A longitudinal model of human neuronal differentiation for functional investigation of schizophrenia disease susceptibility

Authors:

Anil P. S. Ori[1], Merel H. M. Bot[1], Remco T. Molenhuis[1⊥], Loes M. Olde Loohuis[1], Roel A. Ophoff[1,2,3]*

Affiliations:

[1]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California, USA

[2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA

[3]Brain Center Rudolf Magnus, Department of Psychiatry, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

⊥Current affiliation: Department of Translational Neuroscience, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

*Correspondence to Roel A. Ophoff

695 Charles E. Young Drive South, CA 90095, Los Angeles, USA

Tel: (310) 794 9602

E-mail: Ophoff@ucla.edu

Abstract

There is a pressing need for *in vitro* experimental systems that allow for functional investigation of psychiatric disease biology. We developed an analytical framework that integrates genome-wide disease risk from GWAS with longitudinal *in vitro* gene expression profiles of human neuronal differentiation. We demonstrate that aggregate polygenic disease risk of specific psychiatric disorders is significantly associated with genes that are differentially expressed across neuronal differentiation. We find significant evidence for schizophrenia, which is driven by a longitudinal synaptic gene cluster that is upregulated during differentiation. Our findings reveal that *in vitro* neuronal differentiation can be used as an experimental model system to translate genetic findings to schizophrenia disease biology. Overall, this work emphasizes the use of longitudinal *in vitro* transcriptomic signatures as a cellular readout and the application to the genetics of complex traits.

57 **Introduction**

58 Major psychiatric disorders feature a high heritability (h^2) but have a largely unknown
59 etiology[1, 2]. The increasing sample sizes of genome-wide association studies (GWAS)
60 successfully result in identification of more susceptibility loci for these disorders[3]. A major
61 challenge is to understand how genetic findings translate to biological pathways and to
62 develop *in vitro* model systems that recapitulate molecular and cellular processes underlying
63 these devastating diseases[4].

64 Early brain development has been implicated in psychiatric disorders such as
65 schizophrenia (SCZ)[5–8], autism spectrum disorder (ASD)[9, 10], and self-reported depression
66 (SRD)[11]. Differentiation of human embryonic stem cells (hESCs) into neuronal lineages has
67 been demonstrated to hold great promise to model early brain development[12–14], and may
68 thus offer a unique opportunity to study psychiatric disease biology *in vitro*. However, it has
69 remained unclear whether the molecular dynamics underlying *in vitro* human neuronal
70 differentiation are associated with psychiatric disease susceptibility.

71 We set out to investigate *in vitro* human neuronal differentiation in the context of
72 polygenic psychiatric disease risk. To accomplish this, we performed a densely sampled
73 time series experiment to robustly detect transcriptome-wide changes across neuronal
74 differentiation. By integrating longitudinal gene expression signatures with GWAS summary
75 statistics we observe significant enrichment of genetic risk for SCZ in genes that are
76 differentially expressed across differentiation. We further show that this enrichment is driven
77 by a specific longitudinal gene cluster that is involved in synaptic functioning. These findings
78 support the use of *in vitro* genome-wide gene expression profiles to study psychiatric
79 disease processes and establish *in vitro* neuronal differentiation as a promising model
80 system to investigate molecular mechanisms that underlie schizophrenia based on evidence
81 from GWAS.

82

83

84

85 **Material and Methods**

86 ***Approval for stem cell research***

87 This study and all described work was approved by the University of California, Los
88 Angeles Embryonic Stem Cell Research Oversight (ESCRO) committee.

89

90 ***In vitro human neuronal differentiation***

91 WA09(H9)-derived hNSCs were commercially obtained (Gibco) as neural progenitors
92 and subsequently expanded as adherent culture according to the manufacturer's guidelines
93 (Supplementary Methods). Low passage hNSCs (< 4 passage rounds) were plated in 12-
94 well plates coated with poly-D-lysine (0.1 mg/mL, VWR) and laminin (4.52ug/cm^2, Corning^{TM})
95 at 1.5x10^5 cells, which were equally distributed and subsequently cultured in expansion
96 medium as described above. After 24h of proliferation, media was changed to neuronal
97 differentiation medium consisting of Neurobasal® Medium (Gibco), 2% B-27® Serum-Free
98 Supplement (Gibco), 2mM GlutaMax™-I Supplement, 0.05 mM β-mercaptoethanol (Gibco),
99 and 1x Pen Strep. Media was changed every 2-3 days.

100

101 ***Experimental design and assessment of gene expression***

102 Human neural stem cells were differentiated over a course of 30 days and RNA
103 harvested at seven time points (day 0, 2, 5, 10, 15, 20, and 30) in triplicates or
104 quadruplicates (n = 24) (Supplementary Methods). Genome-wide array-based transcriptome
105 data was collected at the UCLA Neuroscience Genomics Core using Illumina's HumanHT-12
106 v4 Expression BeadChip Kit.

107

108 ***Data preprocessing and quality control***

109 Gene expression data was extracted using the Gene Expression Module in
110 GenomeStudio Software 2011.1. Data was background corrected with subsequent variance-
111 stabilizing transformation and robust spline normalization was applied[15, 16]. We excluded low
112 quality probes and subsequently performed sample outlier detection by Euclidean distance

113 and standardized connectivity (see Supplementary Methods). The FactoMineR package
114 (v1.28) in R was used to perform principal component analysis (PCA). For subsequent
115 downstream analyses, we used the normalized expression values of 19,012 high quality
116 filtered probes for all 24 samples.
117
118 ***In vitro cellular identity***
119 We identified cell-type specific genes of neurons, astrocytes, oligodendrocyte
120 precursor cells (OPC), newly formed oligodendrocytes (NFO), myelinating oligodendrocytes
121 (MO), microglia, and endothelial cells from mouse cerebral cortex[17] (Supplementary
122 Methods). Next, we extracted normalized gene expression values of these genes for each
123 cell type from our own *in vitro* dataset. We then standardized expression values to time point
124 zero and calculated mean standardized expression levels of cell type-specific genes for
125 these seven cell types across time points to investigate cellular identity across differentiation.
126
127 ***Transition mapping to a spatiotemporal atlas of early human brain development***
128 To investigate global transcriptomic matching between *in vitro* gene expression
129 profiles and *in vivo* gene expression profiles of neocortical brain regions, we applied
130 transition mapping (TMAP), which is implemented in the online CoNTExT bioinformatic
131 pipeline (https://context.semel.ucla.edu)[14]. Analyses were run for *in vitro* time points day-0 vs
132 day-30, day-0 vs day-5, day-5 vs day-15, and day-15 vs day-30 across both temporal and
133 spatial dimensions of human cortical development (see Supplementary Methods).
134
135 ***Time-series differential gene expression analysis***
136 Two multivariate empirical Bayes models were used to identify differentially
137 expressed genes across *in vitro* neuronal differentiation. The first method was implemented
138 in the Timecourse package (v 1.42) in R. We used the mb.long() function to calculate the
139 one-sample T^2 statistic that ranks genes based on their log10 probability to have differential
140 expression over time[18]. Bayesian Estimation of Temporal Regulation (BETR), an extension

141 of the first approach, uses a flexible random-effect model that allows for correlations
142 between the magnitude of differential expression at different time points[19]. BETR is
143 implemented in the betr package (v 1.26) in R. Differentially expressed genes were classified
144 as the union of the set of genes with a probability of 1.0 using BETR and an equally-sized
145 set of top ranked genes using the T^2-statistic.
146
147 ***Fuzzy c-means cluster analysis***
148 To identify probes with similar expression patterns across differentiation, we applied
149 fuzzy c-means clustering to all differentially expressed probes. We calculated cluster
150 membership values using the fclusList() and membership() function in the Mfuzz package in
151 R[20, 21]. See Supplementary Methods for more details. Each probe receives a membership
152 value for each cluster. Probe membership values represent gene affiliation to a cluster and
153 highlights the extent of similarity in expression between genes. These values were used for
154 subsequent downstream analyses. We annotated clusters using Database for Annotation,
155 Visualization, and Integrated Discovery (DAVID, v6.8) and probes with a membership > 0.5
156 (Supplementary Methods).
157
158 ***Integration of GWAS data with in vitro transcriptomic signatures***
159 We first mapped Illumina probe IDs to Ensembl gene IDs using NCBI build 37.3,
160 removed duplicate Ensemble IDs, and extended gene boundaries symmetrically by 10kb to
161 include regulatory regions. Probe T^2-statistic and cluster membership values were collapsed
162 per gene ID using the mean value across probes. The mean gene-level T^2-statistic was then
163 log-transformed and the mean cluster membership values rank-transformed. These mean
164 gene values were then used to integrate *in vitro* signatures with GWAS data using Multi-
165 marker Analysis of GenoMic Annotation (MAGMA) and stratified LD score regression
166 (sLDSR).
167
168 *GWAS summary statistics and ancestry matched reference panels*

169 GWAS summary statistics were obtained for SCZ[22], major depressive disorder
170 (MDD)[23], SRD[11], bipolar disorder (BPD)[24], ASD[25], attention deficit hyperactivity disorder
171 (ADHD)[26], cross disorder[27], Alzheimer's disease (AD)[28], and adult human height[29]
172 (Supplementary Methods and Table S2). The 1000 Genomes Project Phase 3 release (1KG)
173 was used as reference panel to model LD[30]. We used 503 individuals of European ancestry
174 and 301 individuals of East Asian ancestry in analyses of GWAS data derived from target
175 population of Europeans and Han Chinese, respectively.
176
177 *MAGMA gene-set analysis*
178 Multi-marker Analysis of GenoMic Annotation (MAGMA v1.06)[31] was used to run
179 "gene property" analyses, which uses a multiple regression framework to associate a
180 continuous gene variable to GWAS gene level p-values. SNPs were mapped to genes using
181 Ensembl gene IDs and NCBI build 37.3 gene boundaries +/- 10kb extensions using the --
182 annotate flag. For each phenotype, we generated gene-level p-values by computing the
183 mean SNP association using the default gene model ('snp-wise=mean'). We only included
184 SNP with MAF > 5% and dropped synonymous or duplicate SNPs after the first entry
185 ('synonym-dup=drop-dup'). For each annotation, we then regressed gene-level GWAS test
186 statistics on the corresponding gene annotation variable using the '--gene-covar' function
187 while adjusting for gene size, SNP density, and LD-induced correlations ('--model
188 correct=all'), which is estimated from an ancestry-matched 1KG reference panel. In all
189 analyses, we included only genes for which we had both the gene variable and GWAS gene
190 level test statistic available. Testing only for a positive association, i.e. enrichment of GWAS
191 signal, we report one-sided p-values along with the corresponding regression coefficient.
192
193 *Stratified LD Score Regression*
194 We applied a recent extension to stratified LD score regression (sLDSR), a statistical
195 method that partitions h^2 from GWAS summary statistics[8]. This extension allows us to
196 partition h^2 by continuous-valued annotations[32]. For each annotation, we first estimated

197 partitioned LD scores using using the ldsc.py --l2 function with MAF > 5%, a 1 centimorgan
198 (cm) window, and an ancestry-match 1KG reference panel (Supplementary Methods). To
199 partition h^2 of each phenotype by our *in vitro* transcriptomic signatures, we ran sLDSR
200 (ldsc.py --h2) for each annotation of interest while accounting for the full baseline model, as
201 recommended by the developers[8, 32], and an extra annotation of all genes detected in our *in*
202 *vitro* model (n = 12,414). That is, for each annotation we partitioned h^2 with the following
203 annotations;

204 1. Full baseline model with 53 annotations
205 2. Annotation of all genes detected during *in vitro* neuronal differentiation.
206 3. Annotation of interest (e.g. cluster membership).

207 Stratified LDSR defines enrichment of h^2 of an annotation of interest (3) as the
208 proportion of h^2 explained by a category divided by the proportion of SNPs in that category.
209 To determine if this enrichment is significant and specific to this annotation, it estimates the
210 contribution of that annotation to the per-SNP h^2 while accounting for the baseline and the all
211 genes detected annotation (1 + 2). As we only test for enrichment, we report the contribution
212 to the per-SNP h^2 ( ) and the associated one-sided p-value, which is calculated using
213 standard errors that are obtained via a block jackknife procedure[8, 33].
214
215
216
217
218
219
220
221
222
223
224

225 **Results**

226 ***Longitudinal in vitro gene expression profiling confirms neuron-specific***
227 ***differentiation and matches in vivo human cortical development***

228 To study the molecular dynamics underlying *in vitro* human neuronal differentiation,
229 we differentiated an hNSC line (WA09/H9) to a neuronal lineage across 30 days. Genome-
230 wide gene expression profiles were assayed densely at seven time points in at least
231 triplicates (n=24 samples). Principle component analysis (PCA) on normalized gene
232 expression values shows a large proportion of the variance in expression to be explained by
233 the differentiation process, with minimal effects of technical variation (Figure 1A & S1).
234 Investigation of cell type-specific gene expression signatures of major classes of cell types in
235 the cerebral cortex showed that relative neuronal gene expression increases as neuronal
236 differentiation progresses over time (Figure 1B). There is no evidence of glial- or endothelial-
237 specific gene expression, which confirms a broadly neuronal *in vitro* cellular identity.

238

239 ***[Figure 1 about here]***

240

241 Having established that the *in vitro* differentiation process is predominantly neuronal,
242 we applied transition mapping (TMAP) to assess the correspondence of longitudinal *in vitro*
243 transcriptome data to *in vivo* signatures of human cortical development. TMAP uses a
244 spatiotemporal transcriptome atlas of the human neocortex and laminar expression data to
245 assess global overlap in differential gene expression (DGE) profiles between *in vitro* time
246 points and *in vivo* brain developmental stages or laminae of the human neocortex. We find
247 significant matching between the *in vitro* longitudinal DGE profiles (day-0 vs day-30) and *in*
248 *vivo* developmental stage from 4 weeks post-conception (PCW) to 24 PCW (Figure S2). This
249 overlaps with the primary period of neurogenesis in the neocortex, which starts around 6
250 PCW[34, 35]. To gain more insight into this overlap, we partitioned the TMAP analyses in three
251 comparisons and examined how *in vitro* to *in vivo* matching progressed over time across *in*
252 *vitro* neuronal differentiation. We see a clear progression in matching from early

253 developmental stages to later stages (Figure 2). For example, *in vitro* day-0 vs day-5 show
254 strong overlap with *in vivo* period-1 (4-8 PCW) vs period-4 (13-16 PCW), while *in vitro* day-
255 15 vs day-30 shows stronger overlap with *in vivo* period-2 (8-10 PCW) vs period-8 (birth-
256 6M). Similarly, *in vitro* longitudinal DGE shows progression from overlap of early time points
257 with inner laminae, to overlap with more upper cortical layers as *in vitro* neuronal
258 differentiation advances (Figure 2 and S2).
259
260 ***In vitro neuronal differentiation reveals specific longitudinal gene clusters***
261 To identify biological pathways associated with neuronal differentiation, we applied
262 an analysis framework specifically tailored to time-series gene expression data (see
263 Methods and Supplementary Methods). A total of 7,734 probes, mapping to 5,818 genes,
264 were differentially expressed over time (Figure S3). We find that these genes are, on
265 average, more constraint to genetic variation compared to non-differentially expressed
266 genes (Supplementary Results). Using only differentially expressed probes, we next applied
267 fuzzy c-means clustering and identified eight distinct longitudinal gene clusters (Figure 2 and
268 S4). For each probe, we generated a corresponding cluster membership value, representing
269 the degree to which a gene belongs to a cluster. To identify most informative biological
270 interpretation of each cluster, we analyzed genes with high cluster membership for
271 enrichment of functional annotations using DAVID (Supplementary Methods and Table S1).
272
273 ***[Figure 2 about here]***
274
275 We identified three clusters with decreasing gene expression over time that are
276 significantly enriched for cell division and RNA regulation and processing genes, reflective of
277 stem cell proliferation and cell fate determination that is tightly controlled and regulated by
278 RNA dependent processes[36]. Second, there are three clusters showing increased gene
279 expression levels over time that are primarily enriched for neuronal processes, such as
280 neuron formation and synaptic function. Another independent cluster shows an inverted U-

281 shaped expression pattern during development, enriched for genes involved in
282 transcriptional regulation. The final cluster is enriched for genes involved in extracellular
283 region and cell adhesions. These processes are important for cell connectivity and have also
284 been implicated in cell proliferation and neuronal migration[37, 38]. Together, these eight gene
285 clusters reveal different biological mechanisms that are associated with neuronal
286 differentiation and consistent with known biology of neurodevelopment. We hypothesize that
287 the study of these longitudinal gene expression clusters can help decipher disease
288 mechanisms involved in psychiatric phenotypes.
289
290 ***Differentially expressed genes are associated with GWAS disease risk of***
291 ***schizophrenia***
292 To investigate how aggregate psychiatric disease risk is distributed across genes that
293 are important for neuronal differentiation, we applied gene-set analysis and partitioning of h^2
294 with MAGMA and sLDSR, respectively. We examined the distribution of genetic risk for
295 major psychiatric disorders with GWAS results from large-scale studies. See Supplementary
296 Table S2 for details on all included phenotypes. For MDD, we included GWAS results from
297 the China Oxford and VCU Experimental Research on Genetic Epidemiology (CONVERGE)
298 consortium[23] and 23andMe Inc., a personal genetics company[11]. The latter uses SRD as a
299 proxy for major depression. Alzheimer's disease (AD) and adult human height served as
300 non-psychiatric control phenotypes that are heritable and polygenic. We used a two-step
301 approach where we first investigated disease associations on overall differential expression
302 level and subsequently proceeded to deconstruct these associations across the longitudinal
303 gene clusters. Together, these analyses allow us to integrate *in vitro* longitudinal
304 transcriptomic signatures with polygenic disease risk and assess if our model is relevant to
305 study the etiology of psychiatric disorders.
306 We find that genes that are differentially expressed across *in vitro* neuronal
307 differentiation are enriched for multiple psychiatric disorders. We find significant MAGMA
308 enrichment for SCZ (P=0.001), ADHD (P=0.002), and SRD (P=0.003) (Table 1 and Table

309 S3). With sLDSR, we find nominally significant enrichment for SCZ (P=0.01) and SRD
310 (P=0.02) and a suggestive association for ADHD (P=0.06) (Table 1 and Table S4). We
311 observed suggestive enrichment for BPD, and no enrichment for the cross disorder, ASD,
312 MDD CONVERGE or for adult height and AD.

