Unperturbed Expression Bias of Imprinted Genes in Schizophrenics

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

Icahn School of Medicine at Mount Sinai

‡ equal contribution; * correspondence: andrew.chess@mssm.edu

Contents

1 Main text

How inter-individual differences in gene regulation correlates with disease is beginning to be examined through analyses of RNA-seq from post-mortem brains of individuals with schizophrenia and from control brains [5]. Here we focus on differences in allele-specific expression, following up on the CommonMind Consortium (CMC http://www.synapse.org/CMC) RNA-seq analyses of 579 human dorsolateral prefrontal cortex (DLPFC) samples. We find that the fraction of imprinted human genes is consistent with lower ($\approx 0.5\%$) [10, 4, 2] as opposed to higher [7] estimates in mice. The handful of novel potentially imprinted genes we find are all in close genomic proximity to known imprinted genes. Analyzing the extent of allelic expression bias—a hallmark of imprinting—across hundreds of individuals allowed us to examine its dependence on various factors. We find that allelic bias is independent of the diagnosis of schizophrenia. In contrast age up or down-regulates allelic bias of some imprinted genes and genetic ancestry also has an impact.

The observation [9, 11] that maternally derived microduplications at 15q11-q13—harboring the imprinted gene UBE3A—may not only cause Prader-Willi syndrome, but are also highly penetrant for schizophrenia has raised the possibility that perturbation of regulation of imprinted genes in general may play a role in psychotic disorders. As it is known that the extent of imprinting of individual genes varies over different tissues we chose the DLPFC region, which controls complex cognitive and executive functions and is known to display functional abnormalities in schizophrenia. We obtained pre-publication DLPFC RNA-seq data from the CMC and analyzed allele-specific expression with the idea of (i) identifying imprinted genes in the adult human brain and (ii) explaining the variability in allelic bias across 579 individuals in terms of their psychiatric diagnosis, age at death, etc. This was facilitated by the balanced case-control groups (258 SCZ, 267 Control, 54 bipolar or other affective/mood disorder, AFF) and the large age variability.

For each individual i and gene g we quantified allelic bias based on RNA-seq reads using a statistic called *read count ratio* S_{ig} (Fig. 1, Methods), which ranges from 0.5 to 1 indicating unbiased biallelic expression (at 0.5), some allelic bias (at intermediate values) or strictly monoallelic expression (at 1). We corrected for a number of factors this approach is known to be sensitive to. We quality-filtered RNA-seq reads and helped distinguish allele-specific reads using DNA genotyping data before calculating S and then applied post hoc corrections for mapping bias (Methods).

A total of 5307 genes passed our filters designed to remove genes with scarce RNA-seq data reflecting low expression and/or low coverage of RNA-seq. Fig. 2 presents the conditional empirical distribution of $S_{\cdot g}$ across all individuals given each gene g. The observed wide $S_{\cdot g}$ distributions suggest large across-individuals variation of allelic bias for all genes, even if a substantial component of the $S_{\cdot g}$ variation originates from technical sources. Still, as expected, for many genes known to be imprinted in mice or in other human tissues (referred to as *known imprinted* genes like PEG10, ZNF331) the distribution of $S_{\cdot g}$ was shifted to the right signaling strong allelic bias (Fig. 2, upper half).

To identify imprinted genes in the human adult DLPFC we defined the score of each gene g as the fraction of individuals i for whom $S_{ig} > 0.9$. We ranked all 5307 genes according to their score (Fig. 2 bottom right). The heat map of the $S_{\cdot g}$ distribution for ranked genes (Fig. 2, lower left) shows that the top 50 genes, which constitute $\approx 1\%$ of all genes in our analysis, are qualitatively different from the bottom $\approx 99\%$ exhibiting strongly right-shifted distribution of $S_{\cdot g}$ characteristic to imprinting. Consistently 29 of the top-scoring 50 genes fell into previously described imprinted gene clusters (Fig. S1).

Fig. 3 shows that 21 of the 29 within-cluster genes are *known imprinted* and 8 are not (termed

nearby candidates for imprinting); the 21 genes falling outside clusters are termed *distant candidates*. For 37 of the top 50 genes data was sufficient to test for mapping bias (Methods). The test results (see symbols to the right of gene names in Fig. 3) suggested that the strongly right-shifted $S_{\cdot g}$ distribution is due only to mapping bias for *distant candidates* but predominantly to imprinting for *known imprinted* and *nearby candidate* genes. This was further supported by the finding that only for *distant candidate* genes do many individuals display read count ratio consistent with nearly unbiased allelic expression (black bars in Fig. 3, Methods), which suggests that without the likely mapping bias in the remaining individuals the distribution of $S_{\cdot g}$ in these genes would be shifted to the left indicating no imprinting.

Based on these results we called all *known imprinted* and *nearby candidate* genes within the top 50 imprinted in the adult human DLPFC and included also the *known imprinted* gene UBE3A, which ranked below 50 but whose score was still substantial (Fig. S2). These 30 imprinted genes are listed in the panels of Fig. 4, which shows that read count ratio is similarly distributed in the Control, SCZ and AFF group suggesting independence between allelic bias and diagnosis of schizophrenia.

To confirm this rigorously we performed statistical inference using mixed effects models (Methods), which can capture much of the complex pattern of dependencies in genomic data including those we observed within and between technical and biological explanatory variables (Table S1, Fig. S3), and which gain power from letting genes "borrow strength from each other" (Fig. S4). We fitted many alternative mixed models, selected the best-fitting one using the Akaike Information Criterion (AIC) and confirmed the goodness of fit with diagnostic plots (Methods, Fig. S5, S6).

Based on the best fitting model (henceforth "the model") we could formally reject the hypotheses that read count ratio depends on diagnosis as either main effect or interaction (see term $(1 \mid \mathrm{Dx})$ and $(1 \mid \mathrm{Dx} : \mathrm{Gene})$ in Table 1, respectively). That this key result is not due to low power is indicated by the highly significant dependence of read count ratio on gene identity (see $(1 \mid \mathrm{Gene})$ in Table 1 and compare panels in Fig. 4).

Fig. 5 suggests that the read count ratio depends negatively on age for some imprinted genes, depends positively for others, and is independent of age for the rest of imprinted genes. Yhis apparent dependence might be indirect, i.e. one that is mediated by some variable(s) "inbetween" age and read count ratio (Fig. S3, S4) but the model allowed us to isolate the direct component of age dependence: we found that the gene-specific random age effect is indeed significant even if no fixed effect—which would be shared by all imprinted genes—was supported (see $(\mathrm{Age} \mid \mathrm{Gene})$ and Age, respectively, in Table 1).

Based on the model we also predicted gene-specific regression coefficients mediating the direct component of age effect (Fig. S7 top middle). The predicted coefficients agreed well with all but a few panels of Fig. 5 the latter of which (e.g. UBE3A) therefore represent purely indirect dependence.

The same type of analysis on the effects of ancestry principal components and gender gave similar results: while the fixed effect, shared by all genes, of these variables was negligible, three of the random, gene-specific, effects received significant support. These three, ordered by decreasing statistical significance, are $(\mathrm{Ancestry.1} \mid \mathrm{Gene})$, $(\mathrm{Ancestry.3} \mid \mathrm{Gene})$ and $(1 \mid \mathrm{Gender} : \mathrm{Gene})$ (Table 1). The corresponding predicted random coefficients are presented in Fig. S7.

In summary age, ancestry, and to a lesser extent gender, are suggested by our model-based analysis to exert effect on allelic bias in a way that the direction and magnitude of the effect varies across genes.

