

# Identification of novel candidate markers of type 2 diabetes and obesity in Russia by exome sequencing with a limited sample size

Yury A. Barbitoff<sup>1,2,3,4</sup>, Elena A. Serebryakova<sup>1,5,6</sup>, Yulia A. Nasykhova<sup>1,5</sup>, Alexander V. Predeus<sup>2</sup>; Dmitrii E. Polev<sup>1</sup>, Anna R. Shuvalova<sup>1</sup>, Evgenii V. Vasiliev<sup>6</sup>, Stanislav P. Urazov<sup>6</sup>, Andrey M. Sarana<sup>4,6</sup>, Sergey G. Scherbak<sup>4,6</sup>, Maria S. Pokrovskaya<sup>7</sup>, Oksana V. Sivakova<sup>7</sup>, Aleksey N. Meshkov<sup>7</sup>, Oxana M. Drapkina<sup>7</sup>, Oleg S. Glotov<sup>5,6</sup>, Andrey S. Glotov<sup>1,5\*</sup>

<sup>1</sup> Biobank of the Research Park, Saint Petersburg State University, Russia, 199034, St. Petersburg, Universitetskaya nab., 7-9-11

<sup>2</sup> Bioinformatics Institute, Saint Petersburg, Russia, 194100, Kantemirovskaya st., 2A

<sup>3</sup> Department of Genetics and Biotechnology, Saint Petersburg State University, Russia, 199034, St. Petersburg, Universitetskaya nab., 7-9-11

<sup>4</sup> Institute of Translation Biomedicine, Saint Petersburg State University, Russia, 199034, St. Petersburg, Universitetskaya nab., 7-9-11

<sup>5</sup> Laboratory of prenatal diagnostics of hereditary diseases, FSBSI «The Research Institute of Obstetrics, Gynaecology and Reproductology named after D.O. Ott», Russia, 199034, St. Petersburg, Mendeleyevskaya line, 3

<sup>6</sup> City hospital №40, Russia, 197706, St. Petersburg, Sestroretsk, Borisov Str., 9

<sup>7</sup> Federal State Institution «National Medical Research Center for Preventive Medicine» of the Ministry of Healthcare of the Russian Federation, 101990, Moscow, Petroverigskiy lane, bld. 10

For correspondence:

Corresponding author: Director of Biobank of the Research Park, Saint Petersburg State University, Russia, 199034, St. Petersburg, Universitetskaya nab., 7-9-11,

Tel.: + 7-911-783-20-03

E-mail address: anglotov@mail.ru (Andrey Glotov)

## Supporting Information

## Supplementary note

### Calculation of allelic frequency in the cases and controls

To calculate the minor allele frequency in case and control groups given known true population minor allele frequency (tMAF) and true odds ratio (tOR), we first reconstructed the contingency table to obtain proportions of case and control individuals with reference or non-reference genotype:

	Reference genotype	Non-reference genotype	
<b>Control</b>	$p_{11}$	$p_{10}$	$p_1$
<b>Diseased</b>	$p_{01}$	$p_{00}$	$p_0$
	$p_1$	$p_0$	

Under the dominant inheritance model,  $p_1$  is the frequency of reference-homozygous genotype, i.e.  $(1 - tMAF)^2$  under Hardy-Weinberg equilibrium. On the other hand,  $p_1$  is equal to  $1 - P_D$ , where  $P_D$  is the disease prevalence (0.08 in all simulations presented in the manuscript).

Given these estimates and tOR, we can use functional properties of the odds ratio to obtain cell probabilities:

$$p_{11} = \frac{1 + ((1 - P_D) + (1 - tMAF)^2)(tOR - 1) - S}{2(tOR - 1)}$$

$$S = \sqrt{(1 + ((1 - P_D) + (1 - tMAF)^2)(tOR - 1))^2 + 4tOR(1 - tOR)(1 - P_D)(1 - tMAF)^2}$$

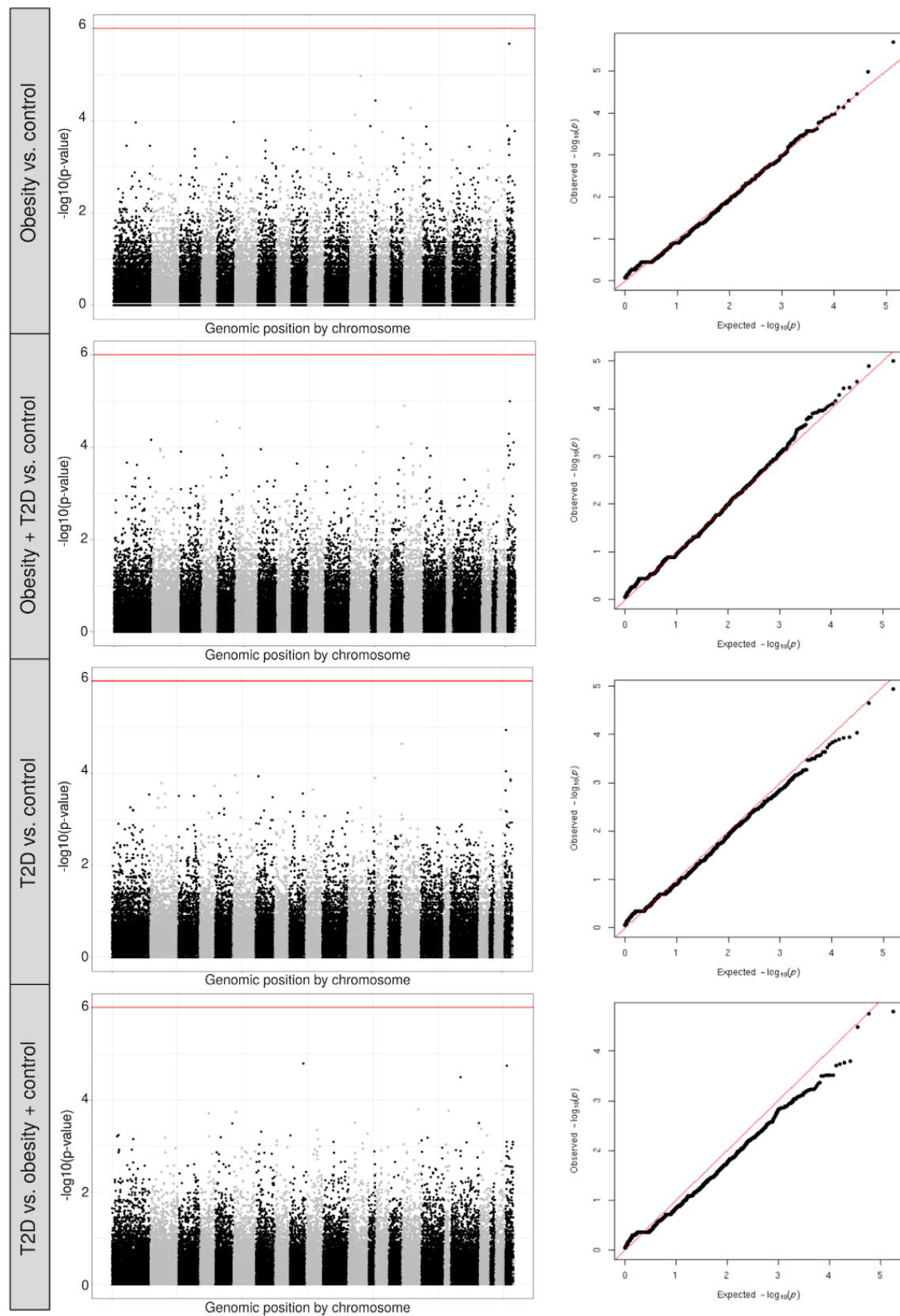
Then, the cell probabilities can be calculated as follows:

$$p_{01} = 1 - P_D - p_{11}; p_{10} = (1 - tMAF)^2 - p_{11}; p_{00} = 1 - p_{11} - p_{10} - p_{01}$$

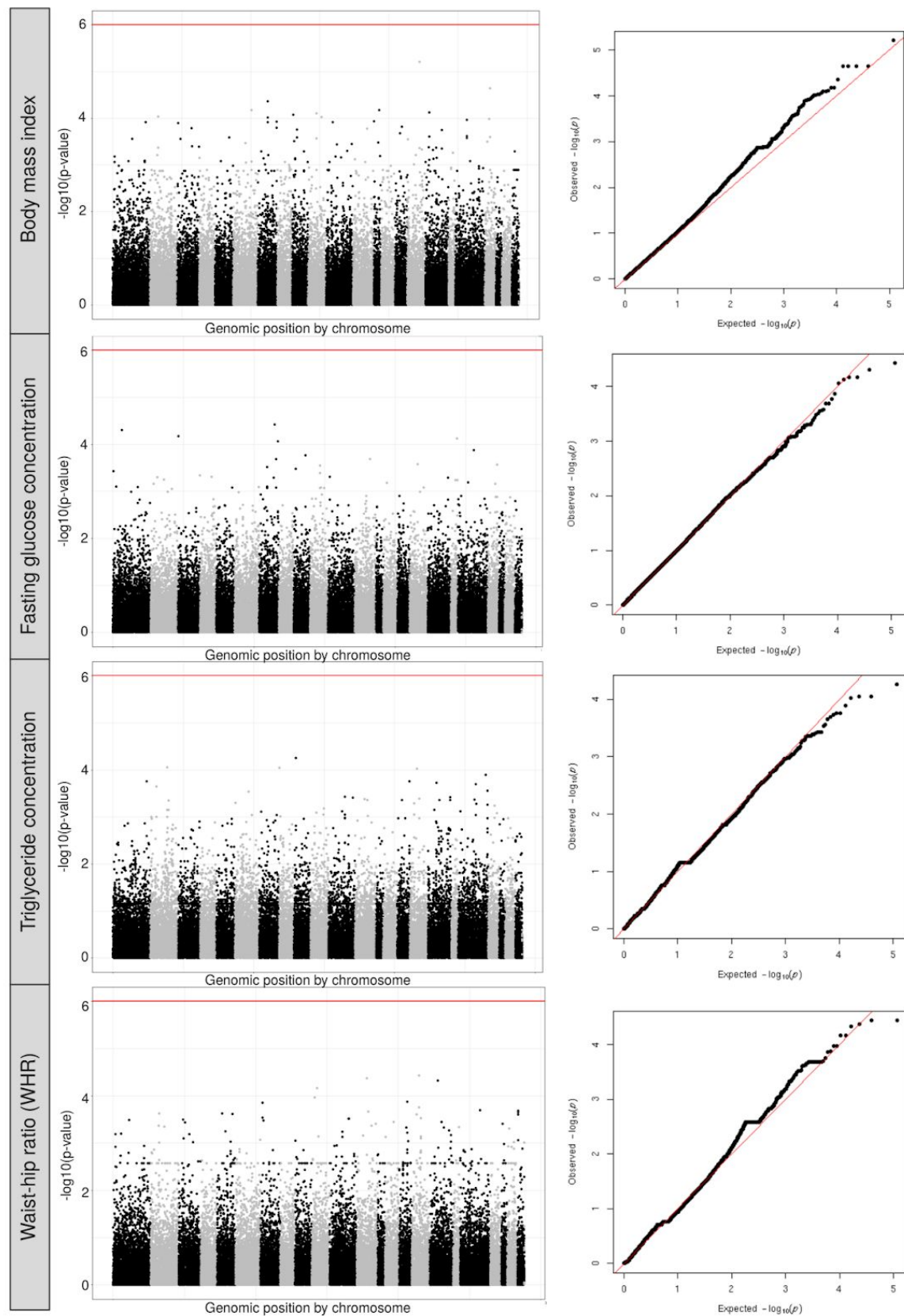
Then, non-reference allele frequency can be obtained by subtracting the square root of reference genotype frequency in each subpopulation:

$$MAF_{controls} = 1 - \sqrt{\frac{p_{11}}{p_{11} + p_{10}}}$$

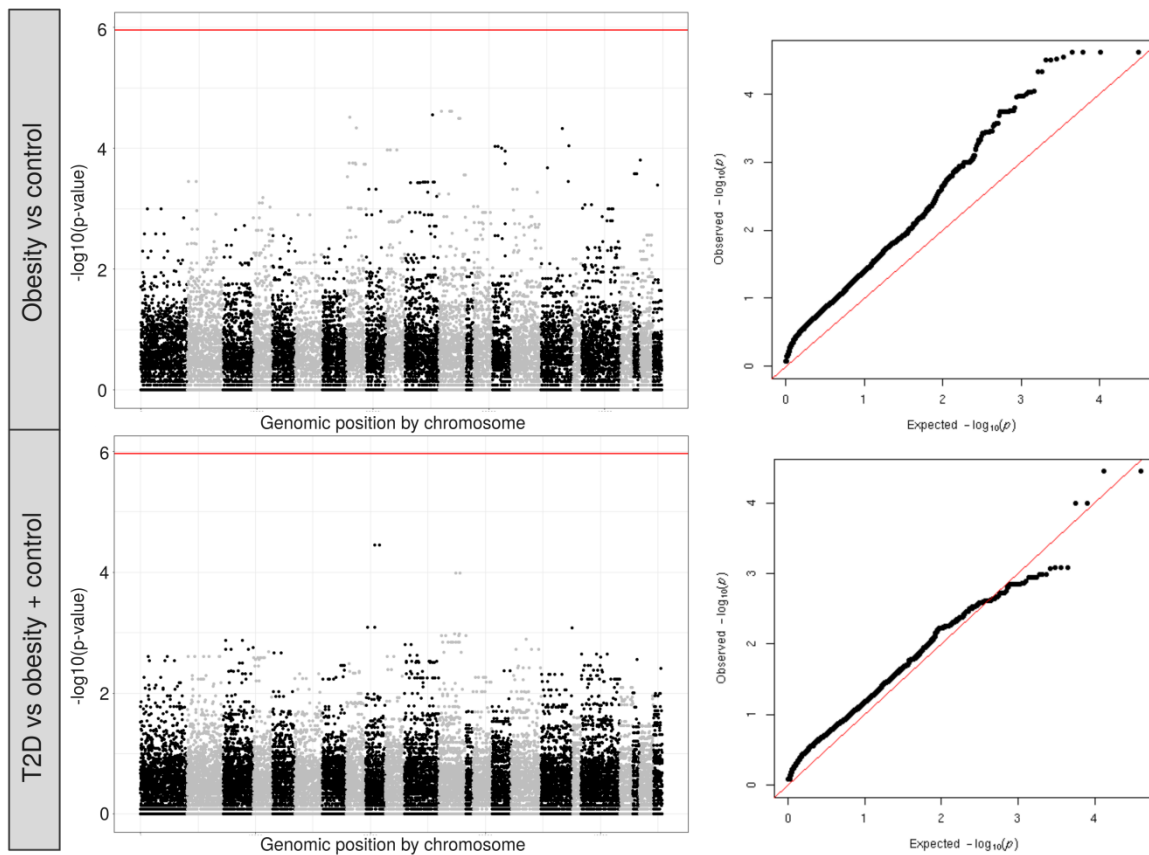
$$MAF_{diseased} = 1 - \sqrt{\frac{p_{01}}{p_{01} + p_{00}}}$$



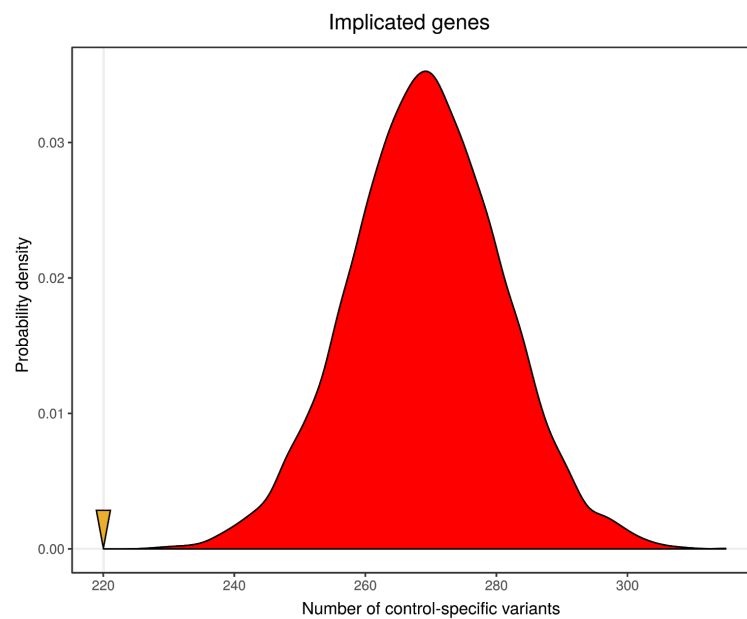
**Figure S1.** Statistics of SNP-level association in 4 comparisons in binary trait analysis (indicated on the left). Left panels, Manhattan plots of association p-values; right panels, quantile-quantile (Q-Q) plots of p-value distribution.



**Figure S2.** Statistics of SNP-level association with 4 selected quantitative traits (indicated on the left). Left panels, Manhattan plots of association p-values; right panels, quantile-quantile (Q-Q) plots of p-value distribution. Note that inflation of p-value distribution is observed only for body mass index (BMI).



**Figure S3.** Statistics of locus-level association with binary traits in indicted comparisons. Left panels, Manhattan plots of association p-values; right panels, quantile-quantile (Q-Q) plots of p-value distribution.



**Figure S4.** Distributions of the random expectation numbers of control-specific protein-altering variants inside implicated genes. Yellow arrowhead indicates observed values.