Machine learning shows association between genetic variability in *PPARG* and cerebral connectivity in preterm infants

Michelle L. Krishnan[a], Zi Wang[b], Paul Aljabar[a], Gareth Ball[a], Ghazala Mirza[c], Alka Saxena[c], Serena J. Counsell[a], Joseph V. Hajnal[a,b], Giovanni Montana[b,1], and A. David Edwards[a,1,2]

[a]Centre for the Developing Brain, King's College London, St. Thomas' Hospital, London SE1 7EH, United Kingdom; [b]Department of Biomedical Engineering, King's College London, St. Thomas' Hospital, London SE1 7EH, United Kingdom; and [c]Genomics Core Facility, National Institute for Health Research Biomedical Research Centre, Guy's and St. Thomas' National Health Service Foundation Trust, London SE1 9RT, United Kingdom

Edited by Marcus E. Raichle, Washington University in St. Louis, St. Louis, MO, and approved November 8, 2017 (received for review March 27, 2017)

Preterm infants show abnormal structural and functional brain development, and have a high risk of long-term neurocognitive problems. The molecular and cellular mechanisms involved are poorly understood, but novel methods now make it possible to address them by examining the relationship between common genetic variability and brain endophenotype. We addressed the hypothesis that variability in the Peroxisome Proliferator Activated Receptor (PPAR) pathway would be related to brain development. We employed machine learning in an unsupervised, unbiased, combined analysis of whole-brain diffusion tractography together with genomewide, single-nucleotide polymorphism (SNP)-based genotypes from a cohort of 272 preterm infants, using Sparse Reduced Rank Regression (sRRR) and correcting for ethnicity and age at birth and imaging. Empirical selection frequencies for SNPs associated with cerebral connectivity ranged from 0.663 to zero, with multiple highly selected SNPs mapping to genes for *PPARG* (six SNPs), *ITGA6* (four SNPs), and *FXR1* (two SNPs). SNPs in *PPARG* were significantly overrepresented (ranked 7–11 and 67 of 556,000 SNPs; $P < 2.2 \times 10^{-7}$), and were mostly in introns or regulatory regions with predicted effects including protein coding and nonsense-mediated decay. Edge-centric graph-theoretic analysis showed that highly selected white-matter tracts were consistent across the group and important for information transfer ($P < 2.2 \times 10^{-17}$); they most often connected to the insula ($P < 6 \times 10^{-17}$). These results suggest that the inhibited brain development seen in humans exposed to the stress of a premature extrauterine environment is modulated by genetic factors, and that PPARG signaling has a previously unrecognized role in cerebral development.

brain development | preterm | magnetic resonance imaging | machine learning | PPARG

Preterm birth accounts for 11% of all births (1), and is the leading global cause of death and disability under 5 y of age (2). Over 30% of survivors experience neurocognitive problems from early life (3) lasting into adulthood (4), including anxiety, inattention, and social and communication problems (5), and socioemotional problems (6). Psychiatric disorders are present in around 25% of preterm adolescents, constituting a 3–4-fold increased risk compared with term-born peers (review in ref. 7), including a risk ratio of 7.4 for bipolar affective disorder and 2.5 for nonaffective psychosis (8), and a threefold increase in the prevalence of autism-spectrum disorders (ASD) (9).

Imaging studies have shown that adverse functional outcomes are associated with changes in brain structure, connectivity, and function (10, 11), but while this phenotype has been extensively investigated in recent years, few studies have addressed the cellular or molecular mechanisms involved. Recent advances in machine learning and imaging genomics now make it possible to investigate potential mechanisms by studying the genetic variability associated with the cerebral endophenotypes.

Previously, our preliminary candidate gene and genomewide pathway-based studies have suggested an association between white-matter development and a number of metabolic pathways, with the strongest link to the Peroxisome Proliferator Activated Receptor (PPAR) pathway (12, 13), raising the hypothesis that the PPAR pathway modulates brain development in preterm infants. To test this, and to explore genetic influences on preterm brain development further, we collected a large cohort of linked diffusion MRI (d-MRI) and genomic data, and undertook an unsupervised, unbiased machine-learning analysis of whole-brain diffusion tractography together with genomewide, single-nucleotide polymorphism (SNP)-based genotypes.

Results

Participants. A cohort of 272 infants born at less than 33 wk gestational age (GA) (mean 29 wk + 4 d) had suitable imaging at term-equivalent age [mean age at scan (SA) 42 wk + 4 d] and allied genomic DNA available (*SI Appendix*, *Supplementary Methods*).

Population Stratification. Relatedness between individuals in the cohort was assessed by calculating pairwise identity by state (IBS) values and using this distance matrix to perform principal component analysis. This revealed a degree of stratification along the first two components, corresponding to parental self-reported ethnicity (*SI Appendix*, Fig. S1). The first principal

Significance

Preterm birth affects 11% of births globally; 35% of infants develop long-term neurocognitive problems, and prematurity leads to the loss of 75 million disability adjusted life years per annum worldwide. Imaging studies have shown that these infants have extensive alterations in brain development, but little is known about the molecular or cellular mechanisms involved. This imaging genetics study found a strong association between abnormal cerebral connectivity and variability in the *PPARG* gene, implicating PPARG signaling in abnormal white-matter development in preterm infants and suggesting a tractable new target for therapeutic research.

Author contributions: M.L.K., Z.W., P.A., G.B., S.J.C., J.V.H., G. Montana, and A.D.E. designed research; M.L.K., Z.W., P.A., G.B., G. Mirza, and A.S. performed research; Z.W. and G. Montana contributed new reagents/analytic tools; M.L.K., Z.W., P.A., and G.B. analyzed data; and M.L.K., P.A., G.B., S.J.C., G. Montana, and A.D.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]G. Montana and A.D.E. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: ad.edwards@kcl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1704907114/-/DCSupplemental.

component of the IBS matrix was used as a covariate for adjustment of the phenotype.

