The American Journal of Human Genetics, Volume *106*

Supplemental Data

UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test

Zhangchen Zhao, Wenjian Bi, Wei Zhou, Peter VandeHaar, Lars G. Fritsche, and Seunggeun Lee

Supplemental Figures

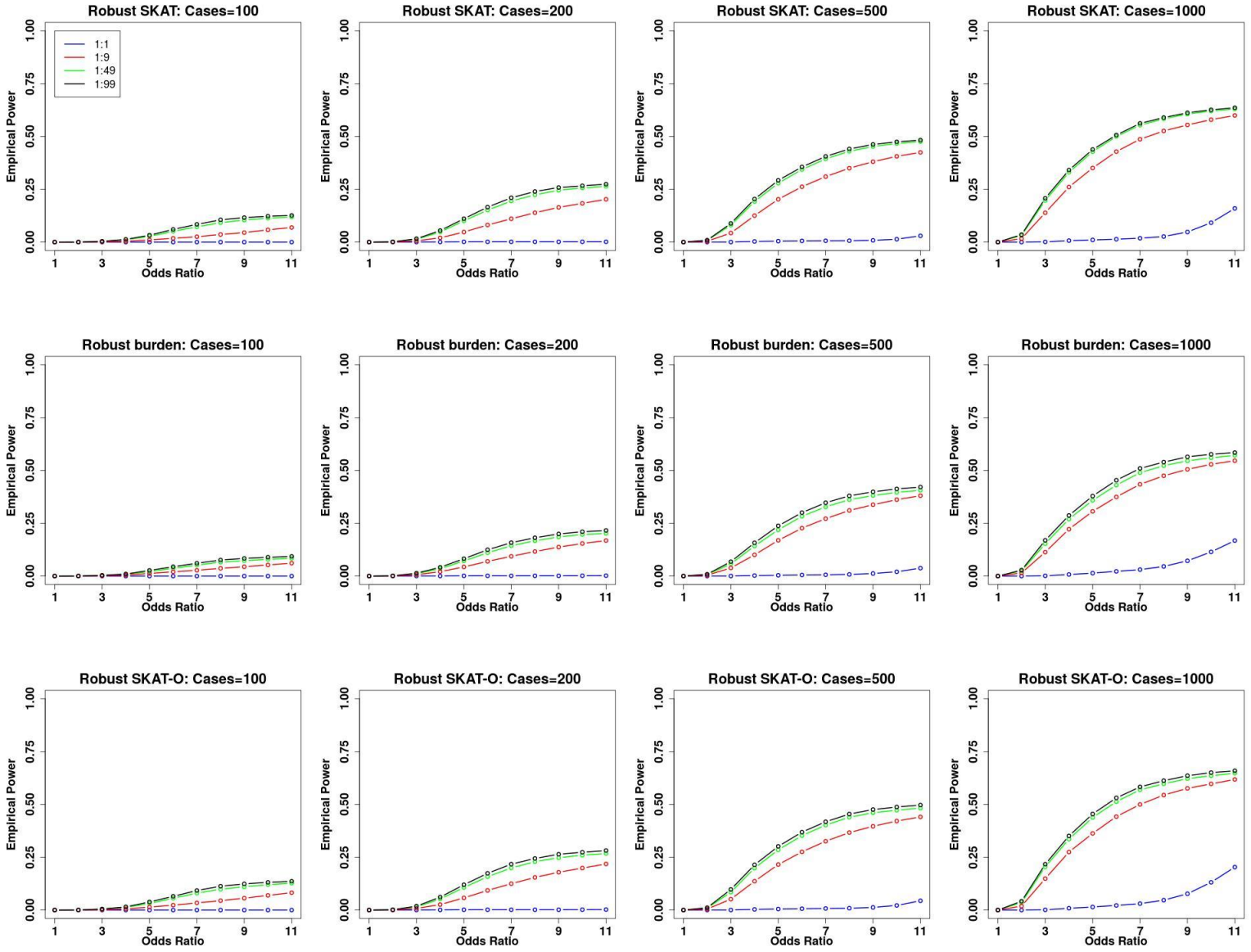


Figure S1. Empirical power estimates for robust SKAT, burden, SKAT-O with the same number of cases across different case control ratios. 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

