Regulators and Psychiatric Associates of Genomic Imprinting in the Human Brain

Attila Gulyás-Kovács*, Ifat Keydar*, ..., Andrew Chess

Mount Sinai School of Medicine

1 Supplementary Figures

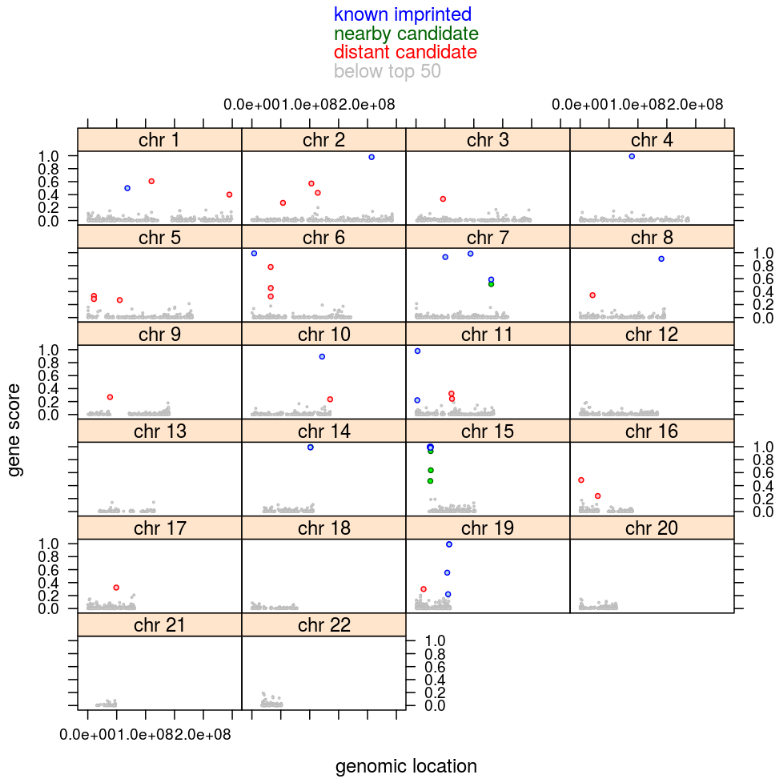

Figure S1: Clustering of top-scoring genes in the context of human DLPFC around genomic locations that had been previously described as imprinted gene clusters in other contexts.

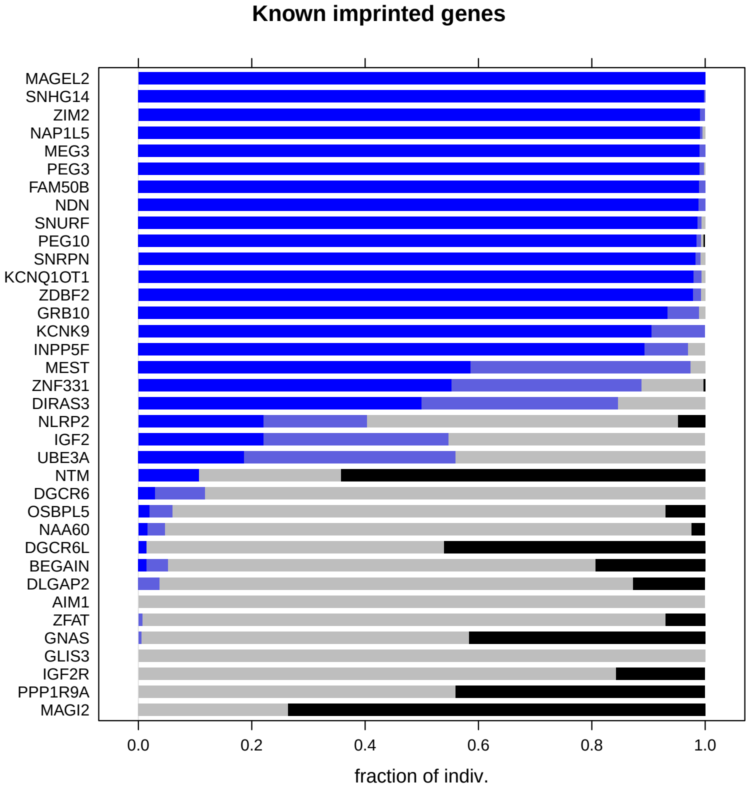

Figure S2: Known imprinted genes ranked by the gene score (dark blue bars). "Known imprinted" refers to prior studies on imprinting in the context of any tissue and developmental stage. The length of the black bars indicates the fraction of individuals passing the test of nearly unbiased expression.

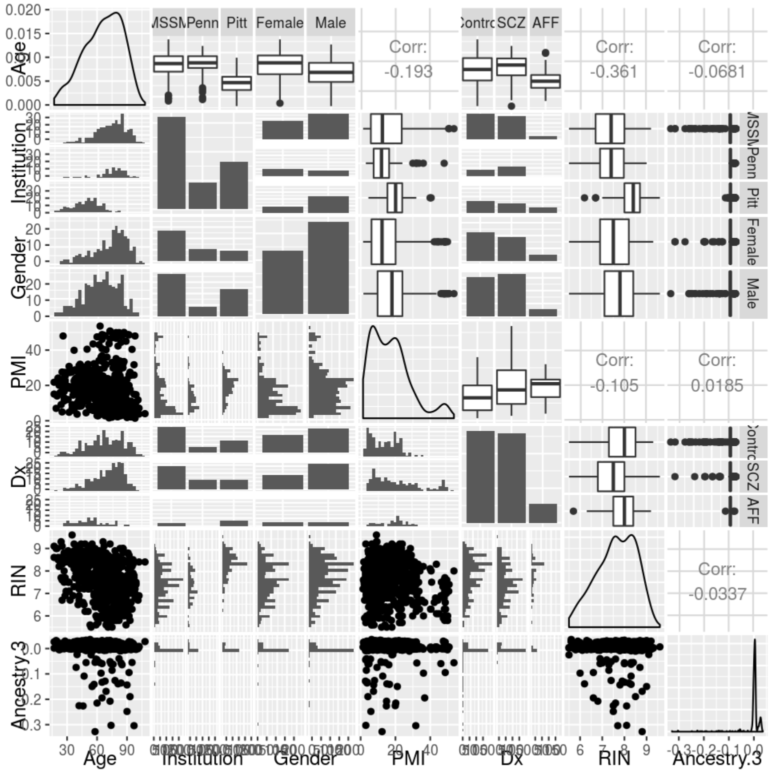

Figure S3: Pairwise dependencies among predictors.

