The American Journal of Human Genetics, Volume *110*

Supplemental information

Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank

Wenjian Bi, Wei Zhou, Peipei Zhang, Yaoyao Sun, Weihua Yue, and Seunggeun Lee

Figure S1. Minor allele counts distribution of the randomly selected 1,000 genes.

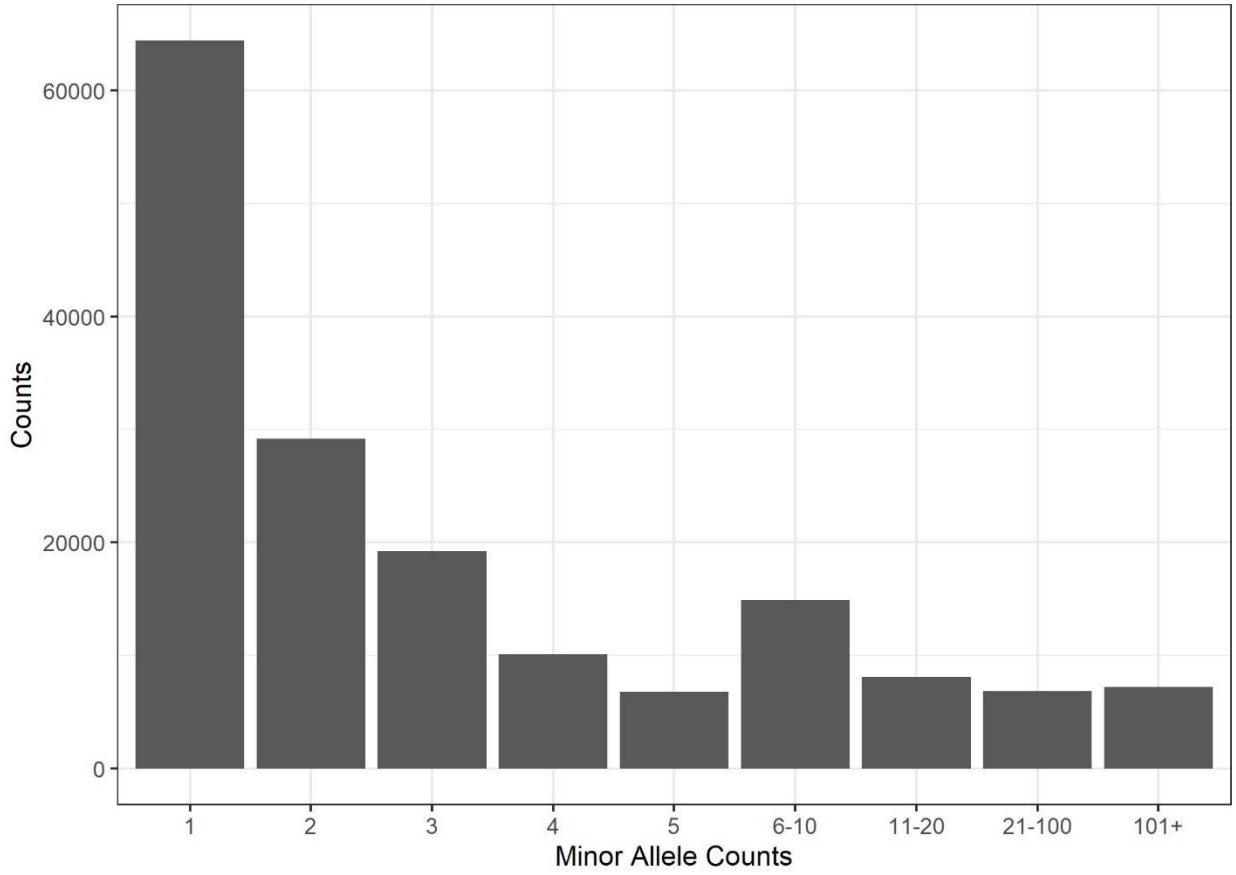

Figure S2. Sample size distribution for the four phenotypes in UK Biobank data analysis

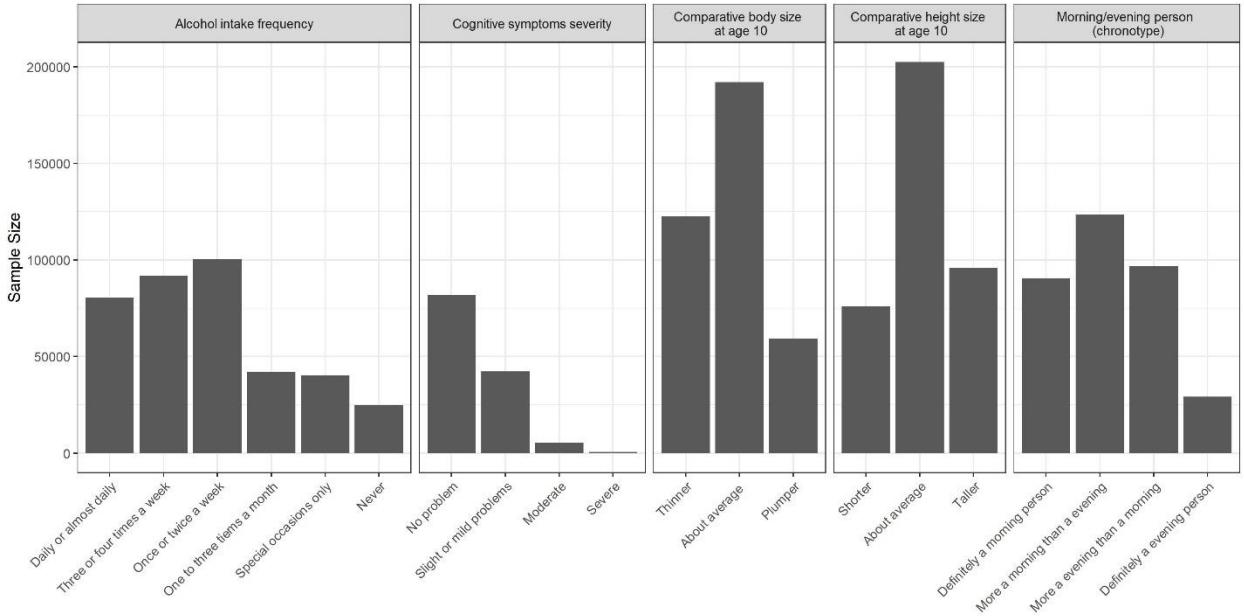

Figure S3. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=10:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=10:2=5:1). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

