Supplementary Information

The iPSYCH-GEMS samples

The Danish iPSYCH-GEMS data consists of three sub-samples (GEMS1, GEMS2, and iPSYCH-SCZ). In all sample subsets, cases were identified from the Danish Psychiatric Central Research Register[99], and diagnosed with SCZ by a psychiatrist according to ICD10. Eligible were singletons born to a known mother and resident in Denmark on their one-year birthday. Samples were linked using the unique personal identification number to the Danish Newborn Screening Biobank at Statens Serum Institute, where DNA was extracted from Guthrie cards and whole genome amplified in triplicates as described previously[100,101]. These cards have been systematically stored for individuals born in Denmark since 1 May 1981.

The GEMS1 and 2 samples are subsets of the original discovery and replication samples, respectively, from a Danish genome-wide association study[100], comprising after quality control and exclusion of population outliers a total of 1597 SCZ cases and 1669 controls.
The iPSYCH-SCZ sample is a population based case-cohort sample extracted from a baseline cohort consisting of all children born in Denmark between May 1st 1981 and December 31st 2005. Controls were selected at random among eligible children from the same birth cohort who did not have an SCZ diagnosis by 2013.

GEMS1 and 2 were genotyped using, respectively, the Illumina Human 610-quad and the Illumina HumanCoreExome beadchip (Illumina, San Diego, CA, USA), while the iPSYCH sample was genotyped using the PsychChip array from Illumina (Illumina Infinium PsychArray-24-v1.1, CA, San Diego, USA) at the Broad Institute of Harvard and MIT (Cambridge, MA, USA). Genotype calling of markers with minor allele frequency (maf) > 0.01 was performed by merging callsets from GenCall[102] and Birdseed[103] while less frequent variants were called with zCall[104]. For the iPSYCH sample genotyping and data processing was carried out in 25 waves. Only waves with at least 40 cases were included in the GWAS analysis. Genotypes were processed using the Ricopili pipeline[22], performing stringent quality control of data, removal of related individuals, exclusion of genetic outliers based on principal component analysis, and imputation using the 1000 genomes phase 3 as reference panel. After this processing, genotypes from a total of 4,133 cases and 24,788 controls from the iPSYCH-GEMS samples were included for analysis.

The study was approved by the Danish regional scientific ethics committee and the Danish data protection agency.

Replication of CMC DLPFC association results

We tested for replication of our CMC DLPFC associations in an independent dataset of 4,133 cases and 24,788 controls obtained through collaboration with the iPSYCH-GEMS schizophrenia working group (effective sample size 14,169.5). Replication data was genotyped and imputed in 25 waves. Predicted CMC DLPFC expression was calculated separately for each wave, and merged for association testing. For each gene, the statistical difference in the imputed gene expression levels between cases and controls were computed using a logistic regression analysis. The wave membership of the samples was used as a covariate in the regression. Principal component analysis was done in order to remove genetic outliers. The phenotype specific PCs that are significantly different between cases and controls were included as covariates as well, to account for the population stratification. Related individuals were identified by pairwise IBD analysis and one of every pair (preferably controls) identified as related (piHAT > 0.2) was removed.

Regression formula:
Disease ~ gene-expression + wave1 + wave2 +…..+ wave22 + PC1+PC2+...

We tested for nominal significance and consistent direction of effect between the discovery (PGC + CLOZUK2) and replication (iPSYCH-GEMS) sample, across all genes reaching global genome-wide significance, excluding the MHC region. Additionally, we tested for correlation of association statistics and directions of effect across the entire discovery and replication samples.

We found significant correlation of effect sizes ($p=1.784 \text{ x}10^{-04}$) and –log10 p-values ($p=1.073 \text{ x}10^{-05}$) between our discovery (PGC+CLOZUK2) and replication (iPSYCH-GEMS) samples. Non-MHC Genes reaching genome-wide significance in our discovery sample (49 genes) were significantly more likely to reach nominal significance in the replication sample, and had significantly more consistent directions of effect than might be expected by chance ($p=2.42 \times 10^{-05}$, p=0.044).

Differentially expressed CMC genes have consistent direction of effect in our analysis

The original CMC analysis identified 694 genes with significant differential expression (after FDR-correction). We observed a significantly higher than chance overlap between our two studies. The current DLPFC meta-analysis includes 460 genes of the prior genes; 76 were nominally significant (p<0.05), and five were genome-wide significant (binomial test, $p=7.83 \times 10^{-17}$, 3.36×10^{-16} respectively).

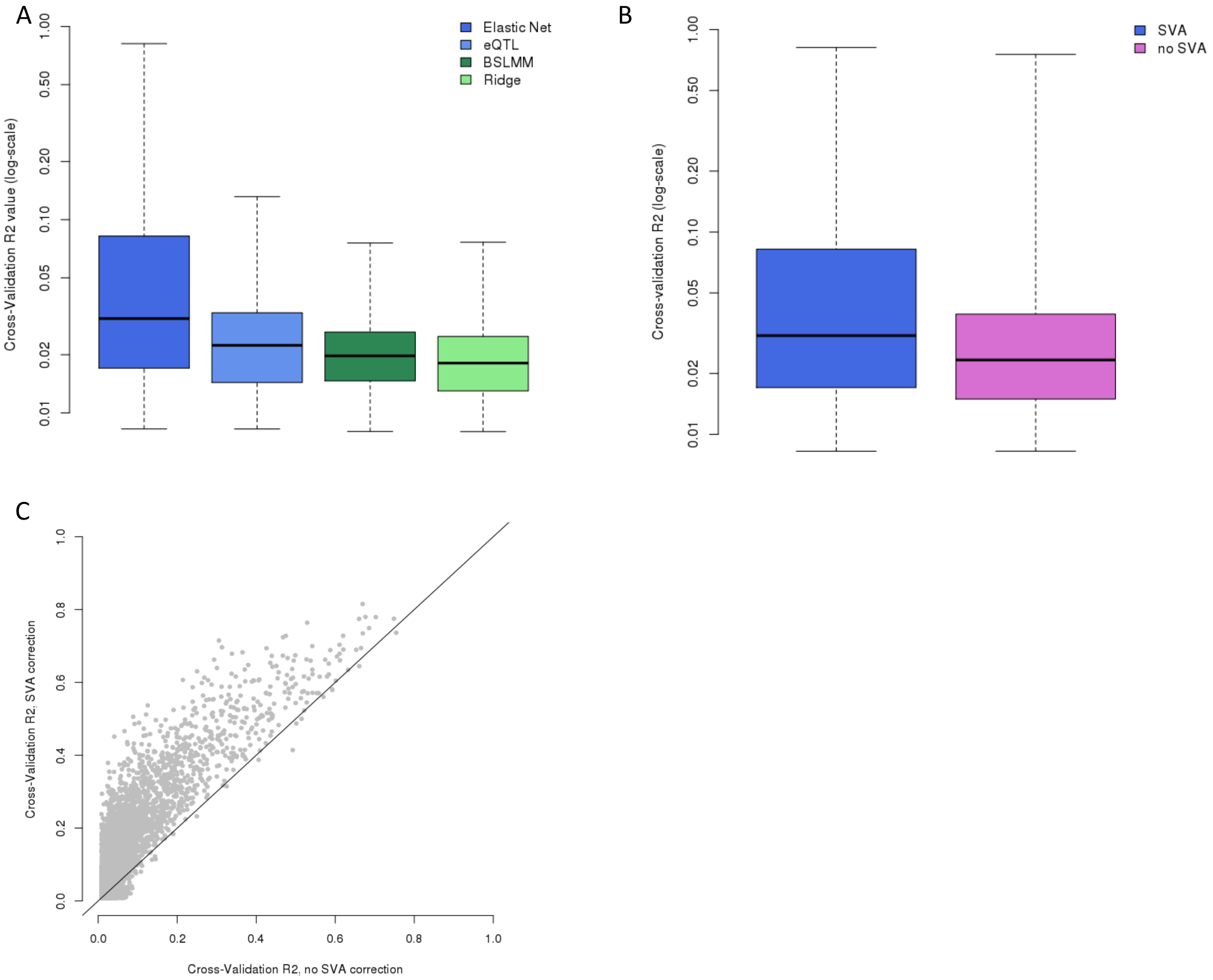
We tested for overlap between our SCZ-associated genes and 36 gene co-expression modules derived from CommonMind Consortium controls. We found significant overlap with six modules (M1c, M8c, M9c, M13c, M14c, M15c), of which two (M9c, M13c) had significant overlap with differentially expressed genes in the original CMC analysis.

Colocalisation of eQTL and GWAS signal

For all genes in the CMC DLPFC predictor database, we calculated posterior probabilities of GWAS and eQTL signals co-localizing (PPA4), using COLOC2. This application is described in detail elsewhere[2,11]. PPA4 values were significantly correlated with prediXcan –log10 p-values (rho=0.31, $p=4.5 \times 10^{-247}$), and absolute Z-Scores (rho=0.30, $p=1.37 \times 10^{-227}$).

Exploring MHC region associations

Genes in the MHC region were analyzed separately, using stepwise conditional analysis as in the non-MHC region. 188 genes were included in the analysis. After conditioning, four genes were genome-wide significant; *BTN1A1, VARS2, HIST1H3B, NUDT3.* Effect size estimates were highly consistent across tissues. For example, effect sizes for all nominally significant genes (p<0.05 in any tissue type) were significantly correlated between all tissue pairs. This effect remained after correction for multiple testing, and regardless of the inclusion or exclusion of genes from the MHC region.

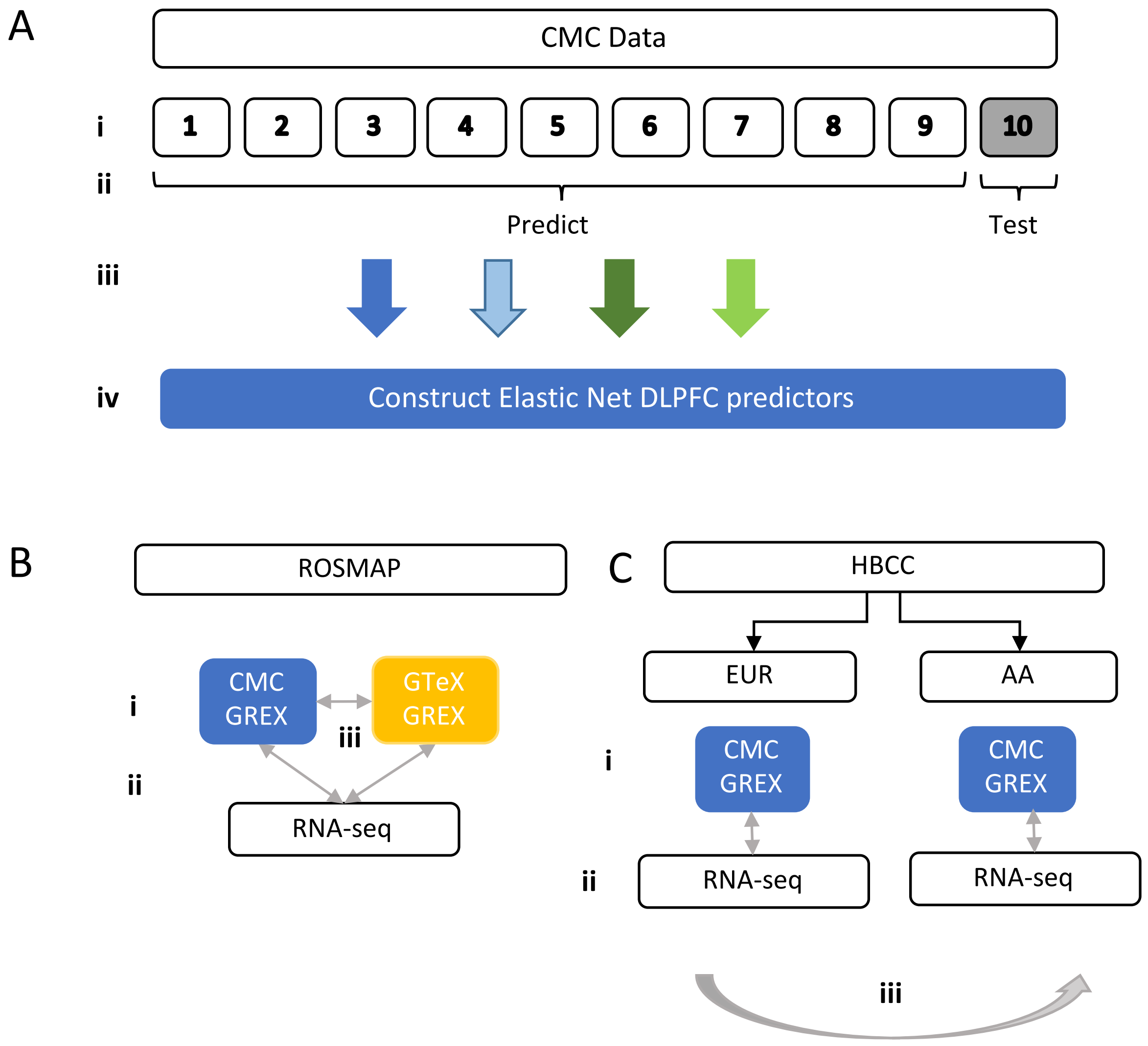


Supplementary Figure 1: Creation of CMC DLPFC Predictors.

A) Cross-validation R^2 values and distributions are significantly higher for Elastic Net regression than for three other approaches. R^2 values are shown only for genes with significant R^2 in cross-validation (~ R^2 >0.008). Number of genes varied according to approach. N elastic net= 15,224; N BSLMM = 16,512; N Ridge = 14,464; N eQTL = 12,835.