The number of imprinted genes in the mammalian brain has been controversial: some early genome wide studies [7, 6] estimated over a thousand, suggesting that the number of imprinted

genes in the brain is an order of magnitude greater than in other tissues. Later work cast doubt on the methodology used and found that the number of imprinted genes in brain is in line with expectations from studies of other tissues, identifying only a handful of new candidate imprinted genes in brain [10, 4, 2]. Based on 579 postmortem human DLPFC samples we find evidence supporting only a handful of novel imprinted genes all of which reside in genomic locations nearby to known imprinted genes. Thus our results support those more recent studies that found no large excess of imprinted genes in the brain.

We have performed the most in-depth analysis up till now on the how imprinting depends on schizophrenia, as well as on age and other variables. This was made possible by (i) more study individuals than in previous work [2] distributed in well-balanced case-control study groups, and (ii) powerful statistical inference based on a mixed model. Despite these advantages, technical variation is still large and consequently so is the uncertainty of our statistical inferences.

Although our approach gave strong support for dependence of imprinting on age and ancestry, no dependence on schizophrenia was detected either when we assumed that the dependence is the same for all imprinted genes or that it varies across genes. This might seem to suggest that imprinting in the DLFPC plays no significant role in schizophrenia and psychotic disorders contradicting the "imprinted brain" hypothesis [3]. Alternatively, imprinting does play a role in schizophrenia but only very strong perturbations of some imprinted genes increase the risk significantly, perturbations that are too rare to detect in even at our relatively large sample size. Additionally, the more subtle perturbations in our data might still have a significant effect when considered together with other genetic, epigenetic or environmental risk factors that were absent in our model. The complex genetic architecture of schizophrenia [12] makes these alternative explanations quite plausible.

We found that imprinting depends on ancestry in a gene specific manner but the type of dependence that is shared by all imprinted genes was not supported. This is expected because the studied ancestry variables must incorporate some of the cis expression QTLs in imprinted genes such that those eQTLS perturb allelic bias in a gene specific manner.

Our discovery that imprinting depends on age in later adulthood is rather intriguing. Age dependence during embryonic development and childhood is both well-supported experimentally and well-understood but that during later adulthood has so far only been predicted [13] based on a hypothesis that links "genomic imprinting and the social brain" [8]. Previous genomics studies [2] were statistically underpowered to address this question in humans and the only experimental hints were gained from young mice [10]. Although our age-related finding supports the "social brain" hypothesis, it leaves the possibility open that the observed age related changes indicate merely the loss of tight regulation of those genes after because they loose their functional significance in aging.

2 Methods

2.1 Defining the read count ratio to quantify allelic bias

We quantified allelic bias based on RNA-seq reads using a statistic called *read count ratio* S, whose definition we based on the total read count T and the *higher read count* H, i.e. the count of reads carrying only either the reference or the alternative SNP variant, whichever is higher. The definition is

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

where i identifies an individual, g a gene, and the summation runs over all SNPs s for which gene g is heterozygous in individual i (Fig. 1). Note that if B_{ig} is the count or reads that map to the b_{ig} allele (defined as above) and if we make the same distributional assumption as above, namely that $B_{ig} \sim \text{Binom}(p_{ig}, T_{ig})$, then $\Pr(H_{ig} = B_{ig}|p_{ig})$, the probability of correctly assigning the reads with the higher count to the allele towards which expression is biased, tends to 1 as $p_{ig} \to 1$. We took advantage of this theoretical result in that we subjected only those genes to statistical inference, whose read count ratio was found to be high and, therefore, whose p_{ig} is expected to be high as well.

Fig. 1 illustrates the calculation of S_{ig} for the combination of two hypothetical genes, g_1, g_2, and two individuals, i_1, i_2. It also shows an example for the less likely event that the lower rather than the higher read count corresponds to the SNP variant tagging the higher expressed allele (see SNP s_3 in gene g_1 in individual i_2).

Before we carried out our read count ratio-based analyses, however, we cleaned our RNA-seq data by quality-filtering and by improving the accuracy of SNP calling with the use of DNA SNP array data and imputation. In the following subsections of Methods we describe the data, these procedures, as well as our regression models in detail.

2.2 Brain samples, RNA-seq

Human RNA samples were collected from the dorsolateral prefrontal cortex of the CommonMind consortium 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. RNA-seq was performed on Illumina HiSeq 2000.

2.3 Mapping, SNP calling and filtering

We mapped 100bp, paired-end RNA-seq reads ($\approx$ 50 million reads per sample) using Tophat to Ensembl gene transcripts of the human genome (hg19; February, 2009) with default parameters and 6 mismatches allowed per pair (200 bp 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. For the 579 samples there were 203 million reads overlapping one of the 364,509 SNPs defined above. Of those 158 million (78%) had genotype data available from either SNP array or imputation.

2.4 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. We calibrated the imputation parameters to find a reasonable balance between the number of genes assessable for allelic bias and the number false positive calls since the latter can arise if a SNP is incorrectly called heterozygous.

We first examined how many SNPs were heterozygous in DNA calls and had a discordant RNA call (i.e. homozygous SNP call from RNA-seq) using different imputation parameters. Known imprinted genes were excluded. We examined RNA-seq reads overlapping array-called heterozygous SNPs which we assigned a heterozygosity score L_{het} of 1, separately from RNA seq data overlapping imputed heterozygous SNPs, where the L_{het} score could range from 0 to 1. After testing different thresholds we selected an 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.

The higher rate of discordance for the imputed SNPs is due to imputation error. These were taken into account in two ways. First, we considered all imputed SNPs for a gene g and individual i jointly. Second, we excluded any individual, for which one or more SNPs supported biallelic expression.

2.5 Quality filtering

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 the read count threshold t_{rc} was set to 15. The individual-based filter removes any genes g (across all individuals) if read count data involving g are available for less than t_{ind} number of individuals, set to 25. These final filtering procedures decreased the number of genes in the data from 15584 to $n = 5307$.

2.6 Test and post hoc correction for mapping bias

We performed a test to see if there is mapping bias for the top ranking genes (Fig. 3). For any gene (known imprinted, or candidate) the expectation is that when some allelic bias is detected, that should equally favor the reference or non-reference allele since for a given individual who is

heterozygous at a given SNP in the genome it is reasonable to assume that the chances are equal that the mother or that the father has the reference allele. The basis of the test is that for genes unaffected by mapping bias the distribution of the number of individuals with reference allele is expected to be binomial with probability parameter 0.5. The results of this and the following test guided our post hoc correction that aims to remove certain genes despite their high score reflecting right-shifted distribution of read count ratio.

2.7 Testing nearly unbiased allelic expression

The null hypothesis of this test is that the higher read count $H_{ig} = S_{ig}T_{ig}$ for gene g and individual i is drawn from a binomial distribution with a probability parameter $p_{ig} \approx 0.5$ suggesting nearly unbiased allelic expression. More specifically, the test was defined by the criteria

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

where the 95% upper confidence limit UCL_{ig} for the expected read count ratio p_{ig} was calculated assuming that the higher read count $H_{ig} \sim \mathrm{Binom}(p_{ig}, T_{ig})$, on the fact that binomial random variables are asymptotically (as $T_{ig} \to \infty$) normal with $\mathrm{var}(H_{ig}) = T_{ig}p_{ig}(1 - p_{ig})$, and on the equalities $\mathrm{var}(S_{ig}) = \mathrm{var}(H_{ig}/T_{ig}) = \mathrm{var}(H_{ig})/T_{ig}^2$. Therefore

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

where z_p is the p quantile of the standard normal distribution.

