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

Marc Jan Bonder, Craig Smail, Michael J. Gloudemans, Laure Frésard, David Jakubosky, Matteo D’Antonio, Xin Li, Nicole M. Ferraro, Ivan Carcamo-Orive, Bogdan Mirauta, Daniel D. Seaton, Na Ca, Dara Vakili, Danilo Horta, Chunli Zhao, Erin N. Smith, Kelly A. Frazer, Stephen B. Montgomery, Danilo Horta, Xin Li, Nicole M. Ferraro, Ivan Carcamo-Orive, Joshua W. Knowles, Erin N. Smith, and Oliver Stegle.

Induced pluripotent stem cells (iPSCs) are an established cellular system to study the impact of genetic variants in derived cell types and developmental contexts. However, in their pluripotent state, the disease impact of genetic variants is less well known. Here, we integrate data from 1,367 human iPSC lines to comprehensively map common and rare regulatory variants in human pluripotent cells. Using this population-scale resource, we report hundreds of new colocalization events for human traits specific to iPSCs, and find increased power to identify rare regulatory variants compared with somatic tissues. Finally, we demonstrate how iPSCs enable the identification of causal genes for rare diseases.

The regulatory effects of common disease loci identified from genome-wide association studies (GWAS), and rare variants for rare genetic disorders, have been increasingly linked to expression changes using large population-scale gene expression resources. Existing efforts have focused on blood somatic tissues collected postmortem, as well as transformed lymphoblastoid cell lines. However, detecting the regulatory effects of variants can be limited by tissue or cell accessibility. Complementary to somatic cells, human iPSCs combined with differentiation protocols provide powerful model systems for a growing range of mature cell states and types, which have been applied to study both molecular mechanisms of common and rare diseases. Pluripotent cells themselves can provide unique insights into regulation of gene expression in cell states that mimic early development, with relevance to diseases that manifest in utero or in transient states throughout development. However, the regulatory landscape of genetic variation in human pluripotent cells and its relationship to common and rare genetic diseases remain poorly understood, mainly due to the lack of appropriately powered genomic resources in human iPSCs.

To address this, we integrated data from five major iPSC genetic studies within the ‘integrated iPSC QTL’ (i2QTL) consortium, establishing a large-scale resource of iPSCs with matched genotype and RNA-sequencing (RNA-seq) data from 1,367 lines. We characterize regulatory effects of common variants using expression quantitative trait loci (eQTL) mapping of a comprehensive set of RNA phenotypes, including gene-level abundance, exon level, transcript ratio, splicing ratio and alternative polyadenylation (APA) ratio. This identifies hundreds of new expression quantitative trait loci (eQTLs), which are implicated in colocalization events across a broad range of human traits and diseases.

We further leverage the unique opportunity posed by a large whole-genome sequencing (WGS) resource combined with...
RNA-seq to probe for rare variants that are associated with gene expression outliers in human iPSCs. Previous work has demonstrated that aberrant gene expression can enable detection of rare variants when analyzed against a large reference cohort; however, despite their use in rare-disease research, no such reference exists for human pluripotent cells. We show that iPSCs provide increased power for identifying rare, large-effect expression variants compared to previous findings using somatic tissues, and improve the prioritization of rare variants implicated in a range of common traits and diseases. We also demonstrate the use of the i2QTL resource in modeling gene expression outlier effects linked to pathogenic rare variants across a range of rare diseases, including monogenic diabetes, Bardet–Biedl syndrome and hereditary cerebellar ataxia. Finally, we present an individual with global developmental delay, demonstrating the rapid improvement in resolution of candidate disease genes using a joint gene expression outlier analysis of blood and iPSC tissues.

Results

To generate the i2QTL resource, we collected previously published (~60%) and newly generated (~40%) data from human iPSC lines across five major iPSC resources (Supplementary Tables 1 and 2), spanning genotype and RNA-seq data from 1,367 iPSC lines, derived from 948 primarily healthy donors (65 rare-disease samples). We included additional data from fibroblast cell lines and embryonic stem cells (ESCs) from the HipSci cohort and Choi et al. (Supplementary Tables 1 and 2). We uniformly reprocessed genotype array, WGS and RNA-seq data across all samples (Methods). Joint multidimensional scaling of our data and samples from the Genotype-Tissue Expression (GTEx) project (v7) revealed a high homogeneity of iPSCs, within and between studies, compared to between-sample and tissue variations observed in GTEx (Fig. 1a). Furthermore, iPSCs clustered together with ESCs, supporting the quality of the i2QTL resource (Fig. 1a).

A high-resolution map of cis-eQTLs in human pluripotent cells.

We mapped cis-eQTLs, considering proximal common variants (gene body + 250 kb on both sides; minor allele frequency (MAF) > 1%) and paired-end stranded RNA-seq data available for 936 samples (n = 682 donors) of European ancestry (Methods). For 18,430 genes of 27,046 Ensembl genes expressed in iPSCs (Methods), we identified at least one cis-eQTL (Supplementary Table 3) in the following denoted eGenes. This corresponds to a 2.5-fold increase compared to the largest previous gene-level cis-eQTL map in human pluripotent cells (Fig. 1b) while replicating previous studies (Supplementary Methods and Extended Data Fig. 1). Iterative eQTL mapping using stepwise regression (Methods) identified two or more independent effects for 39.0% of eGenes, with a maximum of 12 independent cis-eQTLs for PTGR1 (Extended Data Fig. 2). In addition to gene-level expression, we considered further RNA-seq-derived traits for eQTL mapping: transcript ratio, exon level, splicing ratio and APA ratio (Methods and Supplementary Table 4–7). In aggregate, we report genetic effects for 21,548 genes spanning genotype and RNA-seq data from 1,367 iPSC lines, derived from 948 primarily healthy donors (65 rare-disease samples), spanning 2.5 Mb from gene body (median tissuerate; Methods and Extended Data Fig. 3a). We also ranked eQTLs from GTEx by the number of GTEx tissues in which these effects were replicated, observing that eGenes with low eQTL replication in GTEx were again enriched for cancer (n = 3) and developmental gene sets (n = 1; q < 10%; pre-ranked gene-set enrichment analysis (GSEA) v4.1; Methods and Supplementary Table 8). Finally, we extended the mashR analysis, including single-cell data from an iPSC differentiation study using HipSci lines, which demonstrated that the observed pattern of iPSC-specific eQTLs diminishes rapidly as cells exit a pluripotent state (Extended Data Fig. 3b).

Identification of trans-eQTL in pluripotent cells.

We tested for trans effects for 11,682 genes (replicated effect and consistent effect direction; Methods). Globally, this identified iPSCs as markedly distinct from GTEx tissues (iPSC versus GTEx, on average, 68.1% versus 88.7% within GTEx tissues replication rate; Methods and Extended Data Fig. 3a). We also ranked eQTLs discovered in iPSCs by the number of GTEx tissues in which these effects were replicated, observing that eGenes with low eQTL replication in GTEx were again enriched for cancer (n = 3) and developmental gene sets (n = 1; q < 10%; pre-ranked gene-set enrichment analysis (GSEA) v4.1; Methods and Supplementary Table 8). Finally, we extended the mashR analysis, including single-cell data from an iPSC differentiation study using HipSci lines, which demonstrated that the observed pattern of iPSC-specific eQTLs diminishes rapidly as cells exit a pluripotent state (Extended Data Fig. 3b).

We used held-out samples (237 lines from 186 donors) to assess the replication of trans-eQTLs, observing evidence for 17.1% of the individual associations in the hold-out fraction (nominal P < 0.05 and same effect direction). We applied the same replication strategy using DNA methylation data available for a subset of lines (n = 841), replicating 26.9% of the effects (considering methylation probes proximal to target genes; adjusted (adj) P < 0.05; Supplementary Methods). These replication rates exceeded the chance expectation (expression: 7%; methylation: 20%; Methods and, collectively, provided evidence for nominal replication of 37.8% of the trans associations (Supplementary Table 9).