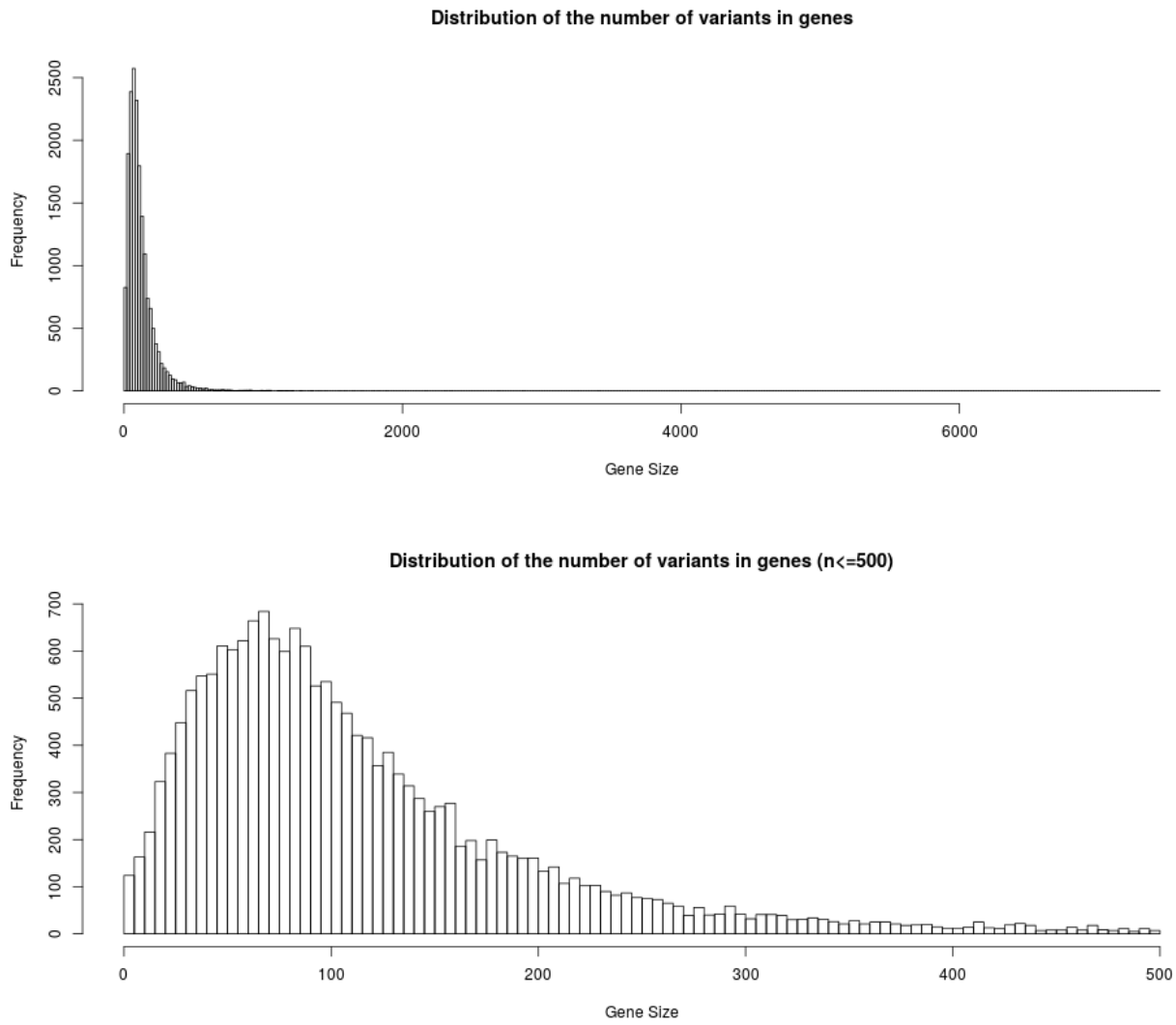


Figure S2. The distribution of the number of variants in genes in the UK-Biobank WES data.

