin fibroblasts were obtained from five patients currently enrolled at the Stanford Center for Undiagnosed Diseases. Fibroblasts were grown from a skin punch biopsy and maintained in DMEM medium (Sigma-Aldrich) supplemented with 10% FBS (ThermoFisher). When cells reached confluency, cells were removed from culture medium washed with PBS (Mg++ and Ca++ free), and 0.05% trypsin/EDTA was added in sufficient quantity to cover the dish. Cells were trypsinized in a 37C, CO2 incubator for 3 minutes. One ml of culture medium was used to stop trypsin, pipetted up and down to break the cells to single cells. Cells were transferred into 15 ml or 50 ml tubes depending on the amount and cells. Culture dish was washed once with 5 ml culture medium and transferred to the same tube. Tubes were spun at 270g (1000 rpm) for 5 minutes and the medium was discarded by pipetting (making sure no cells were discarded). Tube was finger-flicked to loosen up the cell pellet and cells were resuspended completely in 1-2 ml culture medium. Cells were diluted with additional culture medium to the density according to the size of the dish to be plated on. Dishes were plated into a 37C, CO2 incubator until confluency for further passaging or reprogramming.

Fibroblast-iPSC reprogramming: iPSCs were reprogrammed with Sendai Virus from patient biopsy skin fibroblast. It took 26-30 days to form the iPSC colonies. Once the iPSC colonies were formed and ready to be picked, pipette tips were used to pick the colonies under a microscope. The iPSC colonies were then cultured in serum-free/feeder free medium-hStemSFM (Stemmera, ST02001) on matrigel coated plates for around 20 passages.

Genotyping information

Array-based genotyping

The generation of the genotype data is described in the original publications, but were homogeneously reprocessed in this study. For all eQTL analyses, we considered homogeneously imputed genotypes using a per-sample imputation and phasing pipeline based on IMPUTE2 v2.3.1[22] for imputation and SHAPEIT2 v2.r790[23] for phasing. Imputation and phasing were performed using a combined genotyping reference, encompassing the haplotypes from the UK10K cohorts and 1000G Phase 1 data[23,24]. The imputation was run in chunks with an average size of 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[25] (v1.5) and BCFtools[26] (v.1.31). After imputation, data from each study were combined and cohort-, and therefore chip-, wise SNP-QC was performed using Genotype Harmonizer. SNPs with either a call rate below 90% or imputation quality (Mach R2) below 0.4 were discarded. After study-wide QC, we merged the datasets and performed a combined SNP QC filtering, considering a minimum call rate of 1.0 and an imputation score of 0.4. The final dataset for the QTL analyses spanned 2,533 samples and 7,188,631 variants.

Super population assignment

All imputed samples were assigned to 1000G super-populations by projecting genotypes onto principal components from the genotype matrix of individuals from the 1000 Genomes Project (phase 3[21]). Briefly, we selected variants with a minor allele frequency (MAF) of at least 5% within 1000G (phase 3)[21] and that passed QC in the joint imputed i2QTL genotype dataset. We then performed principal component analysis on the 1000G samples and considered the first 40 components to project i2QTL samples onto the calculated principal components. Next, for each i2QTL samples, we calculated the distance between the sample and the centroid value for each 1000G super-population. 40 components were chosen as leave-one out cross validation on the 1000G samples showed this to be the optimal number.

Whole genome sequencing

For a subset of the samples obtained from HipSci and iPSCORE, whole genome sequencing (WGS) data were available. Within the HipSci consortium whole genome sequencing data was available from fibroblasts from 201 donors, iPSC lines of 152 donors (out of the 201) were also sequenced using WGS, a total of 242 iPSC lines were WGS sequenced within the HipSci project. From iPSCORE, 273 individuals had whole genome sequencing data available from blood (254 donors) or fibroblasts (19 donors). We reprocessed the WGS data from both cohorts, calling both SNVs and structural variants, as described in Jakubosky et al[27]. Briefly, 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 716 high-quality samples, 425 of these samples had iPSC RNA-seq.

