Variation of Genomic Imprinting in the Human Population, and Analysis of its Sources and Psychiatric Consequences

Attila Gulyás-Kovács*, Ifat Keydar*,
Eva Xia, Menachem Fromer, Doug Ruderfer,
Ravi Sachinanandam, Andrew Chess

Mount Sinai School of Medicine

TODOs

1. schematic figure on study design (Fig. 1)
2. finish the rest of the Results section

Abstract

Lorem ipsum...

1 Introduction

Variation of expression level across genes, lifetime, cells, tissue types, as well as individuals in the human population is clearly a major determinant of phenotype REF. A distinct, though related, question is the role of the *ratio* between maternal (or paternal) transcripts and those from both alleles of a given gene. Besides the random thermal fluctuations, which occur in all genes, some genes display systematic *parental bias* in allelic expression. The extreme form of this bias is known as monoallelic expression, which may be non-random if expression is biased towards only one parent or random otherwise [1].

Genomic imprinting is the classic, non-random, case of monoallelic expression, which is mediated by epigenetic mechanisms that either concern single genes or entire imprinted gene clusters [6, 7]. Having emerged late in evolution [8], imprinting of several genes have been implicated in biological functions of mainly neurological, psychological, behavioral and social type. Mutations disrupting the expression of certain imprinted genes are deeply penetrant causes of rare syndromes displaying psychological, cognitive, and social dysfunction. But it remains to be established to what extent more subtle perturbations of parental bias might contribute to common, highly polygenic psychiatric disorders such as schizophrenia or autism.

Also of interest is how and why parental bias for a given imprinted gene might vary between individuals. Age is a prime candidate for regulator given that several imprinted genes have been found to be functionally associated to either perinatal stage (e.g. suckling) or young adulthood (e.g. maternal care). Gender and ancestry are obvious further candidates.

A modern approach to study the variation of parental bias is to differentiate maternal and paternal transcripts based on those single nucleotide polymorphisms (SNPs) that contribute to the heterozygosity of a given individual and gene. Coupled with high throughput techniques such as RNA-seq this approach permits the quantification of both genome-wide and among-individual variation. Several research groups [3, 5, 2] combined this approach with F1 hybrids of crossed mouse strains and found, for instance, that $\approx 1\%$ of all genes are imprinted, although some [5] of the same researchers previously estimated $> 5\%$, leaving this question controversial. Another finding is the differential effect of age, but not gender, on parental bias: shifting from neonatal age to young adulthood down-regulated bias in ca. 20% of detected imprinted genes, up-regulated in 6% and had no effect on the remaining nearly 75%.

While such designed mouse experiments afford high statistical power their relevance is at best unclear to human neuropsychological function, ageing and ancestry. The present work directly addresses these points by utilizing post-mortem tissue samples from the dorsolateral prefrontal cortex (DLFPC) from nearly 600 individuals of different age, psychiatric condition, and ancestry, and by performing the RNA-seq based quantification of parental bias.

2 Methods

2.1 Brain samples

[1]

Human RNA samples were collected from the dorsolateral prefrontal cortex of the CommonMind consortium (CMC), from a total of 579 individuals after quality control. Subjects included 267 control individuals, as well as 258 with schizophrenia (SCZ) and 54 with affective spectrum disorder (AFF). RNA-seq library preparation uses Ribo-Zero (which selects against ribosomal RNA) to prepare the RNA, followed by Illumina paired end library generation. RNAseq was performed on Illumina HiSeq 2000.

2.2 RNA-seq, mapping and SNP calling

We mapped 100bp, paired-end reads ($\approx$ 50 million reads per sample) using Tophat to Ensembl gene transcripts of the human genome (hg19; February, 2009) using default parameters with 6 mismatches allowed per pair (200bp total). We required both reads in a pair to be successfully mapped and we removed reads that mapped to > 1 genomic locus. Then, we removed PCR replicates using the Samtools rmdup utility; around one third of the reads mapped (which is expected, given the parameters we used and the known high repeat content of the human genome). We used Cufflinks to determine gene expression of Ensembl genes, using default parameters. Using the BCFtools utility of Samtools, we called SNPs (SNVs only, no indels). Then, we invoked a quality filter requiring a Phred score > 20 (corresponding to a probability for an incorrect SNP call < 0.01).

We annotated known SNPs using dbSNP (dbSNP 138, October 2013). Considering all 579 samples, we find 936,193 SNPs in total, 563,427 (60%) of which are novel. Further filtering of this SNP list removed the novel SNPs and removed SNPs that either did not match the alleles reported in dbSNP or had more than 2 alleles in dbSNP. We also removed SNPs without at least 10 mapped reads in at least one sample. Read depth was measured using the Samtools Pileup utility. After these filters were applied, 364,509 SNPs remained in 22,254 genes. These filters enabled use of data with low coverage, as described below. For the 579 samples there are 203 million data points (reads overlapping one of the 364,509 SNPs defined above), of which 158 million (78%) have genotype data available (array or imputation), which is used later in the pipeline.

2.3 Genotyping and calibration of imputed SNPs

DNA samples were genotyped using the Illumina Infinium SNP array. We used PLINK with default parameters to impute genotypes for SNPs not present on the Infinium SNP array using 1000 genomes data. To maximize the number of genes assessable for monoallelic expression, while minimizing false positive monoallelic expression calls which can arise if the underlying SNP has been incorrectly called as heterozygous by the imputation, we calibrated the imputation parameters.

We first examined how many SNPs were heterozygous in DNA calls and had a discordant RNA call (i.e. homozygous RNA-SNP call) using different imputation parameters. Known imprinted genes were excluded. We examined RNA-seq reads overlapping array-called heterozygous SNPs

[1] This and the following two subsections have been taken apart from a few minor modifications, literally from Ifat's version of the manuscript. They require double-checking because I was not involved with the work described in them.

which we assigned a heterozyous likelihood value, L_{het} of 1, separately from RNA seq data overlapping imputed heterozygous SNPs, where L_{het} values could range from 0 to 1. Based on iterative examination with different thresholds, we selected a L_{het} cutoff of 0.95 (i.e. imputation confidence level of 95%), and a minimal coverage of 7 reads per SNP. With these parameters, the discordance rate (monoallelic RNA genotype in the context of a heterozygous DNA genotype) was 0.71% for array-called SNPs and 3.2% for imputed SNPs.

While undoubtedly, a portion of the excess of discordance for the imputed SNPs is due to imputation error, downstream parts of analytic pipeline are designed taking into account the possibility of imputation error, as described below. One key is that for most genes there are multiple imputed SNPs and we consider data for all of them. Another key is that if even one SNP has evidence for biallelic expression (whether or not there is imputation data), we exlcude that individual due to the conflict. At this point, the matrix includes 147 million data points covering 213,208 SNPs, of which 114 million (77%) have imputation data.

2.4 The read count ratio and related quality filtering

The central quantity of this work is an $m \times n$ matrix of read count ratios $\mathbf{S} = [S_{ig}]_{ig};\ i = 1, ..., m;\ g \in \mathcal{G}$, where $m = 579$ is the number of individuals and n is the number of genes in a set $\mathcal{G}$ of unfiltered or filtered genes ($n = 15584$ and 5307, respectively). The read count ratio for individual i and gene g is defined as

$$S_{ig} = \frac{H_{ig}}{T_{ig}} = \frac{\sum_s H_s}{\sum_s T_s}, \tag{1}$$

where the summation runs through all heterozygous SNPs s that occur in (i, g). The statistic H_s and T_s in Eq. 1 are the higher and total RNA-seq read count at SNP s; H_s is higher in the sense that if the alleles at s are a, b and the corresponding read counts X_a, X_b, then $H_s = X_a$ if $X_a \geq X_b$ and $H_s = X_b$ otherwise. The total read count is simply $T_s = X_a + X_b$.

Two kind of data filters were applied sequentially: (1) a *read count-based* and (2) an *individual-based*. The read count-based filter removes any such pair (i, g) of individual i and genes g for which the total read count $T_{ig} < t_{\text{rc}}$, where t_{rc} is the read count threshold and was set to 15. The individual-based filter removes any genes g (across all individuals) if read count data involving g are available on less than t_{ind} number of individuals, set to 25. These filtering procedures were preceded by an initial read count-based filter, which removed each combination (i, s) of individual i and SNP s for which fewer than 7 reads had quality score ≥ 20. After the initial filtering the number of genes was $n = 15584$, which decreased to $n = 5307$ after the final, individual-based, filtering step.

The test for nearly unbiased expression of parental transcripts was defined by the criterion

$$S_{ig} \leq 0.6 \text{ and } \text{UCL}_{ig} \leq 0.7, \tag{2}$$

where the 95% upper confidence limit was calculated from a normal approximation to likelihood:

$$\text{UCL}_{ig} = S_{ig} + z_{0.975}\sqrt{\frac{S_{ig}(1 - S_{ig})}{T_{ig}}}, \tag{3}$$

such that z_p is the p quantile of the standard normal distribution and T_{ig} is, as before, the total read count.

model	transformation τ	weights w_{ig}
unlm.S	none	1
unlm.R	rank transf.	1
wnlm.S	none	T_{ig}
wnlm.R	rank transf.	T_{ig}

Table 1: Specificiation of four normal linear models based on read count transformation τ and weights w_{ig}.

2.5 Regression models to explain population-wide variation

Let m denote the number of individuals/samples and $\mathcal{G}$ the set of $n = 5307$ genes that passed quality filtering. Regression analysis involved a subset $\mathcal{G}_1 \subset \mathcal{G}$ of $n_1 = 30$ genes called as imprinted.

The basic model, unlm.S, is

$$\mathbf{S} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

$$\varepsilon_{ig} \overset{\text{iid}}{\sim} \text{Norm}(0, \sigma_g^2) \tag{5}$$

where the response $\mathbf{S}$ is an $m \times n_1$ matrix of read count ratios, X is an $m \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times n_1$ matrix of regression coefficients (Table 2), the random error $\boldsymbol{\varepsilon}$ has the same dimension as $\mathbf{S}$, and gene $g \in \mathcal{G}_1$. Eq. 4 may be given as

$$S_g = \mathbf{X}\beta_g + \varepsilon_g, \tag{6}$$

where the vectors $S_g, \beta_g, \varepsilon_g$ are single columns taken from their respective matrix counterparts.

unlm.S was extended in several ways, yielding

1. four normal linear models unlm.S, unlm.R, wnlm.S, wnlm.S
2. two logistic models logi.S and logi2.S.

The general form of the four normal linar models (cf. 6) is

$$\mathbf{W}_g^{1/2}\tau(S_g) = \mathbf{W}_g^{1/2}\mathbf{X}\beta_g + \varepsilon_g. \tag{7}$$

The extension here consists of $\mathbf{W}_g$, an $m \times m$ diagonal matrix of weights w_{ig} on the i-th diagonal position, and τ, a transformation on read counts. These quantities specify the four normal linear models as laid out in Table 1.

The logistic models, logi.S or logi2.S, share the general form

$$S_g = \mu_g + c\,\varepsilon_g \tag{8}$$

$$\mu_g = h(\mathbf{X}\beta_g) \tag{9}$$

$$\varepsilon_{ig} + \mu_g \overset{\text{iid}}{\sim} \text{Binom}(\mu_g, T_{ig}). \tag{10}$$

The link function h is $h(u) = e^u/(1+e^u)$ for logi.S and $h(u) = e^u/(2+2e^u)] + 1/2$ for logi2.S, and the scaling constant $c = 1$ and $1/2$, respectively. Thus, the response S_g under logi2.S is scaled and shifted relative to that under logi.S such that (with probabilty one) $1/2 \leq S_{ig} \leq 1$ under the former and $0 \leq S_{ig} \leq 1$ under the latter.

Each of the six models has $p \times n_1$ regression parameters corresponding to the dimension of $\boldsymbol{\beta}$. This allows different behavior for different genes since $\beta_1 \neq ... \neq \beta_{n_1}$ in general. Therefore, the estimated regression coefficients are reported as $\hat{\beta}_g = (\hat{\beta}_{1g}, ..., \hat{\beta}_{jg}, ..., \hat{\beta}_{pg})$ for each gene g, often replacing index j with the name of the parameter such as *Age* or *InstitutionPitt*.

A second set of six models was also fitted, for which $\boldsymbol{\beta}$ was constrained such that $\beta_1 = ... = \beta_{n_1}$. This was achieved by aggregating over genes $g \in \mathcal{G}_1$ the higher read count $H'_i = \sum_g H_{ig}$, the total read count $T'_i = \sum_g T_{ig}$, redefining the read count ratio as $S'_i = H'_i/T'_i$, and replacing S_g by $S' = (S'_1, ..., S'_m)$ in Eq. 6, 7, 8, and T_{ig} by T'_i in Table 1 and Eq. 10. Note that such aggregation simplifies the matrix variables in Eq. 4 to the corresponding vector variables in Eq. 6. Because S'_i is a weighted average of $\{S_{ig}\}_i$, results under these models are reported as $\hat{\beta}_{\text{WA}} = (\hat{\beta}_{1\text{WA}}, ..., \hat{\beta}_{j\text{WA}}, ..., \hat{\beta}_{p\text{WA}})$. A third set of models is a slight variation of this second set in that aggregation was done on a smaller subset of 8 genes selected at the initial stage of the study. Under these models the results are reported using the WA.8 subscript instead of WA.

These 3×6 models are all multiple regression ones with $p < 1$ parameters. Three corresponding sets of models with a single Age parameter ($p = 1$) were also fitted but the results were only used for graphical comparison of model fits in terms of predictions Fig. 5 but not for quantitative inference.

3 Results

3.1 Study design

The genome- and population-wide variation of parental bias was assessed using DLPFC tissue samples, one from each of $m = 579$ study individuals $i = 1, ..., m$ (Fig. 1). For each combination (i, g) of individuals and 15584 genes $g \in \{g_1, ..., g_{15584}\}$ (which passed an initial quality filter, see Section 2.4) the set of all heterozygous SNPs was identified with SNP-array genotyping, and expression was quantified at each SNP separately for the two alleles by counting RNA-seq reads noting the allele associated with the *higher read count* (as opposed to the *lower read count*). For a given (i, g) combination higher and lower read counts were then separately aggregated across the heterozygous SNPs yielding the statistics H_{ig} and L_{ig}, as well as the *total read count* $T_{ig} = H_{ig} + L_{ig}$. Genes were then quality filtered based on the conditional distribution of total read count across the 579 individuals for any given gene, leaving $n = 5307$ genes in the analysis. The *read count ratio* statistic, defined as $S_{ig} = H_{ig}/T_{ig}$, was used to quantify parental bias towards the more highly expressed parental allele (see also Eq. 1 in Section 2.4).

Taking a genome-wide viewpoint, it is helpful to regard the read count ratio as a set of random variables $\{S_{g_1}, ..., S_{g_n}\}$, each of which varies across the human population described by its own distribution. The difference (or similarity) among the corresponding set of distributions is an indicator of biological mechanisms that differentiate genes' parental bias. For any given gene g the observed $S_{1g}, ..., S_{mg}$ from the present data on $m \leq 579$ individuals estimates the distribution of S_g. That distribution, in turn, informs us on further biological mechanisms that differentiate individuals' parental bias for that gene but at the same time also reflects variation of S_g that arise from technical sources.