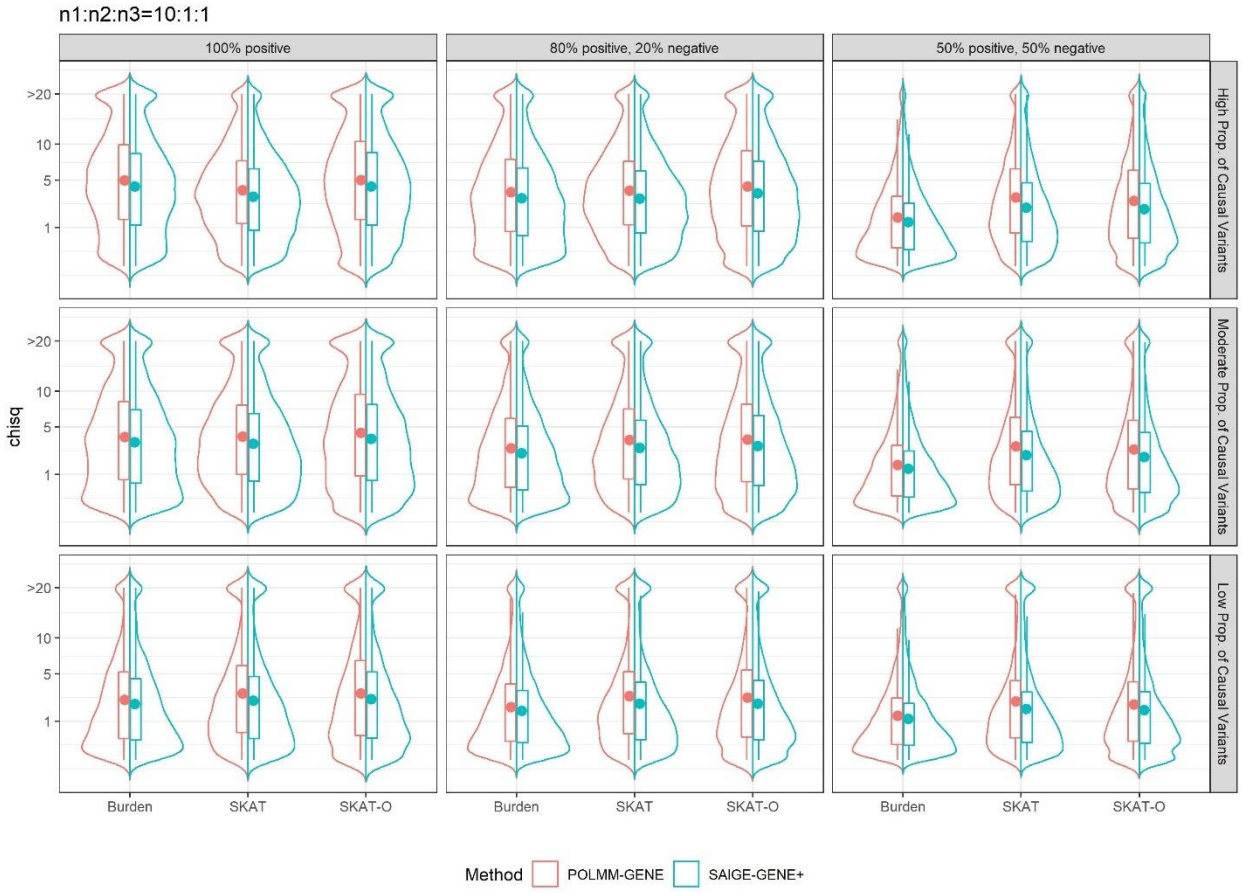

Figure S4. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=30:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=30:2=15:1). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