313 We next investigated whether enrichment of differentially expressed genes was
314 driven by up- or downregulation of genes during differentiation. We find that the enrichment
315 of SCZ across differentially expressed genes is driven by genes that are upregulated
316 (MAGMA $P=5.0x10^{-7}$, sLDSR $P=6.1x10^{-5}$) and not by genes that are downregulated
317 (MAGMA P=0.98, sLDSR P=0.61) (Figure 3 and Figure S5). For SRD, we only find a
318 stronger enrichment in upregulated genes with MAGMA ($P=3.5x10^{-4}$), while ADHD shows no
319 evidence for enrichment driven by either up or downregulated genes.

320

321 ***[Table 1 about here]***

322

323 ***SCZ GWAS disease risk aggregates to specific temporal gene clusters***

324 Next, we explored the relationship between differentially expressed genes and
325 disease risk on cluster level. For this analysis, we only included traits that show significant
326 disease enrichment across differentially expressed genes using MAGMA after correcting for
327 multiple testing (SCZ, ADHD, SRD) and our control traits (AD, height). These disease traits
328 showed at least suggestive enrichment with sLDSR as well. Using both MAGMA and
329 sLDSR, we integrated cluster membership values with GWAS summary statistics (n=5) and
330 assessed whether genome-wide disease risk aggregates to any of the eight experimentally
331 identified longitudinal gene clusters. Overall, MAGMA and sLDSR show a strong
332 concordance across phenotypes and clusters (rho = 0.92, $p<2.2x10^{-16}$, n=40, see also Figure
333 S6). After Bonferroni correction (n=40), we find five significant phenotype-cluster
334 associations with MAGMA and three with sLDSR (Figure 4 and Table S5/S6).

335

336 We find that multiple upregulated clusters show enrichment for SCZ with the
337 strongest evidence for the *synaptic function* cluster (MAGMA $P=1.8x10^{-7}$, sLDSR $P=7.2x10^{-}$
338 ^{5}) (see Figure S7). For SRD, we find significant associations in the *transcription regulation*
339 ($P=2.5x10^{-5}$) and the *neuron formation* ($P=1.2x10^{-4}$) gene cluster with MAGMA only. While
340 the analysis of adult height using all differentially expressed genes did not yield any
341 evidence for enrichment of genetic signal, enrichment is observed at the cluster level. The
342 *cell connectivity* cluster ($P=3.7x10^{-4}$) is enriched for height, in addition to suggestive
343 enrichments in the *cell division* and *RNA regulation* cluster, which are not present for any of
344 the psychiatric phenotypes. Remarkably, across all 8 clusters the enrichments of SCZ and
345 height are inversely correlated (rho=-0.85, P=0.011, n=8, see also Figure S8).
346
347 ***[Figure 4 about here]***
348
349 Finally, in order to take into account the full spectrum of correlations and
350 dependencies between clusters (Figure S9), we performed a conditional analysis for SCZ,
351 the trait for which the strongest cluster enrichments are observed with both methods. Using
352 the same MAGMA model, for each cluster, we conditioned on the highest gene members
353 (membership > 0.5) of the other seven clusters (Table 2). We find that the SCZ enrichment
354 is driven by the *synaptic function* cluster ($p=2.88x10^{-3}$) only. The same conditional analysis
355 for SRD, which only showed significant enrichment with MAGMA, shows that this enrichment
356 is primarily driven by the *transcription regulation* cluster ($p=5.42x10^{-3}$) (Table S7).
357
358 ***[Table 2 about here]***
359
360
361
362
363

364 **Discussion**

365 We investigated a longitudinal *in vitro* stem cell model of human neuronal differentiation
366 to study psychiatric disease susceptibility based on evidence from GWAS. Among five major
367 psychiatric disorders, we observe that SCZ disease susceptibility is significantly enriched in
368 a set of genes relevant to *synaptic functioning* that are upregulated during differentiation. We
369 therefore propose *in vitro* human neuronal differentiation as an experimental system to
370 further understand and decipher SCZ disease biology.

371 We confirmed that our *in vitro* model recapitulates neuronal signatures of *in vivo*
372 cortical development across specific developmental time periods and laminae of the human
373 neocortex. This is in line with previous findings[14] and highlights that longitudinal gene
374 expression dynamics underlying our model of human neuronal differentiation can be
375 informative to study genes and pathways involved in *in vivo* human cortical development.
376 SCZ is such a disorder whose susceptibility has been hypothesized to lie in neuronal cell
377 types[39, 40] and in early brain development[7, 22, 41]. Here, we observe that genes differentially
378 expressed across neuronal differentiation are significantly enriched for genome-wide disease
379 risk of SCZ and that this risk mainly aggregates in genes involved in synaptic functioning
380 during development. Although not the only pathogenic process contributing to SCZ, synaptic
381 dysfunction is most strongly supported by genetic data, postmortem expression studies, and
382 animal models[40, 42–45]. We are the first to provide evidence for this hypothesis using a
383 longitudinal *in vitro* cell-based model and genome-wide disease risk. Two of the highest
384 gene members of the *synaptic function* gene cluster enriched for SCZ include Calcium
385 Voltage-Gated Channel Subunit Alpha 1C (*CACNA1C*) and Solute Carrier Family 45
386 Member 1 (*SLC45A1*), both located at a genome-wide significant SCZ locus[22]. We find no
387 evidence of enrichment for AD, a late-onset non-psychiatric brain disease, nor for adult
388 human height in this neuronal cluster. Together, our findings demonstrate that longitudinal
389 transcriptomic signatures important for neuronal differentiation recapitulate the *in vivo*
390 context *and* align with the genetic basis of the disease. SCZ disease biology can thus be
391 studied through these molecular processes captured by this *in vitro* model.

392 We also observed a significant enrichment of genetic signal with MAGMA for SRD in
393 genes upregulated during differentiation, and show that this enrichment is predominantly
394 driven by genes in the *transcription regulation* gene cluster. Interestingly, the SRD GWAS
395 reported that the top SNPs were enriched for transcription regulation related to
396 neurodevelopment[11], which is in line with our *in vitro* findings. We observed no enrichment
397 for GWAS signal from recurrent and severe MDD in Han-Chinese women[23]. The latter
398 sample represents the most genetically and phenotypically homogeneous GWAS of MDD.
399 The fact that for these results no enrichment for any of our gene sets was observed may
400 suggest that neurodevelopmental processes play a lesser role in MDD[46]. Alternatively, larger
401 sample sizes are needed to better capture the genome-wide genetic risk associated with
402 MDD. Self-reported depression is a much broader phenotype that may include other
403 psychiatric traits, which could drive the observed neurodevelopment and transcription
404 enrichments. It is therefore unclear how our findings and the application of the model
405 extrapolate to the MDD phenotype.

406 We did not find any evidence of significant association in the neuronal clusters for
407 ADHD. This could be due to the smaller sample size in the GWAS studies and thereby lack
408 of power to find a significant association with our transcriptomic signatures (Figure S10). As
409 GWAS sample sizes are expected to increase, these gene cluster associations should be
410 revisited.

411 For height, we found enrichment in opposite direction of psychiatric traits in the
412 downregulated gene clusters. Strikingly, we observe an inverse correlation between SCZ
413 and height enrichment stratified across gene clusters (Supplementary Results, Figure S8
414 and S11), despite the absence of any evidence of a genetic correlation across the whole-
415 genome[47]. These observations not only illustrate the added value of individual longitudinal
416 gene clusters, but also highlight a complex genetic relationship between these two
417 phenotypes.

418 A strength of our approach is the longitudinal analysis framework that we developed.
419 We implemented an experimental design across a dense and repeatedly sampled time-

series and integrated longitudinal transcriptomic signatures with genome-wide disease risk using available GWAS summary statistics. This increases statistical power to directly investigate the role of disease variants on genes important to our model system. While we specifically chose to perform our experiments across an isogenic background to minimize variation and maximize statistical power to identify transcriptomic signatures, our framework can easily be extended to a multi-sample design (e.g. cases vs controls)[19, 48], which makes it relevant for many disease-specific experimental settings.

Our experimental procedure applied differentiation towards a broad neuronal phenotype. Our work does not exclude disease associations with specific subtypes of neuronal cells nor with other major brain cell types. We provide a proof-of-concept of an *in vitro* model of neuronal cells for studying complex diseases, such as SCZ, and present an analytical framework that includes longitudinal assessment of gene expression profiles. This approach can readily be extended to study *in vitro* differentiation of other major brain cell types, such as astrocytes or oligodendrocytes. Although we show strong evidence for SCZ risk in early prenatal neurodevelopment, our findings do not preclude an additional contribution of postnatal neurodevelopment to the etiology of the disease[49–51].

In summary, the current study establishes WA09 neuronal differentiation as an *in vitro* genomic tool to study SCZ. Overall, this work contributes to understand the functional mechanisms that underlie psychiatric disease heritability and polygenicity in the post GWAS era. Our work highlights specific gene clusters involved in disease susceptibility during development and thereby provides a framework for experimental and analytical follow-up functional analyses in a model that is robust with standardized experimental procedures. These can now be extended to incorporate model perturbations, such as genomic manipulation, to study disease processes in finer detail across an isogenic background in a controlled environment. Our findings suggest that *in vitro* longitudinal transcriptomic signatures across neuronal differentiation could then serve as a functional readout to investigate schizophrenia.

448 **Data Availability**

449 The Illumina HT-12 v4 gene expression data is available through the Gene
450 Expression Omnibus (GEO) archive (*Accession number is available for review*). This dataset
451 has the raw and normalized gene expression values on all samples. Supplementary table 8
452 furthermore has specific probe annotations, such as probabilities of differential expression
453 and probe membership values for all identified clusters.

454

455 **Author Contribution:**

456 The project was led by R.O. Experiments were designed and conceived by A.O. and
457 R.O. Experiments were optimized, conducted and samples processed by A.O., M.B., and
458 R.M. Analysis of the data was performed by A.O and M.B and feedback provided by L.O.
459 and R.O. The main findings were interpreted by A.O., M.B., R.M., L.O., and R.O. Primary
460 drafting of the manuscript was performed by A.O. and main feedback provided by R.O and
461 L.O. All authors contributed to the production and approval of the final manuscript.

462

463 **Conflict of Interest**

464 The authors declare no competing interests.

465

466 **Acknowledgement**

467 We thank all research participants and researchers involved in making each GWAS
468 summary statistic available and this work possible, including the 23andMe Research Team.
469 We thank C. de Leeuw for his helpful input and troubleshooting with MAGMA analyses and
470 thank the LD score regression team for their input and helpful troubleshooting with stratified
471 LDSR. This research was supported by NIH/NIMH R01 MH090553.

472

473

474

475

476 **References**

477 1. Geschwind DH, Flint J. Genetics and genomics of psychiatric disease. *Science* 2015;
478 **349**: 1489–1494.

479 2. Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher
480 PM *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin
481 studies. *Nat Genet* 2015; **47**: 702–709.

482 3. Sullivan PF, Agrawal A, Bulik C, Andreassen OA, Borglum A, Breen G *et al.* Psychiatric
483 Genomics: An Update and an Agenda. Am J Psychiatry 2017 (ahead of print).

484 4. Falk A, Heine VM, Harwood AJ, Sullivan PF, Peitz M, Brüstle O *et al.* Modeling
485 psychiatric disorders: from genomic findings to cellular phenotypes. *Mol Psychiatry*
486 2016; 1167–1179.

487 5. Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S *et al.* Spatial and
488 temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical
489 network. *Cell* 2013; **154**: 518–529.

490 6. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P *et al.* A polygenic
491 burden of rare disruptive mutations in schizophrenia. *Nature* 2014; **506**: 185–190.

492 7. Olde Loohuis LMO, Vorstman JAS, Ori AP, Staats KA, Wang T, Richards AL *et al.*
493 Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nat*
494 *Commun* 2015; **6**: 7501.

495 8. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R *et al.* Partitioning
496 heritability by functional annotation using genome-wide association summary statistics.
497 *Nat Genet* 2015; **47**: 1228–1235.

498 9. Geschwind DH. Genetics of autism spectrum disorders. *Trends in Cognitive Sciences*.
499 2011; **15**: 409–416.

500 10. Rubeis SD, He X, Goldberg AP, Poultney CS, Samocha K. Synaptic, transcriptional,
501 and chromatin genes disrupted in autism A. *Nature* 2014; **515**: 209–215.

502 11. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR *et al.* Identification of
503 15 genetic loci associated with risk of major depression in individuals of European
504 descent. *Nature Publishing Group* 2016; **48**: 1031–1036.

505 12. Shi Y, Kirwan P, Smith J, Robinson HPC, Livesey FJ. Human cerebral cortex
506 development from pluripotent stem cells to functional excitatory synapses. *Nat Neurosci*
507 2012; **15**: 477–486.

508 13. van de Leemput J, Boles NC, Kiehl TR, Corneo B, Lederman P, Menon V *et al.*
509 CORTECON: A temporal transcriptome analysis of in vitro human cerebral cortex
510 development from human embryonic stem cells. *Neuron* 2014; **83**: 51–68.

511 14. Stein JL, de la Torre-Ubieta L, Tian Y, Parikshak NN, Hernández IA, Marchetto MC *et*
512 *al.* A quantitative framework to evaluate modeling of cortical development by neural
513 stem cells. *Neuron* 2014; **83**: 69–86.

514 15. Du P, Kibbe W a., Lin SM. lumi: a pipeline for processing Illumina microarray.
515 *Bioinformatics* 2008; **24**: 1547–1548.

516 16. Lin SM, Du P, Huber W, Kibbe W a. Model-based variance-stabilizing transformation for
517 Illumina microarray data. *Nucleic Acids Res* 2008; **36**: 1–9.

518 17. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S *et al.* An RNA-
519 Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells
520 of the Cerebral Cortex. *Journal of Neuroscience* 2014; **34**: 11929–11947.

521 18. Tai YC, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time
522 course data. *Ann Stat* 2006; **34**: 2387–2412.

523 19. Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J. An improved
524 empirical bayes approach to estimating differential gene expression in microarray time-
525 course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*
526 2009; **10**: 409.

527 20. Kumar L, E Futschik M. Mfuzz: a software package for soft clustering of microarray
528 data. *Bioinformation* 2007; **2**: 5–7.

529 21. Schwämmle V, Jensen ON. A simple and fast method to determine the parameters for
530 fuzzy c-means cluster analysis. *Bioinformatics* 2010; **26**: 2841–2848.

531 22. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological
532 insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014; **511**: 421–427

533 23. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for
534 major depressive disorder. *Nature* 2015; **523**: 588.

535 24. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-
536 wide association analysis of bipolar disorder identifies a new susceptibility locus near
537 ODZ4. *Nat Genet* 2011; **43**: 977–983.

538 25. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics
539 Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum
540 disorder highlights a novel locus at 10q24.32 and a significant overlap with
541 schizophrenia. *Mol Autism* 2017; **8**: 21.

542 26. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E *et al.* Discovery of
543 the first genome-wide significant risk loci for ADHD. *bioRxiv* 2017;
544 https://doi.org/10.1101/145581

545 27. Consortium C-DG of TPG. Identification of risk loci with shared effects on five major
546 psychiatric disorders: a genome-wide analysis. *Lancet* 2013; **381**: 1371–1379.

547 28. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C *et al.* Meta-
548 analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's
549 disease. *Nat Genet* 2013; **45**: 1452–1458.

550 29. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S et al. Defining the role
551 of common variation in the genomic and biological architecture of adult human height.
552 *Nat Genet* 2014; **46**: 1173–1186.

553 30. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP,
554 Kang HM *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–
555 74.

556 31. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set
557 Analysis of GWAS Data. *PLoS Comput Biol* 2015; **11**.