Next, to explore the tissue specificity of trans-eQTLs, we assessed evidence for tissue-specific regulation of cis-eQTLs that drive trans-eQTLs (using the mashR analysis; Methods). We observed that cis-eQTLs with downstream trans effects were associated with a lower degree of tissue sharing than other cis-eQTLs (median tissue sharing: 7 versus 16; R = 0.04, Wilcoxon test; Supplementary Table 3). These putative iPSC-specific eGenes were enriched for cancer (COSMIC genes; qu = 2.4 × 10^{-5}, Fisher’s exact test) and embryonic development (q = 0.03).
Methods). Additionally, we assessed the replication of trans-eQTLs across a single-celled RNA-seq differentiation time course from 125 HipSci lines. In undifferentiated iPSCs, 11.9% of the trans effects were replicated ($P < 0.05$ and same direction; 86.5% of trans-eGenes expressed), while we observed a marked decrease in the replication rate following 1 d of differentiation toward endoderm, and even lower replication rates following 3 d of differentiation toward definitive endoderm (4.1%; Supplementary Table 9). Consistent with these global statistics, we also observed reduced trans effects for individual targets of the trans-eQTL hot spot at ELF2 (Supplementary Results, Supplementary Tables 9 and 10 and Extended Data Fig. 5).

iPSC analysis improves identification of rare variants associated with aberrant gene expression. Given the sample size of our cohort and the availability of high-quality SNP and structural variant (SV) calls based on WGS data ($n = 425$ lines), we sought to identify rare variants with large effects on iPSC gene expression. Adapting strategies previously used in cohorts of somatic tissues, we classified iPSCs with outlying gene expression levels for each gene (underexpression or overexpression of probabilistic estimation of expression residuals (PEER)-adjusted gene expression levels; z-score based criterion; Methods), which identified at least one outlying iPSC line for 17,514 genes. Next, we computed burden scores for gene-proximal rare variants (within the gene body or ±10 kb around gene), comparing outlier and non-outlier lines. Notably, both SNPs and indels were enriched in underexpression outliers for both SNPs and indels were enriched in underexpression outliers for

 needles classified iPSCs with outlying gene expression levels for each gene (underexpression or overexpression of probabilistic estimation of expression residuals (PEER)-adjusted gene expression levels; z-score based criterion; Methods), which identified at least one outlying iPSC line for 17,514 genes. Next, we computed burden scores for gene-proximal rare variants (within the gene body or ±10 kb around gene), comparing outlier and non-outlier lines. Notably, both SNPs and indels were enriched in underexpression outliers for

 needles classified iPSCs with outlying gene expression levels for each gene (underexpression or overexpression of probabilistic estimation of expression residuals (PEER)-adjusted gene expression levels; z-score based criterion; Methods), which identified at least one outlying iPSC line for 17,514 genes. Next, we computed burden scores for gene-proximal rare variants (within the gene body or ±10 kb around gene), comparing outlier and non-outlier lines. Notably, both SNPs and indels were enriched in underexpression outliers for
To place this enrichment of rare variants into context with previous studies that linked rare genetic variants to outlying gene expression in somatic tissues, we repeated the outlier analysis for singleton, high-CADD (CADD > 25) SNPs using consistently processed data from GTEx (v7). We considered 35 tissues with at least 50 samples (after removal of globally outlying samples as for iPSCs; Fig. 2b and Methods) and calculated the enrichment scores across 10,000 random draws of equal sample size for each tissue (n samples = 50). This identified iPSCs as the cell type with the largest enrichment score (median ~9), followed by GTEx fibroblasts (median ~7) and GTEx testis (median ~5; Supplementary Table 12). Enrichments were moderately correlated with the number of expressed genes in each tissue (Pearson r = 0.34); however, the overall patterns were retained when controlling for this effect (Extended Data Fig. 6).

**Leveraging iPSC transcriptome reference data to improve rare-disease diagnostics.** RNA-seq of blood and other accessible tissues has been used to prioritize putatively causal genes for rare diseases, by identifying genes with outlying expression patterns and pathogenic variants. Despite the prevalence of iPSC models in rare-disease research, such strategies have not previously been deployed to iPSCs due to lack of sufficiently large reference collections. To assess the potential of i2QTL for this task, we used a set of 65 iPSC lines derived from individuals with rare genetic disorders for which the causal gene was clinically annotated (n = 15 unique...
genes, n = 3 unique diseases; Supplementary Table 1) and that were part of the HipSci collection.

When considering known causal genes in the rare-disease samples, we observed outlier (absolute (abs) value of the z-score > 2) gene expression in 12.3% of disease gene-sample pairs, compared to 3.75% of gene-sample pairs for non-disease-associated genes matched for expression level (n = 1,138,345; Fisher’s exact test: odds ratio, 3.59 (CI: 1.48–7.57); P = 0.002; Methods and Fig. 2c,d). We performed the same analysis for splicing outliers; we compared the fraction of splicing outliers for individuals with rare disease in known causal genes compared to noncausal genes, and observed a twofold enrichment for splicing outliers (abs(z-score) > 2) in disease causal genes (Fisher’s exact test: odds ratio, 1.97 (CI: 1.15–3.18); P = 0.01). Finally, we computed an integrated odds ratio by combining gene- and splicing-level outliers, and observed an almost fivefold enrichment for known rare-disease genes compared to non-disease genes (Fisher’s exact test: odds ratio, 4.83 (CI: 1.76–13.43); P = 0.0009). Focusing on the rare disease hereditary cerebellar ataxia as an example, we compared expression of the known disease gene (CACNA1A) in iPSCs to that in GTEx, and found that the gene was only expressed (fragments per transcript kilobase per million fragments mapped ≥ 1) in iPSCs and GTEx tissues that are difficult to biopsy clinically (including brain tissue, testis and fallopian tube; Extended Data Fig. 7a). More generally, considering a broad range of curated disease genes (OMIM32), we observed a larger number of disease genes expressed in iPSCs compared to whole blood and other GTEx tissues (Extended Data Fig. 7b), highlighting the utility of iPSC transcriptomes to model the effects of pathogenic variants across diverse rare diseases.

We further tested whether i2QTL transcriptomic data can improve the prioritization of putatively causal genes when combined with blood RNA-seq profiles. Briefly, we generated RNA-seq data from an iPSC line derived from an individual with a validated KCNQ1 splicing defect (Methods), for whom blood expression profiles had previously been generated31. Outlying gene expression patterns in blood alone (compared to RNA-seq data from 244 reference samples from Fréard et al.3) yielded 626 candidate disease genes with at least one outlier splicing junction (abs(z-score) > 2).

Notably, the intersection of outlier splicing pattern in blood and iPSCs resulted in a set of only 44 genes—an approximately 14-fold reduction in the number of candidate disease genes for further curation, and containing the known causal disease gene (Fig. 2e). This highlights a generalizable approach enabled by i2QTL reference data to enhance outlier detection in individuals with rare diseases.

GWAS variants from multiple diseases have molecular impacts in pluripotent cells. A major opportunity provided by eQTL maps in iPSCs is to identify colocalization events with GWAS loci, which could point to developmental or transient regulatory mechanisms. Using a combination of FINEMAP46 and eCAVIAR47 (Methods), we systematically assessed colocalization between eQTLs for all five RNA traits and a broad range of previously reported genetic associations obtained from diverse GWAS studies, the Phenome Scanner Database (v2)48, the NHGRI-EBI GWAS catalog49 and GWAS curated in LocusCompare50 for a combined total of 350 GWAS and, additionally, the 1,740 traits from the UK Biobank (UKBB) phase 1 GWAS (http://www.nealelab.is/uk-biobank/).

In total, we identified 4,336 colocalization events (Methods), linking 608 disease and phenotype loci to 10,794 cis-eQTL (Fig. 3a and Supplementary Table 13). Although gene-level eQTLs represented the majority of colocalizations, 41% of these exclusively colocalized with non-gene-level eQTLs (Fig. 3b). For example, 36 of 93 GWAS loci for a coronary artery disease (CAD) GWAS51 had evidence for colocalization with an iPSC eQTL, involving different RNA traits (Extended Data Fig. 8a). Next, we assessed which diseases had the largest numbers of colocalization in iPSCs, relative to the total number of GWAS loci, which identified primary biliary cirrhosis (10/12 GWAS loci), followed by triglyceride levels (12/15 GWAS loci), as iPSC-linked traits. The colocalized genes for primary biliary cirrhosis were enriched for mitogen-activated protein kinase, nuclear factor-κB and tumor necrosis factor receptor 1 signaling pathways (gProfiler P adj: 1.1 × 10−2, 4.9 × 10−2 and 5.725 × 10−3, respectively), with known functions in the immune system matching the disease. Enrichment analysis for triglyceride colocalized genes showed significant overlap with metabolic pathways (gProfiler: 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 × 10−3), again matching the known disease biology. We also observed colocalization events for genes that were identified as trans regulators in iPSCs (Methods, Supplementary Results and Supplementary Table 14). For example, the most significant trans-eQTL (P adj: 1.08 × 10−15) associated with changes in expression of NBPF14, known to be frequently mutated in breast cancer52, colocalized to a GWAS hit (rs11249433:A>G) for breast cancer.