2.8 Mixed and fixed regression models

We modeled the dependence of read count ratio of imprinted genes on biological and technical explanatory variables (Table S1) using mixed and fixed generalized linear models (GLMs).

GLMs in general describe a conditional distribution of a response variable y given a linear predictor η such that the distribution is from the exponential family and that $\mathrm{E}(y|\eta) = g^{-1}(\eta)$, where g is some link function. In the present context the response y is the observed read count ratio that is possibly transformed to improve the model's fit to the data. We performed fitting with the lme4 and stats R packages and tested several combinations of transformations, link functions, and error distributions (Table S2). For inference we used the best fitting combination (unlm.Q, Table S2) as assessed by the normality (Fig. S5) and homoscedasticity (Fig. S6) of residuals as well as by monitoring convergence.

In mixed GLMs the linear predictor $\eta = X\beta + Zb$ and in fixed GLMS $\eta = X\beta$, where X, Z are design matrices containing data on explanatory variables whereas β and b are fixed and random vectors of regression coefficients that mediate fixed and random effects, respectively (see Section 2.9 and Fig. S4 for details).

Besides the random effects term Zb the key difference between the mixed and fixed models in this study is that the former describes read count ratio *jointly* for all imprinted genes and the latter *separately* for each imprinted gene. An important consequence is that our mixed models are more powerful because they can utilize information shared by all genes. Therefore we preferred mixed models for final inference and used fixed models only to guide selection among possible mixed models (Section 2.10).

2.9 Formulation and interpretation of mixed models

Here we describe the detailed syntax and semantics of the normal linear mixed models combined with a quasi-log transformation Q of read count ratio as this combination was found to provide the best fit (Fig. S5, S6). We have data on 579 individuals and 30 imprinted genes and so the response vector is $y = (Q_{i_1 g_1}, ..., Q_{i_{579} g_1}, Q_{i_1 g_2}, ..., Q_{i_{579} g_2}, ..., Q_{i_1 g_{30}}, ..., Q_{i_{579} g_{30}})$. The model in matrix notation is

$$\begin{aligned} y &= X\beta + Zb + \varepsilon && (4)\\ \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, ..., mn && (5)\\ b &\sim \mathcal{N}(0, \Omega_b), && (6) \end{aligned}$$

where the size of the covariance matrix Ω_b depends on the number of terms with random effects (the columns of Z). Simply put: errors and random coefficients are all normally distributed.

To clarify the semantics of Eq. 4 let us consider a simple toy model with just a few terms in the linear predictor. But before expressing it in terms of Eq. 4 it is easier to cast it in the compact "R formalism" of the stats and lme4 packages of the R language as

$$y \sim \overbrace{1 + \text{Age}}^{\text{fixed effect}} + \overbrace{\underbrace{(1 + \text{Age} + \text{Ancestry.1} \,|\, \text{Gene})}_{k=1} + \underbrace{(1 \,|\, \text{Dx} : \text{Gene})}_{k=2}}^{\text{random effects}}. \quad (7)$$

First note that the random effect term labeled with $k = 1$ can be expanded into $(1 \,|\, \text{Gene}) + (\text{Age} \,|\, \text{Gene}) + (\text{Ancestry.1} \,|\, \text{Gene})$. The '1's mean intercept terms: one as a fixed effect and two as random effects. The first random intercept term $(1 \,|\, \text{Gene})$ expresses the gene-to-gene variability in read count ratio (compare panels in Fig. 4 and 5), in other words the random effect of the Gene variable. The second random intercept term $(1 \,|\, \text{Dx} : \text{Gene})$ corresponds to the interaction between psychiatric diagnosis Dx and Gene; it can be interpreted as the Gene specific effect of Dx or—equivalently—as Dx specific gene-to-gene variability. This term is not likely to be informative as Fig. 4 suggests little Gene specific effect of Dx.

We see that Age appears twice: first as a fixed slope effect on y and second as a Gene specific random slope effect, denoted as $(\text{Age} \,|\, \text{Gene})$. The random effect appears to be supported by Fig. 5 because the dependence of read count ratio on Age varies substantially among genes but the fixed effect is not supported because the negative dependence seen for several genes is balanced out by the positive dependence seen for others. The model includes another random slope effect: $(\text{Ancestry.1} \,|\, \text{Gene})$ with a similar interpretation as $(\text{Age} \,|\, \text{Gene})$ but lacks a fixed effect of Ancestry.1.

Now we are ready to write the toy model as an expanded special case of Eq. 4 as

$$y_i = \overbrace{\beta_0 + \text{Age}_i \beta_1}^{\text{fixed effects}} + \overbrace{\underbrace{b_0^{(1)} + \text{Age}_i b_1^{(1)} + \text{Ancestry.1}_i b_2^{(1)}}_{\text{Gene}_i} + \underbrace{b_0^{(2)}}_{\text{Dx}_i:\text{Gene}_i}}^{\text{random effects}} + \varepsilon_i. \quad (8)$$

As in the earlier R formalism the terms of the linear predictor are grouped into fixed and random effects. Within the latter group we have two batches of terms indicated by the k superscripts on the random regression coefficients $b_j^{(k)}$. The first batch $\{b_0^{(1)}, b_1^{(1)}, b_2^{(1)}\}$ corresponds to

$\{(1 \,|\, \text{Gene}), (\text{Age} \,|\, \text{Gene}), (\text{Ancestry.1} \,|\, \text{Gene})\}$ in Eq. 7, the second batch contains only $b_0^{(2)}$ corresponding to $(1 \,|\, \text{Dx} : \text{Gene})$.

Within the kth batch Eq. 8 contains only a single intercept coefficient $b_0^{(k)}$ and, if random slope terms are also present in the batch, only a single slope coefficient associated with the variable Age or Ancestry.1. This is because only a single level of the factor Gene or the composite factor Dx : Gene needs to be considered for the ith observation; these levels are denoted as Gene_i and $\text{Dx}_i : \text{Gene}_i$, respectively. Implicitly however, Eq. 8 contains the respective coefficients for all levels of these factors. For example, there are $n = 30$ intercept coefficients $b_j^{(1)}$ each of which corresponds to a given gene. So to generalize Eq. 8 we need J_k coefficients in the kth batch, where J_k is the product of the number of factor levels and one plus the number of random slope variables. This way we can provide the expansion of the general formula Eq. 4 using the semantics of the toy model (Eq. 7, 8) as

$$y_i = \overbrace{\sum_{j=0}^{J} x_{ij}\beta_j}^{\text{fixed effects}} + \overbrace{\sum_{k=1}^{K}\sum_{j=0}^{J_k} z_{ij}^{(k)} b_j^{(k)}}^{\text{random effects}} + \varepsilon_i. \tag{9}$$

2.10 Model fitting and selection

Eq. 9 describes a large set of mixed models that differ in one or more individual terms that constitute their linear predictor. From this set we aimed to select the best fitting model under the Akaike Information Criterion (AIC).

We used a heuristic search strategy in order to restrict the vast model space to a relatively small subset of plausible models. The search was started at a model whose relatively simple linear predictor was composed of terms using our prior results based on fixed effects models. The same results suggested a sequence in which further terms were progressively added to the model to test if they improve fit. Improvement was assessed by ΔAIC and the χ^2-test on the degrees of freedom that correspond the evaluated term. If fit improved the term was added otherwise it was omitted. Next, further terms were tested. This iterative procedure lead to the following model.

$$\begin{aligned} Q \;&\sim\; \text{RIN} + (1 \,|\, \text{RNA\_batch}) + (1 \,|\, \text{Institution}) + (1 \,|\, \text{Institution} : \text{Individual}) \\ &+\; (1 \,|\, \text{Gene} : \text{Institution}) + (1 \,|\, \text{Gender} : \text{Gene}) \\ &+\; (\text{Age} + \text{RIN} + \text{Ancestry.1} + \text{Ancestry.3} \,|\, \text{Gene}) \end{aligned}$$

We refer to this as the "best fitting model" even thought it may represent only a local optimum in model space.

