Manuscript Information

Journal name: Molecular psychiatry
NIHMS ID: NIHMS923166
Manuscript Title: New considerations for hiPSC-based models of neuropsychiatric disorders
Submitter: Nature Publishing Group (repositorynotifs@nature.com)

Manuscript Files

<table>
<thead>
<tr>
<th>Type</th>
<th>Fig/Table #</th>
<th>Filename</th>
<th>Size</th>
<th>Uploaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>manuscript</td>
<td>1</td>
<td>article_1.docx</td>
<td>787438</td>
<td>2017-11-28 06:13:29</td>
</tr>
<tr>
<td>figure</td>
<td>1</td>
<td>figure_1.pdf</td>
<td>196063</td>
<td>2017-11-28 06:13:33</td>
</tr>
<tr>
<td>figure</td>
<td>2</td>
<td>figure_2.pdf</td>
<td>45501</td>
<td>2017-11-28 06:13:33</td>
</tr>
<tr>
<td>figure</td>
<td>3</td>
<td>figure_3.pdf</td>
<td>30931</td>
<td>2017-11-28 06:13:33</td>
</tr>
<tr>
<td>figure</td>
<td>4</td>
<td>figure_4.pdf</td>
<td>176723</td>
<td>2017-11-28 06:13:34</td>
</tr>
</tbody>
</table>

This PDF receipt will only be used as the basis for generating PubMed Central (PMC) documents. PMC documents will be made available for review after conversion. Any corrections that need to be made will be done at that time. No materials will be released to PMC without the approval of an author. Only the PMC documents will appear on PubMed Central -- this PDF Receipt will not appear on PubMed Central.
New considerations for hiPSC-based models of neuropsychiatric disorders

Gabriel E. Hoffman¹,⁴, Nadine Schröde¹,²,⁵, Erin Flaherty²,⁵, Kristen J. Brennand¹,²,³,⁵#

Department of Genetics and Genomics¹, Neuroscience² and Psychiatry³, Icahn Institute of Genomics and Multiscale Biology⁴, Friedman Brain Institute⁵, Icahn School of Medicine at Mount Sinai, New York, NY 10029
#Correspondence: kristen.brennand@mssm.edu

Human induced pluripotent stem cells; psychiatric disease; complex genetic disorders; transcriptomics

The development of human induced pluripotent stem cells (hiPSCs) has made possible patient-specific modeling across the spectrum of human disease. Here we discuss recent advances in psychiatric genomics and post-mortem studies that provide critical insights concerning cell type composition and sample size that should be considered when designing hiPSC-based studies of complex genetic disease. We review recent hiPSC-based models of SZ, in light of our new understanding of critical power limitations in the design of hiPSC-based studies of complex genetic disorders. Three possible solutions are a movement towards genetically stratified cohorts of rare variant patients, application of CRISPR technologies to engineer isogenic neural cells to study the impact of common variants, and integration of advanced genetics and hiPSC-based datasets in future studies. Overall, we emphasize that to advance the reproducibility and relevance of hiPSC-based studies, stem cell biologists must contemplate statistical and biological considerations that are already well accepted in the field of genetics. We conclude with a discussion of the hypothesis of biological convergence of disease - through molecular, cellular, circuit and patient level phenotypes - and how this might emerge through hiPSC-based studies.

I. Introduction.

We are each unique, comprised of distinct genetic, epigenetic and environmental risk factors that predispose us to some diseases and confer resilience to others. As expanding genetic studies increasingly demonstrate that both rare variants of large impact and common variants of small effect contribute to a variety of neuropsychiatric disorders, it becomes increasingly critical that we unravel how these risk factors interact within and between the diverse cell types populating the brain. While mouse models are uniquely suited for demonstrating how altered function of single gene products contribute to aberrant neuronal function or behavior, genetic studies of penetrance and complex gene interactions are extremely difficult to address using inbred mouse lines. Similarly, the challenges of human post-mortem tissue, coupled with the inability to conduct in vivo functional validations, has to date left us with a very limited understanding of how rare and common variants impact gene expression or cellular function in patients. By
developing human induced pluripotent stem cell (hiPSC)-based models for the study of predisposition to neurological disease, stem cell scientists have established a new approach by which to systematically test the impact of causal variants in human cells.\textsuperscript{1-3}

While familial mutations in a fraction of cases result in neurological diseases from autism spectrum disorder (ASD)\textsuperscript{4} to Alzheimer’s disease (AD)\textsuperscript{5}, a large majority of the patient population is designated idiopathic, arising from unknown genetic risk factor(s). Well established for schizophrenia (SZ)\textsuperscript{6} and increasingly accepted across the breadth of neuropsychiatric disease, is the model that a combination of rare and common variants contributes to disease with differing frequencies and penetrance, with highly penetrant rare variants being particularly relevant for diseases with major fitness consequences.\textsuperscript{7}

Although the hiPSC-based models discussed herein are specifically applied to SZ, the tools and insights acquired are being rapidly and concurrently applied to other neuropsychiatric disorders.

II. Advances in psychiatric genomics.

Genetic and environmental effects together contribute to neuropsychiatric disease risk (Figure 1). Genetic epidemiology, including twin studies,\textsuperscript{9} provides substantial evidence that inherited and de novo genetic variants contribute substantially to disease liability (well-reviewed across SZ\textsuperscript{10}, bipolar disorder (BD)\textsuperscript{11}, ASD\textsuperscript{12}, intellectual disability (ID)\textsuperscript{13}, obsessive-compulsive disorder (OCD)\textsuperscript{14}, depression\textsuperscript{15} and attention-deficit hyperactivity disorder (ADHD)\textsuperscript{16}). The current model posits that multiple types of genetic risk influence a continuum of behavioral and developmental traits, the severe tail of which results in neuropsychiatric disease\textsuperscript{17}.

Rare variants associated with SZ currently include sixteen high confidence large copy number variations (CNVs) that are enriched for genes associated with synaptic function, and frequently shared with other neurodevelopmental disorders.\textsuperscript{18} Unlike whole-exome sequencing studies for ASD\textsuperscript{19,20} and intellectual disability\textsuperscript{21} that have identified a series of rare coding mutations enriched for synaptic genes, transcription factors and chromatin modifiers, similar studies for SZ have implied a role for functional gene sets such as voltage-gated calcium channels, ARC-associated scaffold and N-methyl-D-aspartate receptor (NMDAR)\textsuperscript{22-24} but not yet conclusively identified specific genes. Although there is pathway level genetic convergence across neuropsychiatric diseases, it seems that the spatiotemporal activity of the precise genes involved reflects the specific phenotypes; for example, de novo mutations in ID are enriched for fetal cortex genes, ASD for fetal cortex, cerebellum, and striatum genes, and SZ for adolescent cortex genes\textsuperscript{25}. Moreover, while missense mutations in neuronal development genes contribute to ID, ASD or SZ, loss-of-function mutations, particularly in chromatin genes, bias towards ID and ASD\textsuperscript{25}. In summary, although there is substantial convergence at the pathway level of rare mutations across neuropsychiatric disease, clinical presentation reflects the precise gene(s) involved, timing of developmental expression of the affected gene(s) and type of mutation.