558 32. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X *et al.* Linkage
559 disequilibrium–dependent architecture of human complex traits shows action of negative
560 selection. *Nat Genet* 2017.

561 33. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Consortium SWG of TPG *et*
562 *al.* LD Score regression distinguishes confounding from polygenicity in genome-wide
563 association studies. *Nat Genet* 2015; **47**: 291–295.

564 34. Clancy B, Darlington RB, Finlay BL. Translating developmental time across mammalian
565 species. *Neuroscience* 2001; **105**: 7–17.

566 35. Stiles J, Jernigan TL. The basics of brain development. *Neuropsychology Review*. 2010;
567 **20**: 327–348.

568 36. Hattori A, Buac K, Ito T. Regulation of Stem Cell Self-Renewal and Oncogenesis by
569 RNA-Binding Proteins. In: RNA Processing Disease and Genome-wide probing. 2016.
570 p. 153–188.

571 37. Barros CS, Franco SJ, Muller U. Extracellular Matrix: Functions in the nervous system.
572 *Cold Spring Harb Perspect Biol* 2011; **3**: 1–24.

573 38. Bikbaev A, Frischknecht R, Heine M. Brain extracellular matrix retains connectivity in
574 neuronal networks. *Sci Rep* 2015; **5**: 14527.

575 39. Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar H *et al.* Genetic
576 Identification Of Brain Cell Types Underlying Schizophrenia. bioRxiv 2017;
577 http://dx.doi.org/10.1101/145466

578 40. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M *et al.*
579 Increased burden of ultra-rare protein-altering variants among 4,877 individuals with
580 schizophrenia. *Nat Neurosci* 2016; **19**: 1433–1441.

581 41. Finucane H, Reshef Y, Anttila V, Slowikowski K, Gusev A, Byrnes A *et al.* Heritability
582 enrichment of specifically expressed genes identifies disease-relevant tissues and cell
583 types. *bioRxiv* 2017;https://doi.org/10.1101/103069

584 42. Hall J, Trent S, Thomas KL, O'Donovan MC, Owen MJ. Genetic risk for schizophrenia:
585 Convergence on synaptic pathways involved in plasticity. *Biological Psychiatry*. 2015;
586 **77**: 52–58.

587 43. Lips ES, Cornelisse LN, Toonen RF, Min JL, Hultman CM, Holmans P a. *et al.*
588 Functional gene group analysis identifies synaptic gene groups as risk factor for
589 schizophrenia. *Mol Psychiatry* 2011; **4**: 1–11.

590 44. Pocklington AJ, O'Donovan M, Owen MJ. The synapse in schizophrenia. *Eur J Neurosci*
591 2014; **39**: 1059–1067.

592 45. Schwarz E, Izmailov R, Lio P, Meyer-Lindenberg A. Protein Interaction Networks Link
593 Schizophrenia Risk Loci to Synaptic Function. *Schizophr Bull* 2016; **42**: 1334–1342.

594 46. Peterson RE, Cai N, Bigdeli TB, Li Y, Reimers M, Nikulova A *et al.* The Genetic
595 Architecture of Major Depressive Disorder in Han Chinese Women. *JAMA Psychiatry*
596 2017; **74**: 162–168.

597 47. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R *et al.* An Atlas
598 of Genetic Correlations across Human Diseases and Traits. *Nat Genet* 2015; **47**: 1237–
599 1241.

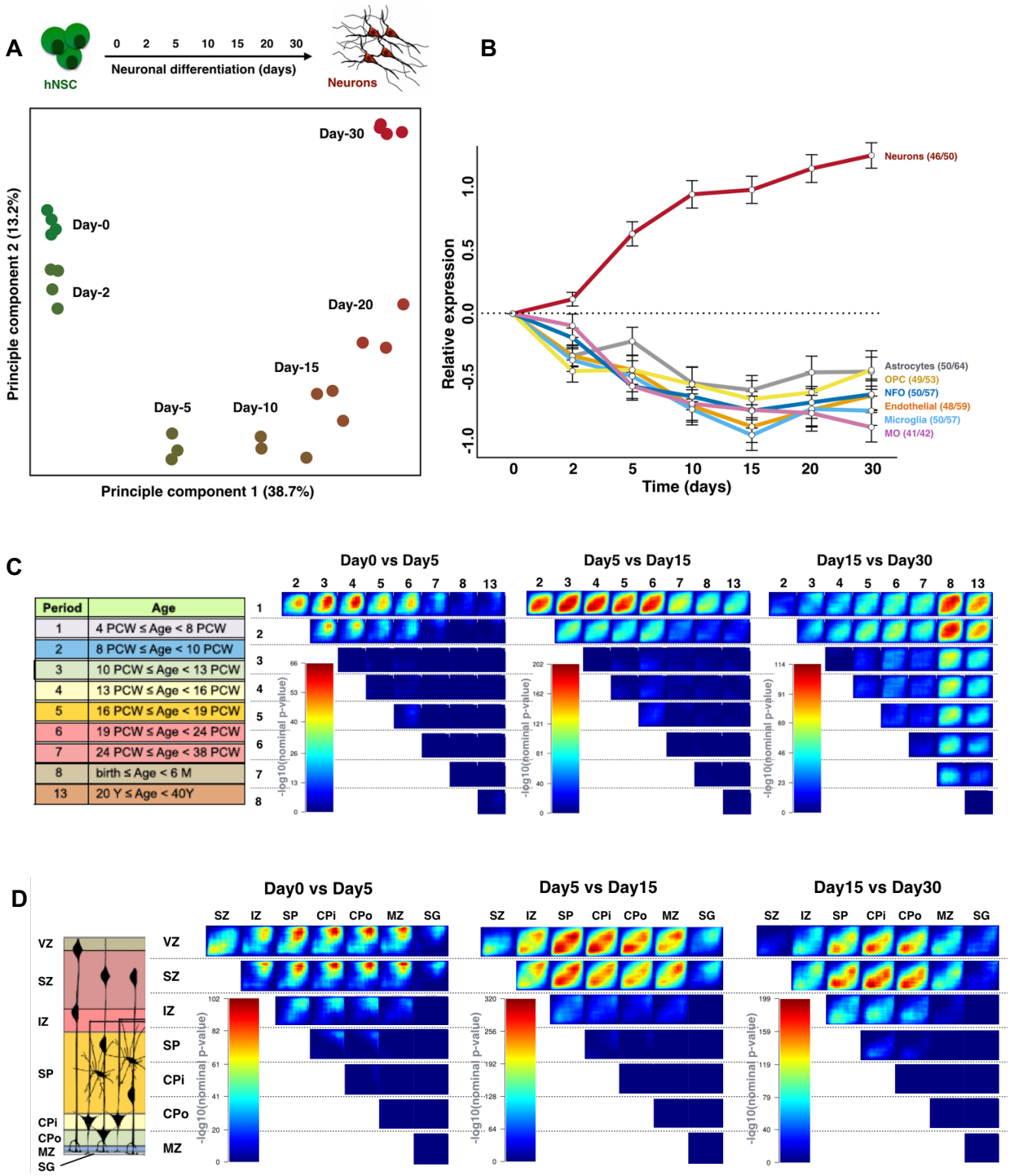
600 48. Tai YC, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time
601 course data. *Ann Stat* 2006; **34**: 2387–2412.

602 49. Birnbaum R, Jaffe AE, Hyde TM, Kleinman JE, Weinberger DR. Prenatal expression
603 patterns of genes associated with neuropsychiatric disorders. *Am J Psychiatry* 2014;
604 **171**: 758–767.

605 50. Pers TH, Timshel P, Ripke S, Lent S, Sullivan PF, O'Donovan MC *et al.* Comprehensive
606 analysis of schizophrenia-associated loci highlights ion channel pathways and
607 biologically plausible candidate causal genes. *Hum Mol Genet* 2015; **25**: 1247–1254.

608 51. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N *et al.*
609 Schizophrenia risk from complex variation of complement component 4. *Nature* 2016;
610 **530**: 177–183.

611



Period	Age
1	4 PCW ≤ Age < 8 PCW
2	8 PCW ≤ Age < 10 PCW
3	10 PCW ≤ Age < 13 PCW
4	13 PCW ≤ Age < 16 PCW
5	16 PCW ≤ Age < 19 PCW
6	19 PCW ≤ Age < 24 PCW
7	24 PCW ≤ Age < 38 PCW
8	birth ≤ Age < 6 M
13	20 Y ≤ Age < 40Y

Figure 1. *In vitro* gene expression profiles confirm a neuron-specific differentiation process and match *in vivo* human cortical development. (A) PCA of *in vitro* transcriptome data with PC1 (x-axis) and PC2 (y-axis) visualized. Variance explained per component is shown in parentheses. Time points are color-coded and labeled by days across differentiation. (B) Cellular identity is shown by average expression of cell type specific genes across days of differentiation. Cell types are highlighted by their name and corresponding color. The first number in the parentheses represents the number of genes for which the average expression is plotted.

644 The second number represents the corresponding number of probes assayed. OPC =
645 oligodendrocyte precursor cells, NFO = newly formed oligodendrocytes, MP = myelinating
646 oligodendrocytes. (C+D) TMAP output visualizes the amount of overlap between *in vitro* and *in*
647 *vivo* DGE profiles colored by –log10(p-value) (see figure S2 for more details on interpretation).
648 Abbreviations and numbering above maps correspond to schematic representations on the left
649 (adopted from Stein et al., 2014) of different developmental stages (C) and laminae (D). VZ =
650 ventricular zone, SZ = subventricular zone, IZ = intermediate zone, SP=subplate zone, CPi=
651 inner cortical plate, CPo = outer cortical plate, MZ = marginal zone, PCW = post conception
652 weeks, M = months, Y = years, Period = developmental stage.

	Genes (n)	Functional annotations	Enrichment	Cluster label
1	737	Mitotic cell cycle and division Chromosome and kinetochore DNA repair	23.90 17.41 15.67	Cell division
2	457	RNA binding and ribonucleoproteins RNA processing and ribosome biogenesis RNA splicing	15.77 10.05 6.18	RNA regulation
3	221	RNA processing and binding ncRNA processing *Regulation of cell cycle*	4.25 3.40 *2.13*	RNA processing
4	891	Neuronal outgrowth Synaptic formation Neuron differentiation	6.29 4.60 4.23	Neuron formation
5	779	Neuronal projections Synaptic signaling Synapse	16.85 13.56 10.16	Synaptic function
6	253	Membrane-bound vesicles Glycoprotein signaling Transmembrane region	4.17 3.95 2.72	Cell signaling
7	396	Regulation of transcription Transcription factor activity Zinc finger and ion binding	4.21 3.02 2.97	Transcription regulation
8	584	Extracellular region Cell junctions and adhesion Regulation of cell motility	8.20 7.78 7.19	Cell connectivity

Figure 2. Identified gene clusters highlight biological pathways important for neuronal differentiation. Significant functional annotations and corresponding enrichment score are shown for each gene cluster. Longitudinal gene expression is visualized for high member genes only (black line represents mean gene expression). Each cluster is color-coded with the number of genes at membership > 0.5 denoted. See table S1 for full annotation results.

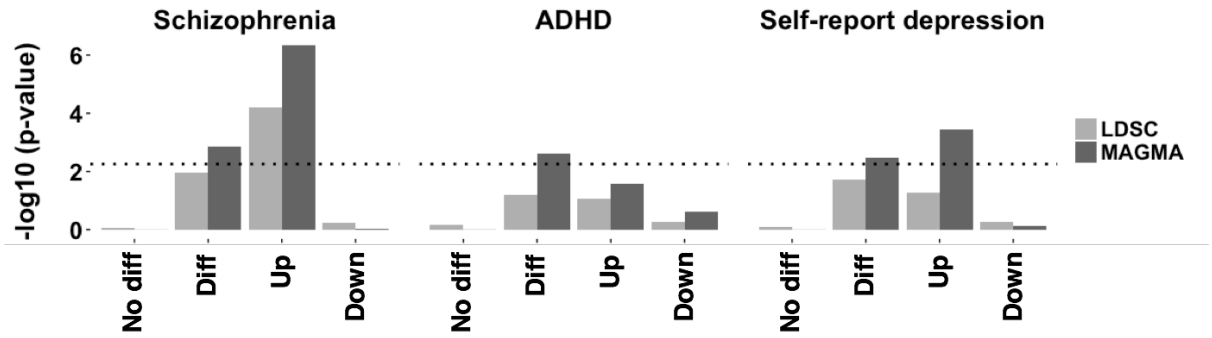


Figure 3. Schizophrenia GWAS enrichment lies in genes up-regulated during neuronal differentiation. A more detailed investigation of the enrichment of h^2 of SCZ, ADHD, and self-reported depression across differentially expressed genes. The y-axis denotes the –log10 P-value of the enrichment. No diff = genes that are not differentially expressed; Diff = log (T^2-statistic) as shown in Table 1; Up = genes up-regulated during differentiation; Down = genes down-regulated during differentiation. The dotted line represents the threshold for P=0.0056 (n=9 traits).

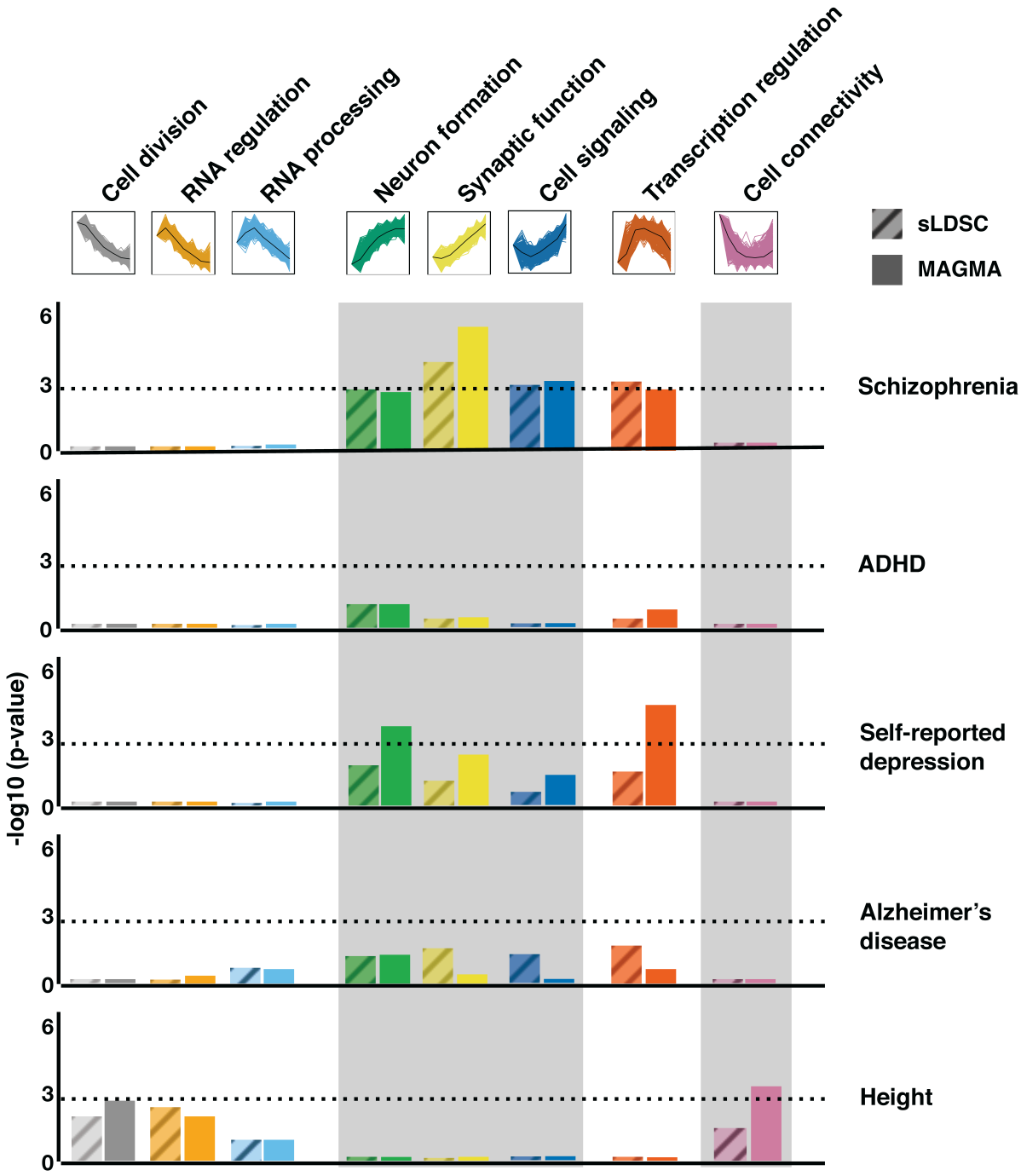


Figure 4. Psychiatric GWAS enrichment is distributed across specific longitudinal gene clusters. Results from sLDSC (diagonal pattern) and MAGMA (solid colors) are shown for each phenotype (labels on the right) colored by gene cluster. Gene cluster annotation and cluster expression pattern are shown on top. The y-axis states the –log10 (p-value). The dotted horizontal line represents the threshold for Bonferroni correction (p=0.05/40).

