Supplementary Material

Newborn Differential DNA Methylation and Subcortical Brain Volumes as Early Signs of Severe Neurodevelopmental Delay in a South African Birth Cohort Study

Supplementary Methods

DNA methylation

Sample Filtering

Samples were determined to be outliers if detected using two or more of the following methods: detectOutlier function from the lumi package [2], Hannum et al. (2013) method [3] using the locFDR package [4] and both the outlyx and pfilter functions from the watermelon package [5]. However, no samples we detected in more than one method and so none were removed for this reason. Samples containing maternal blood contamination (n = 33) were removed[1]. After the completion of pre-processing technical replicates (n = 7) and samples where reported sex didn't match sex chromosome methylation signatures (n = 3) were removed leaving a total of 273 samples remaining for downstream analysis.

Probe Filtering

This dataset contains 59 probes which detect single nucleotide polymorphisms for quality control purposes and so once observed, were removed. Probes with NAs in ≥ 1% of samples or had a detection p value ≥ $1x10^{-16}$ in ≥ 1% of samples were removed (n = 10,868). Probes which bind to the sex chromosomes were removed due to the distribution differences observed (n = 9,896). Probes whose sequence contains a SNP either at the CpG site being measured or at the site of the single base pair extension with a minor allele frequency ≥ 1%[6,7] were removed (n = 13,598). Autosomal probes which

were *in silico* predicted to non-specifically bind to sex chromosomes in the genome were also removed (n = 9,698) leaving a total of 409,033 probes remaining for downstream analysis [6,7].

References

1. Morin AM, Gatev E, McEwen LM, et al. Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs. *Clin Epigenetics*. 2017;9(1):1-9. doi:10.1186/s13148-017-0370-2

2. Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. *Bioinformatics*. 2008;24(13):1547-1548. doi:10.1093/bioinformatics/btn224

3. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49(2):359-367. doi:10.1016/j.molcel.2012.10.016

4. Efron B, Brit Turnbull B, Narasimhan B. locfdr: Computes Local False Discovery Rates. R package version 1.1-8. 2015. https://cran.r-project.org/package=locfdr.

5. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14(1). doi:10.1186/1471-2164-14-293

6. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):1-17. doi:10.1186/s13059-016-1066-1

7. Price ME, Cotton AM, Lam LL, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics and Chromatin*. 2013;6(1):1-15. doi:10.1186/1756-8935-6-4