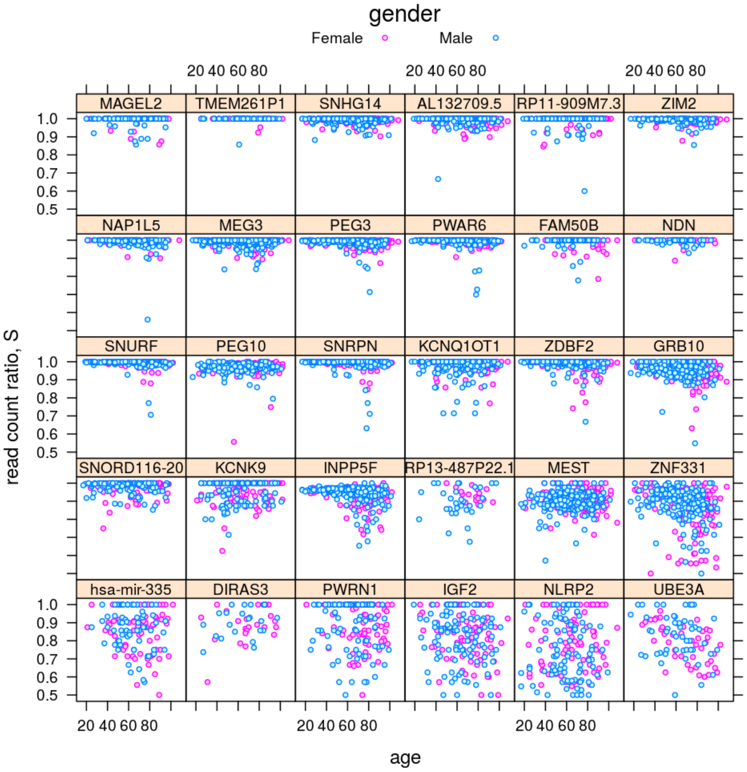

Figure S4: Variation of the read count ratio S_{ig} with age and gender across hundreds of individuals i (dots) and 30 genes g that have been considered as imprinted in the DLPFC brain area in this study.

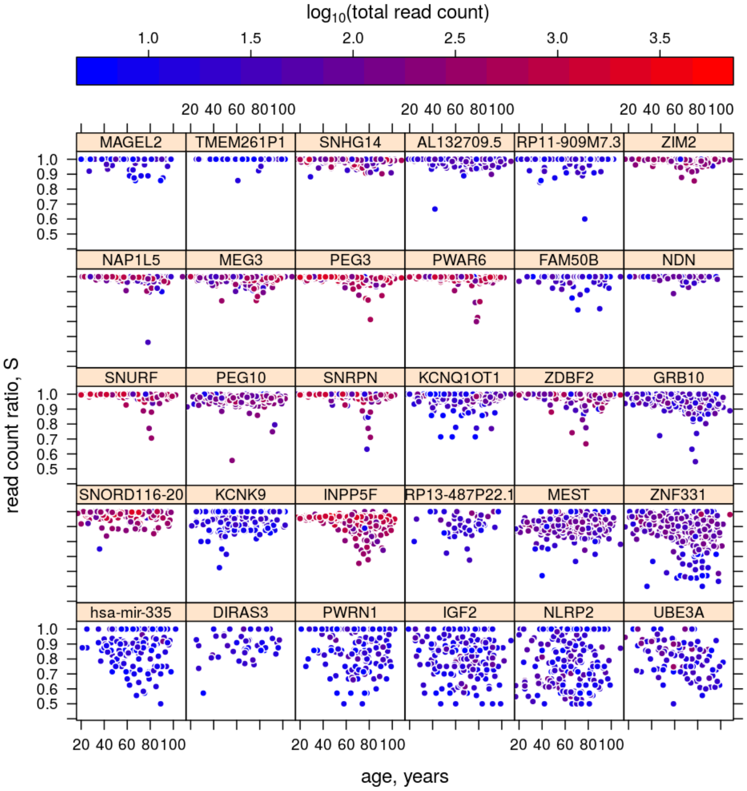

Figure S5: Variation of total RNA-seq read count across genes and individuals.

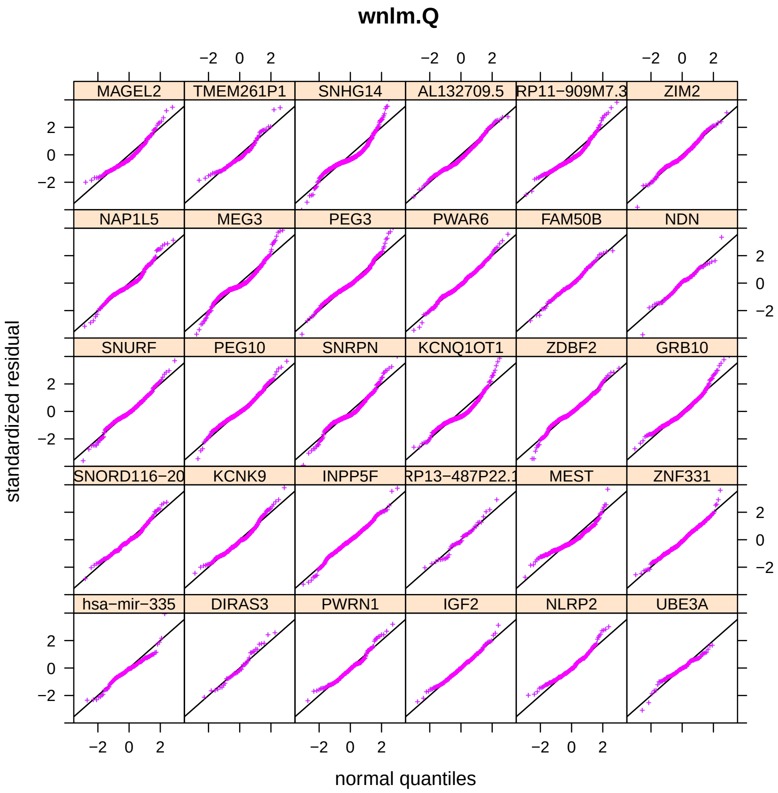

Figure S6: Checking the fit of wnlm.Q model: analysis of the normality of residuals.

