Title: Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87 740 individuals

Jonathan R. I. Coleman PhD [1,2], Julien Bryois PhD [3], Héléna A. Gaspar PhD [1], Philip R. Jansen MD MSc [4,5], Jeanne E. Savage PhD [4], Nathan Skene PhD [6], Robert Plomin PhD [1], Ana B. Muñoz-Manchado PhD [6], Sten Linnarsson PhD [6], Greg Crawford PhD [7,8], Jens Hjerling-Leffler PhD [6], Patrick F. Sullivan MD FRANZCP [3,9], Danielle Posthuma PhD [4,10,¶,*], Gerome Breen PhD[1,2,¶,*]

Author Affiliations:

[1] MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 8AF, UK.
[2] NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Trust, London SE5 8AF, UK.
[3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-17177 Stockholm, Sweden.
[4] Department of Complex Trait Genetics, VU University, Center for Neurogenomics and Cognitive Research, Amsterdam, 1081 HV, The Netherlands.
[5] Department of Child and Adolescent Psychiatry, Erasmus University Medical Center, Rotterdam, the Netherlands
[6] Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden.
[7] Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA
[8] Department of Pediatrics, Medical Genetics Division, Duke University, Durham, NC 27708, USA
[9] Departments of Genetics, University of North Carolina, Chapel Hill, NC, 27599-7264, USA.
[10] Department of Clinical Genetics, VU University Medical Center (VUMC), Neuroscience Campus Amsterdam, Amsterdam, 1081 HV, The Netherlands.
[¶] These authors contributed equally to the work

[*] Correspondence to:
Gerome Breen (gerome.breen@kcl.ac.uk, +442078480409, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, Denmark Hill, London, SE5 8AF, UK) or Danielle Posthuma (danielle.posthuma@vu.nl, +31205982823, Dept Complex Trait Genetics, Center for Neurogenomics and Cognitive Research (CNCR) & Dept Clinical Genetics, VUMC, W&N building, De Boelelaan 1085, 1081 HV Amsterdam)

Running title: Brain expression of intelligence-associated genes

Abstract

Variance in IQ is associated with a wide range of health outcomes, and 1% of the population are affected by intellectual disability. Despite a century of research, the fundamental neural underpinnings of intelligence remain unclear. We integrate results from genome-wide association studies (GWAS) of intelligence with brain tissue and single cell gene expression data to identify tissues and cell types associated with intelligence. GWAS data for IQ (N = 78 308) were meta-analyzed with a study comparing 1 247 individuals with mean IQ ~170 to 8 185 controls. Genes associated with intelligence implicate pyramidal neurons of the somatosensory cortex and CA1 region of the hippocampus, and midbrain embryonic GABAergic neurons. Tissue-specific analyses find the most significant enrichment for frontal cortex brain expressed genes. These results suggest specific neuronal cell types and genes may be involved in intelligence and provide new hypotheses for neuroscience experiments using model systems.

Introduction

Genome-wide association studies (GWAS) have successfully identified statistical associations with a wide range of behavioral phenotypes and neuropsychiatric disorders [1–3]. Increasing sample sizes has begun to yield findings for intelligence [4–6]. The largest and most recent study of intelligence reported 18 loci significantly associated with intelligence [4]. Significant genetic correlations were observed between intelligence and a variety of behavioral (educational attainment, smoking behaviors), anthropometric (cranial morphology, height, body composition), and psychiatric phenotypes (schizophrenia, autism, depressive symptoms), mirroring epidemiological evidence for correlations between intelligence and a broad range of health-related outcomes [4, 7].

Considered alone, not all associations identified by GWAS precisely localize biological mechanisms amenable to subsequent experimentation. For instance, the most associated variant in a significant locus may not be the causal variant [8, 9], there may be multiple causal variants in a locus [10], loci may act through altering the expression of distant genes [11], and the associations identified by GWAS of complex traits are often spread across the genome [12–14]. To extract meaningful biological inferences from GWAS results, it is necessary to integrate data from other sources, such as studies of gene expression [15]. Results from gene-wise analyses in Sniekers et al. (2017) identified expression predominant in the brain for 14 of the 44 genes with significant association, although some transcription was inferred for most genes across most tissues [4]. Integration of genomic results with data on biological pathways suggested that the most common function of genes associated with intelligence was in the development of the nervous system.

In silico functional annotation of GWAS results is dependent on high-quality biological reference data. Recently, data from the Karolinska Institutet (KI) mouse superset of single-cell RNA sequencing (scRNAseq) of ~10 000 single cells from multiple brain regions was used to map schizophrenia GWAS results to brain cell types [16]. Genes that previously showed association with schizophrenia were expressed with higher specificity in pyramidal cells, medium spiny neurons, and interneurons than in 20 other brain cell types. This demonstrates the potential of cell-type specific annotation to enable the construction of new functional hypotheses for complex traits.

We sought to develop a better understanding of the neurobiological underpinnings of intelligence through combining GWAS results with a number of data sources concerned with tissue and cell-type specific gene expression. To this end, we meta-analyzed the most recent GWAS of intelligence [4] with an extreme-trait GWAS that compared individuals of very high intelligence to a group from the general population (HiQ) [17]. By doing this, we increased the effective sample size of Sniekers et al by 25%. We then analyzed the enrichment of associations with intelligence in single cell expression data from the KI mouse superset (the largest and most comprehensive brain single-cell expression dataset available) [16], and in genomic and transcriptomic data from the GTEx project [18]. Finally, we performed a novel analysis combining genomic and tissue-specific expression data with information from biological, disease-relevant and drug-target pathway databases to further assess the potential impact of biological mechanisms explaining variance in intelligence.