Sparse Reduced Rank Regression Selects a Consistent Cerebral Endophenotype. White-matter tracts, defined as the edges in the tractography connectivity matrix, were ranked according to their selection probabilities in the Sparse Reduced Rank Regression (sRRR) model. Selection frequencies ranged from 0.817 to zero; Fig. 1 shows the 100 most frequently selected edges. Two separate approaches were employed to eliminate individual relationships between imaging and genetic data and achieve a set of null results: (*i*) permutation of individuals within the imaging dataset; and (*ii*) replacing the original phenotype matrix with a matrix of randomly generated values with standard normal distribution and the same dimensions. Fig. 1 shows the empirical distribution together with these two null sets. The random replacement, but not the permutation of individuals, produced a null distribution with uniformly low selection frequency, which demonstrates a level of similarity between individuals and a consistent endophenotype.

Consistently Selected White-Matter Tracts Have a Significant Role in Information Flow. To understand the cerebral endophenotype further, the 10 tractography connections selected most often in the sRRR model were visualized according to University of North Carolina Automated Anatomical Labeling atlas coordinates using BrainNet Viewer software (14), and the group median connectivity matrix was examined from an edge-centric perspective, applying graph theoretical measures to detect link communities (15). These tracts were distributed within middle-frontal, middle-temporal, and parahippocampal/entorhinal link communities (15), and integrated fractional anisotropy ranged from 0.06 to 0.2 (Fig. 1, *Inset* and *SI Appendix*, Figs. S3 and S4). Their impact on information flow in the network was then studied using an "edge lesioning" strategy (15), in which the effect on global communicability of removing them was compared with the effect of removing other sets of 10 edges at random. Global communicability captures information flow in the network and accounts for both shortest paths and all other paths connecting two nodes (16). Removal of the top 10 tracts decreased global communicability from an unlesioned baseline of 1–0.858, significantly more than in 1,000 random lesions (*SI Appendix*, Fig. S5, $P < 2.2 \times 10^{-17}$, Wilcoxon rank sum test), supporting the importance of highly selected white-matter tracts in the cerebral network architecture. The cortical regions linked by the 100 most highly selected tracts were extracted and counted (Fig. 2), with the insular cortex occurring significantly more frequently than predicted ($P < 6 \times 10^{-17}$, Fisher's Exact Test).

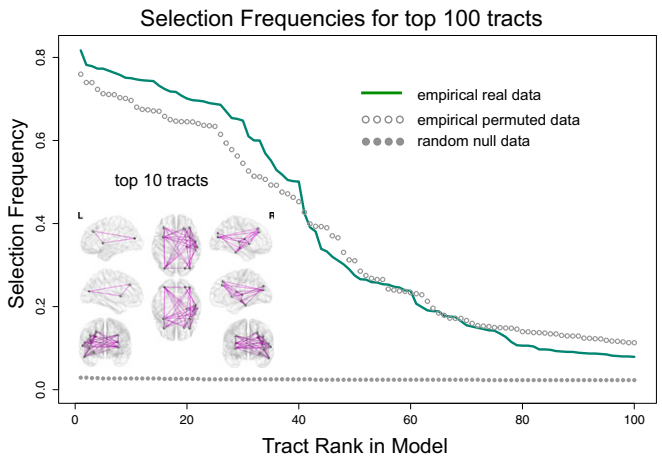

Fig. 1. Selection frequencies of imaging variables with sRRR. Solid green line: Empirical selection frequencies of imaging variables (edges from the probabilistic tractography connectivity matrix) as ranked by the sRRR model with 1,000 subsamples of size 2/3 total number of subjects and convergence criterion = 1×10^{-6}. Empty gray circles: A permuted distribution was computed with the same parameters by permuting the order of individual imaging datasets between each subsample of the data. Solid gray circles: A null distribution was computed using a randomly generated matrix of standard normally distributed values with the same dimensions as the empirical data, using 20,000 subsamples of 2/3 of samples. *Inset* (larger in *SI Appendix*): Surface anatomical location of the 10 edges most highly ranked by the sRRR method, using cortical atlas coordinates from the UNC AAL neonatal atlas (71), rendered with the BrainNet viewer (14). Surface views, cortical regions shown as gray circles and brain surface semitransparent. First row from left to right: lateral view of left hemisphere, view from above, lateral view of right hemisphere. Second row from left to right: medial view of left hemisphere, inferior aspect, medial view of right hemisphere. Third row: anterior aspect and posterior aspect. Top 10 ranked tracts (directionality not implied), subscripts "_L" and "_R" indicate laterality: (*RegionA* * *RegionB*): Insula_R*Temporal_Inferior_R; Insula_R*Occipital_Middle_R; Insula_L*Occipital_Middle_L; Frontal_Superior_R*Temporal_Superior_R; Frontal_Middle_R*Insula_R; Occipital_Superior_R*Temporal_Inferior_R; Hippocampus_R*Occipital_Middle_R; Rolandic_Operculum_R*Insula_R; Frontal_Superior_R*Insula_R; Frontal_Middle_L*Insula_L.

Genetic Variation Is Associated with the Preterm Cerebral Endophenotype. Genetic associations with image features were identified using the sRRR method as previously reported (17, 18). Genomewide SNPs were ranked according to their selection probabilities in the sRRR model, and the empirical and null distributions were inspected. This revealed empirical selection frequencies between 0.663 and zero, with a steeper rate of decrease after the top 100 ranked SNPs, and uniformly low null distribution (Fig. 3). The top 100 SNPs were thus examined in more detail as a stringent subset of genetic variables most highly and stably associated with the tractography features (*SI Appendix*, Table S4). These top 100 SNPs mapped to 47 genes (*SI Appendix*, Table S1), mostly in linkage equilibrium with each other apart from three separate hotspots of linkage disequilibrium centered on the genes *PPARG* (six SNPs), Integrin Subunit Alpha 6 (*ITGA6*) (four SNPs), and Fragile X Mental Retardation, Autosomal Homolog 1 (*FXR1*) (two SNPs).

