Title: Heritability enrichment in open chromatin reveals cortical layer contributions to schizophrenia

Authors: Paul W. Hook[1] and Andrew S. McCallion[1, 2, 3, *]

Affiliations:

[1] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

[2] Department of Comparative and Molecular Pathobiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

[3] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

* To whom correspondence should be addressed: andy@jhmi.edu

One sentence summary: Chromatin regions from discrete mouse cell populations predict schizophrenia functional variants and cells wherein they act.

Abstract: Genome-wide association studies have implicated thousands of non-coding variants across common human phenotypes. However, these studies provide no direct information regarding the cellular context in which disease-associated variants act. We applied stratified linkage disequilibrium score regression to open chromatin signatures from mouse cell populations and evaluated their relationship to 18 phenotypes, emphasizing schizophrenia. We demonstrate signatures from discrete subpopulations of cortical excitatory and inhibitory neurons are significantly enriched for schizophrenia heritability along with retinal cone populations. Enrichment is maximal in features derived from cortical layer V excitatory neurons. These enrichment patterns are iterated across other neuropsychiatric and behavioral phenotypes. We use these data to predict functional variants at 92/144 established schizophrenia loci and further identify the cellular context in which they may modulate risk.

MAIN TEXT

Although genome-wide association studies (GWAS) have implicated thousands of variants in an array of human phenotypes, the cellular context in which these variants act has remained largely unclear. Discernment of disease-relevant cell populations is essential for comprehensive functional investigation of their contribution to risk. Several recent studies have begun to address this question by leveraging stratified linkage disequilibrium (LD) score regression (LDSC-SEG) to partition heritability from GWAS summary statistics to sets of cell-dependent biological signatures in order to identify cell types relevant to disease(*1*, *2*). Predominantly, these studies have focused on transcription data in order to infer the biological and cellular origins of association due to the depth and diversity of available data sets in the public domain.

The resulting studies have inferred a central neuronal origin of schizophrenia and other neuropsychiatric disorders and highlight a diversity of neuronal, glial and immune populations across other common human phenotypes. Although these studies have supported a role for cortical excitatory and inhibitory neurons in schizophrenia risk(*1*, *2*), they rely upon the imposition of variable window sizes around the transcription start sites of genes with cell-dependent expression. As such, they have been limited by the potential to capture only signal driven by common variants within haplotypes extending into selected windows instead of the entire regulatory landscape. Further they largely lack the immediate capacity to use these observations to directly construct hypotheses indicting putative cis-regulatory elements through which identified variants may act.

Here we have applied LDSC-SEG to cell-dependent open chromatin data generated through the assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq) from *ex vivo* cell populations isolated from mice. We processed raw read ATAC-seq data from over 25 populations of cells through a common analytical pipeline. Peaks of open chromatin were lifted over from the mouse to the human genome and used for LDSC-SEG in evaluation of two separate schizophrenia GWAS, as well as an additional 17 common human traits including neuropsychiatric, neurological and immune phenotypes.

We report that open chromatin signatures from specific cortical excitatory and inhibitory neuron sub-populations are most enriched for schizophrenia heritability. Further, we demonstrate that contribution to risk may be rank ordered by cortical layer position. In addition, we find enriched schizophrenia heritability in retinal cone populations. These observations are reflected among but distinct to neuropsychiatric and behavioral traits. By contrast, our analyses indict a distinct spectrum of neural, glial and immune subpopulations contributing to a subset of neurological, immune and other common phenotypes. Further, we take a critical next step in obtaining insight into common disease, predicting which variants within 92/144 schizophrenia-associated loci (64%) may be considered candidates for functional testing in specific cellular contexts.

Results

Common pipeline for processing of mouse ATAC-seq data

In order to use regulatory DNA present in specific cell populations to explore the heritability of schizophrenia, we obtained publicly available ATAC-seq data derived from lineage identified cell-types sorted *ex vivo* from mice and from mouse brain single nuclei analyses (Table S1)(*3–9*). In total, we obtained 25 mouse ATAC-seq datasets encompassing subclasses of six broader cell types (Table S1). The cell populations were selected to maximize the range of cell types for analyses while ensuring inclusion of classes with predicted roles in schizophrenia(*10–12*) and facilitating comparison with single-cell RNA-seq populations analyzed in previous heritability studies (Table S2)(*1*, *2*). Among neural populations, we included broadly identified populations such as cortical excitatory neurons ("Excitatory Camk2a(+ve)") as well as layer defined populations such as layer V cortical excitatory neurons ("Excitatory LV"). We include tyrosine hydroxylase positive dopaminergic populations from both forebrain and midbrain. Glial populations are represented by distinct datasets from microglia, astrocytes and oligodendrocytes. We also included CD4+ and CD8+ T cell populations as a potential negative control; despite prior evidence suggesting schizophrenia enrichment in broad T cell transcriptional profiles, this was not observed for CD4+ or CD8+ T cell chromatin data(*1*). Additionally, we obtained cell population datasets that had not been previously assayed for schizophrenia heritability: retinal cell types ("Rods", "Cones (blue)", "Cones (green)").

To compile comparable open chromatin profiles, all ATAC-seq data were processed in a uniform manner. Sequencing for each cell population was aligned to the mouse genome (mm10), replicates were combined and peaks of open chromatin were called (see Materials and Methods for more details). This resulted in an average of 85,134 peaks called per cell type (range: 21,983 to 183,162; median: 74,446) with profiles derived from the single-cell data having less peaks called in general (Table S3). Finally, the union of all peaks called in all populations resulted in a global set of 629,845 unique peaks (Table S4).

To ensure that the open chromatin profiles reflected the expected cell population identities, read counts for each cell population in the union set of peaks were compared. Principal component analysis (PCA) of counts revealed that the vast majority of variation (68.61%) in the data could be explained by whether the ATAC-seq data was single-cell or bulk (Fig. S1). Additional stepwise quantile normalization and batch correction abolished the variation caused by this technical effect (Fig. S1). By projecting cell population PCA data onto t-stochastic neighbor embedding (t-SNE) space, cells primarily clustered by cell type group (excitatory neurons, inhibitory neurons, etc.) (Fig. S2). Only single-cell data from inhibitory medium spiny neurons ("Inhibitory MSN*") clustered incorrectly in our analyses. Consistent with the original analysis of this data(*3*), they clustered with excitatory neurons. This analysis established that the open chromatin profiles reflected cell-dependent biology.

In order to apply these data in a human context, we lifted all open chromatin profiles for these cell types from the mouse genome (mm10) to the human genome (hg19) (Table S5), successfully mapping ~62-92% of peaks from each population to the human genome (Table S3). These data sets were then ready to be used to partition common disease heritability using LDSC-SEG.

Determination of cell types underlying schizophrenia genetic association

We explored these 25 chromatin signatures genome-wide in an agnostic effort to determine whether one or more cell types or classes of cell types displayed significant enrichment for schizophrenia heritability. To facilitate direct comparison with prior transcription-based analyses, we made use of the large CLOZUK schizophrenia GWAS meta-analysis that identified 145 loci(*13*). Results were analyzed at two levels of significance: within trait (Bonferroni corrected P-value; 0.05/25 = 0.002; -log10(P) = 2.69897) and correcting for 25 cell types and 19 sets of summary statistics tested (Bonferroni corrected P-value; 0.05/(25*19 tests) = 0.00018; -log10(P) = 3.929419).

Of the 25 chromatin data sets, 14, including 12 neuronal data sets, achieve significant "within trait" enrichment (Fig. 1; Table S6). Furthermore, 10 cell populations achieved significance among all traits tested. By contrast, no glial data sets reach significant enrichment. These observations are similar to recent predictions of schizophrenia cellular context based upon the application of LDSC-SEG to transcriptional data(*2*) (Table S2). Further, our analyses largely indict cortical neurons; with open chromatin region (OCR) data from both excitatory and inhibitory populations displaying significant enrichment of schizophrenia heritability (Fig. 1; Table S6). Although the extent to which enrichment is seen in neurons broadly classed as excitatory or inhibitory does not differ markedly, enrichment among these populations is higher in *Gad2*+ inhibitory neurons ("Inhibitory Gad2 (+ve)").

Within cortical excitatory neurons, the availability of data from layer-identified populations allows detection of a clear and progressive increase in the extent of enrichment, progressing from layers II-III and IV, reaching an apex with OCR signatures derived from layer V excitatory neurons, and then diminishing slightly in layer VI (Fig. 1; Table S6). This pattern is mirrored in single-cell OCR data wherein enrichment in layer III/IV/V cortical excitatory neurons ("Excitatory LII-V*") exceeds that for layer VI cortical excitatory neurons ("Excitatory LVI*") and both are significant (Fig. 1, Table S6). Beyond the cortex, significant enrichment for excitatory neurons of the dentate gyrus ("Excitatory DG*") provide similar evidence for the additional contribution of common variation acting within hippocampal excitatory neurons.

Subpopulations of inhibitory neurons also reveal differing levels of enrichment (Fig. 1; Table S6). While none reach the level of enrichment seen in the broadly defined *Gad2*+ GABAergic population, enrichment in parvalbumin positive neurons ("Inhibitory PV") exceeds that seen in *Vip* positive neurons ("Inhibitory VIP"). We detect no enrichment of schizophrenia heritability in *Drd1* positive medium spiny neurons ("Inhibitory MSN*"), in contrast with previous RNA-seq based studies(*2*) (Table S2). Of additional note, we observe that chromatin signatures from both green and blue cones

pass the threshold for significance when evaluating schizophrenia alone (Fig. 1, Table S6). Although this was initially unexpected, it is consistent with observed reduction in light evoked potentials in the cones of patients with schizophrenia(*14–16*).

In an effort to replicate these observations, we similarly explored the summary statistics from the independently performed Psychiatric Genomics Consortium (PGC) schizophrenia GWAS(*17*). The resulting data closely mirrored our results for the CLOZUK GWAS data, with the same 14 cell populations showing significant enrichment for schizophrenia heritability (Fig. S3; Table S6). The top two cell populations remain the same (layer V excitatory neurons followed by layer IV excitatory neurons), however the rank order after those differ slightly from the CLOZUK analysis (Fig. S3; Table S6).

In summary, we find that genome-wide open chromatin from excitatory neurons, inhibitory neurons, and retinal cone populations show significant enrichment for schizophrenia heritability. The most enriched populations are layer V and layer IV cortical excitatory neurons and this is consistent across independently performed genetic studies. Our data further suggest that the these OCR datasets may be applied more broadly to explore heritability enrichment across other common phenotypes.