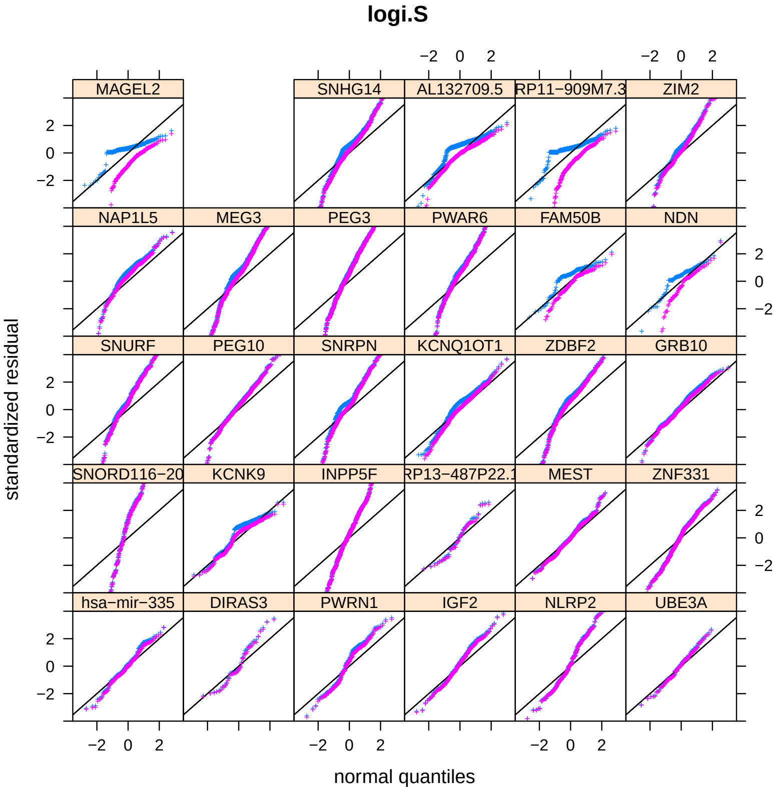

Figure S7: Checking the fit of logi.S model: analysis of the normality of standardized deviance residuals.

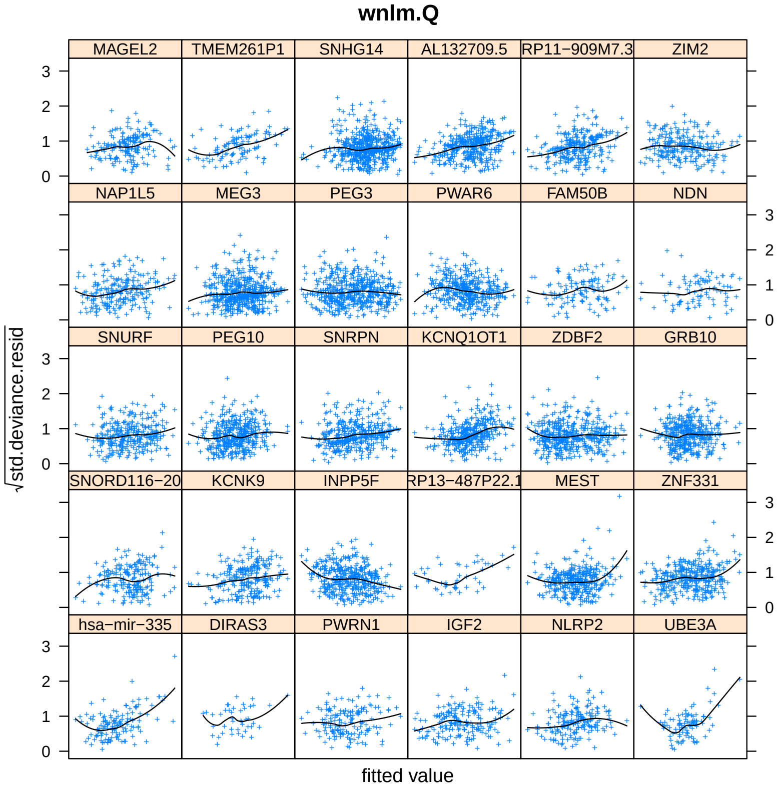

Figure S8: Checking the fit of wnlm.Q model: analysis of homoscedasticity.

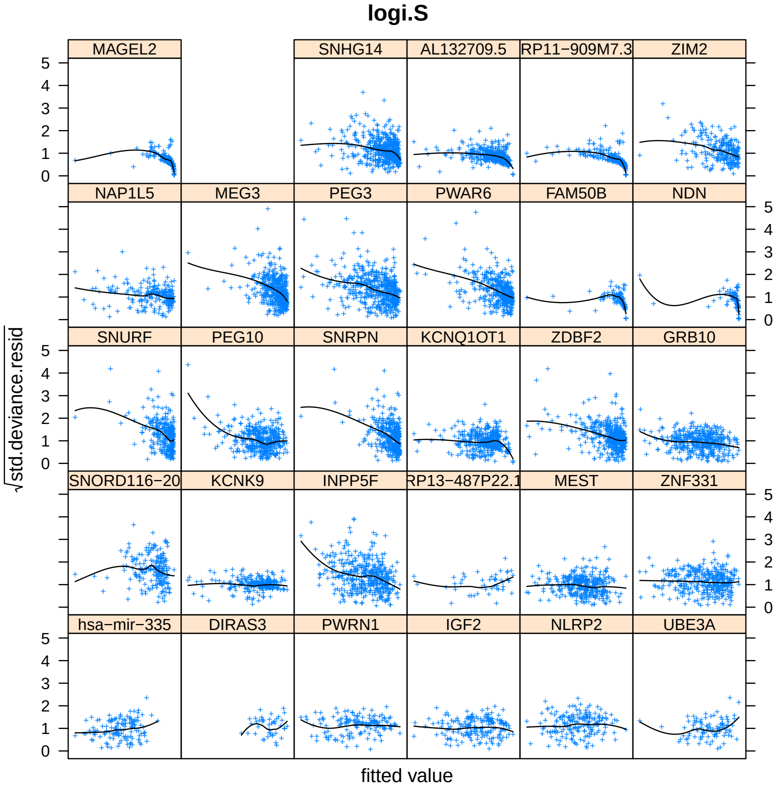

Figure S9: Checking the fit of logi.S model: analysis of homoscedasticity.

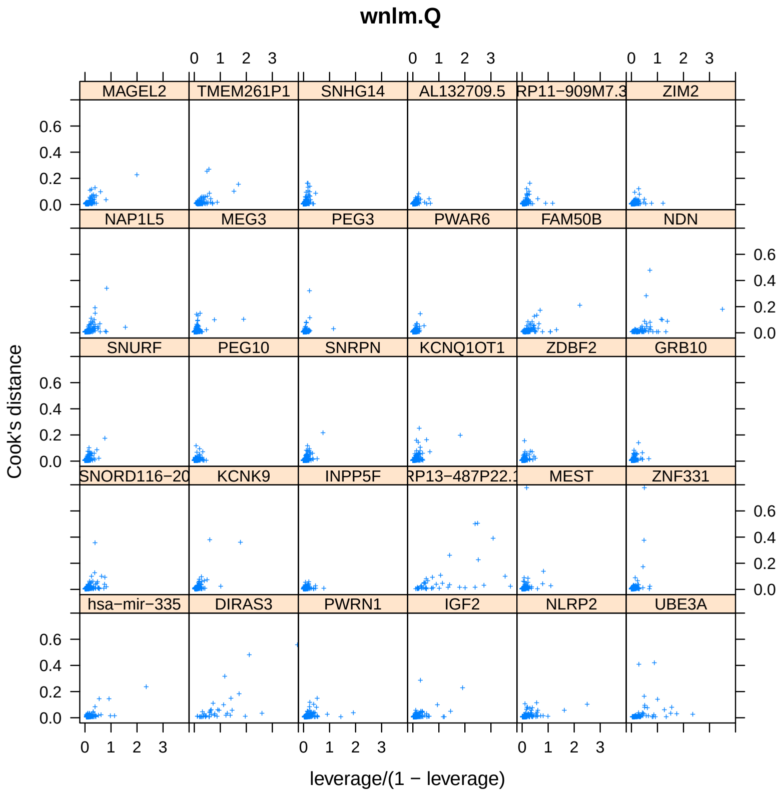

Figure S10: Checking the fit of wnlm.Q model: analysis of the impact of outliers.