SNPs in the PPARG Gene Are Most Highly Associated with Variability in Imaging Features. SNPs in *PPARG* were significantly overrepresented among the top 100 SNPs ($P < 2.2 \times 10^{-7}$, Fisher's Exact Test), ranked by the sRRR model according to strength of association in positions 7–11 (rs17036282, rs6801982, rs4135334, rs4135336, rs4135342) and position 67 (rs6442313) of 556,000, with uniform selection frequencies of 0.663. The *PPARG* SNPs were mostly in intronic or regulatory regions (promoter flanking regions and open chromatin regions) (*SI Appendix*, Fig. S6), with predicted effects on processes including protein coding, retained introns, and nonsense-mediated decay (*SI Appendix*, Table S2).

Highly Associated Genes Are Involved in Biological Processes Including Lipid Metabolism. To explore the 47 top-ranked genes the Gene Ontology (GO) framework was surveyed using the Cytoscape tool ClueGO (19) to create a functionally organized GO term network (*SI Appendix*, Fig. S7). This revealed several significantly overrepresented themes of interest (hypergeometric test, adjusted $P < 0.05$ with Benjamini–Hochberg correction) including lipid metabolism (*PPARG*), neuron projection regeneration (*ADM*, *PRRX1*), response to nerve growth factor stimulation (*BPTF*, *EP300*), acetylcholine biosynthesis (*CHAT*), and presynaptic membrane assembly (*PTPRD*) (full annotation in *SI Appendix*, Table S3).

Given the significant overrepresentation of *PPARG* SNPs among the most highly ranked SNPs and our previous finding of association between lipid metabolism genes and white-matter integrity in preterms (12, 13), the top 100 SNPs were tested for significant overlap with a list of SNPs mapping to genes classified

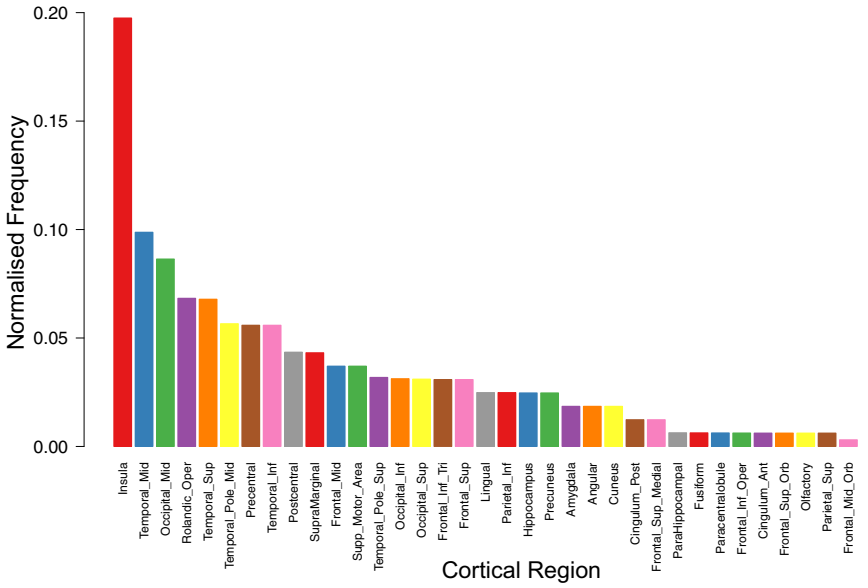

Fig. 2. Cortical regions in top 100 tracts by ranked sRRR. Frequency of participation of cortical regions among the top 100 ranked tracts, normalized by total frequency in the network. Insula occurrence is significantly higher than expected ($P < 6 \times 10^{-17}$, Fisher's Exact Test). Regions (x axis, left to right): Insula; Temporal Middle; Occipital Middle; Rolandic Operculum; Temporal Superior; Temporal Pole Middle; Precentral; Temporal Inferior; Postcentral; Supramarginal; Frontal Middle; Superior Motor Area; Temporal Pole Superior; Occipital Inferior; Occipital Superior; Frontal Inferior Trigone; Frontal Superior; Lingual; Parietal Inferior; Hippocampus; Precuneus; Amygdala; Angular; Cuneus; Cingulum Posterior; Frontal Superior Medial; ParaHippocampal; Fusiform; Paracentral Lobule; Frontal Inferior Operculum; Cingulum Anterior; Frontal Superior Orbital; Olfactory; Parietal Superior; Frontal Middle Orbital.

with the GO term "lipid metabolism" (GO: 0006629), using the R package SuperExactTest (20). Four genes (*PPARG*, *ADM*, *CHAT*, and *PNPL6*) involved in lipid metabolism according to the GO classification were present among the top 100 SNPs ranked by sRRR, more than would be expected by chance ($P < 0.005$) (Fig. 4). As a null frame of reference, this result was compared with the overlap between the bottom-ranked 100 SNPs and the GO lipid list, which was not significant.

