nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0129-5

Genetic identification of brain cell types underlying schizophrenia

Nathan G. Skene[1,2,11], Julien Bryois[3,11], Trygve E. Bakken[4], Gerome Breen[5,6], James J. Crowley[7], Héléna A. Gaspar[5,6], Paola Giusti-Rodriguez[7], Rebecca D. Hodge[4], Jeremy A. Miller[4], Ana B. Muñoz-Manchado[1], Michael C. O'Donovan[8], Michael J. Owen[8], Antonio F. Pardiñas[8], Jesper Ryge[9], James T. R. Walters[8], Sten Linnarsson[1], Ed S. Lein[4], Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium[10], Patrick F. Sullivan[3,7]* and Jens Hjerling-Leffler[1]*

With few exceptions, the marked advances in knowledge about the genetic basis of schizophrenia have not converged on findings that can be confidently used for precise experimental modeling. By applying knowledge of the cellular taxonomy of the brain from single-cell RNA sequencing, we evaluated whether the genomic loci implicated in schizophrenia map onto specific brain cell types. We found that the common-variant genomic results consistently mapped to pyramidal cells, medium spiny neurons (MSNs) and certain interneurons, but far less consistently to embryonic, progenitor or glial cells. These enrichments were due to sets of genes that were specifically expressed in each of these cell types. We also found that many of the diverse gene sets previously associated with schizophrenia (genes involved in synaptic function, those encoding mRNAs that interact with FMRP, antipsychotic targets, etc.) generally implicated the same brain cell types. Our results suggest a parsimonious explanation: the common-variant genetic results for schizophrenia point at a limited set of neurons, and the gene sets point to the same cells. The genetic risk associated with MSNs did not overlap with that of glutamatergic pyramidal cells and interneurons, suggesting that different cell types have biologically distinct roles in schizophrenia.

Knowledge of the genetic basis of schizophrenia has markedly improved in the past five years[1]. We now know that much of the genetic basis and heritability of schizophrenia is due to common variation[2,3]. However, identifying 'actionable' genes in sizable studies[4,5] has proven difficult, with a few exceptions[6–8]. For example, there is aggregated statistical evidence for diverse gene sets including genes expressed in brain or neurons[3,5,9], genes highly intolerant of loss-of-function variation[10], genes involved in synaptic function[11] (hereafter referred to as synaptic genes), genes whose mRNA bind to FMRP[12] and glial genes[13] (Supplementary Table 1). Several gene sets have been implicated by both common- and rare-variant studies of schizophrenia, and this convergence strongly implicates these gene sets in the pathophysiology of schizophrenia. However, the gene sets in Supplementary Table 1 often contain hundreds of functionally distinct genes that do not immediately suggest reductive targets for experimental modeling.

Connecting the genomic results to cellular studies is crucial, as it would allow us to prioritize for cells that are fundamental to the genesis of schizophrenia. Enrichment of schizophrenia genomic findings in genes expressed in macroscopic samples of brain tissue has been reported[3,14,15], but these results are insufficiently specific to guide subsequent experimentation.

A more precise approach has recently become feasible. Single-cell RNA sequencing (scRNA-seq) can be used to derive empirical taxonomies of brain cell types. We thus rigorously compared genomic results for schizophrenia to brain cell types defined by scRNA-seq. Our goal was to connect human genomic findings with the specific brain cell types defined by gene expression profiles, and to ascertain what specific brain cell types the common-variant genetic findings for schizophrenia best 'fit' to. A schematic of our approach is shown in Fig. 1.

Results

Cell-type specificity of gene expression. We assembled a superset of brain scRNA-seq data from the Karolinska Institutet (KI; Supplementary Tables 2 and 3). Brain regions in the KI superset included the neocortex[16], hippocampus[16], hypothalamus[17], striatum and midbrain[18], as well as samples enriched for oligodendrocytes, dopaminergic neurons and cortical parvalbuminergic interneurons (total of 9,970 cells; Fig. 1c). These data were generated using identical methods from the same labs, with unique molecular identifiers that allowed for direct comparison of transcription across regions. Quality control and alignment are described elsewhere[16]. We did not identify important batch effects (Supplementary Fig. 1).

[1]Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [2]UCL Institute of Neurology, Queen Square, London, UK. [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [4]Allen Institute for Brain Science, Seattle, WA, USA. [5]King's College London, Institute of Psychiatry, Psychology and Neuroscience, MRC Social, Genetic and Developmental Psychiatry (SGDP) Centre, London, UK. [6]National Institute for Health Research Biomedical Research Centre, South London and Maudsley National Health Service Trust, London, UK. [7]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [8]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [9]Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [10]A list of members and affiliations appears in the Supplementary Note. [11]These authors contributed equally: Nathan G. Skene, Julien Bryois. *e-mail: patrick.sullivan@ki.se; jens.hjerling-leffler@ki.se

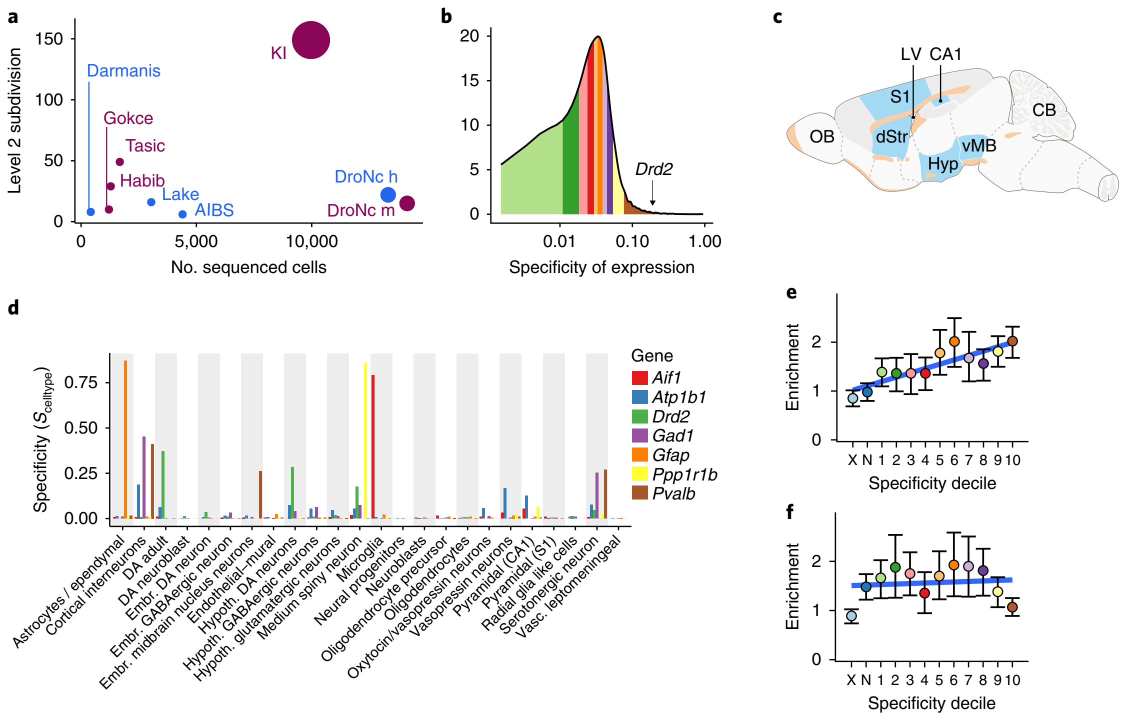

Fig. 1 | Specificity metric calculated from single-cell transcriptome sequencing data can be used to test for increased burden of schizophrenia-SNP heritability in brain cell types. a, Comparison of level 2 cell-type categories and number of cells subjected to snRNA-seq or scRNA-seq from brain tissue. Plum-colored circles are mouse studies, and blue circles are human studies. The number of different tissues is reflected in the size of the circle (see Supplementary Table 2 for citations). AIBS, Allen Institute for Brain Science; KI, Karolinska Institutet. **b**, Histogram of specificity metric ($S_{\mathrm{MSN,KI}}$) for MSNs from the KI superset level 1. Colored regions indicate deciles (the brown region contains the genes most specific to MSNs). Specificity value for dopamine receptor D2 (*Drd2*, $S_{\mathrm{MSN,KI,Drd2}} = 0.17$) is indicated by the arrow. **c**, Schematic highlighting the brain regions sampled in the KI dataset in blue. **d**, Specificity values in the KI level 1 dataset for a range of known cell type markers. Embr., embryonic; Hypoth., hypothalamic. **e**, Enrichment of schizophrenia-SNP heritability in each of the specificity deciles for MSNs (calculated using LDSC). Color of dots corresponds to regions of the specificity matrix in **b**. The light blue dot (marked "X") represents all SNPs that map onto named transcripts that are not MGI-annotated genes or that map onto a gene which does not have a 1:1 mouse:human ortholog. The dark blue dot (marked "N") represents all SNPs that map onto genes not expressed in MSNs. Blue line shows the linear regression slope fitted to the enrichment values. **f**, Enrichment of height-SNP heritability in each of the specificity deciles for MSNs. Colors as in **e**. In **e** and **f**, error bars indicate the 95% confidence intervals.

On the basis of the scRNA-seq data and subsequent clustering analysis, each cell was assigned to a level 1 classification (for example, pyramidal cell, microglia or astrocyte). Level 2 classifications were subtypes of a level 1 grouping (for example, medium spiny neurons expressing Drd1 or Drd2). Clustering was based on patterns of correlations across hundreds of genes and not on single markers. After clustering, cell-type identities were derived using known expression patterns, histology and/or molecular studies[16–18] (Supplementary Table 2). The KI mouse superset identified 24 level 1 brain cell types (Supplementary Fig. 2) and 149 level 2 cell types (all subgroupings of level 1), which were far more than any other brain scRNA-seq or single-nucleus RNA-seq (snRNA-seq) dataset presently available (Fig. 1a).