Finally, we compared the gene-level colocalizations in iPSCs to GTEx, focusing on 452 GWAS that were assessed in the LocusCompare study46, which has used consistent colocalization methodology using eQTLs from GTEx tissues. Among the 7,042 colocalization events in aggregate across all eQTL maps, 836 events were exclusively detected in iPSCs (Fig. 3c and Supplementary Table 15). Notably, 47 of these iPSC-specific colocalization events were associated with genes that lacked an eQTL in GTEx tissues, and 231 colocalizations were due to iPSC-specific eQTL signals as identified from the mashR analysis. For example, we identified an iPSC-specific colocalization event for POLR1B and a GWAS variant for height (rs7586668:C>T53, Extended Data Fig. 8b). Recently, a mutation in this gene in zebrafish has been shown to give rise to altered body size54. Collectively, these new colocalizations substantially increased the number of linkages between GWAS loci and eQTLs for traits such as CAD55 (50% new colocalizations; Fig. 3d), Parkinson’s disease56 (20% new colocalizations) and Alzheimer’s disease59 (7.5% new colocalizations). Moreover, 74 of the iPSC-specific colocalizations were linked to 31 traits that had no prior evidence for eQTL colocalization.

Outlier rare variants in iPSCs have large impacts on diverse complex traits. We leveraged the map of GWAS and eQTL colocalization to prioritize genes that are more likely to harbor rare variants that are associated with expression outliers in iPSCs and affect a specific trait.

Specifically, we intersected our catalog of outlier-associated rare variants with GWAS summary statistics for matched traits contained in the UKBB phase 1 GWAS, resulting in 10,103 outlier-associated variants linked to 779 genes (Methods). We then compared these outlier variants with matched non-outlier variants stratified by the colocalization posterior probability (CLPP) score of the corresponding gene (Methods). Globally, outlier-associated rare variants for genes with evidence of GWAS and eQTL colocalization were associated with more significant GWAS P values for the corresponding traits (Fig. 4a). This enrichment was stronger for increasing CLPP score (Methods; CLPP > 0, P = 0.02; CLPP ≥ 0.2, P < 1 × 10−16), and consistent trends were observed when considering GWAS effect sizes instead of statistical significance (Extended Data Fig. 9).

Overall, from the starting list of outlier and matched non-outlier variants with a CLPP score > 0 (genes: n = 319; traits: n = 543), we identified 48 (8.8%) traits with at least weak evidence of colocalization (CLPP ≥ 0.01), comprising 58 unique outlier-associated variants proximal to 35 genes (Supplementary Table 16).

Among these, we observed an example of an outlier-associated rare variant rs189811790:A>G in HSD17B12, a gene known to be involved in type 2 diabetes mellitus57, for basal metabolic rate

NATURE GENETICS  |  VOL 53  |  MARCH 2021  | 313–321  |  www.nature.com/naturegenetics

ARTICLES
Fig. 3 | Colocalization of disease and trait variants with iPSC eQTLs. a. Overview of colocalization events with iPSC eQTLs, depicting the total number of colocalization events across 350 GWAS and the UKBB, with colors denoting the trait categories. ICD-10, International Classification of Disease (10th revision). b. Colocalization events for each cis-eQTL type, displaying the number of GWAS loci with colocalization events that were either specific to a given eQTL type (light color; not detected by any other eQTL type) or shared with at least one other eQTL type (dark color). c. Overlap between i2QTL and GTEx GWAS colocalization events for gene-level eQTLs, considering the number of unique gene-colocalization pairs. d. For the Howsen et al. CAD GWAS, iPSC eQTLs resulted in a 50% increase in the number of colocalization events with disease loci compared to GTEx eQTLs alone. iPSC associations are shown in orange, GTEx in blue, and GTEx tissues with two or more genes implicated are shown in the dashed box.
The red dashed line indicates the Bonferroni $P$-value cutoff. Points are colored by LD (1000 Genomes European) relative to the lead variant (smallest $P$ value; purple diamond) in the gene locus.

**Fig. 4 | Integration of common-variant colocalization analyses with outlier-associated rare variants.**

**a.** Negative log$_{10}$ $P$ values of GWAS trait associations for iPSC outlier- and non-outlier-associated variants, considering genes with varying degree of evidence for colocalization with common eQTL variants. Genes were stratified by CLPP score. Overall, outlier-associated variants had more significant effects in GWAS in which there was evidence for colocalization of the same genes compared to matched non-outlier variants, increasing with CLPP. Dots denote median values; error bars indicate 95% CIs of empirical data range. $P$ values were derived from a one-sided Wilcoxon test (CLPP $> 0$, $P = 1.3 \times 10^{-2}$; CLPP $\leq 0.1$, $P = 4.0 \times 10^{-10}$; CLPP $\geq 0.2$, $P = 7.1 \times 10^{-4}$).

**b.** Example gene locus with two outlier-associated variants (highlighted in blue), which exhibit the largest protective effect sizes among all outlier and non-outlier samples (gray) mapping to a gene. Color denotes LD (1000 Genomes European) relative to the lead variant (smallest $P$ value; purple diamond) in the gene locus.

**c.** Example gene locus highlighting an outlier-associated variant (blue) with the largest protective effect sizes among all outlier and non-outlier samples (gray) mapping to DENND1B. Points are colored by LD (1000 Genomes European) relative to the lead variant (smallest $P$ value; purple diamond) in the gene locus.

**d.** $P$ value rank for SNP rs11589930:C>A across all UKBB phase 1 GWAS ($n = 2,419$). The red dashed line indicates the Bonferroni $P$-value cutoff.

The outlier-associated rare variant in this gene had one of the largest protective effect sizes within 1 Mb around the locus (overall SNP rank: 30/4,065; top 0.73%), and it was among the top effects genome wide (top 0.8% across all SNPs). However, owing to its low frequency, this variant did not pass conventional thresholds for genome-wide significance ($P = 0.003$). Another candidate was observed for outlier-associated rare variant rs11589930:C>A linked to gene DENND1B, previously implicated in cholangitis, for which we identified a rare variant associated with gene- and transcript-level effects in human iPSCs (rs189811790:G>A). The GWAS effect size ($-\log_{10}(P) = 1.3 \times 10^{-7}$) for this variant was genome-wide significant ($P = 7.2 \times 10^{-4}$; CLPP $\geq 0.1$, $P = 4.0 \times 10^{-10}$; CLPP $\geq 0.2$, $P = 7.1 \times 10^{-4}$).

### Discussion

Genetic effects in pluripotent cells can elucidate the spectrum of traits that may manifest during development and across cell differentiation. To maximize the power for such genetic analyses, we harmonized population-scale iPSC genetic and transcriptomic datasets across five studies. The scale of our resource, spanning transcriptomic and genomic profiles from iPSCs derived from close to one thousand unique donors, has enabled the mapping of cis-eQTLs across a comprehensive range of RNA traits, the identification of trans-eQTLs and the study of rare-variant effects and their collective impacts on genetic traits and diseases.

We identified cis-eQTLs across five RNA traits, yielding regulatory variants for 67.2% of expressed genes in human iPSCs ($n = 21,548$). This included 995 cis-eGenes that were not previously reported in eQTL maps from somatic tissues. Next to cis-eQTLs, we identified 193 trans-eQTLs, a substantial fraction of which (91/193) were linked to non-gene-level eQTLs, which supports the relevance of eQTL variants acting on splicing, transcript isoforms, exons or APA.

