Variation of Genomic Imprinting in the Human Brain

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

Mount Sinai School of Medicine

Abstract

We present a genome-wide analysis of human genomic imprinting based on RNA-seq measurements of parental bias in allele-specific expression in the dorsolateral prefrontal cortex. We find that the fraction of imprinted human genes is consistent with lower ($\approx 0.5\%$) as opposed to higher ($\approx 5\%$) estimates in mice. Our analysis reveals that age up or down-regulates allelic bias of some, but not all, imprinted genes in adulthood and that allelic bias depends also on genetic variation. These results support the hypothesized role of imprinted genes in social interactions, which dynamically change in the life course and depend on neuropsychological function.

1 Introduction

Genomic imprinting, leading to repression of either the maternal or paternal allele, has reached its highest prevalence in humans and other placental organisms [11]. In line with this, well-known physiological functions of imprinted genes include embryonic and placental development, body growth, suckling, and maternal behavior [10, 9]. Genomic imprinting requires the placement of different epigenetic marks, such as DNA methylation, at the respective alleles residing on the chromosomes originating from the mother or father [10]. Indeed, imprinted genes typically reside in clusters spanning hundreds of kilobases and allele-specific differential epigenetic marks are found near specific genes as well as in shared regulatory elements called imprinting control regions. For non-imprinted genes expression is balanced such that the two alleles are roughly equally expressed. By contrast, for imprinted genes epigenetic marks lead to substantial degrees of imbalance in the expression level of the two alleles. We refer to the degree of expression imbalance as *allelic bias*; at its maximum expression is completely monoallelic.

Why natural selection favors allelic bias for imprinted genes remains debated [14, 7, 6] but the most mature of all theories, kinship or conflict theory [14], provides a flexible framework for interpreting past studies on imprinted genes and formulating hypotheses and predictions regarding their detailed regulation and physiological function. The theory assumes that all imprinted genes contribute to inter-individual interactions in a highly dose-dependent fashion, and explains allelic bias with the conflicting interests of paternal and maternal genes, which arise from sexual asymmetries in those interactions [14]. A well-known asymmetry is the disproportionate role of mothers in nurturing offspring in Placentalia. Kinship theory thus explains why overexpression disorders of paternally or maternally biased genes in children abnormally promote or inhibit, respectively, their growth [10, 9].

Since different inter-individual interactions take place in various developmental stages and are mediated by various organs, kinship theory explains the non-uniform pattern of "imprintedness" and

allelic bias over various ages [2] and tissue types, which is seen for several imprinted genes [10, 9]. For other genes such patterns await discovery. The theory quantitatively predicts relaxation of allelic conflict with age [13] and so raises the hypothesis that change in allelic bias is linked to aging. A study on newborn and young adult mice partially supports that hypothesis [8] but experimental evidence from humans, including older individuals, is missing.

In the framework of kinship theory the question of aging is closely tied to the roles of imprinted genes in social interactions and in the underlying psychiatric functions [13, 14], whose importance has been increasing in human evolution. Indeed, most human imprinted gene syndromes are characterized by not only growth disorder but also mental retardation and psychiatric dysfunction [10, 9]. More precisely, paternally and maternally biased genes are suggested to play antagonistic roles not only in growth but also in psychiatric functions [3], since overexpression of the former is associated with autistic while that of the latter with psychotic spectrum disorders. For example, maternally derived microduplications at 15q11-q13 may not only cause the Prader-Willi syndrome [9]—whose symptoms include obesity—but also highly penetrant for schizophrenia [5, 12], which is perhaps the most devastating psychotic spectrum disorder.

Recently the CommonMind Consortium produced and shared[1] genome-wide data sets on genotype and gene expression in the dorsolateral prefrontal cortex (DLPFC) of hundreds of schizophrenic and control individuals, and identified some 650 differentially expressed genes [4]. Our present work extends that analysis with a focus on allele-specfic expression across the genome, allowing us to determine allelic bias for each gene. Based on that, we find that $\approx 0.6\%$ of all genes are imprinted in the human DLPFC, the majority of which had been reported to be imprinted in the context of one or another tissue and/or species. We find a number of genes with allele specific expression residing near clusters of known imprinted genes. Furthermore, our data suggest that for several imprinted genes the variation of allelic bias across individuals is explained by differences in ancestry and age in a manner that depends on the gene.

2 Methods

2.1 Defining the read count ratio to quantify allelic bias

We quantifieied 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.

[1] www.commonmind.org

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 Data

2.2.1 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.

2.2.2 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.

2.2.3 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. 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.

2.2.4 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 $n = 5307$.

2.3 Test for nearly unbiased expression

This test 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 UCL_{ig} for the expected read count ratio p_{ig} was calculated based on the assumption that the higher read count $H_{ig} = S_{ig}T_{ig} \sim \text{Binom}(p_{ig}, T_{ig})$, on the fact that binomial random variables are asymptotically (as $T_{ig} \to \infty$) normal with $\text{var}(H_{ig}) = T_{ig}p_{ig}(1-p_{ig})$, and on the equalities $\text{var}(S_{ig}) = \text{var}(H_{ig}/T_{ig}) = \text{var}(H_{ig})/T_{ig}^2$. Therefore

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

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

2.4 Regression analysis

2.4.1 Formulation of models

Given a top-scoring gene g, we sought to identify biological regulators and psychiatric consequences of allelic bias by studying the conditional distribution of $S_{\cdot g}$ given observations $x_{\cdot 1}, ..., x_{\cdot p}$ on features of study individuals that are not gene-specific. We call these features *predictors* because we used

them in a regression model framework. Predictors and their various levels (if any) are listed in Table S1.

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.