<u>Materials and Methods</u>

Cohort descriptions

Sniekers intelligence GWAS [4]

The cohort analyzed in Sniekers et al. (2017) was drawn from 7 cohorts (13 sub-cohorts), primarily consisting of data from the UK Biobank pilot genotyping release (N = 54 119) and the Childhood Intelligence Consortium (N = 12 441) as well as five additional cohorts (N = 11 748). The phenotype for analysis was Spearman's *g*, or a primary measure of fluid intelligence that correlates highly with *g* [4, 19]. Summary statistics from this analysis are available at

https://ctg.cncr.nl/software/summary_statistics. Full details on cohort characteristics, genotyping and analysis are supplied in the Supplementary Material.

HiQ high-intelligence GWAS

The Duke University Talent Identification Program (TIP) cohort has been described previously [17]. In brief, TIP is a non-profit organization that recruits and nurtures academically-gifted children of extremely high intelligence (top 3%) from the US population. For genomic study, 1 247 participants from the top 1% of TIP (top 0.03% of population) were selected as a high-intelligence cohort (HiQ). IQ was inferred from performance on the Scholastic Assessment Test (SAT) or American College Test (ACT) taken at age 12 rather than the usual age of 18 years. A population comparison cohort (N = 8 185) was obtained from the The University of Michigan Health and Retirement Study (HRS). Participants were assumed to be drawn from the normal distribution of intelligence. Full details on genotyping and analysis are supplied in the Supplementary Material.

Meta-analysis

Summary statistics from Sniekers et al. (2017) and HiQ were meta-analyzed using METAL [4, 20]. The extreme-sampling method used in TIP means the high-intelligence individuals included in Zabaneh et al (2017) are likely to be enriched for variants contributing to intelligence in the population distribution [17]. As such, standard weighting by sample size would not account for the increased discovery power afforded by the extreme-sampling method. Instead, analyses were weighted by their respective non-centrality parameters (NCP), estimated using the Genetic Power Calculator [21–23]. Specifically, NCPs were estimated assuming a causal variant of 20% frequency, capturing 0.1% of variance in each phenotype, assuming HiQ controls are drawn from the normal distribution (+/- 2 SD from the mean) and HiQ cases are sampled from 4 SD above the mean, consistent with IQ > 160 in 99% of the cohort [17]. The NCP of the Sniekers cohort (N = 78 308) was 78.4, while the NCP of the HiQ cohort was 21.6, suggesting the HiQ cohort contributes equivalently to a population cohort of ~21 000 individuals (calculated as ~1000 x NCP), increasing the effective sample size of Sniekers et al (2017) by 27%. Only variants present in both cohorts were retained for analysis.

Following association analysis, genome-wide significant loci were defined via clumping in PLINK2 [24]. Index variants ($p < 5 \times 10^{-4}$) were merged into loci if in linkage disequilibrium ($r^2 > 0.1$ within 500kb) with a variant with a lower p-value. Loci within 50kb of each other were merged. Manhattan and QQ plots were generated using FUMA [25]. Annotation of genomic results with: data from the EBI GWAS catalog; OMIM; GENCODE genes; genes previously implicated in autism and in intellectual disability; copy-number variants previously implicated in psychiatric disorders; and mouse knockout phenotypes was performed with RegionAnnotator version 1.63

(https://github.com/ivankosmos/RegionAnnotator). Summary statistics will be made available at [link available upon publication].

Heritability and partitioned heritability

The heritability of intelligence accounted for by common variants was estimated using LD Score, limited to the HapMap3 variants and pre-computed LD scores provided with the package [26]. Heritability was then partitioned across the 53 genomic annotations provided with the package [27], with the addition of 5 annotations: open chromatin regions (ATAC and ATAC Bryois extend 500, which increases the window around the region by 500 bases in both directions), the intersection between ATAC and conserved regions of the genome (ATAC-conserved) and regions present in the Neanderthal genome (Neanderthal and Neanderthal extend 500 [28, 29]). Regions of open chromatin were identified in prefrontal cortical tissue from 135 schizophrenic individuals and 137 controls using ATAC sequencing, which identifies stretches of DNA free of nucleosomes and other DNA-binding proteins [28, 30].

Gene-wise analyses

Results from the meta-analysis were filtered to retain only single nucleotide variants (SNPs) present in the European superpopulation of 1000 Genomes Phase 3 [31]. SNPs were annotated to a gene using MAGMA v1.06, assigning SNPs to genes if they lay between 35kb upstream and 10kb downstream of the gene location (as supplied on the MAGMA website, build 37 [32]). Gene-wise p-values were obtained from MAGMA as the aggregate of the mean and smallest p-value across all SNPs annotated to the gene. MAGMA accounts for possible confounders such as gene size, gene density, linkage disequilibrium and minor allele count, with reduced type 1

error compared to alternative software [33]. The threshold for genome-wide significance was defined as p = 2.65 x 10^{-6}, the Bonferroni correction for the 18 839 genes tested. Genes passing genome-wide significance were defined as coming from the same locus if their locations were within 50kb of each other, or if they lay within clumped loci from the single variant analysis. Significant genes were cross-referenced to the intellectual disability (ID) gene list provided with RegionAnnotator. The significance of the observed overlap was quantified as a hypergeometric test in R [34], using as background 1 366 ID/DD genes in 18 839 autosomal genes.