Analysis of 17 complex traits across multiple cell types

We then evaluated whether these chromatin data could inform our understanding of the cell types in which variation contributes to common phenotypes. We iterated analyses in these 25 datasets across 17 other common phenotypes, including neuropsychiatric, neurological, immunological and behavioral traits(*18*) (Table S7).

Among neuropsychiatric phenotypes, major depressive disorder (MDD) and bipolar disorder (BD) echo our observations for schizophrenia (Fig. 2A, Table S6). MDD reveals similar patterns of enrichment to those observed for both schizophrenia GWAS studies (CLOZUK, PGC). Although with slightly lower magnitude, cortical layer V again displays greatest significance with layers II-III, IV and broadly defined GABAergic neurons all reaching significance (Fig. 2A, Table S6). When we turn attention to BD, we observe significance across all discrete cortical layer chromatin signatures with the top cell population again being layer V excitatory neurons (Fig. 2A, Table S6). Again reflective of our observations in schizophrenia, blue cones are also significant in BD. The close relationship among these observations is consistent with recent reports of high levels of correlation in the genetic architecture of neuropsychiatric disease(*18*) and the cell populations previously implicated by heritability analyses(*1*, *2*). This reported correlation is diminished between schizophrenia and attention deficit hyperactivity disorder (ADHD). Indeed in our analyses, only dopaminergic neurons derived from the forebrain (posterior hypothalamic) reach significance (Fig. 2A, Table S6). This observation, however, is consistent with report of reduced hypothalamic dopamine release associated with inattention in ADHD patients(*19*).

The established correlation amongst the architecture of neuropsychiatric and behavioral traits leads to additional noteworthy observations. We detect enrichment in

overlapping neuronal cell populations between schizophrenia, MDD, BD and behavioral phenotypes including education years, intelligence (IQ), and neuroticism (Fig. 2B; Table S6). However, although cortical excitatory neuron populations are all significantly enriched for heritability for all three traits, the most enriched cortical layer population for each trait differs (Fig. 2B; Table S6). At least one inhibitory neuron population is significant in each trait with four of five GABAergic populations reaching significance for education years while IQ and neuroticism are more limited. In addition to these shared patterns with schizophrenia, dopaminergic neuron populations are enriched for heritability in both IQ (midbrain and forebrain) and neuroticism (forebrain) (Fig. 2B, Table S6).

Although the genetic architecture of neuropsychiatric diseases have been shown to correlate, little correlation is seen with neurological or other common phenotypes(*18*). Indeed, we see patterns of cortical neuron enrichment do not extend beyond neuropsychiatric and behavioral phenotypes (Fig. 2C; Table S6). In Alzheimers disease (AD), we detect significant enrichment for chromatin signatures derived from microglia (Fig. 2C; Table S6). In multiple sclerosis (MS), we observe significance for T cell populations (both CD4 and CD8) with microglia OCRs approaching within-trait significance (Fig. 2C, Table S6). No cell populations are enriched for heritability across all epilepsy phenotypes studied (Fig. 2C, Table S6). Very simply, this may reflect an absence of data from the most relevant cellular contexts or that epilepsy GWASs have been, so far, underpowered. Further, for immune traits we see significance for OCR signatures derived from microglia (Crohns disease), CD4+ T cells (Crohns, ulcerative colitis and asthma) and CD8+ T cells (ulcerative colitis and asthma) (Fig. S4A, Table S6). In general, the evaluated neurological and immune disorders implicate different subsets of cells, still consistent with what is understood of their biology(*20-24*).

We also analyzed three other phenotypes: height, body mass index (BMI), and age-related macular degeneration (Fig. S4B; Table S6). As expected, inclusion of height as a phenotype which is unrelated to the brain, immune system, or retina, reveals no enrichment in OCR signature from any cell type studied. Unexpectedly, no cell populations were significantly enriched for macular degeneration, although blue cones were the top cell population. Finally, we included BMI as heritability enrichment for brain cell populations had previously been reported(*1*). In contrast to those studies, however, although the top cell population is layer V cortical neurons, no assayed cell population achieves significant enrichment of heritability for this trait.

Our study design provides the opportunity for increased biological resolution both at the cellular and the sequence level. The use of genome-wide OCR data in this way facilitates the direct development of hypotheses concerning in which cell populations candidate risk variants might mediate their effect.

<u>Designation of cell-type schizophrenia risk by locus</u>

We considered that SNPs found in open chromatin from cell populations enriched for schizophrenia heritability would be of special interest (Fig. 1; Fig. S3; Table S6). LDSC-SEG naturally provides the opportunity to explore SNP data in this way.

Overall, a total of 5,961,159 common SNPs were used as input to the LDSC-SEG framework with 539,143 SNPs (~9%) overlapping with an open chromatin peak in at least one cell population. We see ~45% (243,126/539,143) of overlapping SNPs resided in an open chromatin peak in a single cell population while only 2099 SNPs overlapped with an open chromatin region in all 25 populations studied (2099/539,143; ~0.39%)(Fig. S5A).

In order to explore loci unambiguously associated with schizophrenia, we extracted all common SNPs (minor allele frequency $\geq$ 0.05 in 1000 Genomes European data) in LD blocks (lead SNP and SNPs within $r^2 \geq 0.8$ in 1000 Genomes European data) from the CLOZUK schizophrenia GWAS loci(*13*) (Table S8). After merging with the binary open chromatin matrix from LDSC-SEG, and excluding the complex MHC locus, this included 6110 SNPs across 144 loci (Fig. S5B). Overall, 92/144 loci studied (64%) contained at least one SNP in at least one open chromatin domain with a total of 547 SNPs found in an open chromatin peak (Fig. 3A; Fig. S5B). 17 loci contained a single SNP with overlap while 5 loci contained > 20 SNPs with overlaps (Table S9).

For many of these loci (Fig. 3A), the hypotheses arising from our analyses are clear, consistent with known biology and limited in scope; for others, the ongoing challenge is laid out in the breadth of SNPs highlighted by OCRs across a variety of cell types. We describe a few examples below.

The *FOXG1* locus provides an example of a relatively straightforward emerging hypothesis. This locus is marked by lead SNP rs1191551 with 15 total SNPs in LD. Only one SNP, rs61979156, resides within an OCR in one cell population, cortical inhibitory VIP neurons (Fig. 3B; Fig. S6). This SNP lies within an interval predicted to interact with the *FOXG1* promoter and modulate expression in human cortical chromatin interaction data (10 kb resolution; hg19, chr14:3000000-30010000)(*25*). Furthermore, *FOXG1* plays a vital role in GABAergic neuron biology. Overexpression of *FOXG1* causes an overproduction of GABAergic neurons in organoids derived from probands with autism spectrum disorder and is correlated with increased autism phenotype severity(*26*). This overproduction of inhibitory neurons may lead to an excitatory/inhibitory imbalance, a mechanism postulated to underlie both autism and schizophrenia(*27*). Thus our observation and literature evidence leads to the hypothesis that the schizophrenia risk in the rs1191551 locus is mediated through a regulatory element in *Vip*+ GABAergic neurons whose development is altered by rs61979156, leading to an excitatory/inhibitory imbalance.

Similarly, the *PDE4B* gene locus is tagged by rs12129719 and contains 21 total common SNPs, all contained within an intron of *PDE4B* (Fig. S7; Table S9). Only one SNP, rs12081185, is encompassed by OCRs present in cortical excitatory neurons and cortical layer V excitatory neurons (Fig. 3C; Fig. S7). The established genetic association of the *PDE4B* locus with schizophrenia is supported by its structural

disruption in a familial schizophrenia cohort and its interaction with DISC1(*28*) and *PDE4B* is broadly expressed in the nervous system including cortical layer V neurons(*29*). Our observation, which is supported by literature evidence, leads to the hypothesis that schizophrenia risk in the *PDE4B* locus is mediating its effect in cortical excitatory neurons, specifically those in layer V, through rs12081185.

We recognize that not all hypotheses to emerge from these data are as immediately straightforward. The *CACNA1C* locus (lead SNP rs2007044; 10 total reference SNPs) contains three SNPs that overlap with an OCR (Table S9). Two of these SNPs lie in excitatory neuron OCRs: rs1006737 lies within OCRs identified in cortical excitatory neurons including layers II-IV while rs2239038 lies within OCRs identified only in cortical layers II-III and V (Fig. 3D; Fig. S8). rs1006737 modulates the expression of *CACNA1C* in the brain (cerebral hemisphere and cerebellum; GTEx V7) and the interval encompassing the corresponding risk haplotype has been shown to interact with the *CACNA1C* promoter in cells from prefrontal cortex and other neurons(*30*, *31*). By contrast, the neighboring SNP, rs2238057, resides within an OCR in cortical GABAergic neurons and more specifically in *Vip*+ GABAergic neurons (Fig. 3C; Fig. S8). Although cortical expression of *CACNA1C* predominates in glutamatergic neurons, it is found to be co-expressed with inhibitory neuron markers in a small number of cortical neurons(*32*). These kinds of observations lead to more complicated hypotheses in which different variants within a risk haplotype exert their effects, potentially simultaneously, in different cell populations.

The binary matrix for SNPs in schizophrenia GWAS loci overlapping with open chromatin can be found in Supplemental Table 10. Collectively these data provide a systematic biological rationale through which to prioritize SNPs overlapped by OCRs within indicted cell types across all 92/144 schizophrenia loci for functional testing and inform the cellular context in which they might best be assayed.

Discussion

Despite the capacity of GWAS to inform genetic architecture, connecting this framework of disease to the risk, genesis and progression of disease has remained a stubborn challenge. This challenge is particularly stark in schizophrenia, where association of 145 loci found in a single study implicates thousands of noncoding variants in disease risk without providing a systematic and biologically informed strategy to construct hypotheses regarding the cellular contexts in which they act.

Recently, stratified LD score regression (LDSC-SEG) has been used to predict the cellular origins of disease across a range of neuropsychiatric, neurological and immune disorders, primarily leveraging cell-dependent transcription data(*1*, *2*). Although, the resulting data is promising, it falls short of directly connecting specific variants to effects on putative regulatory elements in specific cellular contexts. In an effort to overcome this limitation, we undertook LDSC-SEG across 25 OCR signatures in 18 common phenotypes, with a particular focus on schizophrenia.