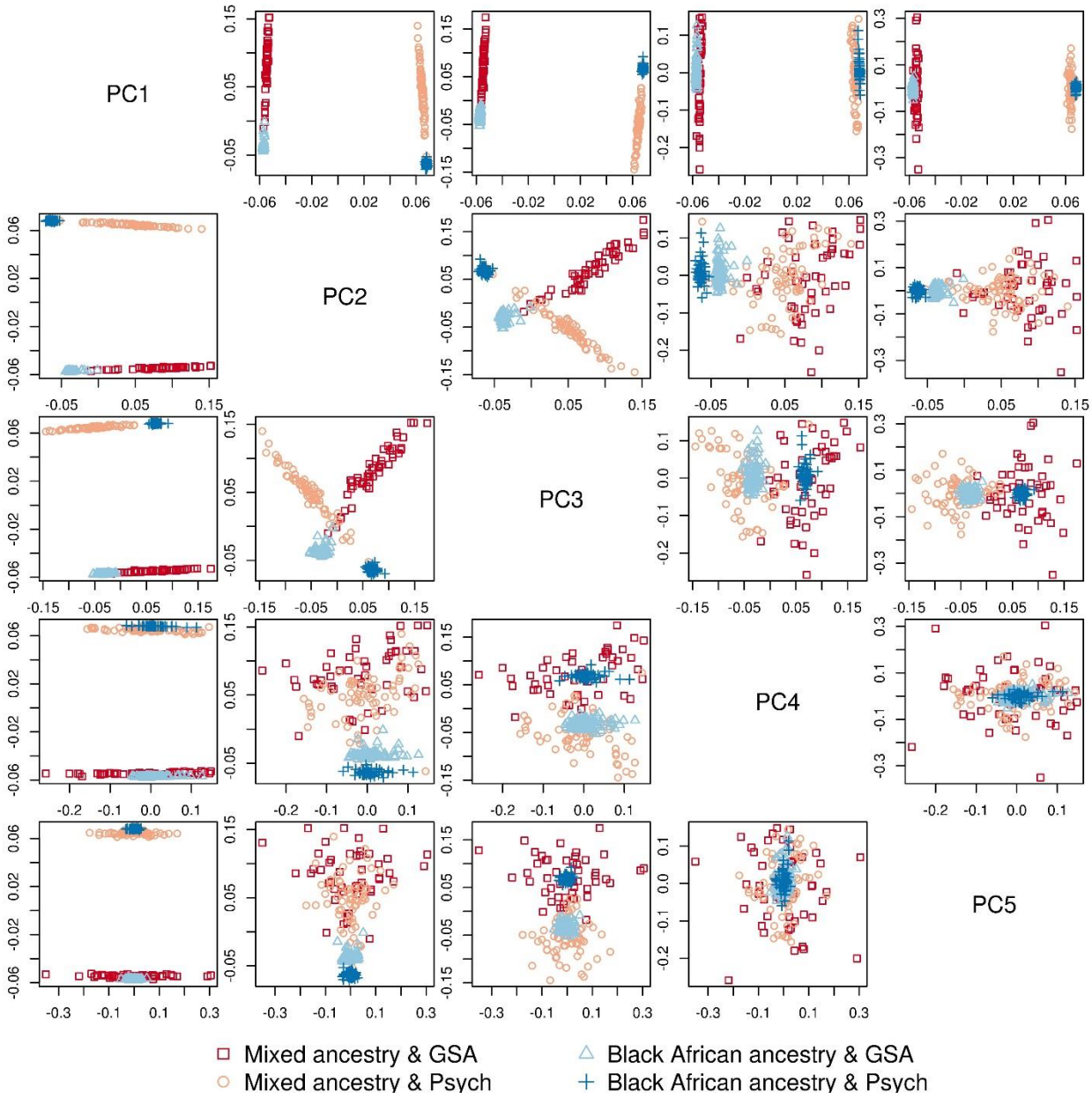

Figure S1. Correlations between the first five genetic principal components (PC1 to PC5). Subgroups defined by ancestry and genotyping array are colored as stated in the legend.