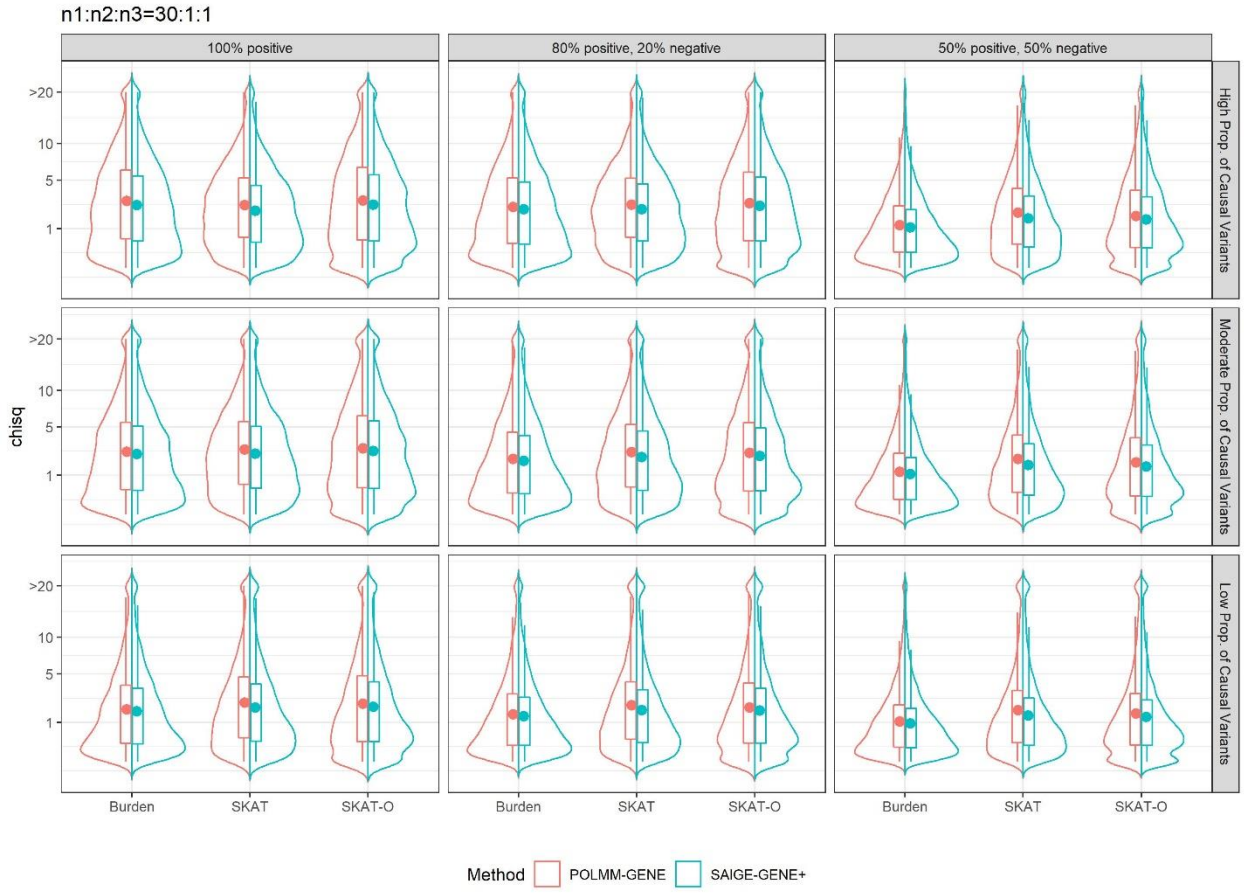

Figure S5. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (RAW). SAIGE-GENE+ (RAW) considered the categorical data as a raw quantitative phenotype of 1, 2, and 3. A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

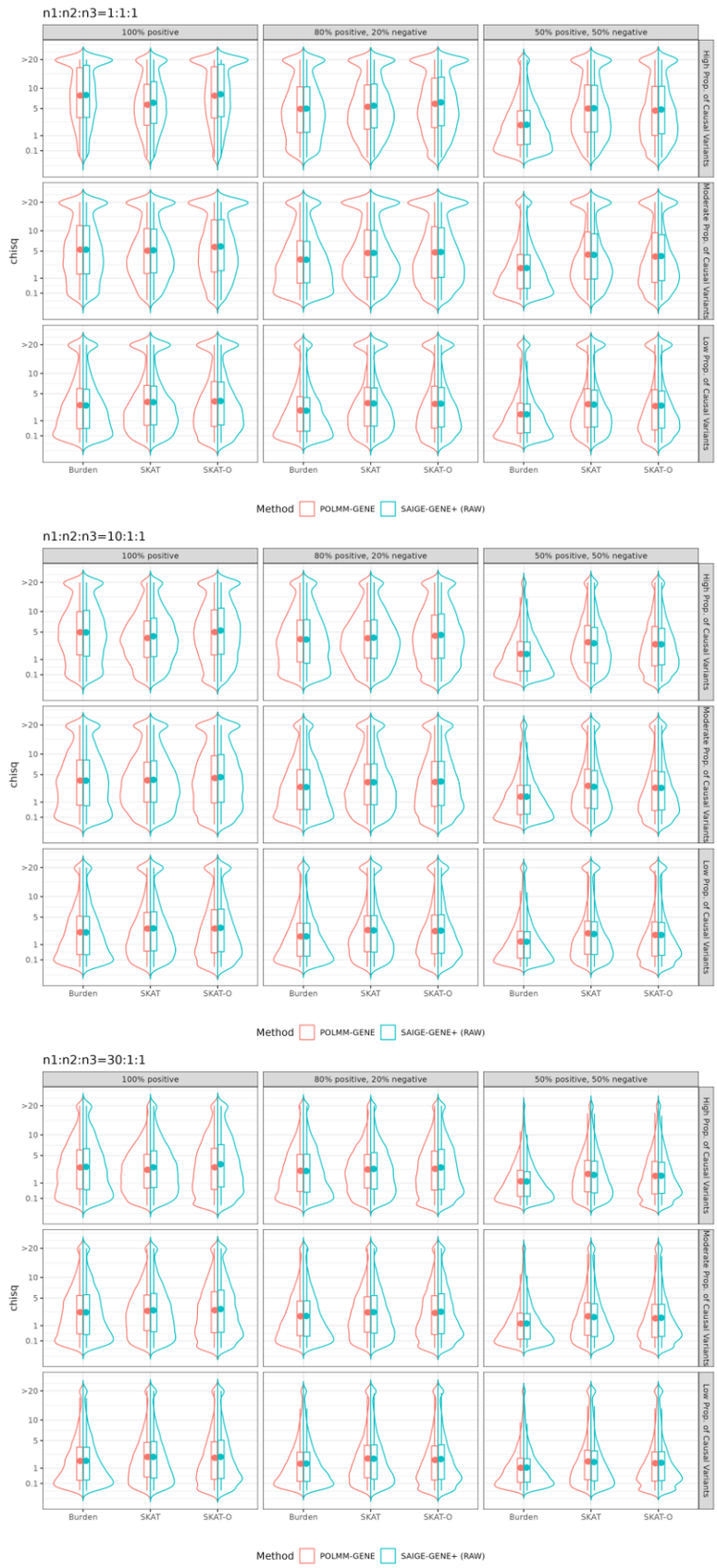

Figure S6. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (INT). SAIGE-GENE+ (INT) considered the categorical data as a quantitative phenotype of 1, 2, and 3. Inverse normalization transformation is conducted for phenotype prior to analysis. A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

