

# Supplement to 'Testing the significance of interactions in genetic studies using interaction information and resampling technique'

Paweł Teisseyre <sup>\*1</sup>, Jan Mielniczuk <sup>1, 2</sup>, and Michał J. Dąbrowski<sup>1</sup>

<sup>1</sup>Institute of Computer Science, Polish Academy of Sciences, Poland

<sup>2</sup>Warsaw University of Technology, Faculty of Mathematics and Information Science, Poland

## 1 Simulation models for power comparison

**Simulation model M1:**

$$s(y) = \begin{cases} p, & \text{if } y = -1, \\ 0, & \text{if } y = 0, \\ p, & \text{if } y = 1. \end{cases}$$

**Simulation model M2:**

$$s(y) = \begin{cases} 0, & \text{if } y = -1, \\ p, & \text{if } y = 0, \\ p, & \text{if } y = 1. \end{cases}$$

**Simulation model M3:**

$$s(y) = \begin{cases} 0, & \text{if } y = -1, \\ p, & \text{if } y = 0, \\ 0, & \text{if } y = 1. \end{cases}$$

**Simulation model M4:**

$$s(y) = \begin{cases} 0, & \text{if } y = -1, \\ 0, & \text{if } y = 0, \\ p, & \text{if } y = 1. \end{cases}$$

## 2 Detailed analysis of gene expression data of CD4+ T cells

In the following we described in detail an experiment on ImmVar data concerning co-receptor CD4+ T lymphocytes [1]. There are 236 gene transcripts measured for five stimulation conditions as well as phenotypic characteristics (1250 variables) for 348 donors of Caucasian (183), African-American (91) and Asian (74) ethnicities. We focused on detection of gene-gene interactions that

---

\*Correspondence to: Paweł Teisseyre, Institute of Computer Science, Polish Academy of Sciences, 5, Jana Kazimierza, 01-248 Warsaw, Poland, e-mail: teisseyrep@ipipan.waw.pl

are associated with the specific ethnicity ( $Y$  variable). We evaluated all pairs of variables using the HYBRID procedure. We detected three interesting interactions among pairs (i) Fatty Acid Desaturase 2 (*FADS2*) and Interferon Induced Transmembrane Protein 3 (*IFITM3*); (ii) *IFITM3* and Steroid 5 Alpha-Reductase 3 (*SRD5A3*); (iii) Interferon Induced Transmembrane Protein 1 (*IFITM1*) and *IFITM3*. We refer to the supplement for detailed description of the analysis.

In the HYBRID method, IICHI option was chosen for 95% of pairs. In the following we focus on the three top interactions, i.e. those having the smallest p-values: (i) (*FADS2*) and (*IFITM3*); (ii) *IFITM3* and (*SRD5A3*); (iii) (*IFITM1*) and *IFITM3*. We stress that due to the method applied we can be reasonably sure that the occurrence of *IFITM3* in all three pairs is *not* due to its main effect. Indeed, the mutual information between *IFITM3* and ethnicity is only 0.041. Figures 7, 8 and 9 show the joint and conditional probabilities for the pairs. Note that the plots show strong dependence of conditional distributions of *FADS2* and *IFITM3* given ethnicity which confirms existence of strong interaction between these predictors. It is worth noting that the dependence is significantly stronger for Africans and Asians than for Caucasians. For Africans and Asians, the low expression of *IFITM3* is associated with the low expression of the second gene in all analyzed pairs. For high *IFITM3* expression there is no evident pattern of the other gene expression.