RNA-sequencing

RNA-sequencing feature quantification

Raw RNA-seq data from all studies were obtained from EGA, dbGaP and SRA (**Supplementary Table 1**). Depending on the data formats provided by the respective source study, the starting files were provided in either CRAM, BAM or FASTQ format. To ensure uniform processing, CRAM and BAM files were converted to FASTQ files. Reads were trimmed to discard adapters and low-quality bases, as well as low quality reads. Trimming was performed using Trim Galore! (https://github.com/FelixKrueger/TrimGalore, v0.6), a wrapper around Cutadapt[28] and FastQc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed and QC reads were aligned using STAR (version: 020201)[29], employing the two-pass alignment mode and the parameters as proposed by ENCODE (c.f. STAR manual). Alignments were performed using the GRCh37 reference genome and Ensembl 75 genome annotations[30]. Based on the aligned reads, we quantified gene-level RNA expression for all samples. For the subset of paired-end stranded data, we additionally 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)[31], which was applied to the primary alignments using the "-B" and "-C" options in stranded mode if applicable. When multiple RNA-seq runs were available for a given iPSC line, count matrices were summed to obtain a consensus gene count table per line. Read counts were normalized by gene length and adjusted for library size using edgeR[32] (V3), yielding adjusted transcript per million counts (TPM). The same workflow and normalization procedure was used to quantify exon expression levels. Given that Ensembl exons can overlap we chose to first merge (strand-specific) overlapping exons into meta-exons and quantified these meta-exons instead of the individual exons.

To facilitate comparisons with GTEx v7[33], we performed a second gene-level quantification using RNA-SeQC[34] (v.1.1.8) using the gencode v19 annotation matching the GTEx quantification pipeline. RNA-SeQC was run on the read and sample QC, and STAR settings as described above; therefore, it should be noted that the processing is not completely matching the GTEx pipeline. This RNA-SeQC gene-level quantification was used for analyses that compare i2QTL data to GTEx. Specifically, these data were used for the alignment of gene expression profile across studies (**Figure 1A),** as well as the GTEx outlier comparison (**Figure 2B**).

Transcript ratios were quantified by determining transcript isoform levels using Salmon[35] (version: 0.8.2). The Salmon transcript database was built based on transcript information from Ensembl (v75). Salmon directly operates on quality controlled FASTQ data; the transcript quantification options "seqBias", "gcBias" and "VBOpt" were used. Transcript count quantifications as returned by Salmon were normalized analogous to the approach taken for gene-level RNA abundance above. Subsequently, we transformed 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 workflow described in Zhernakova et al.[36]. 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[37]. For each 3'UTR, the set of alternative polyadenylation sites within each 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.

Finally, we quantified splicing levels using leafcutter[38] (v0.2.8). Quantifications of splice-events were calculated as described in the leafcutter manual. Junction reads were extracted from the source BAM files, after which leafcutter clusters junction reads into introns. Subsequently, the intron counts were 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 identified based on quality metrics from Picard[39] (V2.9.0) and VerifyBamID[40] (V1.1.3), as well as gene expression statistics. Briefly, we consider the following minimum QC values: > 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. Additionally, we discarded samples that had low expression correlations (<0.6) to the average iPSC expression values across all samples 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. We further used 98 samples from Choi et al[41] and non-iPSC samples from HipSci[11,12] which are included as reference (**Supplementary Table 2**).

After RNA QC, the sample identity was validated using VerifyBamID, whereby for each RNA-sample, the best matching genotype was identified based on the read information and compared to the expected match. This approach identified 36 sample swaps (between 0-26 mismatches per study), and 33 unmatched RNA-seq samples. Where possible, sample swaps were corrected (N=34) and others the corresponding samples were discarded (N=2).

eQTL annotation

The i2QTL genetic maps were annotated by overlapping the eQTL signals with information taken from published eQTL maps, including the source iPSC studies[11,14,17–19], GTEx[33] and BIOS[36].