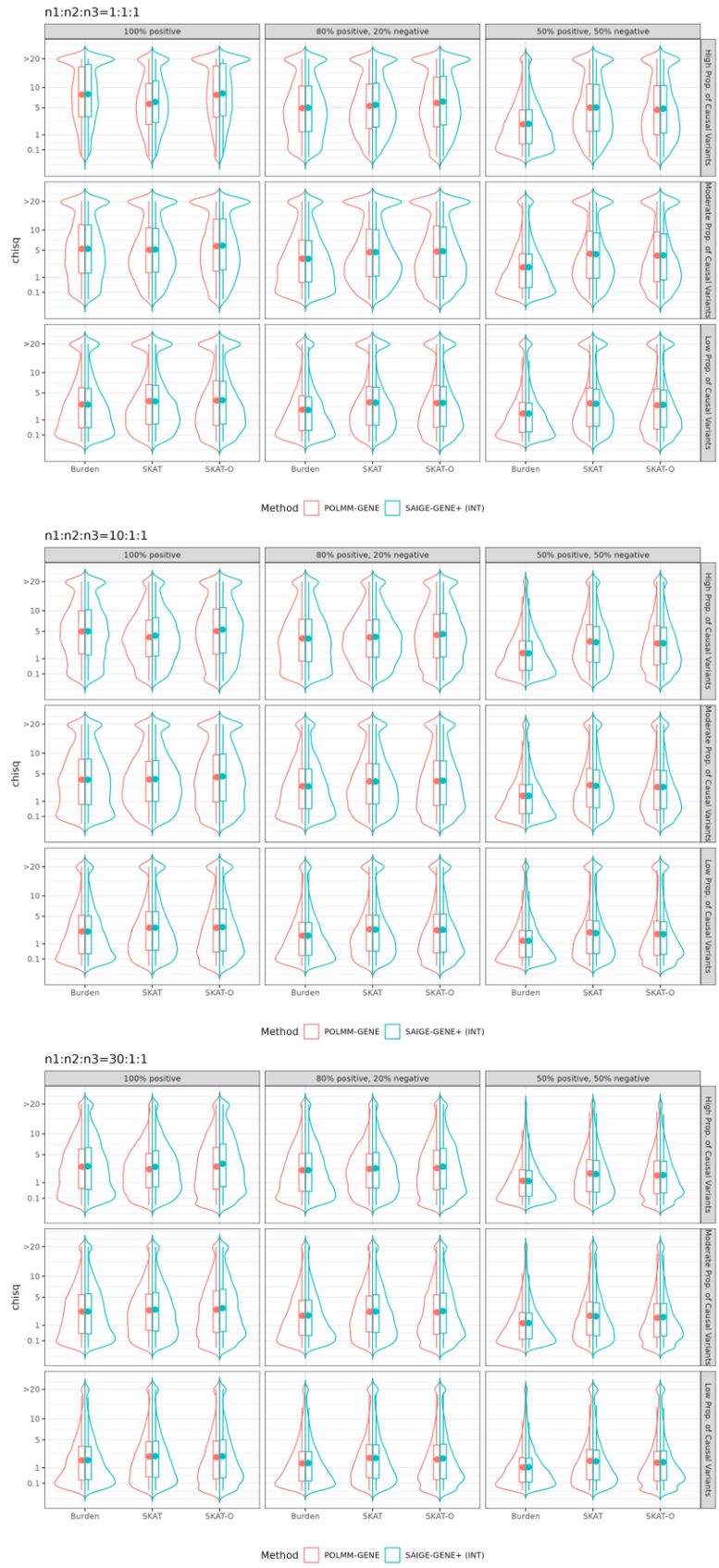

Figure S7. Comparison of p-values using POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=1:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=1:2). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

