Benchmarking Association Analyses of Continuous Exposures with RNA-seq in Observational Studies: Supplementary Information

Tamar Sofer, Nuzulul Kurniansyah, François Aguet, Kristin Ardlie, Peter Durda, Deborah A. Nickerson, Joshua D. Smith, Yongmei Liu, Sina A. Gharib, Susan Redline, Stephen S. Rich, Jerome I. Rotter, Kent D. Taylor

RNA sequencing in MESA

RNA-seq was generated from peripheral blood mononuclear cells (PBMCs) of MESA participants obtained in the fifth clinic visit. PBMC samples were sequenced at the Broad Institute (n=498), and at the North West Genomics Center (NWGC; n=468). Both centers used harmonized protocols. RNA samples quality was assessed using RNA Integrity Number (RIN, Agilent Bioanalyzer) prior to shipment to sequencing centers. QC was re-performed at sequencing centers by RIN analysis at the NWGC and by RNA Quality Score analysis (RQS, Caliper) at the

Broad Institute. A minimum of 250ng RNA sample was required as input for library construction, performed using the Illumina TruSeq™ Stranded mRNA Sample Preparation Kit. RNA was sequenced as 2x101bp paired-end reads on the Illumina HiSeq 4000 according to the manufacturer's protocols. Target coverage was of ≥40M reads. Comprehensive information about the RNA-seq pipeline used for TOPMed can be found in https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md under MESA RNA-seq pilot commit 725a2bc. Here we used transcript-level expected counts quantified using RSEM v1.3.0 [1].

Characteristics of MESA participants

Table S1: Characteristics of participants from MESA with RNA-seq and sleep study, by sex.

Characteristic	Females	Males
n	250	212
Race (%):		
White	108 (43.2)	92 (43.4)
Black	57 (22.8)	41 (19.3)
Hispanic	85 (34.0)	79 (37.3)
Age (mean(SD))	68.37 (9.44)	67.56 (9.30)
BMI (mean (SD))	30.73 (5.90)	28.79 (4.57)
AvgO2 (mean (SD))	94.24 (1.78)	94.00 (2.05)
AHI (mean (SD))	15.41 (15.22)	22.36 (20.10)
MinO2 (mean (SD))	82.87 (8.39)	83.04 (7.56)

Study of filtering based on distributional characteristics of transcripts

We studied multiple approaches to filter transcripts and reduce the number of transcripts in the association analysis. The goal was to identify distributional characteristics of the transcripts that result in low power. Based on these, one may be able to retain a smaller number of genes in the analysis and increase power by reducing the multiple testing burden. We note that filtering also affects type 1 error, because the overall distribution of p-values under the null changes depending on the set of transcripts used.

1. Considered filters

The considered filters (and characteristics) were, as conditions applied on transcript j:

F1. Expression sum filter: The sum of expression counts across all people need to satisfy $\sum_{i=1}^{n} t_{ij} > C_1$.

F2. Median filter: The median expression count across all people need to satisfy $median(t_{1j}, \ldots, t_{nj}) > C_2$.

F3. Max to Median Ratio filter: The maximum to median expression count across all people need to satisfy: $\left[\frac{median(t_{ij}, i=1,\ldots n)}{\max(t_{ij}, i=1,\ldots n)}\right] > C_3$.

F4. Maximum filter: The maximum expression count across all people need to satisfy: $max(t_{1j}, \ldots, t_{nj}) > C_4$.

F5. Range filter: The range of read counts for a transcript needs to satisfy: $(t_{max,j} - t_{min,j}) > C_5$, where $t_{max,j}$ and $t_{min,j}$ are the highest and lowest read counts observed for this transcript across all people in the sample.

F6. Proportion zero count filter: The proportion of individuals with zero expression count needs to satisfy $\frac{1}{n}\sum_{i=1}^{n} 1(t_{ij} = 0) \leq C_6$.

F7. Coefficient of variation filter: The standard deviation divided by the mean expression value $\frac{sd(t_{1j},\ldots,t_{nj})}{mean(t_1,\ldots,t_{nj})} = C_7$ need to be within a specified range.