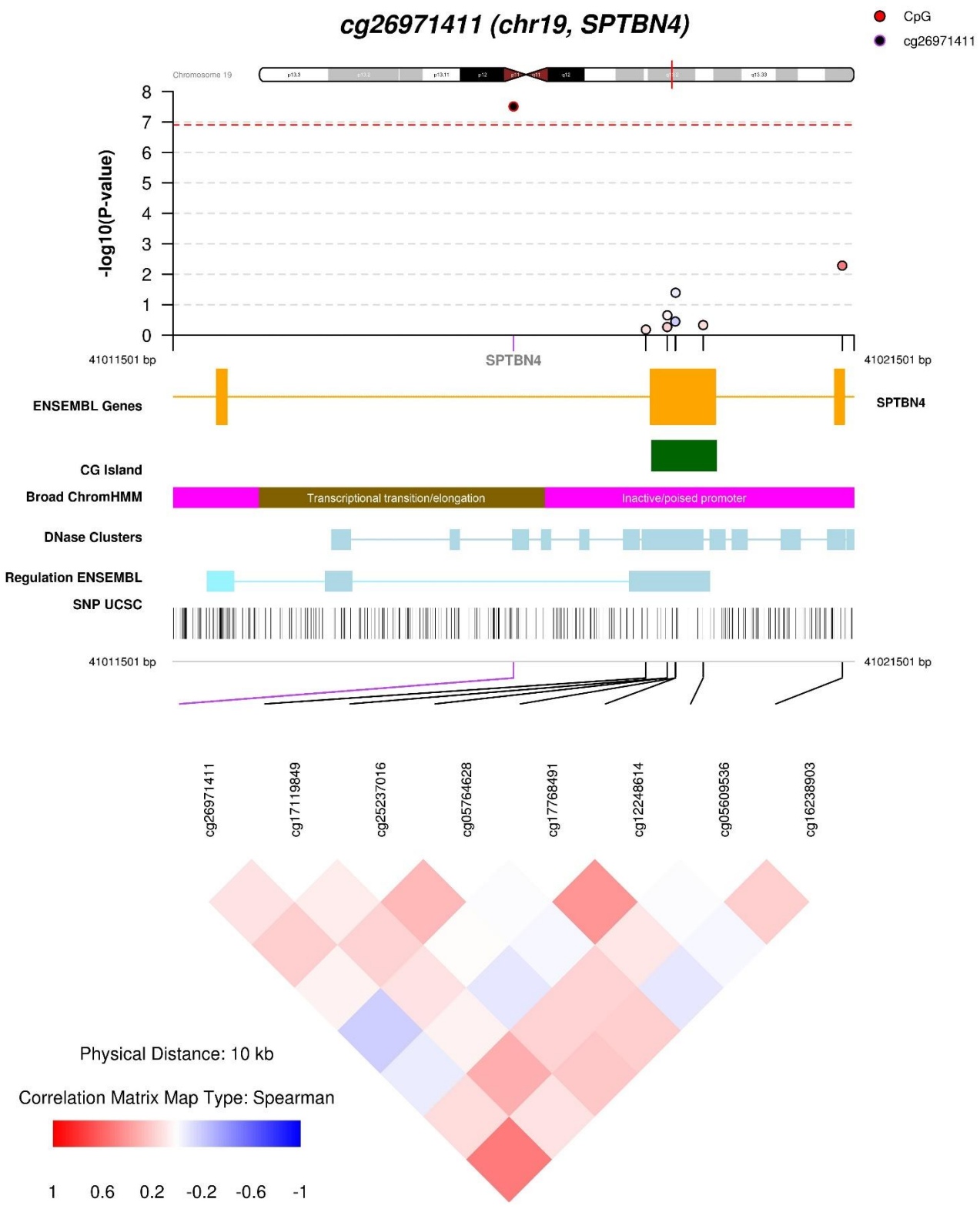

Figure S2. Fine mapping of the association between DNA methylation in cg26971411 (*SPTBN4)* and severe neurodevelopmental delay in motor function (adjusted for sex, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation). cg26971411 is marked in purple. The y-axis indicates the strength of association in terms of negative logarithm of the association P value. Each circle represents a CpG site. Red dashed line within the graph indicates the genome-wide significance threshold (Bonferroni threshold: 1.24e-07). The regulatory information and correlation matrix of other CpG sites in the region with the top hit are shown below the x-axis. Color intensity marks the strength of the correlation and color indicates the direction of the correlation.

