A Statistical View on the Monoallelic Expression Study

Attila G.K.

February 16, 2016

1 Rationale

In my understanding, the aims of the project[1] may be summarized as

1. classify (addressable) genes as mono or biallelically expressed in human DLFPC
2. investigate dependence on explanatory variables like age (regression)
3. estimate the fraction of monoallelically expressed genes

Some shortcomings (listed below) of the used approaches appear to hinder interpretation of results or weaken the conclusions already drawn from those.

- implicit specification of statistical models and underlying assumptions
- no model checking/comparison of alternative models
- no error rates were calculated for the detection of monoallelic expression
- consequently, only a semi-quantitative statement could be made on the small fraction of monoallelically expressed genes

2 Preliminaries

Genome-wide observations on m genes are based on post mortem tissue samples from the DLPFC (dorsolateral prefrontal cortex) of n individuals. The $n \times p$ design matrix X contains observations on all individuals and p *explanatory variables* including age of death and psychological condition (e.g. schizophrenia).

For each gene g and individual i inferred to be heterozygous for g a statistic S_{ig} was derived from the SNP-array and RNA-seq data based on read counts that contain any of the inferred SNPs in the (i, g) pair. Let N_{ig} be the total number of such counts (based on both alleles) and $H_{ig} = \sum_s H'_{is}$, where H'_{is} is the greater of the read counts for the two variants at SNP s; the summation runs over all inferred SNPs s (for individual i and gene g). Using the notations just introduced, the definition in the manuscript reads

$$S_{ig} = \frac{H_{ig}}{N_{ig}}.$$

[1] working title: Novel monoallelically-expressed genes and relaxation of imprinting with advanced age in the dorsolateral prefrontal cortex. See the text of manuscript under this link and the corresponding figures here

In this note I follow the previous analysis in that I assume that given X the random variables $\{S_{ig}\}_{ig}$ are sufficient statistics for the parameters θ (including the level of allelic exclusion) of all model structures under consideration. This admittedly false but attractively simple assumption means that the complete data (from the SNP-array and RNA-seq measurements) carry no more information on θ than $\{S_{ig}\}_{ig}$ do, so it is sufficient to draw inferences solely from the latter (in combination with X, if X is informative). I will not discuss sufficiency of $\{S_{ig}\}_{ig}$ further in this document.

3 Towards explicit model specification

To make the implicit model specification of the manuscript explicit, I will briefly expose here a few plausible model families. I will start from the simplest one, progressing towards generalized linear models. Along the way I will list mechanistic/biological assumptions behind that model, and sketch various modeling directions to relax some of those assumptions.

3.1 The simplest model

For all the model families considered here the following assumptions are made on the mechanism of allelic exclusion

- $\{S_{ig}\}$ are sufficient statistics (see above) for θ
- individuals are independent of each other

In case of the simplest model family, $\{S_{ig}\}$ are independently distributed according to two likelihood functions $f(\cdot|a)$ and $f(\cdot|b)$, corresponding to mono and biallelic expression, respectively:

$$\{S_{ig}\}_{ig} \overset{i.i.d.}{\sim} f(s|\theta_g) \tag{1}$$

$$\theta_g = \begin{cases} a & \text{when } g \text{ is monoallelically expressed} \\ b & \text{when } g \text{ is biallelically expressed.} \end{cases} \tag{2}$$

For this model framework the following additional assumptions must be made on allelic exclusion:

1. it takes only two levels (resulting in fully biallelic or monoallelic expression) of the same alleles in all cells on which the data are based on
2. all genes g are independent
3. X has no impact (i.e. age independence)

3.2 Directions for generalization

When assumption 1 above is relaxed to allow *multiple levels of allelic exclusion* within single cells and/or variation among cells, there is a separate θ_G parameter for each level, expressing the strength of allelic exclusion:

$$\{S_{ig}\}_{ig} \overset{i.i.d.}{\sim} f(s|\theta_G) \quad \forall g \in G.$$

The difficulty with this model family is twofold. First, the biological significance of different levels of θ_G seems vague. Second, the extent of monoallelic expression (genome-wide or restricted to addressable genes) cannot be expressed by a single number such as the expected frequency $\hat{\pi}_{\mathrm{m}}$ of monoallelically expressed genes that is applicable to the two level case. Here I will not pursue this direction further and continue by assuming two levels as before.

Dependence among genes (see assumption 2.) is known to exist because of extensive epigenetic marks for imprinting spanning multiple neighboring genes. The simplest model family for such dependence is a HMM framework. Emission probabilities for $\theta_{ig} \to S_{ig}$ are specified by $f(s|\theta_{ig})$ (cf. Eq. 1), and the hidden Markov chain is $\theta_{i1} \to \theta_{i2} \to ...$, where each θ_{ig} may only take the two values as in Eq. 2. Neither this direction is followed further here and I return to the independent genes scenario.

Until this point $\{S_{1g}, ..., S_{ng}\}$ were distributed identically and independently across individuals for any given gene g. That allows aggregating $\{S_{1g}, ..., S_{ng}\}$ by taking the average $\bar{S}_g = n^{-1} \sum_i S_{ig}$, which results an estimator whose standard error is diminished by $n^{-1/2}$ relative to S_{ig}. Given $n = 579$ this is a great improvement, yet the previous analysis did not take advantage of this.

3.3 Regression models

The following regression framework achieves a similar effect to averaging. Importantly, it also allows X to *impact allelic exclusion* (assumption 3.). The most simple model family in this case is normal linear regression. Let $f(\cdot|\mu, \sigma^2)$ denote the p.d.f. of the normal distribution with mean μ and variance σ^2. Then for a given individual i

$$\{S_{ig}\}_g \overset{i.i.d.}{\sim} f(s|x_i\beta_g, \sigma_g^2) \tag{3}$$

$$\beta_g, \sigma_g^2 = \begin{cases} a_1, a_2 & \text{when } g \text{ is monoallelically expressed} \\ b_1, b_2 & \text{when } g \text{ is biallelically expressed,} \end{cases} \tag{4}$$

where x_i is the i-th row of X and β_g is a p-length vector of regression coefficients.

In this model family the multi level explanatory variables x_i enter the model specification as scaling factors of the regression parameters β_g. Therefore S_{jg} may be distributed according to (many) more distributions than just two because f in Eq. 3 incorporates x_i. Yet, this model family does support binary classification since X is known and, as in Eq. 2, the unknown parameter(s) may only take two values.

Linearity and normal distribution may not hold[2] for S_{jg} and X. Generalized linear model families (among which normal linear models comprise just one family) may offer solutions then. In this more general model family Eq. 3 modifies to $\{S_{ig'}\} \overset{i.i.d.}{\sim} f(s|\theta_{g'}(x_i), \phi_{g'})$, such that $\theta_{g'}$ is a function of the explanatory variables x_i, and $g(\mathrm{E}[S_{ig'}]) = x_i\beta_{g'}$, where g is a link function.

3.4 The model family used in the project

The manuscript describes the used model in bits of varying details, using two different sets of interrelated summary statistics, in the context of different sets of genes and different tasks. Some details only turn out from the scrutiny of related R code[3].

[2] In fact they cannot exactly hold given theoretical considerations such as the boundedness or discrete nature of S_{jg}. Despite this, they may hold approximately well in some sense.

[3] I received Ifat's code from Andy via email on 2/4/16

statistic	S_{ig}	T_i (=LOI_R)
definition	based on the data (Eq. 1)	based on $\{S_{ig}\}_{ig}$ (Eq. 5)
distribution	$\{S_{ig}\}_{ig} \overset{i.i.d.}{\sim} f(s\|\theta)$	$T_i = x_i\beta + \epsilon_i;\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
model family	Eq. 1 with unspecified f	normal linear
parameters	a =?, b =? (Eq. 2)	least sq. estimate $\hat{\beta}, \hat{\sigma}^2$
applied to	all genes	8 selected monoallelic genes
task	calling monoallelic e.	age dependence

Table 1: Properties of the models used in the project

Taken together, the following model framework emerges (summarized in Table 3.4):

- two levels of allelic exclusion: monoallelic and biallelic expression (well suited for binary classification)
- for biallelic expression it is not clear (for me) that a model was specified at all during the course of the study; a confidence interval is mentioned on p6 of the manuscript but that appears to refer to distributions of read counts instead of those of $\{S_{ig}\}_{ig}$
- since the manuscript leaves the likelihood function f unspecified and mentions none of the generalizations of Section 3.2 and 3.3, I assume that the simplest model (Eqs. 1-2) describes best the intentions of the manuscript
- for monoallelic expression a normal linear model was used[4], not on $\{S_{ig}\}_{ig}$, but instead on a "loss of imprinting" statistic LOI_R, which I rename here to T_i to emphasize that each T_i is specific to individual i. T_i aggregates $\{S_{ig}\}_{ig}$ for a few (8) genes classified as monoallelically expressing
- the definition of T_i is essentially

$$\frac{1}{2}\left(\sum_{g=1}^{8} \hat{F}_g(s) + 1\right) \tag{5}$$

  where $\hat{F}_g$ is the empirical cumulative distribution function (e.c.d.f.) based on $\{S_{ig}\}_i$ for one of the 8 selected genes g; thus Eq. 5 gives the average e.c.d.f. whose range is scaled to $[0.5, 1]$ to match the same interval of possible values S_{ig} may take

As Table 3.4 shows, the two kind of statistics ($\{S_{ig}\}_{ig}$ and $\{T_i\}$) differ in their relation to the full data set and to each other, the type of probability model family and how exactly those probabilities are specified. Furthermore, they differ in what kind of analysis (task) they were used in and how broad set of genes they were applied to.

In section 4 I discuss those consequences of not specifying the *null distribution* $f(s|\theta = b)$ for the simple model (Eqs. 1-2), a term that in the present context names the distribution of $\{S_{ig}\}_{ig}$ for biallelically expressed genes g. These consequences concern two tasks: (i.) calling monoallelically expressed genes and (ii.) estimating their fraction in all (addressable) genes (i.e. the extent of monoallelic expression). Here I only suggest revision of the interpretation

[4]implemented in the `glm` R function, which was called in Ifat's code without `family` argument thus defaulting to `gaussian`, which is the normal linear model family.

		FDR	
λ	threshold α	$\pi_1 = 0.008$	$\pi_1 = 0.052$
20	10^{-2}	0.87	0.50
	10^{-4}	0.86	0.48
	10^{-8}	0.86	0.47
2000	10^{-2}	0.55	0.15
	10^{-4}	0.064	0.010
	10^{-8}	0.058	0.0090

Table 2: False discovery rate calculated using Eq under the mixture distribution function (Eq. 6) at various π_1 and λ values and various significance thresholds α.

of results (conclusions) but not that of the analysis itself, which would be desirable if stronger conclusions were to be made.

In section 5 I turn to the dependence of monoallelic expression on the explanatory variables X and suggest ways not only for strengthening and generalizing the result on age dependence, but also for unifying the modeling framework to allow the most direct, genome wide, biological interpretation of parameters.

4 Consequences of unspecified null distribution

4.1 Extent of monoallelic expression

Estimating the expected fraction [1]

$$h(p) = \pi_0 + \pi_1 \lambda (1 - e^{-\lambda}) e^{-\lambda p} \tag{6}$$

for $0 \leq p \leq 1;\ \lambda > 0;\ 0 < \pi_0 < 1$

4.2 Calling monoallelic expression

$$\Pr(\text{TP}) = \pi_1 [1 - (1 - e^{-\lambda})(e^{-\lambda\alpha} - e^{-\lambda})] \tag{7}$$

$$\text{FDR} = \frac{\Pr(\text{FP})}{\Pr(\text{FP}) + \Pr(\text{TP})} \tag{8}$$

5 More ideal selection

Scatter plots from the previous analysis indicate that the normal linear model would not fit well (see Figure 3). This qualitative result explains why that work transformed, for each g, $\{S_{1g}, ..., S_{ng}\}$ into percentiles before fitting the normal linear model. Still before fitting, that analysis averaged each percentile across 8 selected monoallelically expressed genes, which has a similar effect as constraining all monoallelically expressed to share parameters (Eq. 4)

References

[1] JD Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National …*, 100(16):9440–9445, aug 2003.