3 References

[1] Yu Bai, Min Ni, Blerta Cooper, Yi Wei, and Wen Fury. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*, 15(1):325, may 2014.

[2] Yael Baran, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K. Tsang, Manuel A. Rivas, Matti Pirinen, Maria Gutierrez-Arcelus, Kevin S. Smith, Kim R. Kukurba, Rui Zhang, Celeste Eng, Dara G. Torgerson, Cydney Urbanek, Jin Billy Li, Jose R. Rodriguez-Santana, Esteban G. Burchard, Max A. Seibold, Daniel G. MacArthur, Stephen B. Montgomery, Noah A. Zaitlen, and Tuuli Lappalainen. The landscape of genomic imprinting across diverse adult human tissues. *Genome Research*, 25(7), 2015.

[3] Bernard Crespi. Genomic imprinting in the development and evolution of psychotic spectrum conditions. *Biological reviews of the Cambridge Philosophical Society*, 83(4):441–93, nov 2008.

[4] 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.

[5] Menachem Fromer, Panos Roussos, Solveig K Sieberts, Jessica S Johnson, David H Kavanagh, Thanneer M Perumal, Douglas M Ruderfer, Edwin C Oh, Aaron Topol, Hardik R Shah, Lambertus L Klei, Robin Kramer, Dalila Pinto, Zeynep H Gümü, A Ercument Cicek, Kristen K Dang, Andrew Browne, Cong Lu, Lu Xie, Ben Readhead, Eli A Stahl, Jianqiu Xiao, Mahsa Parvizi, Tymor Hamamsy, John F Fullard, Ying-Chih Wang, Milind C Mahajan, Jonathan M J Derry, Joel T Dudley, Scott E Hemby, Benjamin A Logsdon, Konrad Talbot, Towfique Raj, David A Bennett, Philip L De Jager, Jun Zhu, Bin Zhang, Patrick F Sullivan, Andrew Chess, Shaun M Purcell, Leslie A Shinobu, Lara M Mangravite, Hiroyoshi Toyoshiba, Raquel E Gur, Chang-Gyu Hahn, David A Lewis, Vahram Haroutunian, Mette A Peters, Barbara K Lipska, Joseph D Buxbaum, Eric E Schadt, Keisuke Hirai, Kathryn Roeder, Kristen J Brennand, Nicholas Katsanis, Enrico Domenici, Bernie Devlin, and Pamela Sklar. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, sep 2016.

[6] Christopher Gregg, Jiangwen Zhang, James E. Butler, David Haig, and Catherine Dulac. Sex-Specific Parent-of-Origin Allelic Expression in the Mouse Brain. *Science*, 329(5992):682–685, aug 2010.

[7] 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.

[8] Anthony R Isles, William Davies, and Lawrence S Wilkinson. Genomic imprinting and the social brain. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1476):2229–37, dec 2006.

[9] Abdul Noor, Lucie Dupuis, Kirti Mittal, Anath C. Lionel, Christian R. Marshall, Stephen W. Scherer, Tracy Stockley, John B. Vincent, Roberto Mendoza-Londono, and Dimitri J. Stavropoulos. 15q11.2 duplication encompassing only the UBE3a gene is associated with developmental delay and neuropsychiatric phenotypes. 36(7):689–693.

[10] 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.

[11] Elliott Rees, James T R Walters, Lyudmila Georgieva, Anthony R Isles, Kimberly D Chambert, Alexander L Richards, Gerwyn Mahoney-Davies, Sophie E Legge, Jennifer L Moran, Steven A McCarroll, Michael C O'Donovan, Michael J Owen, and George Kirov. Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry : the journal of mental science*, 204(2):108–14, feb 2014.

[12] Patrick F Sullivan, Mark J Daly, and Michael O'Donovan. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, 13(8):537–551, aug 2012.

[13] Francisco Ubeda and Andy Gardner. A model for genomic imprinting in the social brain: elders. *Evolution; international journal of organic evolution*, 66(5):1567–81, may 2012.

[14] X Zheng, J Shen, C Cox, J C Wakefield, M G Ehm, M R Nelson, and B S Weir. HIBAG–HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, 14(2):192–200, apr 2014.

[15] Lillian M Zwemer, Alexander Zak, Benjamin R Thompson, Andrew Kirby, Mark J Daly, Andrew Chess, and Alexander A Gimelbrant. Autosomal monoallelic expression in the mouse. *Genome Biology*, 13(2):R10, feb 2012.

4 Acknowledgements

5 Author information

5.1 Affiliation

Attila Gulyás-Kovács[a]
Ifat Keydar[ab]
...
Andrew Chess[ac]

a Department of Genetics and Genomic Sciences, Department of Developmental and Regenerative Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029-6574

b current address: TODO

c Fishberg Department of Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029-6574

5.2 Contributions

5.3 Competing financial interests

None

5.4 Corresponding author

Andrew Chess
email: andrew.chess@mssm.edu

6 Figures with legends

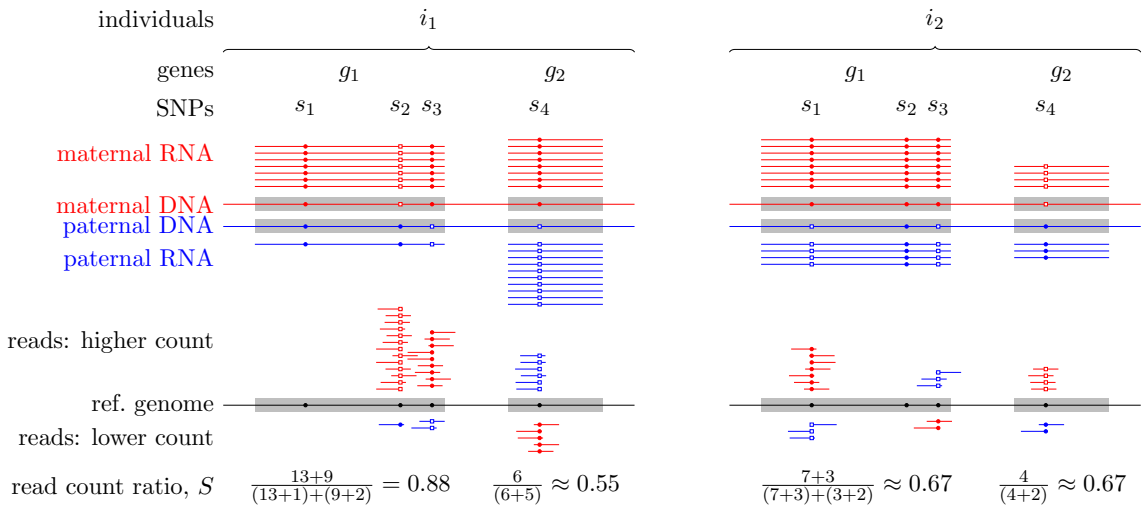

Figure 1: Quantifying allelic bias of expression in human individuals using the RNA-seq read count ratio statistic S_{ig}. The strength of bias towards the expression of the maternal (red) or paternal (blue) allele of a given gene g in individual i is gauged with the count of RNA-seq reads carrying the reference allele (small closed circles) and the count of reads carrying an alternative allele (open squares) at all SNPs for which the individual is heterozygous. The allele with the higher read count tends to match the allele with the higher expression but measurement errors may occasionally revert this tendency as seen for SNP s_3 in gene g_1 in individual i_2.