Gene *SRD5A3* encodes a protein that is essential for production of androgen 5-alpha-dihydrotestosterone (*DHT*) from testosterone. It is as well involved in conversion of polyprenol into dolichol [2]. Dolichol and its analogues play a significant role in the organization and packaging of phospholipids in the biological membrane. They increase its fluidity and permeability. It seems that because of its impact on the membrane, dolichol can play an important role in the transport between Golgi apparatus, cytoplasmic membrane and lysosomes. Gene *FADS2* encodes a protein which has enzymatic properties. It regulates desaturation of fatty acids through the introduction of double bonds between defined carbons of the fatty acyl chain. *FADS2* is involved in biosynthesis of unsaturated fatty acids including metabolism of alpha-linolenic (omega3) and linoleic (omega6) acids [3]. Each of the presented here genes might be found in ethnicity related studies exhibiting their natural variability among populations as well as research focused on diseases. To confirm genes relation with ethnicity, population and diseases we performed a search using National Center for Biotechnology Information (NCBI) resources using R package `rentrez`. We checked the co-occurrence of each of the four genes, within a title or an abstract of the scientific papers, with each of the following terms: 'disorder', 'ethnicity', 'population'. Sum of all papers for all terms was 144 for *FADS2*, 70 for *IFITM3*, 47 for *IFITM1* and 14 for *SRD5A3*.

## References

- [1] C. J. Ye, T. Feng, H.-K. Kwon, T. Raj, M. T. Wilson, N. Asinovski, C. McCabe, M. H. Lee, I. Frohlich, H.-i. Paik, N. Zaitlen, N. Hacohen, B. Stranger, P. De Jager, D. Mathis, A. Regev, and C. Benoist. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*, 345(6202):1254665–1254665, 2014.
- [2] V. Cantagrel et al. SRD5A3 Is Required for Converting Polyprenol to Dolichol and Is Mutated in a Congenital Glycosylation Disorder. *Cell*, 142(2):203–217, 2010.
- [3] G. Lan et al. Identification of the  $\Delta$ -6 desaturase of human sebaceous glands: Expression and enzyme activity. *Journal of Investigative Dermatology*, 120(5):707 – 714, 2003.

