# Supplementary figures of "a large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes"

*Xiang Zhu*[1,2] *and Matthew Stephens*[2,3]

*13 July 2018*

1. Department of Statistics, Stanford University, Stanford, CA, USA
2. Department of Statistics, University of Chicago, Chicago, IL, USA
3. Department of Human Genetics, University of Chicago, Chicago, IL, USA

This document contains all supplementary figures of the manuscript entitled "A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes" (`https://doi.org/10.1101/160770`). Correspondence should be addressed to X.Z. (`xiangzhu@stanford.edu`) and M.S. (`mstephens@uchicago.edu`).

## Contents

# Supplementary Figure 1

**Simulation details and additional results of Figure 2(a).** Here we use real genotypes of 12,758 SNPs on chromosome 16 from 1458 individuals in the UK Blood Service Control Group (Wellcome Trust Case Control Consortium 2007) to simulate phenotype data, and then compute single-SNP association summary statistics. On these summary data, we compare RSS with existing enrichment methods.

We use *Signal Transduction Pathway* (Biosystem, Reactome) to create SNP-level annotation for these 12,758 SNPs. Specifically, we let $a_j = 1$ if SNP $j$ is within ±100 kb of the transcribed region of any *Signal Transduction Pathway* gene, and let $a_j = 0$ otherwise. There are 36 *Signal Transduction Pathway* genes on chromosome 16. There are 676 SNPs assigned to this gene set with annotation $a_j = 1$.

We simulate baseline and enrichment datasets in a paired way. We first simulate the causal indicator $\gamma_j$ of each SNP $j$ for an enrichment dataset as follows:

$$\gamma_j \sim \text{Bernoulli}(\pi_j), \text{ where } \pi_j = (1 + 10^{-(\theta_0 + a_j\theta)})^{-1},$$

where $\theta_0$ is the background parameter and $\theta$ is the enrichment parameter. We count the number of causal SNPs in this enrichment dataset as $n_c := \sum_j \gamma_j$, and then randomly choose $n_c$ SNPs from chromosome 16 as causal SNPs for the baseline dataset.

Given the causal indicators $\{\gamma_j\}$, we simulate the genetic effect $\beta_j$ of each SNP $j$ as follows:

$$\beta_j|\gamma_j = 0 \sim \delta_0 \text{ and } \beta_j|\gamma_j = 1 \sim \text{Normal}(0, 1),$$

where $\delta_0$ denotes point mass at zero. Next, we simulate the phenotype $y_i$ of Individual $i$ as follows:

$$y_i = \sum_j x_{ij}\beta_j + \epsilon_i,\ \epsilon_i \sim \text{Normal}(0, \tau^{-1}),$$

where $x_{ij}$ is the genotype of SNP $j$ for Individual $i$, $j = 1, \ldots, 12{,}758$ and $i = 1, \ldots, 1458$. The true value of residual variance $\tau^{-1}$ is determined by the true value of PVE (total proportion of variance in phenotype $y$ explained by effects of all available SNPs $X$) as follows:

$$\text{PVE} = V(X\beta)/(\tau^{-1} + V(X\beta)),$$

where $V(X\beta)$ is the sample variance of $X\beta$. We adopt this phenotype simulation scheme from previous work, notably Guan and Stephens (2011), Zhou, Carbonetto, and Stephens (2013) and Zhu and Stephens (2017).

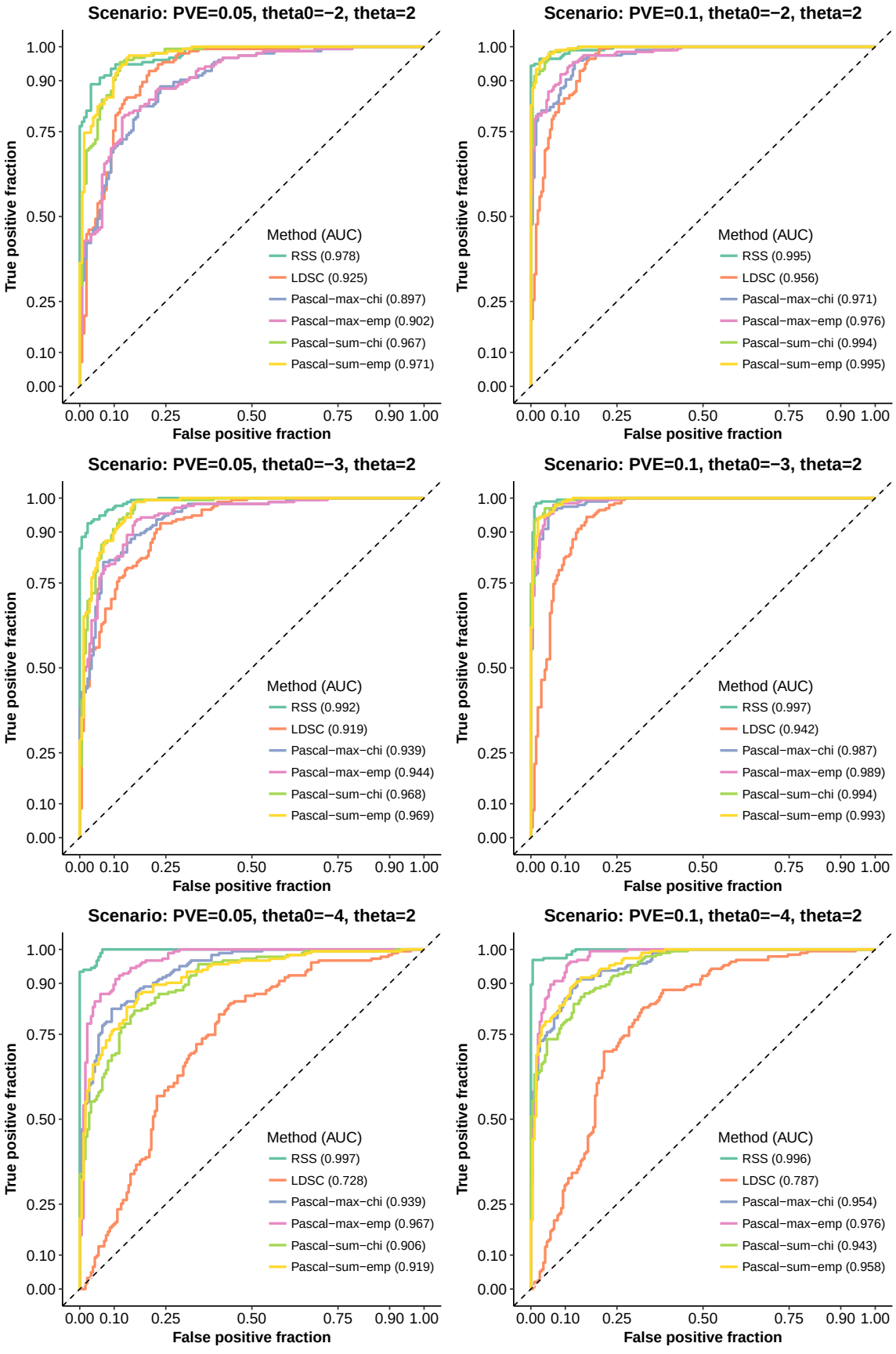
In this set of simulations, the true values of background parameter $\theta_0$ are $\{-2, -3, -4\}$, the true value of enrichment parameter $\theta$ is 2, and the true values of PVE are $\{0.05, 0.1, 0.2, 0.6\}$. For each combination of $(\theta_0, \theta, \text{PVE})$, we simulate 200 baseline and 200 enrichment independent datasets. The "sparse scenario" in **Figure 2(a)** corresponds to simulations with true $\theta_0 = -4$ and PVE = 0.05. The "polygenic scenario" in **Figure 2(a)** corresponds to simulations with true $\theta_0 = -2$ and PVE = 0.05.

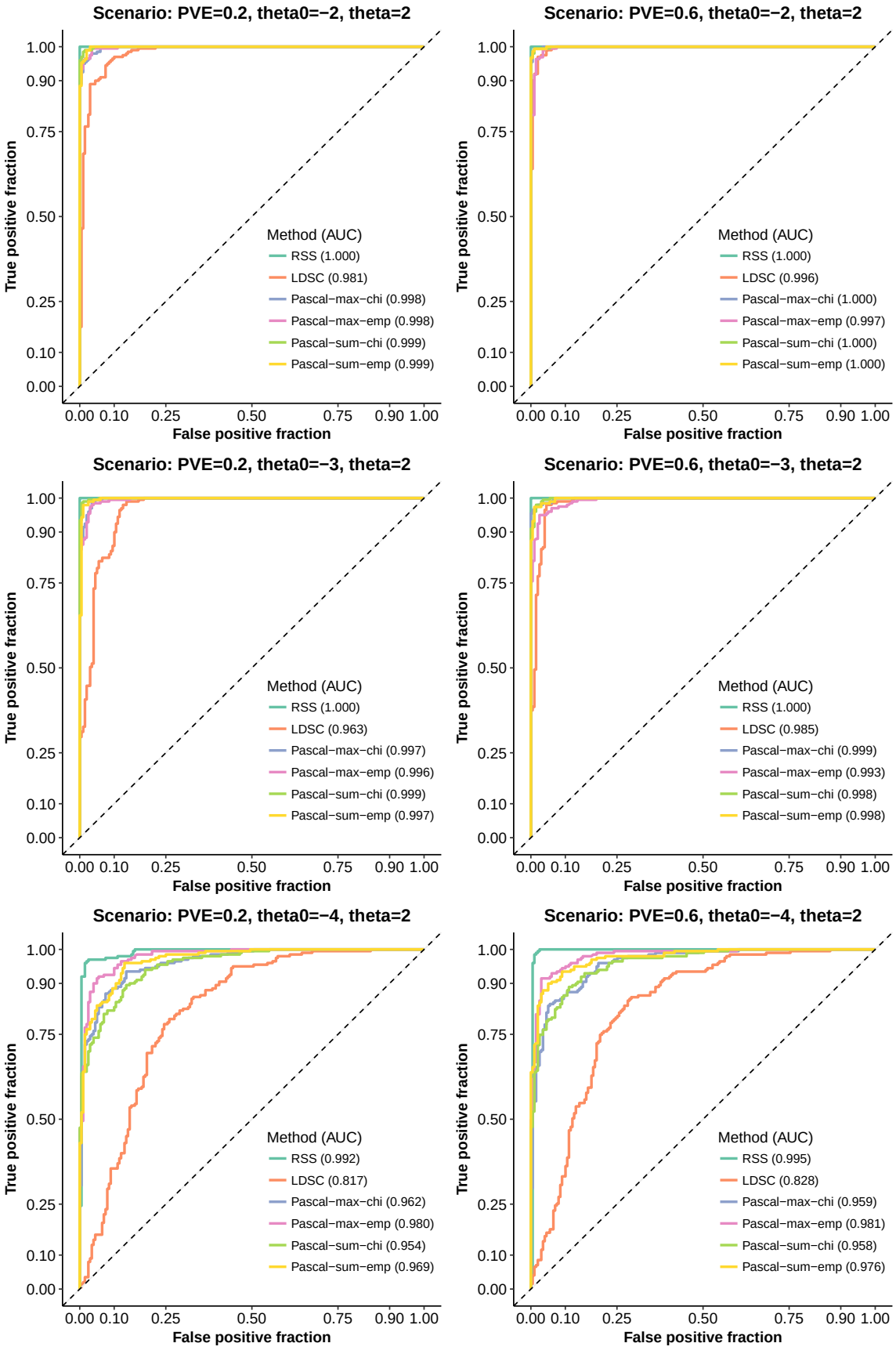
We apply RSS to the simulated datasets, using the program `rss_varbvsr_squarem.m` available at `https://github.com/stephenslab/rss`. The input LD matrix is estimated from the 1480 control individuals in the WTCCC 1958 British Birth Cohort, using a shrinkage estimator from Wen and Stephens (2010). The grid on the enrichment parameter $\theta$ is `(0:0.05:1)` for baseline datasets, and is `(1:0.05:3)` for enrichment datasets. The grid on the background parameter $\theta_0$ is `(true_theta0-0.5:0.05:true_theta0+0.5)`, where `true_theta0` is the true value of $\theta_0$ that was used to generate the dataset.

For each simulated dataset, we also perform enrichment analysis using two existing approaches with their default settings: LDSC (Finucane et al. 2015) and Pascal (Lamparter et al. 2016). Note that Pascal includes two gene scoring options: maximum-of-$\chi^2$ (`-max`) and sum-of-$\chi^2$ (`-sum`), and two pathway scoring options: $\chi^2$ approximation (`-chi`) and empirical sampling (`-emp`).

The candidate pathway for testing is *Signal Transduction Pathway* (Biosystem, Reactome) in both baseline and enrichment datasets. If a method identifies a baseline dataset as enriched, then it is a "false positive". If a method identifies an enrichment dataset as enriched, then it is a "true positive".

We evaluate the performance of these enrichment methods by plotting the receiver operating characteristic (ROC) curve and computing the area under the curve (AUC) for each method. Both metrics are implemented in the package `plotROC` (Sachs 2017).


Scenario: PVE=0.05, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (0.978)
LDSC (0.925)
Pascal−max−chi (0.897)
Pascal−max−emp (0.902)
Pascal−sum−chi (0.967)
Pascal−sum−emp (0.971)
Scenario: PVE=0.1, theta0=−2, theta=2
Method (AUC)
RSS (0.995)
LDSC (0.956)
Pascal−max−chi (0.971)
Pascal−max−emp (0.976)
Pascal−sum−chi (0.994)
Pascal−sum−emp (0.995)
Scenario: PVE=0.05, theta0=−3, theta=2
Method (AUC)
RSS (0.992)
LDSC (0.919)
Pascal−max−chi (0.939)
Pascal−max−emp (0.944)
Pascal−sum−chi (0.968)
Pascal−sum−emp (0.969)
Scenario: PVE=0.1, theta0=−3, theta=2
Method (AUC)
RSS (0.997)
LDSC (0.942)
Pascal−max−chi (0.987)
Pascal−max−emp (0.989)
Pascal−sum−chi (0.994)
Pascal−sum−emp (0.993)
Scenario: PVE=0.05, theta0=−4, theta=2
Method (AUC)
RSS (0.997)
LDSC (0.728)
Pascal−max−chi (0.939)
Pascal−max−emp (0.967)
Pascal−sum−chi (0.906)
Pascal−sum−emp (0.919)
Scenario: PVE=0.1, theta0=−4, theta=2
Method (AUC)
RSS (0.996)
LDSC (0.787)
Pascal−max−chi (0.954)
Pascal−max−emp (0.976)
Pascal−sum−chi (0.943)
Pascal−sum−emp (0.958)


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (1.000)
LDSC (0.981)
Pascal−max−chi (0.998)
Pascal−max−emp (0.998)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.999)
Scenario: PVE=0.6, theta0=−2, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.996)
Pascal−max−chi (1.000)
Pascal−max−emp (0.997)
Pascal−sum−chi (1.000)
Pascal−sum−emp (1.000)
Scenario: PVE=0.2, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.963)
Pascal−max−chi (0.997)
Pascal−max−emp (0.996)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.997)
Scenario: PVE=0.6, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.985)
Pascal−max−chi (0.999)
Pascal−max−emp (0.993)
Pascal−sum−chi (0.998)
Pascal−sum−emp (0.998)
Scenario: PVE=0.2, theta0=−4, theta=2
Method (AUC)
RSS (0.992)
LDSC (0.817)
Pascal−max−chi (0.962)
Pascal−max−emp (0.980)
Pascal−sum−chi (0.954)
Pascal−sum−emp (0.969)
Scenario: PVE=0.6, theta0=−4, theta=2
Method (AUC)
RSS (0.995)
LDSC (0.828)
Pascal−max−chi (0.959)
Pascal−max−emp (0.981)
Pascal−sum−chi (0.958)
Pascal−sum−emp (0.976)

# Supplementary Figure 2

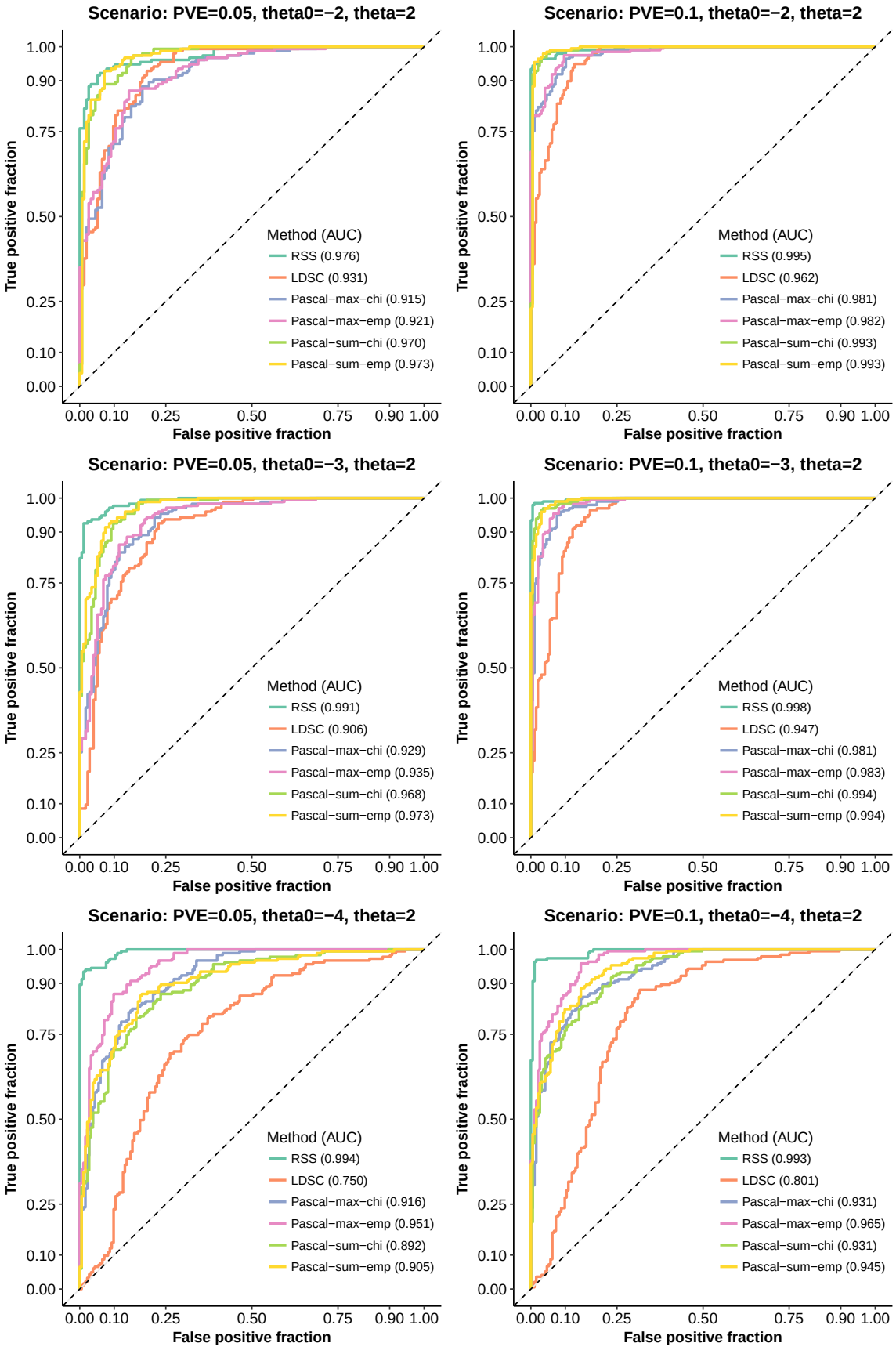
**Simulation details and additional results of Figure 2(b).** Details of this set of simulations are almost identical to those in **Supplementary Figure 1**. Here we only highlight the differences.

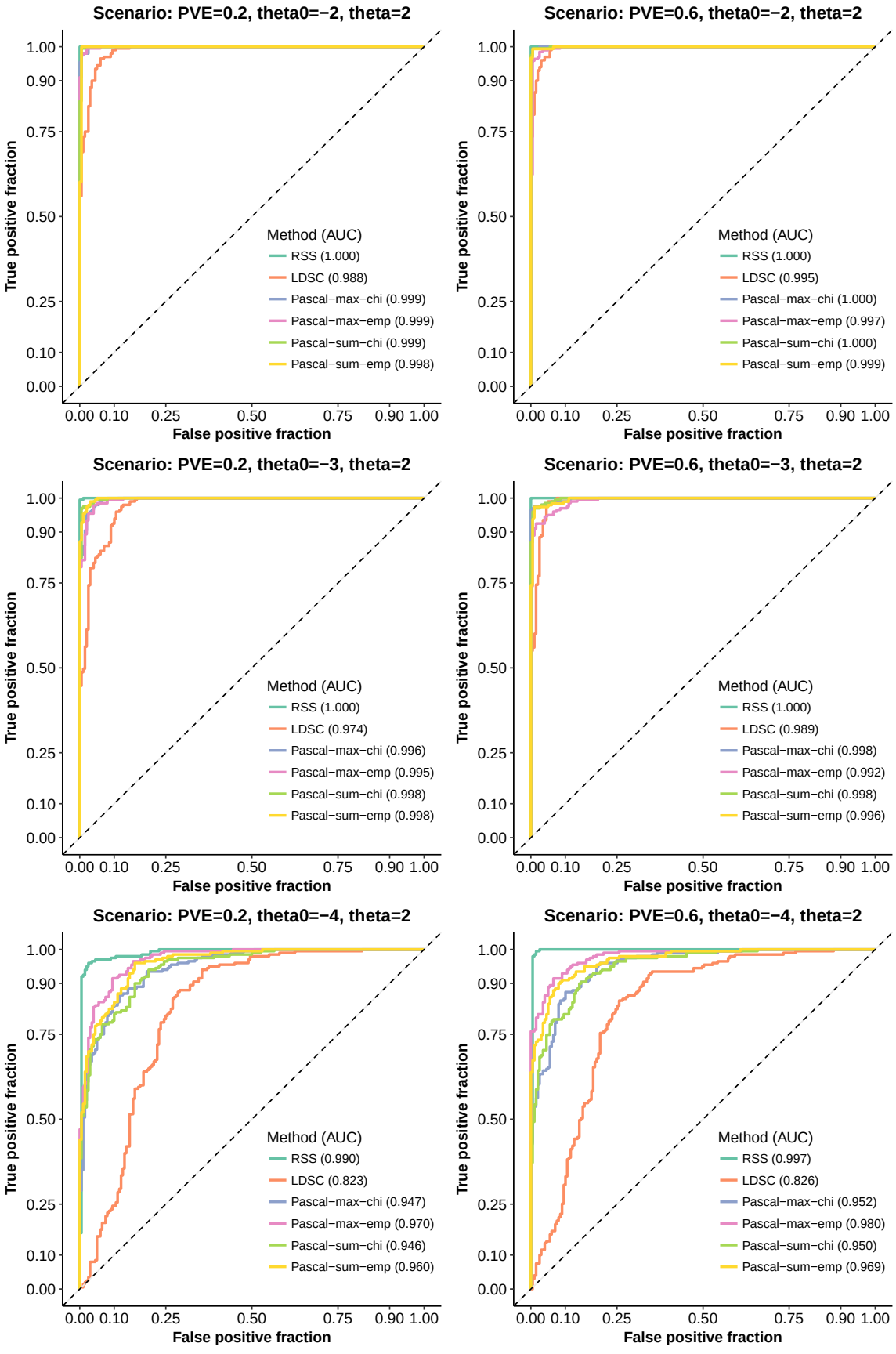
This set of simulations aims to assess the robustness of RSS to model misspecification where a random set of "near-gene" SNPs is enriched for genetic association in baseline datasets.

We define a SNP as "near-gene" if this SNP is within ±100 kb of the transcribed region of any gene. In total, there are 878 genes and 9356 near-gene SNPs on chromosome 16.

We first simulate an enrichment dataset as in **Supplementary Figure 1**. For this enrichment dataset, we count the total number of causal SNPs as $n_c = \sum_j \gamma_j$, and count the number of causal SNPs in the target pathway as $n_p = \sum_j \gamma_j a_j$. We randomly choose $n_p$ SNPs from the 9356 near-gene SNPs and $n_c - n_p$ SNPs from the remaining 3402 SNPs, and use them as causal SNPs to create a baseline dataset.

The "sparse scenario" in **Figure 2(b)** corresponds to simulations with true $\theta_0 = -4$ and PVE = 0.05. The "polygenic scenario" in **Figure 2(b)** corresponds to simulations with true $\theta_0 = -2$ and PVE = 0.05.


Scenario: PVE=0.05, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (0.976)
LDSC (0.931)
Pascal−max−chi (0.915)
Pascal−max−emp (0.921)
Pascal−sum−chi (0.970)
Pascal−sum−emp (0.973)
Scenario: PVE=0.1, theta0=−2, theta=2
Method (AUC)
RSS (0.995)
LDSC (0.962)
Pascal−max−chi (0.981)
Pascal−max−emp (0.982)
Pascal−sum−chi (0.993)
Pascal−sum−emp (0.993)
Scenario: PVE=0.05, theta0=−3, theta=2
Method (AUC)
RSS (0.991)
LDSC (0.906)
Pascal−max−chi (0.929)
Pascal−max−emp (0.935)
Pascal−sum−chi (0.968)
Pascal−sum−emp (0.973)
Scenario: PVE=0.1, theta0=−3, theta=2
Method (AUC)
RSS (0.998)
LDSC (0.947)
Pascal−max−chi (0.981)
Pascal−max−emp (0.983)
Pascal−sum−chi (0.994)
Pascal−sum−emp (0.994)
Scenario: PVE=0.05, theta0=−4, theta=2
Method (AUC)
RSS (0.994)
LDSC (0.750)
Pascal−max−chi (0.916)
Pascal−max−emp (0.951)
Pascal−sum−chi (0.892)
Pascal−sum−emp (0.905)
Scenario: PVE=0.1, theta0=−4, theta=2
Method (AUC)
RSS (0.993)
LDSC (0.801)
Pascal−max−chi (0.931)
Pascal−max−emp (0.965)
Pascal−sum−chi (0.931)
Pascal−sum−emp (0.945)


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (1.000)
LDSC (0.988)
Pascal−max−chi (0.999)
Pascal−max−emp (0.999)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.998)
Scenario: PVE=0.6, theta0=−2, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.995)
Pascal−max−chi (1.000)
Pascal−max−emp (0.997)
Pascal−sum−chi (1.000)
Pascal−sum−emp (0.999)
Scenario: PVE=0.2, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.974)
Pascal−max−chi (0.996)
Pascal−max−emp (0.995)
Pascal−sum−chi (0.998)
Pascal−sum−emp (0.998)
Scenario: PVE=0.6, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.989)
Pascal−max−chi (0.998)
Pascal−max−emp (0.992)
Pascal−sum−chi (0.998)
Pascal−sum−emp (0.996)
Scenario: PVE=0.2, theta0=−4, theta=2
Method (AUC)
RSS (0.990)
LDSC (0.823)
Pascal−max−chi (0.947)
Pascal−max−emp (0.970)
Pascal−sum−chi (0.946)
Pascal−sum−emp (0.960)
Scenario: PVE=0.6, theta0=−4, theta=2
Method (AUC)
RSS (0.997)
LDSC (0.826)
Pascal−max−chi (0.952)
Pascal−max−emp (0.980)
Pascal−sum−chi (0.950)
Pascal−sum−emp (0.969)

# Supplementary Figure 3

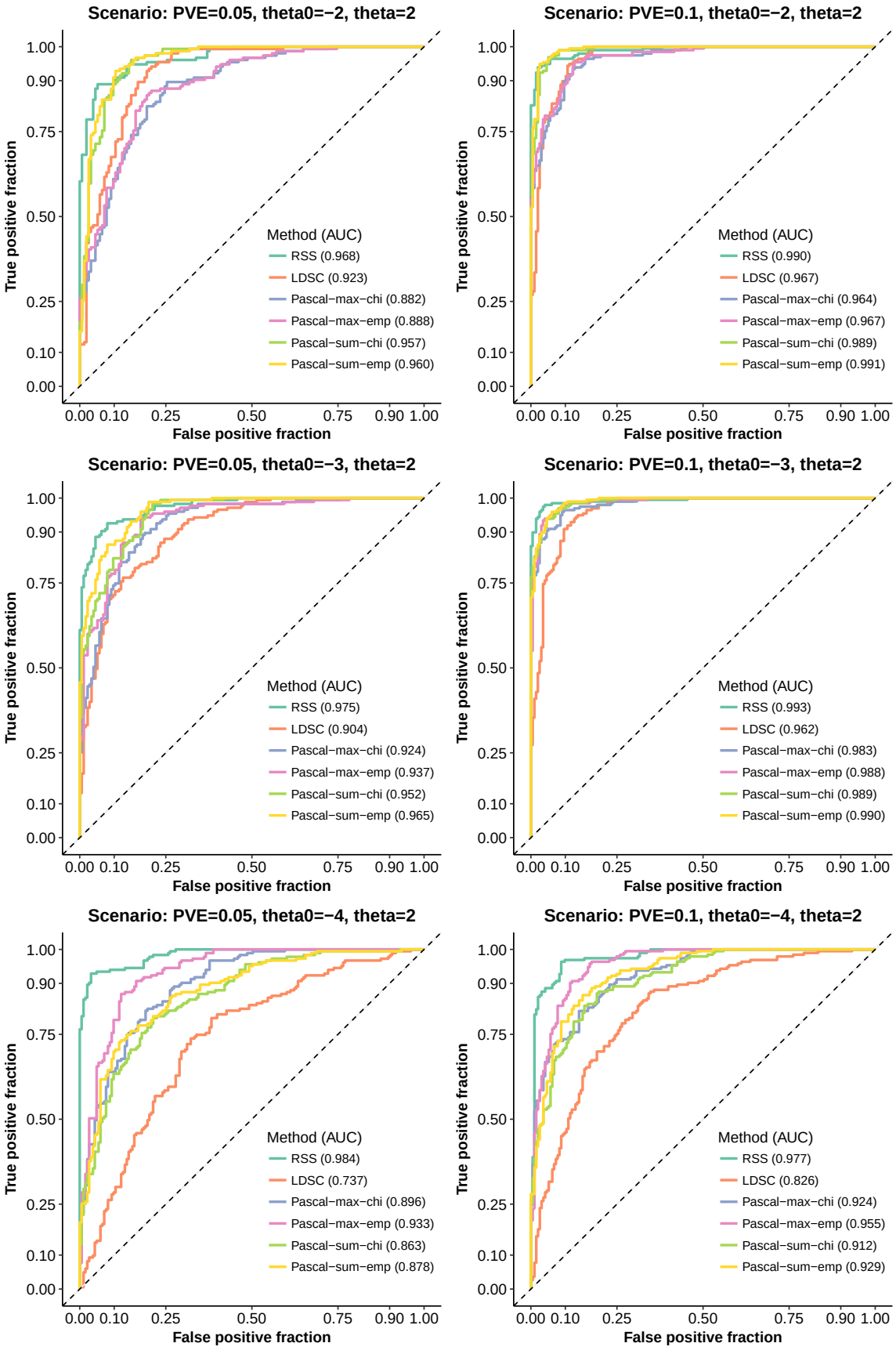
**Simulation details and additional results of Figure 2(c).** Details of this set of simulations are almost identical to those in **Supplementary Figure 1**. Here we only highlight the differences.

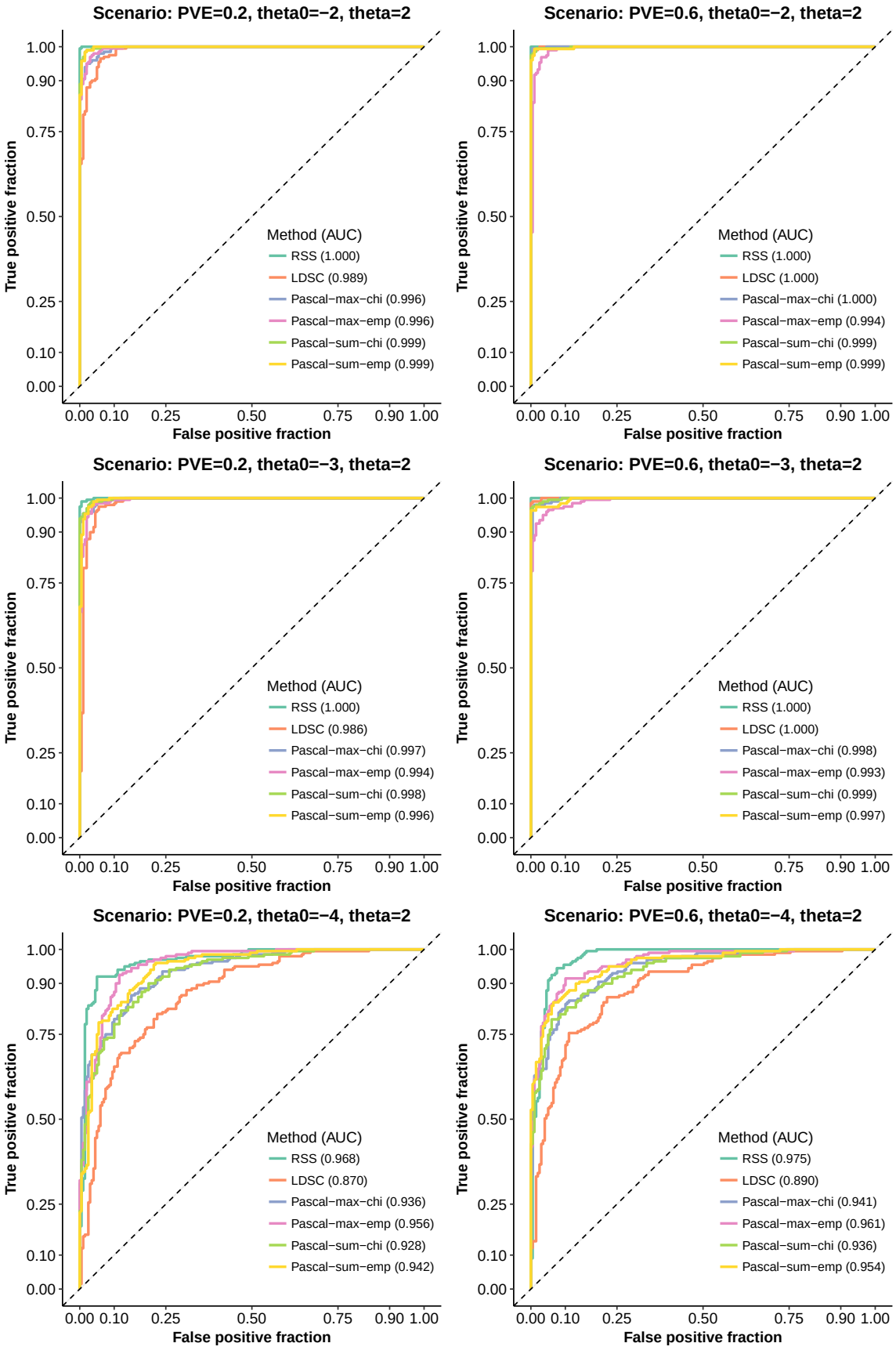
This set of simulations aims to assess the robustness of RSS to model misspecification where a random set of "coding" SNPs is enriched for genetic association in baseline datasets.

We define a SNP as "coding" if this SNP is under the Sequence Ontology term `coding_sequence_variant` or its children terms (`http://www.sequenceontology.org/miso/release_2.5/term/SO:0001580`). In total, there are 124 coding SNPs on chromosome 16.

We first simulate an enrichment dataset as in **Supplementary Figure 1**. For this enrichment dataset, we count the total number of causal SNPs as $n_c = \sum_j \gamma_j$, and count the number of causal SNPs in the target pathway as $n_p = \sum_j \gamma_j a_j$. We randomly choose $\min(n_p, 124)$ SNPs from the 124 coding SNPs and $n_c - \min(n_p, 124)$ SNPs from the 12,634 non-coding SNPs, and use them as causal SNPs to create a baseline dataset.

The "sparse scenario" in **Figure 2(c)** corresponds to simulations with true $\theta_0 = -4$ and PVE = 0.05. The "polygenic scenario" in **Figure 2(c)** corresponds to simulations with true $\theta_0 = -2$ and PVE = 0.05.


Scenario: PVE=0.05, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (0.968)
LDSC (0.923)
Pascal−max−chi (0.882)
Pascal−max−emp (0.888)
Pascal−sum−chi (0.957)
Pascal−sum−emp (0.960)
Scenario: PVE=0.1, theta0=−2, theta=2
Method (AUC)
RSS (0.990)
LDSC (0.967)
Pascal−max−chi (0.964)
Pascal−max−emp (0.967)
Pascal−sum−chi (0.989)
Pascal−sum−emp (0.991)
Scenario: PVE=0.05, theta0=−3, theta=2
Method (AUC)
RSS (0.975)
LDSC (0.904)
Pascal−max−chi (0.924)
Pascal−max−emp (0.937)
Pascal−sum−chi (0.952)
Pascal−sum−emp (0.965)
Scenario: PVE=0.1, theta0=−3, theta=2
Method (AUC)
RSS (0.993)
LDSC (0.962)
Pascal−max−chi (0.983)
Pascal−max−emp (0.988)
Pascal−sum−chi (0.989)
Pascal−sum−emp (0.990)
Scenario: PVE=0.05, theta0=−4, theta=2
Method (AUC)
RSS (0.984)
LDSC (0.737)
Pascal−max−chi (0.896)
Pascal−max−emp (0.933)
Pascal−sum−chi (0.863)
Pascal−sum−emp (0.878)
Scenario: PVE=0.1, theta0=−4, theta=2
Method (AUC)
RSS (0.977)
LDSC (0.826)
Pascal−max−chi (0.924)
Pascal−max−emp (0.955)
Pascal−sum−chi (0.912)
Pascal−sum−emp (0.929)


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (1.000)
LDSC (0.989)
Pascal−max−chi (0.996)
Pascal−max−emp (0.996)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.999)
Scenario: PVE=0.6, theta0=−2, theta=2
Method (AUC)
RSS (1.000)
LDSC (1.000)
Pascal−max−chi (1.000)
Pascal−max−emp (0.994)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.999)
Scenario: PVE=0.2, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.986)
Pascal−max−chi (0.997)
Pascal−max−emp (0.994)
Pascal−sum−chi (0.998)
Pascal−sum−emp (0.996)
Scenario: PVE=0.6, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (1.000)
Pascal−max−chi (0.998)
Pascal−max−emp (0.993)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.997)
Scenario: PVE=0.2, theta0=−4, theta=2
Method (AUC)
RSS (0.968)
LDSC (0.870)
Pascal−max−chi (0.936)
Pascal−max−emp (0.956)
Pascal−sum−chi (0.928)
Pascal−sum−emp (0.942)
Scenario: PVE=0.6, theta0=−4, theta=2
Method (AUC)
RSS (0.975)
LDSC (0.890)
Pascal−max−chi (0.941)
Pascal−max−emp (0.961)
Pascal−sum−chi (0.936)
Pascal−sum−emp (0.954)

# Supplementary Figure 4

**Simulation details and additional results of Figure 2(d).** Details of this set of simulations are almost identical to those in **Supplementary Figure 1**. Here we only highlight the differences.

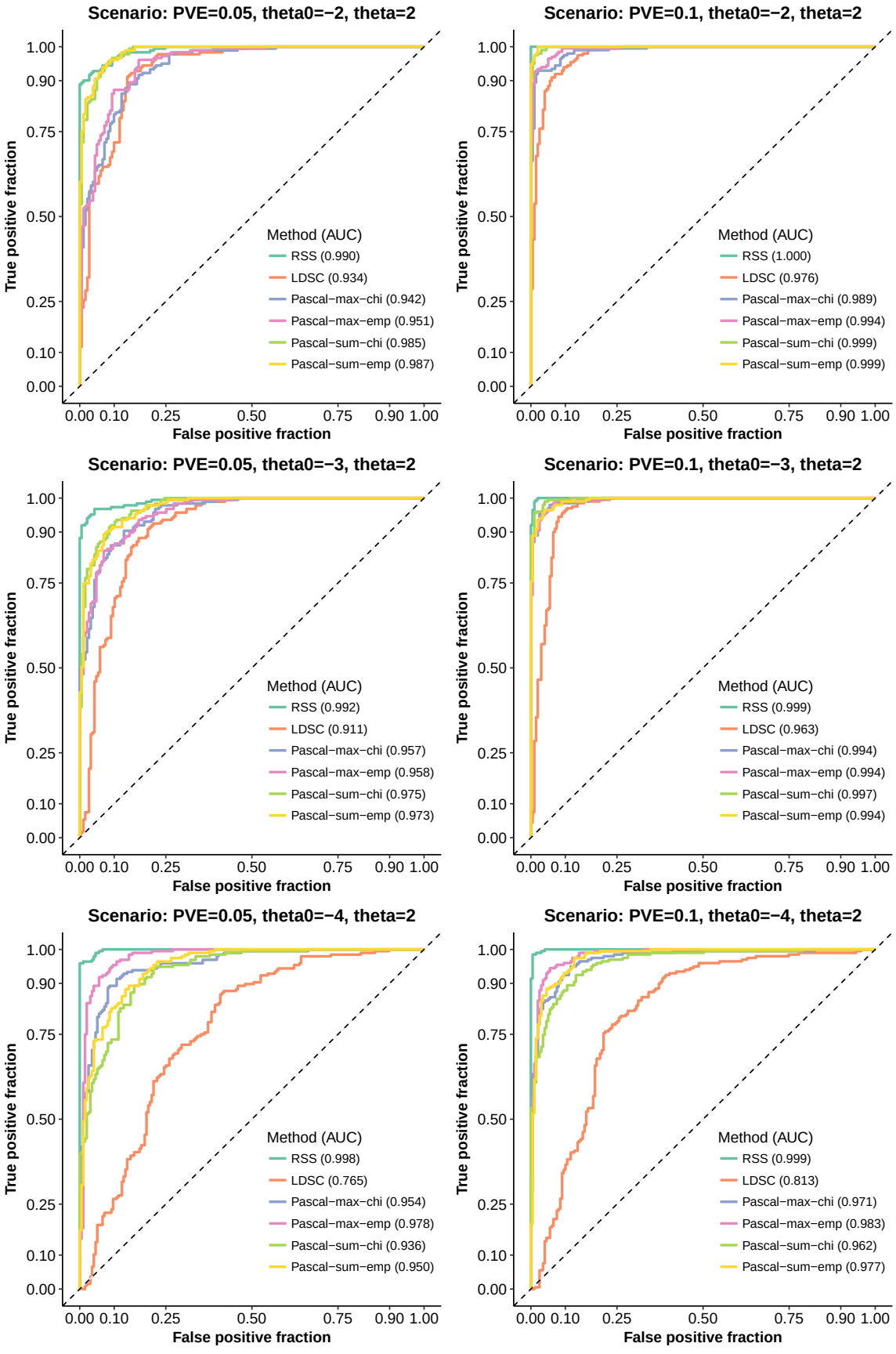
This set of simulations aims to assess the robustness of RSS to model misspecification where SNPs inside the target pathway are not only more likely to be associated with the trait, but also have larger effect on the trait, compared with SNPs outside the target pathway.

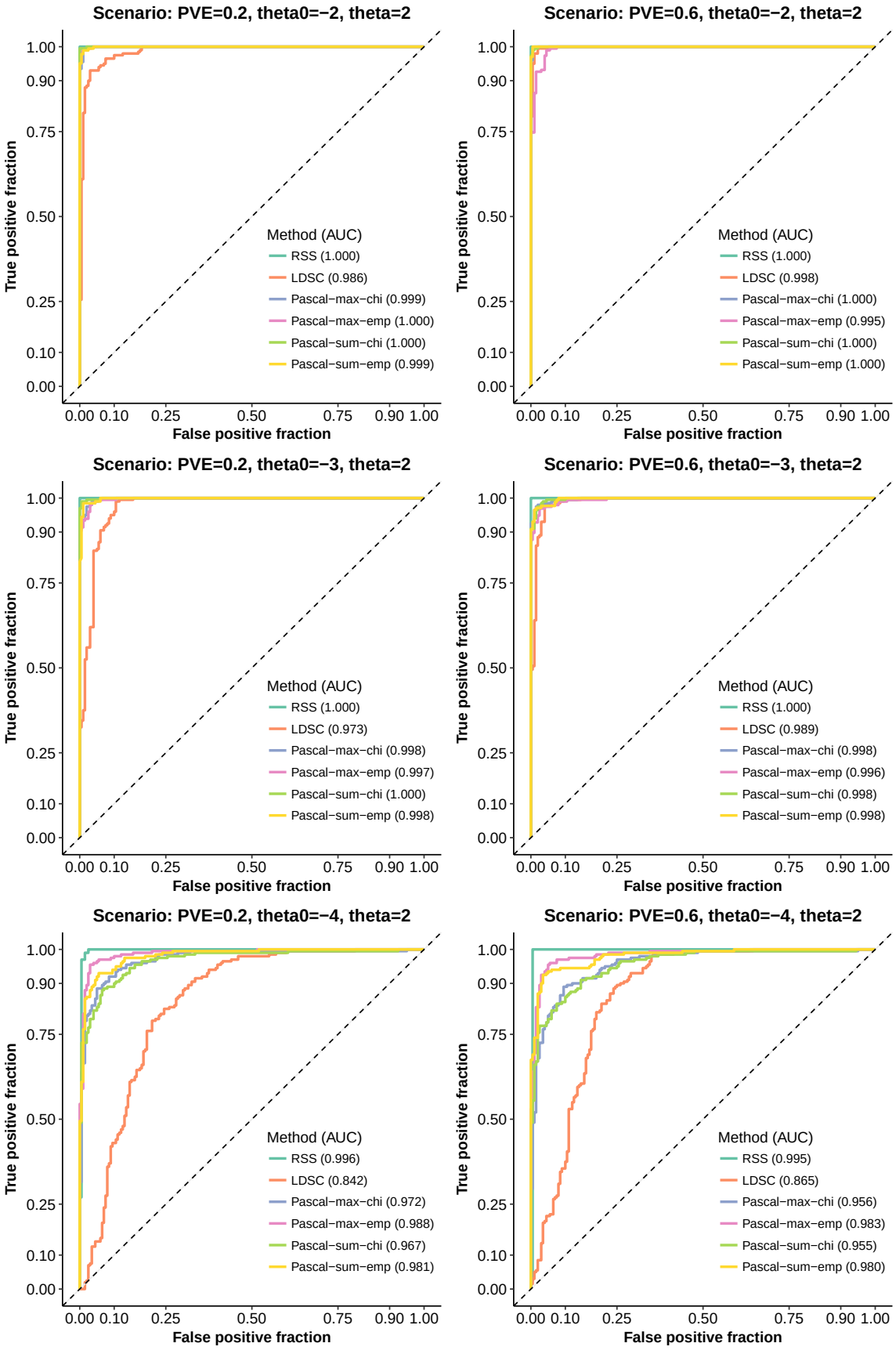
We create this type of enrichment data as follows. We first simulate the causal indicators for an enrichment dataset as in **Supplementary Figure 1**, and then simulate the genetic effects as follows:

$$\begin{aligned} \beta_j|\gamma_j = 0 &\sim \delta_0, \\ \beta_j|\gamma_j = 1, a_j = 0 &\sim \text{Normal}(0, 0.01^2), \\ \beta_j|\gamma_j = 1, a_j = 1 &\sim \text{Normal}(0, 1). \end{aligned}$$

The corresponding baseline dataset is simulated as in **Supplementary Figure 1**.

The "sparse scenario" in **Figure 2(d)** corresponds to simulations with true $\theta_0 = -4$ and PVE = 0.05. The "polygenic scenario" in **Figure 2(d)** corresponds to simulations with true $\theta_0 = -2$ and PVE = 0.05.


Scenario: PVE=0.05, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (0.990)
LDSC (0.934)
Pascal−max−chi (0.942)
Pascal−max−emp (0.951)
Pascal−sum−chi (0.985)
Pascal−sum−emp (0.987)
Scenario: PVE=0.1, theta0=−2, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.976)
Pascal−max−chi (0.989)
Pascal−max−emp (0.994)
Pascal−sum−chi (0.999)
Pascal−sum−emp (0.999)
Scenario: PVE=0.05, theta0=−3, theta=2
Method (AUC)
RSS (0.992)
LDSC (0.911)
Pascal−max−chi (0.957)
Pascal−max−emp (0.958)
Pascal−sum−chi (0.975)
Pascal−sum−emp (0.973)
Scenario: PVE=0.1, theta0=−3, theta=2
Method (AUC)
RSS (0.999)
LDSC (0.963)
Pascal−max−chi (0.994)
Pascal−max−emp (0.994)
Pascal−sum−chi (0.997)
Pascal−sum−emp (0.994)
Scenario: PVE=0.05, theta0=−4, theta=2
Method (AUC)
RSS (0.998)
LDSC (0.765)
Pascal−max−chi (0.954)
Pascal−max−emp (0.978)
Pascal−sum−chi (0.936)
Pascal−sum−emp (0.950)
Scenario: PVE=0.1, theta0=−4, theta=2
Method (AUC)
RSS (0.999)
LDSC (0.813)
Pascal−max−chi (0.971)
Pascal−max−emp (0.983)
Pascal−sum−chi (0.962)
Pascal−sum−emp (0.977)


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (1.000)
LDSC (0.986)
Pascal−max−chi (0.999)
Pascal−max−emp (1.000)
Pascal−sum−chi (1.000)
Pascal−sum−emp (0.999)
Scenario: PVE=0.6, theta0=−2, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.998)
Pascal−max−chi (1.000)
Pascal−max−emp (0.995)
Pascal−sum−chi (1.000)
Pascal−sum−emp (1.000)
Scenario: PVE=0.2, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.973)
Pascal−max−chi (0.998)
Pascal−max−emp (0.997)
Pascal−sum−chi (1.000)
Pascal−sum−emp (0.998)
Scenario: PVE=0.6, theta0=−3, theta=2
Method (AUC)
RSS (1.000)
LDSC (0.989)
Pascal−max−chi (0.998)
Pascal−max−emp (0.996)
Pascal−sum−chi (0.998)
Pascal−sum−emp (0.998)
Scenario: PVE=0.2, theta0=−4, theta=2
Method (AUC)
RSS (0.996)
LDSC (0.842)
Pascal−max−chi (0.972)
Pascal−max−emp (0.988)
Pascal−sum−chi (0.967)
Pascal−sum−emp (0.981)
Scenario: PVE=0.6, theta0=−4, theta=2
Method (AUC)
RSS (0.995)
LDSC (0.865)
Pascal−max−chi (0.956)
Pascal−max−emp (0.983)
Pascal−sum−chi (0.955)
Pascal−sum−emp (0.980)

# Supplementary Figure 5

**Simulations to assess the impact of "HapMap3 SNP subsetting" on enrichment analyses.** Details of this set of simulations are almost identical to those in **Supplementary Figure 1**. Here we only highlight the differences.

This set of simulations aims to assess the impact of "SNP subsetting" on enrichment analyses. The "SNP subsetting" strategy uses only the summary data of HapMap3 SNPs in the analysis, even though the summary data of all 1000 Genomes SNPs are available. This strategy has been widely used in recent analyses of GWAS summary statistics (Bulik-Sullivan et al. 2015; Finucane et al. 2015), and it is used in real data analyses of the present study to reduce computations. (We do not use this strategy in any simulation study except this one.)

To mimic "SNP subsetting" used in our real data analyses, we use genotypes of 255,584 SNPs on chromosome 16 from 503 individuals of European ancestry in the 1000 Genomes Phase 3 panel (1000 Genomes Project Consortium 2015) to simulate phenotype data, and then compute single-SNP association summary statistics for these 255,584 1000 Genomes SNPs. The remaining simulation details are the same as those in **Supplementary Figure 1**.

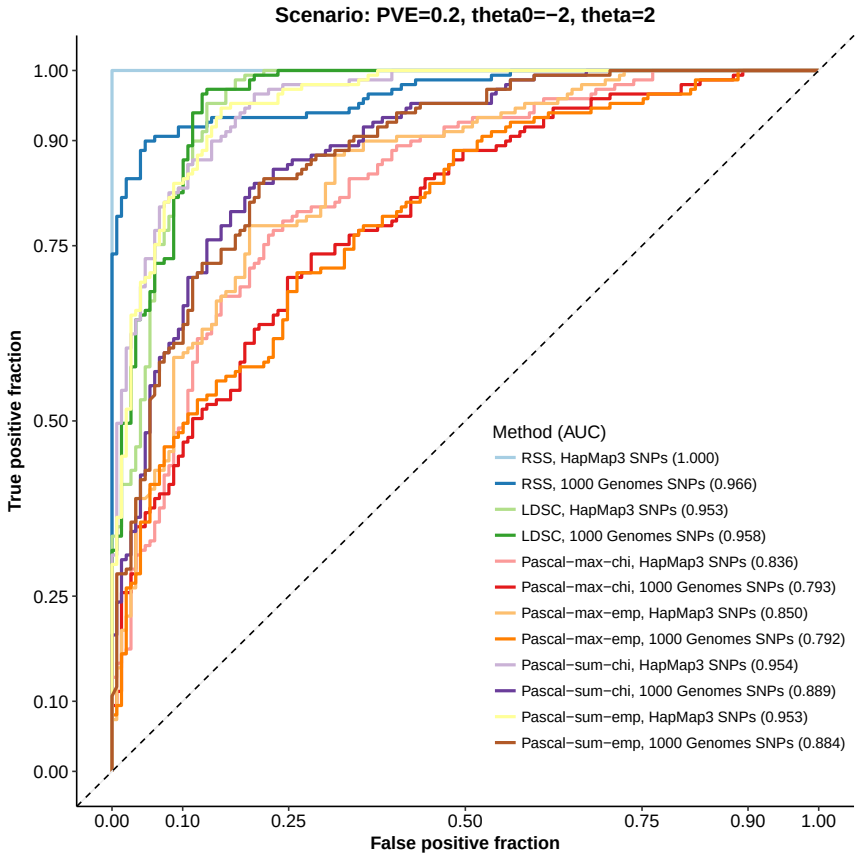
For each simulated dataset, we use each enrichment method to perform two analyses.

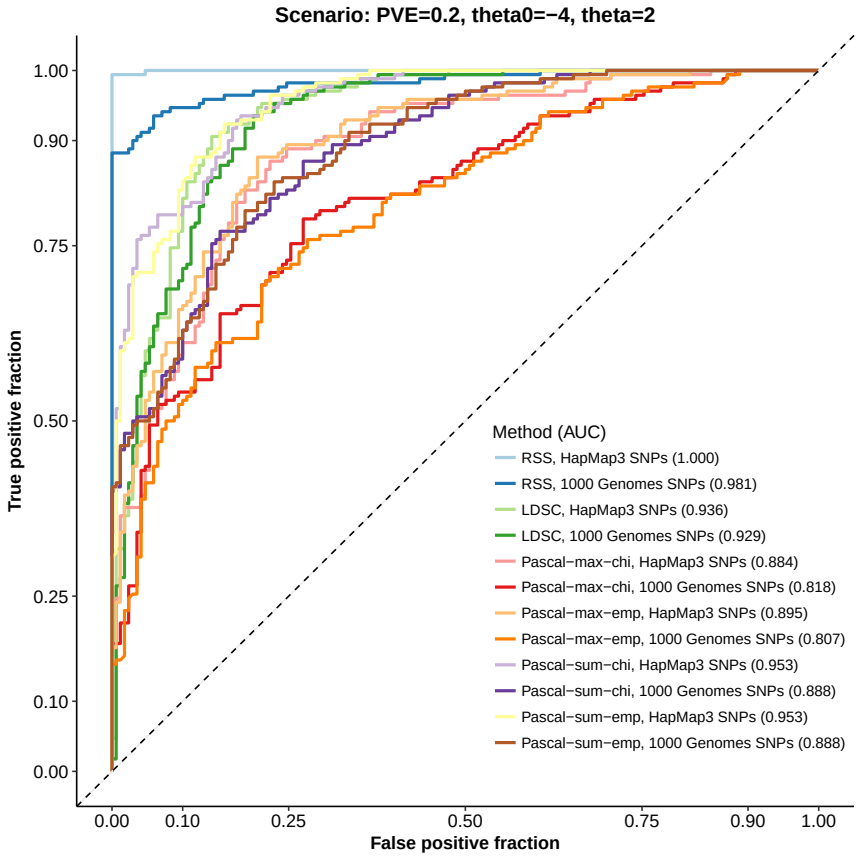
- Analysis "with SNP subsetting": apply the method to the summary data of 36,121 HapMap3 SNPs **only**.
- Analysis "without SNP subsetting": apply the method to the summary data of **all** 255,584 1000 Genomes SNPs.

RSS with SNP subsetting slightly outperforms RSS without SNP subsetting. This is expected, since the variational inference algorithm underlying RSS performs better when less variables (SNPs) are highly correlated (in high LD), and there are more highly correlated SNPs among 255,584 1000 Genomes SNPs than 36,121 HapMap3 SNPs. See Carbonetto and Stephens (2012) for more extensive discussion.

Similarly, Pascal with SNP subsetting also outperforms Pascal without SNP subsetting. We speculate that the large number of highly correlated 1000 Genomes SNPs has a negative impact on the gene-score calculations in Pascal.

For LDSC, results with and without SNP subsetting are almost identical. This is because for both analyses, LDSC only uses LD scores of HapMap3 SNPs as regression weights (downloaded from `https://data.broadinstitute.org/alkesgroup/LDSCORE/weights_hm3_no_hla.tgz`).


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
0.00
0.10
0.25
0.50
0.75
0.90
1.00
Method (AUC)
RSS, HapMap3 SNPs (1.000)
RSS, 1000 Genomes SNPs (0.966)
LDSC, HapMap3 SNPs (0.953)
LDSC, 1000 Genomes SNPs (0.958)
Pascal−max−chi, HapMap3 SNPs (0.836)
Pascal−max−chi, 1000 Genomes SNPs (0.793)
Pascal−max−emp, HapMap3 SNPs (0.850)
Pascal−max−emp, 1000 Genomes SNPs (0.792)
Pascal−sum−chi, HapMap3 SNPs (0.954)
Pascal−sum−chi, 1000 Genomes SNPs (0.889)
Pascal−sum−emp, HapMap3 SNPs (0.953)
Pascal−sum−emp, 1000 Genomes SNPs (0.884)


Scenario: PVE=0.2, theta0=−4, theta=2
True positive fraction
False positive fraction
0.00
0.10
0.25
0.50
0.75
0.90
1.00
Method (AUC)
RSS, HapMap3 SNPs (1.000)
RSS, 1000 Genomes SNPs (0.981)
LDSC, HapMap3 SNPs (0.936)
LDSC, 1000 Genomes SNPs (0.929)
Pascal−max−chi, HapMap3 SNPs (0.884)
Pascal−max−chi, 1000 Genomes SNPs (0.818)
Pascal−max−emp, HapMap3 SNPs (0.895)
Pascal−max−emp, 1000 Genomes SNPs (0.807)
Pascal−sum−chi, HapMap3 SNPs (0.953)
Pascal−sum−chi, 1000 Genomes SNPs (0.888)
Pascal−sum−emp, HapMap3 SNPs (0.953)
Pascal−sum−emp, 1000 Genomes SNPs (0.888)

# Supplementary Figure 6

**Simulation details and additional results of Figure 3 "Baseline" panels.** Here we use real genotypes of 12,758 SNPs on chromosome 16 from 1458 individuals in the UK Blood Service Control Group (Wellcome Trust Case Control Consortium 2007) to simulate phenotype data, and then compute single-SNP association summary statistics. On these summary data, we compare RSS with existing gene-level testing methods.

We first simulate the genetic effect $\beta_j$ of each SNP $j$ as follows:

$$\beta_j \sim (1-\pi)\cdot\delta_0 + \pi\cdot\text{Normal}(0,1),$$

where $\delta_0$ denotes point mass at zero and $\pi = (1+10^{-\theta_0})^{-1}$. We then simulate the phenotype $y_i$ of Individual $i$ as

$$y_i = \textstyle\sum_j x_{ij}\beta_j + \epsilon_i,\ \epsilon_i \sim \text{Normal}(0,\tau^{-1}),$$

where $x_{ij}$ is the genotype of SNP $j$ for Individual $i$, $j = 1,\ldots,12{,}758$ and $i = 1,\ldots,1458$. The true value of residual variance $\tau^{-1}$ is determined by the true value of PVE (total proportion of variance in phenotype $y$ explained by effects of all available SNPs $X$) as follows:

$$\text{PVE} = V(X\beta)/(\tau^{-1} + V(X\beta)),$$

where $V(X\beta)$ is the sample variance of $X\beta$. We adopt the phenotype simulation scheme from previous work, notably Guan and Stephens (2011), Zhou, Carbonetto, and Stephens (2013) and Zhu and Stephens (2017).

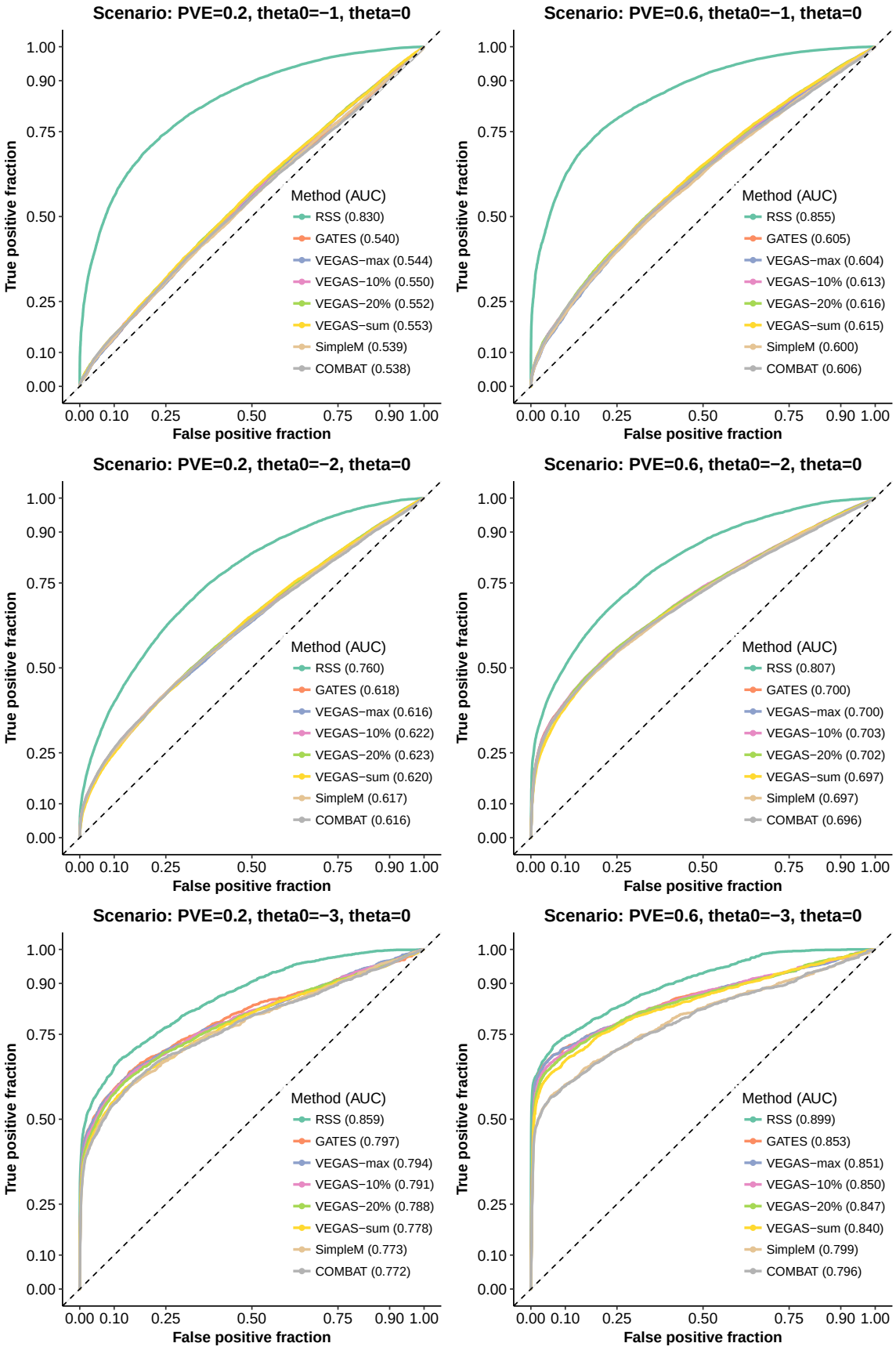
In this set of simulations, the true values of background parameter $\theta_0$ are $\{-1,-2,-3\}$, and the true values of PVE are $\{0.2, 0.6\}$. For each combination of true values of $(\theta_0, \text{PVE})$, we simulate 100 independent datasets. The "Sparse & Baseline" panel in **Figure 3** corresponds to simulations with true $\theta_0 = -3$ and PVE = 0.2. The "Polygenic & Baseline" panel in **Figure 3** corresponds to simulations with true $\theta_0 = -1$ and PVE = 0.2.

We apply RSS to the simulated datasets, using the program `rss_varbvsr_squarem.m` available at `https://github.com/stephenslab/rss`. The input LD matrix is estimated from the 1480 control individuals in the WTCCC 1958 British Birth Cohort, using a shrinkage estimator from Wen and Stephens (2010). The enrichment parameter $\theta$ is set as zero. The grid on the background parameter $\theta_0$ is `(-5:0.1:-1)`.

For each simulated dataset, we also perform gene-level association analysis using four existing approaches with their default settings: SimpleM (Gao, Starmer, and Martin 2008), VEGAS (Liu et al. 2010), GATES (Li et al. 2011) and COMBAT (Wang et al. 2017) that combines SimpleM, VEGAS and GATES results. Note that VEGAS is applied to the full set of SNPs within a gene (`-sum`), on a specified percentage of the most significant SNPs (`-10%` and `-20%`), or on the single most significant SNP (`-max`).

For each simulated dataset, we define a gene as "trait-associated" if at lease one SNP within ±100 kb of the transcribed region of this gene has non-zero effect ($\beta_j \neq 0$). For each gene in each simulated dataset, RSS produces $P_1$, the posterior probability that the gene is trait-associated. whereas the other methods produce p-value with the null hypothesis that the gene is not trait-associated; these statistics are used to rank the significance of gene-level associations. If a method identifies association between a non-trait-associated gene and the trait, then it is a "false positive". If a method identifies association between a trait-associated gene and the trait, then it is a "true positive".

We evaluate the performance of these gene-level association methods by plotting the receiver operating characteristic (ROC) curve and computing the area under the curve (AUC) for each method. Both metrics are implemented in the package `plotROC` (Sachs 2017).


Scenario: PVE=0.2, theta0=−1, theta=0
True positive fraction
False positive fraction
Method (AUC)
RSS (0.830)
GATES (0.540)
VEGAS−max (0.544)
VEGAS−10% (0.550)
VEGAS−20% (0.552)
VEGAS−sum (0.553)
SimpleM (0.539)
COMBAT (0.538)
Scenario: PVE=0.6, theta0=−1, theta=0
Method (AUC)
RSS (0.855)
GATES (0.605)
VEGAS−max (0.604)
VEGAS−10% (0.613)
VEGAS−20% (0.616)
VEGAS−sum (0.615)
SimpleM (0.600)
COMBAT (0.606)
Scenario: PVE=0.2, theta0=−2, theta=0
Method (AUC)
RSS (0.760)
GATES (0.618)
VEGAS−max (0.616)
VEGAS−10% (0.622)
VEGAS−20% (0.623)
VEGAS−sum (0.620)
SimpleM (0.617)
COMBAT (0.616)
Scenario: PVE=0.6, theta0=−2, theta=0
Method (AUC)
RSS (0.807)
GATES (0.700)
VEGAS−max (0.700)
VEGAS−10% (0.703)
VEGAS−20% (0.702)
VEGAS−sum (0.697)
SimpleM (0.697)
COMBAT (0.696)
Scenario: PVE=0.2, theta0=−3, theta=0
Method (AUC)
RSS (0.859)
GATES (0.797)
VEGAS−max (0.794)
VEGAS−10% (0.791)
VEGAS−20% (0.788)
VEGAS−sum (0.778)
SimpleM (0.773)
COMBAT (0.772)
Scenario: PVE=0.6, theta0=−3, theta=0
Method (AUC)
RSS (0.899)
GATES (0.853)
VEGAS−max (0.851)
VEGAS−10% (0.850)
VEGAS−20% (0.847)
VEGAS−sum (0.840)
SimpleM (0.799)
COMBAT (0.796)

# Supplementary Figure 7

**Simulation details and additional results of Figure 3 "Enrichment" panels.** Details of this set of simulations are almost identical to those in **Supplementary Figure 6**. Here we only highlight the differences.

For this set of simulations, we generate the genetic effect of each SNP $j$ as follows:

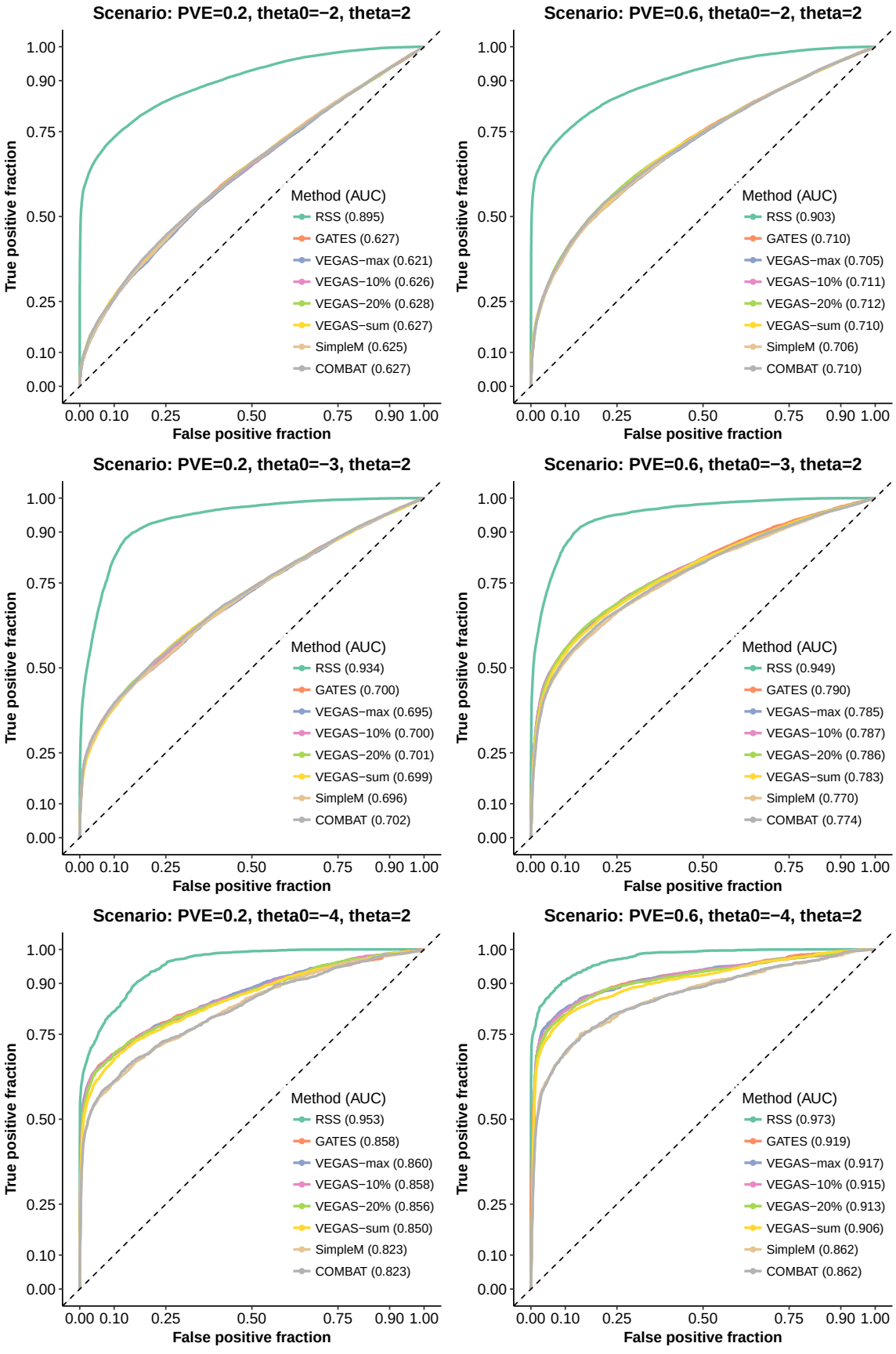
$$\beta_j \sim (1-\pi_j)\cdot\delta_0 + \pi_j \cdot \text{Normal}(0,1),$$
$$\pi_j = (1+10^{-(\theta_0+a_j\theta)})^{-1},$$

where $\{\theta_0, \theta, a_j\}$ are defined in **Supplementary Figure 1**.

In this set of simulations, the true values of background parameter $\theta_0$ are $\{-2,-3,-4\}$, the true value of enrichment parameter $\theta$ is 2, and the true values of PVE are $\{0.2, 0.6\}$.

When applying RSS to the simulated datasets, we fix the the background parameter $\theta_0$ as the true value, and use a grid `(1:0.05:3)` for the enrichment parameter $\theta$.

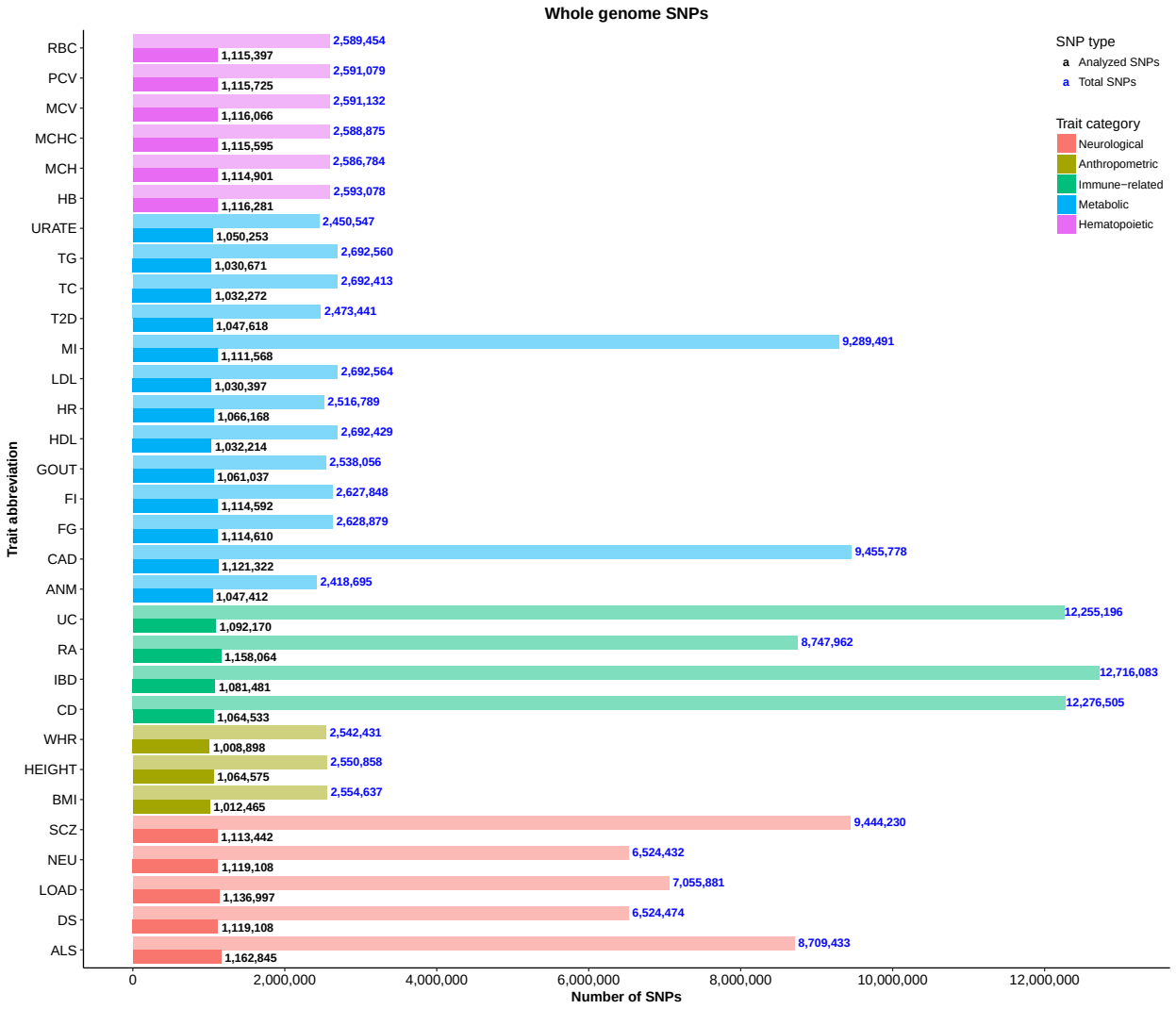
The "Sparse & Enrichment" panel in **Figure 3** corresponds to simulations with true $\theta_0 = -4$ and PVE = 0.2. The "Polygenic & Enrichment" panel in **Figure 3** corresponds to simulations with true $\theta_0 = -2$ and PVE = 0.2.


Scenario: PVE=0.2, theta0=−2, theta=2
True positive fraction
False positive fraction
Method (AUC)
RSS (0.895)
GATES (0.627)
VEGAS−max (0.621)
VEGAS−10% (0.626)
VEGAS−20% (0.628)
VEGAS−sum (0.627)
SimpleM (0.625)
COMBAT (0.627)
Scenario: PVE=0.6, theta0=−2, theta=2
RSS (0.903)
GATES (0.710)
VEGAS−max (0.705)
VEGAS−10% (0.711)
VEGAS−20% (0.712)
VEGAS−sum (0.710)
SimpleM (0.706)
COMBAT (0.710)
Scenario: PVE=0.2, theta0=−3, theta=2
RSS (0.934)
GATES (0.700)
VEGAS−max (0.695)
VEGAS−10% (0.700)
VEGAS−20% (0.701)
VEGAS−sum (0.699)
SimpleM (0.696)
COMBAT (0.702)
Scenario: PVE=0.6, theta0=−3, theta=2
RSS (0.949)
GATES (0.790)
VEGAS−max (0.785)
VEGAS−10% (0.787)
VEGAS−20% (0.786)
VEGAS−sum (0.783)
SimpleM (0.770)
COMBAT (0.774)
Scenario: PVE=0.2, theta0=−4, theta=2
RSS (0.953)
GATES (0.858)
VEGAS−max (0.860)
VEGAS−10% (0.858)
VEGAS−20% (0.856)
VEGAS−sum (0.850)
SimpleM (0.823)
COMBAT (0.823)
Scenario: PVE=0.6, theta0=−4, theta=2
RSS (0.973)
GATES (0.919)
VEGAS−max (0.917)
VEGAS−10% (0.915)
VEGAS−20% (0.913)
VEGAS−sum (0.906)
SimpleM (0.862)
COMBAT (0.862)

# Supplementary Figure 8

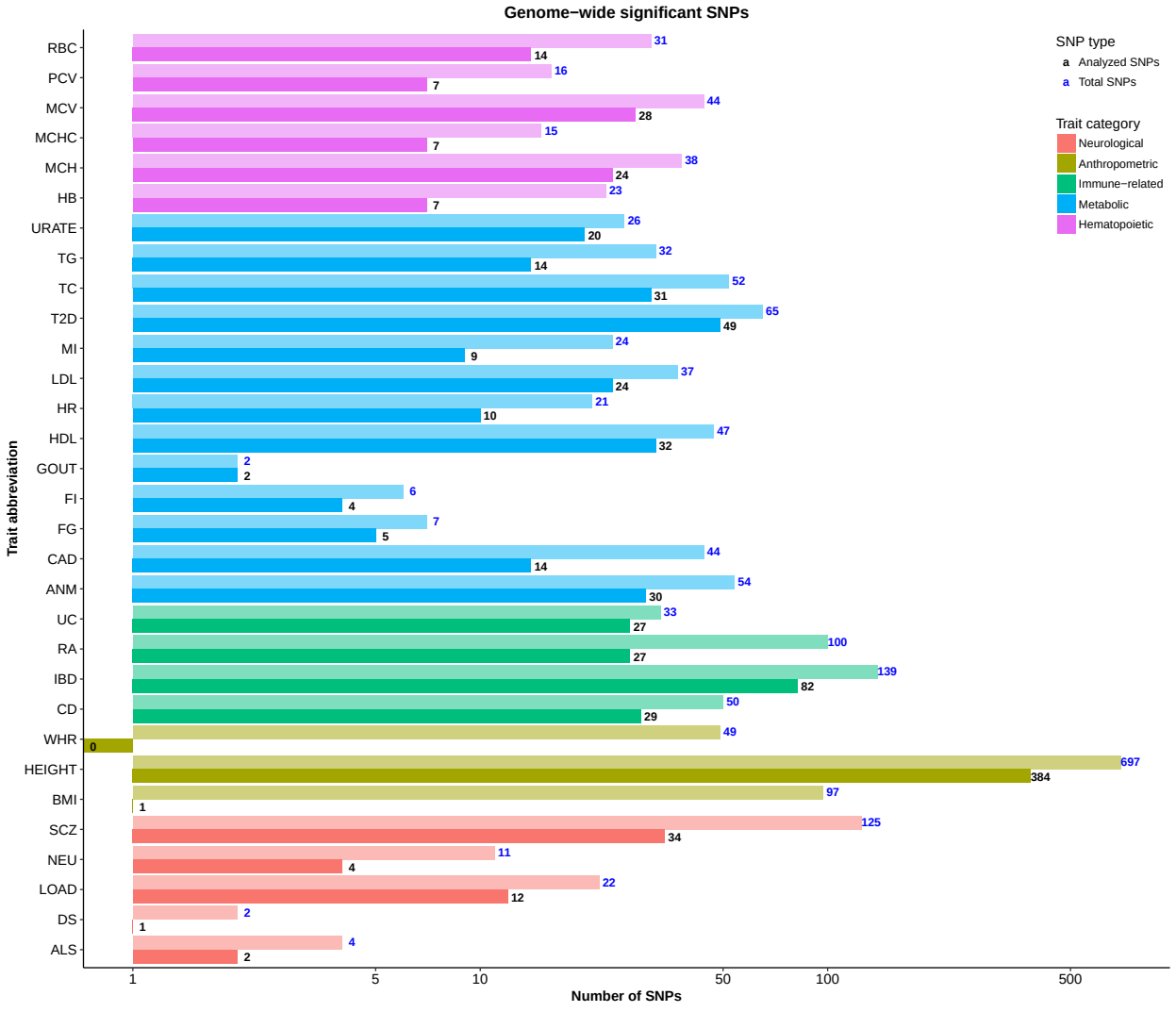
**Summary of genetic variants.** Panel **(a)** reports numbers of all genetic variants available in corresponding GWAS. Panel **(b)** reports numbers of GWAS hits (i.e. loci or SNPs reaching genome-wide significance) reported in corresponding publications. For both panels, "total SNPs" denote SNPs that were available in corresponding publications and/or summary data files (bar charts with higher transparency; blue numbers); "analyzed SNPs" denote SNPs analyzed in the present study (bar charts with lower transparency; black numbers).

**(a)** Total numbers of SNPs available in corresponding GWAS.



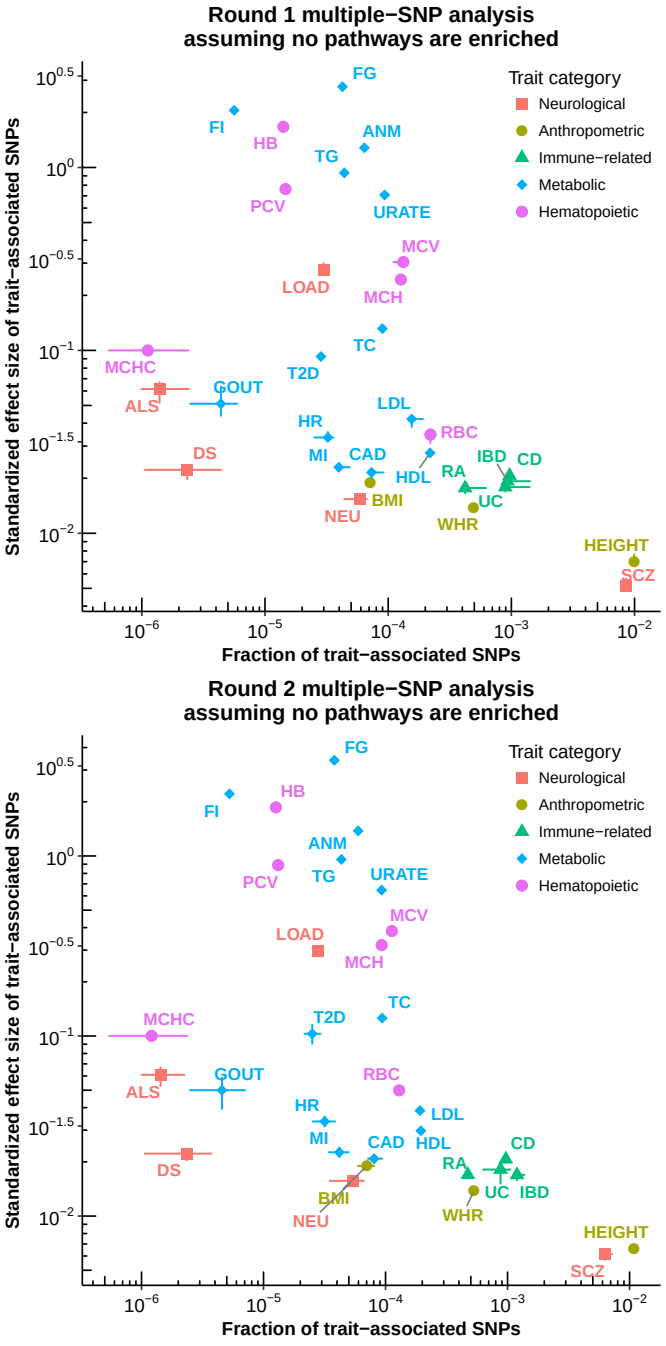
**(b)** The "genome-wide significant SNPs" are sentinel SNPs reaching genome-wide significance ($p < 5 \times 10^{-8}$), which are reported in corresponding publications of GWAS. The $x$-axis uses a logarithmic scale (base 10).