	MAGMA			sLDSC	
Phenotype	**Beta (SE)**	**Beta_std**	**P-value**	**τ (SE)**	**P-value**
Psychiatric					
Schizophrenia	0.022 (0.007)	0.094	**0.001**	1.70×10^{-9} (7.45×10^{-10})	0.01
ADHD	0.014 (0.005)	0.059	**0.002**	1.92×10^{-9} (1.25×10^{-9})	0.06
Self-reported depression	0.013 (0.005)	0.057	**0.003**	4.34×10^{-10} (2.10×10^{-10})	0.02
Bipolar disorder	0.007 (0.005)	0.032	0.06	6.16×10^{-9} (3.64×10^{-9})	0.05
Cross disorder	0.005 (0.005)	0.020	0.16	1.19×10^{-9} (1.00×10^{-9})	0.12
MDD CONVERGE	0.000 (0.004)	-0.001	0.51	6.07×10^{-9} (4.39×10^{-9})	0.08
ASD	0.000 (0.004)	-0.002	0.54	2.97×10^{-9} (3.48×10^{-9})	0.20
Neurodegenerative					
Alzheimer's disease	0.003 (0.004)	0.015	0.22	1.30×10^{-10} ($1.02 \times 10^{-9)}$	0.45
Non-brain					
Height	0.009 (0.011)	0.037	0.21	-1.62×10^{-9} ($1.36 \times 10^{-9)}$	0.88

Table 1. Polygenic psychiatric disease enrichment across differentially expressed genes. Shown are enrichment results for MAGMA and sLDSR for gene differential expression. P-values highlighted in bold show phenotypes that survive multiple testing correction (n=9). See Table S3 and S4 for more details. Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given a change of one standard deviation in log T2 statistic, τ (tau) = the contribution to the per-SNP h^2.

	MAGMA Primary		MAGMA Conditional	
Schizophrenia - clusters	**Beta (SE)**	**P-value**	**Beta (SE)**	**P-value**
Cell division	-0.045 (0.017)	1.00	-0.047 (0.027)	0.96
RNA regulation	-0.040 (0.017)	0.99	-0.044 (0.027)	0.95
RNA processing	-0.006 (0.017)	0.64	-0.011 (0.024)	0.68
Neuron formation	0.048 (0.017)	2.12×10^{-3}	0.018 (0.036)	0.30
Synaptic function	**0.077 (0.017)**	$\mathbf{1.82 \times 10^{-6}}$	**0.070 (0.026)**	$\mathbf{2.88 \times 10^{-3}}$
Cell signaling	0.052 (0.016)	6.88×10^{-4}	0.032 (0.023)	0.08
Transcription regulation	0.048 (0.016)	1.67×10^{-3}	0.019 (0.025)	0.22
Cell connectivity	-0.061 (0.017)	1.00	-0.076 (0.026)	1.00

Table 2. Schizophrenia MAGMA cluster conditional analysis. Gene level association signal is regressed on cluster membership while adjusting for high membership genes of all other seven clusters. Shown are the results of the primary analysis (not adjusted for other clusters) and the conditional analysis. Beta = regression coefficient, SE = standard error.

Supplementary Information

A longitudinal model of human neuronal differentiation for functional investigation of schizophrenia disease susceptibility

Authors:

Anil P. S. Ori, Merel H. M. Bot, Remco T. Molenhuis, Loes M. Olde Loohuis, Roel A. Ophoff

SUPPLEMENTARY MATERIAL AND METHODS 2

APPROVAL FOR STEM CELL RESEARCH 2
HUMAN NEURAL STEM CELL LINE 2
EXPERIMENTAL DESIGN AND RNA EXTRACTION 2
DATA PREPROCESSING AND QUALITY CONTROL 3
IN VITRO CELLULAR IDENTITY 3
TRANSITION MAPPING TO A SPATIOTEMPORAL ATLAS OF EARLY HUMAN BRAIN DEVELOPMENT 4
TIME-SERIES DIFFERENTIAL GENE EXPRESSION ANALYSIS 4
FUZZY C-MEANS CLUSTER ANALYSIS 5
FUNCTIONAL ANNOTATION OF CLUSTERS 6
INTOLERANCE OF LOSS-OF-FUNCTION VARIATION ACROSS CLUSTERS 6
GWAS SUMMARY STATISTICS USED 7
STRATIFIED LD SCORE REGRESSION - GENERATING ANNOTATION FILES AND LD SCORES 7

SUPPLEMENTARY RESULTS 9

UPREGULATED GENES ARE MORE LIKELY TO BE INTOLERANT TO LOSS-OF-FUNCTION FUNCTIONAL VARIATION 9
CLUSTER ENRICHMENTS OF SCHIZOPHRENIA AND HEIGHT ARE INVERSELY CORRELATED 9

REFERENCES 11

Supplementary Material and Methods

Approval for stem cell research

The University of California, Los Angeles Embryonic Stem Cell Research Oversight (ESCRO) committee approved this work. Their policy is based on the recommendations of the National Bioethics Advisory Commission, the National Academies of Science-Institute of Medicine guidelines, and standards created by the California Institute for Regenerative Medicine.

Human neural stem cell line

WA09(H9)-derived hNSC is a commercially available and commonly used neural stem line with standardized and well-documented neuronal differentiation protocols[1–3]. These cells originate from a donated human embryo (F), produced by *in vitro* fertilization for clinical purposes, that was cultured to a blastocyst after which an ESC line was established[4, 5]. This cell line is of European ancestry[6] and has a normal karyotype. It was in addition successfully tested for stem cell characteristics and approved by NIH for stem cell research[7]. WA09 ESCs were differentiated to NSCs by the vendor and obtained by us as neural progenitors. Tissue culture plates were coated with CELLstart CTS™ (Thermo Fisher Scientific) diluted (1:50) in DPBS with Ca^{2+} and Mg^{2+} and hNSCs cells expanded in KnockOut™ DMEM/F-12 Basal Medium (Gibco) with 2% StemPro® Neural Supplement (Gibco), 2mM GlutaMax™-I Supplement (Gibco), FGF Basic and EGF Recombinant proteins (Gibco, both at 20 ng/ml), and 1x Pen Strep (Thermo Fisher Scientific). Cells were plated at 1.0×10^5 cells per 3.8 cm^2, dissociated with preheated StemPro Accutase (Gibco) and subsequently passaged at ~90% confluency. This cell line tested negatively for mycoplasma contamination both at the vendor and in our lab.

Experimental design and RNA extraction

Cells all originated from the same batch of hNSCs differentiated from the WA09 hESC line. We specifically chose to perform our experiments across an isogenic background

to minimize variation and maximize statistical power to identify transcriptomic signatures across differentiation. Each sample was cultured in a separate well and represents an independent differentiation process, which makes for semi-technical replicates. After RNA extraction, samples were quantified using the Quant-iT™ RiboGreen® RNA Assay Kit (Thermo Fisher Scientific). RNA integrity was assessed through RIN scores using the Agilent 2100 Bioanalyzer (mean +/- sd = 9.26 +/- 0.63).

Data preprocessing and quality control

We select for probes present in at least 1 sample at detection p-value of <0.01. Probes were in addition filtered for quality by "perfect" or "good" annotation using the illuminaHumanv4.db package (v1.26) in R. Network adjacency by Euclidean distance and standardized connectivity (Z.K) were calculated on filtered probes values using the WGCNA package to detect outliers, defined as having Z.K. < -2[8, 9]. All samples survived this exclusion threshold. As RNA samples were randomized across gene expression arrays, batch has no explanatory value on days of differentiation (R^2=0.0, p=1.0, see also Figure S1).

In vitro cellular identity

An RNA-sequencing (RNA-Seq) transcriptome database of major classes of cell types present in the cerebral cortex was used to assess cell type-specific gene expression across neuronal differentiation. Briefly, gene expression data of purified populations of neurons, astrocytes, oligodendrocyte precursor cells (OPC), newly formed oligodendrocytes (NFO), myelinating oligodendrocytes (MO), microglia, and endothelial cells from mouse cerebral cortex was downloaded from the database[10]. Fold changes in gene expression values, using fragments per kilobase of exon per million fragments mapped (FPKM), for each gene in each cell type were compared to the mean expression level across the other six cell types. To enrich for cell type-specific genes, we selected the top genes sorted by fold change, with a minimal fold change of 2 and FPKM < 5 in the other brain cell types.

Transition mapping to a spatiotemporal atlas of early human brain development

To investigate global transcriptomic matching between *in vitro* gene expression profiles and *in vivo* gene expression profiles of neocortical brain regions, we applied transition mapping (TMAP)[11]. This method uses a spatiotemporal transcriptome atlas of the human brain[12] and laminar expression data dissected via Laser Capture Microdissection from fetal human brain as *in vivo* input[13]. Both data sets contain brain samples from multiple individuals. TMAP only includes neocortical regions in the analyses. The method performs serial differential gene expression (DGE) analysis between any developmental stages or cortical laminae in the *in vivo* datasets and DGE analysis between two *in vitro* time points of choice. Both DGE lists are sorted on –log10(p-value) and multiplied by the sign of the beta coefficient from the DGE analysis. TMAP subsequently implements the Rank Rank Hypergeometric Overlap (RRHO) test to determine overlap between the *in vitro* and *in vivo* DGE ranked lists and produces RRHO Difference maps that visualizes the extent of overlap[14]. The TMAP and RRHO analyses are implemented in the online CoNTExT bioinformatic pipeline (https://context.semel.ucla.edu). Analyses were run for *in vitro* time points day-0 vs day-30, day-0 vs day-5, day-5 vs day-15, and day-15 vs day-30 across both temporal and spatial dimensions of human cortical development.

Time-series differential gene expression analysis

Two multivariate empirical Bayes models are used to identify differentially expressed genes across *in vitro* neuronal differentiation. The first method exploits the correlation structure among time points and replicates to identify non-constant genes and applies moderation by borrowing the information across genes into the analyses to reduce type-I and type-II errors due to poorly estimated variance-covariance matrices[15]. This method is implemented in the Timecourse package (v 1.42) in R. We used the mb.long() function to calculate the one-sample T^2 statistic that ranks genes based on their log10 probability to have differential expression over time. The second method, Bayesian Estimation of Temporal Regulation (BETR), is an extension of the first approach and uses a flexible

random-effect model that allows for correlations between the magnitude of differential expression at different time points[16]. This method explicitly models the joint distribution of the samples across time points and calculates the probability of a gene being differentially expressed using Bayes rule. BETR is implemented in the betr package (v 1.26) in R. These two methods complement each other as the first approach has increased sensitivity for transient expression differences while BETR has increased sensitivity to detect genes with non-constant expression that is small but sustained over multiple consecutive time points[16]. To maximize our power to detect differentially expressed genes across time points and replicates, we applied both methods to rank genes by their probability of having non-constant gene expression across *in vitro* neuronal differentiation.

Fuzzy c-means cluster analysis

Fuzzy c-means clustering is a soft clustering approach that allows probes to obtain fuzzy memberships to all clusters, minimizes the effect of noise in the data, and avoids erroneous detection of clusters generated by random gene expression patterns. Fuzzy c-means clustering is performed in Euclidian space on standardized gene expression values. This ensures that genes with similar changes in expression cluster together. Membership values represent cluster affiliations and highlight the extent of similarity in expression between genes. To calculate cluster membership values, we first have to estimate a fuzzifier, which determines the level of cluster fuzziness, and the optimal cluster number to use. These two parameters were empirically estimated from the data (fuzzifier = 1.55, number of clusters = 8) as previously described using the Mfuzz package in R[17, 18] (Figure S12). We used these two optimal estimates and subsequently calculated cluster membership with the fclusList() and membership() function in the Mfuzz package. Because these functions only take gene expression values of a single-replicate time series as input, we randomly sampled 100 single-replicate time series from our data and calculated cluster membership values using standardized gene expression values for each independent time series (Figure S13). We then proceeded to calculate average cluster membership for each

probe for each cluster across our 100 independently sampled time series (Figure S14). These average cluster membership values were then used for all downstream analyses.

Functional annotation of clusters

The Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6. 8) was used for functional annotation of each cluster[19]. We restricted our analysis to probes with high membership, i.e. cluster membership > 0.5, to identify most informative functional annotations (Table S1). At a membership value of > 0,5, there is no overlap in genes between clusters (Figure S15). With this setting, 4,318 genes were assigned to a cluster with an average cluster size of 540 genes with the smallest and largest cluster having 221 and 891 genes, respectively. DAVID was run using unique Ensembl IDs and the following databases: UP_KEYWORDS, UP_SEQ_FEATURE, GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT, BIOCARTA, KEGG_PATHWAY, INTERPRO, UCSC_TFBS. Genes significantly detected during differentiation (n = 12,414) were set as background to determine gene overrepresentation in clusters. The functional annotation clustering tool was applied at default settings to group gene list with overlapping gene IDs. Cluster annotations were called significant if the enrichment > 1.0 and at least 1 gene list in the annotation cluster survived Bonferroni correction ($P < 0.05$).

Intolerance of loss-of-function variation across clusters

The probability of being loss-of-function (LoF) intolerant (pLI) was used to infer functional gene constraint across clusters. pLI measures were downloaded (April 2017) for 18,225 genes from the ExAC Browser (http://exac.broadinstitute.org/downloads). The statistical framework underlying the pLI metric is described by others in more detail elsewhere[20]. The Wilcoxon Rank-Sum test was used to test if cluster constraint was statistically different between groups.

GWAS summary statistics used

GWAS summary statistics were checked and reformatted using the *munge_sumstats.py* program within the ldsc software, which removes low quality and ambiguous variants[21]. SNPs in the MHC region (hg19 - chr6: 28477797 – 33448354) were filtered out due to extensive linkage disequilibrium (LD) between markers in this region. The APOE locus (hg19 – chr19: 44,409,039–46,412,650) was removed from analysis of AD to minimize the effect of variants with large effect sizes in downstream regression analyses. For MDD, we included GWAS results from the China Oxford and VCU Experimental Research on Genetic Epidemiology (CONVERGE) consortium[22] and 23andMe[23]. The latter uses a proxy of self-reported depression as a phenotype. We did not include the MDD GWAS of the PGC[24] in our analyses as it has a strong genetic correlation with the self-reported depression GWAS (rg=0.72)[23] but a lower h^2 z-score.

Stratified LD Score Regression - generating annotation files and LD scores

For sLDSR, we used a recent extension to the method that partitions h^2 by continuous-valued annotations[25]. This extension relies on the assumption that if a continuous annotation is associated to increased h^2, LD to SNPs with larger values of this annotation will increase the ^{2} statistic of a SNP more than LD to a SNP with smaller values. We first generated sLDSR annotation files and computed LD scores for each continuous-valued annotation. We mapped gene log(T^2-statistic) and standardized cluster memberships to SNPs in 1KG reference panel BIM files. To increase the number of SNPs in our analyses, we extended gene boundaries with 100kb on each end, similar to here[26]. SNPs that intersected with a gene were annotated with the corresponding gene variable, while SNPs that did not map to genes were annotated with zero. For each annotation, we then estimated partitioned LD scores using using the ldsc.py --l2 function with MAF > 5% and a 1 centimorgan (cm) window. As recommended, only HapMap3 SNPs IDs, with the MHC region removed, were written and used in the final regression model. In case of binary gene annotations, a 1 (in the annotation) and 0 (non in annotation) coding was used. In a similar

fashion, we computed LD scores for all 53 annotations in the baseline model (see Supplementary Methods for details). We in addition generated weight files that contain non-partitioned LD scores using only SNPs that will be included in the final regression model. These are LD scores computed from the HapMap3 SNPs with the MHC region removed. Frequency files were generated with the --freq flag in PLINK 1.9[26–28].

We next generated baseline annotation files using BED files of 52 functional annotations, which were downloaded from the LDSC web portal. Genomic interval coordinates in each BED file were intersected with SNPs present in 1KG reference panel BIM files. If a SNP intersected with an interval in a BED file it was annotated as 1 for that particular annotation. If a SNP did not intersect, it was annotated as 0. In addition to 52 annotations, we also added a recommended base annotation that coded a 1 for every SNP. These 53 annotation makeup the baseline model. With the generated sLDSR annotation files and 1KG reference panels we estimated LD scores for each annotation using the ldsc.py --l2 function with MAF > 5% and a 1 centimorgan (cm) window. As recommended, only HapMap3 SNPs, with the MHC region removed, were written and used in downstream analyses. As a sanity check, we correlated our estimated CEU baseline LD scores to the baseline LD scores that can be downloaded from the LDSC web portal and found a high concordance. For example, the mean Pearson correlation between computed LD scores across baseline annotations on chromosome 22 is 0.99 (n=53, sd=0.002). Thus, we proceeded and used the baseline model in our analyses as it has been shown to provide more accurate mean estimates of enrichment. The baseline model and the details of each annotation are described elsewhere[26, 29].