Common variation assessed by genome-wide association studies (GWAS) of single nucleotide polymorphisms (SNPs) have identified a growing list of risk loci significantly associated with SZ\textsuperscript{26}, BD\textsuperscript{27} and ASD\textsuperscript{28}, which together account for most of the genetic risk for these disorders. Risk loci identified by GWAS are enriched for neuronal genes\textsuperscript{29,30} and show substantial overlap between these disorders\textsuperscript{27,31}. In SZ, these risk loci map to genes expressed in pyramidal excitatory neurons and a subset of GABAergic interneurons, substantially less to progenitor or glial cells\textsuperscript{29,30}. Transcriptomic profiling of
post-mortem brain tissue supports this shared molecular neuropathology, demonstrating that the degree of shared transcriptional dysregulation strongly correlates to the observed common variant overlap across psychiatric disorders 32.

While disease associated loci still explain only a small fraction of the predicted genetic liability to psychiatric disease, the "missing heritability" is now believed to largely reside amongst the common variants with effect sizes well below genome-wide statistical significance 33. Taken to the fullest extent, a proposed "omnigenic" model suggests that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes 34.

Overall, highly penetrant rare mutations tend to confer risk of neurodevelopmental disorders with earlier onset, while more common variants with much lower effect sizes contribute risk towards later onset disorders (reviewed in 35). Different types of risk factors can interact; a higher polygenic risk score is thought to increase disease risk in carriers of rare mutations, suggesting cumulative effects between both types of variants 36, 37. Although variants differ in penetrance and prevalence across the spectrum of neurodevelopmental disorders, there is a growing consensus that genetic risk will converge not only between psychiatric disorders, but also between common and rare variants within disorders - focused on genes expressed during fetal cortical development 38-40 and converging on common pathways such as synaptic function 18, 22, 23, 41 and epigenetic processes 41, 42. In ASD, a shared functional signal between common and rare variants was detected after accounting for the different effect sizes of the genes involved 43. Critically, this is not a literal overlap in risk genes or pathways; rather, functions weakly associated with common variants tend to show stronger effects in rare variants 43. The role of effect size on these functional consequences, the degree of overlap of genetic risk and precise neural cell type(s) impacted by this growing list of disease-associated mutations remains unclear. Throughout this review we will consider the extent to which in vitro stem cell-derived populations can model the impact of different perturbations on these causal genes on cellular and molecular neural phenotypes. The ultimate goal of psychiatric genetics is to study the impact of patient-specific mutations, facilitating a transition into precision psychiatry 8 and personalized medicine 44.

III. Findings from recent hiPSC-based models of SZ.

Although typically diagnosed in late adolescence, SZ reflects abnormal neurodevelopmental processes that begin years before the onset of symptoms (reviewed 45). We and others have shown that hiPSC-derived neurons most resemble fetal brain tissue 46-48, indicating that these models are most appropriate for studying aspects of disease predisposition rather than the disease state itself. (This may be changing, as recent studies have shown that long-term organoid culture can yield cells more similar to neonatal tissue 49, novel media formulations can improve the maturity 50 and genetic strategies can accelerate the aging 3, 51 of hiPSC-derived neurons, facilitating studies of adult-onset diseases). hiPSC-based models have great potential to help unravel the functional impact of the risk loci identified by genetic studies. Idiopathic cohorts capture all of the risk elements, known and unknown, that contribute to disease in any specific patient, whereas genetic cohorts (whether recruited on the basis of a shared genetic mutation (discussed in more detail in section VI) or engineered on isogenic backgrounds (discussed in section VII) provide the opportunity to test the impact of a defined genetic variant shared by a subset of patients.
Early hiPSC-based studies of SZ focused primarily on idiopathic patient collections, largely reflecting the availability of patient-derived fibroblasts during this period. More recently, there has been a transition to more defined cohorts, in an attempt to reduce the heterogeneity between patients. Despite the obvious limitations in these early studies with respect to limited cohort size and heterogeneous neuronal populations, there was a striking convergence of findings across both idiopathic and genetic cohorts. These shared findings include aberrant migration / cell polarity (idiopathic 46, 22q11.2 deletion 54, CNTNAP2 55, and 15q11.2 microdeletion 56 patients), proliferation (DISC1 57, 58) and WNT signaling (idiopathic 59 and DISC1 58) in hiPSC-derived neural progenitor cells (NPCs). Moreover, three groups, across independently reprogrammed and characterized idiopathic SZ cohorts, reported increased oxidative stress and perturbed responses to environmental stressors 48, 60 in patient-derived NPCs. Moreover, patient-derived neurons exhibit decreased neurite outgrowth (idiopathic 2 and 22q11.2 deletion 54), synaptic maturation (idiopathic 2, 53, DISC1 61, and 15q11.2 microdeletion 56 patients) and activity (idiopathic 62, CNTNAP2 63, and DISC1 61) and altered neurotransmitter release (idiopathic 64, and DISC1 61). Global gene expression studies confirmed aberrant expression of synaptic genes (idiopathic 2, DISC1 61, 22q11.2 deletion 65), queried differential microRNA expression (idiopathic 66 and 22q11.2 deletion 67) and demonstrated blunted activity-dependent changes of gene expression (idiopathic 68). A limited number of studies further explored cell type specific effects, focusing specifically on hippocampal neurons 62, 69 or astrocytes 70. Taken together, these reports mirror findings in postmortem pathological studies (reviewed in 71) and animal models (reviewed in 72) and provide a convincing proof-of-concept demonstration that at least some of the cellular and molecular factors underlying SZ are conserved between hiPSC cohorts, suggesting that, at least in vitro, the myriad genetic mechanisms contributing to SZ predisposition may manifest through a more limited number of cellular outcomes.