Please note that the numbers of "analyzed" genome-wide significant SNPs for body mass index (BMI, Locke et al. 2015) and waist-to-hip ratio (WHR) adjusted for BMI (Shungin et al. 2015) are extremely small. This is because summary data from both studies were combined results of GWAS arrays and custom arrays (e.g. Metabochip, Voight et al. 2012), and we excluded SNPs on custom arrays from our analyses. Custom arrays harbored almost all GWAS hits for these two traits, so there are only **zero** and **one** GWAS hit left for WHR and BMI in our analyses. This explains why the number of "analyzed hits" is **zero** for WHR and **one** for BMI as shown in **Supplementary Figure 12**.



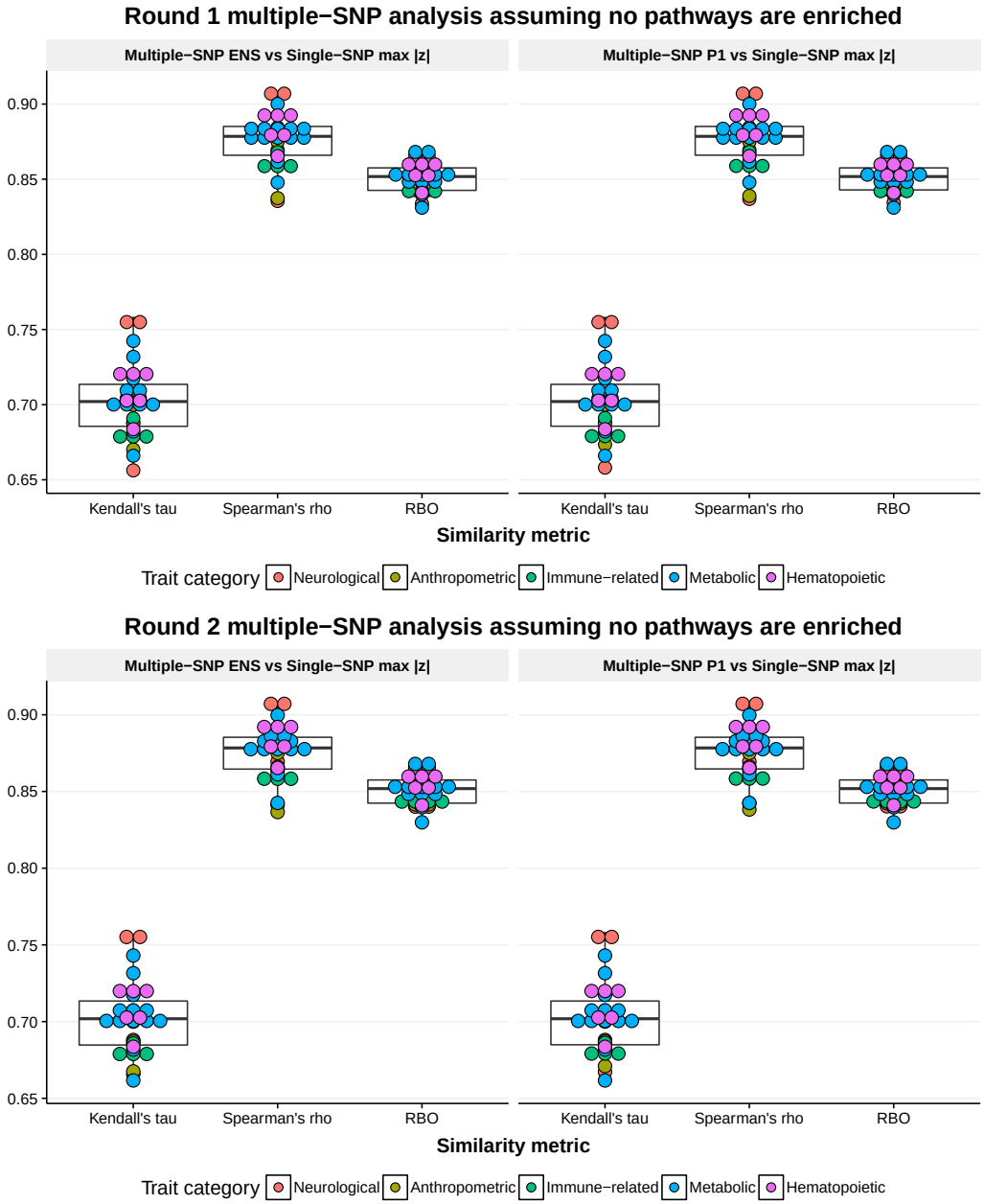
# Supplementary Figure 9

**Inferred effect size distributions of 31 human phenotypes, assuming that no pathways are enriched.** For each trait, its effect size distribution is summarized as the fraction of trait-associated SNPs and the standardized effect size of trait-associated SNPs. See **Supplementary Note** for details of these two quantities. Each dot represents a trait, where the horizontal point range indicates the posterior mean and 95% credible interval (C.I.) of fraction of trait-associated SNPs, and the vertical point range indicates the posterior mean and 95% C.I. of standardized effect size. Both axes use a logarithmic scale (base 10). The numerical values of these posterior statistics are provided in **Supplementary Table 2**. The details of "Round 1" and "Round 2" analyses are provided in **Supplementary Table 6**. **Figure 4(a)** is the same as the "Round 2" panel below.



# Supplementary Figure 10

**Ranking similarity between genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched.** We first divide the entire genome into overlapped loci of 50 SNPs (with an overlap of 25 SNPs between neighboring loci). For each trait and each locus, we then use the same summary data to compute i) the maximum single-SNP $|z|$-score; ii) the posterior probability that the locus contains at least one trait-associated SNP ($P_1$); and iii) the posterior expected number of trait-associated SNPs in the locus (ENS). Note that the posterior statistics $P_1$ and ENS are obtained from the Bayesian genome-wide multiple-SNP analyses. Based on these three locus-level statistics, we obtain three ranked lists of loci for each trait, and then evaluate their similarity via i) Spearman's $\rho$ statistic; ii) Kendall's $\tau$ statistic and iii) rank biased overlap (RBO, Webber, Moffat, and Zobel 2010). The Spearman's $\rho$ and Kendall's $\tau$ statistic are computed by `R` function `cor`. The RBO is computed by the function `rbo` in the package `gespeR` (Schmich 2015). The details of "Round 1" and "Round 2" analyses are provided in **Supplementary Table 6**.

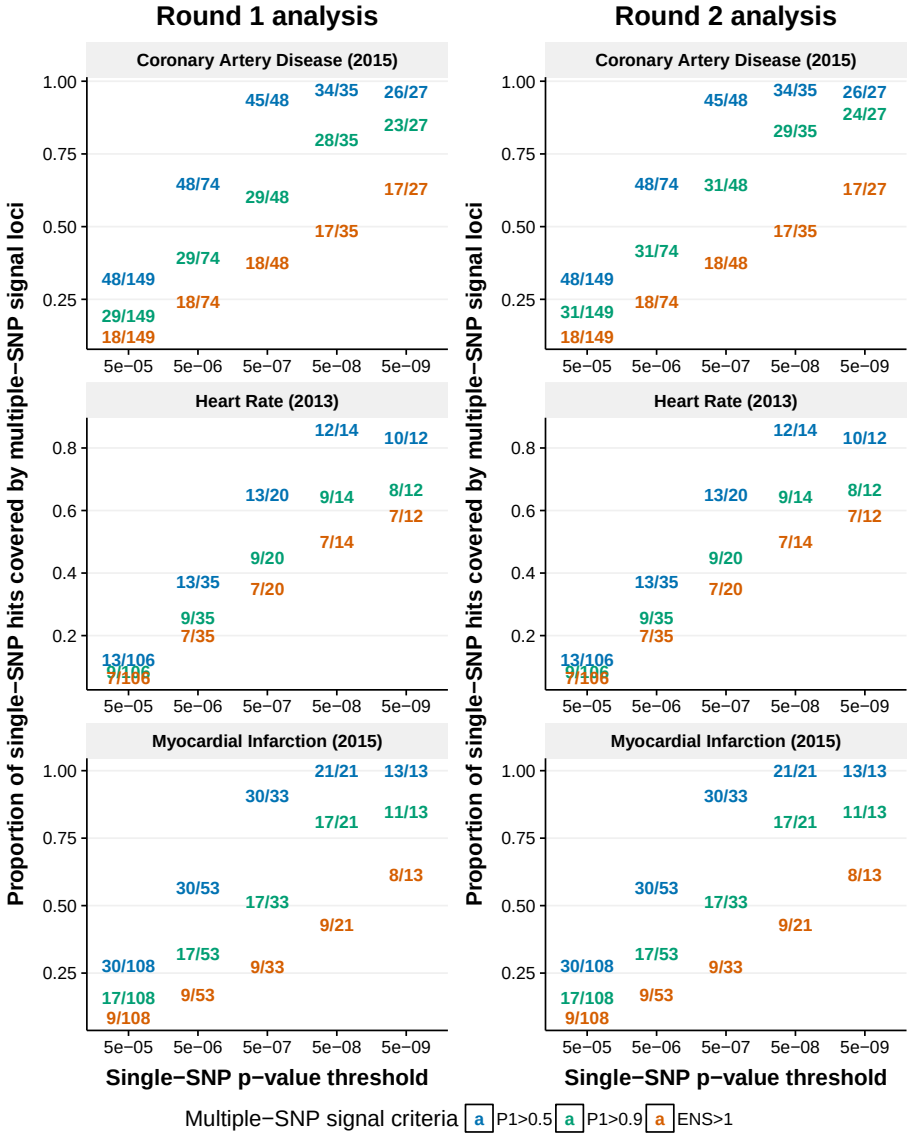


# Supplementary Figure 11

**Concordance between genome-wide single-SNP and multiple-SNP analyses of 31 phenotypes, both assuming that no pathways are enriched.** For a given trait, the concordance between single-SNP and multiple-SNP analyses is measured by the proportion of "single-SNP hits" covered by "multiple-SNP signal loci" ($y$ axis). The "single-SNP hits" are SNPs reaching significance (for a given $p$-value threshold shown in $x$ axis) and separated by at least 1 Mb. The "multiple-SNP signal loci" are predefined genomic regions satisfying certain criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS > 1). For each trait, both single-SNP and multiple-SNP analyses are performed on the **same** summary-level data. See **Supplementary Figure 10** for the definition of locus and multiple-SNP posterior statistics ($P_1$ and ENS). See **Supplementary Table 6** for the details of "Round 1" and "Round 2" analyses.

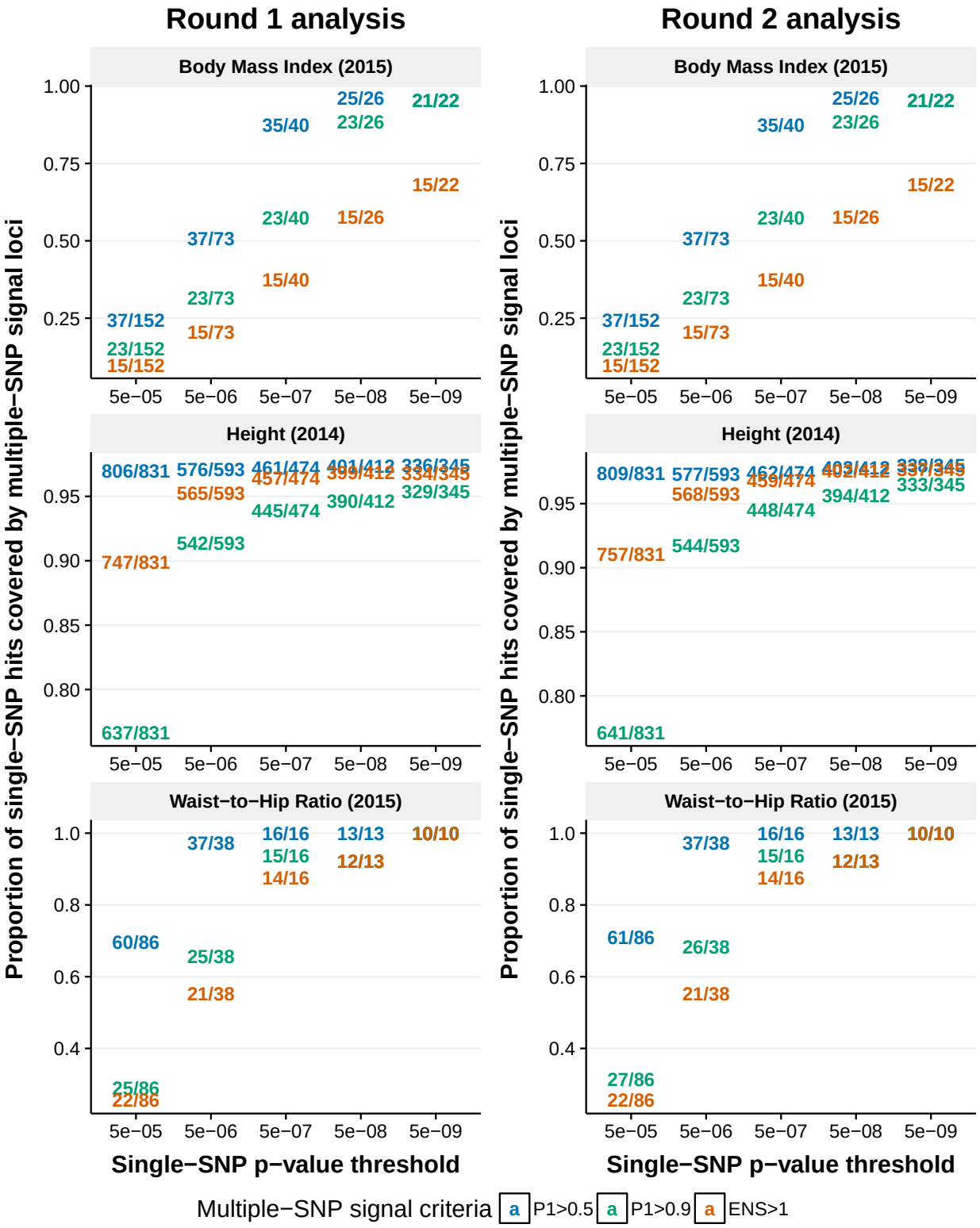
## Heart-related traits

The three heart-related traits are heart rate (Den Hoed et al. 2013), coronary artery disease and myocardial infarction (Nikpay et al. 2015).



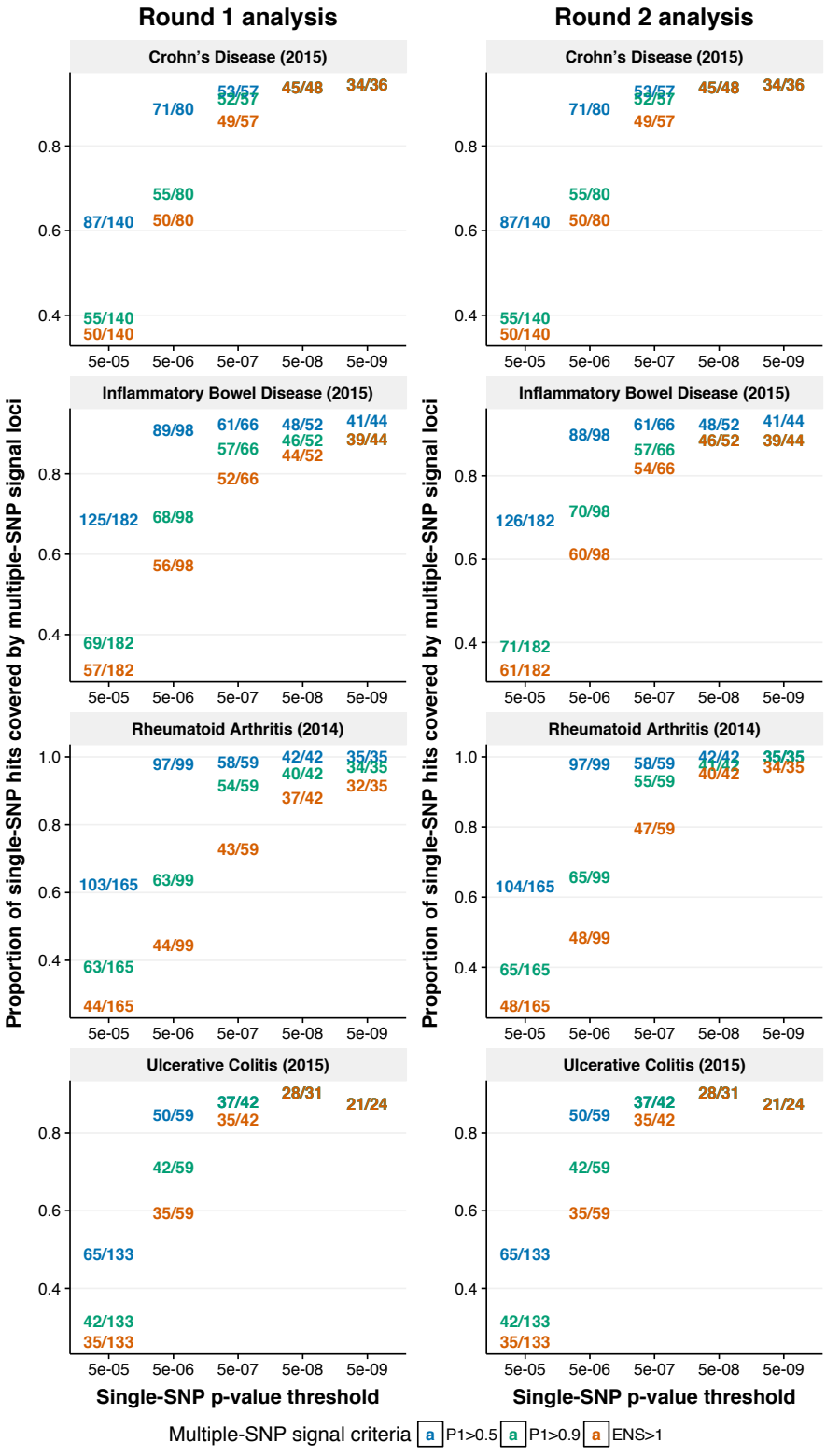
## Anthropometric traits

The three anthropometric traits are adult height (Wood et al. 2014), body mass index (Locke et al. 2015) and waist-to-hip ratio after adjusting for body mass index (Shungin et al. 2015).



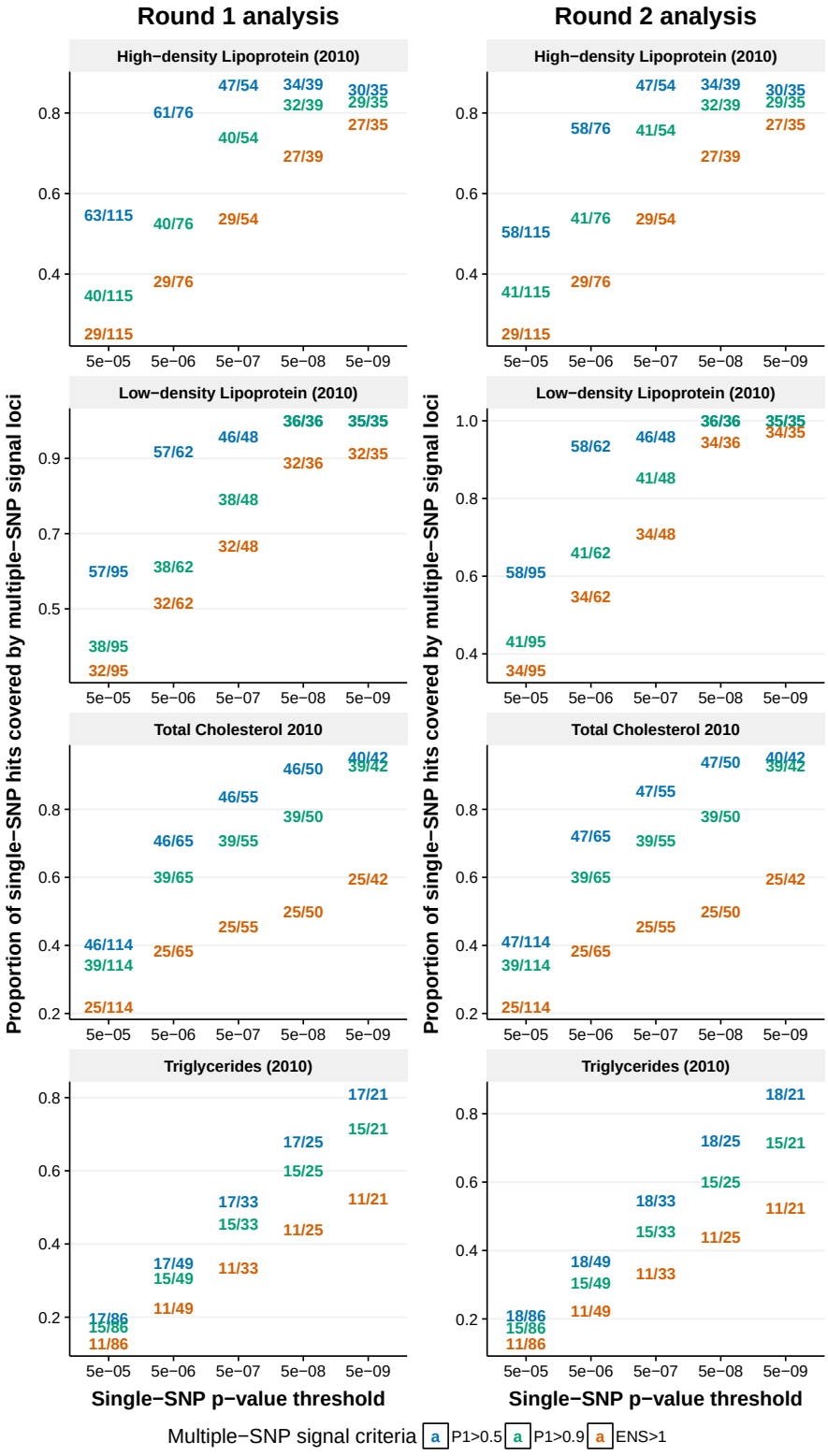
## Immune-related traits

The four immune-related traits are rheumatoid arthritis (Okada et al. 2014), inflammatory bowel disease, Crohn's disease and ulcerative colitis (Liu et al. 2015).



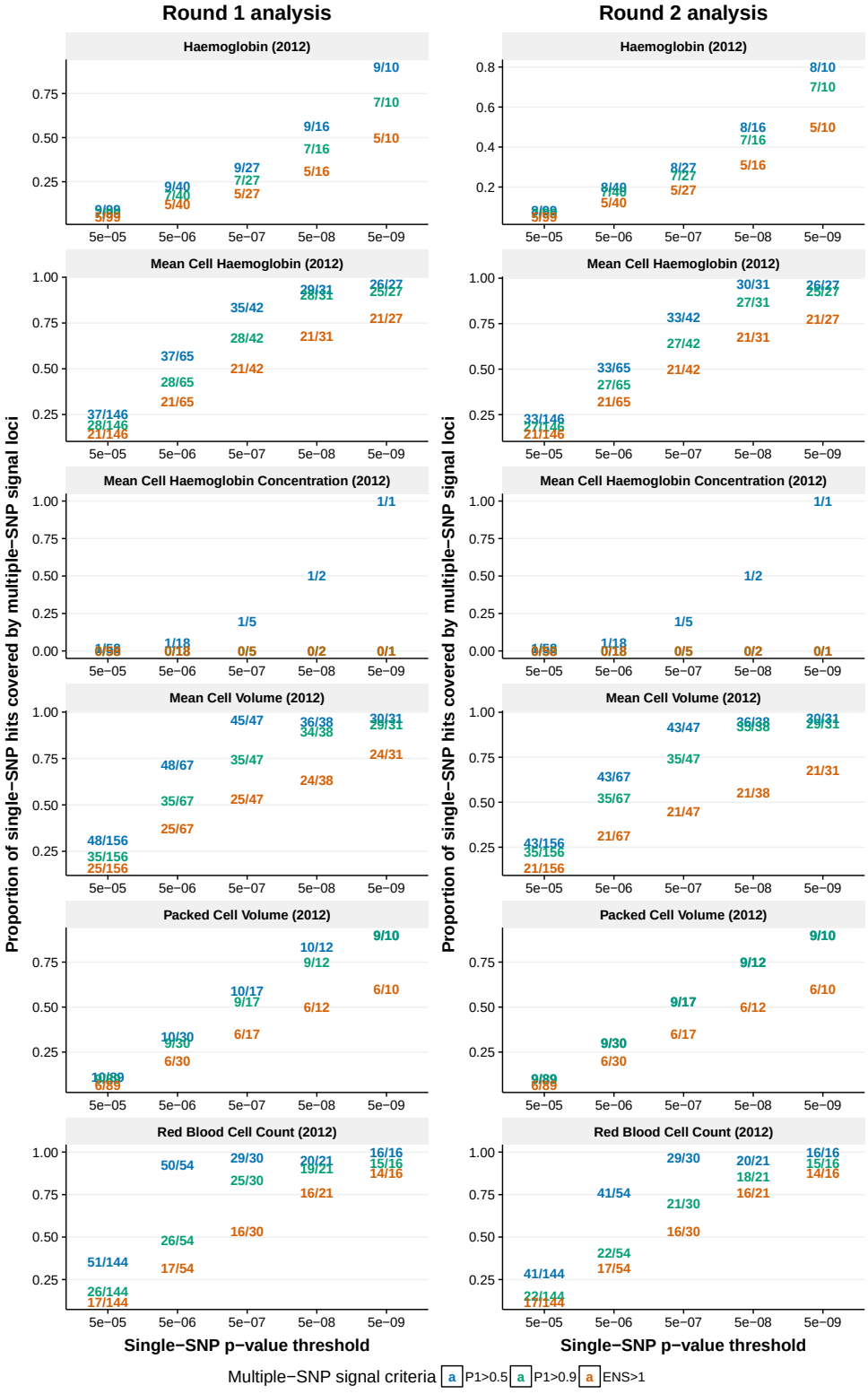
## Blood lipid traits

The four blood lipid traits are total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglycerides (Teslovich et al. 2010).



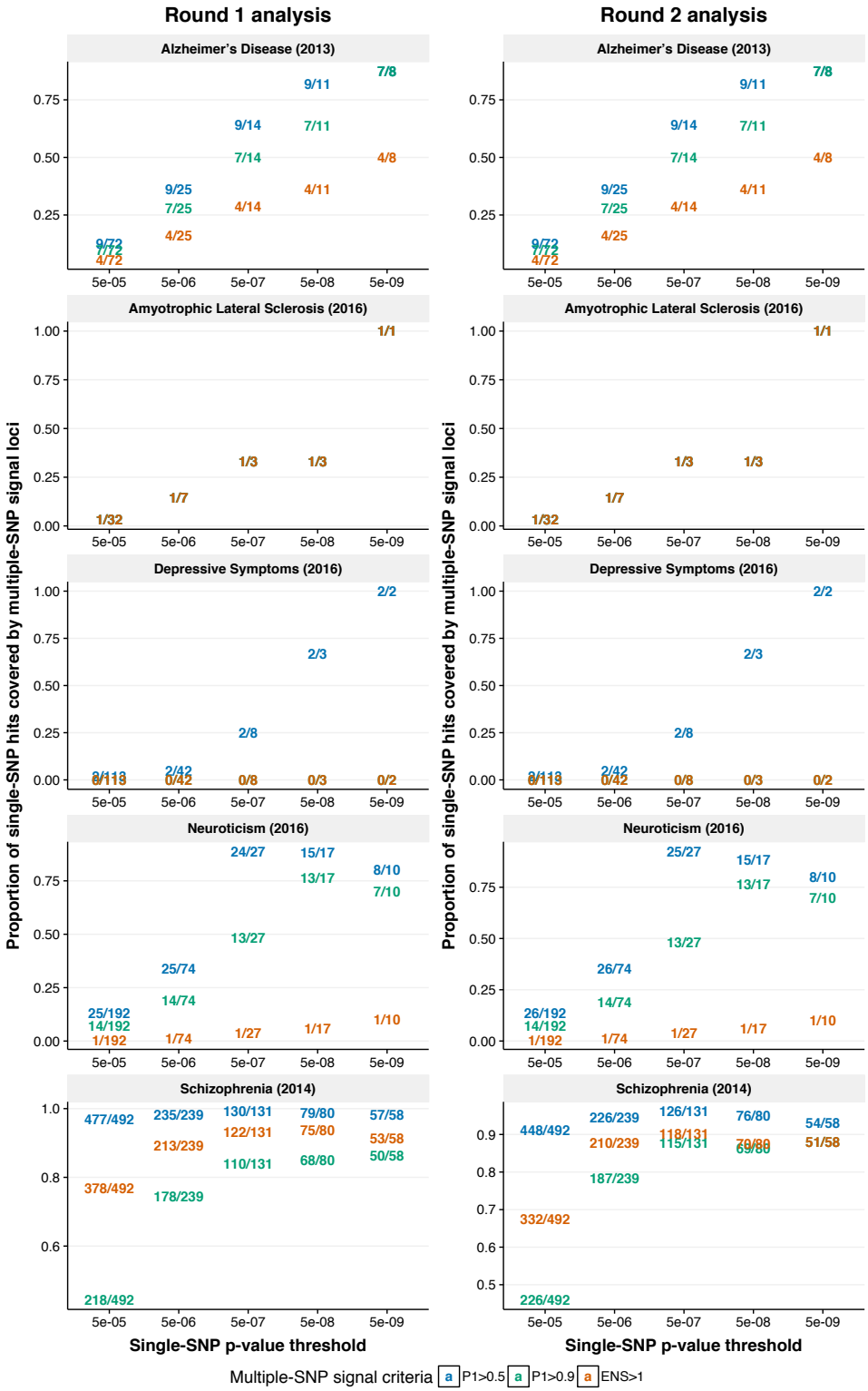
## Hematopoietic traits

The six red blood cell traits are haemoglobin, mean cell haemoglobin, mean cell haemoglobin concentration, mean cell volume, packed cell volume and red blood cell count (Harst et al. 2012).



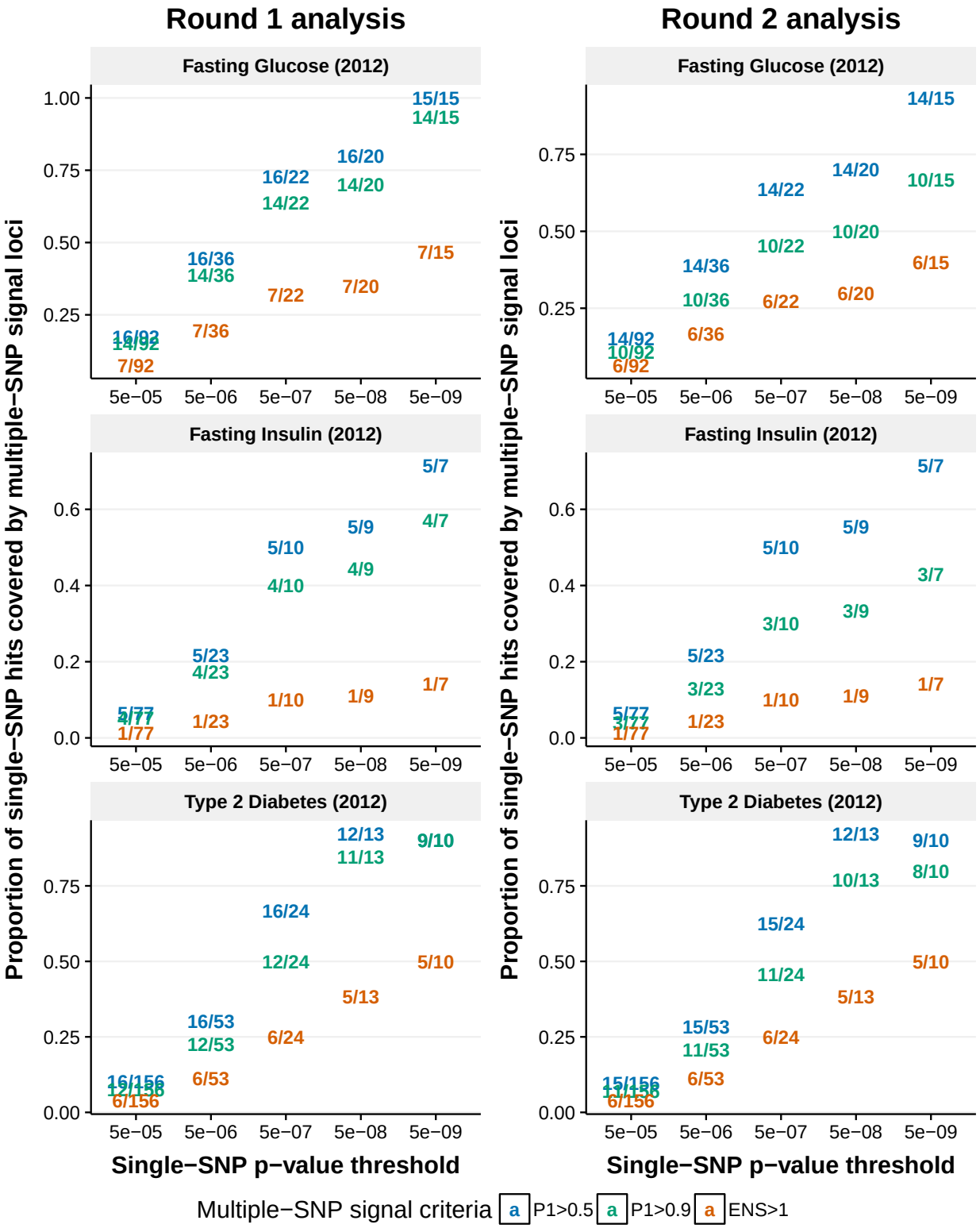
## Neurological phenotypes

The five neurological phenotypes: Alzheimer's disease (Lambert et al. 2013), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), amyotrophic lateral sclerosis (Rheenen et al. 2016), depressive symptoms and neuroticism (Okbay et al. 2016).



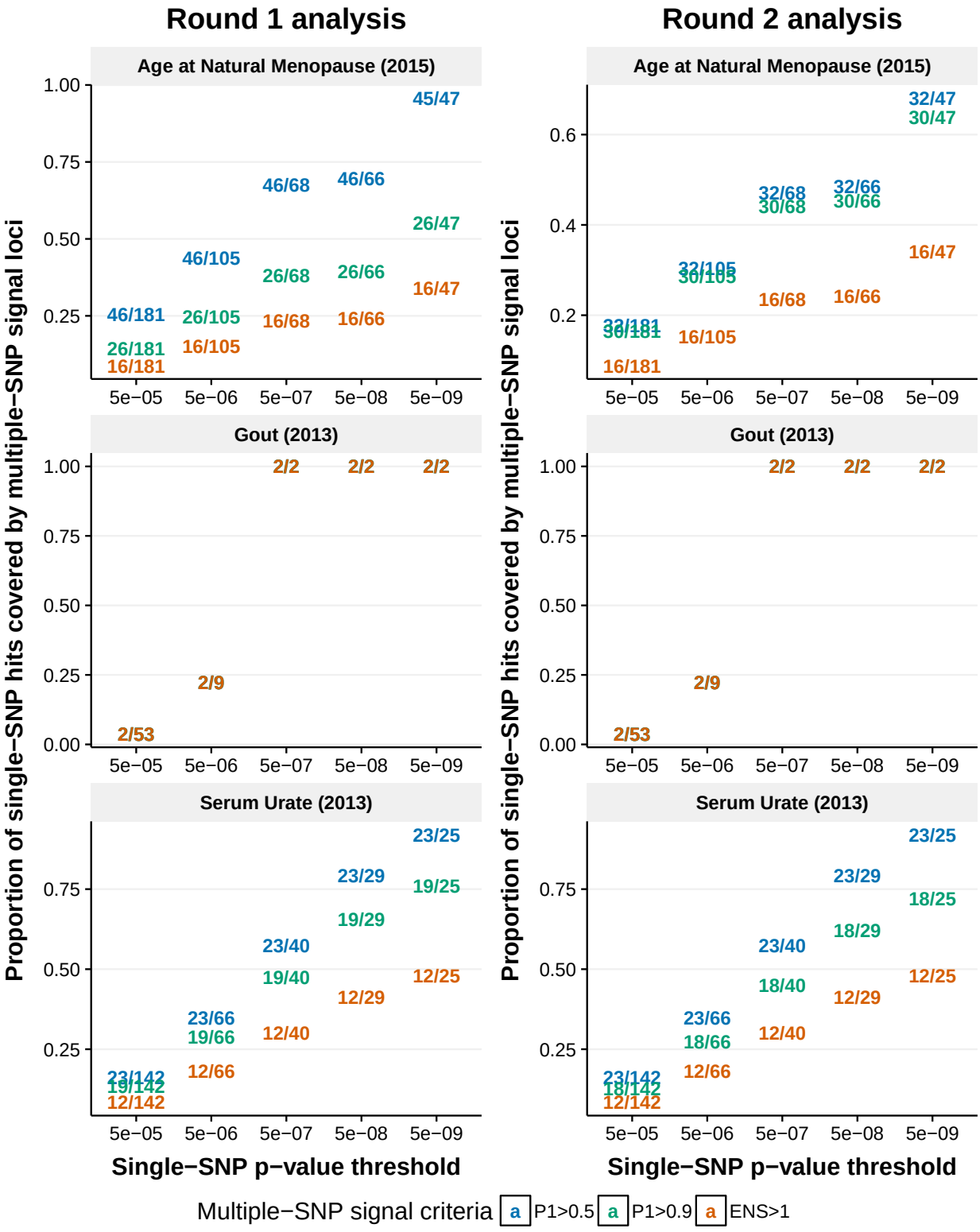
## Insulin-related traits

The three insulin-related traits are fasting glucose, fasting insulin (Manning et al. 2012) and type 2 diabetes (Morris et al. 2012).



## Miscellaneous traits

The three miscellaneous traits are serum urate, gout (Köttgen et al. 2013) and age at natural menopause (Day et al. 2015).



# Supplementary Figure 12

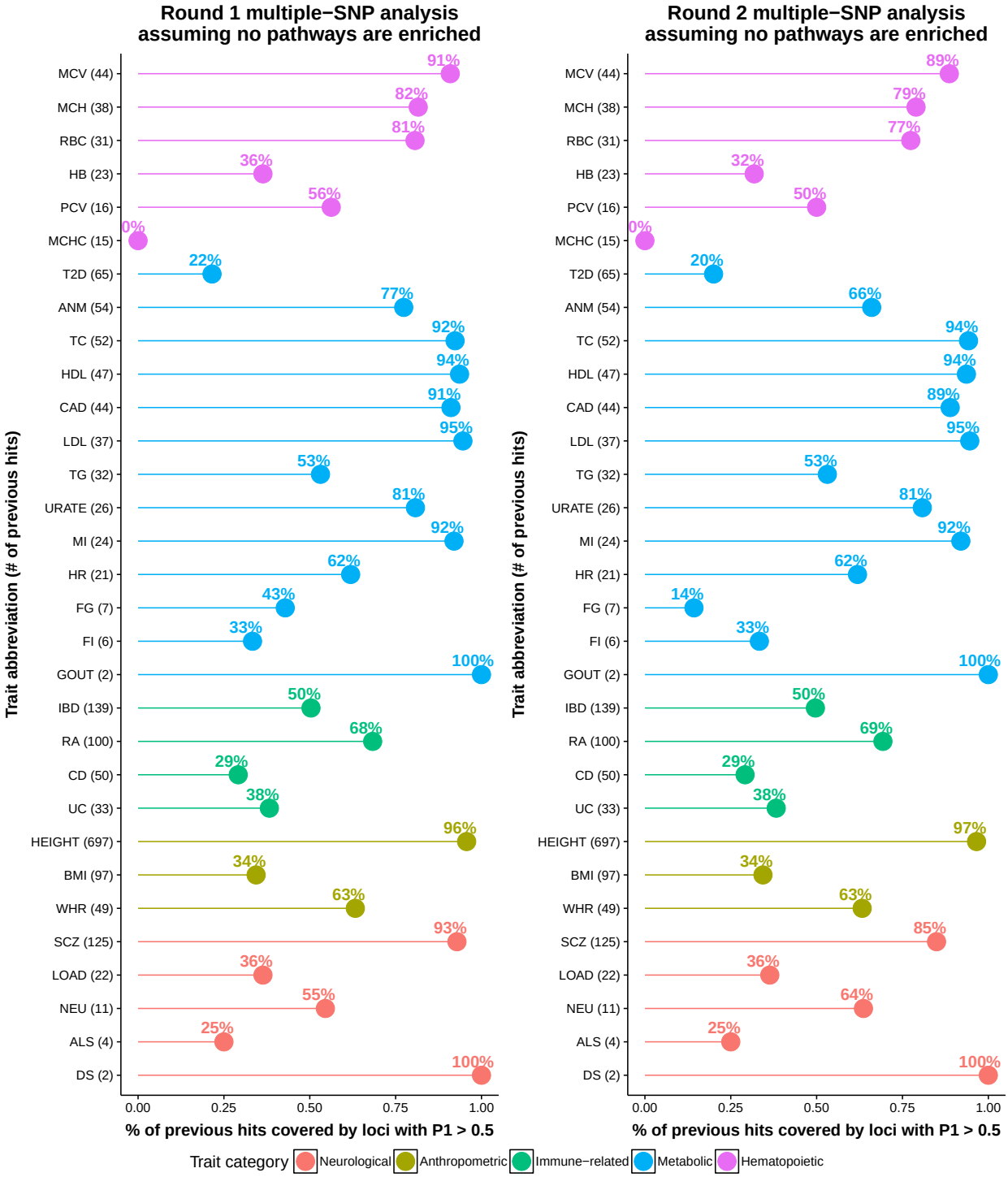
**Proportion of previously-reported genome-wide significant variants that are detected by genome-wide multiple-SNP analyses, assuming that no pathways are enriched.** For each phenotype, we manually extract the genome-wide significant variants (a.k.a. GWAS hits) from corresponding publications (e.g. Tables, Supplementary Files). This is different from **Supplementary Figure 11**, where we derive GWAS hits from the raw summary statistics of the same set of SNPs that were used in our Bayesian multiple-SNP analyses. We call a GWAS hit "detected by multiple-SNP analyses" if this SNP is covered by a predefined locus satisfying certain multiple-SNP association criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS $> 1$). See **Supplementary Figure 10** for the definitions of locus and the multiple-SNP association posterior statistics. For each panel, phenotypes are ordered first by trait category, then the number of reported GWAS hits.

It is important to note that some published GWAS hits of certain traits are not necessarily (genome-wide) significant in the corresponding summary data file. For example, rs34856868 shows $p = 9.80 \times 10^{-9}$ for association with inflammatory bowel disease in Table 2 of Liu et al. (2015); however, the $p$ value of the same SNP is 0.27 in the corresponding summary data file (`EUR.IBD.gwas.assoc.gz`). For this SNP, the result in Table 2 of Liu et al. (2015) was indeed obtained from a combined analysis of data on both GWAS and custom arrays (Immunochip, Cortes and Brown 2011), whereas the result in the summary data file was only based on GWAS arrays. Because of this type of potential discrepancy between summary data files and corresponding publications, the concordance rates shown here are lower than those shown in **Supplementary Figure 11** for certain traits.

Note that **Supplementary Figure 12** is included to highlight the need for careful usage of GWAS summary statistics. To fairly assess the concordance between traditional single-SNP analyses and our multiple-SNP analyses, please **only** use **Supplementary Figure 11**.

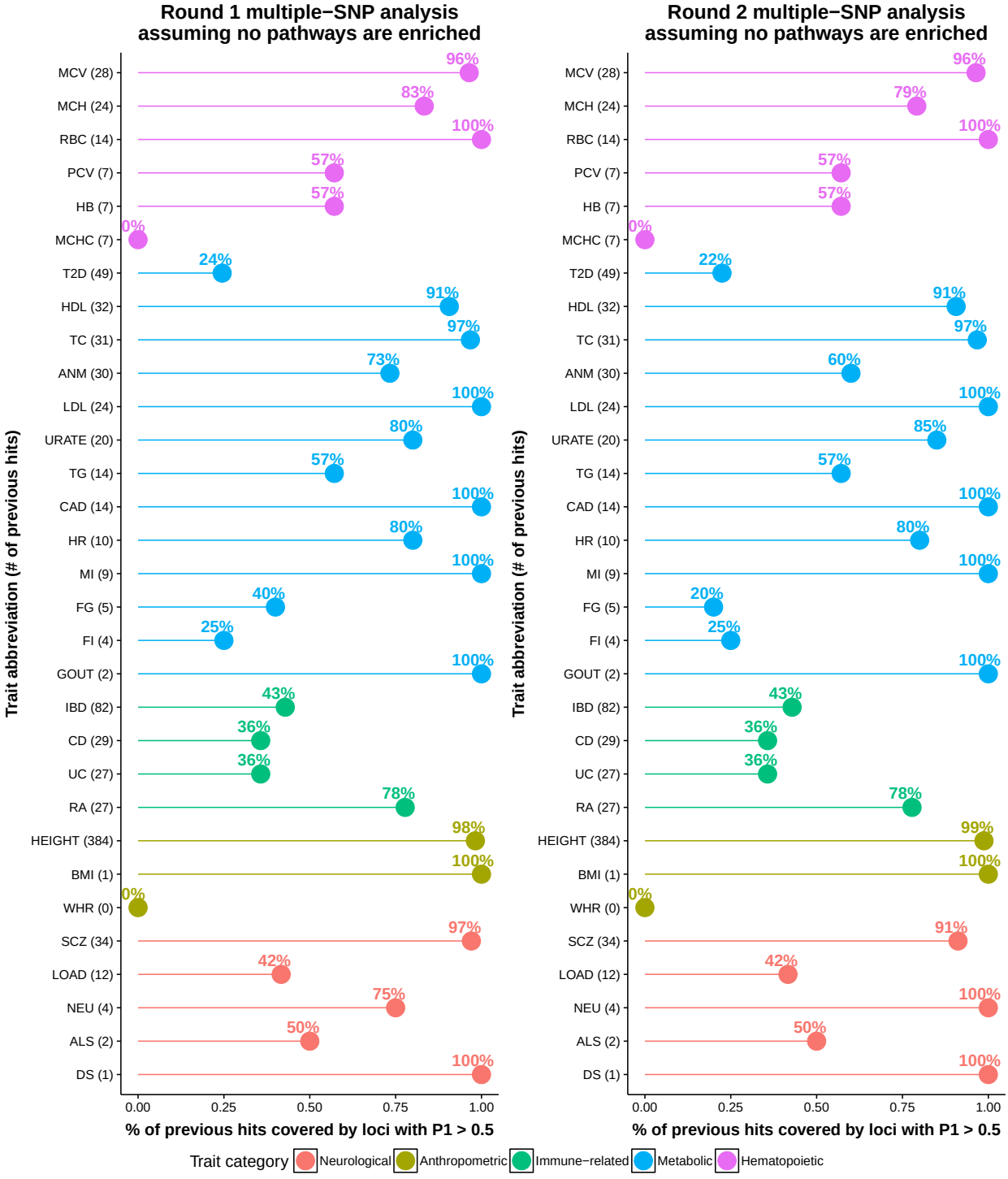
## All hits & $P_1 > 0.5$

Proportion of all previous GWAS hits that are covered by loci with $P_1 > 0.5$.



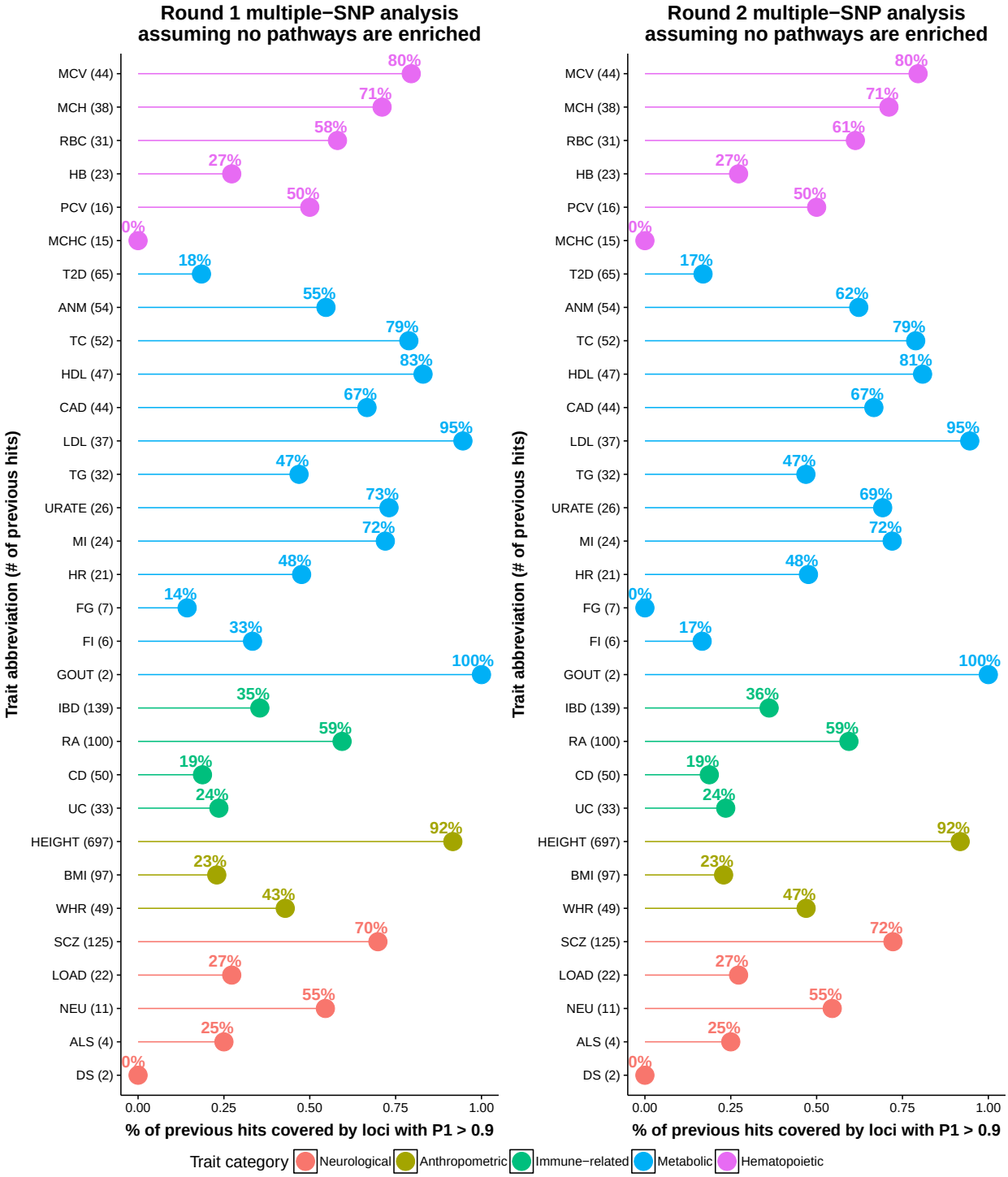
## Analyzed hits & $P_1 > 0.5$

Proportion of previous GWAS hits that are included in genome-wide multiple-SNP analyses and are covered by loci with $P_1 > 0.5$.



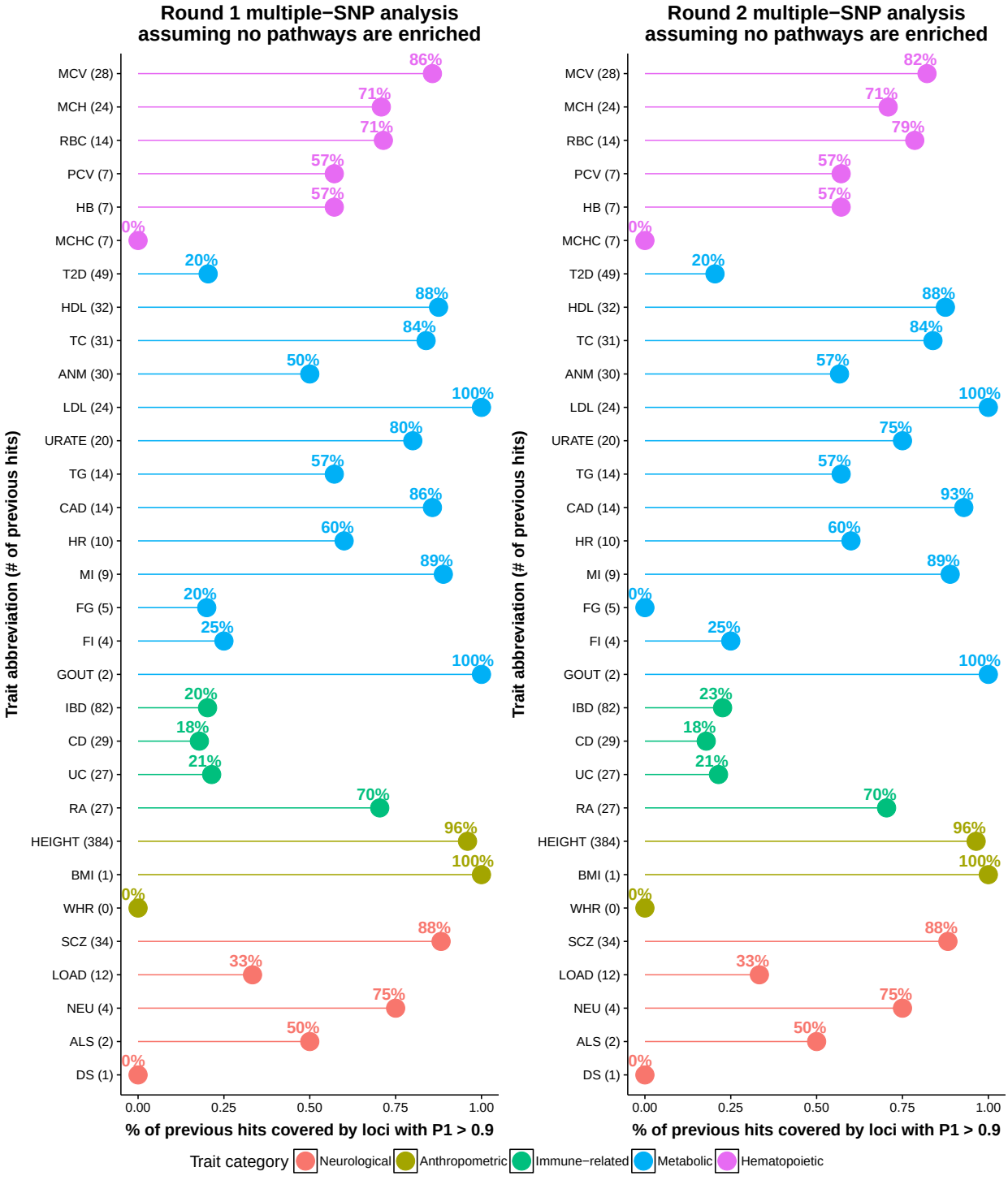
## All hits & $P_1 > 0.9$

Proportion of all previous GWAS hits that are covered by loci with $P_1 > 0.9$.



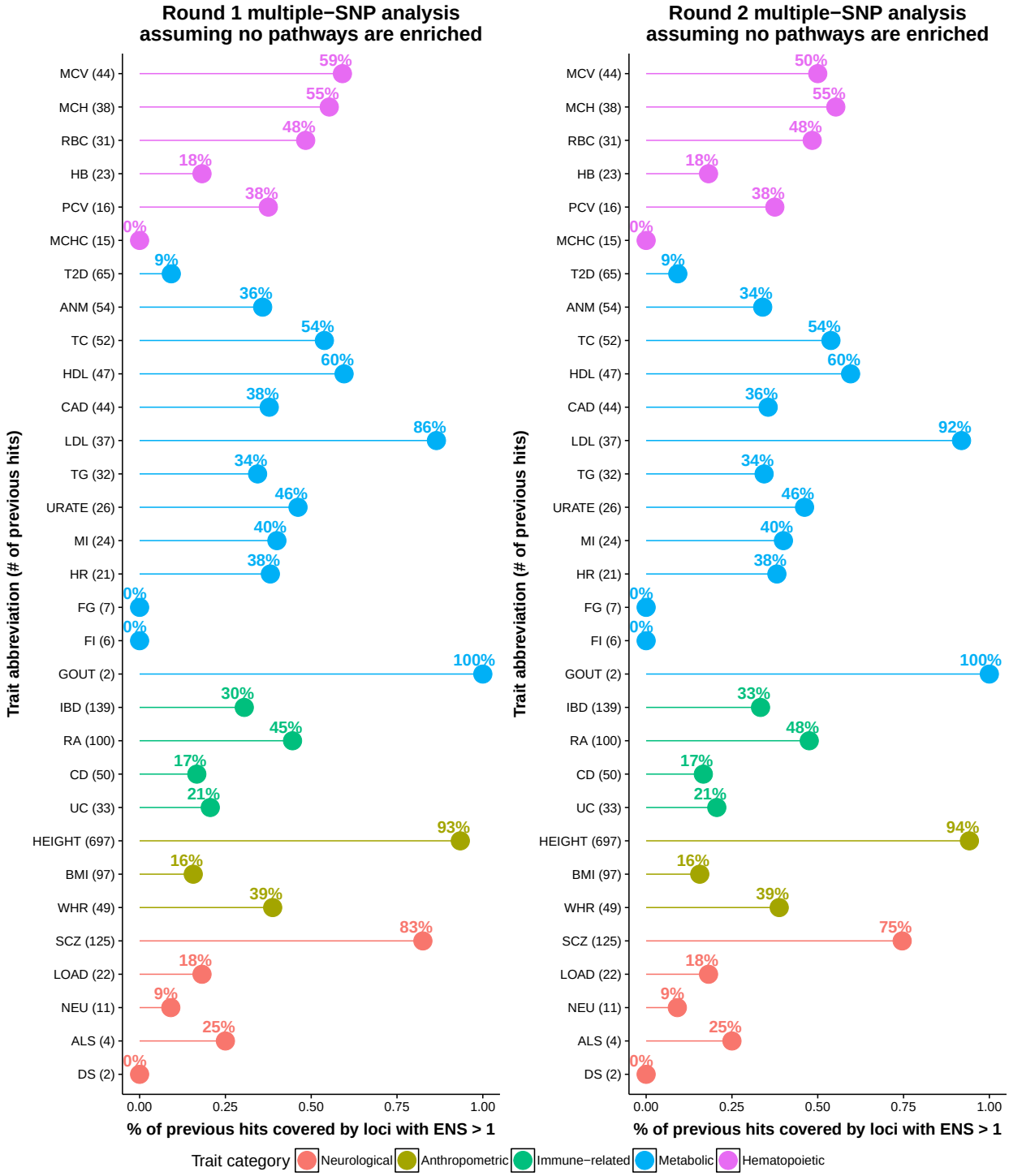
## Analyzed hits & $P_1 > 0.9$

Proportion of all previous GWAS hits that that are included in genome-wide multiple-SNP analyses and are covered by loci with $P_1 > 0.9$.



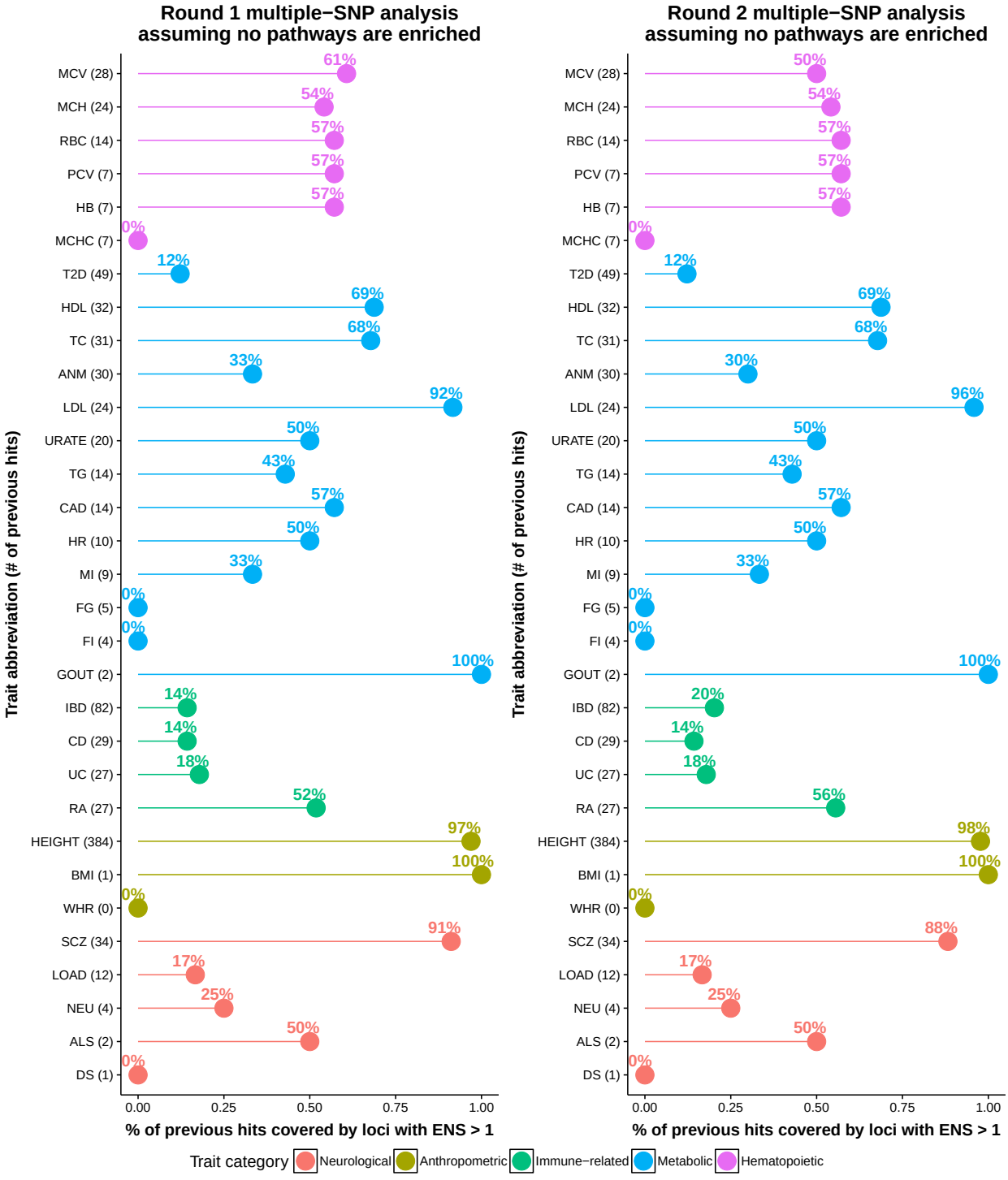
## **All hits & ENS** > 1

Proportion of all previously-reported GWAS hits that are covered by loci with ENS > 1.



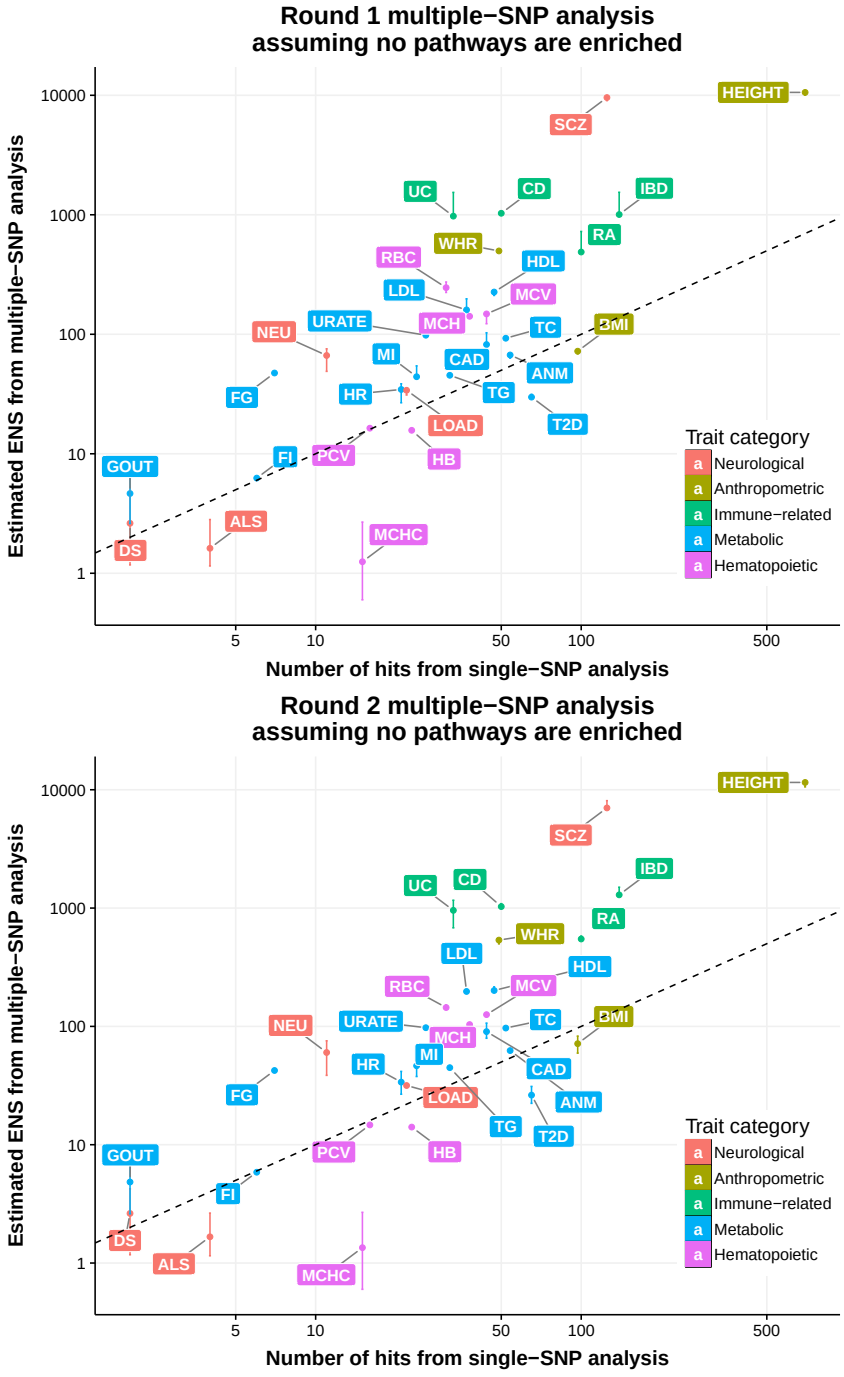
## **All hits & ENS** > 1

Proportion of previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses and are covered by loci with ENS > 1.



# Supplementary Figure 13

**Compare the number of signals from genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched.** Both axes use a logarithmic scale (base 10). Dashed lines are reference lines with intercept zero and slope one. Each point range along $y$-axis denotes the posterior mean and 95% credible interval of the posterior expected total number of trait-associated SNPs (ENS). See **Supplementary Note** for the detail of computing this quantity. As **Supplementary Figure 12**, here the "hits from single-SNP analysis" are the genome-wide significant SNPs reported in corresponding publications.

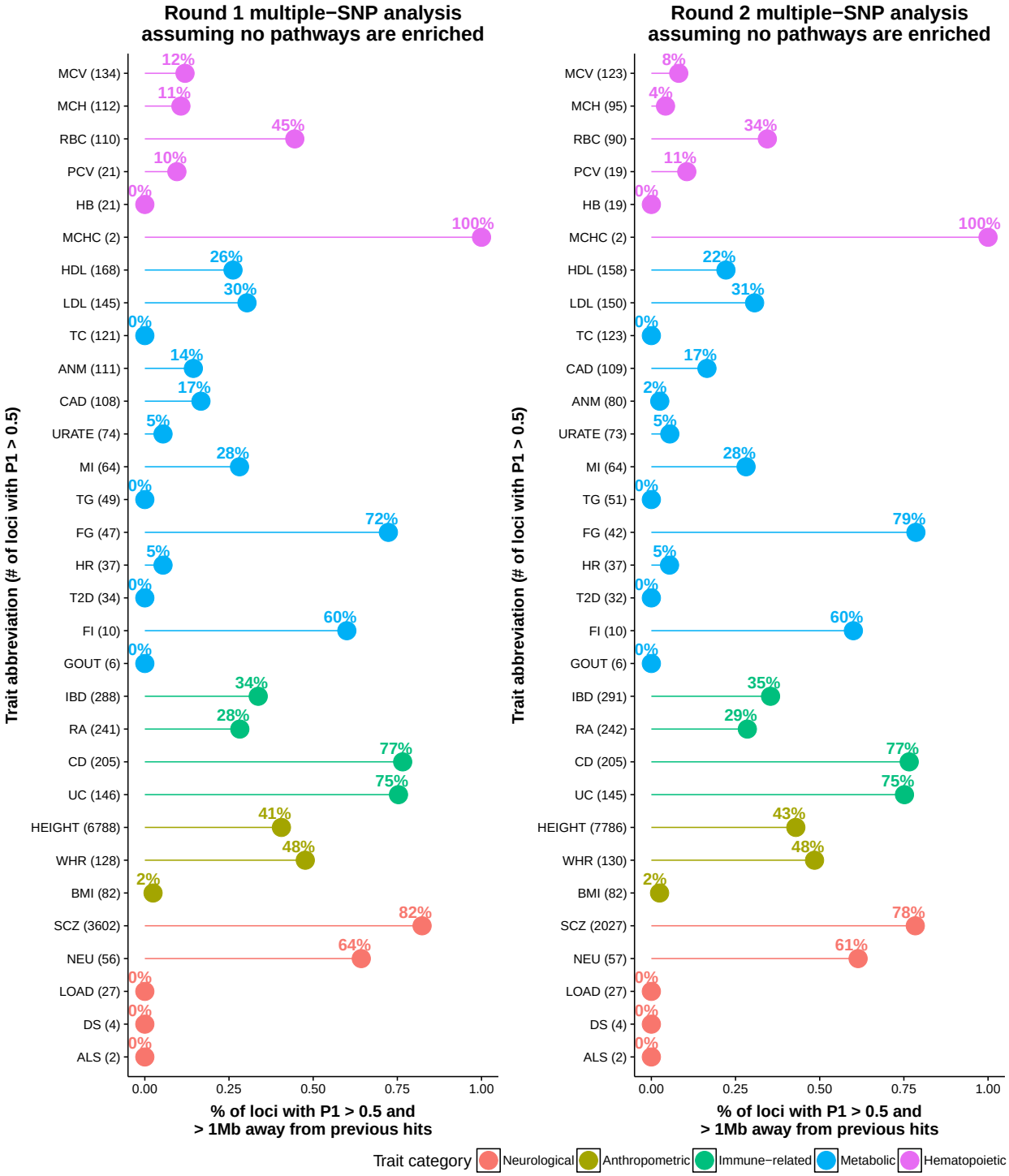


## Supplementary Figure 14

**Proportion of loci identified by genome-wide multiple-SNP analyses that are at least 1 Mb away from previously-reported GWAS hits, assuming that no pathways are enriched.** We call a predefined locus "detected by multiple-SNP analyses" if the locus satisfies certain multiple-SNP association criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS $> 1$). See **Supplementary Figure 10** for the definition of locus and the multiple-SNP association posterior statistics. See **Supplementary Figure 12** for the definition of "previously-reported GWAS hits". For each panel, phenotypes are ordered first by category, then the number of loci identified from multiple-SNP analyses.

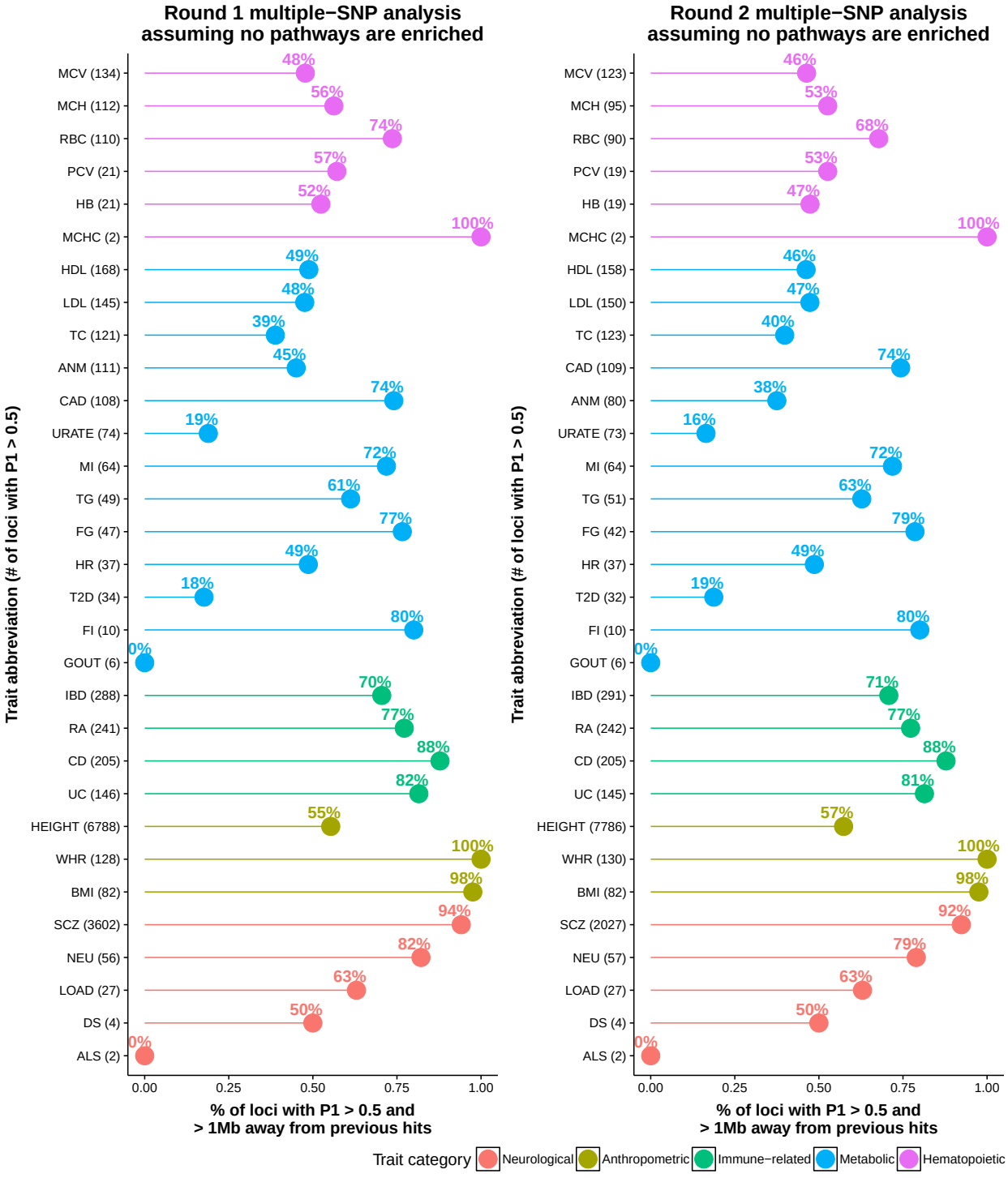
## All hits & $P_1 > 0.5$

Proportion of loci with $P_1 > 0.5$ that are at least 1 Mb away from all previously-reported GWAS hits.



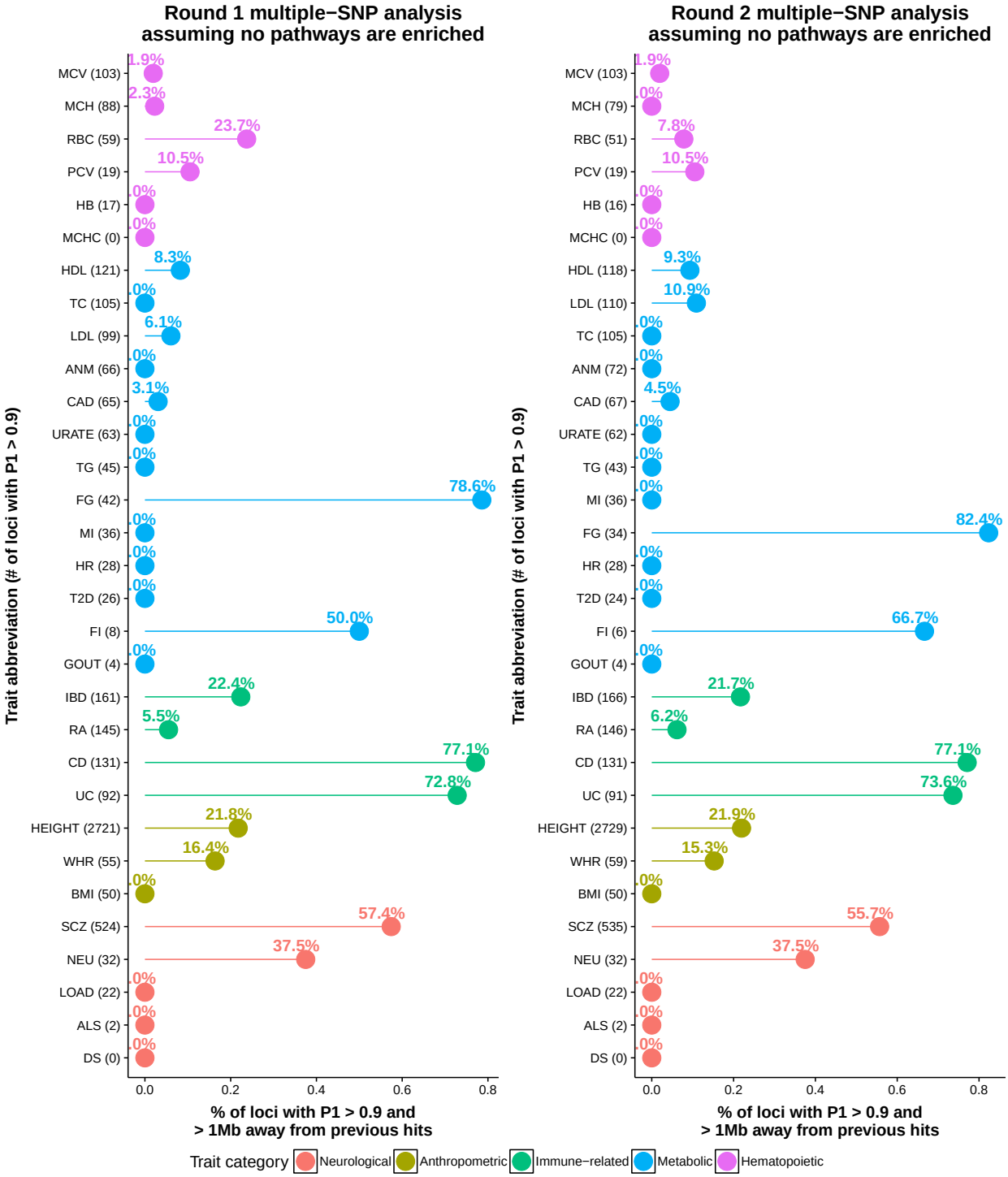
## Analysed hits & $P_1 > 0.5$

Proportion of loci with $P_1 > 0.5$ that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.



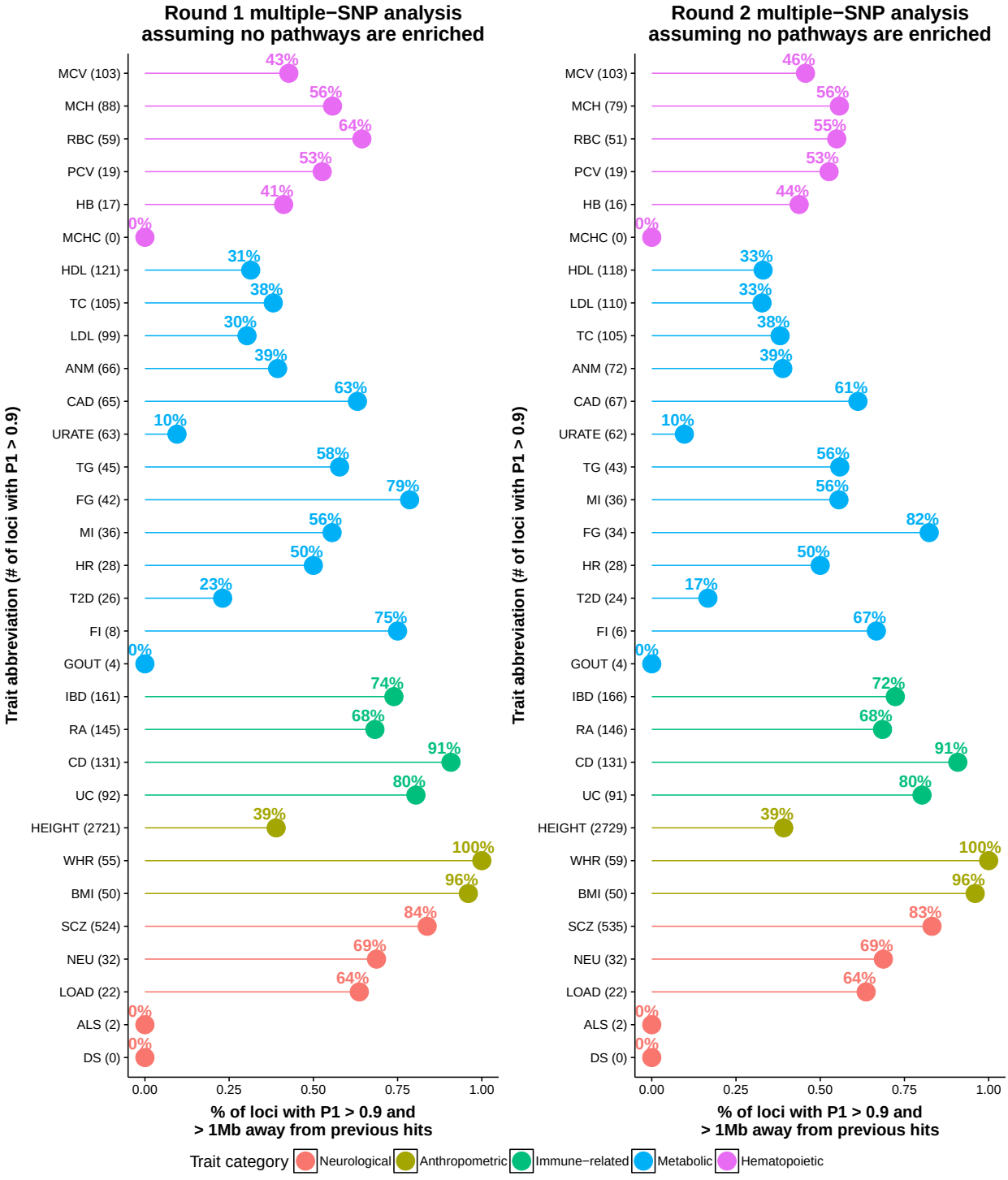
## All hits & $P_1 > 0.9$

Proportion of loci with $P_1 > 0.9$ that are at least 1 Mb away from all previously-reported GWAS hits.



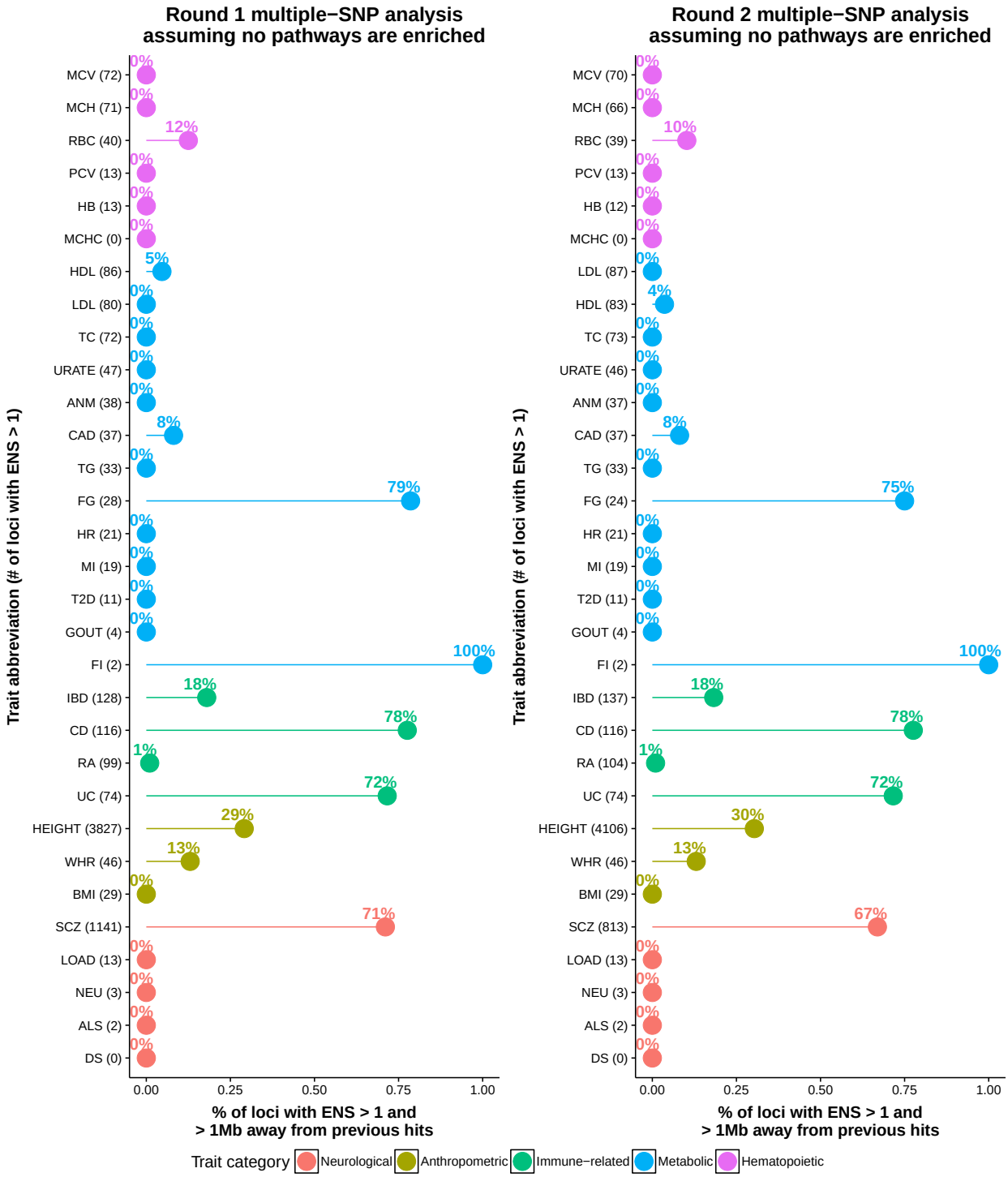
## Analyzed hits & $P_1 > 0.9$

Proportion of loci with $P_1 > 0.9$ that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.



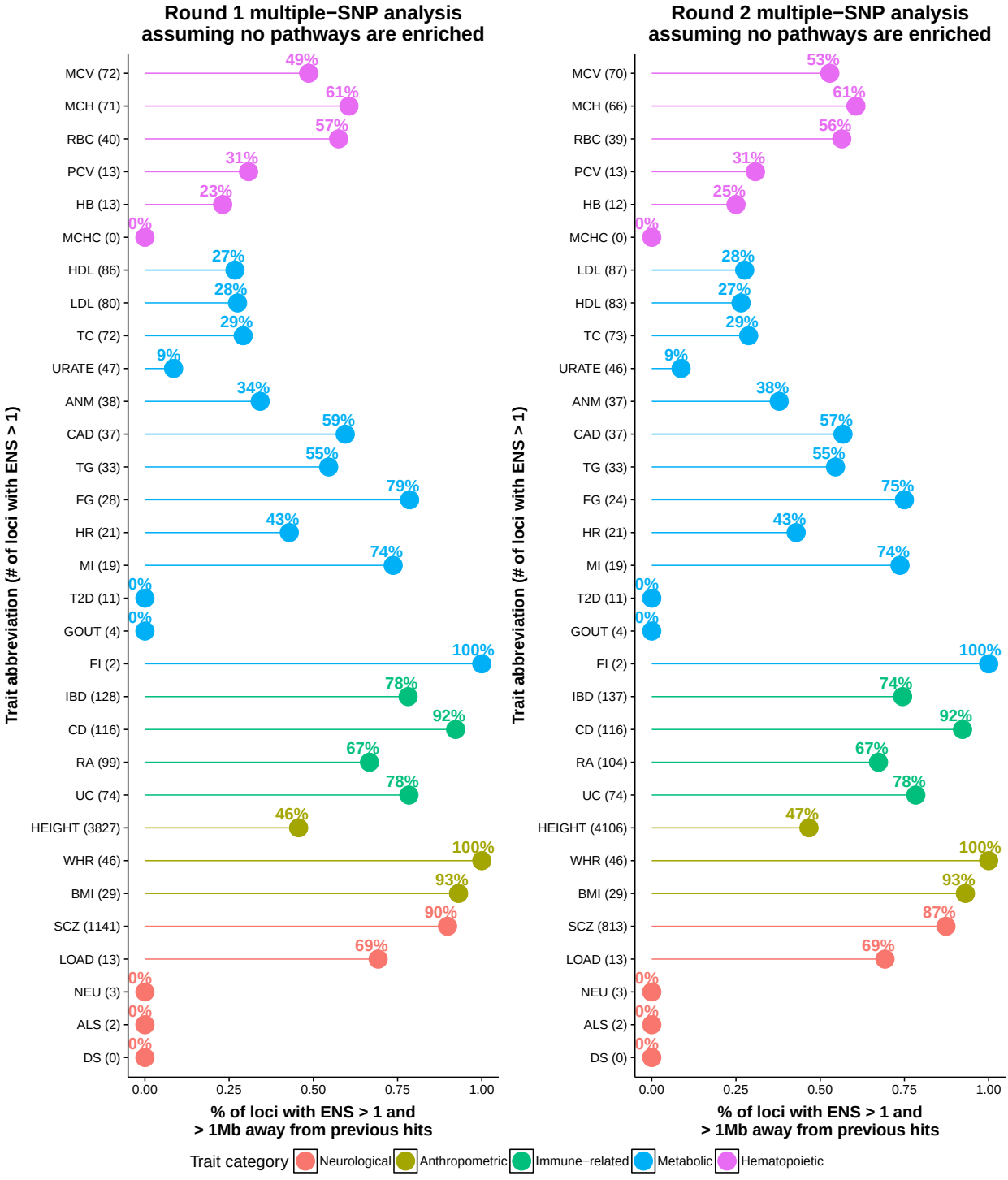
## **All hits & ENS** > 1

Proportion of loci with ENS > 1 that are at least 1 Mb away from all previously-reported GWAS hits.



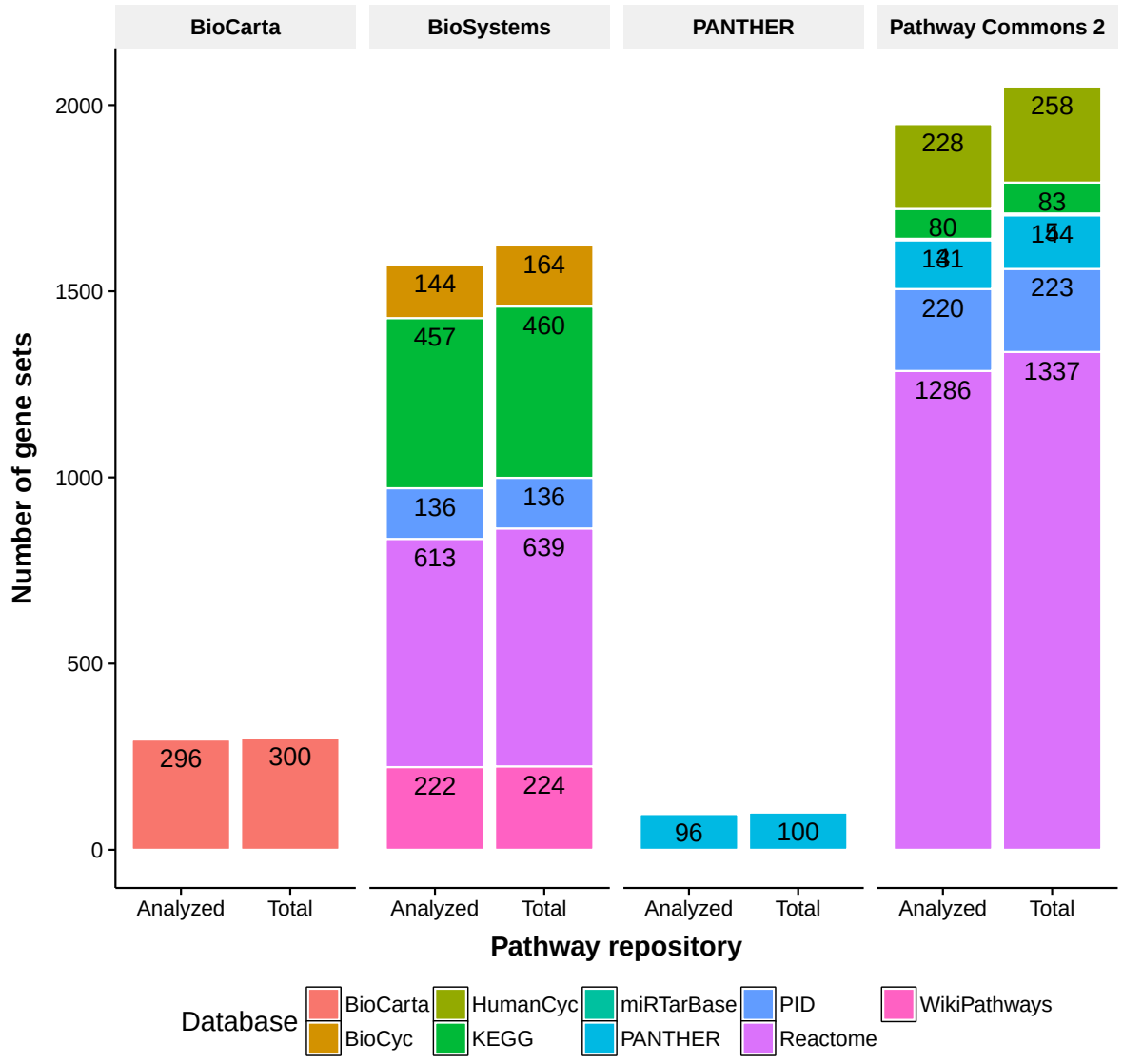
## Analyzed hits & ENS > 1

Proportion of loci with ENS > 1 that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.



## Supplementary Figure 15

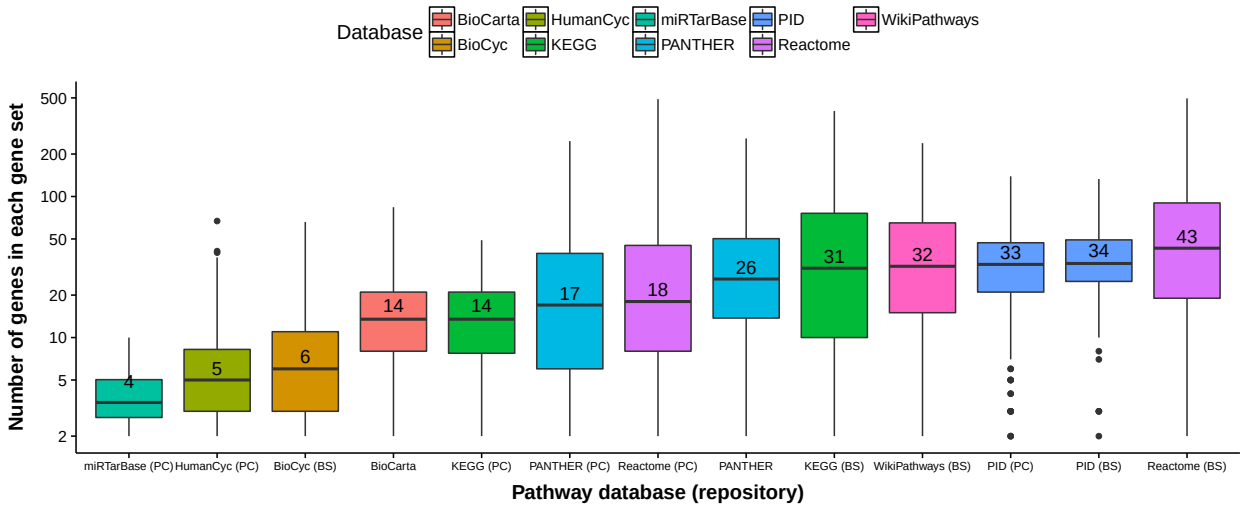
**Summary of biological pathways.** Biological pathway definitions are retrieved from the Pathway Commons 2 (Cerami et al. 2011, version 7), NCBI Biosystems (Geer et al. 2010), PANTHER (Mi and Thomas 2009, version 3.3) and BioCarta used in Carbonetto and Stephens (2013). The Pathway Commons 2 database includes gene sets derived from Reactome (Croft et al. 2014), Nature Pathway Interaction Database (PID, Schaefer et al. 2009), HumanCyc (Romero et al. 2004), PANTHER, miRTarBase (Hsu et al. 2014) and Kyoto Encyclopedia of Genes and Genomes (KEGG, Wrzodek et al. 2013) pathways. The NCBI BioSystem database contains pathways from KEGG, BioCyc (Caspi et al. 2014), PID, Reactome and WikiPathways (Pico et al. 2008). See Supporting Information Text S1 of Carbonetto and Stephens (2013) for the rationale for choosing these pathway databases (`https://doi.org/10.1371/journal.pgen.1003770.s015`). For each panel, the bar chart on the right side (labelled as "Total") shows the total number of pathways retrieved from the corresponding databases, and the one on the left side (labelled as "Analyzed") shows the number of pathways used in the present enrichment analyses.



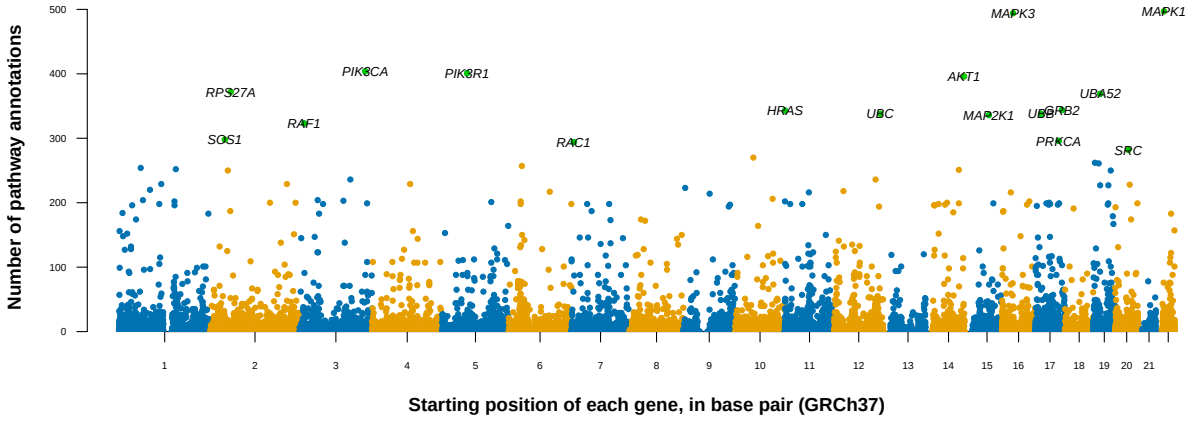
# Supplementary Figure 16

**Summary of genes.** Genomic definitions for genes are derived from *Homo sapiens* reference genome GRCh37. In the present study, we consider 18,313 autosomal protein-coding genes that are mapped to the reference sequence.

**(a)** Distributions of gene set sizes for each pathway database-repository combination. Combinations are ordered by median numbers of genes in pathways, which are displayed in each box plot. The vertical axis uses a logarithmic scale (base: 10). PC: Pathway Commons. BS: NCBI BioSystems.




**(b)** Manhattan plot of the number of pathway annotations for each gene. The highlighted genes (colored in green and labelled by their HGNC symbols) belong to more than 270 of 3,913 analyzed biological pathways. The Manhattan plot is produced by the package `qqman` (Turner 2014).



## Supplementary Figure 17

**Sanity checks of top-ranked gene set enrichments for 31 phenotypes.** To quickly evaluate whether the strong enrichments identified in our model-based analyses can possibly be true, we develop two sanity checks.

The first sanity check is an "eyeball test" that visualizes the distribution of GWAS single-SNP $z$-scores for a target trait, stratified by SNP-level annotations of a target gene set. Specifically, we plot two estimated density curves for each pair of trait and gene set:

- a **solid red curve** estimated from $z$-scores of SNPs within ± 100 kb of the transcribed region of a gene in the gene set ("inside gene set" SNPs);
- a **dashed black curve** estimated from $z$-scores of remaining SNPs ("outside gene set" SNPs).

For a typical pair of trait and gene set that is deemed to pass the "eyeball test", its dashed black curve is often more "spiky" at zero, and its solid red curve is more spread out. The density curves are produced by the function `geom_density` (default setting) in the package `ggplot2` (Wickham 2009).

The second sanity check computes a likelihood ratio (LR) for the following two models:

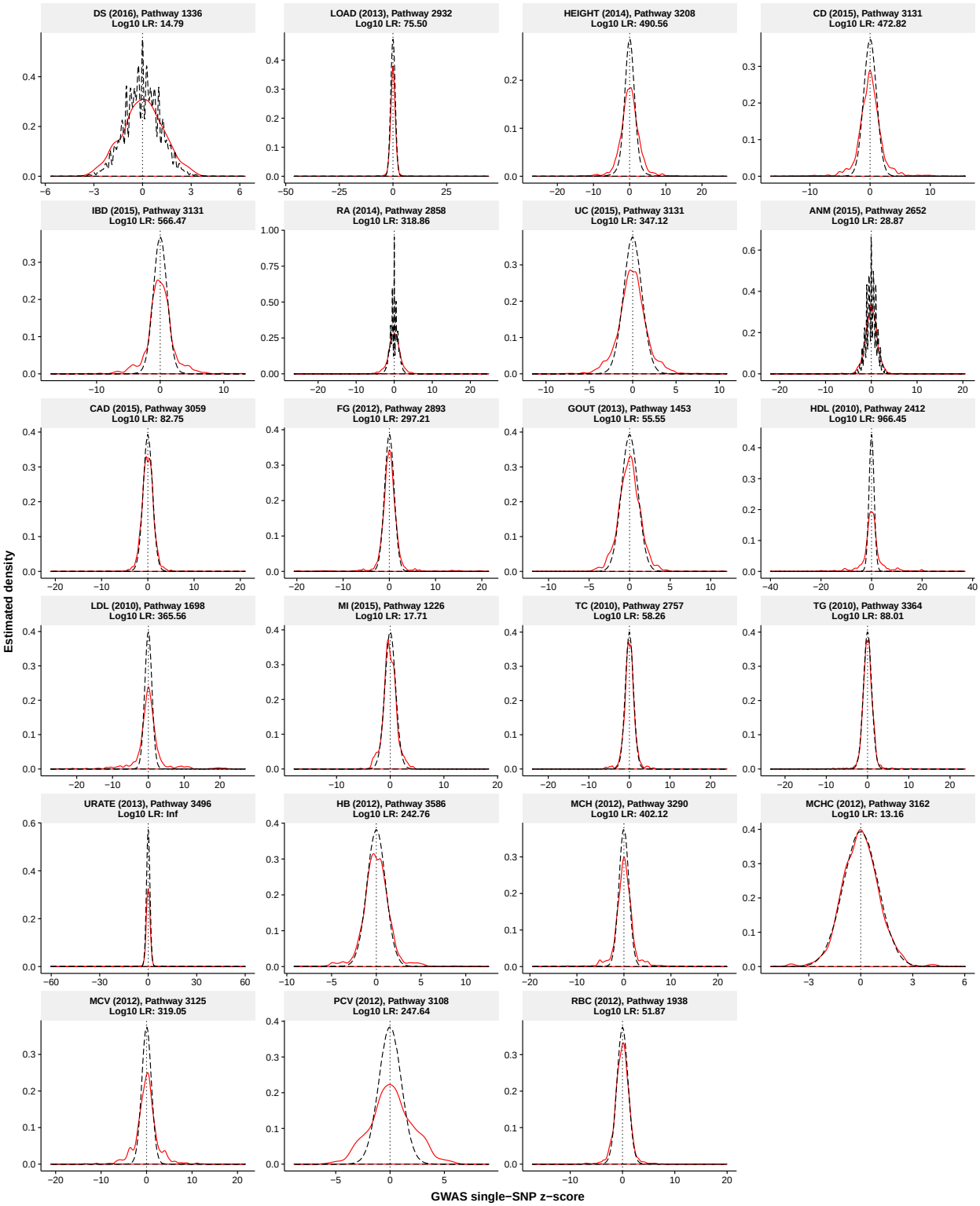
- **baseline model** (`a2`): "inside gene set" SNPs have the same effect size distribution as "outside gene set" SNPs, which can be estimated by `a1$fitted_g` based on the (merged) whole genome data;
- **enrichment model** (`a3`): "inside gene set" and "outside gene set" SNPs have different effect size distributions, which should be estimated separately.

For a strongly enriched gene set, its LR value tends to be very large, since the data should favor the enrichment (`a3`) over the baseline model (`a2`). The second check based on LR computation complements the first visual check in cases where the "eyeball test" results are not clearly visible. The LR calculation is based on the package `ashr` (Stephens 2017). Below are some `R` codes that illustrate the LR calculation.

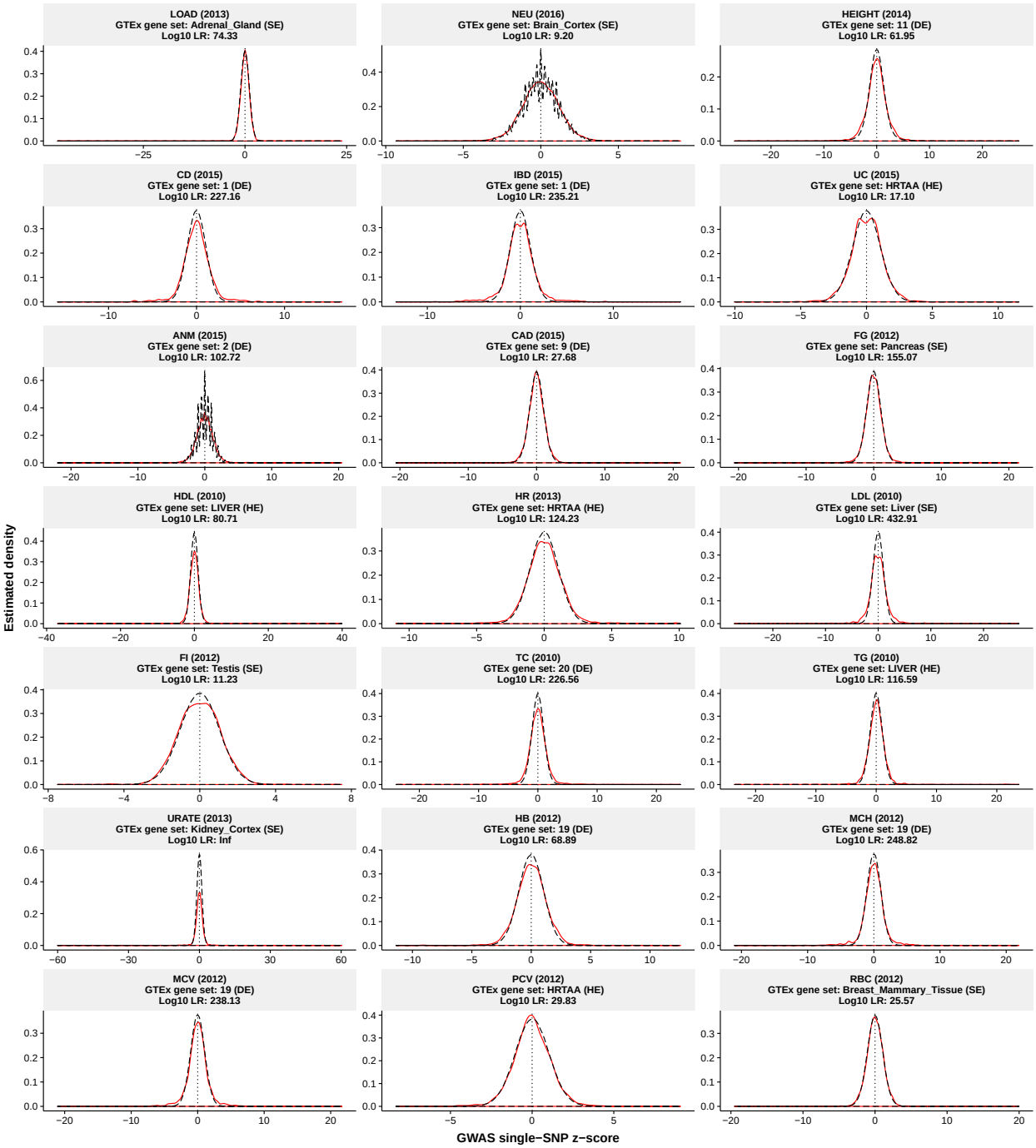
```
suppressPackageStartupMessages(library(ashr))

# load GWAS summary statistics and SNP-level annotations
betahat <- c(sumstat$betahat)
se <- c(sumstat$se)
snps <- c(sumstat$snps)

# analyze summary data of the whole genome
a1 <- ashr::ash(betahat=betahat, sebetahat=se,
                mixcompdist="halfuniform", method="shrink")

# analyze summary data of SNPs that are "inside gene set"
# where the prior is estimated from data
a2 <- ashr::ash(betahat=betahat[snps], sebetahat=se[snps],
                mixcompdist="halfuniform", method="shrink")

# analyze summary data of SNPs that are "inside gene set"
# where the prior is fixed as the one estimated in a1
a3 <- ashr::ash(betahat=betahat[snps], sebetahat=se[snps],
                mixcompdist="halfuniform", method="shrink", fixg=T, g=a1$fitted_g)

# compute log10 likelihood ratio statistics
log10LR <- (a2$logLR - a3$logLR) / log(10)
```

We first check the trait-pathway pairs reported in **Table 1**, and then check the trait-tissue pairs reported in **Table 2**. We also check the top 10 most enriched pathways with at least 10 member genes for each of the 31 traits in our Round 2 analyses (see **Supplementary Table 7** for details). Full information about these top enriched pathways and tissue-based gene sets is available at `http://xiangzhu.github.io/rss-gsea/`.

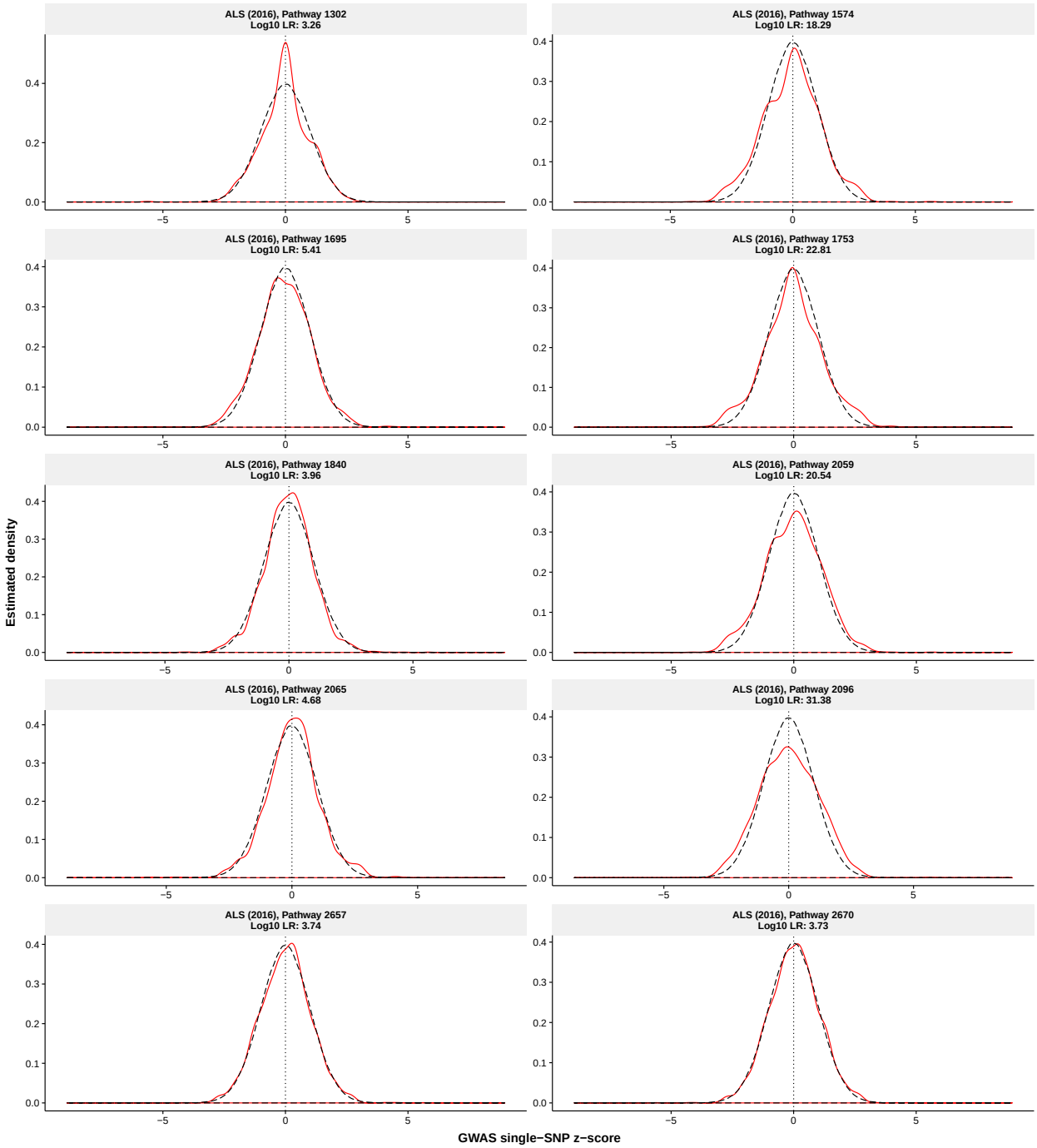
## Pathway-trait pairs reported in Table 1



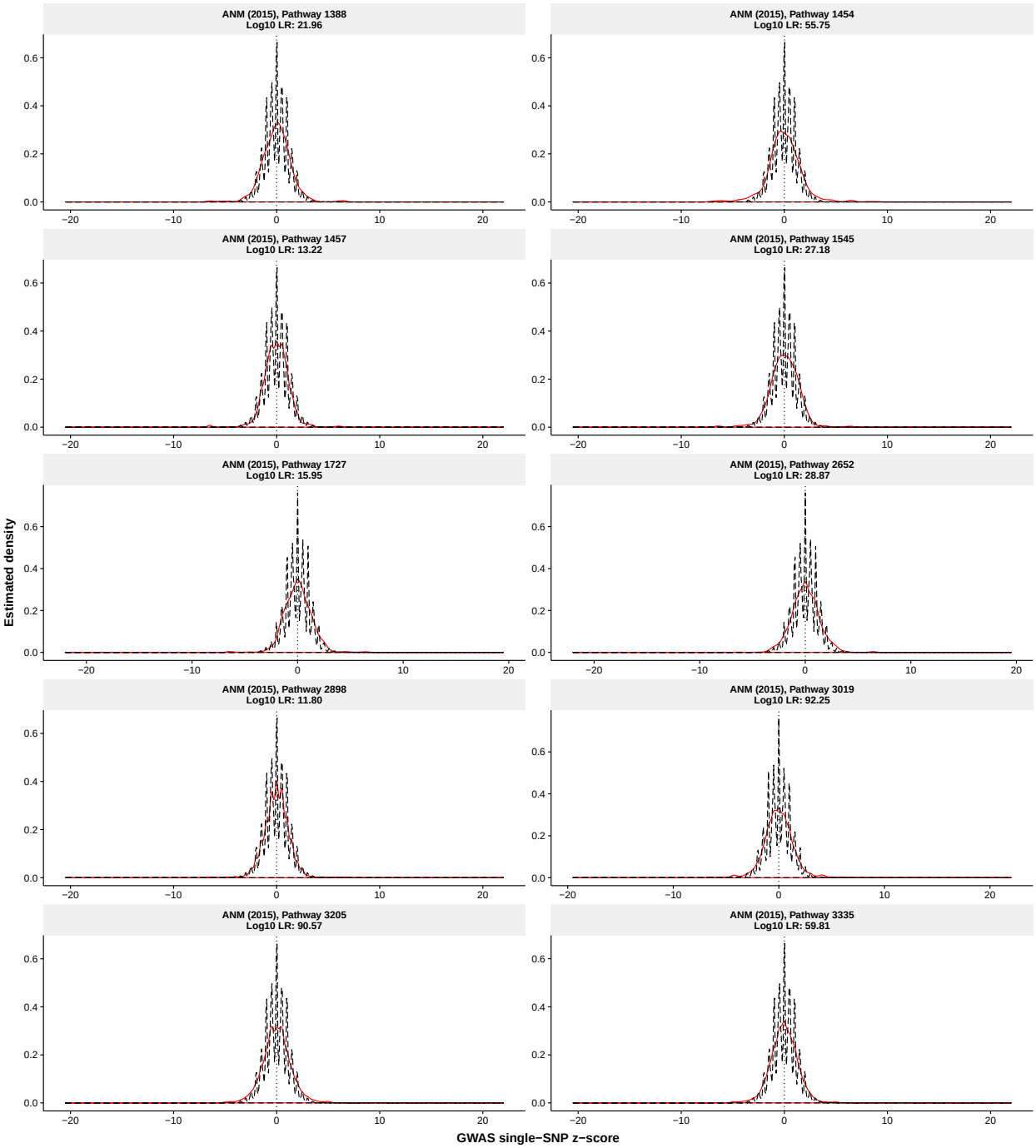
## Tissue-trait pairs reported in Table 2



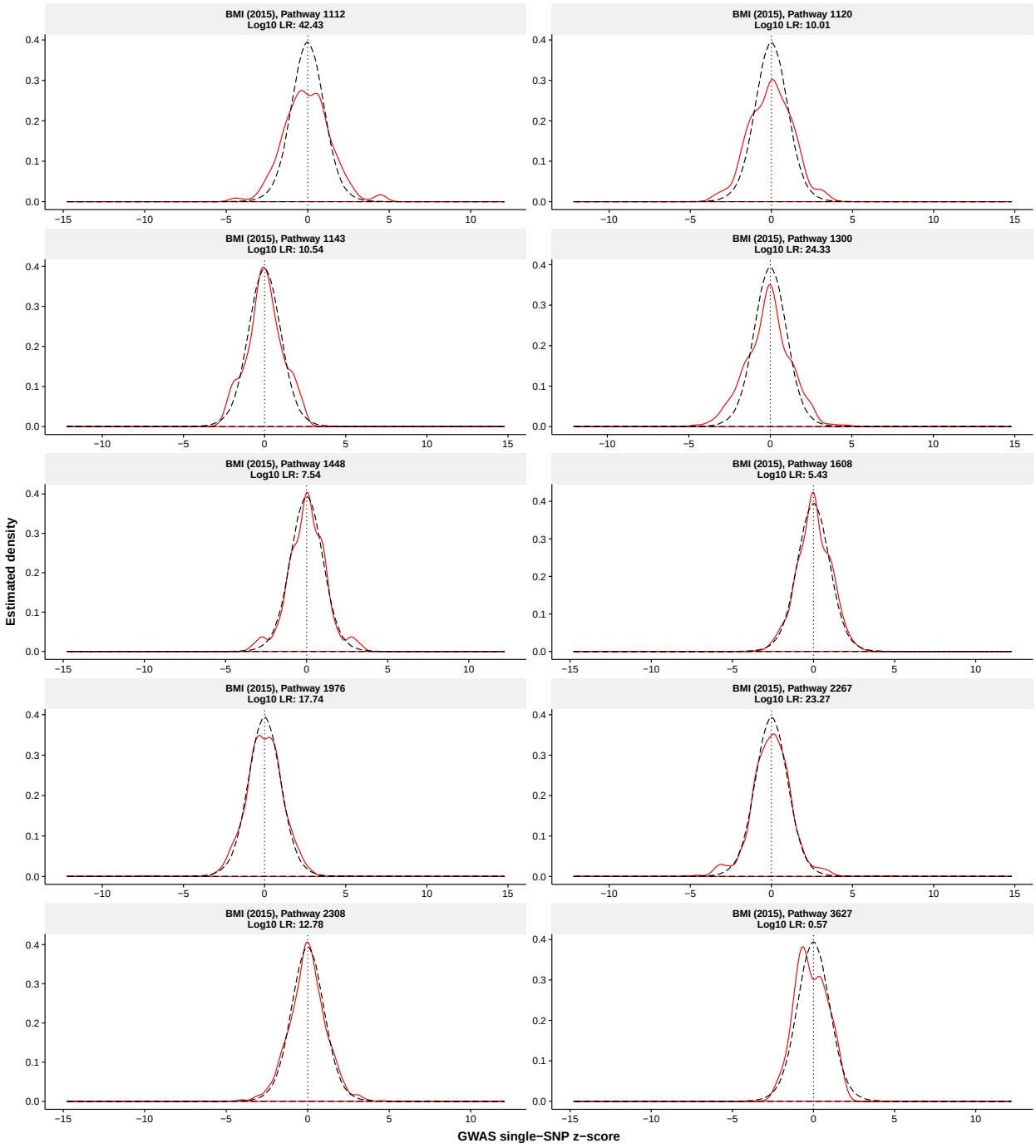
## Amyotrophic lateral sclerosis (Rheenen et al. 2016)



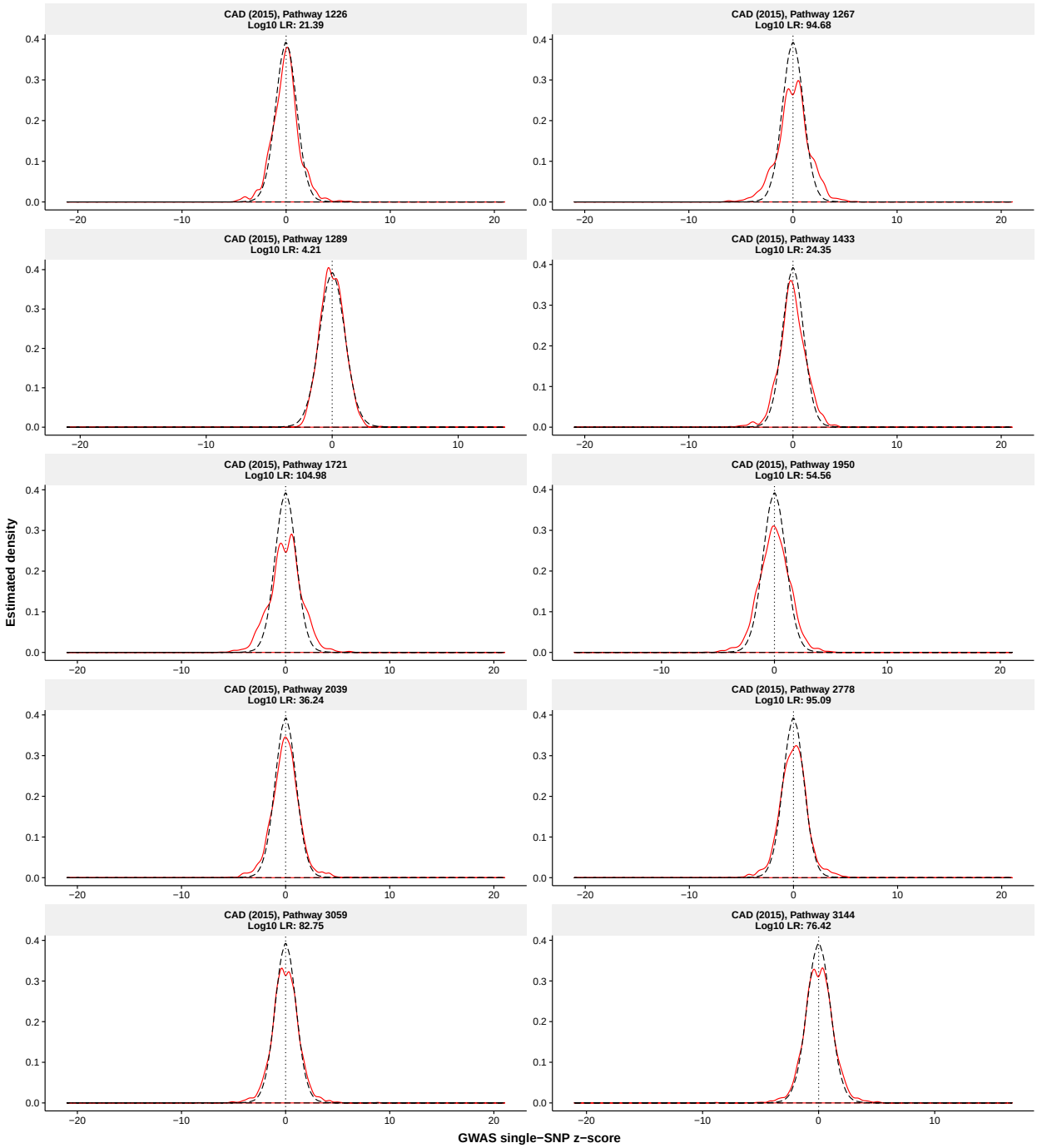
## Age at natural menopause (Day et al. 2015)



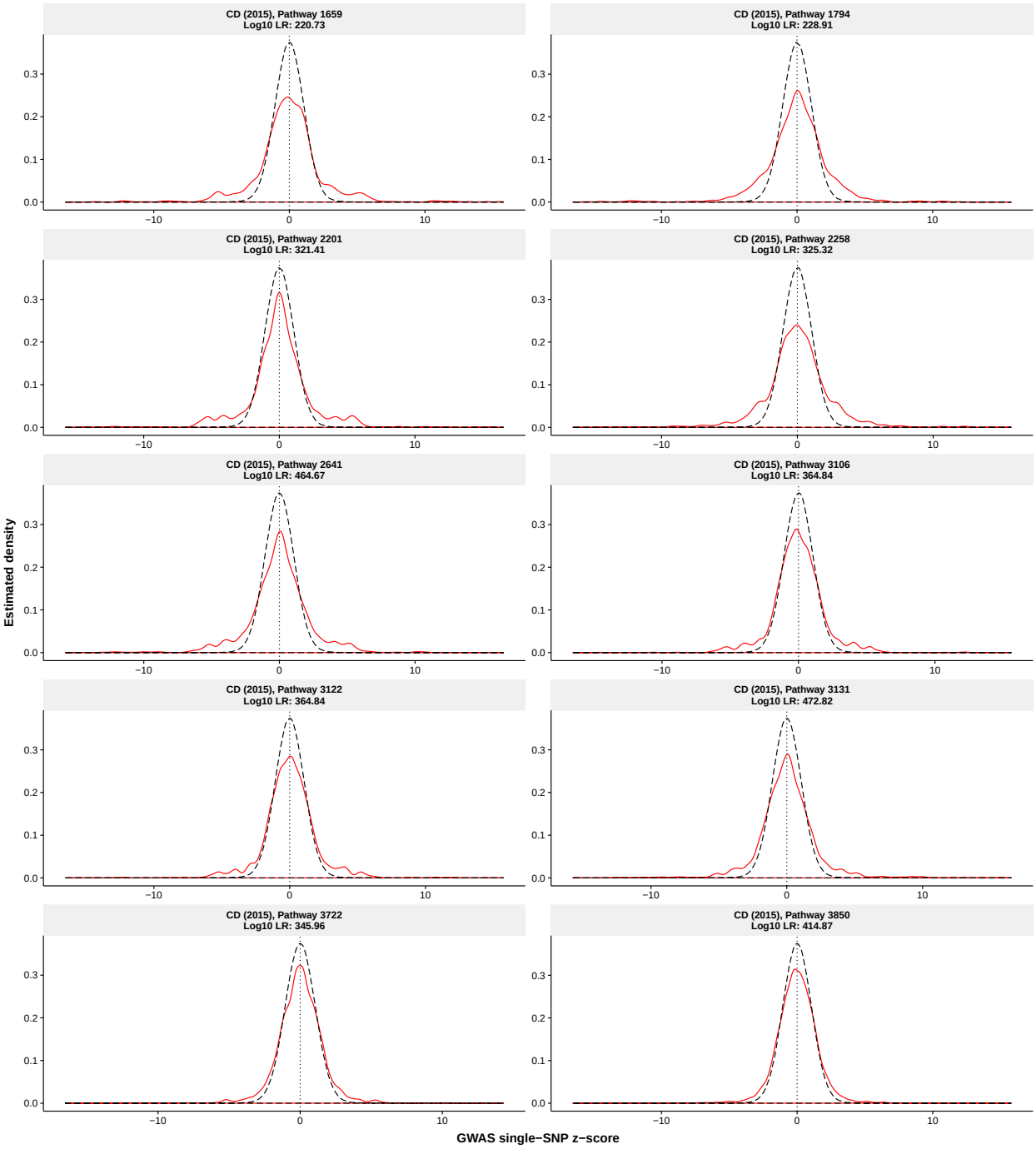
## Body mass index (Locke et al. 2015)



## Coronary artery disease (Nikpay et al. 2015)

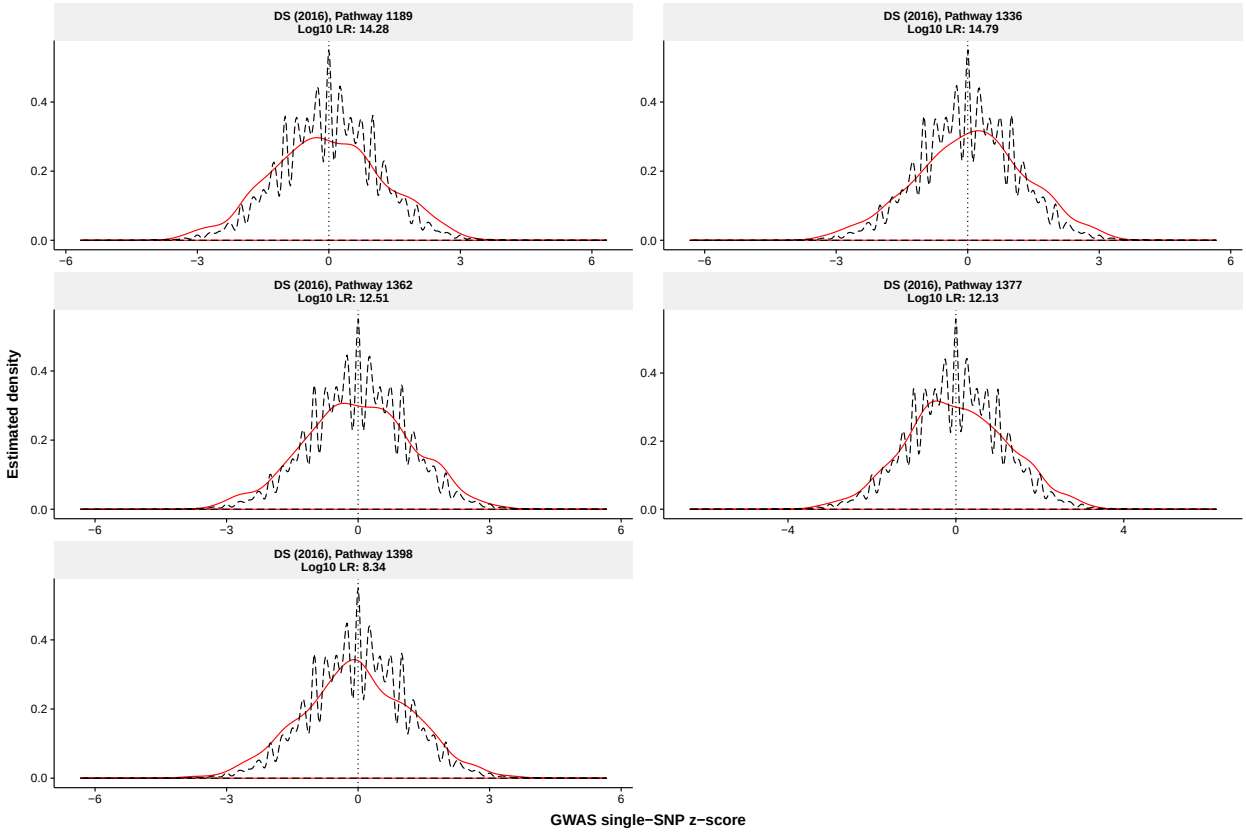


## Crohn's disease (Liu et al. 2015)

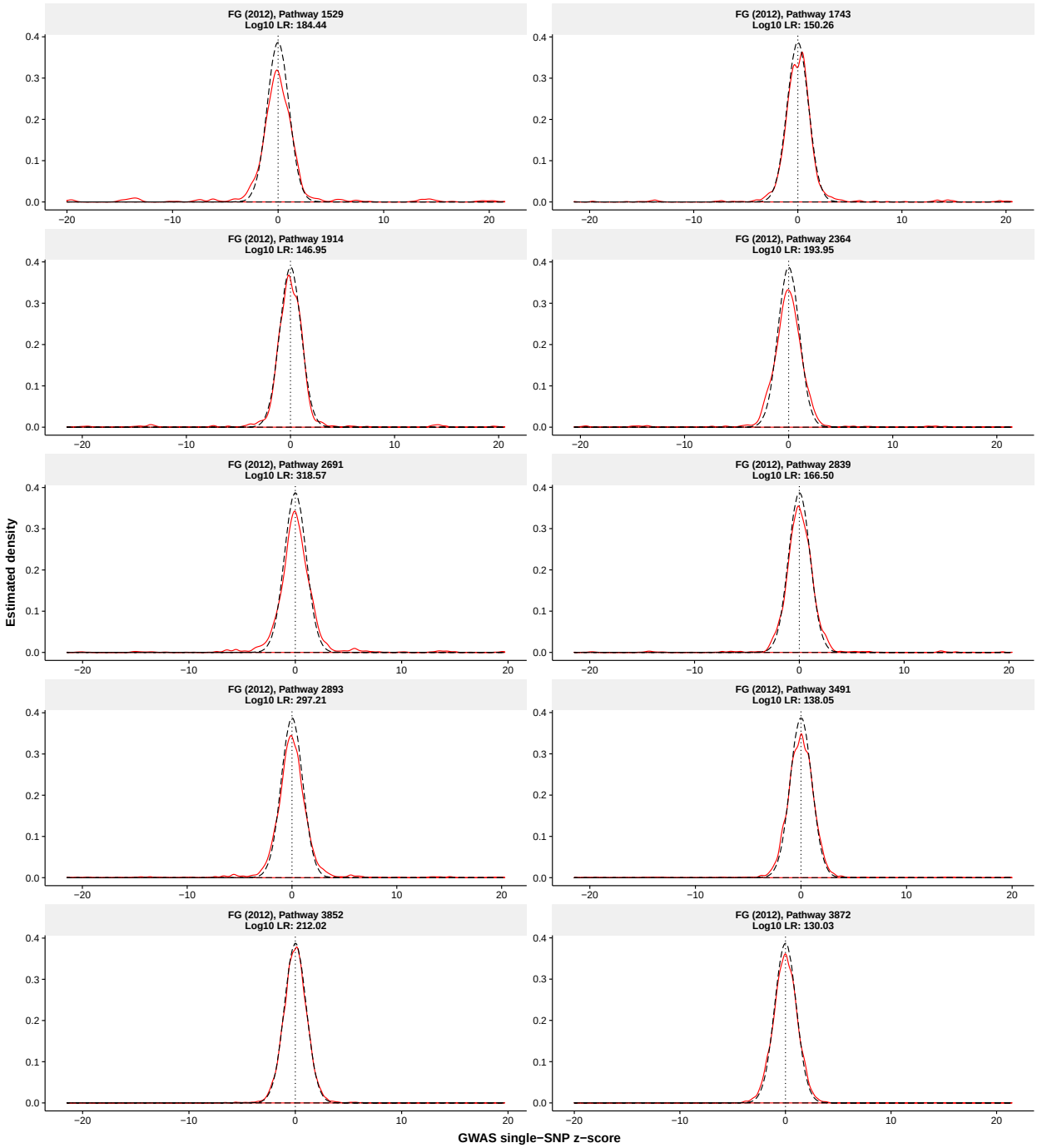


## Depressive symptoms (Okbay et al. 2016)

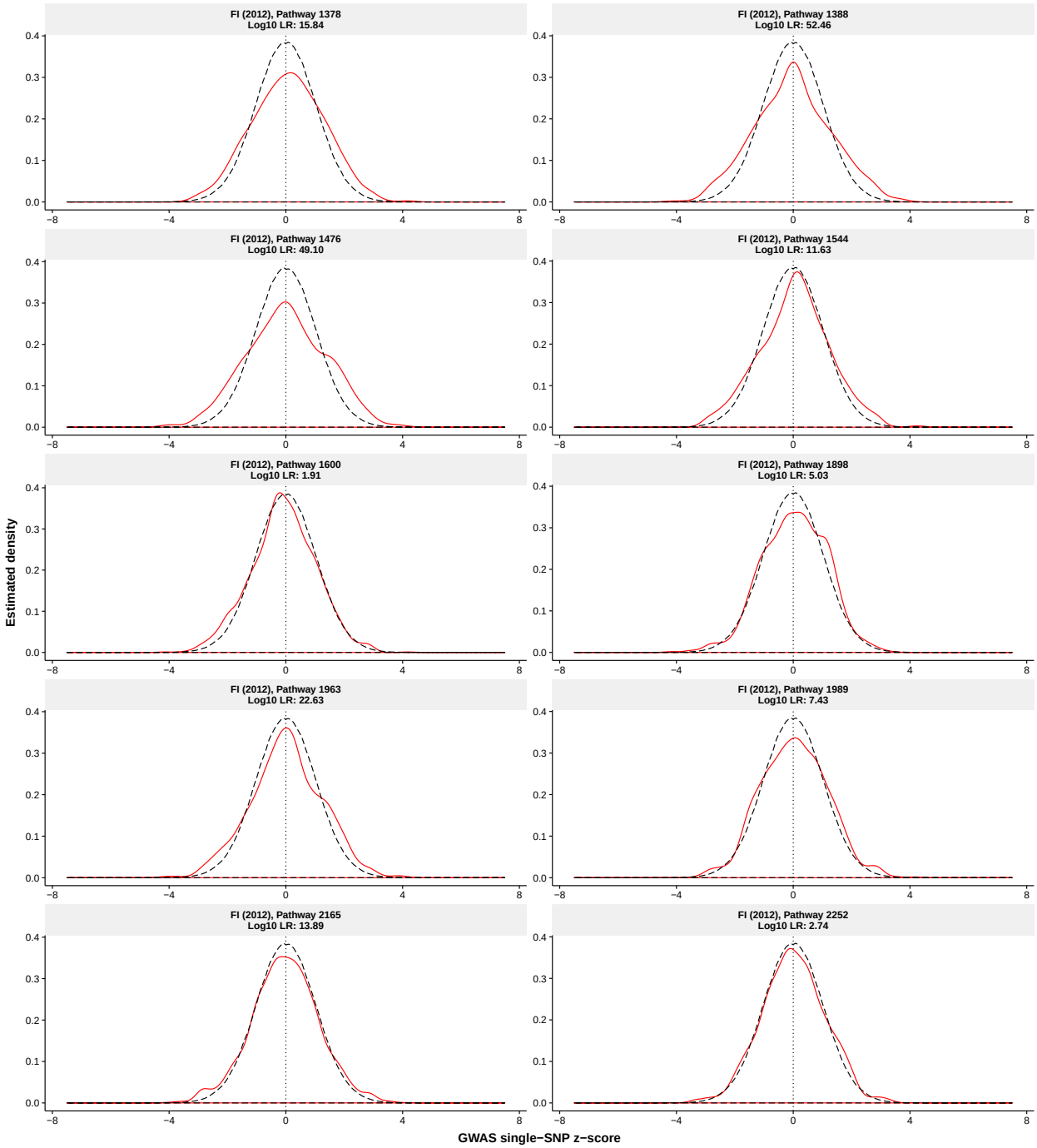
Note that there are only five pathways with at least 10 member genes in Round 2 enrichment analysis of depressive symptoms; see `http://xiangzhu.github.io/rss-gsea/results/ds2016-2.html`.



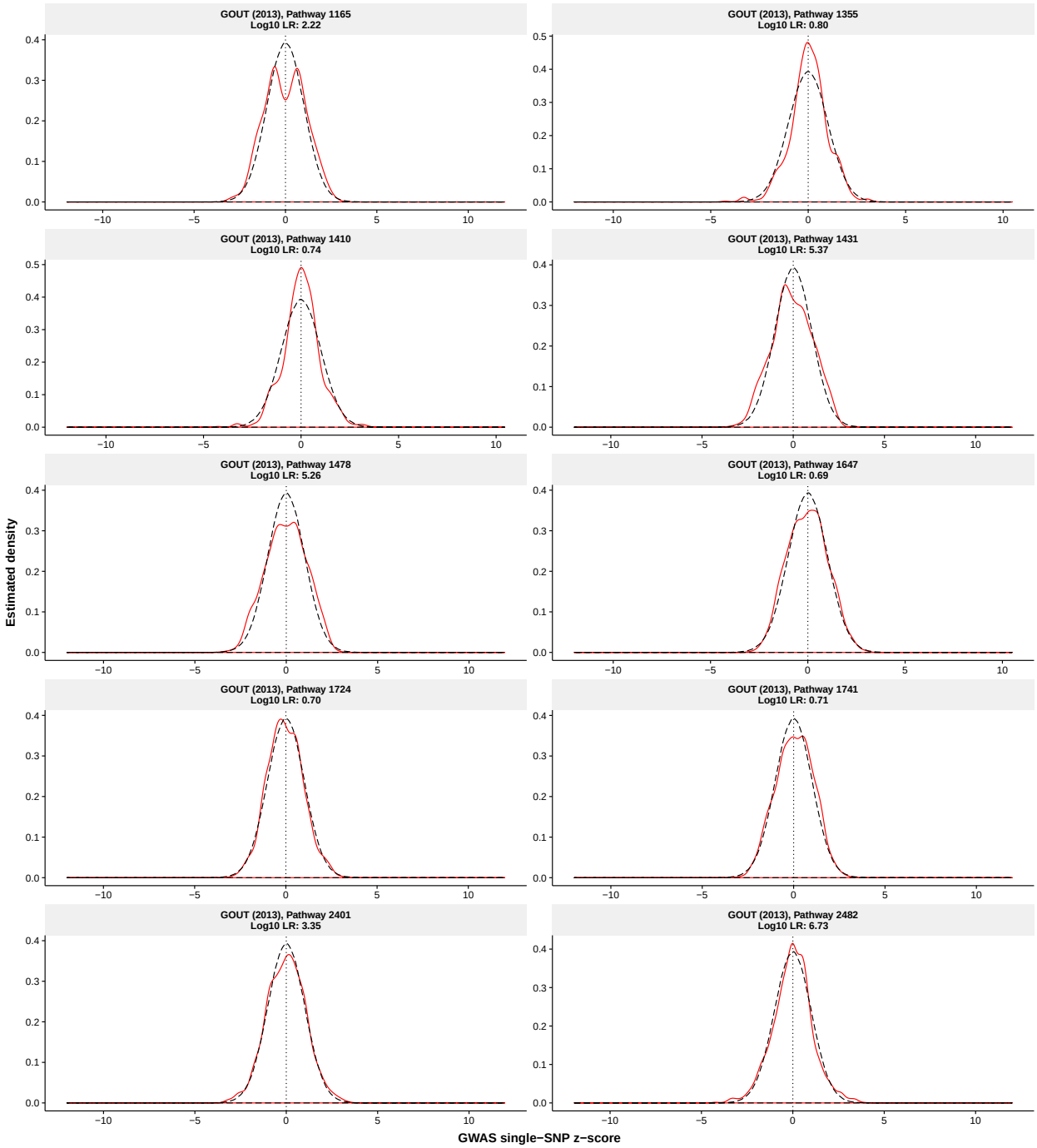
## Fasting glucose levels (Manning et al. 2012)



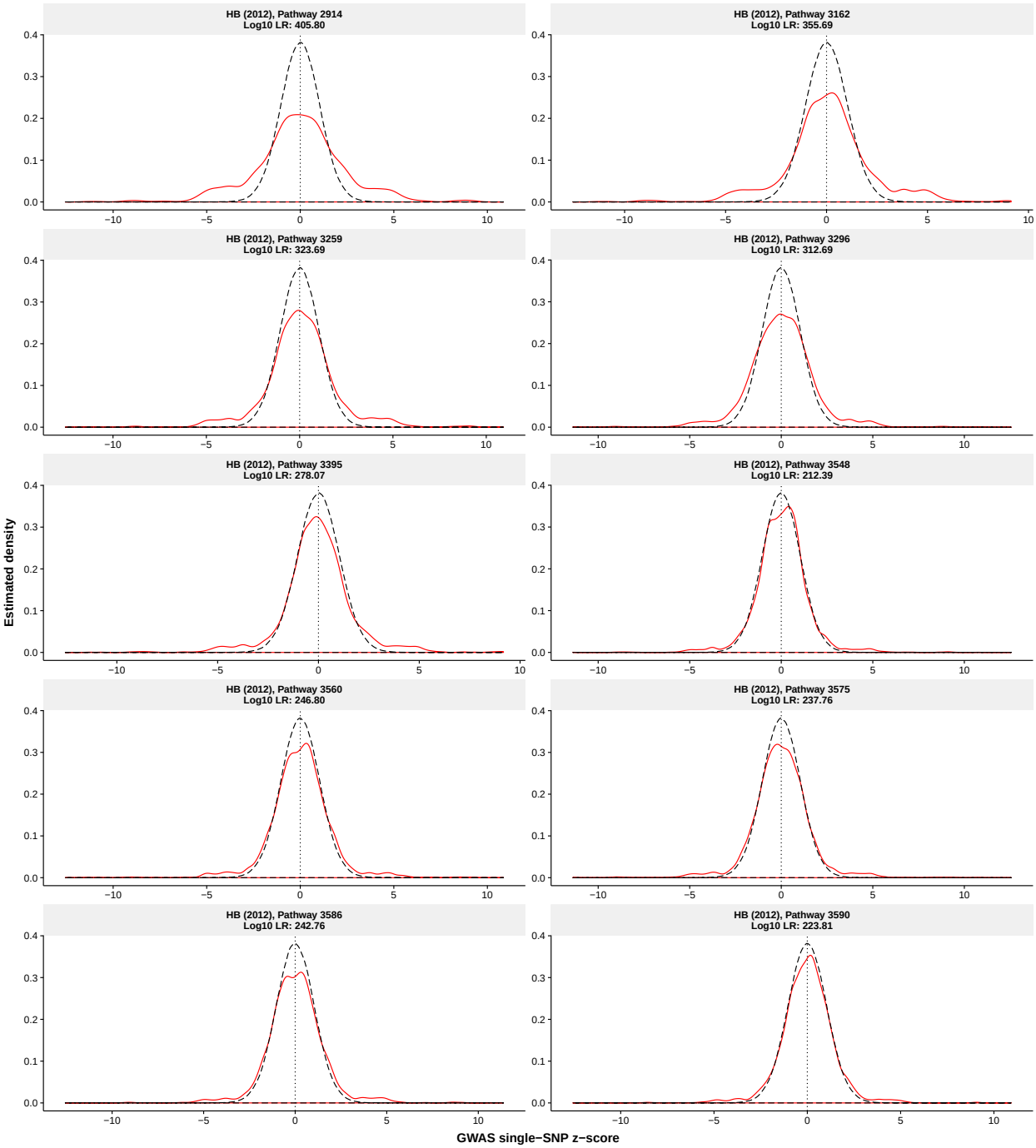
## Fasting insulin levels (Manning et al. 2012)



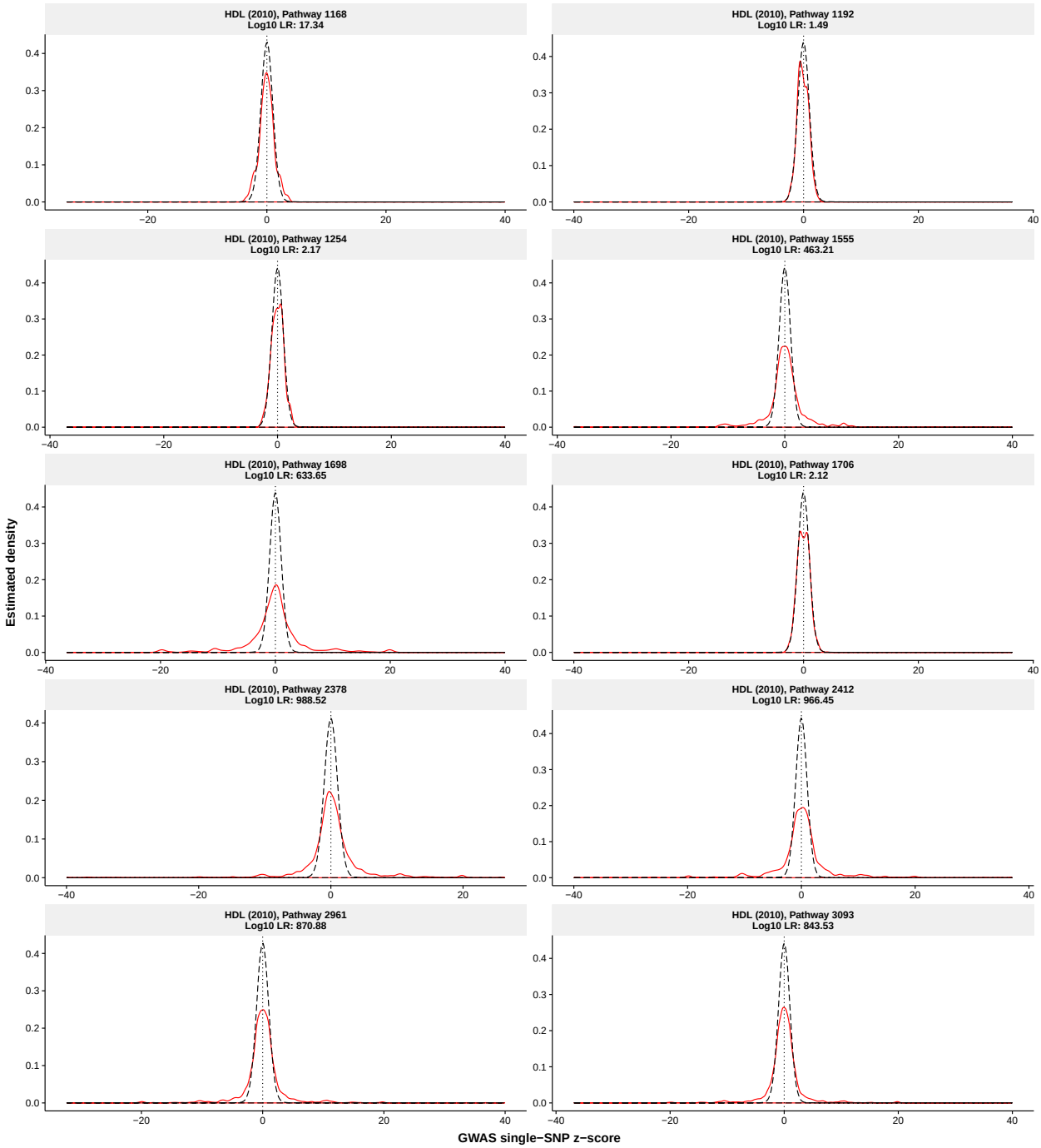
## Gout (Köttgen et al. 2013)



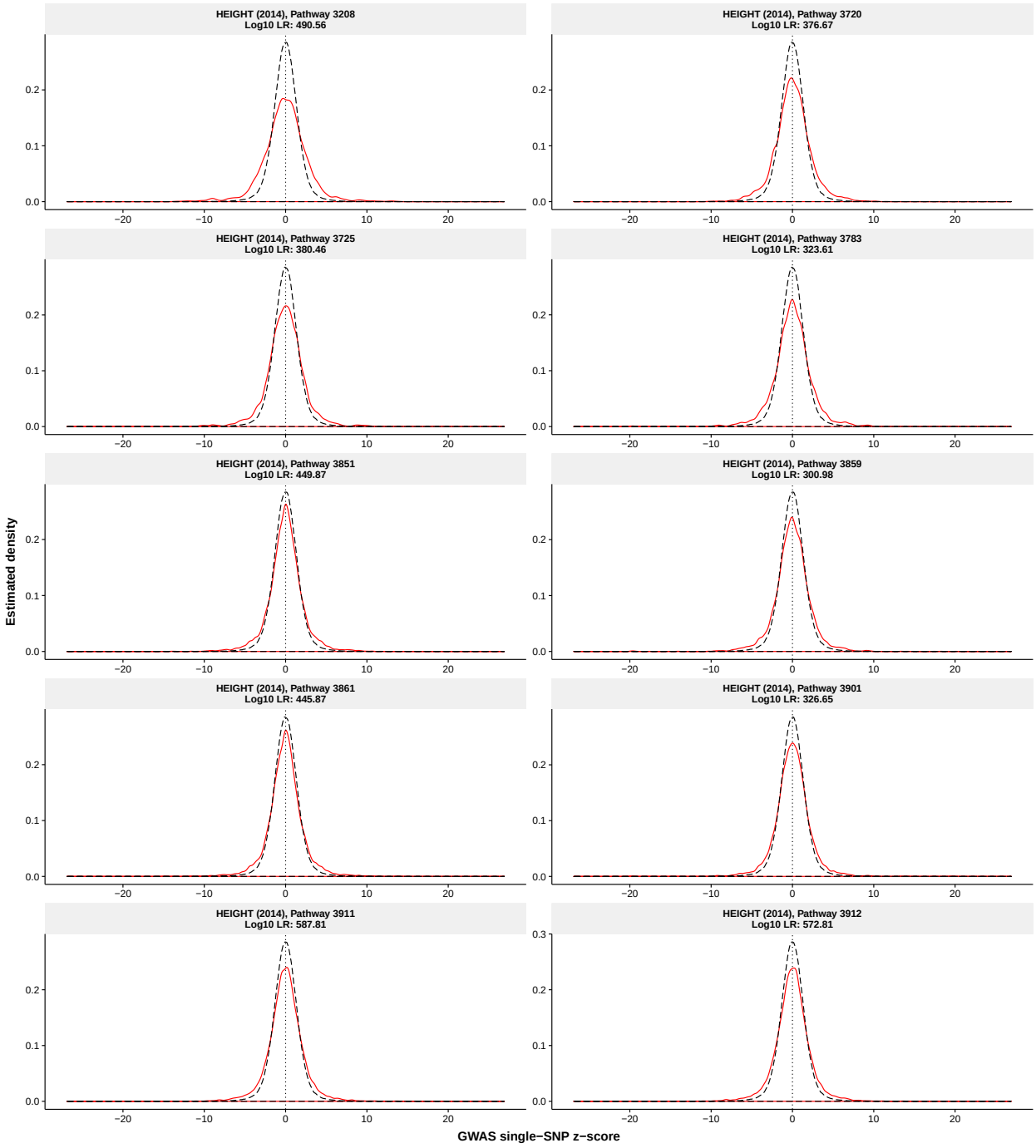
## Haemoglobin (Harst et al. 2012)



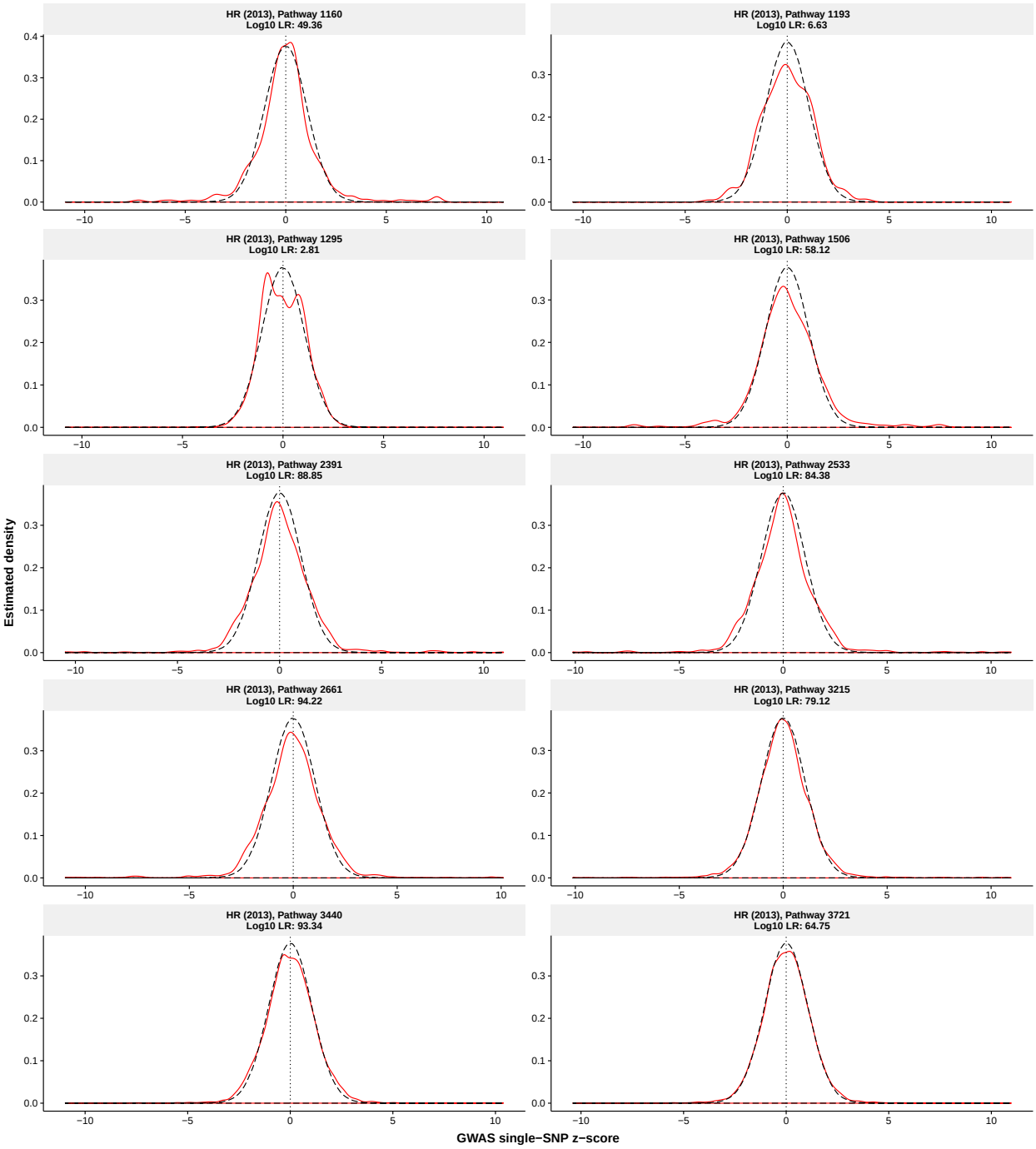
## High-density lipoprotein (Teslovich et al. 2010)



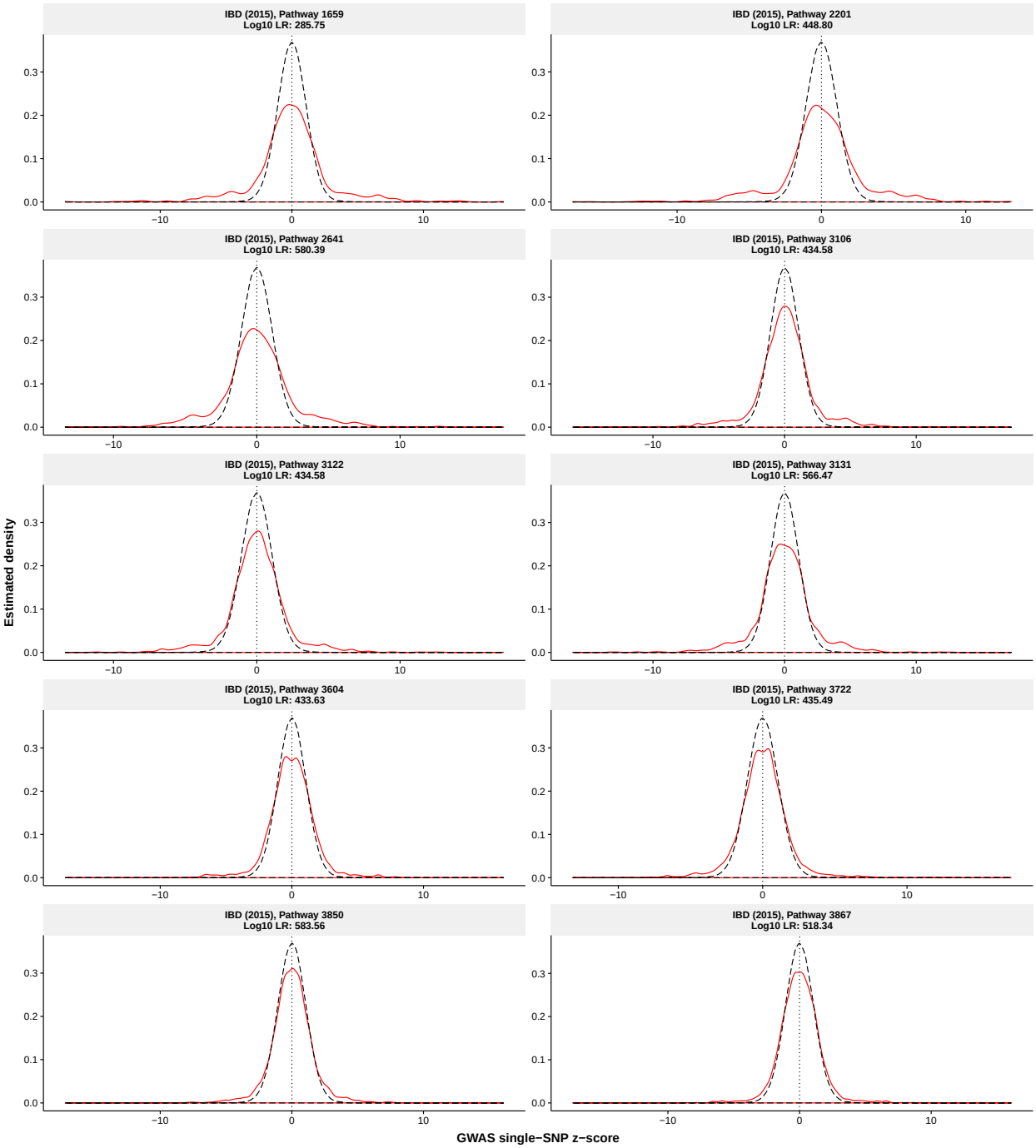
## Adult height (Wood et al. 2014)



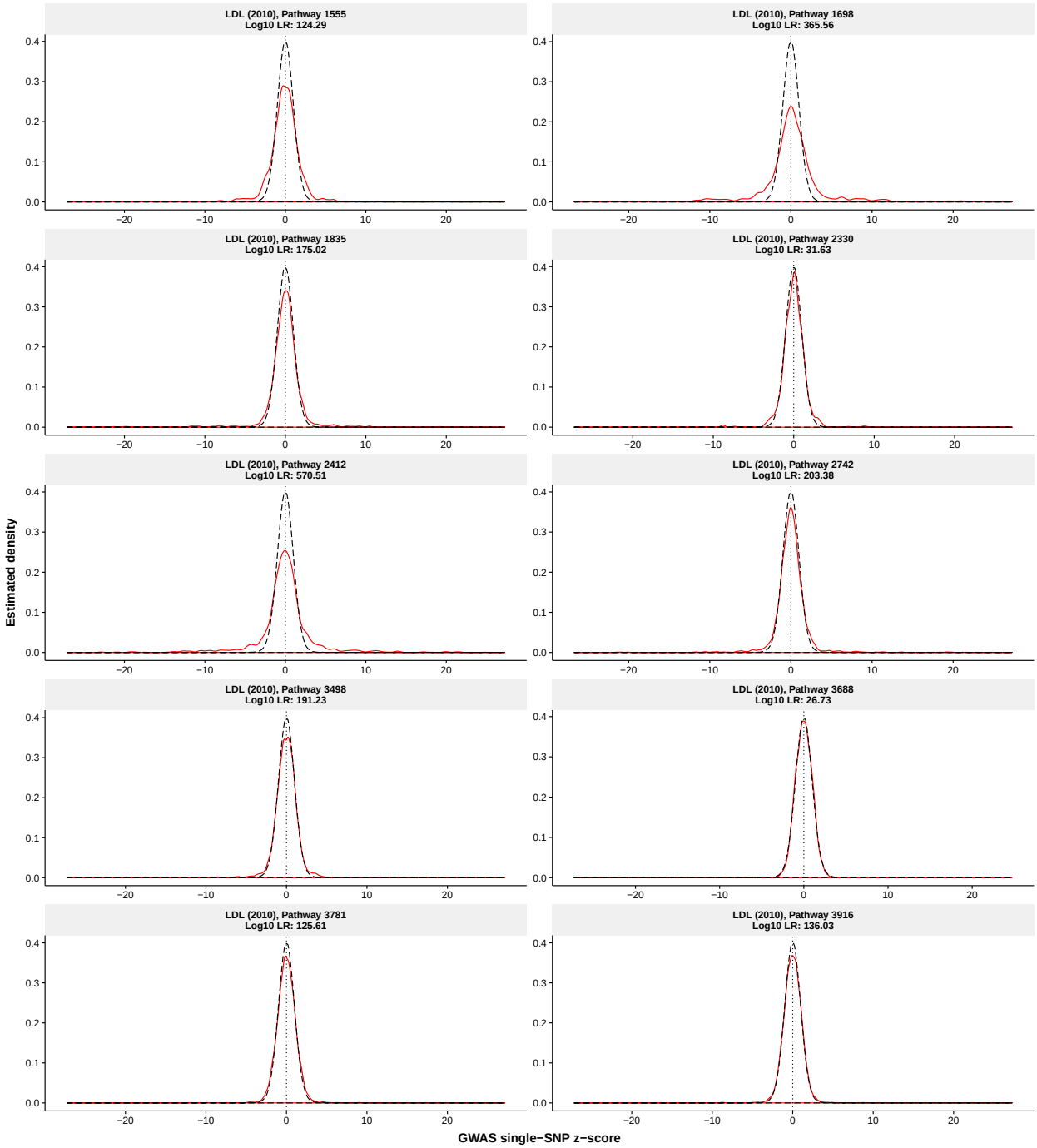
## Heart rate (Den Hoed et al. 2013)



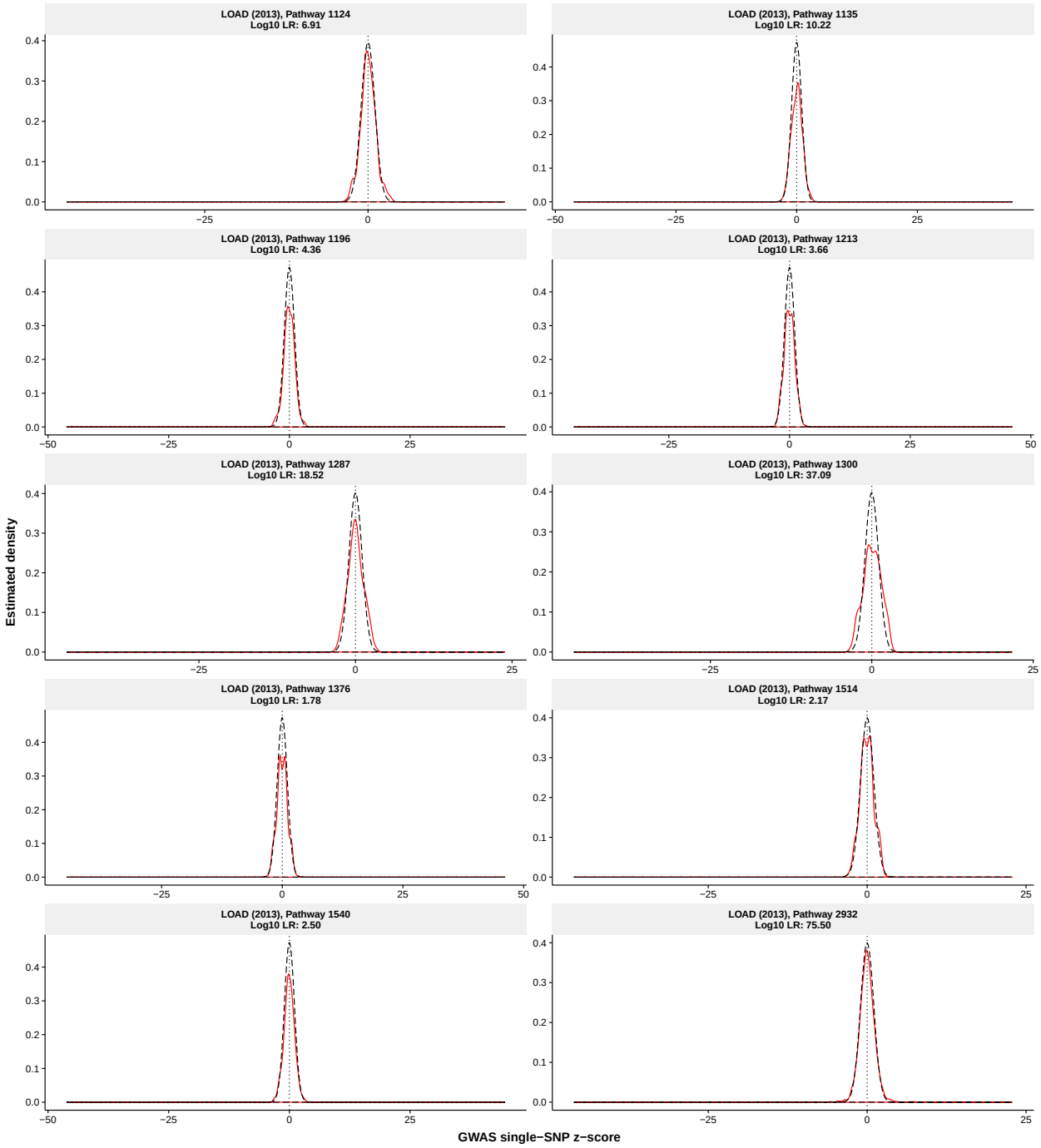
## Inflammatory bowel disease (Liu et al. 2015)