Tissue- and cell-specific gene expression

Tissue-specific and cell-type specific proportions of gene expression were calculated following the method described in detail in the Supplementary Methods and in [16]. Tissue expression data was drawn from the GTEx Consortium [18], and brain cell-type expression data was drawn from scRNAseq data from mouse brain (KI mouse superset) [16]. For each gene with a 1:1 homolog between mouse and human (Supplementary Methods) in each cell-type (or tissue), an enrichment score was calculated by dividing the mean Unique Molecular Identifier (UMI) counts for all cells in the cell-type by the summed mean UMI counts across all tissues [16]. Unique molecular identifier counts are an absolute count of transcripts (rather than a relative count), and so can be compared directly across cells [16]. Differences in the number of cells assessed of each type does not considerably bias the score, because it is reflective of the average, rather than of the total expression.

Associations between gene-wise p-values from the meta-analysis and cell-type specific (or tissue-specific) gene expression were calculated using two methods, implemented in MAGMA [32] and in LD Score [26]. In MAGMA, genes were

grouped into 40 equal bins by ranked enrichment score within the cell-type (0-2.5th centile, 2.5-5th centile, up to 97.5th-100th centile) plus an additional bin for genes not expressed in the cell-type. Bin membership was regressed on gene-wise association with intelligence in the meta-analysis (For these analyses, gene-wise association was defined as the mean p-value across all SNPs assigned to the gene.) In LD Score, the 10% of genes with the highest enrichment score within each cell-type were used as a gene set for partitioned heritability analysis. That is, variants were annotated to genes in the top 10% (as described above for performing gene-wise association analyses). Heritability enrichment of this partition was calculated (in LD Score) by dividing the proportion of common-variant heritability captured by the variants by the proportion of all SNPs included in the annotation [27]. The significance of this enriched was determined by a block jackknife approach [27]. Results were considered significant if the association p-values were smaller than the relevant Bonferroni threshold for both methods. Accordingly, in order for a cell-type to be declared significant, genes associated with intelligence must be enriched roughly linearly with bin membership (MAGMA analyses) and enriched in the top 10% compared to the whole distribution (LD Score). These two criteria mean that the observed enrichment has to be robust to the complexity of gene expression in the cell-type.

Conditional cell-specific analyses were performed as a secondary analysis to test whether each enriched cell-type observed was independent of all others. Full details of the method implemented are provided in Skene et al., 2017. In brief, for each enriched cell-type in turn (the target cell-type), z-scores from gene-wise association analyses with intelligence were randomly resampled without replacement. The mean z-score within each expression-specificity decile of the

target cell-type was held constant, but the mean z-score of each specificity decile of other cell types was randomized. Empirical p-values were derived for each of the other cell-types, and this procedure was repeated 10 000 times. Expression in each of the other cell-types is considered to be associated with intelligence independently of expression in the target cell type if the observed p-value is lower than the 500th empirical p-value (i.e. 95% of the empirical distribution [16]).

Predicted tissue-specific gene expression

Results from the variant-level meta-analysis were used to predict gene expression using MetaXcan and genomic and transcriptomic reference data from the brain regions assayed in the GTEx project [18, 35]. Associations between predicted gene expression levels and intelligence were calculated. Significance was set at 1.60×10^{-6}, the Bonferroni correction for the 31303 gene-tissue pairs tested [35]. Significant genes were cross-referenced to the intellectual disability gene list provided with RegionAnnotator.

Pathway analysis

A pathway matrix **P** was generated with elements $P_{g,p} = 1$ if gene g was in pathway p and $P_{g,p} = 0$ otherwise. The elements in the matrix were multiplied by binned gene expression weights obtained from GTEx data (as in the foregoing section on tissue-specific expression) for 13 brain regions and 32 tissues, generating 45 weighted gene/pathway matrices. Only genes with expression data were taken into account. Association between these tissue-weighted pathways and gene-wise associations with intelligence was computed using MAGMA. 13 564 pathways were drawn from OpenTargets (downloaded January 2017 [36]), GO ontologies, canonical

pathways drawn from MSigSB v5.2 C2 and C5 datasets [37], and biological pathways related to psychiatric disorders found in various scientific publications (a link to each pathway source is provided in Supplementary Table 4). The pathways assessed by this approach are related to each other in a complex fashion; GO pathways are hierarchical and MSigDB and OpenTargets pathways capture related gene sets. Accordingly, in order to control for multiple testing, the effective number of pathways tested was established by computing the number of principal components accounting for 99.5% of explained variance in the pathway similarity matrix, obtained by computing the Tanimoto similarity between pathways. This results in a Bonferroni-corrected threshold of $p = 5.34 \times 10^{-6}$ for 9 361 effectively independent tests for each matrix. A more stringent threshold was applied secondarily, taking into account all tissue-specific pathway matrices for a total for 9 361 x 46 tests and threshold $p = 1.16 \times 10^{-7}$.