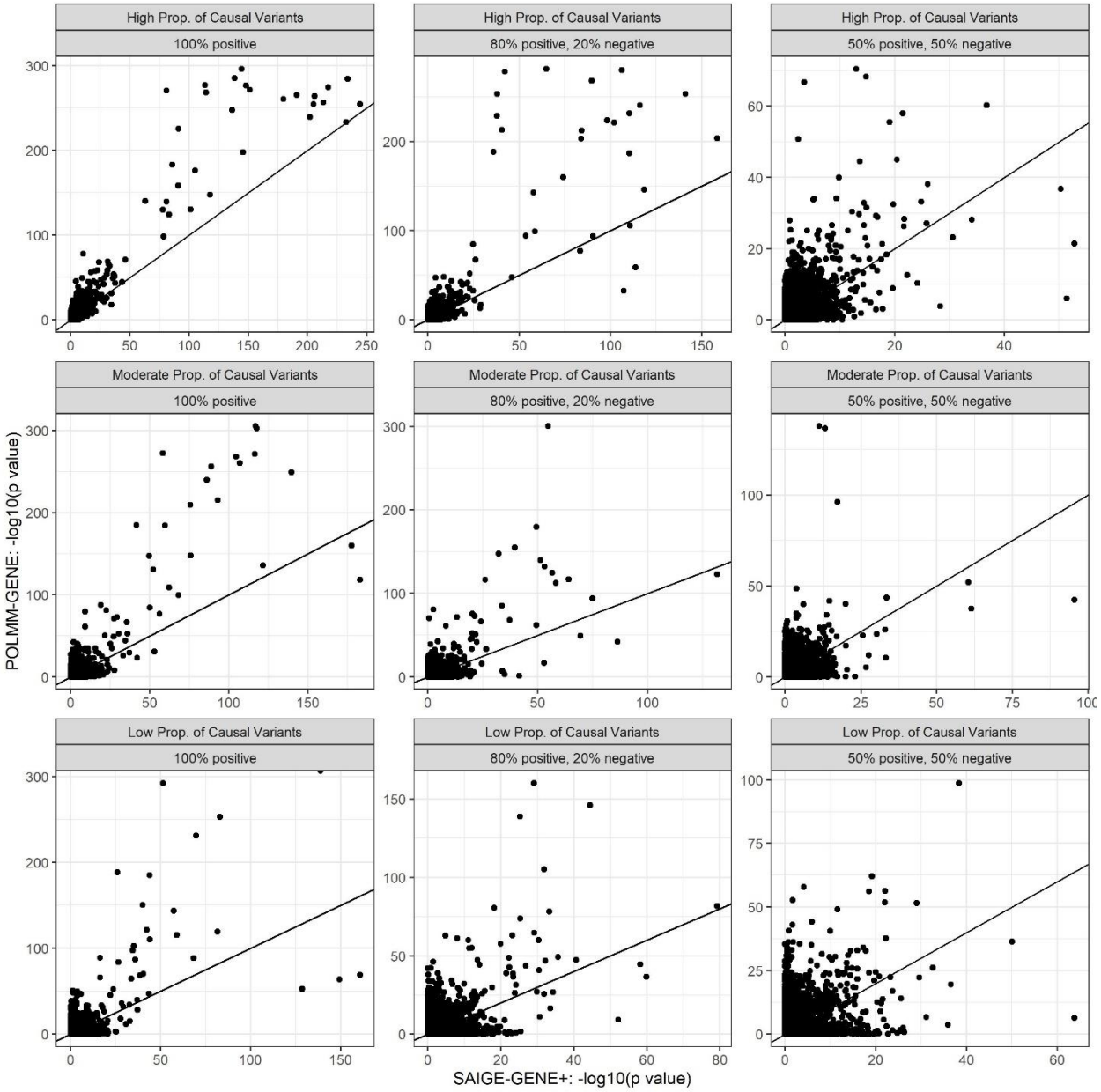

Figure S8. Empirical power of POLMM-GENE, SAIGE-GENE+ (BINA), SAIGE-GENE+ (RAW), and SAIGE-GENE+ (INT) at a significance level of 2.5e-6.

