Project on Monoallelic Expression: a Statistical View

Attila Gulyás-Kovács

February 18, 2016

1 Introduction

The scope of this document is the research project on monoallelic expression in the human dorsolateral prefrontal cortex (DLPFC); I will refer to it as the MAE project and the corresponding draft as the MAE manuscript[1]. The term *allelic exclusion* will refer to mechanisms resulting in mono or biallelic expression and *imprinting* will denote the parent-of-origin specific subtype of allelic exclusion. My goal here is to evaluate the current state of the this study in order to facilitate discussion and propel this project to completion.

In the MAE project two kinds of analysis (task) was carried out:

1. *classification* to call monoallelic expressing genes in each individual
2. *regression*[2] to assess the impact of explanatory variables that vary across individuals

In my understanding, the main results and conclusions may be summarized as follows:

1. relatively few genes were called strongly significant suggesting the total number of monoallelically expressed/imprinted genes is consistent with the earlier conservative estimate of ca. 200 [1] as opposed to the liberal estimate of ca. 1300 [2]
2. the called genes agree well but not completely with previous gene sets suggesting some variation accross tissue types and/or organisms
3. the called genes varied across individuals; regression using 8 of the called genes suggests loss of imprinting with age

The present evaluation finds that the first two conclusions are at best weakened by the lack of estimated error rates of classification. The third point stands out as an interesting but only weakly supported novel finding that calls for improvements in terms of both its statistical significance and its generality.

Section 2 will introduce some quantities and concepts on the data and their summary statistics used by the MAE project. In terms of those statistics, Section 3 will present some plausible models of allelic exclusion that are both genome-wide and population-wide. These are not considered in the MAE manuscript but I include them here to help formalize the rather implicitly stated statistical frameworks of the MAE project in Section 4. That will

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

[2] Regression at this point means the inference of the parameters of some regression model while at some later points it will refer to the model itself. The meaning will hopefully be clear from the context.

pave the way for the reevaluation of the classification results in Section 5 and the regression analysis in Section 6. In both cases suggestions will be made for reinterpretation of results or reanalysis of data.

2 Preliminaries

Genome-wide observations on m genes were 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 (addressable) gene g, and for each individual i inferred to be heterozygous for g, a statistic S_{ig} was used for classification resulting in the set $\{S_{ig}\}_{ig}$. Each 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 MAE manuscript reads as

$$S_{ig} = \frac{H_{ig}}{N_{ig}}. \tag{1}$$

LOI_R[3] is another statistic defined by the MAE project, which was utilized in the regression analysis. I rename it here to T_i to emphasize that each T_i is specific to individual i. Scrutiny of some R code[4] related to the MAE project revealed that T_i was defined in terms of $\{S_{ig}\}_g$, where g is one of 8 selected genes among the previously known imprinted genes that the classification of the MAE project called as monoallelically expressing. For each of those genes g take the set $\{S_{ig}\}_i$ across all individuals i and, based on that, let $\hat{F}_g$ be the empirical cumulative distribution function (e.c.d.f.) evaluated at each data point; note the linear relation of $\hat{F}_g$ to ranks of individuals. Then the definition based on the R code is essentially

$$T \equiv \{T_i\}_i = \frac{1}{2}\left(\sum_{g=1}^{8} \hat{F}_g + 1\right) \tag{2}$$

In words, Eq. 2 shows that the e.c.d.f. was averaged over the 8 selected genes and after which it was scaled to $[0.5, 1]$ presumably in order to match the same interval as that containing the possible values of S_{ig}. Thus T is gene specific in the sense that it is based on only 8 genes sharing a property (inferred imprinting) but it is also gene unspecific in that it aggregates $\{S_{ig}\}_{ig}$ over those genes. Due to the limited applicability of T (only to those 8 genes) I will base on $\{S_{ig}\}_{ig}$ all statistical models described in Section 3.

There is a second reason motivated by the fact that the classification analysis of MAE project is entirely based on $\{S_{ig}\}_{ig}$ suggesting to consider them as sufficient statistics for the model parameter(s) θ. This 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 on θ solely from the latter (in combination with X, if X is informative). It is likely that sufficiency does not hold but will still be assumed for consistency with the MAE project and the assumption's simplicity. Sufficiency will not be discussed here further.

[3] loss of imprinting ratio?

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

3 Some plausible statistical models

I will start from the simplest model family, progressing towards generalized linear models. Along the way I will list biological-mechanistic assumptions behind each model family, and sketch various modeling directions to relax some of those assumptions.