2. Simulations of transcript characteristics and power

For each transcript, we used the exposure phenotype AvgO2 to generate residuals, followed by the residual permutation scheme with correlation parameter $\rho = 0.3$ to generate a simulated exposure variable that is associated with the transcript. We tested the phenotype-transcript association using linear regression after log applied on SubHalfMin transformation. We repeated the simulations 100,000 times, and computed power for each transcript based on the proportion of simulations in which the raw p-value from testing the association of the transcript with the simulated phenotype was smaller than 10^{-6}. Based on these simulations, only one clear pattern emerges: a high number of zero expression values across samples for a given transcript leads to low power. Based on this, we proceeded requiring that at least 50% of samples have non-zero values for a transcript for it to be included in the analysis, or equivalently, median expression values higher than 0. Figure S1 demonstrates the loss in power for analyzing a continuous phenotype when the median expression value is 0.

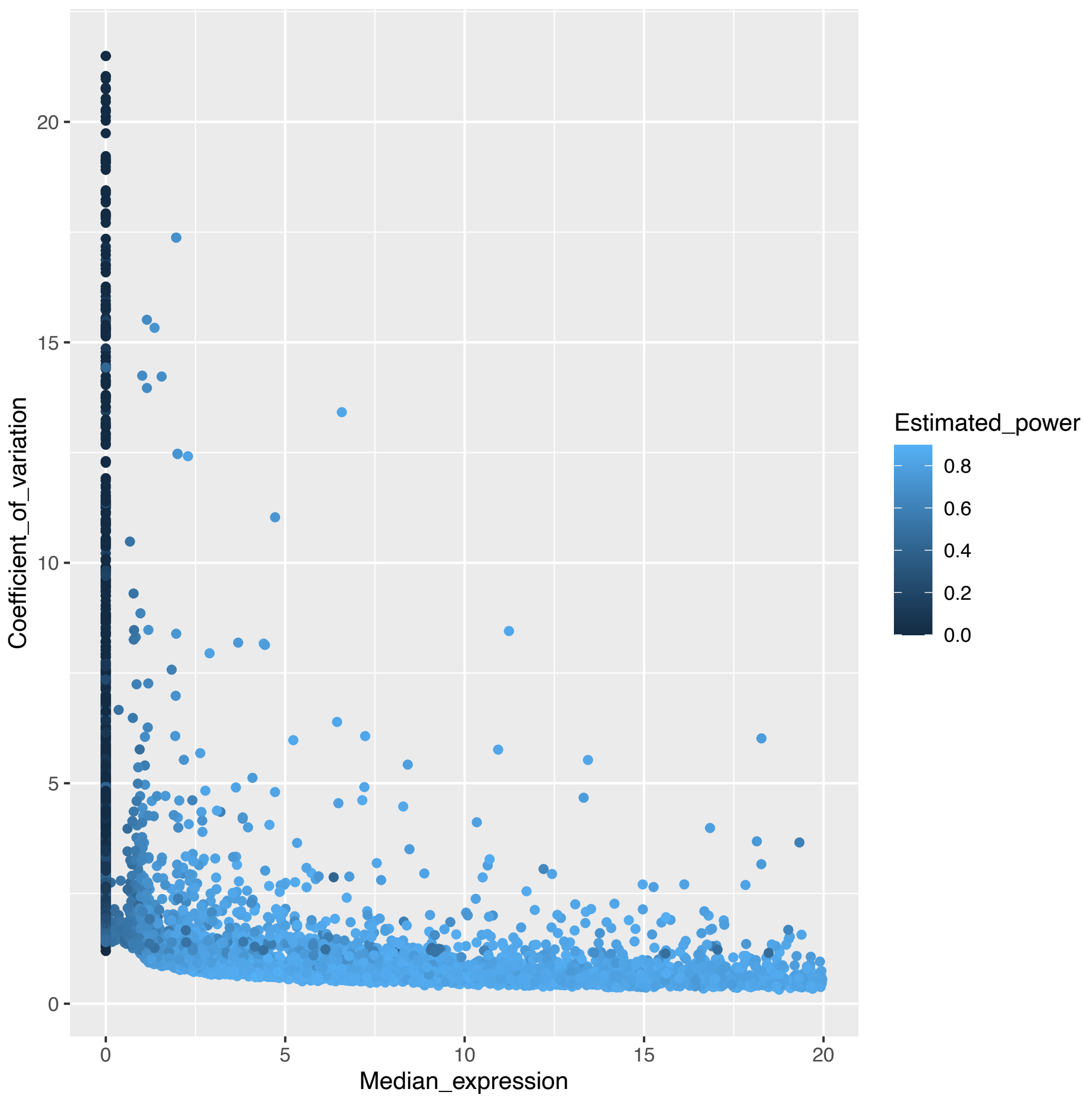

Figure S1: Estimated power to detect associations using linear regression analysis for a given transcript as a function of median expression value (after normalization) and coefficient of variation. Median expression values were capped at 20 in the figure. Power was estimated in 1,000 simulations using the residual permutation approach with $\rho = 0.3$, and 100,000 residual permutations with no associations were used to compute permutation p-values for each transcript.

Permutation and empirical P-values

A standard, well-known approach to computing p-values when the null distribution cannot be specified is permutation. We refer to these p-values as "permutation p-values" and here we provide details in order to contrast them with empirical p-values. We note here that in the statistical literature permutation p-values are sometimes called empirical p-values, but here we use the term empirical p-value as often applied in the gene expression analysis literature.

1. Permutation p-values

The idea is that for a given transcript j, the transcript expression values, or the exposure (or the residuals) are permuted B times across individuals, and a p-value is computed from each study permutation $p_b^j, b = 1, \dots, B$. Then, the permutation p-value is

$$p_{perm}^j = \frac{1}{B}\sum_{b=1}^{B} 1\left(p_b^j < p_{true}^j\right),$$