Our data confirms the contribution of cortical and interneuron populations in genetically related neuropsychiatric disorders. In schizophrenia specifically, OCR signatures from discrete cortical (both excitatory and inhibitory) populations are the most enriched for schizophrenia heritability among the spectrum of cell types we evaluate. The depth of the available data among these populations allows us to demonstrate that contribution to risk may be ranked according to cortical layer position. We demonstrate a clear increase in the significance of enriched heritability for schizophrenia increasing from layer II, reaching a maximum at layer V with significance diminishing once more in layer VI neurons. In addition, we implicate previously untested cell populations for schizophrenia heritability, retinal cone cells.

These data define testable hypotheses across schizophrenia, indicting specific GWAS-implicated common variants as contributing to risk through modulating activity within specific cellular contexts. Further, we use them to take a critical next step by being able to biologically and systematically illuminate which variants may be predicted to act in which cellular context. Inherently, these data establish a tableau of testable hypotheses that may be taken off the "shelf" into the lab environment. Regarding schizophrenia, we identify 92/144 disease-associated loci (64%) that may now be considered candidates for functional testing in their specific cellular contexts. This illumination of disease-relevant cell types may prove essential for treatment, as demonstrated recently with treatment of PV inhibitory neurons in a mouse model of schizophrenia(*33*).

These results extend their significance beyond schizophrenia alone; they are also highly informative for and reflective of noncoding variant contributions to bipolar disorder and major depressive disorder, consistent with prior expectation(*18*). Similarly, we predict a contrasting spectrum of cellular contexts to contribute to a distinct collection of evaluated neurological, immune and other common phenotypes. These data, as a whole, take a critical next step in obtaining functional insight to common disease.

Although clearly powerful, the capacity to observe this enrichment remains dependent on the availability or generation of datasets of corresponding biological relevance – cell type, developmental stage or physiological state. The fact that these observations are facilitated by datasets generated in mice only serves to expand the potential application of this approach reinforcing the use of mice as a lens by which to study the genetics underlying common human phenotypes(*34*). It establishes the feasibility of generating lineage identified ATAC-seq data where access in humans is more challenging. It makes available an almost limitless collection of cell populations with potential disease relevance; across the spectrum of developmental stages; in the presence or absence of genetic, chemical, and behavioral perturbation. Our data is consistent with recently published work that similarly queried the application of single-cell ATAC-seq data(*35*). Here we demonstrate that, although useful, single-cell acquisition is not necessary to achieve the cell layer-based resolution resulting from this study. Furthermore, the sparse nature of chromatin data obtained from individual cells necessitates sequencing of from large numbers of cells(*35*) or additional

information from RNA to optimize cell identification(*36*). Even with multiple levels of information, some cell populations delineated through single-cell assays cannot be reliably identified(*3*, *35*, *36*). Bulk ATAC-seq, at this time, may thus prove more immediately feasible and flexible than single-cell assays.

While we generate results that support previous studies, there are limitations to the breadth and depth of data we have assembled. Not least of which, we cannot come to any conclusions about the disease relevance of any cell-type that was not tested here. Additionally, open chromatin profiles lack the biological interpretation provided by histone marks when trying to identify functional regulatory DNA. We expect these issues to be solved progressively by increasing the resolution, quality and completeness of chromatin and histone data in tandem with decreasing the cell numbers needed to obtain high quality data.

Further, our results rely on the lifting over of mouse data to the human genome. While the vast majority of mouse peaks identified have a human syntenic ortholog, the entire landscape of regulatory DNA present in these cell populations in humans cannot be fully queried. Although the extent to which this has the potential to limit immediate progress is unclear, initial efforts to obtain with single-cell data from human brain samples make future human datasets a possibility(*37*).

Finally, our analysis focuses on SNPs that are common (minor allele frequency $\geq$ 5%) due to the underlying model used in LDSC-SEG(*1*). We, therefore, cannot come to any conclusions about the contribution of rare variation and other common variation. This may cause these kinds of analyses to exclude true causal variants. We acknowledge that this may be further be compounded by an exclusive focus on SNPs. Other more complex variation has been shown to be important in schizophrenia loci and cannot yet be fully accounted for with this method(*38*, *39*).

The study of common genetic disease is entering a new era, in which genomic data increasingly guides the construction of testable hypotheses for common variant involvement in disease risk genome wide. Although a significant challenge, systematic functional testing of disease associated non-coding variation will be facilitated by the development of biologically informed hypotheses. This study emphasizes the value of strategies to seek cellular context for disease associated and functional non-coding variation as a prelude to massively parallel functional studies in cell types whose selection is truly biologically informed. Our data define a spectrum of immediately testable hypotheses, implicating specific variants as potentially modulating the activity of candidate cis-regulatory elements in discrete cellular contexts across all phenotypes evaluated. Taken collectively the capacity to move directly from GWAS data to design of functional test represents a significant step forward in the dissection of common human phenotypes.

REFERENCES

1. H. K. Finucane *et al.*, Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

2. N. G. Skene *et al.*, Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

3. S. Preissl *et al.*, Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).

4. O. Matcovitch-Natan *et al.*, Microglia development follows a stepwise program to regulate brain homeostasis. *Science*. **353**, aad8670 (2016).

5. L. T. Gray *et al.*, Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *Elife*. **6** (2017), doi:10.7554/eLife.21883.

6. A. Mo *et al.*, Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*. **86**, 1369–1384 (2015).

7. A. E. O. Hughes, J. M. Enright, C. A. Myers, S. Q. Shen, J. C. Corbo, Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. *Sci. Rep.* **7**, 43184 (2017).

8. T. Hosoya *et al.*, Global dynamics of stage-specific transcription factor binding during thymocyte development. *Sci. Rep.* **8**, 5605 (2018).

9. S. A. McClymont, P. W. Hook, A. I. Soto, X. Reed, W. D. Law, Parkinson-associated SNCA enhancer variants revealed by open chromatin in mouse dopamine neurons. *bioRxiv* (2018) (available at https://www.biorxiv.org/content/early/2018/07/08/364257.abstract).

10. O. D. Howes, S. Kapur, The dopamine hypothesis of schizophrenia: version III--the final common pathway. *Schizophr. Bull.* **35**, 549–562 (2009).

11. J. T. Coyle, Glutamate and schizophrenia: beyond the dopamine hypothesis. *Cell. Mol. Neurobiol.* **26**, 365–384 (2006).

12. K. Nakazawa *et al.*, GABAergic interneuron origin of schizophrenia pathophysiology. *Neuropharmacology*. **62**, 1574–1583 (2012).

13. A. F. Pardiñas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

14. R. Warner, J. Laugharne, M. Peet, L. Brown, N. Rogers, Retinal function as a marker for cell membrane omega-3 fatty acid depletion in schizophrenia: a pilot study. *Biol. Psychiatry*. **45**, 1138–1142 (1999).

15. M. Hébert *et al.*, Light evoked potentials measured by electroretinogram may tap into the neurodevelopmental roots of schizophrenia. *Schizophr. Res.* **162**, 294–295 (2015).

16. S. M. Silverstein, R. Rosen, Schizophrenia and the eye. *Schizophr Res Cogn*. **2**, 46–55 (2015).

17. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. **511**, 421–427 (2014).

18. Brainstorm Consortium *et al.*, Analysis of shared heritability in common disorders of the brain. *Science*. **360** (2018), doi:10.1126/science.aap8757.

19. N. D. Volkow *et al.*, Evaluating dopamine reward pathway in ADHD: clinical implications. *JAMA*. **302**, 1084–1091 (2009).

20. H. Sarlus, M. T. Heneka, Microglia in Alzheimer's disease. *J. Clin. Invest.* **127**, 3240–3249 (2017).

21. M. W. Salter, B. Stevens, Microglia emerge as central players in brain disease. *Nat. Med.* **23**, 1018–1027 (2017).

22. C. Baecher-Allan, B. J. Kaskow, H. L. Weiner, Multiple Sclerosis: Mechanisms and Immunotherapy. *Neuron*. **97**, 742–768 (2018).

23. E. O'Loughlin, C. Madore, H. Lassmann, O. Butovsky, Microglial Phenotypes and Functions in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **8** (2018), doi:10.1101/cshperspect.a028993.

24. I. Jelcic *et al.*, Memory B Cells Activate Brain-Homing, Autoreactive CD4+ T Cells in Multiple Sclerosis. *Cell*. **175**, 85–100.e23 (2018).

25. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*. **538**, 523–527 (2016).

26. J. Mariani *et al.*, FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell*. **162**, 375–390 (2015).

27. R. Gao, P. Penzes, Common mechanisms of excitatory and inhibitory imbalance in schizophrenia and autism spectrum disorders. *Curr. Mol. Med.* **15**, 146–167 (2015).

28. J. K. Millar *et al.*, DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science*. **310**, 1187–1191 (2005).

29. J. A. Cherry, R. L. Davis, Cyclic AMP phosphodiesterases are localized in regions of the mouse brain associated with reinforcement, movement, and affect. *J. Comp. Neurol.* **407**, 287–301 (1999).

30. N. Eckart *et al.*, Functional Characterization of Schizophrenia-Associated Variation in CACNA1C. *PLoS One*. **11**, e0157086 (2016).

31. P. Roussos *et al.*, A role for noncoding variation in schizophrenia. *Cell Rep.* **9**, 1417–1429 (2014).

32. N. Dedic *et al.*, Cross-disorder risk gene CACNA1C differentially modulates susceptibility to psychiatric disorders during development and adulthood. *Mol. Psychiatry*. **23**, 533 (2017).

33. T. Marissal *et al.*, Restoring wild-type-like CA1 network dynamics and behavior during adulthood in a mouse model of schizophrenia. *Nat. Neurosci.* (2018), doi:10.1038/s41593-018-0225-y.

34. D. A. Cusanovich *et al.*, A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*. **174**, 1309–1324.e18 (2018).

35. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* (2018), doi:10.1126/science.aau0730.

36. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

37. A. Sekar *et al.*, Schizophrenia risk from complex variation of complement component 4. *Nature*. **530**, 177–183 (2016).

38. J. H. T. Song, C. B. Lowe, D. M. Kingsley, Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am. J. Hum. Genet.* (2018), doi:10.1016/j.ajhg.2018.07.011.

39. P. W. Hook *et al.*, Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. *Am. J. Hum. Genet.* **102**, 427–446 (2018).

40. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

41. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

42. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

43. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

44. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**, 923–930 (2014).

45. K. D. Hansen, R. A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. **13**, 204–216 (2012).

46. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. **28**, 882–883 (2012).

47. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

ACKNOWLEDGEMENTS

We would like to give special thanks to Sebastian Preissl, David U. Gorkin, and Rongxin Fang, for providing the information necessary to process single-nuclei ATAC-seq data. **Funding:** This research undertaken at Johns Hopkins University School of Medicine was supported in part by awards from NIH (NS62972 and MH106522) to ASM. **Author contributions:** PWH and ASM designed the study and wrote the manuscript. PWH implemented the computational algorithms to process the raw data and conduct analyses thereof. PWH and ASM analyzed and interpreted the resulting data. PWH contributed novel computational pipeline development. Correspondence to ASM (andy@jhmi.edu). **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The locations for publicly available ATAC-seq sequencing data can be found in Supplementary Table 1 and are described in the Materials and Methods. Data from analyses will be available prior to publication at the Gene Expression Omnibus (GEO) under the accession number GSEXXXXXX. Documentation of code will be made available prior to publication on GitHub (https://github.com/pwh124/open_chromatin).

Supplementary Materials:

Materials and Methods

URLs

Figures S1-S8

Tables S1-S10

References (*40-47*)

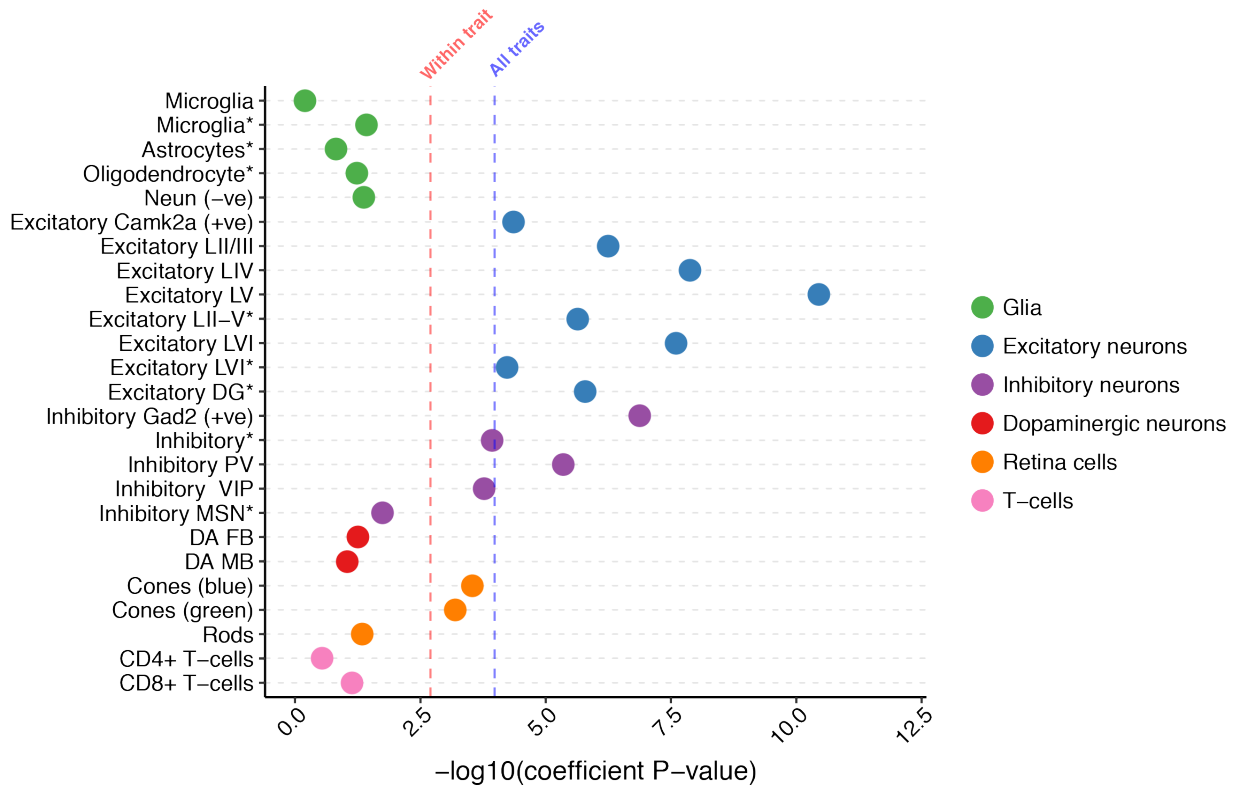


Fig.1. Stratified LDSC results for CLOZUK schizophrenia heritability. Dotplot displaying the -log10(p-values of heritability enrichment coefficient) for each cell population analyzed. Two significance levels for enrichment are shown: within trait (-log10(P) = 2.69897; red) and across traits (-log10(P) = 3.929419; blue). Populations are colored and ordered by broader cell-type category. Asteriks in the cell population name indicate single-nuclei ATAC-seq data. Positive (+ve) or negative (-ve) refers to the gene mentioned in the cell population name. All results can be found in Table S6.

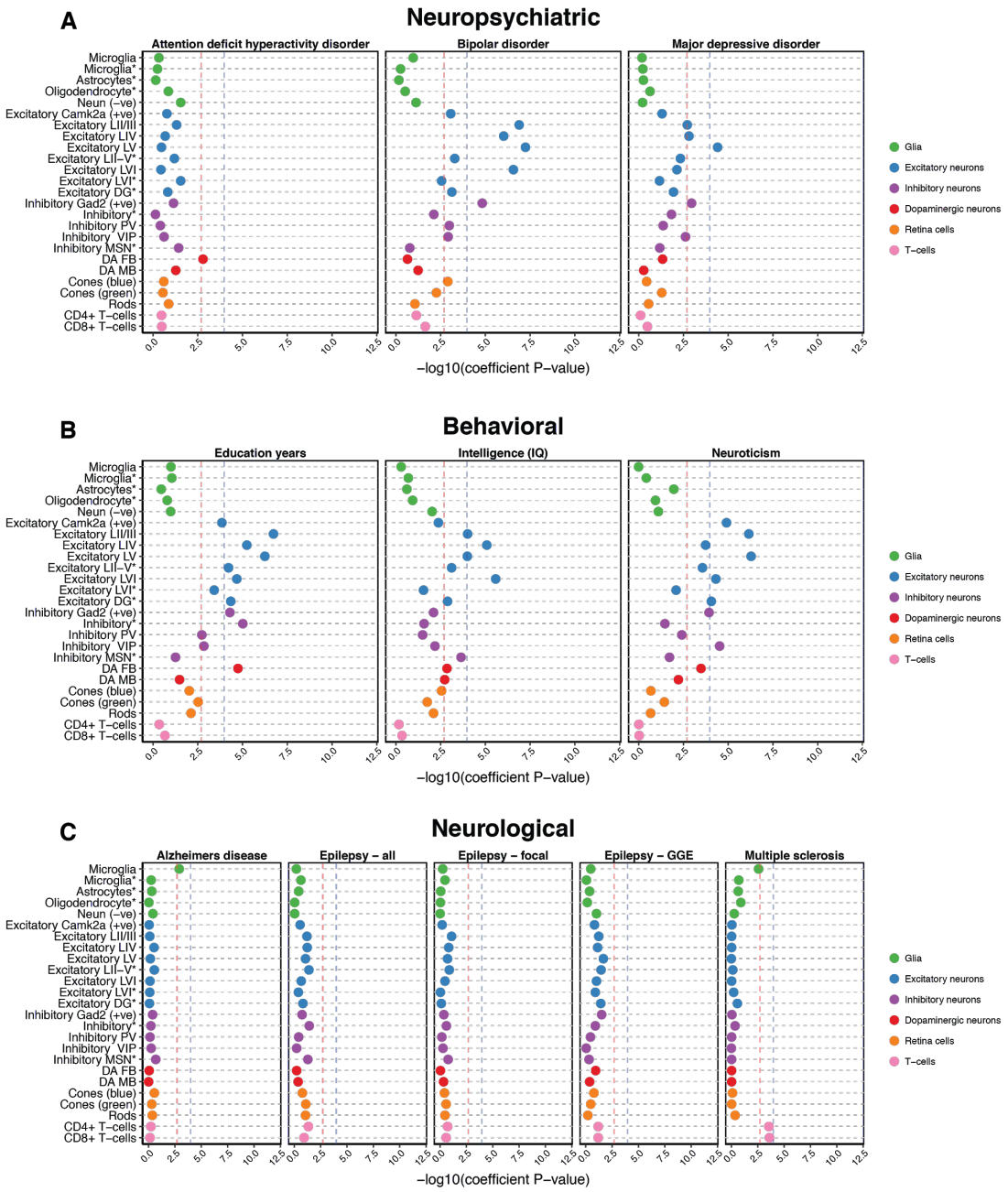


Fig. 2. Stratified LDSC results for select neuropsychiatric, behavioral, and neurological traits. Dotplots for each trait displaying the -log10(p-values of heritability enrichment coefficient) for each cell population analyzed. Two significance levels for enrichment are shown: within trait (-log10(P) = 2.69897; red) and across traits (-log10(P) = 3.929419; blue). Populations are colored and ordered by broader cell-type category. Asteriks in the cell population name indicate single-nuclei ATAC-seq data. Positive (+ve) or negative (-ve) refers to the gene mentioned in the cell population name. All results can be found in Table S6. GGE – genetic generalized epilepsy.