To assess the replication of the eQTL effects in the original iPSC studies[11,14,17–19], we assessed the replication of published eQTL variants at P<0.05 in our study, and assessed effect the consistency of effect direction **(Extended data 1)**. When assessing the replication within the iPSCORE study we exclusively considered SNP effects and excluded the effects of structural variants.

To assess the similarity of eQTL identified for different RNA traits, we assessed the pairwise replication between the corresponding eQTL maps. We quantified the replication at the level of individual genes, i.e. by mapping the eQTL types to the respective gene and considering the most significant association (at gene-level FDR) if multiple traits were tested for a given gene. If an eQTL effect discovered in the first eQTL type was replicated at FDR 10% in a second eQTL type, the eQTL was considered to be replicated (**Figure 1D**).

To identify eQTL specific to iPSCs, we first considered the overlap of eGenes using eQTL maps from GTEx[33] and BIOS[36] (**Figure 1E;** at FDR<5% as provided by the respective studies (i.e. gene selection and FDR methods are taken as reported in the original studies). As a more refined measure of eQTL sharing, we used MASHR[42] (V0.2.21) to probe for shared eQTL effects of individual variants. For this analysis, we considered genes expressed in i2QTL and each of the 48 GTEx tissues (n=11,682), as well as variants tested in both studies. We ran MASHR on the lead eQTL variant per gene, as determined by the largest absolute beta per gene across the assessed tissues, and included four random eQTL per genes for the calculation of the local false sign ratio (LSFR). Using the posterior beta's and the LSFR values we estimated pairwise sharing levels between tissues. eQTL effects were considered as shared if they were deemed significant in both tissues (LSFR<0.05), and if the respective effect sizes had the same sign and were within a factor of two of each other. This analysis was performed twice: i) considering i2QTL iPSC eQTL and eQTL maps from 48 GTEx tissues, ii) additional considering single cell eQTL maps from Cuomo et al.[4] (**Extended data 3**).

We assessed the link between MSigDB gene sets and tissue specificity of eQTL signals from the MASHR analyses using the GSEA tool[43,44] (V4.1), using GSEA MSigDB (v7.1) for gene annotation. Specifically, we used the pre-ranked mode of the tool on the sign only comparison of the MASHR eQTL specificity, linking gene sets with the level top eQTL sharing.

Trans-eQTL replication

We assessed replication of *trans*- eQTL associations using both expression and DNA-methylation information. First, we considered bulk RNA-seq profiles from other holdout samples (n=186 donors, n=253 lines) **Supplementary Table 2**, to assess the replication of the identified *trans*-eQTL in independent data. Second, we considered DNA methylation data available for a subset of lines (n=572 donors, n=841 lines) and performed cross-omic replication of the identified *trans*-eQTL. Third, we considered replication based on the Cuomo et al.[4] single-cell RNA-seq (scRNA-seq) data across multiple days of iPS differentiation towards definitive endoderm to assess the tissue specificity of *trans*-eQTLs.

The replication of the *trans*-eQTL using holdout bulk RNA-seq data as well as scRNA-seq data from Cuomo et al.[28] was performed using the same workflow as used for *trans*-eQTL discovery analysis, with the exception of the adjustment for PEER factors. To rule out the any risk of synthetic associations by adjustment for PEER factors in the bulk replication, only known factors were considered to adjust for confounding: inferred ancestry (from genotyped data), sequencing type (paired-end (yes/no), stranded (yes/no)) and hot-encoded vectors describing the dataset of origin. For the scRNA-seq replication, we reprocessed the reads analogously to the bulk RNA-seq data and we quantified expression with featureCounts. QC steps and the aggregation strategy of cells from the same cell state and line as well as the batch correction were implemented according to Cuomo et al[4]. We deemed *trans*-eQTL as replicated if 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[19] and HipSci[11,12] 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 previously[45], based on the DASEN[46] normalization. After joint normalization, we corrected 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 linked 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 were within 250Kb of the gene TSS and TES, i.e. our *cis*-eQTL window. 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 the 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 assess if we replicated more effects than expected by chance, we repeated the *trans*-eQTL and *trans*-meQTL replication on random *trans*-eQTL pairs. To 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 as well as the single cell replication.