assuming that each $p_b^j, b = 1, \dots, B$ is drown from the null distribution of p_{true}^j. The challenge with permutation p-values for transcriptomics is the computation burden. The number of permutations B has to be very large, because the permutation p-values p_b^j are based on evaluations on the specific transcript permutation. In other words, the entire transcriptomics analysis has to be performed B times, where B is usually at least 100,000, leading to a huge computational burden.

2. Empirical p-values

We study the use of *empirical p-values,* computed while capitalizing over all transcripts for computation of the empirical p-value of every specific transcript. We use a non-parametric, quantile-based approach [2], and the method implemented in the qvalue R package [3], both using the residual permutation approach to obtain the empirical distribution of p-values under the null. The specific implementation of the quantile empirical p-values is as follows: Consider the distribution of permutation p-values for a transcript j, defined by $\{ p_1^j, \dots, p_B^j\}$, as $F_j(x) = \frac{1}{B}\sum_{b=1}^{B} 1\left(p_b^j < x\right)$. Considering all transcripts $j = 1, \dots, k$ passing the filtering criteria and participating in the analysis. Suppose we permute each one of them B times, and estimate an empirical p-value distribution as

$$F_{emp}(x) = \frac{1}{B \times k}\sum_{j=1}^{k}\sum_{b=1}^{B} 1(p_b^j < x).$$

Then, if $F_j(x) \approx F_{emp}(x)$, we can use F_{emp} rather than F_j to compute empirical p-values rather than permutation p-values, with a substantially smaller B, meaning with many less transcriptome-wide association analyses.

3. Comparison of permutation p-values and empirical p-values

We compared the quantile empirical p-values to permutation p-values using the three sleep exposures. We removed all transcripts with maximum counts of 10 and more than 50% zero counts in the sampled, and applied median normalization. We then used the residual permutation approach to generate data for simulation under the null of no phenotype-transcript association, and tested for differential using linear regression after log transformation of subHalfMin approach for handling zero counts. We performed permutation analysis $B = 100{,}000$ times (B transcriptome-wide residual permutation analyses), and compared the resulting permutation p-values to the empirical p-values obtained using $B = 100$ transcriptome-wide residual permutation analyses. Figure 2 provide the comparison, demonstrating that the two p-values are very similar, therefore, it is appropriate to use empirical p-values which are computationally much faster to compute.