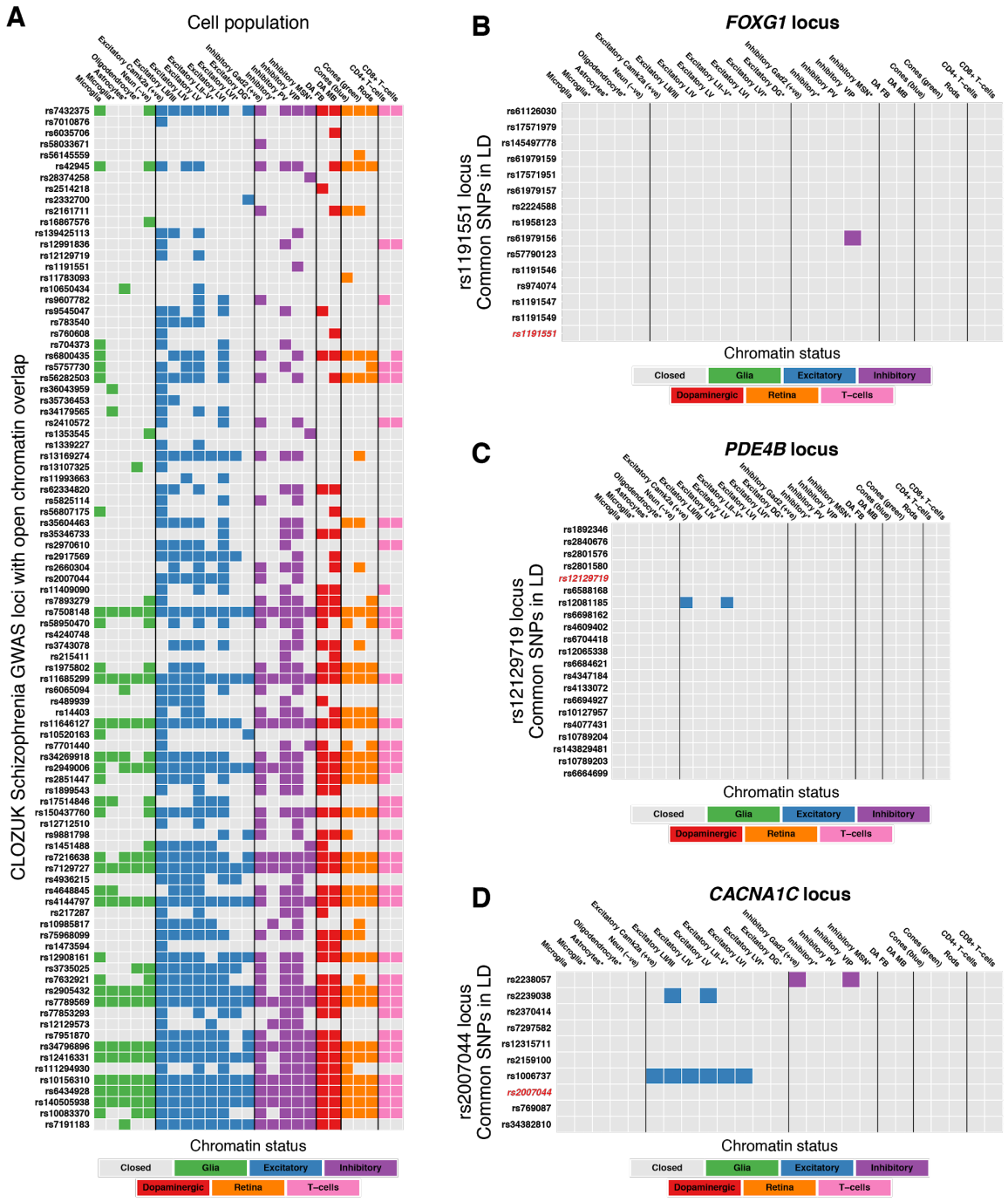


Fig. 3. CLOZUK schizophrenia SNP overlap with open chromatin peaks. A) Heatmap summarizing whether one or more SNPs in each CLOZUK SZ genome-wide significant locus is encompassed by an open chromatin peak in each population. This plot includes only those 92 loci with at least one overlap. Loci are arranged, from top to bottom, in ascending order of number of SNPs in each locus with overlap. Grey indicates that no SNPs in the locus are found in open chromatin whereas cell-type color indicates there was overlap found. Populations are

colored by broader cell-type category. B, C, D) Heatmaps displaying SNP overlap within the LD blocks containing genome-wide significant SNPs rs1191551, rs12129719, and rs200704. SNPs with an $r^2 \geq 0.8$ with the lead SNP, a minor allele frequency ≥ 0.05, and an official reference SNP number are included. Panels are named by genes (*FOXG1*, *PDE4B*, and *CACNA1C*) associated with the loci in the literature. Grey indicates that a SNP is found in open chromatin whereas cell-type color indicates it was not. Lead SNPs in each locus are red and italicized. Images of peaks in loci can be found in Fig. S6.

Supplementary Materials for

Heritability enrichment in open chromatin reveals cortical layer contributions to schizophrenia

Paul W. Hook and Andrew S. McCallion

Correspondence to: andy@jhmi.edu

This PDF file includes:

Materials and Methods
URLs
Figs. S1 to S8
Captions for Tables S1-S10

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S10

Materials and Methods

Obtaining ATAC-seq data

Raw ATAC-seq sequencing data was primarily obtained from the Gene Expression Omnibus (GEO) except for the single-nuclei ATAC-seq data(*3*), which was obtained from the author's website (URLs). All details about the downloaded sequencing data can be found in Table S1.

Additional steps were needed in order to aggregate ATAC-seq reads from individual nuclei into the cell populations to which they belong. In order to extract reads belonging to each cell population, a list of barcode names (cells) and the clusters they belonged to as determined in Preissl, *et*. *al*., were obtained from the authors of the original paper (personal communication). Crucially, these barcodes were included in the name of each sequencing read. Barcodes were grouped according to their cluster identity and paired-end reads belonging to each cell population were extracted via sequencing read name using the BBMap script 'demuxbyname.sh' with the parameters 'substringmode' (URLs). FASTQ files for each barcode in each replicate were then combined into a single FASTQ file for a population. This method had the added advantage of only extracting reads originating from cells that had passed quality control measures in Preissl, *et*. *al*.(*13*).

Alignment and peak calling

Paired-end reads were aligned to the mouse genome (mm10/GRCm38; URLs) using bowtie2 (version 2.2.5; URLs)(*40*) with the following parameters: '-p 15 --local -X2000'. Paired-end reads aligning to the mitochondrial genome as well as random and unknown chromosomes in the reference genome were removed. SAMtools(*41*) was used to remove duplicate reads (v0.1.9), improperly paired reads (v1.3.1), and reads with a mapping quality score of less than or equal to 30 (v1.3.1).

Replicates for each cell population were then merged into a single bam file and peaks were called for each cell population (25 in total) using the MACS2 (v. 2.1.1.20160309)(*42*) 'macs2 callpeak' function with the following parameters: '--nomodel --shift -100 --extsize 200 --keep-dup all --gsize mm'. This resulted in a total of 2,128,339 peaks. Raw MACS2 output can be found on GitHub (URLs).

Relationship between public sets

In order to explore whether or not similar populations of cells have expected similar open chromatin profiles, all peaks called in each population's MACS2 narrowPeak file were merged into a union set of peaks using BEDtools(*43*) 'merge' with the following parameters: '-c 4 -o collapse,count'. Regions that are considered artifacts of ATAC-seq and other chromatin assays in mm10 (so called 'blacklist regions'; see URLs) were removed using BEDtools 'intersect'. This final set contained a total of 629,845 peaks (Table S4).

In order to obtain a count matrix that could allow for cell population comparison, featureCounts from the Subread package was used(*44*). First, the union set of ATAC-seq peaks was converted to an SAF file (see Subread website; URLs) using custom command line scripts. The command 'featureCounts' was used with the '-T 20 -F SAF' parameters in order to obtain a count matrix including read counts for each cell population in each merged open chromatin region. BEDtools 'nuc' was used with a FASTA file of combined mm10 chromosome sequences obtained from the UCSC Genome browser (URLs) in order to calculate GC content for each peak for use in downstream analyses.

The count matrix, the count matrix summary file generated by featureCounts, and the file containing the GC content of the peaks were read into the R statistical environment (URLs). Data was initially transformed into log2(count + 1) counts and analyzed without any prior correction. This first pass principal component analysis (PCA) revealed that the largest source of variation in the counts (PC1; 68.61% variance explained) was explained by the type of ATAC-seq, single-nuclei or bulk (Fig. S1; "log2"). The PC1 variance was decreased, but still clearly separating single-nuclei and bulk ATAC-seq, after performing quantile normalization with the covariates of peak length, peak GC content, and library size using the CQN R package(*45*) (Fig. S1; "log2 + CQN"). Finally, ComBat from the SVA R package(*46*) was used to explicitly correct the CQN normalized counts for type of experiment (single-nuclei or bulk). PC1 variance was again decreased and there was no longer any separation between type of ATAC-seq (Fig. S1; "log2 + CQN + batch"). All PCs (24 in total) were used to generate a t-SNE plot using the tsne (v0.1; URLs) R package function 'tsne' with the parameters 'perplexity = 4, max_iter = 5000, whiten = FALSE' (Fig. S2).

Partitioning heritability with LDSC

All necessary components needed to run LDSC including baseline scores, PLINK files, frequency files, weights, and SNPs, were downloaded from the Broad Institute (URLs). All files were '1000G_Phase3' versions. Additionally, since LDSC was being performed on ATAC-seq data, Roadmap Epigenetic Project LDSC files used as additions to the baseline model in a previous application of LDSC on ATAC-seq data(*1*) were also obtained from the Broad Institute (URLs).

When possible, author-provided GWAS summary statistics were downloaded and processed using the 'munge_sumstats.py' script (LDSC v1.0.0). Otherwise, sumstats files (either pre-processed or not) for Crohns disease, multiple sclerosis, ulcerative colitis, and asthma, were downloaded from the Broad Institute (URLs). The origin and processing pipeline of summary statistics files are described in Supplementary Table 7. Note that the raw, author-provided summary statistics for the PGC schizophrenia GWAS and the CLOZUK schizophrenia GWAS were initially incompatible with 'munge_sumstats.py' and needed minor modifications in order to be processed (see Table S7). Full workflows for processing steps for each set of summary statistics can be found on GitHub (URLs).

In order to perform heritability estimates, mouse peaks (mm10) were lifted over to the human genome (hg19) using the UCSC liftOver utility and the chain file 'mm10ToHg19.over.chain.gz' (URLs) with the default parameter '-minMatch=0.1'. Human regions that are blacklisted either by the ENCODE consortium or ATAC-seq users were removed (URLs). All regions used in this analysis can be found in Table S5. Annotation files needed for analysis were created using the 'make_annot.py' script included in the LDSC software (v1.0.0; URLs).

Cell-type partitioned heritability calculations (also referred to as LDSC-SEG) were performed with the 'ldsc.py' script with the following parameters:

```
--h2-cts {sumstats file path}
--ref-ld-chr {baseline files path}
--ref-ld-chr-cts {cell type test dictionary path}
--w-ld-chr {LD weights file path}
```

The P-values for heritability enrichment are based on a one-sided test for the regression coefficient being greater than 0. This allowed for the direct comparison of the magnitude of enrichment (i.e. higher P-value = higher enrichment). For more information, see Finucane, *et al*., 2015(*47*) and LDSC website (URLs). Partitioned heritability calculations for all diseases were combined and loaded into R for further analysis and creation of plots using custom R scripts (URLs). Two levels of significance were set for LDSC results: Bonferroni correction within a disease (0.05/25 = 0.002; -log10(P) = 2.69897) represented by red dashed lines and Bonferroni correction taking into account all summary statistics tested (0.05/(25*19) = 0.000105; -log10(P) = 3.977723) represented by blue dashed lines.