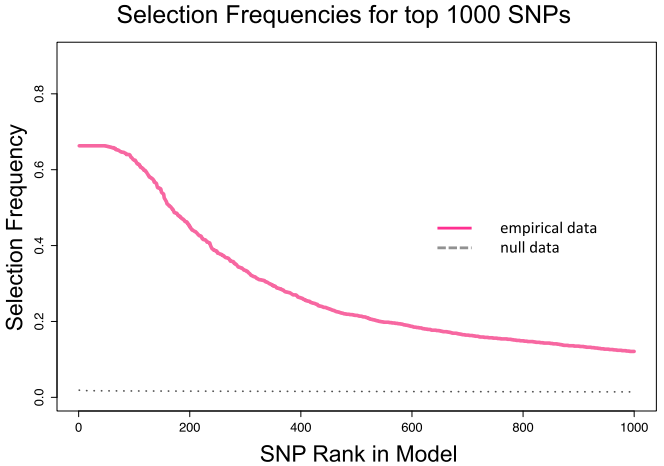

Fig. 3. Selection frequencies of top 1,000 SNPs ranked by sRRR. Selection frequencies of top 1,000 SNPs ranked by the sRRR method over 1,000 subsamples of 2/3 of the individuals and convergence criterion = 1×10^{-6}. Empirical results (solid line) show that there is a plateau of the highest selection frequencies (maximum 0.663), which is stable for a subset of 100 SNPs. The first 10 of 100 equally highly ranked SNPs mapped to six genes (in alphabetical order): AGAP1, POGZ, PPARG, TSEN2, UBE2E1 (full list of mapped genes in *SI Appendix*, Table S1 and a full list of SNPs in *SI Appendix*, Table S4). The null distribution obtained through permutation of the individuals with 20,000 subsamples is very low and uniform (dotted line).

Highly Associated Genes Are Associated with Neuropsychiatric Diseases. Given the uniform selection frequency of the top 100 ranked SNPs, we examined evidence in literature for all their mapped genes. A machine-learning-based text-mining strategy was employed, using the Agilent Literature Search tool in Cytoscape (21) to query text-based search engines and extract associations among the genes, visualizing them as a network with the sentences for each association forming the network edges (*SI Appendix*, Fig. S8). Mentions of at least 2 of the 47 genes of interest were found in 405 Pubmed-indexed abstracts, which were queried for the occurrence of disease terms using the tool pubmed.mineR (22). The most frequently occurring disease terms related to cancer, reflecting a known ascertainment bias in the literature (23). Once cancer-related terms were removed, the most frequently occurring disease terms were "autism spectrum disorder," "intellectual disability," and "schizophrenia" (*SI Appendix*, Fig. S9), neuropsychiatric conditions more common in the preterm population.

Discussion

The molecular and cellular events leading to abnormal brain development in preterm infants are poorly understood, but hypoxia, ischemia, and inflammation are all believed to play a role (11) and the host response to these external insults is modulated by the combined effects of multiple genes (24, 25). The present results are consistent with the hypothesis that changes in white-matter structure that predict adverse outcome are modulated by genetic variability in the PPAR signaling pathway.

The genetic imaging approach relies on heritability and an appropriate endophenotype. Common DNA sequence variation is estimated to account for up to 50% of additive genetic variation in complex traits, including neuroanatomical features (26) and neurocognitive disorders including ASD (27) and schizophrenia (28). Imaging cerebral endophenotypes generally have high heritability and relevance (29, 30): in the neonatal period 60% of the variability between individuals in d-MRI features can be attributed to genetic factors (31, 32), and d-MRI measures of white-matter structure predict neurodevelopmental outcome (33–35). Analysis of the cerebral connections selected by the

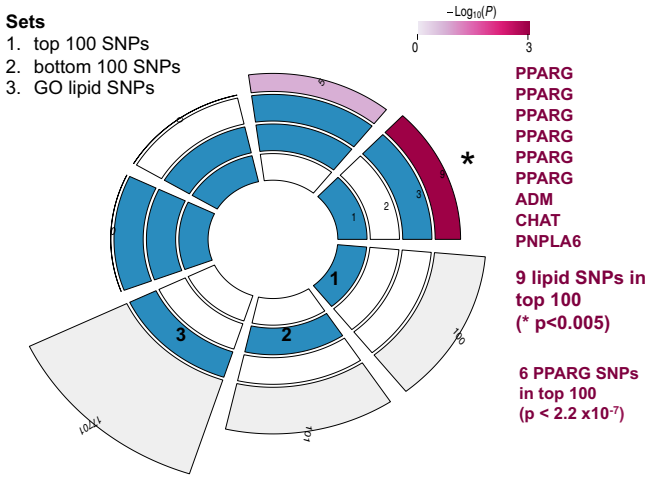

Fig. 4. Representation of PPARG SNPs in the lipid metabolism gene ontology category. Group 1: Top 100 ranked SNPs associated with imaging features by sRRR. Group 2: Bottom 100 SNPs ranked by sRRR. Group 3: All SNPs in the lipid metabolism GO category present on genotyping array. Circular plot illustrating all possible intersections between these three groups and the corresponding statistics. The three tracks in the middle represent the three SNP lists, with individual colored blocks showing "presence" (dark) or "absence" (light) of the SNP groups in each intersection test. The height of the bars in the outer layer is proportional to the log of intersection sizes, indicated by the numbers on the top of the bars. The color intensity of the bars represents the *P*-value significance of the intersections (background = 19,000 protein-coding human genes). The number of SNPs contributing to each intersection is listed above the segment. There is a significant representation of the top 100 SNPs ranked by sRRR (group 1) among the SNPs in the GO lipid metabolism category (group 3) ($P < 0.05$). The SNPs present in both groups 1 and 3 map to four genes as shown (PPARG, ADM, CHAT, PNPLA6).

algorithm showed that they were stable within the group, and virtual lesioning showed that they are important for information transfer in the network. They thus represent structures that are highly relevant to long-term neurological function.

Machine learning using penalized regression provided an unsupervised, unbiased method to address the hypothesis. sRRR is specifically designed to deal with cohorts where the number of individuals is smaller than the number of features, and outperforms mass-univariate linear models when considering genetic effect sizes comparable to those expected here (17). The approach involves fitting a predictive model for the phenotype using all SNPs, while also ranking all SNPs based on their predictive value, and is of benefit in imaging genomics studies where there are many more features than individuals and the number of possible hypotheses is vast (17, 36). The strategy is not impacted by multiple testing concerns since sRRR is based on selecting the variables that contribute most to the relationship between predictors and responses within a multivariate model, rather than performing repeated univariate tests.