Besides the genomic measurements leading to read count ratios our data include observations on variables (Table 2) that we found (see Section 3.3 below) to be informative for the genome-wide dissection of various biological mechanisms and technical effects underlying the observed variation of

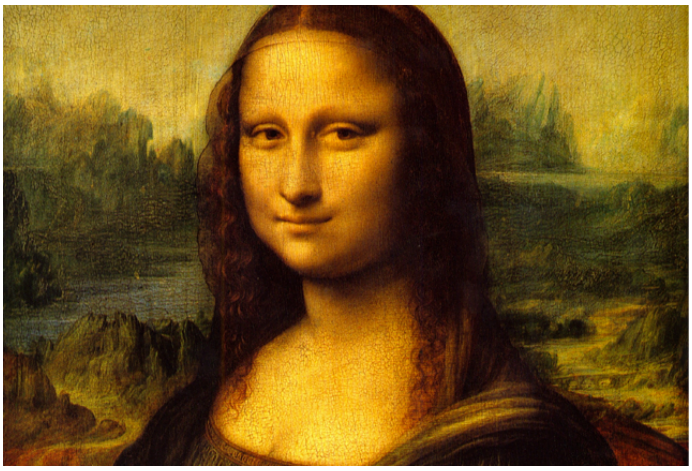

Figure 1: TODO: Study design. This figure (if deemed useful) will schematically illustrate variation of parental bias across a few genes by showing several maternal and paternal transcripts, and will also demonstrate technical sources of variation by depicting the corresponding RNA-seq read counts at heterozygous SNPs. Showing variation across individuals would be desirable but would complicate figure.

predictor	parameter(s)
Age	Age
Institution	[MSSM], Penn, Pitt
Gender	[Female], Male
PMI	PMI
Dx	[AFF], Control, SCZ
RIN	RIN
RIN2	RIN2
RNA_batch	[0], A, B, C, D, E, F, G, H
Ancestry.1	Ancestry.1
⋮	⋮
Ancestry.5	Ancestry.5

Table 2: Variables used as predictors of read count ratio for the study of regulation and consequences of parental bias. The right column lists the corresponding regression parameters and, in square brackets [ ], the baseline level against which other levels are contrasted. PMI: post-mortem interval; Dx: disease status; AFF: affective spectrum disorder; SCZ: schizophrenia; RIN: RNA integrity number; RIN2: the square of RIN; Ancestry.k: the k-th eigenvalue from the decomposition of genotypes indicating population structure

read counts. We carried out a theoretical study[2] that culminated in several probabilistic models of read counts—even at individual SNPs—and other observed variables (Fig. S1). These models may successfully capture the observed complex pattern of correlations among the measured variables (Section 3.3); but our theoretical work also showed that their computational implementation and evaluation of their properties and performance in relevant tasks would reach far beyond the present scope. Therefore, we decided to resort to relatively simple conventional models, some of which were found to fit reasonably well to allow quantitative inferences on the small subset of genes called imprinted (Section 3.3). On the genome-wide scale (next section) we present only an exploratory statistical analysis, which none-the-less permits qualitative conclusions under careful interpretation.

3.2 Genome- and population-wide variation of parental bias

The top three plots of Fig. 2 all show the empirical distribution of $S_{\text{PEG10}}, S_{\text{ZNF331}}$ and S_{AFAP1}, where PEG10 and ZNF331 are *known imprinted genes* based on prior evidence and AFAP1 is referred to as *candidate gene* as it lacks such evidence. For all three genes S_g varies greatly within its theoretical range $[\frac{1}{2}, 1]$. This variation is attributable to both technical and biological effects and is consistent with substantial population-wide variation of parental bias.

The probability density of S_g for the two known imprinted genes is shifted towards the theoretical maximum (relative to the density of the candidate gene), which indicates near monoallelic expression in a great fraction of individuals for these two genes. This is expected if S_g is indeed a useful estimator of the relative level of maternal (or paternal, whichever is greater) transcripts. But the shift is clearly stronger for PEG10 than for ZNF331, suggesting quantitative differences even among imprinted genes. This motivated the ranking of all 5307 genes based on a gene score that quantifies the shift in distribution. We defined the score of gene g as the fraction of individuals with $S_{ig} > 0.9$; as the filled green circles show in Fig. 2 this is equivalent to 1 less the empirical distribution function (ECDF), evaluated at 0.9 (the second and third plots from the top of Fig. 2).

The lower half of Fig. 2 shows the distribution of $S_{g_1}, ..., S_{g_n}$ ordered by rank from the top (rank 1) to bottom (rank 5307). Although the distributional shift is gradual from the top ranking $g_1 =$ MAGEL2 to the lowest ranking genes, Fig. 2 provides a visual argument that the fraction of imprinted genes is no more than 1% of all genes.

Consistent with previously described imprinted gene clusters the top-scoring genes tend to cluster according to genomic position (Fig. S2) and most, but not all, of them are known imprinted genes (Fig. 3). The high scoring candidates were classified on the basis of their distance from known imprinted gene clusters as nearby and distant candidates; the former class was taken as novel imprinted genes and the latter as false positives by considering the epigenetic nature of imprinting mechanisms and the typical, MB-scale, length of those epigenetic marks. Besides this mechanistic argument, the fraction of individuals passing a statistical test for the nearly unbiased expression of alleles (Eq. 2) also supports this classification, as shown by the black bars in Fig. 3. Conversely, more than a third of all known imprinted genes (within the 5307-sized gene set) score low. As the known imprinted genes were identified in different tissue types and organisms, these results indicate the context-dependence of imprinting.

[2] I wonder if we should attach my modeling article confidentially for the reviewers. See that article at: http://bernie.anbg.mssm.edu/~attila/assets/projects/monoallelic-brain/2016-04-14-braim-model.pdf

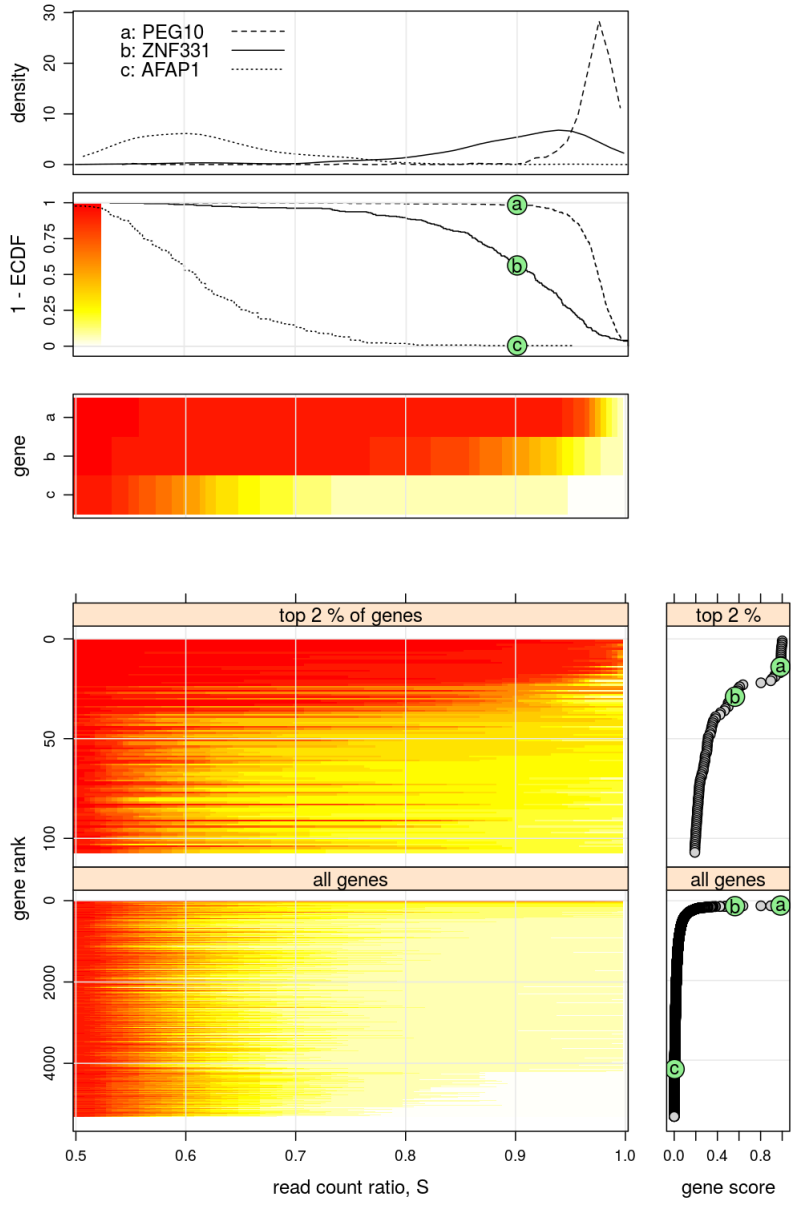
a: PEG10
b: ZNF331
c: AFAP1
density
1 - ECDF
gene
top 2 % of genes
all genes
gene rank
read count ratio, S
top 2 %
all genes
gene score

Figure 2:

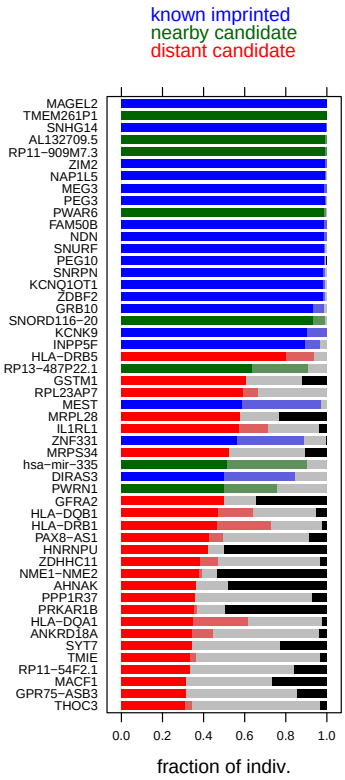
known imprinted
nearby candidate
distant candidate
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
HLA−DRB5
RP13−487P22.1
GSTM1
RPL23AP7
MEST
MRPL28
IL1RL1
ZNF331
MRPS34
hsa−mir−335
DIRAS3
PWRN1
GFRA2
HLA−DQB1
HLA−DRB1
PAX8−AS1
HNRNPU
ZDHHC11
NME1−NME2
AHNAK
PPP1R37
PRKAR1B
HLA−DQA1
ANKRD18A
SYT7
TMIE
RP11−54F2.1
MACF1
GPR75−ASB3
THOC3
0.0
0.2
0.4
0.6
0.8
1.0
fraction of indiv.

Figure 3:

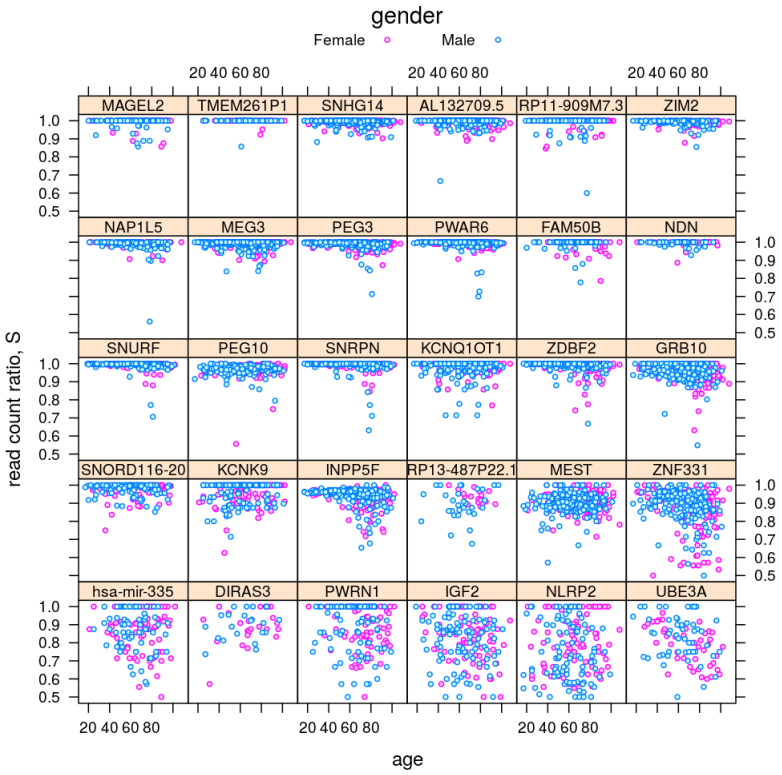
gender
Female
Male
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11-909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116-20
KCNK9
INPP5F
RP13-487P22.1
MEST
ZNF331
hsa-mir-335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
read count ratio, S
age
20 40 60 80
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4:

3.3 Modeling the variation of parental bias for selected genes

We selected the 27 top-scoring genes in the "known imprinted" and "nearby candidate" categories shown in blue and green in Fig. 3 and added three more known imprinted genes (NLRP2, IGF2 and UBE3A) that scored reasonably high (Fig. S3). The sources and psychological consequences of the variation of parental bias in these 30 genes was analyzed further through the detailed characterization how the read count ratio depends on the biological and technical variables—i.e. the predictors—listed in Table 2.

Fig. 4 shows patterns of dependence (or independence) of the read count ratio S_g for a given gene g on age and gender. From this visual inspection it seems that for several genes age is informative to the distribution of S_g in terms of both the location (e.g. the mean of S_g) and scale (e.g. variance); such apparent dependence on gender is not clear.

This qualitative result, however, is greatly complicated by the association among predictors: taking only pairwise associations the situation is already complex given the observed strong association between age and gender with each other and with many other predictors (Fig. S4) but higher order associations might also exist in the data. The correct interpretation of plots like Fig. 4 also depends on the amount of data, i.e. the total read count T_{ig}, based on which the read count ratio S_{ig} was calculated. Fig. S5 shows how T_{ig} varies both within a gene and across genes.

These considerations motivated us to fit various generalized linear regression models, one for each gene, by treating read count ratio as response and the variables in Table 2 as predictors. The thick black curves and colored strips in Fig. 5 represent the predicted read count ratio and prediction intervals, respectively, under four of these models for the gene ZDBF2. Among our models the logistic ones, logi.S and logi2.S, have several desirable theoretical properties: they prohibit values $S_g > 1$, capture at least some of the obvious dependence of the variance of S_g on its mean, and take into account the amount of available data since they are natural extensions of simple binomial models conditioned on the observed total read count T_{ig}. The known robustness and easy interpretation of normal linear models motivated their use for the present data. The flavors of normal linear models of this study are characterized by two features: (i.) whether or not they are weighted by T_{ig} and (ii.) the transformation, if any, that had been applied to S_{ig} before the fit. While weighting had little impact (not shown) the transformation was critical: Fig. 5 demonstrates how a quasi-log-transformation Q dramatically improves fit (compare wnlm.Q to wnlm.S).

An in-depth checking of model fit was carried out involving all models, predictors, and selected genes (except TMEM261P1, for which the iterative weighted least square fitting algorithm did not converge under logistic models). This analysis examined the normality of residuals (Fig. S6, S7, S8, S9, S10), the independence of residuals from the fitted value (i.e. homoscedasticity, Fig. S11, S12, S13, S14, S15), and the influence of each case/individual on the fit to address outliers (Fig. S16, S17, S13, S14, S20). These model checking results revealed the following patterns. First, for all genes the normal linear model seems to fit at least reasonably well to quasi-log-transformed data (wnlm.Q) slightly less well to rank-transformed data (wnlm.R) and much worse to untransformed data (wnlm.S). Second, both logistic models appear to exhibit greatly variable fit across genes but the fit of logi.S is in general better than that of logi2.S. Given these results we made inferences based on wnlm.Q and—if the quality of the fit permits for a given gene—also on logi.S.

Comparing the estimated regression coefficients $\hat{\beta}_{jg}$ for some gene g and parameter under the two logistic models logi.S and logi2.S revealed close agreement between these two models (Fig. S21) suggesting that making inferences under the better fitting logi.S and without logi2.S does not incur information loss. Comparing similarly the best-fitting normal linear model, wnlm.Q, to the less powerful and slightly less well fitting wnlm.R showed also good agreement (Fig. S22) suggesting

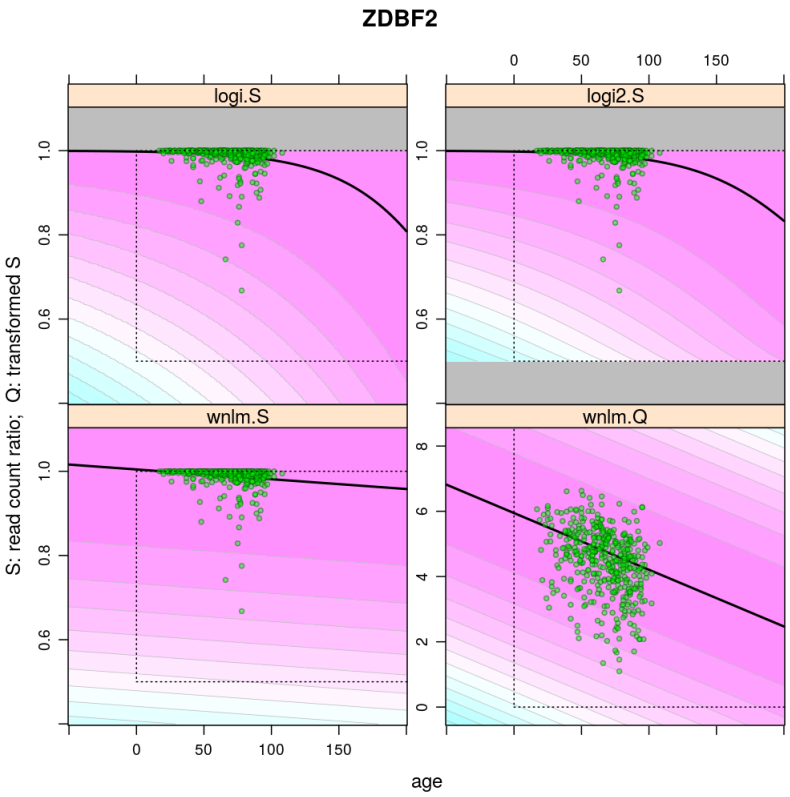
ZDBF2
logi.S
logi2.S
wnlm.S
wnlm.Q
S: read count ratio; Q: transformed S
age
0
50
100
150
0.6
0.8
1.0
0
2
4
6
8

Figure 5:

the use of wnlm.Q without wnlm.R. Comparing logi.S to wnlm.Q showed less close but still good qualitative agreement (Fig. S23) depending on how well logi.S fitted the data for a given gene (cf. S6). A comparison based on 99% confidence intervals (CIs) (Fig. S24, S25) also revealed that logi.S is more powerful in detecting a significant effect and rejecting the null hypothesis $\beta_{jg} = 0$, which is in line with the mentioned theoretical advantages of logistic models. But since the more conservative wnlm.Q provides in general better fit when all selected genes are considered, it was chosen as the preferred model of inference while logi.S is used mainly to confirm findings under wnlm.Q.