3.1 The simplest model family

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

- $\{S_{ig}\}$ are sufficient statistics for θ (Section 2)
- individuals are independent of each other with respect to allelic exclusion

In case of the simplest model family, $\{S_{ig}\}_{ig}$ are independently distributed according to two probability density functions (p.d.f.)[5] $f(\cdot|a)$ and $f(\cdot|b)$, which correspond to mono and biallelic expression, respectively:

$$\{S_{ig}\}_{ig} \overset{i.i.d.}{\sim} f(s|\theta_g) \tag{3}$$

$$\theta_g = \begin{cases} a & \text{when } g \text{ is monoallelically expressed} \\ b & \text{when } g \text{ is biallelically expressed.} \end{cases} \tag{4}$$

Eq. 3 says that all S_{ig} are distributed independently identically (i.i.d.) according to probability distribution function f parametrized by θ_g whenever gene g is monoallelically (or biallelically) expressed (Eq. 4).

For this model family 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. with respect to allelic exclusion all genes g are independent
3. all monoallelically expressed genes are identical, and the same holds for the biallelic case
4. explanatory variables X (including age) have no impact on allelic exclusion

3.2 Directions for generalization

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

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

[5] Mathematical rigor would require the term probability *mass* function because each S_{ig} is discrete but I use p.d.f. for technical reasons too esoteric to be exposed here. f might also be called a likelihood function.

The difficulty with this model family is twofold. First, the biological significance of different levels of θ_G seems vague. Second, the fraction of monoallelically expressed genes (genome-wide or restricted to addressable genes) cannot be expressed by a single number as in the two level case (see π_1 in Section 5). Here I will not pursue this direction further and continue by assuming two levels as before.

Dependence among genes (assumption 2.) is known to exist because of extensive epigenetic marks for imprinting that span multiple neighboring genes. The simplest model family for such dependence is a class of hidden Markov models (HMMs). Emission probabilities for $\theta_{ig} \to S_{ig}$ are specified by $f(s|\theta_{ig})$ (cf. Eq. 3), 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. 4. Neither this direction is followed further here and I return to the independent genes hypothesis.

3.3 Regression models

The following regression models achieve an effect that is similar to averaging. Importantly, this model family also allows X to *impact allelic exclusion* (assumption 4.). The simplest model family in this case is normal linear regression. Let $f(\cdot|\mu, \sigma^2)$ denote the probability density function of the normal distribution with mean μ and variance σ^2. Then for a given individual i

$$\{S_{ig}\}_g \overset{i.i.d.}{\sim} f(s|x_i\beta_g, \sigma_g^2) \tag{5}$$

$$\beta_g, \sigma_g^2 = \begin{cases} a_1, a_2 & \text{when } g \text{ is monoallelically expressed} \\ b_1, b_2 & \text{when } g \text{ is biallelically expressed,} \end{cases} \tag{6}$$

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_{ig} may be distributed according to many more distributions than just two because f in Eq. 5 incorporates x_i. Yet, this model family still allows binary classification (Section 6) since X is known and, as in Eq. 4, the unknown parameter(s) may only take two values.

Linearity and normality may not hold[6] for S_{ig} 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. 5 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 (and g' denotes some gene).

A further generalization would be to allow direct dependence between the explanatory variables, e.g. age of death depends on gender. This would require Bayesian networks with the trade-off of higher model complexity.

4 The statistical frameworks of the MAE project

The MAE project based its previous classification and regression analysis on different, though closely related, statistics: $\{S_{ig}\}_{ig}$ (Eq. 1) and T (Eq. 2), respectively (Section 2). Besides that, the two kinds of analysis differ in their scope, dependencies, model formulation (or the lack of it), and consequently in parameter estimation and formal hypothesis testing. These differences delineate two distinct statistical frameworks, summarized in Table 1.

[6] In fact they cannot hold given the finite sample space of S_{ig}, but assuming them might be useful.

task/analysis	classification	regression on explanatory vars. X
goal/question	call monoall. expr.	impact of X on allelic exclusion
genomic scope	all (addressable) genes	8 selected monoallelic genes
statistic	S_{ig}	$T \equiv \{T_i\}_i$
dependencies	$\{S_{ig}\}_{ig} \overset{i.i.d.}{\sim} f(s\|\theta)$	$T_i = x_i\beta + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
model family	unspecified f	normal linear
estimation	not done	$\hat{\beta}, \hat{\sigma}^2$ by least squares
formal hypothesis test	not done	evidence for $\hat{\beta}_{\text{age}} < 0$

Table 1: The statistical framework of the two main tasks of the MAE project

How do these relate to the genome-wide and population-wide model families introduced in Section 3?