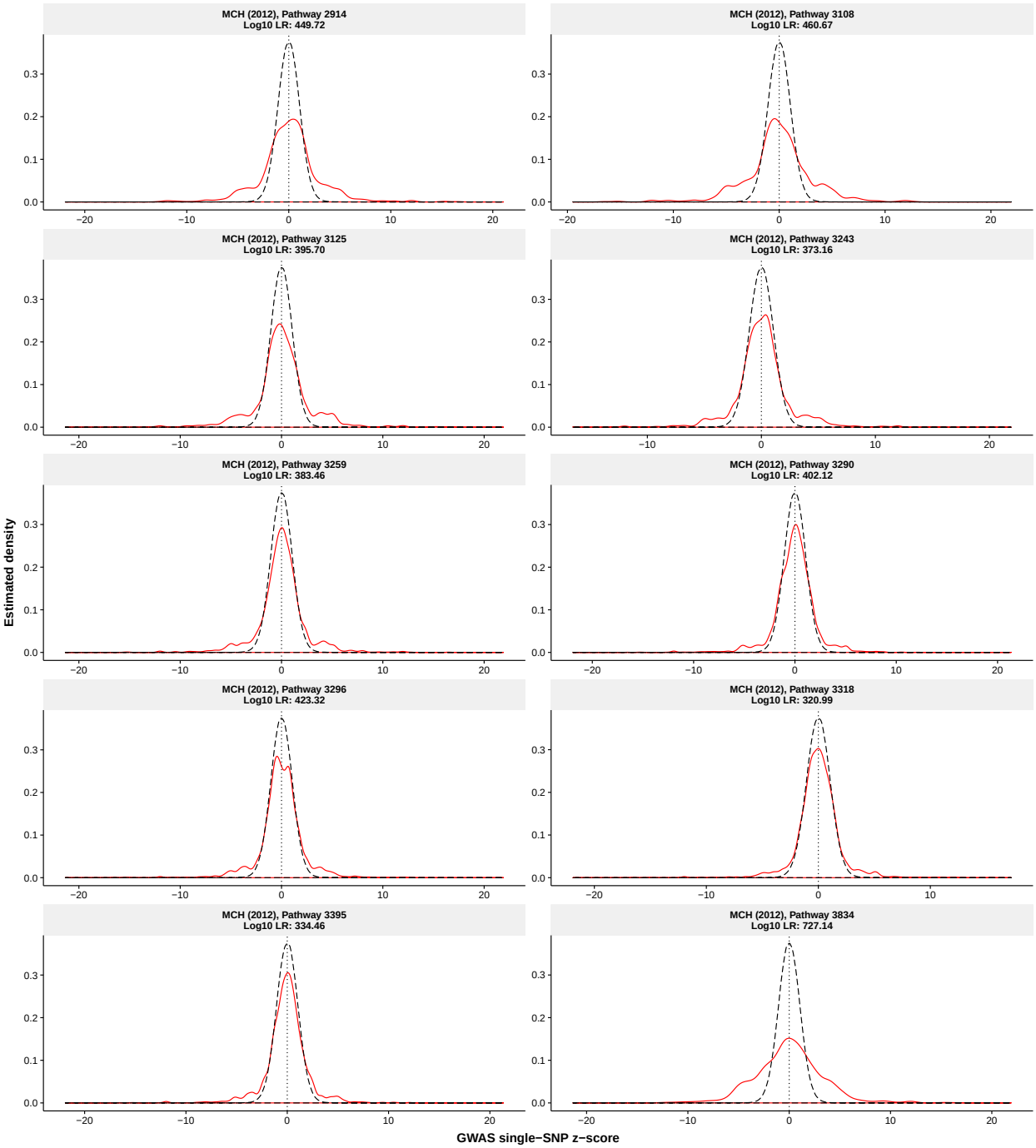
## Low-density lipoprotein (Teslovich et al. 2010)



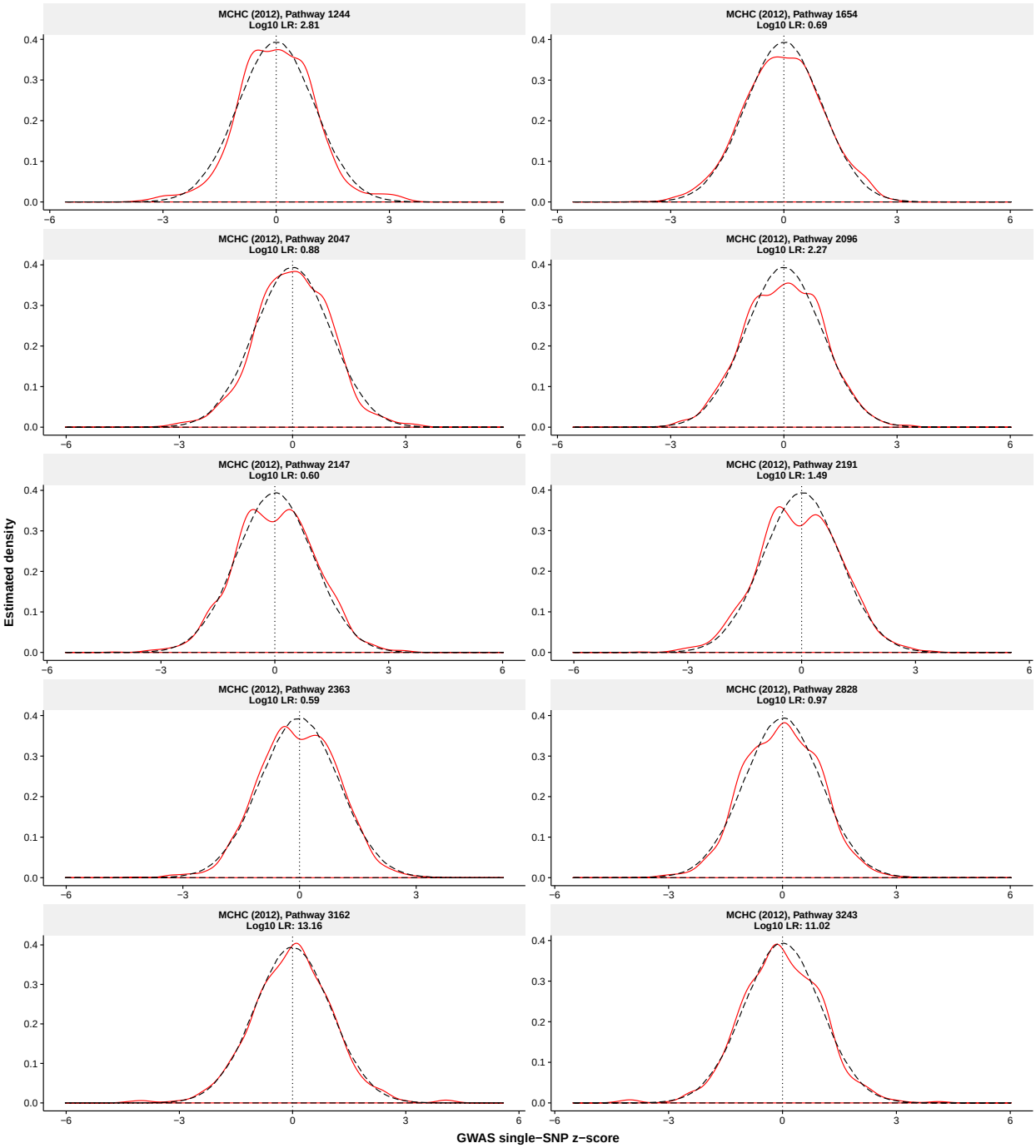
## Alzheimer's disease (Lambert et al. 2013)



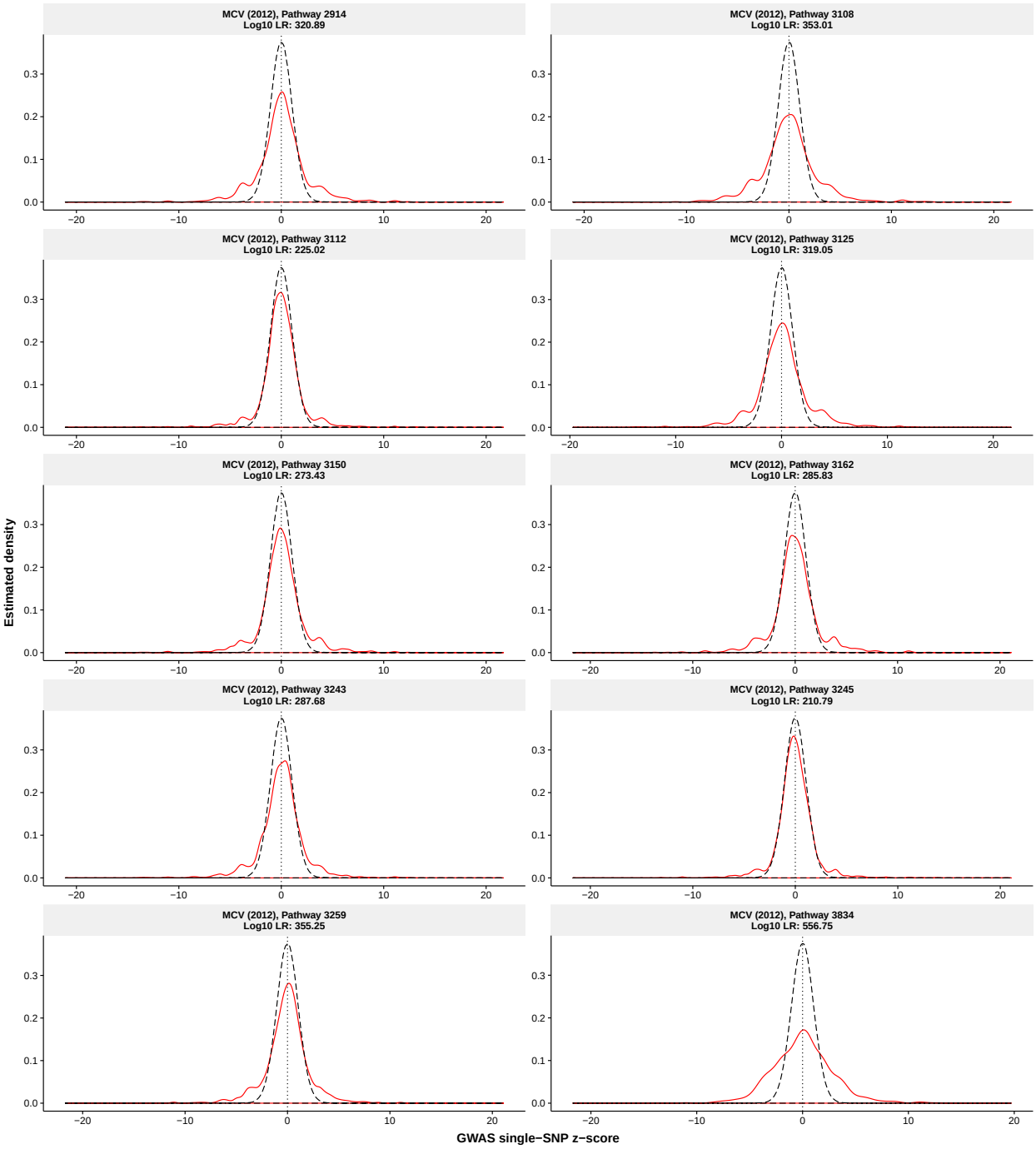
## Mean cell haemoglobin (Harst et al. 2012)



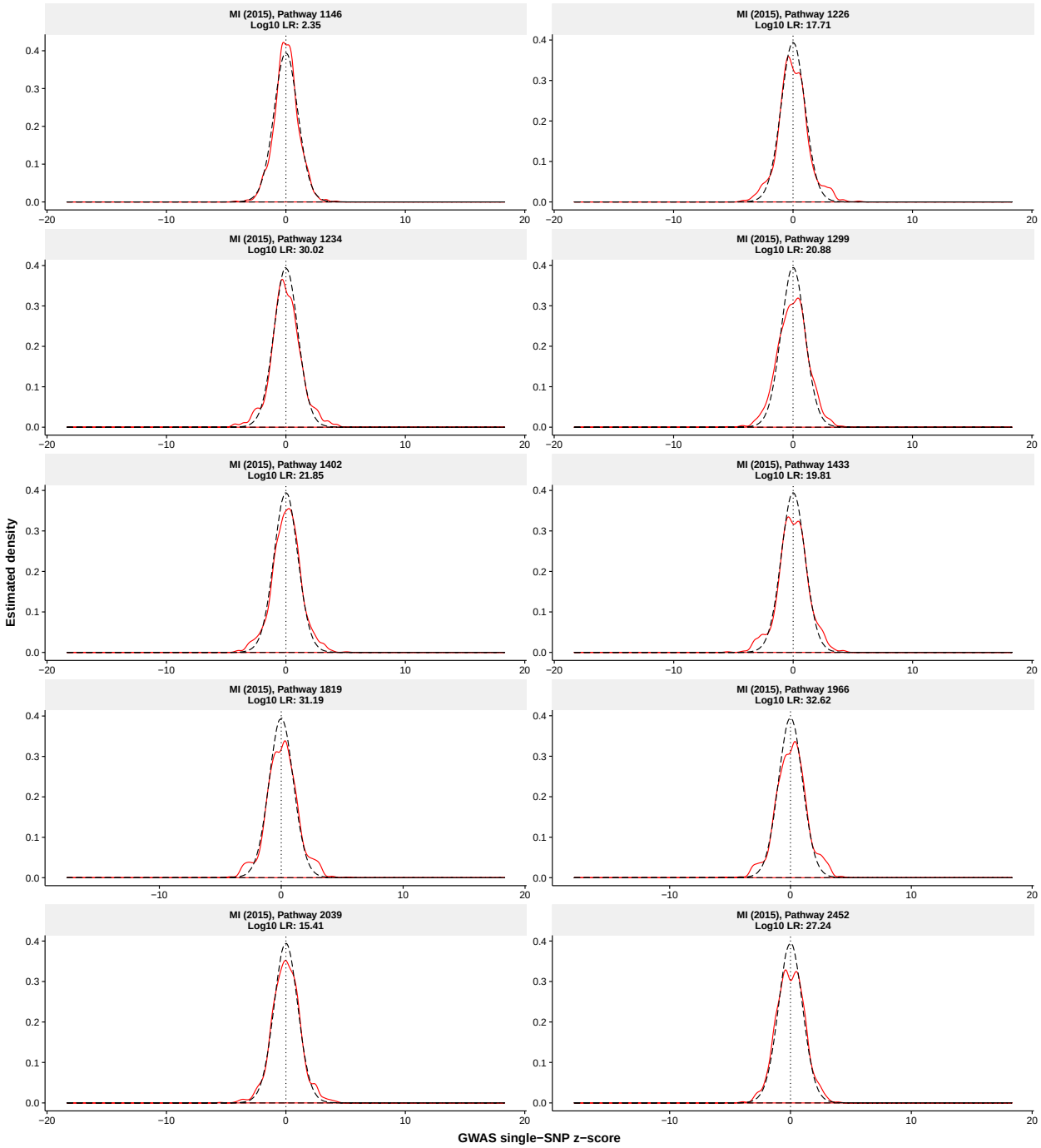
## Mean cell haemoglobin concentration (Harst et al. 2012)



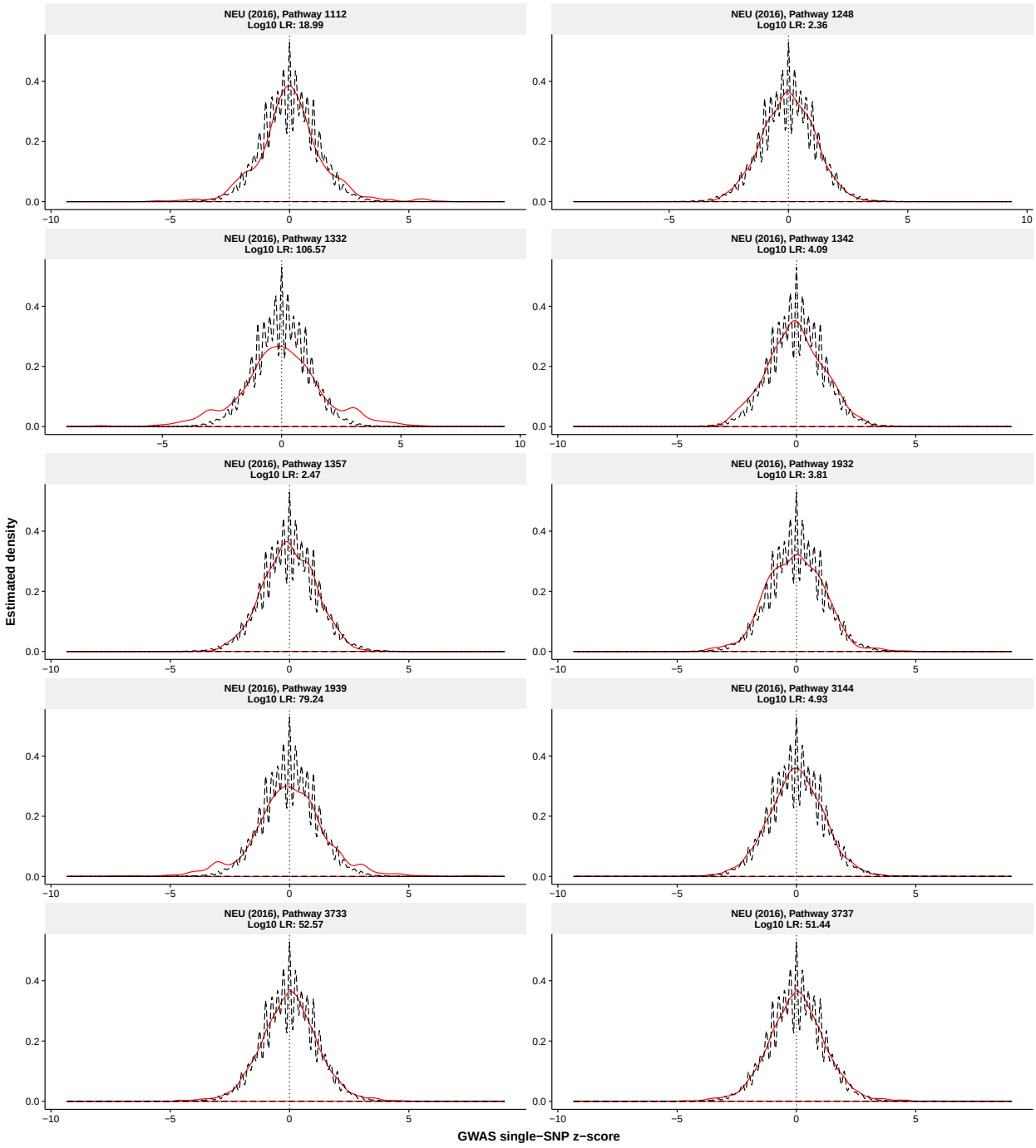
## Mean cell volume (Harst et al. 2012)



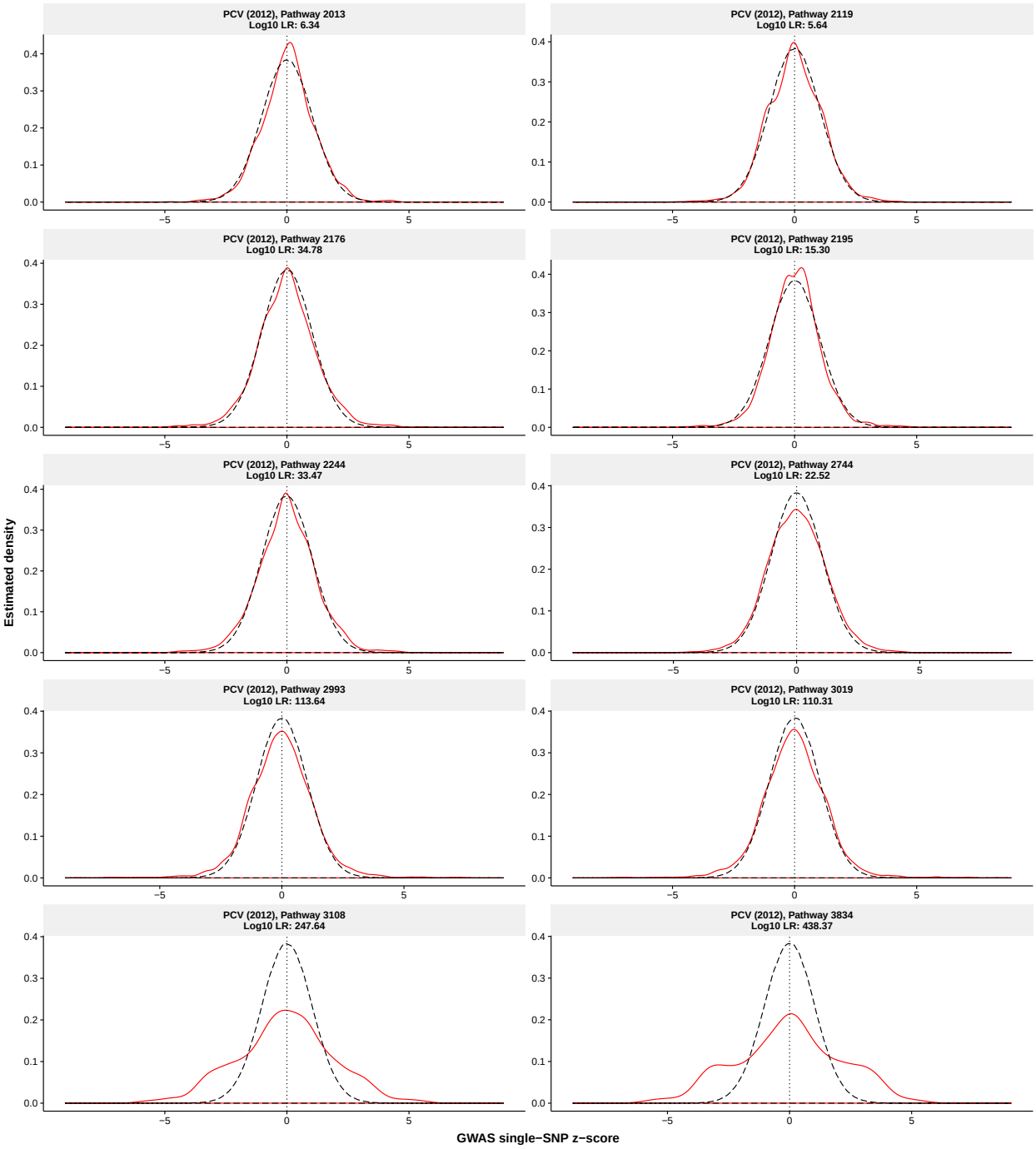
## Myocardial infarction (Nikpay et al. 2015)



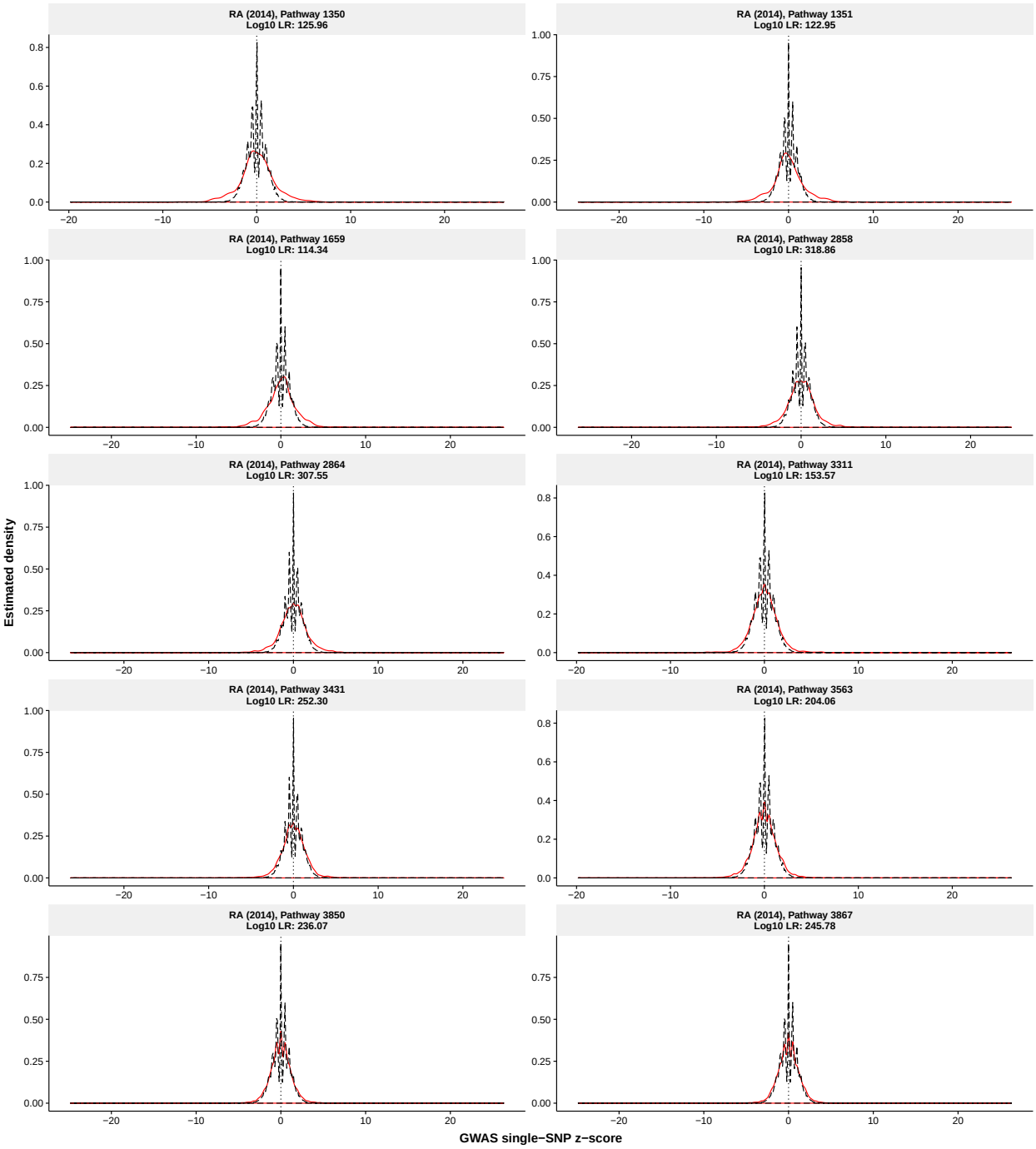
## Neuroticism (Okbay et al. 2016)



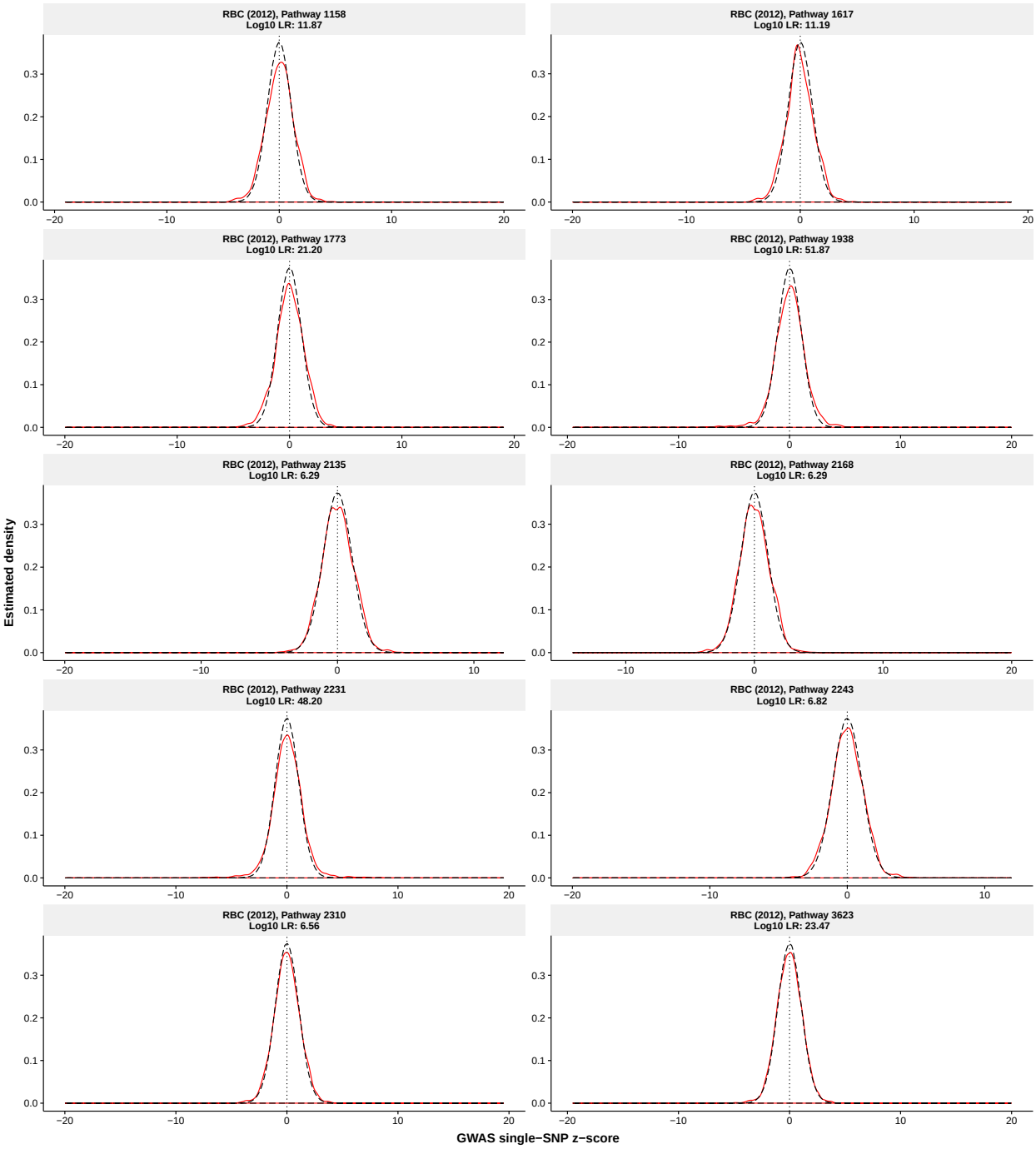
## Packed cell volume (Harst et al. 2012)



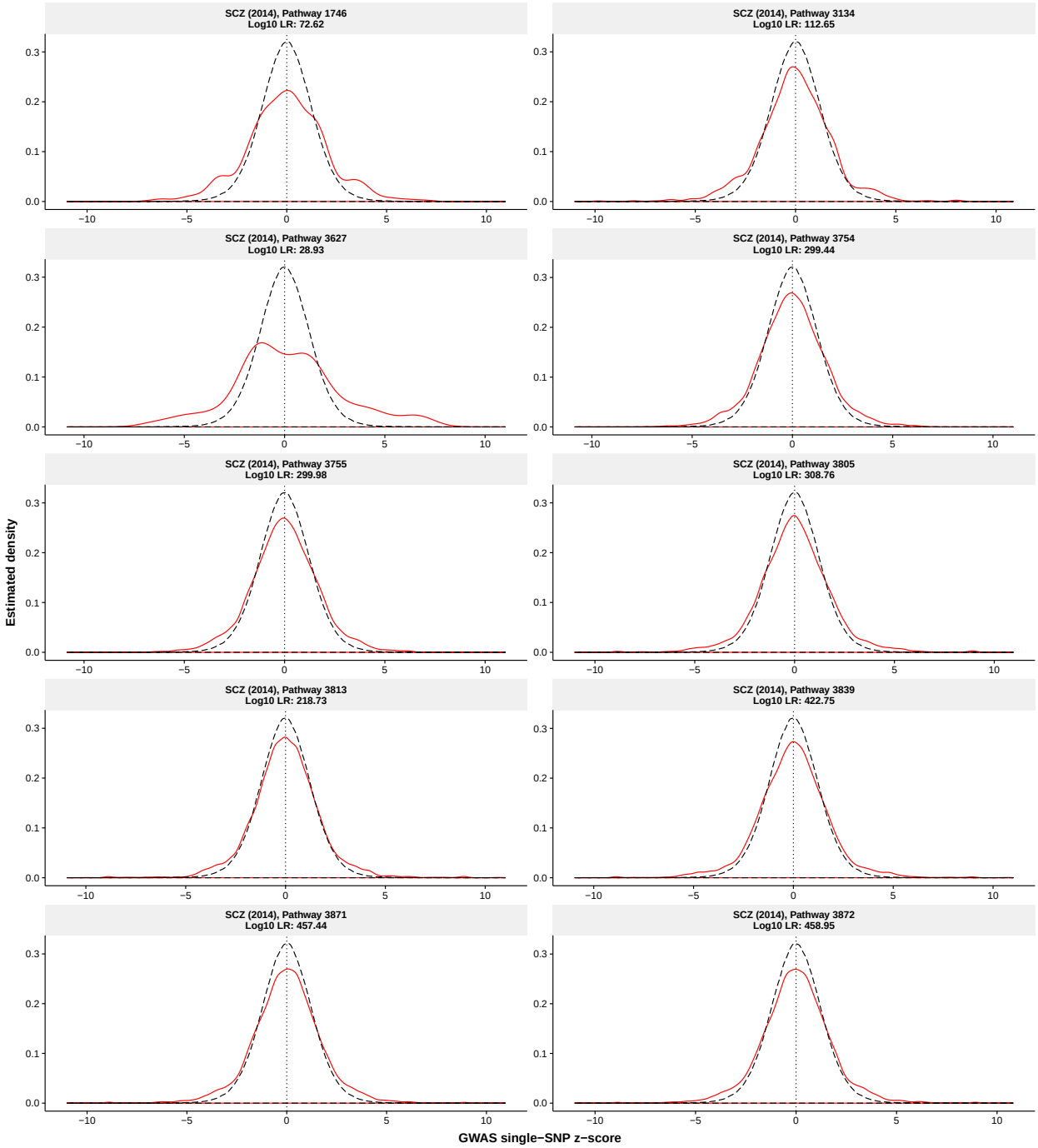
## Rheumatoid arthritis (Okada et al. 2014)



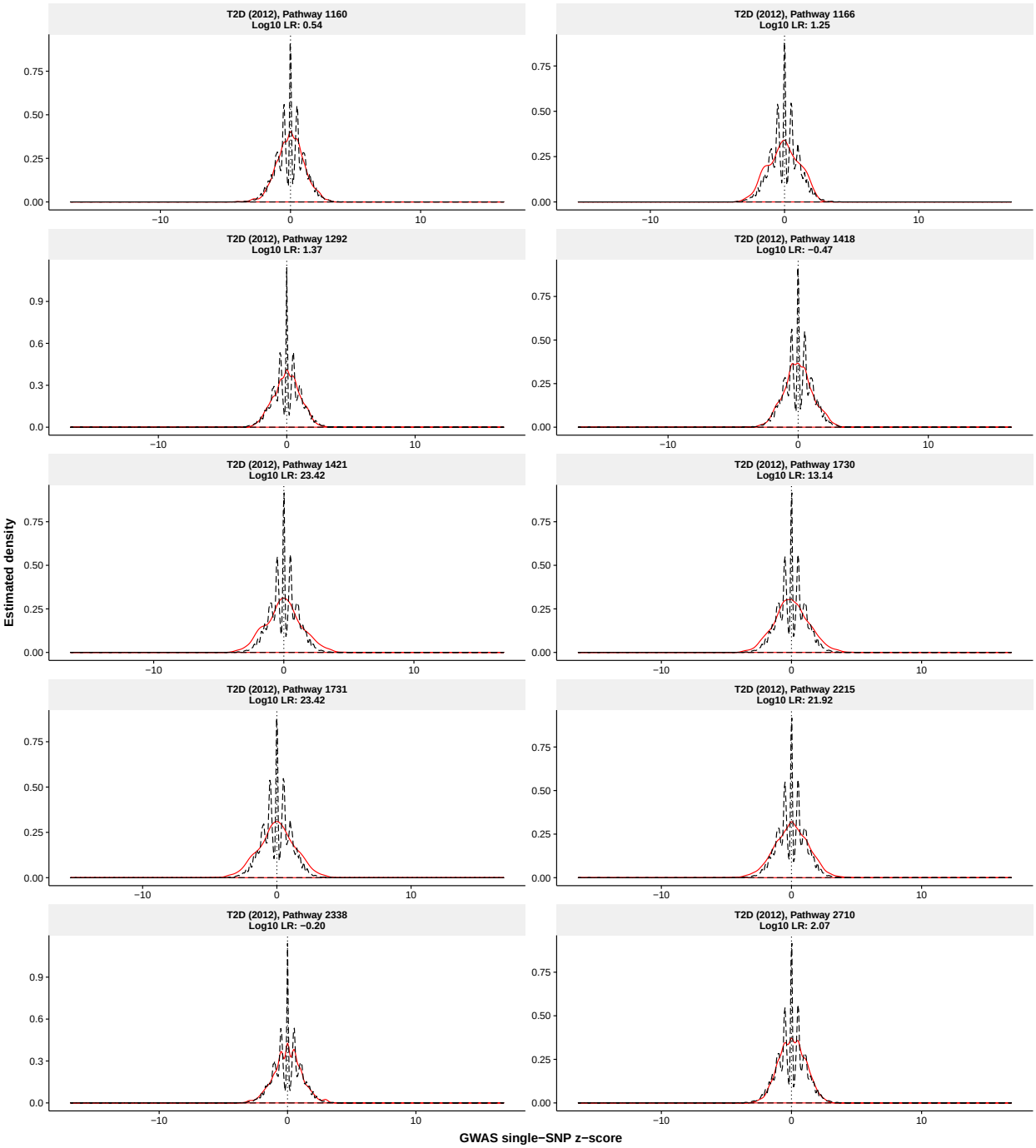
## Red blood cell count (Harst et al. 2012)



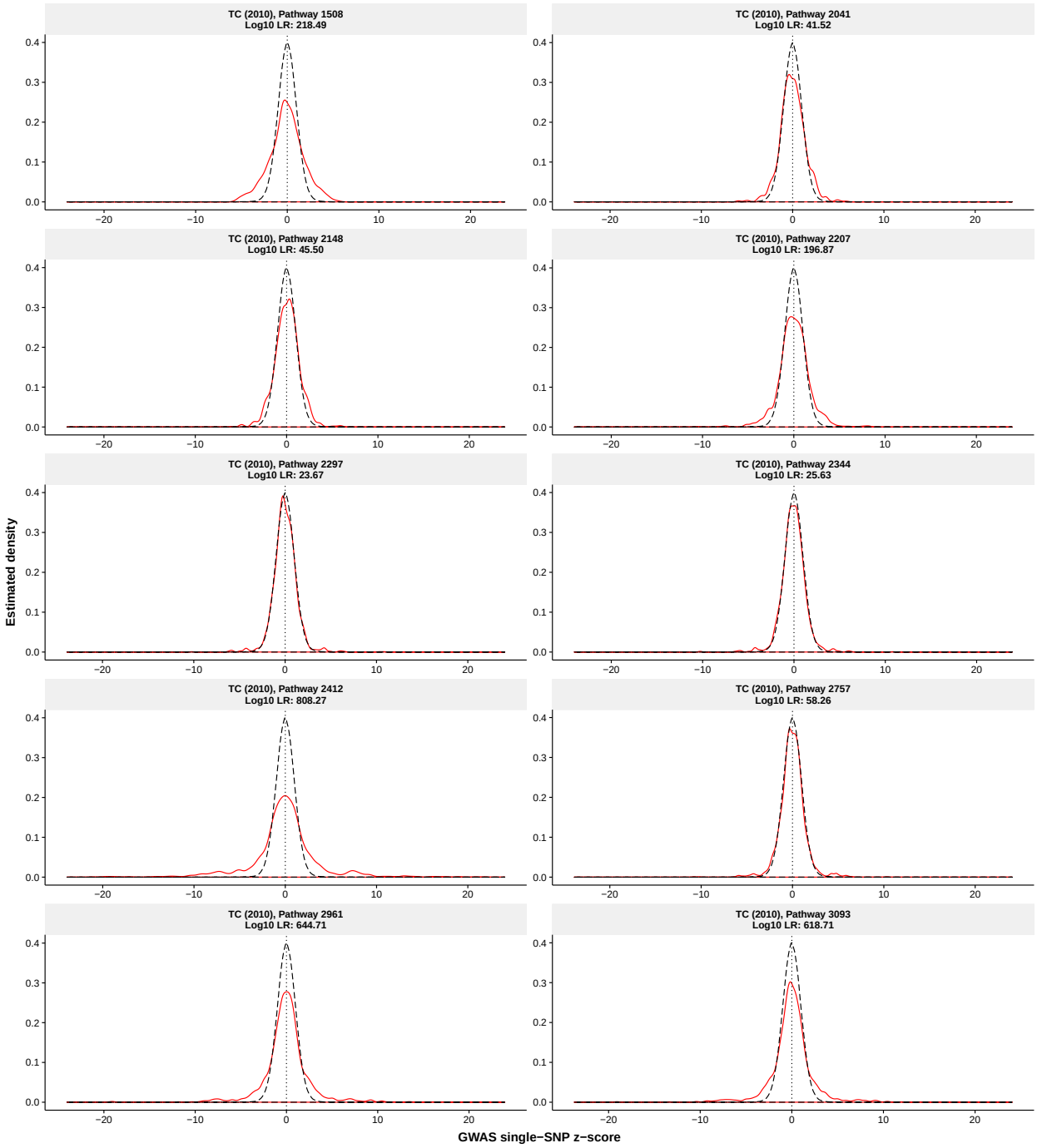
## Schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014)



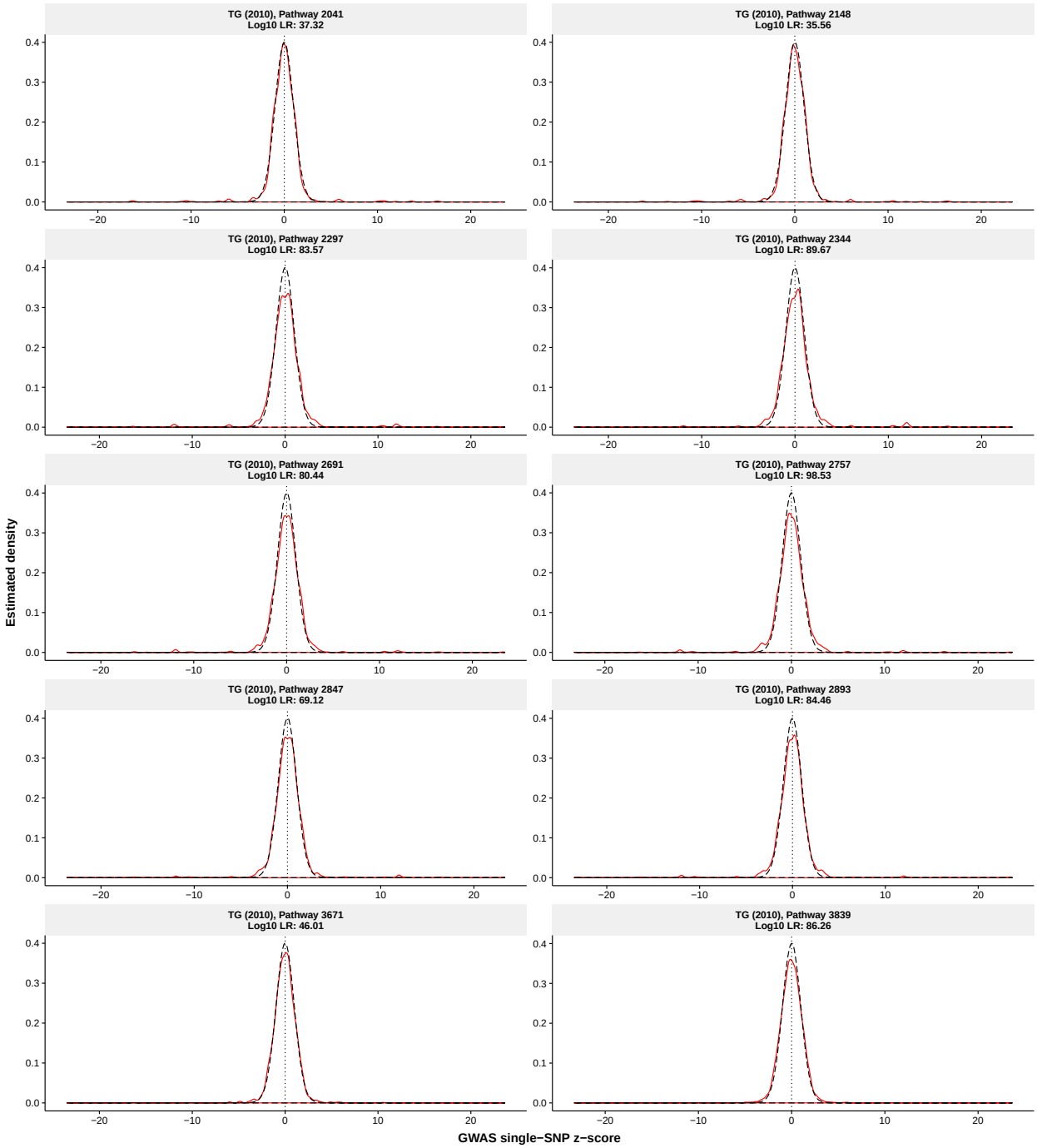
## Type 2 diabetes (Morris et al. 2012)



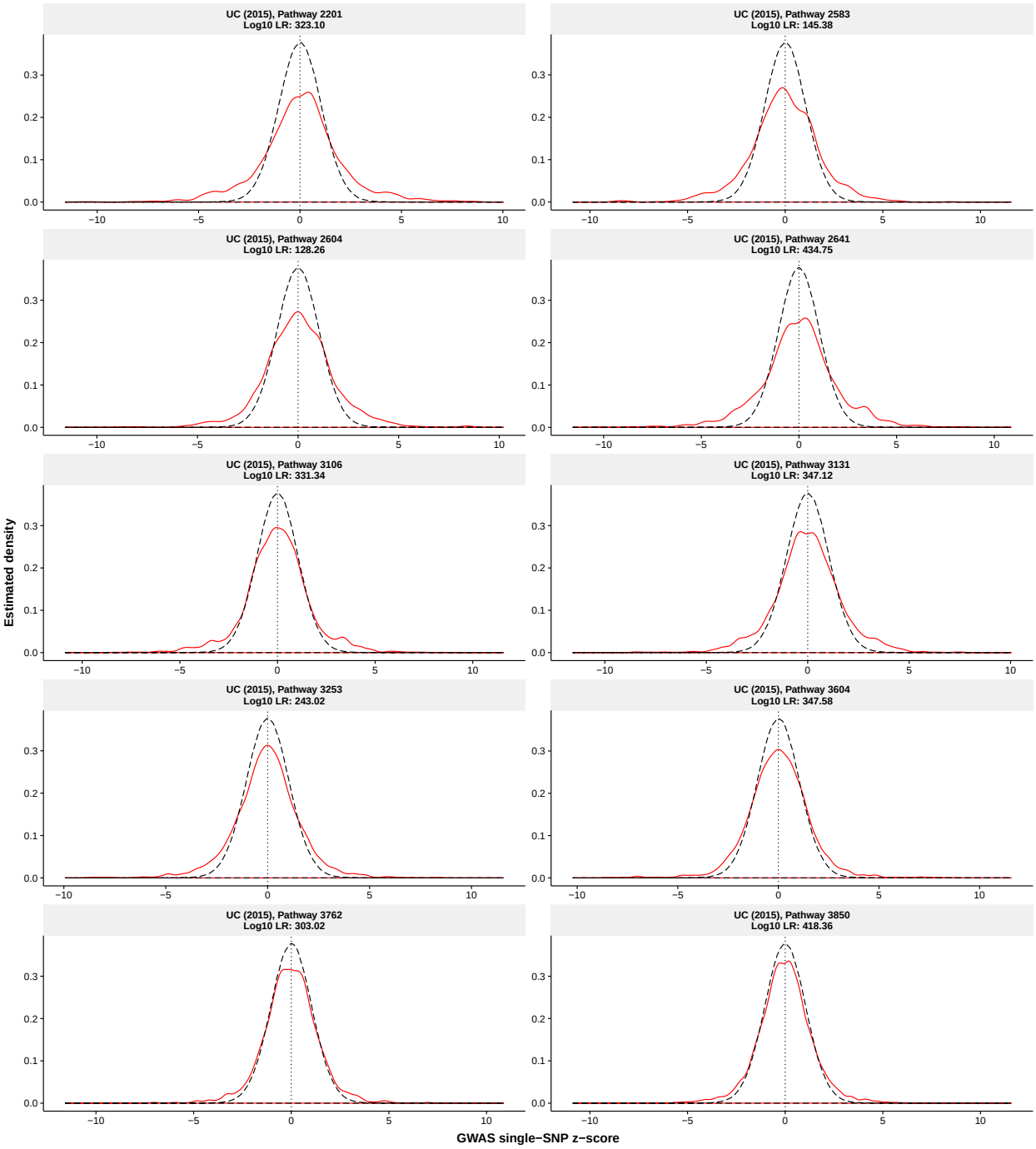
## Total cholesterol (Teslovich et al. 2010)



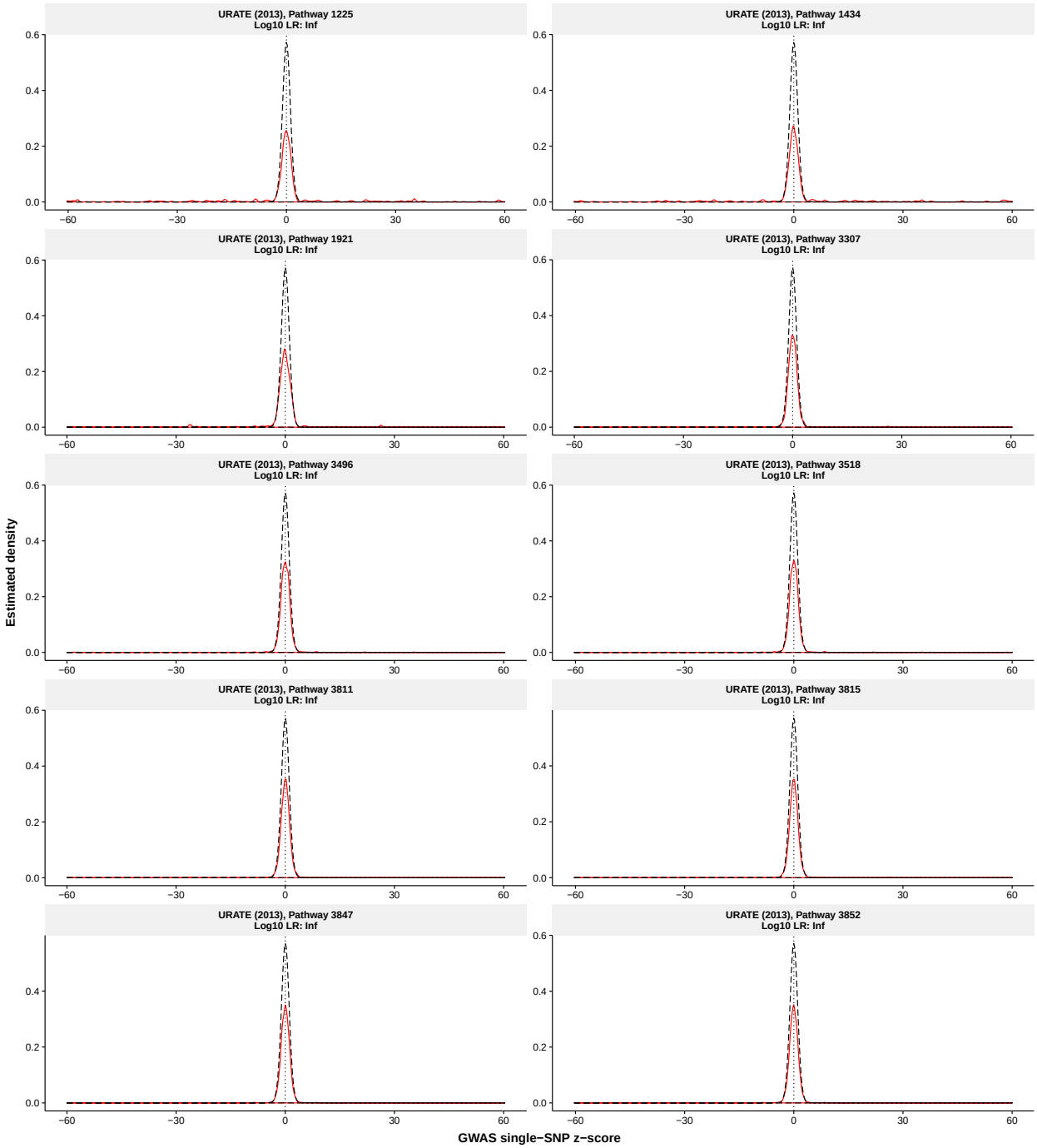
## Triglycerides (Teslovich et al. 2010)



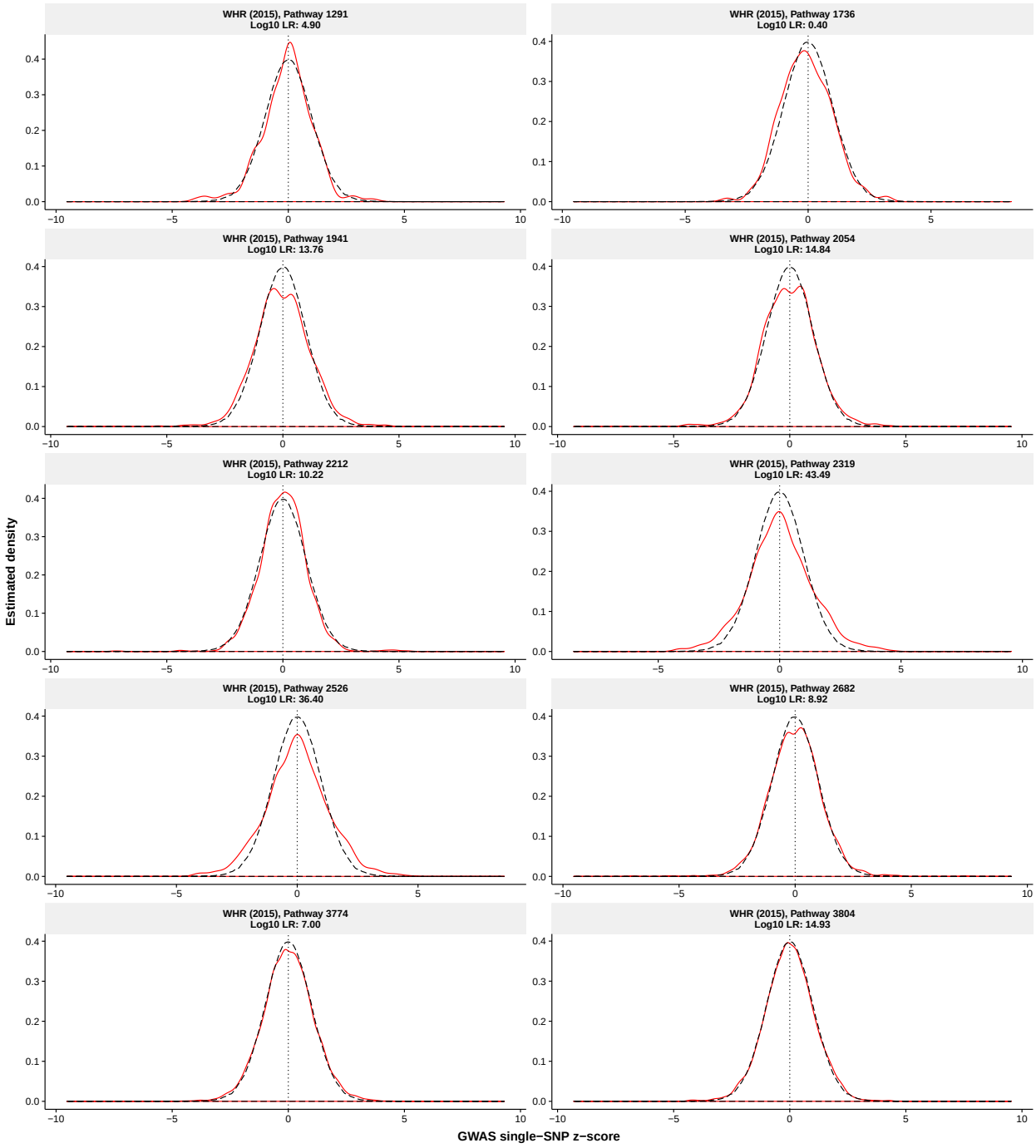
## Ulcerative colitis (Liu et al. 2015)



## Serum urate concentrations (Köttgen et al. 2013)

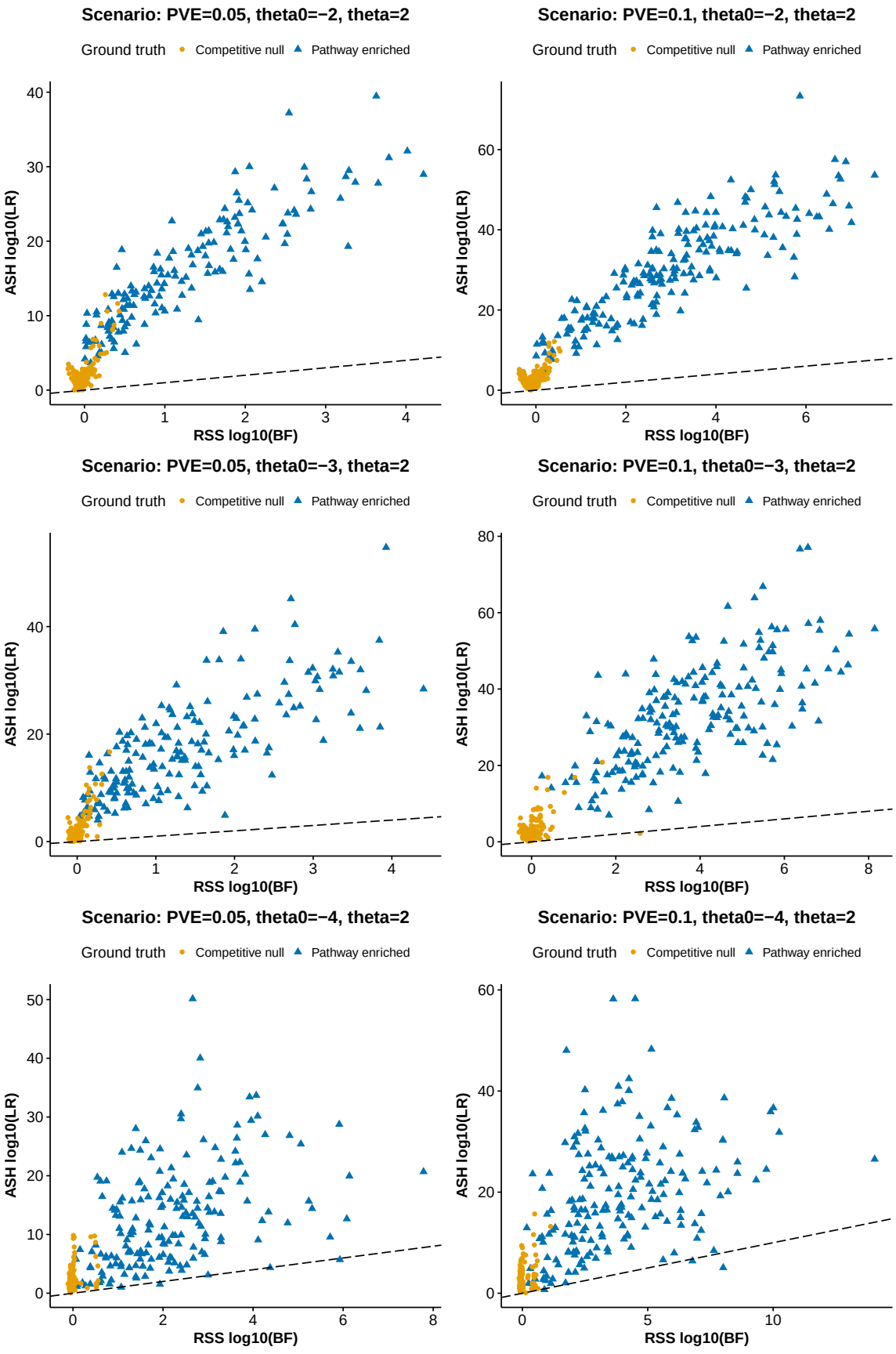


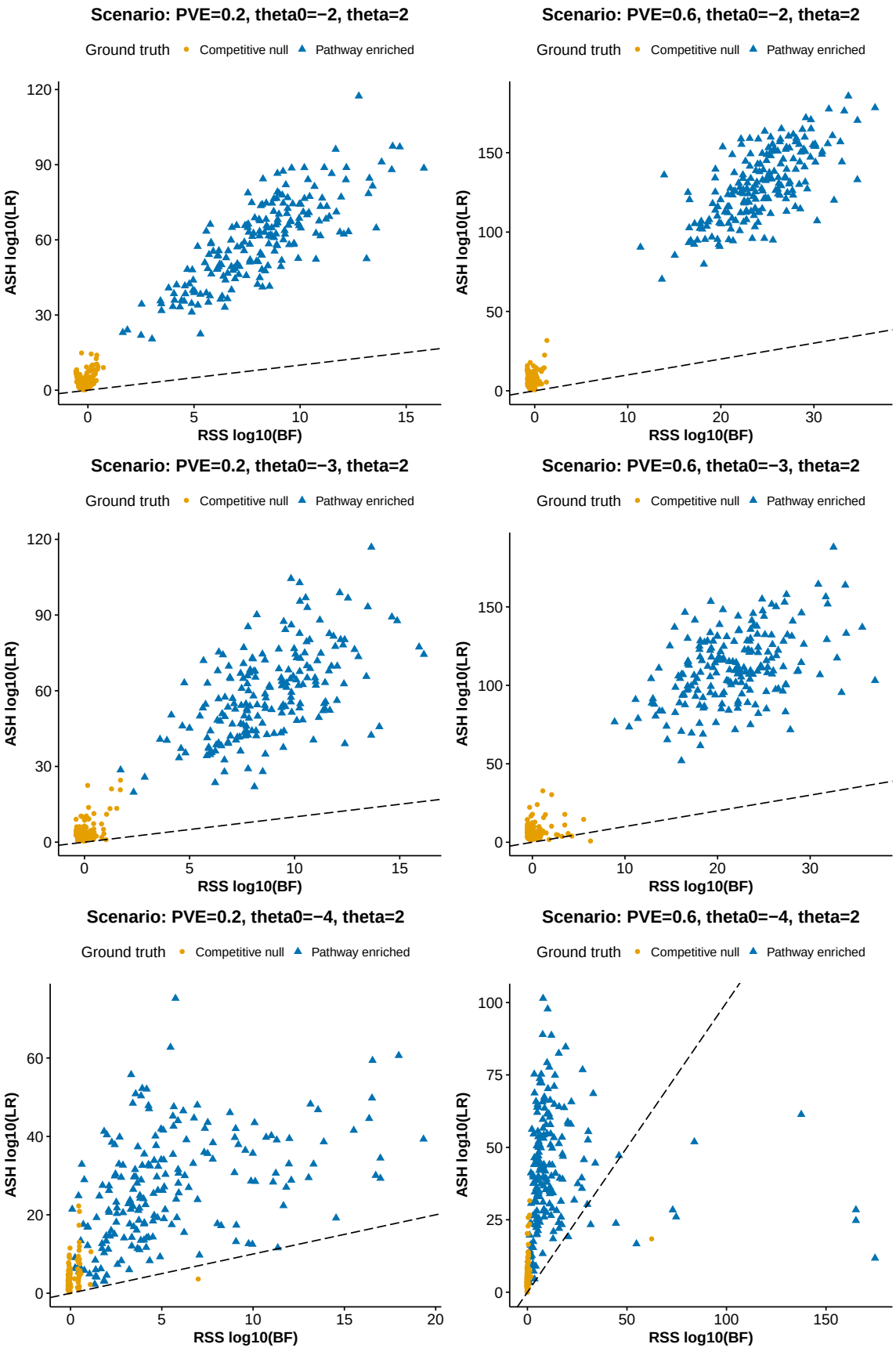
## Waist-to-hip ratio adjusted for body mass index (Shungin et al. 2015)

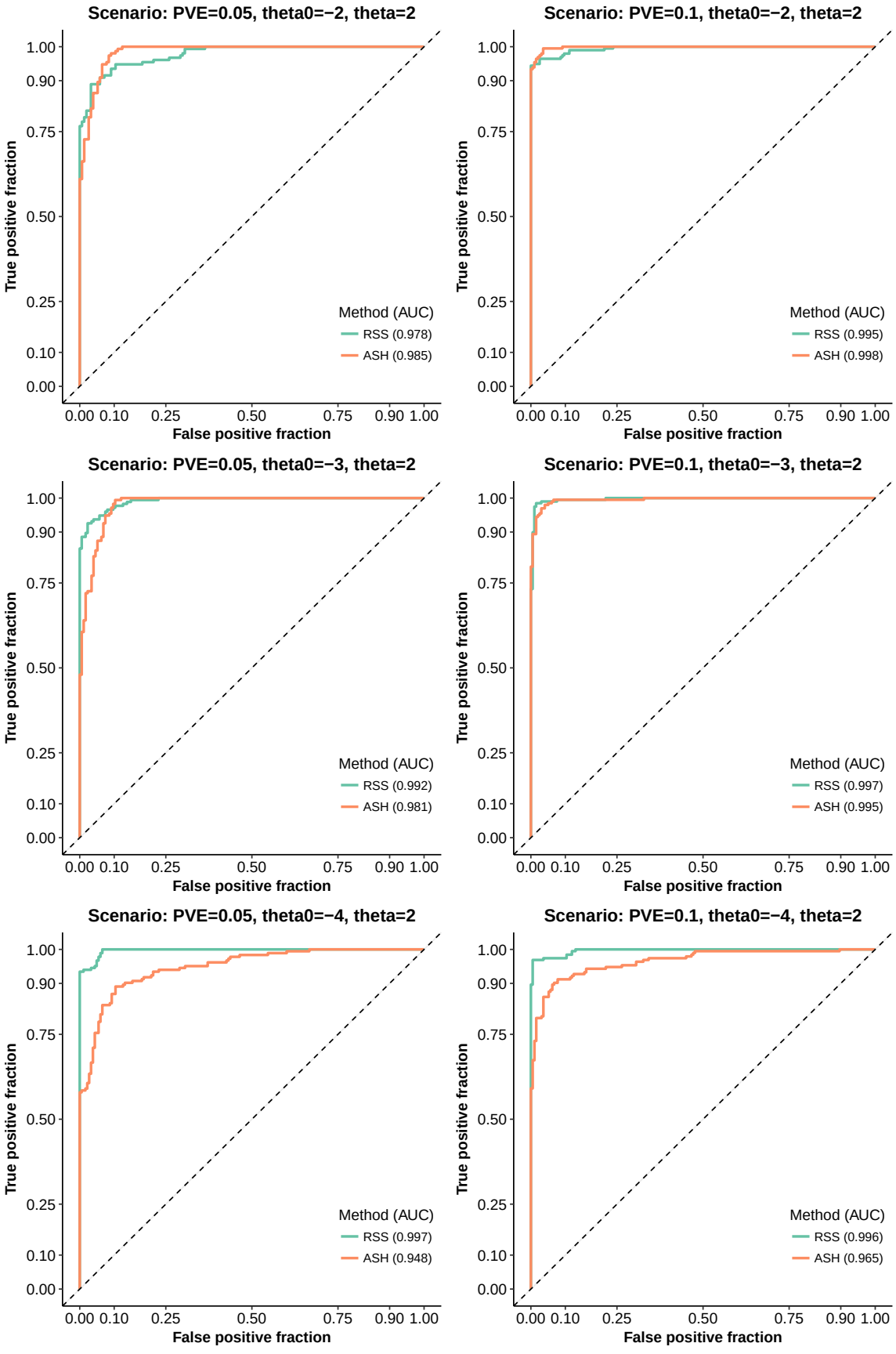


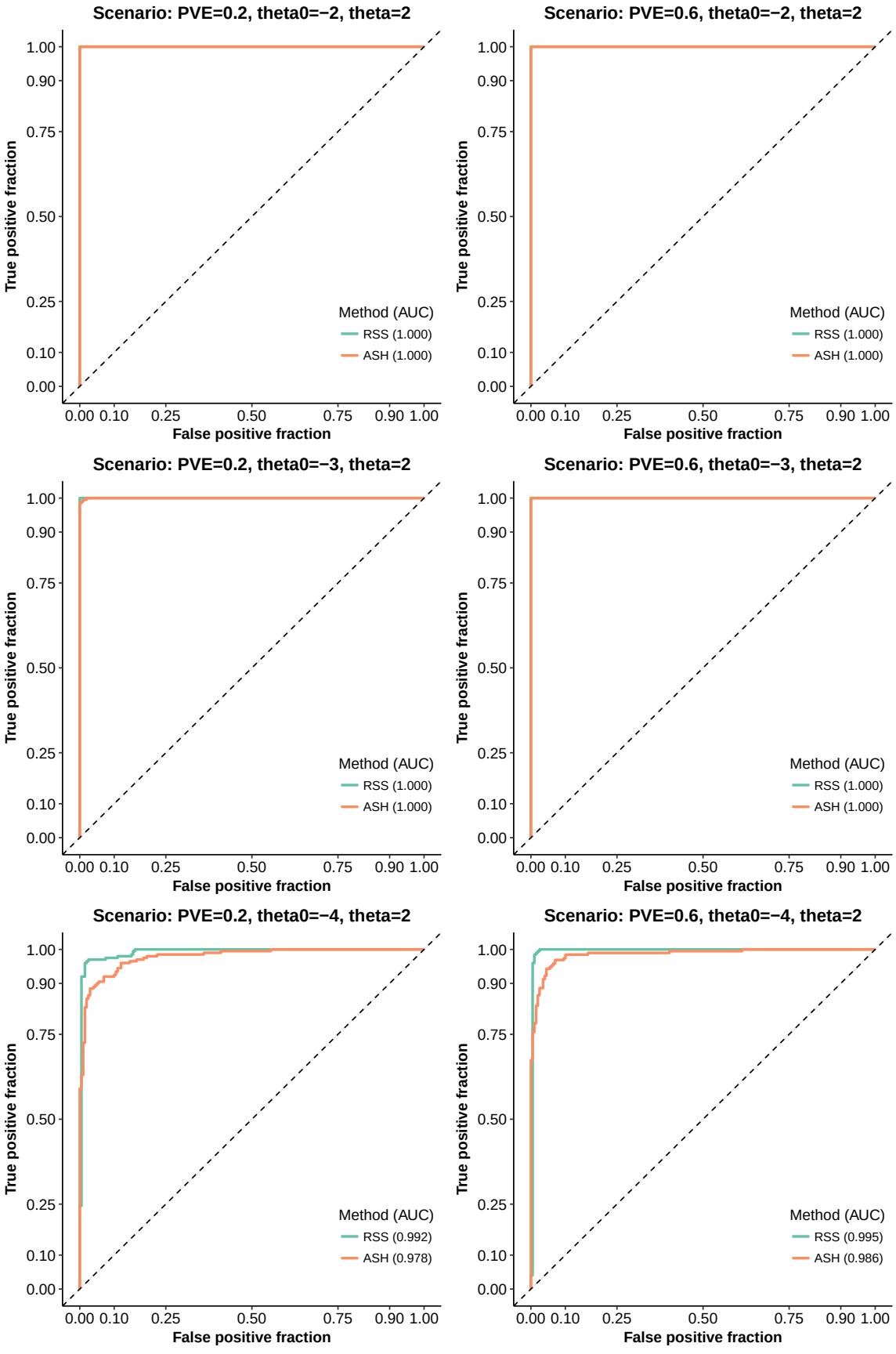
## Supplementary Figure 18

**Comparing the likelihood ratio (LR) with the enrichment Bayes factor (BF).** Simulation details are provided in **Supplementary Figure 1**. For each simulated dataset, we compute a sanity check LR as defined in **Supplementary Figure 17**, and use RSS to obtain an enrichment BF. In each panel, "competitive null" indicates datasets simulated from the baseline model (i.e. no enrichment), and "pathway enriched" indicates datasets simulated from the enrichment model. Dashed lines are reference lines with slope one and intercept zero.


Scenario: PVE=0.05, theta0=−2, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.1, theta0=−2, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.05, theta0=−3, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.1, theta0=−3, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.05, theta0=−4, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.1, theta0=−4, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)


Scenario: PVE=0.2, theta0=−2, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.6, theta0=−2, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.2, theta0=−3, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.6, theta0=−3, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.2, theta0=−4, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)
Scenario: PVE=0.6, theta0=−4, theta=2
Ground truth
Competitive null
Pathway enriched
ASH log10(LR)
RSS log10(BF)


Scenario: PVE=0.05, theta0=−2, theta=2
Scenario: PVE=0.1, theta0=−2, theta=2
Scenario: PVE=0.05, theta0=−3, theta=2
Scenario: PVE=0.1, theta0=−3, theta=2
Scenario: PVE=0.05, theta0=−4, theta=2
Scenario: PVE=0.1, theta0=−4, theta=2
True positive fraction
False positive fraction
0.00
0.10
0.25
0.50
0.75
0.90
1.00
Method (AUC)
RSS (0.978)
ASH (0.985)
RSS (0.995)
ASH (0.998)
RSS (0.992)
ASH (0.981)
RSS (0.997)
ASH (0.995)
RSS (0.997)
ASH (0.948)
RSS (0.996)
ASH (0.965)


Scenario: PVE=0.2, theta0=−2, theta=2
Scenario: PVE=0.6, theta0=−2, theta=2
Scenario: PVE=0.2, theta0=−3, theta=2
Scenario: PVE=0.6, theta0=−3, theta=2
Scenario: PVE=0.2, theta0=−4, theta=2
Scenario: PVE=0.6, theta0=−4, theta=2
True positive fraction
False positive fraction
0.00 0.10 0.25 0.50 0.75 0.90 1.00
Method (AUC)
RSS (1.000)
ASH (1.000)
RSS (0.992)
ASH (0.978)
RSS (0.995)
ASH (0.986)

## Supplementary Figure 19

**Bayes factors for enrichment of genetic associations near all genes in 31 phenotypes.** For each phenotype, we use RSS to assess the genome-wide near-gene enrichment hypothesis that SNPs within ± 100kb of the transcribed region of any autosomal protein-coding gene (18,313 in total in the present study; see **Supplementary Figure 16**) are more likely to be associated with the phenotype. The $x$-axis shows Bayes factor (BF) for each genome-wide near-gene hypothesis. The $x$-axis uses a normal scale inside the range [-1.5, 1.5], and a logarithmic scale (base 10) outside (-1.5, 1.5).

Specifically, we consider

- Model $M_0$: each SNP has equal chance of being associated with the phenotype;
- Model $M_1$: SNPs inside a target gene set are more often associated with the phenotype;
- Model $M_2$: near-gene SNPs are more often associated with the phenotype.

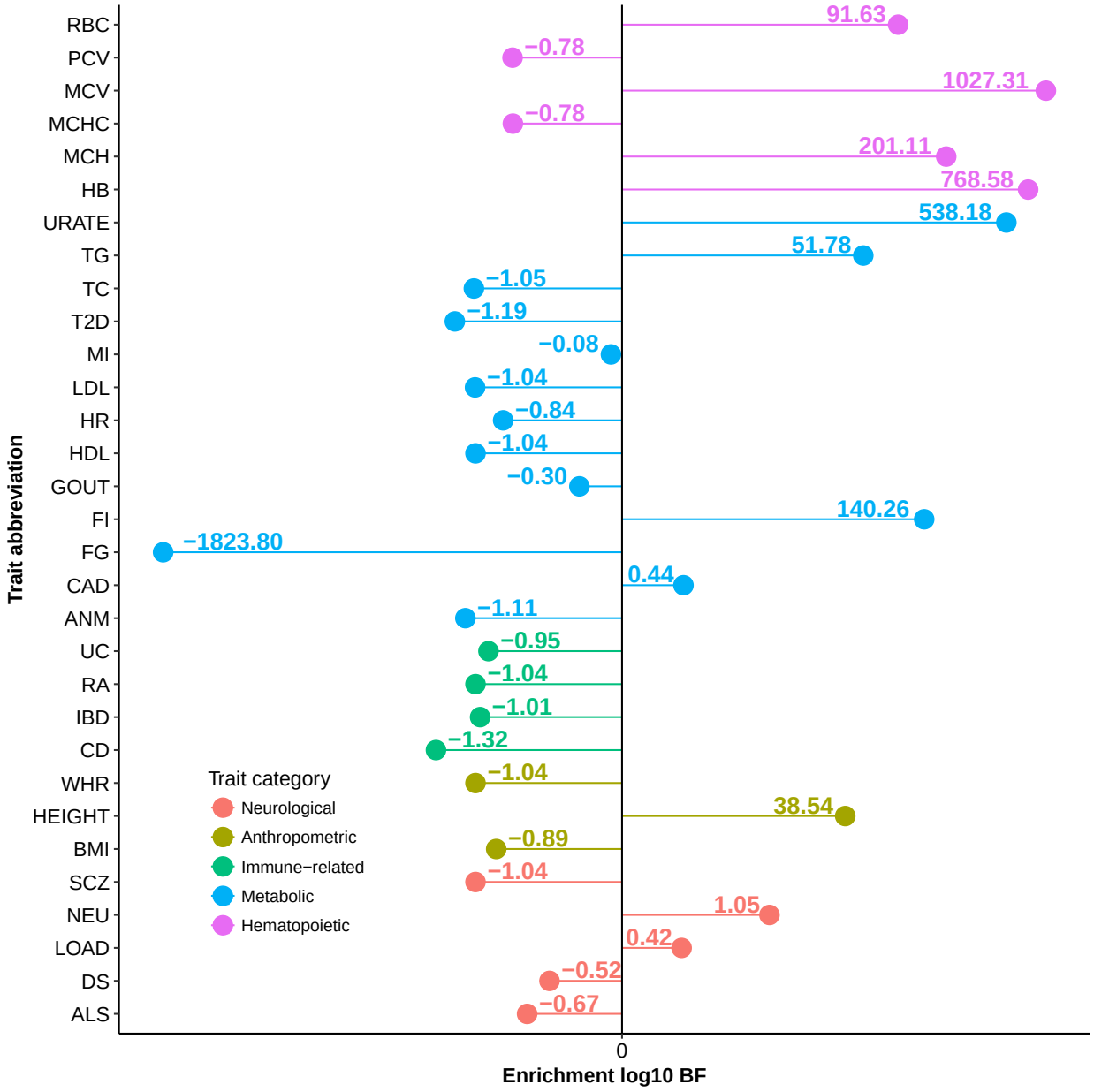
In the main text we compute the following BF to assess the enrichment of target gene set

$$\mathrm{BF}(M_1|M_0) = \Pr(\mathrm{Data}|M_1) \,/\, \Pr(\mathrm{Data}|M_0).$$

Here we compute a BF to assess the genome-wide near-gene enrichment:

$$\mathrm{BF}(M_2|M_0) = \Pr(\mathrm{Data}|M_2) \,/\, \Pr(\mathrm{Data}|M_0).$$

For a given phenotype, if $\mathrm{BF}(M_1|M_0) > \mathrm{BF}(M_2|M_0)$, then the observed GWAS summary statistics show more support for the target gene set enrichment than near-gene enrichment, and vice versa. See Carbonetto and Stephens (2013) for extensive discussions of interpreting enrichment BFs.


RBC
PCV
MCV
MCHC
MCH
HB
URATE
TG
TC
T2D
MI
LDL
HR
HDL
GOUT
FI
FG
CAD
ANM
UC
RA
IBD
CD
WHR
HEIGHT
BMI
SCZ
NEU
LOAD
DS
ALS
Trait abbreviation
91.63
−0.78
1027.31
−0.78
201.11
768.58
538.18
51.78
−1.05
−1.19
−0.08
−1.04
−0.84
−1.04
−0.30
140.26
−1823.80
0.44
−1.11
−0.95
−1.04
−1.01
−1.32
−1.04
38.54
−0.89
−1.04
1.05
0.42
−0.52
−0.67
Trait category
Neurological
Anthropometric
Immune−related
Metabolic
Hematopoietic
0
Enrichment log10 BF

## Supplementary Figure 20

**Enrichment analyses of randomly selected near-gene SNPs based on real GWAS summary statistics.** For a given dataset of GWAS summary statistics, we select a target gene set from the top enriched biological pathways or tissue-based gene sets, and then create "null" sets by randomly drawing near-gene SNPs (defined in **Supplementary Figure 2**), the sizes of which roughly match the target gene set. We perform enrichment analysis of these random sets of near-gene SNPs on the same GWAS summary data, and compare their enrichment Bayes factors (BFs) with BFs of the actual gene sets. We also use these simulated BFs to estimate false discovery rates at a given cutoff of BF (see the R codes below).

```
# a naive FDR estimate function
# "observed": the real gene sets
# "simulated": the random sets of near-gene SNPs

estimate_fdr <- function(log10.bf, case, cutoff) {
  R <- sum(log10.bf[case=="observed"] >= cutoff) / sum(case=="observed")
  V <- sum(log10.bf[case=="simulated"] >= cutoff) / sum(case=="simulated")
  FDR <- V/R
  return(FDR)
}
```

We perform this type of simulation on the following two examples.

**Example 1**

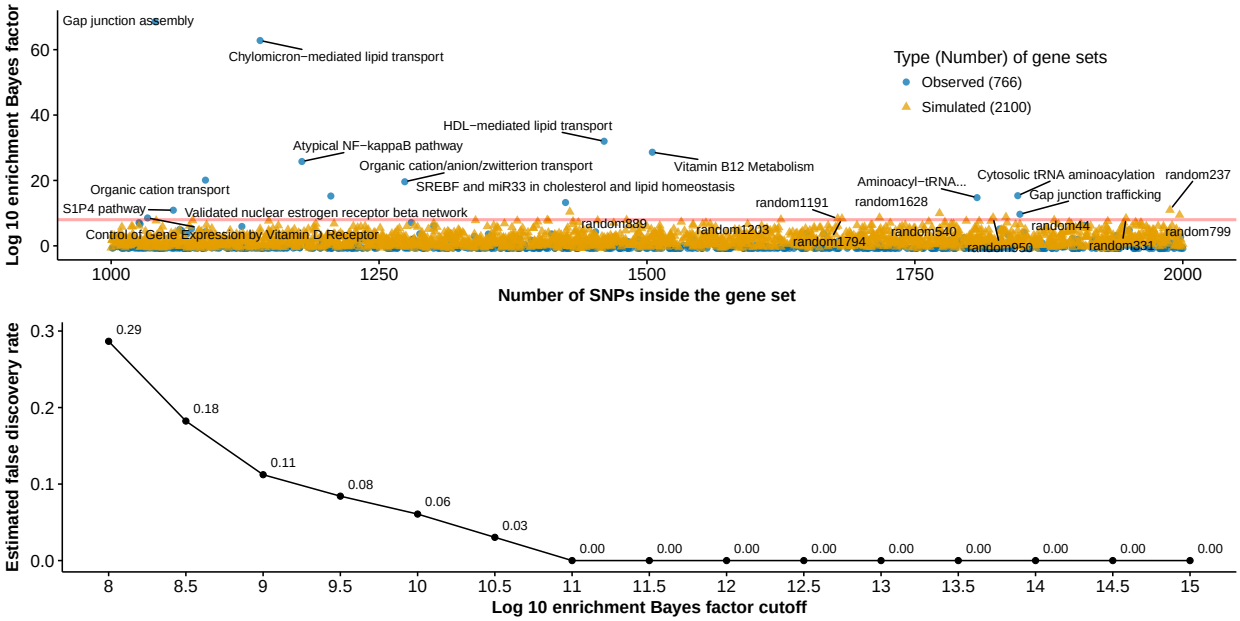
- GWAS summary statistics: low-density lipoprotein (Teslovich et al. 2010).
- Target gene set: *Chylomicron-mediated lipid transport* (Reactome, Pathway Commons 2), 1139 SNPs in this pathway, $\log_{10}\text{BF} = 62.8$.
- Random sets of near-gene SNPs: 2100 sets, set size ~ Uniform(1000, 2000).

**Example 2**

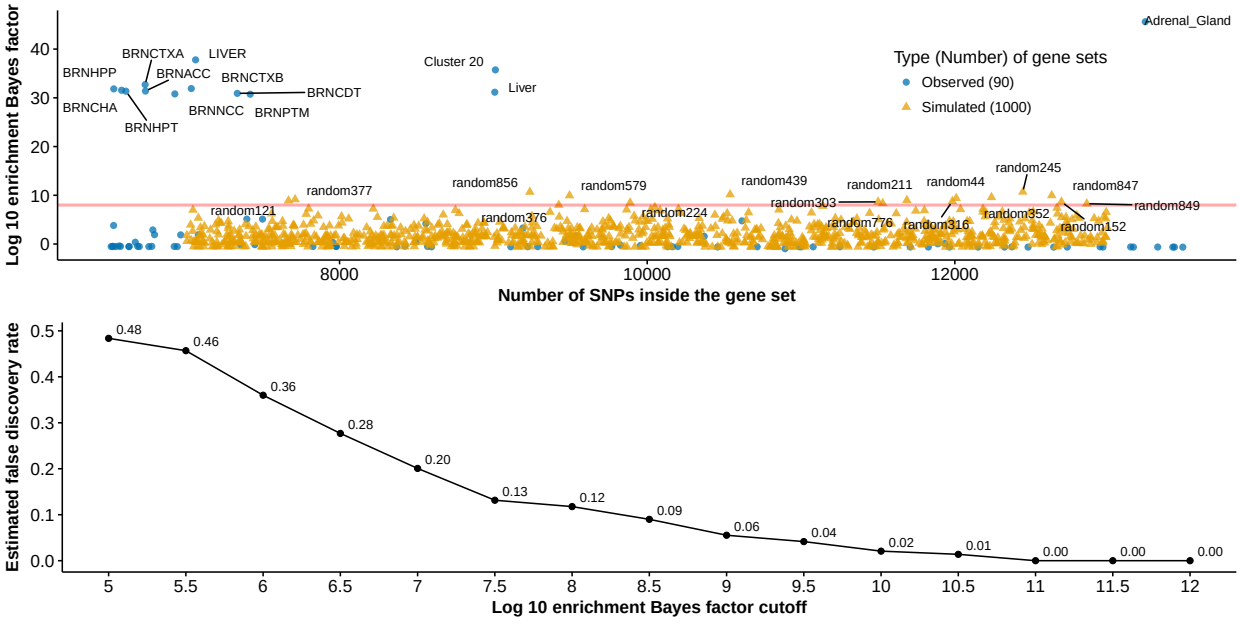
- GWAS summary statistics: late-onset Alzheimer's disease (Lambert et al. 2013).
- Target gene set: *Liver distinctive cluster* (Dey, Hsiao, and Stephens 2017), 9013 SNPs in this gene set, $\log_{10}\text{BF} = 35.7$.
- Random sets of near-gene SNPs: 1000 sets, set size ~ Uniform(7000, 13000).

The blue dots indicate the real gene sets. The yellow triangles indicate the random sets of near-gene SNPs. The red horizontal line indicate $\log_{10}\text{BF} = 8$.

## Example 1




## Example 2



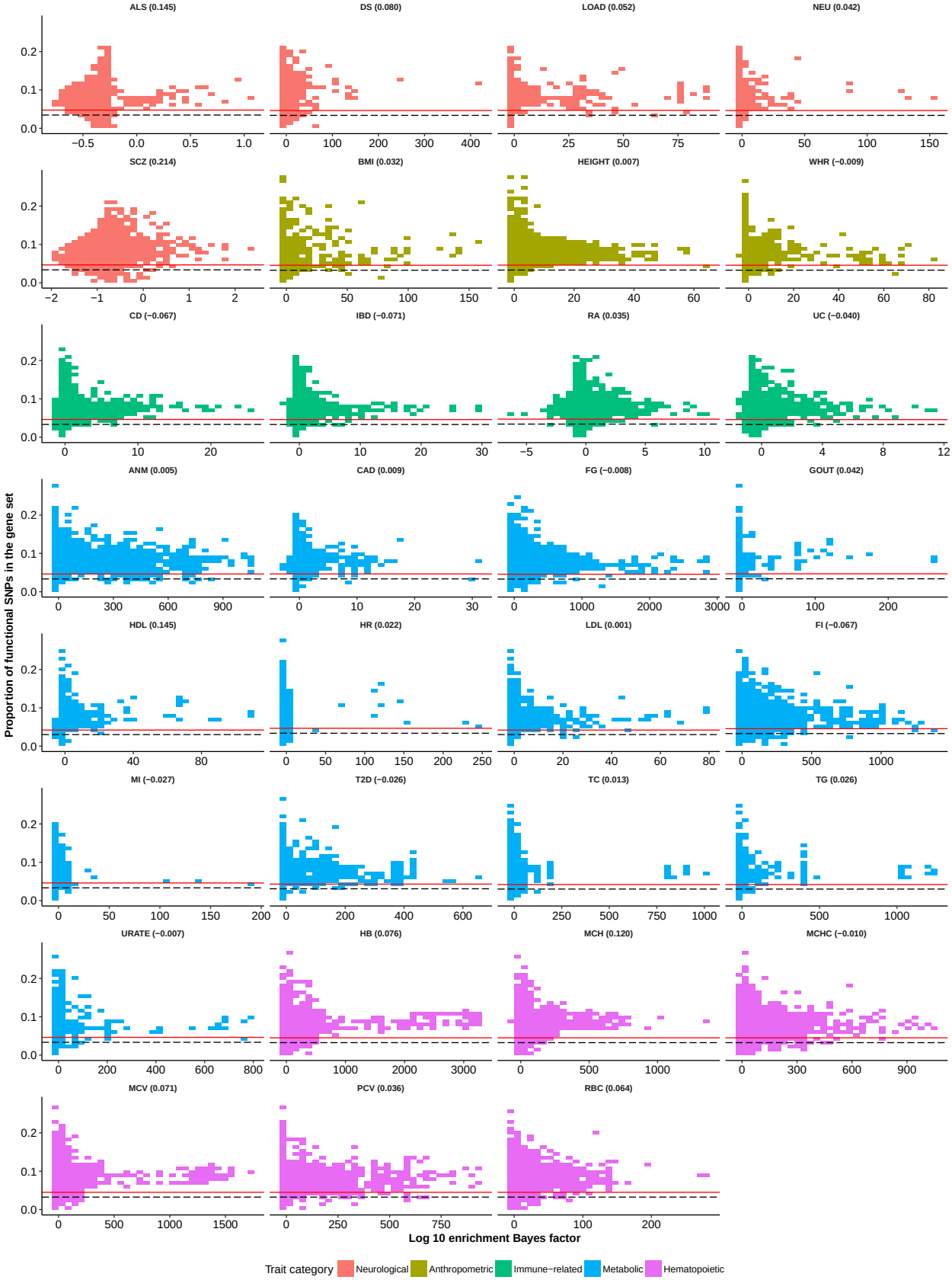
# Supplementary Figure 21

**Correlation between gene set enrichment Bayes factors and 52 functional categories in 31 phenotypes.** For each pair of phenotype and functional category, we plot the log 10 enrichment Bayes factor versus the proportion of functional SNPs for each of 4,026 gene sets (3,913 biological pathways and 113 tissue-based gene sets), and compute their Pearson correlation. In each plot, the black dashed line indicates the proportion of genome-wide SNPs falling into the given functional category; the red solid line indicates the proportion of near-gene SNPs (defined in **Supplementary Figure 2**) falling into the given functional category.

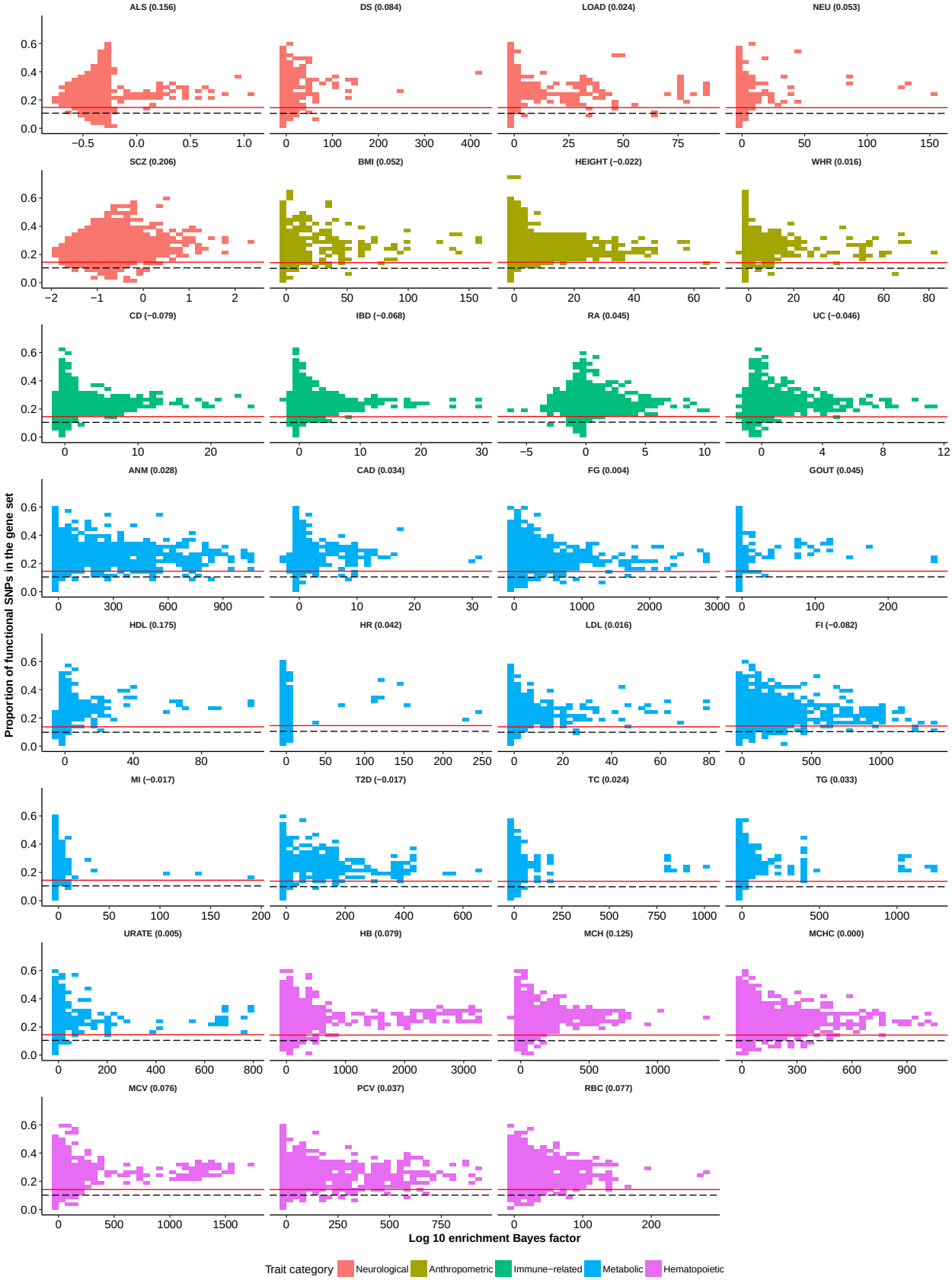
Below are the 52 functional categories used in Finucane et al. (2015).

```
##  [1] "Coding_UCSC"                      "Coding_UCSC.extend.500"
##  [3] "Conserved_LindbladToh"            "Conserved_LindbladToh.extend.500"
##  [5] "CTCF_Hoffman"                     "CTCF_Hoffman.extend.500"
##  [7] "DGF_ENCODE"                       "DGF_ENCODE.extend.500"
##  [9] "DHS_peaks_Trynka"                 "DHS_Trynka"
## [11] "DHS_Trynka.extend.500"            "Enhancer_Andersson"
## [13] "Enhancer_Andersson.extend.500"    "Enhancer_Hoffman"
## [15] "Enhancer_Hoffman.extend.500"      "FetalDHS_Trynka"
## [17] "FetalDHS_Trynka.extend.500"       "H3K27ac_Hnisz"
## [19] "H3K27ac_Hnisz.extend.500"         "H3K27ac_PGC2"
## [21] "H3K27ac_PGC2.extend.500"          "H3K4me1_peaks_Trynka"
## [23] "H3K4me1_Trynka"                   "H3K4me1_Trynka.extend.500"
## [25] "H3K4me3_peaks_Trynka"             "H3K4me3_Trynka"
## [27] "H3K4me3_Trynka.extend.500"        "H3K9ac_peaks_Trynka"
## [29] "H3K9ac_Trynka"                    "H3K9ac_Trynka.extend.500"
## [31] "Intron_UCSC"                      "Intron_UCSC.extend.500"
## [33] "PromoterFlanking_Hoffman"         "PromoterFlanking_Hoffman.extend.500"
## [35] "Promoter_UCSC"                    "Promoter_UCSC.extend.500"
## [37] "Repressed_Hoffman"                "Repressed_Hoffman.extend.500"
## [39] "SuperEnhancer_Hnisz"              "SuperEnhancer_Hnisz.extend.500"
## [41] "TFBS_ENCODE"                      "TFBS_ENCODE.extend.500"
## [43] "Transcribed_Hoffman"              "Transcribed_Hoffman.extend.500"
## [45] "TSS_Hoffman"                      "TSS_Hoffman.extend.500"
## [47] "UTR_3_UCSC"                       "UTR_3_UCSC.extend.500"
## [49] "UTR_5_UCSC"                       "UTR_5_UCSC.extend.500"
## [51] "WeakEnhancer_Hoffman"             "WeakEnhancer_Hoffman.extend.500"
```

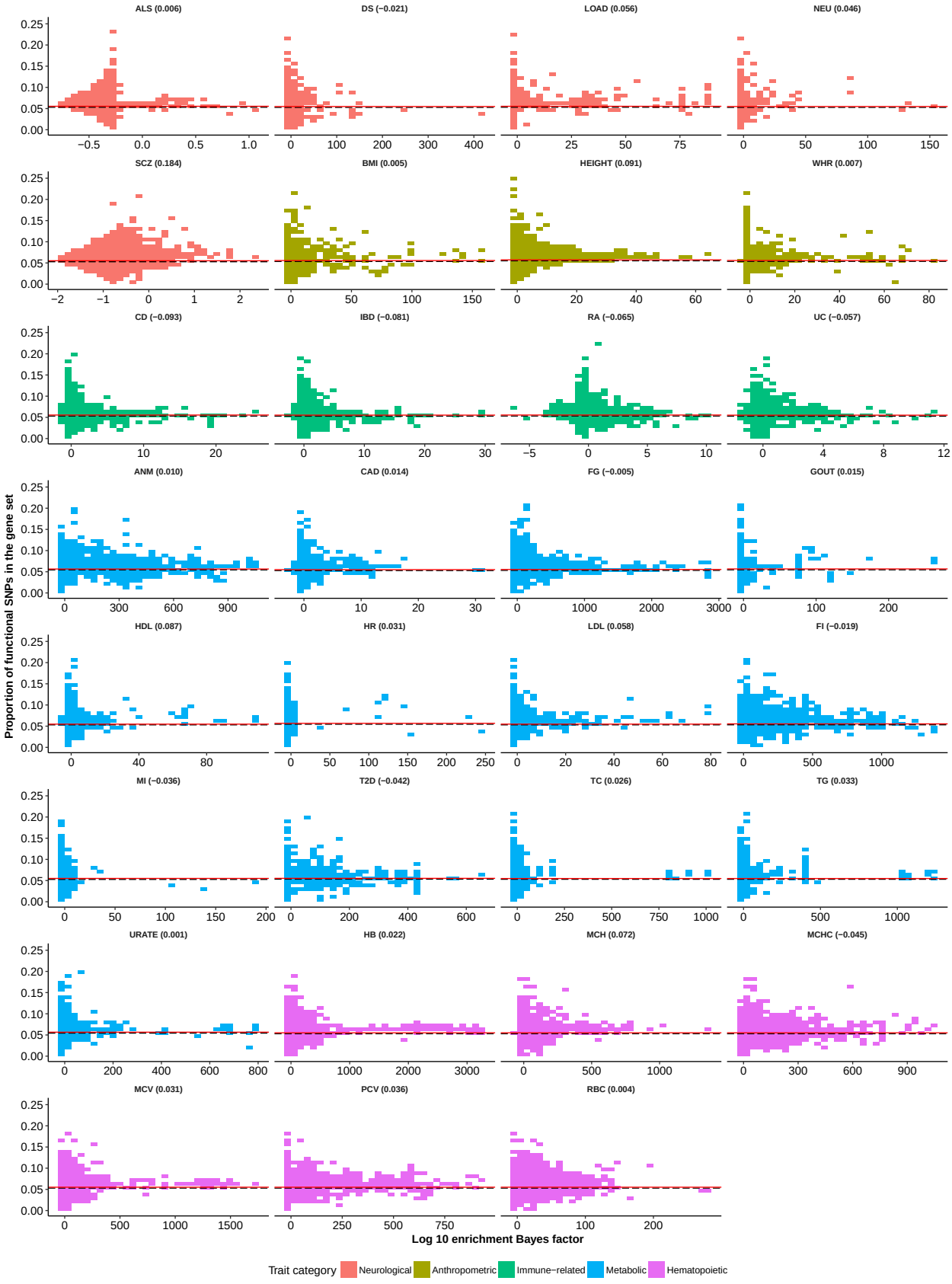
## Functional category: Coding_UCSC



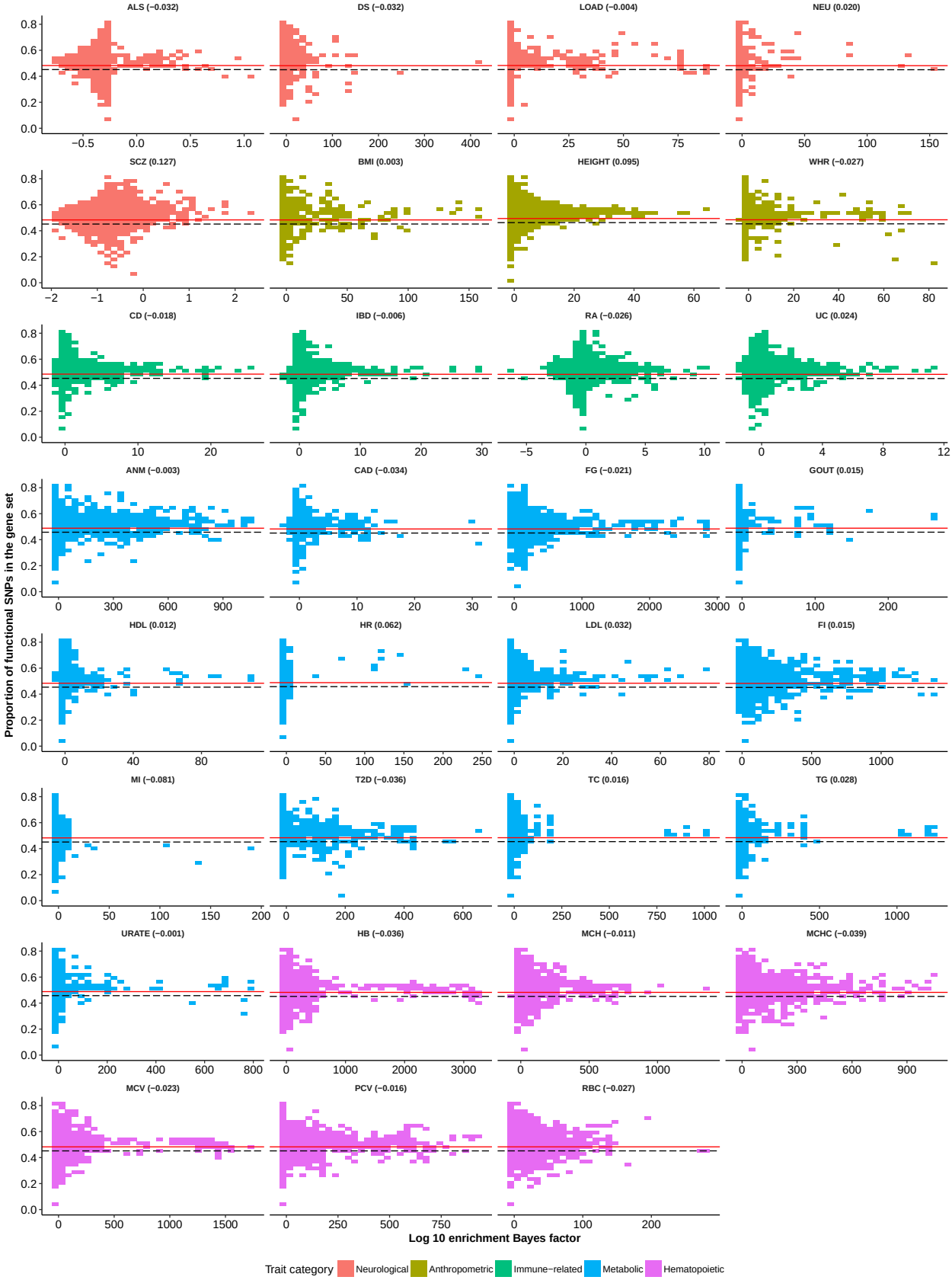
## Functional category: Coding_UCSC.extend.500



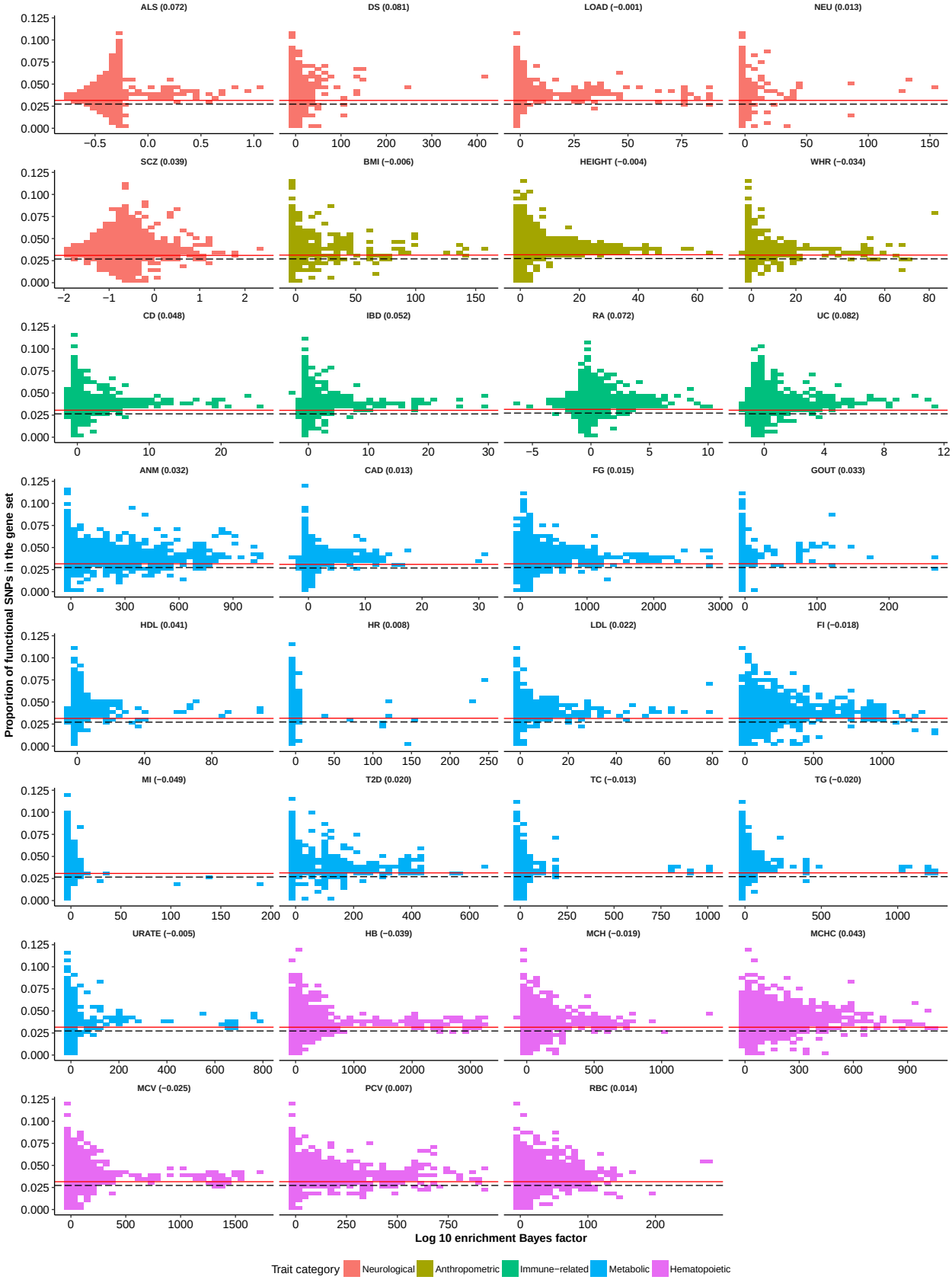
## Functional category: Conserved_LindbladToh



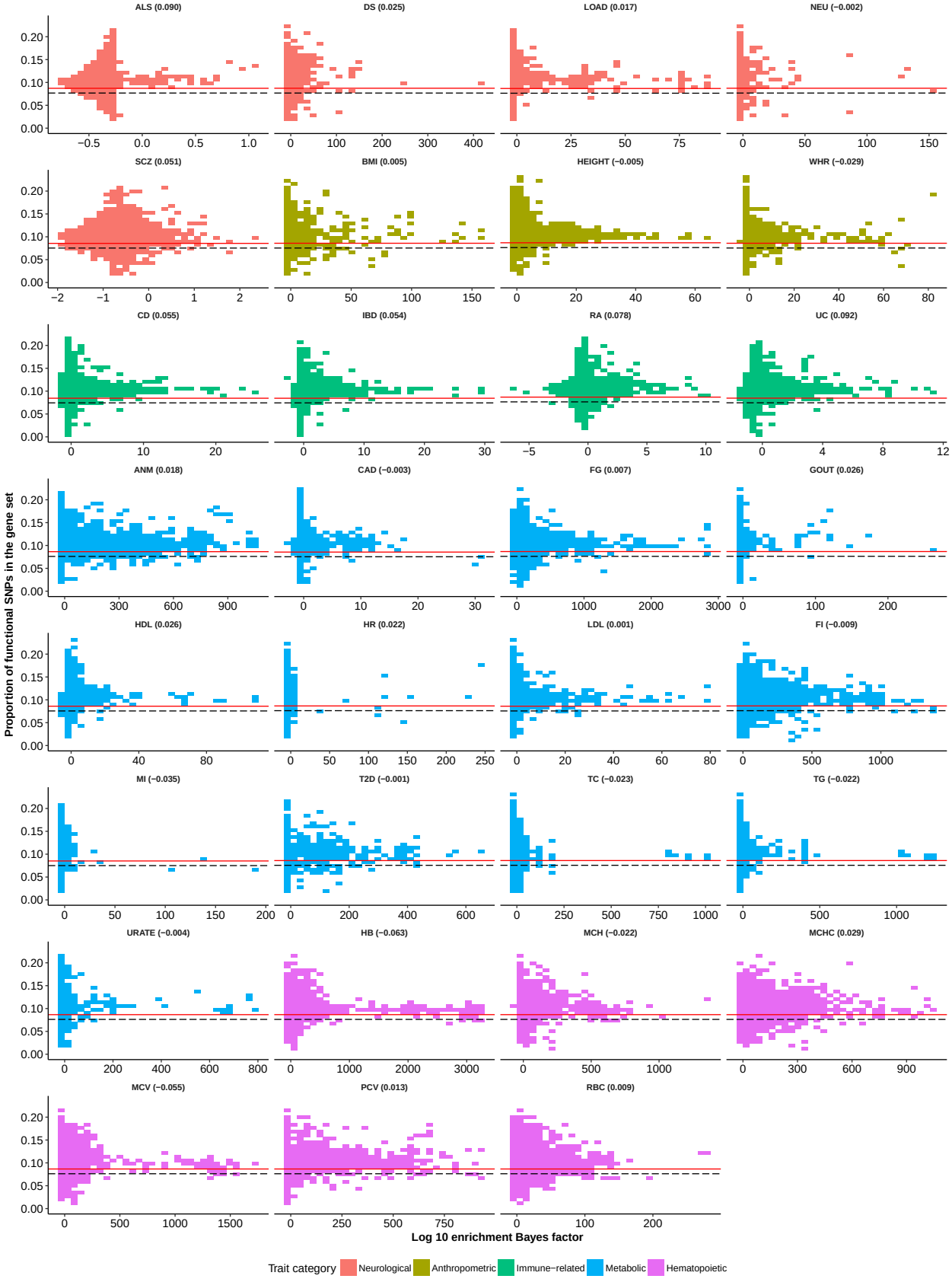
## Functional category: Conserved_LindbladToh.extend.500



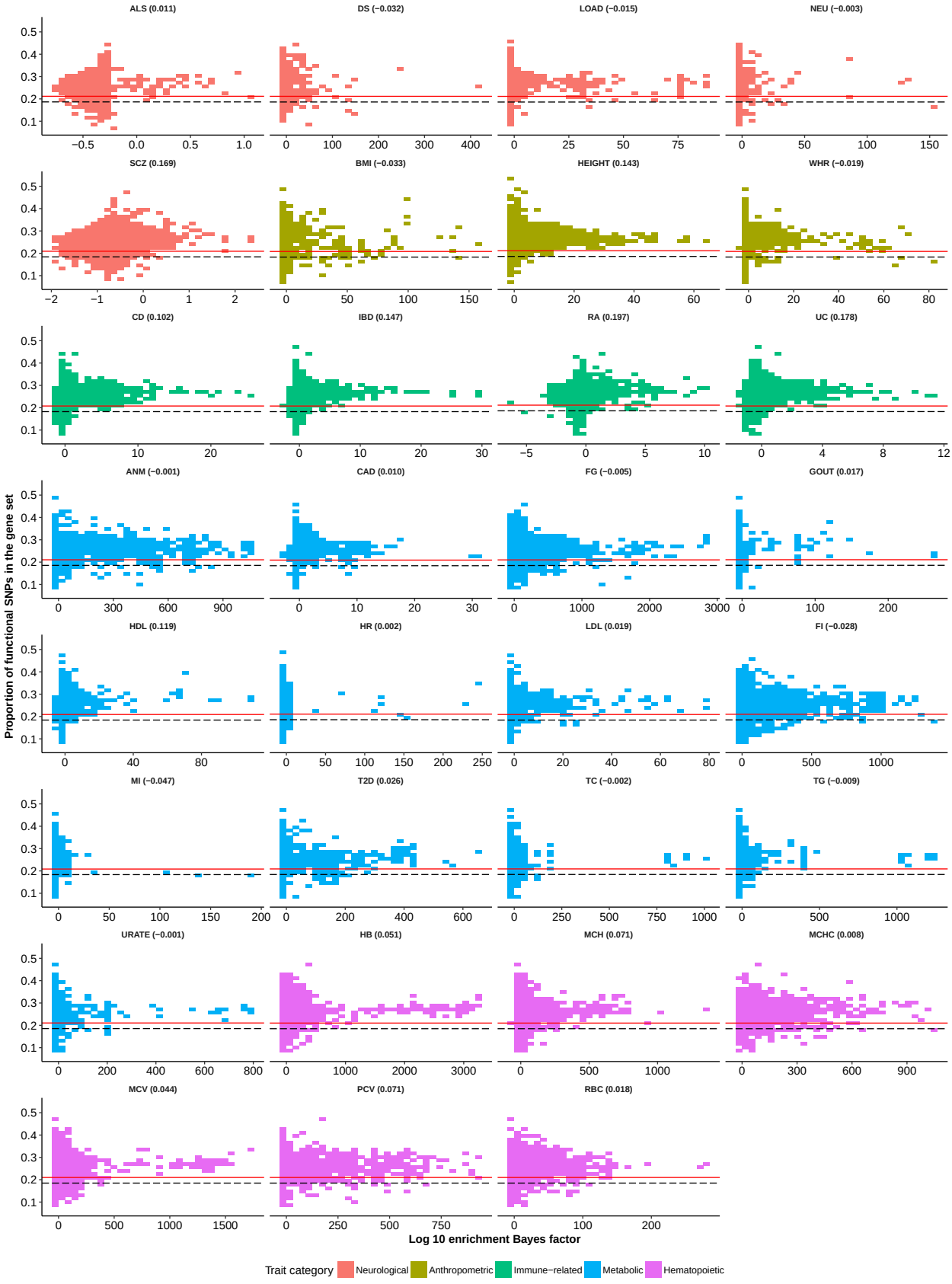
## Functional category: CTCF_Hoffman



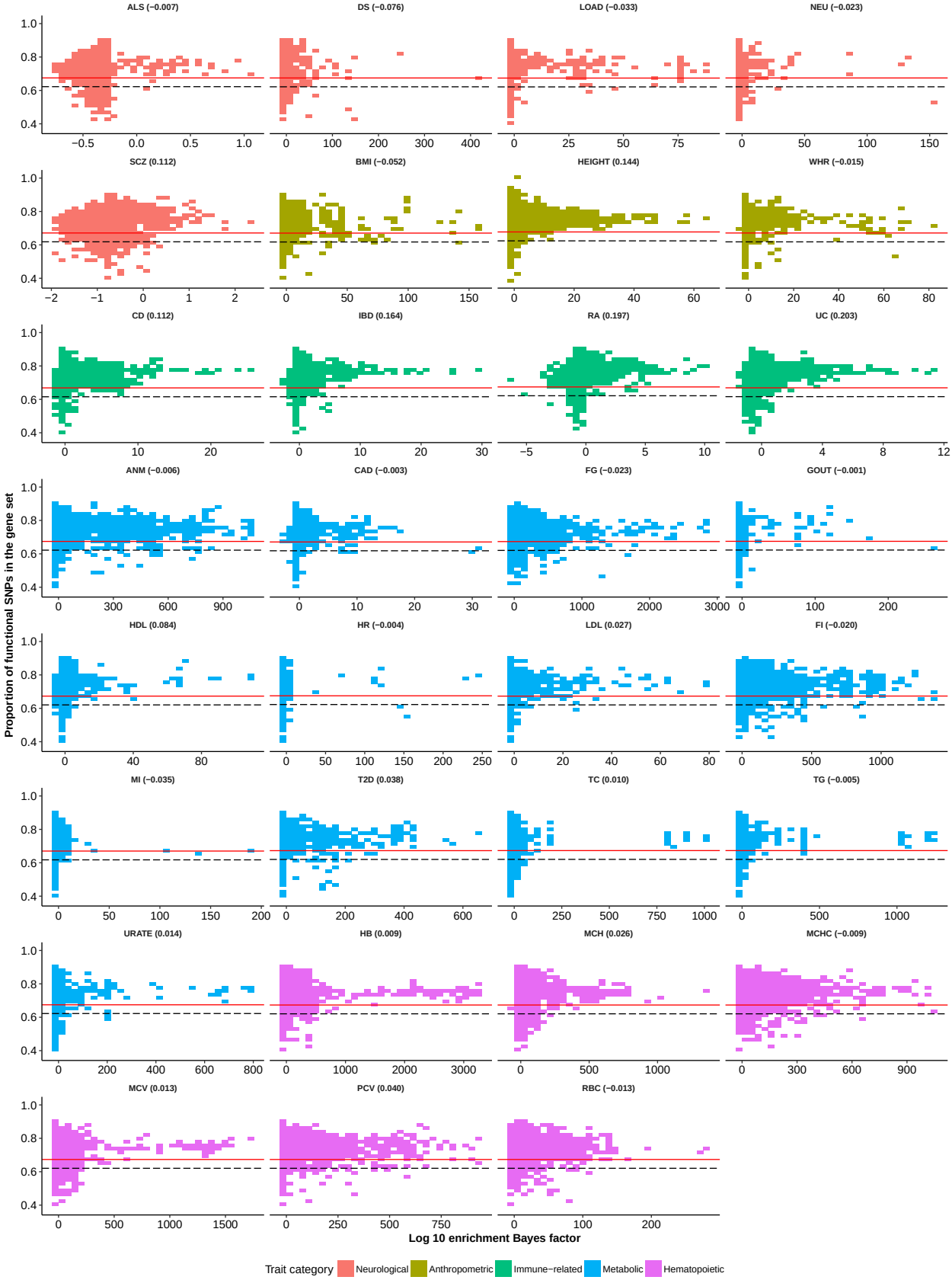
## Functional category: CTCF_Hoffman.extend.500



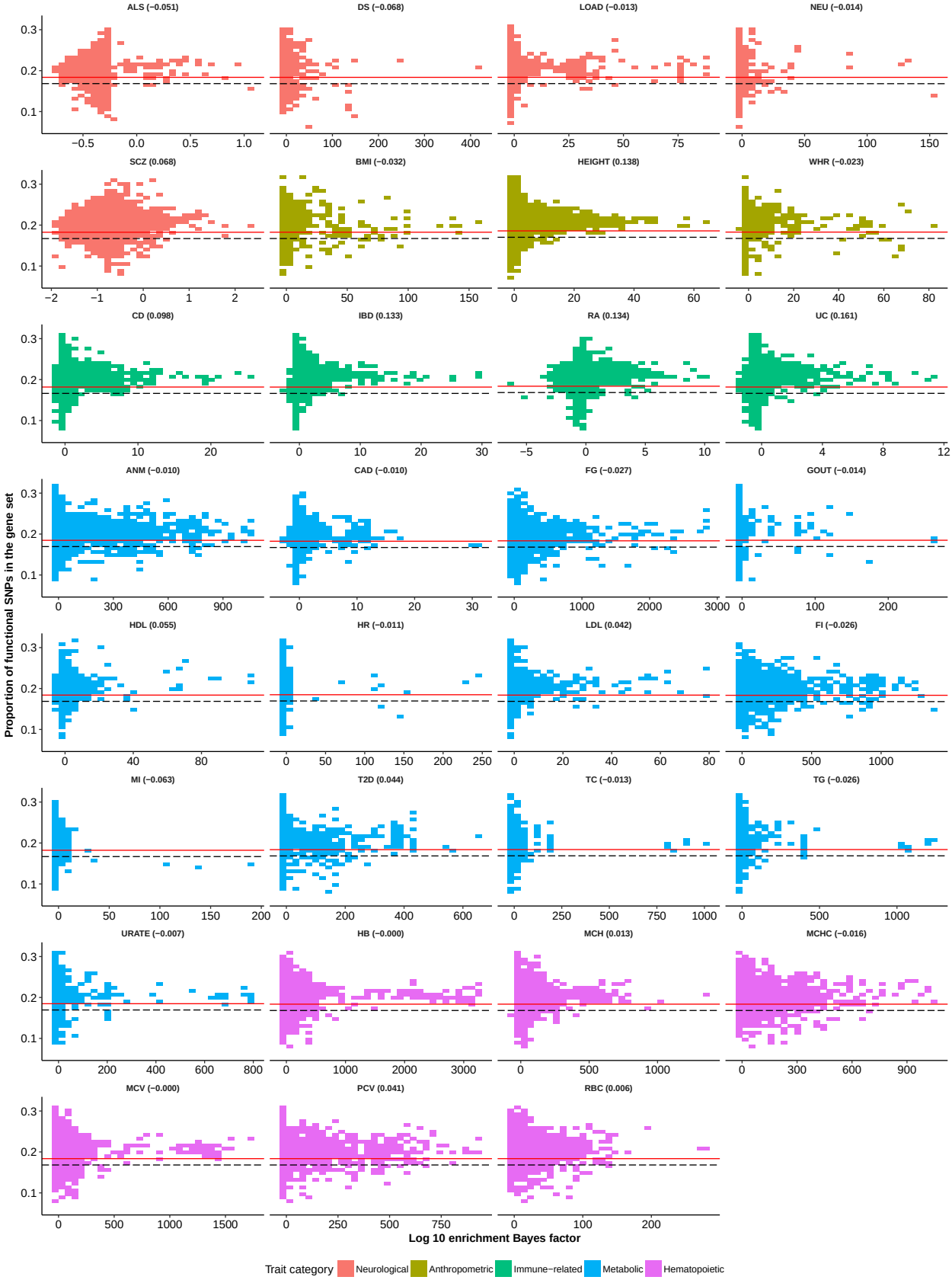
## Functional category: DGF_ENCODE



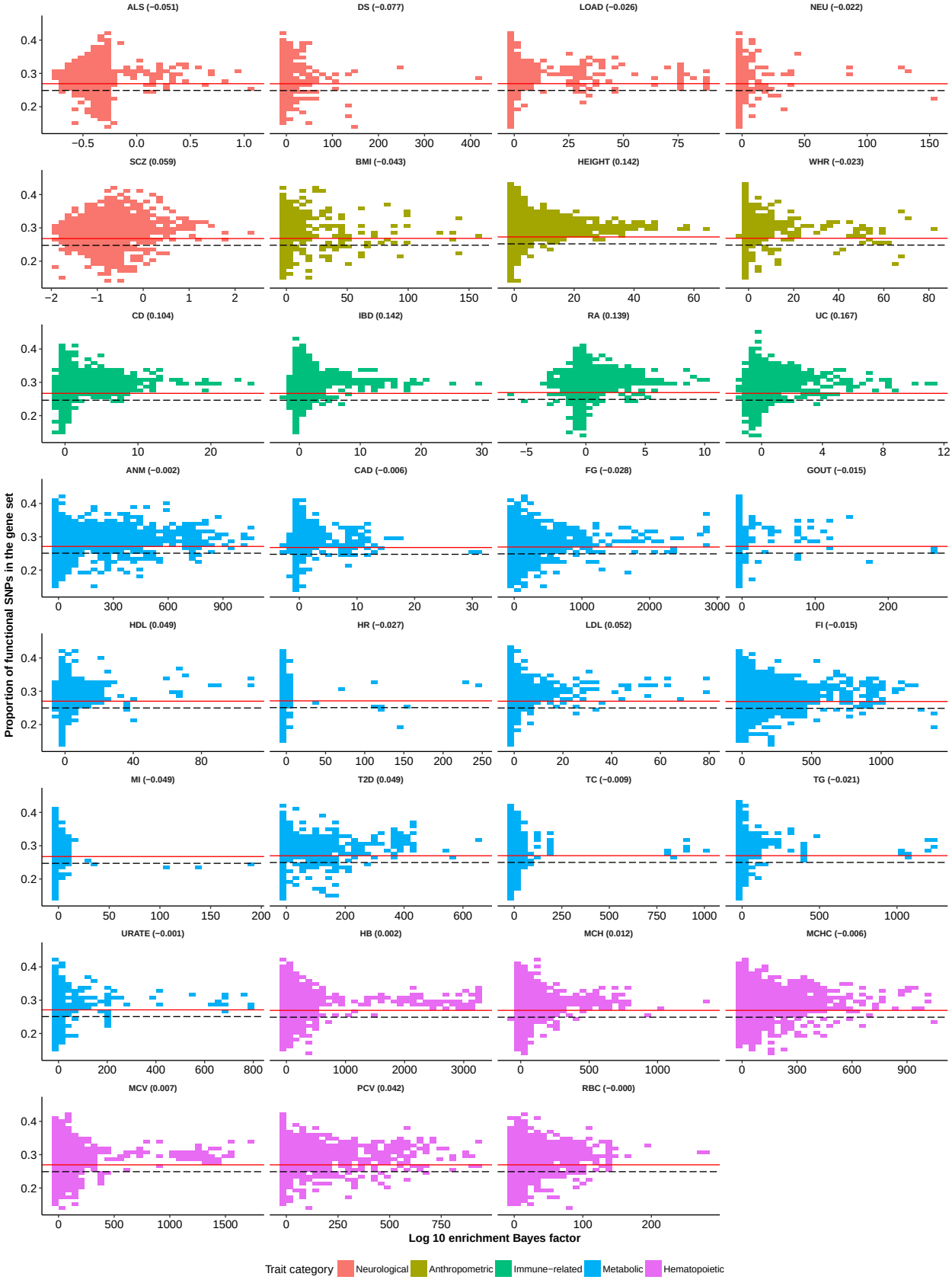
## Functional category: DGF_ENCODE.extend.500



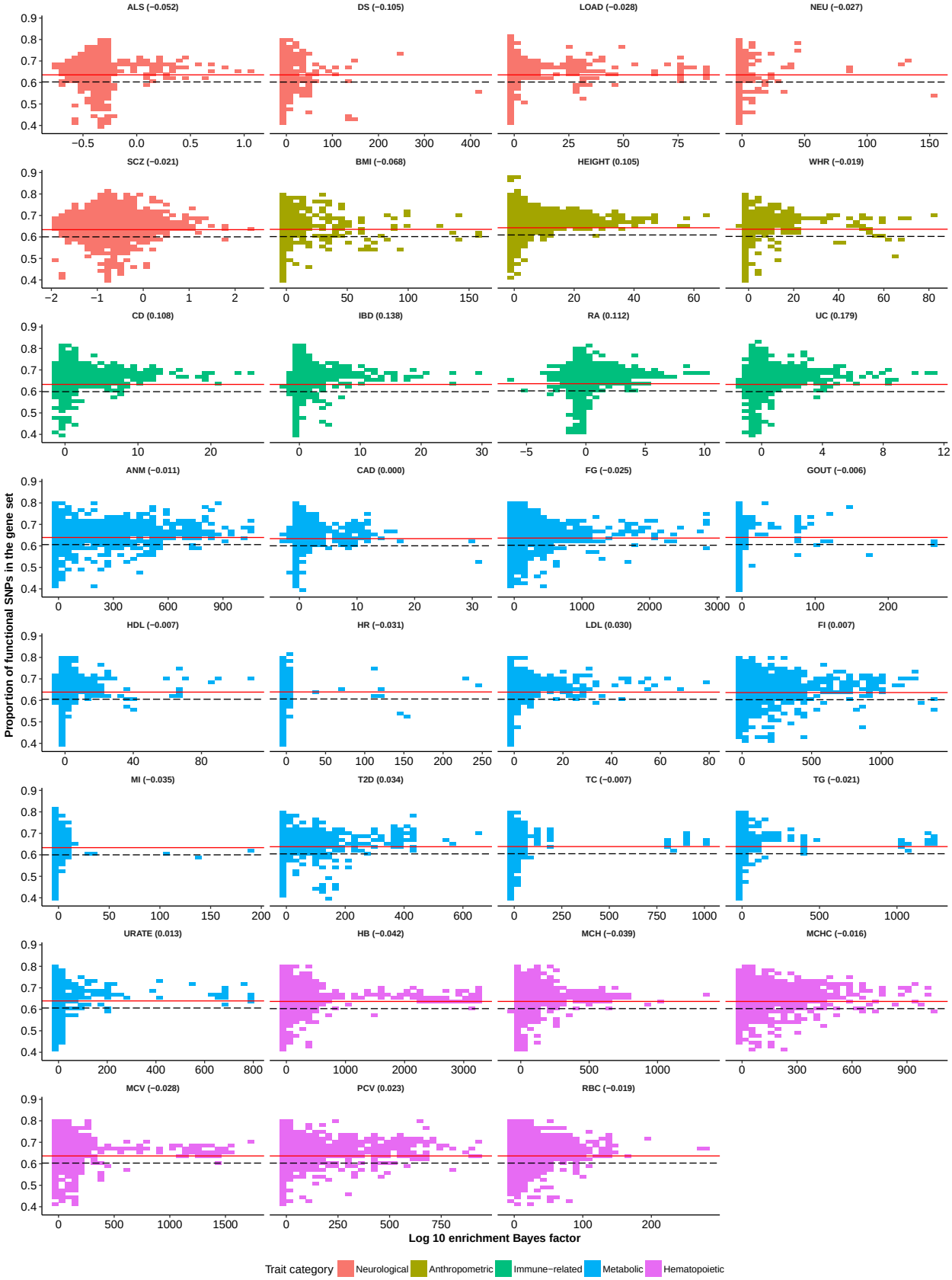
## Functional category: DHS_peaks_Trynka



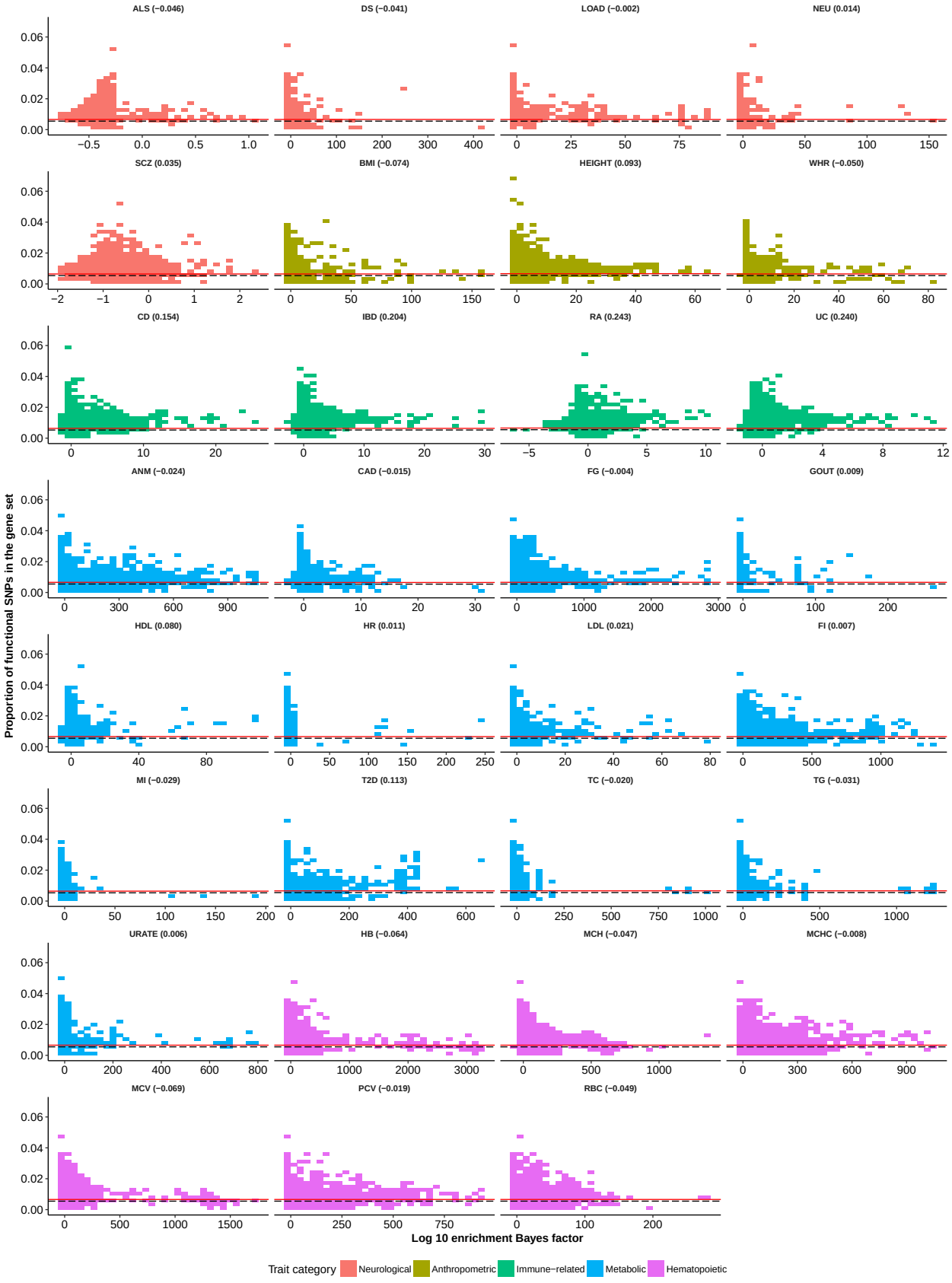
## Functional category: DHS_Trynka



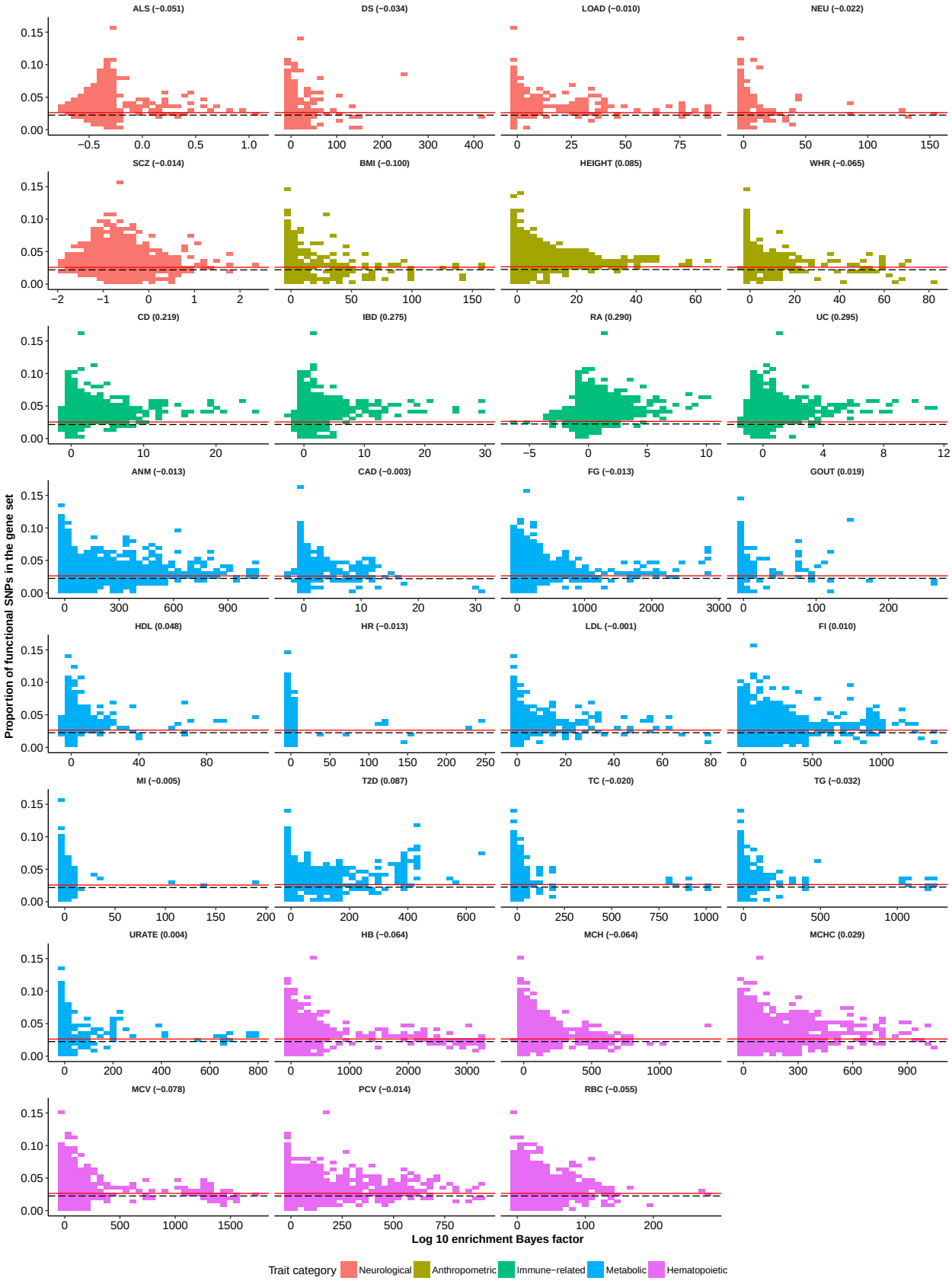
## Functional category: DHS_Trynka.extend.500



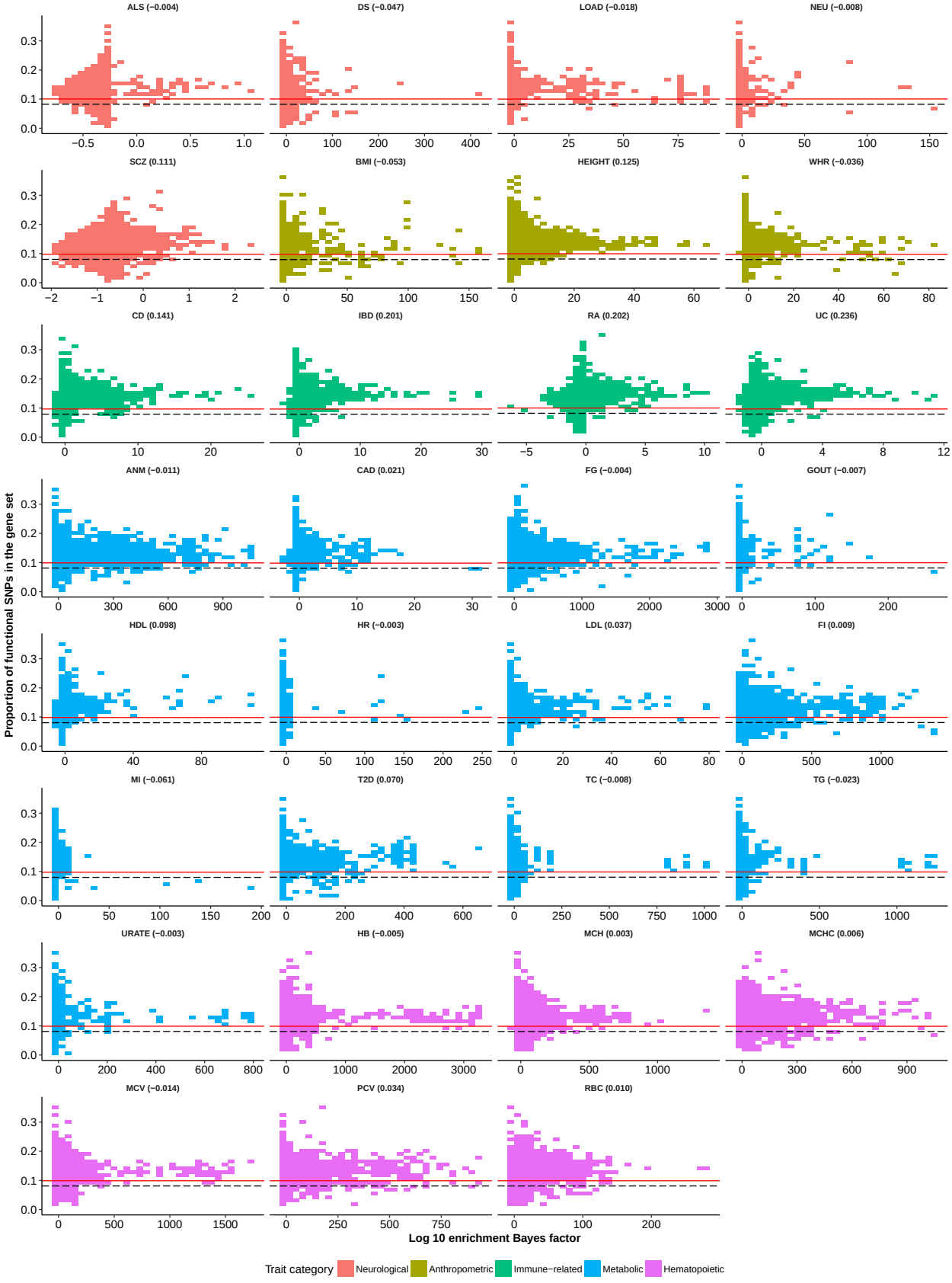
## Functional category: Enhancer_Andersson



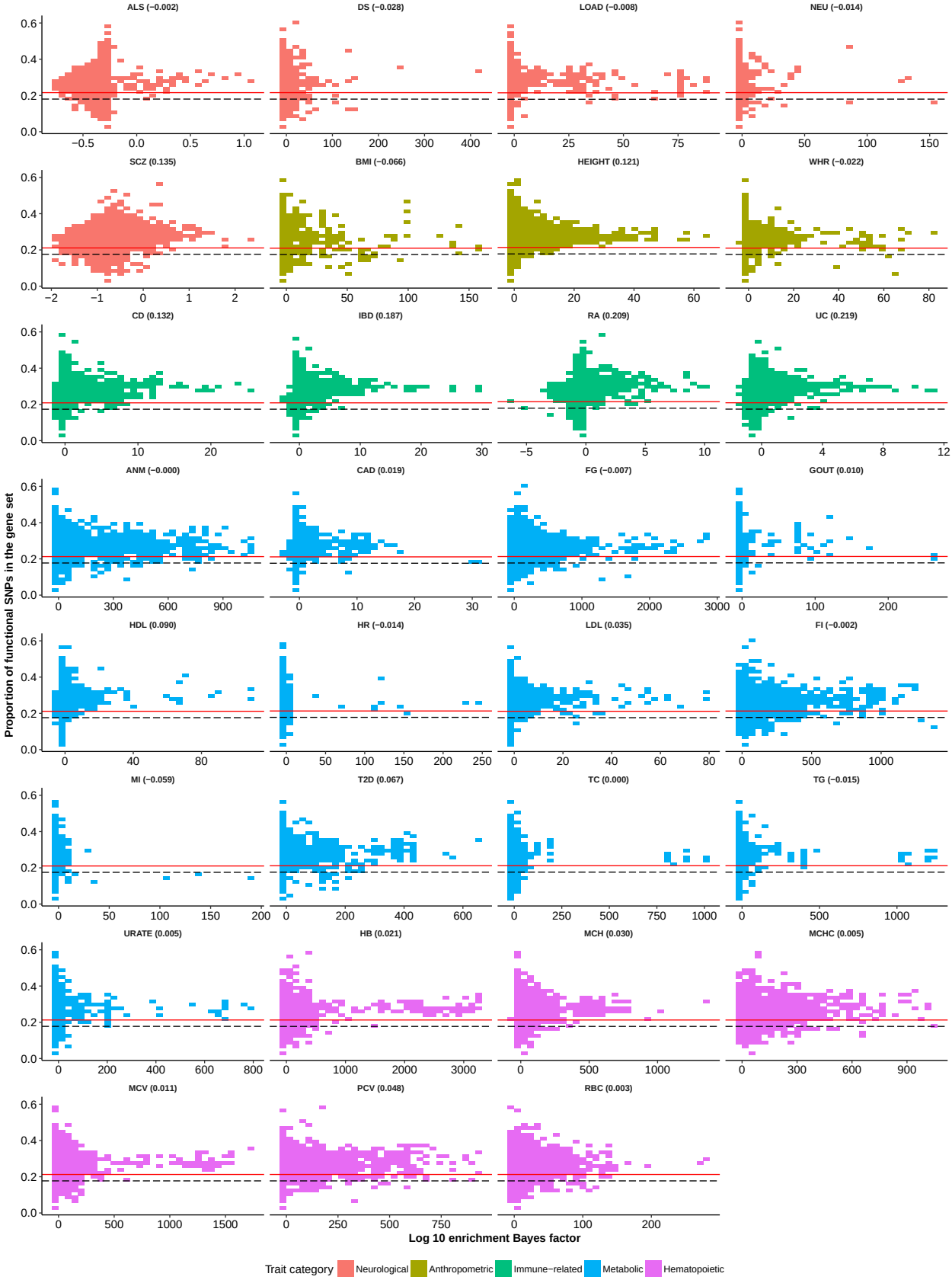
## Functional category: Enhancer_Andersson.extend.500



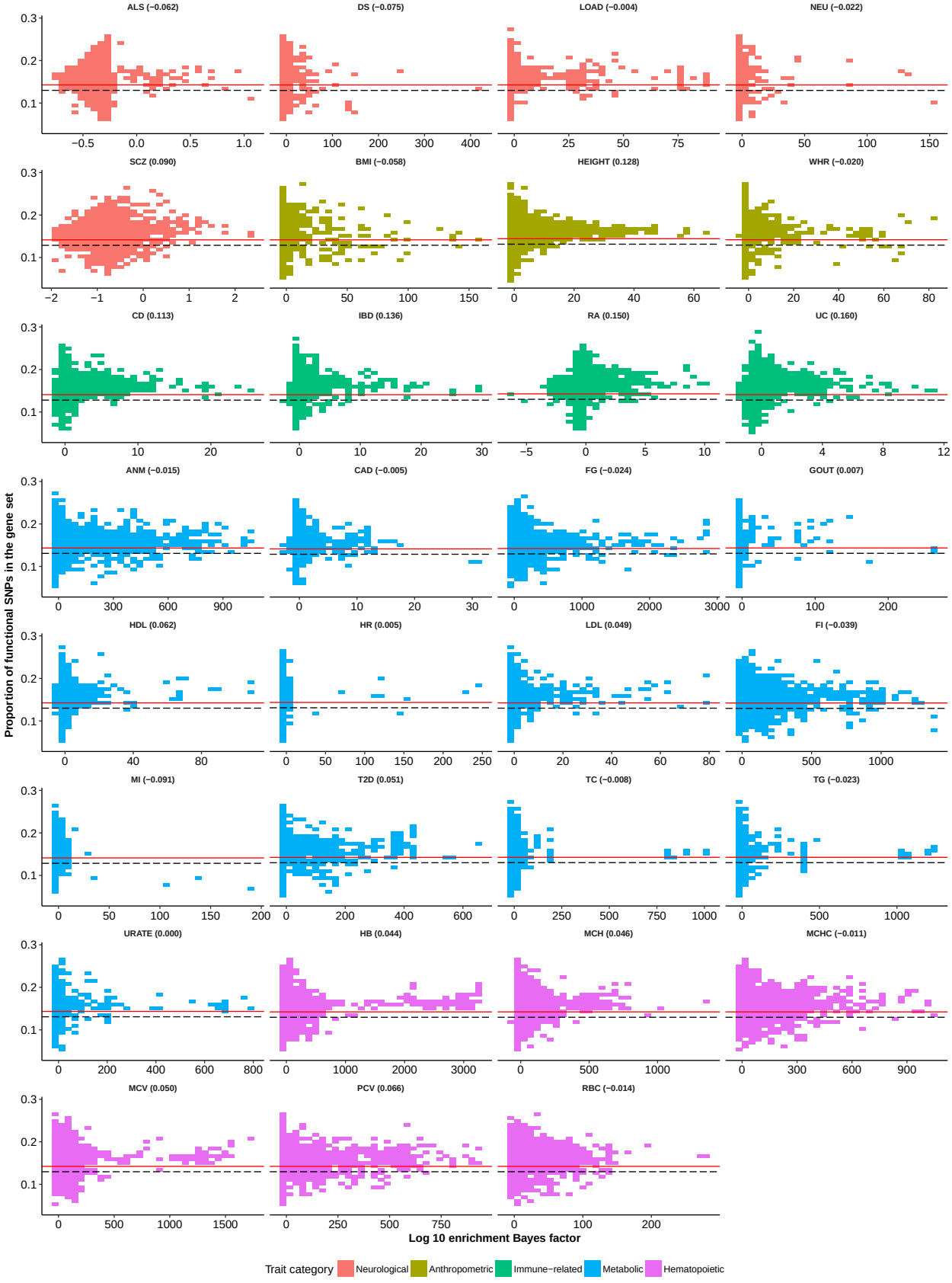
## Functional category: Enhancer_Hoffman



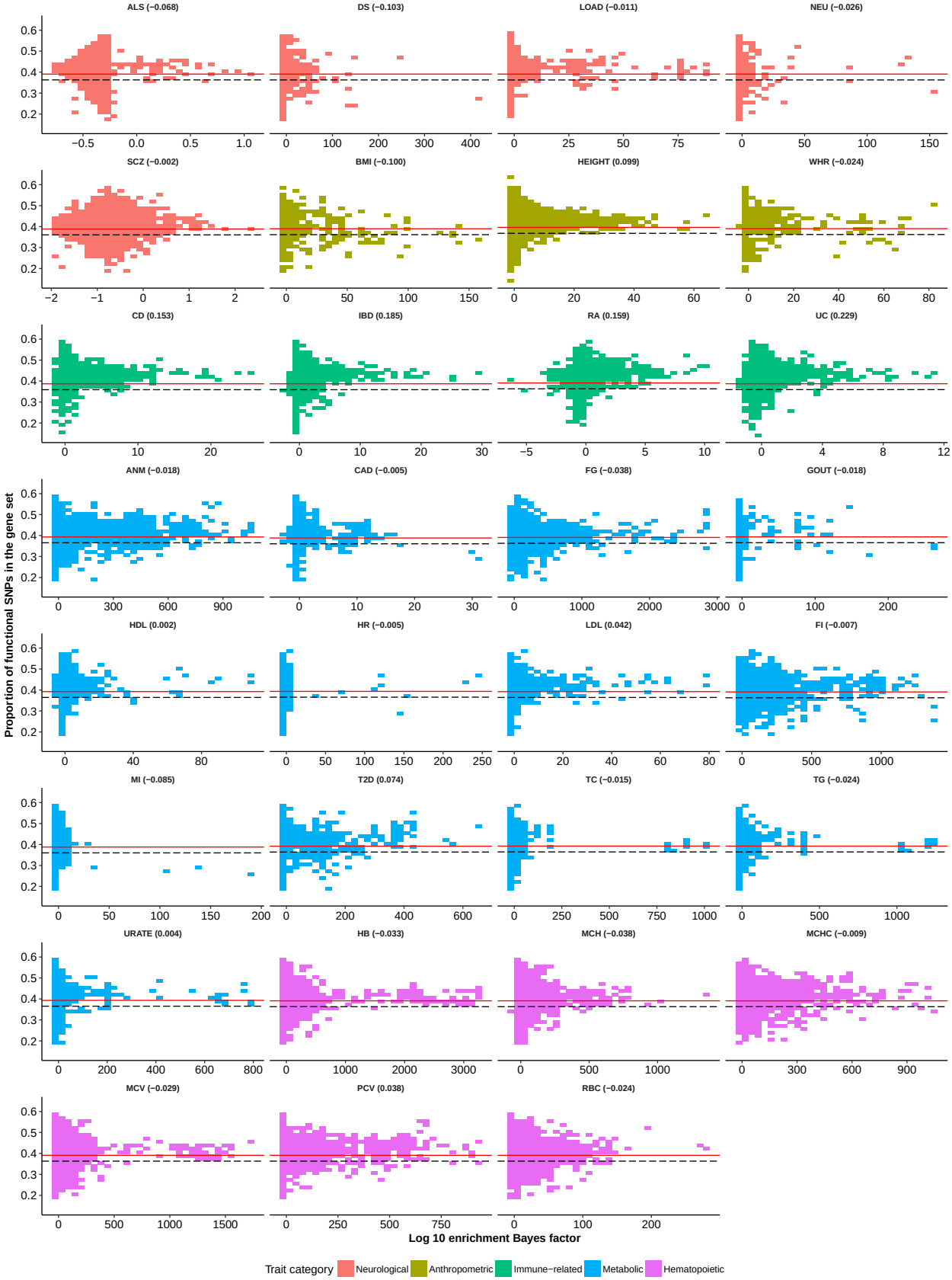
## Functional category: Enhancer_Hoffman.extend.500



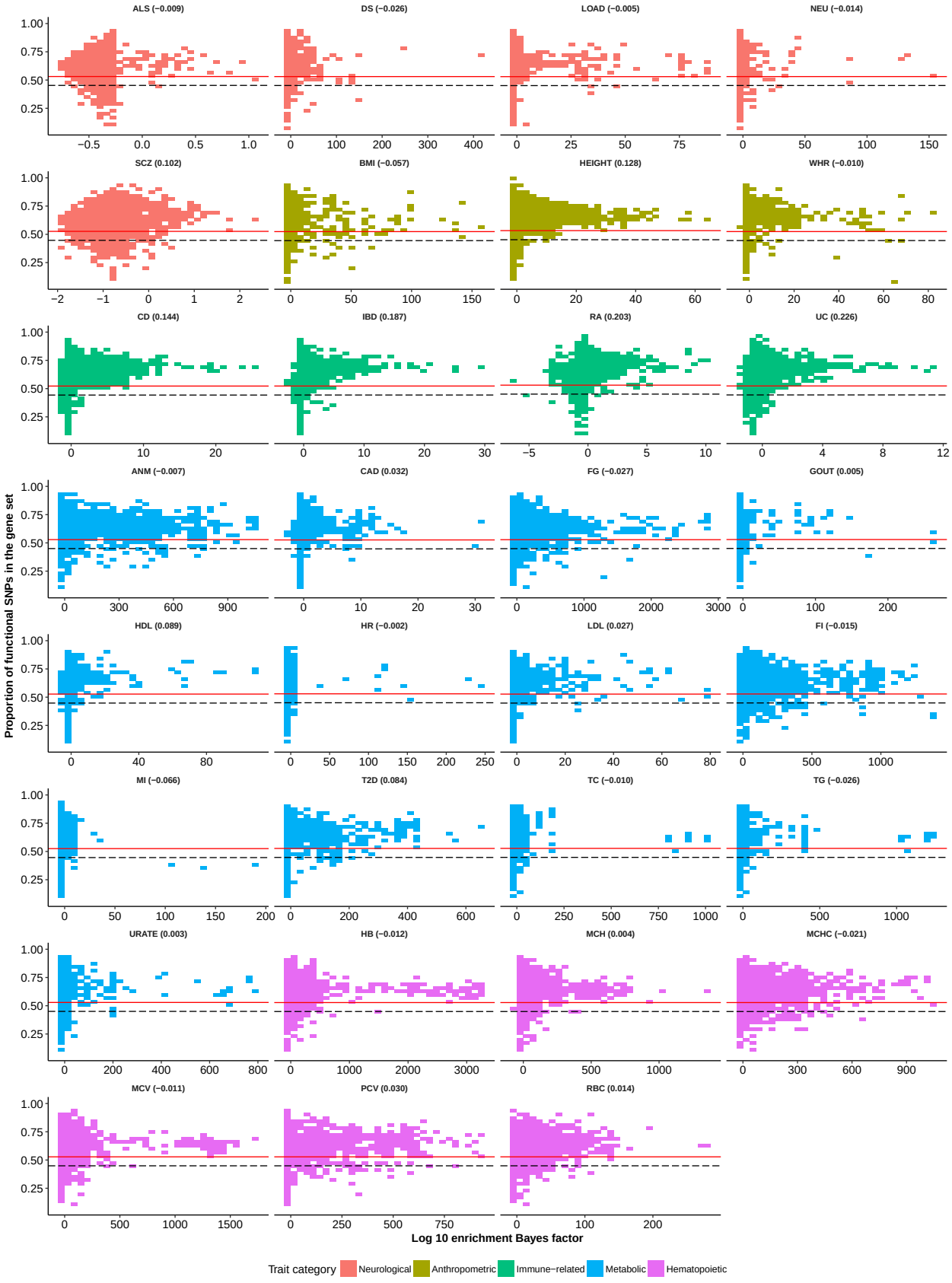
## Functional category: FetalDHS_Trynka



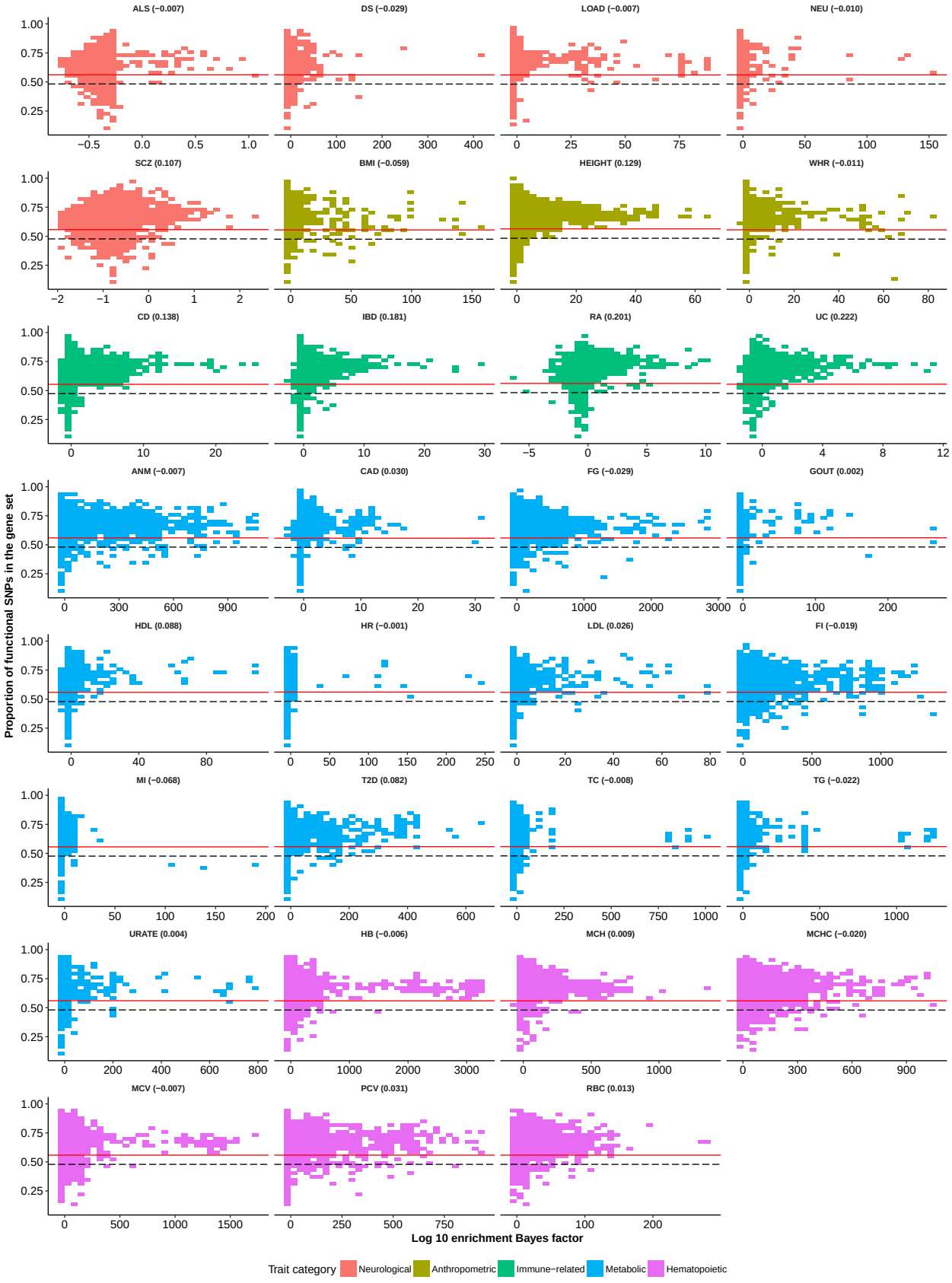
## Functional category: FetalDHS_Trynka.extend.500



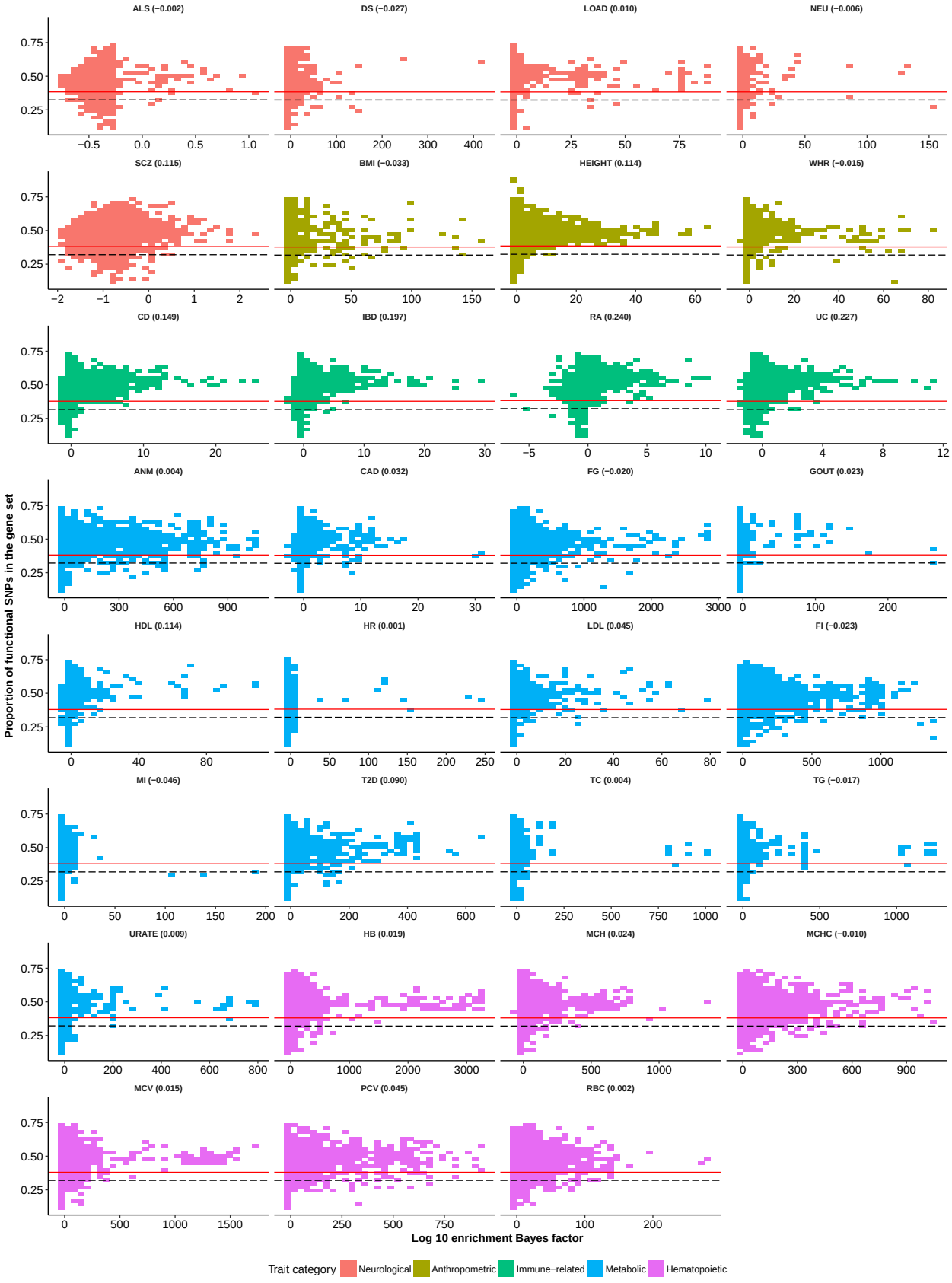
## Functional category: H3K27ac_Hnisz



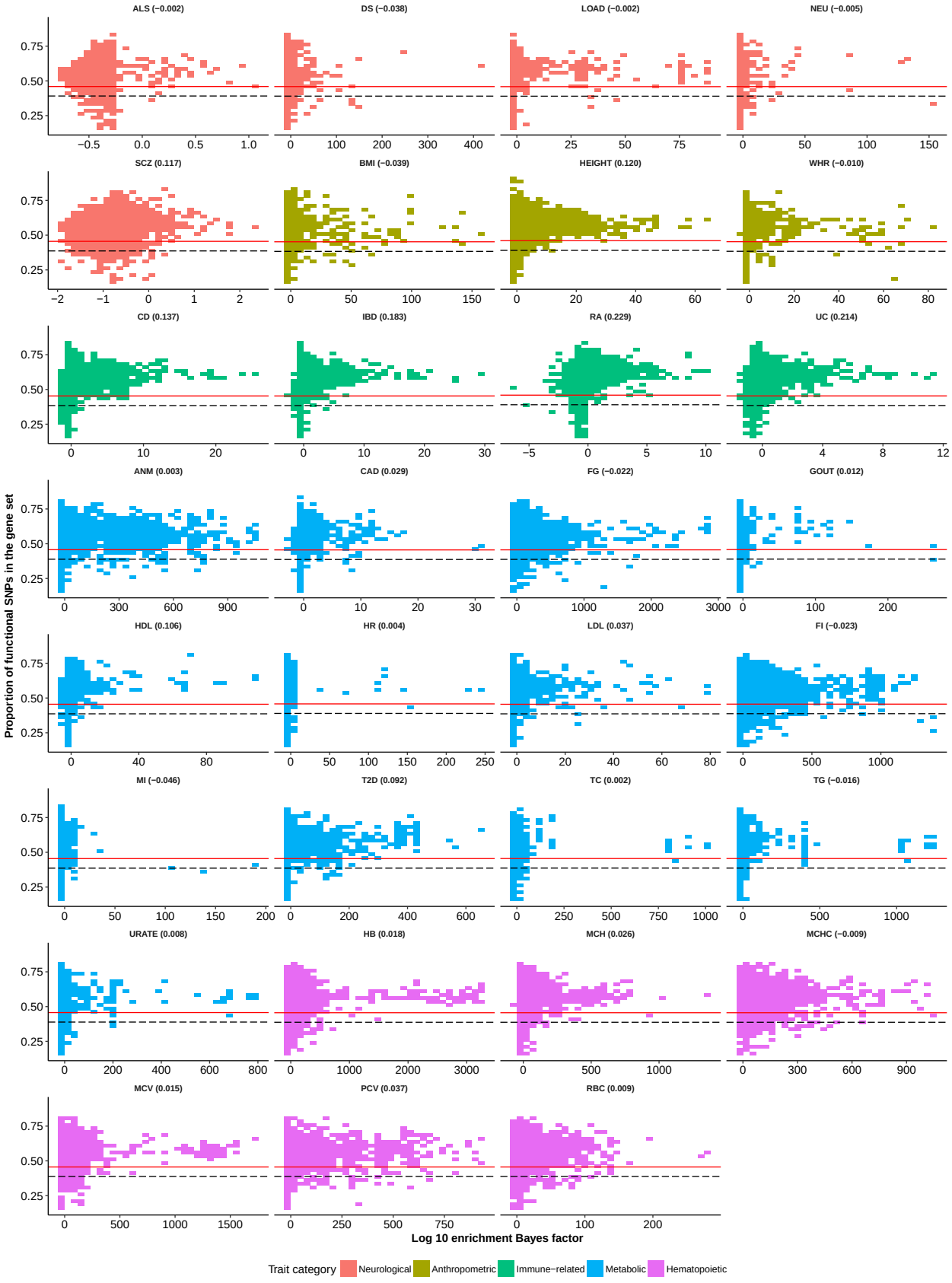
## Functional category: H3K27ac_Hnisz.extend.500



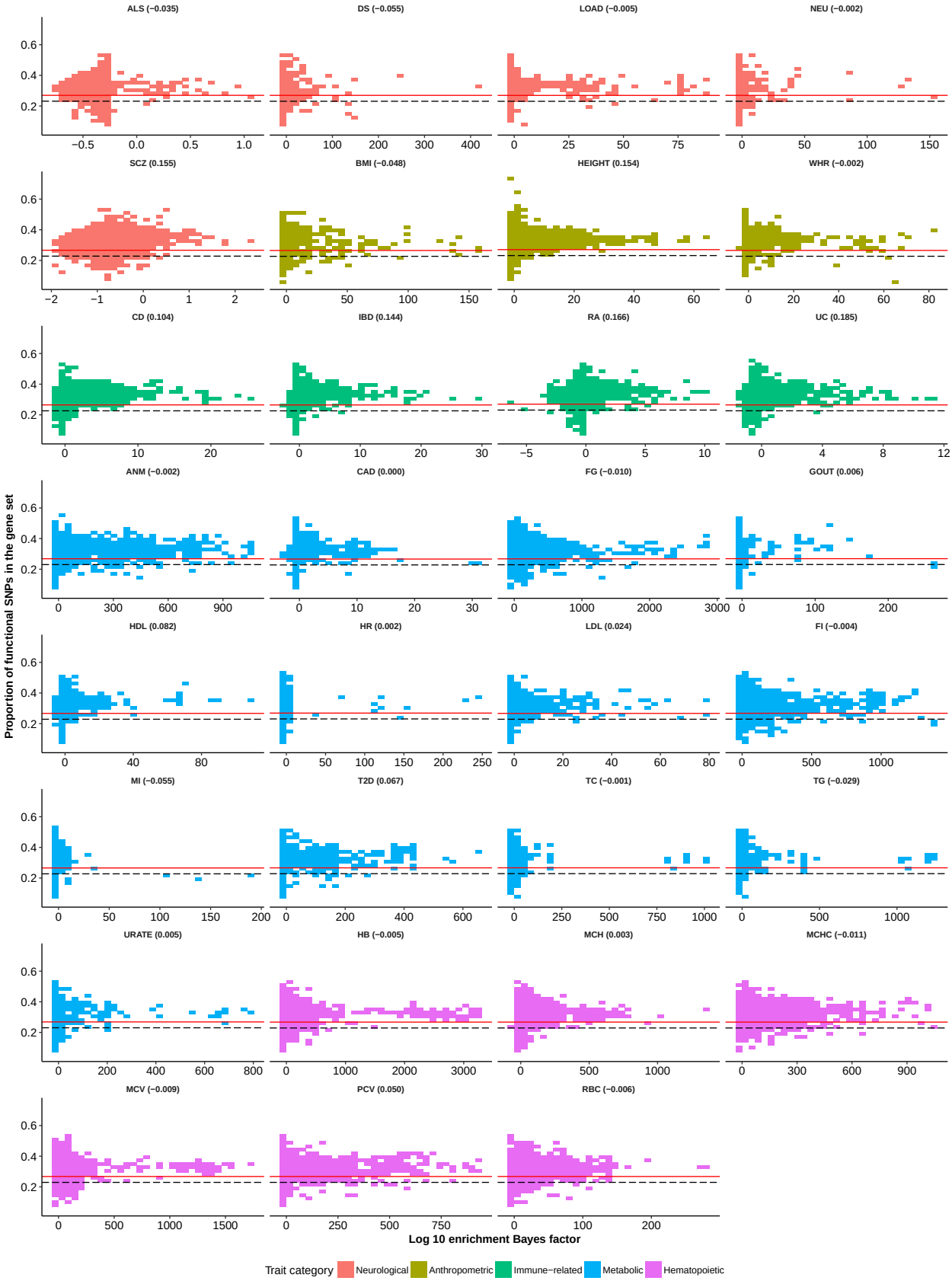
## Functional category: H3K27ac_PGC2



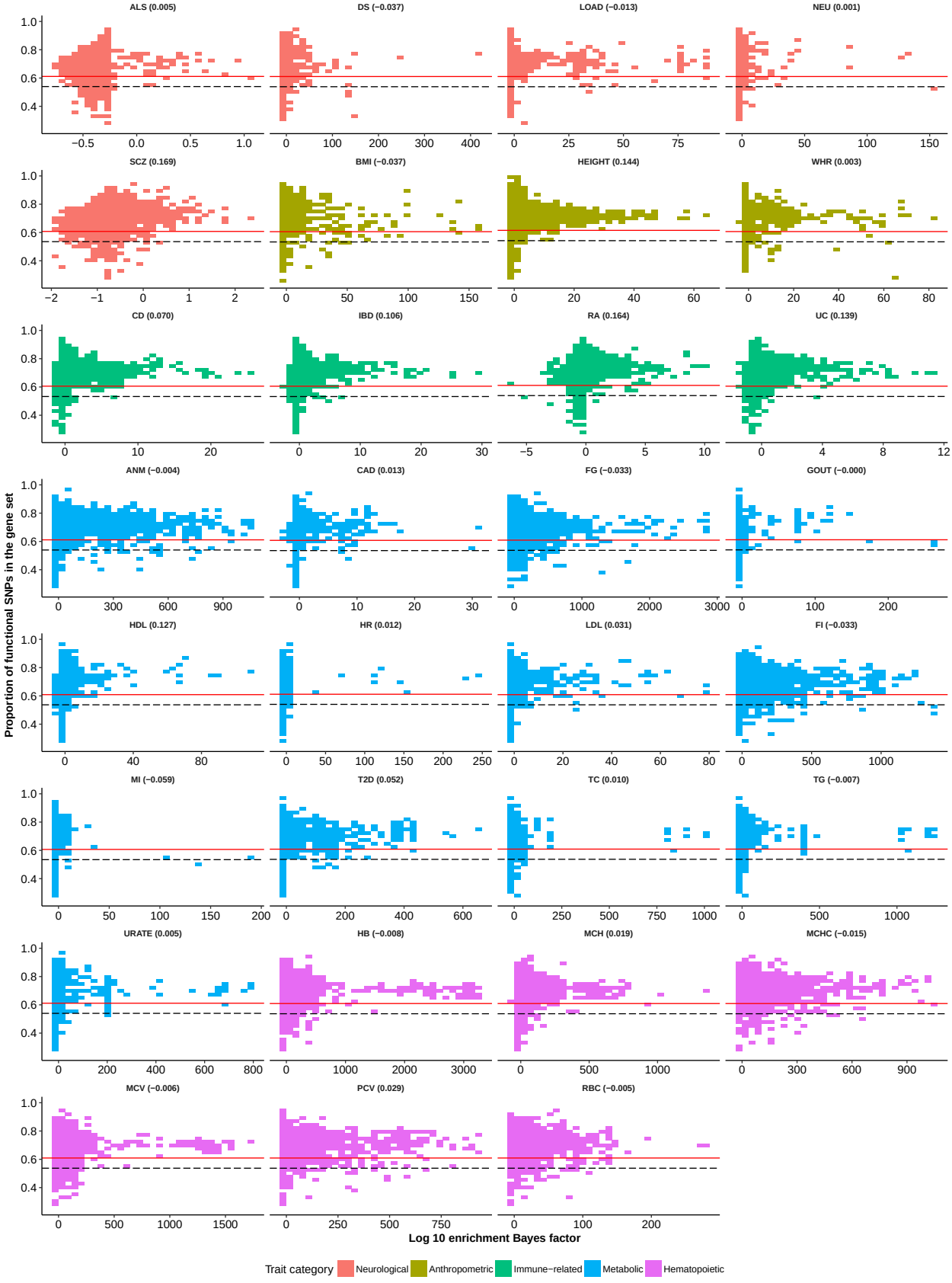
## Functional category: H3K27ac_PGC2.extend.500



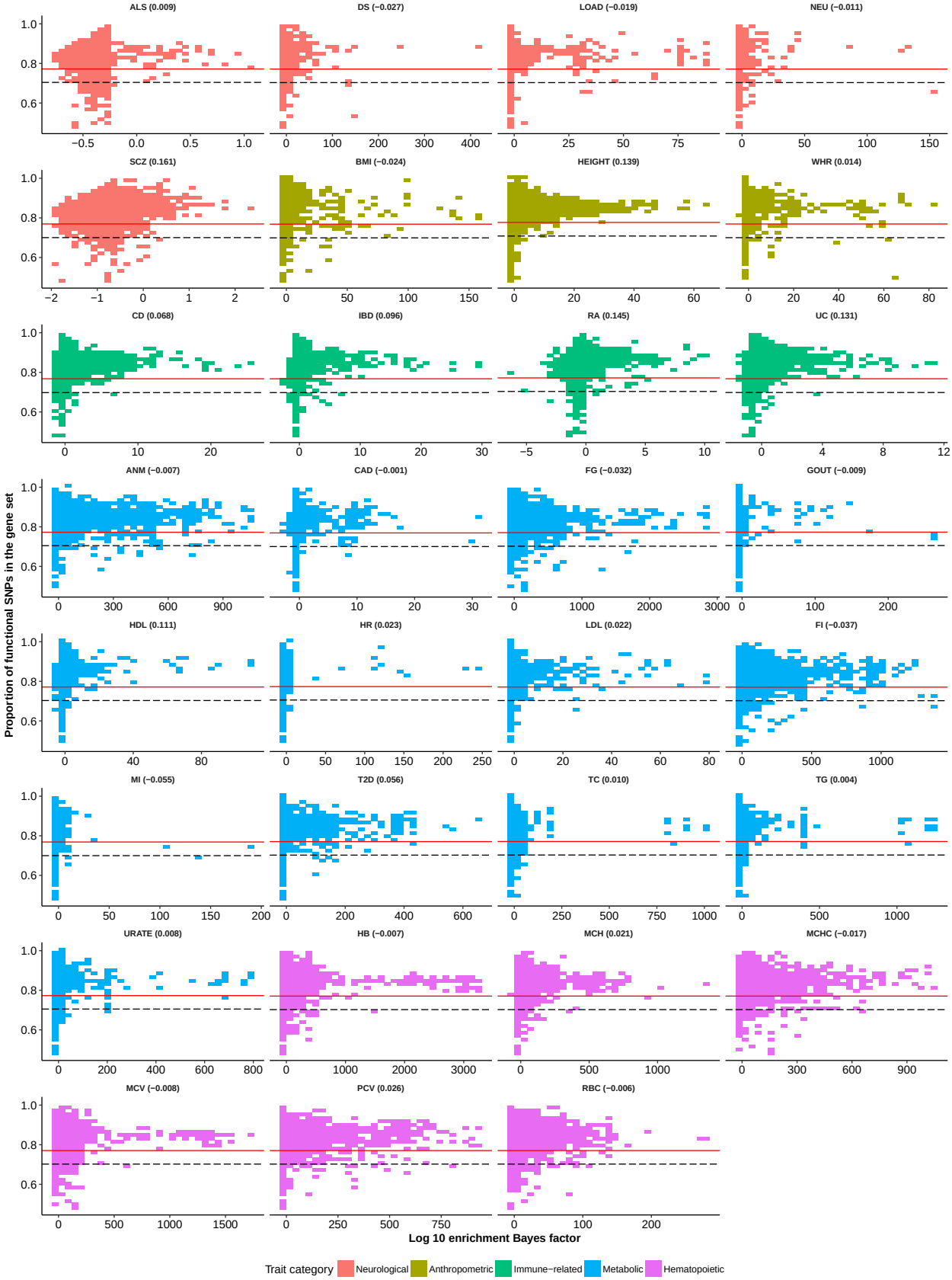
## Functional category: H3K4me1_peaks_Trynka



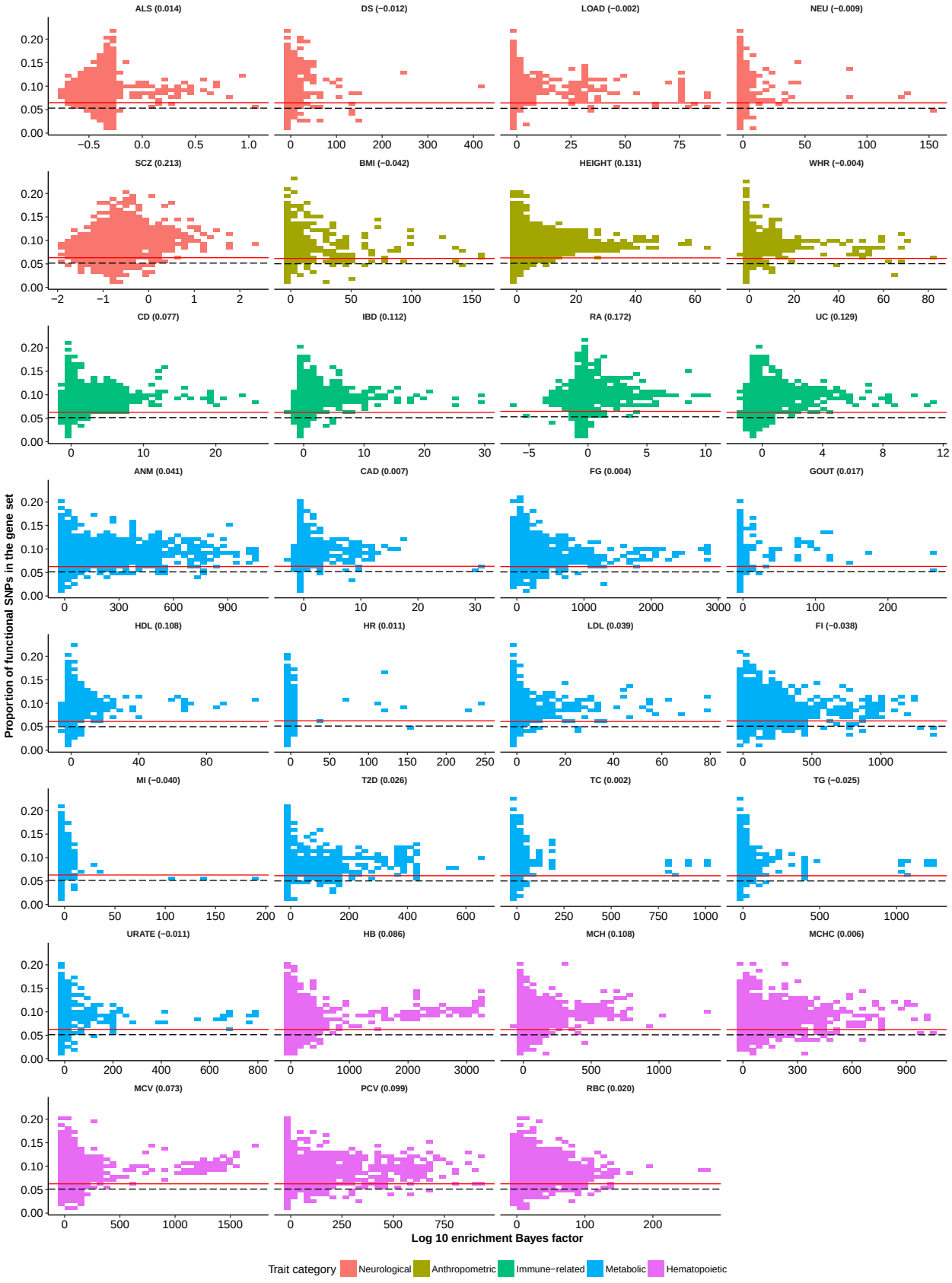
## Functional category: H3K4me1_Trynka



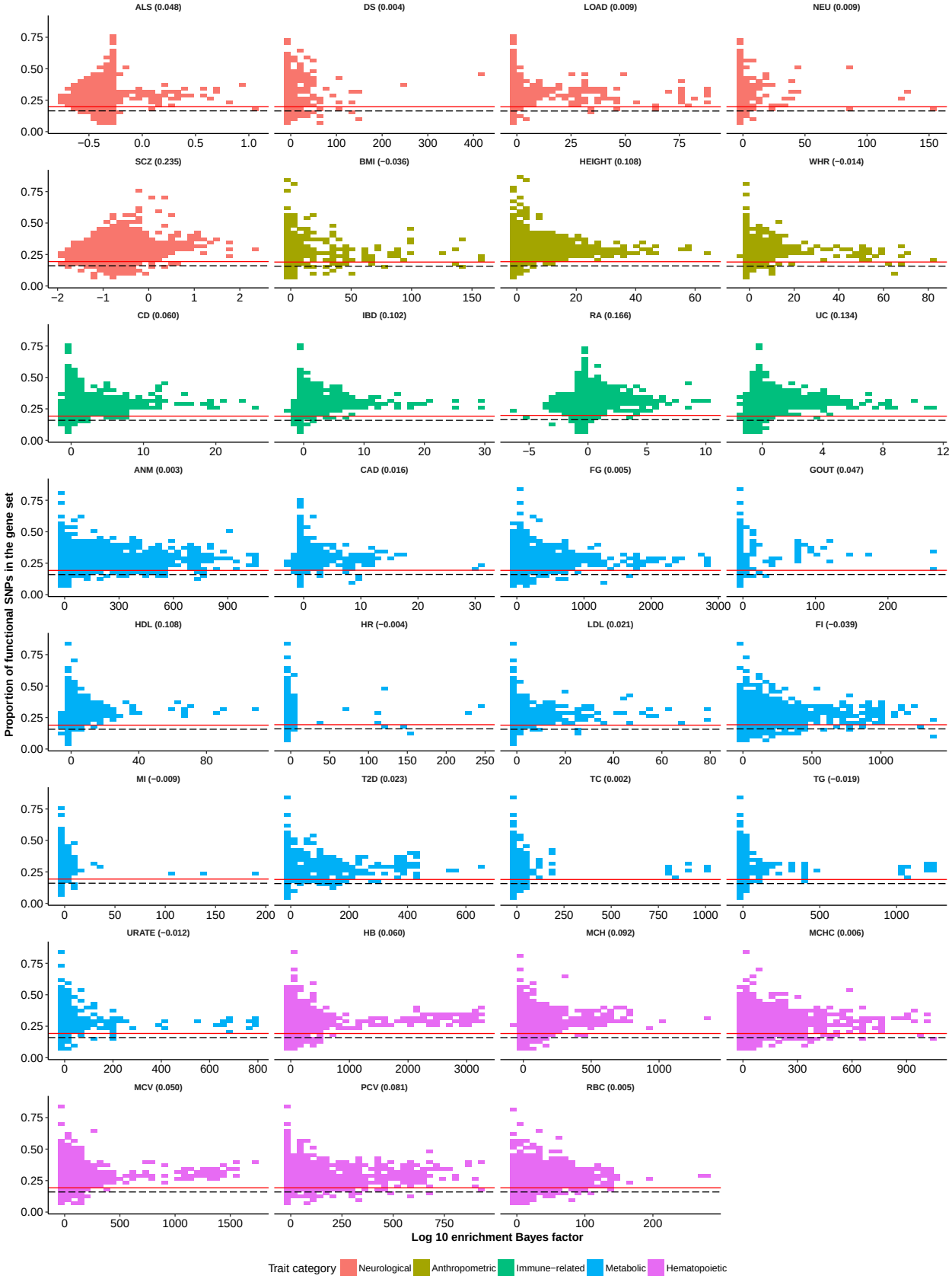
## Functional category: H3K4me1_Trynka.extend.500



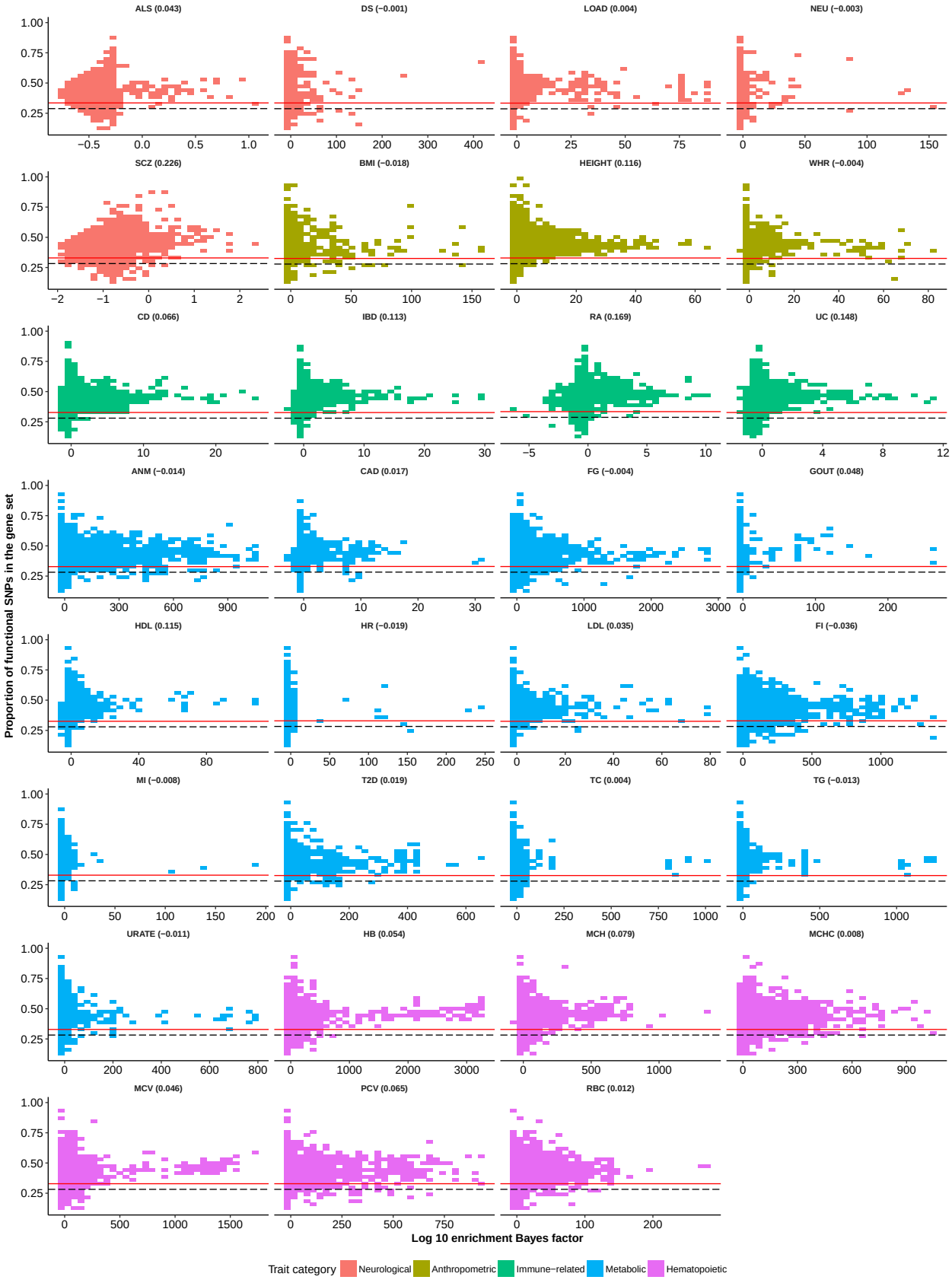
## Functional category: H3K4me3_peaks_Trynka



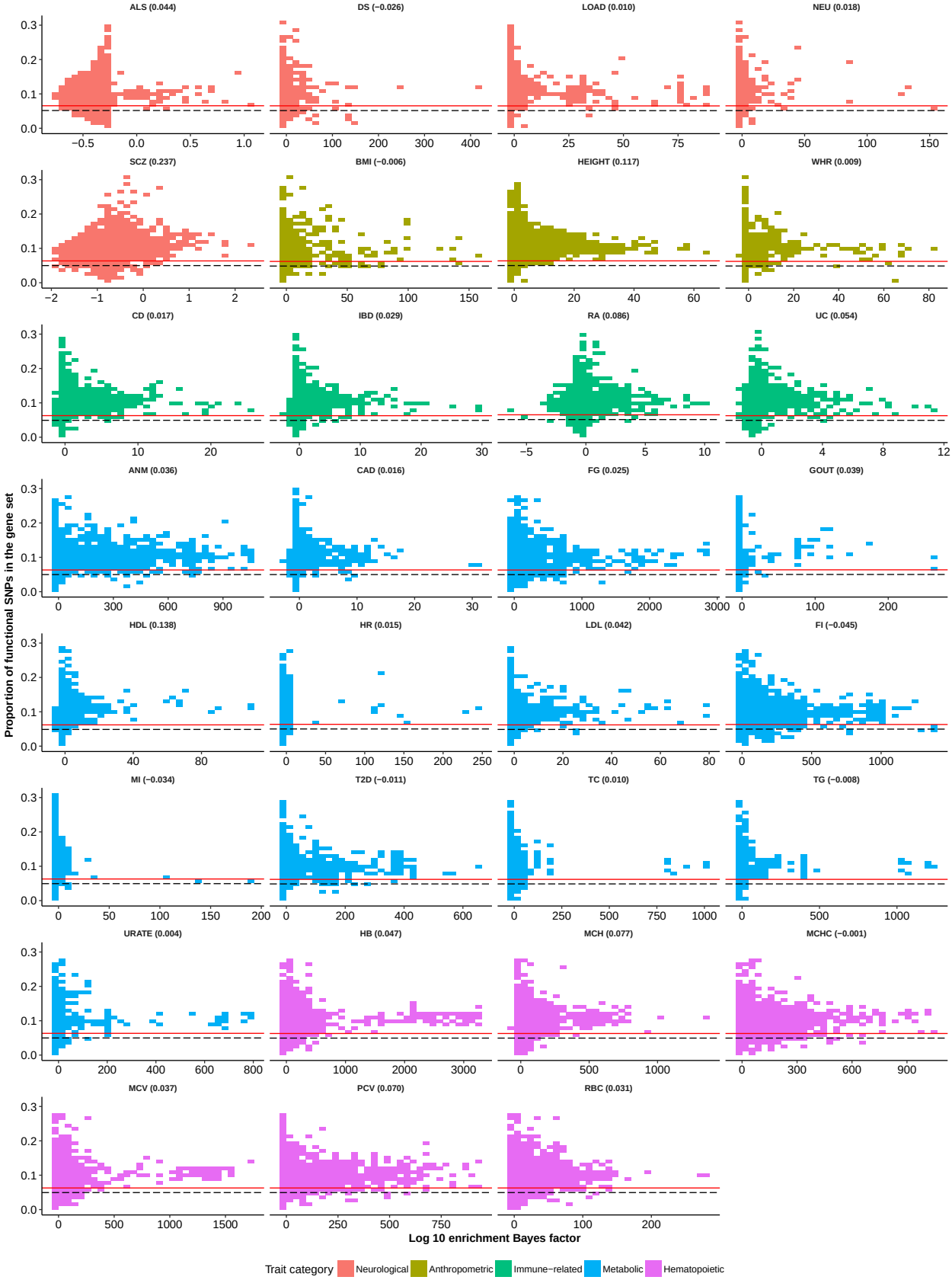
## Functional category: H3K4me3_Trynka



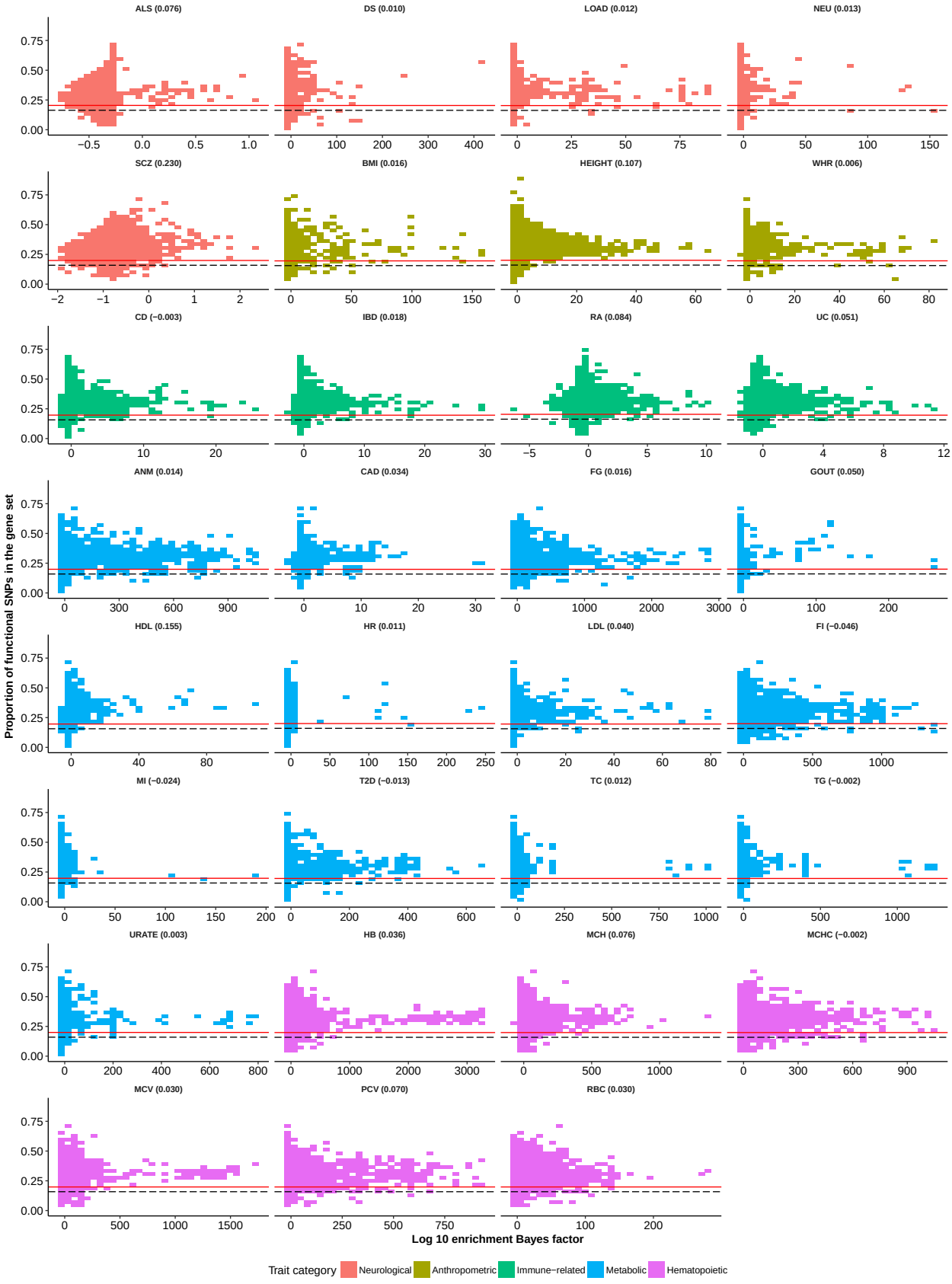
## Functional category: H3K4me3_Trynka.extend.500



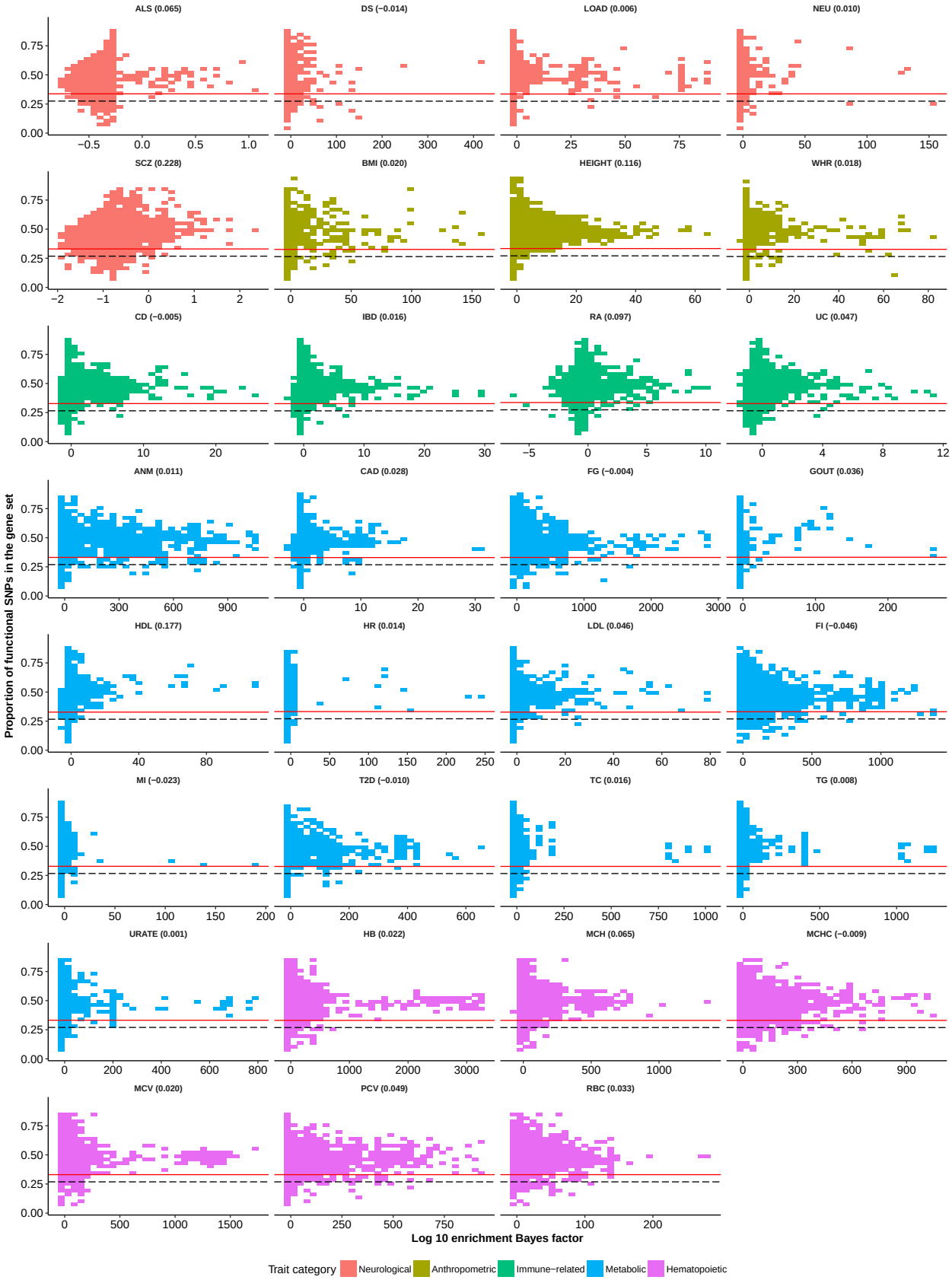
## Functional category: H3K9ac_peaks_Trynka



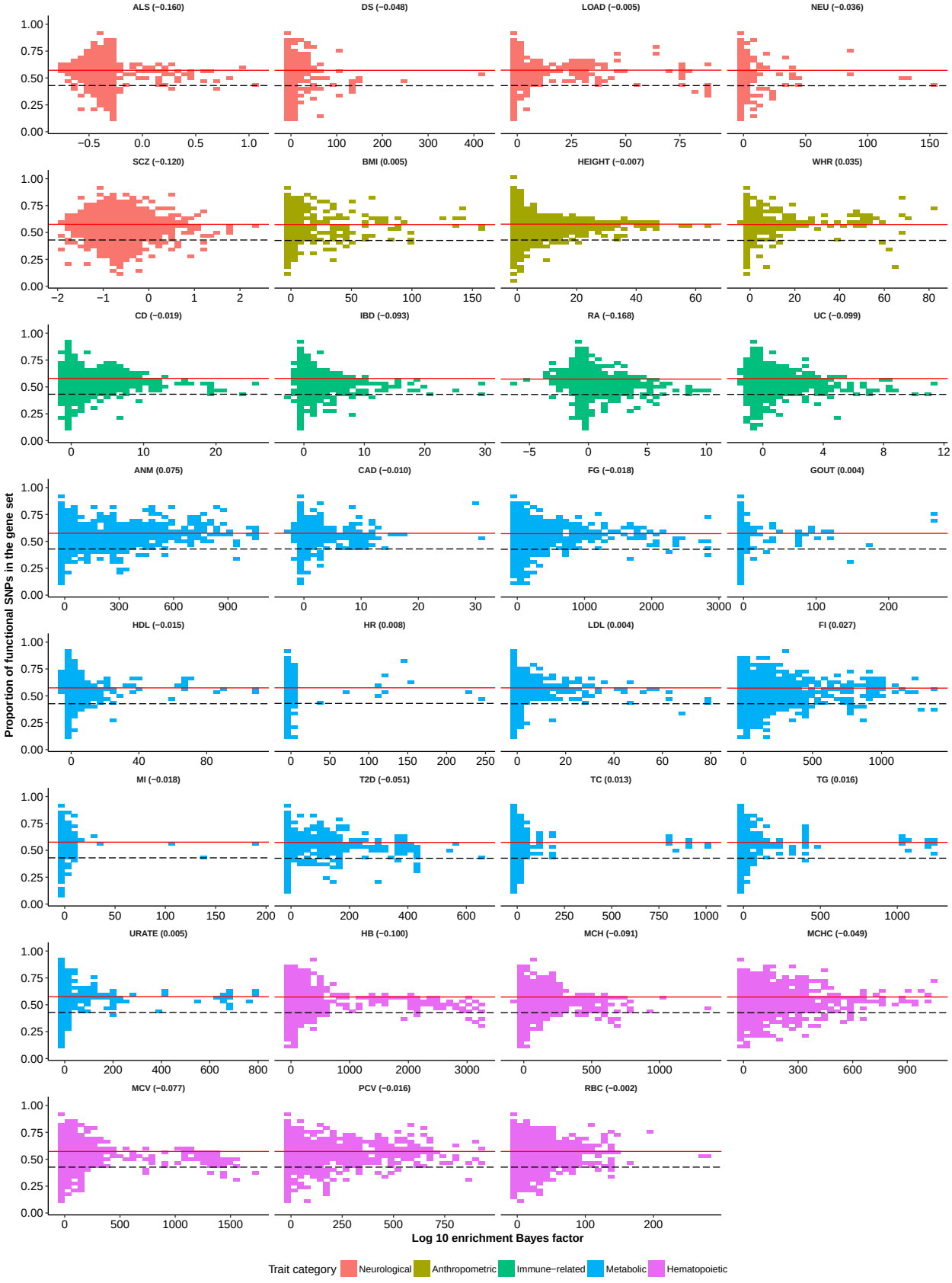
## Functional category: H3K9ac_Trynka



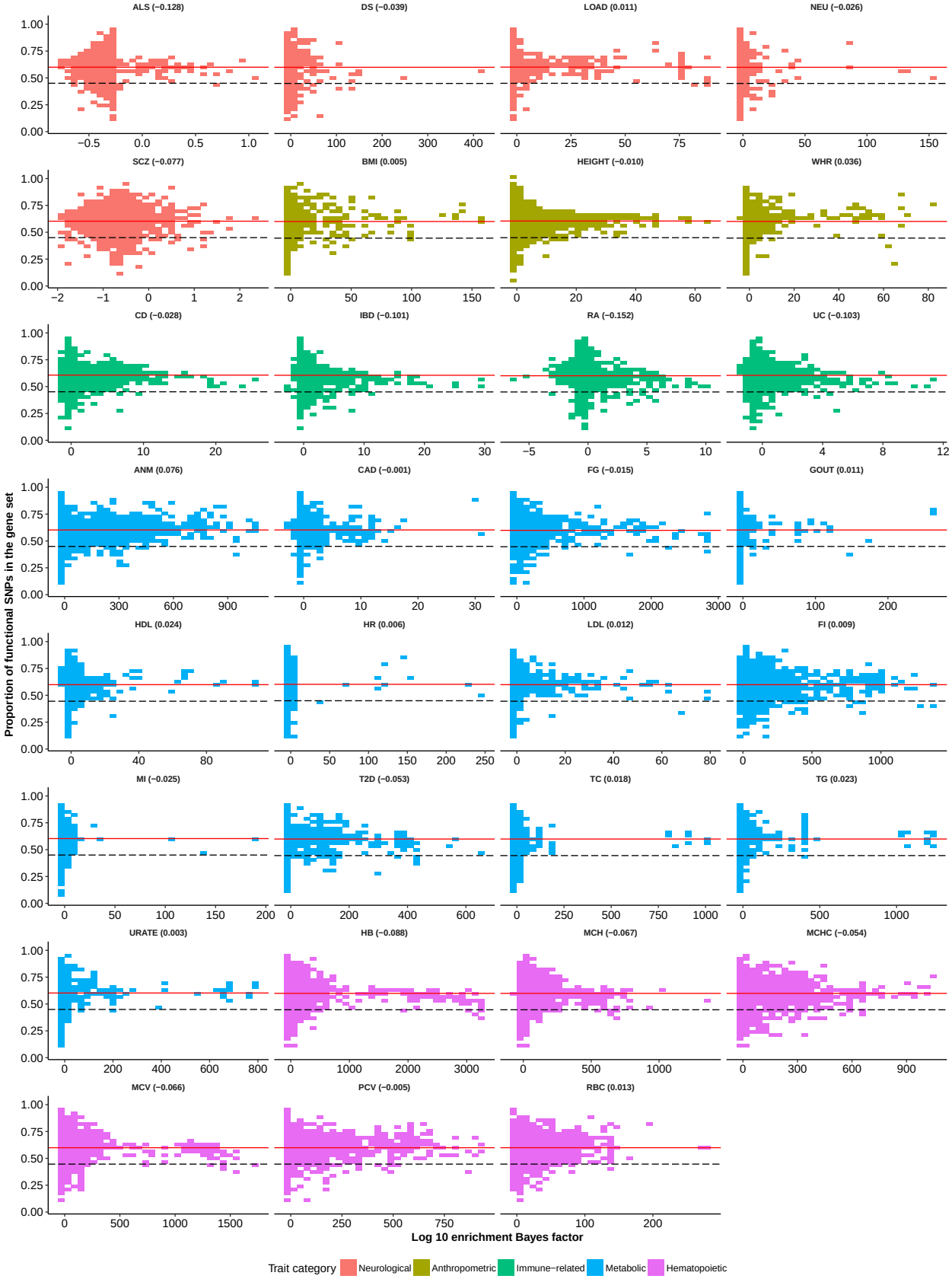
## Functional category: H3K9ac_Trynka.extend.500



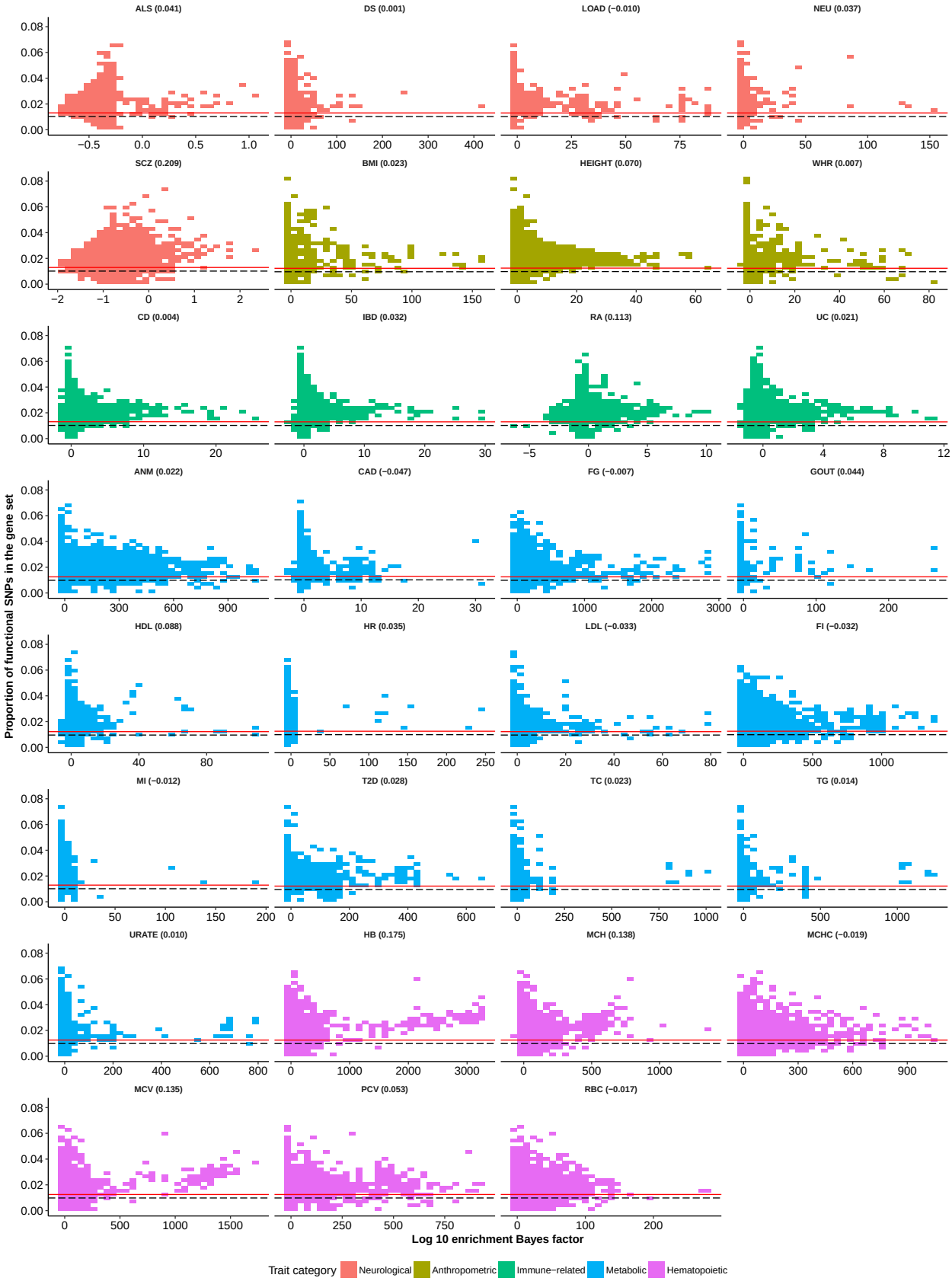
## Functional category: Intron_UCSC



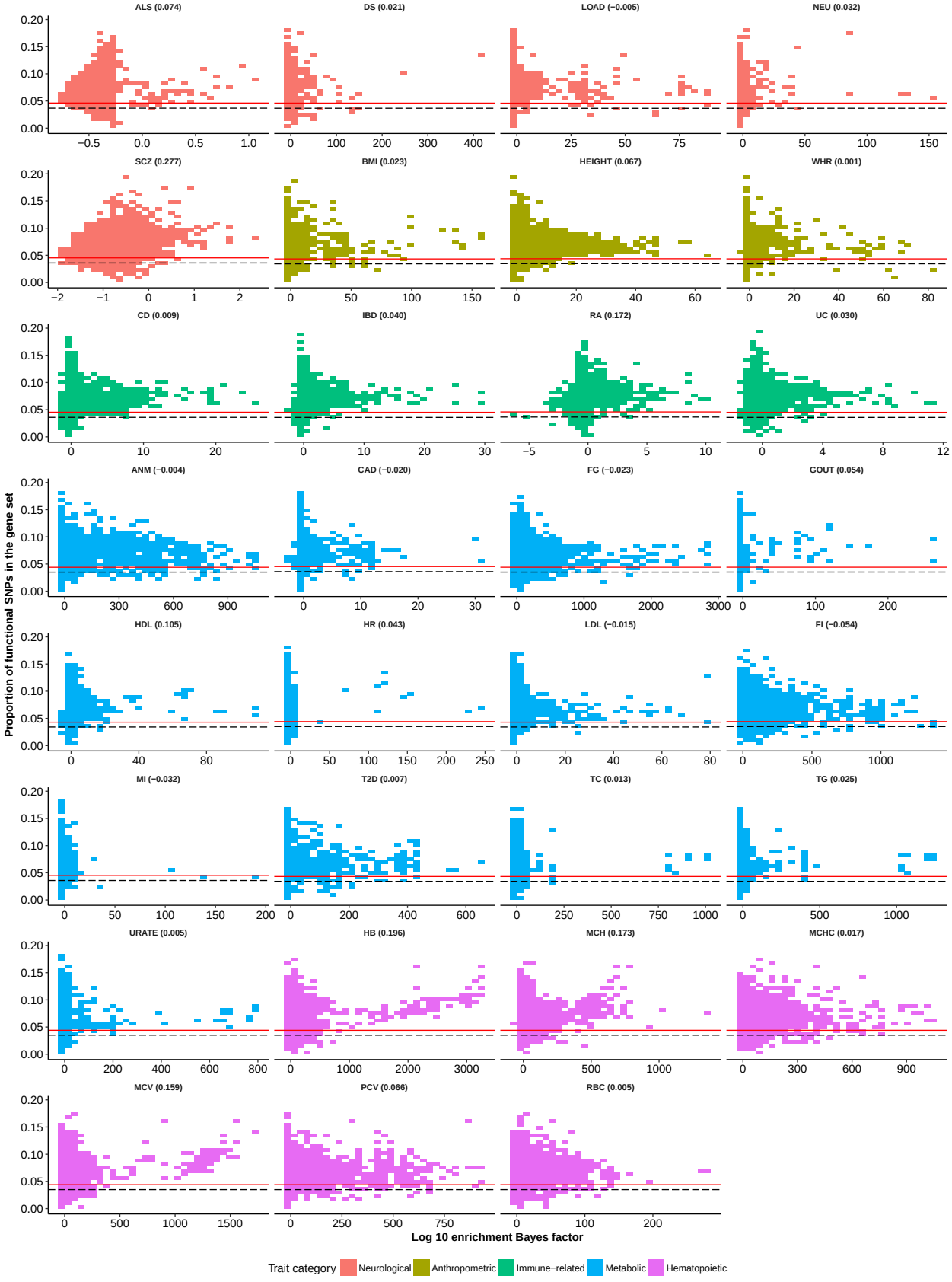
## Functional category: Intron_UCSC.extend.500



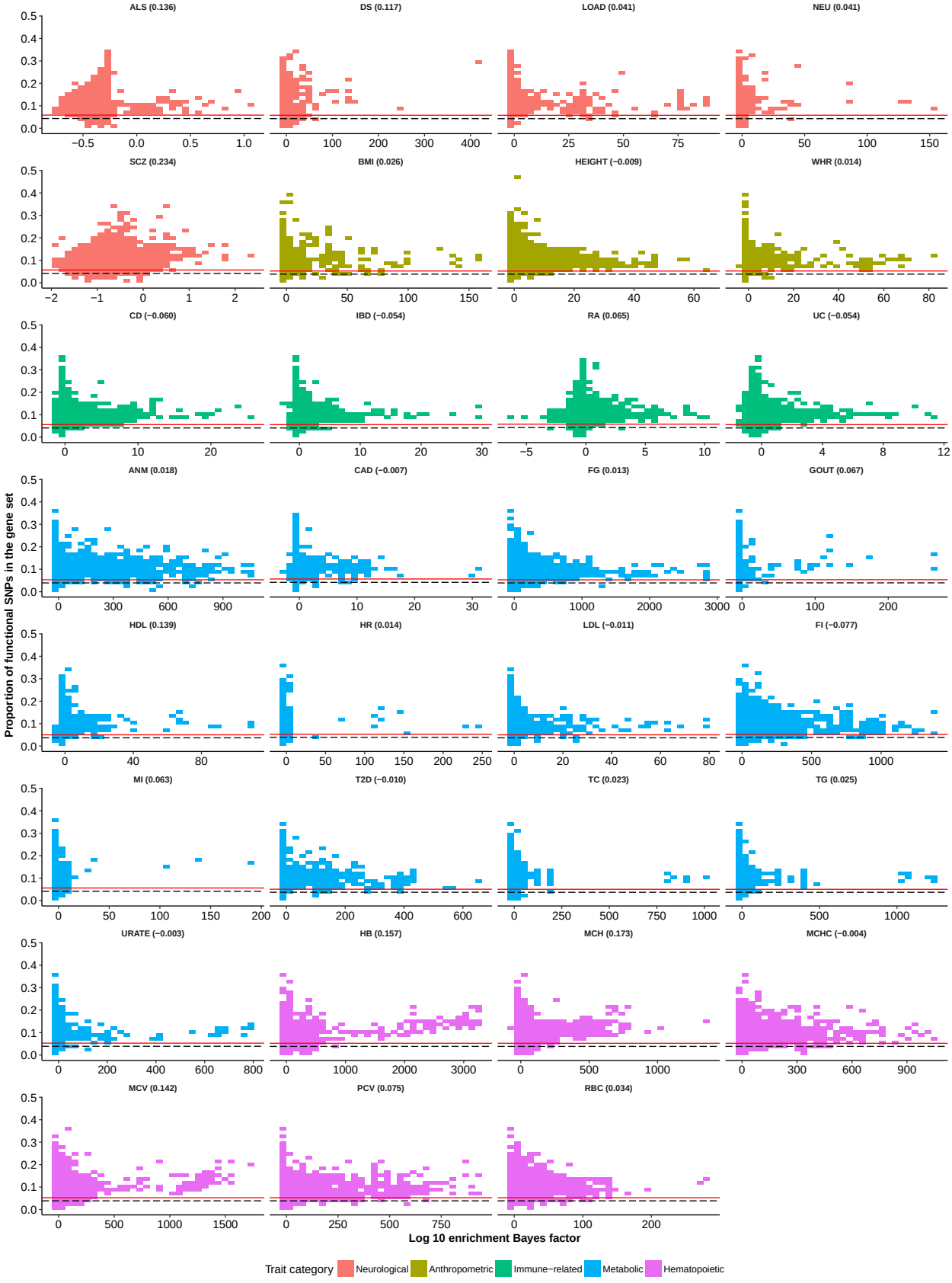
## Functional category: PromoterFlanking_Hoffman



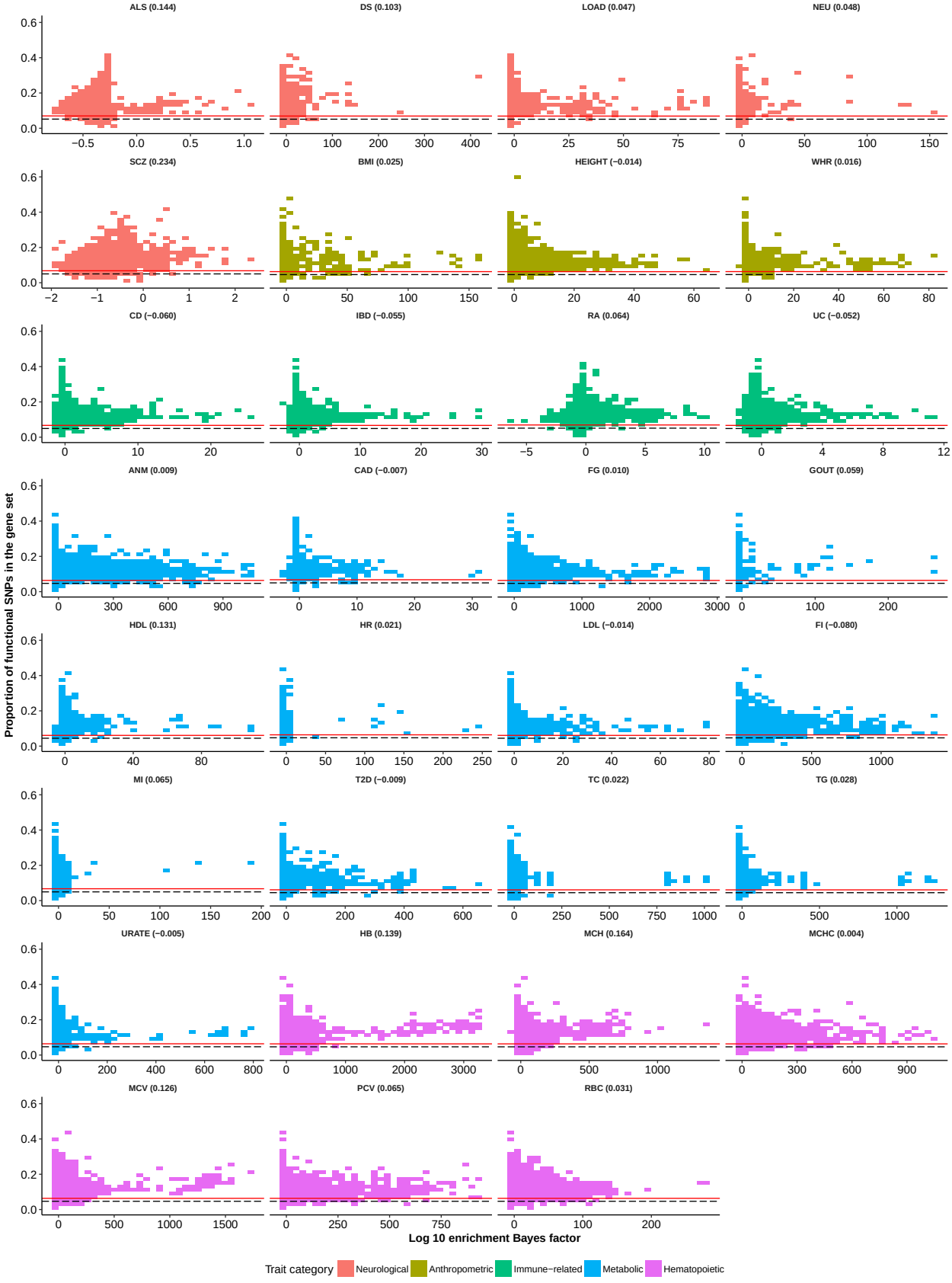
## Functional category: PromoterFlanking_Hoffman.extend.500



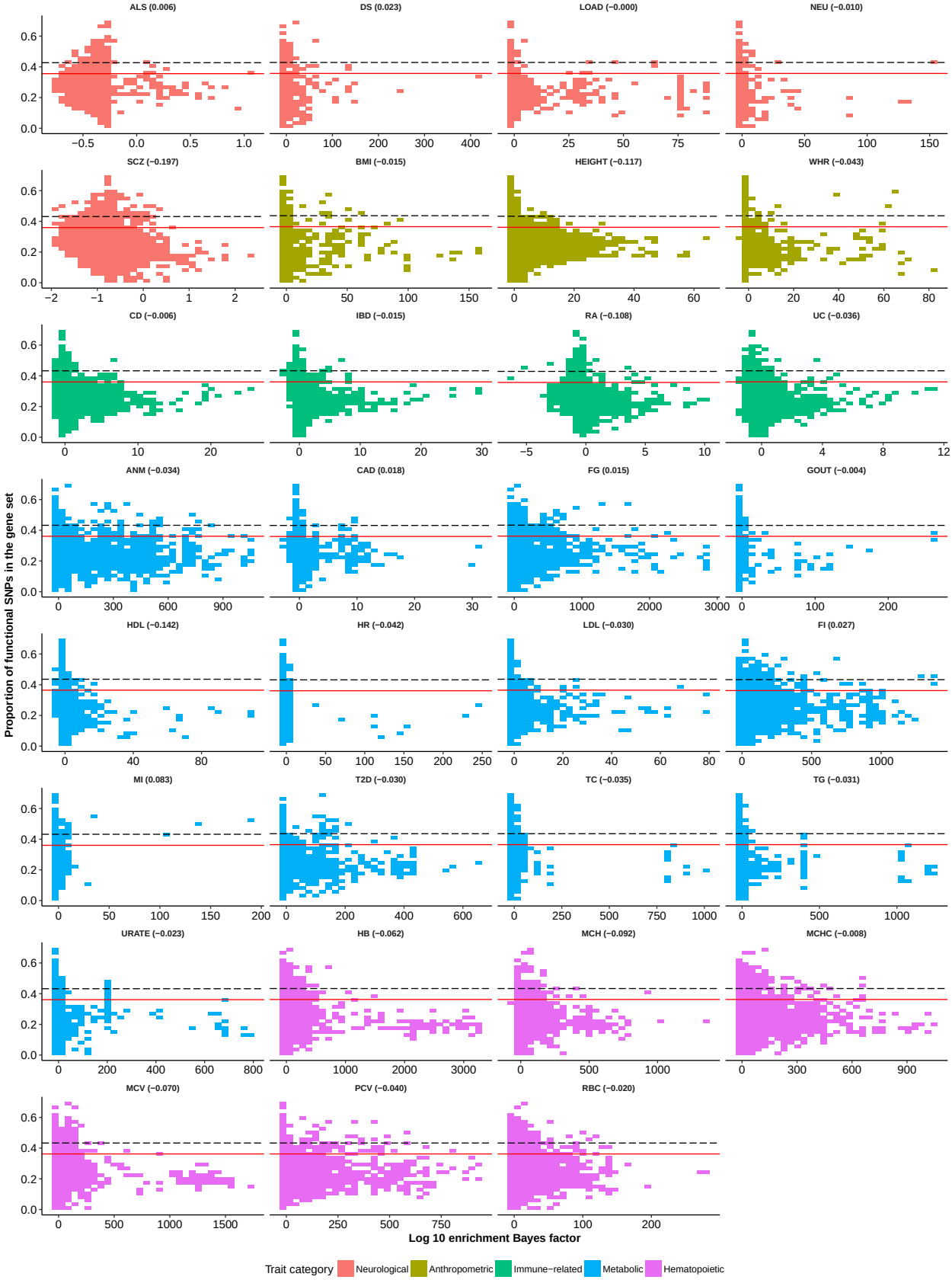
## Functional category: Promoter_UCSC



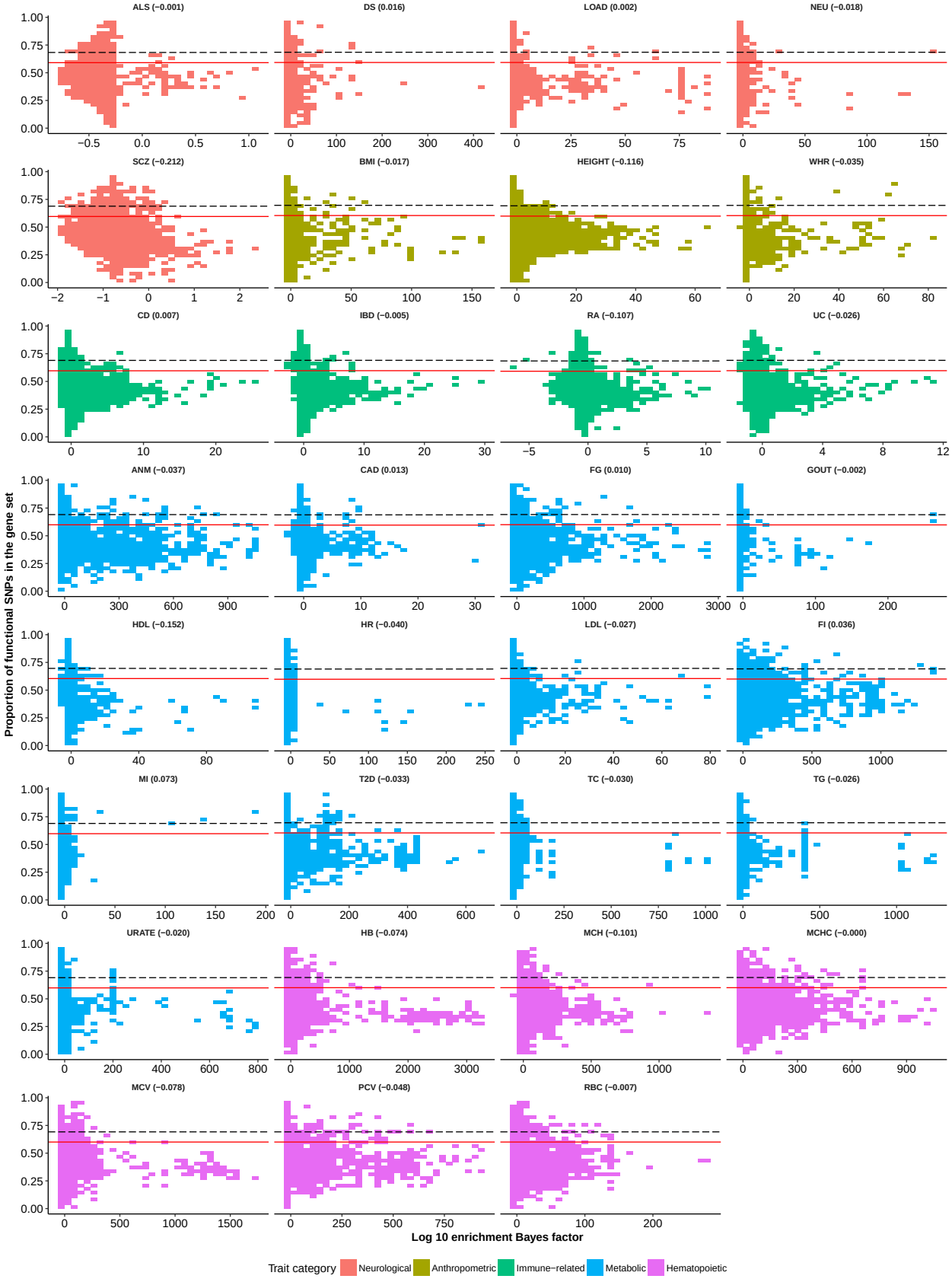
## Functional category: Promoter_UCSC.extend.500



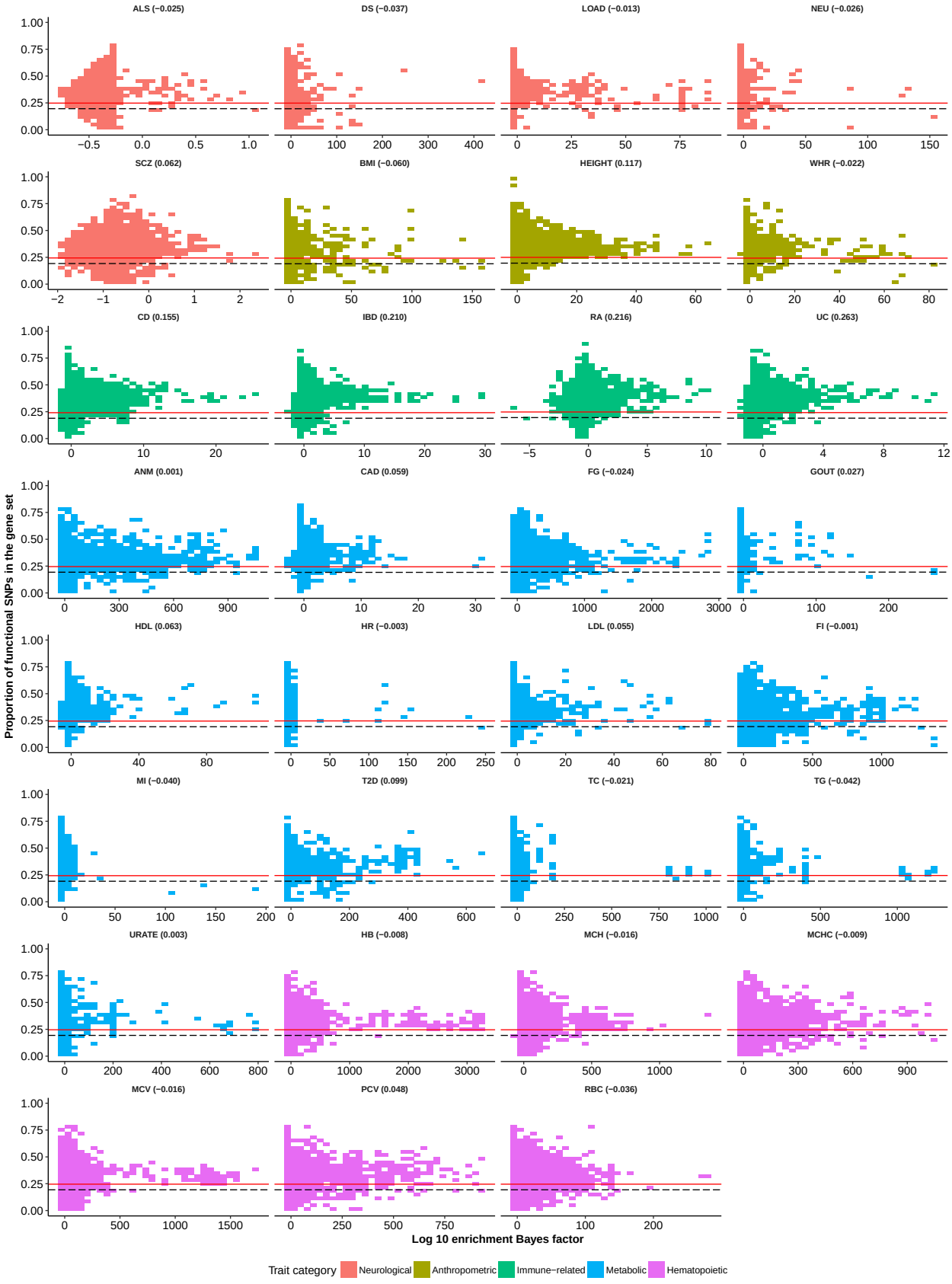
## Functional category: Repressed_Hoffman



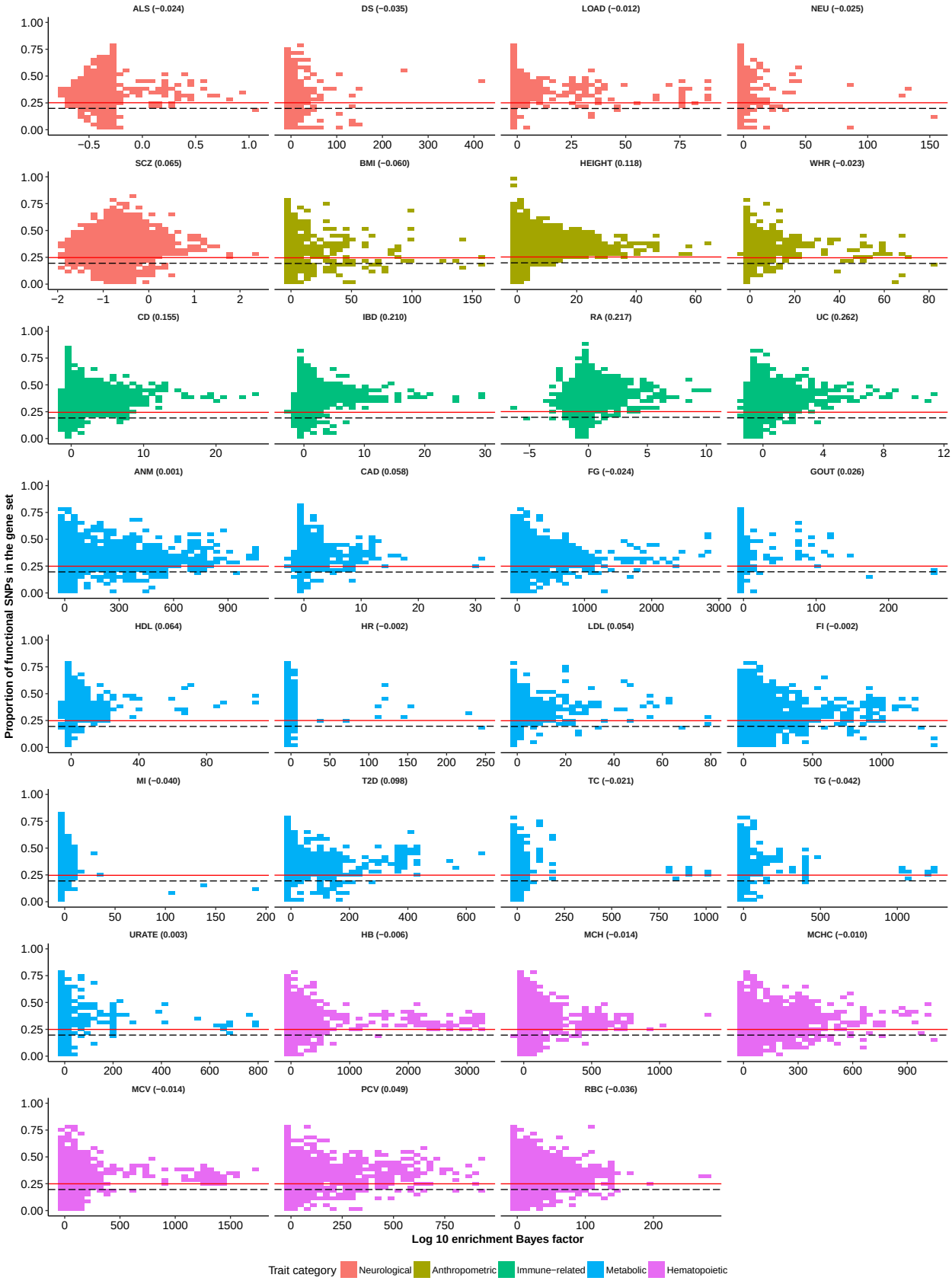
## Functional category: Repressed_Hoffman.extend.500



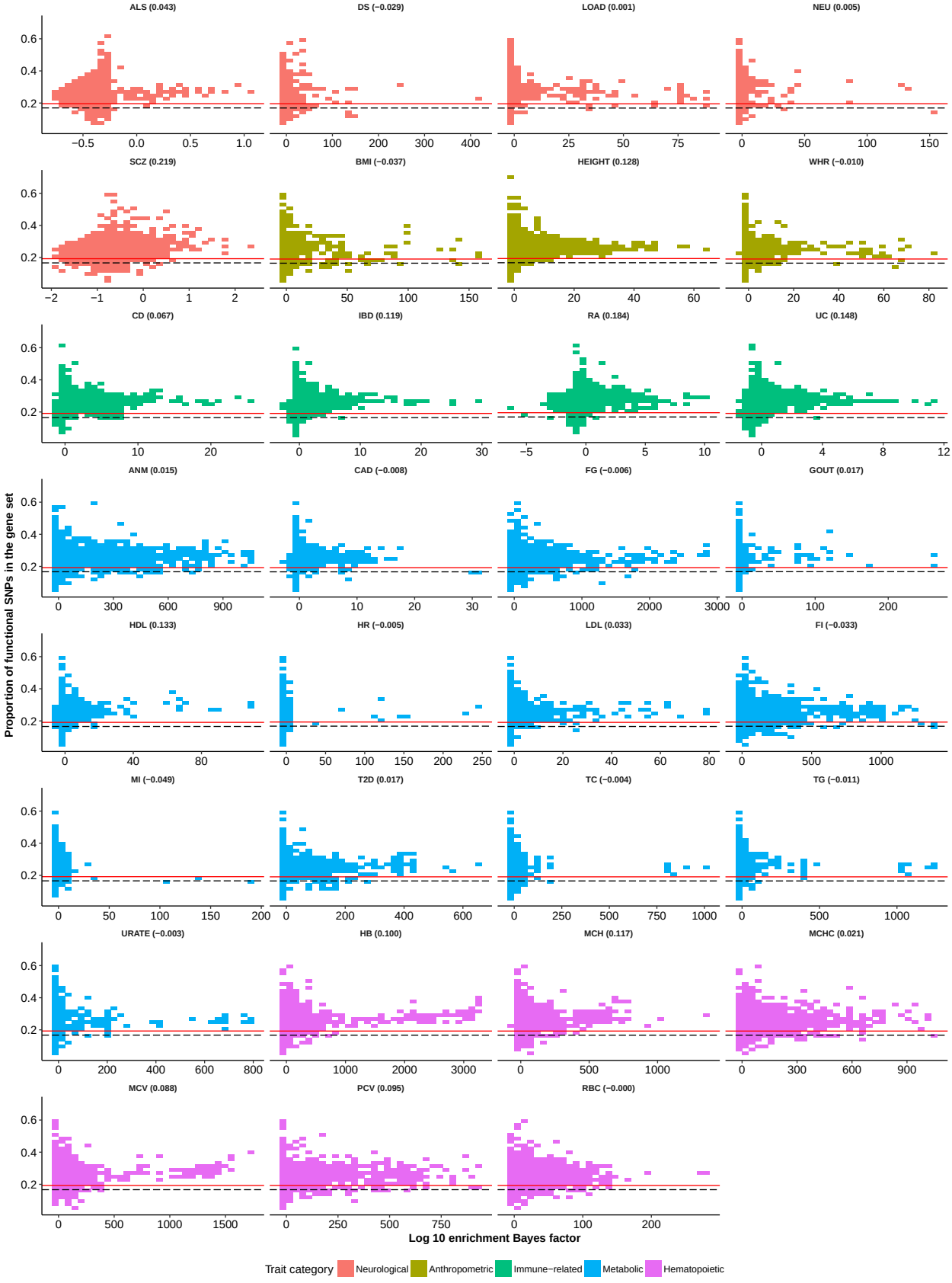
## Functional category: SuperEnhancer_Hnisz



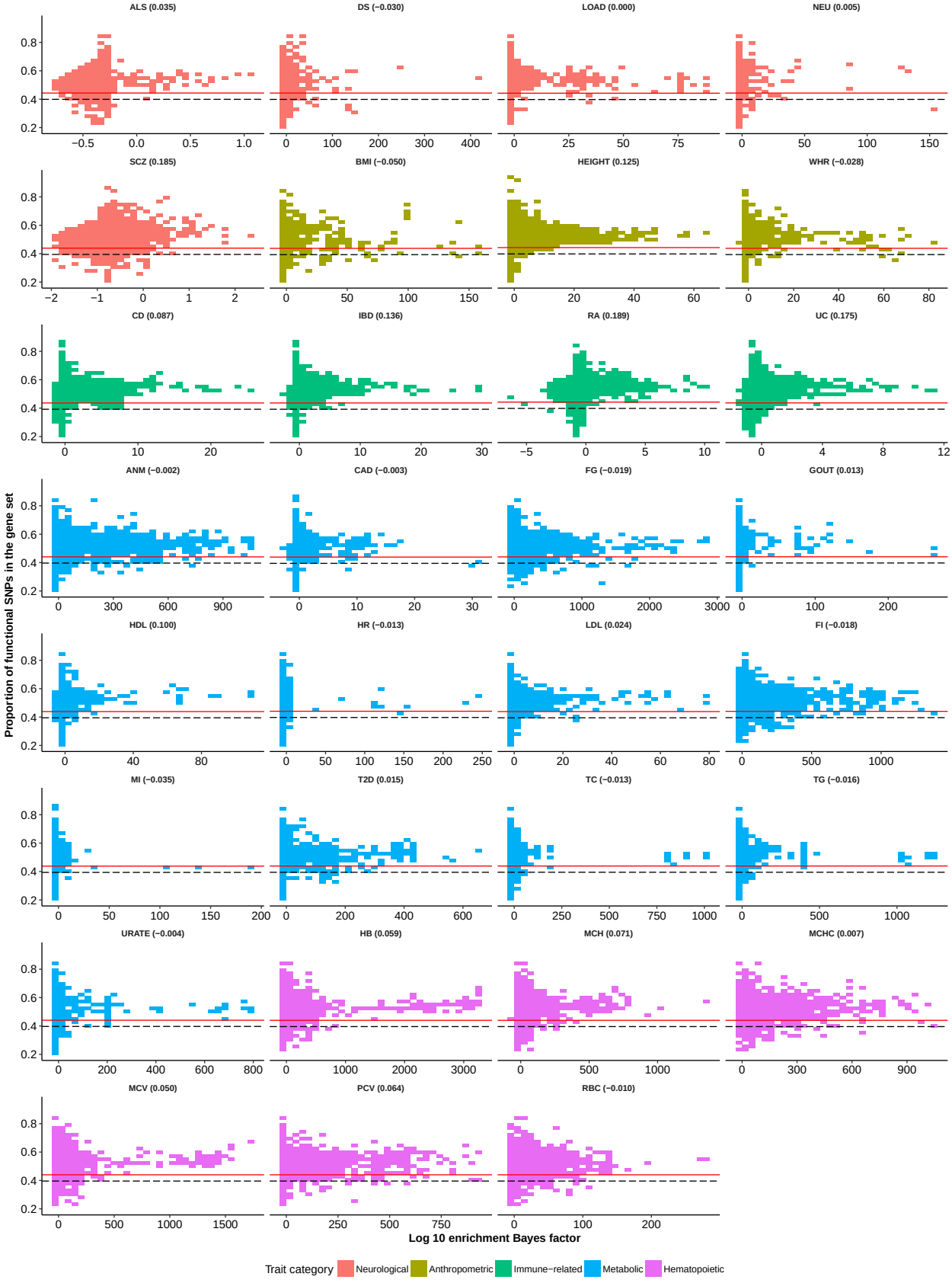
## Functional category: SuperEnhancer_Hnisz.extend.500



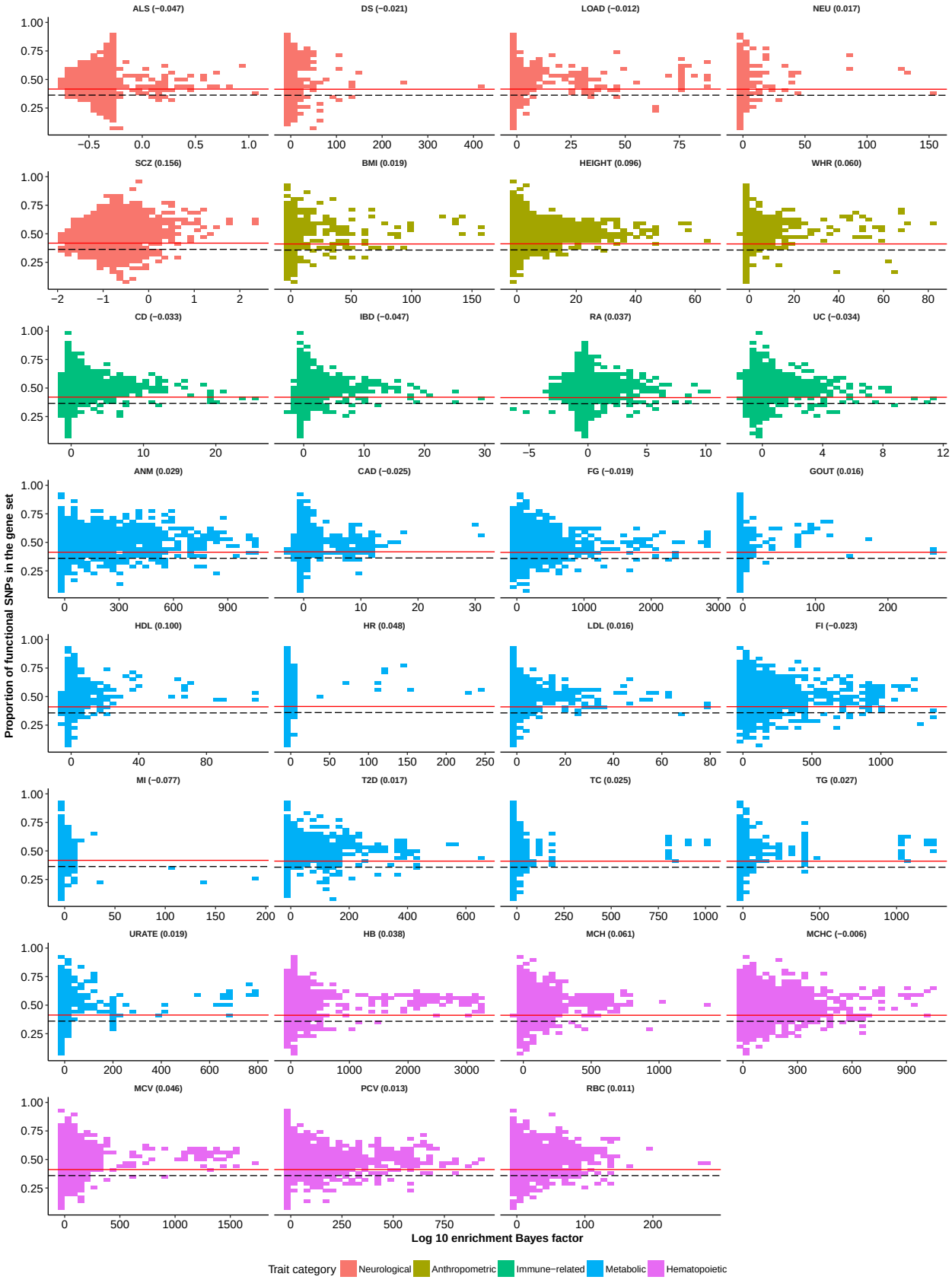
## Functional category: TFBS_ENCODE



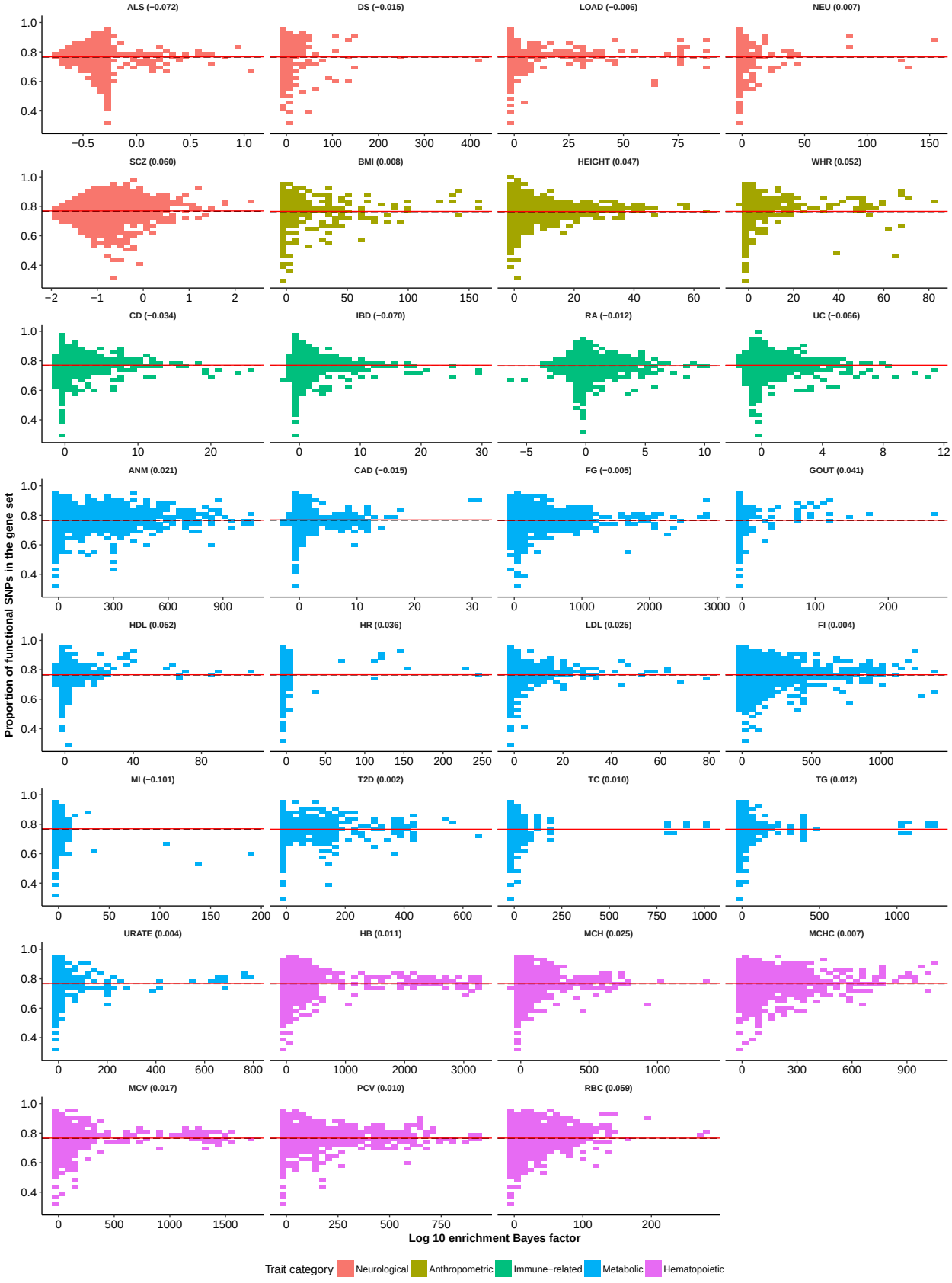
## Functional category: TFBS_ENCODE.extend.500