Conditional gene set enrichment

Gene sets of interest were drawn from the results of pathway analyses. Specifically, the human postsynaptic density proteome gene set [38] was used to capture the effects of synaptic and neuronal pathways. Gene sets of RBFOX and CELF4 targets were included as they showed enrichment that appears to be driven by brain-specificity, and have been previously implicated in other brain-related traits [16, 39, 40]. ID/DD gene sets were tested as they are of specific interest to the study of intelligence. To assess whether the enrichment of these gene sets is independent of gene expression in somatosensory pyramidal neurons, conditional analyses were performed following the method described above for conditional cell-type analyses, modified for the use of gene sets [16].

<u>Code availability</u>

All analyses used publicly available software that were not modified for use.

<u>Results</u>

Meta-analysis

The genetic correlation between Sniekers et al. (2017) and HiQ was 0.92 (SE: 0.07) and was sufficiently high to justify meta-analysis. 25 loci met genome-wide significance with intelligence (i.e., $p < 5 \times 10^{-8}$; Table 1, Figure 1a). Of these, 13 were genome-wide significant in Sniekers et al. (2017) and 12 were novel (Table 1; Supplementary Table 1). The single locus previously identified in the HiQ sample was not genome-wide significant in this analysis (p = 0.00595 [17]).

Assessment of genome-wide inflation yielded a median λ_{GC} of 1.24. The LD score regression intercept was 1.004 (SE: 0.01). A λ_{GC} substantially greater than 1 indicates genome-wide inflation, which can result from confounding factors such as population stratification and batch effects, or from real effects distributed across multiple variants (polygenicity). In contrast, an LD Score regression intercept greater than 1 is indicative of confounding only (as polygenicity contributes to the slope of the regression, not the intercept) [26]. As such, these results taken together suggest that this inflation is caused by polygenicity rather than confounding (Figure 1b) [26].

Annotation of specific genomic loci to databases of interest suggested overlapping loci between intelligence and a variety of phenotypes (Supplementary Table 2). The largest overlap was observed with educational attainment (14/25 loci), but overlap at multiple loci was widespread, including with age at menarche, height, body mass index, autoimmune disease, and schizophrenia. Genes previously

implicated in intellectual disability or developmental delay (ID/DD) were present within 9/25 loci.

Heritability and partitioned heritability

LD Score regression yielded a SNP-heritability estimate of 0.221 (SE: 0.01) in line with the previously reported SNP-heritability in Sniekers et al (2017). Partitioning this heritability across 58 functional SNP annotations identifies conserved regions as significantly enriched contributors to the heritability of intelligence (proportion of heritability = 0.340, enrichment = 13.3 fold, p = 3.26 x 10^{-8}), consistent with previous reports in a subset of our meta-analyzed cohorts [41]. Four extended annotations were also significantly enriched (p < 8.62 x 10^{-4}; Figure 2), suggesting that genetic variation located in the vicinity of conserved regions, enhancers (specifically H3K4me1 elements) and open chromatin in brain dorsolateral prefrontal cortex [28] and fetal cells) are enriched in heritability for intelligence.

Gene-wise analyses

93 genes at 41 loci were identified at genome-wide significance (p < 2.65 x 10^{-6}, Bonferroni threshold for 18 839 genes; Table 2, Supplementary Figure 1). 28 of these genes (20 loci) were also identified in Sniekers et al. (2017). 11 of the 93 genes were previously implicated in intellectual disability or developmental delay, although this overlap does not significantly from what is expected under the null hypothesis of no enrichment (p = 0.073, hypergeometric test; Table 2).

Tissue- and cell-specific gene expression

Tissue-specific enrichment analysis identified an enrichment of genes with high brain-specific expression associated with intelligence (p_{MAGMA} = 4.43 x 10^{-9}, p_{LDSC} = 4.23 x 10^{-6}; Figure 3a, Supplementary Table 3a). Across 10 brain regions in GTEx[18], stronger gene associations with intelligence were associated with increased specificity of gene expression to the frontal cortex (p_{MAGMA} = 0.00305, p_{LDSC} = 2.66 x 10^{-4}; Figure 3b, Supplementary Table 3b).

In the KI level 1 (broad cell groups) cell-type specific analyses, both linear regression (MAGMA) [32] and heritability enrichment-based analyses (LD Score) [27] supported enrichment of genes with high specificity to pyramidal neurons in the somatosensory neocortex (p_{MAGMA} = 1.41 x 10^{-6}, p_{LDSC} = 5.81 x 10^{-4}) and in the CA1 region of the hippocampus (p_{MAGMA} = 9.08 x 10^{-5}, p_{LDSC} = 1.12 x 10^{-3}), as well as to midbrain embryonic GABAergic neurons (p_{MAGMA} = 9.47 x 10^{-5}, p_{LDSC} = 1.61 x 10^{-3}; Figure 3c, Supplementary Table 3c). Level 2 analyses (narrowly-defined cell types) suggested significant enrichment of genes with high specificity to type pyramidal cells of the CA1 region (p_{MAGMA} = 2.16 x 10^{-4}, p_{LDSC} = 1.19 x 10^{-4}), although considerable variability was observed between methods at this level of granularity (Figure 3d, Supplementary Table 3d).

Cell-type specific analyses were repeated conditioning on each enriched cell-type in turn. When controlling for gene expression in pyramidal neurons of the somatosensory neocortex, the previously observed enrichment in CA1 pyramidal cells is lost. In contrast, the patterns of enrichment observed when conditioning on expression in CA1 pyramidal cells or in midbrain embryonic GABAergic neurons were consistent with those observed without conditioning (Figure 4).