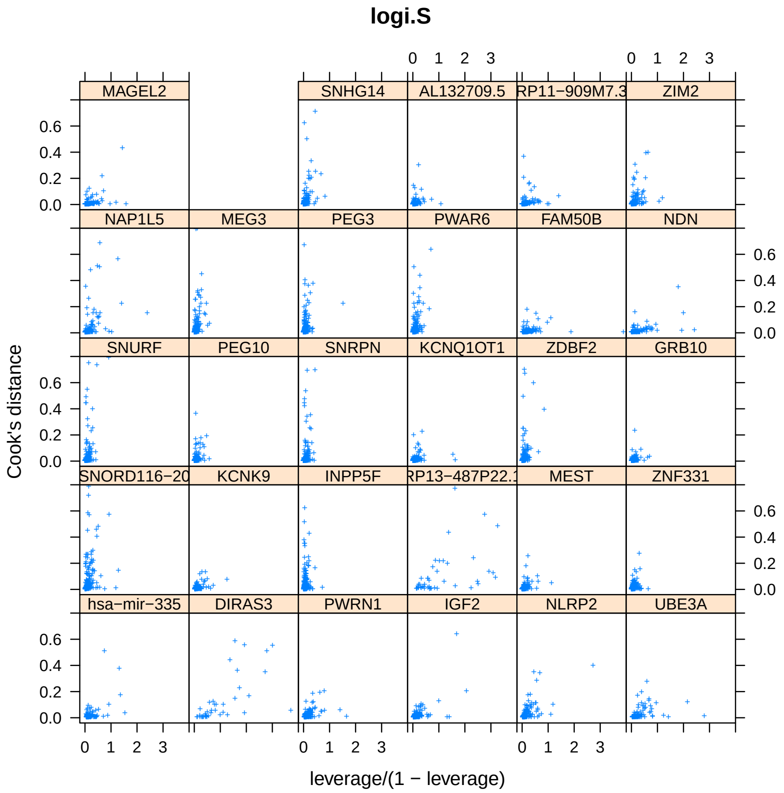

Figure S11: Checking the fit of logi.S model: analysis of the impact of outliers.

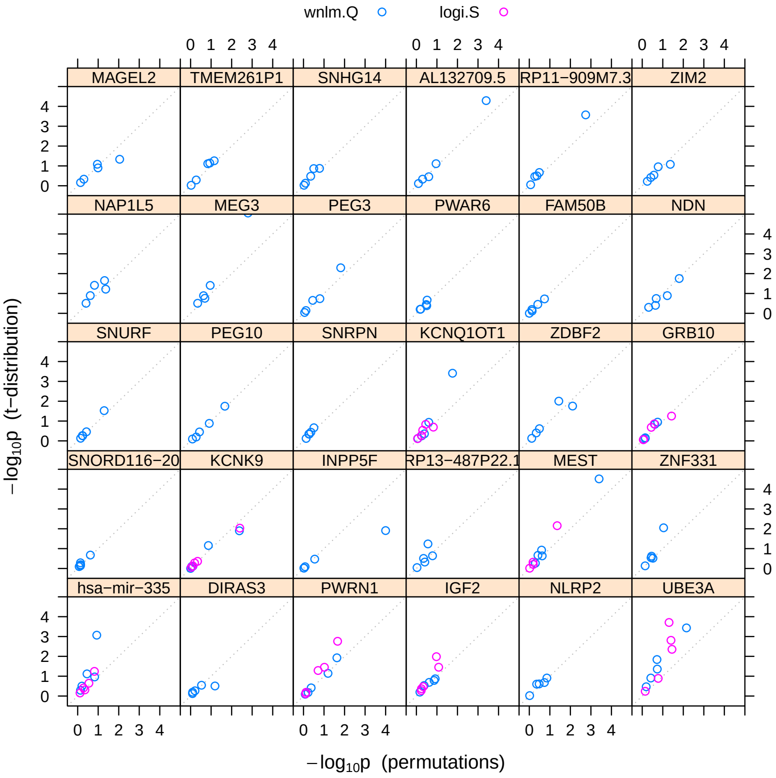

Figure S12: Agreement between a parametric (t-distribution) and non-parametric (permutations) method of estimating p-values.

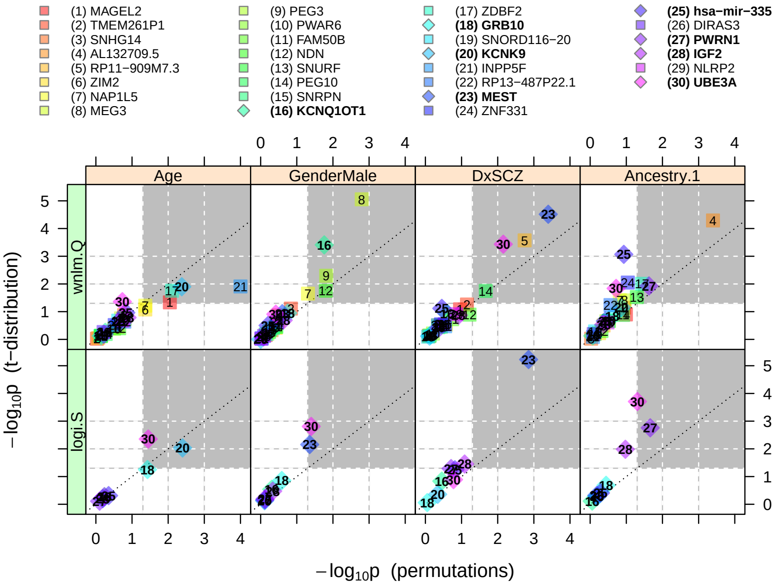

Figure S13: Significance of association between biological predictors and imprinted genes in the DLPFC calculated under wnlm.Q and logi.S. Under logi.S, only those genes are shown for which the model fit was acceptable.