Supplementary Tables

Supplementary Table 1. Overview of study name, data types, and accession IDs for source studies that are contained in the i2QTL resource

Supplementary Table 2. Metadata for iPSCs contained in the i2QTL resource

Supplementary Table 3. Summary statistics of lead gene-level *cis*-eQTL. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 4. Summary statistics of lead trans-cript *cis*-eQTL. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 5. Summary statistics of lead exon *cis*-eQTL. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 6. Summary statistics of lead splicing *cis*-eQTL. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 7. Summary statistics of lead APA *cis*-eQTL. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 8. GSEA enrichment results for the tissue specificity of eQTL in iPSC as derived from the MASHR analysis on GTEx and i2QTL.

Supplementary Table 9. Lead gene-level *trans*-eQTL identified in the i2QTL study. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 10. Full gene-level trans-eQTL identified in the i2QTL study. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q (Methods).

Supplementary Table 11. Outlier rare variant analysis across different outlier Z-score thresholds, gnomad MAF and CADD variant thresholds

Supplementary Table 12. Rare variant enrichment for comparison analysis between i2QTL and GTEx v7

Supplementary Table 13. Summary information on colocalization events identified using the i2QTL *cis*-eQTL summary statistics.

Supplementary Table 14. Summary information on colocalization events identified using the i2QTL *trans*-eQTL summary statistics.

Supplementary Table 15. Combined summary information on colocalization events identified using *cis*-eQTL in iPSC and 48 GTEx tissues.

Supplementary Table 16. GWAS summary statistics for rare variants linked to outlier expression in colocalized genes

Supplementary Table 17. Overview of genes considered across different analyses in the i2QTL study

Supplementary Table 18. Information black listed gene pairs due to sequence homology for *trans*-eQTL mapping.

Banner authors

Consortia ordered by appearance on the author list, the authors are ordered by last name and main affiliations.

HipSci consortium

Chukwuma A. Agu[12], Alex Alderton[12], Shrada Amatya[12],Philip Beales[14], Dalila Bensaddek[21], Ewan Birney[1], Francesco Paolo Casale[1], Laura Clarke[1], Petr Danecek[12], Davide Denovi[21], Rachel Denton[12], Richard Durbin[12], Daniel J. Gaffney[12], Angela Goncalves[12], Reena Halai[12], Sarah Harper[12], Peter W. Harrison[1], Helena Kilpinen[1], Christopher M Kirton[12], Andrew Knights[12], Anja Kolb-Kokocinski[12], Angus I. Lamond[22], Andreas Leha[12], Davis J. McCarthy[1], Shane A. McCarthy[12], Ruta Meleckyte[19], Yasin Memari[12], Natalie Moens[19], Willem H. Ouwehand[23], Minal Patel[12], Oliver Stegle[1], Ian Streeter[1], Ludovic Vallier[23], Fiona M. Watt[19]