It remains unclear to what extent hiPSC-based studies are sufficiently powered to discover the shared molecular mechanism(s) driving phenotypic differences in patient neural cells. Nonetheless, our recent discovery-based approach combining RNA, microRNA and protein analyses found microRNA-9 (miR-9) to be significantly downregulated in a subset of idiopathic SZ NPCs, a finding that was corroborated in a second hiPSC cohort derived from ten childhood-onset-SZ (COS) patients and ten unrelated controls. Overexpression of miR-9 ameliorated a migration deficit in SZ NPCs, whereas knockdown partially phenocopied aberrant migration in control NPCs. Concurrently, a gene-set enrichment analysis of the largest SZ GWAS to date found an enrichment of predicted miR-9 targets among SZ-associated genes, demonstrating a remarkable convergence of independent hiPSC- and genetics-based discovery approaches.

Even within idiopathic hiPSC cohorts, there is a recent effort to focus on patients with shared clinical features such as age of onset, endophenotypes (ie. neurophysiological, biochemical, endocrinological, neuroanatomical, cognitive, or neuropsychological features), or pharmacological response, with the expectation that this may reduce inter-individual variation in vitro. In SZ, COS patients represent a subset of adult onset SZ patients defined by onset and severity, with no genetic or clinical differences to chronic poor outcome adult onset SZ 74, although we have twice applied a COS cohort to replicate findings in our original idiopathic cohort 66, 75, we have no evidence of exaggerated and/or less heterogeneous phenotypes in hiPSC-derived COS neurons, relative to adult-onset SZ neurons. In ASD, two recent studies have now focused specifically on cases with increased head size (macrocephaly) and poorer clinical outcomes 76, 77. Although their differentiation paradigms differed substantially and cohort
sizes were small (four 76 and eight 77 ASD patients), both reported perturbations in synaptogenesis (increased 76 and decreased 77), GABAergic differentiation (increased 76 and decreased 77) and FOXG1 expression (increased 76, 77). Given that these macrocephaly-ASD hiPSC studies were not directly contrasted to a more general idiopathic cohort, one cannot say whether this experimental strategy reduced patient heterogeneity or improved disease signal. This issue is better resolved in a fourth hiPSC-based cohort, this one comparing three BD patients with known clinical lithium responsiveness and three with known nonresponsiveness 78. Here, while hiPSC neurons from all BD patients showed evidence of hyperactive neuronal firing, the phenotype was selectively reversed by lithium treatment only in neurons derived from lithium-responsive BD patients 76, consistent with a genetic mechanism underlying clinical lithium response 79. A similar pharmacological patient stratification is frequently proposed for SZ; while it has not been fully explored to date, one study examined a pair of monozygotic twin cases with treatment-resistant SZ in which one twin responded well to clozapine treatment and the other twin did not, finding subtle gene expression differences between (untreated) twin neurons 80 but failing to ask if the patient-derived neurons differed in response to clozapine treatment. Overall, we posit that only for those endophenotypes with substantial heritability (i.e. cognition 81, 82 and neuroticism 83, 84) or treatment responses with clear pharmacogenomic evidence 85 will cohort heterogeneity be even modestly reduced through this strategy.

IV. Lessons from case-control SZ post-mortem analyses.

Understanding the natural variation in gene expression between control brains, between individuals and across development, is an important first-step before conducting case-control analyses. Towards this, the GTEx project integrates genetic and transcriptomic data across brain regions in the same subjects 86 while the BrainSpan project characterizes gene expression levels across multiple brain regions during human development 87, 88. Large-scale multi-site studies have now established post-mortem brain collections with a comprehensive analysis of the genome, epigenome 89 and transcriptome 90-92 of hundreds of neuropsychiatric disorder patients and controls across multiple brain regions (reviewed in 93).

Post-mortem gene expression studies (and meta-analyses) have identified subtle abnormalities in multiple brain regions and neural cell types, but the results have been inconsistent (reviewed in 94), likely reflecting the small sample sizes involved in the primary cohorts as well as difficulties in accounting for important covariates or accurately combining datasets that were generated at different sites via divergent methods. In addition to the frequently discussed confounds of drug/alcohol abuse, anti-psychotic treatment, cause of death, post-mortem interval and RNA quality, some of this confusion also reflects a failure to account for cellular composition in postmortem brain tissue 95, for example, any loss of neurons (or glia) associated with disease progression could result in many false positives due solely to this changing composition. Moreover, cell-type specific perturbations are frequently missed in tissue level analyses (reviewed 93); for example, genes related to mitochondrial function and ubiquitin-proteasome functions seem to be perturbed in SZ in layer 3 and 5 pyramidal neurons, respectively 96. Until case-control single cell datasets are available, integrating control single cell data into the analysis of case-control comparisons of heterogeneous cell populations will improve compensation for variable cell type composition 97 and resolution of cell-type specific effects 98.
With dramatically expanded sample sizes, consistency of recent results has improved (reviewed). While the CommonMind Consortium (CMC) RNA sequencing (RNA-seq) evaluation of prefrontal cortex (PFC) brain tissue from 537 individuals concluded that current post-mortem studies are dramatically underpowered to detect differential expression directly attributable to SZ risk variants, others hope that this is an overly pessimistic interpretation. In fact, when variation due to degradation of post-mortem RNA was accounted for in the LIBR RNA-seq analysis of 495 post-mortem brains, they identified 237 genes significantly associated with SZ that implicated synaptic processes, were strongly regulated in early development, and replicated in the CMC dataset. Finally, a large cross-disorder meta-analysis of 715 brain samples across five major neuropsychiatric disorders identified shared down-regulation of neuronal gene networks in ASD, SZ, and bipolar disorder (BD), and up-regulation of astrocyte networks in ASD and SZ.

Postmortem studies can be further refined by incorporating genetic data, both from the individual from whom the tissue was obtained as well as the GWAS for the disease being studied. With a sufficiently large cohort, common variants that regulate gene expression can be identified. These expression quantitative trait loci (eQTLs) are regions of the genome containing DNA sequence variants that influence the expression level of one or more genes; they are now understood to vary between cell types and conditions (well reviewed). eQTLs can be integrated with GWAS data to identify genes whose expression levels are associated with disease. One such analysis linked 

\( \text{SNX19} \) and \( \text{NMRAL1} \) to SZ, noting that the genes identified were not generally the nearest annotated gene to the top associated GWAS SNP. The CMC estimated that ~20% of SZ GWAS loci have eQTLs that could regulate gene expression, and for five demonstrated that perturbing gene expression can have effects on neurodevelopment in vivo and/or in vitro. A more refined analysis, considering both cellular and temporal contexts, determined that conditional eQTLs are widespread and revised the estimate of SZ GWAS loci with eQTLs to ~37%. A third group, with a revised analytical strategy that incorporated additional post-mortem brain samples across development, concluded that 42.5% of SZ GWAS variants have eQTLs that converge on gene regulation. A similar strategy substitutes open chromatin state for gene expression, an enrichment of neuronal open chromatin regions at SZ GWAS loci was subsequently confirmed in hiPSC-derived neurons. SZ GWAS risk loci are more likely to present expression differences during the fetal-postnatal developmental transition, with a specific enrichment for both dopaminergic and glutamatergic synapse pathways. Given recent progress in other fields, it is likely that further integration of GWAS and eQTL data to generate a SZ transcriptional risk score will better predict disease risk that using a polygenic risk score alone.