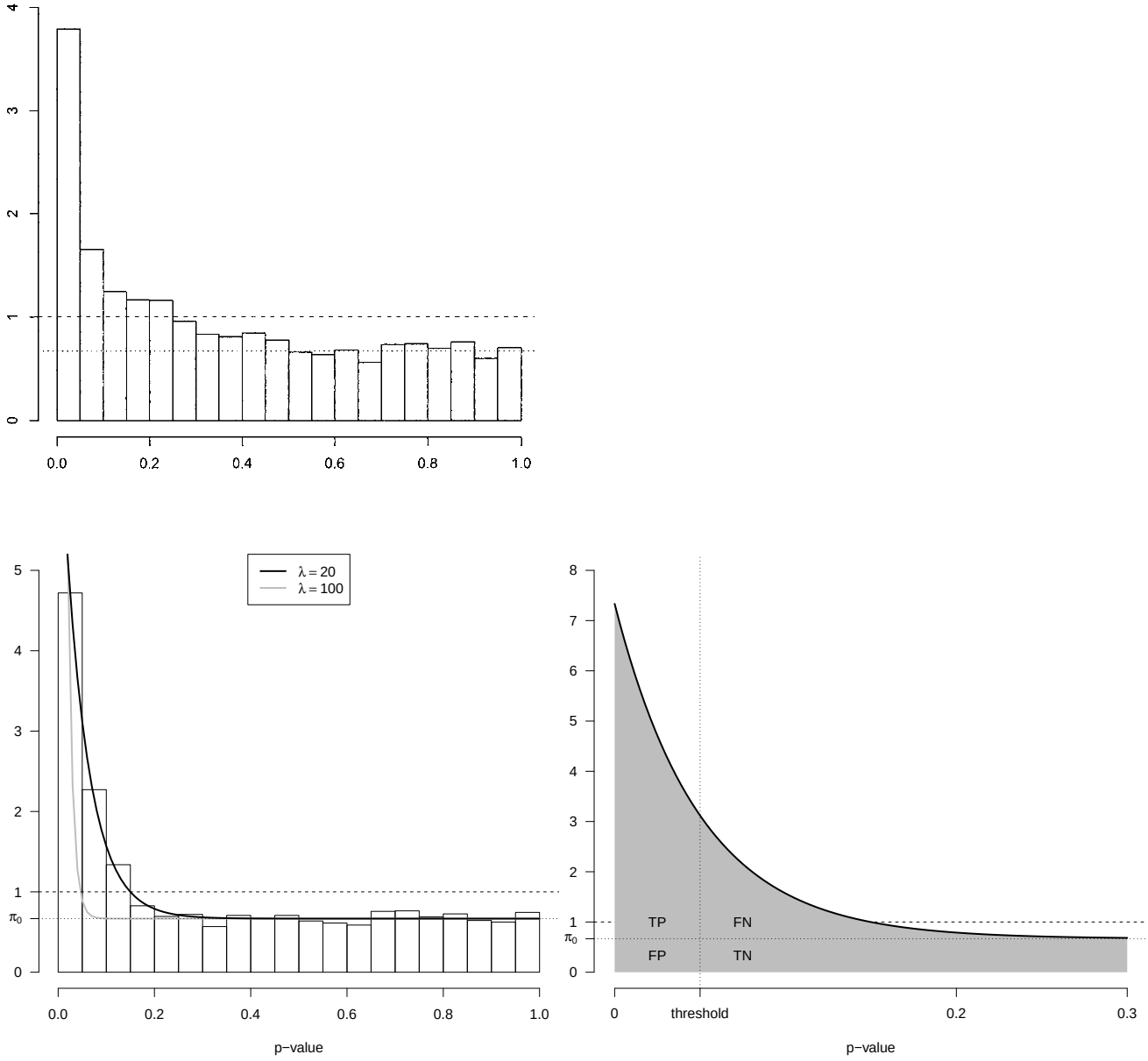

Figure 1: $\pi_0 : \pi_1$ mixtures of null and alternative distributions of p-values. *Top:* figure taken from ref. [1] showing 3170 p-values from a genomwide study with estimated $\pi_0 \approx 2/3$. *Bottom left:* the black and gray thick solid lines show the probability density function for two mixture distributions defined by Eq. 6, with the same $\pi_0 = 2/3$ but different λ values. The bars correspond to the histogram of a 3170-sized sample from the "black" distribution. *Bottom right:* the same "black" distribution function is expanded to illustrate the four outcomes of hypothesis testing; their probabilities equal the gray areas delineated by the dotted lines.