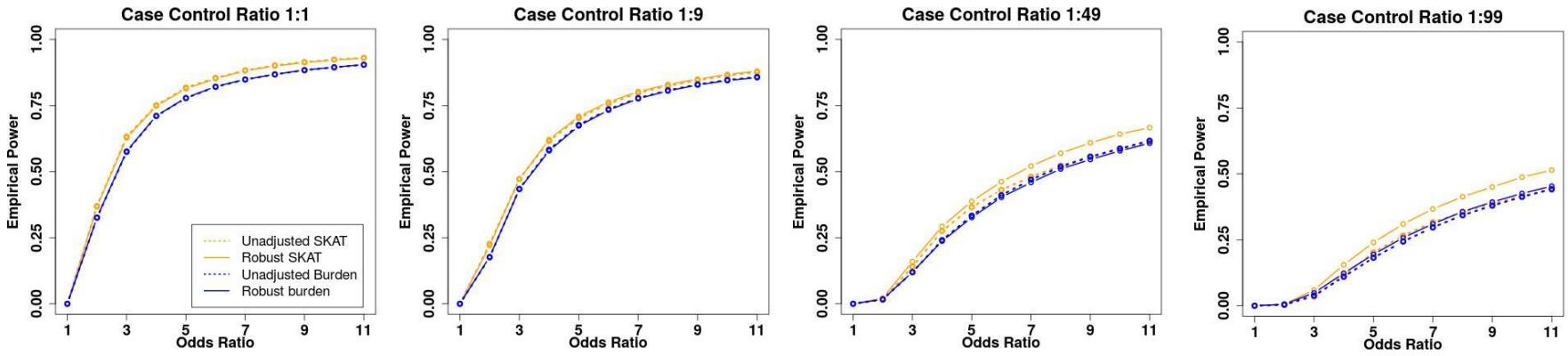
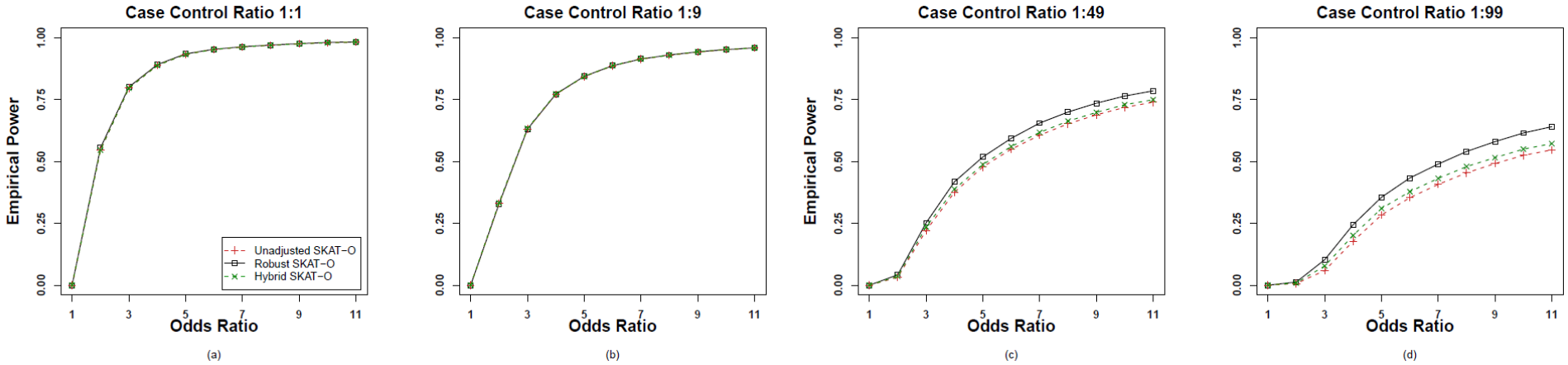


Figure S3. Empirical power estimates for the unadjusted and robust version of SKAT and burden test where 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

(A) Region length 2kb



(B) Region length 3kb

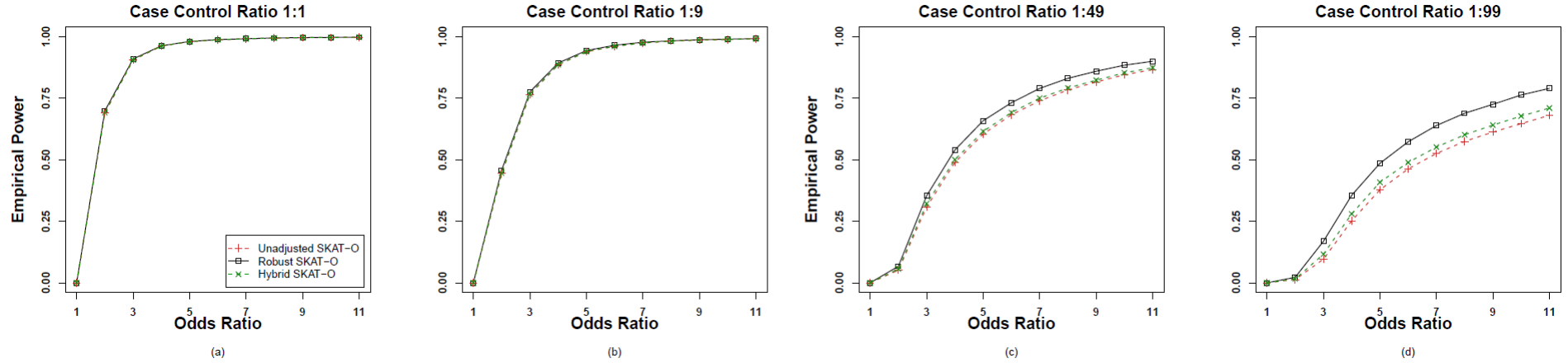


Figure S4. Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. 80% causal variants were risk-increasing and 20% were risk-deceasing. The sample size was 50,000 and 10,000 datasets were generated. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. (A) Region length is 2kb. (B) Region length is 3kb.

