Estimation of the Null Distribution of Fisher's Exact Test P-values in GWA Studies.

Yukinori Okada[1,2], Kazuhiko Yamamoto[2], and Ryo Yamada[1,3].

[1] Functional Genomics, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

[2] Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[3] Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.

Abstract

The assessment of variable inflation of test statistics is an essential process in genome-wide association studies. Variable inflation can be quantified by comparing the observed distribution of test statistics with their expected distribution under the null hypothesis without the inflation. Here, we planned to estimate the null distribution of Fisher's exact test p-values, that is not given arithmetically.

For each table in the table set, we calculated the cumulative distribution function (c.d.f.) of p-values that the marginal counts of the table could give, and averaged the c.d.f.s of all the tables in the table set, which was our estimated null distribution of exact test p-values for the table set.

We generated simulation case-control genotype data in a range of variable inflation, using Wright's F_{ST}. For each genotype data, we calculated Fisher's exact test p-values of the allelic 2 × 2 tables of the SNPs, and estimated the null distribution of Fisher's exact test p-values.

When no variable inflation was induced, observed p-values followed their estimated null distributions. Observed Fisher's exact test p-values indicated higher variable inflation, as F_{ST} increased. Estimated null distributions of Fisher's exact test p-values were identical regardless of F_{ST}.

Our method estimated the null distribution of Fisher's exact test p-values without variable inflation, from the sets of the contingency tables that included the variable inflation.

Methods

Estimation of the null distribution of Fisher's exact test p-values for 2 × 2 contingency tables.

Consider a 2 × 2 contingency table as below.

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

When there is no bias in the distribution of cell counts under the given fixed marginal values, the distribution of the cell counts follow the hypergeometric distribution. The occurrence probability of the observed table is expressed as

$$\Pr(n, n_{1.}, n_{.1}, n_{11}) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!},$$

and the two-tailed Fisher's exact test p-value of the tables is expressed as

$$P(n, n_{1.}, n_{.1}, n_{11}) = \sum_{\Pr(n,n_{1.},n_{.1},n'_{11}) \le \Pr(n,n_{1.},n_{.1},n_{11})} \Pr(n, n_{1.}, n_{.1}, n'_{11}),$$

where n'_{11} satisfies all of $n'_{ij} \geq 0$.

Let $Q(x)$ be the c.d.f. of Fisher's exact test p-values of the tables with marginal values of (n, $n_{1.}$, $n_{.1}$).

$$Q(x \mid n, n_{1.}, n_{.1}) = \sum_{P(n,n_{1.},n_{.1},n'_{11}) \le x} \Pr(n, n_{1.}, n_{.1}, n'_{11}).$$

Suppose sample space Ω which consists of k 2 × 2 contingency tables. We define $F(x|\ \Omega)$, the summation of $Q(x)$ over Ω, as the estimator of the c.d.f. of Fisher's exact test p-values under the null hypothesis without variable inflation.

$$F(x \mid \Omega) = \frac{1}{k} \sum_{(n,n_{1.},n_{.1}) \in \Omega} Q(x \mid n, n_{1.}, n_{.1}).$$