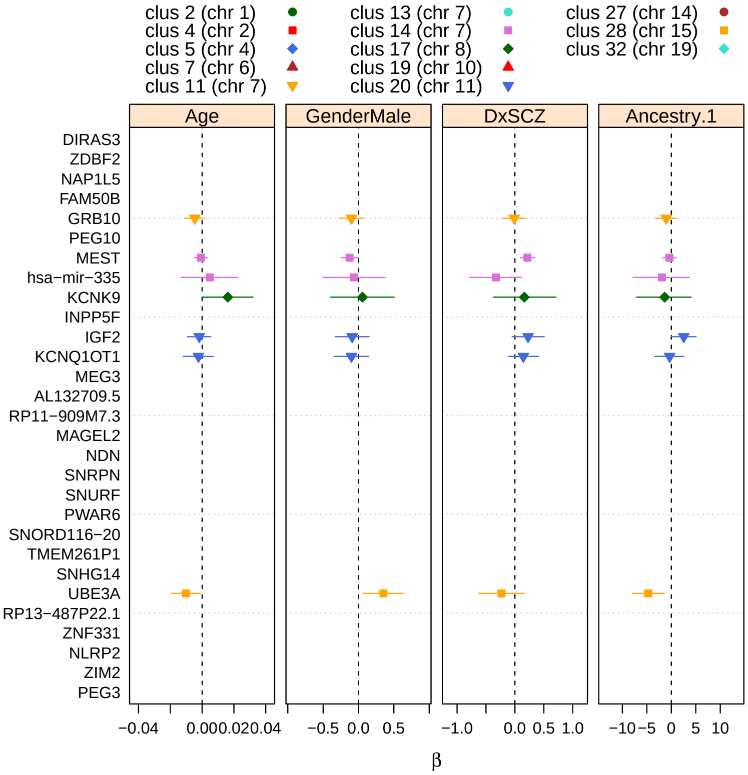

Figure S14: Estimate $\hat{\beta}_{jg}$ and 99% confidence interval of each regression coefficient β_{jg} under the logi.S model, where j is a predictor/term and g is a gene. $\hat{\beta}_{jg}$ and the confidence interval are shown only for those genes are shown for which the model fit was acceptable.

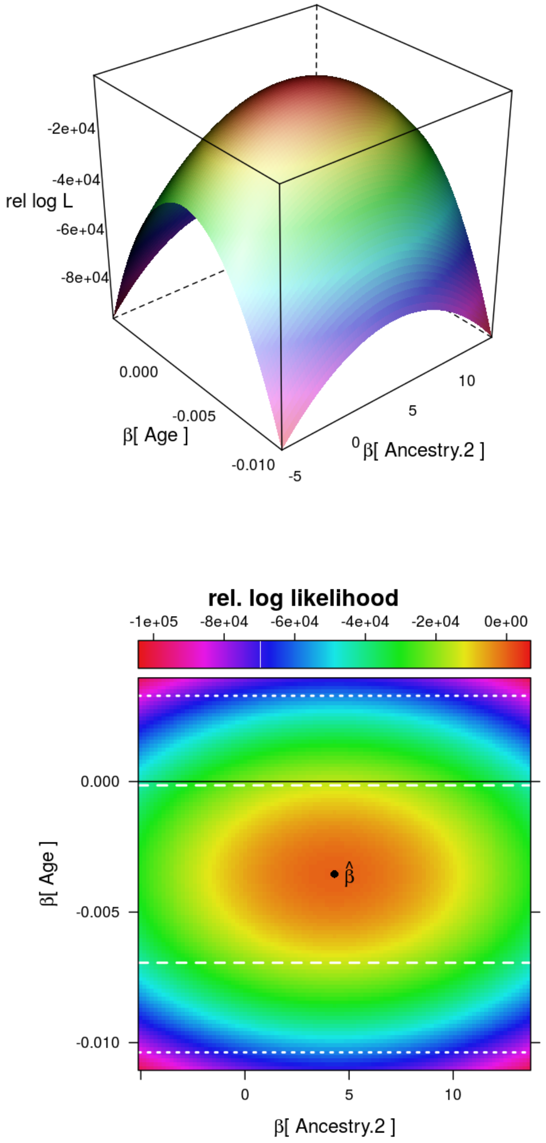

Figure S15: *Top* and *bottom*: two representations of the relative log likelihood on a 2 dimensional section of the $p > 20$ dimensional parameter space given the wnlm.Q model and data for the gene $g =$ PEG3. The section was taken by fixing all but two parameters at their estimates: $\beta_{jg} = \hat{\beta}_{jg}$. A rectangular subspace for these two parameters, $\beta_{\text{Age},g}$ and $\beta_{\text{Ancestry.2},g}$, was chosen around the maximum likelihood estimate $\hat{\beta} = (\hat{\beta}_{\text{Age},g}, \hat{\beta}_{\text{Ancestry.2},g})$. The nearly parabolic shape of the log-likelihood function suggests that the regularity conditions for likelihood-based parametric inference are fulfilled.

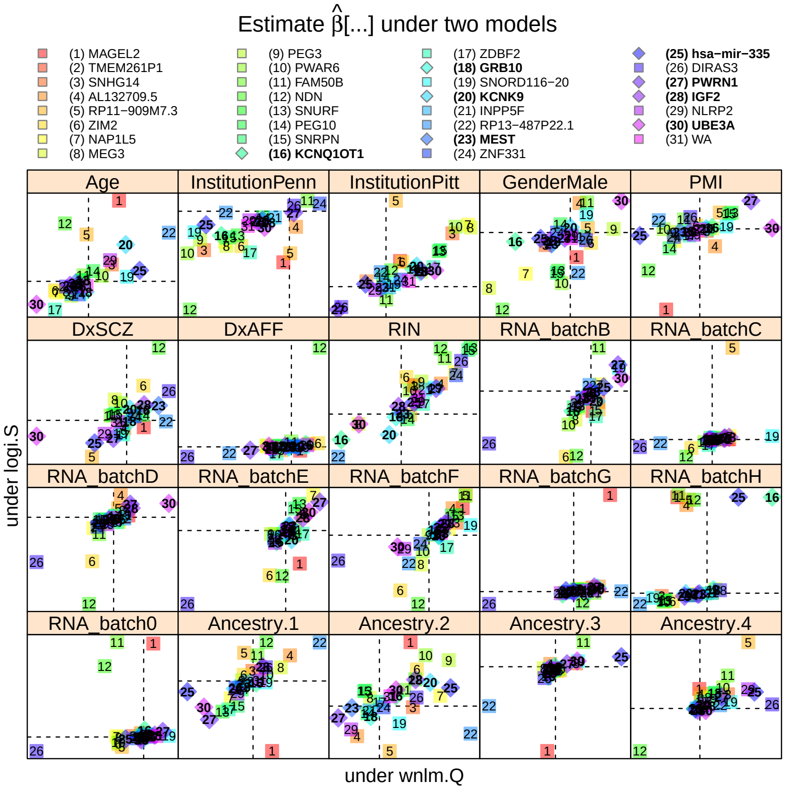

Figure S16: Agreement of the wnlm.Q and logi.S models in terms of estimated regression coefficients.