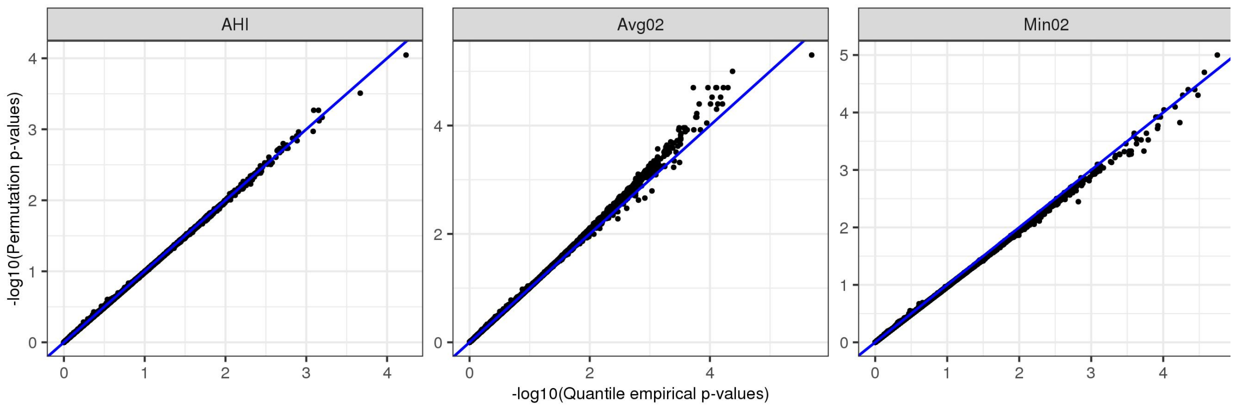

Figure S2: Quantile empirical p-values (computed using $B = 100$ residual permutations) versus standard permutation p-values (computed using $B = 100{,}000$ residual permutations) in simulations.

Results from simulation study 1: assessment of false positive detections

Figure S3: Average number of falsely-detected transcript association with the residual-permuted **AHI** phenotype, estimated over 100 permutations. We compared approaches based on linear regression, DESeq2, limma, and EdgeR packages; raw p-values, quantile-empirical p-values, and Storey's empirical p-values; and associations declared as significant according the arbitrary threshold of p-value < 10^{-5}, FDR adjustment using the Benjamini-Hochberg (BH) procedure and using the local FDR procedure implemented in the qvalue R package, and FWER adjustment using Holm procedure.