For each scRNA-seq and snRNA-seq dataset, we estimated the specificity of each gene and cell type. This measure represents the proportion of the total expression of a gene found in one cell type as compared to that in all cell types (i.e., the mean expression in one cell type divided by the mean expression in all cell types). If the expression of a gene is shared between two or more cell types, it will get a lower specificity measure. For example, *Drd2* was highly expressed in MSNs, adult dopaminergic neurons and hypothalamic interneurons, and its specificity measure in MSNs was 0.17; however, *Drd2* was in the top specificity decile for MSNs (Fig. 1b). Fig. 1c shows cell-type specificity for seven genes with known expression patterns. Because expression was spread over several cell types, the pan-neuronal marker *Atp1b1* had lower specificity than *Ppp1r1b* (DARPP-32, an MSN marker), *Aif1* (a microglia marker) or *Gfap* (an astrocyte marker).

Cell-type specificity of schizophrenia genetic associations. For each cell type, we ranked the expression specificity of each gene into groups (deciles or 40 quantiles). The underlying hypothesis was that if schizophrenia was associated with a particular cell type, then more of the genome-wide association signal would be concentrated in genes with greater cell-type specificity. For example, we plotted the enrichment of single-nucleotide polymorphism (SNP) heritability for schizophrenia and human height in the cell-type specificity deciles for MSNs and found a positive relationship for schizophrenia, but no relationship with human height (Fig. 1d,e). To ensure rigor, we required that two different statistical methods (LDSC[9] and MAGMA[19]) each give strong evidence for connecting schizophrenia genome-wide association studies (GWAS) to a cell type. These two methods are based on different assumptions and algorithms. Linkage disequilibrium score regression (LDSC) assessed enrichment of the common SNP heritability of schizophrenia in the most cell-type-specific genes. MAGMA evaluated whether gene-level genetic association with schizophrenia linearly increased with cell-type expression specificity. Both methods accounted for confounders like gene size and linkage disequilibrium in different ways. We required that both methods give similar results after correcting for

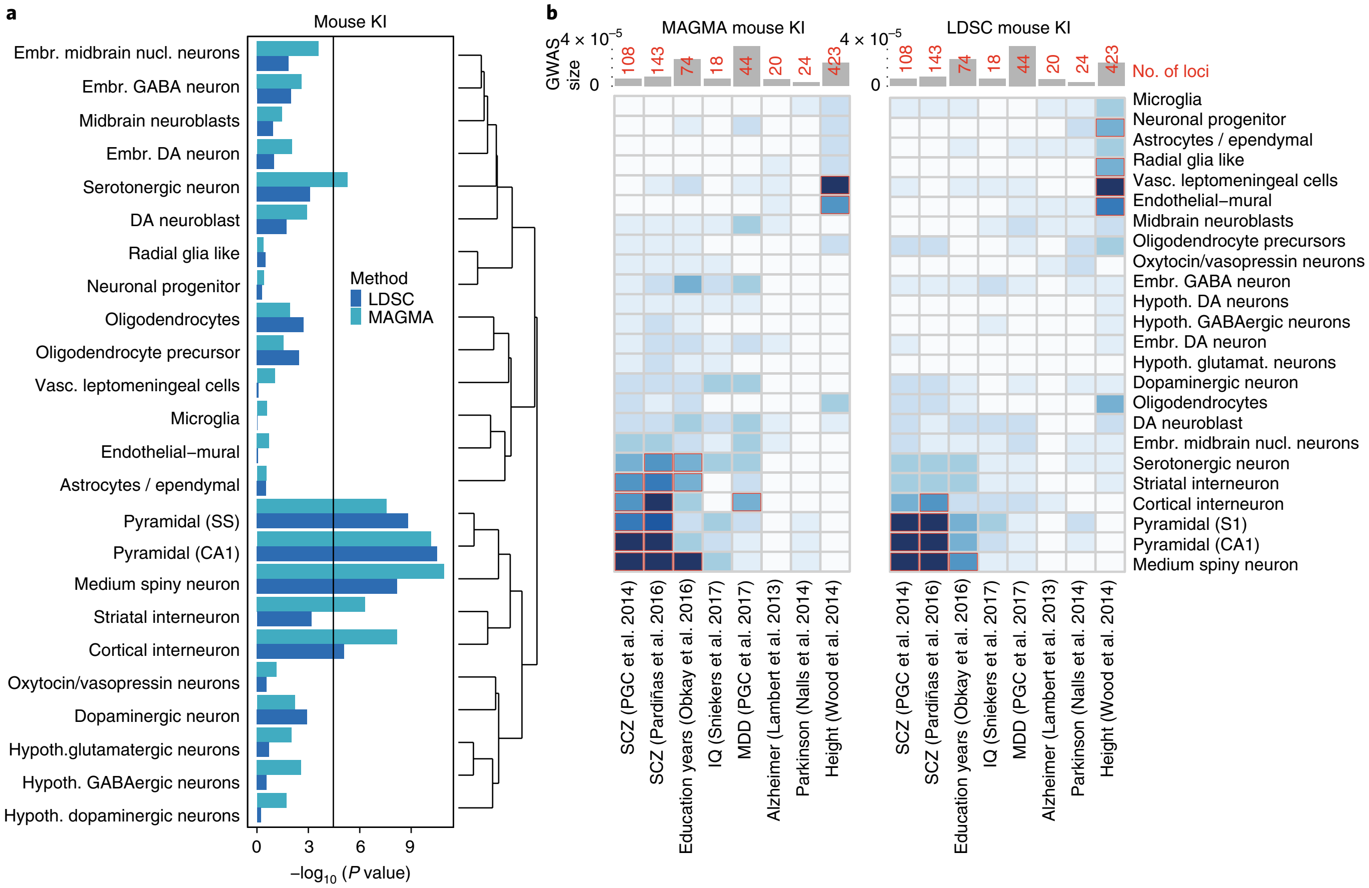

Fig. 2 | Evaluation of enrichment of common-variant CLOZUK schizophrenia GWAS results in the KI brain scRNA-seq dataset from mouse. a, KI level 1 brain cell types. Analyses from both LDSC and MAGMA show enrichment for pyramidal neurons (somatosensory cortex and hippocampus CA1), striatal MSNs and cortical interneurons. The black line is the Bonferroni significance threshold ($P < 0.05/((24+149)\times8)$). **b**, Heat map of association *P* values of diverse human genome-wide association with KI level 1 mouse brain cell types using MAGMA (left) and LDSC (right). Bonferroni-significant results are marked with red borders ($P < 0.05/((24+149)\times8)$). Total number of cases and controls used in the GWAS are shown in the top bar plots, where numbers in red indicate the amount of genome-wide significant loci identified. The CLOZUK results do not generalize indiscriminately across human diseases or traits. In the more-sensitive MAGMA analysis, major depressive disorder (MDD) is primarily enriched in cortical interneurons and embryonic midbrain neurons, unlike in schizophrenia. Similar, but nonsignificant, trends can be observed using LDSC.

multiple comparisons to minimize the chance of a spurious conclusion. As described in the Methods, we evaluated and excluded multiple potential threats to the validity of these analyses.

To identify the brain cell types that were associated with schizophrenia, we used the largest available GWAS of schizophrenia, CLOZUK, which identified ~140 genome-wide significant loci in 40,675 cases and 64,643 controls[20]. We first compared the CLOZUK results to those in the GTEx database (RNA-seq analysis of macroscopic samples from multiple human tissues)[21] by using MAGMA and confirmed[3] that smaller schizophrenia genome-wide association *P* values were substantially enriched in the brain and pituitary (Supplementary Fig. 3).

We evaluated the relation of the CLOZUK genome-wide association schizophrenia results to the 24 KI level 1 brain cell types. Both LDSC and MAGMA analyses strongly highlighted only four cell types: hippocampal CA1 pyramidal cells, striatal MSNs, neocortical somatosensory pyramidal cells and cortical interneurons (Fig. 2a and Supplementary Figs. 4 and 5). Each exceeded a Bonferroni significance level by several orders of magnitude. The results were not pan-neuronal, as multiple other types of neurons did not show enrichment. Schizophrenia risk was greater in mature cells than in embryonic or progenitor cells. We extended the analysis to 149 KI level 2 cell types (subtypes of level 1 cells): for hippocampal CA1 pyramidal cells, both major subgroups were significant; for the striatum, MSNs expressing *Drd2*, MSNs expressing *Drd1* and/or *Drd1*, and striatal *Pvalb*-expressing interneurons were consistently significant; and for neocortical somatosensory pyramidal cells, cortical layers 2/3, 4, 5 and 6 were significant (Supplementary Fig. 6). The cortical level 1 interneuron signal appeared to result from four interneuron subcategories, all of which expressed *Reln*.

Additional analyses showed that these results were not influenced by the total number of molecules detected per cell type or by total number of cells per cell type (Supplementary Table 3). We conducted null simulations and confirmed that there was no type 1 error inflation (Supplementary Fig. 7). We also applied an alternative approach based on differential expression[22] and replicated the association of MSNs, pyramidal CA1 and neocortical somatosensory pyramidal cells with schizophrenia by using a third method (Supplementary Fig. 8). These additional analyses suggested the robustness of our results.