V. Understanding critical power limitations in the design of hiPSC-based studies of complex genetic disorders

Given the limitations of small sample size and intra-donor variation, future hiPSC-based studies should be designed to maximize statistical power. One critical issue is the tradeoff between increasing the number total donors versus increasing the number of independent hiPSC clones per donor. In general, the best way to maximize effective sample size, while controlling the false positive rate, is to use one hiPSC line per donor and increase the number of donors, rather than using multiple replicate clones from a smaller set of donors. We suggest that future hiPSC-based studies of SZ (and other diseases) will achieve the greatest benefit by focusing on three strategies:
increasing the total number of individuals, reducing inter-donor heterogeneity (by focusing on patients with shared genetic variants) and optimizing neuronal protocols to decrease cellular heterogeneity (and thus decreasing intra-donor variation).

Overall, the issues of sample heterogeneity and statistical power are to a large extent shared between post-mortem and hiPSC-based studies; for example, both approaches are equally impacted by the polygenic nature of SZ and genetic heterogeneity between individuals. Unlike the confounds of post-mortem studies discussed above, hiPSC-based studies are impacted by variation in reprogramming (ie. reprogramming method and batch, technician, hiPSC culture) and neuronal differentiation (ie. reagent batch as well as more stochastic experimental effects affecting cell type composition) efficiencies between hiPSC lines derived from both the same (intra-individual) and different (inter-individual) donors. The fraction of expression variation attributable to each of these factors can be quantified using the variancePartition software \(^{106}\). Previous applications have specifically revealed impact of donor, cell type composition and technical artifacts \(^{97,106-108}\).

While it was recently estimated that ~28,500 subjects are required for an adequately powered post-mortem case/control study \(^{91}\), it is unclear whether more or fewer subjects would be necessary for an equivalently powered hiPSC-based cohort. It is simply too early to state with confidence which set of post-mortem and hiPSC specific confounding variables have a greater impact on the observed donor effect in gene expression studies; however, the ability to computationally account for RNA quality \(^{92}\) and cell-type composition \(^{97}\) will improve our ability to resolve biological signals moving forward.

Critically, recent efforts have spurred the establishment of larger NextGen Consortium hiPSC-based studies (discussed below), from which it is now possible to estimate overall sources of variance and observed donor effects in hiPSC-based studies. Both genetic \(^{109-112}\) and epigenetic \(^{113-116}\) errors occur during the reprogramming process. Donor cell type \(^{117,118}\) and age \(^{119}\) can further influence the genome and epigenome of hiPSCs. Moreover, batch effects are seemingly unavoidable in both the reprogramming and differentiation processes. In aggregate, these processes underlie “intra-individual variation”, the subtle differences in gene expression and propensity towards neural differentiation between independent hiPSC lines generated from a given donor. Traditionally, stem cell biologists have attempted to account for these effects by including multiple hiPSCs (up to three) per donor; critically, for complex genetic disorders, we instead propose here that it is more important to power experiments to explore inter- rather than intra-individual variation. For a fixed budget, it is nearly always advisable to add additional donors rather than generate and validate additional hiPSC clones for any given individual (https://gabrielhoffman.shinyapps.io/design_ips_study) \(^{97}\).

Practical limits of cohort size may mean that studies are inadequately powered at the present time for this highly polygenic disorder. Stochastic differences in the differentiation process can result in dramatic differences in cell type composition that can negatively impact analysis. Moreover, the dynamic nature of gene expression means that such differences in cell type composition, neuronal density or other culture conditions can lead to differences in neuronal activity and other physiological events that impact RNA levels. Finally, the constraints of transcriptome-wide multiple testing burden can make it difficult to identify real signals with statistical significance. It is now widely accepted that it is necessary to decrease patient heterogeneity by selecting more (genetically) homogeneous cases and controls. Hoesktra et al \(^{120}\) define this objective more precisely, proposing that future studies either (i) select patients with a rare highly penetrant disease-associated genetic variant with a large effect size, or (ii) select...
patients with high polygenic risk score based on common genetic variants. They further
detail that for both strategies, the ideal design would be to include four groups of
individuals: patients with and without the disease penetrant variant/high polygenic risk
and controls with and without the disease penetrant variant/high polygenic risk 120.
Moreover, a more tenable focus, rather than to continue applying hiPSC-based models
for the discovery of SZ risk factors, may be for the field to apply a combination of genetic
stratification of patient cohorts and gene-edited isogenic hiPSCs in order to evaluate the
functional effects of manipulating putative rare and common causal variants identified
through genetic studies.

To a large extent, inter-individual (genetic) differences explain transcriptional variability
between hiPSCs 107, 121, 122, consistent with the degree of genetic variation in human
gene expression regulation 123-125. Retaining the donor-specific signature is essential to
studying case control differences. In two recent studies of hiPSCs, variance across
donors explained a median of ~6% 126 and 48.8% 107 of expression variation, while we
observed a much smaller (2.2%) donor effect in hiPSC neurons 97. Donor effects in
differentiated cells may be reduced due to stochastic noise in the differentiation from
hiPSCs to neurons. Consistent with this, regulatory variation between individuals is lower
in hiPSCs than in two differentiated cell types 121. It remains unclear whether different
hiPSC-derived cell types will retain more or less donor signal over the course of
differentiation.

Although not focused on neuronal cell types, many of the insights from the large
collection of hiPSCs recently characterized by the NextGen Consortium 127 are very
relevant to neuropsychiatric disease, as these reports quantified the genetic 107, 128 and
epigenetic 129 basis of variation between hiPSC lines and differentiated progeny 130-132.
Separate work focused on the transcriptomes and chromatin accessibility of hiPSC-
derived sensory neurons 108. These large cohorts facilitated both genome-wide insights
into common variants underlying disease 130-132, but also served as a platform to begin
personalized (or at least genome-first) drug screening against disease mechanisms or
phenotypes 133-135.