B) Correcting for technical effects in gene expression; cross-validation R^2 values are significantly higher in data with SVA correction (N=15,224 genes)

C) Correcting for technical effects in gene expression; cross-validation R^2 values are correlated between data with (n=10,930) and without (n=10,044) SVA correction (spearman rho=0.86, p<2.2e-16)

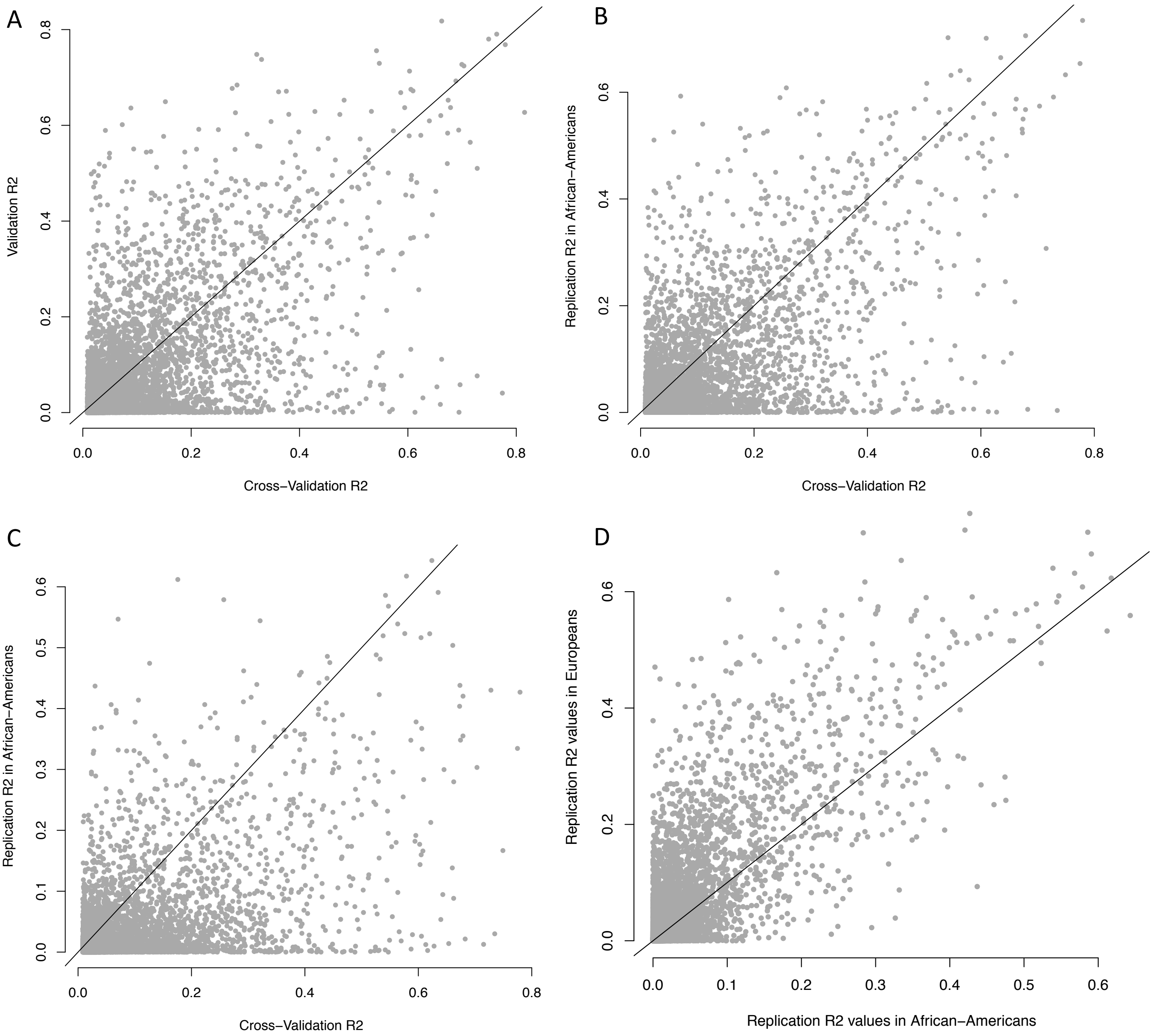


Suppl. Figure 2: Construction and Testing of DLPFC predictors

A. Construction of CMC DLPFC Predictors
i. CommonMind Consortium genotype and gene expression data were partitioned into ten balanced subsets
ii. Predictors were created using a ten-fold cross-validation framework
iii. Predictive accuracy was compared across Elastic net regression, Max eQTL, BSLMM, Ridge Regression

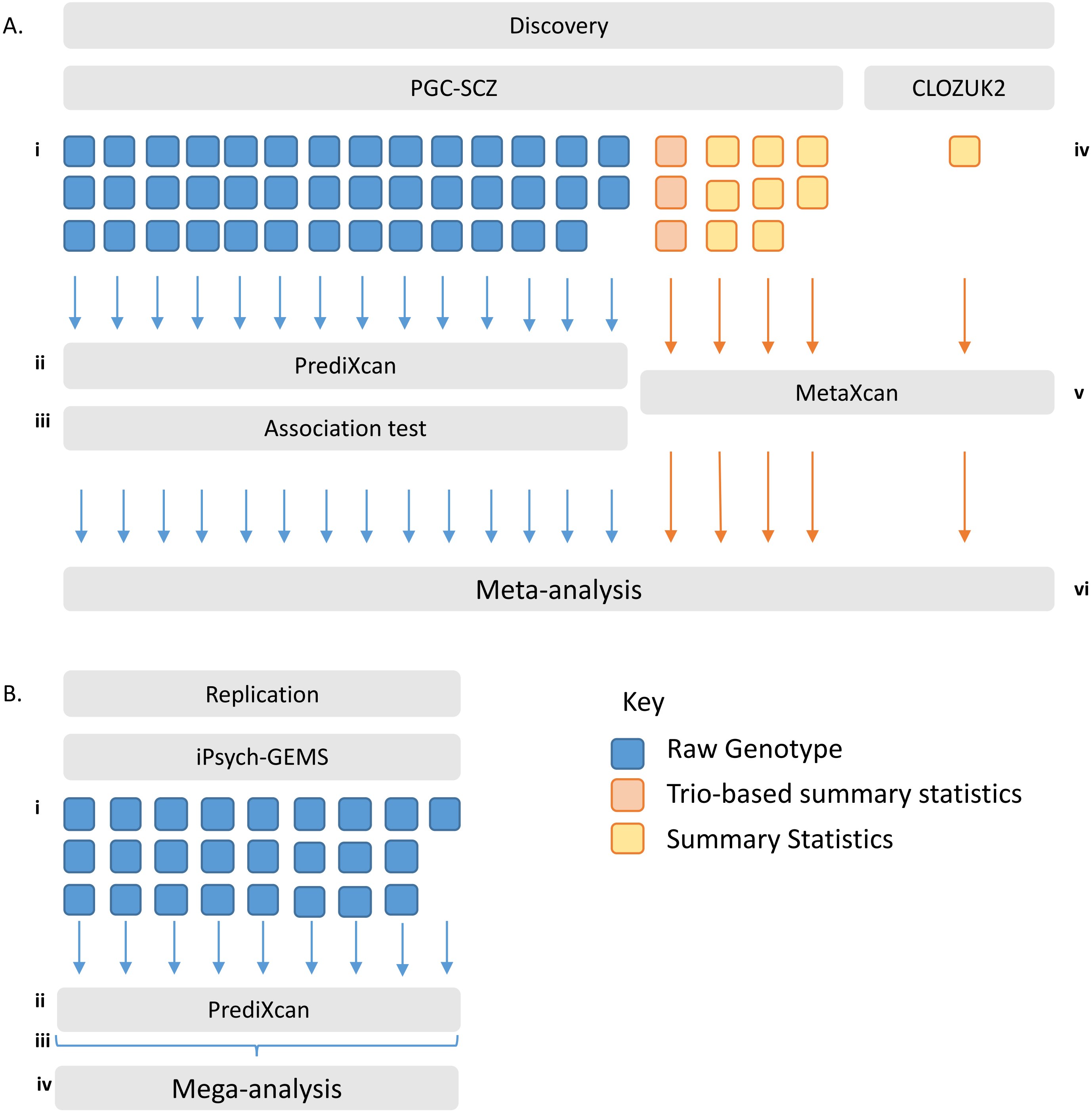
B. Testing and Benchmarking CMC DLPFC and GTeX predictors in ROSMAP data
i. Imputed CMC DLPFC GREX and GTeX GREX in xx ROSMAP samples
ii. Calculate Replication R^2 values by comparing CMC GREX and GTeX GREX to ROSMAP RNA-seq values
iii. Benchmark CMC DLPFC predictor accuracy against GTeX predictor accuracy

C. Comparing predictor accuracy between ethnicities
i. CMC DLPFC GREX was calculated for 280 Europeans (EUR) and 162 African Americans (AA) from the HBCC dataset
ii. Replication R^2 values were calculated by comparing GREX to RNA-seq
iii. Replication R^2 values were compared between EUR and AA analyses



Supplementary Figure 3: Correlation of Cross-validation and replication R^2 values

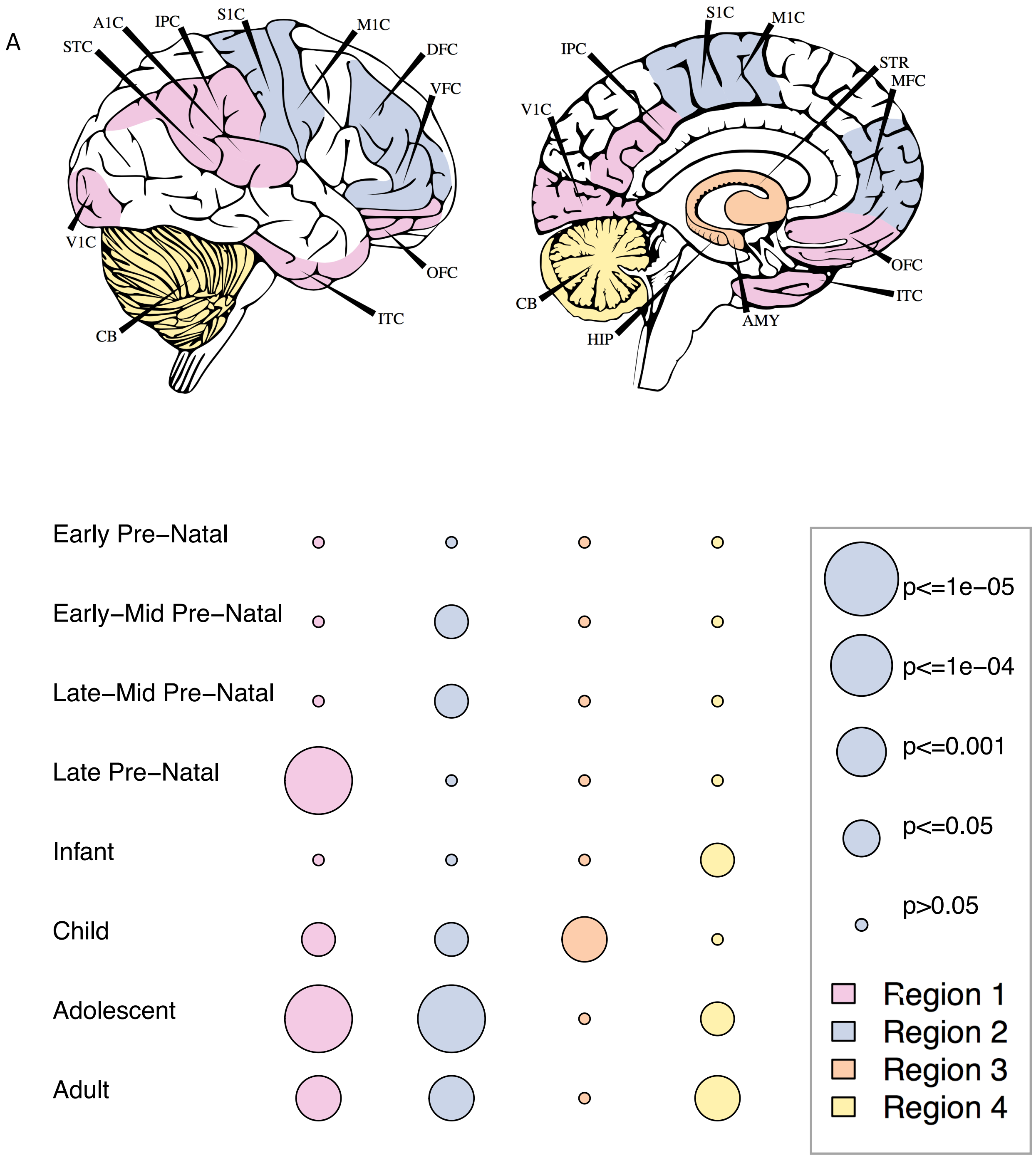
A) Cross-validation R^2 values are significantly correlated with replication R^2 values in ROSMAP (spearman rho=0.62, p<2.2e-16). N=451 post-mortem brain samples, 7,133 genes.
B) Cross-validation R^2 values are significantly correlated with replication R^2 values in HBCC European individuals (spearman rho=0.66, p<2.2e-16). N samples = 280. N genes=4,873.
C) Cross-validation R^2 values are significantly correlated with replication R^2 values in HBCC African-American individuals (spearman rho=0.56, p<2.2e-16). N samples = 162. N genes=4,873.
D) Replication R^2 values are significantly correlated between African-American (N=162) and European individuals (N=280) in HBCC data (microarray) (spearman rho=0.78, p<2.2e-16)



Supplementary Figure 4: Analysis outline.

A) Discovery Samples. 41 PGC-SCZ cohorts had available raw genotypes (i). Predicted DLPFC gene expression was caluclated in each cohort using prediXcan (ii) and tested for association with case-control status (iii). 11 PGC cohorts (3 trio, 8 case-control) and the CLOZUK2 cohort had only summary statistics available (iv). MetaXcan was used to calculate DLPFC associations for each cohort (v). Results were meta-analysed across all 53 cohorts (vi). This procedure was repeated for 12 GTEx prediction models.