We next evaluated whether these results were specific to schizophrenia or whether they resulted from some feature that was common across human traits. Heat maps of KI level 1 enrichment *P* values for genome-wide association results from eight studies of human complex traits are depicted in Fig. 2b. Seven studies evaluated common-variant associations for brain-related diseases or traits with ≥20,000 cases and ≥10 genome-wide significant associations. Human height was included as a non-brain-related comparator. The results from the earlier Psychiatric Genomics Association (PGC) GWAS of schizophrenia[3] were similar to those from CLOZUK. Although we observed cell types being enriched in other sets, none had the specific signal observed in the two schizophrenia sets. For example, for major depressive disorder, we found that GABAergic interneurons, embryonic midbrain neurons and dopaminergic interneurons were the most enriched cell types. For each cell type, we tested whether the enrichment observed in other GWAS was significantly different from that in CLOZUK. We observed no significant difference for SCZ2 (a subset of CLOZUK) and years of education, but all of the other studies contained significantly different cell-type enrichments (Supplementary Fig. 9).

Replication of results in additional single-cell datasets. We replicated most of the findings in independent scRNA-seq and snRNA-seq mouse brain studies. We found significant enrichment for schizophrenia in hippocampal CA1 pyramidal cells, neocortical pyramidal cells, cortical interneurons (although not in all datasets) and MSNs[23–26]. We also saw enrichment in pyramidal neurons from CA3 and dentate gyrus granule cells (Supplementary Fig. 10a–d). Replication of our results in other external datasets again highlights the robustness of our cell-type association results.

We identified an important technical issue for scRNA-seq and snRNA-seq studies of brain. scRNA-seq is readily done in mouse brain but is more difficult in larger and more fragile human brain neurons. Nearly all of the currently available human data have been generated using snRNA-seq. The isolated nuclei used in snRNA-seq lack the cytoplasmic compartment and proximal dendrites, and there are systematic differences between the types and amounts of mRNA in the nucleus versus those in the cell soma[27]. To evaluate the effect of this issue, we analyzed multiple mouse and human datasets. We confirmed that transcripts destined for export to the synaptic neuropil[28] were better captured by scRNA-seq and specifically depleted in snRNA-seq (Fig. 3a). This was important for the purposes of this study because synaptic neuropil transcripts are enriched for genetic associations with schizophrenia ($P = 1.6 \times 10^{-4}$). This places an important caveat on the use of snRNA-seq to evaluate brain cell-type associations with schizophrenia, given that snRNA-seq from human or mouse brain may not comprehensively capture the relevant transcriptome.

With these caveats in mind, we evaluated human snRNA-seq datasets from mid-temporal cortex (Allen Institute for Brain Science, unpublished) and massively parallel snRNA-seq with droplet technology (DroNc-seq) datasets from the prefrontal cortex and hippocampus[26]. Using hierarchical clustering on specificity scores, we found that human and mouse cell types clustered together (Supplementary Fig. 11); level 1 cell types had greater similarity to the same cell type across species than to a different cell type in the same species. We confirmed enrichment of schizophrenia SNP heritability in cortical pyramidal neurons (glutamatergic cells) and cortical interneurons (GABAergic cells) in two different human datasets (Fig. 3b). In the DroNc-seq dataset[26], we confirmed enrichment in hippocampal pyramidal neurons (glutamatergic cells), along with greater enrichment in *Reln*-expressing GABAergic interneurons, as compared to those expressing *Pvalb*. In both human studies, oligodendrocyte precursor cells (OPCs) were significant or close to significance, but it was hard to judge whether this was related to a loss of neuronal-specific signal in snRNA-seq (note that OPCs showed stronger signal in OPCs in mouse snRNA-seq versus that in scRNA-seq) (Fig. 2 and Supplementary Fig. 10d). In a small scRNA-seq study[29], human adult and fetal cortical neurons were significantly enriched for schizophrenia SNP heritability. These were likely pyramidal cells, but the small numbers of cells sequenced precluded further exploration. No significant enrichments were found in another snRNA-seq study of a single human[30], perhaps due to a lack of cellular diversity (data not shown). We are unaware of scRNA-seq and snRNA-seq data from human striatum. The specificity of the human cortical signal for schizophrenia was confirmed in relation to the same set of brain-specific GWAS (Fig. 2d). In summary, all of the major findings from the KI dataset were replicated in independent mouse or human studies.

Cell-type enrichments of schizophrenia-associated gene sets. A major question in the field regards interpretation of the large and diverse gene sets that have been compellingly related to schizophrenia (Supplementary Table 1). These gene sets are highly significant, replicate well and have often been implicated in both common- and rare-variant studies. However, their implications for an experimentalist are unclear: what do these large sets of genes really tell us? These gene sets are large and could be expected to recapitulate the cell-type enrichments found above. However, all neurons have synapses, and NeuN (the protein product of *Rbfox3*) is a widely used neuronal marker, so another possibility is that the RBFOX, PSD95 and FMRP gene sets could simply be pan-neuronal.

We thus evaluated whether gene sets previously implicated in schizophrenia were specifically expressed in the KI level 1 brain cell types (using expression-weighted cell-type enrichment, EWCE)[31]. The inputs to EWCE were a list of genes (for example, FRMP-interacting genes or genes intolerant to loss-of-function variation) and the same scRNA-seq cell type specificity matrix used in the MAGMA and LDSC analyses described above. Association with schizophrenia was not a direct input, although these data were incorporated indirectly (it was why a gene set was selected in the first place). However, these effects were subtle. For instance, there was a CLOZUK-significant genome-wide association hit in only 7.0% of genes that interacted with FMRP versus 4.0% that did not interact with FMRP (using MAGMA gene-wise *P* values), and there was a CLOZUK-significant genome-wide association hit in only 4.1% of genes with a probability of being loss-of-function intolerant determined by the Exome Aggregation Consortium (ExAC pLI) > 0.9 versus 3.3% with low pLI. We also determined that overlap between gene sets was relatively low. For ten key gene sets (antipsychotic targets, CELF4, FMRP, high or low d*N*/d*S*, high pLI, NMDAR, PSD, PSD95 and RBFOX), of 45 pairs of correlations (count of intersection or union), only two correlations exceeded 0.25 (RBFOX–CELF4, 0.31; and RBFOX–high pLI, 0.28); most of the other correlations were near 0 (data not shown).

First, pharmacologically defined molecular targets of antipsychotics (the mainstay of treatment for schizophrenia) have been associated with schizophrenia[32], and we found that targets of antipsychotic medication were associated with the same cell types as those for the schizophrenia genome-wide association results: neocortical S1 pyramidal cells, MSNs and hippocampal CA1 pyramidal cells, whereas cortical interneurons were just above the significance threshold (Fig. 4a). Expanding these analyses, we found that other gene sets associated with schizophrenia were specifically expressed in schizophrenia-relevant cell types (Fig. 4b–d). The gene sets that were consistently associated with schizophrenia—intolerant to loss-of-function variation, NMDA receptor complex, postsynaptic density, PSD95 complex, RBFOX binding, CELF4 binding and FMRP-associated genes—all had more specific expression in neocortical S1 and hippocampal CA1 pyramidal cells, MSNs from the dorsal striatum and cortical interneurons (with the exception of NMDA receptor complex genes). Because some of these gene sets are involved in diverse cellular functions, there were, as expected, associations with other level 1 cell types. For example, genes intolerant to loss-of-function variation had significantly greater expression in progenitor cells (dopaminergic neuroblasts, neuroblasts and embryonic GABAergic neurons). Notably, none of the gene sets previously associated with schizophrenia were pan neuronal. A prior study[13] reported that expert-curated glial gene sets were enriched for schizophrenia associations. We confirmed that those gene sets