VI. Genetically stratified cohorts of rare variant patients

Despite their rarity in the genome (0.01-0.02 per generation 136), CNVs (frequently de
novo) were among the first genetic variants associated with SZ. There is an increased
CNV burden in SZ cases compared to controls 137, 138, so one method to increase the
power of hiPSC-based studies, without increasing sample size, is to focus on genetically
homogenous cohorts harboring such a rare variant of large effect size.

22q11.2 deletion was the first CNV associated with SZ and remains the strongest risk
factor for developing the disorder, with 25% of carriers exhibiting psychotic symptoms
139-141. hiPSC-derived neurons from two 22q11.2 patients diagnosed with SZ
demonstrated significant reductions in a number of cellular phenotypes including
neurosphere size, neuronal differentiation capacity (neuron to glial ratio), neurite
outgrowth and cellular migration 54. Additionally, given that the 22q11.2 deletion
encompasses DGCGR8, a known regulator of miRNA processing, hiPSC cohorts of
22q11.2 deletion cases have been subject to both miRNA and RNA sequencing. miRNA
sequencing, performed on a cohort of six patients compared to six controls, identified a
significant increase in the expression of several miRNAs that overlap with differentially
expressed miRNAs in post-mortem brain and peripheral cells from 22q11.2 cases 67. In
a larger cohort of eight 22q11.2 cases and seven controls, RNA-seq identified changes
in expression of nearly all genes within the deletion region, as well as 745 genes outside of the region implicating apoptosis, cell cycle and MAPK signaling \(^{142}\). These studies demonstrate that rare variant cohorts can identify cellular phenotypes and molecular pathways that are consistent with human post-mortem findings.

A gene-editing approach can also be used to investigate the causal role of a gene at a SZ-associated CNV. Given that \(NRXN1\) deletions are only 6.4% penetrant, Pak et al used CRISPR/Cas9 (CRISPR-associated protein 9) technology to engineer two heterozygous deletions in a control line, to assess if these deletions were sufficient to produce neuronal deficits. Both deletions resulted in deficits in neurotransmitter release, with no changes in synapse number or neuronal differentiation capacity, a finding consistent with mouse models \(^{143}\). Isogenic comparisons are crucial for establishing a causal role for incompletely penetrant rare variants such as \(NRXN1\).

While focusing on rare genetic variants with large effect size will increase the power of hiPSC-based analyses, their incomplete penetrance and pleiotropic effects may complicate the identification of disease-specific phenotypes, which are likely impacted by polygenic risk. Therefore, it is imperative to consider the genetic background of common variants in each donor, even when focusing on variants of large effect size. Ideally, isogenic controls should be used whenever possible to demonstrate causal relationships between genetic variants and their phenotypic consequences.

**VII. CRISPR-editing and manipulating of expression to study common variant effects**

hiPSC-based models can be applied to functionally evaluate common variant risk genes identified by such analyses. For example, induced neurons (iNeurons) generated from over 20 individuals with different genotypes for a non-coding common variant in the voltage-gated calcium channel subunit gene \(CACNA1C\) demonstrated that the homozygous risk allele genotype correlated with increased \(CACNA1C\) mRNA expression and calcium channel current density \(^{144}\). While this study correlated genotype to expression, it did not directly demonstrate the role of the SNP in regulating gene expression.

A combination of CRISPR-based tools now make it possible to precisely engineer hiPSC lines with single nucleotide mutations to recapitulate common variants, finely tuned endogenous gene expression changes, as well as the multiplexing of such modifications \(^{145-147}\). A number of critical issues limit these studies and will be discussed below. First are the decisions of which hiPSC line (genotype and diagnosis) and which precise SNP(s) to engineer, and whether to edit one or both alleles. Second is how to prevent and identify off-target effects, in order to ensure the construct validity of any biological findings.

Obviously, as well established between different mouse strains, genetic background has dramatic impact on gene expression and would be predicted to significantly impact the effects of genetic manipulation \(^{148-152}\). Therefore, one important question is whether to conduct these isogenic experiments in control or SZ hiPSCs. If penetrance is to be first established in control backgrounds, there is the possibility that potential (protective) compensatory processes may limit the impact of the genetic change. If effects are queried in patients, there is the possibility that any of the other risk alleles may interact with the edited SNP to confuse the results. Likely, isogenic comparisons will need to be conducted across a number of genetic backgrounds. Establishing the polygenic risk
scores and haplotypes of different donor cell lines will be critical when choosing a cell line for genetic engineering. Close collaboration between computational geneticists and biologists is essential for the study of complex genetic disorders.

It remains unknown to what extent, and across how many putative causal loci, manipulating eQTLs in control hiPSCs will be sufficient to alter the levels and/or splicing of associated mRNAs, leading to downstream phenotypic changes in one or more neural cell types. To date this has been accomplished comparatively few times, so careful choice of the candidate variant will prove beneficial. Forrest et al. prioritized SZ-associated variants located in neuron-specific open chromatin regions in iNeurons. CRISPR correction of a common variant risk allele near the mir137 locus increased expression of mir137, dendritic complexity and synapse maturation. CRISPR-editing of common variants associated with other diseases has similarly been shown to affect gene expression in hiPSC-derived disease-relevant cell types. In a novel attempt to bridge the study of common and rare variants, Castel et al. tested the potential effect of regulatory variants on the penetrance of pathogenic coding variants. By introducing a pathogenic Mendelian variant into a known eQTL haplotype they elegantly demonstrated a possible mechanism in which common variants can contribute to variable penetrance of genetic disorders.

A significant hindrance to the study of common disease-associated variants is their great likelihood to be in high linkage disequilibrium (LD) with other variants. While the causal SNP may be predicted via fine mapping, it cannot always be identified unequivocally. It may be more simple to directly manipulate endogenous gene expression; in this way, hiPSC-derived neural cells were used to show that the SZ candidate gene ZNF804A impacts inflammatory cytokine response in differentiating neurons, CYFIP1 affects neural polarity in NPCs, and FURIN alter neural migration in NPCs. Not all genes are appropriate targets for traditional overexpression or knockdown experiments, particularly those that are very long or highly alternatively spliced. Fortuitously, gene expression modulation via CRISPR/Cas9 occurs at the
promoter or enhancer, and therefore is predicted to include the full range of alternative
splice isoforms that are frequently overlooked by cDNA overexpression or RNAi
approaches. By introducing nuclease-null mutations into Cas9\textsuperscript{175, 176} and fusing the
catalytically inactive or dead Cas9 (dCas9) to a variety of effector protein domains, the
modulation of transcription\textsuperscript{175, 176}, DNA methylation\textsuperscript{177, 178} and histone modifications\textsuperscript{179}
have all been demonstrated. By testing dCas9-mediated transcriptional modulation using
three such platforms, we recently fully evaluated the efficacy and variability of dCas9-
protein fusion-based transcriptional modulation of seven different SZ-associated risk
genes in three different hiPSC-derived neural cell types (NPCs, neurons and astrocytes),
using hiPSCs from three unique donors\textsuperscript{180}. While this platform has not proven equally
efficacious across all neuronal genes, donors and cell types tested, it can be a fast
method to achieve disease relevant changes in gene expression.