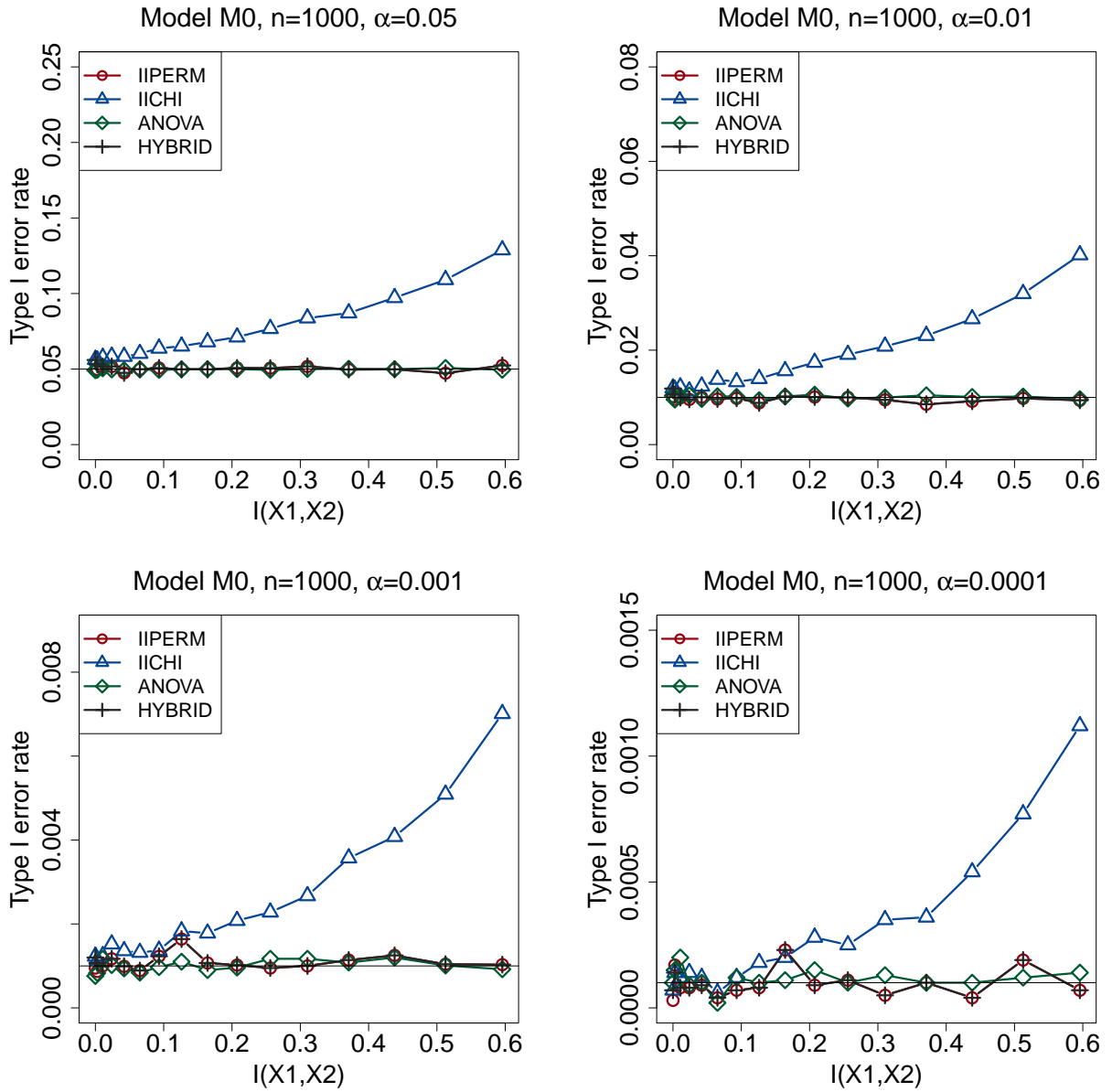


Figure 1: Type I error rate with respect to the mutual information for the simulation model M0, for  $\alpha = 0.05, 0.01, 0.001, 0.0001$  and  $n = 1000$ .