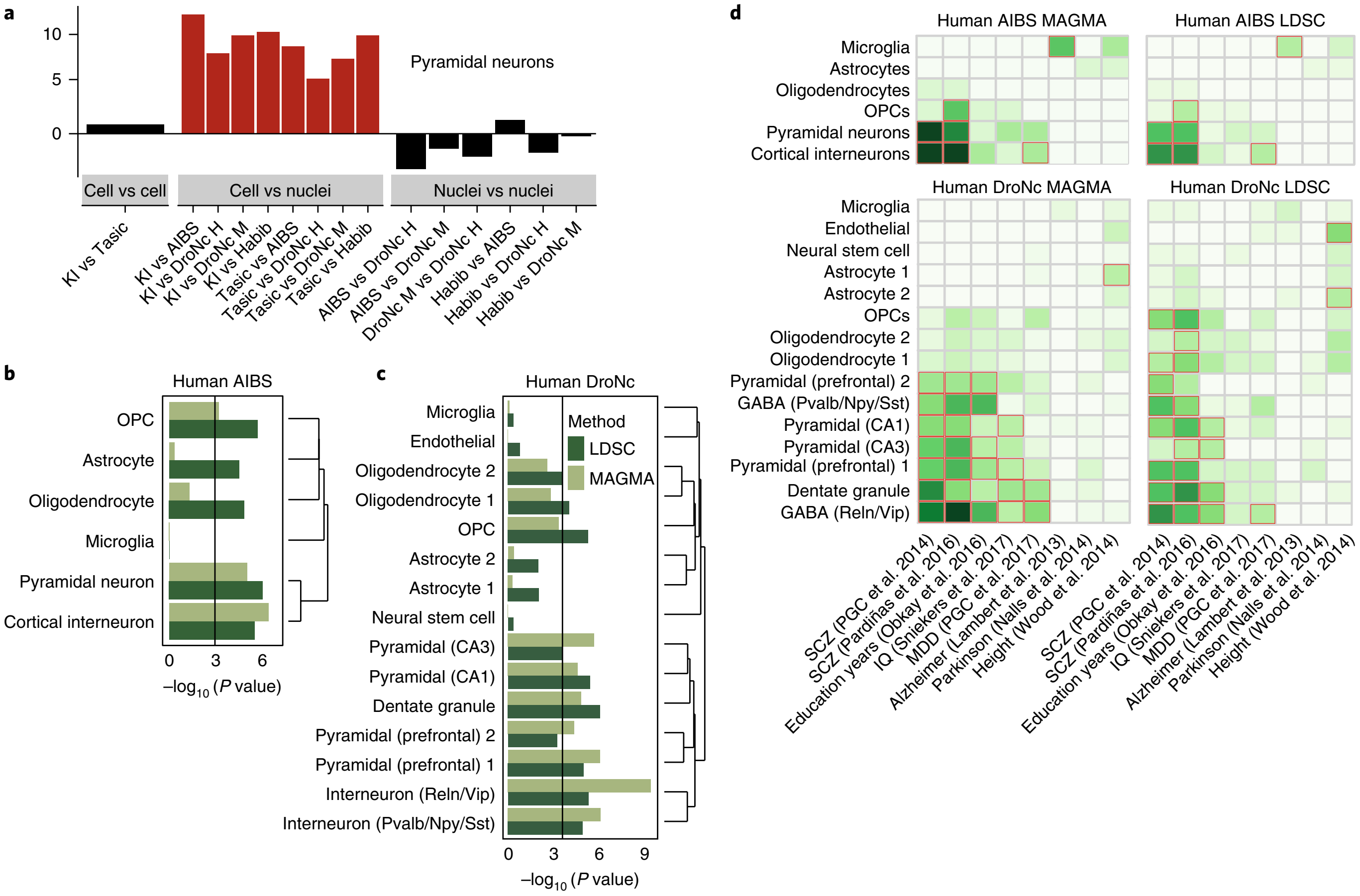

Fig. 3 | Comparison of scRNA-seq and snRNA-seq, and evaluation of enrichment of common-variant CLOZUK schizophrenia genome-wide association results in brain snRNA-seq datasets from adult humans. a, Each bar represents a comparison between two datasets (*X* versus *Y*), with the bootstrapped *z*-scores representing the extent to which dendritically enriched transcripts[28] have lower specificity for pyramidal neurons in dataset *Y* relative to that in dataset *X*. Larger *z*-scores indicate greater depletion of dendritically enriched transcripts, and red bars indicate a statistically significant depletion ($P < 0.05$, by bootstrapping). Supplementary Table 2 describes the studies. **b**, Human mid-temporal cortex brain cell type enrichment. Cortical pyramidal neurons and cortical interneurons show significant enrichment. Oligodendrocyte precursors also show enrichment that was not observed in the KI level 1 data. The black line is the Bonferroni significance threshold ($P < 0.05/(6 \times 8$ comparisons)). **c**, Human prefrontal cortex and hippocampus brain cell type enrichments from ref. [26]. These data show enrichment in cortical and hippocampal glutamatergic (i.e., pyramidal and granule) cells. There was also enrichment in cortical interneurons with the highest level in *Reln*- and *Vip*-expressing cells. The black line is the Bonferroni significance threshold ($P < 0.05/(15 \times 8$ comparisons)). **d**, Heat map of enrichment of diverse human GWAS with human mid-temporal cortex (AIBS), and human prefronal cortex and hippocampus (DroNc) level 1 brain cell types using MAGMA and LDSC. The CLOZUK results do not generalize across human diseases. MDD again shows significant enrichments in cortical interneurons. Common-variant genetic associations for Alzheimer's disease were enriched in microglia. Bonferroni-significant results are marked with red borders (same thresholds as in **b** and **c**).

were significantly associated with glia (Supplementary Fig. 12), but we could not replicate the association of these gene sets with schizophrenia using MAGMA. Finally, we observed that gene sets previously associated with schizophrenia were substantially less associated with schizophrenia after controlling for the pyramidal neurons, MSNs and cortical interneurons (Supplementary Fig. 13). Only loss-of-function intolerant, CELF4-binding and RBFOX-binding gene sets remained significant after controlling for the cell-type enrichments. Our findings highlight that non-overlapping subsets of risk-associated genes each point to the same cell types. Indeed, gene set analysis results can be further subdivided according to cell-type-specific expression. Improved methods are thus needed for gene set analysis that explicitly accounts for cell types, particularly given intensive efforts to conduct a census of the cellular complexity of the human body.

Because neurological diseases are generally not genetically correlated with schizophrenia[33], we evaluated the associations of level 1 cell types with gene sets that were associated with neurological diseases. Genes associated with Alzheimer's disease[34,35] and multiple sclerosis[36] were associated with microglia. Risk-associated genes for leukodystrophy[37] were associated with oligodendrocytes (Fig. 4e). We analyzed genes associated with neurological phenotypes from the Human Phenotype Ontology (HPO) and subcellular localization data from the Human Protein Atlas (Supplementary Figs. 14–19) and found that these mostly targeted cell types distinct from those implicated in schizophrenia. For example, the HPO category "neural tube defect" was associated with neural progenitor cells ($P = 0.0002$) and that for "abnormal myelination" was associated with oligodendrocytes ($P < 0.0001$). We analyzed genes with weak or strong conservation between human and mouse (low or high d*N*/d*S* scores) and found that highly conserved genes were specific to some types of neuron (for example, serotonergic), whereas divergent genes were associated to other cell types (for example, hypothalamic glutamatergic). None of the schizophrenia-associated cell types showed unusually weak or strong evolutionary pressure on their coding sequences (Fig. 4f).

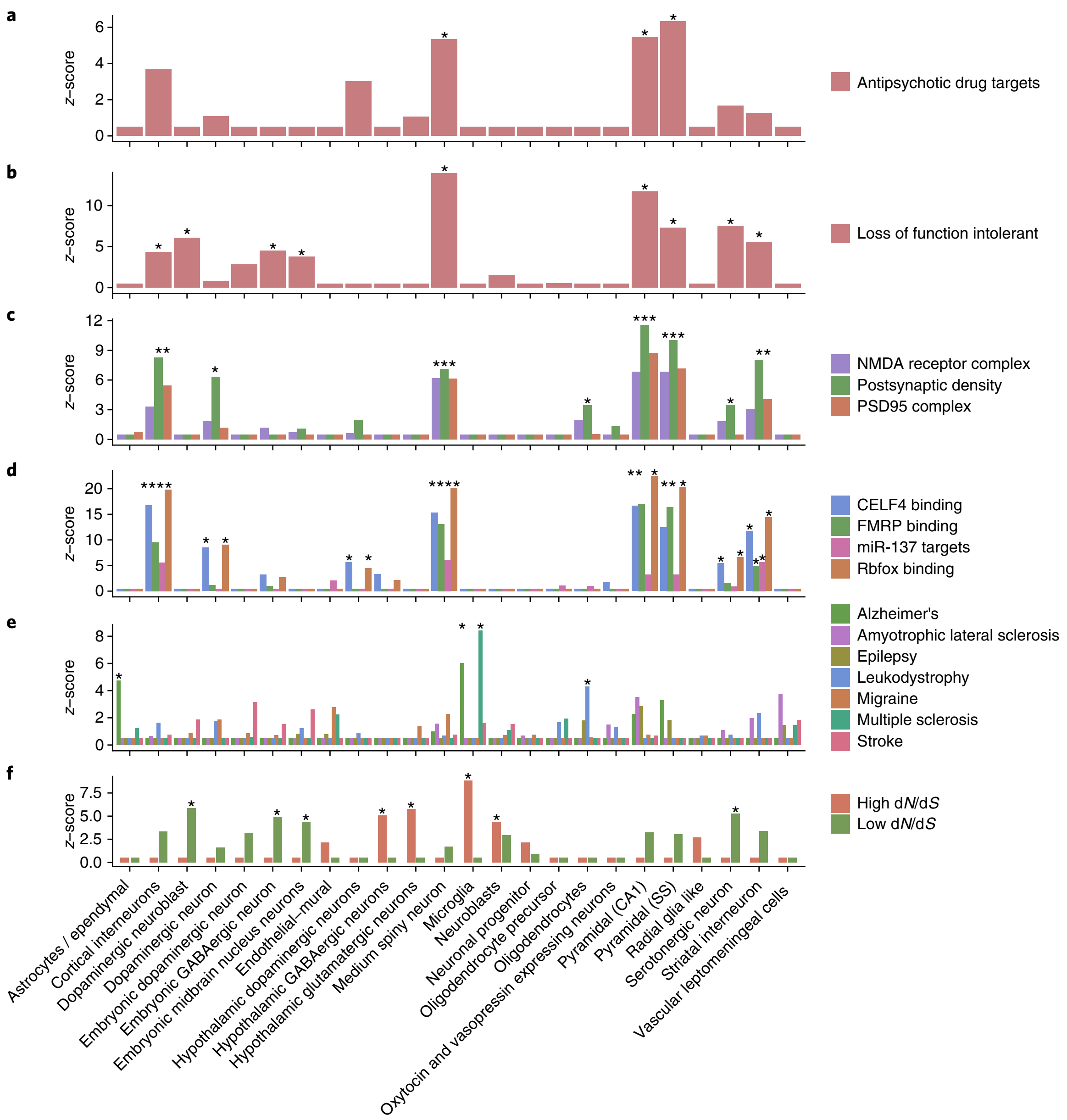

Fig. 4 | Cell-type enrichment of gene sets associated with schizophrenia, neurological disorders and the evolutionary divergence between human and mouse. a, Targets of antipsychotic medications. **b–f**, Gene sets previously shown to be enriched for schizophrenia-SNP heritability, including genes intolerant to loss-of-function variation (**b**), synaptic gene sets (**c**), gene sets mediating DNA or RNA interactions (**d**), gene sets associated with neurological disorders (**e**) and the top 500 genes with lowest or highest d*N*/d*S* ratios between human and mouse (i.e., nonsynonymous-to-synonymous exon changes) (**f**). The level 1 cell types associated with schizophrenia (MSNs, pyramidal CA1, pyramidal SS and cortical interneurons) show enrichment in **a–d**, but neurological diseases do not. Asterisks denote Benjamini–Hochberg-corrected $P<0.05$, as calculated using EWCE.