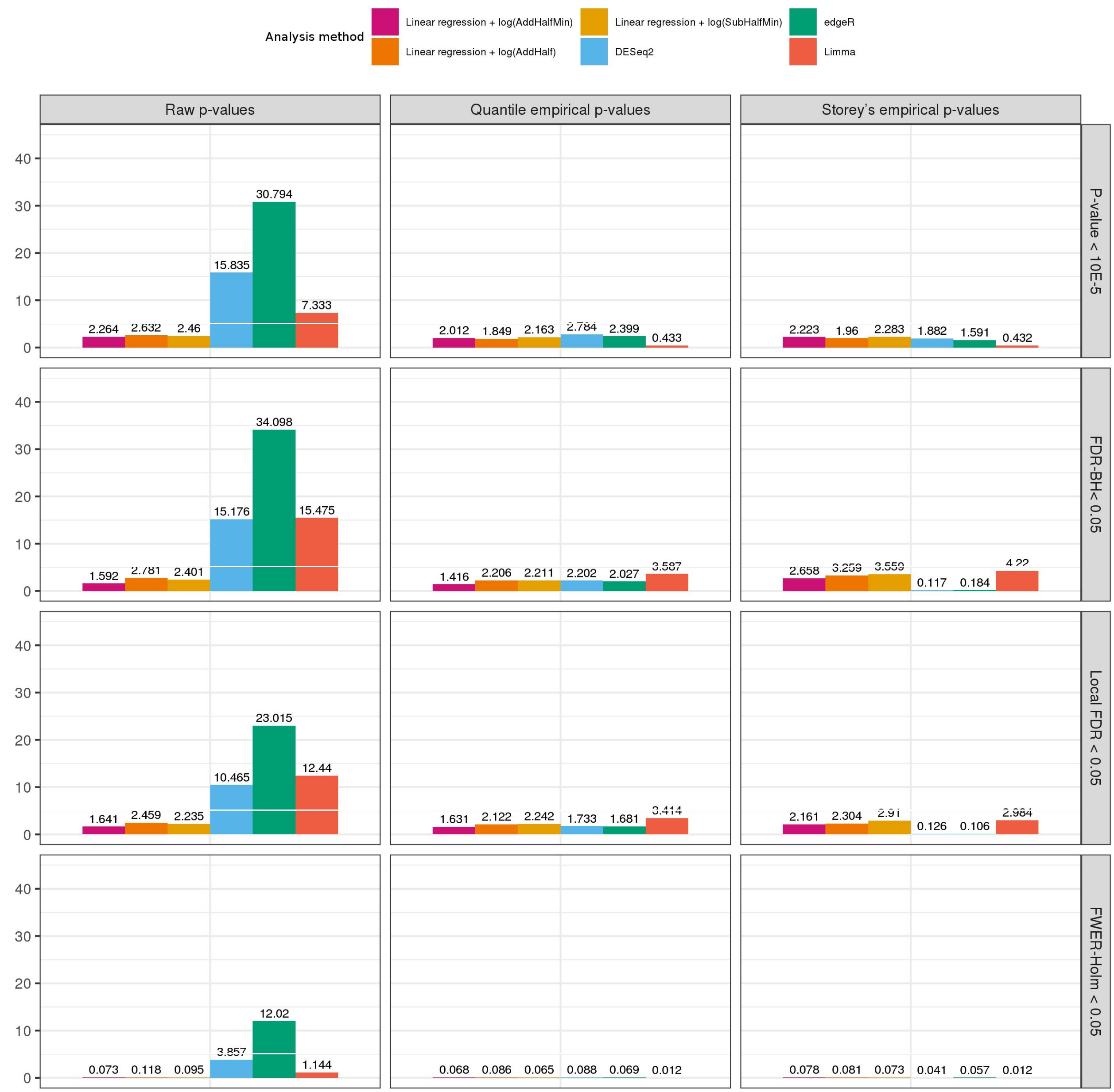

Figure S4: Average number of falsely-detected transcript association with the residual-permuted **AvgO2** phenotype, estimated over 100 permutation. We compared approaches based on linear regression, DESeq2, limma, and EdgeR packages; raw p-values, quantile-empirical p-values, and Storey's empirical p-values; and associations declared as significant according the arbitrary threshold of p-value < 10^{-5}, FDR adjustment using the Benjamini-Hochberg (BH) procedure and using the local FDR procedure implemented in the qvalue R package, and FWER adjustment using Holm procedure.

Figure S5: Average number of falsely-detected transcript association with the residual-permuted **MinO2** phenotype, estimated over 100 permutations. We compared approaches based on linear regression, DESeq2, limma, and EdgeR packages; raw p-values, quantile-empirical p-values, and Storey's empirical p-values; and associations declared as significant according the arbitrary threshold of p-value < 10^{-5}, FDR adjustment using the Benjamini-Hochberg (BH) procedure and using the local FDR procedure implemented in the qvalue R package, and FWER adjustment using Holm procedure.

Comparison of the effect of normalization methods on false positive detections

We also performed simulation study 1 using, instead of Median normalization, two commonly used normalizations: the TMM normalization implemented in the edgeR package (Figure S6), and the size factor normalization implemented in the DESeq2 package (Figure S7). One can see that the results are very similar to those from Figure S4.

Figure S6: Average number of falsely-detected transcript association with the residual-permuted **AvgO2** phenotype, estimated over 100 permutations. The analysis was applied after **TMM normalization**. We compared approaches based on linear regression, DESeq2, limma, and EdgeR packages; raw p-values, quantile-empirical p-values, and Storey's empirical p-values; and associations declared as significant according the arbitrary threshold of p-value < 10^{-5}, FDR adjustment using the Benjamini-Hochberg (BH) procedure and using the local FDR procedure implemented in the qvalue R package, and FWER adjustment using Holm procedure.

Figure S7: Average number of falsely-detected transcript association with the residual-permuted **AvgO2** phenotype, estimated over 100 permutations. The analysis was applied after **size factor normalization**. We compared approaches based on linear regression, DESeq2, limma, and EdgeR packages; raw p-values, quantile-empirical p-values, and Storey's empirical p-values; and associations declared as significant according the arbitrary threshold of p-value < 10^{-5}, FDR adjustment using the Benjamini-Hochberg (BH) procedure and using the local FDR procedure implemented in the qvalue R package, and FWER adjustment using Holm procedure.

Study of false positive detections in lower sample sizes

We performed additional simulations under the null hypothesis of no association between the exposure and the RNA-seq, using the single selected approach, and with lower sample sizes, to assess whether the number of false positive detections may increase as the sample size becomes lower. We used MinO2 as the exposure phenotype, because it yields the highest number of false positive and therefore serves as a more extreme case study. We sampled N=150, 100, 50, and 30 individuals from the dataset used in the primary simulations. We applied the linear regression approach using log(SubHalfMin) transformation, with the same adjustments and the same residual permutation approach used in the main simulations. Figure S8 below visualizes the average number of false positive detections across 1,000 simulations applied on each of the sample sizes when applying BH FDR correction on the quantile empirical p-values followed by (adjusted) p-value threshold of 0.05.

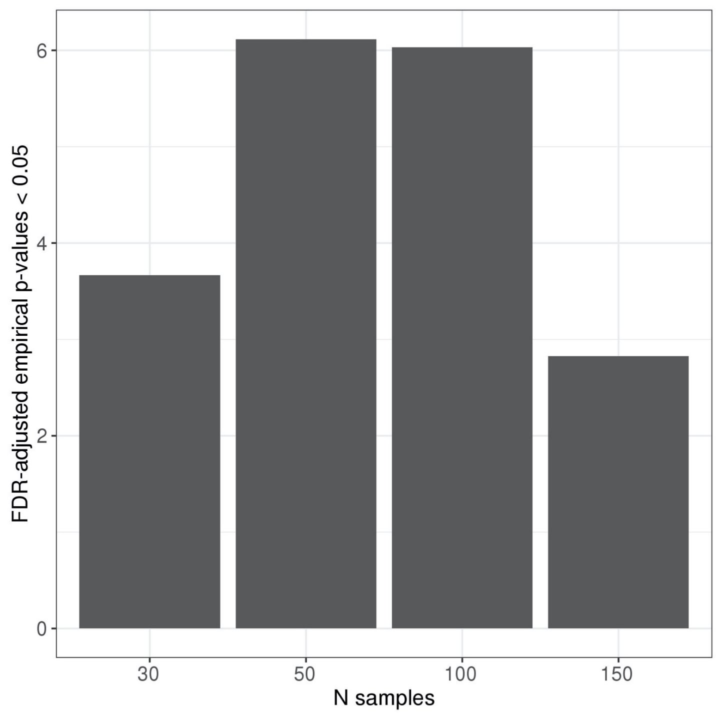

Figure S8: Average number of false positives detected in simulation studies varying the sample size. The exposure phenotype used was MinO2. Linear regression was applied after log(SubHalfMin) transformation.

Results from simulations study 3: power for detecting an association with a simulated transcript association in a transcriptome-wide association analysis