Supplementary Results

Upregulated genes are more likely to be intolerant to loss-of-function functional variation

Recent work has shown that intolerance to loss-of-function (LoF) functional variation (i.e. constraint) in genes and gene sets can highlight core biological processes and likelihood of disease pathogenicity[20, 30]. High constraint genes have been implicated in neurodevelopmental disorders, such as autism spectrum disorder (ASD) and intellectual disability[30], and are in addition more likely to be adjacent to GWAS signal than the average gene[20]. We therefore investigated constraint across clusters and extracted probabilities of LoF intolerance (pLI) from the ExAC database[20]. The median pLI across all 18,225 genes extracted from the browser is 0.027. Differentially expressed genes (n=5,545, median pLI=0.285) have increased average gene constraint compared to non-differentially expressed genes (n=6,839, median pLI=0.085). This difference between the groups is significant ($W=2.09x10^7$, $P < 2.2x10^{-16}$). The increase in constraint is primarily driven by genes that are upregulated during development of neuronal cells in our *in vitro* model. More specifically, genes in clusters that are affiliated to *neuronal maturation* (median pLI = 0.55, n=633) and *synaptic function* (median pLI = 0.52, n=616) show a significant increase in pLI while genes affiliated to *cell division* (median pLI = 0.067, n=543), *RNA binding* (median pLI = 0.046, n=285), and *extracellular matrix* (median pLI = 0.104, n=490) show a significant decrease in pLI relative to differentially expressed genes (see Figure S16 for test statistics). This shows that genes that are upregulated during neuronal differentiation have a lower tolerance to functional disruption than the average gene expressed, which makes these genes interesting to study in the context of disease.

Cluster enrichments of schizophrenia and height are inversely correlated

We find an inverse correlation between enrichments of SCZ and height across eight gene clusters (rho=-0.86, P=0.011, n=8, see also Figure S8), despite the absence of any evidence of a genetic correlation across the whole-genome (rg=-0.002, p=0.95)[31]. Our findings

however do suggest a genetic correlation. Indeed, large-scale epidemiological studies have, for example, reported an inverse relationship between adult height and SCZ[32, 33]. A population-based cohort study of >1 million Swedish men describes a 15% reduction in SCZ risk for tall subjects compared to short subjects[33]. It has therefore been suggested that height and SCZ are likely to have overlapping genetic causes that can be both discordant and concordant[34]. Our results are in line with this hypothesis and suggest that discordant and concordant effects aggregate on pathway levels that are dependent on time and place during development (Figure S11). While future work is needed to further explore the genetic relation between SCZ and height, these results do highlight the strength of our approach to investigate shared and disease-specific genetic contributions among phenotypes and uncover patterns that would otherwise be missed.

References

1. Fedoroff S, Richardson A. *Protocols for Neural Cell Culture*. 4th ed. Humana Press; 2010.

2. Shi Y, Kirwan P, Livesey FJ. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat Protoc* 2012; **7**: 1836–1846.

3. Zhang X-Q, Zhang S-C. Differentiation of Neural Precursors and Dopaminergic Neurons from Human Embryonic Stem Cells. In: Turksen K, editor. Human Embryonic Stem Cell Protocols. Totowa, NJ: Humana Press; 2010. p. 355–366.

4. Amit M, Carpenter MK, Inokuma MS, Chiu CP, Harris CP, Waknitz MA *et al.* Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Dev Biol* 2000; **227**: 271–278.

5. Thomson JA. Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science* 1998; **282**: 1145–1147.

6. Laurent LC, Nievergelt CM, Lynch C, Fakunle E, Harness JV, Schmidt U *et al.* Restricted ethnic diversity in human embryonic stem cell lines. *Nat Methods* 2010; **7**: 5–6.

7. Ware CB, Nelson AM, Blau CA. A comparison of NIH-approved human ESC lines. *Stem Cells* 2006; **24**: 2677–2684.

8. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**: 559.

9. Oldham MC, Langfelder P, Horvath S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst Biol* 2012; **6**: 63.

10. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *Journal of Neuroscience* 2014; **34**: 11929–11947.

11. Stein JL, de la Torre-Ubieta L, Tian Y, Parikshak NN, Hernández I a., Marchetto MC *et al.* A quantitative framework to evaluate modeling of cortical development by neural stem cells. *Neuron* 2014; **83**: 69–86.

12. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011; **478**: 483–489.

13. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A *et al.* Transcriptional landscape of the prenatal human brain. *Nature* 2014; **508**: 199–206.

14. Plaisier SB, Taschereau R, Wong JA, Graeber TG. Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res* 2010; **38**: e169–e169.

15. Tai YC, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann Stat* 2006; **34**: 2387–2412.

16. Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics* 2009; **10**: 409.

17. Kumar L, E Futschik M. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* 2007; **2**: 5–7.

18. Schwämmle V, Jensen ON. A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics* 2010; **26**: 2841–2848.

19. Huang DW, Lempicki R a., Sherman BT. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.

20. Lek M, Karczewski KJ, Samocha KE, Banks E, Fennell T, O AH *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536**: 285–291.

21. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Consortium SWG of TPG *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–295.

22. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 2015; **523**: 588.

23. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Publishing Group* 2016; **48**: 1031–1036.

24. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* 2013; **18**: 497–511.

25. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat Genet* 2017; e-print: http://dx.doi.org/10.1038/ng.3954

26. Finucane H, Reshef Y, Anttila V, Slowikowski K, Gusev A, Byrnes A *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv* 2017; http://biorxiv.org/content/early/2017/01/25/103069

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

28. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.

29. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015; **47**: 1228–1235.

30. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014; **46**: 944–950.

31. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. *Nat Genet* 2015; **47**: 1237–1241.

32. Gunnell D, Harrison G, Whitley E, Lewis G, Tynelius P, Rasmussen F. The association of fetal and childhood growth with risk of schizophrenia. Cohort study of 720,000 Swedish men and women. *Schizophr Res* 2005; **79**: 315–322.

33. Zammit S, Rasmussen F, Farahmand B, Gunnell D, Lewis G, Tynelius P *et al.* Height and body mass index in young adulthood and risk of schizophrenia: A longitudinal study of 1 347 520 Swedish men. *Acta Psychiatr Scand* 2007; **116**: 378–385.

34. Bacanu SA, Chen X, Kendler KS. The genetic overlap between schizophrenia and height. *Schizophr Res* 2013; **151**: 226–228.

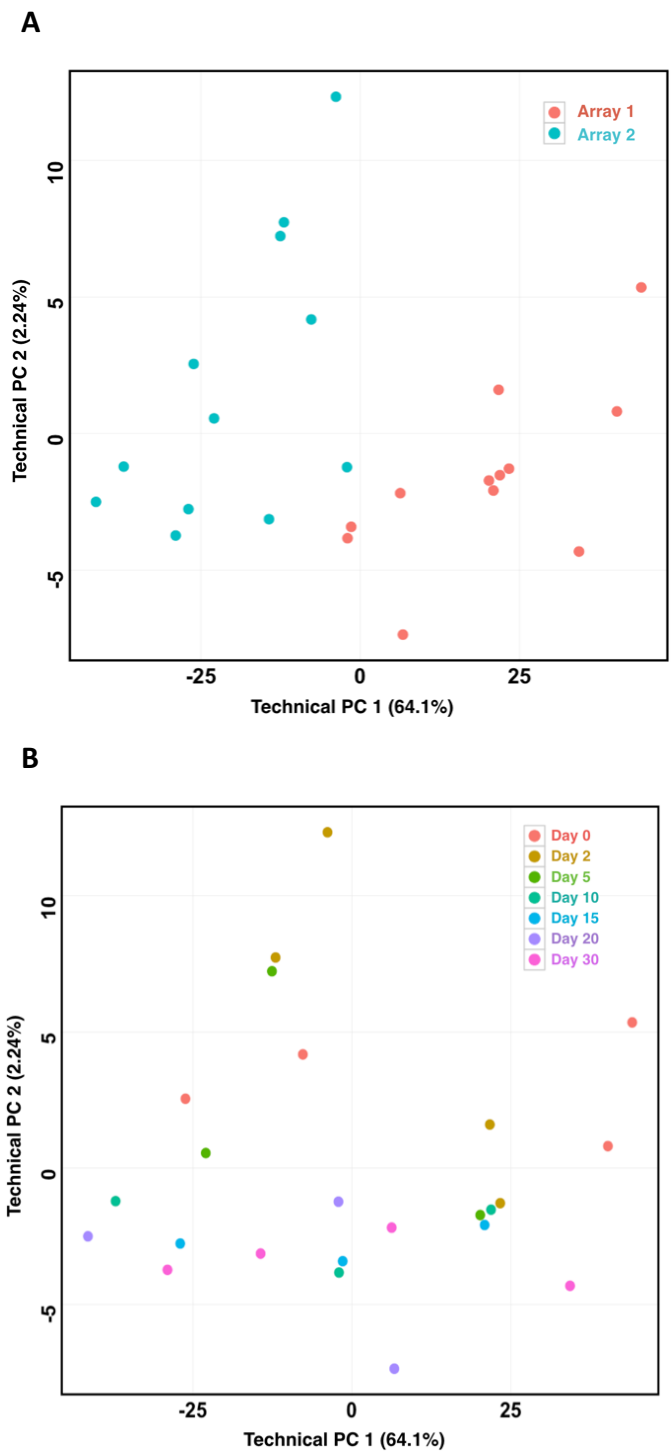


Figure S1. Results of PCA on control probes. The human HT-12 v4 beadchip contains 887 control probes that capture technical variation. Plotted above are PC1 and PC2 with variance explained in parentheses. Dots in the graphs represent samples and are color-coded by (A) array and (B) time. PC1 explains the majority of the information of the control probes but has no correlation with time in culture.

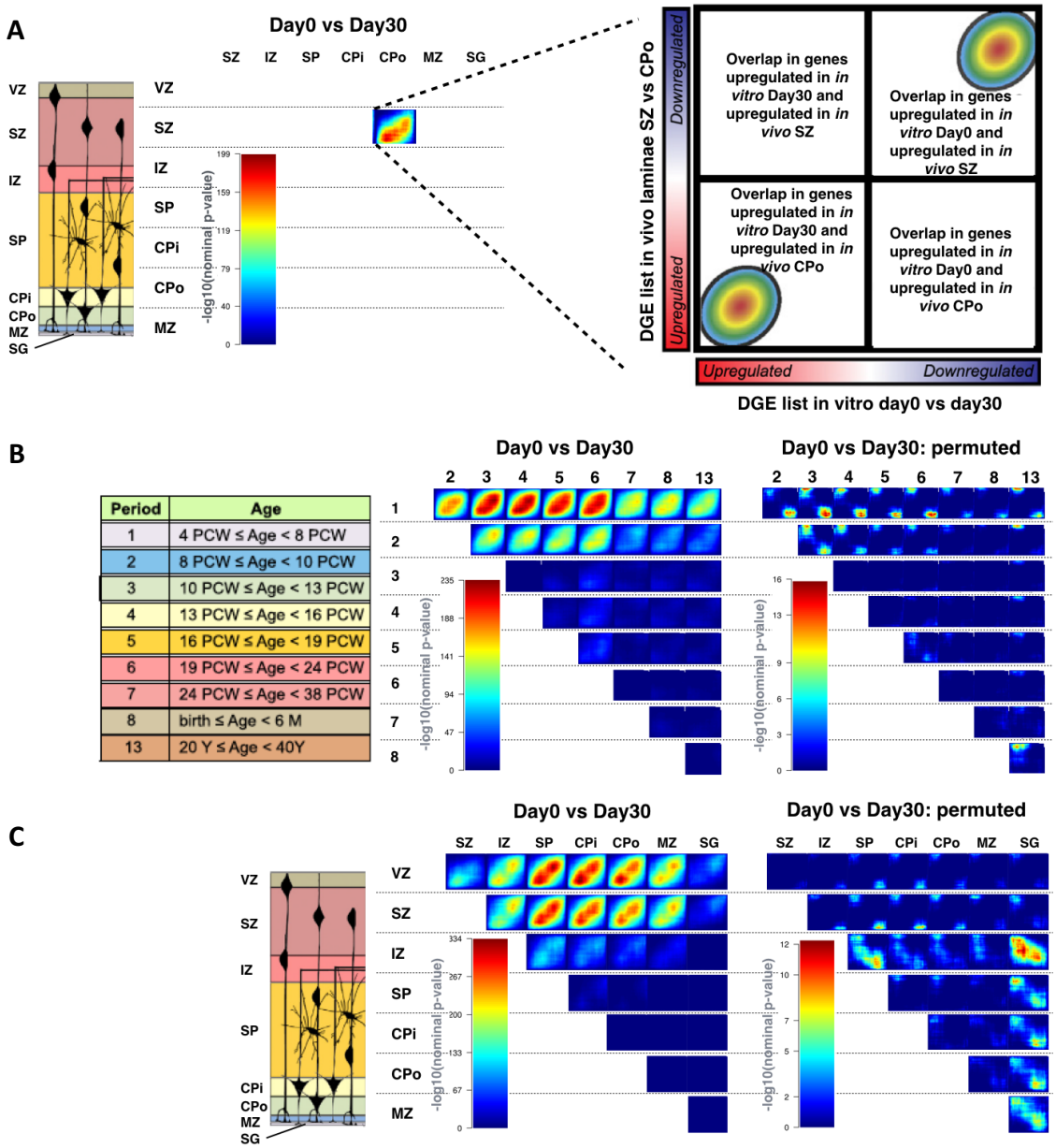


Figure S2. Gene expression overlap between *in vitro* neuronal differentiation and *in vivo* human cortical development. CoNTExT was used to apply transition mapping and generate Rank Rank Hypergeometric Overlap difference maps. (A) Shows a toy example of how to interpret difference maps of overlap between *in vivo* time points and *in vivo* laminae. *In vitro* day-0 vs day-30 differential gene expression (DGE) profile was mapped to serial DGE profiles of (B) human brain developmental stages and (C) laminae of the human cerebral cortex. Difference maps show the amount of matching between *in vitro* and *in vivo* DGE profiles. Maps are colored by –log10(p-value) denoted by each corresponding color bar. On the right of (B) and (C), results are also shown for analyses with permuted *in vitro* sample labels. Abbreviations and numbering above maps correspond to schematic representations on the left (adopted from Stein et al., 2014) of different developmental stages and laminae. VZ = ventricular zone, SZ = subventricular zone, IZ = intermediate zone, SP=subplate zone, CPi= inner cortical plate, CPo = outer cortical plate, MZ = marginal zone, PCW = post conception weeks, M = months, Y = years, Period = developmental stage.

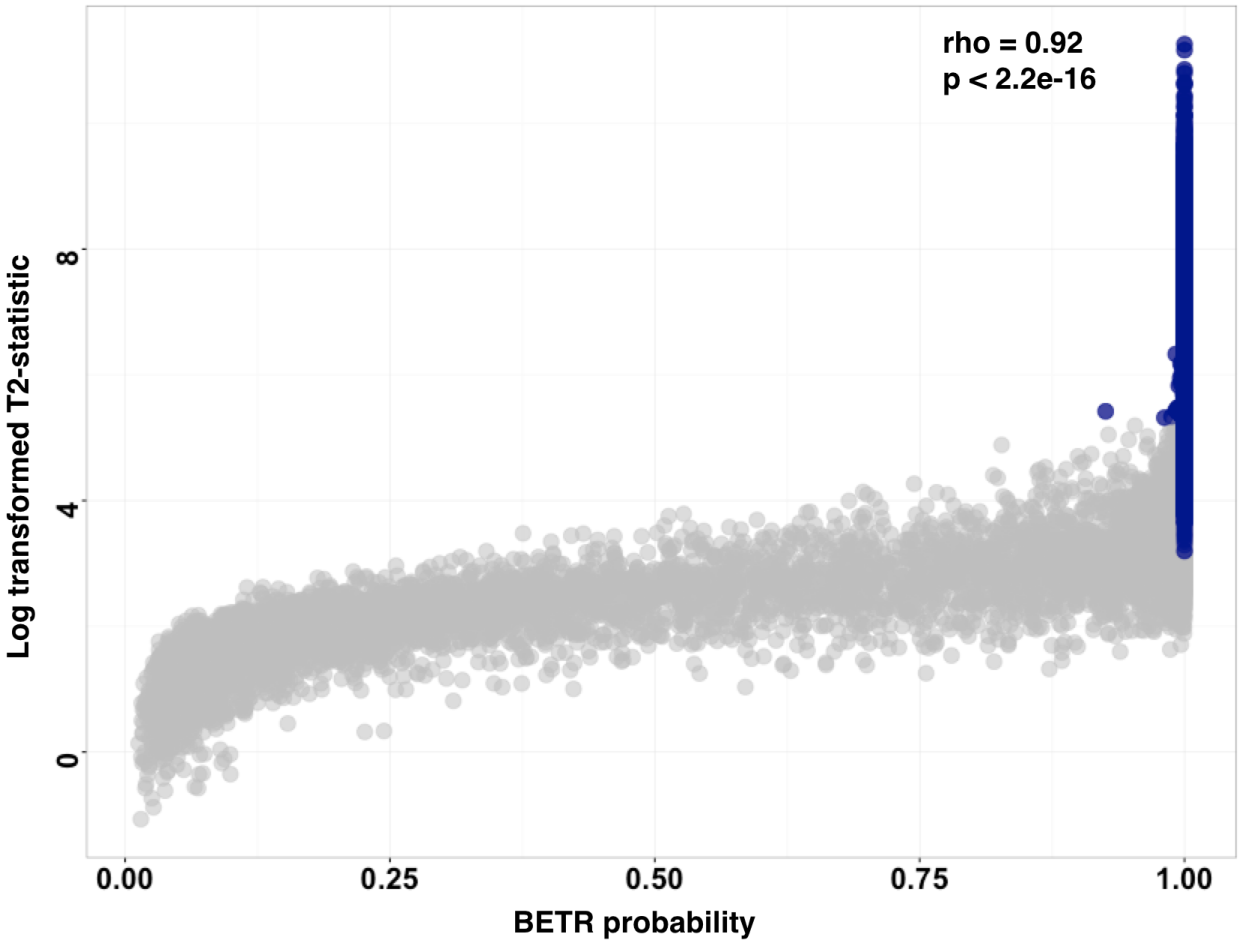


Figure S3. A scatterplot showing the concordance between two methods that identify non-constant genes over time. The x-axis shows the probability from BETR. The y-axis shows the log transformed T^2 statistic from the second method. Each dot represents a probe. Blue color indicates the union of probes that are confidently called as having non-constant expression over time (n=7,734). The Spearman correlation between the ranks is shown in the top right corner.