Independence of genetic association between cell types. Finally, we assessed to what extent cell-type connections to schizophrenia were due to shared gene expression between cell types. For instance, the association of cortical interneurons with schizophrenia was weaker than that for MSNs: we could ask whether these were independent connections to schizophrenia. Alternatively, given that both are GABAergic neurons, we could ask whether both associations were being driven by a common set of genes. We tested this by using resampling without replacement: if the interneuron enrichment was driven solely by overlapping genes with MSNs, then an equivalent level of interneuron association would be found if the schizophrenia association scores of genes within each MSN specificity decile were randomized (Supplementary Fig. 20). We performed 10,000 resamplings for each level 1 cell type while controlling for all four of the significantly associated cell types (Fig. 4a). We found that MSNs, cortical interneurons and hippocampal CA1 pyramidal neurons were independently associated with schizophrenia. However, the association with somatosensory pyramidal neurons was largely due to shared expression with hippocampal CA1 pyramidal neurons. We confirmed this by using conditional analysis (Supplementary Fig. 21a). We then tested whether each cell type remained significant after conditioning on the three other significant cell types together. Notably, only MSNs remained significantly associated with schizophrenia (Supplementary Fig. 21b), indicating that the association of MSNs with schizophrenia is independent from that of pyramidal neurons and cortical interneurons.

To evaluate whether the main sources of enrichment signal in different cell types were from overlapping genes, we used a qualitative

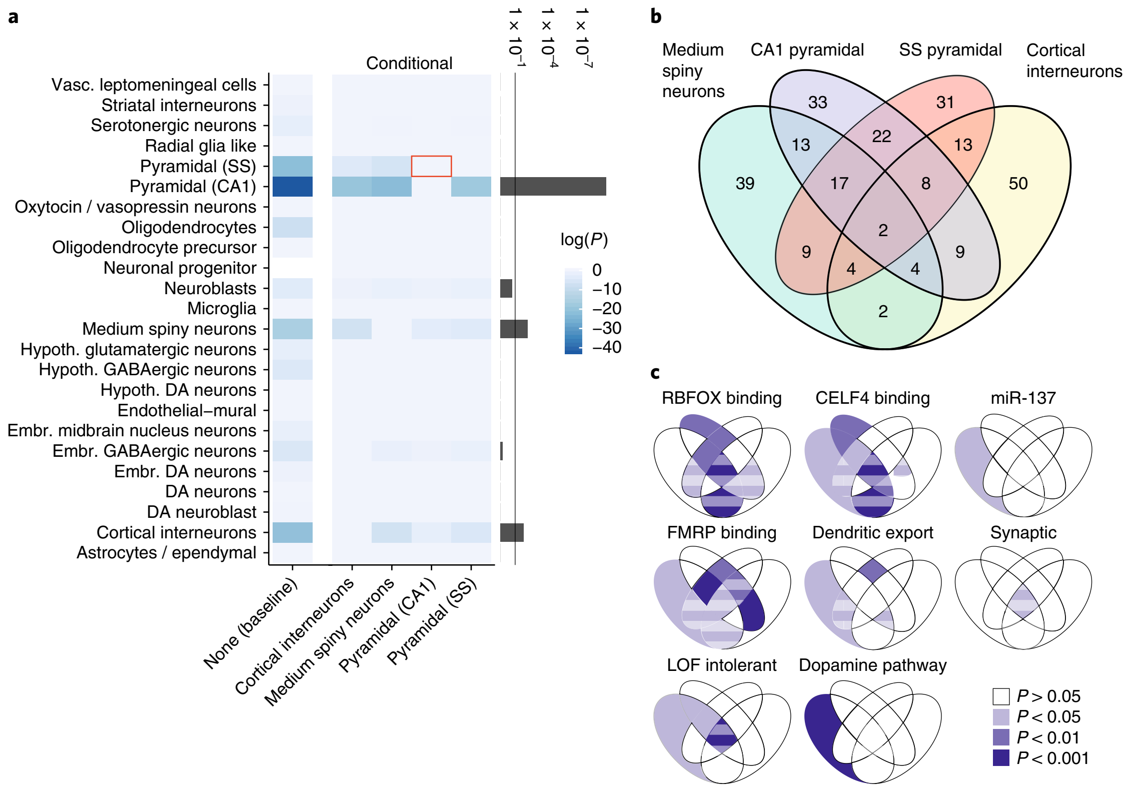

Fig. 5 | CA1 pyramidal neurons, medium spiny neurons and cortical interneurons are independently associated with schizophrenia, and distinct molecular pathways contribute to each cell type. a, Conditional enrichment analysis accounting for correlated gene expression between cell types. The left column shows baseline cell-type enrichment probability values for schizophrenia, calculated by fitting a linear model to specificity deciles against MAGMA gene enrichment *z*-scores. The middle four columns show the enrichment probabilities calculated using bootstrapping to control for correlated expression in other cell types; these probabilities, which approach 0, indicate that, after accounting for expression of the other cell type, there is no enrichment remaining. The red box highlights that there is no longer enrichment in somatosensory pyramidal neurons after accounting for expression in CA1 pyramidal neurons; however, the converse is not true. The bar plot on the right shows the minimum value of the conditional probabilities (excluding self–self comparisons). **b**, Overlap of genes in the schizophrenia-associated cell types. Venn diagram of the top 1,000 schizophrenia-associated genes from the highest-enrichment deciles in the four level 1 cell types. **c**, Benjamini–Hochberg-corrected *P* values for hypergeometric enrichment of genes in **b**. We note enrichment for *Rbfox* in CA1 pyramidal cells, miR-137 targets and dopamine signaling in MSNs, along with shared synaptic genes between pyramidal cells, but separate for GABAergic cells. Areas with striped shading indicate regions with a gene number < 10.

measure. We plotted the overlap of the top 1,000 genes associated with schizophrenia (MAGMA gene-wise *P* values) that were also in the top decile of specificity scores for each of the four main cell types (Fig. 5b). About half of the schizophrenia-associated genes enriched in pyramidal cells and MSNs were shared, but those that conferred risk enrichment in interneurons were, to a larger extent, exclusive. We then evaluated enrichment of gene sets previously associated with schizophrenia (RBFOX-, CELF4- or FMRP-binding genes, loss-of-function intolerant genes, synaptic genes and dendritically transported genes) and genes involved in dopaminergic signaling (Methods) in the different areas of Fig. 5b by using a hypergeometric test. The most associated RBFOX-binding genes were enriched in CA1 pyramidal cells; loss-of-function intolerant genes and genes related to dopamine signaling were specifically enriched in MSNs (Fig. 5c). A subset of synaptic genes associated with schizophrenia was shared by all cell types. These findings show that neuronal classes express a combination of overlapping and non-overlapping functional sets of risk genes.

Discussion

A major issue in schizophrenia genomics is the meaning of the many genome-wide association findings: how do we interpret the hundreds of common-variant associations? Similarly, many sets of genes have been compellingly associated with schizophrenia: what are these diverse functional findings telling us? Thus, we attempted to connect human genomic findings for schizophrenia to specific brain cell types, as defined by their scRNA-seq expression profiles: to what specific brain cell types do the common-variant genetic findings for schizophrenia best fit? Other studies have addressed this question[3,9,14], but by using gene expression data based on aggregates of millions of cells. As described more fully in the Methods ("Rationale"), we used scRNA-seq data to answer this question. We set a high bar—we required that the connections to cell types be identified using two different methods and exceed an appropriately rigorous statistical threshold.

The results were not pan-neural, pan-neuronal or in cell types that were prominent in early development. We found clear connections to just 4 of 24 main brain cell types: MSNs, pyramidal cells in hippocampal CA1, pyramidal cells in the somatosensory cortex and cortical interneurons. Most of the strong results found in the mouse data were replicated in external mouse data and in the more-limited human datasets. Of note, many of the diverse gene sets (for example, antipsychotic drug targets or genes that interacted with the FMRP or RBFOX proteins) that robustly associated with schizophrenia connected to the same cell types. Our results suggest that these discrete cell types are central to the etiology of schizophrenia, and they provide an empirical rationale for deeper investigation of these cell types in regard to the basis of schizophrenia. These results can be used to guide in vivo studies and in vitro modeling (for example, patient-derived neurons from induced pluripotent stem cells) and can provide a basis for analyzing how different risk genes interact to produce the symptoms of schizophrenia.

Our results also suggest that snRNA-seq analysis of neurons leads to systematic underrepresentation of dendritically exported mRNA species. We hypothesize that this is due to destination-specific differences in rates of mRNA decay[38]. Our data on single-nucleus versus single-cell mRNA capture warrants caution when using single-nuclei datasets for the study of neuronal disorders or processes. This fact should be taken into consideration in the design or analysis of future large-scale sequencing efforts.

There are several important caveats, as described more fully in the Methods ("Limitations", including discussion and analyses of gene conservation). Despite our use of multiple statistical methods and efforts to identify and resolve any spurious explanations for our findings, our work has to be considered in light of inevitable limitations. Although the KI scRNA-seq data cover a broad range of brain regions thought to be relevant to the neurobiology of schizophrenia, extensive coverage of cortical and striatal development is lacking at present (gestation, early postnatal or adolescence). The currently available functional genomic data in human brain are limited but improving rapidly via PsychENCODE[39] and similar efforts; however, precisely how schizophrenia GWAS signals impact cell-specific gene expression is not yet a solved problem. Finally, the genetic signals we captured were reflected in the expression levels of hundreds of genes. It is certainly possible for a gene to play an important role in schizophrenia and yet not be in one of the cell types we implicated. For example, genetic polymorphisms in *C4A* appear to be etiologically involved in schizophrenia[7], but the expression of *C4A* is highest in astrocytes, vascular leptomeningeal cells and microglia. We were thus careful with our conclusions: we can implicate a cell type (for example, MSNs show positive evidence), but it is premature to exclude cell types for which we do not have data or those with dissimilar function or under selection pressure between mouse and human.