Determining SNPs in schizophrenia loci contributing to heritability

To create a binary accessibility matrix that included all the SNPs used to partition heritability in LDSC, the annotation matrices used in LDSC were manipulated as follows. All annotation files for each chromosome from the ATAC-seq populations and baseline model were combined together into one matrix. The sum total of ATAC-seq cell populations overlapping each SNP was calculated. All chromosomes were then combined and SNPs with a minor allele frequency $\geq$ 0.05 in 1000 Genomes Phase 3 data were retained because those were the SNPs used to calculate the heritability in LDSC. This resulted in a binary matrix containing 5,961,159 SNPs. In order to make the data easier to work with, only SNPs that were overlapped by open chromatin in at least one ATAC-seq cell population were retained. This resulted in a binary matrix with 539,143 SNPs, ~9% of the original number of SNPs (539143/5961159, ~9%).

In order to focus on SNPs that have strong evidence for contributing to SZ susceptibility, a total of 145 genome-wide significant index SNPs from the CLOZUK SZ study were obtained from Pardinas, *et al*.(*13*). In order to obtain all common SNPs within linkage disequilibrium of the index SNPs, the web-based tool rAggr (URLs) was used with the following parameters:

Genotype database: 1000 Genomes, Phase 3, Oct 2014; hg19
Populations: All European (CEU+FIN+GBR+IBS+TSI)
Min MAF: 0.05
r^2 range: 0.8 to 1.0
Max Distance (kb): 5000
Max # Mendel errors: 1
HW p-value cutoff: 0
Min genotype %: 75

The minor allele frequency of 5% was used in order to match the cutoff used in default LDSC analysis. This method obtained 6536 SNPs in 144 loci and full results can be found in Table S8. Two genome-wide significant loci were initially excluded from the analysis. One locus was initially excluded (the locus containing the SNP rs254782) because it has a minor allele frequency of less than 5% in European populations so no SNPs could be extracted from the European data and none of those SNPs were included in LDSC analysis. The second locus that was excluded was the MHC locus (rs3130820) because that locus is excluded from LDSC analysis(*1*). rs254782 was manually added back in for calculation of overlap statistics. Only common variation with specifically assigned reference SNP ("rs") numbers were used in the analysis in order to match LDSC. This resulted in 6110 SNPs in 144 loci (excluding the MHC locus) being included in the overlap analysis.

The combined LDSC annotation matrix and the SNPs from CLOZUK SZ loci were read-in to the R statistical environment and the locus overlap plots in Fig. 3 were created using custom scripts (URLs).

URLs

GitHub repository: https://github.com/pwh124/open_chromatin
Preissl, *et al*. full data set: http://renlab.sdsc.edu/r3fang/snATAC/
mm10 Bowtie2 index: ftp://ftp.ccb.jhu.edu/pub/data/bowtie2_indexes/mm10.zip
Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Subread FeatureCounts: http://bioinf.wehi.edu.au/featureCounts/
LDSC program components: https://data.broadinstitute.org/alkesgroup/LDSCORE/
Roadmap Epigenetic Project LDSC files:
https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/
Broad Institute LDSC summary statistics files:
https://data.broadinstitute.org/alkesgroup/sumstats_formatted/ and
https://data.broadinstitute.org/alkesgroup/UKBB/
LDSC program: https://github.com/bulik/ldsc
rAggr: http://raggr.usc.edu/
BBMap: https://sourceforge.net/projects/bbmap/
R statistical software, http://www.r-project.org/
tsne R package: https://github.com/jdonaldson/rtsne

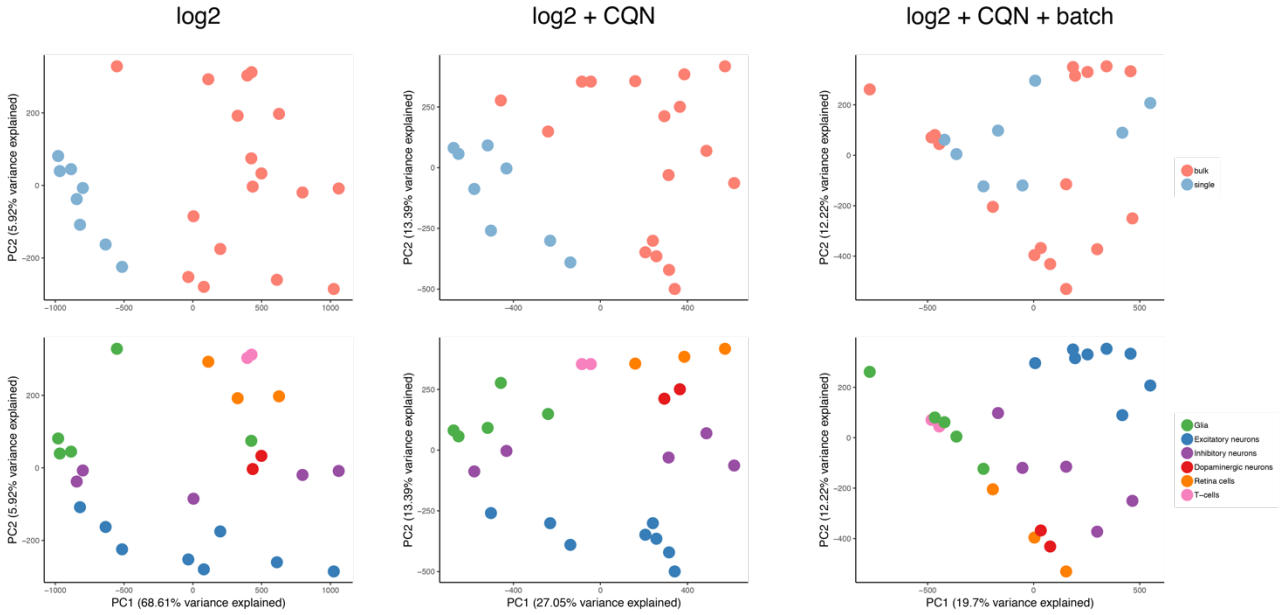
Mouse to human liftOver chain file:
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/liftOver/
mm10 fasta sequence:
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/
ENCODE blacklisted regions:
http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/ (Downloaded May 4, 2018)

mm10: mm10.blacklist.bed.gz
hg19: wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz

ATAC-seq mitochondrial blacklisted regions:
https://sites.google.com/site/atacseqpublic/atac-seq-analysis-methods/mitochondrialblacklists-1 (Downloaded: May 4, 2018)

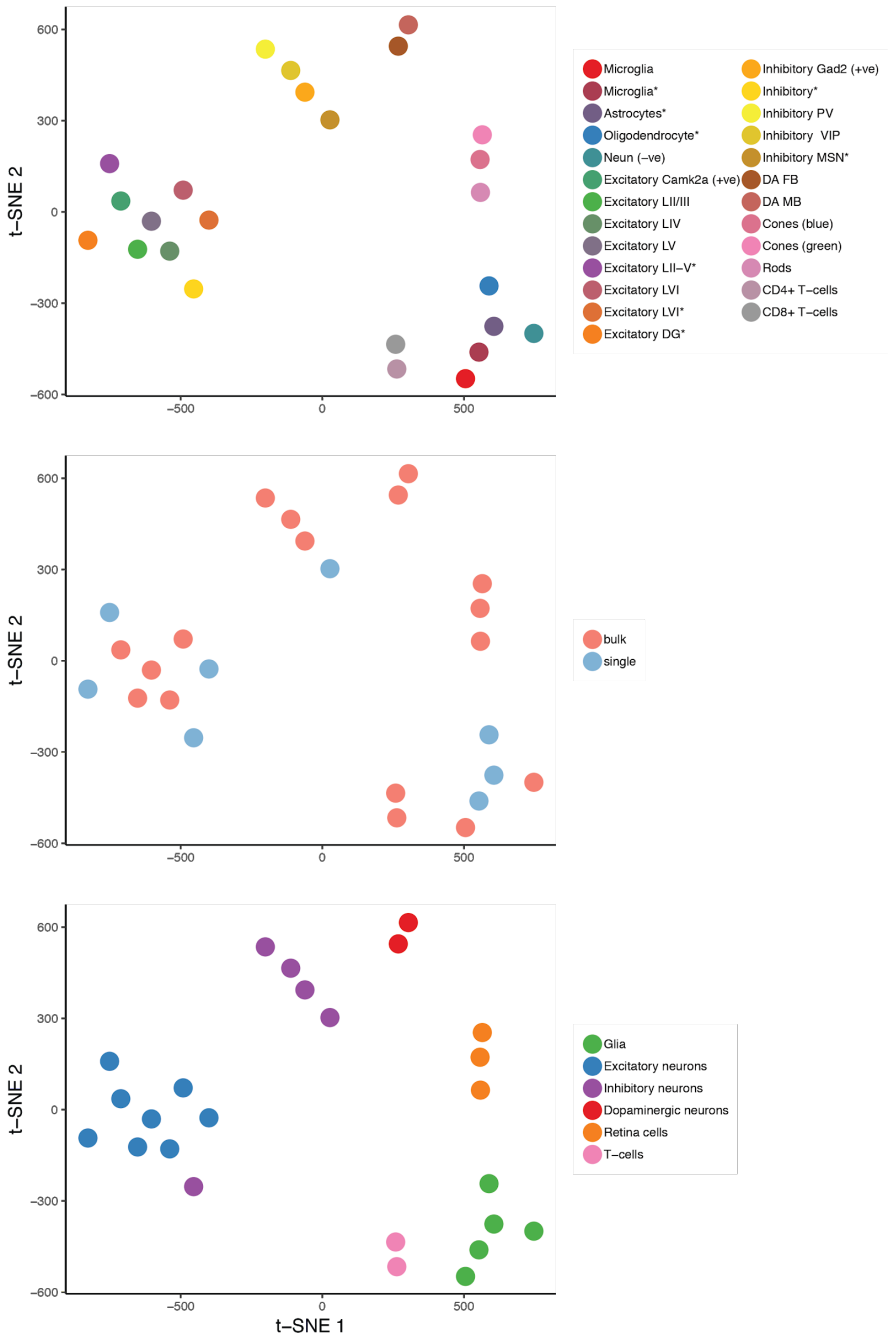
mm10: JDB_BLACKLIST.MM10.BED
hg19: JDB_BLACKLIST.HG19..BED

Fig. S1. Summary of principal component analysis of ATAC-seq cell population peak read counts.