Outline:

- we modeled the dependence of read count ratio of imprinted genes on on biological and technical explanatory variables (Table S1) using two regression frameworks
    1. hierarchical mixed effects models, which describe data jointly for all 30 imprinted genes TODO: equation
    2. non-hierarchical generalized linear models, which describe data separately for each of the 30 imprinted genes TODO: equation
- within each framework several data transformations, link functions, error distributions (Table 5)
- initial inference using non-hierarchical models guided heuristic search for the best fitting mixed model considering the space of linear predictors based on subsets of the explanatory variables
- we prefered mixed models for final inference due to their higher power and robustness against overfitting which they achieve by letting gene-wise data borrow statistical strength from each other

2.4.2 Model fitting and selection

- fitting mixed and non-hierarchical regression models with lme4 and stats R packages, respectively, convergence
- fit during search was evaluated using information criteria AUC, as well as p-value from Chi-squared test on the degrees of freedom that depends on the term(s) added
    1. find the best overall combination of data transformation and model family (Figure S6, S7)
    2. start the search in the space of linear predictors as near the optimum as possible
    3. guide the sequence in which further terms are added
- many of the terms in the linear predictor that significantly improved model fit are technical ones; including them in the model therefore corrects for a large technical noise

3 Results

3.1 Genome- and population-wide variation of allelic bias

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. Examining these genes, we performed exploratory statistical analysis based on the read count ratio statistic S_{ig}, whose results (below) we interpreted in terms of the variation of allelic bias both across genes g and across individuals i. Note that our

later analyses (Section 3.3 and below) used information not only in S_{ig} but also in the total read count as well as in data beyond RNA-seq.

Fig. 2 presents the conditional empirical distribution of $S_{\cdot g}$ given each gene g. Each of the three plots of the upper half show in a distinct representation the same empirical distributions based on data for three genes. The main panels of the lower half present, in the most compact representation, the distributions based on all data (5307 genes). Two of the three genes in the upper half, PEG10 and ZNF331, are *known imprinted* genes in the sense that they had previously been found imprinted in the context of some developmental stage, species, and tissue type other than the adult human DLPFC. The third, AFAP1, has not been reported to be imprinted in any context. For all three genes $S_{\cdot g}$ varies considerably within its theoretical range $[\frac{1}{2}, 1]$. This suggests variation of allelic bias across study individuals, although some component of the variation of $S_{\cdot g}$ must originate from technical sources.

To identify imprinted genes based on the read count ratio we defined the score of each gene g as the location statistic $1 - \mathrm{ECDF}_g(0.9)$, which is fraction of individuals i for which $S_{ig} > 0.9$. We ranked all 5307 genes according to their score shown in the side plots of the lower half of Fig. 2 as gray filled circles. Larger green circles mark the three genes mentioned above. The heat map of empirical distribution of $S_{\cdot g}$ of ranked genes (Fig. 2, lower left) suggests that the top 50 genes, which constitute $\approx 1\%$ of all genes in our analysis, are qualitatively different from the bottom $\approx 99\%$ suggesting that most of them are imprinted. Consistent with this, the top-scoring genes tended to cluster around genomic locations that had been previously described as imprinted gene clusters (Fig. S1).

3.2 Imprinted genes in the adult DLPF cortex

The set of top scoring 50 genes is highly enriched in known imprinted genes, marked by blue in Fig. 3 and in *nearby candidate* genes (green) defined as being within 1Mb of a known imprinted gene. Within the top 50, we find 29 such genes; 21 known imprinted genes and eight nearby candidates. In subsequent analysis (below) we also consider UBE3A as demonstrating allelic bias consistent with imprinting in the context of human adult DLFPC as evidenced by Fig. S2 even though its rank falls outside of the top 50.

The remaining 21 genes in the top 50 are separated by > 1 Mb from some known imprinted gene (termed *distant candidates*, red in Fig. 3). Upon further examination these distant candidate genes are overwhelmingly likely not imprinted. The primary reason for this conclusion is that we performed a test to see if there is reference allele bias for all candidate genes. 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. Most known imprinted genes and the nearby candidates display a reference/non-reference distribution consistent with a binomial distribution with a probability of 0.5 for both the reference and non-reference alleles. However, and in sharp distinction, most distant candidates have distributions of reference/non-reference that are not consistent with equal probabilities (see genes marked with "X" in Fig. 3). Indeed, for most of them the distribution is shifted towards the reference allele strongly suggesting that mistaken genotyping, imputation or a mapping issue led to the presence of these red genes in the list of the top 50 genes. One could argue that we should have left these genes out of Fig. 3, but we thought it was important to show them and to indicate the reasons they are set aside. Note also, that we also tested the hypothesis

for each gene g and individual i that allelic expression is (nearly) unbiased (Eq. 2). The fraction of individuals for which the test was *not* rejected tends to be much higher for the "red" genes in the top 50 (black bars in Fig. 3).

While the shifted distribution of reference/non-reference alleles leads us to discount the possibility of imprinting, random monoallelic expression is still a distinct possibility for these candidates as our studies of random monoallelic expression in mice suggested that a substantial fraction ($40-80\%$) of random monoallelically-expressed genes had a very strong bias towards monoallelic expression of one of the two alleles [16]. Moreover, it is worth noting that three of these candidates are from the major histocompatibility locus (HLA), which is notable for extensive polymorphism and difficulties with allelic identification. For these three genes we also analyzed them more thoroughly with HLA-specific methods for determining haplotype based on RNA-seq [1] and genotype data [15]. The high observed read count ratios for HLA genes appear to be driven by eQTL-like effects, not by random monoallelic expression nor by imprinting (manuscript in preparation). Examining all the assessable known imprinted genes, we find than $\frac{1}{3}$rd of them have a low gene score. This suggests that these genes do not display imprinted expression in the human adult DLPFC, consistent with many reports in the literature indicating that known imprinted genes are often imprinted in some but not all tissues.

3.3 Allelic bias of imprinted genes and schizophrenia

- our overall hypothesis: aberrations in imprinted genes as a risk of SCZ
- the more specific hypothesis: normal allelic bias of imprinted genes is perturbed in SCZ
- Fig. 4 compares read count ratio distribution among control, SCZ and AFF for the 30 imprinted genes and suggests there is no difference
- we performed a more rigorous test: mixed (hierarchical) regression model-based statistical inference, in which read count ratio was taken as the response variable whereas Dx, Gene, Age, RIN,... as explanatory variables
- first we selected the best fitting model among many related ones (Methods)
- then we quantified how significantly the presence of various terms of some explanatory variable improves the fit
- the power of our model is indicated by the extreme significance by which the term $(1 \mid \text{Gene})$ improves fit; this term represents the gene-to-gene variation of read count ratio among imprinted genes that is clear from Fig. 4
- we examined two terms involving Dx, which correspond to two sub-hypotheses:
  1. $(1 \mid \text{Dx})$ represents effect of Dx on read count ratio that is independent of Gene and any other explanatory variable in the model and thus represents an effect shared by all 30 genes
  2. $(1 \mid \text{Dx} : \text{Gene})$ represents gene-specific effect of Dx
- Table 1 shows that neither term improved fit significantly suggesting that perturbation of allelic bias of imprinted genes in the DLPFC is not associated to SCZ