Outlier gene expression can aid in detection of rare variants and disease genes. We observed increased power to discover outlier-associated rare variants in iPSCs compared to somatic tissues. We further demonstrated how population-scale iPSC transcriptomic data enable prioritization of disease genes from individuals with known rare genetic disorders. In a collection of rare-disease samples that are part of our study, we identified a fivefold enrichment of outliers in known rare-disease genes and...
demonstrated detection of gene outliers in individuals with mono-
genetic diabetes, Bardet–Biedl syndrome and hereditary cerebellar ataxia. These results demonstrate how iPSC transcriptomic data from a large control cohort such as i2QTL can be directly utilized for rare-disease identification, even before generating disease-specific differentiated cell types.

The large-scale eQTL maps enabled the generation of a comprehensive colocalization map between regulatory variants in human pluripotent cells and complex human traits. We annotated over 4,400 GWAS-driver loci (from the 29,666 assessed), originating from over 600 traits, to eQTLs in iPSCs. We observed unique colocalized loci across a range of traits, from physical traits to diseases and lab measurements, including 836 colocalizations present exclusively in iPSCs. Among these, we found colocalizations for developmental traits, such as congenital craniofacial abnormalities and heritable cancers. Lastly, by integrating our colocalization results and rare variants linked to expression outliers, we demonstrated prioritization of variants with large impacts on traits measured in the UKBB.

Overall, the genetic maps and colocalization catalogs generated in this study form a valuable reference dataset, further aiding in the interpretation of risk variants in a unique cell type relevant for development, cellular differentiation, cancer and rare-disease research. We expect that the genetic maps presented here, in combination with the constantly growing GWAS and rare-disease resources, will reveal missing molecular underpinnings of complex and rare genetic diseases and traits manifesting during development.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of information, acknowledgements, peer review information; details of data and code availability are available at https://doi.org/10.1038/s41588-021-00800-7.

Received: 4 October 2019; Accepted: 25 January 2021; Published online: 4 March 2021

References

1. Westra, H.-J. et al. Systematic identification of trans-eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238–1243 (2013).
2. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. 49, 131–138 (2017).
3. Zhernovava, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. 49, 139–145 (2017).
4. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017).
5. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013).
6. Alasoos, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. 50, 424–431 (2018).
7. Schwartzentruber, J. et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat. Genet. 50, 54–61 (2018).
8. Cuomo, A. S. E. et al. Single-cell RNA sequencing of differentiating iPSC cells reveals dynamic genetic effects on gene expression. Nat. Commun. 11, 810 (2020).
9. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nat. Genet. https://doi.org/10.1038/s41588-021-00801-6 (2021).
10. Sun, N. et al. Patient-specific induced pluripotent stem cells as a model for familial dilated cardiomyopathy. Sci. Transl. Med. 4, 130ra47 (2012).
11. Lan, F. et al. Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells. Cell Stem Cell 12, 101–113 (2013).
12. Lee, J. et al. Activation of PDGF pathway links LMNA mutation to dilated cardiomyopathy. Nature 572, 335–340 (2019).
13. Kodo, K. et al. iPSC-derived cardiomyocytes reveal abnormal TGF-β signalling in left ventricular non-compaction cardiomyopathy. Nat. Cell Biol. 18, 1031–1042 (2016).
14. Wu, H. et al. Modelling diastolic dysfunction in induced pluripotent stem cell-derived cardiomyocytes from hypertrophic cardiomyopathy patients. Eur. Heart J. 40, 3685–3695 (2019).
15. 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).
16. Sterneckert, J. L., Reinhardt, P. & Schöler, H. R. Investigating human disease using stem cell models. Nat. Rev. Genet. 15, 625–639 (2014).
17. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs, Nature 546, 370–375 (2017).
18. 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 Rep. 8, 1086–1100 (2017).
19. 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 (2017).
20. Banovich, N. E. et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 28, 122–131 (2018).
21. Carcamo-Orrive, I. et al. Analysis of transcriptional variability in a large human iPSC library reveals non-genetic determinants of heterogeneity. Cell Stem Cell 20, 518–532.e9 (2017).
22. Fréard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. 25, 911–919 (2019).
23. Li, X. et al. The impact of rare variation on gene expression across tissues. Nat. Genet. 550, 239–243 (2017).
24. 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).
25. Thomas, S. M. et al. Reprogramming LCLs to iPSCs results in recovery of donor-specific gene expression signature. PLoS Genet. 11, e1005216 (2015).
26. Dong, C.-H., Chornowska, A., D’Antonio, M. & Farnet, K. A. Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants. Nat. Commun. 11, 953 (2020).
27. Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nuclear Acids Res. 39, D945–D950 (2010).
28. Gerrard, D. T. et al. An integrative transcriptomic atlas of organogenesis in human embryos. eLife 5, e15657 (2016).
29. 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).
30. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nuclear Acids Res. 47, D1005–D1012 (2019).
31. Jakubosky, D. et al. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. Nat. Commun. 11, 2928 (2020).
32. 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).
33. 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).
34. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. 7, 500–507 (2012).
35. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. Science 369, eaaz3900 (2020).
36. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. 9, eaal5209 (2017).
37. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. 8, 15824 (2017).
38. Kernohan, K. D. et al. Whole-transcriptome sequencing in blood provides a diagnostic of spinal muscular atrophy with progressive myoclonic epilepsy. Hum. Mutat. 38, 611–614 (2017).
39. McKusick, V. A. Mendelian Inheritance in Man: a Catalog of Human Genes and Genetic Disorders (IU Press, 1998).
40. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016).
41. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99, 1245–1260 (2016).
42. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. Bioinformatics https://doi.org/10.1093/bioinformatics/btz469 (2019).
43. Liu, B., Glaudemons, M., J. Bao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression confound GWAS follow-up. Nat. Genet. 51, 768–769 (2019).
44. van der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. Circ. Res. 122, 433–443 (2018).
45. Liu, J. Z. et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137–1141 (2012).
46. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
47. 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).
48. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
49. Sanchez, E. et al. POLR1B and neural crest cell anomalies in Treacher Collins syndrome type 4. *Genet. Med.* **22**, 547–556 (2020).
50. Howson, J. M. M. et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* **49**, 1113–1119 (2017).
51. Pankratz, N. et al. Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. *Ann. Neurol.* **71**, 370–384 (2012).

52. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
53. Scott, R. A. et al. An expanded genome-wide association study of type 2. *Diabetes* **66**, 2888–2902 (2017).
54. Webb, G. J., Siminovitch, K. A. & Hirschfield, G. M. The immunogenetics of primary biliary cirrhosis: a comprehensive review. *J. Autoimmun.* **64**, 42–52 (2015).
55. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

HipSci Consortium

Marc Jan Bonder¹²³, Helena Kilpinen¹²¹⁴¹⁷¹⁸, Daniel Seaton¹ and Oliver Stegle¹²³¹²²³

A full list of members and their affiliations appears in the Supplementary Information.

iPSCORE consortium

Matteo D'Antonio⁹, Kelly A. Frazer⁹¹⁹, David A. Jakubosky⁷⁸ and Erin N. Smith¹⁹

A full list of members and their affiliations appears in the Supplementary Information.

Undiagnosed Diseases Network

Chunli Zhao¹⁶, Diane B. Zastrow¹⁶, Devon E. Bonner¹⁶ and Matthew T. Wheeler¹¹¹⁶

A full list of members and their affiliations appears in the Supplementary Information.

PhLiPS consortium

Christopher D. Brown²¹ and YoSon Park²¹

²¹Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
**Methods**

**Dataset information.** Within the i2QTL consortium, we reprocessed existing and newly generated transcriptomic and genomic data from iPS cell lines from five studies. A short description on each of the analyzed studies is provided in the Supplementary Methods. The references to the data sources are available in Supplementary Table 1.

**Genotype and RNA data processing.** In brief, all data, array-based genotypes, WGS and RNA-seq, were homogenously reprocessed from the raw data deposited in the respective repositories (Supplementary Table 1). Array-based genotypes from all cohorts were quality controlled and imputed against a combined reference of UK10K and 1000 Genomes. WGS data from HiSeq and iPSCORE were jointly reprocessed to perform joint variant calling across the two cohorts. RNA-seq data were homogenously processed with study-level quality-control metrics and read mapping using STAR, followed by gene and exon expression quantification using featureCounts. Salmon was used to quantify transcript levels and ratios, and leafCutter was used to quantify splicing levels. APA ratios were quantified as described by Zhernakova et al.3. Full details on the raw data processing are provided in Supplementary Methods.

