Supplementary data

Appendix A: FBAT sufficient statistic approach

The FBAT methodology can analyze general pedigrees with missing parental genotype information while maintaining the robustness against population stratification and admixture. The methodology is based on the seminal work by Laird and Rabinowitz [1], describing the computation of the sufficient statistic S_i. The sufficient statistic S_i is a function of the observed genotypes within a nuclear family, which can include offspring genotypes and parental genotypes. The intuition is that the approach identifies a set of potential offspring genotype configurations such that the corresponding conditional probability of offspring genotypes does not depend on any population parameters. For an example, we recall the example B from [2]: We consider a nuclear family where genotype data for two bi-allelic variants is observed for three offspring and the father. Denote the two alleles for the first variant by A and a, and by B and b for the second variant. The observed unphased genotype of the father is (A/a, B/b), whereas the observed offspring genotype is (A/A, B/B), (A/a, B/b), and (A/a, B/b). Following the algorithmic steps described by Horvath et al., the FBAT haplotype algorithm identifies the following conditional distribution. Let g_1= (A/A, B/B) and g_2= (A/a, B/b). Then, the conditional distribution gives equal weight of 1/3 to the three possible ordered offspring genotype distributions (g_1, g_2, g_2), (g_2, g_1, g_2), and (g_2, g_2, g_1). For more technical details, we refer to previous publications [1–3].

Appendix B: estimating empirical p-values and adaptive stopping rules

As described in the example above, the FBAT haplotype distribution provides a conditional offspring genotype distribution that does not depend on any unknown parameters. After computing the test statistic of interest based on observed offspring genotype and phenotype data, it is straightforward to draw new offspring genotype configurations for each nuclear

family, according to the corresponding conditional distribution. In the example above, this would correspond to draw one of the three possible genotype configurations with equal weight of 1/3. Based on this simulated offspring genotype data, test statistics are re-evaluated using the observed phenotype data. By comparison between the observed test statistic and simulated test statistics based on a sufficiently large number of simulations, empirical p-values are estimated as the proportion of simulations where the re-evaluated test statistics attained a more extreme value.

Appendix C: powerful tests for sparse signals

We consider the scenario of an affected offspring trio. Both parental genotypes are observed along the p variants in the analysis region. As noted in Chen et al. [4], if there is no variant for which all three observed genotypes (mother, father, offspring) are heterozygous, the phase information can be recaptured from the observed unphased genotype data. However, as described in Hecker et al. [5], treating inferred haplotypes as observed haplotypes can lead to misspecification.

Nevertheless, more specifically, if there is no variant for which both parental genotypes are heterozygous, haplotypes can be phased, and the resulting conditional genotype distribution obtained by the FBAT haplotype algorithm equals the conditional distribution where we treat the haplotypes as observed. If we restrict the genetic data to rare variants, this is true for most nuclear families.

Let us denote the phased parental mating type for such a trio by $G = (h_1^M, h_2^M)\ x\ (h_1^F, h_2^F)$. The possible offspring genotypes are denoted by $X_1 = (h_1^M + h_1^F)$, $X_2 = (h_1^M + h_2^F), X_3 = (h_2^M + h_1^F)$ and $X_4 = (h_2^M + h_2^F)$. We assume the following, commonly used, disease model that describes the conditional offspring genotype distribution

$$P(X_i|T=1,G)=\frac{\exp(\beta^T X_i)}{\sum_{j=1}^{4}\exp\left(\beta^T X_j\right)}$$

where the p dimensional vector β describes the genetic effects of the variants in the region.

If we denote the inherited offspring haplotypes by (h^M, h^F), this model factors into the product of the two likelihoods

$$P\left(h^g=h_j^g|T=1,G\right)=\frac{\exp\left(\beta^T h_j^g\right)}{\exp\left(\beta^T h_1^g\right)+\exp\left(\beta^T h_2^g\right)},\quad j=1,2,\qquad g=M,F$$

In the likely case that all minor alleles of a parent are located on the same haplotype, this matches the scenario of Weakly Correlated Designs that is described in the paper by Mukherjee et al. [6] about sparse binary regression (Definition 4.1). They showed that in the sparse regime, the higher criticism and the maximum statistic can identify sparse alternatives efficiently (see Theorem 7.4).