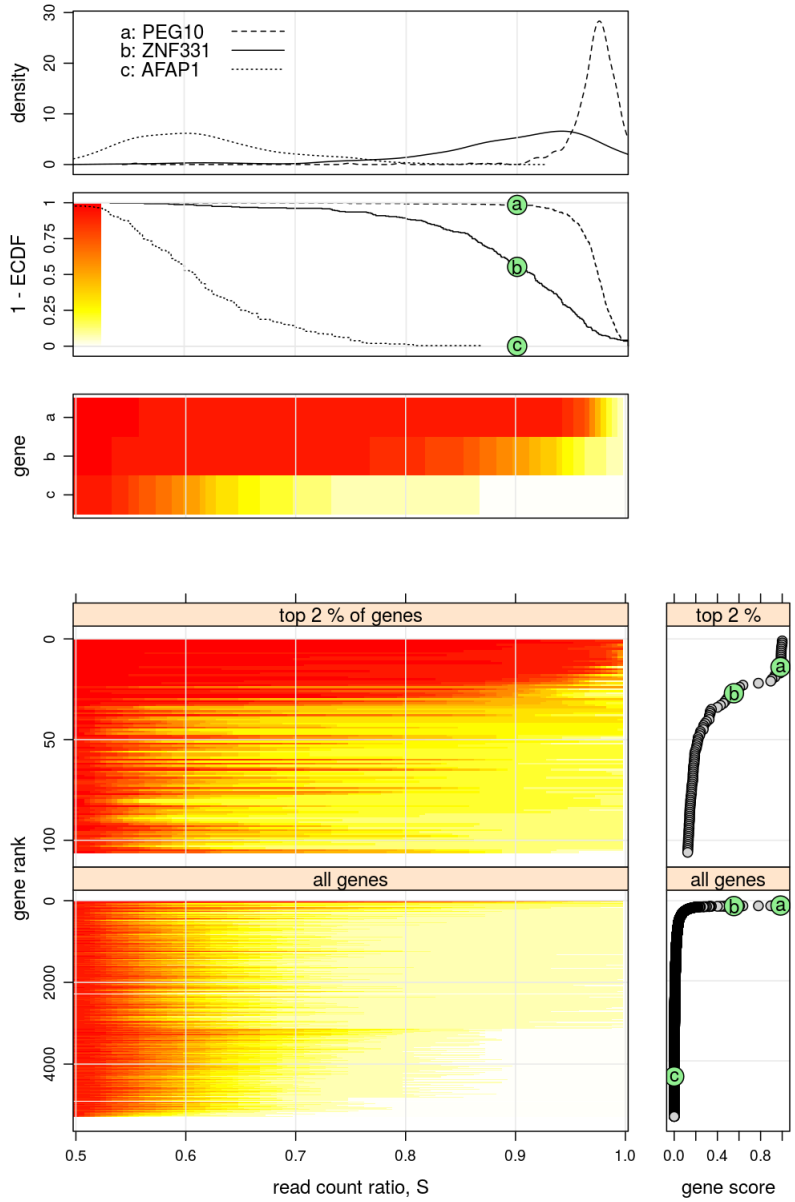

Figure 2: Across-individuals distribution of read count ratio S for each gene indicates substantial variation of allelic bias and that $< 1\%$ of all genes are imprinted. The vertically arranged five main panels present the empirical distribution of $S_{\cdot g}$ across all individuals given each gene g. The *upper three panels* are distinct representations (density plot, survival plot: $1-$ECDF, and "survival heatmap") of the same three distributions: two for PEG10 and ZNF331, previously found to be imprinted in mice or in other human tissues, and one for AFAP1, a gene without prior evidence. Note the color scale in the survival plot (*second panel from top*) for the heatmaps (*third, fourth and fifth main panels from top*). The bottom two survival heatmaps present the distribution of $S_{\cdot g}$ for the top 2% and 100% of the 5307 analyzed genes. These are ranked according to gene score defined as $1 - \text{ECDF}(0.9)$ in the *bottom far right panels*. The score of PEG10, ZNF331, and AFAP1 is marked by a, b, c, respectively, in green circles. As expected, PEG10 and ZNF331 both score high and rank within the top 30 of all genes suggesting they are also imprinted in the present context, the adult human DLFPC. The bottom panels also indicate that $< 1\%$ of all genes might be imprinted.

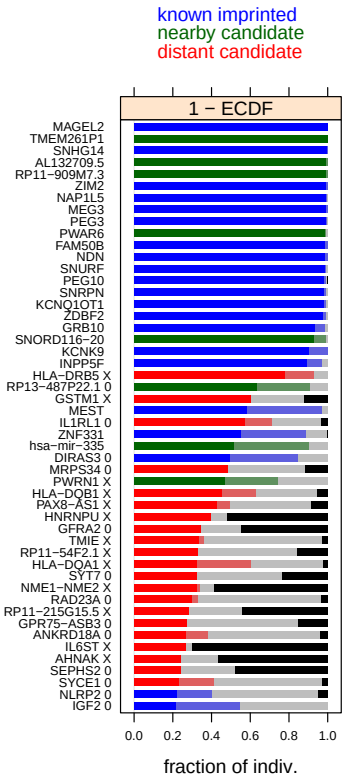

Figure 3: The top 50 genes ranked by the gene score. The score of gene g is $1 - \mathrm{ECDF}_g(0.9)$, the fraction of individuals i for which $S_{ig} > 0.9$ and is indicated by the length of dark blue, dark green or dark red bars. Note that the same ranking and score is shown in the bottom half of Fig. 2. The right border of the light blue, light green and light red bars is at $1 - \mathrm{ECDF}_g(0.8)$. The black bars indicate the fraction of individuals passing the test of nearly unbiased expression (Eq. 2). "X" characters next to gene names indicate mapping bias based on counts of reference and non-reference alleles, while "0" indicate that total allele count was insufficient for this test.

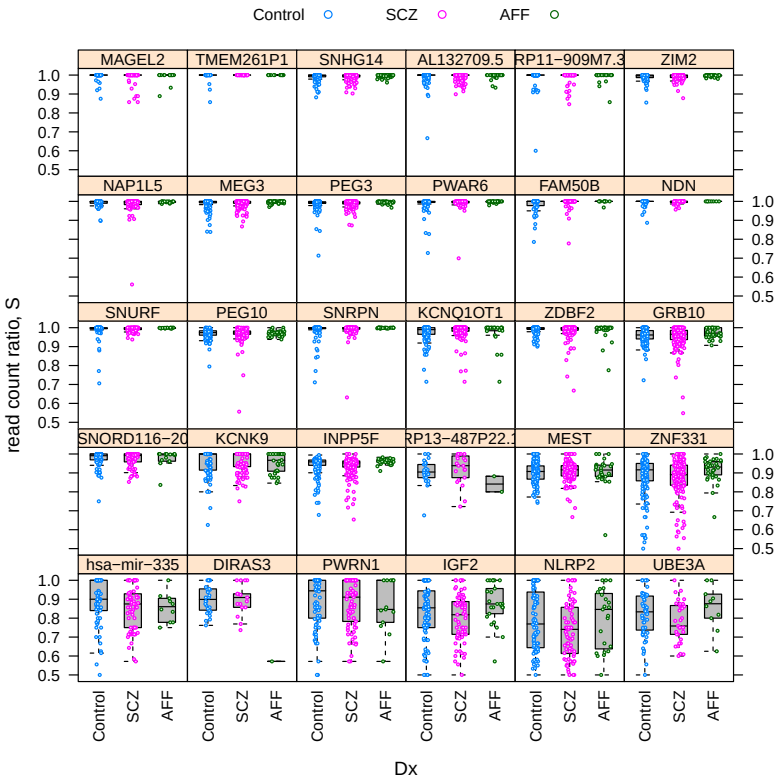

Figure 4: Schizophrenia does not affect allelic bias of imprinted genes. Distribution of read count ratio for Control, schizophrenic (SCZ), and affectic spectrum (AFF) individuals within each gene that has been considered as imprinted in the DLPFC brain area in this study.