All (addressable) genes in all studied individuals were classified but no statistical model was formulated for classification. But if such a model had been given, it could have belonged to the simple model family in Eq. 3-4 for two reasons. Firstly, two levels of allelic exclusion were assumed (thus the binary classification). Secondly, all genes and individuals shared the same set of significance levels (e.g. $S_{ig} > 0.9$), suggesting $\{S_{ig}\}_{ig}$ may have been assumed i.i.d., so all monoallelically (or biallelically) expressing genes within all individuals were considered mechanistically independent and identical.

If at least the *null distribution* $f(s|\theta_g = b)$—the one that represents the biallelic case in Eq. 4 bottom—had been specified, binary classification could have been framed as frequentist hypothesis testing using p-values and other error rates. In Section 5 I discuss the consequences of classification without such error rates focusing on two inseparable questions of classification: (i.) the number of monoallelically expressed genes and (ii.) what fraction of monoallelically called genes are expected to be in fact biallelic, which is known as false discovery rate (FDR).

In contrast, the regression analysis in the MAE project did specify a model class. This is the normal linear family[7] explanatory variables x_i in each individual i. But this regression model must be distinguished from genome-wide and population-wide normal linear models of Section 3.3 on two grounds. First, in this case only 8 of the imprinted genes were modeled and none of the others (that are either mono or biallelically expressed, see Eq 6). Second, the response variable in this case was T_i instead of $\{S_{ig}\}_g$, raising two further questions: (i.) is T_i sufficient (like $\{S_{ig}\}_g$ are assumed to be, Section 2), and (ii.) how is the summation over 8 genes in the definition of T_i (Eq. 2) affect the inference of parameters β, σ^2? Section 6 will elaborate on these questions and possibly useful extensions to the current regression model.

5 Classification and its error rates

5.1 The inseparability of two questions

As mentioned above, the classification framework of the MAE project raises two questions, the first of which is the number of monoallelically expressed genes or, equivalently, their

[7] 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.

fraction

$$\pi_1 = \frac{\#\{\text{monoallelically expressed genes}\}}{m}, \tag{7}$$

where m—as before—is the total number of (addressable) genes.

Let $\pi_0 = 1 - \pi_1$. Then the test statistic S_{ig} is distributed as a $\pi_0 : \pi_1$ mixture of the null and alternative distribution denoted in Eq 3-4 as $f(s|b)$ and $f(s|a)$, respectively. Thus estimation of π_1 or π_0 requires the disentangling of $f(s|b)$ from $f(s|a)$ by fully specifying the former (the null) and making some minimal assumptions on the latter (the alternative). Then estimation of π_0 can be based on comparison between the empirical mixture distribution of S_{ig} and its theoretical null distribution as the classic article by Storey and Tibshirani [3] explains. Equivalently, the comparison may be based on corresponding empirical and theoretical distribution of p-values as shown by Fig. 1 *Top* taken from the same article.

The second question raised by the MAE projects' classification framework is that of misclassification rates. This is *inseparable* from the first question because

$$\pi_1 = (1 - \text{FDR})\frac{\#\{\text{+ calls}\}}{m} + (1 - \text{NPV})\frac{\#\{\text{- calls}\}}{m}, \tag{8}$$

where FDR means false discovery rate and $1 - \text{NPV}$ is known as false omission rate, and a "+" or "-" stands for a mono or biallelic call, respectively. These rates are obtained from the probability of the four outcomes (TP, FP, TN, FN) of binary classification:

$$\text{FDR} = \frac{\Pr(\text{FP})}{\Pr(\text{FP}) + \Pr(\text{TP})} \tag{9}$$

$$\text{NPV} = \frac{\Pr(\text{TN})}{\Pr(\text{TN}) + \Pr(\text{FP})} \tag{10}$$

Those four probabilities $\{\Pr(\text{TP}), ...\}$ are exactly the labeled gray areas in Fig. 1 *Bottom right*.