In addition to testing the primary hypothesis, the study produced further observations. First, other genes were found to be linked to the endophenotype: Integrin Subunit Alpha 6 (*ITGA6*) (four SNPs) which is involved in insulin-like growth factor 1 signaling, and Fragile X Mental Retardation, Autosomal Homolog 1 (*FXR1*) (two SNPs), while genes involved in lipid metabolism, various neural processes, and neuropsychiatric disease appear to be overrepresented among highly selected SNPs. Further work is needed to understand the significance of these observations. Second, insular cortex was more frequently involved in the top 100 ranked tracts than expected by chance. This is consistent with previous observations. This region is highly connected, receiving direct input from the somatosensory cortex and projecting outputs to both cortical and subcortical regions (37) and is part of the rich club in adults (38) and preterm infants (39). In individuals born preterm the volume, surface area, and folding of the insular cortex is reduced in infancy and early adulthood (40–42) accompanied by alterations in functional activation patterns (43, 44), visual function (45), and cognition (46, 47). Insular abnormalities have been implicated in ASD (48, 49) and attention-deficit hyperactivity disorder (50, 51) that are more prevalent in the preterm population, and the insula has recently been shown in preterms to be the source of spontaneous neuronal bursts (delta brushes) that are instructive in neuronal circuit development (52).

Tractography methods aim to provide an insight into in vivo macrostructural brain connectivity via the diffusion features of brain tissue (53–55). Deterministic tractography algorithms propagate streamlines from a seed region along the main estimated fiber orientation, voxel-by-voxel, with one fiber orientation measurement taken in each voxel. This strategy has been successfully employed in the study of a wide range of neurological and psychiatric diseases, but is typically challenged by areas of high uncertainty such as in approaching the gray matter where the anisotropy is typically lower; at areas of crossing fibers where there are fiber populations traveling in different directions; in tracts such as cortico-striatal projections where functionally related anatomical subdivisions of the striatum project to different cortical areas (information funneling); and in the developing brain where there is generally less myelination and lower anisotropy in the white matter (56), resulting in penalization of long-range connections. In our analysis we therefore used advanced probabilistic approaches that provide an integrated probabilistic analysis of whole tracts by estimating the orientation dispersion function at all points in the tract simultaneously (57, 58). This requires significant computing power but provides a much more robust approach, particularly for interhemispheric fibers and in the context of the higher water content and lower myelination of the developing brain (manifesting as much lower fractional anisotropy values in white matter compared with adults), since it allows tracking to overcome areas of high uncertainty (59), and we have previously applied this successfully to the preterm brain (34).

Further work is required to characterize the exact relationship between PPARG and preterm brain development, notably to determine whether the effect is brain specific or systemic. PPARs are ligand-dependent nuclear hormone receptor transcription factors that are highly involved in cell growth, differentiation, inflammation, lipid and glucose metabolism, and homeostasis (review in ref. 60). The *PPARG* gene is expressed in many tissues including the brain (61), within human white matter, and across brain cell types (*SI Appendix*, Fig. S10). *PPARG* gene expression is up-regulated in neurons in response to excitotoxicity and ischemia (62, 63), and modulates the microglial response to injury (64, 65). PPARG agonists improve neuronal and glial survival in a variety of animal models involving ischemia and inflammation (62, 66–69) and it has been suggested that they provide clinical improvement in children with autism (70). The availability of safe drugs modulating PPARG means that this finding has immediate clinical implications for research into neuroprotective strategies for preterm infants.

Methods

Diffusion MR Imaging. All MRI studies were supervised by an experienced pediatrician or pediatric nurse. Pulse oximetry, temperature, and heart rate were monitored throughout the period of image acquisition; ear protection in the form of silicone-based putty was placed in the external ear (President Putty, Coltene; Whaledent) and Minimuffs (Natus Medical Inc.) were used for each infant.

MRI was performed on a Philips 3-Tesla system (Philips Medical Systems) using an eight-channel phased-array head coil, with acquisition of T2-weighted and 32 direction d-MRI images. All MR images were assessed for the presence of image artifacts and severe motion. Acquisition parameters are in *SI Appendix, Supplementary Methods*. The T2-weighted MRI anatomical scans were reviewed to exclude subjects with extensive brain abnormalities, major focal destructive parenchymal lesions, multiple punctate white-matter lesions, or white-matter cysts. All MR images were assessed for

the presence of image artifacts (inferior-temporal signal dropout, aliasing, field inhomogeneity, etc.) and severe motion. All exclusion criteria were designed so as not to bias the study but preserve the full spectrum of clinical heterogeneity typical of a preterm born population.

Diffusion Tractography. Tractography was performed on diffusion MR data using a modified version of probabilistic tractography that gives an idea of the diffusive transfer between voxels (57). Regions of interest for seeding tractography of cortico–cortical connections were obtained by segmentation of the brain based on a 90-node anatomical neonatal atlas (71), and the resulting segmentations were registered to the diffusion space using a custom neonatal pipeline (72). A weighted adjacency matrix of brain regions was produced for each infant, from which self-connections along the diagonal were removed and upon which symmetry was enforced, with removal of the redundant lower triangle.

Tractography data were linearly adjusted for main covariates (GA, SA, ancestry) and used to reconstruct weighted connectivity matrices for each individual. Each individual matrix was converted into a single vector of numerical values corresponding to edge weights, and appended to form the rows of a single group matrix of n individuals by q edges, where $n = 272$ and $q = 4{,}005$. This vectorized group connectivity matrix was then adjusted for the main covariates of GA, SA, and ancestry, and used as the phenotype in the model. Additionally, a group connectivity matrix was obtained from the median of all individual subject connectivity matrices (*SI Appendix*, Fig. S11) and used for selected downstream analyses. Another matrix of the same dimensions $n \times q$ made up of randomly generated, normally distributed values with mean zero and SD 1 was used as the null phenotype.