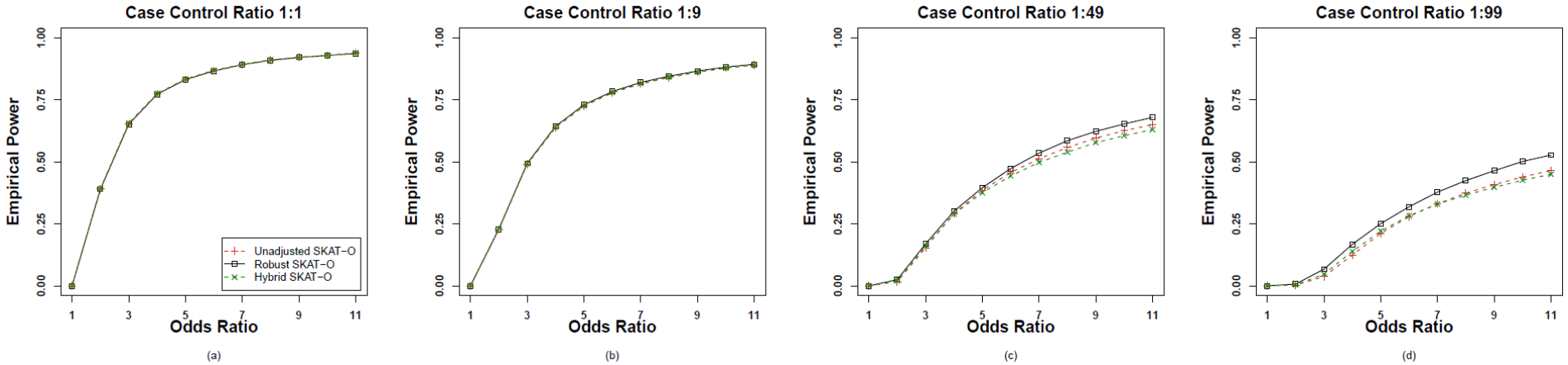


Figure S5. Empirical power estimates for the unadjusted and robust versions of SKAT-O and hybrid method where 30% of variants were causal variants. All causal variants were risk-increasing. The sample size was 50,000 and 10,000 datasets were generated with region length 1kb. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power.

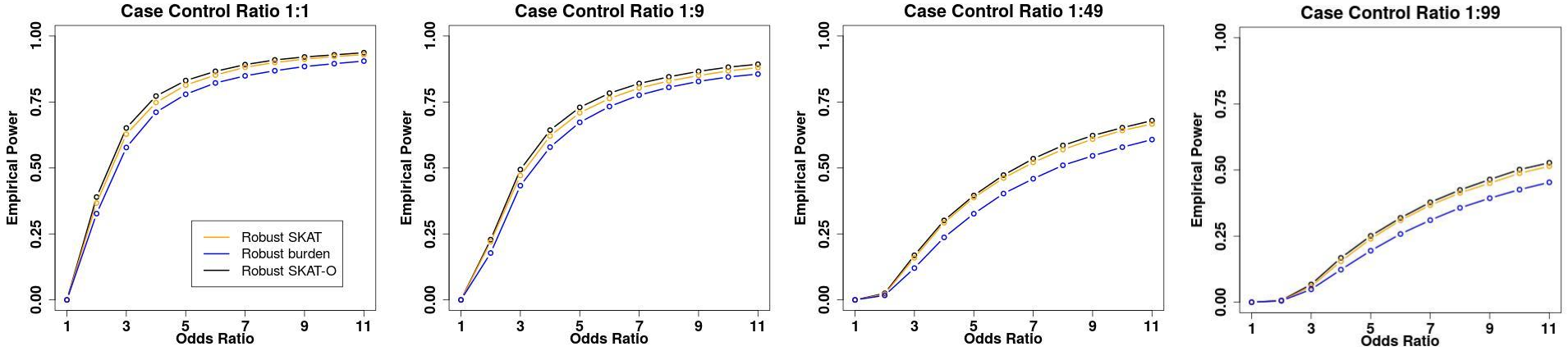


Figure S6. Empirical power estimates for robust SKAT, burden and SKAT-O where 30% of variants were causal variants and all causal variants were risk-increasing. The X-axis represents the genetic effect odds ratio and the Y-axis represents the empirical power. All causal variants had the same odds ratios.