5.2 Illustration

I illustrate the practical impact of inseparability with a toy example by assuming that the theoretical mixture distribution of p-values has a known probability density function

$$h(p) = \pi_0 + \pi_1\lambda(1 - e^{-\lambda})e^{-\lambda p} \tag{11}$$

for $0 \leq p \leq 1;\ \lambda > 0;\ 0 < \pi_0 < 1$. While this simple analytical form is largely motivated by mathematical convenience, comparing the *Top* and *Bottom left* panels of Fig. 1 indicates that similar distributions have already been observed in previous genome-wide studies [3], although in some other context than allelic exclusion.

The form of the density h allows a closed form expression of the four probabilities mentioned earlier and the calculation of the rates FDR and NPV (Eqs. 9-10). For instance,

$$\Pr(\text{TP}) = \pi_1[1 - (1 - e^{-\lambda})(e^{-\lambda\alpha} - e^{-\lambda})], \tag{12}$$

where the classification threshold α is the hypothesis test's significance level[8] shown in Fig. 1 *Bottom right*.

[8] also known as the size of the test, its false positive rate, or the probability of type I error

Table 2 demonstrates that FDR is sensitive to π_1 illustrating the inseparability expressed by Eq. 8. The two levels of π_1 correspond to a previous conservative [1] and liberal [2] estimate of the number of imprinted genes. That the impact of π_1 (and of λ and α) was found to be much weaker on NPV than on FDR (not shown). This is because $\pi_1 \ll \pi_0$ even for the liberal estimate [2].

The following conclusions may be drawn from Table 2:

1. FDR is sensitive to π_1 (see above): it may be several fold lower with the liberal estimate [2] of π_1 than with the conservative one [1]
2. FDR widely varies with λ and α such that
   (a) greater λ corresponds to a greater difference between the null and alternative distributions and hence to a more accurate test (i.e. better classifier)
   (b) making the test stricter by decreasing α improves FDR but the rate of improvement diminishes with α in a way that depends on λ; this behavior follows from the functional form of Eq. 11 and need not be a general property of all hypothesis tests tests

5.3 Impact on current conclusions

In its Figure 1 and S4 the MAE manuscript presents classification results based on a set of three thresholds $\{t_1, t_2, t_3\}$, $t_1 = 0.9, ...$ for S_{ig}. This corresponds to an increasing sequence of *unknown* false positive rates $\{\alpha_1, \alpha_2, \alpha_3\}$ expressing decreasing statistical significance. Counting the genes that passed the highest threshold $t_1 = 0.9$ may seem informative on π_1, the fraction of monoallelically expressed genes. If error rates in Eq. 8 had been known or estimated in the MAE project *independently* of π_1, then that would be true. But such knowledge on error rates from independent source has been lacking. So, the only thing that would be informative on π_1 is the comparison of the empirical distribution of S_{ig} (or, equivalently, of the corresponding p-value) to its theoretical null distribution [3].

For quantifying the false discovery rate (FDR) at some classification threshold, that threshold must be given in terms of false positive rate α as in Table 2. In addition, π_1 must *also* be known (or estimated). So, again, the null distribution would be required for the derivation/estimation of both α and π_1 and consequently also for FDR.

		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 Eqs. 9 and 12 under the mixture distribution function defined in Eq. 11 at various π_1 and λ values and various significance levels α. $\pi_1 = 0.008$ and 0.052 correspond to previous estimate of ca. 200 and 1300 imprinted genes by refs. [1] and [2].

In summary, we cannot even roughly estimate how many of the genes called positive monoallelically expressing at, say, $t_1 = 0.9$ are expected to be biallelically expressing in reality. Strikingly, what prevents us from reaching such estimates is precisely the lack of knowledge on π_1 due to the inseparability discussed in Section 5. This undermines the conclusion of the MAE manuscript on the number of imprinted genes.

5.4 Suggestions on classification

Without knowing FDR what would be the best way of reporting our confidence in any of the novel genes called monoallelically expressing? I propose here a conditional approach given the previously identified imprinted genes. Significance of the novel genes could be quantified using quantiles of the e.c.d.f. of S_{ig} based on known imprinted genes. This would provide, for each gene g and individual i, an estimate for the minimum false negative rate of calling the (i, g) pair biallelically expressing when the monoallelic case is true in reality. However, those quantiles would say nothing about minimal false positive rates that is p-values.

The trade-off of the above approach would be treating the previously identified imprinted genes as an error-free gold standard for an incomplete set of monoallelically expressing genes in the human DLPFC. Besides the impact of possible errors in that set, this would prevent us addressing organism and tissue specificity of allelic exclusion.