3.4 Gene-specific effects of age and ancestry

- Fig. 5 shows that the read count ratio of some imprinted genes do not vary with age whereas for other imprinted genes age has a negative or positive effect on the read count ratio
- however, this apparent age effect may be mediated solely by other explanatory variables to which age is associated with (Fig. S3)
- to investigate the direct effect of age, ancestry, and gender we turned to our mixed models once again, evaluating how significantly predictor terms associated with these variables improve fit (Table 1)
- as for Dx we defined two terms for each variable, one that represents its gene-specific effect, like (Age | Gene), and another one that represents an effect that is shared by all imprinted genes, like Age
- the shared effects were not supported by any criteria of model fit (see Age, Ancestry.1, and (1 | Gene) in Table 1)
- but the gene-specific effects did receive support: the most significant was the effect of Ancestry.1, that of Age followed, while that of Gender was only weakly significant
- TODO: predict the effect of Age that is specific to each gene; note that similar prediction for Ancestry.1 is possible but seems much less interesting because we do not know the genetic mechanisms linked to Ancestry.1

4 Discussion

- we identified imprinted genes in adult DLPF cortex using an RNA-seq read count ratio approach that estimates allelic bias for a given gene and individual
- the resulting set of imprinted genes is in agreement with this and that study but contrasts earlier work (Dulac, Gregg)
- our model-based inference does not support the hypothesis that perturbed allelic bias of imprinted genes increases the risk of SCZ
- implications, alternative hypothetical biological mechanisms involving imprinting and SCZ
- alternatively the effect exists but has been masked by the large technical noise in our study
- we also examined the dependence of allelic bias on variables namely age, gender and ancestry
- we find (so far the strongest?) evidence that age regulates imprinted genes in adulthood
- biological significance in light of kinship theory, social interactions and aging
- we find a strong Ancestry.1:Gene effect; this shows how genetic (Ancestry.1) and epigenetic (imprinting control region of a given gene) mechanisms act together to shape gene expression
- even subtle variation in allelic bias and allele specific expression might substantially affect biological function for some (imprinted) genes while not for others; this must be borne in mind in interpreting our results

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] Andrew F.G. Bourke. Kin Selection and the Evolutionary Theory of Aging. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):103–128, dec 2007.

[3] Bernard Crespi and Christopher Badcock. Psychosis and autism as diametrical disorders of the social brain. *The Behavioral and brain sciences*, 31(3):241–61; discussion 261–320, jun 2008.

[4] 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.

[5] Andrés Ingason, George Kirov, Ina Giegling, Thomas Hansen, Anthony R Isles, Klaus D Jakobsen, Kari T Kristinsson, Louise le Roux, Omar Gustafsson, Nick Craddock, Hans-Jürgen Möller, Andrew McQuillin, Pierandrea Muglia, Sven Cichon, Marcella Rietschel, Roel A Ophoff, Srdjan Djurovic, Ole A Andreassen, Olli P H Pietiläinen, Leena Peltonen, Emma Dempster, David A Collier, David St Clair, Henrik B Rasmussen, Birte Y Glenthøj, Lambertus A Kiemeney, Barbara Franke, Sarah Tosato, Chiara Bonetto, Evald Saemundsen, Stefán J Hreidarsson, GROUP Investigators, Markus M Nöthen, Hugh Gurling, Michael C O'Donovan, Michael J Owen, Engilbert Sigurdsson, Hannes Petursson, Hreinn Stefansson, Dan Rujescu, Kari Stefansson, and Thomas Werge. Maternally derived microduplications at 15q11-q13: implication of imprinted genes in psychotic illness. *The American journal of psychiatry*, 168(4):408–17, apr 2011.

[6] Eric B Keverne. Genomic imprinting, action, and interaction of maternal and fetal genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6834–40, jun 2015.

[7] J F McDonald, M A Matzke, and A J Matzke. Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenetic and genome research*, 110(1-4):242–9, 2005.

[8] 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.

[9] Jo Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nature reviews. Genetics*, 15(8):517–30, aug 2014.

[10] Robert N Plasschaert and Marisa S Bartolomei. Genomic imprinting in development, growth, behavior and stem cells. *Development (Cambridge, England)*, 141(9):1805–13, may 2014.

[11] M. B. Renfree, S. Suzuki, and T. Kaneko-Ishino. The origin and evolution of genomic imprinting and viviparity in mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1609):20120151–20120151, nov 2012.

[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] Jon F Wilkins and David Haig. What good is genomic imprinting: the function of parent-specific gene expression. *Nature reviews. Genetics*, 4(5):359–68, may 2003.

[15] 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.

[16] 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.

predictor term	interpretation: effect due to	ΔAIC	p-value
(1 | Gene)	Gene, independently of other variables	−126.8	8.5×10^{-28}
(1 | Dx)	Dx, independently of other variables	2.0	1.0
(1 | Dx : Gene)	Dx, depending on Gene	0.4	0.21
Age	Age, independently of other variables	1.3	0.39
(Age | Gene)	Age, depending on Gene	−18.9	2.5×10^{-5}
Ancestry.1	Ancestry.1, independently of other variables	0.6	0.24
(Ancestry.1 | Gene)	Ancestry.1, depending on Gene	−71.2	4.6×10^{-16}
(1 | Gender)	Gender, independently of other variables	2.0	1.0
(1 | Gender : Gene)	Gender, depending on Gene	−5.7	5.5×10^{-3}

Table 1: Biological terms and their impact on the fit of the mixed model. The smaller (more negative) ΔAIC, and the smaller the p-value, the more likely that the term has non-zero effect on the read count ratio.

5 Figures, Tables and Supplementary Material

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

Table S1: *Left column:* predictors (explanatory variables) of read count ratio. *Right column:* levels of each factor-valued (i.e. categorical) predictor. 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.