Genomewide Genotyping. Genetic predictors consisted of the genomewide genotype matrix recoded in terms of minor allele counts, including SNPs with minor allele frequency (MAF) ≥5% and 100% genotyping rate. Saliva samples were collected using Oragene DNA OG-250 kits (DNAGenotek Inc.), and genotyped on Illumina HumanOmniExpress-24 v1.1 chip (Illumina). Filtering was carried out using PLINK (73). SNPs with MAF ≥5%, 100% genotyping rate, and Hardy–Weinberg equilibrium exact test $P \geq 1 \times 10^{-6}$ were retained, resulting in 556,227 SNPs for further analysis. This genotype matrix was converted into minor allele counts.

Assessment of Population Stratification. Whole genome SNP data were used for IBS, based on pairwise Euclidean distance as implemented in PLINK 1.9 (73), to assess relatedness between individuals. Dimension reduction in the IBS distance matrix was carried out by principal component analysis (74), and the first principal component was used as a covariate in downstream analyses to adjust for population stratification. Information on self-reported ethnicity (as defined in ISB Standard DSCN 11/2008) was collected by asking mothers (and fathers when present) to define themselves according to a list of options. The terms were drawn from Ethnic Category National Codes as in Department of Health Guidance at the time. Parental self-reported ethnicity was summarized into broader categories for the purposes of data visualization by aggregating all White subcategories into a single group "White," all Black subcategories into "Black," and all Asian subcategories into "Asian." In cases where either one parent self-reported as Mixed or if there was a discrepancy between maternal and paternal ethnicities, the term "Mixed" was applied. Where parents were both from an Association of Southeast Asian Nations member state (two cases), the individual was classified by the authors as "SE Asian." These aggregated ethnic categories were used to label the datapoints of PCA plots of the first two principal components of the IBS variance-standardized relationship matrix (*SI Appendix*, Fig. S1). This illustrates the correspondence between the first two components of genetic ancestry and ethnicity, and provides an overview of the cohort population mixture as well as providing a means for phenotype adjustment.

sRRR. Genetic associations with image features were identified using the sRRR model, which has been previously presented in detail (17, 18). sRRR is a method for multivariate modeling of high-dimensional imaging responses (measurements taken over regions of interest or individual voxels) and genetic covariates (e.g., SNPs) that enforces sparsity in the regression coefficients. Given the assumption that only a subset of genetic markers will be found in statistically meaningful association with a subset of image features (i.e., there is a sparse pattern), the model must be able to select those variables. This is achieved by driving some coefficients in the model to zero by penalizing the l_1 norm of the coefficients for genetic markers and image features (*SI Appendix*, Fig. S2). Such sparsity constraints ensure that the model performs simultaneous genotype and phenotype selection (17). The motivation behind this approach is to improve the power to detect causal genetic variants associated with high-dimensional imaging responses (18).

In the current work, genomewide SNPs were tested for association with white-matter tracts reconstructed by probabilistic tractography. We define an $n \times q$ matrix of phenotypes $\mathbf{Y}$ (where the q elements are the 4,005 vectorized edges from the tractography connectivity matrix), and an $n \times p$ matrix $\boldsymbol{X}$ of minor allele counts for p SNPs (where $P = 556{,}227$ SNPs). Selection frequencies for SNPs were ranked by the sRRR method over 1,000 subsamples of 2/3 of the individuals and convergence criterion = 1×10^{-6}. Model parameters in *SI Appendix*, *Supplementary Methods*.

Computational Literature Search. When a query is entered (e.g., a list of genes), it is submitted to the user-selected search engine, and the retrieved results (documents) are fetched from their respective sources. Each document is then parsed into sentences and analyzed for protein–protein associations. Agilent Literature Search (21) uses a set of "context" files (lexicons) for defining protein names (and aliases) and association terms (verbs) of interest.

Graph Theory Assessment of Top 10 Imaging Variables (Edges) Ranked by sRRR. The 10 tractography edges ranked most highly by sRRR were assessed from an "edge-centric" perspective as previously described for the adult brain (15). In the current approach, the importance of selected edges for information flow in the brain is investigated by removing the edges of interest and assessing the impact of their loss on the "communicability" of the network, compared with removing the same number of other randomly selected edges over many iterations. The communicability measure was introduced (16) as a broad generalization of the concept of shortest path between two nodes in a network, incorporating the concept that information flow in a system can also follow routes other than the shortest path (75). Details in *SI Appendix*, *Supplementary Methods*.

Data Availability. sRRR model code and data are available from the authors upon request, subject to approval of future uses by the National Research Ethics Service. Publicly available data: U.K.BEC (www.braineac.org/); RNA-seq data (76) web.stanford.edu/group/barres_lab/brain_rnaseq.html; GIANT (database giant.princeton.edu/); Brainspan (www.brainspan.org/).

The study was approved by the National Research Ethics Service, and written informed consent was given by all participating families.

ACKNOWLEDGMENTS. Our thanks to the children and families who participated in the study, and the nurses, doctors, and scientists who supported the project. MRI scans of preterm infants were in part obtained during an independent program of research funded by the National Institute for Health Research (NIHR) Programme Grants for Applied Research Programme (RP-PG-0707-10154). Further details will be published in the NIHR Journal. In addition, the authors acknowledge further support from the NIHR comprehensive Biomedical Research Centre Award to Guy's & St. Thomas' National Health Service (NHS) Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust, as well as support from the Medical Research Council (MRC) through a Strategic Grant (to A.D.E.) and a Clinical Training Fellowship (to M.L.K.). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the National Health Service, the National Institute for Health Research, the Medical Research Council, the National Institute for Health Research Evaluation, Trials and Studies Coordinating Centre, the Central Commissioning Facility, the Programme Grants for Applied Research program, or the Department of Health.