21. King's College London, Tower Wing, Guy's Hospital, Great Maze Pond, London, UK
22. University of Dundee, Dundee, UK
23. University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK

iPSCORE consortium

Angelo D. Arias[24], Paola Benaglio[24], W. Travis Berggren[25], Neil C. Chi[24], Matteo D'Antonio[24], Agnieszka D'Antonio-Chronowska[24], Carl T. Dargitz[25], Christopher DeBoever[24], Margaret K.R. Donovan[24], Kenneth E. Diffenderfer[25], Sylvia M. Evans[24], KathyJean Farnam[24], Kelly A. Frazer[24], Rachel Feiring[25], Melvin Garcia[24], Lawrence S.B. Goldstein[24], William W. Greenwald[24], Olivier Harismendy[24], Juan Carlos Izpisua Belmonte[25], David A. Jakubosky[24], Kristen Jepsen[24], He Li[24], Hiroko Matsui[24], Thomas J. McGarry[24], Veronica Modesto[25], Bradley C. Nelson[24], Daniel T. O'Connor[24], Athanasia D. Panopoulos[25], Fangwen Rao[24], Erin N. Smith[24], Gene W. Yeo[24]

24. University of California, San Diego, La Jolla, CA 92093, USA
25. Stem Cell Core, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

PhLiPS consortium

Christopher D. Brown[26], YoSon Park[26]