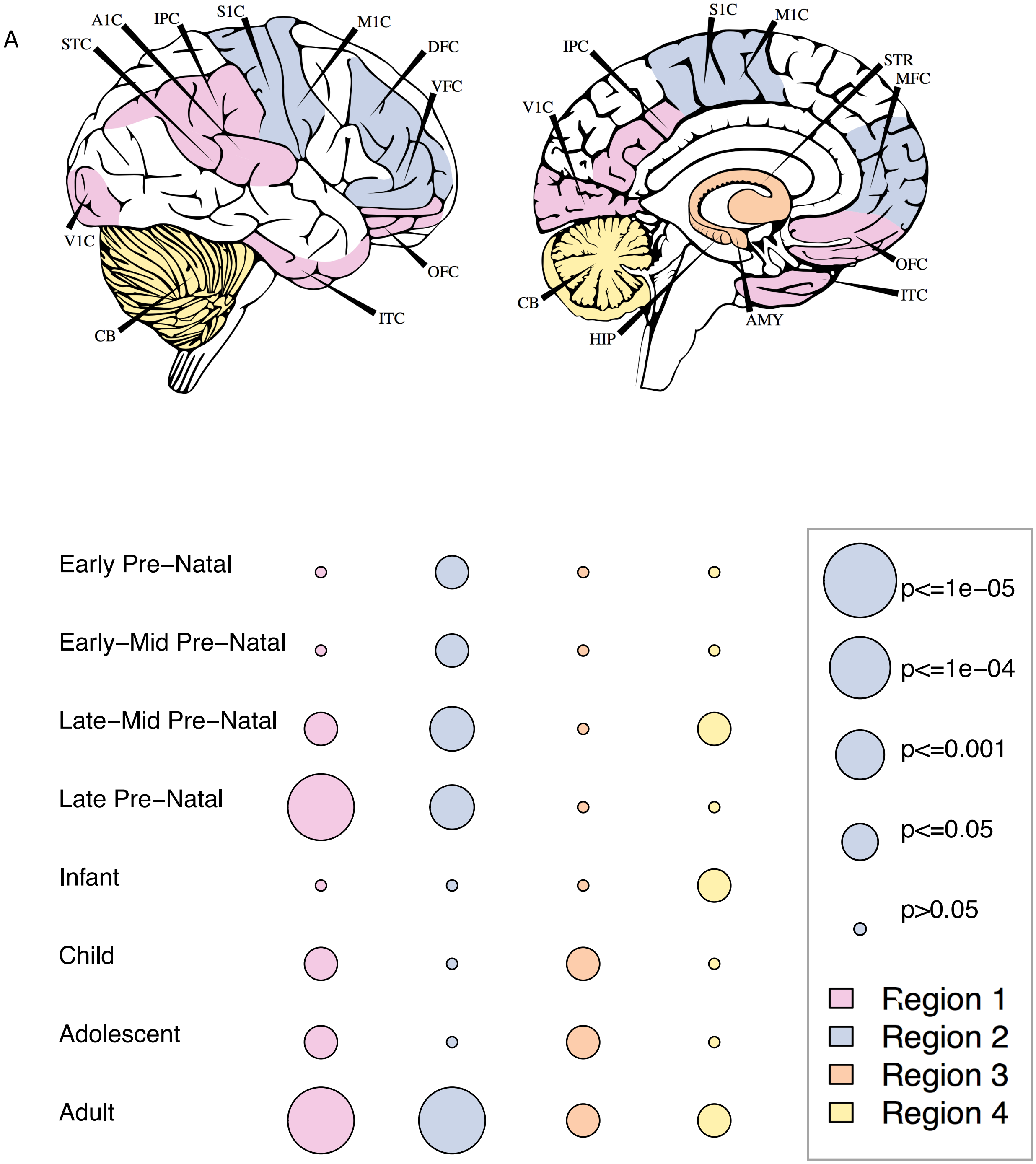
B) Replication Samples. iPsych-GEMS samples were collected in 25 waves (i). Predicted DLPFC gene expression was calculated in each wave separately using prediXcan (ii) and merged for association testing (iii). A mega-analysis was run across all 25 waves, using wave membership as a covariate in the regression (iv)



Suppl. Figure 5: SCZ-associated genes are co-expressed throughout development and across brain regions

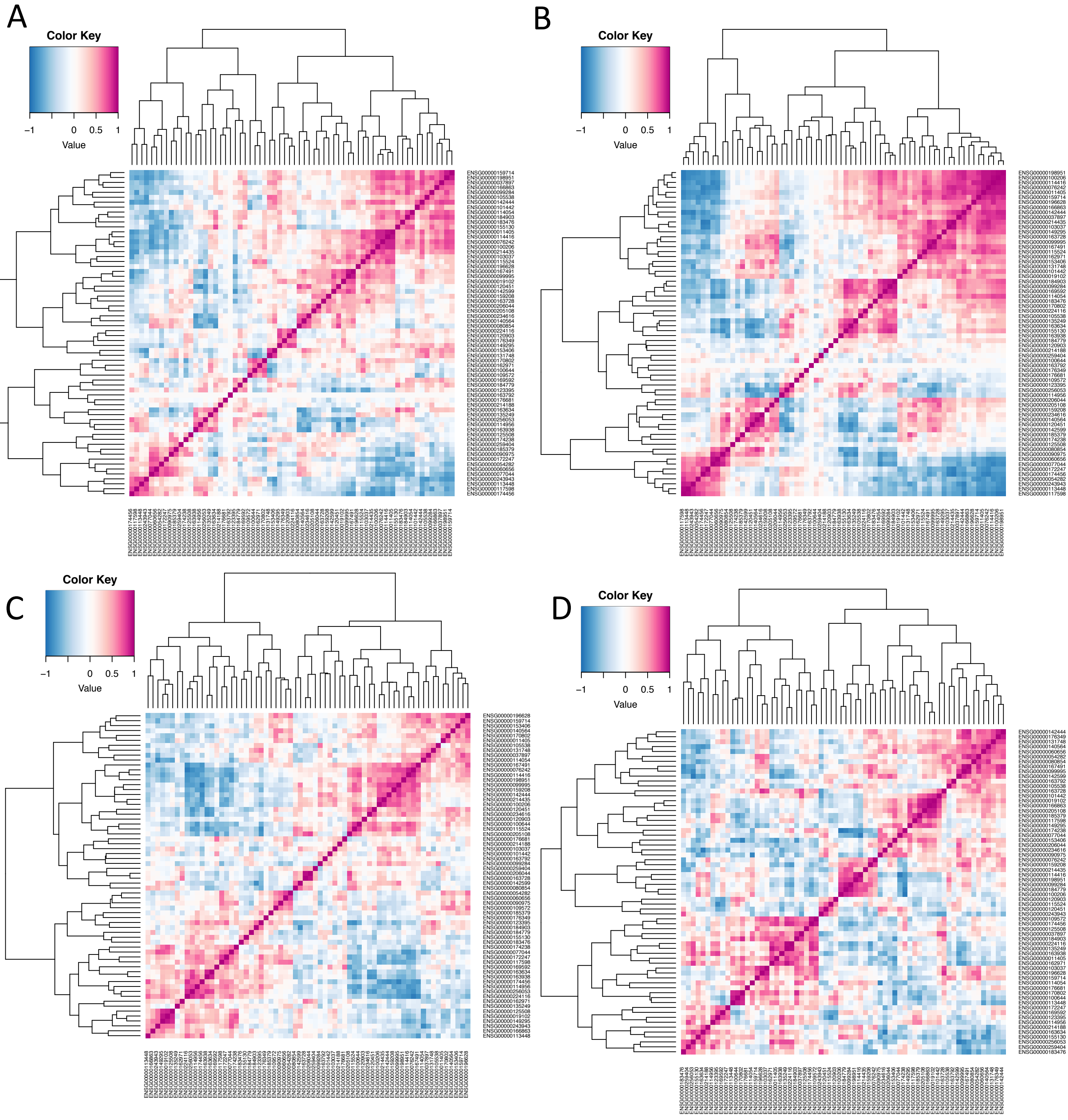
A) Brain tissues selected for each of four brainspan regions. Brainspan includes 525 samples from 43 unique individuals.

B) Average Path lengths were calculated for all pairs of SCZ-associated genes, and compared to permuted gene networks to obtain empirical significance levels.



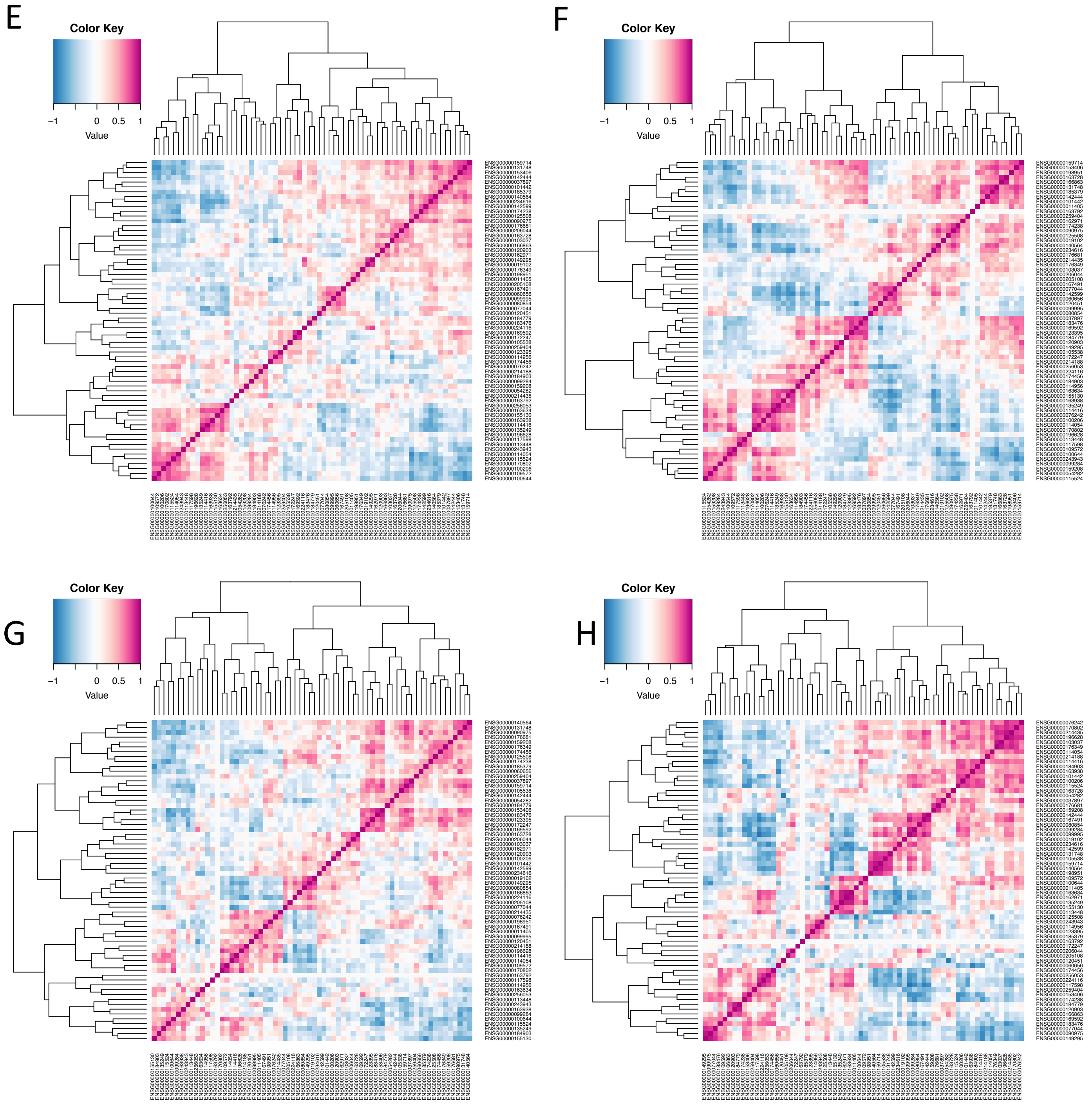
Suppl. Figure 6: SCZ-associated genes are co-expressed throughout development and across brain regions

A) Brain tissues selected for each of four brainspan regions. Brainspan includes 525 samples from 43 unique individuals.
B) Number of edges in a network of SCZ-associated genes, compared to permuted gene networks to obtain empirical significance levels.



Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Early pre-natal period

A) Region 1
B) Region 2
C) Region 3
D) Region 4



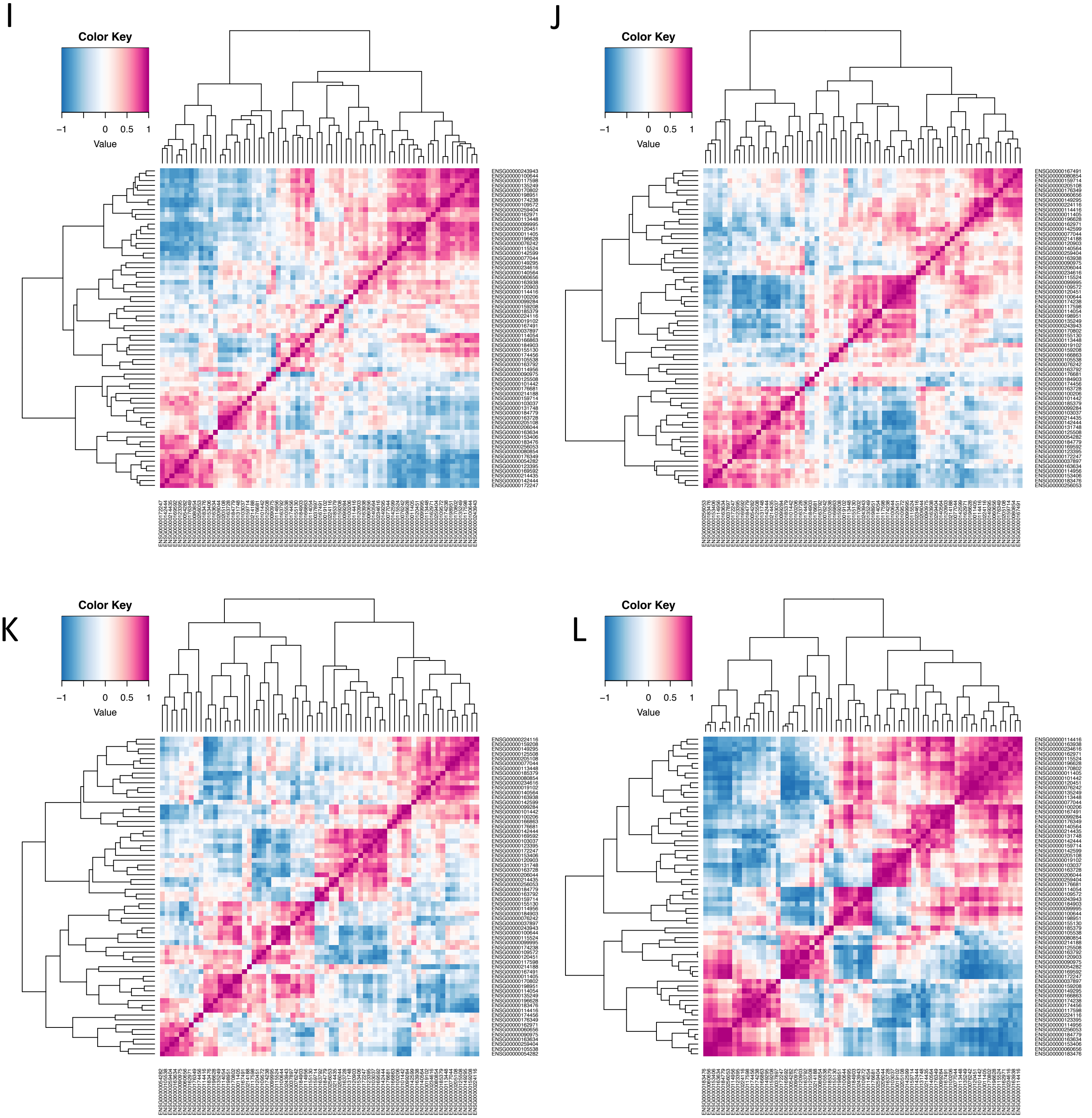
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Early-mid pre-natal period