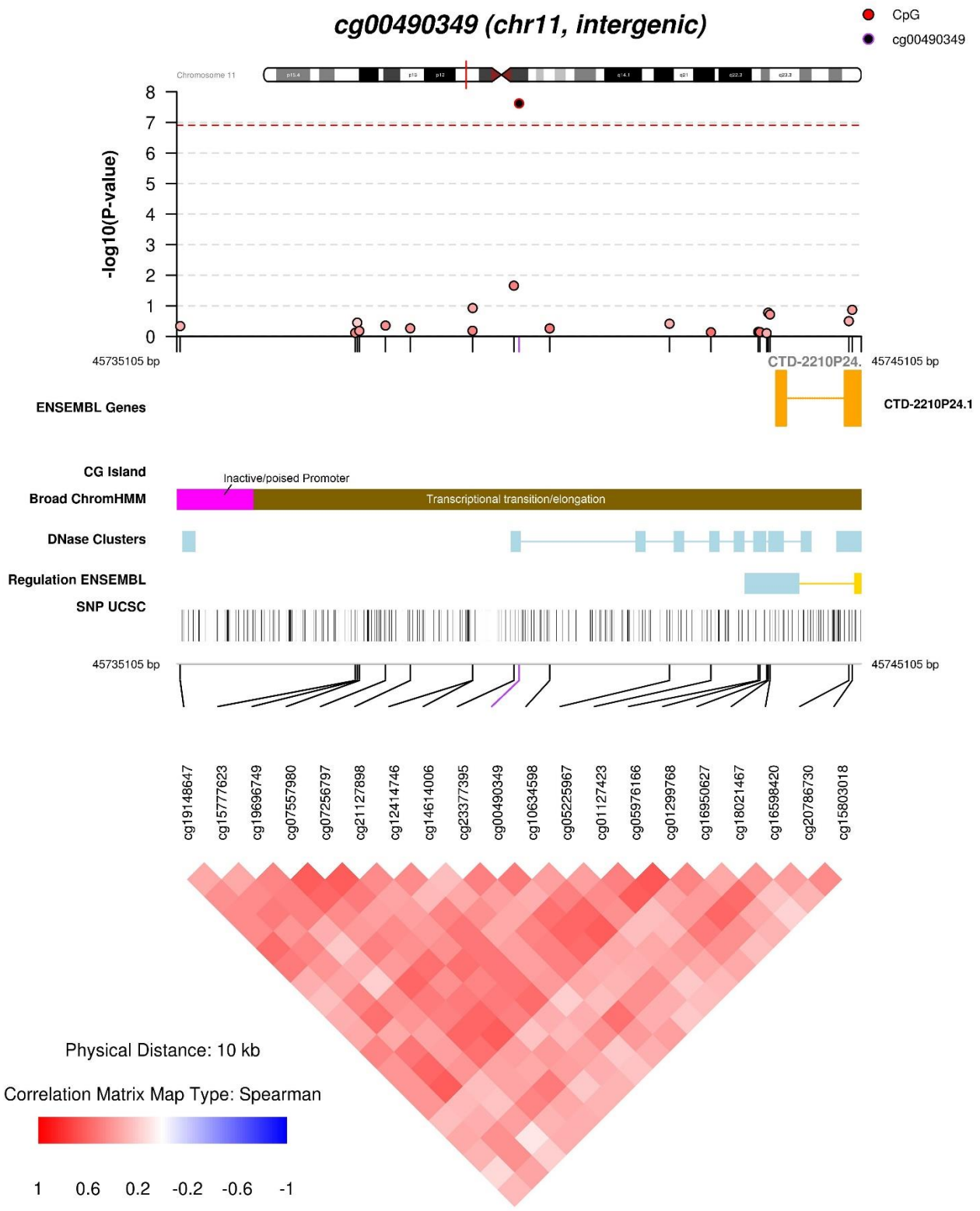

Figure S3. Fine mapping of the association between DNA methylation in cg00490349 and severe delay in language development (adjusted for sex, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation). cg00490349 is marked in purple. The y-axis indicates the strength of association in terms of negative logarithm of the association P value. Each circle represents a CpG site. Red dashed line within the graph indicates the genome-wide significance threshold (Bonferroni threshold: 1.24e-07). The regulatory information and correlation matrix of other CpG sites in the region with the top hit are shown below the x-axis. Color intensity marks the strength of the correlation and color indicates the direction of the correlation.