Predicted tissue-specific gene expression

Across the 10 GTEx brain tissues, results from MetaXcan suggested significant effects on the expression of 16 genes ($p < 1.60 \times 10^{-6}$; Table 3). Eight of these genes are situated at locus 11 (Chr3, 48.7-50.2 Mb). Genetic variation was predicted to upregulate 6 genes and downregulate 10 genes, with both single-region (9 genes) and multiple-region (7 genes) patterns of altered expression implied. Three genes (*NAGA, TUFM* and *GMPPB*) have previously been implicated in ID.

Tissue-specific pathway analysis

Tissue-specific pathway analyses identified 32 nested pathways with $p \leq 5.34 \times 10^{-6}$ (Bonferroni correction for 9 361 effectively independent pathways), of which 29 contained genes that were significant in the gene-wise analysis (Supplementary Table 4; Supplementary Figure 2). 7 pathways remained significant after further correcting for all tissue-specific tests: "self-reported educational attainment", "modulation of synaptic transmission", "neurodegenerative disease", "neuron spine", "schizophrenia", "rare genetic neurological disorder", and "potentially synaptic genes" (Table 4).

Conditional gene set enrichment

Results from tissue-specific and gene-set analyses identified a number of gene sets associated with intelligence. Of specific interest were synaptic genes (post-synaptic density proteome list [37]), RBFOX family binding partners [40], CELF4 binding partners, and previously reported intellectual disability genes. Additional analyses were performed to assess whether the association of these gene sets was independent of the enrichment for gene expression in pyramidal cells of the somatosensory cortex. Bootstrapped analyses confirmed the association of each

gene set prior to conditional analysis (empirical $p < 0.05$; Supplementary Table 5). Conditioning on expression in gene expression in pyramidal cells of the somatosensory cortex, the enrichment for synaptic genes is no longer significant, indicating that this enrichment is not independent of that observed in the pyramidal neurons. This effect was not observed for RBFOX or CELF4 targets. Intellectual disability genes were weakly enriched prior to conditioning, and conditioning had little effect on enrichment. However, subdividing ID/DD genes into those associated with severe and those associated with moderate ID/DD indicated that the enrichment stemmed predominantly from moderate ID/DD genes, and this was not altered by conditioning on somatosensory pyramidal neuron gene expression (Supplementary Table 5).

Discussion

The overall power of our GWAS meta-analysis was equivalent to a sample size of ~99 000 individuals due to the inclusion of the extreme trait HiQ sample, which contributes equivalently to a population cohort of ~21 000 individuals, and is likely to be enriched for alleles associated with intelligence in the normal range. We mapped results from our GWAS to tissue and cell-type specific gene expression data, identifying enrichment of specificity at multiple levels: in the brain, the frontal cortex, midbrain embryonic GABAergic neurons and pyramidal neurons, especially those in the somatosensory cortex. A number of genes previously implicated in intellectual disability or developmental delay are predicted from the GWAS results to show differential gene expression for normal range IQ in different brain regions, and are associated with variation in intelligence in the normal range from gene-wise analyses.

RNA sequencing data suggests that, as expected, genes more strongly associated with intelligence are enriched for brain-specific expression in general. While the dominance of brain-specificity over other body tissue specificity is pronounced, assessing differences within the brain is more difficult. Genes more strongly associated with intelligence showed higher specificity for frontal cortical expression, but the differences in cell composition between brain tissues means that cell-type analyses may be more informative. This can be seen within our results, in that pyramidal neurons of the somatosensory cortex were significant in the cell-type specific analysis, but the cortex as a whole is not significant in the GTEx brain-tissue analysis. This is perhaps due to the fact that the cortex is a highly heterogeneous mixture of cell-types. Our results suggest expression in pyramidal neurons in one area of the cortex is relevant in intelligence, but expression in the other cell-types and areas of the cortex may not be [42]. A further caveat to this interpretation is that the full cellular composition of the cortex (and the brain overall) is not reflected in the KI mouse superset, and as such our conclusions about wider effects must be constrained.

Genes with higher specificity to pyramidal neurons were enriched for associations with intelligence. However, the location of the most interesting neuron population is not yet clear. Observed enrichment in the pyramidal cells of the CA1 region of the hippocampus was lost when accounting for gene expression in pyramidal cells of the somatosensory cortex. An uncaptured population of pyramidal neurons (for example, in the frontal cortex) may exist that similarly overlaps in expression with the somatosensory pyramidal neurons, and accounts for the enrichment observed in the latter population. While the KI superset is the largest

brain scRNAseq resource to date, it covers a limited set of regions and developmental stages [16].

Tissue and cell-specific analyses were performed using related but distinct approaches, namely linear regression of binned specificity and heritability enrichment analysis of the top 10% of specific genes. Linear regression tended to give smaller p-values than heritability enrichment analysis (Supplementary Table 3). While this may be a statistical artefact generated by the differing assumptions of the methods, this could also result from the pattern of enrichment. For example, a small set of genes associated with intelligence, all with very high specificity to a given tissue would result in a lower p-value from heritability enrichment than if the association with intelligence was spread more broadly across genes with generally enriched tissue-specificity.

Increased specificity of expression of genes associated with intelligence was identified across a number of cell-types and tissues. However, increased specificity of expression does not demonstrate a functional role for these genes in these cell-types. Conversely genes with similar expression profiles between tissues might differ in their action in functionally-relevant ways. As such, our results are suggestive of cell-types and tissues of interest for further study, but are by no means a conclusive explanation of the biological action of intelligence-associated genes.