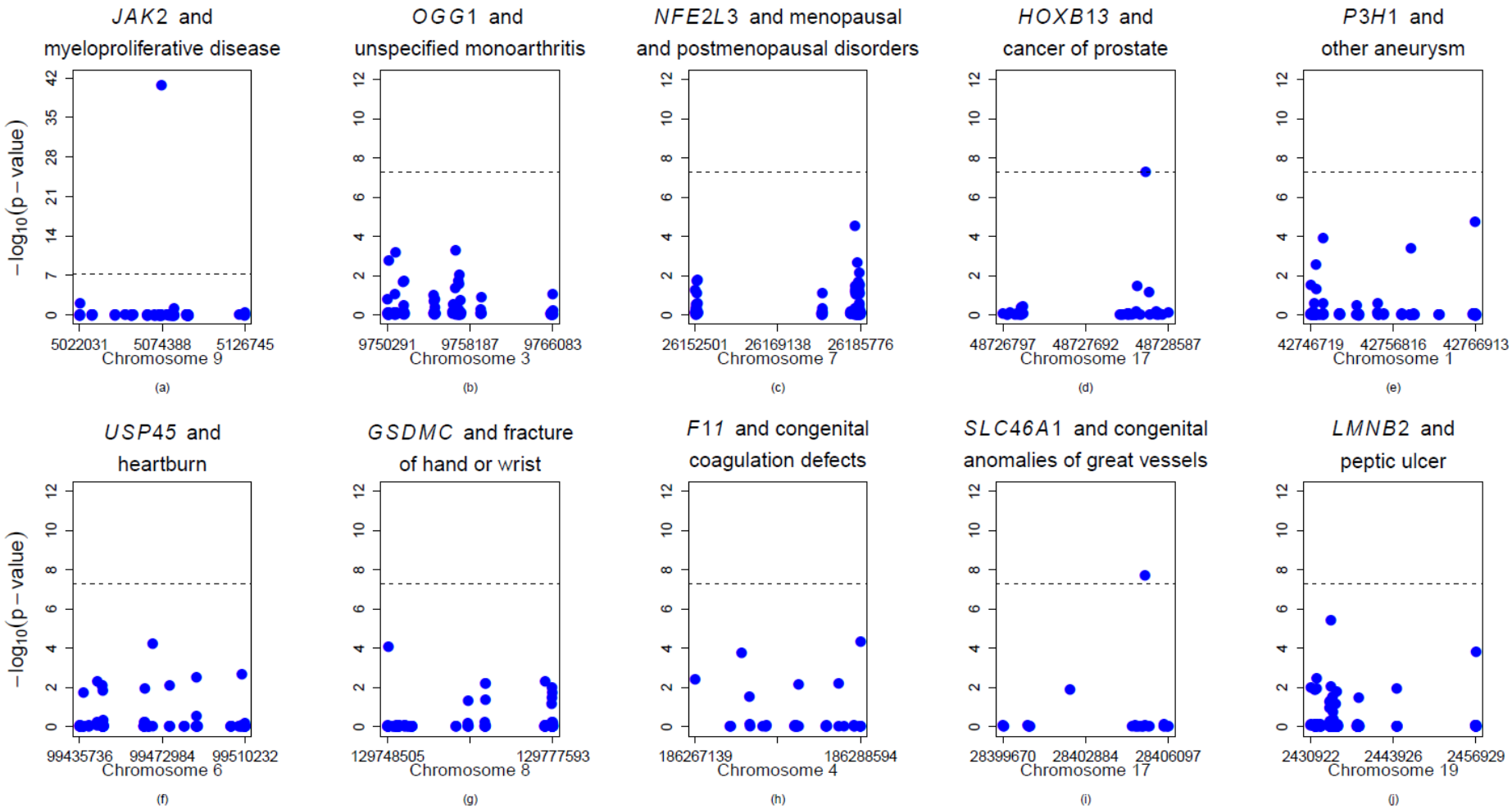


Figure S7. P-values of single variants in 10 significant genes. The X-axis represents the position of each single variant in the gene, and the Y-axis represents the negative log10 p-values of single variants. The dashed line represents the cutoff of 5×10^{-8}.