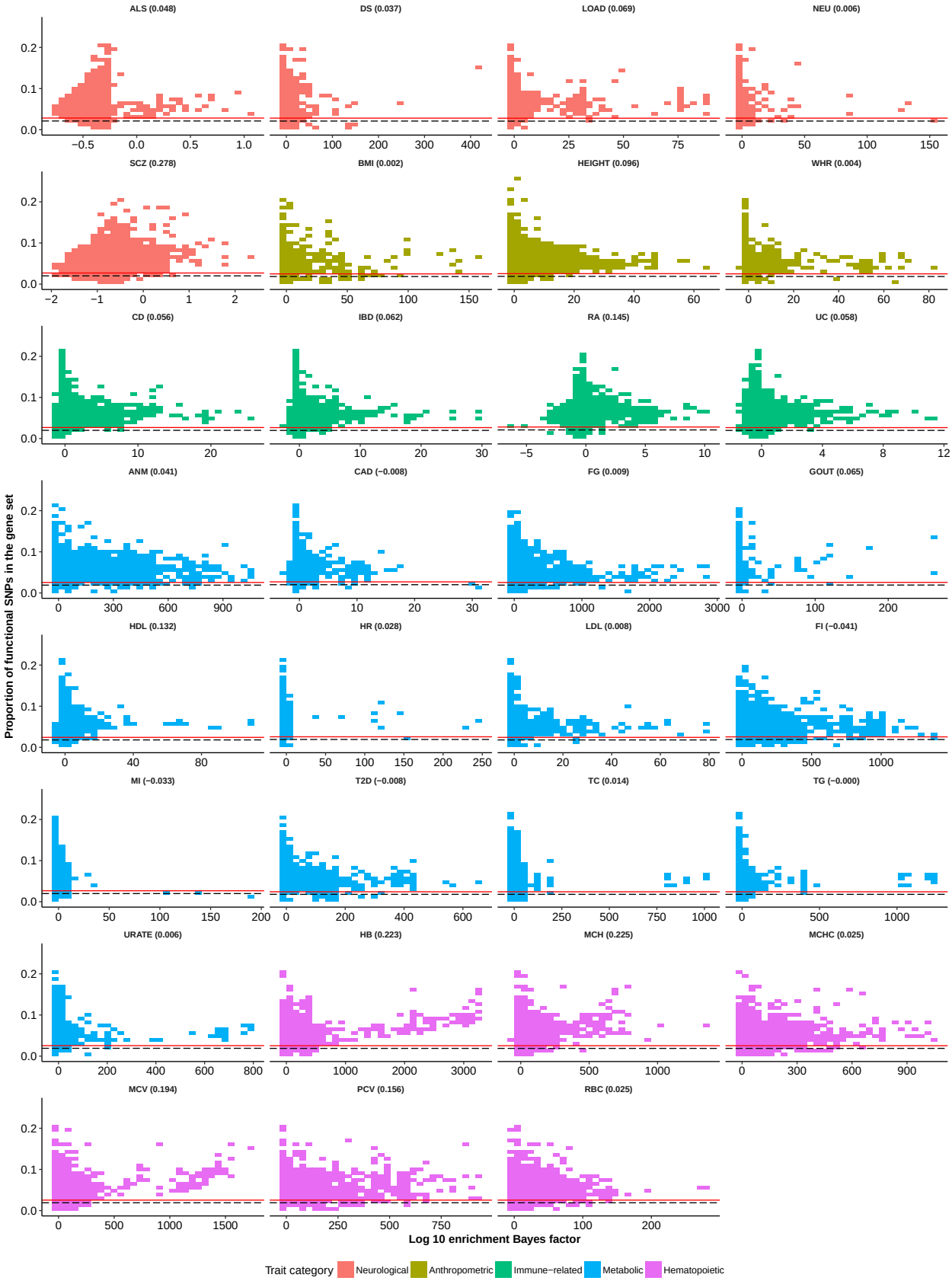
## Functional category: Transcribed_Hoffman



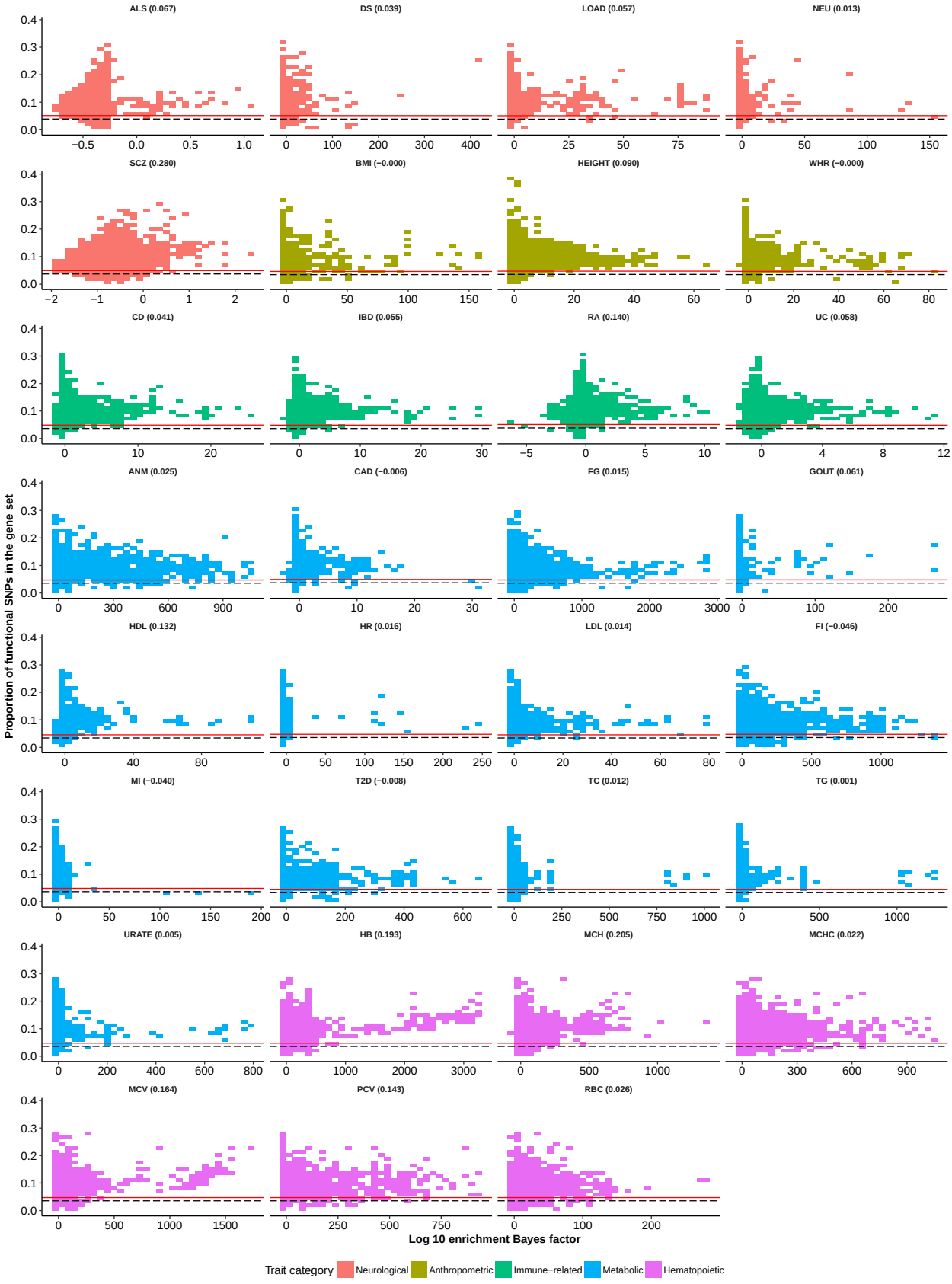
## Functional category: Transcribed_Hoffman.extend.500



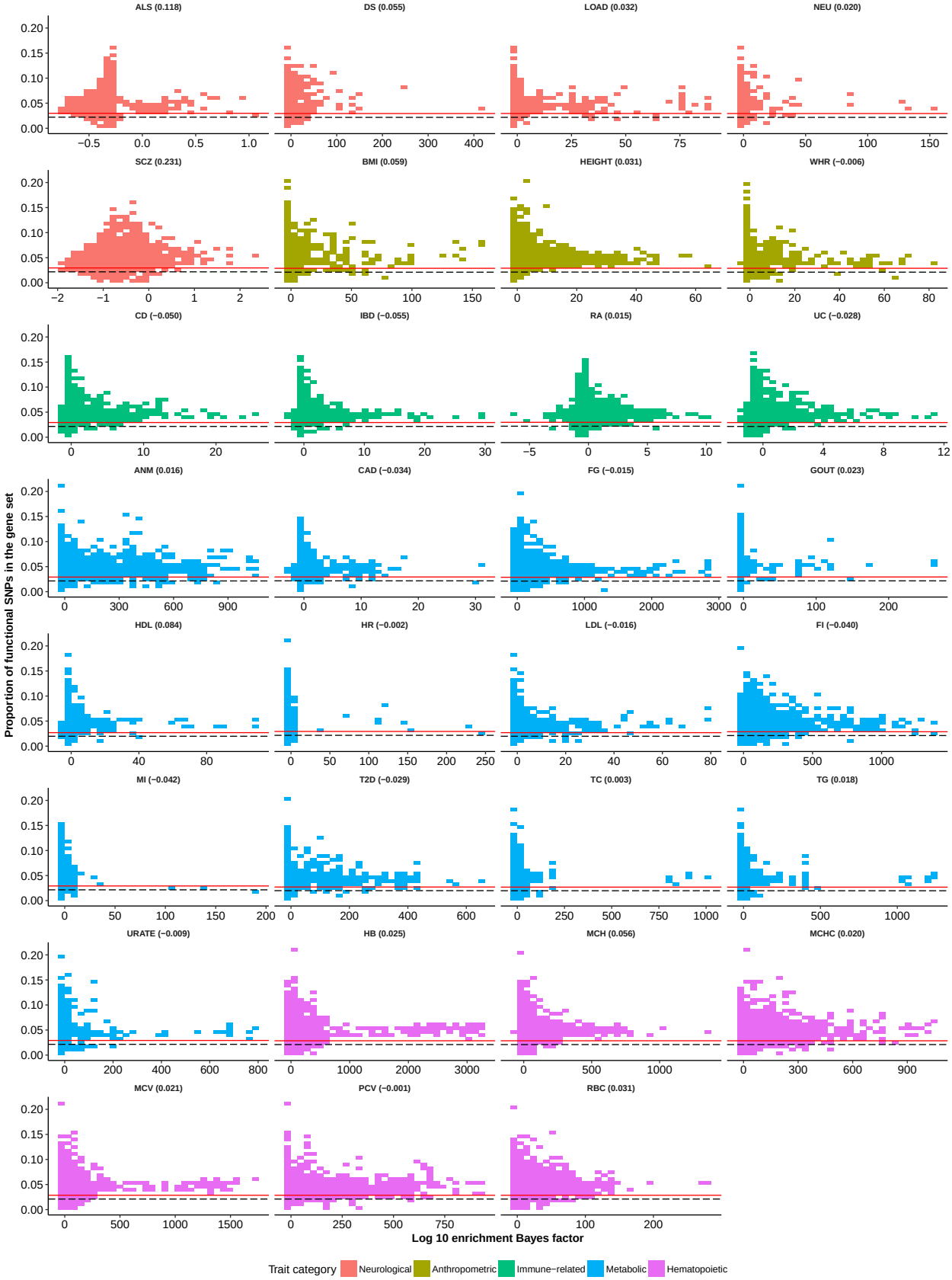
## Functional category: TSS_Hoffman



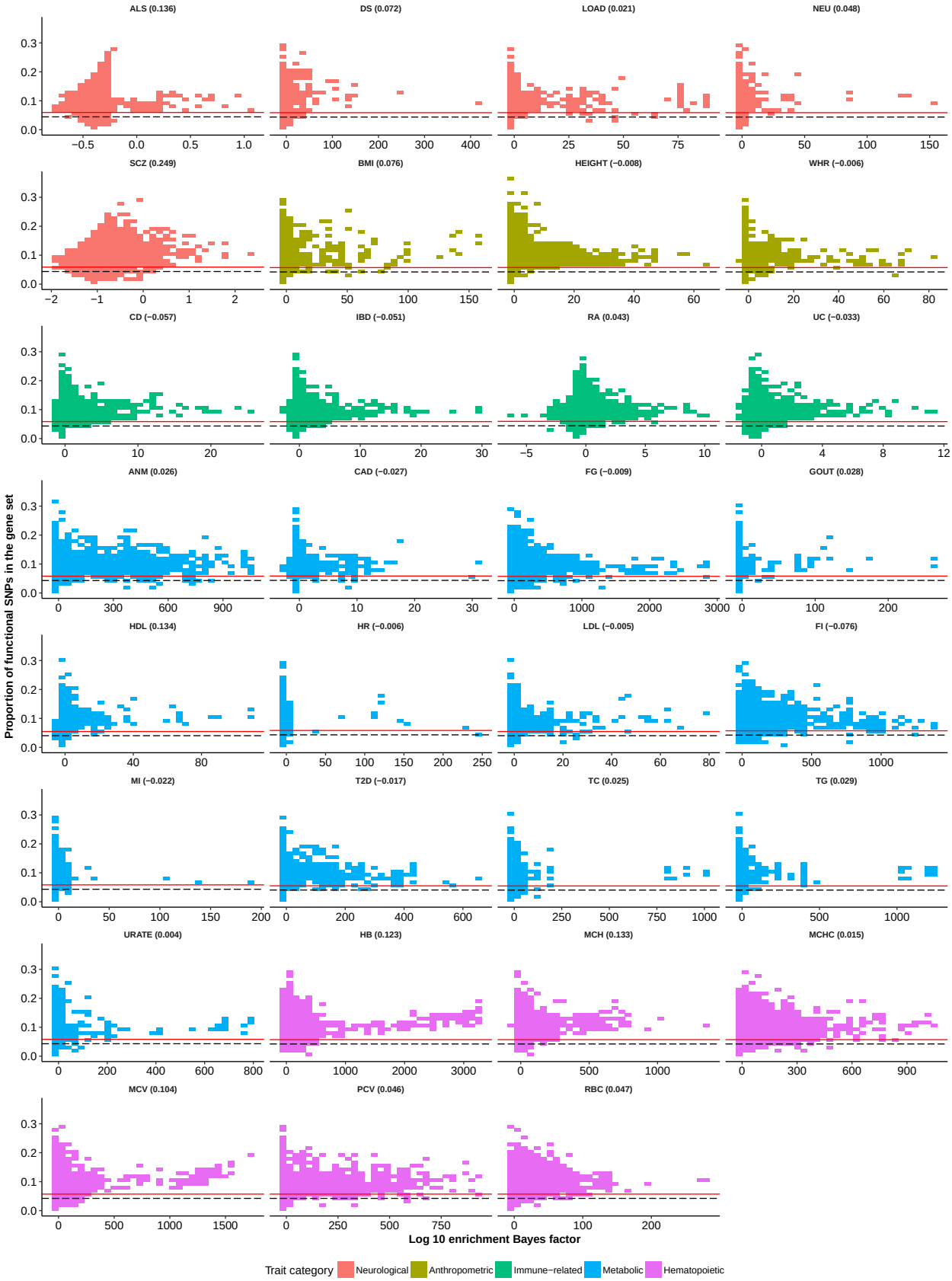
## Functional category: TSS_Hoffman.extend.500



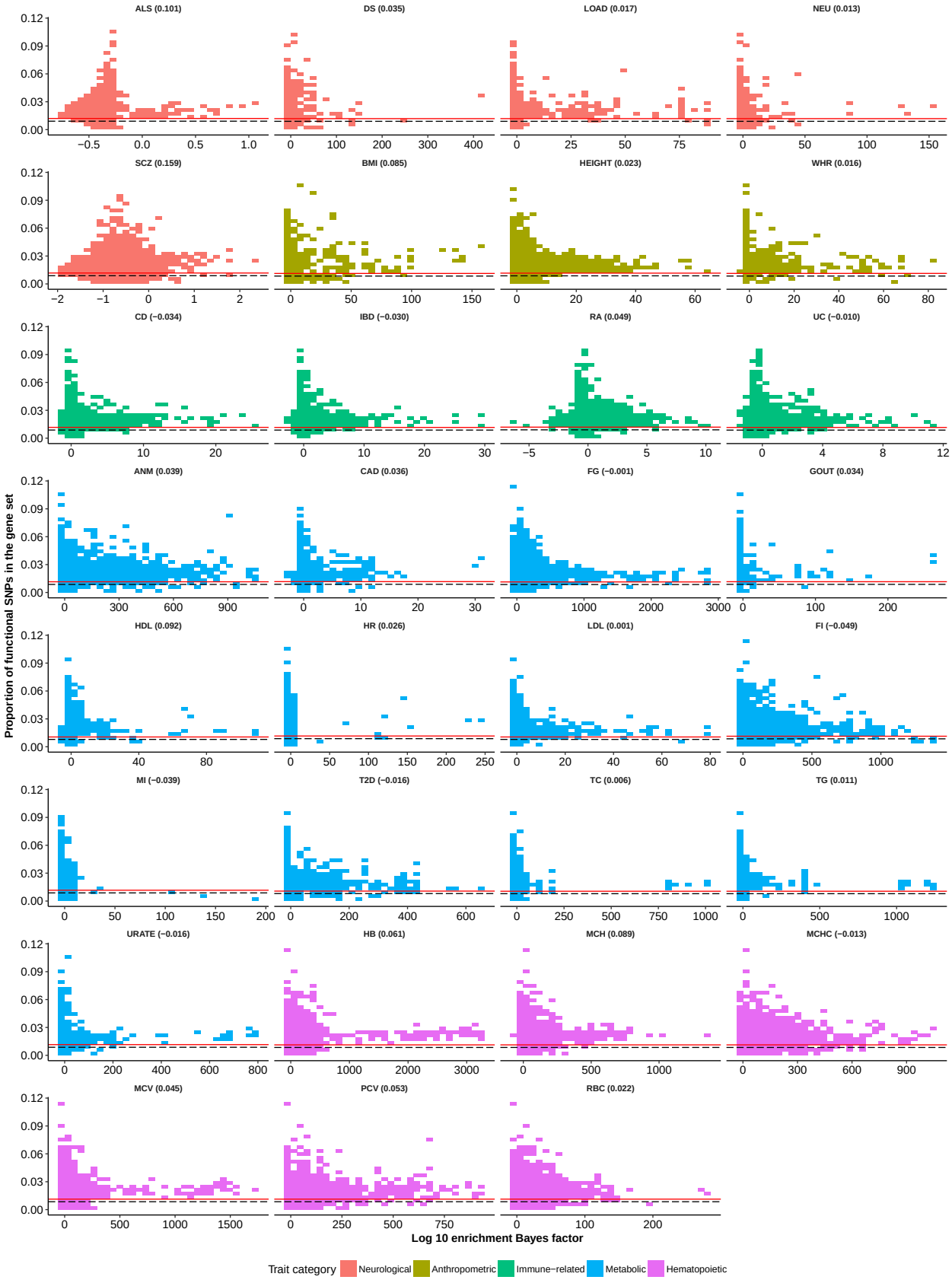
## Functional category: UTR_3_UCSC



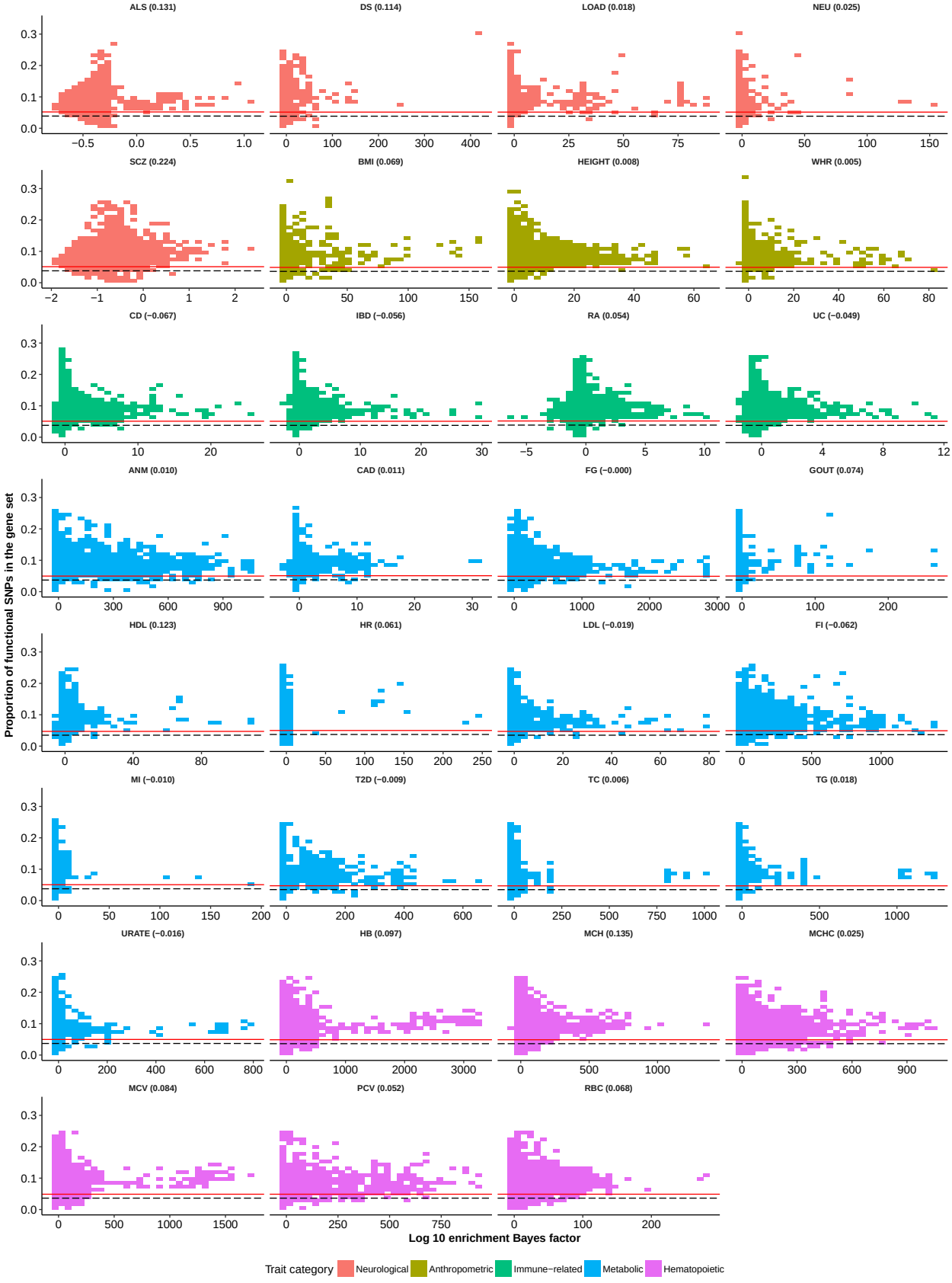
## Functional category: UTR_3_UCSC.extend.500



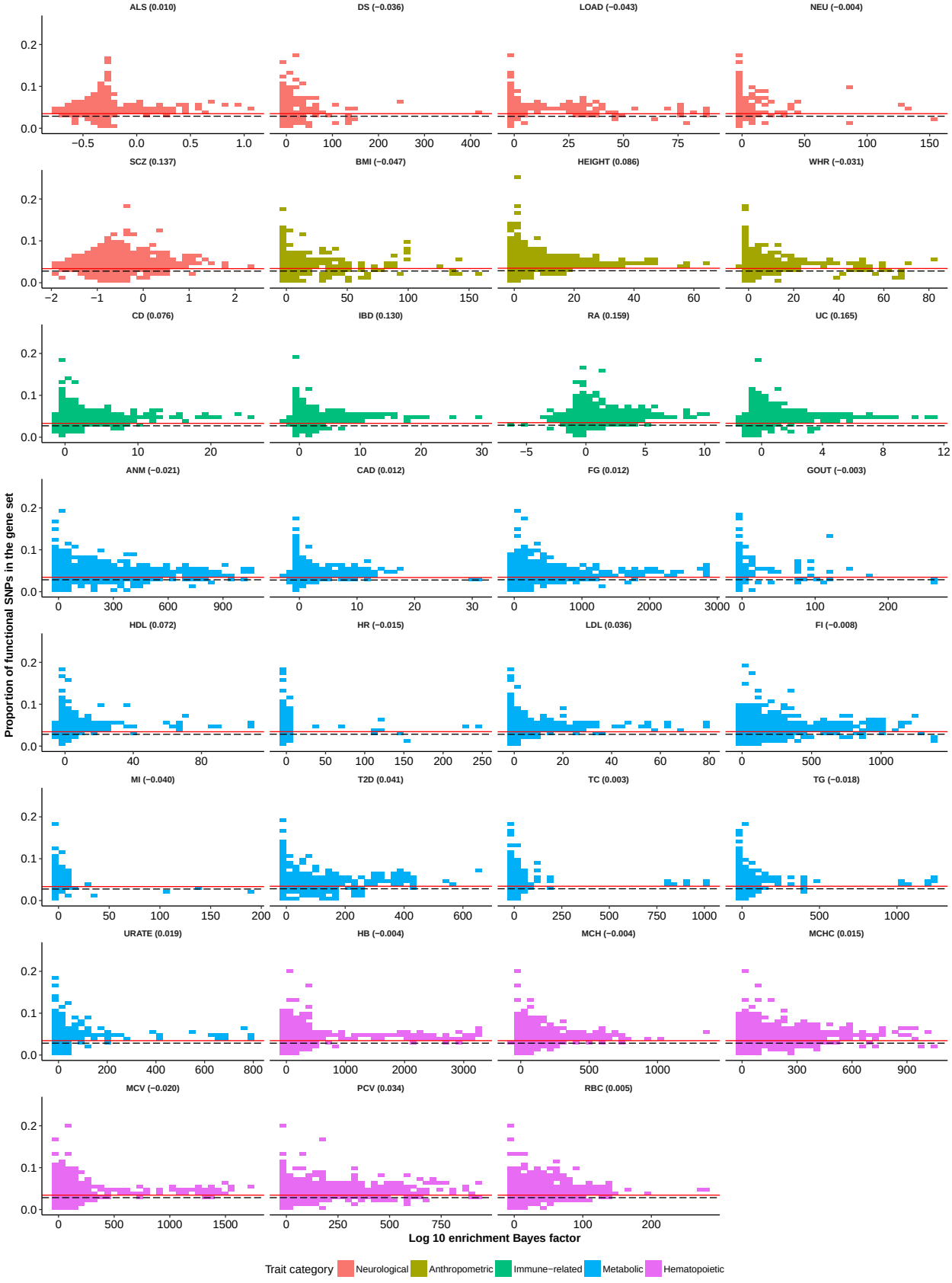
## Functional category: UTR_5_UCSC



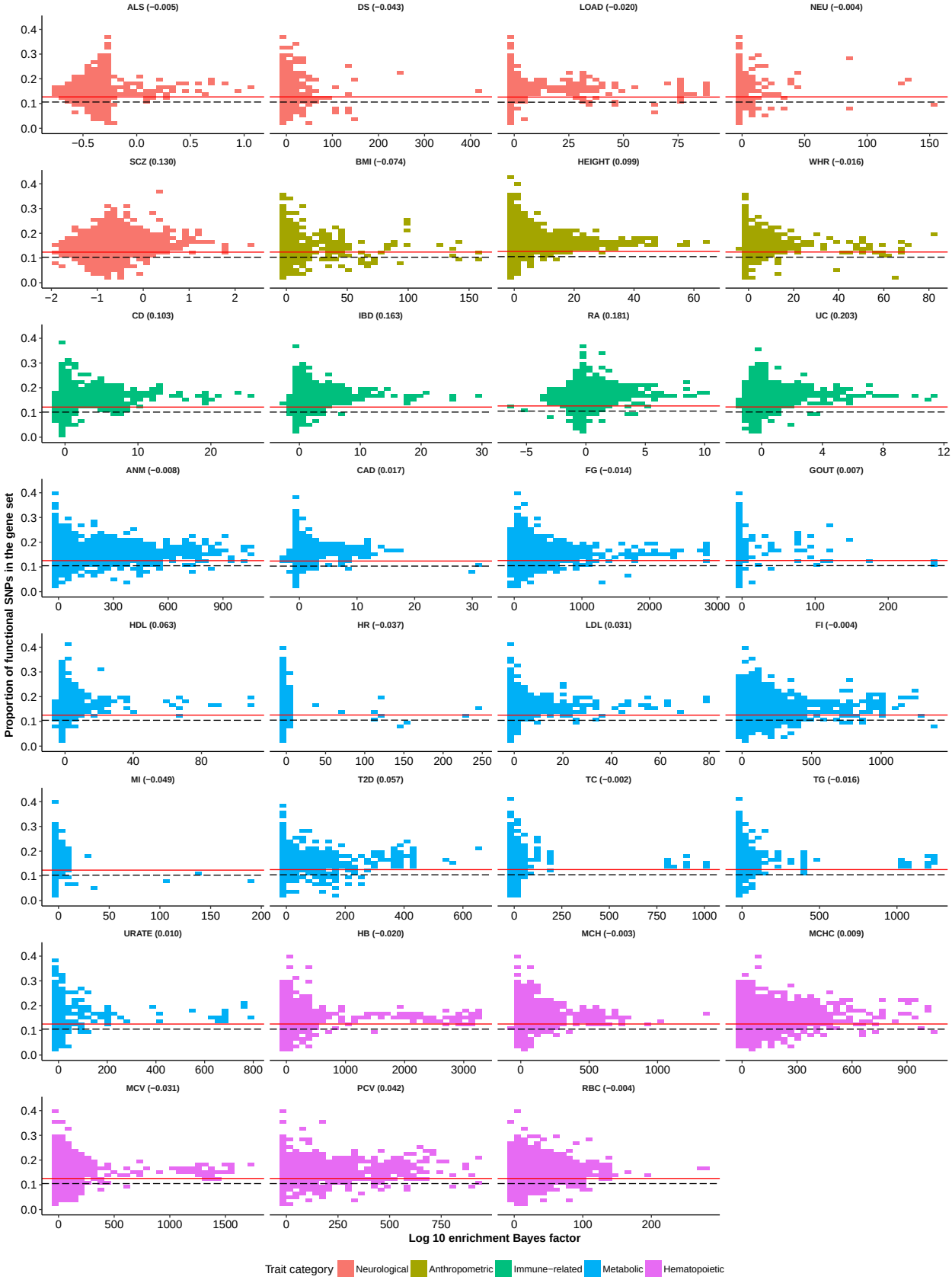
## Functional category: UTR_5_UCSC.extend.500



## Functional category: WeakEnhancer_Hoffman



## Functional category: WeakEnhancer_Hoffman.extend.500



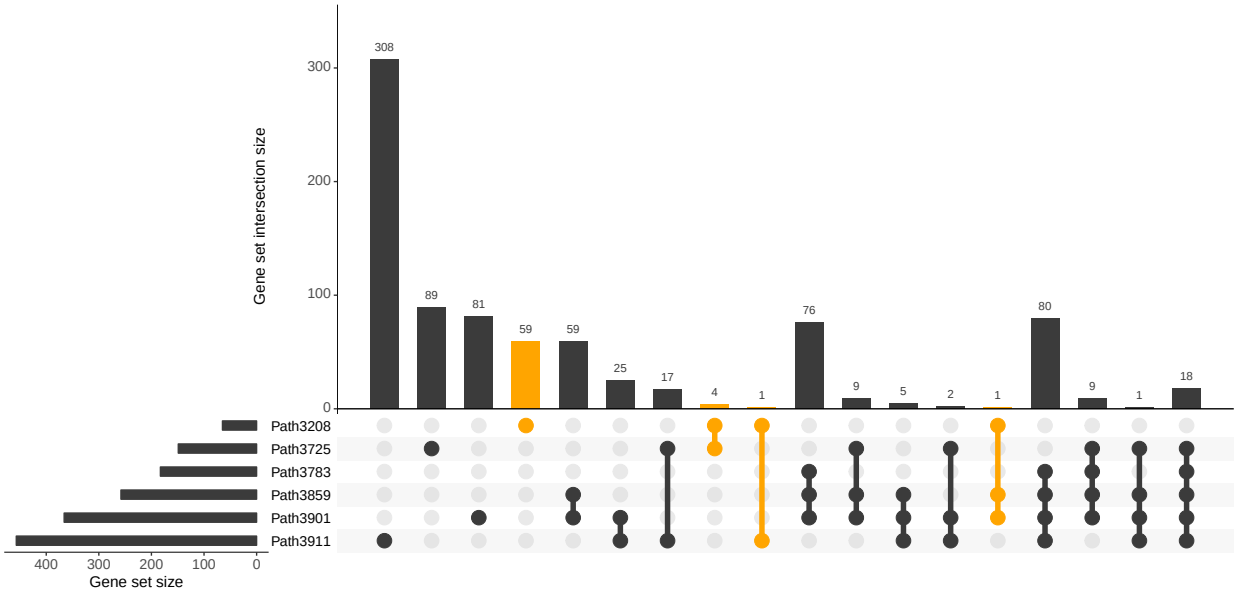
# Supplementary Figure 22

**Distribution of Bayes factors for enrichment of 3,913 biological pathways in 31 phenotypes.** These results are generated from the Round 1 analyses (see **Supplementary Table 7** for details).

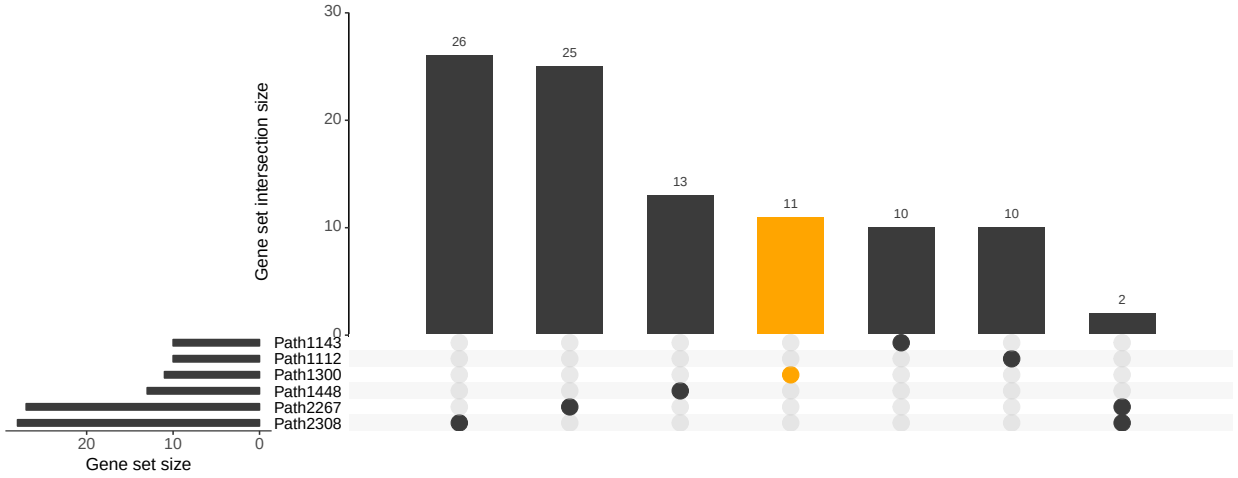
# Supplementary Figure 23

**Gene set overlap among top 6 most enriched pathways for each of 31 phenotypes.** Each barplot below shows the gene set overlap of the top 6 pathways with at least 10 member genes that show the largest enrichment Bayes factors for each trait in Round 2 analyses (see **Supplementary Table 7** for details). If multiple pathways from different databases have the same pathway description, only the one with the largest Bayes factor is displayed here. For each barplot, the yellow bars correspond to the pathway with the largest enrichment Bayes factors among the top 6 pathways. Full information about the enriched pathways can be found at `http://xiangzhu.github.io/rss-gsea/`. Intersections of top six pathways are visualized as UpSet plots (Lex et al. 2014).

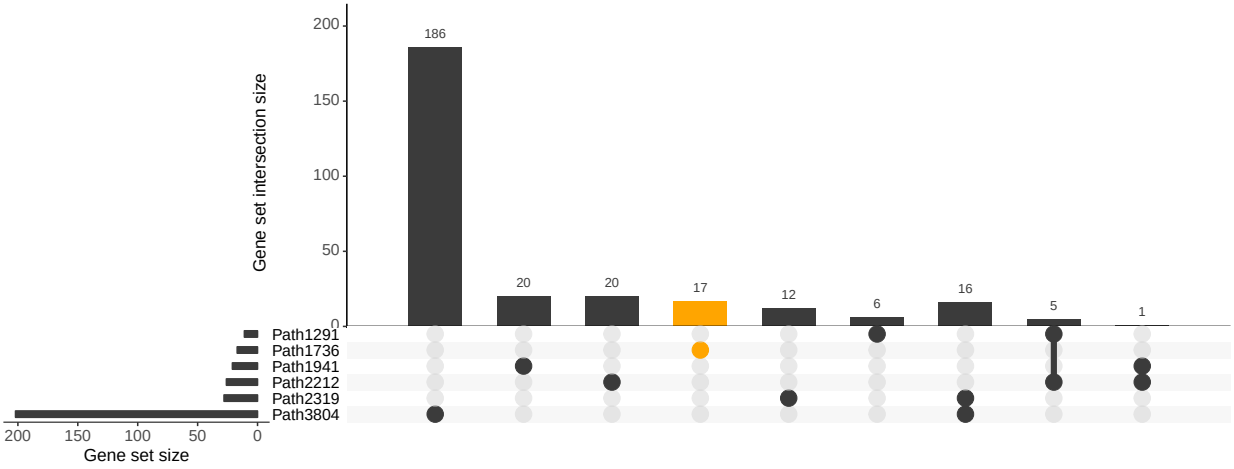
## Adult height (Wood et al. 2014).



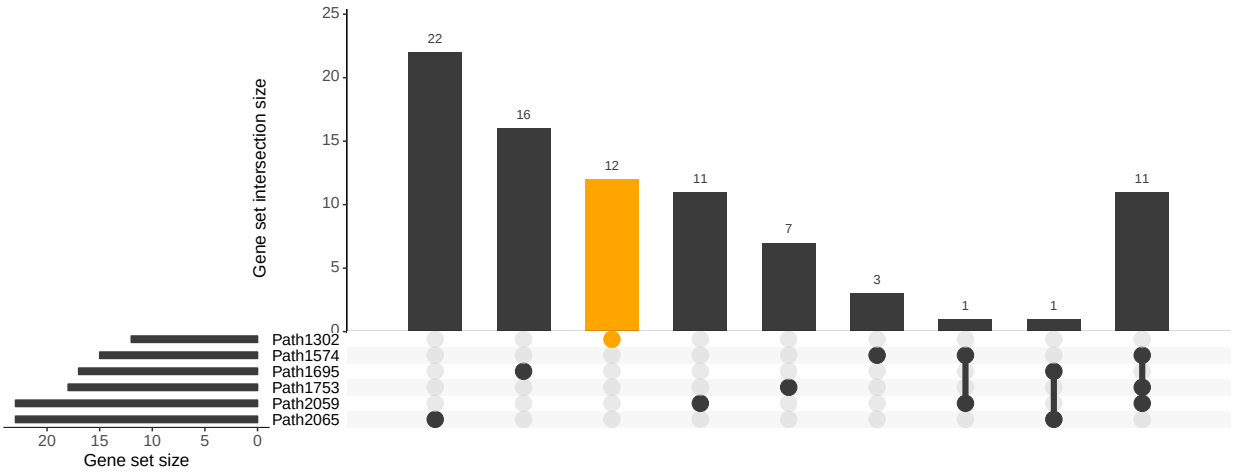

## Body mass index (Locke et al. 2015).



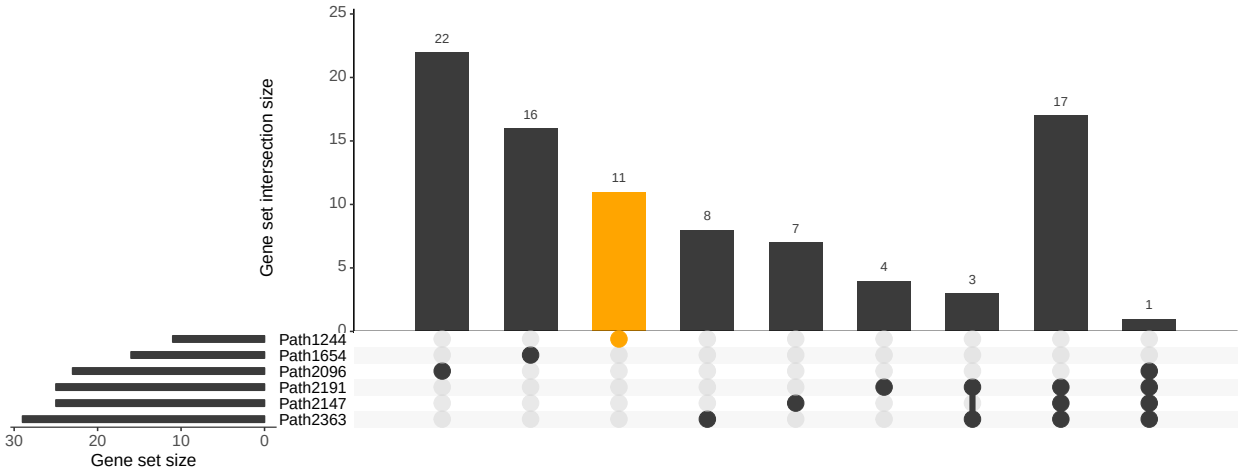
**Waist-to-hip ratio adjusted for body mass index (Shungin et al. 2015).**



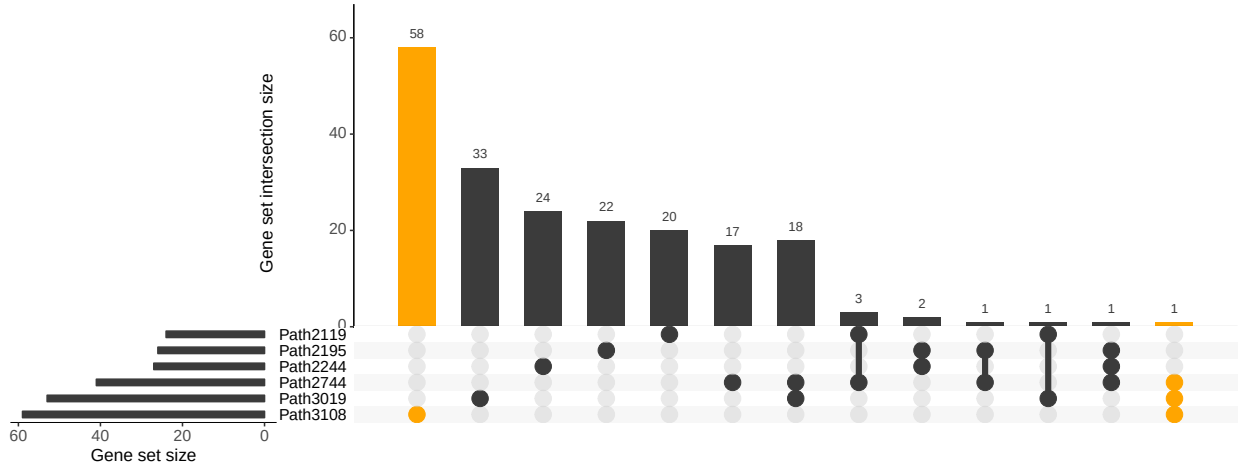

**Amyotrophic lateral sclerosis (Rheenen et al. 2016).**



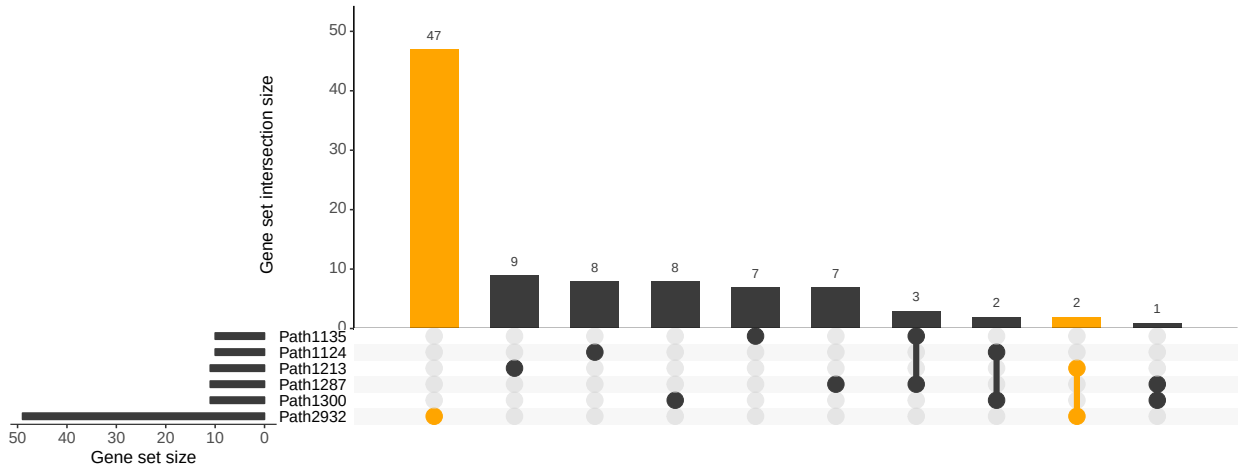
## Mean cell haemoglobin concentration (Harst et al. 2012).



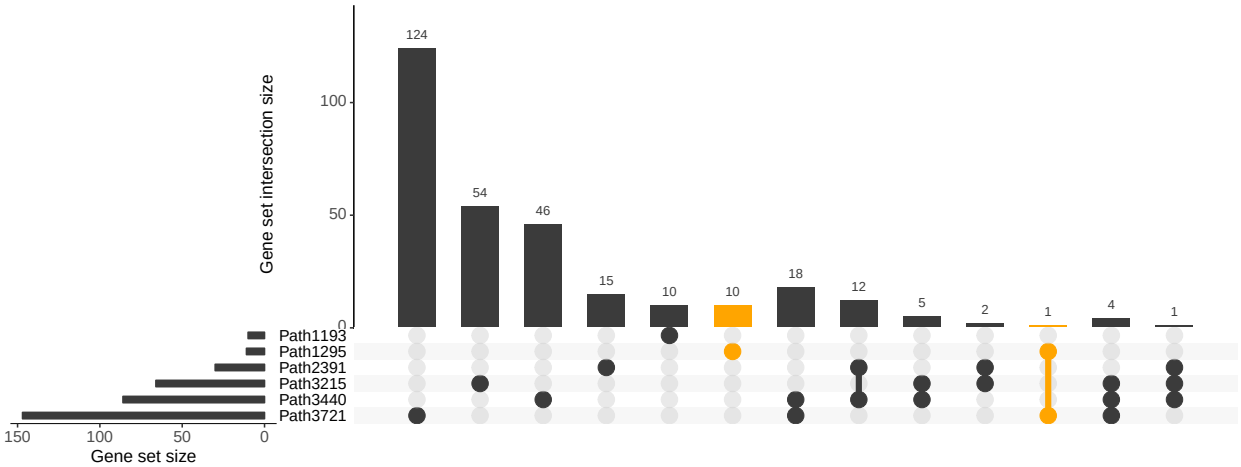

## Packed cell volume (Harst et al. 2012).



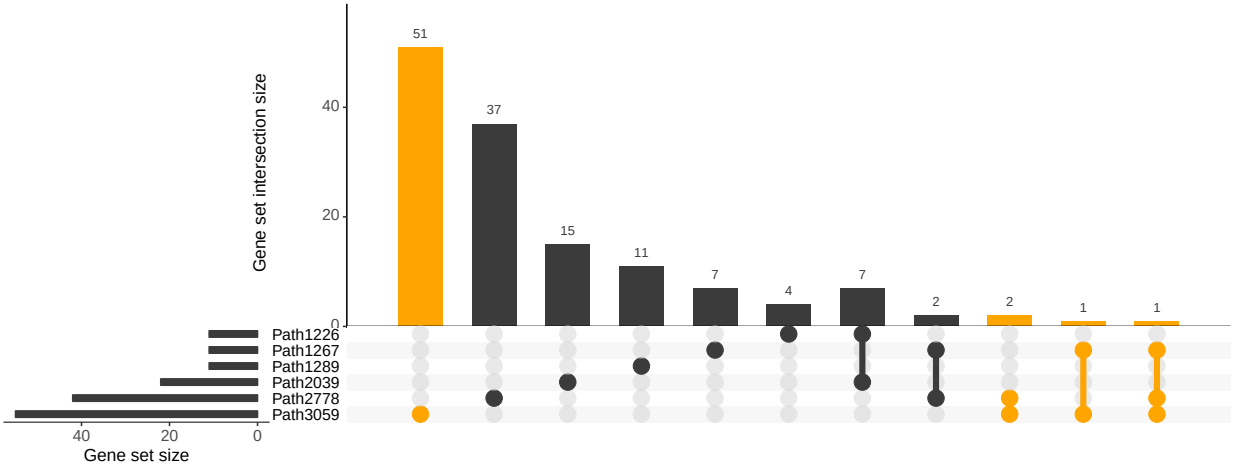
**Alzheimer's disease (Lambert et al. 2013).**



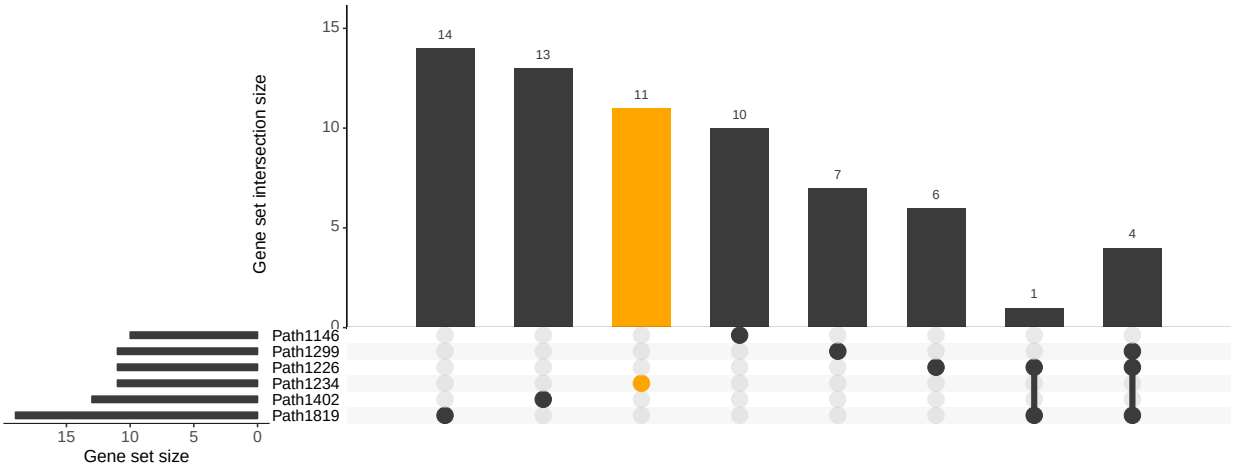

**Heart rate (Den Hoed et al. 2013).**



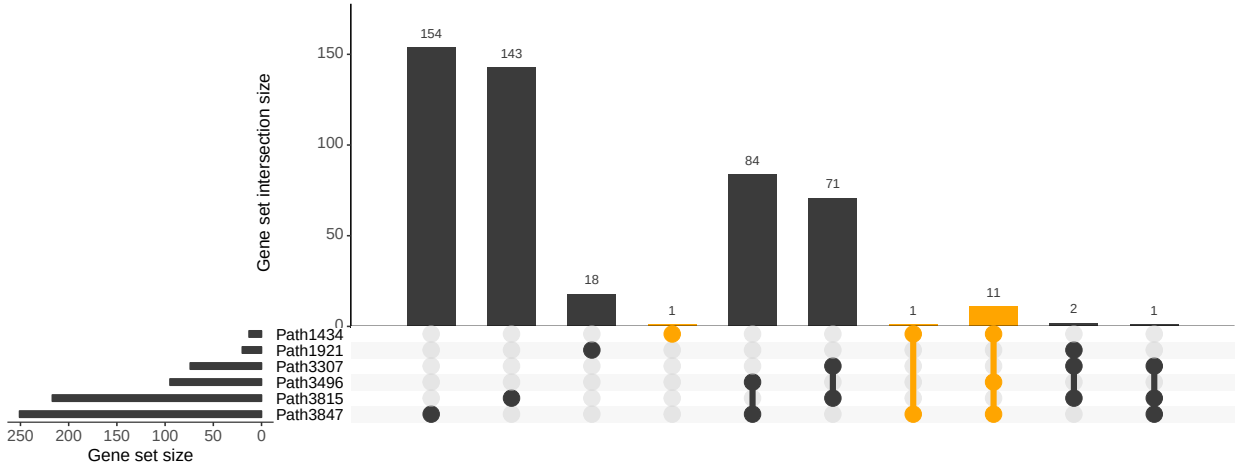
## Coronary artery disease (Nikpay et al. 2015).



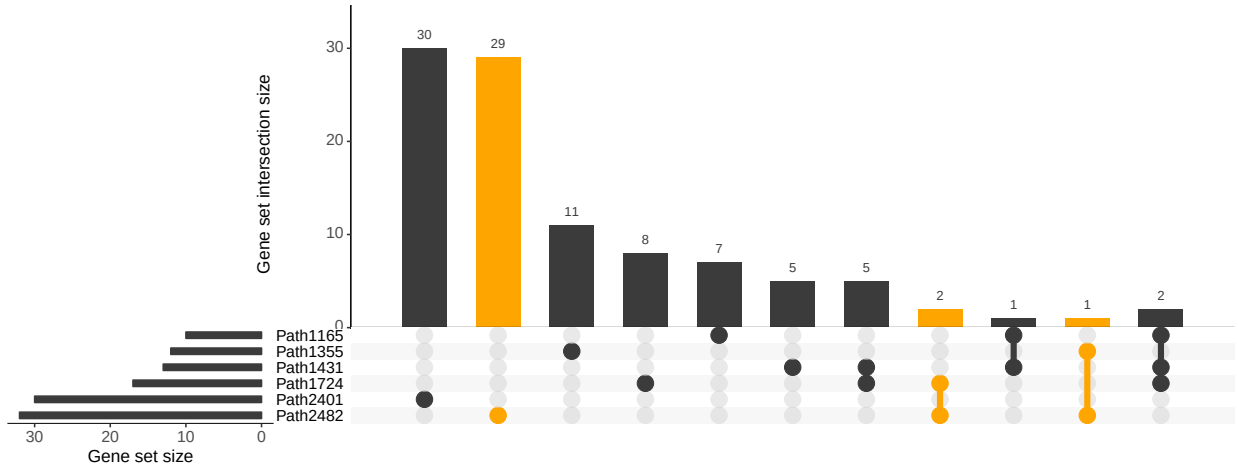

## Myocardial infarction (Nikpay et al. 2015).



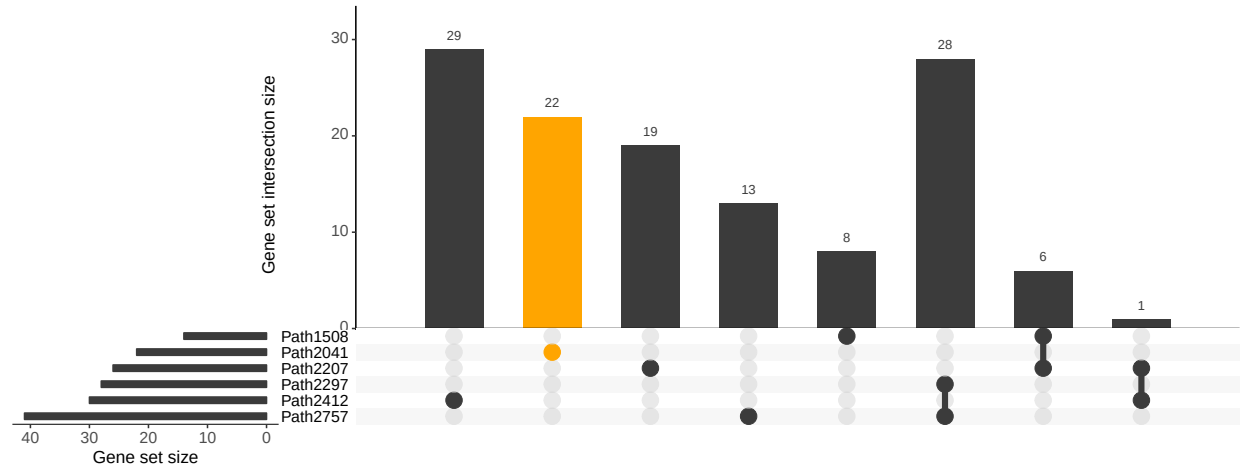
**Serum urate concentrations (Köttgen et al. 2013).**



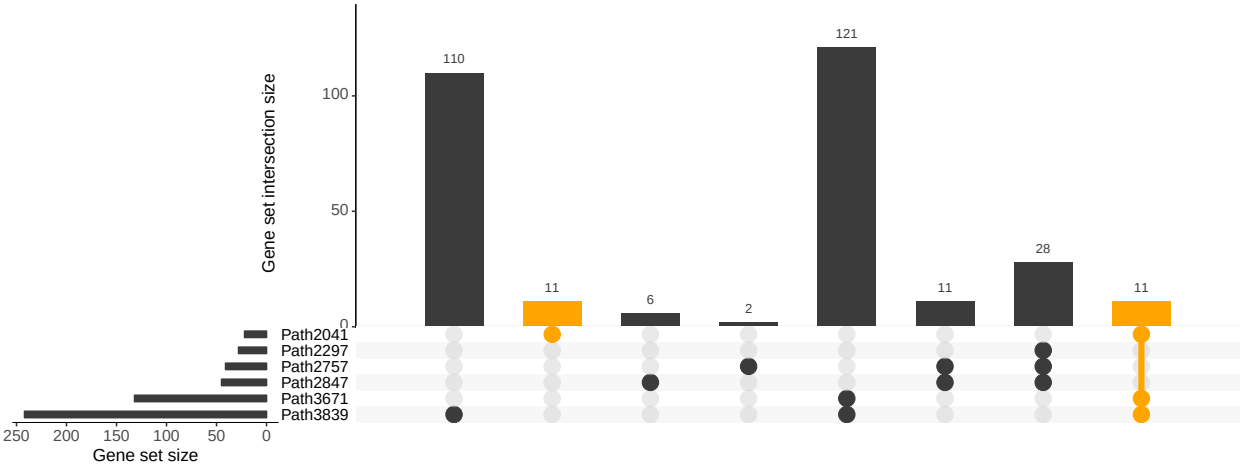

**Gout (Köttgen et al. 2013).**



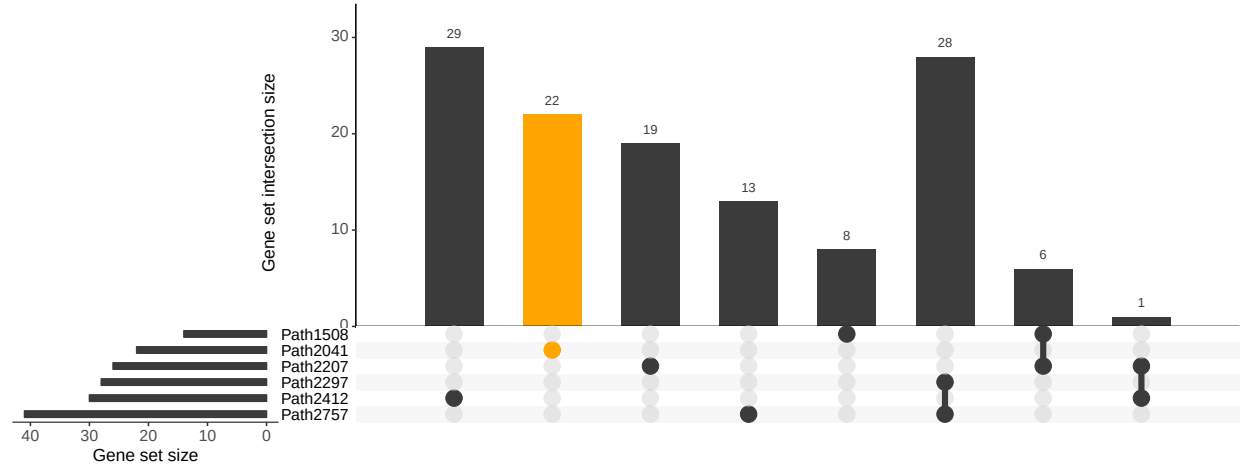
## Total cholesterol (Teslovich et al. 2010).



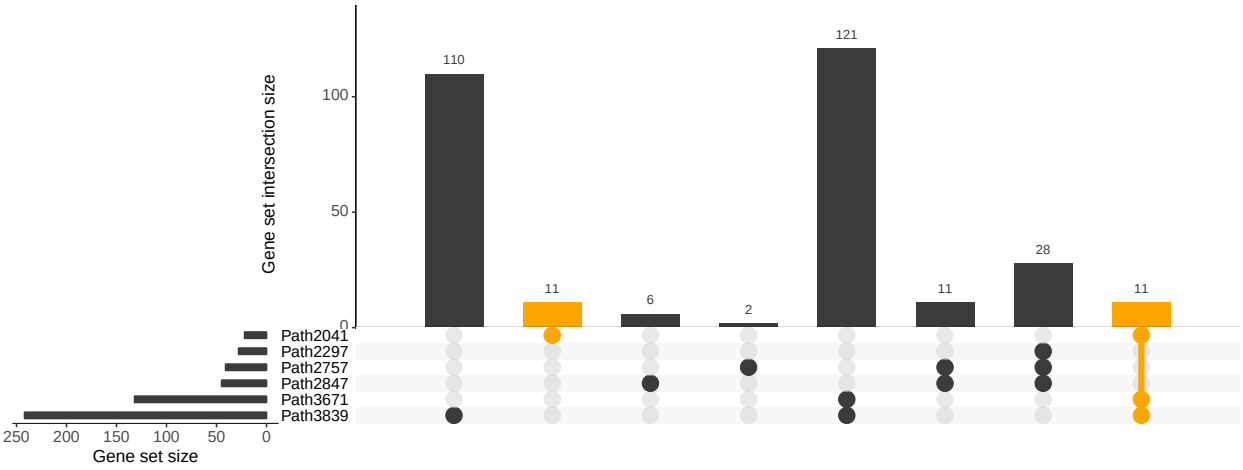

## Triglycerides (Teslovich et al. 2010).



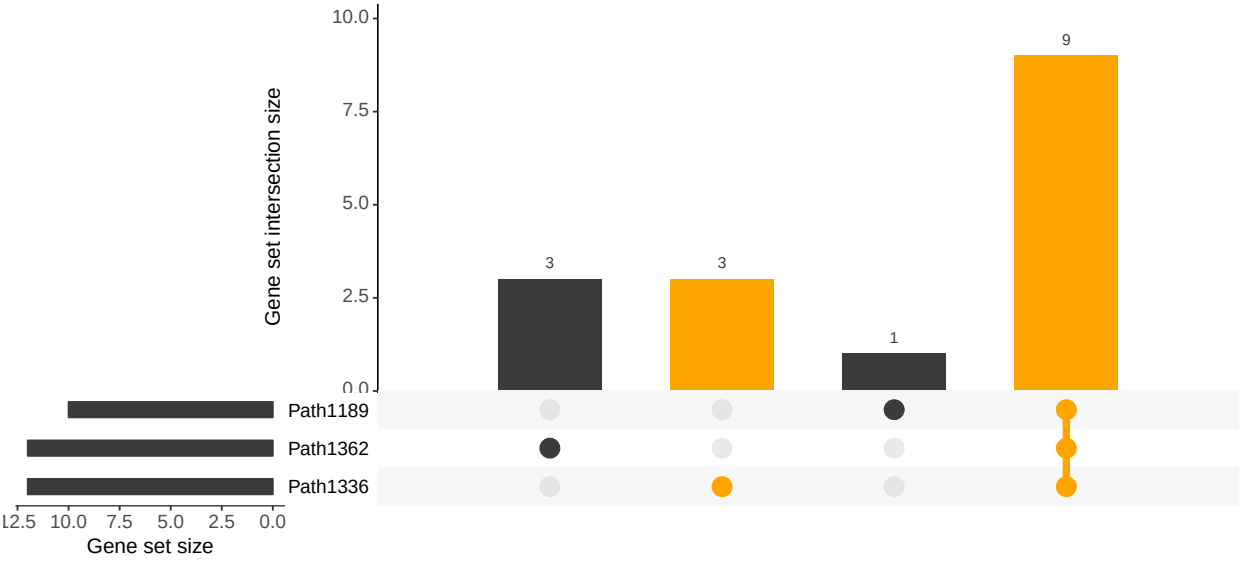
## High-density lipoprotein (Teslovich et al. 2010).



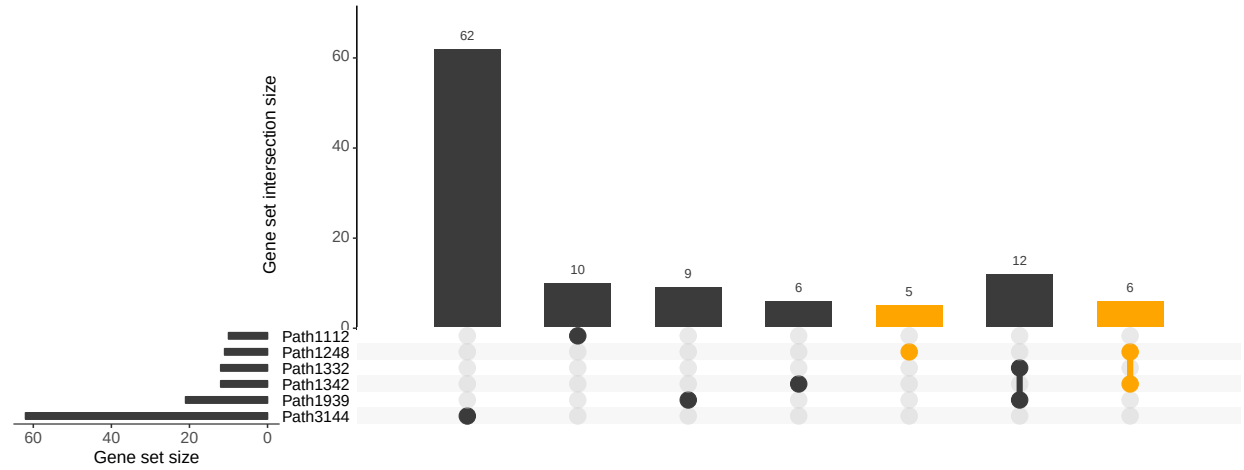

## Low-density lipoprotein (Teslovich et al. 2010).



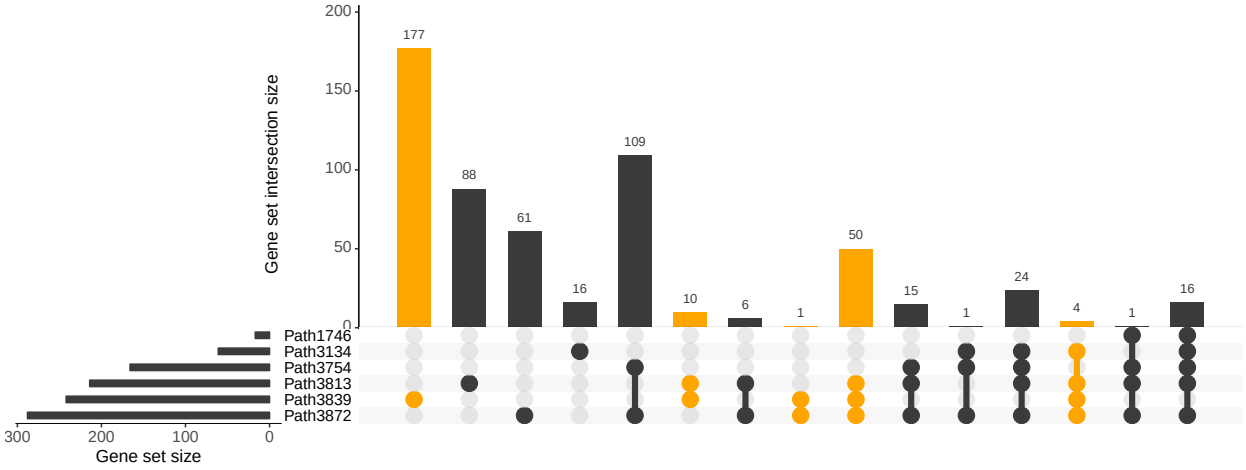
**Depressive symptoms (Okbay et al. 2016).**



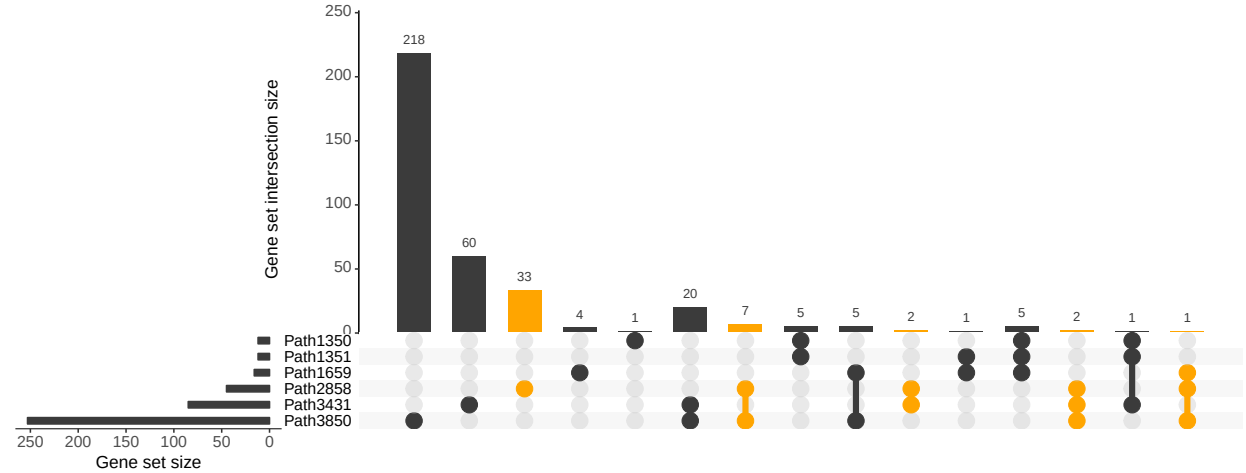

**Neuroticism (Okbay et al. 2016).**



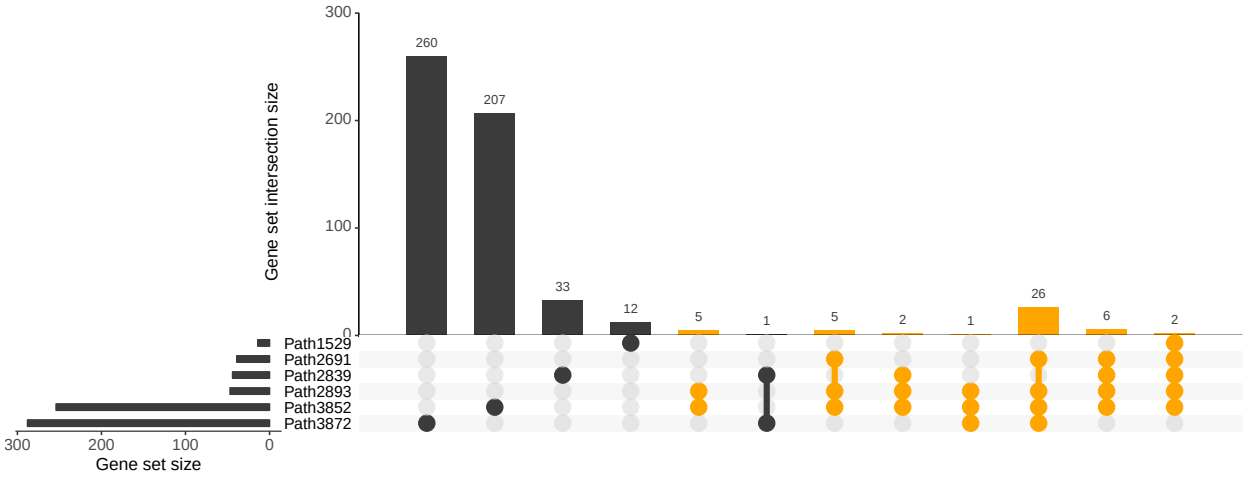
**Schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014).**



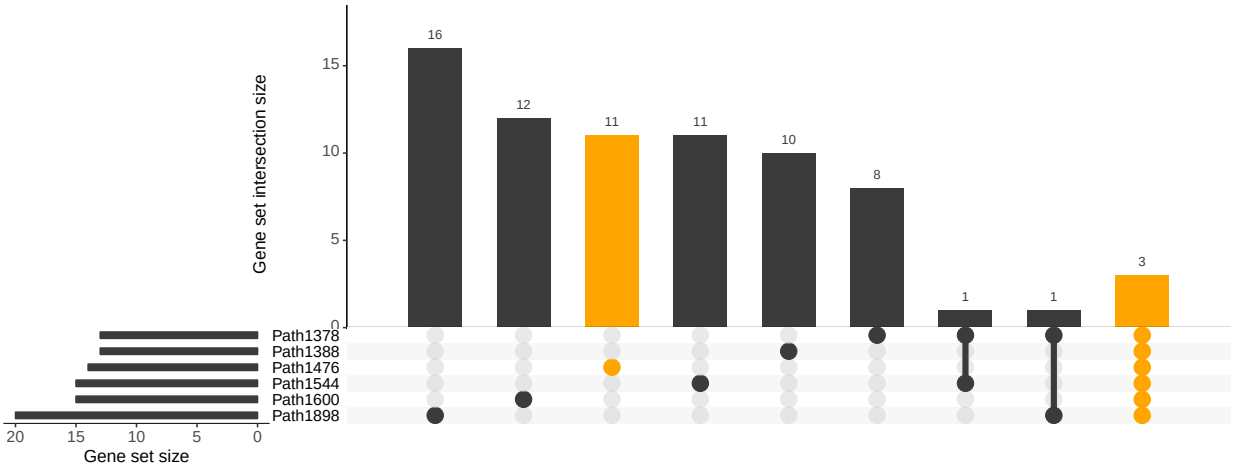

**Rheumatoid arthritis Okada et al. (2014).**



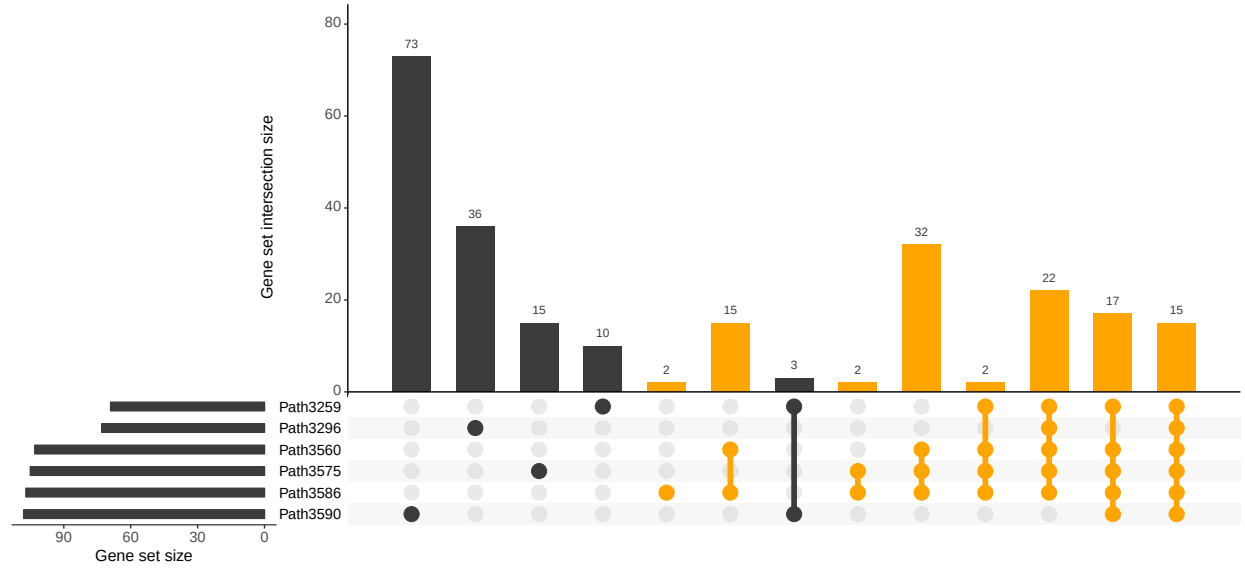
## Fasting glucose levels (Manning et al. 2012).



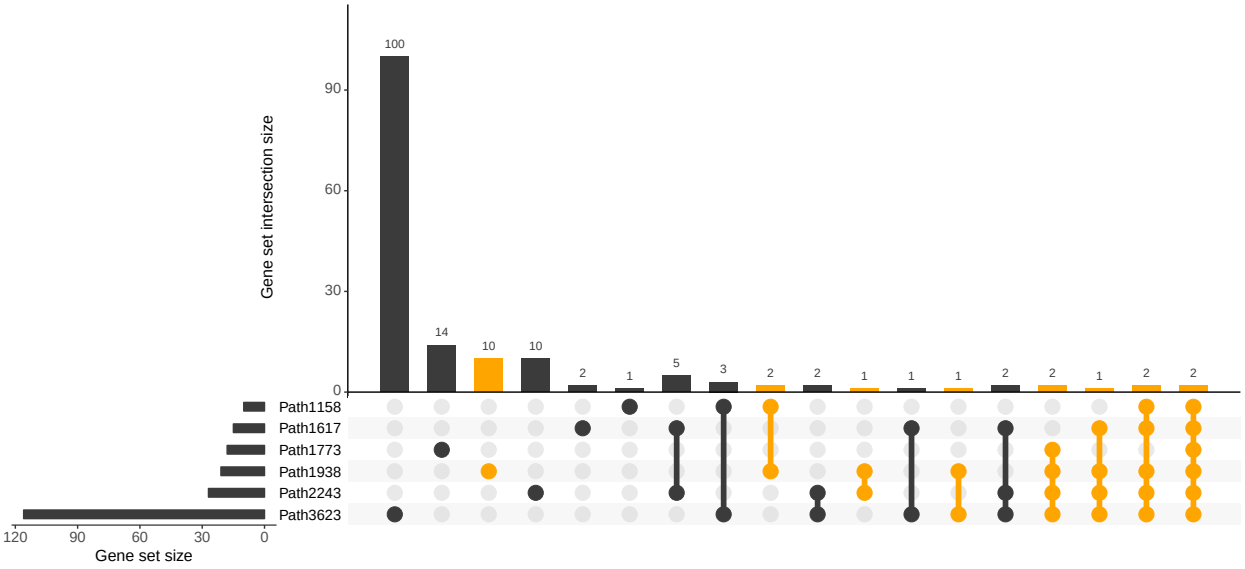

## Fasting insulin levels (Manning et al. 2012).



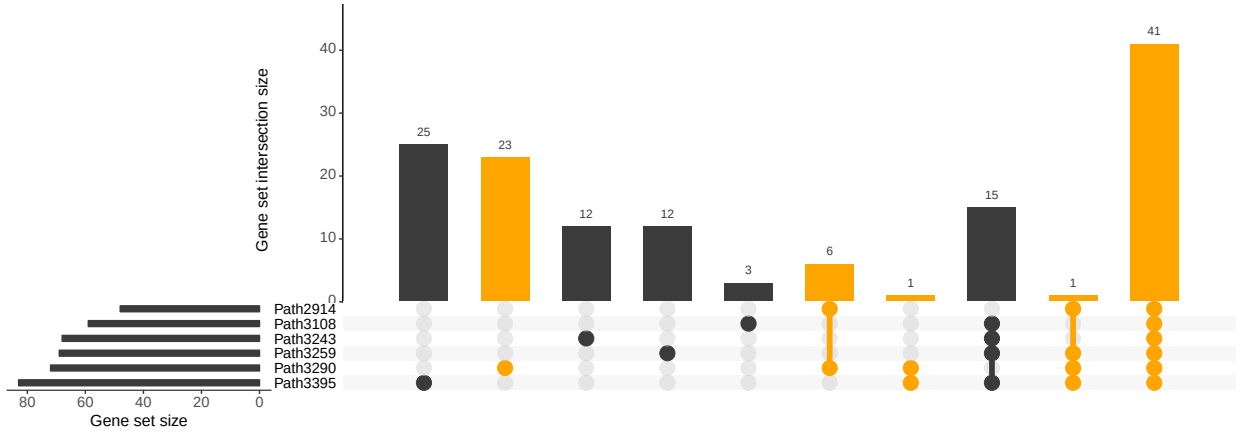
## Haemoglobin (Harst et al. 2012).



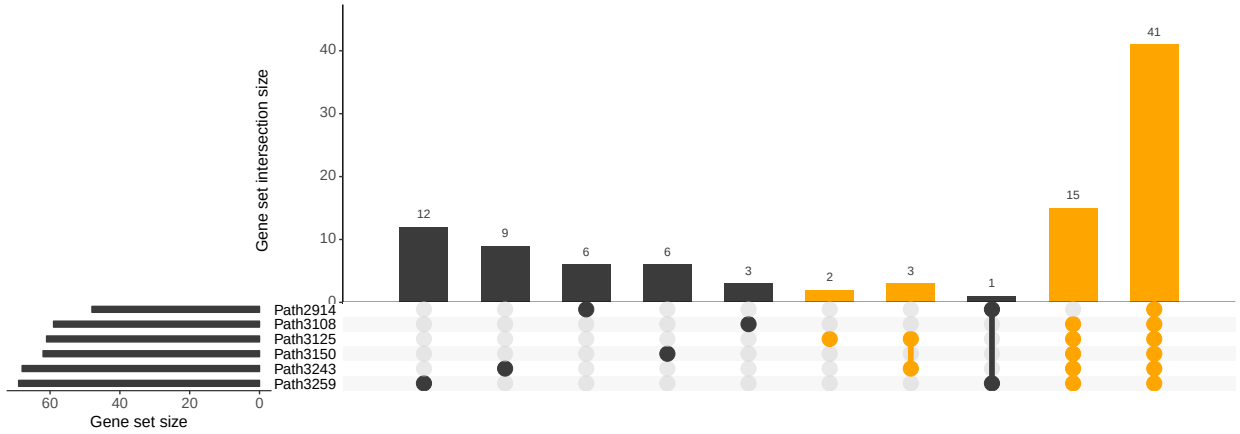

## Red blood cell count (Harst et al. 2012).



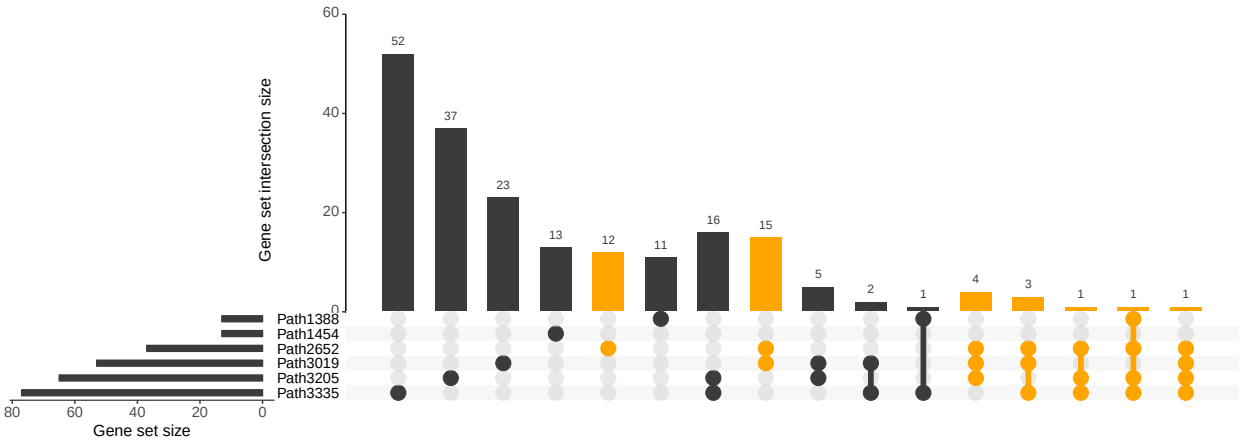
## Mean cell haemoglobin (Harst et al. 2012).