The p-value for each null hypothesis was calculated not only the standard way based on normal distribution theory but also from a random permutation test (Fig. S26). The two approaches to the p-value agreed closely under wnlm.Q for all genes while under logi.S the agreement depended on model fit and in particular on the normality of residuals (cf. Fig. S6).

These findings validate our approach to modeling the variation of parental bias (using wnlm.Q and logi.S models) and how that variation is explained by the observed predictor variables. In what follows we make inferences on the biological predictors.

3.4 Inferring regulators and consequences of parental bias

Since all of these models account for each predictor separately, they are expected, in principle, to successfully dissect their effects although the observed association among the predictors might adversely affect parameter estimation due to collinearity. Inspection of likelihood surfaces suggested that perhaps with the exception of those including the predictor RIN, the associations only moderately distort likelihood surfaces (Fig. S29).

Fig. S24 ,S25, and present the estimated regression coefficients along with their 99% confidence intervals. The more significant results under logistic models indicated that they were indeed more powerful than the weighted or unweighed normal linear models, in agreement with their better fit and favorable theoretical properties and that they avoid the loss of information incurred by rank transformation. But it must be added to this conclusion that logistic models might also be more prone to bias because much of their predicted sigmoidal curve was extrapolated from the data (Fig. 5). Results were very similar under the two logistic models. Moreover, the great variation of regression coefficients across genes suggested that the ".WA" models were inadequate because predictors likely affect genes differentially. These findings made us give priority to the logi.S model and use wnlm.R mainly to check the results' consistency under these two models.

The effects of four biological predictors—age, gender, ancestry and psychiatric condition—on read count ratio of individual genes are depicted in Fig. 6.

TODO: finish the rest of the Results section.

4 Discussion

The main results of this work may be interpreted in terms of the scheme in Fig. 8. The scheme presents all n imprinted genes in a tissue, such as the DLPFC, which is relevant to neural function. Parental bias is putatively regulated by age, gender, ancestry, and possibly other biological factors. The nature of regulation—simple up- or down-regulation, no effect, or some more complex pattern—is captured in the functions $\psi_1, ..., \psi_n$. Function ϕ maps jointly the n parental biases to neuropsychology, and connects therefore the molecular phenotypic level to the organismal one.

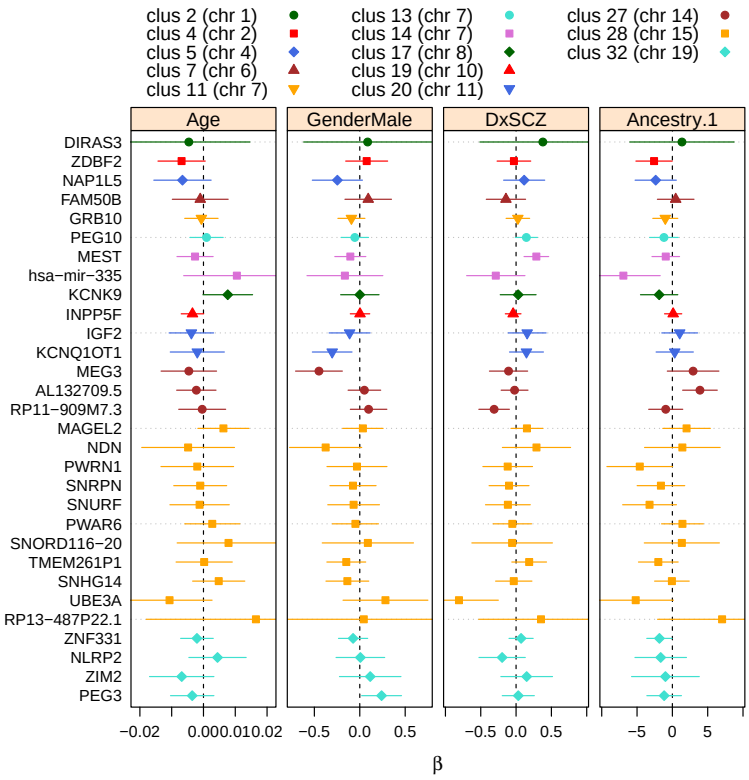

Figure 6: Biological effects: estimate and 99% confidence interval for each regression coefficient β_{jg} under the wnlm.Q model, where g corresponds to a gene and j to a biological covariate (Age, Ancestry.1) or a level of some biological factor (DxSCZ, GenderMale).

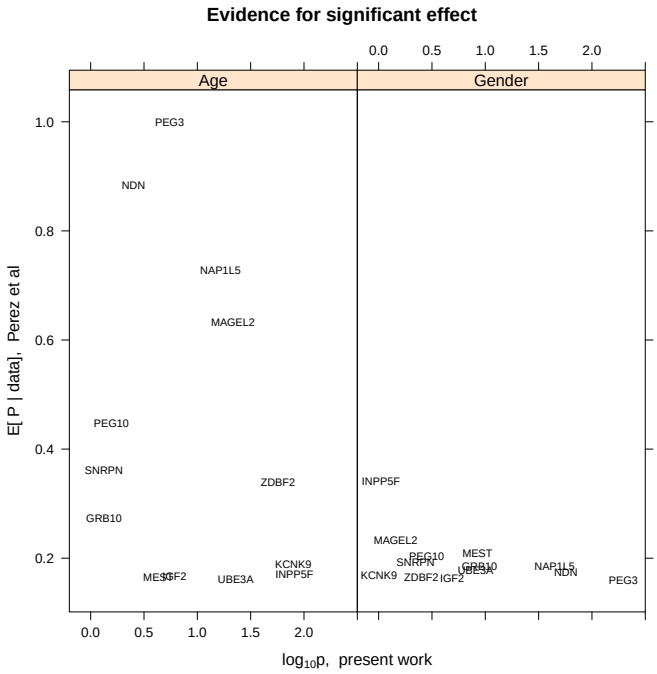

Figure 7: Comparison of the effects of age and gender between the present work and a previous study [5] in the mouse cerebellum.

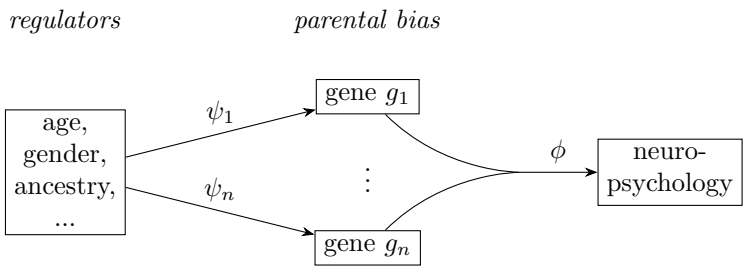

Figure 8:

Our genome-wide analysis suggests that $< 1\%$ of all genes are imprinted, which implies $n \approx 200$. This provides evidence, in addition to previous similarly conservative estimates of n [5, 2], against the controversial estimate of $n \approx 1300$ [3].

More interestingly, we find that several regression coefficients differ significantly from zero, which suggests that age, gender and ancestry do indeed regulate parental bias in at least some imprinted genes in the human DLPFC. We infer that these three regulators exert gene-specific effects, up- or down-regulating parental bias in a subset of imprinted genes while having no (detectable) impact on the rest. That these predictors differ in the subset of genes they affect significantly hints at the potential complexity of regulation. A further layer of that complexity arises from possible interactions among these predictors, which is in fact consistent with the dependence of the age effect on gender in our conditional analysis.

The interpretation of the present results as the effect of (human) ancestry on parental bias points to genetic regulatory mechanisms of imprinting that modulate the known epigenetic mechanisms. Given the late emergence of imprinting in therian phylogeny and its role in neuropsychological and social function, the genetics of parental bias may well be an increasingly important target of natural selection. Also, the ancestry effect is a remarkable novelty of our work since previous studies, all using in-bread mouse strains, failed to address this point. As for age, a similar differential effect was found in the mouse cerebellum to the effect we observe here. Gender, however, had no significant effect in the same mouse study.

The above interpretation certainly depends on our statistical inference, which in turn is based on the present data and regression models, both of which have serious limitations. Some of these—related to differing sequencing and genotyping protocols—might be alleviated by improved across-institute standardization. But others, such as the confounding of age by certain technical variables (e.g. institution), are hopelessly entangled with the observational nature of post mortem human studies, which precludes orthogonal, that is clearly interpretable, decomposition of the variation of the observed measure of parental bias (the response) into separate technical and biological components. In addition, non-orthogonality also limits statistical power. But even if the data fulfilled orthogonality, the present regression models would still remain too rigid to account for the relative overdispersion of RNA-seq read counts, the uncertainty surrounding haplotype, and that genes are neither completely independent nor completely identical in their parental bias. In future work these ought to be tackled either with recently developed hierarchical models built on normal linear model [5, 4] or, if regulators indeed strongly interact as the present work indicates, by the adoption of Bayesian networks.

The molecular mechanism mediating the age effect. TODO: clusters and regression coefficients for gender and ancestry

Even if the regulatory effect of age, gender, ancestry, etc, on parental bias is more firmly established by methodological improvements and more data, it still remains to be determined whether (and how) the corresponding changes in expression phenotype affect neural and psychological properties and might be causal to some common psychiatric disorders. This work may be considered an initial step towards that aim as the estimated regression coefficients, associated with SCZ or AFF, provide statistically weak hints at that causality.

A more general question is how function ϕ in Fig. 8 integrates parental bias across genes and maps that signal to organismal phenotype. That mapping occurs through the intermediate domain of cellular metabolism. Therefore, extending the present "multi-omic" data collection with a metabolic layer appears promising, especially because rather specific metabolism characterizes several imprinted genes [8, 6]. Adopting single-cell RNA-seq in the present framework would be another interesting extension that could possibly elucidate the role of random monoallelic expression.

In fact, the question regarding the map ϕ is even more general, since it is plausible that parental bias acts jointly with overall expression level (this is not indicated in Fig. 8). If so, the resolution of the prevailing system biology is to be refined from mere genes to separate maternal and paternal copies. On the other hand, the present finding that parental bias substantially varies across individuals even withing the same tissue calls for a conceptual shift from the current practice of regarding genes unconditionally as either imprinted or not to considering instead the conditional distribution of their parental bias within the population given regulators such as age, gender, and ancestry.

References

[1] Andrew Chess. Mechanisms and consequences of widespread random monoallelic expression. *Nat. Rev. Genet.*, 13(6):421–8, jun 2012.

[2] Brian DeVeale, Derek van der Kooy, and Tomas Babak. Critical evaluation of imprinted gene expression by RNA-seq: A new perspective. *PLoS Genetics*, 8(3):e1002600, jan 2012.

[3] Christopher Gregg, Jiangwen Zhang, Brandon Weissbourd, Shujun Luo, Gary P Schroth, David Haig, and Catherine Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science (New York, N.Y.)*, 329(5992):643–8, aug 2010.

[4] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, jan 2014.

[5] Julio D Perez, Nimrod D Rubinstein, Daniel E Fernandez, Stephen W Santoro, Leigh A Needleman, Olivia Ho-Shing, John J Choi, Mariela Zirlinger, Shau-Kwaun Chen, Jun S Liu, and Catherine Dulac. Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. *eLife*, 4:e07860, jan 2015.

[6] Jo Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics*, 15(8):517–530, jun 2014.

[7] Robert N Plasschaert, Marisa S Bartolomei, S. Abu-Amero, D. Monk, S. Apostolidou, P. Stanier, G. Moore, U. Albrecht, J. S. Sutcliffe, B. M. Cattanach, C. V. Beechey, D. Armstrong,

G. Eichele, A. L. Beaudet, S. C. Andrews, M. D. Wood, S. J. Tunster, S. C. Barton, A. M. Surani, R. M. John, T. Arima, T. Kamikihara, T. Hayashida, K. Kato, T. Inoue, Y. Shirayoshi, M. Oshimura, H. Soejima, T. Mukai, N. Wake, T. Arima, K. Hata, S. Tanaka, M. Kusumi, E. Li, K. Kato, K. Shiota, H. Sasaki, N. Wake, D. P. Barlow, R. Stöger, B. G. Herrmann, K. Saito, N. Schweifer, M. S. Bartolomei, A. C. Ferguson-Smith, M. S. Bartolomei, S. Zemel, S. M. Tilghman, S. C. Barton, M. Surani, M. L. Norris, S. B. Baylin, P. A. Jones, V. Besson, P. Smeriglio, A. Wegener, F. Relaix, B. N. Oumesmar, D. A. Sassoon, G. Marazzi, D. Bourc'his, G.-L. Xu, C.-S. Lin, B. Bollman, T. H. Bestor, K. D. Broad, J. P. Curley, E. B. Keverne, M. G. Butler, B. W. Carey, S. Markoulaki, J. H. Hanna, D. A. Faddah, Y. Buganim, J. Kim, K. Ganz, E. J. Steine, J. P. Cassady, M. P. Creyghton, T. Caspary, M. A. Cleary, E. J. Perlman, P. Zhang, S. J. Elledge, S. M. Tilghman, S. J. Chamberlain, P.-F. Chen, K. Y. Ng, F. Bourgois-Rocha, F. Lemtiri-Chlieh, E. S. Levine, M. Lalande, F. A. Champagne, J. P. Curley, W. T. Swaney, N. S. Hasen, E. B. Keverne, M. Charalambous, F. M. Smith, W. R. Bennett, T. E. Crew, F. Mackenzie, A. Ward, M. Chotalia, S. A. Smallwood, N. Ruf, C. Dawson, D. Lucifero, M. Frontera, K. James, W. Dean, G. Kelsey, S. Choufani, C. Shuman, R. Weksberg, M. Constância, M. Hemberger, J. Hughes, W. Dean, A. Ferguson-Smith, R. Fundele, F. Stewart, G. Kelsey, A. Fowden, C. Sibley, E. M. Cooper, A. W. Hudson, J. Amos, J. Wagstaff, P. M. Howley, J. P. Curley, S. Barton, A. Surani, E. B. Keverne, T. L. Davis, G. J. Yang, J. R. McCarrey, M. S. Bartolomei, M. M. Dawlaty, A. Breiling, T. Le, G. Raddatz, M. I. Barrasa, A. W. Cheng, Q. Gao, B. E. Powell, Z. Li, M. Xu, T. M. DeChiara, E. J. Robertson, A. Efstratiadis, C. L. Dent, A. R. Isles, C. A. Edwards, A. C. Ferguson-Smith, J. Eggenschwiler, T. Ludwig, P. Fisher, P. A. Leighton, S. M. Tilghman, A. Efstratiadis, N. Engel, A. G. West, G. Felsenfeld, M. S. Bartolomei, A. P. Feinberg, A. C. Ferguson-Smith, B. M. Cattanach, S. C. Barton, C. V. Beechey, M. A. Surani, S. R. Ferrón, M. Charalambous, E. Radford, K. McEwen, H. Wildner, E. Hind, J. M. Morante-Redolat, J. Laborda, F. Guillemot, S. R. Bauer, E. Foulstone, S. Prince, O. Zaccheo, J. L. Burns, J. Harper, C. Jacobs, D. Church, A. B. Hassan, D. Frank, W. Fortino, L. Clark, R. Musalo, W. Wang, A. Saxena, C.-M. Li, W. Reik, T. Ludwig, B. Tycko, A. Gabory, M.-A. Ripoche, A. Le Digarcher, F. Watrin, A. Ziyyat, T. Forné, H. Jammes, J. F. X. Ainscough, M. A. Surani, L. Journot, A. S. Garfield, M. Cowley, F. M. Smith, K. Moorwood, J. E. Stewart-Cox, K. Gilroy, S. Baker, J. Xia, J. W. Dalley, L. D. Hurst, C. Gicquel, S. Rossignol, S. Cabrol, M. Houang, V. Steunou, V. Barbu, F. Danton, N. Thibaud, M. L. Merrer, L. Burglen, P. L. Greer, R. Hanayama, B. L. Bloodgood, A. R. Mardinly, D. M. Lipton, S. W. Flavell, T.-K. Kim, E. C. Griffith, Z. Waldon, R. Maehr, F. Guillemot, A. Nagy, A. Auerbach, J. Rossant, A. L. Joyner, J. A. Hackett, R. Sengupta, J. J. Zylicz, K. Murakami, C. Lee, T. A. Down, M. A. Surani, P. Hajkova, K. Ancelin, T. Waldmann, N. Lacoste, U. C. Lange, F. Cesari, C. Lee, G. Almouzni, R. Schneider, M. A. Surani, Y. Hao, T. Crenshaw, T. Moulton, E. Newcomb, B. Tycko, K. Hata, M. Okano, H. Lei, E. Li, D. H. Heck, Y. Zhao, S. Roy, M. S. LeDoux, L. T. Reiter, A. Henckel, K. Chebli, S. K. Kota, P. Arnaud, R. Feil, K. Higashimoto, H. Soejima, T. Saito, K. Okumura, T. Mukai, H. Hiura, M. Toyoda, H. Okae, M. Sakurai, N. Miyauchi, A. Sato, N. Kiyokawa, H. Okita, Y. Miyagawa, H. Akutsu, T. M. Holm, L. Jackson-Grusby, T. Brambrink, Y. Yamada, W. M. Rideout, R. Jaenisch, L. Holt, K. Siddle, D. Humpherys, M. D. Johnson, X. Wu, N. Aithmitti, R. S. Morrison, S. Kagiwada, K. Kurimoto, T. Hirota, M. Yamaji, M. Saitou, M. Kaneda, M. Okano, K. Hata, T. Sado, N. Tsujimoto, E. Li, H. Sasaki, A. Keniry, D. Oxley, P. Monnier, M. Kyba, L. Dandolo, G. Smits, W. Reik, E. B. Keverne, R. Fundele, M. Narasimha, S. C. Barton, M. A. Surani, T. Kishino, M. Lalande, J. Wagstaff, S. Kumar, A. L. Talis, P. M. Howley, S. Kühnle, B. Mothes, K. Matentzoglu,