Until this point $\{S_{1g}, ..., S_{ng}\}$ have been assumed to be distributed identically and independently across individuals for any given gene g. If $\{S_{1g}, ..., S_{ng}\}$ were to be used as estimators for some gene specific parameter θ_g, then it follows that the average $\bar{S}_g = n^{-1} \sum_i S_{ig}$ would result in an improved estimator in the sense that its standard error is diminished by $n^{-1/2}$ relative to S_{ig}. Given $n = 579$ this would be more than $20\times$ improvement.

This could be combined with the another suggestion under the hypothesis that for any given gene g the set $\{S_{ig}\}_i$ is i.i.d. (Eq. 3) both when g is bi and monoallelically expressed. Then averaging over individuals would give statistic $\bar{S}_g = n^{-1} \sum_i S_{ig}$ with the following benefit. Suppose $\{S_{1g}, ..., S_{ng}\}$ were to be used as estimators for some gene specific parameter θ_g reporting on whether g is bi or monoallelically expressed. Then $\bar{S}_g$ would result an improved estimator in the sense that its standard error is diminished by $n^{-1/2}$ relative to S_{ig}. Given $n = 579$ this would be more than $20\times$ improvement.

Clearly, the i.i.d. hypothesis does not hold if age or some other explanatory variables influence allelic exclusion. The previous and suggested (below) regression analyses in the MAE project test this hypothesis for variables internal to X (but provides no information on those external to X). Even if the i.i.d. hypothesis is rejected by the regression analysis, it may still be beneficial to use the aggregate statistic $\bar{S}_g$ for certain inferred values of parameters of the regression model. In that case, however, a normative way of classification would be framed as Eqs. 5-6 or the corresponding generalized linear models discussed next.

6 Improving the regression analysis

The generalized linear models of Section 3.3 may be used for three distinct tasks in the MAE project (using the notation of normal linear models given by Eqs. 5-6):

1. classification of gene g

   by a frequentist test : given (b_1, b_2) test if $(\beta_g, \sigma^2) = (b_1, b_2)$

by a Bayesian test : given both (b_1, b_2) and (a_1, a_2) test if $(\beta_g, \sigma^2) = (b_1, b_2)$ or if $(\beta_g, \sigma^2) = (a_1, a_2)$

2. inference of (b_1, b_2) (or that of (a_1, a_2)) given that g is known to be biallelically (or monoallelically) expressed

3. joint classification and inference

Without diving into details, the third task is the most challenging to implement yet in theory this would suite MAE project the best since neither sets of conditions of the first two tasks hold *a priori*, at least not on a genome-wide scale. Each of those tasks requires a piece of information (the conditions above). If one of those pieces is known the other may be obtained. But if neither are known then external source of information is needed due to the reciprocity of obtaining the conditions based on each other.

The external information source is available for the second task (inference) in form of previously identified monoallelically expressed genes but is unavailable for classification because nothing is known about the impact of explanatory variables X. I will focus here on the second task for the above reasons and also because the previous regression analysis of the MAE project implemented a special case of this task in several senses. (Section 4).

In the first sense, regression parameters were only inferred for the case when $g \in G_8$ where G_8 is the set of 8 selected genes known to be imprinted. In the second sense, T_i was used instead of $\{S_{ig}\}_{g:g\in G_8}$ raising the question of sufficiency and the effect of the aggregation (see the definition of T_i in Eq. 2) on inference. Finally, a normal linear model was assumed, which is only a special case of generalized linear models.

All three points call for modifications of and extensions to the previous regression analysis. Obviously, assessing the impact of age and other explanatory variables on a set G many more than 8 known imprinted genes appears desirable. The concerns of sufficiency and aggregation could be avoided by using simply $\{S_{ig}\}_{g:g\in G}$ instead of T_i as response variables.

The question of normality and linearity is an important concern given the scatter plots of Figure 3 of the MAE manuscript. This qualitative result might have motivated the transformation of $\{S_{1g}, ..., S_{ng}\}$ into e.c.d.f. as part of the definition of T_i (Eq. 2) in order to improve the fit of the normal linear model.

But normality and linearity could be addressed with some type of generalized linear model without the risk of loosing information with some transformation of $\{S_{ig}\}_{g:g\in G}$. The optimal model type could be addressed by selection based on criteria (like AIC or BIC) that incorporate both model fit to data and model complexity.

References

[1] 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.

[2] 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.