1. Blencowe H, et al. (2012) National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications. *Lancet* 379:2162–2172.
2. World Health Organization (2012) *Born Too Soon: The Global Action Report on Preterm Birth*, eds Howson CP, Kinney MV, Lawn JE (World Health Organization, Geneva).
3. Moore T, et al. (2012) Neurological and developmental outcome in extremely preterm children born in England in 1995 and 2006: The EPICure studies. *BMJ* 345:e7961.
4. Hack M (2009) Adult outcomes of preterm children. *J Dev Behav Pediatr* 30:460–470.
5. Johnson S, Marlow N (2011) Preterm birth and childhood psychiatric disorders. *Pediatr Res* 69:11R–18R.
6. Montagna A, Nosarti C (2016) Socio-emotional development following very preterm birth: Pathways to psychopathology. *Front Psychol* 7:80.
7. Johnson S, Wolke D (2013) Behavioural outcomes and psychopathology during adolescence. *Early Hum Dev* 89:199–207.
8. Nosarti C, et al. (2012) Preterm birth and psychiatric disorders in young adult life. *Arch Gen Psychiatry* 69:E1–E8.
9. Kuzniewicz MW, et al. (2014) Prevalence and neonatal factors associated with autism spectrum disorders in preterm infants. *J Pediatr* 164:20–25.
10. Keunen K, Counsell SJ, Benders MJNL (2017) The emergence of functional architecture during early brain development. *Neuroimage* 160:2–14.