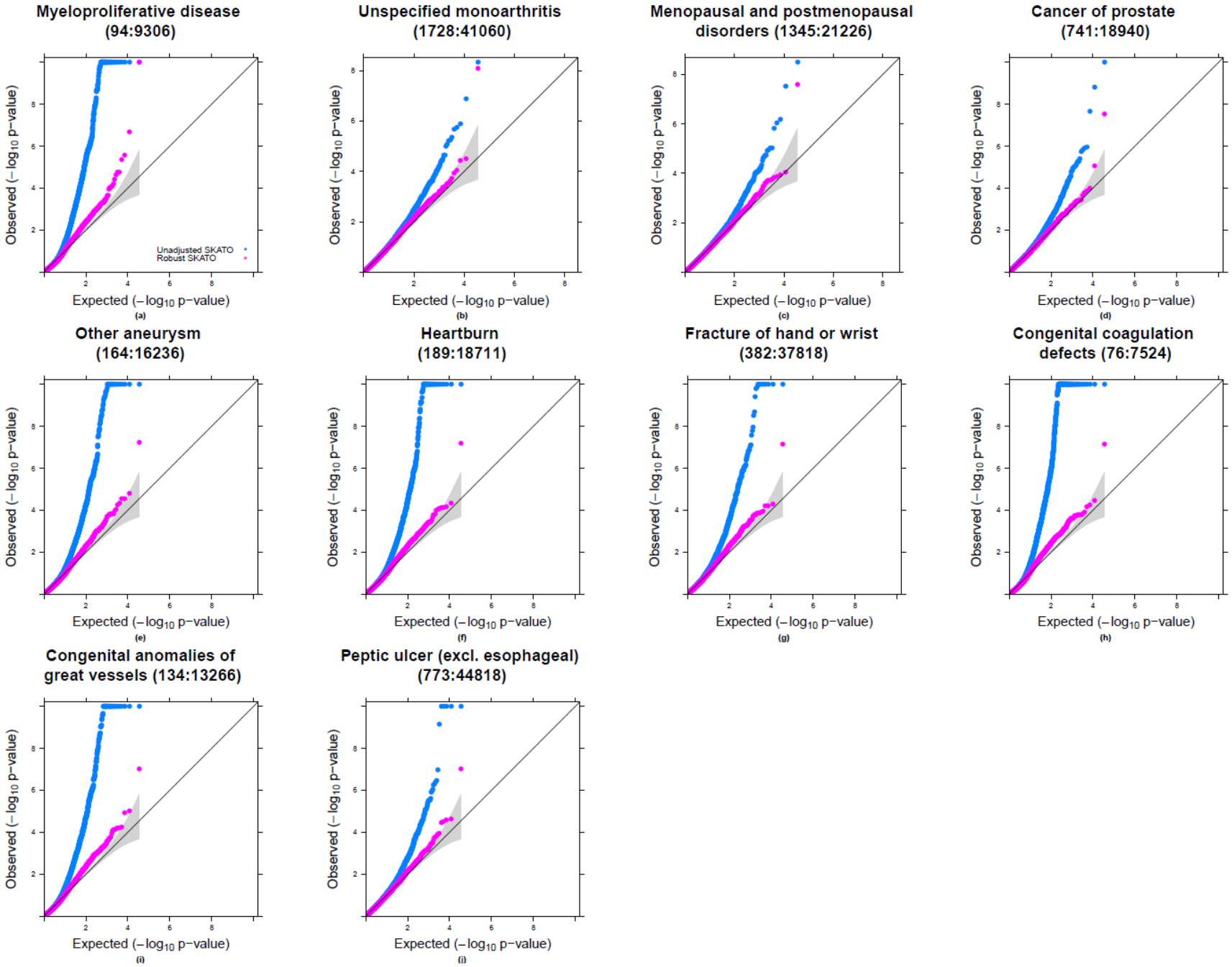


Figure S8. QQ Plots of SKAT-O p-values of 10 selected phenotypes. The X-axis represents the expected negative log10 p-values and the Y-axis represents the observed negative log10 p-values of genes.

Supplemental Tables

Table S1. Type I error rate divided by α of different methods when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6}. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: Control	SKAT	Robust SKAT without additional adjustment	Robust SKAT	Burden	Robust burden without additional adjustment	Robust burden	SKAT-O	Robust SKAT-O without additional adjustment	Robust SKAT-O	Hybrid SKAT-O
10^{-2}	1:1	0.99	0.99	0.99	1.00	1.00	1.00	1.11	1.11	1.11	1.09
	1:9	1.01	1.01	1.01	0.99	0.99	1.00	1.13	1.13	1.13	1.09
	1:49	1.44	1.24	1.22	1.02	0.95	0.95	1.44	1.24	1.23	1.27
	1:99	1.92	1.44	1.41	1.07	0.92	0.91	1.82	1.36	1.33	1.53
	1:199	2.74	1.76	1.71	1.19	0.91	0.88	2.47	1.56	1.52	2.00
	1:399	3.99	2.22	2.14	1.44	0.94	0.89	3.47	1.88	1.82	2.75
10^{-4}	1:1	0.99	0.98	1.03	1.02	1.03	1.00	1.27	1.28	1.32	1.27
	1:9	1.39	1.11	1.14	1.12	1.08	0.99	1.65	1.43	1.40	1.52
	1:49	6.31	1.79	1.65	2.43	1.56	0.97	6.16	2.33	1.79	4.54
	1:99	13.48	2.42	2.13	3.95	2.05	1.02	12.77	3.23	2.17	8.89
	1:199	28.84	3.42	2.89	6.79	2.54	1.01	26.55	4.38	2.72	17.95
	1:399	61.28	4.86	3.79	11.71	2.84	0.95	55.65	5.75	3.37	36.76
2.5×10^{-6}	1:1	1.24	1.03	1.54	1.11	0.94	1.03	1.38	1.19	1.38	1.40
	1:9	2.47	1.19	1.45	1.29	1.41	0.77	2.51	1.61	1.49	2.23
	1:49	28.27	1.80	1.91	6.88	2.91	1.06	23.70	3.24	1.98	16.69
	1:99	89.53	2.67	1.81	16.34	3.80	0.90	71.32	4.87	1.60	42.59
	1:199	262.60	4.62	3.16	39.57	5.99	1.17	211.90	7.93	2.80	126.26
	1:399	715.05	6.70	3.90	86.87	6.65	0.80	577.70	10.70	2.83	347.10