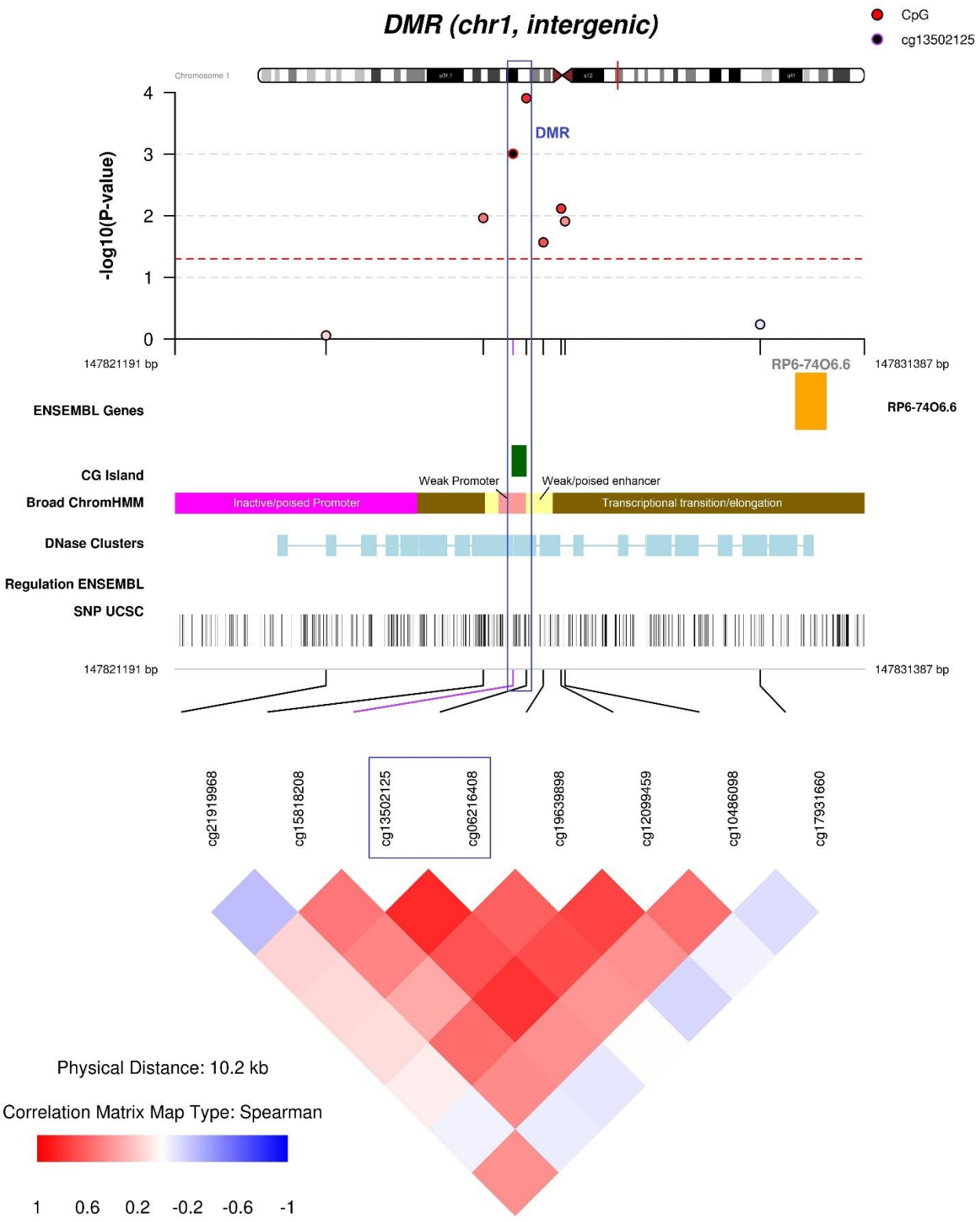

Figure S4. Fine mapping of the differentially methylated region (DMR, highlighted with the blue box) associated with severe delay in language development (adjusted for sex, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation). The CpG which reached the Maximum **Λ** beta for the region is marked in purple (cg13502125). The y-axis indicates the strength of association in terms of negative logarithm of the association P value. Each circle represents a CpG site. Red dashed line within the graph indicates the nominal significance threshold (0.05). The regulatory information and correlation matrix of other CpG sites in the region with the top hit are shown below the x-axis. Color intensity marks the strength of the correlation and color indicates the direction of the correlation.

Table S1. Significant single CpG sites in relation to severe neurodevelopmental delay. Associations tested with multivariate robust linear regression model with empirical Bayes from the R package limma (main model) and linear regression with p-values obtained from normal theory (lm() function in R) as well as from a permutation test with 1 million permutations.

			limma		Linear regression		
Severe neurodevelopmental delay[a]	**CpG**	**chr**	**Δ beta**	**p-value**[a]	**Δ beta**	**p-value normal theory**	**p-value permutation**
Language	cg00490349	11	-0.036	2.41E-08	-0.037	2.67E-08	3.00E-06
Motor function	cg26971411	19	-0.024	3.10E-08	-0.024	3.34E-08	4.30E-05

[a]Bayley score < -2 SD below the mean score

Adjusted for sex, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation).

Δ beta: This coefficient represents the mean difference of DNAm beta values between children with and without severe neurodevelopmental delay. Negative coefficients refer to smaller mean DNAm beta values in children with severe neurodevelopmental delay and positive coefficients refer to larger mean DNAm beta values in children with severe neurodevelopmental.

[a]Bonferroni-threshold (0.05/403933=1.24e-07) was used to correct for multiple testing in the single CpG analyses.

Table S2. Significant single CpG sites and differentially methylated regions (DMRs) in relation to severe neurodevelopmental delay.

A. Severe neurodevelopmental delay[a]				**Cognition**		**Language**		**Motor function**	
CpG	**chr**	**position**	**Gene**	**Δ beta**	**p-value**	**Δ beta group**	**p-value**	**Δ beta group**	**p-value**
cg00490349	11	45740105	N/A	-0.031	1.94E-05	**-0.036**	**2.41E-08**	-0.050	2.73E-07
cg26971411	19	41016501	*SPTBN4*	-0.013	3.93E-05	-0.014	2.57E-06	**-0.024**	**3.10E-08**

B. Mild neurodevelopmental delay[b]				**Cognition**		**Language**		**Motor function**	
CpG	**chr**	**position**	**Gene**	**Δ beta**	**p-value**	**Δ beta group**	**p-value**	**Δ beta group**	**p-value**
cg00490349	11	45740105	N/A	0.0010	1.31E-06	0.0006	0.0002	0.0002	0.1093
cg26971411	19	41016501	*SPTBN4*	0.0004	8.28E-06	0.0002	0.0015	0.0002	0.0149

C. Bayley score				**Cognition**		**Language**		**Motor function**	
CpG	**chr**	**position**	**Gene**	**Δ beta per unit**	**p-value**	**Δ beta per unit**	**p-value**	**Δ beta per unit**	**p-value**
cg00490349	11	45740105	N/A	-0.0118	0.0017	-0.0060	0.1158	-0.0042	0.3191
cg26971411	19	41016501	*SPTBN4*	-0.0044	0.0091	-0.0019	0.2675	-0.0030	0.1100

[a]Bayley score < -2 SD below the mean score, [b]Bayley score < -1 SD below the mean score

Δ beta: This coefficient represents the mean difference of DNAm beta values between children with and without neurodevelopmental delay. Negative coefficients refer to smaller mean DNAm beta values in children with neurodevelopmental delay and positive coefficients refer to larger mean DNAm beta values in children with neurodevelopmental delay.

Δ beta per unit: This coefficient represents the increase of mean DNAm beta values per increase of 1 unit in the Bayley score. Negative coefficients refer to smaller mean DNAm beta values in children with higher Bayley scores and positive coefficients refer to larger mean DNAm beta values in children with higher Bayley scores. Since a low Bayley score refers to neurodevelopmental delay, effect signs in analysis C are different than in A and B.

Adjusted for sex, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation). Bonferroni-threshold (0.05/403933=1.24e-07) was used to correct for multiple testing in the single CpG analyses. Significant p-values in **bold.**

Table S3. Pathway analysis. Top 20 significant gene ontology terms derived from pathway analysis using missMethyl method based on the CpGs with p-values < 0.001 for the association with severe delay in cognition.

	Term	N	DE	P.DE	FDR
GO:0007264	small GTPase mediated signal transduction	536	60	5.12E-05	0.780819
GO:0043547	positive regulation of GTPase activity	375	44	8.31E-05	0.780819
GO:0008047	enzyme activator activity	484	49	0.000151	0.780819
GO:0046578	regulation of Ras protein signal transduction	227	33	0.000165	0.780819
GO:0007265	Ras protein signal transduction	429	48	0.000175	0.780819
GO:0051056	regulation of small GTPase mediated signal transduction	313	40	0.000295	1
GO:0046068	cGMP metabolic process	51	11	0.000378	1
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	151	26	0.000444	1
GO:0005085	guanyl-nucleotide exchange factor activity	307	39	0.000517	1
GO:0060589	nucleoside-triphosphatase regulator activity	332	37	0.000619	1
GO:0004016	adenylate cyclase activity	17	6	0.00068	1
GO:0034199	activation of protein kinase A activity	17	6	0.000906	1
GO:0046069	cGMP catabolic process	6	4	0.000925	1
GO:0005096	GTPase activator activity	261	31	0.001014	1
GO:0043087	regulation of GTPase activity	445	47	0.001034	1
GO:0046872	metal ion binding	3867	264	0.001099	1
GO:0009154	purine ribonucleotide catabolic process	30	8	0.001175	1
GO:0006195	purine nucleotide catabolic process	43	9	0.001207	1
GO:0009261	ribonucleotide catabolic process	31	8	0.001434	1
GO:0003356	regulation of cilium beat frequency	10	4	0.001446	1

N: Number of genes in the GO term; DE: number of genes that are differentially methylated; P.DE: p-value for over-representation of the gene set; FDR: False discovery rate

Table S4. Pathway analysis. Top 20 significant gene ontology terms derived from pathway analysis using missMethyl method based on the CpGs with p-values < 0.001 for the association with severe delay in language development.

	Term	N	DE	P.DE	FDR
GO:0006268	DNA unwinding involved in DNA replication	8	3	0.000725	1
GO:0052735	tRNA (cytosine-3-)-methyltransferase activity	3	2	0.000807	1
GO:0005198	structural molecule activity	682	28	0.001022	1
GO:0060589	nucleoside-triphosphatase regulator activity	332	19	0.00153	1
GO:0043087	regulation of GTPase activity	445	24	0.001653	1
GO:0035983	response to trichostatin A	3	2	0.001767	1
GO:0035984	cellular response to trichostatin A	3	2	0.001767	1
GO:0005883	neurofilament	8	3	0.002357	1
GO:0071526	semaphorin-plexin signaling pathway	29	5	0.00256	1
GO:0044550	secondary metabolite biosynthetic process	26	4	0.002611	1
GO:0043547	positive regulation of GTPase activity	375	20	0.003219	1
GO:0046148	pigment biosynthetic process	48	5	0.003533	1
GO:0002102	podosome	33	5	0.00362	1
GO:0042612	MHC class I protein complex	9	3	0.003761	1
GO:0021544	subpallium development	22	4	0.003817	1
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	9	3	0.004071	1
GO:0045061	thymic T cell selection	19	4	0.004182	1
GO:0030054	cell junction	1188	49	0.00484	1
GO:0006726	eye pigment biosynthetic process	3	2	0.00548	1
GO:0042441	eye pigment metabolic process	3	2	0.00548	1

N: Number of genes in the GO term; DE: number of genes that are differentially methylated; P.DE: p-value for over-representation of the gene set; FDR: False discovery rate

Table S5. Pathway analysis. Top 20 significant gene ontology terms derived from pathway analysis using missMethyl method based on the CpGs with p-values < 0.001 for the association with severe delay in motor function.

	Term	N	DE	P.DE	FDR
GO:0015616	DNA translocase activity	3	2	0.000122	1
GO:0042093	T-helper cell differentiation	50	4	0.000446	1
GO:0002294	CD4-positive, alpha-beta T cell differentiation involved in immune response	52	4	0.000464	1
GO:0005813	centrosome	478	11	0.00048	1
GO:0002287	alpha-beta T cell activation involved in immune response	53	4	0.000488	1
GO:0002293	alpha-beta T cell differentiation involved in immune response	53	4	0.000488	1
GO:0002292	T cell differentiation involved in immune response	59	4	0.000556	1
GO:0043367	CD4-positive, alpha-beta T cell differentiation	61	4	0.000751	1
GO:0002286	T cell activation involved in immune response	79	4	0.001282	1
GO:0035710	CD4-positive, alpha-beta T cell activation	70	4	0.001287	1
GO:0032392	DNA geometric change	77	4	0.001513	1
GO:0005923	bicellular tight junction	111	5	0.001619	1
GO:0003186	tricuspid valve morphogenesis	5	2	0.001683	1
GO:0001695	histamine catabolic process	1	1	0.001763	1
GO:0046539	histamine N-methyltransferase activity	1	1	0.001763	1
GO:0070160	occluding junction	114	5	0.001831	1
GO:0043203	axon hillock	7	2	0.002016	1
GO:0003175	tricuspid valve development	6	2	0.002204	1
GO:0010369	chromocenter	13	2	0.002601	1
GO:0001936	regulation of endothelial cell proliferation	149	5	0.002608	1

N: Number of genes in the GO term; DE: number of genes that are differentially methylated; P.DE: p-value for over-representation of the gene set; FDR: False discovery rate

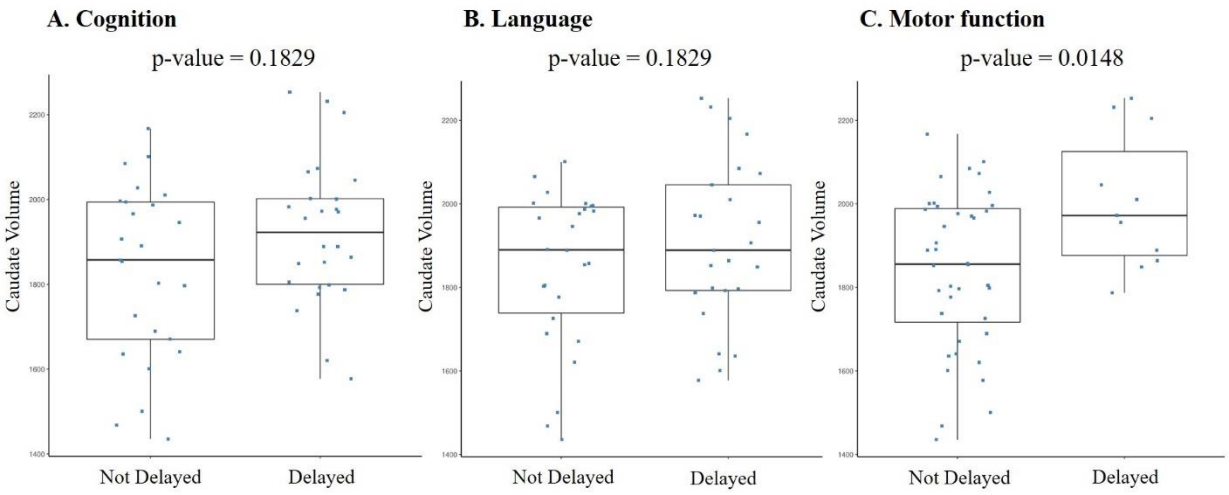

Figure S5. Association between caudate volume in neonates and mild neurodevelopmental delay. Adjusted for age at scan, child sex and intracranial volume.

Table S6. Association between caudate volume (both hemispheres combined) in neonates and cognitive development at two years of age. Adjusted for age at scan, child sex and intracranial volume.

	Cognition		**Language**		**Motor function**	
	Δ caudate volume	**p-value**	**Δ caudate volume**	**p-value**	**Δ caudate volume**	**p-value**
Severe neurodevelopmental delay[a]	117.61	0.2520	**165.30**	**0.0443**	**365.36**	**0.0082**
Mild neurodevelopmental delay[b]	89.42	0.0968	71.73	0.1829	**156.62**	**0.0148**
Bayley score	**-8.15**	**0.0109**	-2.33	0.2592	**-4.94**	**0.0150**

Adjusted for age at scan, child sex and intracranial volume.

[a]Bayley score < -2 SD below the mean score, [b]Bayley score < -1 SD below the mean score

Δ caudate volume for neurodevelopmental delay: This coefficient represents the mean difference of caudate volumes between children with and without neurodevelopmental delay. Negative coefficients refer to smaller caudate volumes in children with neurodevelopmental delay and positive coefficients refer to larger mean caudate volumes in children with neurodevelopmental delay.

Δ caudate volume for Bayley score: This coefficient represents the increase of mean caudate volume per increase of 1 unit in the Bayley score. Negative coefficients refer to smaller mean caudate volumes in children with higher Bayley scores and positive coefficients refer to larger mean caudate volumes in children with higher Bayley scores. Since a low Bayley score refers to neurodevelopmental delay, effect signs in the analyses of the Bayley score are different than for neurodevelopmental delay.

Table S7. Extended adjustment set. Association between MRI imaging data (both hemispheres combined) from neonates and severe delay in motor function at two years of age.

	Main model		Extended model	
MRI	**Δ MRI**	**p-value**	**Δ MRI**	**p-value**
Total grey matter	4605.74	0.6151	3042.67	0.7714
Total white matter	12192.75	0.1408	13193.62	0.1707
Caudate volume	**365.36**	**0.0082**	**349.97**	**0.0144**
Pallidum volume	0.56	0.9806	8.91	0.7288
Putamen volume	34.08	0.5089	36.27	0.5445
Thalamus volume	26.07	0.7347	38.38	0.6473
Amygdala volume	29.38	0.1092	34.46	0.1047
Hippocampus volume	171.76	0.0919	198.63	0.0834

Main model adjusted for age at scan, child sex and intracranial volume. Extended model additionally adjusted for preterm birth, maternal smoking, household income and the first five genetic PCs to control for population stratification.

Δ MRI: This coefficient represents the mean difference of MRI imaging values between children with and without severe neurodevelopmental delay. Negative coefficients refer to smaller MRI imaging values in children with severe neurodevelopmental delay and positive coefficients refer to larger mean MRI imaging values in children with severe neurodevelopmental delay.

Table S8. Extended adjustment set. Association between MRI imaging data (both hemispheres combined) from neonates and severe delay in language development at two years of age.

	Main model		Extended model	
MRI	**Δ MRI**	**p-value**	**Δ MRI**	**p-value**
Total grey matter	4878.53	0.3611	6623.83	0.3334
Total white matter	527.50	0.9143	389.54	0.9516
Caudate volume	**165.30**	**0.0443**	156.83	0.1029
Pallidum volume	-15.72	0.2394	-14.00	0.4057
Putamen volume	-18.37	0.5429	-24.42	0.5348
Thalamus volume	-70.01	0.1149	-119.90	0.0247
Amygdala volume	3.77	0.7289	4.33	0.7606
Hippocampus volume	64.80	0.2814	74.52	0.3301

Main model adjusted for age at scan, child sex and intracranial volume. Extended model additionally adjusted for preterm birth, maternal smoking, household income and the first five genetic PCs to control for population stratification.

Δ MRI: This coefficient represents the mean difference of MRI imaging values between children with and without severe neurodevelopmental delay. Negative coefficients refer to smaller MRI imaging values in children with severe neurodevelopmental delay and positive coefficients refer to larger mean MRI imaging values in children with severe neurodevelopmental delay.

Table S9. Associations between methylation in the significant CpG sites from the EWAS of neurodevelopmental delay (cg26971411 and cg00490349) and MRI imaging data (both hemispheres combined) from neonates.

	cg26971411		**cg00490349**	
MRI	**Δ beta per unit**	**p-value**	**Δ beta per unit**	**p-value**
Total grey matter	-1.28E-07	0.2889	-8.75E-08	0.7759
Total white matter	5.31E-08	0.6919	-2.52E-07	0.4547
Caudate volume	-9.75E-06	0.2416	-2.52E-05	0.2313
Pallidum volume	-4.58E-05	0.3354	2.49E-05	0.8370
Putamen volume	5.27E-06	0.8023	5.70E-05	0.2801
Thalamus volume	1.88E-06	0.8983	-1.36E-05	0.7136
Amygdala volume	-9.15E-05	0.1043	-1.05E-05	0.9426
Hippocampus volume	-9.64E-06	0.4029	-2.01E-05	0.4913

Adjusted for age at scan, child sex, intracranial volume, preterm birth, maternal smoking, household income, the first five genetic PCs to control for population stratification and the first three PCs from cell type proportions (after centered log-ratio transformation).

Δ beta per unit: This coefficient represents the increase of mean DNAm beta values per increase of 1 unit in the MRI imaging values. Negative coefficients refer to smaller mean DNAm beta values in children with higher MRI imaging values and positive coefficients refer to larger mean DNAm beta values in children with higher MRI imaging values.