The results of analyses in the KI mouse superset for intelligence can be contrasted with those for schizophrenia [16]. Both phenotypes initially demonstrate enrichment for the same hippocampal and somatosensory pyramidal neuron populations. However, conditional analyses in schizophrenia demonstrated significant differences in the results between schizophrenia and intelligence - the enrichment observed in the somatosensory cortex in schizophrenia could be fully

18

explained by that in the hippocampus, while the opposite pattern was observed in intelligence. Additionally, unlike in schizophrenia, genes implicated in intelligence also showed specificity in midbrain embryonic cells. Analyses in schizophrenia implicated more level two cell types than were significant in our analyses. Such differences may reflect differences in the common-variant contribution to variance in intelligence and schizophrenia; despite a similar effective sample size to the schizophrenia GWAS (40 675 cases, 64 643 controls, N_{eff} ~ 100K), this GWAS identified only 25 loci compared with the 140 associated with schizophrenia [43]. This highlights that intelligence is potentially more polygenic than schizophrenia, despite similar estimates of heritability captured by common variants.

There is not a direct link between heritability and the molecular genetics of a trait (beyond a non-zero heritability implying that overall genetic similarity is correlated with phenotypic variance). Heritability alone is not informative about the potential biological consequences of statistical associations, except in placing a limit on explanatory power. In this GWAS, common variants explained 22% of the variance in intelligence (and gene-wise associations will only capture a portion of that). A considerable proportion of variance in intelligence is not captured in our analyses, and therefore it should be noted that our conclusions drawn on only one part of the genetic and non-genetic correlates of this trait.

The KI mouse superset has a number of strengths; it is the largest and broadest scRNAseq dataset to date, captures extra-nuclear as well as nuclear transcripts, and was generated using identical methods (Supplementary Note [16]). However, there are also limitations to the use of this dataset. One issue is that the expression data used is derived from mice rather than humans. Gene expression in the brain is conserved across mammals, such that the principal axes of variation in

comparative studies of gene expression capture inter-tissue, rather than inter-species, variance [44]. Furthermore, there is a high degree of conservation of gene expression between mouse and human brains specifically [16, 45, 46]. Previous analyses using the KI mouse superset have made extensive comparison between mouse and human gene expression and found high concordance in mapping mouse to human genes [16]. Further discussion of the use of mouse-derived data, including references for data access, is included in the Supplementary Note and in [16].

Nevertheless, cell types which are not enriched for genes associated with intelligence in this study should not be prematurely excluded, as some cell-types are not present in the dataset and others will have dissimilar functions or have been exposed to different evolutionary pressures in mouse and in human. Intelligence is a major characteristic that differentiates humans from other mammals [47]. As such, it may be the case that genes with higher specificity to regions dissimilar between humans and mice could be enriched for associations with intelligence, which would not be captured by this approach.

Our results highlight potential insights beyond those gleaned from tissue-specific expression patterns. Several genes previously implicated in ID/DD were present in loci associated in the GWAS. An example of this is *GMPPB* (GDP-Mannose Pyrophosphorylase B), which is in locus 11 of the GWAS results, and was genome-wide significant in gene-wise analyses. It is a member of the "rare genetic neurological disorder" gene set (significantly enriched, specifically in neural tissues), and the expression of *GMPPB* was predicted to be downregulated in the cortex of individuals with higher intelligence. Rare loss of function mutations in the *GMPPB* gene have been identified as causal mutations in tens of individuals with muscular dystrophies and myasthenias, many of which present with mild to severe ID [48–52].

The product of this gene is important for the glycosylation of a-DG (alpha-dystroglycan - the dystroglycan gene *DAG* is also present in locus 11 and is a significantly associated gene in this analysis [48]). Glycosylation is required for the interaction of a-DG with extracellular ligands, with a variety of consequences including the organization of axon guidance [53]. Our results, in the context of the medical genetic literature, suggest the effects on axon guidance by glycosylated a-DG may be an area worthy of further exploration to understand the biology of intelligence. Specifically, we might tentatively hypothesize that *GMPPB* acts via a gradated effect, in which too little or no expression of the gene causes intellectual disability, whereas an optimal amount may lead to higher intellectual ability, while too much results in a (mild) decrease in intellectual ability. This hypothesis would require neurobiological evidence to test.

However, the observed overall overlap of all ID/DD genes with loci from the GWAS does not differ from that expected by chance. This lack of significant enrichment is partly reflected in the pathway analysis - there are several overlapping gene sets designed to capture ID/DD genes. Of these, only "rare genetic neurological disorder" was significantly enriched following correction. Further insight is obtained from conditional gene-set analyses - although the overall "intellectual disability" gene set is only nominally associated with intelligence, stratifying this gene set demonstrates considerable enrichment of mild intellectual disability genes, independent of gene expression in somatosensory pyramidal neurons. The presence of genes causative of ID/DD in loci associated with intelligence in the normal range, and the enrichment of specific pathways and gene sets derived from the ID/DD literature, may indicate shared biology between ID/DD and normal intelligence. However, the lack of broad enrichment of all such genes suggests that there may

also be distinct pathways contributing to normal intelligence that are not commonly affected in ID/DD. ID/DD is not a single condition, but a group of disorders with differing etiologies - our results are still consistent with a two-group model of ID/DD etiology [54–56].