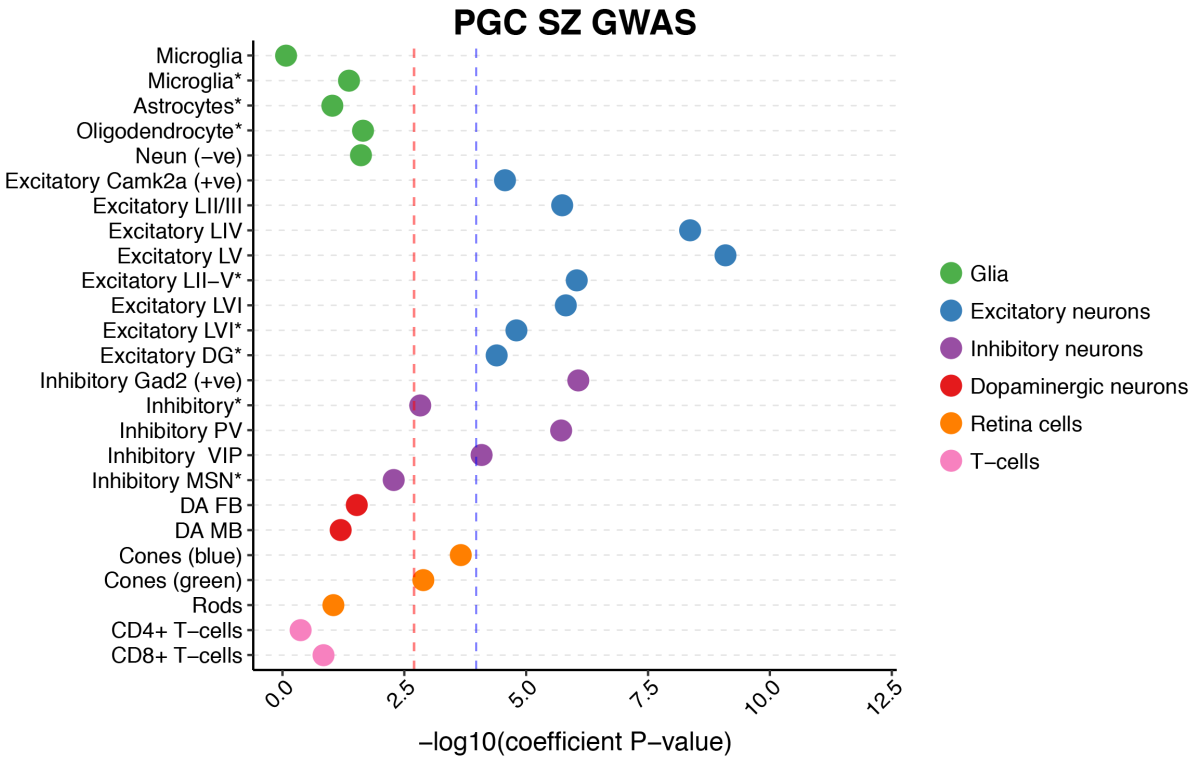
PC1 plotted against PC2 for the progressive analysis of ATAC-seq read counts. Left, PC1 vs. PC2 for log2(counts + 1). Center, PC1 vs. PC2 for CQN quantile normalized log2(counts + 1). Right, PC1 vs. PC2 for CQN quantile normalized and batch corrected log2(counts + 1). Top plots are colored by type of ATAC-seq (bulk or single-cell), and bottom plots are colored by broad cell-type category.

Fig. S2. t-SNE plot of quantile normalized and batch corrected count data.



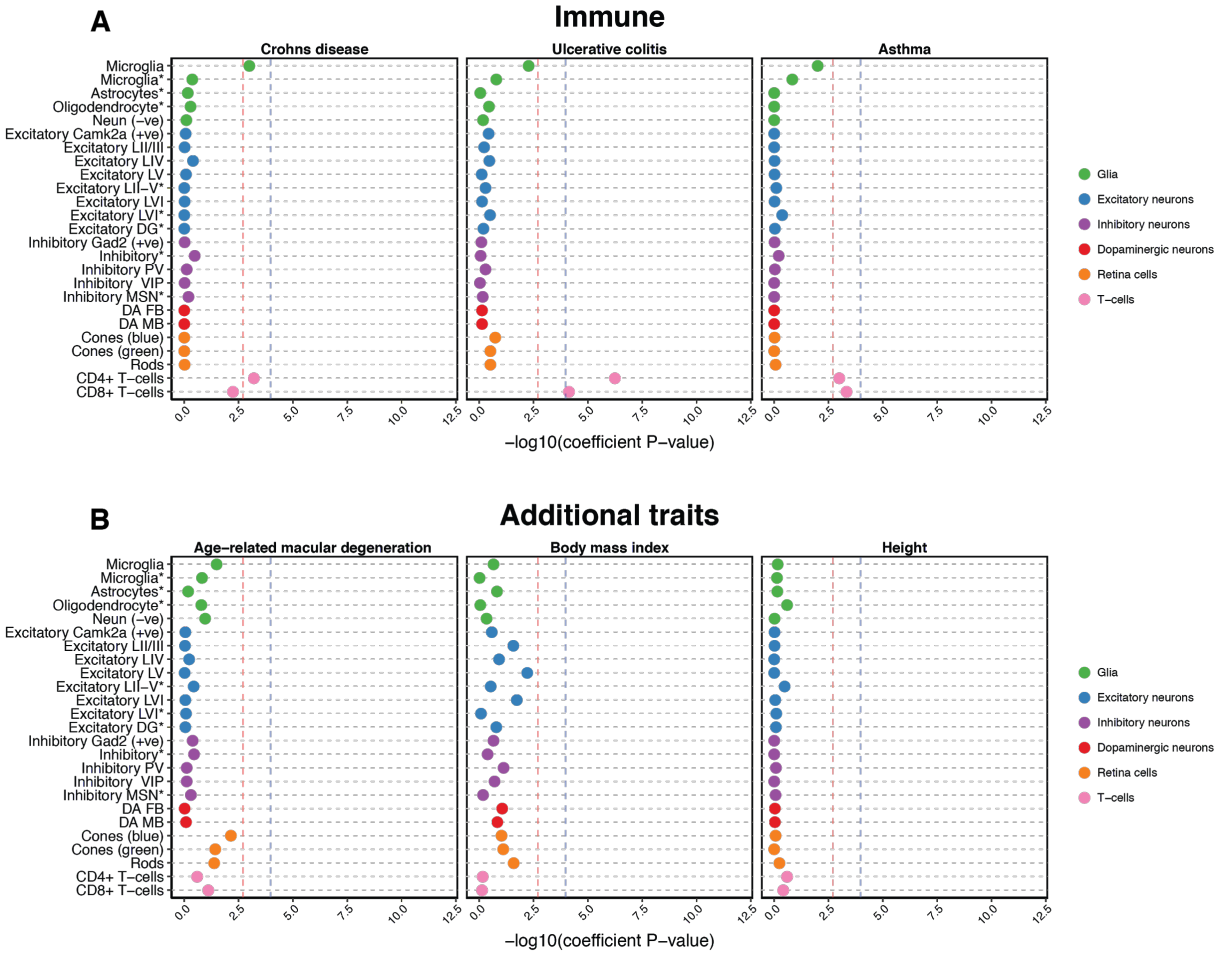
All principal components (n = 24) from the PCA of these counts were used to create the t-SNE plot. Top plot is colored by cell population, central plot is colored by type of ATAC-seq (bulk or single-cell), and bottom plot is colored by broad cell-type category.

Fig. S3. Stratified LDSC results for PGC schizophrenia heritability.



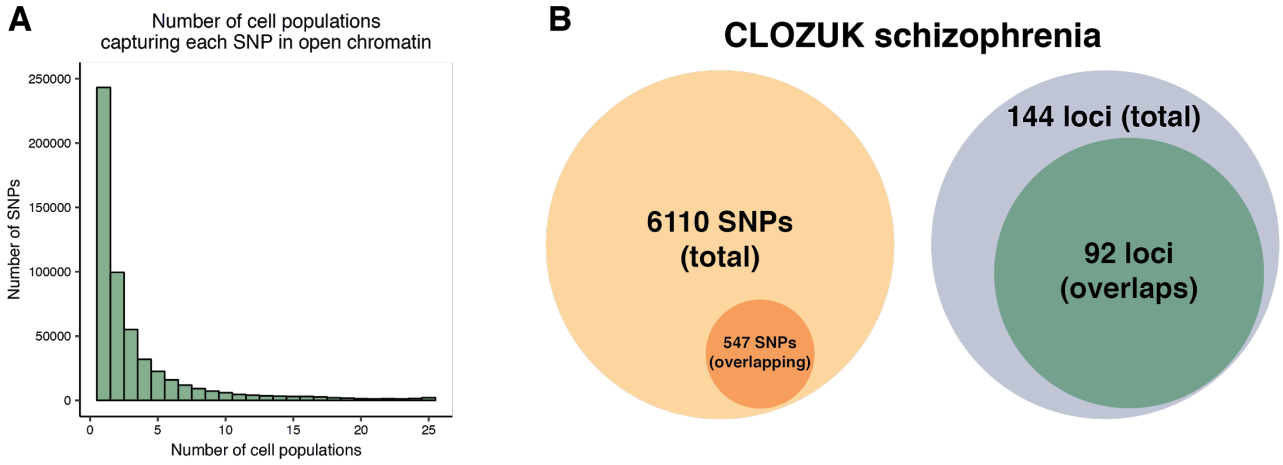
Dotplot displaying the -log10(p-values of heritability enrichment coefficient) for each cell population analyzed. Two significance levels for enrichment are shown: within trait (-log10(P) = 2.69897; red) and across traits (-log10(P) = 3.929419; blue). Populations are colored and ordered by broader cell-type category. Asteriks in the cell population name indicate single-nuclei ATAC-seq data. Positive (+ve) or negative (-ve) refers to the gene mentioned in the cell population name. All results can be found in Table S6.

Fig. S4. Stratified LDSC results for select immune traits and additional traits.



Dotplots for each trait displaying the -log10(p-values of heritability enrichment coefficient) for each cell population analyzed. Two significance levels for enrichment are shown: within trait (-log10(P) = 2.69897; red) and across traits (-log10(P) = 3.929419; blue). Asteriks in the cell population name indicate single-nuclei ATAC-seq data. Populations are colored and ordered by broader cell-type category. Positive (+ve) or negative (-ve) refers to the gene mentioned in the cell population name. All results can be found in Table S6.

Fig. S5. Summary of CLOZUK schizophrenia SNP overlap with cell population open chromatin.