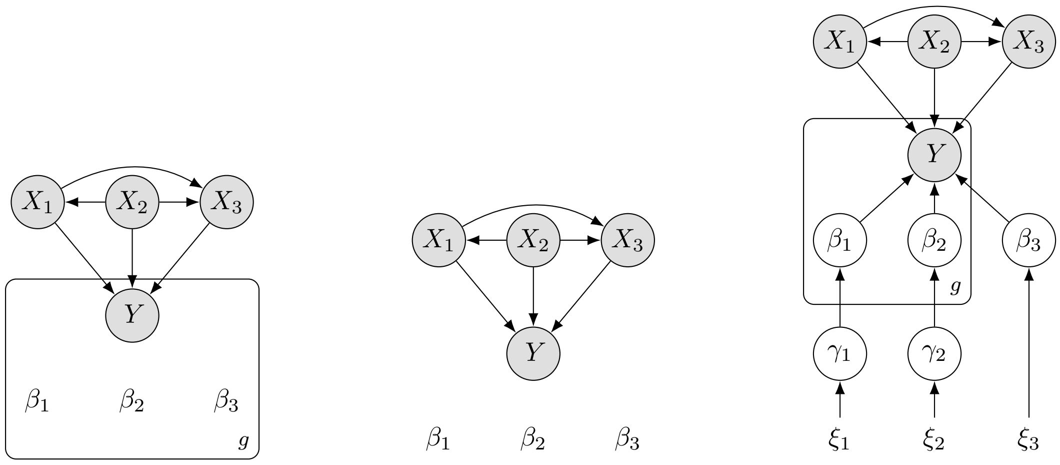

Figure S17: General dependency structure of three regression model frameworks. In all of threse model frameworks the regression coefficients $\beta_{1g}, ..., \beta_{3g}$ mediate, for a given gene g, probabilistic dependencies (arrows) between the response variable Y_g (read count ratio for g) and the corresponding predictors $X_1, ..., X_3$. For simplicity but without loss of generality only 3 predictors are depicted. The model frameworks differ in how $\beta_{jg_1}, \beta_{jg_2}, ...$ relate to each other for a given predictor (or a given j). *Left:* there is no connection among $\beta_{jg_1}, \beta_{jg_2}, ...$ which means that the way Y_g, the read count ratio for gene g depends on predictor X_j is completely separate from how the read count ratio for any other gene g' (i.e. $Y_{g'}$) depends on it. Consequently no information may be shared among gene-specific models. *Middle:* In this case $\beta_{jg_1} = \beta_{jg_2} = ... \equiv \beta_j$ so that all genes are identical with respect to how their read count ratio depends the predictors. Thus genes share all information in the data in the sense that the model forces them to be identical. *Right:* Hierarchical Bayesian model where genes show both variation as well as invariance with regards to depenencies. The variation is described by the dependence of β_j on the hyperparameter γ_j, whereas the invariance by the dependence of γ_j on *its* hyperparameter ξ_j. Only this model framework allows information sharing among genes in a flexible way.

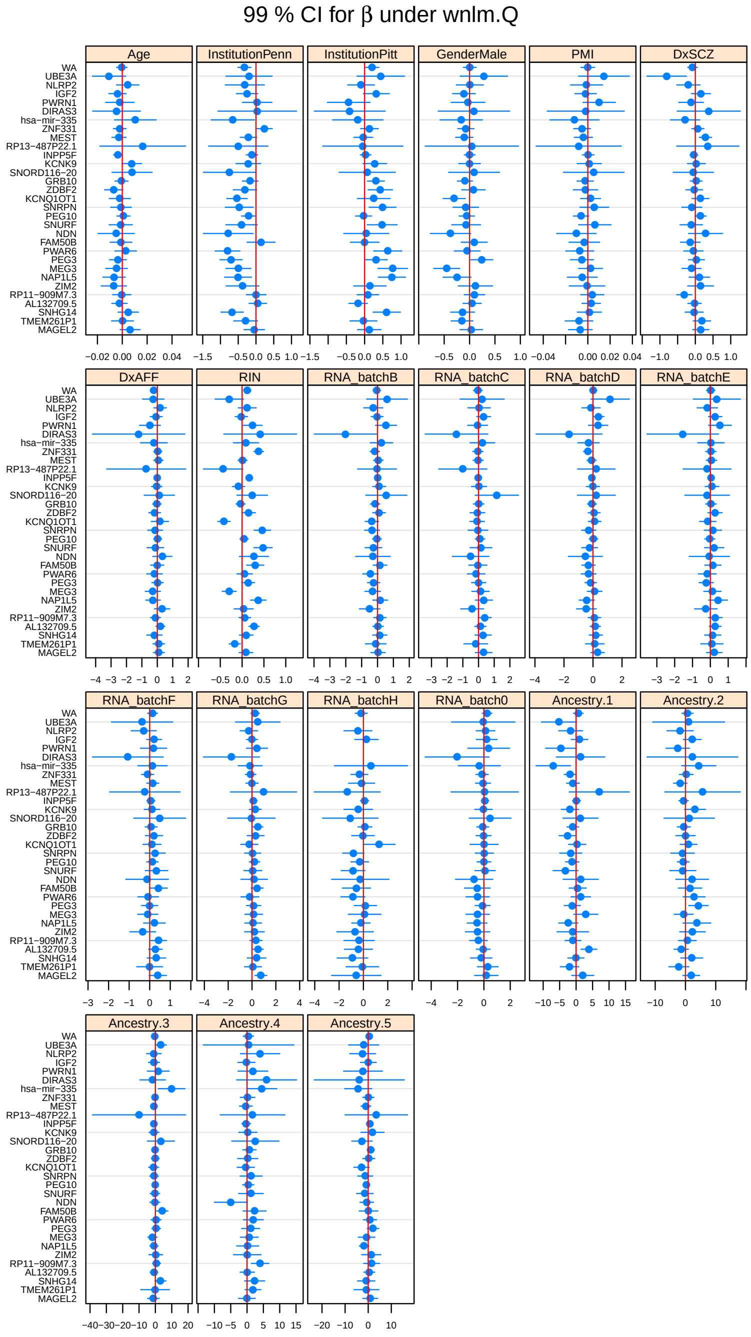

Figure S18: Estimates $\hat{\beta}_{jg}$ and confidence intervals for regression coefficients under the wnlm.Q model concerning for all predictors.