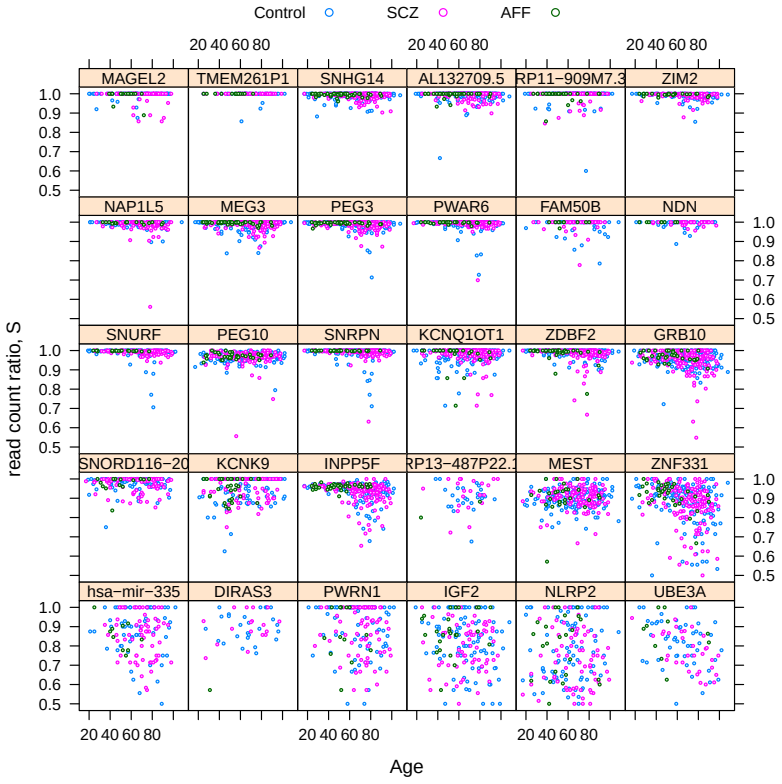

Figure 5: Allelic bias depends differentially on age across imprinted genes. The panels and colors are consistent with the imprinted genes and psychiatric diagnoses presented in Fig. 4. The differential dependence on age is apparent when comparing PEG3 or ZNF331 (negative dependence) to KCNK9 or RP13-487P22.1 (positive dependence) or to NDN or NLRP2 (no dependence).

7 Tables with legends

predictor term	interpretation	ΔAIC	p-value
(1 | Gene)	variability among genes	−126.8	8.5×10^{-28}
(1 | Dx)	variability among Control, SCZ, AFF	2.0	1.0
(1 | Dx : Gene)	Gene specific variability among Ctrl, SCZ, AFF	0.4	0.21
Age	effect of Age	1.3	0.39
(Age | Gene)	Gene specific effect of Age	−18.9	2.5×10^{-5}
Ancestry.1	effect of Ancestry.1	0.6	0.24
(Ancestry.1 | Gene)	Gene specific effect of Ancestry.1	−71.2	4.6×10^{-16}
Ancestry.3	effect of Ancestry.3	1.6	0.54
(Ancestry.3 | Gene)	Gene specific effect of Ancestry.3	−17.9	3.8×10^{-5}
(1 | Gender)	difference between Male and Female	2.0	1.0
(1 | Gender : Gene)	Gene specific difference between M and F	−5.7	5.5×10^{-3}

Table 1: Dependence of read count ratio on various predictor terms in the mixed model; largely negative ΔAIC and small p-values indicate significant dependence (see Methods).

8 Supplementary information

8.1 Supplementary figures with legends

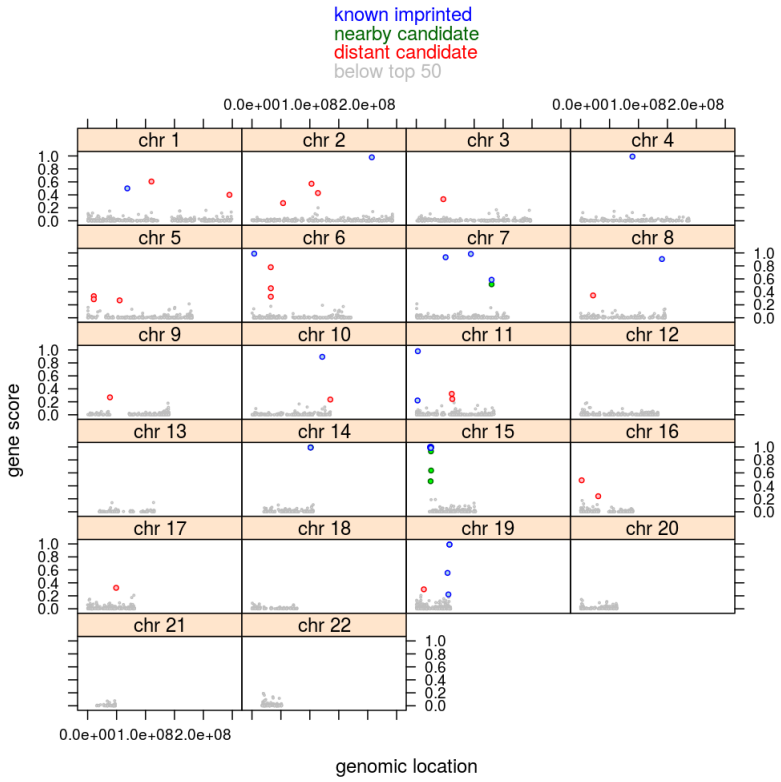

Figure S1: Clustering of top-scoring genes in the context of human DLPFC around genomic locations that had been previously described as imprinted gene clusters in other contexts.

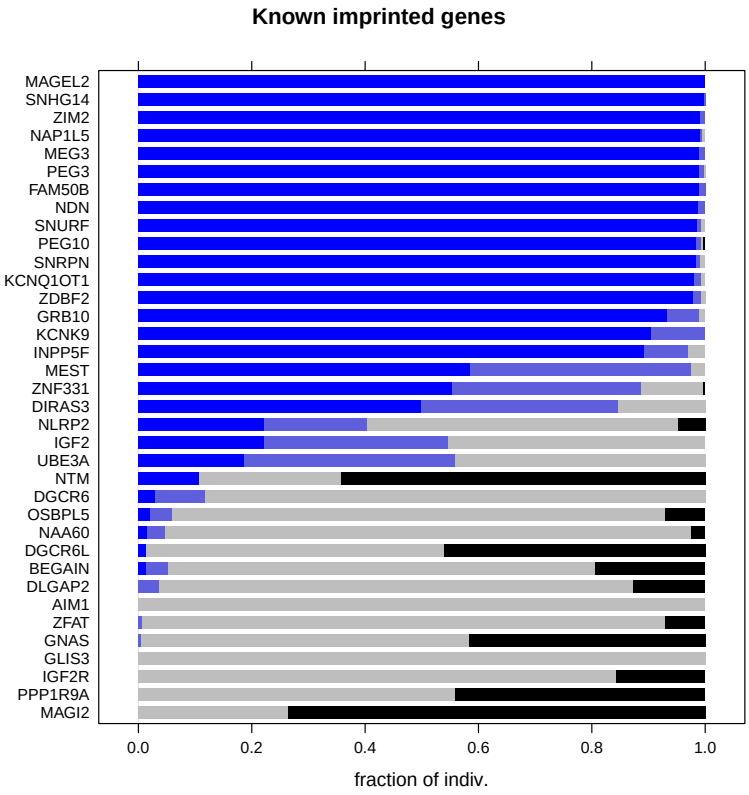

Figure S2: Known imprinted genes ranked by the gene score (dark blue bars). "Known imprinted" refers to prior studies on imprinting in the context of any tissue and developmental stage. The length of the black bars indicates the fraction of individuals passing the test of nearly unbiased expression.