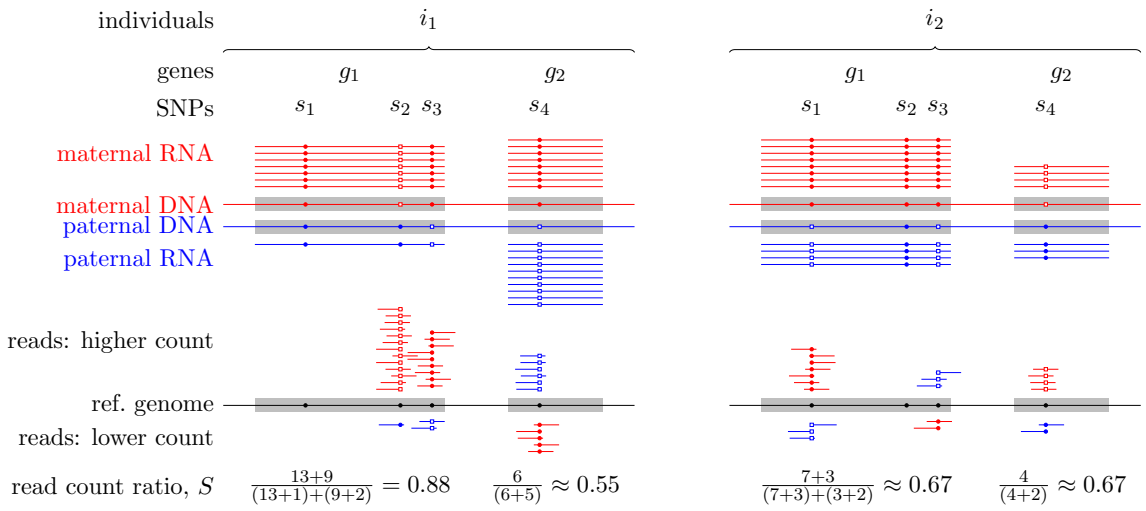

Figure 1: Quantifying allelic bias of expression in human individuals using the RNA-seq read count ratio statistic S_{ig}. The strength of bias towards the expression of the maternal (red) or paternal (blue) allele of a given gene g in individual i is gauged with the count of RNA-seq reads carrying the 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.

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.

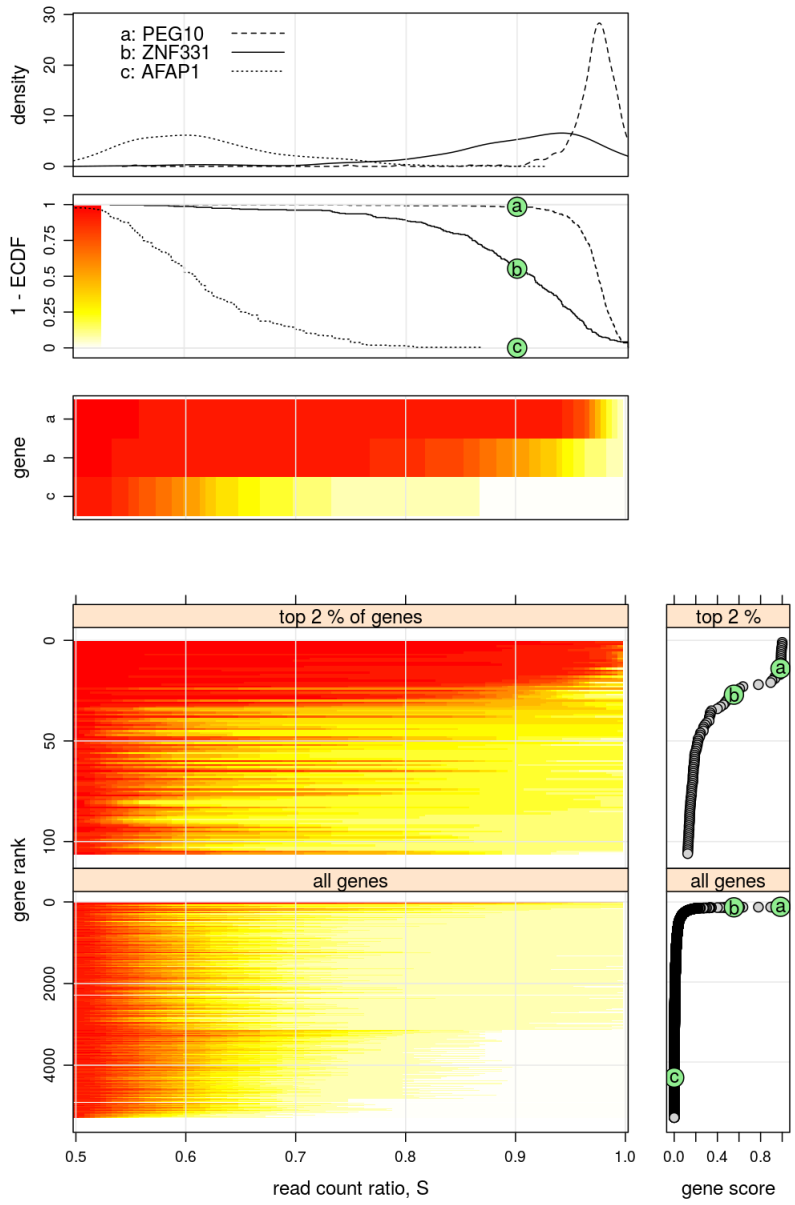

Figure 2: Using the read count ratio statistic S to report on variation of allelic bias across individuals and genes. *Upper half*, from top to bottom: (1) kernel density estimate; (2) the graph of the survival function 1 - ECDF, where ECDF means empirical cumulative distribution function; note color scale for heat map and green filled circles marking genes' score; (3) the heat map representation of the survival function. *Lower half*, main panels: heat map of the survival function for all 5307 analyzed genes ranked according to their score; right side panels: gene score.

