The genetic basis of human brain structure and function: 1,262 genome-wide associations found from 3,144 GWAS of multimodal brain imaging phenotypes from 9,707 UK Biobank participants

Lloyd T. Elliott[1], Kevin Sharp[1], Fidel Alfaro-Almagro[2], Gwenaëlle Douaud[2], Karla Miller[2], Jonathan Marchini[1,3†‡], Stephen Smith[2†‡]

[1] Department of Statistics, University of Oxford, Oxford, UK.

[2] FMRIB, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK.

[3] The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.

[†] These authors jointly directed this work.

[‡] Correspondence to: marchini@stats.ox.ac.uk, steve@fmrib.ox.ac.uk

The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. (H.P. Lovecraft, 1890-1937)

Abstract

The genetic basis of brain structure and function is largely unknown. We carried out genome-wide association studies (GWAS) of 3,144 distinct functional and structural brain imaging derived phenotypes (IDPs), using imaging and genetic data from a total of 9,707 participants in UK Biobank. All subjects were imaged on a single scanner, with 6 distinct brain imaging modalities being acquired. We show that most of the IDPs are heritable and we identify patterns of co-heritability within and between IDP sub-classes. We report 1,262 SNP associations with IDPs, based on a discovery sample of 8,426 subjects. Notable significant and interpretable associations include: spatially specific changes in T2* in subcortical regions associated with several genes related to iron transport and storage; spatially extended changes in white matter micro-structure associated with genes coding for proteins of the extracellular matrix and the epidermal growth factor; variations in pontine crossing tract neural organization associated with genes that regulate axon guidance and fasciculation during development; and variations in brain connectivity associated with 14 genes that contribute broadly to brain development, patterning and plasticity. Our results provide new insight into the genetic architecture of the brain with relevance to complex neurological and psychiatric disorders, as well as brain development and aging.

Introduction

Brain structure, function and connectivity are known to vary between individuals in the human population. Changes in these features have been identified in many neurological and psychiatric disorders such as Alzheimer's disease[1], amyotrophic lateral sclerosis[2], schizophrenia[3], major depression[4], autism[5], bipolar disorder[6] and drug addiction[7]. These effects can be measured non-invasively using Magnetic Resonance Imaging (MRI) and provide intermediate or endo-phenotypes that can be used to assess the genetic architecture of such traits.[8]

MRI is a versatile imaging technique with different brain imaging modalities used to separately assess brain *anatomy*, *function*, and *connectivity*. Measures of brain *anatomy* include brain tissue and structure volumes, such as total grey matter volume and hippocampal volume, while other modalities allow the mapping of different biological markers such as venous vasculature, microbleeds and aspects of white matter (WM) micro-structure. Brain *function* is typically measured using task-based functional MRI (tfMRI) in which subjects perform tasks or experience sensory stimuli, and uses imaging sensitive to local changes in blood oxygenation and flow caused by brain activity in grey matter. Brain *connectivity* can be divided into functional connectivity, where spontaneous temporal synchronisations between brain regions are measured using fMRI with subjects scanned at rest, and structural connectivity, measured using diffusion MRI (dMRI), which traces the physical connections between brain regions based on how water molecules diffuse within white matter tracts".

A new resource for relating neuroimaging measures to genetics is the prospective UK Biobank study. UK Biobank is a rich epidemiological resource including lifestyle questionnaires, physical and cognitive measures, biological samples (including genotyping) and medical records in a cohort of 500,000 volunteers[9]. Participants were 40–69 years of age at baseline recruitment, a major aim being to characterize subjects before disease onset . Identification of disease risk factors and early markers should increase over time with emerging clinical outcomes in significant numbers[10].

An imaging extension to the existing UK Biobank study was funded in 2016 to scan 100,000 subjects from the existing cohort, aiming to complete by 2022. Imaging includes MRI of the brain, heart and body, low-dose X-ray bone and joint scans, and ultrasound of the carotid arteries. Imaging takes place in three centres across the UK, all using identical imaging hardware. The brain imaging component attempts to capture imaging phenotypes relevant to the widest possible range of diseases and hypotheses, including three structural modalities, resting and task fMRI, and diffusion MRI[11].

Unlike most of the measurements included in the UK Biobank resource (for example, tobacco use and body mass index), raw imaging data is not a directly useful source of information. A fully automated image processing pipeline (primarily based on the FMRIB Software Library, FSL[12]) has been developed for UK Biobank that removes artefacts and renders images comparable across modalities and participants. The pipeline also generates thousands of image-derived phenotypes (IDPs), distinct individual measures of brain structure and function[11,13]. Example IDPs include the volume of grey matter in many distinct brain regions, and measures of functional and structural connectivity between specific pairs of brain areas.

Another key component of the UK Biobank resource has been the collection of genome-wide genetic data on every participant using a purpose-designed genotyping array. A custom quality control, phasing and imputation pipeline was developed to address the challenges specific to the experimental design, scale, and diversity of the UK Biobank dataset. The genetic data was publicly released in July 2017 and consists of ~96 million genetic variants in ~500,000 study participants.[14]

Most brain imaging studies consist of small, well defined cohorts of subjects, with sample size being constrained by cost and time of imaging acquisition. To maximise statistical power and biological interpretability, most studies aim to acquire imaging data as consistently as possible across all subjects; this includes minimizing changes to: MRI hardware; MRI software and protocol parameters (when acquiring the raw data); and algorithms used for image processing. The largest brain imaging studies have collected a few thousand subjects at most, and few have collected genetic data in parallel. A notable exception is the ENIGMA meta-analysis project that is pooling data from many independent studies, currently summing to more than 30,000 subjects; while the heterogeneity of modalities and protocols across the many studies combined by ENIGMA is much greater than found within individual focused studies, this is to some degree ameliorated by the very high total subject numbers, and there has been success in finding genetic associations via meta-analysis for some phenotypes such as sub-cortical brain volumes[15,16]. The Human Connectome Project (HCP), while aiming for more modest numbers (but still being very large compared with most studies) has acquired extremely high quality functional and structural imaging in over 1,000 healthy young adults on a common imaging platform, together with genome-wide array genotypes and deep phenotyping, and has recently started expanding that dataset to add similar imaging data from hundreds of participants from ages 4 up to 100 years[17]. In UK Biobank, the combination of very large subject numbers with imaging data collected on a maximally homogeneous imaging platform and protocol is a unique feature.

Joint analysis of the genetic and brain imaging datasets produced by the UK Biobank dataset presents a unique opportunity for starting to uncover the genetic bases of brain structure and function, including genetic factors relating to brain development, aging and disease. In this study we carried out GWAS for 3,144 IDPs at 11,734,353 SNPs (single-nucleotide polymorphisms) in up to 8,428 individuals having both genetic and brain imaging data.

Methods

Imaging data and derived phenotypes

The UK Biobank Brain imaging protocol consists of 6 distinct modalities covering structural, diffusion and functional imaging, summarised in **Table 1**. For this study, we used data from the February 2017 release of ~10,000 participants' imaging data.

Modality name	Type	Description
T1-weighted image (**T1**)	Structural	Measures anatomical features based on contrast between grey and white matter and other tissues.
T2-weighted FLAIR image (**T2 FLAIR**)	Structural	Provides a different contrast between tissues (compared with T1), and is sensitive to some pathologies such as white matter lesions.
Susceptibility-weighted imaging (**swMRI** or **SWI**)	Structural	Can be processed in multiple ways to reflect venous vasculature, microbleeds, aspects of micro-structure (local cellular structure) and biochemical processes such as iron deposition.
Diffusion-weighted imaging (**dMRI**)	Structural connectivity	Measures movement of water molecules within their local tissue environment, allowing for the estimation of long-range structural connectivity and local microstructure.
Resting-state functional MRI (**rfMRI**)	Functional connectivity	Measures dynamic changes in blood oxygenation associated with intrinsic brain activity, to assess functional connectivity via temporal similarities between brain regions.
Task functional MRI (**tfMRI**)	Functional activation	Functional imaging while subject performs a particular task or experiences a sensory stimulus.

Table 1 : **Brain Imaging modalities**. An overview of the 6 different brain imaging types used in the UK Biobank study.

The raw data from these 6 modalities has been processed for UK Biobank to create a set of imaging derived phenotypes (IDPs)[11,13]. Those are available from

UK Biobank, and it is the IDPs from the February 2017 data release that we used in this study.

In addition to the IDPs directly available from UK Biobank, we created two extra sets of IDPs. Firstly, we used the FreeSurfer v6.0.0 software[18,19] to model the cortical surface (inner and outer 2D surfaces of cortical grey matter), as well as modelling several subcortical structures. We used both the T1 and T2-FLAIR images as inputs to the FreeSurfer modelling. FreeSurfer estimates a large number of structural phenotypes, including volumes of subcortical structures, surface area of parcels identified on the cortical surface, and grey matter cortical thickness within these areas. The areas are defined by mapping an atlas containing a canonical cortical parcellation onto an individual subject's cortical surface model, thus achieving a parcellation of that surface. Here we used two atlases in common use with FreeSurfer: the Desikan-Killiany–Tourville atlas (denoted "DKT" [20]) and the Destrieux atlas (denoted "a2009s" [21]). The DKT parcellation is gyral-based, while Destrieux aims to model both gyri and sulci based on the curvature of the surface. Cortical thickness is averaged across each parcel from each atlas, and the cortical area of each parcel is estimated, to create two IDPs for each parcel. Finally, subcortical volumes are estimated, to create a set of volumetric IDPs.

Secondly, we applied a dimension reduction approach to the large number of functional connectivity IDPs. Functional connectivity IDPs represent the network "edges" between many distinct pairs of brain regions, comprising in total 1695 distinct region-pair brain connections. In addition to this being a very large number of IDPs from which to interpret association results, these individual IDPs tend to be significantly noisier than most of the other, more structural, IDPs. Hence, while we did carry out GWAS for each of these 1695 connectivity IDPs, we also reduced the full set into just 6 new summary IDPs using data-driven feature identification. This used independent component analysis (ICA[22]), applied to all functional connectivity IDPs from all subjects, to find linear combinations of IDPs that are independent between the different features (ICA components) identified[23]. The ICA feature estimation was carried out with no use of the genetic data, and was applied to maximize independence between component IDP weights (as opposed to subject weights). Split-half reproducibility (across subjects) was used to optimize both the initial dimensionality reduction (14 eigenvectors from a singular value decomposition was found to be optimal) and also the final number of ICA components (6 ICA components was optimal, with reproducibility of ICA weight vectors greater than r=0.9). The resulting 6 ICA features were then treated as new IDPs, representing 6 independent sets (or, more accurately, linear combinations) of the original functional connectivity IDPs. These 6 new IDPs were added into the GWAS analyses. The 6 ICA features are visualized in **Supplementary Figure S10**.

We grouped all 3,144 IDPs into 9 groups (**Table 2)**, each having a distinct pattern of missing values (not all subjects have usable, high quality data from all modalities)[11]. For the GWAS in this study we did not try to impute missing IDPs due to low levels of correlation observed across classes.

The distributions of IDP values varied considerably between phenotype classes, with some phenotypes exhibiting significant skew (**Supplementary Figure S1**) which would likely invalidate the assumptions of the linear regression used to test for association. To ameliorate this issue we quantile normalized each of the IDPs before association testing. This transformation also helps avoid undue influence of outlier values. We also (separately) tested an alternative process in which an outlier removal process was applied to the un-transformed IDPs; this gave very similar results for almost all association tests, but was found to reduce the significance of a very small number of associations and so this possible alternative method for IDP “preprocessing” was not followed through (data not shown).