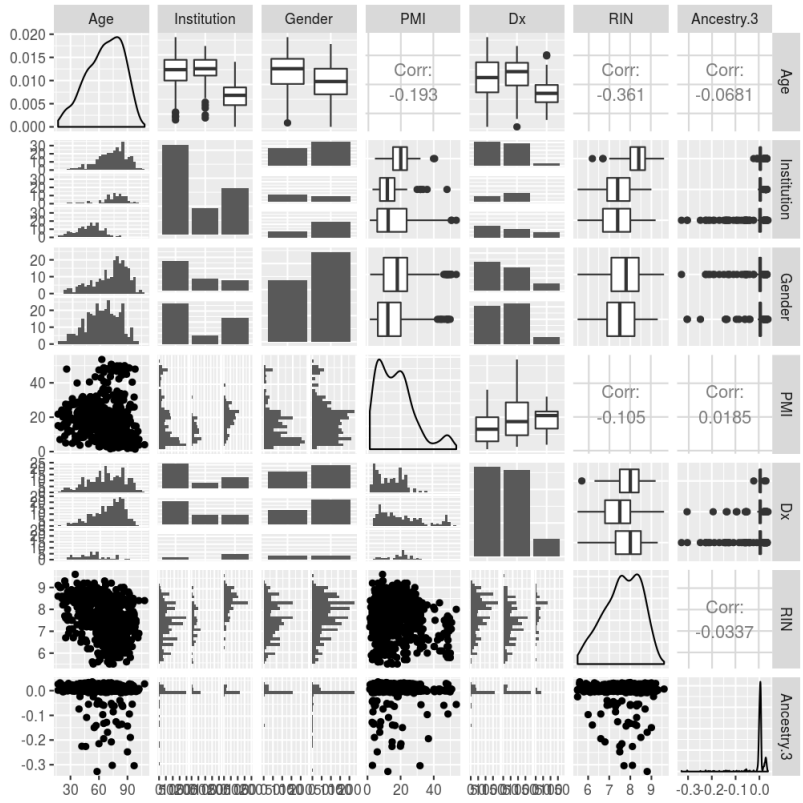

Figure S3: Distribution and inter-dependence of explanatory variables. The diagonal graphs of the plot-matrix show the marginal distribution of six variables (Age, Institution,...) while the off-diagonal graphs show pairwise joint distributions. For instance, the upper left graph shows that, in the whole cohort, individuals' Age ranges between ca. 15 and 105 years and most individuals around 75 years; the bottom and right neighbor of this graph both show (albeit in different representation) the joint distribution of Age and Institution, from which can be seen that individuals from Pittsburg tended to be younger than those from the two other institutions.

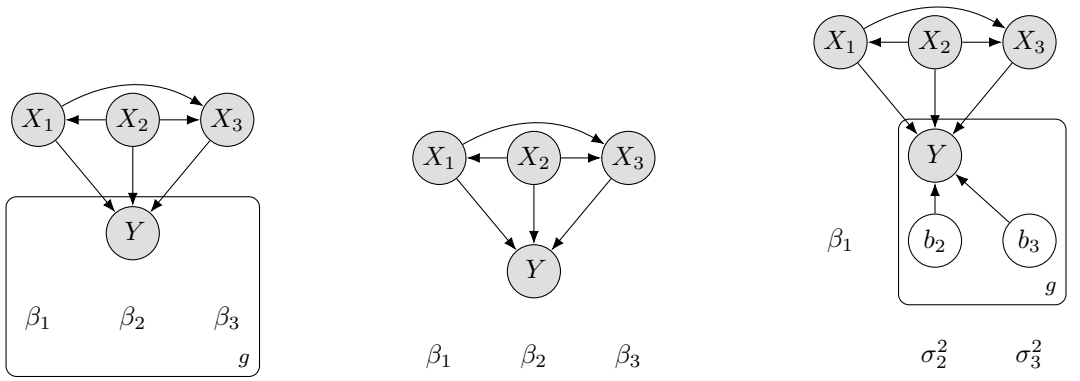

Figure S4: General dependency structures in two fixed effects regression models (*left*, *middle*) and a mixed effects model (*right*). In all three cases the regression coefficients $\beta_{1g}, ..., \beta_{3g}$ or $\beta_{1g}, b_{2g}, b_{3g}$ mediate, for a given gene g, probabilistic dependencies (arrows) between the response variable Y_g (read count ratio for g) and the corresponding explanatory variables $X_1, ..., X_3$. For simplicity but without loss of generality only 3 explanatory variables are depicted. The model frameworks differ in how the coefficients relate to each other for a given explanatory variable (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 variable 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 explanatory variables. Thus genes share all information in the data in the sense that the model forces them to be identical. *Right:* Hierarchical mixed effects model where certain dependencies (β_1) are shared among genes while others ($\{b_{2g}\}_g, \{b_{3g}\}_g$) vary across genes. The variation is controlled by the variance parameters σ_j^2. In this example there is a single set $\{b_{2g}\}_g$ of random coefficients for X_2—and a similar set for X_3—, which are random intercepts. In general, however, one or more set of random slope coefficients may also be present. Given the estimates $\hat{\sigma}_j^2$ and the data the gene-specific random coefficients b_{jg} can be predicted. Among the three only this model framework allows information sharing among genes in a flexible way.

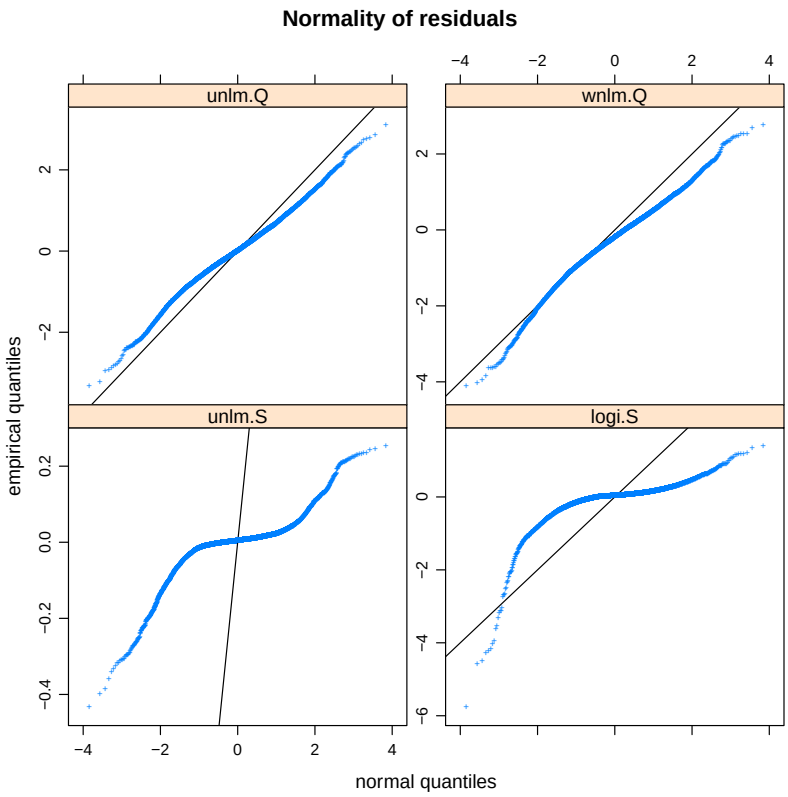

Figure S5: Checking the fit of various model families: analysis of the normality of residuals.