VIII. Integrating advanced genetics and hiPSCs in future approaches

As hiPSC-based studies increasingly incorporate isogenic comparisons, stem cell
biologists must contemplate statistical and biological considerations that are already well
accepted in the field of genetics. When planning CRISPR-mediated isogenic hiPSC-
based comparisons to validate disease-associated eQTLs, care must be taken to design
a well powered study. In general, the power to identify a statistically significant
difference in the expression of a particular gene between two or more groups of samples
depends on the magnitude of the expression differences between the groups (i.e. effect
size) and the amount of expression heterogeneity within each group (i.e. variance).

Standard eQTL analysis divides samples into three allelic categories based on a given
SNP and determines if the number of reference alleles for that SNP is significantly
associated with expression of a particular gene. For a particular SNP, the statistical
power to detect a significant effect on gene expression varies by gene, reflecting both
the effect size of the eQTL and the expression variance within each of the three allelic
groups (Figure 2). Although isogenic hiPSC and post-mortem experiments are predicted
to have equivalent eQTL effect sizes, we strongly believe that isogenic hiPSC-based
approaches will demonstrate reduced expression variance due to the common genetic
background in isogenic comparisons. Therefore, assuming even a 5-fold reduction in
standard deviation (i.e. square root of the variance) for most genes, an isogenic design
is well powered to identify differences in the cis-gene targets (assuming a comparison of
four isogenic hiPSC replicates from a single donor with the CRISPR allele swap to the
homozygous alternative allele to twelve isogenic replicates with the homozygous
reference allele). While we hypothesize that using isogenic lines will reduce biological
and technical variation, until the extent to which this is true is empirically evaluated, it
must be modeled across a range of values of expression variability.

Statistical considerations for CRISPR validation studies

Designing a successful study requires adapting the experiment, question or analysis so
that it is experimentally tractable, biologically plausible and statistically well powered.
Here, although we consider the power to detect a significant expression change
following a targeted CRISPR allele swap of a SNP affecting gene expression, where the
eQTL association was identified in postmortem brain, the same reasoning applies to
testing case/control expression differences. The statistical power to detect expression
differences between two sets of samples depends on four major factors discussed
below: magnitude of expression difference, expression variance, sample size and
significance level.
i) **Magnitude of the expression difference.** For a given SNP-gene pair, a linear regression model can be fit to estimate the effect size ($\beta$) of increasing the number of minor alleles and the variance of the expression residuals ($\sigma^2$) after the SNP effect is considered. Let $\mu_0$ and $\mu_2$ be the average expression levels of the two homozygote classes. Since $\beta$ is the difference in expression from changing a single allele, then $\mu_2 - \mu_0 = 2\beta$. Power calculations depend on the term $2\beta/\sigma$ and generally assume that the effect size remains the same in the discovery and validation datasets. But, when isogenic hiPSC studies are used to validate findings identified in whole brain tissue, this variation in cell type composition can reduce the observed magnitude of a signal present in only one cell type. Therefore, it is obviously desirable to conduct hiPSC-based isogenic experiments in the appropriate neural cell type where the eQTL is most active, increasing the statistical power by increasing the effect size.

ii) **Expression variance.** Gene expression levels measured in postmortem brain can be highly variable due to donor genetics, variation in cell type composition, environmental effects experienced by the donor, and technical variation in sample processing. Increasing the fraction of expression variation explained by the SNP of interest by reducing the variation attributable to extraneous variables will improve the statistical power to validate an initial finding. Using isogenic cell culture model as a validation system should remove some of this variation and increase statistical power.

iii) **Sample size.** Collecting a sufficient number of samples is essential to conducting a well-powered study, especially when effect sizes are small. Of course, findings should always report the number of hiPSC lines (and individuals represented), experimental replicates and technical replicates conducted for each analysis. Observations should be presented by individual, rather than as averaged group effects, so as to most transparently reflect variation within these isogenic comparisons. Interpreting expression variation within each phenotype class is essential to interpreting the biological relevance of a finding.

Obviously, cost constraints are often the major factor limiting sample size. Therefore, one solution is to conduct all isogenic comparisons in a few common hiPSC line(s), allowing all engineered lines to serve as additional isogenic controls for the other pairs. The effects of biological heterogeneity can be further reduced by conducting a paired statistical analysis whereby each perturbed sample is compared to a donor-specific baseline (Figure 3).

iv) **Significance level.** The central goal of most gene expression studies is to identify expression differences that are larger than expected by chance. Findings are generally evaluated based on passing statistical significance, by convention often $p<0.05$. When multiple genes are tested in an analysis, the $p$-value cutoff to attain significance must be adjusted in order to account for the number of statistical hypothesis tests. This is known as the ‘multiple testing problem’. Analyses focusing on a single gene can use a nominal $p$-value cutoff of 0.05. Genome-wide analyses must overcome a higher multiple testing burden in order to control the false positive rate. The Bonferroni correction is widely used and intuitive so that a nominal $p$-value cutoff of 0.05 corresponds to $0.05/k$ when $k$ genes are tested. When 20,000 genes are tested the Bonferroni cutoff becomes $p<2.5e-6$ and is much more stringent that the nominal cutoff.

The scope of the biological question and the statistical analysis has major impact on the multiple testing burden and the power of the study. Thus decreasing the scope of an analysis by only considering genes near a SNP of interest can dramatically increase
statistical power compared to a genome-wide analysis. Overall, focused validation experiments will always be better powered than genome-wide discovery.