IDP Group name	Description	Number of IDPs	Number of complete samples
T1-SIENAX	White, grey and cerebrospinal fluid (CSF) volumes	10	8,428
T1-FIRST	7 Sub-cortical volumes x3 (left, right and left+right); brain-stem volume	22	8,428
T1-FAST_ROIs	Grey matter partial volume for 139 regions of interest (ROIs)	139	8,427
T2-FLAIR-BIANCA	Total white matter hyperintensity volume	1	7,705
SWI-T2*	Signal intensity in 7 distinct subcortical structures x 3 (left, right and left+right)	21	7,778
FreeSurfer	Cortical areas and thicknesses based on 2 different cortical atlases; subcortical volumes	483	8,411
dMRI	6 Diffusion tensor and 3 microstructure modelling measures, on each of 75 white matter tract regions	675	7,532
tfMRI	Signal strength in task activated regions	16	7,612
rfMRI	Resting state fluctuation amplitudes in regions from two functional parcellations, and functional network connectivity between all pairs of regions + ICA dimension reduced connectivity	1,777	7,916

Table 2 : **Imaging derived phenotype (IDP) grouping**. The 3,144 IDPs grouped according to modality and missing data patterns. Column 1: short descriptive name for each IDP group. Column 2: a short description of each IDP grouping. Column 3: the number of IDPs in each group. Column 4: the number of subjects with fully observed IDPs in each group.

Genetic data processing

We used the imputed dataset made available by UK Biobank in its July 2017 release[14]. This dataset consists of >92 million autosomal variants imputed from the Haplotype Reference Consortium (HRC) reference panel[24] and a merged UK10K + 1000 Genomes reference panel[25]. We first identified a set of 12,623 participants who had also been imaged by UK Biobank. We then applied filters to remove variants with minor allele frequency (MAF) below 0.1% and with an imputation information score[26] below 0.3, which reduced the number of SNPs to 18,174,817. We then kept only those samples (subjects) estimated to have white British ancestry using the sample quality control information provided centrally by UK Biobank[14] (using the variable *in.white.British.ancestry.subset* in the file *ukb_sqc_v2.txt*); population structure can be a serious confound to genetic association studies[27], and this type of sample filtering is standard. This reduced the number of samples to 8,522. The UK Biobank dataset contains a number of close relatives (3rd cousin or closer). We therefore created a subset of 8,428 nominally unrelated subjects following similar procedures in Bycroft et al. (2017). After running GWAS on all the (SNP) variants we applied three further variant filters to remove variants with a HWE (Hardy-Weinberg equilibrium) p-value less than 10^{-7}, remove variants with MAF<0.1% and to keep only those variants in the HRC reference panel. This resulted in a dataset with 11,734,353 SNPs.

We constructed a set of 930 additional samples to use for replicating the associated variants found in this study. The 1,279 samples with imaging data that we did not use for the main GWAS had been primarily excluded due to not being in the white British subset. An examination of these samples according the genetic principal components (PCs) revealed that many of those samples are mostly of European ancestry (**Supplementary Figure S2**). We selected 930 samples with a 1st genetic PC < 14 from **Supplementary Figure S2** and these constituted the replication sample.

Potential Confounds for brain IDP GWAS

There are a number of potential confounding variables when carrying out GWAS of brain IDPs. We used three sets of covariates in our analysis relating to (a) imaging confounds (b) measures of genetic ancestry, and (c) non-brain imaging body measures.

We identified a set of variables likely to represent imaging confounds, for example being associated with biases in noise or signal level, corruption of data by head motion or overall size changes. For many of these we generated various numerical versions (for example, using quantile normalization and also outlier removal, to generate two versions of a given variable, as well as including the squares of these to help model nonlinear effects of the potential confounds). This was done in order to generate a rich set of covariates and hence reduce as much as possible potential confounding effects on analyses such as the GWAS, which are particularly of concern when the subject numbers are so high.[11]

Age and sex are can be variables of biological interest, but can also be sources of imaging confounds, and here were included in the confound regressors. Head motion is summarized from the rfMRI and tfMRI as the mean (across timepoints) of the mean (across the brain) estimated displacement (in mm) between one timepoint and the next. Head motion can be a confounding factor for all modalities and not just those comprising timeseries of volumes, but are only readily estimable from the timeseries modalities.

The exact location of the head and the radio-frequency receive coil in the scanner can affect data quality and IDPs. To help account for variations in position in different scanned participants, several variables have been generated that describe aspects of the positioning (see **URLs**). The intention is that these can be useful as “confound variables”, for example these might be regressed out of brain IDPs before carrying out correlations between IDPs and non-imaging variables. TablePosition is the Z-position of the coil (and the scanner table that the coil sits on) within the scanner (the Z axis points down the centre of the magnet).

BrainCoGZ is somewhat similar, being the Z-position of the centre of the brain within the scanner (derived from the brain mask estimated from the T1-weighted structural image). BrainCoGX is the X-position (left-right) of the centre of the brain mask within the scanner. BrainBackY is the Y-position (front-back relative to the head) of the back of brain mask within the scanner.

UK Biobank brain imaging aims to maintain as fixed an acquisition protocol as possible during the 5-6 years that the scanning of 100,000 participants will take. There have been a number of minor software upgrades (the imaging study seeks to minimise any major hardware or software changes). Detailed descriptions of every protocol change, along with thorough investigations of the effects of these on the resulting data, will be the subject of a future paper. Here, we attempt to model any long-term (over scan date) changes or drifts in the imaging protocol or software or hardware performance, generating a number of data-driven confounds. The first step is to form a temporary working version of the full subjects × IDPs matrix with outliers limited (see below) and no missing data, using a variant of low-rank matrix imputation with soft thresholding on the eigenvalues[28]. Next, the data is temporally regularized (approximate scale factor of several months with respect to scan date) with spline-based smoothing. PCA was then applied and the top 10 components kept, to generate a basis set reflecting the primary modes of slowly-changing drifts in the data.

To describe the full set of imaging confounds we use a notation where subscripts “i” indicate quantile normalization of variables, and “m” to indicate median-based outlier removal (discarding values greater than 5 times the median-absolute-deviation from the overall median). If no subscript is included, no normalization or outlier removal was carried out. Certain combinations of normalization and powers were not included, either because of very high redundancy with existing combinations, or because a particular combination was not well-behaved. The full set of variables used to create the confounds matrix are:

- a = age at time of scanning, demeaned (cross-subject mean subtracted)
- s = sex, demeaned

- q = 4 confounds relating to the position of the radio-frequency coil and the head in the scanner (see above), all demeaned
- d = drift confounds (see above)
- m = 2 measures of head motion (one from rfMRI, one from tfMRI)
- h = volumetric scaling factor needed to normalise for head size[29]

The full matrix of imaging confounds is then:
$[\ a\ \ a^2\ \ a{\times}s\ \ a^2{\times}s\ \ a_i\ \ a_i^2\ \ a_i{\times}s\ \ a_i^2{\times}s\ \ m_m\ \ m_m^2\ \ h_m\ \ q_m\ \ q_m^2\ \ d_m\ \ m_i\ \ h_i\ \ q_i\ \ q_i^2\ \ d_i\ ]$
Any missing values in this matrix are set to zero after all columns have had their mean subtracted. This results in a full-rank matrix of 53 columns (ratio of maximum to minimum eigenvalues = 42.6).

Genetic ancestry is a well known potential confound in GWAS. We ameliorated this issue by filtering out samples with non-white British ancestry. However, a set of 40 genetic principal components (PCs) has been provided by UK Biobank[14] and we used these PCs as covariates in all of our analysis. The matrix of imaging confounds, together with a matrix of 40 genetic principal components, was regressed out of each IDP before the analyses reported here.

There exist a number of substantial correlations between IDPs and non-genetic variables collected on the UK Biobank subjects[11]. Based on this, we also included some analyses involving variables relating to Blood Pressure (Diastolic and Systolic), Height, Weight, Head Bone Mineral Density, Head Bone Mineral Content and 2 principal components from the broader set of bone mineral variables available (see **URLs**). **Supplementary Figure S3** shows the association of these 8 variables against the IDPs and shows significant associations. These are variables that will likely have a genetic basis, at least in part. Genetic variants associated with these variables might then produce false positive associations for IDPs. To investigate this we ran GWAS for these 8 traits (conditioned on the imaging confounds and genetic PCs) (**Supplementary Figures S3**). We also ran a parallel set of IDP GWAS with these “body confounds” regressed out of the IDPs.

Heritability and co-heritability of IDPs

We used a multi-trait mixed model to jointly estimate heritability and genetic correlations between traits. If Y is an NxP matrix of P phenotypes (columns) measured on N individuals (rows) then we use the model

$$Y = U + \varepsilon \tag{1}$$

where U is an NxP matrix of random effects and ε is a NxP matrix of residuals and are modelled using Matrix normal distributions as follows

$$U \sim MN\left(0, K, B\right)$$
$$\varepsilon \sim MN\left(0, I_N, E\right)$$

In this model K is the NxN kinship matrix between individuals, B is the PxP matrix of genetic covariances between phenotypes and E is the PxP matrix of residual covariances between phenotypes. We estimate the covariance matrices B and E using a new C++ implementation of an EM algorithm[30] included in the SBAT software (see **URLs**). For the Kinship matrix (K) in the model we used realised relationship matrices (RRMs) that were calculated for the 8,428 nominally unrelated individuals using fastLMM (see **URLs**). We used the subset of imputed SNPs that were both assayed by the genotyping chips and included in the HRC reference panel. In addition, all SNPs with duplicate rsids were removed. PLINK (see **URLs**) was used to convert imputed genotype calls to thresholded genotypes, as required by fastLMM. We fit the model to several of the groupings of IDPs detailed in **Table 2**. The estimated covariance matrices B and E were used to estimate heritability of each IDP and genetic correlation of pairs of IDPs. Specifically, the heritability of ith IDP in a jointly analyzed group of IDPs is estimated as

$$h_i^2 = \frac{B_{ii}}{B_{ii} + E_{ii}}$$

The genetic correlation between the ith and jth IDPs in a jointly analyzed group of IDPs is estimated as

$$r_{ij} = \frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}$$

Genetic association of IDPs

We used BGENIE v1.2 (see **URLs**) to carryout GWAS of imputed variants against each of the processed IDPs. This program was designed to carryout the large number of IDP GWAS required in this analysis. It avoids repeated reading of the genetic data file for each IDP and uses efficient linear algebra libraries and threading to achieve good performance. The program has already been used by several studies to analyze genetic data from the UK Biobank[31,32]. We fit an additive model of association at each variant, using expected genotype count (dosage) from the imputed genetic data. We ran associated tests on the main set of 8,428 and the replication set of 930 samples.

Identifying associated genetic loci

Most GWAS only analyze one or a few different phenotypes, and often uncover just a handful of associated genetic loci, which can be interrogated in detail. Due to the large number of associations uncovered in this study we developed an automated method to identify, distinguish and count individual associated loci from the 3,144 GWAS (each corresponding to one IDP). For each GWAS we first identified all variants with a –log10 p-value > 7.5. We applied an iterative process that starts by identifying the most strongly associated variant, storing it as a lead variant, and then removing it, and all variants within 0.25cM from the list of variants. This process is then repeated until the list of variants is empty. We applied this process to each GWAS using 2 different filters on MAF: (a) MAF > 0.1%, and (b) MAF > 1%.

Results

Heritability and genetic correlations of IDPs

Figure 1 shows the estimated heritability (h^2) of the IDPs analyzed using SBAT. It shows that the majority of the IDPs have estimated levels of h^2 in the range (0.25, 0.65). Notable exceptions are the task fMRI and resting fMRI IDPs which are estimated to have relatively lower heritability than the structural T1, T2* and FreeSurfer IDPs and the structural connectivity diffusion MRI IDPs.

Supplementary Figure S4 shows the estimated genetic correlations, together with the raw phenotype correlations, for several groups of analyzed IDPs. These plots show that there are a range of both strong and weak positive and negative genetic correlations between the IDPs.

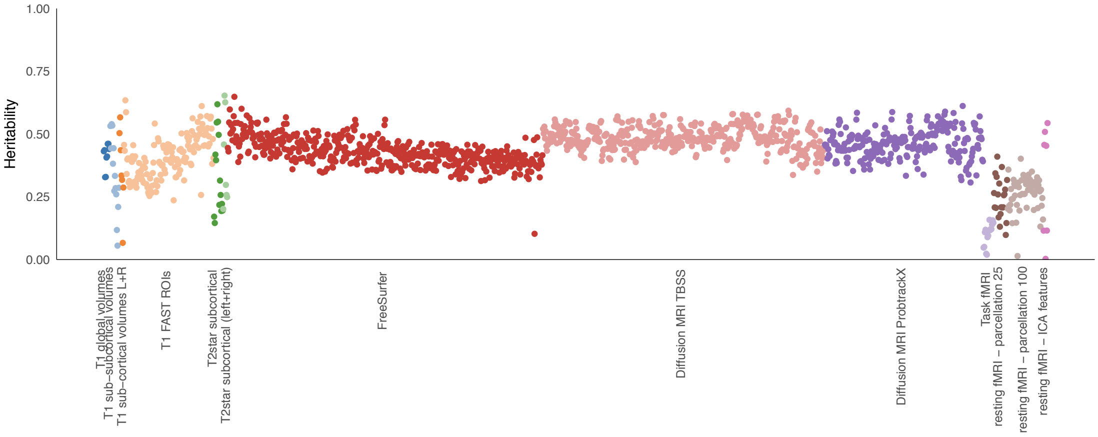

Figure 1 : Estimated heritability of IDPs. Estimated heritability (y-axis) of all of the IDPs analyzed jointly in groups (x-axis). Each point is an IDP. Points are coloured according to IDP groups.

Significant associations between IDPs and SNPs

Using a minor allele frequency filter of 1% and a –log$_{10}$ p-value threshold of 7.5, we found 1,262 significant associations between SNPs and the 3,144 IDPs. The –log10 p-value threshold of 7.5 is an established threshold in the GWAS literature for controlling for the large number of tests carried out and takes into account the correlation structure between test variants, in much the same way that Random Field theory is used to determine brain-wide voxel testing thresholds in fMRI experiments. 455 of these 1,262 associations replicated at the 5% significance level (see **Supplementary Table S1**. Using the estimated effect sizes and allele frequencies at each of the 1,262 SNPs we calculated that the expected number of replications in a sample size of 930 would be 427 (**Supplementary Methods**) which seems to agree well with what we observe, and strongly suggests that increasing the size of the replication sample will improve the number of successful replications. Some associated genetic loci overlap across IDPs; we estimate that there are approximately 427 distinct associated genetic regions, and 91 of these "clusters" has at least one IDP that already replicates at the 5% level.

To correct for the huge number of tests (all IDPs tested against all SNPs) we adjusted the genome-wide significance threshold (-log10 p-value > 7.5) by a factor ($-\log_{10}(3144)=3.5$) to account for the number of IDPs tested, giving a threshold of -log10 p-value > 11. This assumes (incorrectly) that the IDPs are independent and so is likely to be conservative, but we were inclined to be cautious when analyzing so many IDPs. At this threshold we find 368 significant associations between genetic regions and IDPs, which cluster into 38 distinct associated regions (**Table 3, Supplementary Table S2**). 229 of these 368 associated regions replicated at the 5% significance level. Using the estimated effect sizes and allele frequencies at each of the 368 SNPs we calculated that the expected number of replications in a sample size of 930 would be 188 (Supplementary Methods). Taking the most strongly associated SNP in each of the 38 regions, we find that 27 of these replicate at the 5% significance level; on the basis of the actual replication sample size we would expect just 19 to replicate on average (**Supplementary Methods**). We found no appreciable difference between these GWAS results when we included the set of potential body confound measures (**Supplementary Figure S11**).

cluster index	cluster name	# IDPs	top IDP	chr	rsid	pos	ref allele	nonref allele	nonref AF	p value	replication p-value
1	Volume Cerebellum VIIIa (vermis)	1	T1_FAST_ROIs_V_cerebellum_VIIIa	1	rs76934732	76013268	G	A	0.145	8.51E-13	0.052191527
2	dMRI Corpus callosum (genu)	1	dMRI_TBSS_ICVF_Genu_of_corpus_callosum	1	rs2365715	156615114	A	G	0.388	5.38E-12	0.013298419
3	Volume WM lesions	1	T2_FLAIR_BIANCA_WMH_volume	2	rs146896516	56152750	C	A	0.104	4.58E-14	0.35087304
4	rfMRI Cortical and cerebellar motor nodes and edges	1	NODEamps25_0012	2	rs60873293	114092549	G	T	0.217	9.86E-15	0.0950167
5	T2* Pallidum	1	SWI_T2*_pallidum_L+R	2	rs6740926	190326498	C	T	0.038	1.31E-14	0.0003783
6	rfMRI Middle temporal sulcus nodes and edges	1	netmat_ICA_003	3	rs66499884	89659012	T	G	0.402	2.77E-23	0.004825
7	T2* Putamen and pallidum	7	SWI_T2*_putamen_L+R	3	rs4428180	133466374	A	G	0.152	2.23E-22	0.0010337
8	rfMRI Prefrontal and parietal edges	1	netmat_ICA_002	3	rs2279829	147106319	C	T	0.221	8.34E-12	0.002507
9	dMRI Superior cerebellar peduncles	8	dMRI_TBSS_ICVF_Superior_cerebellar_peduncle_L	4	rs4697414	23724255	C	T	0.823	5.83E-24	0.0463126
10	Volume Putamen, ventral striatum, cerebellum VIIIb, IX, X; T2* Pallidum; dMRI Cerebral peduncles	21	IDP_T1_FAST_ROIs_L_ventral_striatum	4	rs13107325	103188709	C	T	0.073	1.04E-42	8.97222E-06
11	dMRI Most WM tracts	199	dMRI_ProbtrackX_ICVF_ilf_r	5	rs67827860	82860485	C	T	0.188	4.06E-37	0.0002194
12	rfMRI Parietal and prefrontal edges	1	netmat_ICA_004	5	rs7442779	92788278	A	G	0.050	8.18E-15	0.0404482
13	dMRI Corpus callosum (genu, body, splenium)	7	dMRI_TBSS_ICVF_Genu_of_corpus_callosum	5	rs4150221	139719991	T	C	0.264	8.43E-20	0.0406349
14	T2* Putamen	3	SWI_T2*_putamen_L+R	6	rs144861591	26072992	C	T	0.074	4.81E-20	0.0035917
15	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	7	rs2286184	84630516	C	T	0.201	5.31E-17	0.0001577
16	dMRI Corpus callosum (genu)	1	dMRI_TBSS_OD_Genu_of_corpus_callosum	7	rs12113919	117612315	C	G	0.416	3.96E-12	0.0018373
17	Volume Brain	2	volume_MaskVol	7	rs2536185	120984041	G	T	0.452	1.30E-16	9.14E-05
18	T2* Putamen	2	SWI_T2*_putamen_L+R	8	rs35469695	23406169	C	G	0.174	2.22E-12	0.2172300
19	Volume Pallidum	3	T1_FIRST_pallidum_volume_L+R	8	rs2923405	42448126	T	G	0.583	3.31E-17	0.0059841
20	T2* Pallidum	2	SWI_T2*_pallidum_L+R	8	rs2978098	101676675	A	C	0.468	6.43E-15	0.3227379
21	Volume Cerebellum	3	T1_FAST_ROIs_L_cerebellum_crus_I	9	rs72754248	119061396	G	A	0.069	1.38E-17	0.2010620
22	T2* Caudate, putamen and pallidum	17	SWI_T2*_caudate_L+R	10	rs10430578	18226714	G	A	0.243	2.73E-31	0.0256921
23	T2* Caudate	3	SWI_T2*_caudate_L+R	10	rs12570727	18425519	G	A	0.394	2.17E-22	0.0006228
24	rfMRI Parietal, temporal and prefrontal nodes	20	NODEamps100_0002	10	rs2274224	96039597	G	C	0.431	6.55E-19	0.0721107
25	rfMRI Prefrontal nodes	6	NODEamps25_0013	10	rs11596664	134280157	C	T	0.439	1.97E-15	0.035958
26	T2* Pallidum	3	SWI_T2*_pallidum_L+R	11	rs11230859	61769972	G	A	0.663	2.31E-17	0.0482947
27	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	11	rs4935898	124742385	G	A	0.048	1.76E-19	0.2465982
28	Volume Mesencephalon (WM cerebellum, brainstem)	3	volume_Right-Cerebellum-White-Matter	12	rs4301837	102336310	T	C	0.501	3.40E-13	0.0122659
29	Volume Hippocampus	2	T1_FAST_ROIs_R_hippocampus	12	rs7315280	117320938	A	G	0.115	7.06E-14	0.6694528
30	Volume Putamen	4	volume_Right-Putamen	14	rs945270	56200473	C	G	0.419	3.67E-14	0.0033166
31	Volume and area of precuneus and cuneus	11	T1_FAST_ROIs_R_intracalc_cortex	14	rs74826997	59628609	T	C	0.125	2.46E-16	0.028780
32	Thickness, area and volume of primary sensorimotor cortex	15	a2009s_lh_S_central_area	15	rs4924345	39639898	A	C	0.081	3.27E-53	1.01E-06
33	Volume 4th ventricle	1	volume_4th-Ventricle	15	rs2464469	58362025	G	A	0.587	3.16E-16	0.2281602
34	dMRI Uncinate	4	dMRI_ProbtrackX_ISOVF_unc_r	16	rs7197215	51449978	A	G	0.566	2.24E-15	0.0001434
35	Volume Cerebellum IX	2	T1_FAST_ROIs_L_cerebellum_IX	17	rs9905515	35261073	G	C	0.230	3.32E-13	0.0002698
36	T2* Caudate and putamen	6	SWI_T2*_putamen_L+R	17	rs668799	40716235	C	T	0.278	1.43E-17	0.0009855
37	Volume WM lesions	1	T2_FLAIR_BIANCA_WMH_volume	17	rs3744020	73871773	G	A	0.188	1.15E-12	0.033604
38	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	18	rs2928990	49421125	T	G	0.898	3.97E-16	0.0022656

Table 3 : Summary of most highly associated SNP-IDP clusters. The table summaries the 38 clusters of SNP-IDP associations. For each cluster the most significant association between a SNP and an IDP is detailed by the chromosome, rsID, base-pair position, SNP alleles, non-reference allele frequency, p-value in the discovery sample and the replication p-value.

Supplementary Figure S6 provides genome-wide association plots (also generally known as Manhattan plots) and QQ-plots for all 3,144 IDPs. **Supplementary Figure S7** provides (for convenience) the same plots for just the subset of IDPs listed in **Table 3**. Having identified a SNP as being associated with a given IDP, it can be useful then to explore the association with all other IDPs. These associations can be visualized in a PheWAS (Phenome wide Association Study) plot. **Supplementary Figure S8** shows the PheWAS plots for all 78 SNPs listed in **Supplementary Table S2**. Examining these plots highlights that it is often the case that a SNP is associated with several IDPs in addition to the IDP that most strongly identified it. In some cases we find SNPs that are associated with IDP measures that span the classes of structural, structural connectivity and functional connectivity measures.

Overall, we found that 4 of the 78 unique SNPs in **Supplementary Table S2** were associated (-$\log_{10}$ p-value > 4.79) with all 3 classes of structural, structural connectivity and functional connectivity measures, and these were all SNPs in cluster 31 of **Table 3** (see pages 61-64 of **Supplementary Figure S8**). This genetic locus is associated with the volume of the precuneus and cuneus, the diffusion MRI measure for the forceps major which is a fibre bundle which connects the occipital lobes and crosses the midline via the splenium of the corpus callosum, and two functional connections (parcellation 100 edges 614 and 619, which connect two cognitive networks - the default mode network and the dorsal attention network). There were 18 SNPs that showed association with both structural and structural connectivity classes but not with the functional measures. There were 2 that showed association with both structural and functional IDP classes but not with the structural connectivity. There were no SNPs that showed association with just the structural connectivity and functional IDP classes. The number of SNPs that showed association with only one of the structural, structural connectivity and functional IDP classes was 30, 12 and 12 respectively. **Supplementary Figure S9** shows the local association plots for the 368 SNP-IDP associations listed in **Supplementary Table S2**

Table 4 shows the pattern of associations stratified by the 9 IDP groups listed in **Table 2**. We found associations at the nominal GWAS threshold of –log10 p-value > 7.5 in all classes except the task fMRI IDPs. Taking account of the number of associations in each group the SWI-T2* group shows a relatively large number of associations.

IDP Group name	Number of IDPs	Number of associated loci with –log10 p > 11	Number of associated loci with –log10 p > 7.5
T1-SIENAX	10	0	14
T1-FIRST	22	5	18
T1-FAST_ROIs	139	24	88
T2-FLAIR-BIANCA	1	2	3
SWI-T2*	21	47	89
FreeSurfer	483	33	185
dMRI	675	225	599
tfMRI	16	0	0
rfMRI	1,777	32	266

Table 4 : Summary of associations by IDP group. The table shows the number of associated genetic loci stratified by IDP group. Column 2: the number of IDPs in each group. Column 3: the number of associated genetic loci (from **Supplementary Table S2**) for each group. Column 4: the number of associated genetic loci (from **Supplementary Table S1**) for each group.

The 368 associations passing –log10 p-value threshold of 11 are listed in **Supplementary Table S2**, where they are organized into 38 distinct clusters; each cluster contains one or more IDPs associated with one or more SNPs within a single genetic locus The strongest IDP-SNP association for each cluster is listed in Table 3. In general, a cluster that contains multiple IDPs tend to contain IDPs of similar feature types; for example, cluster 11 contains 199 IDPs, most of which are primarily reflecting water diffusivity measures in the dMRI data.

Many of these clusters relate to known brain-related genes or relevant GWAS. We replicated (in our 8428 subjects) the same genome-wide significant associations as found in several other GWAS of brain MRI measures. In subcortical regions, we observed a significant association between volumes of the putamen (L/R) and an intergenic region between *KTN1* and *RPL13AP3* (rs945270, P_{min}=3.67E-14, cluster 30). This is the same location identified in a previous GWAS of subcortical volumes as being associated with volume of putamen[15]. Similarly, a significant association was seen between the volume of the hippocampus (L/R) and a region slightly upstream from *HRK* (rs7315280, P_{min}=7.06E-14, cluster 29). This locus was less than 10kb away from two SNPs further upstream from *HRK*, which were previously found to be associated with hippocampal volumes in four separate GWAS of hippocampal and subcortical volumes (rs7294919[33 34]; rs77956314[15 16]). We also identified an association between volume of white matter hyperintensities from T2 FLAIR images ("lesions") and *TRIM47* (rs3744020, P=1.15E-12, cluster 37), the same gene identified in a GWAS of cerebral white matter lesion burden[35], one of the two strongest candidate genes for small vessel disease.

A major source of cross-subject differences seen in T2* data is microscopic variations in magnetic field, often associated with iron deposition in aging and pathology[11]. We identified many associations between T2* measurements in the caudate, putamen and pallidum with genes known to affect iron transport and storage (*TF*[36], rs4428180, P_{min}=2.23E-22, cluster 7; *HFE*[37], rs144861591, P_{min}=4.81E-20, cluster 14; SLC25A37[38], rs35469695, P_{min}=2.22E-12, cluster 18; *FTH1*[39], rs2978098, P_{min}=6.43E-15, cluster 20), as well as neurodegeneration with brain iron accumulation (NBIA) (*COASY*[40], rs668799, P_{min}=1.43E-17, cluster 36). In addition to *TF*, which transports iron from the intestine, and *SLC25A37*, a mitochondrial iron transporter, four further genes were found that are involved in transport of nutrients and minerals: *SLC44A5*[41] (cluster 1), *SLC39A8/ZIP8*[42] (cluster 10), *SLC20A2*[43] (cluster 19) and *SLC39A12/ZIP12*[44] (cluster 22).

While the GWAS results we report here have been based purely on IDPs (which use pre-determined spatial regions over which to summarise the original voxelwise imaging data), it is possible to interrogate the full voxelwise data in order to seek further insight about detailed spatial localization of SNP associations, as well as possibly identifying additional associated areas not already well captured by the IDPs (while keeping in mind the statistical dangers of potential circularity[45]). For instance, we computed the average T2* image across the 5,602 subjects having close to 0 copies of the rs4428180 non-reference allele, and separately computed the average T2* image across the 1,995 subjects having close to 1 copy. (In this case the number of subjects with 2 copies was small (181), resulting in a much noisier average T2* image.) We then subtracted the *copies~1* average T2* image from the *copies~0* image, in order to visualize, voxelwise across the whole brain, areas relating to the significant association found between T2* in the putamen and pallidum and rs4428180. T2* intensity is a quantitative measure and so the group difference image can be thresholded in an interpretable way; the group differences in T2* shown in **Figure 2** were thresholded at 0.8ms.

In **Figure 2** we can therefore see the voxelwise differences (across the whole brain) associated with rs4428180 and 3 additional SNPs, from the 4 most significant GWAS associations with T2* in the putamen (as seen in the Manhattan plot at the top). While the T2* group-average difference images for these 4 SNPs are all thresholded (at 0.8ms) to show the strongest group differences, no further masking of the results was applied, showing the relative spatial specificity of these associations. On the other hand, the 4 SNPs that are all strongly associated with T2* in these subcortical areas, have quite distinct voxelwise patterns, showing that the exact effects of these SNPs are not identical to each other. rs668799 is most strongly associated with T2* changes in posterior putamen (and anterior caudate); rs10430578 most strongly in anterior (and slightly medial) putamen (and anterior caudate); rs144861591 strongly with most of the putamen, both anterior and posterior, (and less so in caudate); rs4428180 most strongly in posterior and inferior putamen, and also in pallidum. These effects of rs4428180 (gene *TF)* were found not just in the lenticular nucleus (putamen and pallidum), but also in much smaller subcortical structures, including caudate nucleus, red nucleus, substantia nigra, subthalamic nucleus, lateral geniculate nucleus of the thalamus (seen in bottom-right of the figure) and the dentate nucleus of the cerebellum.

Figure 3 shows the pattern of association in and around SNP rs4428180 with the IDP of *T2* in the left putamen plus the right putamen*, and **Figure 4** is a PheWAS plot for rs4428180, which shows the overall pattern of association with all 3,144 IDPs.

Interestingly, three clusters relating to white matter: cluster 2 (with one dMRI microstructural measure in the genu of the corpus callosum), cluster 3 (measuring the volume of white matter lesions) and cluster 11 (encompassing multiple dMRI measures of most of the WM tracts), were strongly associated with three different genes coding for proteins of the extracellular matrix (ECM) (*BCAN*, rs2365715, P=5.38E-12, cluster 2; *EFEMP1*, rs146896516, P=4.58E-14, cluster 3; *VCAN*, rs67827860, P_{min}=4.06E-37, cluster 11). In particular, *BCAN* and *VCAN* both code for chondroitin sulfate proteoglycans of the ECM, which are especially crucial for synaptic plasticity[46] and myelin repair[47]. VCAN is, for instance, increased in association with astrocytosis in multiple sclerosis[48], while both BCAN and VCAN are differentially regulated following spinal cord injury[49]. *BCAN*, *EFEMP1* and *VCAN* have been further associated in three separate GWAS with stroke[50], site of onset of amyotrophic lateral sclerosis[51] and major depressive disorder[52]. Incidentally, *EFEMP1* is characterised by tandem arrays of epidermal growth factor (EGF)-like domains and a C-terminal fibulin, and we also identified a strong association between the whole of the corpus callosum (genu, body and splenium) and *HBEGF*, a heparin-binding EGF-like growth factor (rs4150221, P_{min}=8.43E-20, cluster 13). Similarly to *BCAN* and *VCAN*, *HBEGF* plays an important role in oligodendrocytes development and helps recovering WM injury in preterm babies[53]. Remarkably, this means that almost all prosencephalic WM-related dMRI IDPs that were found to be associated with genes in this study (N=219) were with genes coding for proteins involved in either the extracellular matrix, the epidermal growth factor, or both.

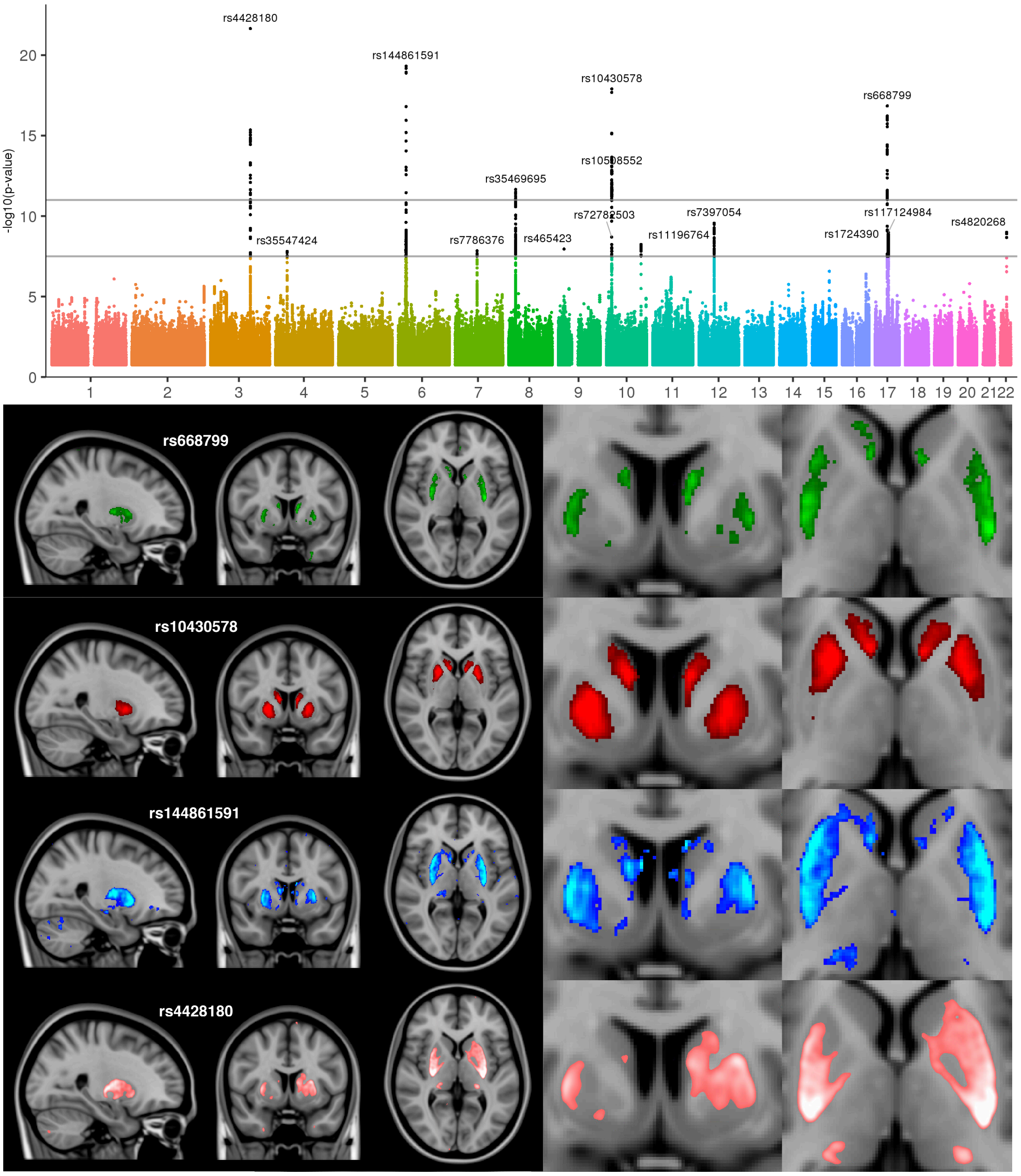

Figure 2: Manhattan plot and detailed spatial investigations into associations between T2* and 4 SNPs. The Manhattan plot relates to the original GWAS for the IDP *T2* in the putamen* (left plus right). The spatial maps show that the 4 SNPs most strongly associated with T2* in the putamen have distinct voxelwise patterns of effect. All T2* data was first transformed into standard (MNI152) space before being averaged within different allele subject groups for different SNPs. The standard MNI152 image is used as the background image for the spatial maps; the left side of the brain is shown on the right, and the slices are located at (26,6,3). All group difference images (colour

overlays) are thresholded at a T2* difference of 0.8ms. The group differences for all 4 SNPS are mean(*copies~0*)-mean(*copies~1*).

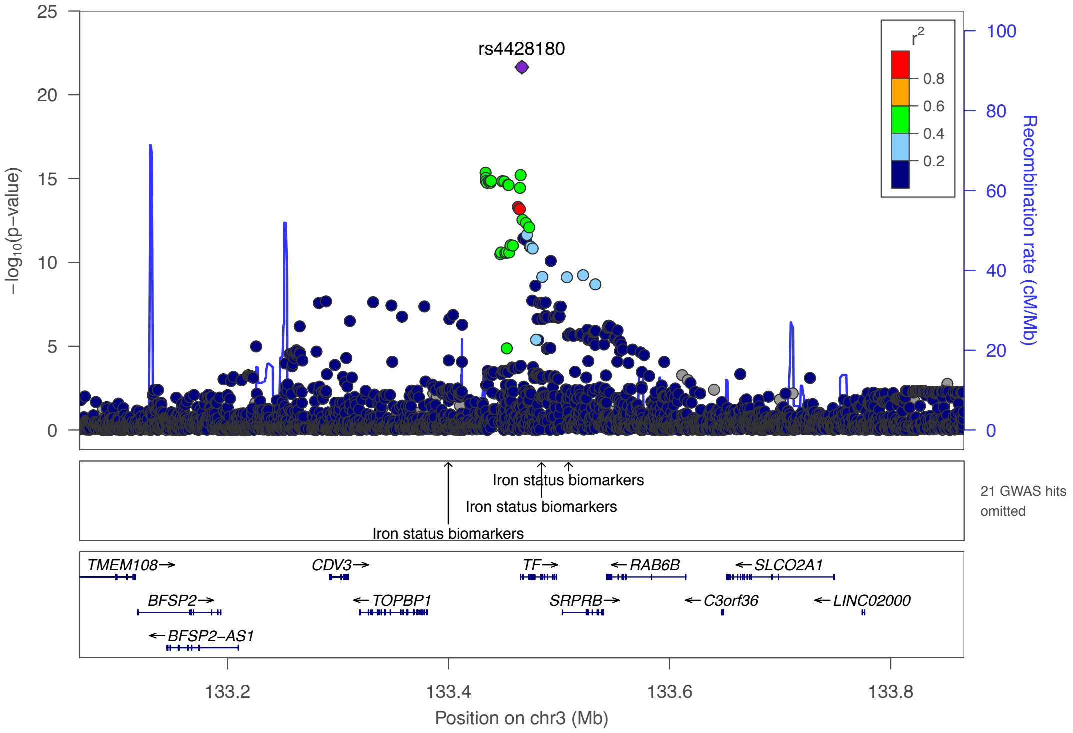

Figure 3 : Genetic association of SNPs with T2* in the left plus right putamen, centered on rs4428180. In the top panel SNPs are plotted by their positions on the chromosome against association with the IDP ($-\log_{10}$ P value) on the left *y* axis. Points are coloured by their local linkage disequilibrium (LD) pattern with the focal SNP rs4428180 (purple diamond). Below the main plot are two tracks that show existing GWAS associations, and position and orientation of local genes.

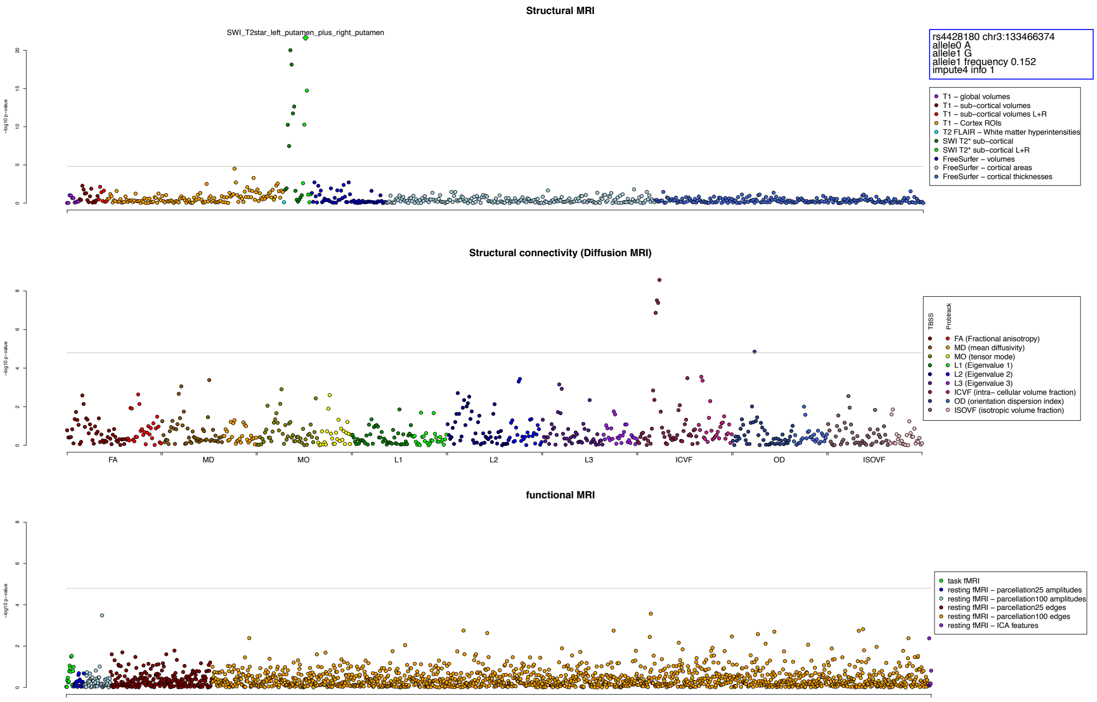

Figure 4 : PheWAS plot for SNP rs4428180. The association (-log$_{10}$ p-value on the y-axis) for the SNP rs4428180 with each of the 3,144 IDPs. The IDPs are arranged on the x-axis in the three panels : (top) Structural MRI IDPs, (middle) Structural connectivity/micro-structure dMRI IDPs, (bottom) functional MRI IDPs. Points are coloured to delineate subgroups of IDPs and detailed in the legends. Summary details of SNP rs4428180 are given in the top right box. The grey line shows the Bonferroni multiple testing threshhold of 4.8.

Figure 5 shows associations between one measure from the dMRI data and SNP rs67827860. The measure most strongly associated is ICVF (intra-cellular volume fraction), estimated from the NODDI modelling (neurite orientation dispersion and density imaging) [54]. The ICVF parameter aims to quantify neurite density, predominantly intra-axonal water in white matter, by estimating where water diffusion is restricted. The more simplistic diffusion tensor model decomposes the same data differently into mean diffusivity (MD, which inversely correlates with ICVF since restricted diffusion manifests as low apparent diffusivity) and diffusion eigenvalues (with the L2 and L3 components corresponding to apparent diffusivity along the directions with greatest restriction - i.e., those directions that drive the ICVF estimates). The figure shows voxelwise mapping of the effect of this SNP. Unlike the previous example of (spatially) very focal effects in T2*, the effect of this SNP are extremely widespread across much of the white matter (hence the size of cluster 11).

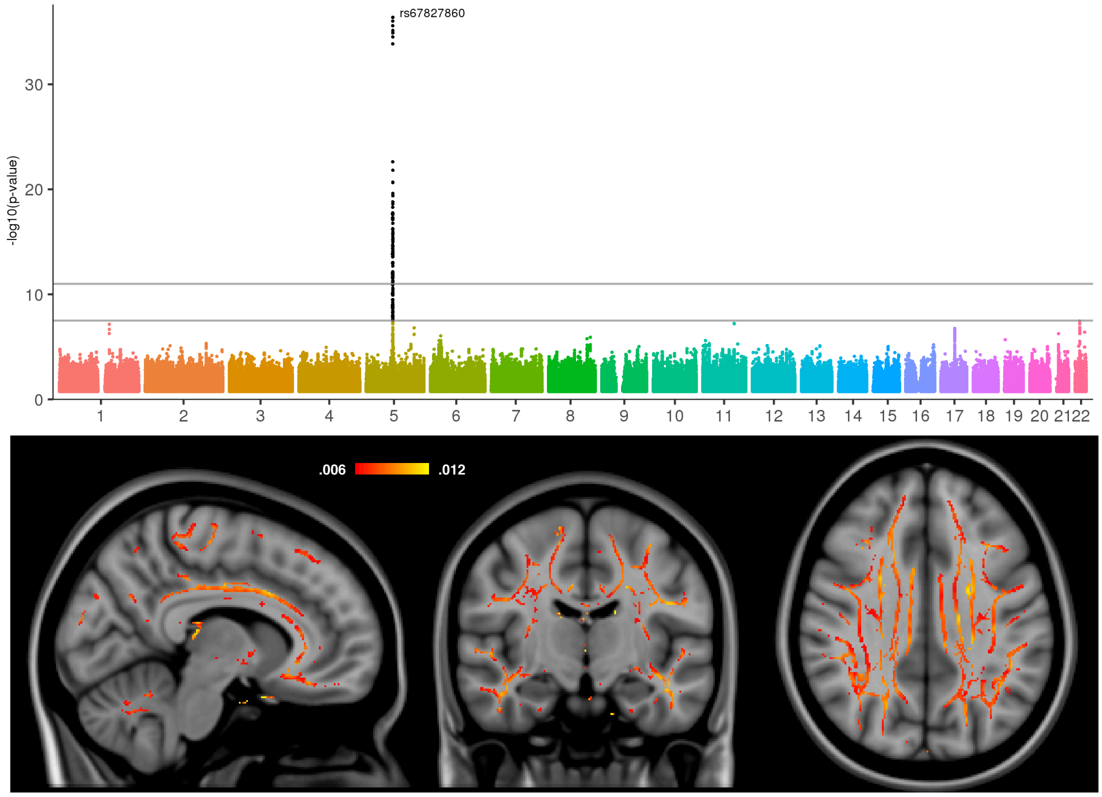

Figure 5: Manhattan plot and detailed spatial mapping of the association between the *intra-cellular volume fraction* micro-structure measure and rs67827860. The Manhattan plot relates to the original GWAS for the IDP *ProbtrackX_ICVF_ilf_r* (inferior longitudinal fasciculus). All ICVF data was transformed into standard space and masked with a white-matter tract-centres skeleton mask. The quantitative (though unitless) ICVF measure was then averaged across all 4,957 subjects with ~0 copies of the non-reference allele, and the average from all 2,304 subjects having ~1 copy was subtracted from that, for display in colour here. The difference was thresholded at 0.006.

Figure 6 shows the pattern of association in and around SNP rs67827860 with the IDP *ProbtrackX_ICVF_ilf_r* (the average value of the intra-cellular volume fraction micro-structure measure, in the left inferior longitudinal fasciculus), and **Figure 7** is a PheWAS plot for rs67827860, which shows the overall pattern of association with all 3,144 IDPs, and how this SNP is broadly associated with many of the dMRI IDPs, as well as the T2 FLAIR white matter lesion volume IDP.

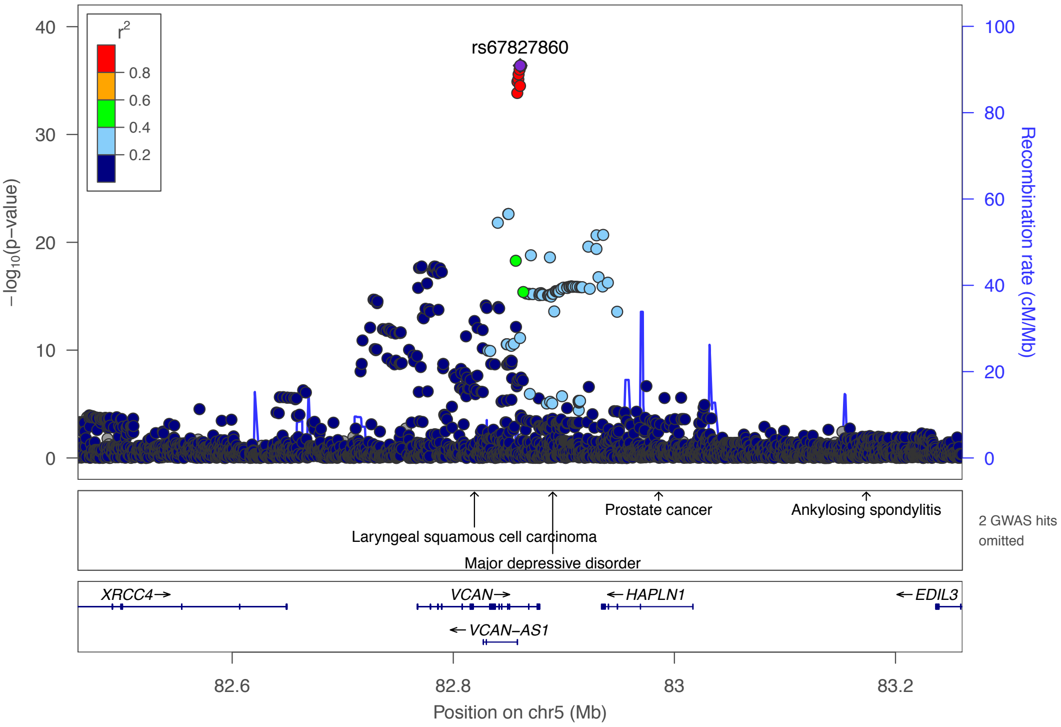

Figure 6 : Genetic association of SNPs with IDP *ProbtrackX_ICVF_ilf_r* (the average value of the intra-cellular volume fraction micro-structure measure, in the left inferior longitudinal fasciculus) centered on rs67827860. In the top panel SNPs are plotted by their positions on the chromosome against association with the IDP ($-\log_{10}$ P value) on the left y axis. Points are coloured by their local linkage disequilibrium (LD) pattern with the focal SNP rs67827860 (purple diamond). Below the main plot are two tracks that show existing GWAS associations, and position and orientation of local genes.

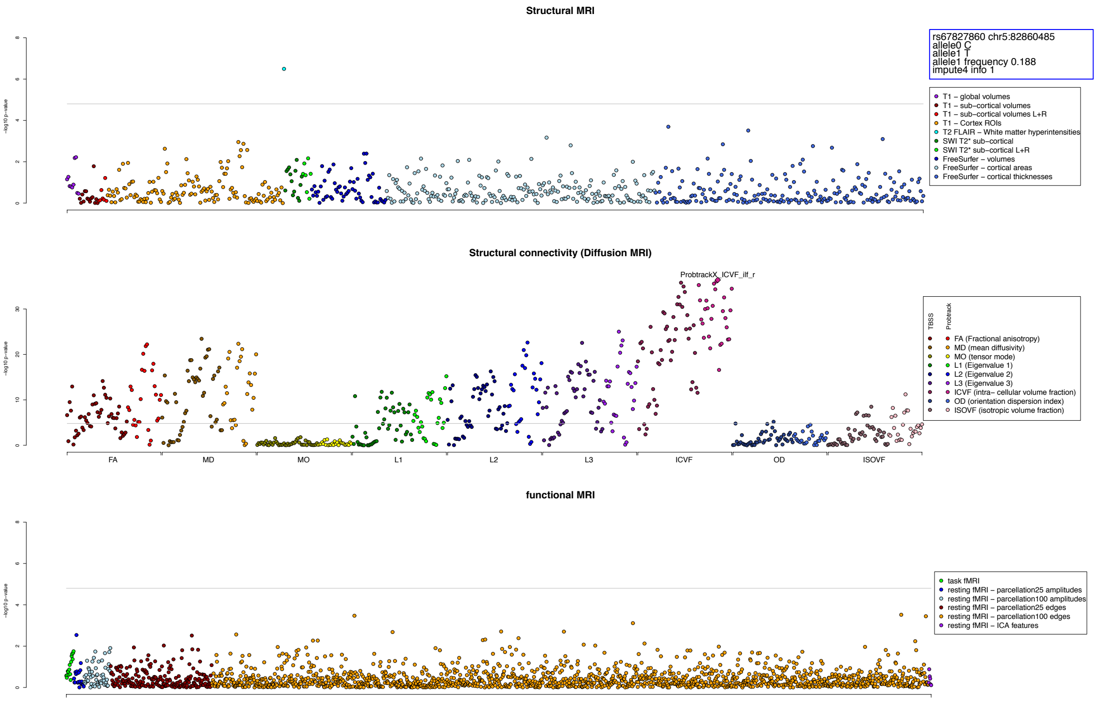

Figure 7 : PheWAS plot for SNP rs67827860. The association (-log$_{10}$ p-value) on the y-axis for the SNP rs67827860 with each of the 3,144 IDPs. The IDPs are arranged on the x-axis in the three panels : (top) Structural MRI IDPs, (middle) Structural connectivity dMRI IDPs, (bottom) functional MRI IDPs. Points are coloured to delineate subgroups of IDPs and detailed in the legends. Summary details of SNP rs67827860 are given in the top right box. The grey line shows the Bonferroni multiple testing threshhold of 4.79.

Two additional examples further illustrate highly meaningful correspondences between locations of our brain IDPs and significantly associated genes. First, the volume of the 4th ventricle, which develops from the central cavity of the neural tube and belongs to the hindbrain, was found to be significantly associated with *ALDH1A2*, which facilitates posterior organ development and prevents human neural tube defects, including spina bifida, a disorder which results from failure of fusion of the caudal neural tube[55] (rs2464469, P=3.15E-16, cluster 33). Second, amongst the three associations we identified for the crossing pontine tract (the part of the pontocerebellar fibres from pontine nuclei that decussate across the brain midline to project to contralateral cerebellar cortex), two were with genes that regulate axon guidance and fasciculation during development (*SEMA3D*, rs2286184, P=5.31E-17, cluster 15 and *ROBO3* (exon), rs4935898, P=1.76E-19, cluster 27). The exact location of our IDP in the crossing fibres of the pons remarkably coincides with the function of *ROBO3*, which is specifically required for axons to cross the midline in the hindbrain (pons, medulla oblongata and cerebellum); mutations in *ROBO3* result in horizontal gaze palsy, a disorder in which the corticospinal and somatosensory axons fail to cross the midline in the medulla[56]. Notably, all three significant associations with the IDP of the crossing pontine tract were found using the dMRI measure of mode of anisotropy (MO), which is a tensor-derived measure particularly sensitive to regions of crossing fibres[57].

These associations are investigated spatially in **Figure 8**. As with the T2* voxelwise results shown above, these spatial maps for the effect of SNPs rs4935898 and rs2286184 on tensor mode are extremely spatially specific, with no extended differences elsewhere in the brain. However, unlike with the 4 distinct maps in the T2* shown above, here these two SNPs had almost identical spatial localizations.

Figure 9 shows the pattern of association in and around SNP rs4935898 with the IDP of *TBSS_MO_Pontine_crossing_tract* (average value of the tensor mode tract measure within the pontine crossing tract), and **Figure 10** is a PheWAS plot for rs4935898 which shows the overall pattern of association with all 3,144 IDPs, and how this SNP is broadly associated with many of the dMRI IDPs, and the T2 FLAIR IDP.