M. Scheffner, T. Kuwajima, I. Nishimura, K. Yoshikawa, P. A. Latos, F. M. Pauler, M. V. Koerner, H. B. Senergin, Q. J. Hudson, R. R. Stocsits, W. Allhoff, S. H. Stricker, R. M. Klement, K. E. Warczok, J. T. Lee, M. S. Bartolomei, L. Lefebvre, S. Viville, S. C. Barton, F. Ishino, E. B. Keverne, M. A. Surani, P. A. Leighton, J. R. Saam, R. S. Ingram, C. L. Stewart, S. M. Tilghman, L. Li, X. Li, M. Ito, F. Zhou, N. Youngson, X. Zuo, P. Leder, A. C. Ferguson-Smith, L. Liu, G.-Z. Luo, W. Yang, X. Zhao, Q. Zheng, Z. Lv, W. Li, H.-J. Wu, L. Wang, X.-J. Wang, D. Lucifero, M. R. W. Mann, M. S. Bartolomei, J. M. Trasler, T. Ludwig, J. Eggenschwiler, P. Fisher, A. J. D'Ercole, M. L. Davenport, A. Efstratiadis, D. J. G. Mackay, J. L. A. Callaway, S. M. Marks, H. E. White, C. L. Acerini, S. E. Boonen, P. Dayanikli, H. V. Firth, J. A. Goodship, A. P. Haemers, S. Matsuoka, J. S. Thompson, M. C. Edwards, J. M. Bartletta, P. Grundy, L. M. Kalikin, J. W. Harper, S. J. Elledge, A. P. Feinberg, J. McGrath, D. Solter, P. Monnier, C. Martinet, J. Pontis, I. Stancheva, S. Ait-Si-Ali, L. Dandolo, F. Muscatelli, D. N. Abrous, A. Massacrier, I. Boccaccio, M. Le Moal, P. Cau, H. Cremer, T. Nagano, P. Fraser, T. Nakamura, Y.-J. Liu, H. Nakashima, H. Umehara, K. Inoue, S. Matoba, M. Tachibana, A. Ogura, Y. Shinkai, T. Nakano, H. Okae, H. Hiura, Y. Nishida, R. Funayama, S. Tanaka, H. Chiba, N. Yaegashi, K. Nakayama, H. Sasaki, T. Arima, M. Okano, D. W. Bell, D. A. Haber, E. Li, R. Ono, K. Nakamura, K. Inoue, M. Naruse, T. Usami, N. Wakisaka-Saito, T. Hino, R. Suzuki-Migishima, N. Ogonuki, H. Miki, S. Pagliardini, J. Ren, R. Wevrick, J. J. Greer, B. Papp, K. Plath, W. A. Pastor, L. Aravind, A. Rao, K. Pelc, S. G. Boyd, G. Cheron, B. Dan, M. Pick, Y. Stelzer, O. Bar-Nur, Y. Mayshar, A. Eden, N. Benvenisty, A. Plagge, A. R. Isles, E. Gordon, T. Humby, W. Dean, S. Gritsch, R. Fischer-Colbrie, L. S. Wilkinson, G. Kelsey, A. R. Prickett, R. J. Oakey, S. Ramamoorthy, Z. Nawaz, A. Rieusset, F. Schaller, U. Unmehopa, V. Matarazzo, F. Watrin, M. Linke, B. Georges, J. Bischof, F. Dijkstra, M. Bloemsma, A. H. Salehi, S. Xanthoudakis, P. A. Barker, M. Scheffner, J. M. Huibregtse, R. D. Vierstra, P. M. Howley, Y. Sekita, H. Wagatsuma, K. Nakamura, R. Ono, M. Kagami, N. Wakisaka, T. Hino, R. Suzuki-Migishima, T. Kohda, A. Ogura, H. Shiura, K. Nakamura, T. Hikichi, T. Hino, K. Oda, R. Suzuki-Migishima, T. Kohda, T. Kaneko-Ishino, F. Ishino, K. S. Srivenugopal, X.-H. Yuan, H. S. Friedman, F. Ali-Osman, M. Stadtfeld, K. Hochedlinger, M. Stadtfeld, E. Apostolou, H. Akutsu, A. Fukuda, P. Follett, S. Natesan, T. Kono, T. Shioda, K. Hochedlinger, M. Stadtfeld, E. Apostolou, F. Ferrari, J. Choi, R. M. Walsh, T. Chen, S. S. K. Ooi, S. Y. Kim, T. H. Bestor, T. Shioda, M. A. Surani, K. Hayashi, P. Hajkova, D. F. Swaab, H. Taniura, K. Matsumoto, K. Yoshikawa, A. Varrault, C. Gueydan, A. Delalbre, A. Bellmann, S. Houssami, C. Aknin, D. Severac, L. Chotard, M. Kahli, A. Le Digarcher, A. Venkatraman, X. C. He, J. L. Thorvaldsen, R. Sugimura, J. M. Perry, F. Tao, M. Zhao, M. K. Christenson, R. Sanchez, J. Y. Yu, M. Wernig, A. Meissner, R. Foreman, T. Brambrink, M. Ku, K. Hochedlinger, B. E. Bernstein, R. Jaenisch, S. Yamaguchi, L. Shen, Y. Liu, D. Sendler, Y. Zhang, Y. Yu, S.-O. Yoon, G. Poulogiannis, Q. Yang, X. M. Ma, J. Villen, N. Kubica, G. R. Hoffman, L. C. Cantley, S. P. Gygi, S. J. Zacharek, C. M. Fillmore, A. N. Lau, D. W. Gludish, A. Chou, J. W. K. Ho, R. Zamponi, R. Gazit, C. Bock, N. Jäger, P. Zhang, C. Wong, R. A. DePinho, J. W. Harper, S. J. Elledge, R. Zwart, S. Verhaagh, M. Buitelaar, C. Popp-Snijders, and D. P. Barlow. Genomic imprinting in development, growth, behavior and stem cells. *Development (Cambridge, England)*, 141(9):1805–13, may 2014.

[8] Valter Tucci, SN Archer, H Oster, S Maret, S Dorsaz, L Gurcel, S Pradervand, B Petit, C Pfister, R Massart, M Freyburger, M Suderman, J Paquet, J El Helou, E Belanger-Nelson, A Azzi, R Dallmann, A Casserly, H Rehrauer, A Patrignani, B Maier, E Mignot, G Tononi, C Cirelli, JM Krueger, MG Frank, JP Wisor, S Roy, G Tononi, C Cirelli, MS Bartolomei, AC Ferguson-

Smith, J Peters, LS Wilkinson, W Davies, AR Isles, WT Powell, JM LaSalle, D Landgraf, CE Koch, H Oster, V Tucci, C Saini, DM Suter, A Liani, P Gos, U Schibler, MD Schwartz, TS Kilduff, M Murphy, R Huber, S Esser, BA Riedner, M Massimini, F Ferrarelli, JM Siegel, M Engle-Friedman, P McNamara, A Vela-Bueno, A Kales, CR Soldatos, B Dobladez-Blanco, J Campos-Castello, P Espino-Hurtado, AN Vgontzas, A Kales, J Seip, MJ Mascari, EO Bixler, DC Myers, G Hertz, M Cataletto, SH Feinsilver, M Angulo, IV Zhdanova, RJ Wurtman, J Wagstaff, J Clayton-Smith, L Laan, JC Ehlen, KA Jones, L Pinckney, CL Gray, S Burette, RJ Weinberg, SQ Shi, TJ Bichell, RA Ihrie, CH Johnson, NC Gossan, F Zhang, B Guo, D Jin, H Yoshitane, A Yao, G Lassi, L Priano, S Maggi, C Garcia-Garcia, E Balzani, N El-Assawy, K Krauchi, T Deboer, SV Kozlov, JW Bogenpohl, MP Howell, R Wevrick, S Panda, JB Hogenesch, G Lassi, ST Ball, S Maggi, G Colonna, T Nieus, C Cero, EJ Van Someren, MB Renfree, S Suzuki, T Kaneko-Ishino, JM Stringer, AJ Pask, G Shaw, MB Renfree, WF Flanigan Jr., RH Wilcox, A Rechtschaffen, KM Hartse, A Rechtschaffen, F Ayala-Guerrero, S Huitron Resendiz, AC Huntley, JM Siegel, JM Siegel, PR Manger, R Nienhuis, HM Fahringer, JD Pettigrew, T Allison, H Van Twyver, T Allison, H Van Twyver, WR Goff, JH Brown, JF Gillooly, AP Allen, VM Savage, GB West, MF Bonetti, JJ Wiens, A Clarke, P Rothery, NJ Isaac, F Tinarelli, C Garcia-Garcia, F Nicassio, V Tucci, SR Ferron, M Charalambous, E Radford, K McEwen, H Wildner, and E Hind. Genomic Imprinting: A New Epigenetic Perspective of Sleep Regulation. *PLOS Genetics*, 12(5):e1006004, may 2016.

5 Supplementary Material

Figure S1: AGK models: M1 (left) and M2 (right)

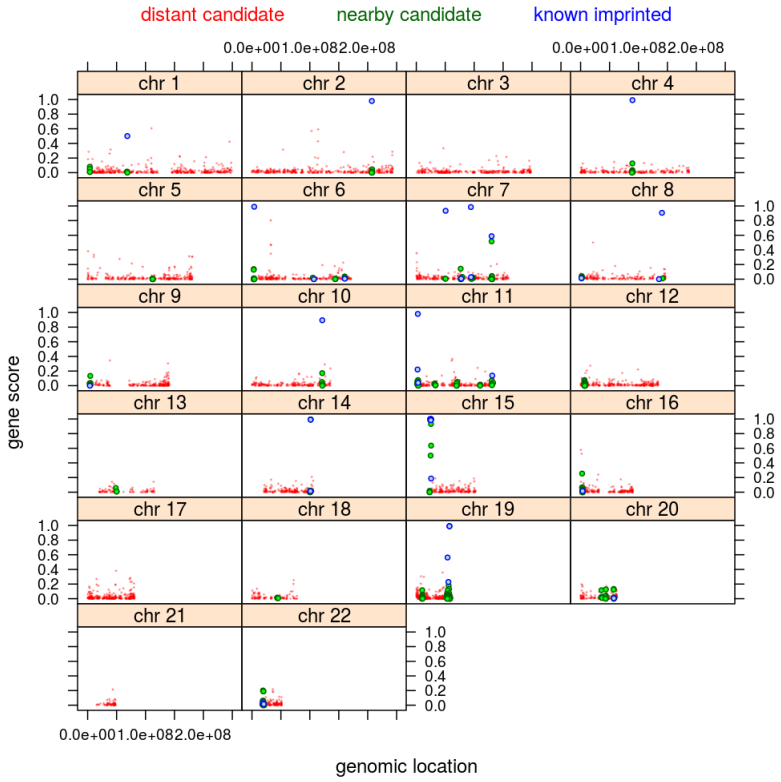
distant candidate
nearby candidate
known imprinted
chr 1
chr 2
chr 3
chr 4
chr 5
chr 6
chr 7
chr 8
chr 9
chr 10
chr 11
chr 12
chr 13
chr 14
chr 15
chr 16
chr 17
chr 18
chr 19
chr 20
chr 21
chr 22
gene score
genomic location

Figure S2:

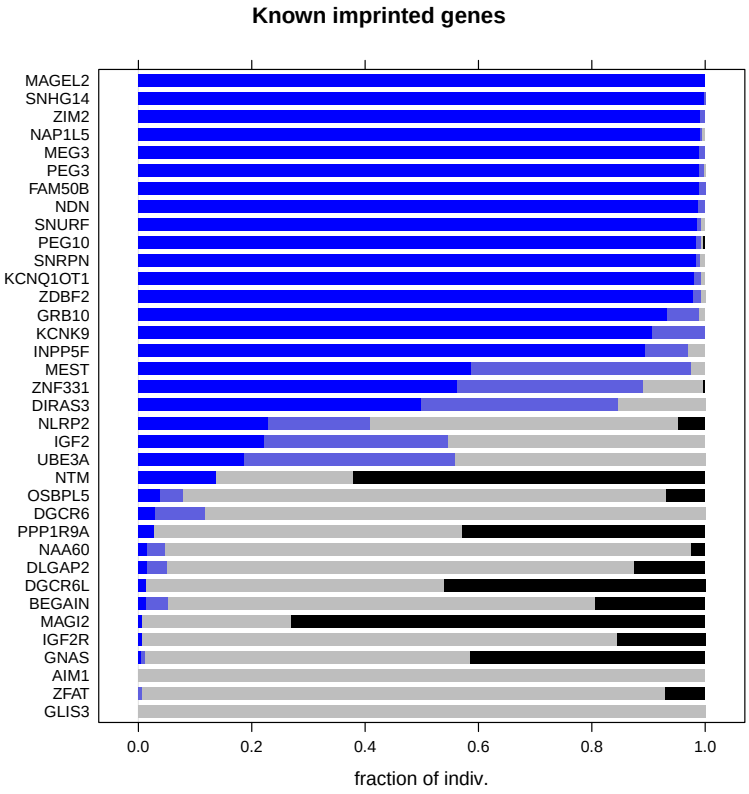
Known imprinted genes
MAGEL2
SNHG14
ZIM2
NAP1L5
MEG3
PEG3
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
KCNK9
INPP5F
MEST
ZNF331
DIRAS3
NLRP2
IGF2
UBE3A
NTM
OSBPL5
DGCR6
PPP1R9A
NAA60
DLGAP2
DGCR6L
BEGAIN
MAGI2
IGF2R
GNAS
AIM1
ZFAT
GLIS3
0.0
0.2
0.4
0.6
0.8
1.0
fraction of indiv.