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

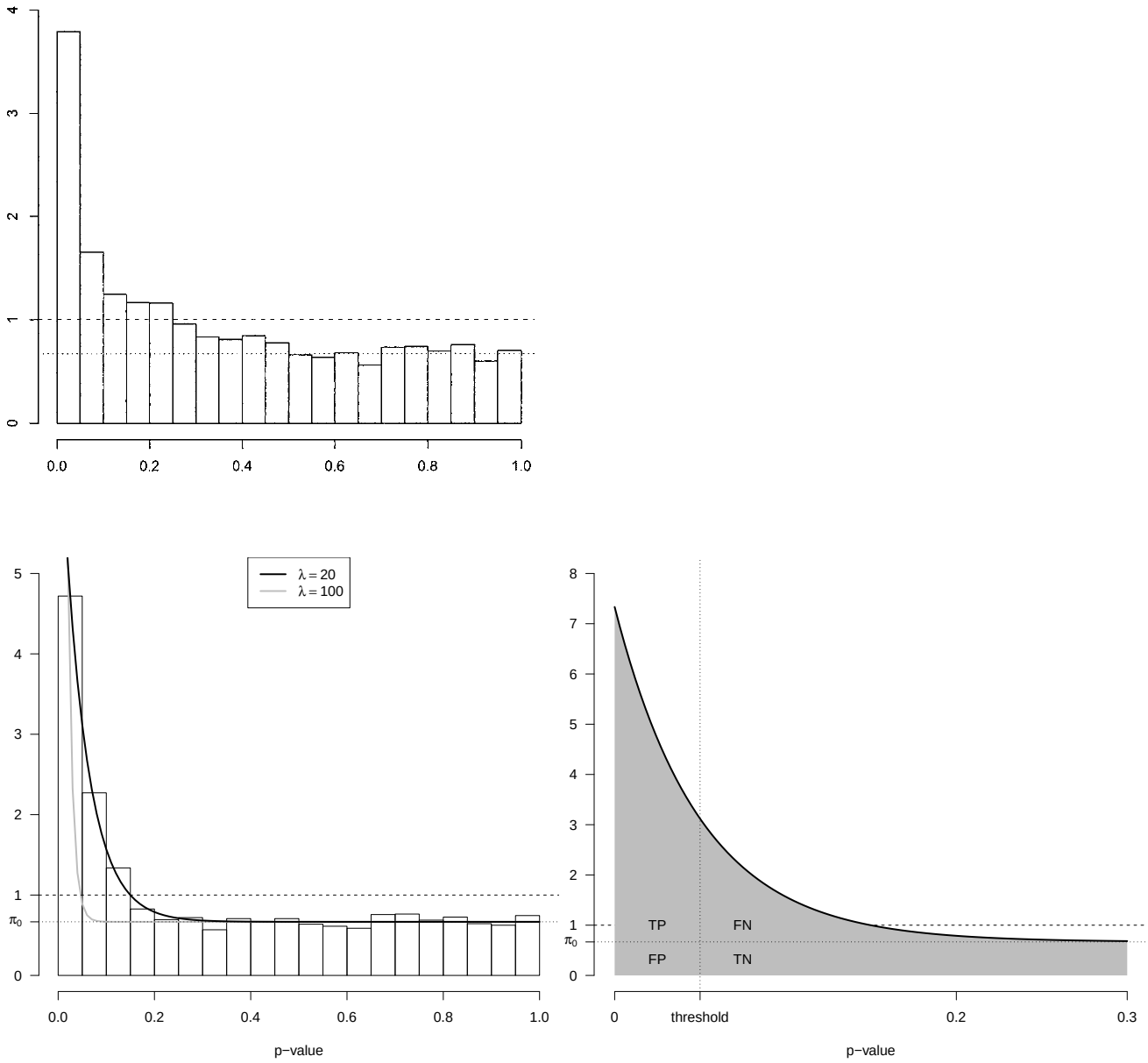

Figure 1: $\pi_0 : \pi_1$ mixtures of null and alternative distributions of p-values. *Top:* figure taken from ref. [3] showing 3170 p-values from a genome-wide study with estimated $\pi_0 \approx 2/3$ marked by the dotted line. *Bottom left:* the black and gray thick solid lines show the probability density function for two mixture distributions defined by Eq. 11, 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 on an expanded scale to illustrate the four outcomes of hypothesis testing; their probabilities equal the gray areas delineated by the dotted lines.