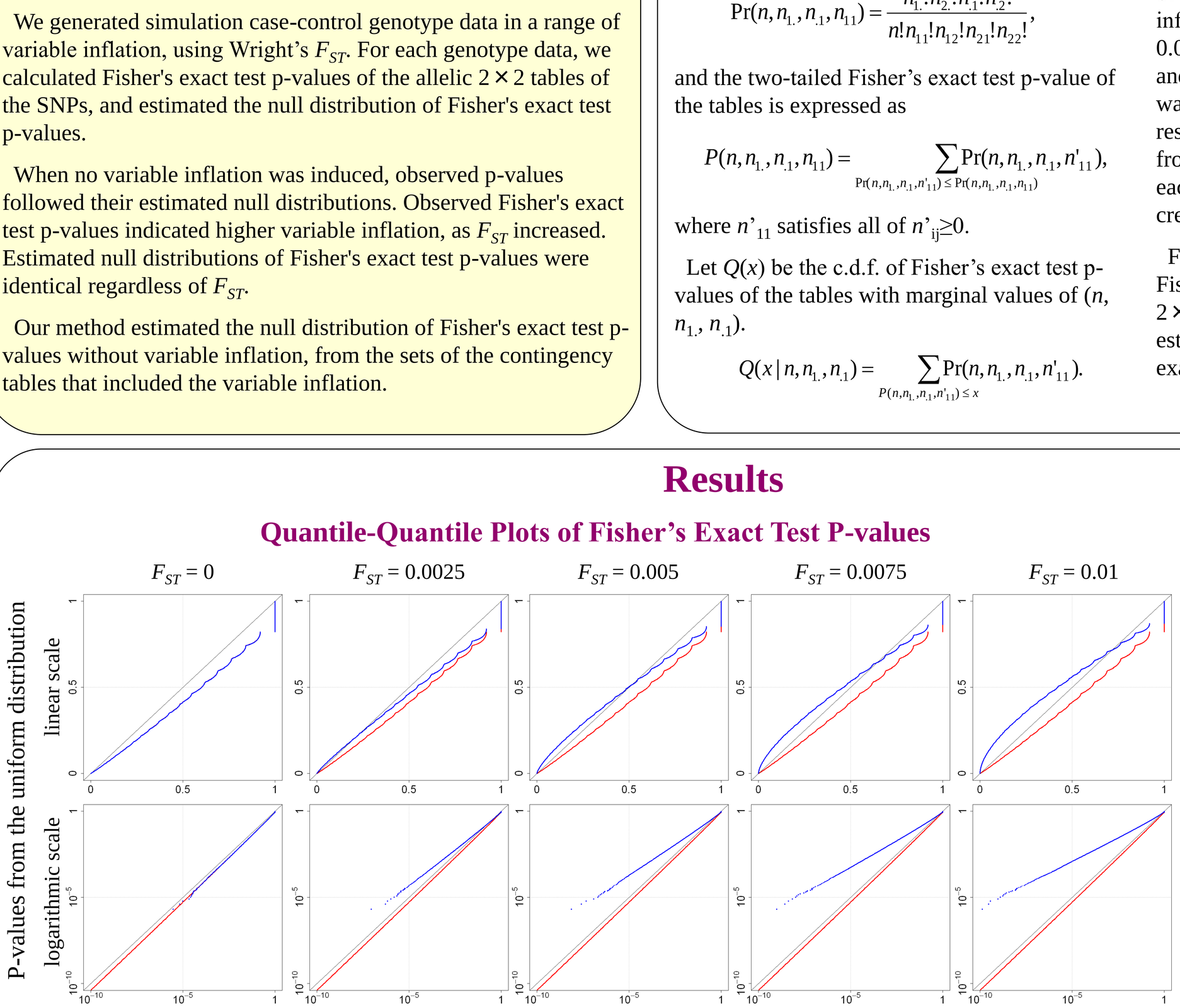
Simulation analysis

We generated simulation case-control SNP genotype data in a range of variable inflation (Wright's F_{ST} = 0, 0.0025, 0.005, 0.0075, 0.01). Sample size was 100 cases and 100 controls, and number of SNPs was 500,000. Allele frequencies of the respective SNPs were randomly obtained from the uniform distribution of (0,1). For each value of F_{ST}, genotype data were created 10 times separately.

For each genotype data, we calculated Fisher's exact test p-values of the allelic 2 × 2 tables of the SNPs, and then estimated the null distribution of Fisher's exact test p-values.

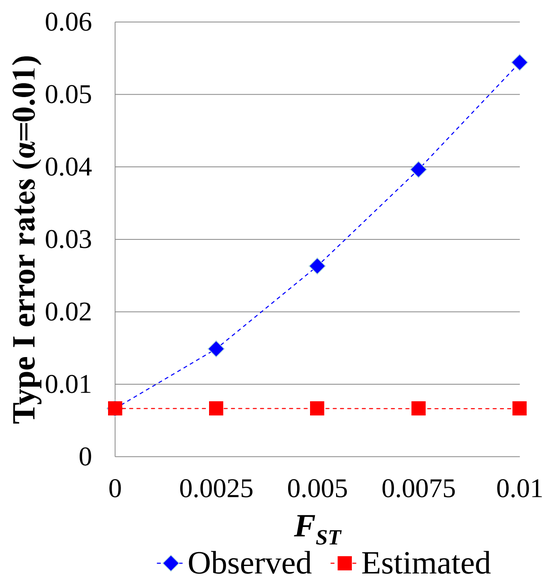
Results

Quantile-Quantile Plots of Fisher's Exact Test P-values



Observed p-values or estimated null distributions of p-values (observed: ——— estimated: ———)

Type I Error Rates of Fisher's Exact Test P-values



(i) When no variable inflation was induced (F_{ST}=0), observed Fisher's exact test p-values followed their estimated null distributions.

(ii) Observed Fisher's exact test p-values indicated higher variable inflation, as the value of F_{ST} increased from 0 to 0.01.

(iii) Estimated null distributions of Fisher's exact test p-values were identical regardless of the value of F_{ST}.

Conclusions

Our method estimated the null distribution of Fisher's exact test p-values without variable inflation, regardless of the degree of the included variable inflation.

Acknowledgements

We thank all the members of the Lab of Functional Genomics, the Lab for Statistical Analysis, CGM, RIKEN, and the Lab for Autoimmune Diseases, CGM, RIKEN.
Any questions or comments are welcome at yokada-tky@umin.ac.jp