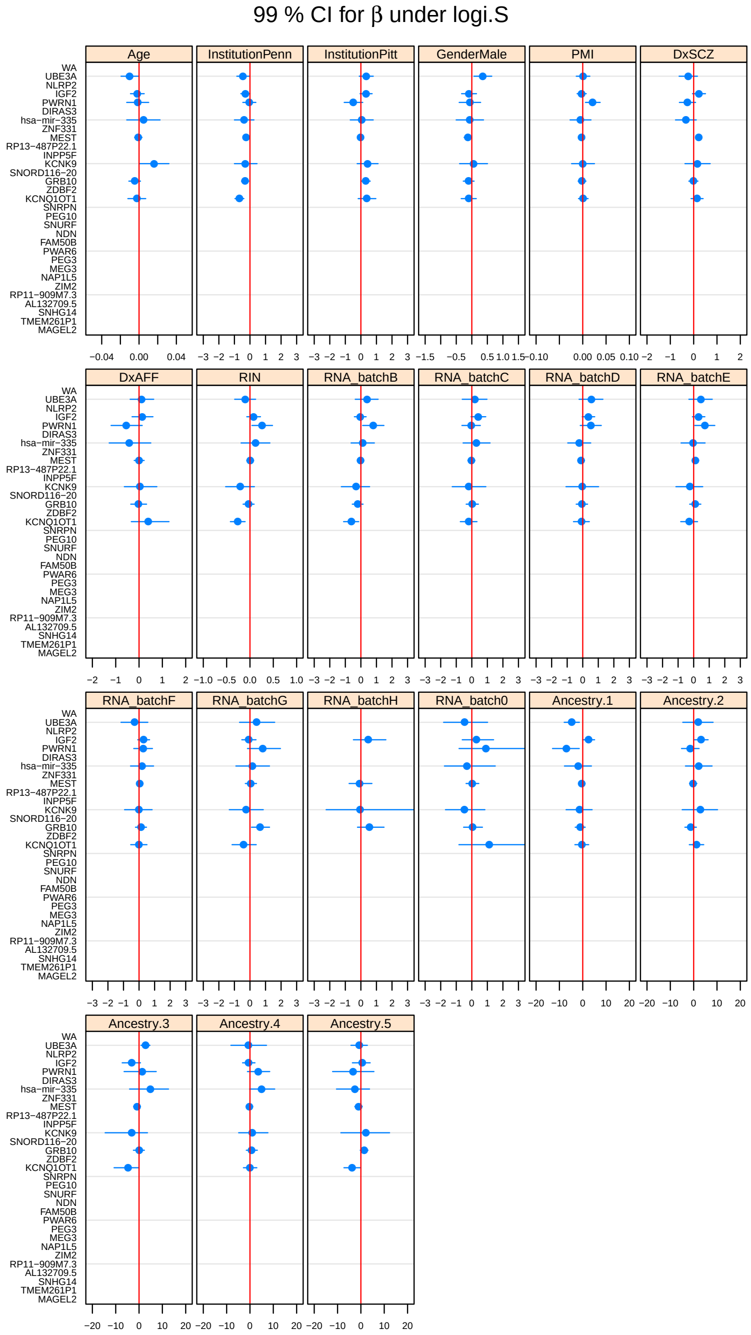

Figure S19: Estimates $\hat{\beta}_{jg}$ and confidence intervals for regression coefficients under the logi.S model concerning for all predictors. Gaps for certain genes indicate unacceptable fit.

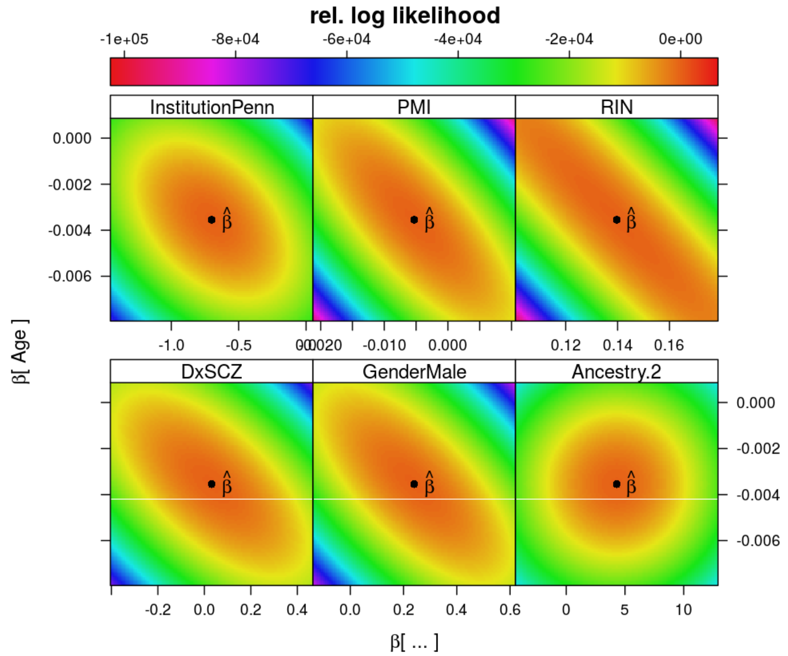

Figure S20: Analysis of orthogonality of regression coefficients. Relative log-likelihood surfaces for the gene PEG3 on various rectangles corresponding to 2D sections through the $p > 20$ dimentional parameter space. The set of points in a rectangle where log-likelihood takes the same value are quasi-ellipses, whose major and minor axes and their tiltedness express association between coefficients. For instance, β_{Age} is not assiciated with (orthogonal to) $\beta_{\mathrm{Ancestry.2}}$ but is strongly associated with β_{RIN}. Such association hinders statistical inference due to the following circularity: If we knew the precise value of β_{RIN} we could estimate the true value of β_{Age} with higher precision and confidence but the precise value of β_{RIN} could only be obtained with high confidence if we precisely knew β_{Age}.