E) Region 1

F) Region 2

G) Region 3

H) Region 4



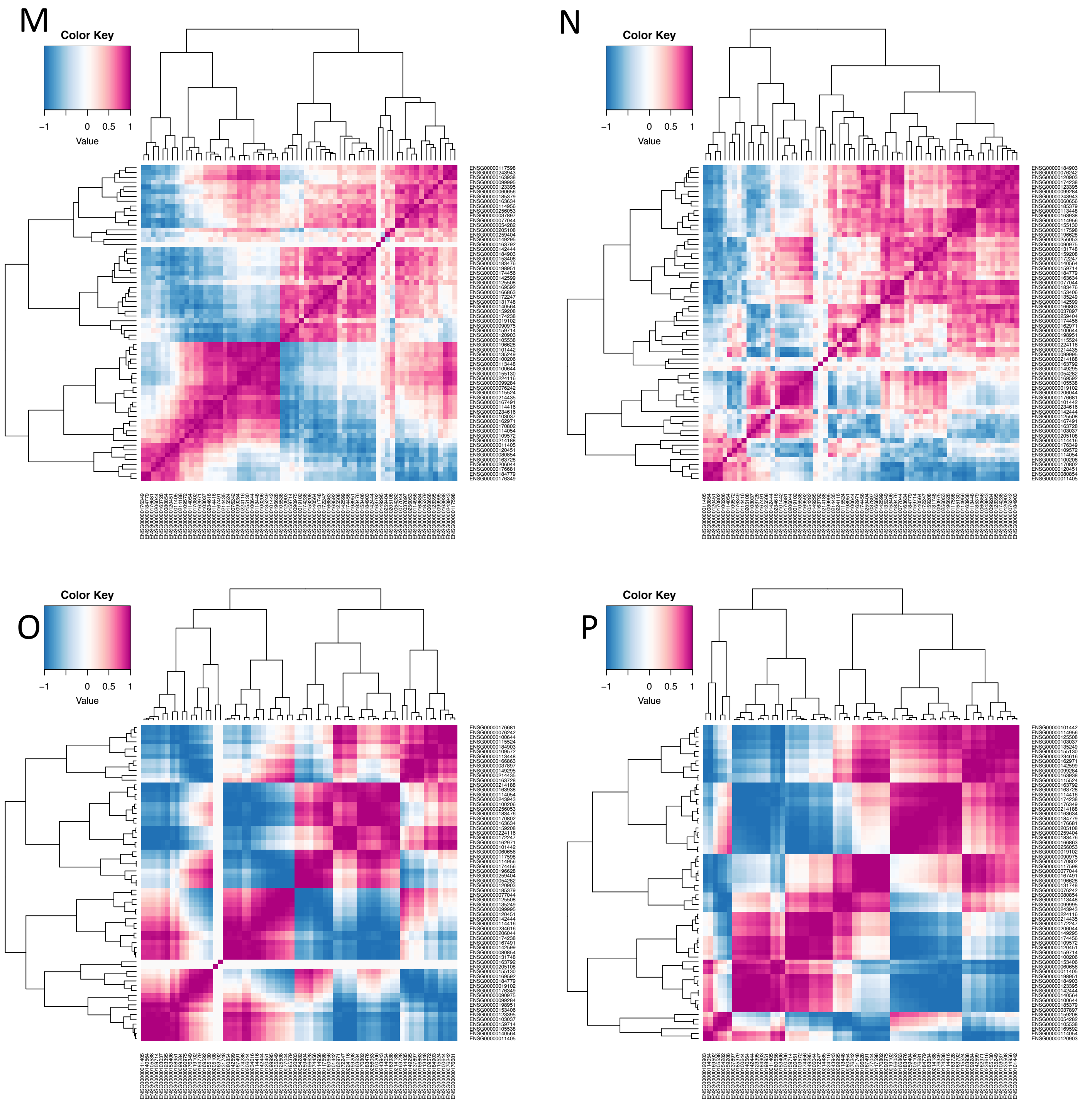
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Late-mid pre-natal period

I) Region 1

J) Region 2

K) Region 3

L) Region 4



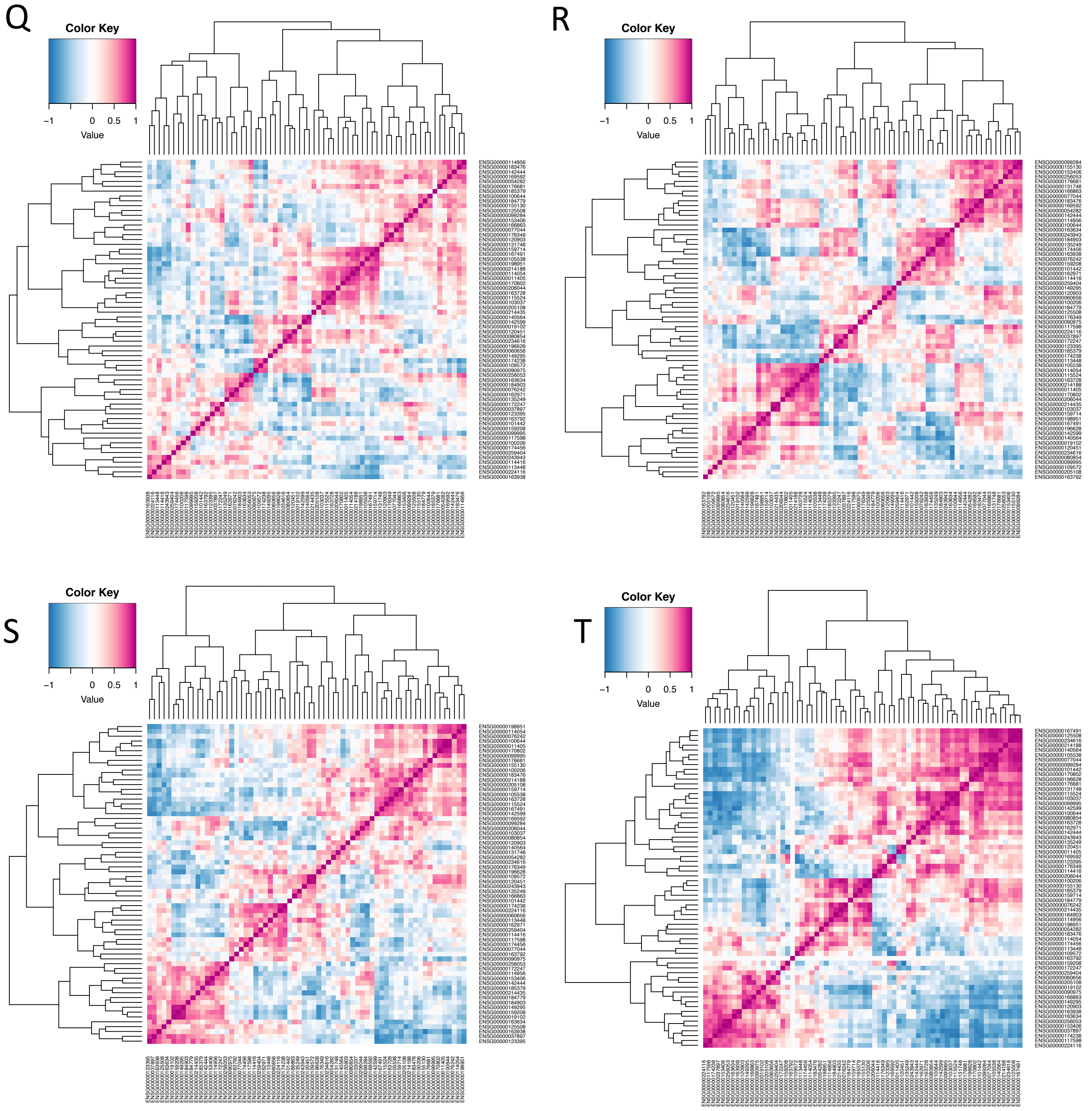
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Late pre-natal period

M) Region 1

N) Region 2

O) Region 3

P) Region 4



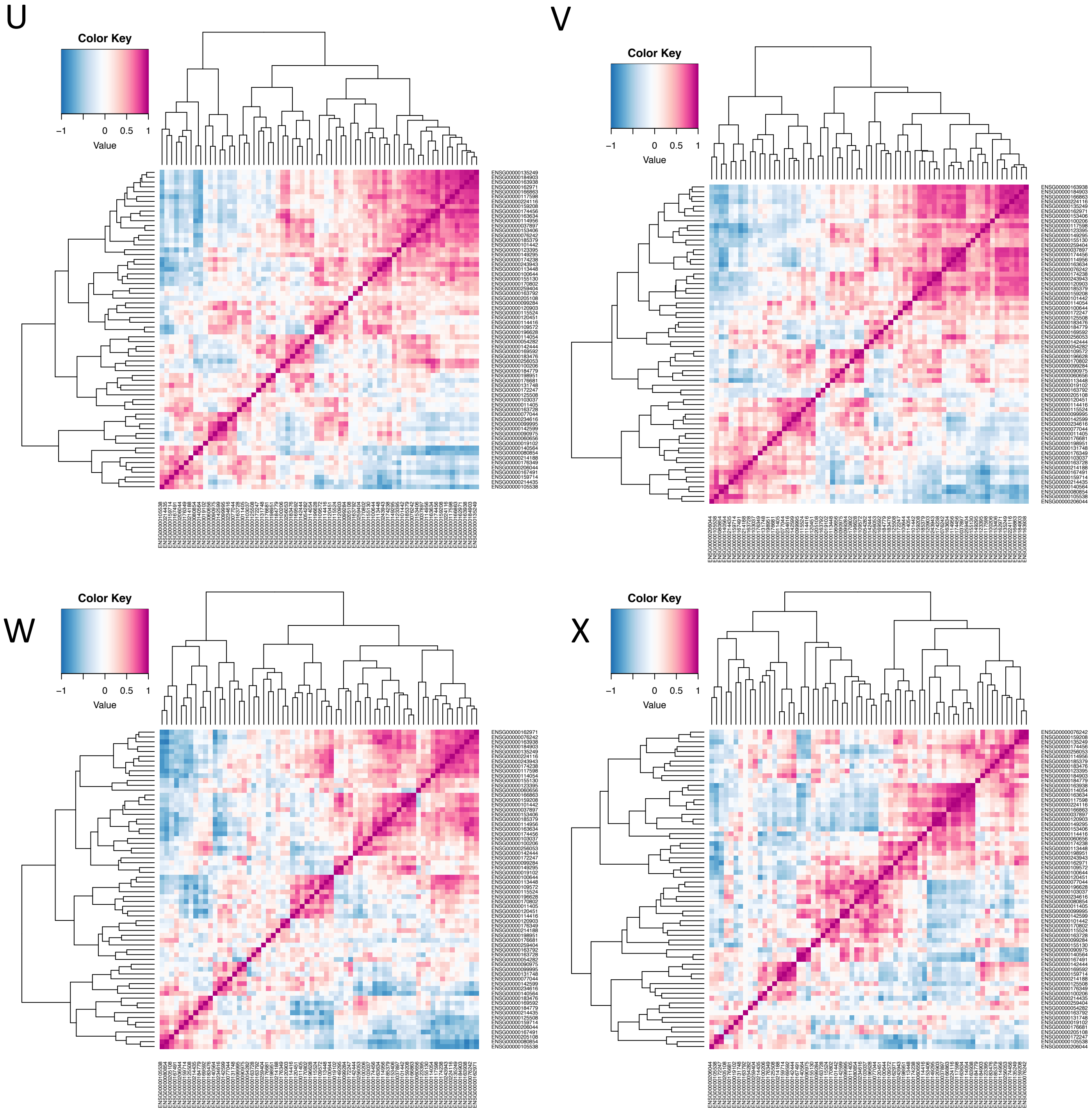
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Infancy

