Molecular and functional variation in iPSC-derived sensory neurons

Jeremy Schwartzentruber[1*], Stefanie Foskolou[2], Helena Kilpinen[3], Julia Rodrigues[1], Kaur Alasoo[1], Andrew Knights[1], Minal Patel[1], Angela Goncalves[1], Rita Ferreira[2], Caroline Louise Benn[2], Anna Wilbrey[2], Magda Bictash[2], Emma Impey[2], Lishuang Cao[2], Sergio Lainez[2], Alexandre Julien Loucif[2], Paul John Whiting[2,4], HIPSCI Consortium (www.hipsci.org), Alex Gutteridge[2*], Daniel J. Gaffney[1*]

1) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom
2) Pfizer Neuroscience and Pain Research Unit, Pfizer Ltd., Great Abington, Cambridge, United Kingdom
3) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
4) AR-UK Drug Discovery Institute, Institute of Neurology, University College London, London, WC1E 6BT, United Kingdom

*Corresponding authors: Jeremy Schwartzentruber (js29@sanger.ac.uk), Alex Gutteridge (alex.x.gutteridge@gsk.com), Daniel Gaffney (dg13@sanger.ac.uk)

Abstract

Induced pluripotent stem cells (iPSCs), and cells derived from them, have become key tools to model biological processes and disease mechanisms, particularly in cell types such as neurons that are difficult to access from living donors. Here, we present the first map of regulatory variants in iPSC-derived neurons. We performed 123 differentiations of iPSCs from 103 unique donors to a sensory neuronal fate, and measured gene expression, chromatin accessibility, and neuronal excitability. Compared with primary dorsal root ganglion, gene expression was more variable across iPSC-derived neuronal cultures, particularly in genes related to differentiation and nervous system development. Single cell RNA-sequencing revealed that, although the majority of cells are neuronal and express the expected marker genes, a substantial fraction have a fibroblast-like expression profile. We found that the fraction of neuronal cells was influenced by the culture conditions of the iPSCs prior to the start of differentiation. Despite this differentiation-induced variability, applying an allele-specific method enabled us to detect thousands of quantitative trait loci influencing gene expression, chromatin accessibility, and RNA splicing. A number of these overlap with common disease associations, including known causal variants at *SNCA* for Parkinson's disease and *TNFRSF1A* for multiple sclerosis, as well as new candidates for Parkinson's disease and schizophrenia. Finally we show that recall by genotype studies of specific variants using iPSC-derived cells are likely to require sample sizes of 20-80 individuals to detect the effects of regulatory variants with moderately large (1.5- to 2-fold) effect sizes.

Introduction

Cellular disease models are critical for understanding the molecular mechanisms of disease and for the development of novel therapeutics. In principle, induced pluripotent stem cell (iPSC) technology enables the development of these models in any human cell type. Initial uses of iPSCs for disease modelling have focused mostly on highly penetrant, rare coding variants with large phenotypic effects (Itzhaki et al. 2011; Liu et al. 2011; Wainger et al. 2014; Lee et al. 2009; Cao et al. 2016). However, there is growing interest in using iPSCs to model the effects of the common genetic variants of modest effect size that drive complex disease (Warren, Jaquish, et al. 2017). A key question is to what extent variability in directed differentiation is a barrier to studying the effects of common disease-associated variants in iPSC-derived cells. In addition, because cultured cells are imperfect models of primary tissues, not all common disease-associated genetic variants also alter cell phenotypes in iPSC-derived systems.

Here, we present the first large-scale study of common genetic effects in a neuronal cell type differentiated from human stem cells, iPSC-derived sensory neurons (IPSDSNs). Peripheral sensory nerve fibres innervate the skin and other organs and are brought together at the dorsal root ganglia (DRG) before synapsing with the spinal cord around the dorsal horn. The development of efficient protocols to differentiate iPSCs into nociceptive (pain-sensing) neurons (Young et al. 2014) provides the opportunity to model common genetic effects on human sensory neuron function, which may underlie individual differences in pain sensitivity and chronic pain. We investigate how power to detect common genetic effects is affected by the variability introduced by differentiation and demonstrate how initial iPSC growing conditions influence cell phenotypes in IPSDSNs. We identify quantitative trait loci (QTLs) for gene expression, RNA splicing, and chromatin accessibility and identify a number of overlaps between molecular QTLs and common disease associations. In generating this gene regulatory map we establish effective techniques for using IPSDSN cells to model molecular phenotypes relevant to common diseases.

Results

Sensory neuron differentiation and characterisation

We obtained 107 IPS cell lines derived from unrelated apparently healthy individuals by the HIPSCI resource (Kilpinen et al. 2017), and followed an established small molecule protocol (Young et al. 2014) to differentiate these into sensory neurons of a nociceptor phenotype (Figure 1a). We performed a total of 123 differentiations; 13 of these were done with an early version of the protocol (P1) which was subsequently refined (P2) to reduce the number of differentiation failures and to yield a higher proportion of neuronal cells in the final cultures. One RNA-seq sample failed sequencing, and four others were outliers based on principal components analysis and were excluded (Supplementary Figure 1). This left a set of 119 differentiations with gene expression data from 100 unique iPSC donors; all subsequent analyses focused on the 106 P2 protocol samples, except for QTL calling, where we used all samples to maximize discovery power.

We clustered our gene expression data with 239 iPSC samples from the many of same donors, as well as 28 post-mortem DRG tissue samples from 10 different donors, and 44 primary tissues from the GTEx project (Mele et al. 2015) (Figure 1b). Globally, IPSDSN samples showed greatest similarity to iPSCs (gene expression correlation Spearman ρ=0.89), followed by DRG (ρ=0.84), and then brain samples from GTEx. However, because different gene expression quantitation methods were used in GTEx, we cannot be certain of relative similarities between GTEx tissues and the samples we uniformly processed (IPSDSNs, iPSCs, DRG). The similarity to iPSCs may reflect lack of maturity in IPSDSNs, which is a well-recognized problem with iPSC-derived cells (Soldner et al. 2016; Pashos et al. 2017; Warren, Sullivan, et al. 2017; Sala, Bellin, and Mummery 2016). We also note that because the same iPSCs were differentiated to IPSDSNs, both donor genetic background and cell culture effects may contribute to the observed similarity. Despite this, key sensory neuronal marker genes were highly expressed in IPSDSNs, while pluripotency genes were not (Figure 1c). Using Ca^{2+} flux measurements on a subset of differentiated cultures (n=31) we confirmed that the cells consistently responded to veratridine (a sodium ion channel agonist) and tetrodotoxin (a selective sodium ion channel antagonist), as expected (Supplementary Figure 2). Patch-clamp electrophysiology on 616 individual neurons from 31 donors (Supplementary Figures 3,4) showed that the distribution of rheobases was comparable to those obtained from primary DRG cells, but showed significant variation between donors (Supplementary Figure 5).

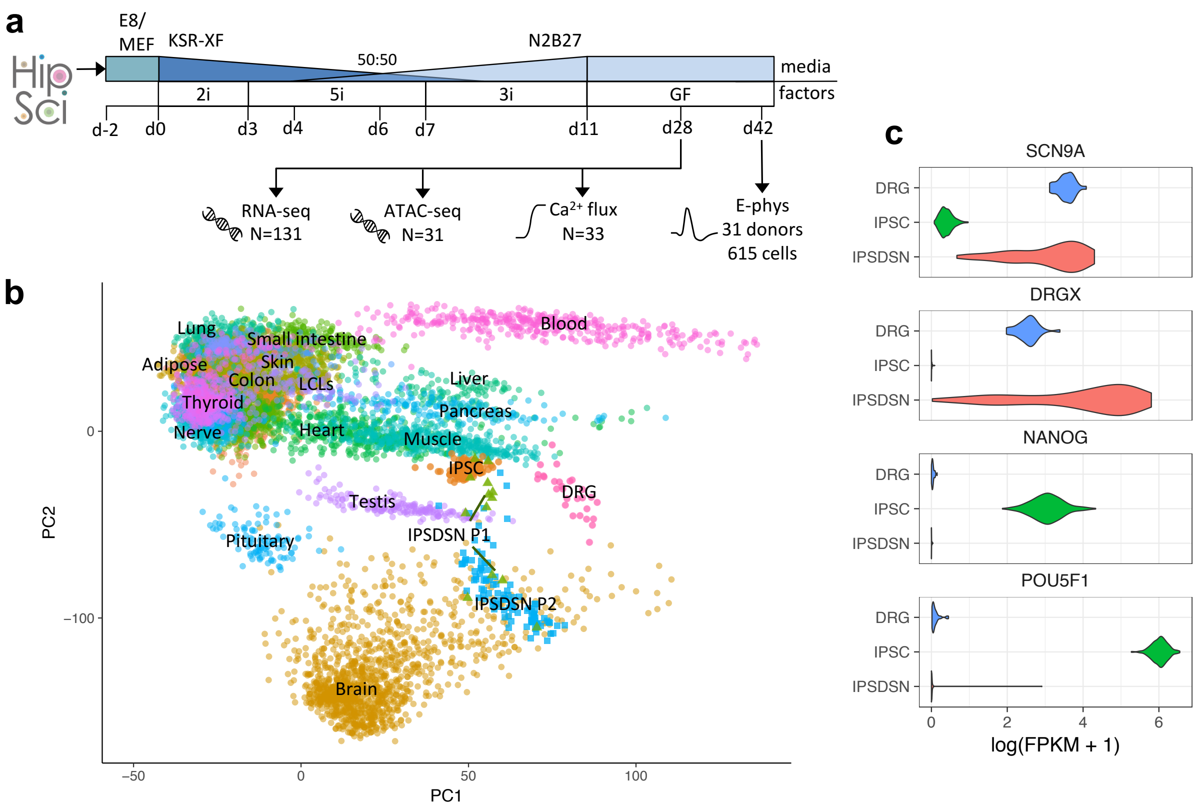

Figure 1 Characterization of molecular phenotypes in iPSC-derived sensory neurons. (**a**) Schematic of IPSDSN differentiation and assays. iPSCs were received in Essential 8 (E8) medium (N=82) or on mouse embryonic fibroblasts (MEFs, N=49), and transferred to KSR-XF medium. Over 11 days, different inhibitor combinations were added (2i, 5i, 3i, see Methods), and N2B27 medium phased in, followed by transfer to growth factor medium at day 11 for neuronal maturation. (**b**) PCA plot projecting IPSDSN, iPSC, and DRG samples onto the first two principal components defined based on RNA-seq FPKMs in GTEx tissues. Some GTEx tissues are unlabelled due to overlapping labels. (**c**) Expression of sensory neuronal marker genes (SCN9A, DRGX) and key iPSC genes (NANOG, POU5F1).

Quantifying differentiation variability using single-cell RNA-seq

In previous work we showed that not all individual cells express neuronal marker genes after differentiation (Young et al. 2014). Samples also appeared to differ visually in the fraction of cells with a neuronal morphology. To further characterize this heterogeneity, we sequenced 177 IPSDSN cells from one individual and clustered them based on expression profiles using SC3 (Kiselev et al. 2016). The data were best explained by two clusters (Figure 2a and Supplementary Figure 6), with 63% of cells forming a tight cluster expressing sensory-neuronal genes (e.g. *SCN9A*, *CHRNB2*), and the remaining 37% of cells forming a looser cluster expressing genes typical of a fibroblastic cell type (e.g. MSN, VIM). The two cell types also separated cleanly in a principal components plot (Supplementary Figure 7), indicating that the cells do not fall on a smooth gradient from more neuronal to less, but rather have differentiated to distinct cell states. Comparing gene expression from each cluster to other tissues showed that the neuronal cluster was most similar to DRG (Spearman's ρ=0.654), followed by iPSCs (ρ=0.609) and GTEx brain (mean ρ=0.599) (Supplementary Figure 8) while the fibroblast-like cluster was most similar to GTEx transformed fibroblasts (ρ=0.683), DRG (ρ=0.662), and iPSCs (ρ=0.653). The similarity of these cells to GTEx fibroblasts could suggest a general similarity of adherent cultured cells, although the neuronal cluster had lower similarity to GTEx fibroblasts (ρ=0.579) than many other tissues.

Next, we used CIBERSORT (Newman et al. 2015) to estimate the fraction of RNA from neuronal cells in our bulk RNA-seq samples, using the single cell gene expression counts with their cluster labels from SC3 as signatures of neuronal or fibroblast-like expression. The estimated neuronal content was strongly correlated (R^2 = 0.75) with the first principal component of gene expression, and this corresponded well with a visual assessment of neuronal content from microscopy images (Figure 2b, Supplementary Figures 9,10). Although a majority of samples appeared by microscopy to have high neuronal content, CIBERSORT estimated relatively high fibroblast-like content for many samples (mean 49%). A factor contributing to this may be the greater RNA content (2.3-fold greater; Supplementary Figure 11) of fibroblast-like cells: indeed when the single cell counts are pooled, CIBERSORT estimates the fibroblast content of this “sample” as 60%, considerably higher than the 37% of single cells in the fibroblast-like cluster. A second consideration is that our scRNA-seq sample was matured for 8 weeks, whereas our bulk RNA-seq samples were matured for 4 weeks. Although gene expression changes are minor after 4 weeks maturation (Young et al. 2014), this difference in maturity means that our single cell reference profiles do not perfectly represent cells in our bulk samples. Despite this, IPSDSN samples estimated to have high fibroblast content still showed greater similarity in genome-wide gene expression with DRG than with any GTEx tissue, including fibroblast cell lines (Supplementary Figure 12). Although these similarities are reassuring, we note that technical factors could contribute to the greater similarity with DRG, as different gene expression quantification tools were used for GTEx (RNASeQC) and for our iPSC, DRG, and IPSDSN samples (featureCounts).