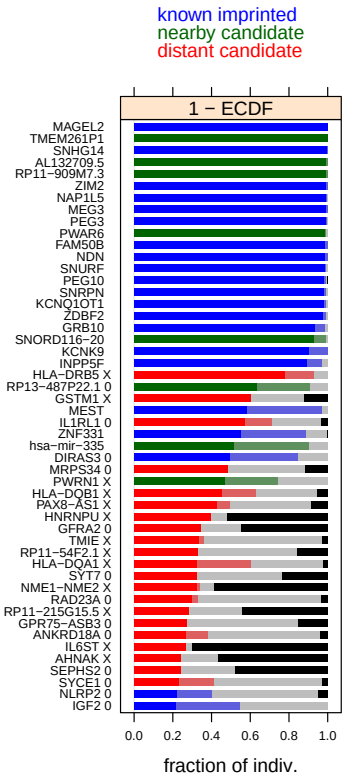

Figure 3: The top 50 genes ranked by the gene score. The score of gene g is $1 - \mathrm{ECDF}_g(0.9)$, the fraction of individuals i for which $S_{ig} > 0.9$ and is indicated by the length of dark blue, dark green or dark red bars. Note that the same ranking and score is shown in the bottom half of Fig. 2. The right border of the light blue, light green and light red bars is at $1 - \mathrm{ECDF}_g(0.8)$. The length of the black bars indicates the fraction of individuals passing the test of nearly unbiased expression (Eq. 2). "X" characters next to gene names indicate reference allele bias, while "0" indicate that reference allele bias could not be determined due to small amount of data.

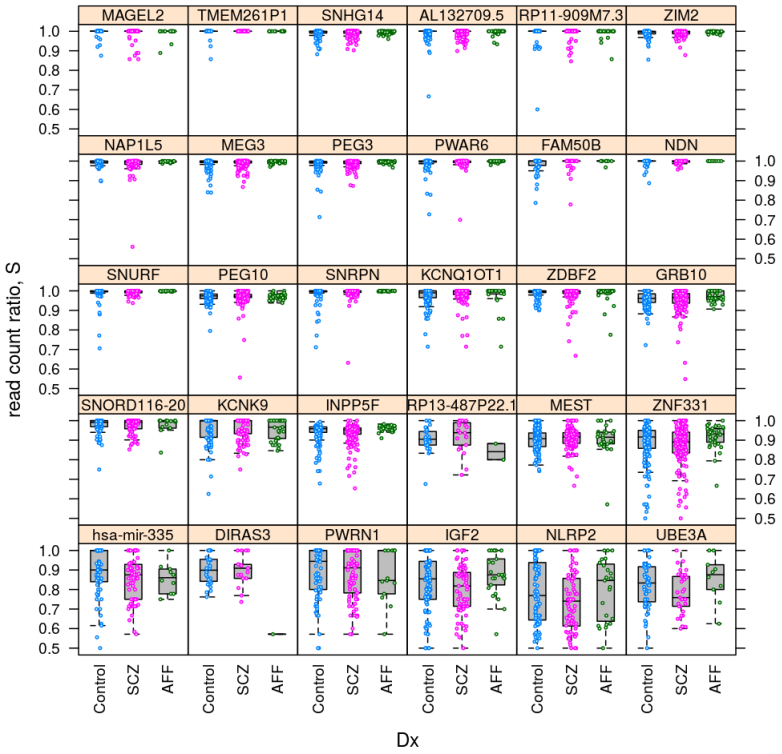

Figure 4: Distribution of read count ratio for control, schizophrenic (SCZ), and affectic spectrum (AFF) individuals within each gene that has been considered as imprinted in the DLPFC brain area in this study.

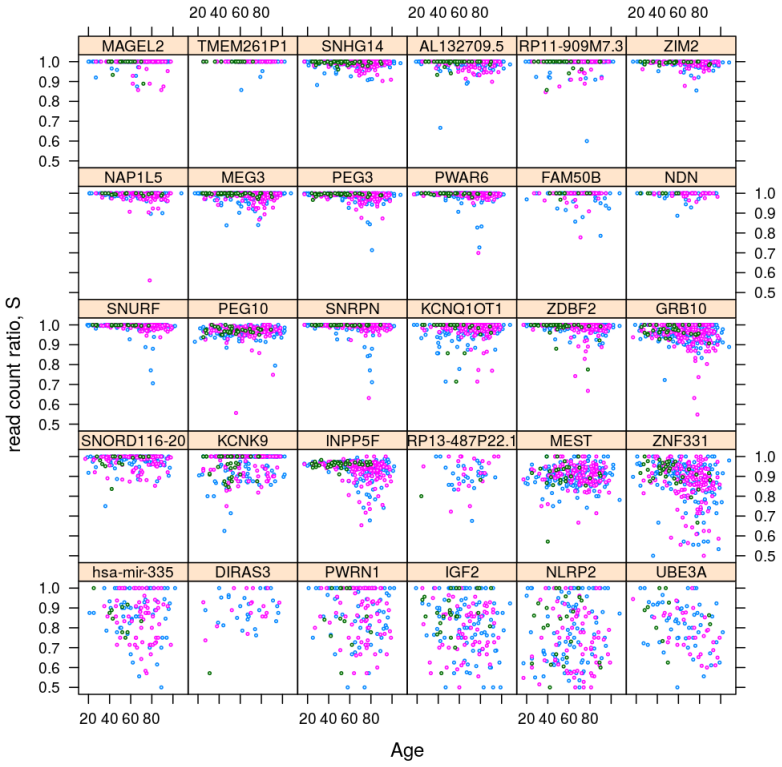

Figure 5: Variation of read count ratio with age. Color codes psychiatric diagnosis (Dx) consistent with Fig. 4.

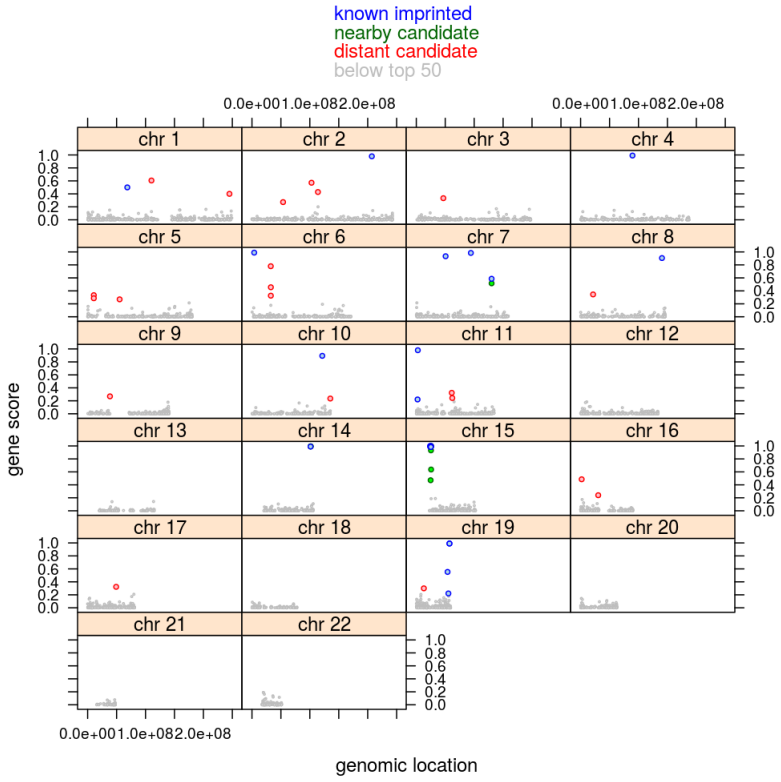

Figure S1: Clustering of top-scoring genes in the context of human DLPFC around genomic locations that had been previously described as imprinted gene clusters in other contexts.

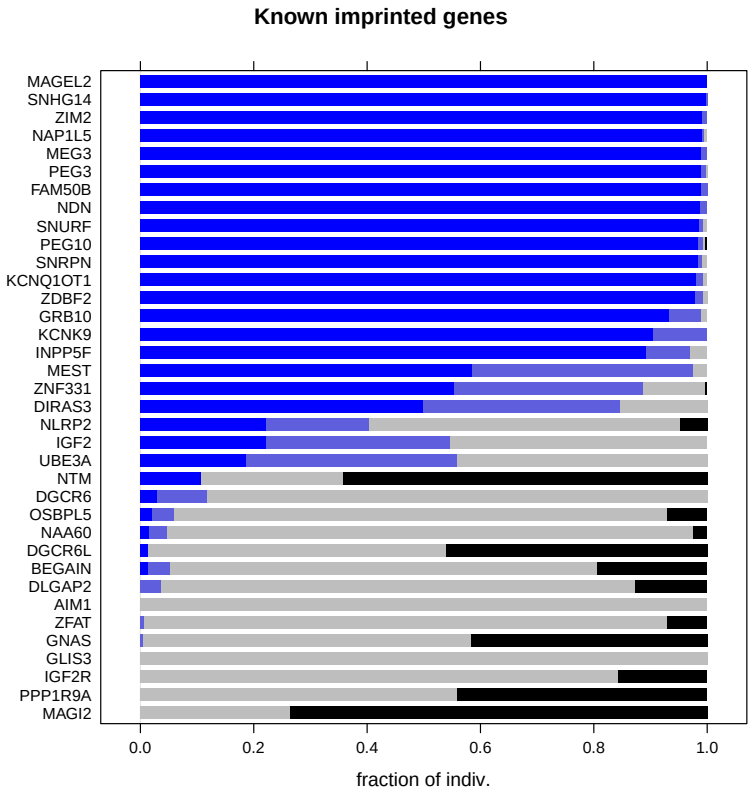

Figure S2: Known imprinted genes ranked by the gene score (dark blue bars). "Known imprinted" refers to prior studies on imprinting in the context of any tissue and developmental stage. The length of the black bars indicates the fraction of individuals passing the test of nearly unbiased expression.

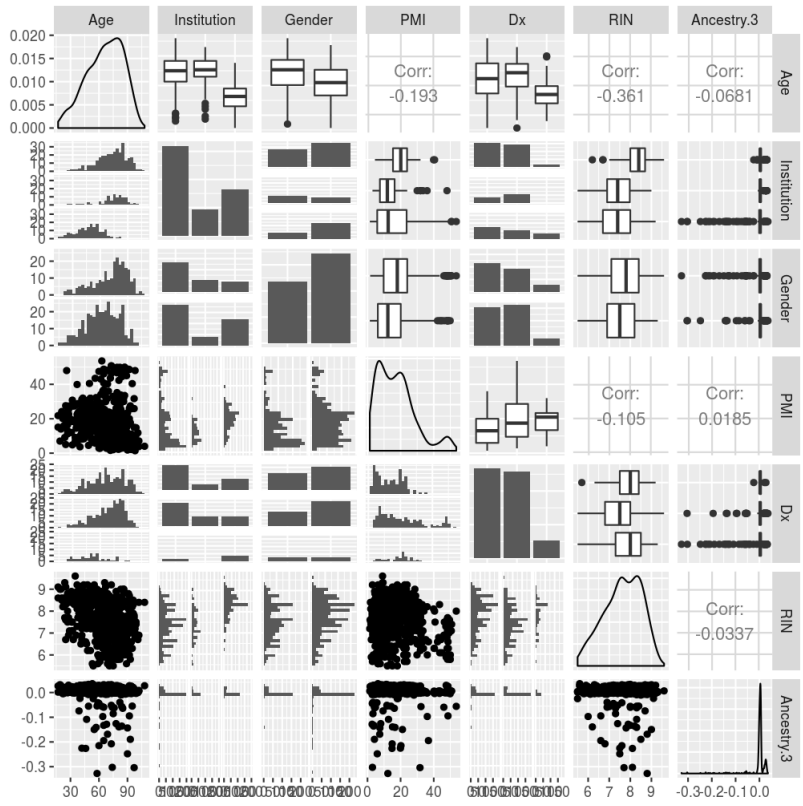

Figure S3: Distribution and inter-dependence of predictors. The diagonal graphs of the plot-matrix show the marginal distribution of six predictors (Age, Institution,...) while the off-diagonal graphs show pairwise joint distributions. For instance, the upper left graph shows that, in the whole cohort, 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.