Q) Region 1

R) Region 2

S) Region 3

T) Region 4



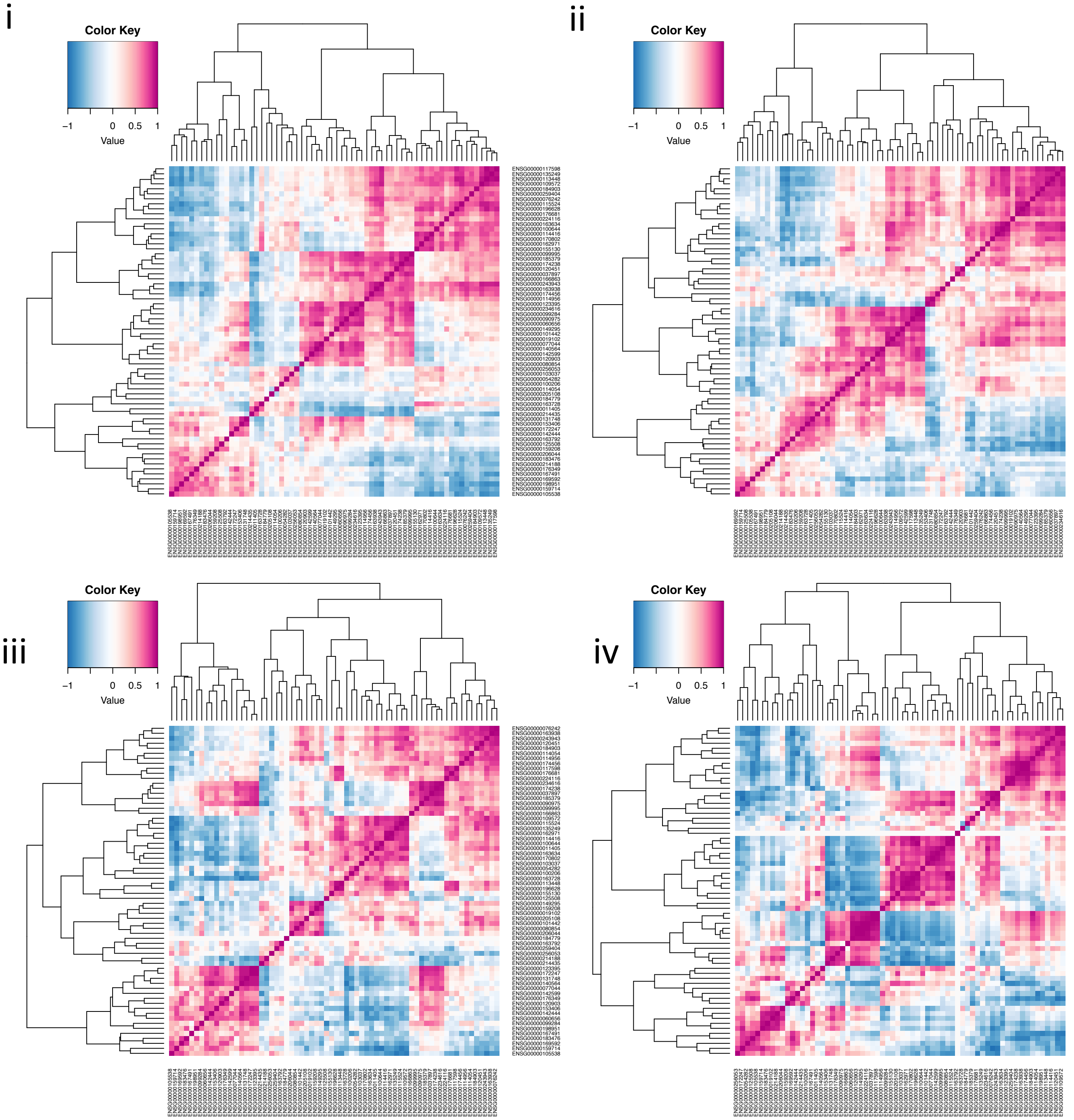
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Childhood

U) Region 1

V) Region 2

W) Region 3

X) Region 4



Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions.

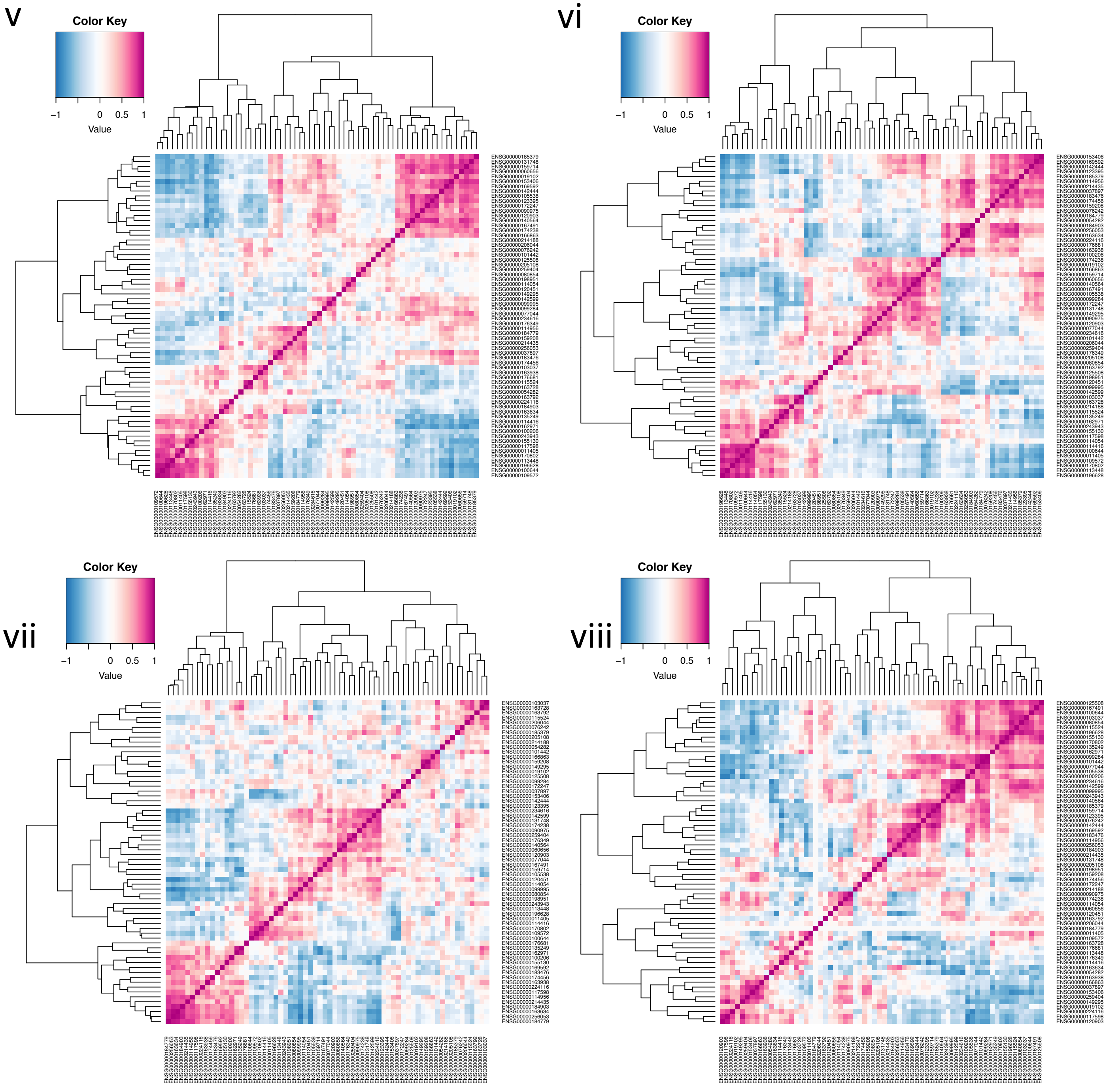
Adolescence

i) Region 1

ii) Region 2

iii) Region 3

iv) Region 4



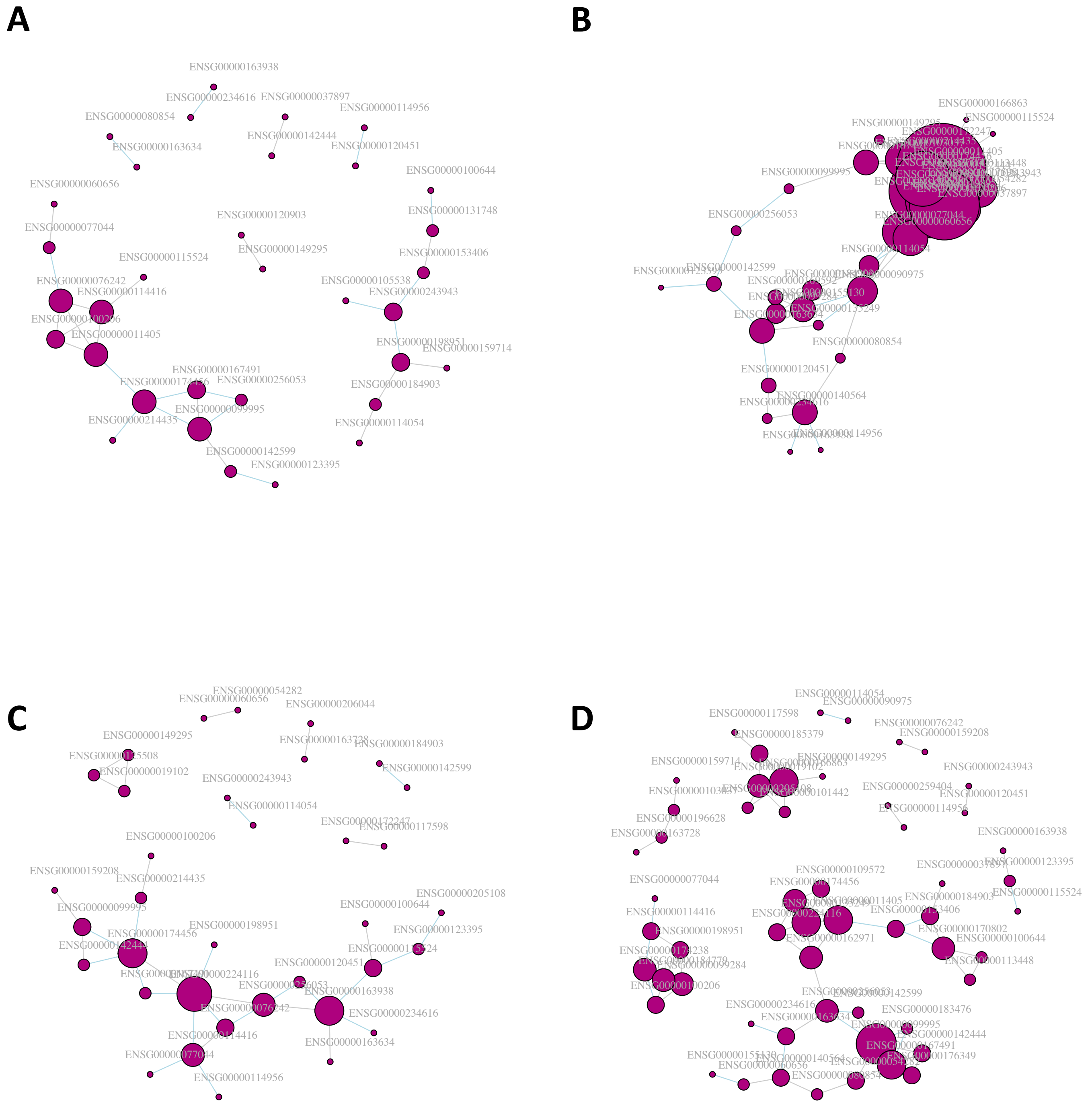
Supplementary Figure 7: Correlation of gene expression levels throughout development and across brain regions. Adulthood