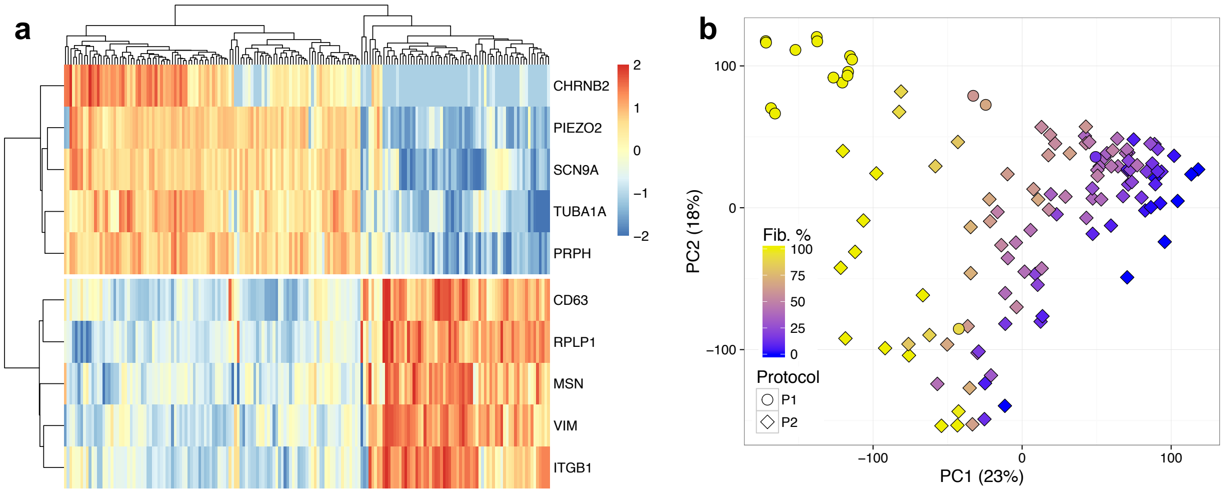

Figure 2 Single-cell sequencing of IPSDSN cells. (**a**) A heatmap of RNA-seq data for ten marker genes of the two cell clusters identified by SC3. Color scale denotes normalised gene expression levels. (**b**) The first two principal components (PCs) of IPSDSN gene expression, with estimated fibroblast-like percentage from CIBERSORT, from samples derived using protocols 1 and 2 (P1 and P2).

Heterogeneity in IPSDSN gene expression

A central issue for genetic studies in iPSC-derived cells is heterogeneity of cellular phenotypes. This heterogeneity could arise from donor genetic background, effects of clonal selection and effects of the cell culture environment during reprogramming and differentiation. Genome-wide gene expression was highly correlated within lines differentiated multiple times (median Spearman ρ=0.96) and reduced slightly between IPSDSNs from different donors (median ρ=0.93) (Supplementary Figure 13). However, differentiation replicates within donor cell lines did not consistently cluster together (Supplementary Figure 14), suggesting that variability due to differentiation was at least as large as that due to donor genetic background and iPSC reprogramming together. Although marker genes specific to sensory neurons and nociceptors were expressed (FPKM > 1) in nearly all samples, we observed a high degree of heterogeneity in the level of expression of some genes compared with DRG (Figure 1c and Supplementary Figure 15), despite the fact that a cell culture system is theoretically more pure in cell type composition than a complex tissue. These observations were independent of sample size, and were robust when comparing with DRG samples from unique donors only, rather than all 28 DRG samples (Supplementary Figure 16).

Next, we examined how between-sample variability in global gene expression of IPSDSNs compared with other somatic tissues and cell lines. The distribution of coefficient of variation (CV) of gene expression in IPSDSNs fell within the range of most GTEx tissues (Figure 3a). However, the median CV of gene expression in IPSDSNs (0.37) was considerably higher than in DRG (0.23), indicating that IPSDSNs have greater between-sample variability in expression than the primary tissue they are intended to model. Highly variable genes in IPSDSNs were enriched for function in neuronal differentiation and development (Supplementary Table 4). Genes that were significantly upregulated between iPSCs and IPSDSNs, which will include those essential for sensory neuronal function, were also more variable than remaining genes (Supplementary Figure 17). Importantly, we did not observe similar levels of expression variability of neuronal or developmental gene groups in DRG, iPSCs, or GTEx nervous tissues (Supplementary Figure 18). These results highlight that expression of neuronal genes varies substantially more in IPSDSNs than in somatic nervous tissue, probably as a result of variability in differentiation. Consistent with this, variance components analysis (Figure 3b, Supplementary Figure 19) showed that as much or more variation was explained by differentiation batch (median 24.7%) as donor/iPSC line of origin (median 23.3%), which would include both donor and reprogramming effects.

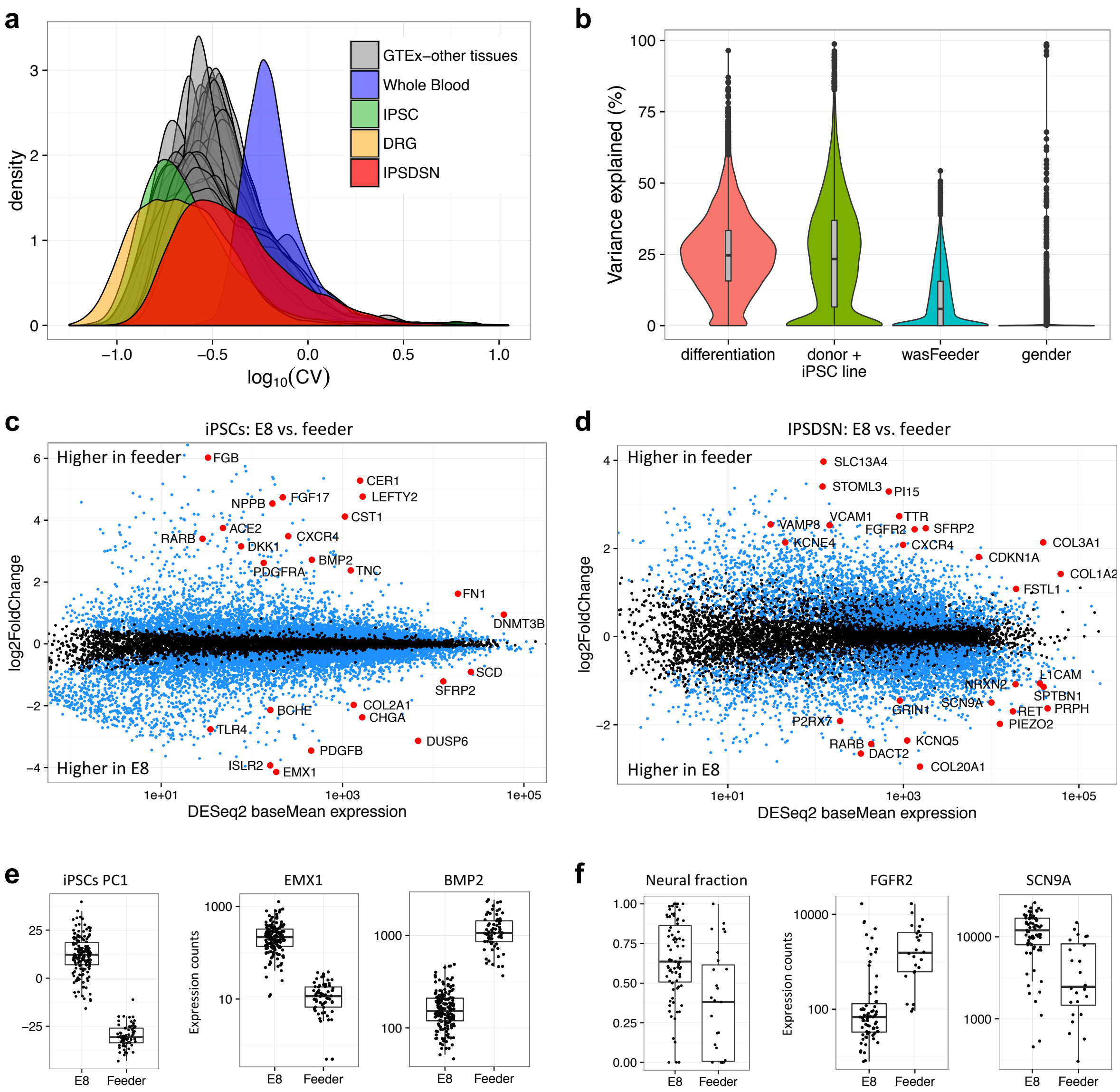

Figure 3 Gene expression variability in IPSDSNs is influenced by differentiation conditions. (**a**) Density plot of the coefficient of variation of genes across samples, separately for each GTEx tissue, IPSDSN samples (n=106, P2 protocol only), iPSC (n=200), and DRG (n=28). (**b**) Violin plot showing, for each gene, the estimated fraction of total expression variability across samples due to differentiation batch, donor genetics or iPSC reprogramming, culture conditions (“wasFeeder”: feeder-dependent vs. E8 medium), and gender. (**c**) Differentially expressed genes (FDR 1%, blue and red points) between iPSC samples grown on feeders (n=68) vs. E8 medium (n=171). (**d**) Differentially expressed genes (FDR 1%) between IPSDSNs from feeder- (n=27) and E8-iPSCs (n=79). Neuronal differentiation genes, such as *RET* and *L1CAM*, are more highly expressed in samples from E8-iPSCs. (**e**) Left barplot: global gene expression differences between feeder- and E8-iPSCs are captured in PC1. Right two barplots: selected differentially expressed genes. (**f**) Left barplot: estimated neural fraction of samples differs in IPSDSNs derived from feeder- and E8-iPSCs. Right two barplots: selected differentially expressed genes.

iPSC culture conditions influence cell fate

Intriguingly our variance components analysis suggested that, although the cell lines for this analysis were differentiated using an identical protocol, starting iPSC cell culture conditions influenced gene expression patterns in the IPSDSNs produced four weeks later (Figure 3). Of the 106 successful P2 protocol differentiations, 27 were from iPSCs maintained on mouse embryonic fibroblast (MEF) feeder cells (feeder-iPSCs), while the remaining 79 were grown in Essential 8 medium (E8-iPSCs). The first principal component (PC) of iPSC gene expression clearly differentiated feeder- and E8-iPSCs (Figure 3e), indicating that culture conditions are among the largest global effects on transcription. Similarly, PC1 of gene expression in IPSDSNs distinguished samples originating from feeder- and E8-iPSCs; moreover, IPSDSNs from E8-iPSCs had higher neuronal content (Figure 3f, 28% higher for E8-iPSCs, t-test $p=1.84\times10^{-5}$). A possible technical explanation for these results is that protocol implementation and batch effects changed subtly over the course of the project. However, the difference in neuronal content between IPSDSNs derived from E8 or feeder-iPSCs remained when sample derivation date was included as an explanatory covariate (linear regression $p=6.5\times10^{-4}$, 36% higher for E8-iPSCs, Supplementary Figure 20).

Next, we determined genes that were differentially expressed between E8- and feeder-iPSCs and IPSDSNs (Figure 3c,d). Genes more highly expressed in feeder-iPSCs were strongly enriched for mesenchyme development, stem cell differentiation, and Wnt and TGF-β signalling, while genes more highly expressed in E8-iPSCs showed less clear enrichment (Supplementary Tables 5-7). Notably, inhibition of TGF-β/SMAD signalling is a key step in sensory neuronal differentiation. Top differentially expressed genes include early developmental regulators such as *EMX1* (15-fold higher in E8-iPSCs), important for specific neuronal cell fates, and *BMP2* (13-fold higher in feeders), which has been shown to suppress differentiation to sensory cell fates by antagonizing Wnt/beta-catenin (Kléber et al. 2005) (Figure 3e). In addition, *SCN9A* and *TAC1*, key markers of sensory neurons, were expressed at low levels in iPSCs, with 2.2-fold and 2.9-fold higher expression in E8-iPSCs. We also considered genes differentially expressed between IPSDSNs derived from E8- and feeder-iPSCs (Figure 3d). Genes more highly expressed in IPSDSN samples from feeder-iPSCs were overrepresented in extracellular matrix components, pattern specification, organ morphogenesis, and Wnt signalling (Supplementary Tables 8-10), and include *FGFR2*, *BMP7*, and *WNT5A* (Figure 3f). Genes more highly expressed in IPSDSN samples from E8-iPSCs were overrepresented in ion channel complexes, peripheral nervous system development, and synapse organisation, and include *SCN9A*, *DRGX*, and *CACNA1A*. These differences likely reflect the increased neuronal content of samples from E8-iPSCs. Together these results suggest that iPSCs are primed towards different cell fates depending on the iPSC culture medium.

Since iPSC culture conditions influenced differentiation outcomes, we examined gene expression variability within subsets of IPSDSN samples. IPSDSNs differentiated from feeder-iPSCs had somewhat higher global gene expression variability, yet those from E8-iPSCs were still highly variable relative to DRG and iPSCs (Supplementary Figure 21), with neuronal and developmental gene sets enriched for highly variable genes (Supplementary Table 11). Among the 79 IPSDSNs from E8-iPSCs, samples with high fibroblast content had somewhat higher variability, but those with low fibroblast content still showed high variability relative to DRG and iPSCs.

Genetic variants influence gene expression, splicing and chromatin accessibility in sensory neurons

Using a linear model (FastQTL (Ongen et al. 2016)), we mapped 1,403 expression quantitative trait loci (eQTLs) at FDR 10%, of which 746 were expressed at a moderate level (FPKM > 1). We noted that we discovered many fewer eQTLs than in GTEx tissues of comparable sample size (Supplementary Figure 23). This suggested that power for eQTL discovery was lower in IPSDSNs than somatic tissues, possibly due to additional variability introduced by differentiation. Using an allele-specific method (Kumasaka, Knights, and Gaffney 2015) we detected 3,778 genes with expression-modifying genetic variants, termed eGenes, at FDR 10% (Supplementary Table 12), with 2,607 of these expressed at FPKM > 1. Notably, it was only using the additional information from allele specific signals that we achieved approximately similar statistical power to GTEx tissues with equivalent sample sizes, and the improvement in power was greatest among genes with high variability across samples (Supplementary Figures 22,23).