**Biological considerations for CRISPR validation studies**

While the effect of modifying an eQTL SNP in an hiPSC-based study of gene expression can be estimated from post mortem data, considering the effect of a SNP identified by GWAS of a psychiatric trait is more challenging and open-ended. Whereas genetic variants have a large effect on direct phenotypes (such as gene expression), the effect on higher-level phenotypes (such as neuronal function or diagnosis) will be attenuated by buffering and/or environmental effects. Thus, the effect size will likely reflect how far the assayed phenotype is from genetics and assumes that one has selected the correct intermediate phenotype to study. For a given eQTL SNP, the sorting of effect sizes might look like: gene expression > protein > neuronal function > psychosis > SZ. A particular SNP might not act through gene expression, or not under the specific conditions of the experiment, or not in a particular cell type.

Moreover, the question remains as to how to link eQTL associations to the complex behavioral phenotypes associated with SZ, ranging from delusions, hallucinations, negative affect and impaired cognition. Obviously, once the constraints of these isogenic studies are better understood, future studies may wish to improve the complexity and maturity of the neural cells being queried, advancing towards circuit-level complexity. Moving forward, it will be increasingly straightforward to incorporate three-dimensional culture techniques and/or generate defined neuronal circuits comprised of specific neuronal cell types, synapsed in a defined orientation, together with oligodendrocytes to provide myelination, and astrocytes and microglia to incorporate critical aspects of inflammation and synaptic pruning. Moreover, either by neuronal stimulation or treatment with physiologically-relevant exposures of stress hormones and/or other environmental factors, it may one day be possible to further unravel the causality of environmental risk factors such as neuroinflammation, stress and drug exposure using isogenic hiPSC-based comparisons. While models, by definition, will always lack the intricacies of human disease, our goal must always be to strive towards the complexity of the human brain.

**IX. A perspective on the hypothesis of biological convergence of disease**

Much of the current work to improve our understanding of disease etiology and develop novel treatment strategies is implicitly predicated on an assumption of biological convergence. Under this model, patients who share high-level psychiatric symptoms are clinically diagnosed with a particular disease and are hypothesized to share a disruption of some lower-level biological function. As discussed in section II, this convergence model has yielded themes of disruption in ion channels and neuronal genes in SZ. Recent psychiatric research has pursued models of biological convergence at multiple levels (Figure 4), including genetic, epigenetic, gene expression, gene module, proteome, braining imaging, drug response, psychiatric endophenotypes and disease subtypes and high-level disease phenotypes.

The best level to study this common hypothesis of functional convergence is unclear; it may be too low level or underpowered to detect at the level of gene expression and more likely to be identified at the level of modules/pathways/gene-sets/subnetworks. For example, while there may be hundreds of genetic perturbations that converge at the level of synaptic function, the majority of these likely have a distinct expression signature
Therefore, detecting a convergence at the molecular level is dependent on having the proper module/pathway/gene-set/subnetwork to test. Whereas these pathways show clear convergence at the higher behavioral/psychiatric level, and deconstructing this convergence to lower level phenotypes can reveal shared molecular etiology and potential therapies, which level to focus on is not clear. The amount of convergence increases with complexity (i.e. toward psychiatry) since the diagnosis is based on psychiatric symptoms, but so does the cost per patient to study. Going down a level (towards DNA) allows increased resolution, sample size and prospects for understanding molecular mechanism, but also dramatically increases multiple testing burden and is susceptible to a lack of biological convergence. Obviously, each has its own technical and logistical challenges.

A major challenge in psychiatric genomics is how best to align a patient cohort, biological assay and analytical approach with a hypothesis of convergence that is both biologically feasible and statistically well powered. While in ASD there is sufficient convergence at the genetic and pathway levels that exome and genome sequencing have already identified multiple genetic disruptions of the same genes and pathways, such clear convergence is not seen in SZ. Since SZ risk is conferred by genetic variants of weaker effect sizes, there is weaker convergence at the genetic and expression levels in SZ, and so convergence has only been observed at the pathway (i.e. ion channel) level. Moving forward, this can be addressed by increasing sample size, developing focused cohorts to increase statistical power and applying integrative statistical methods to detect convergence at new levels. The open questions in psychiatric genetics are therefore, given a set of patients, with a given amount of heterogeneity in genetics, phenotype and technical noise, as well as cost limitations: what level of biological convergence can be expected, is the assay able to detect this level of convergence and are we statistically well-powered to detect a signal of a reasonable effect size? By thoughtfully integrating GWAS, post-mortem and hiPSC-based approaches, we hope that the molecular convergence underlying SZ and other complex genetic disorders will become better resolved, revealing novel points of therapeutic intervention.

ACKNOWLEDGEMENTS

Kristen J Brennand is a New York Stem Cell Foundation - Robertson Investigator. The Brennand Laboratory is supported by a Brain and Behavior Young Investigator Grant, National Institute of Health (NIH) grants R01MH101454 and R01MH106056 and the New York Stem Cell Foundation. Erin Flaherty is supported by F31MH112285. The authors declare no conflict of interest exists.
**FIGURES**

**Figure 1: Genetic contributions to neuropsychiatric disorders.** A) Current state of psychiatric genomics with the assays and analyses linking genetics, gene expression and disease biology. B) Genetic liability to neuropsychiatric disease. Common variants constitute the majority of the genetic contribution to neuropsychiatric disorders. Graph adapted from Gandal et al. C/D) Representative plots illustrating the polygenetic nature of neuropsychiatric diseases. The liability threshold plot shows the normal distribution of disease liability among the population. Genetic and environmental factors combined may lead to crossing of the threshold into disease. The dose-response plot visualizes the same additive effect of genetic liability to disease risk.

**Figure 2: Genetics of gene expression and statistical power.** A) Statistics from eQTL analysis of 3 representative genes from post mortem RNA-seq data from the CommonMind Consortium. The reference SNP ID (rsid) of the SNP with the smallest p-value for each gene is shown along with the corresponding p-value, effect size (β), residual variance (σ², but shown on the scale of standard deviation (σ)), and the percentage of expression variation explained by the SNP. B) Observed expression of the 3 genes from (A) in post mortem brains stratified by the corresponding eQTL SNP. C) Hypothetical gene expression in isogenic hiPSC-derived cell lines substantially reduces the amount of expression variation due to multiple genetic and non-genetic factors. D) Statistical power of targeted CRISPR allele swap from one homozygote to the other as a function of the residual variance, where effect size and variance of each gene were computed from CommonMind Consortium data of 534 post mortem brains. For each gene the power is computed as a function of the residual variance for a study of 4 hiPSC lines that are homozygous for the alternative allele at the relevant SNP compared to 12 hiPSC lines that are homozygous for the reference allele. Power is evaluated at p<0.05 corresponding to a focused analysis testing only a single gene. Power is shown as σ decreases by up to a factor of 10. We note that statistical power varies substantially across genes.