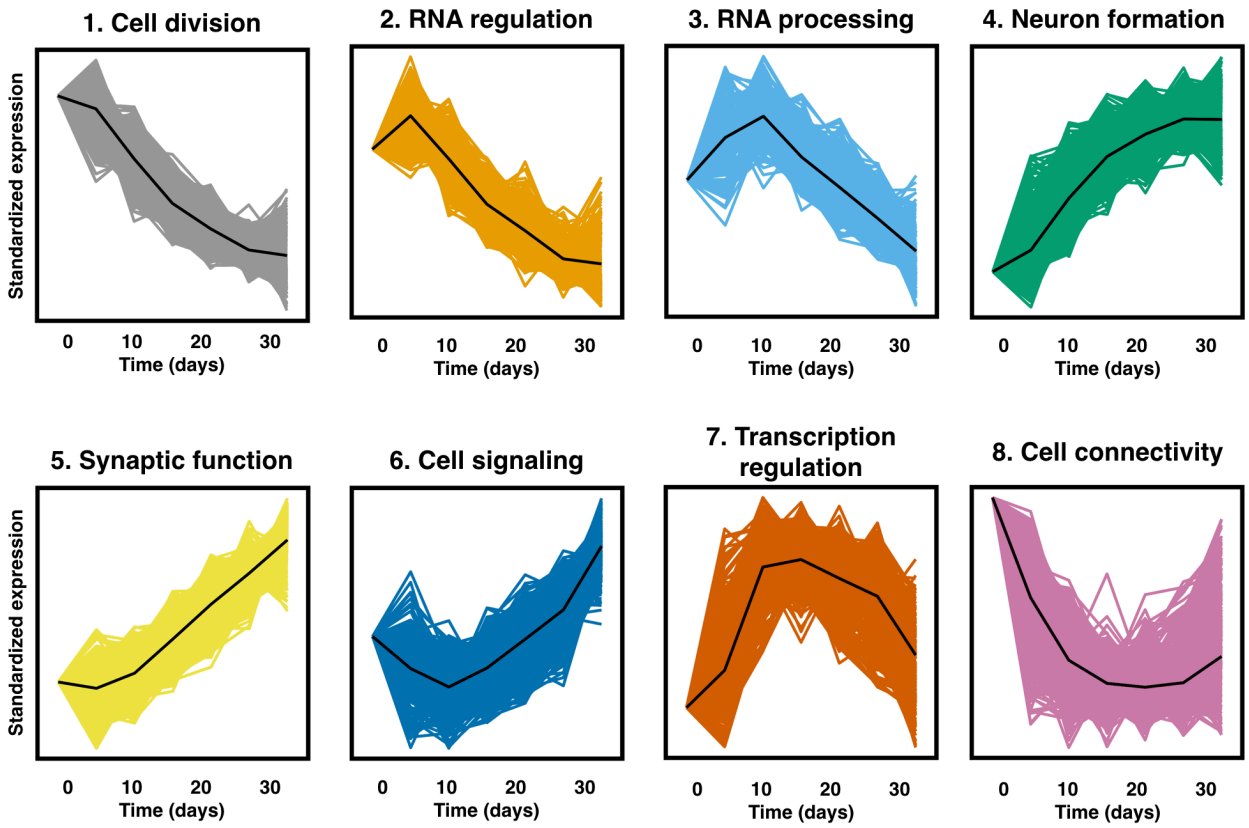


Figure S4. Experimentally-derived longitudinal gene clusters. An enlarged representation of gene expression patterns of high confidence gene members for each cluster (see also figure 3). The x-axis denotes the time across differentiation and the y-axis gene expression values standardized to day-0. The black line highlights the average expression patterns of each cluster.

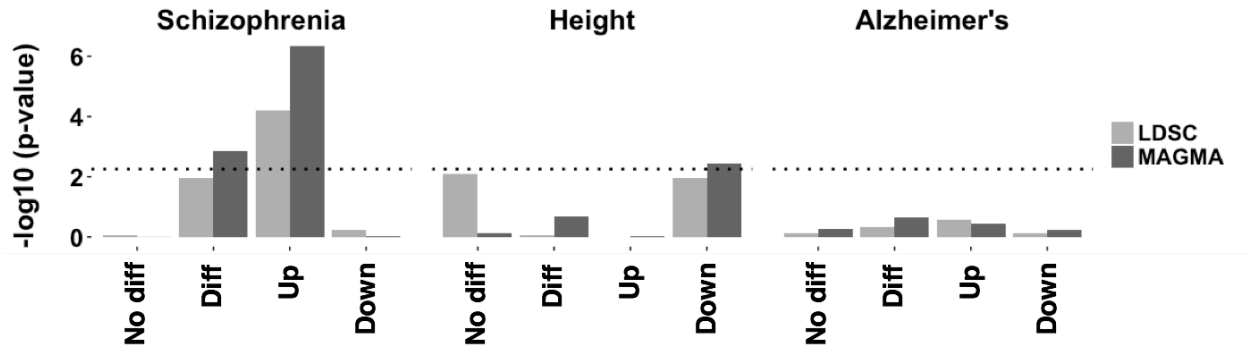


Figure S5. Height and Alzheimer's disease show no h^2 enrichment in up-regulated genes. A more detailed investigation of the enrichment of h^2 of SCZ, height, and Alzheimer's disease across differentially expressed genes. The y-axis denotes the –log10 P-value of the enrichment. No diff = genes that are not differentially expressed; Diff = log (T^2-statistic) as shown in Table 1; Up = genes up-regulated during differentiation; Down = genes down-regulated during differentiation. The dotted line represents the threshold for P = 0.0056 (n=9 tests).

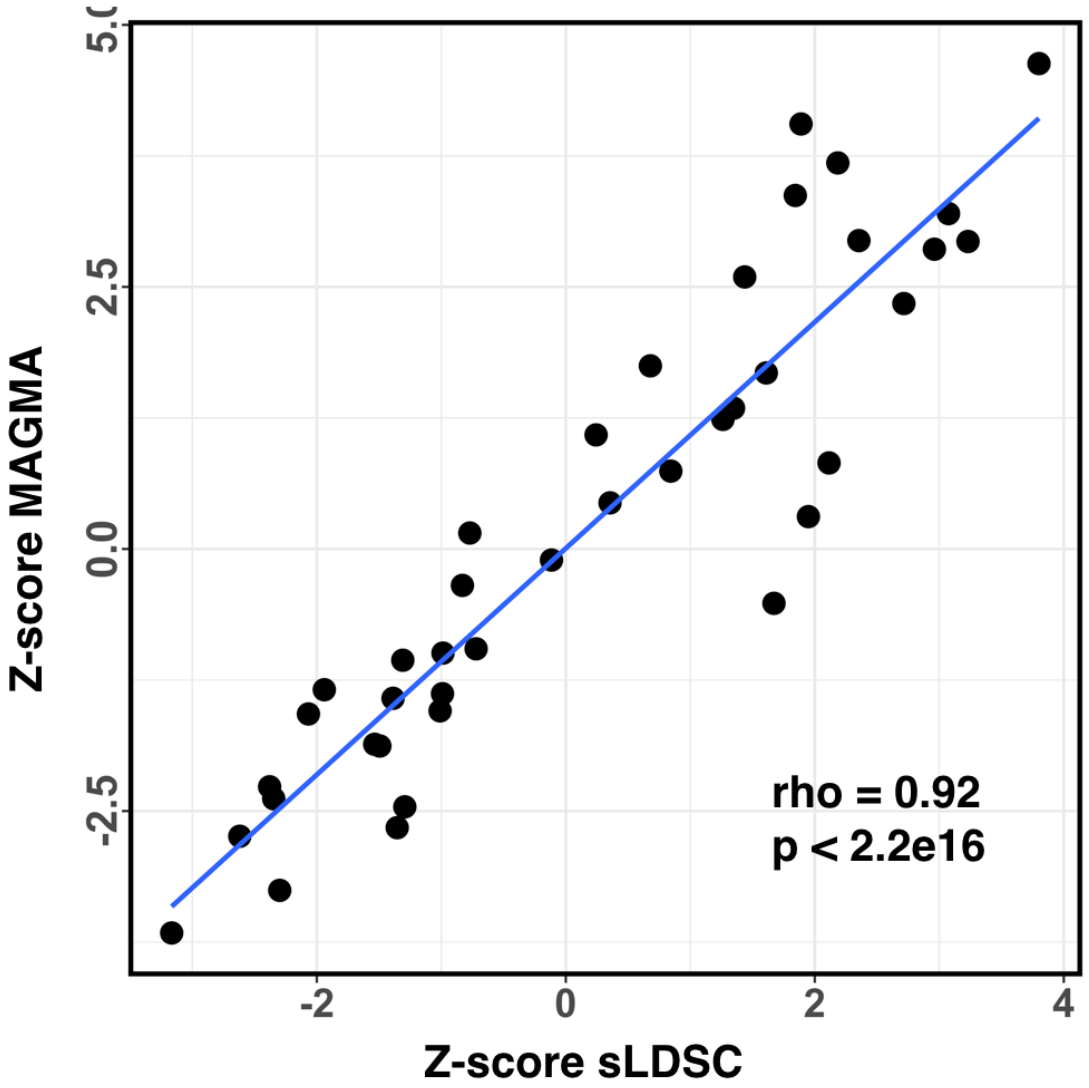


Figure S6. MAGMA and sLDSC show strong concordance in results. Each dot represents the results of phenotype-cluster combination for both MAGMA (y-axis) and sLDSC (x-axis) (n=40). The regression line is shown in blue with the Spearman correlation between the ranks in the bottom right corner.

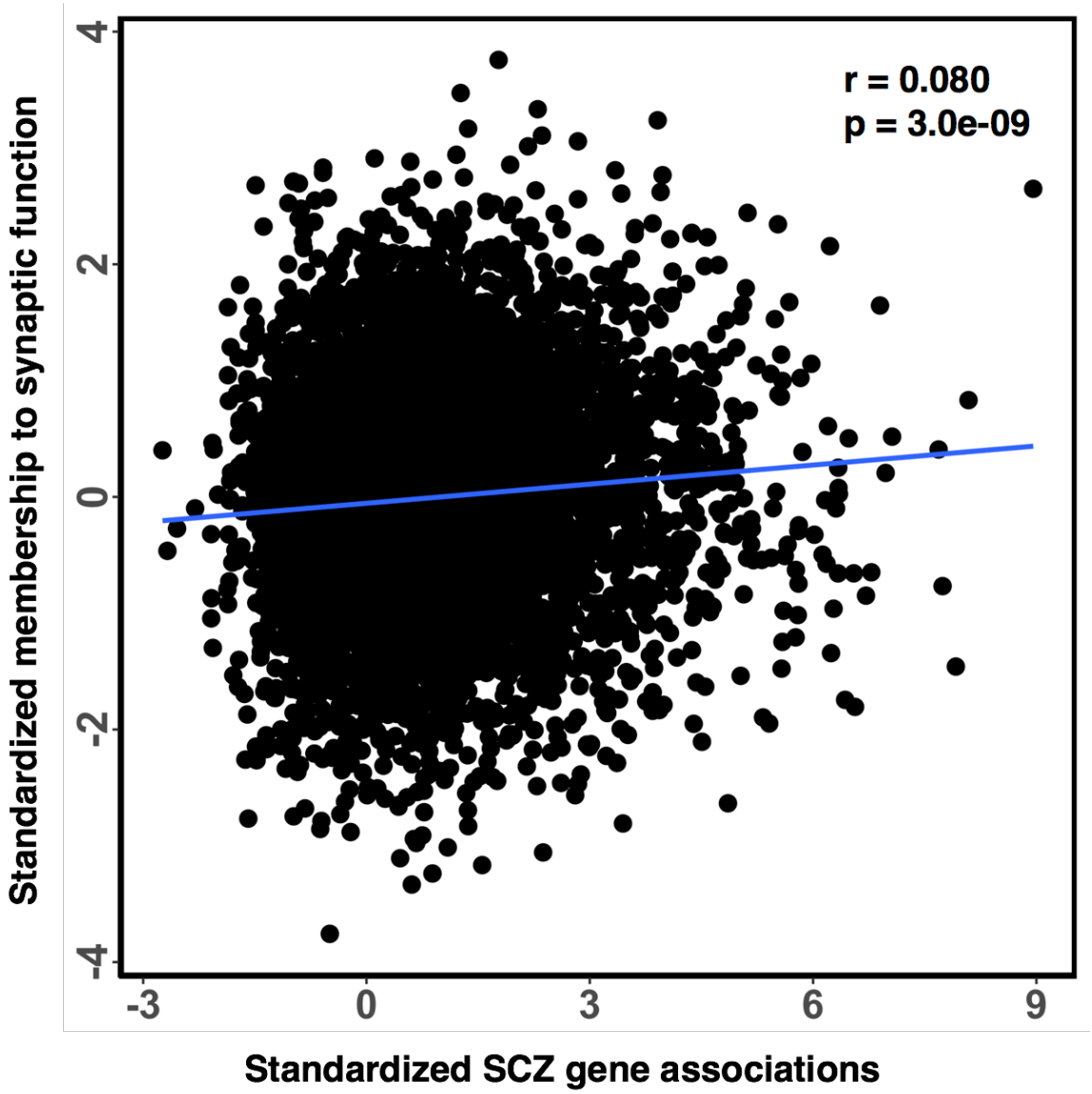


Figure S7. A plot showing the association between SCZ gene-level association statistics and synaptic cluster gene membership. Standardized membership values to the synaptic function cluster and standardized gene level association statistics are shown on the y-axis and x-axis, respectively. The regression line is shown in blue with Pearson correlation test statistics denoted in the top right corner. The plotted association is not yet corrected for gene size, SNP density nor LD.