We next compared our eQTLs with GTEx. When clustering tissues based on the pairwise correlation in eQTL effect sizes, IPSDSNs clustered most closely with GTEx brain tissues, while also showing elevated correlation with GTEx fibroblasts (Supplementary Figure 24). We could not call eQTLs in DRG as the samples were not consented for use of genetic data. To identify eQTLs that were not already reported in GTEx (v6), we used a protocol described previously for the HIPSCI project (Kilpinen et al. 2017). Of all 3,778 eGenes, 954 had tissue-specific associations (Supplementary Table 15), including genes with known involvement in pain or neuropathies, such as *SCN9A*, *GRIN3A*, *P2RX7*, *CACNA1H*/Cav3.2, and *NTRK2*. Because these eQTLs were not seen in any GTEx tissue, this suggests that these are regulatory variants with IPSDSN-specific function.

Variants affecting gene splicing (sQTLs) often change either protein structure or context-dependent gene regulation, and may be more enriched for complex trait loci than are eQTLs (Li et al. 2016). To detect sQTLs we used the annotation-free method LeafCutter (Li, Knowles, and Pritchard 2016) to define 30,591 clusters of alternatively spliced introns. Using FastQTL (Ongen et al. 2016) we discovered QTLs for 2,079 alternative splicing clusters at FDR 10% (Supplementary Table 13). Notably, only 538 (26%) of the lead variants for these splicing associations were in linkage disequilibrium (LD) r^2 >= 0.5 with a lead eQTL variant in our dataset, indicating that the sQTLs extend our catalog of expression-altering variants and are not merely proxies for gene-level eQTLs (or vice versa).

	Number	GWAS overlap
eQTLs	3778	156
sQTLs	2079	129
ATAC QTLs	6318	172
Joint ATAC/eQTLs	177	14

Table 1 QTL associations. Columns show the number of associations and the number of unique overlaps ($r^2 > 0.8$) between lead QTL SNPs and GWAS catalog SNPs after removing duplicates for each GWAS trait.

We collected ATAC-seq data for 31 samples (Buenrostro et al. 2013) and used this to identify active regulatory regions in IPSDSNs and to map 6,318 caQTLs chromatin accessibility QTLs (caQTLs) at FDR 10% (Supplementary Table 14). To identify transcription factors in IPSDSNs whose binding is altered by regulatory variants, we used the LOLA Bioconductor package (Sheffield and Bock 2015) to test for enrichment of our lead QTL SNPs, relative to GTEx lead SNPs, in ENCODE ChIP-seq peaks and JASPAR transcription factor motifs (Supplementary Tables 16,17). Tissue-specific eQTLs were highly enriched within SMARCB1 and SMARCC2 peaks (odds ratios 5.8 and 14.1; $p < 5x10^{-5}$), which are both members of the neuron-specific chromatin remodeling (nBAF) complex (Lessard et al. 2007). Considering all IPSDSN eQTLs, we found enrichments for ELK1 and ELK4, as well as c-Fos, a target of ELK1 and ELK4 which is widely expressed but is known to have specific functions in sensory neurons (Hunt, Pini, and Evan 1987; Kohno et al. 2003). Notably, DNA sequence motifs for REST, ELK1 and ELK4 are also among the most highly enriched motifs in our ATAC-seq peaks (Supplementary Table 18).

Sensory neuron eQTLs and sQTLs overlap with complex trait loci

While we were interested in comparing our set of QTLs with GWAS for pain, the largest GWAS for pain to date included just 1,308 samples and found no associations at genome-wide significance (Peters et al. 2013). We therefore considered all GWAS catalog associations with $p < 5x10^{-8}$ that were in high LD ($r^2 > 0.8$) with a QTL in our dataset, with two purposes in mind: to determine whether any GWAS traits are enriched overall for overlap with sensory neuron QTLs, and to find individual cases where a QTL is a strong candidate as a causal association for the GWAS trait. Overall, IPSDSN eQTLs were significantly enriched for overlap with GWAS catalog SNPs ($p < 0.001$) relative to 1000 random sets of SNPs matched for minor allele frequency (MAF), distance to nearest gene, gene density, and LD (Pers, Timshel, and Hirschhorn 2014), and the overlap was consistent with that seen for eQTL studies in other tissues (Supplementary Figure 25). Although nociceptive neurons are specialized for sensing and relaying pain signals, they share characteristics with other neurons; thus, we might expect enrichment for traits known to involve the nervous system more generally. However, among the 41 traits with at least 40 GWAS catalog associations, we could not detect any trait with significantly greater overlap with our QTL catalog than other traits after correcting for multiple testing (Supplementary Table 19).

Across all traits, we found 156 genes with an eQTL overlapping at least one GWAS association, and similarly 129 sQTLs and 172 caQTLs with GWAS overlap (full catalog in Supplementary Tables 20-22). We examined individual associations, in conjunction with ATAC-seq peaks and LD information, to identify candidate causal variants influencing both a molecular phenotype and a complex trait. For most of these associations we do not expect that sensory neurons are the most relevant cell type; rather the overlaps may reflect either general neuronal mechanisms or non-cell-type-specific functions. We thus focused on traits where neurons are likely to be a relevant cell type.

Among overlapping associations we found a number that relate to neuronal diseases, such as Parkinson's disease, multiple sclerosis, and Alzheimer's disease. One striking overlap is between an eQTL for *SNCA*, encoding alpha synuclein, and Parkinson's disease, for which a likely causal variant has recently been identified (Soldner et al. 2016). The lead GWAS SNP and our lead eQTL are both in perfect LD with rs356168 (1000 genomes MAF 0.39), which lies in an ATAC-seq peak in an intron of *SNCA*. Soldner et al. used CRISPR/Cas9 genome editing in iPSC-derived neurons to show that rs356168 alters both *SNCA* expression and binding of brain-specific transcription factors (Soldner et al. 2016). In IPSDSN cells we find that the G allele of rs356168 increases *SNCA* expression 1.14-fold, in line with Soldner et al. who reported 1.06- to 1.18-fold increases in neurons and neural precursors. However, despite residing in a visible ATAC-seq peak in our data, rs356168 is not detected as a caQTL (SNP p value = 0.22). eQTLs for SNCA have recently been reported in the latest GTEx release (v6p), but none of the tissue lead SNPs are in LD ($r^2 > 0.2$) with rs356168, suggesting that the effect of this SNP can be more readily detected in specific cell and tissue types, including IPSDSNs and the frontal cortex tissue and iPSC derived neurons studied by Soldner et al.

We also find multiple compelling overlaps between splice QTLs and GWAS associations (Figure 4). One known example is a strong sQTL for *TNFRSF1A* ($p=9.9\times10^{-29}$) with the same lead SNP (rs1800693, MAF 0.30) as a multiple sclerosis association. This likely causal SNP is located 10 base pairs from the donor splice site downstream of exon 6, and has been experimentally shown to cause skipping of exon 6, which results in a truncated, soluble form of TNFR1 that appears to reduce TNF (Gregory et al. 2012). *TNFRSF1A* is highly expressed (>15 FPKM) in both IPSDSNs and in DRG. We do not see an effect of this variant on total expression levels in our cells (p > 0.5), but we observe skipping of exon 6 in about 12% of transcripts from individuals homozygous for rs1800693 (Figure 4a). Since these transcripts undergo nonsense-mediated decay, the actual rate of exon skipping is likely to be higher. Given the broad role of TNF in inflammation and immunity, it is interesting that rs1800693 is associated with MS but not with other autoimmune disorders, apart from primary biliary cirrhosis (Gregory et al. 2012). Moreover, whereas TNF inhibitors are effective in many autoimmune disorders, they exacerbate MS, an effect that is mimicked by the reduction in TNF signalling produced by the TNFRSF1A splice variant. These observations suggest an interplay between cells of the CNS and immune system involving TNF signalling. TNF signalling has been shown to have both inflammatory and neuroprotective effects in the CNS and, despite a large body of research, the exact mechanisms and cell types responsible for the genetic risk associated with TNF receptor polymorphisms remain unclear (Probert 2015).

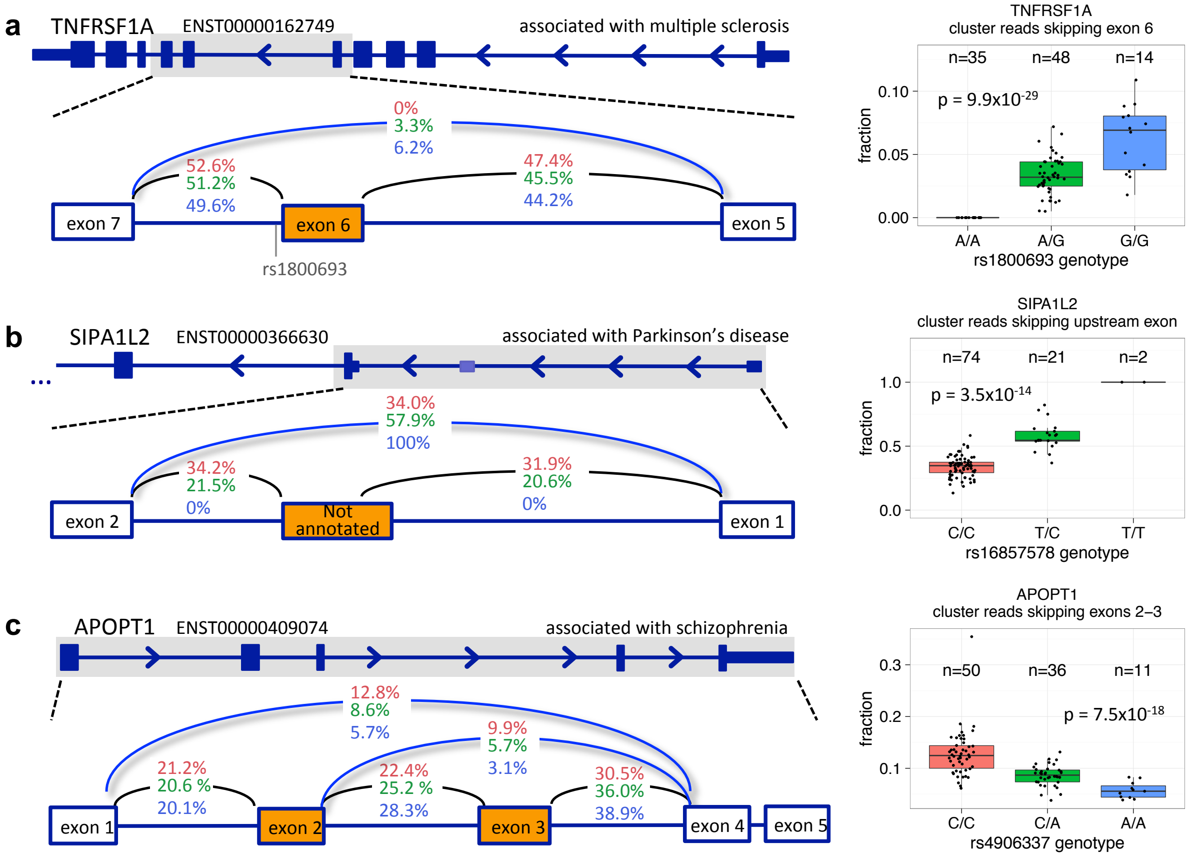

Figure 4 Splicing QTLs overlapping GWAS. **(a)** An sQTL for *TNFRSF1A* leads to skipping of exon 6, and overlaps with a multiple sclerosis association. **(b)** An sQTL for *SIPA1L2* leads to increased skipping of an unannotated exon between alternative promoters, and overlaps with a Parkinson's disease association. **(c)** An sQTL for *APOPT1* alters skipping of exons 2 and 3, and overlaps with a schizophrenia association. P values are from the beta approximation based on 10,000 permutations as reported by FastQTL.

An sQTL for *SIPA1L2* (rs16857578, MAF 0.23) is in LD with associations for both Parkinson's disease (rs10797576, r^2=0.93) and blood pressure (rs11589828, r^2=0.94). An unannotated noncoding exon (chr1:232533490-232533583) between alternative *SIPA1L2* promoters is included in nearly 50% of transcripts in individuals with the reference genotype, but splicing in of the exon is abolished by the variant (Figure 4b). SIPA1L2, also known as SPAR2, is a Rap GTPase-activating protein expressed in the brain and enriched at synaptic spines (Spilker and Kreutz 2010). Although its function is not yet clear, expression is seen in many tissues profiled by GTEx, with highest expression in the peripheral tibial nerve. Interestingly, the related protein SIPA1L1 exhibits an alternative protein isoform with an N-terminal extension that is regulated post-translationally to influence neurite outgrowth (Jordan et al. 2005).

A complex sQTL for *APOPT1* (rs4906337, MAF 0.22) is in near-perfect LD with a schizophrenia association (rs12887734). The splicing events involve skipping either of exon 3 only or both exons 2 and 3 (Figure 4c). At least 20 variants are in high LD (r^2 > 0.9), including rs4906337 which is 40 bp from the exon 3 acceptor splice site, and rs2403197 which is 63 bp from the exon 4 donor splice site. No sQTL is reported in GTEx, and although eQTLs are reported for *APOPT1*, only the thyroid-specific eQTL (rs35496194) is in LD (r^2 = 0.94) with the schizophrenia-associated SNP rs12887734. APOPT1 is localized to mitochondria and is broadly expressed. Homozygous loss-of-function mutations in this gene lead to Cytochrome c oxidase deficiency and a distinctive brain MRI pattern showing cavitating leukodystrophy in the posterior region of the cerebral hemispheres, with affected individuals having variable motor and cognitive impairments and peripheral neuropathy (Melchionda et al. 2014).