In sum, our results support a parsimonious hypothesis: the common-variant genome-wide association results for schizophrenia point to a limited set of brain cells, and that subsets of these genes—the gene sets associated with schizophrenia (including antipsychotic medication targets)—each point at the same cell types.

URLs. Expression-weighted cell-type enrichment (EWCE), https://github.com/NathanSkene/EWCE; Linnarsson lab data, http://linnarssonlab.org/data; Mouse Genome Informatics, Jackson Laboratory, http://www.informatics.jax.org/homology.shtml; LDSC, https://github.com/bulik/ldsc/wiki; PGC results, https://www.med.unc.edu/pgc/results-and-downloads; AlzGene database, http://www.alzgene.org/TopResults.asp; GREAT, http://great.stanford.edu/public/html; Hjerling-Leffler lab website, http://www.hjerling-leffler-lab.org/data/scz_singlecell; Human Phenotype Ontology, http://compbio.charite.de/hpoweb; MAGMA_Celltyping, https://github.com/NathanSkene/MAGMA_Celltyping; NMDA Receptor Complex Genes, http://www.genes2cognition.org/db/GeneList/L00000007.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0129-5.

Received: 8 June 2017; Accepted: 3 April 2018;
Published online: 21 May 2018

References

1. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
2. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
4. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
5. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
6. Singh, T. et al. Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
7. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
8. Marshall, C. R. et al. Contribution of copy-number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2016).
9. Finucane, H. K. et al. Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
10. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. Lips, E. S. et al. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Mol. Psychiatry* **17**, 996–1006 (2012).
12. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
13. Goudriaan, A. et al. Specific glial functions contribute to schizophrenia susceptibility. *Schizophr. Bull.* **40**, 925–935 (2014).
14. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
15. Pers, T. H. et al. Comprehensive analysis of schizophrenia-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Hum. Mol. Genet.* **25**, 1247–1254 (2016).
16. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
17. Romanov, R. A. et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).
18. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human and stem cells. *Cell* **167**, 566–580 (2016).
19. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
20. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
21. GTEx Consortium. The Genotype–Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. *Science* **348**, 648–660 (2015).
22. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
23. Gokce, O. et al. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. *Cell Rep.* **16**, 1126–1137 (2016).
24. Habib, N. et al. Div-Seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
25. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single-cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
26. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
27. Abdelmoez, M.N. et al. Correlation of gene expressions between nucleus and cytoplasm reflects single-cell physiology. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2017/10/20/206672 (2017).
28. Cajigas, I. J. et al. The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. *Neuron* **74**, 453–466 (2012).
29. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single-cell level. *Proc. Natl Acad. Sci. USA* **112**, 7285–7290 (2015).
30. Lake, B. B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
31. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single-cell transcriptomes and expression-weighted cell-type enrichment. *Front. Neurosci.* **10**, 16 (2016).
32. Gaspar, H. A. & Breen, G. Drug enrichment and discovery from schizophrenia genome-wide association results: an analysis and visualisation approach. *Sci. Rep.* **7**, 12460 (2017).
33. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* (in the press).
34. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

35. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).
36. Patsopoulos, N. et al. The Multiple Sclerosis Genomic Map: role of peripheral immune cells and resident microglia in susceptibility. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2017/07/13/143933 (2017).
37. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–843 (2015).
38. Burow, D. A. et al. Dynamic regulation of mRNA decay during neural development. *Neural Dev.* **10**, 11 (2015).
39. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).

Acknowledgements

J.H.-L. was funded by the Swedish Research Council (Vetenskapsrådet, award 2014-3863), StratNeuro, the Wellcome Trust (108726/Z/15/Z) and the Swedish Brain Foundation (Hjärnfonden). P.F.S. gratefully acknowledges support from the Swedish Research Council (Vetenskapsrådet, award D0886501). N.G.S. was supported by the Wellcome Trust (108726/Z/15/Z). J.B. was supported by the Swiss National Science Foundation. The PGC has received major funding from the US National Institute of Mental Health (U01 MH109528 and U01 MH109532). H.A.G. was supported by PGC grant 1U01MH109514-01. Primary schizophrenia GWAS data were generated with support from the Medical Research Council (MRC) Centre (MR/L010305/1), program grant G0800509 and project grant MR/L011794/1, and funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 279227 (CRESTAR Consortium).

Author contributions

N.G.S., J.B., P.F.S. and J.H.-L. designed the study and wrote and reviewed the manuscript; N.G.S. performed the LDSC analyses; J.B. performed the MAGMA analyses; T.E.B., R.D.H., J.A.M. and E.S.L. generated the human mid-temporal cortex data; A.B.M.-M., J.R., S.L. and J.H.-L. generated the KI single-cell data; the Major Depressive Disorder Working Group of the PGC performed the MDD GWAS; J.T.R.W., J.J.C., P.G.-R., M.C.O., M.J.O. and A.F.P. performed the schizophrenia CLOZUK GWAS; G.B. and H.A.G. analyzed the antipsychotic drug targets; and all authors read and approved the manuscript.

Competing interests

P.F.S. is on the advisory committee at Lundbeck, is a Scientific Advisory Board member at Pfizer and has received speaker reimbursement and grant funding from Roche. J.H.-L. is a Scientific Advisor at Cartana and has received grant funding from Roche.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0129-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.F.S. or J.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Mouse-to-human gene mapping. We used the expert curated human–mouse homolog list (Mouse Genome Informatics, The Jackson Laboratory; see URLs, version of 11/22/2016). Only genes with a high-confidence, 1:1 mapping were retained. This is discussed further in the Supplementary Note.

Calculation of cell-type expression specificity. A key metric used for our cell-type analyses was the specificity (proportion of expression) for a given gene. This metric was calculated separately for each single-cell dataset. This is a measure of cell-type specificity scaled so that a value of 1 implies that the gene is completely specific to a cell type and a value of 0 implies the gene is not expressed in that cell type. It was calculated using the 'generate.celltype.data' function of the EWCE package (see URLs). See Supplementary Note for further details.

Thresholding of low-expressed transcripts. Because $s_{g,c}$ (specificity for gene g in cell type c) is independent of the overall expression level of a gene, it was desirable to exclude genes with very low or sporadic gene expression levels, as a small number of reads in one cell could falsely make that gene appear to be a highly specific cell marker. Direct thresholding of low-expressed genes was not ideal for performing this, as thresholds need to be set individually for each dataset, and some individual cells can show exceptionally and anomalously high expression of the sporadically expressed gene. We reasoned that all of the genes we wanted to include in the study should be differentially expressed in at least one level 2 cell type included in the study. We thus excluded sporadically expressed genes via analysis of variance (ANOVA) with the level 2 cell-type annotations as groups and excluded all genes with $P > 0.00001$. Gene filtering was performed separately for each single-cell dataset; notably though, the KI dataset was filtered as a merged superset. A consequence of this (and of differences in sample preparation and sequencing) was that different genes were used, for example, in the analysis of the KI superset than were used for the Habib et al. (Mouse Hippocampus Div-Seq) dataset[24]. For datasets for which level 2 cell-type annotations were not available (for example, the Allan Brain Institute human cortex dataset), we used the same approach but with level 1 cell-type annotations instead.

Linkage disequilibrium score regression (LDSC) and partitioning SNP heritability. To partition SNP heritability using LDSC (see URLs)[9], it was necessary to pass LDSC annotation files (one per chromosome) with a row per SNP and a column for each sub-annotation (1 = a SNP is part of that sub-annotation). To map SNPs to genes, we used the dbSNP SNPContigLocusId file (build 147 and hg19/ NCBI Build 37 coordinates). All SNPs not annotated in this file were given a value of 0 in all sub-annotations. Template annotation files obtained from the LDSC Github repository were used as the basis for all cell-type and gene-set annotations ("cell_type_group.1*"). Only SNPs present in the template files were used. If an annotation had no SNPs, then 50 random SNPs from the same chromosome were selected as part of the annotation (if no SNPs are selected, then the software fails to calculate SNP heritability).

Annotation files were created for each cell type for which we applied partitioned LDSC. Twelve sub-annotations were created for each cell type. The first represented all SNPs that mapped onto named regions that were not MGI-annotated genes or that mapped onto a gene which did not have a 1:1 mouse:human homolog. The second contained all SNPs that mapped onto genes not expressed in a cell type. The other ten sub-annotations were associated with genes with increasing levels of expression specificity for that cell type. To assign these, the deciles of $s_{g,c}$ were calculated over all values of g (separately for each value of c) to give ten equal length sets of genes. These were then mapped to SNPs, as described above. To partition SNP heritability among the gene sets (not the cell types), a single set of annotation files was created with each of the gene sets used as a sub-annotation column.

LDSC was then run using associated data files from phase 3 of the 1000 Genomes Project[40]. We computed LD scores for cell-type annotations by using a 1-cM window (–ld-wind-cm 1). As recommended (LDSC Github Wiki, URLs), we restricted the analysis to using Hapmap3 SNPs and, as in the original report[9], excluded the major histocompatibility complex (MHC) region due to its high gene density and exceptional LD. The LDSC 'munge_sumstats.py' script was used to prepare the summary statistics files. The SNP heritability was then partitioned to each sub-annotation. We used the LD weights calculated for HapMap3 SNPs, excluding those for the MHC region, for the regression weights available from the Github page (files in the 'weights_hm3_no_hla' folder).