26. Perelman School of Medicine University of Pennsylvania

Undiagnosed Diseases Network

Maria T. Acosta[27], Margaret Adam[27], David R. Adams[27], Pankaj B. Agrawal[27], Mercedes E. Alejandro[27], Justin Alvey[27], Laura Amendola[27], Ashley Andrews[27], Euan A. Ashley[27], Mahshid S. Azamian[27], Carlos A. Bacino[27], Guney Bademci[27], Eva Baker[27], Ashok Balasubramanyam[27], Dustin Baldridge[27], Jim Bale[27], Michael Bamshad[27], Deborah Barbouth[27], Pinar Bayrak-Toydemir[27], Anita Beck[27], Alan H. Beggs[27], Edward Behrens[27], Gill Bejerano[27], Jimmy Bennet[27], Beverly Berg-Rood[27], Jonathan A. Bernstein[27], Gerard T. Berry[27], Anna Bican[27], Stephanie Bivona[27], Elizabeth Blue[27], John Bohnsack[27], Carsten Bonnenmann[27], Devon Bonner[27], Lorenzo Botto[27], Brenna Boyd[27], Lauren C. Briere[27], Elly Brokamp[27], Gabrielle Brown[27], Elizabeth A. Burke[27], Lindsay C. Burrage[27], Manish J. Butte[27], Peter Byers[27], William E. Byrd[27], John Carey[27], Olveen Carrasquillo[27], Ta Chen Peter Chang[27], Sirisak Chanprasert[27], Hsiao-Tuan Chao[27], Gary D. Clark[27], Terra R. Coakley[27], Laurel A. Cobban[27], Joy D. Cogan[27], Matthew Coggins[27], F. Sessions Cole[27], Heather A. Colley[27], Cynthia M. Cooper[27], Heidi Cope[27], William J. Craigen[27], Andrew B. Crouse[27], Michael Cunningham[27], Precilla D'Souza[27], Hongzheng Dai[27], Surendra Dasari[27], Joie Davis[27], Jyoti G. Dayal[27], Matthew Deardorff[27], Esteban C. Dell'Angelica[27], Shweta U. Dhar[27], Katrina Dipple[27], Daniel Doherty[27], Naghmeh Dorrani[27], Argenia L. Doss[27], Emilie D. Douine[27], David D. Draper[27], Laura Duncan[27], Dawn Earl[27], David J. Eckstein[27], Lisa T. Emrick[27], Christine M. Eng[27], Cecilia Esteves[27], Marni Falk[27], Liliana Fernandez[27], Carlos Ferreira[27], Elizabeth L. Fieg[27], Laurie C. Findley[27], Paul G. Fisher[27], Brent L. Fogel[27], Irman Forghani[27], Laure Fresard[27], William A. Gahl[27], Ian Glass[27], Bernadette Gochuico[27], Rena A. Godfrey[27], Katie Golden-Grant[27], Alica M. Goldman[27], Madison P. Goldrich[27], David B. Goldstein[27], Alana Grajewski[27], Catherine A. Groden[27], Irma Gutierrez[27], Sihoun Hahn[27], Rizwan Hamid[27], Neil A. Hanchard[27], Kelly Hassey[27], Nichole Hayes[27], Frances High[27], Anne Hing[27], Fuki M. Hisama[27], Ingrid A. Holm[27], Jason Hom[27], Martha Horike-Pyne[27], Alden Huang[27], Yong Huang[27], Laryssa Huryn[27], Rosario Isasi[27], Fariha Jamal[27], Gail P. Jarvik[27], Jeffrey Jarvik[27], Suman Jayadev[27], Lefkothea Karaviti[27], Jennifer Kennedy[27], Dana Kiley[27], Isaac S. Kohane[27], Jennefer N. Kohler[27], Deborah Krakow[27], Donna M. Krasnewich[27], Elijah Kravets[27], Susan Korrick[27], Mary Koziura[27], Joel B. Krier[27], Seema R. Lalani[27], Byron Lam[27], Christina Lam[27], Grace L. LaMoure[27], Brendan C. Lanpher[27], Ian R. Lanza[27], Lea Latham[27], Kimberly LeBlanc[27], Brendan H. Lee[27], Hane Lee[27], Roy Levitt[27], Richard A. Lewis[27], Sharyn A. Lincoln[27], Pengfei Liu[27], Xue Zhong Liu[27], Nicola Longo[27], Sandra K. Loo[27], Joseph Loscalzo[27], Richard L. Maas[27], John MacDowall[27], Ellen F. Macnamara[27], Calum A. MacRae[27], Valerie V. Maduro[27], Marta M. Majcherska[27], Bryan C. Mak[27], May Christine V. Malicdan[27], Laura A. Mamounas[27], Teri A. Manolio[27], Rong Mao[27], Kenneth Maravilla[27], Thomas C. Markello[27], Ronit Marom[27], Gabor Marth[27], Beth A. Martin[27], Martin G. Martin[27], Julian A. Martínez-Agosto[27], Shruti Marwaha[27], Jacob McCauley[27], Allyn McConkie-Rosell[27], Colleen E. McCormack[27], Alexa T. McCray[27], Elisabeth McGee[27], Heather Mefford[27], J. Lawrence Merritt[27], Matthew Might[27], Ghayda Mirzaa[27], Eva Morava[27], Paolo M. Moretti[27], Deborah Mosbrook-Davis[27], John J. Mulvihill[27], David R. Murdock[27], Anna Nagy[27], Mariko Nakano-Okuno[27], Avi Nath[27], Stan F. Nelson[27], John H. Newman[27], Sarah K. Nicholas[27], Deborah Nickerson[27], Shirley Nieves-Rodriguez[27], Donna Novacic[27], Devin Oglesbee[27], James P. Orengo[27], Laura Pace[27], Stephen Pak[27], J. Carl Pallais[27], Christina GS. Palmer[27], Jeanette C. Papp[27], Neil H. Parker[27], John A. Phillips III[27], Jennifer E. Posey[27], Lorraine Potocki[27], Bradley Power[27], Barbara N. Pusey[27], Aaron Quinlan[27], Wendy Raskind[27], Archana N. Raja[27], Deepak A. Rao[27], Genecee Renteria[27], Chloe M. Reuter[27], Lynette