v) Region 1

vi) Region 2

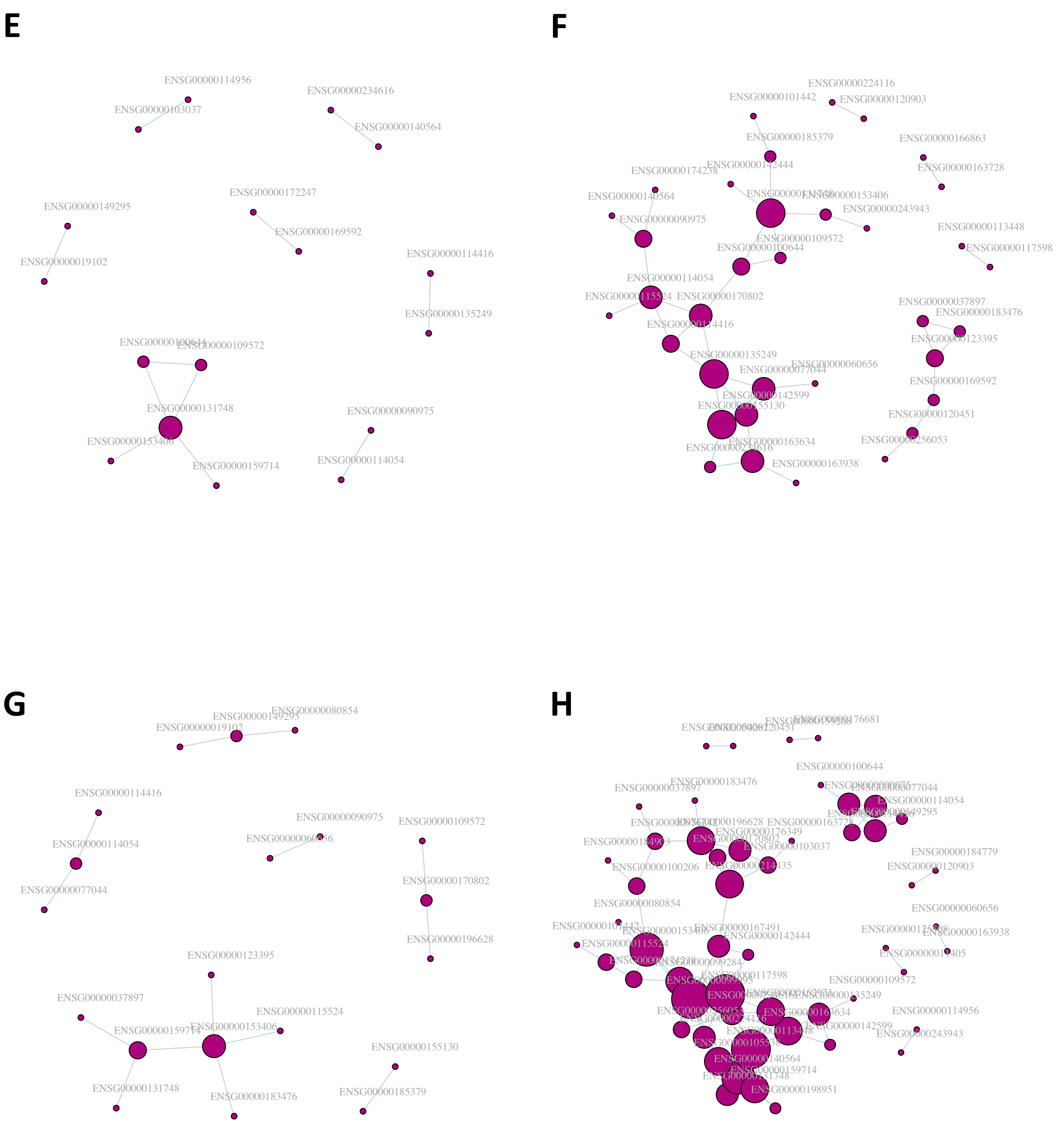
vii) Region 3

viii) Region 4



Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Early pre- natal period

A) Region 1
B) Region 2
C) Region 3
D) Region 4



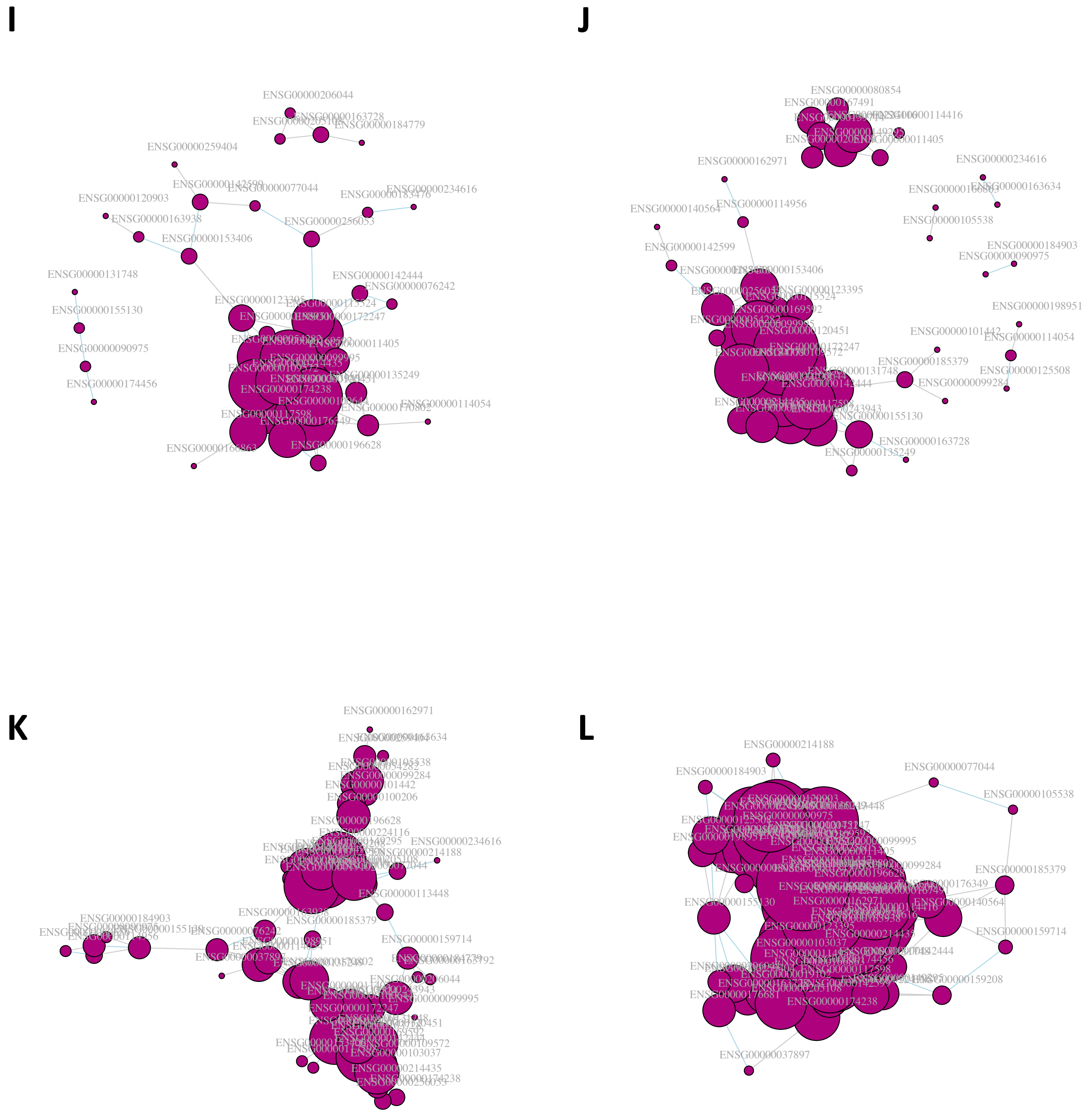
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Early-mid pre- natal period

E) Region 1

F) Region 2

G) Region 3

H) Region 4



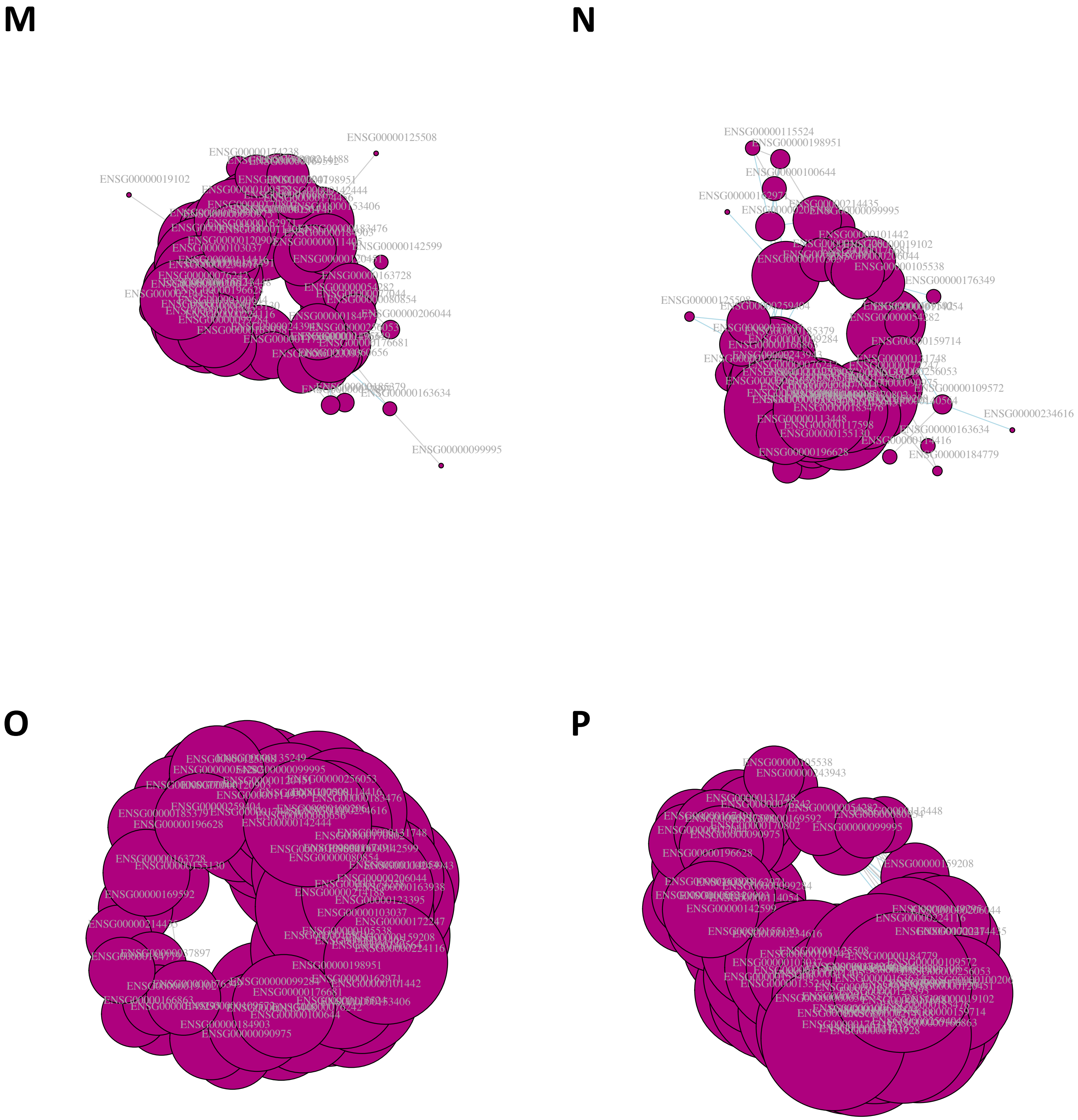
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Late-mid pre- natal period

I) Region 1

J) Region 2

K) Region 3

L) Region 4



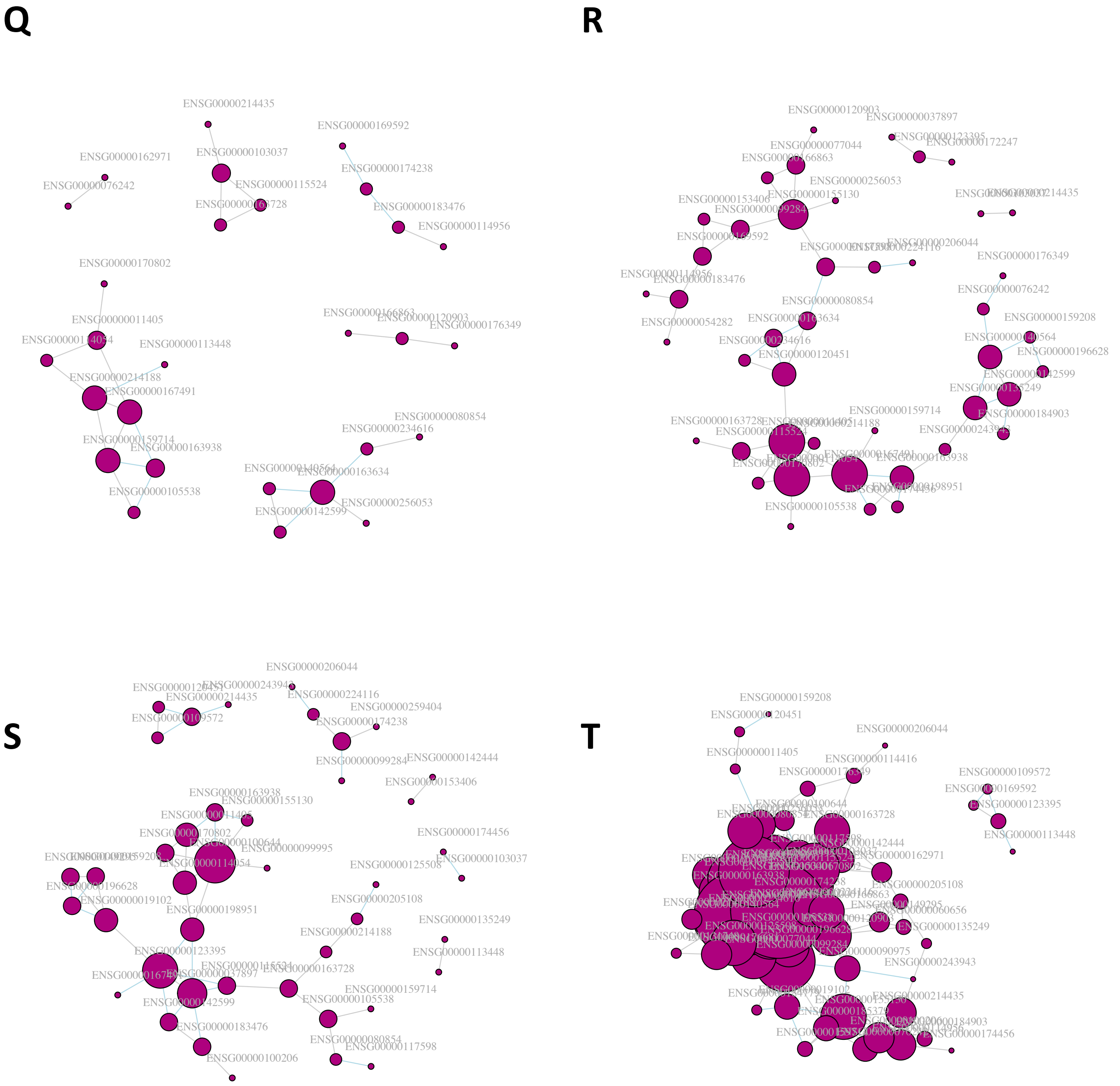
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Late pre- natal period

M) Region 1

N) Region 2

O) Region 3

P) Region 4



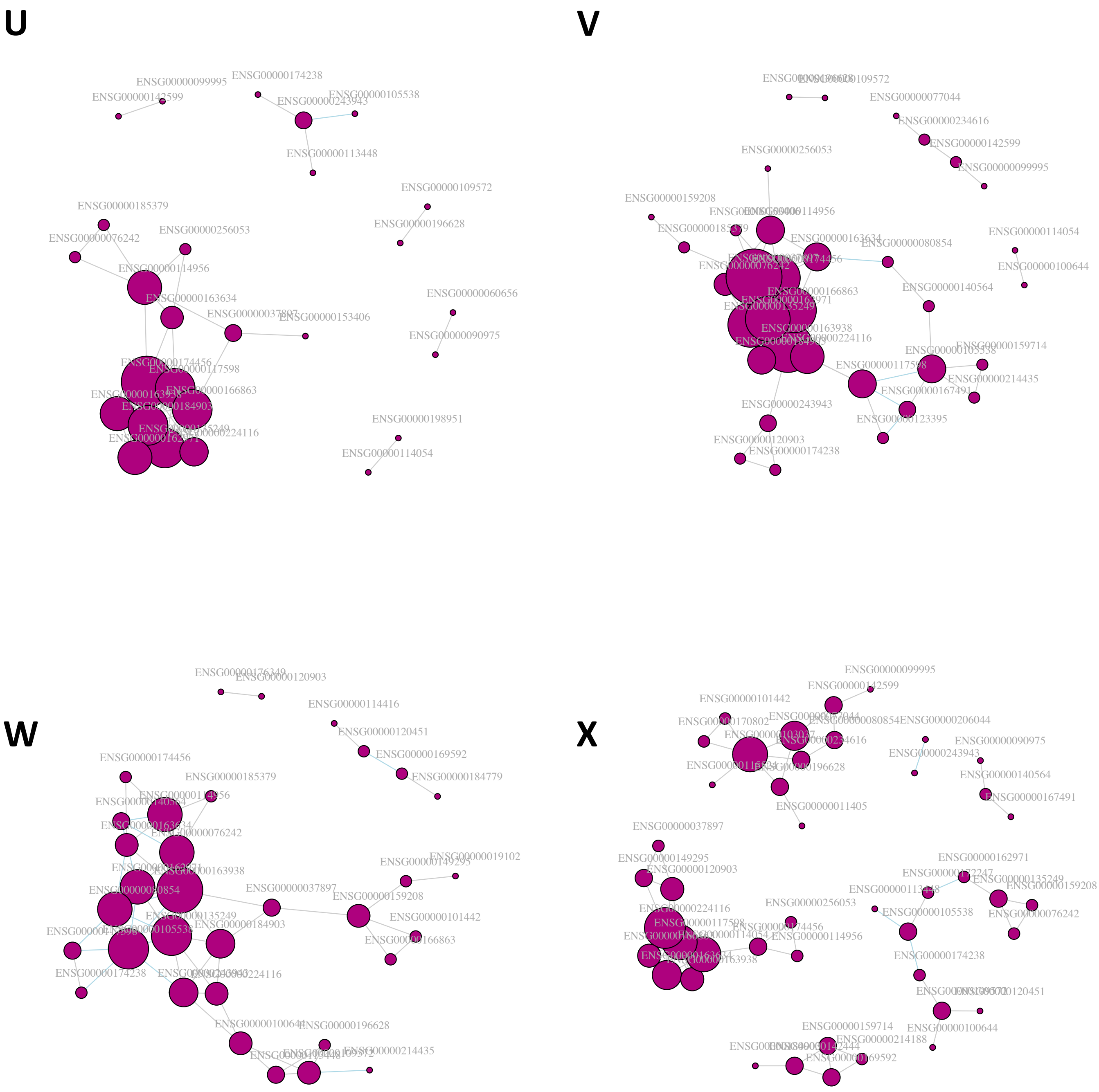
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Infant period