**PEER correction and optimization.** We used PEER3 to adjust for transcriptome-wide confounding sources of variation. We chose to not include known factors when estimating PEER factors, as metadata were sparse and not standardized across studies. We ran PEER (v1.3) on normalized gene-level quantifications, considering genes with a transcripts per million (TPM) > 2. We assessed the impact on number of estimated PEER factors on eQTL mapping as quantified by the number of eGenes detected (genes with at least one eQTL at FDR < 5%; Extended Data Fig. 10). We used 50 PEER factors for all analyses, reflecting a compromise between selecting a compact set of factors while maximizing eQTL detection power. To rule out that PEER factors themselves are subject to genetic regulation, we tested each factor association with genome-wide variants, and found no effect (FDR > 10%).

**Quantitative trait loci mapping.** For eQTL mapping, both in cis and trans, we used a linear mixed model implemented in LIMIX (v2). This model allowed controlling 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)3, considering independent variants with MAF > 5%. Fifty PEER factors, derived from gene-level abundance, were included as fixed-effect covariates in all analyses (see above). eQTL mapping was performed using log-transformed standardized expression levels when considering both gene-level and exon-level data; for the other RNA traits, the ratio-based traits, we used an arcsine transformation to approximately stabilize the variance of each trait. Significance of the eQTL SNP was assessed using a likelihood-ratio test.

To control for multiple testing, we used an approximate permutation scheme as in work by Ongen et al.2. Briefly, for each gene, we obtained an empirical null distribution of P values from 1,000 permutations while retaining covariates, kinship values and expression values. Subsequently, we fit a parametric beta distribution to the most significant P value per gene for each permutation to interpolate the null distribution. Using this null model, we estimated adjusted P values for cis regions for eQTL lead variants. When multiple features per gene were tested (that is, for transcript ratio, exon level, splicing ratio and APA ratio in cis-eQTLs, herein features), the FDR was controlled at a gene level, using an additional Bonferroni correction for the number of features per gene. To control for multiple testing across genes, we used Storey’s Q-value procedure29 to control for the genome-wide FDR.

**cis-eQTL mapping.** For cis-eQTL mapping, we considered common variants (MAF > 1%) in proximal regions of genes (variants within 250kb of the gene body). To limit technical factors of variation, only paired-end stranded samples from European ancestry were used (n = 716 donors; n = 932 lines). Significant eQTLs were reported at a gene-level FDR < 5%. For all RNA traits, two trait inclusion criteria were used for genetic analyses. Traits were considered that were expressed (that is, non-zero expression) in at least 25% of the samples. For the splicing and APA eQTLs, we required at least 50% of the samples to have an expression value (or ratio) > 0. The assessed genes per eQTL type are summarized in Supplementary Table 17. For gene-level eQTLs, we additionally tested for higher-order eQTLs using iterative eQTL mapping. Lead eQTL variants were accounted for as covariates in subsequent mapping iterations until no additional independent cis-eQTLs were identified (Extended Data Fig. 2).

Using information from GTEx, BIOS and results from previous iPSC eQTL studies, we assessed the replication and annotated the identified cis-eQTLs. See Supplementary Methods for details, including alternative replication strategies using mashr (v0.22.1).

**Trans-eQTL mapping.** Trans-eQTLs were identified using an analogous approach as for cis-eQTL mapping, considering common variants (MAF > 1%) in distal regions of genes, defined as at least 2.5 Mb upstream and downstream of the gene transcription start and end sites. Given the potentially large number of tests when assessing all variant gene pairs in an exhaustive manner, we chose to limit the trans-eQTL tests to the union of cis-eQTL lead variants discovered in our study and known GWAS-implicated variants (obtained from the NHGRI-EBI GWAS catalog v92), yielding 115,709 variants to test for trans-eQTLs. These variants were tested for associations with 17,039 expressed protein-coding genes (TPM ≥ 1 in at least 25% of the samples). To maximize power for the trans-eQTL discovery, we used all samples from European ancestry (n = 743; lines = 1,120), and given the even larger number of tests for the other RNA traits and the larger impact of sequencing differences, we chose not to test trans-eQTLs on other RNA traits.

To reduce the possibility of spurious associations, a more conservative quantile normalization to a Gaussian distribution was used, and we included the lead cis-eQTL variant as an additional fixed-effect covariate in the model. To avoid spurious associations caused by read cross-mapping,36 we excluded gene combinations with high sequence similarity from the trans analysis. Briefly, we used primary and secondary mappings of the RNA-seq reads to the genome to construct such a blacklist. Any secondary mapping to another gene was reason to exclude the specific gene pair for trans-eQTL mapping (n = 66,964 gene pairs; Supplementary Table 18). This cross-mapping blacklist was obtained based on all paired-end stranded data only (Supplementary Table 2). Additionally, we excluded variants within the human leukocyte antigen region, due to its complex LD structure.

We considered the left-out RNA-seq samples, single-cell RNA-seq data from the preprocessed single-cell RNA-seq analysis by Cusono et al.27 and iPSC methylation information on the HiSeq samples to assess the replication of the discovered trans-eQTL effects (Supplementary Methods).

**Outlier analysis.** Complementary to eQTL analyses of common variants (MAF > 1%), we considered effects of rare variants linked to transcriptomic outliers. Based on featureCounts TPM quantifications (log TPM), we considered autosomal protein-coding and long noncoding RNA genes for outlier analysis. Cell lines from donors with predicted ancestry other than the European super-population were discarded, and we additionally limited the analysis to lines with paired-end RNA-seq data. Genes were then filtered for minimal expression, defined as gene expression TPM > 0 in 50% or more in each study.

To adjust for transcriptome-wide confounding sources of variation, PEER3 correction was run on the filtered data as described above (n = 50 PEER factors). The resulting residual expression profiles were scaled and centered (z-score normalization). As an additional quality-control step, we tested for consistent overexpression or underexpression, that is, cell lines found to be the most underexpressed or overexpressed across hundreds of genes. Cell lines with an expression absolute z-score > 2 in more than 100 genes were discarded from subsequent analyses (n = 21 lines). Finally, cell lines were retained if WGS data were available in addition to RNA-seq data, leaving data from the HiSeq and iPSCORE projects only. After applying these quality-control steps, 17,514 genes and 377 cell lines remained for full data analysis.

To prepare the WGS genotype data SNP and indel variants for the analysis, variants were filtered based on the variant quality score recalibration method using a tranch of cutoff of 99%. The software vcfanno35 (v0.2.9) was used to annotate the WGS vcf file with MAF from gnomAD (v2.0.2) and CADD score from CADD (v1.3). For trans-eQTLs, variants were filtered or per-sample level to test at least one alternate allele. Variants were then linked to genes using the bcftools68 (v1.11) window command, selecting a maximum distance of 10 kb based on the Ensembl 75 GTF reference. A separate file was produced for each cell line consisting of the following columns: cell line ID; gene ID; chromosome; position; gnomAD MAF; CADD (phred); and CADD (raw).

To facilitate comparative analysis using GTEx (v7) tissues, i2QTL data were reprocessed to match the GTEx (v7) pipeline to limit technical variation. For this specific analysis, RNA-SeQC (v1.1.8) expression quantification was used, and a separate PEER analysis was run to correct for technical variation, including known factors. As before, the top 50 PEER factors were selected to adjust the i2QTL data. For GTEx (v7) tissues with ≤ 150 samples, 15 PEER factors were used; for tissues with ≤ 250 samples, 30 PEER factors; and for tissues with > 250 samples, 35 PEER factors. GTEx (v7) WGS variants were annotated with MAF from gnomAD (v2.0.2) and CADD scores from CADD (v1.3) using vcfanno. GTEx (v7) tissue samples with an expression absolute z-score > 2 in more than 100 genes were discarded. GTEx tissues were considered only if at least 50 samples were available (n = 35 tissues). Profiles of expression residuals from PEER adjustment were centered and scaled to generate expression z-scores.