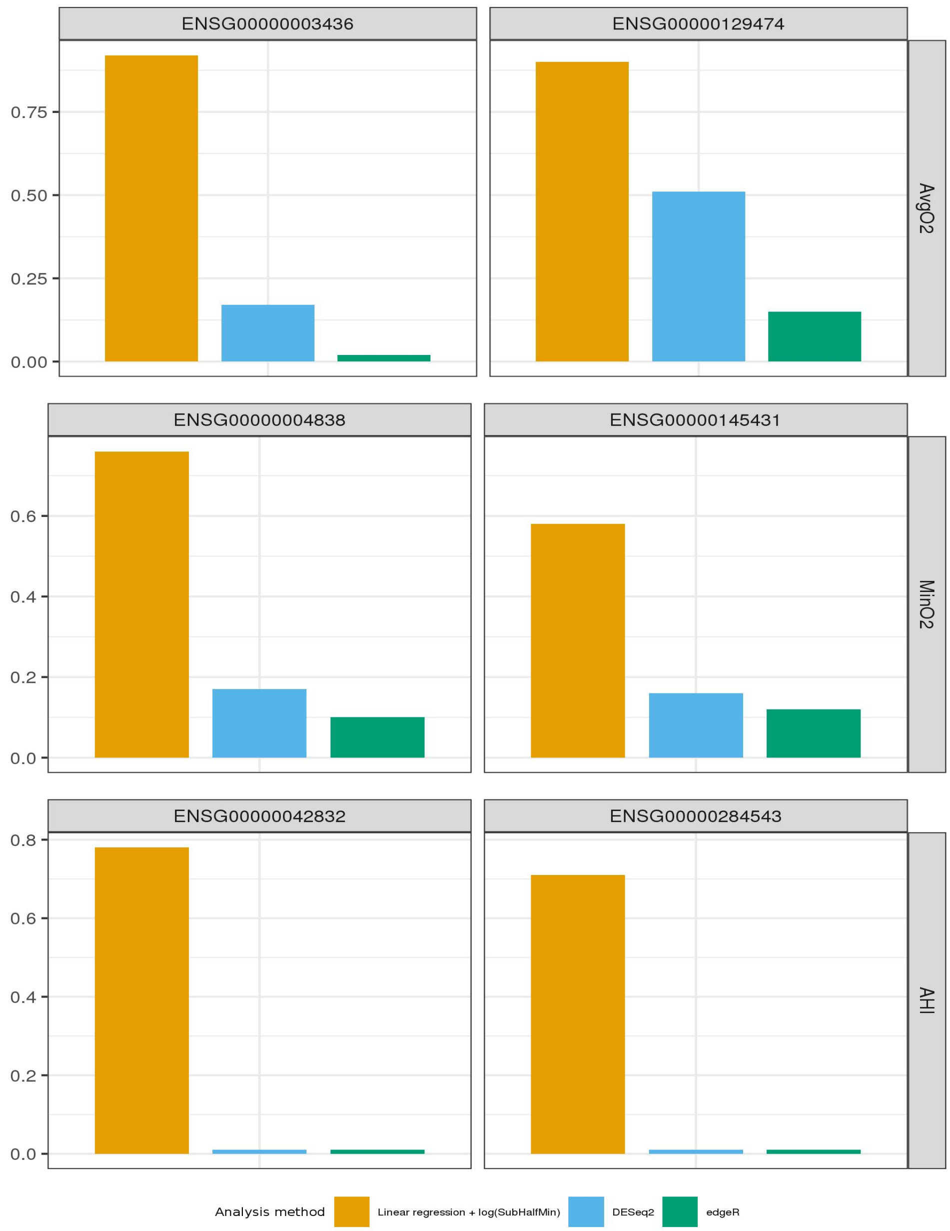

Figure S9: Estimated power for detecting a transcript simulated as associated with the three sleep traits when using **quantile empirical p-values** (compared to Storey empirical p-values in the main manuscript), and association is determined significant if its BH FDR-adjusted p-value is <0.05. We compared logistic regression, DESeq2, and edgeR in transcriptome-wide association analysis for each of the phenotypes.

Comparison of association discovery using a dichotomized versus a continuous phenotype

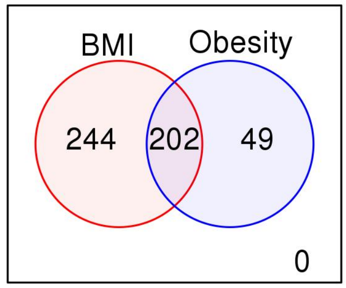

Figure S10: Number of transcripts associated and overlapping between BMI and obesity. Obesity was defined as BMI≥ 30. We used the same filtering criterion for both phenotypes, requiring 50% non-zero transcript values. For obesity, we also relaxed the filtering and required 20% and 30% non-zero values in other analyses, and the results were similar.