The meta-analysis results presented herein extend previous findings [4]. The results are largely consistent with those previously reported, which is unsurprising. More genes were identified in our gene-wise analysis due to an analytical decision to extend the boundaries by which each gene is defined 35kb upstream and 10kb downstream of the coding region [57, 58]. Defining genes using this boundary (as opposed to no boundary extension) captures additional transcriptional elements - these may be specific to the target gene, but may also capture elements with more distal regulatory effects.

Deriving testable biological hypotheses from the statistical associations of GWAS results is one of the central challenges for the immediate future of complex genetics [3, 15]. The provision of high-quality reference datasets encompassing genetic information from variation to translation, and the integration of genomic data to such reference data is invaluable to this aim [16, 18]. We have demonstrated that some insights into the biology of intelligence can be derived from GWAS, and have suggested potential avenues for further exploration. Our results could indicate that intelligence represents optimal pyramidal neuron functioning. Cognitive tests are highly correlated with general intelligence (*g*), which may depend on pyramidal neuron function [7].

Understanding how biology underlies variation in intelligence is an active area of research that is beginning to yield results. Unifying these new genetic results with data from multiple approaches, can increase the power of each approach, has the

potential to yield greater understanding of the biology of intelligence, which in turn could inform the study of many health-related phenotypes with which intelligence is correlated.

Author contributions

GB, DP and PFS conceived the study. JRIC, JB, HAG, PRJ, JS and NS performed statistical analyses. RP, ABM, SL, GC and JH acquired data. JRIC and GB wrote the manuscript. All authors reviewed the manuscript.

Conflict of Interest Statement

PF Sullivan reports the following potentially competing financial interests: Lundbeck (advisory committee), Pfizer (Scientific Advisory Board member), and Roche (grant recipient, speaker reimbursement). G Breen reports consultancy and speaker fees from Eli Lilly and Illumina and grant funding from Eli Lilly. All other authors declare no financial interests or potential conflicts of interest.

References

1. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**: 539–542.

2. Lo M-T, Hinds DA, Tung JY, Franz C, Fan C-C, Wang Y *et al.* Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat Genet* [Internet] 2016;Available from: http://dx.doi.org/10.1038/ng.3736

3. Sullivan PF, Agrawal A, Bulik C, Andreassen OA, Borglum A, Breen G *et al.* Psychiatric Genomics: An Update and an Agenda [Internet]. *bioRxiv.* 2017 [cited 2017]; 115600. Available from: http://biorxiv.org/content/early/2017/03/10/115600

4. Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat Genet* [Internet] 2017 [cited 2017];Available from: http://dx.doi.org/10.1038/ng.3869

5. Davies G, Marioni RE, Liewald DC, Hill WD, Hagenaars SP, Harris SE *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol Psychiatry* 2016; **21**: 758–767.

6. Benyamin B, Pourcain B, Davis OS, Davies G, Hansell NK, Brion M-J *et al.* Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Mol Psychiatry* 2013; **19**: 253–258.

7. Deary IJ. Intelligence. *Annu Rev Psychol* 2012; **63**: 453–482.

8. Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res Hum Genet* 2005; **8**: 87–94.

9. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.

10. Hormozdiari F, Zhu A, Kichaev G, Ju CJ-T, Segrè AV, Joo JWJ *et al.* Widespread Allelic Heterogeneity in Complex Traits. *Am J Hum Genet* 2017; **100**: 789–802.

11. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016; **48**: 481–487.

12. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 2017; **169**: 1177–1186.

13. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* 2016; **99**: 139–153.

14. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* 2015; **47**: 1385–1392.

15. Breen G, Li Q, Roth BL, O'Donnell P, Didriksen M, Dolmetsch R *et al.* Translating genome-wide association findings into new therapeutics for psychiatry. *Nat Neurosci* 2016; **19**: 1392–1396.

16. Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar H *et al.* Genetic Identification Of Brain Cell Types Underlying Schizophrenia [Internet]. *bioRxiv*. 2017 [cited 2017]; 145466. Available from: http://biorxiv.org/content/early/2017/06/02/145466

17. Zabaneh D, Krapohl E, Gaspar HA, Curtis C, Lee SH, Patel H *et al.* A genome-wide association study for extremely high intelligence. *Mol Psychiatry* [Internet] 2017 [cited 2017];Available from: http://dx.doi.org/10.1038/mp.2017.121

18. Consortium, G. TEx. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**: 648–660.

19. Deary IJ, Penke L, Johnson W. The neuroscience of human intelligence differences. *Nat Rev Neurosci* 2010; **11**: 201–211.

20. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–2191.

21. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–150.

22. Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. *Eur J Hum Genet* 2016; **24**: 924–930.

23. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17**: R122–R128.

24. 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.

25. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. FUMA: Functional mapping and annotation of genetic associations [Internet]. *bioRxiv*. 2017 [cited 2017]; 110023. Available from: http://biorxiv.org/content/early/2017/02/20/110023.abstract

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

27. 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.

28. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD *et al.* Evaluation Of Chromatin Accessibility In Prefrontal Cortex Of Schizophrenia Cases And Controls [Internet]. *bioRxiv*. 2017 [cited 2017]; 141986. Available from: http://biorxiv.org/content/early/2017/05/25/141986

29. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 2016; **352**: 235–239.

30. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013; **10**: 1213–1218.

31. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015; **526**: 75–81.

32. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**: e1004219.

33. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. *Nature Reviews Genetics* 2016; **17**: 353-364.