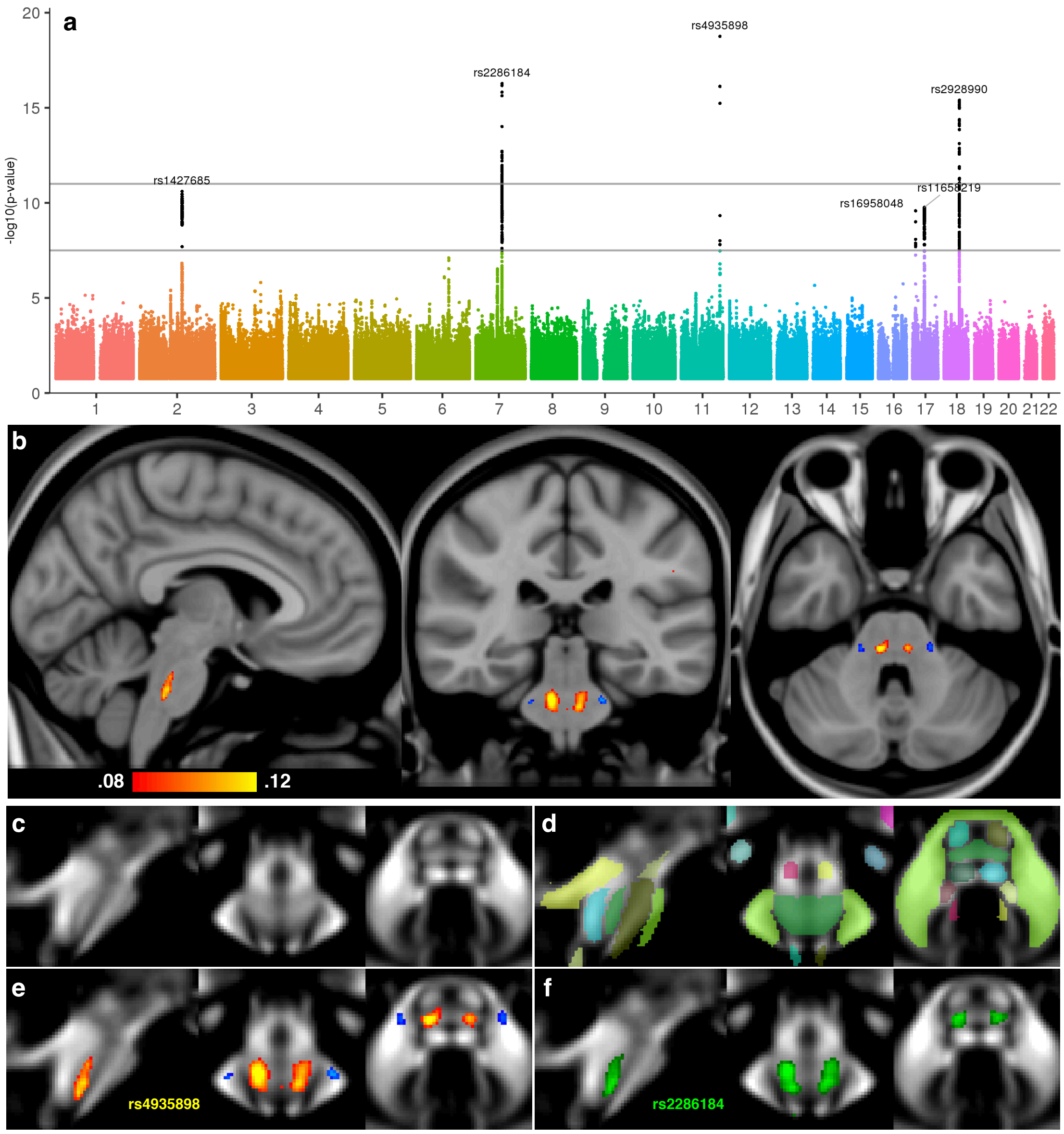

Figure 8: **Manhattan plot and detailed spatial mapping of the association between the *tensor mode* structural connectivity measure and SNPs rs4935898 and rs2286184.** The Manhattan plot relates to the original GWAS for the IDP *TBSS_MO_Pontine_crossing_tract*. All tensor mode data was transformed into standard space and masked with a white-matter tract-centres skeleton mask. For rs4935898**,** the quantitative (though unitless) tensor mode measure was then averaged across all 6,807 subjects with ~0 copies of the non-reference allele, and the average from all 703 subjects having ~1 copy was subtracted from that, for display in red-yellow here, thresholded at 0.08 (**b**,**e**). For rs2286184, the average mode from 4,810 subjects with ~0 copies of the non-reference allele was subtracted from the average from all 2,412 subjects having ~1 copy, for display in green here, thresholded at 0.03 (**f**). No further masking

was applied. In (**b**) the results from mapping rs4935898 are shown overlaid on the MNI152 T1 structural image; it can be seen how little within-white-matter contrast is available in the T1 data. In contrast, images **c-f** show the UK Biobank average FA (fractional anisotropy) image from the dMRI data, with clear tract structure visible within the brainstem. In (**d**) the ICBM-DTI-81 white-matter atlas is overlaid, showing delineation of tracts such as the pontine crossing tract (medium green, spanning the main areas affected by these two SNPs), and the middle cerebellar peduncle (light green, incorporating the small areas shown in blue in **e**, where rs4935898 has the opposite effect on tensor mode compared with its effect in the pontine crossing tract).

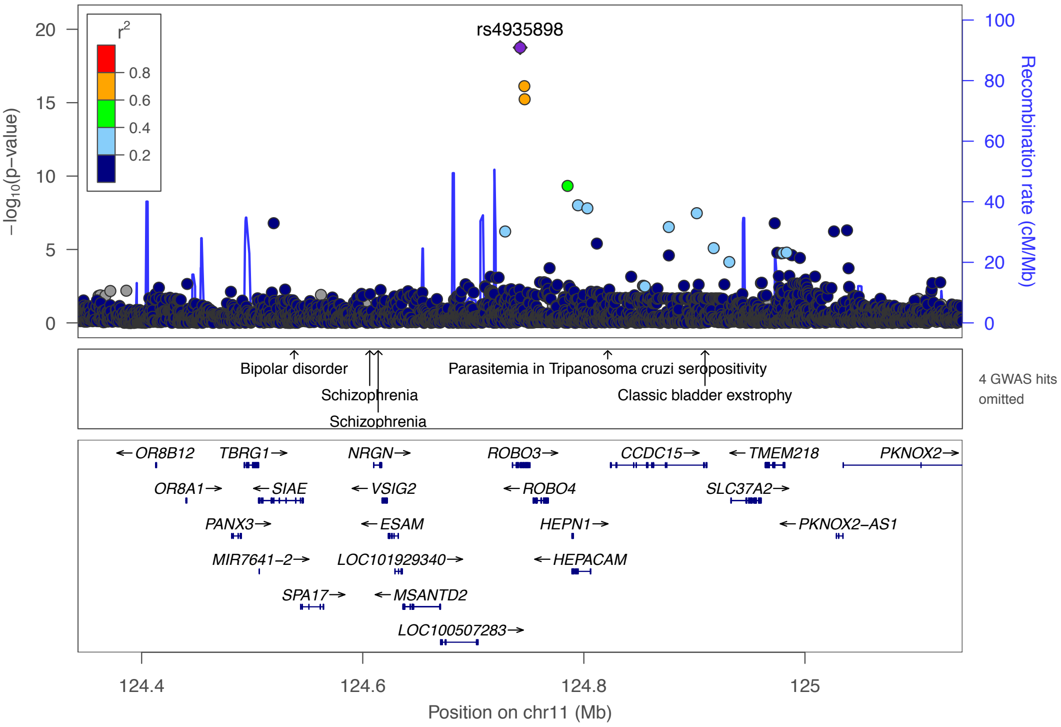

Figure 9 : Genetic association of SNPs with IDP *TBSS_MO_Pontine_crossing_tract* (average value of the tensor mode tract measure within the pontine crossing tract) centered on rs4935898. In the top panel SNPs are plotted by their positions on the chromosome against association with the IDP ($-\log_{10}$ P value) on the left y axis. Points are coloured by their local linkage disequilibrium (LD) pattern with the focal SNP rs4935898 (purple diamond). Below the main plot are two tracks that show existing GWAS associations, and position and orientation of local genes.

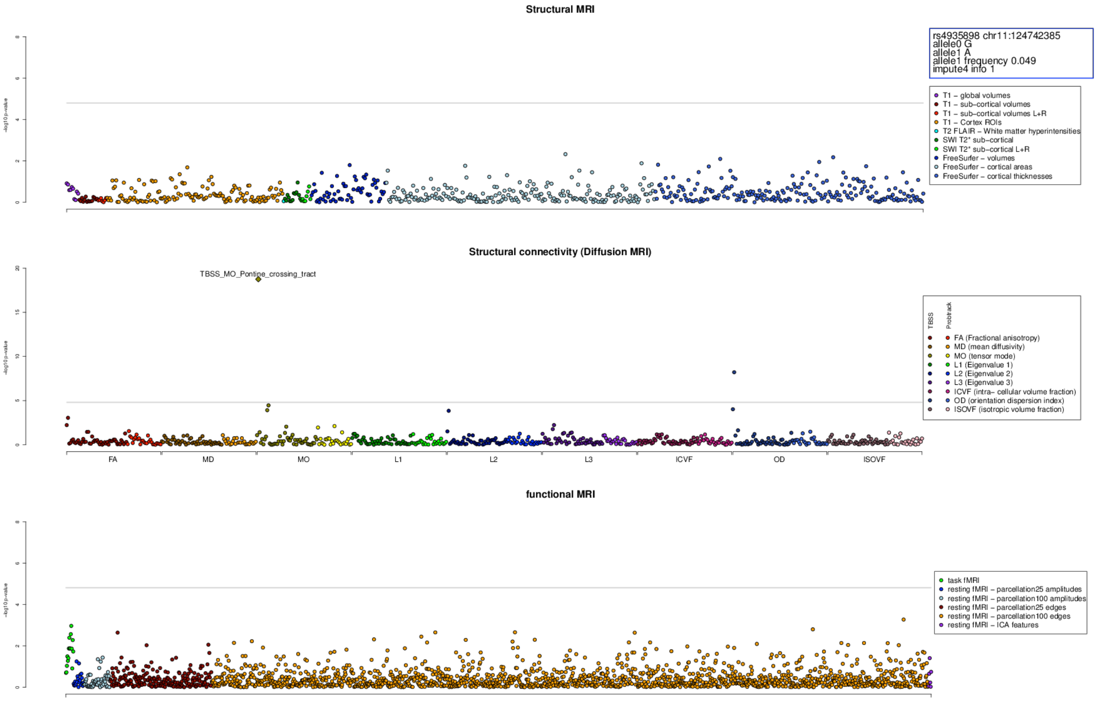

Figure 10 : PheWAS plot for SNP rs4935898. The association (-log$_{10}$ p-value) on the y-axis for the SNP rs4935898 with each of the 3,144 IDPs. The IDPs are arranged on the x-axis in the three panels : (top) Structural MRI IDPs, (middle) Structural connectivity dMRI IDPs, (bottom) functional MRI IDPs. Points are coloured to delineate subrgroups of IDPs and detailed in the legends. Summary details of SNP rs4935898 are given in the top right box. The grey line shows the Bonferroni multiple testing threshhold of 4.79.

14 genes identified here contribute broadly to brain development, patterning and plasticity. Beside *SEMA3D* and *ROBO3*, *BCAN* and *VCAN* have also been involved, as chondroitin sulfate proteoglycans, in axon guidance and signalling pathways in neurons[58], with VCAN co-localising with SEMA3A, a guidance cue[59]. Similarly, *EPHA3* was associated in our GWAS with cluster 6, which included many rfMRI functional connections between the middle temporal sulcus and mainly prefrontal and parietal brain areas (rs66499884, P$_{min}$=2.77E-23). *EPHA3* mediates the regulation of cell migration and axon guidance[60], and regulates trans-axonal signalling[61]. We have discussed above the role of *EFEMP1* in the ECM. The other relevant genes are: *WDR75*, which reduces the expression of

homeobox *NANOG*[62] and was associated with T2* in the pallidum (rs6740926, P_{min}=1.31E-14, cluster 5); *ZIC4* (exon) which can lead to cerebellar malformations and was found associated with multiple rfMRI connections mainly between prefrontal, cerebellar and parietal areas (rs2279829, P=8.34E-12, cluster 8); *ZIP8* (see above) which plays a role in brain development via release from choroid plexus; *NR2F1-AS1* (*COUP-TF1*), a master regulator which interacts with *PAX6* (cluster 12); *HBEGF* which stimulates neurogenesis in proliferative zones of the adult brain (see above); *WNT16* (cluster 17); *ALDH1A2* (cluster 33, see above) and *COASY* (cluster 39, see above).

Conclusions

Bringing together researchers with backgrounds in both brain imaging analysis and genetic association was key to this work. We have uncovered a large number of associations at the nominal level of GWAS significance (-log10 p-value > 7.5) and at a more stringent threshold (-log10 p-value > 11) designed to (probably over-conservatively) control for the number of IDPs tested. We find associations with all the main IDP groups (**Table 4**) except the task fMRI measures. A valuable aspect of this work has been to link the associated SNPs back to spatial properties of the voxel-level brain imaging data. For example, we have linked SNPs associated with IDPs to both highly localized (**Figures 2** and **8**) and distributed spatial properties (**Figure 5**). In addition, looking at PheWAS plots has been useful when working with so many phenotypes. It has allowed investigation of the overall patterns of association and has led to the identification of SNP associations that span multiple modalities.