We note that this derivation only motivates the application to affected offspring trios, the statistics can be applied in all FBAT scenarios and none of the assumptions described here must be satisfied to obtain a valid test. The Type I error is preserved because we utilize a simulation-based approach.

Appendix D: Tables simulation study

Table S1. Power estimates at a significance level of 5% for the FBAT, gTDT and RV-GDT statistics. We considered seven scenarios, separately for $p = 30$ and $p = 50$ variants. All results based on 1,000 replicates.

		FBAT					gTDT			RV-GDT
		ACAT	Burden	HC	MAX	SKAT	gTDT-AD	gTDT-CH	gTDT-DOM	RV-GDT
$\boldsymbol{p = 30}$	1	84.2%	32.2%	80.4%	77.6%	87.8%	32.7%	17.2%	39.7%	45.7%
	2	80.6%	18.7%	78.2%	75.1%	85.2%	19.7%	10.7%	21.5%	27.5%
	3	78.7%	20.5%	73.4%	72.3%	83.8%	21.9%	11.0%	27.0%	32.7%
	4	48.2%	7.2%	50.8%	48.3%	48.1%	7.7%	5.3%	8.7%	0.16%
	5	29.6%	11.4%	39.7%	32.9%	11.8%	11.6%	7.5%	13.7%	18.5%
	6	85.3%	66.0%	80.2%	71.9%	82.6%	67.5%	13.4%	29.1%	77.2%
	7	84.1%	65.6%	78.9%	70.3%	81.0%	66.2%	11.4%	29.3%	77.9%
$\boldsymbol{p = 50}$	1	87.7%	50.5%	81.6%	77.6%	90.6%	51.3%	17.9%	31.8%	65.6%
	2	82.5%	23.7%	76.7%	73.7%	88.1%	25.2%	9.9%	12.1%	36.3%
	3	78.6%	32.8%	71.4%	68.6%	83.5%	33.4%	11.7%	18.9%	47.8%
	4	46.1%	8.0%	47.0%	43.1%	50.1%	8.7%	6.4%	4.9%	0.8%
	5	24.7%	8.3%	30.5%	27.4%	10.9%	8.6%	7.6%	9.0%	13.8%
	6	87.5%	72.8%	79.7%	69.1%	84.8%	73.7%	13.5%	21.9%	83.4%
	7	84.3%	65.6%	75.4%	65.4%	82.7%	66.9%	13.1%	19.3%	76.9%

Table S2. Type 1 error and power estimates at significance levels of 1% and 5%, all results based on 1,000 replicates. Scenario 1 and 2 are based on 1,000 trios, scenario 3 and 4 are based on 10,000 trios.

		FBAT					gTDT			RV-TDT
	scenario	ACAT	Burden	HC	MAX	SKAT	gTDT-AD	gTDT-CH	gTDT-DOM	RV-TDT BRV
$\alpha = 0.05$	null	5.1%	5.2%	5.0%	3.9%	4.5%	5.2%	5.9%	2.5%	5.5%
	1	66.3%	5.6%	45.0%	81.2%	14.5%	5.6%	4.1%	1.7%	8.0%
	2	91.0%	5.6%	80.8%	94.9%	85.0%	5.6%	6%	1.5%	2.1%
	3	77.7%	43.8%	82.3%	67.4%	14.2%	43.8%	5.2%	3.3%	56.2%
	4	48.3%	7.1%	60.4%	51.9%	6.3%	7.1%	4.1%	2.8%	12.0%
$\alpha = 0.01$	null	1.0%	1.3%	1.2%	0.7%	1.5%	1.3%	1.1%	0.4%	1.3%
	1	46.2%	1.3%	39.9%	61.2%	3.4%	1.3%	0.8%	0.2%	2.3%
	2	81.8%	0.9%	77.7%	87.5%	54.6%	0.9%	1.3%	0.2%	0.5%
	3	43.3%	23.0%	48.3%	39.2%	2.7%	23.0%	1.3%	0.5%	32.2%
	4	21.4%	1.8%	31.9%	27.3%	0.9%	1.8%	0.8%	0.5%	3.1%

Table S3. Type 1 errors at a significance level of 0.5% for the FBAT, gTDT and RV-GDT statistics. We considered four scenarios, separately for $p = 30$ and $p = 50$ variants. All results based on 10,000 replicates.

		FBAT					gTDT			RV-GDT
		ACAT	Burden	HC	MAX	SKAT	gTDT-AD	gTDT-CH	gTDT-DOM	RV-GDT
$\boldsymbol{p = 30}$	null	0.53%	0.49%	0.43%	0.42%	0.51%	0.51%	0.35%	0.65%	0.49%
	adm1	0.5%	0.53%	0.34%	0.33%	0.55%	0.57%	0.56%	0.53%	0.53%
	adm2	0.32%	0.41%	0.36%	0.39%	0.50%	0.51%	0.50%	0.49%	0.00%
	adm3	0.5%	0.47%	0.45%	0.44%	0.52%	0.54%	0.43%	0.37%	96.58%
$\boldsymbol{p = 50}$	null	0.58%	0.46%	0.47%	0.47%	0.57%	0.47%	0.36%	0.44%	0.54%
	adm1	0.49%	0.53%	0.53%	0.53%	0.43%	0.57%	0.51%	0.48%	0.57%
	adm2	0.58%	0.4%	0.55%	0.54%	0.63%	0.44%	0.45%	0.61%	0.0%
	adm3	0.41%	0.5%	0.42%	0.44%	0.35%	0.53%	0.47%	0.45%	9.67%

Appendix E: Supplementary Figures

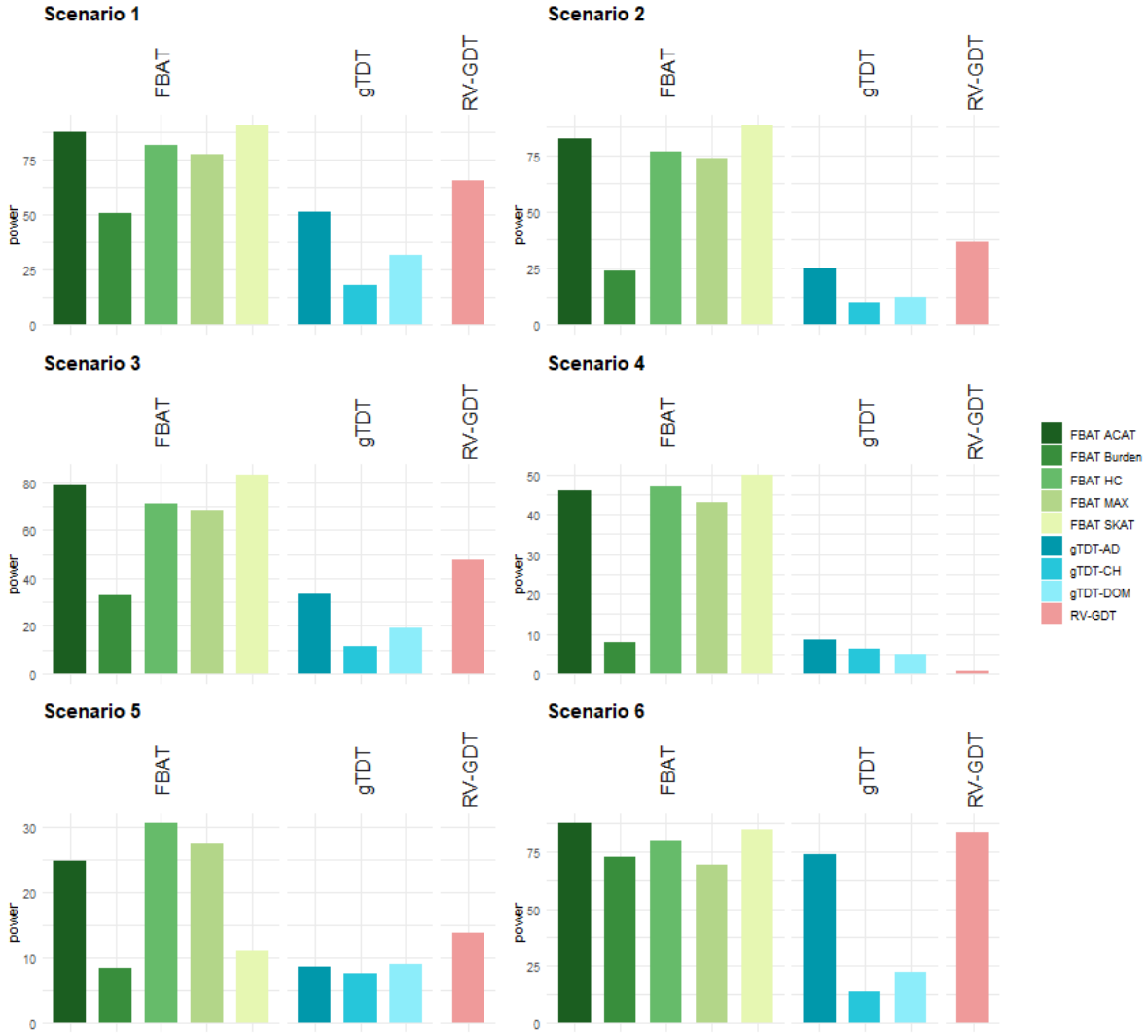

Figure S1. Power results in six different scenarios for genetic regions consisting of 50 variants at a significance level of $\alpha = 0.05$. All results based on 1,000 replicates.

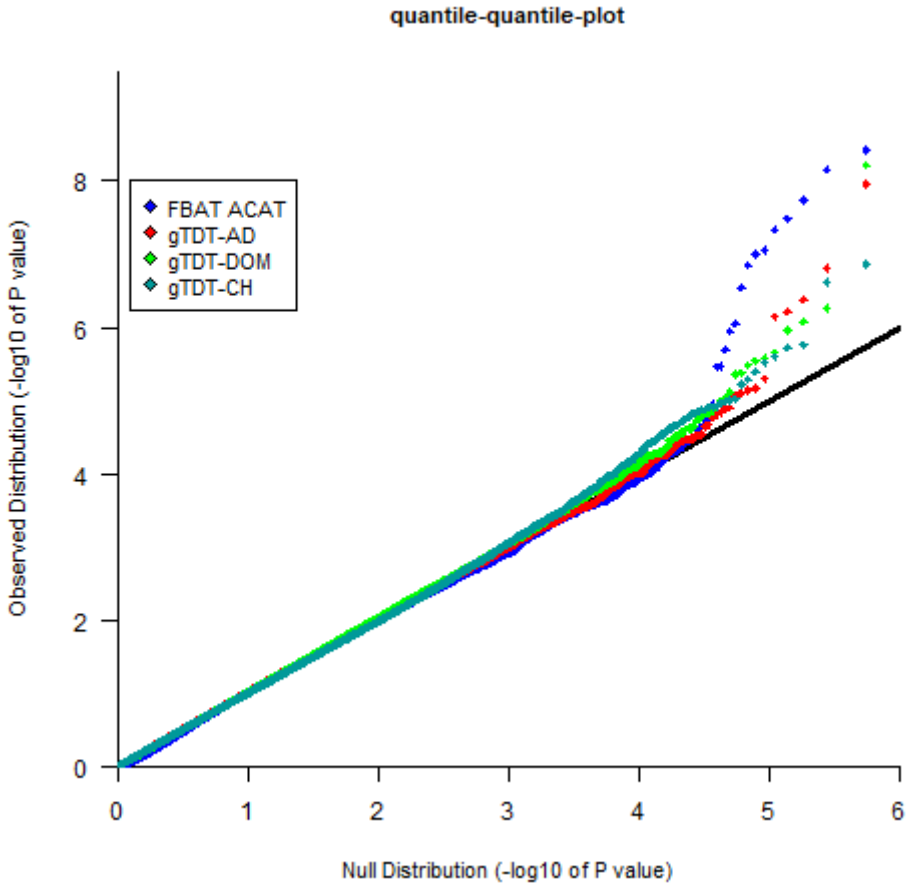

Figure S2. Quantile-quantile plot for Burden, SKAT, MAX, HC, and ACAT test statistics based on approximately 547,000 windows of 50 consecutive rare variants in the analysis of 897 asthmatic trios from Costa Rica.