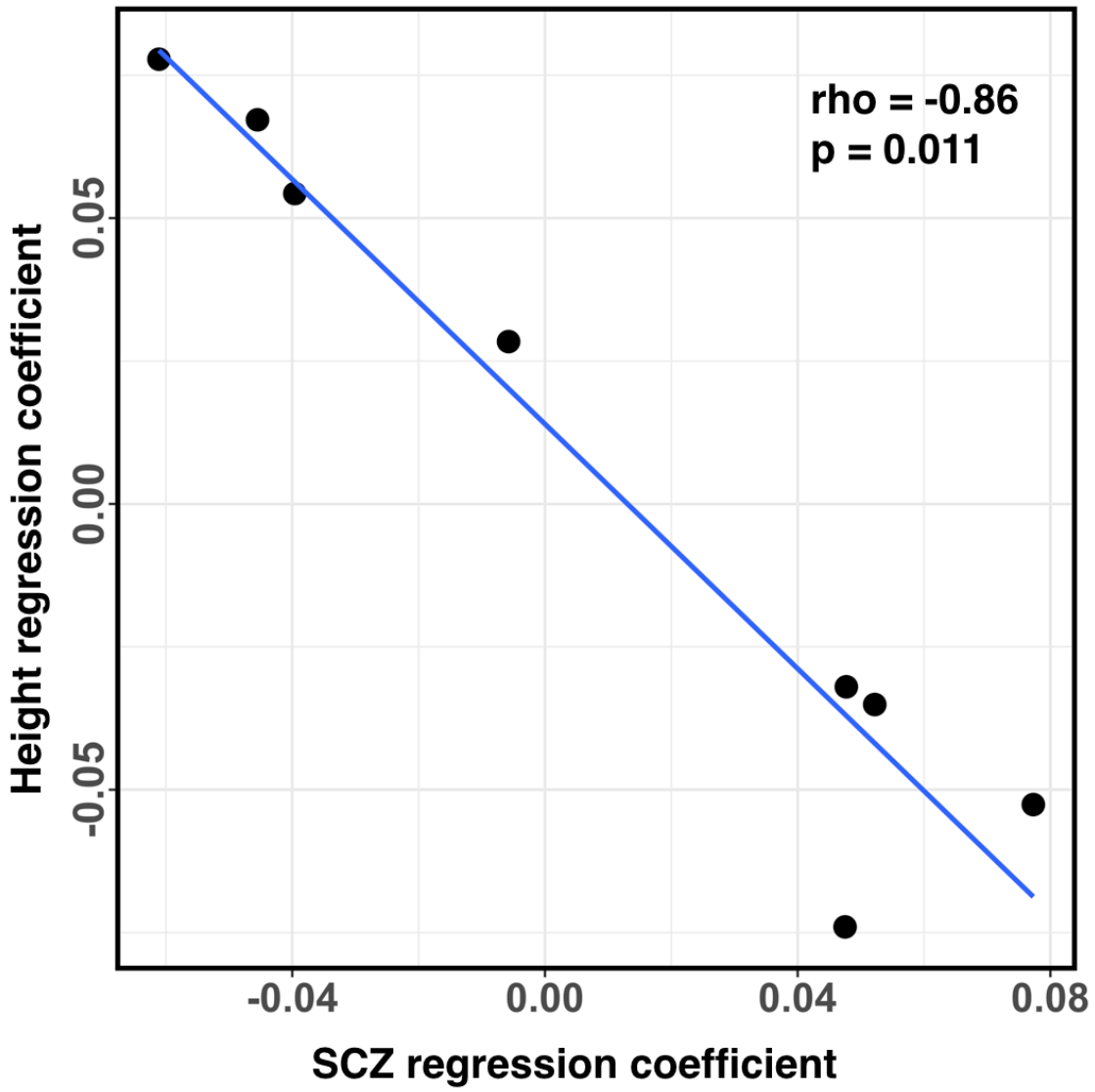


Figure S8. Schizophrenia and height show an inversely correlated pattern of enrichment results. Shown are MAGMA results with each dot representing the regression coefficients of enrichment for schizophrenia and height on the x-axis and y-axis, respectively. The Spearman correlation between the ranks of both methods is shown in the top right corner along with the corresponding significance level.

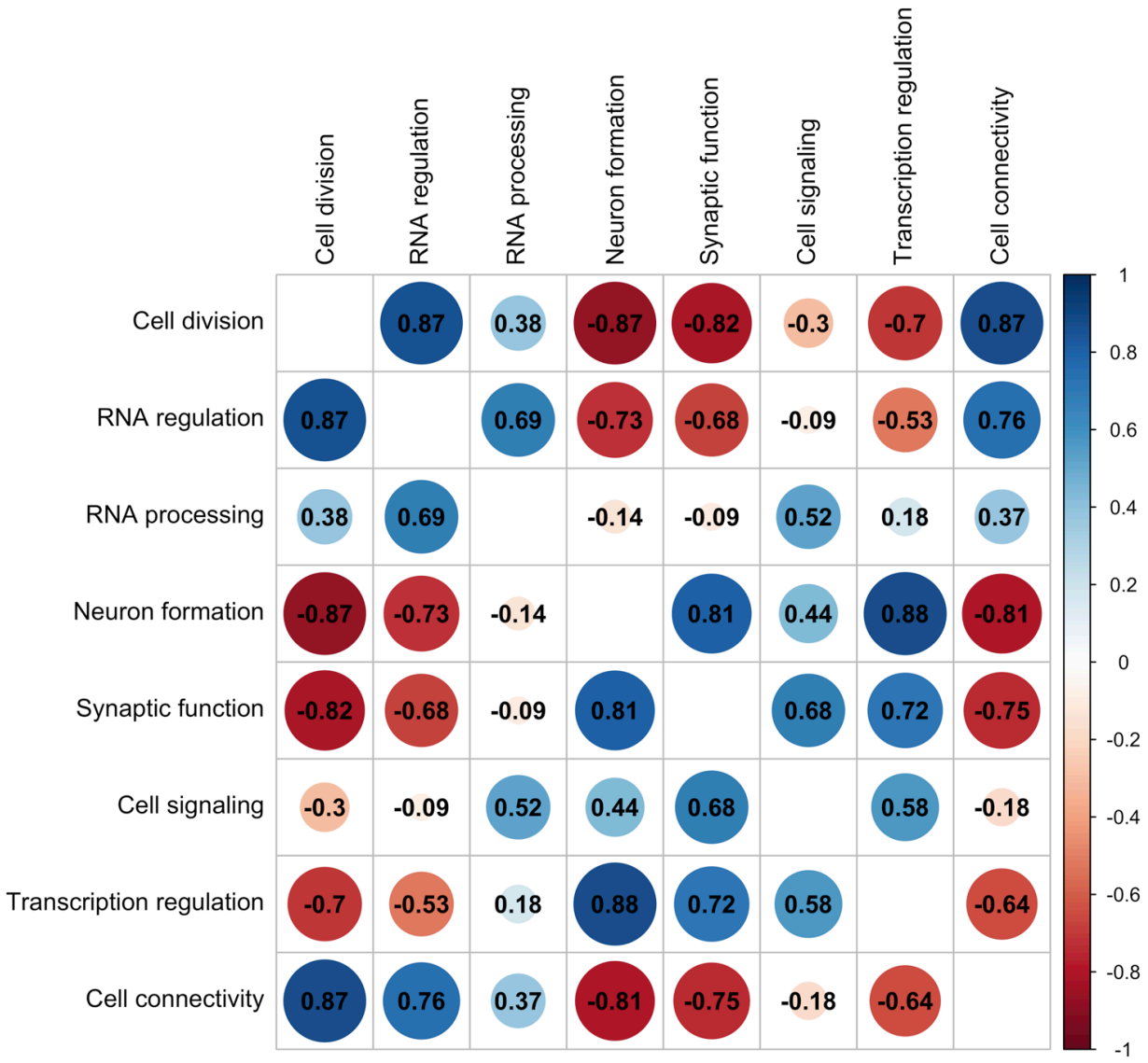


Figure S9. The correlation structure across clusters. A matrix with spearman's correlations calculated between gene membership values across clusters. The rho is denoted in each cell and the strength of the correlation color coded according to the bar on the right.

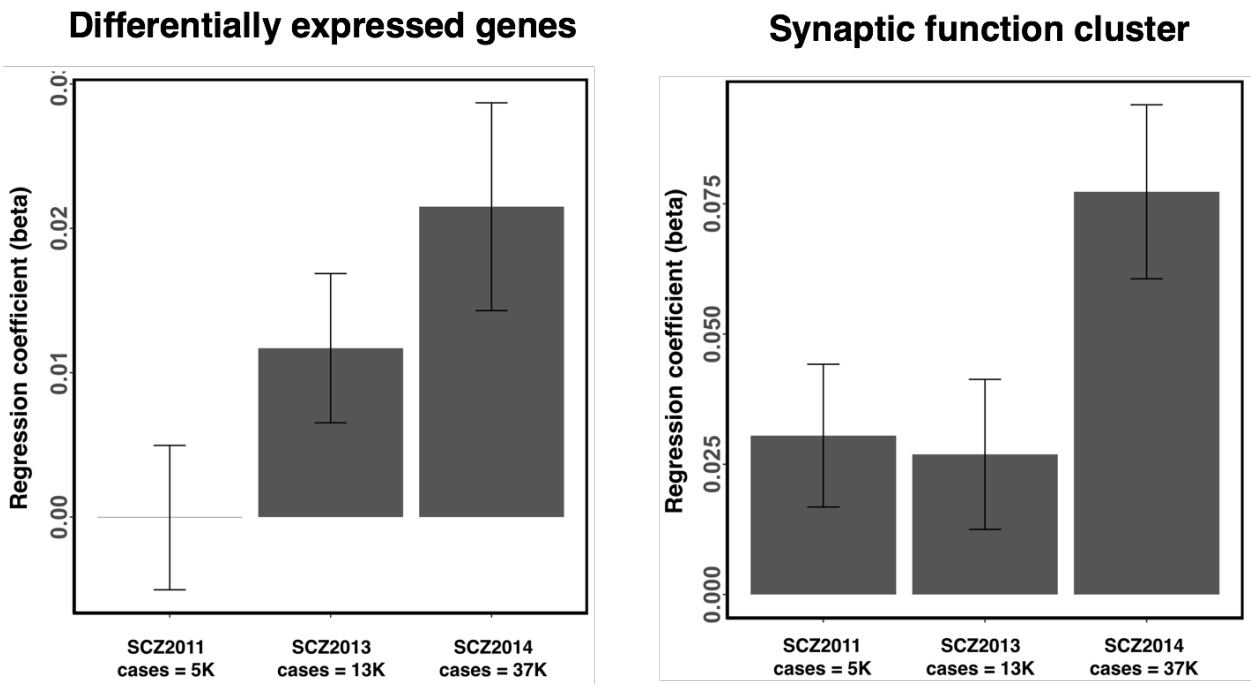


Figure S10. GWAS sample size matters. (left) A bar plot showing the regression coefficient (MAGMA) of the association between the T2 statistic (likelihood of being differentially expressed) and SCZ gene level test statistics for three SCZ GWAS studies of increasing sample sizes. The numbers of cases for each study are denoted on the x-axis labels. (right) A similar plot showing the association of SCZ risk and membership to the synaptic function cluster for each GWAS. Regression coefficients are shown with corresponding standard errors.

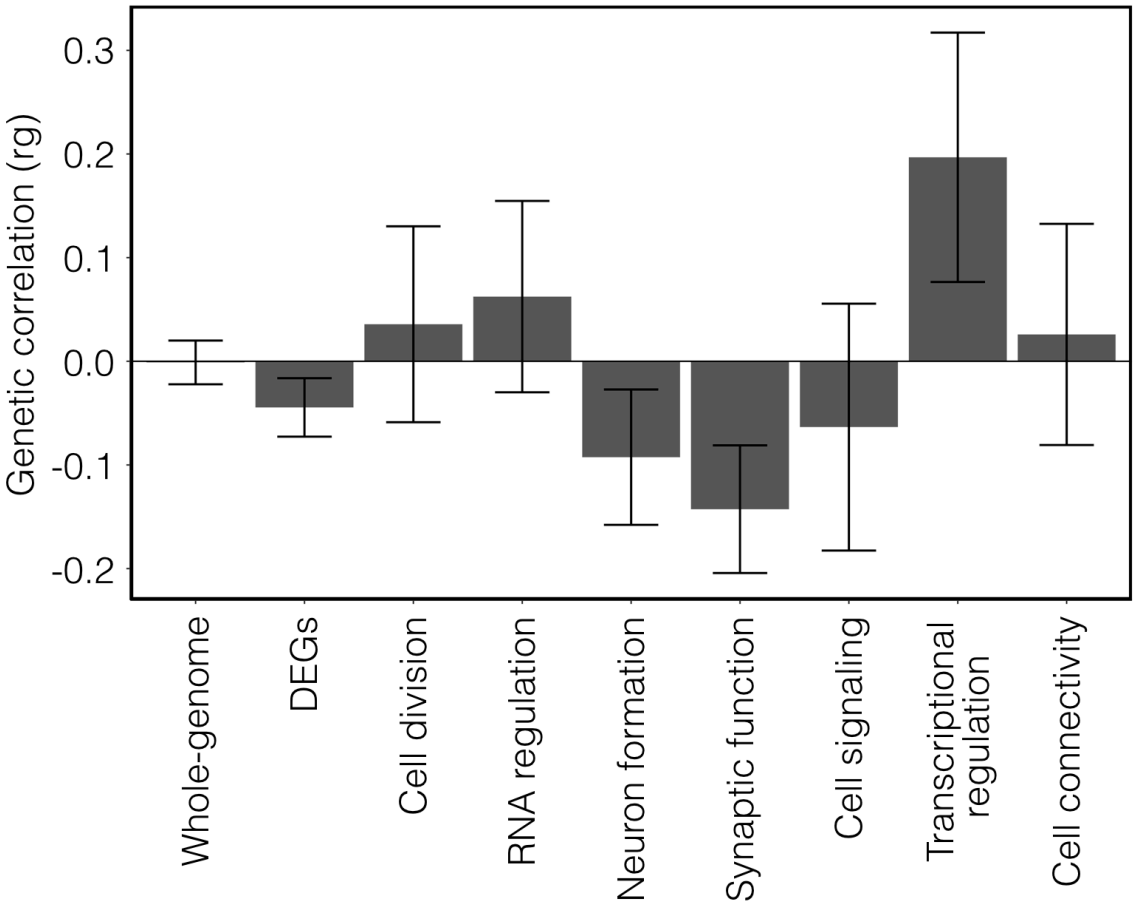


Figure S11. The genetic correlation between schizophrenia and height varies across cluster while absent across the whole genome. Genetic correlations were determined using cross-trait LD score regression and SNPs with MAF > 5%. Stratified correlations were computed using only a subset of SNPs that overlap with genomic coordinates of the highest gene members of that cluster (membership > 0.5). For one cluster (RNA processing), the subset of SNPs was too few to compute a genetic correlation. For differentially expressed genes (DEGs), the correlation was computed on SNPS overlapping the union of DEGs (n=5,818). Error bars represent the standard error.

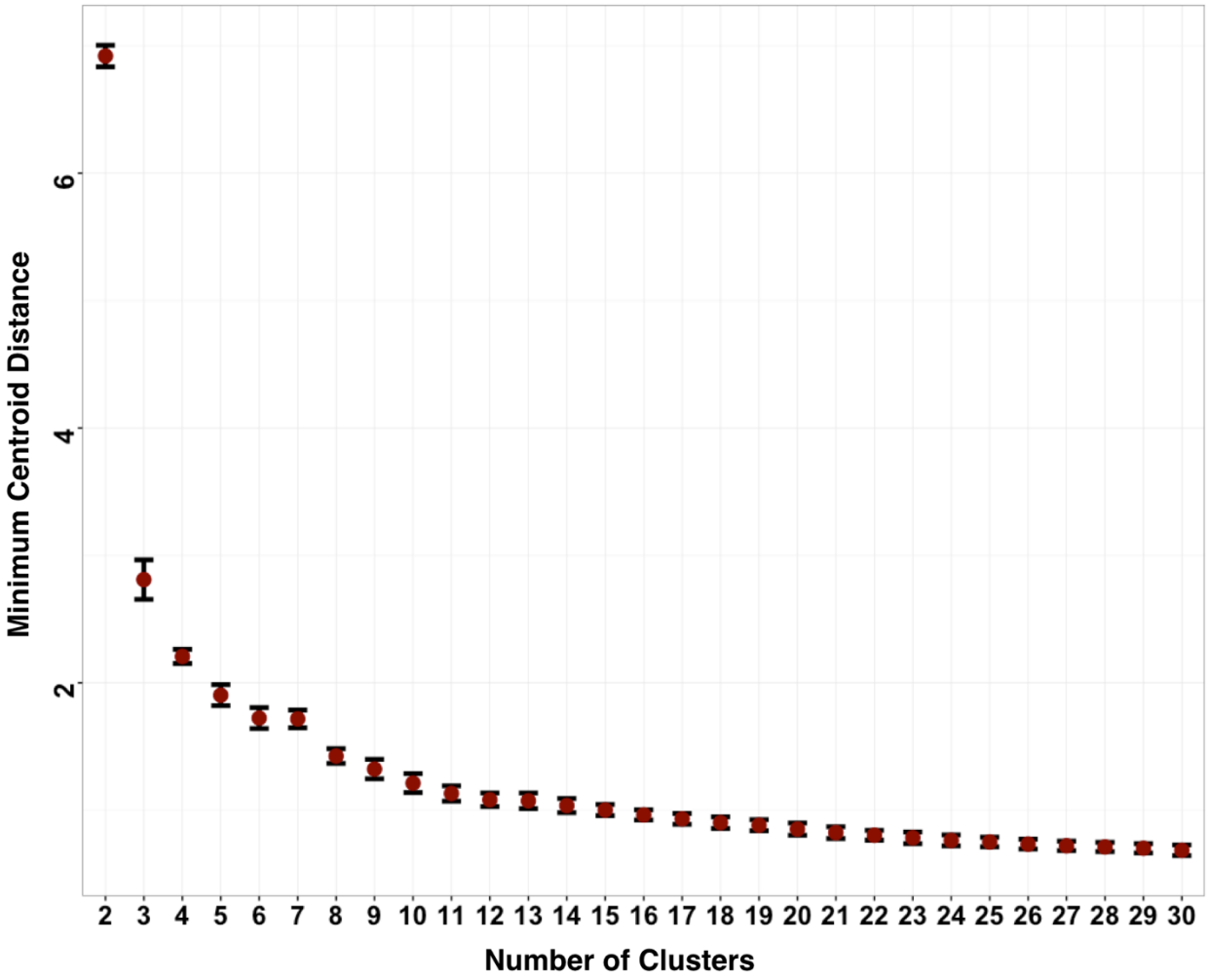


Figure S12. A plot showing minimum centroid distance against increasing numbers of clusters. We sampled 100 independent single-replicate time series (see supplementary figure 5) and performed fuzzy c-means clustering for each time series across various numbers of clusters with a fuzzifier of 1.55. For each we calculated the minimum centroid distance across clusters. Shown above in red are the mean across time series with corresponding standard errors in black. The x-axis shows the number of clusters and the y-axis the minimum centroid distance. The optimal cluster number is chosen as the number before which there starts a gradual decrease in minimum centroid distance as cluster number increases. This indicates that additional clusters add little information. The optimal cluster number was set at 8.