A) Histogram displaying the number of cell populations that capture each SNP in open chromatin. Only SNPs that have overlap in at least one population are plotted. B) Left, diagram displaying the number of SNPs (both lead and those within LD) from CLOZUK schizophrenia genome-wide significant loci that are encompassed by open chromatin in at least one cell population analyzed. Right, diagram displaying the number of CLOZUK schizophrenia genome-wide significant loci containing at least one SNP that is encompassed by open chromatin in at least one cell population analyzed.

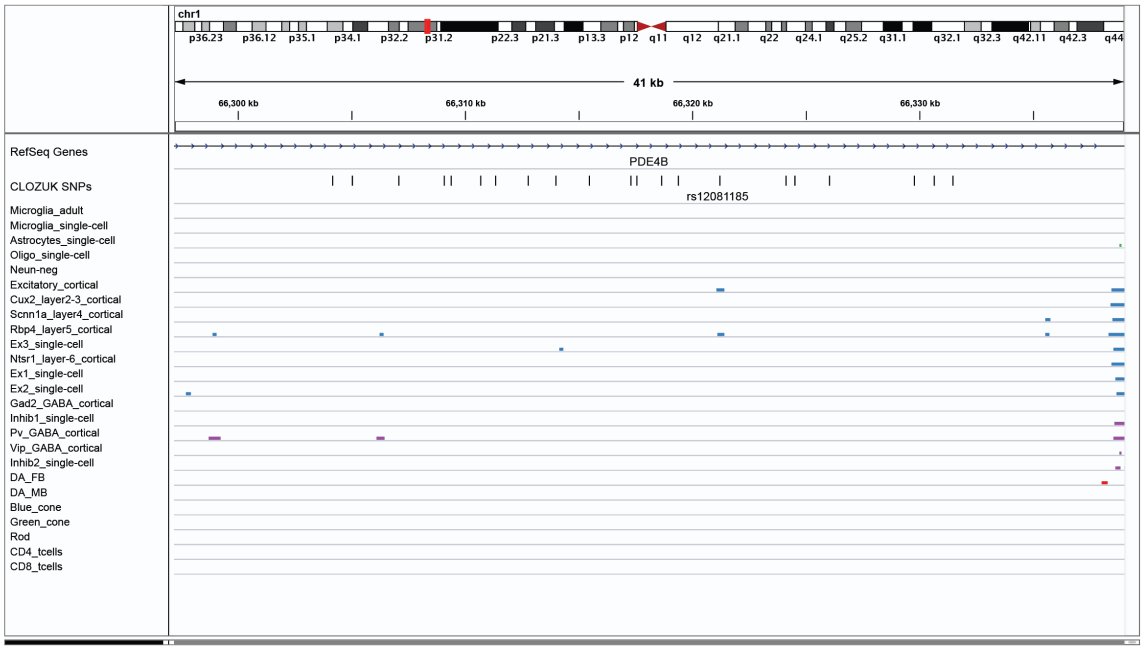
Fig. S6. Integrative Genomics Viewer (IGV) images of rs1191551 CLOZUK schizophrenia locus.

FOXG1 locus
rs1191551 locus
hg19; chr14:29999824-30020902

Refseq genes, CLOZUK loci displaying lead SNPs and all SNPs in LD, and peaks from all 25 cell populations analyzed are displayed. SNPs that overlap with an open chromatin peak have their reference SNP numbers displayed. Corresponding heatmap for this locus is found in Fig. 3B. Populations are labeled by their ATAC-seq file name found in Table S1.

Fig. S7. Integrative Genomics Viewer (IGV) images of rs12129719 CLOZUK schizophrenia locus.

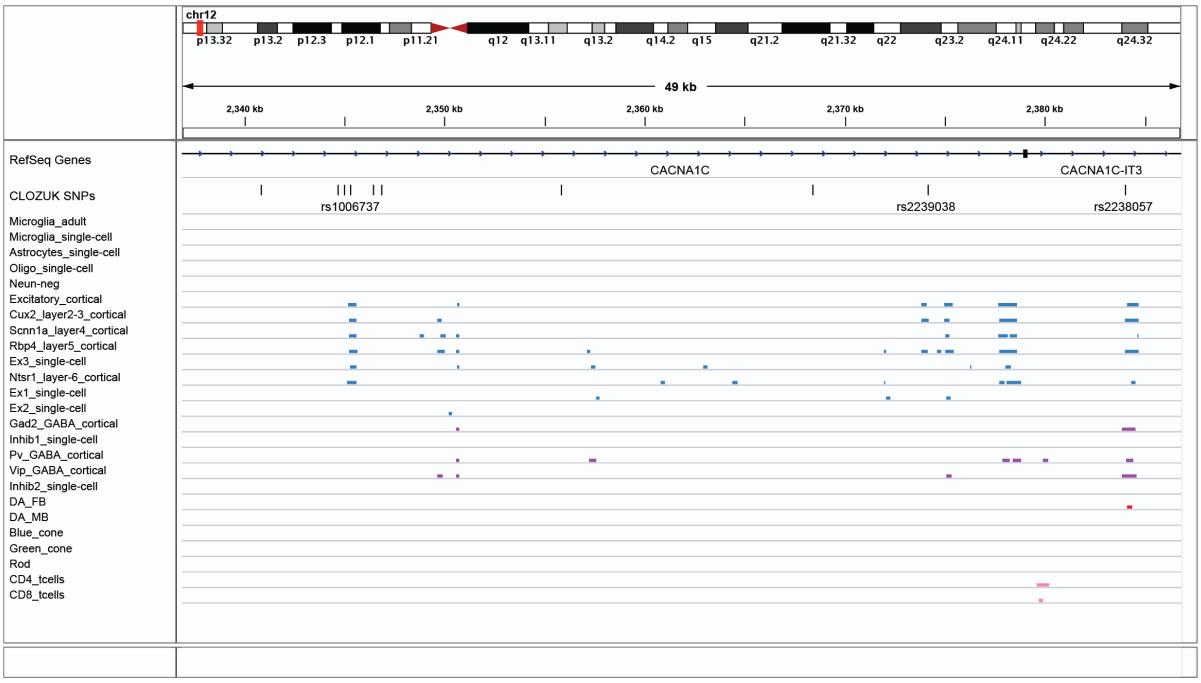
PDE4B locus
rs12129719 locus
hg19; chr1:66297216-66339029



Refseq genes, CLOZUK locus displaying lead SNPs and all SNPs in LD, and peaks from all 25 cell populations analyzed are displayed. SNPs that overlap with an open chromatin peak have their reference SNP numbers displayed. Corresponding heatmap for this locus is found in Fig. 3C. Populations are labeled by their ATAC-seq file name found in Table S1.

Fig. S8. Integrative Genomics Viewer (IGV) images of rs2007044 CLOZUK schizophrenia locus.

CACNA1C locus
rs2007044 locus
hg19; chr12:2336865-2386816



Refseq genes, CLOZUK locus displaying lead SNPs and all SNPs in LD, and peaks from all 25 cell populations analyzed are displayed. SNPs that overlap with an open chromatin peak have their reference SNP numbers displayed. Corresponding heatmap for this locus is found in Fig. 3D. Populations are labeled by their ATAC-seq file name found in Table S1.

Table S1. Description of all publicly available ATAC-seq used in this study. Included is the cell population name used in the paper, the file name used for the cell population data, a description of the cell population, the type of ATAC-seq, a broader cell type classification of cell populations, the Pubmed or bioRxiv ID for the publication of the data, source of the data, and any associated file names.

Table S2. Comparison of ATAC-seq populations to previous populations studied using stratified LDSC and RNA-seq data. Included is the cell population name used in the paper, whether or not the population was found to be enriched for CLOZUK schizophrenia heritability, Pubmed ID for previous transcription based LDSC studies, the population from Skene, 2018 that matches the ATAC-seq population analyzed, the cell type level from Skene, 2018, whether or not the population was significant in Skene, 2018, the population from Finucane, 2018 that matches the ATAC-seq population analyzed, whether or not the population was significant in Finucane, 2018, the figure or table the significance was found in, and any notes about the populations.

Table S3. Summary of peak data. This includes the cell population name, the number of peaks called in the mouse genome ("mm10_peaks"), the number of peaks lifted over to the human genome ("hg19_peaks"), and the percentage of peaks that mapped from mouse to human ("perc_mapped"). Mean, median, minimum, and maximum values for each category are calculated at the bottom of the table.

Table S4. BED file of a merged set of 629845 peaks called across all 25 cell populations. Includes peak coordinates as well as a comma separated list of merged peak names for each peak (4th column) and the number of populations merged for each peak (5th column).

Table S5. BED file of human peaks lifted over from peaks called in each mouse population. Note, peaks are not merged across cell populations. Includes peak coordinates as well as the cell population the peak was lifted over from, both the file name (column 4) and the name of the cell population used in the paper (column 5). The peaks listed were used as annotation input to stratified LDSC. Includes 1743034 peaks.

Table S6. Raw stratified LDSC output for all traits analyzed. Includes cell population name (column 1), the broad category of cell (column 2), the trait analyzed (column 3), stratified LDSC coefficient (column 4), stratified LDSC coefficient standard error (column 5), p-value of coefficient (column 6), -log10(p-value of the coefficient) (column 7), and whether or not the p-value was significant within a trait (column 8) or among all traits analyzed (column 9).

Table S7. Description of GWAS summary statistics used in this study. This includes the trait name the name of the summary statistics, the paper where the summary statistics were generated, the reference of the paper, the PMID or bioRxiv ID for the paper, any notes about

modification of the summary statistics before munging, the original name of the summary statistics file, link or instructions for download for all the summary statistics.

Table S8. rAggr output for CLOZUK schizophrenia lead SNPs. In the table, “SNP1” refers to the lead SNP in each locus and “SNP2” refers to a SNP in LD with the lead SNP. Note that the “SNP2” data also includes data on the lead SNP.

Table S9. Summary of CLOZUK schizophrenia SNP overlap with cell population open chromatin. Includes the lead SNP for each locus (“index.snp”), the total number of SNPs in the CLOZUK locus (“ld.snps”), the number of SNPs that overlap with an open chromatin domain in at least one cell population (“snp.overlap”).

Table S10. The binary matrix for SNPs in schizophrenia GWAS loci overlapping with open chromatin in 25 cell populations. “1” indicates that there was overlap between a SNP and an open chromatin region in the cell type. “0” indicates that there was no overlap. This matrix includes 6111 SNPs from 144 genome-wide significant loci found in Pardinas, *et al*.(*13*).