Table S2. Type I error rate divided by α of robust SKAT-O when testing an association with dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6} with three different region length (1) 1kb; (2) 2kb; (3) 3kb. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: control	Region length 1kb Mean:16.33 SD: 4.05	Region length 2kb Mean:32.69 SD: 5.65	Region length 3kb Mean:49.05 SD: 6.71
10^{-2}	1:1	1.11	1.11	1.10
	1:9	1.13	1.11	1.11
	1:49	1.23	1.19	1.18
	1:99	1.33	1.29	1.27
10^{-4}	1:1	1.32	1.15	1.14
	1:9	1.40	1.18	1.20
	1:49	1.79	1.73	1.67
	1:99	2.17	2.09	1.99
2.5×10^{-6}	1:1	1.38	0.96	1.04
	1:9	1.49	1.56	1.08
	1:49	1.98	2.24	1.60
	1:99	1.60	2.44	2.08

Table S3. Type I error rate divided by α of different methods when testing an association between all variants, including both common and rare variants, and dichotomous traits at stringent α levels $\alpha = 10^{-2}, 10^{-4}$ and 2.5×10^{-6}. The sample size was 50,000 and 10^7 datasets were generated.

α	Case: control	SKAT-CommonRare	Robust SKAT-CommonRare without additional adjustment	Robust SKAT-CommonRare
10^{-2}	1:1	1.00	0.99	0.99
	1:9	1.00	1.01	1.01
	1:49	1.11	1.22	1.22
	1:99	1.26	1.42	1.41
	1:199	1.52	1.71	1.72
	1:399	1.97	2.15	2.15
10^{-4}	1:1	0.98	0.98	1.04
	1:9	1.09	1.11	1.13
	1:49	2.42	1.73	1.69
	1:99	4.20	2.26	2.21
	1:199	7.84	3.02	3.03
	1:399	16.02	4.01	4.04
2.5×10^{-6}	1:1	0.94	1.05	1.66
	1:9	1.53	1.17	1.45
	1:49	7.48	1.57	1.94
	1:99	18.54	2.50	1.93
	1:199	50.43	3.87	3.37
	1:399	140.09	4.47	4.46

Table S4. Gene-phenotype associations detected by Robust SKAT-O across 791 phenotypes at $\alpha=2.5\times10^{-6}$ (Number of associations=111)

See excel file “Table S4.xlsx”.

Table S5. Top 3 single rare-variant signals of associations with p-value < 10^{-7} in the UK Biobank WES data.