Figure S3:

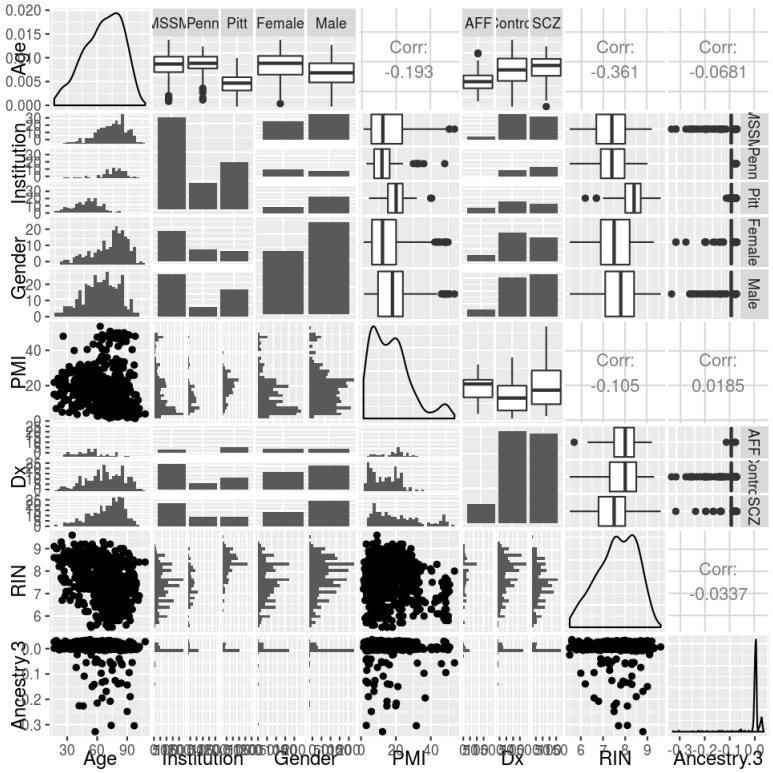
Age
Institution
Gender
PMI
Dx
RIN
Ancestry.3
Penn
Pitt
Female
Male
SCZ
Corr:
-0.193
Corr:
-0.361
Corr:
-0.0681
Corr:
-0.105
Corr:
0.0185
Corr:
-0.0337

Figure S4:

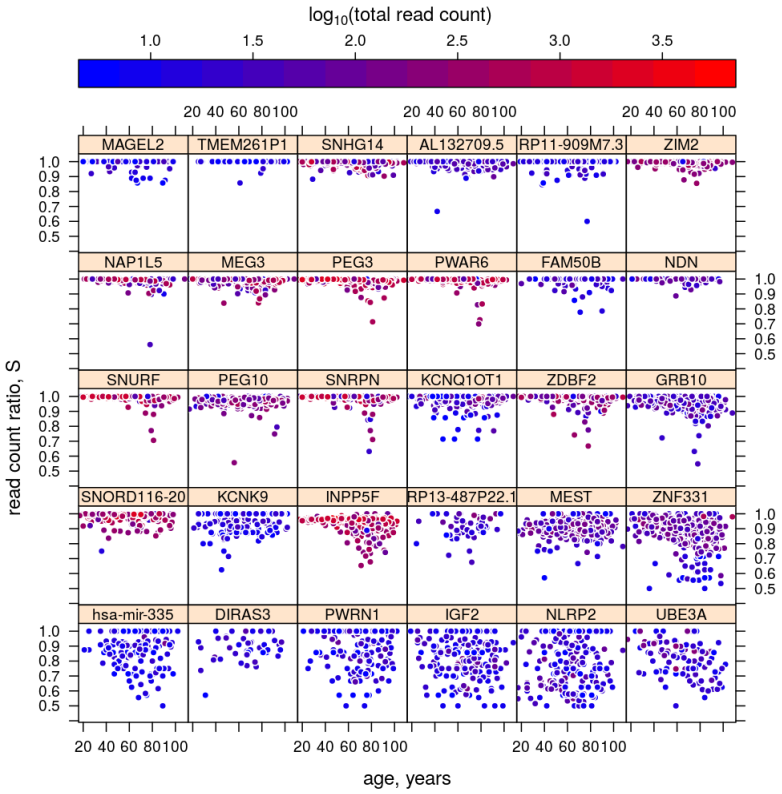
log10(total read count)
1.0
1.5
2.0
2.5
3.0
3.5
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11-909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116-20
KCNK9
INPP5F
RP13-487P22.1
MEST
ZNF331
hsa-mir-335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
read count ratio, S
age, years

Figure S5:

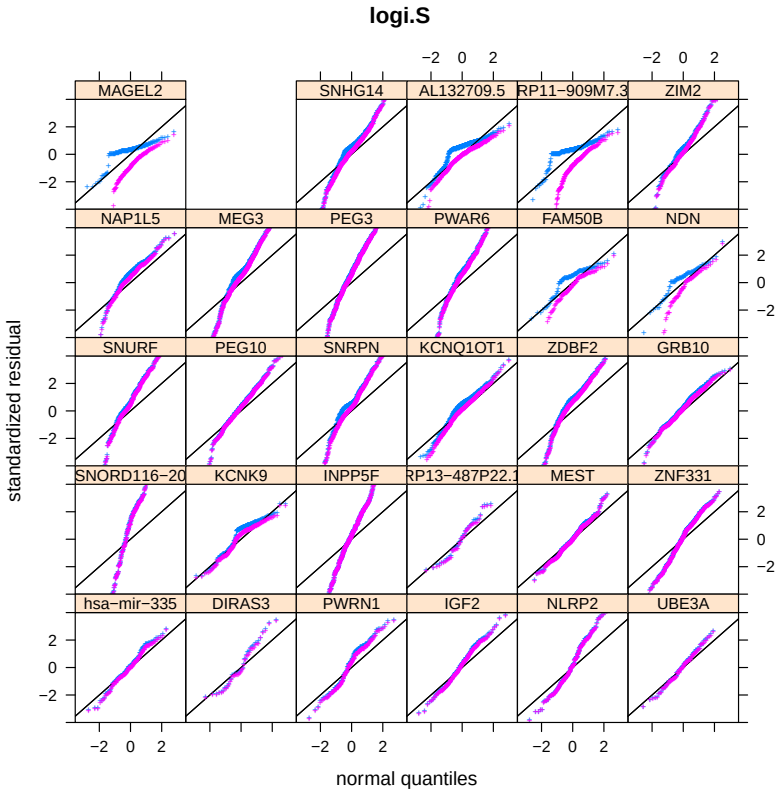
logi.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
standardized residual
normal quantiles

Figure S6:

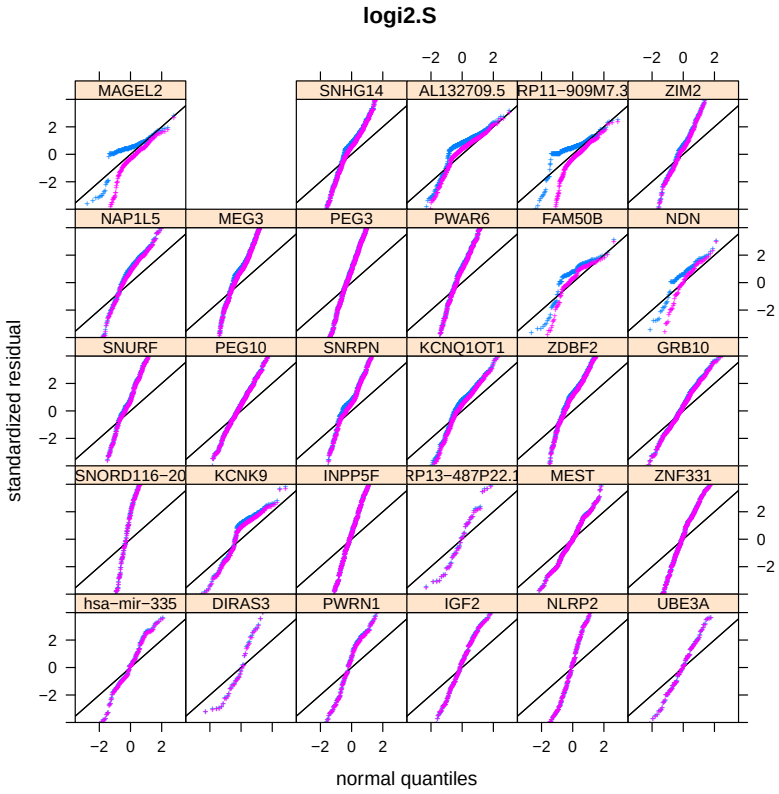
logi2.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
standardized residual
normal quantiles

Figure S7:

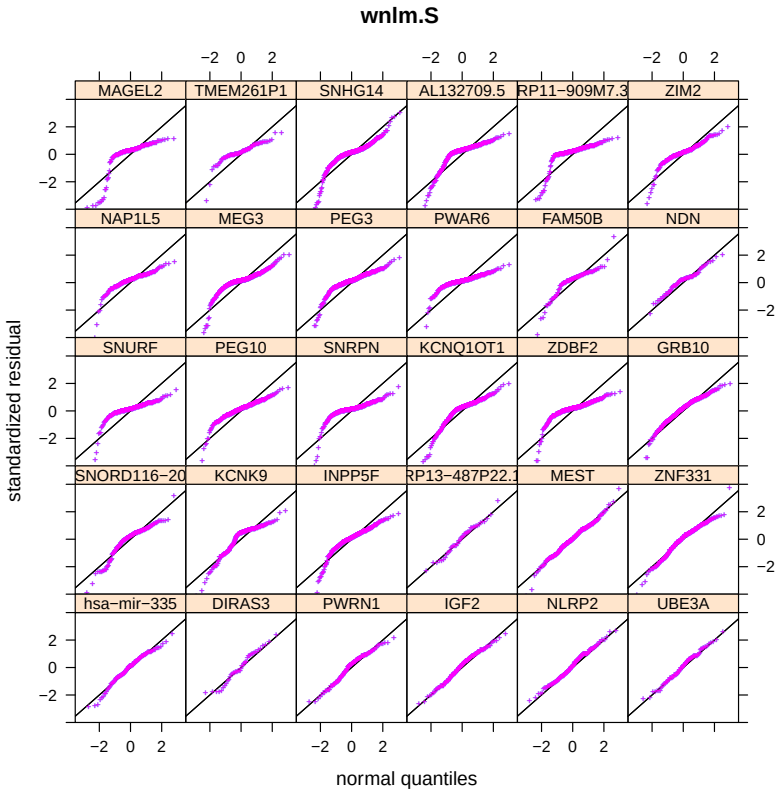
wnlm.S
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
standardized residual
normal quantiles

Figure S8:

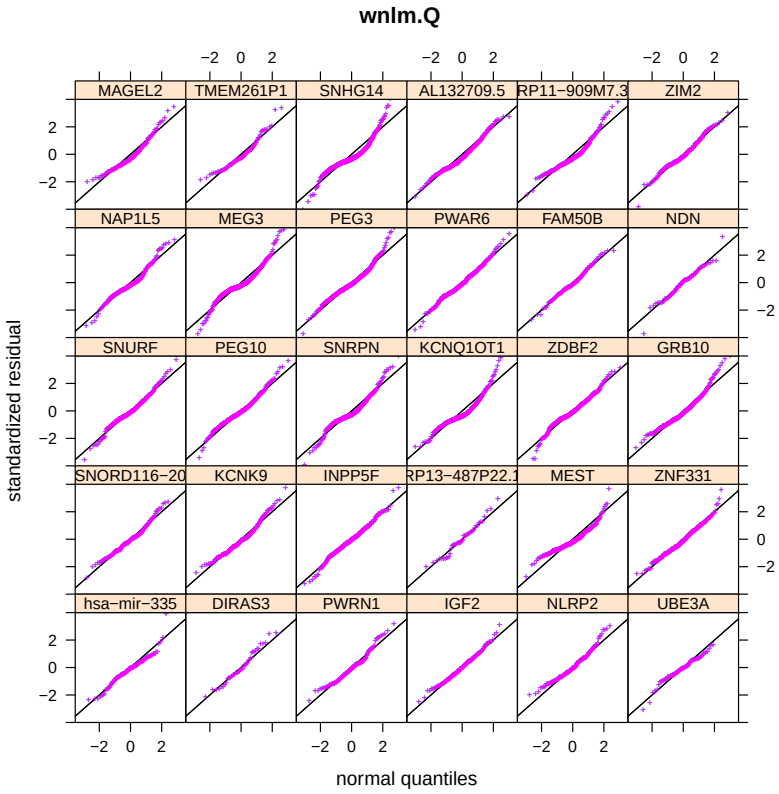
wnlm.Q
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
standardized residual
normal quantiles

Figure S9:

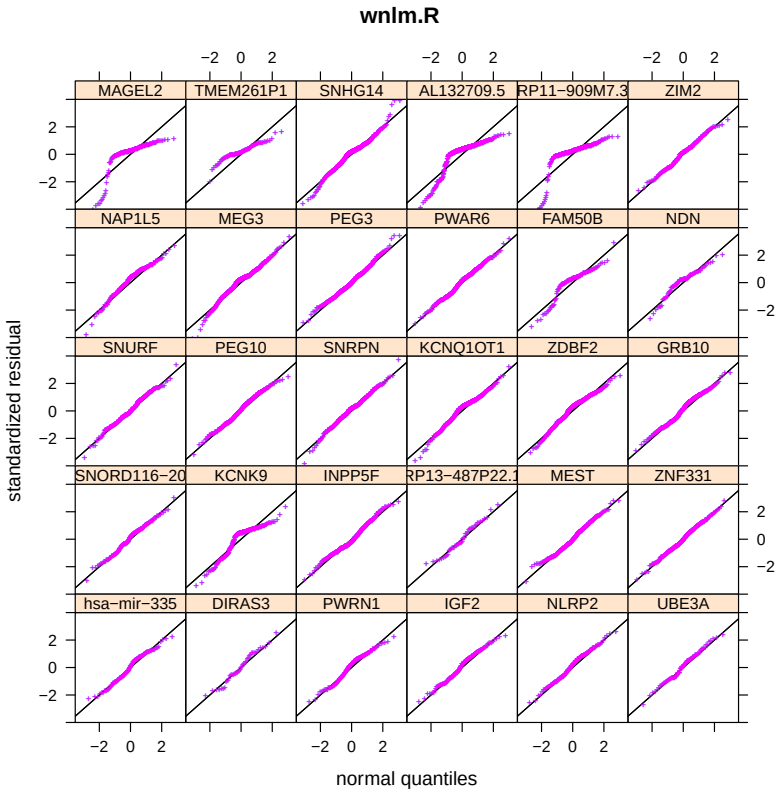
wnlm.R
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
standardized residual
normal quantiles

Figure S10:

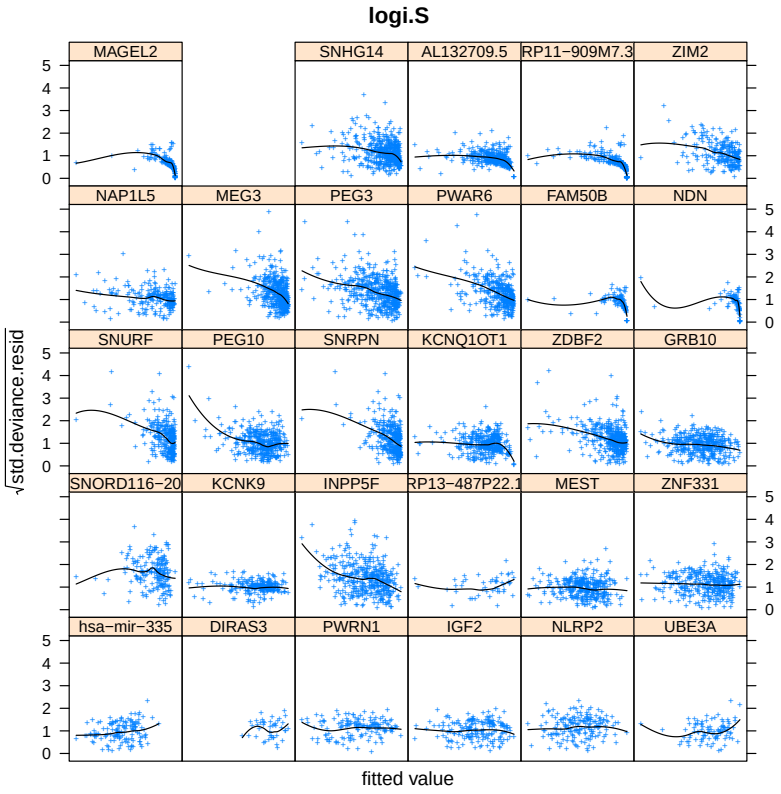
logi.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
√std.deviance.resid
fitted value

Figure S11:

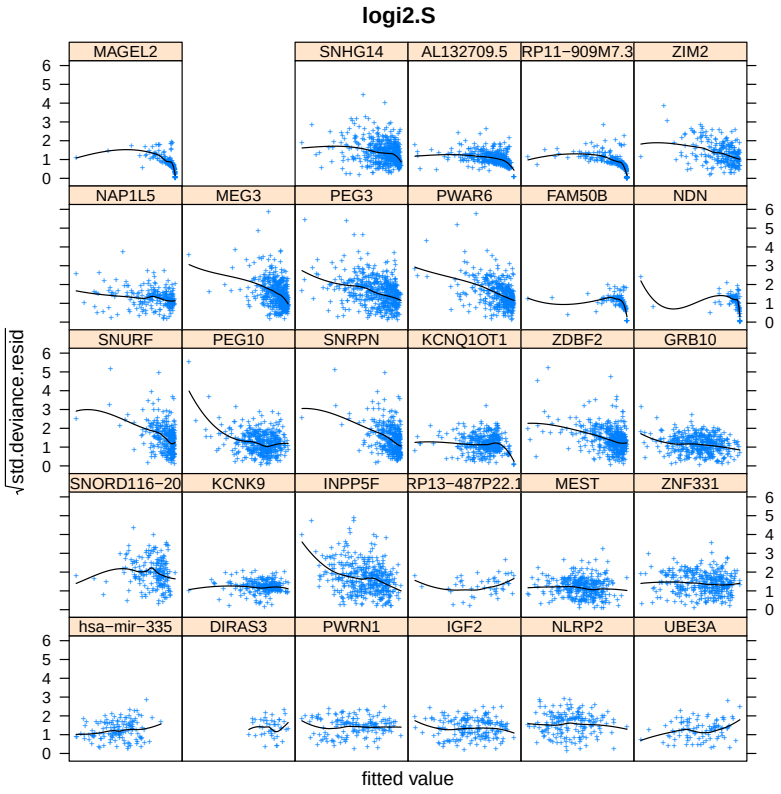
logi2.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
√std.deviance.resid
fitted value

Figure S12:

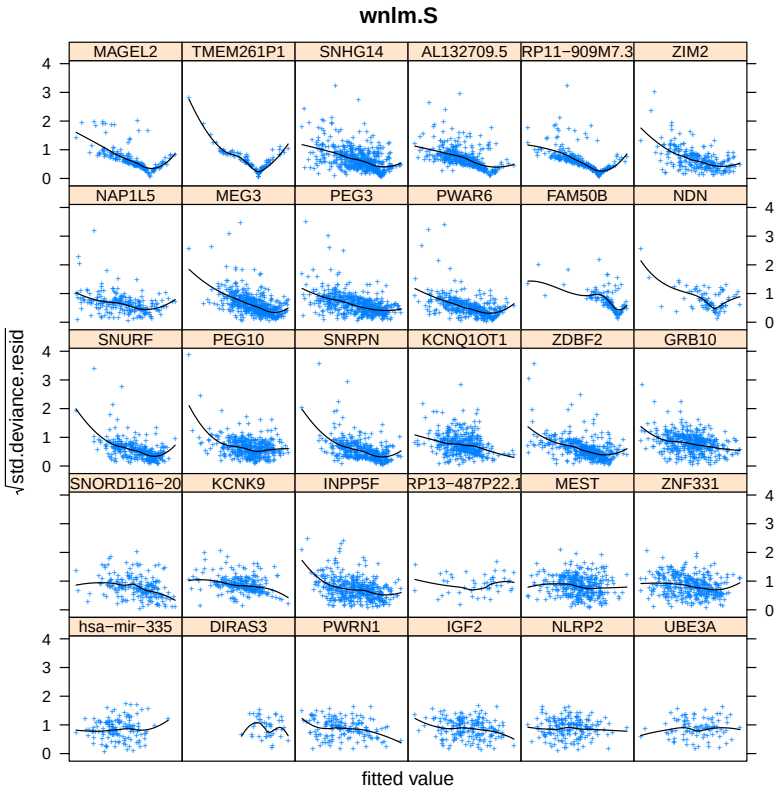
wnlm.S
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
√std.deviance.resid
fitted value

Figure S13:

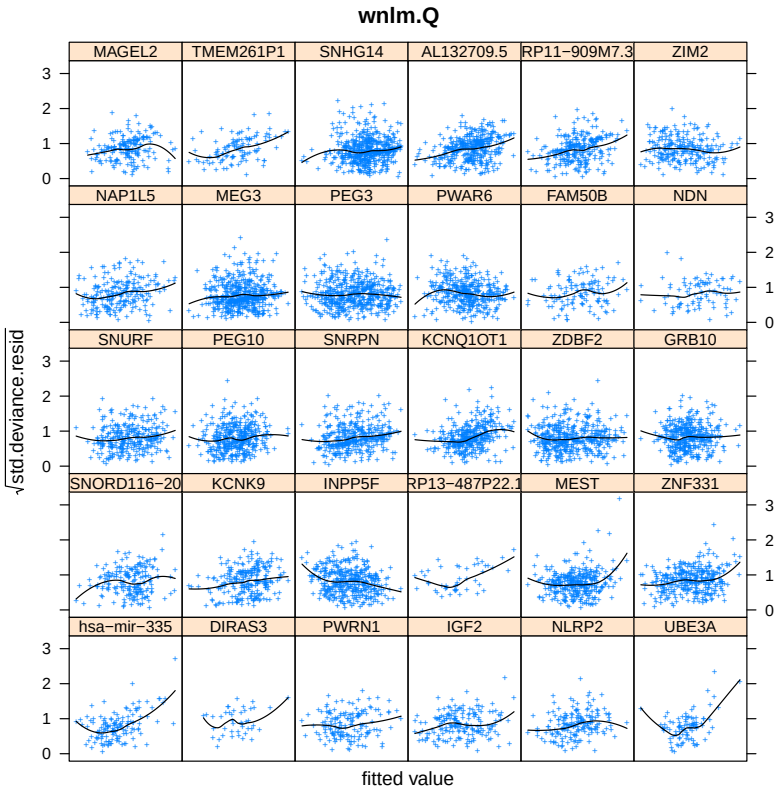
wnlm.Q
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
√|std.deviance.resid|
fitted value

Figure S14:

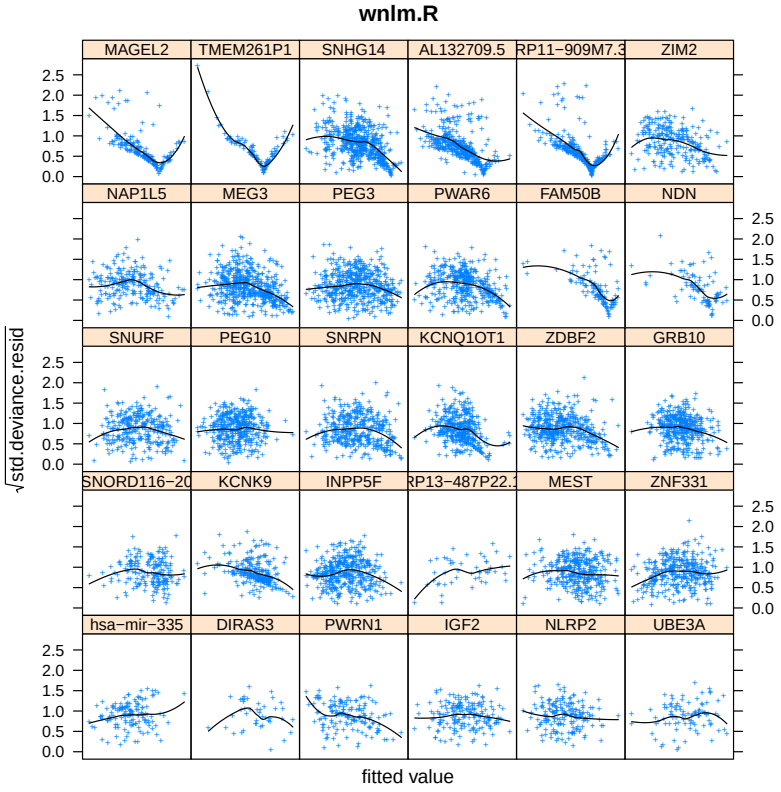
wnlm.R
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
√std.deviance.resid
fitted value

Figure S15:

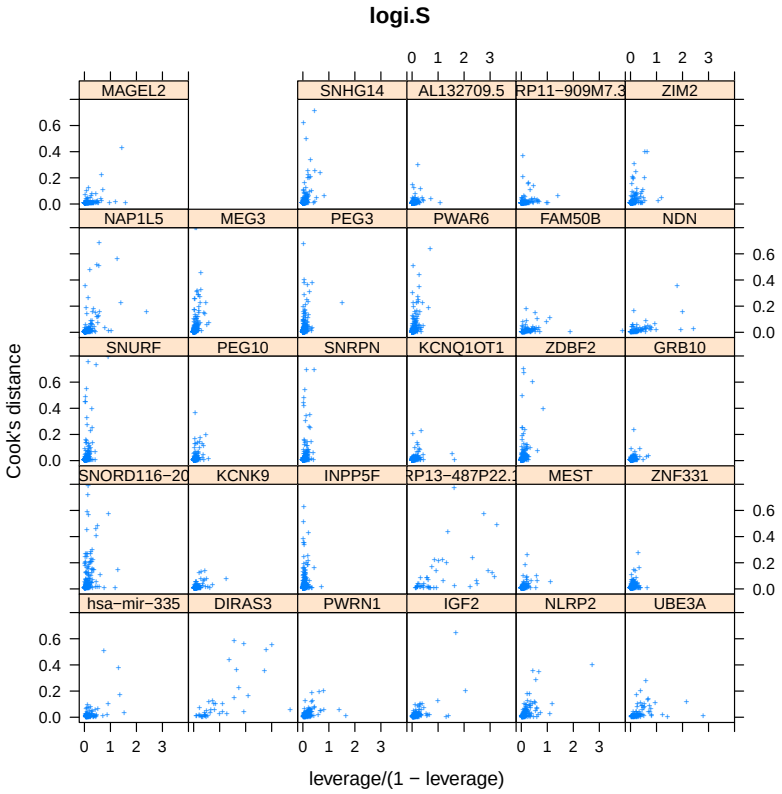
logi.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
Cook's distance
leverage/(1 − leverage)

Figure S16:

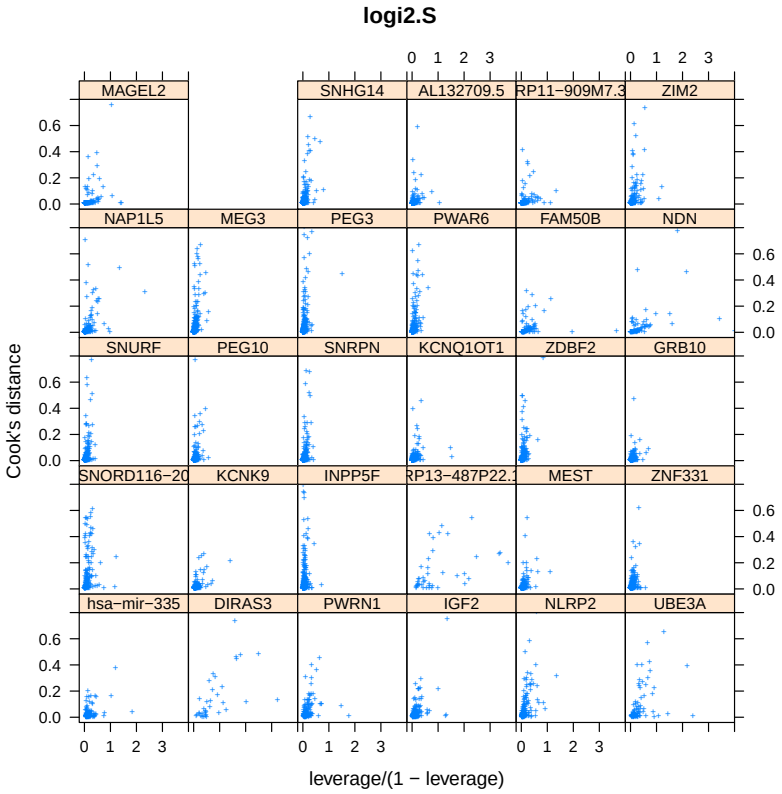
logi2.S
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
Cook's distance
leverage/(1 − leverage)

Figure S17:

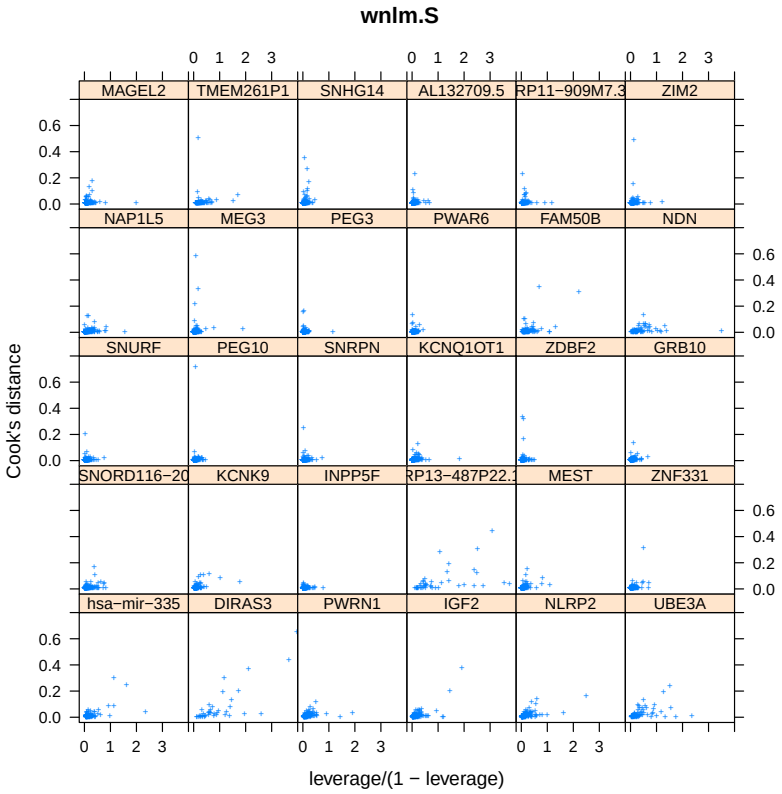
wnlm.S
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
Cook's distance
leverage/(1 − leverage)

Figure S18:

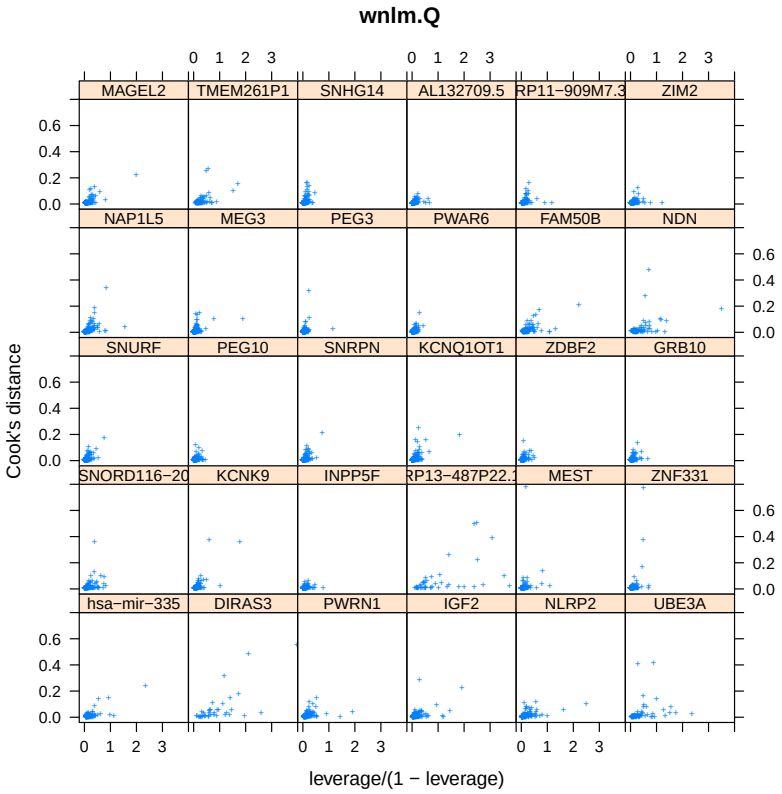
wnlm.Q
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
Cook's distance
leverage/(1 − leverage)

Figure S19:

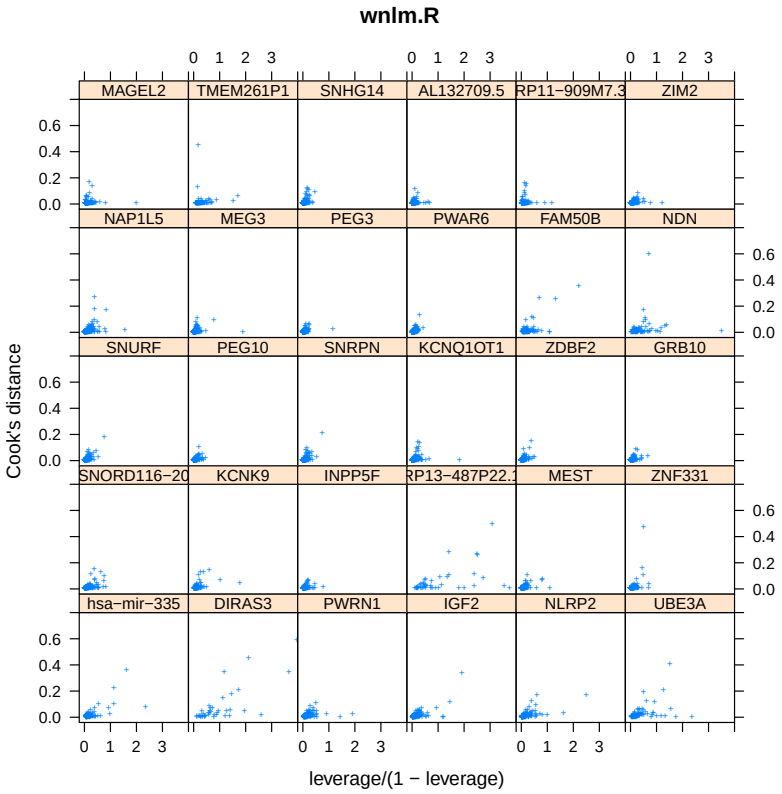
wnlm.R
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
Cook's distance
leverage/(1 − leverage)