Q) Region 1

R) Region 2

S) Region 3

T) Region 4



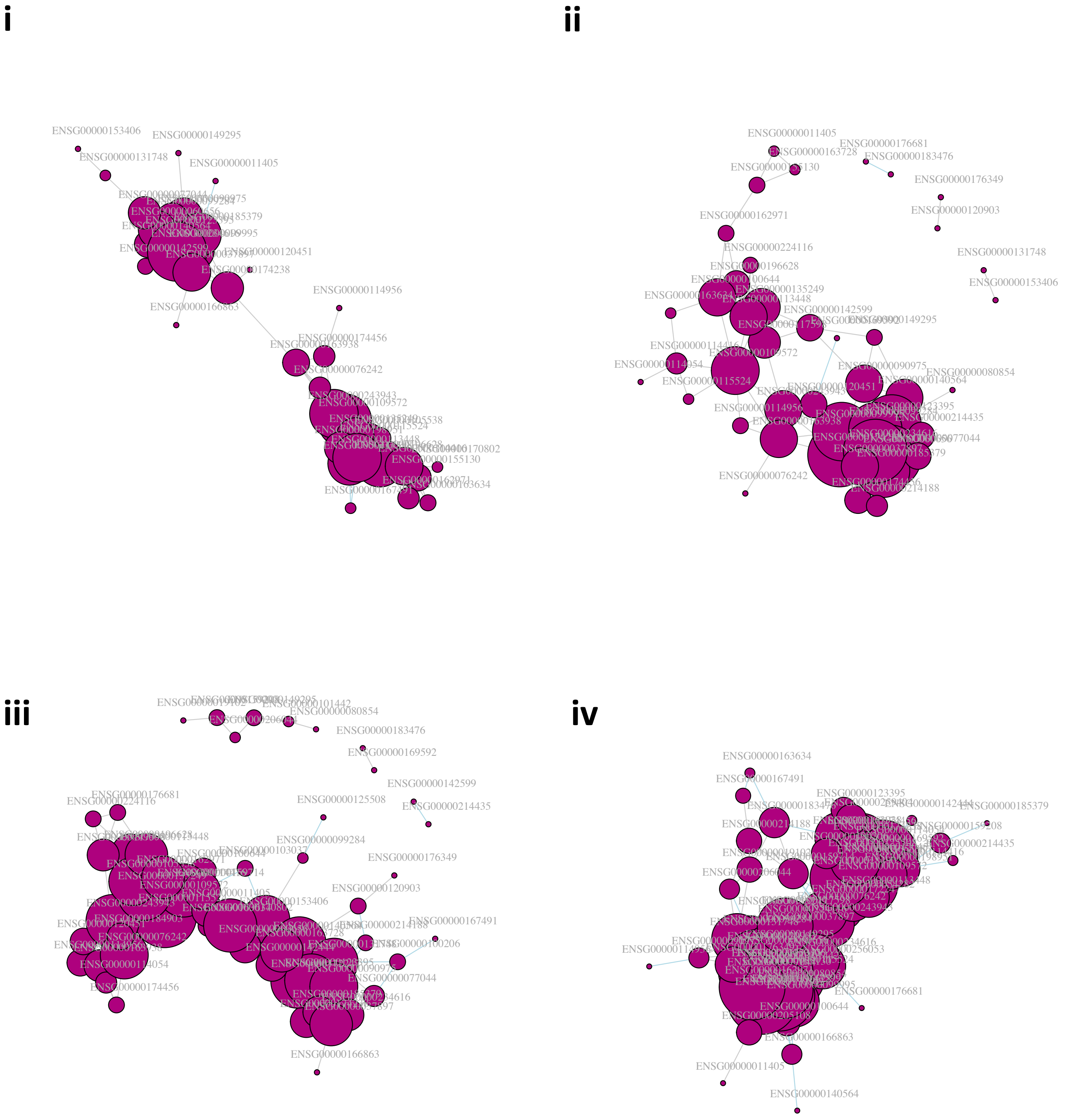
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Child.

U) Region 1

V) Region 2

W) Region 3

X) Region 4



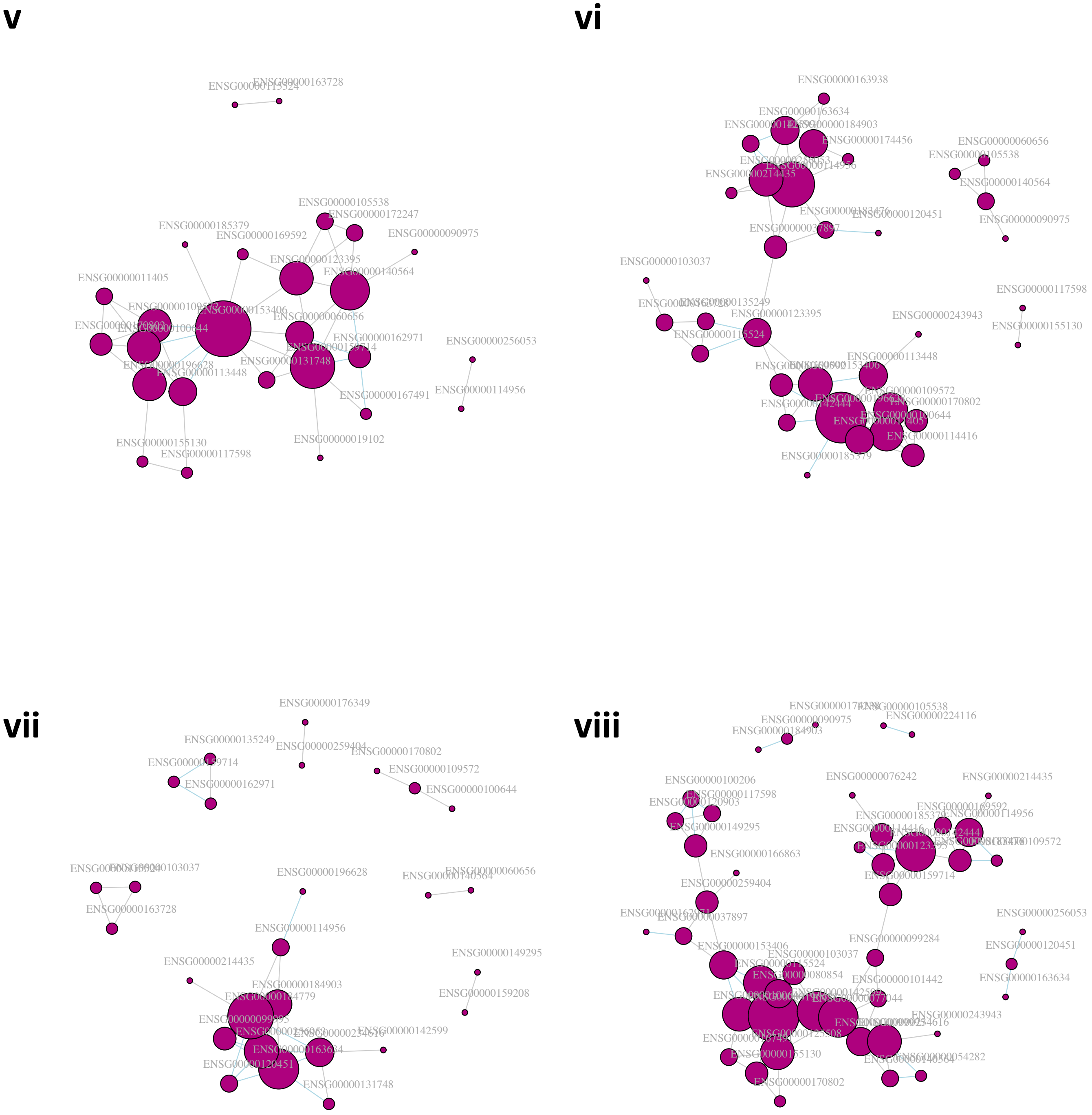
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Adolescent.

i) Region 1

ii) Region 2

iii) Region 3

iv) Region 4



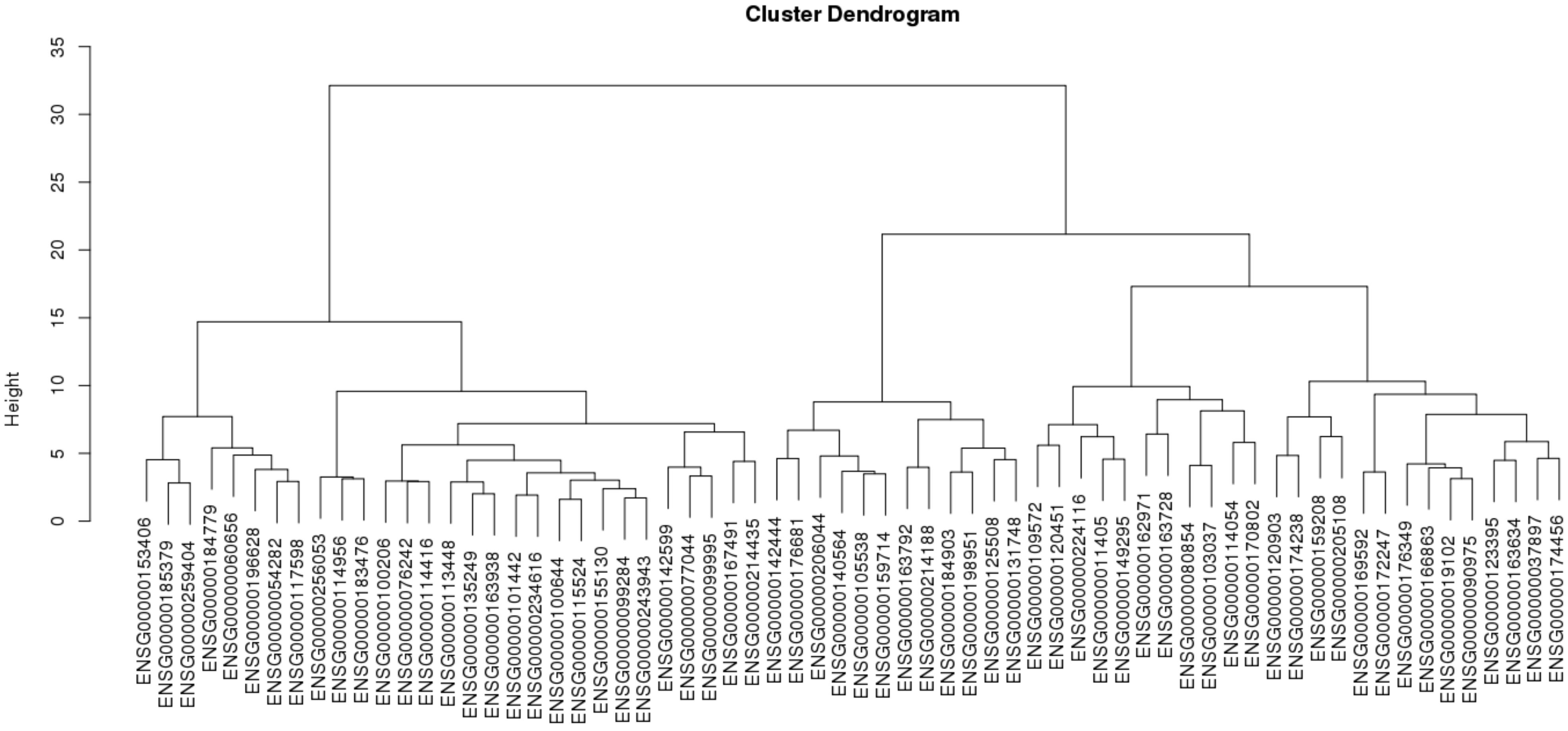
Supplementary Figure 8: Gene co-expression networks throughout development and across brain regions. Adult.