34. Team RC. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. *Document freely available on the internet at: http://www r-project org* 2015;

35. Barbeira A, Shah KP, Torres JM, Wheeler HE, Torstenson ES, Edwards T *et al.* MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results [Internet]. *bioRxiv*. 2016 [cited 2017]; 045260. Available from: http://biorxiv.org/content/early/2016/03/23/045260.full-text.pdf+html

36. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* [Internet] 2016;Available from: http://dx.doi.org/10.1093/nar/gkw1055

37. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015; **1**: 417–425.

38. Bayés A, van de Lagemaat LN, Collins MO, Croning MDR, Whittle IR, Choudhary JS *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci* 2011; **14**: 19–21.

39. Wagnon JL, Briese M, Sun W, Mahaffey CL, Curk T, Rot G *et al.* CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet* 2012; **8**: e1003067.

40. Lee JA, Damianov A, Lin CH, Fontes M, Parikshak NN, Anderson ES *et al.* Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron* 2016; **89**: 113–128.

41. Hill WD, Davies G, Harris SE, Hagenaars SP, neuroCHARGE Cognitive Working group, Liewald DC *et al.* Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions. *Transl Psychiatry* 2016; **6**: e980.

42. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015; **347**: 1138–1142.

43. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection [Internet]. *bioRxiv*. 2016 [cited 2016]; 068593. Available from: http://biorxiv.org/content/early/2016/08/09/068593

44. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P *et al.* The evolution of gene expression levels in mammalian organs. *Nature* 2011; **478**: 343–348.

45. La Manno G, Gyllborg D, Codeluppi S, Nishimura K, Salto C, Zeisel A *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* 2016; **167**: 566–580.e19.

46. Zeng H, Shen EH, Hohmann JG, Oh SW, Bernard A, Royall JJ *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* 2012; **149**: 483–496.

47. Flinn MV, Geary DC, Ward CV. Ecological dominance, social competition, and coalitionary arms races. *Evol Hum Behav* 2005; **26**: 10–46.

48. Carss KJ, Stevens E, Foley AR, Cirak S, Riemersma M, Torelli S *et al.* Mutations in GDP-mannose pyrophosphorylase B cause congenital and limb-girdle muscular dystrophies associated with hypoglycosylation of α-dystroglycan. *Am J Hum Genet* 2013; **93**: 29–41.

49. Jensen BS, Willer T, Saade DN, Cox MO, Mozaffar T, Scavina M *et al.* GMPPB-Associated Dystroglycanopathy: Emerging Common Variants with Phenotype Correlation. *Hum Mutat* 2015; **36**: 1159–1163.

50. Rodríguez Cruz PM, Belaya K, Basiri K, Sedghi M, Farrugia ME, Holton JL *et al.* Clinical features of the myasthenic syndrome arising from mutations in GMPPB. *J Neurol Neurosurg Psychiatry* 2016; **87**: 802–809.

51. Belaya K, Rodríguez Cruz PM, Liu WW, Maxwell S, McGowan S, Farrugia ME *et al.* Mutations in GMPPB cause congenital myasthenic syndrome and bridge myasthenic disorders with dystroglycanopathies. *Brain* 2015; **138**: 2493–2504.

52. Cabrera-Serrano M, Ghaoui R, Ravenscroft G, Johnsen RD, Davis MR, Corbett A *et al.* Expanding the phenotype of GMPPB mutations. *Brain* 2015; **138**: 836–844.

53. Wright KM, Lyon KA, Leung H, Leahy DJ, Ma L, Ginty DD. Dystroglycan organizes axon guidance cue localization and axonal pathfinding. *Neuron* 2012; **76**: 931–944.

54. Roberts JA. The genetics of mental deficiency. *Eugen Rev* 1952; **44**: 71–83.

55. McGue M, Gottesman II. Classical and Molecular Genetic Research on General Cognitive Ability. *Hastings Cent Rep* 2015; **45**: S25–S31.

56. Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* 2016; **17**: 9–18.

57. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat Neurosci* 2015; **18**: 199–209.

58. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006; **7**: 29–59.

Figure Legends

Figure 1: a) Manhattan plot and b) QQ plot of meta-analysis results.

Figure 2: Heritability enrichment of genomic annotations (dots) +/- 1 standard error (bars). Horizontal line is p = 8.62×10^{-4}, Bonferroni-corrected threshold for 58 annotations. Vertical line is enrichment = 1 (that is, no enrichment).

Figure 3: Results of tissue- and cell-specific analyses. a) Whole-body analyses; b) Brain tissue analyses; c) KI level 1 cell analyses; d) KI level 2 cell analyses (only those significant in MAGMA competitive analyses after correction for multiple testing shown). Vertical lines are the Bonferroni threshold for each analysis. Results shown are from MAGMA analyses - tissues and cell-types also significant in LDSC analyses are indicated with †. Full results are shown in Supplementary Table 3.

Figure 4: Conditional cell-type enrichment analysis. X-axis lists target cell-types. Y-axis lists other cell-types. Colors correspond to the enrichment probability of the other cell type after conditioning (P). Values of log(P) approaching zero indicate no enrichment after conditioning. Barplot on the right shows the minimum value of P for each cell-type across all conditional analyses (excluding analyses of the target cell-type with itself); the vertical line marks p=0.05. Red box highlights the loss of significant enrichment in CA1 pyramidal neurons when conditioning on somatosensory (SS) neurons.