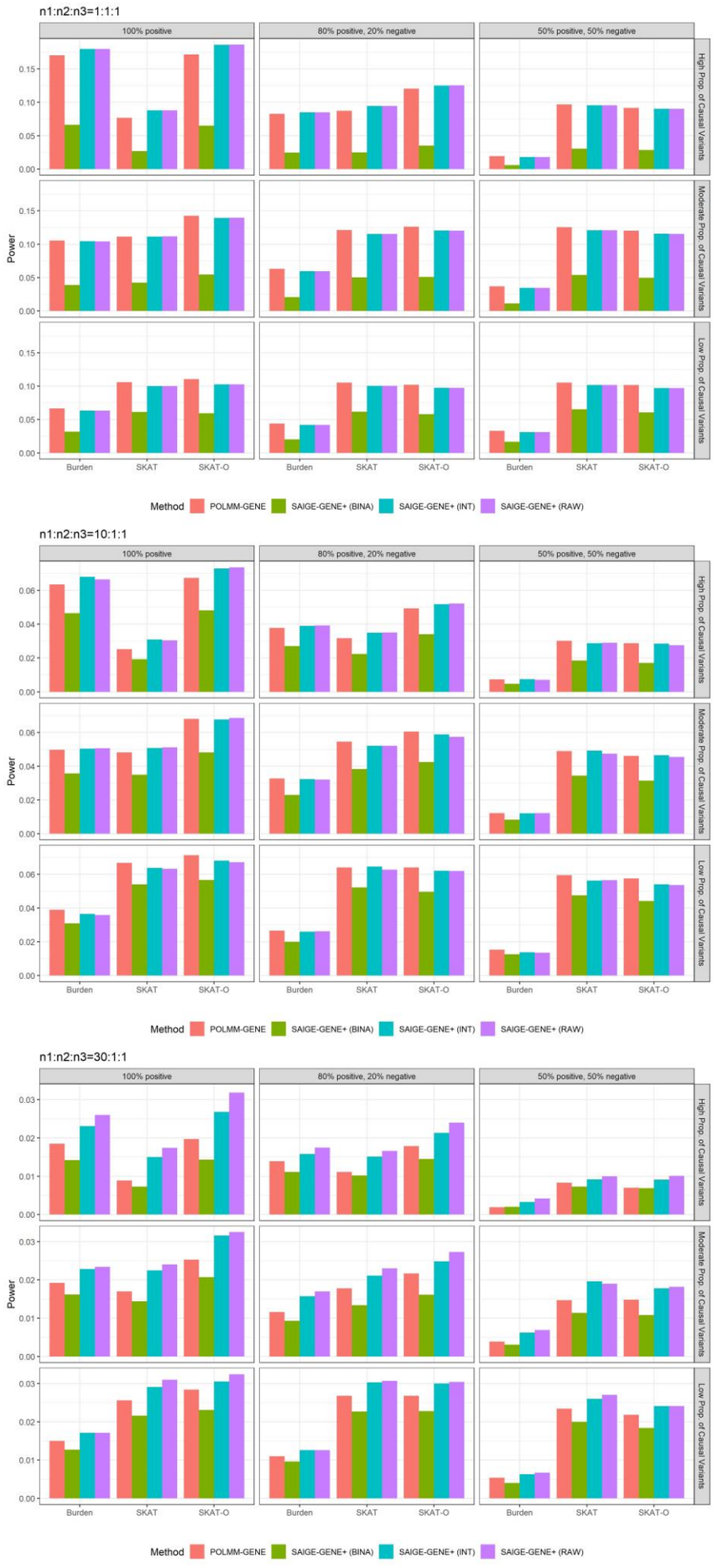

Figure S9. SAIGE-GENE+ (BINA): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The categorical traits were transformed to a binary trait prior to analysis. For trait of "alcohol intake frequency", categories "daily or almost daily" and "three or four times a week" were grouped, and other categories were grouped. For trait of "cognitive symptoms severity", categories "slight or mild problems", "moderate", and "severe" were grouped. For trait of "comparative body size at age 10", categories "about average" and "plumper" were grouped. For trait of "comparative height size at age 10", categories "shorted" and "about average" were grouped. For trait of "morning/evening person", categories "definitely a morning person" and "more a morning than a evening" were grouped, and the other categories were grouped.

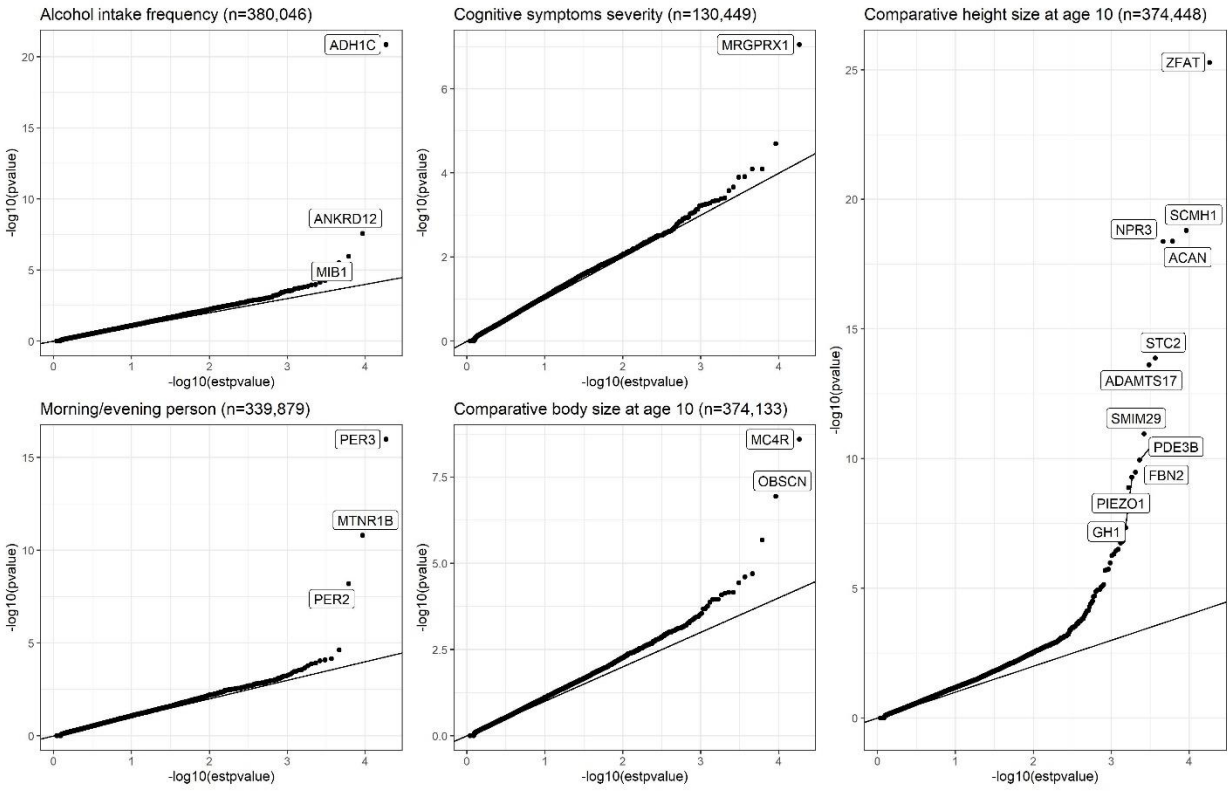

Figure S10. SAIGE-GENE+ (RAW): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The traits were recorded as $1,2,\ldots,m$ where m is the number of categories.