**Figure 3: Statistical tests for independent and paired samples.** A) Testing expression differences between two genotypes in independent samples involves comparing the mean expression between the two genotypes. The statistical analysis considers whether the difference in the mean expression in the homozygous reference genotype (i.e. Ref/Ref) is statistically different from the mean expression in the homozygous alternative genotype (i.e. Alt/Alt). This corresponds to testing if the difference in means is statistically different from zero. Here the color of the observation indicates the donor, but the fact that each donor is measured twice is not used. B) When a study design involves paired samples as in a perturbation experiment with treated and control samples from the same individual, an individual-specific baseline can be used to compute the expression change between the two genotypes within each individual. In this case, the statistical test considers the difference between the treated and control samples for each individual, and then tests if the mean of the differences is statistically different from zero. The colors of the observations indicate the donor and the statistical model explicitly considers the fact that the paired observations are from the same donor. In R, the paired model is used by including the donor in the regression formula.

**Figure 4: Hypothesis of multiple levels of biological convergence.** A) Multiple genetic variants directly affect the same risk gene. B) Risk variants have an indirect
effect on the expression of a downstream gene. C) Risk variants indirectly affect a higher level molecular process by disrupting the function of a biological module of genes. D) Risk variants disrupt intermediate genes to indirectly affect a psychiatric phenotype. E) Biological variation is traced from low level DNA variants to high level disease phenotype through intermediate levels of increasing complexity. The possibility of performing biological assays at each level is indicated for hiPSCs, post mortem brains and living patients. Gradients indicate the level of biological convergence expected for idiopathic versus a genetically selected cohort. The feasible sample size is largest for studies of DNA and disease phenotype (i.e. GWAS), but is much lower for intermediate phenotypes.
REFERENCES


7. Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Current opinion in genetics & development 2016; 41: 150-158.


1266  Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell*
1268
1269  Panopoulos AD, D’Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt
1270  BM et al. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional
1271  Characterization of Genetic Variation across a Variety of Cell Types. *Stem
1273
1274  Cayo MA, Mallanna SK, Di Furio F, Jing R, Tolliver LB, Bures M et al. A
1275  Drug Screen using Human iPSC-Derived Hepatocyte-like Cells Reveals
1276  Cardiac Glycosides as a Potential Treatment for Hypercholesterolemia.
1278
1279  Bi X, Pashos EE, Cuchel M, Lyssenko NN, Hernandez M, Picataggi A et al.
1280  ATP-Binding Cassette Transporter A1 Deficiency in Human Induced
1281  Pluripotent Stem Cell-Derived Hepatocytes Abrogates HDL Biogenesis
1283
1284  Gu M, Shao NY, Sa S, Li D, Termglinchan V, Ameen M et al. Patient-
1285  Specific iPSC-Derived Endothelial Cells Uncover Pathways that Protect
1286  against Pulmonary Hypertension in BMPR2 Mutation Carriers. *Cell Stem
1287  Cell* 2017; 20(4): 490-504 e495.
1288
1290  Mapping copy number variation by population-scale genome sequencing.
1292
1293  Rees E, Kirov G, O'Donovan MC, Owen MJ. De novo mutation in
1294  schizophrenia. *Schizophr Bull* 2012; 38(3): 377-381.
1295
1296  Szatkiewicz JP, O'Dushlaine C, Chen G, Chambert K, Moran JL, Neale
1297  BM et al. Copy number variation in schizophrenia in Sweden. *Mol
1298  Psychiatry* 2014; 19(7): 762-773.
1299
1300  Karayiorgou M, Simon TJ, Gogos JA. 22q11.2 microdeletions: linking DNA
1301  structural variation to brain dysfunction and schizophrenia. *Nat Rev
1303
1304  Schneider M, Debbane M, Bassett AS, Chow EW, Fung WL, van den Bree
1305  M et al. Psychiatric disorders from childhood to adulthood in 22q11.2
1306  deletion syndrome: results from the International Consortium on Brain and
1307  Behavior in 22q11.2 Deletion Syndrome. *Am J Psychiatry* 2014; 171(6):
1308  627-639.


174. Nebel RA, Zhao D, Pedrosa E, Kirschner J, Lachman HM, Zheng D et al. Reduced CYFIP1 in Human Neural Progenitors Results in Dysregulation


A) Common SNP → Rare SNV, CNV

Gene expression → Genetics

GWAS

Differential expression

Disease

B) Liability (%)

common genetic variation

rare genetic variation

C) Population frequency

threshold

D) Disease probability

liability

Disease

Gene expression

Common SNP

Rare SNV, CNV

eQTL

Outlier analysis

GWAS

Family and population WES/WGS
Type of file: figure
Label: 2
Filename: figure_2.pdf
A) eQTL statistics

<table>
<thead>
<tr>
<th>gene</th>
<th>rsid</th>
<th>P_eQTL</th>
<th>$\beta$</th>
<th>$\sigma$</th>
<th>% var</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATAD2A</td>
<td>rs2965199</td>
<td>5.52e-27</td>
<td>0.1103</td>
<td>0.1334</td>
<td>6.08</td>
</tr>
<tr>
<td>FURIN</td>
<td>rs4702</td>
<td>1.65e-12</td>
<td>0.0749</td>
<td>0.1345</td>
<td>1.74</td>
</tr>
<tr>
<td>TCF20</td>
<td>rs5751341</td>
<td>2.24e-02</td>
<td>0.0163</td>
<td>0.0901</td>
<td>1.83</td>
</tr>
</tbody>
</table>

B) Observed in post-mortem brains

C) Hypothetical expression in isogenic hiPSC comparisons

D) Power of targeted CRISPR allele swap

Gene
- Blue: GATAD2A
- Green: FURIN
- Red: TCF20

Reducing expression variation
Type of file: figure
Label: 3
Filename: figure_3.pdf
Δ

≠

expression ~ genotype ~ genotype + donor

Statistical test

R regression formula

\[ \mu_{\text{Alt/Alt}} - \mu_{\text{Ref/Ref}} \neq 0 \]

\[ \bar{\Delta} \neq 0 \]

expression ~ genotype

expression ~ genotype + donor
A) Genetic convergence

B) Expression convergence

C) Module convergence

D) Psychiatric convergence

E) Biological convergence

Phenotype

Biological convergence

- Disease phenotype
- Psychiatric endophenotypes
- Drug response
- Brain imaging
- Cellular electrophysiology
- Proteome
- Gene expression
- Epigenetics
- DNA