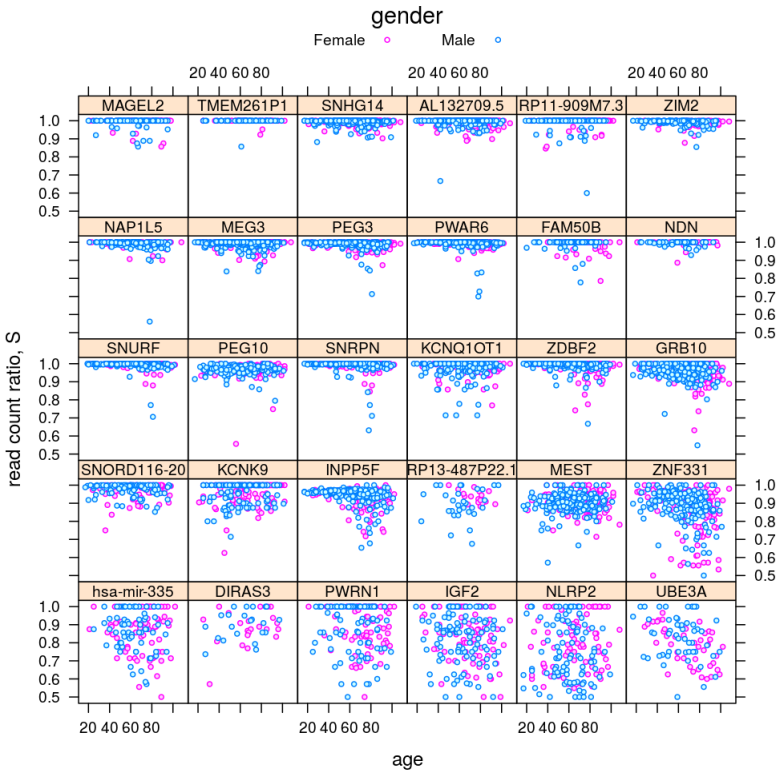

Figure S4: Variation of the read count ratio S_{ig} with age and gender across hundreds of individuals i (dots) and 30 genes g that have been considered as imprinted in the DLPFC brain area in this study.

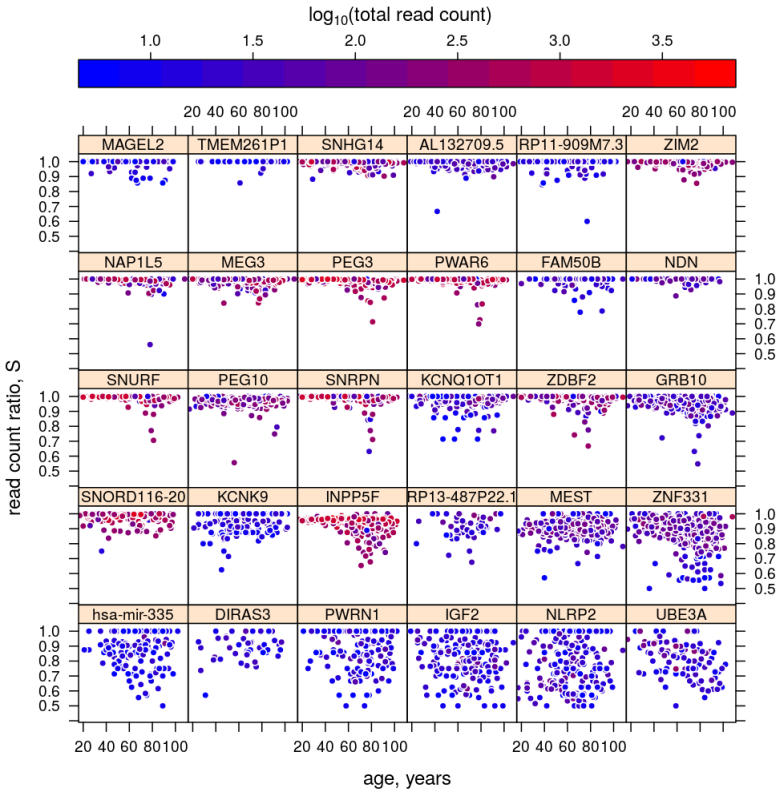

Figure S5: Variation of total RNA-seq read count across genes and individuals.

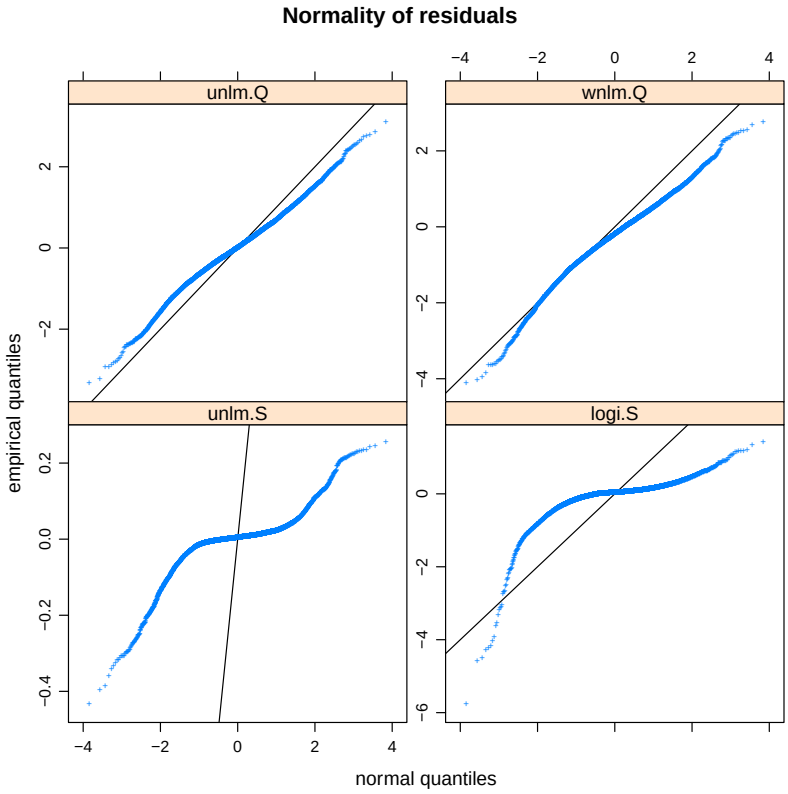

Figure S6: Checking the fit of wnlm.Q model: analysis of the normality of residuals.