Day 0	Day 2	Day 5	Day 10	Day 15	Day 20	Day 30
Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1
Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2
Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3
Replicate 4	Replicate 4	X	X	X	X	Replicate 4

Day 0	Day 2	Day 5	Day 10	Day 15	Day 20	Day 30
Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1
Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2
Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3
Replicate 4	Replicate 4	X	X	X	X	Replicate 4

Figure S13. A schematic example of sampling independent single-replicate time series. We calculated average cluster membership for each probe for each cluster across 100 independently sampled single-replicate time series. Give the data we can sample 5,184 independent single-replicate time series (4^3 X 3^4). Above are two dummy examples shown of how a single-replicate time series could look like. The yellow color denotes the sampled samples and the red line shows a path that defines the single-replicate time series that these samples make up.

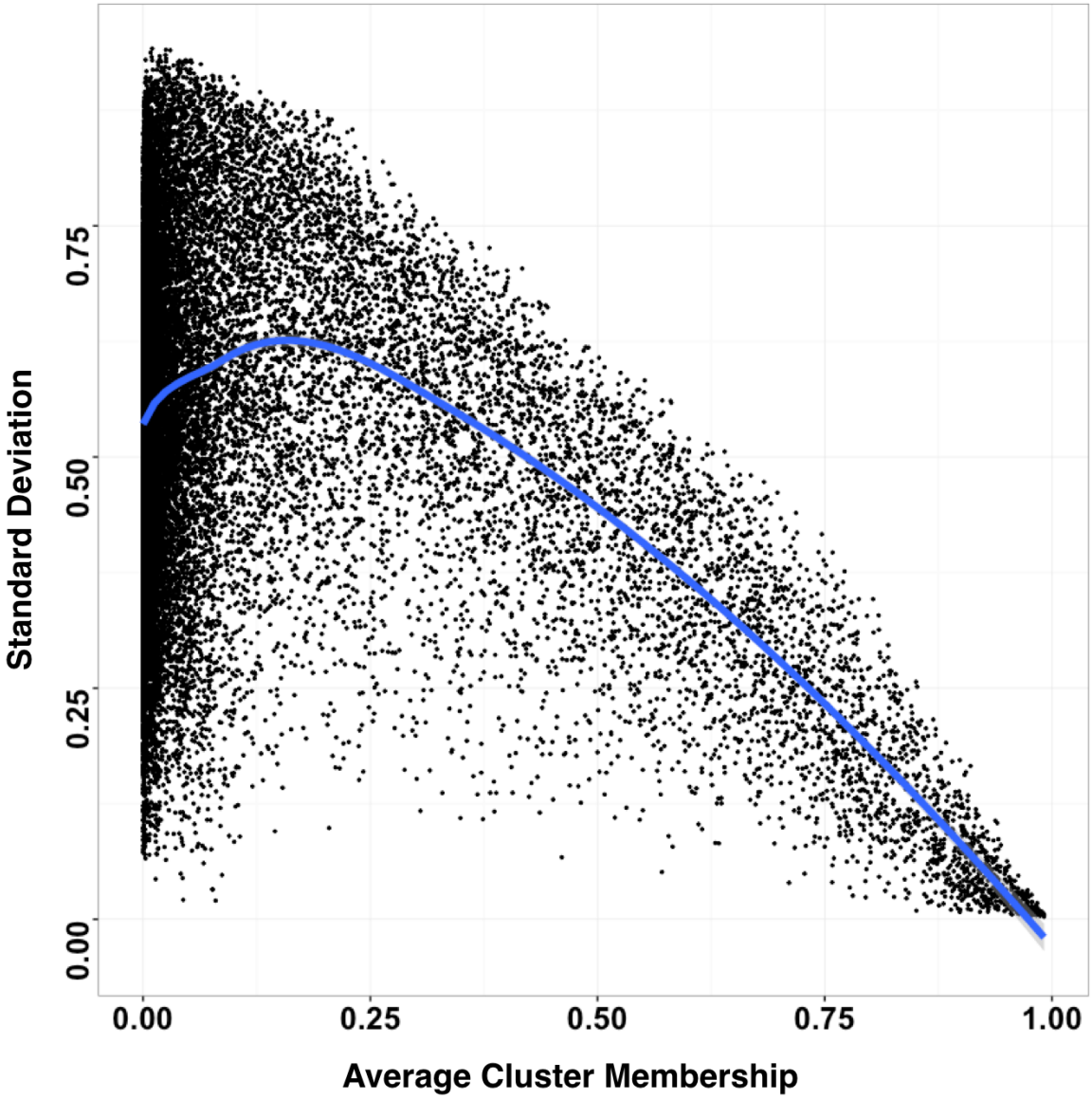


Figure S14. A plot showing the variation in cluster membership values across 100 independently sampled time series. We performed soft clustering on 7,734 probes using fuzzy c-means clustering with a fuzzifier of 1.55 and a cluster number of 8. Cluster memberships were calculated as the average membership determined across 100 independently sampled time series. The x-axis above shows average cluster membership and the y-axis the standardized standard deviation. Data is shown for a specific cluster with each dot representing a probe. The blue line represents a smoothened curve representing the relationship between standard deviation and average membership with 95% confidence intervals in grey. This relationship is consistent across all 8 clusters.

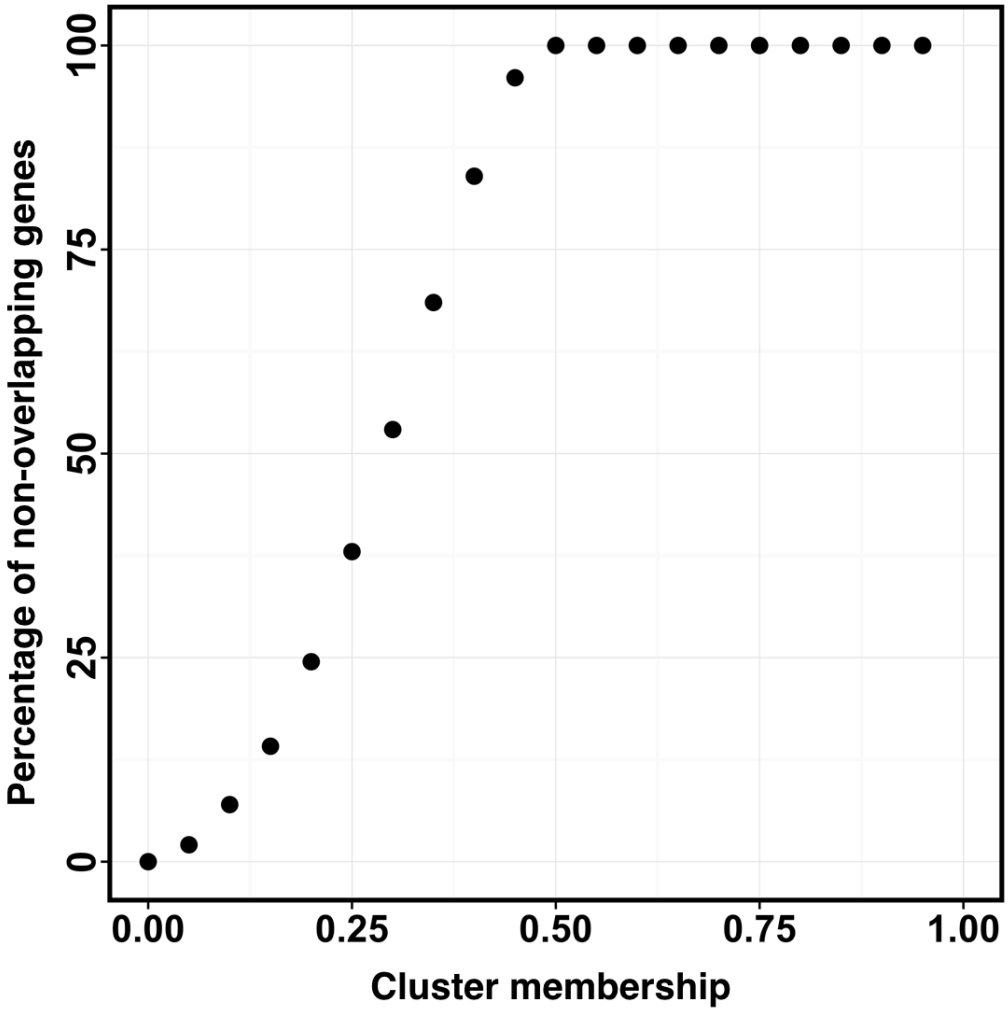


Figure S15. The overlap between clusters across the membership range. The percentage of unique genes by Ensembl ID was calculated across different membership values for each cluster. These percentages were subsequently averaged across 8 clusters. The y-axis shows the average percentage of unique genes (i.e. no overlap between clusters) with membership value on the x-axis.

	Genes (n)	Functional annotations	Enrichment	Cluster label	LoF constraint (median pLI)	LoF constraint (p-value)
1	737	Mitotic cell cycle and division Chromosome and kinetochore DNA repair	23.90 17.41 15.67	Cell division	0.067 ↓	8.54E-11
2	457	RNA binding and ribonucleoproteins RNA processing and ribosome biogenesis RNA splicing	15.77 10.05 6.18	RNA regulation	0.046 ↓	7.64E-06
3	221	RNA processing and binding ncRNA processing *Regulation of cell cycle*	4.25 3.40 *2.13*	RNA processing	0.166	0.48
4	891	Neuronal outgrowth Synaptic formation Neuron differentiation	6.29 4.60 4.23	Neuron formation	0.551 ↑	7.54E-08
5	779	Neuronal projections Synaptic signaling Synapse	16.85 13.56 10.16	Synaptic function	0.524 ↑	1.02E-06
6	253	Membrane-bound vesicles Glycoprotein signaling Transmembrane region	4.17 3.95 2.72	Cell signaling	0.180	0.20
7	396	Regulation of transcription Transcription factor activity Zinc finger and ion binding	4.21 3.02 2.97	Transcription regulation	0.207	0.35
8	584	Extracellular region Cell junctions and adhesion Regulation of cell motility	8.20 7.78 4.85	Cell connectivity	0.104 ↓	3.31E-05

Figure S16. Genes upregulated during neuronal differentiation are intolerant for loss-of-function genetic variation. Cluster annotations shown with average gene constraint shown across clusters.

Table S2. Overview of GWAS summary statistics used

Phenotype	Cases	Controls	λ	Intercept	h^2	h^2 SE	Z	Reference
ADHD	19,099	34,194	1.25	1.03	0.24	0.015	15.6	Demontis et al., 2017[1]
Alzheimer's	17,008	37,154	1.09	1.04	0.06	0.011	5.6	Lambert et al., 2013[2]
Autism	6,197	7,377	1.07	0.99	0.34	0.042	7.9	ASD PGC, 2017[3]
Bipolar disorder	7,481	9,250	1.16	1.01	0.45	0.042	10.7	BPD PGC, 2011[4]
Cross Disorder	33,332	27,888	1.22	1.01	0.17	0.012	13.9	Smoller et al., 2013[5]
Depression self-report	75,607	231,747	1.26	1.01	0.05	0.003	17.2	23andMe Inc., 2016[6]
Height	NA	253,288	2.00	1.32	0.31	0.014	22.4	GIANT 2014[7]
Major Depression CONVERGE	5,303	5,337	1.09	1.04	0.21	0.055	3.76	CONVERGE, 2015[8]
Schizophrenia	36,989	113,075	1.59	1.05	0.24	0.009	25.7	SCZ PGC, 2014[9]

From left to right; numbers of cases and controls included in the GWAS; the lambda GC (λ) as outputted by LDSR, the intercept of LDSR, the h^2 on the observed scale, the h^2 standard error (SE), the h^2 z-score (h2/SE), and the reference of the GWAS study. Heritability was estimated using filtered and processed summary statistics with ancestry matched 1,000 Genome reference panel using MAF > 5%.

References

1. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* (2017). doi:https://doi.org/10.1101/145581
2. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–8 (2013).
3. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8,** 21 (2017).
4. Group, P. G. C. B. D. W. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43,** 977–983 (2011).
5. Smoller, J. W. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381,** 1371–9 (2013).
6. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Publ. Gr.* **48,** 1031–1036 (2016).
7. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, Lo KS, Locke AE, Mägi R, Mihailov E, Por, F. T. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* **46,** 1173–86 (2014).
8. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523,** 588 (2015).
9. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–7 (2014).

Full research team 23andMe Inc. self-reported depression GWAS:
Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, David A. Hinds, Karen E. Huber, Aaron Kleinman, Nadia K. Litterman, Jennifer C. McCreight, Matthew H. McIntyre, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A.M. Northover, Steven J. Pitts, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Vladimir Vacic, and Catherine H. Wilson.

URLs:
PGC downloads: https://www.med.unc.edu/pgc/results-and-downloads
GIANT: http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

Table S3. MAGMA results across differentially expressed genes

Phenotype	Genes (N)	Beta (SE)	Beta_std	P-value	P.adj
Psychiatric					
Schizophrenia	11432	0.022 (0.007)	0.094	0.001	**0.012**
ADHD	11567	0.014 (0.005)	0.059	0.002	**0.021**
Self-report depression	11628	0.013 (0.005)	0.057	0.003	**0.030**
Bipolar disorder	10922	0.007 (0.005)	0.032	0.063	0.566
Cross disorder	11047	0.005 (0.005)	0.020	0.164	1.00
MDD CONVERGE	11584	0.000 (0.004)	-0.001	0.514	1.00
ASD	11406	0.000 (0.004)	-0.002	0.548	1.00
Neurodegenerative					
Alzheimer's disease	11570	0.003 (0.004)	0.015	0.224	1.00
Non-brain					
Height	11580	0.009 (0.011)	0.037	0.210	1.00

Note: Gene level association signal is regressed on log transform Hotelling T^2 statistic while adjusting for gene size, SNP density, and LD between genes. The number of genes included in the regression model can vary across phenotypes based on the number of markers reported in the summary statistics. Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given a change of one standard deviation in the predictor. P-values are adjusted (P.adj) for the number of phenotypes tested (n=9).

Table S4. Stratified LDSR results across differentially expressed genes

Phenotype	τ	SE	P-value	P.adj
Psychiatric				
Schizophrenia	1.70×10^{-9}	7.45×10^{-10}	0.011	0.102
ADHD	1.92×10^{-9}	1.92×10^{-9}	0.062	0.559
Self-report depression	4.34×10^{-10}	4.34×10^{-10}	0.019	0.174
Bipolar disorder	6.16×10^{-9}	6.16×10^{-9}	0.045	0.407
Cross disorder	1.19×10^{-9}	1.19×10^{-9}	0.118	1.00
MDD CONVERGE	6.07×10^{-9}	6.07×10^{-9}	0.084	0.752
ASD	2.97×10^{-9}	2.97×10^{-9}	0.197	1.00
Neurodegenerative				
Alzheimer's disease	1.30×10^{-10}	1.02×10^{-9}	0.449	1.00
Non-brain				
Height	-1.62×10^{-9}	1.36×10^{-9}	0.884	1.00

Note: Heritability of each phenotype is partitioned along the Hotelling T^2 statistic while accounting for the full baseline model and all genes detected. For each trait, we report the contribution to the per-SNP heritability (τ). SE = standard error. P-values are adjusted (P.adj) for the number of phenotypes tested (n=9).

Table S7. Self-reported depression MAGMA cluster conditional analysis

	MAGMA Primary		MAGMA Conditional	
Self-reported depression	**Beta (SE)**	**P-value**	**Beta (SE)**	**P-value**
Cell division	-0.036 (0.014)	1.00	-0.031 (0.022)	0.92
RNA regulation	-0.034 (0.014)	0.99	-0.019 (0.022)	0.81
RNA processing	-0.001 (0.014)	0.54	0.013 (0.020)	0.26
Neuron formation	0.050 (0.014)	1.15×10^{-4}	0.053 (0.029)	0.035
Synaptic function	0.036 (0.014)	4.72×10^{-3}	0.015 (0.021)	0.23
Cell signaling	0.023 (0.013)	0.040	0.017 (0.019)	0.27
Transcription regulation	**0.054 (0.013)**	**2.51×10^{-5}**	**0.052 (0.020)**	**5.42×10^{-3}**
Cell connectivity	-0.026 (0.014)	0.67	-0.010 (0.022)	0.67

Note: Gene level association signal is regressed on cluster membership while adjusting for high membership genes of other seven clusters. Shown are the results of the primary analysis (not adjusted for other clusters) and the conditional analysis. Beta = regression coefficient, SE = standard error.