11. Volpe JJ (2009) Brain injury in premature infants: A complex amalgam of destructive and developmental disturbances. *Lancet Neurol* 8:110–124.
12. Boardman JP, et al. (2014) Common genetic variants and risk of brain injury after preterm birth. *Pediatrics* 133:e1655–e1663.
13. Krishnan ML, et al. (2016) Possible relationship between common genetic variation and white matter development in a pilot study of preterm infants. *Brain Behav* 6: e00434.
14. Xia M, Wang J, He Y (2013) BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS One* 8:e68910.
15. de Reus MA, Saenger VM, Kahn RS, van den Heuvel MP (2014) An edge-centric perspective on the human connectome: Link communities in the brain. *Philos Trans R Soc Lond B Biol Sci* 369:20130527.
16. Estrada E, Hatano N (2008) Communicability in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 77:036111.
17. Vounou M, Nichols TE, Montana G; Alzheimer's Disease Neuroimaging Initiative (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage* 53:1147–1159.
18. Vounou M, et al.; Alzheimer's Disease Neuroimaging Initiative (2012) Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* 60:700–716.
19. Bindea G, et al. (2009) ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25:1091–1093.
20. Wang M, Zhao Y, Zhang B (2015) Efficient test and visualization of multi-set intersections. *Sci Rep* 5:16923.
21. Vailaya A, et al. (2005) An architecture for biological information extraction and representation. *Bioinformatics* 21:430–438.
22. Rani J, Shah AB, Ramachandran S (2015) pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. *J Biosci* 40:671–682.
23. Taşan M, et al. (2015) Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat Methods* 12:154–159.
24. Dempfle A, et al. (2008) Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges. *Eur J Hum Genet* 16:1164–1172.
25. Leviton A, Gressens P, Wolkenhauer O, Dammann O (2015) Systems approach to the study of brain damage in the very preterm newborn. *Front Syst Neurosci* 9:58.
26. Toro R, et al. (2015) Genomic architecture of human neuroanatomical diversity. *Mol Psychiatry* 20:1011–1016.
27. Gaugler T, et al. (2014) Most genetic risk for autism resides with common variation. *Nat Genet* 46:881–885.
28. Arnedo J, et al. (2015) Uncovering the hidden risk architecture of the schizophrenias: Confirmation in three independent genome-wide association studies. *Am J Psychiatry* 172:139–153.
29. Gottesman II, Gould TD (2003) The endophenotype concept in psychiatry: Etymology and strategic intentions. *Am J Psychiatry* 160:636–645.
30. Glahn DC, et al. (2014) Arguments for the sake of endophenotypes: Examining common misconceptions about the use of endophenotypes in psychiatric genetics. *Am J Med Genet B Neuropsychiatr Genet* 165B:122–130.
31. Shen KK, et al. (2014) Investigating brain connectivity heritability in a twin study using diffusion imaging data. *Neuroimage* 100:628–641.
32. Geng X, et al. (2012) White matter heritability using diffusion tensor imaging in neonatal brains. *Twin Res Hum Genet* 15:336–350.
33. van Kooij BJ, et al. (2012) Neonatal tract-based spatial statistics findings and outcome in preterm infants. *AJNR Am J Neuroradiol* 33:188–194.
34. Ball G, et al. (2015) Thalamocortical connectivity predicts cognition in children born preterm. *Cereb Cortex* 25:4310–4318.
35. Counsell SJ, et al. (2008) Specific relations between neurodevelopmental abilities and white matter microstructure in children born preterm. *Brain* 131:3201–3208.
36. Silver M, Montana G; Alzheimer's Disease Neuroimaging Initiative (2012) Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat Appl Genet Mol Biol* 11:7.
37. Augustine JR (1996) Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Res Brain Res Rev* 22:229–244.
38. van den Heuvel MP, Sporns O (2011) Rich-club organization of the human connectome. *J Neurosci* 31:15775–15786.
39. Ball G, et al. (2014) Rich-club organization of the newborn human brain. *Proc Natl Acad Sci USA* 111:7456–7461.
40. Nosarti C, et al. (2008) Grey and white matter distribution in very preterm adolescents mediates neurodevelopmental outcome. *Brain* 131:205–217.
41. Engelhardt E, et al. (2015) Regional impairments of cortical folding in premature infants. *Ann Neurol* 77:154–162.
42. Kersbergen KJ, et al. (2016) Relation between clinical risk factors, early cortical changes, and neurodevelopmental outcome in preterm infants. *Neuroimage* 142: 301–310.
43. Taylor MJ, Donner EJ, Pang EW (2012) fMRI and MEG in the study of typical and atypical cognitive development. *Neurophysiol Clin* 42:19–25.
44. Kalpakidou AK, et al. (2014) Functional neuroanatomy of executive function after neonatal brain injury in adults who were born very preterm. *PLoS One* 9:e113975.
45. Sripada K, et al. (2015) Visual-motor deficits relate to altered gray and white matter in young adults born preterm with very low birth weight. *Neuroimage* 109:493–504.
46. Ullman H, et al. (2015) Neonatal MRI is associated with future cognition and academic achievement in preterm children. *Brain* 138:3251–3262.
47. Botellero VL, et al. (2017) A longitudinal study of associations between psychiatric symptoms and disorders and cerebral gray matter volumes in adolescents born very preterm. *BMC Pediatr* 17:45.
48. Odriozola P, et al. (2016) Insula response and connectivity during social and non-social attention in children with autism. *Soc Cogn Affect Neurosci* 11:433–444.
49. Caria A, de Falco S (2015) Anterior insular cortex regulation in autism spectrum disorders. *Front Behav Neurosci* 9:38.
50. Sripada CS, Kessler D, Angstadt M (2014) Lag in maturation of the brain's intrinsic functional architecture in attention-deficit/hyperactivity disorder. *Proc Natl Acad Sci USA* 111:14259–14264.
51. Adisetiyo V, et al. (2014) Attention-deficit/hyperactivity disorder without comorbidity is associated with distinct atypical patterns of cerebral microstructural development. *Hum Brain Mapp* 35:2148–2162.
52. Arichi T, et al. (2017) Localization of spontaneous bursting neuronal activity in the preterm human brain with simultaneous EEG-fMRI. *Elife* 6:e27814.
53. Conturo TE, et al. (1999) Tracking neuronal fiber pathways in the living human brain. *Proc Natl Acad Sci USA* 96:10422–10427.
54. Mori S, Crain BJ, Chacko VP, van Zijl PC (1999) Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Ann Neurol* 45:265–269.
55. Basser PJ, Pajevic S, Pierpaoli C, Duda J, Aldroubi A (2000) In vivo fiber tractography using DT-MRI data. *Magn Reson Med* 44:625–632.
56. Ciccarelli O, Catani M, Johansen-Berg H, Clark C, Thompson A (2008) Diffusion-based tractography in neurological disorders: Concepts, applications, and future developments. *Lancet Neurol* 7:715–727.
57. Robinson EC, et al. (2008) Multivariate statistical analysis of whole brain structural networks obtained using probabilistic tractography. *Med Image Comput Comput Assist Interv* 11:486–493.
58. Robinson EC, Hammers A, Ericsson A, Edwards AD, Rueckert D (2010) Identifying population differences in whole-brain structural networks: A machine learning approach. *Neuroimage* 50:910–919.
59. Johansen-Berg H, Rushworth MF (2009) Using diffusion imaging to study human connectional anatomy. *Annu Rev Neurosci* 32:75–94.
60. Agarwal S, Yadav A, Chaturvedi RK (2016) Peroxisome proliferator-activated receptors (PPARs) as therapeutic target in neurodegenerative disorders. *Biochem Biophys Res Commun* 483:1166–1177.
61. Warden A, et al. (2016) Localization of PPAR isotypes in the adult mouse and human brain. *Sci Rep* 6:27618.
62. Zhao X, et al. (2009) Neuronal PPARgamma deficiency increases susceptibility to brain damage after cerebral ischemia. *J Neurosci* 29:6186–6195.
63. San YZ, Liu Y, Zhang Y, Shi PP, Zhu YL (2015) Peroxisome proliferator-activated receptor-γ agonist inhibits the mammalian target of rapamycin signaling pathway and has a protective effect in a rat model of status epilepticus. *Mol Med Rep* 12: 1877–1883.
64. Zhao X, et al. (2015) Neuronal interleukin-4 as a modulator of microglial pathways and ischemic brain damage. *J Neurosci* 35:11281–11291.
65. Song GJ, et al. (2016) A novel small-molecule agonist of PPAR-γ potentiates an anti-inflammatory M2 glial phenotype. *Neuropharmacology* 109:159–169.
66. De Nuccio C, et al. (2015) Peroxisome proliferator activated receptor-γ agonists protect oligodendrocyte progenitors against tumor necrosis factor-alpha-induced damage: Effects on mitochondrial functions and differentiation. *Exp Neurol* 271:506–514.
67. Han L, et al. (2015) Rosiglitazone promotes white matter integrity and long-term functional recovery after focal cerebral ischemia. *Stroke* 46:2628–2636.
68. Lan LF, et al. (2015) Peroxisome proliferator-activated receptor-γ agonist pioglitazone ameliorates white matter lesion and cognitive impairment in hypertensive rats. *CNS Neurosci Ther* 21:410–416.
69. Flores JJ, et al. (2016) PPARγ-induced upregulation of CD36 enhances hematoma resolution and attenuates long-term neurological deficits after germinal matrix hemorrhage in neonatal rats. *Neurobiol Dis* 87:124–133.
70. Boris M, et al. (2007) Effect of pioglitazone treatment on behavioral symptoms in autistic children. *J Neuroinflammation* 4:3.
71. Shi F, et al. (2011) Infant brain atlases from neonates to 1- and 2-year-olds. *PLoS One* 6:e18746.
72. Ball G, et al. (2010) An optimised tract-based spatial statistics protocol for neonates: Applications to prematurity and chronic lung disease. *Neuroimage* 53:94–102.
73. Chang CC, et al. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4:7.
74. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
75. Hromkovič J, Klasing R, Pelc A, Ruzicka P, Unger W (2005) *Dissemination of Information in Communication Networks. Broadcasting, Gossiping, Leader Election, and Fault-Tolerance* (Springer, Berlin), pp 317–339.
76. Zhang Y, et al. (2014) An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* 34:11929–11947.