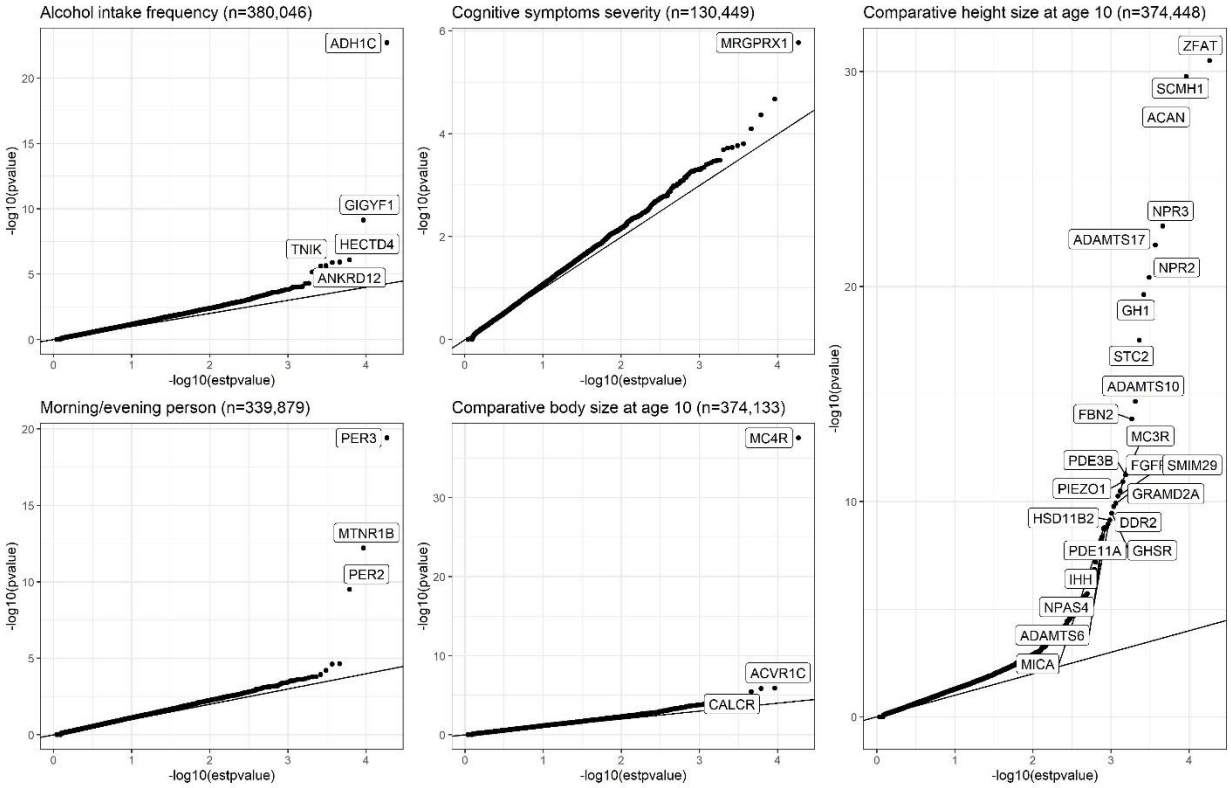

Figure S11. SAIGE-GENE+ (INT): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The traits were recorded as $1,2,\ldots,m$ where m is the number of categories.

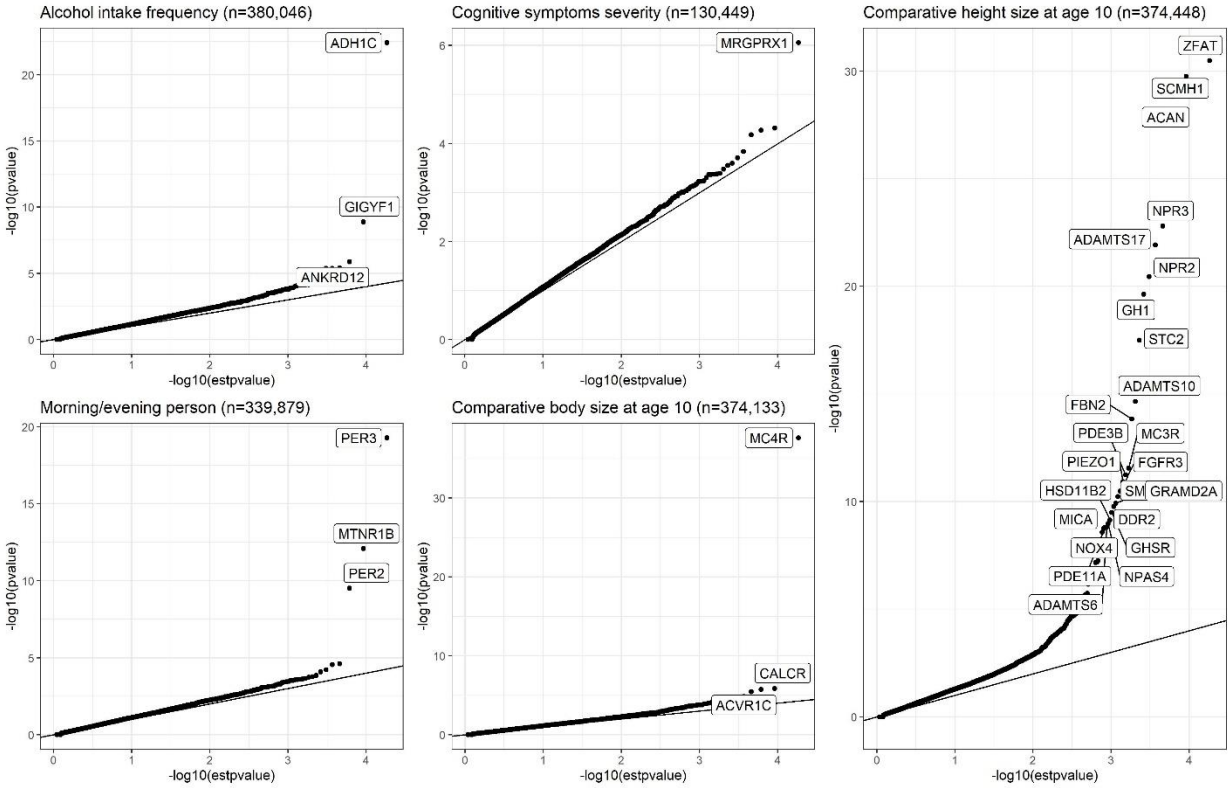

Figure S12. Comparison of POLMM-GENE and SAIGE-GENE+ approaches when analyzing "Comparative height size at age 10"