**Outlier enrichment.** We considered the subset of lines with both RNA-seq and WGS data available (n = 625 cell lines in filtering; Supplementary Table 2), and focused on variants up to 10 kb upstream and downstream of long noncoding and long noncoding RNA genes. Gene expression outliers for a given gene were defined as samples with a minimum gene expression z-score (z-score < −2; underexpression outlier) or a maximum gene expression z-score (z-score > 2; overexpression outlier). Separate scores were computed for gene-level underexpression outliers and overexpression outliers. The reported enrichment score was calculated as the
ratio of the proportion of outlier lines with variants across several MAF/CADD bins compared to non-outlier lines. Specifically, enrichment here refers to the relative risk (RR):

\[
RR = \frac{O'}{O}
\]

where \(O'\) denotes the number of outlier lines with \(\geq 1\) variant in or near (\(\pm 10\) kb upstream or downstream of a gene body) a gene passing given MAF and CADD thresholds; \(O\) is the total number of outlier lines; \(O'\) is the number of non-outlier lines with \(\geq 1\) variant in/near a gene passing given MAF and CADD thresholds; and \(O'\) is the number of non-outlier lines. The RR is reported with 95% Wald CIs derived from the asymptotic distribution of the log RR:

\[
\log(\text{s.e.}) = \left(\frac{1}{O'} - \frac{1}{O}\right) + \left(\frac{1}{2O'} - \frac{1}{O}\right)
\]

where s.e. is the standard error. The bounds on the CIs are defined as follows:

- \(\text{maxCI} = RR \times \exp(1.96 \times \log(\text{s.e.}))\)
- \(\text{minCI} = RR \times \exp(-1.96 \times \log(\text{s.e.}))\)

The analysis was performed separately for SNPs, indels and SVs, and across different MAF bins (from common to rare) and CADD bins (progressively more deleterious variants). Expression outlier direction (that is, underexpression and overexpression) was tested separately. For example, for underexpression outliers, an outlier line was defined as the least-expressed line in a given gene that also had a \(z\)-score \(< -2\). Consequently, a gene was defined as having at most 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 i2QTL and GTEx data.

Colocalization of GWAS loci with induced pluripotent stem cell eQTLs.

For colocalization analyses, we considered two sets of curated GWAS summary statistics: (1) UKBB rapid GWAS results \((n = 1,740\) traits\)\(^{23}\) and (2) publicly available GWAS results from the NHGRI-EBI GWAS catalog\(^{30}\) and PHENOMESCANNER (v2) \((n = 350\) studies), obtained through a wide variety of studies and consortia\(^{39}\). For each trait and study, we iteratively selected loci with a GWAS \(P\) value \(< 5 \times 10^{-8}\) and located at least 1 Mb away from previously selected (more significant) loci for the same trait and study. Among these GWAS loci, we selected those with at least one significant cis-eQTL, for any of the quantified RNA traits, within 10 kb of the lead GWAS variant at a FDR \(< 5\%\). Owing 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 with a known eQTL, for computational feasibility. Given the presence of several different types of eQTLs 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 eQTL traits originating from single or multiple genes.

We next tested each pair of GWAS locus and eQTL feature in our set. For each locus pair, we considered variants that were contained in both the GWAS and eQTL summary statistics. Loci with less than five common SNPs were discarded. We additionally discarded loci for which the minimal GWAS \(P\) value for the intersecting variants was greater than \(5 \times 10^{-6}\), and loci for which the adjusted \(\epsilon\)-TL \(P\) value was greater than 0.05. LD between SNP pairs was estimated based on 1000 Genomes phase 3 \((n = 2,504)\)\(^{32}\). We then applied FINEMAP\(^{30}\) separately to the GWAS and eQTL summary data to estimate posterior probabilities of causality for each SNP, and we combined these probabilities to compute a CLPP score following the approach outlined in eCAVIAR\(^{40}\).

We note that there was a relationship between the LocusCompare CLPP scores and the number of intersecting variants observed at a locus. To improve the comparability of colocalization events from different studies, we used an adaptive threshold, which accounts for the observed differences in the number of overlapping variants. Specifically, we considered a locus to be colocalized if it passed one of three CLPP and SNP thresholds: (1) \(5\) or more variants at a locus: CLPP of 0.5; (2) \(10\) or more variants at the locus: CLPP of 0.1; or (3) \(25\) or more variants at a locus: CLPP of 0.01. Further details and a discussion on the relation between CLPP and the number of SNPs can be found in work by Hormozdiari et al.\(^{12}\).

To compare colocalization results between GTEx tissues and i2QTL, we obtained colocalization results from LocusCompare and applied the same filtering strategies to select high-confidence colocalizations. Overlap of colocalization events between GTEx tissues and i2QTL was assessed based on the level of study, trait and eGene pairs.

To link trans-eQTL to GWAS loci, we performed an extended trans-eQTL mapping, linking all variants within 250 kb around the identified trans-eQTL variants to the identified downstream genes. This trans-eQTL information was subsequently used to perform a colocalization analysis as detailed above for cis-eQTL (Supplementary Results).

Annotating rare variants using UK Biobank GWAS summary statistics. To test for differences in rare variants associated with outliers and non-outliers and risk for disease and traits, we overlapped i2QTL WG3 variants (as described in ‘Outlier analysis’) with those measured or imputed in UKBB GWAS\(^7\). Specifically, we considered variants with gnomAD MAF < 1\% and CADD > 0. Multi-allelic variants were discarded. Variants were retained if they were observed in only one iQTL individual, for outlier-associated variants. This has the effect of isolating the set of variants putatively driving observed outlier gene expression (that is, should the same variant be observed in both an outlier and non-outlier sample; by definition, this would suggest that the variant is less likely to be causing the observed outlier expression). From this list of unique variants, outlier-associated variants were identified separately for underexpression and overexpression outlier samples; therefore, there could be a maximum of two outlier samples per gene.

For each gene with \(\geq 1\) outlier sample, non-outlier-associated variants were chosen for each non-outlier sample if a variant had a CADD score within a range of \(\pm 5\) of outlier variants. Non-outlier samples were defined as samples with an expression absolute \(z\)-score \(< 1\) for a given gene. If a non-outlier sample had a larger 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 (that is, underexpression and overexpression). For integration of colocalization results, we used a subset of variants that linked to the set of genes with evidence of colocalization (CLPP score \(> 0\)). The final list of variants was linked to UKBB GWAS using the exact size and \(P\) values for each trait in UKBB phase 1 GWAS. After intersecting these datasets, we obtained 10,703 outlier- and non-outlier-associated variants linked to 779 genes and 2,419 traits.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used in the study are available from the Sequence Read Archive, dbGaP or the European Nucleotide Archive; the full data availability is provided in Supplementary Table 1. Supplementary Table 2 provides a description of the samples used in the study. Full summary statistics on significant eQTLs can be obtained from Zenodo (https://doi.org/10.5281/zenodo.4005576). The colocalization results are accessible from the LocusCompare portal (http://locuscompare.com/). We used eQTL summary statistics from GTEx (v7; https://gtexportal.org/home/datasets/) and the BIOS cohort (https://genenetwork.nl/biositolbrowser/). We further used GWAS summary statistics from the NHGRI-EBI GWAS catalog (Ensembl v92; https://www.ensmbi uk/), dbGaP (https://www.ncbi.nlm nih.gov/dbgap/), and GWAS studies aggregated in the LocusCompare study. We downloaded the GTEx colocalization results via LocusCompare. Other external data sources are referenced in the Methods and the main text. Source data are provided with this paper.

Code availability

Code produced for the following analyses is available on GitHub: eQTL mapping (https://github.com/VMBio/hipsc_pipeline/tree/master/limax_QTL_pipeline/), genetic expression outlier analysis (https://github.com/pscook/omniQTL) and colocalization analysis (https://github.com/mikegloudemans/ipsc-coloco).

References

56. Streeter, I. et al. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).
57. 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).
58. 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).
59. Knowles, J. W., Hao, K., Xie, W., Weedon, M. N. & Zhang, Z. Genetic and functional analyses identify NAT2 as a human insulin sensitivity gene. *Circulation* **128**, A10906 (2018).
60. 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).
61. 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).
62. 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).
63. Storey, J. D., Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
64. Ongen, H., Gloudemans, M., Hall, H., Weedon, M. N. & Loesch, D. Z. Do Battle, A. Fused transcript and coexpression analyses arising from RNA-sequencing alignment errors. *F1000Res.* **7**, 1860 (2018).
65. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcanno: fast, flexible annotation of genetic variants. *Genome Biol.* **17**, 118 (2016).
66. 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).

67. 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).

68. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 1000, 2078–2079 (2009).