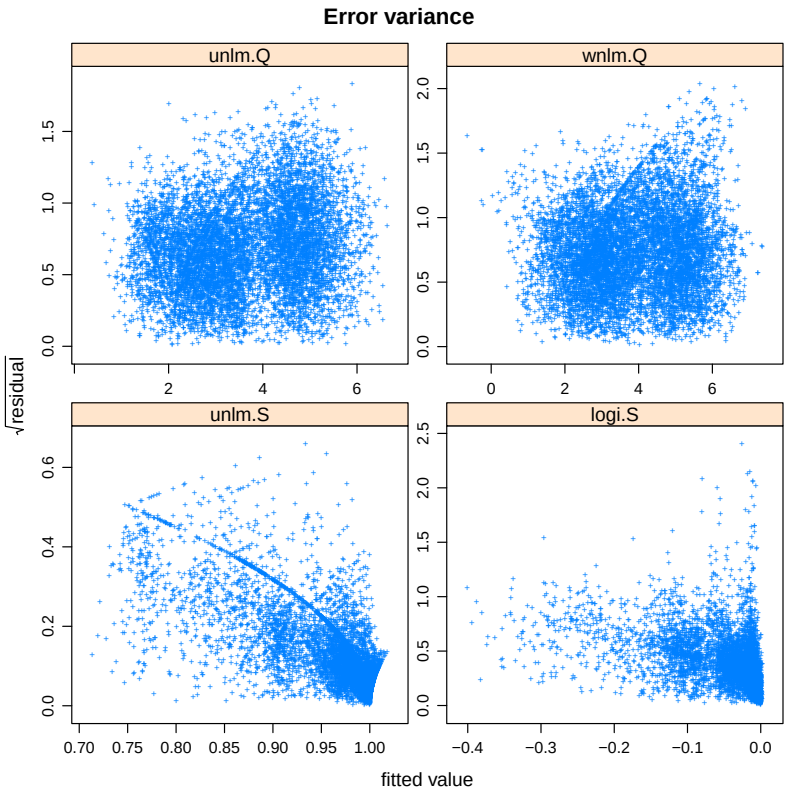

Figure S6: Checking the fit of various model families: analysis of homoscedasticity.

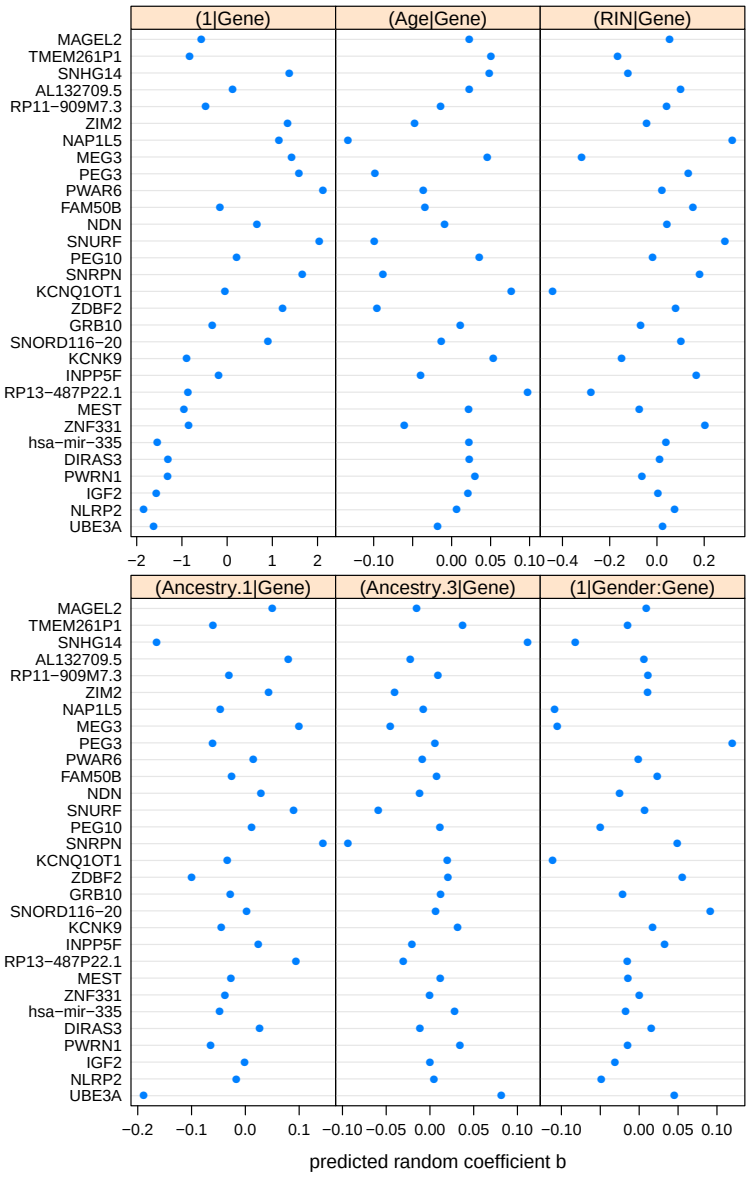

Figure S7: Predicted random coefficients b_{gj} for gene g (y-axis) and random effect j (panel headers). Positive and negative coefficient indicates direct positive and negative dependence of the given gene's read count ratio on age, respectively, while zero coefficient suggests independence of age. Compare with Fig. 5.

8.2 Supplementary tables with legends

explanatory variable	levels
Age	
Institution	[MSSM], Penn, Pitt
Gender	[Female], Male
PMI	
Dx	[Control], SCZ, AFF
RIN	
RNA_batch	[A], B, C, D, E, F, G, H, 0
Ancestry.1	
⋮	
Ancestry.5	

Table S1: *Left column:* explanatory variables of read count ratio. *Right column:* levels of each factor-valued (i.e. categorical) variable. Square brackets [...] surround the baseline level against which other levels are contrasted. *Abbreviations:* PMI: post-mortem interval; Dx: disease status; AFF: affective spectrum disorder; SCZ: schizophrenia; RIN: RNA integrity number; Ancestry.k: the k-th eigenvalue from the decomposition of genotypes indicating population structure.

model family	abbrev.	response var.
*u*nweighted *n*ormal *l*inear	unlm	S, Q, or R
*w*eighted *n*ormal *l*inear	wnlm	S, Q, or R
*logi*stic	logi	S
*logi*stic, $\frac{1}{2}\times$ down-scaled link fun.	logi2	S

Table S2: Fitted regression model families, in which the response variable is the read count ratio with or without some transformation: S—untransformed, Q—*q*uasi-log-transformed, and R—*r*ank-transformed read count ratio. Diagnostic plots (Fig. S5, S6) and monitoring convergence suggested that the unlm.Q combination allows the best fit for several linear predictors tested.