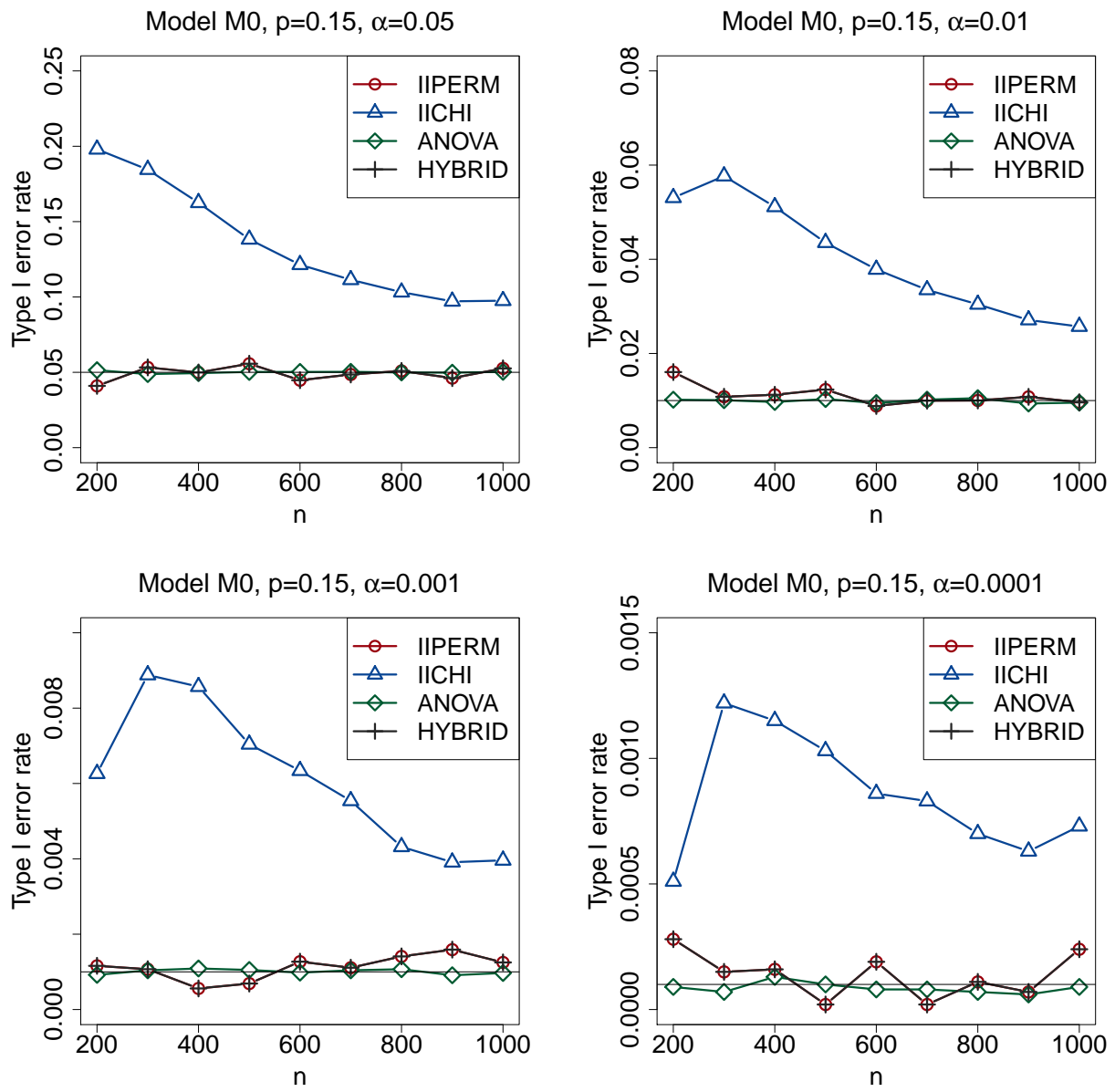


Figure 2: Type I error rate with respect to sample size  $n$  for simulation model  $M_0$ , for  $\alpha = 0.05, 0.01, 0.001, 0.0001$  and  $p = 0.15$ .

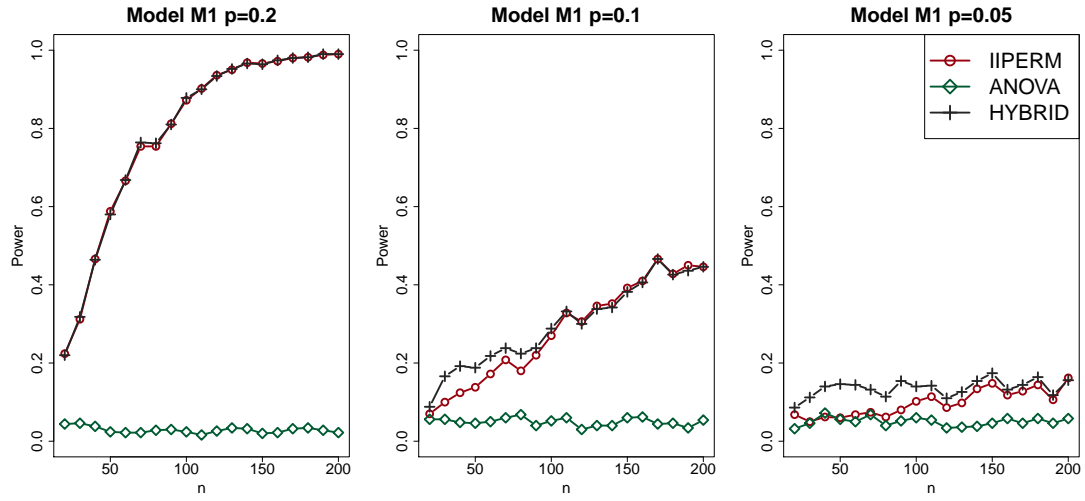


Figure 3: Power with respect to the sample size  $n$  for a simulation model M1.  $II(X_1, X_2, Y) = 0.1891, 0.0368, 0.0085$ , for  $p = 0.2, 0.1, 0.05$ , respectively.