v) Region 1

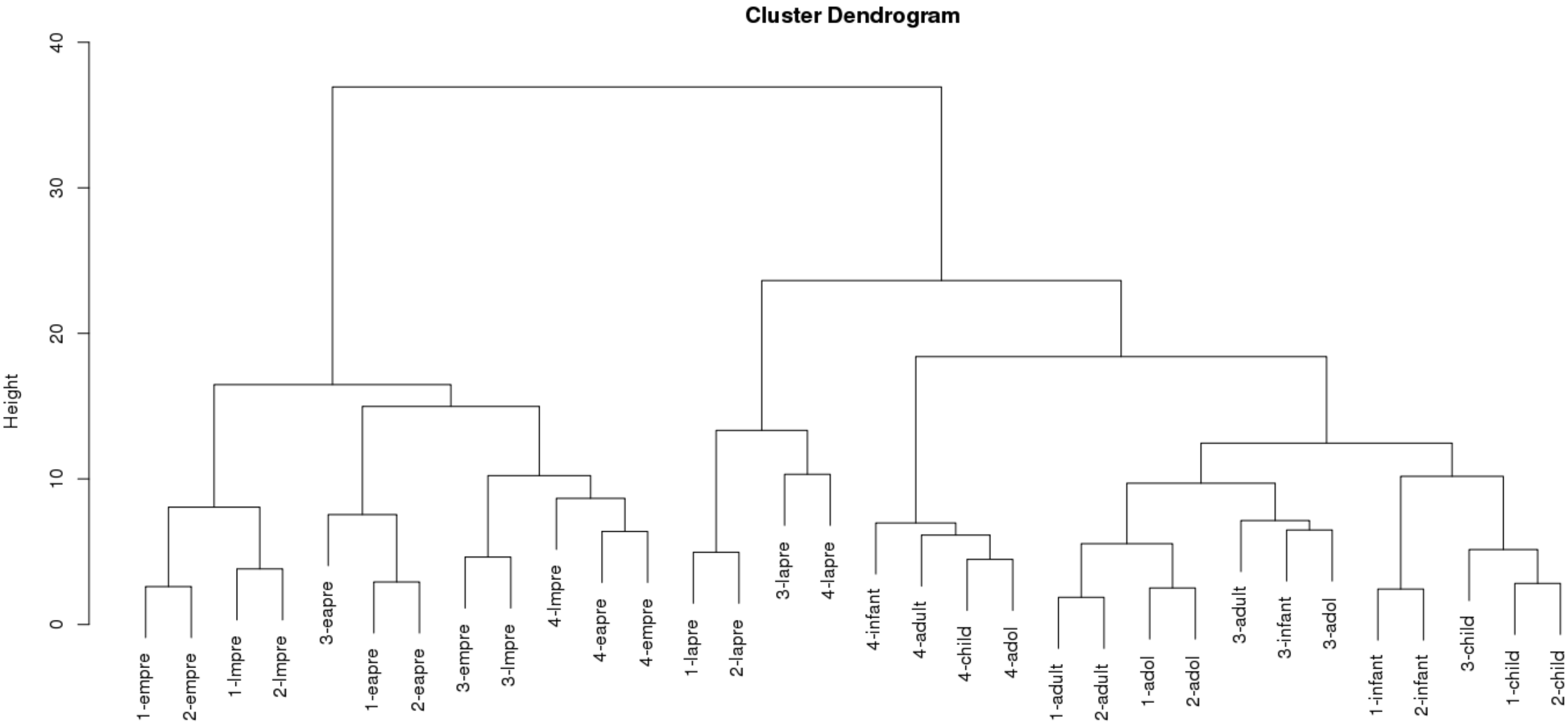
vi) Region 2

vii) Region 3

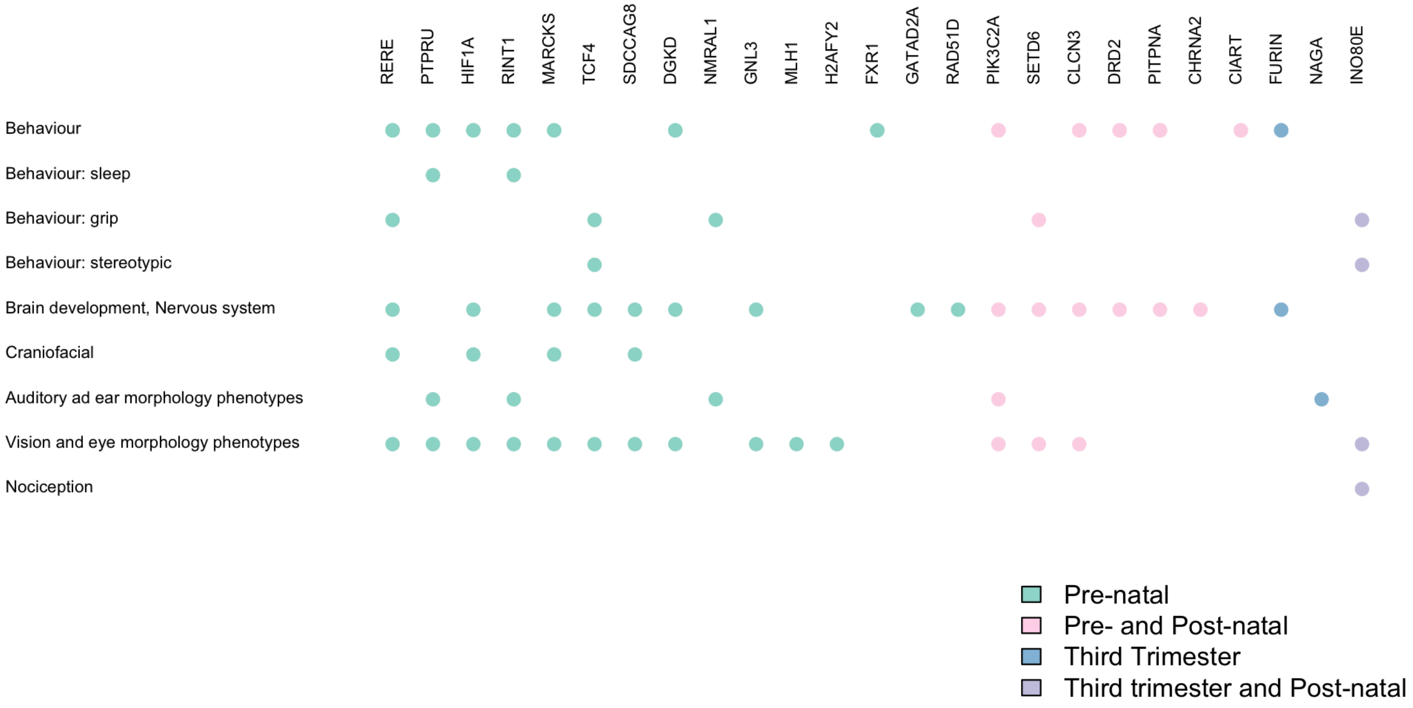
viii) Region 4



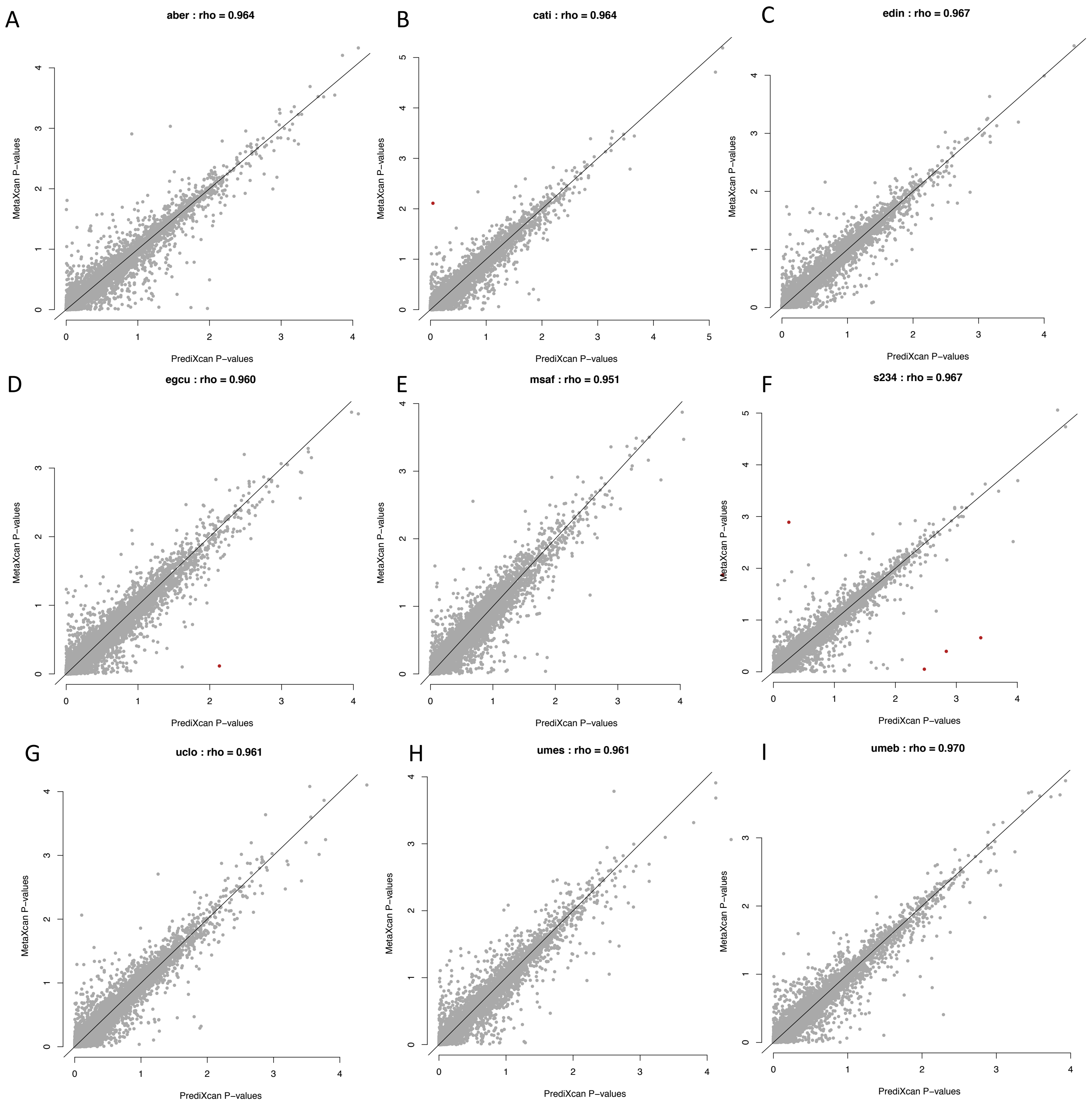
Suppl. Fig 9 A: Dendrogram of spatio-temporal clustering for 67 SCZ-associated genes.
Genes were clustered using Ward's hierachical clustering method (ward.D2) in R.
Individual clusters were defined using a cut at height 10.



Suppl. Fig 9B: Dendrogram of spatio-temporal clusters.
Spatio-temporal points were clustered using Ward's hierachical clustering method (ward.D2) in R.

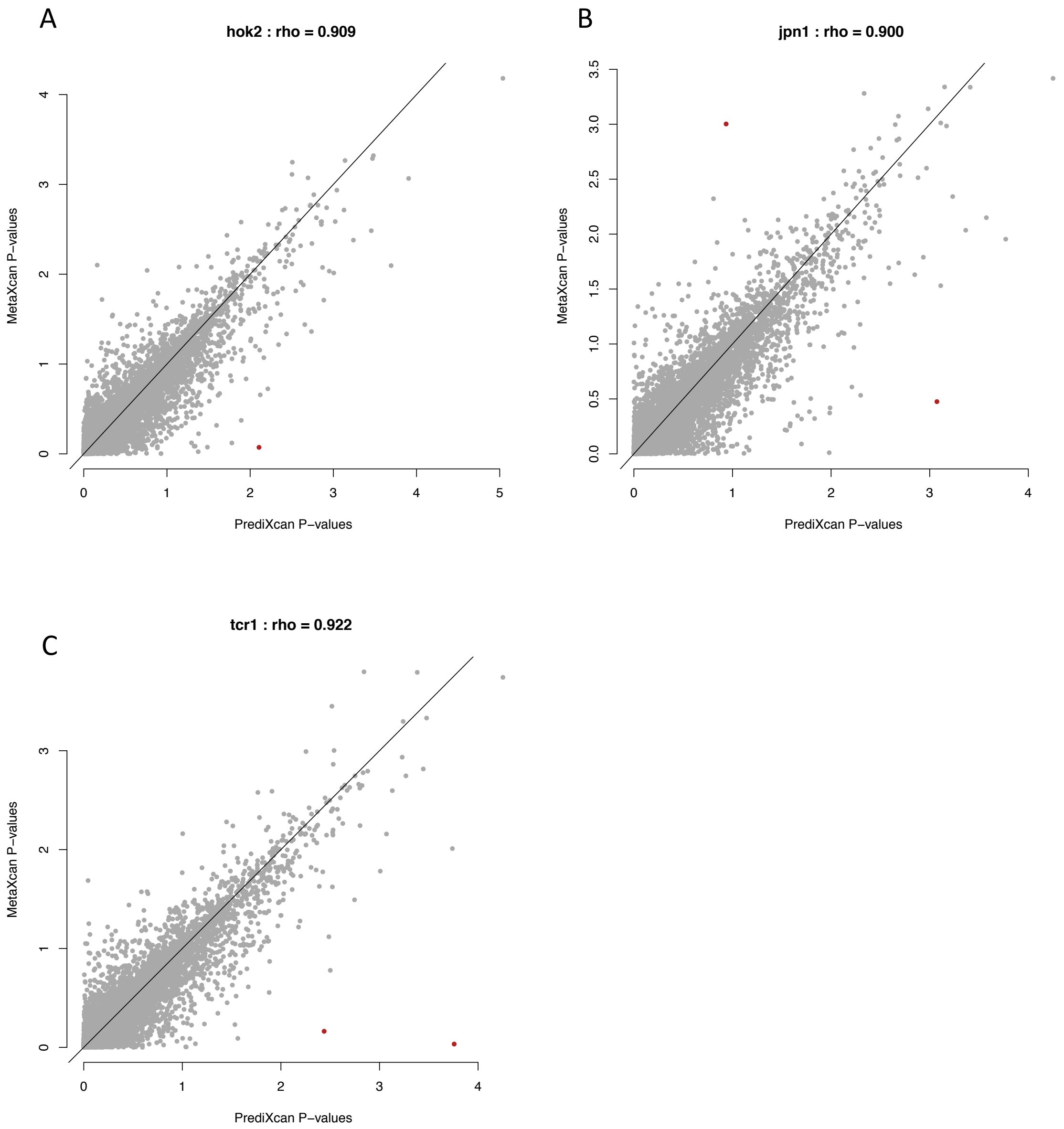


Supplementary figure 10: Specific behavioural and neurological phenotypes occurring in mouse mutant models. Mutant mouse lines are delineated according to spatio-temporal expression of a given gene in humans.



Supplementary Figure 11: Spearman correlation between PrediXcan p-values (X axis) and MetaXcan p-values (y-axis) for nine European PGC-SCZ cohorts
Discordant genes are shown in red. N genes =10,929 for all cohorts. Sample sizes for each cohort are as follows:

	A	B	C	D	E	F	G	H	I
Cohort PGC name	aber	cati	edin	egcu	msaf	s234	uclo	umes	umeb
Number of genotyped samples	1420	802	653	1417	467	4419	1016	911	960



Supplementary Figure 12: Spearman correlation between PrediXcan p-values (X axis) and MetaXcan p-values (y-axis) for three Asian PGC-SCZ cohorts

Discordant genes are shown in red. N genes for all cohorts =10,929. Number of genotyped samples in each cohort is as follows:

	A	B	C
Cohort PGC name	hok2	jpn1	tcr1
Number of genotyped samples	2495	920	1872