Rives[27], Amy K. Robertson[27], Lance H. Rodan[27], Jill A. Rosenfeld[27], Natalie Rosenwasser[27], Francis Rossignol[27], Maura Ruzhnikov[27], Ralph Sacco[27], Jacinda B. Sampson[27], Susan L. Samson[27], Mario Saporta[27], C. Ron Scott[27], Judy Schaechter[27], Timothy Schedl[27], Kelly Schoch[27], Daryl A. Scott[27], Vandana Shashi[27], Jimann Shin[27], Rebecca Signer[27], Edwin K. Silverman[27], Janet S. Sinsheimer[27], Kathy Sisco[27], Edward C. Smith[27], Kevin S. Smith[27], Emily Solem[27], Lilianna Solnica-Krezel[27], Ben Solomon[27], Rebecca C. Spillmann[27], Joan M. Stoler[27], Jennifer A. Sullivan[27], Kathleen Sullivan[27], Angela Sun[27], Shirley Sutton[27], David A. Sweetser[27], Virginia Sybert[27], Holly K. Tabor[27], Amelia L. M. Tan[27], Queenie K.-G. Tan[27], Mustafa Tekin[27], Fred Telischi[27], Willa Thorson[27], Audrey Thurm[27], Cynthia J. Tifft[27], Camilo Toro[27], Alyssa A. Tran[27], Brianna M. Tucker[27], Tiina K. Urv[27], Adeline Vanderver[27], Matt Velinder[27], Dave Viskochil[27], Tiphanie P. Vogel[27], Colleen E. Wahl[27], Stephanie Wallace[27], Nicole M. Walley[27], Chris A. Walsh[27], Melissa Walker[27], Jennifer Wambach[27], Jijun Wan[27], Lee-kai Wang[27], Michael F. Wangler[27], Patricia A. Ward[27], Daniel Wegner[27], Mark Wener[27], Tara Wenger[27], Katherine Wesseling Perry[27], Monte Westerfield[27], Matthew T. Wheeler[27], Jordan Whitlock[27], Lynne A. Wolfe[27], Jeremy D. Woods[27], Shinya Yamamoto[27], John Yang[27], Muhammad Yousef[27], Diane B. Zastrow[27], Wadih Zein[27], Chunli Zhao[27], Stephan Zuchner[27]

27. NIH Undiagnosed Disease Network, National institutes of health, bethesda, MD, USA

References

1. 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).
2. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* vol. 47 W191–W198 (2019).
3. 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).
4. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
5. Coutts, F. *et al.* The polygenic nature of telomere length and the anti-ageing properties of lithium. *Neuropsychopharmacology* **44**, 757–765 (2019).
6. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat. Genet.* **44**, 260–268 (2012).
7. Bartzela, T. N., Carels, C. & Maltha, J. C. Update on 13 Syndromes Affecting Craniofacial and Dental Structures. *Front. Physiol.* **8**, (2017).
8. Iwata, J.-I. *et al.* Modulation of noncanonical TGF-β signaling prevents cleft palate in Tgfbr2 mutant mice. *J. Clin. Invest.* **122**, 873 (2012).
9. Yao, Y. & Dai, W. Shugoshins function as a guardian for chromosomal stability in nuclear division. *Cell Cycle* **11**, 2631–2642 (2012).
10. André, F. *et al.* Exonic expression profiling of breast cancer and benign lesions: a retrospective analysis. *Lancet Oncol.* **10**, 381–390 (2009).
11. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
12. Streeter, I. *et al.* The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).
13. 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).

14. 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).
15. 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).
16. 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).
17. 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).
18. 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).
19. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).
20. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412 (2011).
21. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
22. 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).
23. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
24. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

25. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
26. Li, H. *et al.* Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **1000**, 2078–2079.
27. Jakubosky, D. *et al.* Discovery and Quality Analysis of a Comprehensive Set of Structural Variants and Short Tandem Repeats. *Nat. Commun*. 11, 2928 (2020).
28. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
30. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
31. 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).
32. 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).
33. Consortium, G. & GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* vol. 550 204–213 (2017).
34. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
35. 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).
36. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
37. Müller, S. *et al.* APADB: a database for alternative polyadenylation and microRNA regulation events. *Database* **2014**, (2014).
38. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat.*

Genet. **50**, 151–158 (2018).

39. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/.
40. 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).
41. 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).
42. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
43. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
44. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
45. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
46. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).