Recall by genotype studies in iPSC-derived cells will require large sample sizes

One attractive future use of iPSCs is to experimentally characterise GWAS loci using a “recall by genotype” approach. Here, iPSC lines with specific genotypes are chosen from a large bank and differentiated into target cell types (for example, see (Warren, Sullivan, et al. 2017)). Our observations suggested that, for certain protocols, the additional cellular heterogeneity introduced by differentiation could impact the power of these studies to detect the effects of common genetic variants. Importantly, our large set of differentiations gave us accurate genome-wide estimates of effect size and expression variability in an IPS-derived cell type, for use as a benchmark “ground truth”. We investigated the performance of iPSC-based recall by genotype studies by bootstrap resampling from a stringent (FDR 1%) IPSDSN eQTL call set. For each eQTL gene we sampled expression counts from an equal number of major and minor homozygotes for the lead SNP, sampling with replacement to achieve a specific sample size. We then estimated power as the fraction of 100 bootstrap replicates where we found a significant difference ($p < 0.05$, Wilcoxon rank sum test) in expression between the homozygotes.

Our results illustrate important trends. First, recall by genotype studies in IPS-derived cells are likely to require relatively large sample sizes, typically 20-80 unrelated individuals, for variants with a 1.5-2-fold effect size (Figure 5a). Second, as expected, highly variable genes are more challenging (Figure 5b) with power below 40% in a sample size of 20 for even moderately variable genes (CV 0.5 - 0.75). While expression noise will not typically be known accurately a priori, an estimate of effect size may be available from previous eQTL studies in specific tissues. This could enable estimating the number of samples needed to achieve a desired power (Figure 5a).

Note that these power estimates assume that a single gene is being tested, which is only likely to be the case when there is a very strong prior belief in the causal gene and few genes in the region. Where multiple genes are tested, power will be lower. These results also suggest that large sample sizes will be required when using genome editing to identify causal GWAS-associated variants: although genetic background can be controlled in such an experiment, differentiation noise will continue to be a major contributor to gene expression variability.

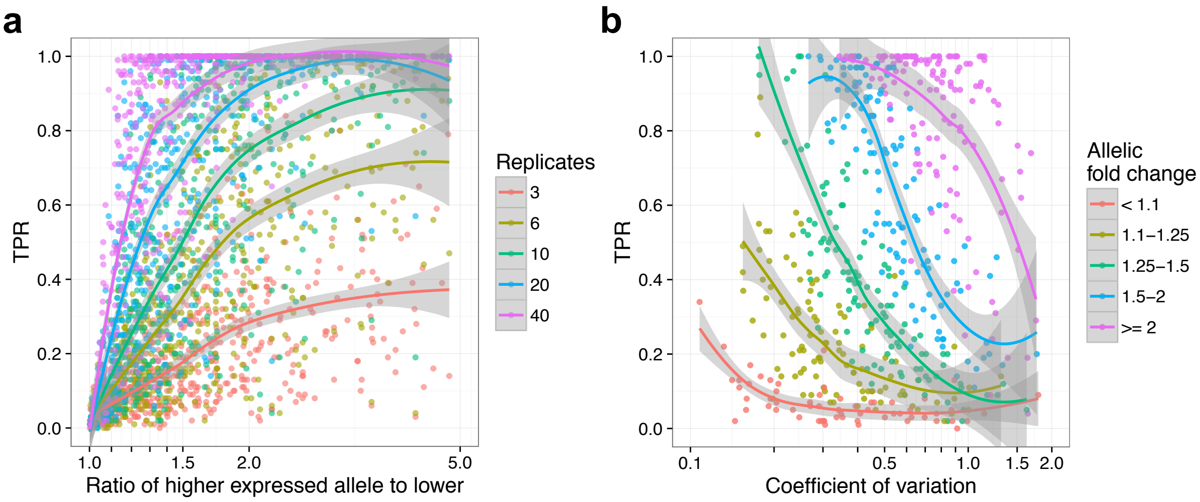

Figure 5 Power to detect a genetic effect in a single-variant single-gene test depends on sample size, allelic effect size, and gene expression variability. (**a**) TPR as a function of allelic fold change for five different numbers of replicates (half the total sample size). (**b**) TPR as a function of CV for five bins of allelic fold change, with 10 samples of each genotype.

Discussion

iPSC-derived cells enable the molecular mechanisms of disease to be studied in relevant human cell types, including those which are inaccessible as primary tissue samples. Because the effect sizes of common disease-associated risk alleles tend to be small, observing their effects in cellular models is challenging (Soldner et al. 2016; Pashos et al. 2017). In an iPSC-based system, this difficulty is compounded by variability between samples in the success of differentiation, as described for hepatocytes (Dianat et al. 2013), hematopoietic progenitors (Smith et al. 2013), and neurons (Handel et al. 2016; Hu et al. 2010).

Our study is the first that we are aware of to perform iPSC differentiation to a neuronal cell type and functionally characterise the resulting cells at scale. Sample-to-sample variability in gene expression in the iPSC-derived cells was greater than in DRGs, with highly variable genes enriched in processes relating to neuronal differentiation and development. This highlights that genes likely to be of particular interest and relevance for the function of these cells are also among the most variable, a challenge which may be broadly true of iPSC-derived cells. Despite the observed sample-to-sample variability in gene expression, we detected thousands of eQTLs, sQTLs, and caQTLs in IPSDSNs, most of which were discovered only with a model that statistically combines both allele-specific and between individual differences in expression to improve power for association mapping. Some of these overlap known expression-modifying variants that are associated with disease, such as an eQTL for SNCA associated with Parkinson’s disease. However, for most of these disease overlaps the causal variants are not known. This QTL map is thus a starting point for in-depth dissection of individual loci in iPSC-derived neurons where we have shown that a genetic effect is present.

Although our study highlights the potential power of IPSC derived cells as model systems for studying human genetic variation, our results also illustrate the limitations of this approach. First, despite expressing key marker genes and exhibiting neuronal morphology and electrophysiology, it is clear from our data that IPSDSNs are transcriptionally distinct from their primary counterparts, DRGs. This reflects a limitation of existing in vitro differentiation protocols, which produce cells that are not as functionally or transcriptionally mature as primary tissues. Second, our differentiations did not produce pure populations of neurons, nor could we measure the purity of the resulting cultures precisely. A portion of the sample-to-sample variability that we observed is likely due to this mixture of cell types, which varied across differentiations. Although mature neurons can be labeled for marker genes, they are not easily sorted by automated systems, which limits the high-throughput options available for purifying neuronal populations. As a result, the eQTLs that we discovered do not represent those of a pure sensory neuronal cell type. For many cell types, sorting is more feasible, and could provide one solution to the variable maturity and heterogeneity of differentiated cell populations.

We used single-cell RNA-seq from three differentiation batches to characterise IPSDSN heterogeneity, which showed that they cluster into neuronal cells and cells with more fibroblast-like gene expression. Using reference profiles from these clusters enabled us to estimate a proxy measure of neuronal cell purity in our bulk RNA-seq samples, and these estimates qualitatively agreed with the neuronal content in images from the cell cultures. Our method is similar to a deconvolution approach described recently using bulk and single-cell sequencing of primary human and mouse pancreas (Baron et al. 2016).

The similarity of the fibroblast-like single cells to DRG raises the important question of whether these cells are immature sensory neurons. Single-cell sequencing at multiple time points during MYOD-mediated myogenic reprogramming has suggested that some individual cells traverse a desired course, while others terminate at incomplete or aberrant reprogramming outcomes (Cacchiarelli et al. 2017). Such an approach in IPSDSNs could reveal determinants of neuronal differentiation trajectories, and may yield useful insights for protocol changes to improve the purity of differentiated neurons, or to specify more precise neuronal subtypes. More generally, replacing bulk RNA-seq with single cell sequencing across many samples could enable *in silico* sorting of cells based on their transcriptome, and better characterisation of the sources of variation within a differentiated population of cells. Further, culturing cells from multiple donors in a pool, along with an scRNA-seq readout, could reduce differentiation-related batch effects while retaining the ability to identify donor-specific genetic effects on gene expression. These advantages suggest to us that a move towards scRNA-seq will be extremely useful in iPSC-derived cell models.

For iPSC models of common disease associated variants to be used effectively, it is critical to know which candidate disease associated variants exhibit a detectable cellular phenotype in an in vitro model. We used *in silico* resampling to estimate the sample sizes needed to detect the effects of noncoding regulatory variants in iPSC-derived cells using a recall by genotype design. Power above 80% is only achieved with surprisingly large (40+) samples, even for alleles with a fold change of 1.5 to 2. Further, the power we report may be overestimated, due to ascertainment bias in defining a set of eQTLs as “true positives”, which fails to include true genetic effects that we did not discover in our samples. Even larger samples will be needed when multiple genes, for example in a single GWAS interval, are to be tested. These observations are consistent with a recent genome-editing experiment that required 136 differentiations in hepatocyte-like cells to discover an effect of rs12740374 on *SORT1* gene expression (Warren, Sullivan, et al. 2017). Notably, the modest effect of this variant on expression in hepatocyte-like cells (1.3-fold increase) stands in contrast to the large effect of the variant (4- to 12-fold increase) observed previously in primary liver (Musunuru et al. 2010). Where it is possible to use a coding SNP to assess the allele-specific effect of a genome edit, as done for *SNCA* (Soldner et al. 2016), this may prove a more efficient approach to detecting causal effects of individual regulatory variants.

In summary, we have measured multiple molecular phenotypes in a large panel of iPSC-derived neurons. The catalog of QTLs we provide reveals a large set of common variants and target genes with detectable effects in IPSDSNs. These associations provide promising targets for functional studies to fine-map causal disease-associated alleles, such as by allelic replacement using CRISPR-Cas9, and our study describes the importance of considering differentiation-induced variability when planning these studies in iPSC-derived cells.

URLs

OpenTargets, www.targetvalidation.org.
CIBERSORT, cibersort.stanford.edu.
ENCODE, www.encodeproject.org.
GTEx, www.gtexportal.org.
HIPSCI, www.hipsci.org.

Data Availability

Code used for processing and analysing data is available at https://github.org/js29/ipsdsn. RNA-seq and ATAC-seq data for open access samples are deposited in the European Nucleotide Archive under accession ERP020576. These data for managed access samples are deposited in the European Genome Archive under accession EGAD00001003145. Summary statistics and gene expression counts are available at https://www.ebi.ac.uk/biostudies/studies/S-BSST16. Sample genotypes and accession numbers are available at http://www.hipsci.org/data.

Acknowledgments

The iPSC lines were generated under the Human Induced Pluripotent Stem Cell Initiative (HIPSCI) funded by a grant from the Wellcome Trust and Medical Research Council, supported by the Wellcome Trust (WT098051) and the NIHR/Wellcome Trust Clinical Research Facility. HIPSCI funding was used for sensory neuron RNA-sequencing. We acknowledge Life Science Technologies Corporation as the provider of Cytotune. Pfizer Neuroscience (Pfizer Ltd.) funded neuronal differentiation, functional assays, single-cell RNA-sequencing, and collection and sequencing of dorsal root ganglion samples. The authors gratefully acknowledge Natsuhiko Kumasaka for help with RASQUAL. We thank Florian Merkle for comments on the manuscript. JS gratefully acknowledges support from the Wellcome Trust for his PhD studentship.

Author contributions

JS analyzed data, and JS and DJG wrote the manuscript. SF performed all differentiations. AGu analyzed data; AGu, DJG, and PJW conceived and supervised the project. HK compared eQTLs with GTEx and identified tissue-specific eQTLs. JR and MP cultured iPSC samples. AJK performed all ATAC-seq. KA and AGon assisted with data analysis. AW performed single cell RNA work and assisted with data analysis. RF and CLB performed RNA extraction and quantification. EI performed cell culture and Ca^{2+} flux assays. MB assisted with experimental design and Ca^{2+} flux assays. LC, SL, and AJL performed electrophysiology measurements. All authors reviewed the manuscript.

Conflicts of Interest

SF, RF, CB, AW, MB, EI, LC, SL, AJL, PJW and AGu were all employees of Pfizer at the time the experiments were performed.

Online methods

IPS cell lines

A summary of iPSC lines used is available in Supplementary Table 2, and details of processes and assays for these iPSCs generated by the HIPSCI project are available at www.hipsci.org. Briefly, 107 human induced pluripotent stem cells (iPSCs) from 103 healthy donors were obtained from the HIPSCI resource (Kilpinen et al. 2017). We reproduce an abridged version of their methods here:

> For each donor, primary human fibroblasts were derived from 2 mm skin punch biopsies. Dissected biopsy fragments were cultured in fibroblast growth medium until fibroblast outgrowths appeared, which took 14 days on average. Fibroblasts were then transduced using Sendai vectors expressing hOCT3/4, hSOX2, hKLF4, and hc-MYC (CytoTuneTM, Life Technologies, Cat. no. A1377801). Transduced cells were cultured on an irradiated mouse embryonic fibroblast (MEF-CF1) feeder layer in iPSC medium consisting of Advanced DMEM (Life technologies, UK) supplemented with 10% Knockout Serum Replacement (KOSR, Life technologies, UK), 2 mM L-glutamine (Life technologies, UK), 0.007% 2-mercaptoethanol (Sigma-Aldrich, UK), 4 ng/mL of recombinant Zebrafish Fibroblast Growth Factor-2 (CSCR, University of Cambridge), and 1% Pen/Strep (Life technologies, UK). Cells with an iPSC morphology appeared approximately 25 to 30 days post-transduction. The undifferentiated colonies (6 per donor) were picked between days 30-40, transferred onto 12-well MEF-CF1 feeder plates and cultured in iPSC medium with daily media change until ready to passage.
>
> Between passages 4 to 8, selected feeder-dependent iPSC lines were transferred to feeder-free culture, while other lines continued to be cultured on MEF-CF1 feeder plates. Feeder-free lines were cultured in Essential 8 (E8) medium on tissue culture dishes coated with 10 μg/ml Vitronectin XF (StemCell Technologies, UK, 07180). E8 complete medium consists of basal medium DMEM/F-12(HAM) 1:1(Life technologies, UK, A1517001) supplemented with E8 supplement (50X) (Life technologies, UK, A1517001) and 1% Pen/Strep (Life technologies, UK, 15140122).