Using the modest replication set of 930 samples, we were only able here to replicate a sizeable, but not complete subset of these associations. It will shortly be possible to increase the size of the replication sample (as UK Biobank is regularly releasing additional imaging data), and substantially increase the number of replicated associations. Combining the discovery and replication samples will likely also lead to novel associations, as will use of methods that can analyze multiple IDPs together, both from the raw genetic data, and from the IDP by SNP matrix of summary statistics of association. Over the next few years the number of UK Biobank participants with imaging data will gradually increase to 100,000, which will allow a much more complete discovery of the genetic basis of human brain structure, function and connectivity. A potential avenue of research will involve attempting to uncover causal pathways that link genetic variants to IDPs and then onto a range of neurological, psychiatric and developmental disorders.

Acknowledgements

The data used in this work was obtained from UK Biobank under Data Access Application 8107. We are grateful to UK Biobank for making the resource data available, and are extremely grateful to all UK Biobank study participants, who generously donated their time to make this resource possible. F.A-A acknowledges funding from the UK Medical Research Council and the Wellcome Trust via UK Biobank. K.L.M. and S.M.S. receive further support from the Wellcome Trust. JM acknowledges funding for this work from the European Research Council (ERC; grant 617306). We are grateful to Bruce Fischl, Doug Greve and Matt Glasser for advice on FreeSurfer processing.

Author contributions

J.M and S.S conceived and supervised the work. F.A-A, K.M, S.S created the new IDPs and confound covariates. L.E, K.S and J.M carried out the genetic association analysis. J.M, S.S, G.D, K.M and L.E wrote the paper.

Conflicts of interest

J.M is a co-founder and director of Gensci Ltd. S.S is a co-founder of SBGneuro.

URLs

SBAT https://jmarchini.org/sbat/

fastLMM https://github.com/MicrosoftGenomics/FaST-LMM

PLINK http://www.cog-genomics.org/plink/2.0/

BGENIE https://jmarchini.org/bgenie/

UK Biobank showcase variables used for head positioning confounds:

http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25756

http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25757

http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25758

http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25759

Head bone density and mineral content measures

https://biobank.ctsu.ox.ac.uk/crystal/docs/DXA_explan_doc.pdf

References

1. Karas, G. B. *et al.* A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *Neuroimage* **18,** 895–907 (2003).
2. Douaud, G., Filippini, N., Knight, S., Talbot, K. & Turner, M. R. Integration of structural and functional magnetic resonance imaging in amyotrophic lateral sclerosis. *Brain* **134,** 3470–3479 (2011).
3. van Erp, T. G. M. *et al.* Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* **21,** 547–553 (2016).
4. Schmaal, L. *et al.* Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol. Psychiatry* **21,** 806–812 (2016).
5. Ecker, C., Bookheimer, S. Y. & Murphy, D. G. M. Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan. *The Lancet Neurology* **14,** 1121–1134 (2015).
6. Strakowski, S. M. *et al.* Brain Magnetic Resonance Imaging of Structural Abnormalities in Bipolar Disorder. *Arch Gen Psychiatry* **56,** 254–260 (1999).
7. Ersche, K. D. *et al.* Abnormal Brain Structure Implicated in Stimulant Drug Addiction. *Science* **335,** 601–604 (2012).
8. Fornito, A. & Bullmore, E. T. Connectomic Intermediate Phenotypes for Psychiatric Disorders. *Front. Psychiatry* **3,** (2012).
9. Sudlow, C., Gallacher, J., Allen, N., Beral, V. & Burton, P. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS …* (2015).
10. Allen, N. *et al.* UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology* **1,** 123–126 (2012).
11. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank

prospective epidemiological study. *Nat. Neurosci.* **19,** 1523–1536 (2016).

12. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62,** 782–790 (2012).
13. Alfaro-Almagro, F. *et al.* Image Processing and Quality Control for the first 10,000 Brain Imaging Datasets from UK Biobank. *bioRxiv* 130385 (2017). doi:10.1101/130385
14. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 1–36 (2017). doi:10.1101/166298
15. Hibar, D. P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520,** 224–229 (2015).
16. Hibar, D. P. *et al.* Novel genetic loci associated with hippocampal volume. *Nature Communications* **8,** 13624 (2017).
17. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *Neuroimage* **80,** 62–79 (2013).
18. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9,** 179–194 (1999).
19. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9,** 195–207 (1999).
20. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical labeling protocol. *Front Neurosci* **6,** 171 (2012).
21. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53,** 1–15 (2010).
22. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw* **10,** 626–634 (1999).
23. Duff, E. P. *et al.* Learning to identify CNS drug action and efficacy using multistudy fMRI data. *Sci Transl Med* **7,** 274ra16–274ra16 (2015).
24. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv* 035170 (2015). doi:10.1101/035170
25. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications* **6,** 8111 (2015).
26. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
27. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36,** 512–517 (2004).
28. Cai, J.-F., Candès, E. J. & Shen, Z. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization* **20,** 1956–1982 (2010).
29. Smith, S. M. *et al.* Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* **17,** 479–489 (2002).
30. Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* 1–9 (2016). doi:10.1038/ng.3513
31. Luciano, M. *et al.* 116 independent genetic variants influence the neuroticism personality trait in over 329,000 UK Biobank individuals. 1–32 (2017). doi:10.1101/168906
32. Davies, G. *et al.* Ninety-nine independent genetic loci influencing general

cognitive function include genes associated with brain health and structure (N = 280,360). *bioRxiv* 1–35 (2017). doi:10.1101/176511

33. Mizoguchi, T. *et al.* Behavioral abnormalities with disruption of brain structure in mice overexpressing VGF. *Sci Rep* **7,** 593 (2017).
34. Hass, J. *et al.* A Genome-Wide Association Study Suggests Novel Loci Associated with a Schizophrenia-Related Brain-Based Phenotype. *PLoS ONE* **8,** e64872 (2013).
35. Tran, T. *et al.* Candidate-gene analysis of white matter hyperintensities on neuroimaging. *J Neurol Neurosurg Psychiatry* **87,** jnnp–2014–309685–266 (2015).
36. Leitner, D. F. & Connor, J. R. Functional roles of transferrin in the brain. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1820,** 393–402 (2012).
37. Ali-Rahmani, F., Schengrund, C.-L. & Connor, J. R. HFE gene variants, iron, and lipids: a novel connection in Alzheimer's disease. *Frontiers in Pharmacology* **5,** 953 (2014).
38. Richardson, D. R. *et al.* Mitochondrial iron trafficking and the integration of iron metabolism between the mitochondrion and cytosol. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 10775–10782 (2010).
39. Thakurela, S. *et al.* The transcriptome of mouse central nervous system myelin. *Sci Rep* **6,** 971 (2016).
40. Hogarth, P. Neurodegeneration with Brain Iron Accumulation: Diagnosis and Management. *Journal of Movement Disorders* **8,** 1–13 (2015).
41. Pelis, R. M. & Wright, S. H. SLC22, SLC44, and SLC47 transporters–organic anion and cation transporters: molecular and cellular properties. *Curr Top Membr* (2014).
42. Saunders, N. R. *et al.* Influx mechanisms in the embryonic and adult rat choroid plexus: a transcriptome study. *Front Neurosci* **9,** 123 (2015).
43. Wang, C. *et al.* Mutations in SLC20A2 link familial idiopathic basal ganglia calcification with phosphate homeostasis. *Nat. Genet.* **44,** 254–256 (2012).
44. Scarr, E. *et al.* Increased cortical expression of the zinc transporter SLC39A12 suggests a breakdown in zinc cellular homeostasis as part of the pathophysiology of schizophrenia. *npj Schizophrenia 2016 2:null* **2,** npjschz20162 (2016).
45. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science* **4,** 274–290 (2009).
46. Dityatev, A., Schachner, M. & Sonderegger, P. The dual role of the extracellular matrix in synaptic plasticity and homeostasis. *Nature Reviews Neuroscience* **11,** 735–746 (2010).
47. Lau, L. W., Cua, R., Keough, M. B., Haylock-Jacobs, S. & Yong, V. W. Pathophysiology of the brain extracellular matrix: a new target for remyelination. *Nature Reviews Neuroscience* **14,** 722–729 (2013).
48. Sobel, R. A. & Ahmed, A. S. White matter extracellular matrix chondroitin sulfate/dermatan sulfate proteoglycans in multiple sclerosis. *Journal of Neuropathology & …* (2001).
49. Shih, C.-H., Lacagnina, M., Leuer-Bisciotti, K. & Pröschel, C. Astroglial-Derived Periostin Promotes Axonal Regeneration after Spinal Cord Injury. *J. Neurosci.* **34,** 2438–2443 (2014).

50. Matarin, M. *et al.* A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. *The Lancet Neurology* **6,** 414–420 (2007).
51. Clark, J. A., Yeaman, E. J., Blizzard, C. A., Chuckowree, J. A. & Dickson, T. C. A Case for Microtubule Vulnerability in Amyotrophic Lateral Sclerosis: Altered Dynamics During Disease. *Frontiers in Cellular Neuroscience* **10,** 2910 (2016).
52. Lewis, C. M. *et al.* Genome-Wide Association Study of Major Recurrent Depression in the U.K. Population. *American Journal of Psychiatry* **167,** 949–957 (2010).
53. Scafidi, J. *et al.* Intranasal epidermal growth factor treatment rescues neonatal brain injury. *Nature* **506,** 230–234 (2013).
54. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* **61,** 1000–1016 (2012).
55. Deak, K. L. *et al.* Analysis of ALDH1A2, CYP26A1, CYP26B1, CRABP1, and CRABP2 in human neural tube defects suggests a possible association with alleles in ALDH1A2. *Birth Defects Research Part A: Clinical and Molecular Teratology* **73,** 868–875 (2005).
56. Jen, J. C. *et al.* Mutations in a Human ROBO Gene Disrupt Hindbrain Axon Pathway Crossing and Morphogenesis. *Science* **304,** 1509–1513 (2004).
57. Douaud, G. *et al.* DTI measures in crossing-fibre areas: Increased diffusion anisotropy reveals early white matter alteration in MCI and mild Alzheimer's disease. *Neuroimage* **55,** 880–890 (2011).
58. Ohtake, Y., Wong, D., Abdul-Muneer, P. M., Selzer, M. E. & Li, S. Two PTP receptors mediate CSPG inhibition by convergent and divergent signaling pathways in neurons. *Sci Rep* **6,** srep37152 (2016).
59. Vo, T. *et al.* The chemorepulsive axon guidance protein semaphorin3A is a constituent of perineuronal nets in the adult rodent brain. *Molecular and Cellular Neuroscience* **56,** 186–200 (2013).
60. Shi, G., Yue, G. & Zhou, R. EphA3 Functions are Regulated by Collaborating Phosphotyrosine Residues. *Cell research* **20,** 1263–1275 (2010).
61. Gallarda, B. W. *et al.* Segregation of Axial Motor and Sensory Pathways via Heterotypic Trans-Axonal Signaling. *Science* **320,** 233–236 (2008).
62. You, K. T., Park, J. & Kim, V. N. Role of the small subunit processome in the maintenance of pluripotent stem cells. *Genes Dev.* **29,** 2004–2009 (2015).