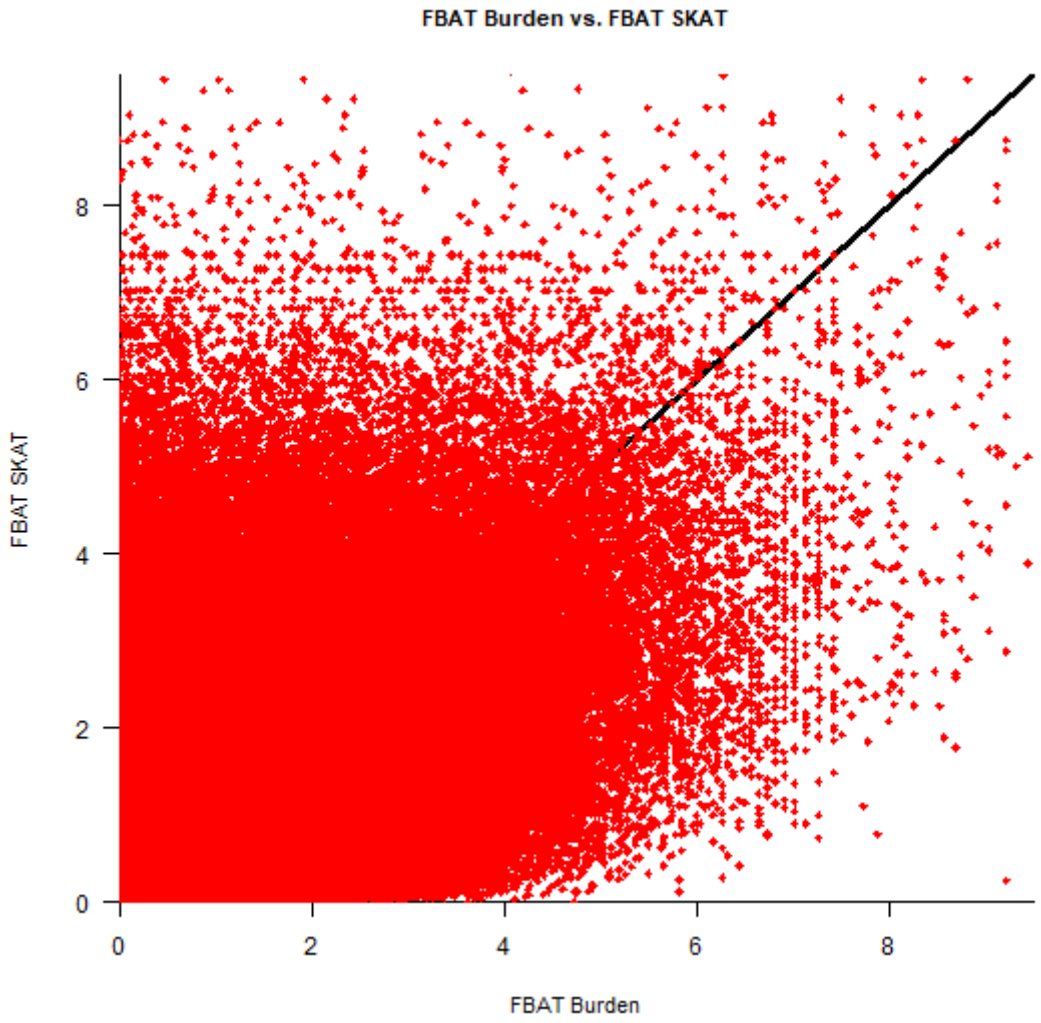

Figure S3: Comparison between the association p-values for FBAT Burden and FBAT-SKAT based on the approximately 547,000 regions in the analysis of the 897 asthmatic WGS offspring trios from Costa Rica.