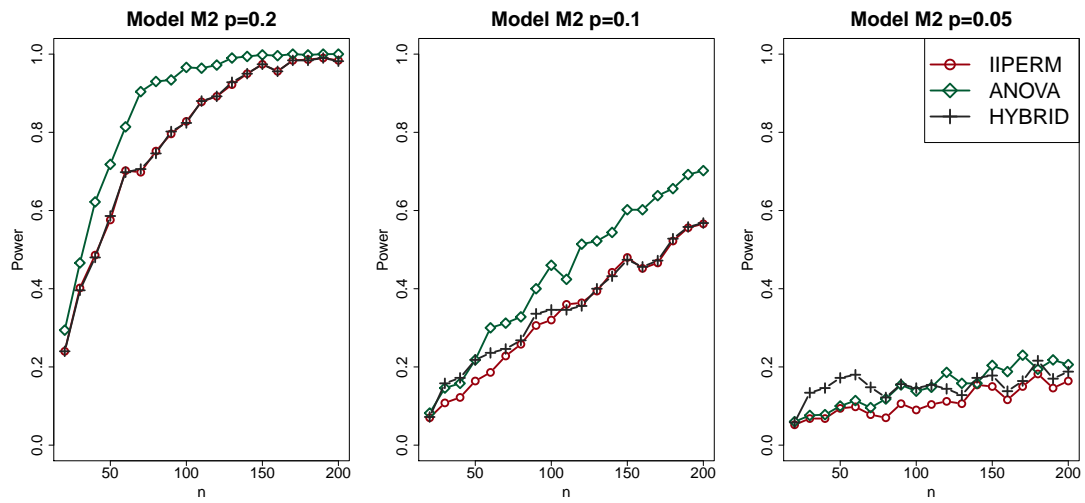


Figure 4: Power with respect to the sample size  $n$  for a simulation model M2.  $II(X_1, X_2, Y) = 0.1891, 0.0368, 0.0085$ , for  $p = 0.2, 0.1, 0.05$ , respectively.

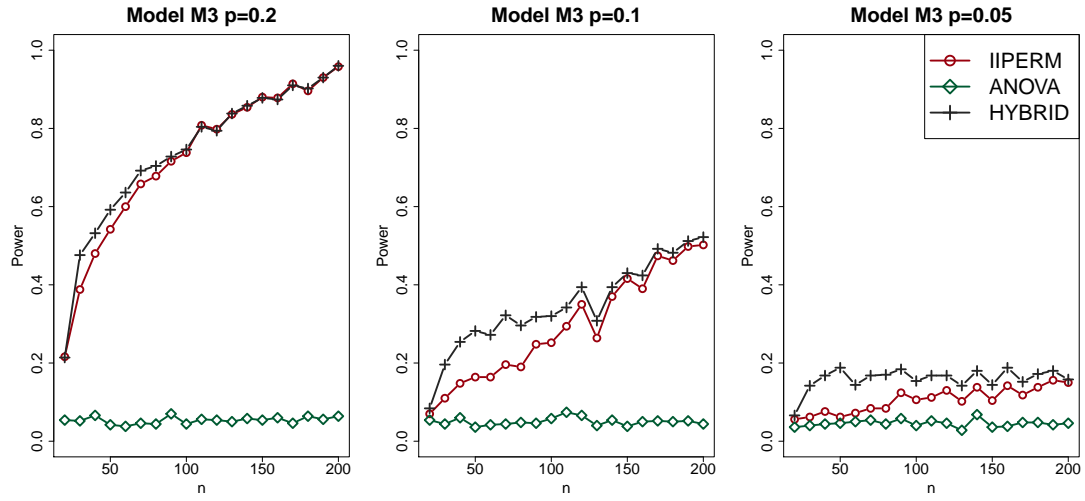


Figure 5: Power with respect to the sample size  $n$  for a simulation model M3.  $II(X_1, X_2, Y) = 0.173, 0.0371, 0.0086$ , for  $p = 0.2, 0.1, 0.05$ , respectively.