Of the 107 lines, 38 were initially grown in feeder-dependent medium and the remainder were grown in feeder-free E8 medium. All HIPSCI samples were collected from consented research volunteers recruited from the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk). Samples were collected initially under existing Cambridge BioResource ethics for iPSC derivation (REC Ref:

09/H0304/77, V2 04/01/2013), with later samples collected under a revised consent (REC Ref: 09/H0304/77, V3 15/03/2013).

Sensory neuron differentiation

All differentiations in this study were performed by a single individual, and a summary of the IPSDSN cell lines is in Supplementary Table 1. Two differentiation protocols were used, named P1 (13 differentiations) and P2 (110 differentiations). Note that P1 protocol samples were used only for QTL calling, and other analyses used P2 protocol samples exclusively. The P1 protocol (described in detail in (Young et al. 2014)) was developed prior to this study using a small number of cell lines. It involved the addition of "2i" inhibitors (LDN193189 and SB-431542) for 5 days, followed by "5i" inhibitors (LDN193189, SB-431542, CHIR99021, DAPT, SU5402) for a further 6 days. When applying this protocol to a larger number of samples we observed an excessive rate of cell death prior to obtaining neural progenitors (days 9-12). A separate study was undertaken to optimise the robustness of the protocol. We altered the protocol to make it more similar to that of Chambers et al. (Chambers et al. 2012), and differentiated 17 replicates using both the new P2 protocol and the P1 protocol (these samples are not used for this manuscript). All 17 replicates successfully differentiated with the P2 protocol, whereas only 7 of 17 (41%) were successful with the P1 protocol.
The P2 protocol differed by:

- using E8 rather than mTeSR1 media when maintaining iPSCs prior to differentiation;
- phasing in neurobasal media beginning at day 4, and gradually increasing this to 100% by day 11, to support neurons during differentiation;
- beginning addition of inhibitors 5i two days earlier (day 3 rather than day 5);
- stopping addition of small molecule inhibitors LDN193189 (1µmol/l) and SB-431542 (10 µmol/l) beginning at day 7 (rather than day 11), referred to as "3i" in the main text for the 3 inhibitors that continued to be added.

We measured cell culture endpoints, including:

- Total cell numbers at multiple points during differentiation
- Population doubling time
- Viability using Trypan blue staining

Functional assays (Ca^{2+} flux, response to Veratridine) confirmed that response of the sensory neurons produced by each protocol was equivalent; however, the P2 protocol performed more consistently across cell lines and culture parameters.

In general, for each differentiation from iPSCs of a given donor multiple flasks were cultured in parallel. The first successful flask was used for RNA-seq. Subsequent flasks were used for electrophysiology measurements, Ca flux or pharmacological measurements. If an additional flask was available then it was used for ATAC-seq.

P2 protocol details

Clump passaged iPSCs were single cell seeded in E8 media (Life Technologies) on growth factor-reduced Matrigel (BD Biosciences, San Jose, CA) 48 hours prior to neural induction (day 0). KSR Media was prepared as 500ml DMEM-KO (Life Technologies 10829-018), 130 ml Knockout Serum Replacement Xeno-Free (Life Technologies 12618-013), 1x NEAA (Life Technologies 11140-068), 1x Glutamax (Life Technologies 35050-087), 0.01 mM β-mercaptoethanol (Sigma M6250-100ml). KSR media containing small molecule inhibitors LDN193189 (100 nM) and SB-431542 (10 µM) was added to cells from day 0 to 3 to drive anterior neuroectoderm specification. From day 3, CHIR99021 (3 µM), DAPT (10 µM) and SU5402 (10 µM) were also added to further enable the emergence of neural crest phenotypes. N2B27 media was progressively phased in every two days from D4. N2B27 Media was prepared as 500 ml Neurobasal medium (Life Technologies 21103-049), 5 ml N2 supplement (Life

Technologies 17502-048), 10 ml B27 supplement without vitamin A (Life Technologies 12587-010), 0.01mM β-mercaptoethanol (Sigma M6250-100 ml) and 1x Glutamax (Life Technologies 35050-087). On day 7, inhibitors LDN193189 and SB-431542 were no longer used, while CHIR99021, DAPT, and SU5402 continued to be added. On day 11 cells were harvested and reseeded at 150,000 cells/cm2 in maturation media containing N2B27 media with human-b-NGF (25 ng/ml), BDNF (25 ng/ml), NT3 (25 ng/ml) and GDNF (25 ng/ml). Mitomycin C treatment (1 µg/ml) was used once at day 14 for 2 hrs to reduce the non-neuronal population. Cells were differentiated in T25 flasks for RNA and nuclei isolation, and onto coverslips and 96 well plates for electrophysiology and Ca2+ flux assays.

P1 protocol details

All media and inhibitors and concentrations used were identical to the P2 protocol described above; the difference was timing of addition. Clump passaged iPSCs were single cell seeded in mTeSR1 iPSC (StemCell Technologies, Vancouver) media on growth factor-reduced Matrigel (BD Biosciences, San Jose, CA) 48 hours prior to neural induction (day 0). KSR media containing LDN193189 and SB-431542 was added to cells from day 0 to 5. From day 5, CHIR99021, DAPT and SU5402 were also added. On day 11 cells were harvested and reseeded at 150,000 cells/cm^2 in maturation media containing N2B27 media with human-b-NGF, BDNF, NT3 and GDNF. Mitomycin C treatment (1 µg/ml) was used once at day 14 for 2 hrs to reduce the non-neuronal population.

Single-cell RNA sequencing

Blood-derived iPSCs from a single individual, who was not a HIPSCI donor, were differentiated to sensory neurons in 3 separate batches using the P2 protocol. These samples were matured for 8 weeks, whereas the RNA-seq samples were matured 4 weeks. Previous work showed only minor changes in gene expression between 4 and 8 weeks maturation (Young et al. 2014). Each batch of dissociated cells was loaded onto a Fluidigm C1 system for automatic cell separation, reverse transcription and amplification. Libraries were only prepared from C1 chambers that contained single cells, using the Illumina Nextera XT kit as per the Fluidigm C1 protocol. These were quantified with the Qubit dsDNA HS assay (Thermo Fisher) and KAPA Library Quantification Kit (KAPA Biosystems) and size-checked with the Agilent Bioanalyser DNA 1000 assay (Agilent), as per manufacturers' recommendations. Libraries were 96-way multiplexed and sequenced paired end on an Illumina Nextseq500 (75bp reads). Reads for each cell were aligned to GRCh38 and Ensembl 80 transcript annotations using STAR v2.4.0d with default parameters.

We had gene expression counts for ~56,000 genes (including noncoding RNAs) for 186 cells, although many of these were zeros. We excluded 9 cells expressing fewer than 20% of the quantified genes, and then used SC3 (Kiselev et al. 2016) to cluster the remaining 177 cells based on expression counts. Note that when clustering cells from complex tissues there is often a hierarchy of clusters, and no specific number of clusters can be considered correct. Allowing that the same could be true of IPS-derived cells, we examined alternative numbers of clusters from k=2 to 5 (Supplementary Figure 6), specifying k (the number of clusters) ranging from 2 to 5. With two clusters, the marker genes reported by SC3 clearly identified one cluster (111 cells) as neuronal, whereas the other cluster (66 cells) had high expression of extracellular matrix genes reminiscent of fibroblasts. With 3 and 4 clusters, the sensory-neuronal cell cluster remained unchanged, and the fibroblast-like cluster became further subdivided. This suggests that a majority of the cells in this sample were terminally differentiated into sensory neurons, whereas the remaining cells were more heterogeneous in their gene expression.

To display marker gene expression we selected 5 neuronal and 5 fibroblast marker genes based on the literature. After DESeq2's variance stabilizing transformation, we used R's "scale" function to mean-center and normalize expression values across cells for these genes, and plotted the result using the pheatmap R package.

To compare gene expression between single cell clusters and bulk RNA-seq samples, we computed the mean FPKM expression for each gene separately in single neurons and fibroblast-like cells. We subsetted to genes with nonzero expression in at least one GTEx tissue and in at least one of our tissues (iPSC, DRG, IPSDSN bulk, IPSDSN single cells), and computed the Spearman correlation between each pair of tissues for the remaining genes.

Genotypes

We obtained imputed genotypes for all of the samples from the HIPSCI project. We used CrossMap (http://crossmap.sourceforge.net/) to convert variant coordinates from GRCh37 reference genome to GRCh38. We then used bcftools (http://samtools.github.io/bcftools/) to retain only bi-allelic variants (SNPs and indels) with INFO score > 0.8 and MAF > 0.05 in the 97 samples used for QTL calling. This filtered VCF file was used for all subsequent analyses.

RNA sequencing

Cells growing in T25 flasks were washed twice with PBS followed by addition of 600 mL of RLTPlus buffer. Cells were gently lifted from the flask and transferred to 1.5 ml tubes. Lysates were transferred to 1.5 mL tubes. RNA and gDNA were isolated using AllPrep DNA/RNA Minikit (Qiagen). RNA was eluted in 33 uL of DNAse free water and DNA eluted in 53 uL EB buffer.

RNA libraries were prepared using the Illumina TruSeq strand-specific protocol, and were sequenced with paired-end reads (2x75) on Illumina Hiseq with V4 chemistry. There were 131 RNA samples, which corresponded with 103 unique HIPSCI cell lines, as some of the samples were differentiation replicates or RNA-extraction replicates. One sample failed in sequencing and was excluded.

Two sets of analyses were done with different genome builds:

- QTL analyses and GWAS overlaps were done with reads aligned to GRCh38;
- all other analyses, including comparisons with GTEx, iPSCs, and DRG, and expression variability, were done with reads aligned to GRCh37. This was so that comparisons were done with identical alignment and counting methods.

For QTL analyses, reads for each sample were aligned to GRCh38 and Ensembl 79 transcript annotations using STAR v2.4.0j with default parameters. We used VerifyBamID v1.1.2 (Jun et al. 2012) to check that RNA-seq sample BAM files matched the corresponding sample genotypes in the core HIPSCI VCF files. This revealed 5 mislabeled RNA samples, for which the correctly matching sample genotypes could be easily determined and corrected, as well as two samples for which no match could be found in HIPSCI genotype data and which were thus excluded (these had been labeled as problematic samples in HIPSCI). For comparisons among tissues, reads for each sample were aligned to the 1000 Genomes GRCh37 reference genome with human decoy sequence 37d5 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz), and with Gencode v19 transcript annotations (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz) using STAR 2.5.3a.

Gene expression quantification, quality control and exclusions

Gene expression counts for QTL calling

GTF files for the Gencode Basic transcript annotations, GRCh38 release 79, were downloaded from www.gencodegenes.org. Gene expression counts were determined using the featureCounts tool of the subread package v1.5.0 (Liao, Smyth, and Shi 2014) with options (-s 2 -p -C -D 2000 -d 25); only uniquely mapping reads were counted. A median of 45 million reads were generated per sample, with median 32.8 million reads (72%) uniquely mapping and assigned to genes. We subsequently excluded short RNAs, pseudogenes, and genes not mapping to chromosomes 1-22, X, Y, or MT, leaving 35,033 unique genes. Expression counts were normalised using conditional quantile normalisation with the R package cqn v5.0.2 (Hansen, Irizarry, and Wu 2012). We defined expressed genes as the 14,215 genes with mean CQN-normalised expression across samples > 1.

We determined pairwise correlation between samples using normalized counts for expressed genes and plotted these as a heatmap. We also plotted the first five principal components of gene expression against each other. These plots identified four outlier samples, which were excluded from subsequent analyses (Supplementary Figure 1). After all exclusions and corrected sample labels, we retained 126 samples from 99 unique donors. For gene expression quantification for QTL calling (both eQTL and sQTL), replicate BAM files from same donor were merged together using samtools. Because genotypes were not available from HIPSCI for two donors, we retained gene expression data for 97 donors for QTL calling.

Gene expression counts for sample comparisons

For all between-tissue comparisons, gene expression counts were determined using featureCounts, as for QTL calling, except that GTF files for Gencode v19 transcript annotations were used, along with BAM files with reads aligned to GRCh37 as described above. 131 sensory neuron samples, 28 DRG samples, and 239 iPSC samples were quantified in this way.

Assessing gene expression replicability

We used R with ggplot2 to plot the CQN-normalized expression for pairs of sample replicates. We excluded 13 samples differentiated using the first version protocol (P1), as most samples (110) were differentiated with the second version (P2), which gave us sufficient samples to consider variability between differentiations without including protocol effects. We determined the spearman correlation coefficient across all genes for (a) extraction replicates, (b) differentiation replicates, and (c) all possible pairs of samples from different donors. The histogram of correlation coefficients for these categories is shown in Supplementary Figure 13.

Dorsal root ganglion samples and sequencing

Human tissue acquisition and handling was performed at Pfizer Neuroscience and Pain Research Unit in accordance with regulatory guidelines and ethical board approval. Postmortem human dorsal root ganglia (DRG) were obtained in dissected form from Anabios or as an encapsulated sheath together with sensory/afferent axons from National Disease Research Interchange which were subsequently dissected to isolate the cell-body rich ganglion. The tissue was homogenised in an appropriate volume QIAzol Lysis Reagent according to weight and processed according to the manufacturer's instructions for the Qiagen RNeasy Plus lipid-rich kit. RNAseq library preparation and sequencing was performed using the Illumina TruSeq Stranded mRNA Library Prep Kit and an Illumina HiSeq 2500 generating 2 x 100 bp reads by Aros Inc. according to the manufacturer's instructions. Sequencing reads were aligned to the GRCh37 reference human genome using STAR and gene counts and FPKMs obtained using featureCounts and Ensembl v75 gene annotations.