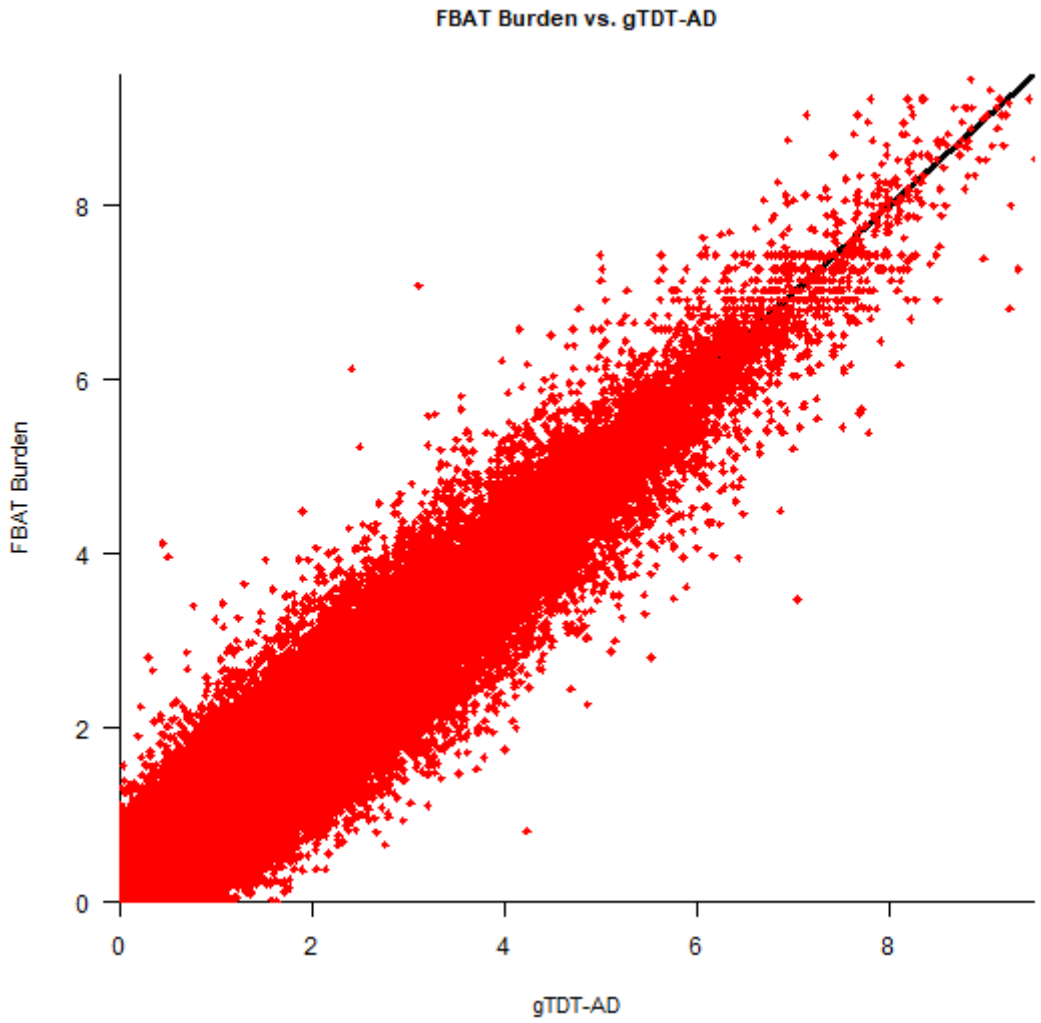

Figure S4: Comparison between the association p-values for FBAT Burden and gTDT-AD based on the approximately 547,000 regions in the analysis of the 897 asthmatic WGS offspring trios from Costa Rica.

Appendix F: whole-genome sequencing data

DNA samples were sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) whole-genome sequencing (WGS) program [7]. WGS was performed at the Northwest Genomics Center. Details on DNA sample handling, quality control, library construction, clustering and sequencing, read processing, and sequence data quality control are described on the TOPMed website (https://www.nhlbiwgs.org/methods). Variant calls were obtained from TOPMed data freeze 7 variant call format files aligned to the GRCh38 genome reference. In our analyses, we included only biallelic SNPs with a minimal depth of coverage of 10 reads that were marked as PASS in the VCF FILTER column.

References

1. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000;19 Suppl 1: S36-42. doi:10.1002/1098-2272(2000)19:1+<::AID-GEPI6>3.0.CO;2-M

2. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol. 2004;26: 61–69. doi:10.1002/gepi.10295

3. Hoffmann TJ, Vansteelandt S, Lange C, Silverman EK, DeMeo DL, Laird NM. Combining disease models to test for gene-environment interaction in nuclear families. Biometrics. 2011;67: 1260–1270. doi:10.1111/j.1541-0420.2011.01581.x

4. Chen R, Wei Q, Zhan X, Zhong X, Sutcliffe JS, Cox NJ, et al. A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis. Bioinformatics. 2015;31: 1452–1459. doi:10.1093/bioinformatics/btu860

5. Hecker J, Laird N, Lange C. A comparison of popular TDT-generalizations for family-based association analysis. Genet Epidemiol. 2019;43: 300–317. doi:10.1002/gepi.22181

6. Mukherjee R, Pillai NS, Lin X. HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL SPARSE BINARY REGRESSION. Ann Stat. 2015;43: 352–381. doi:10.1214/14-AOS1279

7. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv. 2019; 563866. doi:10.1101/563866