69. Churchhouse, C. & Neale, B. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. https://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank/ (2017).

Acknowledgements

The authors 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: a strategic award from the Wellcome Trust and UK Medical Research Council (WT098503), the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant no. 664726 to M.J.B. and D.S.); National Institutes of Health (NIH: T32 LM012409 to C.S.; T35 LM01127 to D.J.; U01 HL107388-01, P30DK116074 and SPO 130829 to I.C.-O.; HL107442, DK105541, DK112155 to K.A.F.); Stanford Graduate Fellowship (to M.J.G.); National Natural Science Foundation of China (31970554 to X.L.); National Key R&D Program of China (2019YFC1315804 to X.L.); Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 to X.L.); EBI–Sanger Postdoctoral Fellowship (to N.C.); National Science Foundation Graduate Research Fellowship (to N.M.F.); and California Institute for Regenerative Medicine (GC1R-06673 to K.A.F.). Research in the Knowles laboratory is supported by NIH grants DK116750, DK120565, DK106236, DK107437 and HL107388. We thank the staff in the Cellular Genetics and Phenotyping and Sequencing core facilities at the Wellcome Trust Sanger Institute. Work at the Wellcome Trust Sanger Institute was further supported by Wellcome Trust grant WT090851. Several datasets used in this study were sourced from the NextGen Consortium, available from the NCBI database of Genotypes and Phenotypes (dbGaP): iPSCORE (phs000924.v4.p1 and phs001325.v3.p1); PhLiPS (phs001341.v1.p1); and GENSEIPS (phs001139.v1.p1). This work, in part, used supercomputing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by NIH S10 instrumentation grant S10OD23452. Research reported in this manuscript was in part supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under award no. U01HG007708. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions

The main analyses and data preparations were performed by M.J.B. and C.S.; D.J. and M.D. performed SV calling and analysis; N.M.F. performed GTEx (v7) data processing; I.C.-O. and X.L. performed rare-variant annotation and interpretation; H.K. and D.V. annotated and validated the rare-disease variants and genes for the HipSci rare-disease samples; M.I.G. performed the colocalization analysis; M.J.B., D.S., B.M. and D.H. developed the eQTL software; C.Z. generated iPSC lines for Undiagnosed Diseases Network rare-disease samples; N.C., I.C.-O. and Y.P. assisted with data processing and analysis; M.J.B., C.S., M.J.G., S.B.M. and O.S. wrote the manuscript; I.C.-O., N.C., N.M.F., K.A.F., I.F., M.J.G., D.J., M.T.W., D.B.Z. and B.M. assisted in editing the manuscript; M.J.B., C.S., E.S., K.A.F., O.S. and S.B.M. conceived and oversaw the study.

Competing interests

S.B.M. is a member of the Scientific Advisory Board of Myome. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00800-7. Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00800-7.

Correspondence and requests for materials should be addressed to M.J.B., C.S., S.B.M. or O.S.

Peer review information *Nature Genetics* thanks Gabriel Hoffman and Vijay Sankaran for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Replication and consistency of effect sizes of eQTL discovered in the the original iPSC studies and replicated in i2QTL. Shown are scatter plots between eQTL effect size estimates in this study (i2QTL, x-axis) versus effect size estimates from previous studies (y-axis). Dots correspond to eQTL discovered in the respective study. Black: eQTL with consistent effect direction. Red: eQTL with discordant effect direction. Replication defined at nominal $P < 0.05$ in i2QTL. 

**a**, Replication of HipSci in i2QTL. 70.4% of the effects are replicated; 98% of the eQTL have concordant effect direction. Differences in the approach for estimating effect sizes result in the observed variation. 

**b**, Replication of iPSCORE in i2QTL. 81% of the effects are replicated; 98.7% of the eQTL have concordant effect direction. Notably only SNP eQTL were considered whereas SVs were not considered for replication. 

**c**, Replication of GENESiPS in i2QTL; 76.7% of the effects are replicated; 99.5% of the effects have concordant effect direction. 

**d**, Replication of PhiLiPS in i2QTL. 76.8% of the effects are replicated; 97.5% of the effects have concordant effect direction.
Extended Data Fig. 2 | Identification of multiple independent eQTL for the same gene using stepwise regression. a, Histogram of the number of independent eQTL effects identified for individual eGenes. Up to 12 independent effects were identified. b, Zoom-in view displaying the number of eGenes with 5 to 8 independent effects. c, Zoom-in view displaying the number eGenes with 8 to 12 independent effects.
Extended Data Fig. 3 | Sharing of lead eQTL signals between cell types and studies, considering i2QTL, GTEX and the iPS differentiation study from Cuomo et al. 

a, Pairwise sharing of lead eQTL signals in i2QTL (iPSC) and 48 GTEX tissues. Shown is the fraction of shared eQTL signals relative to the total number of common genes and lead eQTL variants in the two respective maps. Shared eQTL signals are defined as eQTL with concordant effect direction and absolute effect size within a factor of two (Methods).

b, Distribution of pairwise sharing as in A of iPSCs versus GTEX tissues (blue, \(N = 48\) comparisons) versus pairwise sharing between GTEX tissues (red, \(N = 2,401\) comparisons).

c, Pairwise sharing of lead eQTL signals in i2QTL (iPSC) and 48 GTEX tissues as in A, however additionally including single-cell eQTL from Cuomo et al. in iPSCs (iPSCsc), differentiated cell types (mesendo, endoderm).

d, Distribution of pairwise sharing as in C, considering iPSCs and differentiated cell types (bulk left, followed by iPSC single cell), mesendoderm (cyan), and endoderm (green) versus GTEX tissues (\(N = 48\) comparisons). During differentiation genetic signals in iPSC become more similar to those in GTEX tissues. Data in panels b and d are displayed as violin- and boxplot with the midpoint corresponding to the median, the lower and upper edges of the box to the first and third quartiles and the whiskers corresponding to the IQR \(\times 1.5\).
Extended Data Fig. 4 | Properties of distal (trans) gene-level eQTL. 

**a**, Dot plot of N = 862 trans-eQTL detected in iPSC (FDR < 10%). Dots correspond to individual trans-eQTL, with color denoting the variant category (blue: cis-eQTL, red: GWAS variant, purple: cis-eQTL and GWAS variant).

**b**, Breakdown of unique trans-eQTL variants (N = 193) across different variant annotations. Darker shaded colors denote trans-eQTL linked to variants that have more than one annotation; lighter shades correspond variants that can be unique assigned to a given variant category.

**c**, Mediation analysis of trans-eQTL, considering variants that are linked to a cis-eQTL. The outer pie denotes the annotation of the underlying trans variant: GWAS only variants (n = 7, white), cis-eQTL variants (n = 186, dark-blue). The inner pie displays results from the mediation analysis for individual trans variants: n = 7 trans-eQTL are exclusively linked to GWAS variants and hence not mediated ('GWAS only trans-eQTL'); n = 58 trans-eQTL variants are not significantly linked to mediation with any RNA trait ('Non-significant link to any cis-eQTL'); 86 are exclusively linked to gene-level abundance ('Only linked to gene-level cis-eQTL'); 7 are exclusively linked a RNA trait other than gene-level abundance ('Linked only to non gene-level cis-eQTL'); 35 are linked to gene-level abundance and at least one additional RNA trait('Linked both to gene-level and other level cis-eQTL').
Extended Data Fig. 5 | Analysis of the trans-eQTL hotspot at ELF2. a, Schematic representation of the genetic loci around ELF2 (left) and NAA15 (right). SNPs are annotated by evidence of cis-eQTL regulation on different traits (blue: transcript-ratio eQTL on ELF2, purple: gene-level eQTL on ELF2, green: splice eQTL on ELF2, brown: the APA eQTL on NAA15). b, LD structure between cis-eQTL variants implicated in the hotspot, annotated by cis-eQTL type as in a. c-f, Lead cis-eQTL effects by RNA trait for ELF2 and NAA15 across SNPs linked to downstream trans-eQTLs (n = 682 samples). Data are represented as a violin- and boxplot with the midpoint corresponding to the median, the lower and upper edges of the box to the first and third quartiles, the whiskers represent the interquartile range × 1.5. C. Splicing cis-eQTL on ELF2. d, Gene level cis-eQTL on ELF2. E. Transcript-ratio cis-eQTL on ELF2. F, Alternative polyadenylation cis-eQTL on NAA15. g, Co-expression network of genes that are controlled by the trans-eQTL linked to the hotspot at ELF2 (N = 37 genes), including ELF2 itself (center). Color of the bounding box around Genes denotes the cis-eQTL variant that drives the corresponding trans effect (colors as in d-g). Genes with multiple trans regulators are depicted with multiple colored rings. h, Replication of trans effects at ELF2 in a single cell differentiation study (Cuomo et al.). Shown are trans-eQTL effect sizes (beta’s) for the 12 ELF2-linked trans targets that show significant replication (defined as P < 0.05 and consistent effect direction) in any of the Cuomo et al tissues. From left to right: eQTL effect size in; the i2QTL study (discovery); in undifferentiated iPSC profiled using scRNA-seq; in mesendoderm profiled using scRNA-seq and in definitive endoderm profiled using scRNA-seq. Significant replication are indicated with a red asterisk.
Extended Data Fig. 6 | Enrichment for rare, deleterious SNPs in iPSC and GTEx tissues. Comparison of enrichments for singleton, highly-deleterious (CADD > 25) SNPs in iPSC versus GTEx v7 tissues analogous to Fig. 2b, however with an additional adjustment for the number of expressed genes. Displayed are enrichments for 10,000 random draws of 50 samples, controlling for the number of expressed genes (genes subset at a fixed number across tissues, N = 500 genes). Strongest enrichment is observed in iPSC. The data are represented as a boxplot where the middle line corresponds to the median, the lower and upper edges of the box corresponding to the first and third quartiles, the whiskers represent the interquartile range (IQR) ×1.5 and beyond the whiskers are outlier points.
**Extended Data Fig. 7 | Expression level of rare disease genes in iPSC versus GTEx tissues.**