ATAC library preparation and sequencing

Nuclei isolation

Media was removed from T25 flasks and washed twice with 10 mL of room temperature D-PBS without calcium and magnesium. The adherent neuronal cultures were lifted by treating with 3 mL of Accutase (Millipore – SCR005) at room temperature for four minutes. The Accutase was quenched by adding 6 mL of 2% foetal bovine serum in D-PBS. The cells were transferred to a 15 mL conical tube and centrifuged at 300 g for 5 minutes at 4 °C. The cell pellet was resuspended in 1 mL of ice-cold sucrose buffer (10 mM tris-Cl pH 7.5, 3 mM $CaCl_2$, 2 mM $MgCl_2$ and 320 mM sucrose) and pipetted briefly to break up the large clumps before incubating on ice for 12 minutes. 50 µL of 10% Triton-X 100 was added to the sucrose-treated cells and mixed briefly before incubating on ice for a further 6 minutes. Nuclei were released by performing 30 strokes with a tight dounce homogeniser on ice. Approximately 1 x 10^5 nuclei were transferred to a 1.5 mL microfuge tube and centrifuged at 300 g for 5 minutes at 4 °C. All traces of the lysis buffer were removed from the nuclei pellet.

Tagmentation, PCR amplification and size selection

The tagmentation and PCR methods used here are in principle the same as that described in Buenrostro et al., 2013, but with some modifications as described in Kumasaka et al., 2016. The nuclei pellet was resuspended in 50 µL of Nextera tagmentation master mix (Illumina FC-121-1030) (25 µL 2x Tagment DNA buffer, 20 µL nuclease-free water and 5 µL Tagment DNA Enzyme 1) and incubated at 37 °C for 30 minutes. The tagmentation reaction was stopped by the addition of 500 µL Buffer PB (Qiagen) and purified using the MinElute PCR purification kit (Qiagen 28004), according to the manufacturer's instructions and eluting in 10 µL of Buffer EB (Qiagen). 10 µL of the tagmented chromatin was mixed with 2.5 µL Nextera PCR primer cocktail and 7.5 µL Nextera PCR mastermix (Illumina FC-121-1030) in a 0.2 mL low-bind PCR tube. The indexing primers used for amplification were from the Nextera Index kit (Illumina FC-121-1011), using 2.5 µL of an i5 primer and 2.5 µL of an i7 primer per PCR, totalling 25 µL. PCR amplification was performed as follows: 72 °C for 3 minutes and 98 °C for 30 seconds, followed by 12 cycles of 98 °C for 10 seconds, 63 °C for 30 seconds and 72 °C for 3 minutes. To remove the excess of unincorporated primers, dNTPS and primer dimers, Agencourt AMPure XP magnetic beads (Beckman Coulter A63880) were used at a ratio of 1.2 AMPure beads:1 PCR sample (v/v), according the manufacturer's instructions, eluting in 20 µL of Buffer EB (Qiagen). Finally, size selection was performed by 1 % agarose TAE gel electrophoresis, selecting library fragments from 120 bp to 1 kb. Gel slices were extracted with the MinElute Gel Extraction kit (Qiagen 28604), eluting in 20 µL of Buffer EB.

Illumina sequencing

A total of 31 ATAC-seq libraries each prepared with a unique Nextera i5 and i7 tag combination were pooled. Index tag ratios were assessed by a single MiSeq run and were balanced before being sequenced at two per lane with paired-end reads (2x75) on a HiSeq with V4 chemistry. However, rebalancing did not appear to work correctly, as the number of reads varied greatly between samples, from a minimum of 17 million to a maximum of 987 million. However, 22 samples had over 100 million reads, and 30 samples had over 40 million reads. Across samples, a median of 56% of reads mapped to mitochondrial DNA. For calling ATAC QTLs we used all sample counts as-is.

Read alignment

We aligned reads to GRCh38 human reference genome using bwa mem v0.7.12 . Reads mapping to the mitochondrial genome and alternative contigs were excluded from all downstream analysis. As for RNA-seq data, we used VerifyBamID v1.1.2 (Jun et al. 2012) to detect sample swaps. This revealed one mislabeled sample, which we then corrected. We used Picard v1.134 MarkDuplicates (https://broadinstitute.github.io/picard/) to mark duplicate fragments. We constructed fragment coverage BigWig files using bedtools v2.21.0 (Quinlan and Hall 2010).

Peak calling

We used MACS2 v2.1.1 (Zhang et al. 2008) to call ATAC-seq peaks individually on sample BAM files with parameters '--nomodel --shift -25 --extsize 50 -q 0.01'. We then constructed a consensus set of peaks by determining regions in which peaks overlapped in at least 3 samples. At regions of overlap, the consensus peak was defined as the union of overlapping peaks. This resulted in 381,323 peaks, with 98% of peaks ranging in size from 82 - 1191 base pairs.

PCA plot clustering samples with GTEx tissues

We downloaded the GTEx v6 gene RPKM file (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz) as well as sample metadata (GTEx_Data_V6_Annotations_SampleAttributesDS.txt) from the GTEx web portal (http://www.gtexportal.org/home/datasets). We computed RPKMs for all genes for the 28 DRG samples, the 119 sensory neuron samples (5 outliers removed), and 239 HIPSCI IPS samples. We used genes that were quantified in all of these sample sets, and where at least 50 GTEx samples had RPKM > 0.1. We passed log2(RPKM + 1) for 8553 GTEx samples to the bigpca R package to compute the first 5 PCs using the SVD method. We the determined sample loadings for each PC using the PC weights and log2(RPKM + 1) values for GTEx samples as well as for our in-house samples, and plotted sample PC1 vs. PC2 values as Figure 1b.

Highly variable genes in IPSDSNs and GTEx

We obtained GTEx v6 RPKM files for all genes as described above. For each of the 44 tissues, as well as IPSDSNs, DRG, and HIPSCI iPSCs, we calculated the coefficient of variation (CV) of each gene among samples with the same detailed tissue type (SMTSD in GTEx sample metadata). We then subsetted the genes considered in each tissue to those expressed at RPKM > 1 in that tissue. We plotted the distribution of CVs across all genes for each tissue as a density plot (Figure 3a).

We used GeneTrail2 (https://genetrail2.bioinf.uni-sb.de) to do a gene set over-representation analysis for the top 1000 most highly variable genes in IPSDSNs by CV, which are included in Supplementary Table 4. Similarly, gene set over-representation analysis in E8-IPSDSN subsets was done using Genetrail2 and the top 1000 most variable genes with RPKM > 1 (Supplementary Table 11).

Variance components analysis

For Figure 3b, we selected the 106 P2 protocol IPSDSN samples after QC exclusions, and used DESeq2 to get FPKM values for each gene after size factor normalization. We included all genes with mean FPKM > 1, and input log2-transformed counts per sample into the variancePartition Bioconductor R package, with design formula ~ (1|donor) + (1|differentiation) + (1|gender) + (1|wasFeeder). We used ggplot2 to plot the distribution of variance explained for each gene across the four above factors, with unexplained variance shown as “residuals”. For Supplementary figure 19a, we included 119 QC-passed samples, and used variancePartition as above, but with protocol in the design formula: ~ (1|donor) + (1|differentiation) + (1|gender) + (1|wasFeeder) + (1|protocol). For Supplementary Figure 19b, we used 18 samples, for which we had 3 differentiation replicates from each of 6 donor cell lines; all 6 iPSC lines were from females and had been cultured in E8 medium. We therefore included only donor and differentiation in the design formula.

Estimation of neuronal purity

We used CIBERSORT (Newman et al. 2015) to estimate the fraction of RNA from neuronal cells in our bulk RNA-seq samples. We used the 14,786 genes whose CQN expression in bulk RNA samples was greater than zero, and retrieved raw counts for these genes in our single cell RNA-seq data. We labeled the single cells as “neuron” or “fibroblast-like” as determined based on the SC3 clustering, and specified these single cell counts as the reference samples for CIBERSORT to generate a custom signature genes file during its analysis. We used raw expression counts for the same genes for our 126 bulk RNA-seq samples as the mixture file for CIBERSORT to use in estimating the relative fractions of neuron and fibroblast-like cell RNA.

Electrophysiological recordings

Six coverslips per line were placed singularly into a 12-well plate and washed 1x with 1 ml DPBS (+/+). After removal of DPBS, the coverslips were coated with 1 ml of 0.33 mg/ml growth factor reduced matrigel for > 3 hr at room temperature. D14 cells were prepared at a suspension of 1.6e6/ml in 15 ml media. The cells were then diluted in NB media to create a 0.3e6/ml suspension. The coverslips were transferred into a clean 12-well plate and 1 ml of the cell suspension was added. Plates were incubated at 37°C (5% CO2) in a cell culture incubator for 24hrs, after which the coverslips were transferred into a clean 12-well plate containing 2 ml media. Cells were then treated with Mitomycin C (0.001 mg/ml for 2hr hours at 37°C) post plating on day 4 and day 10. Media was changed twice weekly.

Patch-clamp experiments were performed in whole-cell configuration using a patch-clamp amplifier 200B for voltage clamp and Multiclamp 700A or 700B for current clamp controlled by Pclamp 10 software (Molecular Devices). Experiments were performed at 35°C or 40°C as noted controlled by an in-line solution heating system (CL-100 from Warner Instruments). Temperature was calibrated at the outlet of the in-line heater daily before the experiments. Patch pipettes had resistances between 1.5 and 2 MΩ. Basic extracellular solution contained (mM) 135 NaCl, 4.7 KCl, 1 $CaCl_2$, 1 $MgCl_2$, 10 HEPES and 10 glucose; pH was adjusted to 7.4 with NaOH. The intracellular (pipette) solution for voltage clamp contained (mM) 100 CsF, 45 CsCl, 10 NaCl, 1 $MgCl_2$, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with CsOH. For current clamp the intracellular (pipette) solution contained (mM) 130 KCl, 1 $MgCl_2$, 5 MgATP, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with KOH. The osmolarity of solutions was maintained at 320 mOsm/L for extracellular solution and 300 mOsm/L for intracellular solutions. All chemicals were purchased from Sigma. Currents were sampled at 20 kHz and filtered at 5 kHz. Between 80% and 90% of the series resistance was compensated to reduce voltage errors. The voltage protocol used for the compounds testing on voltage gated sodium channels consisted of steps from a holding potential of -110 mV to -70 mV for 5 seconds, followed by step to -110 mV for 100 millisecond then currents were measured at step to 0 mV for 20 milliseconds. Intersweep intervals were 15 seconds. Rheobase was measured in current clamp mode by injecting increasing 30 milliseconds current steps until a single action potential was evoked. Intersweep intervals were 2 seconds. Membrane potential was set at either free-resting or held at -70 mV as noted. Current clamp data was analyzed using Spike2 software (Cambridge Electronic Device, UK) and Origin 9.1 software (Originlab).

Correlation of iPSC and IPSDSN gene expression with cell culture conditions

We selected the 106 IPSDSN samples differentiated with the P2 protocol, as well as the 87 iPSC samples these were derived from and for which we had RNA-seq data, and we used DESeq2's variance stabilising transformation on the raw gene expression counts. We computed the first 5 principal components of gene expression separately in iPSC and IPSDSNs with Bioconductor's pcaMethods package, and used corrplot to compute pairwise correlations among these PCs and sample metadata of interest: gender, iPSC passage number, iPSC culture conditions (wasFeeder), iPSC PluriTest score, IPSDSN fibroblast content, and IPSDSN processing date.

We determined differentially expressed genes between feeder-iPSCs and E8-iPSCs using DESeq2, using gene expression counts for all genes with median expression > 0.1 FPKM across iPSC samples (Supplementary Table 5). We removed associations driven by outliers, defined as a maximum Cook's distance >= 5. Similarly, we determined differentially expressed genes in IPSDSNs derived from either feeder-iPSCs or E8-iPSCs (Supplementary Table 8), again for genes with median expression > 0.1 FPKM across samples. We used GeneTrail2 (https://genetrail2.bioinf.uni-sb.de) to do a gene set over-representation analysis for the 717 genes with expression at least 2-fold higher in feeder-iPSCs relative to E8-iPSCs, and similarly for the 631 genes at least 2-fold higher in E8-iPSCs (Supplementary Tables 6, 7). We did an equivalent gene set over-representation analysis for the 1159 genes with expression at least 2-fold higher in IPSDSNs differentiation from feeder-iPSCs, and also for the 958 genes at least 2-fold higher in IPSDSNs from E8-iPSCs (Supplementary Tables 9, 10).

To determine genes upregulated on differentiation from iPSCs to IPSDSNs, we first selected the 19,658 genes with expression FPKM > 1 in at least two samples (iPSC or IPSDSN). We used DESeq2 as before, removing genes with maximum Cook's distance > 5, and identifying 4246 differentially expressed genes at FDR <= 1%.

QTL calling

Expression QTLs

To call cis-eQTLs we used RASQUAL (Kumasaka, Knights, and Gaffney 2015), which leverages allele-specific reads in heterozygous individuals to improve power for QTL discovery, while accounting for reference mapping bias and a number of other potential artifacts. With RASQUAL a feature is defined by a set of start and end coordinates; for calling a gene eQTL these are the start and end coordinates for exons, whereas for an ATAC-seq peak these are the peak coordinates. RASQUAL requires as input the allele-specific read counts at each SNP within a feature. We used the Genome Analysis Toolkit (GATK) program ASEReadCounter (Castel et al. 2015) with options '-U ALLOW_N_CIGAR_READS -dt NONE --minMappingQuality 10 -rf MateSameStrand' to count allele-specific reads at SNPs (and not indels). We then annotated the AS read counts in the INFO field of the VCF used as input for RASQUAL. We used custom scripts to determine the number of feature SNPs in gene exons.