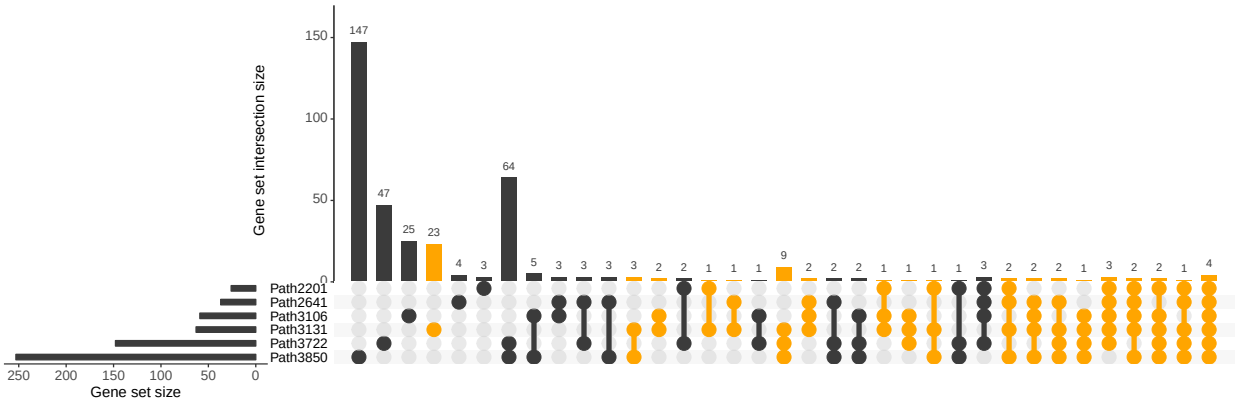

## Mean cell volume (Harst et al. 2012).



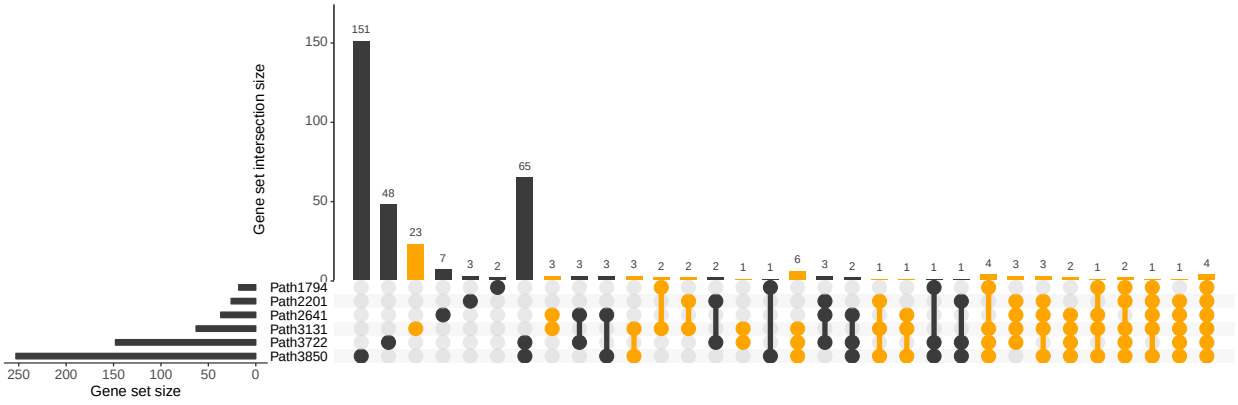

## Age at natural menopause (Day et al. 2015).



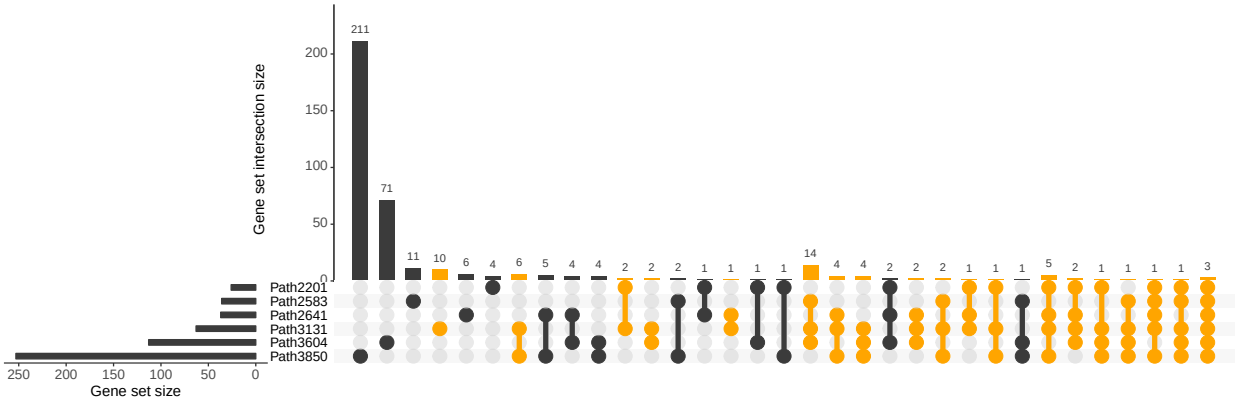
**Inflammatory bowel disease (Liu et al. 2015).**




**Crohn's disease (Liu et al. 2015).**

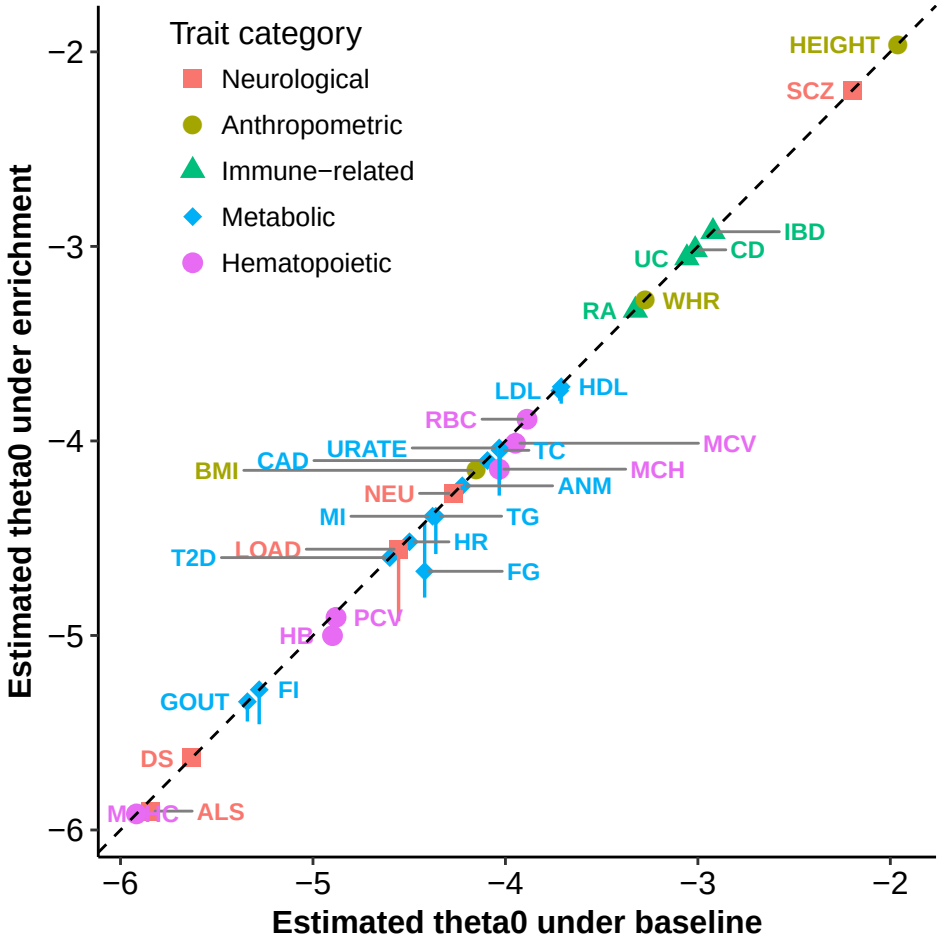



**Ulcerative colitis (Liu et al. 2015).**



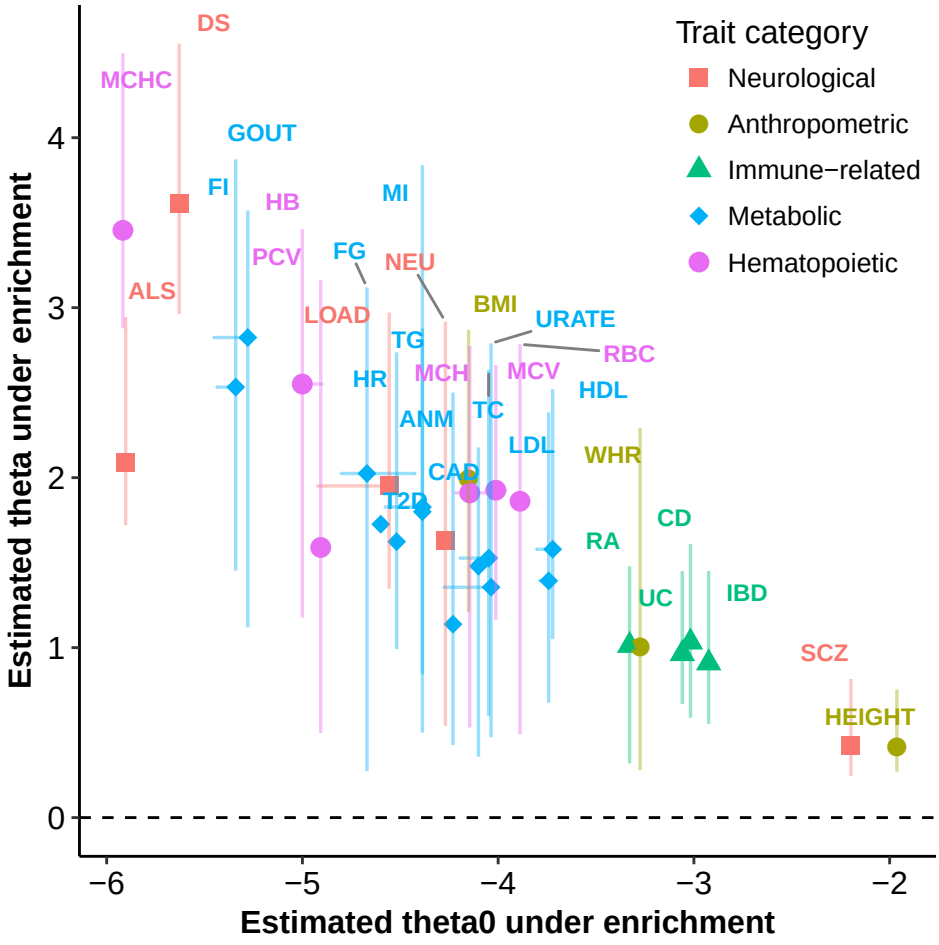
## Supplementary Figure 24

**Estimated genome-wide background parameter ($\theta_0$) under the baseline ($M_0$) and enrichment ($M_1$) models across 31 traits.** Each dot represents a trait, with the $x$-axis value being the posterior mean of $\theta_0$ estimated under $M_0$ and a vertical point range indicating the (2.5, 50, 97.5)-percentile for posterior means of $\theta_0$ across 100 top-ranked gene sets under $M_1$. The dashed line has slope one and intercept zero. The tight vertical point ranges demonstrate that RSS gives almost identical estimates of the background parameter $\theta_0$ in both baseline and enrichment analyses.



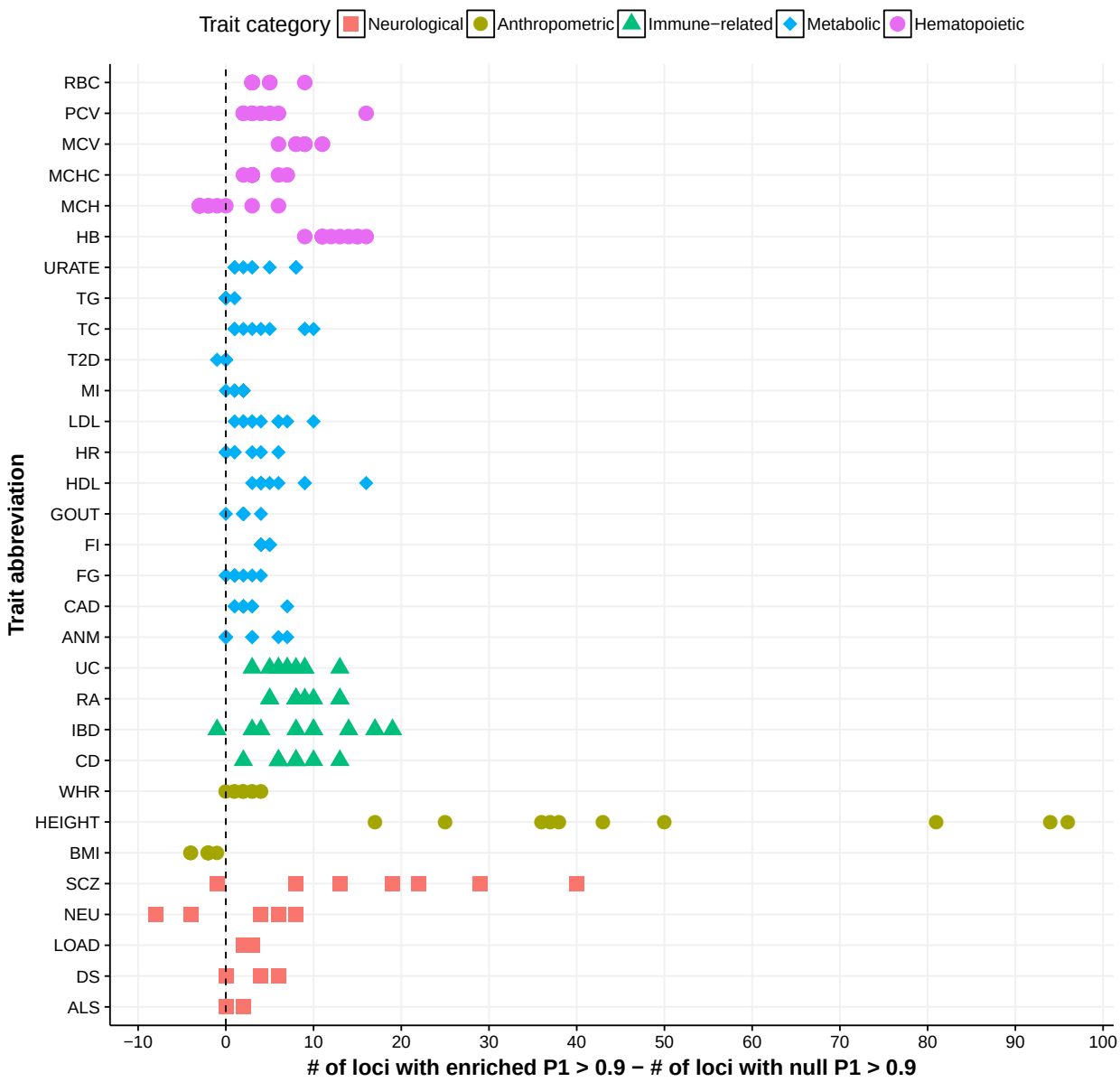
## Supplementary Figure 25

**Estimated enrichment parameter ($\theta$) versus the background parameter ($\theta_0$) under the enrichment model ($M_1$) across 31 traits.** Each dot represents a trait, with horizontal and vertical point ranges indicating the (2.5, 50, 97.5)-percentiles for posterior means of $\theta_0$ and $\theta$ across 100 top-ranked gene sets. This plot demonstrates that RSS yields a positive estimate of enrichment parameter $\theta$ if the gene set is identified as enriched.



# Supplementary Figure 26

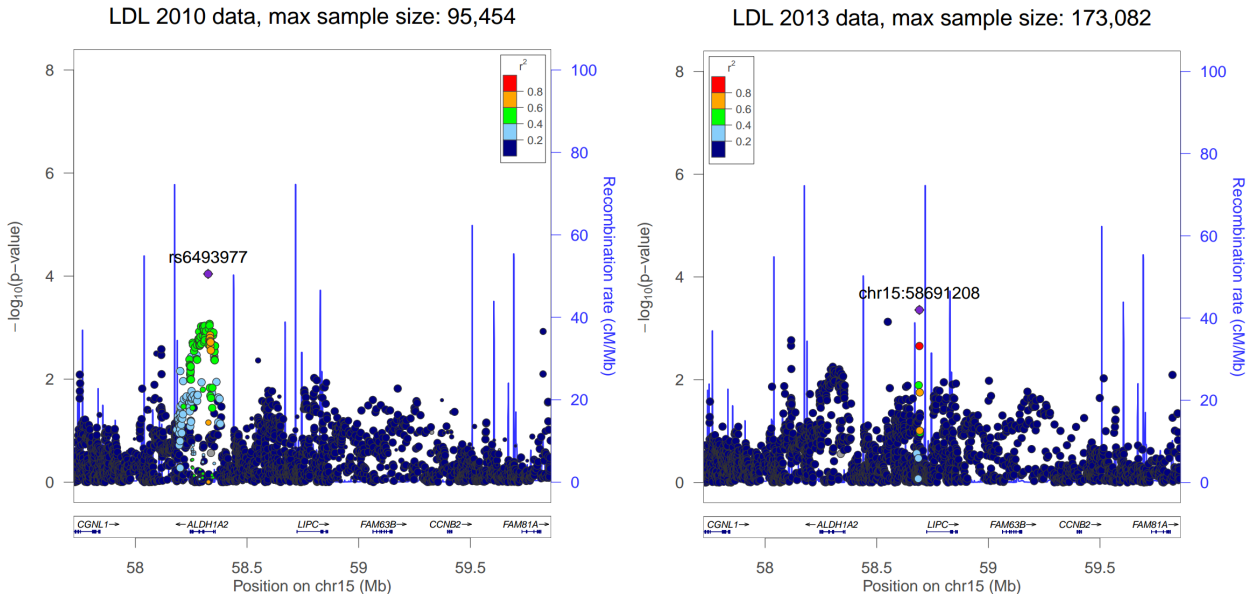
**Compare the number of trait-associated loci detected under the baseline hypothesis with the number of trait-associated loci detected under the enrichment hypothesis.** The baseline hypothesis assumes that no pathways are enriched. The enrichment hypothesis assumes that a candidate pathway is enriched. For each phenotype, each dot corresponds to one of the top 10 most enriched pathways in Round 2 analyses (shown in **Supplementary Figure 17**). A positive value in $x$-axis indicates that more trait-associated loci ($P_1 > 0.9$) are identified under the enrichment hypothesis than under the null hypothesis. See **Supplementary Figure 10** for the definition of locus and $P_1$. Note that some dots are overlapped due to their similar values in $x$-axis.



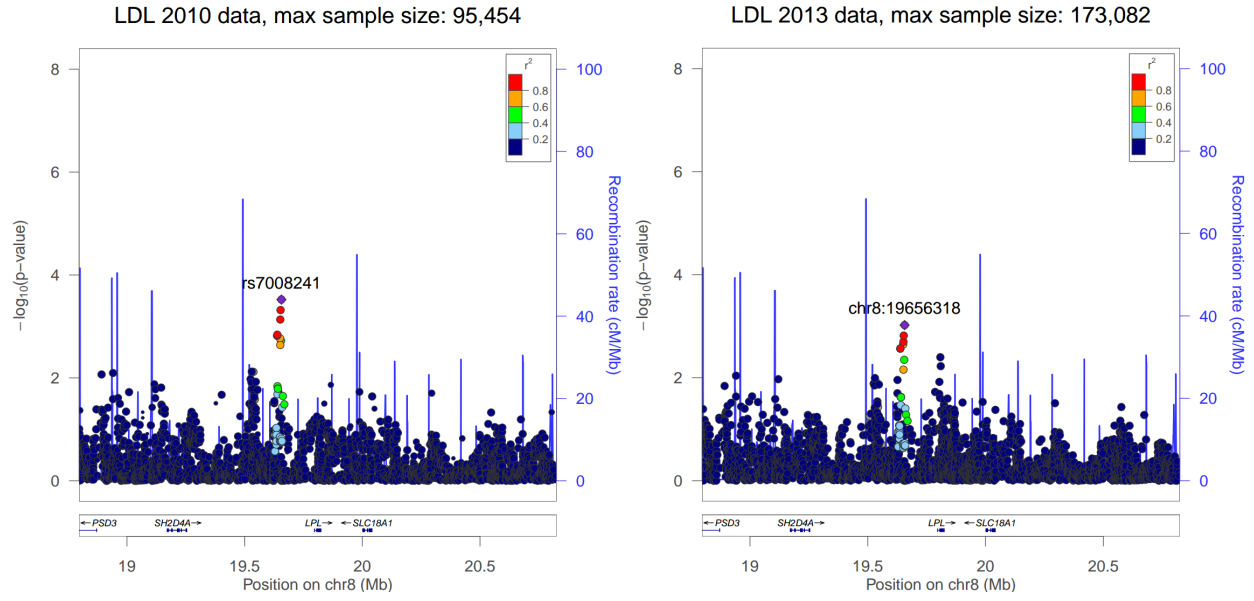
# Supplementary Figure 27

**Regional association plots of genes *LIPC* and *LPL* based on single-SNP summary data of low-density lipoprotein cholesterol levels.** The "2010 data" were reported in Teslovich et al. (2010). The "2013 data" were reported in Global Lipids Genetics Consortium (2013).

**(a)** Regional association plot of *LIPC*.

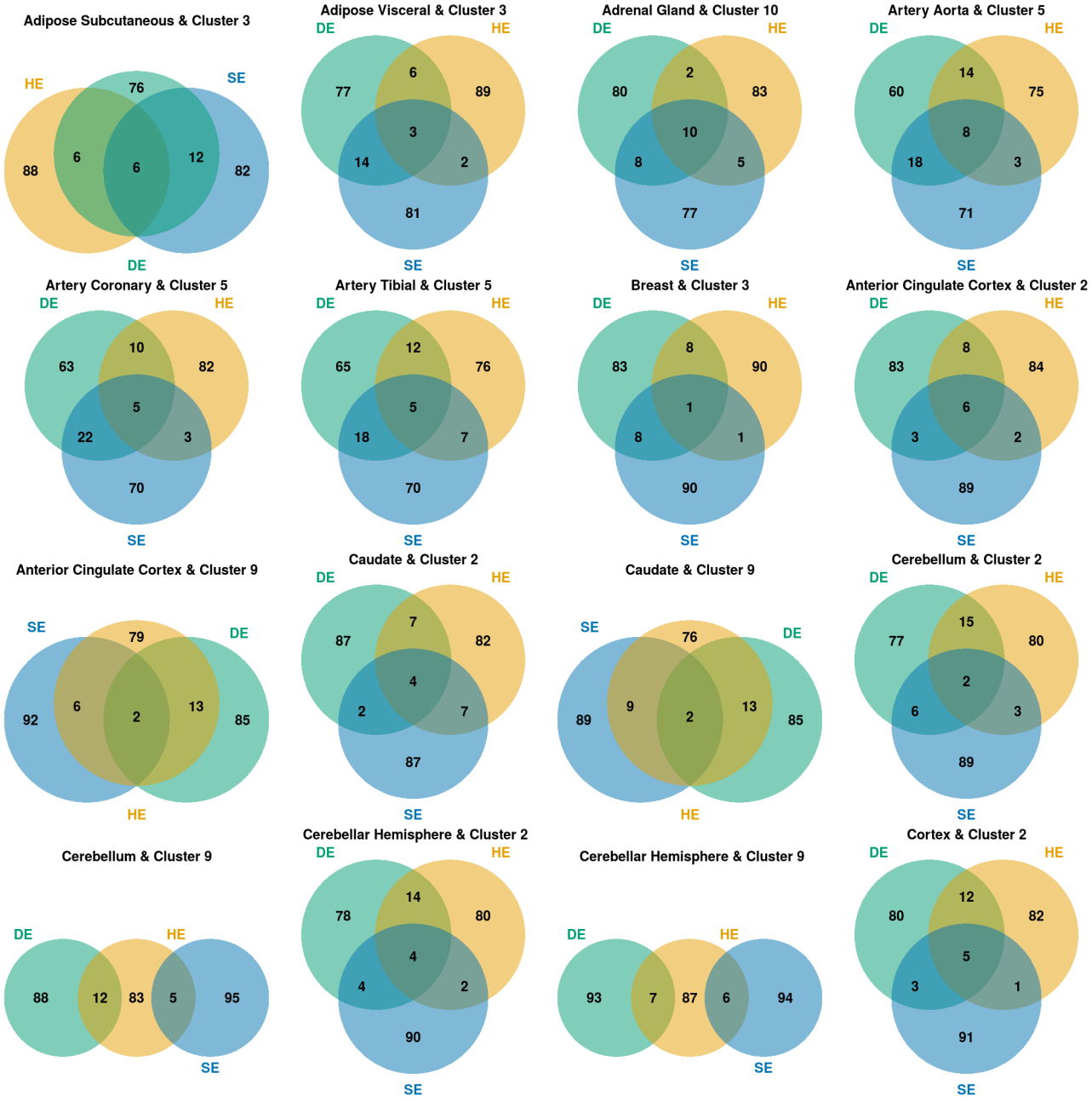


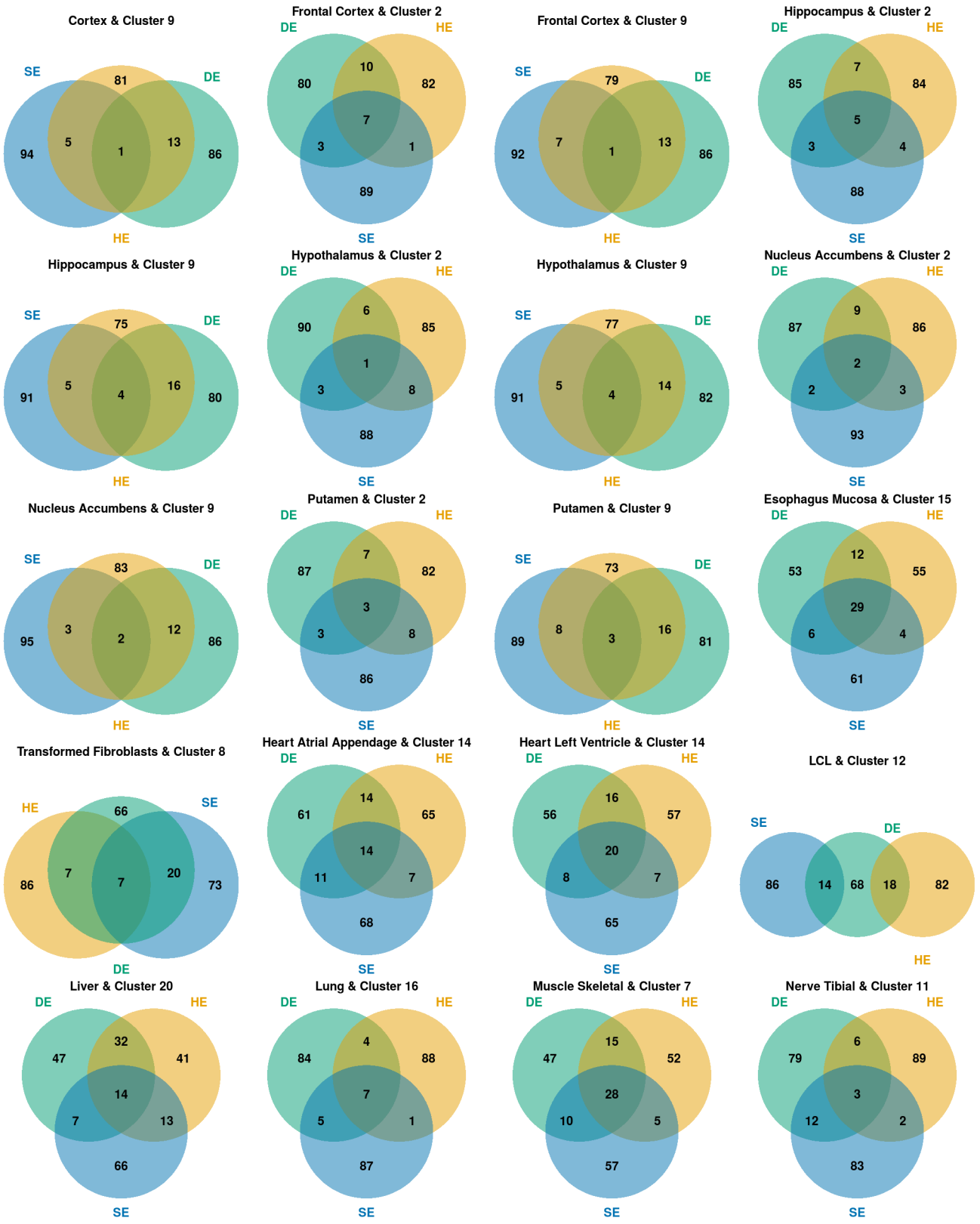

**(b)** Regional association plot of *LPL*.



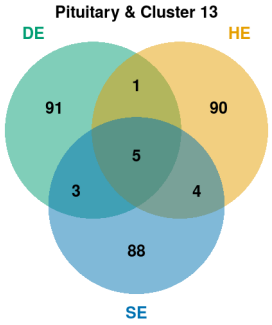
# Supplementary Figure 28

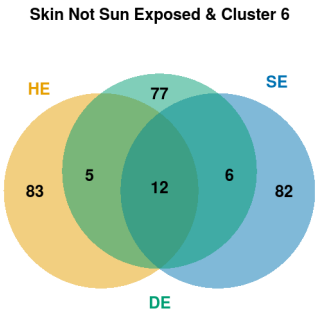
**Venn diagrams showing that three methods yield different sets of genes for a given tissue.** For each tissue, we use three complementary approaches to defining tissue-relevant genes based on GTEx RNA-seq data. The first approach ("highly expressed", HE) uses the highest expressed genes in each tissue. The second approach ("selectively expressed", SE) uses a tissue-selectivity score designed to identify genes that are much more strongly expressed in that tissue than in other tissues (S. Xi, personal communication). The third approach ("distinctively expressed", DE) clusters the tissue samples and identifies genes that are most informative for distinguishing each cluster from others (Dey, Hsiao, and Stephens 2017). Note that the HE and SE methods define unique gene sets for a given tissue, whereas the DE method sometimes produce a gene set shared by multiple tissues. All tissue-based gene sets in the present study contain 100 member genes.

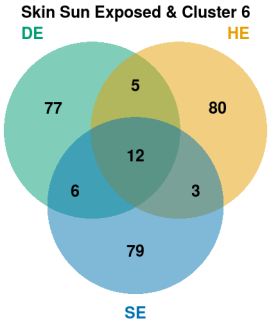


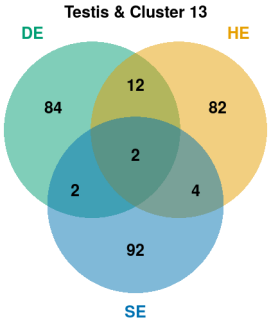

Cortex & Cluster 9
SE
DE
HE
81
94
5
1
13
86
Frontal Cortex & Cluster 2
DE
HE
SE
10
80
82
7
3
1
89
Frontal Cortex & Cluster 9
SE
DE
HE
79
92
7
1
13
86
Hippocampus & Cluster 2
DE
HE
SE
7
85
84
5
3
4
88
Hippocampus & Cluster 9
SE
DE
HE
75
91
5
4
16
80
Hypothalamus & Cluster 2
DE
HE
SE
6
90
85
1
3
8
88
Hypothalamus & Cluster 9
SE
DE
HE
77
91
5
4
14
82
Nucleus Accumbens & Cluster 2
DE
HE
SE
9
87
86
2
2
3
93
Nucleus Accumbens & Cluster 9
SE
DE
HE
83
95
3
2
12
86
Putamen & Cluster 2
DE
HE
SE
7
87
82
3
3
8
86
Putamen & Cluster 9
SE
DE
HE
73
89
8
3
16
81
Esophagus Mucosa & Cluster 15
DE
HE
SE
12
53
55
29
6
4
61
Transformed Fibroblasts & Cluster 8
HE
SE
DE
66
86
7
7
20
73
Heart Atrial Appendage & Cluster 14
DE
HE
SE
14
61
65
14
11
7
68
Heart Left Ventricle & Cluster 14
DE
HE
SE
16
56
57
20
8
7
65
LCL & Cluster 12
SE
DE
HE
86
14
68
18
82
Liver & Cluster 20
DE
HE
SE
32
47
41
14
7
13
66
Lung & Cluster 16
DE
HE
SE
4
84
88
7
5
1
87
Muscle Skeletal & Cluster 7
DE
HE
SE
15
47
52
28
10
5
57
Nerve Tibial & Cluster 11
DE
HE
SE
6
79
89
3
12
2
83

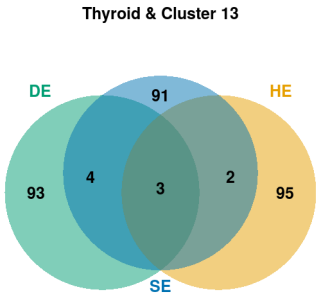

Pancreas & Cluster 18
DE
HE
SE
61
7
61
28
4
4
64

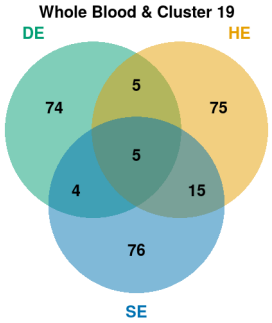

Pituitary & Cluster 13
DE
HE
SE
91
1
90
5
3
4
88


Skin Not Sun Exposed & Cluster 6
HE
SE
DE
83
77
5
12
6
82


Skin Sun Exposed & Cluster 6
DE
HE
SE
77
5
80
12
6
3
79


Testis & Cluster 13
DE
HE
SE
84
12
82
2
2
4
92


Thyroid & Cluster 13
DE
HE
SE
93
91
4
3
2
95

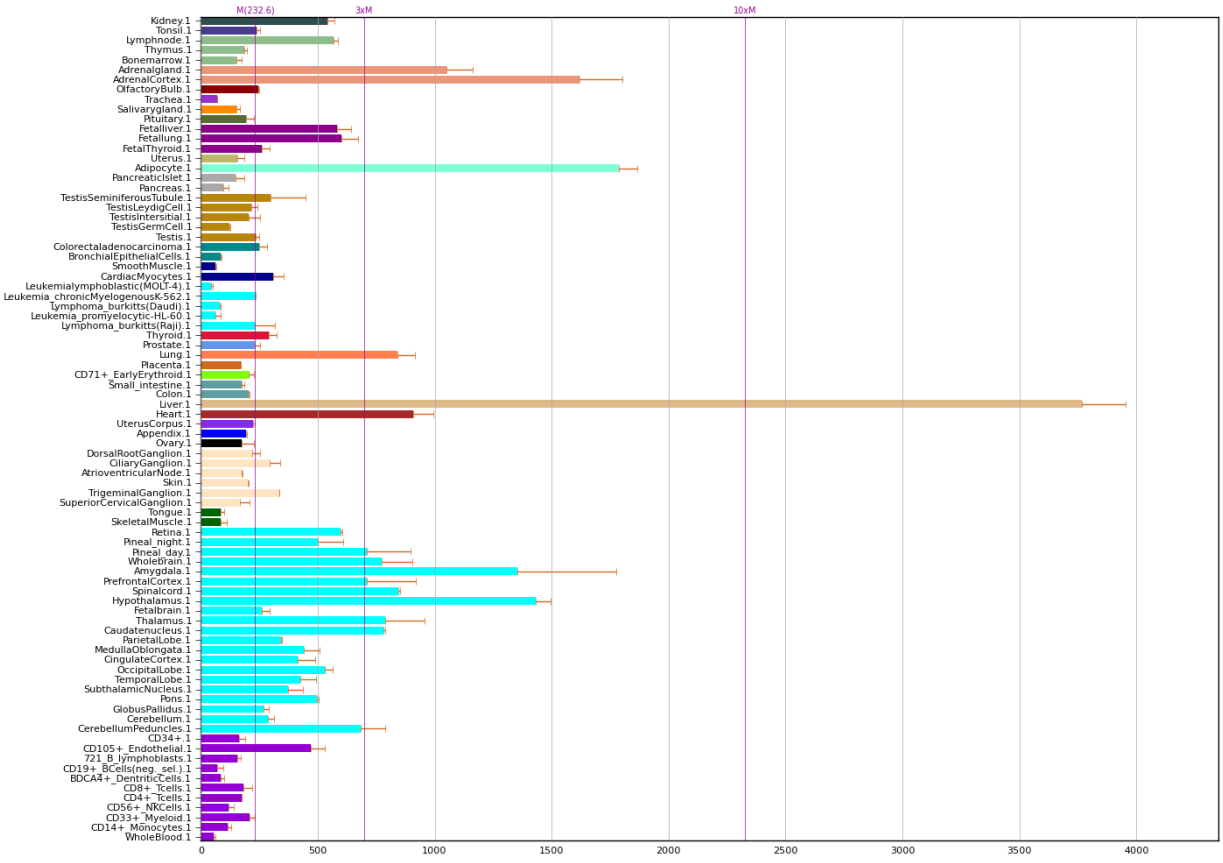

Whole Blood & Cluster 19
DE
HE
SE
74
5
75
5
4
15
76

# Supplementary Figure 29

**Expression pattern of gene *APOE* across human tissues.** The GeneAtlas microarray data (Su et al. 2004) are retrieved from `http://biogps.org/#goto=genereport&id=348`. The RNA-seq data of Illumina bodyMap2 transcriptome project, Duff et al. (2015) and Fagerberg et al. (2014) are retrieved from `https://www.ncbi.nlm.nih.gov/gene/348`. The GTEx RNA-seq data are retrieved from `http://www.gtexportal.org/home/gene/APOE`.

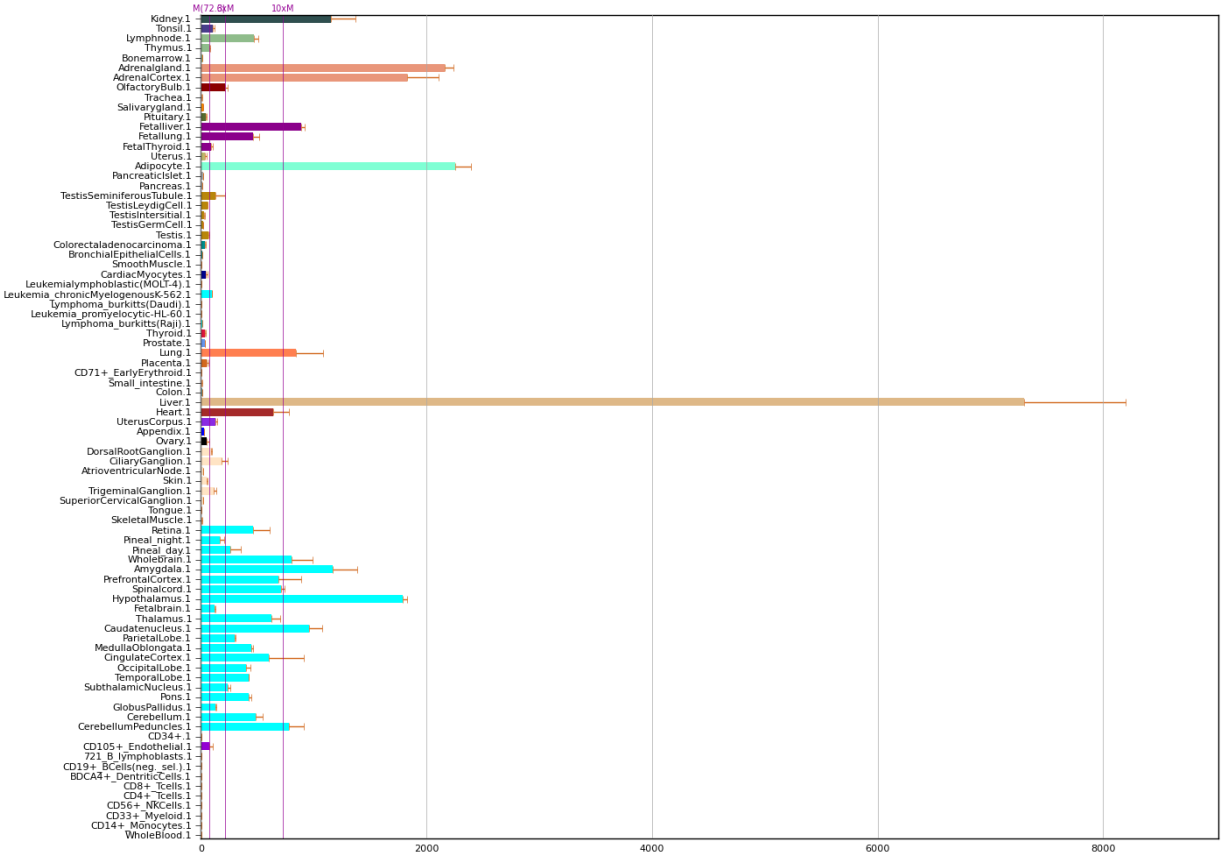
## GeneAtlas microarray probeset `212884_x_at`

Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset `212884_x_at` (Su et al. 2004).



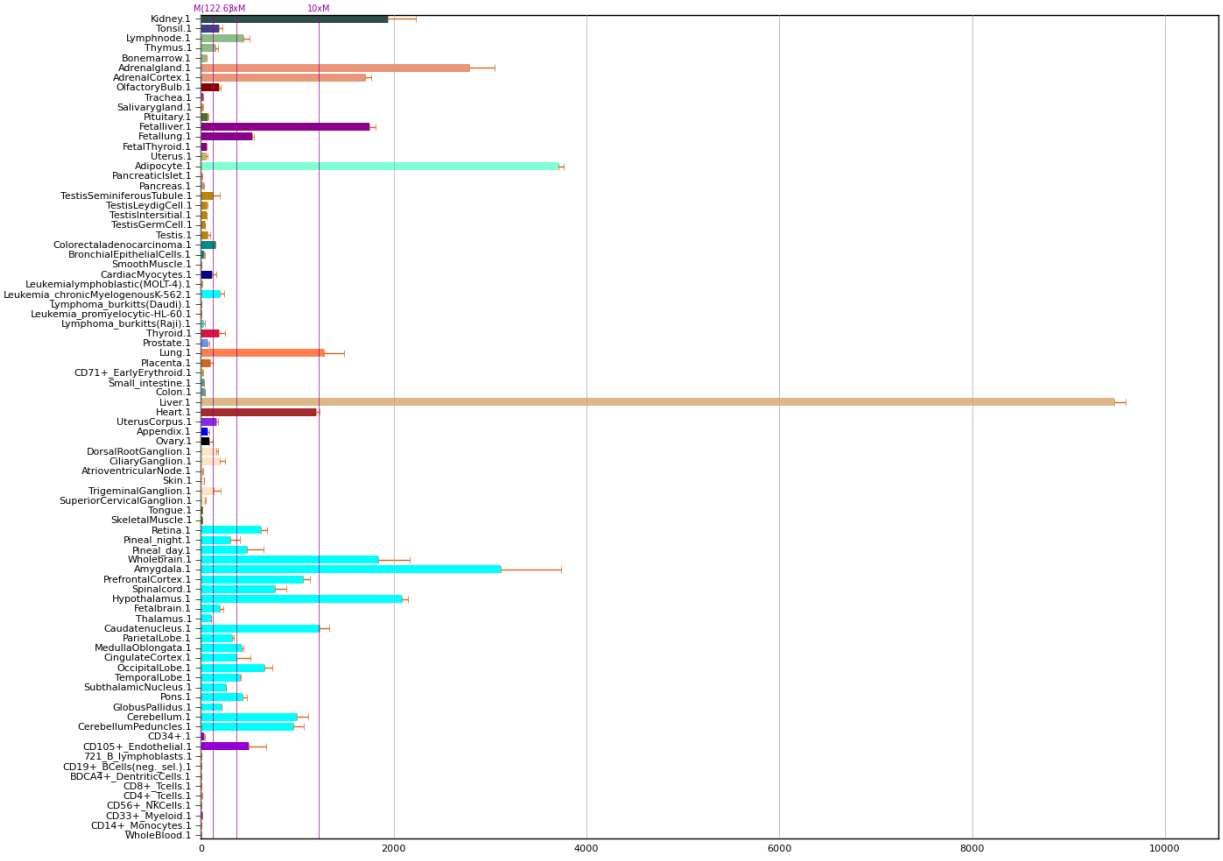
## GeneAtlas microarray probeset 203382_s_at

Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset 203382_s_at (Su et al. 2004).



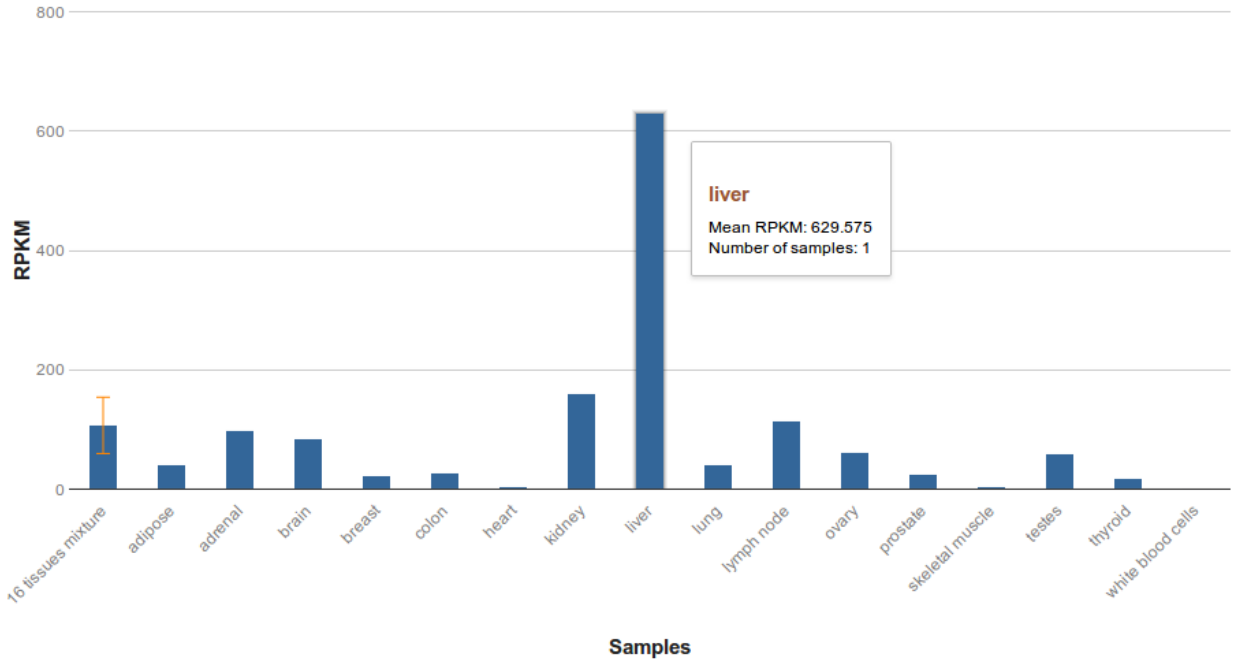
## GeneAtlas microarray probeset `203381_s_at`

Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset `203381_s_at` (Su et al. 2004).



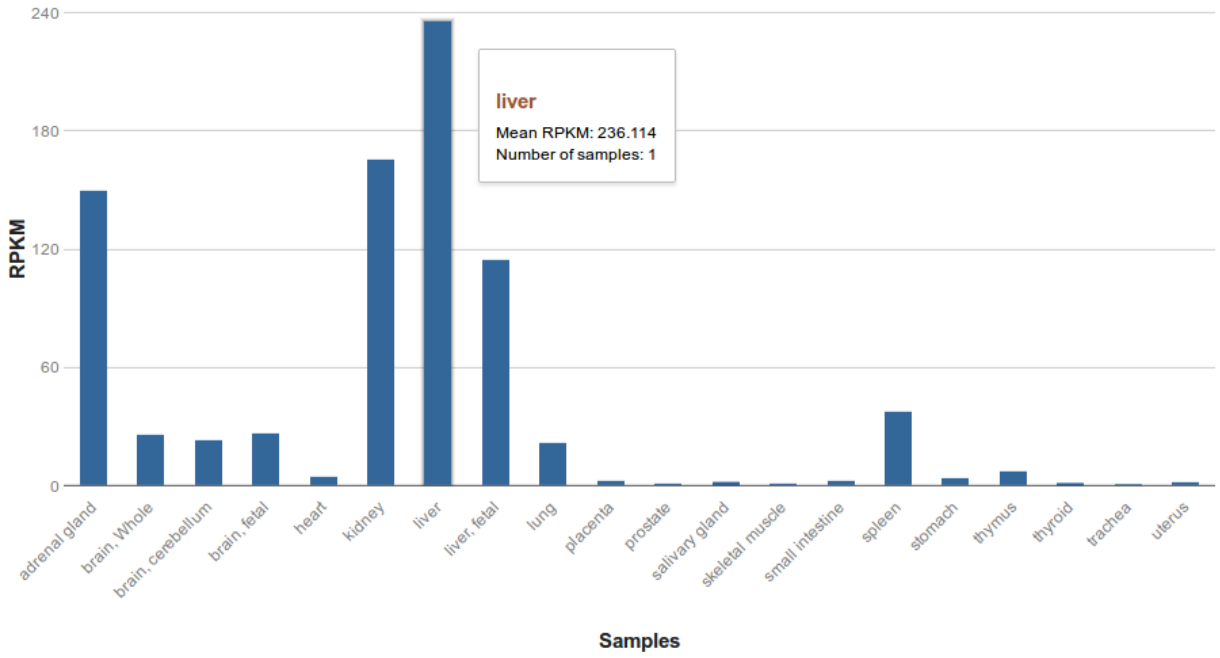
## Illumina bodyMap2 RNA-seq

Expression pattern of *APOE* across 16 human tissues, based on RNA-seq data from Illumina bodyMap2 transcriptome project (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2445/`).



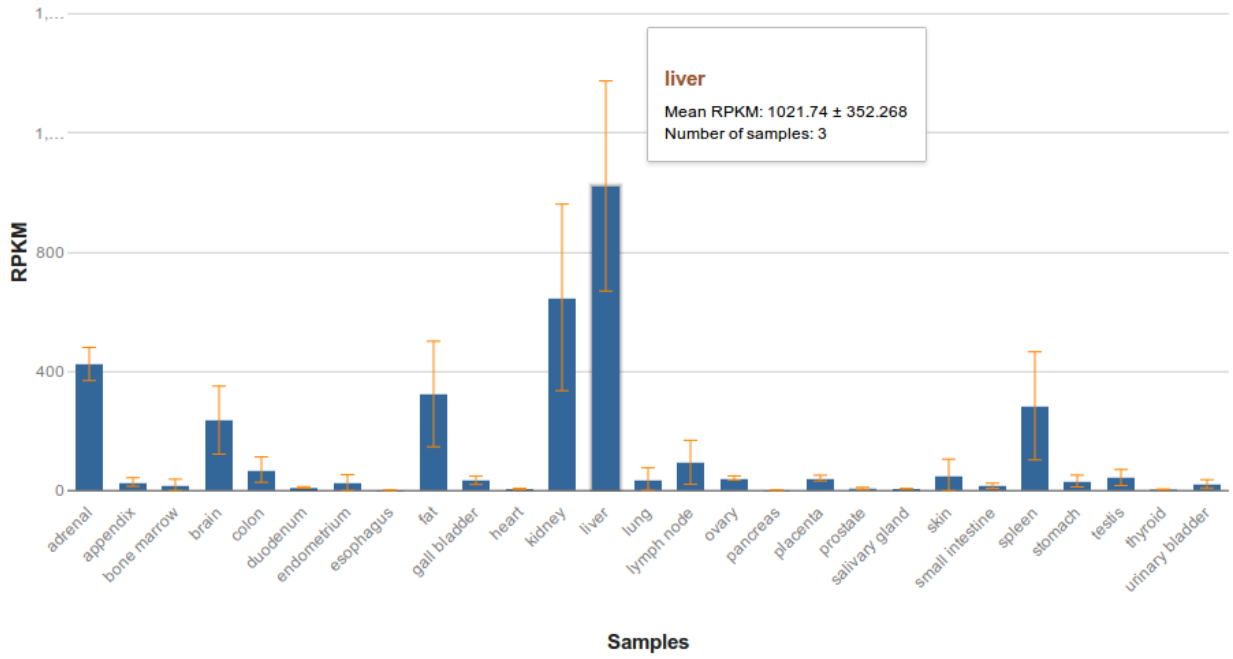

## Duff et al. (2015) RNA-seq

Expression pattern of *APOE* across 20 human tissues based on RNA-seq data (Duff et al. 2015, `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA280600/`).



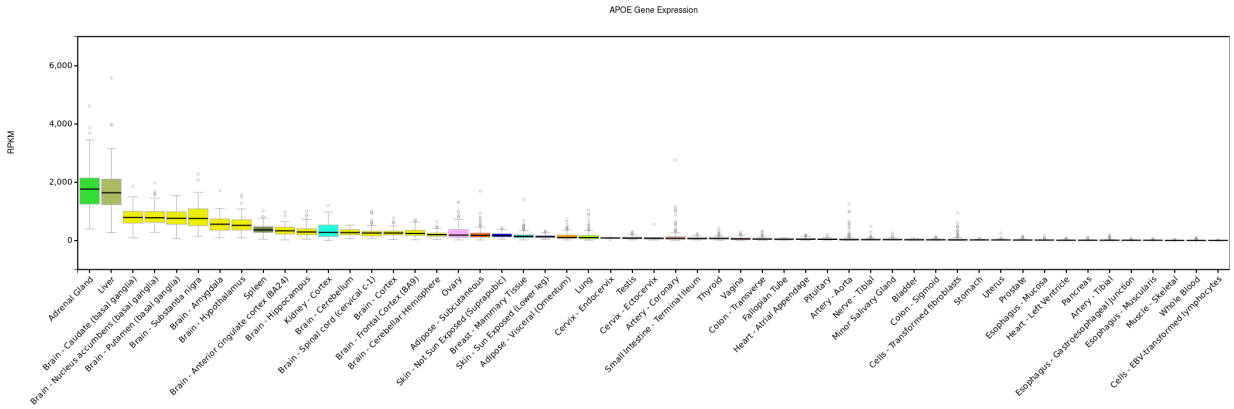
## Fagerberg et al. (2014) RNA-seq

Expression pattern of *APOE* across 27 human tissues, based on RNA-seq data from 95 human individuals (Fagerberg et al. 2014, `https://www.ncbi.nlm.nih.gov/bioproject/PRJEB4337/`).




## GTEx V6p RNA-seq

Expression pattern of *APOE* across 53 human tissues, based on RNA-seq data from GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1).

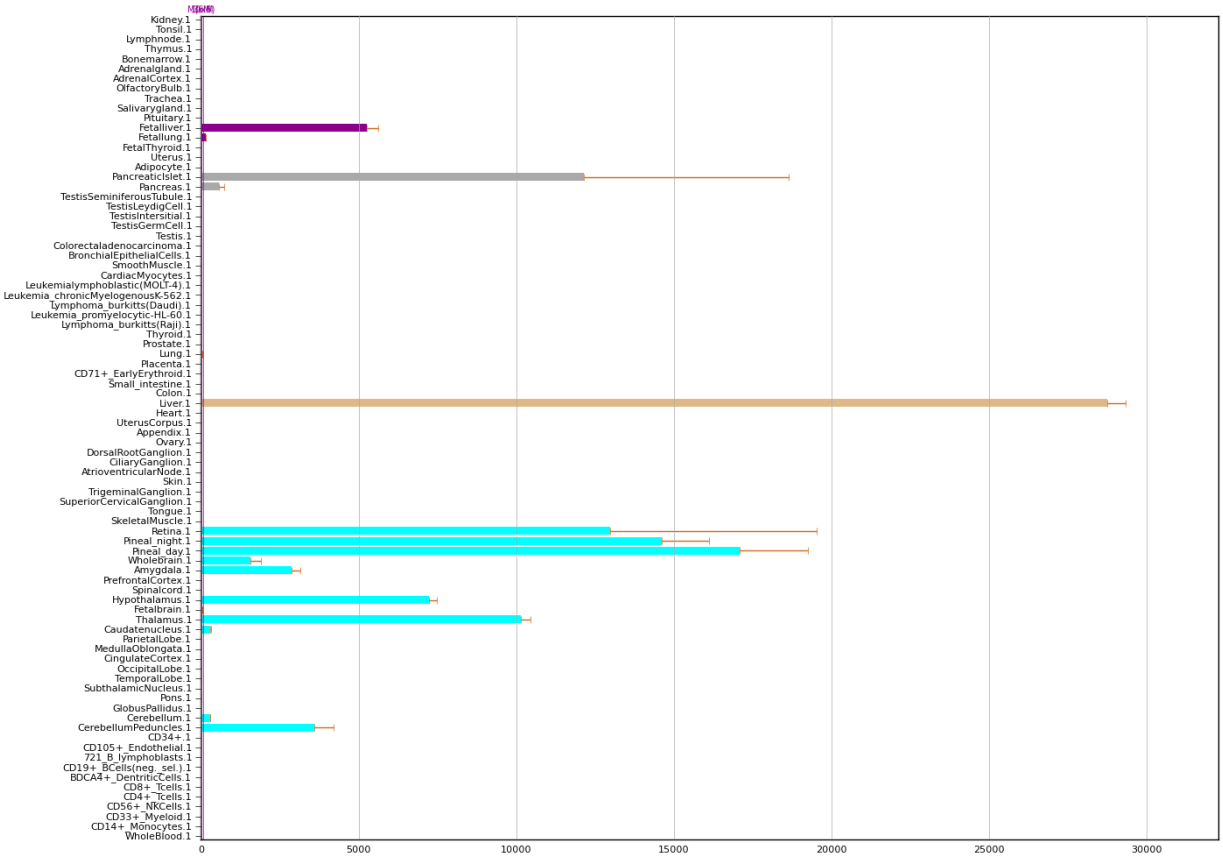


# Supplementary Figure 30

**Expression pattern of gene *TTR* across human tissues.** The GeneAtlas microarray data (Su et al. 2004) are retrieved from `http://biogps.org/#goto=genereport&id=7276`. The RNA-seq data of Illumina bodyMap2 transcriptome project, Duff et al. (2015) and Fagerberg et al. (2014) are retrieved from `https://www.ncbi.nlm.nih.gov/gene/7276`. The GTEx RNA-seq data are retrieved from `http://www.gtexportal.org/home/gene/TTR`.

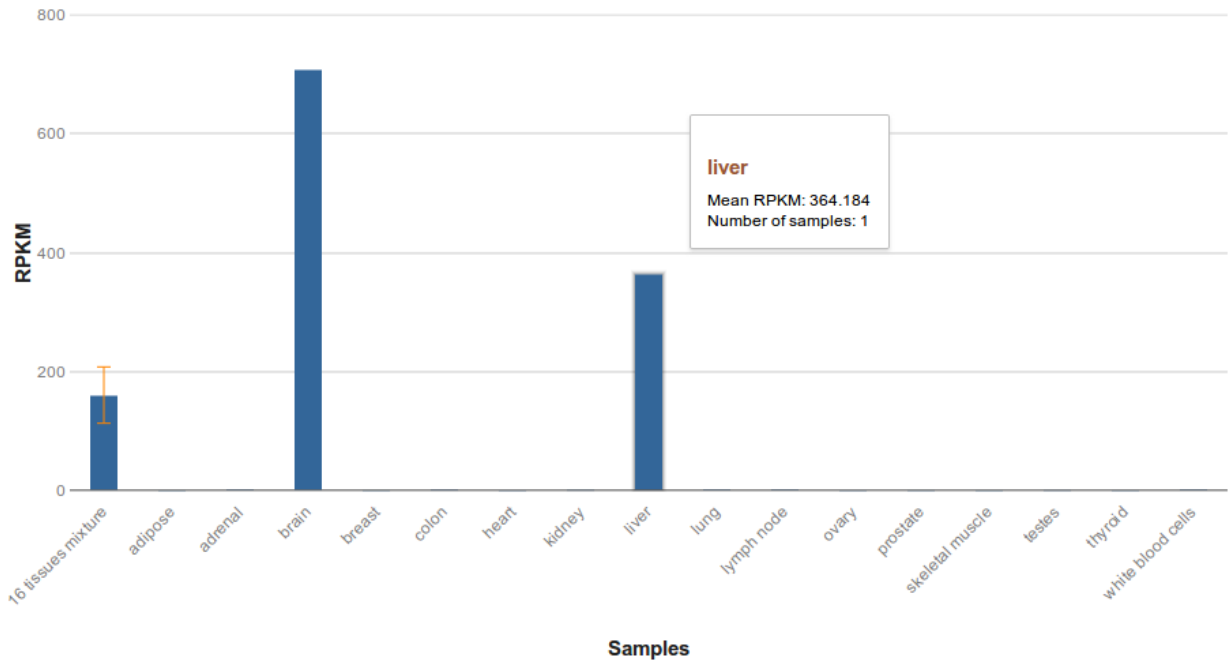
## GeneAtlas microarray probeset `209660_at`

Expression pattern of *TTR* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset `209660_at` (Su et al. 2004).



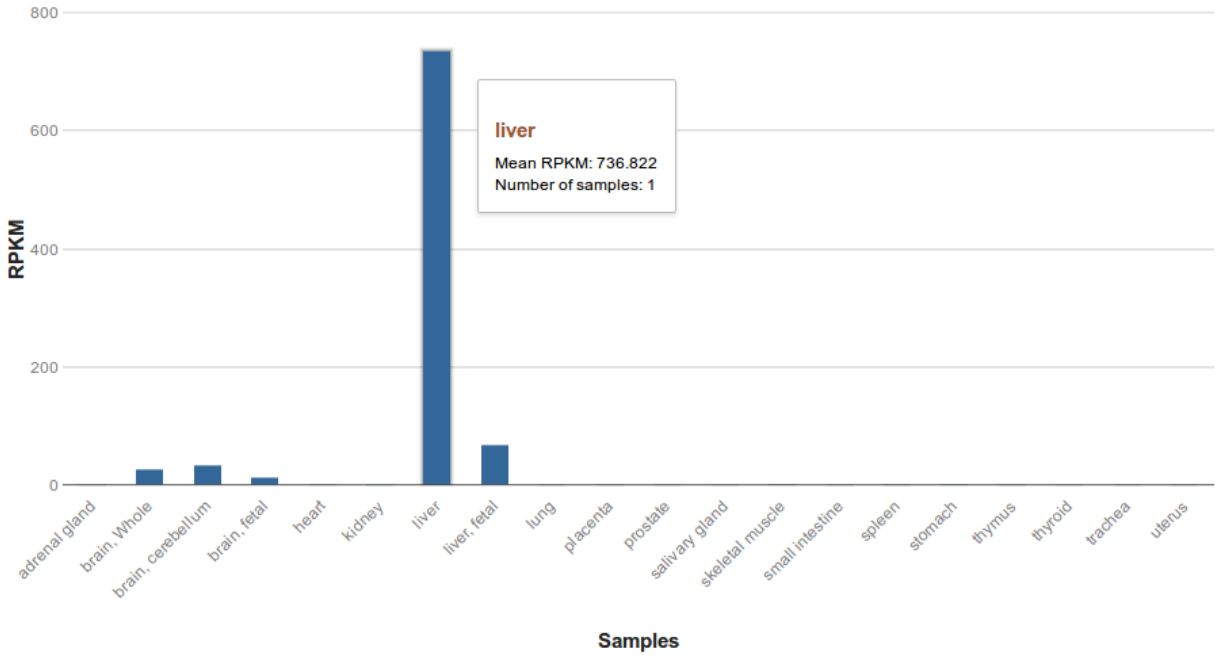
## Illumina bodyMap2 RNA-seq

Expression pattern of *TTR* across 16 human tissues, based on RNA-seq data from Illumina bodyMap2 transcriptome project (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2445/`).



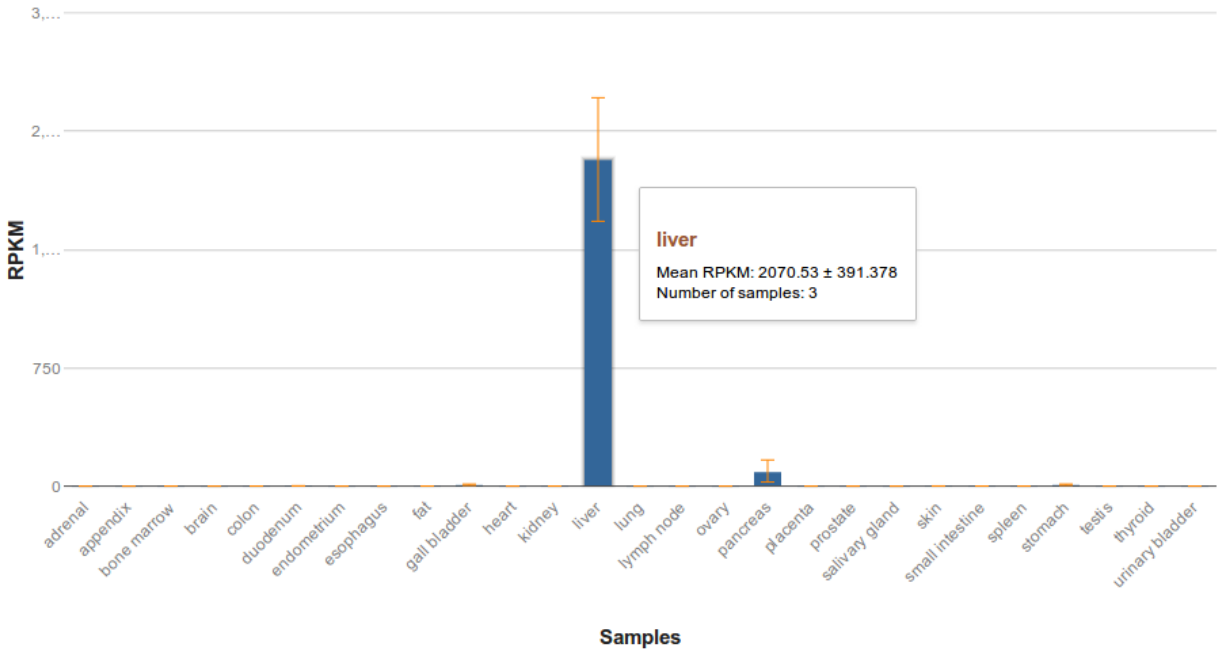

## Duff et al. (2015) RNA-seq

Expression pattern of *TTR* across 20 human tissues based on RNA-seq data (Duff et al. 2015, `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA280600/`).



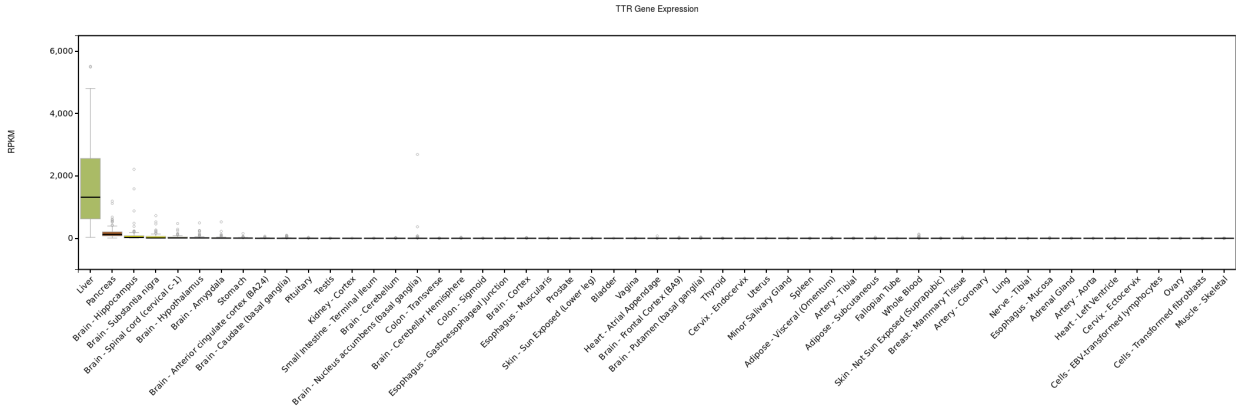
## Fagerberg et al. (2014) RNA-seq

Expression pattern of *TTR* across 27 human tissues, based on RNA-seq data from 95 human individuals (Fagerberg et al. 2014, `https://www.ncbi.nlm.nih.gov/bioproject/PRJEB4337/`).



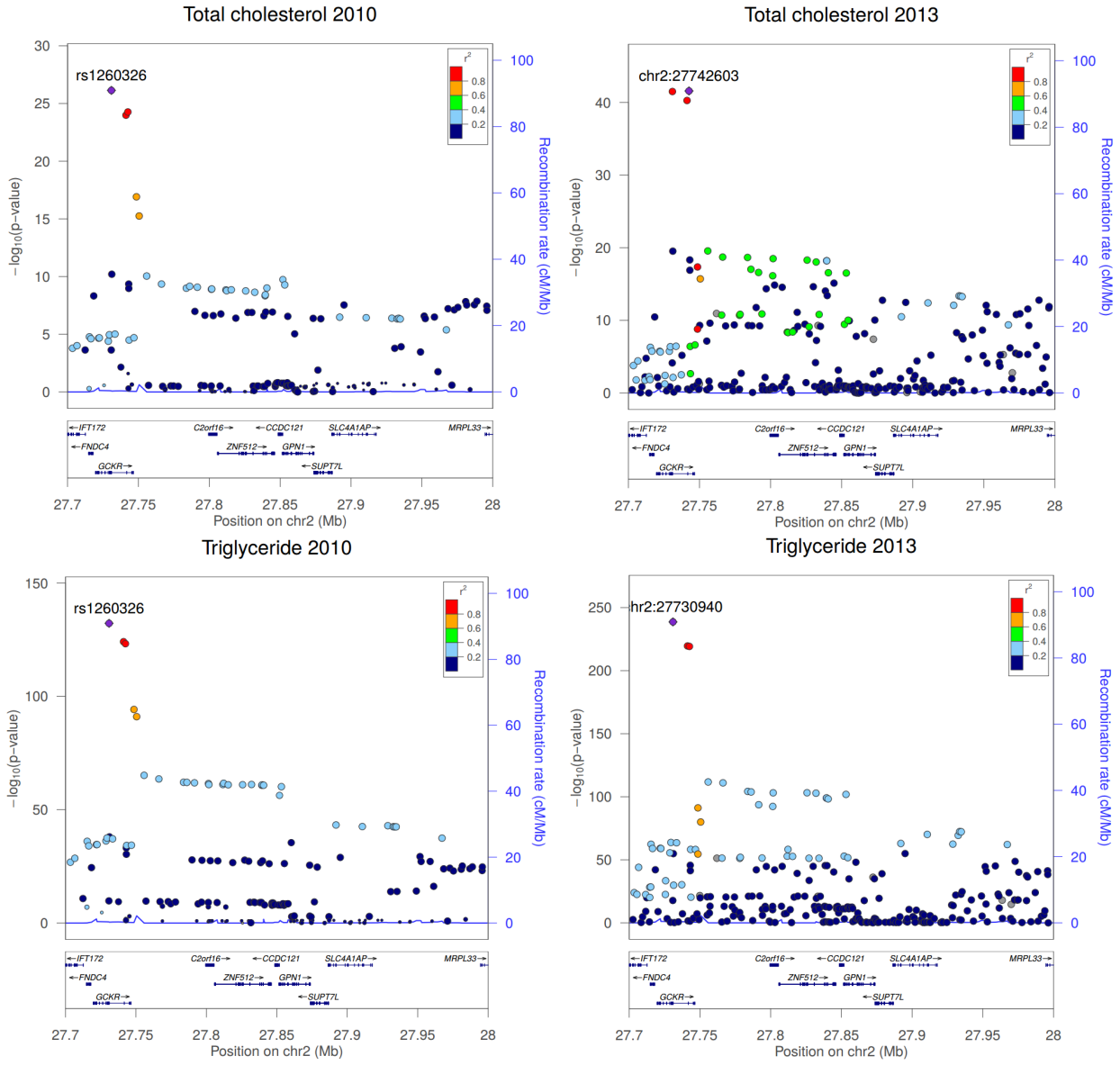

## GTEx V6p RNA-seq

Expression pattern of *TTR* across 53 human tissues, based on RNA-seq data from GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1).



# Supplementary Figure 31

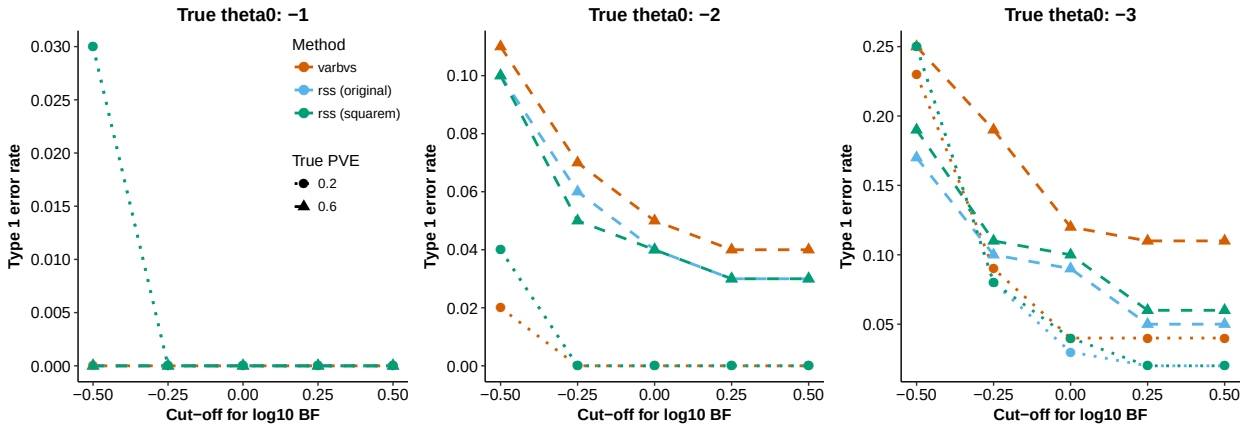
**Regional association plots of genes *C2orf16* and *GCKR* based on single-SNP summary data of total cholesterol and triglycerides levels.** The "2010 data" were reported in Teslovich et al. (2010). The "2013 data" were reported in Global Lipids Genetics Consortium (2013).



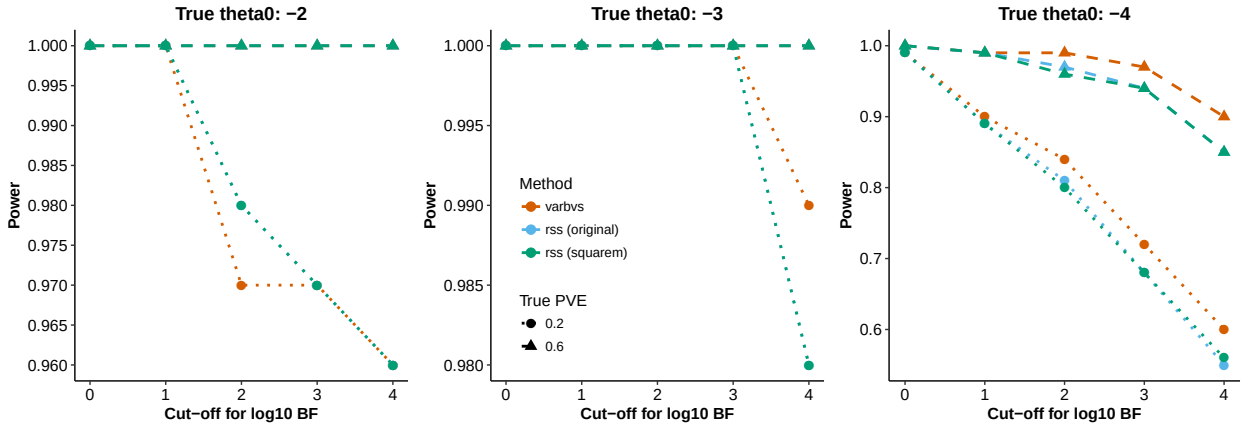
# Supplementary Figure 32

**Comparing analyses of individual-level data (Carbonetto and Stephens 2013) with analyses of summary-level data.** Simulation details of Panel **(a)** are provided in **Supplementary Figure 6**. Simulation details of Panel **(b)** are provided in **Supplementary Figure 1**. For each simulated dataset, `varbvs` analyzes individual-level data and `rss` analyzes summary-level data. Note that `rss (original)` (shown in blue) is included here merely for testing purpose; `rss (squarem)` (shown in green) is the **only** algorithm used in all simulations and data analyses for the present study.

**(a)** Type 1 error rates of `varbvs` (orange), `rss (original)` (blue) and `rss (squarem)` (green) in baseline simulations (see **Supplementary Figure 6**). For each simulated dataset, a type 1 error is made if the enrichment Bayes factor (BF) is greater than the given cutoff for BF.




**(b)** Power of `varbvs` (orange), `rss (original)` (blue) and `rss (squarem)` (green) in enrichment simulations. For each scenario, the power is computed as the fraction of datasets whose enrichment BFs are greater than the given cutoff for BF.



# References


1000 Genomes Project Consortium. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571). Nature Publishing Group: 68–74.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, et al. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3). Nature Research: 291–95.

Carbonetto, Peter, and Matthew Stephens. 2012. “Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies.” *Bayesian Analysis* 7 (1). International Society for Bayesian Analysis: 73–108.

———. 2013. “Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn’s Disease.” *PLoS Genetics* 9 (10). Public Library of Science: e1003770.

Caspi, Ron, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, et al. 2014. “The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases.” *Nucleic Acids Research* 42 (D1): D459–D471.

Cerami, Ethan G., Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Özgön Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. 2011. “Pathway Commons, a Web Resource for Biological Pathway Data.” *Nucleic Acids Research* 39 (suppl 1): D685–D690.

Cortes, Adrian, and Matthew A Brown. 2011. “Promise and Pitfalls of the Immunochip.” *Arthritis Research & Therapy* 13 (1). BioMed Central: 101.

Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, et al. 2014. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 42 (D1): D472–D477.

Day, Felix R, Katherine S Ruth, Deborah J Thompson, Kathryn L Lunetta, Natalia Pervjakova, Daniel I Chasman, Lisette Stolk, et al. 2015. “Large-Scale Genomic Analyses Link Reproductive Aging to Hypothalamic Signaling, Breast Cancer Susceptibility and BRCA1-Mediated DNA Repair.” *Nature Genetics* 47 (11). Nature Publishing Group: 1294–1303.

Den Hoed, Marcel, Mark Eijgelsheim, Tõnu Esko, Bianca JJM Brundel, David S Peal, David M Evans, Ilja M Nolte, et al. 2013. “Identification of Heart Rate-Associated Loci and Their Effects on Cardiac Conduction and Rhythm Disorders.” *Nature Genetics* 45 (6). Nature Publishing Group: 621–31.

Dey, Kushal K, Chiaowen Joyce Hsiao, and Matthew Stephens. 2017. “Visualizing the Structure of RNA-Seq Expression Data Using Grade of Membership Models.” *PLoS Genetics* 13 (3). Public Library of Science: e1006599.

Duff, Michael O, Sara Olson, Xintao Wei, Sandra C Garrett, Ahmad Osman, Mohan Bolisetty, Alex Plocik, Susan E Celniker, and Brenton R Graveley. 2015. “Genome-Wide Identification of Zero Nucleotide Recursive Splicing in Drosophila.” *Nature* 521 (7552). Nature Research: 376–79.

Fagerberg, Linn, Björn M Hallström, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, et al. 2014. “Analysis of the Human Tissue-Specific Expression by Genome-Wide Integration of Transcriptomics and Antibody-Based Proteomics.” *Molecular & Cellular Proteomics* 13 (2). ASBMB: 397–406.

Finucane, Hilary K, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” *Nature Genetics* 47 (11). Nature Publishing Group: 1228–35.

Gao, Xiaoyi, Joshua Starmer, and Eden R Martin. 2008. “A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms.” *Genetic Epidemiology* 32 (4). Wiley Online Library: 361–69.

Geer, Lewis Y., Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. 2010. "The NCBI BioSystems Database." *Nucleic Acids Research* 38 (suppl 1): D492–D496.

Global Lipids Genetics Consortium. 2013. "Discovery and Refinement of Loci Associated with Lipids Levels." *Nature Genetics* 45 (11): 1274–83.

Guan, Yongtao, and Matthew Stephens. 2011. "Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems." *The Annals of Applied Statistics* 5 (3). Institute of Mathematical Statistics: 1780–1815.

Harst, Pim van der, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S Paul, et al. 2012. "Seventy-Five Genetic Loci Influencing the Human Red Blood Cell." *Nature* 492 (7429). Nature Publishing Group: 369–75.

Hsu, Sheng-Da, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, et al. 2014. "miRTarBase Update 2014: An Information Resource for Experimentally Validated miRNA-Target Interactions." *Nucleic Acids Research* 42 (D1): D78–D85.

Köttgen, Anna, Eva Albrecht, Alexander Teumer, Veronique Vitart, Jan Krumsiek, Claudia Hundertmark, Giorgio Pistis, et al. 2013. "Genome-Wide Association Analyses Identify 18 New Loci Associated with Serum Urate Concentrations." *Nature Genetics* 45 (2). Nature Publishing Group: 145–54.

Lambert, Jean-Charles, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, et al. 2013. "Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's Disease." *Nature Genetics* 45 (12). Nature Publishing Group: 1452–8.

Lamparter, David, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. 2016. "Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics." *PLoS Computational Biology* 12 (1). Public Library of Science: e1004714.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. "UpSet: Visualization of Intersecting Sets." *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)* 20 (12): 1983–92.

Li, Miao-Xin, Hong-Sheng Gui, Johnny SH Kwan, and Pak C Sham. 2011. "GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure." *The American Journal of Human Genetics* 88 (3). Elsevier: 283–93.

Liu, Jimmy Z, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, et al. 2010. "A Versatile Gene-Based Test for Genome-Wide Association Studies." *The American Journal of Human Genetics* 87 (1). Elsevier: 139–45.

Liu, Jimmy Z, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, et al. 2015. "Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk Across Populations." *Nature Genetics* 47 (9). Nature Publishing Group: 979–86.

Locke, Adam E, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, et al. 2015. "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology." *Nature* 518 (7538). Nature Publishing Group: 197–206.

Manning, Alisa K, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, et al. 2012. "A Genome-Wide Approach Accounting for Body Mass Index Identifies Genetic Variants Influencing Fasting Glycemic Traits and Insulin Resistance." *Nature Genetics* 44 (6). Nature Publishing Group: 659–69.

Mi, Huaiyu, and Paul Thomas. 2009. "PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools." *Protein Networks and Pathway Analysis*. Springer, 123–40.

Morris, Andrew P, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, et al. 2012. "Large-Scale Association Analysis Provides Insights into

the Genetic Architecture and Pathophysiology of Type 2 Diabetes." *Nature Genetics* 44 (9). Nature Publishing Group: 981–90.

Nikpay, Majid, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, et al. 2015. "A Comprehensive 1000 Genomes–Based Genome-Wide Association Meta-Analysis of Coronary Artery Disease." *Nature Genetics* 47 (10). Nature Publishing Group: 1121–30.

Okada, Yukinori, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, et al. 2014. "Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery." *Nature* 506 (7488). Nature Publishing Group: 376–81.

Okbay, A, BM Baselmans, JE De Neve, P Turley, MG Nivard, MA Fontana, SF Meddens, et al. 2016. "Genetic Variants Associated with Subjective Well-Being, Depressive Symptoms, and Neuroticism Identified Through Genome-Wide Analyses." *Nature Genetics* 48 (6): 624–33.

Pico, Alexander R, Thomas Kelder, Martijn P Van Iersel, Kristina Hanspers, Bruce R Conklin, and others. 2008. "WikiPathways: Pathway Editing for the People." *PLoS Biology* 6 (7): e184.

Rheenen, Wouter van, Aleksey Shatunov, Annelot M Dekker, Russell L McLaughlin, Frank P Diekstra, Sara L Pulit, Rick AA van der Spek, et al. 2016. "Genome-Wide Association Analyses Identify New Risk Variants and the Genetic Architecture of Amyotrophic Lateral Sclerosis." *Nature Genetics* 48 (9). Nature Publishing Group: 1043–8.

Romero, Pedro, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. 2004. "Computational Prediction of Human Metabolic Pathways from the Complete Human Genome." *Genome Biology* 6 (1). BioMed Central Ltd: R2.

Sachs, Michael C. 2017. "plotROC: A Tool for Plotting Roc Curves." *Journal of Statistical Software, Code Snippets* 79 (2): 1–19. https://doi.org/10.18637/jss.v079.c02.

Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009. "PID: The Pathway Interaction Database." *Nucleic Acids Research* 37 (suppl 1): D674–D679.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." *Nature* 511 (7510). Nature Publishing Group: 421–27.

Schmich, Fabian. 2015. *GespeR: Gene-Specific Phenotype Estimator*. http://www.cbg.ethz.ch/software/gespeR.

Shungin, Dmitry, Thomas W Winkler, Damien C Croteau-Chonka, Teresa Ferreira, Adam E Locke, Reedik Mägi, Rona J Strawbridge, et al. 2015. "New Genetic Loci Link Adipose and Insulin Biology to Body Fat Distribution." *Nature* 518 (7538). Nature Publishing Group: 187–96.

Stephens, Matthew. 2017. "False Discovery Rates: A New Deal." *Biostatistics* 18 (2): 275–94.

Su, Andrew I, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, et al. 2004. "A Gene Atlas of the Mouse and Human Protein-Encoding Transcriptomes." *Proceedings of the National Academy of Sciences of the United States of America* 101 (16). National Acad Sciences: 6062–7.

Teslovich, Tanya M, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, et al. 2010. "Biological, Clinical and Population Relevance of 95 Loci for Blood Lipids." *Nature* 466 (7307). Nature Publishing Group: 707–13.

Turner, Stephen D. 2014. "qqman: An R Package for Visualizing GWAS Results Using QQ and Manhattan Plots." *bioRxiv*. Cold Spring Harbor Labs Journals, 005165.

Voight, Benjamin F, Hyun Min Kang, Jun Ding, Cameron D Palmer, Carlo Sidore, Peter S Chines, Noël P Burtt, et al. 2012. "The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits." *PLoS Genetics* 8 (8). Public Library of Science: e1002793.

Wang, Minghui, Jianfei Huang, Yiyuan Liu, Li Ma, James B Potash, and Shizhong Han. 2017. "COMBAT: A Combined Association Test for Genes Using Summary Statistics." *Genetics* 207 (3). Genetics Soc America: 883–91.

Webber, William, Alistair Moffat, and Justin Zobel. 2010. “A Similarity Measure for Indefinite Rankings.” *ACM Transactions on Information Systems (TOIS)* 28 (4). ACM: 20.

Wellcome Trust Case Control Consortium. 2007. “Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls.” *Nature* 447: 661–78.

Wen, Xiaoquan, and Matthew Stephens. 2010. “Using Linear Predictors to Impute Allele Frequencies from Summary or Pooled Genotype Data.” *The Annals of Applied Statistics* 4 (3). Institute of Mathematical Statistics: 1158–82.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, et al. 2014. “Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height.” *Nature Genetics* 46 (11). Nature Publishing Group: 1173–86.

Wrzodek, Clemens, Finja Büchel, Manuel Ruff, Andreas Dräger, and Andreas Zell. 2013. “Precise Generation of Systems Biology Models from KEGG Pathways.” *BMC Systems Biology* 7 (1). BioMed Central Ltd: 15.

Zhou, Xiang, Peter Carbonetto, and Matthew Stephens. 2013. “Polygenic Modeling with Bayesian Sparse Linear Mixed Models.” *PLoS Genetics* 9 (2). Public Library of Science: e1003264.

Zhu, Xiang, and Matthew Stephens. 2017. “Bayesian Large-Scale Multiple Regression with Summary Statistics from Genome-Wide Association Studies.” *Annals of Applied Statistics* 11 (3): 1561–92. https://doi.org/10.1214/17-AOAS1046.