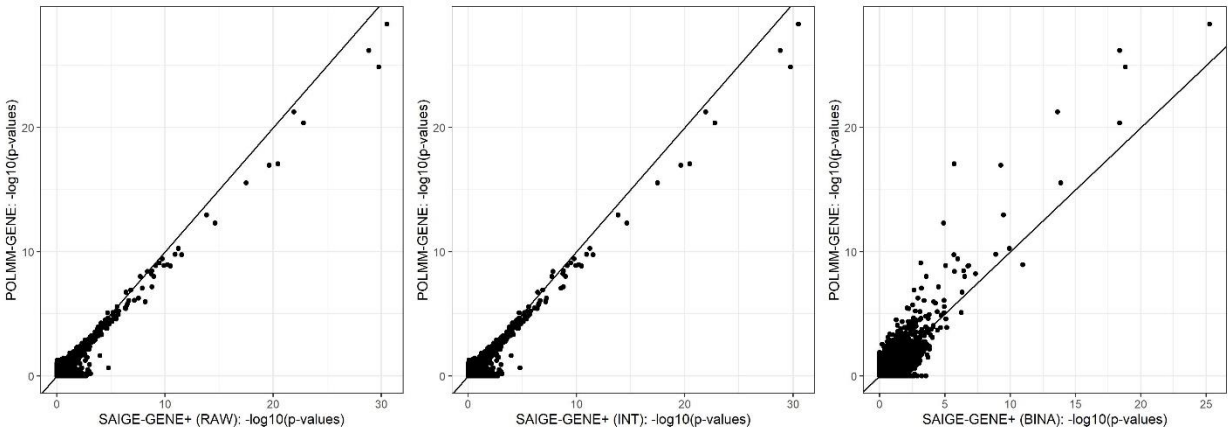

Table S1. Computation time and cost of the 5 ordinal categorical phenotypes analysis in UK Biobank RAP.

		Alcohol intake frequency		Morning/evening person (chronotype)		Comparative height size at age 10		Cognitive symptoms severity		Comparative body size at age 10	
chrom	instance type	Computation time	Cost	Computation time	Cost	Computation time	Cost	Computation time	Cost	Computation time	Cost
chr1	mem1_ssd2_v2_x4	14:05:53	£0.94	7:46:13	£0.52	7:45:53	£0.51	3:37:38	£0.24	7:58:04	£0.53
chr2	mem1_ssd2_v2_x5	9:35:13	£0.64	5:25:19	£0.36	5:26:52	£0.36	2:22:11	£0.16	6:16:15	£0.42
chr3	mem1_ssd2_v2_x4	8:05:40	£0.54	4:18:55	£0.28	4:14:07	£0.28	2:00:11	£0.13	4:29:32	£0.30
chr4	mem2_ssd2_v2_x2	12:24:50	£0.66	3:44:32	£0.12	3:51:08	£0.13	1:46:22	£0.06	3:55:12	£0.13
chr5	mem2_ssd2_v2_x2	7:51:12	£0.26	4:17:44	£0.14	4:08:02	£0.14	2:01:42	£0.07	4:10:49	£0.14
chr6	mem1_ssd2_v2_x4	7:13:12	£0.48	3:52:57	£0.26	3:42:33	£0.24	1:44:56	£0.11	3:44:54	£0.25
chr7	mem1_ssd2_v2_x4	6:44:47	£0.45	3:39:05	£0.24	3:46:44	£0.25	1:36:48	£0.10	3:42:59	£0.24
chr8	mem2_ssd2_v2_x2	6:30:03	£0.22	3:28:35	£0.11	3:27:20	£0.11	1:34:19	£0.05	3:23:22	£0.11
chr9	mem2_ssd2_v2_x2	7:48:58	£0.26	4:08:28	£0.14	4:19:22	£0.14	2:00:00	£0.07	4:19:49	£0.14
chr10	mem2_ssd2_v2_x2	6:48:18	£0.23	3:51:18	£0.13	3:45:04	£0.12	1:51:24	£0.06	3:54:31	£0.13
chr11	mem1_ssd2_v2_x4	9:20:37	£0.62	4:46:54	£0.32	4:54:08	£0.32	2:10:05	£0.14	4:51:50	£0.32
chr12	mem1_ssd2_v2_x4	7:45:54	£0.52	4:00:31	£0.26	3:52:35	£0.26	1:43:22	£0.11	3:47:06	£0.25
chr13	mem2_ssd2_v2_x2	3:27:18	£0.11	1:45:17	£0.06	1:42:03	£0.06	0:51:32	£0.03	1:43:48	£0.06
chr14	mem2_ssd2_v2_x2	6:04:14	£0.20	3:00:39	£0.10	3:02:43	£0.10	1:26:58	£0.05	3:00:47	£0.10
chr15	mem2_ssd2_v2_x2	6:18:19	£0.21	3:28:58	£0.11	3:21:35	£0.11	1:34:35	£0.05	3:14:50	£0.11
chr16	mem1_ssd2_v2_x4	7:14:18	£0.48	3:46:29	£0.25	3:47:20	£0.25	1:42:28	£0.11	3:41:26	£0.24
chr17	mem1_ssd2_v2_x4	8:30:11	£0.56	4:26:48	£0.29	4:24:21	£0.29	2:01:29	£0.13	4:36:51	£0.30
chr18	mem2_ssd2_v2_x2	2:54:58	£0.10	1:30:12	£0.05	1:33:30	£0.05	0:45:34	£0.02	1:39:17	£0.05
chr19	mem1_ssd2_v2_x4	11:04:40	£0.74	5:53:23	£0.39	5:39:49	£0.37	2:30:19	£0.16	6:50:59	£0.45
chr20	mem2_ssd2_v2_x2	4:51:36	£0.16	2:28:39	£0.08	2:26:22	£0.08	1:13:21	£0.04	2:30:49	£0.08
chr21	mem2_ssd2_v2_x2	2:03:10	£0.07	1:08:40	£0.04	1:05:50	£0.04	0:35:45	£0.02	1:08:04	£0.04
chr22	mem2_ssd2_v2_x2	4:30:42	£0.15	2:14:31	£0.07	2:17:47	£0.08	1:07:09	£0.04	2:18:44	£0.08
Total			£8.60		£4.32		£4.29		£1.95		£4.47
The allocated instance: "on-demand" for job of chr4 to analyze "Alcohol intake frequency", "spot" for the other jobs											