We used RASQUAL's makeCovariates.R script to determine principal components (PCs) to use as covariates, which determined 12 PCs as appropriate from the expression count data. We ran RASQUAL separately for each of 35,033 genes (19,796 protein-coding genes and 15,237 noncoding RNAs), passing in VCF lines for all SNPs and indels (MAF > 0.05, INFO > 0.8) within 500 kb of the gene transcription start site. We used the --no-posterior-update option in RASQUAL, as we found that not doing so led to some genes having miniscule p values, even with permuted data. To correct for multiple testing we used permutations; however, because RASQUAL is computationally intensive, it would not be possible to run a thousand or more permutations for every gene. Therefore we used an approach to balance power and computational time. To correct for the number of SNPs tested per gene, we used EigenMT (Davis et al. 2016) to estimate the number of independent tests per gene, and then performed Bonferroni correction on a gene-by-gene basis. To estimate the false discovery rate (FDR) across genes, we used the --random-permutation option of RASQUAL and re-ran it once for every gene, saving the minimum p value (after eigenMT correction) of the SNPs tested for each gene. This gave a distribution of minimum p values across genes for the permuted data. To determine the FDR for eQTL discovery at a given gene, we use R to compute (#permuted data min pvalues < p) / (#real data min p values < p), where p is the minimum p value among SNPs for the gene in question. With this procedure we obtained 3,586 genes with a cis-eQTL at FDR 10% (2,628 at FDR 5%).

For QTL calling with FastQTL, we first computed principal components from the CQN-transformed gene expression matrix (cqn v5.0.2 (Hansen, Irizarry, and Wu 2012)). We ran FastQTL with permutations 31 separate times, in each run including the first N principal components (N=0...30) as covariates. For each run we used a cis-window of 500 kb, and included SNPs and indels with MAF > 0.05, INFO > 0.8, as we did for RASQUAL. We plotted the number of eGenes found in each of these runs, which plateaued and remained relatively stable at ~1,400 eGenes (FDR 10%) when anywhere from 16 to 30 PCs were used. We arbitrarily chose to use the FastQTL run with 20 PCs in downstream analyses.

ATAC QTLs

As we did for gene expression, we used featureCounts v1.5.0 to count fragments overlapping consensus ATAC-seq peaks and ASEReadCounter to count allele-specific reads at SNPs (and not indels) within peaks. We ran RASQUAL separately for each of 381,323 peaks, passing in VCF lines for SNPs and indels (MAF > 0.05, INFO > 0.8) within 1 kb of the center of the peak. Since >99.9% of peaks were less than 2 kb in size, this meant that we tested effectively all SNPs within peaks. As we did when calling eQTLs, we ran RASQUAL with the --random-permutation option for every gene, and determined FDR as described above. Note that in this case we used Bonferroni correction based on the number of SNPs tested, without using EigenMT, due to the small size of the windows tested. With this procedure we obtained 6,318 ATAC peaks with a cis-QTL at FDR 10%.

Splice QTLs

We downloaded LeafCutter from Github (https://github.com/davidaknowles/leafcutter) on April 17, 2016. We used the LeafCutter bam2junc.sh script to determine junction counts for each sample, followed by leafcutter_cluster.py. This resulted in 254,057 junctions in 59,736 clusters. To focus on splicing events likely to be significant, we applied a number of filters, including: (a) removing junctions accounting for less than 2% of the cluster reads, (b) removing introns used (i.e. having at least 1 supporting read) in fewer than 5 samples, (c) retaining only clusters where at least 10 samples had 20 or more reads in the cluster. This yielded a filtered set of 95,786 junctions in 30,591 clusters. We first determined the read proportions for all junctions within alternatively excised clusters. We then Z-score standardised each junction read proportion across samples, and then quantile-normalised across introns. We used this as our phenotype matrix for input to FastQTL to test for associations between intron usage and variants within 15 kb of the center of each intron. We chose a cis-window size of 30 kb (2 x 15 kb) because >91% of introns are < 30 kb in size, and so this tests variants near exon/intron boundaries for the great majority of introns, while maximising power.

We ran FastQTL in nominal pass mode 31 times specifying the first 0 to 30 principal components as covariates, and examined the number of intron QTLs with minimum SNP p value < 10^{-5}. This showed that the number of QTLs plateaued when 5 PCs were used, and so we used 5 PCs in subsequent runs. We next ran FastQTL with 10,000 permutations to determine empirical p values for each alternatively excised intron. To correct for the number of introns tested per cluster, we used Bonferroni correction on the most significant intron p value per cluster. We then used the Benjamini-Hochberg method to estimate FDR across tested clusters. This yielded 2,079 significant SNP associations for intron usage (sQTLs) at FDR 10%.

For significant sQTLs we used bedtools closest with GRCh38 release 84 to annotate the gene(s) nearest the lead SNP for the association. To ensure we had relevant genes, we filtered the annotation to include only genes where one of the exon boundaries matched the intron boundary for the sQTL.

Similarity of eQTLs with GTEx

Both GTEx samples and IPSDSNs had QTLs called using FastQTL. We selected lead eQTL variants in IPSDSNs for genes with expression >= 1 FPKM. We identified effect sizes for the same variants in each GTEx tissue, where these were available. Because only genes passing certain expression cutoffs were tested in GTEx, each tissue had a different number of values obtained. We next determined the pairwise similarity between tissues in effect sizes for these variants (in R, cor() with option “pairwise.complete.obs”). IPSDSNs were a significant outlier, having lower pairwise similarity with all GTEx tissues than they had with each other. Although FastQTL was used for all tissues, different expression quantification methods used; therefore, a significant batch effect is expected. Therefore we used the relative similarity across tissues by Z-scaling each row of the tissue correlation matrix, and plotted the result in Supplementary Figure 24. IPSDSNs are relatively more similar to GTEx brain in their effect sizes than to other GTEx tissues.

Identifying tissue-specific eQTLs

We determined the set of tissue-specific eQTLs using the same procedure and code as in the HIPSCI project (Kilpinen et al. 2017). Briefly, we considered the full cis eQTL output of sensory neuron eQTLs and 44 tissues analyzed by the GTEx Project (Consortium et al. 2015). To enable comparison, lead SNP positions for sensory neuron eQTLs were first lifted back from GRCh38 to GRCh37 using Crossmap (Zhao et al. 2014). For each discovery tissue (including sensory neurons), we tested for the replication of all lead eQTL - target eGene pairs reported at FDR 5%. If the lead eQTL variant was not reported in the comparison tissue, then the best high-LD proxy of the lead variant ($r^2 > 0.8$ in the UK10k European reference panel) was used as the query variant. Replication was defined as the query variant having a nominal eQTL $p < 2.2x10^{-4}$ (corresponding to p = 0.01 / 45, where 45 refers to the total number of tissues tested) for the same eGene. We then extracted eGenes for which the lead eQTL did not show evidence of replication in any other tissue ($p > 2.2x10^{-4}$) or could not be tested (i.e. was not measured or reported as expressed in any other tissue).

This analysis gave 954 eGenes where the eQTL is specific to sensory neurons (Supplementary Table 15). We note that some of these “tissue-specific” eGenes could be due to the difference in QTL-calling methods used, notably that we used RASQUAL, a method incorporating both allele-specific and population-level expression variation. Therefore, some of the tissue-specific eGenes we report may actually be present more broadly in GTEx tissues but missed by the linear QTL model used in GTEx. Among the 1403 eGenes called by FastQTL, 208 were tissue-specific to IPSDSNs.

Pain-associated genes

We identified a set of pain-associated genes by searching for the term “pain” in the OpenTargets web site (https://www.targetvalidation.org/) on August 22, 2016, and downloading the reported gene associations and scores. We chose a score cutoff of 0.05 to designate a gene as pain-associated, which resulted in 617 genes.

Motif enrichment analyses

We used the R Bioconductor package LOLA (Sheffield and Bock 2015) to identify enrichments in transcription factor binding sites (TFBS) and motifs. We defined three sets of loci to consider for enrichment: 1) tissue-specific eQTL SNPs with a window of 50 bp (+/- 25) around the SNP position, 2) all eQTL SNPs (50 bp window), and 3) all ATAC-seq peaks. For the QTLs we used all GTEx eQTL lead SNPs as the “universe” set against which we were testing TFBS for enrichment. For this we downloaded all GTEx QTL files (*_Analysis.snpgenes), loaded them in R and used the liftOver function from the rtracklayer package to convert their coordinates to the GRCh38 genome version. We tested for enrichment against the LOLA core database but considered only ENCODE TFBS enrichments. These enrichments are reported in Supplementary Tables 16 and 17. We also tested for enrichment against the LOLA extension database and considered JASPAR motif enrichments. No motif enrichments were found for IPSDSN eQTLs relative to GTEx eQTLs. We also tested ATAC-seq peaks for enrichment relative to DNase hypersensitive sites for many tissues from Sheffield et al. (Sheffield et al. 2013), which are available in the LOLA catalog. Many of the same TFBS enrichments were seen for ATAC-seq peaks as for eQTLs (data not shown), although with a skew towards general transcription factors (e.g. CTCF, ATF3, MYC, JUN) as might be expected. Motif enrichments in ATAC-seq peaks are reported in Supplementary Table 18.

Power simulations

Gene expression values were normalized to counts per million. We selected the 544 eGenes discovered by RASQUAL at FDR 1% which met the following criteria:

- at least 10 P2-protocol samples homozygous for each allele of the lead eQTL variant,

- mean expression among homozygous carriers was consistent with RASQUAL's reported direction of effect, and
- CV < 2 (this filter removed only 8 eGenes)

For each gene we resampled the normalized expression values, with replacement, from IPSDSN samples to achieve a specified number N of samples (N ∈ {4,6,10,20,40}) with each homozygous genotype category. From 100 such resamplings, we defined the power (true positive rate, TPR) to discover a given variant's effect as the fraction of cases with $p < 0.05$ from a Wilcoxon rank sum test comparing mean expression in each genotype category. A minimum sample size of 4 in each group is needed for the Wilcoxon rank sum test, as otherwise no difference can be significant at $p < 0.05$. Note that we did the same resampling procedure using Student's t-test, and the results were nearly identical. We determined the allelic fold change between genotypes using RASQUAL's effect size (pi), as:

fold change = max(pi / (1-pi), (1-pi) / pi)

We used ggplot2 with geom_smooth to display the 95% confidence interval around the fitted mean TPR at each parameter combination. As can be seen on the plots, the deviation about this mean for individual genes is larger than the standard error of the mean.

QTL overlap with GWAS catalog

The GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/ on 2016-5-08. To determine overlap between variants in the GWAS catalog and our lead QTLs, we first extracted all lead variants (both QTLs and GWAS catalog variants) from the full VCF file. We used vcftools v0.1.14 (Danecek et al. 2011) to compute the correlation R^2 between all lead variants within 500 kb of each other among our samples. We determined overlap separately for eQTLs, sQTLs, and ATAC QTLs, and retained only overlaps with $R^2 > 0.8$ between lead variants. Note that a given GWAS variant may be in LD with an eQTL for more than one gene, and vice versa, an eQTL for a single gene may be in LD with more than one GWAS catalog entry.

We used QTL-GWAS overlap for two purposes: first, to find individual cases where a QTL is a strong candidate as a causal association for the GWAS trait, and second, to determine whether any GWAS catalog traits are enriched overall for overlap with sensory neuron QTLs. For the first goal, we considered all overlaps with GWAS catalog associations having $p < 5x10^{-8}$, i.e. did not filter any redundant overlaps. These overlaps are reported in Supplementary Tables 20 (for eQTLs), 21 (for sQTLs), and 22 (for ATAC QTLs).

To determine whether our QTL overlaps were enriched in any specific GWAS catalog traits relative to other traits, we computed overlap with all GWAS catalog SNPs ($p < 5x10^{-8}$) but sought to eliminate redundant overlaps. For traits that were reported with differing names (e.g. "Alzheimer's disease (cognitive decline)" and "Alzheimer's disease in APOE e4- carriers"), we grouped these into a single trait name (e.g. "Alzheimer's disease"). We then sorted overlaps by decreasing LD R^2, and kept the single overlapping QTL with the highest R^2 for each GWAS catalog entry. Similarly, we removed duplicates with the same reported GWAS catalog SNP and trait, such as when successive GWAS of the same trait report the same SNP association. We counted the number of such unique GWAS-QTL overlaps separately for eQTLs, sQTLs, and caQTLs, and we report these in Table 1. To avoid bias due to correlation between GWAS power and LD patterns, we restricted our analysis to the 41 traits with at least 40 GWAS catalog associations. We then considered the binomial probability of the observed overlap with each trait, with the expected overlap frequency being the proportion of QTL overlaps among all trait associations (6.2%). After correcting for multiple testing, no traits showed significantly greater overlap with our QTL catalog than other traits.

To test for overall enrichment of QTL overlapping with GWAS catalog SNPs, we downloaded the 1000 genomes VCF files (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and subsetted these to the EUR samples. We used vcftools to identify all SNPs in LD $R^2 > 0.8$ with a GWAS catalog SNP and removed duplicate SNPs. We used our IPSDSN eQTL lead SNPs as input to SNPsnap (https://data.broadinstitute.org/mpg/snpsnap/), and computed 1000 random sets of SNPs using default parameters to match for LD partners, MAF, gene density, and distance to nearest gene. We determined the number of occurrences of eQTL lead SNPs in the GWAS catalog SNP + LD partners, and did the same for the 1000 matched SNP sets. The IPSDSN eQTL lead SNPs had more overlaps (92) than any of the matched sets (median: 58, range 37-87). Note that this number of overlaps is fewer than the number we report in Supplementary Table 20; this is because we detect more overlaps when using LD from our own samples than when using 1000 genomes LD patterns, which is expected since 1000 genomes EUR LD does not perfectly reflect LD in our data. We performed the same overlapping process for lead eQTL SNPs from each GTEx tissue, and plotted the number of overlaps per tissue in Supplementary Figure 25.