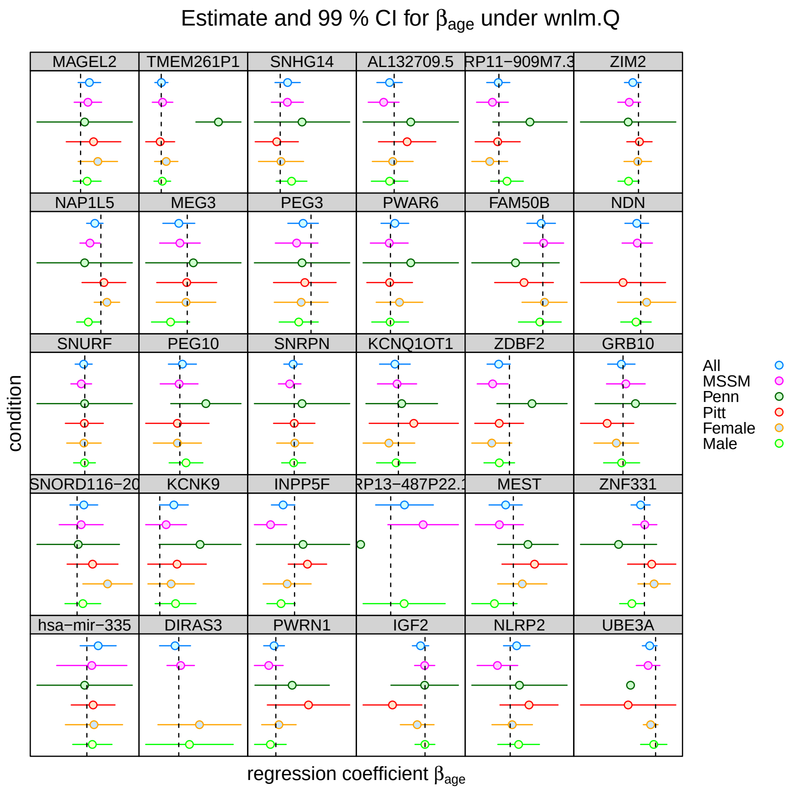

Figure S21: Analysis of interactions among predictors under the wnlm.Q model. Contextual depenendence of the read count ratio on age, where the context is given by some specific level of Institution (MSSM, Penn, Pitt) or Gender (Female, Male).

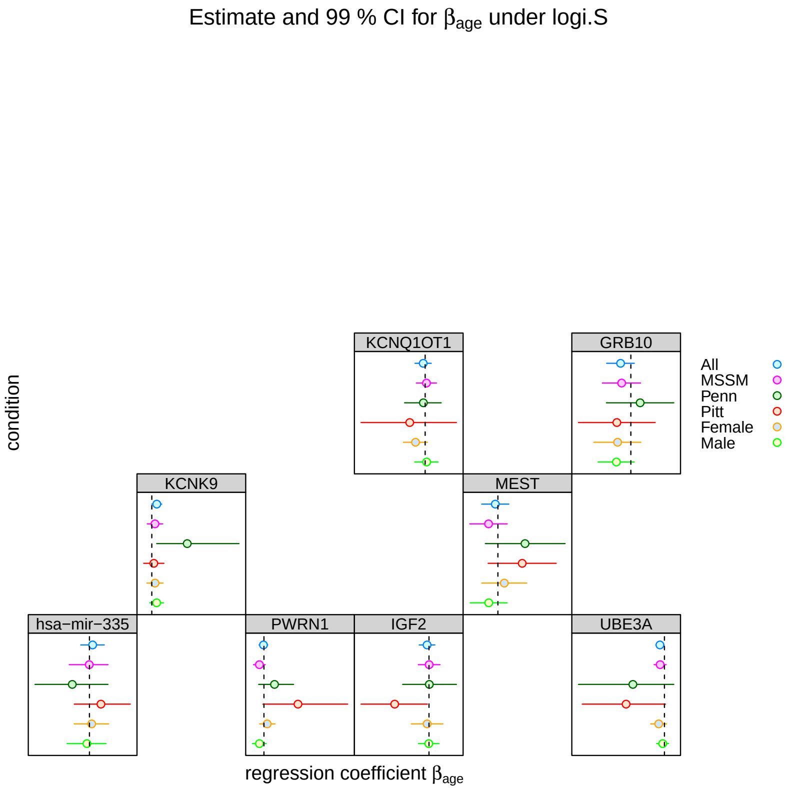

Figure S22: Analysis of interactions among predictors under the logi.S model. The missing panels correspond to genes for which logi.S did not provide acceptable fit.

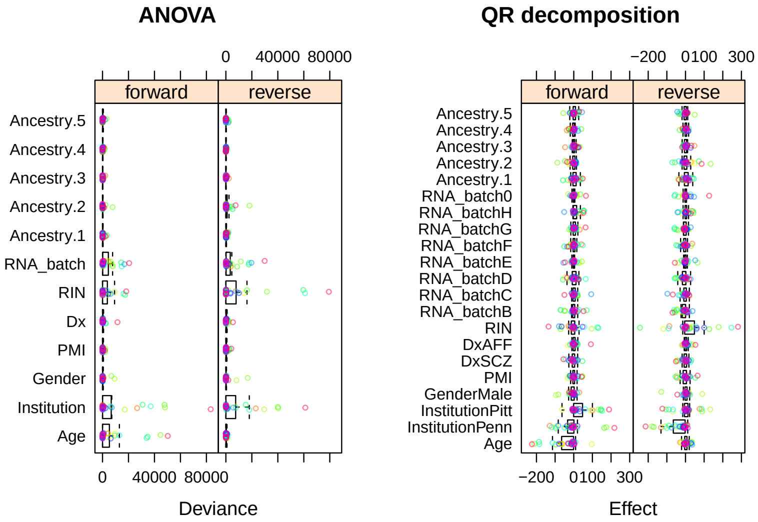

Figure S23: *Left:* analysis of variance (ANOVA) is undermined by the non-orthogonality of predictors because the reduction in deviance (i.e. in residual sum of squares) for a term (predictor) depends on the sequence in which terms are added to the model. *Right:* the same concept is demonstrated using QR decomposition.

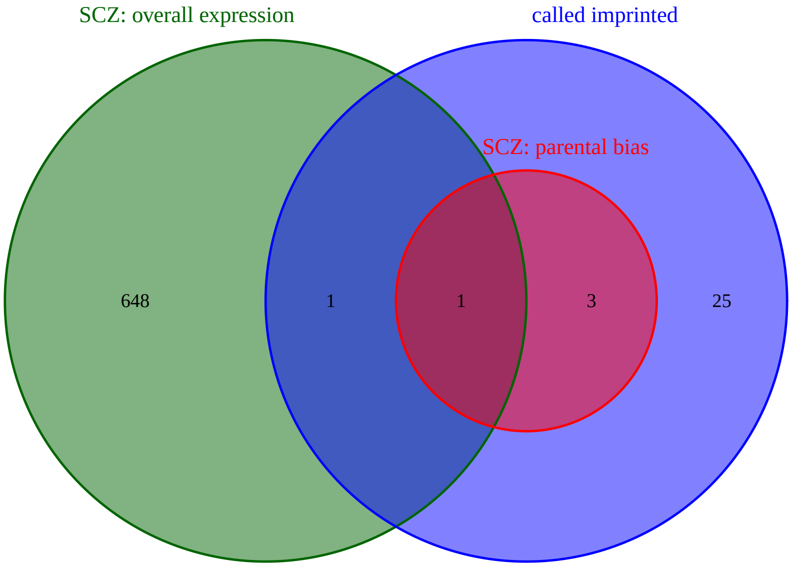

Figure S24: Association of genes' expression to schizophrenia (SCZ) assayed by two RNA-seq based approaches: total read count (overall expression, Nat Neurosci. 2016 Nov;19(11):1442-1453.) and read count ratio (allelic bias, present work). When these approaches are compared for only those genes that we find imprinted in the DLPFC in this study, 1 gene is found associated to schizophrenia by both approaches, 1 by only overall expression, and 3 by only allelic bias.