**a**, Distribution of gene expression level of XX rare disease genes (log10(FPKM + 1)) in i2QTL iPSC and 17 GTEx tissues with a median expression level of at least 1 FPKM (red dashed line). Expression in i2QTL highlighted in teal, GTEx tissues in yellow. Disease genes are expressed in iPSCs and only difficult to biopsy tissues in GTEx display higher expression levels. n = 2,952 biologically independent samples. Data are represented as boxplots with the middle line corresponding to the median, the lower and upper edges of the box to the first and third quartiles, the whiskers to the interquartile range (IQR) ×1.5.

**b**, Expression level of genes in different curated gene lists, comparing i2QTL iPSC (left), all GTEx tissues (middle) and GTEx whole blood (right). Shown is the fraction of genes in each category for two expression bins: [0,1) FPKM expression absent or lowly expressed; [1,1e+12) FPKM gene expressed.
Extended Data Fig. 8 | Additional results from the colocalization analysis of eQTL and GWAS traits. a. Summary of colocalization results for the Coronary artery Disease GWAS (van der Harst et al 2018). 36 out of 93 GWAS loci were identified as colocalized with an eQTL of at least one RNA trait (38.7%). Shown are the number of colocalization events for eQTL of different RNA traits. From left to right: any eQTL type (All), gene-level eQTL (Gene), exon eQTL (Exon), transcript eQTL (Transcript), splicing eQTL (Splicing), APA eQTL (APA). The number of GWAS colocalization that are uniquely linked to a given RNA trait eQTL are displayed using a triangle and the total number of colocalizations per trait is depicted as a circle. b. Colocalization between a gene-level eQTL for POLR1B and a GWAS hit (rs7586668:C>T) for Height. Left: Manhatten plots displaying the local association signal for the eQTL on POLR1B (bottom) and the GWAS signal on Height (top). Right: Scatter plot of negative log P-values for the GWAS signal (x-axis) for BMI versus the POLR1B eQTL signal (y-axis) for the corresponding region.
Extended Data Fig. 9 | Enrichment for large-effect outlier-associated rare variants in colocalized genes. Absolute effect size of GWAS trait associations for iPSC outlier- and non-outlier-associated variants, considering genes with varying degree of evidence for colocalization with common eQTL variants.
Extended Data Fig. 10 | Optimization of the number of PEER factors to adjust for confounding expression heterogeneity. Shown is the number of genes with at least one gene-level eQTL (eGenes) for the top 3,000 highest expressed genes in iPSC for increasing number of PEER factors. PEER factors are adjusted for as additional fixed effect covariates. The vertical red line denotes the number of PEER factors considered in all i2QTL analyses.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size \((n)\) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. \(F\), \(t\), \(r\)) with confidence intervals, effect sizes, degrees of freedom and \(P\) value noted
  - Give \(P\) values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's \(d\), Pearson's \(r\)), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

| Data collection            |
|----------------------------|
| vcfanno 0.2.9              |
| bcftools 1.11              |
| mashr_0.2.21               |
| limix v2                   |
| verifyBamID 1.1.3          |
| picard 2.9.0               |
| edgeR V3                   |
| trimgalore 0.4.3           |
| GATK 3.7                   |
| featureCounts 1.6.0        |
| STAR 020201                |
| RNA-SeQC 1.1.8             |
| Salmon 0.8.2               |
| leafcutter 0.2.8           |
| GSEA 4.1                   |
| MSigDB 7.1                 |
| PEER 1.3                   |
| PLINK 1.07                 |
| FINEMAP 1.3.1              |
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in the study is available via SRA, dbGaP or ENA; the full data availability is provided in Table S1. Full summary statistics on significant eQTL can be obtained from https://zenodo.org/record/4005576 (doi:10.5281/zenodo.4005576). The colocalization results are accessible from the LocusCompare portal (http://locuscompare.com). We used eQTL summary statistics from GTEx (v7, available at: https://gtexportal.org/home/datasets); the BIOS cohort (available at: https://genenetwork.nl/biosqtlbrowser/). We further used GWAS summary statistics from the NHGRI-EBI GWAS catalog (Ensembl V92, available at: https://www.ebi.ac.uk/gwas/); Phenome scanner v2 (available at: http://www.phenoscanner.medschl.cam.ac.uk/); and GWAS studies aggregated in the LocusCompare study. Via LocusCompare we downloaded the GTEx colocalization results. Other external data sources are referenced in Methods and the main text.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
We aggregated all publicly-available iPSC data resources, without a per-determined sample size threshold. We compared eQTL discovery with previous studies (GTEx) and are confident that the current sample size provides sufficient power for detecting eQTLs and outlier expression in iPSC.

Data exclusions
Sample exclusion was performed based on low quality RNA/genotype information data, as described in the methods. No other sub-selections were performed. Exclusion criteria was pre-established.

Replication
Replication of the trans-eQTLs are performed based on left-out samples from other populations and DNA methylation data. Both described in the methods. cis-eQTL replication was assessed in GTEx, BIOS, and the original iPSC studies comprising i2QTL. trans-eQTLs were replicated in left-out samples in the Cuomo et al. studies and DNA methylation study in HipSci. Refer to manuscript for metrics on replication rates.

Randomization
N/A, a case/control design was not utilized in the present study.

Blinding
N/A, a case/control design was not utilized in the present study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | Antibodies            |
| ☒   | Eukaryotic cell lines |
| ☒   | Palaeontology and archaeology |
| ☒   | Animals and other organisms |
| ☒   | Human research participants |
| ☒   | Clinical data         |
| ☒   | Dual use research of concern |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | ChIP-seq              |
| ☒   | Flow cytometry        |
| ☒   | MRI-based neuroimaging |
## Eukaryotic cell lines

**Cell line source(s):** HipSci, iPSCORE, GENESiPS, PhLiPS, Banovich. Links provided in the manuscript.

**Authentication:** NA. Re-analysis of samples collected in previous projects.

**Mycoplasma contamination:** NA. Re-analysis of samples collected in previous projects.

**Commonly misidentified lines**

(See ICLAC register)

Human research participants

**Population characteristics**

The present study re-used existing data. Full information on how samples were collected, population characteristics, recruitment, and ethics can be found at the following references:

- **HipSci**
  Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature 546, 370–375 (2017)

- **iPSCORE**
  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)

- **GENESiPS**
  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)

- **PhLiPS**
  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)

- **Banovich**
  Banovich, N. E. et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 28, 122–131 (2018)

**Recruitment:** As above

**Ethics oversight:** As above

Note that full information on the approval of the study protocol must also be provided in the manuscript.