For the LD score files that were used as independent variables in LD score regression, we used the full baseline model[9] and the annotations described above. We used the '–overlap-annot' argument and the minor-allele frequency files ('1000G_Phase3_frq' folder via the '–frqfile-chr' argument).

Partitioned LDSC computes the proportion of SNP heritability associated with each annotation column while taking into account all other annotations. Based on the proportion of total SNPs in an annotation, LDSC calculates an enrichment score and an associated enrichment P value (one-tailed, as we were only interested in annotations showing enrichments of SNP heritability). All figures showing partitioned LDSC results show P values associated with the enrichment of the most specific decile for each cell type.

Cell-type identification using MAGMA. We used MAGMA (v1.04)[19], a leading program for gene set analysis[41], to evaluate the association of gene-level schizophrenia-association statistics with cell-type-specific expression under the hypothesis that, in relevant cell types, genes with greater cell-type specificity would be more associated with schizophrenia. Gene-level association statistics were obtained using MAGMA (window size 10 kb upstream and 1.5 kb downstream of each gene; see below for discussion of window size), using an approach based on Brown's method[42] (model: 'snpwise-unweighted'). This approach allows one to combine P values in the specified windows surrounding each gene into a gene-level P value, while accounting for LD (computed using the European panel of 1000 Genomes Project phase 3)[40].

The tissue-specific expression metric for each gene in each cell type was obtained by dividing the gene expression level in a particular cell type by the sum of the expression of the gene in all cell types (see $s_{g,c}$, defined above). The distributions of $s_{g,c}$ were complex (point mass at zero expression, substantial right-skewing). For each cell type, we transformed s into 41 bins (0 = not expressed, 1 = below 2.5th percentile, 2 = 2.5–5th percentile, …, 40 = above 97.5th percentile), so that each cell type would be comparable.

MAGMA was then used to test for a positive association (one-sided test) between the binned fractions in each cell type and the gene-level associations (option–gene-covar onesided). For a given mouse or human brain cell type, this tested whether increasing tissue specificity of gene expression was associated with increasing common-variant genetic findings for schizophrenia using information from all of the genes. By default, the linear regression performed by MAGMA is conditioned on the following covariates: gene size, log(gene size), gene density (representing the relative level of LD between SNPs in that gene) and log(gene density). The model also takes into account gene–gene correlations. For the conditional analysis, we used the condition modifier of the 'gene-covar' parameter to condition on each of the significant cell types.

Random permutations of MAGMA. For the analysis in Supplementary Fig. 7, we randomly permuted gene labels of gene-level association statistics of MAGMA and looked for cell-type association with schizophrenia using 1,000 permutations. We observed a mean of 24.8 significant results across cell types at $P < 0.05$, indicating that MAGMA was conservative using our approach (50 significant results expected by chance).

Schizophrenia association using alternative cell-type-specificity method. We tested another recent approach to associate cell types with traits using differentially expressed genes[22]. We computed a normalization factor for each single cell using the scran R package[43] by using the 50% of the genes with mean expression higher than the median. The normalization factors were computed after clustering cells using the scran 'quickcluster' function to account for cell-type heterogeneity. We then performed 24 differential expression analyses using BPSC[44], testing each cell type against the 23 other cell types with the normalization factors as covariate. For each cell type, we then selected the 10% most-upregulated genes and created bed files with the coordinates of these genes extended by 100 kb upstream and 100 kb downstream. SNPs of the baseline model from Finucane et al.[9] that were located in the top 10% of the genes were used to create a cell-type-specific annotation that was added to the 'baseline' model. We then used LDSC[9] to test for association between the cell-type-specific annotations and schizophrenia using a one-sided P value based on the coefficient z-score from the output of LDSC.

Enrichment analyses of gene sets and antipsychotic drug targets. EWCE (see URLs)[31] was used to test for cell types that showed enriched expression of genes associated with particular schizophrenia-associated gene sets. These analyses used the same specificity (s) values for the KI level 1 data that were used for the MAGMA and LDSC analyses. EWCE was run with 10,000 bootstrap samples. Enrichment P values were corrected for multiple testing using the Bonferroni method calculated over all cell types and gene lists tested. EWCE returned a z-score that assessed s.d. from the mean. Values < 0 (depletion of expression) were recoded to 0.

Schizophrenia common-variant association results. The schizophrenia GWAS results were from the CLOZUK and PGC studies[3,20]. CLOZUK is the largest currently obtainable GWAS for schizophrenia (40,675 cases and 64,643 controls), and the authors identified ~150 genome-wide significant loci. It includes the schizophrenia samples from earlier PGC papers. For selected analyses, we also included the PGC schizophrenia results from the *Nature* 2014 report (see URLs)[3]. This paper included 36,989 cases and 113,075 controls and identified 108 loci that were associated with schizophrenia. Results from the published PGC and CLOZUK studies were qualitatively similar, with the CLOZUK data generally showing increased significance owing to its larger sample size.

Comparison of GWAS results for other traits. We included comparisons for a selected set of brain-related traits, as well as height as a negative control. As power to identify cell types is directly proportional to the sample size of the GWAS, we only included traits with at least 20,000 samples that discovered at least 20 genome-wide significant loci. The genome-wide assoication results were from the indicated sources: schizophrenia[3] from the PGC, Alzheimer's disease[34], educational attainment[45], IQ[46], MDD from the PGC (unpublished), Parkinson's disease[47] and height[48].

Test of cell-type association differences between traits. We tested whether the beta coefficient in MAGMA was significantly different between two traits for each cell type, using the approach described in Paternoster et al.[49]. We first computed a z-score for each cell type: $Z = \frac{\beta_1 - \beta_2}{\sqrt{SE\beta_1^2 + SE\beta_2^2}}$, where β_1 and β_2 are the SNP heritability enrichments for traits 1 and 2 (or beta coefficients in MAGMA) and SEβ_1 and SEβ_2 are the s.e.m. values. A two-sided P value was then computed based on the z-score using the R 'pnorm' function.

Gene sets associated with schizophrenia. The gene set results for schizophrenia are summarized in Supplementary Table 1. For CELF4-binding genes, we used genes with iCLIP occupancy > 0.2 from Supplementary Table 4 in ref. [50]. For FMRP-binding genes, we used genes from Supplementary Table 2a in ref. [12]. Genes intolerant to loss-of-function variation were from the Exome Aggregation Consortium (pLI > 0.9)[10]. Genes containing predicted miR-137 target sites were from http://www.microrna.org. NMDA receptor complex genes came from Genes-to-Cognition database entry L00000007[51] (see URLs). The human postsynaptic density gene set was from Supplementary Table 2 in ref. [52]. The PSD95 complex came from Supplementary Table 1 in ref. [53], using all genes marked with a cross in the 'PSD-95 core complex' column. For RBFOX binding, we took all genes with *RBFOX2* count > 4 or summed *RBFOX1* and *RBFOX3* > 12 from Supplementary Table 1 in ref. [54]. For antipsychotic drug targets, we used a gene list provided by G.B. and H.A.G., as reported in ref. [32]. The oligodendrocyte and astrocyte gene lists came from Supplementary Table 4 in ref. [13]. All EWCE P values were corrected with the Benjamini–Hochberg method.

Gene sets for neurological disorders, human phenotype ontology and d*N*/d*S*. For multiple sclerosis, we used results from the largest available GWAS (the Multiple Sclerosis Genomic Map); we used the genes listed in the Supplementary Table of ref. [36]. For Alzheimer's disease, we used the top results from the AlzGene database[35] (see URLs), as well as genome-wide significant genes[34]. For genes associated with leukodystrophy (HP:0002415) we used the Human Phenotype Ontology[37] (see URLs). For amyotrophic lateral sclerosis, we used the top results from the ALSGene database (see URLs). For epilepsy, migraine and stroke, we used the EBI GWAS catalog. For the Human Phenotype Ontology (HPO) gene sets, we downloaded the 'ALL_SOURCES_ALL_FREQUENCIES_phenotype_to_genes.txt' file from build 133. To obtain the genes with the top 500 highest/lowest d*N*/d*S* between humans and mice, we obtained the d*N* and d*S* values through BioMart.

Gene sets associated with subcellular localization. Subcellular localization data were downloaded from the Human Protein Atlas website (HPA, v.17; https://www.proteinatlas.org/)[55]. Only gene lists with > 100 genes were used. Lysosomal genes were downloaded from the Human Lysosome Gene Database[56]. Mitochondrial genes were obtained from Human MitoCarta2.0[57]. Axonal (adult) and axonal (embryonic day E17) were obtained from a study which used axon-TRAP-RiboTags to capture the mRNAs from retinal ganglion cell axons that projected to the superior colliculus (Supplementary Table 1 in ref. [58]). Presynaptic genes came from Supplementary Table 1 in ref. [59]. Synaptic vesicle genes came from Supplementary Table 1 in ref. [60].

Depletion of dendritically enriched transcripts in nuclei datasets. Dendritically enriched transcripts were obtained from Supplementary Table 10 of ref. [28]. This list was produced from pyramidal cells from rat hippocampus, and human 1:1 homologs were obtained. We refer to this set of genes as $L_{\text{dendritic}}$. To enable direct comparisons between datasets, all datasets were reduced to contain a common set of six KI level 1 cell types: pyramidal neurons, interneurons, astrocytes, interneurons, microglia and oligodendrocyte precursors. For the KI dataset, we used S1 pyramidal neurons. The specificity metric (denoted as $s_{g,c}$) was recalculated for each dataset by using this reduced set of cell types. Comparisons were then made between datasets (denoted in the graph with the format 'X versus Y'). We denoted the mean pyramidal neuron specificity scores for dendritically enriched genes in dataset X as $\overline{S_{D=X,L_{dendritic},Pyramidal}}$. We then obtained the difference in pyramidal specificity for list L between two datasets as $D_{X,Y,\mathrm{L}} = \overline{S_{D=X,L,Pyramidal}} - \overline{S_{D=Y,L,Pyramidal}}$. We then calculated values of wenriched gene list, with the genes randomly selected from the background gene set. We denoted the n^{th} random gene list as R_n. The mean and s.d. of the bootstrapped $D_{X,Y,L}$ values were denoted as $\mu_{D_{X,Y,R}}$ and $\sigma_{D_{X,Y,R}}$, respectively. The depletion z-score was then calculated as: $Z_{X,Y,\mathrm{L}_{dendritic}} = \frac{D_{X,Y,L_{dendritic}} - \mu_{D_{X,Y,R}}}{\sigma_{D_{X,Y,R}}}$. A large positive z-score thus indicated that dendritically enriched transcripts were specifically depleted from pyramidal neurons from dataset Y relative to dataset X.