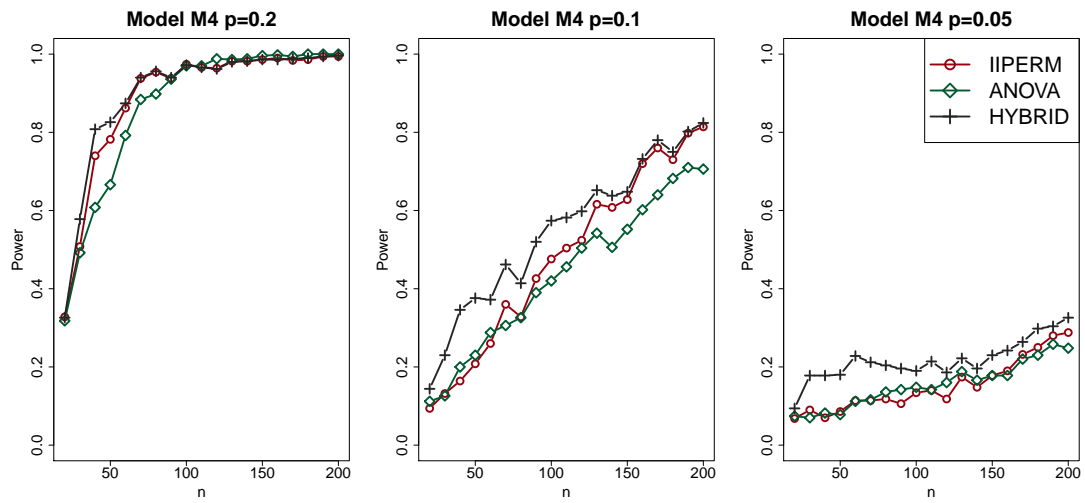
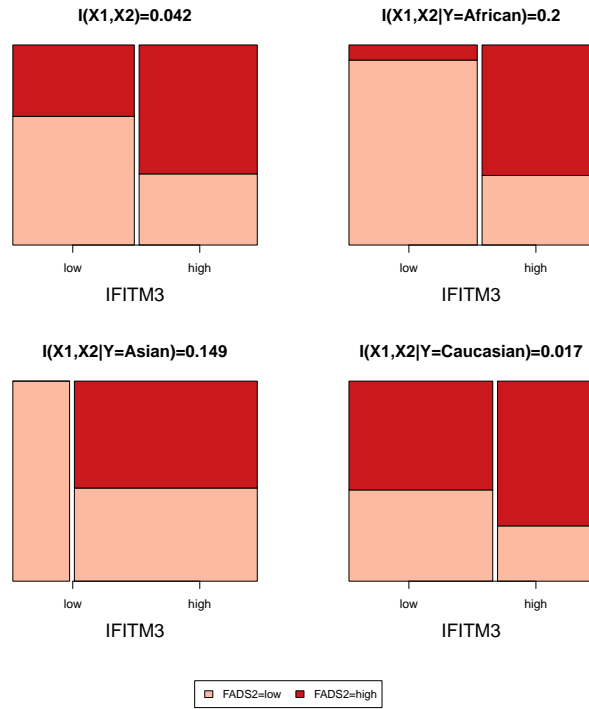
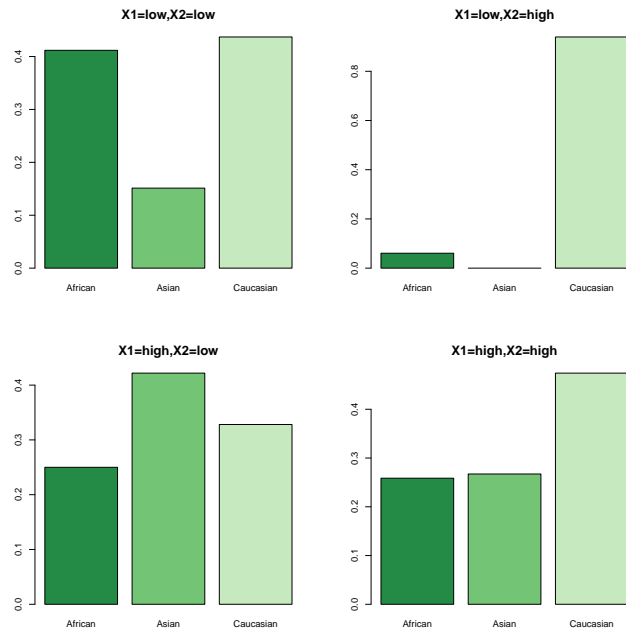


Figure 6: Power with respect to the sample size  $n$  for a simulation model M4.  $II(X_1, X_2, Y) = 0.173, 0.0371, 0.0086$ , for  $p = 0.2, 0.1, 0.05$ , respectively.



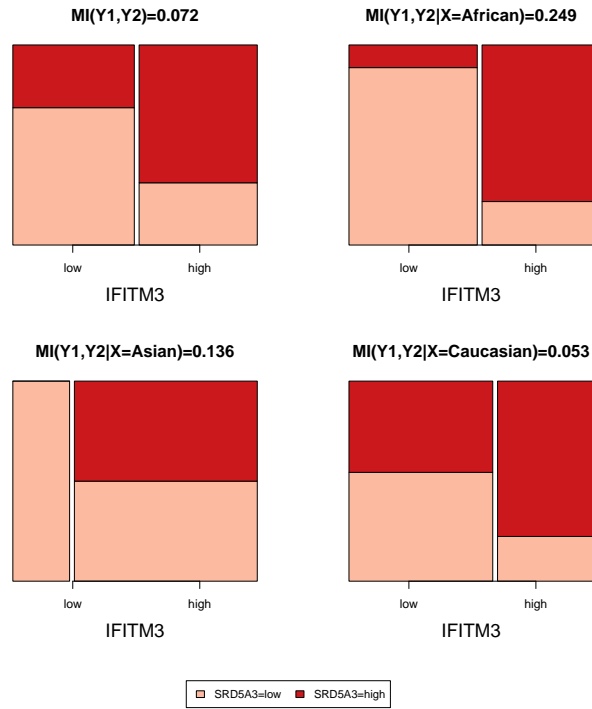
(a)



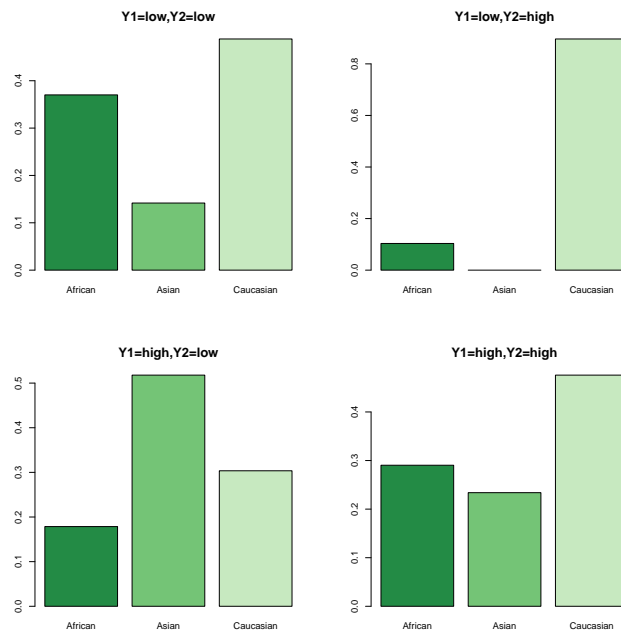
(b)

Figure 7: CD4+ data set. (a) Joint and conditional probabilities for the pair  $(X_1, X_2) = (IFITM3, FADS2)$ . (b) Distribution of ethnicity for four combinations of expression levels of genes  $(X_1, X_2) = (IFITM3, FADS2)$ .



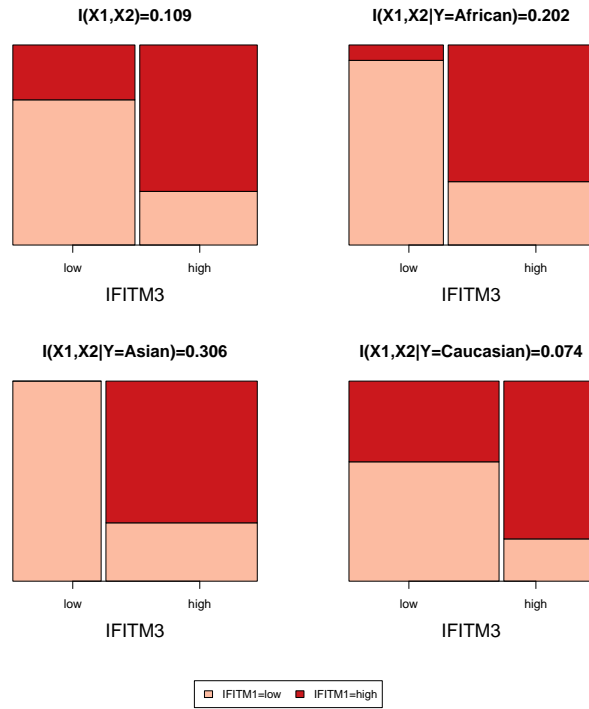


(a)

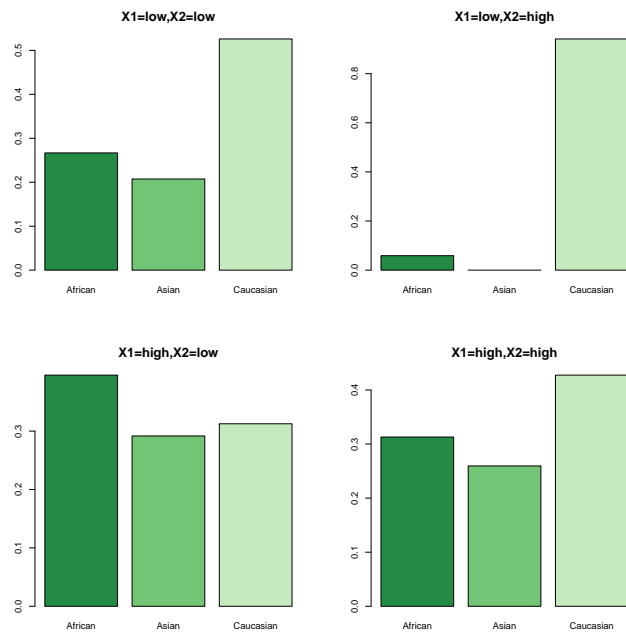


(b)

Figure 8: CD4+ data set. (a) Joint and conditional probabilities for the pair  $(X_1, X_2) = (IFITM3, SRD5A3)$ . (b) Distribution of ethnicity for four combinations of expression levels of genes  $(X_1, X_2) = (IFITM3, SRD5A3)$ .



(a)



(b)

Figure 9: CD4+ data set. (a) Joint and conditional probabilities for the pair  $(X_1, X_2) = (IFITM3, IFITM1)$ . (b) Distribution of ethnicity for four combinations of expression levels of genes  $(X_1, X_2) = (IFITM3, IFITM1)$ .