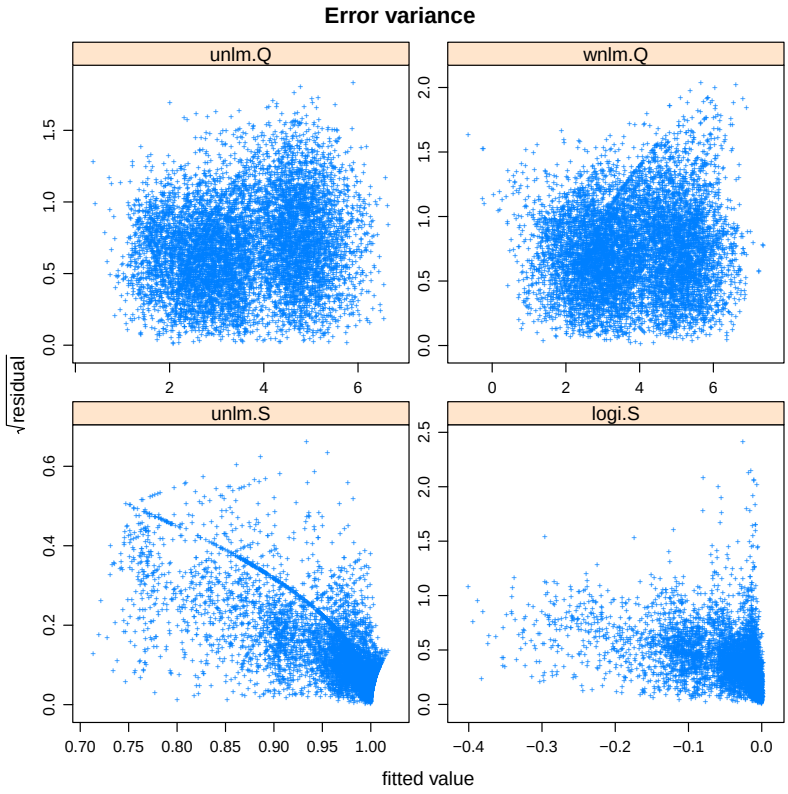

Figure S7: Checking the fit of wnlm.Q model: analysis of homoscedasticity.

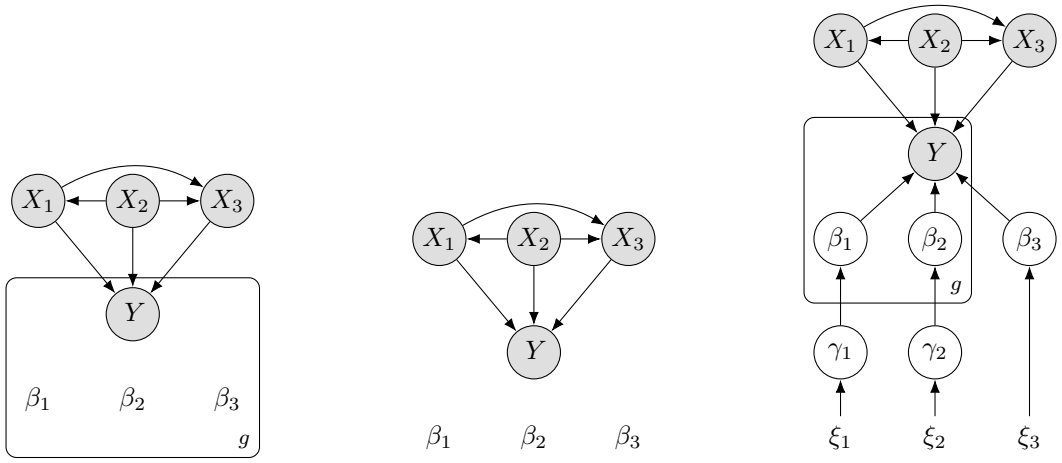

Figure S8: General dependency structure of three regression model frameworks. In all of threse model frameworks the regression coefficients $\beta_{1g}, ..., \beta_{3g}$ mediate, for a given gene g, probabilistic dependencies (arrows) between the response variable Y_g (read count ratio for g) and the corresponding predictors $X_1, ..., X_3$. For simplicity but without loss of generality only 3 predictors are depicted. The model frameworks differ in how $\beta_{jg_1}, \beta_{jg_2}, ...$ relate to each other for a given predictor (or a given j). *Left:* there is no connection among $\beta_{jg_1}, \beta_{jg_2}, ...$ which means that the way Y_g, the read count ratio for gene g depends on predictor X_j is completely separate from how the read count ratio for any other gene g' (i.e. $Y_{g'}$) depends on it. Consequently no information may be shared among gene-specific models. *Middle:* In this case $\beta_{jg_1} = \beta_{jg_2} = ... \equiv \beta_j$ so that all genes are identical with respect to how their read count ratio depends the predictors. Thus genes share all information in the data in the sense that the model forces them to be identical. *Right:* Hierarchical Bayesian model where genes show both variation as well as invariance with regards to depenencies. The variation is described by the dependence of β_j on the hyperparameter γ_j, whereas the invariance by the dependence of γ_j on *its* hyperparameter ξ_j. Only this model framework allows information sharing among genes in a flexible way.

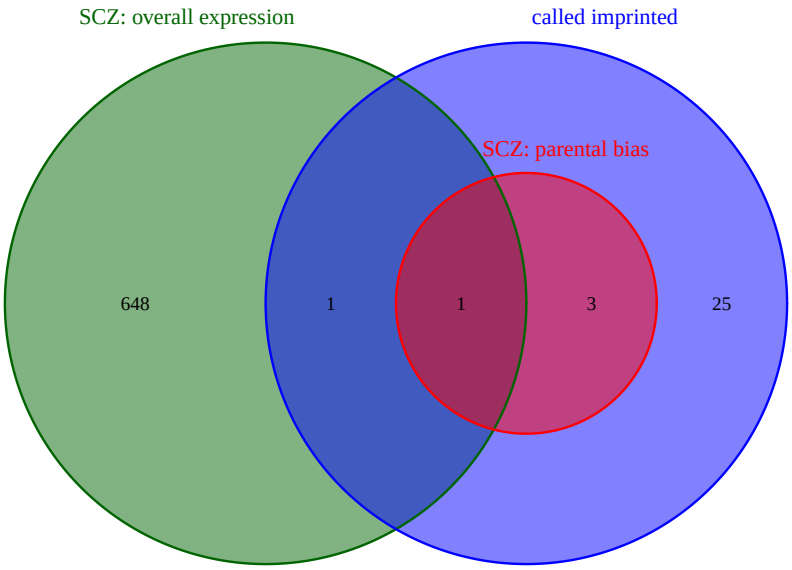

Figure S9: Association of genes' expression to schizophrenia (SCZ) assayed by two RNA-seq based approaches: total read count (overall expression, Nat Neurosci. 2016 Nov;19(11):1442-1453.) and read count ratio (allelic bias, present work). When these approaches are compared for only those genes that we find imprinted in the DLPFC in this study, 1 gene is found associated to schizophrenia by both approaches, 1 by only overall expression, and 3 by only allelic bias.