Results from gene set enrichment analysis of sleep disordered breathing phenotypes and RNA-seq

We performed gene set enrichment analysis using the `fgsea` R Bioconductor package [4] (version 3.12) using results from the multivariate Wald test of association of the three sleep disordered breathing phenotypes AHI, MinO2, and AvgO2, and using the empirical p-values. We used the Hallmark gene set collection [5]. Table S2 provides the two genes with BH adjusted empirical p-value < 0.1. Table S3 provides enriched gene sets, having FDR adjusted p-value < 0.05.

Table S2. Top genes associated with sleep disordered breathing phenotypes.

ID	Gene	adjLogFC AvgO2	adjLogFC MinO2	adjLogFC AHI	Empirical P-value	BH empirical P-value
ENSG00000145431	PDGFC	0.016	0.002	-2.95E-05	3.55E-06	0.07
ENSG00000172059	KLF11	0.014	0.001	9.16E-05	7.09E-06	0.07

Table S3. Enriched Hallmark gene sets (FDR p-value<0.05) in association with multiple sleep disordered breathing phenotypes modelled jointly.

Pathway	**P-value**	**BH p-value**	**Enrichment Score**	**Negative Enrichment Score**	**N genes**
Hypoxia	4.58E-06	6.12E-05	-0.43	-2.17	168
Inflammatory response	4.65E-06	6.12E-05	-0.47	-2.37	173
Heme metabolism	4.80E-06	6.12E-05	-0.38	-1.95	185
TNFA signaling via NFKB	4.89E-06	6.12E-05	-0.53	-2.74	192
MTORC1 signaling	9.91E-06	8.49E-05	-0.34	-1.78	196
MYC targets v2	1.02E-05	8.49E-05	0.54	2.03	58
Cholesterol homeostasis	1.34E-05	9.58E-05	-0.48	-2.11	70
Apoptosis	1.98E-04	1.10E-03	-0.33	-1.66	154
Glycolysis	1.98E-04	1.10E-03	-0.32	-1.63	177
Complement	2.66E-04	1.16E-03	-0.32	-1.61	181
MYC targets v1	2.66E-04	1.16E-03	0.36	1.62	199
IL-6 JAK STAT3 signaling	2.79E-04	1.16E-03	-0.41	-1.82	78
P53 pathway	3.27E-04	1.22E-03	-0.31	-1.59	186
Coagulation	3.41E-04	1.22E-03	-0.37	-1.73	103
KRAS signaling up	5.85E-04	1.95E-03	-0.31	-1.58	162
Interferon gamma response	6.82E-04	2.13E-03	-0.3	-1.53	198
IL-2 STAT5 signaling	1.04E-03	3.05E-03	-0.29	-1.51	187

UV response up	2.91E-03	8.08E-03	-0.3	-1.5	143
Epithelial mesenchymal transition	4.90E-03	1.29E-02	-0.3	-1.46	142
Adipogenesis	5.91E-03	1.48E-02	-0.27	-1.4	184
G2M checkpoint	6.44E-03	1.53E-02	0.32	1.44	192

References

1. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**(1): p. 323.
2. van der Laan, M.J. and A.E. Hubbard, *Quantile-function based null distribution in resampling based multiple testing.* Stat Appl Genet Mol Biol, 2006. **5**: p. Article14.
3. Storey, J., A. Bass, A. Dabney, et al., *qvalue: Q-value estimation for false discovery rate control*, in *R package version 2.18.0.* 2019.
4. Korotkevich, G., V. Sukhov, and A. Sergushichev, *Fast gene set enrichment analysis.* bioRxiv, 2019: p. 060012.
5. Liberzon, A., C. Birger, H. Thorvaldsdóttir, et al., *The Molecular Signatures Database Hallmark Gene Set Collection.* Cell Systems, 2015. **1**(6): p. 417-425.