Figure S20:

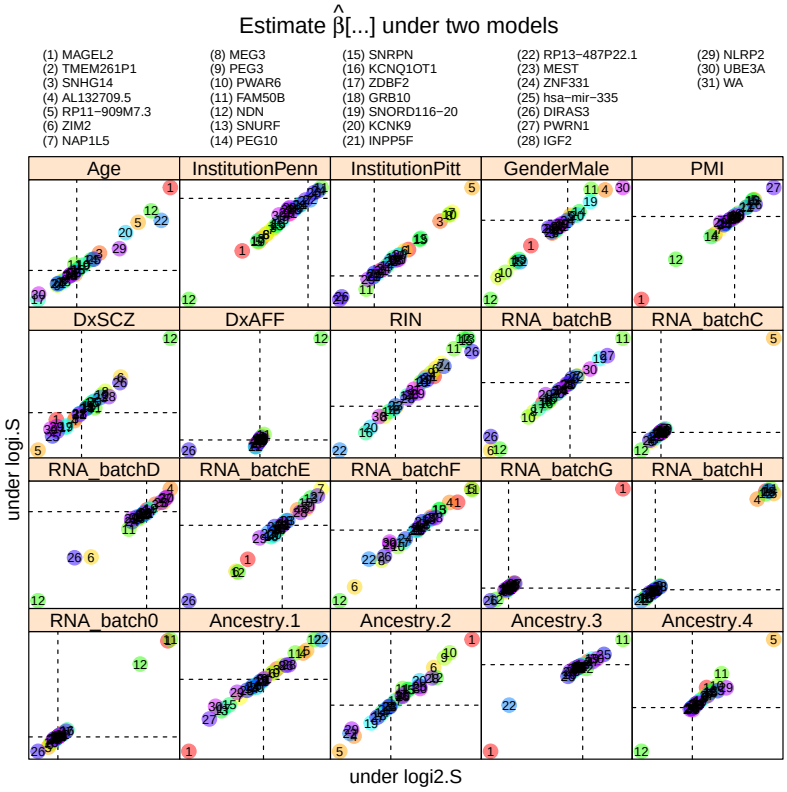
Estimate β̂[...] under two models
(1) MAGEL2
(2) TMEM261P1
(3) SNHG14
(4) AL132709.5
(5) RP11−909M7.3
(6) ZIM2
(7) NAP1L5
(8) MEG3
(9) PEG3
(10) PWAR6
(11) FAM50B
(12) NDN
(13) SNURF
(14) PEG10
(15) SNRPN
(16) KCNQ1OT1
(17) ZDBF2
(18) GRB10
(19) SNORD116−20
(20) KCNK9
(21) INPP5F
(22) RP13−487P22.1
(23) MEST
(24) ZNF331
(25) hsa−mir−335
(26) DIRAS3
(27) PWRN1
(28) IGF2
(29) NLRP2
(30) UBE3A
(31) WA
Age
InstitutionPenn
InstitutionPitt
GenderMale
PMI
DxSCZ
DxAFF
RIN
RNA_batchB
RNA_batchC
RNA_batchD
RNA_batchE
RNA_batchF
RNA_batchG
RNA_batchH
RNA_batch0
Ancestry.1
Ancestry.2
Ancestry.3
Ancestry.4
under logi.S
under logi2.S

Figure S21:

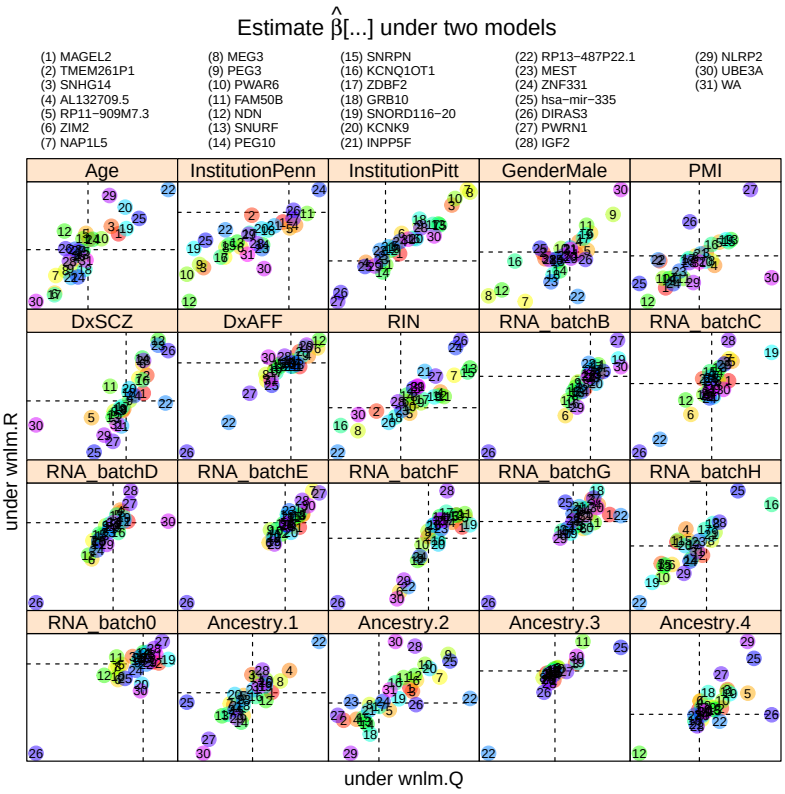
Estimate β̂[...] under two models
(1) MAGEL2
(2) TMEM261P1
(3) SNHG14
(4) AL132709.5
(5) RP11−909M7.3
(6) ZIM2
(7) NAP1L5
(8) MEG3
(9) PEG3
(10) PWAR6
(11) FAM50B
(12) NDN
(13) SNURF
(14) PEG10
(15) SNRPN
(16) KCNQ1OT1
(17) ZDBF2
(18) GRB10
(19) SNORD116−20
(20) KCNK9
(21) INPP5F
(22) RP13−487P22.1
(23) MEST
(24) ZNF331
(25) hsa−mir−335
(26) DIRAS3
(27) PWRN1
(28) IGF2
(29) NLRP2
(30) UBE3A
(31) WA
Age
InstitutionPenn
InstitutionPitt
GenderMale
PMI
DxSCZ
DxAFF
RIN
RNA_batchB
RNA_batchC
RNA_batchD
RNA_batchE
RNA_batchF
RNA_batchG
RNA_batchH
RNA_batch0
Ancestry.1
Ancestry.2
Ancestry.3
Ancestry.4
under wnlm.R
under wnlm.Q

Figure S22:

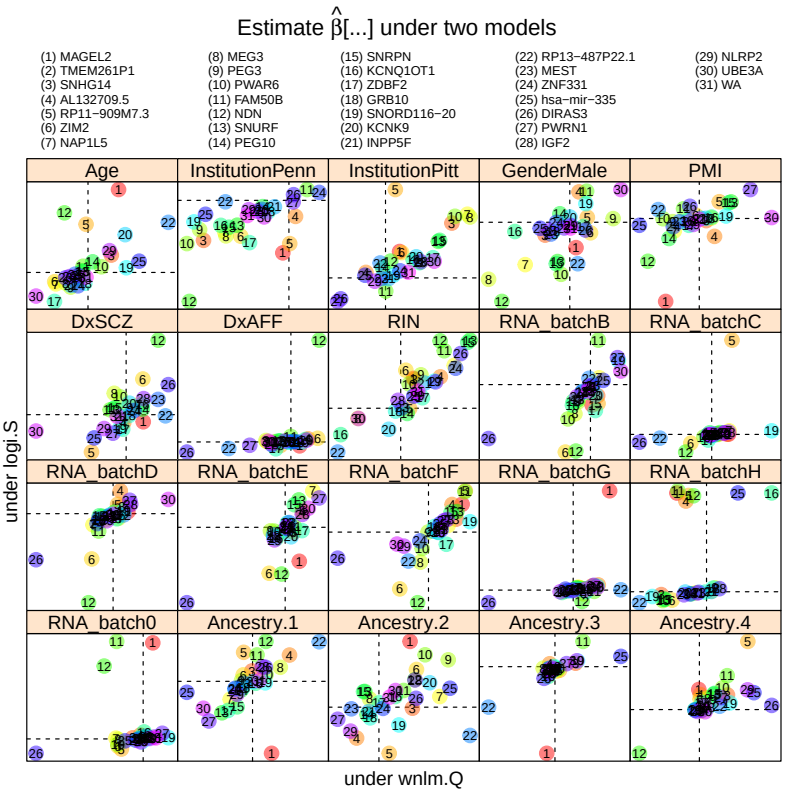
Estimate β̂[...] under two models
(1) MAGEL2
(2) TMEM261P1
(3) SNHG14
(4) AL132709.5
(5) RP11−909M7.3
(6) ZIM2
(7) NAP1L5
(8) MEG3
(9) PEG3
(10) PWAR6
(11) FAM50B
(12) NDN
(13) SNURF
(14) PEG10
(15) SNRPN
(16) KCNQ1OT1
(17) ZDBF2
(18) GRB10
(19) SNORD116−20
(20) KCNK9
(21) INPP5F
(22) RP13−487P22.1
(23) MEST
(24) ZNF331
(25) hsa−mir−335
(26) DIRAS3
(27) PWRN1
(28) IGF2
(29) NLRP2
(30) UBE3A
(31) WA
Age
InstitutionPenn
InstitutionPitt
GenderMale
PMI
DxSCZ
DxAFF
RIN
RNA_batchB
RNA_batchC
RNA_batchD
RNA_batchE
RNA_batchF
RNA_batchG
RNA_batchH
RNA_batch0
Ancestry.1
Ancestry.2
Ancestry.3
Ancestry.4
under logi.S
under wnlm.Q

Figure S23:

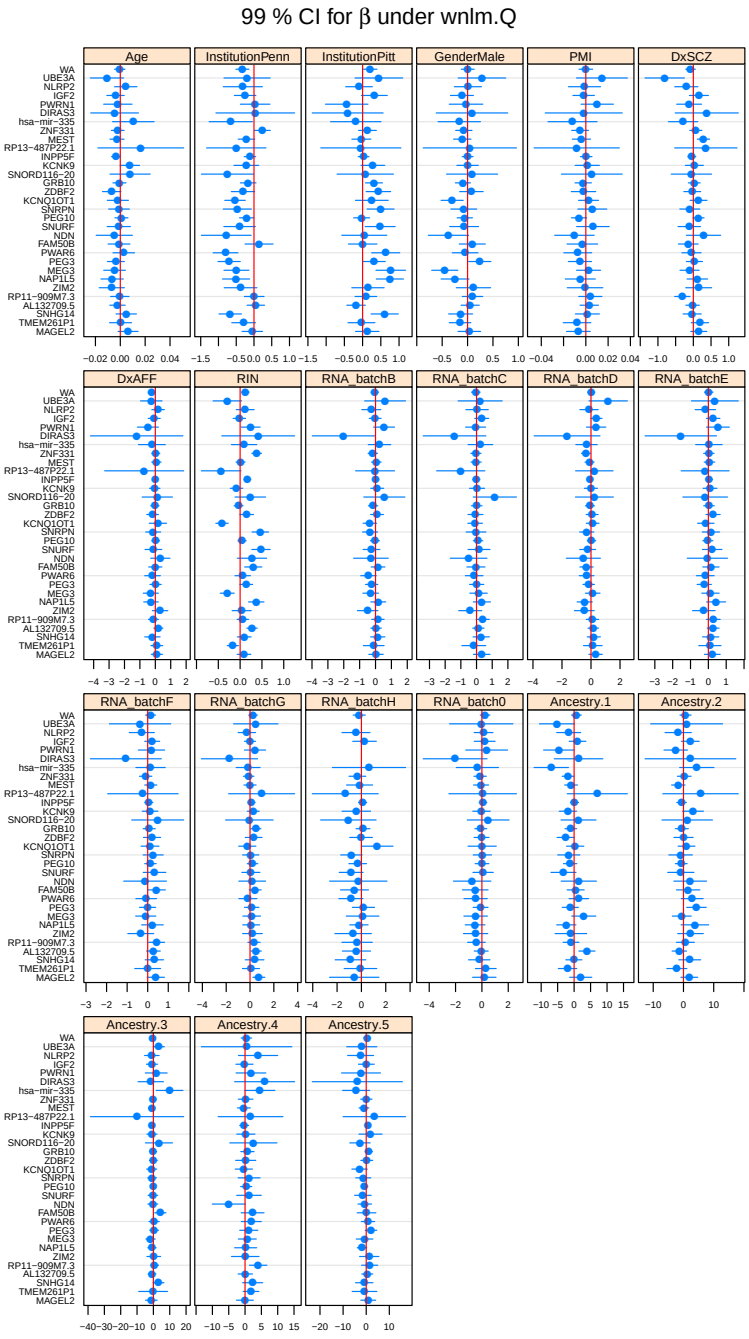
99 % CI for β under wnlm.Q
Age
InstitutionPenn
InstitutionPitt
GenderMale
PMI
DxSCZ
DxAFF
RIN
RNA_batchB
RNA_batchC
RNA_batchD
RNA_batchE
RNA_batchF
RNA_batchG
RNA_batchH
RNA_batch0
Ancestry.1
Ancestry.2
Ancestry.3
Ancestry.4
Ancestry.5
WA
UBE3A
NLRP2
IGF2
PWRN1
DIRAS3
hsa−mir−335
ZNF331
MEST
RP13−487P22.1
INPP5F
KCNK9
SNORD116−20
GRB10
ZDBF2
KCNQ1OT1
SNRPN
PEG10
SNURF
NDN
FAM50B
PWAR6
PEG3
MEG3
NAP1L5
ZIM2
RP11−909M7.3
AL132709.5
SNHG14
TMEM261P1
MAGEL2

Figure S24:

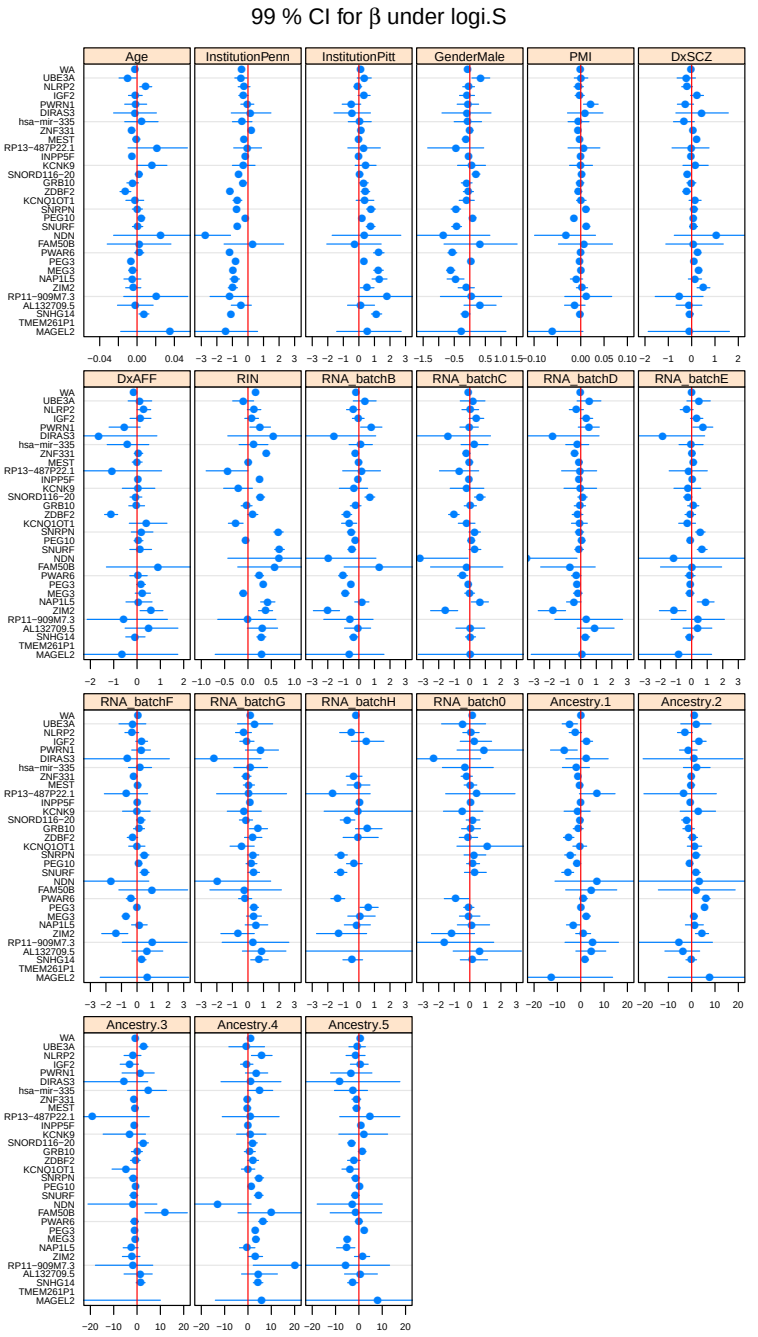
99 % CI for β under logi.S
Age
InstitutionPenn
InstitutionPitt
GenderMale
PMI
DxSCZ
DxAFF
RIN
RNA_batchB
RNA_batchC
RNA_batchD
RNA_batchE
RNA_batchF
RNA_batchG
RNA_batchH
RNA_batch0
Ancestry.1
Ancestry.2
Ancestry.3
Ancestry.4
Ancestry.5
WA
UBE3A
NLRP2
IGF2
PWRN1
DIRAS3
hsa−mir−335
ZNF331
MEST
RP13−487P22.1
INPP5F
KCNK9
SNORD116−20
GRB10
ZDBF2
KCNQ1OT1
SNRPN
PEG10
SNURF
NDN
FAM50B
PWAR6
PEG3
MEG3
NAP1L5
ZIM2
RP11−909M7.3
AL132709.5
SNHG14
TMEM261P1
MAGEL2

Figure S25:

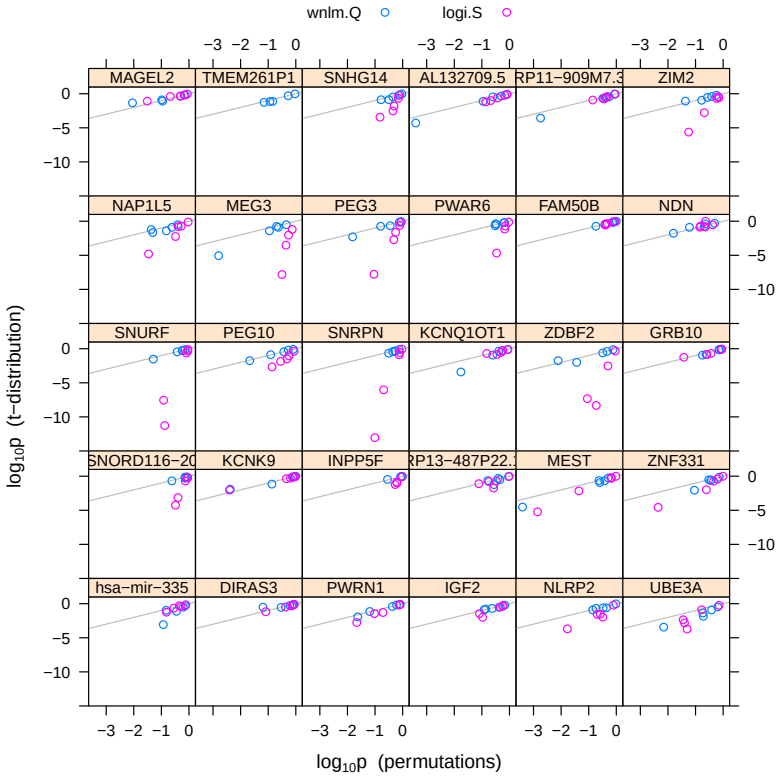
wnlm.Q
logi.S
MAGEL2
TMEM261P1
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
log10p (t−distribution)
log10p (permutations)

Figure S26:

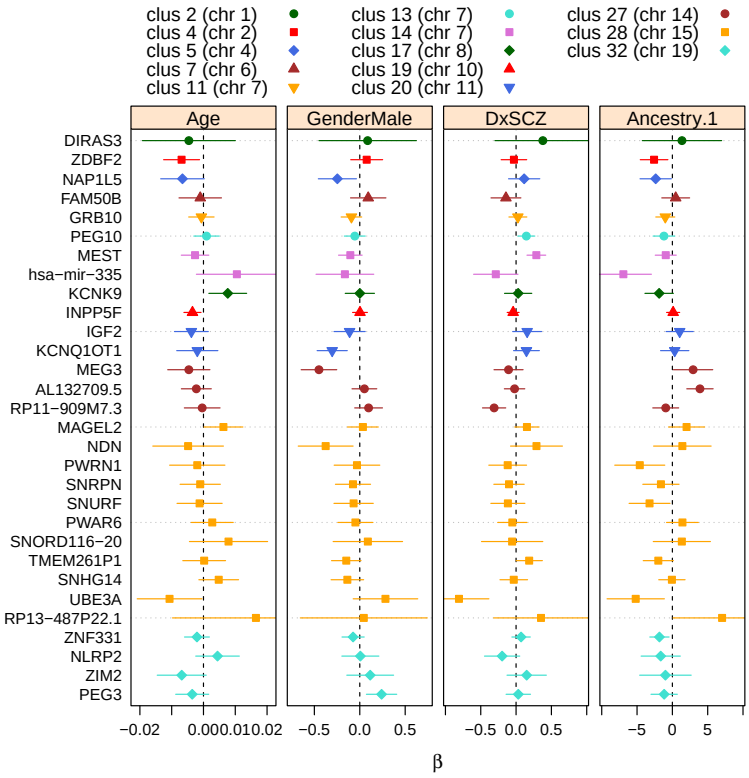

Figure S27: Biological effects: estimates and 95% confidence intervals under the wnlm.Q model

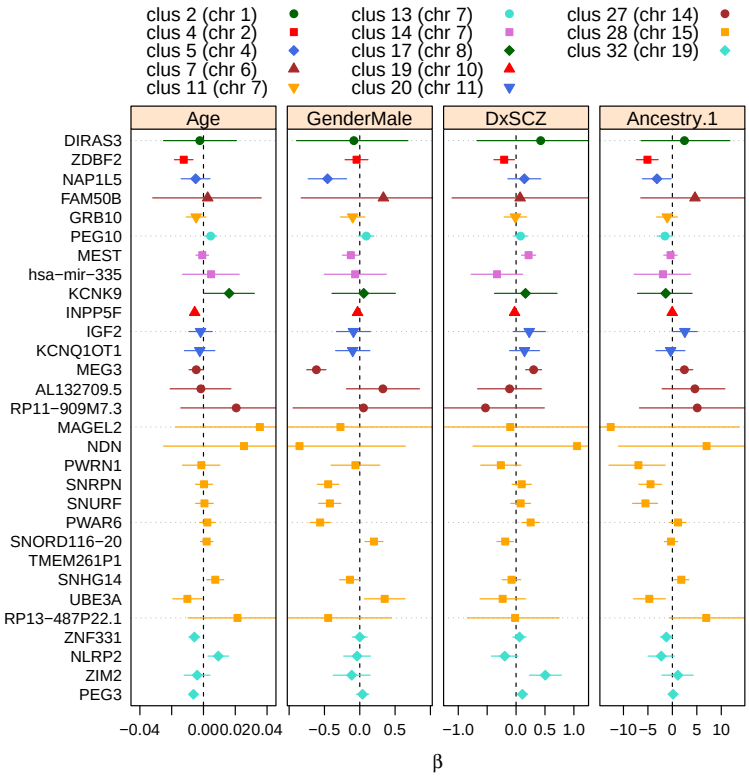

Figure S28: Biological effects: estimates and 99% confidence intervals under the logi.S model

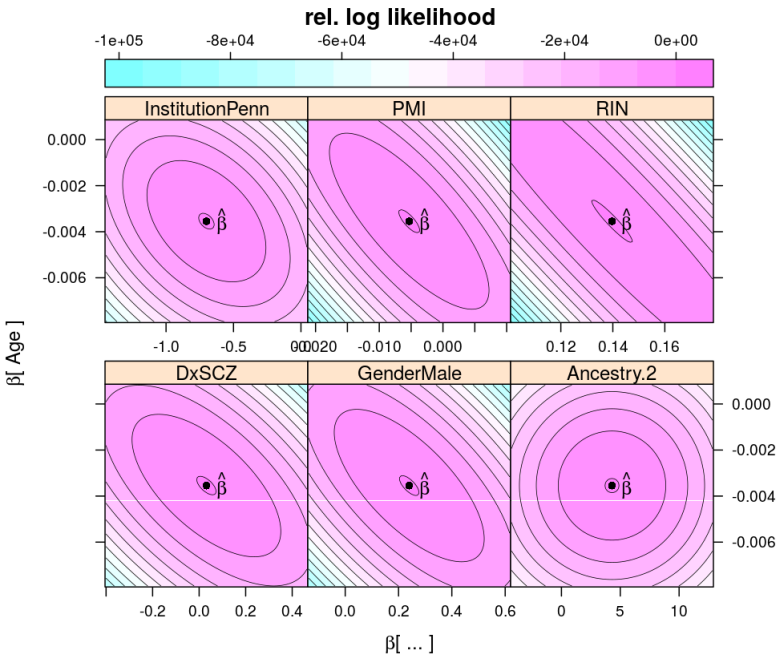
rel. log likelihood
-1e+05
-8e+04
-6e+04
-4e+04
-2e+04
0e+00
InstitutionPenn
PMI
RIN
DxSCZ
GenderMale
Ancestry.2
β̂
β[Age]
β[...]

Figure S29:

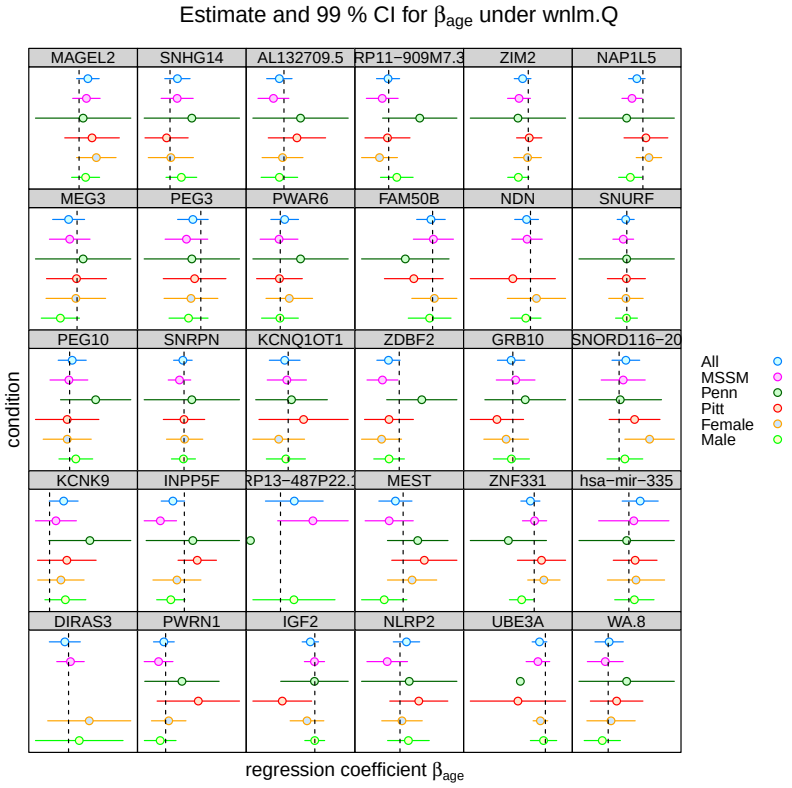
Estimate and 99 % CI for β_{age} under wnlm.Q
MAGEL2
SNHG14
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
RP13−487P22.1
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
WA.8
condition
All
MSSM
Penn
Pitt
Female
Male
regression coefficient β_{age}

Figure S30:

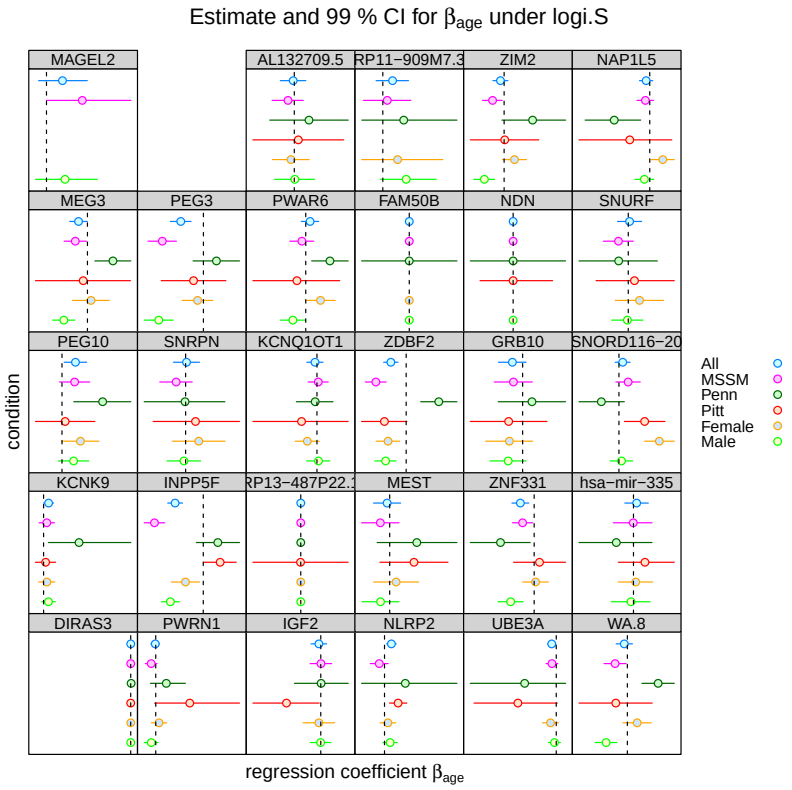
Estimate and 99 % CI for β_{age} under logi.S
MAGEL2
AL132709.5
RP11−909M7.3
ZIM2
NAP1L5
MEG3
PEG3
PWAR6
FAM50B
NDN
SNURF
PEG10
SNRPN
KCNQ1OT1
ZDBF2
GRB10
SNORD116−20
KCNK9
INPP5F
MEST
ZNF331
hsa−mir−335
DIRAS3
PWRN1
IGF2
NLRP2
UBE3A
WA.8
condition
regression coefficient β_{age}
All
MSSM
Penn
Pitt
Female
Male

Figure S31:

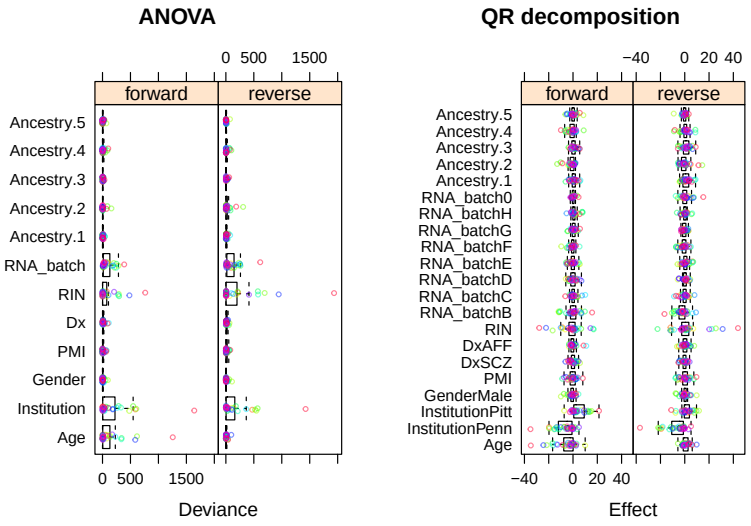
ANOVA
QR decomposition
forward
reverse
Ancestry.5
Ancestry.4
Ancestry.3
Ancestry.2
Ancestry.1
RNA_batch
RIN
Dx
PMI
Gender
Institution
Age
0 500 1500
Deviance
RNA_batch0
RNA_batchH
RNA_batchG
RNA_batchF
RNA_batchE
RNA_batchD
RNA_batchC
RNA_batchB
DxAFF
DxSCZ
GenderMale
InstitutionPitt
InstitutionPenn
−40 0 20 40
Effect

Figure S32: