Adaptive combination of Bayes factors as a powerful method for the joint analysis of rare and common variants

Wan-Yu Lin, Wei J. Chen, Chih-Min Liu, Hai-Gwo Hwu, Steven A. McCarroll, Stephen J. Glatt, Ming T. Tsuang

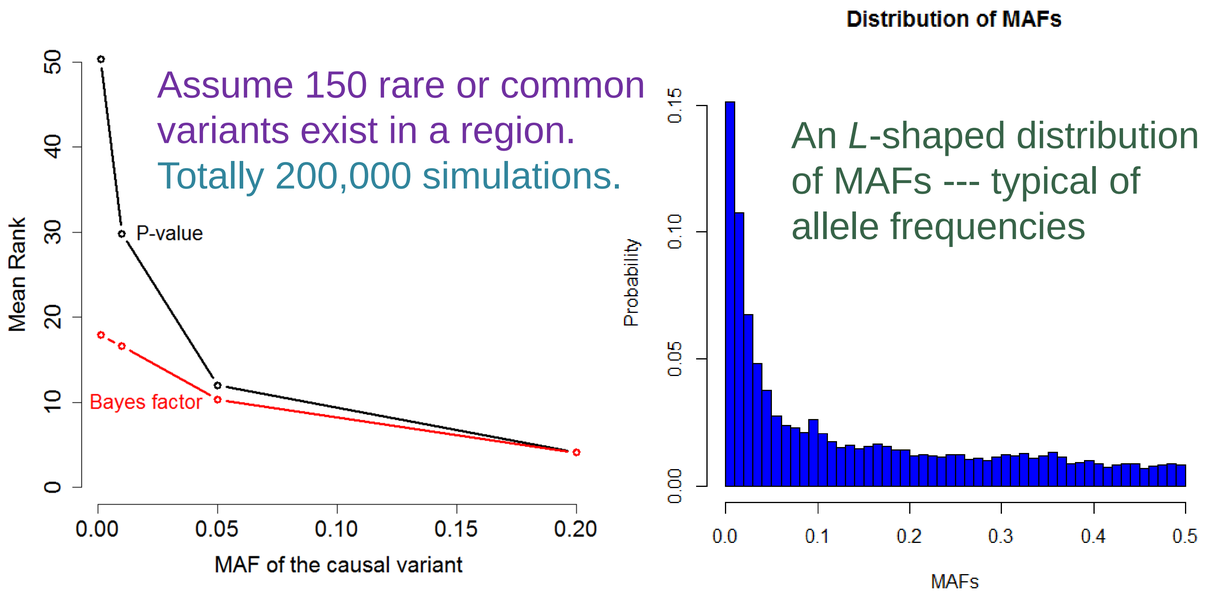
Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

Background: Multi-marker association tests can be more powerful than single-locus analyses because they aggregate the variant information within a gene/region. However, combining the association signals of multiple markers within a gene/region may cause noise due to the inclusion of neutral variants, which usually compromises the power of a test. To reduce noise, the "adaptive combination of *P*-values" (ADA) method [1] removes variants with larger *P*-values. However, when both rare and common variants are considered, it is not optimal to truncate variants according to their *P*-values.

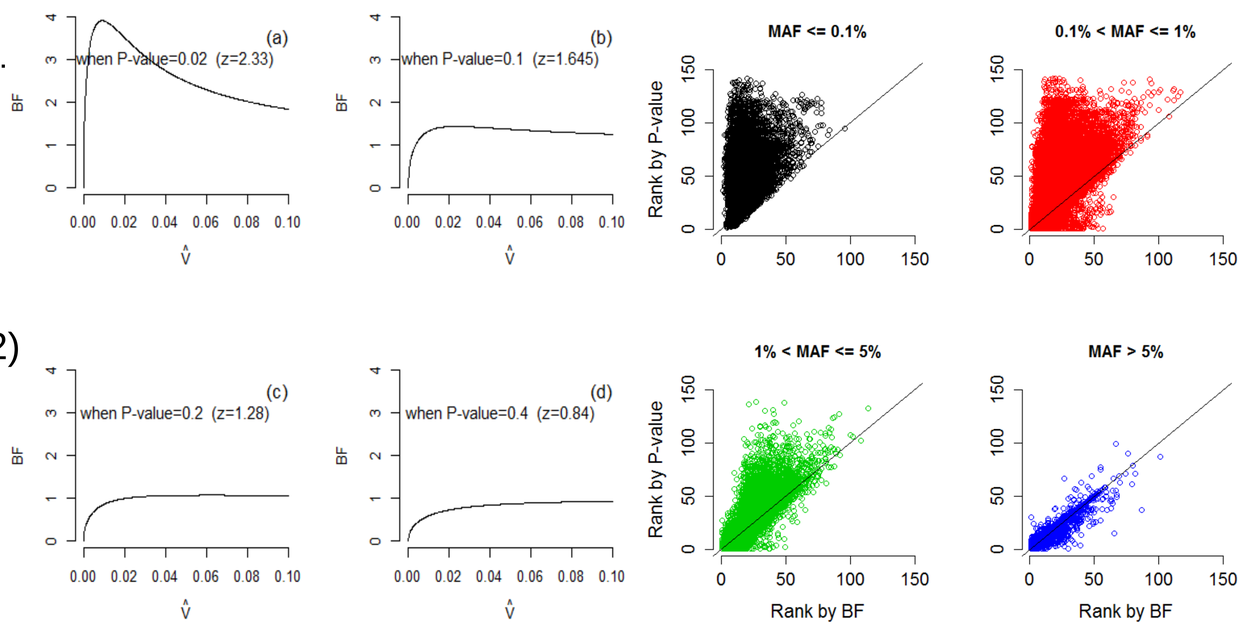
Methods: An alternative summary measure, the Bayes factor (BF), is defined as the ratio of the probability of the data under the alternative hypothesis to that under the null hypothesis. The BF quantifies the "relative" evidence supporting the alternative hypothesis. Here, we propose an "adaptive combination of Bayes factors" (ADABF) method that can be directly applied to variants with a wide spectrum of minor allele frequencies.

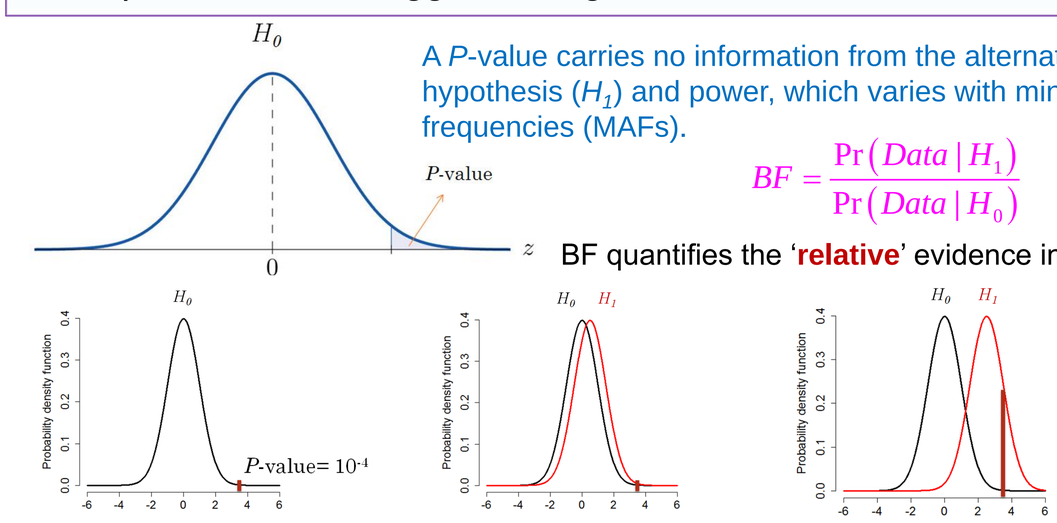
Conclusions: The simulations show that ADABF is more powerful than single-nucleotide polymorphism (SNP)-set kernel association tests and burden tests. We also analyzed 1,109 case-parent trios from the Schizophrenia Trio Genomic Research in Taiwan. Three genes on chromosome 19p13.2 were found to be associated with schizophrenia at the suggestive significance level of $5*10^{-5}$.

This work is forthcoming in the *Scientific Reports*. The paper can be downloaded from http://homepage.ntu.edu.tw/~linwy/ADABF.pdf
The R code of our ADABF method can be downloaded from http://homepage.ntu.edu.tw/~linwy/ADABF.html



- ✓ A smaller rank would be better, meaning that the causal variant would be ranked in priority order.
- ✓ The mean rank of the causal variant by the BF was smaller than (or equal to) that by the *P*-value, across all ranges of causal-allele frequencies.
- ✓ As the MAF of the causal variant increased, the power to detect that causal variant also increased and both mean ranks improved.
- ✓ Compared with the *P*-value ranking, rare causal variants will benefit from the BF ranking. (also can be seen from the bottom-right figure)





A *P*-value carries no information from the alternative hypothesis (H_1) and power, which varies with minor allele frequencies (MAFs).

$$BF = \frac{\Pr(Data \mid H_1)}{\Pr(Data \mid H_0)}$$

BF quantifies the '**relative**' evidence in favor of H_1.

- ➢ A rare causal variant generally has a larger *P*-value (say, *P*-value = 0.2) and a larger $\hat{V}$ (say, 0.1). Its BF will be larger than that of a common neutral variant with the same *P*-value but a smaller $\hat{V}$ (say, 0.005). (Please see the right figure, (c))
- ➢ That is, a common variant with a *P*-value = 0.2 may actually be a neutral variant, because this large *P*-value is obtained from reliable information (smaller $\hat{V}$).
- ➢ However, a rare variant with a *P*-value = 0.2 may still be causal, because this large *P*-value is obtained from less reliable information (larger $\hat{V}$). Rare variants seldom have small *P*-values, and therefore, our previous ADA method [1] prioritizes the rare variants with *P*-values smaller than 0.2.
- ➢ However, in a region with a mixture of rare and common variants, a *P*-value threshold of 0.2 is too liberal for common variants. In this situation, it will be better to consider the "relative" evidence in favor of (i.e., BF), instead of *P*-values.

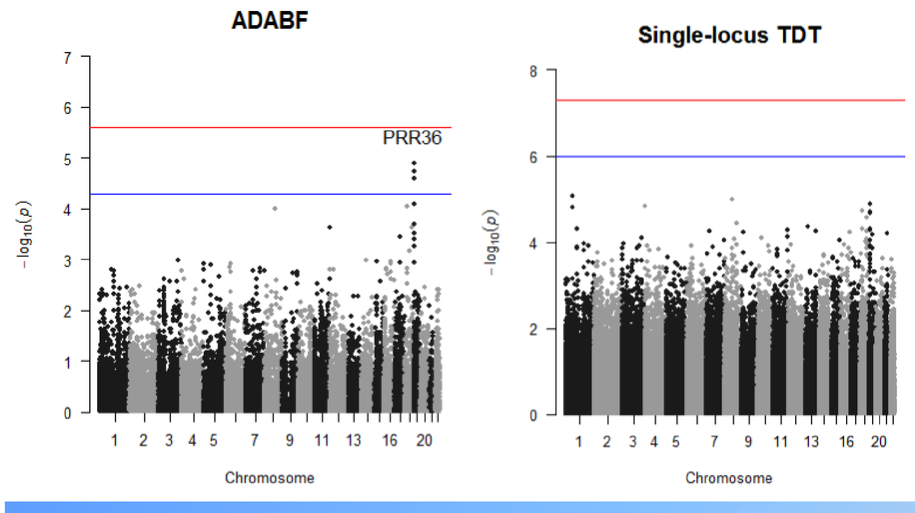
- How to obtain the Bayes factor (BF)?
- Usually we need to adjust for some covariates. For example,

$$E(Y) = \beta_0 + \beta G_l + \beta_A Age + \beta_S Smoking$$

- According to the asymptotic normality of MLE: $\hat{\beta} \sim N(\beta, V)$
- The prior distribution of the true effect sizes: $\beta \sim N(0, W)$
- We follow the WTCCC GWAS to specify the prior variance *W*=0.04 [2]

- How to assess the significance of a gene/region?
- The highest *k* BFs in favor of H_1 are combined, in the observed sample and in each of the resamples, respectively.
- The optimal *k* that achieves the strongest signal is allowed to vary in the observed sample and in each of the resamples.
- Then, the significance of the gene/region is assessed by comparing the strongest signal in the observed sample with its counterparts in the resampling replicates.

$$BF = \sqrt{\frac{\hat{V}}{\hat{V}+W}} \exp\left(\frac{\hat{\beta}^2 W}{2\hat{V}(\hat{V}+W)}\right) \quad [3]$$



- ✓ Three genes on chromosome 19p13.2, including *EVI5L* (ecotropic viral integration site 5 like), *PRR36* (proline rich 36), and *LYPLA2P2* (lysophospholipase II pseudogene 2), were detected to be associated with schizophrenia at the suggestive significance level of $5*10^{-5}$.
- ✓ Chromosome 19p13.2 has been found to be associated with panic disorder [4].

Conclusions:

[A] Compared with ADA [1], our ADABF method is recommended for its applicability to variants with a wide spectrum of MAFs.

[B] Compared with other multi-marker association tests, our ADABF method is recommended for its robustness to the inclusion of neutral variants.

References:

[1] Lin, W. Y., Lou, X. Y., Gao, G. & Liu, N. (2014). *Plos One* **9**, e85728.
[2] WTCCC. (2007). *Nature* **447**, 661-678.
[3] Wakefield, J. (2009). *Genetic epidemiology* **33**, 79-86.
[4] Gregersen, N. O. *et al.* (2016). *Psychiat Genet* **26**, 287-292.