Conditional cell-type enrichments. Gene-association z-scores for schizophrenia were calculated in MAGMA as described above. To enable randomization of the z-scores and recalculation of the associations to be done programmatically, these were then loaded into R, and associations with disease were calculated within this environment without external calls to MAGMA. All genes within the extended MHC region (chromosome 6; 25–34 Mb) were removed due to its confounding effects. We controlled for gene size and gene density by regressing out the effect of the NSNPS and NDENSITY parameters (and the log of each) on the z-score. To ensure that a meaningful number of genes were randomized within each group, associations were calculated over deciles rather than the smaller percentile bins used earlier with MAGMA. Probabilities of association were calculated using the lmFit and ebayes functions from the limma package, to enable rapid computation. We denoted the set of cells studied as c, such that c_i represented the ith cell type. The original z-scores were denoted z, such that z_i was the z-score of the ith gene, whereas the randomized z-scores were denoted as R. The set of genes in the ith specificity decile of the controlled cell type, c_x and the jth specificity decile of target cell type, c_y were denoted $S_{i,j}^{x,y}$ and thus $\bigcup_{k \in C} S_{i,k}^{x,y}$ contained all of the genes in the ith specificity decile of cell type c_x.

The basis of the approach (Supplementary Fig. 23) was to randomize the z-scores with respect to the specificity deciles of the target cell type c_y, but not with respect to the specificity deciles of the controlled cell type c_x. Thus, for each of the deciles indexed by i, we randomly resampled without replacement the z-scores such that $\{R_g\}_{g \in \cup_{k \in C} S_{i,k}^{x,y}} = \{Z_g\}_{g \in \cup_{k \in C} S_{i,k}^{x,y}}$ and yet $R_g \neq Z_g$. In practical terms, this would mean that if we controlled for MSNs and targeted cortical interneurons, then the mean z-score in the 10th MSN decile would remain the same but would be different in cortical interneurons; the question being tested was the degree to which this equated to total randomization in terms of the schizophrenia association found in cortical interneurons.

The baseline association values shown in the leftmost column in Fig. 4a (described as $P_{\text{celltypey,baseline}}$) were calculated using Z. The values of $P_{\text{celltype}y\text{,celltype}x}$ (probability of cell type y being associated with schizophrenia controlling for cell type x) were calculated using intermediate probabilities: 10,000 association P values are calculated for resampled values of R. We selected the 500th lowest of these P values (equivalent to the value that the baseline association probability would need to exceed to be declared independently associated with a probability of 95%) and denoted this $p_{x,y}^{bootstrap}$. The value of $P_{\text{celltype}y\text{,celltype}x}$ was then calculated as $\exp(\log(P_{\text{celltype}y\text{,celltype}x}) - \log(p_{x,y}^{bootstrap}))$. If the value of $P_{\text{celltype}y\text{,celltype}x} > 1$ (indicating that the randomized samples were actually more significantly associated than was found to be the case), then it was set to 1. We were also able to evaluate whether the probability of schizophrenia association in cell type y was greater than would be expected based solely on the expression in cell type x by asking whether the actual association P value was < 95% of the bootstrapped P values. As expected, all self–self comparisons were found to be nonsignificant by this metric (i.e., after accounting for expression in CA1 pyramidal neurons, CA1 pyramidal neurons were no longer significant). In Fig. 4a, a red box was placed around the CA1 pyramidal versus somatosensory pyramidal square because this was the only comparison that involved the four significantly associated cell types in which controlling for expression of a different cell type abolished the enrichment.

Venn diagram enrichments. The Venn diagram shown in Fig. 5 was generated by selecting the top 1,000 genes most associated with schizophrenia based on the MAGMA gene-specific z-scores. All genes within the extended MHC region (chromosome 6; 25–34 Mb) were dropped from the analysis. We controlled for gene size and gene density by regressing out the effect of the NSNPS and NDENSITY parameters (and the log of each) on the z-score. We then took the intersection of the top 1,000 genes with the top decile for each of the four significantly associated level 1 cell types and generated the Venn diagram using the R 'VennDiagram' package. The dopamine gene set included all genes associated with any of the following Gene Ontology (GO) terms: GO:0090494 ("dopamine uptake"), GO:0090493 ("catecholamine uptake"), GO:0051584 ("regulation of dopamine uptake involved in synaptic transmission"), GO:0032225 ("regulation of synaptic transmission, dopaminergic"), GO:0001963 ("synaptic transmission, dopaminergic") and GO:0015872 ("dopamine transport"). The synaptic gene list comprised a combination of three published gene lists: the human postsynaptic density[52]; presynaptic active vesicle docking sites[59] and synaptic vesicle genes[60]. For the presynaptic gene list, the data came from Supplementary Table 1 of ref. [59]; the geneInfo numbers were converted from genInfo accessions to Refseq IDs using Entrez Batch then from Rat RefSeq to HGNC symbols keeping only 1:1 homologs. The synaptic vesicle gene list came from Supplementary Table 1 of ref. [60] and were converted from Rat RefSeq to HGNC symbols using only 1:1 homologs. Enrichment probabilities were calculated using a hypergeometric test against a background set of all MGI genes with 1:1 homologs in human (as described above).

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

Code availability. An R package that can be used for running the cell type association analysis can be obtained from https://github.com/NathanSkene/MAGMA_Celltyping.

Data availability. The RNA-seq data used in this report can be obtained from the Hjerling-Leffler lab website (see URLs), and they include the KI scRNA-seq superset, processed versions of the human and mouse snRNA-seq DroNc-seq data, and the Allan Brain Institute human snRNA-seq data. The specificity values for the KI scRNA-seq superset are included in Supplementary Table 4. The dataset has also been made available in the 'MAGMA_Celltyping' R package (see URLs).

References

40. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
41. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
42. Brown, M. B. A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987–992 (1975).
43. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
44. Vu, T. N. et al. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).
45. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
46. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
47. Nalls, M. A. et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
48. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
49. Paternoster, R., Brame, R., Mazerolle, P. & Piquero, A. Using the correct statistical test for the equality of regression coefficients. *Criminology* **36**, 859–866 (1998).
50. Wagnon, J. L. et al. CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet.* **8**, e1003067 (2012).
51. Collins, M. O. et al. Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J. Neurochem.* **97**, 16–23 (2006). (Suppl. 1).
52. Bayés, A. et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* **14**, 19–21 (2011).
53. Fernández, E. et al. Targeted tandem affinity purification of PSD95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Mol. Syst. Biol.* **5**, 269 (2009).
54. Weyn-Vanhentenryck, S. M. et al. HITS–CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).
55. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
56. Brozzi, A., Urbanelli, L., Germain, P. L., Magini, A. & Emiliani, C. hLGDB: a database of human lysosomal genes and their regulation. *Database* **2013**, bat024 (2013).
57. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**, D1251–D1257 (2016). (D1).
58. Shigeoka, T. et al. Dynamic axonal translation in developing and mature visual circuits. *Cell* **166**, 181–192 (2016).
59. Boyken, J. et al. Molecular profiling of synaptic vesicle docking sites reveals novel proteins but few differences between glutamatergic and GABAergic synapses. *Neuron* **78**, 285–297 (2013).
60. Takamori, S. et al. Molecular anatomy of a trafficking organelle. *Cell* **127**, 831–846 (2006).

nature research

Corresponding author(s): Patrick Sullivan & Jens Hjerling-Leffler

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

We used summary statistics from the largest available schizophrenia GWAS. We only used data from other GWAS studies with over 20000 subjects and 10 genome wide significant loci. We used the largest available high quality single cell transcriptome dataset.

2. Data exclusions

Describe any data exclusions.

Glial cells not from cortex were excluded from the main analyses as they are shared across brain regions but may have been clustered differently

3. Replication

Describe whether the experimental findings were reliably reproduced.

We reproduced the core findings across multiple human and mouse single cell transcriptome datasets. We also replicated it using a smaller Schizophrenia GWAS dataset.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not applicable. All GWAS summary statistics and single cell transcriptome data used were associated with independent publications. No other samples / organisms or participants used which could have been subject to randomisation.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable. All GWAS summary statistics and single cell transcriptome data used were associated with independent publications; nonetheless, as these datasets were generated prior to this study being conceived, they were defacto generated in a 'blind' manner.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	LDSC (https://github.com/bulik/ldsc). MAGMA v1.05 (https://ctg.cncr.nl/software/magma) R EWCE (R package): https://github.com/NathanSkene/EWCE MAGMA_Celltyping (R package): available at https://github.com/NathanSkene/magma_celltyping upon publication

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Not applicable. No unique materials were used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable. No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable. No cell lines were used in this study.
b. Describe the method of cell line authentication used.	Not applicable. No cell lines were used in this study.
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable. No cell lines were used in this study.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable. No cell lines were used in this study.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Not applicable. No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Not applicable. No human research participants were used in this study.