Phenotype (Phecode)	Gene Name	RS ID	Location	MAF	P-value	Annotation	Polyphen	SIFT
Myeloproliferative disease (200)	*JAK2*	rs77375493	9:5073770:G:T	1.38E-03	1.81E-41	nonsynonymous SNV	probably_damaging	deleterious
	-		9:5022213:G:T	1.06E-04	7.16E-03	nonsynonymous SNV	probably_damaging	deleterious
		rs150221602	9:5081828:G:C	1.12E-03	5.07E-02	nonsynonymous SNV	benign	deleterious
Unspecified monoarthritis (716.2)	*OGG1*	rs113561019	3:9756791:G:A	5.67E-03	4.67E-04	nonsynonymous SNV	probably_damaging	deleterious
		rs17050550	3:9751060:G:T	2.66E-03	5.96E-04	nonsynonymous SNV	benign	tolerated
		rs104893751	3:9750423:G:A	3.89E-03	1.69E-03	nonsynonymous SNV	probably_damaging	deleterious
Menopausal and postmenopausal disorders (627)	*NFE2L3*	rs148235978	7:26184630:A:G	7.93E-03	2.72E-05	nonsynonymous SNV	benign	deleterious
		rs148159120	7:26185140:G:A	5.91E-03	2.18E-03	nonsynonymous SNV	benign	tolerated
		rs144789579	7:26185559:A:G	3.33E-04	6.71E-03	nonsynonymous SNV	benign	tolerated
Cancer of prostate (185)	*HOXB13*	rs138213197	17:48728343:C:T	2.16E-03	5.24E-08	nonsynonymous SNV	probably_damaging	deleterious
		rs764401781	17:48728250:G:A	7.62E-05	3.37E-02	nonsynonymous SNV	benign	deleterious
		rs774579054	17:48728379:C:A	1.02E-04	7.05E-02	nonsynonymous SNV	possibly_damaging	tolerated
Other aneurysm (442)	*P3H1*	rs372710498	1:42766778:C:T	6.71E-04	1.71E-05	nonsynonymous SNV	probably_damaging	tolerated
		rs140254470	1:42748243:C:T	3.35E-04	1.21E-04	nonsynonymous SNV	benign	tolerated
		rs372301077	1:42758970:C:G	9.15E-05	3.88E-04	nonsynonymous SNV	probably_damaging	deleterious
Heartburn (530.9)	*USP45*	rs554927779	6:99468543:D:1	4.52E-03	5.39E-05	frameshift deletion	LoF: High-confidence	-
		rs201110065	6:99508768:G:T	3.70E-04	2.13E-03	nonsynonymous SNV	probably_damaging	deleterious_low_confidence
		rs144269307	6:99488237:G:A	2.65E-05	2.78E-03	nonsynonymous SNV	possibly_damaging	tolerated
Fracture of hand or wrist (804)	*GSDMC*	rs149748731	8:129748706:G:A	3.17E-03	8.17E-05	nonsynonymous SNV	possibly_damaging	tolerated
		rs79403769	8:129776220:C:A	1.14E-03	4.95E-03	nonsynonymous SNV	benign	deleterious
		rs16904151	8:129765749:C:T	6.54E-04	5.81E-03	nonsynonymous SNV	benign	tolerated
Congenital coagulation defects (286.1)	*F11*	rs281875276	4:186288589:T:G	1.32E-04	4.52E-05	nonsynonymous SNV	probably_damaging	deleterious
		rs140190776	4:186273178:G:A	1.97E-04	1.73E-04	splicing	LoF: High-confidence	-
		-	4:186267139:G:T	6.58E-05	3.85E-03	nonsynonymous SNV	probably damaging	deleterious
Congenital anomalies of great vessels (747.13)	*SLC46A1*	rs189103810	17:28405185:A:T	1.75E-03	1.86E-08	nonsynonymous SNV	possibly_damaging	deleterious
		rs281875211	17:28402276:C:T	3.73E-05	1.24E-02	nonsynonymous SNV	possibly_damaging	deleterious
		rs41297071	17:28405926:C:G	5.22E-04	7.13E-01	nonsynonymous SNV	benign	tolerated
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	rs752957340	19:2434094:G:A	2.30E-04	3.83E-06	nonsynonymous SNV	probably_damaging	deleterious
		-	19:2456926:G:T	2.19E-05	1.47E-04	nonsynonymous SNV	-	-
		-	19:2431845:C:A	1.10E-05	3.45E-03	nonsynonymous SNV	benign	deleterious

Table S6. The most significant nearby variant association signals (± 100 Kbp up and down stream) in the UK-Biobank imputed datasets of 400,000 British samples.

Phenotype (Phecode)	Gene Name	RS ID	Location	Ref Allele	Alt Allele	MAF	p-value
Myeloproliferative disease (200)	*JAK2*	rs10283564	chr9:5075628	C	G	2.51E-01	2.30E-17
Unspecified monoarthritis (716.2)	*OGG1*	rs75924392	chr3:9797369	A	G	4.25E-02	4.28E-04
Menopausal and postmenopausal disorders (627)	*NFE2L3*	rs35838658	chr7:26175486	T	A	3.48E-01	2.14E-04
Cancer of prostate (185)	*HOXB13*	rs145922598	chr17:48733224	C	T	3.27E-02	1.17E-04
Other aneurysm (442)	*P3H1*	rs148797090	chr1:42818294	T	C	1.35E-02	1.22E-03
Heartburn (530.9)	*USP45*	rs75597991	chr6:99367094	G	C	2.38E-02	4.08E-02
Fracture of hand or wrist (804)	*GSDMC*	rs409790	chr8:129833249	A	G	8.28E-01	1.84E-02
Congenital coagulation defects (286.1)	*F11*	rs115583874	chr4:186260648	G	T	1.37E-02	6.30E-03
Congenital anomalies of great vessels (747.13)	*SLC46A1*	rs111306228	chr17:28425402	C	A	2.06E-02	2.29E-03
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	rs146283067	chr19:2441769	G	T	1.62E-02	1.31E-03