References

Baron, Maayan, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. 2016. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure.” *Cell Systems* 3 (4): 346–60. doi:10.1016/j.cels.2016.08.011.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12). Nature Publishing Group: 1213–18. doi:10.1038/nmeth.2688.

Cacchiarelli, Davide, Xiaojie Qiu, Sanjay Srivatsan, Michael Ziller, Eliah Overbey, Tarjei Mikkelsen, Regenerative Biology, Cellular Biology Program, and South San Francisco. 2017. “Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Reprogramming Outcome.”

Cao, Lishuang, Aoibhinn McDonnell, Anja Nitzsche, Aristos Alexandrou, Pierre-philippe Saintot, Alexandre J C Loucif, Adam R Brown, et al. 2016. “Pharmacological Reversal of a Pain Phenotype in iPSC-Derived Sensory Neurons and Patients with Inherited Erythromelalgia.” *Science Translational Medicine* 8 (335): 335ra56. doi:10.1126/scitranslmed.aad7653.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (1): 195. doi:10.1186/s13059-015-0762-6.

Chambers, Stuart M, Yuchen Qi, Yvonne Mica, Gabsang Lee, Xin-jun Zhang, Lei Niu, James Bilsland, et al. 2012. “Combined Small-Molecule Inhibition Accelerates Developmental Timing and Converts Human Pluripotent Stem Cells into Nociceptors.” *Nature Biotechnology* 30 (7). Nature Publishing Group: 715–20. doi:10.1038/nbt.2249.

Consortium, The GtEx, Kristin G. Ardlie, David S. Deluca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, et al. 2015. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60. doi:10.1126/science.1262110.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. doi:10.1093/bioinformatics/btr330.

Dianat, Noushin, Clara Steichen, Ludovic Vallier, Anne Weber, and Anne Dubart-Kupperschmitt. 2013. “Human Pluripotent Stem Cells for Modelling Human Liver Diseases and Cell Therapy.” *Current Gene Therapy* 13 (2): 120–32. doi:10.2174/1566523211313020006.

Gregory, Adam P, Calliope A Dendrou, Kathrine E Attfield, Aiden Haghikia, Dionysia K Xifara, Falk Butter, Gereon Poschmann, et al. 2012. “TNF Receptor 1 Genetic Risk Mirrors Outcome of Anti-TNF Therapy in Multiple Sclerosis.” *Nature* 488 (7412): 508–11. doi:10.1038/nature11307.

Handel, Adam E., Satyan Chintawar, Tatjana Lalic, Emma Whiteley, Jane Vowles, Alice Giustacchini, Karene Argoud, et al. 2016. “Assessing Similarity to Primary Tissue and Cortical Layer Identity in Induced Pluripotent Stem Cell-Derived Cortical Neurons through Single-Cell Transcriptomics.” *Human Molecular Genetics* 25 (5): 989–1000. doi:10.1093/hmg/ddv637.

Hansen, Kasper D., Rafael A. Irizarry, and Zhijin Wu. 2012. “Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization.” *Biostatistics* 13 (2): 204–16. doi:10.1093/biostatistics/kxr054.

Hu, Bao-Yang, Jason P Weick, Junying Yu, Li-Xiang Ma, Xiao-Qing Zhang, James a Thomson, and Su-Chun Zhang. 2010. “Neural Differentiation of Human Induced Pluripotent Stem Cells Follows Developmental Principles but with Variable Potency.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (9): 4335–40. doi:10.1073/pnas.0910012107.

Hunt, S P, A Pini, and G Evan. 1987. “Induction of c-Fos-like Protein in Spinal Cord Neurons Following Sensory Stimulation.” *Nature* 328 (6131): 632–34. doi:10.1038/328632a0.

Itzhaki, Ilanit, Leonid Maizels, Irit Huber, Limor Zwi-Dantsis, Oren Caspi, Aaron Winterstern, Oren Feldman, et al. 2011. “Modelling the Long QT Syndrome with Induced Pluripotent Stem Cells.” *Nature* 471 (7337): 225–29. doi:10.1038/nature09747.

Jordan, J. Dedrick, John Cijiang He, Narat J. Eungdamrong, Ivone Gomes, Wasif Ali, Tracy Nguyen, Trever G. Bivona, Mark R. Philips, Lakshmi A. Devi, and Ravi Iyengar. 2005. “Cannabinoid Receptor-Induced Neurite Outgrowth Is Mediated by Rap1 Activation through Gαo/i-Triggered Proteasomal Degradation of Rap1GAPII.” *Journal of Biological Chemistry* 280 (12): 11413–21. doi:10.1074/jbc.M411521200.

Jun, Goo, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. 2012. “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data.” *American Journal of Human Genetics* 91 (5): 839–48. doi:10.1016/j.ajhg.2012.09.004.

Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Sofie Ashford, Sendu Bala, Dalila Bensaddek, et al. 2017. “Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs.” *Nature*. Nature Publishing Group, 55160. doi:10.1101/055160.

Kiselev, Vladimir Yu., Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. 2016. “SC3 - Consensus Clustering of Single - Cell RNA - Seq Data.” *Bioarxiv*. doi:http://dx.doi.org/10.1101/036558.

Kohno, Tatsuro, Kimberly a Moore, Hiroshi Baba, and Clifford J Woolf. 2003. “Peripheral Nerve Injury Alters Excitatory Synaptic Transmission in Lamina II of the Rat Dorsal Horn.” *The Journal of Physiology* 548 (Pt 1): 131–38. doi:10.1113/jphysiol.2002.036186.

Kumasaka, Natsuhiko, Andrew Knights, and Daniel Gaffney. 2015. “Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq.” *bioRxiv* 48 (2): 18788. doi:10.1101/018788.

Lee, G, E P Papapetrou, H Kim, S M Chambers, M J Tomishima, C A Fasano, Y M Ganat, et al. 2009. “Modelling Pathogenesis and Treatment of Familial Dysautonomia Using Patient-Specific iPSCs.” *Nature* 461 (7262): 402–6. doi:10.1038/nature08320.

Lessard, Julie, Jiang I. Wu, Jeffrey A. Ranish, Mimi Wan, Monte M. Winslow, Brett T. Staahl, Hai Wu, Ruedi Aebersold, Isabella A. Graef, and Gerald R. Crabtree. 2007. “An Essential Switch in Subunit Composition of a Chromatin Remodeling Complex during Neural Development.” *Neuron* 55 (2): 201–15. doi:10.1016/j.neuron.2007.06.019.

Li, Yang I, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. 2016. “RNA Splicing Is a Primary Link between Genetic Variation and Disease.” *Science (New York, N.Y.)* 352 (6285): 600–604. doi:10.1126/science.aad9417.

Li, Yang I, David A Knowles, and Jonathan K Pritchard. 2016. “LeafCutter: Annotation-Free Quantification of RNA Splicing.” *bioRxiv*, 44107. doi:10.1101/044107.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. doi:10.1093/bioinformatics/btt656.

Liu, Guang-Hui, Basam Z. Barkho, Sergio Ruiz, Dinh Diep, Jing Qu, Sheng-Lian Yang, Athanasia D. Panopoulos, et al. 2011. “Recapitulation of Premature Ageing with iPSCs from Hutchinson–Gilford Progeria Syndrome.” *Nature* 472 (7342): 221–25. doi:10.1038/nature09879.

Melchionda, Laura, Tobias B. Haack, Steven Hardy, Truus E M Abbink, Erika Fernandez-Vizarra, Eleonora Lamantea, Silvia Marchet, et al. 2014. “Mutations in APOPT1, Encoding a Mitochondrial Protein, Cause Cavitating Leukoencephalopathy with Cytochrome c Oxidase Deficiency.” *American Journal of Human Genetics* 95 (3): 315–25. doi:10.1016/j.ajhg.2014.08.003.

Mele, M., P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, et al. 2015. “The Human Transcriptome across Tissues and Individuals.” *Science* 348 (6235): 660–65. doi:10.1126/science.aaa0355.

Musunuru, Kiran, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V

Sachs, Xiaoyu Li, et al. 2010. “From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus.” *Nature* 466 (7307): 714–19. doi:10.1038/nature09266.

Newman, Aaron M, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash a Alizadeh. 2015. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles.” *Nature Methods* 12 (5): 1–10. doi:10.1038/nmeth.3337.

Ongen, Halit, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. 2016. “Fast and Efficient QTL Mapper for Thousands of Molecular Phenotypes.” *Bioinformatics* 32 (10): 1479–85. doi:10.1093/bioinformatics/btv722.

Pashos, Evanthia E, Yoson Park, Xiao Wang, Daniel J Rader, Christopher D Brown, and Kiran Musunuru. 2017. “Large , Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci Resource Large , Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Gen.” *Stem Cell* 20 (4). Elsevier Inc.: 558–570.e10. doi:10.1016/j.stem.2017.03.017.

Pers, Tune H., Pascal Timshel, and Joel N. Hirschhorn. 2014. “SNPsnap: A Web-Based Tool for Identification and Annotation of Matched SNPs.” *Bioinformatics* 31 (3): 418–20. doi:10.1093/bioinformatics/btu655.

Peters, Marjolein J, Linda Broer, Hanneke L D M Willemen, Gudny Eiriksdottir, Lynne J Hocking, Kate L Holliday, Michael A Horan, et al. 2013. “Genome-Wide Association Study Meta-Analysis of Chronic Widespread Pain: Evidence for Involvement of the 5p15.2 Region.” *Annals of the Rheumatic Diseases* 72 (3): 427–36. doi:10.1136/annrheumdis-2012-201742.

Probert, L. 2015. “TNF and Its Receptors in the CNS: The Essential, the Desirable and the Deleterious Effects.” *Neuroscience*. doi:10.1016/j.neuroscience.2015.06.038.

Quinlan, Aaron R, and Ira M Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Sala, Luca, Milena Bellin, and Christine L. Mummery. 2016. “Integrating Cardiomyocytes from Human Pluripotent Stem Cells in Safety Pharmacology: Has the Time Come?” *British Journal of Pharmacology*, 1–17. doi:10.1111/bph.13577.

Sheffield, Nathan C., and Christoph Bock. 2015. “LOLA: Enrichment Analysis for Genomic Region Sets and Regulatory Elements in R and Bioconductor.” *Bioinformatics* 32 (4): 587–89. doi:10.1093/bioinformatics/btv612.

Sheffield, Nathan C., Robert E. Thurman, Lingyun Song, Alexias Safi, John A. Stamatoyannopoulos, Boris Lenhard, Gregory E. Crawford, and Terrence S. Furey. 2013. “Patterns of Regulatory Activity across Diverse Human Cell Types Predict Tissue Identity, Transcription Factor Binding, and Long-Range Interactions.” *Genome Research* 23 (5): 777–88. doi:10.1101/gr.152140.112.

Smith, Brenden W., Sarah S. Rozelle, Amy Leung, Jessalyn Ubellacker, Ashley Parks, Shirley K. Nah, Deborah French, et al. 2013. “The Aryl Hydrocarbon Receptor Directs Hematopoietic Progenitor Cell Expansion and Differentiation.” *Blood* 122 (3): 376–85. doi:10.1182/blood-2012-11-466722.

Soldner, Frank, Yonatan Stelzer, Chikdu S. Shivalila, Brian J. Abraham, Jeanne C. Latourelle, M. Inmaculada Barrasa, Johanna Goldmann, Richard H. Myers, Richard A. Young, and Rudolf Jaenisch. 2016. “Parkinson-Associated Risk Variant in Distal Enhancer of α-Synuclein Modulates Target Gene Expression.” *Nature* 533 (7601). Nature Publishing Group: 1–20. doi:10.1038/nature17939.

Spilker, Christina, and Michael R. Kreutz. 2010. “RapGAPs in Brain: Multipurpose Players in Neuronal Rap Signalling.” *European Journal of Neuroscience*. doi:10.1111/j.1460-9568.2010.07273.x.

Wainger, Brian J., Evangelos Kiskinis, Cassidy Mellin, Ole Wiskow, Steve S W Han, Jackson Sandoe, Numa P. Perez, et al. 2014. “Intrinsic Membrane Hyperexcitability of Amyotrophic Lateral Sclerosis Patient-Derived Motor Neurons.” *Cell Reports* 7 (1): 1–11. doi:10.1016/j.celrep.2014.03.019.

Warren, Curtis R., Cashell E. Jaquish, Chad A. Cowan, C.E. Becker, X. Zhang, P. Liu, Y. Wakabayashi, et al. 2017. “The NextGen Genetic Association Studies Consortium: A Foray into In Vitro Population Genetics.” *Cell Stem Cell* 20 (4). Elsevier Inc.: 431–33. doi:10.1016/j.stem.2017.03.021.

Warren, Curtis R, John F O Sullivan, Max Friesen, Ramachandran S Vasan, Christopher J O Donnell, Chad A Cowan, Curtis R Warren, et al. 2017. “Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Resource Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease.” *Stem Cell* 20 (4). Elsevier Inc.: 547–557.e7. doi:10.1016/j.stem.2017.01.010.

Young, Gareth T, Alex Gutteridge, Heather D E Fox, Anna L Wilbrey, Lishuang Cao, Lily T Cho, Adam R Brown, et al. 2014. “Characterizing Human Stem Cell-Derived Sensory Neurons at the Single-Cell Level Reveals Their Ion Channel Expression and Utility in Pain Research.” *Molecular Therapy : The Journal of the American Society of Gene Therapy* 22 (8): 1530–43. doi:10.1038/mt.2014.86.

Zhang, Yong, Tao Liu, Clifford a Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137. doi:10.1186/gb-2008-9-9-r137.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean Pierre Kocher, and Liguo Wang. 2014. “CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies.” *Bioinformatics* 30 (7): 1006–7. doi:10.1093/bioinformatics/btt730.